2022

Mini-Analytics Report

MODULE BEMM457J: TOPICS IN BUSINESS ANALYTICS

ANAND PHADTARE, STUDENT ID: 720005065

UNIVERSITY OF EXETER | ap1001@exeter.ac.uk

Contents

Introduction

The report is prepared based on datasets collected from a fictional Telecommunication (Telco) company providing phone and Internet services to 7043 customers situated in California, USA. The data is specifically from months: July, August, and September, which are part of the third quarter(Q3) as per the concept of Fiscal year for any company performing their accounting, budgeting, and making financial reports.

The information in the data is focused on customers who've left, stayed or signed-up for any of the services the Telco company has to offer. Primary focus of the data is on the following variables from the dataset: Churn Score, Satisfaction Score, Customer Lifetime Value (CLTV) index. Along with these there are multiple customer demographics included to better understand the customer base.

Objectives

- Analysing service usage based on the contract, tenure, monthly charges, and total charges of those services
- Detailed study of customer base of the Telco Company with respect to types of services provided by the company
- Understand customer churn rate of the company and analyse which customer left the services and why they could've left
- Contrary to above objective, studying which customer continued to stay with the company and analyse why they did it
- Identifying gaps in services provided, if any, based on the satisfaction and churn score of the customers

Understanding the Business

Internet service

Phone Service

Online Security

Online Backup

Streaming Media

Tech Support

Based on sample data obtained for Q3, we can see that the telco company provides the following paid services to their customers:

- Phone service
  - Multiple Lines
- Internet Service
  - DSL
  - Fiber optic
- Online Security
- Online Backup
- Device Protection
- Tech Support
- Streaming TV
- Streaming Movies

Customers have an option to choose any kind of subscription they want with respect to 3 types of contracts:

- Month-to-month
- One year
- Two Year

Customers can choose to receive their subscription bills either on paper or in digital format. They're allowed to pay these bills with following methods:

- Bank Withdrawal
- Credit Card
- Mailed Check

Customers can opt for any combination of the services as per their requirements and the bills will be generated depending on choice of services, type of contract on a monthly cycle.

The Telco company provides a unique Customer ID to each of their customers, while collecting the following demographic information from them:

- Gender
- Senior Citizen or not
- Marital Status
- Dependents
- Address Details

All the above details are stored in the database by the Telco Company to keep track of their customer base. Customer feedback is conducted to ensure good quality of service and also to improve on any shortcomings from the company's end to make sure their customers are getting satisfactory experience while using any of the services.

Data collection and understanding

The data used in this report is downloaded from the 'IBM Data and AI Accelerators' which can be found on the community webpage of the IBM where publicly available datasets are allowed to use for commercial or non-commercial purposes. (1)

Datasets collected:

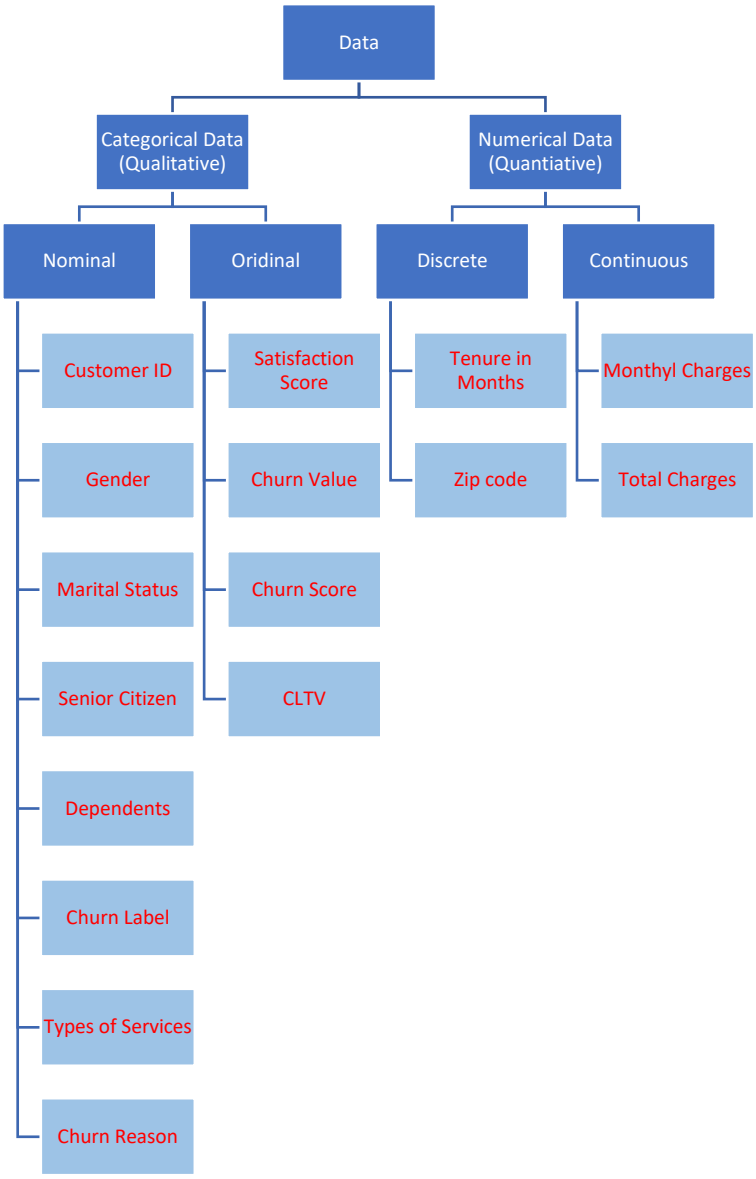
1. **Telco customer Churn:** Sample dataset file containing a fictional telco company's customer churn information including multiple factors relating to it. The 'Churn Label' or 'Churn Value' indicates if any customer left within the last month. The dataset also includes other columns like gender, status of dependents, monthly charges applied to each customer and categories of services provided by the company and which of them are opted by the customers. (2)

2. **Telco Customer Churn status and reason for leaving:** This dataset has some common values as of the first dataset including the reasons why a customer left or stopped using services of the company. The data also provides information about overall satisfaction scores customers have given for company's services and the reasons for those scores. (3)

Understanding the data:

1. First Dataset primarily focuses following parameters: Customer demographics and location details; types of services provided, billing cycle, type of contracts, methods of payment used, and churn status categories.
2. Customer demographics and their locations will help in exploring insights on the type of customer base the company has and from which area of population the customers are subscribing to the company's services.
3. Billing related information will provide details about which type of services customers can choose to subscribe to; tenure of those services, which payment methods are available, and type of contract chosen will decide the amount of monthly billing cycles along with total charges at the end of the quarter.
4. Finally, churn status and different categories of it will help in understanding the customer churn rate on basis of billing information, overall experience of using their choice of service/services and how much satisfied or unsatisfied they are will decide if the respective customer would continue with their subscription or move out to look for alternative solutions.
5. The datasets had data in structed format consist of both qualitative and quantitative types of data in them.

Dividing all the variables of both datasets into the Data type they belong to by using a simple hierarchical diagram as follows:



Definitions of variables present in the dataset

- **Churn Label:** This column consists of two categorical values: 'Yes' OR 'No',
    a. **Yes:** Customers who left in the current quarter
    b. **No:** Customers who stayed with the company
- **Churn Value:** This column is simple conversion of values from 'Churn Label' column, where Yes = 1 and No = 0
- **Churn Reason:** Customer's reasons for leaving the company or stopping their services
- **Churn Score:** This consists of values between 0-100, values under this column were calculated using IBM's predictive tool. The higher the value of Churn score, higher the chances of customer leaving
- **CLTV:** Customer Lifetime Value, these values are calculated using predictive model and formulas on existing data. Higher the CLTV, more valuable the customer is for the company
- **Satisfaction Score:** Overall Satisfaction score given by each customer ranging from 1 being 'Very Unsatisfied' and 5 for 'Very Satisfied'. (4)

Data Preparation

- The primary tool used to prepare both the datasets which were in '.xlsx' format: Microsoft Excel
- First step executed while preparing the datasets was cleaning the datasets to ensure smooth operations during data modelling phase of the report.
- Investigation was conducted to look for any missing or incorrect values to avoid any kind of obstructions for processes ahead in the report.

