**Module Interpretations**

This document is meant to assist the instructor in interpreting some of the subjective parts of the module content. These sample interpretations should serve as a frame of reference for the instructor to assist the students in interpreting their findings.

**Module Part 1:**

Question 3 - for this question we ask students to assess normality of the different delay variables and pax\_per\_outcome for American Airlines with histograms, boxplots, and qq-plots. Note that histograms are best for assessing the general shape of the data, boxplots are best for assessing outliers as well as shape, and qq-plots are the best test for normality. A normally distributed sample of data should roughly follow a bell curve shape and generally not have outliers, but the ultimate test is with the qq-plot. In reality, very little data is likely to perfectly follow a normal distribution, so students should be liberal with the normal assumption and only indicate that a sample of data is not normal if there is **strong deviation from the 45 degree line of the qq-plot, many outliers on the boxplot, and/or the histogram is clearly skewed, bimodal, or some other deviation from a bell curve.** In general, none of the delay variables and pax\_per\_flight for American Airlines are normally distributed aside from Carrier Delay Percent and NAS Delay Percent, which have some outliers and deviation on the qq-plot, respectively, but they are close enough. The other variables have more egregious deviations from normality. If students have questions, make sure to emphasize that they can be liberal with claiming normality. Ultimately, it doesn’t matter a ton since the Central Limit Theorem can be invoked with data of this size.

Question 5 - here we ask students to assess which variables are correlated with one another and think about why this makes sense. There is a “hint” in the question (shown below), but we provide another interpretation here as well:

* *“Hint: Think about why two delay variables might be highly (or lowly) correlated. For instance, the output below will show you that the correlation between carrier\_delay\_percent and security\_delay\_percent is 0.1911, which is fairly small. This might be because carrier delays refer to delays that are in the airline's control (e.g., crew problems), whereas security delays are delays not in the airline's control (e.g., airport security breaches), so it makes sense that the two are not closely related.”*
* Second interpretation (for instructor reference): Late\_aircraft\_delay\_percent is the variable with the second highest correlation value (0.3692) with pax\_per\_flight, implying that with greater proportions of late aircraft delays, there are more passengers per flight. This might confuse some students, as they might be led to believe that more late aircraft delays actually improve airline performance via increasing the number of passengers. However, this is likely an artifact of some other process that may be occurring. The data description table in “DIFUSE\_Stats\_Module\_Part\_1\_Assignment” file describe Late Aircraft Delays as instances “where delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas.” Thus, for instance, if an aircraft is running late, they may decide to shove all passengers from that flight onto another and potentially even cancel the flight running late, thereby increasing the number of passengers per flight. This may be an incorrect/partially correct interpretation, but the goal is for students to reach some logical conclusion based on what they know about the data (**the instructor should tell students to revisit the data description table in the “DIFUSE\_Stats\_Module\_Part\_1\_Assignment” file for this purpose).**

**Module Part 2:**

Task 2a - here we ask students to produce a linear best fit for several variables on pax\_per\_flight. The linearBestFit() function is meant to provide a correlation coefficient between the independent and response variable (pax\_per\_flight), where the p-value (in the printed output) is that of the two-way t-test between these variables. For all of the delay variables, the output can be read as, “Changes in the predictor variable results in (positive or negative) movements in pax\_per\_flight.” **This coefficient should be used to determine direction of correlation, not magnitude. If the absolute value of the correlation coefficient is greater than 0.5, then there is a significant correlation between the variables.**

Question 5 asks students to reason through the correlations between variables, and Task 2a is an extension of that thinking between pax\_per\_flight and the delay variables. While each variable may differ, the following are surface-level interpretations of why a positive or negative correlation may exist between these variables. Students are more than welcome to offer their own suggestions, **this module does not attempt to identify the mechanism(s) behind these relationships**:

| **Delay Type** | **Sample Positive Correlation Interpretation** | **Sample Negative Correlation Interpretation** |
| --- | --- | --- |
| carrier\_delay\_percent | Airlines may pad their flights with buffer time to increase pax\_per\_flight **without reducing** their controllable delays (i.e. their carrier delays) | If an airline provides uniquely poor service and has greater delays than similar carriers (for delays within their control), then passengers may book with another airline |
| weather\_delay\_percent | In the event of bad weather, a low-cost carrier may experience an increase in pax\_per\_flight as passengers cancel their original flight | In the event of bad weather, a full-service may experience an decrease in passengers per flight as passengers book cheaper flights |
| nas\_delay\_percent | This airline is relatively small by market share, and an increase in NAS delays (which are structural delays related to the logistics of the airport, not the airline) may not decrease their pax\_per\_flight | This airline is relatively large by market share, and an increase in NAS delays (which are structural delays related to the logistics of the airport, not the airline) may increase their pax\_per\_flight |
| security\_delay\_percent | Passengers view security delays as a means to an end, or perceive a greater payoff to going through security  e.g. an airline that flies internationally, like Delta, may have greater security delays because they bring passengers through customs, but these delays do not detract from Delta’s pax\_per\_flight if they are equal or on par with other international carriers’ security delays | Passengers view security delays as an encumbrance and decide to book on another airline, e.g. for a domestic airline |
| late\_aircraft\_delay\_percent | If an aircraft is running late, the airline may decide to shove all passengers from that flight onto another and potentially even cancel the flight running late, thereby increasing the number of passengers per flight | If an aircraft is running late and the time-cost benefit of booking another flight exceeds remaining on their current flight, passengers may decide to book another flight |

Task 2b - here we ask students to use a parsimonious approach to find the best model that predicts pax\_per\_flight. Students should experiment with the value of k, and see how the value of Rsquared changes. In the first note,

*Note 1: Parsimonious model creation isn't just about dropping the least predictive variables, it's about finding a predictive model. Is there a large change in Rsquared when you drop 3 variables instead of 2, or 4 variables instead of 3? If Rsquared is 0.60 with 2 variables, but 0.15 with 1 variable, which model (which set of predictor variables) would we say better predicts the outcome? Sometimes dropping all variables isn't the best idea.*

We suggest to students that they should look at the Rsquared value each time while adjusting the value of k. The variables they choose as the most predictive of pax\_per\_flight will have the greatest Rsquared value with the fewest variables possible.

Task 2c - here we ask students to use Support Vector Machine (SVM) to assess whether airlines can be distinguished by their delay types. The first function, predictAirlineSVM(), is a single trial of SVM; the second function, predictAirlineSVM\_N(), performs *n* trials of SVM. Students should use these functions **in tandem** to interpret the accuracy of their model.

The section, **Conclusions from Task 2c**, gives an overview of what students should be thinking after performing these tests. To reiterate, SVM uses all independent variables to attempt to classify an airline as distinct from another (e.g. distinguishing Southwest and American airlines).

If the model is highly accurate, then a student may suggest that PHX should examine these airlines (e.g. Southwest and American) more closely. SVM may speak to underlying characteristics of an airline and allow PHX to implement targeted strategies for improving each airline (which would be more cost-effective than applying a blanket approach to both airlines that may not address the unique problems of each).