# Data Analystics Project

Before talking about the full-fledged data analysis process and diving into the details of individual methods, this chapter demonstrates some typical pitfalls one encounters when analyzing real-world data. We start our journey through the data analysis process by looking over the shoulders of two (pseudo) data analysts, Stan and Laura, working on some hypothetical data analysis problems in a sales environment. Being differently skilled, they show how things should and should not be done. Throughout the chapter, a number of typical problems that data analysts meet in real work situations are demonstrated as well. We will skip algorithmic and other details here and only briefly mention the intention behind applying some of the processes and methods. They will be discussed in depth in subsequent chapters.

## 2.1 The Setup

**Disclaimer** The data and the application scenario used in this chapter are fictional. However, the underlying problems are motivated by actual problems which are encountered in real-world data analysis scenarios. Explaining particular applicational setups would have been entirely out of the scope of this book, since in order to understand the actual issue, a bit of domain knowledge is often helpful if not required. Please keep this in mind when reading the following. The goal of this chapter is to show (and sometimes slightly exaggerate) pitfalls encountered in real-world data analysis setups and not the reality in a supermarket chain. We are painfully aware that people familiar with this domain will find some of the encountered problems strange, to say the least. Have fun.

**The Data** For the following examples, we will use an artificial set of data sources from a hypothetical supermarket chain. The data set consists of a few tables, which have already been extracted from an in-house database:[1]

[1] Often just getting the data is a problem of its own. Data analysis assumes that you have access to the data you need—an assumption which is, unfortunately, frequently not true.

M.R. Berthold et al., *Guide to Intelligent Data Analysis,*
Texts in Computer Science 42,
DOI 10.1007/978-1-84882-260-3_2, © Springer-Verlag London Limited 2010

- **Customers**: data about customers, stemming mostly from information collected when these customers signed up for frequent shopper cards.
- **Products**: A list of products with their categories and prices.
- **Purchases**: A list of products together with the date they were purchased and the customer card ID used during checkout.

**The Analysts** Stan and Laura are responsible for the analytics of the southern and northern parts, respectively, of a large supermarket chain. They were recently hired to help better understand customer groups and behavior and try to increase revenue in the local stores. As is unfortunately all too common, over the years the stores have already begun all sorts of data acquisition operations, but in recent years quite a lot of this data has been merged—however, still without a clear picture in mind. Many other stores had started to issue frequent shopping cards, so the directors of marketing of the southern and northern markets decided to launch a similar program. Lots of data have been recorded, and Stan and Laura now face the challenge to fit existing data to the questions posed. Together with their managers, they have sat down and defined three data analysis questions to be addressed in the following year:

- differentiate the different customer groups and their behavior to better understand their impact on the overall revenue,
- identify connections between products to allow for cross selling campaigns, and
- help design a marketing campaign to attract core customers to increase their purchases.

Stan is a representative of the typical self-taught data analysis newbie with little experience on the job and some more applied knowledge about the different techniques, whereas Laura has some training in statistics, data processing, and data analysis process planning.

## 2.2 Data Understanding and Pattern Finding

The first analysis task is a standard data analysis setup: customer segmentation—find out which types of customers exist in your database and try to link them to the revenue they create. This can be used later to care for clientele that are responsible for the largest revenue source or foster groups of customers who are under-represented. Grouping (or *clustering*) records in a database is the predominant method to find such customer segments: the data is partitioned into smaller subsets, each forming a more coherent group than the overall database contains. We will go into much more detail on this type of data analysis methods in Chap. 7. For now it suffices to know that some of the most prominent clustering methods return one typical example for each cluster. This essentially allows us to reduce a large data set to a small number of representative examples for the subgroups contained in the database.

**Table 2.1** Stan's clustering result

| Cluster-id | Age | Customer revenue |
|---|---|---|
| 1 | 46.5 | € 1,922.07 |
| 2 | 39.4 | € 11,162.20 |
| 3 | 39.1 | € 7,279.59 |
| 4 | 46.3 | € 419.23 |
| 5 | 39.0 | € 4,459.30 |

**The Naive Approach** Stan quickly jumps onto the challenge, creates a dump of the database containing customer purchases and their birth date, and computes the age of the customers based on their birth date and the current day. He realizes that he is interested in customer clusters and therefore needs to somehow aggregate the individual purchases to their respective "owner." He uses an aggregating operator in his database to compute the total price of the shopping baskets for each customer. Stan then applies a well-known clustering algorithm which results in five prototypical examples, as shown in Table 2.1.

