# A Statistical Analysis of Flight Price Data

**Student Name: Prativa Basnet**

**Course Name: STAT 7100 - Statistical Methods**

**Professor Name: Dr. Kimberly Gardner**

**Fall 2022**

KENNESAW STATE
UNIVERSITY

## Table of Contents



## List of Tables



## List of Figures

## 1. Introduction

The main objective of this project report is to provide the statistical analysis of collection of flight prices data of different destinations from Hartsfield Jackson Atlanta International Airport using different statistical methods. The flight prices data of four different destinations from Atlanta Airport were extracted from Kaggle. This report provides the detailed data analysis using one-way ANOVA, simple linear regression, and Chi-square test for homogeneity.

I travel internationally and domestically quite often. I always wonder about the factors that affect the flight price and also wonder about the best reasonable price. Hence, I am interested in analyzing the collection of airfare data to know the important factors that affect the flight price. There are many factors that have appreciable effect on the flight price. The important factors are the distance, season, flight time, flight types (non-stop/stop), competition among airlines at departure location, number of available flights, seats types ( economy/Business class), etc.

Statistical analysis presented in this report only focuses on collected flight price data of a single day of April 17, 2022 for economy airfare (seat type), multiple destinations, and stop /non-stop flight types. This flight price data is a result of search date of April 16th, 2022 in Expedia.

The collected flight price data was analyzed using three methods previously mentioned.

After this project, I anticipate to improve by analytical skills using one-way ANOVA, simple linear regression, and chi-square test for homogeneity.

The Hartsfield Jackson Atlanta International Airport is referred to as ATL hereafter for convenience. The airfare and flight price words are used interchangeably.

## 2. Dataset Description

The flight price data used in this report are extracted from Kaggle, a Data Science Company. This flight price data is only for a flight date of April 17th, 2022 and search date of April16th, 2022 in Expedia. There are total of 8,259 flight price data from different destinations and departures within U.S on this specific day.

Total of 8,259 flight price data are reduced to 188 by selecting a starting ATL airport and four destinations DTW, JFK, LAX, & MIA airports.

The DTW, JFK, LAX, and MIA are airport codes used for Detroit Metropolitan, John F. Kennedy, Los Angeles, and Miami airports, respectively.

The 12 out of 188 flight price data are missing total travel flight distance. The 176 flight price data will be used for analysis that requires flight distance values.

There are a total of 27 variables for the flight price data. The most relevant variables for this study are listed in Table 1.

## 3. Variable Definitions

**Table 1** below shows the flight price data variable names and variable types. The “DestinationAirport” and “IsNonstop” are qualitative variables. The Basefare & TotalTravelDistance are quantitative variables.

The “DestinationAirport” variable is the arrival airports, i.e., DTW, JFK, LAX, & MIA from a starting airport, i.e., ATL. The “isNonStop” variable is the flight types, i.e., stop or nonstop. The baseFare variable is the base price without tax and other associated fees. Similarly, totalTravelDistance is the distance between the starting and destination airports.

The basefare is given in dollars and totalTravelDistance is given in miles. The total distance of a non-stop flight is the actual flight distance between two airports. However, the total distance of a flight with stop is the sum of the flight distances between all the airports.

Table 1: Description of Flight Price Data Variables

| Variable | Variable type | Values |
| --- | --- | --- |
| destinationAirport | Qualitative | Geographical airport Location within USA |
| isNonStop | Qualitative | Flight types : stop or nonstop |
| baseFare | Quantitative | Base price of a flight in dollars |
| totalTravelDistance | Quantitative | Total travel distance in miles |

## 4. Method A: One-way ANOVA

One –way ANOVA test method is used to compare the mean values of two or more independent groups to see if there is statistical evidence showing significant differences in mean values. The assumptions used by One-way ANOVA test method are;

(a) Sample is randomly independent.

(b) Distribution of response variable is normal within each population.

(c) All populations have approximately equal variance.

The collected flight price data is divided into four different groups based on departure and destination airports. The ATL is the departure airport, and DTW, JFK, LAX, & MIA are the four destination airports. The total number of flight price data for each group (destination) is shown in Table **2.**

The software SAS 9.4 is used for this statistical analysis using one-way ANOVA test method. The output of one-way ANOVA using SAS 9.4 is shown in Figure 1.

Table 2: Mean & Standard Deviation of Flight Prices Data of Four Destinations

| Level of destinationAirport | N | baseFare | |
|---|---|---|---|
| | | Mean | Std Dev |
| DTW | 76 | 404.386579 | 159.338245 |
| JFK | 20 | 315.999000 | 154.007285 |
| LAX | 64 | 460.405625 | 154.695811 |
| MIA | 28 | 334.487143 | 132.423228 |

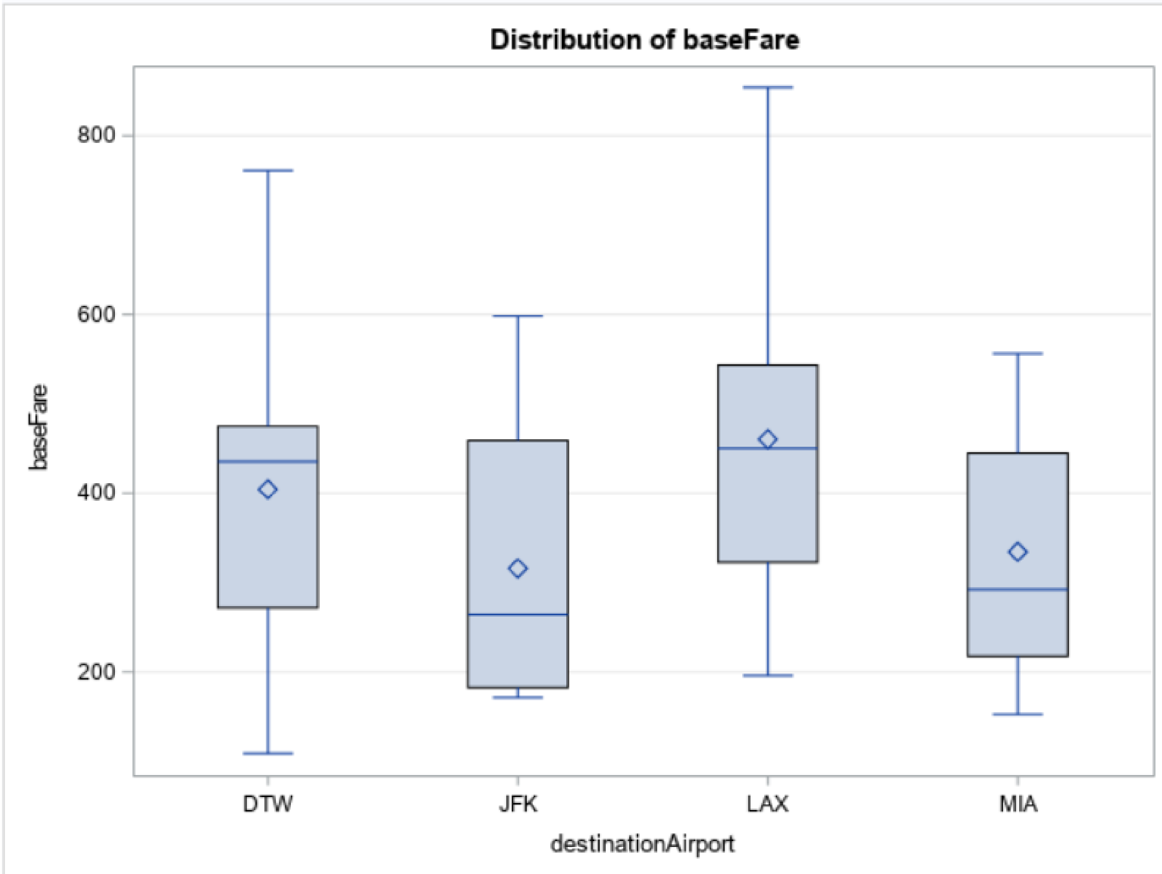


Figure 1: Box Plots of Flight Prices for Four Destination Airports

The samples are assumed to be independent. The boxplots presented in Figure 1 shows evidence of no outliers. The medians within the boxes are not overlapping the quartiles marks. The whiskers are also reasonably the same size. The lengths of the boxplots are also reasonably the same.

