General Data Analyst Interview Questions

In an interview, these questions are more likely to appear early in the process and cover data analysis at a high level.

1. Mention the differences between Data Mining and Data Profiling?

Data Mining	Data Profiting
Data mining is the process of discovering relevant information that has not yet been identified before.	Data profiling is done to evaluate a dataset for its uniqueness, logic, and consistency.
In data mining, raw data is converted into valuable information.	It cannot identify inaccurate or incorrect data values.

2. Define the term 'Data Wrangling in Data Analytics.

Data Wrangling is the process wherein raw data is cleaned, structured, and enriched into a desired usable format for better decision making. It involves discovering, structuring, cleaning, enriching, validating, and analyzing data. This process can turn and map out large amounts of data extracted from various sources into a more useful format. Techniques such as merging, grouping, concatenating, joining, and sorting are used to analyze the data. Thereafter it gets ready to be used with another dataset.

3. What are the various steps involved in any analytics project?

This is one of the most basic data analyst interview questions. The various steps involved in any common analytics projects are as follows:

Understanding the Problem

Understand the business problem, define the organizational goals, and plan for a lucrative solution.

Collecting Data

Gather the right data from various sources and other information based on your priorities.

Cleaning Data

Clean the data to remove unwanted, redundant, and missing values, and make it ready for analysis.

Exploring and Analyzing Data

Use data visualization and business intelligence tools, data mining techniques, and predictive modeling to analyze data.

Interpreting the Results

Interpret the results to find out hidden patterns, future trends, and gain insights.

4. What are the common problems that data analysts encounter during analysis?

The common problems steps involved in any analytics project are:

- Handling duplicate
- Collecting the meaningful right data and the right time
- Handling data purging and storage problems
- Making data secure and dealing with compliance issues

5. Which are the technical tools that you have used for analysis and presentation purposes?

As a data analyst, you are expected to know the tools mentioned below for analysis and presentation purposes. Some of the popular tools you should know are:

MS SQL Server, MySQL

For working with data stored in relational databases

MS Excel, Tableau

For creating reports and dashboards

Python, R, SPSS

For statistical analysis, data modeling, and exploratory analysis

MS PowerPoint

For presentation, displaying the final results and important conclusions

6. What are the best methods for data cleaning?

- Create a data cleaning plan by understanding where the common errors take place and keep all the communications open.
- Before working with the data, identify and remove the duplicates. This will lead to an easy and effective data analysis process.
- Focus on the accuracy of the data. Set cross-field validation, maintain the value types of data, and provide mandatory constraints.
- Normalize the data at the entry point so that it is less chaotic. You will be able to ensure that all information is standardized, leading to fewer errors on entry.

7. What is the significance of Exploratory Data Analysis (EDA)?

- Exploratory data analysis (EDA) helps to understand the data better.
- It helps you obtain confidence in your data to a point where you're ready to engage a machine learning algorithm.
- It allows you to refine your selection of feature variables that will be used later for model building.
- You can discover hidden trends and insights from the data.

8. Explain descriptive, predictive, and prescriptive analytics.

Descriptive	Predictive	Prescriptive
It provides insights into the past to answer "what has happened"	Understands the future to answer "what could happen"	Suggest various courses of action to answer "what should you do"
Uses data aggregation and data mining techniques	Uses statistical models and forecasting techniques	Uses simulation algorithms and optimization techniques to advise possible outcomes
Example: An ice cream company can analyze	Example: An ice cream company can analyze	Example: Lower prices to increase the sale of ice creams,

how much ice cream was sold, which flavors were sold, and whether more or less ice cream was sold than the day before	how much ice cream was sold, which flavors were sold, and whether more or less ice cream was sold than the day before	produce more/fewer quantities of a specific flavor of ice cream

9. What are the different types of sampling techniques used by data analysts?

Sampling is a statistical method to select a subset of data from an entire dataset (population) to estimate the characteristics of the whole population.

There are majorly five types of sampling methods:

- Simple random sampling
- Systematic sampling
- Cluster sampling
- Stratified sampling
- Judgmental or purposive sampling

10. Describe univariate, bivariate, and multivariate analysis.

Univariate analysis is the simplest and easiest form of data analysis where the data being analyzed contains only one variable.

Example - Studying the heights of players in the NBA.

Univariate analysis can be described using Central Tendency, Dispersion, Quartiles, Bar charts, Histograms, Pie charts, and Frequency distribution tables.

The bivariate analysis involves the analysis of two variables to find causes, relationships, and correlations between the variables.

Example – Analyzing the sale of ice creams based on the temperature outside.

The bivariate analysis can be explained using Correlation coefficients, Linear regression, Logistic regression, Scatter plots, and Box plots.

The multivariate analysis involves the analysis of three or more variables to understand the relationship of each variable with the other variables.

Example – Analysing Revenue based on expenditure.

Multivariate analysis can be performed using Multiple regression, Factor analysis, Classification & regression trees, Cluster analysis, Principal component analysis, Dual-axis charts, etc.

11. What are your strengths and weaknesses as a data analyst?

The answer to this question may vary from a case to case basis. However, some general strengths of a data analyst may include strong analytical skills, attention to detail, proficiency in data manipulation and visualization, and the ability to derive insights from complex datasets. Weaknesses could include limited domain knowledge, lack of experience with certain data analysis tools or techniques, or challenges in effectively communicating technical findings to non-technical stakeholders.

12. What are the ethical considerations of data analysis?

Some of the most the ethical considerations of data analysis includes:

- Privacy: Safeguarding the privacy and confidentiality of individuals' data, ensuring compliance with applicable privacy laws and regulations.
- Informed Consent: Obtaining informed consent from individuals whose data is being analyzed, explaining the purpose and potential implications of the analysis.
- Data Security: Implementing robust security measures to protect data from unauthorized access, breaches, or misuse.
- Data Bias: Being mindful of potential biases in data collection, processing, or interpretation that may lead to unfair or discriminatory outcomes.
- Transparency: Being transparent about the data analysis methodologies, algorithms, and models used, enabling stakeholders to understand and assess the results.
- Data Ownership and Rights: Respecting data ownership rights and intellectual property, using data only within the boundaries of legal permissions or agreements.
- Accountability: Taking responsibility for the consequences of data analysis, ensuring that actions based on the analysis are fair, just, and beneficial to individuals and society.
- Data Quality and Integrity: Ensuring the accuracy, completeness, and reliability of data used in the analysis to avoid misleading or incorrect conclusions.
- Social Impact: Considering the potential social impact of data analysis results, including potential unintended consequences or negative effects on marginalized groups.
- Compliance: Adhering to legal and regulatory requirements related to data analysis, such as data protection laws, industry standards, and ethical guidelines.

13. What are some common data visualization tools you have used?

You should name the tools you have used personally, however here's a list of the commonly used data visualization tools in the industry:

- Tableau
- Microsoft Power BI

- QlikView
- Google Data Studio
- Plotly
- Matplotlib (Python library)
- Excel (with built-in charting capabilities)
- SAP Lumira
- IBM Cognos Analytics

Data Analyst Interview Questions On Statistics

14. How can you handle missing values in a dataset?

This is one of the most frequently asked data analyst interview questions, and the interviewer expects you to give a detailed answer here, and not just the name of the methods. There are four methods to handle missing values in a dataset.

Listwise Deletion

In the listwise deletion method, an entire record is excluded from analysis if any single value is missing.

Average Imputation

Take the average value of the other participants' responses and fill in the missing value.

Regression Substitution

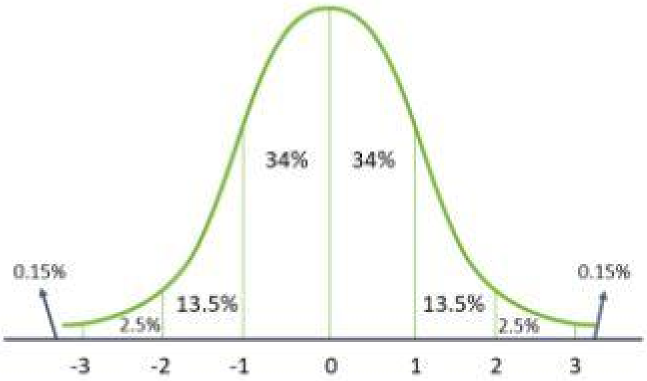
You can use multiple-regression analyses to estimate a missing value.

Multiple Imputations

It creates plausible values based on the correlations for the missing data and then averages the simulated datasets by incorporating random errors in your predictions.

15. Explain the term Normal Distribution.

Normal Distribution refers to a continuous probability distribution that is symmetric about the mean. In a graph, normal distribution will appear as a bell curve.



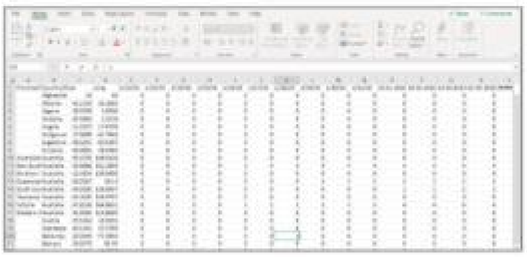
- The mean, median, and mode are equal
- All of them are located in the center of the distribution
- 68% of the data falls within one standard deviation of the mean
- 95% of the data lies between two standard deviations of the mean
- 99.7% of the data lies between three standard deviations of the mean

16. What is Time Series analysis?

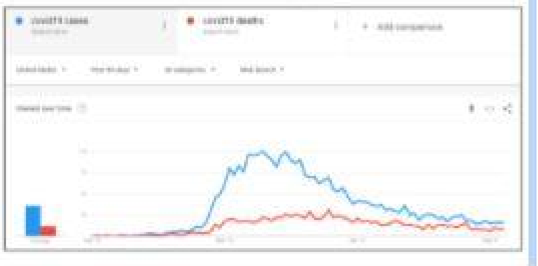
Time Series analysis is a statistical procedure that deals with the ordered sequence of values of a variable at equally spaced time intervals. Time series data are collected at adjacent periods. So, there is a correlation between the observations. This feature distinguishes time-series data from cross-sectional data.

Below is an example of time-series data on coronavirus cases and its graph.

Time series data on Covid19 cases



Time series graph

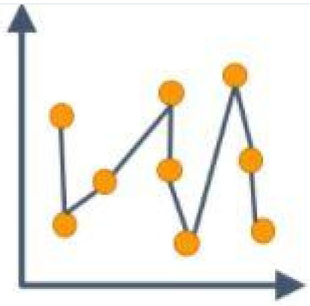


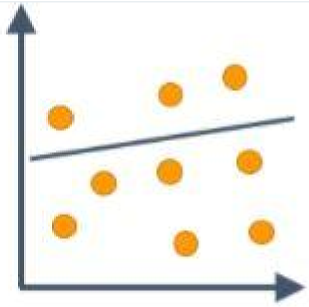
17. How is Overfitting different from Underfitting?

This is another frequently asked data analyst interview question, and you are expected to cover all the given differences!

Overfitting	Underfitting
The model trains the data well using the training set.	Here, the model neither trains the data well nor can generalize to new data.
The performance drops considerably over the test set.	Performs poorly both on the train and the test set.

Happens when the model learns the random fluctuations and noise in the training dataset in detail.	This happens when there is lesser data to build an accurate model and when we try to develop a linear model using non-linear data.

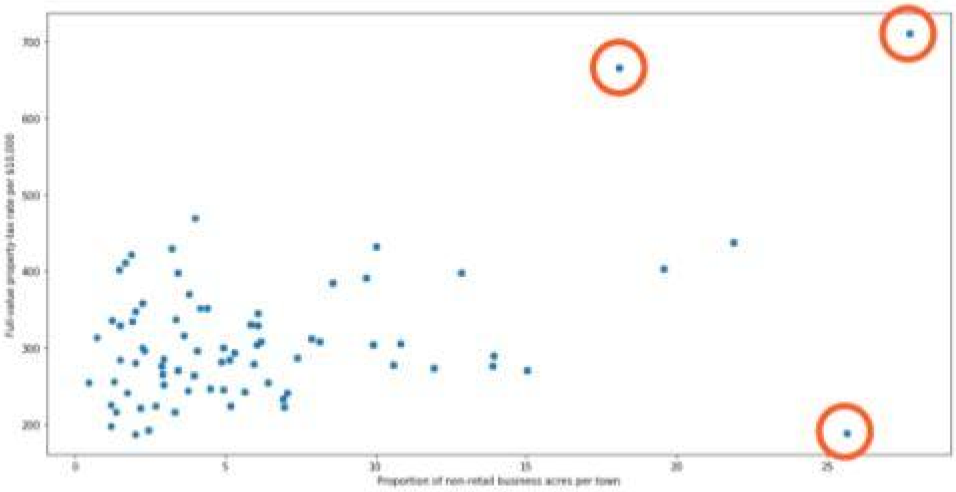




18. How do you treat outliers in a dataset?

An outlier is a data point that is distant from other similar points. They may be due to variability in the measurement or may indicate experimental errors.

The graph depicted below shows there are three outliers in the dataset.



To deal with outliers, you can use the following four methods:

- Drop the outlier records
- Cap your outliers data
- Assign a new value
- Try a new transformation

19. What are the different types of Hypothesis testing?

Hypothesis testing is the procedure used by statisticians and scientists to accept or reject statistical hypotheses. There are mainly two types of hypothesis testing:

- Null hypothesis: It states that there is no relation between the predictor and outcome variables in the population. H0 denoted it.

Example: There is no association between a patient's BMI and diabetes.

- Alternative hypothesis: It states that there is some relation between the predictor and outcome variables in the population. It is denoted by H1.

Example: There could be an association between a patient's BMI and diabetes.

20. Explain the Type I and Type II errors in Statistics?

In Hypothesis testing, a Type I error occurs when the null hypothesis is rejected even if it is true. It is also known as a false positive.

A Type II error occurs when the null hypothesis is not rejected, even if it is false. It is also known as a false negative.

21. How would you handle missing data in a dataset?

Ans: The choice of handling technique depends on factors such as the amount and nature of missing data, the underlying analysis, and the assumptions made. It's crucial to exercise caution and carefully consider the implications of the chosen approach to ensure the integrity and reliability of the data analysis. However, a few solutions could be:

- removing the missing observations or variables
- imputation methods including, mean imputation (replacing missing values with the mean of the available data), median imputation (replacing missing values with the median), or regression imputation (predicting missing values based on regression models)
- sensitivity analysis

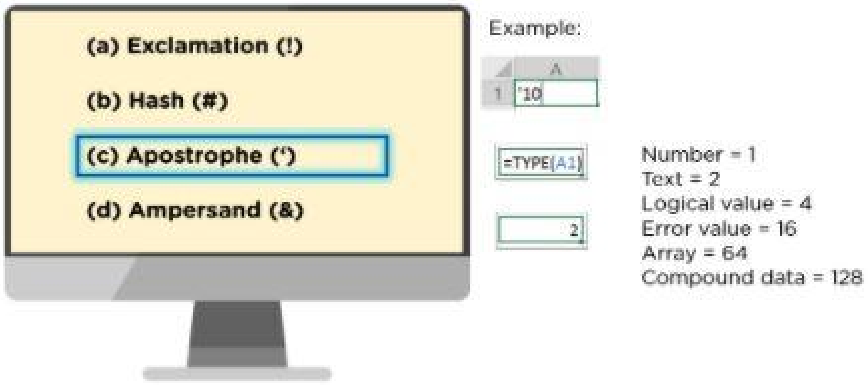
22. Explain the concept of outlier detection and how you would identify outliers in a dataset.

Outlier detection is the process of identifying observations or data points that significantly deviate from the expected or normal behavior of a dataset. Outliers can be valuable sources of information or indications of anomalies, errors, or rare events.

It's important to note that outlier detection is not a definitive process, and the identified outliers should be further investigated to determine their validity and potential impact on the analysis or model. Outliers can be due to various reasons, including data entry errors, measurement errors, or genuinely anomalous observations, and each case requires careful consideration and interpretation.

Excel Data Analyst Interview Questions

23. In Microsoft Excel, a numeric value can be treated as a text value if it precedes with what?



(a) Exclamation (!)

(b) Hash (#)

(c) Apostrophe (')

(d) Ampersand (&)

Number = 1
Text = 2
Logical value = 4
Error value = 16
Array = 64
Compound data = 128

24. What is the difference between COUNT, COUNTA, COUNTBLANK, and COUNTIF in Excel?

- COUNT function returns the count of numeric cells in a range
- COUNTA function counts the non-blank cells in a range
- COUNTBLANK function gives the count of blank cells in a range
- COUNTIF function returns the count of values by checking a given condition

25. How do you make a dropdown list in MS Excel?

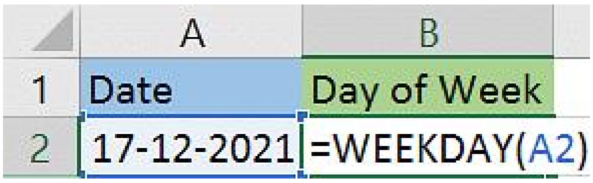
- First, click on the Data tab that is present in the ribbon.
- Under the Data Tools group, select Data Validation.
- Then navigate to Settings > Allow > List.
- Select the source you want to provide as a list array.

26. Can you provide a dynamic range in "Data Source" for a Pivot table?

Yes, you can provide a dynamic range in the "Data Source" of Pivot tables. To do that, you need to create a named range using the offset function and base the pivot table using a named range constructed in the first step.

27. What is the function to find the day of the week for a particular date value?

The get the day of the week, you can use the WEEKDAY() function.



	A	B
1	Date	Day of Week
2	17-12-2021	=WEEKDAY(A2)

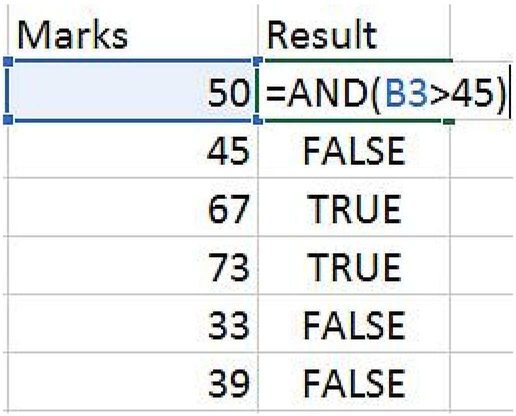
The above function will return 6 as the result, i.e., 17th December is a Saturday.

28. How does the AND() function work in Excel?

AND() is a logical function that checks multiple conditions and returns TRUE or FALSE based on whether the conditions are met.

Syntax: AND(logica1,[logical2],[logical3]....)

In the below example, we are checking if the marks are greater than 45. The result will be true if the mark is >45, else it will be false.



