Capstone Project 2

Telecom Customer Churn Analysis

Churn is a metric that shows customers who stop doing business with a company or a particular service, also known as customer attrition. By following this metric, what most businesses could do was try to understand the reason behind churn numbers and tackle those factors, with reactive action plans.

But what if you could know in advance that a specific customer is likely to leave your business, and have a chance to take proper actions in time to prevent it from happening?

The reasons that lead customers to the cancellation decision can be numerous, coming from poor service quality, delay on customer support, prices, new competitors entering the market, and so on. Usually, there is no single reason, but a combination of events that somehow culminated in customer displeasure.

If your company were not capable to identify these signals and take actions prior to the cancel button click, there is no turning back, your customer is already gone. But you still have something valuable: the data. Your customer left very good clues about where you left to be desired. It can be a valuable source for meaningful insights and to train customer churn models. Learn from the past, and have strategic information at hand to improve future experiences, it's all about machine learning.

When it comes to the telecommunications segment, there is great room for opportunities. The wealth and the amount of customer data that carriers collect can contribute a lot to shift from a reactive to a proactive position. The emergence of sophisticated artificial intelligence and data analytics techniques further help leverage this rich data to address churn in a much more effective manner.

About Data

Churn data for a fictional Telecommunications company that provides phone and internet services to 7,043 customers in California, and includes details about customer demographics, location, services, and current status. Customers' data has only two quarters with no data/monthly data.

Data Dictionary

Each row represents a customer, and each column contains the customer's attributes, as described below. The dataset includes information about:

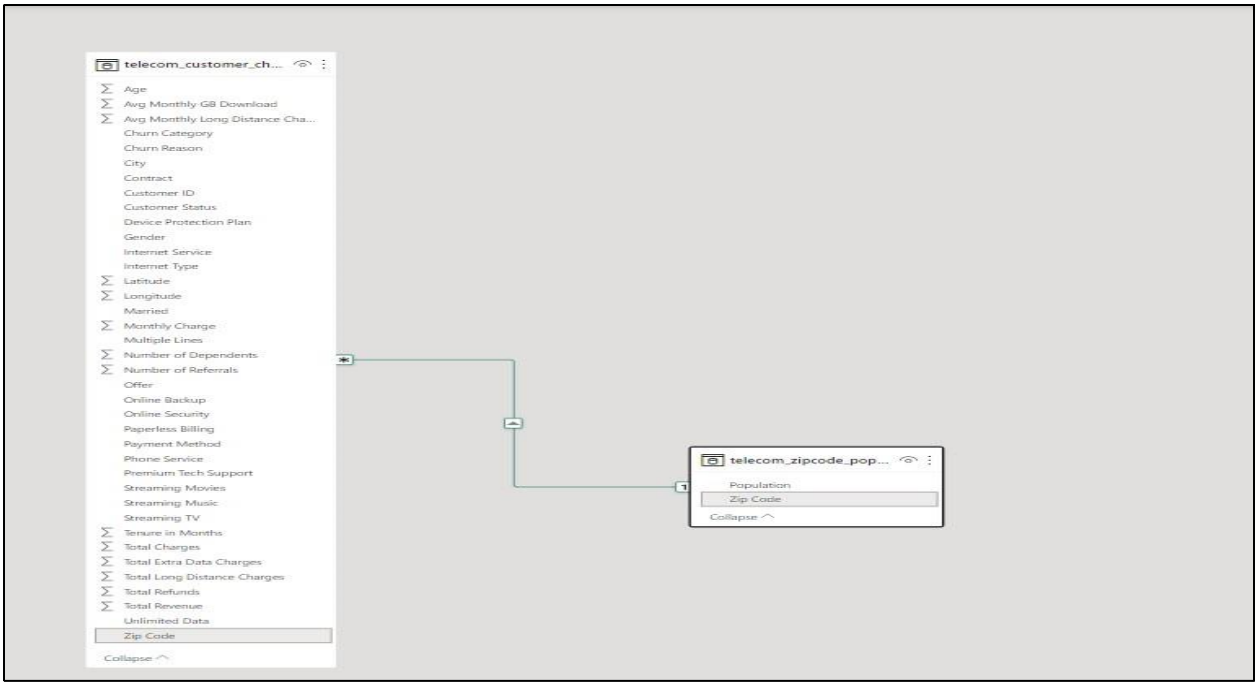
- Customers' demographic info:
    - `gender`: customer's gender: *Male*, *Female*
    - `SeniorCitizen`: customer is 65 or older: *1*, *0* (meaning *Yes* and *No*, respectively)
    - `Partner`: customer is married: *Yes*, *No*
    - `Dependents`: customer lives with any dependents: *Yes*, *No*. Dependents could be children, parents, grandparents, etc.
- Services that each customer has signed up for:
    - `PhoneService`: customer subscribes to home phone service with the company: *Yes*, *No*
    - `MultipleLines`: customer subscribes to multiple telephone lines with the company: *Yes*, *No*, *No internet service*
    - `InternetService`: customer subscribes to Internet service with the company: *No*, *DSL*, *Fiber Optic*
    - `OnlineSecurity`: customer subscribes to an additional online security service provided by the company: *Yes*, *No*, *No internet service*
    - `OnlineBackup`: customer subscribes to an additional online backup service provided by the company: *Yes*, *No*, *No internet service*
    - `DeviceProtection`: customer subscribes to an additional device protection plan for their Internet equipment provided by the company: *Yes*, *No*, *No internet service*
    - `TechSupport`: customer subscribes to an additional technical support plan from the company with reduced wait times: *Yes*, *No*, *No internet service*
    - `StreamingTV`: customer uses their Internet service to stream television programming from a third-party provider: *Yes*, *No*, *No internet service*
    - `StreamingMovies`: customer uses their Internet service to stream movies from a third-party provider: *Yes*, *No*, *No internet service*
- Customer account information:
    - `tenure`: total number of months that the customer has been with the company.
    - `Contract`: customer's current contract type: *Month-to-Month*, *One Year*, *Two Year*.
    - `PaperlessBilling`: customer has chosen paperless billing: *Yes*, *No*
    - `PaymentMethod`: how the customer pays their bill: *Electronic check*, *Credit Card*, *Mailed Check*, *Bank transfer*
    - `MonthlyCharge`: customer's current total monthly charge for all their services from the company
    - `TotalCharges`: customer's total charges, calculated to the end of the quarter
- Finally, each customer has a `CustomerID`, a unique ID that identifies the customer.

Load DataSet:

We load the both dataset “telecom_customer_churn.csv” and “telecom_zipcode_population.csv” by pandas library pd.read_csv.

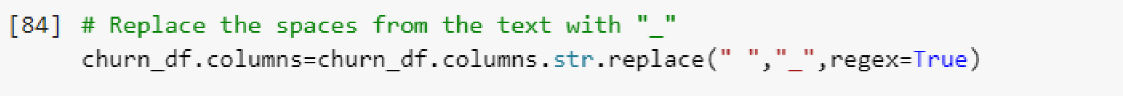
ERD Diagram:

Both the tables were connected with key “Zip_Codes”



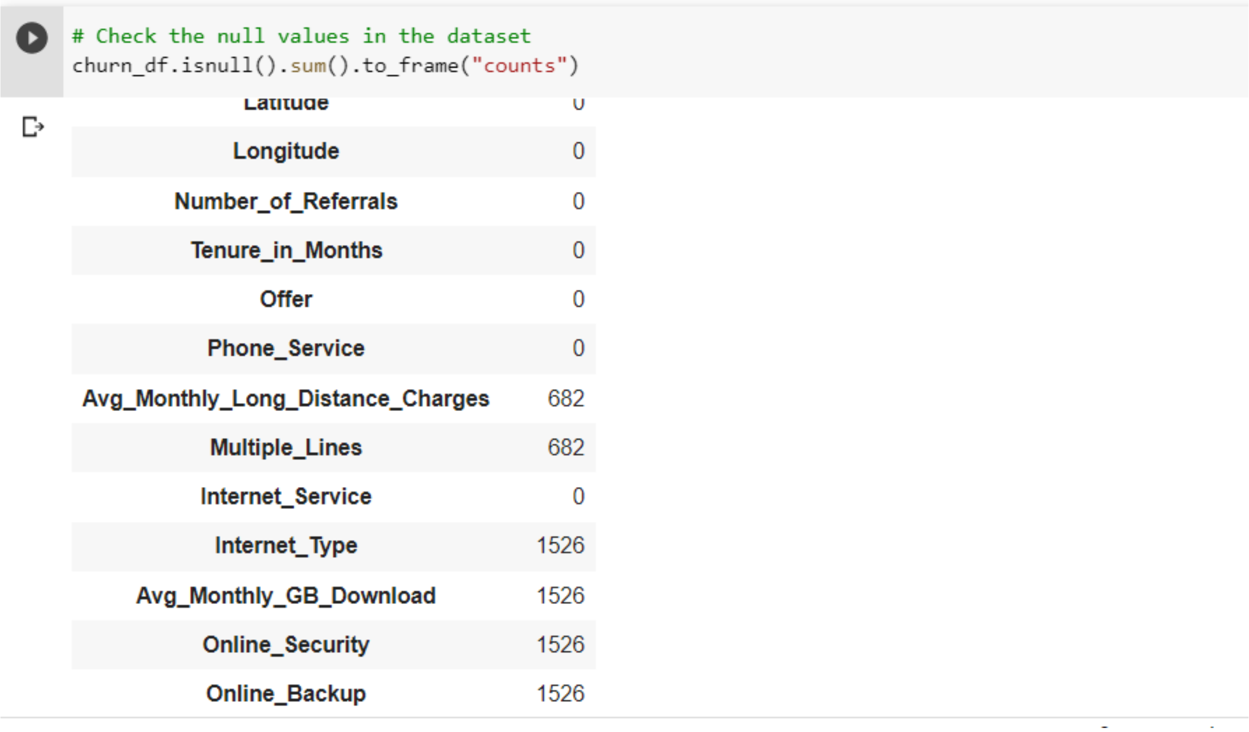
Exploratory Data Analysis & Data Cleaning:

Dataset contains 7043 rows and 39 columns which was checked by **churn_df.shape.** Then check data columns to ensure hat types of columns are there. In the column name, there as spaces between the words which was replace by “_”



```
[84] # Replace the spaces from the text with "_"
     churn_df.columns=churn_df.columns.str.replace(" ","_",regex=True)
```

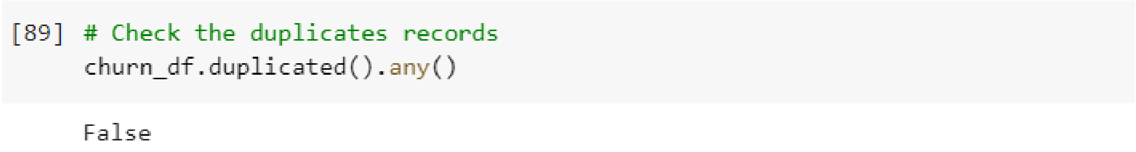
Checking the null values by using this code



```
# Check the null values in the dataset
churn_df.isnull().sum().to_frame("counts")
```

Latitude	0
Longitude	0
Number_of_Referrals	0
Tenure_in_Months	0
Offer	0
Phone_Service	0
Avg_Monthly_Long_Distance_Charges	682
Multiple_Lines	682
Internet_Service	0
Internet_Type	1526
Avg_Monthly_GB_Download	1526
Online_Security	1526
Online_Backup	1526

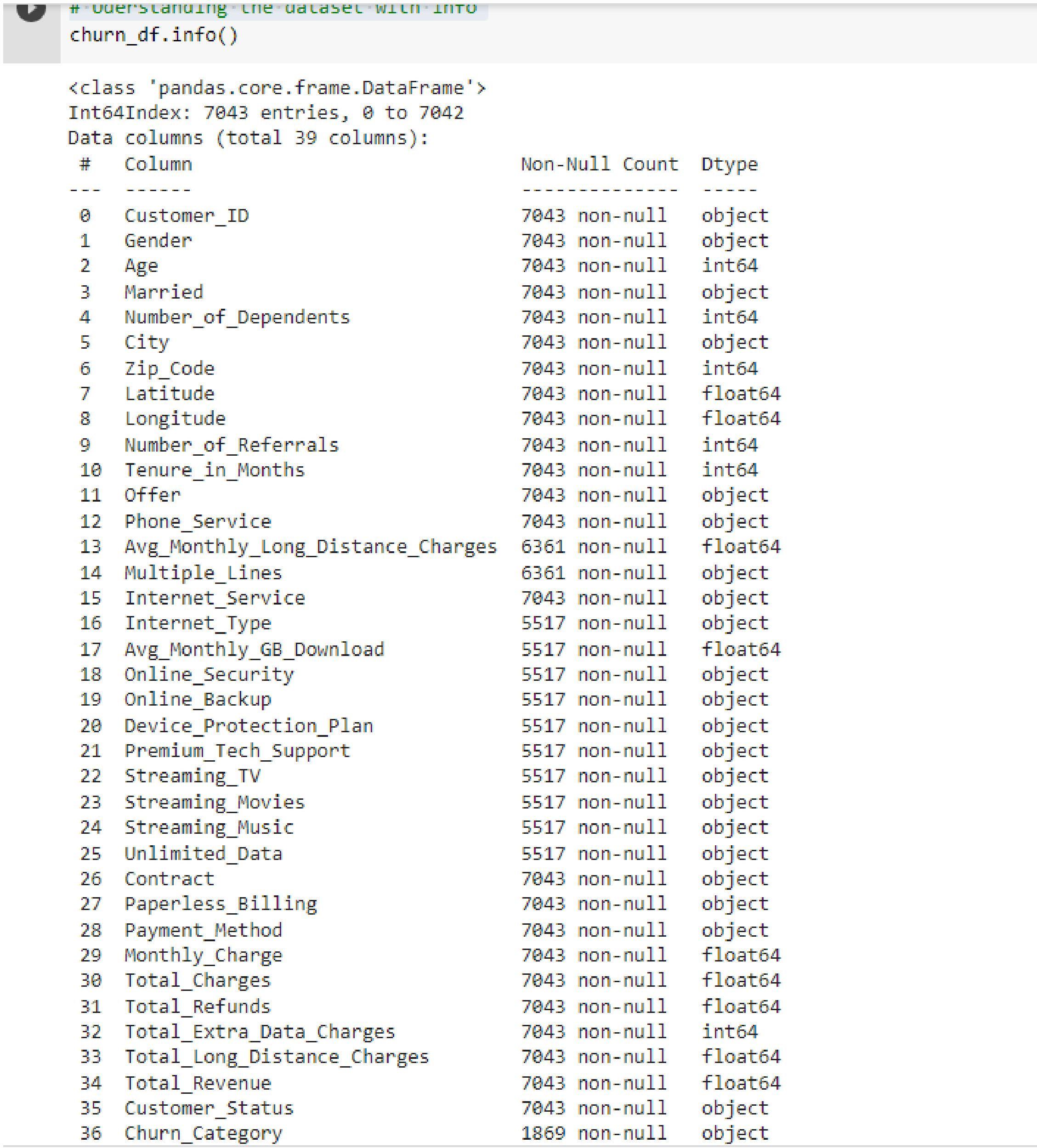
Checking the duplicate records. Our data is being checked to make sure there aren't any duplicate customers. There are no duplicate values found!.



```
[89] # Check the duplicates records
     churn_df.duplicated().any()

     False
```

Understanding the data with info where we got each column names with its datatypes and non null values.



```
# Uderstanding the dataset with info
churn_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7043 entries, 0 to 7042
Data columns (total 39 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Customer_ID                        7043 non-null   object
 1   Gender                             7043 non-null   object
 2   Age                                7043 non-null   int64
 3   Married                            7043 non-null   object
 4   Number_of_Dependents               7043 non-null   int64
 5   City                               7043 non-null   object
 6   Zip_Code                           7043 non-null   int64
 7   Latitude                           7043 non-null   float64
 8   Longitude                          7043 non-null   float64
 9   Number_of_Referrals                7043 non-null   int64
 10  Tenure_in_Months                   7043 non-null   int64
 11  Offer                              7043 non-null   object
 12  Phone_Service                      7043 non-null   object
 13  Avg_Monthly_Long_Distance_Charges  6361 non-null   float64
 14  Multiple_Lines                     6361 non-null   object
 15  Internet_Service                   7043 non-null   object
 16  Internet_Type                      5517 non-null   object
 17  Avg_Monthly_GB_Download            5517 non-null   float64
 18  Online_Security                    5517 non-null   object
 19  Online_Backup                      5517 non-null   object
 20  Device_Protection_Plan             5517 non-null   object
 21  Premium_Tech_Support               5517 non-null   object
 22  Streaming_TV                       5517 non-null   object
 23  Streaming_Movies                   5517 non-null   object
 24  Streaming_Music                    5517 non-null   object
 25  Unlimited_Data                     5517 non-null   object
 26  Contract                           7043 non-null   object
 27  Paperless_Billing                  7043 non-null   object
 28  Payment_Method                     7043 non-null   object
 29  Monthly_Charge                     7043 non-null   float64
 30  Total_Charges                      7043 non-null   float64
 31  Total_Refunds                      7043 non-null   float64
 32  Total_Extra_Data_Charges           7043 non-null   int64
 33  Total_Long_Distance_Charges        7043 non-null   float64
 34  Total_Revenue                      7043 non-null   float64
 35  Customer_Status                    7043 non-null   object
 36  Churn_Category                     1869 non-null   object
```

