Project Summary

We are majorly performing two things in our project namely Customer Segmentation and their corresponding Attrition Rate. Their importance lies in the fact that the Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers (and potential customers), a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators. Since the marketer's goal is usually to maximize the value (revenue and/or profit) from each customer, it is critical to know in advance how any particular marketing action will influence the customer. On the other hand, Customer churn is one of the most important metrics for a growing business to evaluate. While it's not the happiest measure, it's a number that can give your company the hard truth about its customer retention. It's important because it costs more to acquire new customers than it does to retain existing customers. In fact, an increase in customer retention of just 5% can create at least a 25% increase in profit. This is because returning customers will likely spend 67% more on your company's products and services. As a result, your company can spend less on the operating costs of having to acquire new customers. You don't need to spend time and money on convincing an existing customer to select your company over competitors because they've already made that decision. Again, it might seem like a 5% churn rate is solid and healthy. You can still make a vast revenue with that churn rate.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Topic Brief

Customer experience is fast becoming the number one priority for businesses all over the globe. In fact, according to one study, over 80% of organizations expected to compete mainly on customer experience in 2019. The logic is simple. Consumers don't want to give their money to businesses that treat them badly. However, the focus is now shifting and instead of just "not bad", customers want an excellent experience. Customers remember highly positive and frictionless interactions with businesses, and they will continue to seek that experience the next time they think of making a purchase.

A ton of factors contribute to providing a competitive and high-quality customer experience but knowing your customers and how to reach them is the vital first step. This is where the need for customer segmentation and customer churn analysis comes in.

Customer Segmentation - Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach to the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

Customer Churn - Customer churn (also known as customer attrition) refers to when a customer (player, subscriber, user, etc.) ceases his or her relationship with a company. Online businesses typically treat a customer as churned once a particular amount of time has elapsed since the customer's last interaction with the site or service. The full cost of churn includes both lost revenue and the marketing costs involved with replacing those customers with new ones.

Reducing churn is a key business goal of every online business.

Purpose of the Study

- It will help in identifying the most potential customers.
- It will help managers to easily communicate with a targeted group of the audience.
- Also, help in selecting the best medium for communicating with the targeted segment.
- It improves the quality of service, loyalty, and retention.
- Improve customer relationships via better understanding of the needs of segments.
- It provides opportunities for upselling and cross-selling.
- It will help managers to design special offers for targeted customers, to encourage them to buy more products.
- It helps companies to stay a step ahead of competitors.
- It also helps in identifying new products that customers could be interested in.

Significance of the Study

1. **More Customer Retention**

A personalized connection with your customers will help you to win satisfied customers. About 75% of satisfied customers are more likely to remain with that organization who regularly meet up their needs.

2. Enhances Competitiveness

The more customer retention, the more will be revenue generation. And all of this will enhance your competitiveness in the market. If you segment up your market, you are well known to your customers & accordingly. You will work to give them the expected result.

3. Establishes Brand Identity

By segmenting your customers, you can make them well aware of your brand. Identifying your brand will help your customers to directly engage with your products. This will increase your goodwill in the market & a brand value will be established among your other competitors.

4. Better Customer Relationship

Better Customer segmentation will lead to developing a better relationship with your prospective customers. This will leave for developing a better relationship for your client base. Communicate with your customers on their queries & requests. Regular communication will make you update about the new upcoming changes & opportunities to be availed of.

5. Leads to Price Optimization

Try to understand the social and financial status of your customers. This data will assist you to pace up with the price optimization accordingly. Learning “How to price a product?” is equivalent support to boost business when segmenting customers on the basis of their spending psychology.

6. Best Economies of Scale

Customer segmentation helps in better allocation of resources which in return helps you to gain economies of scale. Economies of scale mean when you are able to achieve your desired goal and that too at the most efficient cost. This can be done through customer segmentation. Therefore, try to focus on a limited number of resources which will help to serve with resources.

7. Improves Channel of Distribution

Customer segmentation will improve your channel of distribution. By knowing the right number of customers you can direct the right channel of distribution. This will remove confusion among your team members about whom they have to deliver the product & at what time.You will be

able to choose a precise channel of distribution & that too at minimal cost which eventually serves the best needs of your customers.

Method of Procedure

We have used the telecom customer dataset from Kaggle. Our dataset has 7043 rows and 20 columns.

Each row represents a customer, each column contains customer's attributes described on the column Metadata.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

customerID	gender	SeniorCitizen	Partner	Dependent	tenure	PhoneServ	MultipleLin	InternetS	OnlineSec	OnlineBac	DevicePro	TechSuppo	Streaming	StreamingT	Contract	PaperlessB	PaymentMethod	MonthlyCharge	TotalCharges	Churn
7590-VHVE	Female	0	Yes	No	1	No	No phone s	DSL	No	Yes	No	No	No	No	Month-to-m	Yes	Electronic check	29.85	29.85	No
5575-GNVD	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
3668-QPYB	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-m	Yes	Mailed check	53.85	108.15	Yes
7795-CFOC	Male	0	No	No	45	No	No phone s	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (aut	42.3	1840.75	No
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber opt	No	No	No	No	No	No	Month-to-m	Yes	Electronic check	70.7	151.65	Yes
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber opt	No	No	Yes	No	Yes	Yes	Month-to-m	Yes	Electronic check	99.65	820.5	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber opt	No	Yes	No	No	Yes	No	Month-to-m	Yes	Credit card (auton	89.1	1949.4	No
6713-OKOM	Female	0	No	No	10	No	No phone s	DSL	Yes	No	No	No	No	No	Month-to-m	No	Mailed check	29.75	301.9	No
7892-POOK	Female	0	Yes	No	28	Yes	Yes	Fiber opt	No	No	Yes	Yes	Yes	Yes	Month-to-m	Yes	Electronic check	104.8	3046.05	Yes
6388-TABG	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfer (aut	56.15	3487.95	No
9763-GRSK	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-m	Yes	Mailed check	49.95	587.45	No
7469-LKBC	Male	0	No	No	16	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Two year	No	Credit card (auton	18.95	326.8	No
8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber opt	No	No	Yes	No	Yes	Yes	One year	No	Credit card (auton	100.35	5681.1	No
0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber opt	No	Yes	Yes	No	Yes	Yes	Month-to-m	Yes	Bank transfer (aut	103.7	5036.3	Yes
5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber opt	Yes	No	Yes	Yes	Yes	Yes	Month-to-m	Yes	Electronic check	105.5	2686.05	No
3655-SNQY	Female	0	Yes	Yes	69	Yes	Yes	Fiber opt	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit card (auton	113.25	7895.15	No
8191-XWSZ	Female	0	No	No	52	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Mailed check	20.65	1022.95	No
9959-WOFK	Male	0	No	Yes	71	Yes	Yes	Fiber opt	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank transfer (aut	106.7	7382.25	No
4190-MFLU	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-m	No	Credit card (auton	55.2	528.35	Yes
4183-MYFR	Female	0	No	No	21	Yes	No	Fiber opt	No	Yes	Yes	No	No	Yes	Month-to-m	Yes	Electronic check	90.05	1862.9	No
8779-QRDM	Male	1	No	No	1	No	No phone s	DSL	No	No	Yes	No	No	Yes	Month-to-m	Yes	Electronic check	39.65	39.65	Yes
1680-VDCW	Male	0	Yes	No	12	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	One year	No	Bank transfer (aut	19.8	202.25	No
1066-JKSGK	Male	0	No	No	1	Yes	No	No	No interne	No interne	No interne	No interne	No interne	No interne	Month-to-m	No	Mailed check	20.15	20.15	Yes
3638-WEAB	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit card (auton	59.9	3505.1	No
6322-HRPF	Male	0	Yes	Yes	49	Yes	No	DSL	Yes	Yes	No	Yes	No	No	Month-to-m	No	Credit card (auton	59.6	2970.3	No
6865-JZNK	Female	0	No	No	30	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-m	Yes	Bank transfer (aut	55.3	1530.6	No
6467-CHFZ	Male	0	Yes	Yes	47	Yes	Yes	Fiber opt	No	Yes	No	No	Yes	Yes	Month-to-m	Yes	Electronic check	99.35	4749.15	Yes

Data Source - https://www.kaggle.com/blastchar/telco-customer-churn

Definitions of Terms

Clustering

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples. Clustering is used in market segmentation; where we try to find customers that are similar to each other whether in terms of behaviours or attributes, image segmentation/compression; where we try to group similar regions together, document clustering based on topics, etc.

Unlike supervised learning, clustering is considered an unsupervised learning method since we don't have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. We only want to try to investigate the structure of the data by grouping the data points into distinct subgroups.

K-Means Algorithm

K Means Algorithm is an Iterative algorithm that divides a group of n datasets into k subgroups /clusters based on the similarity and their mean distance from the centroid of that particular subgroup/ formed.

K, here is the pre-defined number of clusters to be formed by the Algorithm. If K=3, It means the number of clusters to be formed from the dataset is 3 .

Algorithm steps Of K Means:-

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the value of K, to decide the number of clusters to be formed.

Step-2: Select random K points which will act as centroids.

Step-3: Assign each data point, based on their distance from the randomly selected points (Centroid), to the nearest/closest centroid which will form the predefined clusters.

Step-4: place a new centroid of each cluster.

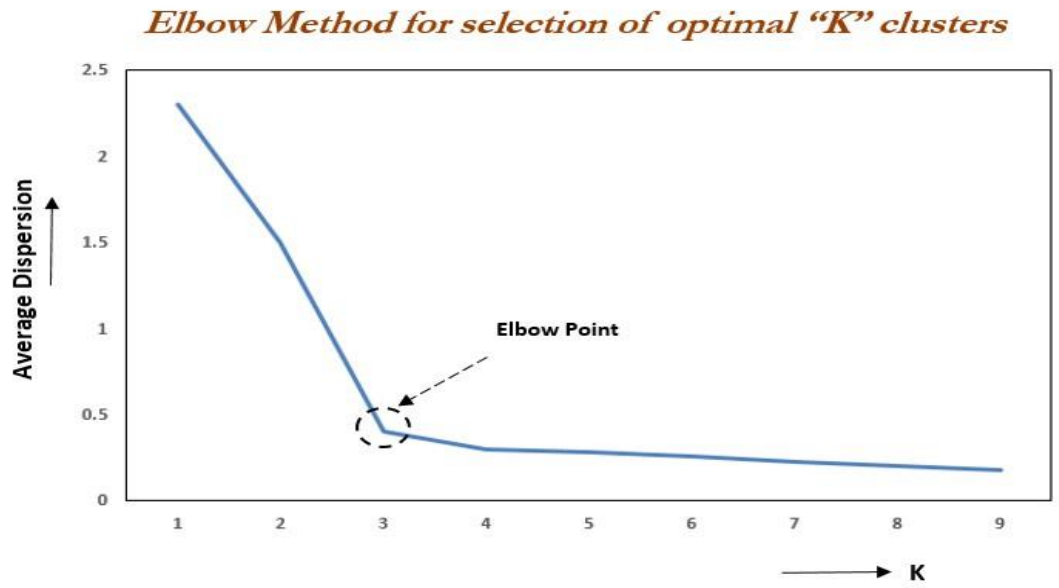
Step-5: Repeat step no.3, which reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to Step 7.

Step-7: FINISH

Elbow method

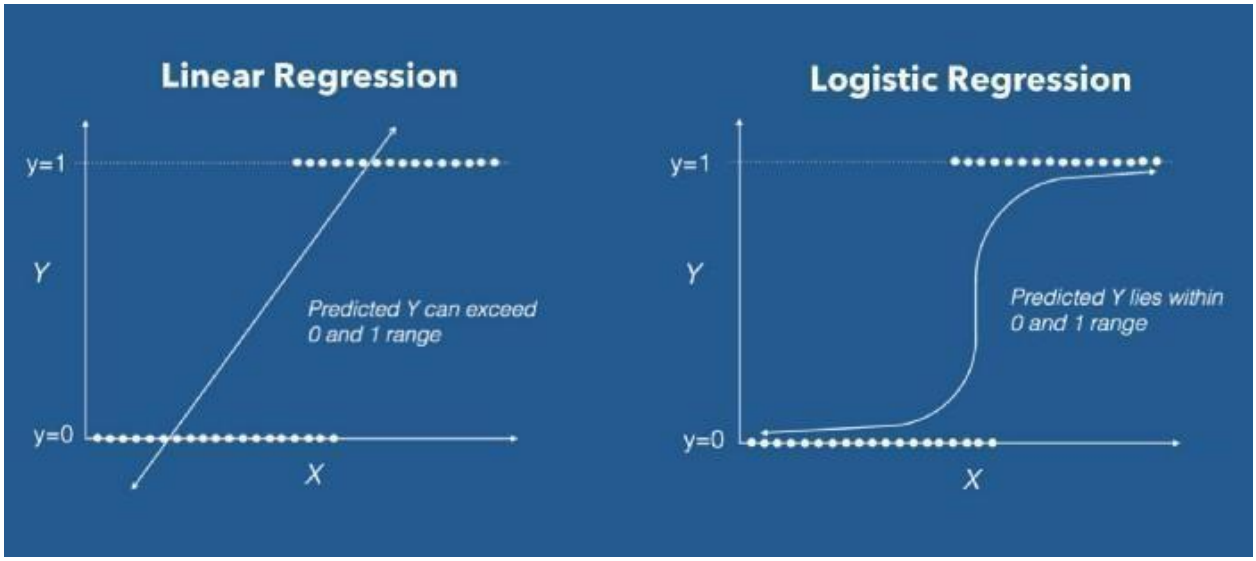
The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of *k*. As you know, if *k* increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as *k* increases. The value of *k* at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.



Logistic Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there

would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.



Random Forest Classifier

The Random forest or Random Decision Forest is a supervised Machine learning algorithm used for classification, regression, and other tasks using decision trees.The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

Assumptions and Delimitations

1}First of all, we are ignoring the competition the business market is facing due to other businesses. We are just taking the dataset, analyzing it and giving the insights based upon that.

2) We have also ignored the satisfaction level of the customers regarding the services and benefits, we are just analyzing their churn rate based on their behavior.

3) We have not taken the demographic data(except gender) and geographic data of the customers.

4) We have also ignored 3 outliers in the Segmentation part of our project

5) We are including only those customers who left and remained with us within the last month ,i.e., this dataset contains the data corresponding to only one particular month.

6) While analyzing the data for Segmentation, we took only a fraction of the dataset i.e., 15% such that the number of rows reduced from 7043 to 1056. This was done in order to decrease the

time complexity of the code and as a result to produce better and clearer insights.