Marks	Result
50	=AND(B3>45)
45	FALSE
67	TRUE
73	TRUE
33	FALSE
39	FALSE

29. Explain how VLOOKUP works in Excel?

VLOOKUP is used when you need to find things in a table or a range by row.

VLOOKUP accepts the following four parameters:

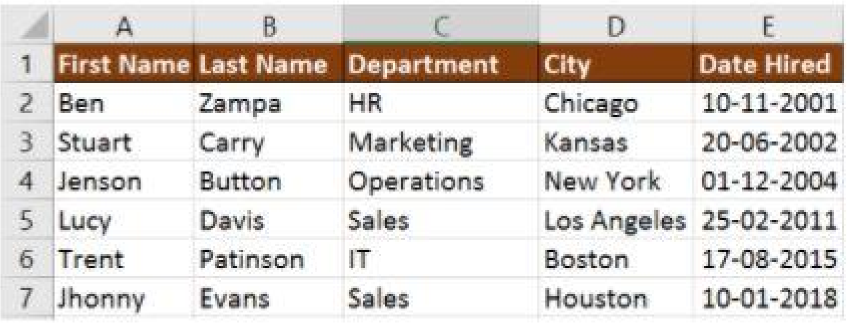
lookup_value - The value to look for in the first column of a table

table - The table from where you can extract value

col_index - The column from which to extract value

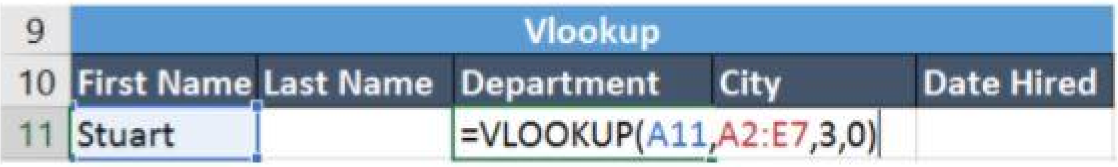
range_lookup - [optional] TRUE = approximate match (default). FALSE = exact match

Let's understand VLOOKUP with an example.



	A	B	C	D	E
1	First Name	Last Name	Department	City	Date Hired
2	Ben	Zampa	HR	Chicago	10-11-2001
3	Stuart	Carry	Marketing	Kansas	20-06-2002
4	Jenson	Button	Operations	New York	01-12-2004
5	Lucy	Davis	Sales	Los Angeles	25-02-2011
6	Trent	Patinson	IT	Boston	17-08-2015
7	Jhonny	Evans	Sales	Houston	10-01-2018

If you wanted to find the department to which Stuart belongs to, you could use the VLOOKUP function as shown below:



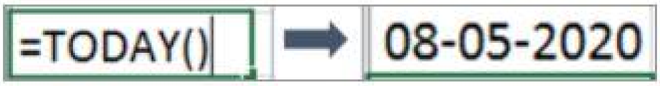
9	Vlookup				
10	First Name	Last Name	Department	City	Date Hired
11	Stuart		=VLOOKUP(A11,A2:E7,3,0)		

Here, A11 cell has the lookup value, A2:E7 is the table array, 3 is the column index number with information about departments, and 0 is the range lookup.

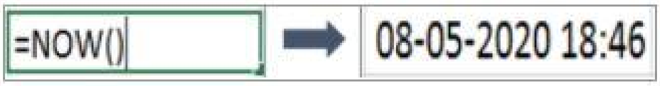
If you hit enter, it will return "Marketing", indicating that Stuart is from the marketing department.

30. What function would you use to get the current date and time in Excel?

In Excel, you can use the TODAY() and NOW() function to get the current date and time.

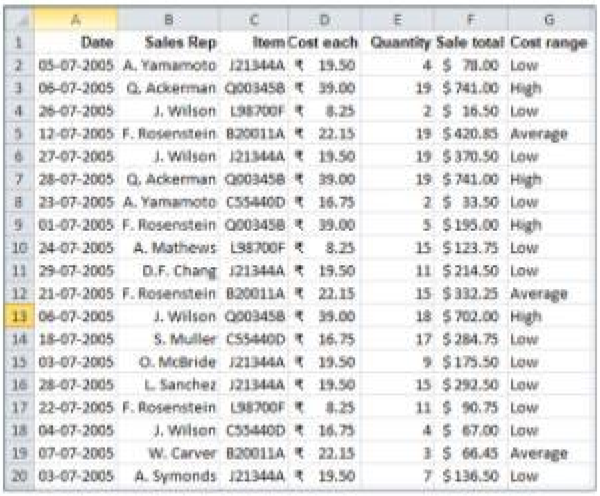


=TODAY() → 08-05-2020



=NOW() → 08-05-2020 18:46

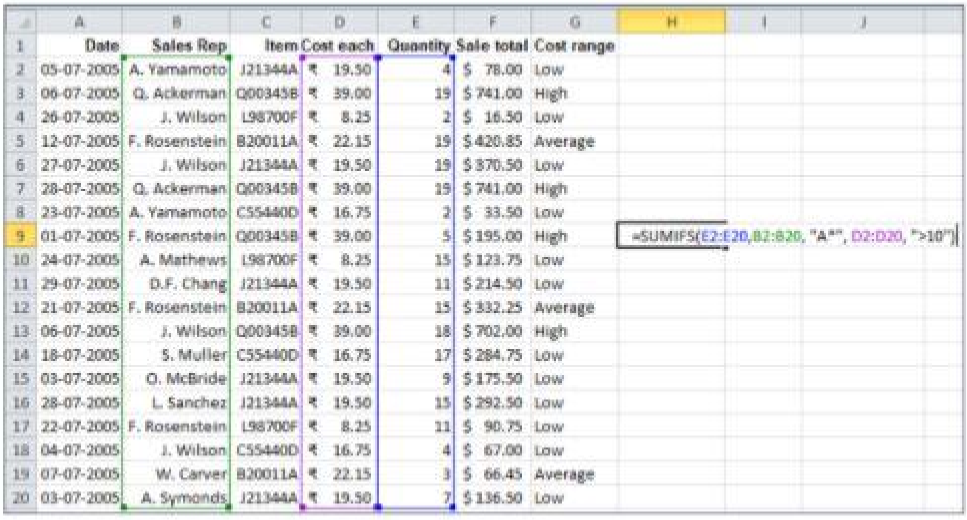
31. Using the below sales table, calculate the total quantity sold by sales representatives whose name starts with A, and the cost of each item they have sold is greater than 10.



	A	B	C	D	E	F	G
1	Date	Sales Rep	Item	Cost each	Quantity	Sale total	Cost range
2	05-07-2005	A. Yamamoto	J21344A	₹ 19.50	4	$ 78.00	Low
3	06-07-2005	Q. Ackerman	Q003458	₹ 39.00	19	$ 741.00	High
4	26-07-2005	J. Wilson	L98700F	₹ 8.25	2	$ 16.50	Low
5	12-07-2005	F. Rosenstein	B20011A	₹ 22.15	19	$ 420.85	Average
6	27-07-2005	J. Wilson	J21344A	₹ 19.50	19	$ 370.50	Low
7	28-07-2005	Q. Ackerman	Q003458	₹ 39.00	19	$ 741.00	High
8	23-07-2005	A. Yamamoto	C55440D	₹ 16.75	2	$ 33.50	Low
9	01-07-2005	F. Rosenstein	Q003458	₹ 39.00	5	$ 195.00	High
10	24-07-2005	A. Mathews	L98700F	₹ 8.25	15	$ 123.75	Low
11	29-07-2005	D.F. Chang	J21344A	₹ 19.50	11	$ 214.50	Low
12	21-07-2005	F. Rosenstein	B20011A	₹ 22.15	15	$ 332.25	Average
13	06-07-2005	J. Wilson	Q003458	₹ 39.00	18	$ 702.00	High
14	18-07-2005	S. Muller	C55440D	₹ 16.75	17	$ 284.75	Low
15	03-07-2005	O. McBride	J21344A	₹ 19.50	9	$ 175.50	Low
16	28-07-2005	L. Sanchez	J21344A	₹ 19.50	15	$ 292.50	Low
17	22-07-2005	F. Rosenstein	L98700F	₹ 8.25	11	$ 90.75	Low
18	04-07-2005	J. Wilson	C55440D	₹ 16.75	4	$ 67.00	Low
19	07-07-2005	W. Carver	B20011A	₹ 22.15	3	$ 66.45	Average
20	03-07-2005	A. Symonds	J21344A	₹ 19.50	7	$ 136.50	Low

You can use the SUMIFS() function to find the total quantity.

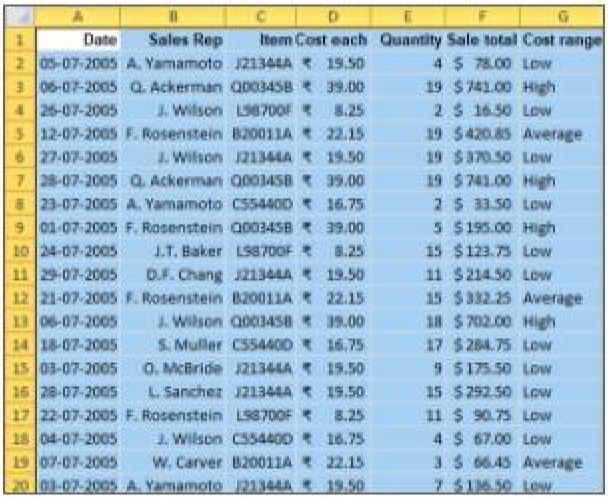
For the Sales Rep column, you need to give the criteria as “A*” - meaning the name should start with the letter “A”. For the Cost each column, the criteria should be “>10” - meaning the cost of each item is greater than 10.



	A	B	C	D	E	F	G	H	I	J
1	Date	Sales Rep	Item	Cost each	Quantity	Sale total	Cost range			
2	05-07-2005	A. Yamamoto	J21344A	₹ 19.50	4	$ 78.00	Low			
3	06-07-2005	Q. Ackerman	Q00345B	₹ 39.00	19	$ 741.00	High			
4	26-07-2005	J. Wilson	L98700F	₹ 8.25	2	$ 16.50	Low			
5	12-07-2005	F. Rosenstein	B20011A	₹ 22.15	19	$ 420.85	Average			
6	27-07-2005	J. Wilson	J21344A	₹ 19.50	19	$ 370.50	Low			
7	28-07-2005	Q. Ackerman	Q00345B	₹ 39.00	19	$ 741.00	High			
8	23-07-2005	A. Yamamoto	C55440D	₹ 16.75	2	$ 33.50	Low			
9	01-07-2005	F. Rosenstein	Q00345B	₹ 39.00	5	$ 195.00	High	=SUMIFS(E2:E20,B2:B20, "A*", D2:D20, ">10")		
10	24-07-2005	A. Mathews	L98700F	₹ 8.25	15	$ 123.75	Low			
11	29-07-2005	D.F. Chang	J21344A	₹ 19.50	11	$ 214.50	Low			
12	21-07-2005	F. Rosenstein	B20011A	₹ 22.15	15	$ 332.25	Average			
13	06-07-2005	J. Wilson	Q00345B	₹ 39.00	18	$ 702.00	High			
14	18-07-2005	S. Muller	C55440D	₹ 16.75	17	$ 284.75	Low			
15	03-07-2005	O. McBride	J21344A	₹ 19.50	9	$ 175.50	Low			
16	28-07-2005	L. Sanchez	J21344A	₹ 19.50	15	$ 292.50	Low			
17	22-07-2005	F. Rosenstein	L98700F	₹ 8.25	11	$ 90.75	Low			
18	04-07-2005	J. Wilson	C55440D	₹ 16.75	4	$ 67.00	Low			
19	07-07-2005	W. Carver	B20011A	₹ 22.15	3	$ 66.45	Average			
20	03-07-2005	A. Symonds	J21344A	₹ 19.50	7	$ 136.50	Low			

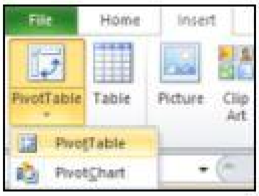
The result is 13.

33. Using the data given below, create a pivot table to find the total sales made by each sales representative for each item. Display the sales as % of the grand total.

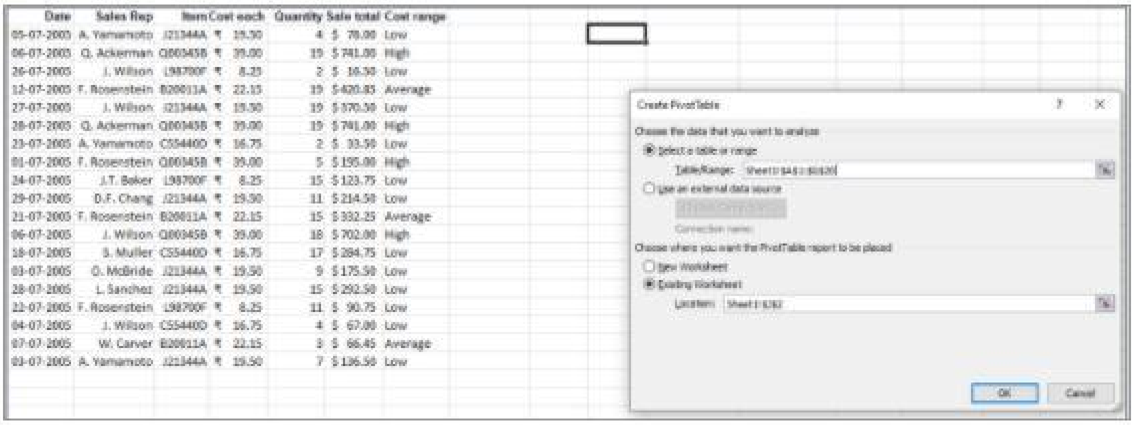


	A	B	C	D	E	F	G
1	Date	Sales Rep	Item	Cost each	Quantity	Sale total	Cost range
2	05-07-2005	A. Yamamoto	J21344A	₹ 19.50	4	$ 78.00	Low
3	06-07-2005	Q. Ackerman	Q00345B	₹ 39.00	19	$ 741.00	High
4	26-07-2005	J. Wilson	L98700F	₹ 8.25	2	$ 16.50	Low
5	12-07-2005	F. Rosenstein	B20011A	₹ 22.15	19	$ 420.85	Average
6	27-07-2005	J. Wilson	J21344A	₹ 19.50	19	$ 370.50	Low
7	28-07-2005	Q. Ackerman	Q00345B	₹ 39.00	19	$ 741.00	High
8	23-07-2005	A. Yamamoto	C55440D	₹ 16.75	2	$ 33.50	Low
9	01-07-2005	F. Rosenstein	Q00345B	₹ 39.00	5	$ 195.00	High
10	24-07-2005	J.T. Baker	L98700F	₹ 8.25	15	$ 123.75	Low
11	29-07-2005	D.F. Chang	J21344A	₹ 19.50	11	$ 214.50	Low
12	21-07-2005	F. Rosenstein	B20011A	₹ 22.15	15	$ 332.25	Average
13	06-07-2005	J. Wilson	Q00345B	₹ 39.00	18	$ 702.00	High
14	18-07-2005	S. Muller	C55440D	₹ 16.75	17	$ 284.75	Low
15	03-07-2005	O. McBride	J21344A	₹ 19.50	9	$ 175.50	Low
16	28-07-2005	L. Sanchez	J21344A	₹ 19.50	15	$ 292.50	Low
17	22-07-2005	F. Rosenstein	L98700F	₹ 8.25	11	$ 90.75	Low
18	04-07-2005	J. Wilson	C55440D	₹ 16.75	4	$ 67.00	Low
19	07-07-2005	W. Carver	B20011A	₹ 22.15	3	$ 66.45	Average
20	03-07-2005	A. Yamamoto	J21344A	₹ 19.50	7	$ 136.50	Low

- Select the entire table range, click on the Insert tab and choose PivotTable



- Select the table range and the worksheet where you want to place the pivot table



Date	Sales Rep	Item	Cost each	Quantity	Sale total	Cost range
05-07-2005	A. Yamamoto	J21344A	₹ 19.50	4	$ 78.00	Low
06-07-2005	Q. Ackerman	Q00345B	₹ 39.00	19	$ 741.00	High
26-07-2005	J. Wilson	L98700F	₹ 8.25	2	$ 16.50	Low
12-07-2005	F. Rosenstein	B20011A	₹ 22.15	19	$ 420.85	Average
27-07-2005	J. Wilson	J21344A	₹ 19.50	19	$ 370.50	Low
28-07-2005	Q. Ackerman	Q00345B	₹ 39.00	19	$ 741.00	High
23-07-2005	A. Yamamoto	C55440D	₹ 16.75	2	$ 33.50	Low
01-07-2005	F. Rosenstein	Q00345B	₹ 39.00	5	$ 195.00	High
24-07-2005	J.T. Baker	L98700F	₹ 8.25	15	$ 123.75	Low
29-07-2005	D.F. Chang	J21344A	₹ 19.50	11	$ 214.50	Low
21-07-2005	F. Rosenstein	B20011A	₹ 22.15	15	$ 332.25	Average
06-07-2005	J. Wilson	Q00345B	₹ 39.00	18	$ 702.00	High
18-07-2005	S. Muller	C55440D	₹ 16.75	17	$ 284.75	Low
03-07-2005	O. McBride	J21344A	₹ 19.50	9	$ 175.50	Low
28-07-2005	L. Sanchez	J21344A	₹ 19.50	15	$ 292.50	Low
22-07-2005	F. Rosenstein	L98700F	₹ 8.25	11	$ 90.75	Low
04-07-2005	J. Wilson	C55440D	₹ 16.75	4	$ 67.00	Low
07-07-2005	W. Carver	B20011A	₹ 22.15	3	$ 66.45	Average
03-07-2005	A. Yamamoto	J21344A	₹ 19.50	7	$ 136.50	Low

Create PivotTable

Choose the data that you want to analyze

● Select a table or range

Table/Range: Sheet1!$A$1:$G$20

○ Use an external data source

Connection name:

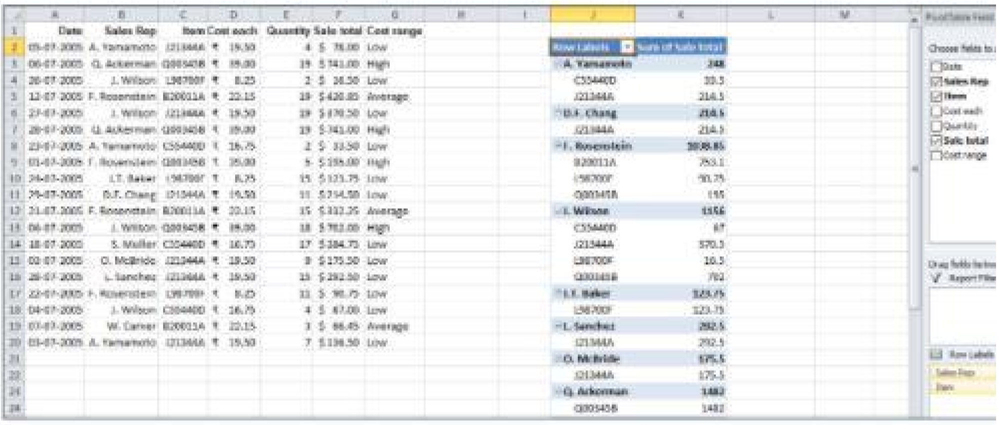
Choose where you want the PivotTable report to be placed

○ New Worksheet

● Existing Worksheet

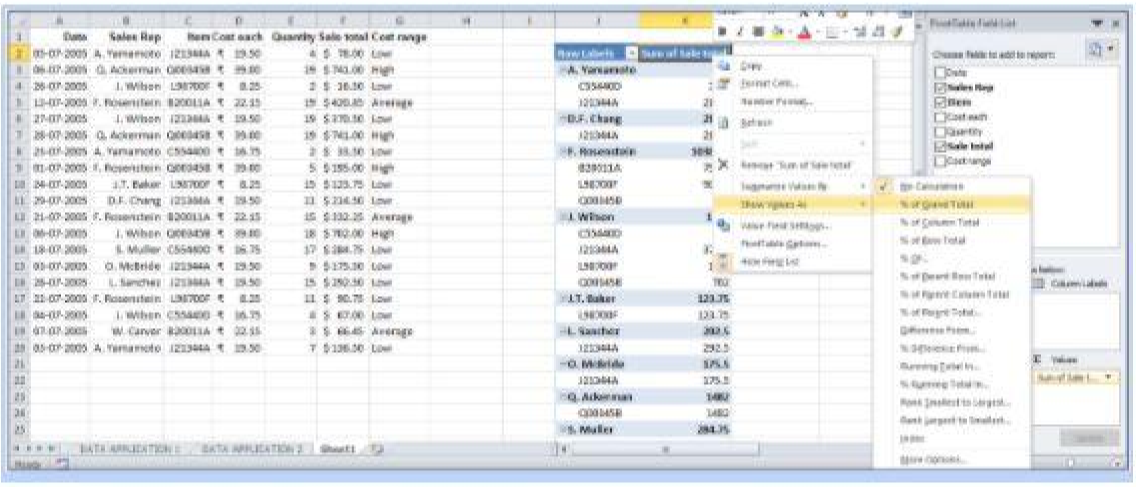
Location: Sheet1!$J$2

- Drag Sale total on to Values, and Sales Rep and Item on to Row Labels. It will give the sum of sales made by each representative for every item they have sold.