Data cleaning:

- Since the data obtained from the main source was already in a structed format with well-defined column names and absence of any extra space and/or incorrect characters, the duration spent for cleaning the data was relatively less.
- There were no repeating or duplicate values in the datasets
- Although there were some blank cells present in following columns:
    - Total Charges: Number of blank cells 11
    - Churn Category: Number of blank cells 5174
    - Churn Reason: Number of blank cells 5174

  To further investigate the reason for missing values in these cells, exploratory data analysis was done using Microsoft Excel, by filtering the whole dataset to highlight only the blank cells under all the columns. Later it was discovered that 11 missing values under the column 'Total Charges' were absent since the tenure for each of values was having a value: 0. Assuming minimum value under 'Tenure' column need be 1 to complete a monthly billing cycle; the blank cells from column 'Total Charges' were converted to 'Zero' values. This was achieved executing following command:

```
df.replace(df['Total_Charges'] == ' ', 0, inplace=True) #Command to convert blank cells  value to : Zero
```

This was done to make sure there will not be any computational error during data modelling process. Alternative approach could've been to simply remove all the rows corresponding to these missing values, but due to active subscription of all customers under this specific condition, their churn value was considered as 'No', referring that they would still be likely to continue the services with the company.

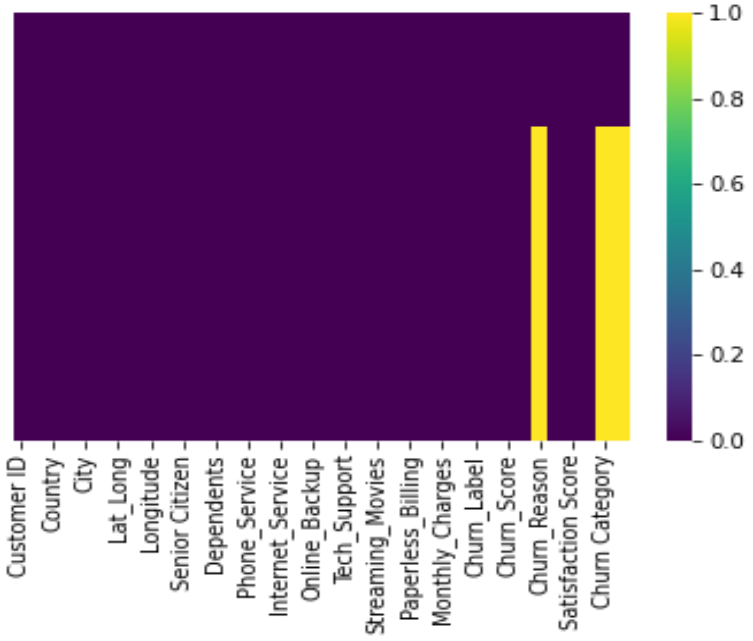
Another issue found in the same column was with the datatype of values in the column, apparently Python was reading all the values under this column as 'object' or 'string' datatype. This was fixed by executing following command:

```
df['Total_Charges'] = df.Total_Charges.astype(float) # Command to convert string objects to float for computations
```

While missing values for 'Total Charges' column could be filled, same could not be achieved for remaining two columns namely: 1. Churn Category 2. Churn Reason. For both the columns exact number of values were missing from the original datasets. Since both columns had categorical data under them with a high number of missing values (5174), all the blank cells of these columns were transformed into 'Null' values. This was achieved by executing following commands:

```
df.replace(df['Churn Reason'] == ' ', np.nan, inplace=True) #Command to convert blank cells  value to : isnull
```

```
df_status.replace(df_status['Churn Category'] == ' ', np.nan, inplace=True) #Command to convert blank cells  value to : isnull
```



Heatmap diagram of all the null values (in yellow colour) present across the dataset

Data Manipulation

- The data manipulation involved renaming some of the column names to make sure they can be used for modelling operations in python.
- For ease of visualisation and computation of the data, the datasets were merged into a single dataframe using following command:

```
df_merged = pd.merge(df, df_status, how='left', on='Customer ID') # command to merge datasets with primary key : Customer ID
```

In order to avoid repeated columns and datapoints, duplicate columns were removed using following command:

```
df_merged = table_success_score.drop(labels=['Count', 'Churn Label','Churn Value','Churn Score','CLTV', 'Churn Reason'], axis=1)
```

Performing both Data Cleaning and Data Manipulation, it was made sure that following criteria are satisfied to go ahead with data modelling, analysis, and visualization phases (5):

Questions	Answers
1. Does the data make sense?	Yes, since it is well structured, cleaned, and ready to be analysed
2. Does the data follow appropriate rules for its fields?	Yes, it is made sure that all blank cells are either filled or removed for seamless computational purposes and datatypes are appropriate for the category of data
3. Does it bring any insights to light?	Yes, shown in data modelling
4. Able to find trends in data to help with next theory?	Yes, the trends will help in making recommendations for the business
5. If not, is that because of quality of the data?	Although there are few limitations in the completeness of the data from original source, but the available data quality is improved through iterative data cleaning process.

Answering these questions ultimately help in achieving following characteristics of quality data (6):

1. **Validity.** The degree to which data obeys set principles or rules of the business	☑
2. **Accuracy.** Data is close to true values	☑
3. **Completeness.** Degree of gathering all required data	☒
4. **Consistency.** Consistency of data even after merging multiple datasets	☑
5. **Uniformity.** Degree of using same unit of measure for respective variables	☑

The limitation in achieving Completeness existed because of high number missing values in two variables of the datasets. Availability of this data would have enabled the report to get more accurate insights corresponding to those two variables.