In addition, using the brute rule of thumb, it is evident that the largest standard deviation of the DTW airport is less than two times the smallest standard deviation of MIA airport. Therefore, the assumption of normality and equal variance is validated.

## Hypotheses for One-way ANOVA

The hypotheses for one-way ANOVA are described below. .

| Notation | Interpretation |
|---|---|
| $H_0$: $\mu_{DTW}=\mu_{JFK}=\mu_{LAX}=\mu_{MIA}$ | The average flight price is statistically same for all the destinations. |
| HA: At least one flight price mean is significantly different | The average price of at least one destination is significantly different from others. |

The level of significance ($\alpha$) used for this analysis is 5%. The null hypothesis is rejected if p-value for the F-statistic is less than 5%.

## Test Statistic and p-value

Per SAS output results shown in Table 3 below, p-value is 0.0002 which is less than $\alpha=0.05$. Hence, the null hypothesis is rejected. Therefore, there is at least one destination airport that has a significantly different mean value of flight prices.

Table 3: Test Statistic and P-value Outputs from SAS

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 493790.066 | 164596.689 | 6.98 | 0.0002 |
| Error | 184 | 4335906.979 | 23564.712 | | |
| Corrected Total | 187 | 4829697.045 | | | |

| R-Square | Coeff Var | Root MSE | baseFare Mean |
|---|---|---|---|
| 0.102240 | 38.03060 | 153.5080 | 403.6434 |

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| destinationAirport | 3 | 493790.0662 | 164596.6887 | 6.98 | 0.0002 |

Since it is proven that there is at least one destination airport with a significantly different mean, the analysis is required to be continued. Next, pair-wise comparisons are used to determine which means are significantly different from others. The Tukey-Kramer method is used to determine which destination airport had a significantly different mean.

## Test decision

There are twelve pair-wise comparisons between four destinations, see Table 4. Per SAS output with Tukey-Kramer method shown in Table 4, there are two pair-wise comparisons with significantly different mean value. The pair-wise comparisons with significantly different mean value don't contain zero in the confidence interval and have stars (***).

The result shows there is a significant difference in mean for LAX-MIA and LAX-JFK in pair-wise comparisons. It can also be seen that the lower limit of the 95% confidence interval is greater than 0 for both pair-wise comparisons (LAX-MIA & LAX-JFK).

The lower and upper limits of 95% confidence interval for LAX-MIA are 35.74 and 216.10. Similarly, the lower and upper limits of 95% confidence interval for LAX-JFK are 42.45 and 246.36.

Table 4: SAS Output Showing Difference between Mean & Confidence Limits

**Tukey's Studentized Range (HSD) Test for baseFare**

**Note:** This test controls the Type I experimentwise error rate.

| | |
|---|---|
| Alpha | 0.05 |
| Error Degrees of Freedom | 184 |
| Error Mean Square | 23564.71 |
| Critical Value of Studentized Range | 3.66658 |

Comparisons significant at the 0.05 level are indicated by ***.

| destinationAirport Comparison | Difference Between Means | Simultaneous 95% Confidence Limits | | |
|---|---|---|---|---|
| LAX - DTW | 56.02 | -11.50 | 123.54 | |
| LAX - MIA | 125.92 | 35.74 | 216.10 | *** |
| LAX - JFK | 144.41 | 42.45 | 246.36 | *** |
| DTW - LAX | -56.02 | -123.54 | 11.50 | |
| DTW - MIA | 69.90 | -18.09 | 157.88 | |
| DTW - JFK | 88.39 | -11.63 | 188.41 | |
| MIA - LAX | -125.92 | -216.10 | -35.74 | *** |
| MIA - DTW | -69.90 | -157.88 | 18.09 | |
| MIA - JFK | 18.49 | -98.03 | 135.01 | |
| JFK - LAX | -144.41 | -246.36 | -42.45 | *** |
| JFK - DTW | -88.39 | -188.41 | 11.63 | |
| JFK - MIA | -18.49 | -135.01 | 98.03 | |

From above one-way ANOVA method, it is proven that some destination airports have significantly different mean flight price.

## 5. Method B: Simple Linear Regression

Simple linear regression analysis is used to study the relationship between two quantitative variables. The totalTravelDistance and baseFare are the two quantitative variables. The assumptions used in simple linear regression are linearity, equal variance, and normality. There are a total of 188 flight price data with baseFare and totalTravelDistance variables. However, twelve flight price data are missing a value of total travel distance.

The scatter plot using SAS 9.4 is shown in Figure 2 which shows the scatter plot of total travel distance vs baseFare. This plot shows base fare increases with inceasing total travel distance. Hence, there is a positive linear association between these two variables. However, data are not close to the regression line resulting in weak data strength.

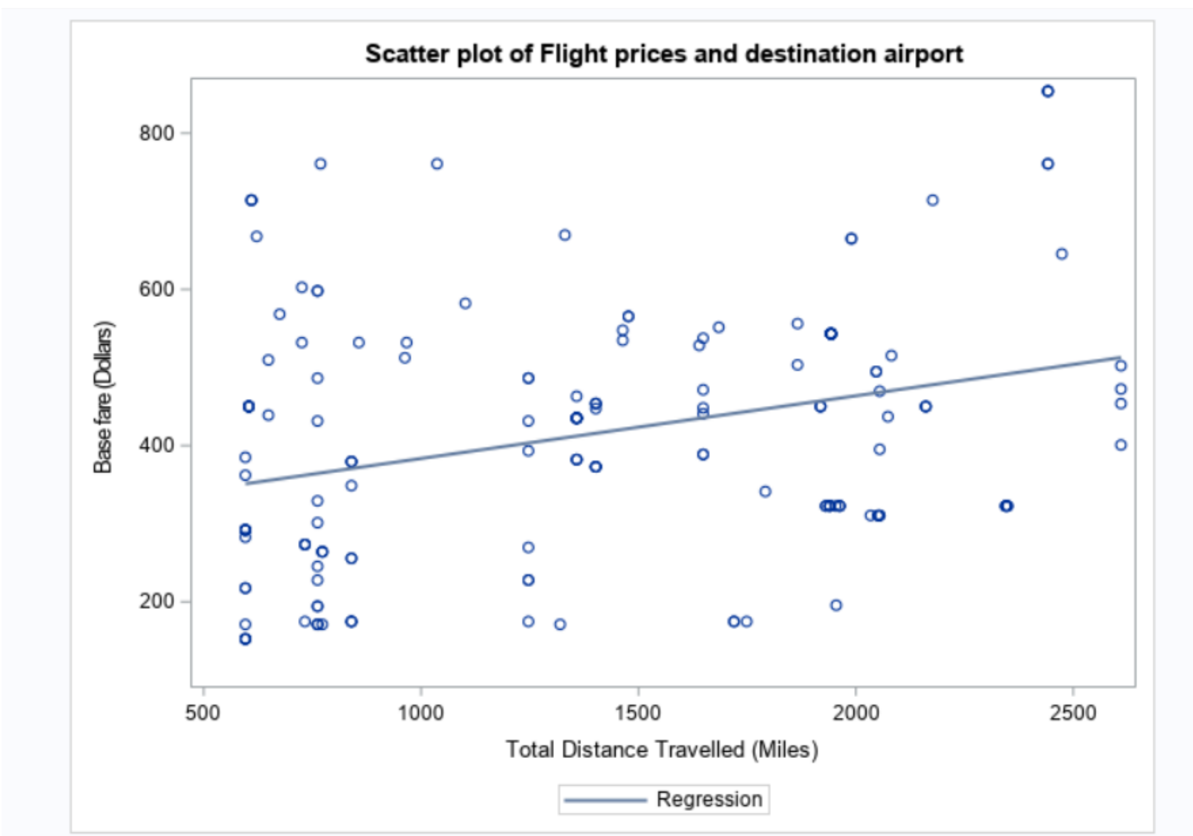


Figure 2: Scatter Plot of totalTravelDistance vs baseFare

In order to see how closely the base fare is related to the total travel distance, the correlation calcualtion is conducted. If the correlation coeffcient is close to -1 or 1 , the variables (baseFare/totalTravelDistance) have strong linear relationship. If the correlation coeffcient is close to zero, the correlation is weak.

The analysis shows the correlation between baseFare and total distance traveled is 0.315 which is between -0.5 and 0.5 and not close to zero. Hence, variables have some moderate correlations.