7) We have also replaced some columns having similar meaning with similar column values.

Chapter 2

PRESENTATION OF FINDINGS

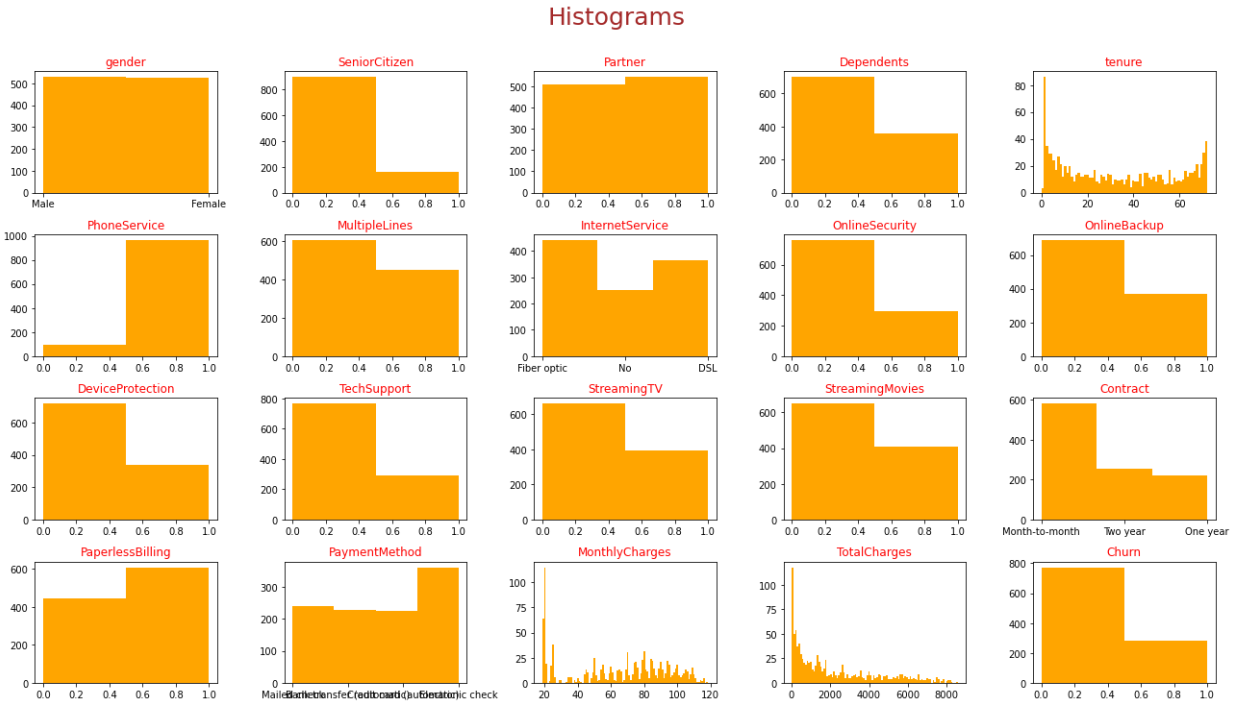
Customer Segmentation

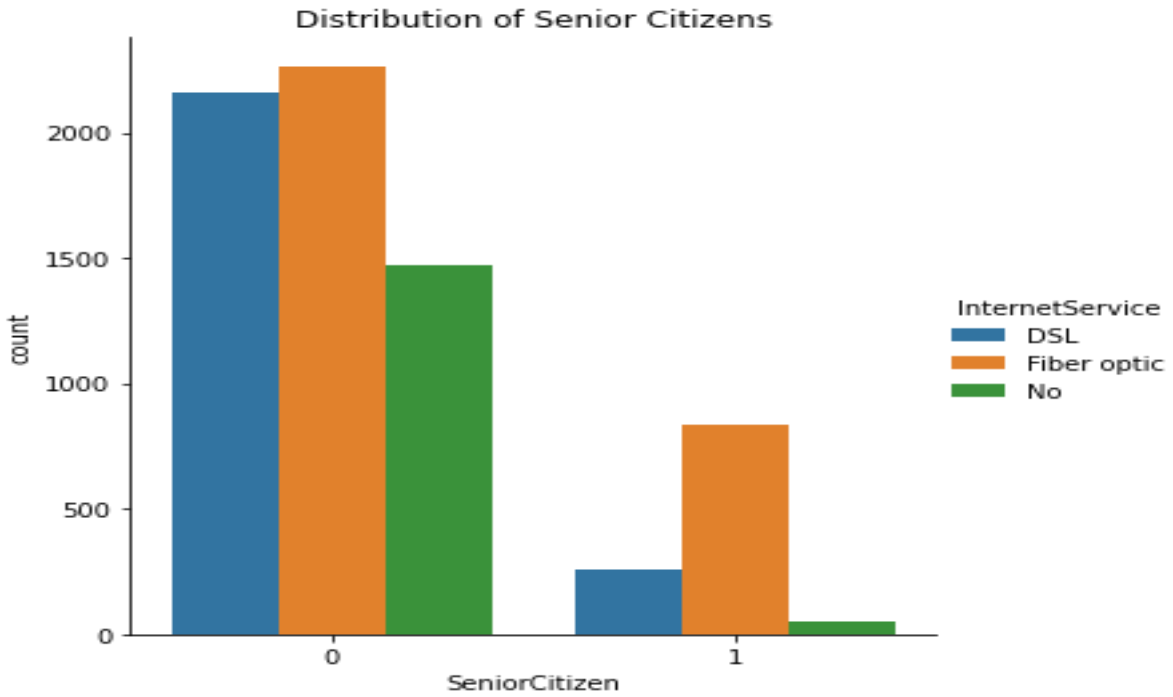
After cleaning the Dataset, we started off with analyzing the data. We started by plotting Histograms of all the features of our dataset to see the variations among their different values.

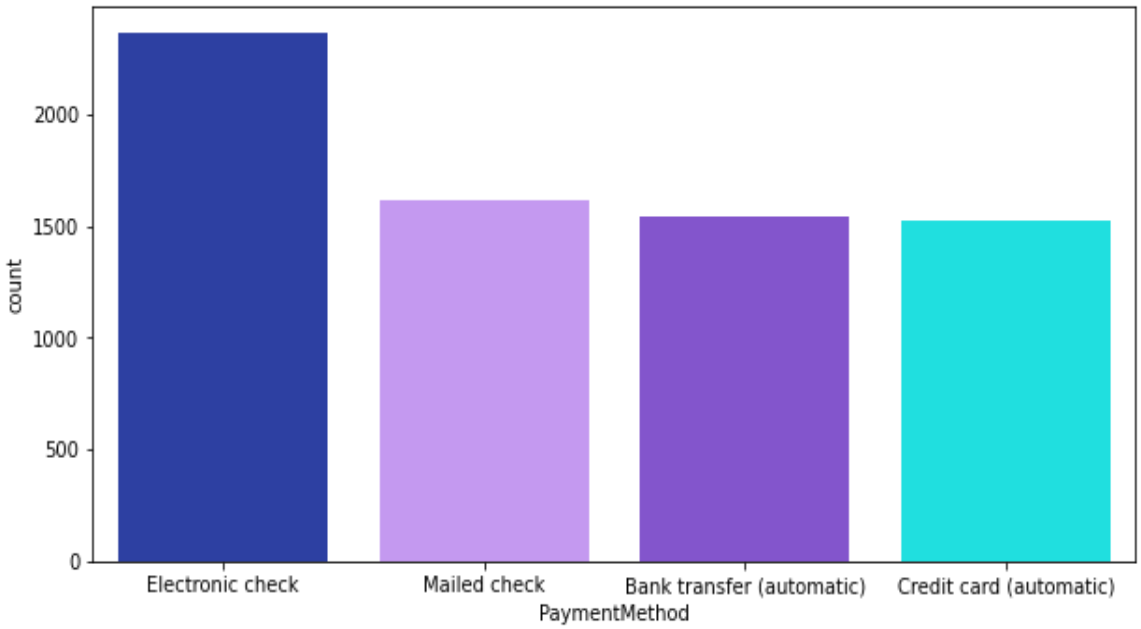
After analyzing, we found out that Gender features had little or no variation between males and females. Similar was the case with "Partners" feature of the dataset. Most variation was found among the features - "tenure", "Contract", "Payment Method", "Monthly Charges" and "Total Charges". Variations in other features didn't matter much because most of them were features containing Yes or No values which were no use to us for the segmentation part of our Project. Tenure feature describes the Loyalty of the Customer, from how much time they have been subscribed to us.The Contract feature describes the length of their current subscription ,i.e., Month-to-month, One-year, two-year. Most people belonged to month-to-month subscription and there was little or no variation in the other two values of the feature. The next feature i.e., Payment Method shows little or no variation among all values except the Electronic Check Method. Lastly The features Monthly charges and Total Charges deal with the payment from the customers and they definitely have huge variations among them. Moving forward we then plotted a Category Plot of Senior Citizens with respect to the types of Internet Services they are having. Here we find out that the maximum number of senior citizens are not subscribed to our services. Apart from that, we found out that the ones with the subscription have more probability of having Fibre Internet

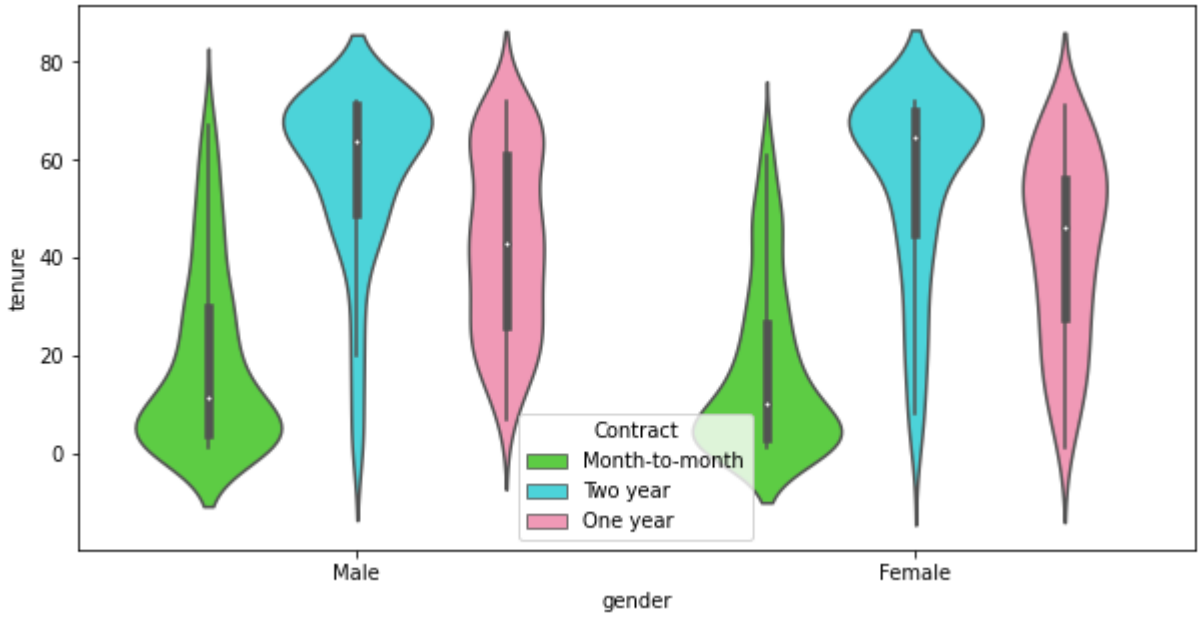
services than DSL and there are very few without Internet Services. As we are done with the category Plot, we move towards plotting a count plot for Payment Method. Here we find out that the maximum number of Payments are done through electronic checks followed by Mailed Check, Bank Transfer and Credit Card with little variation among each other. With this, we move onto building a Violin Plot of tenure vs Gender with respect to the Contract feature. We come to note that males have a higher Tenure than females with respect to month-to-month contracts. Moreover, we also analysed that there are more females having more tenure with respect to Two-year contracts. Also, in terms of one-year contracts, males are the ones with more tenure. While in two-year contract number of males has surged down. After taking some insights from the Violin Plot, we move further to one more Histogram of Monthly Charges with respect to Gender. As we have seen earlier, Variation between males and females is near to zero, there is no such insights from this graph except for the fact that there are considerably more number of males with monthly charges less than 30. Moving on, We reach the Relation Plot between Tenure and Monthly Charges with respect to Total charges. In Spite of the fact that the more the Monthly Charges the more the Total Charges are, we found out that Total charges were still low for maximum Monthly Charges until the tenure period of the customer increased. Same thing is with tenure too. It doesn't matter how much the tenure is, Total charges won't be maximum until the monthly charges are high too and vice versa. So Total Charges depend both on Monthly Charges and Tenure of the Customer such that with any one being less will make the Total Charges less too. After that, we plotted Heatmap where we found out the correlation between different features of the dataset. Then we moved on to building count plots of Tenure, Contract, Monthly Charges and Total Charges.

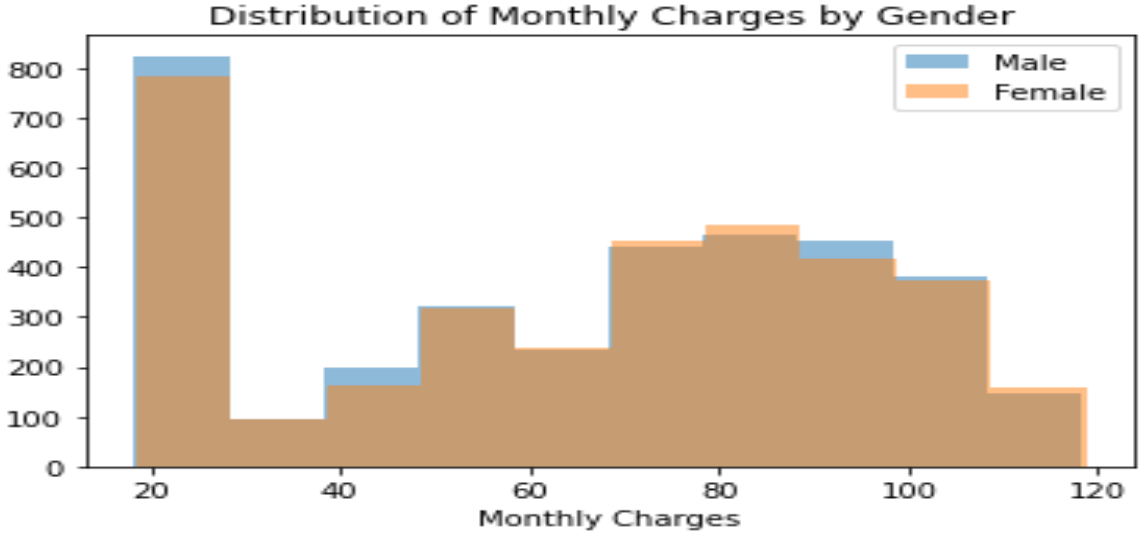
Then we moved ahead towards clustering by first locating the KMeans of our dataset. We took the numerical data for ease of plotting namely Tenure, Monthly Charges and Total Charges. From here, the Elbow point came out to be at 3 thereby giving us our k value or the number of clusters ,i.e., 3. So, lastly we moved on to build the scatter plot to plot the 3 clusters that we made, first in 2-D graph and then in 3-D to gain a better view. There are 3 clusters stacked into each other making a beautiful little pattern of 3 colored-dotted points.

