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.

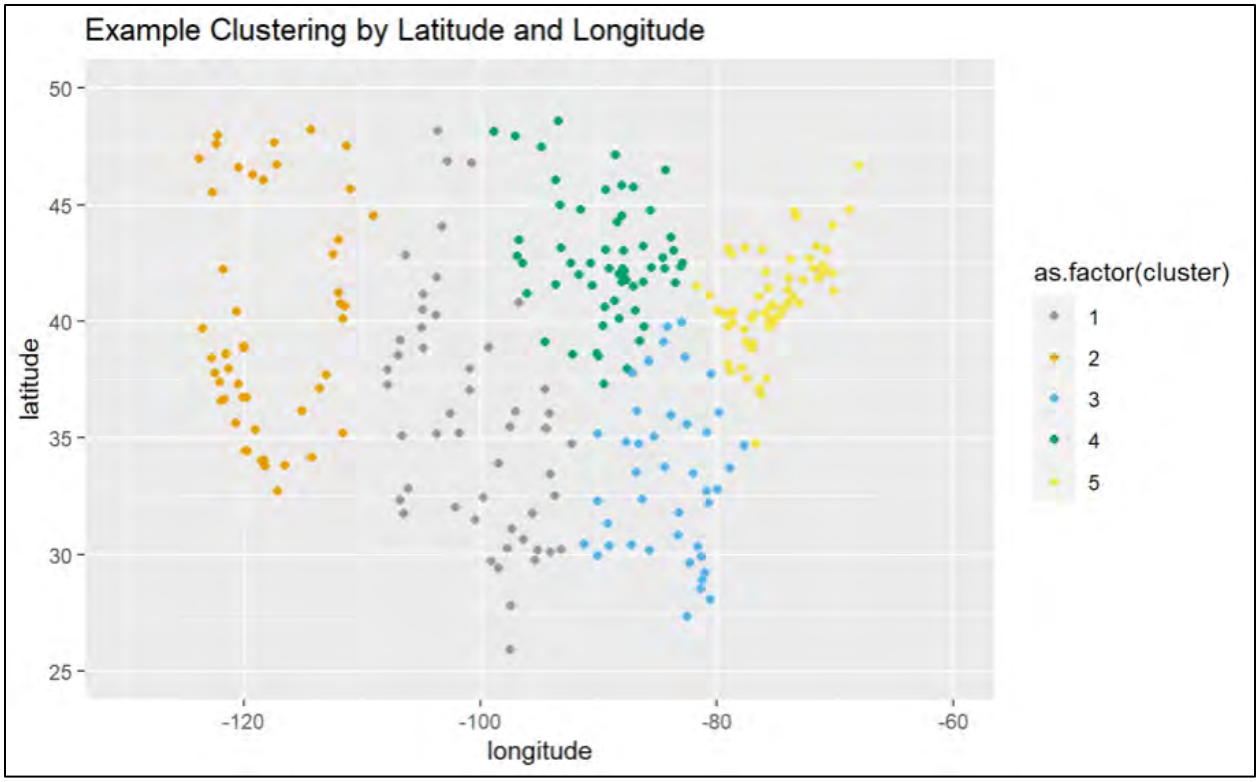
Candidate performed was mixed task. Full-credit responses included a distinct advantage and drawback that were relevant to the data and regions provided. No additional credit was awarded for describing a drawback that was the converse of the advantage provided (e.g. listing interpretability as an advantage of the pre-defined variable and a drawback of generating clusters). Minimal credit was awarded to descriptions of generic limitations of clustering techniques such as the need to scale variables and select the number of clusters.

ANSWER:

Assuming at least as many clusters as there are regions (4), creating clusters based upon the data could result in the cities grouped into each cluster having more in common than those cities with the same region label. The clustered cities would certainly be closer geographically. If other features in the data are included in the clustering, then clustered cities may be similar to each other in additional ways that are not captured by predefined region variables. This similarity will be greater the larger the number of clusters used.

An advantage to using the predefined region variable is that it might be meaningful for reasons that are not represented as variables in the dataset and thus cannot be automatically captured through clustering. It is also possible that clustering could capture noise in the data, grouping cities together that do not actually belong together. In addition, the predefined region variable might be easier to interpret.

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.

Candidates struggled with this task. Full credit was awarded to clear descriptions of the calculation that were mathematically equivalent to the model solution. Partial credit was awarded for minor formula errors like failing to square terms in the formula for Euclidean distance. No credit was awarded for descriptions of clustering algorithms themselves.

ANSWER:

For each cluster, *i*, and city, *j*, the steps for calculating within cluster sum of squares are as follows:

1. Calculate the centroid of each cluster:
    - latitude_centroid_*i* = mean(latitude_*j* in cluster *i*)
    - longitude_centroid_*i* = mean(longitude_*j* in cluster *i*)
2. Calculate the squared Euclidean distance between each city and the centroid of its cluster:
    - sqdistance_*j* = (latitude_*j* – latitude_centroid_*i*) ^ 2 + (longitude_*j* – longitude_centroid_*i*) ^ 2
3. Sum the squared distances for all cities:
    - within_sum_of_squares = sum(sqdistance_*j* ^ 2)

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.

Most candidates earned full credit on this task. Full-credit answers recognized that the variables are on different scales and identified an appropriate method to re-scale the variables.

ANSWER:

The variables should all be standardized prior to running the *K*-means clustering algorithm. Standardizing variables ensures that they are all on the same scale and will get an equal amount of weight in the clustering. For many problems one or more variables could be on vastly different scales from others and thus could dominate the clustering unless standardization is performed first. In this case, latitude and longitude are on similar scales, but average fare is on a different scale. Thus, standardization would avoid average fare getting more weight than it should.

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.

Candidates performed poorly on this task. Few candidates were able to make a connection between the correlation matrix and principal components analysis, which was required for full credit. Many candidates provided generic descriptions of PCA; minimal partial credit was awarded for these types of responses when they made a connection to the problem.

ANSWER:

The goal of Principal Components Analysis is typically to use fewer principal components than original variables to reduce the dimensionality of the problem. However, the correlation matrix shows very little correlation among the variables, indicating that PCA will not be effective for reducing dimensionality. Based on the information provided, there is no reason to believe that clusters based on the principal components will provide any advantage over clusters based on the untransformed variables.

