Table of Contents

A. IBM HR Attrition Analysis

A.1 Objective of case study – Employee attrition

Employee attrition can be defined as the employee leaves the company by personal reasons or retirement. The objective aims to find the important factors that will lead to employees leaving the company. This would allow the HR team of the company to come up with measures to prevent high attrition rates.

A.2 Dataset

The dataset consists of details of employees that is working in this company of the case study, IBM. Such details includes, age, job role and level, salary, total working hours and years working in the company or with their manager. There are a total of 1470 rows and 35 columns in this dataset. The dataset also does not have any missing values, with all rows and columns having values in them, therefore there isn't a need to perform data preprocessing on any missing values. This shows that this dataset have a good quality of data.

```
Missing values in the entire dataframe

Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EmployeeNumber              0
EnvironmentSatisfaction     0
Gender                      0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
JobRole                     0
JobSatisfaction             0
MaritalStatus               0
MonthlyIncome               0
MonthlyRate                 0
NumCompaniesWorked          0
Over18                      0
OverTime                    0
PercentSalaryHike           0
PerformanceRating           0
RelationshipSatisfaction    0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           0
TrainingTimesLastYear       0
WorkLifeBalance             0
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

Figure 1

A.3 Exploratory Data Analysis

Before performing model building and testing, data analysis on the given dataset is required. This will highlight any trend or patterns for us to be able to decide on the features to use for machine learning. Firstly, each of the columns with numerical values in the dataset was plotted on a histogram (figure 2) to analyse the distribution of the values of each category. The object datatype columns are analyse using a bar graph, as seen in figure 3.

