REGION
JAWA TIMUR

DATA SCIENCE CLUB (DSC)
UNIVERSITAS PGRI ADI BUANA SURABAYA

WANDA NUR HAMIDAH

1. **How is the characteristics of the data? Do exploratory data analysis, calculate descriptive statistics, and visualize it.**

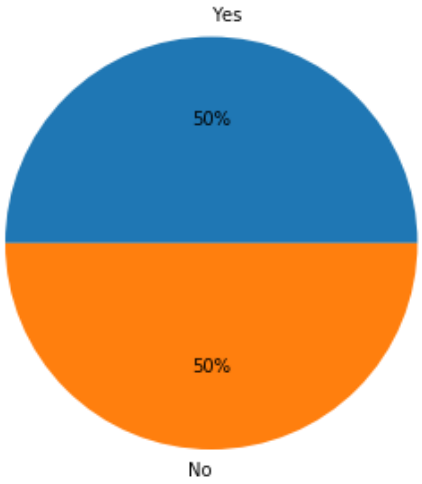
```
Data columns (total 19 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Churn                        1896 non-null   int64
 1   Tenure                       1739 non-null   float64
 2   PreferredLoginDevice         1896 non-null   object
 3   CityTier                     1896 non-null   int64
 4   WarehouseToHome              1744 non-null   float64
 5   PreferredPaymentMode         1896 non-null   object
 6   Gender                       1896 non-null   object
 7   HourSpendOnApp               1766 non-null   float64
 8   DeviceRegistered             1896 non-null   int64
 9   PreferedOrderCat             1896 non-null   object
 10  SatisfactionScore            1896 non-null   int64
 11  MaritalStatus                1896 non-null   object
 12  NumberOfAddress              1896 non-null   int64
 13  Complain                     1896 non-null   int64
 14  OrderIncreaseFromLastYear    1855 non-null   float64
 15  CouponUsed                   1817 non-null   float64
 16  OrderCount                   1837 non-null   float64
 17  DaySinceLastOrder            1812 non-null   float64
 18  CashbackAmount               1896 non-null   float64
dtypes: float64(8), int64(6), object(5)
memory usage: 296.2+ KB
```

The characteristics of the eccomerce churn data are integer, float, and object types.

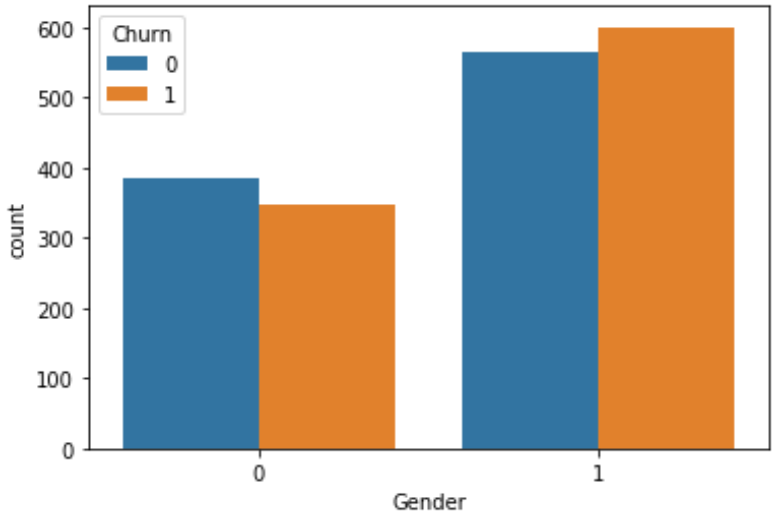
	Churn	Tenure	CityTier	WarehouseToHome	HourSpendOnApp	DeviceRegistered	SatisfactionScore	NumberOfAddress	Complain	Ord
count	1896.000000	1739.000000	1896.000000	1744.000000	1766.000000	1896.000000	1896.000000	1896.000000	1896.000000	
mean	0.500000	7.347901	1.719409	15.922018	2.682899	3.543776	3.275316	4.088080	0.385549	
std	0.500132	8.149302	0.936148	8.498368	0.679286	1.015023	1.269551	2.694888	0.486853	
min	0.000000	0.000000	1.000000	5.000000	0.000000	1.000000	1.000000	1.000000	0.000000	
25%	0.000000	1.000000	1.000000	9.000000	2.000000	3.000000	2.000000	2.000000	0.000000	
50%	0.500000	4.000000	1.000000	14.000000	3.000000	3.000000	3.000000	3.000000	0.000000	
75%	1.000000	13.000000	3.000000	22.000000	3.000000	4.000000	4.000000	6.000000	1.000000	
max	1.000000	50.000000	3.000000	36.000000	4.000000	6.000000	5.000000	21.000000	1.000000	

