



**Table of Contents**

**Regular (MC1)**

Introduction………………………………………………………………………………….1

Goal………………………………………………………………………………………….1

Dataset Characteristics………………………………………………………………………1

Handling Missing Data……………………………………………………………...………4

New Data set5

Vertical Bar Chart………………………………………………………………………….6

Mean Report & Scatter Plot………………………………………………………………...7

Quantitative Variables – Histogram………………………………………………………...8

Pivot tables – Summarization…………………………………………………………….10

**Advance (MC2)**

Data Reduction12

Data Conversion14

Prediction………………………………………………………………………………….16

Challenges…………………………………………………………………………………22

Benefits…………………………………………………………………………………….22

The idea for further development………………………………………………………….22

References ………………………………………………………………………………23

**Regular (MC1):**

**Introduction:**

The flight dataset provided here contains information on all commercial flights that left Washington, D.C. and arrived in New York in January 2004. It includes information on the departure and arrival airports, the route distance, the scheduled time and date of the flight, and so on for each flight. The variable we're seeking to predict is whether or not a flight will be delayed.

**Goal:**

Accurately predict whether a flight will be delayed or not.

**Dataset Characteristics:**

| **Variable** | **Significance** | **Type** |
| --- | --- | --- |
| CRS\_DEP\_TIME | It shows the actual departure time of flight. Broken down into 18 intervals between 6:00 AM and 10:00 PM | Numerical |
| CARRIER | It shows the code of 8 airlines | Character |
| DEP\_TIME | It shows the actual departure time of flight from origin. | Numerical |
| DESTINATION | It shows the code for the destination airport. Three airport codes: JFK (Kennedy), LGA (LaGuardia), EWR (Newark) | Character |
| DISTANCE | It shows the distance covered by the flight. | Numerical |
| FL\_DATE | It shows the date on which flight was departed. | Numerical |
| FL\_NUM | It shows the flight number. | Numerical |
| ORIGIN | It shows the origin airport. Three airport codes: DCA (Reagan National), IAD (Dulles), BWI (Baltimore-Washington Int's) | Character |
| Weather | It shows whether the weather is good or not. | Numerical |
| DAY\_WEEK | It shows the day of the week on which the flight departed. | Numerical |
| DAY\_OF\_MONTH | It shows the day of month on which the flight departed. | Numerical |
| TAIL\_NUM | It shows the tail number of flight. | Character |
| FLIGHT\_STATUS | It shows whether the flight was on time or not. | Character |

To begin with, we imported “FlightDelays” dataset in SAS using the proc import procedure and printed it.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, table

Description automatically generated

We found the missing values in all the character variables using the **proc means procedure.**

Before proceeding to the next step, we found that variable **dep\_time & crs\_dep\_time are in numerical format**, and hence needed to be changed in time format to get the accurate results. We fixed the two variables using the **put & input and formatted them using time5. format.**

Graphical user interface, text, application

Description automatically generated

**Handling Missing Data**

Perform the necessary “Handling Missing Data” operations to the missing values. In order to handle the missing data, firstly, we specified a format for the variables so that the missing values all have one value and the nonmissing values have another value. We used **Missingcount** and **Missingchar** to find out the missing values in numerical and character variables respectively.

Graphical user interface

Description automatically generated with low confidence

After that, we used the frequency procedure on the flight\_new dataset and formatted all the variables using the 2 created values to find the missing and non-missing values.

Text

Description automatically generated

Table

Description automatically generated Table

Description automatically generated Table

Description automatically generated

After finding out the missing values, we deleted them using the if statement and evaluated the result using the proc Means procedure.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

**New Data set:**

Create a new SAS dataset “FlightDelays” containing only one Origin plus a new variable called DelayedFlight with values of 1 for delayed flight and 0 for none.

In the following code, we created a new **Dataset “flightdelays”**. We selected origin as **“DCA”** and created a new variable **“DelayedFlight”** with values of 1 and 0 to check if flight is delayed or not. We also created a new variable **“delay\_time\_minutes”** to find out the difference of time between scheduled departure time and actual departure time i.e., crs\_dep\_time and dep\_time.

Text

Description automatically generated

Table

Description automatically generated

**Vertical Bar Chart**

Generate a table for the average delay per day for each airport and plot the vertical bar chart for the 7 days. To create the vertical bar chart for the 7 days, we selected the **delayed flights before 01/09/2004** in the datasetep. Then we created a table **“delayed\_avg”** using a **“SQL procedure”** and selected average of delayed time and grouped it by destination. Finally, we plotted the vertical bar chart using the **“sgplot procedure”**.

