Jfk to lax

Training

Keeping outliers

Randomized the dataset

If the flight is not non-stop only the takeoff time at JFK and arrival LAX are considered

Stattools only works until 30 categorical variables

|  |  |
| --- | --- |
| **ArrivalAirportCode** | |
| LAX | 1 |
| SFO||LAX | 2 |
| CLT||LAX | 3 |
| SLC||ONT | 4 |
| SFO||ONT | 5 |
| ATL||ONT | 6 |
| DFW||LAX | 7 |
| MIA||LAX | 8 |
| DFW||ONT | 9 |
| ORD||LAX | 10 |
| ONT | 11 |
| FLL||LAX | 12 |
| DFW||ABQ||LAX | 13 |
| DFW||ELP||LAX | 14 |
| LAS||LAX | 15 |
| PHX||LAX | 16 |
| AUS||LAX | 17 |
| PHX||ONT | 18 |
| BOS||LAX | 19 |
| IAD||LAX | 20 |
| PDX||LAX | 21 |
| RDU||LAX | 22 |
| IND||LAX | 23 |
| SLC||LAX | 24 |
| DTW||LAX | 25 |
| PHX||LAS||LAX | 26 |
| BOS||PHX||LAX | 27 |
| CLT||ONT | 28 |
| BUF||LAX | 29 |
| SAV||ATL||ONT | 30 |
| MCO||LAX | 31 |
| TPA||LAX | 32 |
| ORF||ATL||ONT | 33 |

|  |  |
| --- | --- |
| **DepartureAirportCode** | |
| JFK | 1 |
| JFK||SFO | 2 |
| JFK||CLT | 3 |
| JFK||SLC | 4 |
| JFK||ATL | 5 |
| JFK||DFW | 6 |
| JFK||MIA | 7 |
| JFK||ORD | 8 |
| JFK||FLL | 9 |
| JFK||DFW||ABQ | 10 |
| JFK||DFW||ELP | 11 |
| JFK||LAS | 12 |
| JFK||PHX | 13 |
| JFK||AUS | 14 |
| JFK||BOS | 15 |
| JFK||IAD | 16 |
| JFK||PDX | 17 |
| JFK||RDU | 18 |
| JFK||IND | 19 |
| JFK||DTW | 20 |
| JFK||PHX||LAS | 21 |
| JFK||BOS||PHX | 22 |
| JFK||BUF | 23 |
| JFK||SAV||ATL | 24 |
| JFK||MCO | 25 |
| JFK||TPA | 26 |
| JFK||ORF||ATL | 27 |

|  |  |
| --- | --- |
| **Airline** |  |
| American Airlines | 1 |
| Delta | 2 |
| Alaska Airlines||Alaska Airlines | 3 |
| JetBlue Airways | 4 |
| American Airlines||American Airlines | 5 |
| Delta||Delta | 6 |
|  |  |
| Alaska Airlines||United | 7 |
| United||United | 8 |
| Alaska Airlines||Delta | 9 |
| United | 10 |
| JetBlue Airways||JetBlue Airways | 11 |
| American Airlines||American Airlines||American Airlines | 12 |
| Delta||Alaska Airlines | 13 |
| Delta||Delta||Delta | 14 |

|  |  |
| --- | --- |
| **CabinCode** |  |
| coach | 1 |
| coach||coach | 2 |
| coach||coach||coach | 3 |
| first | 4 |
| business | 5 |

Divided the dataset into 75% (5907) and 25% (1969) for model training and validation.

|  |  |  |
| --- | --- | --- |
| **Statistics** | | |
| total fare | | |
| N | Valid | 5907 |
| Missing | 0 |
| Mean | | 337.9151 |
| Median | | 322.6000 |
| Mode | | 328.60 |
| Std. Deviation | | 132.00613 |
| Variance | | 17425.617 |
| Skewness | | 6.037 |
| Std. Error of Skewness | | .032 |
| Range | | 3707.00 |
| Minimum | | 103.60 |
| Maximum | | 3810.60 |
| Sum | | 1996064.44 |
| Percentiles | 25 | 262.6100 |
| 50 | 322.6000 |
| 75 | 406.6000 |

1. **Central Tendency:**
   * **Mean (Average):** The average total fare is 337.92337.92.
   * **Median (50th Percentile):** The median fare is 322.60322.60.
   * **Mode:** The mode is 328.60328.60.
2. **Variability:**
   * **Standard Deviation:** The total fare values are spread out from the mean by approximately 132.01132.01. This indicates a moderate degree of variability.
   * **Variance:** The variance is 17425.6217425.62, providing a measure of the extent to which individual fares differ from the mean.
3. **Distribution Shape:**
   * **Skewness:** The skewness is 6.046.04, indicating a right-skewed distribution. This suggests that there might be a concentration of lower fare values and a long tail of higher fare values.
4. **Range:**
   * **Range:** The range of total fares is 3707.003707.00, showing the difference between the minimum (103.60103.60) and maximum (3810.603810.60) values.
5. **Minimum and Maximum:**
   * **Minimum:** The lowest total fare is 103.60103.60.
   * **Maximum:** The highest total fare is 3810.603810.60.
6. **Sum:**
   * The total sum of all fares is 1996064.441996064.44.
7. **Percentiles:**
   * The 25th percentile (Q1) is 262.61262.61.
   * The 50th percentile (Median or Q2) is 322.60322.60.
   * The 75th percentile (Q3) is 406.60406.60.

