**MCSD2123**

**Group Assignment 3**

|  |  |
| --- | --- |
| **Name** | **Matric Number** |
| CHONG XIAN JUN | MCS211047 |
| TAN FEI ZHI | MCS221016 |
| TEH HUNG WEI | MCS212011 |

**SECTION 1: Exploratory Data Analysis (EDA)**

The dataset we selected to study with is the Dataset-B-TelcoChurn dataset that is developed to predict behaviors to retain customers. This dataset is chosen for having fewer categorical features compared to others, for which distance-based clustering algorithms can perform better. Python was selected as the programming language to be used for this study as Python consists of various useful libraries which could facilitate our work. The dataset is uploaded to our GitHub user profile and imported to pandas framework.

The clustering tasks carried out are to explore different ways to cluster the customers and describe customers with respect to their churn status. In this section, the following aspects of the dataset are discussed.

1. The summary information of the dataset, including statistical description such as mean, median, outliers; structure such as data types, null values, etc.

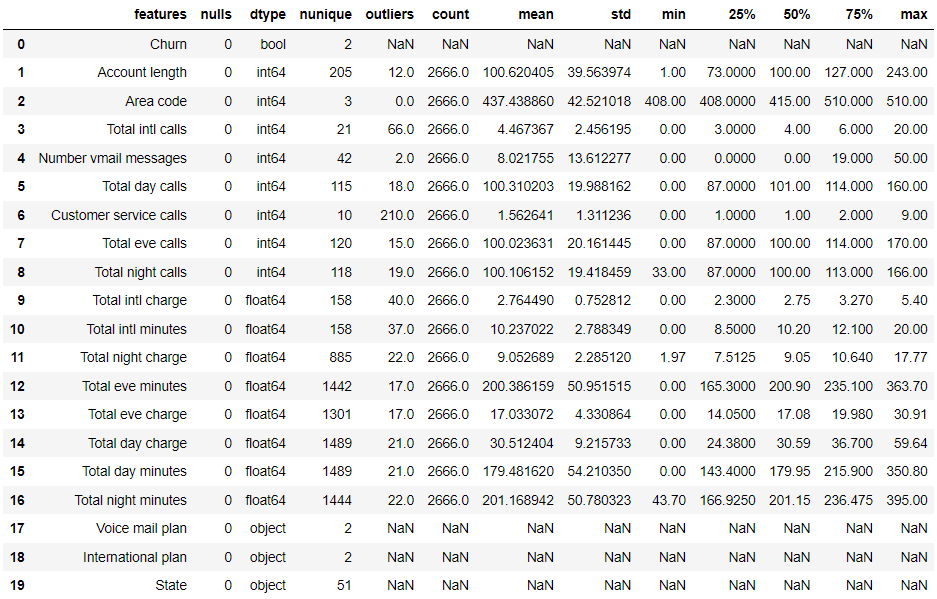


Table 1 Observations of the dataframe information summarised

|  |  |  |
| --- | --- | --- |
| EDA Step | Method | Result |
| 1. Check dataframe shape | df.shape | (2666, 20) |
| 1. Check null values | df.isnull().sum() | No null |
| 1. Check data types | df.dtypes() | 1 boolean, 8 float, 8 integer, 3 object |
| 1. Check unique elements | df.nunique() | Number of unique value in each column, refer Figure 1 |
| 1. Check outliers | IQR method | Many outliers in all numerical features |
| 1. Check statistical summary info | df.describe() | Refer figure above |

This dataset consists of 2666 data with 20 features. Amongst the 20 features, 16 of them are numerical features, 3 of them are categorical features (state,international plan, and voice mail plan), and 1 boolean feaeture (churn). While the `Area code` feature's dtype is `int64`, it has only 3 unique values, indicating that it is more appropriate to be transformed to categorical feature in data preprocessing step.

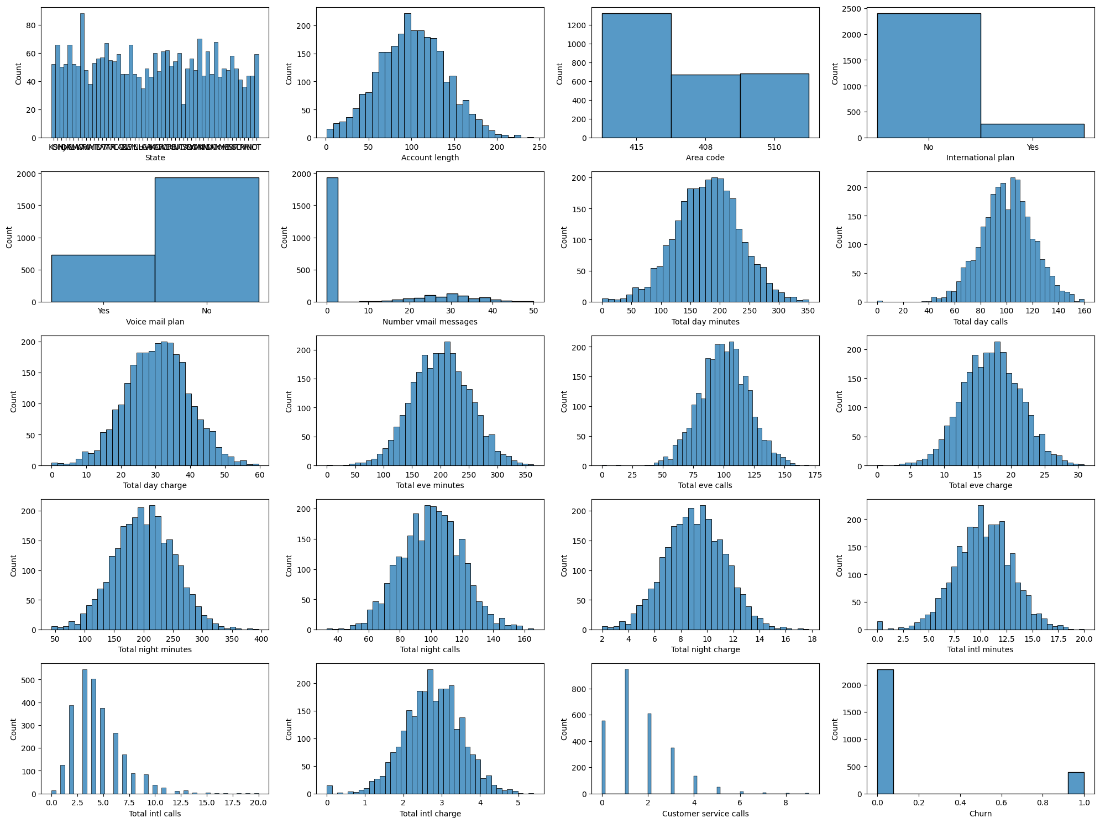
Besides, there are many outliers for the numerical features, assuming outliers are defined by the Interquartile range (IQR) method. It shows that most of the numerical features have long tailed distribution consisting of a noticeable amount of edge cases.

1. Explore features within dataset

To further understand the features in the dataset, each of the features are plotted on a histogram below.

All numerical features have mostly normal distribution, which facilitates the use of clustering models such as k-means as k-means assumes the data is spherical and normally distributed. From the graphs, we can also see that there are no strong outliers observed among the numerical features. Thus, the outliers defined by IQR method are viewed as edge cases and preserved instead of being removed.

As for categorical features, most of them have class imbalance to different extents. However, as our task is unsupervised clustering, it does not require treatment such as oversampling and undersampling of target variables commonly used in classification tasks.

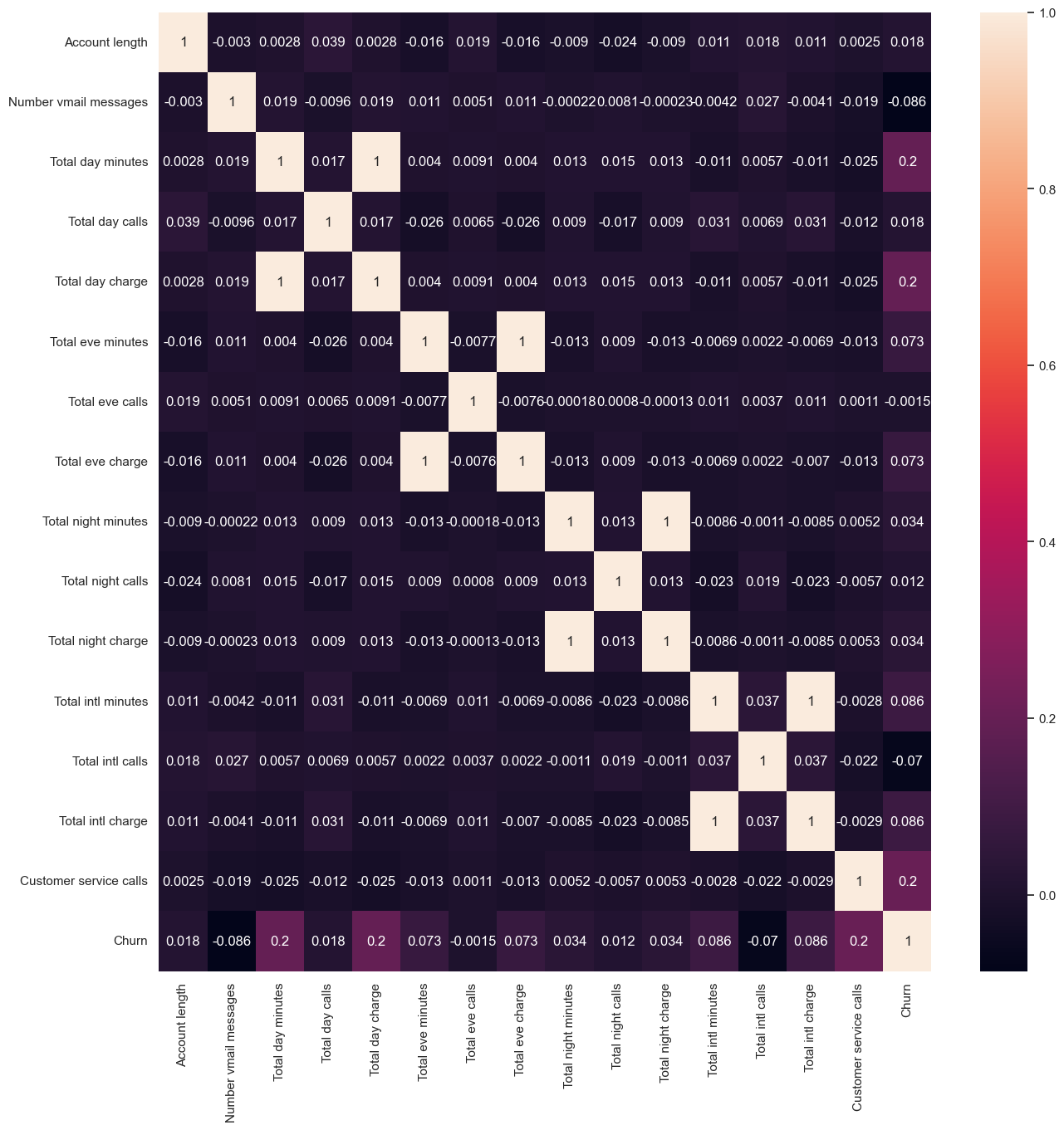


Other than exploring univariate distribution, the correlation between features is explored with Pearson correlation coefficient. In Figure 3, a correlation matrix is plotted to understand the correlation between numerical features of the dataset.

Upon inspecting the data, we observed that most features have very low correlation with one another. However, there are a few features that have correlation coefficient with one another, which might result in multicollinearity. Such highly correlated features are listed as below:

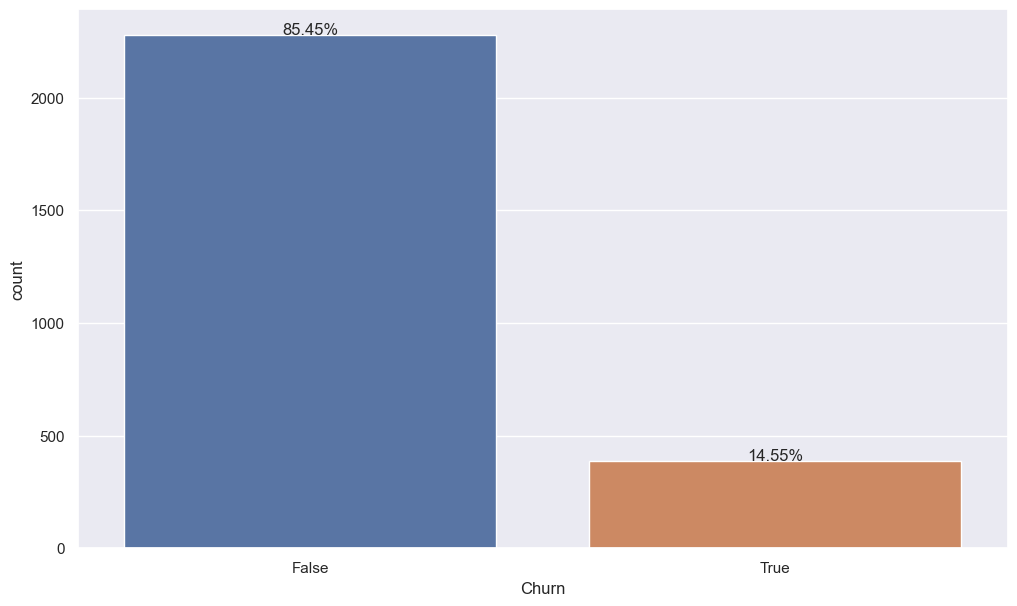
1. “Total day charge” and “Total day minutes”
2. “Total eve charge” and “Total eve minutes”
3. “Total night charge” and “Total night minutes”
4. “Total intl charge” and “Total intl minutes”

These features will be further treated in the data preparation section to eliminate the redundant features.



