PDA Report – Airlines Model: PDA

Section 56

Winter Quarter

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

School of Continuing Studies

Northwestern University

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Program Analyst

Wooddale Church

6630 Shady Oak Road

Eden Prairie, MN 55344

**Executive Summary**

In the current economic climate, keeping total cost down is a matter of survival for any competitive airline industry. Through the initial exploratory data analysis (EDA), it was found that the variable Revenue Passenger Miles has the strongest correlation with total cost and is statistically significant. Through data transformations, the natural log was used to make the data user-friendly; as a result when interpreting the coefficients, please perceive the values accordingly. The variable Load Factor suffers from heteroscedasticity, and given its overall contribution to the model, I would recommend dropping it from the model. While the initial EDA was helpful in establishing a strong predictor variable, I would recommend a further analysis including time and a delineation between the airlines in the EDA.

**Introduction**

In 1978, the US government deregulated the airline industry and as a result over $60 billion dollars has been lost to-date through airlines filing for bankruptcy (Npr.org & Severin Borenstein). For the deregulated airline industry the game is quite simple, cover your total costs or cease to exist in your business model. In order for an airline to stay profitable, it must understand the dynamics between its total cost (C-dependent variable) with revenue (Q), price of fuel (PF), and capacity utilization (LF) -independent variables. When studying profitability in the airline industry, time (T) and airline companies (I) are additional important variables that add depth and breadth to the EDA. The pie chart to the left was created by Charles Najda, from the Department of Economics at Stanford University, and visually breaks down airline operating costs. Fuel only represents 13 percent of operating costs and capacity utilization does not encompass operating costs, thus I suspect neither of these will have a strong correlative relationship with total cost. For this report, revenue is expressed as follows: Revenue Per Passenger Miles, and can be understood as the more miles a passenger accumulates the greater the total cost for the airline. Revenue per passenger miles as a variable encompasses all the operational expenses of an airline and I suspect it will have a strong correlative relationship with total cost. I expect time and total cost to have a positive relationship, such that as time increase cost rises as well. Airline ID is a rather arbitrary variable, and I expect specific ID’s to follow an inclusive pattern with total cost. A further analysis of these dynamics will aid airline executives in better understanding its bottom line and how to remain profitable.

**Analysis**

In order to meet the objective of exploring the relationship between the dependent variable and independent variables, an exploratory data analysis must be conducted. I will be utilizing the EDA paradigm and structure put forth by Bruce Ratner found in his book *Statistical and Machine-Learning Data Mining*:

Problem/Objective: Explore the relationship between Total Cost (C), and the independent variables Revenue Per Passenger Miles (Q), Price of Fuel (PF) Time (T), Airline (I), and Capacity Utilization for load factor (LF). While exploring the relationship, I will conduct both fixed and random panel data analysis on this dataset.

Data: The data has been aggregated and has been supplied from management.

Analysis: Scatter plots and correlation coefficients will be used to study the nature of the relationships between the independent variables and their relation to the dependent variable. I will conduct a pooled analysis and briefly comment on the overall findings.

Model: After assessing the data, a model will be used. Management has recommended using a regression model, but the standard OLS assumptions will need to be validated. In addition, least squares dummy variable (LSDV), fixed and random panel models will be used to draw further conclusions from the data.

Results/Interpretation: Once the model has been validated and iterations complete, a recommendation will be written to management in regard to the relational dynamics amongst the variables listed above. In addition, the Beusch and Pagan Lagrange Multiplier test or the Hausman Specification test will be used to determine which model should be used.

A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives which model is used, and the analyst’s personal bias is mitigated.

