Fun with Data!

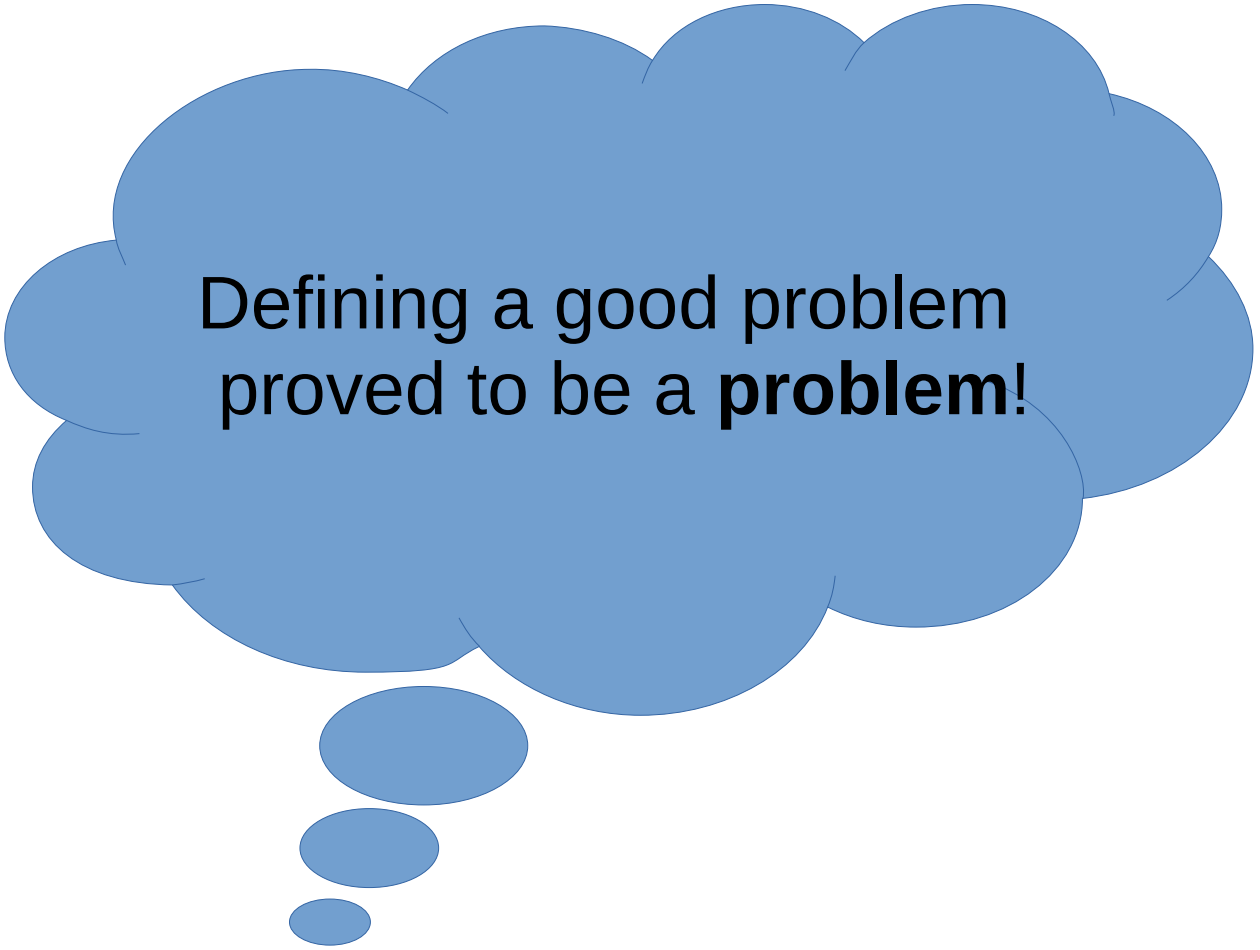
Kalina Spasova

Rebecca Tabad

Anna Wykes

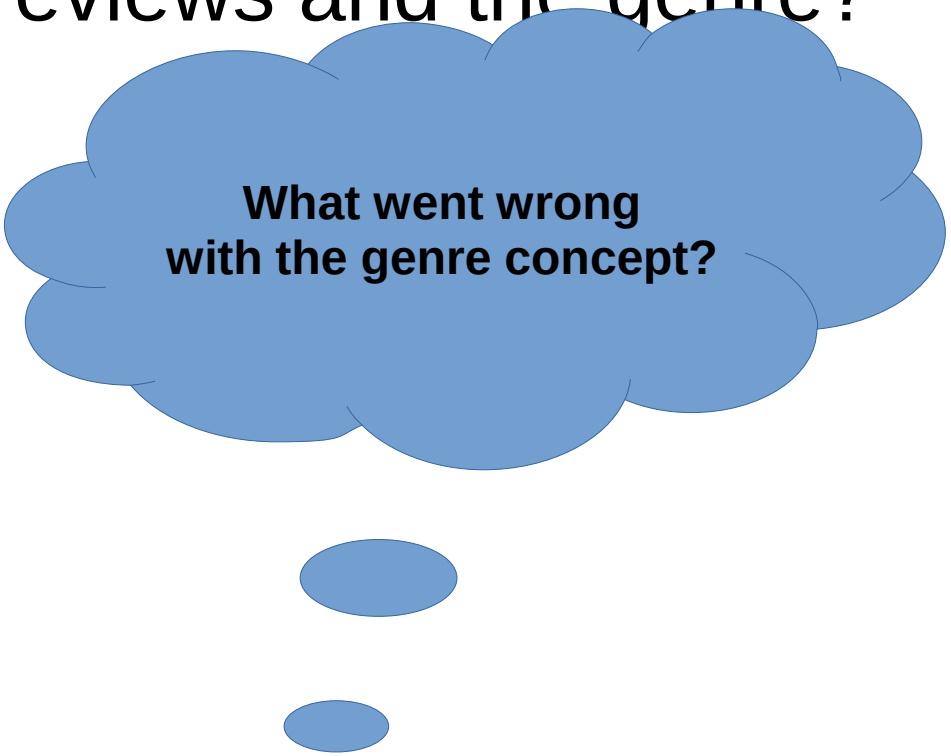
Where were you before this workshop?