1. Exploratory analysis on data features with respect to the churn status of customers

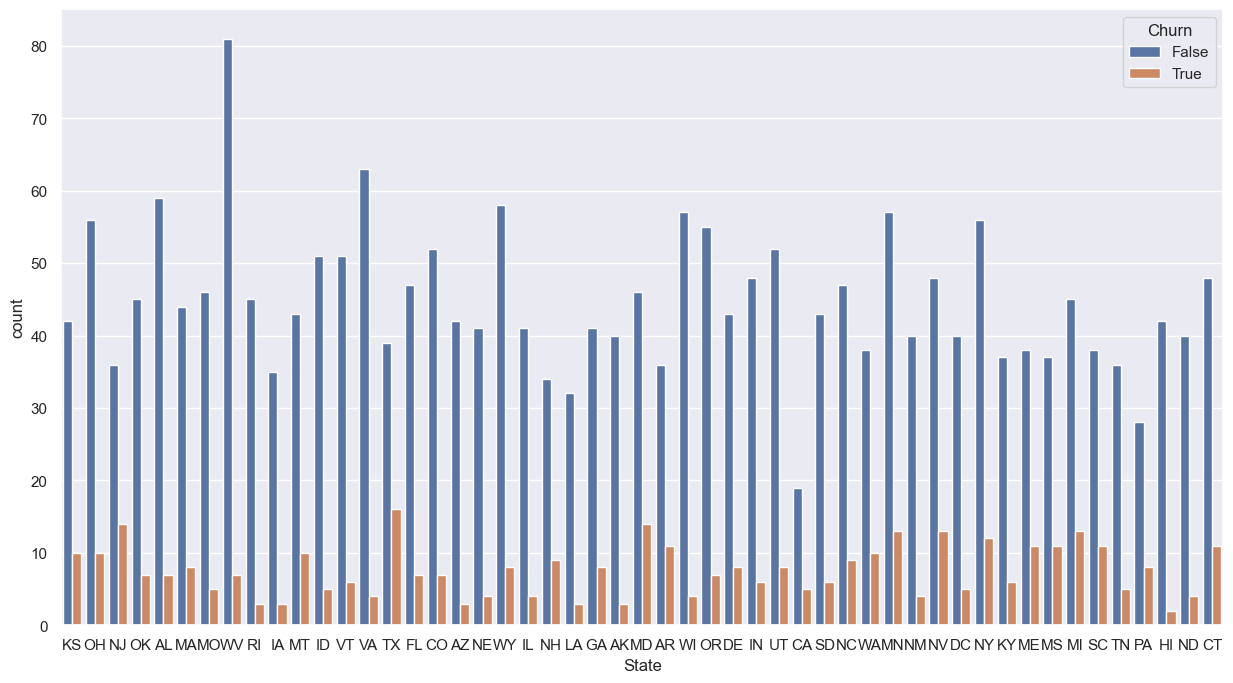
As this dataset is developed to predict customer behaviors and develop customer retention programs, features are explored with respect to the churn status of customer. Firstly, the distribution of churn status in the dataset is shown. It is observed that most customers (85.45%) in the dataset do not churn.

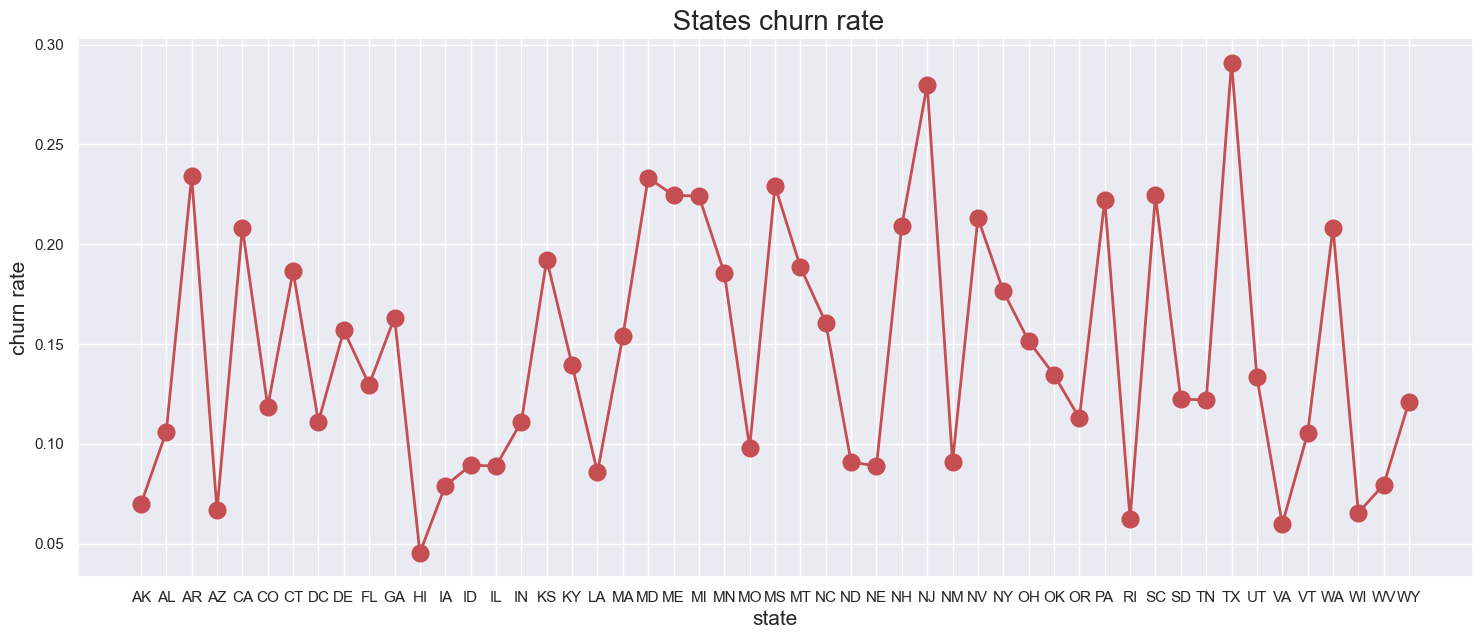


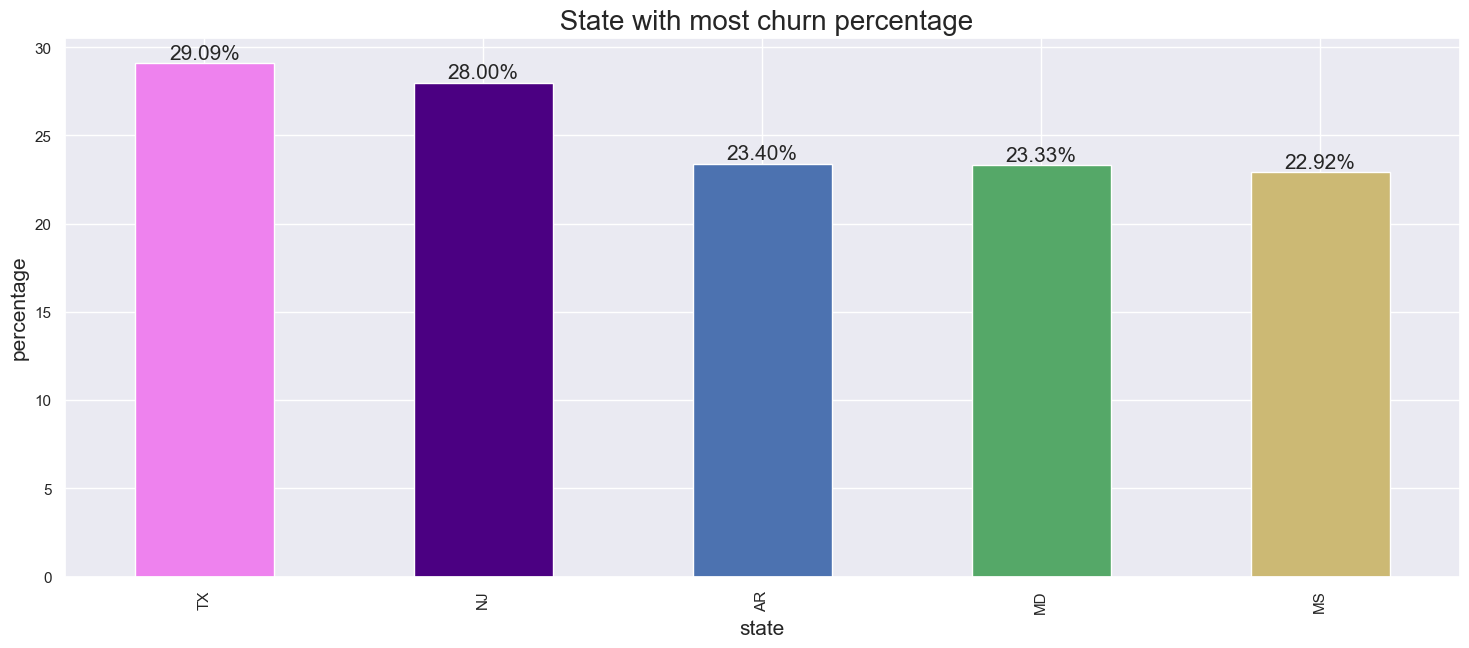
Based on domain experiences, several interesting features are explored with respect to churn, I.e., “States vs Churn”, “Area code vs Churn”, “International plan vs Churn”, “Voice mail plan vs Churn”, “Number vmail messages vs Churn”, and “Customer service calls vs Churn”. From there, we will determine which columns are considered as the important feature or redundant feature which can be removed from the dataset.

1. States vs Churn:

In figure below, it shows the number of users in every state. For each state, the blue bar represents non-churn user, and orange bar represents churn user. In short, in WV state, the number of users is the highest, whereas in CA state, the number of users is the lowest. Next, the churn rate for every state is shown below. We observed there is big difference in churn rate for every state, and states with top 5 highest churn rate are TX (29.1%), NU (28.0%), AR (23.4%), MD (23.3%), and MS (22.9%). From this information obtained, we can predict that the churn rate is affected by the network coverage area. If the network condition is not good for that state, the churn rate is higher and vice versa.

