The Art of Data Visualization

What story can data tell us ?

Kasper Kurzyński

What is Data Visualization?

- According to definition: „data visualization is the graphical representation of information and data, by using visual elements like charts, graphs and maps to provide an accesible way to see and understand trends, outliers and patterns in data.

Why do we need to visualize data?

EASIER FOR BRAIN TO UNDERSTAND

GAIN INSIGHTS

BETTER COMMUNICATION ABOUT DATA FINDINGS

MANY FIELDS CAN BENEFIT FROM IT (FINANCE, HEALTH, SPORT…)

BUILD CREATIVITY

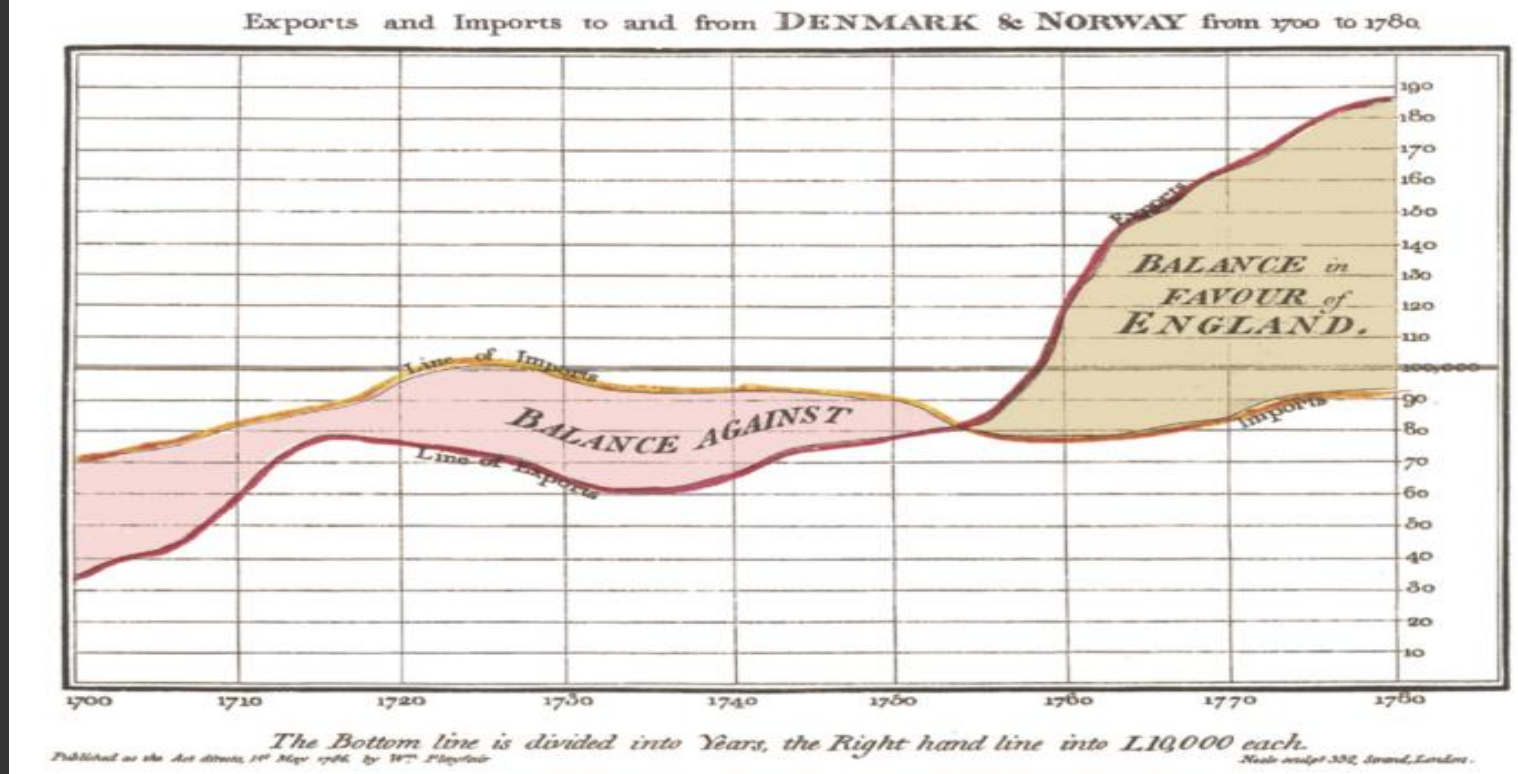
A little bit of history

- Charles Joseph Minard – was a French engineer born in 1781 who made significant innovations in combining the fields of statistics and cartography. He is known for his depiction of Napoleon’s invasion of Russia in 1812.

- William Playfair – was born in 1759. He was Scottish engineer and economist who also took part in the storming of the Bastille during the French Revolution. He is a founder of graphical methods of statistics. Playfair presented data using for example: bar charts or time series charts.

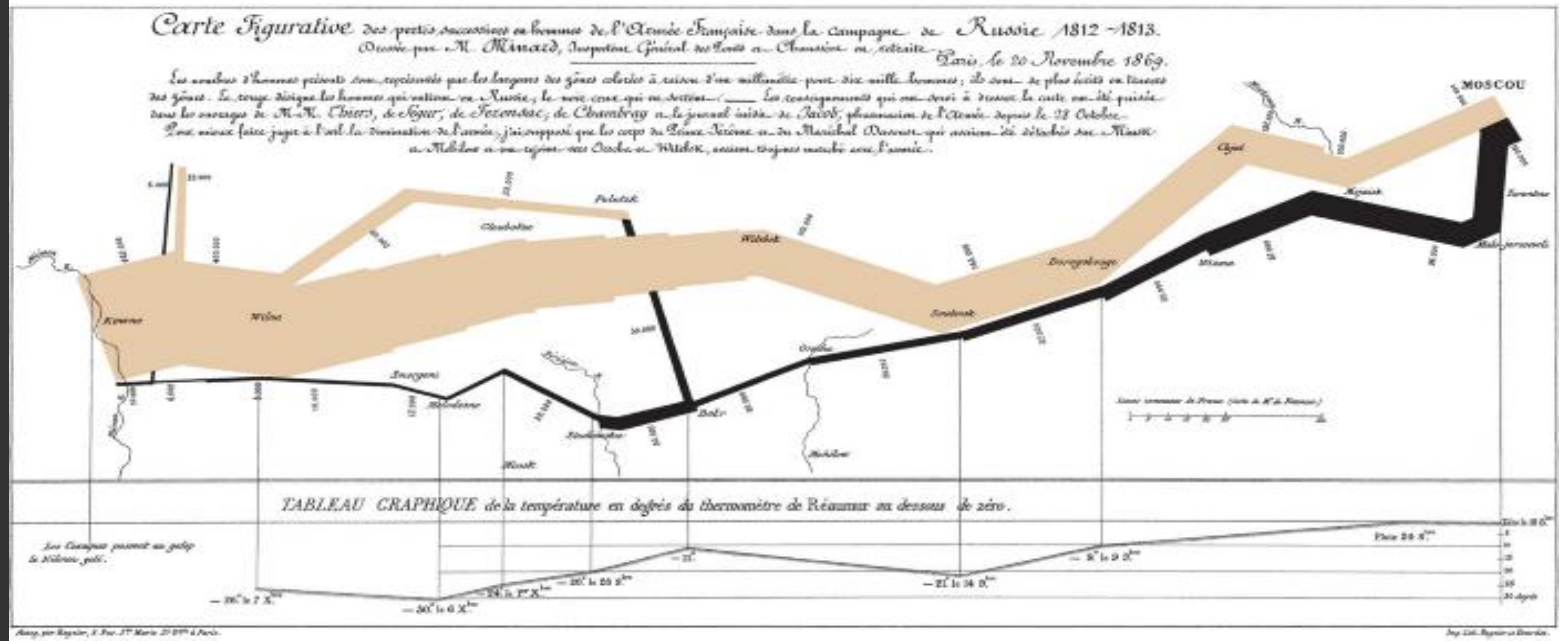
Balance of trade for England by W. Playfair

