1 Task

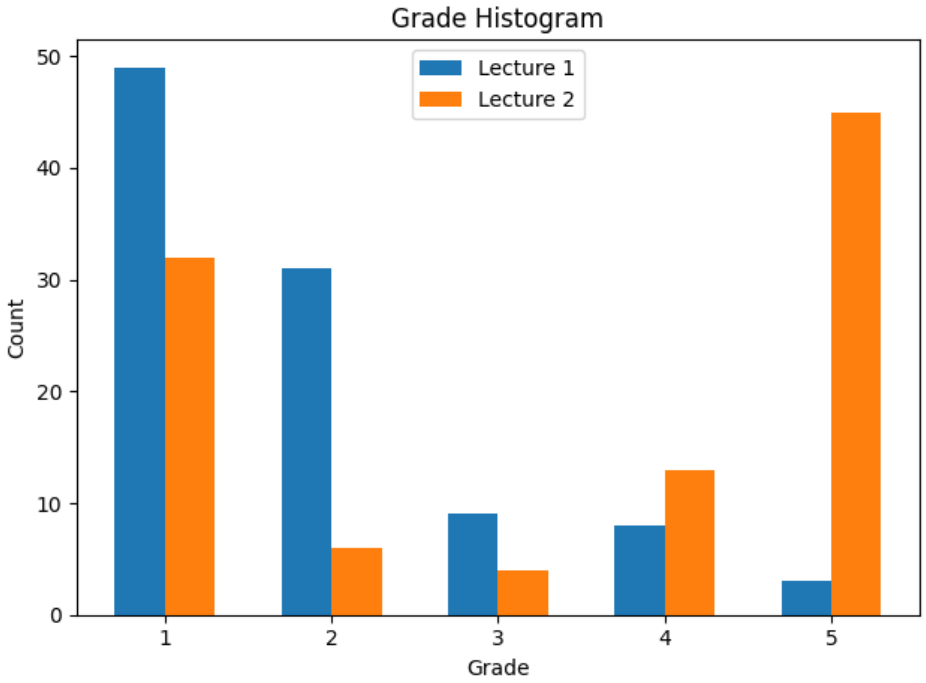

Figure 1: This histogram shows a descete number of students for each given grade in two exams.

Histogram in figure 1

- What are the characteristics of a histogram?

  This histogram shows how many instances exist for (in this case) each possible grade. It is also applicable for continuous values using bins, but the grades are descrete.

- Why do you think this plot was best to demonstrate the data?

  The histogram shows how many students have which grade for both exams. This opens the possibility for further manual analysis regarding the distribution of grades for each exam.

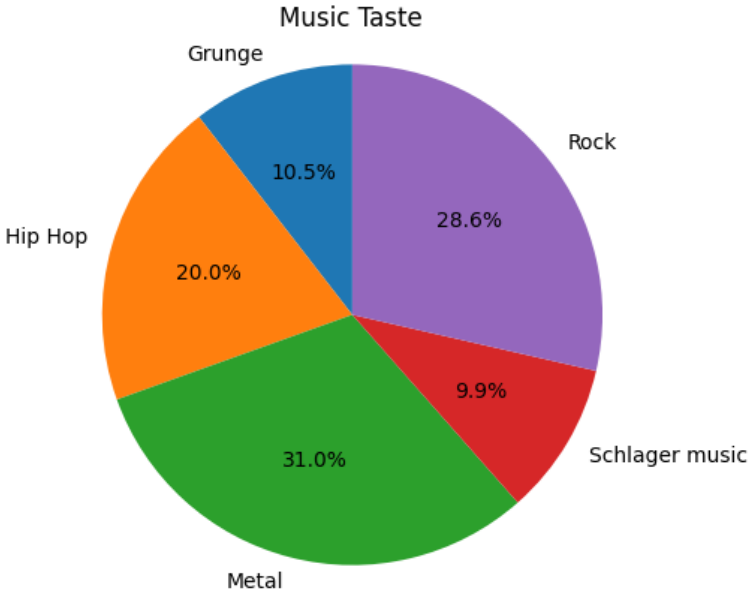

Figure 2: This pie chart shows the music taste among 1000 asked people.