Figure 2

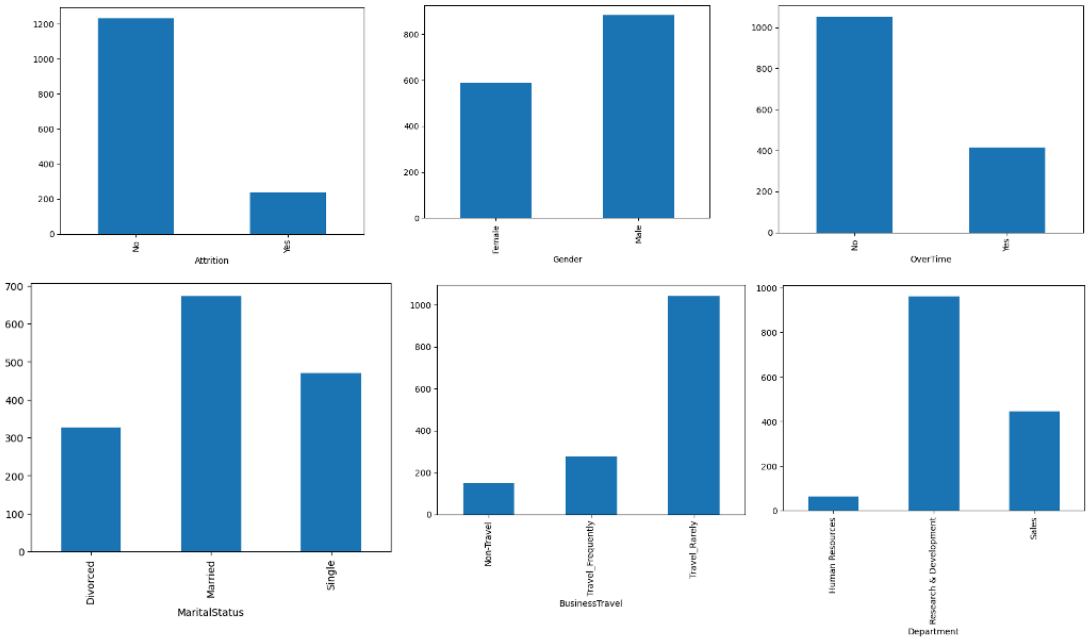


Figure 3

Figure 2 shows that some columns have only 1 value and these value is the constant on all the rows, such as employee count, Over18, Standard hours. Therefore, I have chosen to opt these columns out by using the .dropna() function on python.

From these analysis, I was able to pick out some factors that could lead to employees attrition. Some factors includes overtime, frequency of business travels, marital status, age, years in current role and total working hours. These factors are then compared against attrition rates so that the most influential feature that causes attrition can be picked out.

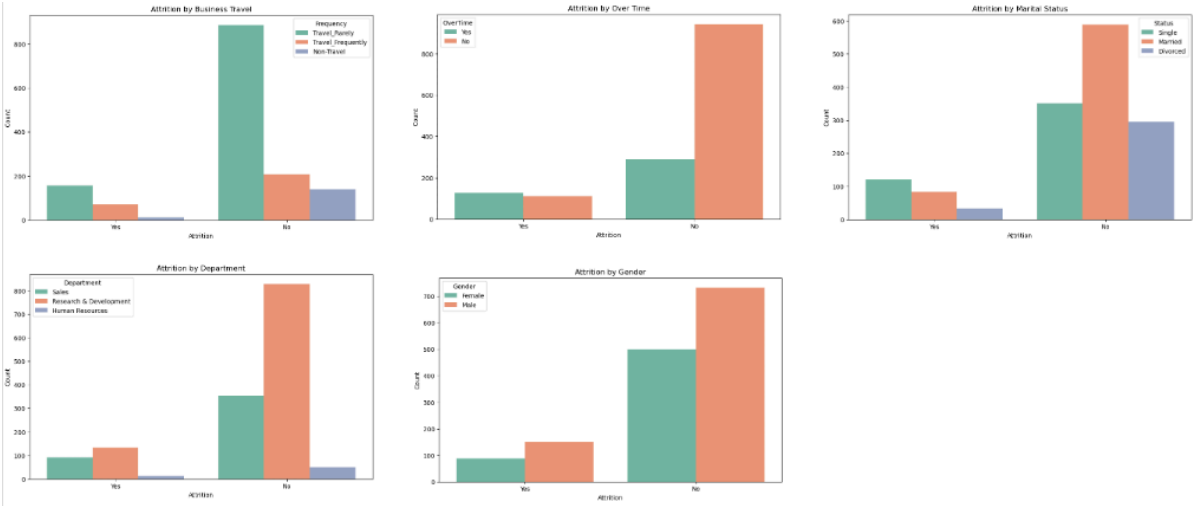


Figure 4

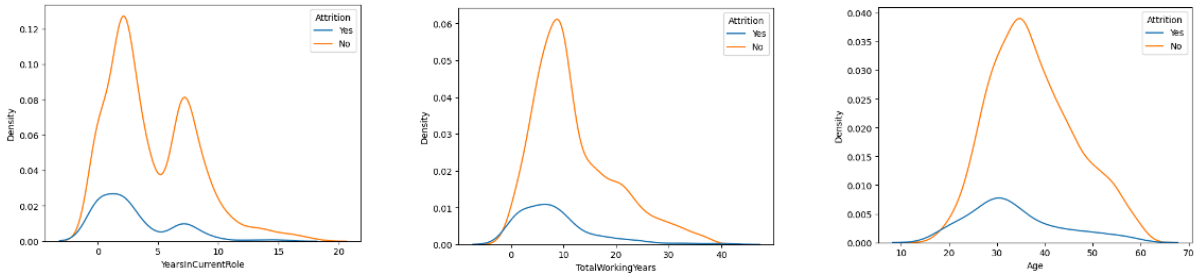


Figure 5

Some trends and pattern can be picked out from analysing figure 4 and 5 graphs. There is a higher attrition counts when the employees are single, have to work overtime or are on regular business travels. It also shows that the attrition happens the most on employees around the age of 25 – 40 with the highest count being employees at 30. The total years working years and years in current role are also factors on attrition. Employees with more than 10 years of working tend to not leave the company. It is also observed that employees leaves the company when they have less than 5 years in current role and around 7 – 8 years in the current role. The former suggests that the employees, after gaining experience for a few years in the company, are looking elsewhere for a promotion or change in role. The latter suggest that employees leaves the company because they are stuck in the same role in the same role for many year.

A.4 Correlation Matrix

