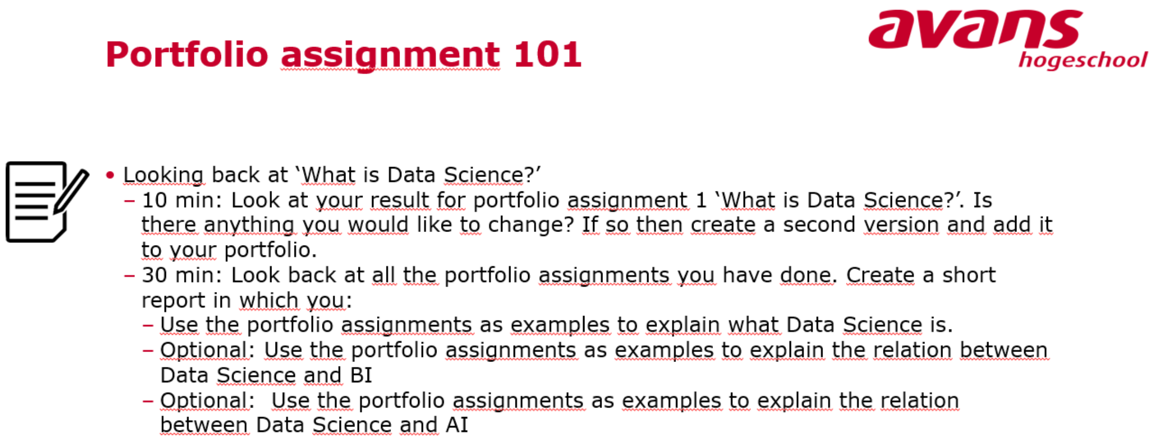


Portfolio assignment 101

- Looking back at 'What is Data Science?'
  - 10 min: Look at your result for portfolio assignment 1 'What is Data Science?'. Is there anything you would like to change? If so then create a second version and add it to your portfolio.
  - 30 min: Look back at all the portfolio assignments you have done. Create a short report in which you:
    - Use the portfolio assignments as examples to explain what Data Science is.
    - Optional: Use the portfolio assignments as examples to explain the relation between Data Science and BI
    - Optional: Use the portfolio assignments as examples to explain the relation between Data Science and AI

Data science is a method of transforming data into valuable business strategies. It is a tool for making sound decisions to increase profits, improve the customer journey and spot opportunities.

For example, we had to compare two completely different rows of values to determine if there is a correlation between the two. This is something you can't always spot when looking at data yourself, especially when there's a huge amount of it.

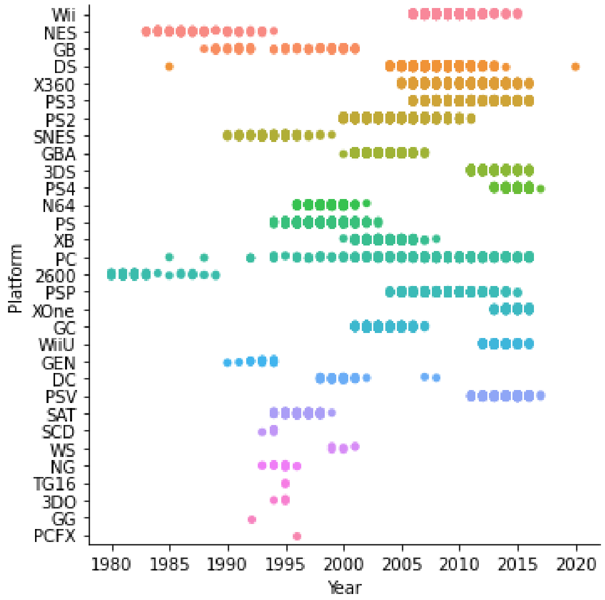
Something that really explained data science for me was machine learning. A very basic example of this is Decision Trees. You give the tree set a lot of data, and the tree set determines what condition it is gong to use to split up data.

Portfolio assignment 102

- 30 min: Look back at all the portfolio assignments you have done. Create a short report in which you use the portfolio assignments as examples to explain the tasks and process of a Data Scientist.

Visualizing / Modelling data

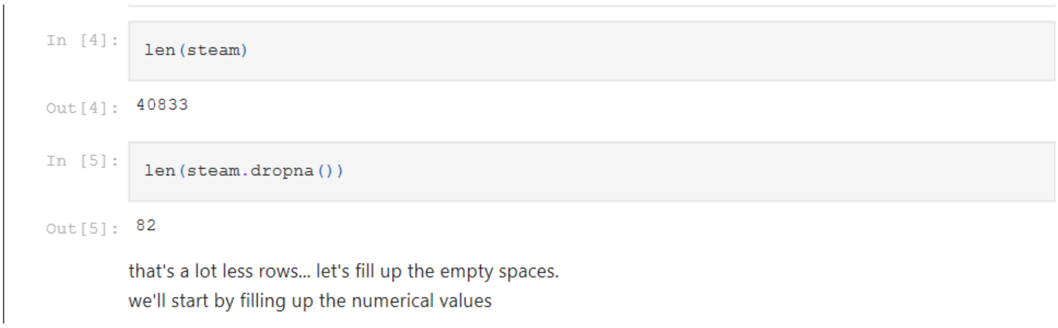
One of the necessary things in Data Science is visualizing data for the user. This way you can present the data you've viewed.



