**Statistical Computing**

**Midterm Project**

Nidhi Mavani – M10786158

**Abstract**

Landing distance plays a vital role for flights to avoid the mishaps; ICAO SARPs define the term 'Landing Distance' as "the horizontal distance traversed by the airplane by the airplane from a point on the approach path at a selected height above the landing surface to the point on the landing surface at which the airplane comes to a complete stop". In order to ensure safety for flights our goal for this project is the determine the factors that affect the landing distance. We have several observations of the flights with parameters measured are height, pitch, speed\_air, speed\_ground, aircraft type, number of passengers, and the duration of the flight. After cleaning and exploring the data a regression model was fitted. It was observed that landing distance is highly dependent on factors like aircraft type, height and speed\_air and a relationship between these variables is quantified.

**Chapter 1: Data Exploration and Cleaning**

1. **Goal of this chapter (Why are you doing this?)**

Getting, exploring and cleaning the data is the primary step to be under taken in any analysis.

Here since the data is already available for aircraft Boeing and Airbus, next step for us becomes to explore, know more about the data and clean it or manipulate in order to make it consumable for analysis

Steps that are followed in exploration

1. Understanding the variable and their limits
2. Merging the dataset from two excel files imported in SAS
3. Quantifying the missing values for each variable
4. Identifying the abnormality for each of the variable (Missing value is however retained)
5. Removing the Duplicate observation
6. Quantifying the missing values for each variable
7. Getting the Statistics like Mean, Median, Min and Maximum for all the variable
8. Plotting a Box Plot and Histogram for each of the variables
9. **The SAS code**

1 PROC IMPORT

2 DATAFILE = '/home/mavaninm0/sasuser.v94/FAA1.xls'

3 DBMS = xls

4 OUT =FlightDetailsB; /\* Flight details belonging to sheet 1\*/

5

6 PROC IMPORT

7 DATAFILE = '/home/mavaninm0/sasuser.v94/FAA2.xls'

8 DBMS = xls

9 OUT =FlightDetailsA; /\* Flight details belonging to sheet 2\*/

10

11 options missing=' ';

12 DATA FlightDetails;

13 Set FlightDetailsB FlightDetailsA;

14 IF missing(cats(of \_all\_)) THEN delete; /\* removing empty rows\*/

15 PROC print Data=FlightDetails;

16 RUN;

17

18 PROC MEANS DATA=FLightDetails NMISS N; /\* finding the number of missing values for each variable\*/

19 RUN;

20

21 PROC UNIVARIATE DATA=flightdetails;

22 RUN;

23

24 /\*validating the data for abnormality in values \*/

25 option missing ='.';

26 DATA FlightDetailsValidated;

27 SET FlightDetails;

28 IF (Duration ^=.)and Duration < 40 then validity = 'abnormal' ; Else validity = 'normal';

29 IF (validity = 'normal') and (Speed\_ground < 30 OR Speed\_ground > 140) then validity = 'abnormal' ;

30 IF (validity = 'normal') and (speed\_air ^=.) and (Speed\_air <30 OR Speed\_air > 140) then validity = 'abnormal' ;

31 IF (validity = 'normal') and Height< 6 then validity = 'abnormal' ;

32 IF (validity = 'normal') and Distance >= 6000 then validity = 'abnormal' ;

33 Proc print data=flightdetailsvalidated;

34 run;

35 proc freq data=flightdetailsvalidated;

36 tables validity;

37 run;

38

39 /\* check for duplication of observation \*/

40 PROC SORT DATA=FlightDetailsValidated

41 OUT = FlightDetailsRemovedDuplication

42 NODUPKEY ;

43 BY Speed\_ground Speed\_air height Distance;

44 run;

45

46 /\* check for duplication of observation \*/

47 PROC FREQ DATA=FlightDetailsRemovedDuplication;

48 TABLES VALIDITY;

49 RUN;

50

51 /\* finding the number of missing values for each variable\*/

52 PROC MEANS DATA=FlightDetailsRemovedDuplication NMISS N;

53 RUN;

54

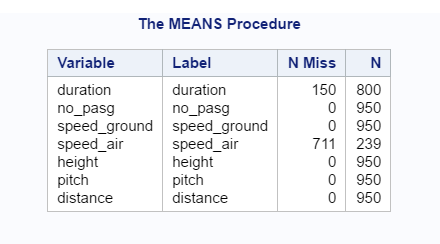
55 PROC MEANS DATA=FlightDetailsRemovedDuplication MEAN MEDIAN MIN MAX;

56

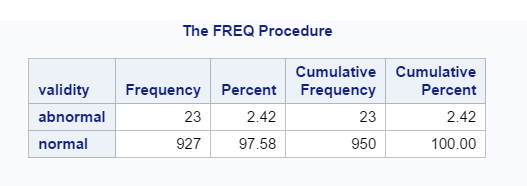
57 PROC UNIVARIATE DATA=FlightDetailsRemovedDuplication PLOT;