These findings provide an overview of the distribution and characteristics of total fares. The skewness and the difference between mean and median suggest that there might be some extreme values pulling the distribution to the right. Further analysis, such as identifying and handling outliers, may be beneficial for a more detailed understanding.

A graph of a person

Description automatically generated

Linear Regression

|  |  |  |
| --- | --- | --- |
| **Case Processing Summary** | | |
|  | N | Percent |
| Included | 5907 | 100.0% |
| Excluded | 0 | 0.0% |
| Total | 5907 | 100.0% |

A screen shot of a chart

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with a blue and white line

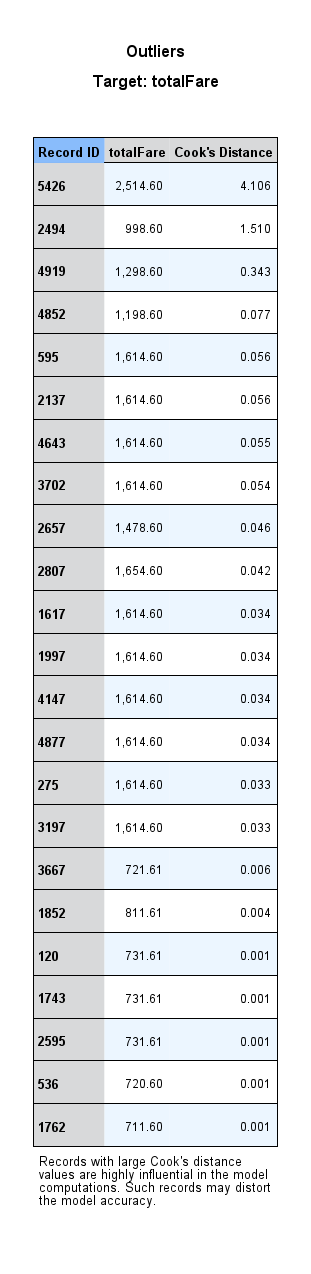
Description automatically generated with medium confidence

A graph of a graph

Description automatically generated

A diagram of a diagram

Description automatically generated with medium confidence



A graph of a line

Description automatically generated with medium confidence

Validation

|  |  |
| --- | --- |
| ArrivalAirportCode | |
| IND||LAX | 1 |
| ORD||LAX | 2 |
| LAX | 3 |
| SFO||ONT | 4 |
| SFO||LAX | 5 |
| CLT||LAX | 6 |
| LAS||LAX | 7 |
| FLL||LAX | 8 |
| DFW||ONT | 9 |
| ONT | 10 |
| ATL||ONT | 11 |
| DFW||LAX | 12 |
| SLC||ONT | 13 |
| IAD||LAX | 14 |
| PDX||LAX | 15 |
| MIA||LAX | 16 |
| PHX||ONT | 17 |
| AUS||LAX | 18 |
| CLT||ONT | 19 |
| PHX||LAX | 20 |
| RDU||LAX | 21 |
| BOS||LAX | 22 |
| MCO||LAX | 23 |
| ORF||ATL||ONT | 24 |
| SLC||LAX | 25 |
| PHX||LAS||LAX | 26 |

|  |  |
| --- | --- |
| DepartureAirportCode | |
| JFK||IND | 1 |
| JFK||ORD | 2 |
| JFK | 3 |
| JFK||SFO | 4 |
| JFK||CLT | 5 |
| JFK||LAS | 6 |
| JFK||FLL | 7 |
| JFK||DFW | 8 |
| JFK||ATL | 9 |
| JFK||SLC | 10 |
| JFK||IAD | 11 |
| JFK||PDX | 12 |
| JFK||MIA | 13 |
| JFK||PHX | 14 |
| JFK||AUS | 15 |
| JFK||RDU | 16 |
| JFK||BOS | 17 |
| JFK||MCO | 18 |
| JFK||ORF||ATL | 19 |
| JFK||PHX||LAS | 20 |

|  |  |
| --- | --- |
| Airline |  |
| American Airlines||American Airlines | 1 |
| JetBlue Airways | 2 |
| American Airlines | 3 |
| Delta | 4 |
| United||United | 5 |
| Alaska Airlines||Delta | 6 |
| United | 7 |
| Alaska Airlines||United | 8 |
| Alaska Airlines||Alaska Airlines | 9 |
| JetBlue Airways||JetBlue Airways | 10 |
| Delta||Delta | 11 |
| Delta||Alaska Airlines | 12 |
| Delta||Delta||Delta | 13 |
| American Airlines||American Airlines||American Airlines | 14 |

|  |  |
| --- | --- |
| CabinCode | |
| coach||coach | 1 |
| coach | 2 |
| first | 3 |
| business | 4 |
| coach||coach||coach | 5 |

Handling outliers

Based on the cook’s distance we removed two instances.

After removing outliers

A screenshot of a computer screen

Description automatically generated

A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated

A diagram of different colored lines

Description automatically generated

A graph of a graph

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Statistics** | | |
| totalFare | | |
| N | Valid | 5905 |
| Missing | 0 |
| Mean | | 337.4346 |
| Std. Error of Mean | | 1.67438 |
| Median | | 322.6000 |
| Mode | | 328.60 |
| Std. Deviation | | 128.66568 |
| Variance | | 16554.856 |
| Skewness | | 5.689 |
| Std. Error of Skewness | | .032 |
| Kurtosis | | 108.688 |
| Std. Error of Kurtosis | | .064 |
| Range | | 3707.00 |
| Minimum | | 103.60 |
| Maximum | | 3810.60 |
| Percentiles | 25 | 262.6100 |
| 50 | 322.6000 |
| 75 | 406.6000 |

A graph of a person

Description automatically generated

1. **Skewness:** The skewness is positive at 5.689, indicating a right-skewed distribution. This means that the distribution has a long tail to the right, with a concentration of lower values and a few higher values pulling the mean to the right.
2. **Kurtosis:** The kurtosis is exceptionally high at 108.688, indicating heavy tails in the distribution. This suggests the presence of extreme values or outliers.

