# Exam PA April 2024 Project Statement

## IMPORTANT NOTICE – THIS IS THE APRIL 16, 2024, PROJECT STATEMENT. IF TODAY IS NOT APRIL 16, 2024, SEE YOUR TEST CENTER ADMINISTRATOR IMMEDIATELY.

### General Information for Candidates

This examination has 12 tasks numbered 1 through 12 with a total of 70 points. The points for each task are indicated at the beginning of the task, and the points for subtasks are shown with each subtask.

Each task pertains to the business problem described below. Additional information on the business problem may be included in specific tasks—where additional information is provided, including variations in the target variable, it applies only to that task and not to other tasks. This exam includes an Excel data file with information for Task 9(e). You may use Excel for calculation for this or any of the other tasks, but all answers must be submitted in the Word document. *If you upload the Excel document, it will not be looked at by the graders.* Neither R nor RStudio are available.

The responses to each specific subtask should be written after the subtask and the answer label, which is typically ANSWER, in this Word document. Each subtask will be graded individually, so be sure any work that addresses a given subtask is done in the space provided for that subtask. Some subtasks have multiple labels for answers where multiple items are asked for—each answer label should have an answer after it.

Each task will be graded on the quality of your thought process (as documented in your submission), conclusions, and quality of the presentation. The answer should be confined to the question as set. No response to any task needs to be written as a formal report. Unless a subtask specifies otherwise, the audience for the responses is the examination grading team and technical language can be used.

Prior to uploading your Word file, it should be saved and renamed with your five-digit candidate number in the file name. If any part of your exam was answered in French, also include "French" in the file name. Please keep the exam date as part of the file name.

The Word file that contains your answers must be uploaded before the five-minute upload period time expires.

## Business Problem

*You have recently joined the analytics team of a consultancy that serves the U.S. aviation industry. You serve a number of clients that are interested in developing strategies that are responsive to changing demand patterns and competitive pressures in the market.*

*Your team will use data from the US Department of Transportation's Domestic Airline Consumer Airfare Report,[1] focusing on information on the Top 1,000 Contiguous State City-Pair Markets. This dataset will provide insights into market dynamics, fare trends, and consumer behavior patterns.*

Notes on the data set:

The city data is structured so that one record represents the data for flights in both directions between two identified cities and the designation between **city1** and **city2** is arbitrary. For example,

- For flights that go between New York City and Phoenix, New York City will always be designated as **city1**
- For flights that go between New York City and Los Angeles, New York City will always be designated as **city2**

| year | quarter | city1 | city2 | passengers | Fare |
|---|---|---|---|---|---|
| 2021 | 1 | New York City, NY | Phoenix, AZ | 1,019 | $207.99 |
| 2021 | 1 | Los Angeles, CA | New York City, NY | 3,157 | $233.00 |

- Data is at the city level, not the airport level, so for cities served by multiple airports, the data consolidates records for all airports within that city.

- Summary statistics in the data dictionary are all aggregated at the per city pair per quarter, unless otherwise specified.

[1] *Source: United States Department of Transportation*

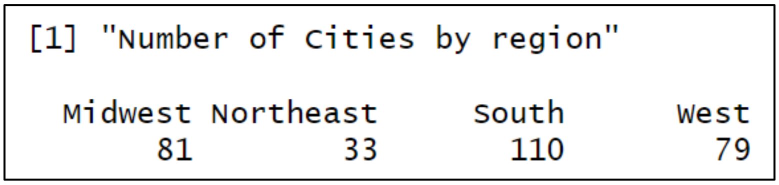
## Data Dictionary

| Variable | Data Type / Range / Example | Unique Values | Description |
|---|---|---|---|
| year | Numeric: 2016 to 2022 | 7 | Year of the flight departure |
| quarter | Numeric: 1 to 4 | 4 | Calendar quarter of the flight departure |
| citymarketid_1 | Numeric: 30135 to 35550 | 266 | 5-digit identification number assigned by the US Department of Transportation to identify a city market for city1 |
| citymarketid_2 | Numeric: 30140 to 36133 | 284 | 5-digit identification number assigned by the US Department of Transportation to identify a city market for city2 |
| latitude_1 | Numeric: 25.90 to 70.13 | 233 | Latitude of city1 |
| longitude_1 | Numeric: -160.97 to -68.77 | 233 | Longitude of city1 |
| latitude_2 | Numeric: 25.90 to 70.13 | 247 | Latitude of city2 |
| longitude_2 | Numeric: -160.97 to -68.02 | 247 | Longitude of city2 |
| city1 | Character: “Charlotte, NC” | 266 | City and State name used to consolidate airports serving the same city market for city1 |
| city2 | Character: “Salt Lake City, UT” | 284 | City and State name used to consolidate airports serving the same city market for city2 |
| state_1 | Character: “NC” | 47 | Associated state in which city1 resides |
| region_1 | Character: “South” | 4 | Associated region in which city1 resides |
| state_2 | Character: “UT” | 48 | Associated state in which city2 resides |

|  |  |  |  |
|---|---|---|---|
| region_2 | Character: “West” | 4 | Associated region in which city2 resides |
| nsmiles | Numeric:  67 to 2,783 | 2,256 | Non-Stop market miles between the cities (using radian measure) |
| passengers | Numeric:  10 to 24,734 | 20,990 | Average passengers per day |
| fare | Numeric:  $61.77 to $730.71 | 154,421 | Average fare in US dollars |
| carrier_lg | Character: “DL” | 10 | Abbreviation for the airline with the largest market share |
| large_ms | Numeric:  18.42% to 100% | 79,080 | Market share for the city pair for the carrier with the largest market share |
| fare_lg | Numeric:  $60.36 to $769.80 | 153,727 | Average fare for the carrier with the largest market share |
| carrier_low | Character: “AA” | 16 | Abbreviation for the airline with the lowest fare |
| lf_ms | Numeric:  1.00% to 100% | 84,324 | Market share for the carrier with the lowest average fare |
| fare_low | Numeric:  $54.00 to $696.20 | 149,328 | Average fare for the carrier with the lowest average fare |
| connections_1 | Numeric:  1 to 236 | 122 | Number of flight connections that city1 has **across the entire dataset** |
| connections_2 | Numeric:  1 to 236 | 123 | Number of flight connections that city2 has **across the entire dataset** |

## Task 1 (8 *points*)

Your client is interested in being able to predict the number of passengers who will travel between two cities. Your manager believes that the geographic location variables may be important, especially the region variable. The cities are already labeled with a region variable:



```
[1] "Number of Cities by region"

  Midwest Northeast     South      West 
       81        33       110        79 
```

