BIBA: Business Intelligence and Big Data

2018-09-26

Jens Ulrik Hansen

Exploratory Data Analysis

Questions

- Any questions from last time?

Today’s program

- Piping in R
- Recap on data and data types
- Exploratory Data Analysis and Descriptive Statistics
- Plotting in R with ggplot2
- Work on the synopsis and the second hand-in of it

Piping in R

- See the Jupyter notebook “Piping in R.ipynb”
  - https://notebooks.azure.com/jensuh/libraries/BIBA-2018/html/Piping%20in%20R.ipynb

Today’s program

- Piping in R
- Recap on data and data types
- Exploratory Data Analysis and Descriptive Statistics
- Plotting in R with ggplot2
- Work on the synopsis and the second hand-in of it

The data science/BI process recap

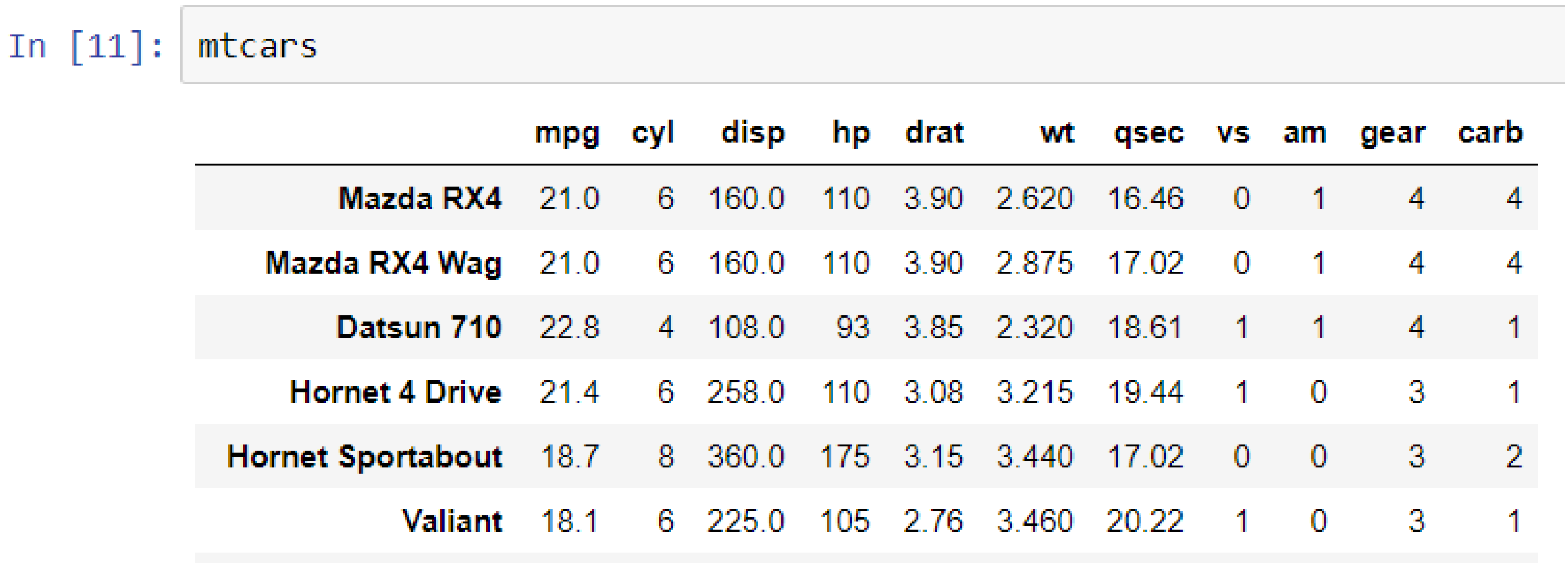
- Using data (and data analysis) to solve business problems
  1. identify business problem
  2. collect data
  3. prepare data
  4. analyze data
  5. conclude and communicate

The data science/BI process recap

- Using data (and data analysis) to solve business problems
  1. identify business problem
  2. collect data
  3. prepare data
     - Load the data
     - Transform the data
     - Clean the data
     - **Explore and understand the data**
  4. analyze data
  5. conclude and communicate