Stan is puzzled—he was expecting the clustering algorithm to return reasonably meaningful groups, but this result looks as if all shoppers are around 40–50 years old but spend vastly different amount of money on products. He looks into some of the customers' data in some of these clusters but cannot seem to find any interesting relations or any reason why some seem to buy substantially more than others. He changes some of the algorithm's settings, such as the number of clusters created, but the results are similarly uninteresting.

**The Sound Approach** Laura takes a different approach. Routinely she first tries to understand the available data and validates that some basic assumptions are in fact true. She uses a basis data summarization tool to report the different values for the string attributes. The distribution of first names seems to match the frequencies she would expect. Names such as "Michael" and "Maria" are most frequent, and "Rosemarie" and "Anneliese" appear a lot less often. The frequencies of the occupations also roughly match her expectations: the majority of the customers are employees, while the second and third groups are students and freelancers, respectively. She proceeds to checking the attributes holding numbers. In order to check the age of the customers, she also computes the customers' ages from their birth date and checks minimum and maximum. She spots a number of customers who obviously reported a wrong birthday, because they are unbelievably young. As a consequence, she decides to filter the data to only include people between the ages of 18 and 100. In order to explore the data more quickly, she reduces the overall customer data set to 5,000 records by random sampling and then plots a so-called histogram, which shows different ranges of the attribute *age* and how many customers fall into that range. Figure 2.1 shows the result of this analysis.

This view confirms Laura's assumptions—the majority of shoppers is middle aged, and the number of shoppers continuously declines toward higher age groups.