Modelling and Analysis

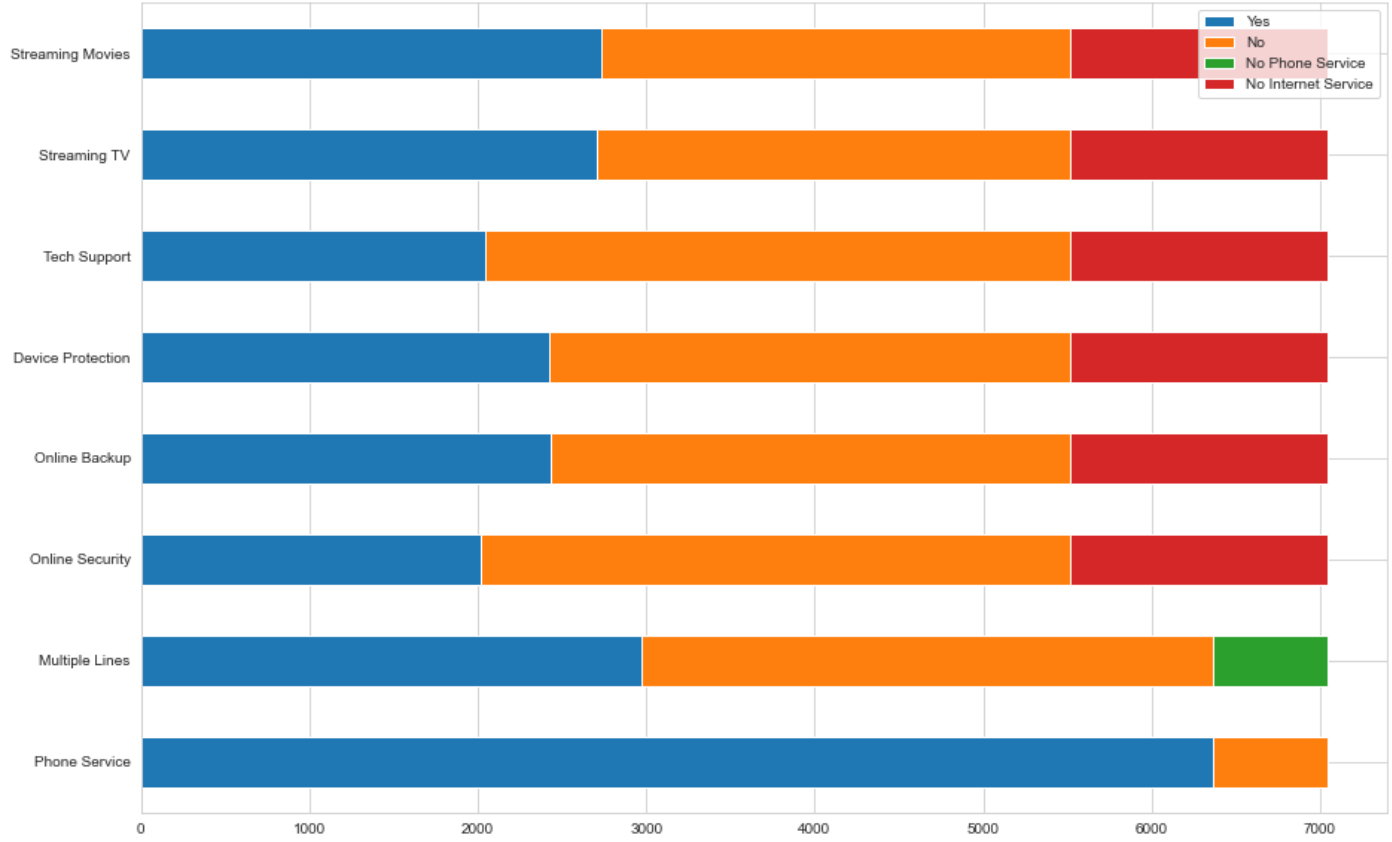
In this section, initially various aspects of the dataset are described using descriptive statistics and analytics approaches.

This table gives Descriptive statistics about numerical data from multiple columns.

	Tenure_Months	Monthly_Charges	Total_Charges	Churn_Value	Churn_Score	CLTV	Satisfaction Score
count	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000
mean	32.371149	64.761692	2279.734304	0.265370	58.699418	4400.295755	3.244924
std	24.559481	30.090047	2266.794470	0.441561	21.525131	1183.057152	1.201657
min	0.000000	18.250000	0.000000	0.000000	5.000000	2003.000000	1.000000
25%	9.000000	35.500000	398.550000	0.000000	40.000000	3469.000000	3.000000
50%	29.000000	70.350000	1394.550000	0.000000	61.000000	4527.000000	3.000000
75%	55.000000	89.850000	3786.600000	1.000000	75.000000	5380.500000	4.000000
max	72.000000	118.750000	8684.800000	1.000000	100.000000	6500.000000	5.000000

Analysis of service distribution and their charges

Following diagram illustrates the distribution of types of services subscribed by the customers:



Before going into analysis part, let us understand exact functions of these services in brief:

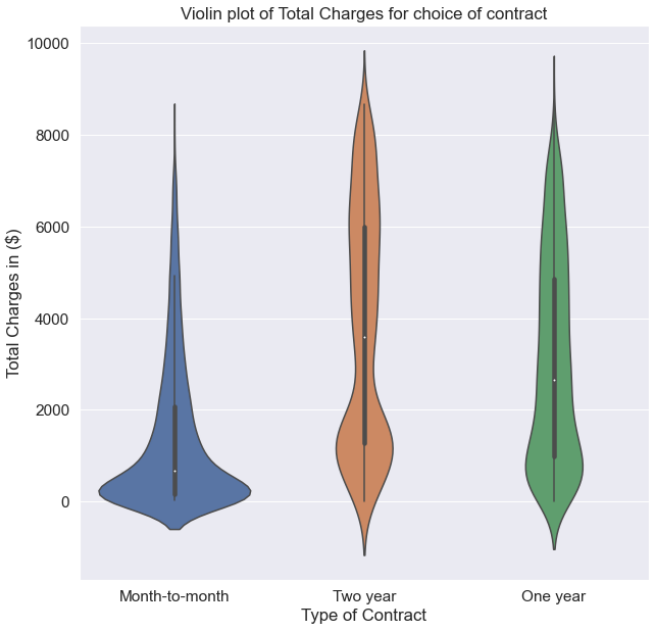
- **Phone Service**: Indicates if customers subscribed to home phones service by 'Yes' or 'No' Values.
- **Multiple Lines:** Indicates if customers subscribed to an additional feature of multiple telephone lines by 'Yes', 'No' values.
- **Online Security:** Indicates if customers subscribed to an additional Online security feature provided by the company by 'Yes' or 'No' values.
- **Online Backup:** Indicates if customers subscribed to an additional Online backup feature provided by the company by 'Yes' or 'No' values.
- **Device Protection:** Indication of the customers subscribed to an additional feature which protects the internet device or equipment provided by the company using 'Yes' or 'No' values.
- **Tech Support:** Indicates if the customers opted for premium tech support to reduce the wait time by using 'Yes' or 'No' values.
- **Streaming TV:** Indicates if customers their use internet services to stream TV programs from third-party provider, the company does not charge extra fees for this feature.
- **Streaming Movies:** Indicates if customers their use internet services to stream movies from a third-party provider, the company does not charge extra fees for this feature.

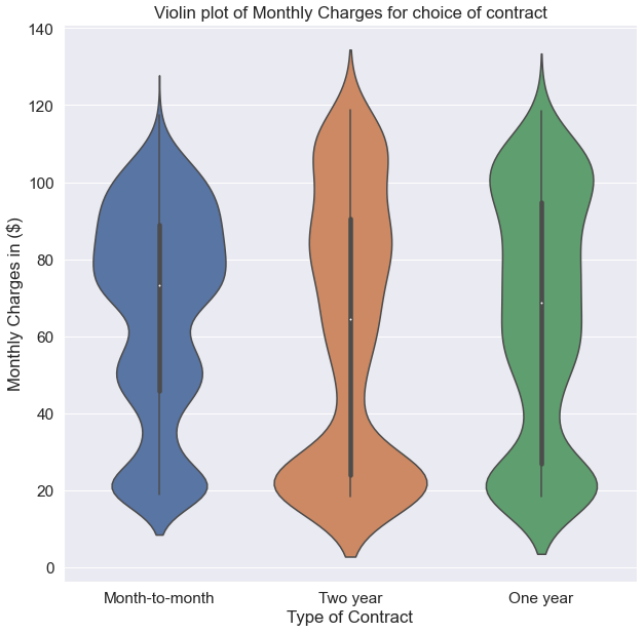
Note: 'No Phone Service' represents type of customers who have neither a 'Phone Service' nor its additional feature of 'Multiple Lines', whereas 'No Internet Service' indicates number of customers who've not opted for the internet services namely: 'DSL' and 'Fiber Optic' provided by the company.

