CUSTOMER RETENSTION CASE STUDY PROJECT

Problem statement:

Customer satisfaction has emerged as one of the most important factors that guarantee the success of online store; it has been posited as a key stimulant of purchase, repurchase intentions and customer loyalty. A comprehensive review of the literature, theories and models have been carried out to propose the models for customer activation and customer retention. Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit. The research furthermore investigated the factors that influence the online customers repeat purchase intention. The combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively. The data is collected from the Indian online shoppers. Results indicate the e-retail success factors, which are very much critical for customer satisfaction.

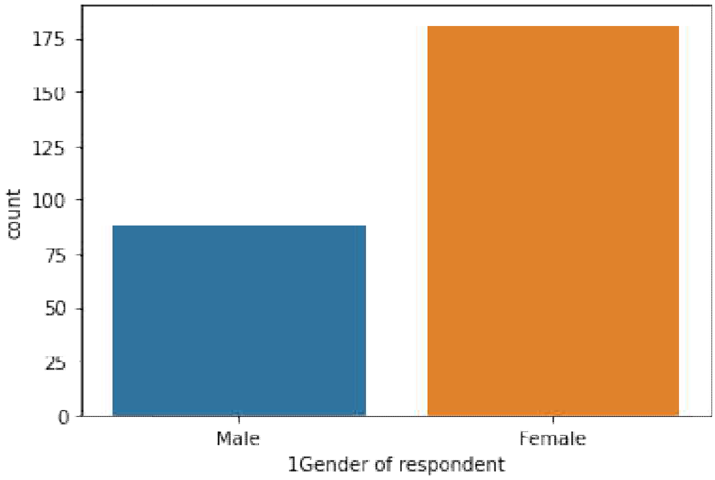
Problem understanding:

Customer segmentation is a process where we divide the consumer base of the company into subgroups. We need to generate the subgroups by using some specific characteristics so that the company sells more products with less marketing expenditure. Before moving forward, we need to understand the basics, for example, what do I mean by customer base? What do I mean by segment? How do we generate the consumer subgroup? What are the characteristics that we consider while we are segmenting the consumers? Let's answers these questions one by one. When we have different segments, we can design a customized marketing strategy as well as customized products that suit the customer of the particular segment. This segment-wise marketing will help the company sell more products with lower marketing expenses. Thus, the company will make more profit. This is the main reason why companies use customer segmentation analysis nowadays. Customer segmentation is used among other domain such as the retail domain, finance domain, and in customer relationship management (CRM)-based products.

EDA Concluding Remarks:

- Find patterns of data through visualization and reveal the hidden trends from data.
- Using both matplotlib and seaborn library to visualize the data.
- Finding relationships between features using bar graphs, histograms, box plots, heat map.
- Analyzing both the numerical and the categorical columns separately.

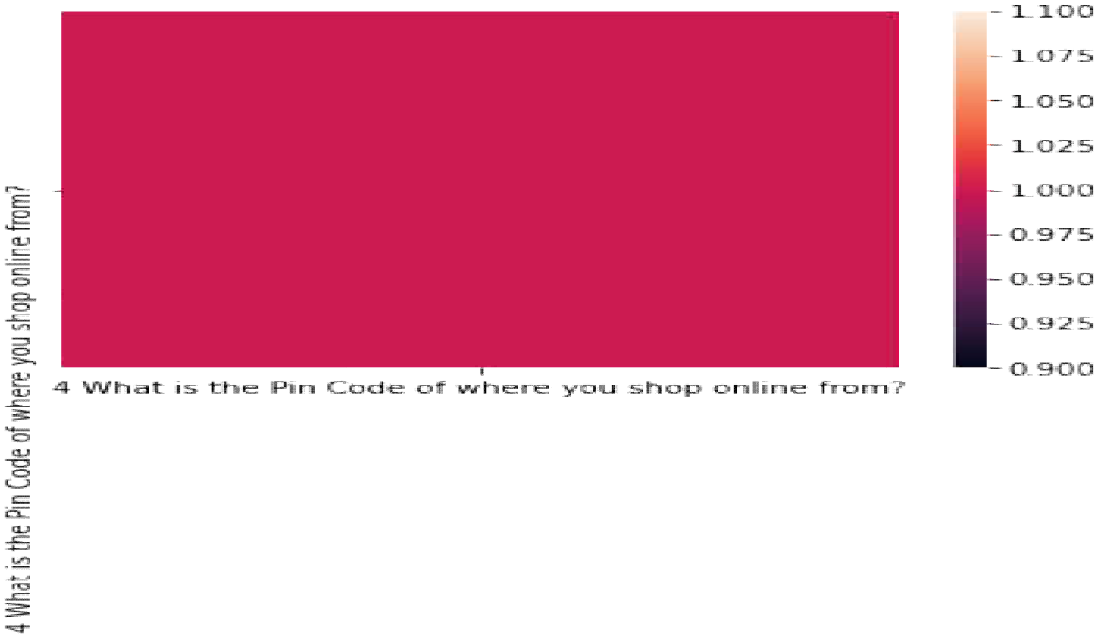
```
sns.countplot(df['1Gender of respondent'])
```