Data recap

- ***Tidy data***
  - A row column spreadsheet kind of format where:
    - Each rows represent one ***case/observation/object***
    - Each column represent an ***attribute/variable/feature*** of the cases



```
In [11]: mtcars
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1

Types of Attributes (scales of measurement)

- There are different types of attributes

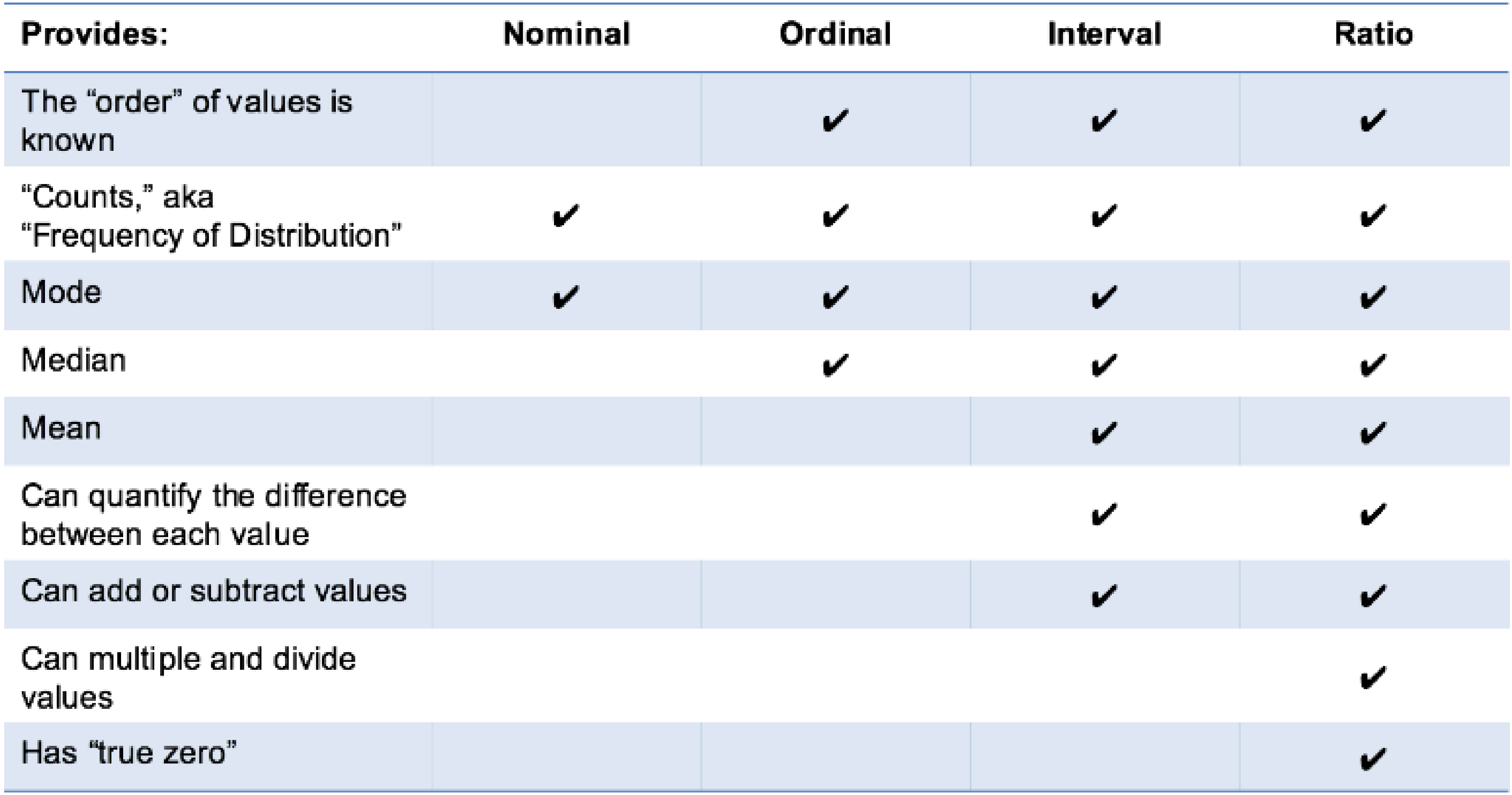
  categorical

  - ***Nominal***
    - Examples: ID numbers, eye color, zip codes
  - ***Ordinal***
    - Examples: rankings (e.g., taste of potato chips on a scale from 1 to 10), grades, height in {tall, medium, short}

  numerical

  - ***Interval***
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - ***Ratio***
    - Examples: temperature in Kelvin, length, time, counts
- The numeric and graphical exploration differs from the two types of variables

Types of Attributes (scales of measurement)



Provides:	Nominal	Ordinal	Interval	Ratio
The “order” of values is known		✔	✔	✔
“Counts,” aka “Frequency of Distribution”	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has “true zero”				✔

Different data types in R

- Variables can be mapped to the following different data types in R:
  - ***Numeric*** (floating point numbers) – most used data type for Interval and Ration
    - 5.34, pi, 3.3333, 1, 5.0, as.numeric(5L)
  - ***Integer***
    - 5L, as.integer(10.0)
  - ***Character (strings)***
    - “In this book”, “10.4”, as.character(10.4), paste(“Hello”, “my”, “name”)
  - ***Logical*** (true or false)
    - TRUE, FALSE, T, F, 10 == 10, 10 == 2, 10 == 20/10
  - ***Factor (categorical)***
    - “yes”/”no”, “Country of origin” (numbers underneath)
    - Can be given an order
  - ***Dates***
    - “2016-09-22” (from the Sys.Date() command)
    - (can internally be stored as integers or floating point numbers)
  - ***Time-Date***
    - "2016-09-22 12:06:30 CEST" (from the Sys.time() command)
- The `class` function in R can tell the data type (fx `class(10.4)`)

Exploratory Data Analysis (EDA)

What is “Exploratory data analysis”

- It is not a completely agreed upon and well-defined term
  - R for Data Science: *“EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind.”*
  - Wikipedia*: “In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual method.”*
- EDA is about
  - Getting to know/understand your data
  - Getting to know and improve the quality of your data
- EDA often involves
  - A lot of plotting and visual exploration
  - summary statistics (descriptive statistics) of the data
- ***EDA is the initial necessary stage of getting to know your data***

EDA step 1 – the basics

- What variables are in your data set?
- I What does the different variables represent?
- I What observations does the data set contain?