Row Labels	Sum of Sale total
A. Yamamoto	248
C55440D	33.5
J21344A	214.5
D.F. Chang	214.5
J21344A	214.5
F. Rosenstein	1038.85
B20011A	753.1
L98700F	90.75
Q00345B	195
J. Wilson	1156
C55440D	67
J21344A	370.5
L98700F	16.5
Q00345B	702
J.T. Baker	123.75
L98700F	123.75
L. Sanchez	292.5
J21344A	292.5
O. McBride	175.5
J21344A	175.5
Q. Ackerman	1482
Q00345B	1482

- Right-click on “Sum of Sale Total’ and expand Show Values As to select % of Grand Total.



- Below is the resultant pivot table.

Date	Sales Rep	Item	Cost each	Quantity	Sale total	Cost range
05-07-2005	A. Yamamoto	J21344A	₹ 19.50	4	$ 78.00	Low
06-07-2005	Q. Ackerman	Q003458	₹ 39.00	19	$ 741.00	High
26-07-2005	J. Wilson	L98706F	₹ 8.25	2	$ 16.50	Low
12-07-2005	F. Rosenstein	B20011A	₹ 22.15	19	$ 420.85	Average
27-07-2005	J. Wilson	J21344A	₹ 19.50	19	$ 370.50	Low
28-07-2005	Q. Ackerman	Q003458	₹ 39.00	19	$ 741.00	High
23-07-2005	A. Yamamoto	C55440D	₹ 16.75	2	$ 33.50	Low
01-07-2005	F. Rosenstein	Q003458	₹ 39.00	5	$ 195.00	High
24-07-2005	J.T. Baker	L98706F	₹ 8.25	15	$ 123.75	Low
29-07-2005	D.F. Chang	J21344A	₹ 19.50	11	$ 214.50	Low
21-07-2005	F. Rosenstein	B20011A	₹ 22.15	15	$ 332.25	Average
08-07-2005	J. Wilson	Q003458	₹ 39.00	18	$ 702.00	High
18-07-2005	S. Muller	C55440D	₹ 16.75	17	$ 284.75	Low
03-07-2005	O. McBride	J21344A	₹ 19.50	9	$ 175.50	Low
28-07-2005	L. Sanchez	J21344A	₹ 19.50	15	$ 292.50	Low
22-07-2005	F. Rosenstein	L98706F	₹ 8.25	11	$ 90.75	Low
04-07-2005	J. Wilson	C55440D	₹ 16.75	4	$ 67.00	Low
07-07-2005	W. Carver	B20011A	₹ 22.15	3	$ 66.45	Average
03-07-2005	A. Yamamoto	J21344A	₹ 19.50	7	$ 136.50	Low

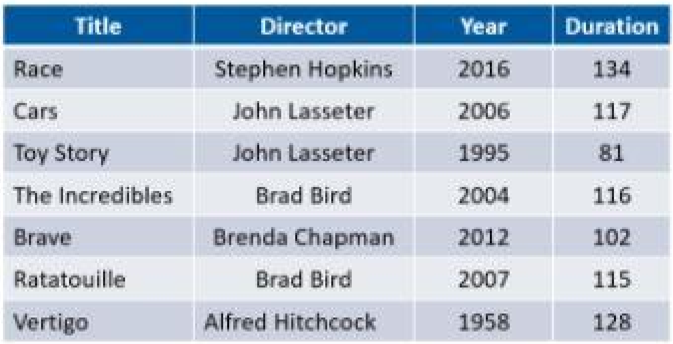
Row Labels	Sum of Sale total
A. Yamamoto	4.88%
C55440D	0.66%
J21344A	4.22%
D.F. Chang	4.22%
J21344A	4.22%
F. Rosenstein	20.44%
B20011A	14.82%
L98706F	1.79%
Q003458	3.84%
J. Wilson	22.75%
C55440D	1.32%
J21344A	7.29%
L98706F	0.32%
Q003458	13.81%
J.T. Baker	2.43%
L98706F	2.43%
L. Sanchez	5.76%
J21344A	5.76%
O. McBride	3.45%
J21344A	3.45%
Q. Ackerman	29.16%
Q003458	29.16%

SQL Interview Questions for Data Analysts

34. How do you subset or filter data in SQL?

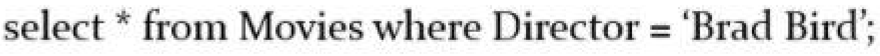
To subset or filter data in SQL, we use WHERE and HAVING clauses.

Consider the following movie table.

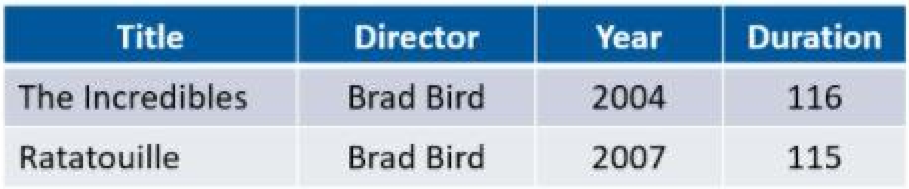


Title	Director	Year	Duration
Race	Stephen Hopkins	2016	134
Cars	John Lasseter	2006	117
Toy Story	John Lasseter	1995	81
The Incredibles	Brad Bird	2004	116
Brave	Brenda Chapman	2012	102
Ratatouille	Brad Bird	2007	115
Vertigo	Alfred Hitchcock	1958	128

Using this table, let’s find the records for movies that were directed by Brad Bird.

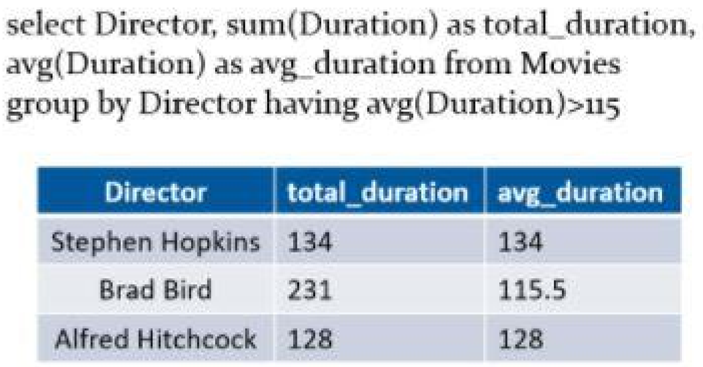


```
select * from Movies where Director = ‘Brad Bird’;
```



Title	Director	Year	Duration
The Incredibles	Brad Bird	2004	116
Ratatouille	Brad Bird	2007	115

Now, let’s filter the table for directors whose movies have an average duration greater than 115 minutes.



```
select Director, sum(Duration) as total_duration,
avg(Duration) as avg_duration from Movies
group by Director having avg(Duration)>115
```

Director	total_duration	avg_duration
Stephen Hopkins	134	134
Brad Bird	231	115.5
Alfred Hitchcock	128	128

35. What is the difference between a WHERE clause and a HAVING clause in SQL?

Answer all of the given differences when this data analyst interview question is asked, and also give out the syntax for each to prove your thorough knowledge to the interviewer.

WHERE	HAVING
WHERE clause operates on row data.	The HAVING clause operates on aggregated data.
In the WHERE clause, the filter occurs before any groupings are made.	HAVING is used to filter values from a group.

Aggregate functions cannot be used.	Aggregate functions can be used.

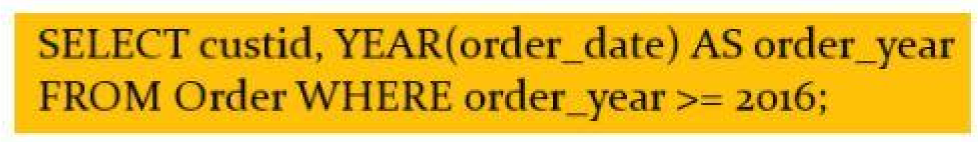
Syntax of WHERE clause:

```
SELECT column1, column2, ...
FROM table_name
WHERE condition;
```

Syntax of HAVING clause;

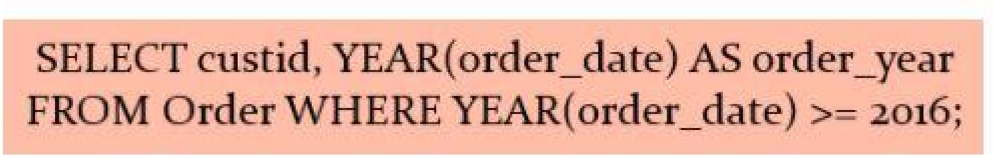
```
SELECT column_name(s)
FROM table_name
WHERE condition
GROUP BY column_name(s)
HAVING condition
ORDER BY column_name(s);
```

36. Is the below SQL query correct? If not, how will you rectify it?



```
SELECT custid, YEAR(order_date) AS order_year
FROM Order WHERE order_year >= 2016;
```

The query stated above is incorrect as we cannot use the alias name while filtering data using the WHERE clause. It will throw an error.



```
SELECT custid, YEAR(order_date) AS order_year
FROM Order WHERE YEAR(order_date) >= 2016;
```

37. How are Union, Intersect, and Except used in SQL?

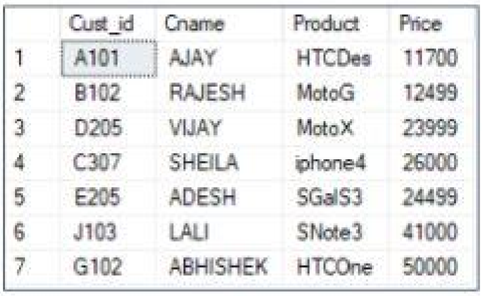
The Union operator combines the output of two or more SELECT statements.

Syntax:

```
SELECT column_name(s) FROM table1
UNION
SELECT column_name(s) FROM table2;
```

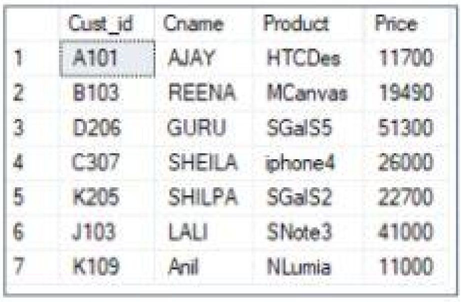
Let’s consider the following example, where there are two tables - Region 1 and Region 2.

Region 1



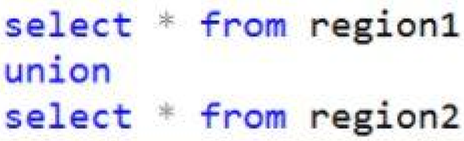
	Cust_id	Cname	Product	Price
1	A101	AJAY	HTCDes	11700
2	B102	RAJESH	MotoG	12499
3	D205	VIJAY	MotoX	23999
4	C307	SHEILA	iphone4	26000
5	E205	ADESH	SGalS3	24499
6	J103	LALI	SNote3	41000
7	G102	ABHISHEK	HTCOne	50000

Region 2

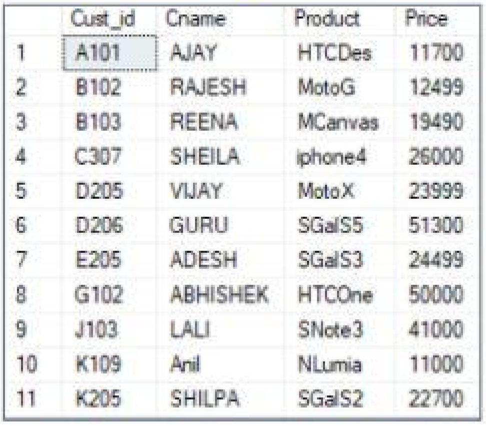


	Cust_id	Cname	Product	Price
1	A101	AJAY	HTCDes	11700
2	B103	REENA	MCanvas	19490
3	D206	GURU	SGalS5	51300
4	C307	SHEILA	iphone4	26000
5	K205	SHILPA	SGalS2	22700
6	J103	LALI	SNote3	41000
7	K109	Anil	NLumia	11000

To get the unique records, we use Union.



```
select * from region1
union
select * from region2
```

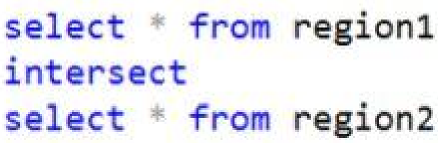


	Cust_id	Cname	Product	Price
1	A101	AJAY	HTCDes	11700
2	B102	RAJESH	MotoG	12499
3	B103	REENA	MCanvas	19490
4	C307	SHEILA	iphone4	26000
5	D205	VIJAY	MotoX	23999
6	D206	GURU	SGalS5	51300
7	E205	ADESH	SGalS3	24499
8	G102	ABHISHEK	HTCOne	50000
9	J103	LALI	SNote3	41000
10	K109	Anil	NLumia	11000
11	K205	SHILPA	SGalS2	22700

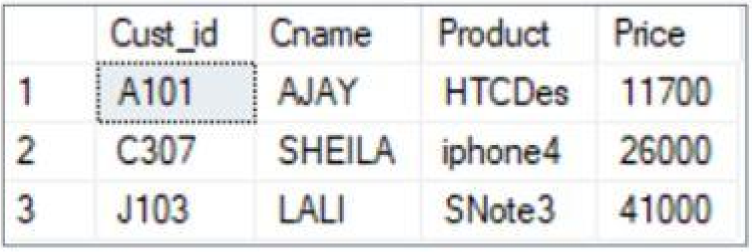
The Intersect operator returns the common records that are the results of 2 or more SELECT statements.

Syntax:

SELECT column_name(s) FROM table1
INTERSECT
SELECT column_name(s) FROM table2;



```
select * from region1
intersect
select * from region2
```

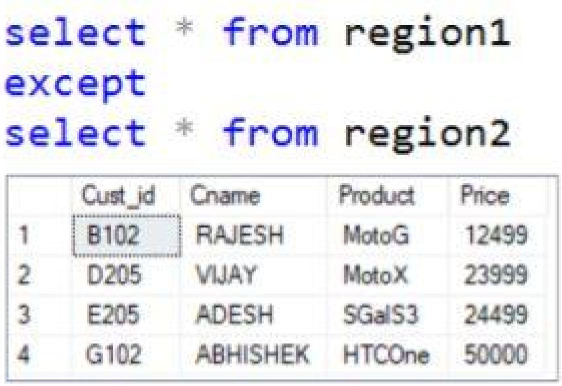


	Cust_id	Cname	Product	Price
1	A101	AJAY	HTCDes	11700
2	C307	SHEILA	iphone4	26000
3	J103	LALI	SNote3	41000

The Except operator returns the uncommon records that are the results of 2 or more SELECT statements.

Syntax:

SELECT column_name(s) FROM table1
EXCEPT
SELECT column_name(s) FROM table2;



```
select * from region1
except
select * from region2
```

	Cust_id	Cname	Product	Price
1	B102	RAJESH	MotoG	12499
2	D205	VIJAY	MotoX	23999
3	E205	ADESH	SGalS3	24499
4	G102	ABHISHEK	HTCOne	50000

Below is the SQL query to return uncommon records from region 1.

38. What is a Subquery in SQL?

A Subquery in SQL is a query within another query. It is also known as a nested query or an inner query. Subqueries are used to enhance the data to be queried by the main query.

It is of two types - Correlated and Non-Correlated Query.

Below is an example of a subquery that returns the name, email id, and phone number of an employee from Texas city.