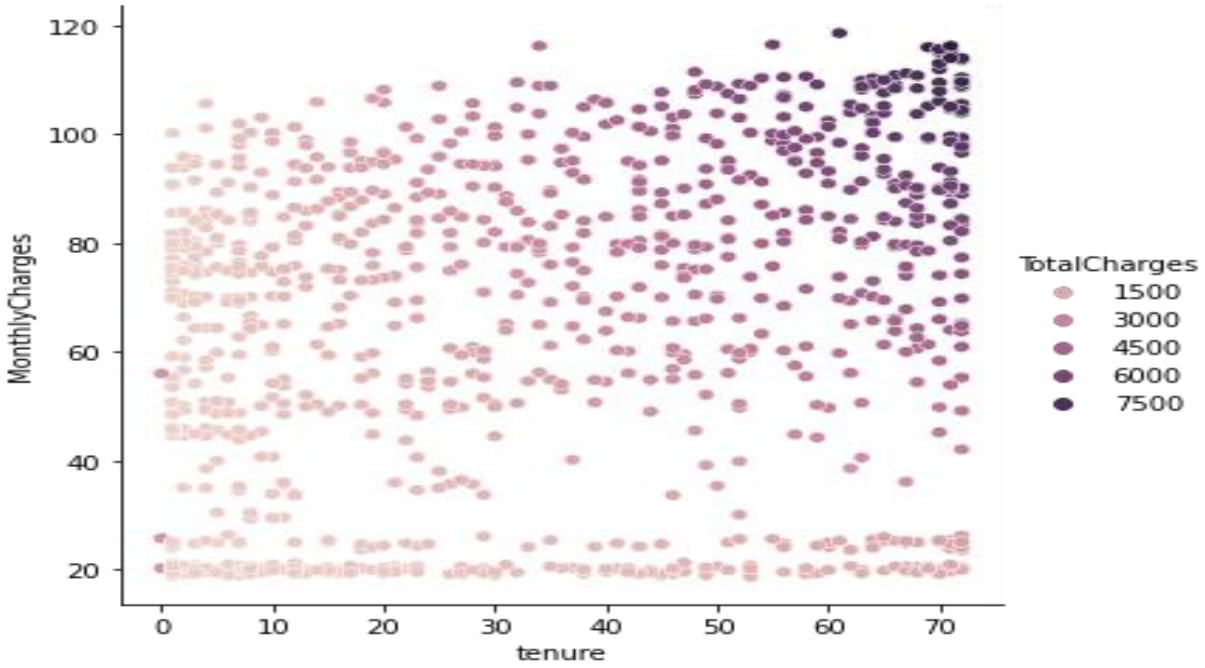

Histograms
gender
SeniorCitizen
Partner
Dependents
tenure
PhoneService
MultipleLines
InternetService
OnlineSecurity
OnlineBackup
DeviceProtection
TechSupport
StreamingTV
StreamingMovies
Contract
PaperlessBilling
PaymentMethod
MonthlyCharges
TotalCharges
Churn
Male
Female
Fiber optic
No
DSL
Month-to-month
Two year
One year

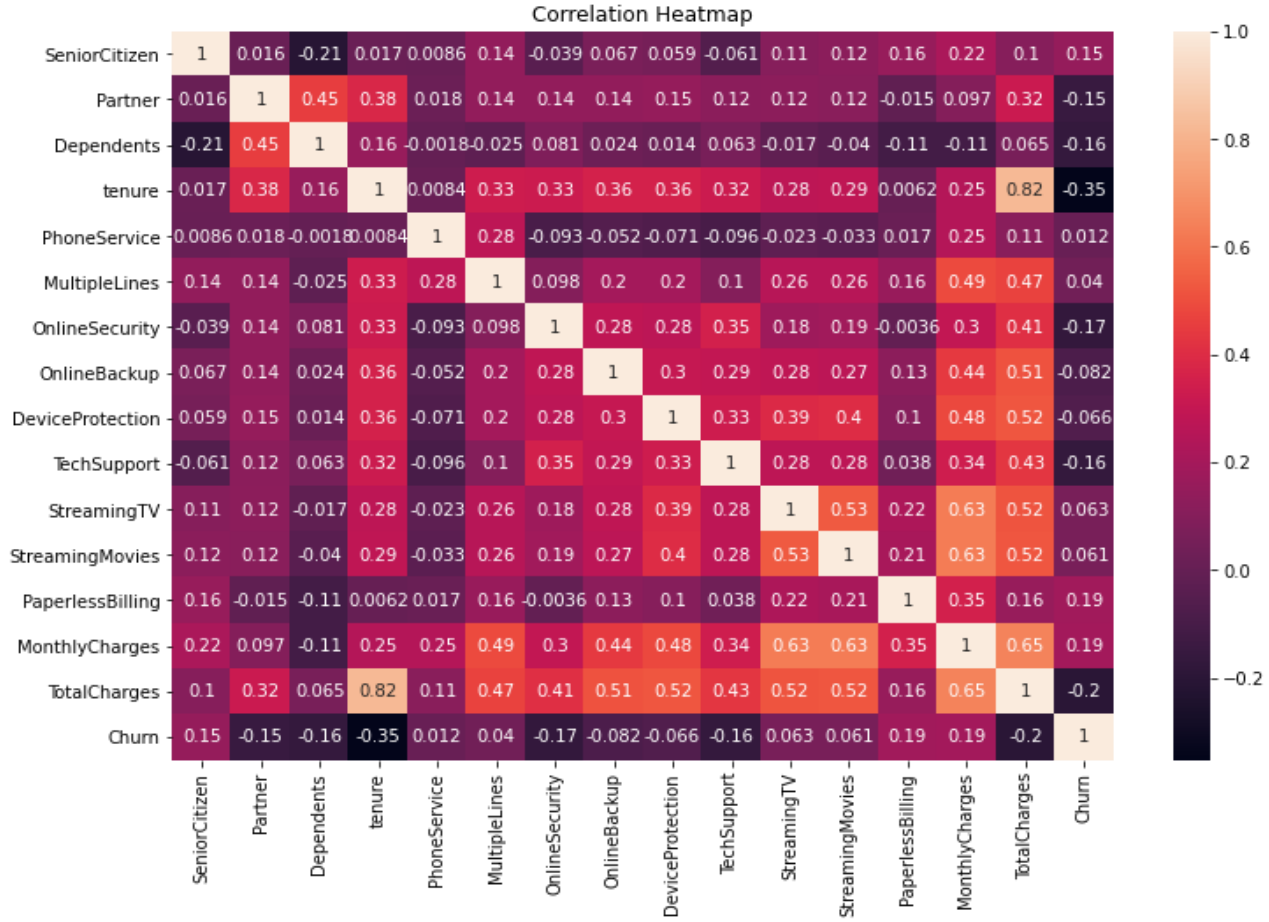

Distribution of Senior Citizens
count
SeniorCitizen
InternetService
DSL
Fiber optic
No

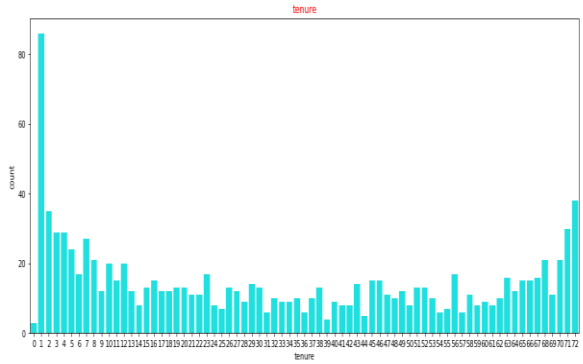

2000
1500
1000
500
0
count
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
PaymentMethod

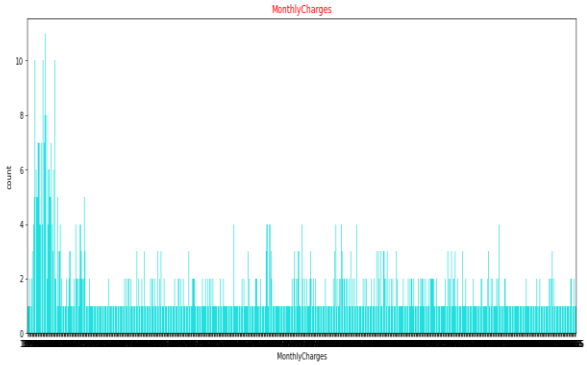

80
60
40
20
0
tenure
Contract
Month-to-month
Two year
One year
Male
Female
gender

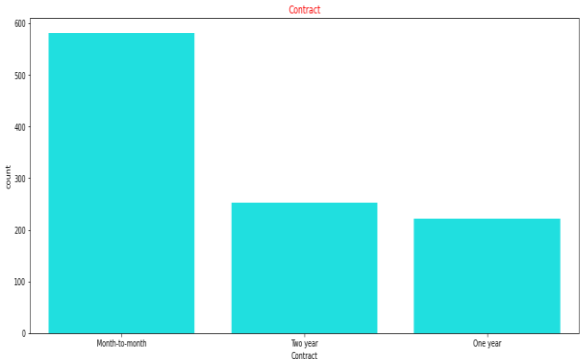

Distribution of Monthly Charges by Gender
Male
Female
800
700
600
500
400
300
200
100
0
20
40
60
80
100
120
Monthly Charges