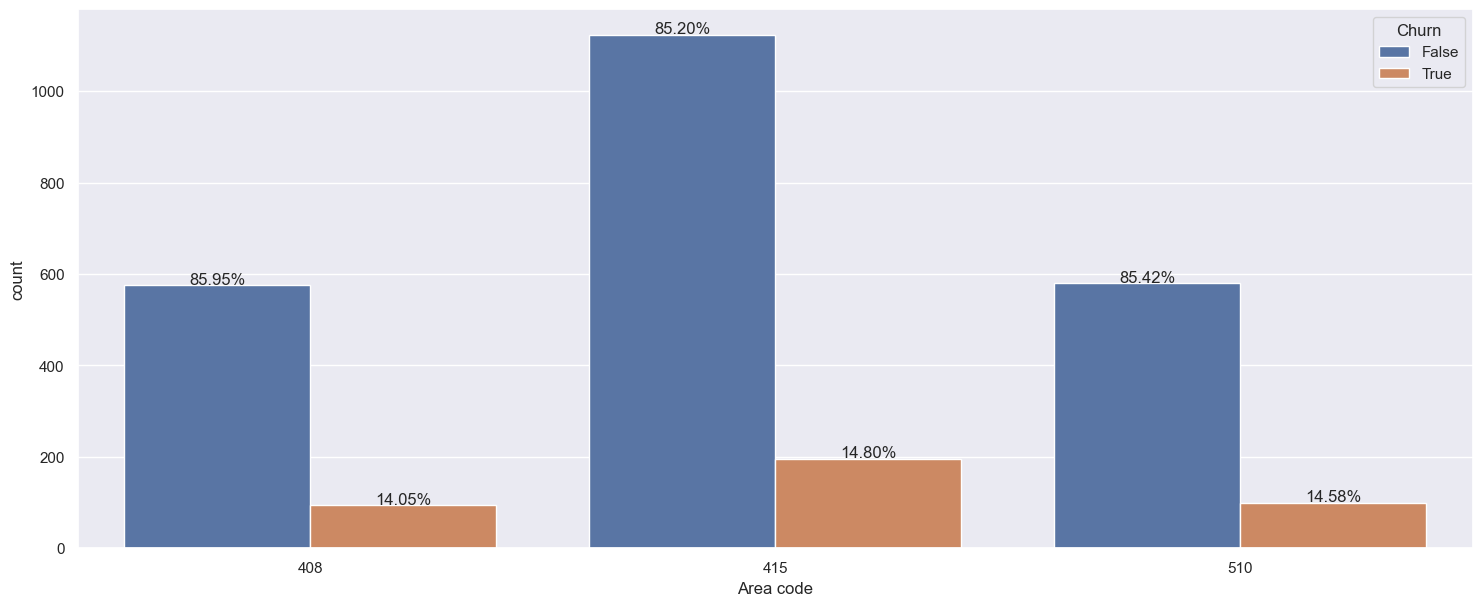






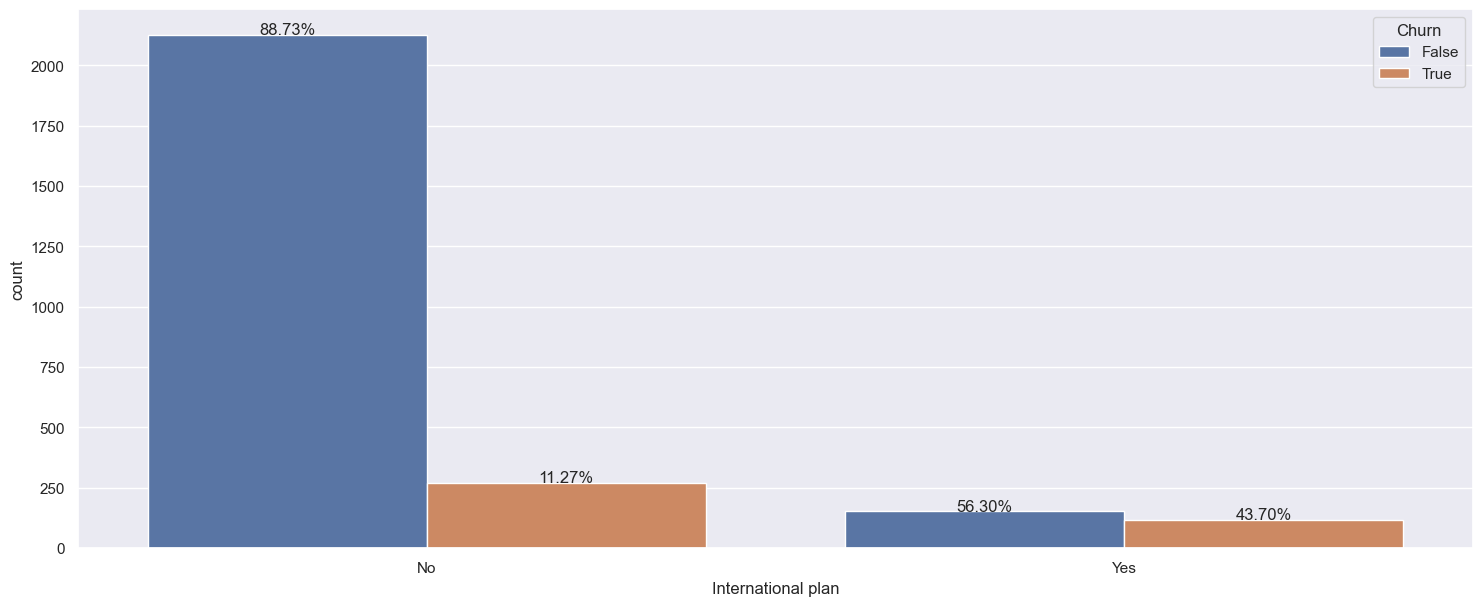
1. Area code vs Churn:

In the figure below, it shows the number of counts for different area codes with respect to churn rate. Basically, we observed that area code 415 has a higher number of users as compared to that in 408 and 510. Secondly, the churn rate for all area codes is roughly the same for 408, 415, and 510, with churn rates of 14.05%, 14.80%, and 14.58% respectively. This means that different area codes do not show much difference on churn rates, in other words, area code is not a significant feature for this dataset.



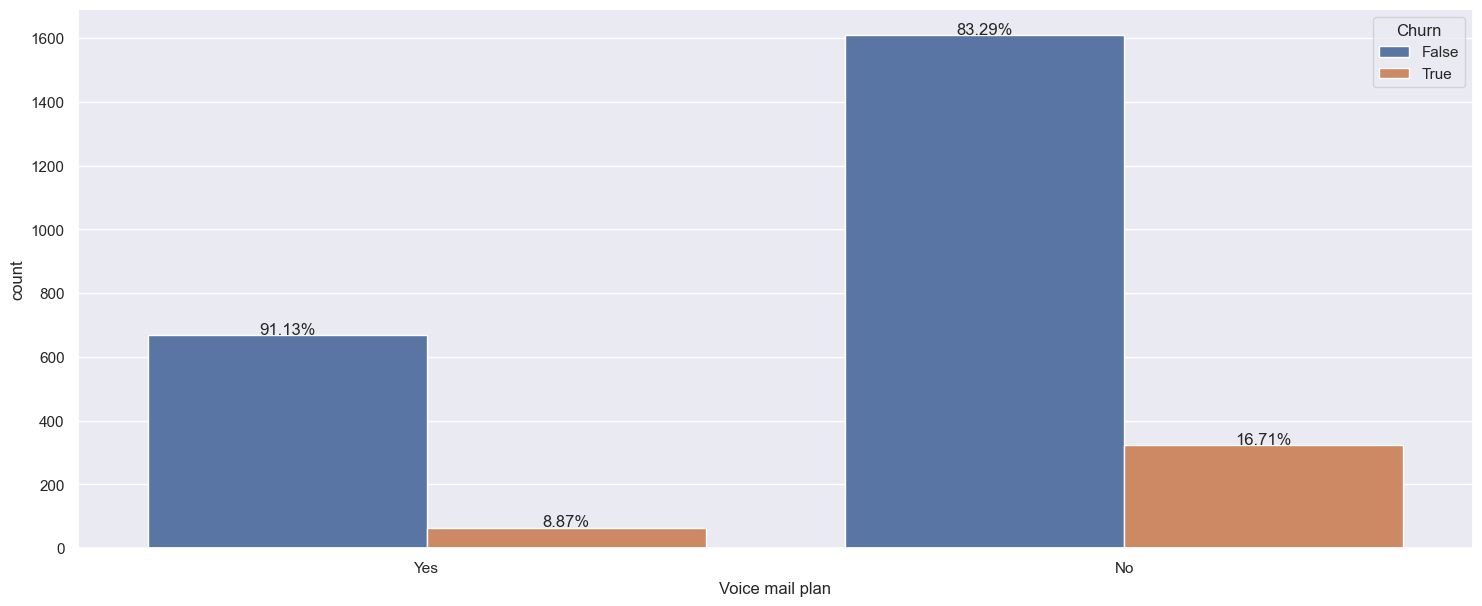
1. International plan vs Churn:

In the figure below, it shows the number of counts for international plan with respect to churn. We can easily determine that most of the users did not subscribe to the international plan. However, the churn rate is much higher for users that subscribed to international plan (43.70%) as compared to users that did not subscribe (11.27%). The other hypothesis we can make about it is the service for international plans is not good, or the charge is expensive, which could possibly lead to higher churn rate.



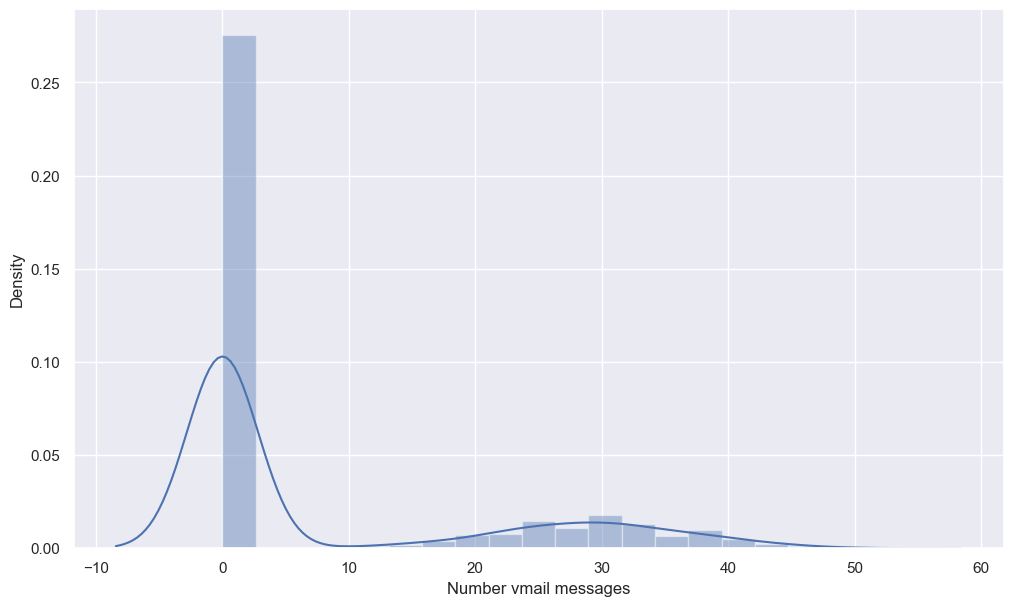
1. Voice mail plan vs Churn:

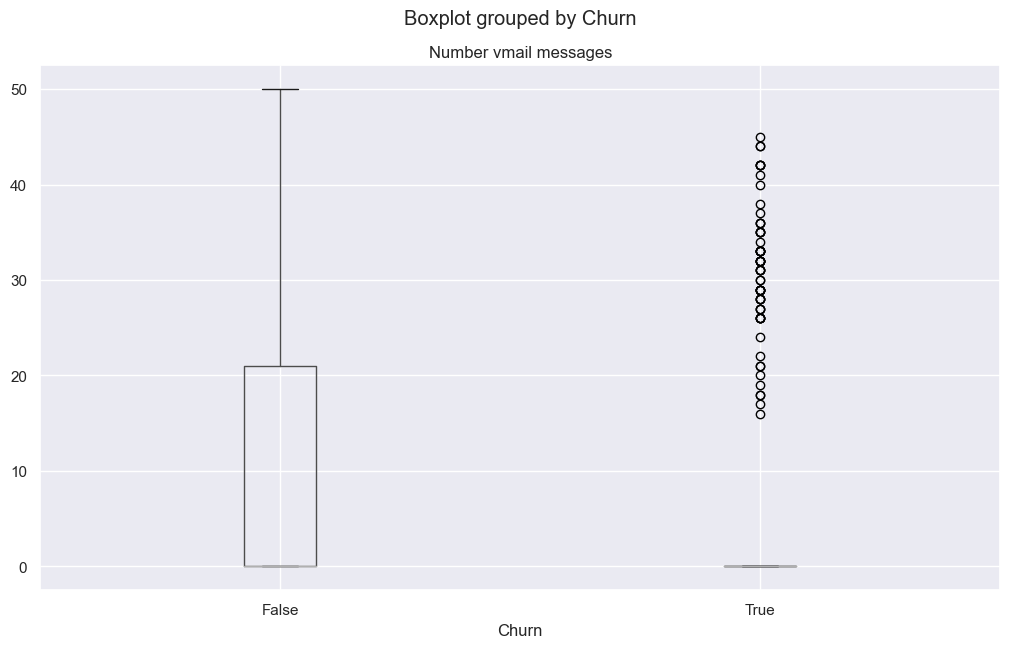
In the figure below, it shows the number of counts for voice mail plan with respect to churn. Generally, there are more users who did not subscribe to voice mail plan, and the churn rate is also higher as compared to user that subscribed voice mail plan. For this feature, it is harder for us to gain meaningful insight using univariate analysis, as we see that the information obtained from the graph is not that useful. Why does the churn rate is lower when a user tends to subscribe to a voice mail plan? Perhaps the service provided is satisfying the customer's needs and hence resulting in a lower churn rate. Next, we will further analyze the churn rate for users that subscribed voice mail plan with the help of other feature, which is number vmail messages.



1. Number vmail messages vs Churn:

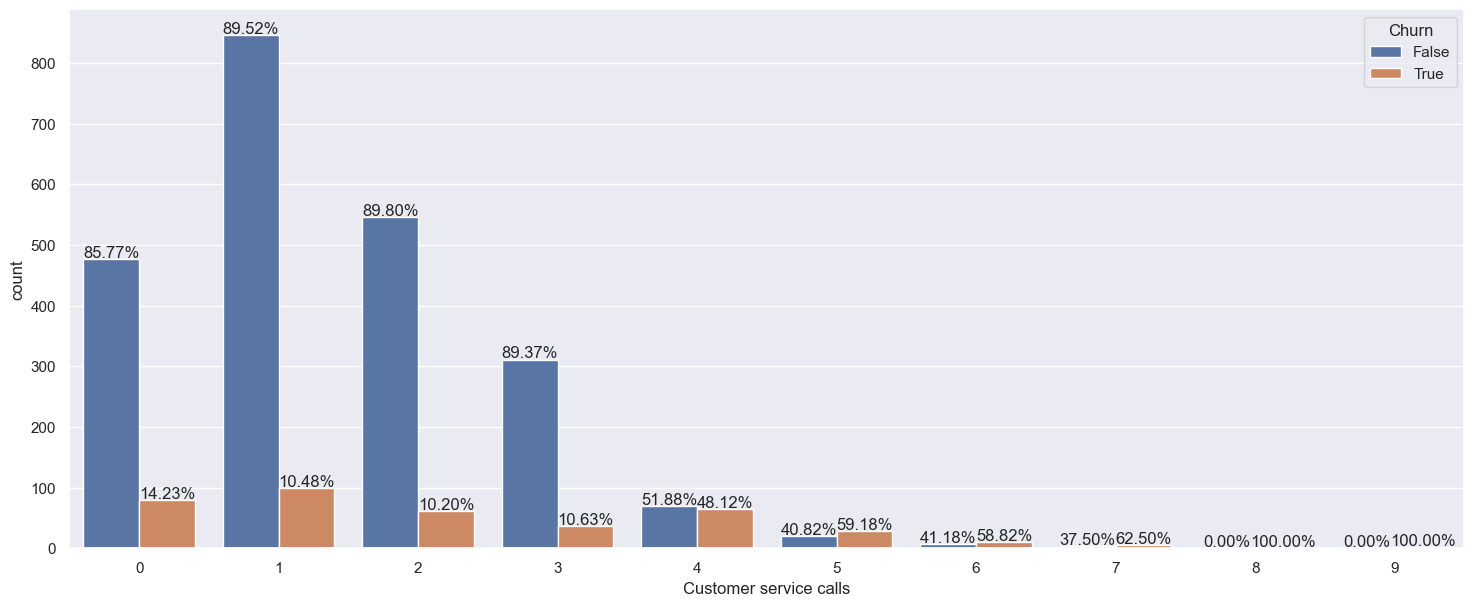
For number vmail messages feature, it is related to the voice mail plan. Basically, if the user did not subscribe voice mail plan, the number vmail messages will be zero. If they subscribed voice mail plan, there will be a different number of vmail messages. This scenario is illustrated in the below figures. Basically, boxplot for the number vmail messages with respect to churn has been plotted. From there, we can obtain a useful information, where the users are more likely to be churned as the number vmail messages approximately higher than 20.





1. Customer service calls vs Churn:

In the figure below, it shows the number of customer service calls with respect to churn. We can easily observe that as the number of customer service calls is increasing, the number of users is decreasing. This possibly indicates that the customer is able to solve their query or problem through the customer service call, therefore they did not call the customer service hotline again. Besides, another interesting information gained from the figure is that, as the number of customer service calls are more than 3, the churn rate increases sharply from 10.63% to 48.12%, and it keeps on going up as the number of customer service call increases. This shows that maybe the quality of customer service calls is not good, or the employee was not able to solve customer queries or problems, which lead to a higher churn rate.



**SECTION 2: Preparation of Dataset**

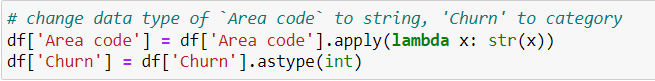
Data preparation is carried out before actual clustering analysis. There are several tasks to be completed in this section:

1. Format data type.
2. Perform feature engineering.
3. Feature scaling to standardize numerical features.

Note: As different clustering algorithms have different requirements on data representation, data encoding will be conducted at each clustering stage.

1. **Format data type**

The “Area code” is converted into string so that it can be recognized as categorical data in the unsupervised learning models. The “Churn” feature is converted to integer to ease the following preprocessing steps.



1. **Perform feature engineering**

As the clustering analysis is conducted with respect to the churn of teleco service, feature engineering process is carried out with two objectives:

* Reduce redundant features in the dataset to prevent the curse of dimensionality when doing clustering analysis.
* Ensure that the engineered features are useful predictors of whether a client will churn.

These objectives can be checked by fitting the engineered feature against a classification model.

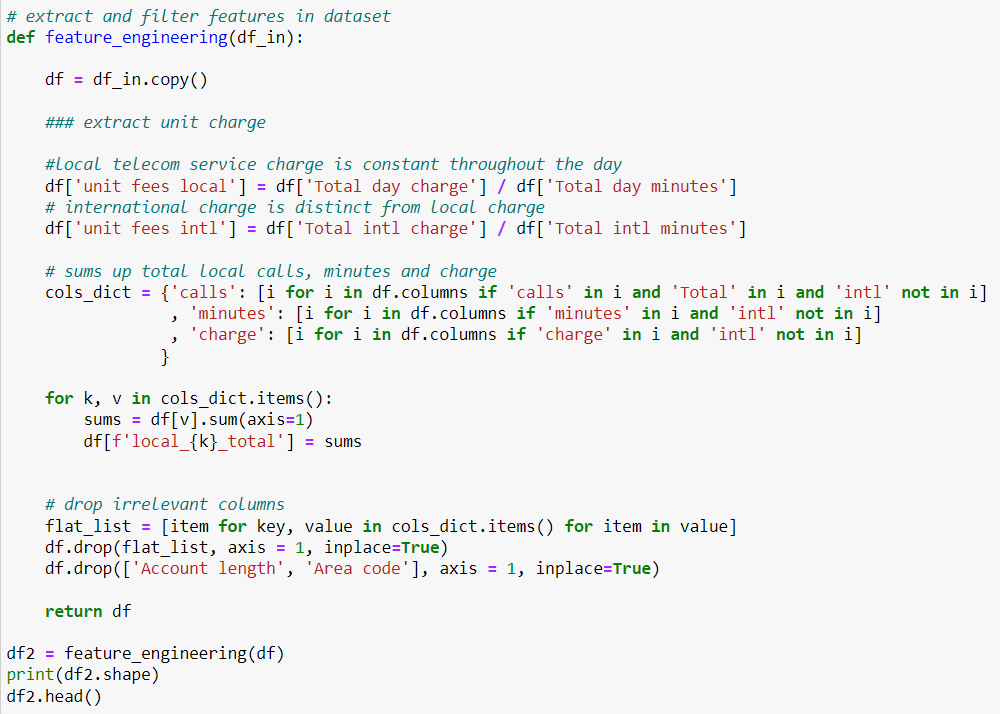
The feature engineering process is shown below. Feature extraction and selection are carried out to meet the objectives of feature engineering mentioned above.

Extracted columns:

* “unit fees local” and “unit fees intl” to find out the charge of service per minute of use. With this, we can remove columns "Total day charge", "Total eve charge", "Total night charge", "Total intl charge" that are colinear with "Total day minutes", "Total eve minutes ", "Total night minutes ", "Total intl minutes " respectively.
* “local\_calls\_total”, “local\_charge\_total”, “local\_minutes\_total” are extracted by adding up the local calls, charge and minutes throughout day, evening and night respectively. It allows us to further cut down the redundant features.

Dropped columns:

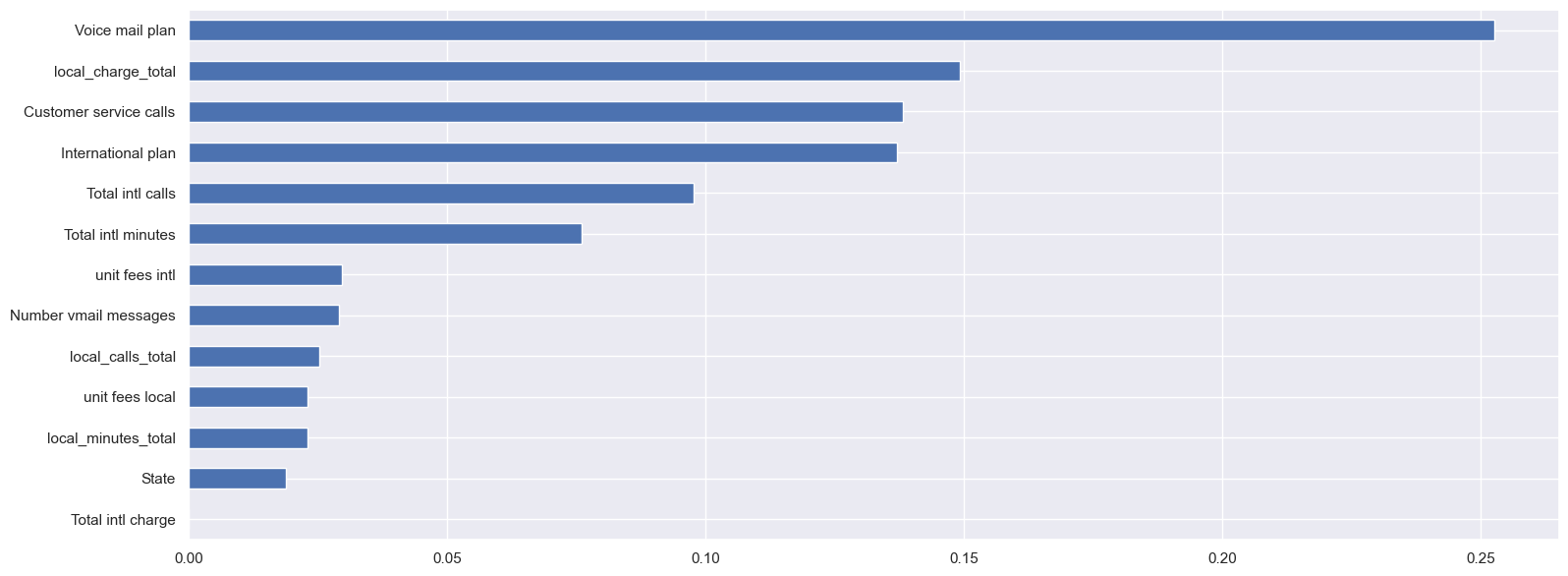
* 'Total day calls', 'Total eve calls', 'Total night calls', 'Total day minutes', 'Total eve minutes', 'Total night minutes', 'Total day charge', 'Total eve charge', 'Total night charge', 'Account length', 'Area code' are dropped from the dataset. (11 features)



The output features are fed into a XGBoost classifier, and it shows a significant improvement in its ability to detect churn in the dataset as observed by the increase in F1-score. F1-score is chosen as the measurement metric as there is class imbalance in the target feature of “Churn”.

baseline F1-score: 0.8127845436056559

F1-score (feature engineered): 0.8955737075350252



The output dataframe after feature engineering is illustrated in the figure below. It now consists of only 14 features.

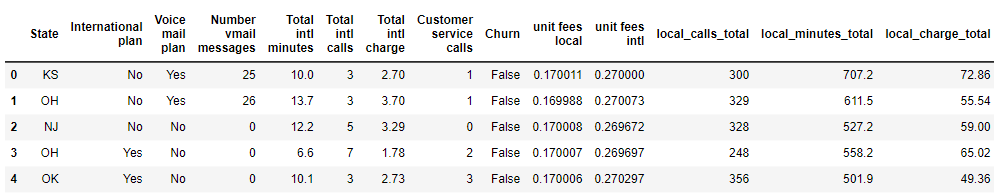
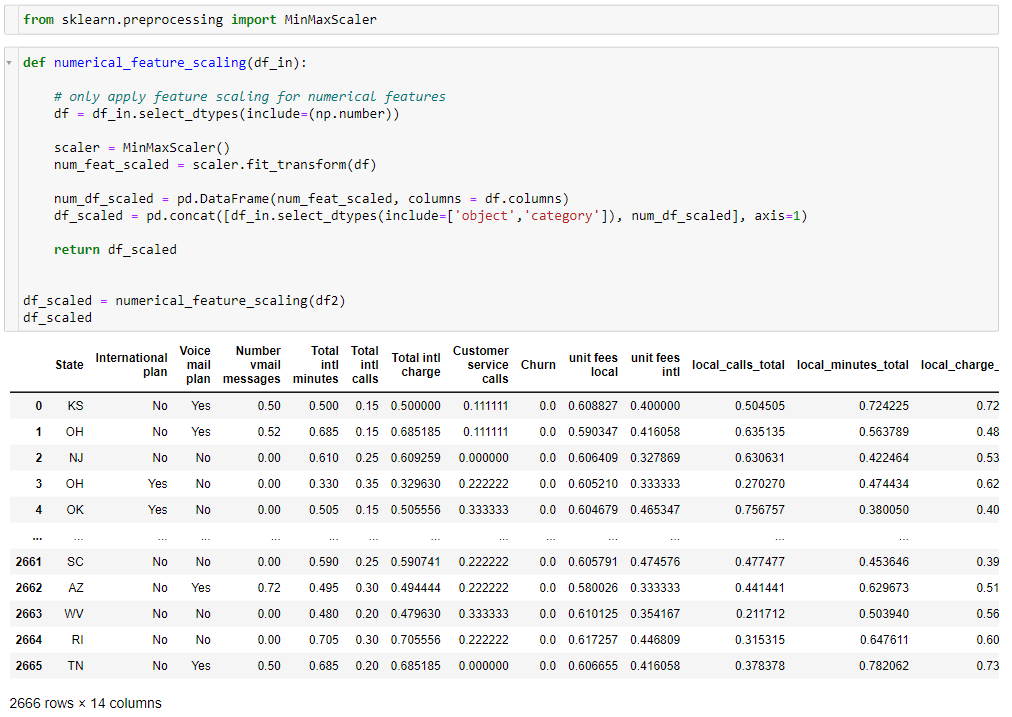


Figure 2.4 Output dataframe after feature engineering.

1. **Feature scaling to standardize numerical features.**

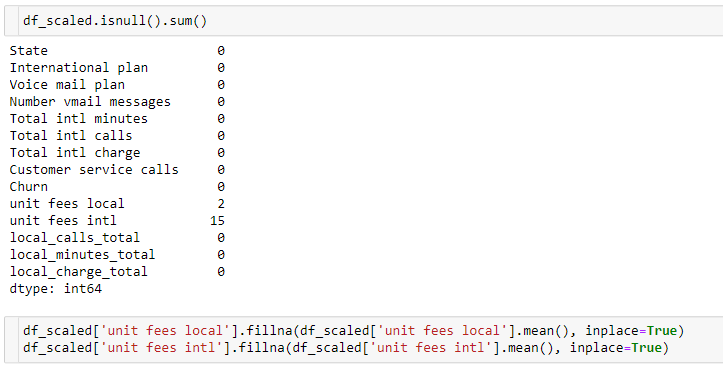
As there is no target feature to guide the learning process, most unsupervised learning models are based on the distances between different data points. Thus, feature scaling is essential prior to any unsupervised learning task.

As there are several categorical features in our data, of which distances are computed in the scale of 0 and 1, a *MinMaxScaler* is used to rescale the numerical data to a range between 0 and 1. By scaling the feature to a similar range, we can avoid the learning models from overly bias against categorical data. Figure below shows the details of feature scaling process which we transform numerical features to the range of 0 and 1.



1. **Impute null values**

In the feature engineering and scaling process, there might be some errors such as divide by 0 happening, which results in a few null values. As there are not many null values in the data, they are simply imputed with the mean values of their respective features.



**SECTION 3: Clustering Methodology**

As this study involves mixed data that contains both categorical and numerical data, Euclidean distance-based models such as K-means cannot be applied to the dataset. Even if we encode the categorical data, it remains mathematically illegal to apply Euclidean distance-based models on the dataset. Therefore, several less popular approaches are deployed.

In this study, two types of clustering methods were chosen in accordance with the brief requirement:

* **k-prototype** (partition-based clustering)
* Gower distance + **dbscan** (density-based clustering).

**Method 1: K-prototype**

K-prototype is a clustering model which includes a combination of k-means and k-modes models to achieve clustering of data points around certain prototypes (similar concept as centroids). Similar with K-means, the objective of k-Prototype is to group data points into k clusters in a way that minimizes the “distance” between data points and centroid. However, due to the existence of mixed data, a composite measure that integrates Euclidean distances for numerical variables and similarities for categorical variables is used.

Nonetheless, similarity and Euclidean distance are inevitably very different. To counter this challenge, an assumed gamma as weightage for categorical variables that decides its preference towards categorical or numerical variables is applied in the model. As there is no mathematically proven way- to verify the value of gamma, domain knowledge along with heavy trial and error are required to find out the appropriate gamma value for each clustering task. Smaller values (closer to 0) of gamma favour the numerical variables while those larger favour (away from 0) the categorical variables (Huang, 1997).

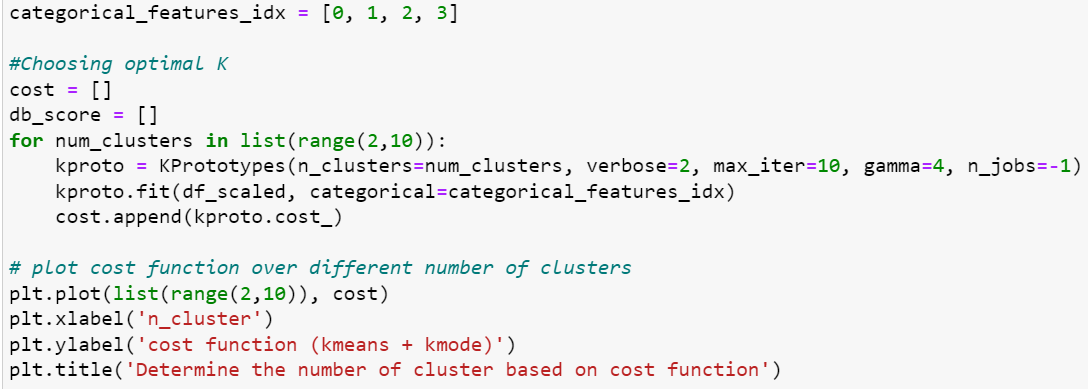
The clustering process is carried out in following steps:

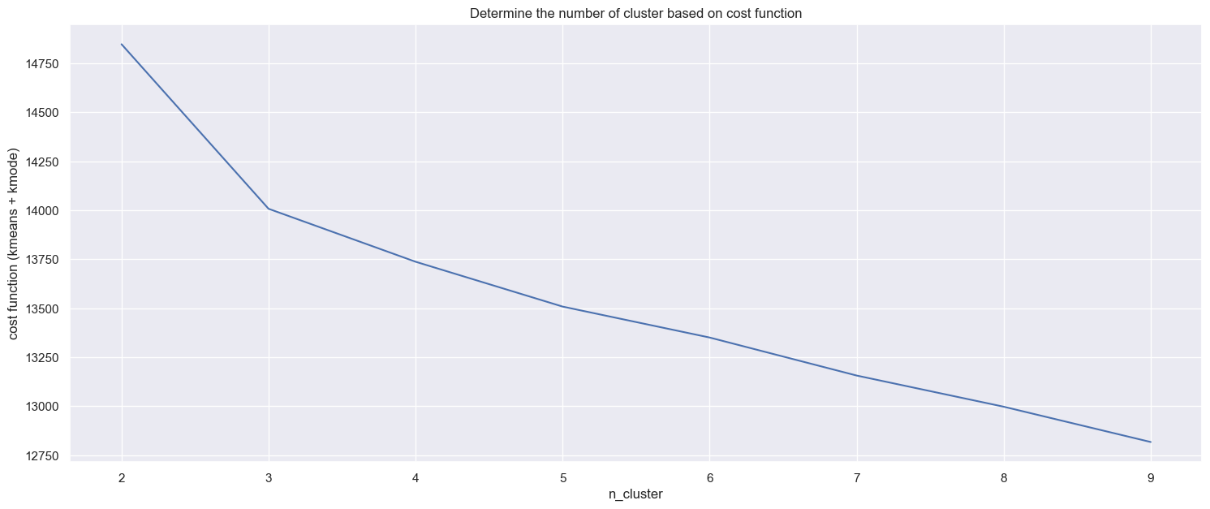
1. Parameter tuning (determine the number of clusters and gamma with elbow method).
2. Use the clustering model to predict clusters in dataset.
3. Visualize the clustering output to verify the quality of cluster.
4. **Parameter tuning with elbow method**

The K-prototype can be executed with *kprototype* package from the kmodes library.



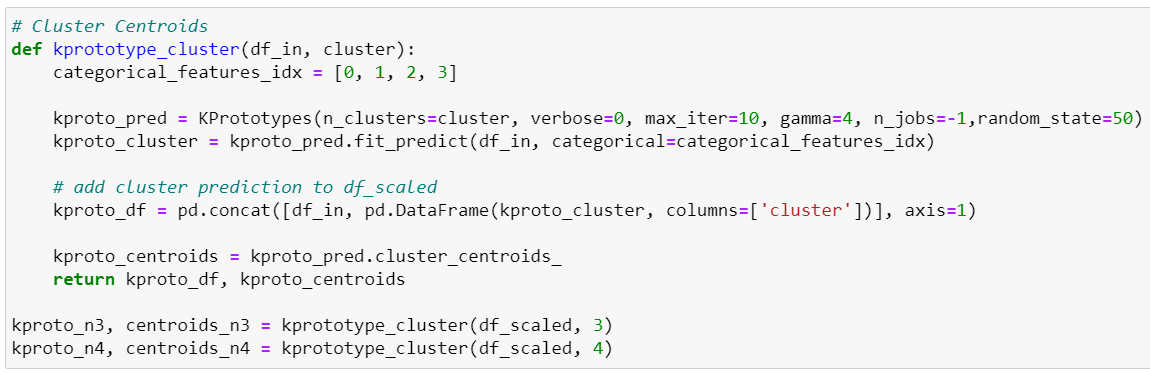
After importing the *kprototypes* package, elbow method is used to find out the optimal number of clusters by plotting the cost function of K-prototype model against the number of clusters. At the same time, various *gamma* values are tested until an elbow shape is formed, which indicates the quality of cluster vaguely. In our case, a gamma value of 4 is used.





An elbow is formed at 3 clusters. Heuristically, the data should be segregated into 3 clusters based on the elbow method shown above. Nonetheless, to better decide the optimal k for k-prototype clustering, n\_clusters of 3 and 4 are both used to cluster the dataset and visualised to cross validate whether 3 clusters are good enough to cluster the data.

1. **Predict clusters in dataset**

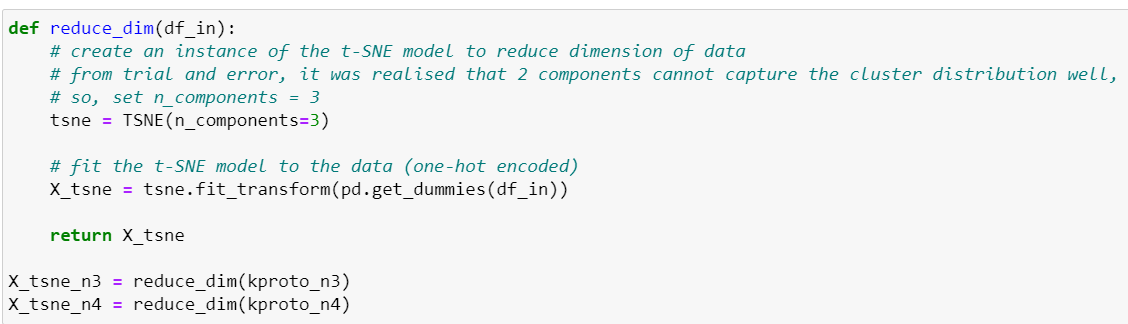


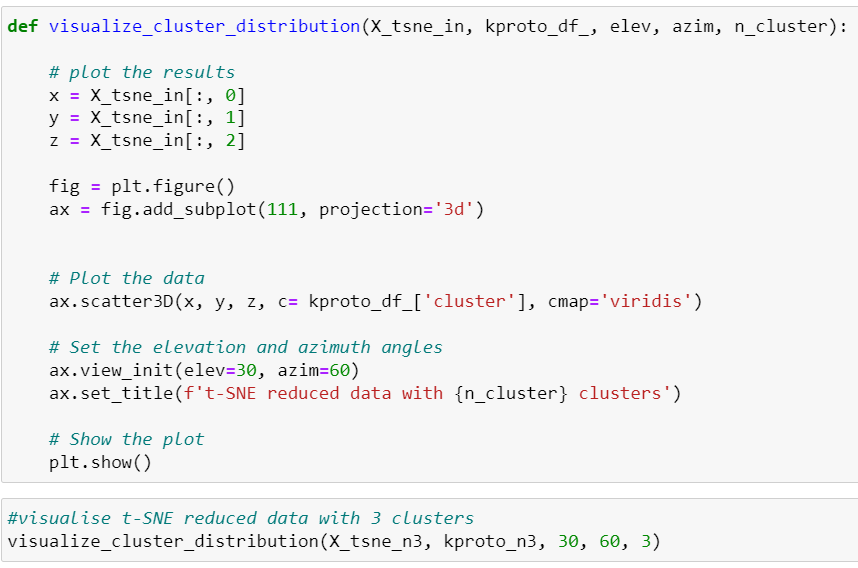
1. **Visualize the clustering output to verify the quality of cluster.**

