Exam—UK gender pay gap

1. Loading, formatting and scaling data

- Loading Data
- Standardise
- Scaling data
- Normalization
- Format dataframe

2. Making new features (combining / aggregating variables)

- Aggregating variables

  Aggregation is the process by which a GroupBy object efficiently slices the data as the groupby method groups it, uses the specified operations on each slice, and assembles the results into a final result.

- Multi-column aggregation operations for selected GroupBy objects

3. Summarising variables (means, distributions, data types)

- df.describe().T
- Means---use two methods:

  The mean of male bottom quartile % + The mean of female bottom quartile %= 100%
- Notable differences
  The top quartile betweent male and female has notable difference.
- Distribution of Organisation size
- Distributions

  a. Histogram

  b. Probability densities that all sum to 1

  c. Gaussian distribution

  d. Exponential distribution

  e. Skewed normal distribution

  f. Lognormal distribution

  The values of Male bottom quartile % are basically concentrated in the range of 20-50

  As we can see from the picture, a lognormal distribution fits these data best and peaks at around 40.

- Data types: two methods
  a. arr1.dtype
  b. df.info()

4. Plotting data

- Plot the data of "Male bottom quartile %" and "Female bottom quartile %".

  There is a negative correlation between Male bottom quartile and Female bottom quartile.
- PCA and calculate BIC scores---k and cov_type
- Cluster and Plot---There is a clear group.

5. Segregating/ Filtering data

Use df.drop() delete some columns which aren't important for this dataset.

6. Investigating the relationships between variables

- Correlation--- The r value is close to -1, indicating a good negative correlation between Male bottom quartile and Female bottom quartile.
- Making models
  a. .stats.linregress()---returns us a value for the slope (a) and intercept (b)
  b. Female bottom quartile %= -Male bottom quartile %+100
  c. R^2=0.9999999999999996----a perfect model
  d. p-value=0.0---A p-value of less than 0.05 allows us the reject the null hypothesis and conclude there is a relationship between the two variables

7. Build predictive models

- Multiple Regression--- Male upper middle quartile % = 0.83794144Male lower middle quartile % + 0.08805007Male bottom quartile %
- Predicting New Values, .predict()---if Male lower middle quartile % is 24, Male bottom quartile % is 26, I expect that Male upper middle quartile % is 30.68.

8. Investigating the relationships between variables

- general_info.corr()---We can see here that the bottom, lower middle, upper middle, top quartile of men have strong negative relationship with women's. Male bonus has strong negative relationship with women bonus.

There is a great negative relationship between Male bottom quartile and Female bottom quartile, Male lower middle quartile and Female lower middle quartile, Male upper middle quartile and Female upper middle quartile, Male top quartile and Female top quartile. There is a great relationship between Mean wage difference and Median wage difference (0.75), Male bonus and Female bonus (0.94).

9. Write some critical reflections on what data is and isn't being recorded

10. Now lets fit our multiple linear regression model to the data

11. Create and printing a dictionary that shows us the correspondences between ingredients and their indexes in the dataframe used for the regression model

12. See the male lower middle quartile of different organisation size

- df.groupby('Organisation size')['Male lower middle quartile %'].mean()
- plot
- See the organisation size which has most and less male lower middle quartile
- See the organisation size which has most and less female lower middle quartile

13. Identify any trends based on geographic location

I couldn't distinguish between London and non-London areas to determine trends.