Customer Churn Analysis

2023

Table of Contents

CONTENTS

INTRODUCTION

The area of exploration for this project is customer churn for a telecom company in the California region.

The problem objective can be categorized into two broad domains.

1) Identifying the key components that constitute a customer leaving a company, finding a relationship between the said factors

.

2) Provide recommendations on improvement of retention of the customers.

Additionally other attributes are also investigated and have revealed their impact on churn of customers.

By working on the areas of lacking we can decrease the churn rate and increase the customer loyalty.

DATA DESCRIPTION

The data set used for the project is acquired from Kaggle, the source that collected the data is Maven Analytics which is free to use by the public and can be accessed through the website: https://www.mavenanalytics.io/blog/maven-churn-challenge

The Customer Churn table contains information on all 7,043 customers from a Telecommunications company in California in Q2 2022

Each record represents one customer, and contains details about their demographics, location, tenure, subscription services, status for the quarter (joined, stayed, or churned).

The Zip Code Population table contains complimentary information on the estimated populations for the California zip codes in the Customer Churn table.

The data is open source and free to use by the public.

PROCESS

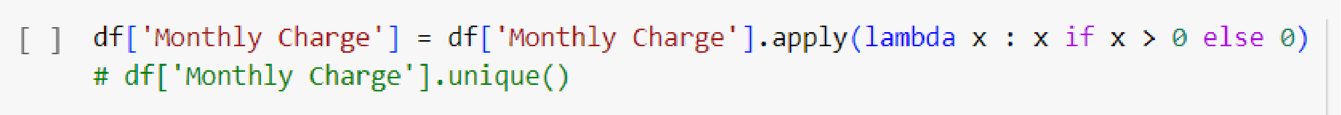
The tools used for the project include

> Python programming language to preprocess the data, and perform EDA, power BI for data viz

Few of the steps involved in preprocessing of the data are.

- Handling Missing Values
- Correction of Data Type
- Fixing Data Range
- Checking for regular expressions
- Cross field Ref

Some values in Monthly charge were negative and had to be fixed, we assumed it to be zero instead of negative.



```
[ ]  df['Monthly Charge'] = df['Monthly Charge'].apply(lambda x : x if x > 0 else 0)
     # df['Monthly Charge'].unique()
```

The columns that had huge ranges needed to be converted into bins for that we used Sturges formula as follows:

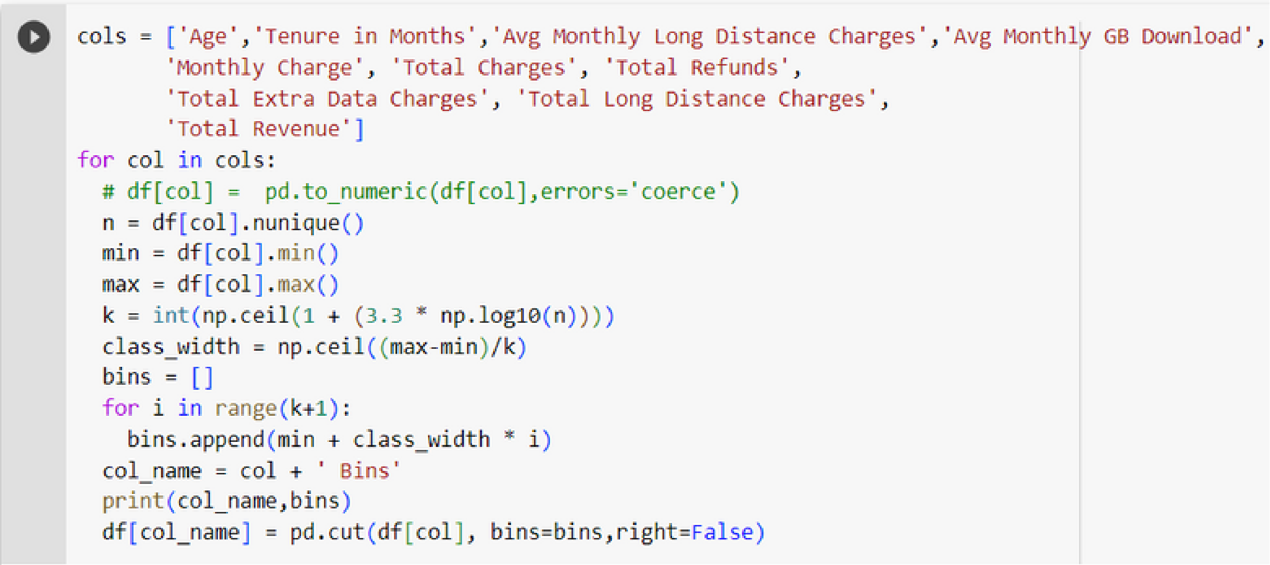
$$K = 1 + 3.3 \log_{10}(n)$$

And then determined the column width as follows:

$$\text{Col Width} = \frac{(\text{Largest val - smallest val })}{K}$$

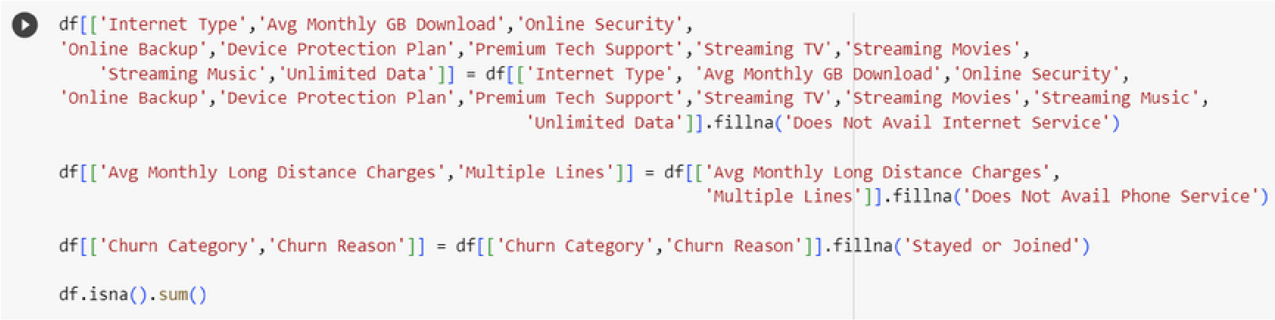
The table as follows shows the detailed calculations:

Col Name	Min	Max	Distinct Values	No of bins	Col Width
Age	19	80	1.62	1.7	1.9
Tenure in months	1	1.72	1.72	1.7	1.10
Avg Monthly Long Distance Charges	1.01	1.49.99	1.3583	1.13	1.4
Avg Monthly GB Download	2	1.85	1.49	1.7	1.12
Monthly Charge	0	1.118.75	1.1582	1.12	1.10
Total Charge	18.8	1.8684.8	1.6540	1.14	1.619
Total Refund	0	1.49.79	1.500	1.10	1.5
Total Long Distance Charge	0	1.3564.72	1.6068	1.13	1.274
Total Revenue	21.36	1.11979.34	1.6975	1.14	1.854



```python
cols = ['Age','Tenure in Months','Avg Monthly Long Distance Charges','Avg Monthly GB Download',
       'Monthly Charge', 'Total Charges', 'Total Refunds',
       'Total Extra Data Charges', 'Total Long Distance Charges',
       'Total Revenue']
for col in cols:
  # df[col] =  pd.to_numeric(df[col],errors='coerce')
  n = df[col].nunique()
  min = df[col].min()
  max = df[col].max()
  k = int(np.ceil(1 + (3.3 * np.log10(n))))
  class_width = np.ceil((max-min)/k)
  bins = []
  for i in range(k+1):
    bins.append(min + class_width * i)
  col_name = col + ' Bins'
  print(col_name,bins)
  df[col_name] = pd.cut(df[col], bins=bins,right=False)
```

Then we fixed the missing values in columns as follows:



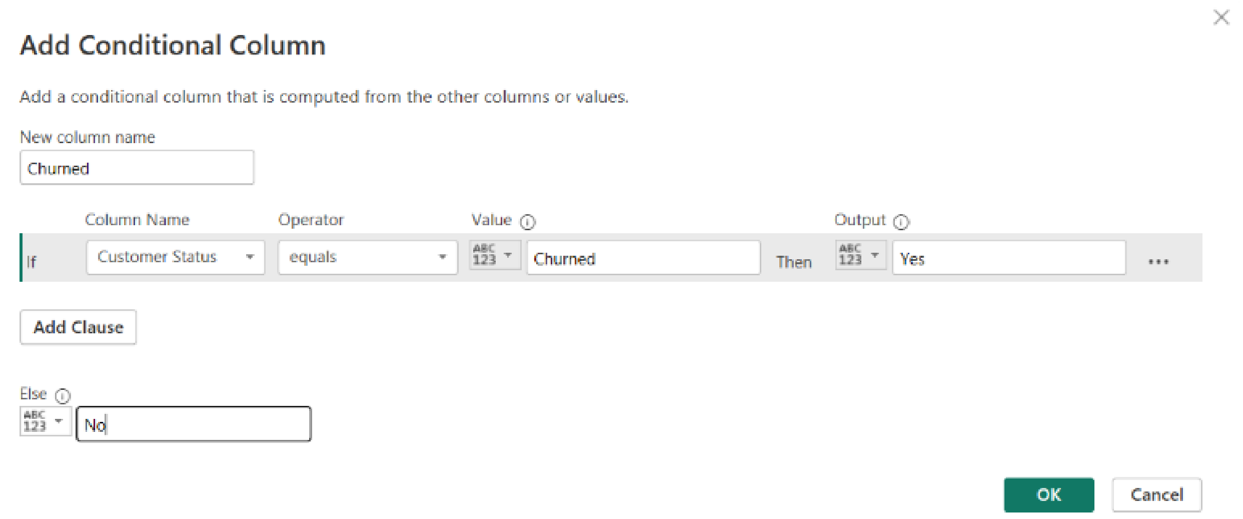
```python
df[['Internet Type','Avg Monthly GB Download','Online Security',
'Online Backup','Device Protection Plan','Premium Tech Support','Streaming TV','Streaming Movies',
    'Streaming Music','Unlimited Data']] = df[['Internet Type', 'Avg Monthly GB Download','Online Security',
'Online Backup','Device Protection Plan','Premium Tech Support','Streaming TV','Streaming Movies','Streaming Music',
                                           'Unlimited Data']].fillna('Does Not Avail Internet Service')

df[['Avg Monthly Long Distance Charges','Multiple Lines']] = df[['Avg Monthly Long Distance Charges',
                                                             'Multiple Lines']].fillna('Does Not Avail Phone Service')

df[['Churn Category','Churn Reason']] = df[['Churn Category','Churn Reason']].fillna('Stayed or Joined')

df.isna().sum()
```

Then we loaded the data into power query and performed some final pre-processing steps there before proceeding to visualization phase:



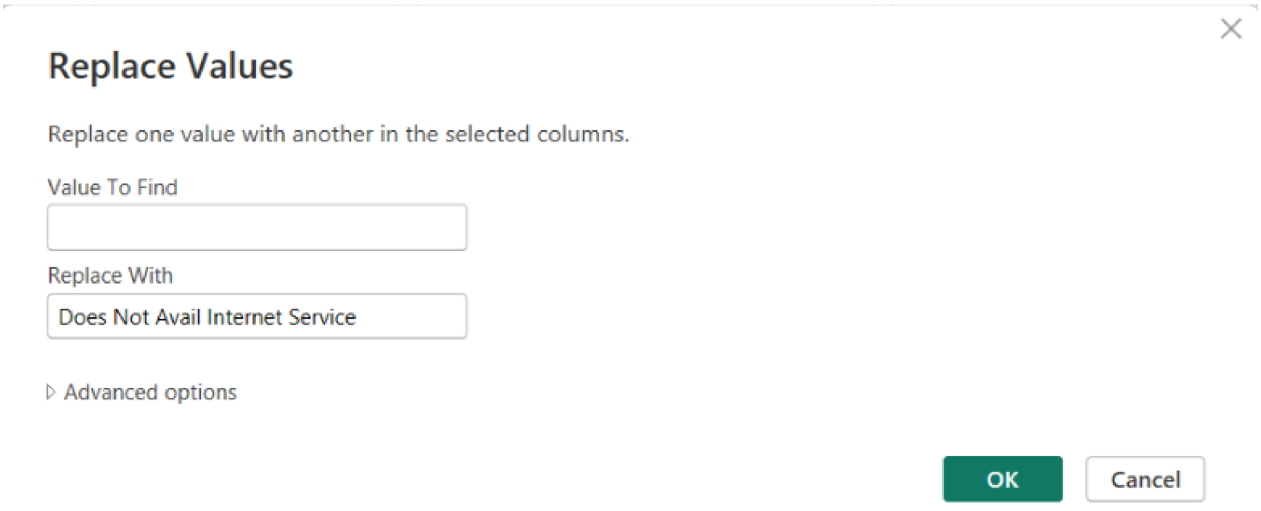
Add Conditional Column

Add a conditional column that is computed from the other columns or values.