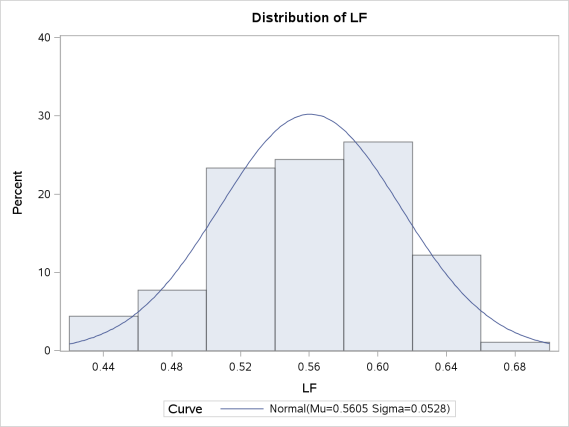
**Data**

Following the outline above, exploring the data is the next step for the EDA. There are a total of 90 observations with 0 missing values in the data set for each variable. The response variable along with two of the independent variables requires a data transformation in order to better study the variables. Utilizing log transformations alter the data to a fairly standard shape (Ajmani). Specifically, this technique is used for positively skewed data and the result of the log transformation moves the majority of the data such that it follows a normal distribution.

Each variable has its own descriptive breakdown explained in a subsection below.

Capacity Utilization (LF): This variable did not require a data transformation and represents the utilization of overall capacity for the airplane load factor.

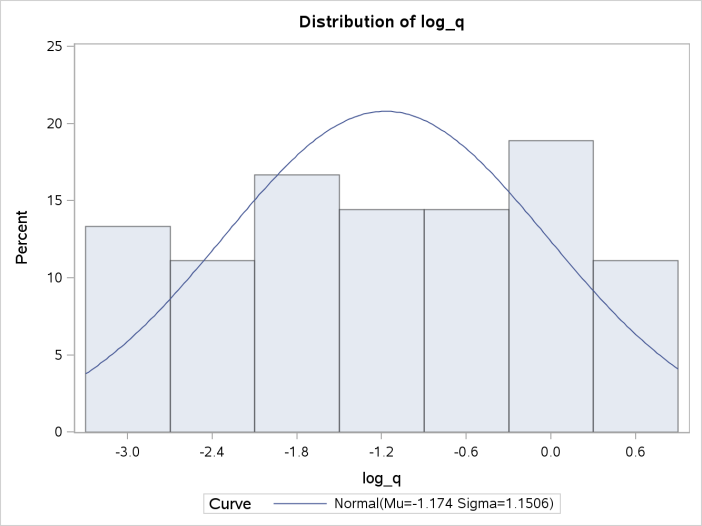
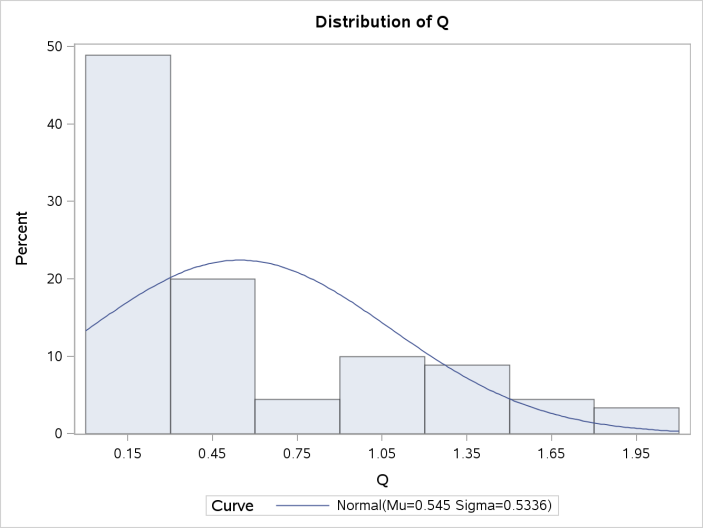
| **Descriptive Stats for Variable: Capacity Utilization as LF** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **N** | **Miss** | **Minimum** | **Maximum** | **Median** | **Mean** | **Variance** | **Std Dev** |
| 90 | 0 | 0.432 | 0.676 | 0.566 | 0.560 | 0.003 | 0.053 |

LF is the easiest variable to understand given that the minimum and maximum values are less than one and the difference is .244. In addition, the mean and median are relatively close to one another which would lead me to believe there is a small standard deviation (SD). The variance is small, of which the SD is based. The visual demonstration, via the histogram to the right, reveals exactly what would be expected from the table above. This variable is slightly negatively skewed, but overall is an excellent variable to conduct analysis.