Given the right-skewed nature of the distribution, you might consider applying a transformation to make the data more symmetrical and to reduce the impact of extreme values. Common transformations include:

* **Log Transformation:** This is often used for right-skewed distributions.
* **Square Root Transformation:** Another option for reducing skewness.
* **Box-Cox Transformation:** A family of power transformations that includes the logarithm as a special case.

Before deciding on a transformation, it's crucial to understand the nature of your data and the context of your problem. Log transformations are common, especially when dealing with financial data like fares, as they tend to be right-skewed.

Applying a log transformation to the "totalFare" variable is a reasonable approach given its right-skewed distribution. This transformation can help make the distribution more symmetrical, which aligns with the assumptions of linear regression. It's a common practice, especially when dealing with financial data.

|  |  |  |
| --- | --- | --- |
| **Statistics** | | |
| log(totalFare) | | |
| N | Valid | 5905 |
| Missing | 0 |
| Mean | | 2.503905118516786 |
| Std. Error of Mean | | .001866510743217 |
| Median | | 2.508664363052943 |
| Mode | | 2.516667559099043 |
| Std. Deviation | | .143430147471146 |
| Variance | | .021 |
| Skewness | | .054 |
| Std. Error of Skewness | | .032 |
| Kurtosis | | 1.510 |
| Std. Error of Kurtosis | | .064 |
| Range | | 1.565633607713504 |
| Minimum | | 2.015359755409215 |
| Maximum | | 3.580993363122718 |
| Sum | | 9223.372036854777000 |
| Percentiles | 25 | 2.419311259689980 |
| 50 | 2.508664363052943 |
| 75 | 2.609167374302020 |

A graph of a graph

Description automatically generated

These insights suggest that the log transformation has helped in reducing the skewness of the distribution. The nearly symmetrical shape is beneficial for linear regression analysis, as it aligns with the assumption of normality. The kurtosis indicates a distribution with moderately heavy tails, but it is less extreme compared to the original distribution.

When using this log-transformed variable in multiple linear regression, it's important to interpret the coefficients in the context of the log scale. For example, a one-unit increase in a predictor variable may be associated with a percentage change in the predicted total fare rather than a linear change. The interpretation will depend on the specifics of your regression model and the variables involved.

A screen shot of a chart

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A graph of a column

Description automatically generated

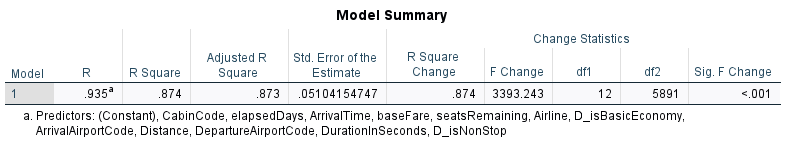
A diagram of a company

Description automatically generated with medium confidence

A graph of a number of dots

Description automatically generated with medium confidence

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model Summary** | | | | | | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
| R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .935a | .874 | .873 | .051041547473814 | .874 | 3393.243 | 12 | 5891 | <.001 |
| a. Predictors: (Constant), CabinCode, elapsedDays, ArrivalTime, baseFare, seatsRemaining, Airline, D\_isBasicEconomy, ArrivalAirportCode, Distance, DepartureAirportCode, DurationInSeconds, D\_isNonStop | | | | | | | | | |



1. **Model Fit:**
   * **R (Multiple Correlation Coefficient):** 0.935
     + This indicates the strength and direction of the linear relationship between the dependent variable and the combination of independent variables.
   * **R Square (Coefficient of Determination):** 0.874
     + The proportion of variance in the dependent variable (totalFare) that can be explained by the independent variables in the model. In this case, approximately 87.4% of the variance is explained.
2. **Model Adjustments:**
   * **Adjusted R Square:** 0.873
     + A version of R square that adjusts for the number of predictors in the model. It helps account for the possibility of overfitting.
3. **Error of the Estimate:**
   * **Std. Error of the Estimate:** 0.051
     + Represents the standard deviation of the residuals (the differences between observed and predicted values). A lower value indicates a better fit.
4. **Change Statistics:**
   * **R Square Change:** 0.874
     + The change in R square from a null model with no predictors to the current model with predictors. It represents the improvement in explained variance.
   * **F Change:** 3393.243
     + The F statistic for the overall significance of the model. It tests whether the addition of predictors significantly improves the model fit.
   * **df1, df2:** 12, 5891
     + Degrees of freedom associated with the F Change statistic.
5. **Significance of F Change:**
   * **Sig. F Change:** <.001
     + The p-value associated with the F Change statistic. In this case, the p-value is less than 0.001, suggesting that the overall model is statistically significant.

**Conclusion:** The regression model, which includes predictors such as CabinCode, elapsedDays, ArrivalTime, baseFare, seatsRemaining, Airline, D\_isBasicEconomy, ArrivalAirportCode, Distance, DepartureAirportCode, DurationInSeconds, and D\_isNonStop, demonstrates a strong overall fit. The predictors collectively explain a significant portion (87.4%) of the variance in the dependent variable (totalFare). The p-value associated with the F statistic is highly significant, indicating that the model as a whole is statistically meaningful.