- 90.3% of customers have subscribed to the telephone service, and out of that 2971 customers have subscribed to multiple telephone lines feature. This indicates 'Phone service' is either most popular or best quality service provided by the Telco company.
- Almost similar number of customers have subscribed to both 'Online Backup' and 'Device Protection' services each contributing count just above 34%, which is nearly 9% lesser than customers who've not opted for these services.
- Number of customers who've not opted for any kind of Internet services provided by the company (neither DSL nor Fiber optic) remains constant at 1526, indicates that the company is missing out on these customers who are using internet services with other providers.
- Only 29% of customers use 'Tech Support' service, which could either mean the quality of other service is good enough to not reach out to get help of support staff or the 'Tech Support' is unable to meet good standards. Following table illustrates exact percentage of usages of the services:

Types of Services	Yes	No	No Internet Service	No Phone Service
Phone Service	90.3%	9.7%		
Multiple Lines	42.2%	48.1%		9.7%
Online Security	28.7%	49.7%	21.7%	
Online Backup	34.5%	43.8%	21.7%	
Device Protection	34.4%	43.9%	21.7%	
Tech Support	29.0%	49.3%	21.7%	
Streaming TV	38.4%	39.9%	21.7%	
Streaming Movies	38.8%	39.5%	21.7%	

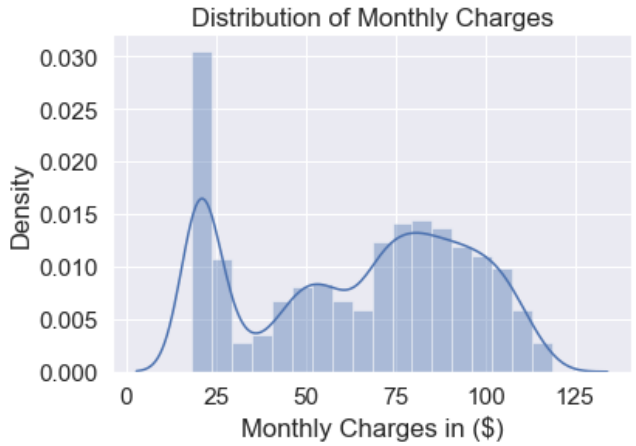
Distribution of monthly and total charges with respective to type of contact

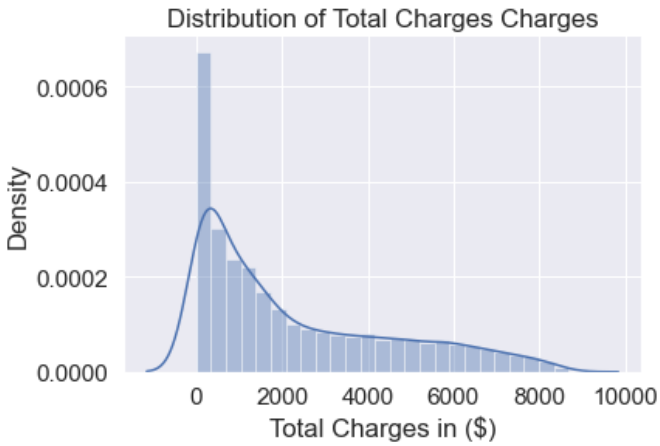




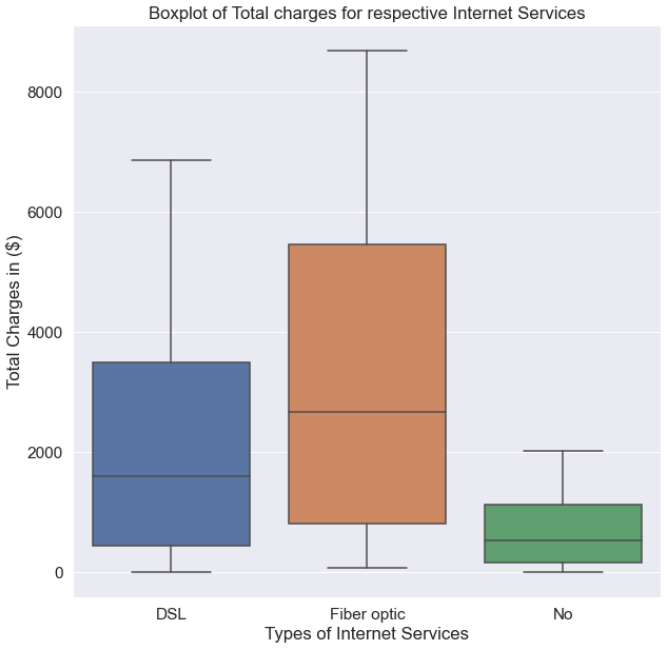
- We can that there is high density of total charge for month-to-month contract, meaning large number of customers are opting for month-to-month basis, and from the graph next to it we can see the distribution of monthly charges for the same type of contract. Population density is varying for the range of amount they are paying for monthly charges, having highest density within price range $70-$90.
- For two-year contract high number customer count normalizes around $20 for monthly charges, having median value just above $60, whereas for total charges the spread of customer base is relatively even from low to high price range stretching upto $8000.
- Same could be said about spread of customers with one-year contract, having peak of density around $1000 starting from 0 and steadily decreasing till $8000. Monthly charges have similar density of customers for prices $20 and just above $100

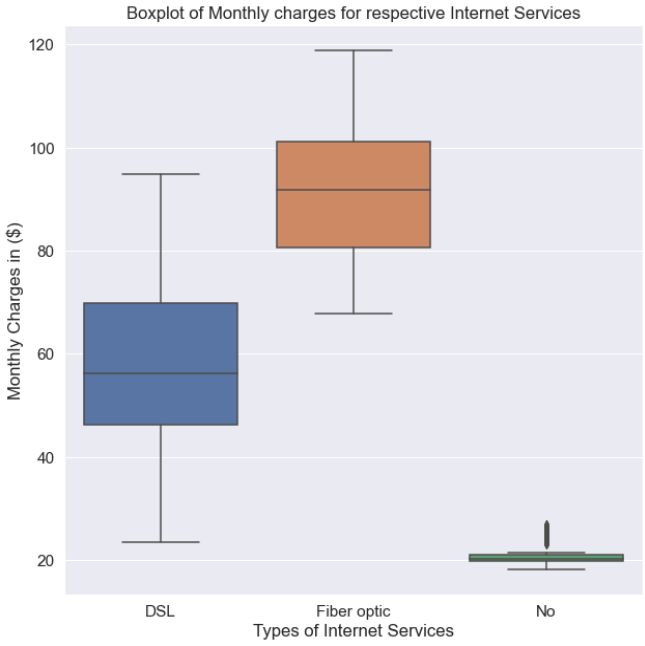
Following graphs shows distribution of monthly and total charges regardless of their contract type





Distribution of monthly and total charges for choice of Internet Service



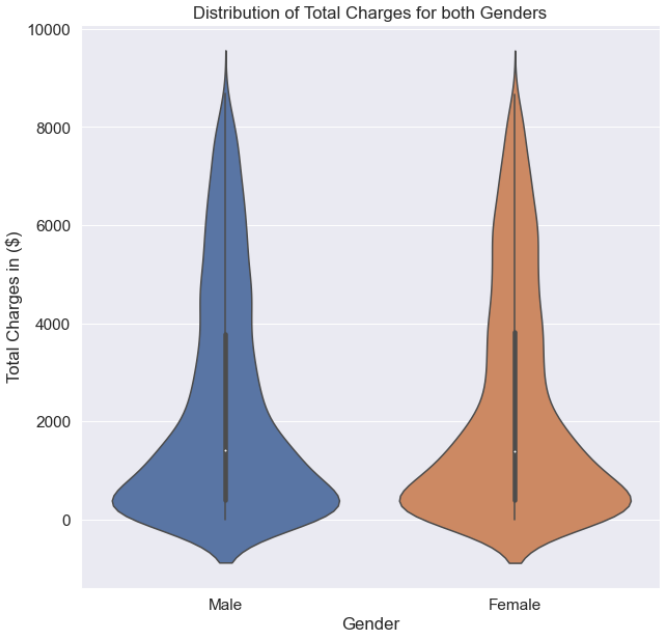


- It is evident that total charges for customers who are using 'Fiber optic' services are relatively high compared to customers who either use 'DSL' or customers who're not using any internet service with the company. This indicates popularity of 'Fiber optic' service, having median value around $3000
- Customers who are not using any internet services are charged significantly low total charges indicating that Internet services contributes large portion of total charges applied at the end of the quarter.
- Similar can be observed from the boxplot monthly charges with respect to type of internet service used. We can see that monthly charges are extremely low around $20 per month.