- Chart presents shaping of the volume of exports, imports and the mutual balance over the period of several dozen years.



Napoleon's invasion of Russia in 1812 by Ch. J. Minard

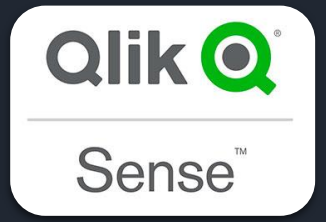
- Map shows the fate of Napoleon's army during the invasion of Russia. The graphics include army movements, important battles and the number of soldiers that Napoleon had at his disposal.



Popular data visualization tools

Python Libraries (matplotlib, pandas, seaborn…)

R Libraries (ggplot2)



QlikView / QlikSense

Microsoft PowerBI

Tableau

Visme

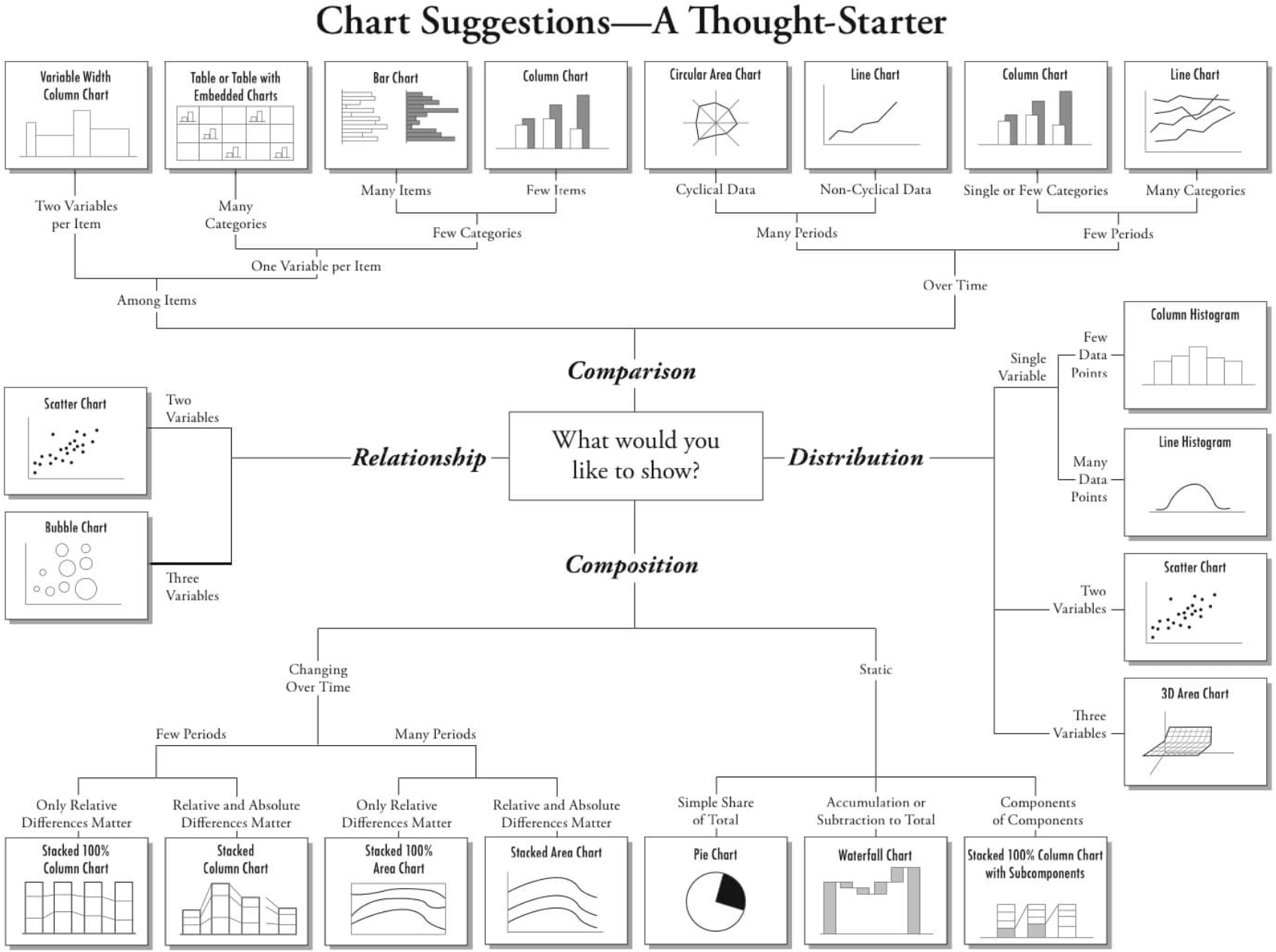
IBM Cognos Analytics

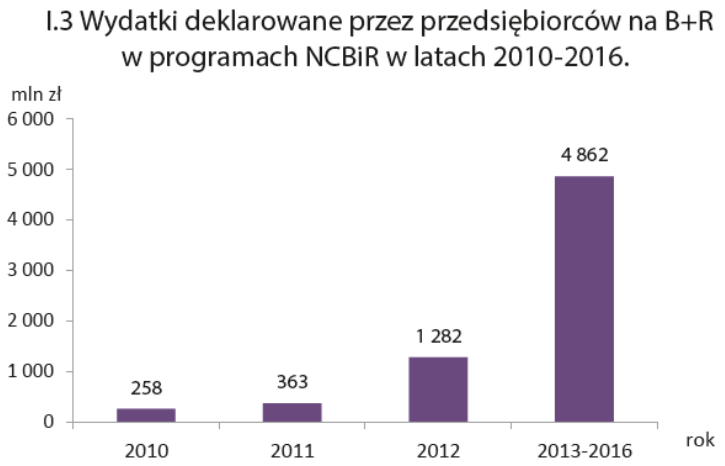
