**RELATIONSHIP BETWEEN VARIABLES:**

Now that we have cleaned the dataset, we would now like to understand the impact of every factor on the Distance variable using linear regression modelling.

1. **PLOTS:**

First, let us try plotting the distribution of Distance with respect to some of the variables we have in our data set.

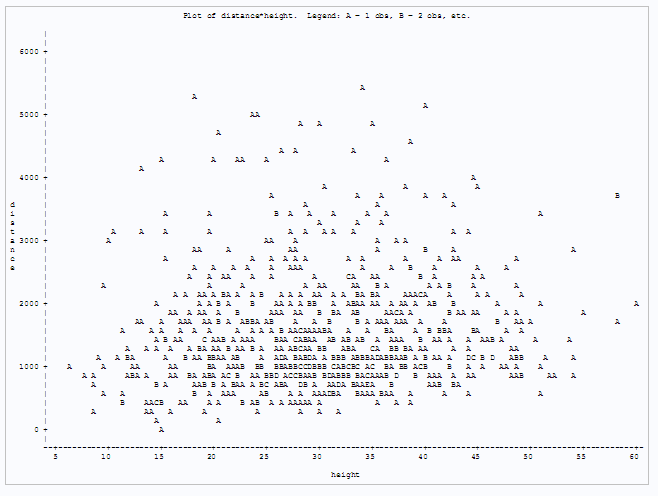
1. DISTANCE VS HEIGHT

CODE

proc plot data = proj1sc.FLIGHTCLEANED;

plot distance\*height;

run;



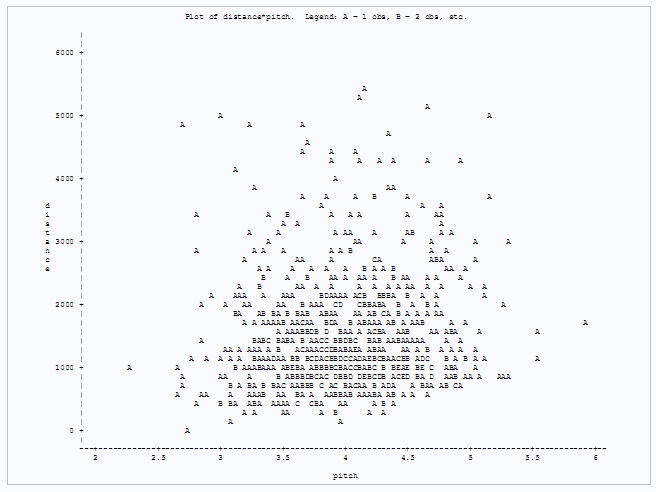
Looking at the graph, we are not able to make any inference about the relationship between distance and height.

1. DISTANCE VS PITCH

proc plot data = proj1sc.FLIGHTCLEANED;

plot distance\*pitch;

run;



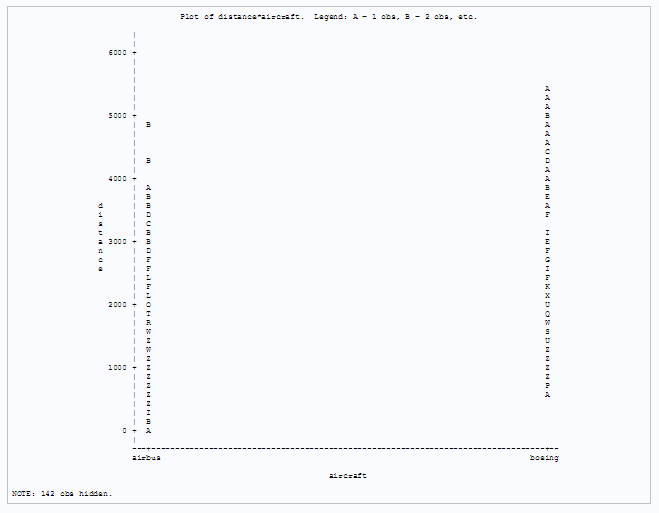
Looking at the graph, we are not able to make any inference about the relationship between distance and pitch.

1. DISTANCE VS TYPE OF AIRCRAFT

proc plot data = proj1sc.FLIGHTCLEANED;

plot distance\*aircraft;

run;



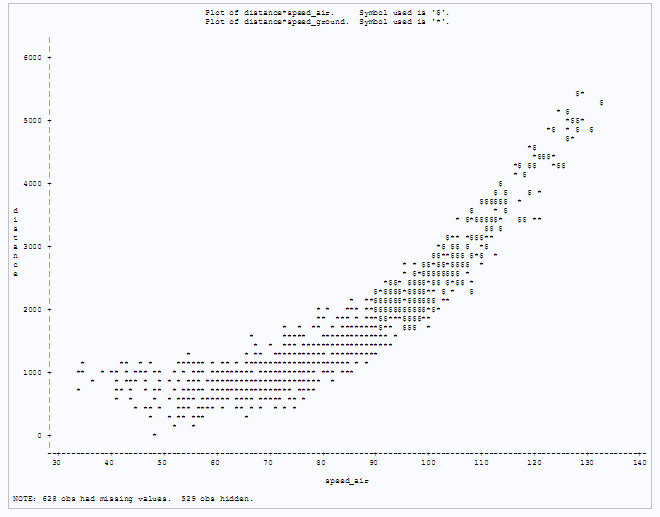
Looking at the graph, we can say that the distribution of distance between the two types of aircrafts is slightly different. So this variable might have an impact on the prediction of landing distance.