- What is the (intended) type/scale of measurement of each of the variables/features in your data set
- Map each variable to the right data type in R
  - What is the data type of the variable once you loaded it into R?
    - R is smart so it can map most variables to their right data type in R
    - However, it often does not do a complete job, especially dates can cause troubles, thus use the data transformation we learned last time to transform the variable into the right data type

Next step: Asking questions about your data

- Standard questions about you data
  - What type of variation occurs within my variables?
  - What type of variation occurs between my variables
  - What is the quality of my data? Are there outliers, missing data etc.

Variation within a variable

- The variation with a variable is the tendency for the variable to change from observation to observation
- A variation within a variable can be due to several things
  - The variable simply varies within the population. If we measure the height of people in this room, we will get different heights for different people
  - There can be measurement errors or noise. If we measure the same person several times with very high accuracy, we are bound to get a little bit of different heights every time
- The variation of a variable can be understood through its ***distribution***, which can be visualized as well as quantified . . .

The distribution of a variable

- The distribution of a variable is how the actual values of the values are distributed across all possible values
- The distribution of a variable can be visualized in several ways depending on its type:
  - For categorical variables use a ***bar plot***
    - In R: `barplot(summary(myVariable))` or just `plot(myVariable)`
  - For numerical variables: Use a ***histogram*** or a ***boxplot***
    - Histogram in R: `hist(myVariable)`
    - Boxplot in R: `boxplot(myVariable)`
- See the Jupyter notebook “Exploratory Data Analysis.ipynb” for examples

Quantifying what we see in distribution plots

- With descriptive statistics we can measure the tendencies we see in the distribution plots
  - For categorical variables we can just want to see a table of the number of each category (in R using the `summary` or `table` functions)
  - For numerical variables there many measures to use. They fall in two types:
    - ***Centrality tendencies***: Where do most values fall? What values am I most likely to get if I draw a random value from the distribution?
    - ***Variation/spread tendencies***: How spread out is my data? What are the most extreme values I can get by a random draw? How far from the “centrality” of the distribution am I like to end up by a random draw?

Centrality tendencies