Revenue Passenger Miles (Q and LogQ): Variables often need transformation in order to be better understood and presented in a form that is conducive to iterative analysis. In this analysis, variable Q needed a log transformation.

| **Descriptive Stats for Variable: Log\_Q and Q** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **N** | **Miss** | **Minimum** | **Maximum** | **Median** | **Mean** | **Variance** | **Std Dev** |
| log\_q Q | Q | 90 90 | 0 0 | -3.279 0.038 | 0.661 1.936 | -1.187 0.305 | -1.174 0.545 | 1.324 0.285 | 1.151 0.534 |

At first glance, this variable might appear to not need a transformation based on the descriptive statistics, the min, max, variance and SD all look fine. The red flag that caught my eye was the difference between the median and mean, which suggests that the observations are not normally distributed. After the log transformation, the mean and median are much closer.

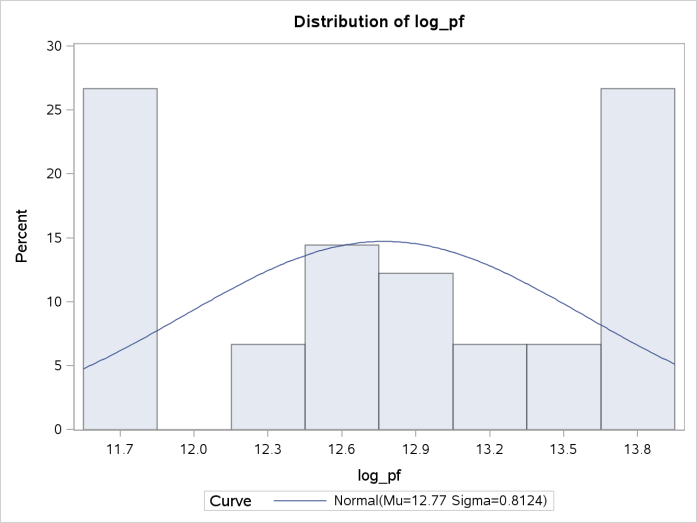
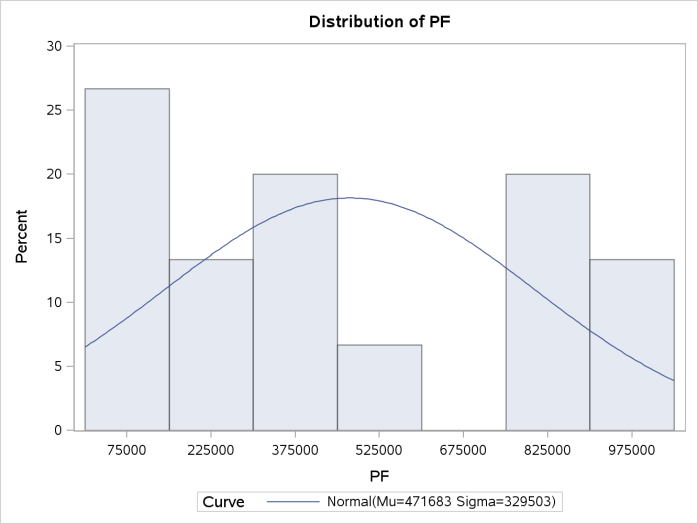


The histogram reveals the severely skewed variable Q, which is why visual statistics are an important asset to EDA. After the log transformation, variable Q follows a normal distribution with only a slight negative skew of -.1.

Price of Fuel (PF and LogPF): Similar to variable Q, variable PF needed a log transformation.

| **Descriptive Stats for Variable: Log PF and PF** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **N** | **Miss** | **Minimum** | **Maximum** | **Median** | **Mean** | **Variance** | **Std Dev** |
| log\_pf PF | PF | 90 90 | 0 0 | 11.550 103795.000 | 13.831 1015610.000 | 12.787 357433.500 | 12.770 471683.011 | 0.660 108572166191 | 0.812 329502.908 |

In its original form, variable PF is very hard to understand. Grasping the variance and (SD) is rather trivial given the sheer size of the numbers. In addition, one should note the difference between mean and median. After the log transformation, the min and max are not far apart. The median and mode would lead me to believe there is a relatively normal distribution, and the variance/SD fits the variable.