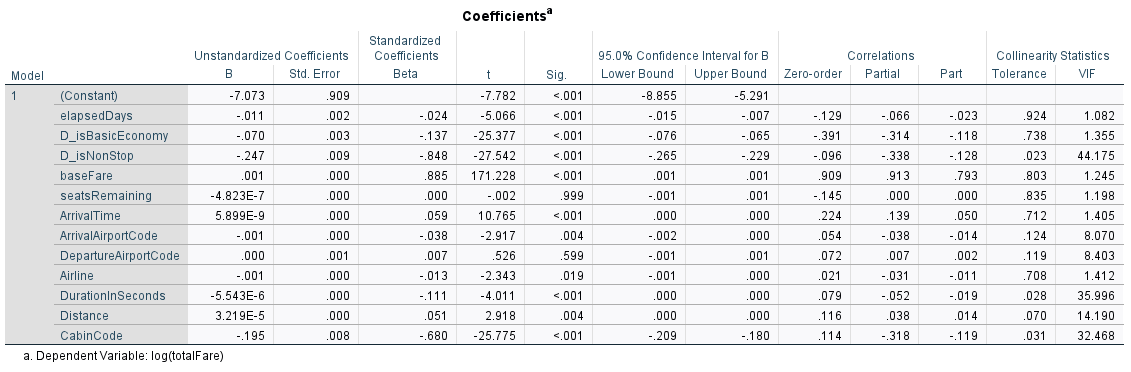
Keep in mind that while the overall model is significant, individual predictors may contribute differently to the model, and their significance should be examined separately. Additionally, diagnostics such as residual analysis can provide insights into the model's assumptions and potential areas for improvement.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 106.083 | 12 | 8.840 | 3393.243 | <.001b |
| Residual | 15.347 | 5891 | .003 |  |  |
| Total | 121.430 | 5903 |  |  |  |
| a. Dependent Variable: log(totalFare) | | | | | | |
| b. Predictors: (Constant), CabinCode, elapsedDays, ArrivalTime, baseFare, seatsRemaining, Airline, D\_isBasicEconomy, ArrivalAirportCode, Distance, DepartureAirportCode, DurationInSeconds, D\_isNonStop | | | | | | |

The ANOVA table provides evidence that the regression model is statistically significant in predicting the log-transformed total fare. The low p-value (<.001) associated with the F-statistic suggests that the model explains a significant amount of variance in the dependent variable. The regression model is preferred over a null model with no predictors.

This information supports the findings from the model summary and reinforces the idea that the chosen predictors collectively contribute to the prediction of the log-transformed total fare.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | | Correlations | | | Collinearity Statistics | |
| B | Std. Error | Beta | Lower Bound | Upper Bound | Zero-order | Partial | Part | Tolerance | VIF |
| 1 | (Constant) | -7.073 | .909 |  | -7.782 | <.001 | -8.855 | -5.291 |  |  |  |  |  |
| elapsedDays | -.011 | .002 | -.024 | -5.066 | <.001 | -.015 | -.007 | -.129 | -.066 | -.023 | .924 | 1.082 |
| D\_isBasicEconomy | -.070 | .003 | -.137 | -25.377 | <.001 | -.076 | -.065 | -.391 | -.314 | -.118 | .738 | 1.355 |
| D\_isNonStop | -.247 | .009 | -.848 | -27.542 | <.001 | -.265 | -.229 | -.096 | -.338 | -.128 | .023 | 44.175 |
| baseFare | .001 | .000 | .885 | 171.228 | <.001 | .001 | .001 | .909 | .913 | .793 | .803 | 1.245 |
| seatsRemaining | -4.823E-7 | .000 | .000 | -.002 | .999 | -.001 | .001 | -.145 | .000 | .000 | .835 | 1.198 |
| ArrivalTime | 5.899E-9 | .000 | .059 | 10.765 | <.001 | .000 | .000 | .224 | .139 | .050 | .712 | 1.405 |
| ArrivalAirportCode | -.001 | .000 | -.038 | -2.917 | .004 | -.002 | .000 | .054 | -.038 | -.014 | .124 | 8.070 |
| DepartureAirportCode | .000 | .001 | .007 | .526 | .599 | -.001 | .001 | .072 | .007 | .002 | .119 | 8.403 |
| Airline | -.001 | .000 | -.013 | -2.343 | .019 | -.001 | .000 | .021 | -.031 | -.011 | .708 | 1.412 |
| DurationInSeconds | -5.543E-6 | .000 | -.111 | -4.011 | <.001 | .000 | .000 | .079 | -.052 | -.019 | .028 | 35.996 |
| Distance | 3.219E-5 | .000 | .051 | 2.918 | .004 | .000 | .000 | .116 | .038 | .014 | .070 | 14.190 |
| CabinCode | -.195 | .008 | -.680 | -25.775 | <.001 | -.209 | -.180 | .114 | -.318 | -.119 | .031 | 32.468 |
| a. Dependent Variable: log(totalFare) | | | | | | | | | | | | | |



The coefficient table provides information about the regression coefficients for each predictor variable in the model predicting the log-transformed total fare. Here's the interpretation:

1. **Constant:**
   * The constant term represents the estimated intercept when all predictor variables are zero.
   * The coefficient for the constant is -7.073.
   * The p-value associated with the constant is less than 0.001, indicating that the intercept is significantly different from zero.
2. **Predictor Variables:**
   * **elapsedDays:** Coefficient = -0.011, Beta = -0.024
     + For each additional elapsed day, the log-transformed total fare is expected to decrease by 0.011 units.
   * **D\_isBasicEconomy:** Coefficient = -0.070, Beta = -0.137
     + Flights with basic economy fare are associated with a decrease of 0.070 units in the log-transformed total fare.
   * **D\_isNonStop:** Coefficient = -0.247, Beta = -0.848
     + Non-stop flights are associated with a significant decrease in log-transformed total fare by 0.247 units.
   * **baseFare:** Coefficient = 0.001, Beta = 0.885
     + For each additional unit increase in base fare, the log-transformed total fare is expected to increase by 0.001 units.
   * **seatsRemaining:** Coefficient = -4.823E-7, Beta = 0.000
     + The coefficient is very close to zero, indicating that seats remaining has a negligible effect on log-transformed total fare.
   * **ArrivalTime:** Coefficient = 5.899E-9, Beta = 0.059
     + A small positive effect of ArrivalTime on log-transformed total fare, where an increase in ArrivalTime is associated with a slight increase in total fare.
   * **ArrivalAirportCode:** Coefficient = -0.001, Beta = -0.038
     + Flights arriving at certain airport codes are associated with a decrease of 0.001 units in the log-transformed total fare.
   * **DepartureAirportCode:** Coefficient = 0.000, Beta = 0.007
     + Departure airport code has a negligible effect on log-transformed total fare.
   * **Airline:** Coefficient = -0.001, Beta = -0.013
     + Certain airlines are associated with a decrease of 0.001 units in the log-transformed total fare.
   * **DurationInSeconds:** Coefficient = -5.543E-6, Beta = -0.111
     + Longer flight durations are associated with a decrease of 5.543E-6 units in the log-transformed total fare.
   * **Distance:** Coefficient = 3.219E-5, Beta = 0.051
     + For each additional unit increase in distance, the log-transformed total fare is expected to increase by 3.219E-5 units.
   * **CabinCode:** Coefficient = -0.195, Beta = -0.680
     + Certain cabin codes are associated with a decrease of 0.195 units in the log-transformed total fare.
3. **Correlations and Collinearity:**
   * The table includes correlations and collinearity statistics, such as zero-order correlations, partial correlations, and variance inflation factors (VIF). These statistics help assess the relationships between predictor variables and the potential multicollinearity.

**Conclusion:** The coefficients provide insights into the direction and strength of the relationships between each predictor variable and the log-transformed total fare. Interpretation should consider both the magnitude of the coefficients and their statistical significance. The Beta values (standardized coefficients) indicate the relative importance of each predictor in explaining the variance in log-transformed total fare.

* A low tolerance and high VIF suggest potential collinearity, indicating that the predictor variable may be redundant in the presence of other predictors. In your table, most VIF values seem reasonable, indicating a relatively low risk of multicollinearity.
* It's crucial to monitor these statistics because high collinearity can lead to unstable coefficient estimates and challenges in interpreting the individual contributions of predictors.
* While a single high VIF might not be problematic, it's essential to consider the overall pattern across predictors and potentially address multicollinearity if it becomes a concern.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Excluded Variablesa** | | | | | | | | |
| Model | | Beta In | t | Sig. | Partial Correlation | Collinearity Statistics | | |
| Tolerance | VIF | Minimum Tolerance |
| 1 | flightDate | -1.867b | -3.965 | <.001 | -.052 | 9.654E-5 | 10358.790 | 9.654E-5 |
| DepartureTime | -12.630b | -8.129 | <.001 | -.105 | 8.791E-6 | 113757.161 | 8.791E-6 |
| a. Dependent Variable: log(totalFare) | | | | | | | | |
| b. Predictors in the Model: (Constant), CabinCode, elapsedDays, ArrivalTime, baseFare, seatsRemaining, Airline, D\_isBasicEconomy, ArrivalAirportCode, Distance, DepartureAirportCode, DurationInSeconds, D\_isNonStop | | | | | | | | |

Excluded Variablesa

Model Beta In t Sig. Partial Correlation Collinearity Statistics

Tolerance VIF Minimum Tolerance

1 flightDate -1.867b -3.965 <.001 -.052 9.654E-5 10358.790 9.654E-5

DepartureTime -12.630b -8.129 <.001 -.105 8.791E-6 113757.161 8.791E-6

a Dependent Variable: log(totalFare)

b Predictors in the Model: (Constant), CabinCode, elapsedDays, ArrivalTime, baseFare, seatsRemaining, Airline, D\_isBasicEconomy, ArrivalAirportCode, Distance, DepartureAirportCode, DurationInSeconds, D\_isNonStop

**Interpretation:**

* The exclusion of "flightDate" and "DepartureTime" from the model was likely a result of high multicollinearity issues with these variables.
* The Beta values and t-values suggest that, even though statistically significant, these variables were associated with collinearity issues that made their inclusion problematic in the presence of the other predictors.
* The partial correlation values give an idea of the relationship each excluded variable has with the dependent variable, taking into account the other predictors in the model.