Graphical user interface, text, application, email

Description automatically generated

Chart, bar chart

Description automatically generated

**Mean Report & Scatter Plot:**

Mean Report

Graphical user interface, text, application

Description automatically generated

To get the mean report we have used proc SQL to create a new table ‘flight\_count’ where total number of flight count for each carrier and flight date would be recorded which will be used to derive the mean number of flights for which carrier which is stored in ‘mean\_flight’ table.

Table

Description automatically generated

Above is the report which shows the mean flights for every carrier. From the report it can seen that US and DL are the two busiest carriers where DH and CO have the lowest mean in the number of flights amongst all the carriers

Scatter plot

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

From the above scatterplot we can clearly comprehend that most days of the week Carrier RU has 6 flights per day but almost every week there is one day where it is 5 or less number of flights and during the end of the January 2004 there was a day where only 1 flight was operational which brings it mean to 5 flights per day

**Quantitative Variables – Histogram**

In the given data the given 2 variables are Quantitative Variables:

* distance
* delay\_time\_minutes

We have used the following code to get the histograms for both the variables

A picture containing logo

Description automatically generated

**Histogram for Distribution of Distance**

Chart, histogram

Description automatically generated

**Histogram for Distribution of delay\_time\_minutes**

Graphical user interface, application

Description automatically generated

To find out which of the variables have the highest variabilities we need to find the minimum and the maximum values for each of variables. Proc Univariate also provides us with the extreme values of distance and delay\_time\_minutes as show below.

| Distance | Delay\_time\_minutes |
| --- | --- |
|  |  |

To find the variabilities we need to subtract the value of lowest observation from the value of highest observation

Variability of Distance = 214(Value of highest observation) - 199(Value of lowest observation) = 15

Variability of delay\_time\_minutes = 187(Value of highest observation) – (-13)(Value of lowest observation) = 200

Hence, delay\_time\_minutes has the largest variabilities.

**Pivot tables – Summarization**

The below code shows the data summarization using pivot table

Graphical user interface, text, application, email

Description automatically generated

Pivot tables

|  |  |
| --- | --- |
|  |  |

* From the pivot table ‘Flight Status by Origin’ we can determine that the flights originating from BWI are having the highest number of delayed flights till the time they reach their destination whereas flights originating for DCA have the lowest delay flights amongst flights originating from BWI, DCA and IAD.
* Pivot table ‘Flight Status by different carrier’ depicts that the US carrier has the least number of delayed flights as compared to all the other carriers.
* It can be determined that as the distanced of the flights increases the possibility of the flight being delayed by the time it reaches its destination also increased, from the pivot table ‘Destination and Distance affecting flight table’
* The pivot table ‘Weather affecting flight status’ concludes that the number of delays in flights increase when the weather condition on a given is not good as compared to when it is good.

**Advance (MC2):**

We have already done regular programming with our Flight Delay dataset. Moving forward, we will attempt to use an advanced programming strategy for our data set to anticipate the elements that cause flight delays.

**Data Reduction:**

           Data reduction is a strategy for reducing the volume of data while maintaining data integrity. When we utilise a large quantity of data for our research, it causes storage challenges, increases execution time, and makes it harder to identify the data we want. We must be cautious while reducing data because the reduction will not effect our outcomes. The findings before and after the reduction must be identical or nearly identical, and the quality of our data must be maintained ("What Is Data Reduction? Techniques").

To increase the quality of our dataset, we used the data reduction approach for our Flight Delay data. We have analyzed our data to how we can improve its performance of it without disturbing our results. From our Analysis, we came up with the below points.

**Variables used in our Programming or Analysis:**

Below are the variables we utilized in our programming. Directly or indirectly, we used them to calculate the delay time of different flights. Since they are useful for our analysis, we can’t remove them for data reduction.

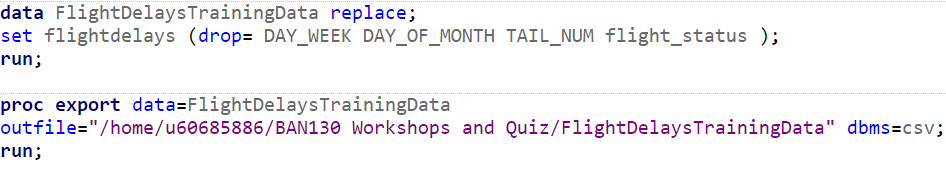
* CRS - Departure time – Departure time and CRS departure time are the main variables that help us to calculate the flights' delays. So, we can’t use these variables for data reduction, if we do it will affect the accuracy of our results.
* Origin – Represents the Flight start place which has been used for our analysis.
* Destination – Destination for the flights
* Carrier – The flight carrier that carries passengers from origin to destination
* Flight Number – Number for the Flights
* Weather – Weather faced by the flights during their travel time
* DelayedFlight - Whether the flight reaches its destination on time or is delayed.

