**To:** Prof. Yu Du

**From:** Avi Manawat, Arpit Agarwal, Puneet Kochar

**Subject**: EDA and Regression Analysis for Non-Airline Revenues

**Date**: 11/12/2019

To forecast and predict 4 types of non-airline revenues using regression analysis.

**EXECUTIVE SUMMARY**

**Data Pre-Processing-**

By analyzing the Denver Airport dataset, we found 1 NA value and 1 negative value in Parking and concession respectively in the data of airport revenue. A combined column of ‘month and year’ and an individual column of ‘month’ was observed in data. In ‘year and month’ column the year was interpreted as day.

Since, it is time series data, we cannot remove the NA values from the data so in order to deal with the NA value we took the moving average of that individual month and verified with time series data. To deal with negative value, we took the absolute value of that column. Finally, we segregated the column of ‘month and year’ into an individual column of ‘year’ and individual column of ‘month’.

**EDA Analysis-**

On analyzing the Revenue for each month of five years we observed an increasing pattern in Non-Airline Revenues with every passing month from January to December. December being the most productive month while January the least. Examining Total Non-Airline Revenue by Year, revenue increased with each passing year.

In winter, the revenues from Parking, Rental and concession reduces while in summer they have peak revenue generation period. However, the revenue from ground remain constant throughout year.

Parking was the major source of Non- Airline revenue while Ground was the least in given 5 years. Increasing trend in Non-Airline revenues was observed in successive years. Most revenue comes in the month of December and sudden slump is noted in the following month, January.

**Regression Analysis-**

For our 4 models, we divided our data into training and testing data in order to train our model and make predictions on our test dataset. We created plots; interpreted model coefficients, p-values, Adjusted R-squared values and analyzed residual plots in order to validate regression model assumptions. We created different models taking p-values, adjusted R-Squared and VIF values into account. Lastly, we compared models based on these parameters and chose the optimum model to forecast each Non-Airline Revenue for the months of March through June 2017. We build the correlation matrix to analyze the correlation among different variables and selected the independent variable to fit in our model which had the least correlation among them and highly correlated with the dependent variable.

The aim of this process is to determine the effect on our dependent variable with respect to the independent variables and to analyze which independent variables to fit into our model.

For rental cars, we observed that revenue from rental car was mostly affected by the lagged data, revenue from parking, revenue from concession, seasonality and revenue from ground. After transforming our model into best fitted model based on our verification of assumptions and other parameters such as R-square, coefficients of p-values and VIF, there was an error of 20.4% in our predicted values for the month of Mar 2017 through June 2017. The R-squared value is 0.8943, Thus 89% of the dependent variable value variation around the regression line can be explained by independent variables in the model.

For Ground, we observed that revenue from ground was affected most by the Year, Enplaned and Origin+Destination. After transforming our model into best fitted model, there was an error of 8.87% in our predicted values for the month of Mar 2017 through June 2017. The R-squared value is 0.873, Thus 87% of the dependent variable value variation around the regression line can be explained by the independent variables in the model.

For Concession, the revenue from Concession was affected most by the Year, Month\_Jan, Origin+Destination and revenue from Rental Cars. After transforming our model into best fitted model, there was an error of 9.56% in our predicted values for the month of Mar 2017 through June 2017. The R-squared value is 0.711, Thus 71% of the dependent variable value variation around the regression line can be explained by the independent variables in the model

For Parking, the revenue from Parking was affected most by Years, Transfer and Month. After transforming our model into best fitted model, there was an error of 1% in our predicted values for the month of Mar 2017 through June 2017. The R-squared value is 0.8093, Thus 81% of the dependent variable value variation around the regression line can be explained by the independent variables in the model

