FIT5196 - Assignment 2

Task 1

I fixed issues with the data as categorised below.

**Irregularities and Typos**

My first action was to improve the presentation of the data in job title. I removed asterisks as I saw, for whatever reason, they permeated many of the job title values without adding any meaning.

I also checked value counts for irregular spellings or variations, such as ‘ ‘ or ‘-‘ used in place of NaN. These were removed.

**Duplications**

I removed duplications by choosing a uniqueness constraint; this would merely be four columns which, alone, informed the user about the job sufficiently without using Salary, which would be rife with errors and arbitrarily unique. I felt it necessary to remove two jobs identical except in Salary as this would likely be an error, or at least not useful to the user, who would only want to see the highest available salary.

**Inconsistency**

I wanted to remove all Salaries not expressed as annual figures, so I multiplied per-hour jobs by 20 hours per week (all per-hour salary jobs were part time and I assumed a 20 hour work week for these jobs) and then by 48 weeks a year assuming they get annual leave same as full-time staff. I also removed non-numeric figures following this treatment, and this also included numeric figures following a non-numeric which followed another numeric, so for a value like ‘50000 to 55000’, the substring ‘50000 to’ would be removed leaving only ’55000’.

**Missing Values**

I assumed default values for many columns, i.e. contract type would be full-time in absence of any other data to suggest otherwise (data regarding this would be searched for in job title). I applied the same logic to contract time and permanent. Salary missing values would attract a linear regression analysis of other variables. This required dummy variables for categorical predictor variables. I attempted many different predictor variables, but ultimately, no linear model was acceptable as a way to predict salary. The scatter plot around the model was sufficiently spread, random and nebulous that I chose to reject a static average as well. I chose to leave 0 as a default for these missing values, as it allowed me to retains ‘float’ as a datatype in compliance with requirements, and inform the user that a salary will be raised upon successful application.

**Outliers**

There were two outlying salaries, both in excess of 10,000,000 pounds per year. These were set to zero as the default value.

Task 2

I largely had an easier time with task 2, which required fixes to salary format, datetime formats in opening and closing dates, and the order of columns. I also added default values in columns for consistency with jobs.df. I also had to create a randomly generated numeric ID for the second table.