**Data Analytics Assignment Documentation**

­

Abhinna Manandhar

BSc. (Hons) Computing, Softwarica College of IT and E-Commerce,

Coventry University

SP6000COM Data Analytics

Nahakul Prasad Jaishi

March-01-2023

# Keywords

Figure : Keywords

Timeline

Description automatically generated with medium confidence

Contents

[Keywords 2](#_Toc128563053)

[Table of figures 4](#_Toc128563054)

[EDA 6](#_Toc128563055)

[Analysis 2 7](#_Toc128563056)

[Data Cleaning 13](#_Toc128563057)

[Removing Outliers 15](#_Toc128563058)

[EDA on cleaned data. 17](#_Toc128563059)

[Tableau Dashboard 25](#_Toc128563060)

# Table of figures

[Figure 1: Keywords 2](#_Toc128563021)

[Figure 2: Average middle school percent according to gender. 6](#_Toc128563022)

[Figure 3: Average high school percentage according to gender. 7](#_Toc128563023)

[Figure 4: Average undergraduate percentage according to gender. 7](#_Toc128563024)

[Figure 5: Average MBA percentage based on genders. 8](#_Toc128563025)

[Figure 6: Average salary based on work experience. 8](#_Toc128563026)

[Figure 7: Average salary based on specialisation. 9](#_Toc128563027)

[Figure 8: Relationship between degree percentage, MBA percentage and salary. 10](#_Toc128563028)

[Figure 9: Relationship between experience test percentage, MBA percentage, and salary. 11](#_Toc128563029)

[Figure 10: Correlation matrix before data cleaning. 12](#_Toc128563030)

[Figure 11: Removing sl\_no column. 13](#_Toc128563031)

[Figure 12: Replacing NaN values in salary column. 13](#_Toc128563032)

[Figure 13: Replacing ssc\_b and hsc\_b with binary indecators. 13](#_Toc128563033)

[Figure 14: Splitting hsc\_s column. 14](#_Toc128563034)

[Figure 15: Splitting degree\_t column. 14](#_Toc128563035)

[Figure 16: Implementing Binary Indicators. 14](#_Toc128563036)

[Figure 17: Function to remove outliers. 15](#_Toc128563037)

[Figure 18: Removing outliers 16](#_Toc128563038)

[Figure 19: Boxplots before data cleaning. 16](#_Toc128563039)

[Figure 20: Boxplot after data cleaning. 17](#_Toc128563040)

[Figure 21: Correlation matrix 18](#_Toc128563041)

[Figure 22: Average salary based on different factors. 19](#_Toc128563042)

[Figure 23:Placement based on specialization 19](#_Toc128563043)

[Figure 24: Placement based on degree type and gender. 20](#_Toc128563044)

[Figure 25:Placement status based on degree\_t and gender 21](#_Toc128563045)

[Figure 26: Average work experience by gender. 21](#_Toc128563046)

[Figure 27: Salary distribution by specialization. 22](#_Toc128563047)

[Figure 28: Placement count by degree type. 23](#_Toc128563048)

[Figure 29: Average salary by degree type 23](#_Toc128563049)

[Figure 30: Average salayr by highschool specialization. 24](#_Toc128563050)

[Figure 31: Placement status based on degree type. 24](#_Toc128563051)

[Figure 32: Tableau dashboard. 25](#_Toc128563052)