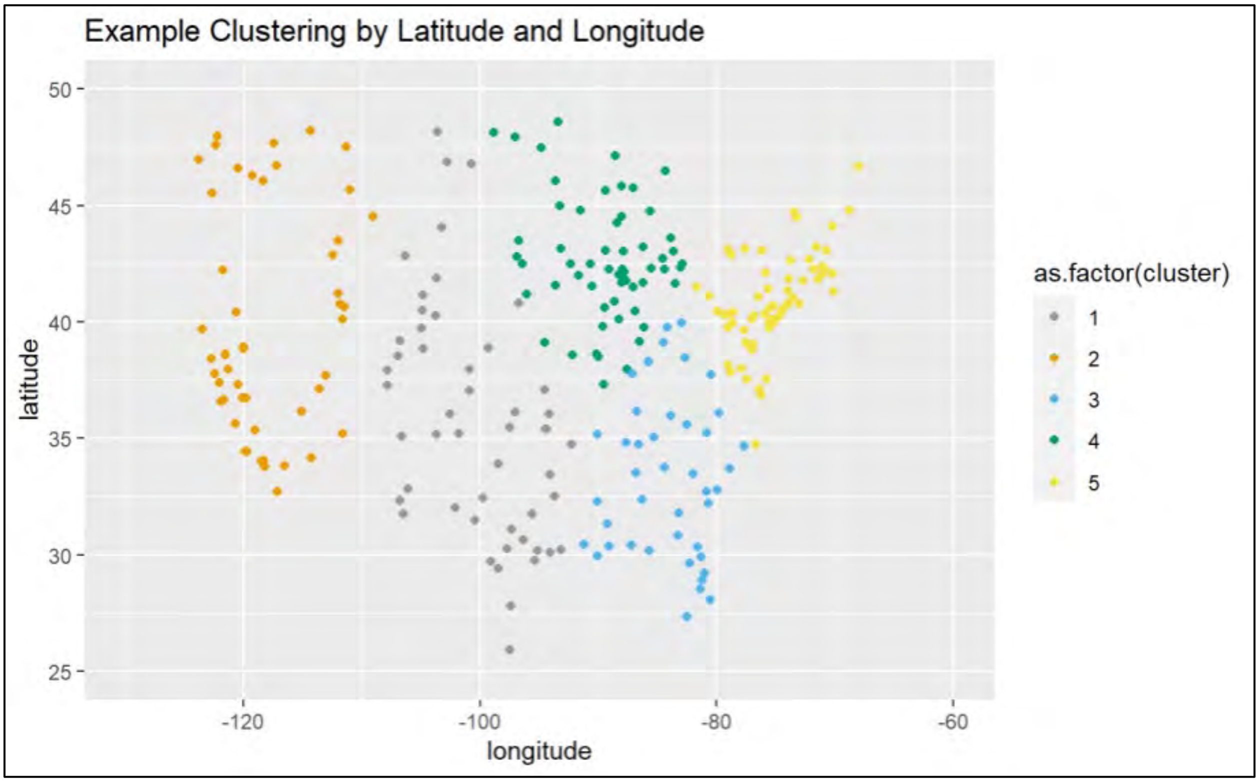
However, your assistant suggests clustering other geographic location variables to create new data-driven groupings of the cities. The clustering would include the longitude and latitude of each city and potentially other variables.

(a) (*2 points*) Describe one advantage and one drawback of creating clusters based upon the data compared to using predefined regions.

**ANSWER:**

---

Your assistant has produced the following graph showing the output of a *K*-means clustering algorithm using only latitude and longitude.

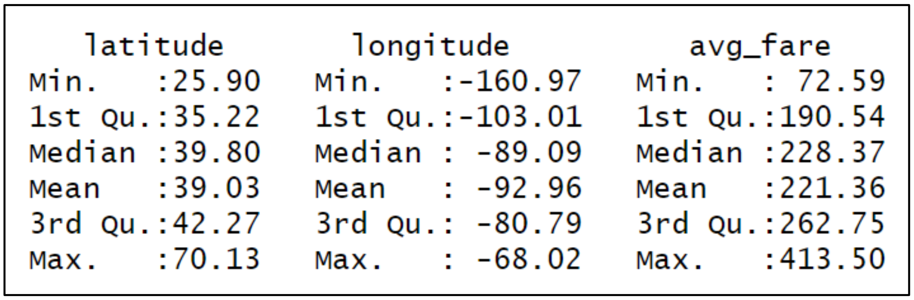


Your assistant suggests it might be useful to compare how homogenous each cluster is.

(b) (*2 points*) Describe the steps to calculate the within cluster sum of squares using latitude and longitude.

**ANSWER:**

Your assistant creates a new variable called **avg_fare**, which represents the average fare for a city across all routes and time periods for that city. Your assistant wants to include the new variable as one of the variables in the *K*-means clustering.

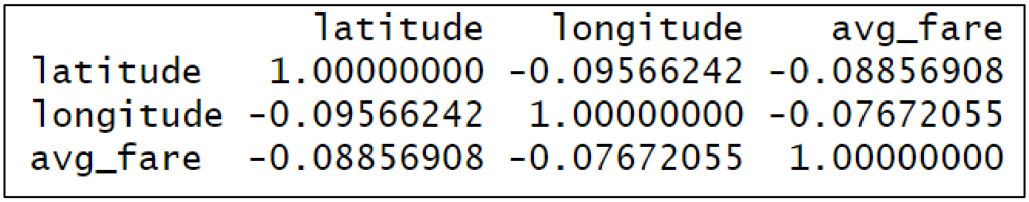


```
   latitude        longitude          avg_fare
Min.   :25.90   Min.   :-160.97   Min.   : 72.59
1st Qu.:35.22   1st Qu.:-103.01   1st Qu.:190.54
Median :39.80   Median : -89.09   Median :228.37
Mean   :39.03   Mean   : -92.96   Mean   :221.36
3rd Qu.:42.27   3rd Qu.: -80.79   3rd Qu.:262.75
Max.   :70.13   Max.   : -68.02   Max.   :413.50
```

(c) (*1 point*) Identify an important modification to the variables **latitude**, **longitude**, and **avg_fare** before beginning *K*-means clustering.

**ANSWER:**

Your assistant provides the following correlation matrix for the features being used in the clustering:



```
             latitude   longitude    avg_fare
latitude   1.00000000 -0.09566242 -0.08856908
longitude -0.09566242  1.00000000 -0.07672055
avg_fare  -0.08856908 -0.07672055  1.00000000
```

(d) (*2 points*) Evaluate the value of clustering principal components derived from these variables as opposed to clustering the untransformed variables, based on the results of the correlation matrix.

**ANSWER:**

Your assistant suggests that the resulting clusters based on the three variables are to be used as a response variable in a subsequent model that will be used to predict the average fare between cities.

(e) (*1 point*) Evaluate your assistant's approach.

**ANSWER:**

## Task 2 (6 *points*)

You are working with a client that supplies beverages to airlines for flights flying into and out of the state of Washington. Your client explains that on flights shorter than 300 miles, no beverages are provided, for flights over 300 miles, beverages are provided. The client is interested in expanding their sales and wants to better understand the size of the total market for their products in the state.