**Appendix**

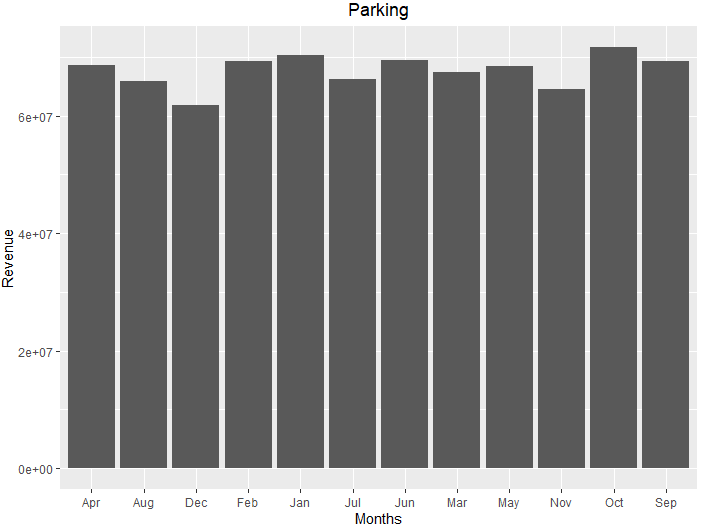
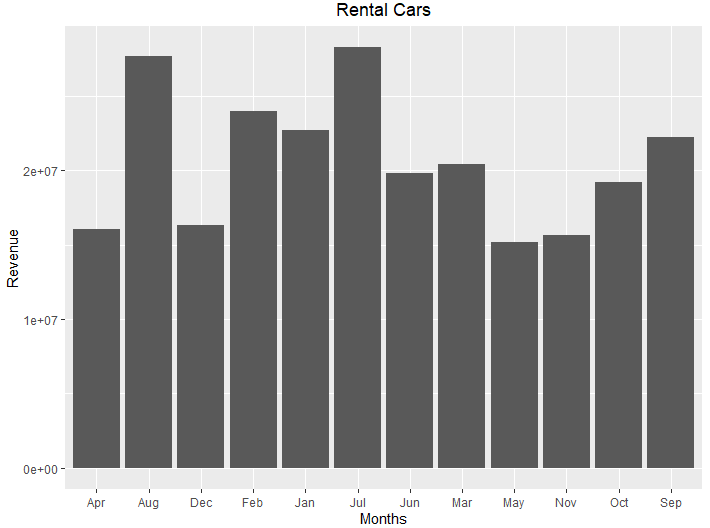
**EDA**

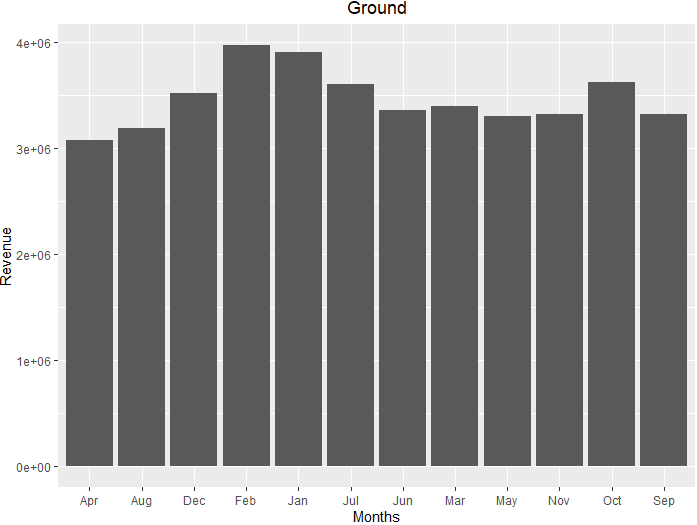
**Non Airline Revenues by time**

A close up of text on a white background

Description automatically generated

**Total Revenues During the 5 years by Month**





Model 1 – Rental Cars

**Equation**

Log(Rental Cars) = 0.0003(umscentlag2\_tr)+0.93595log(parking)-0.3596log(Concession\_tr)+month+0.025exp(ground\_tr)-5.9121+0.3035\*Month

**P-value of Coefficients Interpretation**

As all p-values are less than 0.05, we reject the null hypothesis and conclude that our coefficients are statistically significant.

**Uncertainty of Coefficients**

We are 95% confident that the intercept and coefficients of our independent variables fall within the intervals.

**R-squared Interpretation**  
The R-squared value is 0.8943, Thus 89% of the dependent variable value variation around the regression line can be explained by independent variables in the model and the rest is explained by other factors.

