Fundamentals of Information Visualisation

Module Code: COMP3021

Module Convenor: Ke Zhou

Student ID Number: 20299113

Word count: 1901

1. Introduction

Data Description

The dataset is collected from a survey of 1470 IBM employee by IBM data scientists. It contains basic information, background, position, working environment, payment, satisfaction score and employee attrition state from each employee, a total of 35 variables.

Aim

This report aims to analyse the possible factors of employee attribution and several exploratory insights of jobs by data visualisation. Initial and further questions have been created to show the whole analysis procedure. Furthermore, visualisation strategies are explained following each graph.

Figure 1 is an interactive sunburst graph of all the variables in six categories. Questions are deliberately selected from each category.

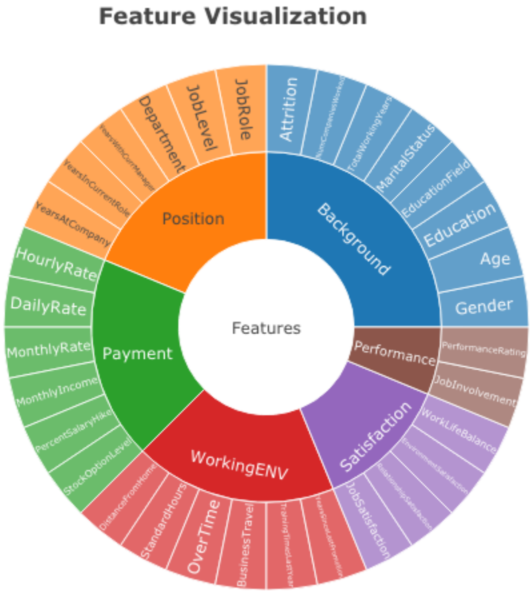

Figure 1: Feature Visualisation by six categories

2. Initial Questions

The initial questions are about how these variables influence employee attrition from these six categories.

1. What is the distribution of employee attrition?

This question asks the proportion of resignation and non-resignation in the survey, aiming to have an overview of employee attrition.

2. How does age influence employee attrition?

Age is an important feature of each employee. The question aims to find the difference between resignation and non-resignation on age.

3. How does job level influence employee attrition?

IBM define a job in 5 levels, from low to high. This question aims to compare employee attrition at each level, then drag insights from it.

4. How does working overtime influence employee attrition?

Working overtime is supposed to be a crucial factor of employee attrition. This question aims to find the difference between resignation and non-resignation employee on overtime condition.

5. How do job satisfaction and working-life balance influence employee attrition?

There are four levels in job satisfaction and balance of work and life, from low to high. This question is to compare the difference between resignation and non-resignation on job satisfaction and working-life balance.

6. How does salary influence employee attrition?

Salary is supposed to be a significant factor of jobs. The question aims to make a comparison between resignation and non-resignation on salary.

3. Fitness of the data

The fitness of data is assessed by the process of "DataSet Check" and "Data Type Check".

In "DataSet Check", library mice is used to find the missing observations in the dataset. It proves there is no missing data in the dataset.

There are four types of data which are nominal, ordinal, interval and ratio. Variables of *Education*, *EnvironmentSatisfaction*, *JobInvolvement*, *JobSatisfaction*, *PerformanceRating*, *RelationshipSatisfaction* and *WorkLifeBalance* are all ordinal data, although the dataset uses the numerical value represent them. Therefore, it is necessary to transform these numerical data into categorical data before the visualisation. In general, this dataset is usable.

4. Virtualization to Initial Questions

4.1. What is the distribution of employee attrition?

Figure 2 shows that 16.1% of employees resigned from the job, which is a totally of 237 employees. 83.9% of employees still stay at company, which is a totally of 1233 employees.

