**BANA 6043 PROJECT**

**Name: Madhava Chandra UCID: M12399645**

**Background**: Flight landing.

**Motivation**: To reduce the risk of landing overrun.

**Goal**: To study what factors and how they would impact the landing distance of a commercial flight.

**Data**: Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). See two Excel files ‘FAA-1.xls’ (800 flights) and ‘FAA-2.xls’ (150 flights).

**Summary**:

The purpose of this study was to determine a safe landing distance. The data was obtained from FAA and contained about 850 unique values pertaining to flight landing. Few of the key variables involved were flight speeds in air and ground, height of the flight prior to landing and duration of the flight. The data was first analyzed for outliers and then correlations were examined. Upon then weeding out the nonsignificant variables, the landing distance was regressed over the key variables and it was concluded that the speed in air and ground explain the landing distance. Also two models have been built and depending on the user preference, a more complex mix of models(based on whether the air speed is captured) or a simpler model based only on ground speed can be used to predict the landing distance.

1. Data exploration and data cleaning

**Goal:** This module focuses on understanding the variables present.

**Observations and Decisions:** (SAS code and outputs to follow)

* First summary of the loaded data was observed for each of FAA1 and FAA2.
* It was then concluded that there are duplicates in this data along with empty observations.
* Post removing these, the proc means was studied for each of these to conclude that these values belong to the same population, based on which the data was combined.
* Duration and Air speed were still observed to have missing values.
* Since we are not yet sure how critical duration is to the landing distance, removing the entire set of data for a missing duration might lead to loss of significant information.
* For air speed, once the distributions were plotted, it became evident that the speed has a truncated distribution with a clear lower bound of 90, implying that the observations below that threshold were not being observed for a reason rather than randomly being missing/not being collected. Thus, the air speed values were concluded to be critical and were not to be removed.
* Other variable distributions were also observed and were largely normal
* Correlation between variables was also looked at and for low correlations, plots were drawn to spot any non-linear correlation
* Since passenger count and duration didn’t have much correlation with distance, they were dropped from the data
* Prior to modeling, it was to be tested if the aircraft class has any impact on distance. A t test was run to test the difference in means of the distance and it was proved that it has a significant impact. As a result of which, the data was imputed with a dummy variable for the airline class.

**Modeling**

* In the first run, with pitch included, since the coefficient could not be concluded as non-zero, regression was rerun with that variable excluded.
* A choice of 2 models is being presented herewith to the end user. This is to not discard the relatively few air speed values.
* One model will be used when speed air is captured, other when speed air is missing. If the end user wants a more simplistic approach, the model with speed ground could be used. This has a slightly lower R squared but not significantly different. The 2 or 1 model approach appears to be a better option than imputing the air speed values.
* Since air speed had higher correlation to the distance, when it was present and was also resulting in much higher R squared, the speed ground was ignored since it would cause multicollinearity.

**SAS Code:**

*/\*1-Importing Input files FAA1 and FAA2\*/*

*FILENAME REFFILE '/home/satyasmc0/Stat computing class/FAA1.xls';*

*PROC IMPORT DATAFILE=REFFILE*

*DBMS=XLS*

*OUT=WORK.faa1;*

*GETNAMES=YES;*

*sheet=FAA1;*

*run;*

*FILENAME REFFILE '/home/satyasmc0/Stat computing class/FAA2.xls';*

*PROC IMPORT DATAFILE=REFFILE*

*DBMS=XLS*

*OUT=WORK.faa2;*

*GETNAMES=YES;*

*sheet=faa2;*

*proc means data=faa1 n mean std range min max nmiss;*

*proc means data=faa2 n mean std range min max nmiss;*

*run;*

*/\*2-from the means procedure, it is evident that the faa1 and faa2 belong to the same population*