# Introduction

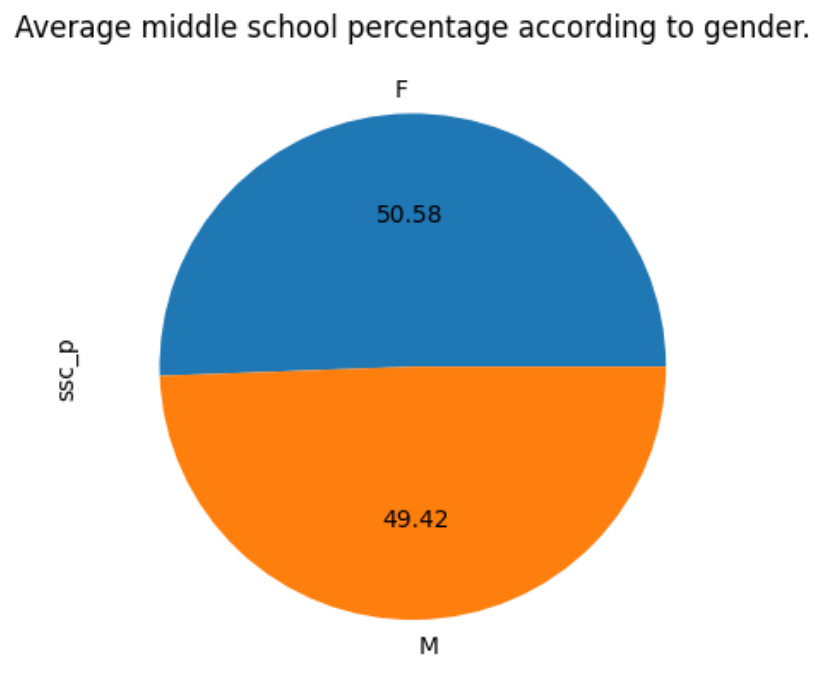
In this documentation we will analyze the provided dataset about the student’s education history, and their job placement status, salary and other factors. This will allow us to analyze the scope of different specialization fields and their benefits to make better decisions. We will start by analyzing the data and then continue to clean it. After cleaning the data, we will again perform EDA to compare the changes in the data and generate our desired results and perform visualization using different tools like Tableau.

# EDA

Exploratory data analysis helps data analysts to gain more detail about the data and perform further analysis on the data itself.

Performing EDA on our dataset helps in discovering the relationship between different entities and create appropriate matrices accordingly.

Figure : Average middle school percent according to gender.



The above analysis displays the average percentage obtained by each genders in middle school. We can conclude that that the differences in percentage based on genders is significantly low.

## Analysis 2

Figure : Average high school percentage according to gender.

Chart, pie chart

Description automatically generated

The analysis of percentage scored in high school is also similar.

Figure : Average undergraduate percentage according to gender.

Chart, pie chart

Description automatically generated

Figure : Average MBA percentage based on genders.

Chart, pie chart

Description automatically generated

The average percentage difference between the genders in undergraduate is only of 1%. And 2.92 in MBA.

Therefore, we can conclude that gender has very less impact on the average percentage scored in all levels.

Figure : Average salary based on work experience.

Chart, pie chart

Description automatically generated

The average salary of an individual based on their work experience has shown a fairly good amount of difference with individuals with previous work experience having more chances of getting higher salary jobs.

Figure : Average salary based on specialisation.

Chart, bar chart

Description automatically generated

Individuals with specialization in marketing and finance have shown to get paid higher salary in jobs than individuals with specialization in Marketing and HR.

Figure : Relationship between degree percentage, MBA percentage and salary.

Chart, scatter chart

Description automatically generated

The percentage obtained by individuals in ther undergraduate degree and MBA have also shown a good relationship with each other with salaries identified by the color range.

Because the outliers have not been removed currently and the analysis is performed in the raw dataset without any cleaning or manipulation, the salary ranges display a noticeable outlier with color yellow between 60-65 percentage in MBA in the above scatter plot.

Figure : Relationship between experience test percentage, MBA percentage, and salary.

Chart, scatter chart

Description automatically generated

The scatterplot displaying the relationship between the experience test percentage, MBA percentage and salary also has a somewhat similar output.

Figure : Correlation matrix before data cleaning.

Chart

Description automatically generated

Likewise, the above heatmap visualizes the correlation between all the numeric variables.

# Data Cleaning

Data cleaning is the process of filtering our data to further support our analysis. The cleaning process may involve multiple steps. Such as replacing or removing the null values, converting categorical data into binary standards either by simply replacing the values if possible or by splitting the column into multiple columns and then providing binary values.

We start cleaning our data by first removing the “sl\_no” column as it is unnecessary in our case.

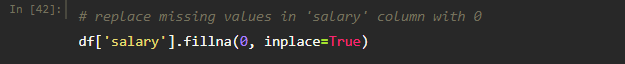
Figure : Removing sl\_no column.

Text

Description automatically generated

We then continue to replace the null values in the salary column with 0 as they represent that the students are not placed for any job or are not working.

Figure : Replacing NaN values in salary column.



Since the 'ssc\_b' and 'hsc\_b' columns simply have two unique values, we replace the categorical data types with binary inficators.