Infogram

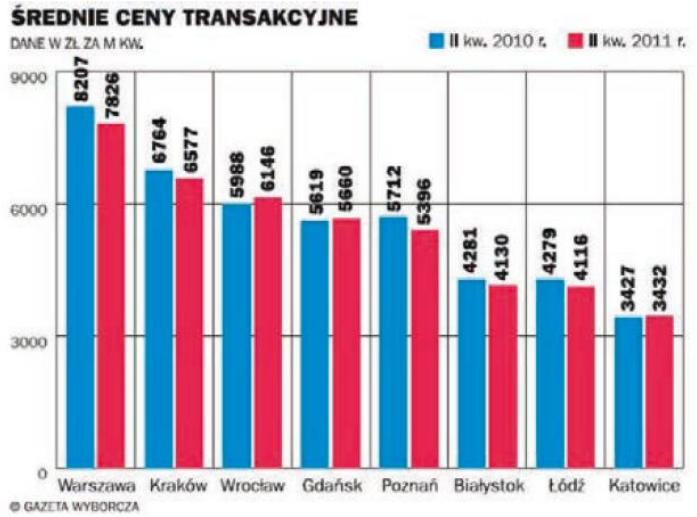
Google Charts

Excel 😂

Types of charts and their application

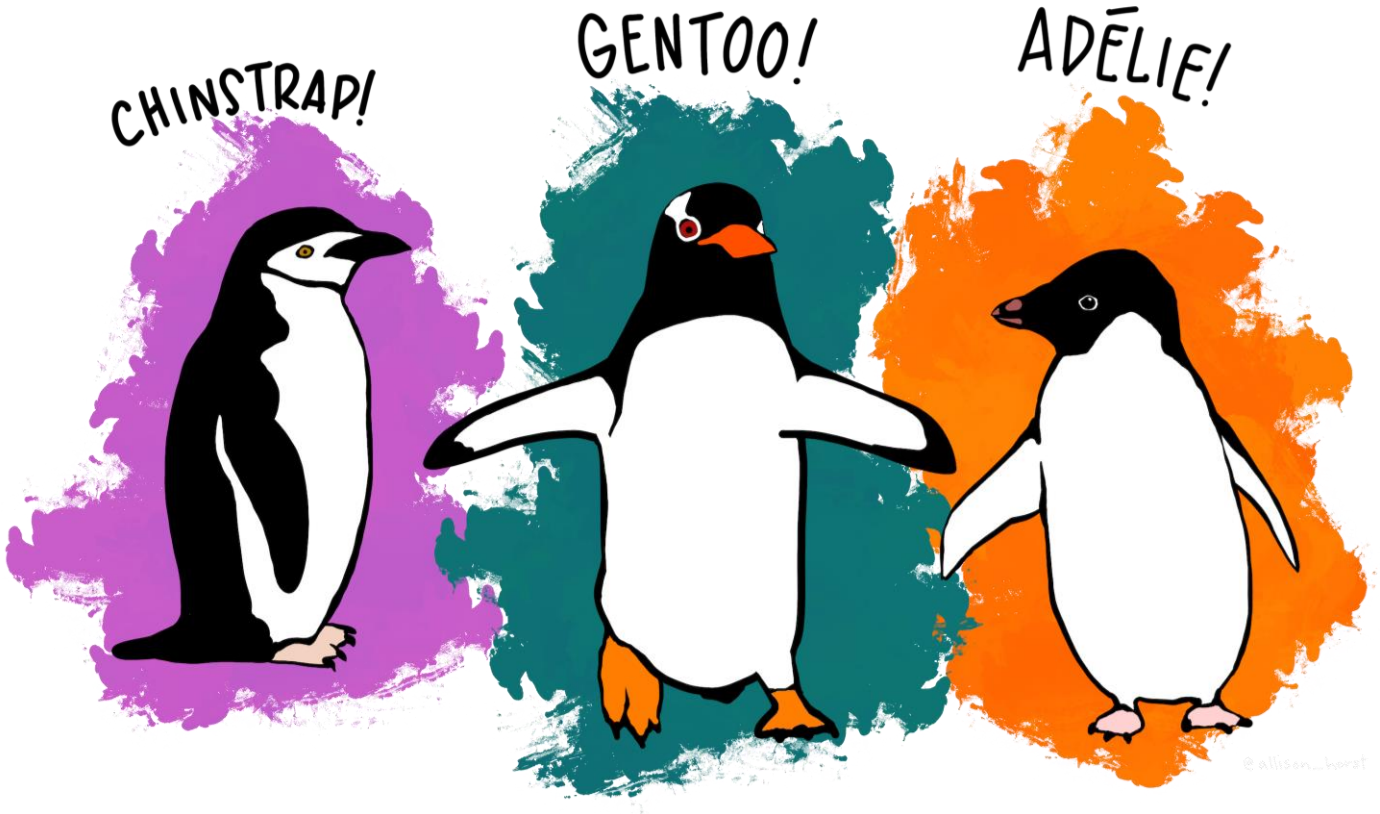




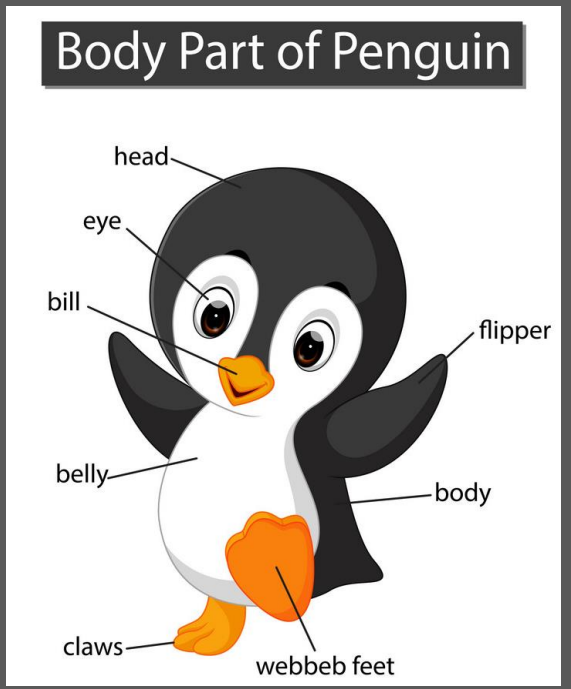


Examples of incorrect visualizations

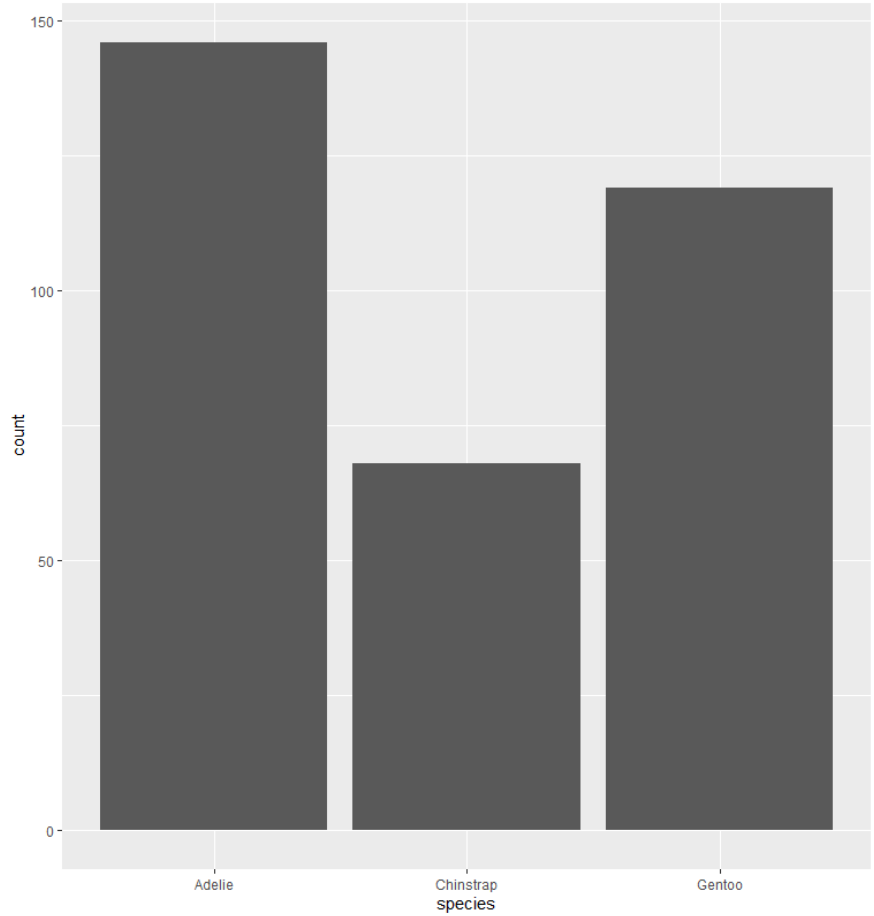
Quick examples of different stories on the same data

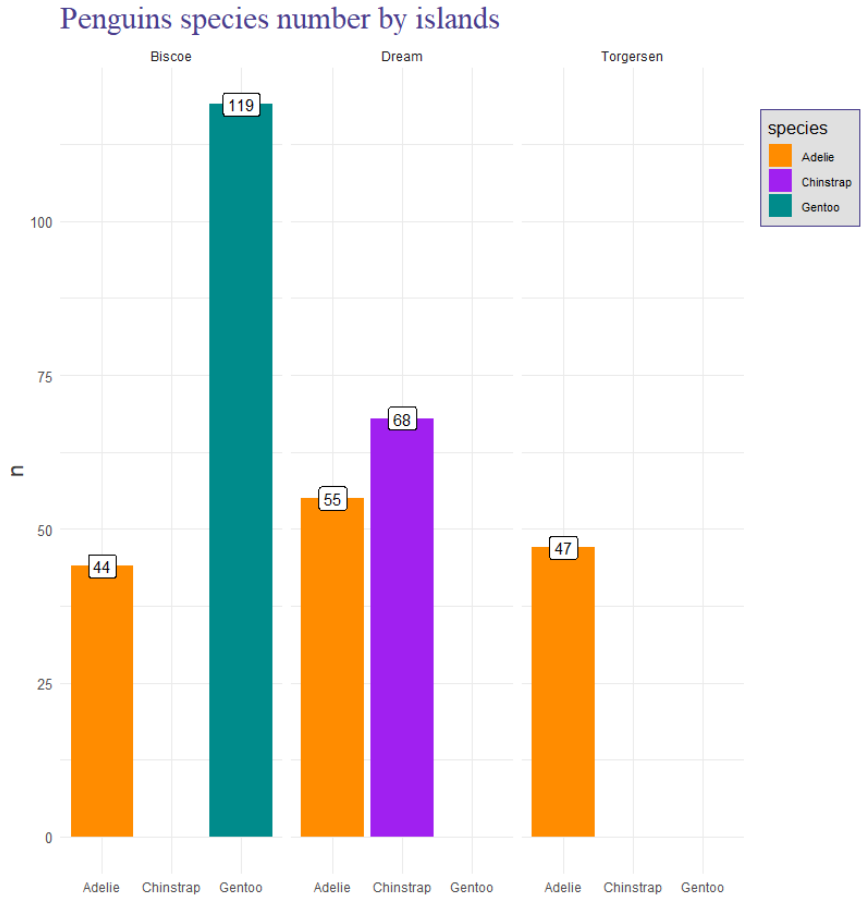


- Dataset used to create charts is called Palmer Archipelago (Antarctica) penguin data and it can be found on kaggle.com. The dataset contains 7 variables: sex, species, island, body weight, flipper length, culmen length and depth. Culmen is the other name for bill.