After performing EDA on the dataset, a correlation matrix was plotted to observe the correlation between the columns, in the case we are seeing the correlation of attrition to other columns. However, more columns, such as daily, hourly and monthly rate and Employee number got to be drop first. It is then transform using the label encoder to encode the values of columns with a string datatype into a integer. The first correlation matrix can be observed from figure 6.

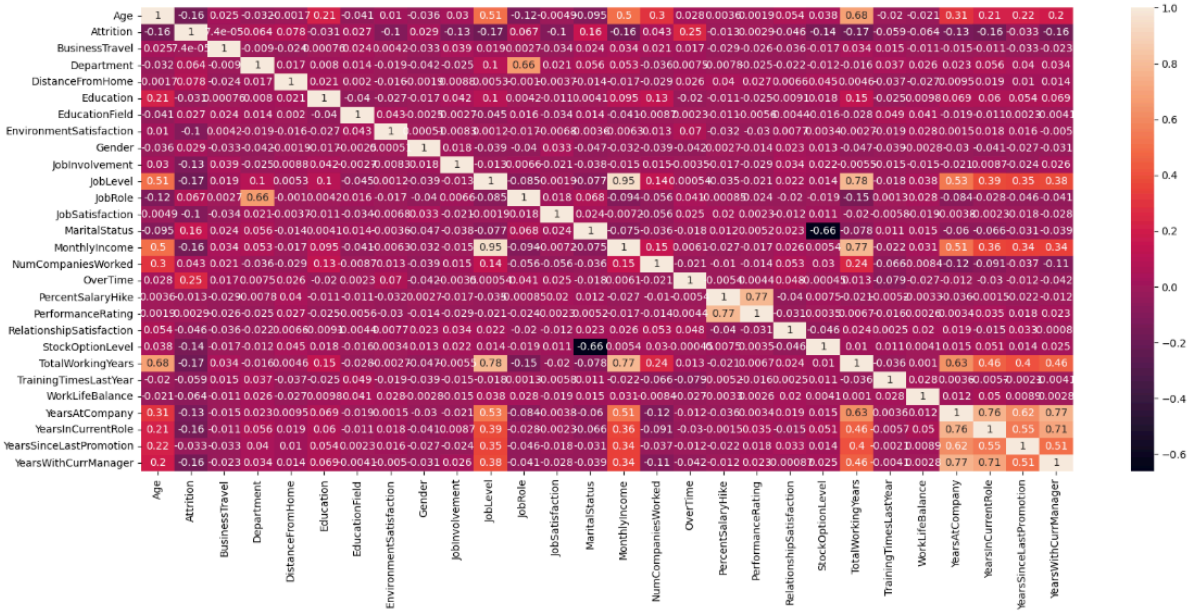


Figure 6

From the correlation matrix, there can be a further reduction of columns. The columns that are between -0.1 and 0.1 in relation to attrition are dropped as it shows that it is either a weak or no correlation to attrition. The business travel, which was first thought as a factor to attrition, was also dropped as it shows a very weak correlation to attrition with only 7.4 e-0.5. Thus, a new correlation matrix was plotted as seen in figure 7.

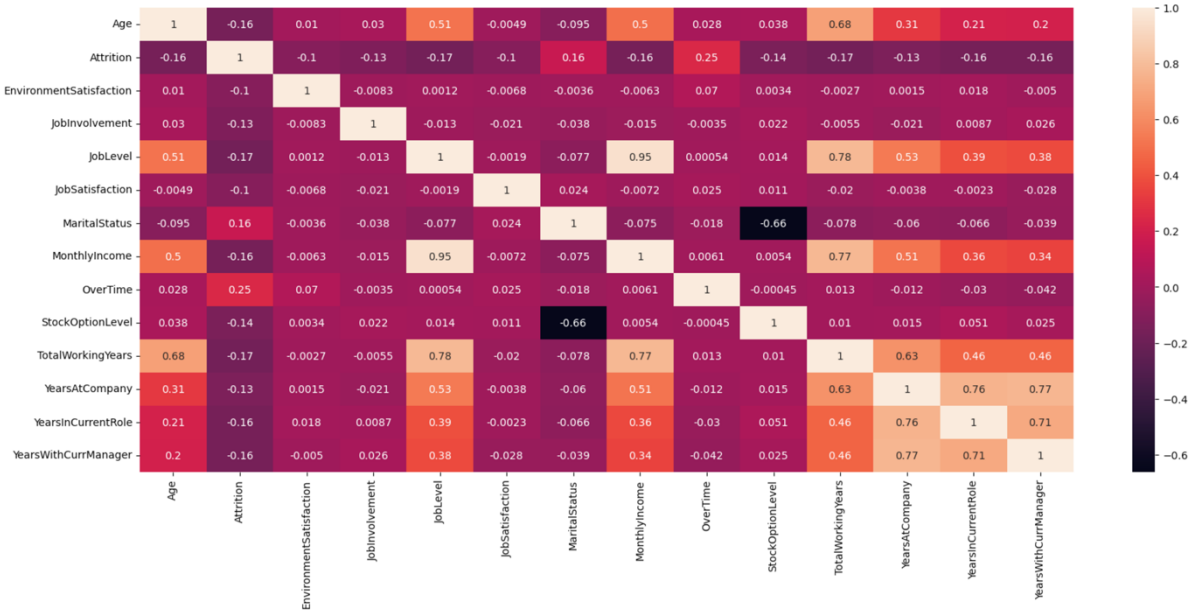


Figure 7

From the EDA and correlation matrix, I was able to select the feature that could possible cause attrition. This feature is overtime. Although only having 0.25 of correlation to attrition, which is considered weak, it was the highest compared to the other columns. This means that when an employee got to work overtime, it is likely that the employee will leave the company.

A.5 Classification Model

Classification is one method of supervised learning to provide predictions through algorithms. In this case, model is created to predict if the employee is likely to quit the company. The 3 different classification techniques used were K-Nearest Neighbour (KNN), Decision Tree and Random Forest.

A.5.1 Handling Class Imbalance

Before building the models, there is a need to first handle any class imbalance. For this dataset, there is a class imbalance where there are more 'No' than 'Yes'. Thus, there's a need to further overpopulate more 'Yes' into the dataset so that there is sufficient amount for training and testing on the models build. Synthetic minority Oversampling Technique (SMOTE) was used. After performing SMOTE, there should be the same amount of count, as seen in figure 8.