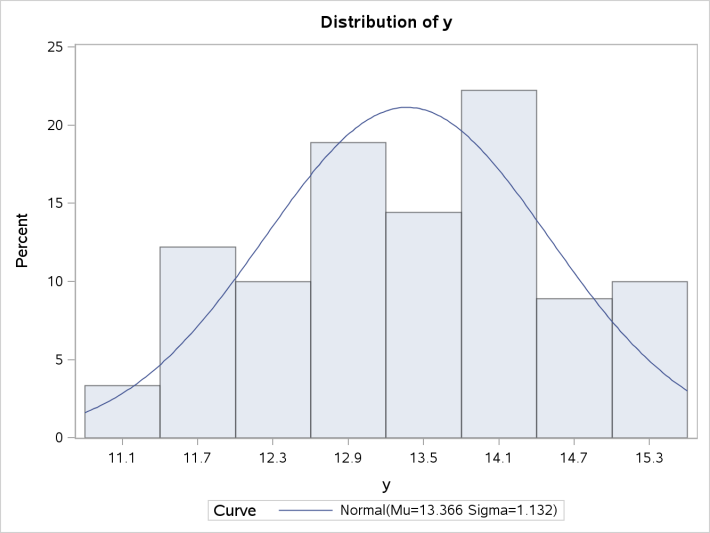
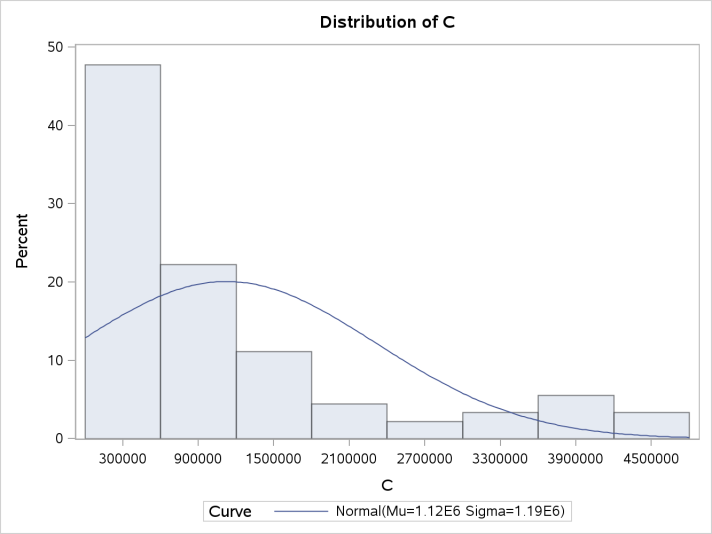


Variable PF in its original form is positively skewed .40, and appears to not follow a normal distribution. After the log transformation, the skew is only -.14, and the distribution allows one to conduct further analysis.

Total Cost (C and LogC expressed as y): This variable represents the dependent variable, and is expressed in millions of dollars.

| **Variable** | **Label** | **N** | **Miss** | **Minimum** | **Maximum** | **Median** | **Mean** | **Variance** | **Std Dev** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| y C | C | 90 90 | 0 0 | 11.142 68978.000 | 15.373 4748320.000 | 13.365 637001.000 | 13.366 1122523.833 | 1.281 1.4210421E12 | 1.132 1192074.704 |

Variable C is very similar to variable PF in that its large numbers are hard to understand. In addition, the large difference between median and mean suggest that a log transformation is necessary. After the log transformation, expressed as y, the numbers are perceivable and the median and mean fall close to one another.