**Recommendation:**

* High collinearity often poses challenges in interpreting individual predictor effects. You might consider further exploring the reasons for collinearity, potentially through correlation matrices or variance inflation factor (VIF) analysis.
* Additionally, you may need to make decisions about whether to retain or exclude these variables based on the context of your analysis and the goals of your regression model. If these variables are essential, you might explore ways to address collinearity, such as combining related variables or using dimensionality reduction techniques.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficient Correlationsa** | | | | | | | | | | | | | | |
| Model | | | CabinCode | elapsedDays | ArrivalTime | baseFare | seatsRemaining | Airline | D\_isBasicEconomy | ArrivalAirportCode | Distance | DepartureAirportCode | DurationInSeconds | D\_isNonStop |
| 1 | Correlations | CabinCode | 1.000 | -.034 | .026 | -.290 | .047 | -.064 | -.100 | .005 | .036 | -.021 | .006 | .845 |
| elapsedDays | -.034 | 1.000 | -.060 | .082 | -.023 | -.038 | -.035 | -.018 | .102 | .032 | -.170 | -.125 |
| ArrivalTime | .026 | -.060 | 1.000 | -.094 | -.226 | .033 | .144 | .001 | -.444 | .004 | .442 | .208 |
| baseFare | -.290 | .082 | -.094 | 1.000 | .038 | .013 | .274 | .044 | -.045 | -.023 | .018 | -.224 |
| seatsRemaining | .047 | -.023 | -.226 | .038 | 1.000 | -.112 | -.304 | -.014 | .047 | .012 | -.028 | -.004 |
| Airline | -.064 | -.038 | .033 | .013 | -.112 | 1.000 | .079 | -.110 | .088 | -.006 | -.089 | -.030 |
| D\_isBasicEconomy | -.100 | -.035 | .144 | .274 | -.304 | .079 | 1.000 | .052 | .049 | .000 | -.046 | -.037 |
| ArrivalAirportCode | .005 | -.018 | .001 | .044 | -.014 | -.110 | .052 | 1.000 | -.019 | -.875 | .059 | .052 |
| Distance | .036 | .102 | -.444 | -.045 | .047 | .088 | .049 | -.019 | 1.000 | -.030 | -.921 | -.316 |
| DepartureAirportCode | -.021 | .032 | .004 | -.023 | .012 | -.006 | .000 | -.875 | -.030 | 1.000 | .048 | .059 |
| DurationInSeconds | .006 | -.170 | .442 | .018 | -.028 | -.089 | -.046 | .059 | -.921 | .048 | 1.000 | .460 |
| D\_isNonStop | .845 | -.125 | .208 | -.224 | -.004 | -.030 | -.037 | .052 | -.316 | .059 | .460 | 1.000 |
| Covariances | CabinCode | 5.698E-5 | -5.499E-7 | 1.092E-13 | -1.360E-8 | 1.064E-7 | -1.533E-7 | -2.095E-6 | 1.480E-8 | 3.031E-9 | -8.281E-8 | 6.555E-11 | 5.720E-5 |
| elapsedDays | -5.499E-7 | 4.613E-6 | -7.027E-14 | 1.089E-9 | -1.487E-8 | -2.589E-8 | -2.084E-7 | -1.451E-8 | 2.409E-9 | 3.638E-8 | -5.052E-10 | -2.401E-6 |
| ArrivalTime | 1.092E-13 | -7.027E-14 | 3.003E-19 | -3.192E-16 | -3.716E-14 | 5.687E-15 | 2.187E-13 | 2.790E-16 | -2.683E-15 | 1.281E-15 | 3.351E-16 | 1.022E-12 |
| baseFare | -1.360E-8 | 1.089E-9 | -3.192E-16 | 3.855E-11 | 7.145E-11 | 2.562E-11 | 4.710E-9 | 1.050E-10 | -3.087E-12 | -7.523E-11 | 1.556E-13 | -1.248E-8 |
| seatsRemaining | 1.064E-7 | -1.487E-8 | -3.716E-14 | 7.145E-11 | 9.006E-8 | -1.072E-8 | -2.524E-7 | -1.594E-9 | 1.570E-10 | 1.847E-9 | -1.162E-11 | -1.179E-8 |
| Airline | -1.533E-7 | -2.589E-8 | 5.687E-15 | 2.562E-11 | -1.072E-8 | 1.011E-7 | 6.970E-8 | -1.333E-8 | 3.074E-10 | -1.085E-9 | -3.901E-11 | -8.689E-8 |
| D\_isBasicEconomy | -2.095E-6 | -2.084E-7 | 2.187E-13 | 4.710E-9 | -2.524E-7 | 6.970E-8 | 7.645E-6 | 5.450E-8 | 1.504E-9 | 7.070E-10 | -1.771E-10 | -9.226E-7 |
| ArrivalAirportCode | 1.480E-8 | -1.451E-8 | 2.790E-16 | 1.050E-10 | -1.594E-9 | -1.333E-8 | 5.450E-8 | 1.453E-7 | -7.866E-11 | -1.762E-7 | 3.115E-11 | 1.781E-7 |
| Distance | 3.031E-9 | 2.409E-9 | -2.683E-15 | -3.087E-12 | 1.570E-10 | 3.074E-10 | 1.504E-9 | -7.866E-11 | 1.217E-10 | -1.756E-10 | -1.404E-11 | -3.125E-8 |
| DepartureAirportCode | -8.281E-8 | 3.638E-8 | 1.281E-15 | -7.523E-11 | 1.847E-9 | -1.085E-9 | 7.070E-10 | -1.762E-7 | -1.756E-10 | 2.791E-7 | 3.529E-11 | 2.793E-7 |
| DurationInSeconds | 6.555E-11 | -5.052E-10 | 3.351E-16 | 1.556E-13 | -1.162E-11 | -3.901E-11 | -1.771E-10 | 3.115E-11 | -1.404E-11 | 3.529E-11 | 1.910E-12 | 5.700E-9 |
| D\_isNonStop | 5.720E-5 | -2.401E-6 | 1.022E-12 | -1.248E-8 | -1.179E-8 | -8.689E-8 | -9.226E-7 | 1.781E-7 | -3.125E-8 | 2.793E-7 | 5.700E-9 | 8.043E-5 |
| a. Dependent Variable: log(totalFare) | | | | | | | | | | | | | | |