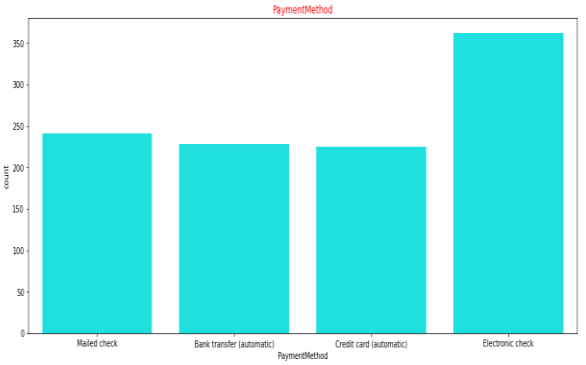

MonthlyCharges
tenure
TotalCharges
1500
3000
4500
6000
7500

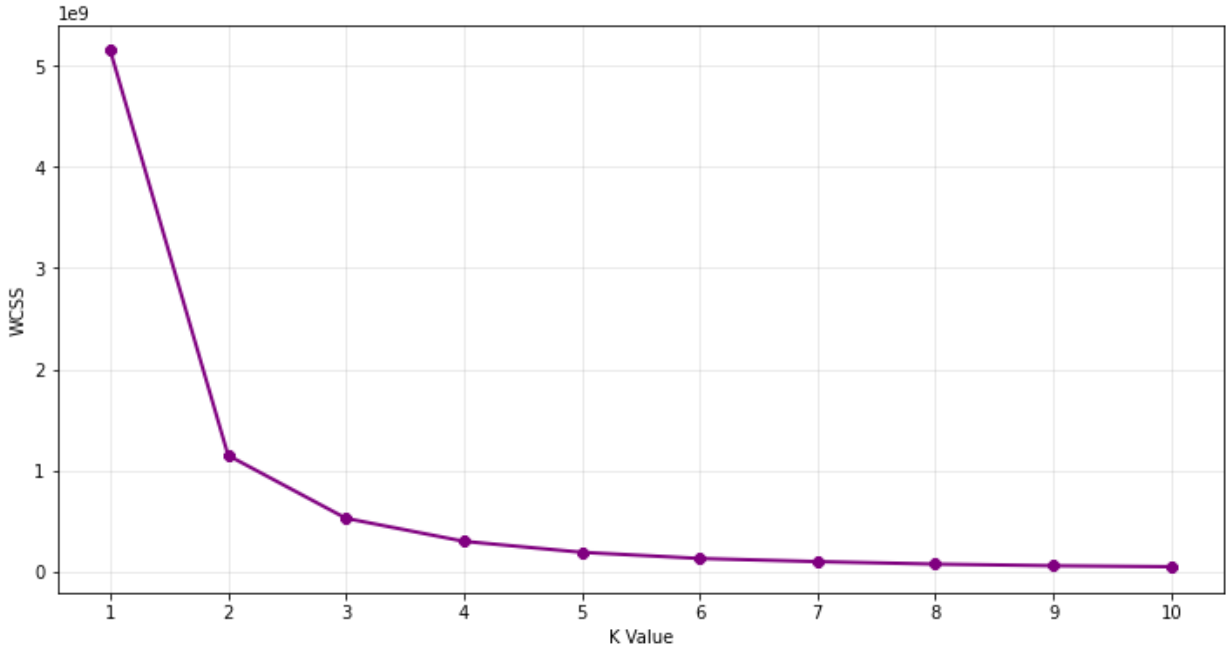

Correlation Heatmap
SeniorCitizen 1 0.016 -0.21 0.017 0.0086 0.14 -0.039 0.067 0.059 -0.061 0.11 0.12 0.16 0.22 0.1 0.15
Partner 0.016 1 0.45 0.38 0.018 0.14 0.14 0.14 0.15 0.12 0.12 0.12 -0.015 0.097 0.32 -0.15
Dependents -0.21 0.45 1 0.16 -0.0018 -0.025 0.081 0.024 0.014 0.063 -0.017 -0.04 -0.11 -0.11 0.065 -0.16
tenure 0.017 0.38 0.16 1 0.0084 0.33 0.33 0.36 0.36 0.32 0.28 0.29 0.0062 0.25 0.82 -0.35
PhoneService 0.0086 0.018 -0.0018 0.0084 1 0.28 -0.093 -0.052 -0.071 -0.096 -0.023 -0.033 0.017 0.25 0.11 0.012
MultipleLines 0.14 0.14 -0.025 0.33 0.28 1 0.098 0.2 0.2 0.1 0.26 0.26 0.16 0.49 0.47 0.04
OnlineSecurity -0.039 0.14 0.081 0.33 -0.093 0.098 1 0.28 0.28 0.35 0.18 0.19 -0.0036 0.3 0.41 -0.17
OnlineBackup 0.067 0.14 0.024 0.36 -0.052 0.2 0.28 1 0.3 0.29 0.28 0.27 0.13 0.44 0.51 -0.082
DeviceProtection 0.059 0.15 0.014 0.36 -0.071 0.2 0.28 0.3 1 0.33 0.39 0.4 0.1 0.48 0.52 -0.066
TechSupport -0.061 0.12 0.063 0.32 -0.096 0.1 0.35 0.29 0.33 1 0.28 0.28 0.038 0.34 0.43 -0.16
StreamingTV 0.11 0.12 -0.017 0.28 -0.023 0.26 0.18 0.28 0.39 0.28 1 0.53 0.22 0.63 0.52 0.063
StreamingMovies 0.12 0.12 -0.04 0.29 -0.033 0.26 0.19 0.27 0.4 0.28 0.53 1 0.21 0.63 0.52 0.061
PaperlessBilling 0.16 -0.015 -0.11 0.0062 0.017 0.16 -0.0036 0.13 0.1 0.038 0.22 0.21 1 0.35 0.16 0.19
MonthlyCharges 0.22 0.097 -0.11 0.25 0.25 0.49 0.3 0.44 0.48 0.34 0.63 0.63 0.35 1 0.65 0.19
TotalCharges 0.1 0.32 0.065 0.82 0.11 0.47 0.41 0.51 0.52 0.43 0.52 0.52 0.16 0.65 1 -0.2
Churn 0.15 -0.15 -0.16 -0.35 0.012 0.04 -0.17 -0.082 -0.066 -0.16 0.063 0.061 0.19 0.19 -0.2 1
SeniorCitizen
Partner
Dependents
tenure
PhoneService
MultipleLines
OnlineSecurity
OnlineBackup
DeviceProtection
TechSupport
StreamingTV
StreamingMovies
PaperlessBilling
MonthlyCharges
TotalCharges
Churn
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


tenure
count
tenure

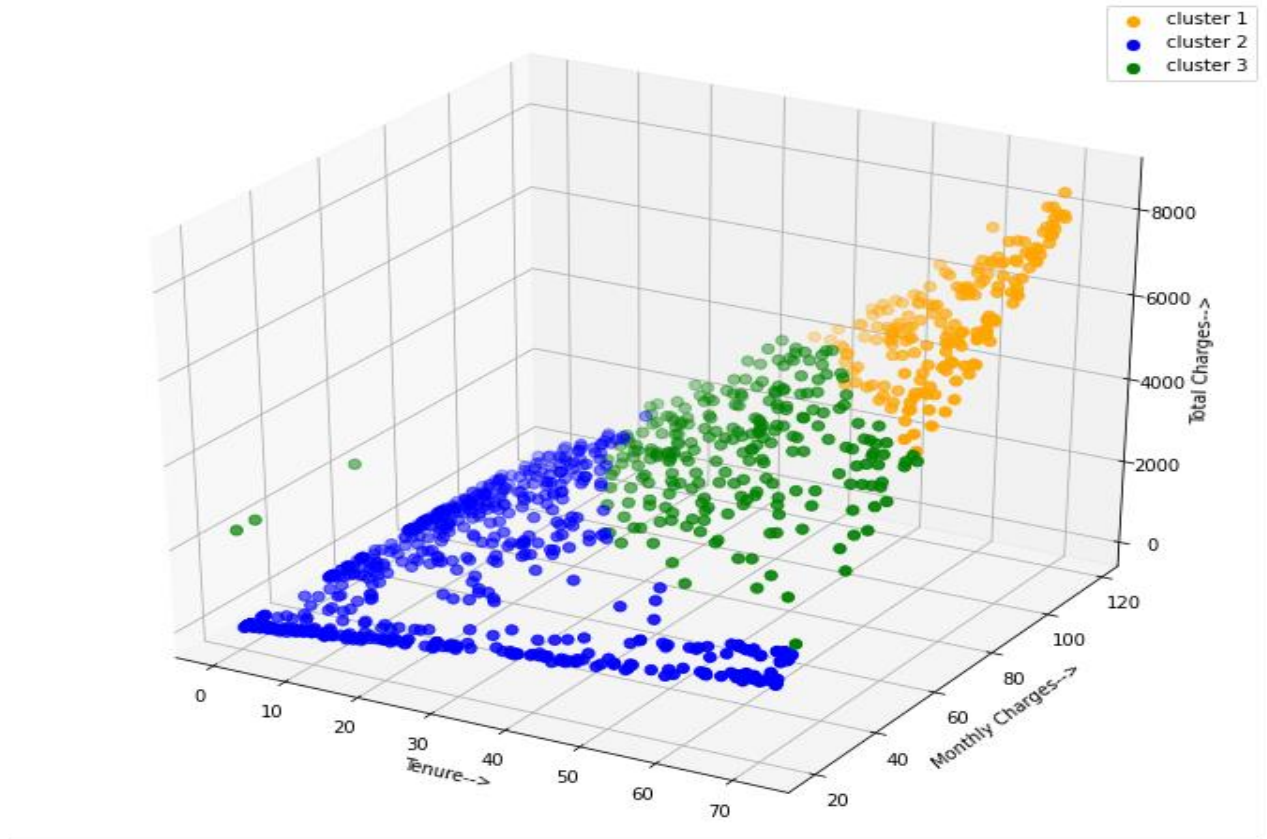

MonthlyCharges
count
MonthlyCharges


Contract
count
Month-to-month
Two year
One year
Contract

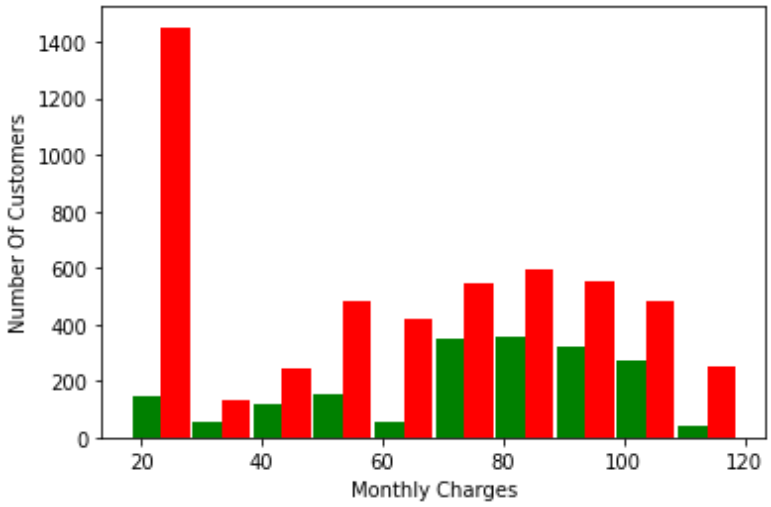

PaymentMethod
count
Mailed check
Bank transfer (automatic)
Credit card (automatic)
Electronic check
PaymentMethod

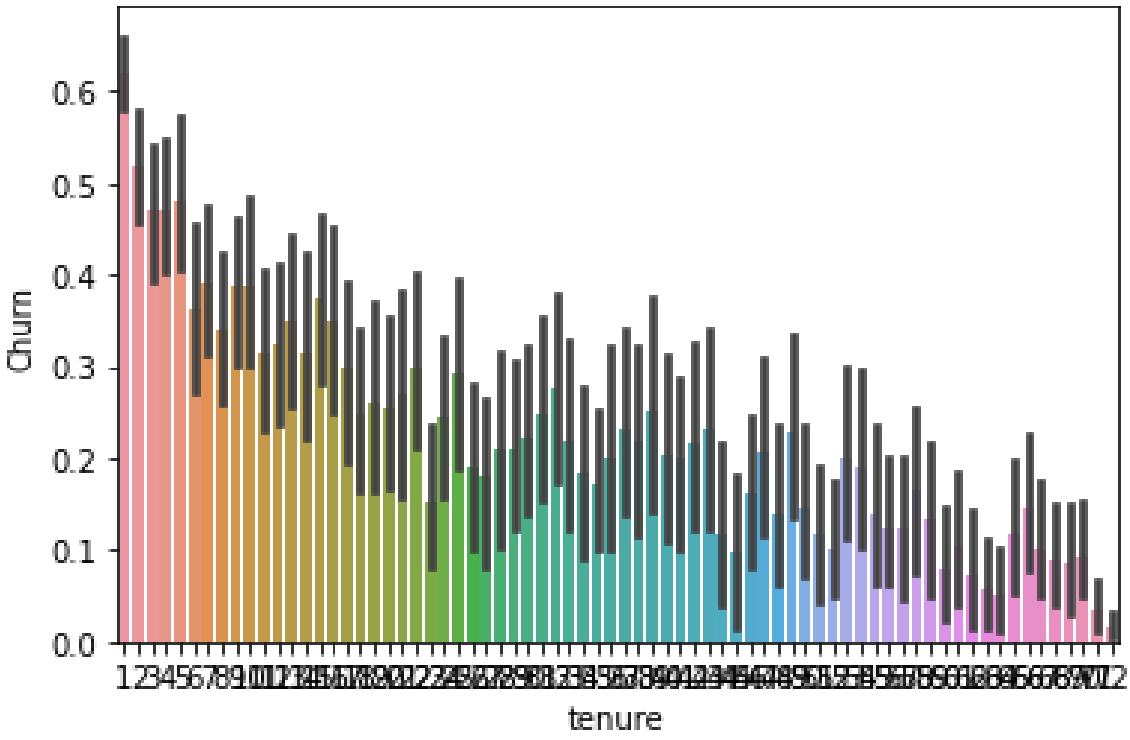

1e9
WCSS
K Value

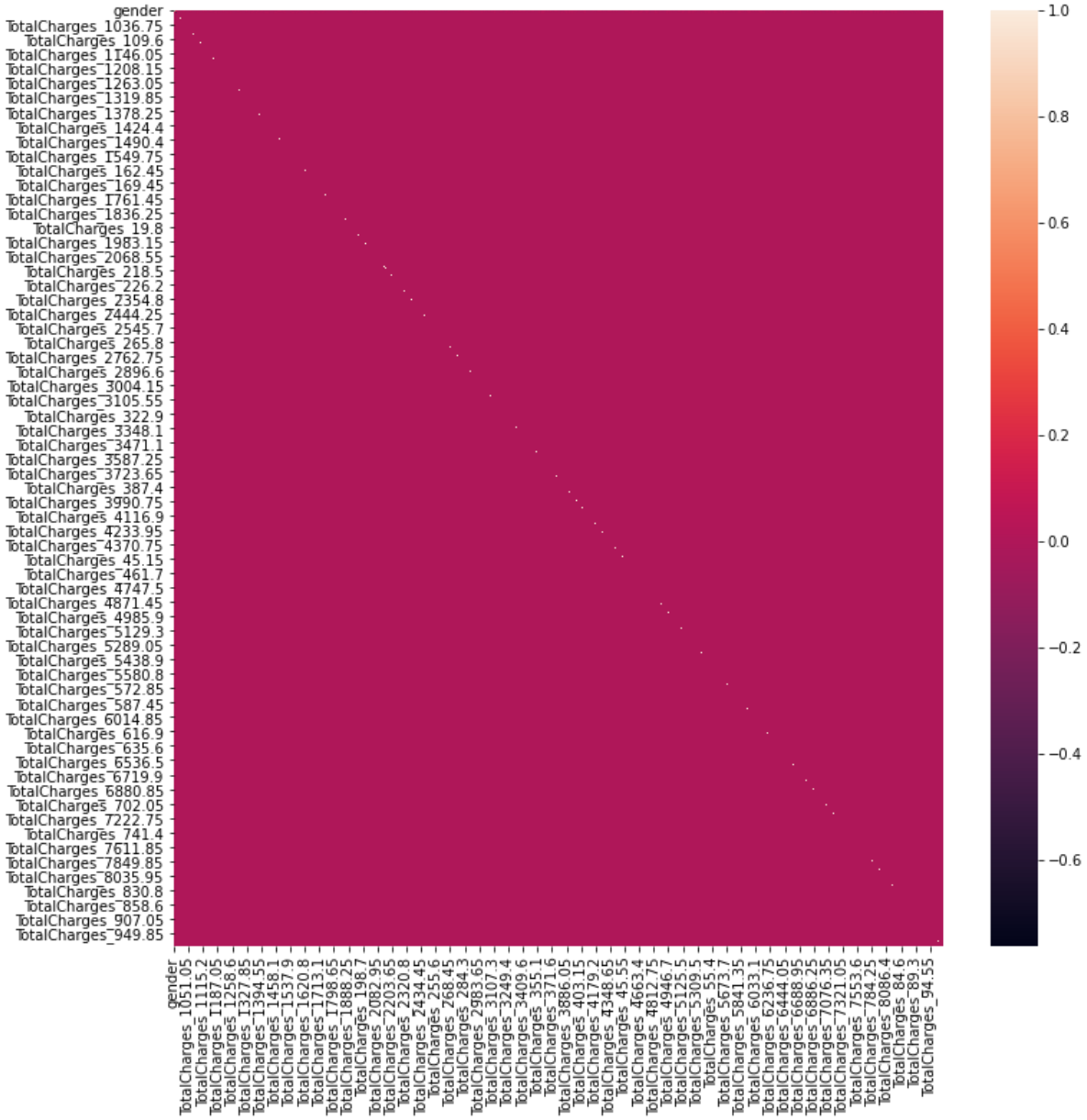

Clusters of Customers
Monthly Charges
Tenure
120
100
80
60
40
20
0
10
20
30
40
50
60
70