New column name: Churned

	Column Name	Operator	Value		Output
If	Customer Status	equals	Churned	Then	Yes

Else: No

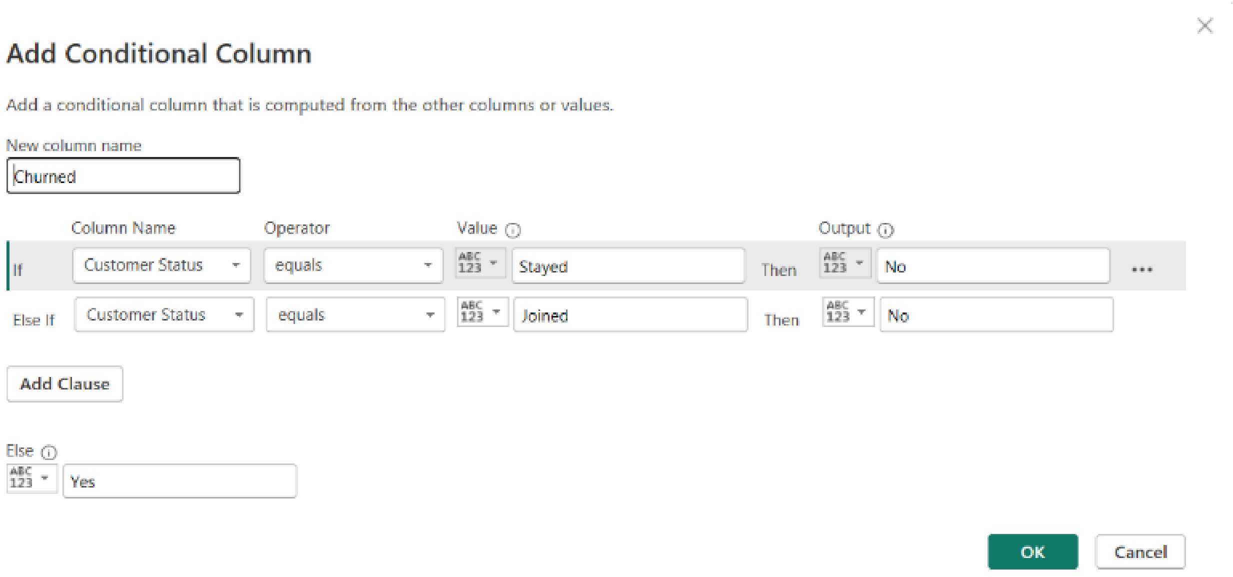


Replace Values

Replace one value with another in the selected columns.

Value To Find

Replace With: Does Not Avail Internet Service



Add Conditional Column

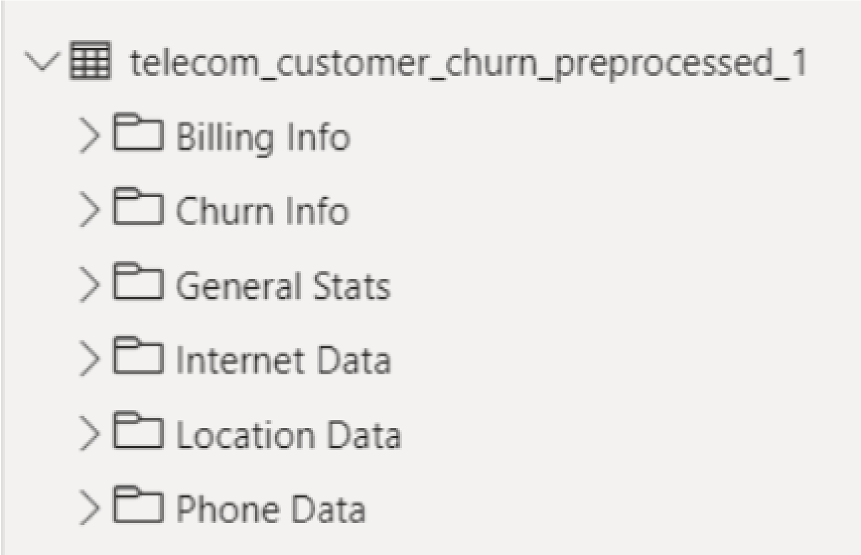
Add a conditional column that is computed from the other columns or values.

New column name: Churned

	Column Name	Operator	Value		Output
If	Customer Status	equals	Stayed	Then	No
Else If	Customer Status	equals	Joined	Then	No