Here Data is not normally distributed.

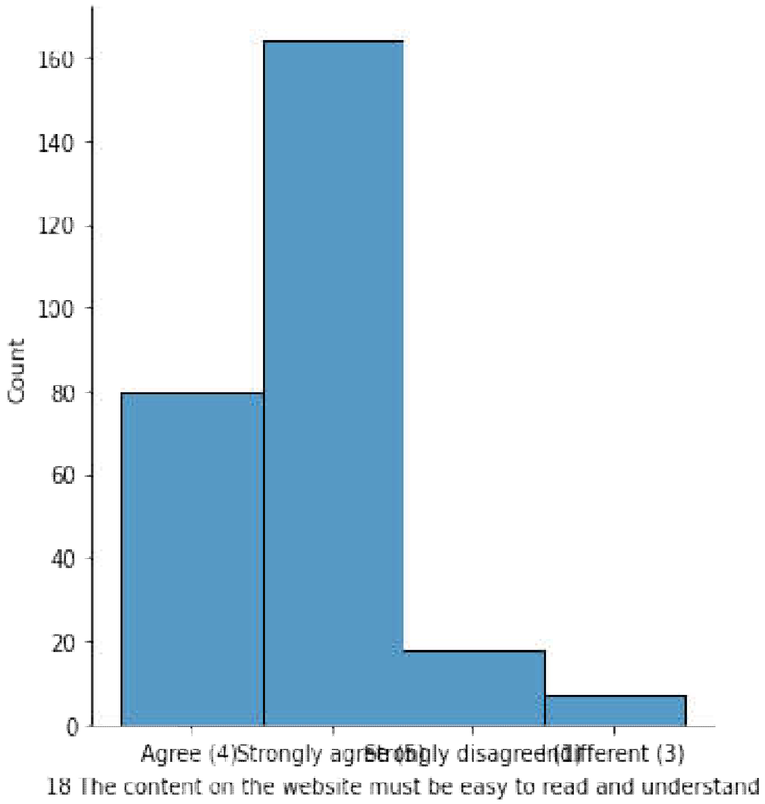
- Are employees leaving because they are poorly paid. Employees are paid an hourly rate of $30 to $100, and attrition seems to happen at every level regardless of employee hourly rate. This can be confirmed later at feature importance.
- Education Field seems to be one of the key factors to attrition, as a larger proportion of education field employees has departed.

sns.heatmap(dfcor)