A table with numbers and letters

Description automatically generated

1. **Correlations:**
   * Each cell in the table represents the correlation coefficient between the coefficients of two predictor variables.
   * Positive correlations indicate that as the coefficient of one variable increases, the coefficient of the other variable also tends to increase, and vice versa.
   * Negative correlations indicate that as the coefficient of one variable increases, the coefficient of the other variable tends to decrease, and vice versa.
2. **Interpretation of Specific Correlations:**
   * **CabinCode and D\_isNonStop (0.845):**
     + A strong positive correlation between the coefficients of CabinCode and D\_isNonStop. This suggests that changes in the coefficient of one variable are associated with similar changes in the coefficient of the other variable.
   * **ArrivalTime and DurationInSeconds (0.442):**
     + A moderate positive correlation between the coefficients of ArrivalTime and DurationInSeconds. Changes in the coefficient of ArrivalTime tend to be associated with similar changes in the coefficient of DurationInSeconds.
   * **ArrivalAirportCode and DepartureAirportCode (-0.875):**
     + A strong negative correlation between the coefficients of ArrivalAirportCode and DepartureAirportCode. Changes in the coefficient of one variable are associated with opposite changes in the coefficient of the other variable.
   * **Distance and DurationInSeconds (-0.921):**
     + A strong negative correlation between the coefficients of Distance and DurationInSeconds. Changes in the coefficient of one variable are associated with opposite changes in the coefficient of the other variable.

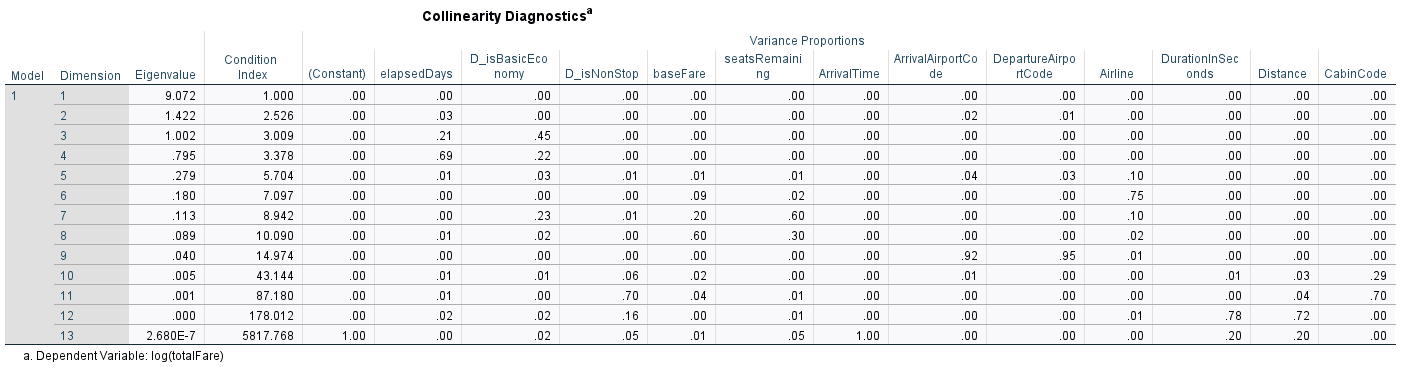
**Interpretation:**

* Strong correlations between certain coefficients may suggest redundancy in the information provided by those variables. For example, the strong positive correlation between CabinCode and D\_isNonStop may imply that these variables convey similar information.
* The strong negative correlation between ArrivalAirportCode and DepartureAirportCode, as well as between Distance and DurationInSeconds, suggests potential collinearity. These variables might be capturing similar aspects of the variance in the dependent variable.
* While correlations don't necessarily imply causation, identifying these patterns can guide further investigation into the relationships among predictor variables and potential adjustments to the model.

