Project

Introduction

With the development of our modern society, we noticed that there is an increasing number of people exploring the world relying on airplanes, which effectively shortens the time and distance for journeys. However, we still hear a lot of negative feedback about the ride experience such as complaints about flight delays, terrible food, or in-flight services. In our project, we are going to explore the factors that influence passenger satisfaction in both positive and negative ways.

To achieve this goal, we will use the Airline Passenger Satisfaction dataset to analyze the correlation between passenger satisfaction and possible factors. We will sort each factor into different ranking levels, then look at the impact of ease of booking, cabin class, and flight distance which are decided before boarding. We will see whether the customer type and the cabin class can raise satisfaction effectively or their impact can be countervailed by short flight distance. Furthermore, we will examine the influence that the existence of delay, comfort, and cleanliness of the environment, quality of food, and inflight entertainment have on the satisfaction of passengers during the flight. The conclusions reached at the end of the project can be used to predict the choice of the airline companies that customers will make during planning the journey and help airline companies improve the service and passengers' experience.

Background

Our dataset, can be downloaded from https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction, it contains 129880 observations and 24 variables. The data set contains two customer type: disloyal customer and loyal customer, it has two travel type: business travel and personal travel. All the flights has three classes: business economy and economy plus. This dataset contains an airline passenger satisfaction survey with following variables:

- Satisfaction:Airline satisfaction level(Satisfaction, neutral or dissatisfaction)
- Age:The actual age of the passengers
- Gender:Gender of the passengers (Female, Male)
- Type of Travel:Purpose of the flight of the passengers (Personal Travel, Business Travel)
- Class:Travel class in the plane of the passengers (Business, Eco, Eco Plus)
- Customer Type:The customer type (Loyal customer, disloyal customer)
- Flight distance:The flight distance of this journey
- Inflight wifi service:Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
- Ease of Online booking:Satisfaction level of online booking
- Inflight service:Satisfaction level of inflight service
- Online boarding:Satisfaction level of online boarding
- Inflight entertainment:Satisfaction level of inflight entertainment
- Food and drink:Satisfaction level of Food and drink
- Seat comfort:Satisfaction level of Seat comfort
- On-board service:Satisfaction level of On-board service

- Leg room service:Satisfaction level of Leg room service
- Departure/Arrival time convenient:Satisfaction level of Departure/Arrival time convenient
- Baggage handling:Satisfaction level of baggage handling
- Gate location:Satisfaction level of Gate location
- Cleanliness:Satisfaction level of Cleanliness
- Check-in service:Satisfaction level of Check-in service
- Departure Delay in Minutes:Minutes delayed when departure
- Arrival Delay in Minutes:Minutes delayed when Arrival
- Flight cancelled:Whether the Flight cancelled or not (Yes, No)
- Flight time in minutes:Minutes of Flight takes

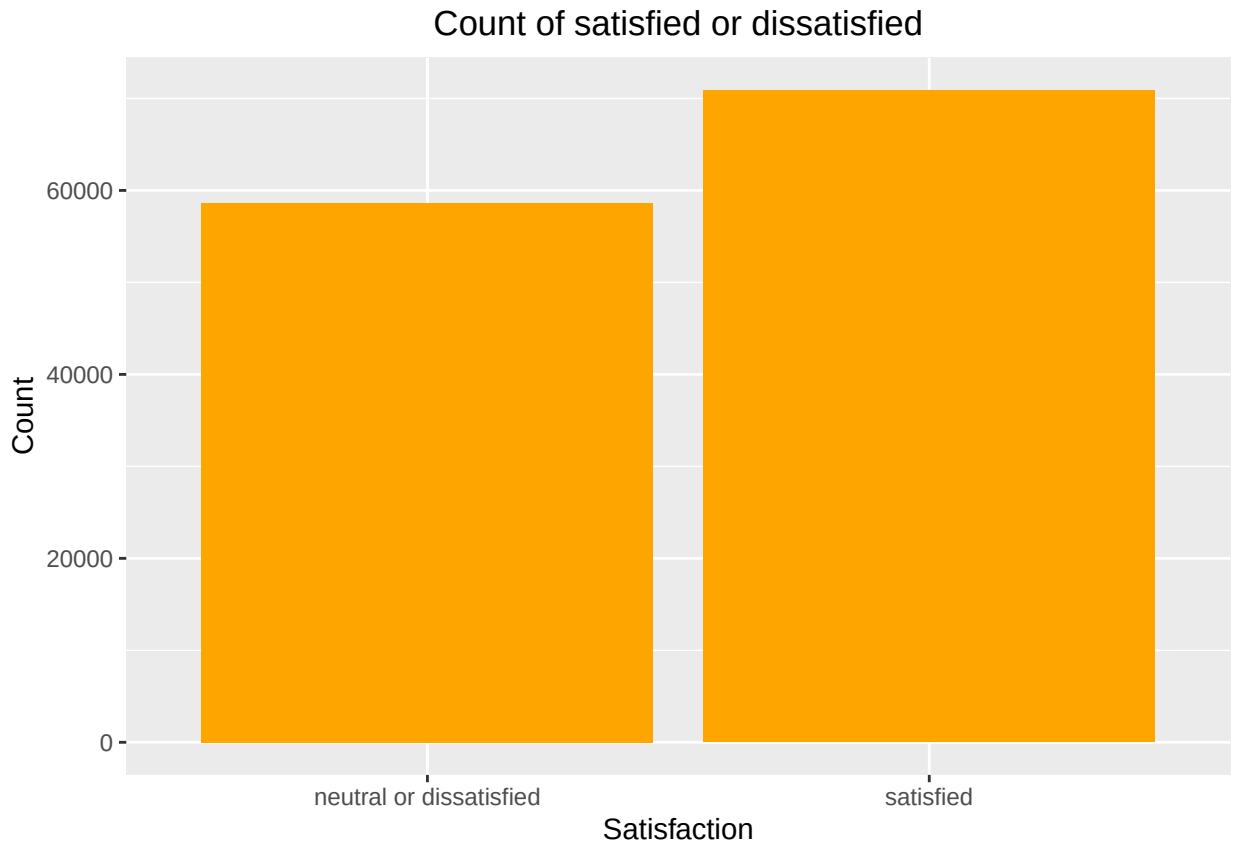
Data Exploration

```
## [1] 129880     24
```

```
## [1] 129487     24
```