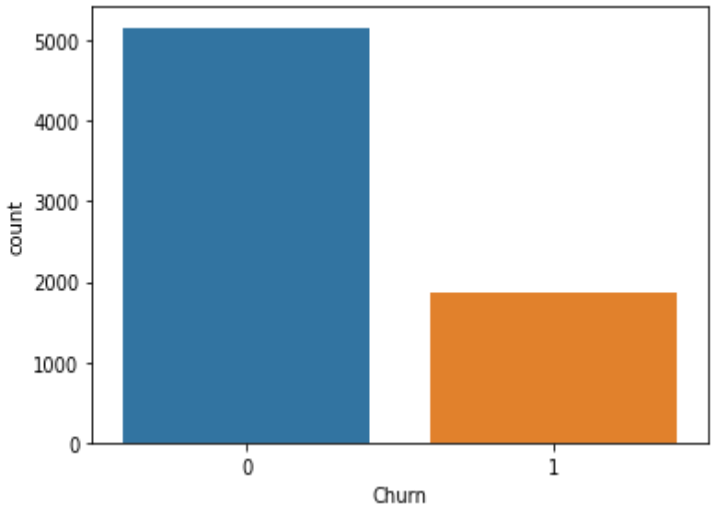

cluster 1
cluster 2
cluster 3
Tenure-->
Monthly Charges-->
Total Charges-->
8000
6000
4000
2000
0
120
100
80
60
40
20
0
10
20
30
40
50
60
70

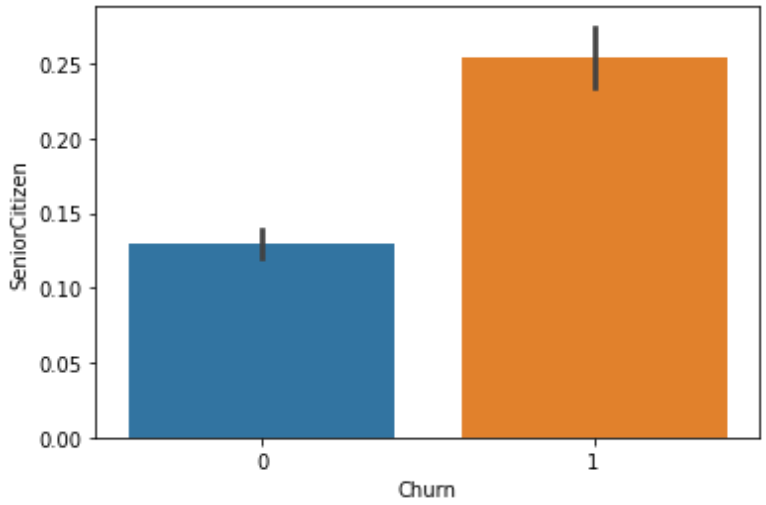
Customer Churn

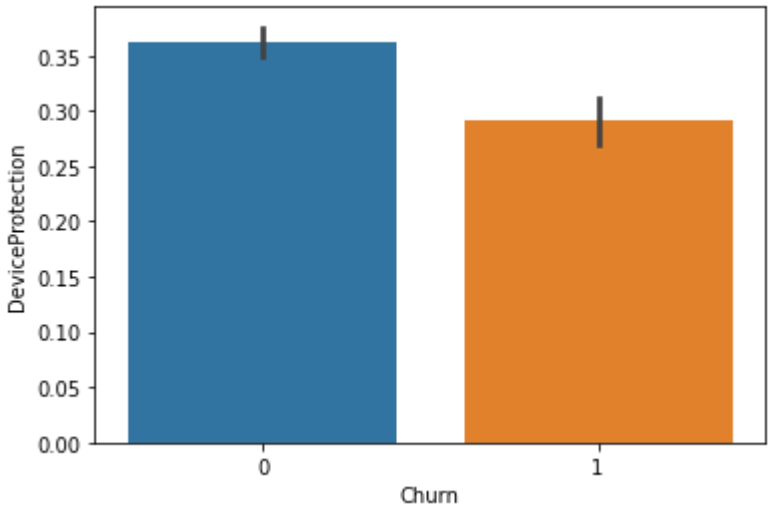


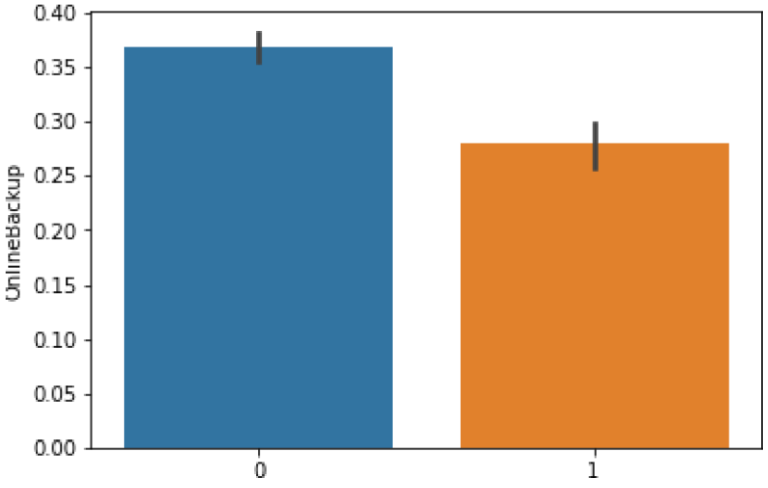


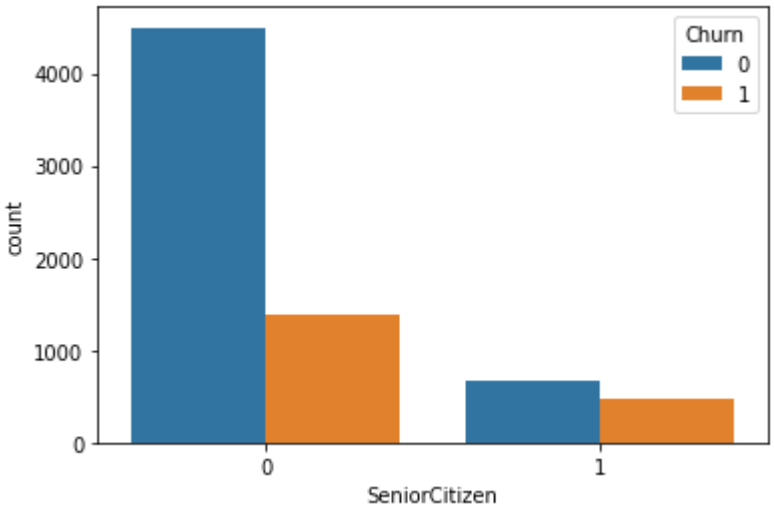

gender
TotalCharges_1036.75
TotalCharges_109.6
TotalCharges_1146.05
TotalCharges_1208.15
TotalCharges_1263.05
TotalCharges_1319.85
TotalCharges_1378.25
TotalCharges_1424.4
TotalCharges_1490.4
TotalCharges_1549.75
TotalCharges_162.45
TotalCharges_169.45
TotalCharges_1761.45
TotalCharges_1836.25
TotalCharges_19.8
TotalCharges_1983.15
TotalCharges_2068.55
TotalCharges_218.5
TotalCharges_226.2
TotalCharges_2354.8
TotalCharges_2444.25
TotalCharges_2545.7
TotalCharges_265.8
TotalCharges_2762.75
TotalCharges_2896.6
TotalCharges_3004.15
TotalCharges_3105.55
TotalCharges_322.9
TotalCharges_3348.1
TotalCharges_3471.1
TotalCharges_3587.25
TotalCharges_3723.65
TotalCharges_387.4
TotalCharges_3990.75
TotalCharges_4116.9
TotalCharges_4233.95
TotalCharges_4370.75
TotalCharges_45.15
TotalCharges_461.7
TotalCharges_4747.5
TotalCharges_4871.45
TotalCharges_4985.9
TotalCharges_5129.3
TotalCharges_5289.05
TotalCharges_5438.9
TotalCharges_5580.8
TotalCharges_572.85
TotalCharges_587.45
TotalCharges_6014.85
TotalCharges_616.9
TotalCharges_635.6
TotalCharges_6536.5
TotalCharges_6719.9
TotalCharges_6880.85
TotalCharges_702.05
TotalCharges_7222.75
TotalCharges_741.4
TotalCharges_7611.85
TotalCharges_7849.85
TotalCharges_8035.95
TotalCharges_830.8
TotalCharges_858.6
TotalCharges_907.05
TotalCharges_949.85
gender
TotalCharges_1051.05
TotalCharges_1115.2
TotalCharges_1187.05
TotalCharges_1258.6
TotalCharges_1327.85
TotalCharges_1394.55
TotalCharges_1458.1
TotalCharges_1537.9
TotalCharges_1620.8
TotalCharges_1713.1
TotalCharges_1798.65
TotalCharges_1888.25
TotalCharges_198.7
TotalCharges_2082.95
TotalCharges_2203.65
TotalCharges_2320.8
TotalCharges_2434.45
TotalCharges_255.6
TotalCharges_268.45
TotalCharges_284.3
TotalCharges_2983.65
TotalCharges_3107.3
TotalCharges_3249.4
TotalCharges_3409.6
TotalCharges_355.1
TotalCharges_371.6
TotalCharges_3886.05
TotalCharges_403.15
TotalCharges_4179.2
TotalCharges_4348.65
TotalCharges_45.55
TotalCharges_4663.4
TotalCharges_4812.75
TotalCharges_4946.7
TotalCharges_5125.5
TotalCharges_5309.5
TotalCharges_55.4
TotalCharges_5673.7
TotalCharges_5841.35
TotalCharges_6033.1
TotalCharges_6236.75
TotalCharges_6444.05
TotalCharges_6688.95
TotalCharges_6886.25
TotalCharges_7076.35
TotalCharges_7321.05
TotalCharges_7553.6
TotalCharges_784.25
TotalCharges_8086.4
TotalCharges_84.6
TotalCharges_89.3
TotalCharges_94.55
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