**Recommendation:**

* Consider assessing multicollinearity more formally using tolerance and VIF values. High correlations can be indicative of multicollinearity, but they don't provide the complete picture.
* If multicollinearity is a concern, you may need to explore ways to address it, such as removing redundant variables, combining variables, or using dimensionality reduction techniques.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Collinearity Diagnosticsa** | | | | | | | | | | | | | | | | |
| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | | | | | | | | | | | |
| (Constant) | elapsedDays | D\_isBasicEconomy | D\_isNonStop | baseFare | seatsRemaining | ArrivalTime | ArrivalAirportCode | DepartureAirportCode | Airline | DurationInSeconds | Distance | CabinCode |
| 1 | 1 | 9.072 | 1.000 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 2 | 1.422 | 2.526 | .00 | .03 | .00 | .00 | .00 | .00 | .00 | .02 | .01 | .00 | .00 | .00 | .00 |
| 3 | 1.002 | 3.009 | .00 | .21 | .45 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 4 | .795 | 3.378 | .00 | .69 | .22 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 5 | .279 | 5.704 | .00 | .01 | .03 | .01 | .01 | .01 | .00 | .04 | .03 | .10 | .00 | .00 | .00 |
| 6 | .180 | 7.097 | .00 | .00 | .00 | .00 | .09 | .02 | .00 | .00 | .00 | .75 | .00 | .00 | .00 |
| 7 | .113 | 8.942 | .00 | .00 | .23 | .01 | .20 | .60 | .00 | .00 | .00 | .10 | .00 | .00 | .00 |
| 8 | .089 | 10.090 | .00 | .01 | .02 | .00 | .60 | .30 | .00 | .00 | .00 | .02 | .00 | .00 | .00 |
| 9 | .040 | 14.974 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .92 | .95 | .01 | .00 | .00 | .00 |
| 10 | .005 | 43.144 | .00 | .01 | .01 | .06 | .02 | .00 | .00 | .01 | .00 | .00 | .01 | .03 | .29 |
| 11 | .001 | 87.180 | .00 | .01 | .00 | .70 | .04 | .01 | .00 | .00 | .00 | .00 | .00 | .04 | .70 |
| 12 | .000 | 178.012 | .00 | .02 | .02 | .16 | .00 | .01 | .00 | .00 | .00 | .01 | .78 | .72 | .00 |
| 13 | 2.680E-7 | 5817.768 | 1.00 | .00 | .02 | .05 | .01 | .05 | 1.00 | .00 | .00 | .00 | .20 | .20 | .00 |
| a. Dependent Variable: log(totalFare) | | | | | | | | | | | | | | | | |



The "Collinearity Diagnostics" section provides information about the Eigenvalues, Condition Indices, and Variance Proportions for each dimension of the model. Let's interpret the key components:

1. **Eigenvalues:**
   * Eigenvalues represent the variance of each dimension in the model. Larger eigenvalues indicate dimensions that explain more variance in the data.
   * For example, in the first dimension, the eigenvalue is 9.072, suggesting that this dimension explains the majority of the variance in the data. As you move to subsequent dimensions, the eigenvalues decrease.
2. **Condition Indices:**
   * Condition indices measure how much the variance in the regression coefficients is inflated due to multicollinearity.
   * A condition index close to 1 indicates little or no multicollinearity. As the condition index increases, it suggests increasing multicollinearity.
   * For instance, in the first dimension, the condition index is 1.000, indicating no multicollinearity. In subsequent dimensions, the condition index increases, suggesting potential multicollinearity issues.
3. **Variance Proportions:**
   * Variance proportions indicate the proportion of variance in each predictor variable that is shared with other predictor variables. High values indicate high redundancy.
   * In the first dimension, the variance proportions for each variable are 0.00, suggesting little shared variance. As you move to subsequent dimensions, you see increasing values, indicating increasing shared variance.

**Interpretation:**

* The initial dimensions (lower eigenvalues) are likely the most important in explaining the variance in the data. As you move to higher dimensions, you are capturing less and less unique variance.
* The condition indices provide insight into multicollinearity. If the condition indices are very high (e.g., close to or above 30), it suggests potential multicollinearity issues in the model.
* The variance proportions highlight which variables contribute the most to the shared variance in each dimension. Higher proportions suggest more redundancy.
* It's common for the first few dimensions to explain the majority of the variance, and subsequent dimensions to capture diminishing amounts.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Residuals Statisticsa** | | | | | |
|  | Minimum | Maximum | Mean | Std. Deviation | N |
| Predicted Value | 2.190294742584229 | 5.342311382293701 | 2.503933612184823 | .134055785998938 | 5904 |
| Residual | -1.761318206787109 | .449131697416306 | .000000000000010 | .050989640801935 | 5904 |
| Std. Predicted Value | -2.340 | 21.173 | .000 | 1.000 | 5904 |
| Std. Residual | -34.508 | 8.799 | .000 | .999 | 5904 |
| a. Dependent Variable: log(totalFare) | | | | | |

**1. Predicted Value:**

* **Mean:** The average predicted value of the dependent variable is 2.5039.
* **Minimum:** The smallest predicted value is 2.1903.
* **Maximum:** The largest predicted value is 5.3423.
* **Standard Deviation:** The variability of the predicted values around the mean is 0.1341.
* **N:** The total number of observations is 5904.

**2. Residual:**

* **Mean:** The mean of the residuals is practically zero (very close to 0.0000), indicating that, on average, the model's predictions are accurate.
* **Minimum:** The smallest residual is -1.7613.
* **Maximum:** The largest residual is 0.4491.
* **Standard Deviation:** The standard deviation of the residuals is 0.0510, reflecting the average distance of individual data points from the mean residual.
* **N:** The total number of observations is 5904.

**3. Std. Predicted Value:**

* **Mean:** The mean standardized predicted value is essentially zero (very close to 0.000), indicating that, on average, the predictions are close to the mean.
* **Minimum:** The smallest standardized predicted value is -2.340.
* **Maximum:** The largest standardized predicted value is 21.173.
* **Standard Deviation:** The standard deviation of the standardized predicted values is 1.000.

**4. Std. Residual:**

* **Mean:** The mean standardized residual is essentially zero (very close to 0.000), indicating that, on average, the residuals are close to the mean.
* **Minimum:** The smallest standardized residual is -34.508.
* **Maximum:** The largest standardized residual is 8.799.
* **Standard Deviation:** The standard deviation of the standardized residuals is 0.999.

**Interpretation:**

1. The mean residuals being close to zero suggests that, on average, the model predictions are accurate.
2. The minimum and maximum residuals provide insights into the range of errors in prediction.
3. Standardized values (Std. Predicted Value and Std. Residual) help in assessing the relative position of each observation in the context of the entire dataset.