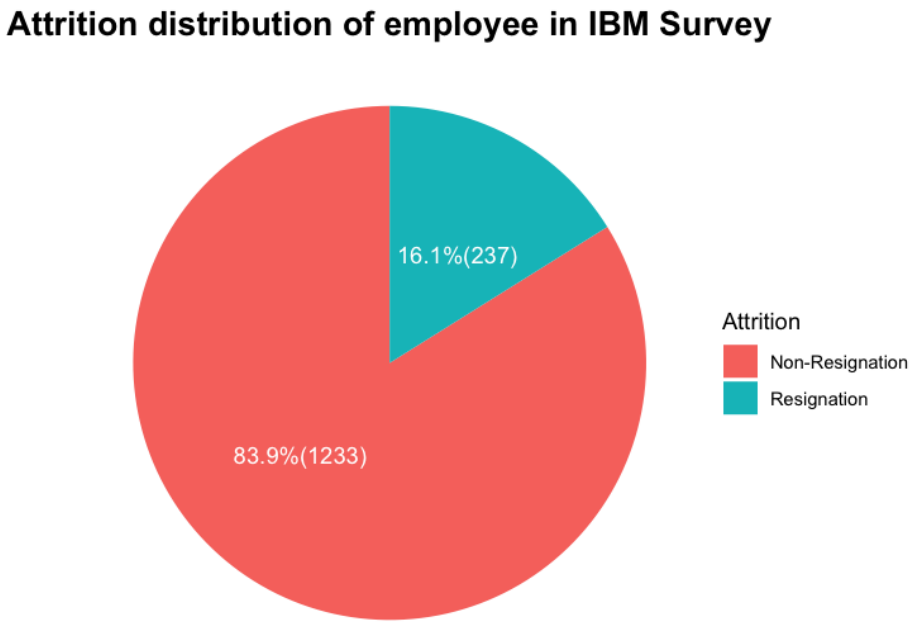

Figure 2: Attrition distribution of employee in IBM Survey

Visualisation Strategies

Ajibade and Adediran (2016) state that the best way to utilise a pie chart is when there are not many components in the data and add percentage and text on the graph. Therefore, as there are only two categories that need to display, a pie chart is used to illustrate the proportion of resignation and non-resignation (Abela, 2009).

The feature *Attrition* was transformed into an R-dataframe with frequency and percentage. Then, use this R-dataframe to create the pie chart. The style and size of the title and legend are adjusted to show a user-friendly graph.

4.2. How does age influence employee attrition?

Figure 3 shows that the younger employee has a higher possibility of resigning, especially those younger than approximate 33 years old.

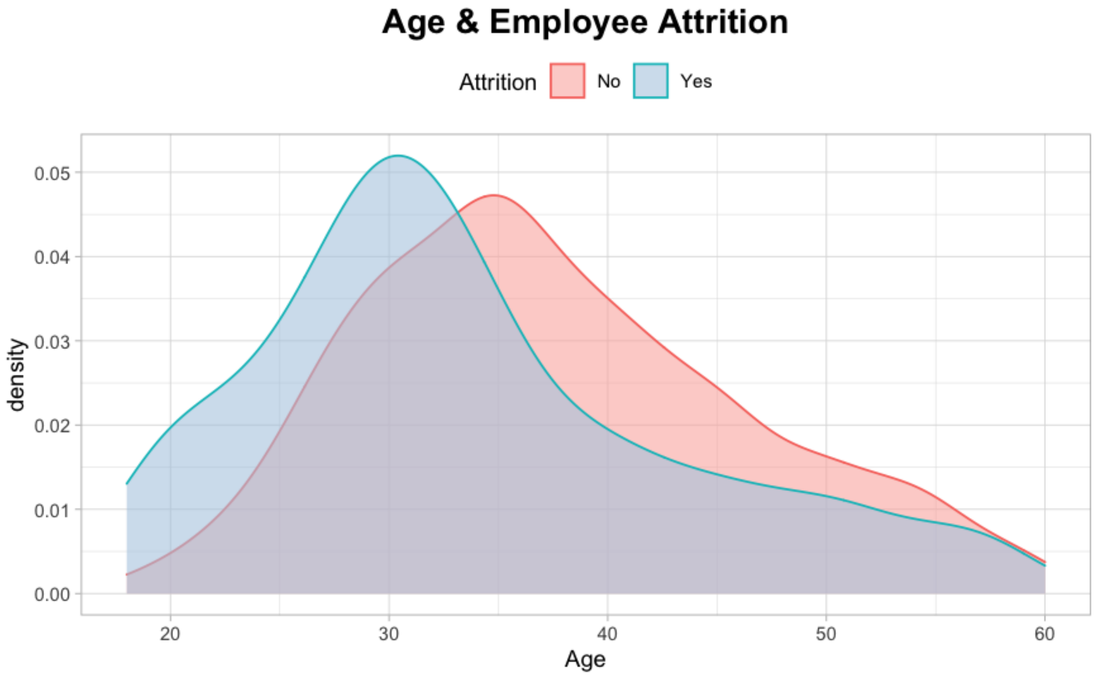

Figure 3: Employee Age and Employee Attrition

Visualisation Strategies

The density plot is a smooth version of the histogram plot, and it displays the distribution of continuous variables(datavizcatalogue.com, n.d.). It is easier comparable than a histogram because the difference and the overlap area is easily recognisable.