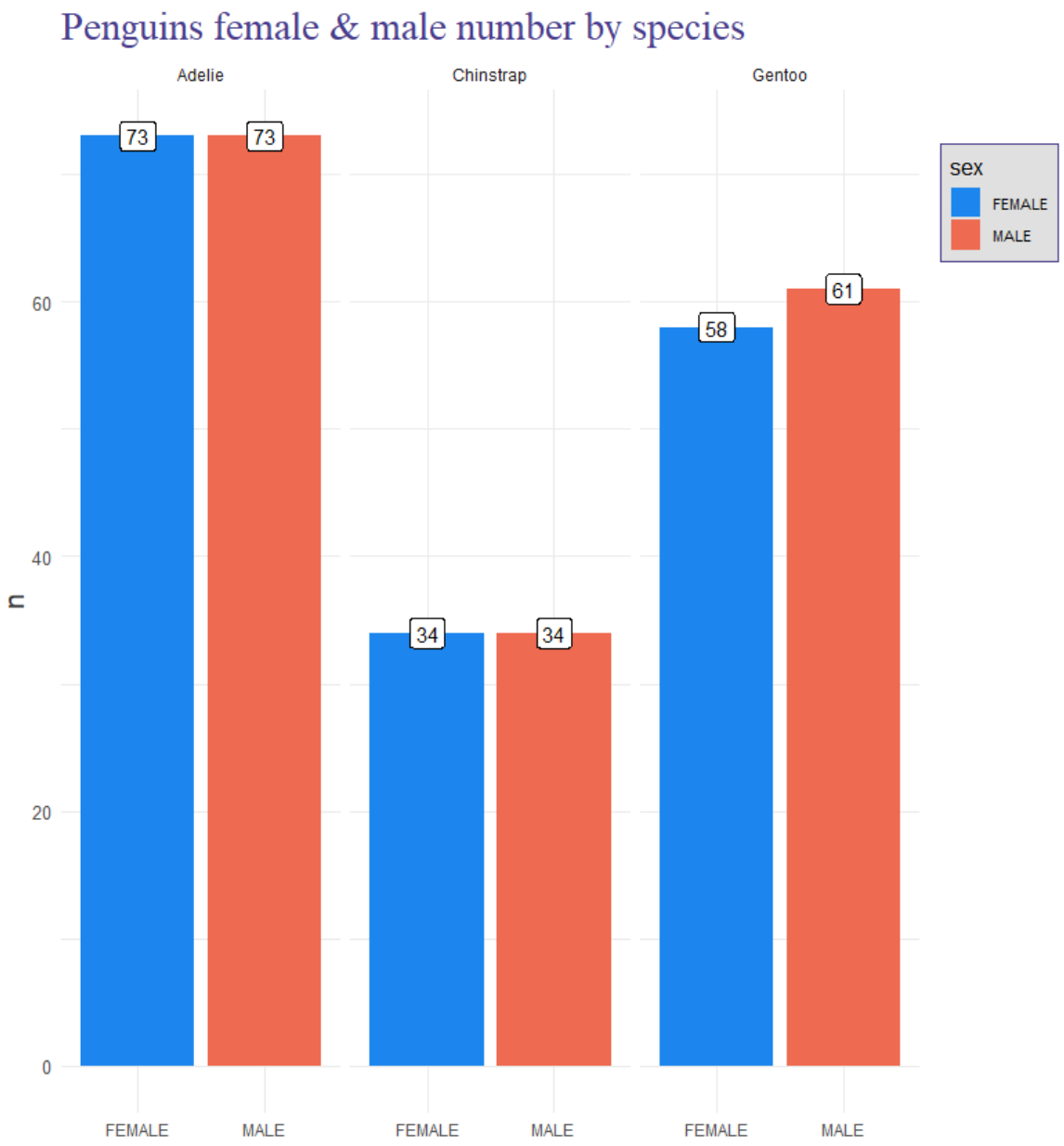


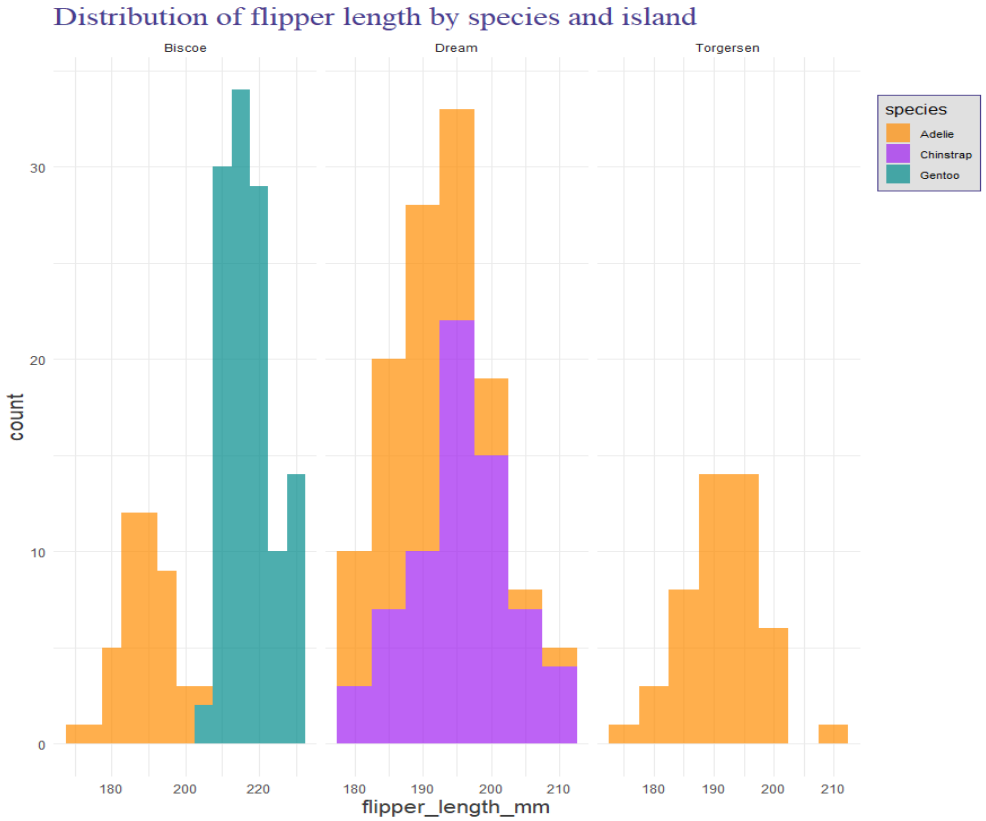
- Bar chart on the left shows only the number of penguins by species. Version on the right also includes split into different islands and is visually improved.