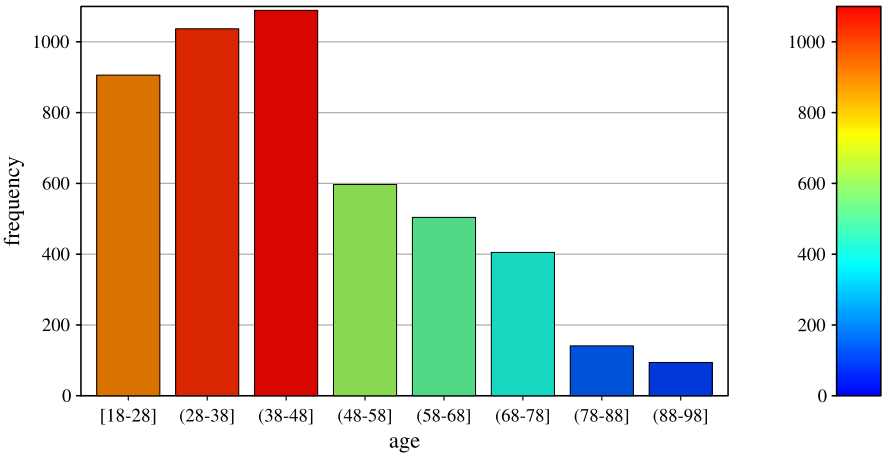


**Fig. 2.1** A histogram for the distribution of the value of attribute *age* using 8 bins

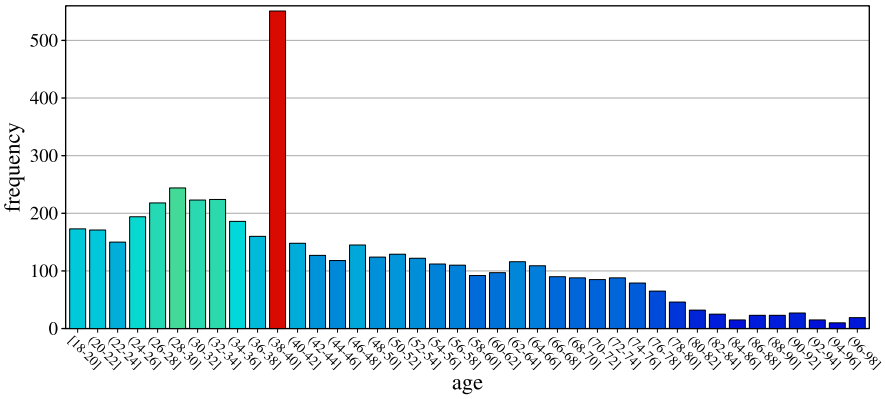


**Fig. 2.2** A histogram for the distribution of the value of attribute *age* using 40 bins

She creates a second histogram to better inspect the subtle but strange cliff at around age 48 using finer setting for the bins. Figure 2.2 shows the result of this analysis.

Surprised, she notices the huge peak in the bin of ages 38–40. She discusses this observation with colleagues and the administrator of the shopping card database. They have no explanation for this odd concentration of 40-year-old people either. After a few other investigations, a colleague of the person who—before his retirement—designed the data entry forms suspects that this may have to do with the coding of missing birth dates. And, as it turns out, this is in fact the case: forms where people entered no or obviously nonsensical birth dates were entered into the form as zero values. For technical reasons, these zeros were then converted into the Java 0-date which turns out to be January 1, 1970. So these people all turn up with the same birth date in the customer database and in turn have the same age after the

**Table 2.2** Laura's clustering result

| Cluster | Age | Avg. cart price | Avg. purchases/ month |
|---|---|---|---|
| 1 | 75.3 | € 19.- | 5.6 |
| 2 | 42.1 | € 78.- | 7.8 |
| 3 | 38.1 | € 112.- | 9.3 |
| 4 | 30.6 | € 16.- | 4.8 |
| 5 | 44.7 | € 45.- | 3.7 |

conversion Laura performed initially. Laura marks those entries in her database as "missing" in order to be able to distinguish them in future analyses.

Similarly, she inspects the shopping basket and product database and cleans up a number of other outliers and oddities. She then proceeds with the customer segmentation task. As in her previous data analysis projects, Laura first writes down her domain knowledge in form of a cognitive map, indicating relationships and dependencies between the attributes of her database. Having thus recalled the interactions between the variables of interest, she is well aware that the length of customer's history and the number of overall shopping trips affect the overall basket price, and so she settles on the average basket price as a better estimator for the value of a particular customer. She considers also distinguishing the different product categories, realizing that those, of course, also potentially affect the average price. For the first step, she adds the average number of purchases per month, another indicator for the revenue a customer brings in. Data aggregation is now a bit more complex, but the modern data analysis tool she is using allows her to do the required joining and pivoting operations effortlessly. Laura knows that clustering algorithms are very sensitive to attributes with very different magnitudes, so she normalizes the three attributes to make sure they all three contribute equally to the clustering result. Running the same clustering algorithm that Stan was using, with the same setting for the number of clusters to be found, she gets the result shown in Table 2.2.

Obviously, there is a cluster (#1) of older customers who have a relatively small average basket price. There is also another group of customers (#4) which seems to correlate to younger shoppers, also purchasing smaller baskets. The middle-aged group varies wildly in price, however. Laura realizes that this matches her assumption about family status—people with families will likely buy more products and hence combine more products into more expensive baskets, which seems to explain the difference between clusters #2/#3 and cluster #5. The latter also seem to shop significantly less often. She goes back and validates some of these assumptions by looking at shopping frequency and average basket size as well and also determines the overall impact on store revenues for these different groups. She finally discusses these results with her marketing and campaign specialists to develop strategies to foster the customer groups which bring in the largest chunk of revenue and develop the ones which seem to be under-represented.

## 2.3 Explanation Finding

The second analysis goal is another standard shopping basket analysis problem: find product dependencies in order to better plan campaigns.