Music Pie Chart in figure 2

- What are the characteristics of a pie chart?

  The pie chart shows the non overlapping classes as the partitions of a circle. It gives an good overview of how set is splited into these classes and how it adds up. It is not so practical for very big class size differences.

- Why do you think this plot was best to demonstrate the data?

  The given data was not overlapping and splitted the set of 1000 people into 5 classes which are good displayable on a pie chart. It easily visualizes the size of each class in contrast to the total size and the other classes.

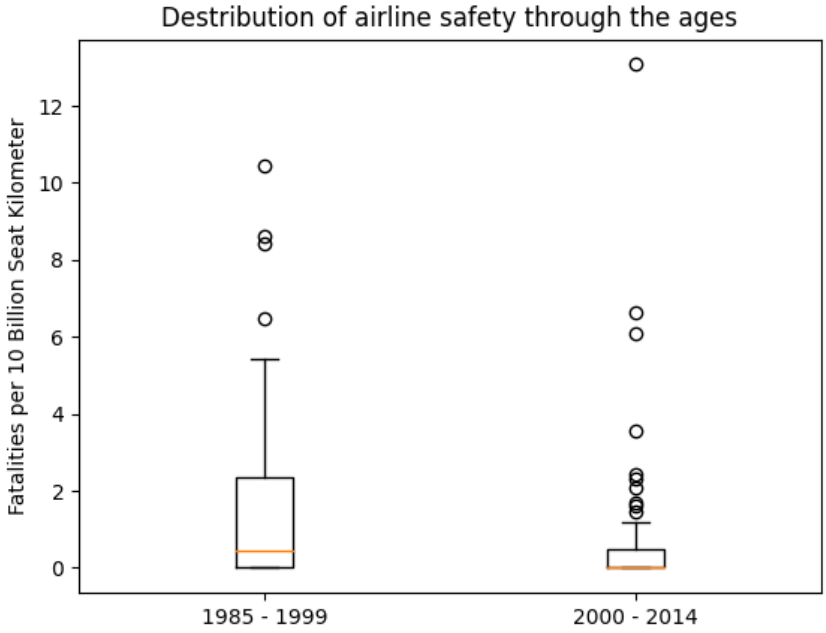

Figure 3: This box plot shows the number of fatalaties per 10 billion seat kilometers for 57 airlines in the time from 1985-1999 and 2000-2015.

Box plot in figure 3

- What are the characteristics of a box plot?

  The box plot shows the a box which represent the first and third quartile of the data, inside the box is a line which represents the median. The 'T' like lines on the top and bottom of the box are the minimum and maximum values without outliers and the little circles are the outliers. The box plot abstracts the data plot whithout hiding the distribution of the data like in a bar chart where only the mean is displayed.

- Why do you think this plot was best to demonstrate the data?

  In this case the fatalaties per (10 billion) seat kilometers is shown at two equal sized time intervals. First of all the fatalaty values is corrected by the flown seat kilometers of the airline. This enshures a fair rating for bigger/smaller airlines. This values are used in the box plot, where each data point represents one airline. The box plot shows the mean fatalaties for all airlines in the two intervals, which is good to see the overall improvement over time in the two plot parts. The plot also shows that not only the airlines got saver, there safety distribution is

also narrowed in contrast to the first interval. Showing the outliers displays the few critical airlines which could also be labeled and analyzed further more. All in all the plot summerizes the data from the 57 airlines with many data points in an smaller but informative plot.