SELECT name, email, phone

FROM employee

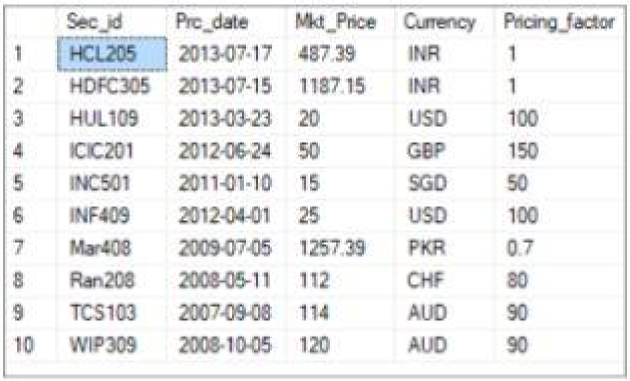
WHERE emp_id IN (

SELECT emp_id

FROM employee

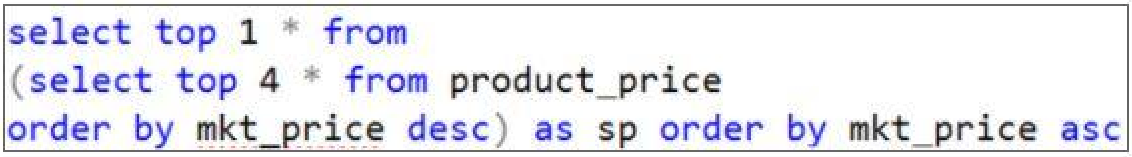
WHERE city = 'Texas');

39. Using the product_price table, write an SQL query to find the record with the fourth-highest market price.



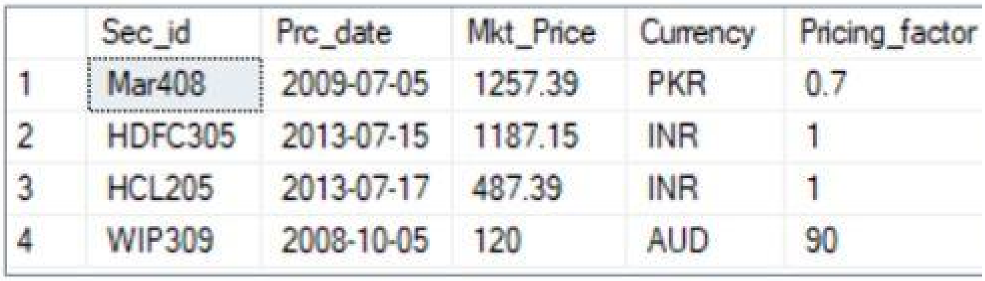
	Sec_id	Prc_date	Mkt_Price	Currency	Pricing_factor
1	HCL205	2013-07-17	487.39	INR	1
2	HDFC305	2013-07-15	1187.15	INR	1
3	HUL109	2013-03-23	20	USD	100
4	ICIC201	2012-06-24	50	GBP	150
5	INC501	2011-01-10	15	SGD	50
6	INF409	2012-04-01	25	USD	100
7	Mar408	2009-07-05	1257.39	PKR	0.7
8	Ran208	2008-05-11	112	CHF	80
9	TCS103	2007-09-08	114	AUD	90
10	WIP309	2008-10-05	120	AUD	90

Fig: Product Price table



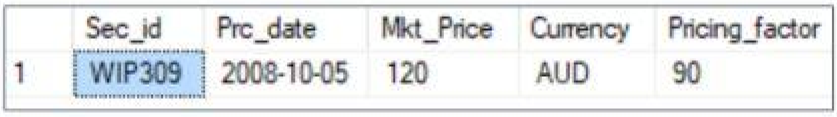
```
select top 1 * from
(select top 4 * from product_price
order by mkt_price desc) as sp order by mkt_price asc
```

select top 4 * from product_price order by mkt_price desc;



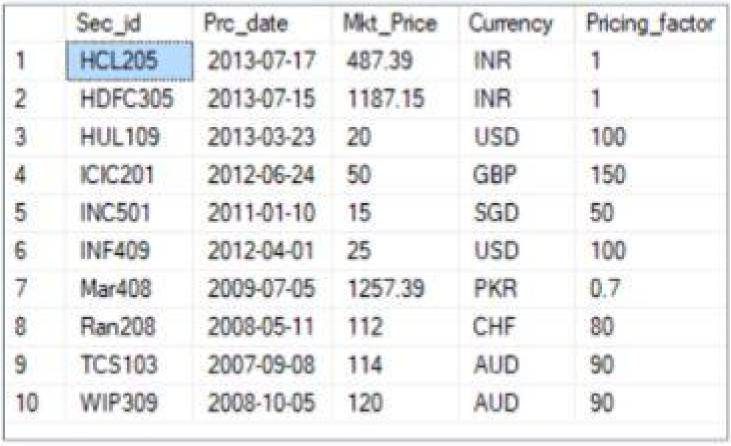
	Sec_id	Prc_date	Mkt_Price	Currency	Pricing_factor
1	Mar408	2009-07-05	1257.39	PKR	0.7
2	HDFC305	2013-07-15	1187.15	INR	1
3	HCL205	2013-07-17	487.39	INR	1
4	WIP309	2008-10-05	120	AUD	90

Now, select the top one from the above result that is in ascending order of mkt_price.



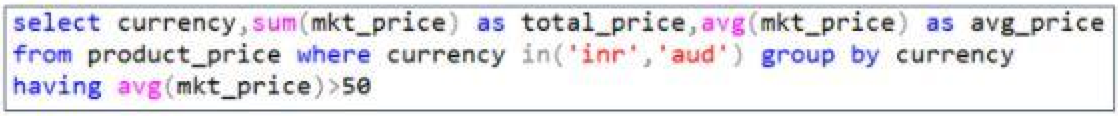
	Sec_id	Prc_date	Mkt_Price	Currency	Pricing_factor
1	WIP309	2008-10-05	120	AUD	90

40. From the product_price table, write an SQL query to find the total and average market price for each currency where the average market price is greater than 100, and the currency is in INR or AUD.



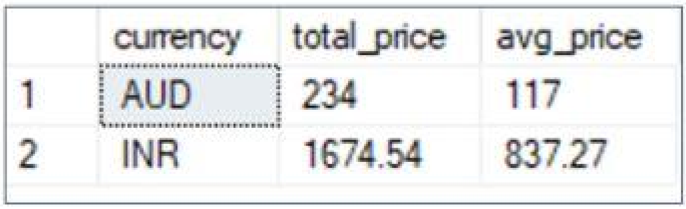
	Sec_id	Prc_date	Mkt_Price	Currency	Pricing_factor
1	HCL205	2013-07-17	487.39	INR	1
2	HDFC305	2013-07-15	1187.15	INR	1
3	HUL109	2013-03-23	20	USD	100
4	ICIC201	2012-06-24	50	GBP	150
5	INC501	2011-01-10	15	SGD	50
6	INF409	2012-04-01	25	USD	100
7	Mar408	2009-07-05	1257.39	PKR	0.7
8	Ran208	2008-05-11	112	CHF	80
9	TCS103	2007-09-08	114	AUD	90
10	WIP309	2008-10-05	120	AUD	90

The SQL query is as follows:



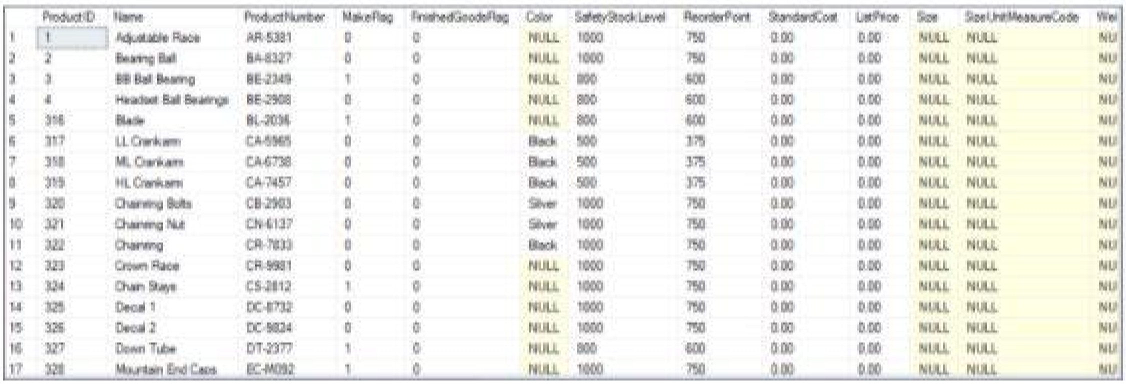
```
select currency,sum(mkt_price) as total_price,avg(mkt_price) as avg_price
from product_price where currency in('inr','aud') group by currency
having avg(mkt_price)>50
```

The output of the query is as follows:



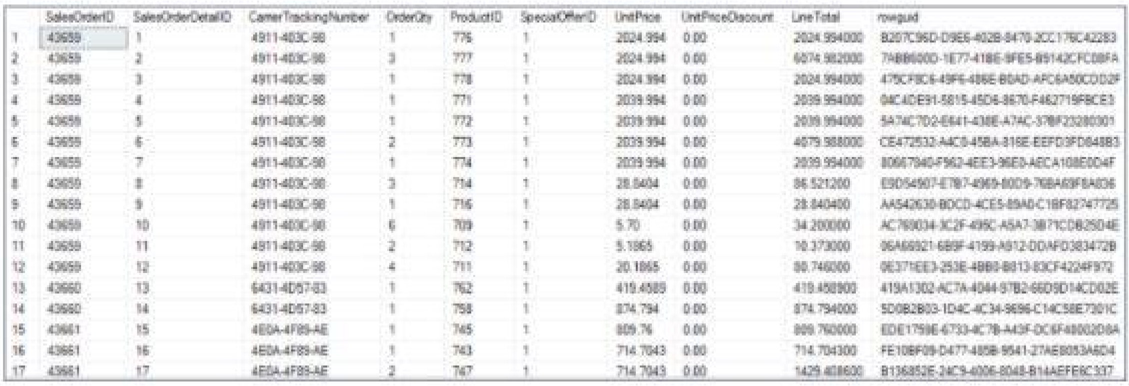
	currency	total_price	avg_price
1	AUD	234	117
2	INR	1674.54	837.27

41. Using the product and sales order detail table, find the products with total units sold greater than 1.5 million.



	ProductID	Name	ProductNumber	MakeFlag	FinishedGoodsFlag	Color	SafetyStockLevel	ReorderPoint	StandardCost	ListPrice	Size	SizeUnitMeasureCode	Wei
1	1	Adjustable Race	AR-5381	0	0	NULL	1000	750	0.00	0.00	NULL	NULL	NU
2	2	Bearing Ball	BA-8327	0	0	NULL	1000	750	0.00	0.00	NULL	NULL	NU
3	3	BB Ball Bearing	BE-2349	1	0	NULL	800	600	0.00	0.00	NULL	NULL	NU
4	4	Headset Ball Bearings	BE-2908	0	0	NULL	800	600	0.00	0.00	NULL	NULL	NU
5	316	Blade	BL-2036	1	0	NULL	800	600	0.00	0.00	NULL	NULL	NU
6	317	LL Crankarm	CA-5965	0	0	Black	500	375	0.00	0.00	NULL	NULL	NU
7	318	ML Crankarm	CA-6738	0	0	Black	500	375	0.00	0.00	NULL	NULL	NU
8	319	HL Crankarm	CA-7457	0	0	Black	500	375	0.00	0.00	NULL	NULL	NU
9	320	Chainring Bolts	CB-2903	0	0	Silver	1000	750	0.00	0.00	NULL	NULL	NU
10	321	Chainring Nut	CN-6137	0	0	Silver	1000	750	0.00	0.00	NULL	NULL	NU
11	322	Chainring	CR-7833	0	0	Black	1000	750	0.00	0.00	NULL	NULL	NU
12	323	Crown Race	CR-9981	0	0	NULL	1000	750	0.00	0.00	NULL	NULL	NU
13	324	Chain Stays	CS-2812	1	0	NULL	1000	750	0.00	0.00	NULL	NULL	NU
14	325	Decal 1	DC-8732	0	0	NULL	1000	750	0.00	0.00	NULL	NULL	NU
15	326	Decal 2	DC-9824	0	0	NULL	1000	750	0.00	0.00	NULL	NULL	NU
16	327	Down Tube	DT-2377	1	0	NULL	800	600	0.00	0.00	NULL	NULL	NU
17	328	Mountain End Caps	EC-M092	1	0	NULL	1000	750	0.00	0.00	NULL	NULL	NU

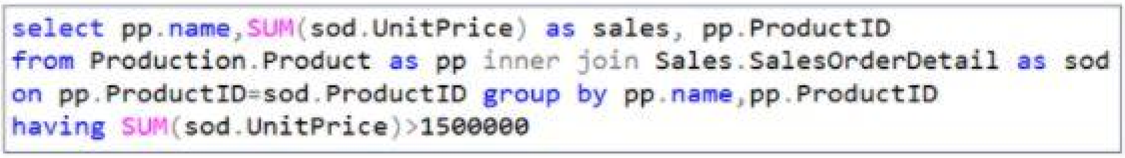
Fig: Products table



	SalesOrderID	SalesOrderDetailID	CarrierTrackingNumber	OrderQty	ProductID	SpecialOfferID	UnitPrice	UnitPriceDiscount	LineTotal	rowguid
1	43659	1	4911-403C-98	1	776	1	2024.994	0.00	2024.994000	B207C96D-D9E6-402B-8470-2CC176C42283
2	43659	2	4911-403C-98	3	777	1	2024.994	0.00	6074.982000	7ABB600D-1E77-41BE-9FE5-B9142CFC08FA
3	43659	3	4911-403C-98	1	778	1	2024.994	0.00	2024.994000	475CF8C6-49F6-486E-B0AD-AFC6A50CDD2F
4	43659	4	4911-403C-98	1	771	1	2039.994	0.00	2039.994000	04C4DE91-5815-45D6-8670-F462719FBCE3
5	43659	5	4911-403C-98	1	772	1	2039.994	0.00	2039.994000	5A74C7D2-E641-438E-A74C-37BF23280301
6	43659	6	4911-403C-98	2	773	1	2039.994	0.00	4079.988000	CE472532-A4C0-45BA-816E-EEFD3FD848B3
7	43659	7	4911-403C-98	1	774	1	2039.994	0.00	2039.994000	80667840-F962-4EE3-96E0-AECA108E0D4F
8	43659	8	4911-403C-98	3	714	1	28.8404	0.00	86.521200	E9D54907-E7B7-4969-80D9-76BA69F8A836
9	43659	9	4911-403C-98	1	716	1	28.8404	0.00	28.840400	AA542630-BDCD-4CE5-89A0-C1BF82747725
10	43659	10	4911-403C-98	6	709	1	5.70	0.00	34.200000	AC769034-3C2F-495C-A5A7-3B71CDB25D4E
11	43659	11	4911-403C-98	2	712	1	5.1865	0.00	10.373000	06A66921-6B9F-4199-A912-DDAFD383472B
12	43659	12	4911-403C-98	4	711	1	20.1865	0.00	80.746000	0E371EE3-253E-4BB0-8813-83CF4224F972
13	43660	13	6431-4D57-83	1	762	1	419.4589	0.00	419.458900	419A1302-AC7A-4044-97B2-66D9D14CD02E
14	43660	14	6431-4D57-83	1	758	1	874.794	0.00	874.794000	5D0B2B03-1D4C-4C34-9696-C14C58E7301C
15	43661	15	4E0A-4F89-AE	1	745	1	809.76	0.00	809.760000	EDE1759E-6733-4C78-A43F-DC6F48002D8A
16	43661	16	4E0A-4F89-AE	1	743	1	714.7043	0.00	714.704300	FE10BF09-D477-485B-9541-27AEB053A6D4
17	43661	17	4E0A-4F89-AE	2	747	1	714.7043	0.00	1429.408600	B136852E-24C9-4006-8048-B14AEFE8C337

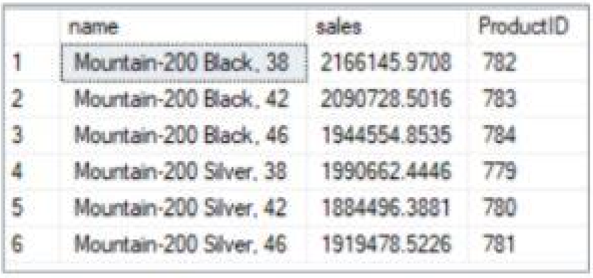
Fig: Sales order detail table

We can use an inner join to get records from both the tables. We'll join the tables based on a common key column, i.e., ProductID.



```sql
select pp.name,SUM(sod.UnitPrice) as sales, pp.ProductID
from Production.Product as pp inner join Sales.SalesOrderDetail as sod
on pp.ProductID=sod.ProductID group by pp.name,pp.ProductID
having SUM(sod.UnitPrice)>1500000
```

The result of the SQL query is shown below.



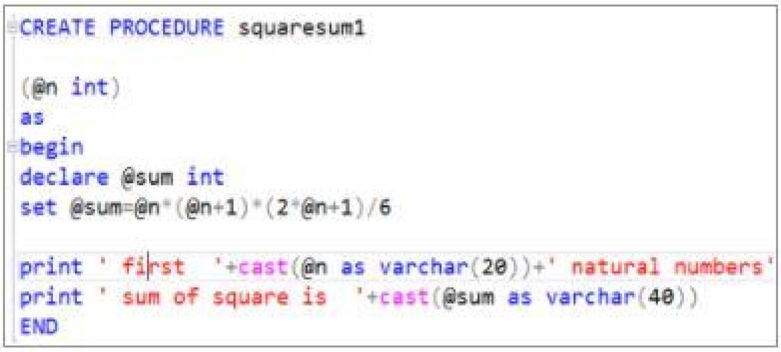
	name	sales	ProductID
1	Mountain-200 Black, 38	2166145.9708	782
2	Mountain-200 Black, 42	2090728.5016	783
3	Mountain-200 Black, 46	1944554.8535	784
4	Mountain-200 Silver, 38	1990662.4446	779
5	Mountain-200 Silver, 42	1884496.3881	780
6	Mountain-200 Silver, 46	1919478.5226	781

42. How do you write a stored procedure in SQL?

You must be prepared for this question thoroughly before your next data analyst interview. The stored procedure is an SQL script that is used to run a task several times.

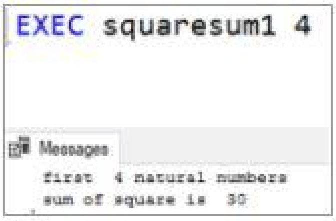
Let's look at an example to create a stored procedure to find the sum of the first N natural numbers' squares.