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.

Candidate performance was mixed on this task. Full-credit was awarded for responses that addressed the appropriateness of using the clusters as a response variable or as a predictor variable since the wording in the task is ambiguous. Most full-credit responses discussed data leakage in some way, as is shown in the model solution.

ANSWER:

Since the intended purpose of the clusters is to predict the average fare between two cities, including a city's average fare in the clustering could result in overfitting and an optimistic measurement of model fit. Thus, in general we would not recommend including average fare in the clustering. However, there might be a legitimate way to do this by using average fare to cluster *only* on the training data, and then applying the clustering to the validation data using only the latitude and longitude variables.

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.

Candidates performed poorly on this task. Full-credit answers identified passengers as an appropriate variable to use for measuring market size and described reasonable data exploration steps. Most full-credit answers discussed analysis of seasonality or bivariate analysis of passengers and carrier.

ANSWER:

The average daily passengers variable can be used to measure market size since passengers are the ultimate recipient of the product. To measure market size, I would look at the most recent quarter of data available and filter to city pairs flying into or out of Washington state, excluding any flights of fewer than 300 miles. For those pairs, I would then look at the results over time to see how the market size has varied, if there is quarterly seasonality, and if the market appears to be growing or shrinking over time.

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.

Candidate performance was mixed on this task. Most candidates struggled to demonstrate an understanding of how market concentration can be analyzed using the dataset for part (i). Candidates generally performed better on part (ii), with most candidates identifying at least one reason why the data is insufficient. Performance on part (iii) task was mixed; full-credit answers identified that the city1 and city2 are assigned arbitrarily, as discussed in the notes on the dataset.

ANSWER:

i. Yes, we can answer this type of question. While we don't have the level of concentration and cost for every airline, we do have the concentration and cost for the

airline with the highest market share for each city to city grouping. We can also assess how changes in market concentration over time affect average fare pricing.

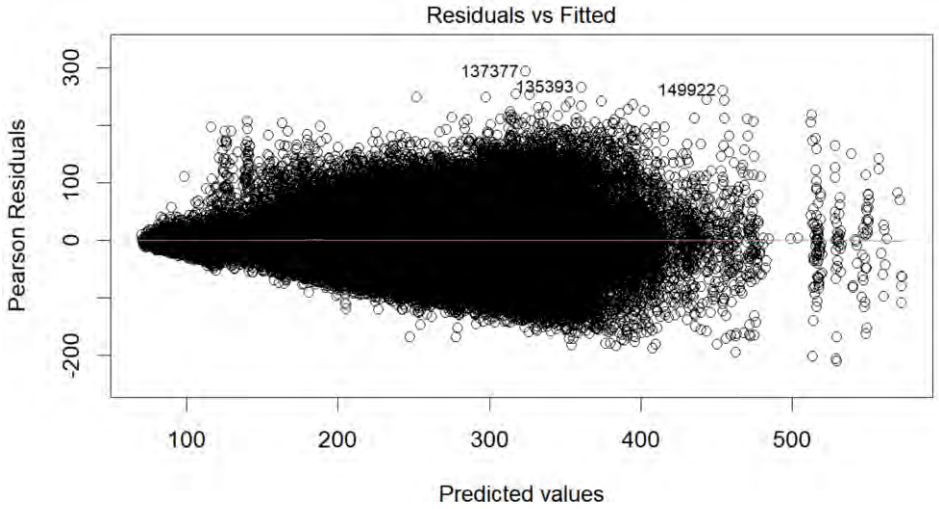
ii. No, we do not have the data to answer this type of question. Our data set is quarterly, so it would not be possible to match cancellations to specific flights and the temperature and precipitation at those times. We also don't have cancellation data or temperature data or precipitation data. To answer this question, we would need a more granular data set and a record for each flight rather than aggregated data across a quarter. We would also need additional variables representing whether a flight was cancelled or not and precipitation and temperature data for each airport at the time of each flight.

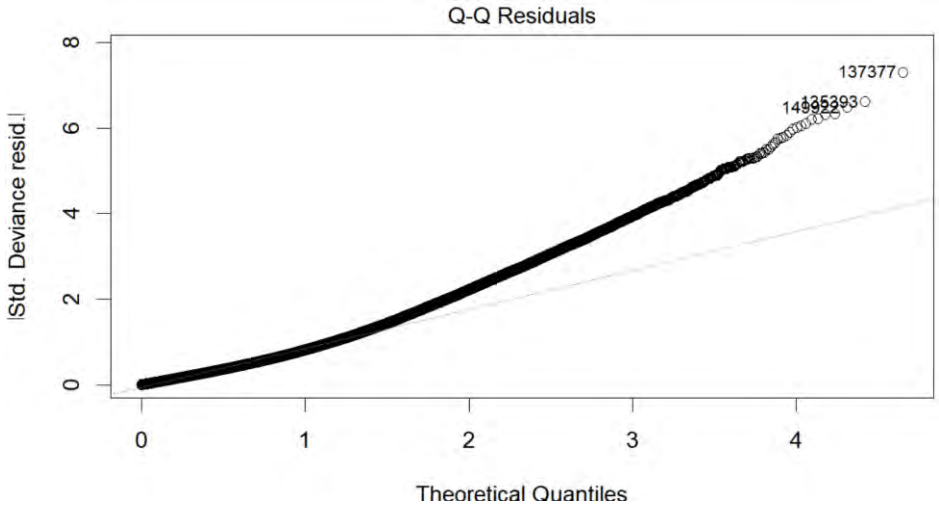
iii. No, we do not have the data to answer this type of question. Our current data set aggregates data at the city to city level, which means that if a city is served by two different airports, the data from both airports is aggregated into a single record. Without airport-specific pricing data we would not be able to recommend when it makes sense to look at an alternative nearby airport to save money. In order to answer this question, we would need a similar data set, but rather than aggregating results across each city to city pair, we would want to aggregate results at the airport level. This would let us recommend when a nearby airport might offer cost savings.

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.