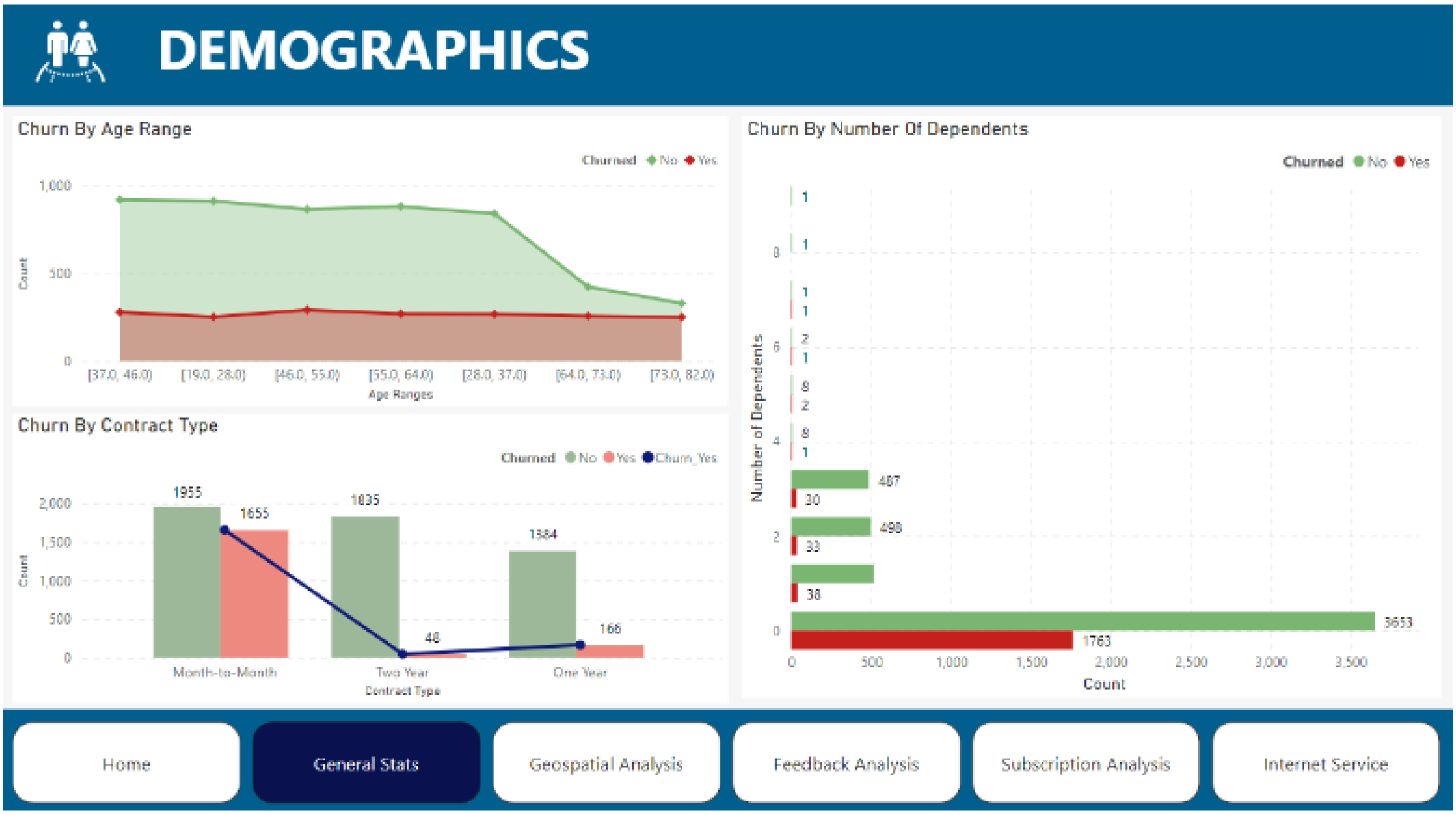
Else: Yes

Finally we stored the different columns into a designated folder to tidy up the work space.



telecom_customer_churn_preprocessed_1
- Billing Info
- Churn Info
- General Stats
- Internet Data
- Location Data
- Phone Data

ANALYZE

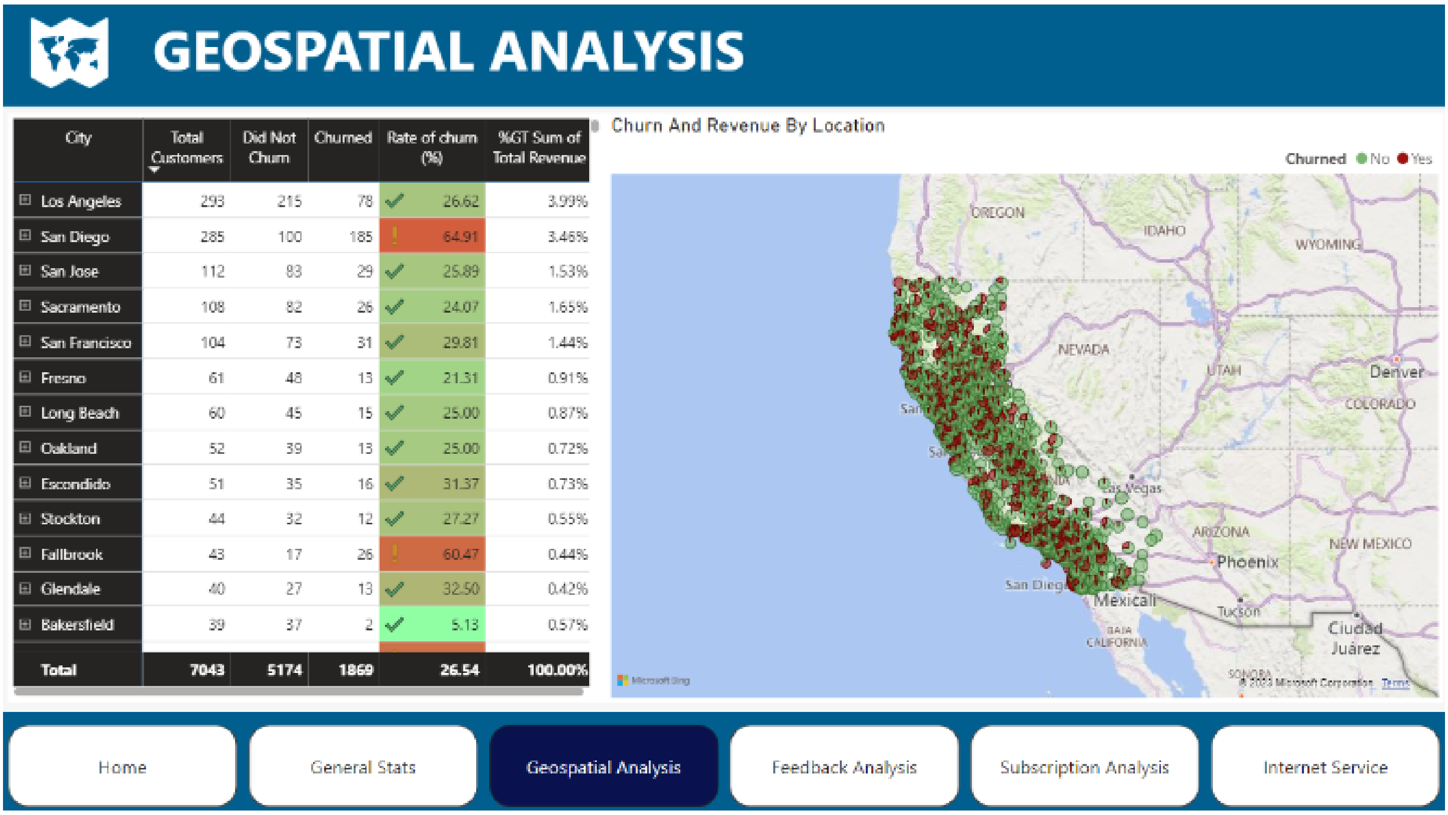


The demographics page depict the relationship between the attributes: age, dependents and contract type with churn.