Figure : Replacing ssc\_b and hsc\_b with binary indecators.

Text

Description automatically generated

For the columns 'hsc\_s', and 'degree\_t', we added a new binary indicator column. This is because the column consists of more than two unique values. So we cannot directly convert them into binary. Thus, we split the columns into multiple columns named after the unique values and provide them with binary values accordingly.

Figure : Splitting hsc\_s column.

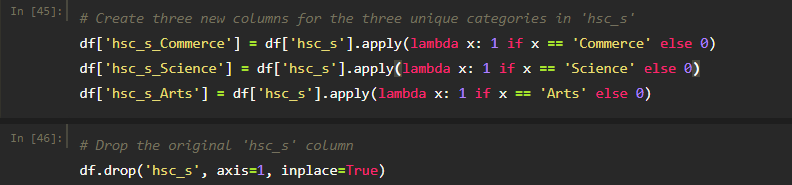


Figure : Splitting degree\_t column.

Text

Description automatically generated

We then continue to implement binary indicators in as many columns as possible.

Figure : Implementing Binary Indicators.

A screenshot of a computer

Description automatically generated with medium confidence

## Removing Outliers

Outliers are special types of data that fall outside the range of the variable. Outliers have a significant impact on our analysis as they are capable of changing our visualizations and machine learning models completely. Detecting and removing outliers are a vital part of data cleaning as they enable us to produce more accurate results from our analysis.

The columns containing quantitative data are “salary”, “ssc\_p”, “ssc\_p”, “degree\_p”, “etest\_p” and “mba\_p”. Boxplot is one of the best visualization techniques to identify outliers in a dataset. Therefore, we move forward by producing boxplots for these quantitative data.

In order to remove the outliers, we have first created a function that finds and remove the outliers for any given quantitative column. This is done by calculating the first and the third quartile and then calculating the interquartile range. Then by further calculating the upper and lower bounds, we can remove the outliers that fall outside the bounds.

Figure : Function to remove outliers.

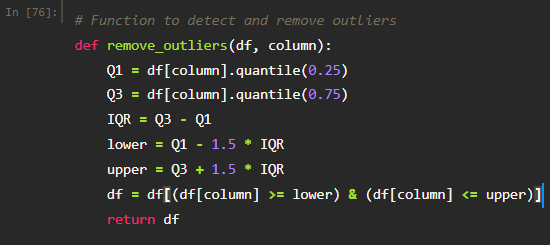


Figure : Removing outliers

Text

Description automatically generated

Figure : Boxplots before data cleaning.

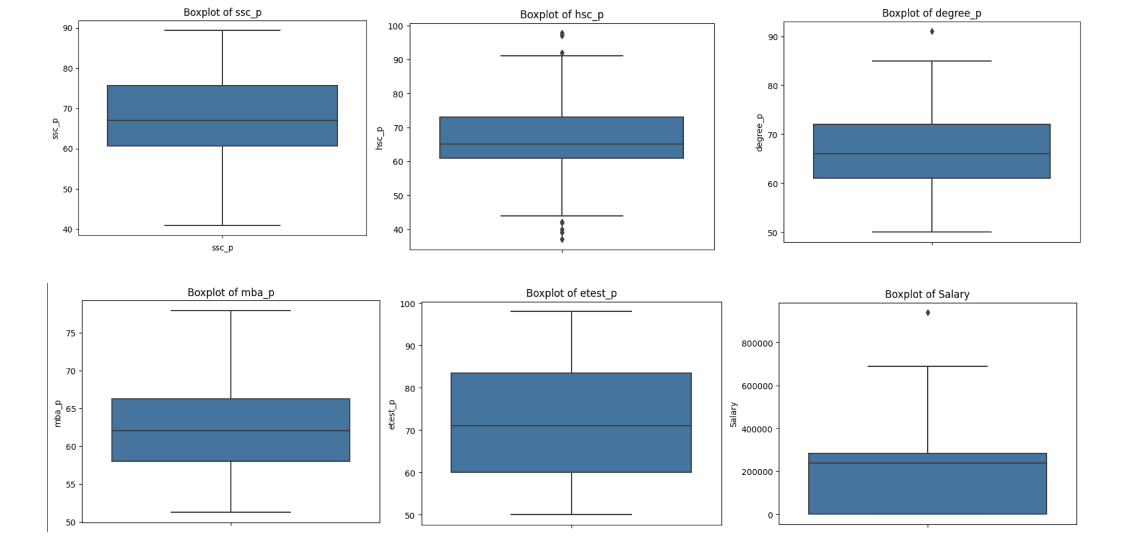


Figure : Boxplot after data cleaning.