- The ***mean*** of a distribution:
  - (Also known as the arithmetic mean)
  - The sum of all values divided by the number of values: $sum(x_1 + x_2 + ... + x_n)/n$
  - In R: `mean(myVariable)`
- The ***median*** of a distribution:
  - The value such that half the of value are below that value and the other half is above that value
  - If you sort the list of values, the median is the middle value
  - In R: `median(myVariable)`

Centrality tendencies

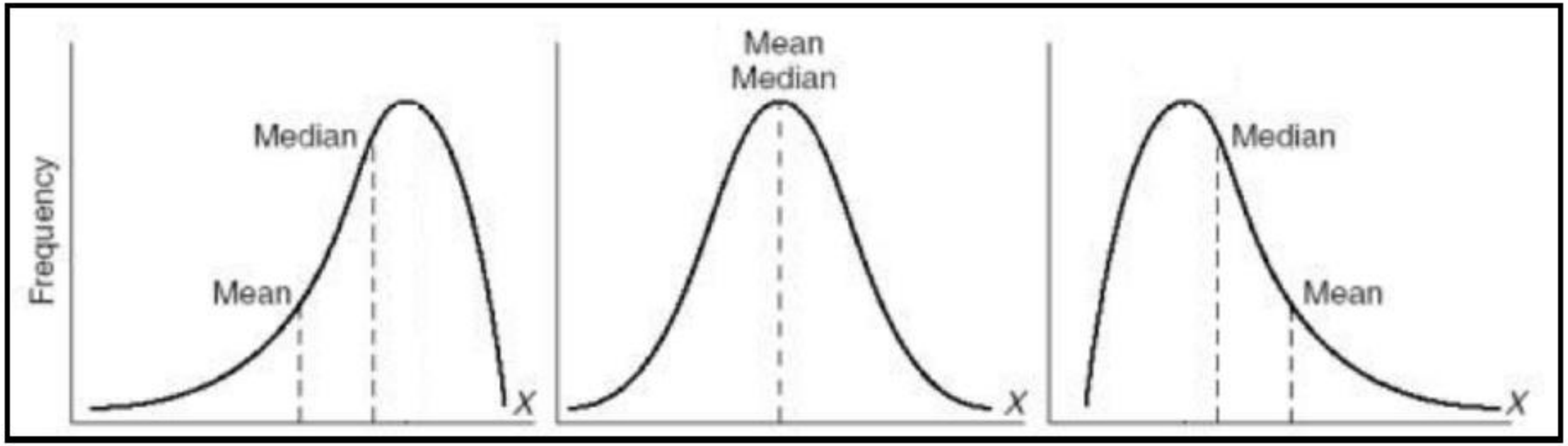
- The ***mode*** of a distribution:
  - The mode is the value that occur most often
  - For numerical variables this depends on the precission/binwidth and is not always of interest
  - For categorical variables it make more sense
  - Note the mode function in R is NOT this statistical mode
  - In R first install and load the “modeest” package (`install.packages("modeest"); library(modeest)`), then: mfv(myVariable)

Centrality tendencies

- ***Quantiles*** of a distribution:
  - Like the mode, but with different location than half way.
  - The "First quartile" is such that 25% of the values are below this value (and 75% above)
  - The second quartile is the median
  - The third quartile is such that 75% of the values are below this value (and 25% above)
  - We can also talk about quantiles in general, such as the 35% quantile, i.e. the value such that 35% of the values are below this value
  - In R the `summary` function provide the 1st and 3rd quartiles, while specific quantiles can be calculated by the `quantile` function, for instance:
    - `quantile(diamonds$carat, c(0.25, 0.35, 0.5, 0.75))`

Mean vs median

- These values can be very different, especially when the distribution is ***skewed***
  - ***Left skewed*** (negative skewed): The tail to the left is longer than the tail to the right, the mean is less than the median
  - ***Right skewed*** (positive skewed): The tail to right is longer than the tail to the left, the mean is greater than the median

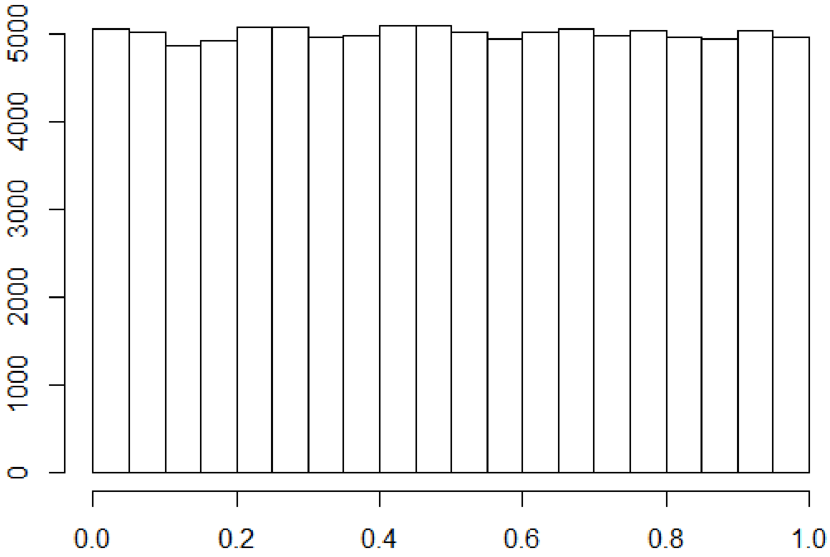


Variation/spread tendencies

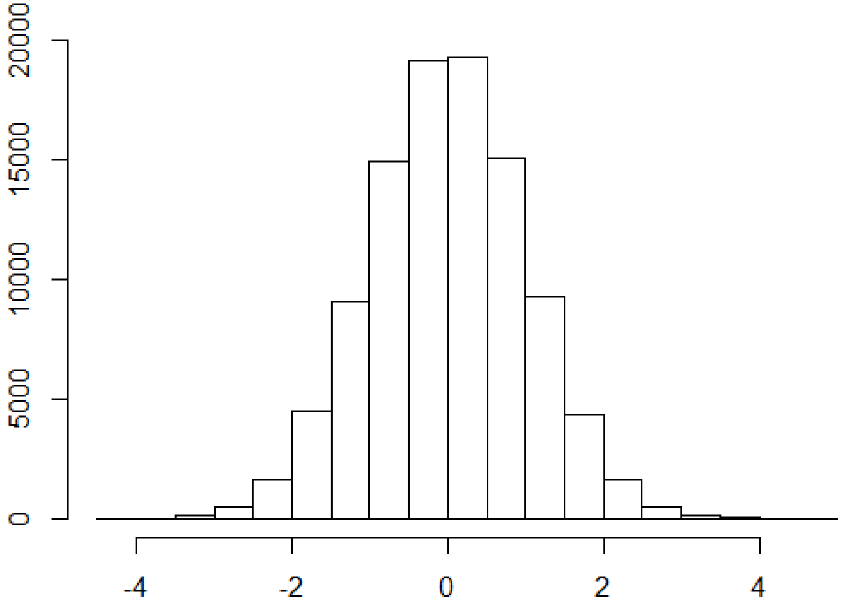
- The ***range*** of a distribution:
  - The maximum value minus the minimum value, or just the maximum and minimum values
  - In R: `min(myVariable)`, `max(myVariable)`, and `range(myVariable)`
- The ***variance*** of a distribution:
  - A measure for how far from the mean values of variable are:
  - *sum*$((x_1 - \mathit{mean})^2 + (x_2 - \mathit{mean})^2 + \ldots +(x_n - \mathit{mean})^2) / (n - 1)$
  - Note the “(n - 1)” it is not an error, it has to do with the *degrees of freedom* (we will not get into this any futher)
  - Note, the square makes the signs not important and make extreme values more important
  - In R: `var(myVariable)`
