Lab Sheet 2

This labsheet introduces using the `pandas` library for data exploration, `numpy` for transforming data and `matplotlib` for simple plots. These three are the most popular Python libraries for descriptive statistics in data science.

Task 1: Import libraries and read data

1. Import `pandas`, `numpy` and `matplotlib`.
2. Read `motor.csv` into a `DataFrame`.
3. Print a summary of the motor insurance and fraud claim data.

Task 2: Formatting, Cleaning and Filtering DataFrames

Often when dealing with a large number of features it is nice to see the first row, or the names of all the columns, using the `columns` property and `head(nRows)` function. However if we are interested in the types of values for a categorical such as the `modelLine`, we can access the column using the square bracket syntax and use `.unique()` to inspect the options.

1. Inspect the columns and print the first five rows.
2. Check categorical variables for errors and all variables for NaNs.
3. Remove NaNs if needed.
4. Merge categories with different spellings, if needed.

Task 3: Using Group-by and Merge

Group-by can be used to build groups of rows based on a specific feature in your dataset eg. the `Injury Type` categorical column. We can then perform an operation suc as `mean`, `min`, `max`, or `std` on the individual groups to help describe the sample data.

1. Group-by `Injury Type` and print `mean` and `count`.
2. Show a table of `income` and `availabilty` based on `Injury Type`.

Task 4: Visualising the data

Use the `seaborn` library to create the following graphs: 1. Create a `FacetGrid` of `Income of Policy Holder` and `Claim Amount` broken down to each of the three different `Marital Status`. 2. Create a `PairGrid` for the whole dataset. 3. Create a `Correlation Heatmap` for the whole dataset.