Table 5: SAS Output of Correlation Data between totalTravelDistance and baseFare

**The CORR Procedure**

| | |
|---|---|
| **2 Variables:** | totalTravelDistance baseFare |

**Simple Statistics**

| Variable | N | Mean | Std Dev | Sum | Minimum | Maximum | Label |
|---|---|---|---|---|---|---|---|
| totalTravelDistance | 176 | 1386 | 620.85409 | 243922 | 596.00000 | 2610 | totalTravelDistance |
| baseFare | 188 | 403.64340 | 160.70861 | 75885 | 109.00000 | 853.95000 | baseFare |

**Pearson Correlation Coefficients**
**Prob > |r| under H0: Rho=0**
**Number of Observations**

| | totalTravelDistance | baseFare |
|---|---|---|
| **totalTravelDistance**<br>totalTravelDistance | 1.00000<br><br>176 | 0.31494<br><.0001<br>176 |
| **baseFare**<br>baseFare | 0.31494<br><.0001<br>176 | 1.00000<br><br>188 |

The scatter plot follows general linear pattern and variables have sufficient corelation. Hence, the least square regression line is fitted to perform the inference tests.

Figure 3a shows the residual vs TotalTravelDistance plot. It can be seen that the residuals are randomly scattered about the zero line and have no distinguishable patterns. The residuals are normal. Hence, equal variance assumption is met.

Figure 3b shows the residual vs quantile plot. It can be seen that the data are close to the regression line and don’t appear to be curvy. Similarly, Figure 3c shows the percent vs residual plot. It can be seen that the histogram is approximately normally distributed.

Hence, the assumption of linearity, normality, and equal variance is met.

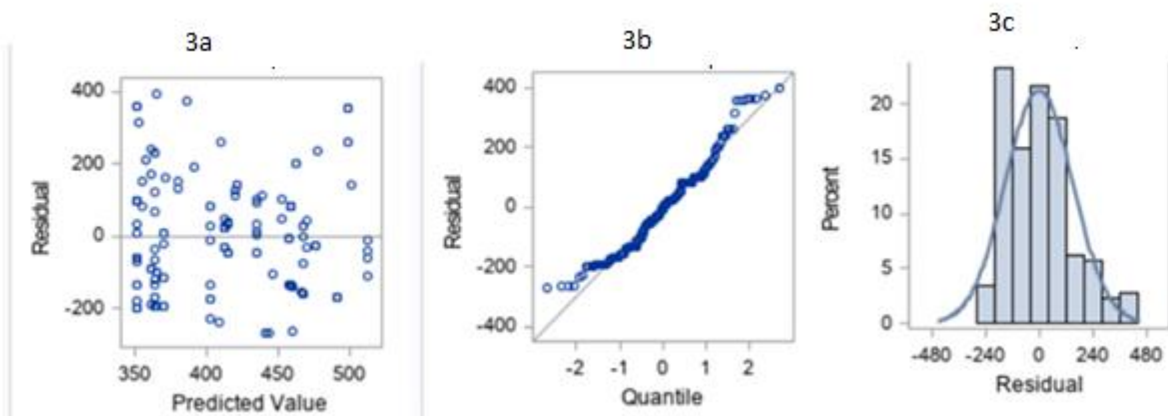


Figure 3: Residual Plots – Simple Regression Analysis

Next, the ANOVA test is used to determine whether the linear relationship is statistically significant. The ANOVA test is often called a model utility test. The hypotheses of the model utility test are given below.

$H_0$: $\beta0=\beta1=0$, All betas are zeros if no significant linear relationship.

$H_A$: At least one $\beta i \neq 0$, there is a significant linear relationship.

The level of significance ($\alpha$) is 5%. The null hypothesis is rejected if p-value for the F-statistic is less than 5%.

Table 6 shows SAS output for simple linear regression analysis. The p-value < 0.0001 which is less than $\alpha$ = 0.05. Hence, null hypothesis is rejected. Therefore, at least one of the betas does not equal 0. At least one has good ability to predict the response.

Table 6: SAS Output of Regression Analysis

The REG Procedure
Model: MODEL1
Dependent Variable: baseFare baseFare

| | |
|---|---|
| Number of Observations Read | 188 |
| Number of Observations Used | 176 |
| Number of Observations with Missing Values | 12 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 434541 | 434541 | 19.16 | <.0001 |
| Error | 174 | 3946609 | 22682 | | |
| Corrected Total | 175 | 4381151 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 150.60432 | R-Square | 0.0992 |
| Dependent Mean | 414.46000 | Adj R-Sq | 0.0940 |
| Coeff Var | 36.33748 | | |

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
|---|---|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 303.22400 | 27.83392 | 10.89 | <.0001 | 248.28844 | 358.15957 |
| totalTravelDistance | totalTravelDistance | 1 | 0.08026 | 0.01834 | 4.38 | <.0001 | 0.04407 | 0.11645 |

Since model utility test verified at least one beta has good ability to predict the response variable, slope test and confidence interval are conducted. The hypotheses of the slope test are given below.

$H_0$: $\beta_1=0$, No significant contribution to the linear relationship.

$H_A$: $\beta_1 \neq 0$, Contributes significantly to the linear relationship.

Since the p-value < 0.0001 (see Table 6) which is less than alpha=0.05, reject the null hypothesis. Therefore, it is concluded that at least one of the slope parameter is not zero and it contributes significantly to the linear relationship. The slope = .0802 is within the 95% confidence limits 0.044 and 0.116.

The equation of the estimated regression line is given below.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 303.22 + 0.080x$$

Since it is confirmed that the model sufficiently fits the data, the reliable prediction is made next. Let x be the predicted value of the independent variable (totalTravelDistance) in the estimated regression line. The mean predicted value is assumed to be x=1385.92.

The 95% confidence interval of the true mean price of baseFare is between 392.05 and 436.87, see Table **7**. When the total distance travel is 1385.92 miles, the base fare is between $392.05 and $436.87.

The 95% predicted confidence interval of baseFare is between 116.37 and 712.55, see Table **7**. When the total distance travel is 1385.92 miles, the predicted base fare is between $116.37 and $712.55.

Table 7: SAS Output of REG Procedure

**Model: MODEL1**
**Dependent Variable: baseFare baseFare**

**Output Statistics**

| Obs | name | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Mean | | 95% CL Predict | | Residual |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | 109 | . | . | . | . | . | . | . |
| 2 | | 109 | . | . | . | . | . | . | . |
| 3 | | 175 | 362.0557 | 16.4990 | 329.4917 | 394.6196 | 63.0308 | 661.0805 | -187.1757 |
| 4 | | 175 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -195.7636 |
| 5 | | 175 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -195.7636 |
| 6 | | 175 | 443.6013 | 13.1605 | 417.6265 | 469.5761 | 145.2221 | 741.9805 | -268.7213 |
| 7 | | 175 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | -228.4300 |
| 8 | | 175 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -195.7636 |
| 9 | | 175 | 441.2737 | 12.8997 | 415.8138 | 466.7337 | 142.9389 | 739.6085 | -266.3937 |
| 10 | | 175 | 441.2737 | 12.8997 | 415.8138 | 466.7337 | 142.9389 | 739.6085 | -266.3937 |
| 11 | | 228 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | -175.4000 |
| 12 | | 228 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | -175.4000 |
| 13 | | 256 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -114.8336 |
| 14 | | 256 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -114.8336 |
| 181 | | 441 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | 5.3549 |
| 182 | | 448 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | 12.7949 |
| 183 | | 472 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | 36.0549 |
| 184 | | 504 | 452.9919 | 14.3656 | 424.6387 | 481.3451 | 154.3962 | 751.5875 | 50.5181 |
| 185 | | 535 | 420.7268 | 11.4422 | 398.1435 | 443.3101 | 122.6237 | 718.8299 | 114.1532 |
| 186 | | 538 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | 102.1049 |
| 187 | | 548 | 420.7268 | 11.4422 | 398.1435 | 443.3101 | 122.6237 | 718.8299 | 127.1732 |
| 188 | | 556 | 452.9919 | 14.3656 | 424.6387 | 481.3451 | 154.3962 | 751.5875 | 103.2881 |
| 189 | x1 | . | 414.4600 | 11.3522 | 392.0542 | 436.8658 | 116.3703 | 712.5497 | . |