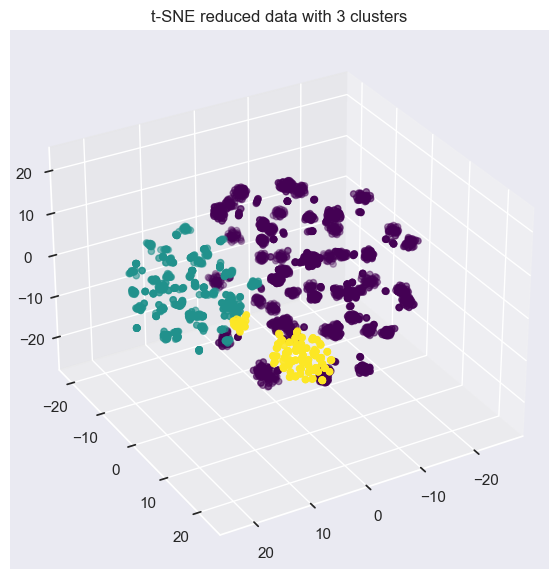
Visually assess the distribution of cluster by using:

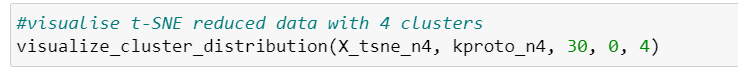
* dimensionality reduction. In this case, t-SNE is used as it can deal with categorical data
* visualise plot

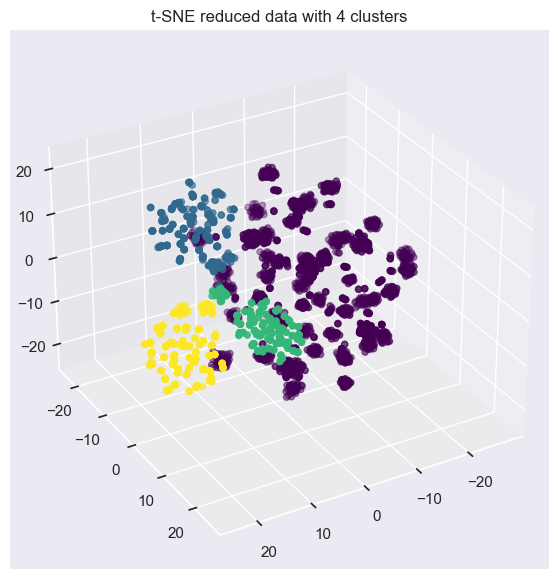




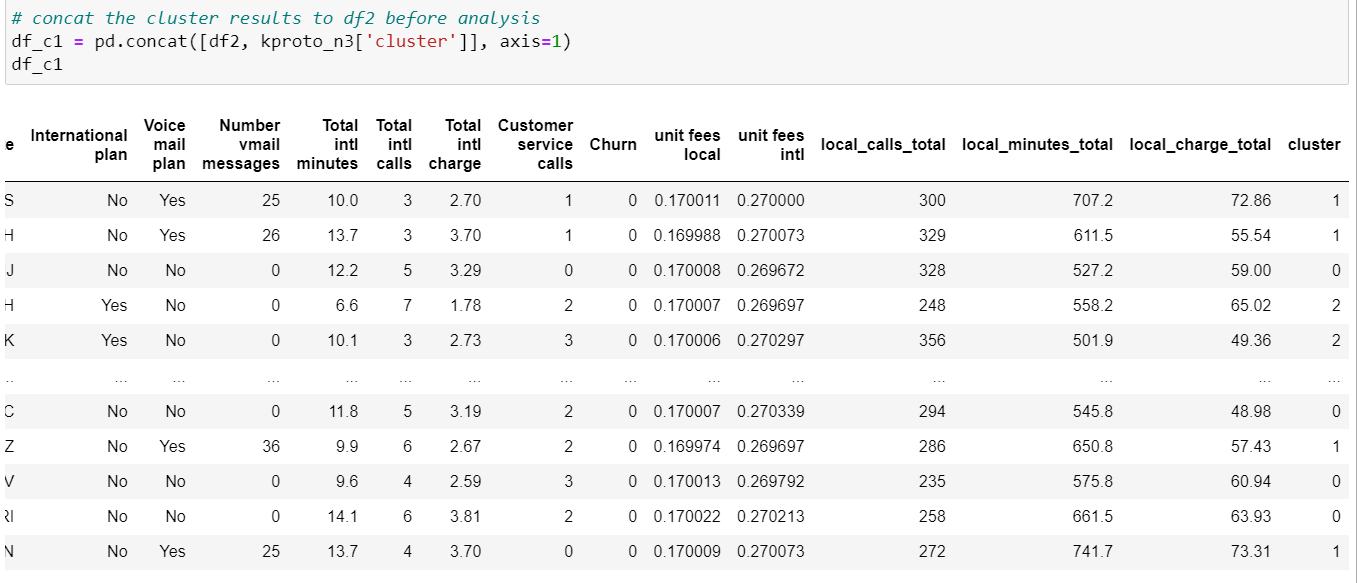








From the visualisation above, it appears that n\_cluster of 3 is able to cluster the data succinctly based on elbow method and visualisation.



**Method 2: Gower distance + DBSCAN**

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm. It groups together data points that are closely packed together (points with many nearby neighbors) and separates out points that are sparsely located (points with few or no nearby neighbors). The algorithm starts by defining a neighborhood around each point, and then groups together all points that are in dense neighborhoods. The parameters of the DBSCAN algorithm are the radius of the neighborhood (eps) and the minimum number of points required to form a dense neighborhood (minPts). Points that are not in any dense neighborhood are considered noise and are ignored.

**Gower Distance**

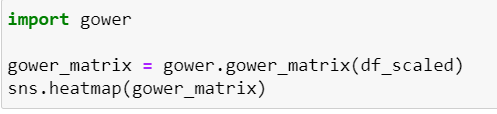
As the data contains a mixture of categorical and numerical features, DBSCAN is unable to group find the distance between data points accurately. To tackle this challenge, a composite distance measure- Gower distance is used. Gower distance is a measure to find similarity between two rows of dataset consisting of mixed data (categorical and numerical). It works by combining composite distance measures for continuous and discrete variables with Manhattan distance and Dice distance respectively. By applying Gower distance, we can compute the similarity or dissimilarity between any two data points on the same metric, of which the distances are returned in the form of a matrix with shape (n x n), where n is the number of samples in the dataset. This allows further clustering to occur in Euclidean space.

Gower distance lies in the range of 0 and 1. The closer the distance is to 1, the larger the difference between two data points. The closer the distance is to 0, the more similar the two data points are.

The learning algorithm is executed with steps below:

1. Compute Gower distance.
2. Apply DBSCAN on the output Gower matrix
3. **Compute Gower distance**

Gower distance and its matrix can be computed with *gower* library.





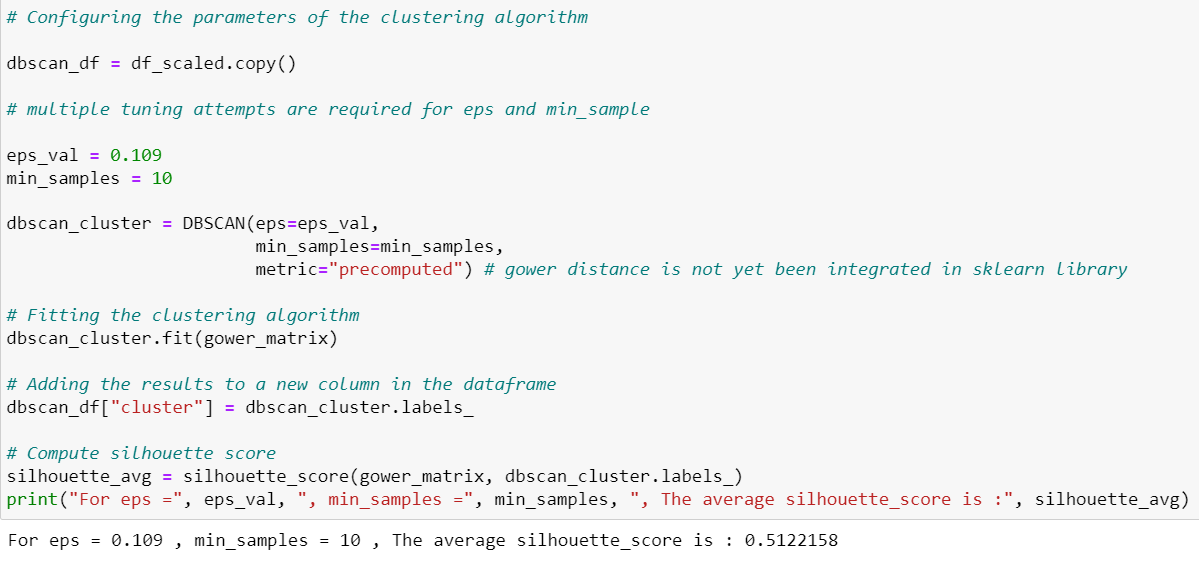
By visually inspecting the heatmap, we can observe that the dataset has a relatively small Gower distance from one another, meaning that they are similar to one another.

1. **Apply DBSCAN on the output Gower matrix**

The parameters for DBSCAN: eps and min\_samples are hand-tuned to optimize two objectives:

1. maximise the silhouette score

2. clusters are visualised to ensure that the clusters are interpretable.

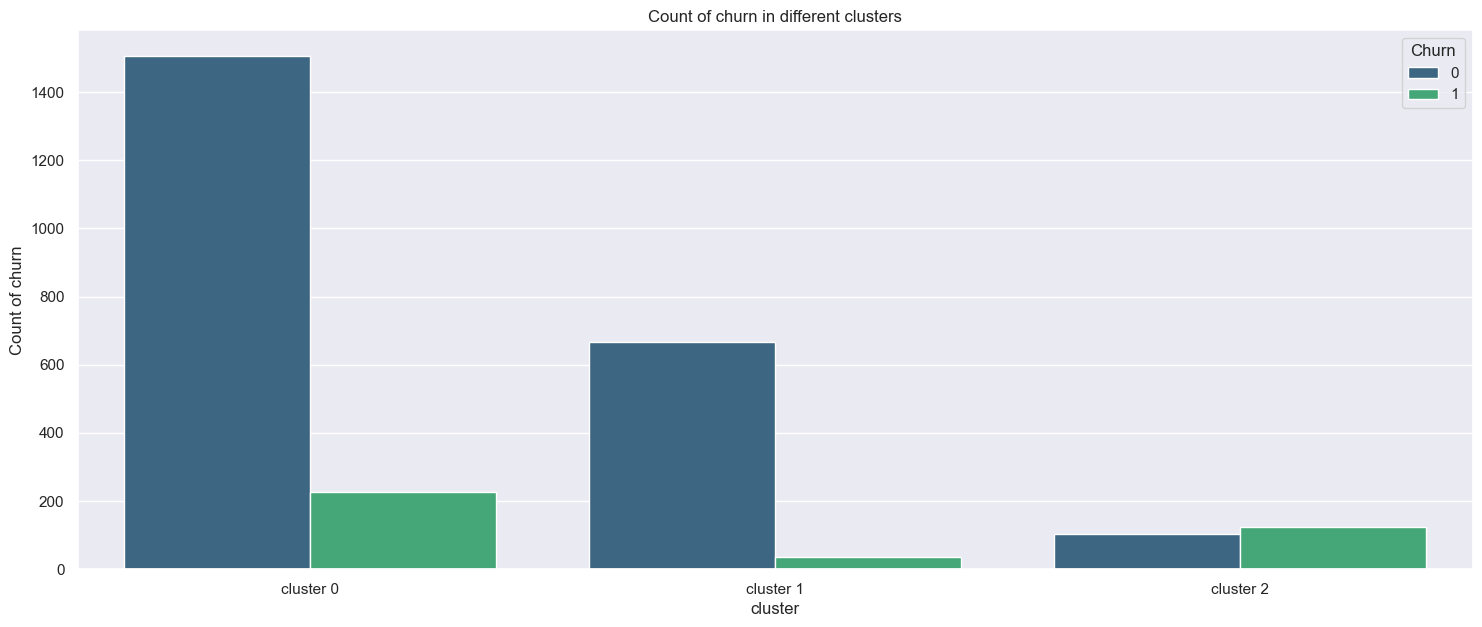


The fine-tuned epsilon and min\_samples are 0.109 and 10 respectively, resulting in a silhouette score of 0.512.