- Create a procedure by giving a name, here it's squaresum1
- Declare the variables
- Write the formula using the set statement
- Print the values of the computed variable
- To run the stored procedure, use the EXEC command



```
CREATE PROCEDURE squaresum1

(@n int)
as
begin
declare @sum int
set @sum=@n*(@n+1)*(2*@n+1)/6

print ' first  '+cast(@n as varchar(20))+' natural numbers'
print ' sum of square is  '+cast(@sum as varchar(40))
END
```

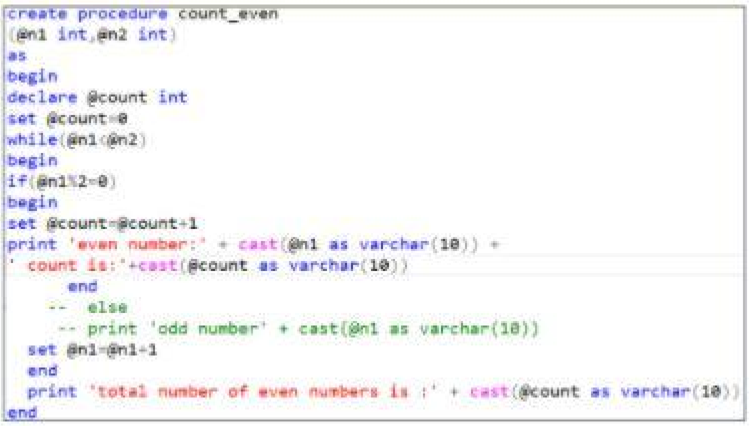
Output: Display the sum of the square for the first four natural numbers



```
EXEC squaresum1 4
```

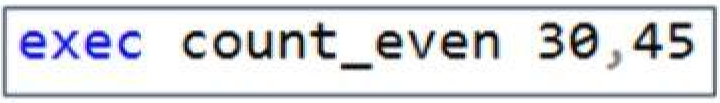
```
first  4 natural numbers
sum of square is  30
```

43. Write an SQL stored procedure to find the total even number between two users given numbers.

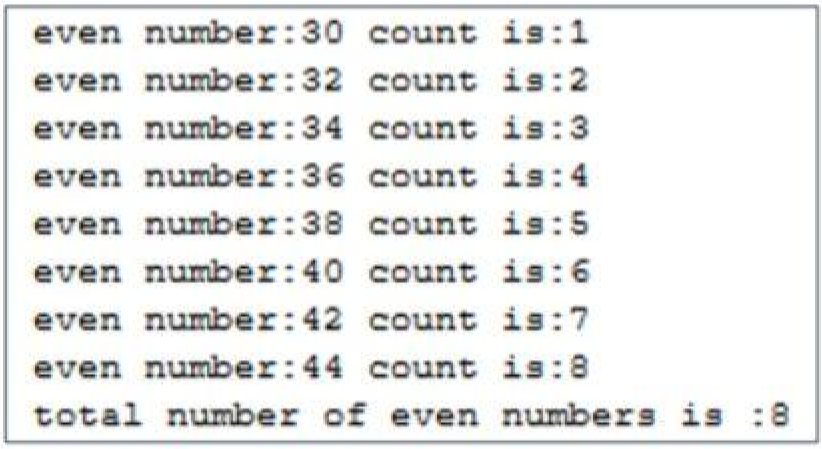


```
create procedure count_even
(@n1 int,@n2 int)
as
begin
declare @count int
set @count=0
while(@n1<@n2)
begin
if(@n1%2=0)
begin
set @count=@count+1
print 'even number:' + cast(@n1 as varchar(10)) +
' count is:'+cast(@count as varchar(10))
      end
    --  else
     -- print 'odd number' + cast(@n1 as varchar(10))
  set @n1=@n1+1
  end
  print 'total number of even numbers is :' + cast(@count as varchar(10))
end
```

Here is the output to print all even numbers between 30 and 45.



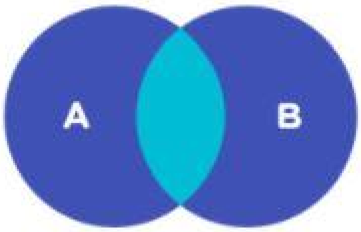
```
exec count_even 30,45
```



```
even number:30 count is:1
even number:32 count is:2
even number:34 count is:3
even number:36 count is:4
even number:38 count is:5
even number:40 count is:6
even number:42 count is:7
even number:44 count is:8
total number of even numbers is :8
```

Tableau Data Analyst Interview Questions

44. How is joining different from blending in Tableau?



Data Joining	Data Blending

Data joining can only be carried out when the data comes from the same source.	Data blending is used when the data is from two or more different sources.
E.g: Combining two or more worksheets from the same Excel file or two tables from the same databases. All the combined sheets or tables contain a common set of dimensions and measures.	E.g: Combining the Oracle table with SQL Server, or combining Excel sheet and Oracle table or two sheets from Excel. Meanwhile, in data blending, each data source contains its own set of dimensions and measures.

45. What do you understand by LOD in Tableau?

LOD in Tableau stands for Level of Detail. It is an expression that is used to execute complex queries involving many dimensions at the data sourcing level. Using LOD expression, you can find duplicate values, synchronize chart axes and create bins on aggregated data.

46. Can you discuss the process of feature selection and its importance in data analysis?

Feature selection is the process of selecting a subset of relevant features from a larger set of variables or predictors in a dataset. It aims to improve model performance, reduce overfitting, enhance interpretability, and optimize computational efficiency. Here's an overview of the process and its importance:

Importance of Feature Selection:

- Improved Model Performance: By selecting the most relevant features, the model can focus on the most informative variables, leading to better predictive accuracy and generalization.
- Overfitting Prevention: Including irrelevant or redundant features can lead to overfitting, where the model learns noise or specific patterns in the training data that do not generalize well to new data. Feature selection mitigates this risk.
- Interpretability and Insights: A smaller set of selected features makes it easier to interpret and understand the model's results, facilitating insights and actionable conclusions.
- Computational Efficiency: Working with a reduced set of features can significantly improve computational efficiency, especially when dealing with large datasets.

47. What are the different connection types in Tableau Software?

There are mainly 2 types of connections available in Tableau.

Extract: Extract is an image of the data that will be extracted from the data source and placed into the Tableau repository. This image(snapshot) can be refreshed periodically, fully, or incrementally.

Live: The live connection makes a direct connection to the data source. The data will be fetched straight from tables. So, data is always up to date and consistent.

48. What are the different joins that Tableau provides?

Joins in Tableau work similarly to the SQL join statement. Below are the types of joins that Tableau supports:

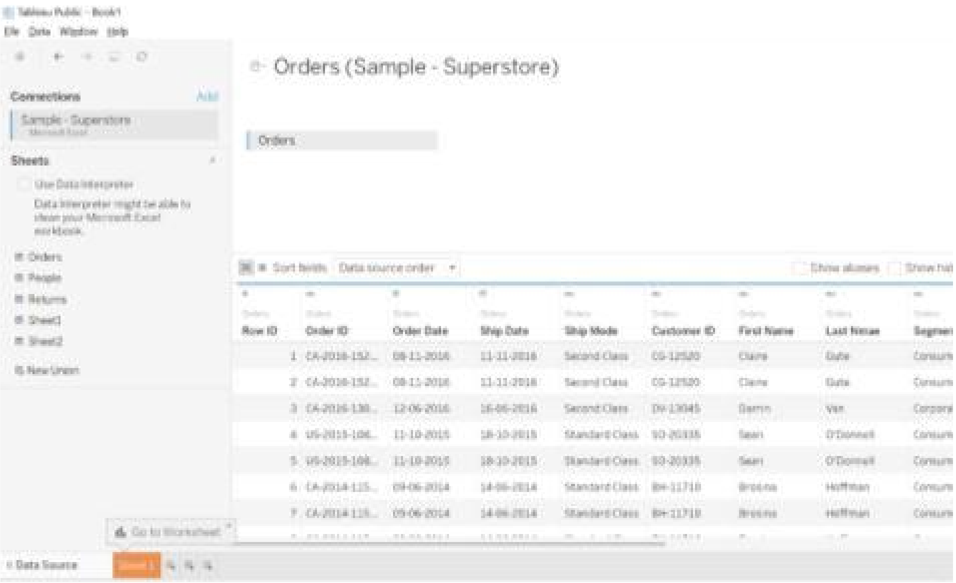
- Left Outer Join
- Right Outer Join
- Full Outer Join
- Inner Join

49. What is a Gantt Chart in Tableau?

A Gantt chart in Tableau depicts the progress of value over the period, i.e., it shows the duration of events. It consists of bars along with the time axis. The Gantt chart is mostly used as a project management tool where each bar is a measure of a task in the project.

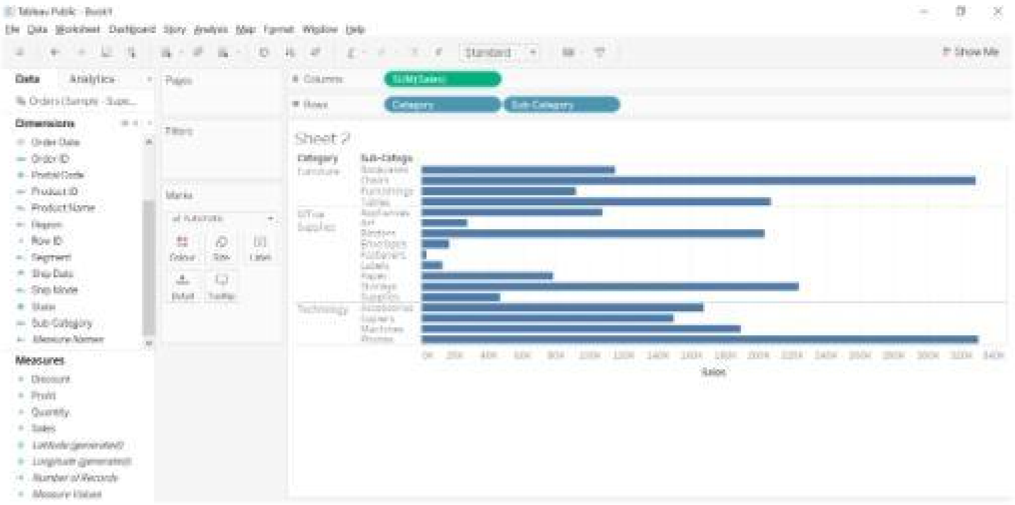
50. Using the Sample Superstore dataset, create a view in Tableau to analyze the sales, profit, and quantity sold across different subcategories of items present under each category.

- Load the Sample - Superstore dataset



Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	First Name	Last Name	Segmen
1	CA-2016-152...	08-11-2016	11-11-2016	Second Class	CG-12520	Claire	Gute	Consum
2	CA-2016-152...	08-11-2016	11-11-2016	Second Class	CG-12520	Claire	Gute	Consum
3	CA-2016-138...	12-06-2016	16-06-2016	Second Class	DV-13045	Darrin	Van	Corpora
4	US-2015-108...	11-10-2015	18-10-2015	Standard Class	SO-20335	Sean	O'Donnell	Consum
5	US-2015-108...	11-10-2015	18-10-2015	Standard Class	SO-20335	Sean	O'Donnell	Consum
6	CA-2014-115...	09-06-2014	14-06-2014	Standard Class	BH-11710	Brosina	Hoffman	Consum
7	CA-2014-115...	09-06-2014	14-06-2014	Standard Class	BH-11710	Brosina	Hoffman	Consum

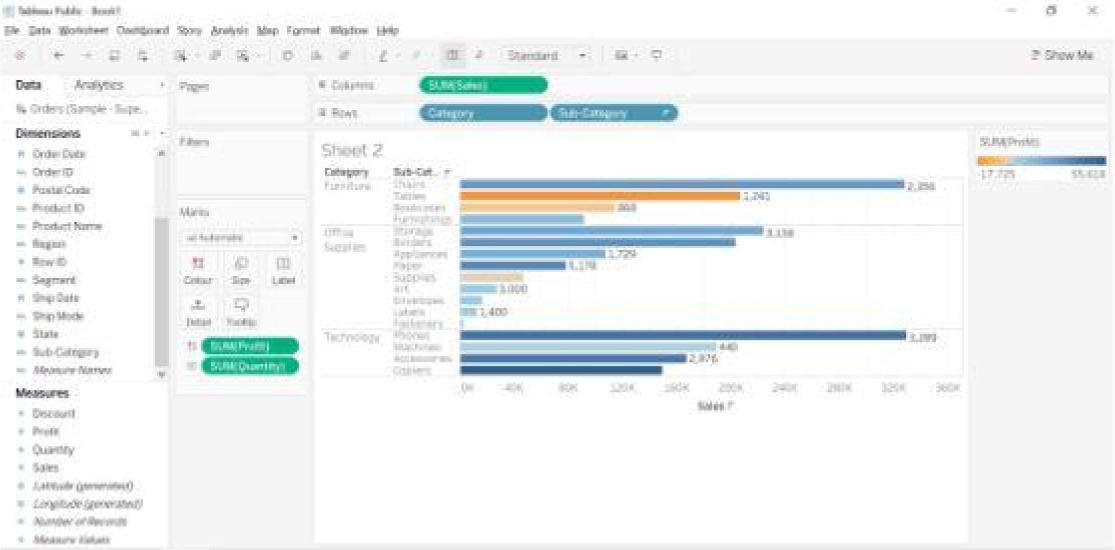
- Drag Category and Subcategory columns into Rows, and Sales on to Columns. It will result in a horizontal bar chart.



Columns: SUM(Sales)

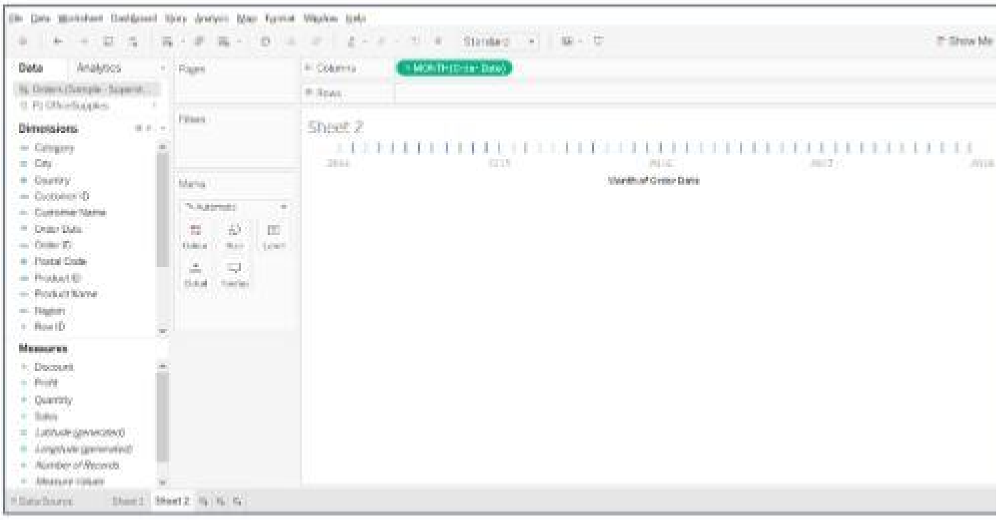
Rows: Category, Sub-Category

- Drag Profit on to Colour, and Quantity on to Label. Sort the Sales axis in descending order of the sum of sales within each sub-category.

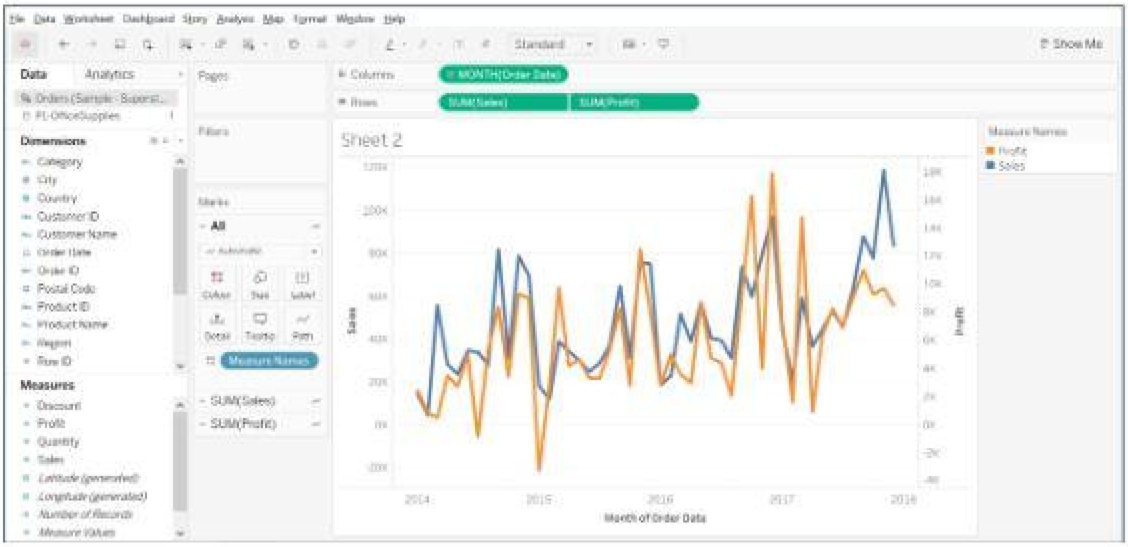


51. Create a dual-axis chart in Tableau to present Sales and Profit across different years using the Sample Superstore dataset.

- Drag the Order Date field from Dimensions on to Columns, and convert it into continuous Month.



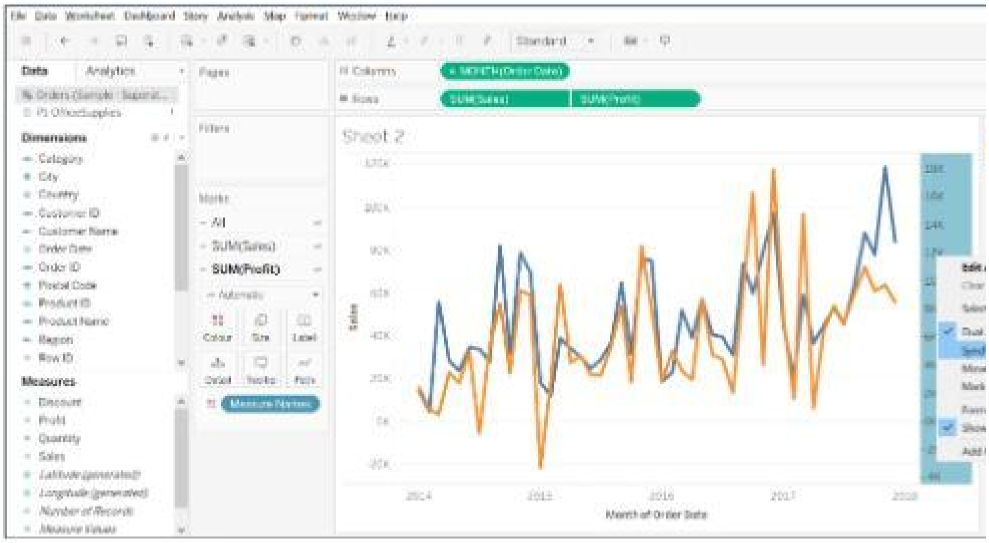
- Drag Sales on to Rows, and Profits to the right corner of the view until you see a light green rectangle.