```
Before OverSampling, counts of label '1': 190
Before OverSampling, counts of label '0': 986

After OverSampling, counts of label '1': 986
After OverSampling, counts of label '0': 986
```

Figure 8

A.5.2 Model Evaluation

After building the models, it got to be evaluated to see which model gives the best and accurate results so that it can be used. For all 3 models the dataset was split into 80% train and 20% test.

The models are evaluated by:

- Accuracy: Ratio of correctly predicted data to the dataset
- Precision: Shows how much positively predicted data is actually positive
- Recall: Shows how much data should be predicted positive and is predicted positively
- F1 score: The harmonic between precision and recall

The overall results can be seen from the figures 9 to 12. It shows that the random forest model is the most suitable to be used to predict the likelihood of employee attrition as it gives an accuracy score of 82%. Based on the confusion matrix, the random forest model also gave the least total count of false positive and negative.

```
The KNN model accuracy score is: 62.585034013605444 %
The Decision Tree model accuracy score is: 71.08843537414967 %
The Randam Forest Model accuracy score is: 81.63265306122449 %
```

Figure 9

	precision	recall	f1-score	support
0	0.86	0.66	0.75	247
1	0.19	0.43	0.27	47
accuracy			0.63	294
macro avg	0.53	0.54	0.51	294
weighted avg	0.75	0.63	0.67	294

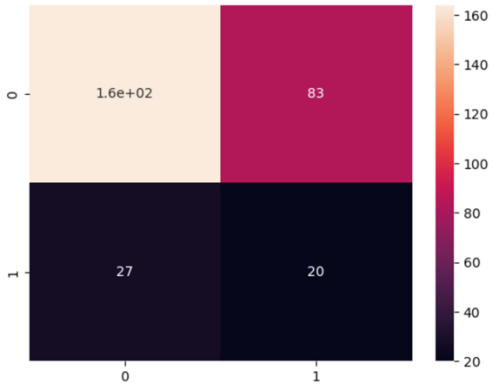


Figure 10. Classification report and Confusion matrix for KNN model

	precision	recall	f1-score	support
0	0.88	0.76	0.82	247
1	0.26	0.45	0.33	47
accuracy			0.71	294
macro avg	0.57	0.60	0.57	294
weighted avg	0.78	0.71	0.74	294

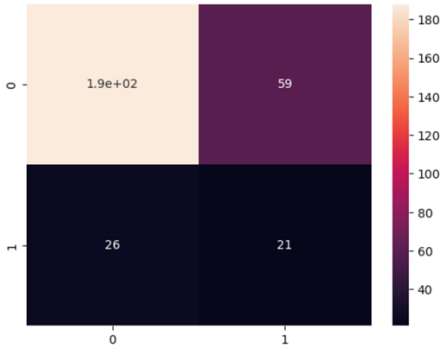


Figure 11. Classification report and Confusion matrix for Decision Tree model

	precision	recall	f1-score	support
0	0.88	0.90	0.89	247
1	0.41	0.36	0.39	47
accuracy			0.82	294
macro avg	0.65	0.63	0.64	294
