

**ANL252**

**Python for Data Analytics**

**Tutor-Marked Assignment (TMA01)**

**July 2022 Presentation**

**Submitted by:**

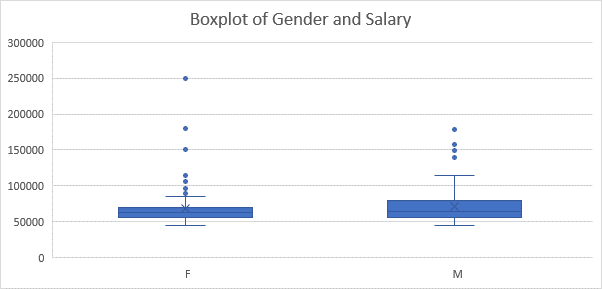
|  |  |
| --- | --- |
| **Name** | **PI No.** |
| **Jeffrey Lim Dao Fong** | **Y2111639** |

**Tutorial Group : T09**

**Instructor’s Name : Mr Kumar Munish**

**Submission Date : 14 August 2022**

**Q1a) Chart 1: Boxplot of Gender and Salary**



|  |  |  |
| --- | --- | --- |
|  | **Male** | **Female** |
| **Count** | **106** | **144** |
| **Average Salary** | **$69,389.50** | **$69,118.50** |
| **Median** | **$63,025** | **$63,001.50** |
| **Mode** | **$63,025.00** | **$63,025.00** |

Male

From the chart, it is evident that the salary of the male population has a larger interquartile range between the 25th percentile and 75th percentile. The male population have also lesser outliers as compared to the female population. The male population although have a larger variability in salary as compared to the female, the outliers present are significant lesser than the female population.

Female

From the chart, it is evident that the salary of the female population has a lesser interquartile range between the 25th and 75th percentile. This means that between the 25th and 75th percentile, there is lesser variability between the salary drawn. However, there is significant outliers present in the chart. From the chart, it is evident that there are more females that are drawing higher salary as compared to the males.

Overall, the average salary drawn for males is higher than the female population. But there is more females drawing higher salaries as compared to the males outside of the whiskers.

**Chart 2: Histogram of Citizenship of employees**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Citizen** | **PR** | **Non-resident** |
| **Counts** | **197** | **37** | **16** |

From the chart, it is evident many employees are Citizens, standing at 78.8%. Whereby 37 employees are Permanent residents (PR), at 14.8% and lastly 16 Non-residents at 6.4%.

**Q1b)**

**Data Cleaning & Processing**

In this section, the data will be cleaned and processed to be used for the data study.

|  |
| --- |
| In [1]: *# importing libraries*  **import** warnings  warnings**.**filterwarnings("ignore")  **import** numpy **as** np  **import** pandas **as** pd  **import** io  **import** requests  pd**.**set\_option('display.max\_columns', **None**)  **import** matplotlib.pyplot **as** plt  **import** seaborn **as** sns  **%matplotlib** inline  **%config** InlineBackend.figure\_format = 'retina'  plt**.**style**.**use("ggplot")  In [2]: df**.**sample(5, random\_state**=**999) |

**Data Cleaning**

* Check and rename/modify some column names
* Check for missing values
* Standardising the format of the dates

|  |
| --- |
| In [4]: df['LeftDate']**.**fillna('1/5/2022', inplace**=True**)  In [5]: df['BirthYear'] **=** pd**.**to\_datetime(df['BirthYear'])  df['JoinDate'] **=** pd**.**to\_datetime(df['JoinDate'])  df['LeftDate'] **=** pd**.**to\_datetime(df['LeftDate'])  In [6]: df**.**sample(5, random\_state**=**999) |

**Data Exploration and Visualisation**

The dataset is now considered to be cleaned and we are ready to start visualising and exploring the features.

|  |
| --- |
| In [7]: plt**.**figure(figsize **=** (20,15))  sns**.**boxplot(df['Gender'], df['Salary']);  plt**.**title('Figure 1: Boxplot of Gender and Salary', fontsize **=** 15)  plt**.**show();    In [8]: plt**.**figure(figsize **=** (20,15))  fig **=** sns**.**countplot(x **=** 'Citizenship', data **=** df, palette **=** 'magma',  order **=** df['Citizenship']**.**value\_counts()**.**index)  fig **=** plt**.**title('Figure 2: Citizenship', fontsize **=** 15)  plt**.**show() |

**Q1c) Finding the length of service**

|  |
| --- |
| In [9]: df.head()  In [10]: df['ServiceDays'] = (df['LeftDate'] - df['JoinDate'])  In [11]: df['ServiceYears'] = round(df['ServiceDays'] / np.timedelta64(1, 'Y'),1)  In [12]: df.head()  ### Maximum number of years in service  In [13]: df['ServiceYears'].max()  Out [13]: 16.0  ###Minimum number of years in service  In [14**]:** df['ServiceYears'.min()  Out [14]: 0.1  ### Average number of years in service    In [15]: round(df['ServiceYears']**.**mean(),1)  Out [15]: 6.6 |

**Q1d) Interactive user input which allows the user to query if a particular person was/is a staff of the organization**

|  |
| --- |
| In[] : **def** check(name):  **if** name **==** 'exit':  **return**  **elif** name **in** df['Staff']**.**unique():  print(‘He/she is a staff.\n\nEnter name or "exit" to quit.’)  check(input())  **else**:  print(‘He/she is not a staff.\n\nEnter name or "exit" to quit.’)  check(input())  print('Enter name or "exit" to quit.')  name **=** input()  check(name) |