Before the log transformation, C had a massive positive skew of 1.53 and would be a difficult variable to analyze. After the log transformation, the skew is only -.10 and the variable is easier to understand.

| **I = Airline Companies** | | | | |
| --- | --- | --- | --- | --- |
| **I** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **1** | 15 | 16.67 | 15 | 16.67 |
| **2** | 15 | 16.67 | 30 | 33.33 |
| **3** | 15 | 16.67 | 45 | 50.00 |
| **4** | 15 | 16.67 | 60 | 66.67 |
| **5** | 15 | 16.67 | 75 | 83.33 |
| **6** | 15 | 16.67 | 90 | 100.00 |

Airline Companies (I): In this data set, there are a total of six airline companies being studied. As demonstrated in the table to the left, each airline company accounts for 16.67 percent of the total data being studied. In addition, each airline has 15 individual observations that represent a year. If one were to conduct an individual analysis on each airline, the results would not be statistically significant, but pooling the airlines together creates 90 data points, which is statistically more significant.

Time (T): This is the first EDA that is analyzing time in conjunction with other

| **T = Time** | | | | |
| --- | --- | --- | --- | --- |
| **T** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **1** | 6 | 6.67 | 6 | 6.67 |
| **2** | 6 | 6.67 | 12 | 13.33 |
| **3** | 6 | 6.67 | 18 | 20.00 |
| **4** | 6 | 6.67 | 24 | 26.67 |
| **5** | 6 | 6.67 | 30 | 33.33 |
| **6** | 6 | 6.67 | 36 | 40.00 |
| **7** | 6 | 6.67 | 42 | 46.67 |
| **8** | 6 | 6.67 | 48 | 53.33 |
| **9** | 6 | 6.67 | 54 | 60.00 |
| **10** | 6 | 6.67 | 60 | 66.67 |
| **11** | 6 | 6.67 | 66 | 73.33 |
| **12** | 6 | 6.67 | 72 | 80.00 |
| **13** | 6 | 6.67 | 78 | 86.67 |
| **14** | 6 | 6.67 | 84 | 93.33 |
| **15** | 6 | 6.67 | 90 | 100.00 |

independent/dependent variables. The time variable is split into 15 years. It can be seen that each year accounts for 6.67 percent of the total time variable. In addition, each airline company is represented once though each year.

The data has been analyzed in its original form and transformed for better analysis. From analyzing variable I, it can be seen that an individual airline only produces 15 data points. It would be unwise to conduct a thorough EDA on an individual airline because the sample dataset is very small. Thus, using a method that utilizes panel data is desired since this method increases the sample size and includes unobserved heterogeneity effects in the model. Unobserved heterogeneity is a new concept increases the sample size and includes unobserved heterogeneity effects in the model. uld be unwise to conduct

**Results**

| **Analysis of Variance** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 3 | 112.70545 | | 37.56848 | 2419.34 | <.0001 |
| **Error** | 86 | 1.33544 | | 0.01553 |  |  |
| **Total** | 89 | 114.04089 | |  |  |  |
| **Root MSE** | 0.12461 | | **R-Square** | | 0.9883 |
| **Dependent Mean** | 13.36561 | | **Adj R-Sq** | | 0.9879 |
| **Coeff Var** | 0.93234 | |  | |  |

From the data analysis, management has encouraged initially using a pooled regression model utilizing using anyear. ime. rs. It can be seen that each year ling the airlines together creates 90 data points, ut utsdfjkl;asdfjkl;Ordinary Least Squares (OLS) to estimate the parameters for the data. Through using this model and method the paramters for the , a brief overview will be given for the overall findings.

Preliminarily, the model has a strong R-squared but the Adj R-squared is preferred given that the model has more than one variable. The F-value is very significant based on three degrees of freedom. This can be interpreted as at least one variable is explanative of the dependent variable in the model.

Statistically the variables are significant, and Log Q has the largest r-squared. The variance inflation factors (VIFs) do not warrant concern for multi-collinearity. The overall model has strong predictive qualities and is statistically significant, but in order to use this model the OLS assumptions need to be validated. Load factor is the only variable to have an inverse relationship, while LogQ and LogPF have positive relationships with total cost given the caveat than when interpreted the other variables are held constant.