I was fond of this graph; it shows a categorical plot between the different consoles and the years.

Missing data

Something small, but important thing that came up was NaN data, these were missing fields in data sets. You'd think this is easily solvable by dropping rows with this data.



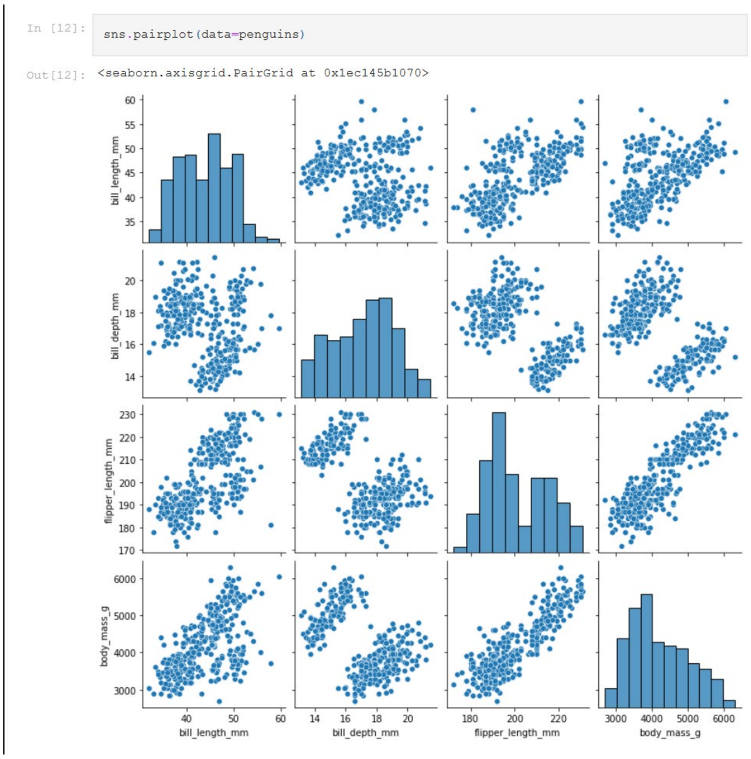
```
In [4]:  len(steam)

Out[4]:  40833

In [5]:  len(steam.dropna())

Out[5]:  82
```

that's a lot less rows... let's fill up the empty spaces.
we'll start by filling up the numerical values

But this doesn't work when you have one column that barely gets filled. You'll have to either drop specific rows or fill up the missing data. This is different in every data set, but I replaced NaN values with 0 in my case.

Clustering data

Clustering data is something we mostly did at the end of these assessments.

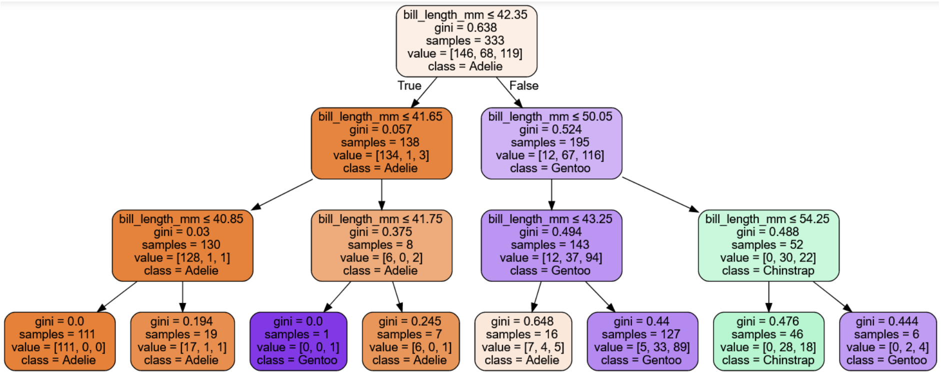


```
In [12]: sns.pairplot(data=penguins)

Out[12]: <seaborn.axisgrid.PairGrid at 0x1ec145b1070>
```

We did this by using seaborn's pairplot(). This function compares all numerical data and visualizes it for the user. You can easily see a division between a few comparisons.

Predicting data

Another important thing is to predict the outcome of new data. Predicting data uses the idea of clustering data to be able to make assumptions in the future. You generally want a training set that can accurately predict what new data is going to be. This way a company can make general assumptions.



With this treeset, we can pipe through new data and see which class they would end up on.