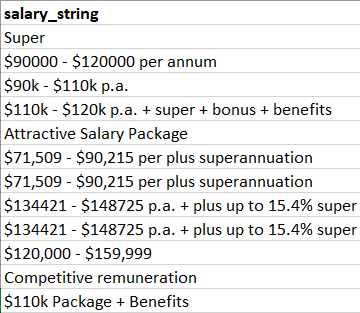
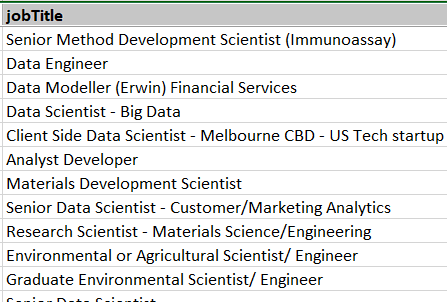
**IS THE DATA FIT FOR USE?**

**1.1 Data pre-processing**

After the data has been collected, we have to check and ensure that our Data is fit for use. After loading the data using python, we can see that our data is present in an inconsistent and complex format for some columns like salary, job title etc.

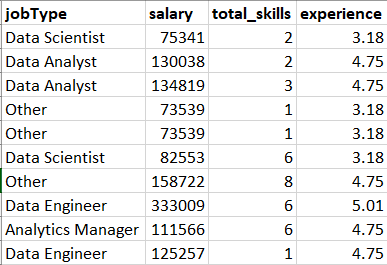




*Figure 1.1: Data in complex format*

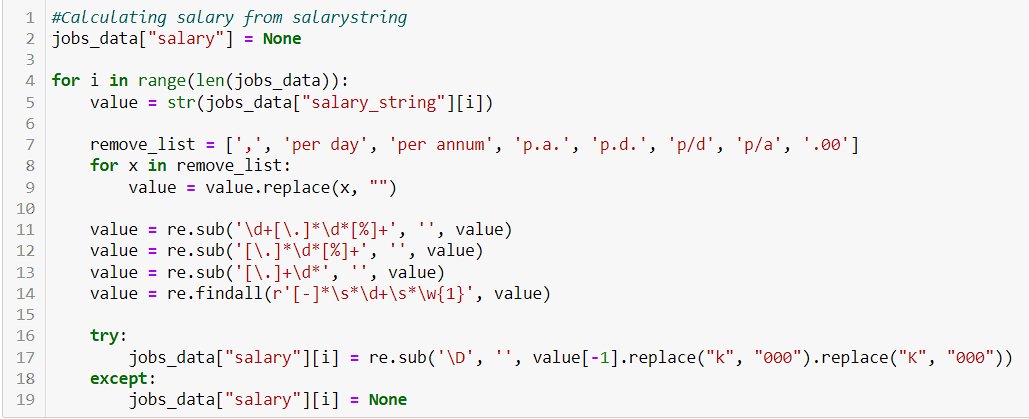
To get some useful insights from the data, we need to apply some pre-processing techniques to get data in a clean format. At the start, we remove duplicate and irrelevent rows of tuples from the dataset.

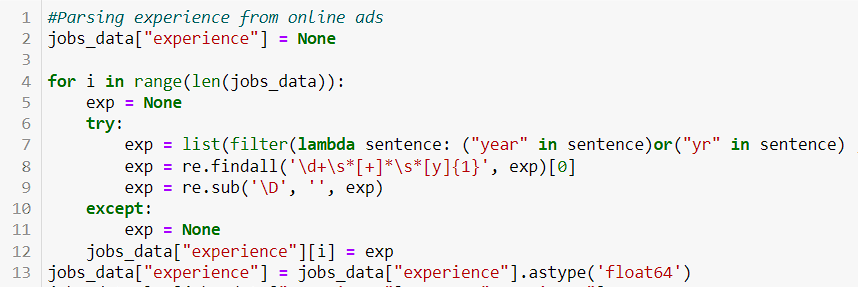
Regex is used to get cleaned data and create or replace columns by processing the already available data.