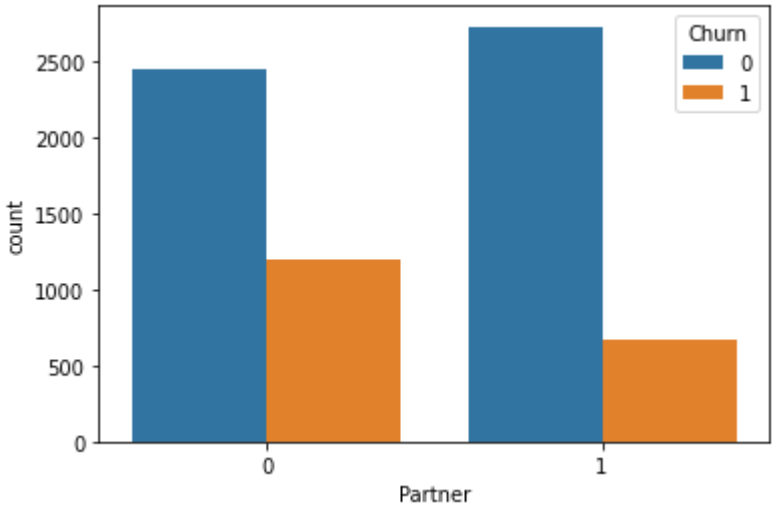

5000
4000
3000
2000
1000
0
count
0
1
Churn

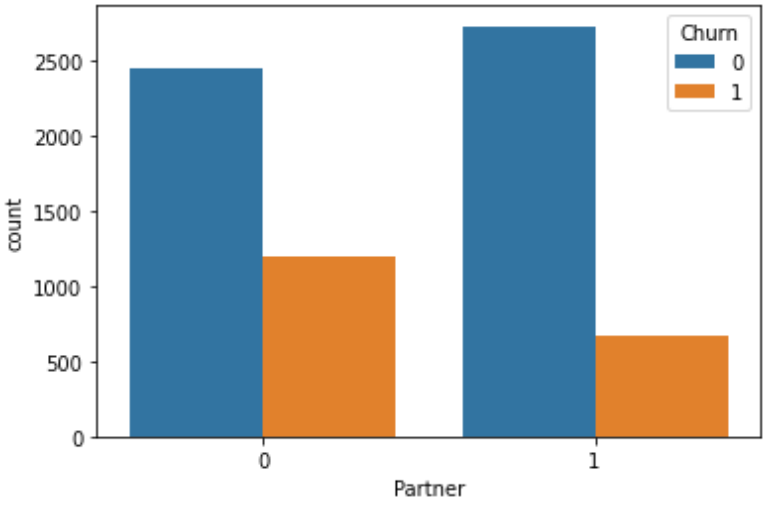

SeniorCitizen
0.25
0.20
0.15
0.10
0.05
0.00
0
1
Churn

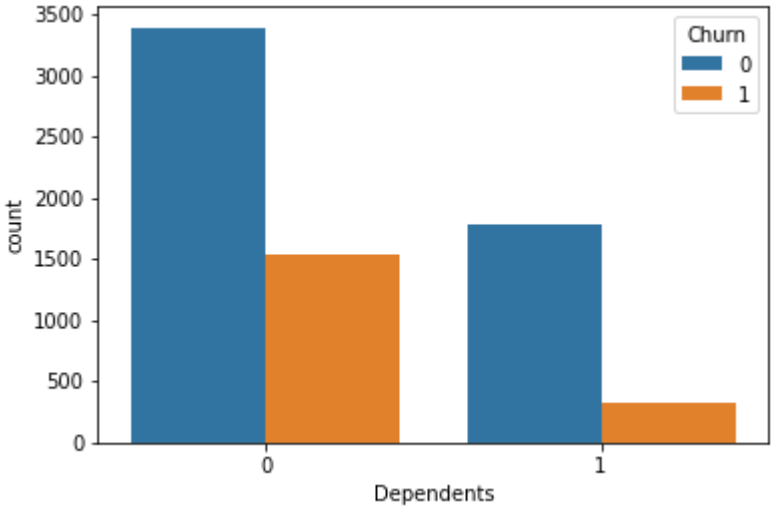

DeviceProtection
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
Churn

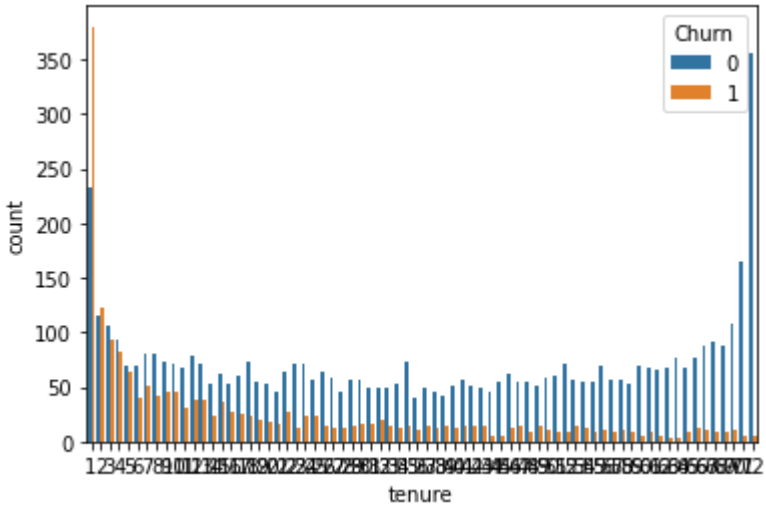

OnlineBackup
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1

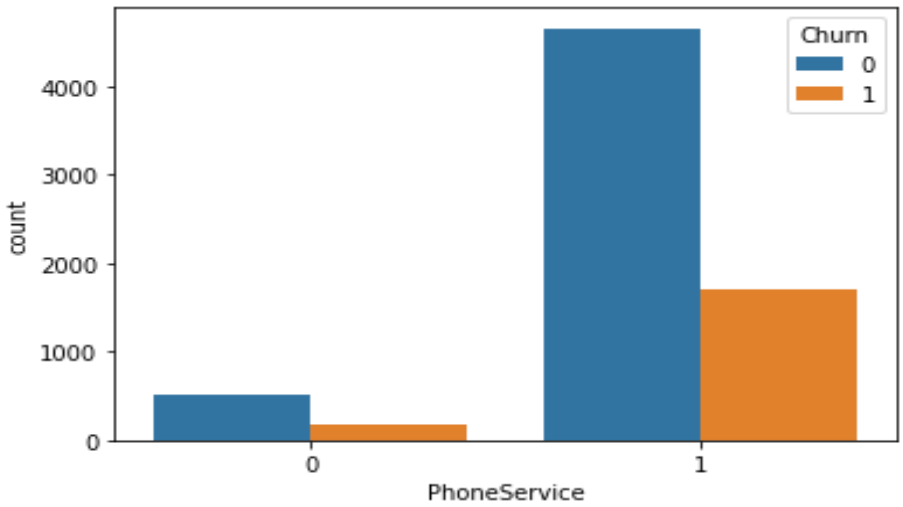

Churn
0
1
4000
3000
2000
1000
0
count
0
1
SeniorCitizen

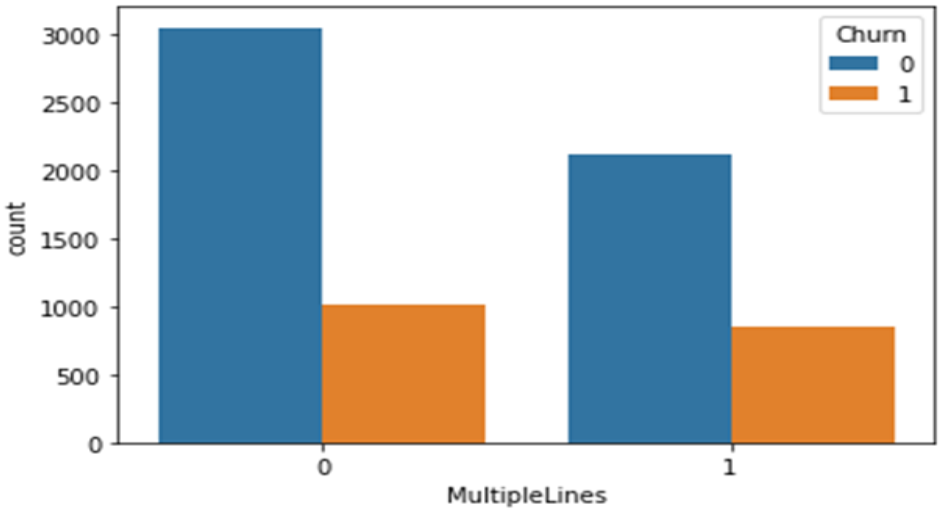

Churn
0
1
2500
2000
1500
1000
500
0
count
0
1
Partner

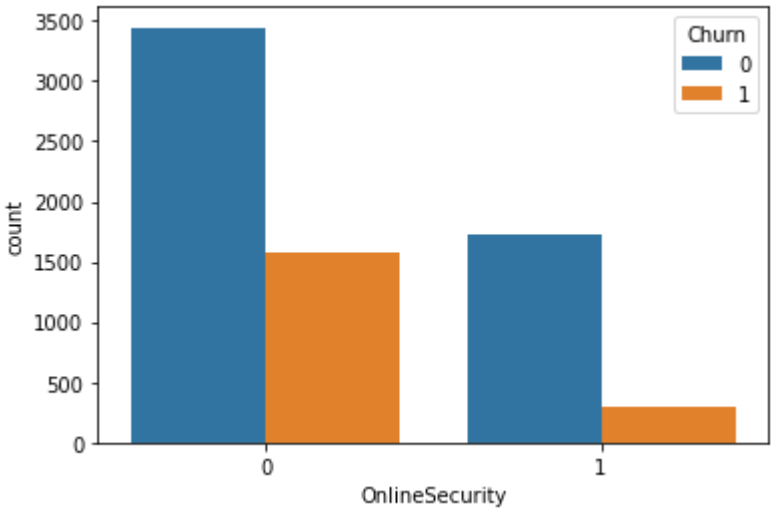

Churn
0
1
2500
2000
1500
1000
500
0
count
0
1
Partner

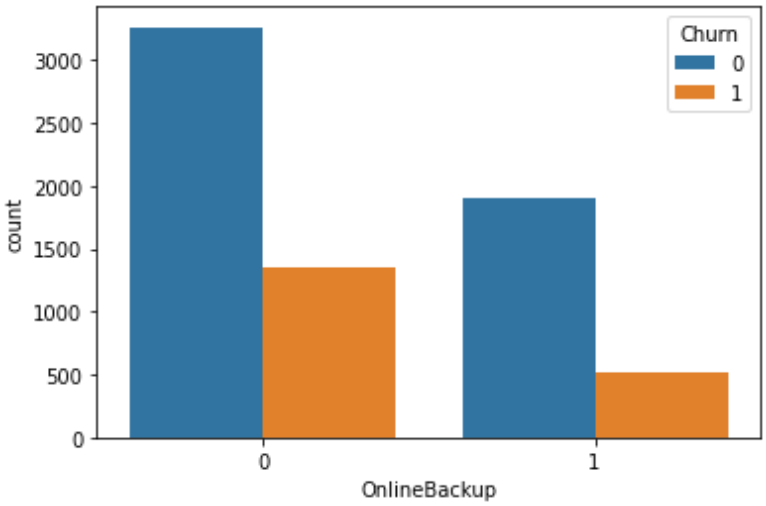

Churn
0
1
count
Dependents
0
1
500
1000
1500
2000
2500
3000
3500

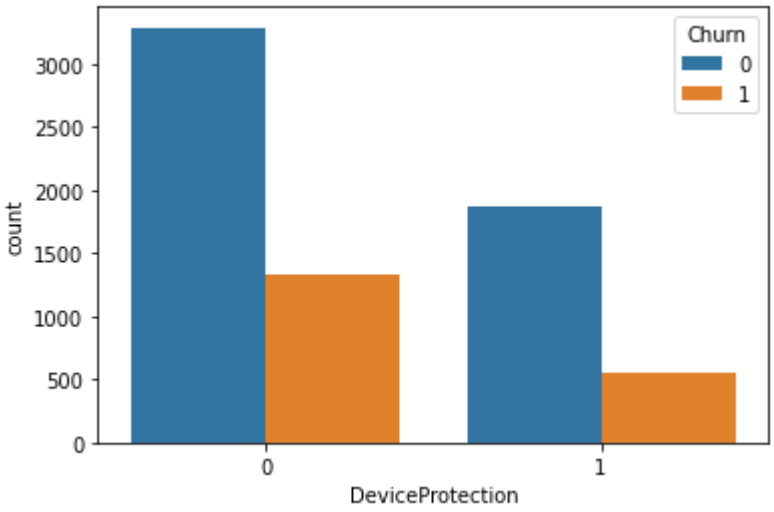

Churn
0
1
count
tenure
0
50
100
150
200
250
300
350

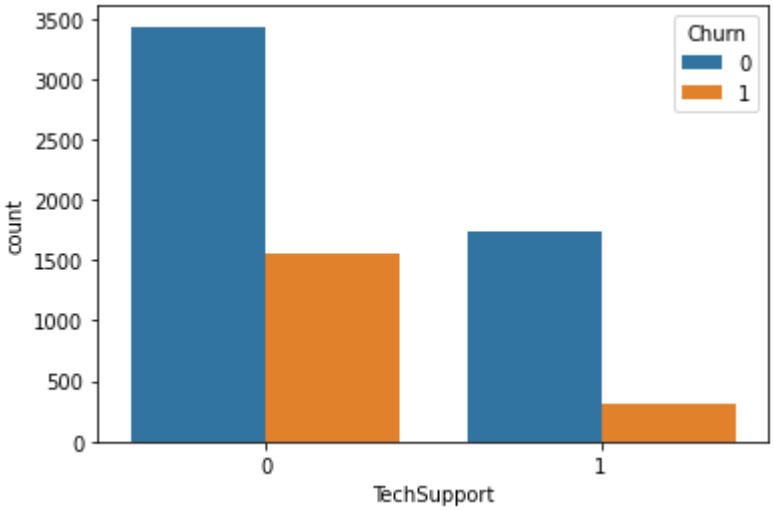

Churn
0
1
count
PhoneService
0
1
1000
2000
3000
4000

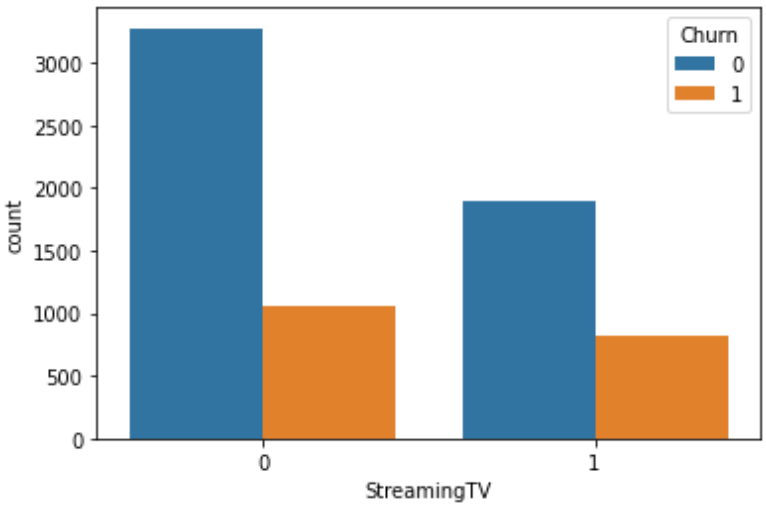

Churn
0
1
count
3000
2500
2000
1500
1000
500
0
0
1
MultipleLines

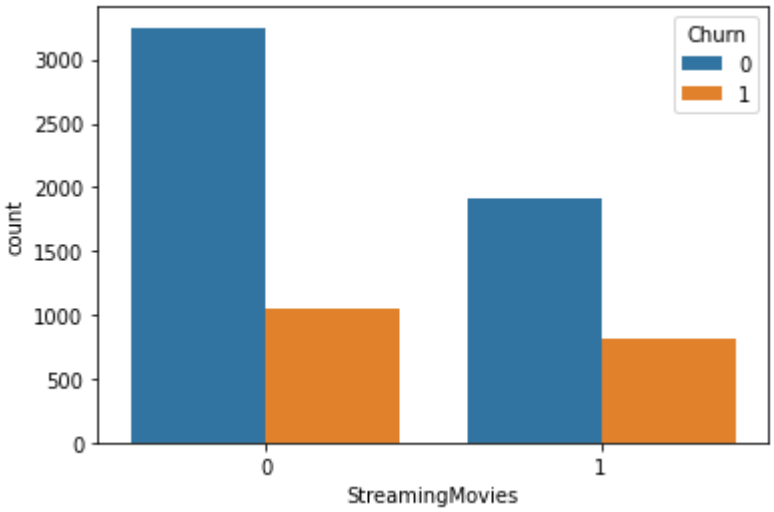

Churn
0
1
count
3500
3000
2500
2000
1500
1000
500
0
0
1
OnlineSecurity

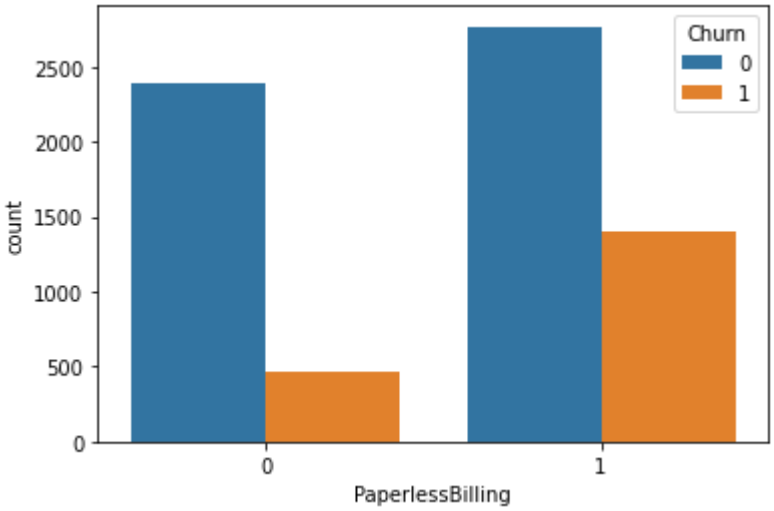

Churn
0
1
count
3000
2500
2000
1500
1000
500
0
0
1
OnlineBackup

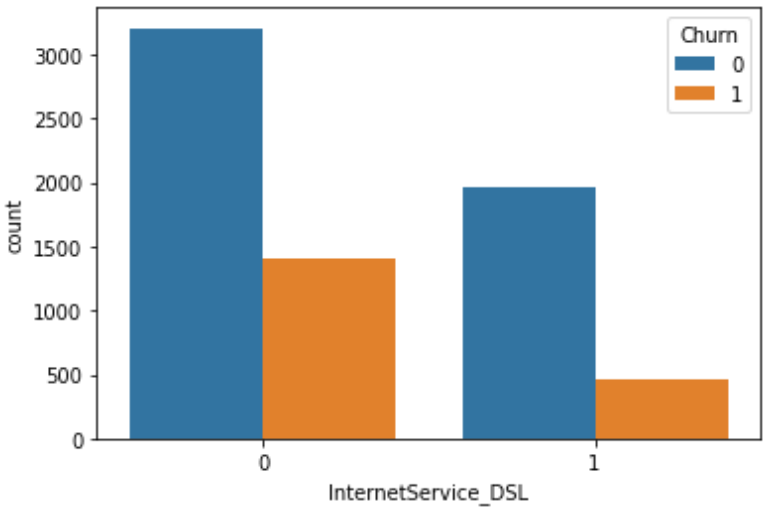

Churn
0
1
count
3000
2500
2000
1500
1000
500
0
0
1
DeviceProtection

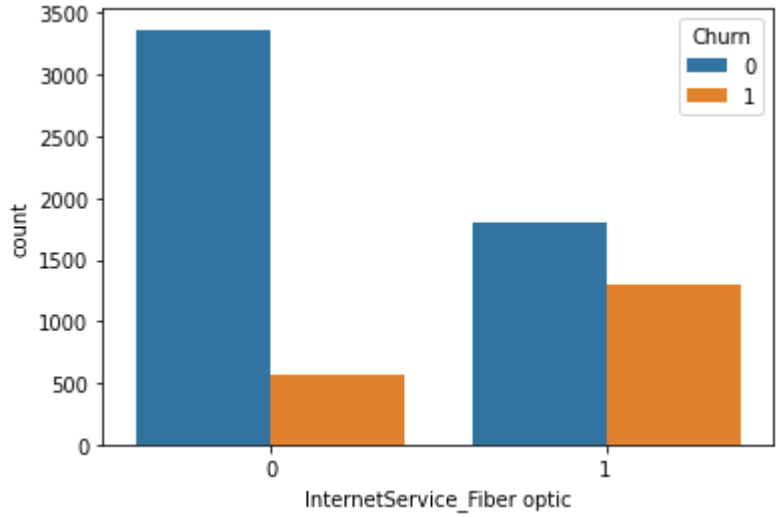

Churn
0
1
count
3500
3000
2500
2000
1500
1000
500
0
0
1
TechSupport

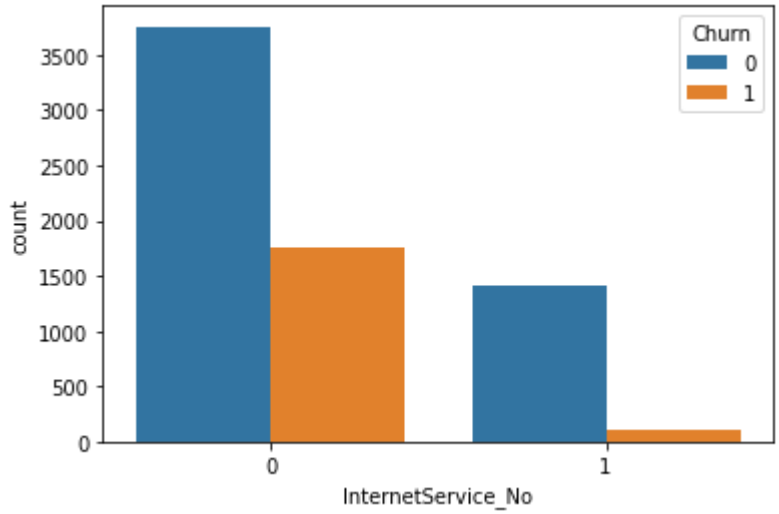

Churn
0
1
count
3000
2500
2000
1500
1000
500
0
0
1
StreamingTV

