Methodolody

Data Planning

1. I started by first identifying all datatypes and looking into all data types if they were relevant to the particular column (data field). Of particular interest was “CustomerID”, which came in as pandas dtype “float64” but should be “int64”, thus I changed this data type.
2. In terms of data processing, data visualisation and other statistical methods I have a full personalised module called “data_analysis_functions.py” where we have barplotting, histogram plotting, scatterplotting and pairplotting as well as aggregation methods.
3. Following this, I loaded the data.

Analysing the Data

- First the key iis to print some summary statistics, which showed that there were negative values for the Quantities (returns), thus, if these returns happened in means it took away from some the sales. Thus, to capture the effective sale quantity, I subtracted adjacent sales and returns quantities and concatenated to the original datasheet.
- I then masked the datasheet for only positive values even after summing the returned values.
- EDA:
    - Printed the means, medians and modes of the Quantity and UnitPrice fields.
    - Printed the standard deviation and skew coefficients of the Quantity fields.
    - Plotted a pairplot to see if there are any important data relationships, there were none whatsoever.
    - Plotted the histograms and box plots of the “Quantity” field, it was clear that there were severe outliers that affected the skew of the data and the mean to a smaller extent.
- Handing outliers:
    - Took the 90^{th} percentile occurrences in the “Quantity” field and used this to mask the datasheet and thus remove outliers.

Conclusions/Inferences

- We clearly see that the United Kingdom has the highest sales volume

- We can also see that there are outliers in quantities, but these are mitigated for and removed, the negative quantities and negative unit prices, we see that some of these are discounts and some of them were returns of a product.
- We also see that there is little correlation amongst any of the data fields here.
- We also see that the month with the highest sales volume was November 2011.