Columns: MONTH(Order Date)

Rows: SUM(Sales) SUM(Profit)

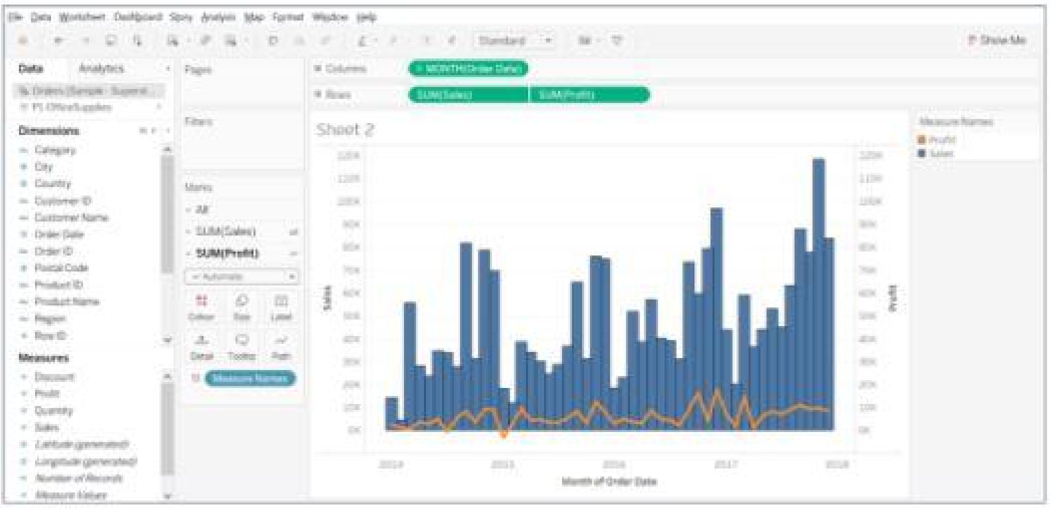
- Synchronize the right axis by right-clicking on the profit axis.



Columns: MONTH(Order Date)

Rows: SUM(Sales) SUM(Profit)

- Under the Marks card, change SUM(Sales) to Bar and SUM(Profit) to Line and adjust the size.

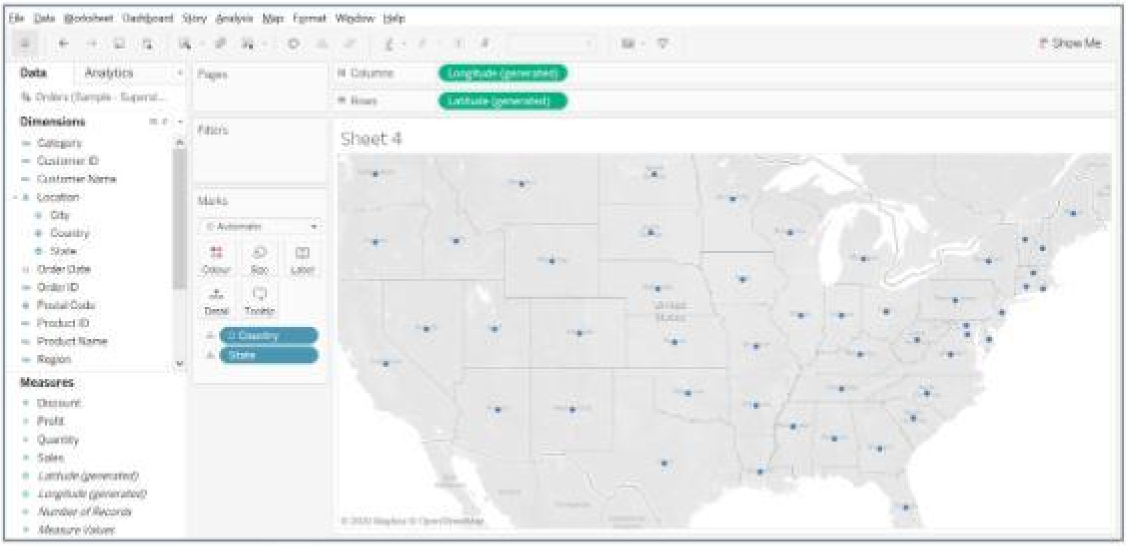


Columns: MONTH(Order Date)

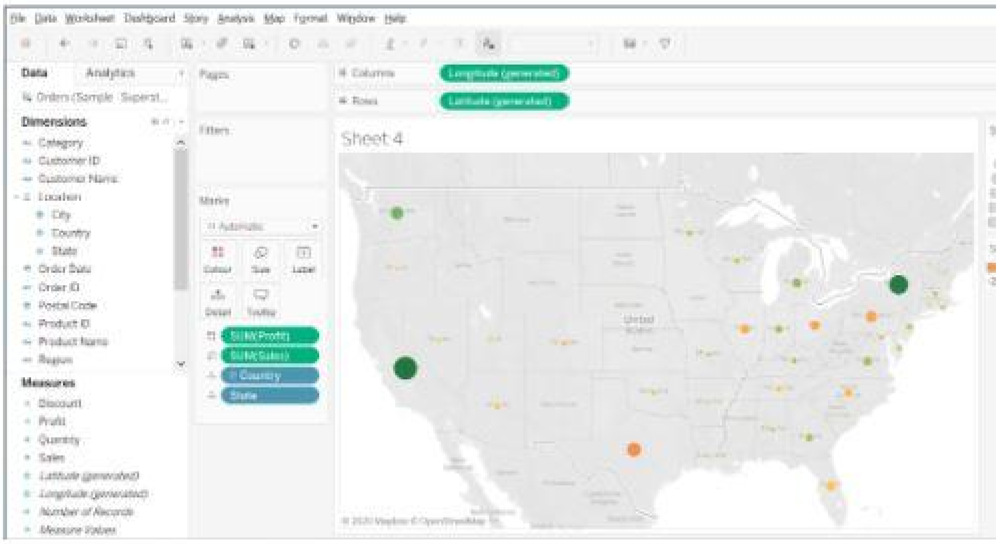
Rows: SUM(Sales) SUM(Profit)

52. Design a view in Tableau to show State-wise Sales and Profit using the Sample Superstore dataset.

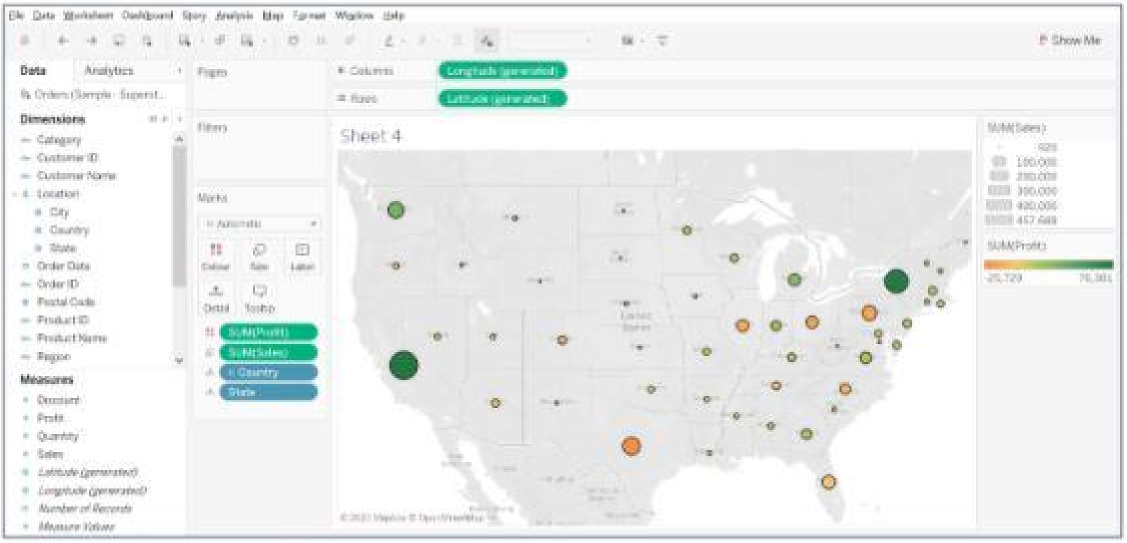
- Drag the Country field on to the view section and expand it to see the States.



- Drag the Sales field on to Size, and Profit on to Colour.

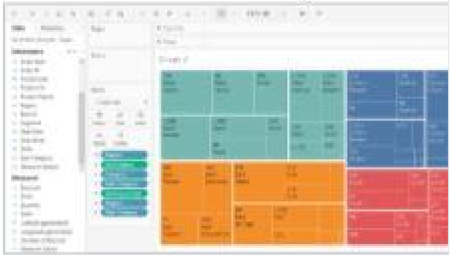
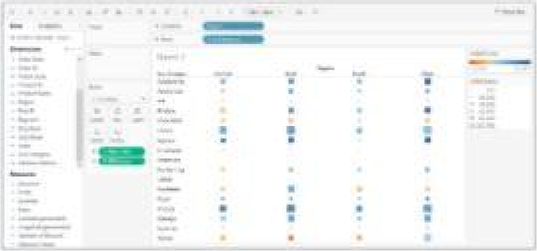


- Increase the size of the bubbles, add a border, and halo color.



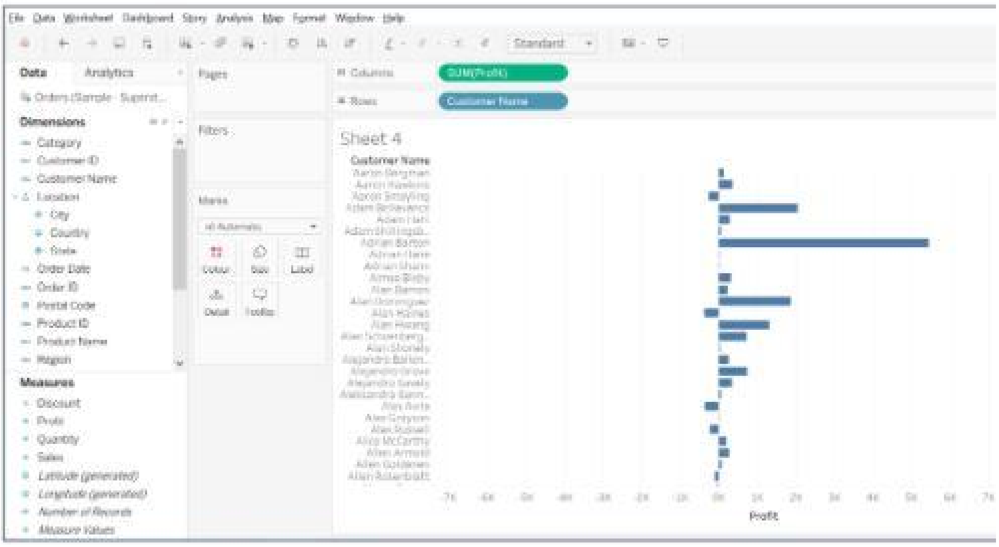
From the above map, it is clear that states like Washington, California, and New York have the highest sales and profits. While Texas, Pennsylvania, and Ohio have good amounts of sales but the least profits.

53. What is the difference between Treemaps and Heatmaps in Tableau?

Treemaps	Heatmaps
	
Treemaps are used to display data in nested rectangles.	Heat maps can visualize measures against dimensions with the help of colors and size to differentiate one or more dimensions and up to two measures.

You use dimensions to define the structure of the treemap, and measures to define the size or color of the individual rectangles.	The layout is like a text table with variations in values encoded as colors.
Treemaps are a relatively simple data visualization that can provide insight in a visually attractive format.	In the heatmap, you can quickly see a wide array of information.

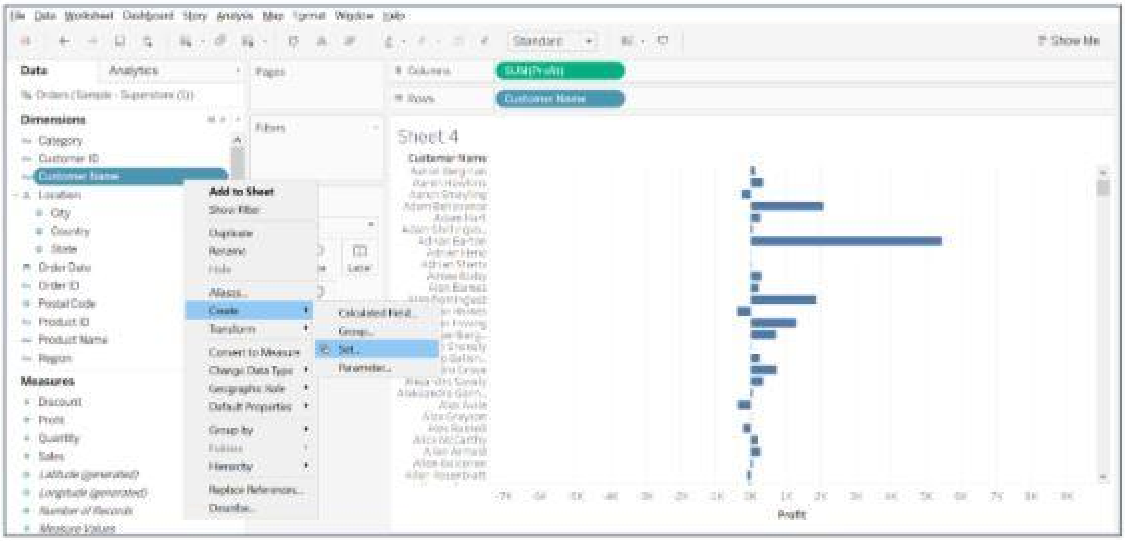
54. Using the Sample Superstore dataset, display the top 5 and bottom 5 customers based on their profit.



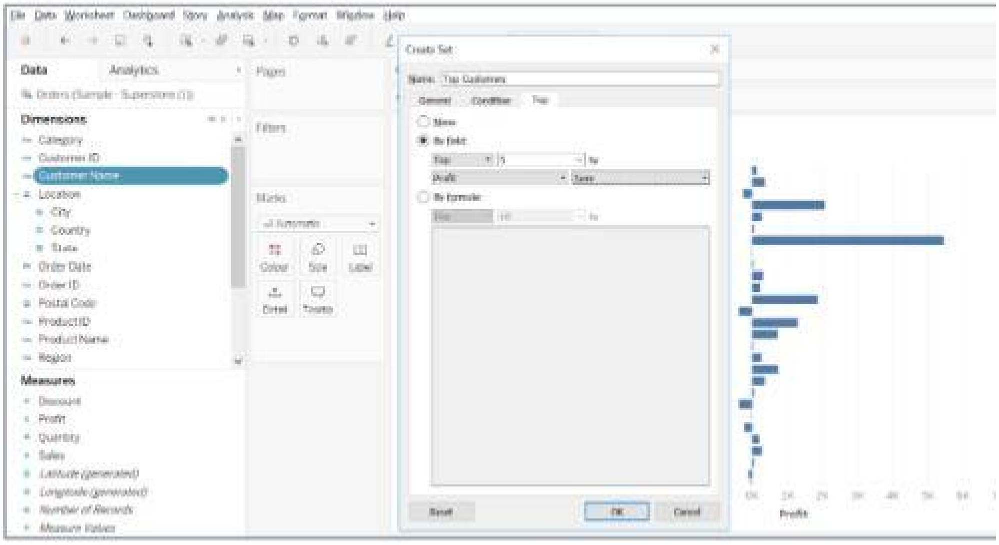
Columns: SUM(Profit)

Rows: Customer Name

- Drag Customer Name field on to Rows, and Profit on to Columns.



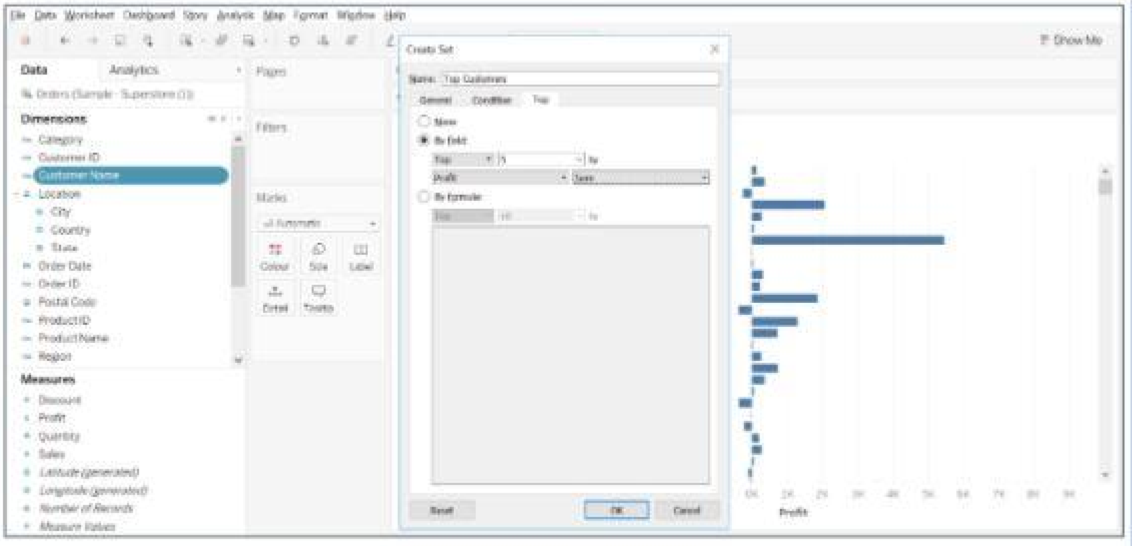
- Right-click on the Customer Name column to create a set



Name: Top Customers

○ None
● By field
Top 5 by
Profit Sum
○ By formula

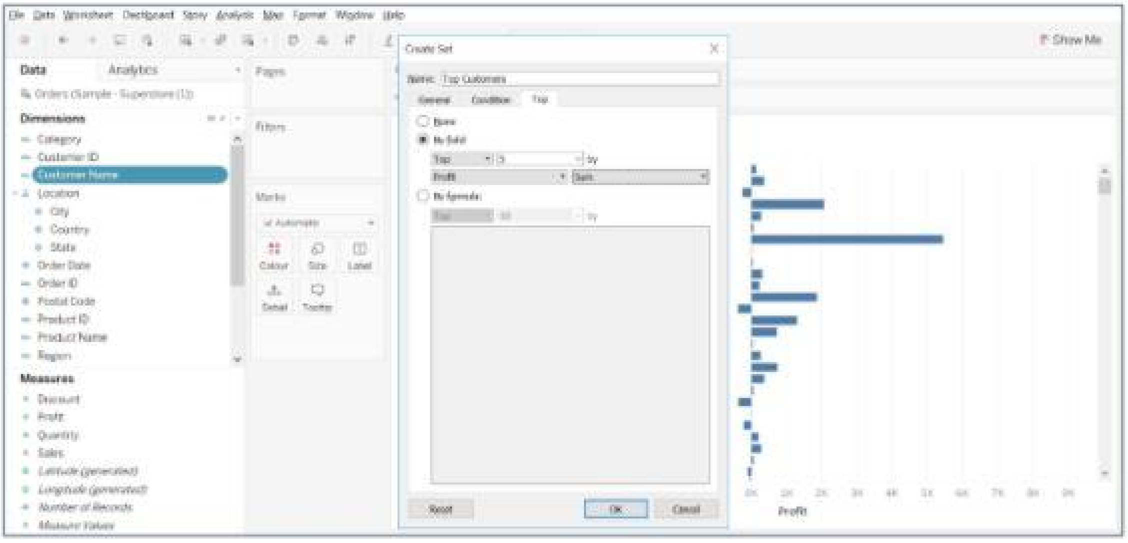
- Give a name to the set and select the top tab to choose the top 5 customers by sum(profit)



