Airline Pricing Dynamics Analysis

Ha Doan, Lena Le, Chi Vu

DA350 - Advanced Methods of Data Analytics

Dr. Zhe Wang

Fall 2023

1. Introduction

The U.S. Department of Transportation's Air Travel Consumer Report is a comprehensive monthly publication that offers insights into various aspects of airline service quality, including flight delays, mishandled baggage, oversales, and a range of other consumer complaints. Our dataset provides a focused summary by city, detailing key metrics such as the count of city-pair markets linked to each city, passenger traffic volumes, average fares, cost per mile, and the typical travel distance.

The dataset name is *"Consumer Airfare Report: Table 2 - Top 1,000 City-Pair Markets"*.

For this data set, our main focus is to investigate the following questions:

- What are the underlying patterns or dimensions in the airfare data based on the numbers of current passengers and the current fares for different states? [PCA] From that, can we identify distinct groups of states based on numbers of current passengers and the current fares? [K-means]
- Is the impact of distance on airfare uniform across all states? [Linear Mixed Effect Model]
- How can distance and number of passengers could influence the affordability of air flights, thus helping airlines make better pricing strategies [Logistics Regression]

Below are the variables from the main dataset:

Variable	Type	Meaning
Year	Numerical	Data year
quarter	Numerical	Data quarter
city	Categorical	City is used to consolidate airports serving the same city market
cur_passengers	Numerical	Current year number of passengers traveling to and from each city
cur_fare	Numerical	Current year average fare
cur_yield	Numerical	Current year yield (average fare per mile) in Cents
distance	Numerical	Current year average distance traveled
ly_passengers	Numerical	Last year number of passengers traveling to and from each city

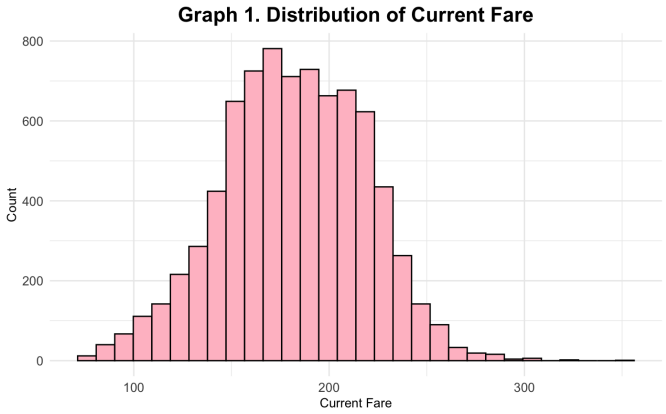
ly_fare	Numerical	Last year average fare
ly_yield	Numerical	Last year yield
ly_distance	Numerical	Last year average distance traveled

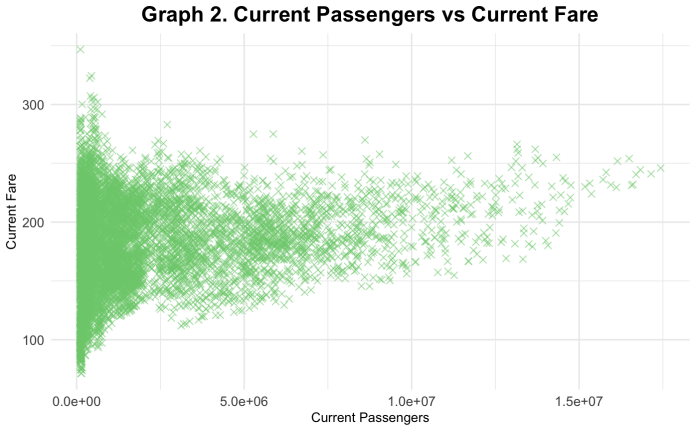
2. Ethical Consideration

The data does not contain any personal identifiable information, but the intent behind using this data still remains crucial. We want to use this information for scholarly purposes as well as serve the public interests. As a result, it is our responsibility to interpret and portray the data correctly, avoiding any skewed representations or misleading conclusions. When presenting findings or creating visualizations, clarity about the methods used, the data's limitations, and any underlying assumptions should be maintained to uphold transparency and trustworthiness in the analysis.