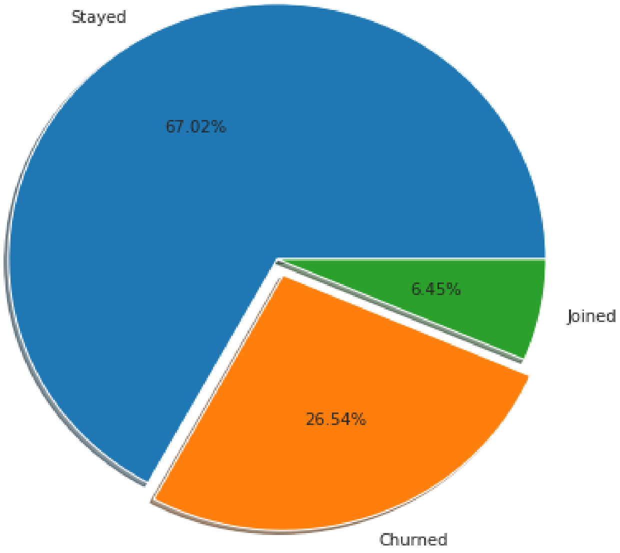
There was total 11 columns which had a null values, so do so. Multiple_Lines blank values replaced by No_Phone_Services. While "Internet_Type","Avg_Monthly_GB_Download","Online_Security","Online_Backup","Device_Protection_Plan","Premium_Tech_Support","Streaming_TV","Streaming_Movies","Streaming_Music","Unlimited_Data" as replaced by “No_Internet_Services”. Churn_Category & Churn_Reasons replaced by “No_churn”.After cleaning we saved the dataset.

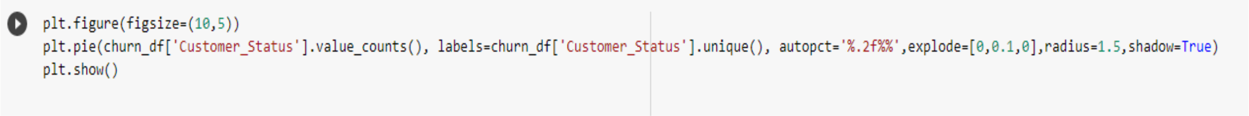
The categorical & numerical column length were 23 & 15.

Business Problems:

1.What were the customer churn & retention rates?

- We note that a big proportion of our clients did not abandon the services. Therefore, only **26.54% (1869)** of the customers was churned.
- **Churn Rate**= 1869/7043 = 26.54%
- Retention Rate = 1 - Churn Rate = 100 - 26.54 = 73.46%

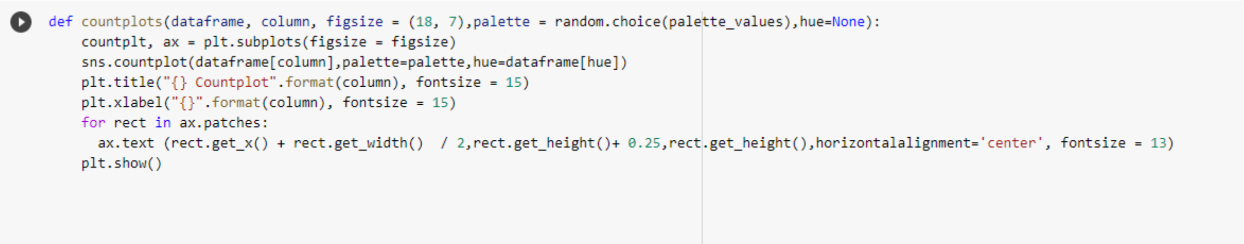




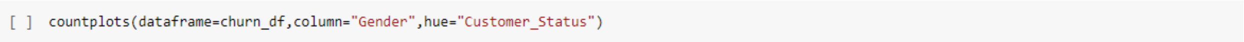
```
plt.figure(figsize=(10,5))
plt.pie(churn_df['Customer_Status'].value_counts(), labels=churn_df['Customer_Status'].unique(), autopct='%.2f%%',explode=[0,0.1,0],radius=1.5,shadow=True)
plt.show()
```

2. What was the customer churn pattern according to Gender?

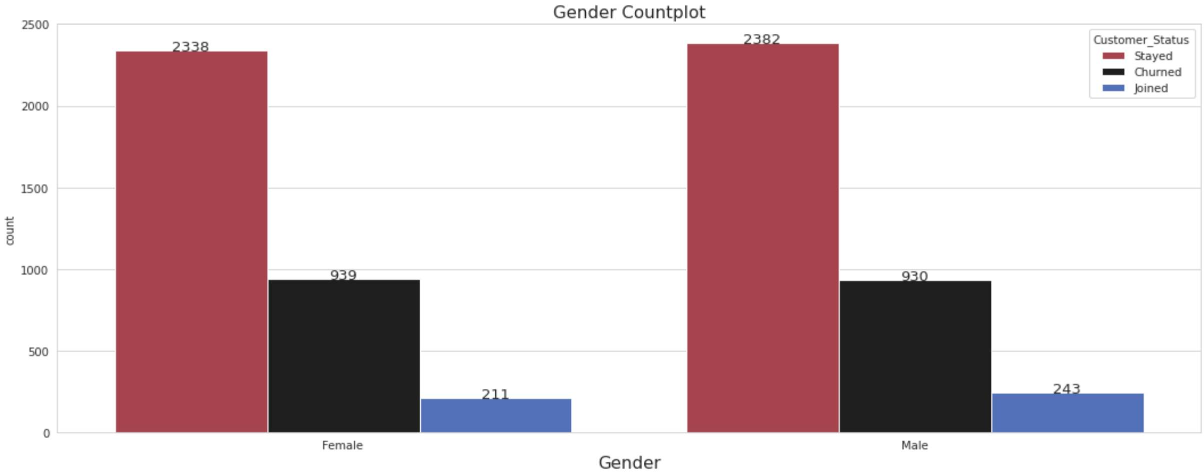
From the chart we can surely depict that same number of both genders were moved toward attrition. So there are equal numbers of genders contributing toward churn.



```
def countplots(dataframe, column, figsize = (18, 7),palette = random.choice(palette_values),hue=None):
    countplt, ax = plt.subplots(figsize = figsize)
    sns.countplot(dataframe[column],palette=palette,hue=dataframe[hue])
    plt.title("{} Countplot".format(column), fontsize = 15)
    plt.xlabel("{}".format(column), fontsize = 15)
    for rect in ax.patches:
      ax.text (rect.get_x() + rect.get_width()  / 2,rect.get_height()+ 0.25,rect.get_height(),horizontalalignment='center', fontsize = 13)
    plt.show()
```

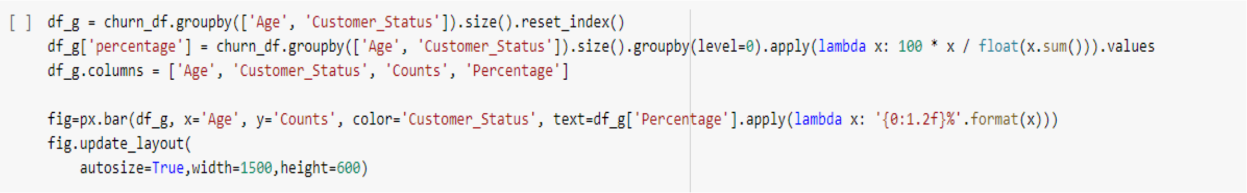


```
[ ] countplots(dataframe=churn_df,column="Gender",hue="Customer_Status")
```



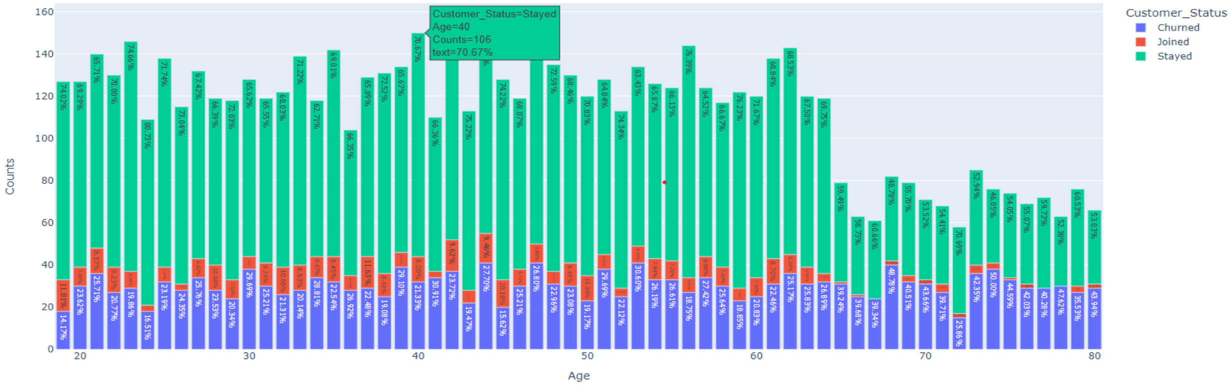
3. What was the behavior of churn customers with age?