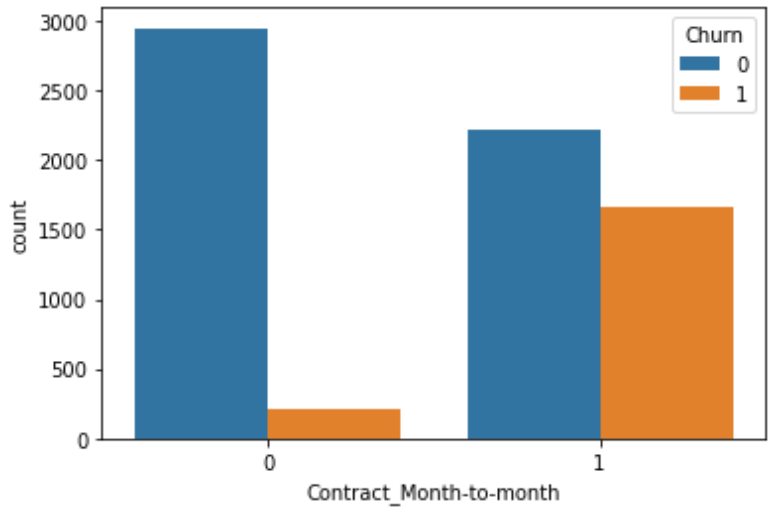

Churn
0
1
3000
2500
2000
1500
1000
500
0
count
0
1
StreamingMovies

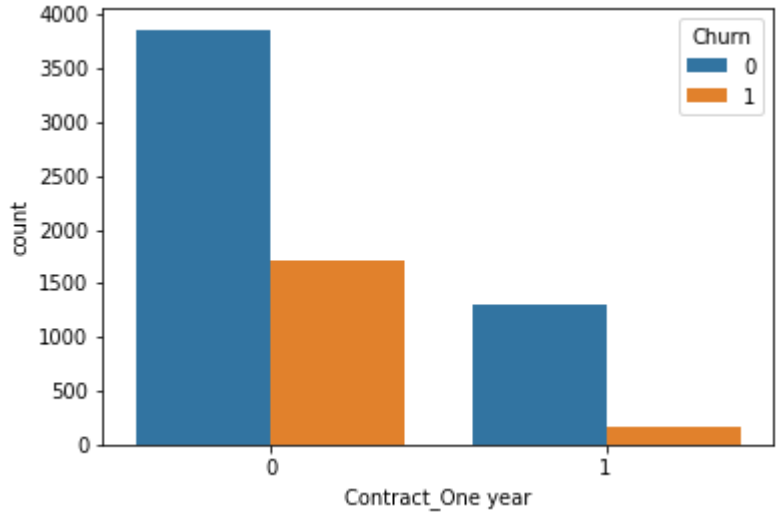

Churn
0
1
2500
2000
1500
1000
500
0
count
0
1
PaperlessBilling

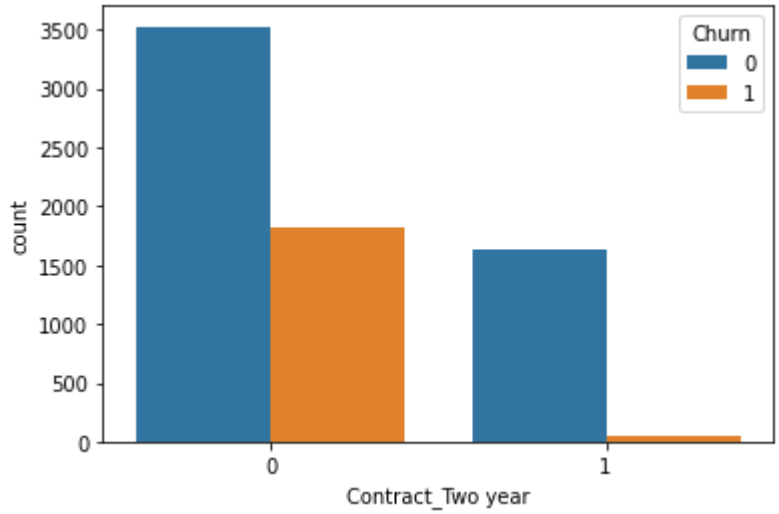

Churn
0
1
3000
2500
2000
1500
1000
500
0
count
0
1
InternetService_DSL

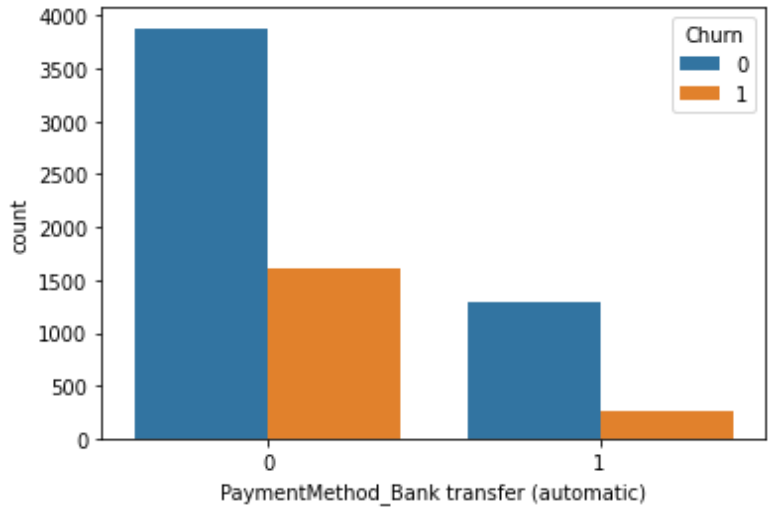

Churn
0
1
count
3500
3000
2500
2000
1500
1000
500
0
0
1
InternetService_Fiber optic

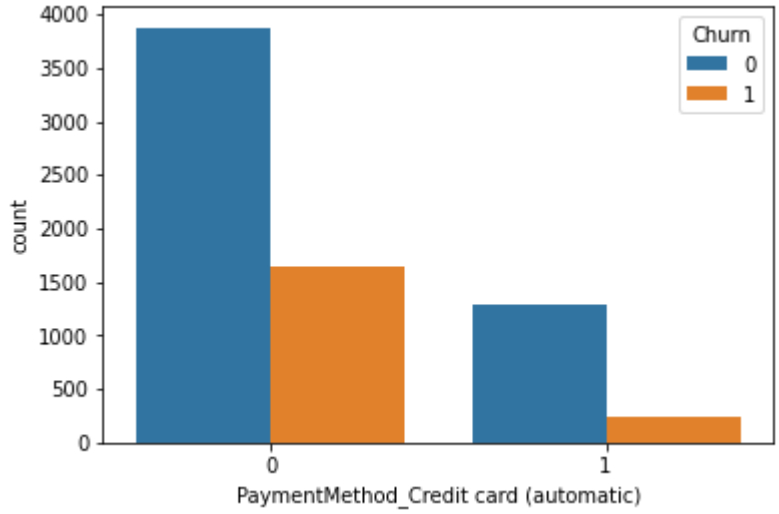

Churn
0
1
count
3500
3000
2500
2000
1500
1000
500
0
0
1
InternetService_No

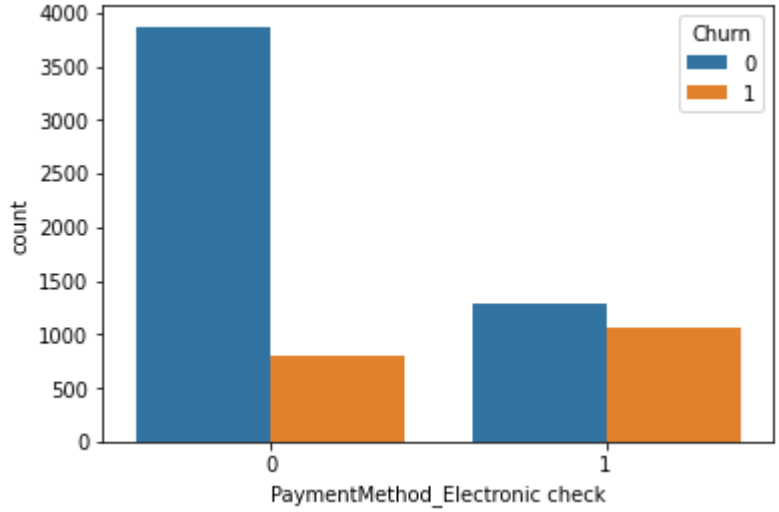

Churn
0
1
count
3000
2500
2000
1500
1000
500
0
0
1
Contract_Month-to-month

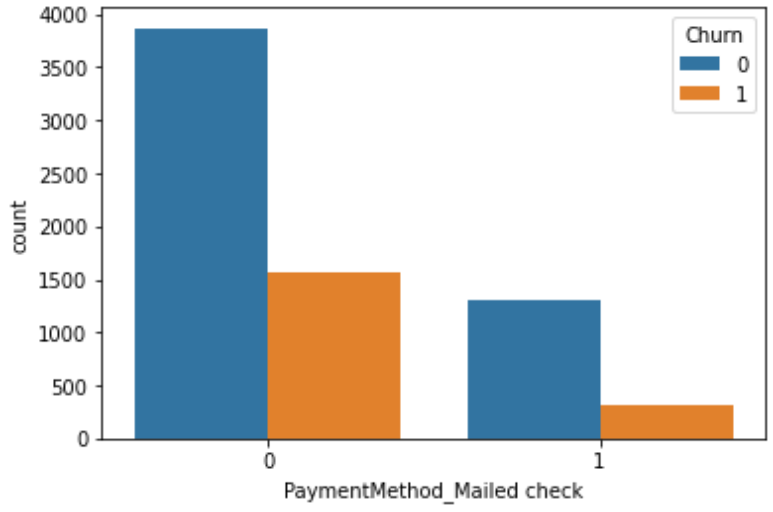

Churn
0
1
count
Contract_One year

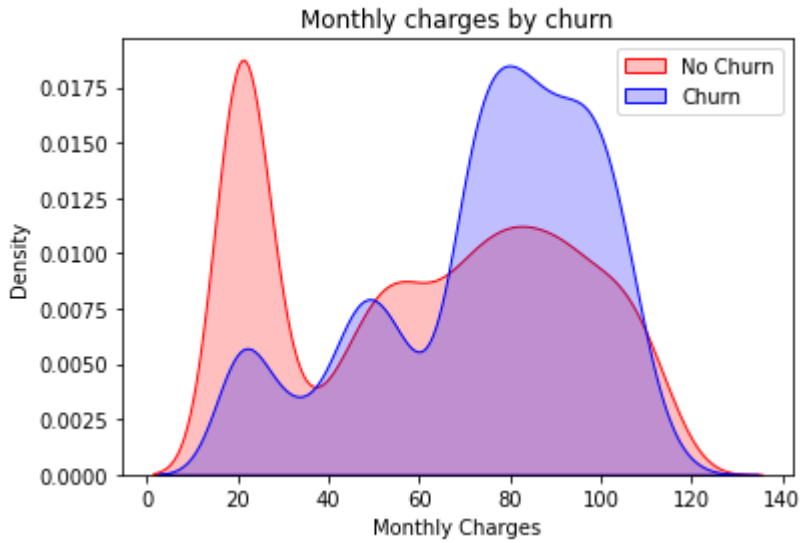

Churn
0
1
count
Contract_Two year


Churn
0
1
count
PaymentMethod_Bank transfer (automatic)


Churn
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
PaymentMethod_Credit card (automatic)


Churn
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
PaymentMethod_Electronic check


Churn
0
1
count
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
PaymentMethod_Mailed check


Monthly charges by churn
No Churn
Churn
Density
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
0
20
40
60
80
100
120
140
Monthly Charges

CHAPTER 3

From the dataset and its analysis, we have concluded that there are majorly 3 groups of customers with similar behavior. From now on the company can orient its product and service in the direction of these groups. This will enable them to reach out to more customers and easily promote their products and services in different parts of the society.