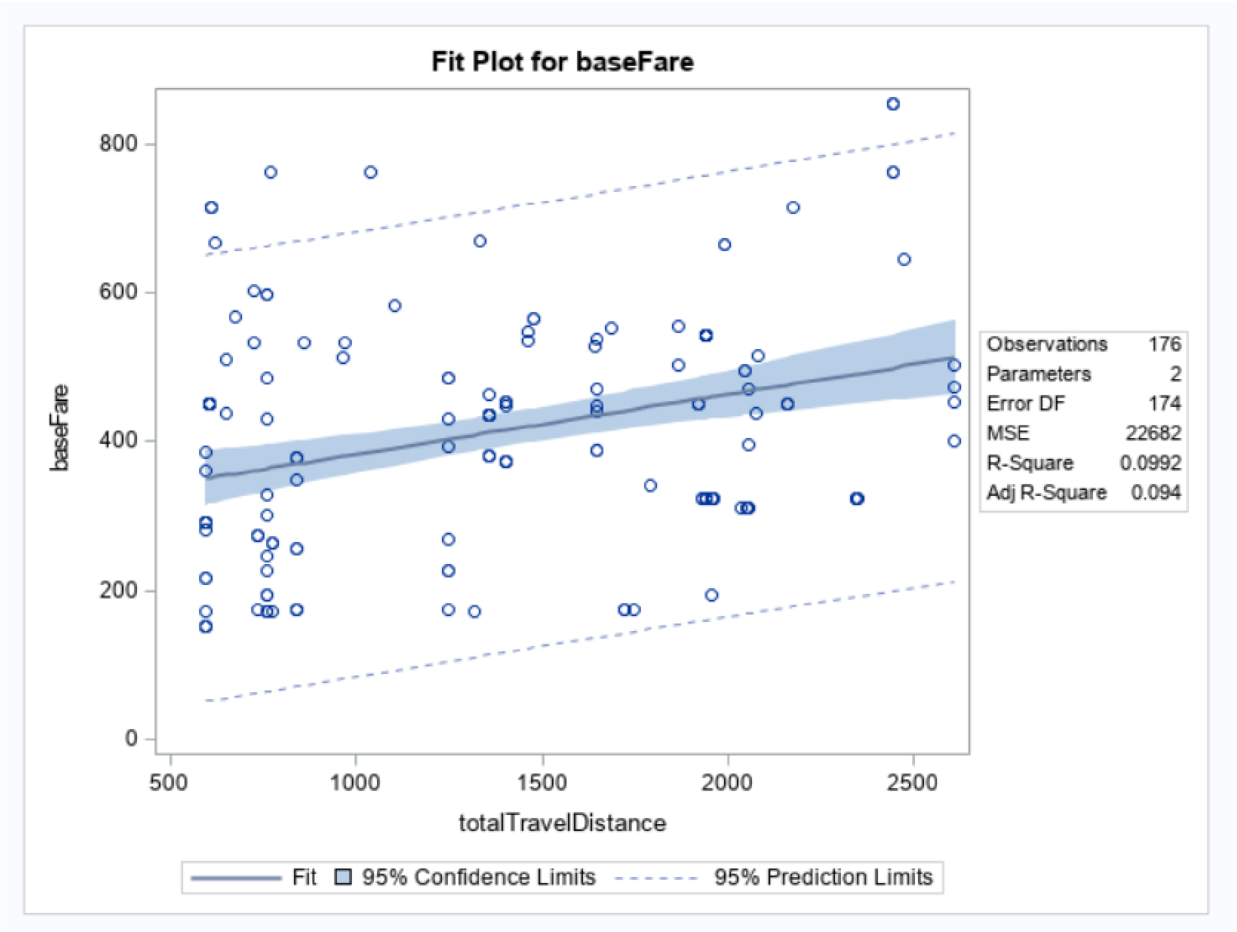


Figure 4: Fit Plot for baseFare vs totalTravelDistance

From this simple regression analysis, it is proven that there is a linear relation between distance travel and base fare. The base price increases with the increase in travel distance.

## 6. Method C: Chi-square Test for Homogeneity

The Chi-square test for homogeneity is used to compare the distribution of different populations. The observed values ($O_{ij}$) are recorded by counting the number of stop and nonstop flights separately. The flight price data used for this chi-square test have two categorical variables which are "isNonstop" and "Destination".

The assumption of chi-square test of homogeneity is that the samples are drawn from different populations and expected value (Eij) is at least 5.

Table 8 below shows the number of observed values (Oij) for all four destinations from ATL airport. A total count of observed values (Oij) for a given destination is determined based on whether the flight type is stop or nonstop. The Dataset for DTW destination, as an example, shows there are total of 76 flight price data from ATL airport with 66 stop and 10 nonstop flights. The counting of nonstop and stop is automatically done for each destination separately by SAS program.

The expected value (Eij) is calculated using a formula shown below.

$$E_{ij} = \frac{\text{Row Total x Column Total}}{\text{Grand Total}}$$

The grand total is equal to 188 and is basically the same as total number of flight price data including all four destinations. The column total is shown in Table 8 below which depends on the flight types stop or nonstop. The row total is the same as the total flight price data of each destination.

The calculation shows the expected value (Eij) of all four destinations regardless of the flight type is greater than 5. Hence, the assumption of chi-square test of homogeneity is met.

Table 8: Observed ($O_{ij}$) & Expected ($E_{ij}$) Values of Categorical Variables

| Starting-Destination airport | Total Flights | IsNonstop Observed Value ($O_{ij}$) | | Row Total | Expected value (Eij ) | |
|---|---|---|---|---|---|---|
| | | False (Stop flights) | True (nonstop flights) | | False (Stop flights | True (nonstop flights) |
| ATL-DTW | 76 | 66 | 10 | 76 | 54 | 21 |
| ATL-JFK | 20 | 10 | 10 | 20 | 14 | 6 |
| ATL-LAX | 64 | 46 | 18 | 64 | 46 | 18 |
| ATL-MIA | 28 | 12 | 16 | 28 | 20 | 8 |
| | Gran total = 188 | Column total = 134 | Column total = 54 | | | |

## Hypotheses for Chi-square Test of Homogeneity

The hypotheses for Chi-square test of homogeneity are given below.

Null hypothesis ($H_0$): Distribution of flight price data of different destinations is the same.

Alternative ($H_A$): At least one of the distribution of flight price data has different distribution.

The level of significance ($\alpha$) used for this analysis is 5%. The null hypothesis is rejected if p-value for the chi-square statistic is less than 5%.

## Test statistic and p-value

Table 9 shows SAS output for chi-square of homogeneity. The table shows frequency, percent, row Pct, and column Pct calculations for each destination. Only the row Pct is used for this analysis.

The Chi-square Statistic is calculated using a formula shown below.

$$\chi 2 = \sum_{all\ cells} \frac{(Observed(\mathrm{Oij}) - Expected(\mathrm{Eij}))^2}{Expected(\mathrm{Eij})}$$

The DF (degree of freedom) is given as, DF = (I-1) (J-1), where I is number of rows and J is number of column.

From Table 9,$\chi 2 = 24.47$, DF = 3, & p-value < 0.0001.

The p-value is less than 0.05. Hence, null hypothesis is rejected. Therefore, there is at least one destination airport that has a different distribution than others.

Table 9: Chi-square Test of Association between Destinations and Flight Types

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of StartDestinationAirport by IsNonstop

| StartDestinationAirport | IsNonstop False | IsNonstop True | Total |
|---|---|---|---|
| ATL-DTW | 66<br>35.11<br>86.84<br>49.25 | 10<br>5.32<br>13.16<br>18.52 | 76<br>40.43 |
| ATL-JFK | 10<br>5.32<br>50.00<br>7.46 | 10<br>5.32<br>50.00<br>18.52 | 20<br>10.64 |
| ATL-LAX | 46<br>24.47<br>71.88<br>34.33 | 18<br>9.57<br>28.13<br>33.33 | 64<br>34.04 |
| ATL-MIA | 12<br>6.38<br>42.86<br>8.96 | 16<br>8.51<br>57.14<br>29.63 | 28<br>14.89 |
| Total | 134<br>71.28 | 54<br>28.72 | 188<br>100.00 |

Statistics for Table of StartDestinationAirport by IsNonstop

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 3 | 24.4737 | <.0001 |
| Likelihood Ratio Chi-Square | 3 | 24.2681 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 15.0490 | 0.0001 |
| Phi Coefficient | | 0.3608 | |
| Contingency Coefficient | | 0.3394 | |
| Cramer's V | | 0.3608 | |

## Test decision

Figure 5 shows the distribution plot of all the flight price data of all four destinations. The flight price ATL-DTW and ATL-LAX have different distributions than others.