(a) (*3 points*) Develop an approach to estimating the total market size in Washington state for your client. You should state how you would measure market size and describe the analysis and variables you would use based on the available data.

**ANSWER:**

______________________________________________________________________________

Your manager is looking for new clients to offer analytics consulting to using the Domestic Airline Consumer Airfare Report. They are interested in exploring what other questions can be answered with the data set.

(b) (*3 points*) Explain whether the data is sufficient to answer each of the use cases below. If the data is not sufficient suggest additional data or changes to the data you would need to answer the question.

i. Explain how market concentration affects fare prices.
ii. Predict flight cancellation rates based on temperature and precipitation.
iii. Recommend when flying out of a different nearby airport may save money on flight costs.

**ANSWER:**

i.

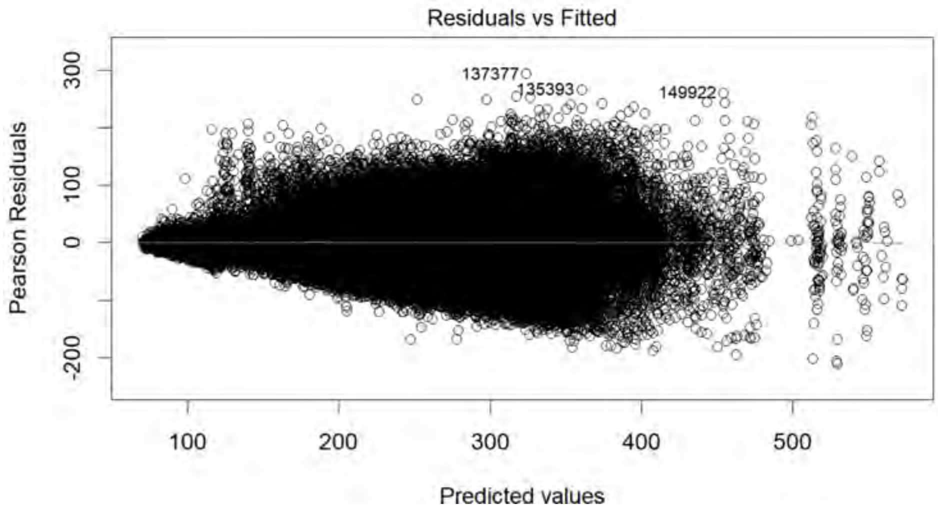
ii.

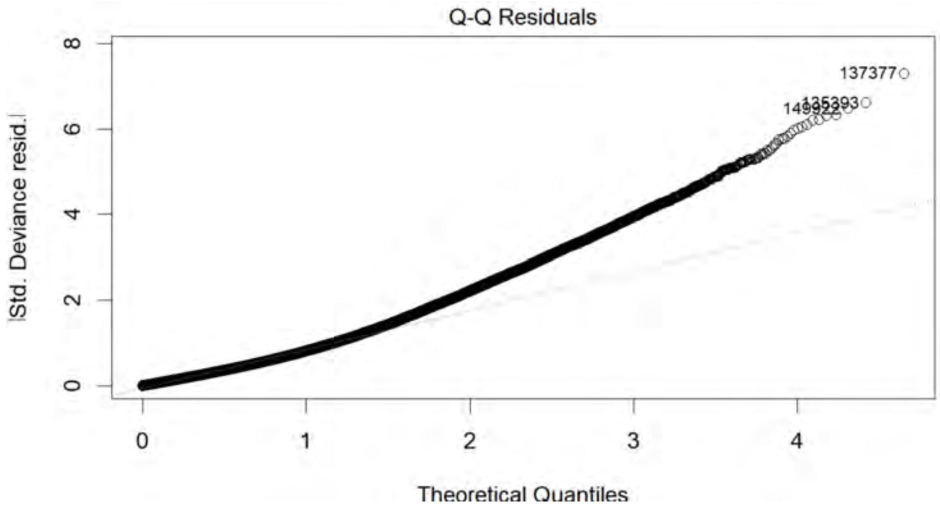
iii.

## Task 3 (7 *points*)

You are guiding your assistant to additional transformations that may help fit a generalized linear model to predict the **fare** for any city pair within any particular **quarter**.

When fitting a linear model to predict **fare**, the following diagnostics are produced:



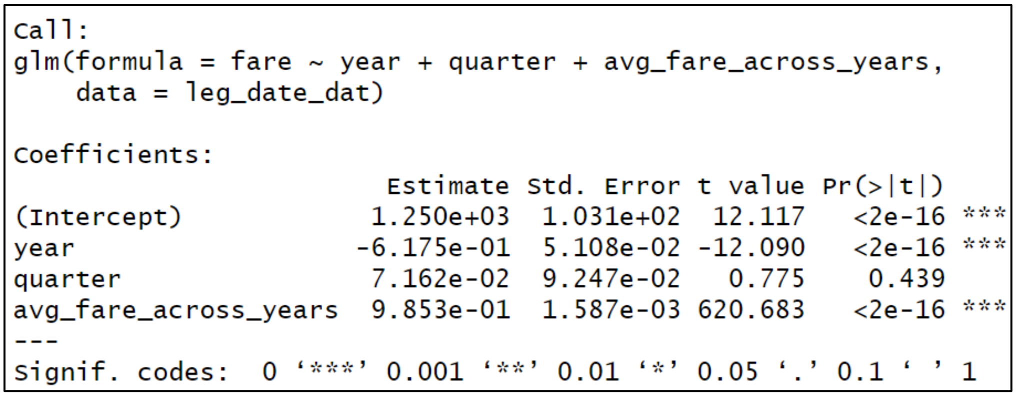


(a) (*3 points*) Evaluate the appropriateness of the linear model based on these plots and recommend a more appropriate GLM distribution and link function.