**SECTION 4: Result and Interpretation**

**Method 1: K-prototype**

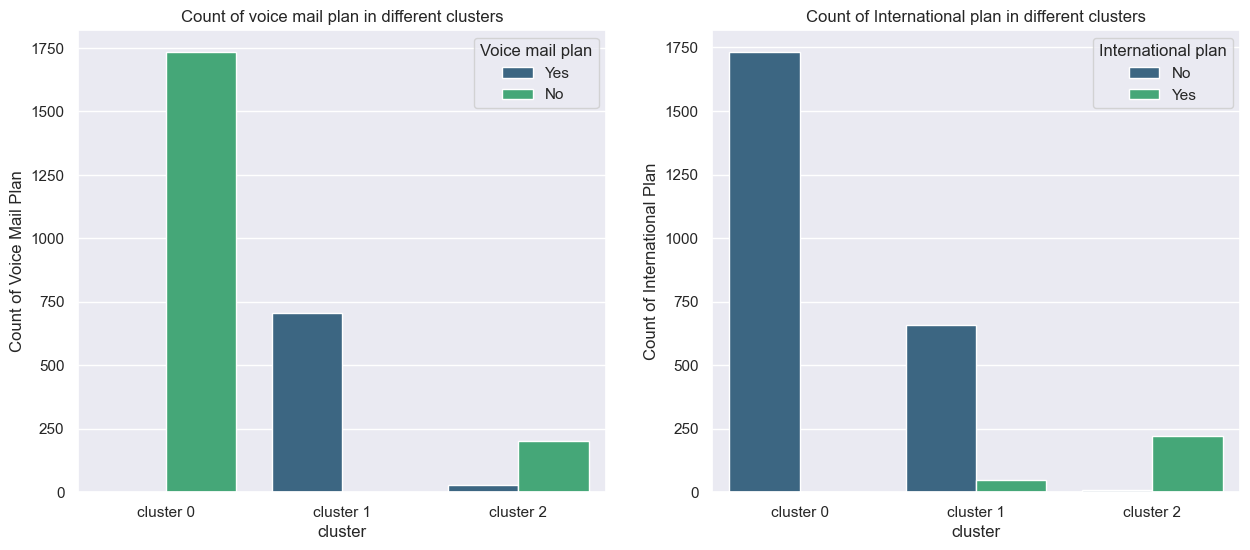
**Explore the clusters with respect to Churn**

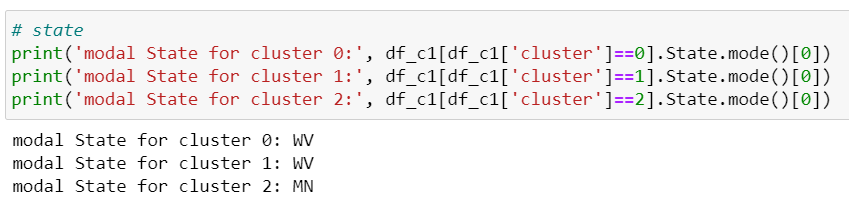


From the distribution of churn status in each cluster, we can notice cluster pattern below:

* **Cluster 0**:
  + more than half of the samples are in this cluster
  + A considerable amount, more than 10% of the customers churn their services
* **Cluster 1**:
  + the second largest cluster
  + A very small portion of the customers in this cluster churn their services.
* **Cluster 2**:
  + the smallest cluster
  + More than half of the customers in this cluster churn their services.

**Categorical Features Analysis**





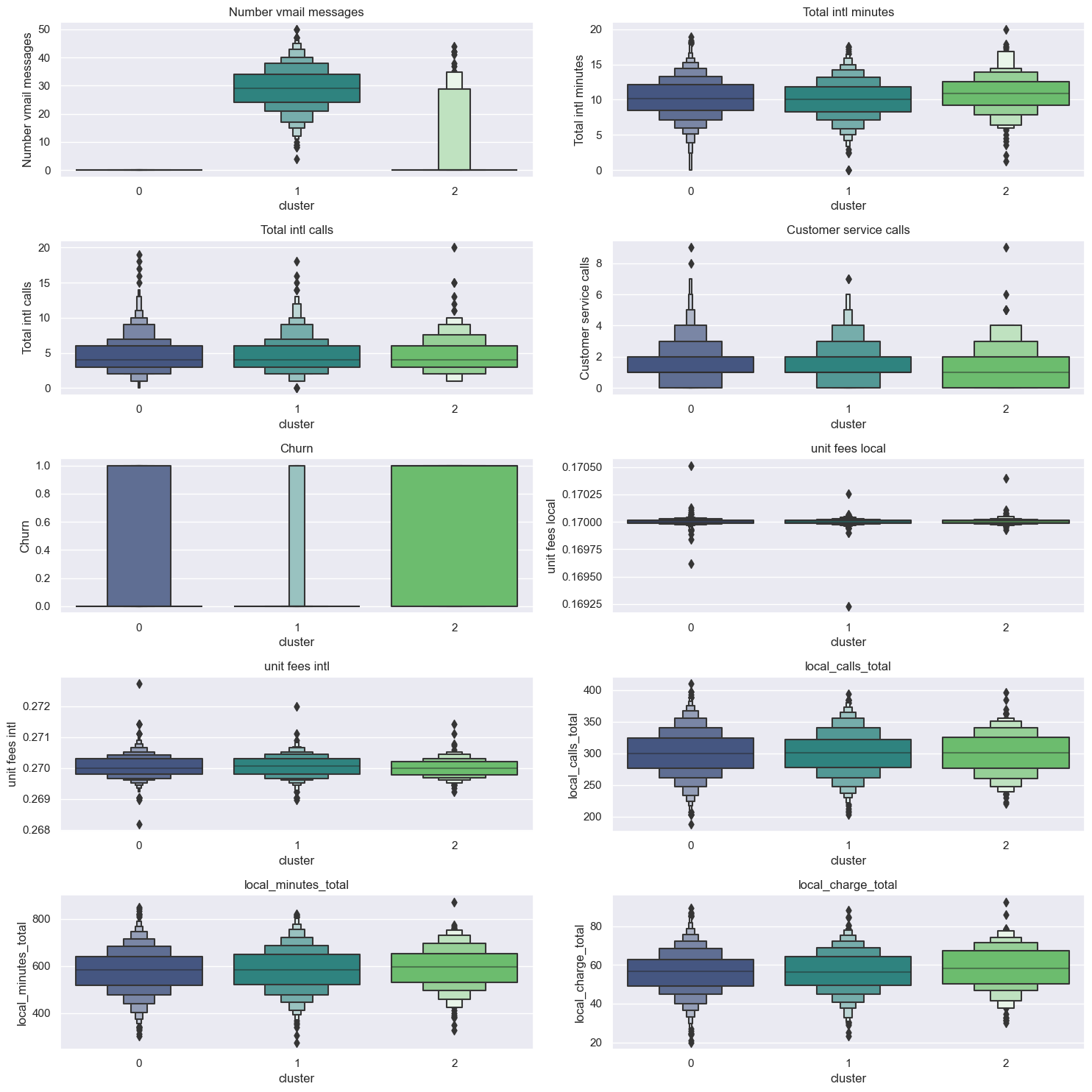
**Observations:**

* **Cluster 0:**
  + All of them do not have voice mail plan.
  + All of them do not have an international plan.
  + Most of the customers are in the state of 'WV'.
* **Cluster 1**:
  + All of them have voice mail plan.
  + A negligibly small portion of them have international plan.
  + Most of the customers are in the state of 'WV'.
* **Cluster 2**:
  + Almost all of them have no voice mail plan.
  + All of them have international plan.
  + Most of the customers are in the state of 'MN'.

**Discussion**

From the observations, we can see that the clustering is largely dependent on theVoice mail plan feature as all customers from cluster 0 and 2 have no voice mail plan but all of them in cluster 1 do. As for the International plan, we can notice that none of customers in cluster 1 subscribe to an international plan while all of the customers in cluster 2 hold an international plan. This observation is consistent with our common sense that people tend to cancel their international plan as cluster 2 has a significant higher churn rate than cluster 0. Besides, most of the customers in cluster 0 and 1 (clusters that seldom churn) are in the state of 'WV'. This is an interesting feature that can be further investigated in the future discussion.

**Numerical Features Analysis**



**Observations:**

For the numerical features, most features have very similar distributions in every cluster. To prevent cluttering, the observation in this section only focuses on the features with interesting difference between clusters.

**Cluster 0**:

* Have 0 “Number vmail message”
* Similar `Total intl minutes` and `Total intl calls` with cluster 1
* Similar number of `Customer service calls` compared with cluster 1
* Similar `local\_minutes\_total` and `local\_charge\_total` with cluster 1

**Cluster 1**:

* Have the most “Number vmail message”
* Similar `Total intl minutes` and `Total intl calls` with cluster 0
* Similar number of `Customer service calls` compared with cluster 0
* Similar `local\_minutes\_total` and `local\_charge\_total` with cluster 0

**Cluster 2**:

* Slightly higher median of `Total intl minutes` compared with the other two clusters but similar `Total intl calls` with the other two clusters.
* Have significant lower median for `Customer service calls` compared with the other two clusters.
* Significantly higher `local\_calls\_total` and `local\_charge\_total` compared with cluster 0 and 1 in terms of lower limit

#### **Conclusion**

**Cluster 0**:

* more than half of the samples are in this cluster
* A considerable amount, more than 10% of the customers chuen their services
* All of them do not have voice mail plan.
* All of them do not have an international plan.
* Most of the customers are in the state of 'WV'.
* Have 0 “Number vmail message”
* Similar `Total intl minutes` and `Total intl calls` with cluster 1
* Similar number of `Customer service calls` compared with cluster 1
* Similar `local\_minutes\_total` and `local\_charge\_total` with cluster 1

**Cluster 1**:

* the second largest cluster
* A very small portion of the customers in this cluster churn their services.
* All of them have voice mail plan.
* A negligibly small portion of them have international plan.
* Most of the customers are in the state of 'WV'.
* Have the most “Number vmail message”
* Similar `Total intl minutes` and `Total intl calls` with cluster 0
* Similar number of `Customer service calls` compared with cluster 0
* Similar `local\_minutes\_total` and `local\_charge\_total` with cluster 0

**Cluster 2**:

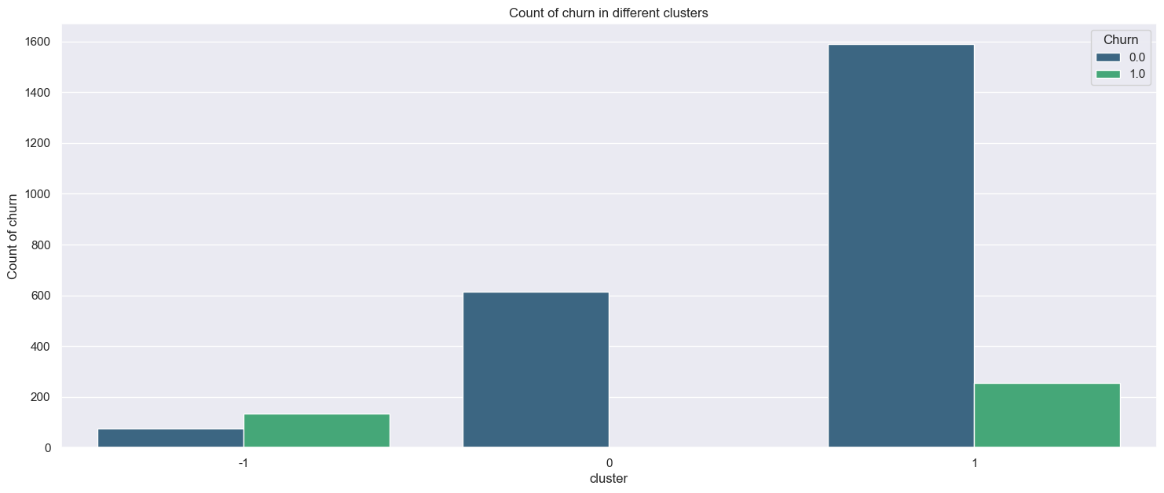
* the smallest cluster
* More than half of the customers in this cluster churn their services.
* Almost all of them have no voice mail plan.
* All of them have international plan.
* Most of the customers are in the state of 'MN'.
* Slightly higher median of `Total intl minutes` compared with the other two clusters but similar `Total intl calls` with the other two clusters.
* Have significant lower median for `Customer service calls` compared with the other two clusters.
* Significantly higher `local\_calls\_total` and `local\_charge\_total` compared with cluster 0 and 1 in terms of lower limit

**Discussion**

Compared with customers from cluster 1 who hardly churn at all, customers from cluster 0 are noticeably more likely to churn their services. While we notice that characteristics of customers in cluster 0 is almost the same with cluster 1, there is one single difference between the two clusters that might result in the difference: All of the customers in cluster 1 have a voice mail plan and use them frequently while all the customers in cluster 0 do not.