*Figure 1.2: Data in a cleaned format*

**Python code:**

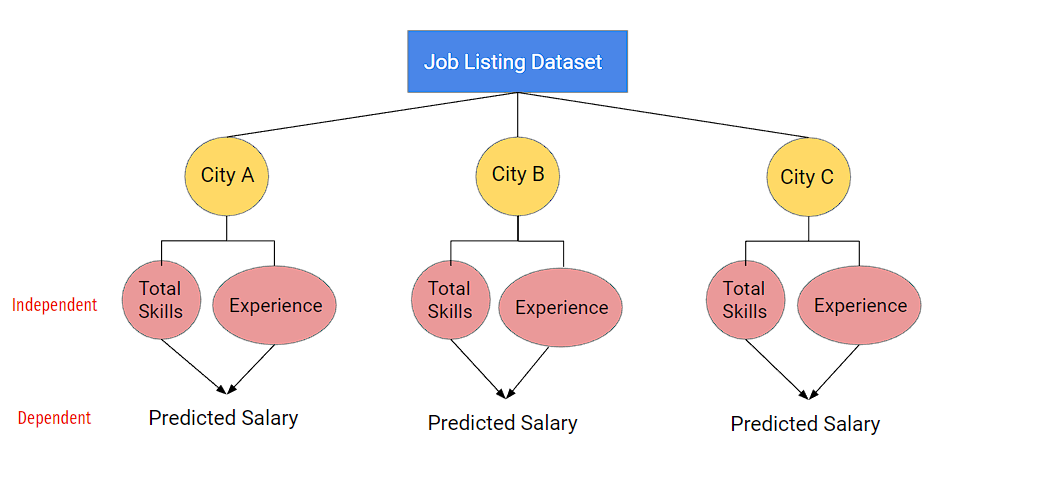




*Figure 1.3: Python data to parse salary and experience*

**1.2 Imputation for missing values**

After the Data has been cleaned, we move on to impute value for the missing values in numerical columns. For imputation, we fill the missing values with the mean of the column. For the salary column, we will develop and fit a linear model on total\_skills, experience, city to predict the value of missing values for salaries. Here a different model is developed for each different location. This is done to not weaken the relationship between salary and location.



*Figure 1.4: Linear Model to predict salary*

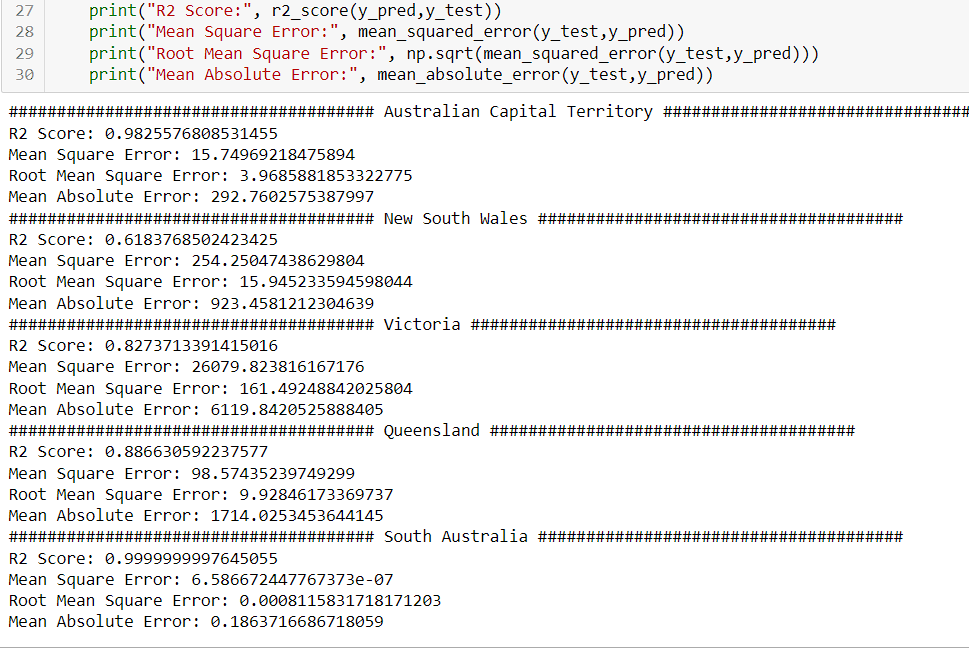
A linear regression model describes the relationship between a dependent variable, y, and one or more independent variables, X. The dependent variable is also called the response variable. Independent variables are also called explanatory or predictor variables. Continuous predictor variables are also called covariates, and categorical predictor variables are also called factors. The matrix X of observations on predictor variables is usually called the design matrix.