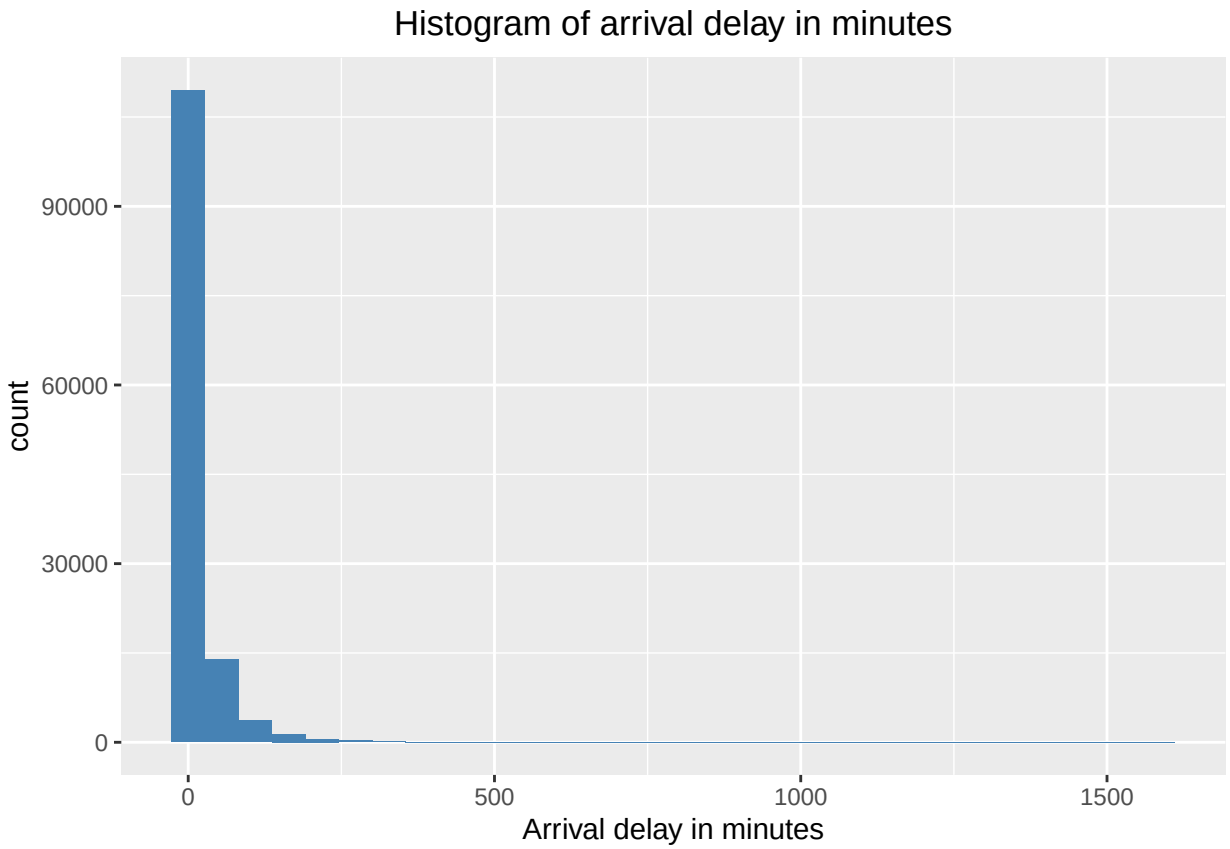
Find the number of customers who are satisfied with the flights and who are netural or dissatisfied with the flights, the draw a bar plot, we can see from the plot, the number of customers who are satisfied with the flights is larger than the number of customers who are dissatisfied with the flights.

```
ggplot(data=sat, aes(x=satisfaction_v2)) + geom_bar(fill="orange") +
  labs(x="Satisfaction", y="Count", title="Count of satisfied or dissatisfied") +
  theme(plot.title = element_text(hjust = 0.5))
```