**Not used in our Programming or Analysis:**

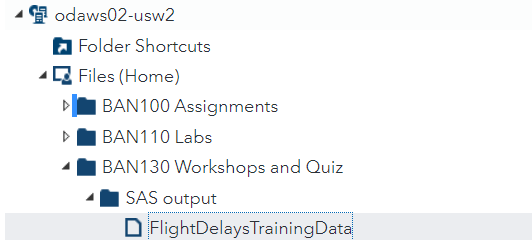
The below variables are not references in our program. For Air transport, the week and month will not have any influence on their travel time. We already have DelayedFlight variable, showing 0 for False and 1 for True, we don’t need Flight\_Status for our analysis.

* Day of the Week
* Day of the Month
* Tail number
* Flight\_Status

For our data set, the removal of the above variables will not affect the accuracy of our results. We came to this conclusion after analyzing any specific patterns followed for flight delays based on the above variables. Since the days and tail number are not contributing much to flights delays. We removed the above variables from our dataset and created a new file as **the Flights Delays Training Data** CSV file.



After executing the above statements, new csv file was generated and stored in above mentioned path.





**Data Conversion:**

While applying our data to the analytics tool, sometimes the machine learning algorithms didn’t comply with the string of characters. Here, we convert the Character data into Numeric to avoid such types of issues. Below character variables will be converted to numeric.

* Carrier – There are 8 distinct values (CO, DH, DL, MQ, OH, RU, US, UA)
* Origin – There are 3 distinct values (BWI, DCA, IAD)
* Destination – There are 3 distinct values (EWR, JFK, LGA)

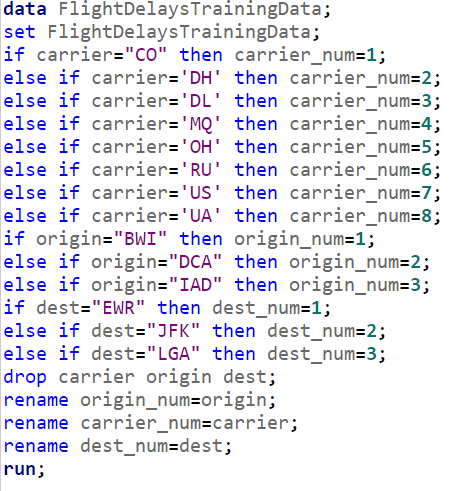
At first, we converted the data of the above 3 variables by sequence order. We transformed their data in the below order.

CO -1, DH -2, DL -3, MQ -4, OH -5, RU -6, US -7, UA -8

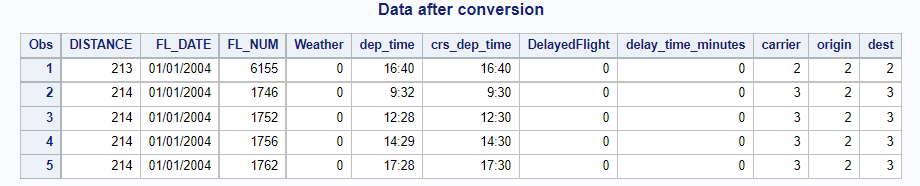
BWI -1, DCA -2, IAD -3

EWR -1, JFK -2, LGA -3

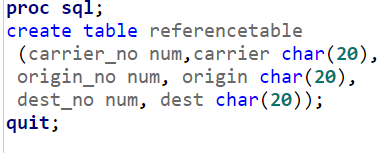
We have used the below code to convert our character data into numeric values.



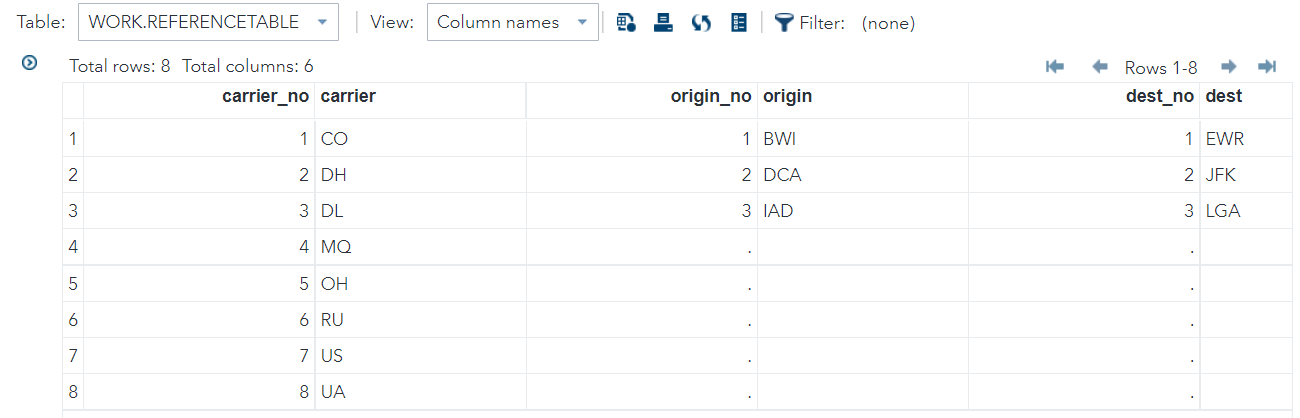
After executing these data step, we got the below output. Here we can see the data for the carrier, origin and destination has been changed to numeric values.