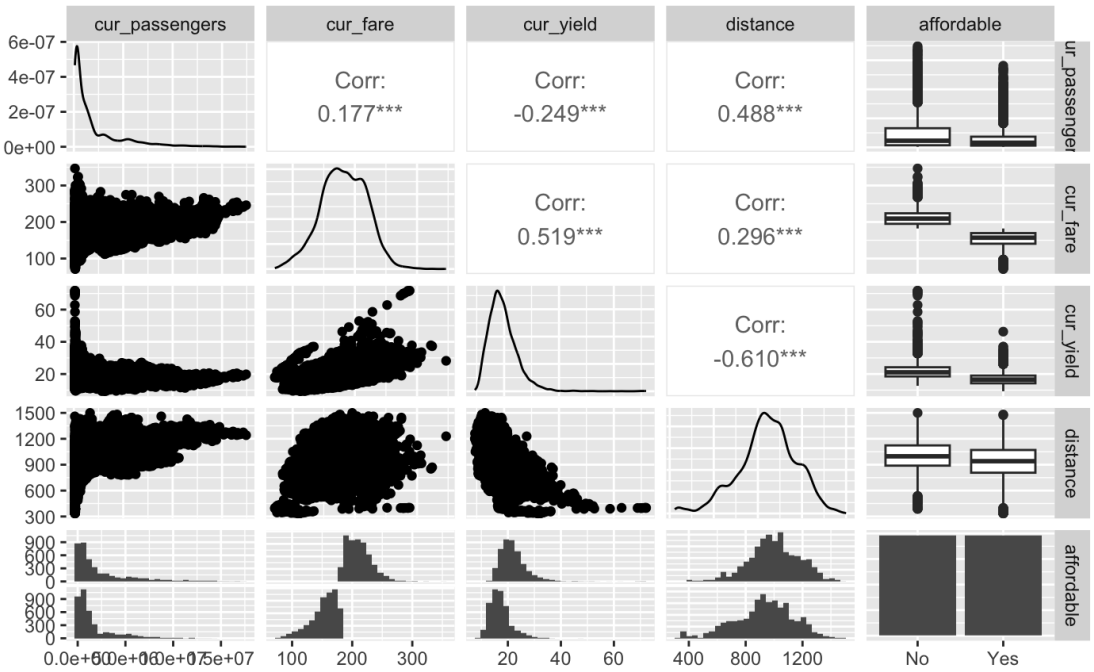
3. Data Exploration

To understand the data, first we will visualize some of the variables we want to focus on.





As shown by the Graph 1, current fare is quite normally distributed. In Graph 2, there does not appear to be a clear linear relationship between the number of passengers and the fare. Instead, the points are dispersed widely, suggesting variability in fare prices that is not strictly determined by the number of passengers alone. There is a dense clustering of data points towards the lower end of both 'Current Passengers' and 'Current Fare', which could indicate that a majority of flights have fewer passengers and lower fares.



Overall we could see that these variables (that we want to focus on) do not have meaningful multicollinearity to each other.

4. Statistical Analysis and Interpretation

Model 1: K-Means Clustering

Question 1: What are the underlying patterns or dimensions in the airfare data based on the numbers of current passengers and the current fares for different states? [PCA] From that, can we identify distinct groups of states based on numbers of current passengers and the current fares? [K-means]

Data Diagnostics & PCA

The current number of passengers have a positively moderate linear, yet statistically significant, correlation with the distances (r = .488 with significant p-value). Therefore, we used PCA to reduce any potential multicollinearity, impact of noise as well as reveal any pattern of these variables' combinations. Using the principle components, we conducted our k-means model to cluster the states based on their similarities in passenger counts and distances.