Key Findings:

01 From the visuals you can see that people who avail the month-to-month contract are prone to leave the company as compared to those who have longer contracts.

02 Another key insight is that the individuals in the age bracket of 75 to 82 are more likely to churn then those in other categories.



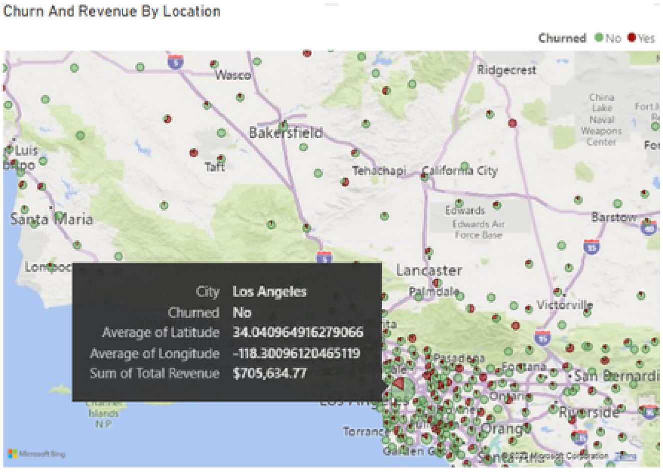
GEOSPATIAL ANALYSIS

City	Total Customers	Did Not Churn	Churned	Rate of churn (%)	%GT Sum of Total Revenue
⊞ Los Angeles	293	215	78	✓ 26.62	3.99%
⊞ San Diego	285	100	185	! 64.91	3.46%
⊞ San Jose	112	83	29	✓ 25.89	1.53%
⊞ Sacramento	108	82	26	✓ 24.07	1.65%
⊞ San Francisco	104	73	31	✓ 29.81	1.44%
⊞ Fresno	61	48	13	✓ 21.31	0.91%
⊞ Long Beach	60	45	15	✓ 25.00	0.87%
⊞ Oakland	52	39	13	✓ 25.00	0.72%
⊞ Escondido	51	35	16	✓ 31.37	0.73%
⊞ Stockton	44	32	12	✓ 27.27	0.55%
⊞ Fallbrook	43	17	26	! 60.47	0.44%
⊞ Glendale	40	27	13	✓ 32.50	0.42%
⊞ Bakersfield	39	37	2	✓ 5.13	0.57%
Total	7043	5174	1869	26.54	100.00%

Churn And Revenue By Location

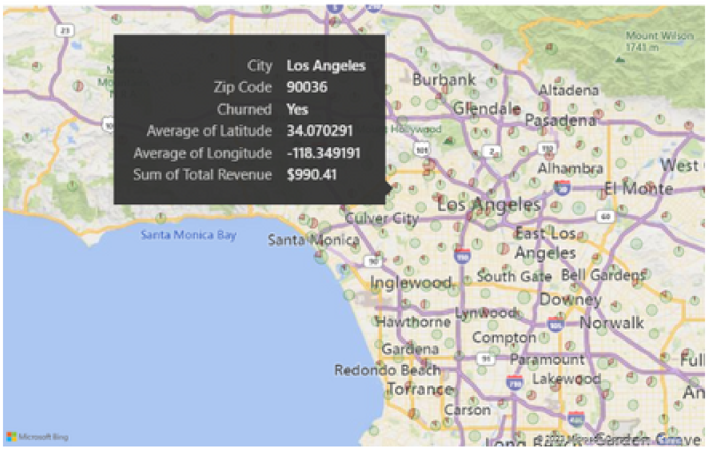
Firstly the map visualization shows the city and their proportion to churn rate, moreover the size of the bubble indicate how much a city contributes towards the total revenue generated by the company.

Moreover you can further drill down and see how does a particular zip-code within the city contributes to the total revenue and the proportion of churn.



Churn And Revenue By Location

City	Los Angeles
Churned	No
Average of Latitude	34.040964916279066
Average of Longitude	-118.30096120465119
Sum of Total Revenue	$705,634.77

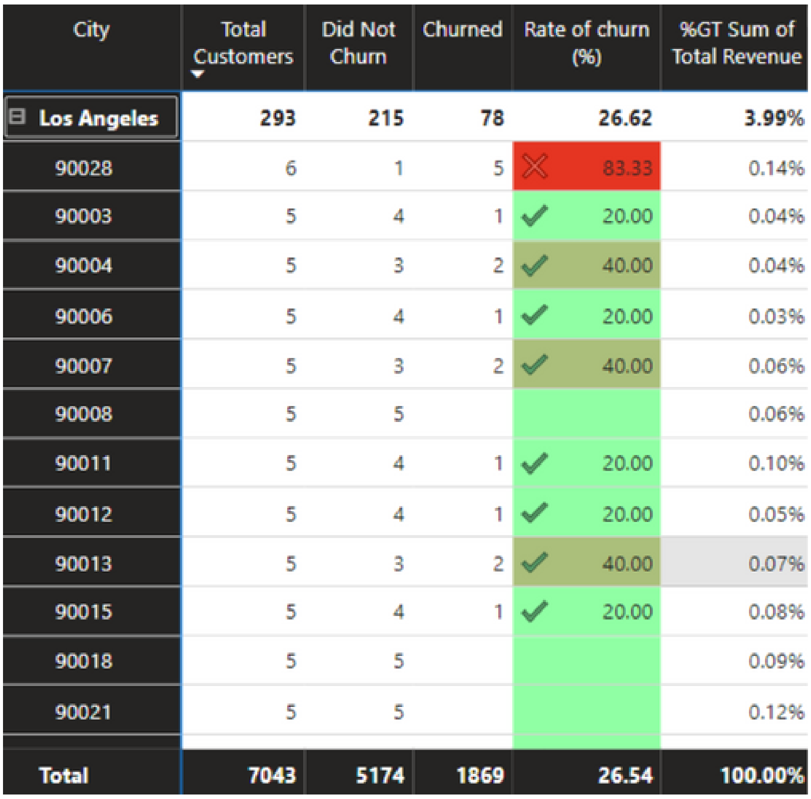


City	Los Angeles
Zip Code	90036
Churned	Yes
Average of Latitude	34.070291
Average of Longitude	-118.349191
Sum of Total Revenue	$990.41

The table on the left has four columns to indicate the total customers in a particular city, how many of them did not churn, how many churned, **rate of churn***, how much each city contribute to the total revenue (in percentage).