- The ***standard deviation*** of a distribution:
  - Note that the unit of the variance is the square of the unit of the variable. Sometimes it is nice to have measure in the same unit as the variable
  - The standard deviation is the square root of the variance
  - I In R: `sd(myVariable)`

Special distributions

- The uniform distribution:



- The normal distribution:



Variation between variables

- Covariance/correlation/causation
  - We will talk much more about this next time when we talk about linear models and linear regression
  - Also, read more about plotting to variables against each other in chapter 7 of “R for Data Science”
  - How to plot and what descriptive statistics to look at, depends on what are the types of the involved variables. There are the following three cases:
    - Two categorical variables
    - Two numerical variables
    - One categorical variable and one numerical variable

Two categorical variables

- Plotting
  - ***Mosaic plot***
    - This is not always that usefull a plot
    - In R: `mosaicplot(myFactorVar1 ~ myFactorVar2)`
- Descriptive statistics
  - A ***table***
    - Showing the number of cases in each combination of values of the categorical variables
    - In R: `table(myFactorVar1, myFactorVar2)`

Two numerical variables

- Plotting
  - ***scatter plot***
    - The classic x-y-plot as you know the from kindergarten…
    - In R: `plot(myNumericVar1, myNumericVar2)`
- Descriptive statistics
  - ***Pearson's correlation coefficient***
    - The standard correlation coefficient to measure *linear* correlation
    - In R: `cor(myNumericVar1, myNumericVar2)`
    - Exactly how to interpret this and different types of correlation as well as problems measuring correlation, we will talk a lot about next time
    - We will also skip the issue of significance today (see the chapter in the book)

A categorical variable and a numerical variable

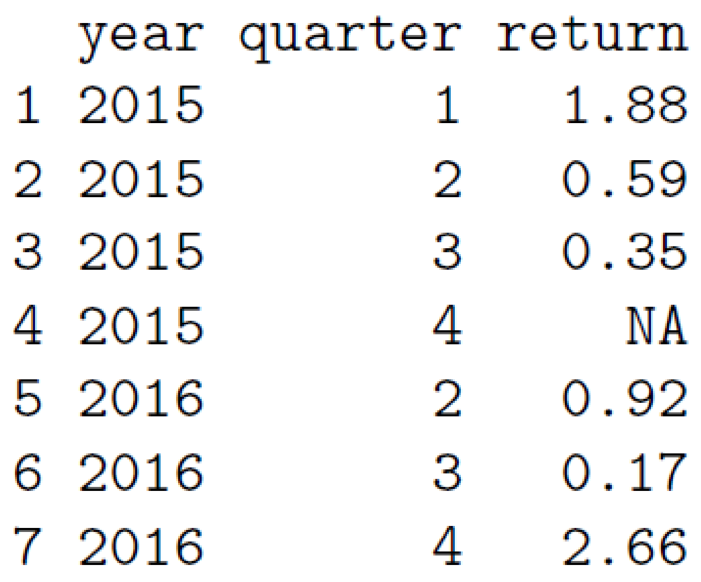
- Plotting
  - ***boxplot***
    - Like the previously, just one box for each value of the categorical variable
    - In R: `boxplot(myNumericVar ~ myFactorVar)`
- Descriptive statistics
  - numeric descriptive statistics for each group of the factor values

Data quality, missing values, and outliers

- ***Garbage-in-garbage-out (GIGO)***: If your data is of poor quality so will any analysis you base on it be, as well
  - Thus, you need to make the data as good as it can get before you start your analysis
  - During an analysis, you might need to go back and improve the quality of your data
- ***Missing values*** can affect you analysis greatly
- So can ***outliers***

Missing values

- Values can be:
  - ***Explicitly missing*** (in R represented by NA)
  - ***Implicitly missing*** (there is no record of it in the data)