According to the graph, interest in telecommunications services declines after the age of 65, and California has a high attrition rate across the board.



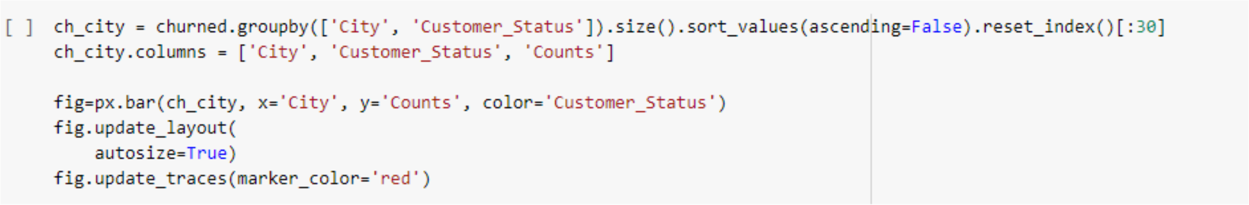
```
[ ] df_g = churn_df.groupby(['Age', 'Customer_Status']).size().reset_index()
    df_g['percentage'] = churn_df.groupby(['Age', 'Customer_Status']).size().groupby(level=0).apply(lambda x: 100 * x / float(x.sum())).values
    df_g.columns = ['Age', 'Customer_Status', 'Counts', 'Percentage']

    fig=px.bar(df_g, x='Age', y='Counts', color='Customer_Status', text=df_g['Percentage'].apply(lambda x: '{0:1.2f}%'.format(x)))
    fig.update_layout(
        autosize=True,width=1500,height=600)
```



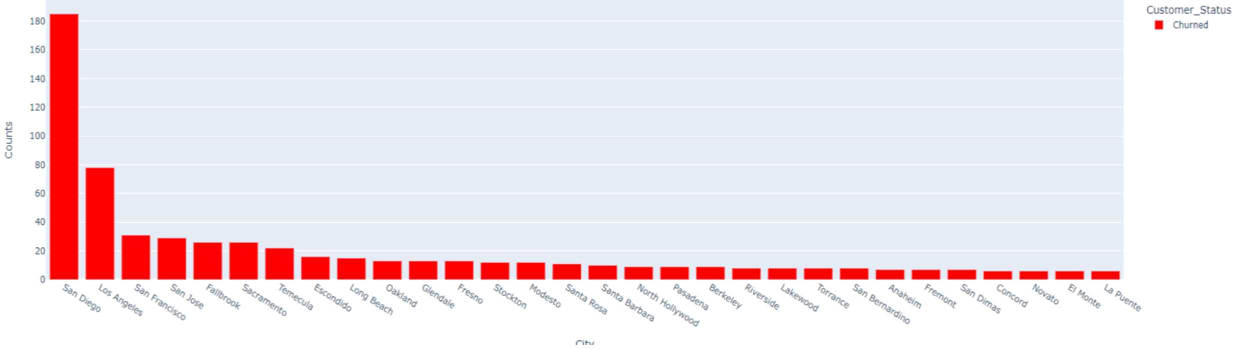
4. Which city has the most and least churn rate in California state?

We can surely depicted that San Diego is the most prominent churned city. Around 777 are the most low churn cities.



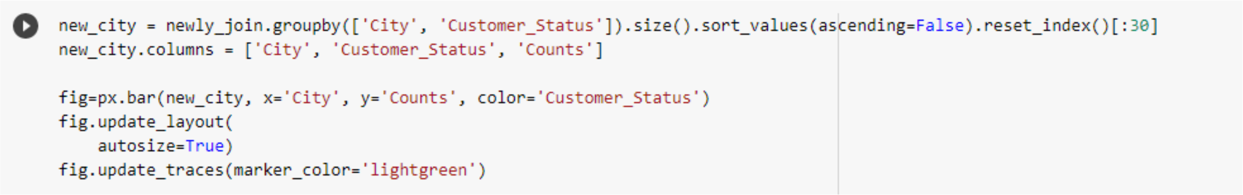
```
[ ]  ch_city = churned.groupby(['City', 'Customer_Status']).size().sort_values(ascending=False).reset_index()[:30]
     ch_city.columns = ['City', 'Customer_Status', 'Counts']

     fig=px.bar(ch_city, x='City', y='Counts', color='Customer_Status')
     fig.update_layout(
         autosize=True)
     fig.update_traces(marker_color='red')
```



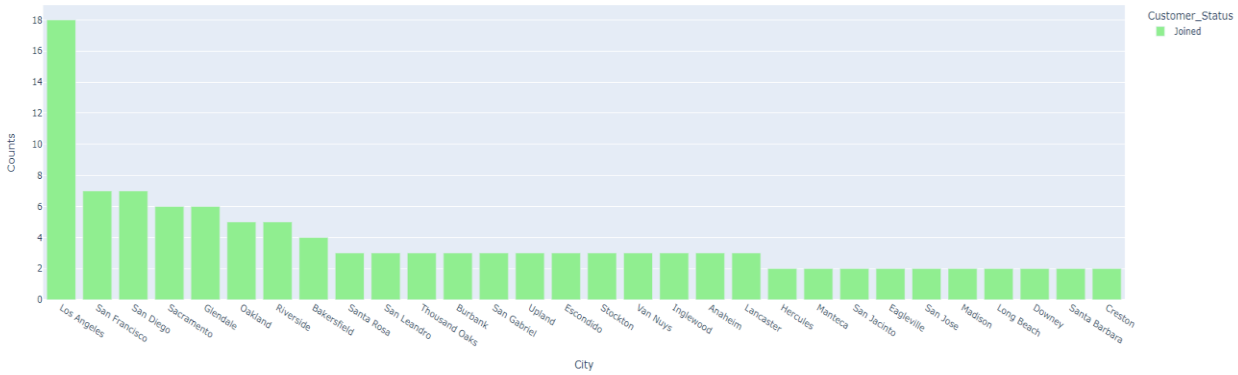
5. How many new clients did the business welcome in the most recent quarter? How much more from there?

It is shown that Los Angeles is the most demanding city where new customers are newly joined. On the other side second one is San Francisco and San Diego.



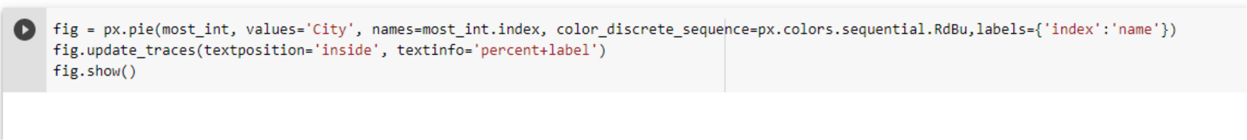
```python
new_city = newly_join.groupby(['City', 'Customer_Status']).size().sort_values(ascending=False).reset_index()[:30]
new_city.columns = ['City', 'Customer_Status', 'Counts']

fig=px.bar(new_city, x='City', y='Counts', color='Customer_Status')
fig.update_layout(
    autosize=True)
fig.update_traces(marker_color='lightgreen')
```

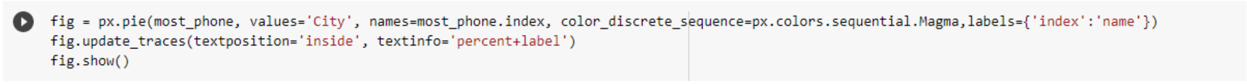


6. Which city has the most and least phone/internet churn rate in California state?

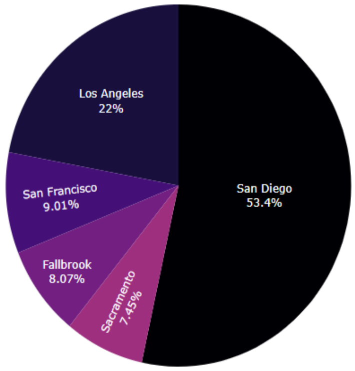
Both San Diego and Los Angeles provide their customers both services, whether we are talking about the best or the worst. Least phone/ internet churn counts are 736/740.

