Data deep dive

Chinedu and Emre

What do we know?

- We have been given two dataset, giving information about multiple samples of wines(red and white)
- The categories include: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality
- The only subjective category is the quality
- Quality was determined using the median value of a score (1:10) given by 3 wine experts
- The dependent variable is quality, and the rest are independent variables

What do we want to know?

- The reasons for the subjective scores
- Potential connections between any of the variables, in relation to the quality scores given by the wine tasters
- Understanding which of these variables aid in a good tasting wine
- And vice versa for a bad tasting wine
- Understanding the bias of wine tasters for certain qualities either positive or negative

Our Regression building process

- Finding where single variables that have a strong linear relationship with the quality variable
- Finding where multiple variables combine to result in a high quality score
- Keep a lookout for potential outliers that can skew the data
- Ensure regression assumptions are all true
- Making sure the chosen model is properly fitted to the given dataset

Model Types

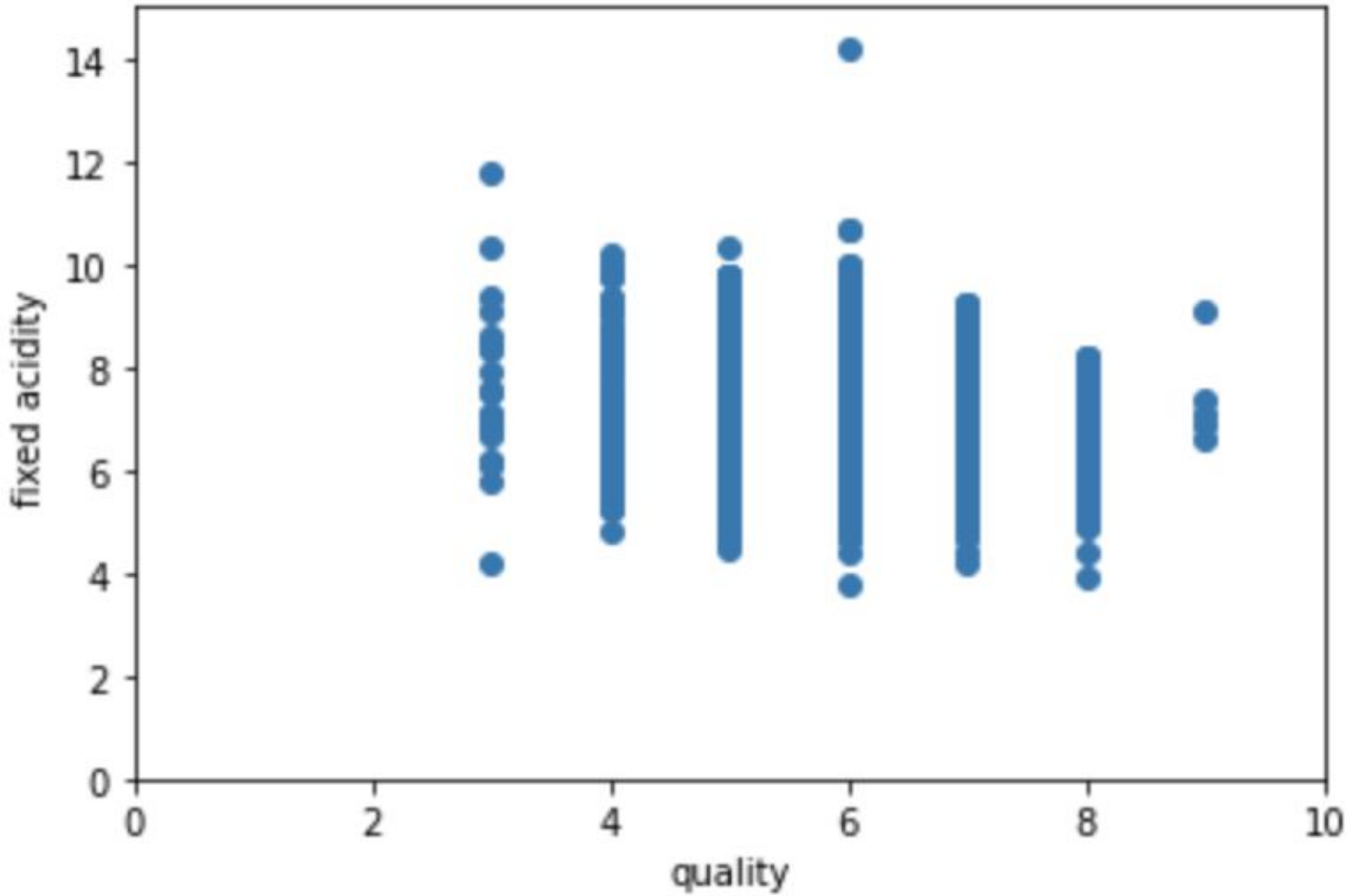
- Single variable linear regression model
- Multivariable linear regression model

Ethical considerations

- Respect for human autonomy
  - While this model may be a predictor of future wine ratings, this type of data will always ultimately be subjective
- Transparency and Explainability
  - The model should be open for improvement after production and should be documented thoroughly so others can take over

Some Data trends

- White:
  - Negative correlation with Quality: Total Sulfur Dioxide, Fixed Acidity, Density
  - Normally Distributed: pH, Chloride, Sugar, Citric Acid
  - Positive: Alcohol
- Red:
  - Negative:
  - Normally Distributed: pH, Chloride, Sugar, Density, Total Sulfur Dioxide, Fixed Acidity
  - Positive: Alcohol, Citric Acid


fixed acidity
quality
0
2
4
6
8
10
12
14
0
2
4
6
8
10