DWDM Mini-Project: Prediction of Employee Attrition

About Dataset:

The Employee Attrition Dataset consists of following information for 1500 employees (column names):

EmployeeNumber, Age, BusinessTravel, Department, EducationLevel, EducationField, Gender, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, NumCompaniesWorked, OverTime, PercentSalaryHike, PerformanceRating, TotalWorkingYears, YearsAtCompany, YearsSinceLastPromotion, Attrition

Source: kaggle

Preprocessing:

1.Reduced unnecessary columns

2.For data cleaning

- Checked for null values
- Checked for duplicate values
- Checked for outliers
- Plotted heatmap and other graphs

Libraries used:

1. pandas (For manipulating numerical data)
2. Sklearn (For various classification algorithms)
3. Matplotlib (Plotting graphs)
4. Seaborn (For statistical graphs)
5. IPython (To display the decision tree)
6. Sklearn.preprocessing (For label encoder)

Association Rule Mining:

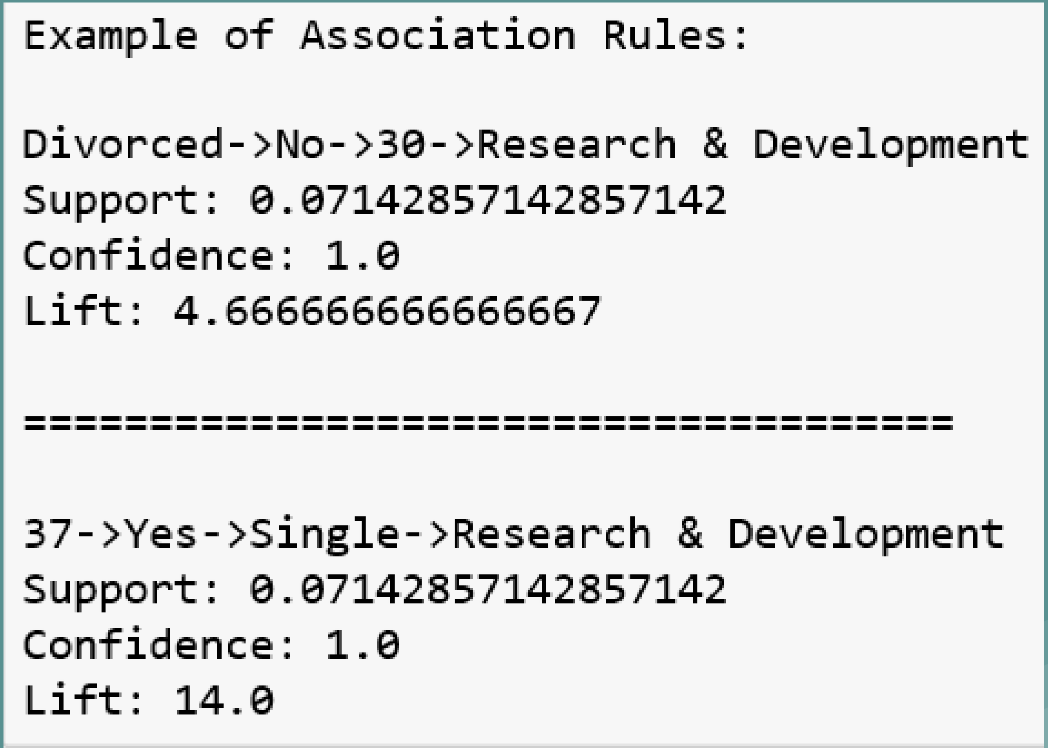
Apriori Algorithm using Apyori Library in Python.

A consequent is an item found in combination with the antecedent.

Association rules are created by searching data for frequent if-then patterns and using the criteria support and confidence to identify the most important relationships.

Metrics:

Support is an indication of how frequently the items appear in the data. Confidence indicates the number of times the if-then statements are found true. A Lift can be used to compare confidence with expected confidence.



```
Example of Association Rules:

Divorced->No->30->Research & Development
Support: 0.07142857142857142
Confidence: 1.0
Lift: 4.666666666666667

=====================================

37->Yes->Single->Research & Development
Support: 0.07142857142857142
Confidence: 1.0
Lift: 14.0
```