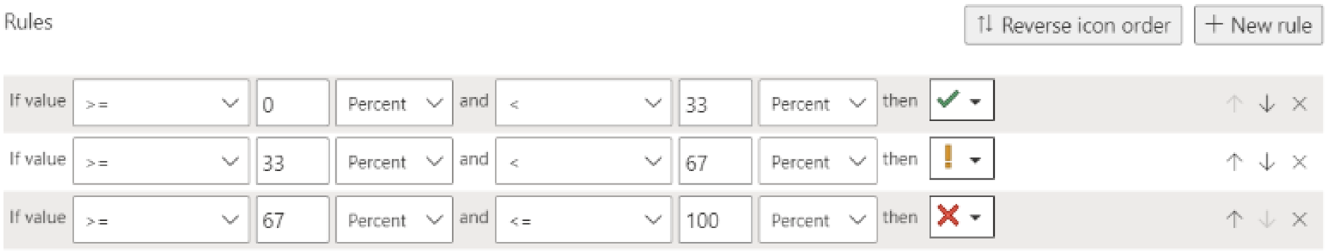
$$*\text{Rate of churn} = \frac{\text{churned customer}}{\text{total customer}}$$

The city can be further drilled down to zip-codes to see the same attributes at a granular level.



City	Total Customers	Did Not Churn	Churned	Rate of churn (%)	%GT Sum of Total Revenue
⊟ **Los Angeles**	**293**	**215**	**78**	**26.62**	**3.99%**
90028	6	1	5	✗ 83.33	0.14%
90003	5	4	1	✓ 20.00	0.04%
90004	5	3	2	✓ 40.00	0.04%
90006	5	4	1	✓ 20.00	0.03%
90007	5	3	2	✓ 40.00	0.06%
90008	5	5			0.06%
90011	5	4	1	✓ 20.00	0.10%
90012	5	4	1	✓ 20.00	0.05%
90013	5	3	2	✓ 40.00	0.07%
90015	5	4	1	✓ 20.00	0.08%
90018	5	5			0.09%
90021	5	5			0.12%
Total	**7043**	**5174**	**1869**	**26.54**	**100.00%**

Moreover there is conditional formatting on the Rate of churn column to indicate that as the rate reaches towards 100 the column background color changes to red to indicate that the region requires immediate attention.
Moreover the symbols indicate:

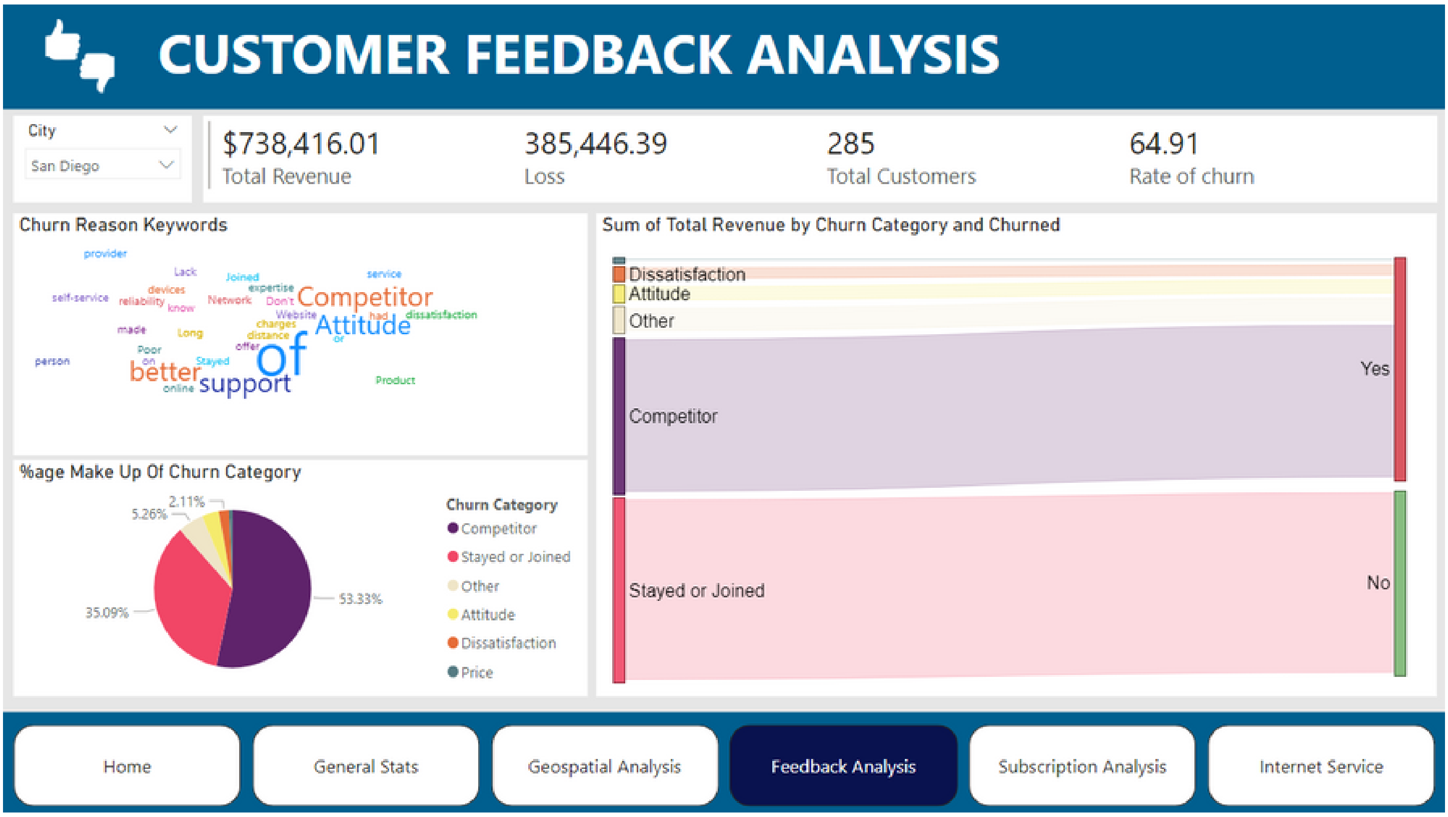


If value	>=	0	Percent	and	<	33	Percent	then	✓	↑ ↓ ✕
If value	>=	33	Percent	and	<	67	Percent	then	!	↑ ↓ ✕
If value	>=	67	Percent	and	<=	100	Percent	then	✗	↑ ↓ ✕

Key Findings:

01

Among the top ten revenue generating regions San Diego stands on 2nd with a contribution of 3.46% total revenue but at the same time has a high churn rate of 64.91% and hence require immediate attention otherwise the telecom revenue would take a massive hit.