The graph's transparency and colour are customised by coding to see how the displayed area is different clearly. An interactive graph has also been created to show the exact value of any position on the graph. Furthermore, it can display only one category by a simple click on the name of the category so that people can focus on a chosen one.

4.3. How does job level influence employee attrition?

As shown in Figure 4, the employee attrition rate is lower when people are at a high job level. The job at the beginner level has the most employee attrition rate - 26.3%, which is almost three times than level 2, two times than level 3. It is also the only level beyond the overall employee attrition rate, which is 16.1% in the Q1's pie plot.

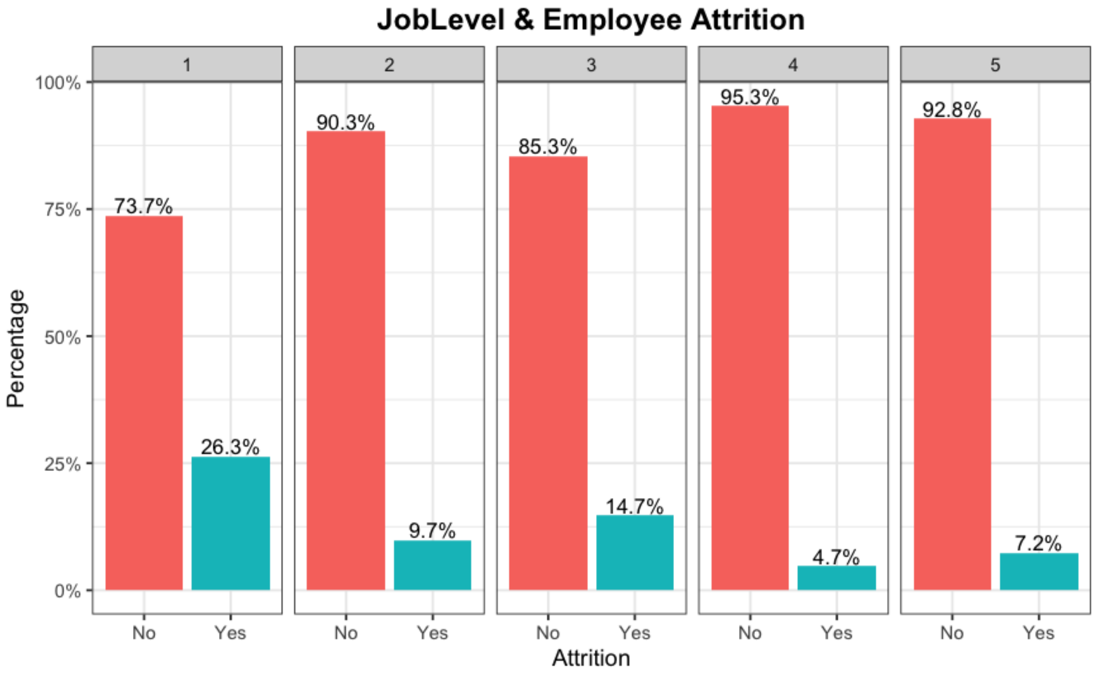

Figure 4: JobLevel and Employee Attrition

Visualisation Strategies

Bar chat is commonly used in comparing data in few categories. A recommended number of categories is less than ten because it will let people feel messy when there are too many categories in one graph (Klipfolio, 2019). Therefore, a clustered bar chart is used here as there are only four categories and two groups in each category. Due to the comparison object is employee attrition rate, the proportion is shown on each bar. In this way, the employee attrition rate can be compared in and between each job level.

4.4. How does working overtime influence employee attrition?

It can be seen from Figure 5 that more than half of the employee who chose to resign has encountered overtime working. On the contrary, the proportion in the non-resignation group is fewer than a quarter.