Candidates performed well on this task. Full credit responses provided a recommendation for both the distribution and link function, and justified the recommendation with an observation from the plots.

ANSWER:

The normal distribution with identity link does not seem appropriate based upon these results. In the Residuals vs. Fitted plot, we see significant heteroskedasticity, with the residuals clearly having greater spread on the right side of the plot than on the left. We see this also in the Q-Q Residual plot, where

standard deviation of the residuals in the right tail is higher than would be expected from the normal distribution.

I recommend a GLM with gamma distribution and log link function. With gamma distribution, standard deviation of the residuals is higher for higher predicted values, which matches what we observe in the plots. The log link function is more appropriate than the identity since it only permits positive values, matching the domain of the fare variable and the mean of the gamma distribution.

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.

Most candidates received partial credit on this task. Full-credit responses used the model output provided as part of the evaluation. Partial credit was awarded for insightful responses that only discussed general properties of the variables without considering information that can be gathered from the model output.

ANSWER:

i. The table above shows that the year variable as represented as numeric has coefficient that is statistically significant, but with a small coefficient which implies the effect is not very large. It would be good to do some bivariate data analysis on year and fare to determine if the decrease is constant through time. If it isn't, including year as a categorical variable should be considered. One disadvantage of encoding year as categorical would be an increase in the number of variables.

ii. The coefficient for the quarter variable is not statistically significant and there are even more reasons to think that quarter may not exhibit a consistent monotonic relationship. For example, if the holiday season that straddles the end of the year and beginning of the next one affects fair prices, using quarter as a numeric variable may not be a good idea. Using quarter as a categorical variable would allow non-linear, non-monotonic relationships to be captured.

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

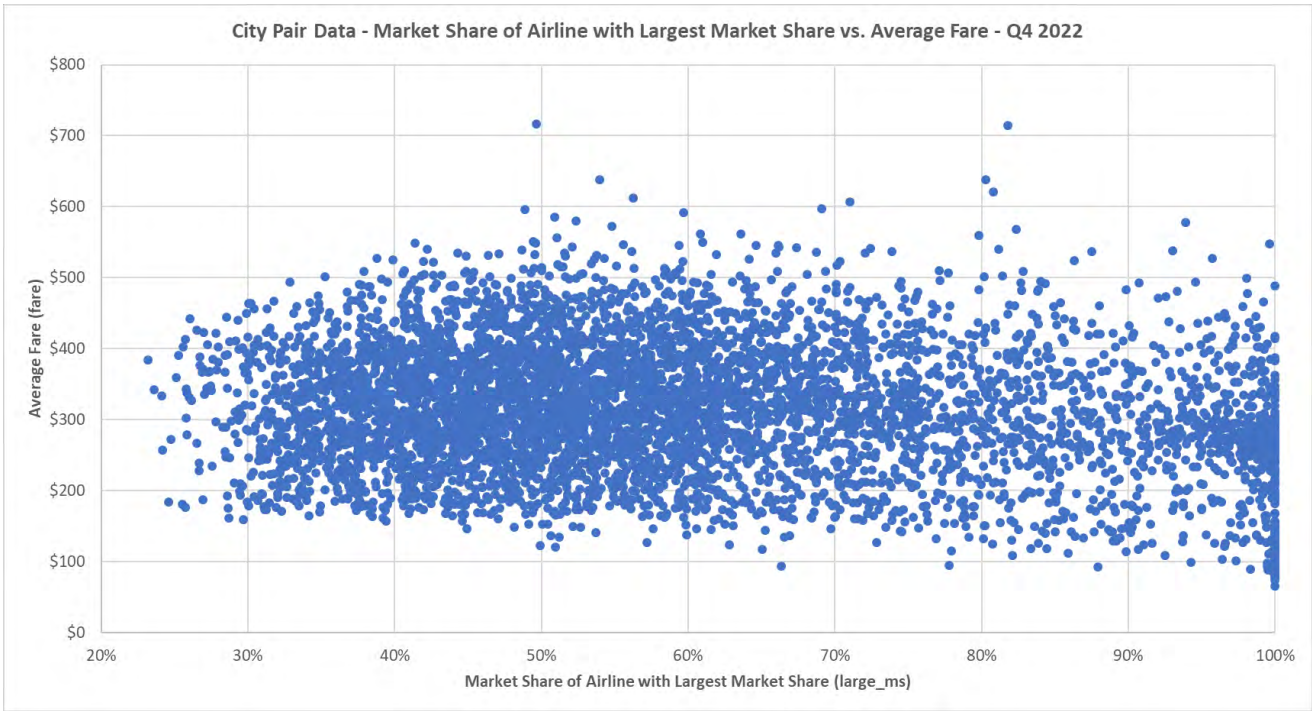
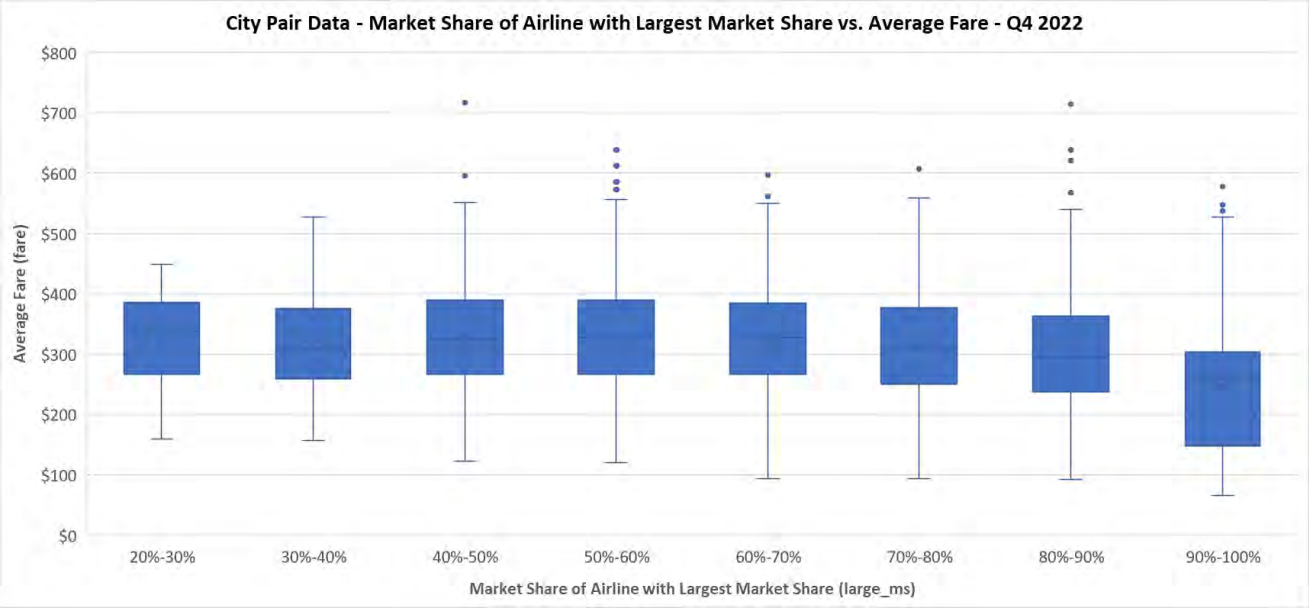