- Here we can see a correlation or regression diagram of Monthly Charges versus Total Charges
- From the diagram we can see that the variables do not have a perfect relationship with each other
- Although there is no strong relationship between these variables, we can observer a weak relationship, which exists cause of other deciding factors like, choice of services, types of contracts and varying tenure etc.

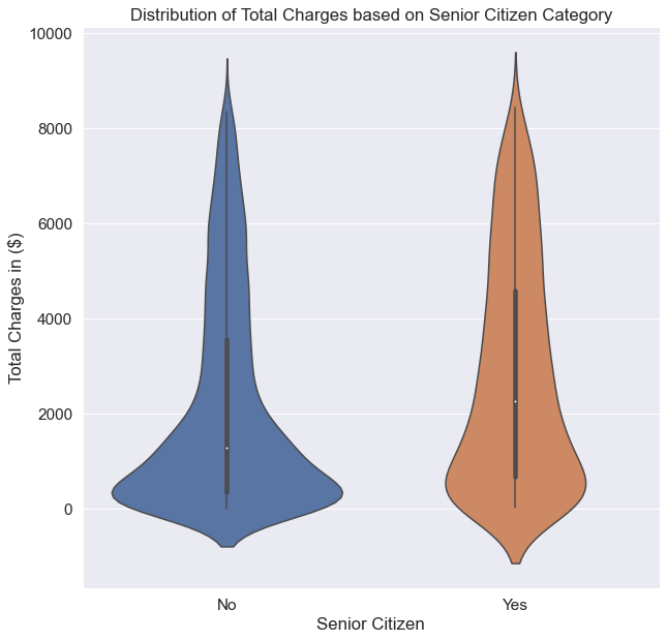
Customer Demographic Analysis

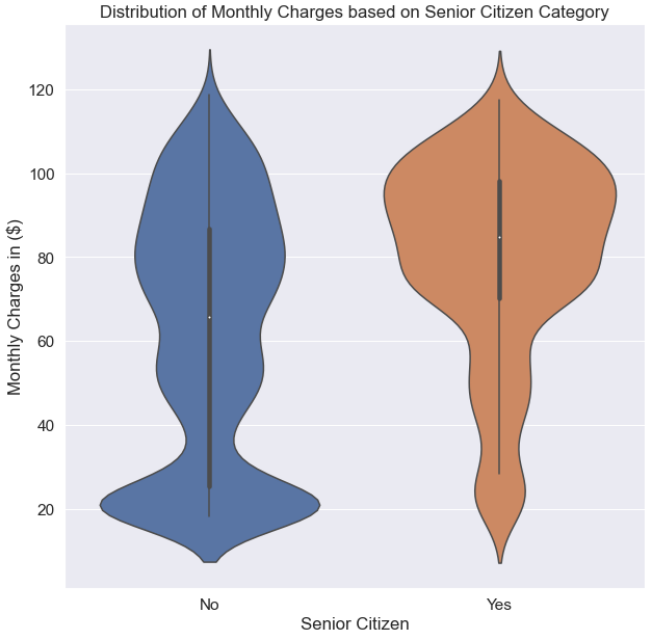
Analysis based on Gender of the customers



- From both these diagrams it is very clear that service usage of both the Male and Female gender is almost identical with only a minor dissimilarties among both. It can also be infered that customer base of the company is equally when considering Gender of the customers who are signing up for the services. But this diagnostic only apply for the Q3 of the financial year, since the data sample only respresents data from Q3.

Analysis based on Senior Citizen crietria



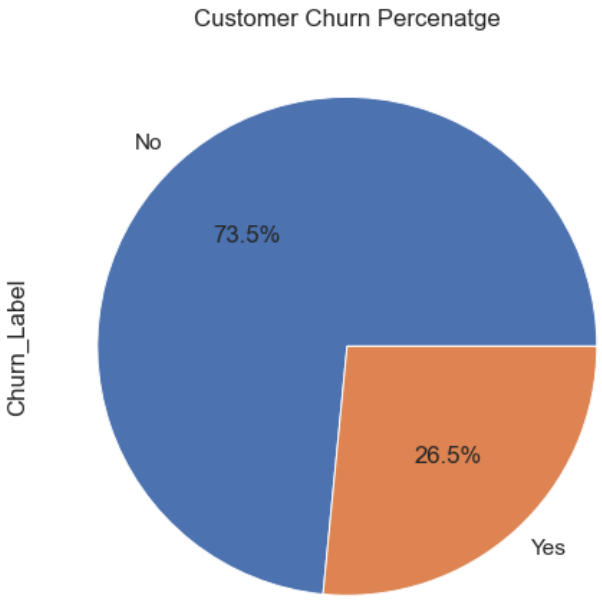


- One common observation about both total charges and monthly charges, senior citizen tend to have higher median amount compared to customers who are yonger than them. Another observation is that large number of Senior citizens have subscribed to monthly services which cost from $80 to $100, conversely large number of younger customers are subscribing to monthly services whose values are just above $20

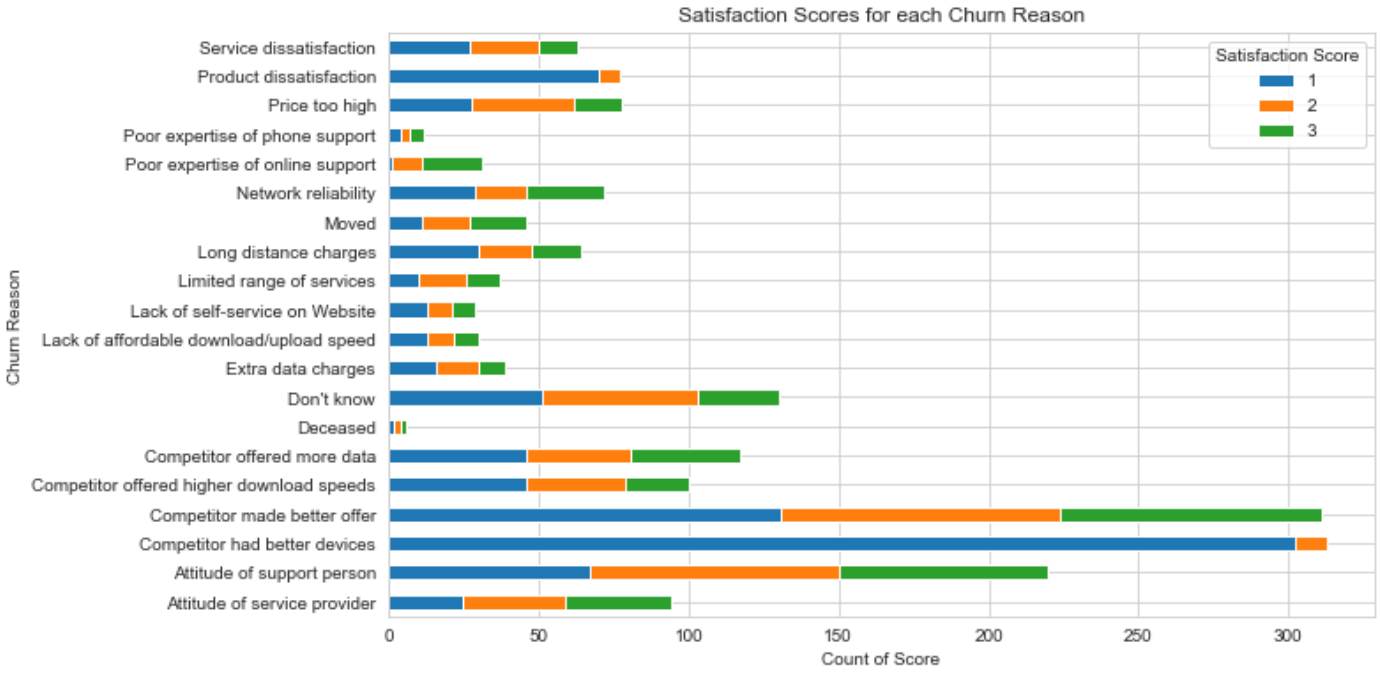
Customer Churn Analysis

Now let us first look at the Customer Churn percentage with the help of ‘Churn Label’