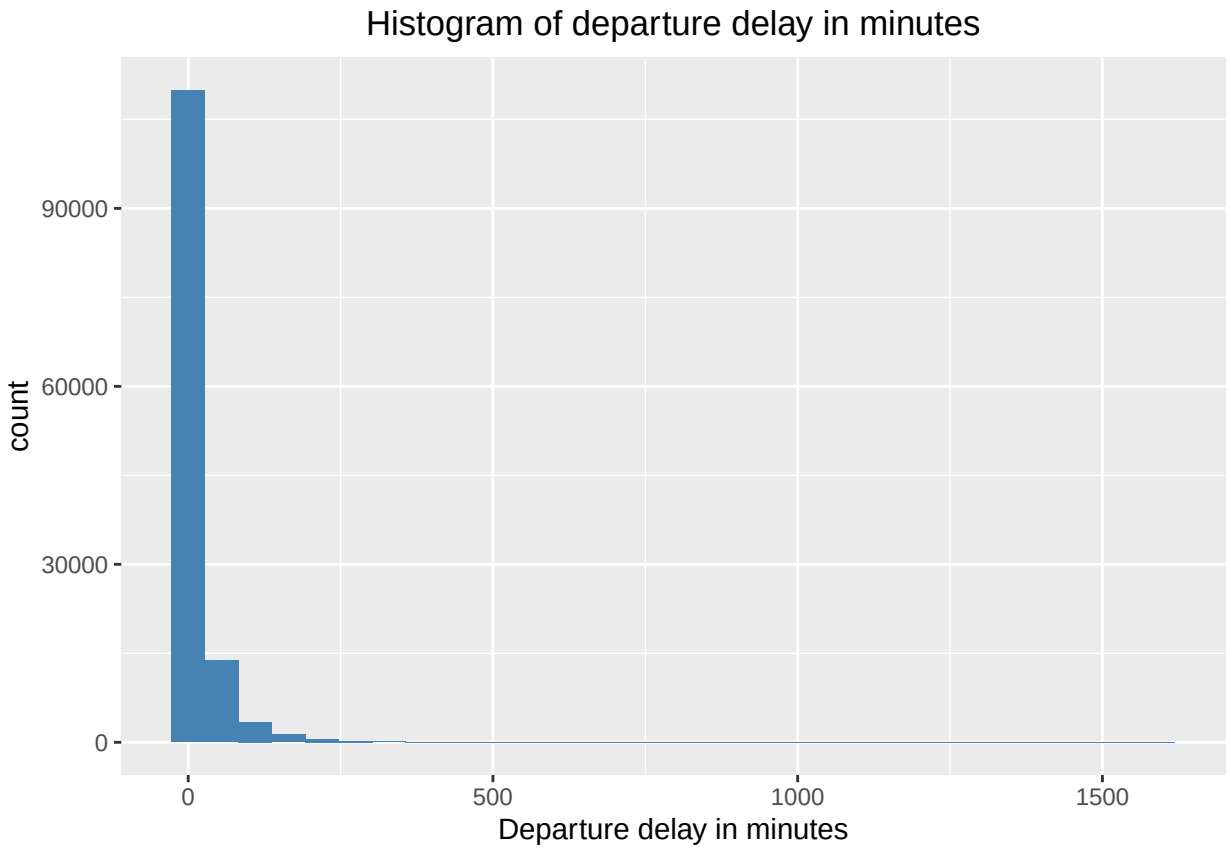
The distribution of arrival delay in minutes is right skewed and uni-model, most of the arrival delay in minutes are below 200 minutes.

```
ggplot(data=sat, aes(x=arrival_delay_in_minutes)) + geom_histogram(bins=30, fill="steelblue") +
  labs(x="Arrival delay in minutes", title="Histogram of arrival delay in minutes") +
  theme(plot.title = element_text(hjust = 0.5))
```



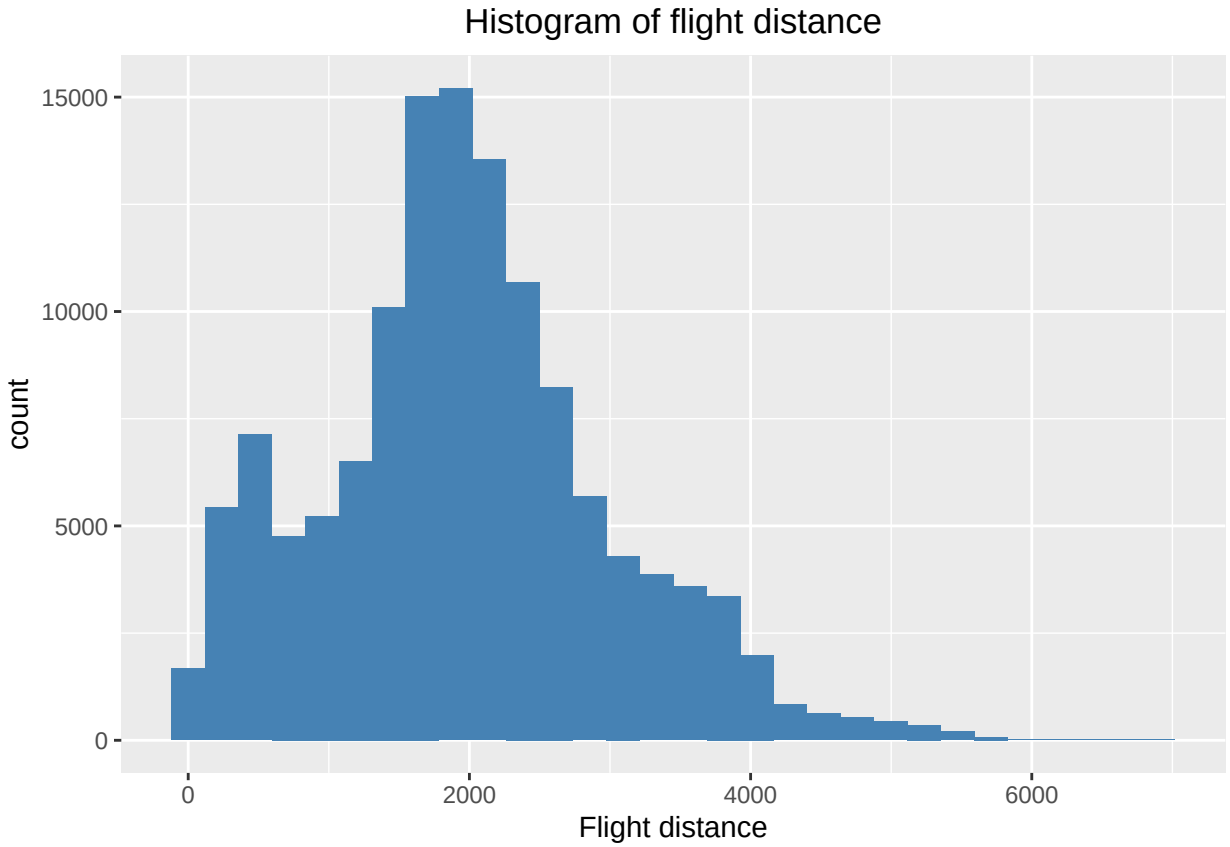
The distribution of departure delay in minutes is right skewed and uni-model, most of the departure delay in minutes are below 200 minutes.

```
ggplot(data=sat, aes(x=departure_delay_in_minutes)) + geom_histogram(bins=30, fill="steelblue") +
  labs(x="Departure delay in minutes", title="Histogram of departure delay in minutes") +
  theme(plot.title = element_text(hjust = 0.5))
```