Name: Top Customers

○ None
● By field
Top 5 by
Profit Sum
○ By formula

- Similarly, create a set for the bottom five customers by sum(profit)



Create Set

Name: Top Customers

General | Condition | Top

○ None

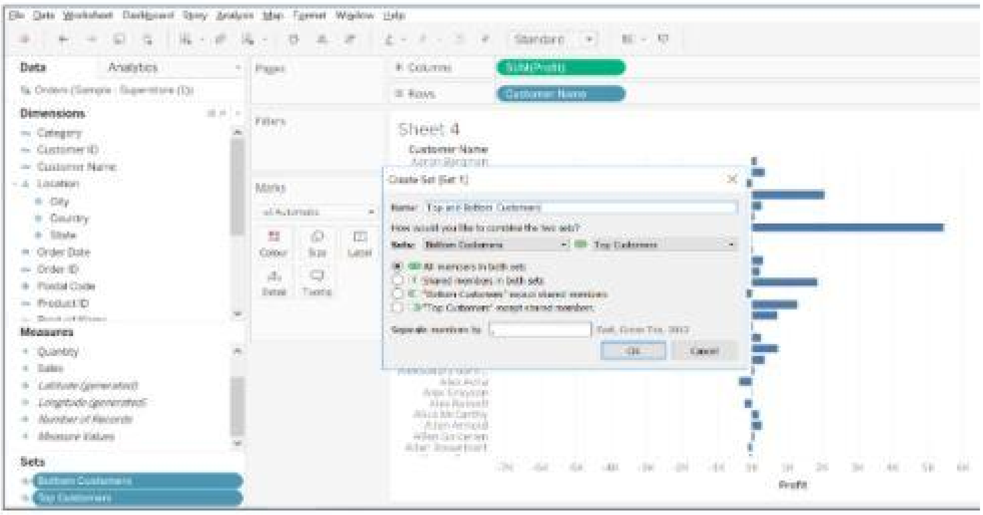
● By field:

Top 5 by Profit Sum

○ By formula:

Reset | OK | Cancel

- Select both the sets, right-click to create a combined set. Give a name to the set and choose All members in both sets.



Create Set (Set 1)

Name: Top and Bottom Customers

How would you like to combine the two sets?

Sets: Bottom Customers | Top Customers

● All members in both sets

○ Shared members in both sets

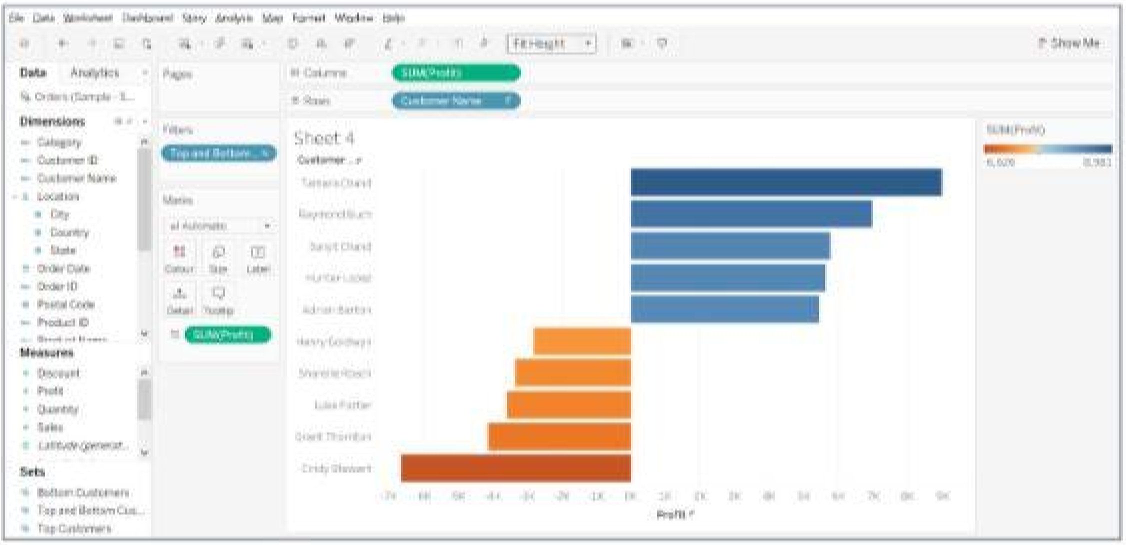
○ "Bottom Customers" except shared members

○ "Top Customers" except shared members

Separate members by:

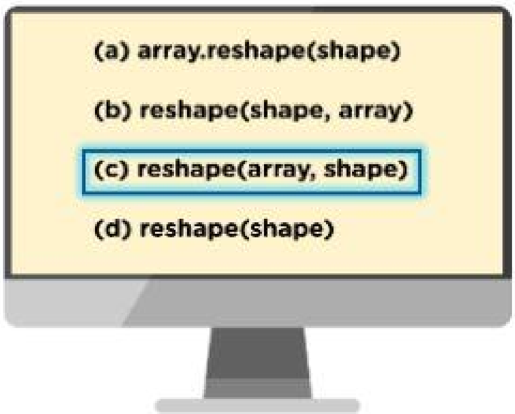
OK | Cancel

- Drag top and bottom customers set on to Filters, and Profit field on to Colour to get the desired result.



Data Analyst Interview Questions On Python

55. What is the correct syntax for reshape() function in NumPy?

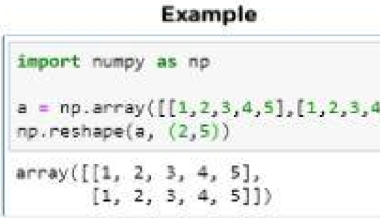


(a) array.reshape(shape)

(b) reshape(shape, array)

(c) reshape(array, shape)

(d) reshape(shape)



Example

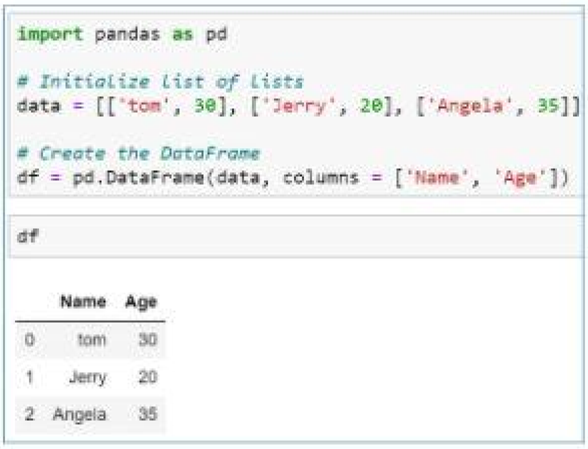
```
import numpy as np

a = np.array([[1,2,3,4,5],[1,2,3,4
np.reshape(a, (2,5))
```

```
array([[1, 2, 3, 4, 5],
       [1, 2, 3, 4, 5]])
```

56. What are the different ways to create a data frame in Pandas?

There are two ways to create a Pandas data frame.

- By initializing a list

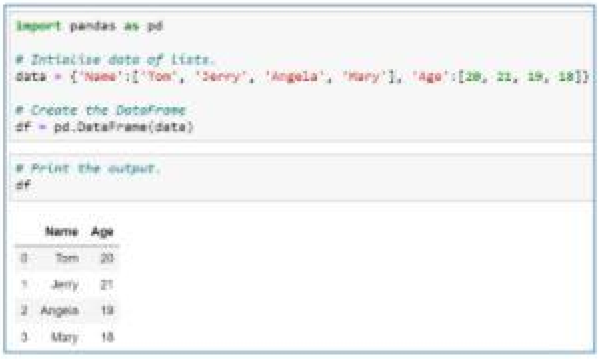


```
import pandas as pd

# Initialize list of lists
data = [['tom', 30], ['Jerry', 20], ['Angela', 35]]

# Create the DataFrame
df = pd.DataFrame(data, columns = ['Name', 'Age'])
```

```
df
```

```
     Name  Age
0     tom   30
1   Jerry   20
2  Angela   35
```

- By initializing a dictionary



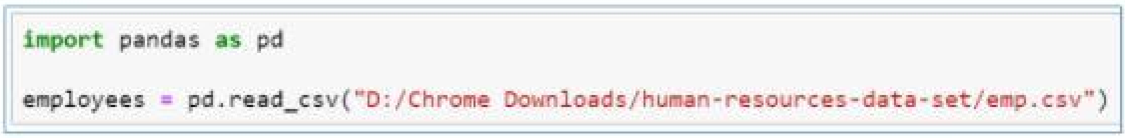
```
import pandas as pd

# Intialize data of lists.
data = {'Name':['Tom', 'Jerry', 'Angela', 'Mary'], 'Age':[20, 21, 19, 18]}

# Create the DataFrame
df = pd.DataFrame(data)
```

```
# Print the output.
df
```

```
     Name  Age
0     Tom   20
1   Jerry   21
2  Angela   19
3    Mary   18
```

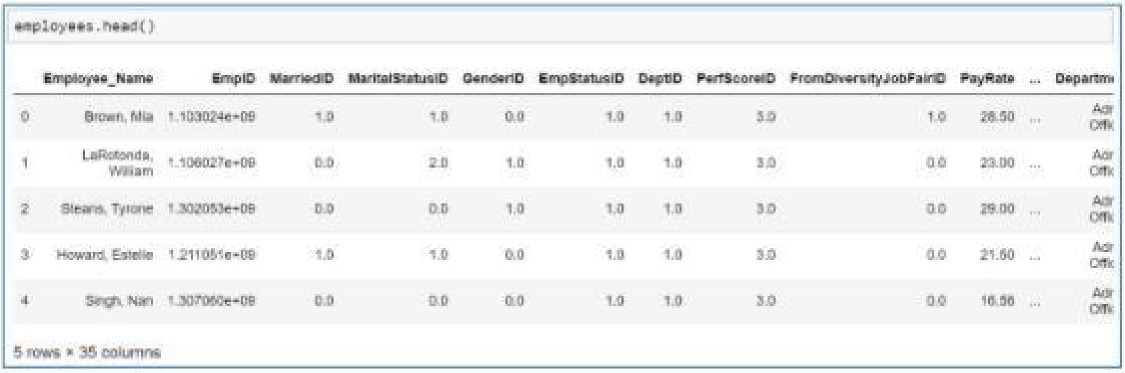
57. Write the Python code to create an employee's data frame from the "emp.csv" file and display the head and summary.

To create a DataFrame in Python, you need to import the Pandas library and use the read_csv function to load the .csv file. Give the right location where the file name and its extension follow the dataset.



```
import pandas as pd

employees = pd.read_csv("D:/Chrome Downloads/human-resources-data-set/emp.csv")
```

To display the head of the dataset, use the head() function.

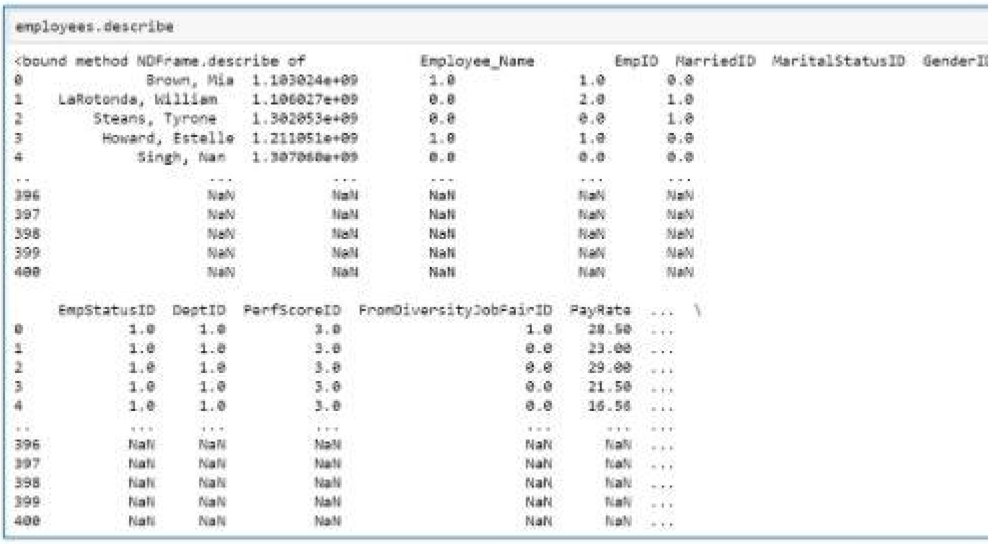


```
employees.head()
```

	Employee_Name	EmpID	MarriedID	MaritalStatusID	GenderID	EmpStatusID	DeptID	PerfScoreID	FromDiversityJobFairID	PayRate	...	Departm
0	Brown, Mia	1.103024e+09	1.0	1.0	0.0	1.0	1.0	3.0	1.0	28.50	...	Adr Offic
1	LaRotonda, William	1.106027e+09	0.0	2.0	1.0	1.0	1.0	3.0	0.0	23.00	...	Adr Offic
2	Steans, Tyrone	1.302053e+09	0.0	0.0	1.0	1.0	1.0	3.0	0.0	29.00	...	Adr Offic
3	Howard, Estelle	1.211051e+09	1.0	1.0	0.0	1.0	1.0	3.0	0.0	21.50	...	Adr Offic
4	Singh, Nan	1.307060e+09	0.0	0.0	0.0	1.0	1.0	3.0	0.0	16.56	...	Adr Offic

5 rows × 35 columns

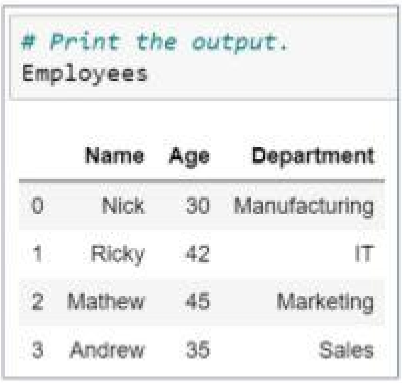
The ‘describe’ method is used to return the summary statistics in Python.



```
employees.describe
```

```
<bound method NDFrame.describe of           Employee_Name         EmpID  MarriedID  MaritalStatusID  GenderII
0             Brown, Mia  1.103024e+09       1.0              1.0      0.0
1    LaRotonda, William   1.106027e+09       0.0              2.0      1.0
2        Steans, Tyrone   1.302053e+09       0.0              0.0      1.0
3       Howard, Estelle  1.211051e+09       1.0              1.0      0.0
4            Singh, Nan   1.307060e+09       0.0              0.0      0.0
..                  ...           ...       ...              ...      ...
396                 NaN           NaN       NaN              NaN      NaN
397                 NaN           NaN       NaN              NaN      NaN
398                 NaN           NaN       NaN              NaN      NaN
399                 NaN           NaN       NaN              NaN      NaN
400                 NaN           NaN       NaN              NaN      NaN

     EmpStatusID  DeptID  PerfScoreID  FromDiversityJobFairID  PayRate  ...  \
0            1.0     1.0          3.0                     1.0    28.50  ...
1            1.0     1.0          3.0                     0.0    23.00  ...
2            1.0     1.0          3.0                     0.0    29.00  ...
3            1.0     1.0          3.0                     0.0    21.50  ...
4            1.0     1.0          3.0                     0.0    16.56  ...
..           ...     ...          ...                     ...      ...  ...
396          NaN     NaN          NaN                     NaN      NaN  ...
397          NaN     NaN          NaN                     NaN      NaN  ...
398          NaN     NaN          NaN                     NaN      NaN  ...
399          NaN     NaN          NaN                     NaN      NaN  ...
400          NaN     NaN          NaN                     NaN      NaN  ...
```

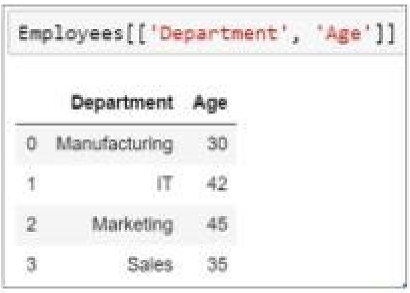
58. How will you select the Department and Age columns from an Employee data frame?



```
# Print the output.
Employees
```

	Name	Age	Department
0	Nick	30	Manufacturing
1	Ricky	42	IT
2	Mathew	45	Marketing
3	Andrew	35	Sales

You can use the column names to extract the desired columns.