- Total number of customers in the datasets are 7043, which come under dataset from Q3.
- Based on the visualisation of the following pie chart, the percentage of customers who are leaving the company or unsubscribing to all the services is 26.54%
- Number of customers who are continuing with the company corresponds to 73.46%



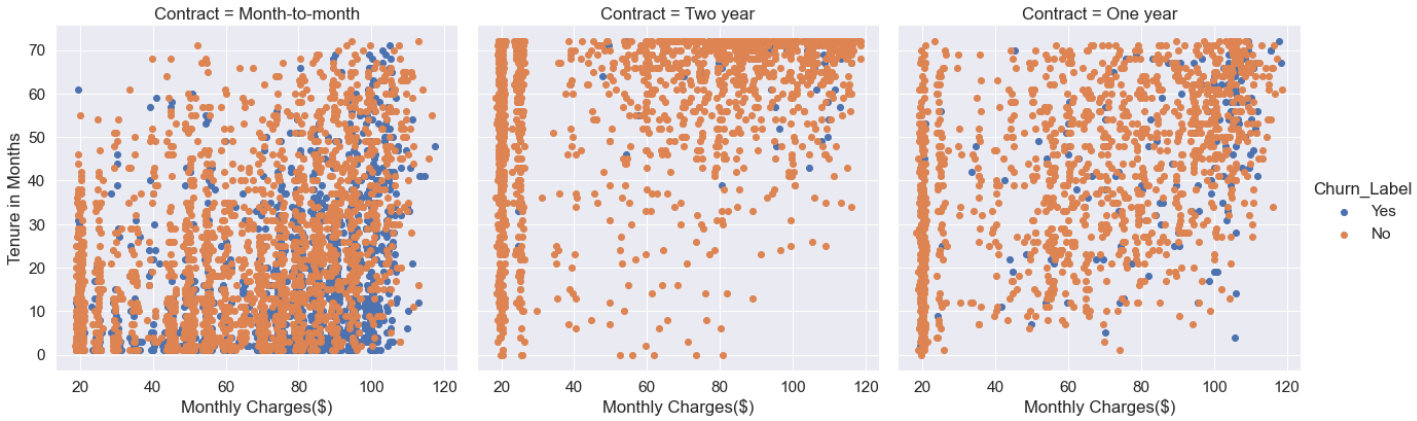
Customer Churn analysis based on Churn reason and Satisfaction score



- Highest number of unsatisfied customers are leaving on basis reason which says: Competitor had better devices

- Second highest contributing reason for leaving is because competitor made better offer according to the customers.
- According to around 220 customers, attitude of the technical support agent made them to leave the company and unsubscribe to their services, contrary to that lesser number of customers experienced poor support when it comes to expertise of phone and online services.
- Considerable number of customers stopped the services for reason unknown to them, which indicates absence of crucial feedback and overall experience of using the services with the company.

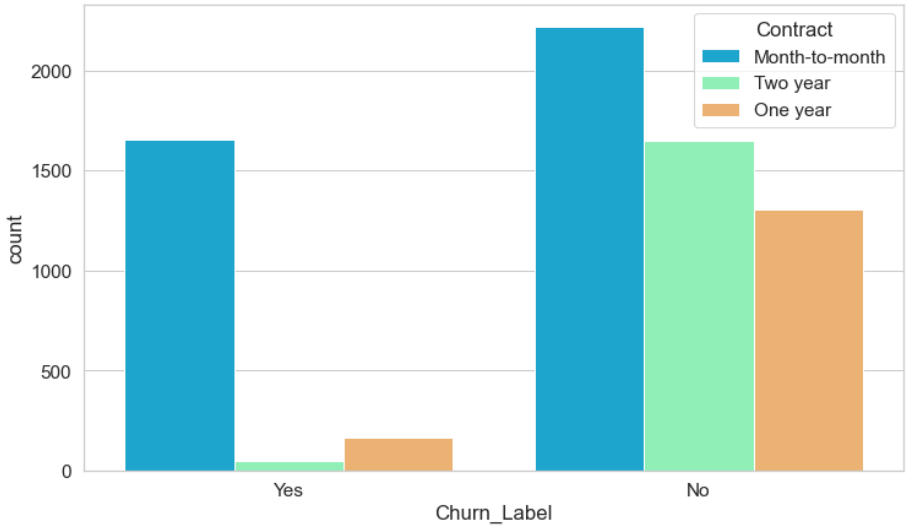
Following diagram gives overview of Churn distribution across the type of contracts with respect to monthly charges paid by those customers over the tenure of months customer stayed with the company or left in the current quarter.



- Based on graph above, we can see that high number of customers left who were subscribed to services on month-to-month contract agreement. And huge proportion of such customers were getting monthly charges between $70-$110.
- When it comes to two years contract very few customers are choosing to leave the company, and the churn rate for one year contract is slightly higher than that. Monthly charges of such customers have formed clusters near low charges as well as high charges.

Similar conclusions can be justified on basis of following bar chart:

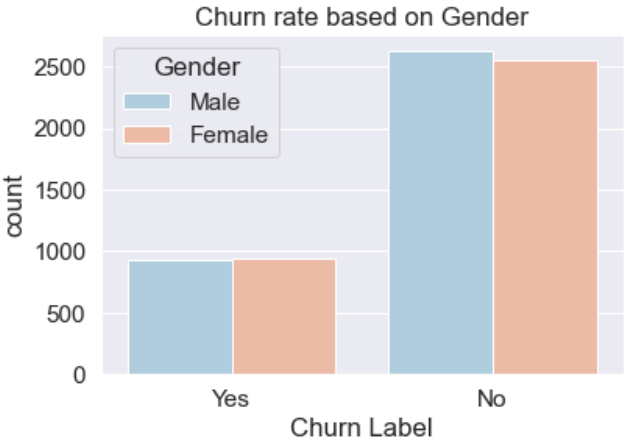
- But this diagram gives more clarity on the fact that highest number of customers who continue to stay with the company also prefer to stay on month-to-month contract.



