Table of Contents

[Chapter 1: Data Cleaning and Exploration 2](#_Toc463052107)

[Goal 2](#_Toc463052108)

[SAS Code 2](#_Toc463052109)

[SAS Output 4](#_Toc463052110)

[Observations 5](#_Toc463052111)

[Conclusion 5](#_Toc463052112)

[Questions 5](#_Toc463052113)

[Chapter 2: Data Visualization 6](#_Toc463052114)

[Goal 6](#_Toc463052115)

[SAS Code 6](#_Toc463052116)

[SAS Output 8](#_Toc463052117)

[Observations 13](#_Toc463052118)

[Conclusions 13](#_Toc463052119)

[Chapter 3: Data Exploration 14](#_Toc463052120)

[Goal 14](#_Toc463052121)

[SAS Code 14](#_Toc463052122)

[SAS Output 14](#_Toc463052123)

[Observations 18](#_Toc463052124)

[Conclusions 18](#_Toc463052125)

[Chapter 4: Modeling 19](#_Toc463052126)

[Goal 19](#_Toc463052127)

[SAS Code 19](#_Toc463052128)

[SAS Output 19](#_Toc463052129)

[Observations 21](#_Toc463052130)

[Conclusion 21](#_Toc463052131)

[Chapter 5: Model Checking 21](#_Toc463052132)

[Goal 21](#_Toc463052133)

[SAS Code 21](#_Toc463052134)

[SAS Output 22](#_Toc463052135)

[Observations 24](#_Toc463052136)

[Conclusion 24](#_Toc463052137)

# Chapter 1: Data Cleaning and Exploration

## Goal

To detect and remove abnormal values in the data.

## SAS Code

/\*Chap 1 Data Exploration and Data Cleaning\*/  
  
**PROC IMPORT** datafile='/folders/myfolders/Stat\_Computing/FAA1.xls' dbms=XLS out=FAA1 replace;  
getnames=yes;  
title 'FAA1.xls';  
**run;**  
**PROC IMPORT** datafile='/folders/myfolders/Stat\_Computing/FAA2.xls' dbms=XLS out=FAA2 replace;  
getnames=yes;  
title 'FAA2.xls';  
**run;**

/\*Merging the two files\*/  
**DATA** FAA\_MERGED;  
SET FAA1 FAA2;  
**run;**  
  
/\*To get the variables names and data type present in the dataset\*/  
**proc contents** data=FAA\_MERGED; title 'Proc Contents';  
**run ;**

**DATA FAA\_AbnrmlRM;**  
set FAA\_MERGED;  
array cvars(\*) distance duration height no\_pasg pitch speed\_air speed\_ground;  
 /\*Delete observation if all variables are blank\*/  
 if ((cmiss(of cvars(\*))+cmiss(aircraft)) eq (dim(cvars)+1))then delete;  
  
 /\*Delete row if Duration <40 mins\*/   
 if(not missing(duration) and Duration <=40) then delete;  
   
 /\*Delete row if Speed\_ground<30 mph or >140 mph\*/  
 if(not missing(speed\_ground) and (speed\_ground<30 or speed\_ground>140)) then delete;  
   
 /\*Delete row if speed\_air<30 mph or >140 mph\*/  
 if(not missing(speed\_air) and (speed\_air <30 or speed\_air>140)) then delete;  
   
 /\*Delete row if height<6 m\*/  
 if(not missing(height) and height <6) then delete;  
   
 /\*Delete row if distance>6000ft\*/  
 if(not missing(distance) and distance>=6000) then delete;  
**run;**

/\*Remove Duplicates\*/

**PROC SORT** data=faa\_abnrmlrm out=FAA nodupkey dupout=Duplicate;  
 by \_all\_;**run;**

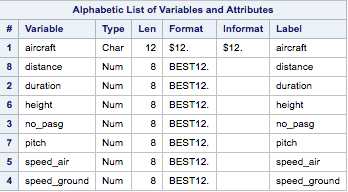
**proc print** data=FAA; title 'FAA Data Cleaned Up';**run;**