CUSTOMER FEEDBACK ANALYSIS

City: San Diego

$738,416.01 Total Revenue

385,446.39 Loss

285 Total Customers

64.91 Rate of churn

The feedback page shows insights like; total revenue, loss that the company faced by the customers that left and how much they were contributing to the total revenue, total customers and the total churn rate.Moreover there is a word cloud that shows the most used words in the review given by the customers regarding why they left.

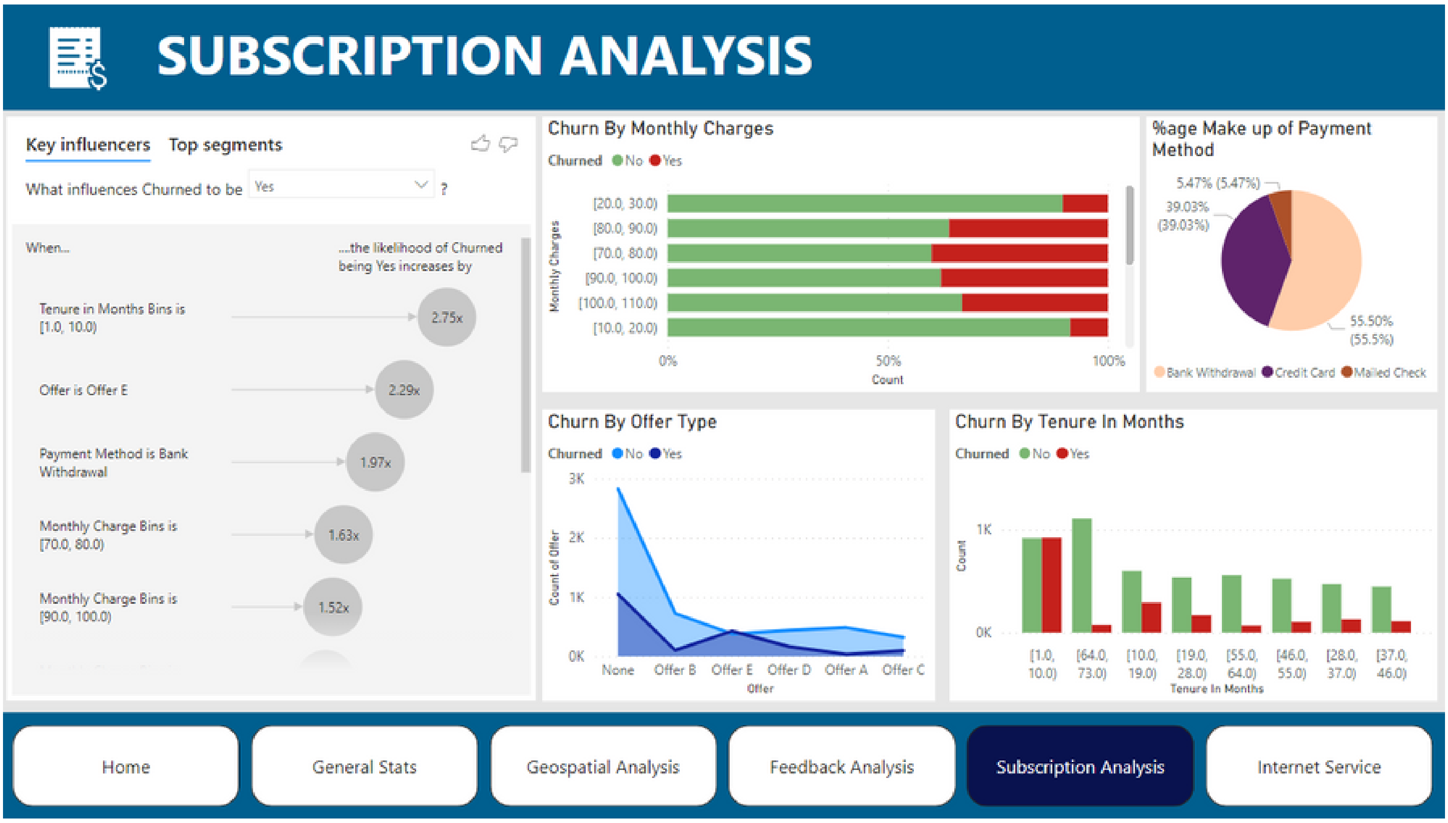
The Sankey chart shows the churn category attribute, relationship to churned attribute and how much revenue each category adds up to. There is also a filter present with city attribute in-order to gauge the popular categories of churn with respect to a particular city.

Finally the pie chart reveals the current churn category of each customer

Key Findings:

01

From the Sankey chart it is clear that telecom company is losing most of it's revenue to it's competitor, moreover the dissatisfaction and the attitude of the employees towards the customers are also decent contributors to the loss of revenue.