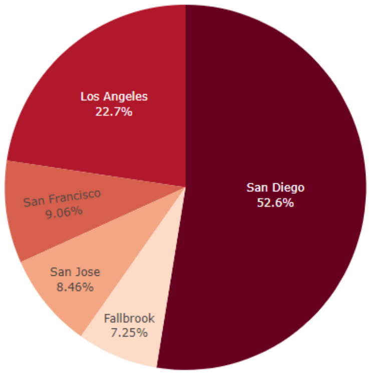


```python
fig = px.pie(most_int, values='City', names=most_int.index, color_discrete_sequence=px.colors.sequential.RdBu,labels={'index':'name'})
fig.update_traces(textposition='inside', textinfo='percent+label')
fig.show()
```



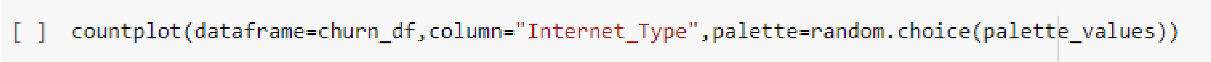
```python
fig = px.pie(most_phone, values='City', names=most_phone.index, color_discrete_sequence=px.colors.sequential.Magma,labels={'index':'name'})
fig.update_traces(textposition='inside', textinfo='percent+label')
fig.show()
```



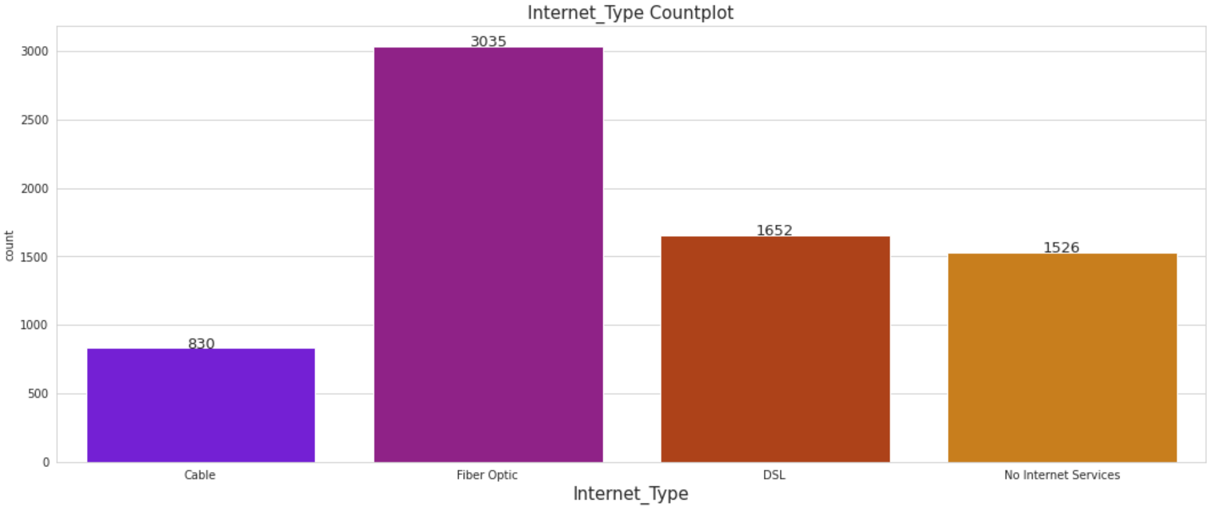


7. Which internet services are people utilizing the most frequently? Where?

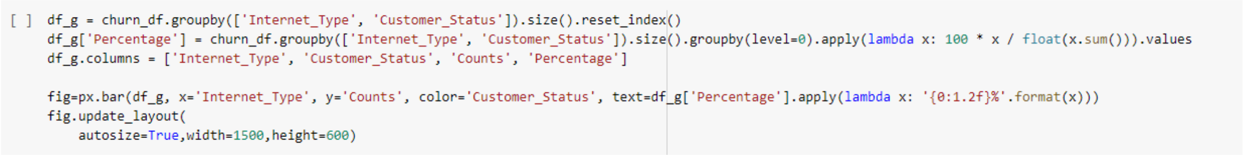
Fiber Optic is the leading internet service around 41% (3035 customers) of the revenue is generated by this service, where as churn rate is almost equal to that.



```
[ ]  countplot(dataframe=churn_df,column="Internet_Type",palette=random.choice(palette_values))
```

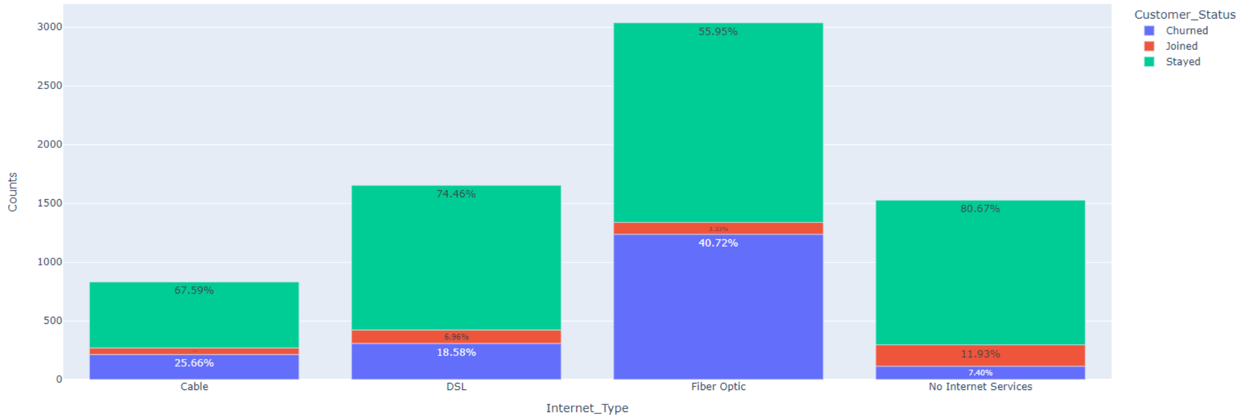


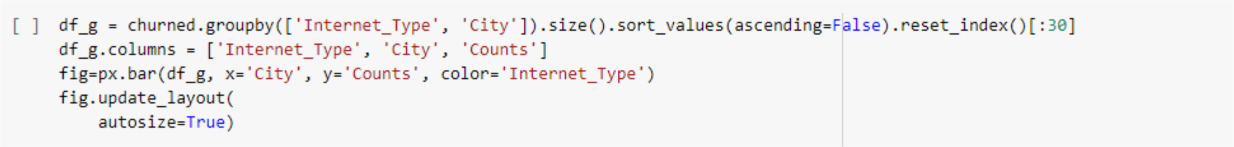
Fiber Optic is the leading internet service around 41% (3035 customers) of the revenue is generated by this service, where as churn rate is almost equal to that.. Fiber Optics is making a name for itself in San Diego and Los Angeles.



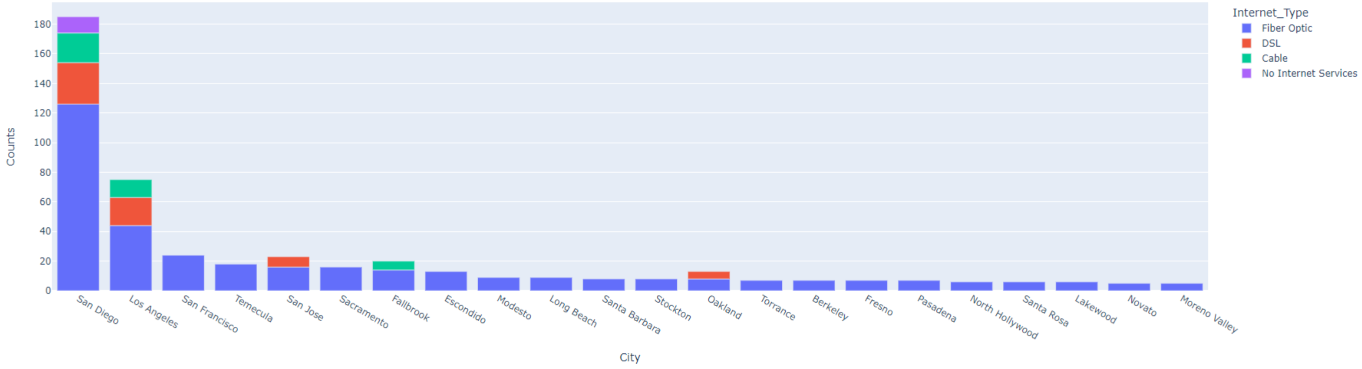
```
[ ]  df_g = churn_df.groupby(['Internet_Type', 'Customer_Status']).size().reset_index()
     df_g['Percentage'] = churn_df.groupby(['Internet_Type', 'Customer_Status']).size().groupby(level=0).apply(lambda x: 100 * x / float(x.sum())).values
     df_g.columns = ['Internet_Type', 'Customer_Status', 'Counts', 'Percentage']

     fig=px.bar(df_g, x='Internet_Type', y='Counts', color='Customer_Status', text=df_g['Percentage'].apply(lambda x: '{0:1.2f}%'.format(x)))
     fig.update_layout(
         autosize=True,width=1500,height=600)
```