*thus data can be combined\*/*

*data faa3;*

*set faa1 faa2;*

*proc means data=faa3 n mean std range min max nmiss;*

*run;*

*/\*3- Checking for duplicates\*/*

*proc sort data= faa3 nodupkey*

*out = faa4;*

*by pitch;*

*proc print data=faa4;*

*run;*

*/\*4- Removing missing values\*/*

*data faa4;*

*set faa4;*

*if missing(aircraft) then delete;*

*run;*

*proc sort data=faa4;*

*by aircraft;*

*proc means data=faa4 n mean std range min max nmiss;*

*title 'All data summary- unique';*

*run;*

*/\*No missing values found in summary- duration is missing for 50 values\*/*

*/\*5- Checking and removing abnormal values\*/*

*data faa\_normal;*

*set faa4;*

*if duration=. then miss='yes';*

*if speed\_ground=. then miss='yes';*

*if speed\_air=. then miss='yes';*

*if height=. then miss='yes';*

*if distance=. then miss='yes';*

*if pitch=. then miss='yes';*

*if duration<=40 and duration <> . then abnormal='yes';*

*if speed\_ground<30 or speed\_ground>140 and speed\_ground <> . then abnormal='yes';*

*if (speed\_air<30 or speed\_air>140) and speed\_air <> . then abnormal='yes';*

*if height<6 and height <> . then abnormal='yes';*

*if distance>6000 and distance <> . then abnormal='yes';*

*run;*

*proc sort data=faa\_normal;*

*by abnormal miss;*

*proc print data=faa\_normal;*

*proc means data = faa\_normal n nmiss min max ;*

*run;*

*/\*Since abnormal values are a very small percentage of the entire data, deleting them\*/*

*data faa\_normal;*

*set faa\_normal;*

*if abnormal='yes' then delete;*

*drop abnormal;*

*drop miss;*

*proc means data = faa\_normal n nmiss min max ;*

*run;*

*/\*We end up with 831 values with their summary\*/*

*/\* Comparing distributions indicates the speed\_air is a truncated dist so better to seggregate\*/*

*proc chart data= faa\_normal;*

*vbar speed\_air;*

*run;*

*proc chart data= faa\_normal;*

*vbar speed\_ground;*

*run;*

*data faa\_normal;*

*set faa\_normal;*

*if speed\_air= . then Group = 0; else Group = 1;*

*proc print data=faa\_normal;*

*proc means data = faa\_normal n nmiss min max;*

*run;*

*/\*Exploring Variable distributions\*/*

*proc univariate data=faa\_normal;*

*class group;*

*var speed\_ground;*

*histogram speed\_ground;*

*proc univariate data = faa\_normal;*

*class group;*

*var height;*

*histogram height;*

*proc univariate data = faa\_normal;*

*class group;*

*var pitch;*

*histogram pitch;*

*proc univariate data = faa\_normal;*

*class group;*

*var no\_pasg;*

*histogram no\_pasg;*

*/\*Height, pitch and passenger count variables are nearly normal\*/*

*/\*Exploring Variable Correlations\*/*

*proc corr data=faa\_normal;*

*var duration speed\_air speed\_ground no\_pasg pitch height distance;*

*run;*

*/\*The correlation matrix shows us that there is no impact of passenger count and duration on distance*