Data Mining Algorithm: Decision Tree Classification

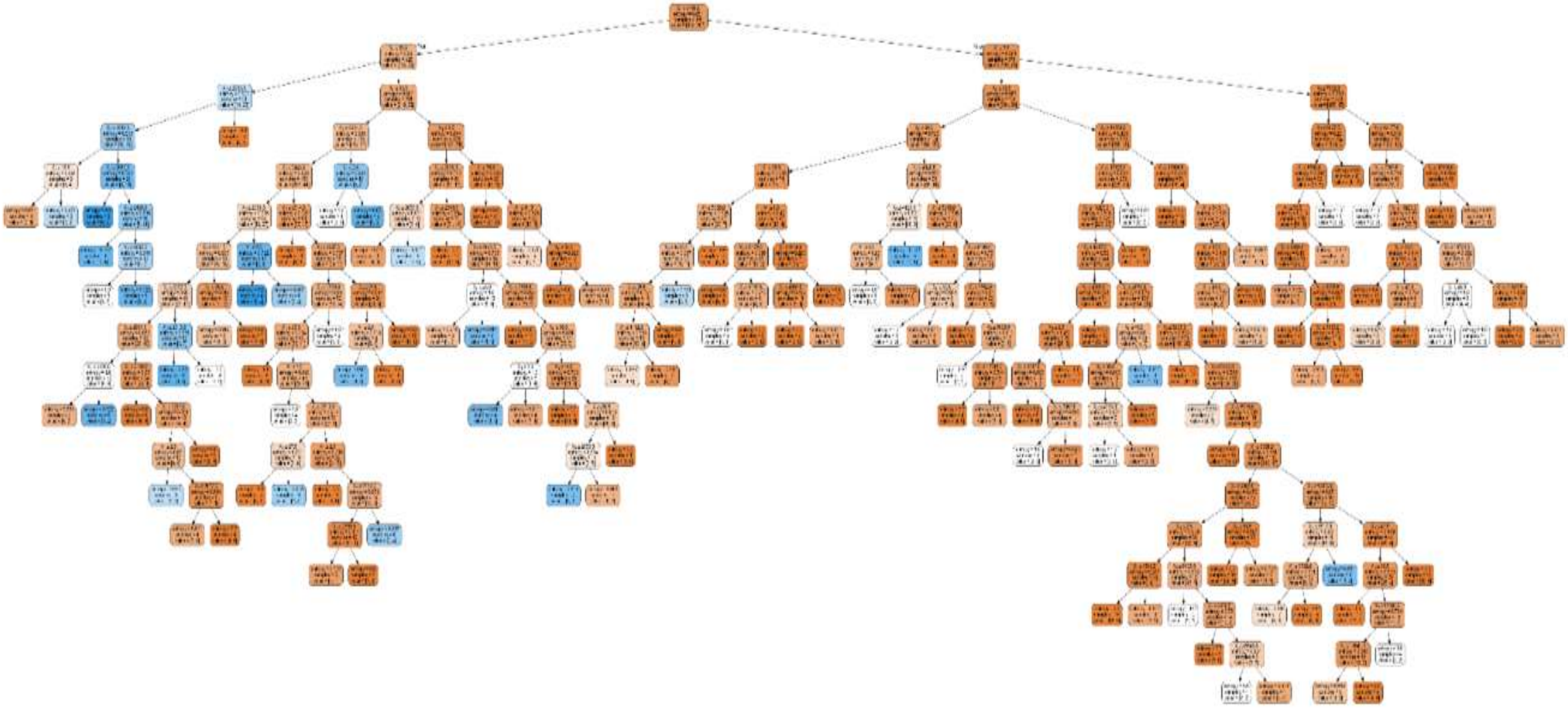
Decision tree builds classification models in the form of a tree structure.

It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node has two or more branches. Leaf node represents a classification or decision.

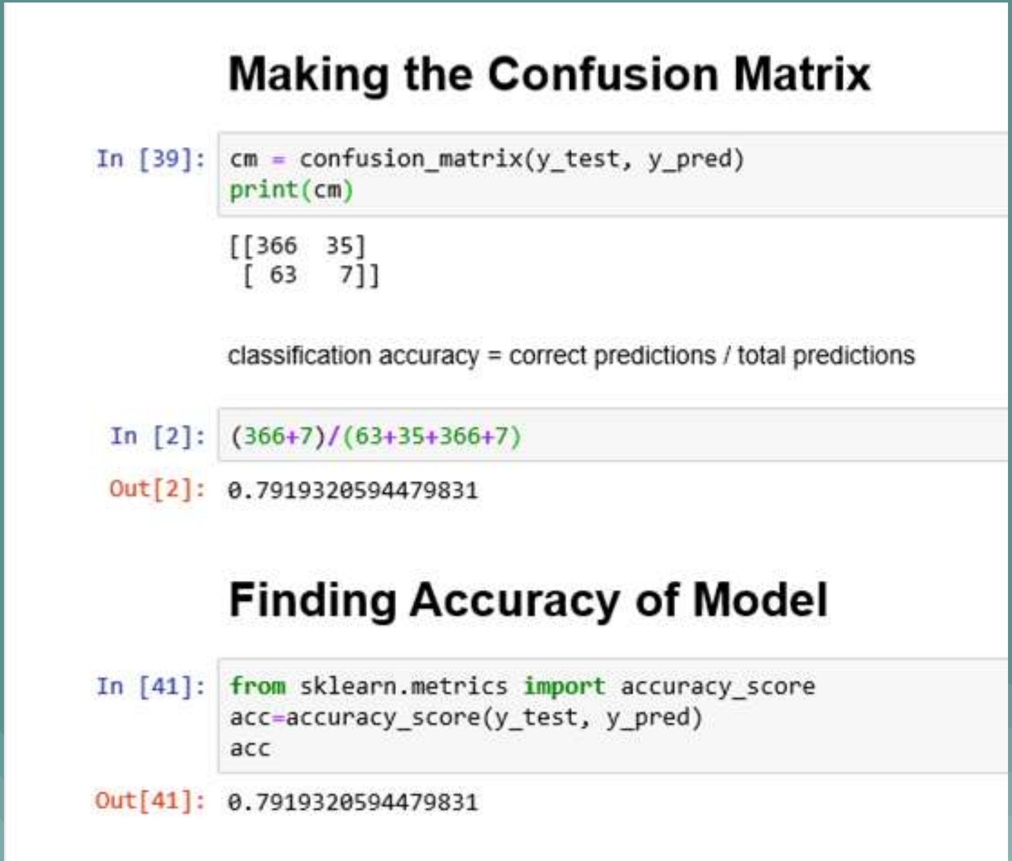
The topmost decision node in a tree which corresponds to the best predictor called **root node**.

Decision trees can handle both categorical and numerical data.

Decision Tree:



Result:



Making the Confusion Matrix

```
In [39]: cm = confusion_matrix(y_test, y_pred)
         print(cm)

         [[366  35]
          [ 63   7]]
```

classification accuracy = correct predictions / total predictions

```
In [2]: (366+7)/(63+35+366+7)

Out[2]: 0.7919320594479831
```

Finding Accuracy of Model

```
In [41]: from sklearn.metrics import accuracy_score
         acc=accuracy_score(y_test, y_pred)
         acc

Out[41]: 0.7919320594479831
```

Our Classification model gave an accuracy of 79.2%.
We have used Confusion Matrix to calculate the accuracy, as well as accuracy function from sklearn.metrics library.

Ways to improve accuracy:

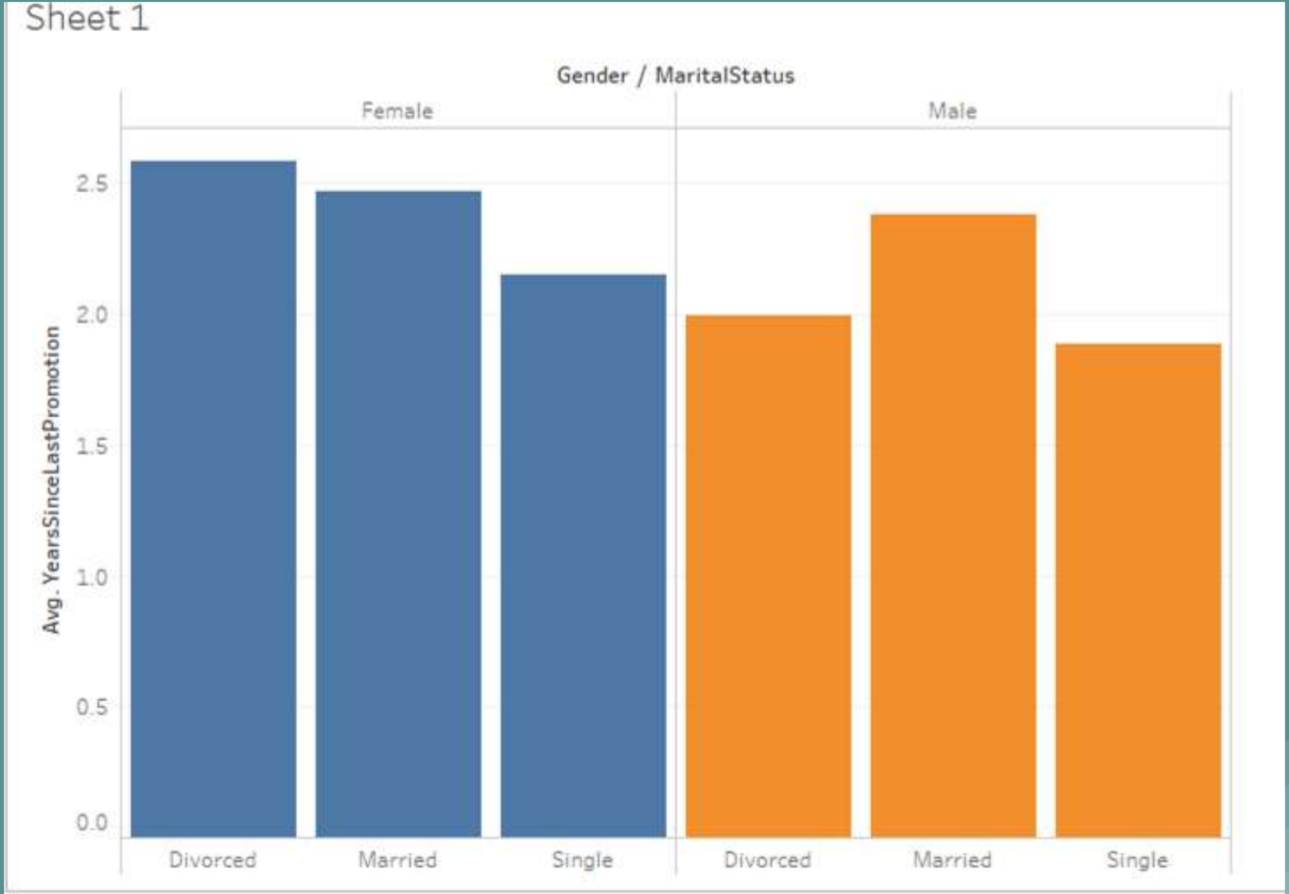
1. Add more data.
2. Treat missing and Outlier values.
3. Feature Engineering.
4. Feature Selection.
5. Multiple algorithms.
6. Algorithm Tuning.
7. Cross Validation.

Visualization using Tableau

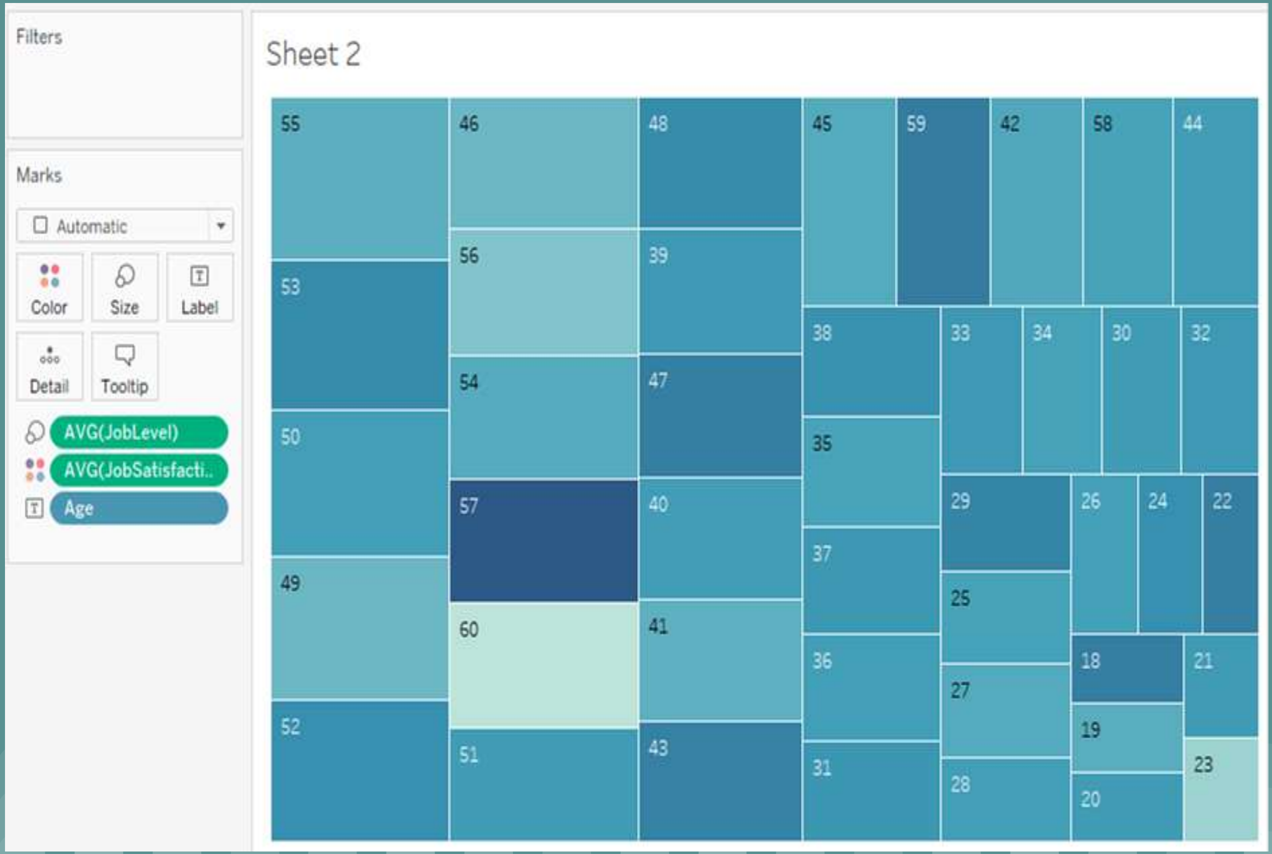
Bar Graph:

Depending on their marital status, the average of their years since last promotion are shown.

For example, we can observe that, divorced females have had more years since last promotion as compared to single women.



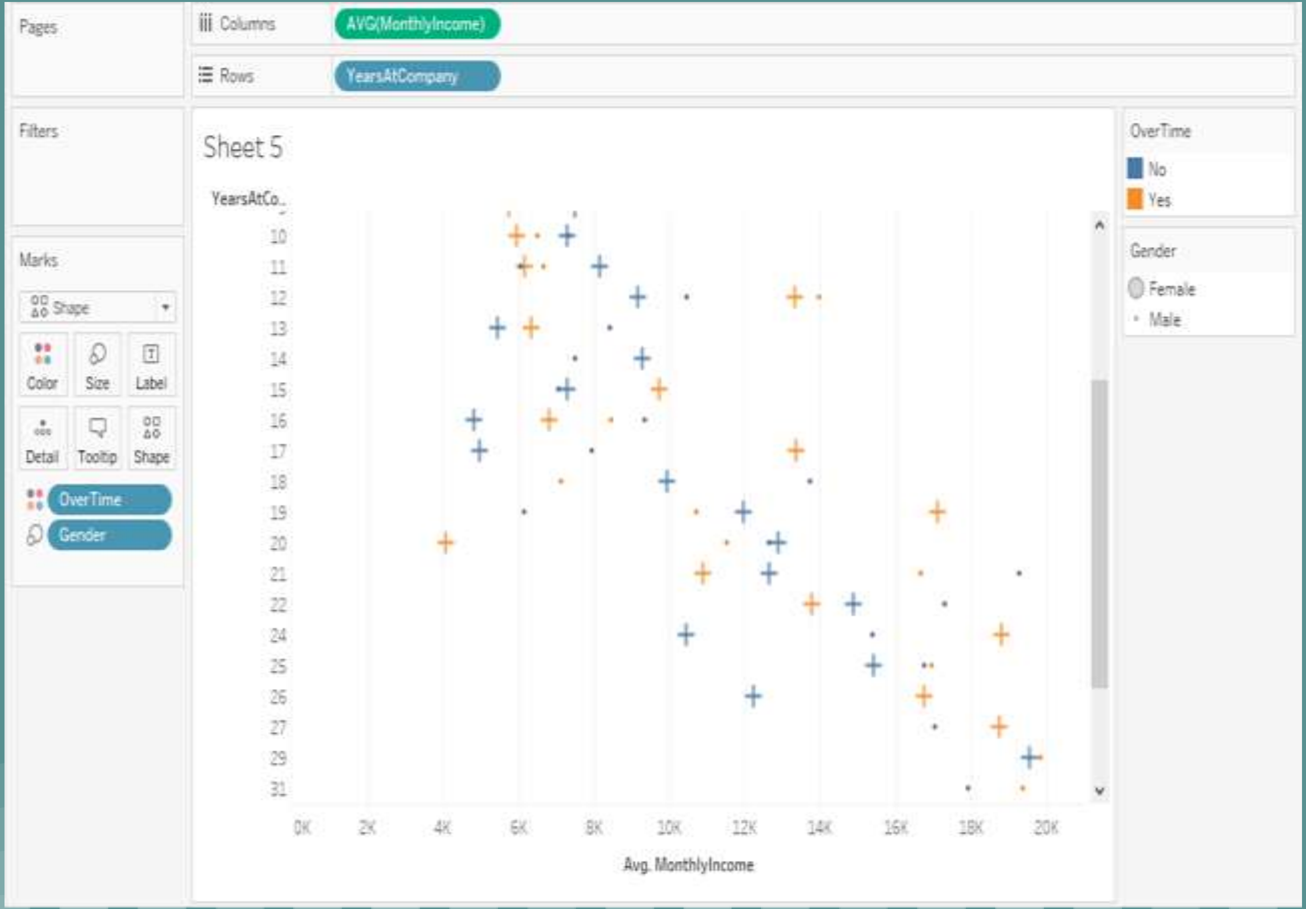
Heat Map:



The number present in each box represents the age of the employee, the size of a box shows the average job level and the shade of blue represents the average of job satisfaction (dark denotes most satisfied).

For example, we can observe that at the age of 57 the job level is substantial and the job satisfaction level is high.