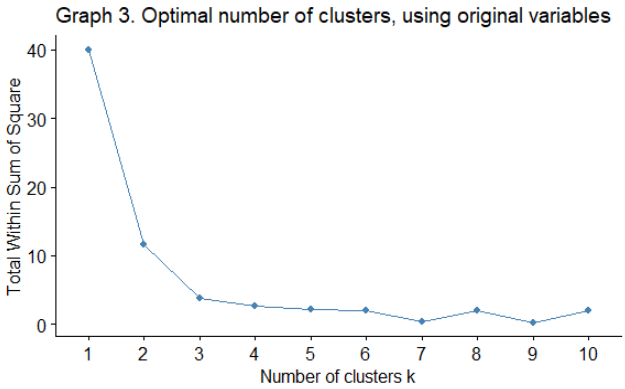
We created a new variable called 'state' based on 'city' in our dataframe and aggregated these values by state: cur_passengers, cur_yield, distance. We then conducted PCA and K-means clustering models.

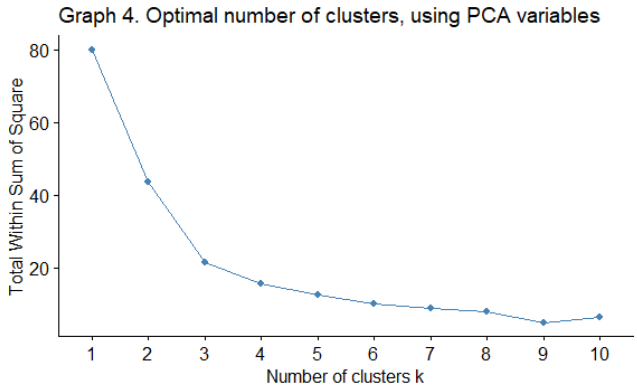
Using the variables "cur_passengers" and "distance", our Principal Component Analysis (PCA) showed a matrix of the "loadings" or coefficients of those variables on the principal components (PC1 and PC2). It's important to understand that PC1 and PC2 are linear combinations of the original variables. Specifically, in PC1, both cur_passengers and distance

have positive loadings of approximately 0.7071, indicating that they contribute positively to PC1. In other words, this means that higher values of both variables contribute positively to PC1. In PC2, cur_passengers has a positive loading, while distance has a negative loading, both approximately 0.7071. This implies an inverse relationship between cur_passengers and distance in PC2. PC1 has a standard deviation of 1.20 and for PC2, it's 0.75. PC1 explains 72% of the total variance, the majority of the variability in the data, and PC2 explains 28%. This indicates the proportion of the total variance explained by each principal component.

Model 1 Validation

We are going to assess the quality of k-means clusters by using the elbow method. We calculated the within-cluster sum of squares (WSS) for different values of k (number of clusters), plotted the WSS against k, and looked for an "elbow" point where the rate of decrease in WSS slows down.





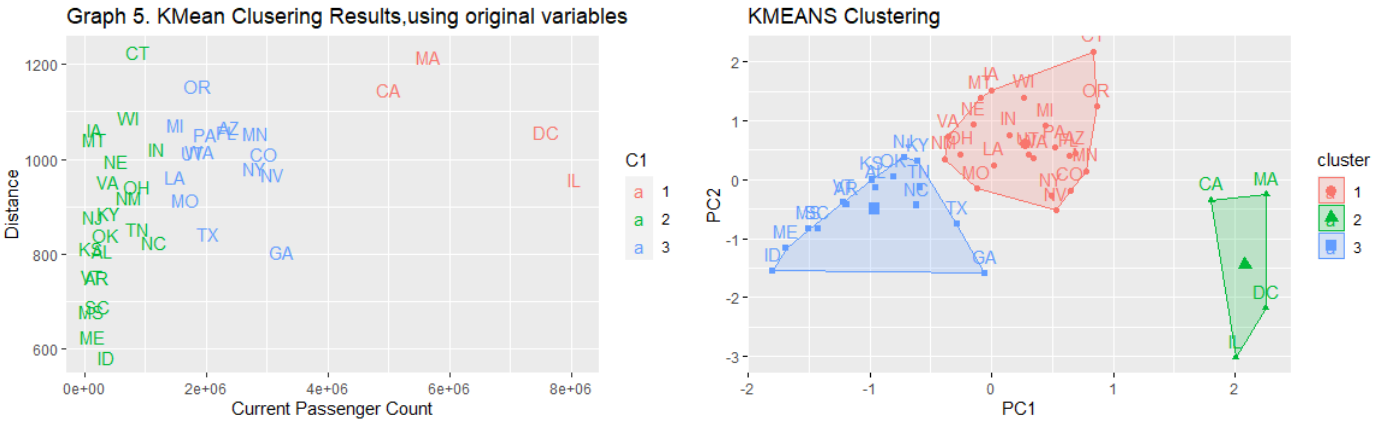
Graph 3. *Optimal number of clusters for K-means model, using original variables ("cur_passengers", "distance")*