Meanwhile, we observe that customers in cluster 2 are the most likely to churn, and they show a great deal of differences in characteristics compared with the other two clusters. The most significant differences are: (1) they are more likely to call often locally and spend more time on international calls compared with the other two clusters. (2) They deal with fewer customer service calls compared with the other two. These differences might result in almost all customers in cluster 2 churning their services, with reasons to be further investigated.

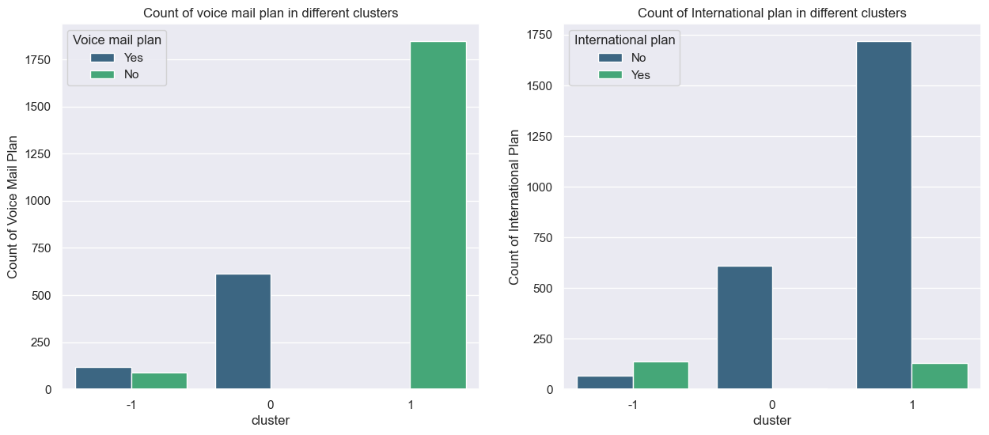
**Method 2: DBSCAN**

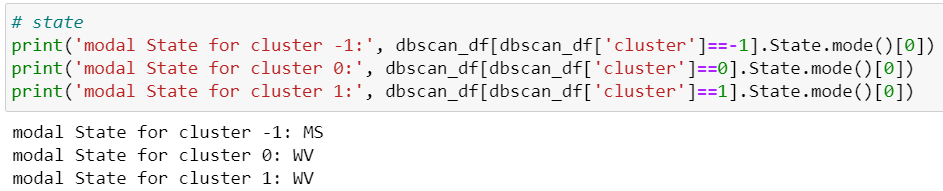


With DBSCAN, 3 clusters are generated from the dataset, I.e., -1, 0 and 1, with –1 regarded as the group consisting of ‘noise’ according to the model. However, we believe that the ‘noise’ can be informative too as there are many edge cases in this dataset as discussed in Section 1. From the distribution of churn status in each cluster, we can notice cluster pattern below:

* **Cluster -1**:
  + the smallest cluster (133 samples)
  + More than half of the customers in this cluster churn their services.
* **Cluster 0**:
  + the second largest cluster (613 samples)
  + All the customers in this cluster do not churn.
* **Cluster 1**:
  + more than half of the samples are in this cluster (1590 samples)
  + Most of the customers do not churn.

**Categorical Features Analysis**





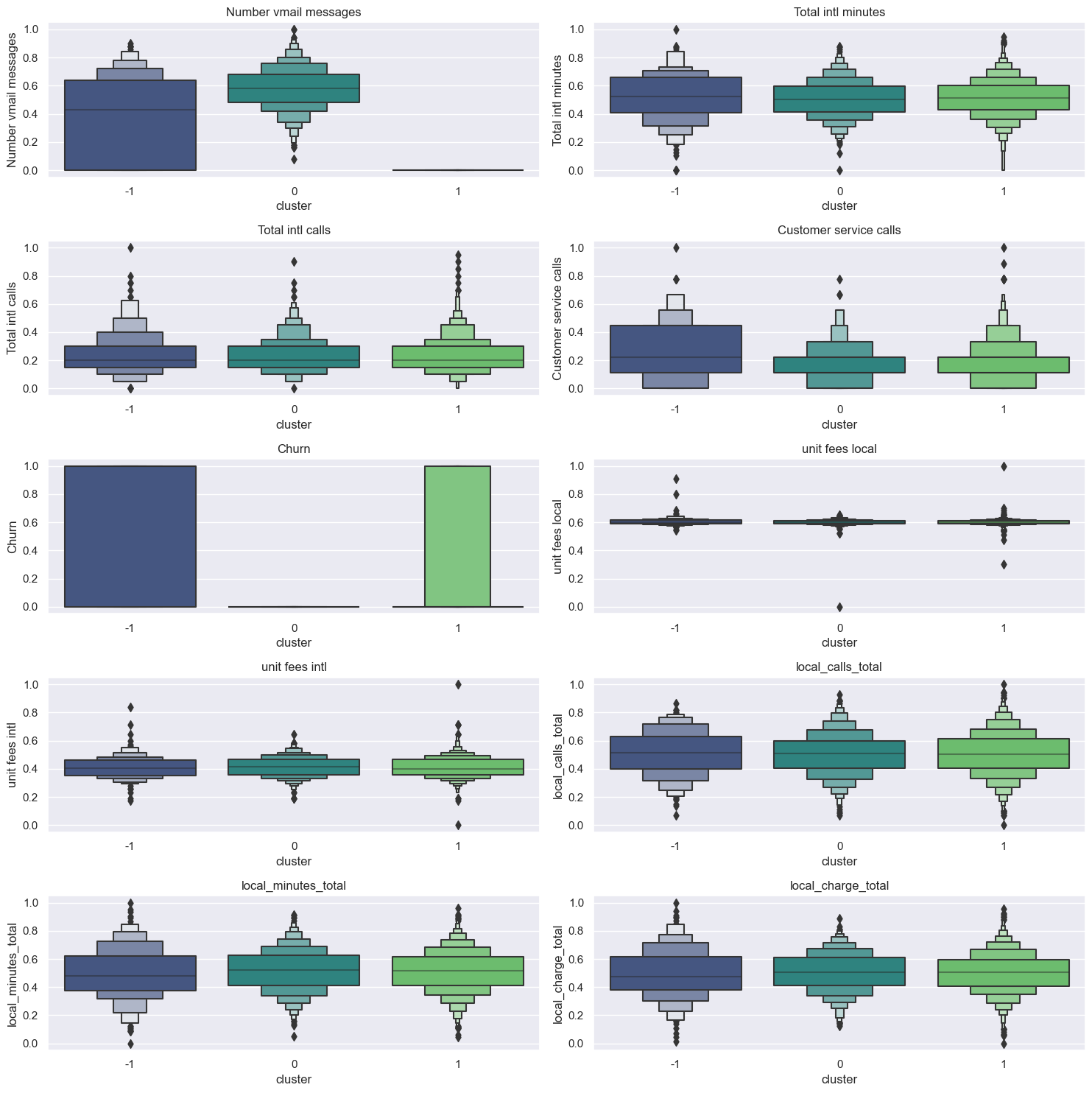
**Observations:**

* **Cluster -1**:
  + More than half of them have voice mail plan.
  + More than 50% of them have international plan.
  + Most of the customers are in the state of 'MS'.
* **Cluster 0**:
  + All of them have voice mail plan.
  + Only 3 of them have international plan.
  + Most of the customers are in the state of 'WV'.
* **Cluster 1**:
  + All of them do not have voice mail plan.
  + A small portion of them have international plan.
  + Most of the customers are in the state of 'WV'.

**Discussion**

From the observations, we can see that the clustering is largely dependent on theVoice mail plan feature as most customers from cluster -1 and all customers from cluster 0 have voice mail plan but none of them in cluster 1 have it. As for the International plan, we can notice that only a small portion of customers in cluster 0 and cluster 1 subscribe to an international plan. Meanwhile, more than half of the customers in cluster -1 hold an international plan. This observation is consistent with our common sense that people tend to cancel their international plan. Besides, most of the customers in cluster 0 and 1 (clusters that seldom churn) are in the state of 'WV'. This is an interesting feature that can be further investigated in the future discussion.

**Numerical Features Analysis**



**Observations:** For the numerical features, most features have very similar distributions in every cluster. To prevent cluttering, the observation in this section only focuses on the features with interesting difference between clusters.

* **Cluster -1**:
  + A wide distribution of the number of vmail message
  + slightly higher median of Total intl minutes compared with the other two clusters.
  + have a lot more Customer service calls compared with the other two clusters.
* **Cluster 0**:
  + Have the most number of vmail messages
  + similar Total intl minutes and Total intl calls with cluster 1
  + similar number of Customer service calls compared with cluster 1
  + similar local\_minutes\_total with cluster 1
  + slightly lower median of local\_charge\_total compared with cluster -1 and 1
* **Cluster 1**:
  + similar Total intl minutes and Total intl calls with cluster 0
  + similar number of Customer service calls compared with cluster 0
  + similar local\_minutes\_total with cluster 0
  + similar local\_charge\_total compared with cluster –1

#### **Conclusion**

* **Cluster -1**:
  + more than half of the samples have international plan.
  + Most of the customers churn their services.
  + the smallest cluster (133 samples).
  + **All of them have voice mail plan.**
  + Most of the customers are in the state of 'MS'.
  + A wide distribution of the number of vmail message
  + slightly higher median of Total intl minutes compared with the other two clusters. Ironically, they have significantly fewer Total intl calls than the other two clusters.
  + have a lot more Customer service calls compared with the other two clusters.
* **Cluster 0**:
  + the second largest cluster (613 samples)
  + All the customers in this cluster do not churn their services.
  + Only 3 of them have international plan.
  + **All of them have voice mail plan and** they havethe greatest number of vmail messages
  + Most of the customers are in the state of 'WV'.
* **Cluster 1**:
  + the largest cluster (1590 samples)
  + Most customers in this cluster do not churn their services.
  + A small portion of them have international plan.
  + **All of them have voice mail plan.**
  + Most of the customers are in the state of 'WV'.

**Discussion**

Compared with customers from cluster 0 who do not churn at all, customers from cluster 1 are noticeably more likely to churn their services. While we notice that characteristics of customers in cluster 0 is almost the same with cluster 1, there is one single difference between the two clusters that might result in the difference: All the customers in cluster 0 have a voice mail plan while all the customers in cluster 1 do not.

Meanwhile, we observe that most people who churn are in cluster -1, and customers from cluster -1 shows a great deal of differences in characteristics compared with the other two clusters. The most significant differences are: (1) they spend a lot more time on calls compared with the other two clusters, locally and internationally. (2) They deal with much more customer service calls compared with the other two. These differences might result in almost all customers in cluster -1 churning their services, with reasons to be further investigated.

**SECTION 5: Evaluation and Comparison between Models**

Due to the restrictions in the brief requirement to use both partition-based clustering and density-based clustering, on top of the dataset being a mixed dataset consisting of both categorical and numerical features, the two clustering models used do not exist in the same metric scale. Besides, there is also no external measure acting as an objective measure of accuracy. This makes both methods used impossible to be evaluated and compared objectively with a single metric.

Therefore, when comparing the two methods used, it is important to think about: “What is the clustering results used for?” combined with domain knowledge to generate useful insights. In our case, the clustering models are used to construct specific customer profiles based on customer segments with respect to whether they will churn their service. This allows us to fine tune our strategies in retaining the customers.

In this project where the number of clusters is consistent, both methods have showed very similar clustering patterns. For instance, the cluster with vmail plan have a very low likelihood to churn, while cluster with international plan is more likely to churn. The number of customer service calls plays a significant role in clustering in both models, where we notice that either people with a lot more or very few customer service calls are likely to be in the cluster that has higher probability to churn.

These insights and observations show that both models show a very similar pattern of clustering in this dataset with only slight variations. Both models are useful for us to achieve our purpose. As clustering analysis is highly subjective, it is important that we cluster our data on various clustering models and compare them with one another to inform decisions from multiple perspectives.