1. **The SAS output**

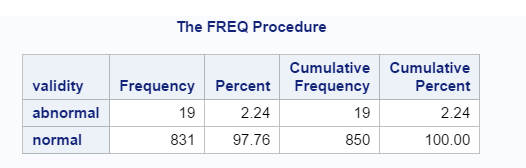
* Number of missing values variable wise



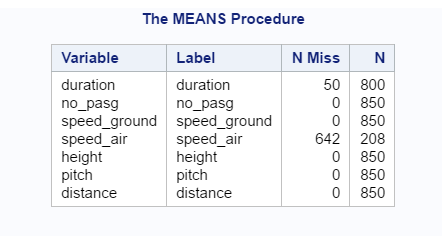
* Frequency of abnormal and normal observations before removing the duplicates



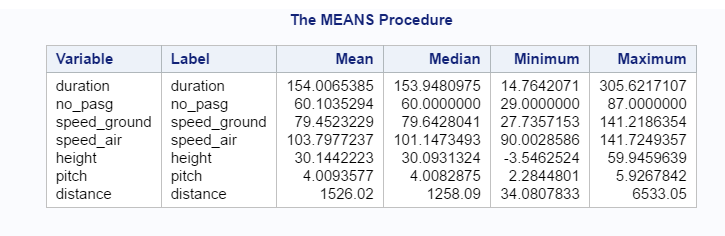
* Frequency of abnormal and normal observations after removing the duplicates