Graph 4. *Optimal number of clusters for K-means model, using PCA variables (PC1, PC2)*

The elbow point gives us a good number of clusters to choose to balance between cluster separation and avoiding overfitting. Based on the graphs, we choose 3 as our number of state clusters.

Model 1 Results, Discussion, & Conclusion

We now conduct our K-means clustering model based on PCA variables, producing 3 clusters.



Graph 5. *K-Means Clustering plot using original variables (cur_passengers, distance).*

Graph 6. *K-Means Clustering plot using PCA variables (PC1, PC2). States are grouped into 3 categories (1,2,3). We will use graph 6 for their interpretations.*

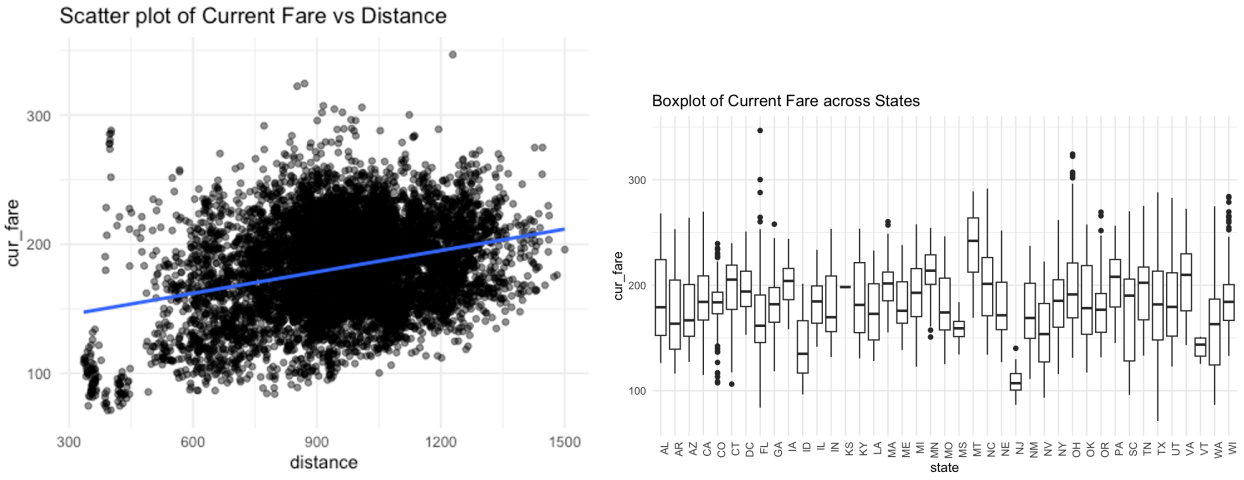
In group 1, states have moderate to positive values of PC1 and PC2. In other words, group 1 represents states with a combination of moderate passenger traffic and distance. Examples of states in this group are MI, IN, LA, UT, and PA.

In group 2, the states have relatively high values of PC1 and low values of PC2. So group 2 represents states with high passenger traffic and relatively shorter distances. Some examples are DC, IL, and CA.