Exhibit B



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

Performance was mixed on this task. Full-credit responses addressed what the visualizations convey for the given variables. Generic pros and cons about the plot types received partial credit.

ANSWER:

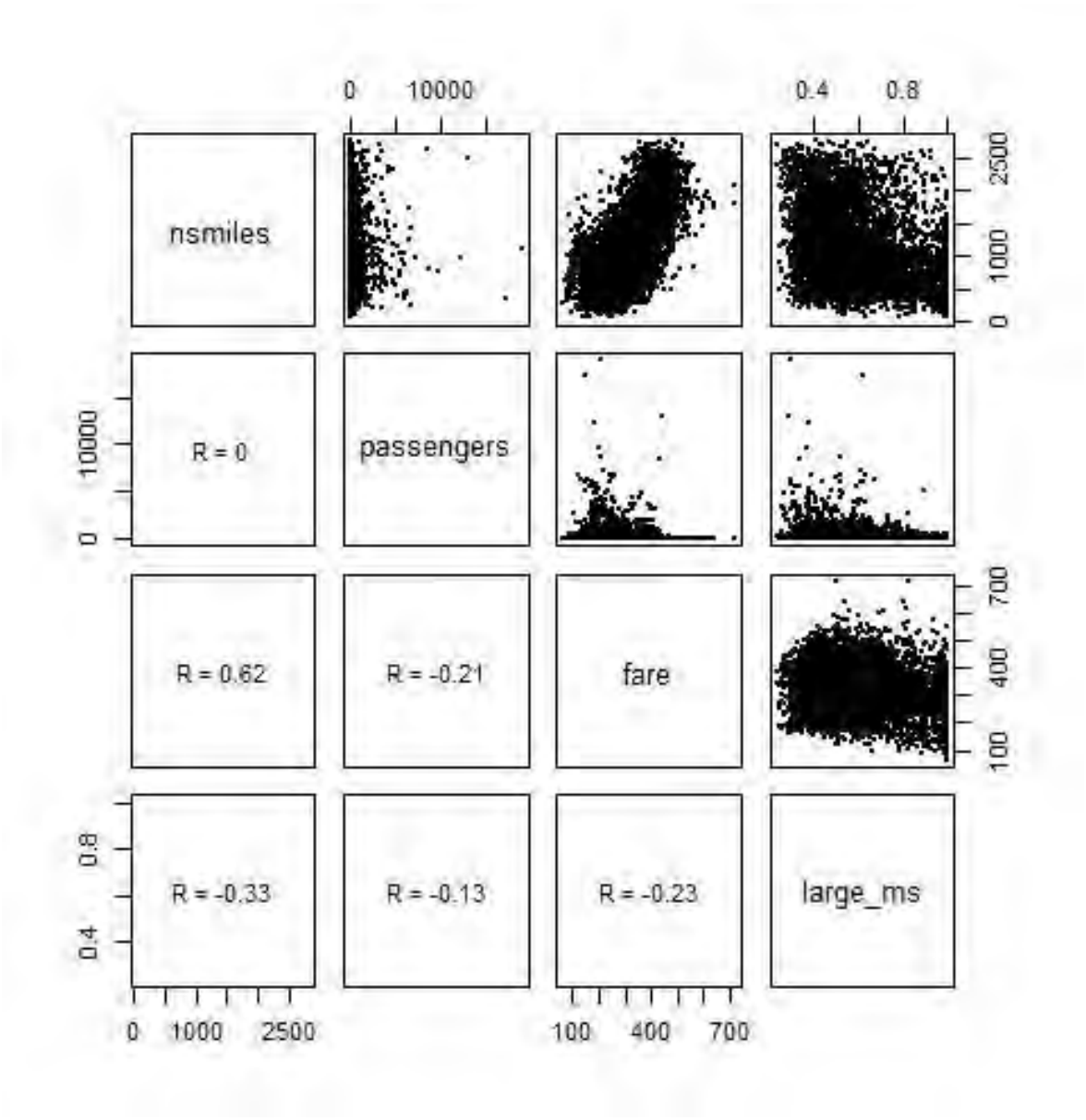
Exhibit A – Pro:
The scatterplot in Exhibit A shows the full range of the data and makes it easy to identify specific features, such as the fact that the lowest value for large_ms is around 23% and that there are a large number of city pairs that have a value of 100% for large_ms.

Exhibit A – Con:
Given the large volume of data visualized in Exhibit A it is difficult to see how many data points are in certain ranges, especially for values of large_ms between 40% and 60%. It is also difficult to tell if there is an overall trend to the dataset.

Exhibit B – Pro:
The box and whisker plots for deciles of large_ms make it very easy to see that the mean fare is lower for values of large_ms between 80%-90% and 90%-100%. It is also easy to pick out key features for each decile such as the outlier and quartile values.

Exhibit B – Con:
By dividing the data into deciles, it isn't possible to tell the overall number of values that are within each decile. There is no easy way to see that there are fewer data points between large_ms values between 70% and 90% as compared to values between 40% and 60%.

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.