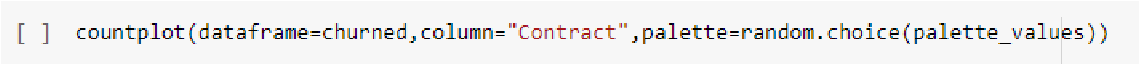


```
[ ]  df_g = churned.groupby(['Internet_Type', 'City']).size().sort_values(ascending=False).reset_index()[:30]
     df_g.columns = ['Internet_Type', 'City', 'Counts']
     fig=px.bar(df_g, x='City', y='Counts', color='Internet_Type')
     fig.update_layout(
         autosize=True)
```

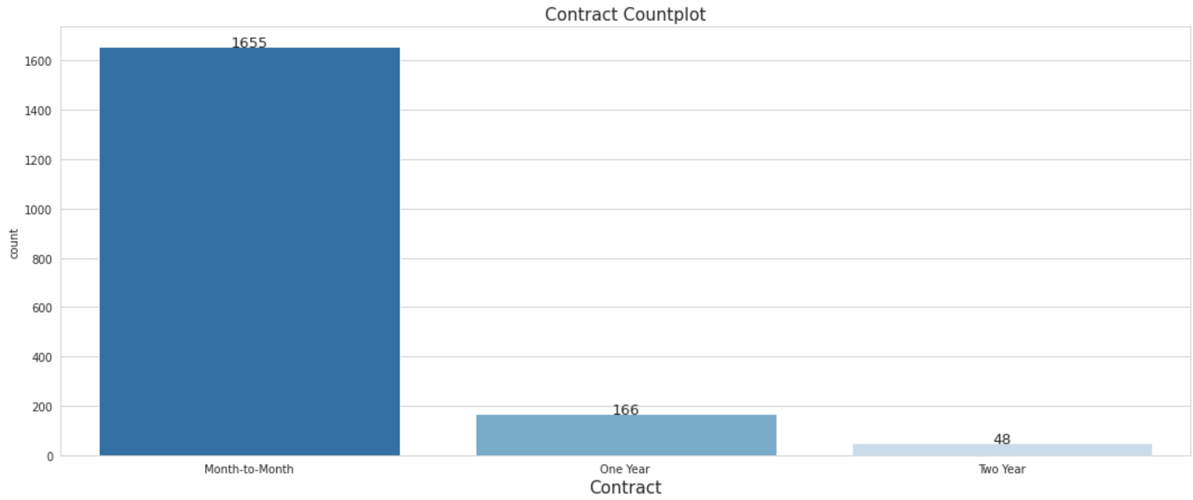


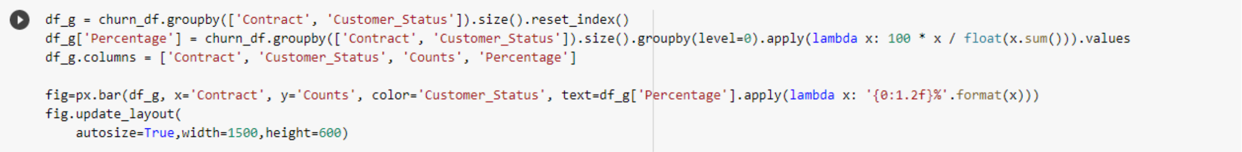
8. What are the type of contracts that contributes the customer churn? And which city are up on the board?

Customers can conveniently pay their costs under month-to-month contracts. A large percent of customers with monthly subscription have left when compared to customers with one or two years.



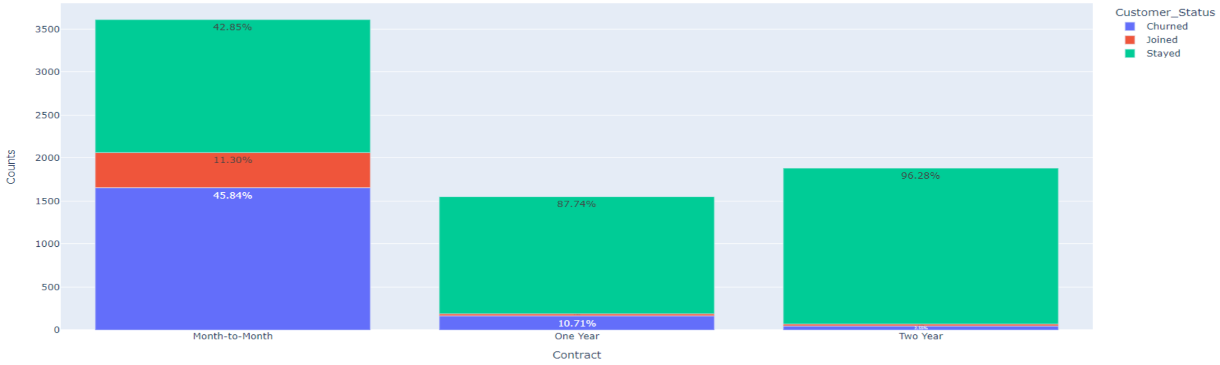
```
[ ]  countplot(dataframe=churned,column="Contract",palette=random.choice(palette_values))
```

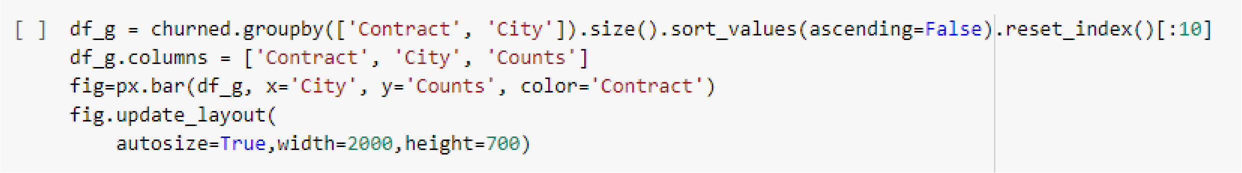




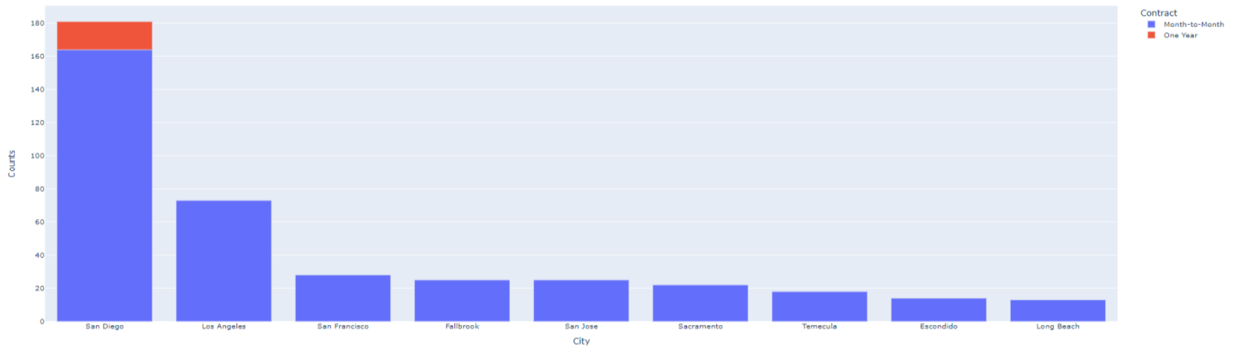
```
df_g = churn_df.groupby(['Contract', 'Customer_Status']).size().reset_index()
df_g['Percentage'] = churn_df.groupby(['Contract', 'Customer_Status']).size().groupby(level=0).apply(lambda x: 100 * x / float(x.sum())).values
df_g.columns = ['Contract', 'Customer_Status', 'Counts', 'Percentage']

fig=px.bar(df_g, x='Contract', y='Counts', color='Customer_Status', text=df_g['Percentage'].apply(lambda x: '{0:1.2f}%'.format(x)))
fig.update_layout(
    autosize=True,width=1500,height=600)
```