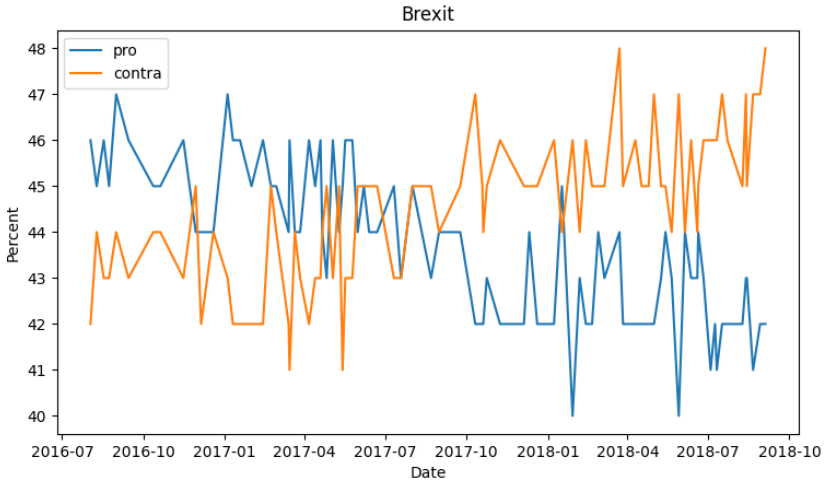

Figure 4: This plot shows the percentage of people who think Brexit was wether right or wrong over time.

Plot in figure 4

- What are the characteristics of this line plot?

  The plot shows the movement of the two values over time.

- Why do you think this plot was best to demonstrate the data?

  It nicely shows the pro and contra values throughout the time. Different events could be labeled in the plot.

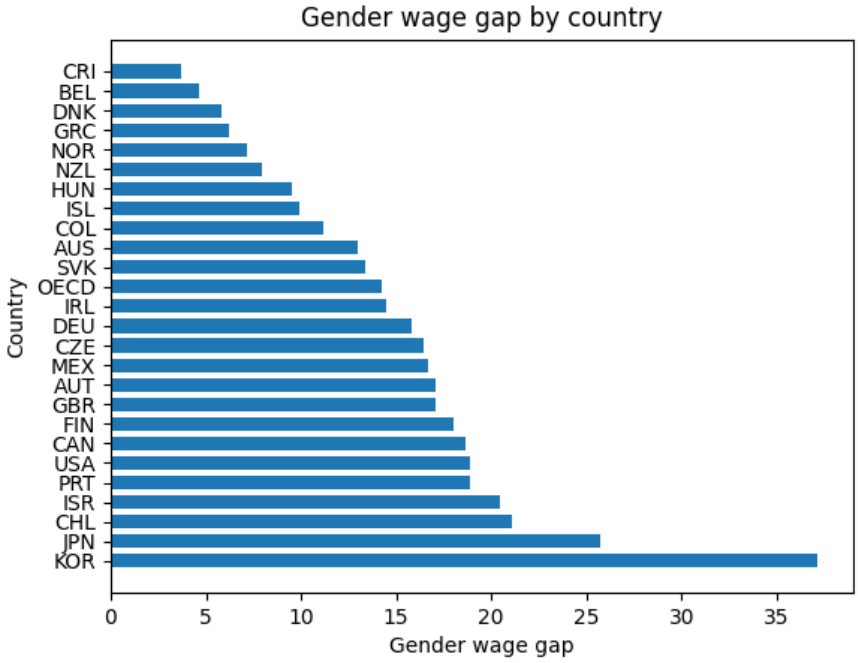

Figure 5: This bar chart shows the deviation of wages for female employees with respect to wages for male employees determined by oecd for a number of countries.

Bar plot in figure 5

- What are the characteristics of this line plot?

  It shows a value for each of the countries without furter information about e.g. the distribution.

- Why do you think this plot was best to demonstrate the data?

  It allows to compare the gab between the different countries, whitout beeing too big and confusing or losing e.g. the country data.

2 Task

- Plot 1

  The data is not corrected by the amount of costline the states have and also a log scale would be nice to differentiate the small values without haveing a big spike.