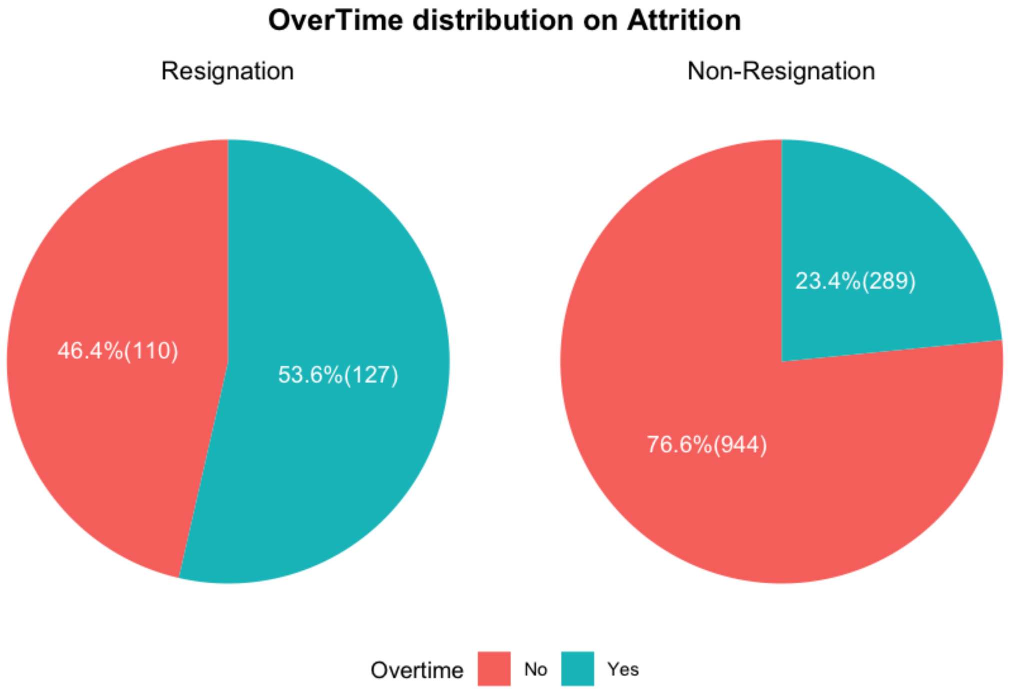

Figure 5: OverTime distribution on Employee Attrition

Visualisation Strategies

As there are only two groups in two categories, a pie chart is introduced to compare the working overtime proportion. An alternative solution can be a bar chart because the comparison scenario also applied. However, as there are only two categories and the main focus is on proportion, the pie chart works better.

Firstly, a transformation of the data is applied to divide the dataset by whether or not resigning. Then, make a calculation to get the frequency and proportion. At last, store this information in an R-dataframe. The size, style and colour are customised. A shared legend is created at the bottom of the graph to illustrate the representation of the colour.

4.5. How does job satisfaction and working-life balance influence employee attrition?

Figure 6 shows that the lower job satisfaction the employee has, the higher possibility the employee will resign. Statistically, the interquartile range of the job satisfaction in the resignation group is from 1~3, 2~4 in the non-resignation group. The data is skewed, and it exactly meets the conclusion.