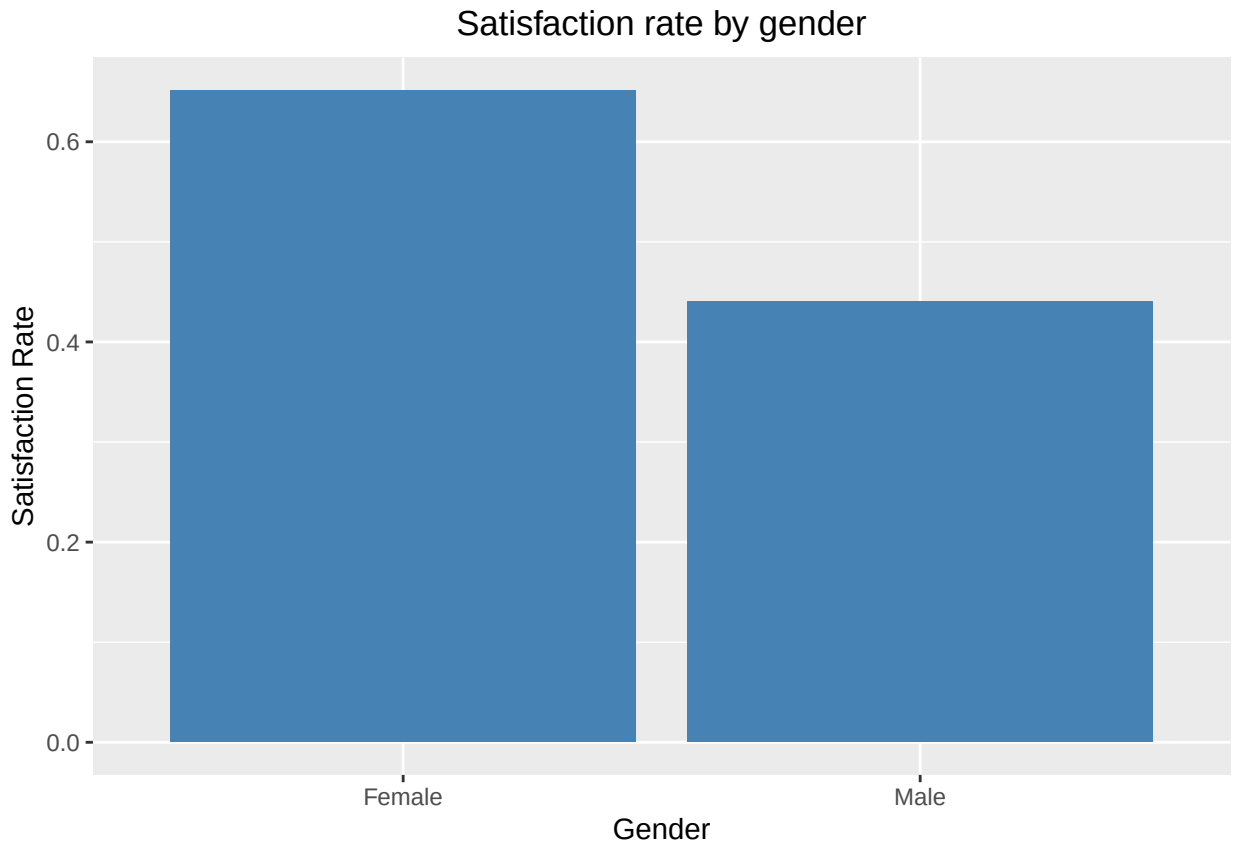
The distribution of flight distance is bi-model and right skewed, most of the flight distances are distributed around 2000 km.

```
ggplot(data=sat, aes(x=flight_distance)) + geom_histogram(bins=30, fill="steelblue") +
  labs(x="Flight distance", title="Histogram of flight distance") +
  theme(plot.title = element_text(hjust = 0.5))
```