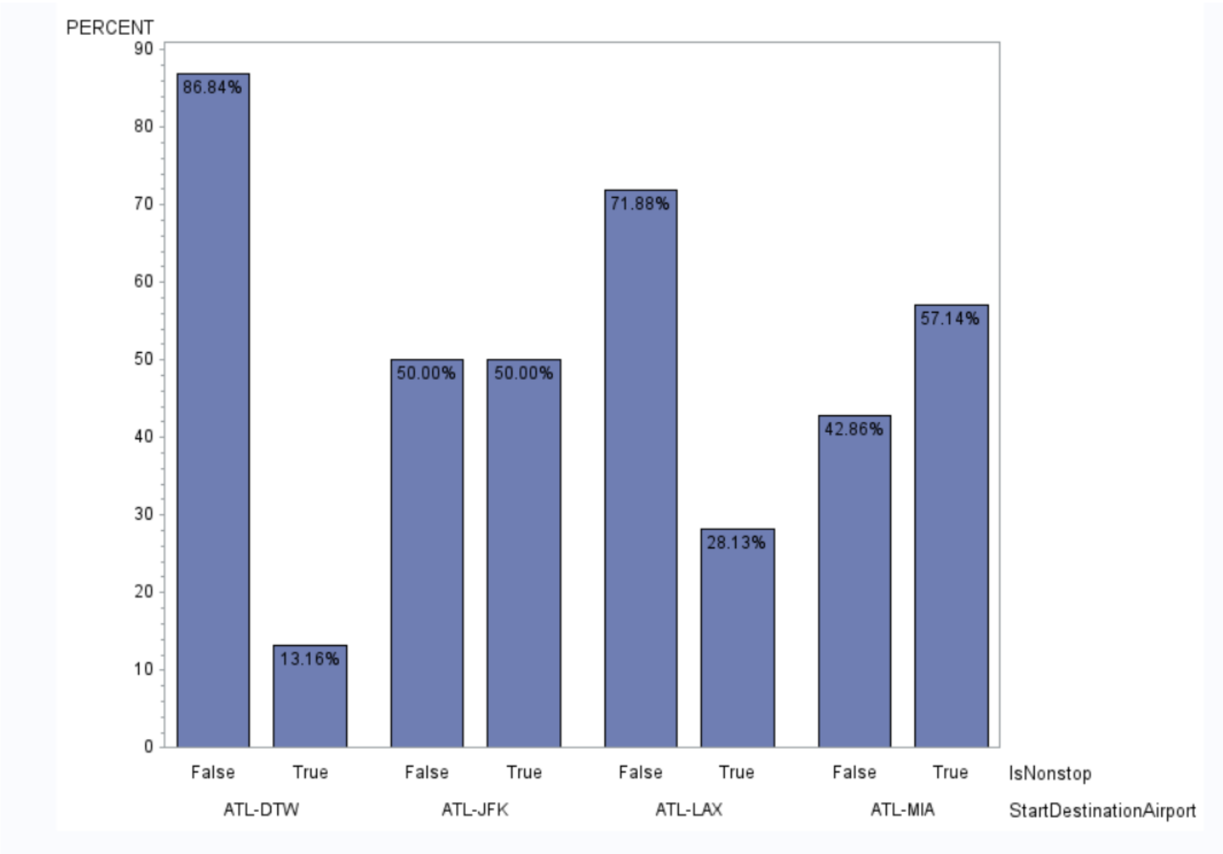


Figure 5: Bar Graph of Different Destinations with IsNonstop (True/False)

The chi-square test shows there is at least one destination airport with different distribution. As shown previously, ATL-DTW and ATL-LAX have different distribution than others

## 7. Conclusions

A total of 188 flight price data was analyzed using one-way ANOVA, simple linear regression, and Chi-square test for homogeneity.

First, the one-way ANOVA test was performed to verify some destinations have significantly different mean base fare. One-way ANOVA test shows the mean flight price of LAX is significantly different from MIA. Similarly, the mean flight price of LAX is also significantly different from JFK.

Secondly, simple linear regression analysis was performed to verify there is a linear relationship between the response and explanatory variables. This analysis shows there is a linear relationship between the distance travel (explanatory) and base fare (response) variables. It also shows increasing travel distance increases the base fare.

Thirdly, chi-square test for homogeneity is performed to verify that some destinations have different flight price distribution. This test shows there are two destinations with different distribution. The ATL to DTW and ATL to LAX flight price data have different distribution than others.

From this project, i increased my knowledge on data analysis using one-way ANOVA test, simple linear regression, and chi-square for homogeneity using SAS code. I also learned how the flight price changes with total travel distance and flight types (stop/nonstop).

As mentioned previously, there are total of twelve missing data of total travel distance from all four destinations. This is the defect that exists in the input data. One could make some improvement by adding the missing data to improve one-way ANOVA test.

The following are three futures recommend studies that can be done to expand this project analysis.

I. Analyze the flight price data based on specific airlines and compare the outcomes

II. Analyze the flight price data based on specific flight time and compare the outcomes

III. Analyze data using others starting (Departure) airports.

## 8. References

Navidi, W. (2020). *Statistics for Engineers and Scientists* (5th ed.). New York, NY: McGraw Hill Education.

Waller, J. L., & Johnson, M. H. (2013). *SAS global forum 2 0 1 3 statistics and data anal y sis*. https://support.sas.com/en/support-home.html. Retrieved December 7, 2022, from https://support.sas.com/resources/papers/proceedings13/430-2013.pdf

Wong, D. (2022, October 18). *Flight prices*. Kaggle. Retrieved December 7, 2022, from https://www.kaggle.com/datasets/dilwong/flightprices

## 9. Appendix A: SAS CODE

SAS Code for Cleaning Data:

```
/* Create the permanent library*/

libname FinalPro"C:\Users\prati\OneDrive\Desktop\MSAS\KSU_Fall_22\STAT 
7100\Project";

/* Import dataset from excel*/

options validvarname=v7;
proc import
datafile="C:\Users\prati\OneDrive\Desktop\MSAS\KSU_Fall_22\STAT 
7100\Project\Flight_Prices_Project.xlsx"
out=FinalPro.FlightPrice
dbms=xlsx
replace;
getnames=yes;
run;

/*Delete all destination airport except DTW, JFK, LAX, & MIA */

data FinalPro.FlightPrice1(keep= startingAirport destinationAirport 
travelDuration isBasicEconomy isNonStop baseFare totalFare 
totalTravelDistance
segmentsAirlineName segmentsDepartureTimeRaw segmentsArrivalTimeRaw 
segmentsArrivalAirportCode segmentsDepartureAirportCode);
set FinalPro.FlightPrice;
      if destinationAirport ="CLT"  then delete;
      else if destinationAirport ="DEN"  then delete;
      else if destinationAirport ="DFW"  then delete;
      else if destinationAirport ="IAD"  then delete;
      else if destinationAirport ="LGA"  then delete;
      else if destinationAirport ="ORD"  then delete;
      else if destinationAirport ="PHL"  then delete;
      else if destinationAirport ="BOS"  then delete;
      else if destinationAirport ="SFO"  then delete;
      else if destinationAirport ="OAK"  then delete;
      else if destinationAirport ="EWR"  then delete;
run;
proc sort data=FinalPro.FlightPrice1 out=FinalPro.FlightPrice2;
by destinationAirport;
run;

/*Delete all starting airport expect Atlanta*/
data FinalPro.FlightPrice2(drop=list);
set FinalPro.FlightPrice2;
      if startingairport ne"ATL" then delete;
      travelDuration=substr(travelDuration,3,5);
      do list = "2022-04-17T",".000-04:00" ,"2022-04-18T",".000-05:00",".000-
06:00",".000-07:00";
      segmentsDepartureTimeRaw=tranwrd(strip(segmentsDepartureTimeRaw),strip(
list),"");
```

```
      segmentsArrivalTimeRaw=
tranwrd(strip(segmentsArrivalTimeRaw),strip(list),"");

      end;
run;
/* create format*/

proc format;
      value myform
            0 ="False"
            1 ="True";
run;

/*Apply Format*/
data FinalPro.FlightPrice2;
set FinalPro.FlightPrice2;
format isBasicEconomy myform. isNonStop myform. ;
run;

/*Just use the airways that have one stop and nonstop*/
data FinalPro.FlightPrice3;
set FinalPro.FlightPrice2;
      if segmentsAirlineName ="Delta||Cape Air||Cape Air||Delta"  then
delete;
      else if segmentsAirlineName ="Frontier Airlines||Frontier
Airlines||Frontier Airlines"  then delete;
      else if segmentsAirlineName ="Spirit Airlines||Spirit Airlines||Spirit
Airlines"  then delete;
      else if segmentsAirlineName ="United||United||Delta"  then delete;

      run;
```