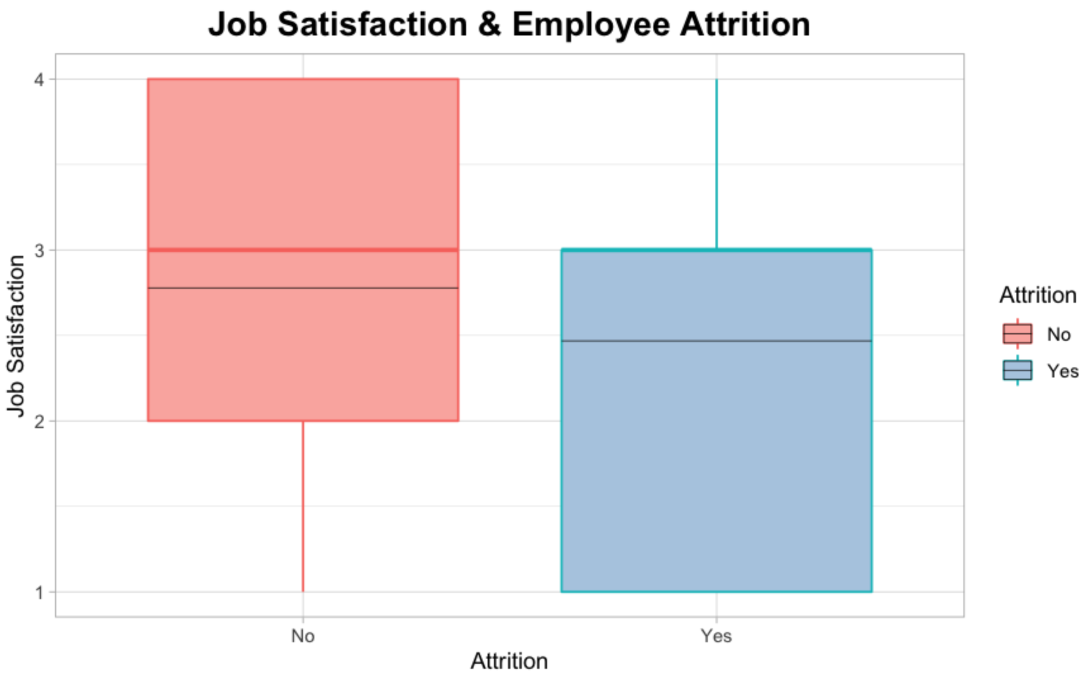

Figure 6: Job Satisfaction and Employee Attrition

Visualisation Strategies

The box plot is introduced due to the dispersion of the data can be displayed intuitively, such as the minimum, first quartile, median, third quartile and maximum (Galarnyk, 2018). It only displays several critical values of the distribution of data without showing the shape of the distribution so that the comparison can be more straightforward.

As the original box plot does not contain the mean value of the data, an extra line in black represents the mean value added in the box plot. In this way, the centrality of the data can also be covered in the graph. That increase the effectiveness of the visualisation.

Figure 7 shows that the worse the working-life balance, the higher the employee attrition rate. However, the trend is slightly different in the "Best" category of WorkLifeBalance. It may be that those people pursue a better opportunity as they have already done great at their current job.

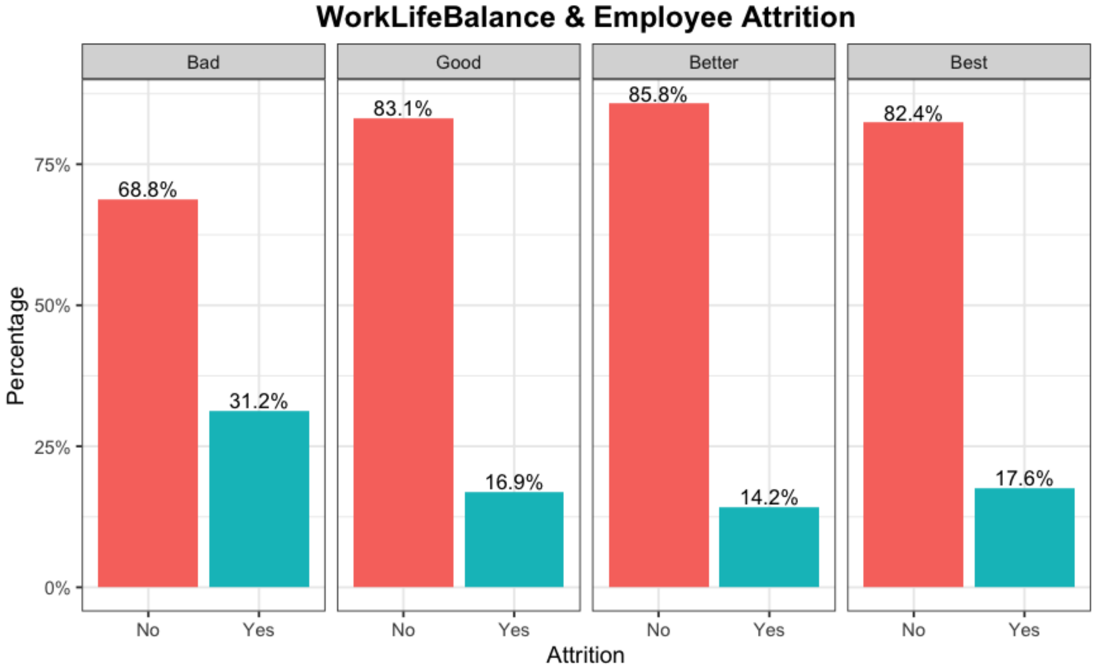

Figure 7: WorkLifeBalance and Employee Attrition

Visualisation Strategies

A clustered bar chart is used here as there are only four categories and two groups in each category. It is a similar scenario with Q3. The proportion value is shown at the top of each bar for convenience.

4.6. How does salary influence employee attrition?

Figure 8 shows that the employee with a higher salary will have a lower probability to resign. Although fewer employees resign from the job with a high salary, they are outliers shown as blue points on the graph. They do not affect the overall trend. The first quartile salary at the non-resignation group is greater than the median salary at the resignation group. The median salary at the non-resignation group is greater than the mean salary at the resignation group. The mean salary is larger than the median salary indicates that a few employees earn much greater than the overall level (Rudy, 2019). The salary distribution is skewed to the right.