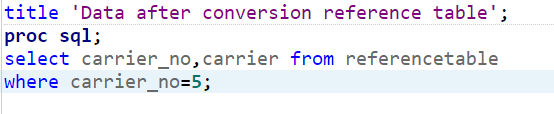


We have created a reference table using PROC SQL to show the values of the transformed variables.

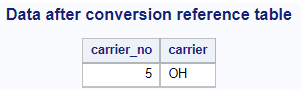


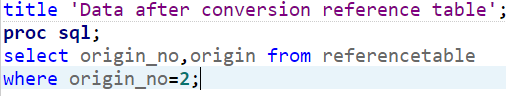
Below is the table created by the above statement.

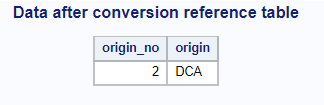




After executing the above select statement, we can see the result show for carrier number 5 with its corresponding character variable.





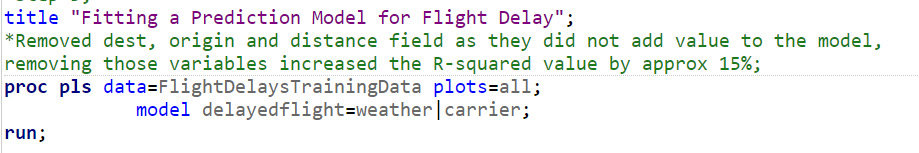


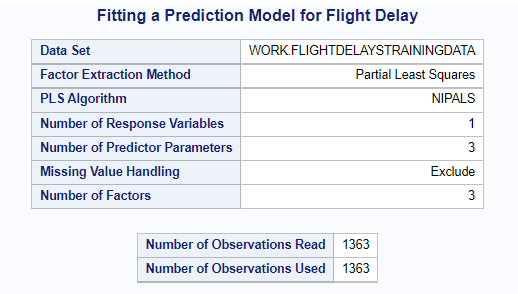
**Prediction:**

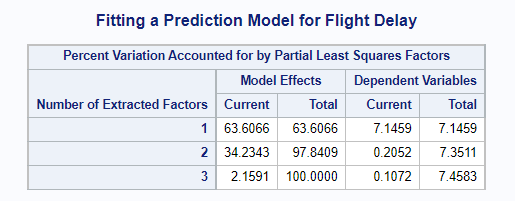
When flights are delayed, airlines must give additional services to customers to alleviate their complaints. They occasionally have to return money for their tickets, and the crucial thing is that this does not lead to client unhappiness. The impact of a flight delay on an airline's business might be significant. They risk incurring significant losses if they give delayed services to their passengers. The precise estimate of flight delays is critical for airlines (Yazdi et al.). They may strengthen that particular area to solve this issue and raise customer happiness if they can forecast the cause of flight delays.

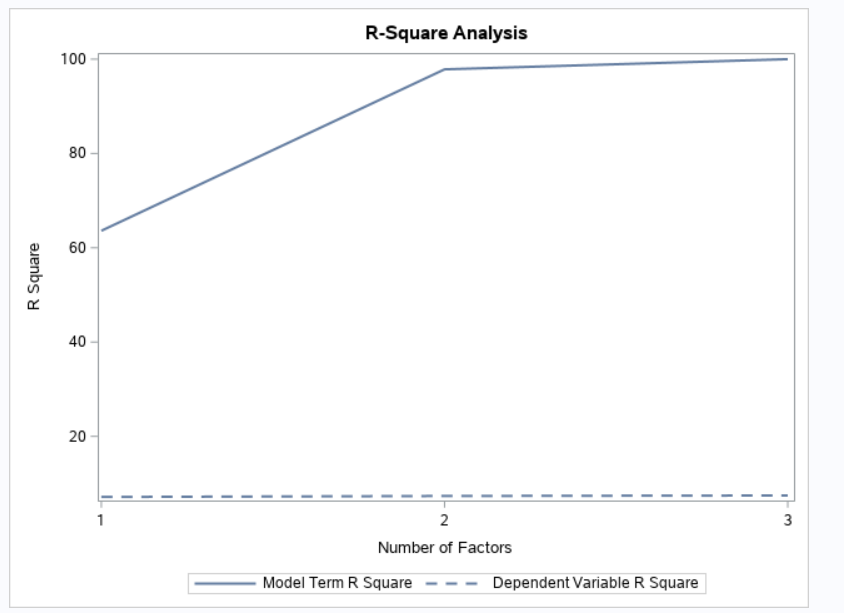
The flight delay can be calculated in a variety of ways. For our study, we employed the PLS technique to calculate our forecast. PLS (partial least squares) is a procedure similar to PROC GLM and PROC REG that reduces prediction error and explains response variations. Proc PLS will limit the predictor set to a minimal level and perform least square regression on these variables rather than all variables. PLS, unlike other Procs, has no fixed hypothesis. Because the predictors may be assessed with inaccuracy, PLS becomes more resilient to measurement uncertainty (“What Is Partial Least Squares Regression?”).