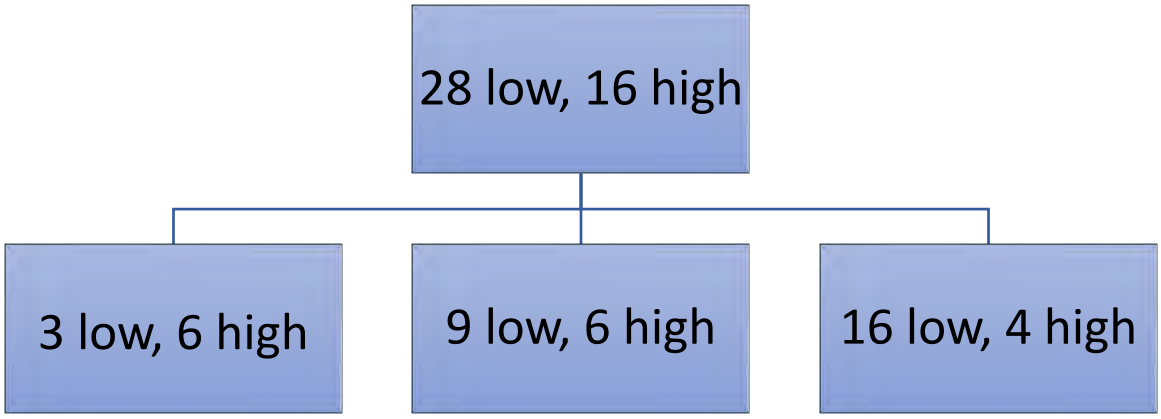
Candidates struggled with this task. Full-credit recommendations were grounded in information from the visualization provided.

ANSWER:

Given the variables nsmiles and fare are strongly correlated, I would recommend my assistant look at the relationship between fare and large_ms while controlling for nsmiles. The approach I recommend for this analysis is to group the data into deciles based on nsmiles and look at the scatterplots between fare and large_ms within each decile of data for nsmiles.

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.

Candidates performed poorly on this task. There is an error in the formula provided; $log_2(p_1)$*should instead be* $log_2(p_i)$*. Full credit was awarded for accurately calculating information gain based on either the formula provided or the correct formula. A common error was to add or average the log values instead of weighting by* p_i*.*

ANSWER:

Split 1 E = -(3/9)*$\log_2$(3/9) - (6/9)*$\log_2$(6/9) = 0.91829
Split 2 E = -(9/15)*$\log_2$(9/15) - (6/15)*$\log_2$(6/15) = 0.97095
Split 3 E = -(16/20)*$\log_2$(16/20) - (4/20)*$\log_2$(4/20) = 0.72193

The weighted average root node entropy after the split = (9/44)*0.91829 + (15/44)*0.97095 + (20/44)*0.72193 = 0.84699
The root node entropy before the split = -(28/44) *$\log_2$(28/44) -(16/44) *$\log_2$(16/44) = 0.94566
The Information Gain = 0.94566 – 0.84699 = 0.0987

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).

Candidate performance was mixed on this task. Full-credit responses identified the key difference that (b) is a regression tree whereas (a) is a classification tree and also described how the impurity calculations differ. Most candidates who identified the classification/regression difference were able to adequately describe the regression impurity formula for full credit.

ANSWER:

This tree is a regression tree, while the tree in part (a) is a classification tree. For a regression tree, residual sum of squares is used rather than Gini or entropy. Each node's impurity is the average value of the residual sum of squares for all observations in the node.

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**.

Most candidates received full credit on this task. Some mistakes included exponentiating the prediction and switching fare and passengers in the interpretation.

ANSWER:

With a fare of $0, the predicted number of passengers is the intercept of 593.94020. For every increase in fare of $1, the predicted number of passengers goes down by 1.54073.

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.

Candidates performed well on this task. For full credit, responses needed to select a model based on AIC with a discussion of how this indicates that Model 1 does a better job of balancing model fit and complexity. Many candidates discussed other methods of model selection and how they could be applied; no credit was awarded for these discussions as they were considered off-topic.

ANSWER:

When using AIC for model selection, the model with the smallest value should be selected. This provides the best balance between model fit and complexity.

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

Candidate performance was mixed on this task. Many candidates failed to demonstrate an understanding that AIC is calculated using the same data as the model is trained on.

ANSWER:

AIC is calculated on the training data, which is data that the model has already seen. A better approach to assessing the model's predictive power is to measure the model fit on a set of validation or test data which the model did not have during the fitting process. This will better detect the degree of overfitting, if any, that may be happening on the training data. If insufficient data is available for validation and testing splits, cross validation could be used instead.

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.

Candidates performed well on this task. Full-credit responses demonstrated an understanding that the numeric values assigned to the clusters are labels and that their numeric values are arbitrary.

ANSWER:

Treating cluster numbers as numeric features does not make sense because the ordering of the cluster numbers is arbitrary, as are the magnitudes of the cluster numbers. Treating cluster number as a categorical feature forces the model to ignore the numbers themselves; for example, a GLM will fit a separate coefficient for each unique cluster (except the number that is designated as the base level).

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

Candidate performance was mixed on this task. Many responses lacked sufficient detail, explaining that the process creates new variables, but not describing how. A common error was describing a process of creating three new variables instead of two, which failed to acknowledge that a variable isn't needed for the base level.

ANSWER:

When a categorical variable is used in a regression, it is binarized into multiple numeric features each taking the value of 0 or 1. To apply binarization to the city_cluster variable, first select a base level, say cluster 1. Then create variables for the other two levels, say cluster_2 and cluster_3, setting each cluster_x variable to 1 for observations in cluster x and 0 otherwise.

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.

Candidates performed well on this task. While most candidates demonstrated sound understanding of LASSO regression, fewer candidates were able to effectively describe ridge regression. Some candidates provided a recommendation, which was not required for full credit.

ANSWER:

Ridge and LASSO regression are both regularization techniques used to address overfitting in linear regression models. They work by adding a penalty term based on the magnitude of the coefficients to the linear regression cost function.

LASSO uses an absolute value penalty function, which can reduce coefficients to 0, which has the benefit of reducing the dimensionality of the model. Ridge regression uses a square penalty function. A consequence of this is that ridge regression does not shrink coefficients all the way to 0, and therefore does not reduce dimensionality.