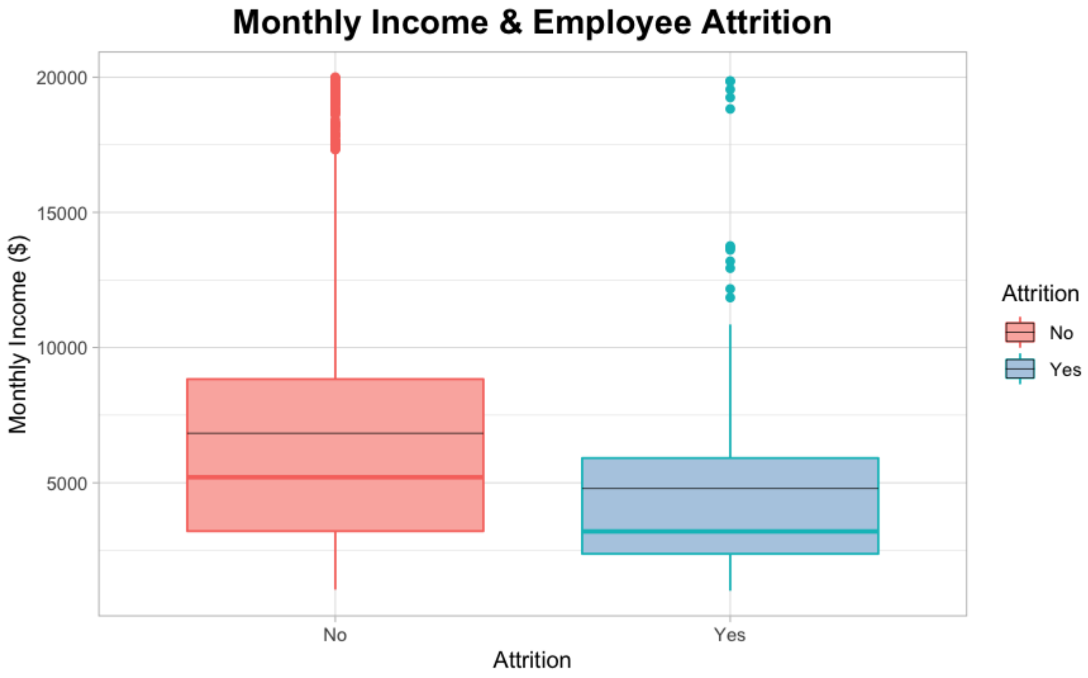

Figure 8: Monthly Income and Employee Attrition

Visualisation Strategies

The box plot is used to compare the difference between resignation and non-resignation employee on salary by dispersion and centrality information. It is crucial to calculate and show the statistical value of the salary so that more insights can be discovered. The exact value of these statistical value can be found in the interactive version of the graph to make the analysis more precisely.

5. Further Questions

5.1. Which job earn the most salary overall?

As shown in Figure 9, the manager earns the most salary overall. Moreover, the minimum salary of the manager is even larger than the third quartile salary of all the job except the research director.