Scatter Plot:



Columns AVG(MonthlyIncome)

Rows YearsAtCompany

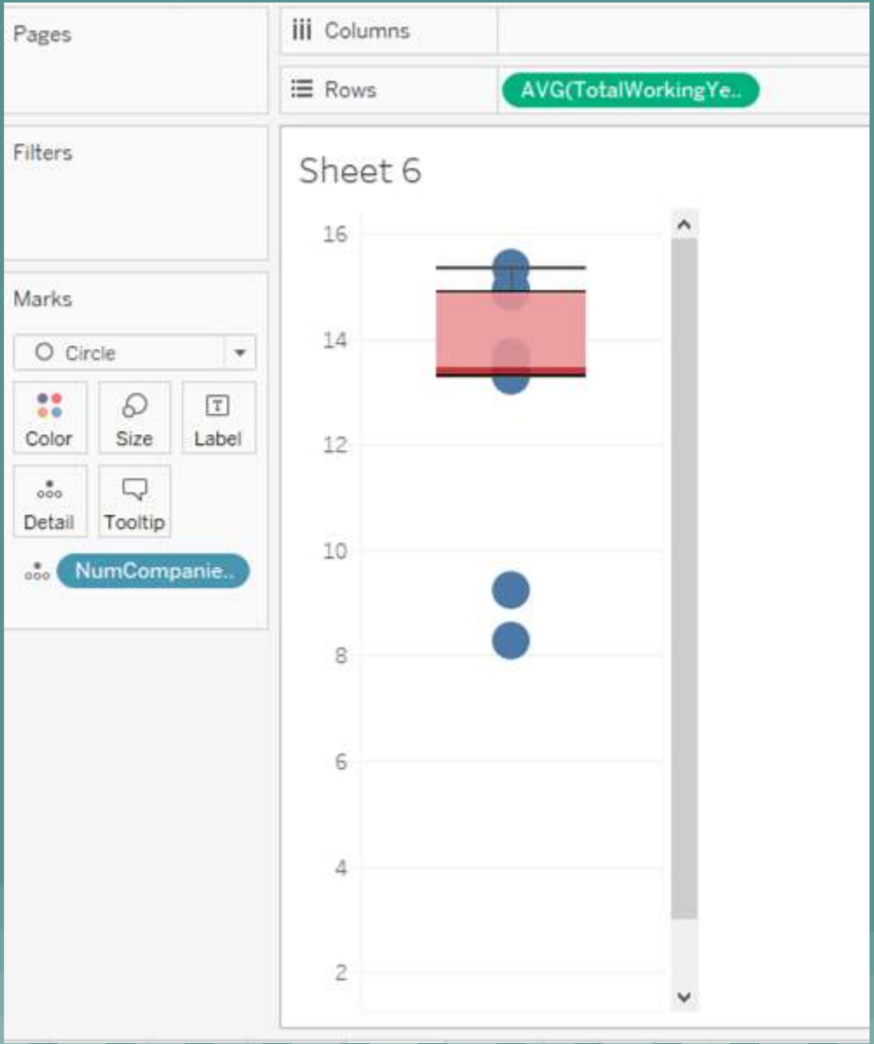
In this data visualisation, blue colour represents no overtime and orange shows over time.

The y-axis shows the number of years the employee has worked at the company. The + sign shows females and the dots represent males. Here we can see that, men do more overtime than women.

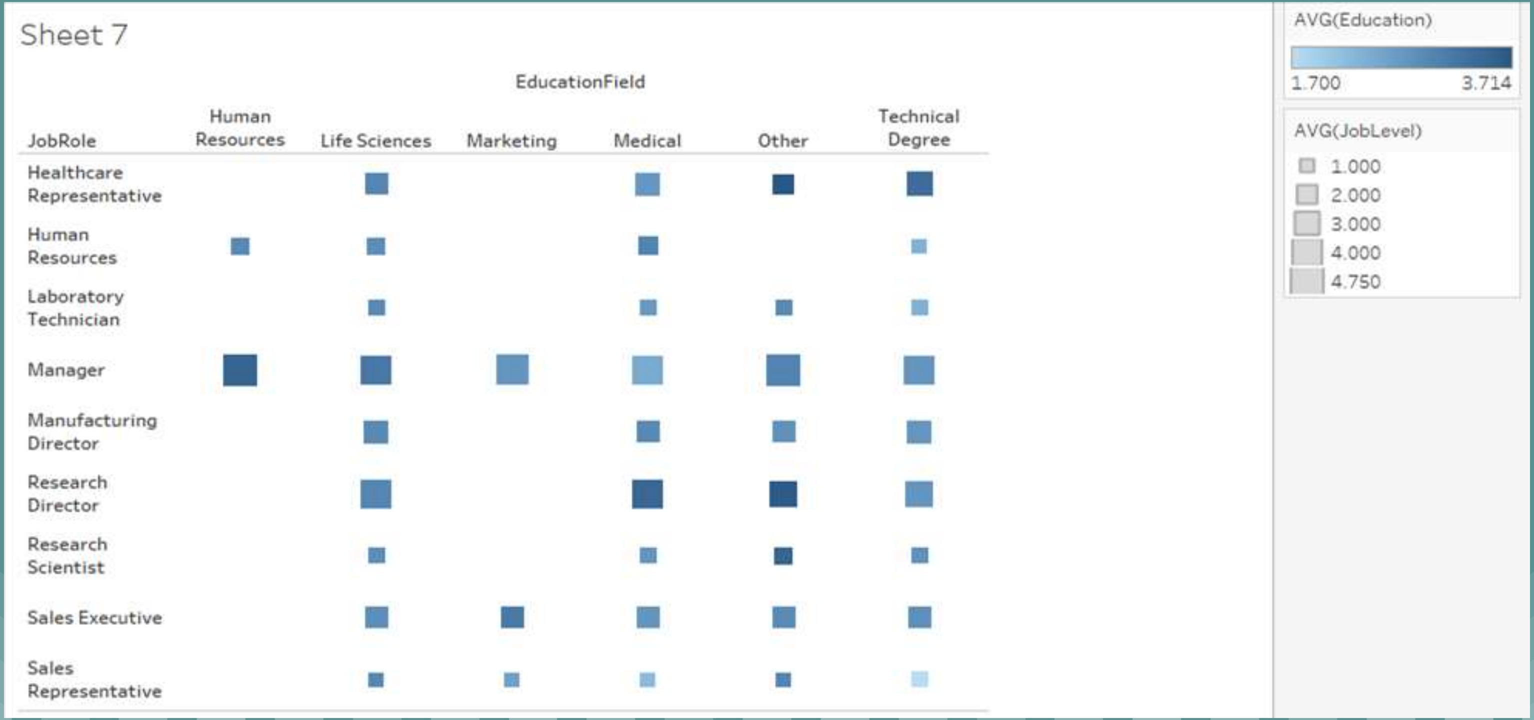
Box Plot:

This data visualisation shows the outliers.
Y axis represents total number of working years.
Each point shows Number of Companies the employee worked before at.

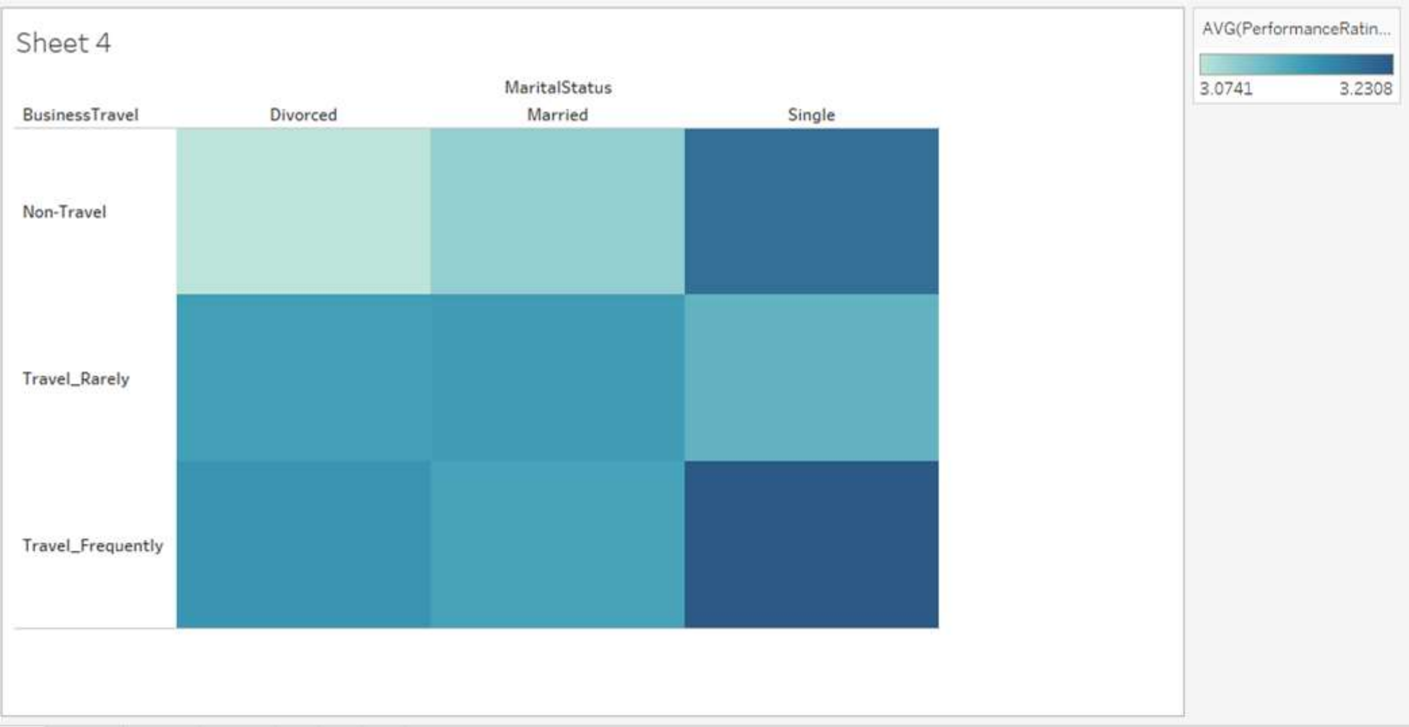
Here, one can observe that there are two outliers with this being their first and second company respectively and have spent many years there.