**The Naive Approach** Stan recently read in a book on practical data analysis how association rules can find arbitrary such connections in market basket data. He runs the association rule mining algorithm in his favorite data analysis tool with the default settings and inspects the results. Among the top-ranked generated rules, sorted by their confidence, Stan finds the following output:

```
'foie gras' (p1231) <- 'champagne Don Huberto' (p2149),
      'truffle oil de Rossini' (p578) [s=1E-5, c=75%]
'Tortellini De Cecco 500g' (p3456)'
   <- 'De Cecco Sugo Siciliana' (p8764) [s=1E-5, c=60%]
```

He quickly infers that this representation must mean that foie gras is bought whenever champagne and truffle oil are bought together and similarly for the other rule. Stan knows that the confidence measure $c$ is important, as it indicates the strength of the dependency (the first rule holds in 3 out of 4 cases). He considers the second measure of frequency $s$ to be less important and deliberately ignores its fairly small value. The two rules shown above are followed by a set of other, similarly luxury/culinary product-oriented rules. Stan concludes that luxury products are clearly the most important products on the shelf and recommends to his marketing manager to launch a campaign to advertise some of the products on the right side of these rules (champagne, truffle oil) to increase the sales of the left side (foie gras). In parallel, he increases orders for these products, expecting a recognizable increase in sales. He proudly sends the results of his analysis to Laura.

**The Sound Approach** Laura is puzzled by those nonintuitive results. She reruns the analysis and notices the support values of the rules extracted by Stan—some of the rules Stan extracted have indeed a remarkably high confidence, and some do almost forecast shopping behavior. However, they have very low support values, meaning that only a small number of shopping baskets containing the products were ever observed. The rules that Stan found are not representative at all for his customer base. To confirm this, she runs a quick query on her database and sees that, indeed, there is essentially no influence on the overall revenue.

She notices that the problem of low support is caused by the fact that Stan ran the analysis on product IDs, so in effect he was forcing the rules to differentiate between brands of champagne and truffle oil. She reruns the analysis based on the product categories instead, ranks them by a mix of support and confidence, and finds a number of association rules with substantially higher support:

```
tomatoes <- capers, pasta  [s=0.007, c=32%]
tomatoes <- apples  [s=0.013, c=22%]
```

Laura focuses on rules with a much higher support measure $s$ than before and also realizes that the confidence measure $c$ is significantly higher than one would expect

by chance. The first rule seems to be triggered by a recent fashion of Italian cooking, whereas the apple/tomato-rule is a known aspect.

However, she is still irritated by one of the rules discovered by Stan, which has a higher than suspected confidence despite a relatively low support. Are there some gourmets among the customers who prefer a very specific set of products? Rerunning this analysis on the shopping card owners yields almost the same results, so the (potential) gourmets appear among their regular customers. Just to be sure, she inspects how many different customers (resp. shopping cards) occur for baskets that support this rule. As she had conjectured, there is a very limited number of customers that seem to have a strong affection for these products. Those few customers have bought this combination frequently, thus inflating the overall support measure (which refers to shopping baskets). This means that the support in terms of the *number of customers* is even smaller than the support in terms of *number of shopping baskets*. The response to any kind of special promotion would fall even shorter than expected from Stan's rule.

Apparently the time period in which the analyzed data has been collected influences the results. Thinking about it, she develops an idea how to learn about changes in the customers shopping behavior: She identifies a few rules, some rather promising other well-known facts, and decides to monitor those combinations on a regular basis (say quarterly). She got to know that a chain of liquor stores will soon open a number of shops close to the own markets, so she picks some rules with beverages in their conclusion part to see if the opening has any impact on the established shopping patterns of the own customers. As she fears a loss of potential sales, she plans a comparison of rules obtained not only over time but also among markets in the vicinity of such stores versus the other markets. She wonders whether promoting the products in the rule's antecedent may help to bring back the customer and decides to discuss this with the marketing&sales team to determine if and where appropriate campaigns should be launched, once she has the results of her analysis.

## 2.4 Predicting the Future

The third and final analysis goal we consider in this brief overview is a forecasting or prediction problem. The idea is to find some relationship in our existing data that can help us to predict if and how customers will react to coupon mailings and how this will affect our future revenue.