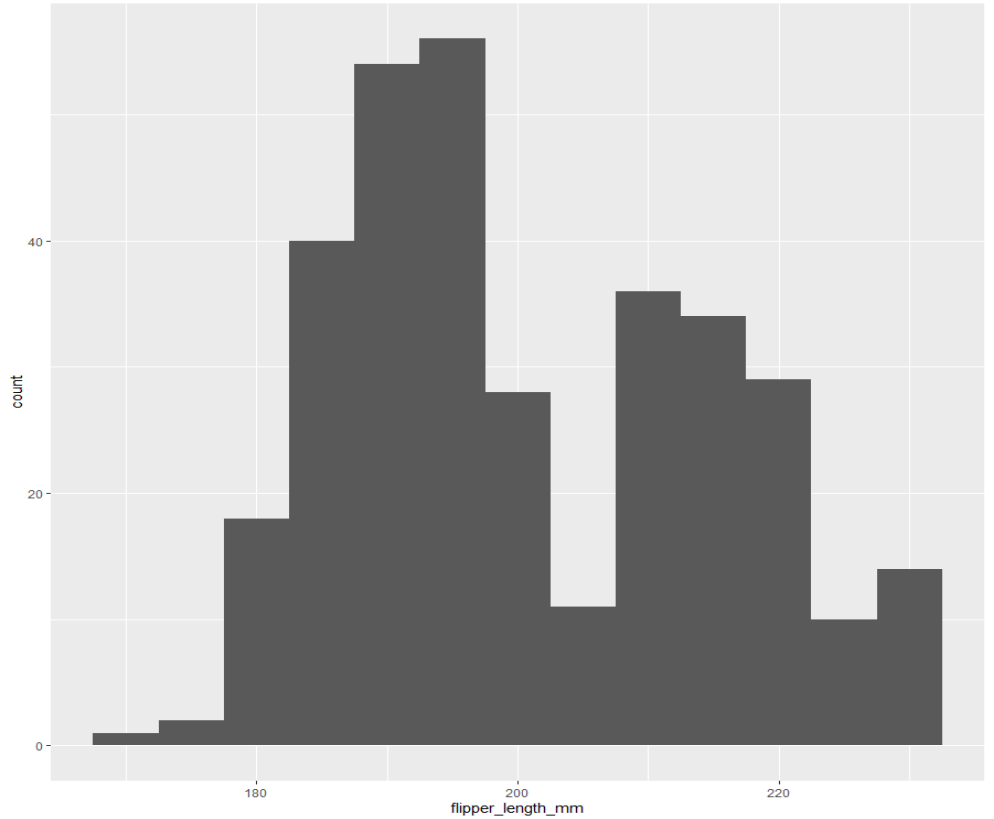




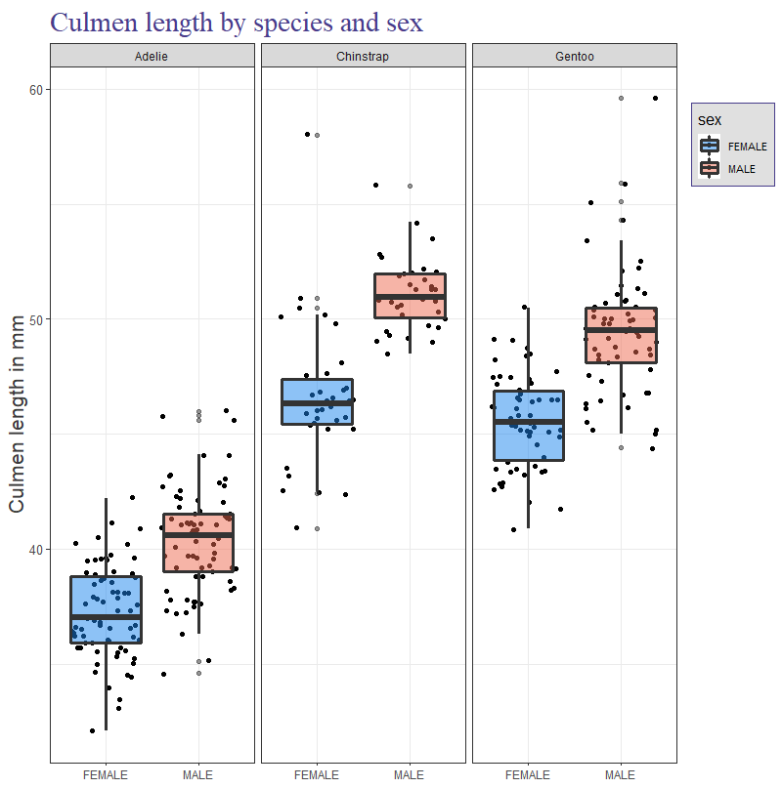
- Also, we can check whether there are differences in case of sex between species.