- Plot 2

  The amount of medals shown does not scale with the emount of medals earned. The USA won nearly twice as many medals as Russia but have only one sixth more in this plot.

- Plot 3

  The number of cases seems to go down in 2016 but the plot only changes the time interval on the right side while still showing the absolute value of cases in this shorter time interval.

3 Task

Notebook:

```
[[1.         0.81761114]
 [0.81761114 1.        ]]
<matplotlib.collections.PathCollection at 0x7fe2f2242630>
```

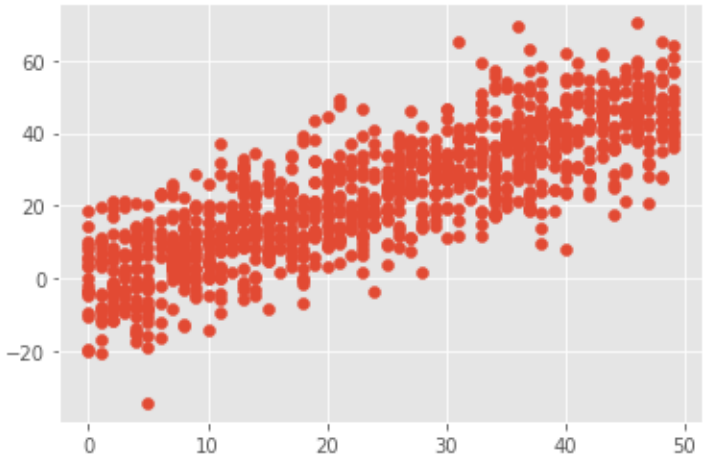

Figure 6: The first correlation matrix shows a correlation 1 between (x, x) and also (y, y) (obviously). More interesting is the correlation cov of 0.817 between x and y in both ways. It incicates that a bigger x value could imply a higher y value and the other way around.

```
[[ 1.         -0.94957116]
 [-0.94957116  1.        ]]
<matplotlib.collections.PathCollection at 0x7fe2f29873c8>
```

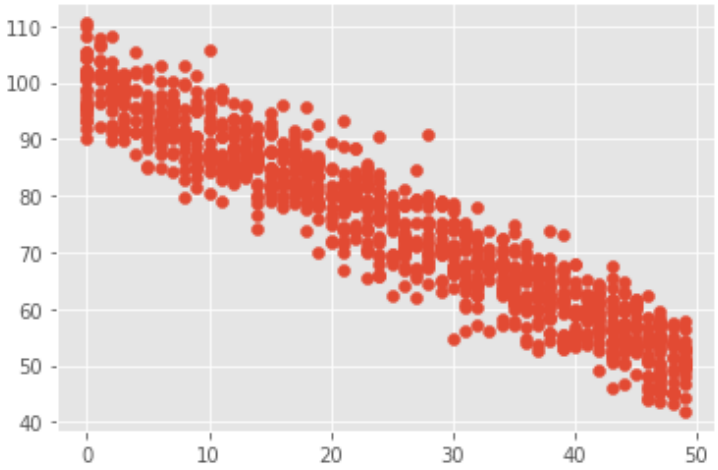

Figure 7: The first correlation matrix shows a correlation 1 between (x, x) and also (y, y) (obviously). More interesting is the correlation cov of -0.949 between x and y in both ways. This indicates a negative correlation, so that a smaller x value results in a larger y value.

```
[[1.         0.03774225]
 [0.03774225 1.        ]]
<matplotlib.collections.PathCollection at 0x7fe2f13b0b38>
```

