Key Findings from Exploratory Data Analysis Task

First priority is to get a feel for the data and what it looks like.

- 'Age' seems likely to be capped at 90. It also appears to follow an unusual distribution (The 75th percentile is only 48)
- 'Background diversity' also appears long-tailed
- 'Qualification level ID' is obviously a categorical variable in numerical form, both because of its name and the low range of values. When processing it, it should be treated as a categorical variable.
- 'Investment profit' and 'Investment losses' both appear very sparse, with low means but huge variance. Profit also appears to be capped.
- Finally, hours worked per week looks to be quite realistic, centred around 40 hours. There is a maximum of 99, however, which might be an outlier.

All of these distributions will be looked at more closely later.

Next missing values were examined:

- At first look, it does not appear that there are any missing values. However, there are a few variables of note here, the Investment related ones and the Qualification Level related ones.
- Upon inspecting this, one can see that, generally, as 'Qualification_Level_ID' increases, so does 'Qualification_Level' become higher. This means that the IDs are not just random IDs and they could possibly be used to predict whether the salary is higher or lower. After all, one would expect someone with a 'Doctorate' to earn more than someone with just 'A-Levels'.
- Since these columns contain the same data in a different form, one of them can be removed.
- There is a relationship the whole way, from '0' being 'Un-qualified' to '16' being 'Doctorate'. As mentioned before, one would expect the salaries to increase as the qualifications increase.

Note: if one wanted to be extremely thorough, a detailed analysis on the distribution of qualification could be written up here, as well as an analysis of its effects on salary (simple correlation).

In order to check if there is ever a case when both of the investment variables are non-zero:

This means that whenever there is a non-zero value in one of the two columns, the other is always 0. Therefore, they can be safely combined into one variable, thus achieving multiple goals:

1.Reduced over-representation of investment

2.Reduced overall sparseness by combining two sparse variables into one.

The most obvious way of combining the two is by a simple subtraction. This will maintain all of the information in the two variables.

In the numerical variables, it can be said with certainty that there are no missing or unexpected values, because there are no nulls (from info) and being of a numeric data type, there are only numbers.

In the categorical variables, however, that is not a certainty.

- It appears that the number of '?' in Workclass and Profession almost exactly match. Does that mean that whenever one of them is ' ?', the other one is also? If so, what about the 7 cases where they don't match?
- In all of these cases, the person's Workclass is 'Never-worked'. This makes sense. If they have never worked, their profession cannot be anything. It seems, then, that in this case ' ?' stands for 'None'.
- As suspected, for the rest of the 1836 rows, the ' ?' for Profession and Workclass match exactly. There does not seem to be any other connection between them but this 'coincidence' alone is enough to investigate further.
- The percentage of ' ?' in the two columns is 5.63%. That is almost exactly equal to the percentage of unemployment in the UK in 2015. If the researcher knew when this data was collected (perhaps by asking the provider of the data), then a reasonable assumption could be made. A similar process could be applied to 'Residency' to determine whether the ' ?' could possibly stand for 'Homeless/Unknown Residency'.
- Going forward, we will assume that '? ' stands for 'Unemployed/Homeless'. Therefore, all these values can be left as is. For aesthetic purposes, they can be replaced by 'Unknown' or 'Unemployed' or 'Homeless' using the .replace() function.

Note: A lot of interesting visualisations can be drawn and talked through here. I will only go through a few. The most important bit here is the commentary with the plots, not the plots themselves. I.e. your ability to draw conclusions from visual information.

Interesting. It seems that as the qualification level increases, so does the average number of hours worked per week. The big dip at number 7 ('Apprenticeship') can be explained by the low number of required hours.The distributions seem to point towards non-normal data. There are also quite a few apparent outliers in 'Investment' and 'Background_diversity'.