1. DISTANCE VS SPEED VARIABLES

proc plot data = proj1sc.FLIGHTCLEANED;

plot distance\*speed\_air = "$" distance\*speed\_ground = "\*" / overlay;

run;



Looking at the graph, we can see that the Speed\_air and Speed\_ground almost has the same relationship with the distance variable but the value of speed\_air starts only after a value of 90. One take away from the graph is that both the speed variables might be useful in predicting the landing distance.

1. CORRELATION ANALYSIS:

Next step would be to run a correlation analysis on the data set to identify the significant variable to be considered for the prediction of distance.

* 1. CODING THE CATEGORICAL VARIABLE

Now to run a correlation analysis, firstly lets convert the categorical variable aircraft into numeric values. Only then we will be able to run a regression analysis.

/\* Convert the categorical variable into a numerical condition \*/

data proj1sc.Flight;

set proj1sc.flightcleaned;

if (aircraft = "boeing") then type = 0;

else type = 1;

drop aircraft;

run;

1. PAIRWISE CORRELATION TO UDERSTAND CORRELATION BETWEEN ALL VARIABLES

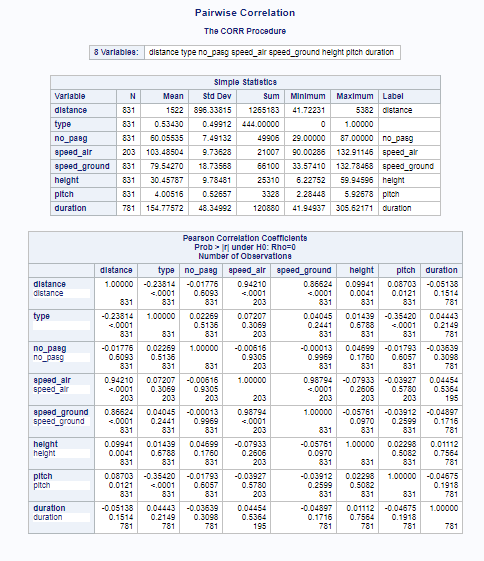
COMPUTING PAIRWISE CORRELATION:

proc corr data = proj1sc.flight;

var distance type no\_pasg Speed\_air Speed\_ground height pitch duration;

title "Pairwise Correlation";

run;



We find that the variables Speed\_air and Speed\_ground are highly correlated by the order of 98.7%. While building the model we surely need to consider their impact on inflating the predictions due to this multicollinear relationship.

1. CORRELATION OF VARIABLES WITH THE DISTANCE VARIABLE

The major area of interest is in understanding the correlation of all the variables to the distance variable.

CORRLATION WITH DISTANCE:

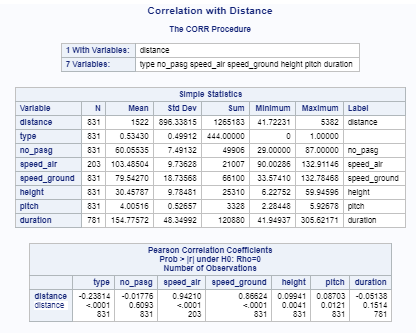
proc corr data = proj1sc.flight;

var type no\_pasg Speed\_air Speed\_ground height pitch duration;

with distance;

title "Correlation with Distance";

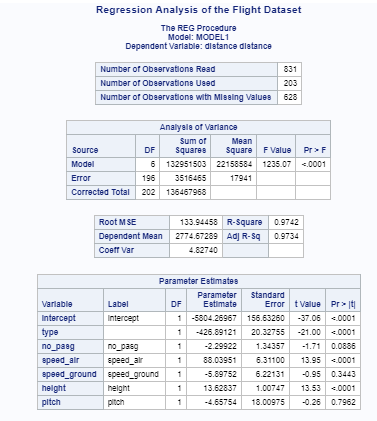
run;



We find that excepting the duration variable, all other variables have a significant correlation with the distance variable at 95% confidence level. Since the NULL hypothesis that rho = 0 for the distance variable cannot be rejected, we can leave the duration variable from our analysis.