**ANSWER:**

---

Your assistant creates a new variable, avg_fare_across_years, which reflects the average fare for a route across all time periods. Including the numeric variables **year** and **quarter,** and the new avg_fare_across_years variable in an ordinary linear regression to predict fare, yields the following result:



```
Call:
glm(formula = fare ~ year + quarter + avg_fare_across_years,
    data = leg_date_dat)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.250e+03  1.031e+02  12.117   <2e-16 ***
year                  -6.175e-01  5.108e-02 -12.090   <2e-16 ***
quarter                7.162e-02  9.247e-02   0.775    0.439
avg_fare_across_years  9.853e-01  1.587e-03 620.683   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b) *(4 points)*

i. (*2 points*) Evaluate using **year** as a categorical variable vs. numeric variable.

ii. (*2 points*) Evaluate using **quarter** as a categorical variable vs. numeric variable.

**ANSWER:**

i.

ii.

## Task 4 (5 *points*)

Your manager is interested in the relationship between the market share held by the airline with the highest market share between two cities (**large_ms**) and the average fare between those cities (**fare**). Your assistant produces the graphs below.

**Exhibit A**

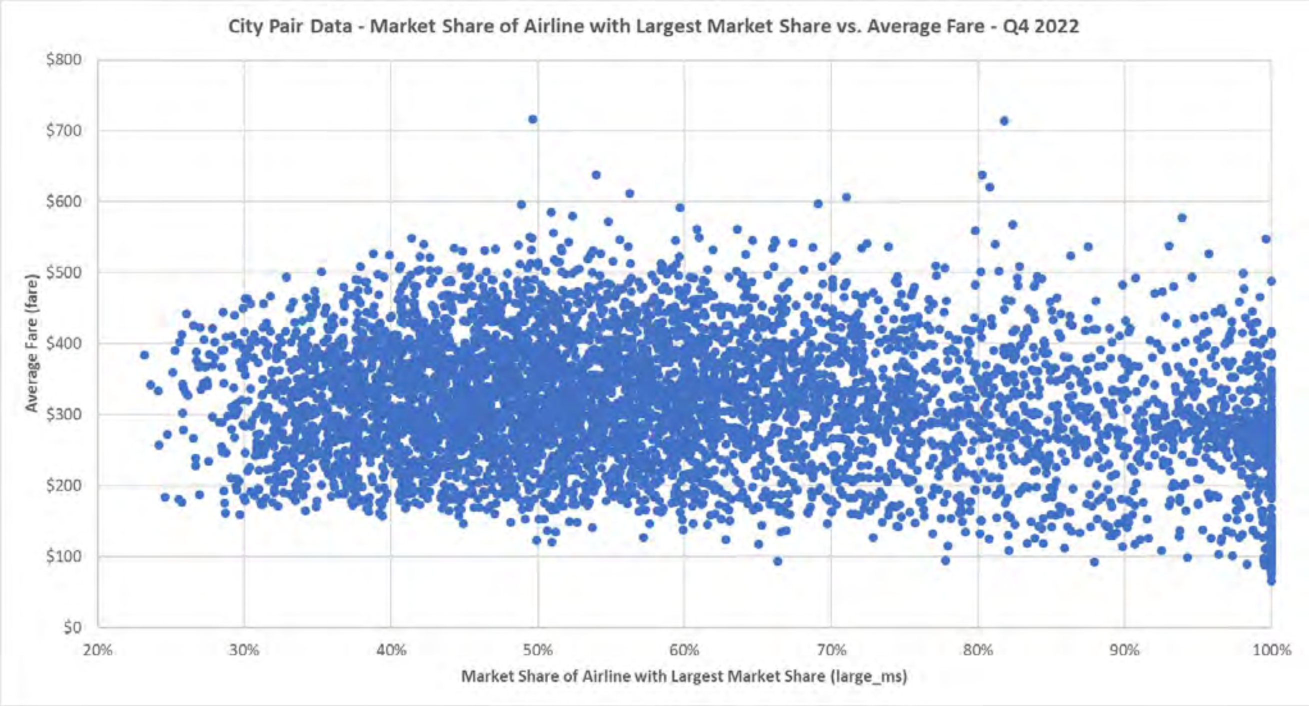
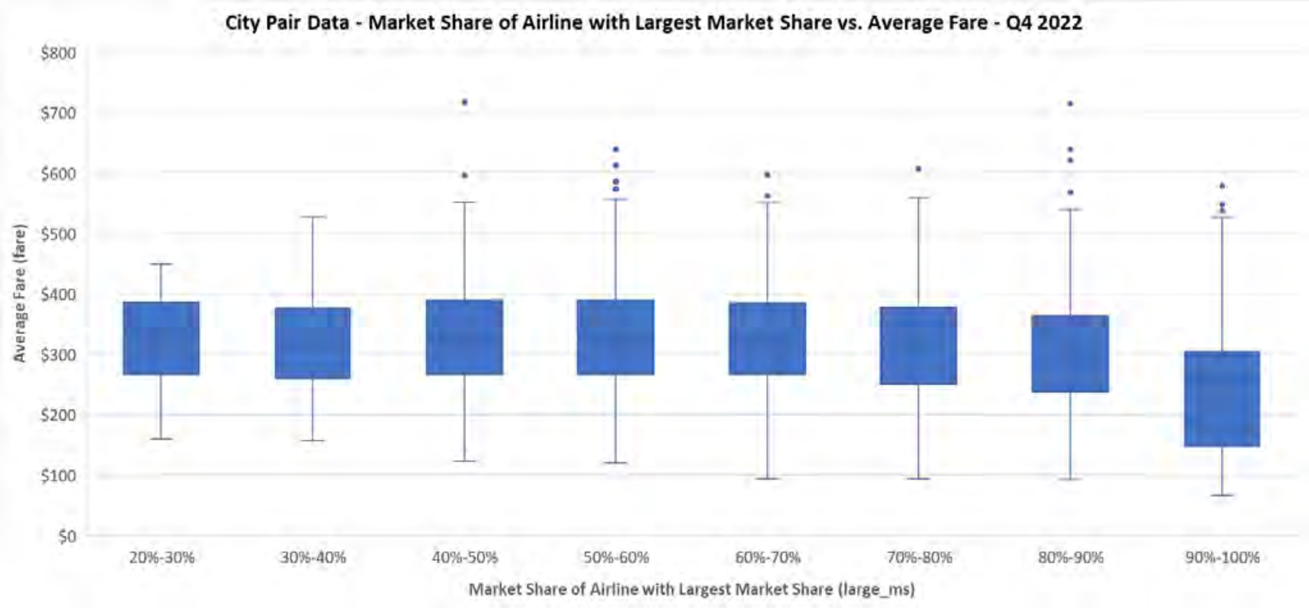


**Exhibit B**



(a) (*3 points*) Describe one pro and one con of each visualization in explaining the relationship between **large_ms** and **fare**.

**ANSWER:**

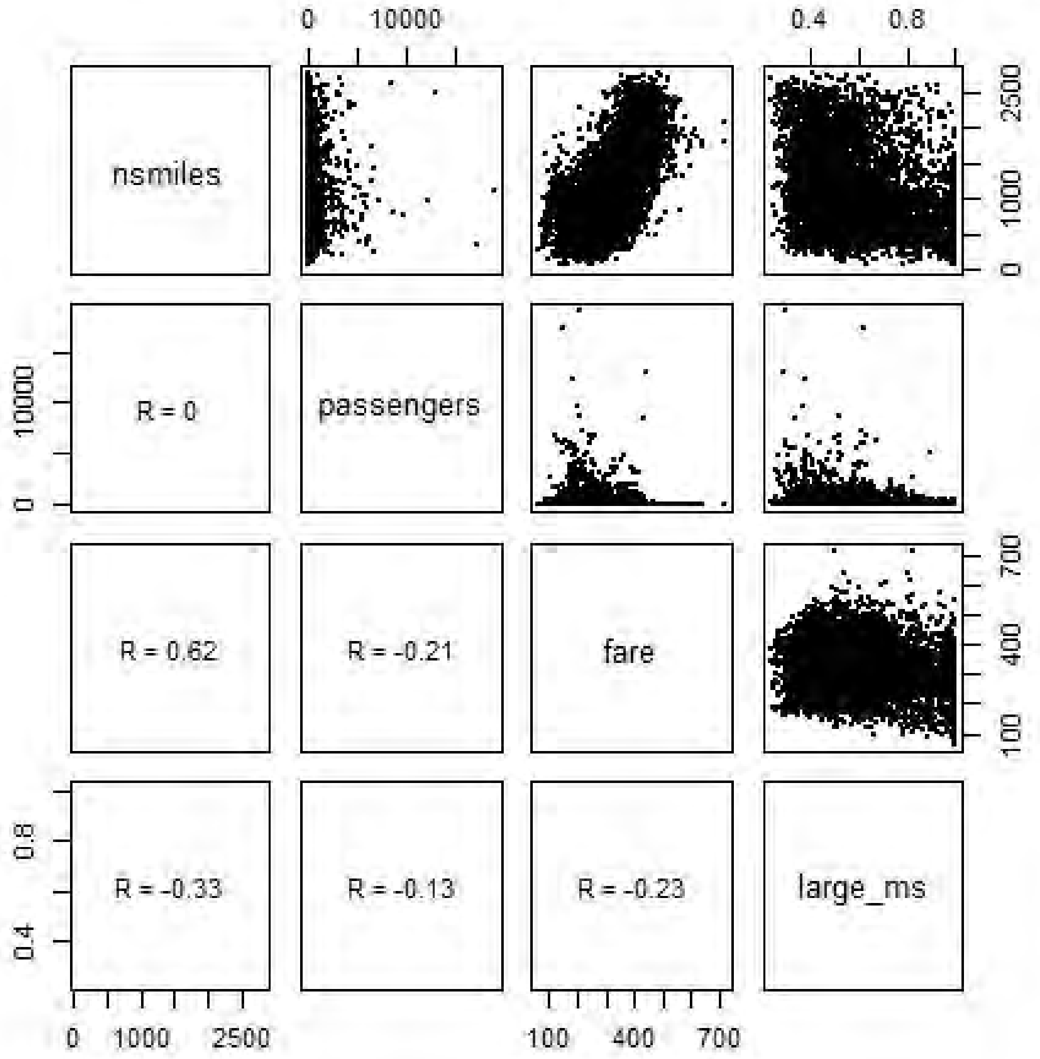
**Exhibit A – Pro:**

**Exhibit A – Con:**

**Exhibit B – Pro:**

**Exhibit B – Con:**

Based on the graphs produced in part (a), your assistant concludes that **fare** is negatively correlated with **large_ms**. You recommend your assistant look at comparisons between **fare**, **large_ms** and some of the other variables in the dataset, such as the number of