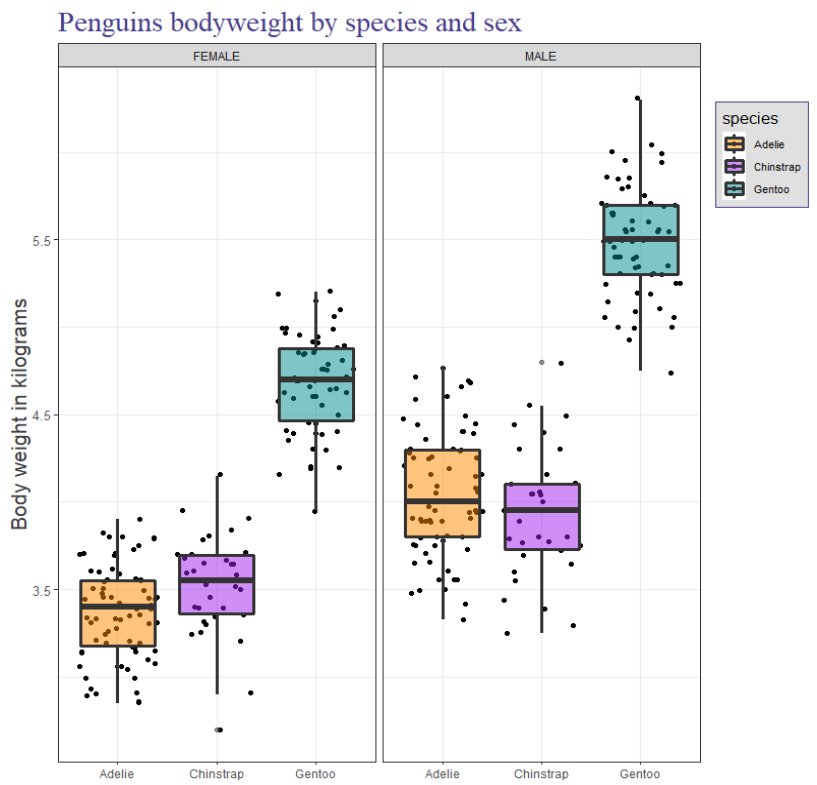


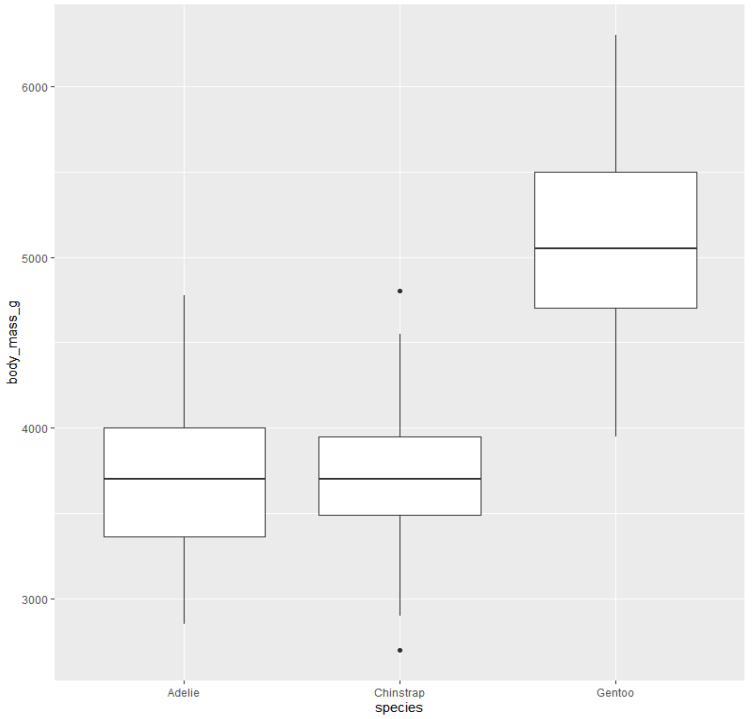




- These histograms show different aproach to distribution of flipper length. On the right we can see distribution for every penguin, but on the left it is separeted by species and island.