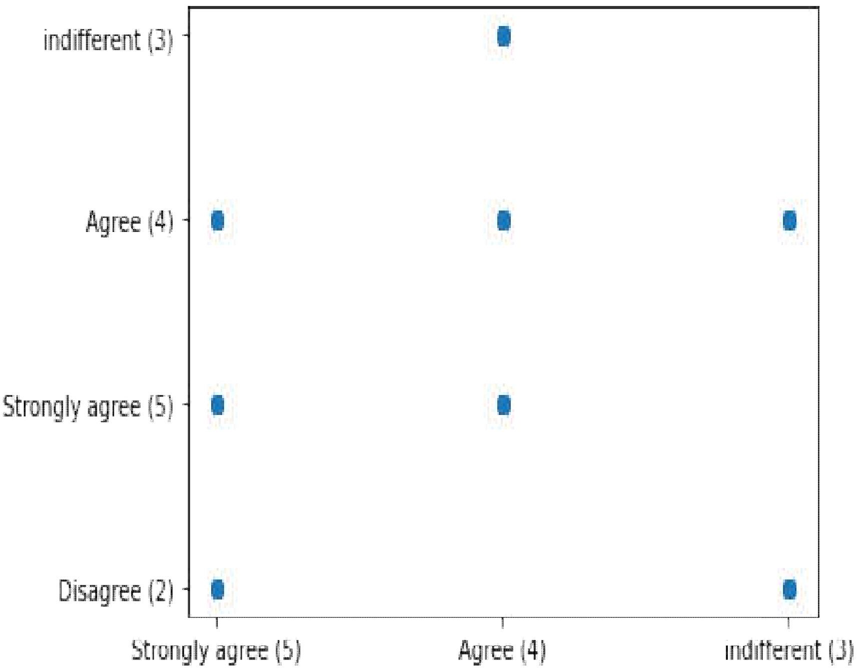
The correlation matrix does not indicate any high degree of correlation with the dependent variable. However, it does provide us with a holistic view off all the factors.

```
sns.displot(df['18 The content on the website must be easy to read and understand'])
```



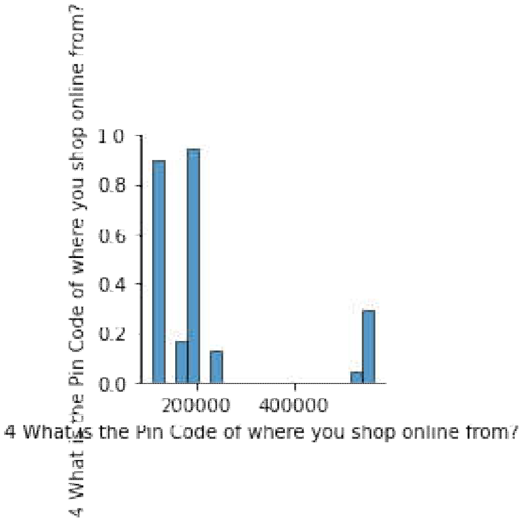
Here data is not normally distributed in all columns because of viscous problem.

```
plt.scatter(df['47 Getting value for money spent'],df['41 Monetary savings'])
```



Scattering the plots of the above code.

```
sns.pairplot(df)
```



Remove the missing values.

Drop the negativity correlated columns.

Remove the outliers.

Pre-processing pipeline:

- For the model to proceed with the data efficiently, the categorical variables salary and department have been encoded. As the values of salary have an order, they have been encoded into integers within the same variable. For department, as the values have no specific order, they have been encoded into individual variables with Boolean values.
- Thus, the dataset has been transformed from 10variables to 19 variables. Numerical variables scaled between 0 and 1 to remove any influence of their difference in value ranges on the model. They have also been checked for skewness, without a real change on their shape.

Building machine learning model:

- As the dataset is imbalance, use cross validation when training the models,and each baseline model performance can be tabulated.
- The model will be cross-validated using a 10-fold cross validation method returning the average accuracy. This method will be applied at every modelling step, to ensure that the model is not biased by the training set split.
- According to the classification report the accuracy of the model is 87% however its recall is lower at 43% of positive cases. The RandomForestClassifiermodel is providing excellent results, however the purpose of the problem is to identify employees that are likely to leave. This is the reason that recall then becomes a very important measure.
- Recall measures the fraction of values that are identified correctly. Random Forest Classifier has emerged as the final winning model with F1-score 100.0% and highest **Recall 100.0%**. This could be the highest possible score achieved with the inherent limitations in the dataset.
- Machine learning models are as good as the data to feed it, and more data would strengthen the model. For example, in this

dataset, the feature ‘Performance Rating’ has been restricted to scores of 3 and 4 only.

- More insights could be generated if the full spectrum of performance ratings is included. In the real-life situation, getting the right data is often more challenging than the analytics itself.

Concluding Remarks:

- Customer retention case study is gaining traction in organizations that embrace digital transformation. The scope has expanded from analytics of employee work performance to providing insights so that decisive improvements can be made to organizational processes. While some level of attrition is inevitable, it should be kept at the minimal possible level.
- This model will allow the company to calculate the probability of an employee to leave the company and to act On key-factors to avoid departures. The satisfaction of employees and the amount of workload they have to bear seem To be important causes of withdrawals. A particular attention on the work-life balance would be crucial to improve the turnover rate.