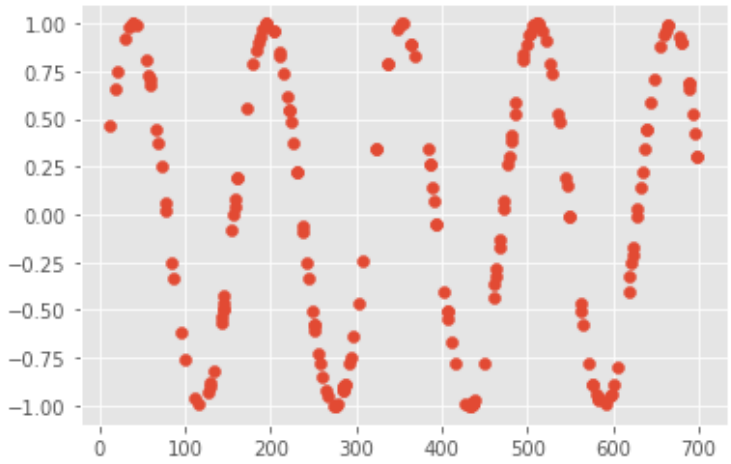

Figure 8: The first correlation matrix shows a correlation 1 between (x, x) and also (y, y) (obviously). The non linear correlation can not be determined.

Chi-Square-Test:

	Likes Zombie movies	Does not like Zombie movies	Total
Plays harp	24 (18)	6 (12)	30
Does ot play harp	8 (14)	16 (10)	24
Total	32	22	54

Table 1: Table with completed values

There is $(I-1)*(J-1) = 1*1 = 1$ degree of freedom.
The X^2-Value is: $X^2 = \frac{(24-17.7)^2}{17.7} + \frac{(6-12.2)^2}{12.2} + \frac{(8-14.2)^2}{14.2} + \frac{(16-9.7)^2}{9.7} = \frac{1323}{110} \approx 12$.
For an $\alpha = 0.0595\%$ of the values are below the threshold of $3,841$. Because 12 is greater than $3,841$, the null hypothesis gets rejected and the data is propably corrected.

4 Task

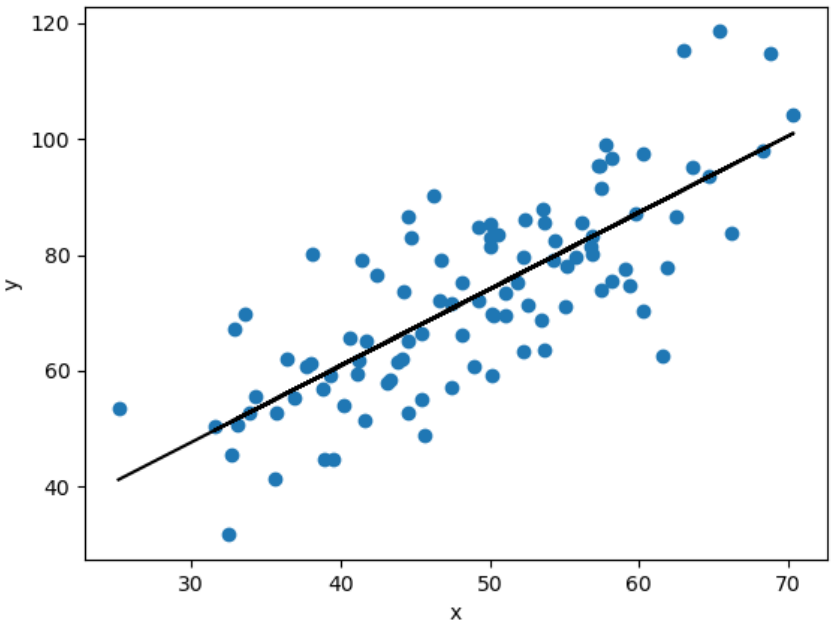

Figure 9: Line fitted in the data using the given regression method.

- Why is a regression also referred to as ordinary least square?

  Because the squared error beween the fitted line and the data points is minimal.

- The regression results as seen in figure 9 are:

  $\beta_1 = 1.322431022755357$
  $\beta_0 = 7.991020982270527.$

  In the model β_1 is the line slope of the fitted line and β_0 is the y-interception.

- For which data distribution a regression would not be a good fit?

  This given linear regression is not good for data that follows higher polinomial functions, exponential function or sinus like functions.

5 Code

Github:
Plot Code
Regression Code