weighted avg	0.81	0.82	0.81	294

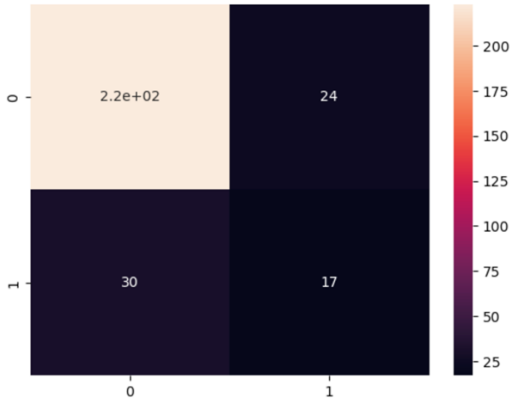


Figure 12. Classification report and Confusion matrix for Random Forrest model

A.6 Clustering Model

Clustering is an unsupervised learning method to give predictions through algorithms from the various model. The 2 clustering model that I have used are hierarchical clustering and DBSCAN clustering.

A.6.1 Hierarchical Clustering

Hierarchical Clustering is an algorithm to cluster that group of data into a tree of nested clusters. To find the number of cluster for the algorithm, a dendrogram was plotted using the ward method to show the total number of clusters form (figure 13). From the dendrogram, it shows that there are 2 clusters for the attrition caused by overtime.

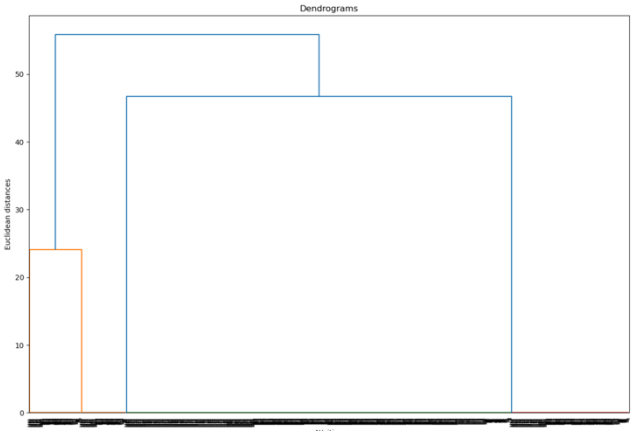


Figure 13

After finding the number of clusters, it can be proceed to initiating the model, followed by fitting and predicting the model. The prediction can then be visualise using scatterplot, which shows each clusters as seen from figure 14.

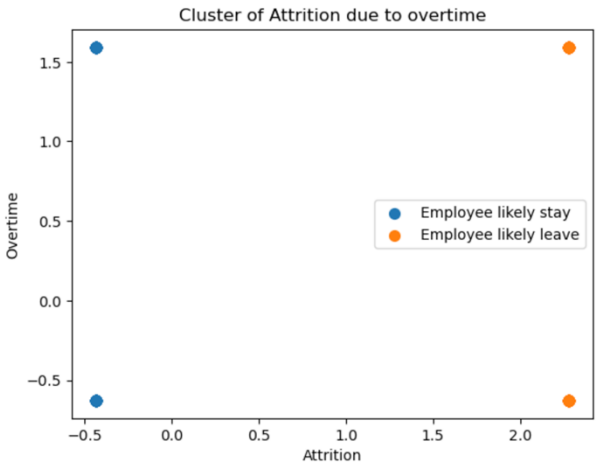
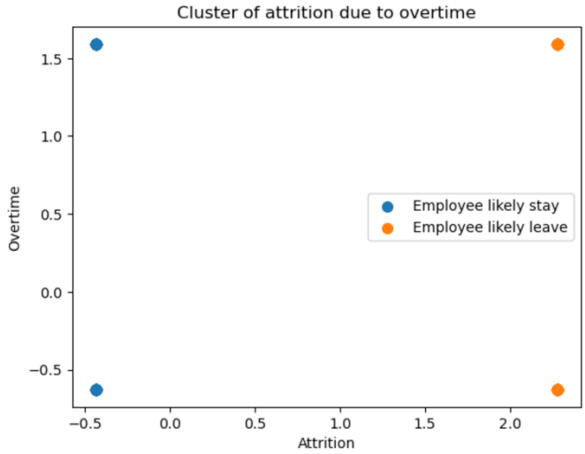


Figure 14

A.6.2 DBSCAN Clustering

DBSCAN which meant at density-based spatial clustering of applications with noise, is another clustering algorithm that can be used. This algorithm will partition data into clusters based on their distance to other points. This method can be useful in removing noise from the data set. However, in this IBM Attrition Dataset, there isn't any noise to be removed.

Firstly, dbscan cluster was applied. It was then followed by fitting the model for prediction and importing it back into the dataset. Evaluation can be done by printing the number of cluster groups and the number of instances in each cluster group. From these cluster groups, it can be then plotted into a scatterplot to visualise the prediction. The results can be seen from figure 15.