In group 3, the states have negative values of PC1 and PC2. So group 3 represents states with lower passenger traffic and a different distance profile compared to the other groups. This group shows more special patterns in the combination of passenger counts and distances. Some examples are ME, VT, TX, and NC.

Overall, group 1 encapsulates states with a balanced combination of moderate passenger traffic and distance. Meanwhile, Group 2 stands out for its representation of states with high passenger traffic and shorter distances. Lastly, Group 3 signifies states with lower passenger traffic and a unique distance profile. These findings illuminate the diverse patterns in passenger influxes and flight distances across different states. Group 2, with high passenger traffic and shorter distances, might benefit from promotions targeting frequent travelers. States in Group 3 with lower passenger traffic but unique distance profiles could be assessed for potential adjustments in flight schedules or route planning; for example, they might have more connecting flights than others, potential for economic initiatives to capitalize on their strengths and attract more air traffic.

Model 2: Linear Mixed Effects



Graph 7 & 8. *Exploratory visualizations on the relationship between fare, distance, and state*

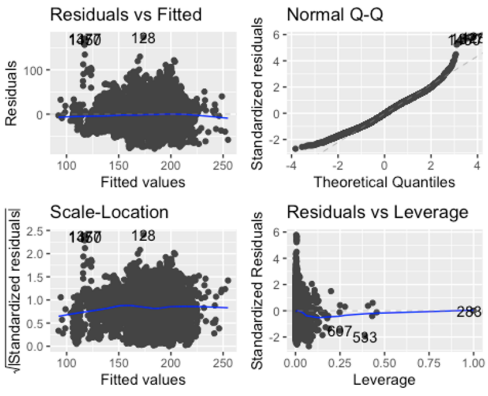
The scatter plot displays the relationship between current fare (cur_fare) and distance. There's a general trend indicating that as distance increases, so does the fare, depicted by the blue line. However, the spread of the points suggests considerable variability around this trend, which could be influenced by many factors not captured solely by distance, such as differences in state, time of booking, or airline pricing strategies. Meanwhile, the boxplot displays current fare variations across different states. As shown, there's considerable overlap in fare ranges between states, but some show distinct differences in medians and variability, indicating state-specific factors may influence airfare.

From these two visualizations, we are inspired to ask a question: Is the impact of distance on airfare uniform across all states? To answer this question, we first employ the *linear fixed effects model* and assign the covariates as following:

- state: Categorical variable -> random effect (since each state might have its own unique influence on the fare).
- distance: numerical variable -> Fixed slope

```
model_full <-
  lm(formula = cur_fare ~ distance*state, data = myData)

summary(model_full)
```

```
##
## Call:
## lm(formula = cur_fare ~ distance * state, data = myData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.019 -22.944  -1.038  21.197 173.176
##
## Coefficients: (1 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.254e+02  5.497e+01   2.281 0.022565 *
## distance          7.809e-02  6.800e-02   1.148 0.250848
## stateAR          -2.808e+02  6.203e+01  -4.526 6.09e-06 ***
## stateAZ           1.245e+00  6.141e+01   0.020 0.983829
## stateCA          -2.497e+01  5.557e+01  -0.449 0.653228
```



Overall, the model seems to be statistically significant enough as the p-value is small, and it reveals several important findings. The estimated baseline fare is approximately $125.40, representing the expected fare when distance and state variables are zero.The effect of distance on fare is positive, with each mile increasing the fare by approximately $0.0781. However, this effect is not statistically significant. Different states have significantly different fares compared to the reference state, with variations in fare levels. The model explains around 33% of the variability in airfare, but the distance effect is not statistically significant, suggesting that other factors may play a more significant role in predicting airfare. From the diagnostic plots, we also see a non-random pattern in the residuals, as they do not appear to be evenly scattered around the horizontal line, which would indicate homoscedasticity. The points in the Q-Q plot also deviate from the straight dashed line, especially at the ends, suggesting that the residuals may not be normally distributed. These mean that our model needs to be adjusted. We do so by using *LASSO regression* to identify key predictors for current airfare.

```
# best lambda to minimize the MSE
bestlam_lasso <- cv.out.lasso $lambda.min
grid <- 10 ^ seq(10, -2, length = 100)
bestlam_lasso
```

```
## [1] 0.002374286
```

```
out.lasso = glmnet(x, y, alpha = 1, lambda = grid)

# Coefficients from lambda chosen by cross validation
lasso_coef = predict(out.lasso, type = "coefficients", s = bestlam_lasso)
lasso_coef
```

```
## 82 x 1 sparse Matrix of class "dgCMatrix"
##                               s1
## (Intercept)       1.264523e+02
## distance          7.511852e-02
## stateAR          -2.652692e+02
## stateAZ          -9.945602e-02
## stateCA          -2.588288e+01
## stateCO          -1.078105e+01
```

