VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY
HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
FACULTY OF TRANSPORTATION

PROBABILITY AND STATISTICS PROJECT

EVALUATE THE AVIATION ACTIVITIES FROM 2005 TO 2016 AND THE RELATIONSHIP WITH PRICE CATEGORY

Instructor: Doc.Nguyễn Tiến Dũng

Class: MT2013 – CC01

Group: 13

Full Name	Student's ID	Faculty
Triệu Quốc Khải	2153451	Transportation engineering
Đinh Văn Việt Hùng	2153401	Transportation engineering
Phạm Đình Khang	2252309	Mechanical Engineering
Nguyễn Võ Trung Tài	2053414	Mechanical Engineering
Nguyễn Quốc Hưng	2153411	Civil Engineering

Ho Chi Minh City – Thursday, 2nd May, 2024

TABLE OF CONTENTS

I. Dataset overview

The data set provides details on air traffic passenger statistics categorized by airlines, airports, and regions for departing and arriving flights. It also covers aspects like activity type, price category, terminal, boarding area, and passenger numbers. Air traffic data is valuable for comprehending the airline industry and planning trips. This dataset sourced from Open Flights spans from 2005 to 2008 and includes passenger numbers, operating airlines, published airlines, geographic regions, activity type codes, price category codes, terminals, boarding areas, as well as flight years and months.

II. Knowledge basis

Logistic Regression

In statistics, the logistic model (or logit model) is a statistical model that models the log-odds of an event as a linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names[1].

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression. Many other medical scales used to

[1] Wikipedia, (20/04/2024), *Logistic regression*, Link: https://en.wikipedia.org/wiki/Logistic_regression

assess severity of a patient have been developed using logistic regression. Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.). Another example might be to predict whether a Nepalese voter will vote Nepali Congress or Communist Party of Nepal or Any Other Party, based on age, income, sex, race, state of residence, votes in previous elections, etc. The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product. It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription, etc. In economics, it can be used to predict the likelihood of a person ending up in the labor force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing. Disaster planners and engineers rely on these models to predict decision take by householders or building occupants in small-scale and large-scales evacuations ,such as building fires, wildfires, hurricanes among others. These models help in the development of reliable disaster managing plans and safer design for the built environment.

Logistic regression function[2]:

$$h_\beta(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \cdots + \beta_k x_k)}}$$

We also have the likehood function:

$$L(\beta) = P(y_1, y_2, \ldots, y_n | x, \beta) = \prod_{i=1}^{n} P(y_i | x_i, \beta) = \prod_{i=1}^{n} (h_\beta(x_i))^{y_i} \, (1 - h_\beta(x_i))^{1-y_i}$$

Then, we calculate the Log of the previous likehood function:

$$Log\big(L(\beta)\big) = \sum_{i=1}^{n} [y_i \log\Big(h_\beta(x_i)\Big) + (1 - y_i) \log\Big(1 - h_\beta(x_i)\Big)]$$

After that, we find the vecto β where $Log\big(L(\beta)\big)$ is maximum.

[2] Wikipedia, (15/03/2024), *Multinomial Logistic Regression*, Link: https://en.wikipedia.org/wiki/Multinomial_logistic_regression

III. Preprocessing of data

While the logistic regression method need a single binary dependent variable, then we have to process the data to the correct fomat. First, we need to remove columns with independent variables that have high correlation value with others because these variables are not only unnecessary, which makes the regression model more complicate, but also affect the accuracy of the final model. Here, we find out these columns are as same as the ones which could show you more details exiting in the data. Then, we change all the characters data following an order to numbers before we import it to R to calculate the probability:

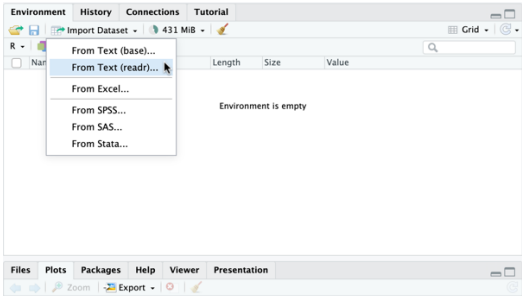
	A	B	C	D	E	F	
1	index	Activity Period	Operating Airline	Operating Airline IATA Code	Published Airline	Published Airline IATA Code	GE
2	0	200507	ATA Airlines	TZ	ATA Airlines	TZ	Don
3	1	200507	ATA Airlines	TZ	ATA Airlines	TZ	Don
4	2	200507	ATA Airlines	TZ	ATA Airlines	TZ	Don
5	3	200507	Air Canada	AC	Air Canada	AC	Inte
6	4	200507	Air Canada	AC	Air Canada	AC	Inte
7	5	200507	Air China	CA	Air China	CA	Inte
8	6	200507	Air China	CA	Air China	CA	Inte
9	7	200507	Air France	AF	Air France	AF	Inte
10	8	200507	Air France	AF	Air France	AF	Inte
11	9	200507	Air New Zealand	NZ	Air New Zealand	NZ	Inte
12	10	200507	Air New Zealand	NZ	Air New Zealand	NZ	Inte
13	11	200507	AirTran Airways	FL	AirTran Airways	FL	Don
14	12	200507	AirTran Airways	FL	AirTran Airways	FL	Don
15	13	200507	Alaska Airlines	AS	Alaska Airlines	AS	Don
16	14	200507	Alaska Airlines	AS	Alaska Airlines	AS	Don
17	15	200507	Alaska Airlines	AS	Alaska Airlines	AS	Don
18	16	200507	Alaska Airlines	AS	Alaska Airlines	AS	Inte
19	17	200507	Alaska Airlines	AS	Alaska Airlines	AS	Inte
20	18	200507	Alaska Airlines	AS	Alaska Airlines	AS	Inte