A multiple linear regression model is:

***yi* = *β*0+*β*1*Xi*1+*β*2*Xi*2+⋯+*βpXip*+*εi*, *i*=1,⋯,*n***



*Figure 1.5: python Linear Model to predict salary*

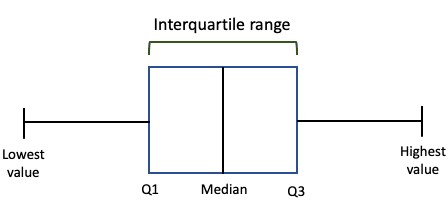


*Figure 1.6: Model Statistics*

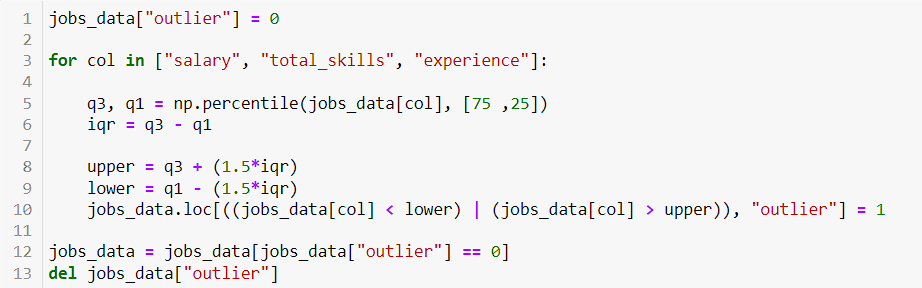
**1.3 Outlier Detection**

After the data has been imputed, outlier detection is used to remove outliers to get consistent data which will be used to build visualisations and gain insights. For outlier detection, quantile method is used, where the data between the range of max and min calculated using inter quartile range is only kept. For detection we have assumed that the data is normally distributed. The rule of data falling between the upper and lower limit of box plot is considered, where every value outside of this range is removed from the data set.

|  |  |
| --- | --- |
| Point | Definition |
| Q3 | 75th percentile |
| Q2 | Middle Value ( Median) |
| Q1 | 25th percentile |
| IQR | Q3 – Q1 |
| Highest | Q1 + 1.5 x IQR  (UPPER Whisker) |
| Lowest | Q1 - 1.5 x IQR (LOWER Whisker) |
| OUTLIERS | Value lying outside Maximum or Minimum |

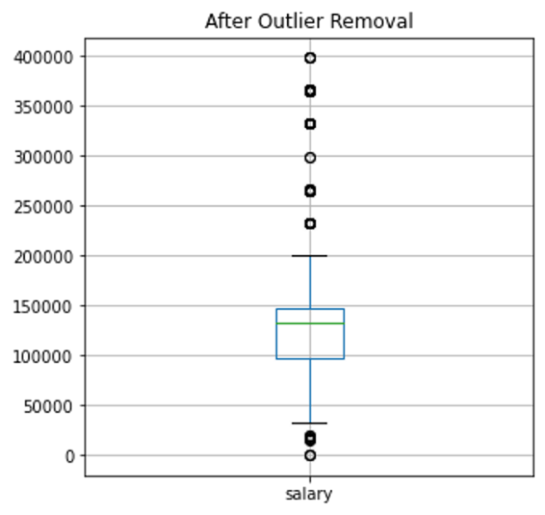
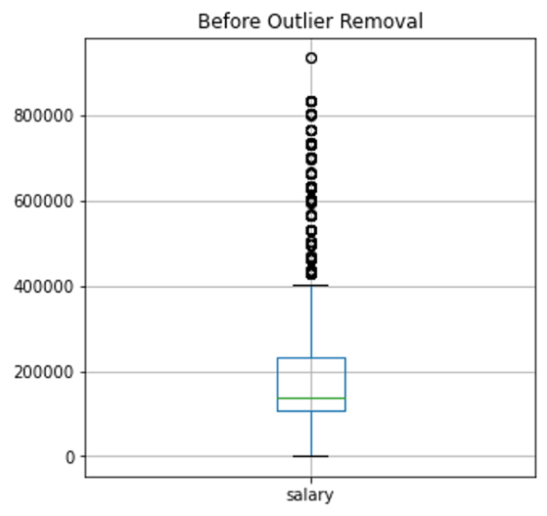


*Figure 1.6: Boxplot definitions*



*Figure 1.7: Python code to remove outliers*

Here, we will implement the quartile method to remove outliers from salary, total\_skills and experience column. Only the values between 25th and 75th percentile are kept other values are removed from the dataset.



*Figure 1.8: Box Plot for Before and After outlier Removal*

**REFERENCES**

[1] What is a linear regression model? - MATLAB & Simulink. (n.d.). MathWorks - Makers of MATLAB and Simulink - MATLAB & Simulink. <https://www.mathworks.com/help/stats/what-is-linear-regression.html>

[2] (n.d.). https://cdn.scribbr.com/wp-content/uploads/2020/09/iqr\_boxplot.png.