* Missing Values for each observation after removing duplication

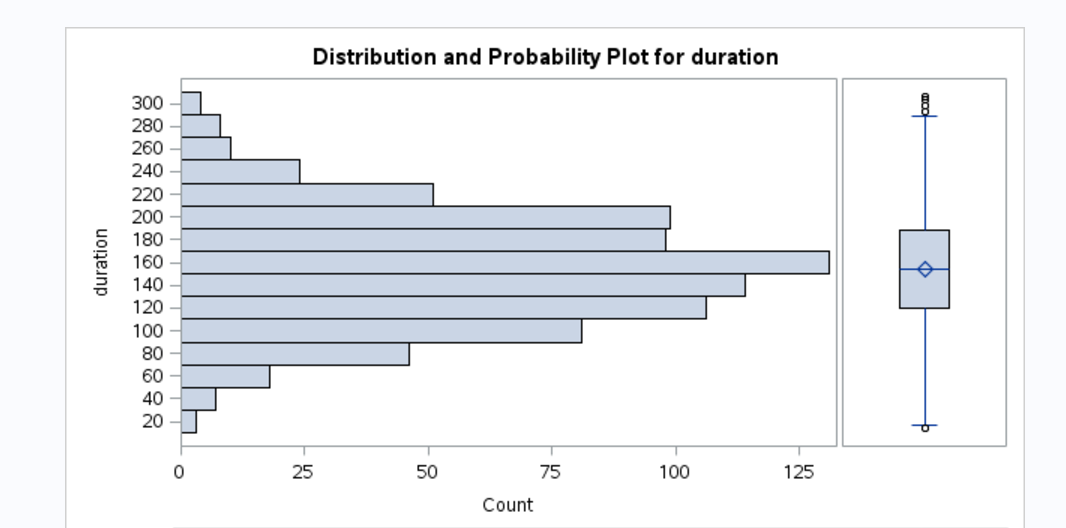


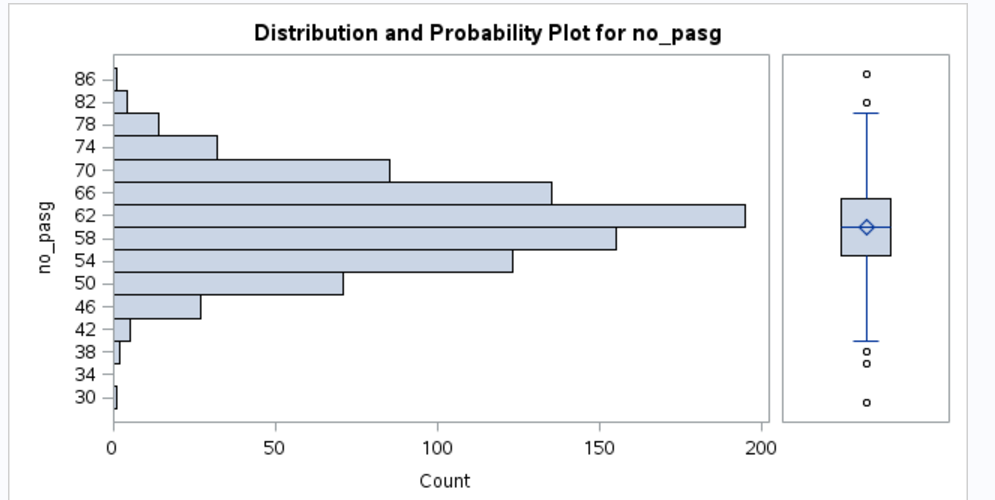
* Mean, Median, Minimum and Maximu

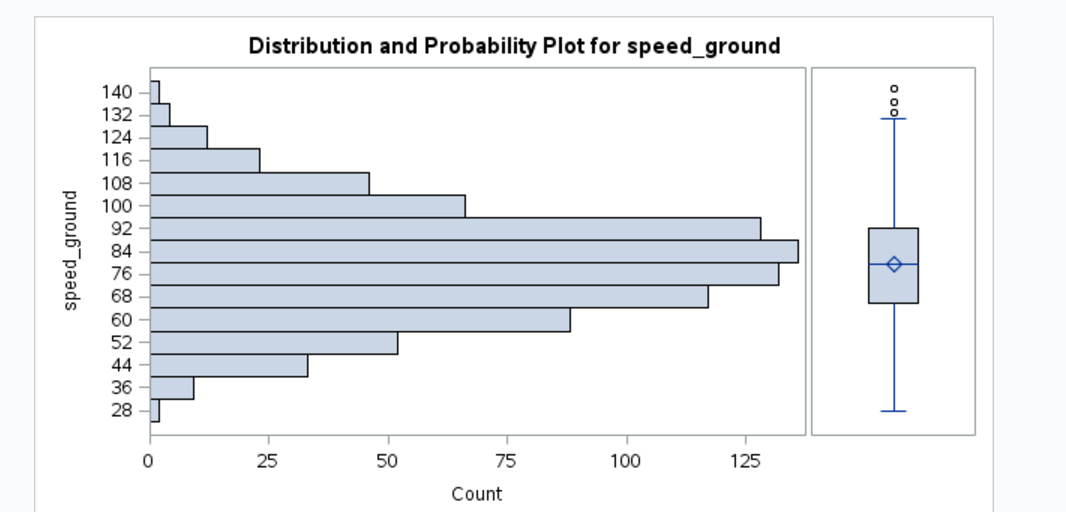


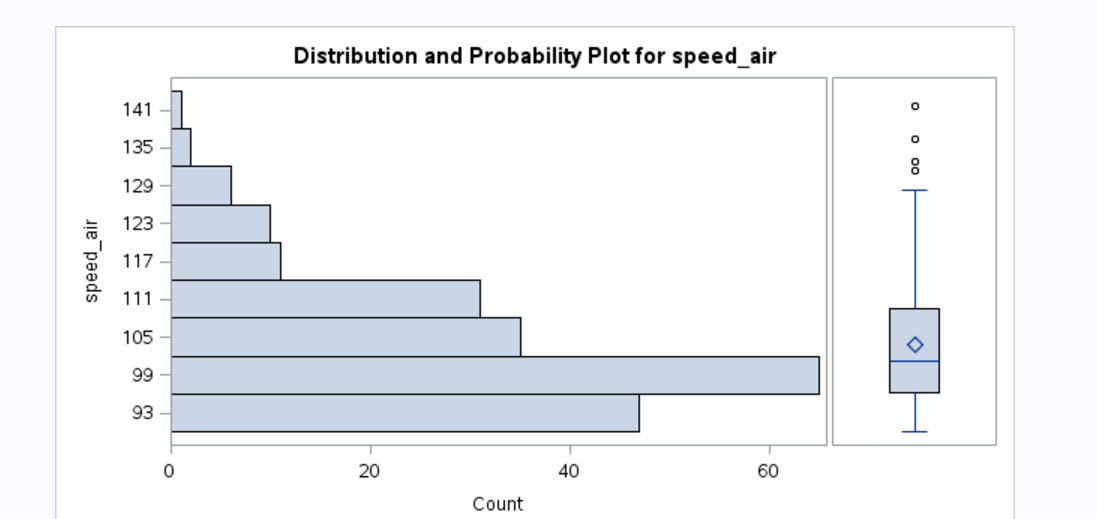
1. **PLOTS**

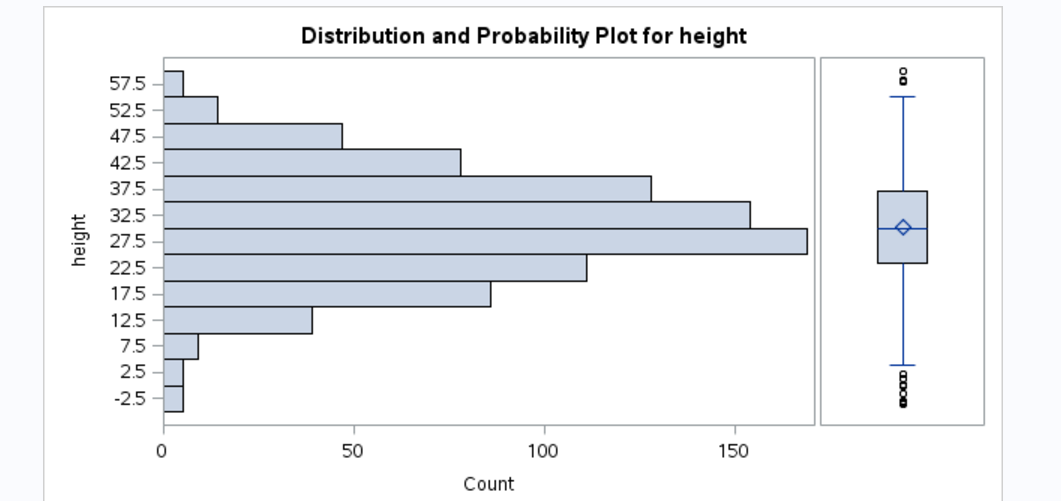
* **Normal and Box plot for Variables like Duration, Speed\_air, Speed\_ground, no\_pasg, Height, pitch**