Initially, we have to try to extract the flights' delays by using all the variables like origin, destination, distance, weather, and carrier. But when we tried to implement the procedure, the R-Square value was very low which shows our model prediction is not up to the mark. We tried to reduce our variables and concentrate on weather and carrier for the flight delays prediction.

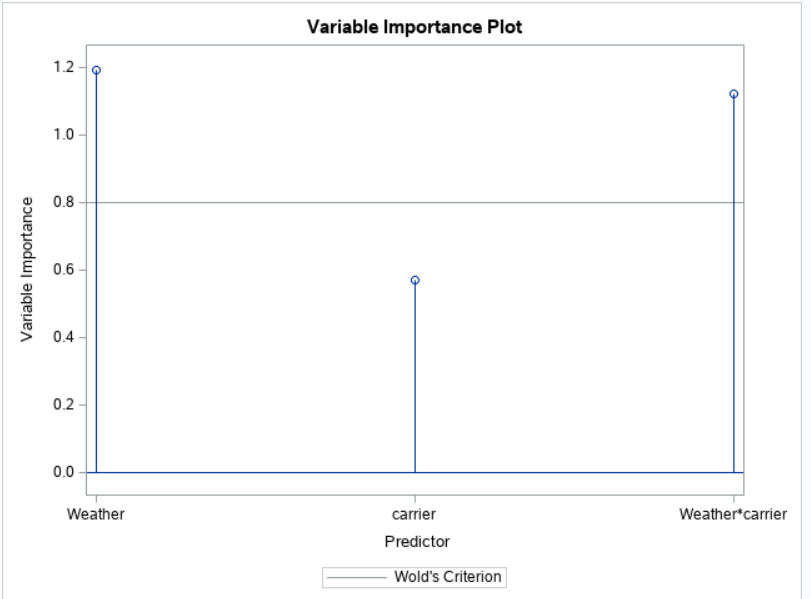
We have used the below codes to find the predictor value for flight delays.