Our best lambda is 0.002374286. LASSO regression performs variable selection by shrinking less important variables towards zero, effectively excluding them from the model. From the matrix, we can see that some states and interaction terms have coefficients that are reduced to zero, suggesting that they do not significantly contribute to predicting airfare in this model. The coefficients that remain in the model provide insights into how the selected variables (e.g., distance, certain states) impact airfare. For example, the coefficient for distance is approximately 0.0751, indicating a positive effect on airfare. Some states like ME or PA have positive coefficients, showing deviations from the baseline fare.

Diving deeper, we are curious to see how random effect of 'state' influenced the model - we move to the *linear mixed effects model.* In the linear mixed model, the random effects table reveals substantial variability in the intercept across the 'state' group, suggesting that fares differ by state beyond what is explained by distance alone. The fixed effect of 'distance' is positive, indicating that as distance increases, so does the fare, and the negative correlation between the intercept and 'distance' (-0.595) suggests that states with higher base fares may have a less steep increase in fare with distance.

```
anova(model_rand)
```

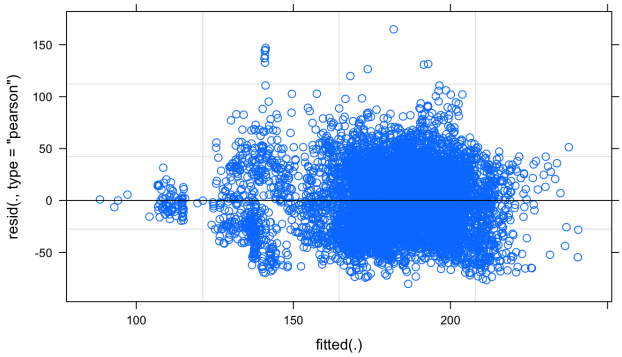
```
## Type III Analysis of Variance Table with Satterthwaite's method
##                Sum Sq Mean Sq NumDF DenDF  F value Pr(>F)
## distance       138446  138446     1  7783 153.8331 <2e-16 ***
## state          216874    5422    40     2   6.0244 0.1524
## distance:state 553379   14189    39  7783  15.7662 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test indicates that the main effect of state alone is not statistically significant in explaining airfare variability. Meanwhile, the main effect of distance and the interaction between distance and state significantly influence airfare are highly significant ($p < 0.001$), indicating that these distance variables have a significant impact on airfare. But we still want to double check with ICC to see if the random effects are significant enough to include in the model. We calculate ICC to be 0.408, meaning the proportion of variance explained by the random intercept (state) is around 41% in this model. The high ICC indicates that the random effect does have a significant impact on the model, thus we should include it. In other words, there are significant differences in airfare between states that cannot be explained by the fixed effects (predictors) in the model.

To summarize, in studying the effects of distance and state in the price of airfare in the US across the years, we have come up with a model that takes into account the stand-alone fixed effect of distance, random effect of state as well as the interaction between distance and state.

```
final_model <- lmer(cur_fare ~ distance + (1 |state), data=myData)

summary(final_model)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: cur_fare ~ distance + (1 | state)
##    Data: myData
##
## REML criterion at convergence: 76533.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5809 -0.7731 -0.0285  0.7239  5.3028
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 355.8    18.86
##  Residual             966.3    31.09
## Number of obs: 7864, groups:  state, 41
##
## Fixed effects:
##              Estimate Std. Error        df t value Pr(>|t|)
## (Intercept) 1.080e+02  3.761e+00 8.966e+01   28.71   <2e-16 ***
## distance    7.830e-02  2.368e-03 7.734e+03   33.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr)
## distance -0.595
```



The linear mixed model with distance as a fixed effect and state as a random effect shows that distance is a significant predictor of current fare ($p < 0.001$), with an average increase of 7.83 cents per unit distance. Meanwhile, the random effects indicate substantial variability in the base fare across states, suggesting that fares do start differently depending on the state. The model fits well, with residuals reasonably distributed, though the presence of outliers is noted.

Overall, going back to our initial question for this model: Is the impact of distance on airfare uniform across all states? We found that distance significantly influences airfare, and there is also significant variation in fares that is attributable to differences between states.

However, while the model demonstrates a significant relationship between distance and airfare, with state as a random effect, the residual plot indicates potential issues with heteroscedasticity and non-linearity, which may impact the precision of our predictions. For future steps, we would look into exploring data transformation, such as log-transformation of the response variable or predictors, to stabilize variance and improve model fit. Additionally, investigating other forms of mixed models or non-linear modeling techniques could provide insights that better capture the complexity of the data. These steps will help in refining the model for more accurate and reliable predictions.

Model 3: Logistics Regression

In this section, I will use logistic regression to investigate the relationship between number of passengers and distance and whether they affect the probability that a fare will be considered affordable. For the purpose of this project, I define affordability as lower than median fare for the current year (or in another word median of the variable cur_fare). From predicting

such relationships, we could even see how airlines can target specific passenger thresholds or flight distances for promotions or pricing adjustments to increase the proportion of fares that are affordable, thus potentially influencing customer behavior and demand.

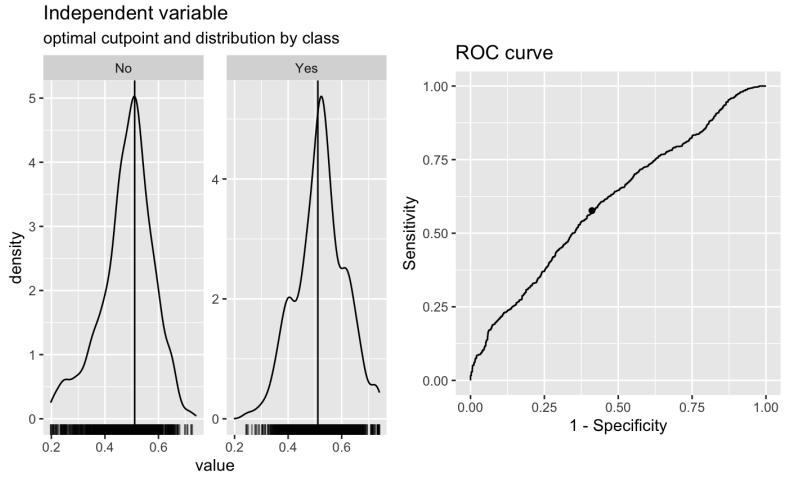
```
Call:
glm(formula = affordable ~ cur_passengers + distance, family = binomial(link = "logit"),
    data = trainData)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.552e+00  1.510e-01  10.279  < 2e-16 ***
cur_passengers -6.596e-08  1.224e-08  -5.388 7.14e-08 ***
distance       -1.471e-03  1.638e-04  -8.977  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7635.7  on 5507  degrees of freedom
Residual deviance: 7424.4  on 5505  degrees of freedom
AIC: 7430.4

Number of Fisher Scoring iterations: 4
```

```
     Confusion Matrix and Statistics

              Reference
     Prediction  No Yes
           No  619 440
           Yes 561 739

                   Accuracy : 0.5757
                     95% CI : (0.5554, 0.5957)
        No Information Rate : 0.5002
        P-Value [Acc > NIR] : 1.213e-13

                      Kappa : 0.1514

      Mcnemar's Test P-Value : 0.0001489

                 Sensitivity : 0.5246
                 Specificity : 0.6268
              Pos Pred Value : 0.5845
              Neg Pred Value : 0.5685
                  Prevalence : 0.5002
              Detection Rate : 0.2624
        Detection Prevalence : 0.4489
           Balanced Accuracy : 0.5757

            'Positive' Class : No
```

From the model result above, we could have a few conclusions:

- While *cur_passengers* has a statistically significant coefficient, its small size suggests that strategies based solely on changing passenger numbers might not significantly impact affordability.
- It is understandable why distance is statistically significant as well as has a more noticeable impact on the fare as the longer one flies the more fuel, energy, labor and cost airlines have to pay. But at the same, fare adjustment is something that airlines could consider for longer flights to increase affordability without compromising overall profitability. (finding a more efficient route to shorten distance, finding alternative energy and so on)
- At the same time, the model accuracy is only 0.5757, which means it correctly predicts affordability about 57.57% of the time. While better than random guessing, there is considerable room for improvement. We could see that other metrics like Sensitivity or Specificity are not considerably high as well.

We also find the optimal cut-off point for the model which is 0.511 with an accuracy of 58.29% (not much improvement in my opinion).



Based on the cut-off point we have the boundary as below.

As accuracy is not very high, there are a lot of misclassifications in this boundary graph. However, I think it is still helpful in a way, for example targeting flights that fall near the decision boundary with promotions or adjustments could potentially make them more appealing by moving them into the affordable category.

5. Conclusions

Our analysis of k-means clustering based on principal components has revealed distinctive patterns among states in terms of passenger traffic and distance. Group 1 encapsulates states with moderate passenger traffic and distance, like MI, IN, LA, UT, and PA. Group 2 stands out with states having high passenger traffic and shorter distances, such as DC, IL, and CA. Lastly, Group 3 signifies states with lower user traffic and unique distance profiles. This group has very few states, including ME, VT, TX, and NC. We now understand the diverse dynamics of passenger influxes and flight distances across the states. We can explore strategies to meet different demands. Group 2 can benefit from promotions targeting frequent travelers. States in Group 3 could be assessed for potential adjustments in flight schedules or route planning as well as economic strengths to attract more air traffic.

Our analysis using a linear mixed model reveals that distance notably impacts airfare, with a significant increase in fare per unit distance. Simultaneously, we observe considerable state-based variability in base fares. Although the model effectively captures these trends, it displays signs of potential heteroscedasticity and non-linearity, as indicated by the residual plot. This finding suggests that the model's current predictions may not be entirely precise. Moving forward, to enhance the model's accuracy and reliability, we plan to explore options like log-transformation of variables and alternative modeling methods, such as different mixed models or non-linear approaches. These refinements aim to better address the complexities

inherent in the data and provide more robust predictions regarding the relationship between distance and airfare across different states.

Lastly, we uncovered the relationship between passenger count, flight distance, and the affordability of airfares using logistic regression. Overall, flight distance exhibits a more substantial influence on fares than the number of passengers, indicating that airlines could enhance affordability through strategic fare adjustments, route optimization, or exploring alternative energy sources. The model, however, is not highly accurate - which means that in the future if we wish to continue this analysis we could find a better alternative model for this question, or try out a few other different thresholds of what's considered affordable. I also think that collecting other information like operational cost like maintenance, staffing, and services could also help give more insights in classifying affordability.

6. References

Department of Transportation Office of the Assistant Secretary for Aviation and International Affairs. (2023, October 16). *Consumer Airfare Report: Table 2 - top 1,000 city-pair markets: Department of Transportation - Data Portal*. DOT Open Data Catalog. https://data.transportation.gov/Aviation/Consumer-Airfare-Report-Table-2-Top-1-000-City-Pai/wqw2-rjgd/about_data

Code Deliverables:
Model1_KMeans Clustering.html
Model2_Linear Mixed Effects.html
Model3_Logistics Regression.html