```
cluster group
1    944
2    289
0    127
3    110
Name: count, dtype: int64
```

Figure 15

A.6.3 Evaluation of clustering results

From the results, these 2 clustering model could not be used for the specific feature that I have chosen. As the datatype of the 2 columns, Attrition and Overtime, are binary. Thus, when plotted in a scatterplot, it will only show 4 distinct dots, as the data are clustered in the same place. However, this clustering algorithm technique is still possible to be performed on other columns other 'Overtime'.

B. Regression model on Insurance Charges

B.1 Objective of this project

The objective is to create a regression model and to come up with an objective for based on the models created on the selected dataset.

B.2 Dataset

The dataset was taken from the Kaggle website, which consists of 1338 rows and 7 columns. There are no missing data in this dataset, thus no missing values data preprocessing were performed on the dataset.

```
Dimension of the data: (1338, 7)

 ------------------------------------------------------

Summary of the data

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
None

 ------------------------------------------------------

Missing values in the entire dataframe

age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64

 ------------------------------------------------------
```

Figure 16

B.3 Exploratory Data Analysis (EDA)

Some EDA can be performed to help picked the right columns to use in our model building. A correlation matrix was plotted to observe any suitable correlation can be used for model building and objective.

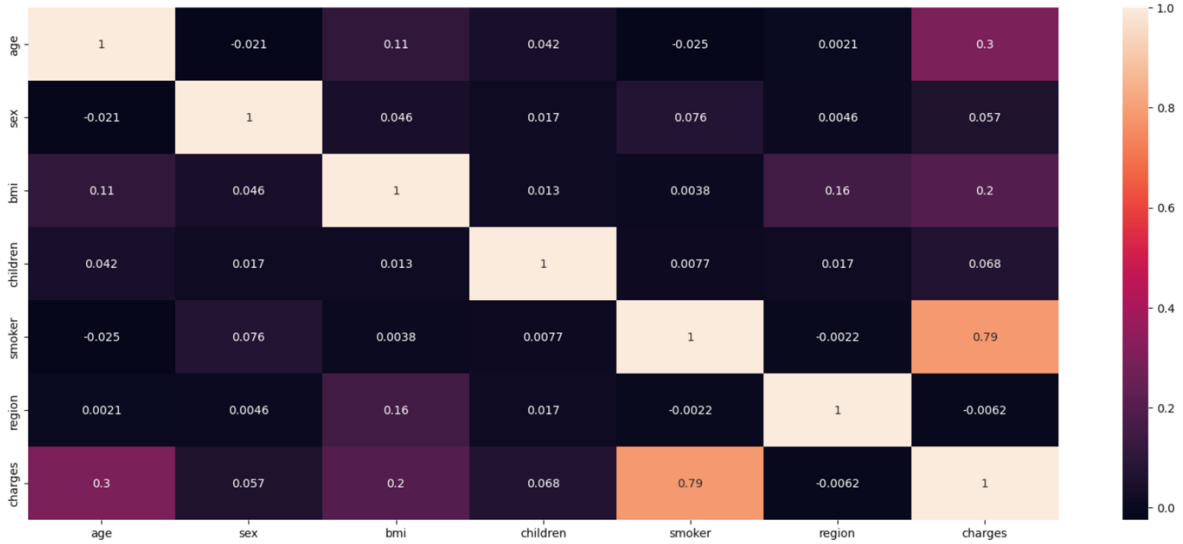


Figure 17

From figure 17, there is a high correlation between smoker and charges of 0.79. This allowed me to use these features in model building.

B.4 Model Building and Evaluation

Regression model is another type of supervised learning in predicting the objective. From the correlation matrix, I have decided that the objective is to predict the insurance charges based on certain factors. There 2 regression used were single linear regression and multiple linear regression.

B.4.1 Single Linear Regression (SLR)

This regression method compares the charges based on a certain factor. As the 'Smoker' column have the highest correlation to charges, I have picked it as the feature in this SLR model. With X = whether they are smokers and Y = charges. Figure 18 shows the regression on a scatterplot, however it could be properly observed as smoker column is a binary data type, there are only 2 answers for it, 'No' = 0 and 'Yes' = 1. This regression model also gives a 67% score (Figure 19).