**proc print** data=Duplicate; title ‘Duplicate Data';**run;**

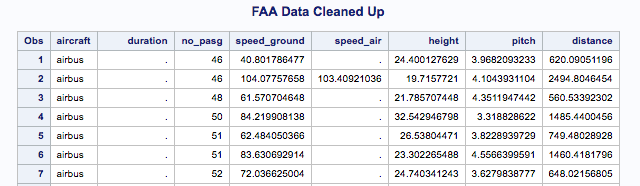
**proc contents** data=FAA; title 'After Clean Up';**run;**

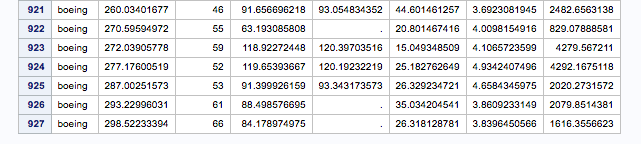
## SAS Output



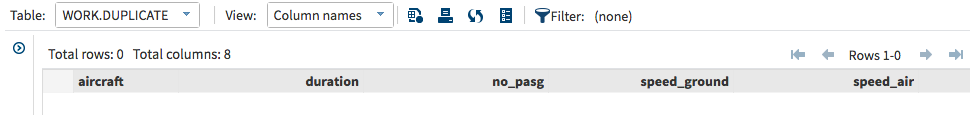


Sample Cleaned Up data:





No Duplicate data found:





## Observations

* **Presence of Blank rows:** After the two excel files were merged there were 50 blank rows with no data in any of the columns
* **Duration column of second excel file:** The Duration column of the second excel file is blank for all rows
* **Presence of Abnormal values:** There were abnormal values in the data as described in the project document
* **Duplicate data:** When the data was sorted and deduped by all variables no duplicate values were found

## Conclusion

* **Final number of observations:** After the clean up process we are left with 927 observations

## Questions

* Should we delete observations with blank variables?
* Why is there no unique key associated with the data set?
* Should we cap or delete abnormal values?

# Chapter 2: Data Visualization

## Goal

To observe the data trends

## SAS Code

/\*distance duration height no\_pasg pitch speed\_air speed\_ground\*/  
ods graphics on / width=4in;  
ods graphics on / height=4in;  
**proc means** data=faa mean median min max; var distance;**run**;  
**proc sgplot** data=faa;  
 histogram distance/scale=count;   
 title Histogram of Distance;  
**run;**  
  
ods graphics on / width=4in;  
ods graphics on / height=4in;  
proc means data=faa mean median min max; var duration;run;  
**proc sgplot** data=faa;  
 histogram duration/scale=count;   
 title Histogram of Duration;  
**run;**  
  
ods graphics on / width=4in;  
ods graphics on / height=4in;  
**proc means** data=faa mean median min max; var height;run;  
**proc sgplot** data=faa;  
 histogram height/scale=count;   
 title Histogram of Height;  
**run;**  
  
ods graphics on / width=4in;  
ods graphics on / height=4in;  
**proc means** data=faa mean median min max; var no\_pasg;run;  
**proc sgplot** data=faa;  
 histogram no\_pasg/scale=count;   
 title Histogram of Number of Passengers;  
**run;**  
  
ods graphics on / width=4in;  
ods graphics on / height=4in;  
**proc means** data=faa mean median min max; var speed\_air;**run**;  
**proc sgplot** data=faa;  
 histogram speed\_air/scale=count;   
 title Histogram of speed\_air;  
**run;**  
  
ods graphics on / width=4in;  
ods graphics on / height=4in;  
**proc means** data=faa mean median min max; var speed\_ground;**run;**  
**proc sgplot** data=faa;  
 histogram speed\_ground/scale=count;   
 title Histogram of speed\_ground;  
**run;**

ods graphics on / width=4in;  
ods graphics on / height=4in;  
**proc means** data=faa mean median min max; var pitch;**run;**  
**proc sgplot** data=faa;  
 histogram pitch /scale=count;   
 title Histogram of Pitch;  
**run;**

**proc chart** data=faa;

vbar distance/subgroup=aircraft;

**run;**

/\*Creating subsets based on aircraft make\*/

**data** Boeing;set faa; if aircraft='boeing';**run;**

**data** Airbus;set faa; if aircraft='airbus';**run;**

/\*Checking for difference in metrics\*/

**proc means** data=boeing mean median min max; vars distance no\_pasg speed\_ground speed\_air height pitch;Title Boeing Metrics;**run;**

**proc means** data=airbus mean median min max; vars distance no\_pasg speed\_ground speed\_air height pitch;Title Airbus Metrics;**run;**

**proc chart** data=boeing;

vbar distance/midpoints=200 to 5400 by 400 axis=0 to 150 by 20;

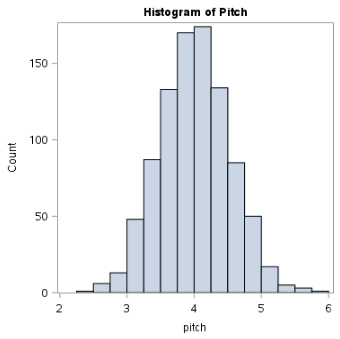
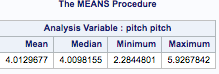
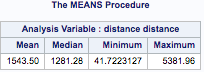
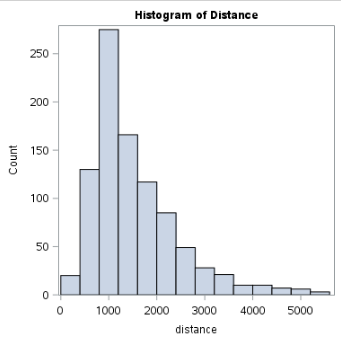
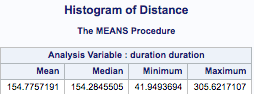
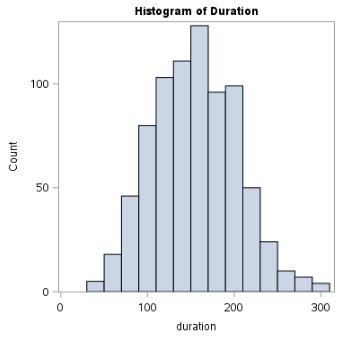
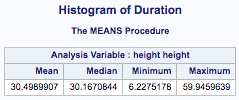
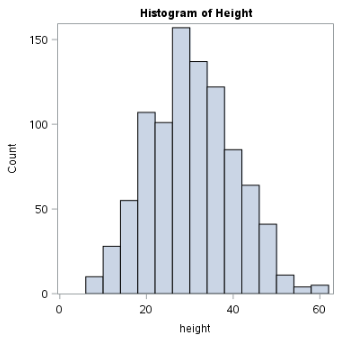
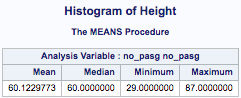
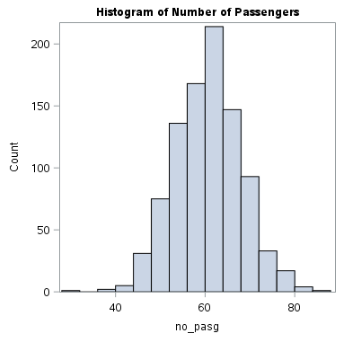
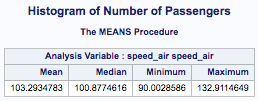
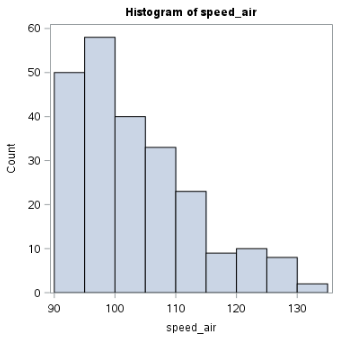
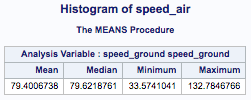
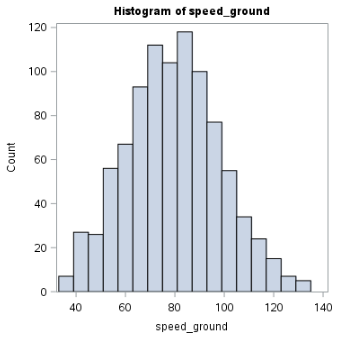
Title Boeing Distance Distribution;**run;**

**proc chart** data=airbus;

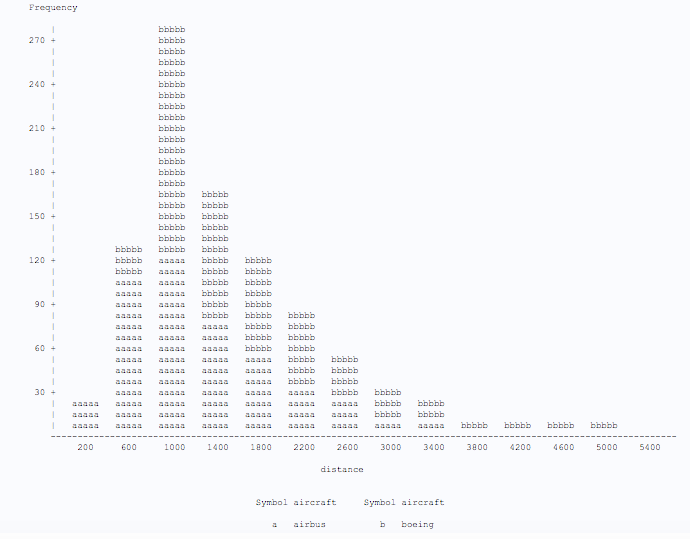
vbar distance/midpoints=200 to 5400 by 400 axis=0 to 150 by 20;

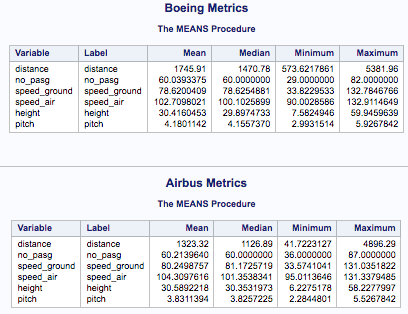
Title Airbus Distance Distribution;**run;**

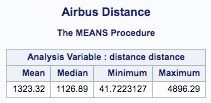
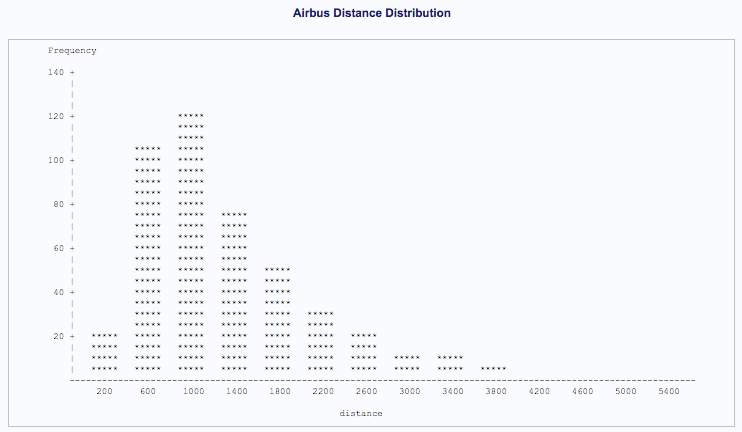
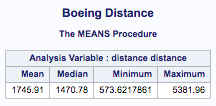
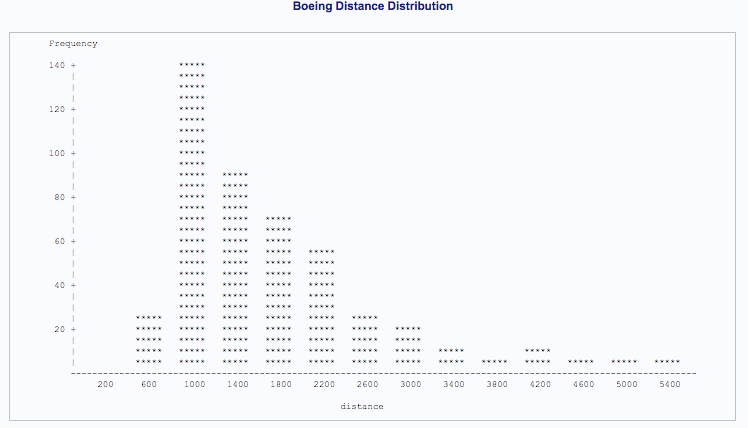
## SAS Output



Boeing Vs. Airbus Distance distribution







## Observations

* The variables Speed air and Distance are right skewed
* The variables Speed ground, Height, Number of Passengers and Duration are normally distributed
* The mean values of Speed air, Speed ground, Height, Number of Passengers and Pitch are similar for both Boeing and Airbus aircrafts
* The mean Distance for landing is different for the two types of aircrafts

## Conclusions

* Since the aircraft capacity and speed are similar for both the aircrafts they are a similar model

# Chapter 3: Data Exploration

## Goal

To look for relationships between variables and form hypothesis

## SAS Code

**proc ttest** data=faa;  
 class aircraft;  
 var distance;  
 Title TTest for diffference in Distance by make;  
**run;**

/\*Checking for relations\*/

**proc plot** data=faa;

plot duration\*distance;

plot no\_pasg\*distance;

plot speed\_ground\*distance;

plot speed\_air\*distance;

plot height\*distance;

plot pitch\*distance;

**run;**

/\*Calculating Correlation coefficients\*/

**proc corr** data=faa;

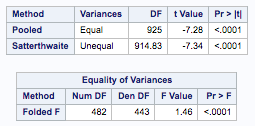
var distance speed\_air speed\_ground duration no\_pasg height pitch;

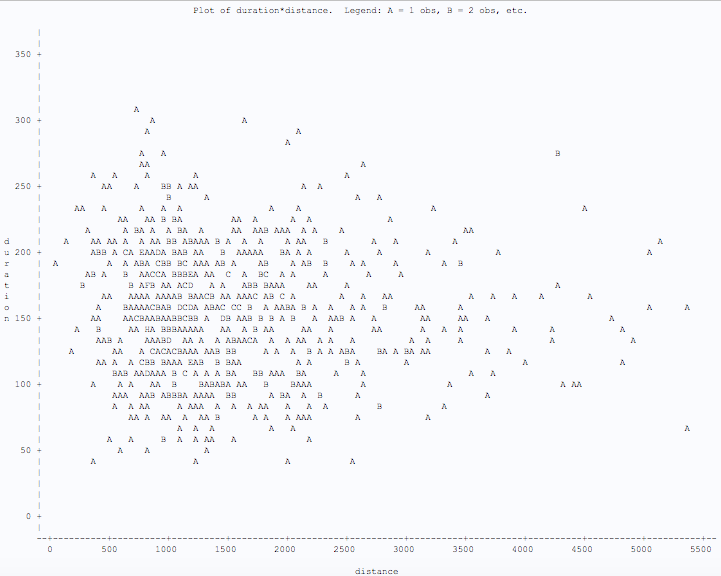
Title Correlation Coefficients;

**run;**

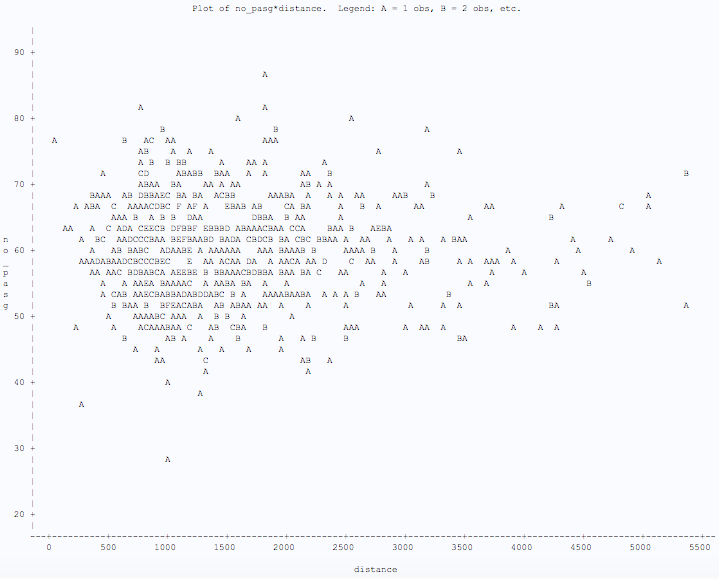
## SAS Output

From the test result we can reject the null hypothesis that the mean distance for Boeing and Airbus aircrafts are equal.

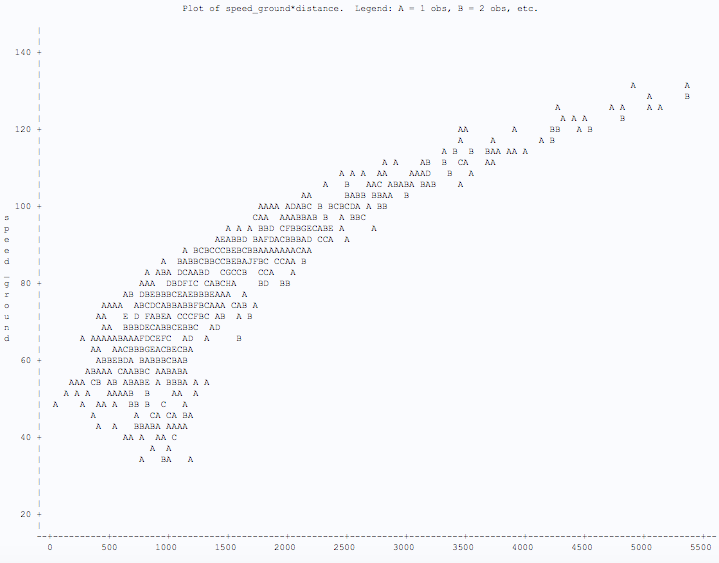




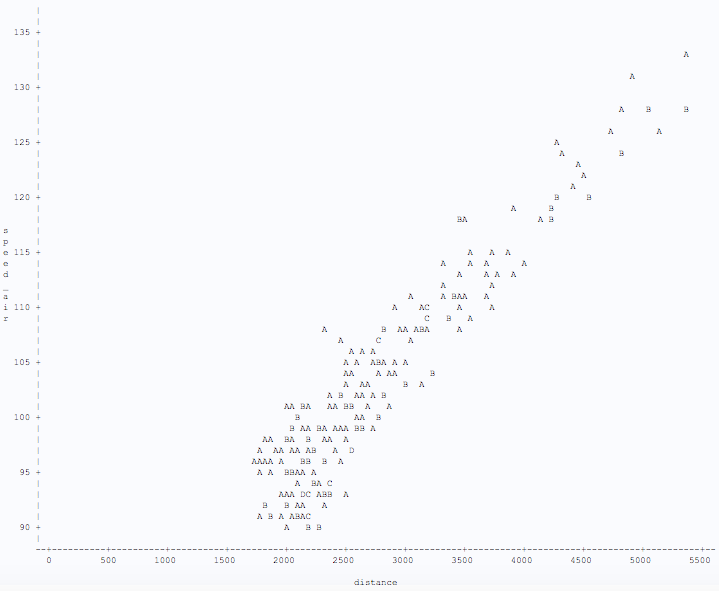
Duration Vs. distance



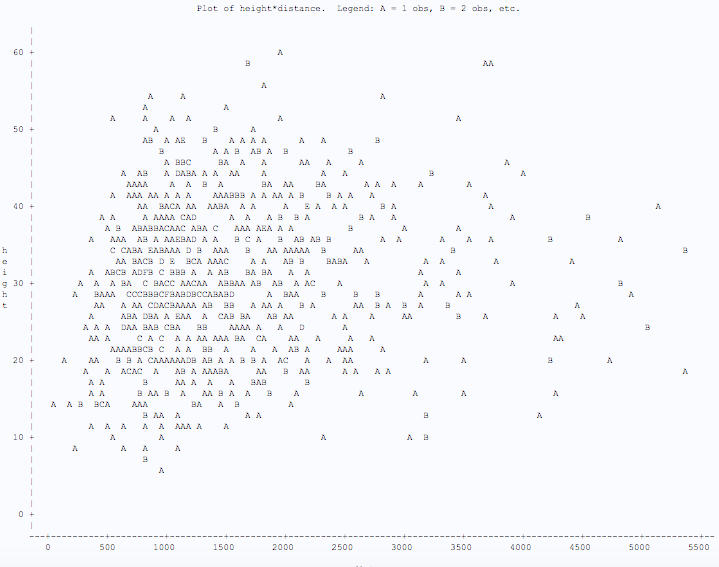
Number of Passengers Vs. Distance



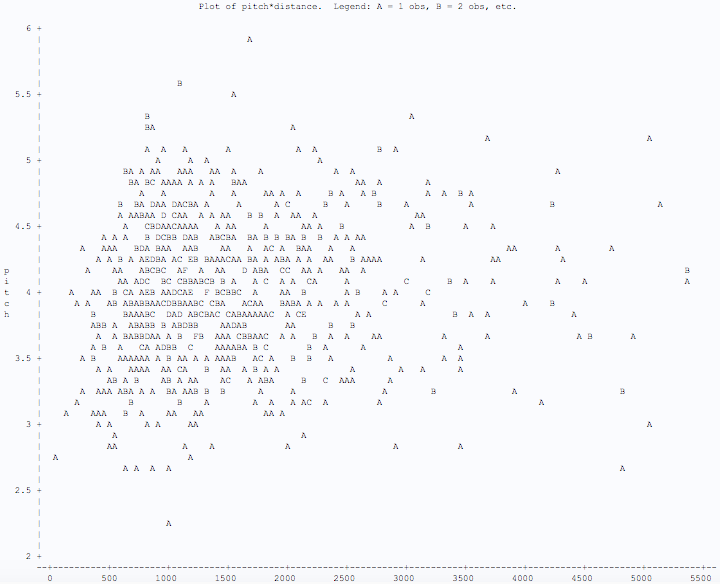
Speed Ground Vs. Distance



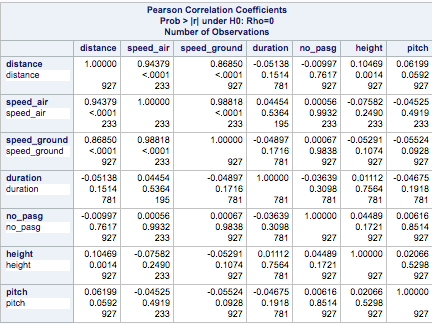
Speed Air Vs. Distance



Height Vs. Distance



Pitch Vs. Distance



## Observations

* Speed in Air and Speed on Ground show a discernable pattern with Distance
* Other variables don’t show any obvious pattern on first glance
* Speed in Air and Speed on Ground show high positive correlation with Distance

## Conclusions

* Distance is probably influenced by the speed in air and speed on ground of the aircraft

# Chapter 4: Modeling

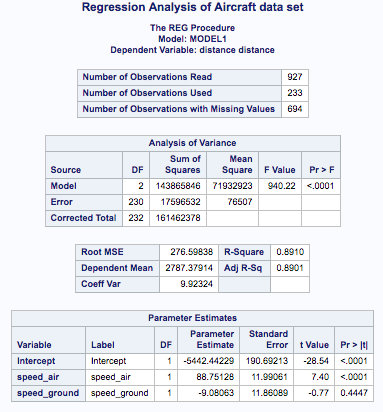
## Goal

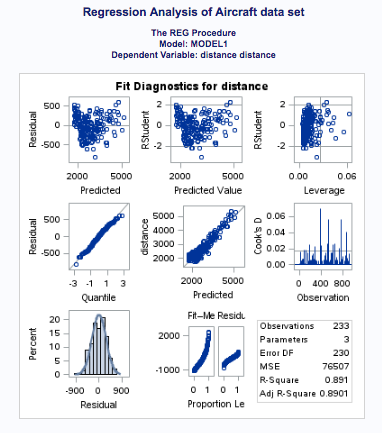
Perform a regression analysis to create a simple linear regression model which predicts the Distance variable using Speed in Air and Speed on Ground variables.

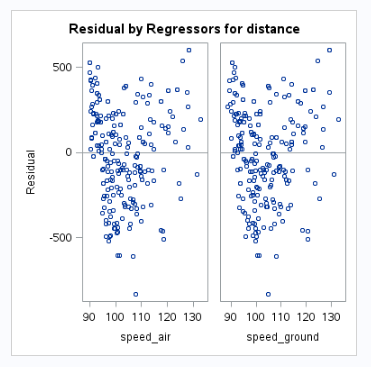
## SAS Code

/\*Performing regression analysis\*/  
**proc reg** data=faa;  
 model distance= speed\_air speed\_ground;   
 title Regression Analysis of Aircraft data set;  
**run;**

## SAS Output







## Observations

* We obtain a model: -5442.44 + 88.75\*Speed\_air -9.08\*Speed\_ground=Distance

## Conclusion

* The model must be validated before being finalized

# Chapter 5: Model Checking

## Goal

To check if the assumptions made on the noise present in the data is valid or not

## SAS Code

**proc reg** data=faa;  
 model distance= speed\_air speed\_ground;

output out=diagnostics r=residual;  
 title Regression Analysis of Aircraft data set;  
**run;**

/\*Checking the distribution of residuals\*/  
**proc plot** data=diagnostics;  
 plot residual\*speed\_air;  
 plot residual\*speed\_ground;  
**run;**  
  
/\*Checking if residuals are normally distributed\*/  
/\*Checking if mean = 0\*/  
**proc univariate** data=diagnostics normal plot;  
 var residual;  
**run;**

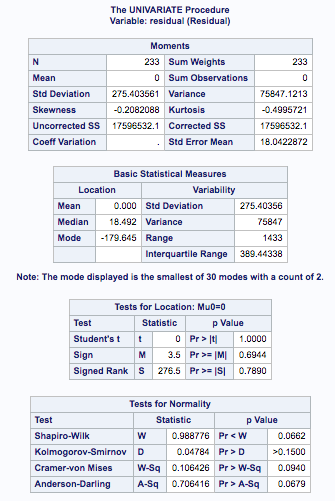
## SAS Output

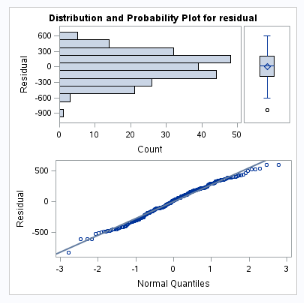


Residual Vs. Speed in Air



Residual Vs. Speed on Ground





## Observations

* The residual values are even spread about 0
* The residual values appear to be normally distributed
* The mean value of residuals is 0, this is verified by hypothesis testing

## Conclusion

* The model satisfies all validity checks and hence can be adopted