**The Naive Approach** Stan believes that no detailed analysis is required for this problem and notices that it is fairly straightforward to monitor success. He has seen at a competitor how discount coupons attract customers to purchase additional products. So he suggests launching a coupon campaign that gives customers a discount of 10% if they purchase products for more than €50. This coupon is mailed to all customers on record. Throughout the course of the next month, he carefully monitors his database and is positively surprised when he sees that his campaign is obviously

working: the average price of shopping baskets is going up in comparison with previous months. However, at the end of the quarter he is shocked to see that overall revenues for the past quarter actually fell. His management is finally fed up with the lack of performance and fires Stan.

**The Sound Approach** Laura, who is promoted to head of analytics for the northern and southern super market chain first cancels Stan's campaign and looks into the underlying data. She quickly realizes that even though quite a number of customers did in fact use the coupons and increased their shopping baskets, their average number of baskets per month actually went down—so quite a number of people seem to have simply combined smaller shopping trips to be able to benefit from the discount offer. However, for some shoppers, the combined monthly shopping basket value did go up markedly, so there might be value here. Laura wonders how she can discriminate between those customers who simply use the coupons to discount their existing purchases and those who are actually enticed to purchase additional items. She notices that one of the earlier generated customer segments correlates better than others with the group of customers whose revenue went up—this fraction of customers is significantly higher than in the other groups. She considers using this very simple, manually designed predictor for a future campaign but wants to first make sure that she cannot do better with some smarter techniques. She decides that in the end it is not so important if she can actually understand the extracted model but only how well it performs.

To provide good starting points for the modeling technique, she decides to generate a few potentially informative attributes first. Models that rely on thousands of details typically perform poor, so providing how often every product has been bought by the customer in the last month is not an option for her. To get robust models, she wants to aggregate the tiny bits of information, but what kind of aggregation could be helpful? She returns to her cognitive map to review the dependencies. One aspect is the availability of competitors: She reckons that customers may have alternative (possibly specialized) markets nearby but have been attracted by the coupon this time, keeping them away from the competitors. She decides to aggregate the money spent by the customer per month for a number of product types (such as *beverages*, thinking of the chain of liquor stores again). She conjectures that customers that perform well on average, but underperform in a specific segment only, may be enticed by the coupon to buy products for the underperforming segment also. Providing the segment performance before and after Stan's campaign should help a predictor to detect such dependencies if they exist.

The cognitive map brings another idea into her mind: people who appreciate the full assortment but live somewhat further away from the own stores may see the coupon as a kind of travel compensation. So she adds a variable expressing a coarse estimation of the distance between the customer home and the nearest available market (which is only possible for the shopping card owners). She continues to use her cognitive map to address many different aspects and creates attributes that may help to verify her hypotheses. She then investigates the generated attributes visually and also technically by means of feature selection methods.

After selecting the most promising attributes, she trains a classifier to distinguish the groups. She uses part of the data to simulate an independent test scenario and thereby evaluates the expected impact of a campaign—are the costs created by sending coupons to customers who do not purchase additional products offset by customers buying additional items? After some additional model fine tuning, she reaches satisfactory performance. She discusses the results with the marketing&sales team and deploys the prediction system to control the coupon mailings for the next quarter. She keeps monitoring the performance of these coupon campaigns over future quarters and updates her model sporadically.

## 2.5 Concluding Remarks

In this chapter we have, very briefly and informally, touched upon a number of issues data analysts may encounter while making sense of real-world data. Many other problems can arise, and many more methods for data analysis exist in the academic literature and in real-world data analysis tools. We will attempt at covering the most prominent and most often used examples in the following chapters.

Note that one of the biggest problems data analysts very often have is that the data they get is not suited to answer the questions they are asked. For instance, if we were supposed to use the data in our customer database to find out how to differentiate Asian shopping behavior from European, we would have a very hard time. This data can only be used to distinguish between different types of European shoppers because it contains data from European markets only. Note also that we are (why ever) assuming that we used a nice, representative sample of all different types of European shoppers to generate the data—very often this is not the case, and the data itself is already biased and will bias our analysis results—in this example we could be heavily biased by the type of supermarket chain we used to record the data in the first place. An upscale delicatessen supermarket will have dramatically different shopping patterns than the low-scale discounter. We will be discussing these points later in more depth as well.