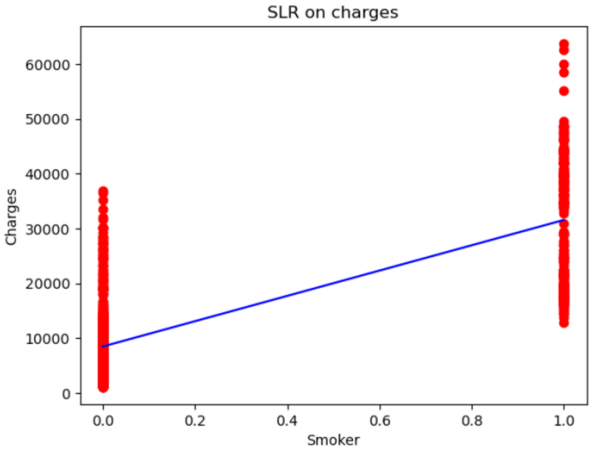


Figure 18

```
SLR R2 score: 67.34292727177755 %
```

Figure 19

B.4.2 Multiple Linear Regression (MLR)

Another regression method used was the MLR model. This predicts the charges based on multiple factors of the dataset. Before the model was build and evaluated, I removed some columns that have a weak correlation to charges so as to allow the model to properly predict the objective. This model gives a score of 79% as seen from figure 20.

```
MLR R2 score: 79.45500805653087 %
```

Figure 20

B.4.3 Model Comparison

Comparing both SLR and MLR model, the MLR is a better algorithm to use in predicting the insurance charges for each person. This is because it gives a better R2 score of 79% as compared to 67% from the SLR model. The could properly give a prediction of the charges because the MLR compares the charges across different factors whereas SLR only compares to whether the patient is a smoker.

C. Association Model on TV shows

C.1 Objective of this project

The objective of this project is to build an association algorithm model from the dataset given and provide insights on real-world application.

C.2 Dataset

The dataset is taken from the Kaggle website, which consist of 9690 rows and 32 columns. The dataset does not come with a header, therefore in the code, it got to state there are no headers, if not the first row will be used as an header (figure 21 and 22).

	Cobra Kai	Lupin	12 Monkeys	Sherlock	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8
0	Lost	Jack Ryan	The Flash	Game of thrones	House of Cards	12 Monkeys	Vikings	Fringe	The Mentalist
1	Sex Education	Dr. House	Kingdom	The Walking Dead	NaN	NaN	NaN	NaN	NaN
2	Ozark	Sex Education	Constantine	Preacher	Vikings	The Tick	NaN	NaN	NaN
3	Naruto	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Sex Education	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 32 columns

```
df = pd.read_csv('TV Shows.csv')

#confirming data
df.head()
```

Figure 21

	0	1	2	3	4	5	6	7	8
0	Cobra Kai	Lupin	12 Monkeys	Sherlock	NaN	NaN	NaN	NaN	NaN
1	Lost	Jack Ryan	The Flash	Game of thrones	House of Cards	12 Monkeys	Vikings	Fringe	The Mentalist
2	Sex Education	Dr. House	Kingdom	The Walking Dead	NaN	NaN	NaN	NaN	NaN
3	Ozark	Sex Education	Constantine	Preacher	Vikings	The Tick	NaN	NaN	NaN
4	Naruto	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 32 columns

```
df = pd.read_csv('TV Shows.csv', header = None)

#confirming data
df.head()
```

Figure 22

C.3 Data Preprocessing

Firstly, this dataset have to undergo a transformation to turn into the array into a list. After transforming the dataset, we can observe whether code was performed properly as shown in figure 23.