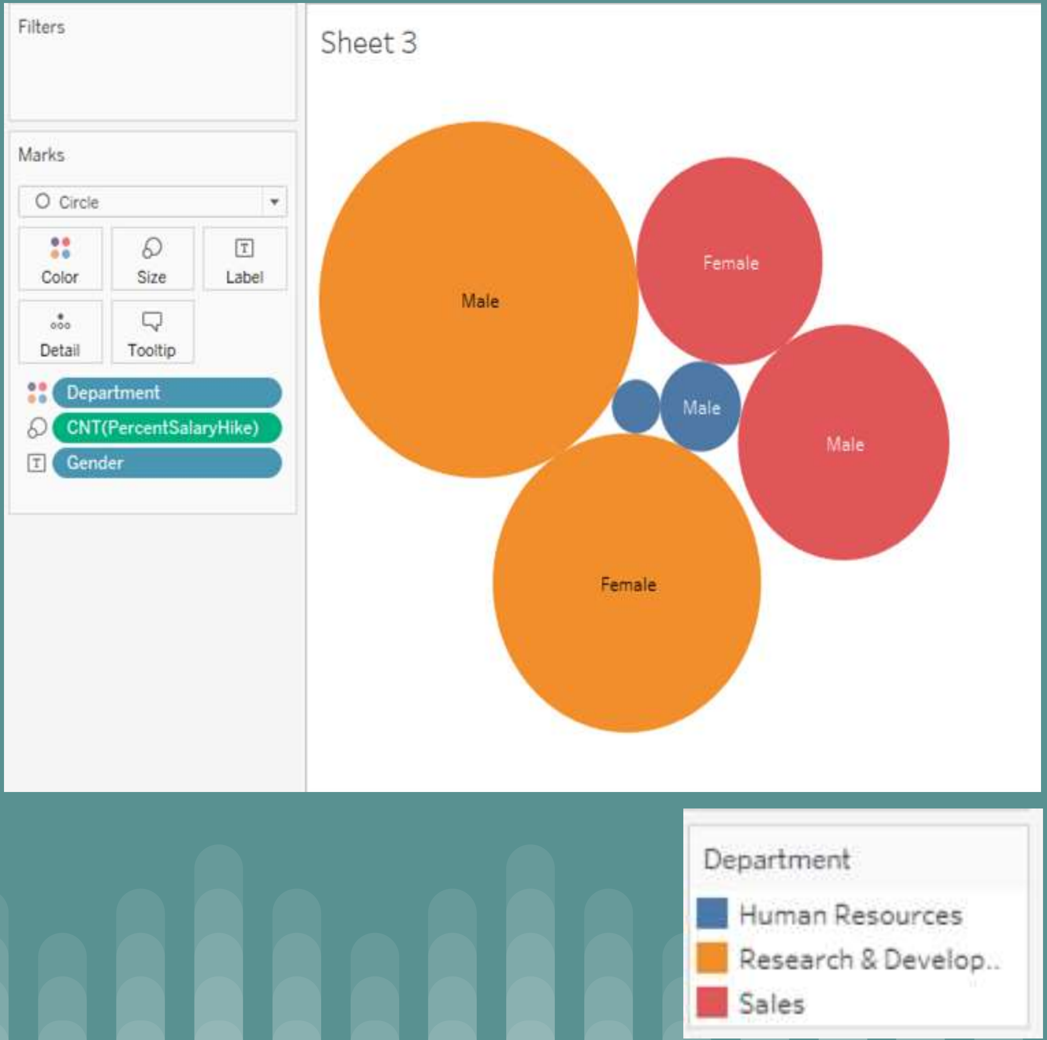
This data visualization shows, various job roles as the different rows. The field of education of employees is shown as columns. The size of the blue boxes shows the Job Level. The shade of blue shows level of education. The deeper the shade of blue, the more the level. For example, an employee who has an education in Human resources is a manager, at a high level of job because of high education level.



In this heat map we can see business travel frequency along the y-axis and the marital status of the employee along the x-axis. The shade of blue represents the performance rating. The deeper the shade of blue, the higher the performance rating. One can observe that, single employee who travel frequently have a high performance rating also.



In this data visualization, three different departments have been depicted using three different colours. The two circles of the same circle, show male and female employees separately. The size of the circle, shows the percent salary hike. For example, one can observe that the men in the Research and development department have maximum salary hike, whereas the women employees in Human resources have the least hike.



Thank You