## SAS Code for Statistical Analysis

```
/* One-way anova test*/

proc anova data=FinalPro.FlightPrice3;
  class destinationAirport; /*The explanatory variable*/
  model baseFare = destinationAirport; /*The model for the response
variable*/
 means  destinationAirport ;
run;
quit;

ods graphics on;
proc anova data= FinalPro.FlightPrice3;
  class destinationAirport;
  model baseFare = destinationAirport;
   /*means type / lsd cldiff alpha= 0.05; /*Fisher*/
   /*means type / bon cldiff alpha= 0.05; /*Bonferroni*/
   means destinationAirport / tukey cldiff alpha= 0.05; /*Tukey*/
run;
quit;
ods graphics off;
quit;

/*Simple linear regression*/
/*Check the preliminary assumptions*/
proc univariate data= FinalPro.FlightPrice3 cibasic plot alpha= 0.05;
var baseFare totalTravelDistance;
run;

*A simple scatter plot;
proc sgplot data=FinalPro.FlightPrice3;
  title 'Scatter plot of Flight prices and destination airport';
  reg x=totalTravelDistance y= basefare;
  xaxis label= 'Total Distance Travelled (Miles)';
  yaxis label= 'Base fare (Dollars)';
run;
*Correlation the descriptive statistics;
proc corr data= FinalPro.FlightPrice3 plot= matrix;
var totalTravelDistance baseFare;
run;
/*Generate the least square regression line*/
proc reg data= FinalPro.FlightPrice3;
model basefare=totaltraveldistance / clb alpha= .05; /*response = explanatory
and provide parameter CIs*/
run;
quit;
proc means data=FinalPro.FlightPrice3  mean;
var totaltraveldistance;
run;

/*Generate a predicted value from a given x and calculate CI and PI*/
*Create a data set for the given x;
data predict;
```

```
   input name$ totaltraveldistance;
   datalines;
   x1 1385.92
   ;
run;
proc print data= predict; run;
*Merge the data sets;
data FinalPro.FlightPrice4;
  set FinalPro.FlightPrice3 predict;
 run;
 proc print data= FinalPro.FlightPrice4; run;
 *Generate the CI and PI of the predicted response;
proc reg data= FinalPro.FlightPrice4;
model basefare =totaltraveldistance / clm cli alpha = 0.05; /*response CI and
PI*/
id name;
run;
quit;


/* Categorical Dataset*/
/* Counting of nonstop and onestop*/
 proc freq data=FinalPro.FlightPrice3;
 tables destinationAirport*isnonstop/ nocol norow nopercent;
 run;

data Destairp;
input StartDestinationAirport$ IsNonstop$ Count;
datalines;
ATL-DTW False 66
ATL-DTW True 10
ATL-JFK False 10
ATL-JFK True 10
ATL-LAX False 46
ATL-LAX True 18
ATL-MIA False 12
ATL-MIA True 16
;
run;
proc print data=destairp; run;

/*Test of homogenity*/
 proc freq data=destairp;
 tables StartDestinationAirport*IsNonstop/chisq ;
 weight count;
 run;
 proc gchart data= destairp;
 vbar IsNonstop/freq=count type= percent group=StartDestinationAirport
 g100 nozero inside=percent width=20;
 run;
 quit;
```

## 10. Appendix B: SAS CODE OUTPUT

**The REG Procedure**
**Model: MODEL1**
**Dependent Variable: baseFare baseFare**

**Output Statistics**

| Obs | name | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Mean | | 95% CL Predict | | Residual |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | 109 | . | . | . | . | . | . | . |
| 2 | | 109 | . | . | . | . | . | . | . |
| 3 | | 175 | 362.0557 | 16.4990 | 329.4917 | 394.6196 | 63.0308 | 661.0805 | -187.1757 |
| 4 | | 175 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -195.7636 |
| 5 | | 175 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -195.7636 |
| 6 | | 175 | 443.6013 | 13.1605 | 417.6265 | 469.5761 | 145.2221 | 741.9805 | -268.7213 |
| 7 | | 175 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | -228.4300 |
| 8 | | 175 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -195.7636 |
| 9 | | 175 | 441.2737 | 12.8997 | 415.8138 | 466.7337 | 142.9389 | 739.6085 | -266.3937 |
| 10 | | 175 | 441.2737 | 12.8997 | 415.8138 | 466.7337 | 142.9389 | 739.6085 | -266.3937 |
| 11 | | 228 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | -175.4000 |
| 12 | | 228 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | -175.4000 |
| 13 | | 256 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -114.8336 |
| 14 | | 256 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -114.8336 |
| 15 | | 199 | . | . | . | . | . | . | . |
| 16 | | 199 | . | . | . | . | . | . | . |
| 17 | | 270 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | -133.5300 |
| 18 | | 273 | 362.0557 | 16.4990 | 329.4917 | 394.6196 | 63.0308 | 661.0805 | -88.5657 |
| 19 | | 273 | 362.0557 | 16.4990 | 329.4917 | 394.6196 | 63.0308 | 661.0805 | -88.5657 |
| 20 | | 273 | 362.0557 | 16.4990 | 329.4917 | 394.6196 | 63.0308 | 661.0805 | -88.5657 |
| 21 | | 209 | . | . | . | . | . | . | . |
| 22 | | 209 | . | . | . | . | . | . | . |
| 23 | | 349 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | -21.8036 |
| 24 | | 373 | 415.7506 | 11.3561 | 393.3372 | 438.1639 | 117.6603 | 713.8408 | -42.7306 |
| 25 | | 373 | 415.7506 | 11.3561 | 393.3372 | 438.1639 | 117.6603 | 713.8408 | -42.7306 |
| 26 | | 373 | 415.7506 | 11.3561 | 393.3372 | 438.1639 | 117.6603 | 713.8408 | -42.7306 |
| 27 | | 373 | 415.7506 | 11.3561 | 393.3372 | 438.1639 | 117.6603 | 713.8408 | -42.7306 |
| 28 | | 380 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | 8.8964 |
| 29 | | 380 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | 8.8964 |
| 30 | | 380 | 370.6436 | 15.1355 | 340.7708 | 400.5165 | 71.8999 | 669.3874 | 8.8964 |
| 31 | | 382 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | -29.8991 |
| 32 | | 382 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | -29.8991 |