We have successfully categorized our data frame into 3 segments of similar Customer Behavior. Using this segmentation algorithm, we can categorize data of new customers into the different clusters they belong to depending on their interests.

We have also concluded that the Electronic check medium are the highest churners, Monthly customers are more likely to churn because of no contract terms, as they are free to go customers, No Online security, No Tech Support category are high churnersNon senior Citizens are high churners. HIGH Churn seen in case of Month to month contracts, No online security, No Tech support, First year of subscription and Fibre Optics Internet. LOW Churn is seen in case of Long term contracts, Subscriptions without internet service and The customers engaged for 5+ years. Factors like Gender, Availability of PhoneService and of multiple lines have almost NO impact on Churn. With Random Forest Classifier, also we are able to get quite good results, in fact better than Decision Tree

If the company is able to improve upon these, it will experience a lesser churn rate thereby increasing more profits than losses.

Future Scope

- This model can be forecasted across various businesses.
- Combining customer demographic and behavioural data with customer feedback data, helps you capture what various persons want and need from your company.
- Business will possess stronger customer focus and market clarity, allowing it to scale in a far more predictable and efficient manner.
- Churn by line of business

- Churn by product group
- Model can be built to support Campaign responses

REFERENCES

https://www.kaggle.com/

https://towardsdatascience.com/

APPENDICES

Appendix 1

Notebook link - Customer Segmentation.ipynb

```
#Importing modules

import numpy as np from matplotlib import pyplot
as plt import pandas as pd import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D from
sklearn.model_selection import train_test_split
import warnings

warnings.filterwarnings('ignore')

#Importing Data

df=pd.read_csv('dataa.csv') #

Viewing Data

df.head() df.head()
df=pd.read_csv("dataa.csv",index_col="customerID")
df.tail() df.shape df.describe() df.columns # Data
Cleaning df.dtypes
df.columns.to_series().groupby(df.dtypes).groups
#Converting dtype of TotalCharges to Float
df['TotalCharges'] =
pd.to_numeric(df['TotalCharges'],errors='coerce')
df['TotalCharges'] = df['TotalCharges'].astype("float")
df.info() df.isnull().sum()

The Dataset has 20 columns and 7043 rows with no empty(NULL)
value df.isna().any()

There are missing values in the column TotalCharges

#Finding the average and filling the missing values
```

```
na_cols = df.isna().any()
 #reset the index of the column with wrong datatype or True value
na_cols = na_cols[na_cols == True].reset_index()

na_cols = na_cols["index"].tolist()
for col in df.columns[1:]:
    if col in na_cols:
       if df[col].dtype != 'object':
           df[col] =  df[col].fillna(df[col].mean()).round(0) #finding
average and filling the missing values

#To revalidate

df.isna().any()

Now no column has missing values df.dtypes

df.columns.to_series().groupby(df.dtypes).groups

#Replacing similar values

df.replace('No internet service','No',inplace=True)
df.replace('No phone service','No',inplace=True)

yes_no=df['Partner','Dependents','PhoneService','MultipleLines','Onli
neSecurity','OnlineBackup',
'DeviceProtection','TechSupport','StreamingTV','StreamingMovies','P
aperlessBilling','Churn'] for col in yes_no:
   df[col].replace({'Yes': 1,'No': 0},inplace=True)
df.tail()

# Analysing Data

#Histograms

df2 = df.sample(frac=0.15)
df2.shape fig =
plt.figure(figsize=(18, 12)) for i
in range(df2.shape[1]):
plt.subplot(4, 5, i + 1) f =
plt.gca()
   f.set_title(df2.columns.values[i],color='red')
   vals = np.size(df2.iloc[:, i].unique()) if vals
   >= 100:
      vals = 100
   plt.hist(df2.iloc[:, i], bins=vals, color = 'orange')
```

```
plt.tight_layout(rect=[0, 0.1, 1, 0.9])
plt.suptitle('\nHistograms\n',fontsize=25,color='brown')

#Category Plot for comparison

sns.catplot(x = "SeniorCitizen", hue="InternetService",
kind="count", data=df) plt.title('Distribution of Senior
Citizens'); plt.show() #Count Plot for number

plt.figure(1, figsize=(10,5))
a=['#1933b5','#c48aff','#7e41e0','cyan'] sns.countplot(x
="PaymentMethod",palette=a, data=df) plt.show()

#Violin Plots for seeing highest occurence

fig = plt.figure(figsize=(10, 5)) a=['#4fe32d','#35e9f0','#ff8ab1']
sns.violinplot(y = "tenure", x = "gender", hue="Contract",palette=a,
data = df2) plt.show()

#Histogram of monthly charges by gender

plt.hist('MonthlyCharges', data=df[df['gender'] == 'Male'], alpha=0.5,
label='Male') plt.hist('MonthlyCharges', data=df[df['gender'] ==
'Female'], alpha=0.5, label='Female') plt.title('Distribution of
Monthly Charges by Gender') plt.xlabel('Monthly Charges')
plt.legend() plt.show()

#Relation plot between Monthly Charges and Tenure with
respect to Total Charges

sns.relplot(x="tenure",hue="TotalCharges",y="MonthlyCharges",da
t a=df2) plt.show()

#Heatmap for correlation

plt.figure(1, figsize=(12,8))
sns.heatmap(df.corr(), annot=True);
plt.title('Correlation Heatmap');

#Countplot of the columns with most variation

df3=df2[['tenure','MonthlyCharges','Contract','PaymentMethod']]
plt.figure(1, figsize=(35,30))
for i in range(df3.shape[1]):
plt.subplot(4, 2, i + 1) f =
plt.gca()
```

```
    f.set_title(df3.columns.values[i],color='red')

    sns.countplot(df3.iloc[:,i], color = 'cyan')
plt.show()
```

#Finding K for KMeans Clustering

```
df4=df2[['tenure','MonthlyCharges','TotalCharges']
] from sklearn.cluster import KMeans wcss = [] for
k in range(1,11):
    kmeans = KMeans(n_clusters=k, init="k-means++")
    kmeans.fit(df4.iloc[:,0:])
wcss.append(kmeans.inertia_) plt.figure(figsize=(12,5))
plt.grid(alpha=0.3) plt.plot(range(1,11),wcss,
linewidth=2, color="purple", marker ="8") plt.xlabel("K
Value") plt.xticks(np.arange(1,11,1)) plt.ylabel("WCSS")
plt.show()
```

#K comes out to be 3, i.e. number of clusters=3

```
X1=df4.iloc[:,0:].values

kmeans=KMeans(n_clusters=3)

label=kmeans.fit_predict(X1)

print(label)

print(kmeans.cluster_centers_) #2-
```

D Scatter Plot for clusters

```
plt.figure(1, figsize=(10,6))
plt.scatter(X1[:,0],X1[:,1],s=25,c=kmeans.labels_,cmap=plt.cm.Paire
d)
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],
color='indigo') plt.xlabel("Tenure") plt.ylabel("Monthly Charges")
plt.title('Clusters of Customers') plt.show()
```

#3-D Scatter Plot for clusters

```
fig = plt.figure(figsize = (12,10)) ax = fig.add_subplot(111,
projection='3d') ax.scatter(X1[label == 0,0],X1[label ==
0,1],X1[label == 0,2], s = 40
, color = 'orange', label = "cluster 1")
ax.scatter(X1[label == 1,0],X1[label == 1,1],X1[label == 1,2], s = 40
, color = 'blue', label = "cluster 2")
ax.scatter(X1[label == 2,0],X1[label == 2,1],X1[label == 2,2], s = 40
```

```
, color = 'green', label = "cluster 3")
ax.set_xlabel('Tenure-->')
ax.set_ylabel('Monthly Charges-->')
ax.set_zlabel('Total Charges-->')
ax.legend() plt.show()
```

Appendix 2

Notebook link - Customer_Churn.ipynb

```
import pandas as pd import numpy as np import
matplotlib.pyplot as plt import seaborn as sns from
sklearn.model_selection import train_test_split from
sklearn.preprocessing import LabelEncoder from
sklearn.linear_model import LogisticRegression from
sklearn.neighbors import KNeighborsClassifier from
sklearn.svm import SVC from sklearn.tree import
DecisionTreeClassifier from sklearn.ensemble import
RandomForestClassifier from sklearn import metrics
from sklearn import linear_model from
sklearn.metrics import accuracy_score from
sklearn.metrics import confusion_matrix from
sklearn.decomposition import PCA
df=pd.read_csv("/content/dataa.csv") df.head()
df.tail()
df.drop('customerID',axis='columns',inplace=True)
df.shape

labels = ['Male','Female'] fig1, ax1 =
plt.subplots() ax1.pie(sizes, labels=labels,
autopct='%1.1f%%')
ax1.axis('equal')
plt.show()

#There is an almost equal gender distribution.

df.info()

#Total charges and monthly charges are of diff. datatypes.


df.describe()
```

```
#75% of the customers have tenure of less than 55 months

df.groupby('Contract').sum()

df[pd.to_numeric(df.TotalCharges,errors='coerce').isnull()]

coerce is used for ignoring spaces in string
returning rows whose total charges are null.

#Dropping those columns which have empty value of Total Charges

df = df[df.TotalCharges!=' ']

df.shape

df[df.Churn=='No'].tenure

no=df[df.Churn=='No'].MonthlyCharges
yes=df[df.Churn=='Yes'].MonthlyCharges
plt.xlabel("Monthly Charges") plt.ylabel("Number
Of Customers") plt.hist([yes, no], rwidth=0.95,
color=['green','red']) plt.show() df.isnull().sum()

df.replace('No internet service','No',inplace=True)

df.replace('No phone service','No',inplace=True)

df['gender'].replace({'Female':1,'Male':0},inplace=True)

yes_no =
['Partner','Dependents','PhoneService','MultipleLines','OnlineSecurity
','OnlineBackup',
'DeviceProtection','TechSupport','StreamingTV','StreamingMovies','
P aperlessBilling','Churn'] for col in yes_no: df[col].replace({'Yes':
1,'No': 0},inplace=True)

#One hot encoding: Finding unique values per feature and transform
data

df= pd.get_dummies(data=df,

columns=['InternetService','Contract','PaymentMethod'])

df.columns df.head()

churn_count=pd.DataFrame(df['Churn'].value_counts())

churn_count #Overall Churn rate is low


df.corr()
```

```
#Heatmap showing correlation

sns.heatmap(df.corr())

#Displaying positively and negatively correlated features.

plt.figure(figsize=(20,8))
df.corr()['Churn'].sort_values(ascending = False).plot(kind='bar')

#1.  Fiber optic is the preferred choice for internet service.
#2.  Most of the customers are on a monthly contract.

sns.countplot(x='Churn',                    data=df)

sns.barplot(x='tenure',              y='Churn',data=df)

sns.barplot(x='Churn',y='SeniorCitizen',data=df)

sns.barplot(x='Churn',y='OnlineBackup',data=df)

sns.barplot(x='Churn',y='DeviceProtection',data=df
```

) **Univariate Analysis**

```
for i, predictor in enumerate(df.drop(columns=['Churn',
'TotalCharges', 'MonthlyCharges'])):
    plt.figure(i)
    sns.countplot(data=df, x=predictor, hue='Churn')

df = pd.get_dummies(df)
df.head()

Mth = sns.kdeplot(df.MonthlyCharges[(df["Churn"] == 0) ],
            color="Red", shade = True)
Mth = sns.kdeplot(df.MonthlyCharges[(df["Churn"] == 1) ],
            ax =Mth, color="Blue", shade= True)
Mth.legend(["No Churn","Churn"],loc='upper right') Mth.set_ylabel('Density')
Mth.set_xlabel('Monthly Charges')
Mth.set_title('Monthly charges by churn')

High Churn at high monthly Charges

sns.countplot(data = df, x= col,
order=df['Partner'].value_counts().index,hue='gender',palette='bright'
)
plt.show()
```

```
sns.countplot(data = df, x= col,
order=df['TechSupport'].value_counts().index,hue='gender',palette='
b right') plt.show()

sns.countplot(data = df, x= col,
order=df['SeniorCitizen'].value_counts().index,hue='gender',palette='
bright') plt.show()

#Non senior citizens are high churners

Y = df['Churn']
X = df.drop('Churn',axis='columns')

le = LabelEncoder()
Y = le.fit_transform(Y)

Train_X, Test_X, Train_Y, Test_Y = train_test_split(X, Y,
test_size=0.2)

#20% records are test data and 80% are training

Knn = KNeighborsClassifier(n_neighbors =

4).fit(Train_X,Train_Y) prediction = Knn.predict(Test_X)
```

Accuracy of KNN

```
accuracy_score(Test_Y, prediction)

reg = linear_model.LogisticRegression()

reg.fit(Train_X,Train_Y) predict =

reg.predict(Test_X)
```

Accuracy of Logistic Regression

```
accuracy_score(Test_Y, predict)

confu_mat = confusion_matrix(Test_Y, predict)
print(confu_mat)

lm1 =

DecisionTreeClassifier().fit(Train_X,Train_Y)

forest_pred = lm1.predict(Test_X)
```

Accuracy of

Decision Tree Classifier accuracy_score(Test_Y,

forest_pred)

confu_mat = confusion_matrix(Test_Y, forest_pred)
print(forest_pred)

lm2 = RandomForestClassifier(n_estimators= 100,random_state=
20).fit(Train_X,Train_Y) forest_pred =

lm2.predict(Test_X) **Accuracy of**

Random Forest Classifier

accuracy_score(Test_Y, forest_pred)

confu_mat = confusion_matrix(Test_Y, forest_pred)
print(confu_mat)

Conclusion : Random Forest Classifier has maximum accuracy

con=[] cat=[] for i in
df.columns: if
df[i].dtypes=='int64':
con.append(i)
elif df[i].dtypes=='object':

cat.append(i) #Continuous

Variables con

#Categorical Variables

cat plt.figure(figsize =
(16,20))
for i in range(0,14):
plt.subplot(6,5,i+1)
sns.distplot(df[con[i]])