*Also as expected speed air and speed ground are heavily correlated\*/*

*/\*We can plot to see if there is any non-linear correlation\*/*

*proc gplot ; plot distance\*height;*

*proc gplot ; plot distance\*pitch;*

*proc gplot ; plot distance\*no\_pasg;*

*/\*Drop duration and passenger count\*/*

*data faa\_trim;*

*set faa\_normal;*

*drop duration no\_pasg;*

*run;*

*proc means data=faa\_trim n nmiss min max;*

*run;*

*/\*Impact of aircraft class\*/*

*proc ttest data=faa\_trim;*

*class aircraft;*

*var distance;*

*title 'Mean distance across Airbus and Boeing';*

*run;*

*/\*pvalue<alpha so refer Satterthwaite section implying unequal variances\*/*

*/\*p value of ttest implies that the mean equality can be rejected*

*Created a dummy variable that can be used in regression\*/*

*data faa\_final;*

*set faa\_trim;*

*if aircraft= 'boeing' then planetype = 0; else planetype = 1;*

*run;*

*/\*Regressing the landing distance against variables*

*- building 2 models in order to not delete out the air speed\*/*

*proc reg data = faa\_final;*

*model distance = planetype speed\_air pitch height;*

*title 'Regression when air speed is available';*

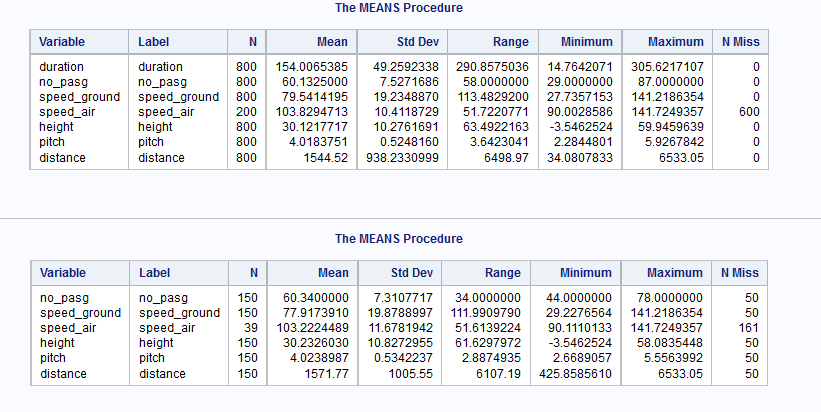
*proc reg data = faa\_final;*

*model distance = planetype speed\_ground pitch height;*

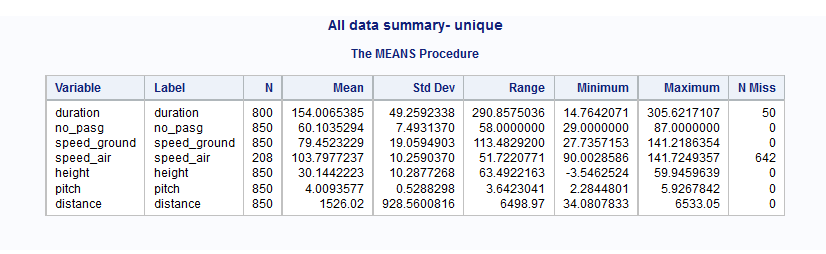
*title 'Regression when air speed is unavailable;*

**SAS Outputs:**

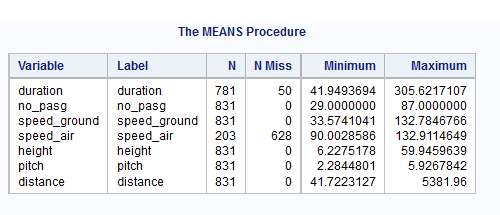
1. Initial summary of data when it was loaded



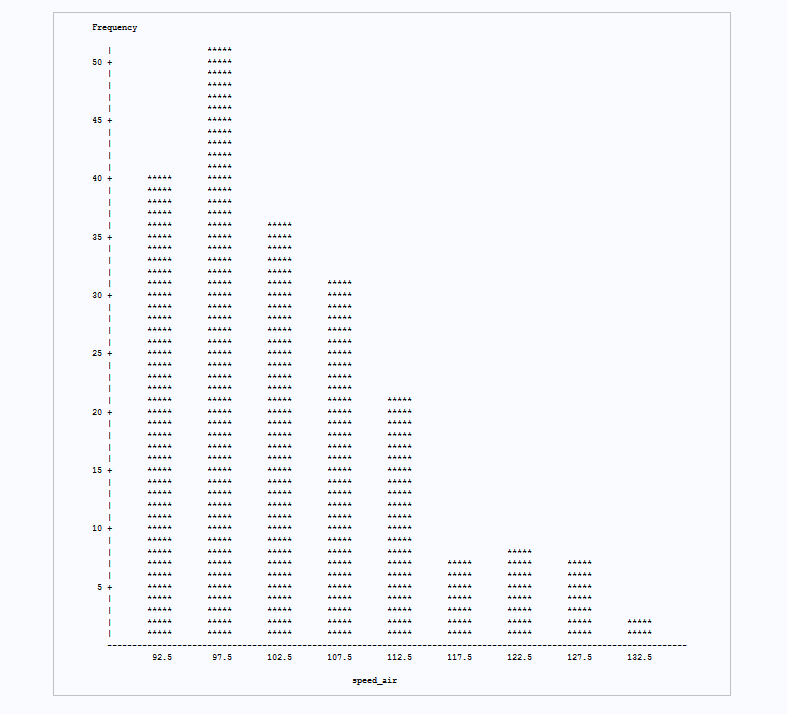
1. Data summary once duplicates were removed

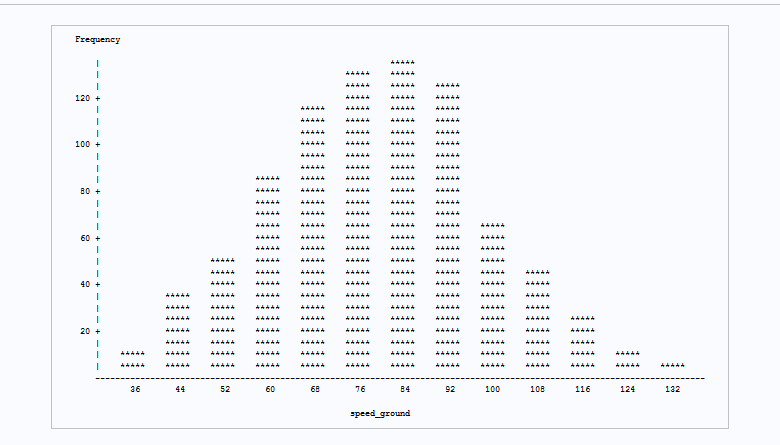


1. Summary when abnormal values were removed- 831 values remain

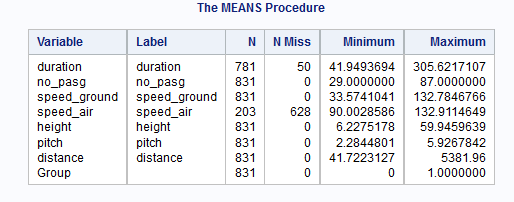


1. Since air speed values had most blanks, examining the distributions of air speed and air ground

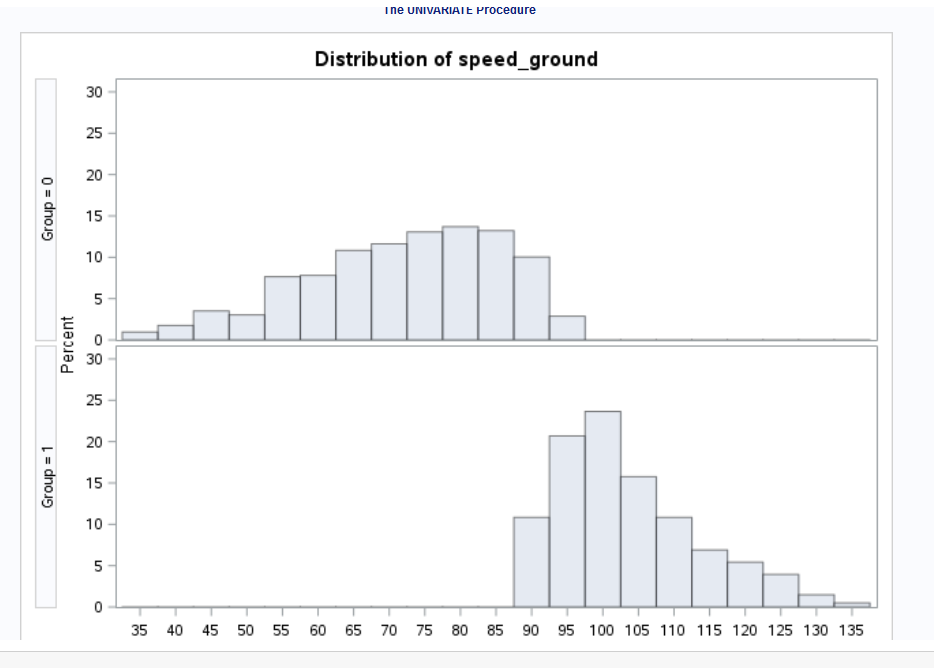


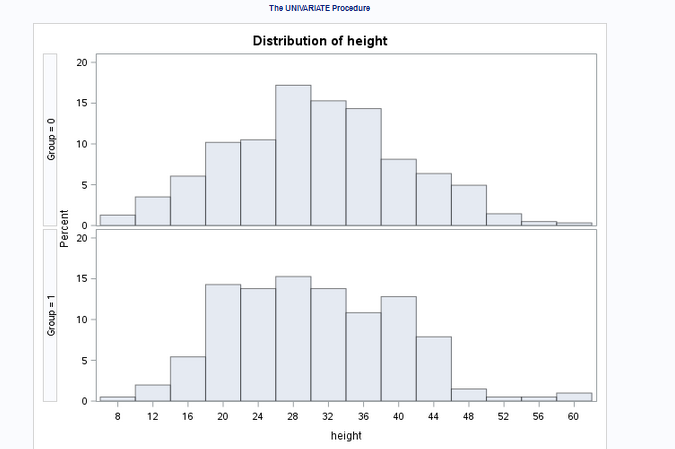


1. Since air speed values were to be retained for the model, created two groups

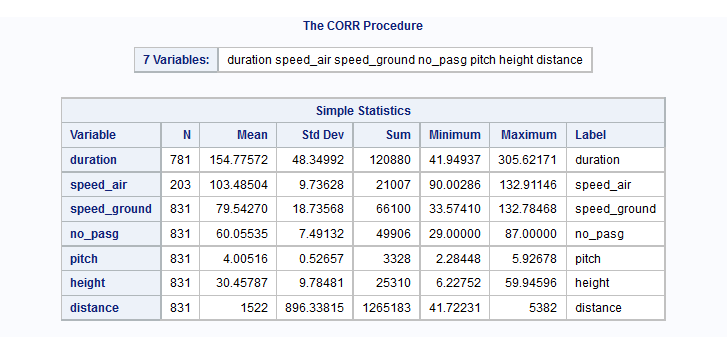


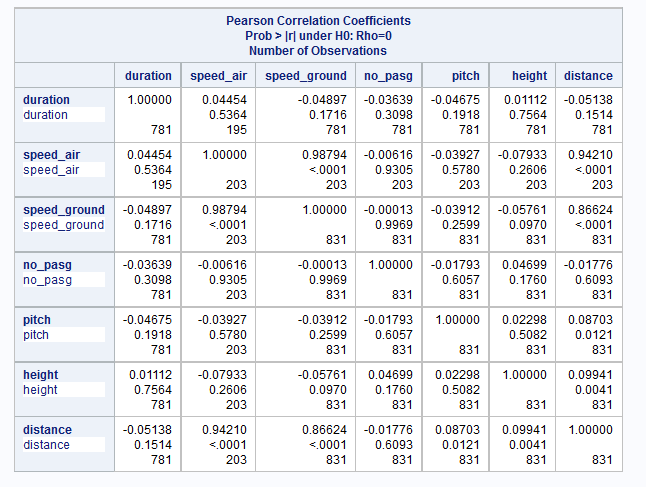
1. Creating groups made the difference in distributions very clear for speed ground, while the rest were largely normal.



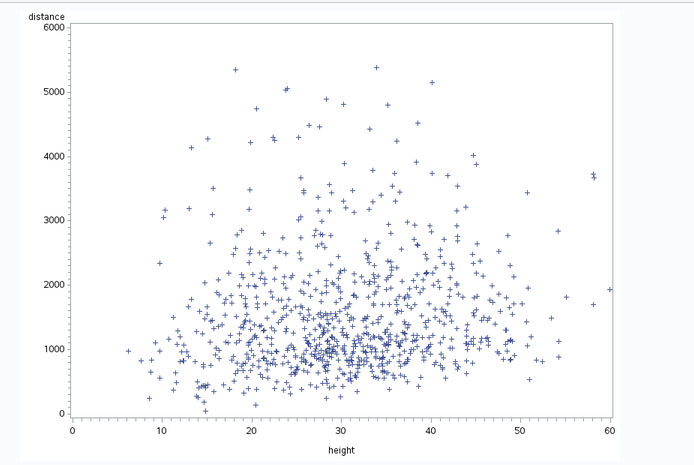


1. Understanding linear correlations

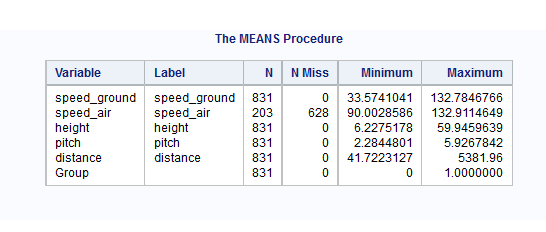




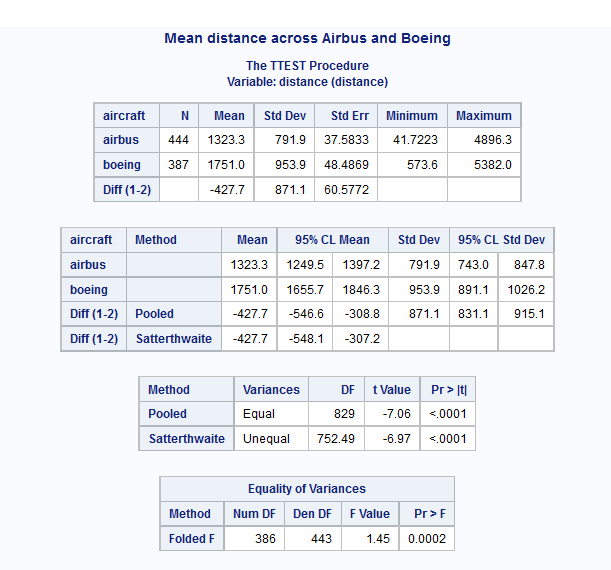
1. Checking for non-linear correlation



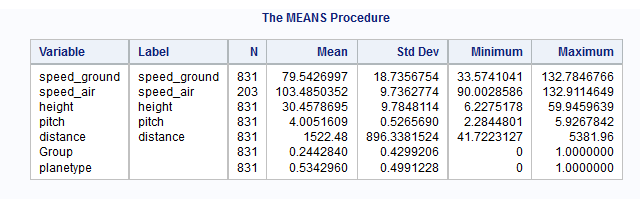
1. Dropping duration and no of passengers due to poor correlation



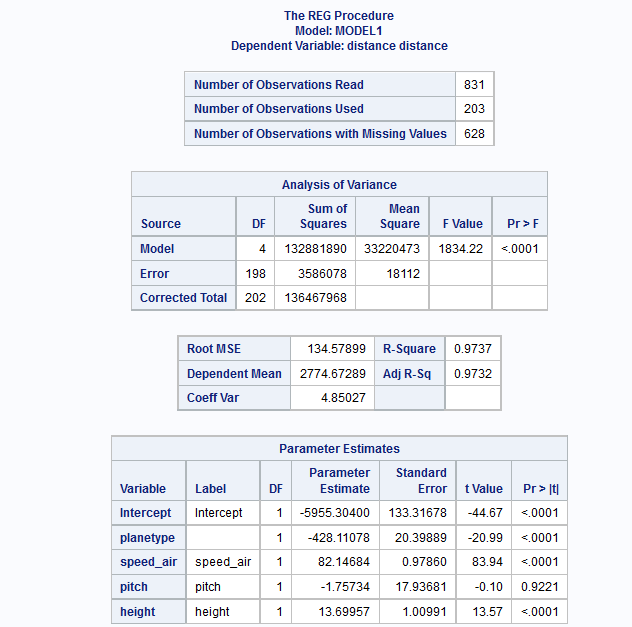
1. Examining impact of aircraft – Boeing and airbus: concluded there is a difference