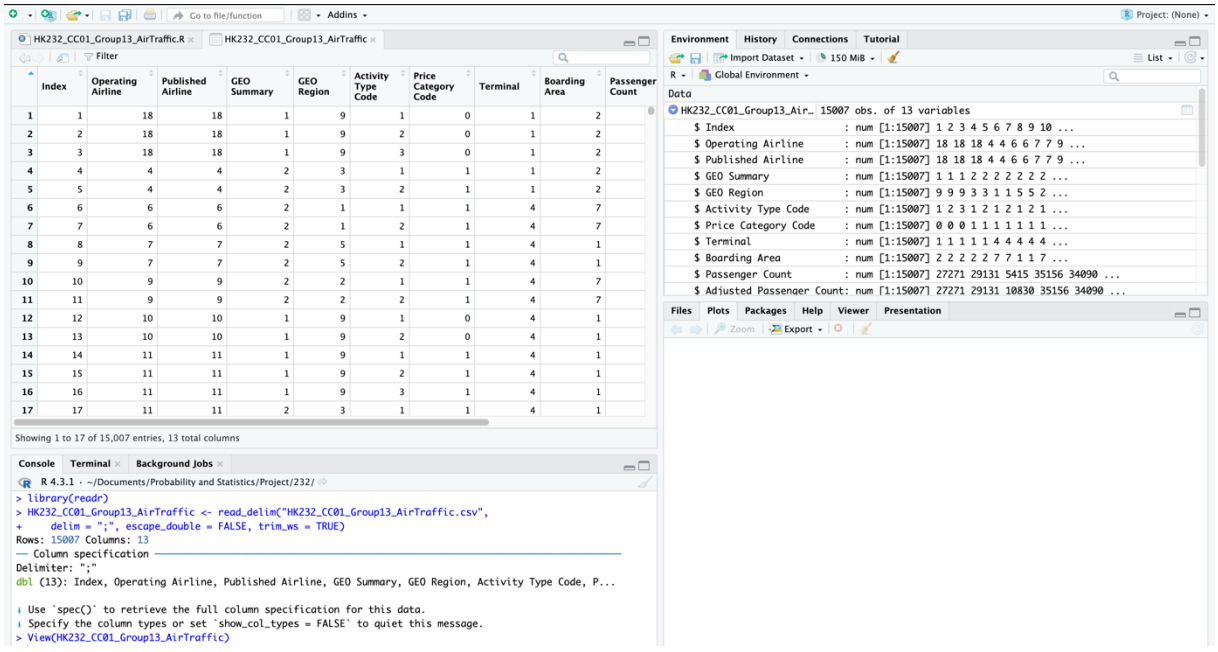
	A	B	C	D	E	F	
1	index	Activity Period	Operating Airline	Published Airline	GEO Summary	GEO Region	Activity T
2	0	200507	ATA Airlines	ATA Airlines	Domestic	US	Deplaned
3	1	200507	ATA Airlines	ATA Airlines	Domestic	US	Enplaned
4	2	200507	ATA Airlines	ATA Airlines	Domestic	US	Thru / Tra
5	3	200507	Air Canada	Air Canada	International	Canada	Deplaned
6	4	200507	Air Canada	Air Canada	International	Canada	Enplaned
7	5	200507	Air China	Air China	International	Asia	Deplaned
8	6	200507	Air China	Air China	International	Asia	Enplaned
9	7	200507	Air France	Air France	International	Europe	Deplaned
10	8	200507	Air France	Air France	International	Europe	Enplaned
11	9	200507	Air New Zealand	Air New Zealand	International	Australia / Oceania	Deplaned
12	10	200507	Air New Zealand	Air New Zealand	International	Australia / Oceania	Enplaned
13	11	200507	AirTran Airways	AirTran Airways	Domestic	US	Deplaned
14	12	200507	AirTran Airways	AirTran Airways	Domestic	US	Enplaned
15	13	200507	Alaska Airlines	Alaska Airlines	Domestic	US	Deplaned
16	14	200507	Alaska Airlines	Alaska Airlines	Domestic	US	Enplaned
17	15	200507	Alaska Airlines	Alaska Airlines	Domestic	US	Thru / Tra
18	16	200507	Alaska Airlines	Alaska Airlines	International	Canada	Deplaned
19	17	200507	Alaska Airlines	Alaska Airlines	International	Canada	Enplaned
20	18	200507	Alaska Airlines	Alaska Airlines	International	Mexico	Deplaned

	A	B	C	D	E	F	G	H	I	J	
1	Index	Operating Airline	Published Airline	GEO Summary	GEO Region	Activity Type Code	Price Category Code	Terminal	Boarding Area	Passenger Count	Adju Pass
2	1	18	18	1	9	1	0	1	2	27271	
3	2	18	18	1	9	2	0	1	2	29131	
4	3	18	18	1	9	3	0	1	2	5415	
5	4	4	4	2	3	1	1	1	2	35156	
6	5	4	4	2	3	2	1	1	2	34090	
7	6	6	6	2	1	1	1	4	7	6263	
8	7	6	6	2	1	2	1	4	7	5500	
9	8	7	7	2	5	1	1	4	1	12050	
10	9	7	7	2	5	2	1	4	1	11638	
11	10	9	9	2	2	1	1	4	7	4998	
12	11	9	9	2	2	2	1	4	7	4962	
13	12	10	10	1	9	1	0	4	1	8055	
14	13	10	10	1	9	2	0	4	1	7984	
15	14	11	11	1	9	1	1	4	1	36641	
16	15	11	11	1	9	2	1	4	1	39379	
17	16	11	11	1	9	3	1	4	1	3678	
18	17	11	11	2	3	1	1	4	1	7977	
19	18	11	11	2	3	2	1	4	1	8837	
20	19	11	11	2	6	1	1	4	1	6969	

The rule for our order is included in a file Excel named "*Air_Traffic_Passenger_Statistics_Order.xlsx*", a chart that we have submitted online.

After making a suitable change on the original data set, we import the chart in file CSV, which is renamed to be "*HK232_CC01_Group13_AirTraffic.csv*", into the environment of R so as to run our code:





We continue to check the information of it in R, we have made it to shown the first six lines of the data set:

```
> head(HK232_CC01_Group13_AirTraffic)
# A tibble: 6 × 13
  Index `Operating Airline` `Published Airline` `GEO Summary` `GEO Region` `Activity Type Code`
  <dbl>               <dbl>               <dbl>         <dbl>        <dbl>                <dbl>
1     1                  18                  18             1            9                    1
2     2                  18                  18             1            9                    2
3     3                  18                  18             1            9                    3
4     4                   4                   4             2            3                    1
5     5                   4                   4             2            3                    2
6     6                   6                   6             2            1                    1
# ℹ 7 more variables: `Price Category Code` <dbl>, Terminal <dbl>, `Boarding Area` <dbl>,
#   `Passenger Count` <dbl>, `Adjusted Passenger Count` <dbl>, Year <dbl>, Month <dbl>
```

Checked the structure of database:

```
> str(HK232_CC01_Group13_AirTraffic)
spc_tbl_ [15,007 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Index                  : num [1:15007] 1 2 3 4 5 6 7 8 9 10 ...
 $ Operating Airline      : num [1:15007] 18 18 18 4 4 6 6 7 7 9 ...
 $ Published Airline      : num [1:15007] 18 18 18 4 4 6 6 7 7 9 ...
 $ GEO Summary            : num [1:15007] 1 1 1 2 2 2 2 2 2 2 ...
 $ GEO Region             : num [1:15007] 9 9 9 3 3 1 1 5 5 2 ...
 $ Activity Type Code     : num [1:15007] 1 2 3 1 2 1 2 1 2 1 ...
 $ Price Category Code    : num [1:15007] 0 0 0 1 1 1 1 1 1 1 ...
 $ Terminal               : num [1:15007] 1 1 1 1 1 4 4 4 4 4 ...
 $ Boarding Area          : num [1:15007] 2 2 2 2 2 7 7 1 1 7 ...
 $ Passenger Count        : num [1:15007] 27271 29131 5415 35156 34090 ...
 $ Adjusted Passenger Count: num [1:15007] 27271 29131 10830 35156 34090 ...
 $ Year                   : num [1:15007] 2005 2005 2005 2005 2005 ...
 $ Month                  : num [1:15007] 7 7 7 7 7 7 7 7 7 7 ...
 - attr(*, "spec")=
  .. cols(
  ..   Index = col_double(),
  ..   `Operating Airline` = col_double(),
  ..   `Published Airline` = col_double(),
  ..   `GEO Summary` = col_double(),
  ..   `GEO Region` = col_double(),
  ..   `Activity Type Code` = col_double(),
  ..   `Price Category Code` = col_double(),
  ..   Terminal = col_double(),
  ..   `Boarding Area` = col_double(),
  ..   `Passenger Count` = col_double(),
  ..   `Adjusted Passenger Count` = col_double(),
  ..   Year = col_double(),
  ..   Month = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

And find out the length of it by seeing the total rows:

```
> nrow(HK232_CC01_Group13_AirTraffic)
[1] 15007
```

In the final data processing, we divided the data set into 2 parts, train data and test data. Train data is used to create a regression model to serve the goal of the topic. Test data is the part to test the results of that regression model:

```
> # Set seed for reproducibility
> set.seed(123)
> # Generate random indices for data partitioning
> indices <- sample(1:nrow(HK232_CC01_Group13_AirTraffic))
> train_indices <- indices[1:round(0.7 * length(indices))]
> test_indices <- indices[(round(0.7 * length(indices)) + 1):length(indices)]
> # Split the data into training and testing sets
> train_data <- HK232_CC01_Group13_AirTraffic[train_indices, ]
> test_data <- HK232_CC01_Group13_AirTraffic[test_indices, ]
```

test_data	4502 obs. of 13 variables
train_data	10505 obs. of 13 variables
Values	
indices	int [1:15007] 2463 2511 10419 8718 12483 2986 1842 9334 3371 13…
test_indices	int [1:4502] 13235 8180 6484 8716 6420 9455 14197 11715 2734 85…
train_indices	int [1:10505] 2463 2511 10419 8718 12483 2986 1842 9334 3371 13…

IV. Inferential statistics

Target

We aim to used the logistic regression for finding which category groups of airport statistics have low price. The logit function returns values between 0 and 1 for the dependent variable, regardless of the value of the independent variable. This is how logistic regression estimates the dependent variable value, which is Price Category Code. By the way, logistic regression also models equations between multiple independent variables on a single dependent variable. To do this, we use glm (Generalized Linear Models) function in R to build our model of train data:

```
> model <- glm(`Price Category Code` ~
+                `Operating Airline` +
+                `Published Airline` +
+                `GEO Summary` +
+                `GEO Region` +
+                `Activity Type Code` +
+                `Terminal` +
+                `Boarding Area` +
+                `Passenger Count` +
+                `Adjusted Passenger Count` +
+                `Year` +
+                `Month`,
+              data = train_data, family = "binomial")
```

```
model                 List of  30
  $ coefficients     : Named num [1:12] 82.1642 0.0242 -0.0497 0.6906 -0.6251 ...
   ..- attr(*, "names")= chr [1:12] "(Intercept)" "`Operating Airline`" "`Published Air…
  $ residuals        : Named num [1:10505] 3.44 1 -1.98 1 1 ...
   ..- attr(*, "names")= chr [1:10505] "2463" "2511" "10419" "8718" ...
  $ fitted.values    : Named num [1:10505] 0.291 0.998 0.496 0.998 0.995 ...
   ..- attr(*, "names")= chr [1:10505] "2463" "2511" "10419" "8718" ...
  $ effects          : Named num [1:10505] -26.69 7.83 7.59 -20.58 -11.25 ...
   ..- attr(*, "names")= chr [1:10505] "(Intercept)" "`Operating Airline`" "`Published …
  $ R                : num [1:12, 1:12] -28.1 0 0 0 0 ...
   ..- attr(*, "dimnames")=List of 2
   .. ..$ : chr [1:12] "(Intercept)" "`Operating Airline`" "`Published Airline`" "`GEO …
   .. ..$ : chr [1:12] "(Intercept)" "`Operating Airline`" "`Published Airline`" "`GEO …
  $ rank             : int 12
  $ qr               :List of 5
   ..$ qr   : num [1:10505, 1:12] -28.09207 0.00149 0.0178 0.00165 0.00238 ...
   .. ..- attr(*, "dimnames")=List of 2
   .. .. ..$ : chr [1:10505] "2463" "2511" "10419" "8718" ...
   .. .. ..$ : chr [1:12] "(Intercept)" "`Operating Airline`" "`Published Airline`" "`G…
   ..$ rank : int 12
   ..$ qraux: num [1:12] 1.02 1 1 1 1 ...
   ..$ pivot: int [1:12] 1 2 3 4 5 6 7 8 9 10 ...
   ..$ tol  : num 1e-11
   ..- attr(*, "class")= chr "qr"
  $ family           :List of 13
   ..$ family    : chr "binomial"
   ..$ link      : chr "logit"
   ..$ linkfun   :function (mu)
   ..$ linkinv   :function (eta)
   ..$ variance  :function (mu)
   ..$ dev.resids:function (y, mu, wt)
   ..$ aic       :function (y, n, mu, wt, dev)
   ..$ mu.eta    :function (eta)
   ..$ initialize: language {     if (NCOL(y) == 1) { ...
   ..$ validmu   :function (mu)
   ..$ valideta  :function (eta)
   ..$ simulate  :function (object, nsim)
```

Results

After building the regression model and testing it using the AIC (Akaike Information Criterion) method, we obtain the regression coefficients of the independent variables that influence the model as well as draw a symbolic graph of the logistic regression function. From there, the function and chart will yield groups with low airfare prices. The AIC value reflects a trade-off between goodness of fit and model complexity.

Lower AIC values indicate better models, with the "best" model being the one with the lowest AIC value among the candidate models:

```
> step_model <- stepAIC(model, direction = "both")
Start:  AIC=5045.94
`Price Category Code` ~ `Operating Airline` + `Published Airline` +
    `GEO Summary` + `GEO Region` + `Activity Type Code` + Terminal +
    `Boarding Area` + `Passenger Count` + `Adjusted Passenger Count` +
    Year + Month

                            Df Deviance    AIC
- Terminal                   1   5022.1 5044.1
- `Activity Type Code`       1   5022.7 5044.7
- Month                      1   5023.1 5045.1
<none>                           5021.9 5045.9
- `GEO Summary`              1   5030.1 5052.1
- Year                       1   5031.7 5053.7
- `Passenger Count`          1   5049.0 5071.0
- `Operating Airline`        1   5049.3 5071.3
- `Adjusted Passenger Count` 1   5049.5 5071.5
- `Published Airline`        1   5166.0 5188.0
- `GEO Region`               1   5185.8 5207.8
- `Boarding Area`            1   5654.0 5676.0
```

```
Step:  AIC=5044.08
`Price Category Code` ~ `Operating Airline` + `Published Airline` +
    `GEO Summary` + `GEO Region` + `Activity Type Code` + `Boarding Area` +
    `Passenger Count` + `Adjusted Passenger Count` + Year + Month

                            Df Deviance    AIC
- `Activity Type Code`       1   5022.8 5042.8
- Month                      1   5023.2 5043.2
<none>                           5022.1 5044.1
+ Terminal                   1   5021.9 5045.9
- `GEO Summary`              1   5030.2 5050.2
- Year                       1   5032.5 5052.5
- `Passenger Count`          1   5049.0 5069.0
- `Adjusted Passenger Count` 1   5049.5 5069.5
- `Operating Airline`        1   5050.6 5070.6
- `Published Airline`        1   5173.7 5193.7
- `GEO Region`               1   5185.8 5205.8
- `Boarding Area`            1   5672.1 5692.1
```

```
Step:  AIC=5042.85
`Price Category Code` ~ `Operating Airline` + `Published Airline` +
    `GEO Summary` + `GEO Region` + `Boarding Area` + `Passenger Count` +
    `Adjusted Passenger Count` + Year + Month

                            Df Deviance    AIC
- Month                      1   5024.0 5042.0
<none>                           5022.8 5042.8
+ `Activity Type Code`       1   5022.1 5044.1
+ Terminal                   1   5022.7 5044.7
- `GEO Summary`              1   5031.2 5049.2
- Year                       1   5033.7 5051.7
- `Passenger Count`          1   5049.7 5067.7
- `Adjusted Passenger Count` 1   5050.2 5068.2
- `Operating Airline`        1   5051.3 5069.3
- `Published Airline`        1   5175.3 5193.3
- `GEO Region`               1   5186.3 5204.3
- `Boarding Area`            1   5673.5 5691.5
```

```
Step:  AIC=5041.99
`Price Category Code` ~ `Operating Airline` + `Published Airline` +
    `GEO Summary` + `GEO Region` + `Boarding Area` + `Passenger Count` +
    `Adjusted Passenger Count` + Year