The REG Procedure
Model: MODEL1
Dependent Variable: baseFare baseFare

| Output Statistics | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Obs | name | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Mean | | 95% CL Predict | | Residual |
| 33 | | 382 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | -29.8991 |
| 34 | | 393 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | -9.8300 |
| 35 | | 432 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | 28.3100 |
| 36 | | 435 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | 23.1209 |
| 37 | | 435 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | 23.1209 |
| 38 | | 435 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | 23.1209 |
| 39 | | 435 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | 23.1209 |
| 40 | | 435 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | 23.1209 |
| 41 | | 435 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | 23.1209 |
| 42 | | 435 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | 23.1209 |
| 43 | | 435 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | 23.1209 |
| 44 | | 450 | 351.7019 | 18.2881 | 315.6069 | 387.7970 | 52.2720 | 651.1319 | 98.5281 |
| 45 | | 450 | 351.7019 | 18.2881 | 315.6069 | 387.7970 | 52.2720 | 651.1319 | 98.5281 |
| 46 | | 450 | 351.7019 | 18.2881 | 315.6069 | 387.7970 | 52.2720 | 651.1319 | 98.5281 |
| 47 | | 450 | 351.7019 | 18.2881 | 315.6069 | 387.7970 | 52.2720 | 651.1319 | 98.5281 |
| 48 | | 450 | 351.7019 | 18.2881 | 315.6069 | 387.7970 | 52.2720 | 651.1319 | 98.5281 |
| 49 | | 450 | 351.7019 | 18.2881 | 315.6069 | 387.7970 | 52.2720 | 651.1319 | 98.5281 |
| 50 | | 450 | 351.7019 | 18.2881 | 315.6069 | 387.7970 | 52.2720 | 651.1319 | 98.5281 |
| 51 | | 450 | 351.7019 | 18.2881 | 315.6069 | 387.7970 | 52.2720 | 651.1319 | 98.5281 |
| 52 | | 439 | 355.3137 | 17.6486 | 320.4808 | 390.1465 | 56.0332 | 654.5941 | 83.7563 |
| 53 | | 447 | 415.7506 | 11.3561 | 393.3372 | 438.1639 | 117.6603 | 713.8408 | 31.6894 |
| 54 | | 454 | 415.7506 | 11.3561 | 393.3372 | 438.1639 | 117.6603 | 713.8408 | 38.1994 |
| 55 | | 454 | 415.7506 | 11.3561 | 393.3372 | 438.1639 | 117.6603 | 713.8408 | 38.1994 |
| 56 | | 454 | 415.7506 | 11.3561 | 393.3372 | 438.1639 | 117.6603 | 713.8408 | 38.1994 |
| 57 | | 463 | 412.2191 | 11.3638 | 389.7905 | 434.6476 | 114.1276 | 710.3105 | 51.0309 |
| 58 | | 487 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | 83.2000 |
| 59 | | 487 | 403.3100 | 11.6345 | 380.3471 | 426.2730 | 105.1779 | 701.4422 | 83.2000 |
| 60 | | 513 | 380.5158 | 13.7483 | 353.3810 | 407.6506 | 82.0334 | 678.9982 | 132.0442 |
| 61 | | 510 | 355.3137 | 17.6486 | 320.4808 | 390.1465 | 56.0332 | 654.5941 | 154.4563 |
| 62 | | 528 | 434.8528 | 12.2711 | 410.6334 | 459.0722 | 136.6213 | 733.0843 | 93.5272 |
| 63 | | 532 | 361.4938 | 16.5924 | 328.7456 | 394.2420 | 62.4488 | 660.5388 | 170.5962 |
| 64 | | 552 | 438.4646 | 12.6075 | 413.5812 | 463.3479 | 140.1784 | 736.7507 | 113.1554 |

**The REG Procedure**
**Model: MODEL1**
**Dependent Variable: baseFare baseFare**

| Output Statistics | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Obs | name | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Mean | | 95% CL Predict | | Residual |
| 65 | | 566 | 421.7702 | 11.4744 | 399.1232 | 444.4171 | 123.6622 | 719.8781 | 143.8098 |
| 66 | | 566 | 421.7702 | 11.4744 | 399.1232 | 444.4171 | 123.6622 | 719.8781 | 143.8098 |
| 67 | | 568 | 357.4005 | 17.2863 | 323.2828 | 391.5182 | 58.2024 | 656.5985 | 210.9695 |
| 68 | | 582 | 391.6721 | 12.4891 | 367.0225 | 416.3218 | 93.4054 | 689.9389 | 190.6579 |
| 69 | | 603 | 361.4938 | 16.5924 | 328.7456 | 394.2420 | 62.4488 | 660.5388 | 241.2962 |
| 70 | | 670 | 410.0520 | 11.3968 | 387.5582 | 432.5458 | 111.9557 | 708.1483 | 259.7080 |
| 71 | | 668 | 353.1466 | 18.0305 | 317.5600 | 388.7332 | 53.7775 | 652.5157 | 314.7634 |
| 72 | | 714 | 352.1835 | 18.2020 | 316.2585 | 388.1085 | 52.7740 | 651.5930 | 362.2365 |
| 73 | | 714 | 352.1835 | 18.2020 | 316.2585 | 388.1085 | 52.7740 | 651.5930 | 362.2365 |
| 74 | | 714 | 352.1835 | 18.2020 | 316.2585 | 388.1085 | 52.7740 | 651.5930 | 362.2365 |
| 75 | | 761 | 364.9451 | 16.0264 | 333.3139 | 396.5762 | 66.0203 | 663.8698 | 395.9849 |
| 76 | | 761 | 386.4551 | 13.0311 | 360.7358 | 412.1745 | 88.0981 | 684.8122 | 374.4749 |
| 77 | | 171 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | -193.2232 |
| 78 | | 171 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | -193.2232 |
| 79 | | 171 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | -193.2232 |
| 80 | | 171 | 365.2661 | 15.9747 | 333.7370 | 396.7953 | 66.3522 | 664.1801 | -194.1061 |
| 81 | | 171 | 409.1691 | 11.4164 | 386.6367 | 431.7016 | 111.0699 | 707.2684 | -238.0091 |
| 82 | | 194 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | -169.9632 |
| 83 | | 194 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | -169.9632 |
| 84 | | 228 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | -136.4732 |
| 85 | | 246 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | -118.8032 |
| 86 | | 264 | 365.2661 | 15.9747 | 333.7370 | 396.7953 | 66.3522 | 664.1801 | -101.0761 |
| 87 | | 264 | 365.2661 | 15.9747 | 333.7370 | 396.7953 | 66.3522 | 664.1801 | -101.0761 |
| 88 | | 264 | 365.2661 | 15.9747 | 333.7370 | 396.7953 | 66.3522 | 664.1801 | -101.0761 |
| 89 | | 301 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | -62.9832 |
| 90 | | 329 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | -35.0832 |
| 91 | | 432 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | 67.2468 |
| 92 | | 487 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | 122.1268 |
| 93 | | 532 | 380.8368 | 13.7070 | 353.7834 | 407.8903 | 82.3618 | 679.3119 | 151.2532 |
| 94 | | 532 | 372.0081 | 14.9312 | 342.5385 | 401.4776 | 73.3044 | 670.7118 | 160.0819 |
| 95 | | 598 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | 233.7468 |
| 96 | | 598 | 364.3832 | 16.1173 | 332.5727 | 396.1937 | 65.4395 | 663.3270 | 233.7468 |