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

Candidates performed well on all three parts of this task. Full credit was awarded for responses that included valid recommendations with justification.

ANSWER:

i. A log link is appropriate for income as it provides strictly positive predictions. A gamma distribution is appropriate since it is continuous and supports positive values.

ii. A logit or probit link would be appropriate for this example as it allows only valid probabilities (values between 0 and 1) as values. The binomial distribution is appropriate since the response variable is binary.

iii. A log link is appropriate for this variable as it provides strictly positive predictions. A Poisson distribution is appropriate for counting variables.

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.

Candidate performance was mixed on this task. Many candidates struggled to articulate what a weighting variable is, leading to partial or no credit.

ANSWER:

If **nsmiles** is used as a weighting variable, the model will adjust the contribution of each observation in estimating the model coefficients based on the value of **nsmiles**. The model will not estimate a coefficient for **nsmiles** if it is used as a weighting variable.

If **nsmiles** is used as a predictor variable, it will be included in the model formula, and the model will estimate a coefficient for **nsmiles**.

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.

Candidates struggled with this task. Full-credit responses explained how the two variables provided interacted to impact the response variable. No credit was awarded for only identifying whether an interaction exists based on the output.

ANSWER:

Both discount and passengers have positive coefficients, meaning market share increases as either variable increases in isolation. Their interaction coefficient is negative, which dampens the impact of each variable's individual effect when both discount and passengers increase. Increasing the discount is more effective in increasing market share for routes with fewer passengers.

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.

Candidates struggled with this task. Few candidates were able to identify any qualities of a good KPI, much less discuss them in the context of annual revenue.

ANSWER:

Annual revenue is appropriate as a Key Performance Indicator for the following reasons:

- It is measurable from the given data and assumptions.
- It is in line with the overall business strategy to start the new flight route.
- It is directly related to the underlying objective of the project to estimate the potential revenue of the new flight route.

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.

Candidates performed well on this task, with most candidates receiving full credit. The most common reason for awarding partial credit was failing to discuss how a performance metric is used.

ANSWER:

For LASSO regression, the hyperparameter to be tuned is lambda, which is the weight given to the regularization term in the objective function.

The most common approach for hyperparameter tuning during model fitting is to use k-fold cross validation. The steps for applying k-fold cross validation to tuning lambda are:

- Set an array of possible lambda values to test
- Partition the training data k times so that for each partition, 1/k of the data is used as a new "testing data" set and remainder is a new "training data" set. Partitioning is done without replacement.
- For each value of lambda to be tested and for each k-fold data partition

    - Fit the model being considered using the current lambda and the current training data set for this partition
    - Predict values using the current test data set for this partition
    - Determine an error measure for the predicted values
- Determine the mean and standard deviation of the prediction error over the k data partitions based on the current lambda
- Select a value of lambda which best balances the mean and standard deviation of the prediction error

---

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.

Candidates struggled with this task. Very few candidates identified the core issue that all folds were used in selecting the hyperparameter, making that data invalid for testing performance of the selected model.

ANSWER:

This approach is not appropriate. Since the k validation data folds are used to select the tuning parameters, this data has been used for model training. Therefore, this is not a true test of the model on previously unseen/unused data.

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:

$$projected\_revenue = \min(predicted\_market\_share \times passengers, maximum\_available\_seats) \times weekly\_flights \times 52 \times (1 - discount) \times fare$$

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.

Candidate performance was mixed on this task; most candidates received either full credit or struggled with the entire task.

ANSWER:

i. Complete the table.

City	Aspen	Fresno	Grand Junction	Montrose	Reno
predicted_market_share	0.1918	0.2466	0.2331	0.2172	0.2565
projected_revenue	1,717,385.72	945,334.10	282,821.12	573,885.48	2,677,294.28

	coefficient	Aspen	Fresno	Grand Junctio	Montrose	Reno
(Intercept)	.					
discount	.	0.2	0.2	0.2	0.2	0.2
passengers	-2.18299E-06	127	88	26	49	226
nsmiles	9.7572E-05	1754	2484	1851	1820	2443
fare	0.000225397	565	349	374	432	370
carrier_no	-0.01286598	3	6	3	3	5
city2Aspen, CO	-0.06761603	1				
city2Denver, CO	0.003398898					
city2Fresno, CA	0.003147759		1			
city2Grand Junction, CO	0.006847131			1		
city2Montrose/Delta, CO	-0.01892722				1	
discount:passengers	-8.68671E-06	25.4	17.6	5.2	9.8	45.2
	model prediction:	0.191779019	0.246639504	0.233051717	0.217235515	0.256549611
	Revenue:	$ 1,717,385.7	$ 945,334.1	$ 282,821.1	$ 573,885.5	$ 2,677,294.3
	(revenue = min(pred*passengers, 150) * 3 * 52 * (1-20%) * fare)					

ii. Recommend a destination.

Since it has the highest projected revenue, I recommended Reno, NV.

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.

Candidates performed well on this task. Full-credit responses discussed how to tune the hyperparameters in a single tree, or in an ensemble tree model (random forest or boosted tree).

ANSWER:

Regularization is the process of building a model that penalizes complexity and is biased towards a simpler model with the goal of avoiding overfitting. Some methods include tuning the complexity parameter using a cost-complexity table, or limiting the maximum depth of the tree to limit the complexity of the tree that is built.

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.

Candidates performed well on this task. Full-credit responses identified that the GBM is more susceptible to overfitting and provided some explanation of how this can result from iteratively fitting to residuals.

ANSWER:

Random forests are good at avoiding overfitting while gradient boosting machines are prone to overfitting. Therefore, regularization would be more important for a gradient boosting machine. GBMs unlike random forests are built sequentially. Each tree tries to correct the remaining errors after fitting the prior one, naturally leading to overfitting. Random forests are built simultaneously from bagged trees and hence are less susceptible to overfitting.

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.

Candidate performance was mixed on this task; most candidates received either full credit or no credit.