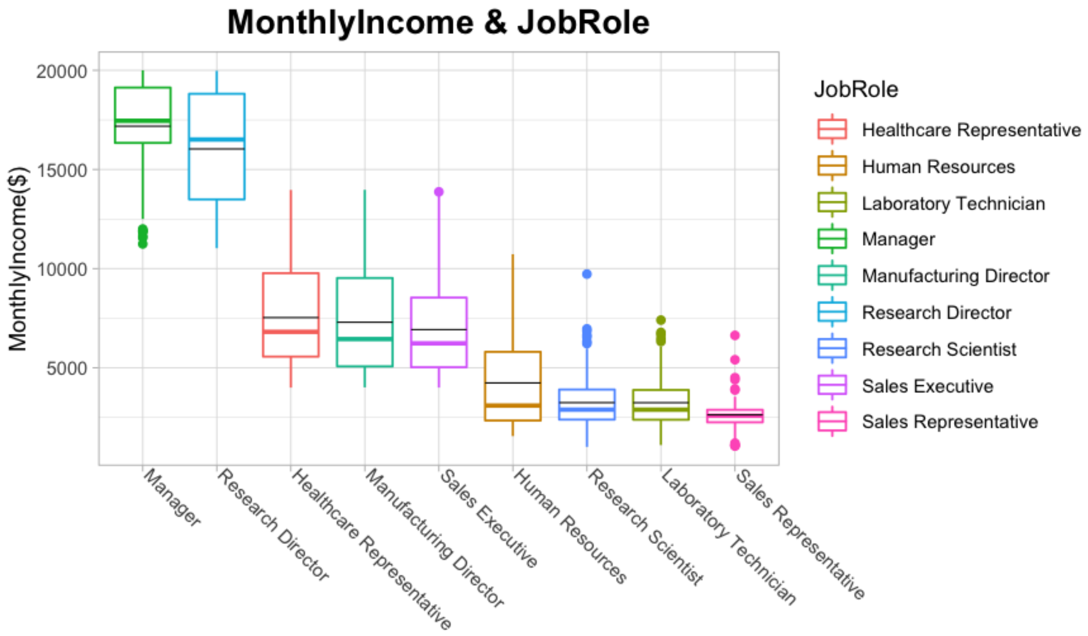

Figure 9: Monthly Income and JobRole

Visualisation Strategies

As the box plot can display the dispersion and centrality of the data, it is more persuasive than a bar chart that can only compare a single value. The order of JobRole is sorted by salary decreasingly. Different job is in different colours to ensure category distinguishability. As several names of JobRole is a little bit longer, all the names are rotated by 45 degrees. Additionally, an interactive graph has also been created for more flexible and precise comparison.

5.2. What's the relationship between total working years and monthly income?

Figure 10 shows that the longer working years, the higher salary. They have a strong positive correlation with each other. A more detailed correlation relationship has been shown in Figure 11.