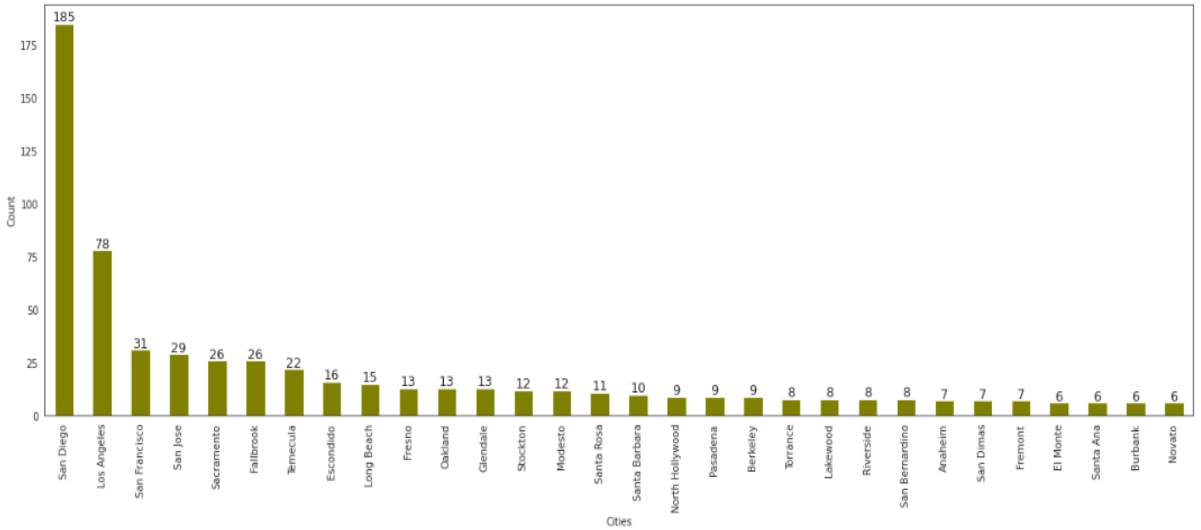




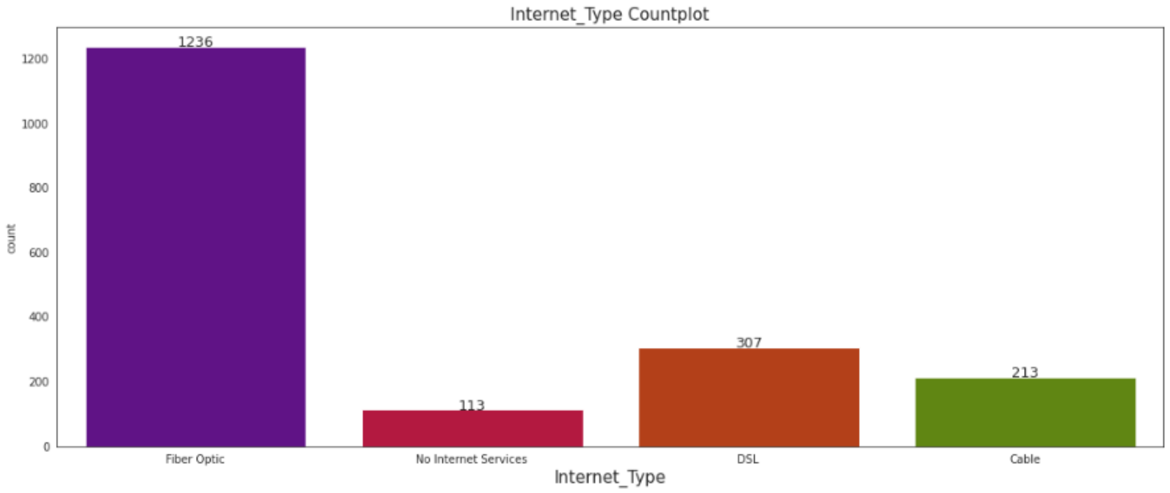
```
df_g = churned.groupby(['Contract', 'City']).size().sort_values(ascending=False).reset_index()[:10]
df_g.columns = ['Contract', 'City', 'Counts']
fig=px.bar(df_g, x='City', y='Counts', color='Contract')
fig.update_layout(
    autosize=True,width=2000,height=700)
```

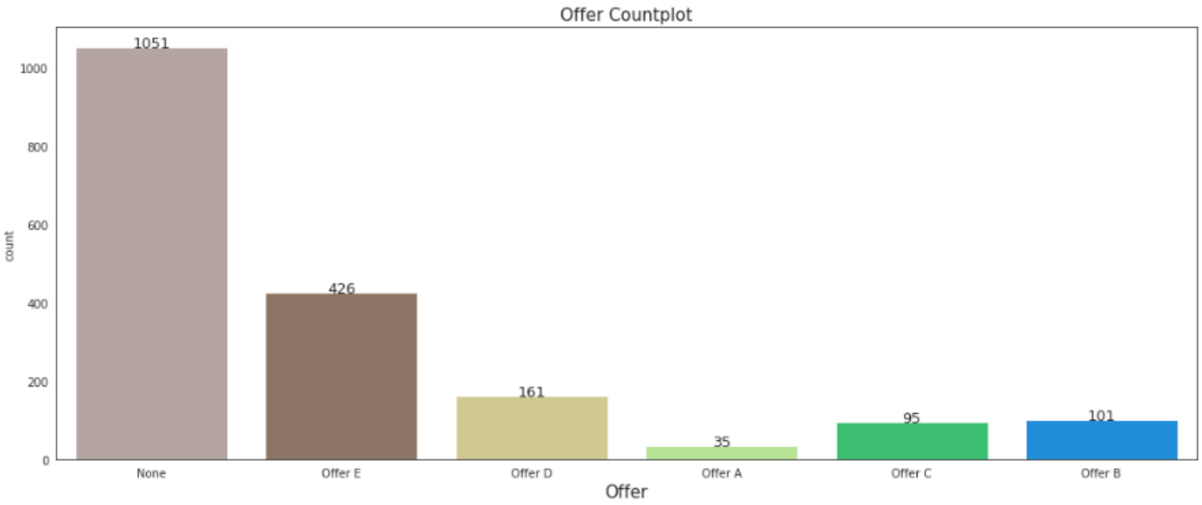


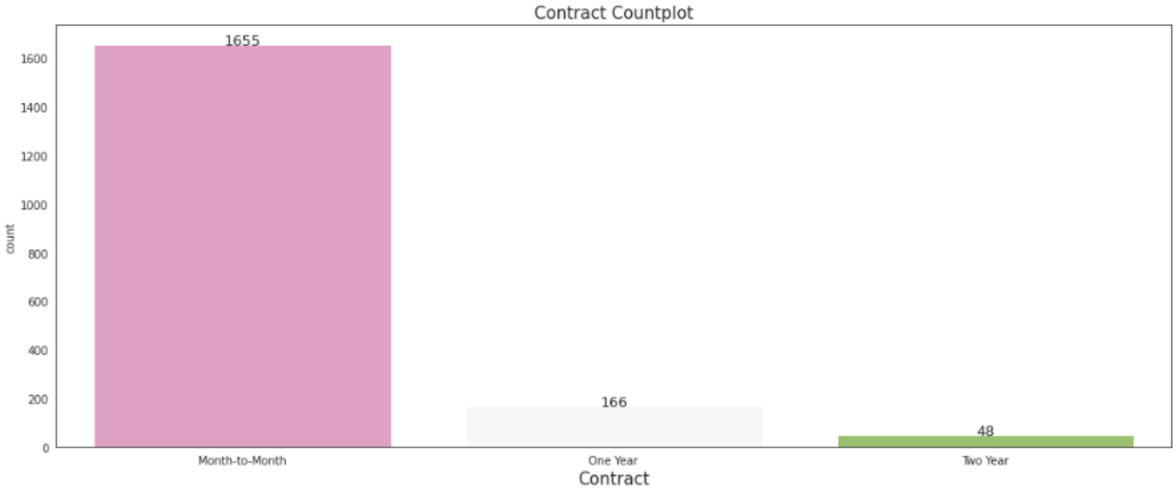
9. What seems to be the key drivers of customer churn?