Customer Churn rate based on Demographics

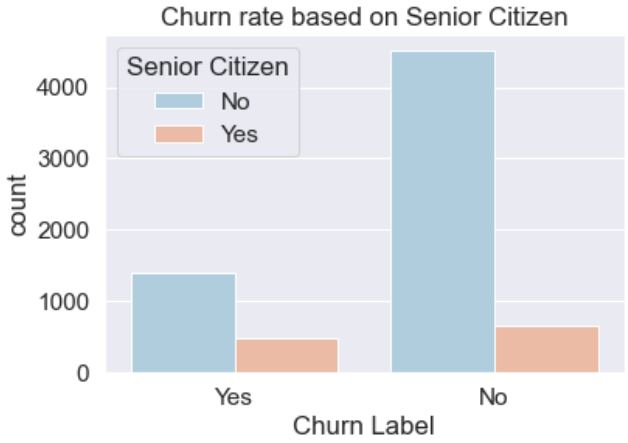
Based on Gender of the customers:

- There is no significant difference in churn rate considering the gender of customers. The proportion of gender among customers who are leaving the services is almost equal to 800 in both male and female categories.
- Male customer numbers are only slightly higher than that of females who continue to stay with the company for Q3



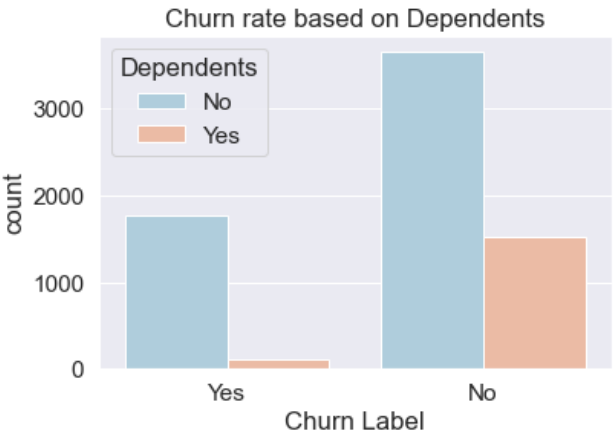
Based on Senior Citizen:

- Interestingly, when it comes to considering if the customer is a senior citizen or not, the churn rate has a noticeable difference.
- Number of senior citizens who are leaving the company is slightly lesser than customers who continue to stay
- Whereas for customers who are younger than those have a relatively higher rate of continuing with the company compared to customers who are parting.

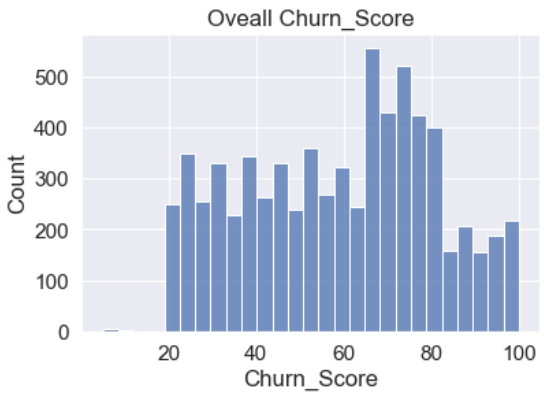


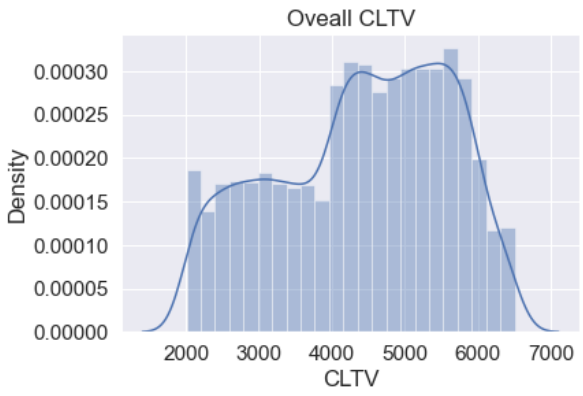
Based on Dependents:

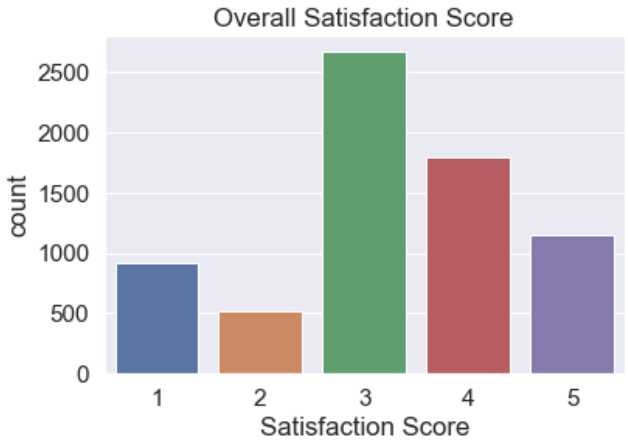
- Based on the graph, we see customers with dependents have a higher rate of continuing with the services compared to those leaving the company
- Whereas for customers without any dependents in the family, their proportion for continuing with the company is double that of customers who are leaving.



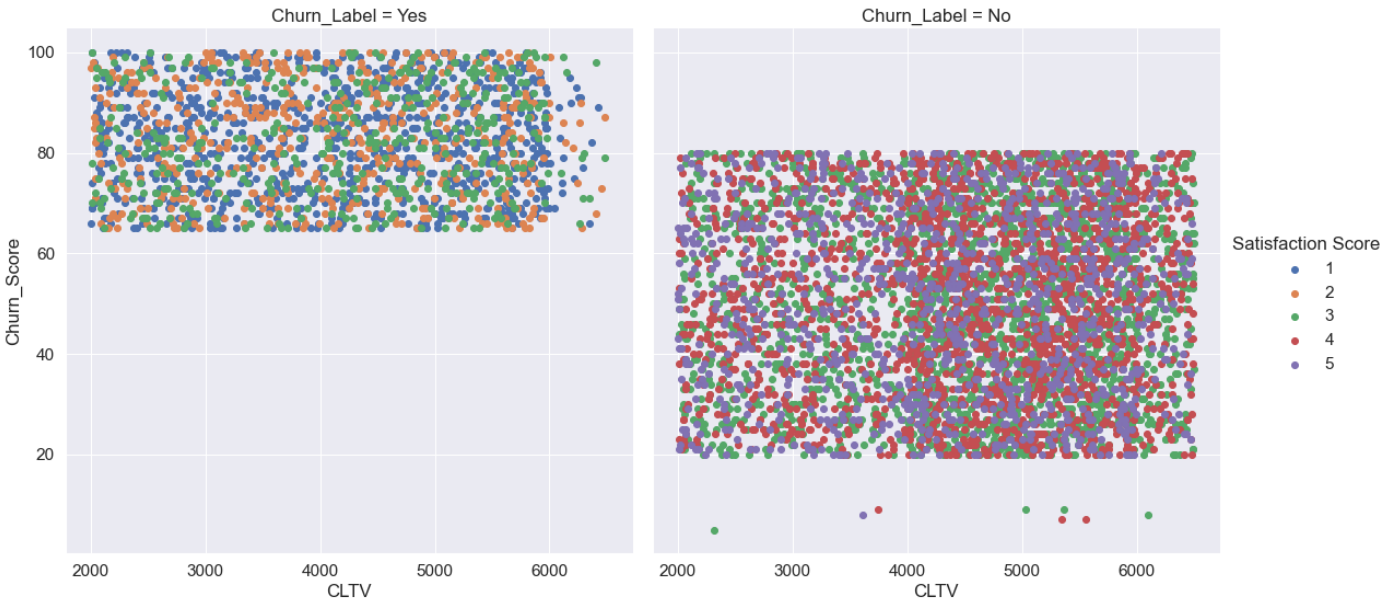
Analysis of Churn Score, CLTV along with Satisfaction Score







Now let us see the correlation between these three parameters by using a scatter plot



- Uniform distribution of CLTV can be observed in customer who left the company, but their Churn Score was predicted to be within a high range from 70-100, with but the outliers above CLTV 6000 could be retainable even with high Churn score
- Most satisfied customers who chose to stay with the company lie between CLTV 4000-6000
- Highest number of customers have given satisfaction score of 3 for their overall experience with the company and its services.