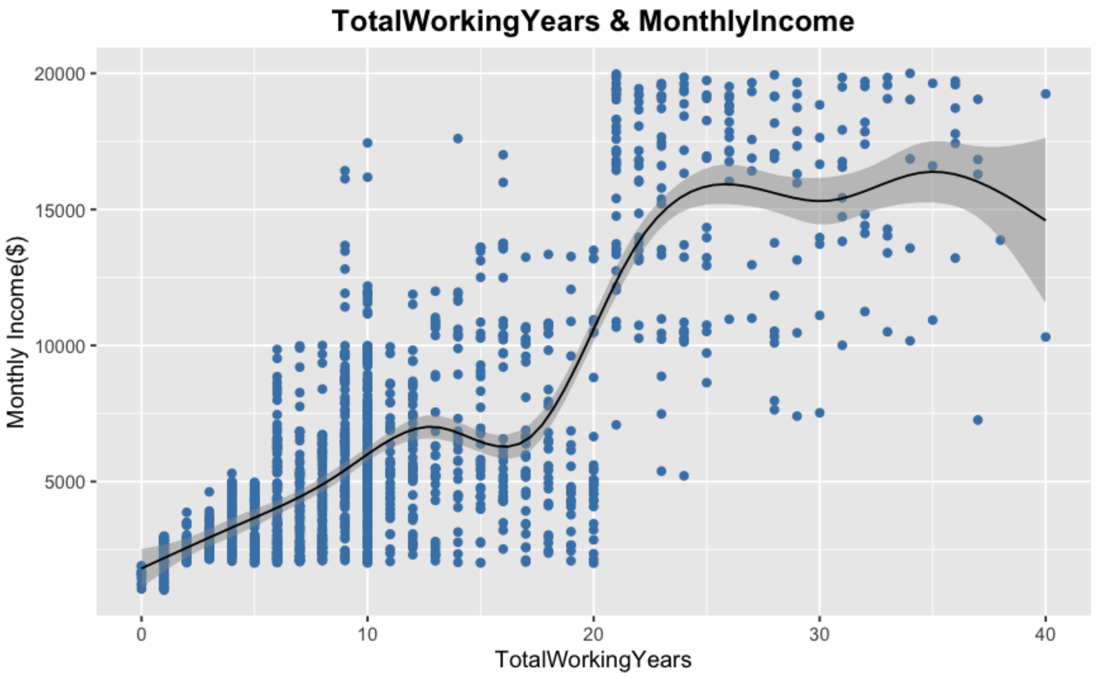

Figure 10: Relationship between TotalWorkingYears and Monthly Income

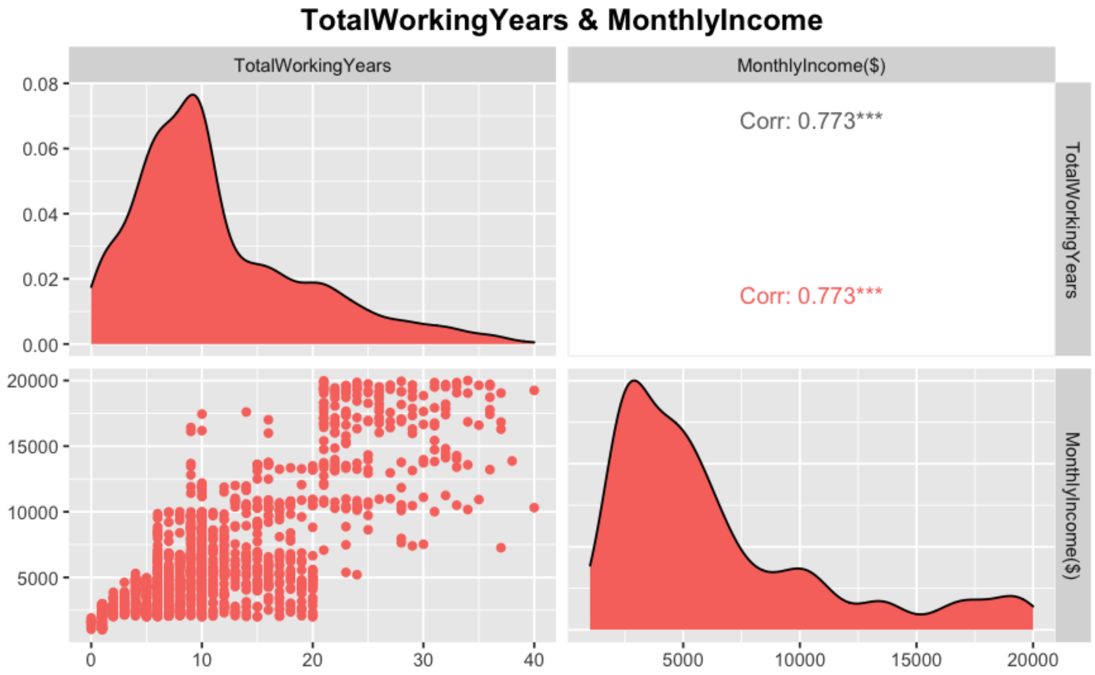

Figure 11: Relationship between TotalWorkingYears and Monthly Income(statistical view)

Visualisation Strategies

The scatter plot is introduced to illustrate the relationship between two numeric variables(Hessing, 2014). In addition, a smoothed conditional regression line is added to represent the possible relationship(ggplot2.tidyverse.org, n.d.). A statistical graph has been created by "ggpairs" to get a deeper insight, such as the correlation value and distribution.

6. Reflection

In summary, all the questions have been answered successfully by data visualisation. More insights from the visualisation also been discussed below each question. Additionally, visualisation strategies have been explained for each question. The visualisation of each question may not be the best, there could exist alternative ways, but it should be the proper one.

Considering the expressiveness and effectiveness, the visualisation implement in this report followed the methodology from Chart Suggestions (Abela, 2009). Several adjustments of the axis, title, legend, shape and colour were applied to the plot. At last, a user-friendly and persuasive graph has been generated.

The interactive graph for each question contains more information than the static graph without directly showing information on it. Besides, the interactive graph is more flexible as it can directly manipulate the graph for different purposes in real-time. It is the trend of data visualisation.

References

Abela, A.V. (2009). *Chart Suggestions-A Thought-Starter*. [online] . Available at: https://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf [Accessed 14 Apr. 2021].

Ajibade, S.S. and Adediran, A. (2016). An Overviewof Big Data Visualization Techniquesin Data Mining. *International Journal of Computer Science and Information Technology Research*, 4(3), pp.105–113.

datavizcatalogue.com. (n.d.). *Density Plot - Learn about this chart and tools to create it*. [online] Available at: https://datavizcatalogue.com/methods/density_plot.html#:~:text=A%20Density%20Plot%20visualises%20the [Accessed 20 Apr. 2021].

Galarnyk, M. (2018). *Understanding Boxplots*. [online] Medium. Available at: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51 [Accessed 21 Apr. 2021].

ggplot2.tidyverse.org. (n.d.). *Smoothed conditional means — geom_smooth*. [online] Available at: https://ggplot2.tidyverse.org/reference/geom_smooth.html [Accessed 20 Apr. 2021].

Hessing, T. (2014). *Scatter Diagrams (Plots), Analysis & Regression*. [online] Six Sigma Study Guide. Available at: https://sixsigmastudyguide.com/scatter-analysis-regression/ [Accessed 22 Apr. 2021].

Klipfolio (2019). *What is Data Visualization? Definitions, Graph Types and How to Use Them*. [online] Klipfolio.com. Available at: https://www.klipfolio.com/resources/articles/what-is-data-visualization [Accessed 16 Apr. 2021].

Rudy, B. (2019). *Defining the Difference Between Average and Median Salary*. [online] Salary.com. Available at: https://www.salary.com/blog/defining-the-difference-between-average-and-median-salary/ [Accessed 23 Apr. 2021].