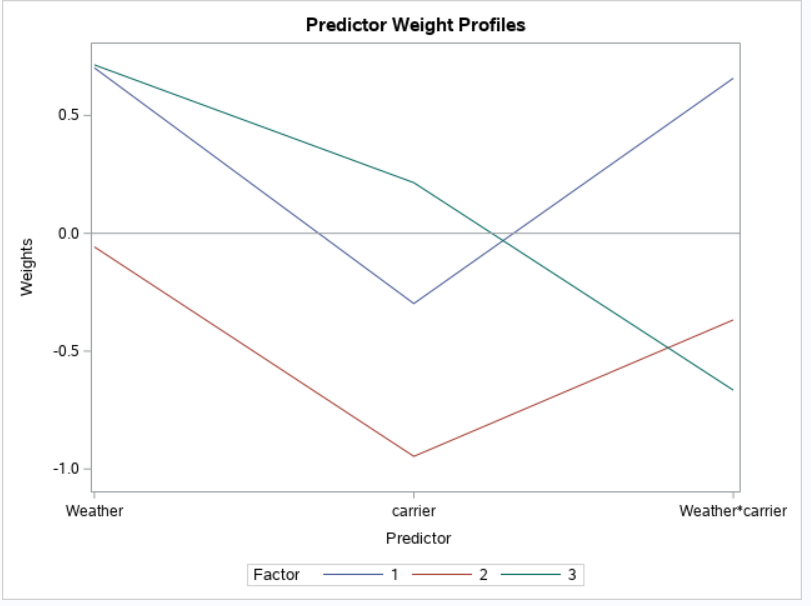


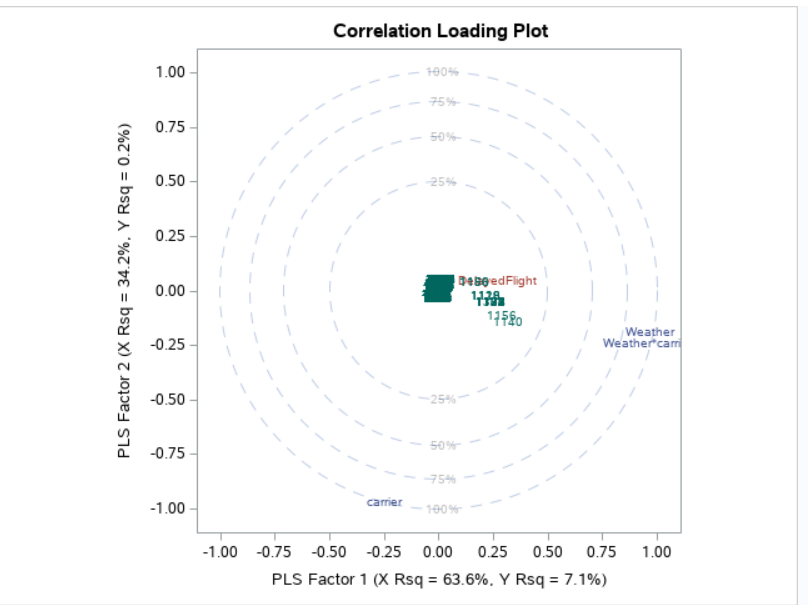


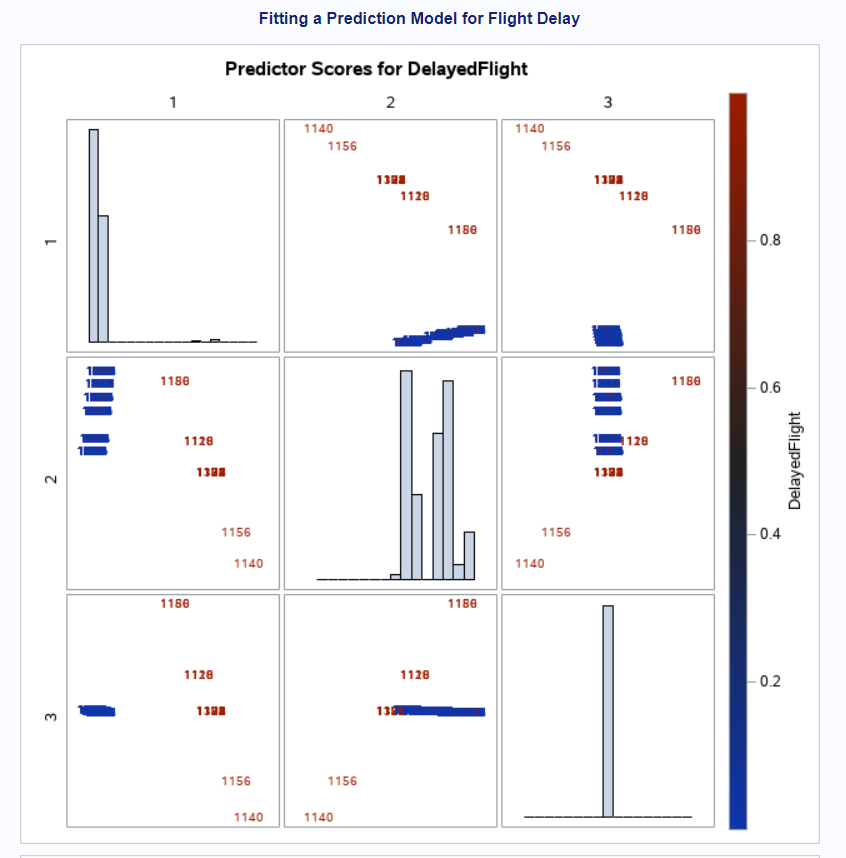


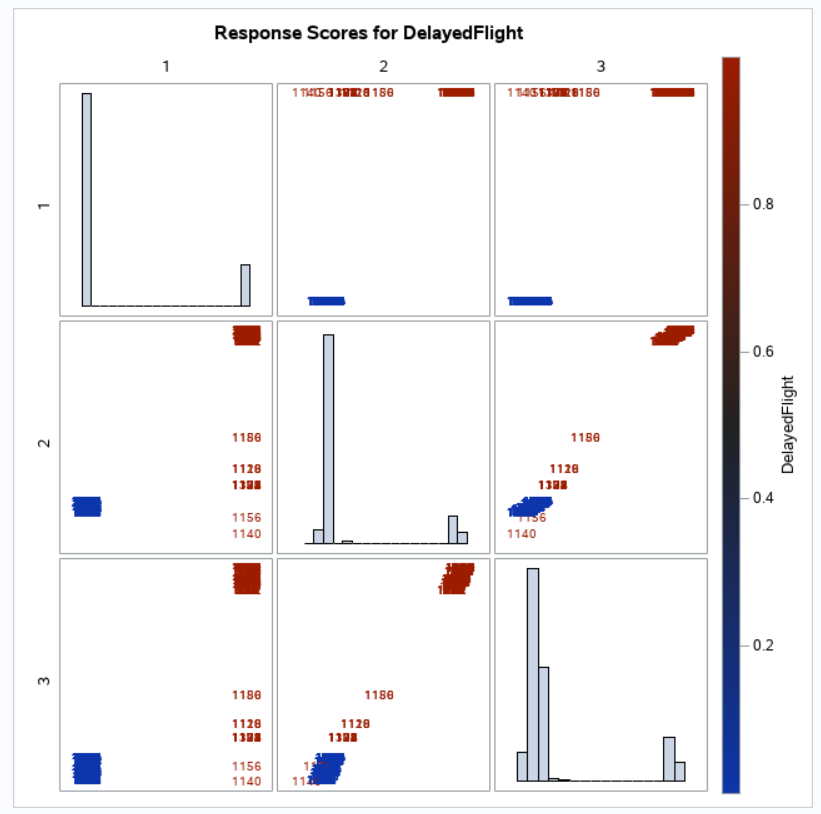
R-Square Analysis for the variables weather and carrier. This model shows us that rate of prediction is high, which is more than 62%. This shows our prediction is robust. Here when we compared them with other variables the output of R Square is less than 45%, which shows the uncertainty of our prediction.