non-stop miles (**nsmiles**) and the number of passengers (**passengers**). Your assistant produces the graphic below showing the scatterplots and correlations between all four variables (each scatterplot and correlation is matched to the corresponding set of variables in the row and column of the graphic).

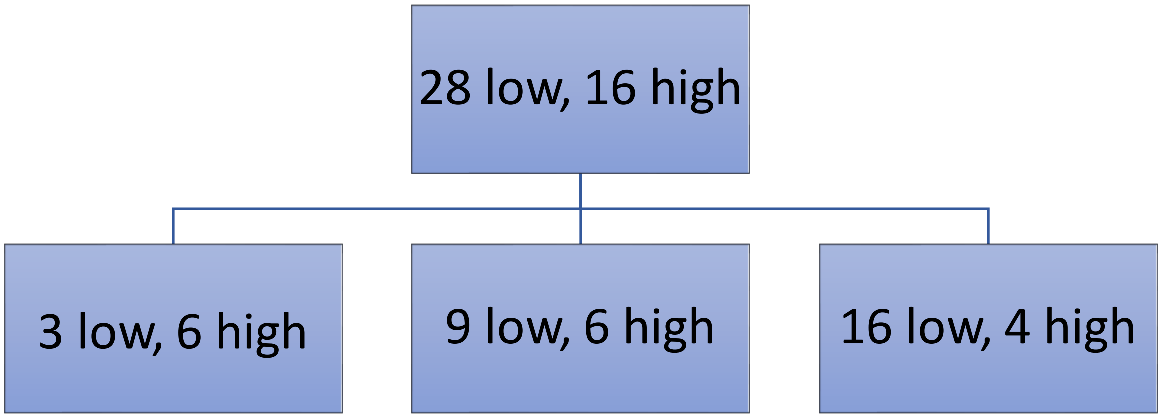


(b) (*2 points*) Recommend a visualization for your assistant to create to understand the modeling implications of the graphic above.

**ANSWER:**

## Task 5 (6 *points*)

Your assistant builds a decision tree to estimate whether a city pair has an average of more than 500 passengers per day (low_traffic vs. high_traffic). Below a root node, there is a single split based on a factor variable with three levels.



i. Node 1 – 3 low_traffic city pair, 6 high_traffic city pair
ii. Node 2 – 9 low_traffic city pair, 6 high_traffic city pair
iii. Node 3 – 16 low_traffic city pair, 4 high_traffic city pair

The formula for entropy is: Entropy (N) = $-\sum_{i=1}^{c} p_i log_2(p_1)$

(a) (*3 points*) Determine the information gain of this split using the entropy measure.

**ANSWER:**

---

Instead of building a tree to estimate whether a city pair was low traffic or high traffic, your assistant decided to build a tree to estimate the average passengers per day.

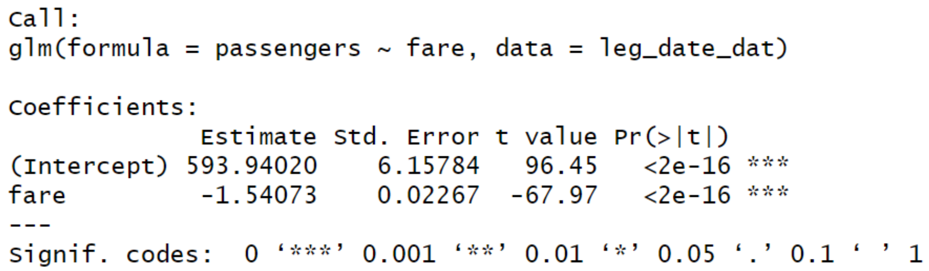
(b) (*3 points*) Explain how the calculation of impurity would differ from part (a).

**ANSWER:**

## Task 6 (3 *points*)

Your client is interested in optimizing revenue and wants to determine the extent to which the additional revenue per passenger generated by a higher fare would be offset by a reduction in the number of passengers. Your manager suggests constructing a model that predicts the **passengers** per day flying between a city pair given a particular **fare** charged.

See the results of a very basic regression of **passengers** on **fare**:



```
Call:
glm(formula = passengers ~ fare, data = leg_date_dat)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 593.94020    6.15784   96.45   <2e-16 ***
fare         -1.54073    0.02267  -67.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) (*1 point*) Interpret the intercept and the coefficient for **fare**.

**ANSWER:**

---

Your assistant fits two different regression models showing you the following fit results from the training data:

| Model | AIC |
|---|---|
| Model 1 | 2,091,364 |
| Model 2 | 2,098,969 |

(b) (*1 point*) Recommend and justify which model is better based upon the output above.

**ANSWER:**

---

(c) (*1 point*) Critique the use of a training-data-based metric such as AIC versus a metric calculated on validation data.

**ANSWER:**

## Task 7 (4 *points*)

One of the features you have available for estimating expected **fare**, **city_cluster**, is the result of a *K*-means clustering exercise that was performed on the cities, with each city receiving a cluster number (such as 1, 2, 3) representing the cluster to which it is assigned.

(a) (*1 point*) Explain why treating the cluster number as a categorical feature may be more appropriate than treating it as a numeric one.

**ANSWER:**

---

(b) (*1 point*) Explain how binarization can be applied to city_cluster to allow it to be used in a regression.

**ANSWER:**

---

Upon examination, the number of clusters is large and you are worried that including the city_cluster in the model could lead to overfitting due to high dimensionality. Because of this, you ask your assistant to try a regularized regression.

(c) (*2 points*) Compare and contrast using ridge vs. LASSO regression to address the overfitting concern.

**ANSWER:**

## Task 8 (5 *points*)

Your assistant is interested in modeling the following three outcomes:

- the average fare
- the probability that the average daily passengers for a given city pair is greater than 500
- the number of connections for a given city

(a) (*3 points*) Recommend a link function and distribution your assistant could use in their generalized linear model to predict each of the three outcomes. Justify your answer.

i. the average fare
ii. the probability that the average daily passengers for a given city pair is greater than 500
iii. the number of connections for a given city

**ANSWER:**

i.

ii.

iii.

---

While predicting average fare your assistant wants to use **nsmiles** as a weighting variable or as a predictor variable.

(b) (*2 points*) Explain the difference between using **nsmiles** as a weighting variable as opposed to using it as a predictor variable.

**ANSWER:**

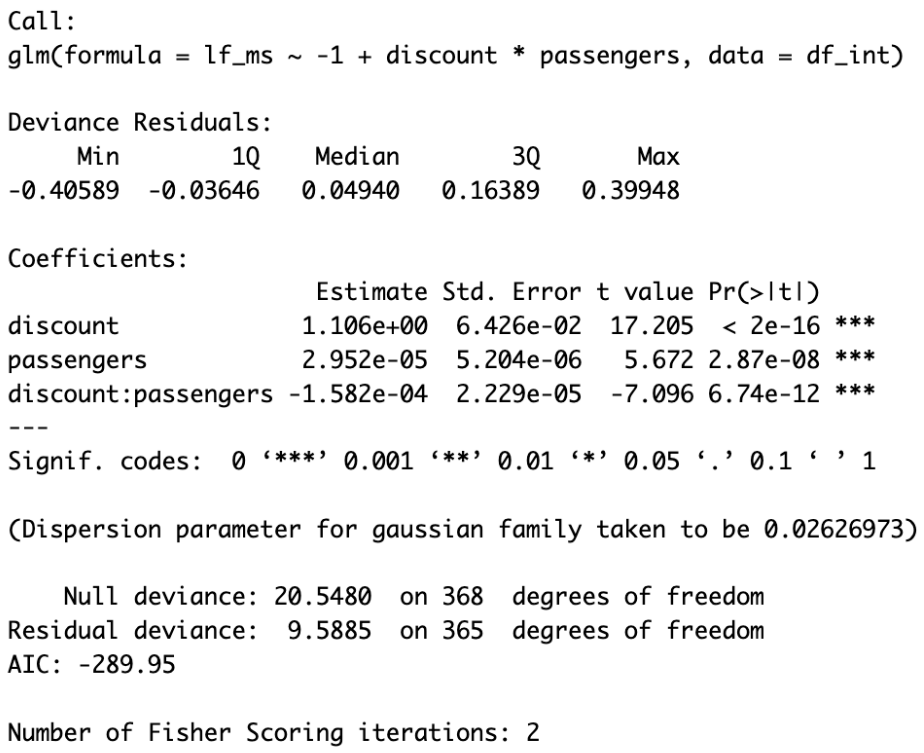
## Task 9 (12 *points*)

Your client runs a budget airline company and wants to start a new transcontinental flight route from New York City to one of the cities in these 5 states of the West region: CA, NV, WA, CO and AZ. Your client asks you to build a model to predict your potential market share if you offered a discount to the average fare for the new route, then forecast the potential revenue.

Your assistant used variable **lf_ms** as a proxy to potential market share, and created the variable **discount** as (1 - **fare_low** / **fare**). Your assistant also built a generalized linear model (Model 1) using **lf_ms** as response. Your assistant pointed out there is an interaction effect between variables **discount** and **passengers**.

You are provided with a model summary.

**Model 1:**



```
Call:
glm(formula = lf_ms ~ -1 + discount * passengers, data = df_int)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-0.40589  -0.03646   0.04940   0.16389   0.39948  

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
discount             1.106e+00  6.426e-02  17.205  < 2e-16 ***
passengers           2.952e-05  5.204e-06   5.672 2.87e-08 ***
discount:passengers -1.582e-04  2.229e-05  -7.096 6.74e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.02626973)

    Null deviance: 20.5480  on 368  degrees of freedom
Residual deviance:  9.5885  on 365  degrees of freedom
AIC: -289.95

Number of Fisher Scoring iterations: 2
```

(a) (*2 points*) Explain how variables **discount** and **passengers** interact in terms of the outcome.

**ANSWER:**

---

Your client is interested in predicting the annual revenue that they will receive for flights between different city pairs based on the discount they offer.

(b) (*2 points*) Assess the appropriateness of annual revenue as a Key Performance Indicator.

**ANSWER:**

---

You are fitting a GLM model by using LASSO regression.

(c) (*2 points*) Identify a hyperparameter that can be tuned and describe how you would tune it using cross validation.

**ANSWER:**

---

Your assistant has used k-fold cross validation for hyperparameter tuning while fitting a model. Your assistant then presents the mean and variance of the prediction error over the k validation data folds as support for assessing how the model will perform on new data.

(d) (2 points) Evaluate your assistant's approach for assessing how the model will perform on new data. Justify your answer.

**ANSWER:**

---

You are provided with the output of a LASSO regression model that predicts the market share for each potential destination to help determine which destination(s) should be considered for the new route.