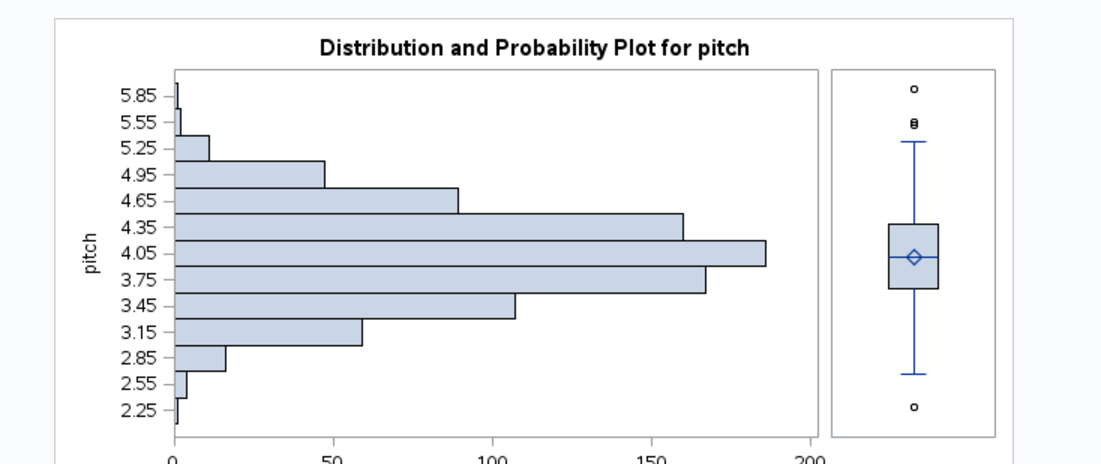












**Chapter 2: Continued Data Exploration**

1. **Goal of this chapter (Why are you doing this?)**

Now that we have the data cleaned with respect to duplications and we are aware of the abnormality, in this chapter we are going to explore the relationship between the dependent variable (landing distance) and independent (explanatory – rest all) variables.

Steps that are followed in exploration

1. Dropping the observations which have abnormality
2. In order to get aircraft variable accounted, adding a variable called make if aircraft=airbus then make is 1 else make=0
3. Using the scatter plot to understand the relationships between variables; We see that speed\_air, speed\_ground has some linear relationship with landing distance.
4. A box plot was plotted to see the differences in landing distance because of its make; there is indeed a difference, for airbus make the landing distance mean, median are lower than that for boeing
5. Next step is to validate the relationships found by plotting the variables; That is finding the correlation
6. From the correlation matrix it is observed that variables make, speed\_air, speed\_ground height have significant correlation with distance variable
7. Also speed\_air and speed\_height are highly correlated
8. **The SAS code**

1 /\* keeping only that data which is normal for further analysis \*/

2 DATA FlightDetailsNormal;

3 SET FlightDetailsRemovedDuplication (where=(validity='normal'));

4 drop validity;

5 PROC PRINT DATA=FlightDetailsNormal;

6 RUN;

7

8 /\* converting the non-numeric make column to binary \*/

9 DATA FlightDetailsNormal1;

10 set FlightDetailsNormal;

11 IF aircraft = "boeing" THEN make = 0;

12 else make = 1;

13 PROC PRINT DATA=FlightDetailsNormal1;

14 RUN;

15

16 /\* using scatter plot to get the relation between variables \*/

17 PROC SGSCATTER DATA=FlightDetailsNormal;

18 MATRIX distance make speed\_air speed\_ground height pitch duration no\_pasg;

19 RUN;

20 /\* Box plot for seeinng the difference in distance w.r.t to make\*/

21 PROC SORT DATA=flightdetailsnormal;

22 BY AIRCRAFT;

23 run;

24 PROC BOXPLOT DATA=FlightDetailsNormal;

25 PLOT DISTANCE\*AIRCRAFT/

26 nohlabel

27 boxstyle = schematic

28 boxwidthscale = 1

29 bwslegend;

30 run;