Map of Churn Score based on customer locations all over the state of California

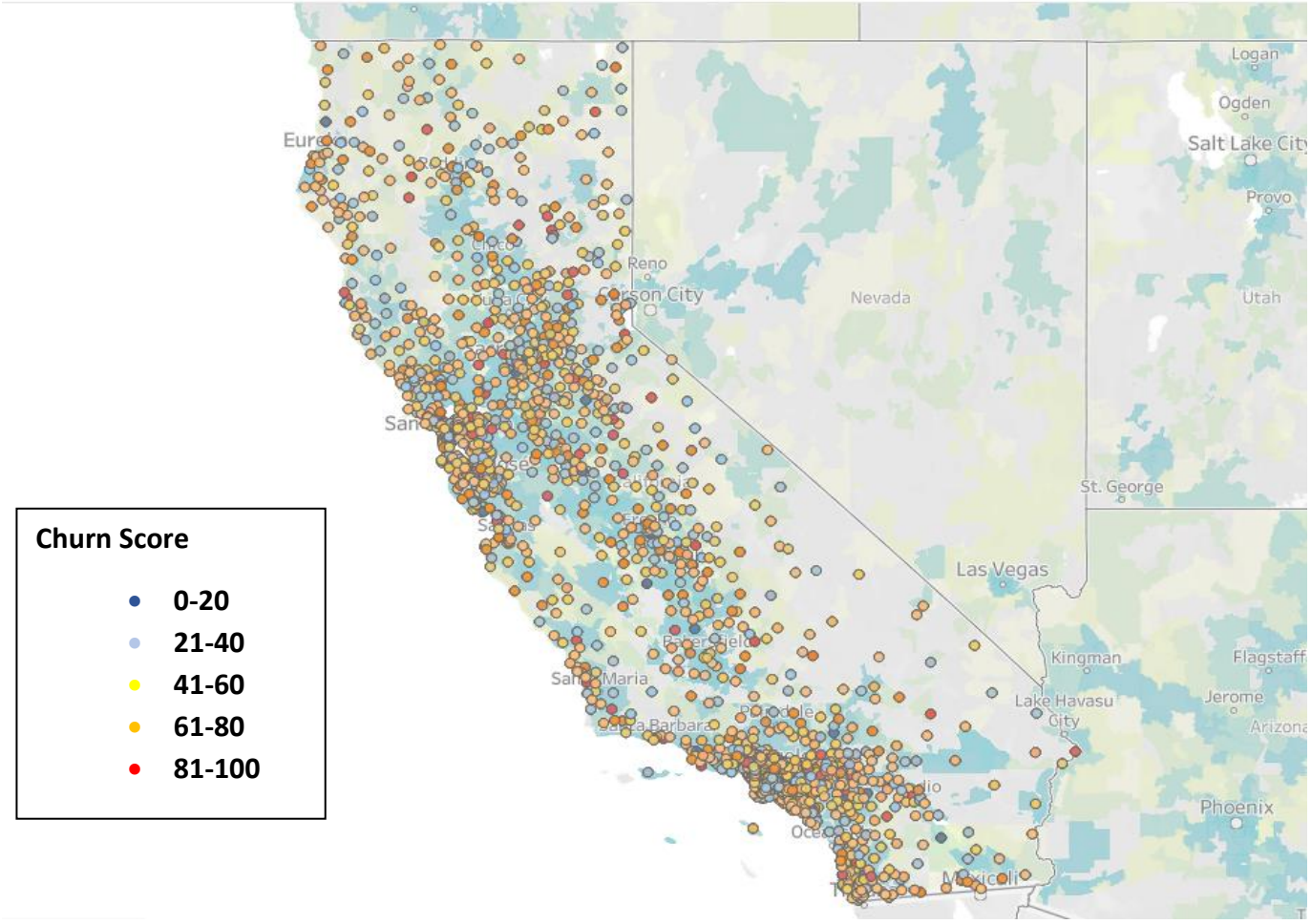


Figure is created using Tableau application (10)

Conclusions and recommendations

- Majority of customer base is subscribed to Phone services, Online backup, and Device protection services.
- Based on analysis of service usage, it can be implicated that considerable number of customers are not subscribing to internet services provided for the Telco company. Since internet services provided are bring significant amount of business into the company, a thorough investigation needs to be conducted to find why they are choosing other service providers.
- 'Online security' being the least subscribed service, security gaps in this service either making customer unsubscribe the service or leave the company as well. Same needs to be investigated.
- Based on customer feedback, the company seems to have poor quality of devices compared to their competitors. The company should consider upgrading their devices to retain existing customers.
- The company should consider introducing better customer centric price offers based on their contract and usage of services to retain customers who are leaving or about to unsubscribe.
- Poor technical support service could lead to loss of loyal customers, gaps in the knowledge and courteousness of tech support staff could be delt with regular training on the required domain.
- Discount offers and marketing campaigns can be conducted to target customers and potential audience based on customer demographics.
- Missing data from two variables limited the accuracy of the analysis driven for churn reason, if the required data was available, the predictions could be made more accurately not just for the current quarter but also for upcoming quarters.
- The report could not present lot of data on regression between price variables, since dataset is majorly consisting of categorical data.

Data Ethics

The data ethics in this report are followed based on UK government data ethics framework, where it is made sure that all three major principles are practised diligently at every phase of the report.

1. Since the original dataset source already anonymised the personal details of customers like their name and address details, ensured accountability of not sharing any sensitive information through the report.
2. The data obtain from a community portal of IBM, where datasets are publicly available to use for commercial and/or non-commercial purposes. (7)
3. There were certain limitations in deriving accurate analysis and predictions due to missing values, but extra care was taken during data cleaning phase to gather all required data for analytical computations. (8)
4. While working on and analysing customer demographics, fair analysis is done throughout the report with regards to gender, age, location of the customers.
5. During data manipulation process, it was made sure that there is no deviation in the information from original datasets, and the process were only executed for computational flexibility. (9)

References:

- IBM developer. (n.d.). https://developer.ibm.com/terms/ibm-developer-terms-of-use/
- Telco customer churn. (n.d.). https://community.ibm.com/accelerators/catalog/content/Telco-customer-churn
- Telco customer churn status and reason for leaving. (n.d.). https://community.ibm.com/accelerators/catalog/content/Telco-customer-churn-status-and-reason-for-leaving
- Tableau. (n.d.). Guide to data cleaning: Definition, benefits, components, and how to clean your data. https://www.tableau.com/learn/articles/what-is-data-cleaning
- Telco customer churn (11.1.3+). (2019, July 11). https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113
- Srip. (n.d.). global icon. Flaticon.com. https://www.flaticon.com/free-icon/global_975645?term=internet%20service&page=1&position=3&page=1&position=3&related_id=975645&origin=style
- Freepik. (n.d.). Call icons. Flaticon.com. https://www.flaticon.com/premium-icon/call_484525?term=telephone&page=1&position=42&page=1&position=42&related_id=484525&origin=style
- Smashingstocks. (n.d.). Online security icons. Flaticon.com. https://www.flaticon.com/premium-icon/online-security_3929728?related_id=3929738&origin=search
- Ultimatearm, (n.d). Backup icons. Flaticon.com. https://www.flaticon.com/free-icon/backup_3043400?related_id=3043400&origin=search
- Freepik, (n.d). Customer service icons. Flaticon.com. https://www.flaticon.com/free-icon/technical-support_4233818?term=tech%20support&related_id=4233818
- Freepik. (n.d). Streaming icons. Flaticon.com. https://www.flaticon.com/premium-icon/live-streaming_1865042?related_id=1865042&origin=search
- Phadtare. (2022, April 4). Telco Customer Churn Score. https://public.tableau.com/app/profile/anand.phadtare/viz/TelcoCustomerChurnScore/Sheet2?publish=yes