**Checking Assumptions**

Through residual plots, we test our assumptions for the regression model. Our linear regression assumptions state that:

The ‘Residuals vs. Fitted’ plot tells us whether our linearity and homoscedasticity assumption is met. In our case, the points are symmetrically distributed around the horizontal line in our plot, with a roughly constant variance. Hence, our assumption of linearity and homoscedasticity is met in the model.

The ‘Normal Q-Q’ plot gives us idea whether our errors are normally distributed. In our case, on visualizing the Normal Q-Q plot, data points are fitting nicely around the red diagonal line, which states that our assumption for Normality of errors is met.

Model 2 – Ground

**Equation**

Ground = -269.83699+ 0.13404\*Years+ -0.03063\*exp(Enplaned\_tr) + 0.89735\*log(Origin...Destin)

**P-value of Coefficients Interpretation**

Since all p-values are less than 0.05, we can reject the null hypothesis and conclude that our coefficients are statistically significant.

**Uncertainty of Coefficients**

We are 95% confident that the intercept and coefficients of our independent variables fall within the intervals.

**Interpret R-squared**  
The R-squared value is 0.873, Thus 87% of the dependent variable value variation around the regression line can be explained by the independent variables in the model and the rest is explained by other factors.

**Checking Assumption**  
Through residual plots, we test our assumptions for the regression model. Our linear regression assumptions state that:

The ‘Residuals vs. Fitted’ plot tells us whether our linearity and homoscedasticity assumption is met. In our case, the points are symmetrically distributed around the horizontal line in our plot, with a roughly constant variance. Hence, our assumption of linearity and homoscedasticity is met in the model.

The ‘Normal Q-Q’ plot gives us idea whether our errors are normally distributed. In our case, on visualizing the Normal Q-Q plot, data points are fitting nicely around the red diagonal line, which states that our assumption for Normality of errors is met.

Model 3 – Parking

**Equation**

Parking = 23.383 +84.13\*Yearstrans -0.483744\*log(transfer)- 0.111\*Month

**P-value of Coefficients Interpretation**

As all p-values are less than 0.05, we reject the null hypothesis and conclude that our coefficients are statistically significant.

**Uncertainty of Coefficients**

We are 95% confident that the intercept and coefficients of our independent variables fall within the intervals.

**R-squared Interpretation**  
The R-squared value is 0.8093, Thus 81% of the dependent variable value variation around the regression line can be explained by the independent variables in the model and the rest is explained by other factors.

**Checking Assumptions**

Through residual plots, we test our assumptions for the regression model. Our linear regression assumptions state that:

The ‘Residuals vs. Fitted’ plot tells us whether our linearity and homoscedasticity assumption is met. In our case, the points are symmetrically distributed around the horizontal line in our plot, with a roughly constant variance. Hence, our assumption of linearity and homoscedasticity is met in the model.

The ‘Normal Q-Q’ plot gives us idea whether our errors are normally distributed. In our case, on visualizing the Normal Q-Q plot, data points are fitting nicely around the red diagonal line, which states that our assumption for Normality of errors is met.

Model 4 – Concession

**Equation**

Parking = 23.383 +84.13\*Yearstrans -0.483744\*log(transfer)- 0.111\*Month

**P-value of Coefficients Interpretation**

As all p-values are less than 0.05, we reject the null hypothesis and conclude that our coefficients are statistically significant.

**Uncertainty of Coefficients**We are 95% confident that the intercept and coefficients of our independent variables fall within the intervals.

**R-squared Interpretation**  
The R-squared value is 0.711, Thus 71% of the dependent variable value variation around the regression line can be explained by the independent variables in the model and the rest is explained by other factors.

**Checking Assumptions**

Through residual plots, we test our assumptions for the regression model. Our linear regression assumptions state that:

The ‘Residuals vs. Fitted’ plot tells us whether our linearity and homoscedasticity assumption is met. In our case, the points are symmetrically distributed around the horizontal line in our plot, with a roughly constant variance. Hence, our assumption of linearity and homoscedasticity is met in the model.

The ‘Normal Q-Q’ plot gives us idea whether our errors are normally distributed. In our case, on visualizing the Normal Q-Q plot, data points are fitting nicely around the red diagonal line, which states that our assumption for Normality of errors is met.