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | 1 | 9.51692 | 0.22924 | 41.51 | <.0001 | 0 |
| **log\_q** | 1 | 0.88274 | 0.01325 | 66.60 | <.0001 | 1.33304 |
| **log\_pf** | 1 | 0.45398 | 0.02030 | 22.36 | <.0001 | 1.55936 |
| **LF** | 1 | -1.62751 | 0.34530 | -4.71 | <.0001 | 1.90468 |

**References**

Borenstein, Severin. Phone interview. 16 Dec. 2011.

Ajmani, Vivek B.. *Applied Econometrics Using the SAS System*. Hoboken, NJ: Wiley, 2008. Print.

Borenstein, Severin. *Why Airlines Keep Going Bankrupt*. Washington DC: Interview - NPR, 2011. Print.

Chatterjee, Samprit, and Ali S. Hadi. *Regression Analysis by Example*. Fifth ed. Hoboken, New Jersey: Wiley, 2012. Print.

Cody, Ronald P.. *SAS Statistics by Example*. Cary, N.C.: SAS Pub., 2011. Print.

Greene, William H.. *Econometric Analysis*. 7th ed. Upper Saddle River, N.J.: Prentice Hall, 2012. Print.

Kenney, Caitlin. "Why Airlines Keep Going Bankrupt : Planet Money : NPR." *NPR : National Public Radio : News & Analysis, World, US, Music & Arts : NPR*. N.p., n.d. Web. 8 Jan. 2013. <http://www.npr.org/blogs/money/2011/12/16/143765367/why-airlines-keep-going-bankrupt>.

"Multicollinearity." *Statistics Solutions*. N.p., n.d. Web. 12 Jan. 2013. <http://www.statisticssolutions.com/resources/dissertation-resources/data-entry-and-management/multicollinearity>.

Najda, Charles. "Low-Cost Carriers and Low Fares: Competition and Concentration in the U.S. Airline Industry." *Stanford University Theses* 1 (2003): 9. *Department of Economics Stanford University*. Web. 8 Jan. 2013.

Ratner, Bruce. *Statistical and Machine-Learning Data Mining* . 2nd ed. Boca Raton, FL: CRC Press, 2012. Print.

Feedback: Daniel: Cover Page: Thanks for including-->awesome job with the formatting.

ES: I am good here. However, to say that the natural log was used to make the data "user-friendly" is incorrect. If this was an interview, you would be given follow-up questions. What you need to say is "The log transformations are typically used to get the data to a somewhat normal shape."

That this is typically used for positive skewed data and that the transformation shifts the bulk of the data in a manner that makes it more normal.

No points deducted here.

Introduction: Awesome job with the introduction write-up on the first page. Love the pie-chart and the reference provided. I suspect that when Greene (2003) said revenue passenger miles---he meant revenue per passenger mile.

Analysis: I am good with the steps you have listed. Nothing for me to add here.

Data: With the LF variable being somewhat normal, there is no sense for utilizing the log transformation. Hopefully, you saw that. I am good with the RPM descriptives and the write-up. In fact, good job here. Trickles down to the pF variable as well. Same holds for the Total Cost dependent variable. Great job in this section.

Results: Fine with the first paragraph. Fine with the correlation analysis as well as the ANOVA and the Global F Test. Fine with the discussions around Multicollinearity.

The interpretations of the Log-Q and Log-pF are fine---However, you needed to say that the other variables need to be held constant (-1/2).

Love the scatter plots for the three independent variables versus the dependent variables. The interpretations are correct as well. You included the correlation analysis table and the coefficients table twice. May want to include them once with a legend and then reference it elsewhere in the document.

The residual diagnosis is fine (Heteroscedasticity detected). The interpretations are fine as well. Normal probability plot, Cook's D, Time Series Plot of the residuals and interpretations are spot on. Thanks for providing the overall model. Future Work section write-up/recommendations make sense. Thanks for the references.