From the table above, it can be seen descriptive statistics by knowing count, mean, standard deviation, minimum and maximum values and we can see that the longest tenure is 50 months and the maximum cashback amount is $323.59. The minimum cashback amount is about $0. The

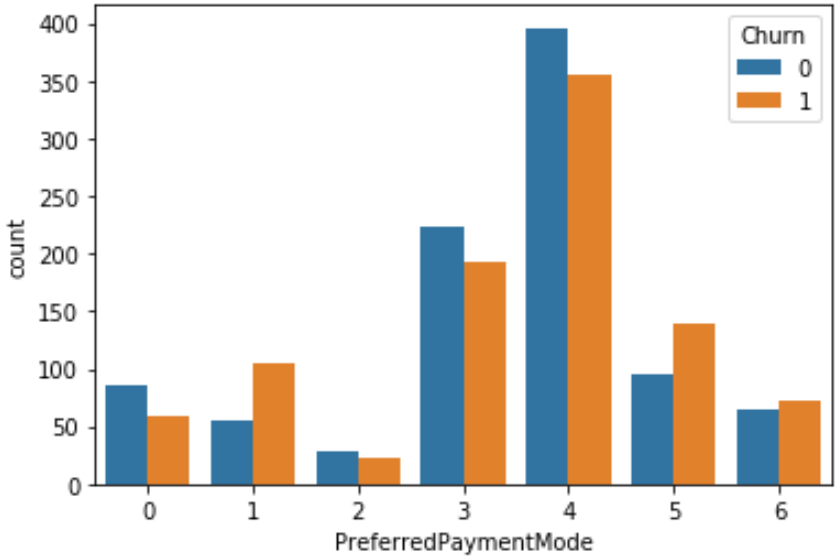
customer can expect to have a cashback amount of about $164.91. I am assuming the charges are in United States Dollars (USD).



Based on the pie chart above, it can be concluded that the distribution of customer data is balanced between those who do not churn and churn, with churn details as much as 50% and no churn as much as 50%.



From the plot above, it looks like gender does not play a role in customer churn. Let's visualize the churn count for Preferred Payment Mode



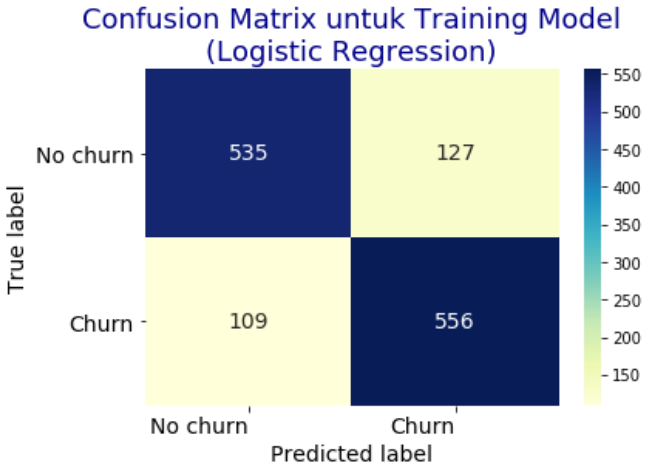
The chart above is interesting, as it helps me to differentiate between retained and churned customers, it shows that most of the customers make the preferred payment method is CC while the less used one is COD method.

2. **Please Do preprocessing data. Is there any missing values or outliers? If yes, solve it and give some explanation. Do variable selection or dimension reduction if needed and give some explanation.**

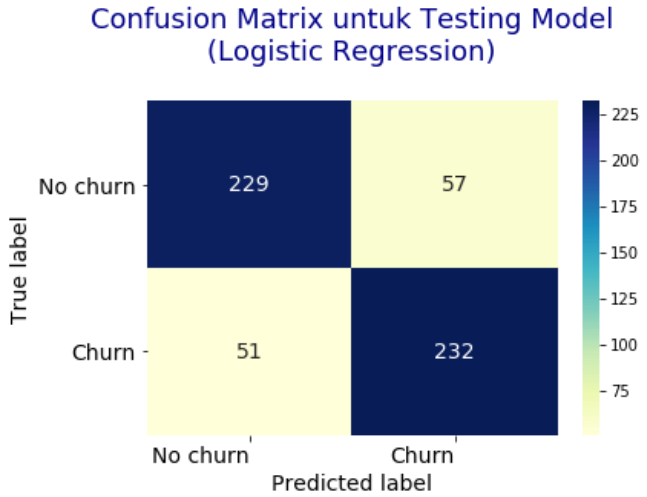
   In the data there are missing values and outliers. The way to handle missing values by means of missing values will be filled with the average of the column, while handling outliers is by normalizing the data.

3. **Find the best model and evaluate the model.**

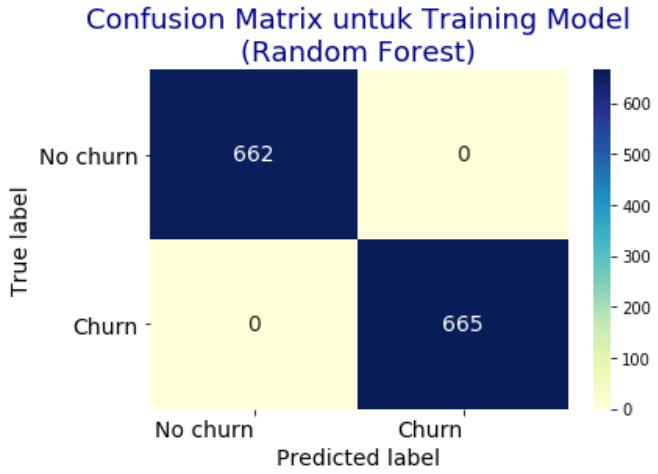
   The dataset used for modeling is divided into 2 types, train data and test data. In this case, I divided 70% data for train data and 30% for test data. Where the variable x is the predictor and the variable y is the target. In this modeling, I used logistic regression and random forest.



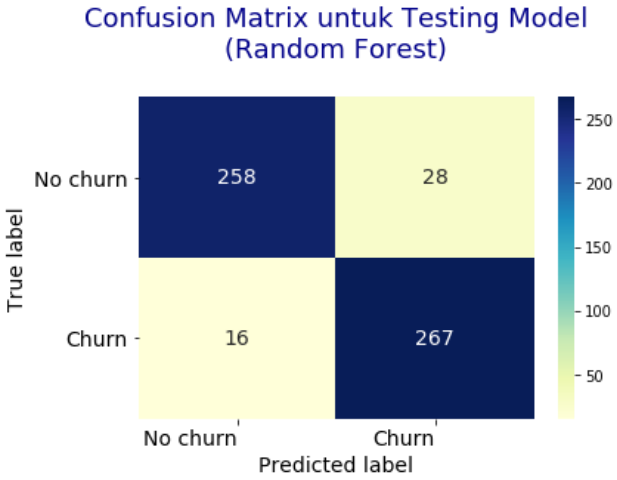
From the training data, it can be seen that the model is able to predict data with an accuracy rate of 82%, with details of the actual churn prediction that churn is 556, the actual churn prediction not churn is 535, the actual churn no churn prediction is 109 and the actual churn prediction no churn is 127.



From the testing data, it can be seen that the model is able to predict data with an accuracy rate of 81%, with details of the actual churn prediction that churn is 232, the actual churn prediction not Churn is 229, the actual churn no churn prediction is 51 and the actual churn prediction no churn is 57.



From the training data, it can be seen that the model is able to predict data with a perfect level of 100% accuracy, with details of the actual churn prediction that churn is 665, the actual churn prediction is not churn is 662, the actual churn prediction is 0 churn and the actual churn prediction no churn is 0.



From the testing data, it can be seen that the model is able to predict data with an accuracy rate of 92%, with details of the actual churn prediction that churn is 267, the actual churn prediction not churn is 258, the actual churn no churn prediction is 16 and the actual churn prediction not churn is 28.

Based on the modeling that has been done using Logistic Regression and Random Forest, it can be concluded that to predict e-commerce churn by using this dataset, the best model is using the Random Forest algorithm. This is because the performance of the Random Forest model tends to be able to predict equally well in the training and testing phases (100% training accuracy, 92% testing accuracy), on the other hand the other algorithms tend to over-fit their performance. However, this does not lead us to conclude that if we use Random Forest for any modeling, we still have to do a lot of modeling experiments to determine which one is the best.