Your client is willing to offer a 20% discount (discount = 0.2) from the average fare, and wants to choose the destination from the table below that maximizes the annual revenue based on 3 weekly one-way flights (3 flights per week) from New York City to the destination on a plane with 150 seats.

Your client also provided that projected revenue should be calculated using the formula below:

$$\begin{aligned} & projected\_revenue \\ & \qquad = \min(predicted\_market\_share \times passengers, maximum\_available\_seats) \\ & \qquad \times weekly\_flights \times 52 \times (1 - discount) \times fare \end{aligned}$$

| city | passengers | nsmiles | large_ms | fare | carrier_no |
|---|---|---|---|---|---|
| **Aspen, CO** | 127 | 1754 | 0.78 | 565 | 3 |
| **Fresno, CA** | 88 | 2484 | 0.43 | 349 | 6 |
| **Grand Junction, CO** | 26 | 1851 | 0.57 | 374 | 3 |
| **Montrose/Delta, CO** | 49 | 1820 | 0.77 | 432 | 3 |
| **Reno, NV** | 226 | 2443 | 0.28 | 370 | 5 |

***The coefficient table and destination inputs are provided in the Excel sheet. Complete the table below. If you upload the Excel document, it will not be looked at by the graders.***

(e) (*4 points*) Complete the table below and recommend a destination to your client.

**ANSWER:**

i. Complete the table.

| City | Aspen | Fresno | Grand Junction | Montrose | Reno |
|---|---|---|---|---|---|
| predicted_market_share | | | | | |
| projected_revenue | | | | | |

ii. Recommend a destination.

## Task 10 (4 *points*)

Your assistant is interested in using a regression tree model to predict the cost of flights.

(a) (*2 points*) Explain how your assistant can address overfitting in a regression tree model, including reference to two specific parameters in your response.

**ANSWER:**

---

Your assistant is trying to choose between using a random forest and a gradient boosting machine.

(b) (*2 points*) Explain which of the two types of models will be more in need of adjustment to avoid overfitting.

**ANSWER:**

## Task 11 (6 *points*)

You are working with a client who is interested in the relative cost of flights between regions within the United States. They ask you for a stratified sample with 10 observations for flights between each pair of regions (they are not interested in flights starting and ending within the same region). The four regions are West, Midwest, Northeast, and South.

(a) (*1 points*) Calculate how many observations your total sample will contain. Assume you can find 10 observations for each pair of regions.

**ANSWER:**

---

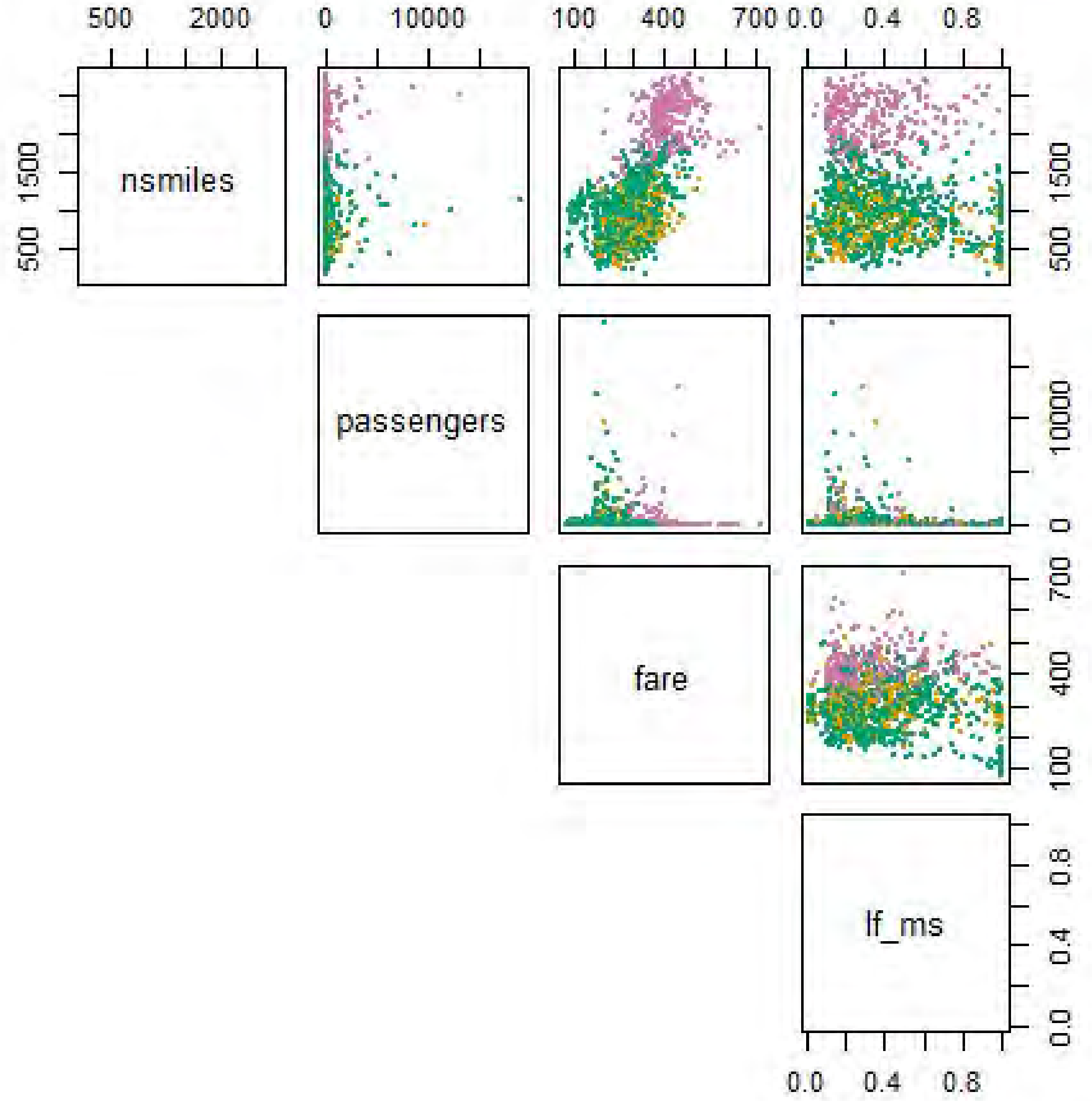
(b) (*3 points*) Describe how you would construct your stratified sample. Include the specific variables you would consider when constructing the stratification groups.

**ANSWER:**

---

Your assistant creates the graphic below to illustrate the costs between region for flights between the Northeast region and the other regions.

(c) (*2 points)* Identify two observations about the costs of flights between different regions based on the graphic below.