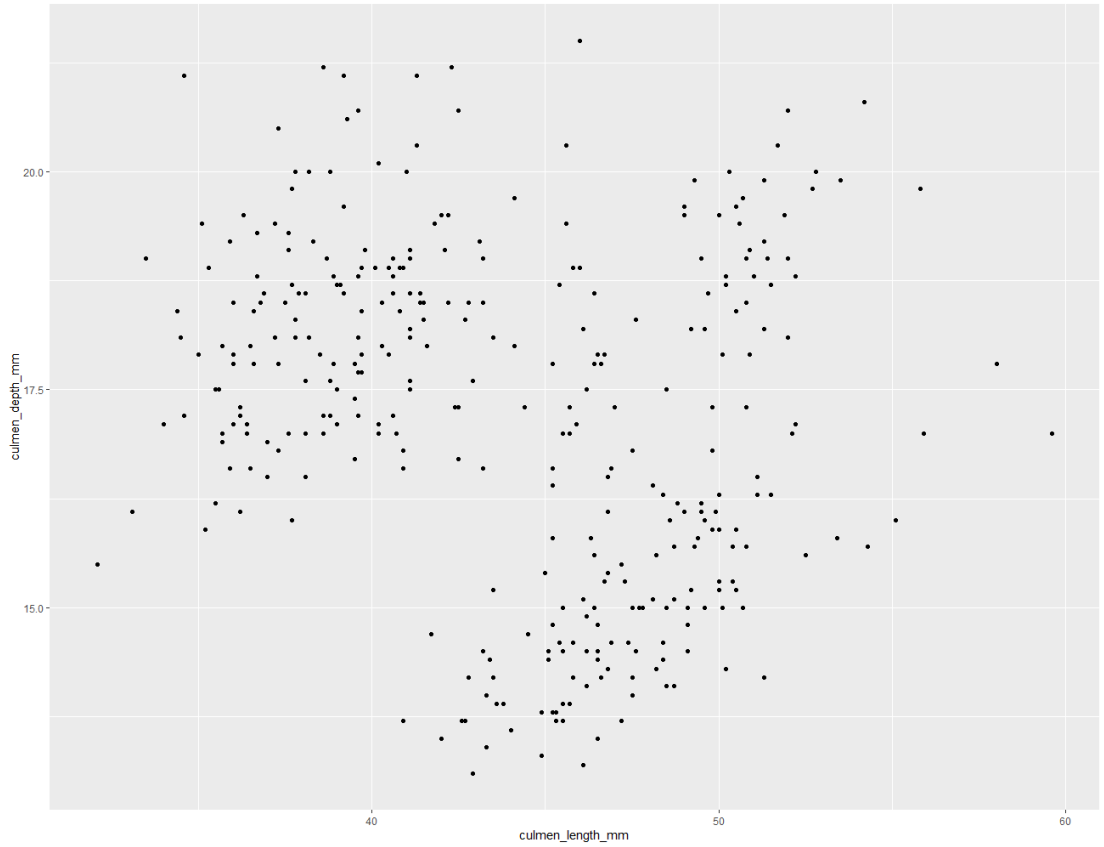


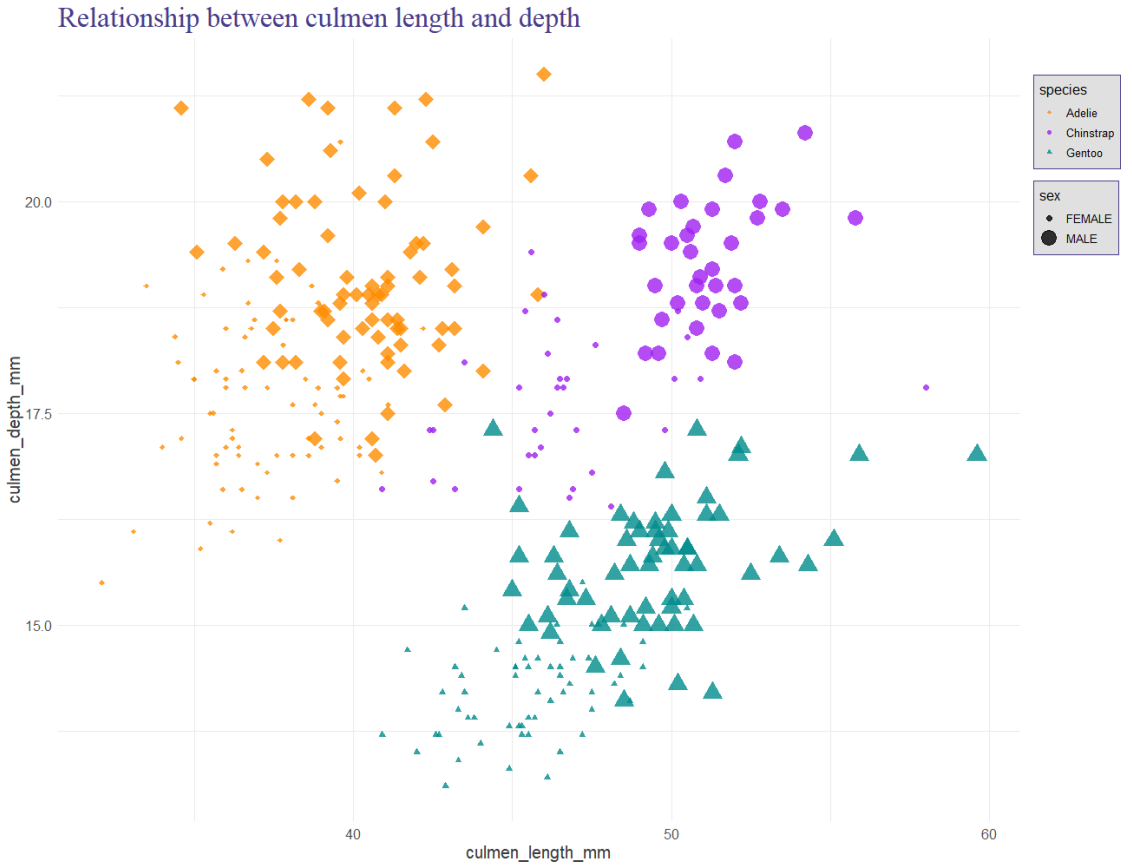




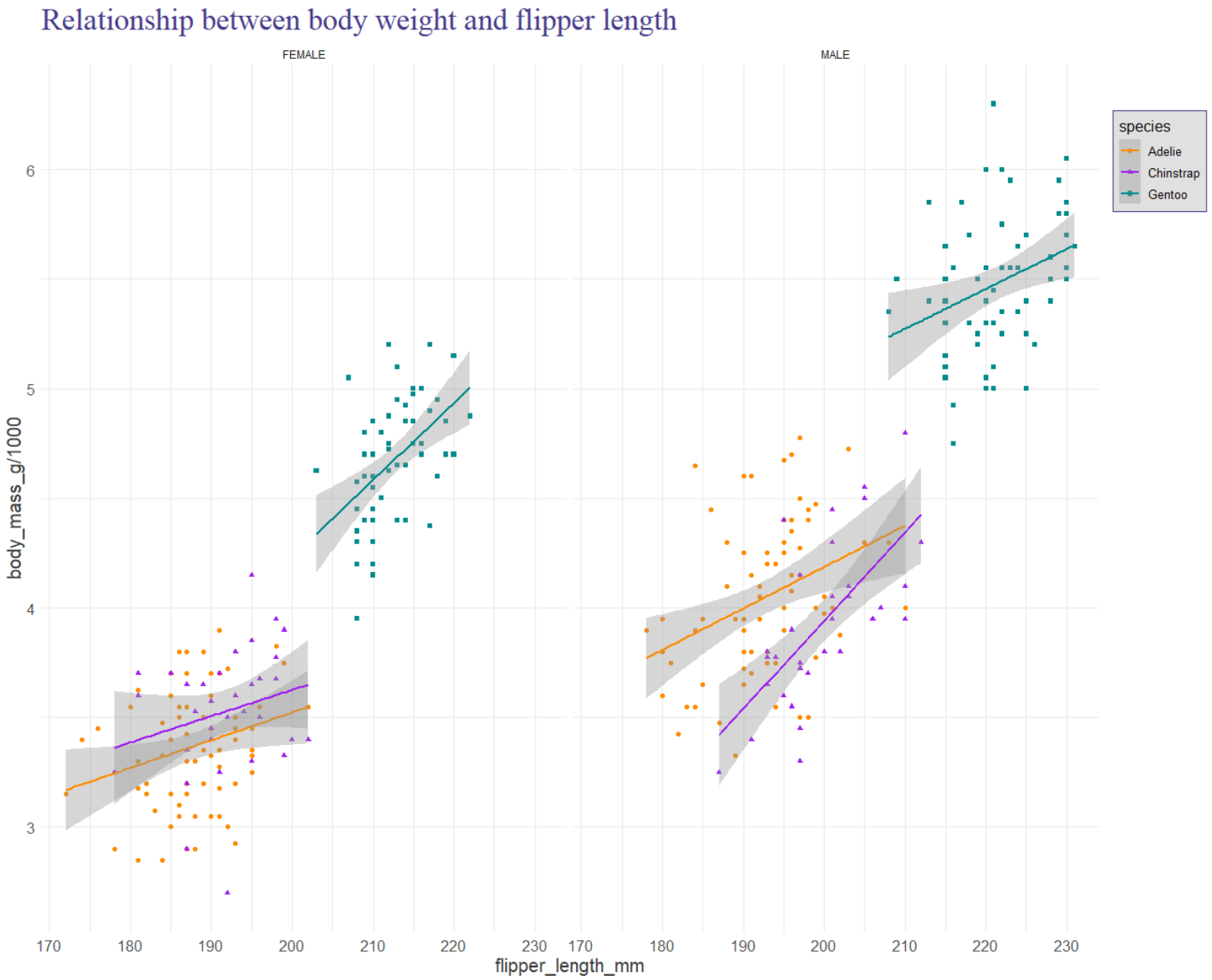
- These charts are called boxplots. They can demonstrate the median, spread and skewness of quantitative variables. Lines which are connected to boxplot are called whiskers and everything outside the whiskers can be classified as outliers. Boxplots on the left and in the center show data divided by species and sex. On the right, on the other hand, there is only presented body mass by species without any graphical improvement.

- Scatter plots are great to present relationships between two variables. In the left chart below there does not seem to be any relationship. The chart on the right, however, shows that there are a clear differences in culmen size for different species.





- The last chart, also a scatter plot, shows relationship between body weight and flipper length by gender. A line and gray area have been added to indicate the trend and confidence interval.



Glossary

Chart – diagram, wykres

Provide – dostarczać, zaopatrywać

Outlier – obserwacja odbiegająca / nietypowa

Pattern - wzór

Insight – wgląd, spostrzeżenie

Depiction - opis

Mutual - wzajemny

Flipper - płetwa

Culmen - dziób

Distribution – rozkład statystyczny

Boxplot – wykres skrzynkowy

Whiskers - wąsy

Skweness - skośnośc

Quantitative variable – zmienna ilościowa

Scatter – rozproszenie, rozrzucenie

Confidence interval – przedział ufności

Thank you for your attention and I invite you to take part in the quiz!