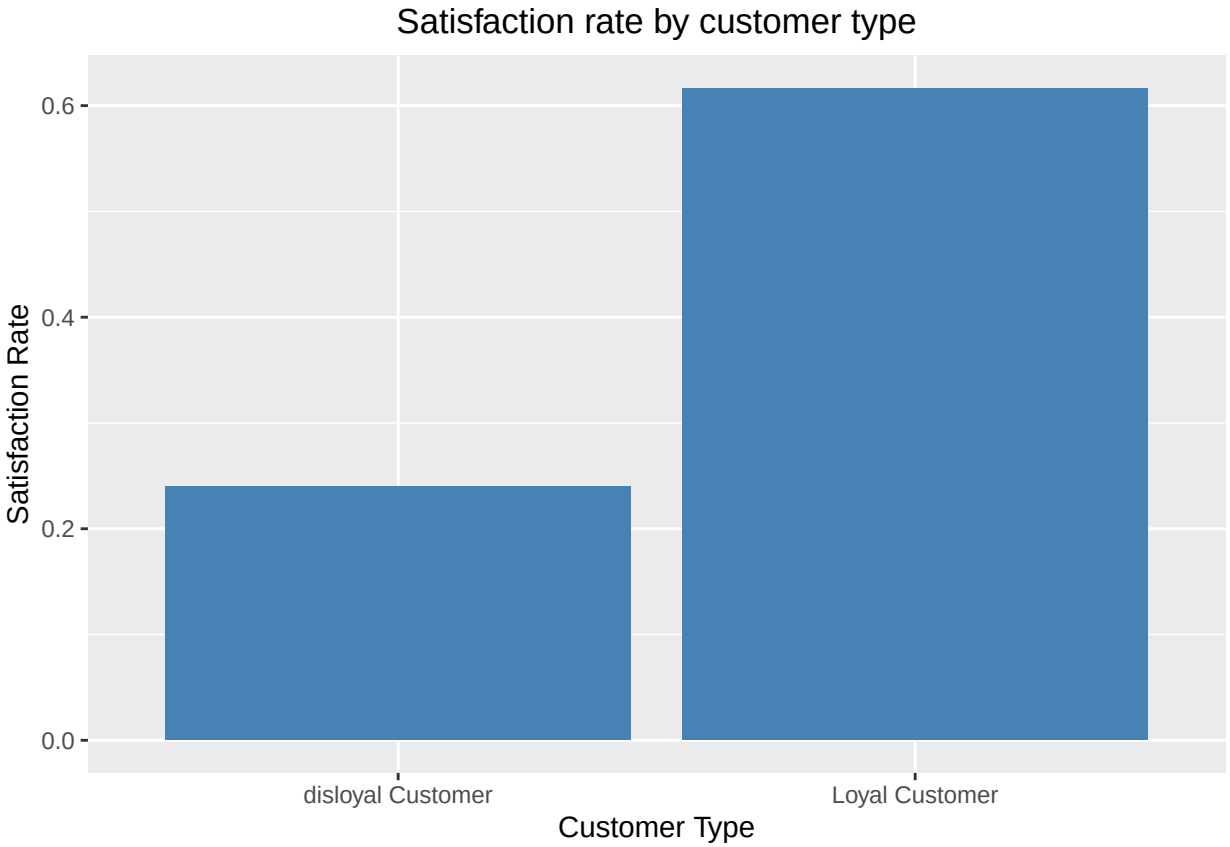
Calculate the proporation of satisfaction rate by gender, the proporation of female customers who are satisfied with the flight is higher than the proporation of male customers who are satisfied with the flight. The proporation of female customers who are satisfied with the flight is about 60%, the proporation of male customers who are satisfied with the flight is about 40%.

```
sat %>% group_by(gender) %>% summarise(percent=mean(satisfaction_v2=="satisfied")) %>%
  ggplot(aes(x=gender, y=percent)) + geom_bar(stat="identity", fill="steelblue") +
  labs(y="Satisfaction Rate", x="Gender", title="Satisfaction rate by gender") +
  theme(plot.title = element_text(hjust = 0.5))
```



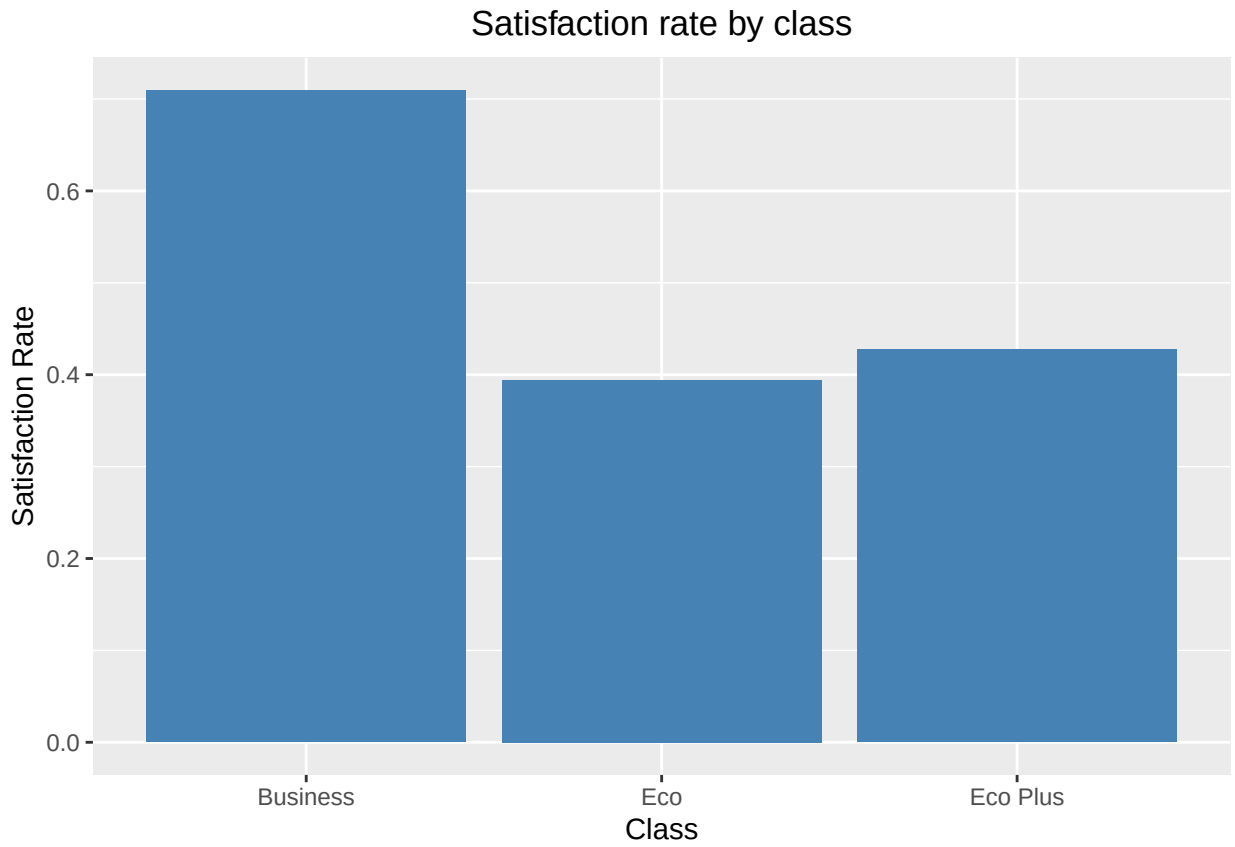
Calculate the proporation of satisfaction rate by customer type, the proporation of loyal customers who are satisfied with the flight is higher than the proporation of disloyal customers who are satisfied with the flight. The proporation of loyal customers who are satisfied with the flight is about 62%, the proporation of disloyal customers who are satisfied with the flight is about 24%.

```
sat %>% group_by(customer_type) %>% summarise(percent=mean(satisfaction_v2=="satisfied")) %>%
  ggplot(aes(x=customer_type, y=percent)) + geom_bar(stat="identity", fill="steelblue") +
  labs(y="Satisfaction Rate", x="Customer Type", title="Satisfaction rate by customer type") +
  theme(plot.title = element_text(hjust = 0.5))
```



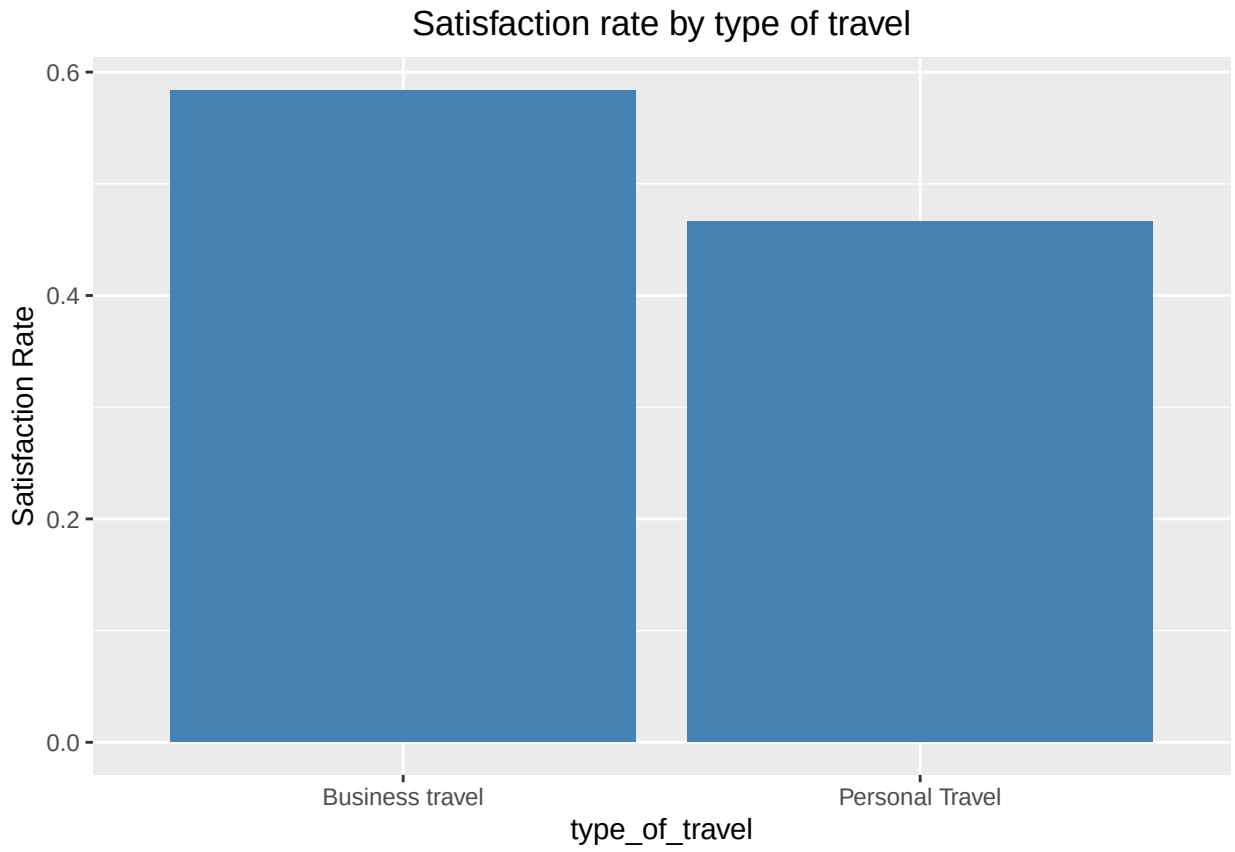
Calculate the proporation of satisfaction rate by class, the proporation of customers from business class who are satisfied with the flight is higher than the proporation of customers from other two classes who are satisfied with the flight. The proporation of customers from business class who are satisfied with the flight is about 71%, the proporation of customers from economy class who are satisfied with the flight is about 39%, the proporation of customers from economy plus class who are satisfied with the flight is about 43%.

```
sat %>% group_by(class) %>% summarise(percent=mean(satisfaction_v2=="satisfied")) %>%
  ggplot(aes(x=class, y=percent)) + geom_bar(stat="identity", fill="steelblue") +
  labs(y="Satisfaction Rate", x="Class", title="Satisfaction rate by class") +
  theme(plot.title = element_text(hjust = 0.5))
```



Calculate the proporation of satisfaction rate by traval type, the proporation of customers from business traval who are satisfied with the flight is higher than the proporation of customers from personal travel who are satisfied with the flight. The proporation of customers from business traval who are satisfied with the flight is about 58%, the proporation of customers from personal travel who are satisfied with the flight is about 47%.

```
sat %>% group_by(type_of_travel) %>% summarise(percent=mean(satisfaction_v2=="satisfied")) %>%
  ggplot(aes(x=type_of_travel, y=percent)) + geom_bar(stat="identity", fill="steelblue") +
  labs(y="Satisfaction Rate", xlab="Type of travel", title="Satisfaction rate by type of travel") +
  theme(plot.title = element_text(hjust = 0.5))
```



Modeling

Split the data into training dataset and test dataset. Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when we separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, we can minimize the effects of data discrepancies and better understand the characteristics of the model. Here we will plit randomly this data set into 70% train and 30% test. After spliting, there are 90640 observations in the train dataset, and 38847 observations in the test dataset.

```
## [1] 90640    24
```

```
## [1] 38847    24
```

Logistic Regression Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Here the satisfacton contains only two unique value, so we can apply logistic here, use the satisfacton as the response variable, and all other variables except ID as the predictors.

```
glm.mod <- glm(satisfaction_v2~., train[, -1], family = binomial)
summary(glm.mod)
```

```
##
## Call:
## glm(formula = satisfaction_v2 ~ ., family = binomial, data = train[,
```

```
##     -1])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0109  -0.5768   0.1917   0.5184   3.5652
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -6.839e+00  7.826e-02 -87.395  < 2e-16 ***
## genderMale                        -9.744e-01  1.970e-02 -49.475  < 2e-16 ***
## customer_typeLoyal Customer        1.986e+00  2.998e-02  66.225  < 2e-16 ***
## age                               -7.381e-03  6.844e-04 -10.784  < 2e-16 ***
## type_of_travelPersonal Travel     -7.648e-01  2.807e-02 -27.251  < 2e-16 ***
## classEco                          -7.461e-01  2.543e-02 -29.338  < 2e-16 ***
## classEco Plus                     -7.596e-01  3.904e-02 -19.455  < 2e-16 ***
## flight_distance                   -1.073e-04  1.032e-05 -10.389  < 2e-16 ***
## seat_comfort                       2.987e-01  1.097e-02  27.228  < 2e-16 ***
## departure_arrival_time_convenient -1.920e-01  8.090e-03 -23.736  < 2e-16 ***
## food_and_drink                    -2.311e-01  1.114e-02 -20.739  < 2e-16 ***
## gate_location                      1.157e-01  9.151e-03  12.647  < 2e-16 ***
## inflight_wifi_service             -8.127e-02  1.069e-02  -7.599 2.97e-14 ***
## inflight_entertainment             6.809e-01  9.917e-03  68.656  < 2e-16 ***
## online_support                     8.521e-02  1.082e-02   7.875 3.41e-15 ***
## ease_of_online_booking             2.278e-01  1.394e-02  16.344  < 2e-16 ***
## on_board_service                   3.157e-01  9.842e-03  32.082  < 2e-16 ***
## leg_room_service                   2.206e-01  8.412e-03  26.225  < 2e-16 ***
## baggage_handling                   9.871e-02  1.114e-02   8.857  < 2e-16 ***
## checkin_service                    2.992e-01  8.313e-03  35.984  < 2e-16 ***
## cleanliness                        8.911e-02  1.160e-02   7.684 1.55e-14 ***
## online_boarding                    1.825e-01  1.191e-02  15.321  < 2e-16 ***
## departure_delay_in_minutes         4.512e-03  9.690e-04   4.657 3.21e-06 ***
## arrival_delay_in_minutes          -9.557e-03  9.545e-04 -10.012  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 124835  on 90639  degrees of freedom
## Residual deviance:  69666  on 90616  degrees of freedom
## AIC: 69714
##
## Number of Fisher Scoring iterations: 5
```

```r
probs <- predict(glm.mod, test, type="response")
preds <- ifelse(probs>0.5, "satisfied", "neutral or dissatisfied")
mean(preds==test$satisfaction_v2)
```

```
## [1] 0.834067
```

All the p-value of the predictors is less than 0.05, so all the predictors are significant, the positive coefficient means that the predictor will increase the probablity of satisfaction, the negative coefficeint means that the predictor will decrease the probability of satisfaction, the auccuray of the model is 0.836.

KNN K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern

recognition already in the beginning of 1970's as a non-parametric technique. K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

Here will use all available numeric variables, apply a KNN model with k=2 to predict the satisfaction of the customers.

```
library(class)
train.x <- as.matrix(train[, -c(1:7)])
test.x <- as.matrix(test[, -c(1:7)])
# k = 2
preds <- knn(train.x, test.x, train$satisfaction_v2, k=2)
cm <- table(preds, test$satisfaction_v2)
mean(preds==test$satisfaction_v2)
```

```
## [1] 0.7151646
```

The accuray of the KNN(k=2) is 0.716.

LDA Discriminant analysis is used to predict the probability of belonging to a given class (or category) based on one or multiple predictor variables. It works with continuous and/or categorical predictor variables.