ANSWER:

Your total sample will contain 60 observations. The pairs of regions are: South-West, South-Northeast, South-Midwest, West-Midwest, Northeast-Midwest, and West-Northeast. You will have 10 observations from each for a total of 10 * 6 = 60 observations.

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

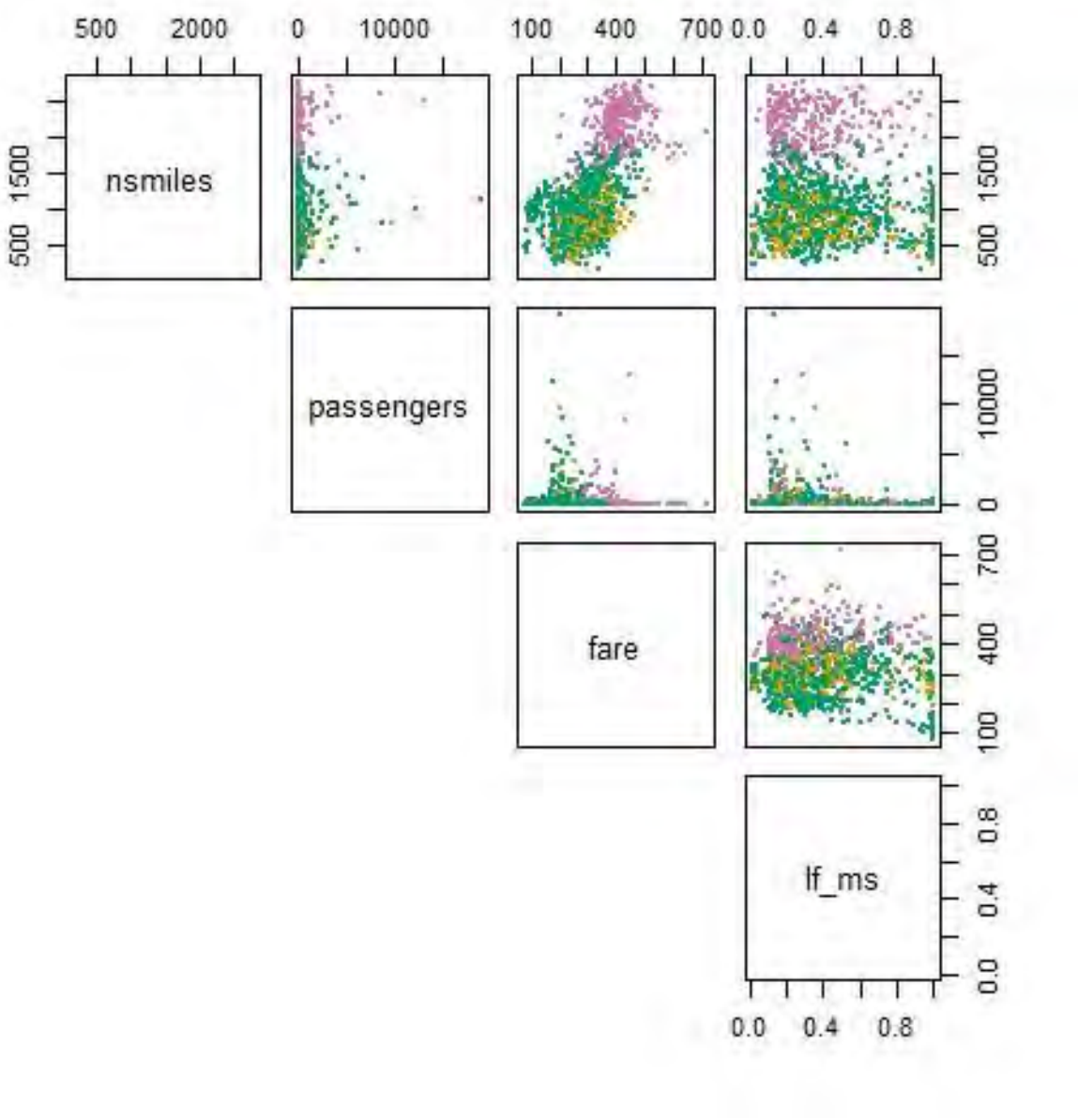
Candidates performed poorly on this task. Most candidates were able to describe stratified sampling in general, receiving partial credit. Fewer candidates discussed how to construct a stratified sample using variables in this dataset, which was required for full credit.

ANSWER:

I would start by constructing an additional variable called “Region_to_region” that contains the two regions that a flight goes between. As an example, if either region_1 = “South” and region_2 = “West” or region_1 = “West” and region_2 = “South”, the variable would have the value “South-West”. I would do the same thing for all other pairs of regions. I would exclude the pairs where both the starting and ending region were the same (e.g., South-South, West-West, etc.).

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

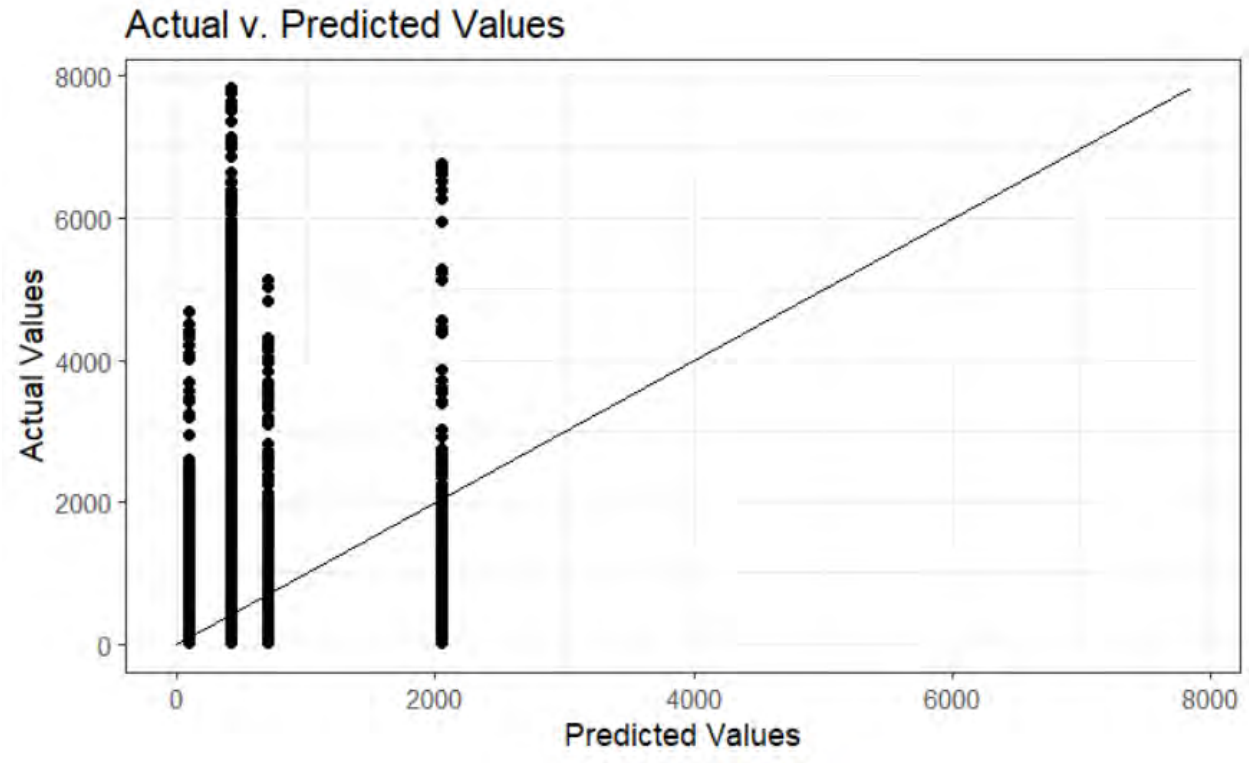
Candidates performed poorly on this task. Most candidates provided a strong observation around the relationship between fare and nsmiles. Few candidates provided a strong second observation.

ANSWER:
The average number of miles flown is much higher for flights between the Northeast and the West, and this is correlated with the higher average fares paid for flights between those regions.

There are many more flights where the lowest cost carrier has 100% of the flight volume for flights between the Northeast and South and Northeast and Midwest regions compared to the Northeast and West regions.

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.

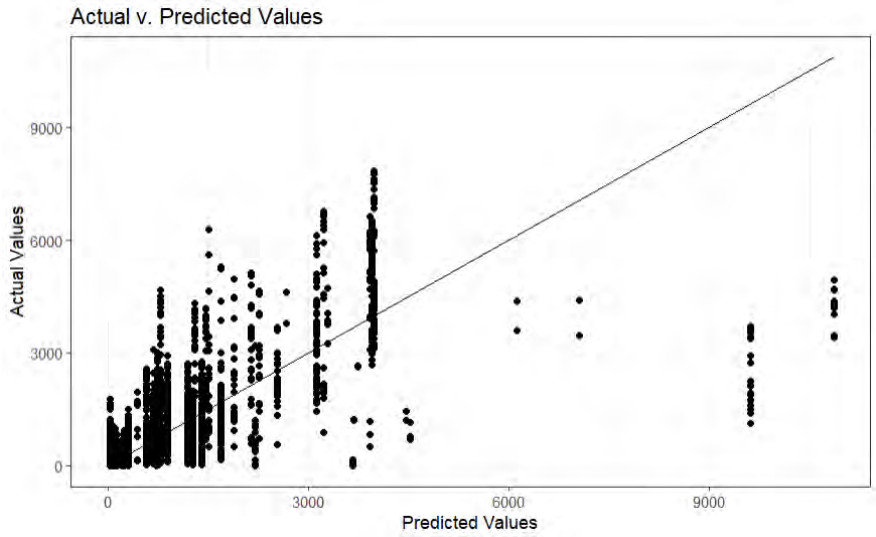
Candidates performed well on this task overall. Full-credit responses generally resembled the model solution, referring to the number of distinct predictions shown in the plot.

ANSWER:

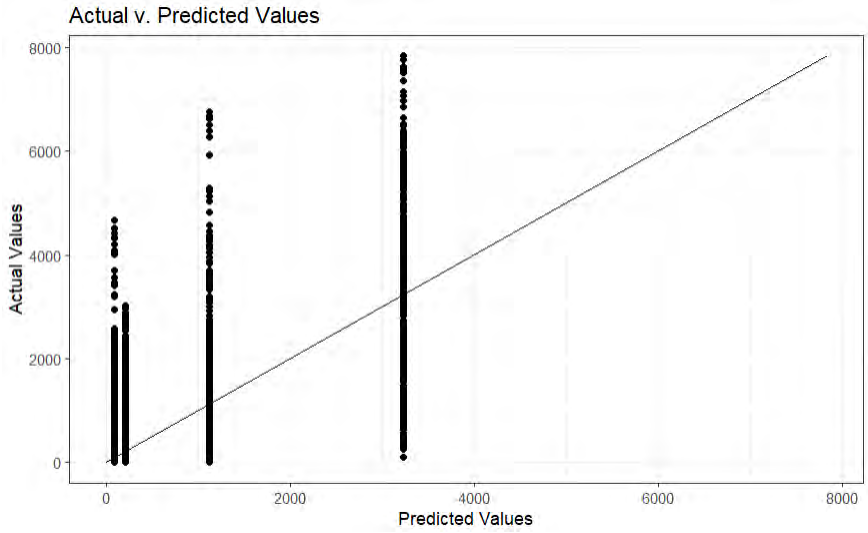
Based on the chart above we can conclude the decision tree has four terminal nodes.
The predicted values from the model, displayed on the x-axis, have only four distinct values indicating the terminal nodes of the model.

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.

Candidates performed well on this task. Full credit was awarded for recommending either model provided the response identified the less complex model from the plots, identified the better-performing model from RMSE, and took both considerations into account.

ANSWER:

From the plots, we see that Model 2 is less complex, having only 4 terminal nodes as compared to Model 1, which has many more terminal nodes. However, Model 1 has better overall performance, as indicated by the lower RMSE on the training data set. This is expected when performance is measured against the training set.

Since we don't have performance on a testing data, I recommend Model 2. Although Model 2 does not perform as well on the training data, it is less complex, and therefore less likely to be overfit to the training data.