```
  year quarter return
1 2015       1   1.88
2 2015       2   0.59
3 2015       3   0.35
4 2015       4     NA
5 2016       2   0.92
6 2016       3   0.17
7 2016       4   2.66
```

What are missing values?

- An explicit `NA` may represent that the data is simply not available
- An implicit missing value might represent an error in the data
- In either case, it usually means that the data is not available.
- That the data is not available can mean:
  - The data was not collected
  - The data is lost
  - The data does not exist
  - The data cannot exist

How to treat missing values

- Whatever is the reason for the missing data and what type of data it is, have implications for how to treat the missing values
  - For instance, in weather data it might make sense to interpolate data
  - In sales data, it might make sense to replace a missing value with 0
  - The context decide!
- Note, however, several functions and models in R can ignore missing values, and depending on the type and frequency of the missing values, the result might be fine by just ignoring the missing values!

Outliers

- Outliers are values that fall well outside the central tendency of your variables (as we saw in the boxplots)
- Outliers can represent different things:
  - Special extreme events. In betting data, the Football world cup final will likely generate an outlier; or a closing day in transaction data
  - Errors in data. A sales entry of a million times the usual sale is probably an error
- Sometimes special tricks can be used to deal with outliers
  - In the first case, we want the information present
  - In the second case, we might want to replace the outliers by a missing value

Plotting in R using ggplot2

Plotting in R with ggplot2

- See the notebook “Plotting in R with ggplot2”

Work on the synopsis and the second hand-in of it

Work on the synopsis and the second hand-in of it

- See also Moodle
- Please hand in a Jupyter notebook with an Exploratory Data Analysis of the dataset for your business case on Moodle
- Deadline is September 30 at 23:55.
- *Do an Exploratory Data Analysis, including the necessary data transformation and cleaning, on the data (or a part of) you have decided to use for your business case. That is, load the data into R in a Jupyter notebook. Look at what variables you have and what their scale of measurement and types are. Do the necessary transformation such that the variables are in a format in R you can work with. Then investigate the individual variables and their distribution using plotting and descriptive statistics. Finally, select some pairs of variable for which you think their could be an interesting relationship and plot the relationship and calculate the relevant descriptive statistics. Finally, download the notebook (right click on in the Azure notebook Library overview) and upload it to Moodle.*