Compared to logistic regression, the discriminant analysis is more suitable for predicting the category of an observation in the situation where the outcome variable contains more than two classes. Additionally, it's more stable than the logistic regression for multi-class classification problems.

The LDA algorithm starts by finding directions that maximize the separation between classes, then use these directions to predict the class of individuals. These directions, called linear discriminants, are a linear combinations of predictor variables.

LDA assumes that predictors are normally distributed (Gaussian distribution) and that the different classes have class-specific means and equal variance/covariance.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

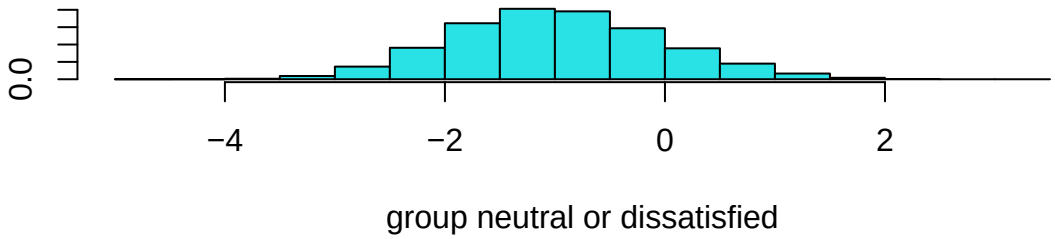
```
## The following object is masked from 'package:dplyr':
##
##     select
```

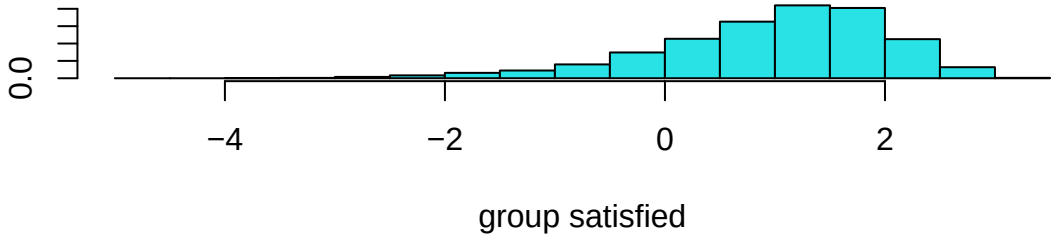
```
lda.mod <- lda(satisfaction_v2~., train[, -1])
lda.mod
```

```
## Call:
## lda(satisfaction_v2 ~ ., data = train[, -1])
##
## Prior probabilities of groups:
## neutral or dissatisfied                satisfied
##               0.4525154                0.5474846
##
## Group means:
##                          genderMale customer_typeLoyal Customer      age
## neutral or dissatisfied  0.6110055                   0.6918032 37.47542
## satisfied                0.3951515                   0.9204014 41.11150
##                          type_of_travelPersonal Travel  classEco classEco Plus
## neutral or dissatisfied                      0.3618344 0.6006924    0.08950166
## satisfied                                    0.2641061 0.3213163    0.05779462
```

```
##                         flight_distance seat_comfort
## neutral or dissatisfied        2025.229     2.467257
## satisfied                      1946.235     3.144386
##                         departure_arrival_time_convenient food_and_drink
## neutral or dissatisfied                          3.011605       2.661961
## satisfied                                        2.965480       3.008101
##                         gate_location inflight_wifi_service
## neutral or dissatisfied      3.008631              2.922274
## satisfied                    2.975314              3.517109
##                         inflight_entertainment online_support
## neutral or dissatisfied               2.611176       2.962722
## satisfied                             4.021320       3.978458
##                         ease_of_online_booking on_board_service
## neutral or dissatisfied               2.850059         2.964112
## satisfied                             3.981864         3.871917
##                         leg_room_service baggage_handling checkin_service
## neutral or dissatisfied         3.048786         3.359104        2.972767
## satisfied                       3.843523         3.968281        3.649565
##                         cleanliness online_boarding departure_delay_in_minutes
## neutral or dissatisfied    3.368393        2.870246                    17.68346
## satisfied                  3.977007        3.749758                    12.20180
##                         arrival_delay_in_minutes
## neutral or dissatisfied                 18.46002
## satisfied                               12.31170
##
## Coefficients of linear discriminants:
##                                             LD1
## genderMale                          -5.609641e-01
## customer_typeLoyal Customer          1.204465e+00
## age                                 -4.586210e-03
## type_of_travelPersonal Travel       -4.223054e-01
## classEco                            -4.377072e-01
## classEco Plus                       -4.393767e-01
## flight_distance                     -5.856361e-05
## seat_comfort                         2.055798e-01
## departure_arrival_time_convenient -1.226482e-01
## food_and_drink                      -1.407165e-01
## gate_location                        5.861508e-02
## inflight_wifi_service               -3.901660e-02
## inflight_entertainment               4.078718e-01
## online_support                       5.476289e-02
## ease_of_online_booking               1.447011e-01
## on_board_service                     1.773196e-01
## leg_room_service                     1.198788e-01
## baggage_handling                     5.057208e-02
## checkin_service                      1.609221e-01
## cleanliness                          4.649826e-02
## online_boarding                      9.299617e-02
## departure_delay_in_minutes           2.372970e-03
## arrival_delay_in_minutes            -4.749928e-03
```

```
plot(lda.mod)
```






```
lda.pred <- predict(lda.mod, test)
mean(lda.pred$class==test$satisfaction_v2)
```

```
## [1] 0.8358174
```

The accuray of the model is 0.837.

Conclusion

Based on the computation above, we can found that the LDA model has the highest accuray rate, so the LDA model is the best classification model among the three models. The accuray is 0.837, it is moderate strong.