                            Df Deviance    AIC
<none>                           5024.0 5042.0
+ Month                      1   5022.8 5042.8
+ `Activity Type Code`       1   5023.2 5043.2
+ Terminal                   1   5023.9 5043.9
- `GEO Summary`              1   5032.3 5048.3
- Year                       1   5034.1 5050.1
- `Passenger Count`          1   5050.9 5066.9
- `Adjusted Passenger Count` 1   5051.4 5067.4
- `Operating Airline`        1   5052.7 5068.7
- `Published Airline`        1   5177.0 5193.0
- `GEO Region`               1   5187.5 5203.5
- `Boarding Area`            1   5675.4 5691.4
```

By summarying the final model, we get the result on regression coefficients of it:

```
> summary(step_model)

Call:
glm(formula = `Price Category Code` ~ `Operating Airline` + `Published Airline` +
    `GEO Summary` + `GEO Region` + `Boarding Area` + `Passenger Count` +
    `Adjusted Passenger Count` + Year, family = "binomial", data = train_data)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                80.9965926 23.8928726   3.390 0.000699 ***
`Operating Airline`         0.0245279  0.0047940   5.116 3.11e-07 ***
`Published Airline`        -0.0500972  0.0045241 -11.073  < 2e-16 ***
`GEO Summary`               0.6749581  0.2340752   2.884 0.003933 **
`GEO Region`               -0.6247807  0.0567870 -11.002  < 2e-16 ***
`Boarding Area`             0.5440001  0.0251025  21.671  < 2e-16 ***
`Passenger Count`          -0.0004415  0.0001230  -3.590 0.000330 ***
`Adjusted Passenger Count`  0.0004448  0.0001230   3.616 0.000300 ***
Year                       -0.0377364  0.0119199  -3.166 0.001546 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8043  on 10504  degrees of freedom
Residual deviance: 5024  on 10496  degrees of freedom
AIC: 5042

Number of Fisher Scoring iterations: 8
```

Comment

We need to process a large amount of raw data, which is a drawback. Additionally, the R code for this job is too lengthy and complicated. The process of verifying with AIC using loops also makes it much longer. However, the above limited process yields a very high accuracy:

```
> # Calculate accuracy
> accuracy <- mean(test_predicted_classes == test_data$`Price Category Code`)
> cat("Accuracy:", accuracy, "\n")
Accuracy: 0.8827188
```

Conclusion

This method is very reasonable, although there are many shortcomings mentioned in the above comments, we get the expected results. They have very high accuracy (as above 88%) and very clear statistics.

V. Discuss and expand

Advantage

Logistic regression is a flexible and powerful method for predicting binary values as “Price Category Code” varible and discrete variables as all other independent variables. It allows us to estimate the probability of an event occurring, helping us better understand the correlation between independent and dependent values.

It also has application widely in some different fields such as health, economics, marketing, …

Disadvantage

It assumes that the independent variables do not have a linear correlation, and the model's result will not be accurate if a linear relationship exists. Moreover, it can be affected by outliers or missing values, skewing the results and losing the correctness of the model. The reason our regression model achieves high accuracy (about 88%) is because we have preprocessed the data and the data set for training and testing is taken from the database itself.

Critically, it is not suitable for predicting continuous value.

Expand

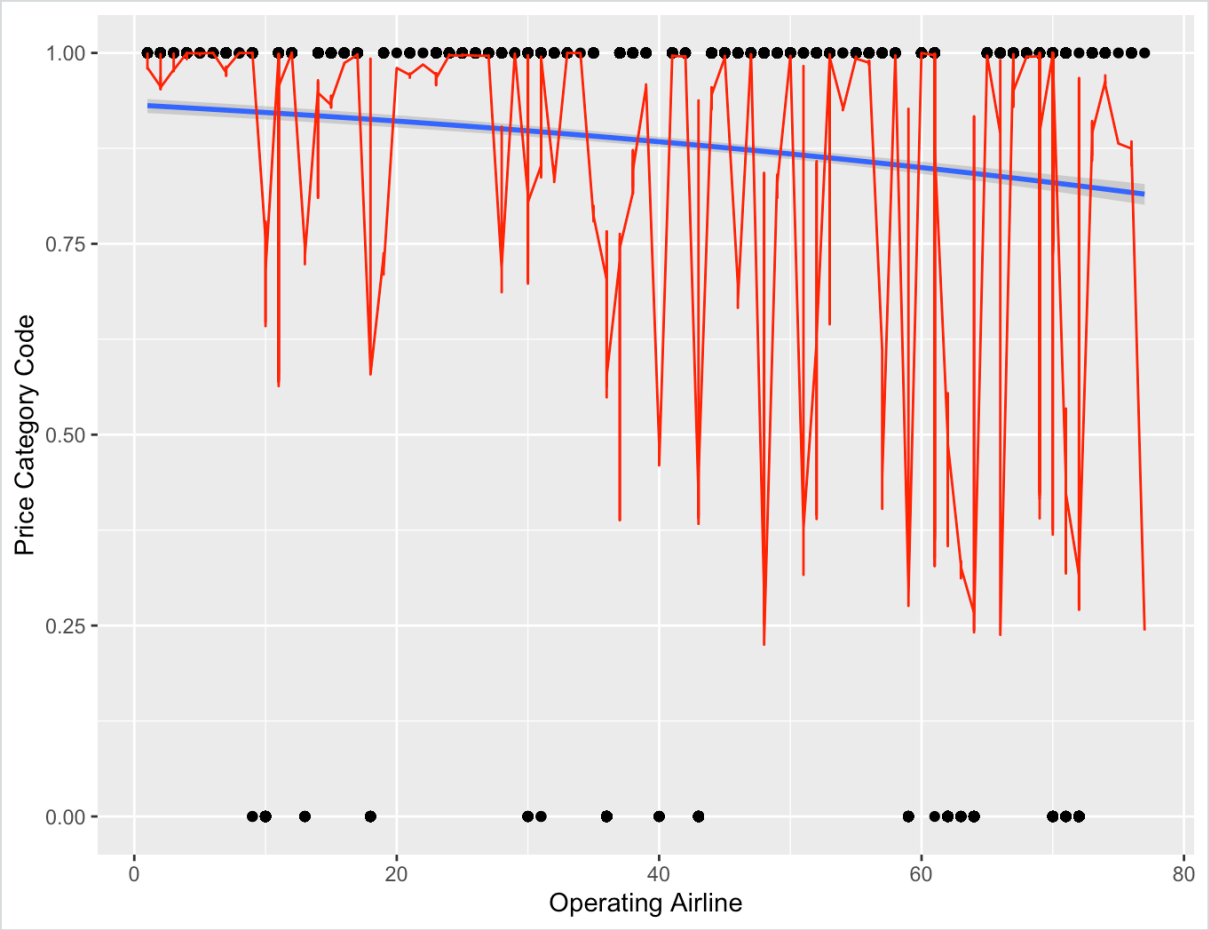
Besides the main goal, if we add new data that wasn't part of the original dataset, we can still calculate the probability of an event (like a 'Low Fare' or 'Other'). Once the model is trained, we can use it to predict the probability of the outcome for new data points. This is done by plugging in the values of the new data into the logistic function,

as we mentioned in the theory above, using the coefficients obtained from the original model.

VI. Descriptive statistics

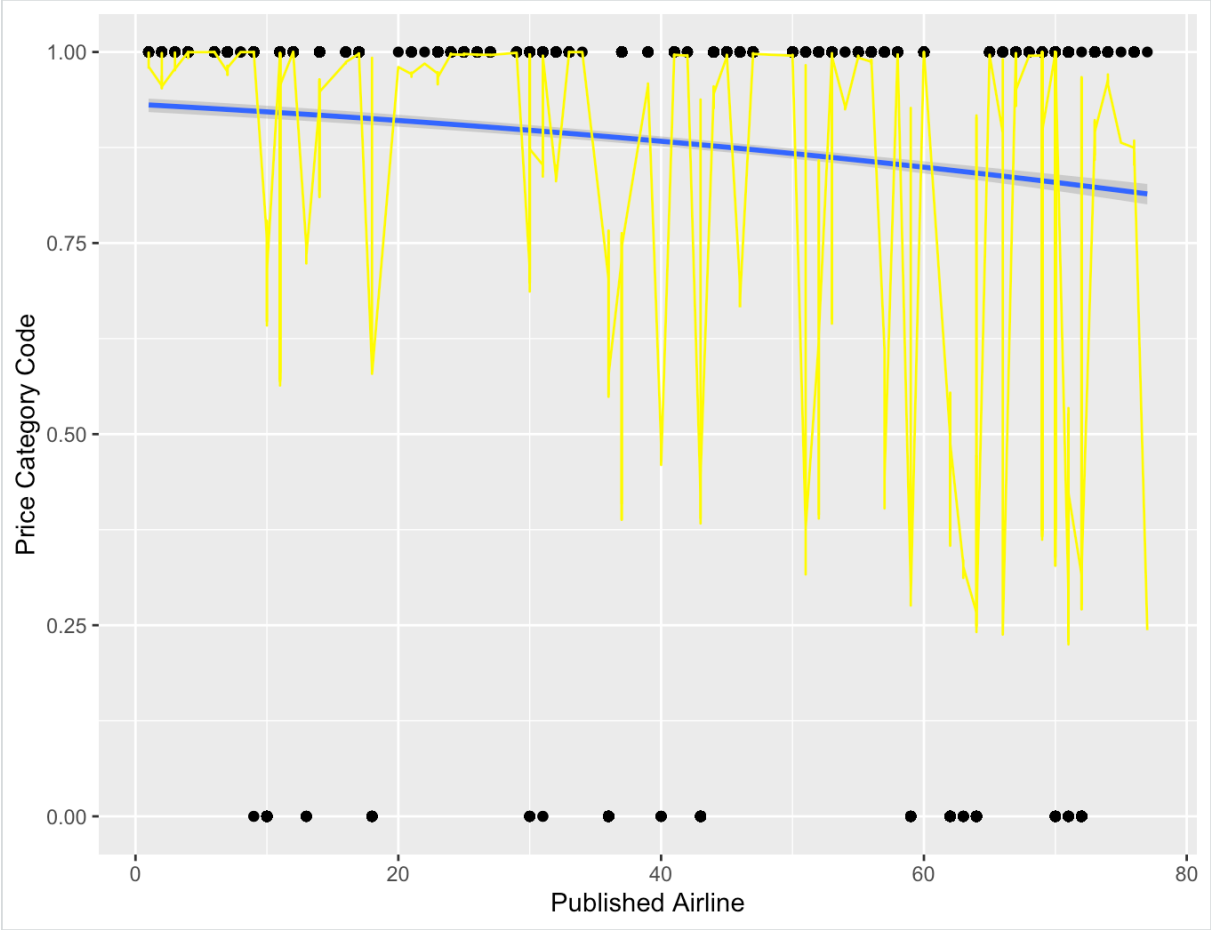
We will discuss on the charts of all final modes that we have checked by code R:

Operating Airline



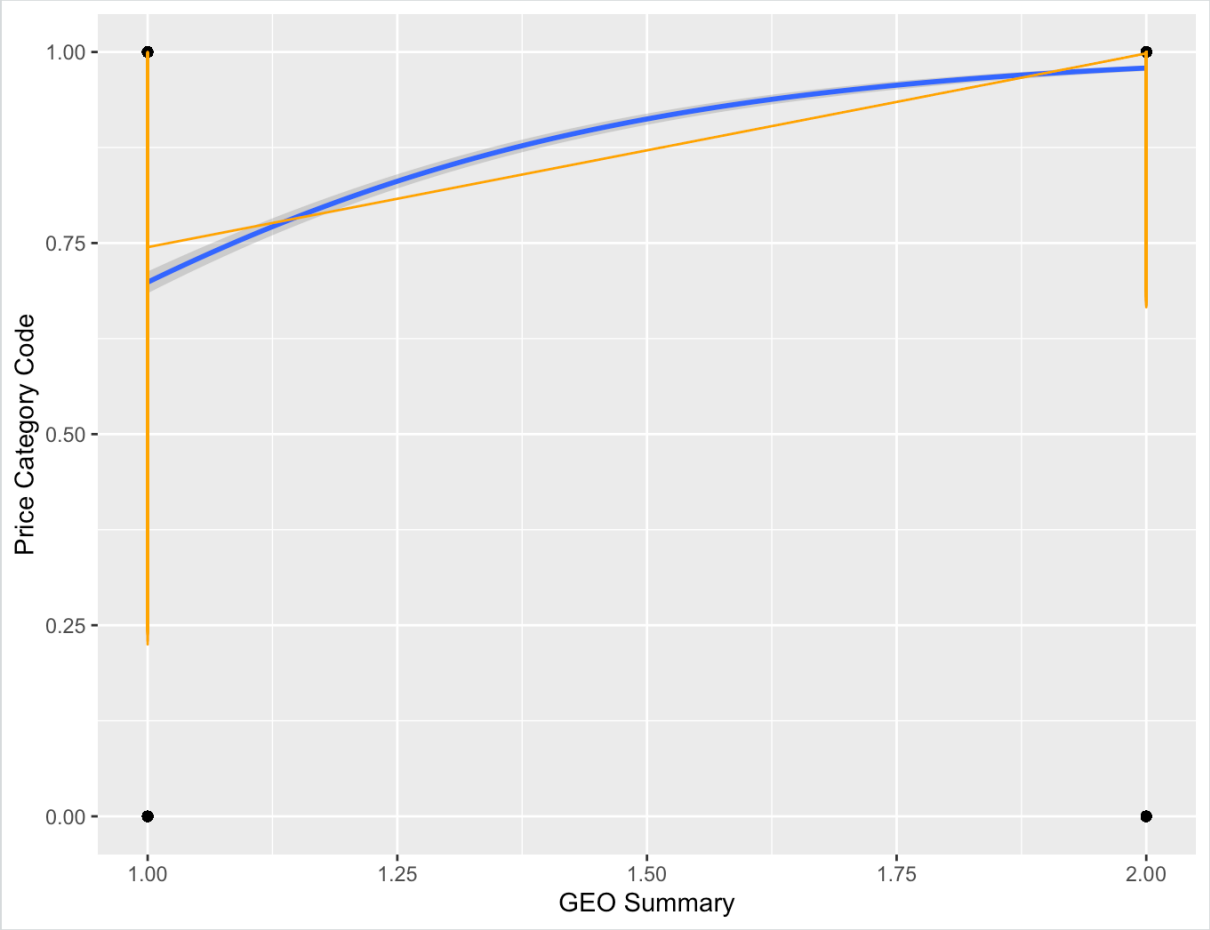