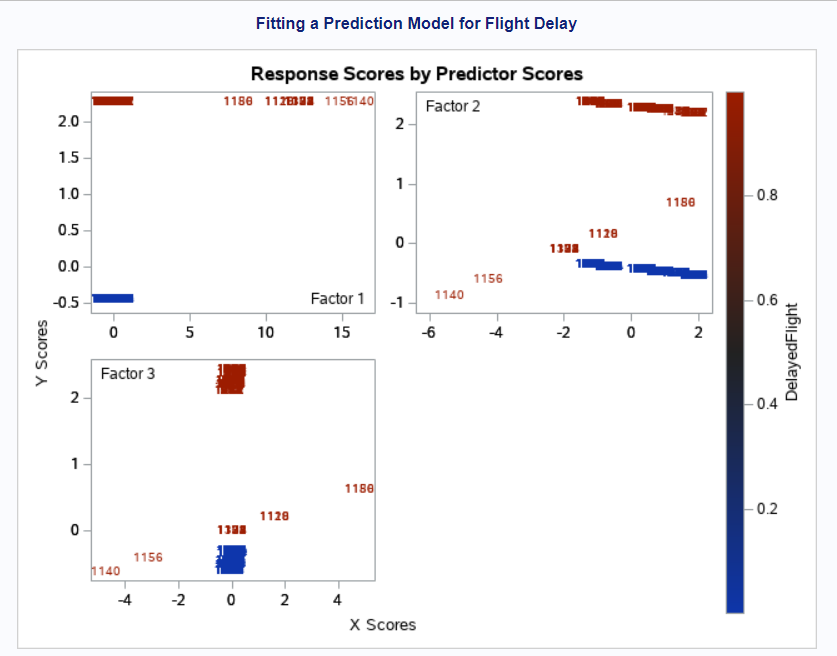


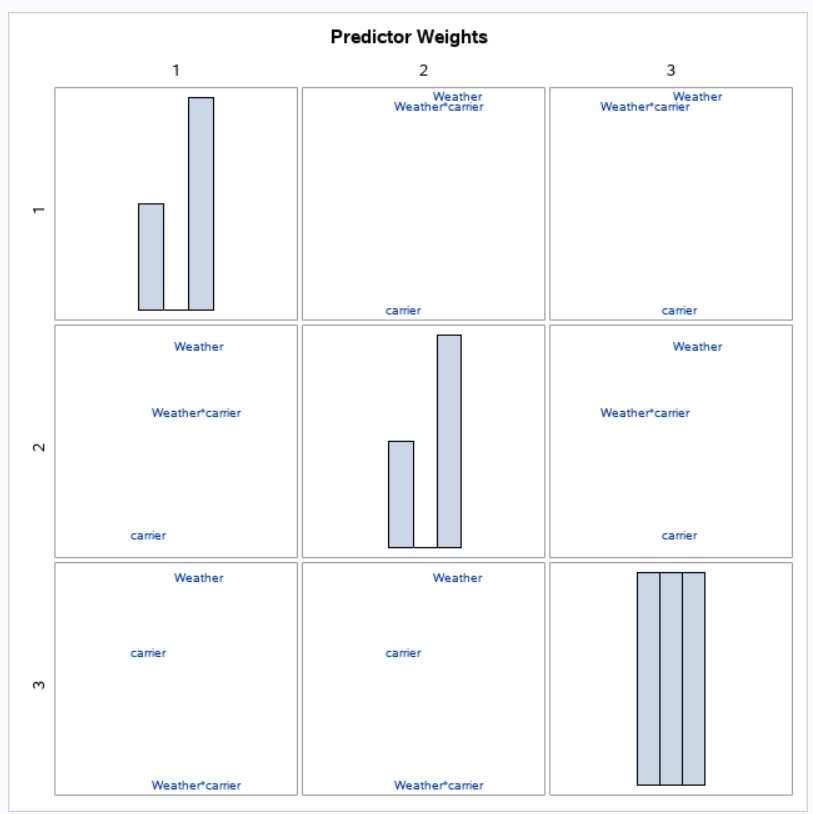


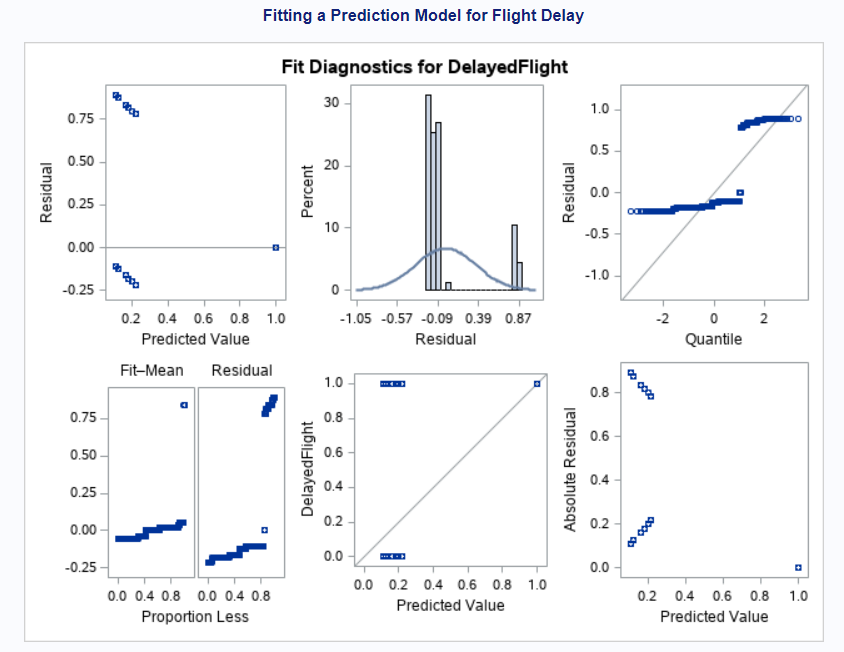












The results of our PLS algorithms are shown above; in essence, they compare weather and carrier as a forecast indication.

Even if the weather is normal, we anticipate that carrier RU will arrive late at the destination based on our research. We could make a variety of assumptions or forecasts here, but we chose weather and carrier because of their high R-square rate.

If we utilize additional factors to make a forecast, we'll conclude that MQ flights from DCA to LGA are frequently delayed. We can't infer that our prediction is right in this situation since the distribution of variables is so close to each other.

**Challenges:**

* Conversion of dep\_time crs\_dep\_time variables to proper time format.
* Faced issue during the calculation of delayed time minutes.
* Prediction

**Benefits:**

* Ability to analyze datasets
* Problem-solving
* Real-time examples of Data reduction and conversion
* Able to experience how the actual analytics can happen in SAS programming.
* Experience of end to end analytics project, from data cleaning to prediction.

**The idea for further development:**

           Prediction is one of the important goals of the analysis. Due to our limited knowledge, we are not able to get deep into the prediction part. If there are more data related to flight routes or detours during their travel can help calculate the prediction.

**References**

1. “What Is Data Reduction? Techniques.” Binary Terms, 14 Sept. 2020, binaryterms.com/data-reduction.html.
2. Yazdi, Maryam Farshchian, et al. “Flight Delay Prediction Based on Deep Learning and Levenberg-Marquart Algorithm.” Journal of Big Data, vol. 7, no. 1, 26 Nov. 2020, 10.1186/s40537-020-00380-z.
3. “What Is Partial Least Squares Regression?” Minitab.com, 2019, support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/partial-least-squares-regression/what-is-partial-least-squares-regression/.