Problems we were trying to solve



Problems we were trying to solve

- Check if there is any correlation between number of critic voters for reviews and the genre?
- Check if there is any correlation between number of user voters for reviews and the genre?



Problems we were trying to solve

- Predicting the IMDB score based on a number of features
    - gross (exchange rate not taken into account)
    - duration
    - title_year
    - movie_facebook_likes
    - imdb_score
    - content_rating

What did we learn?

- A bit of Python
  - Seaborn and Boxplot for visualization
  - Scikit-learn for Machine learning
  - Jupiter Notebook is great for throwing something together! Instant gratification or confusion!

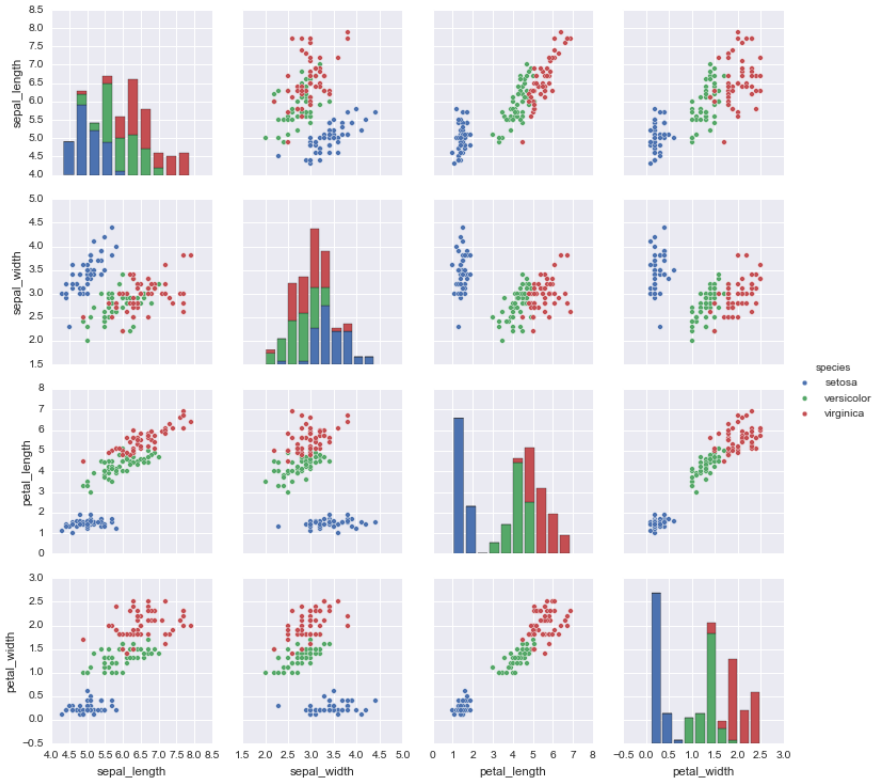
- There are a lot of algorithms out there to use for machine learning
  - Weighted KNN

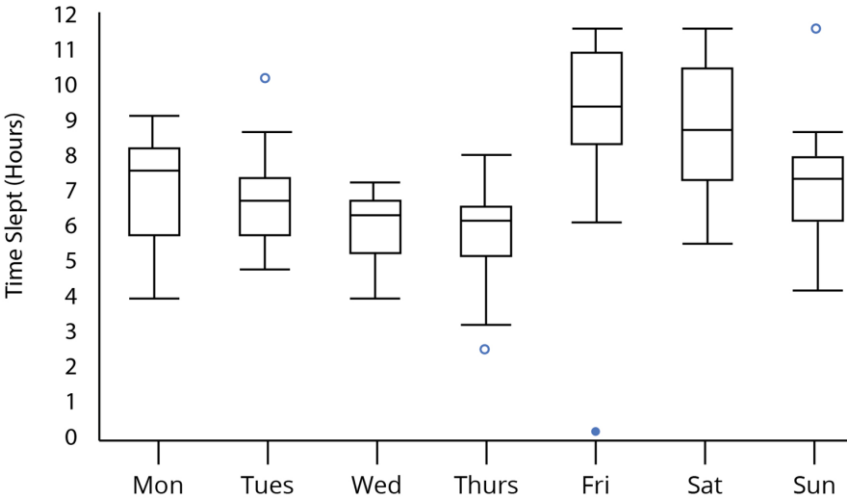
What did we learn?

- There is a lot of data out there
- And even more problems to be solved
- Defining a meaningful problem can be tricky if you do not have experience

What did we learn?

- Data visualization is a great way of getting to know your data and spotting patterns
- Box plot and Scatterplot matrix with the Iris Data





What did we learn?

- Priming your data is just as important as knowing it
    - Exchange rate
    - Integers vs floats
    - Strings
    - NaNs
- Priming it too much can skew your results

What did we learn?

- The concept of training your model and splitting your data so that you can test for accuracy

What next?