The spread of data points and the red vertical lines suggest variability in the price category code across different airlines, with some airlines having a higher probability of being in the higher price category. And the blue line represents the logistic regression model's probability curve, suggesting the likelihood of data points belonging to a certain category based on the 'Operating Airline'. The position of it is close to the y-axis value of 1.00, indicates that the logistic regression model predicts a high probability for the higher price category across most airlines.

Published Airline



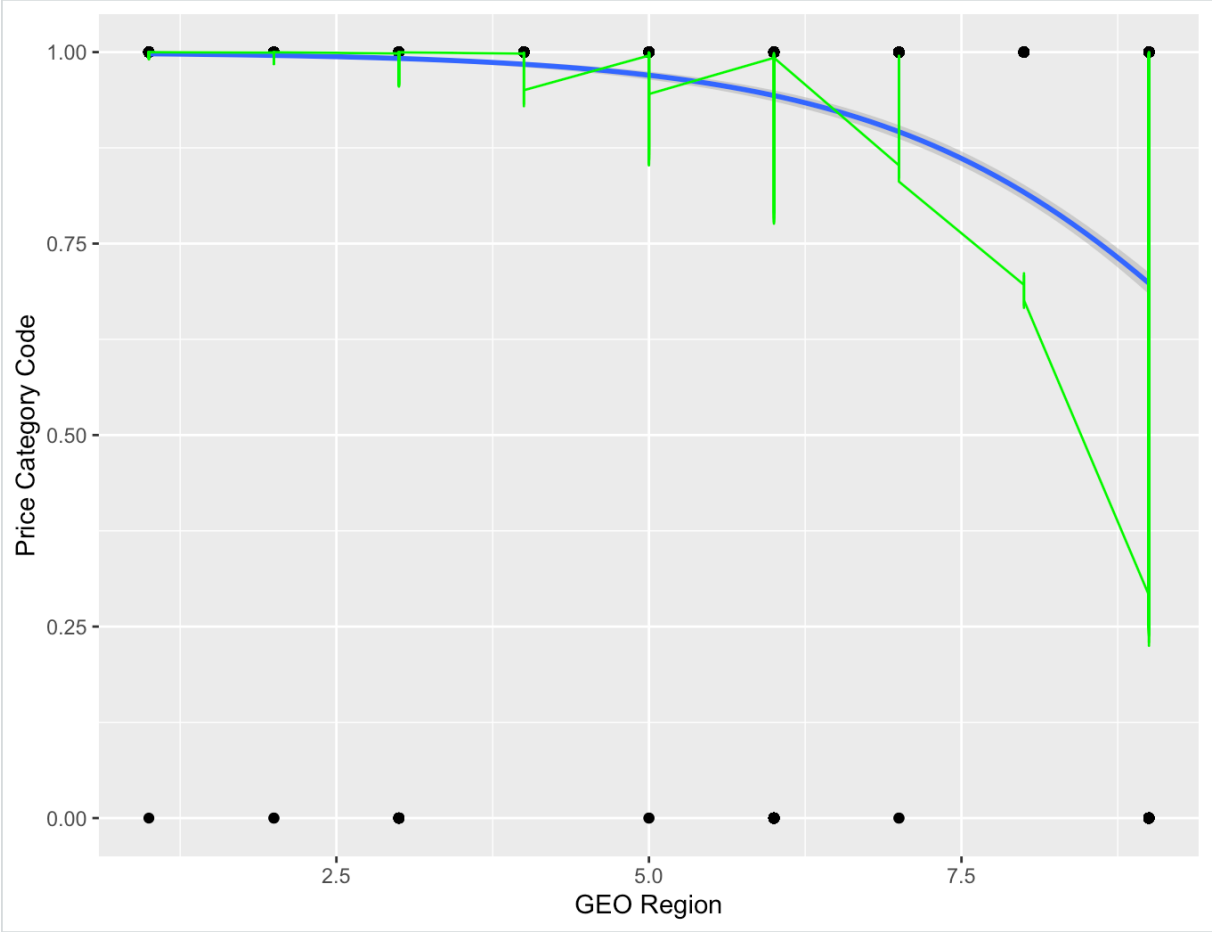
In this graph, the logistic regression line (blue) indicates the probability of each airline being in a specific price category, with the yellow lines representing the residuals or the difference between the observed and predicted values. The data points are scattered, with most lying at the extremes of the y-axis which expresses most of the published airlines don't have the low-fare ticket price.

GEO Summary



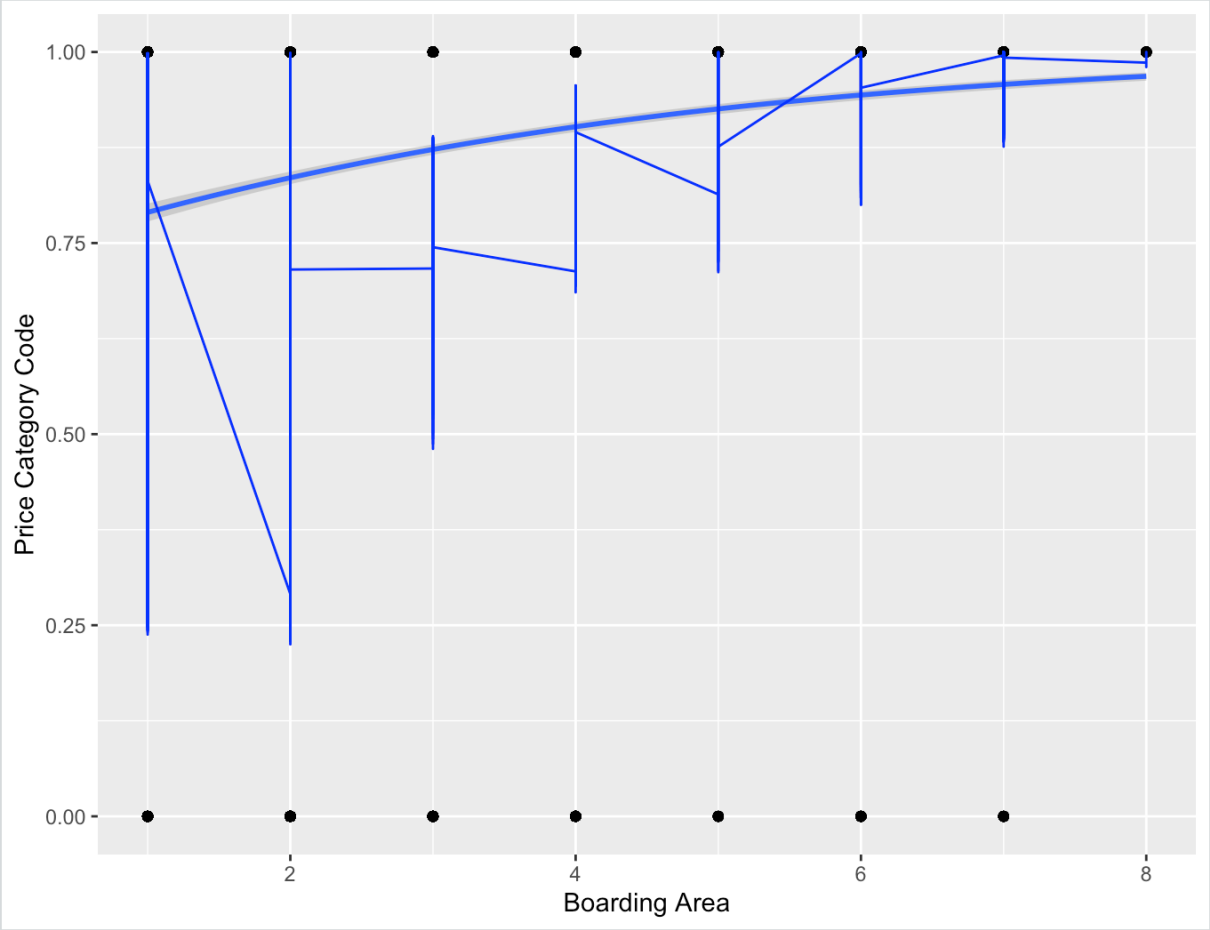
This graph is a visual representation of the logistic regression analysis performed using R, showcasing the model's ability to classify data points into two categories based on the 'GEO Summary' and 'Price Category Code'. The logistic regression line (blue) indicates the probability of each airline being in a specific price category, with the yellow lines representing the residuals or the difference between the observed and predicted values. As presented, most International flights weren't low fare while Domestic flights were widely spread from low fares to others.

GEO Region



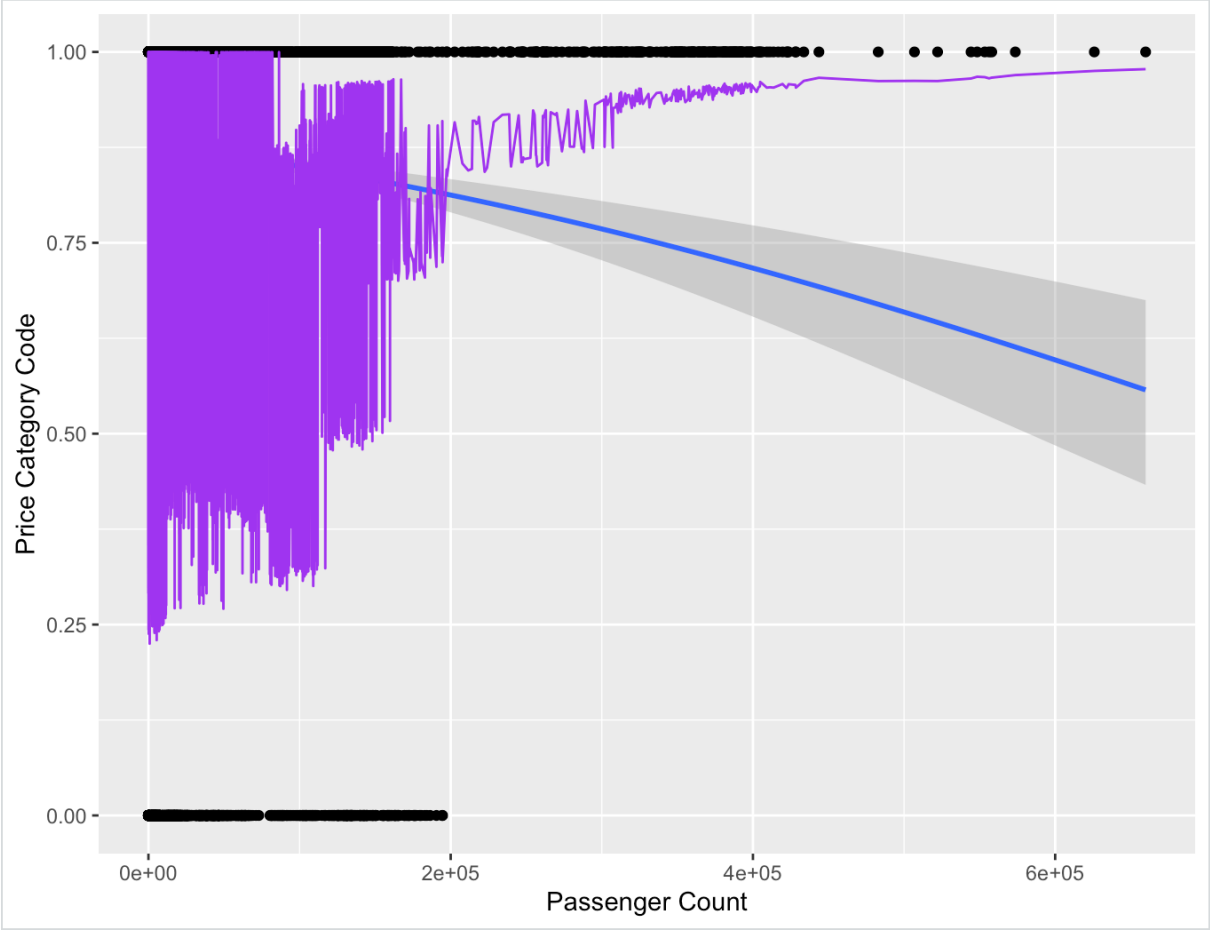
This analysis provides insights into the relationship between geographical regions and price categories, as modeled by logistic regression. Green points scattered along the curve likely represent the region of the flight. Error bars extending from some points suggest variability of flight in that region in which the US is having the most flight. A blue line depicts the logistic regression curve, showing the relationship between the GEO Region and the probability of falling into a particular price category. From the regression line, we can conclude that the more closer the flight's region from the US, the higher chance of getting the low fares flight.

Boarding Area



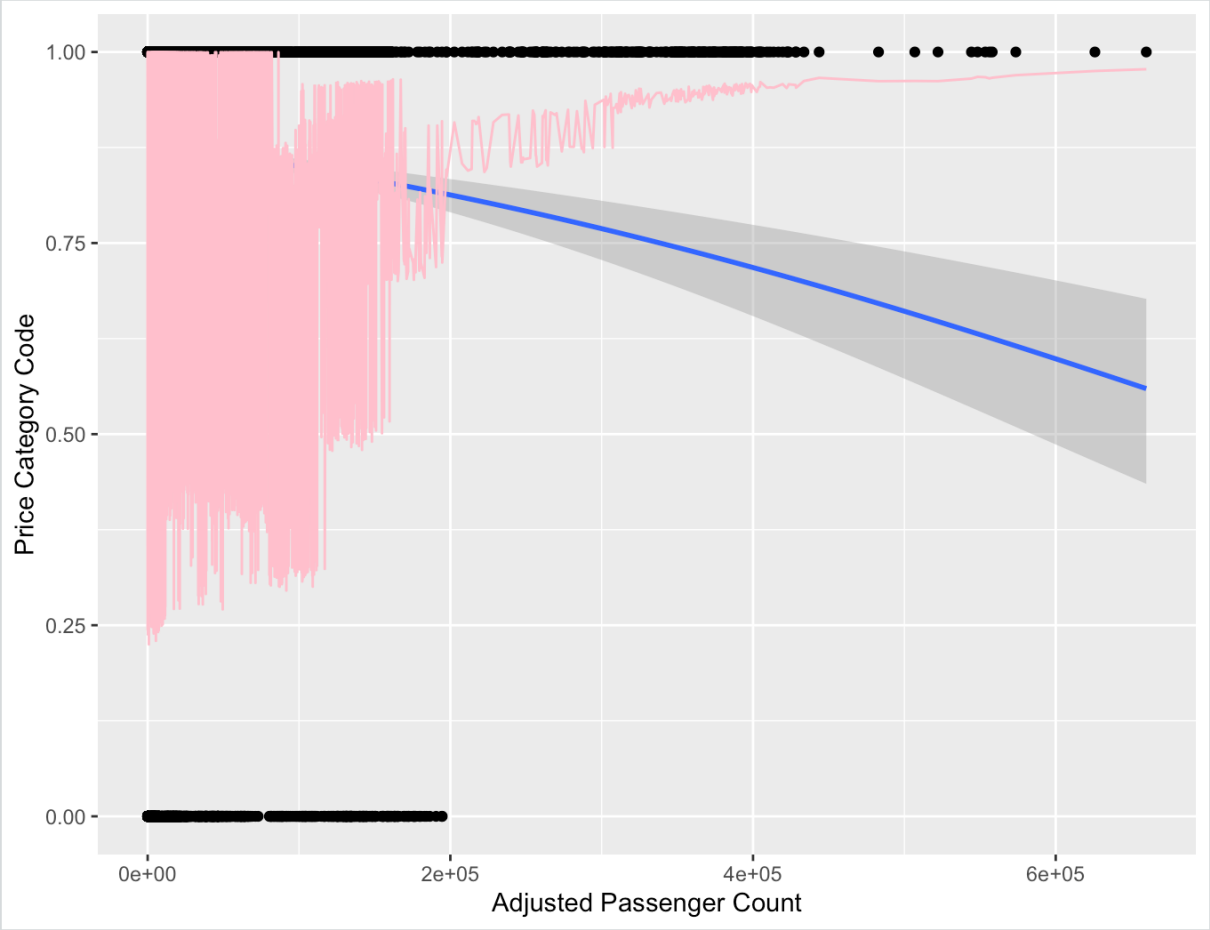
The plot displays a logistic regression analysis examining the relationship between ‘Boarding Area’ (ranging from 1 to 8 which are from A to G to "other") and ‘Price Category Code’. Blue points are connected by lines indicate the predicted probabilities of different ‘Price Category Codes’ for each ‘Boarding Area’. There is a noticeable trend where the probability of a higher chance of getting High Fare in ‘Price Category Code’ as the ‘Boarding Area’ number goes up from A to G to "other" simultaneously.

Passenger Count



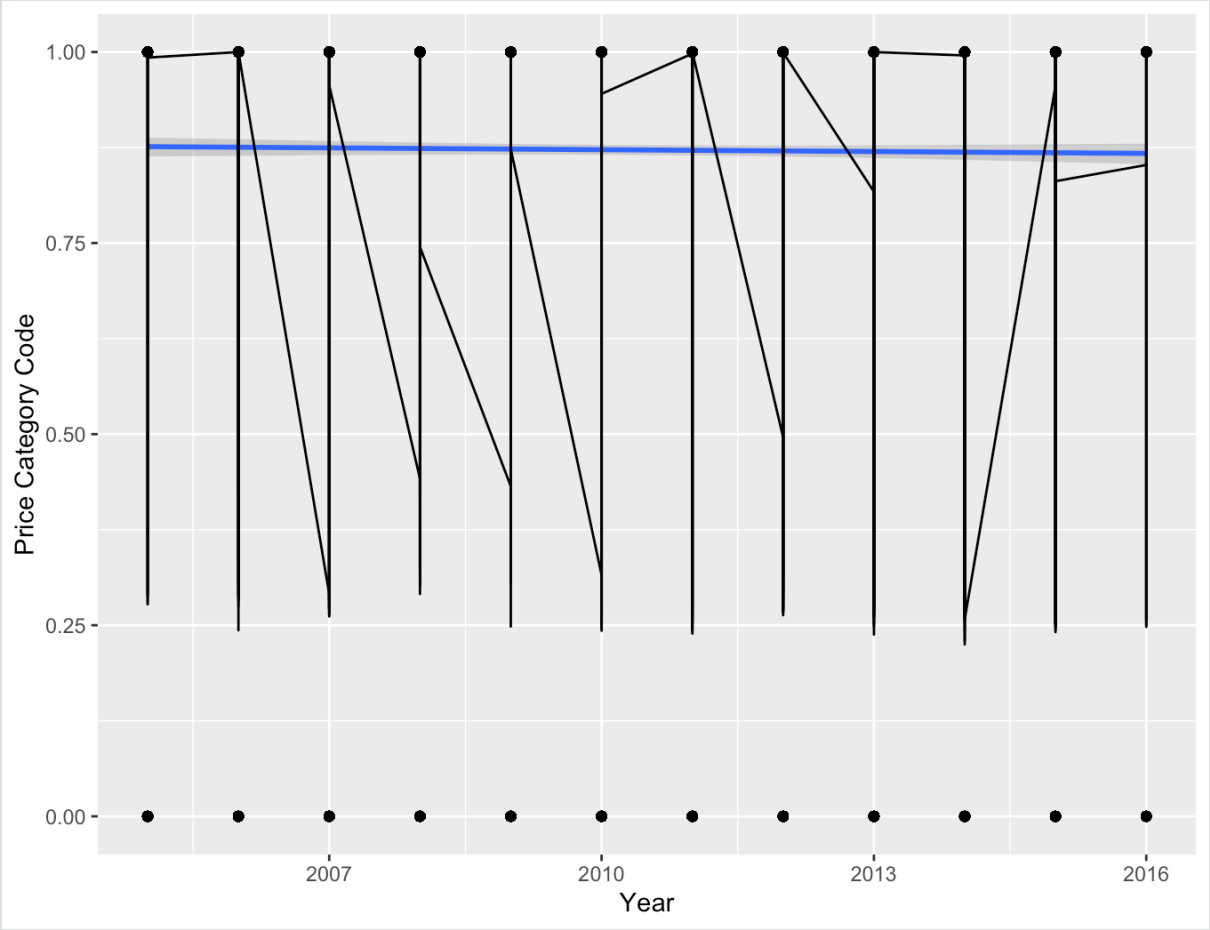
The scatter plot shows a dense cluster of points at the lower end of the “Passenger Count” axis, which are colored in purple. This suggests that there are more flight with a lower passenger count. The shaded grey area around the prediction line represents the confidence interval, providing a range within which the true logistic regression line is likely to fall. As the increases of number of passenger on the flight, the logistic curve descends towards the bottom right corner, suggesting a decrease in the probability of "Other" in “Price Category Code”.

Adjusted Passenger Count



In this graph, the 'Adjusted Passenger Count' is used as the predictor variable, ranging from 0 to approximately 600,000. A logistic regression line is fitted through the data points, indicating the probability of two 'Price Category Code' given the 'Adjusted Passenger Count'. The shaded areas around the regression line represent the confidence intervals, with the pink area indicating passenger counts, which are higher variance in a lower number of 'Adjusted Passenger Count'. The inverse relationship suggests that as the 'Adjusted Passenger Count' increases, the probability of being in a higher price category decreases. This implies that higher passenger counts are associated with lower price categories, due to budget factors.

Year



The horizontal axis shows a time span from 2005 to 2016, indicating the years over which the data was collected and analyzed. Also, the y-axis is labeled “Price Category Code,” suggesting that the binary outcome could be related to a categorization of prices, perhaps indicating whether a price is "Other" (1) or "Low Fare" (0) a certain threshold. We can see that the variance in the number of flights in 10 different years fluctuated around 0.25 or higher, and stayed stable throughout the timeline. This trend made the blue prediction line almost like a horizontal straight line across the graph.

VII. Data source and code source

This original database contains information on air traffic passenger statistics by the airline on Kaggle . It includes information on the airlines, airports, and regions that the flights departed from and arrived at. It also includes information on the type of activity, price category, terminal, boarding area, and number of passengers.

We downloaded the data set "*Air_Traffic_Passenger_Statistics.csv*" from the page and transformed to be the last chart in the part of preprocessing data with the name "*HK232_CC01_Group13_AirTraffic.csv*". File CSV of the chart and the source code R would be subbmitted by official way, but in need of convenience, we additionally provide a GitHub link to all of our data.

VIII. References

1. Josh Starmer, (05/03/2018), *StatQuest: Logistic Regression*, Link: https://www.youtube.com/watch?v=yIYKR4sgzI8&ab_channel=StatQuestwithJoshStarmer
2. The Devastator, (2022), Airlines Traffic Passenger Statistics, Link: https://www.kaggle.com/datasets/thedevastator/airlines-traffic-passenger-statistics
3. Wikipedia, (15/03/2024), *Multinomial Logistic Regression*, Link: https://en.wikipedia.org/wiki/Multinomial_logistic_regression
4. Wikipedia, (20/04/2024), *Logistic regression*, Link: https://en.wikipedia.org/wiki/Logistic_regression
5. Zach Bobbitt, (28/10/2020), *How to Perform Logistic Regression in R (Step-by-Step)*, Link: https://www.statology.org/logistic-regression-in-google-sheets
6. Zach Bobbitt, (04/04/2023), *How to Use predict with Logistic Regression Model in R*, Link: https://www.statology.org/r-logistic-regression-odds-ratio