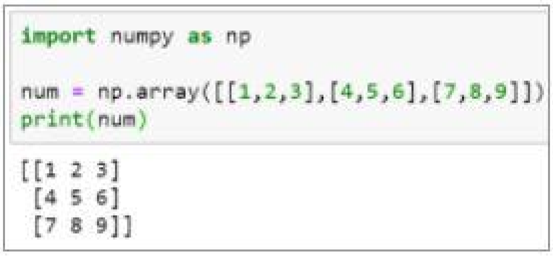


```
Employees[['Department', 'Age']]
```

	Department	Age
0	Manufacturing	30
1	IT	42
2	Marketing	45
3	Sales	35

59. Suppose there is an array, what would you do?

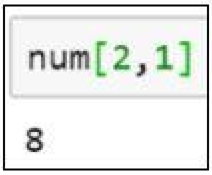
num = np.array([[1,2,3],[4,5,6],[7,8,9]]). Extract the value 8 using 2D indexing.



```
import numpy as np

num = np.array([[1,2,3],[4,5,6],[7,8,9]])
print(num)
```

```
[[1 2 3]
 [4 5 6]
 [7 8 9]]
```

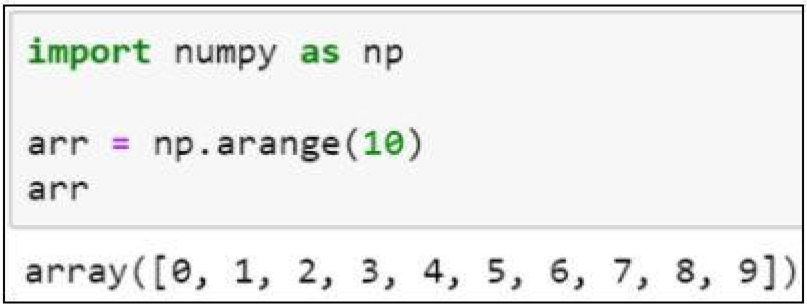
Since the value eight is present in the 2nd row of the 1st column, we use the same index positions and pass it to the array.



```
num[2,1]
```

```
8
```

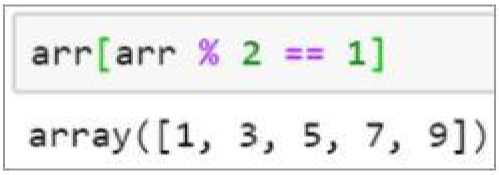
60. Suppose there is an array that has values [0,1,2,3,4,5,6,7,8,9]. How will you display the following values from the array - [1,3,5,7,9]?



```
import numpy as np

arr = np.arange(10)
arr
```

```
array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
```

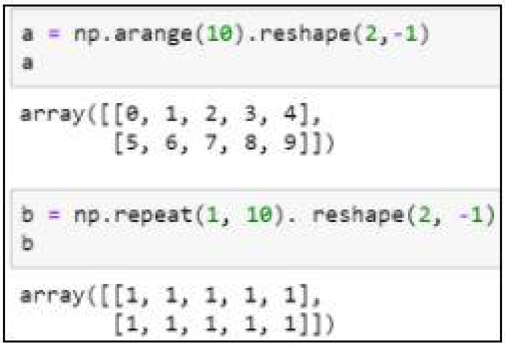
Since we only want the odd number from 0 to 9, you can perform the modulus operation and check if the remainder is equal to 1.



```
arr[arr % 2 == 1]
```

```
array([1, 3, 5, 7, 9])
```

61. There are two arrays, ‘a’ and ‘b’. Stack the arrays a and b horizontally using the NumPy library in Python.



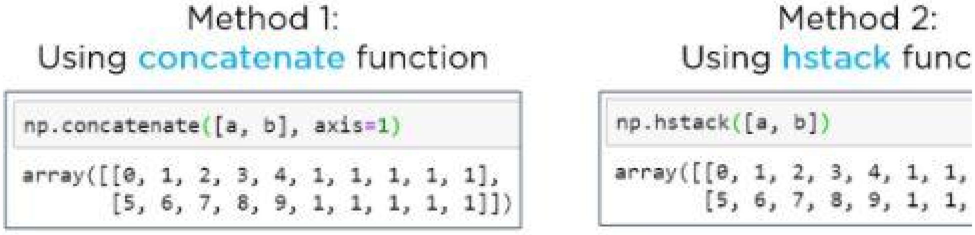
```
a = np.arange(10).reshape(2,-1)
a
```

```
array([[0, 1, 2, 3, 4],
       [5, 6, 7, 8, 9]])
```

```
b = np.repeat(1, 10). reshape(2, -1)
b
```

```
array([[1, 1, 1, 1, 1],
       [1, 1, 1, 1, 1]])
```

You can either use the concatenate() or the hstack() function to stack the arrays.



Method 1:
Using concatenate function

```
np.concatenate([a, b], axis=1)
```

```
array([[0, 1, 2, 3, 4, 1, 1, 1, 1, 1],
       [5, 6, 7, 8, 9, 1, 1, 1, 1, 1]])
```

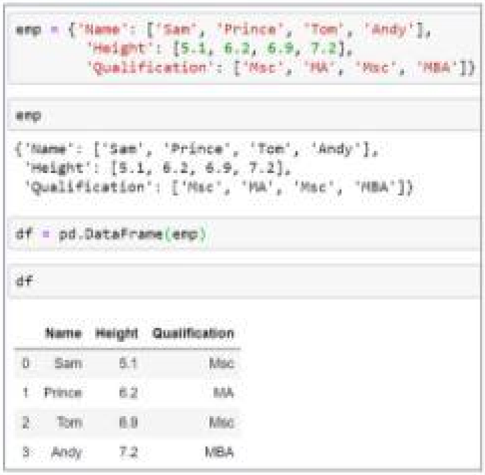
Method 2:
Using hstack func

```
np.hstack([a, b])
```

```
array([[0, 1, 2, 3, 4, 1, 1,
       [5, 6, 7, 8, 9, 1, 1,
```

62. How can you add a column to a Pandas Data Frame?

Suppose there is an emp data frame that has information about a few employees. Let’s add an Address column to that data frame.



```
emp = {'Name': ['Sam', 'Prince', 'Tom', 'Andy'],
       'Height': [5.1, 6.2, 6.9, 7.2],
       'Qualification': ['Msc', 'MA', 'Msc', 'MBA']}
```

```
emp
```

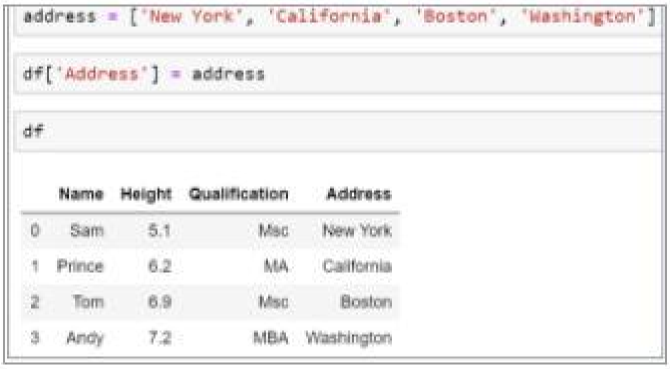
```
{'Name': ['Sam', 'Prince', 'Tom', 'Andy'],
 'Height': [5.1, 6.2, 6.9, 7.2],
 'Qualification': ['Msc', 'MA', 'Msc', 'MBA']}
```

```
df = pd.DataFrame(emp)
```

```
df
```

	Name	Height	Qualification
0	Sam	5.1	Msc
1	Prince	6.2	MA
2	Tom	6.9	Msc
3	Andy	7.2	MBA

Declare a list of values that will be converted into an address column.



```
address = ['New York', 'California', 'Boston', 'Washington']
```

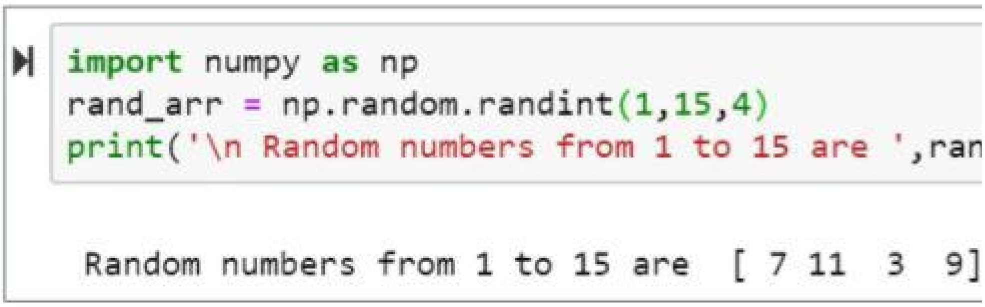
```
df['Address'] = address
```

```
df
```

	Name	Height	Qualification	Address
0	Sam	5.1	Msc	New York
1	Prince	6.2	MA	California
2	Tom	6.9	Msc	Boston
3	Andy	7.2	MBA	Washington

63. How will you print four random integers between 1 and 15 using NumPy?

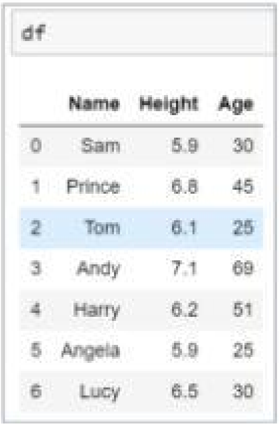
To generate Random numbers using NumPy, we use the random.randint() function.



```
import numpy as np
rand_arr = np.random.randint(1,15,4)
print('\n Random numbers from 1 to 15 are ',ran
```

```
 Random numbers from 1 to 15 are  [ 7 11  3  9]
```

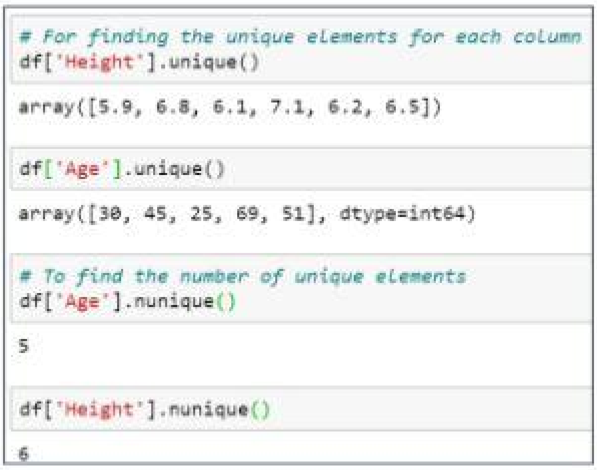
64. From the below DataFrame, how will you find each column's unique values and subset the data for Age<35 and Height>6?



```
df
```

	Name	Height	Age
0	Sam	5.9	30
1	Prince	6.8	45
2	Tom	6.1	25
3	Andy	7.1	69
4	Harry	6.2	51
5	Angela	5.9	25
6	Lucy	6.5	30

To find the unique values and number of unique elements, use the unique() and nunique() function.



```
# For finding the unique elements for each column
df['Height'].unique()
```

```
array([5.9, 6.8, 6.1, 7.1, 6.2, 6.5])
```

```
df['Age'].unique()
```

```
array([30, 45, 25, 69, 51], dtype=int64)
```

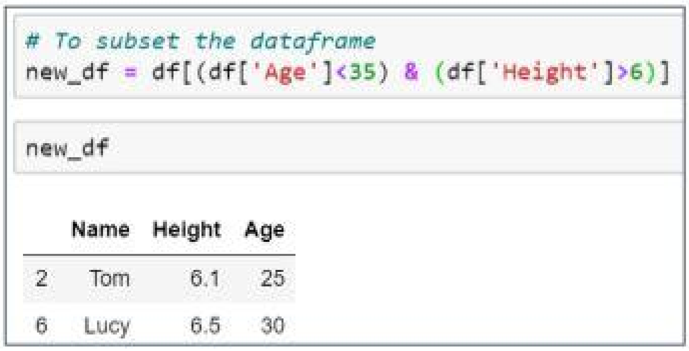
```
# To find the number of unique elements
df['Age'].nunique()
```

```
5
```

```
df['Height'].nunique()
```

```
6
```

Now, subset the data for Age<35 and Height>6.

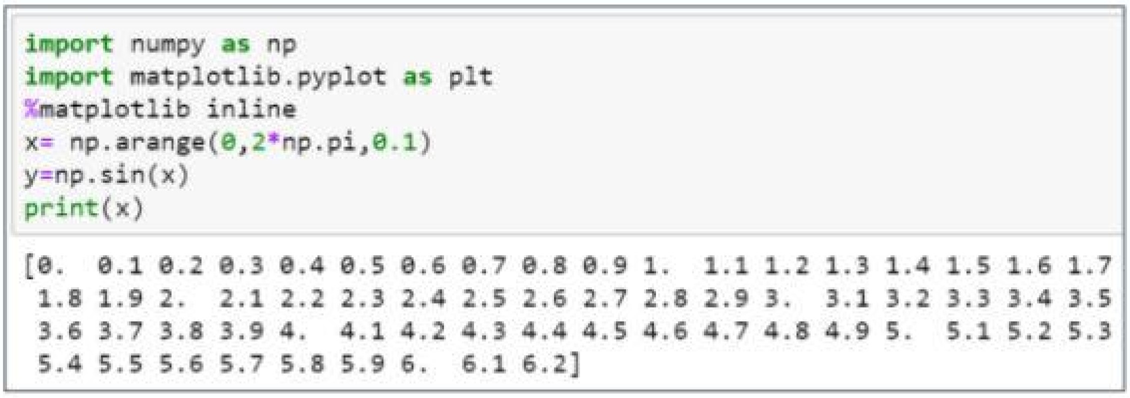


```
# To subset the dataframe
new_df = df[(df['Age']<35) & (df['Height']>6)]
```

```
new_df
```

	Name	Height	Age
2	Tom	6.1	25
6	Lucy	6.5	30

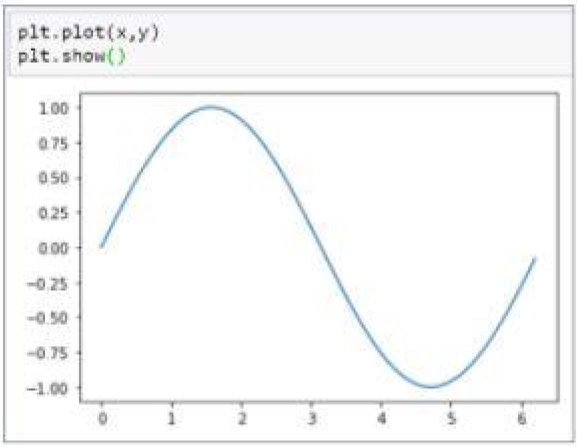
65. Plot a sine graph using NumPy and Matplotlib library in Python.



```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
x= np.arange(0,2*np.pi,0.1)
y=np.sin(x)
print(x)
```

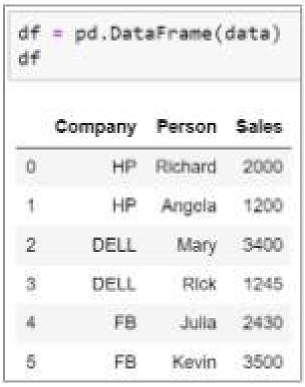
```
[0.  0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.  1.1 1.2 1.3 1.4 1.5 1.6 1.7
 1.8 1.9 2.  2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.  3.1 3.2 3.3 3.4 3.5
 3.6 3.7 3.8 3.9 4.  4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9 5.  5.1 5.2 5.3
 5.4 5.5 5.6 5.7 5.8 5.9 6.  6.1 6.2]
```

Below is the result sine graph.



```
plt.plot(x,y)
plt.show()
```

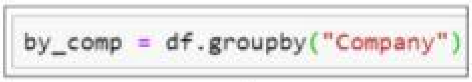
66. Using the below Pandas data frame, find the company with the highest average sales. Derive the summary statistics for the sales column and transpose the statistics.



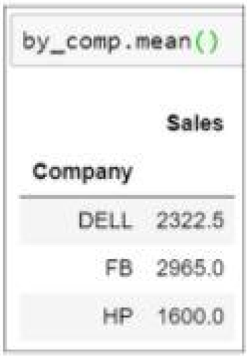
```
df = pd.DataFrame(data)
df
```

	Company	Person	Sales
0	HP	Richard	2000
1	HP	Angela	1200
2	DELL	Mary	3400
3	DELL	Rick	1245
4	FB	Julia	2430
5	FB	Kevin	3500

- Group the company column and use the mean function to find the average sales



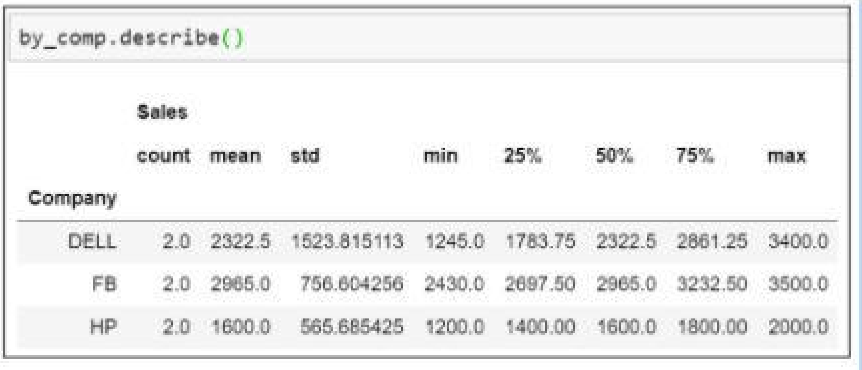
```
by_comp = df.groupby("Company")
```



```
by_comp.mean()
```

	Sales
Company	
DELL	2322.5
FB	2965.0
HP	1600.0

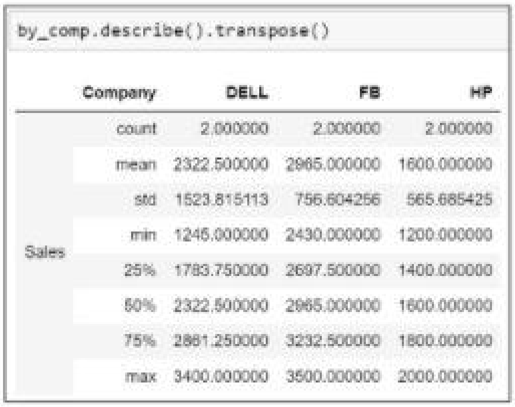
- Use the describe() function to find the summary statistics



```
by_comp.describe()
```

	Sales							
	count	mean	std	min	25%	50%	75%	max
Company								
DELL	2.0	2322.5	1523.815113	1245.0	1783.75	2322.5	2861.25	3400.0
FB	2.0	2965.0	756.604256	2430.0	2697.50	2965.0	3232.50	3500.0
HP	2.0	1600.0	565.685425	1200.0	1400.00	1600.0	1800.00	2000.0

- Apply the transpose() function over the describe() method to transpose the statistics



```
by_comp.describe().transpose()
```

	Company	DELL	FB	HP
Sales	count	2.000000	2.000000	2.000000
	mean	2322.500000	2965.000000	1600.000000
	std	1523.815113	756.604256	565.685425
	min	1245.000000	2430.000000	1200.000000
	25%	1783.750000	2697.500000	1400.000000
	50%	2322.500000	2965.000000	1600.000000
	75%	2861.250000	3232.500000	1800.000000
	max	3400.000000	3500.000000	2000.000000

So, those were the 65+ data analyst interview questions that can help you crack your next data analyst interview and help you become a data analyst.

Conclusion

Now that you know the different data analyst interview questions that can be asked in an interview, it is easier for you to crack for your coming interviews. Here, you looked at various data analyst interview questions based on the difficulty levels. And we hope this article on data analyst interview questions is useful to you.

On the other hand, if you wish to add another star to your resume before you step into your next data analyst interview, enroll in Simplilearn's Data Analyst Master's program, and master data analytics like a pro!

Unleash your potential with Simplilearn's Data Analytics Bootcamp. Master essential skills, tackle real-world projects, and thrive in the world of Data Analytics. Enroll now for a data-driven career transformation!