**The REG Procedure**
**Model: MODEL1**
**Dependent Variable: baseFare baseFare**

| Output Statistics | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Obs | name | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Mean | | 95% CL Predict | | Residual |
| 97 | | 196 | 460.1352 | 15.4197 | 429.7014 | 490.5689 | 161.3348 | 758.9355 | -264.4152 |
| 98 | | 272 | . | . | . | . | . | . | . |
| 99 | | 272 | . | . | . | . | . | . | . |
| 100 | | 311 | 466.4758 | 16.4347 | 434.0388 | 498.9128 | 167.4647 | 765.4869 | -155.7758 |
| 101 | | 311 | 467.8403 | 16.6615 | 434.9556 | 500.7249 | 168.7803 | 766.9002 | -157.1403 |
| 102 | | 311 | 468.1613 | 16.7153 | 435.1705 | 501.1520 | 169.0897 | 767.2329 | -157.4613 |
| 103 | | 311 | 468.1613 | 16.7153 | 435.1705 | 501.1520 | 169.0897 | 767.2329 | -157.4613 |
| 104 | | 311 | 467.8403 | 16.6615 | 434.9556 | 500.7249 | 168.7803 | 766.9002 | -157.1403 |
| 105 | | 323 | 491.7582 | 20.9940 | 450.3224 | 533.1939 | 191.6376 | 791.8788 | -168.9682 |
| 106 | | 323 | 491.7582 | 20.9940 | 450.3224 | 533.1939 | 191.6376 | 791.8788 | -168.9682 |
| 107 | | 323 | 460.7772 | 15.5193 | 430.1468 | 491.4076 | 161.9568 | 759.5977 | -137.9872 |
| 108 | | 323 | 458.9312 | 15.2349 | 428.8622 | 489.0002 | 160.1678 | 757.6947 | -136.1412 |
| 109 | | 323 | 458.9312 | 15.2349 | 428.8622 | 489.0002 | 160.1678 | 757.6947 | -136.1412 |
| 110 | | 323 | 458.2089 | 15.1253 | 428.3561 | 488.0616 | 159.4671 | 756.9506 | -135.4189 |
| 111 | | 323 | 458.9312 | 15.2349 | 428.8622 | 489.0002 | 160.1678 | 757.6947 | -136.1412 |
| 112 | | 323 | 458.2089 | 15.1253 | 428.3561 | 488.0616 | 159.4671 | 756.9506 | -135.4189 |
| 113 | | 323 | 460.7772 | 15.5193 | 430.1468 | 491.4076 | 161.9568 | 759.5977 | -137.9872 |
| 114 | | 323 | 460.0549 | 15.4073 | 429.6457 | 490.4641 | 161.2570 | 758.8528 | -137.2649 |
| 115 | | 323 | 460.7772 | 15.5193 | 430.1468 | 491.4076 | 161.9568 | 759.5977 | -137.9872 |
| 116 | | 323 | 491.4371 | 20.9324 | 450.1231 | 532.7511 | 191.3333 | 791.5409 | -168.6471 |
| 117 | | 323 | 491.4371 | 20.9324 | 450.1231 | 532.7511 | 191.3333 | 791.5409 | -168.6471 |
| 118 | | 323 | 458.9312 | 15.2349 | 428.8622 | 489.0002 | 160.1678 | 757.6947 | -136.1412 |
| 119 | | 323 | 491.4371 | 20.9324 | 450.1231 | 532.7511 | 191.3333 | 791.5409 | -168.6471 |
| 120 | | 272 | . | . | . | . | . | . | . |
| 121 | | 395 | 468.1613 | 16.7153 | 435.1705 | 501.1520 | 169.0897 | 767.2329 | -72.8113 |
| 122 | | 401 | 512.7064 | 25.1534 | 463.0613 | 562.3515 | 211.3427 | 814.0701 | -111.7764 |
| 123 | | 450 | 457.2457 | 14.9808 | 427.6782 | 486.8133 | 158.5323 | 755.9591 | -7.0157 |
| 124 | | 450 | 457.2457 | 14.9808 | 427.6782 | 486.8133 | 158.5323 | 755.9591 | -7.0157 |
| 125 | | 450 | 457.2457 | 14.9808 | 427.6782 | 486.8133 | 158.5323 | 755.9591 | -7.0157 |
| 126 | | 437 | 469.6863 | 16.9726 | 436.1875 | 503.1850 | 170.5582 | 768.8144 | -32.4763 |
| 127 | | 450 | 476.6690 | 18.1899 | 440.7678 | 512.5703 | 177.2623 | 776.0757 | -26.4390 |
| 128 | | 450 | 476.6690 | 18.1899 | 440.7678 | 512.5703 | 177.2623 | 776.0757 | -26.4390 |

**The REG Procedure**
**Model: MODEL1**
**Dependent Variable: baseFare baseFare**

**Output Statistics**

| Obs | name | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Mean | | 95% CL Predict | | Residual |
|---|---|---|---|---|---|---|---|---|---|
| 129 | | 450 | 476.6690 | 18.1899 | 440.7678 | 512.5703 | 177.2623 | 776.0757 | -26.4390 |
| 130 | | 454 | 512.7064 | 25.1534 | 463.0613 | 562.3515 | 211.3427 | 814.0701 | -58.7564 |
| 131 | | 470 | 468.1613 | 16.7153 | 435.1705 | 501.1520 | 169.0897 | 767.2329 | 1.6087 |
| 132 | | 422 | . | . | . | . | . | . | . |
| 133 | | 473 | 512.7064 | 25.1534 | 463.0613 | 562.3515 | 211.3427 | 814.0701 | -40.1564 |
| 134 | | 495 | 467.5192 | 16.6079 | 434.7404 | 500.2980 | 168.4709 | 766.5675 | 27.3608 |
| 135 | | 495 | 467.5192 | 16.6079 | 434.7404 | 500.2980 | 168.4709 | 766.5675 | 27.3608 |
| 136 | | 502 | 512.7064 | 25.1534 | 463.0613 | 562.3515 | 211.3427 | 814.0701 | -10.3764 |
| 137 | | 462 | . | . | . | . | . | . | . |
| 138 | | 515 | 470.3284 | 17.0820 | 436.6138 | 504.0429 | 171.1760 | 769.4807 | 45.0216 |
| 139 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 140 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 141 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 142 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 143 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 144 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 145 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 146 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 147 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 148 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 149 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 150 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 151 | | 543 | 459.1720 | 15.2716 | 429.0305 | 489.3135 | 160.4013 | 757.9428 | 84.0880 |
| 152 | | 646 | 501.7908 | 22.9556 | 456.4835 | 547.0982 | 201.1113 | 802.4704 | 143.7892 |
| 153 | | 665 | 462.9443 | 15.8611 | 431.6395 | 494.2492 | 164.0539 | 761.8347 | 202.1657 |
| 154 | | 665 | 462.9443 | 15.8611 | 431.6395 | 494.2492 | 164.0539 | 761.8347 | 202.1657 |
| 155 | | 714 | 477.9532 | 18.4201 | 441.5977 | 514.3087 | 178.4917 | 777.4147 | 236.4668 |
| 156 | | 761 | 499.2225 | 22.4475 | 454.9181 | 543.5269 | 198.6924 | 799.7526 | 261.7075 |
| 157 | | 761 | 499.2225 | 22.4475 | 454.9181 | 543.5269 | 198.6924 | 799.7526 | 261.7075 |
| 158 | | 854 | 499.2225 | 22.4475 | 454.9181 | 543.5269 | 198.6924 | 799.7526 | 354.7275 |
| 159 | | 854 | 499.2225 | 22.4475 | 454.9181 | 543.5269 | 198.6924 | 799.7526 | 354.7275 |
| 160 | | 854 | 499.2225 | 22.4475 | 454.9181 | 543.5269 | 198.6924 | 799.7526 | 354.7275 |

**The REG Procedure**
**Model: MODEL1**
**Dependent Variable: baseFare baseFare**

| Output Statistics | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Obs | name | Dependent Variable | Predicted Value | Std Error Mean Predict | 95% CL Mean | | 95% CL Predict | | Residual |
| 161 | | 153 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -198.4998 |
| 162 | | 153 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -198.4998 |
| 163 | | 153 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -198.4998 |
| 164 | | 153 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -198.4998 |
| 165 | | 171 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -179.8998 |
| 166 | | 218 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -133.3898 |
| 167 | | 218 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -133.3898 |
| 168 | | 206 | . | . | . | . | . | . | . |
| 169 | | 283 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -68.2698 |
| 170 | | 292 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -58.9698 |
| 171 | | 292 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -58.9698 |
| 172 | | 292 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -58.9698 |
| 173 | | 292 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -58.9698 |
| 174 | | 292 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -58.9698 |
| 175 | | 292 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | -58.9698 |
| 176 | | 341 | 447.0525 | 13.5765 | 420.2568 | 473.8483 | 148.6008 | 745.5043 | -105.6625 |
| 177 | | 362 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | 11.1602 |
| 178 | | 385 | 351.0598 | 18.4033 | 314.7374 | 387.3823 | 51.6024 | 650.5173 | 34.0602 |
| 179 | | 389 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | -46.7451 |
| 180 | | 389 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | -46.7451 |
| 181 | | 441 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | 5.3549 |
| 182 | | 448 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | 12.7949 |
| 183 | | 472 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | 36.0549 |
| 184 | | 504 | 452.9919 | 14.3656 | 424.6387 | 481.3451 | 154.3962 | 751.5875 | 50.5181 |
| 185 | | 535 | 420.7268 | 11.4422 | 398.1435 | 443.3101 | 122.6237 | 718.8299 | 114.1532 |
| 186 | | 538 | 435.5751 | 12.3347 | 411.2302 | 459.9201 | 137.3334 | 733.8169 | 102.1049 |
| 187 | | 548 | 420.7268 | 11.4422 | 398.1435 | 443.3101 | 122.6237 | 718.8299 | 127.1732 |
| 188 | | 556 | 452.9919 | 14.3656 | 424.6387 | 481.3451 | 154.3962 | 751.5875 | 103.2881 |
| 189 | x1 | . | 414.4600 | 11.3522 | 392.0542 | 436.8658 | 116.3703 | 712.5497 | . |