Chart, box and whisker chart

Description automatically generated

We can now see that all the columns don’t have any outliers in them. And with that we can now say that our data has been cleaned.

# EDA on cleaned data.

By performing EDA on our cleaned data, we can now analyze the differences in the dataset before and after cleaning. We can also look for any visible patterns in the dataset by performing EDA.

Figure : Correlation matrix

Timeline

Description automatically generated

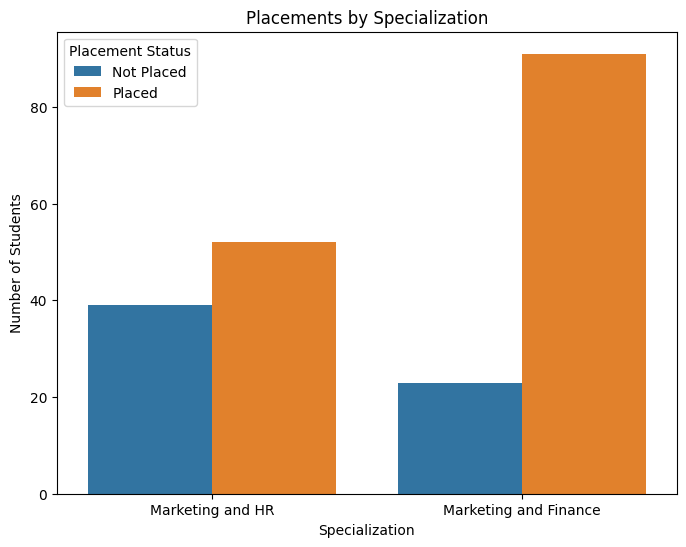
Figure 22: Average salary based on different factors.

Chart, bar chart

Description automatically generated

The difference between the average salary based on genders and specializations have seen a significant difference after the data cleaning with an estimate of 50000 between male and female and between Mkt&HR and Mkt&Fin.

Figure :Placement based on specialization



The plot shows that the students with specialization in Marketing and HR have a comparatively lower placement rate compared to those in Marketing and finance

Figure : Placement based on degree type and gender.

Chart, bar chart

Description automatically generated

The above visualization shows the placement count of different high school specialization and their genders. It is seen that commerce has the placement ratio between the genders in commerce and science are somewhat similar. However, Arts has a 100% female population with the lowest placement count.

Figure :Placement status based on degree\_t and gender

Chart, bar chart

Description automatically generated

The degree type Commerce and management is seen to have the highest placement count with slightly higher females being placed. Then the science and technology stream is seen to have a higher male placement rate and the others having the lowest placement with 100% female population.

Figure : Average work experience by gender.

Chart, bar chart

Description automatically generated

The male population is seen to have a higher average work experience right above 0.35 than the female population i.e around 0.275.

Figure : Salary distribution by specialization.

Chart, bar chart

Description automatically generated

The above visualization summarizes that most students have a salary between 200000 and 300000. It is also seen that most students ini the higher salary range are from Marketing and Finance.

Figure : Placement count by degree type.

Chart, bar chart

Description automatically generated

The above plot shows the placement rate of students based on their degree type.

Figure : Average salary by degree type

Chart, pie chart

Description automatically generated

Te above pie chart shows the average salar of different degree types.

Figure : Average salayr by highschool specialization.

Chart, bar chart

Description automatically generated

The average salary based on highschool specializations shows interesting numbers with Science students with highest average salary followed by commerce and arts respectively.

Figure : Placement status based on degree type.

Chart, bar chart

Description automatically generated

The placement count based on degree types shows that most students in commerce and management and science and technology stream are placed while it is quite the opposite for other degree types.

# Tableau Dashboard

Tableau is a data visualization tool that allows us to appropriately visualize our data to extract useful information. Using tableau, we can now further analyze our data to gain more insights that can allow us to make better business decisions.

Figure : Tableau dashboard.