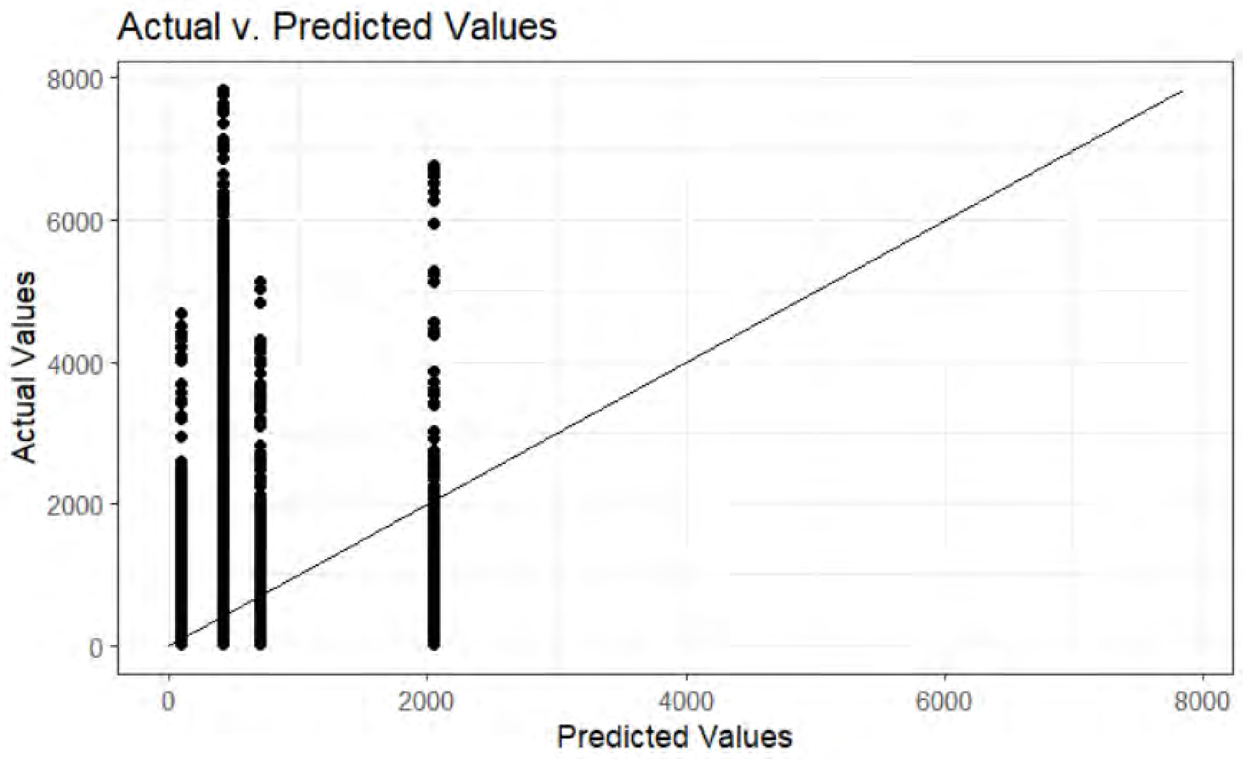


Yellow = Northeast-South,
Pink = Northeast-West,
Green = Northeast-Midwest

**ANSWER:**

## Task 12 (4 *points*)

Your assistant fits a regression decision tree model to predict potential traffic (number of passengers) between city pairs that don't currently have an established route. They graph the actual values vs. the predicted values from the training data set below.



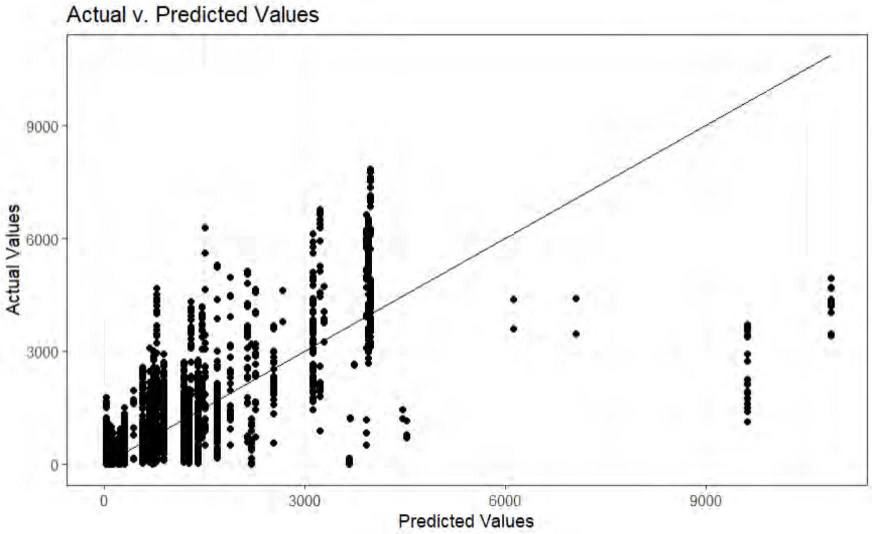
(a) (*2 points*) Evaluate how many terminal nodes the decision tree has. Explain your reasoning.

**ANSWER:**

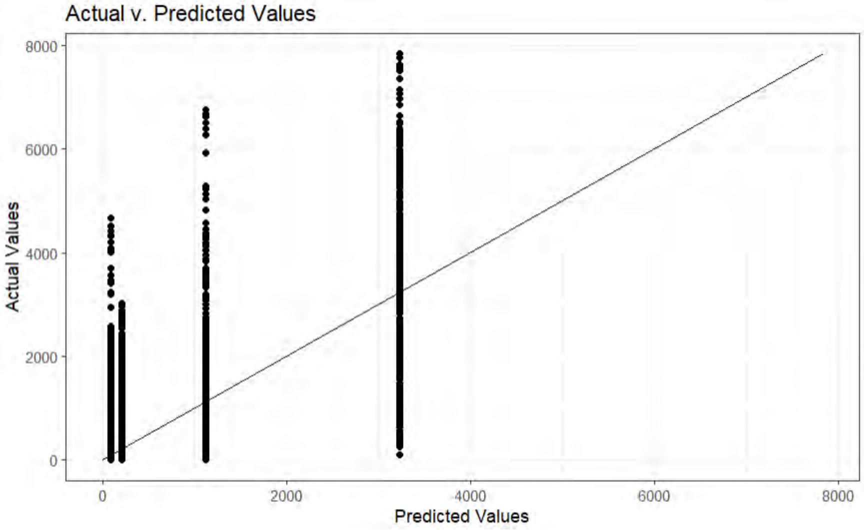
---

Your assistant produces two different models by adjusting the hyperparameters of the decision tree.

Model 1:



Model 2:



Model metrics based on the training data set:

| Model | RMSE |
|---|---|
| Model 1 | 378.10 |
| Model 2 | 381.41 |

(b) (*2 points*) Recommend which model should be used and Justify your recommendation.

**ANSWER:**