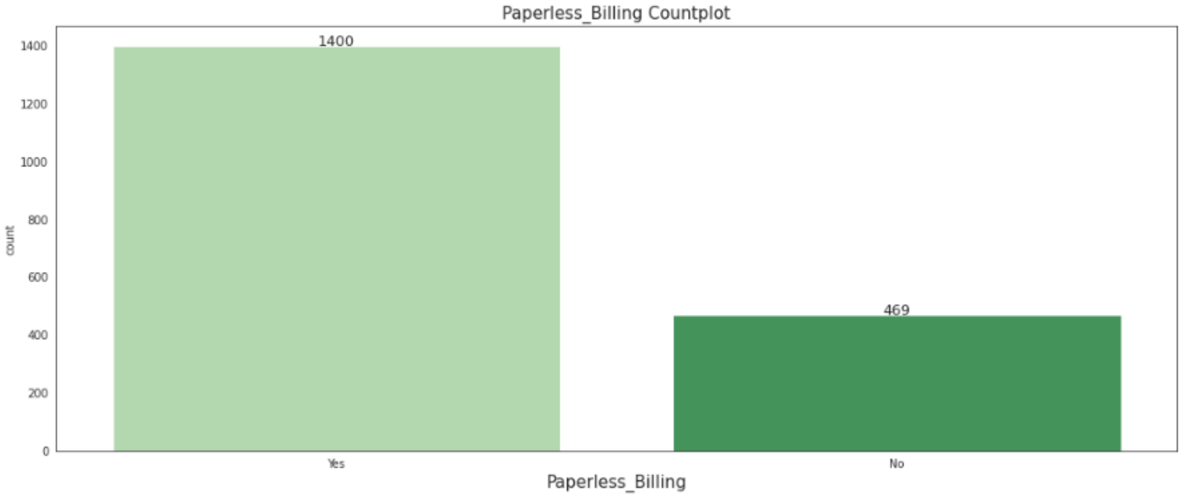


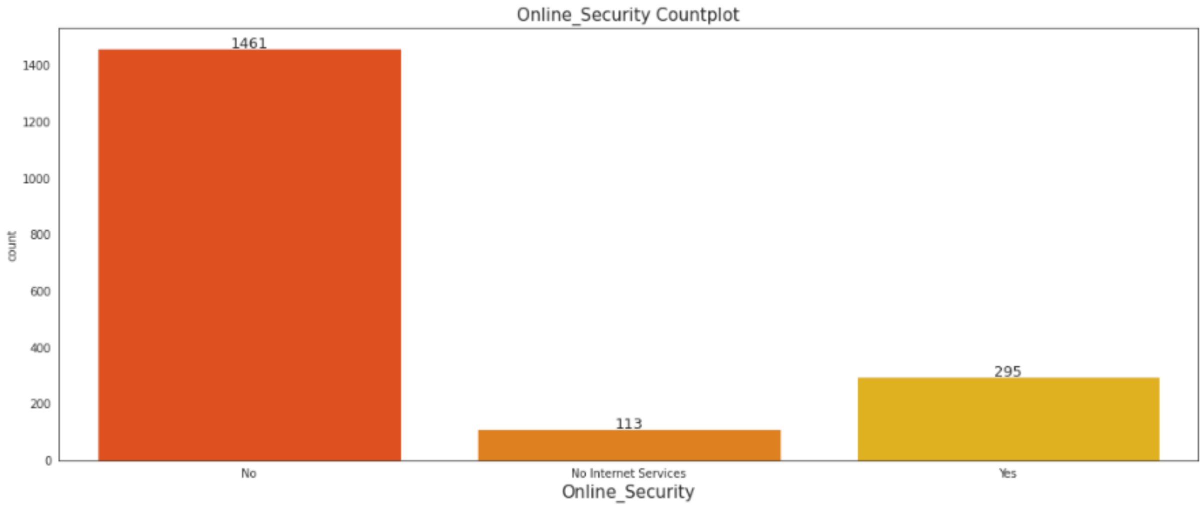
+ Code + Text

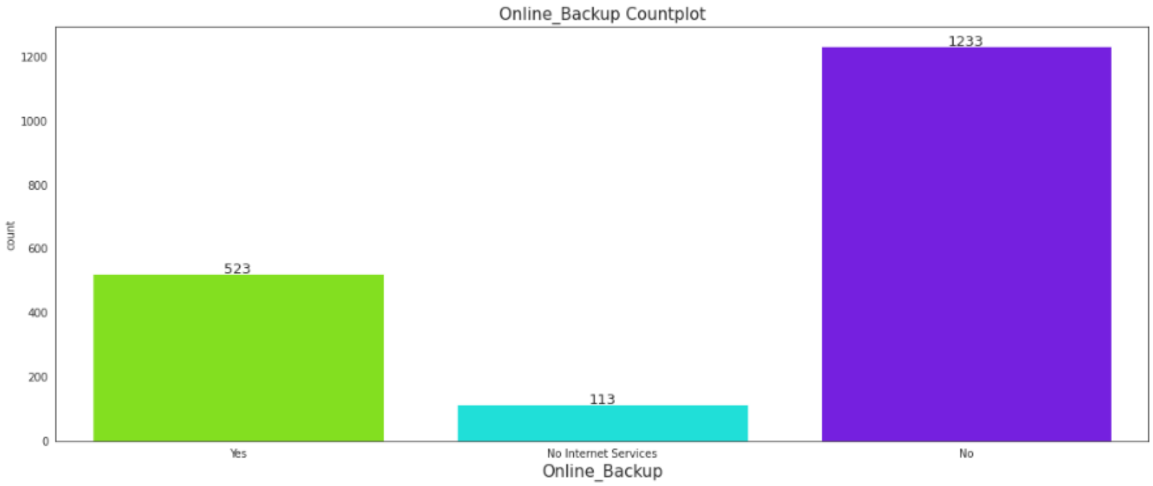



Offer Countplot
count
0
200
400
600
800
1000
1051
426
161
35
95
101
None
Offer E
Offer D
Offer A
Offer C
Offer B
Offer


Contract Countplot
count
0
200
400
600
800
1000
1200
1400
1600
1655
166
48
Month-to-Month
One Year
Two Year
Contract

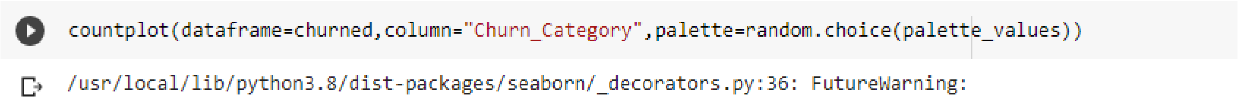

Paperless_Billing Countplot
count
0
200
400
600
800
1000
1200
1400
1400
469
Yes
No
Paperless_Billing





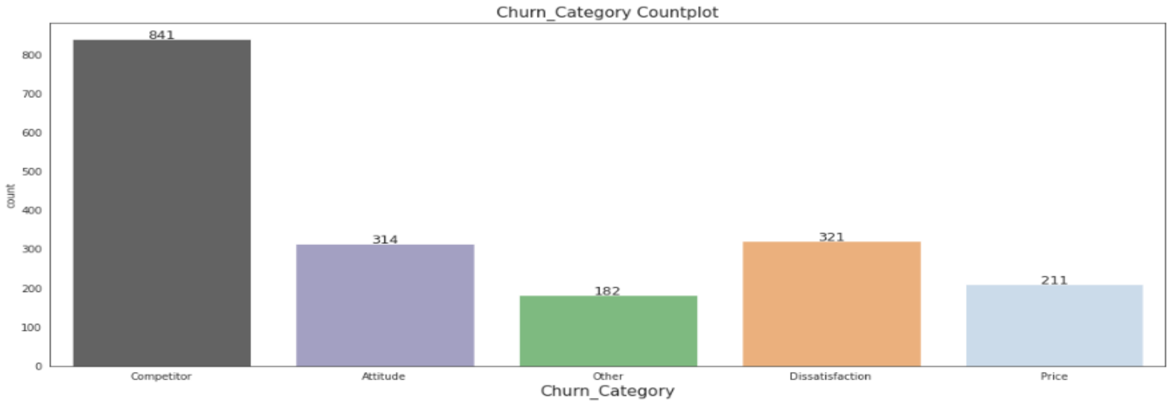
10. Which specific factors cause consumers to leave?

The primary cause of churn in the firm is competitors.



```
countplot(dataframe=churned,column="Churn_Category",palette=random.choice(palette_values))
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning:
```



Revenue Analysis:

- 31% of total revenue was lost by people who left.
- Optical Fiber is responsible for 53% ($16899k) of the monthly revenue, DSL 37% ($11814k)
- 87% of all monthly revenue lost are caused by customers with Month-to-Month contracts.

Recommendation/Limitations:

Target more on young and middle-aged customers since they are more likely to adopt modern technology and have the budget to enjoy.

Offer more discount for the customers who decide to choose the one year or two-year contract so that more customers will be bound with the contract.

Consider an overall discount since the price is always one of the major factors for customers to choose among existing incumbents.

The number of observations is decent, but if we could have more columns of features like the customers' geographic location, competitor's information, and other important factors, we could draw more insights from the result.

The nature of our dataset is a cross-sectional dataset. This means that there are no time series factors inside it.