```
records = []
for i in range(0, 9690):
    records.append([str(df.values[i,j]) for j in range(0, 32)])
```

```
records
```

```
[['Cobra Kai',
  'Lupin',
  '12 Monkeys',
  'Sherlock',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
  'nan',
```

Figure 23

C.4 Model Building and Evaluation

C.4.1 Building of model

For association model, there are 2 algorithm that we can use, Apriori or FP-Growth. In this project, I have selected to Apriori algorithm to build on and evaluated.

Some key association rules that I have used in this algorithm can be seen in figure 24.

```
# association rule apriori
association_rules = apriori(records, min_support=0.0003, min_confidence=0.2, min_lift=3, min_length=2, max_length=2)
association_results = list(association_rules)
```

Figure 24

C.4.2 Evaluation the model

After building the model, we can evaluate on the given dataset. It can be evaluated by these 3 factors:

- Support: Shows how important the item in by calculating the ratio of the item appearing to the total number of transactions
- Confidence: Shows how likely the item will come together with another product
- Lift: Show the strength of the association between the 2 items

Figure 25 display the association rules that the algorithm has come up with and showing the support, confidence and lift. From these array of values, a further sort can be performed based on the lift to show which 2 items have the strongest association. As seen in figure 26, the tv show South Park and Berlin Station have a strongest association. Although Lift value for this 2 item is only 9.0, it has a confidence level of 0.5 which makes it to have a higher association. This means that viewers that watches South Park are likely to watch Berlin Station as well.

```
      Title 1     Title 2  Support Confidence     Lift
0  12 Monkeys  Death Note  0.00051     0.3125  5.33120
==========================================
           Title 1     Title 2  Support Confidence     Lift
0       12 Monkeys  Death Note  0.00051     0.3125  5.33120
1  Game of thrones  12 Monkeys  0.00381    0.22023  3.75723
==========================================
           Title 1     Title 2  Support Confidence     Lift
0       12 Monkeys  Death Note  0.00051     0.3125  5.33120
1  Game of thrones  12 Monkeys  0.00381    0.22023  3.75723
2       Inside Job  12 Monkeys  0.00123    0.24489  4.17792
==========================================
           Title 1     Title 2  Support Confidence     Lift
0       12 Monkeys  Death Note  0.00051     0.3125  5.33120
1  Game of thrones  12 Monkeys  0.00381    0.22023  3.75723
2       Inside Job  12 Monkeys  0.00123    0.24489  4.17792
3      Space Force  12 Monkeys  0.00061    0.27272  4.65268
==========================================
```

Figure 25

	Title 1	Title 2	Support	Confidence	Lift
23	South Park	Berlin Station	0.00051	0.5	9.00557
96	Shadow and Bone	Teen Wolf	0.00030	0.42857	8.59804
61	Game of thrones	Vikings	0.00485	0.27976	8.29019
72	How to get away with murder	Suits	0.00227	0.30985	8.13695
20	South Park	Banshee	0.00030	0.30000	7.85675
..	...	...	...	...	...
30	Breaking Bad	The Originals	0.00030	0.25	14.9537
27	South Park	Big Little Lies	0.00041	0.4	12.2658
60	Lost	Fringe	0.00309	0.37974	11.3924
89	Watchmen	Perception	0.00030	0.2	10.3636
63	Grimm	Perception	0.00030	0.2	10.3636

[103 rows x 5 columns]

Figure 26

C5 Real World Applications

This algorithm can be applied in a real world setting such as entertainment services like Netflix and Disney+. These services can gather the data from the viewers on the shows that they are more likely to watch after watching a certain show. They can put it into a dataset and allow the machine for learning. It will then come up with an algorithm which curate a list of recommended shows for the viewers to watch based on previous watched TV shows history.