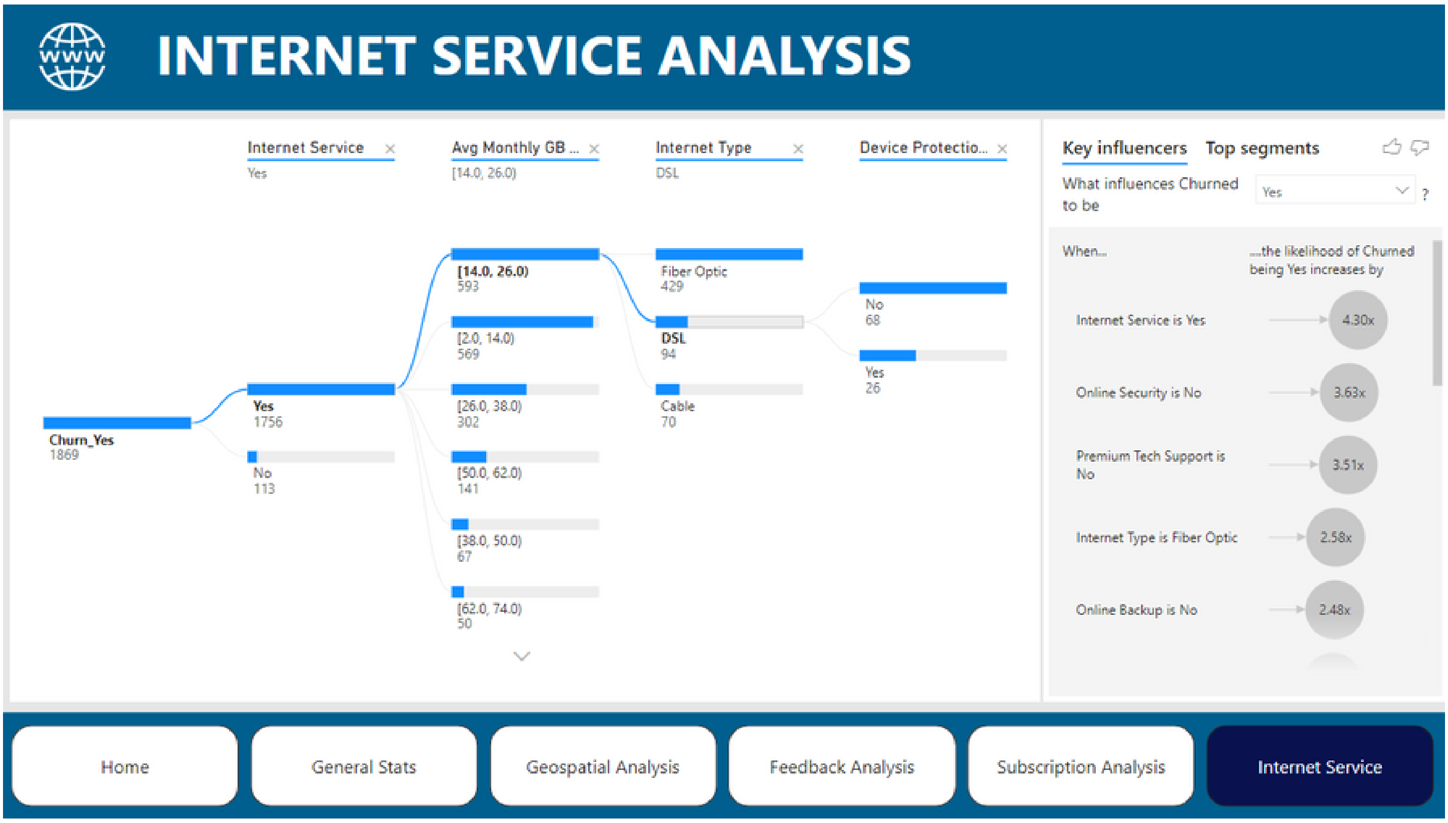
The subscription analysis page, apart from having the usual visuals also contain the key-influences visualization that helped in gaining an idea how a specific attribute affects the churn. The pie chart depicts the preferred method of payment by customers.

Key Findings:

01 First key insight here is that the people who are availing the offer E are most likely to churn also supported by the key-influencer chart which denotes that their churn rate would go up by 2.29 times.

02 Another interesting finding is tenure in months, if it is between 1 to 10 months then the probability of a customer leaving the company is quite high.

03 Finally individuals who have the payment method of bank withdrawal will likely churn.



The internet service page in addition to the key-segments chart consists of the decomposition tree that indicates for the customer who churned, how they were impacted by internet service and the add-ons that come with service.

Key Findings:

01

Upon investigation it was revealed people who avail the internet service are 4.3 times more likely to churn than others.

RECOMMENDATIONS

Following are the recommendations based on the key finding of the data:

01

Month-to-month contracts are prone to leave the company.

A possible solution for this to provide a deal on the long term plans hence persuading the customers to avail the longer option.

02

Age brackets of 75 to 82 are more likely to churn.

A senior citizen discount to this specific age group can be implemented, moreover the seniors are relatively less tech-savvy as compared younger generations hence setting up a designated helpline for them and sending in-person staff to solve their problems can also be taken into consideration.

03

Top ten revenue generating regions San Diego stands on 2nd with a contribution of 3.46% total revenue but at the same time has a high churn rate of 64.91%

Upon further investigation it was revealed that the telecom company is losing most of it's customers to competitors and the words revealed by world cloud are: “better support”, indicating that the company needs to establish better support in this region for retention of the customers.
Moreover further investigation is needed to analyze the competitors strategy

04

Telecom company is losing most of it’s revenue to it’s competitor, moreover the dissatisfaction and the attitude of the employees towards the customers are also decent contributors to the loss of revenue.

The dissatisfaction and attitude of the employees both are an end product of poor training, the recommendation would be review the employee training manual and a rigorous workshops on customer engagement and satisfaction to undo the harm caused.

05

Those who avail offer E are most likely to churn.

Investigate what services offer E entail and try to replicate the services included in other offers.

06

Who avail the internet service are 4.3 times more likely to churn than others

This is a clear indication of the dissatisfactory internet service provided and hence an investigation on why the quality is poor and how can a better quality internet can be provided to the customers is due.

CONCLUSION

The above report consists of detailed steps carried out in the churn analysis of the telecom company. The resources can be found using the following links.

Notebook created for data preprocessing: https://github.com/Shehryar-mallick/GDA-Project-Customer-Churn-Analysis/blob/main/GDA_Project_Preprocessing.ipynb

Preprocessed dataset: https://raw.githubusercontent.com/Shehryar-mallick/GDA-Project-Customer-Churn-Analysis/main/data/telecom_customer_churn_preprocessed_1.csv

GitHub repository: https://github.com/Shehryar-mallick/GDA-Project-Customer-Churn-Analysis

CONTACT

SHEHRYAR MALLICK

✉ shehryarmallick1@yahoo.com

https://www.linkedin.com/in/shehryar-mallick-5b706a194/

https://github.com/Shehryar-mallick

https://medium.com/@shehryarmallick28