31

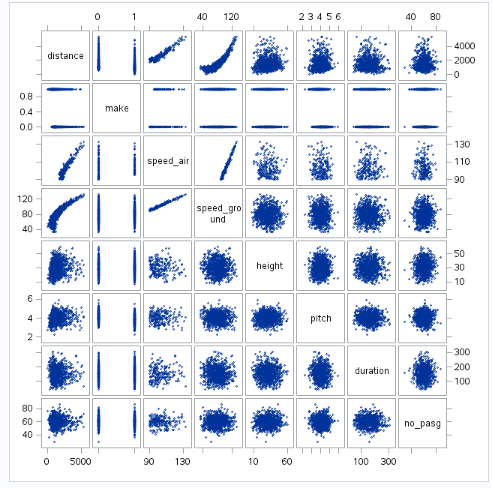
32 /\* finding the correlation between variables \*/

33 PROC CORR DATA=flightdetailsnormal;

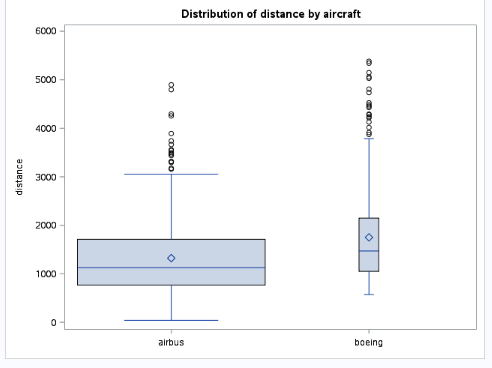
34 VAR distance make duration no\_pasg speed\_ground speed\_air height pitch;

35 RUN;

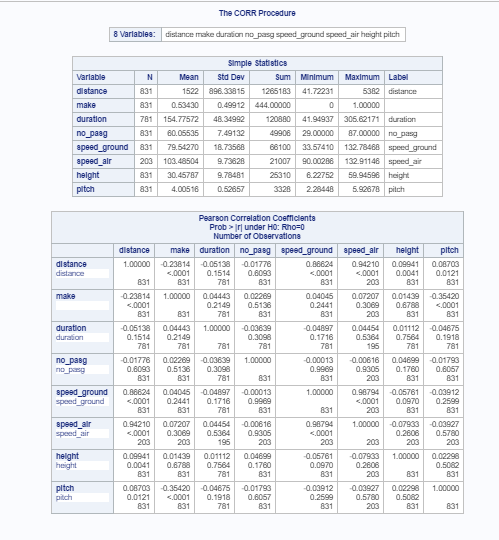
1. **The SAS output**
2. **Scatter plot showing the relationship between variables**



1. **Box plot for the distance with respect to make/aircraft**



**(c ) Correlation Matrix**



**Chapter 3: Modelling**

1. **Goal of this chapter (Why are you doing this?)**

We know which variables seems to have a relationship with landing distance and hence we create the regression model with variables make, speed\_air and height. With trial and error and basis tried model with many variables and the model with above stated variables seemed to be best.

Linear Regression was choosen as we could make out from the scatter plot that there is some kind of linearity between distance and other variables.

1. After seeing the correlation, a regression model was fitted with make, speed\_air, speed\_ground and height; However it turned out that speed\_ground was not statistically significant
2. Therefore dropping that variable; again a regression model was fitted with speed\_air, make and height; All of them turned to be statistically significant(p-value < 0.05)
3. The parameter estimates for each of the independent variable is generated and thus a linear equation can be made to show the relationship



1. **The SAS code**

1 /\* creating a regression analysis for the variables that seemed to be correrated with distance \*/

2

3

4 PROC REG DATA=flightdetailsnormal;

5 MODEL distance = make height speed\_air speed\_ground /r ;

6 output out=diagnostics r=residual;

7 run;

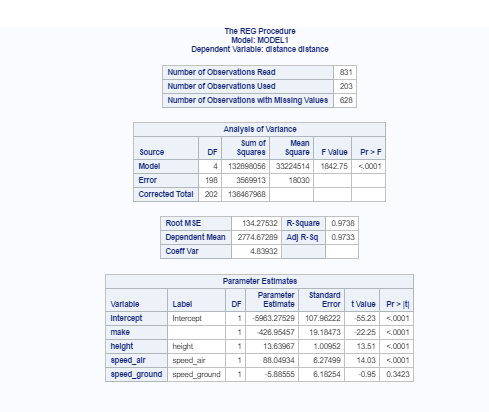
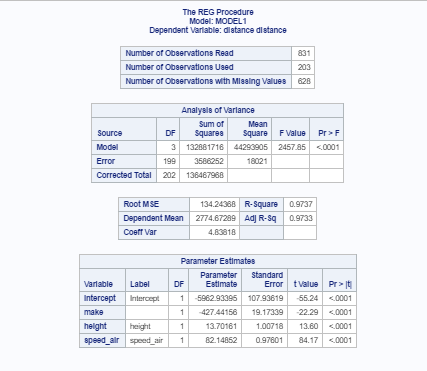
8

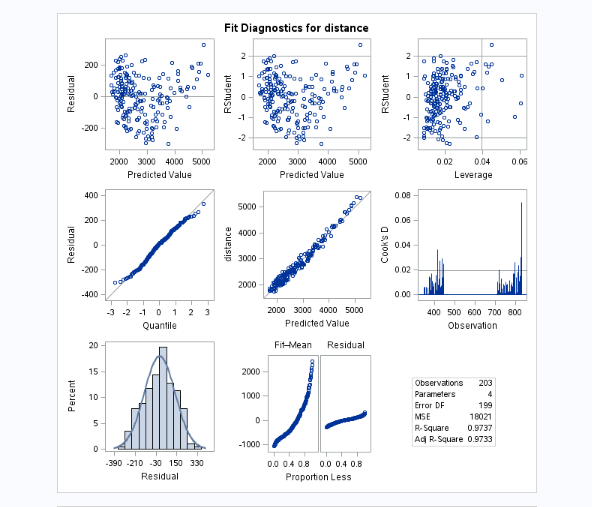
9 PROC REG DATA=flightdetailsnormal;

10 MODEL distance = make height speed\_air speed\_ground /r;

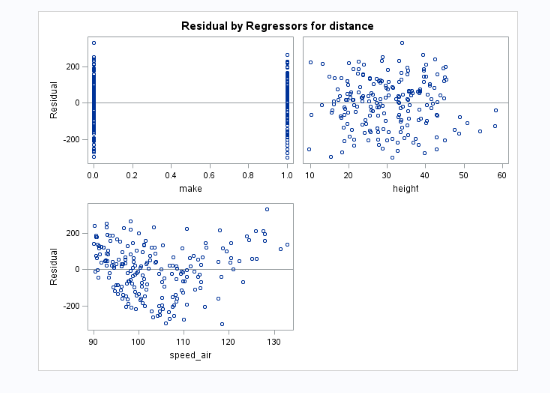
11 output out=diagnostics r=residual;

12 run;

1. **The SAS output**
2. Model with 4 variables 
3. Model with 3 variables
4. Residual Plot with the predicted value and histogram, qqplot for residual



1. Plot of residual vs make, height and speed\_air



**Chapter 4: Model Checking**

1. **Goal of this chapter (Why are you doing this?)**

After fitting the model we will check the model for its authenticity.

We need to check for the assumptions that we made.

1. Given that residuals when plotted against the predicted value, a sign of heteroskedasticity is observed and hence I tried to add the variable speed\_air\_sq in the equation to mitigate it, however with adding a polynomial term the heterskedasticity becomes more evident
2. Also tried using the log of the speed\_air for the model fitting and observed that homoscedasticity is more evident
3. Therefore came to conclusion that the model with speed\_air is better than rest; As residuals are also follow a normal distribution with mean around 0 (p-value 0.06; we fail to reject the null hypothesis)
4. Concluding that this is our final model



1. **SAS Code**

1 /\* creating a regression model uas an alternative to better fit the distance\*/

2

3 DATA FlightDetailsNormal;

4 set FlightDetailsNormal;

5 speed\_air\_sq = speed\_air\*\*2;

6 speed\_air\_log = log(speed\_air);

7 PROC PRINT DATA=FlightDetailsNormal;

8 RUN;

9

10 PROC REG DATA=flightdetailsnormal;

11 MODEL distance = make height speed\_air\_sq /r ;

12 output out=diagnostics r=residual;

13 run;

14

15 PROC REG DATA=flightdetailsnormal;

16 MODEL distance = make height speed\_air\_log /r ;

17 output out=diagnostics r=residual;

18 run;

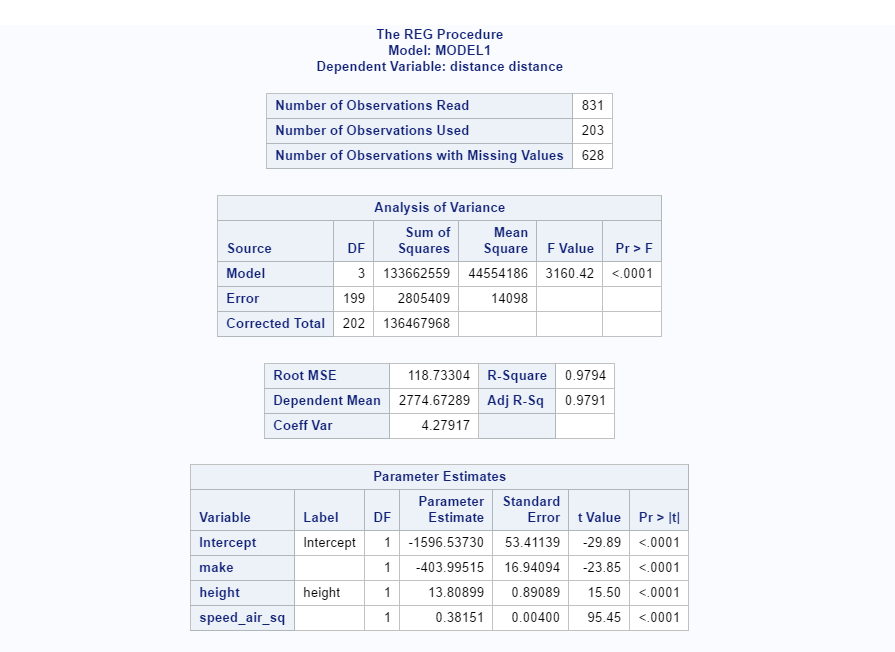
/\* checking the normality of the residuals \*/

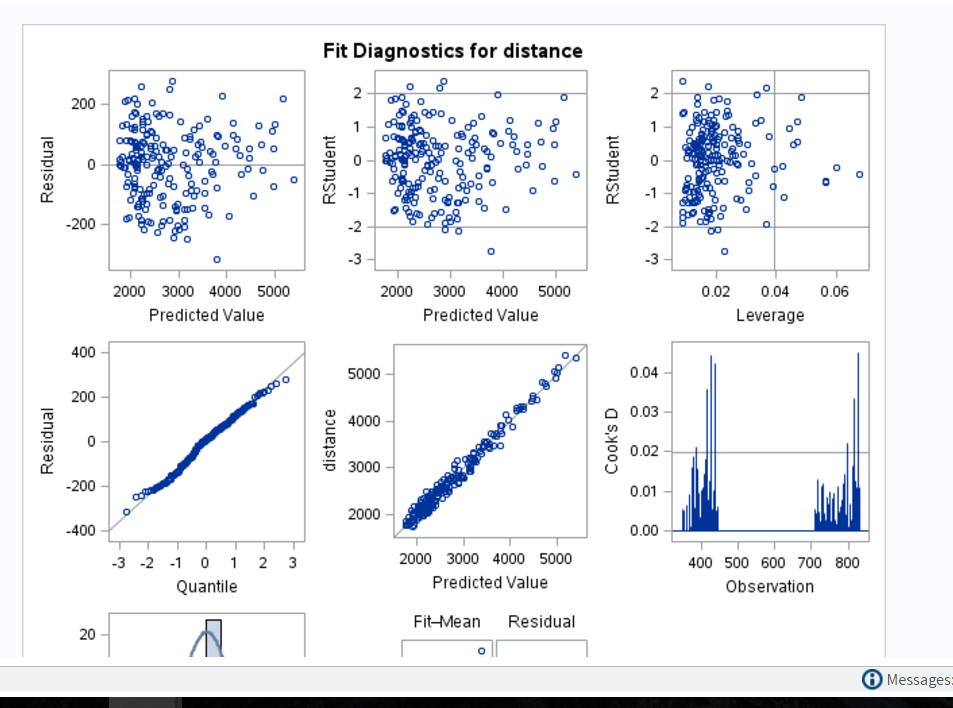
2 proc univariate data=diagnostics normal plot;

3 var residual;

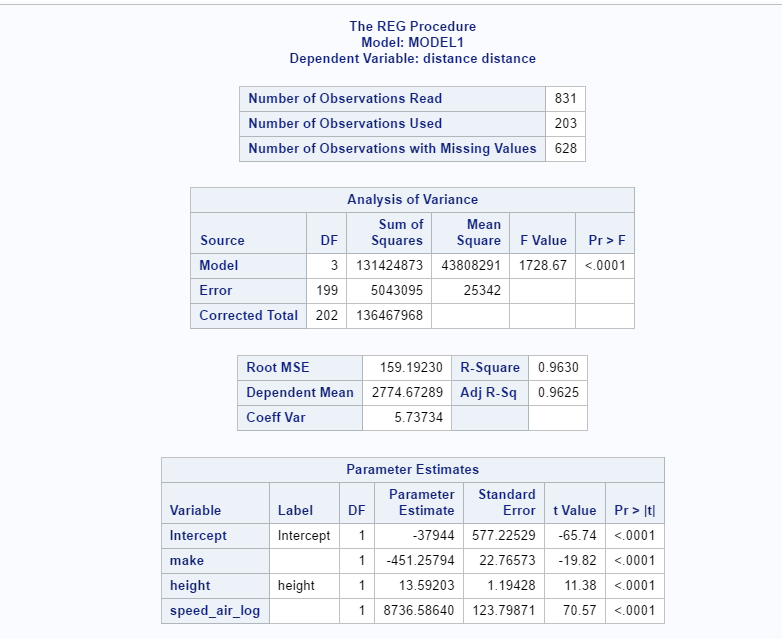
4 run;

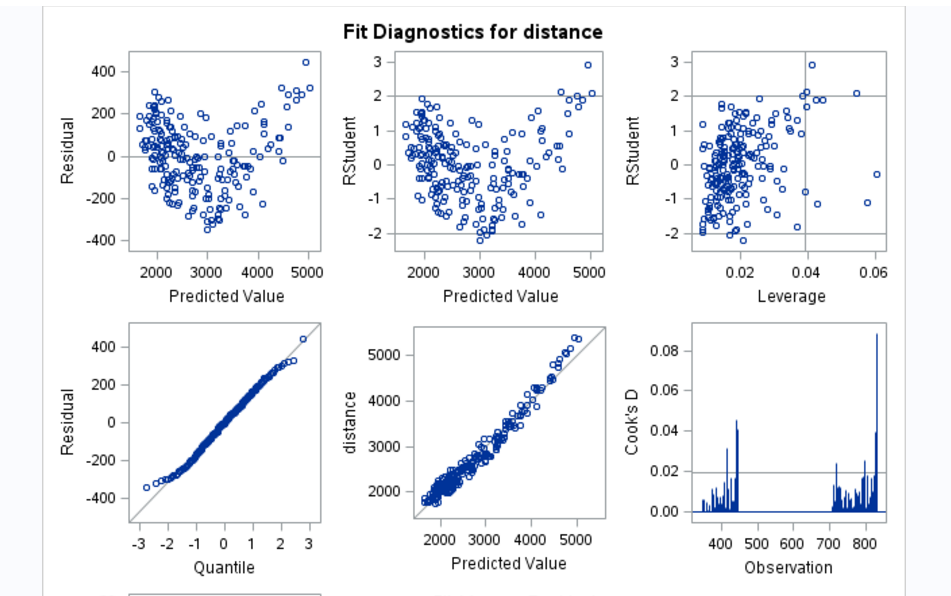
1. **SAS Output**
2. **Model with speed\_air\_sq as a predictor where homoscedasticity becomes more evident**





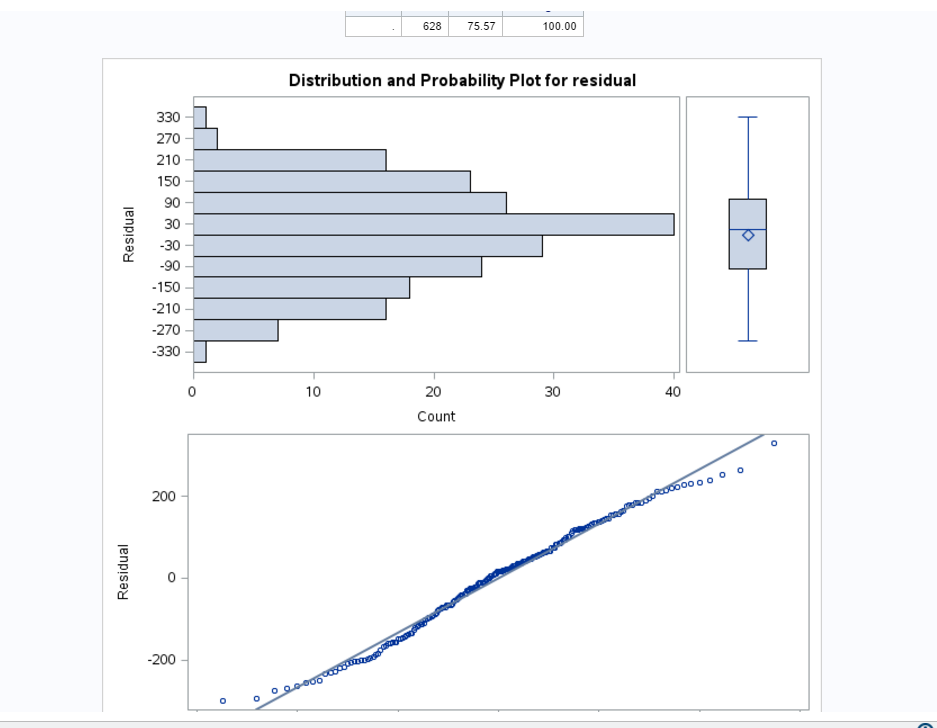
1. **Model with speed\_air\_log as a predictor where homoscedasticity becomes more evident**





1. **Normality check for the residual with model having speed\_air make and height**





**Questions to be answered**

1. **How many observations (flights) do you use to fit your final model? If not all 950 flights, why?**

Final model used only 203 observation out of 950 as speed\_air had many missing values but had a significant impact so could not remove from the model as well.

1. **What factors and how they impact the landing distance of a flight?**

Factors like Aircraft company, height and speed\_air affects the landing distance



As seen from this equation:

With **make** if the value is 1 i.e if the make is airbus the landing distance will decrease by -427.44 keeping the rest values constant

Similarly with variable **height** with every unit increase in height the landing distance will increase by 13.702

Similarly with variable **speed\_air** with every unit increase in it, the landing distance will increase by 82.149

1. **Is there any difference between the two makes Boeing and Airbus?**

Yes there is.

Usually with Airbus it was observed that landing distance is lesser than that with Boeing. We can see that with this plot in chapter 2.3.b