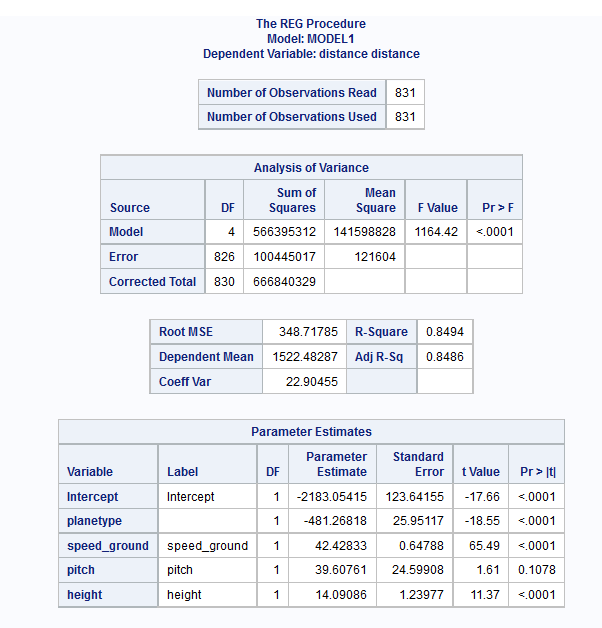
1. Created dummy variable for aircraft and ran correlation check



1. Regression for the group where air speed is available

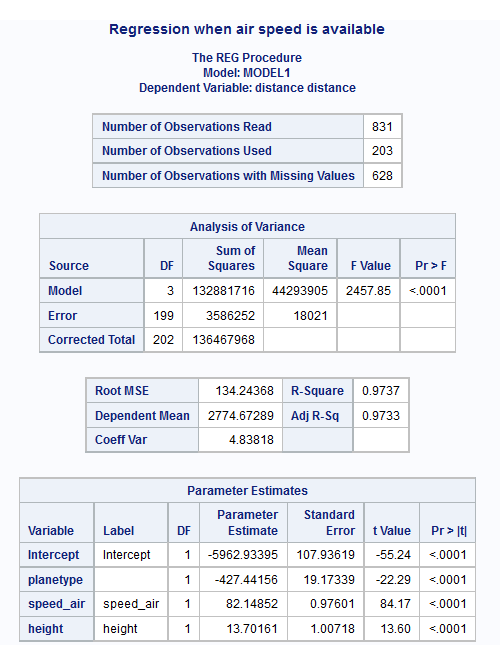


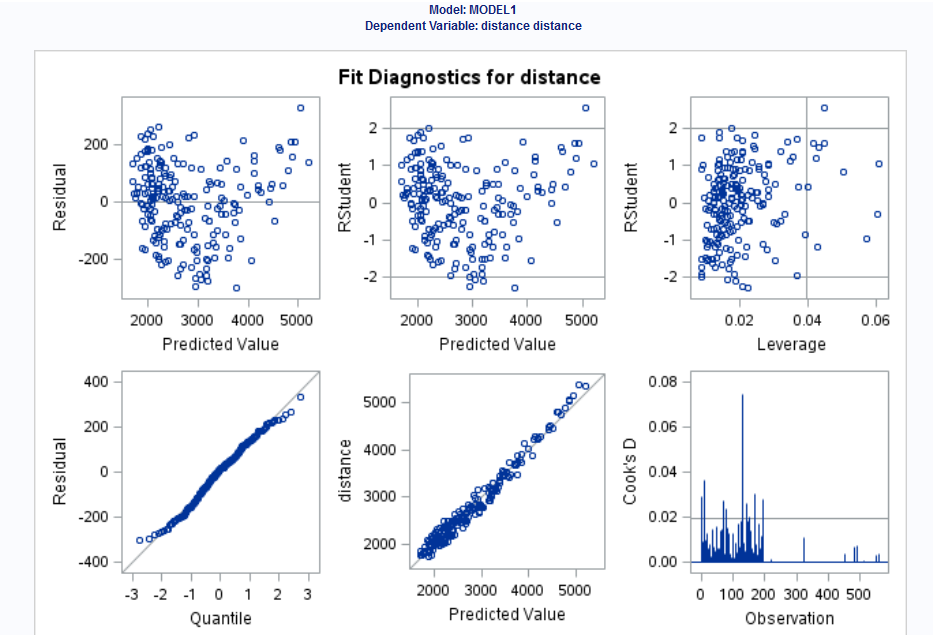
1. Regressing with ground speed



Since pitch is shown to have 0 coefficient in both cases( high p value), rerunning the regression without pitch

1. Revised outputs when pitch is removed





1. Final output of regression with ground speed