REGRESSION MODELLING

proc reg data = proj1sc.flight;  
model distance = type no\_pasg Speed\_air Speed\_ground height pitch;  
title "Regression Analysis of the Flight Dataset";  
run;



ANALYSIS OF RESULTS

When we try interpreting the results, we find the following two conditions:

1. The model has used only 203 observations out of the 831 obseravations in the data set. This is due to the missing values in Speed\_air variable.
2. We also find that despite both the speed variables having a large positive correlation with the distance, we are seeing that one of the speed variables is having a negative coefficient in the regression model which is counter intuitve. This might be due to the correlation between the speed variables. We need to explore the effect of multi collinearity on the regression model.

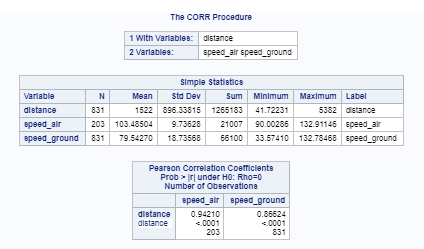
DISTANCE VS Speed Vriables

proc corr data = proj1sc.flight;

var speed\_air speed\_ground;

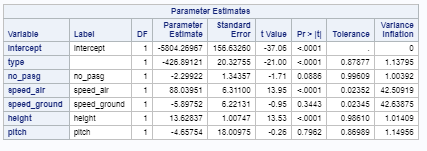
with distance;

run;



COMPUTING VIF AND TOLERANCE:

proc reg data = proj1sc.flight;  
model distance = type no\_pasg Speed\_air Speed\_ground height pitch /vif tol;  
title "Regression Analysis of the Flight Dataset";  
run;



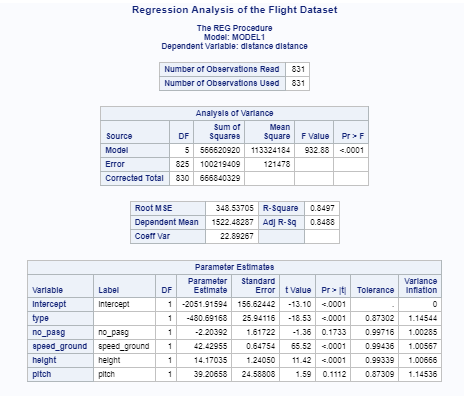
We know that, tolerance (requested by the tol option) is the proportion of variance in a given predictor that is NOT explained by all of the other predictors, while the VIF (or Variance Inflation Factor) is simply 1 / tolerance. The VIF represents a factor by which the variance of the estimated coefficient is multiplied due to the multicollinearity in the model

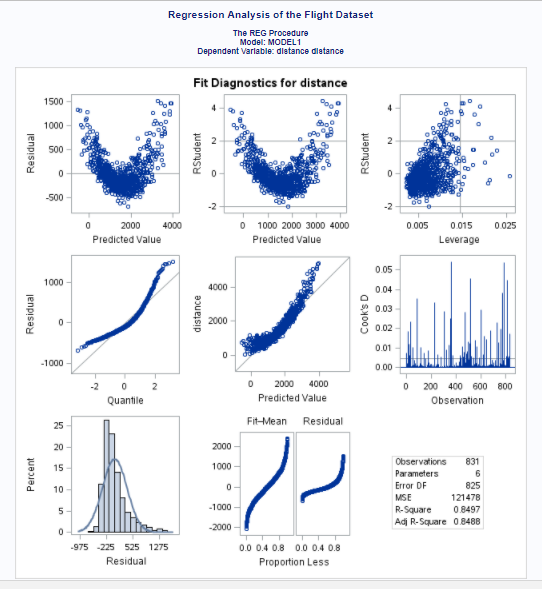
A good “global” check for a multicollinearity problem is to see if the largest condition index is greater than 30.

Here we find that the VIF is 42 for both the speed variables. So, we need to eliminate one of those variables to get a proper fit to the model. We also know that the speed\_air variable has 641 missing values. It is always better to fit a model with more data than less. So we can eliminate Speed\_air from our model and construct the linear regression equation.

MODEL COMPENSATED FOR MULTICOLLINEARITY

/\* Final Model \*/  
  
proc reg data = proj1sc.flight;  
model distance = type no\_pasg Speed\_ground height pitch /vif tol;  
title "Regression Analysis of the Flight Dataset";  
run;





JUSTIFICATION FOR VARIABLE SELECTIONS

As per the results we got in our process of understanding the variahles, we made two choices with respect to the variable:

1. Eliminated duration variable from the list as we found that the NULL Hypothesis for rho=0 couldn’t be rejected for the relationship between duration and distance
2. We eliminated speed\_air variable to compensate for the multicollinearity issue we had on the model due to the correlation between speed\_air and speed\_ground. We chose to eliminate spee\_air among the two variables because it has a lot of missing values and we would always want to use more data for developing the model.

QUESTIONS:

