Statistical Computing : Project Submission

Aniket Sunil Mahapure (M12910618)

Overall Summary of the project

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

Key findings:

- Speed_Air and Speed_Ground are highly corelated and Speed_Air has around 75% missing values
- Landing Distance varies for the two Aircrafts – Boeing & Airbus. The difference between the mean of the two is ~427 meters
- As Landing Distance is different for the two Aircrafts, two models have been built to identify the right factors and the magnitude of their impact
- Speed air and height are highly correlated to distance as compared to other variables.
- The final model equation for Airbus Landing Distance is

  Distance = -2522.89061 + 42.55420 * speed_ground + 14.09773 * height

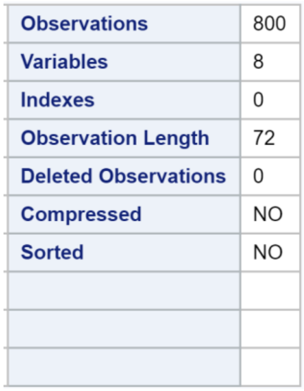
- The final model equation for Boeing Landing Distance is

  Distance = -2008.46764 + 42.28538 * speed_ground + 14.19682 * height

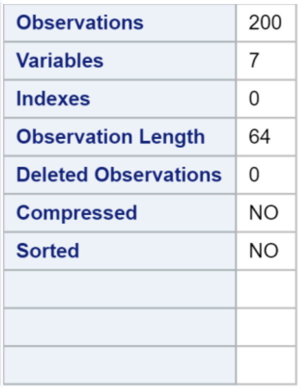
Chapter1: Data Preparation

Importing the data sets

```
/* step 1*/
/* importing data set FAA1 */
FILENAME REFFILE '/folders/myfolders/GASUE34_data/FAA1.xls';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLS
	OUT=WORK.FAA1;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.FAA1; RUN;

/* importing data set FAA2 */
FILENAME REFFILE '/folders/myfolders/GASUE34_data/FAA2.xls';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLS
	OUT=WORK.FAA2;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.FAA2; RUN;
```



Observations	800
Variables	8
Indexes	0
Observation Length	72
Deleted Observations	0
Compressed	NO
Sorted	NO

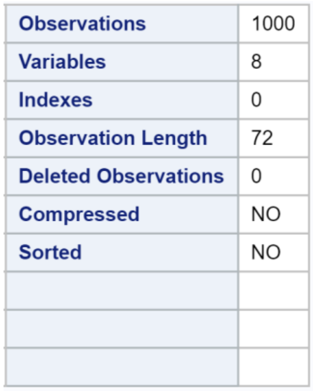


Observations	200
Variables	7
Indexes	0
Observation Length	64
Deleted Observations	0
Compressed	NO
Sorted	NO

FAA1 dataset contains 800 observations and FAA2 contains 200 observations.

Combining 2 data sets

```
/* step 2*/
/* combining two data sets*/
data combined;
       set work.faa1 work.faa2;
proc contents data=combined;
run;
```

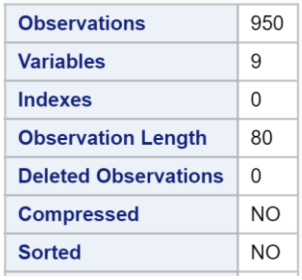


Observations	1000
Variables	8
Indexes	0
Observation Length	72
Deleted Observations	0
Compressed	NO
Sorted	NO

Total 200 observations after combining 2 data sets

Deleting empty rows and duplicates

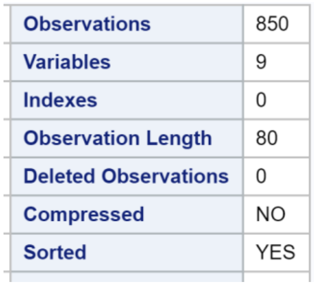
```
/* step 3 */
/*deleting empty rows*/
/*creating a flag for aircraft column to count missing values later*/
data combined_new;
set combined;
if compress(cats(of _all_),'.')=' ' then delete;
if aircraft=' ' then aircraft_1=1;
else aircraft_1=0;
run;
proc contents data=combined_new; run;
```



Observations	950
Variables	9
Indexes	0
Observation Length	80
Deleted Observations	0
Compressed	NO
Sorted	NO

There were 50 empty rows in FAA2 data set which were removed in this step.
Note: In combined data set, no. of variables are 9 because I have created an extra column for aircraft to calculate its missing values by treating it as a numeric.

```
/* removing duplicate entries*/
PROC SORT DATA=combined_new nodupkey OUT=combined_valid_nodup;
 by aircraft distance height no_pasg pitch speed_air speed_ground;
RUN;

proc contents data=combined_valid_nodup;
```



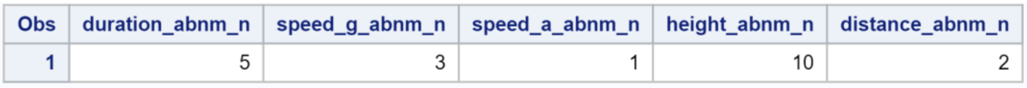
Observations	850
Variables	9
Indexes	0
Observation Length	80
Deleted Observations	0
Compressed	NO
Sorted	YES

100 duplicate records were removed

Treating invalid data rows

```
/*step 4*/
/* creating validity flags for variables */
data combined_flags;
set combined_valid_nodup;
/*keep duration_abnm speed_g_abnm   speed_a_abnm height_abnm distance_abnm;*/
if duration<40 and duration ^=. then duration_abnm=1;
else duration_abnm=0;
if (speed_ground<30 or speed_ground>140) and speed_ground ^=.  then speed_g_abnm=1;
else speed_g_abnm=0;
if (speed_air<30 or speed_air>140) and speed_air ^=. then speed_a_abnm=1;
else speed_a_abnm=0;
if height<6 and height ^=. then height_abnm=1;
else height_abnm=0;
if distance>=6000 and distance ^=. then distance_abnm=1;
else distance_abnm=0;
run;
proc print data=combined_flags;
run;

/* counting invalid entries for every variable*/
proc sql;
```

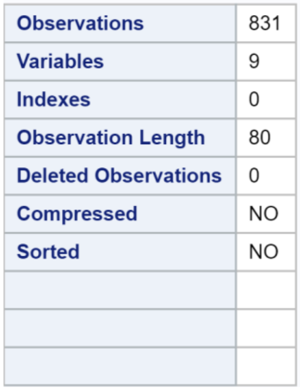
```
create table totals as select
sum(duration_abnm) as duration_abnm_n,
sum(speed_g_abnm) as speed_g_abnm_n,
sum(speed_a_abnm) as speed_a_abnm_n,
sum(height_abnm) as height_abnm_n,
sum(distance_abnm) as distance_abnm_n from combined_flags;
quit;
proc print data=totals;
run;
```



Obs	duration_abnm_n	speed_g_abnm_n	speed_a_abnm_n	height_abnm_n	distance_abnm_n
1	5	3	1	10	2

Variable height has the highest no. of invalid entries while other variables has less no. of invalid entries as compared to height. There are total 19 invalid observations in data.

```
/* removing invalid entries*/
data combined_valid;
keep aircraft aircraft_1 distance duration height no_pasg pitch speed_air speed_ground;
set combined_flags;
if duration_abnm=1 or speed_g_abnm=1 or speed_a_abnm=1 or height_abnm=1 or  distance_abnm=1
then delete;
run;
proc contents data=combined_valid;
run;

proc contents data=combined_valid;
run;
```



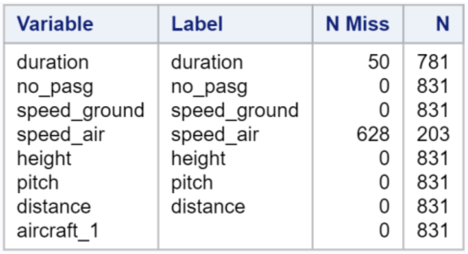
Observations	831
Variables	9
Indexes	0
Observation Length	80
Deleted Observations	0
Compressed	NO
Sorted	NO

Total 19 invalid entries were removed from the dataset; resulting final count to 831 observations.

Treating Missing values

Calculating Missing values

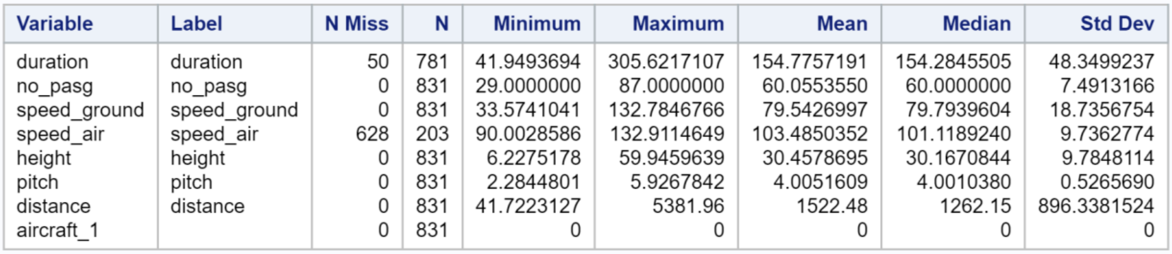
```
/* calculating missing values for all variables */
proc means data=combined_valid NMISS N;
run;
```



Variable	Label	N Miss	N
duration	duration	50	781
no_pasg	no_pasg	0	831
speed_ground	speed_ground	0	831
speed_air	speed_air	628	203
height	height	0	831
pitch	pitch	0	831
distance	distance	0	831
aircraft_1		0	831

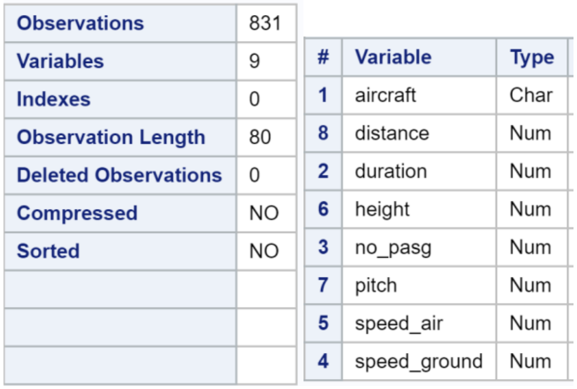
Almost 75% values for speed_air are missing and 6% values for duration are missing.

```
/* analyzing basic details of variables before treating missing values*/
proc means data = combined_valid nmiss N min max mean median std;
var _numeric_;
run;
```



Variable	Label	N Miss	N	Minimum	Maximum	Mean	Median	Std Dev
duration	duration	50	781	41.9493694	305.6217107	154.7757191	154.2845505	48.3499237
no_pasg	no_pasg	0	831	29.0000000	87.0000000	60.0553550	60.0000000	7.4913166
speed_ground	speed_ground	0	831	33.5741041	132.7846766	79.5426997	79.7939604	18.7356754
speed_air	speed_air	628	203	90.0028586	132.9114649	103.4850352	101.1189240	9.7362774
height	height	0	831	6.2275178	59.9459639	30.4578695	30.1670844	9.7848114
pitch	pitch	0	831	2.2844801	5.9267842	4.0051609	4.0010380	0.5265690
distance	distance	0	831	41.7223127	5381.96	1522.48	1262.15	896.3381524
aircraft_1		0	831	0	0	0	0	0

```
/* creating ranking groups */
proc rank data=combined_valid groups=4 descending out=combined_ranks;
var no_pasg speed_ground height pitch distance;
ranks rank_pasg rank_speed_g rank_height rank_pitch rank_dist;
run;

/* replacing missing values for duration by median*/
proc sort data=combined_ranks;
by aircraft rank_pitch;
run;

proc stdize data=combined_ranks reponly method=median out=combined_clean;
var duration;
by aircraft rank_pitch;
run;
proc print data=combined_clean;
run;

/* saving selected data */
data combined_cleaned;
set combined_clean;
keep aircraft duration no_pasg speed_ground speed_air height pitch distance;
run;
proc contents data=combined_cleaned;
run;
```



Observations	831
Variables	9
Indexes	0
Observation Length	80
Deleted Observations	0
Compressed	NO
Sorted	NO

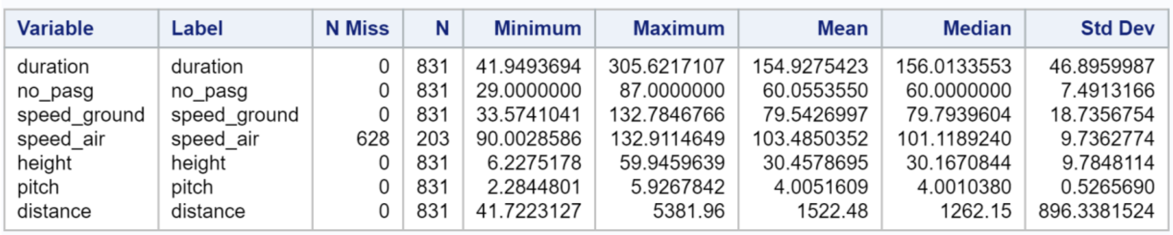
#	Variable	Type
1	aircraft	Char
8	distance	Num
2	duration	Num
6	height	Num
3	no_pasg	Num
7	pitch	Num
5	speed_air	Num
4	speed_ground	Num

First, I am removing invalid entries from the data and then treating the missing values. If missing values are treated first, then calculations I will be doing to replace missing values(mean, median etc.) will be affected by presence of invalid entries in data set.

Creating ranks for each variable to divide its values into 4 groups. This is done, to calculate median for different combination of buckets. This value of each bucket is used to fill the missing value accordingly.

I have preferred median over mean to replace missing values because it is a robust variable; it's not affected by extreme values. Bucketing is done to precisely fill the probable value instead of just calculating median for all combined values.

```
/* calculating missing values for all variables */
proc means data=combined_cleaned nmiss N min max mean median std;
run;
```



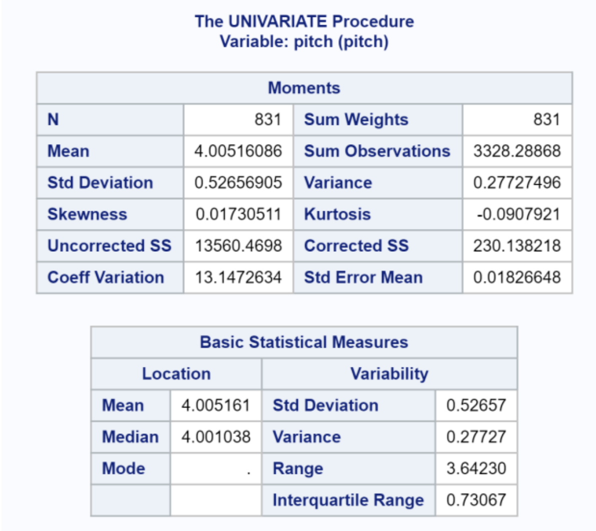
Variable	Label	N Miss	N	Minimum	Maximum	Mean	Median	Std Dev
duration	duration	0	831	41.9493694	305.6217107	154.9275423	156.0133553	46.8959987
no_pasg	no_pasg	0	831	29.0000000	87.0000000	60.0553550	60.0000000	7.4913166
speed_ground	speed_ground	0	831	33.5741041	132.7846766	79.5426997	79.7939604	18.7356754
speed_air	speed_air	628	203	90.0028586	132.9114649	103.4850352	101.1189240	9.7362774
height	height	0	831	6.2275178	59.9459639	30.4578695	30.1670844	9.7848114
pitch	pitch	0	831	2.2844801	5.9267842	4.0051609	4.0010380	0.5265690
distance	distance	0	831	41.7223127	5381.96	1522.48	1262.15	896.3381524

50 missing values of duration variable were filled. I am not treating missing values in speed_air because 75% is a great number to have missing values in any variable. Further actions for speed_air variable will be decided based on the correlation analysis in next steps.

Chapter2: Exploratory Data Analysis

Univariate Analysis

```
/* univariate analysis */
proc univariate data=combined_cleaned;
run;
```



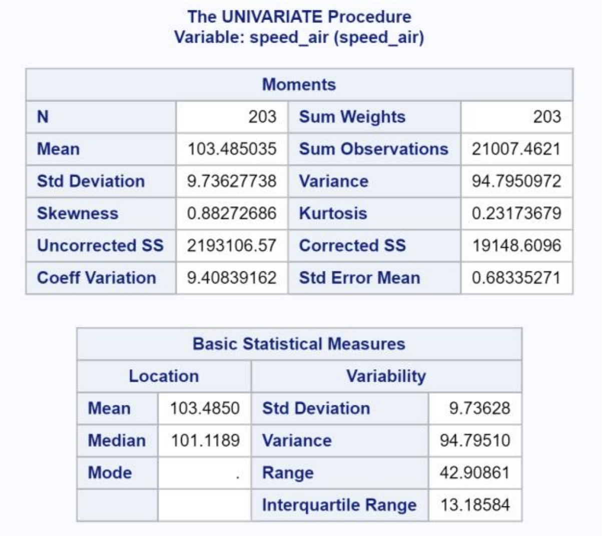
The UNIVARIATE Procedure
Variable: pitch (pitch)

Moments

N	831	Sum Weights	831
Mean	4.00516086	Sum Observations	3328.28868
Std Deviation	0.52656905	Variance	0.27727496
Skewness	0.01730511	Kurtosis	-0.0907921
Uncorrected SS	13560.4698	Corrected SS	230.138218
Coeff Variation	13.1472634	Std Error Mean	0.01826648

Basic Statistical Measures

Location		Variability	
Mean	4.005161	Std Deviation	0.52657
Median	4.001038	Variance	0.27727
Mode	.	Range	3.64230
		Interquartile Range	0.73067



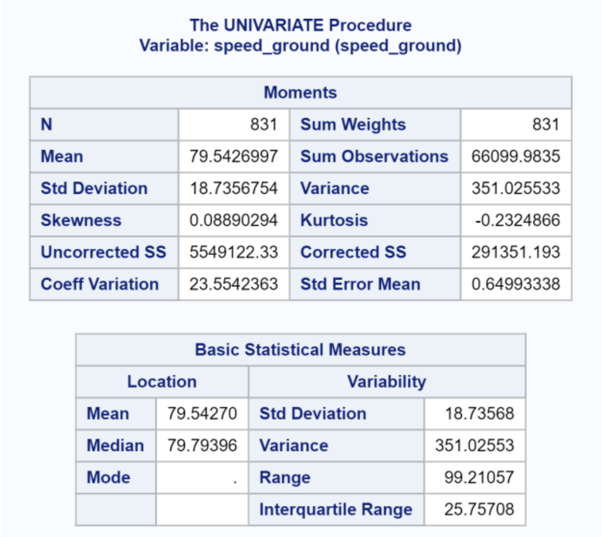
The UNIVARIATE Procedure
Variable: speed_air (speed_air)

Moments

N	203	Sum Weights	203
Mean	103.485035	Sum Observations	21007.4621
Std Deviation	9.73627738	Variance	94.7950972
Skewness	0.88272686	Kurtosis	0.23173679
Uncorrected SS	2193106.57	Corrected SS	19148.6096
Coeff Variation	9.40839162	Std Error Mean	0.68335271

Basic Statistical Measures

Location		Variability	
Mean	103.4850	Std Deviation	9.73628
Median	101.1189	Variance	94.79510
Mode	.	Range	42.90861
		Interquartile Range	13.18584



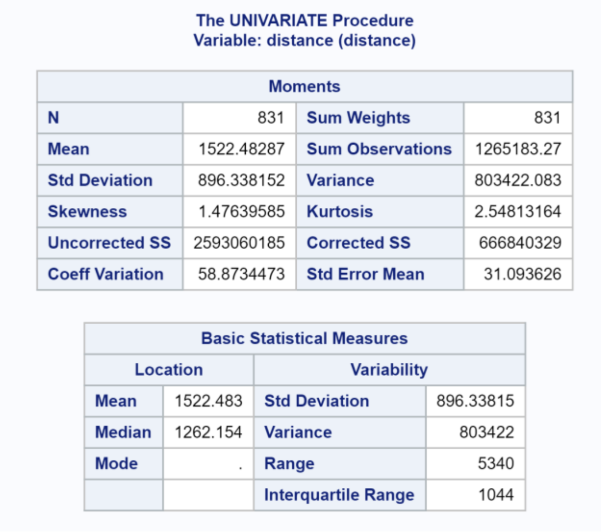
The UNIVARIATE Procedure
Variable: speed_ground (speed_ground)

Moments

N	831	Sum Weights	831
Mean	79.5426997	Sum Observations	66099.9835
Std Deviation	18.7356754	Variance	351.025533
Skewness	0.08890294	Kurtosis	-0.2324866
Uncorrected SS	5549122.33	Corrected SS	291351.193
Coeff Variation	23.5542363	Std Error Mean	0.64993338

Basic Statistical Measures

Location		Variability	
Mean	79.54270	Std Deviation	18.73568
Median	79.79396	Variance	351.02553
Mode	.	Range	99.21057
		Interquartile Range	25.75708



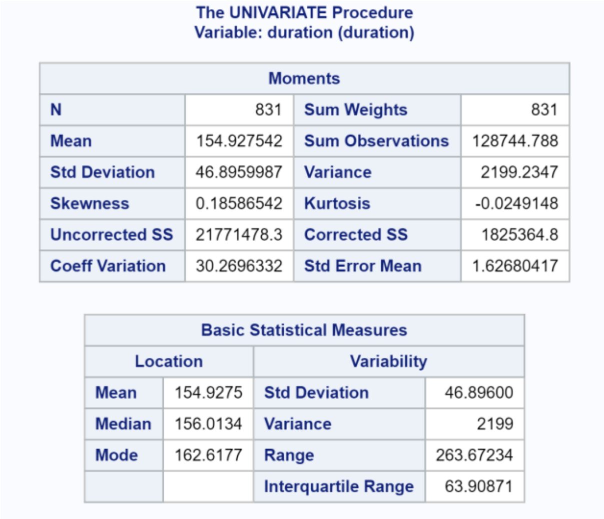
The UNIVARIATE Procedure
Variable: distance (distance)

Moments

N	831	Sum Weights	831
Mean	1522.48287	Sum Observations	1265183.27
Std Deviation	896.338152	Variance	803422.083
Skewness	1.47639585	Kurtosis	2.54813164
Uncorrected SS	2593060185	Corrected SS	666840329
Coeff Variation	58.8734473	Std Error Mean	31.093626

Basic Statistical Measures

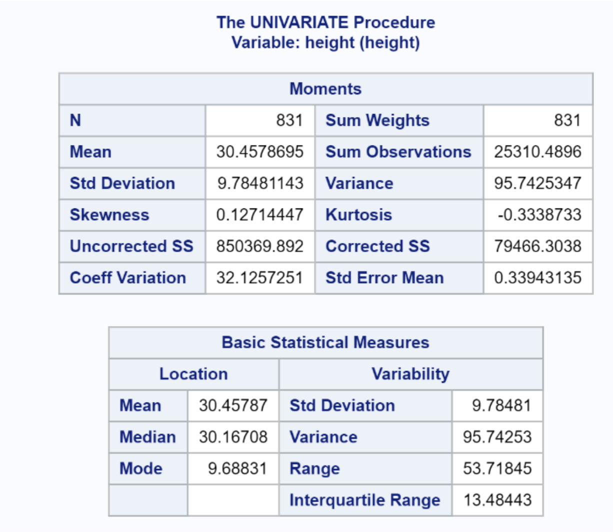
Location		Variability	
Mean	1522.483	Std Deviation	896.33815
Median	1262.154	Variance	803422
Mode	.	Range	5340
		Interquartile Range	1044



The UNIVARIATE Procedure
Variable: duration (duration)

Moments			
N	831	Sum Weights	831
Mean	154.927542	Sum Observations	128744.788
Std Deviation	46.8959987	Variance	2199.2347
Skewness	0.18586542	Kurtosis	-0.0249148
Uncorrected SS	21771478.3	Corrected SS	1825364.8
Coeff Variation	30.2696332	Std Error Mean	1.62680417

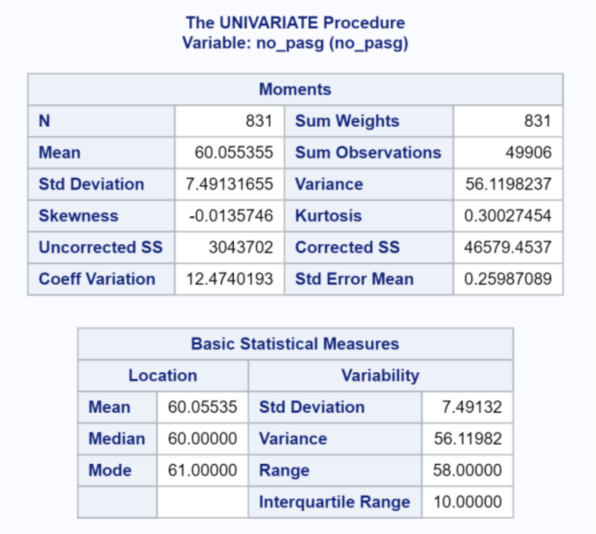
Basic Statistical Measures			
Location		Variability	
Mean	154.9275	Std Deviation	46.89600
Median	156.0134	Variance	2199
Mode	162.6177	Range	263.67234
		Interquartile Range	63.90871



The UNIVARIATE Procedure
Variable: height (height)

Moments			
N	831	Sum Weights	831
Mean	30.4578695	Sum Observations	25310.4896
Std Deviation	9.78481143	Variance	95.7425347
Skewness	0.12714447	Kurtosis	-0.3338733
Uncorrected SS	850369.892	Corrected SS	79466.3038
Coeff Variation	32.1257251	Std Error Mean	0.33943135

Basic Statistical Measures			
Location		Variability	
Mean	30.45787	Std Deviation	9.78481
Median	30.16708	Variance	95.74253
Mode	9.68831	Range	53.71845
		Interquartile Range	13.48443



The UNIVARIATE Procedure
Variable: no_pasg (no_pasg)

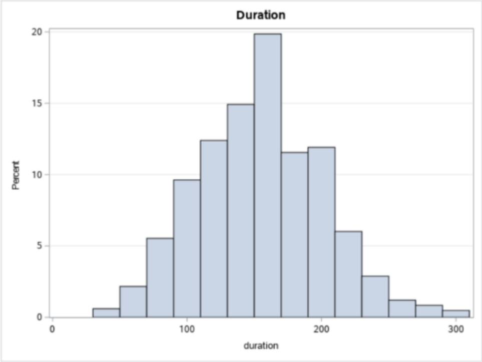
Moments			
N	831	Sum Weights	831
Mean	60.055355	Sum Observations	49906
Std Deviation	7.49131655	Variance	56.1198237
Skewness	-0.0135746	Kurtosis	0.30027454
Uncorrected SS	3043702	Corrected SS	46579.4537
Coeff Variation	12.4740193	Std Error Mean	0.25987089

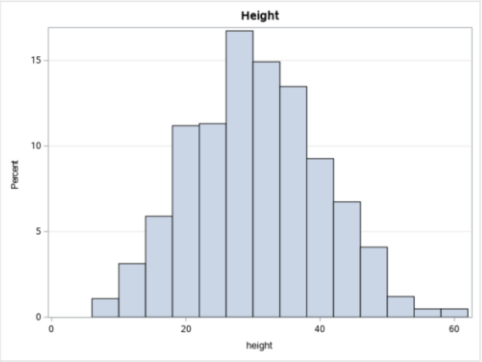
Basic Statistical Measures			
Location		Variability	
Mean	60.05535	Std Deviation	7.49132
Median	60.00000	Variance	56.11982
Mode	61.00000	Range	58.00000
		Interquartile Range	10.00000

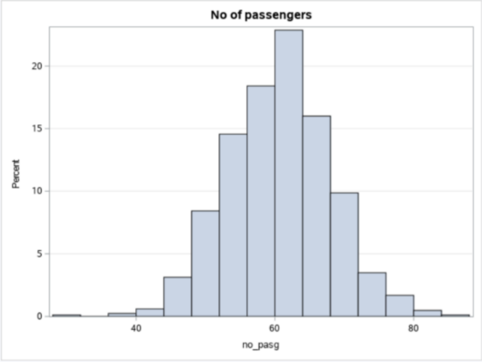
Histograms of all variables to understand the distribution

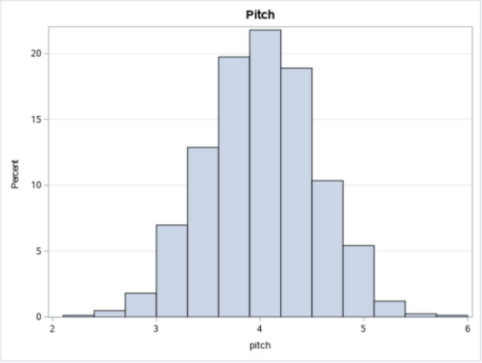
```
proc sgplot data=WORK.COMBINED_CLEANED;
	title height=12pt "Duration";
	histogram duration /;
	yaxis grid;
run;

proc sgplot data=WORK.COMBINED_CLEANED;
	title height=12pt "No of passengers";
	histogram no_pasg /;
	yaxis grid;
run;

proc sgplot data=WORK.COMBINED_CLEANED;
	title height=12pt "Speed ground";
	histogram speed_ground /;
	yaxis grid;
run;


proc sgplot data=WORK.COMBINED_CLEANED;
```

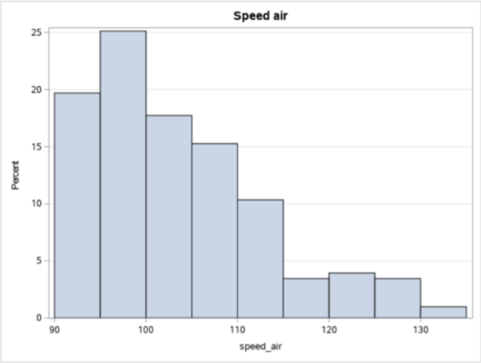
```
        title height=12pt "Speed air";
        histogram speed_air /;
        yaxis grid;
run;


proc sgplot data=WORK.COMBINED_CLEANED;
        title height=12pt "Height";
        histogram height /;
        yaxis grid;
run;


proc sgplot data=WORK.COMBINED_CLEANED;
        title height=12pt "Pitch";
        histogram pitch /;
        yaxis grid;
run;


proc sgplot data=WORK.COMBINED_CLEANED;
        title height=12pt "Distance";
        histogram distance /;
        yaxis grid;
run;
```

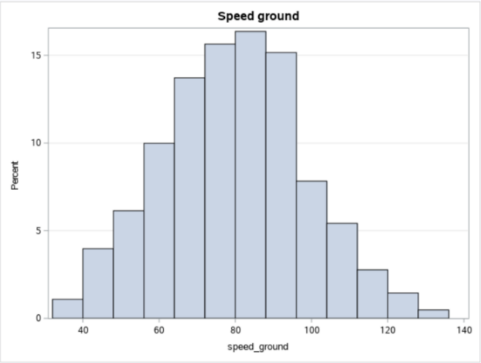


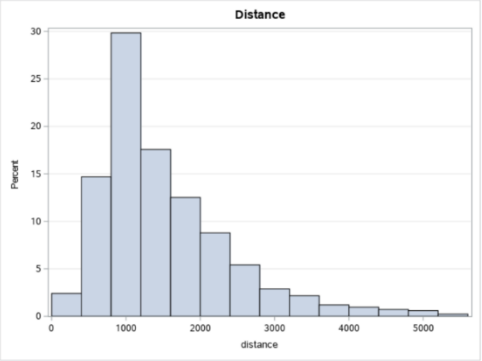












All variables have normal distribution except speed_air and duration. The distribution of these two variables is slightly right skewed.

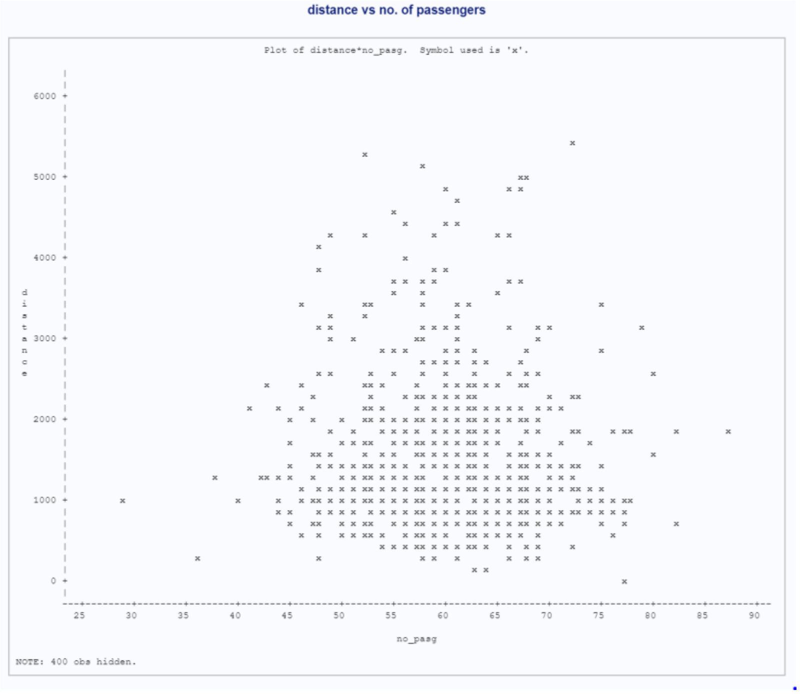
List all the questions during data preparation

- What are the assumptions that are considered while measuring/ producing this data?
- For what time span these observations are recorded?
- Does this data include observations from multiple airports? If yes, what are the geographical differences between them? (weather, location)
- At what time of the day are these observations recorded? (Day, noon, evening, night etc.)
- What are the measurements of the flights from which the data is recorded? Because structural details of the flight can make impact on the landing attributes.
- What are the sources of error responsible for invalid data according to given definitions?
- What were the lengths of the different airport landing tracks considered for data measurement?
- How experienced were the pilots in respective flights? Can we categorize them in term of experience/skills to consider that factor?
- Is there any specific reason for the absence of duration variable for second dataset (FAA2)?

Bi-variate Analysis

```
proc plot data= combined_cleaned;
title "distance vs no. of passengers";
plot distance*no_pasg='x';


proc plot data= combined_cleaned;
title "distance vs speed air";
plot distance*speed_air='x';
```

```
proc plot data= combined_cleaned;
title "distance vs speed ground";
plot distance*speed_ground='x';


proc plot data= combined_cleaned;
title "distance vs height";
plot distance*height='x';

proc plot data= combined_cleaned;
title "distance vs pitch";
plot distance*pitch='x';

proc plot data= combined_cleaned;
title "distance vs duration";
plot distance*duration='x';
```

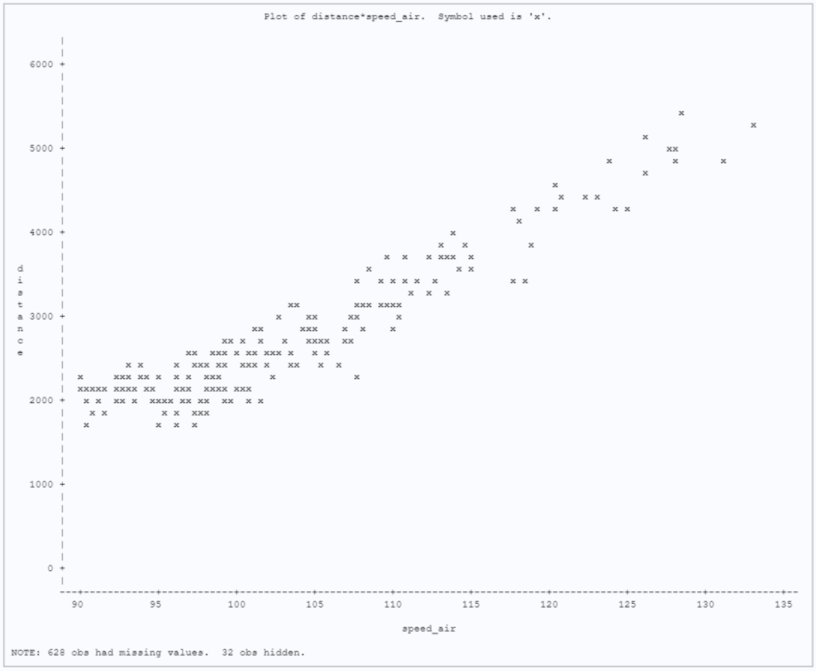


distance vs no. of passengers

```
Plot of distance*no_pasg.  Symbol used is 'x'.
```

```
NOTE: 400 obs hidden.
```

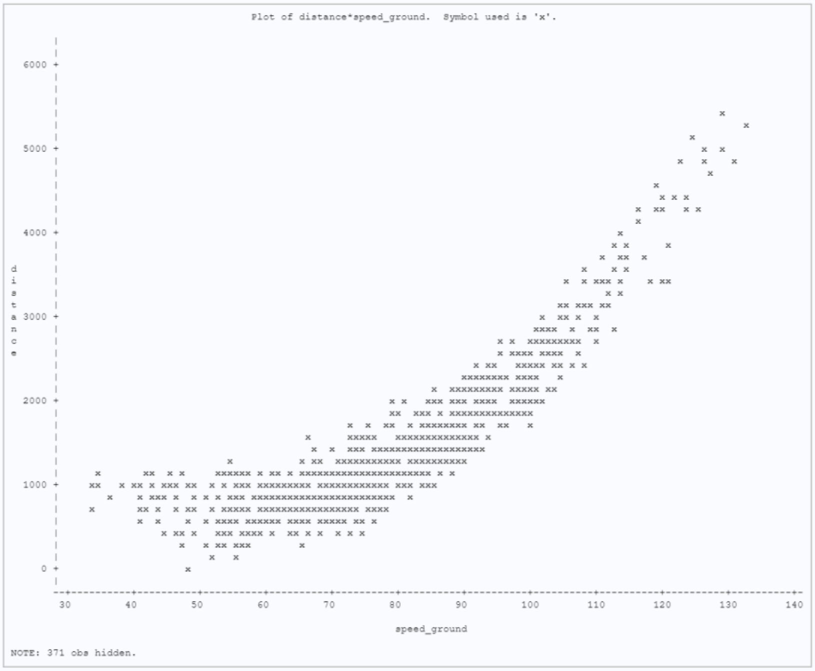
distance vs speed air



Plot of distance*speed_air. Symbol used is 'x'.

NOTE: 628 obs had missing values. 32 obs hidden.

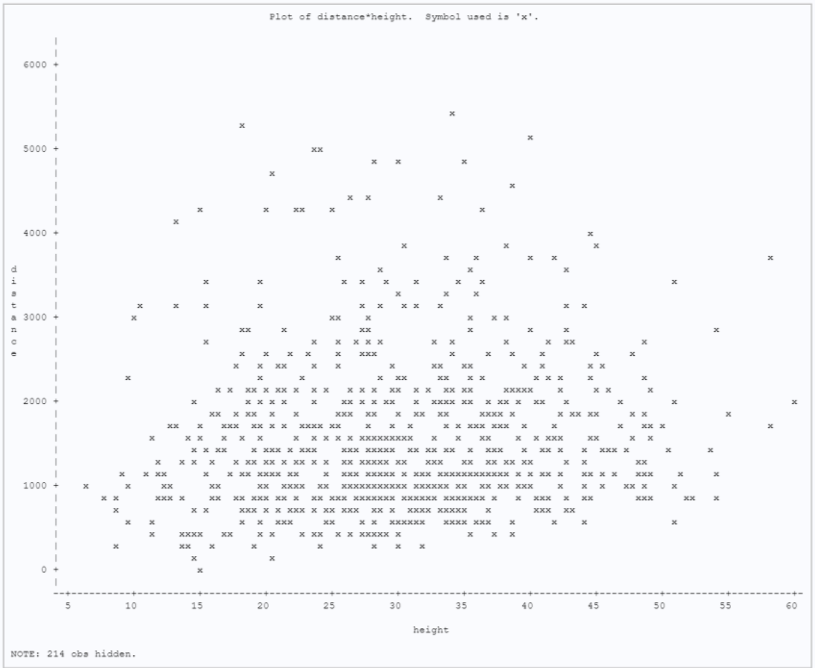
distance vs speed ground



Plot of distance*speed_ground. Symbol used is 'x'.

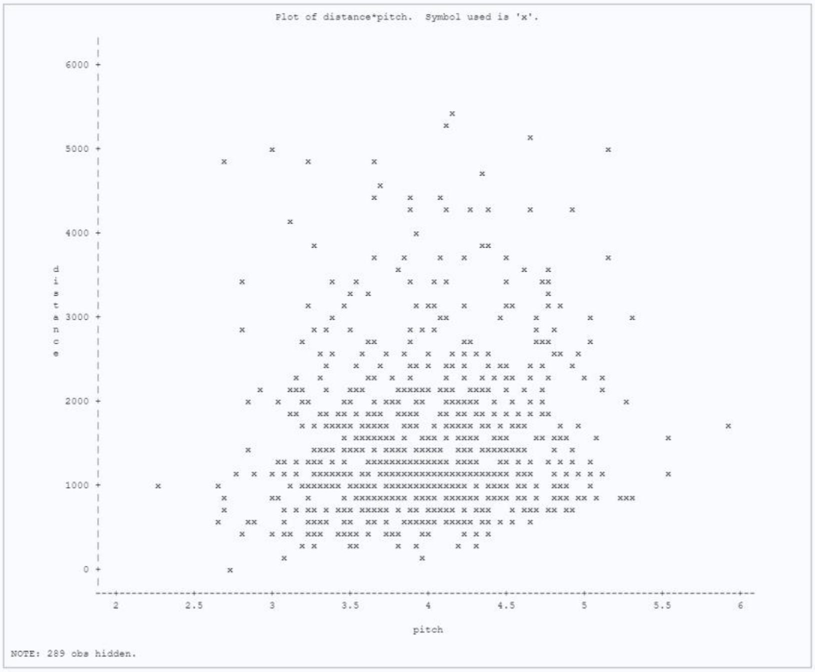
NOTE: 371 obs hidden.

distance vs height

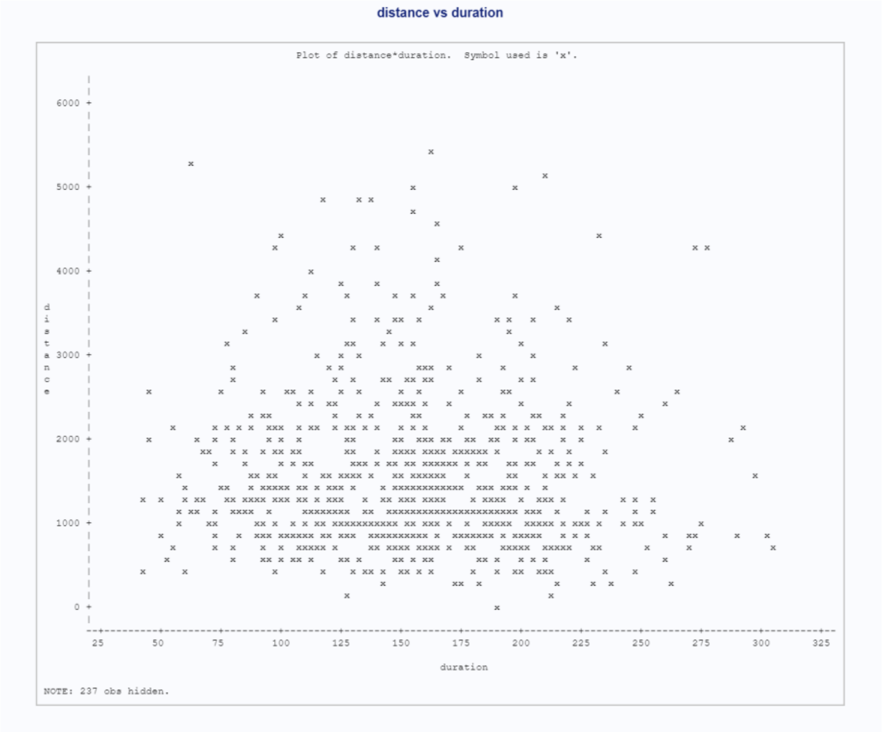


```
Plot of distance*height.  Symbol used is 'x'.

NOTE: 214 obs hidden.
```

distance vs pitch



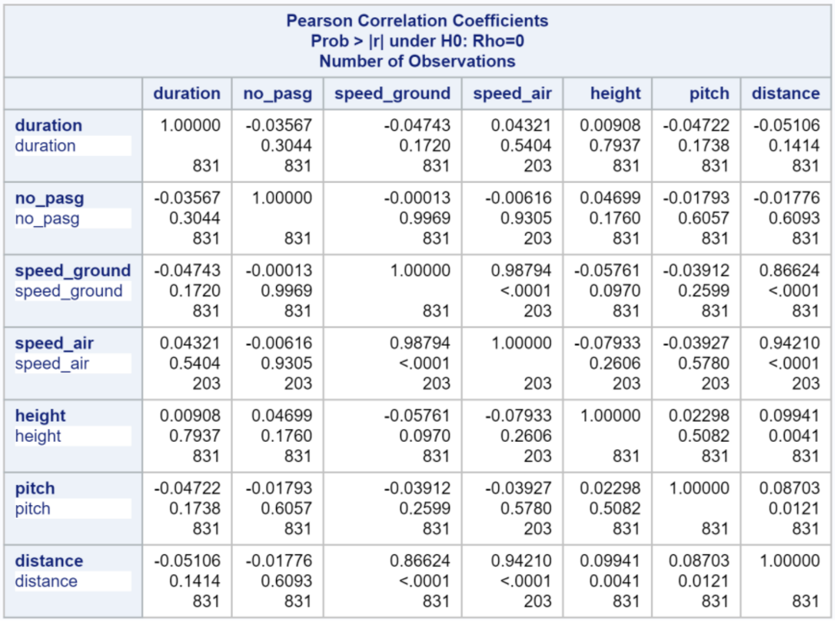
```
Plot of distance*pitch.  Symbol used is 'x'.

NOTE: 289 obs hidden.
```



By looking at above plots, we can say that only speed air and speed ground have a clear linear relationship with distance.

Correlation analysis

```
/* Correlation Coefficients*/
proc corr data = combined_cleaned;
/*where aircraft='airbus';*/
var duration no_pasg speed_ground speed_air height pitch distance;
run;
```



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

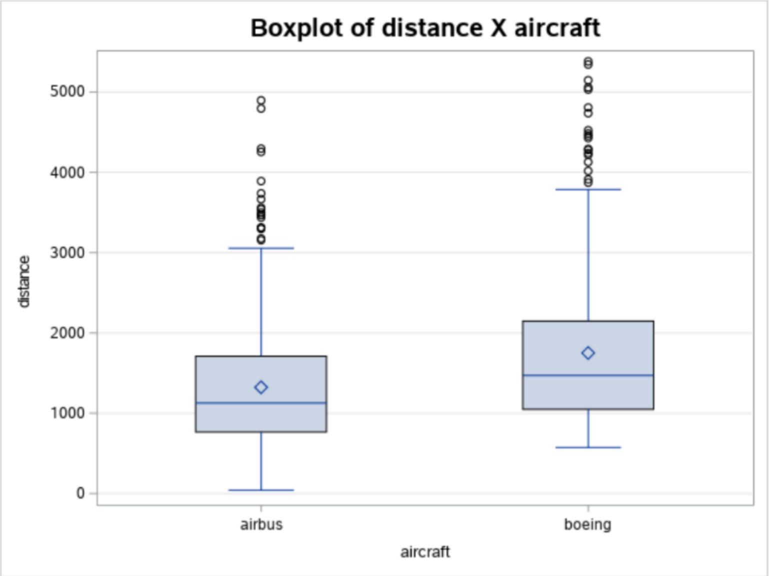
	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
duration duration	1.00000 831	-0.03567 0.3044 831	-0.04743 0.1720 831	0.04321 0.5404 203	0.00908 0.7937 831	-0.04722 0.1738 831	-0.05106 0.1414 831
no_pasg no_pasg	-0.03567 0.3044 831	1.00000 831	-0.00013 0.9969 831	-0.00616 0.9305 203	0.04699 0.1760 831	-0.01793 0.6057 831	-0.01776 0.6093 831
speed_ground speed_ground	-0.04743 0.1720 831	-0.00013 0.9969 831	1.00000 831	0.98794 <.0001 203	-0.05761 0.0970 831	-0.03912 0.2599 831	0.86624 <.0001 831
speed_air speed_air	0.04321 0.5404 203	-0.00616 0.9305 203	0.98794 <.0001 203	1.00000 203	-0.07933 0.2606 203	-0.03927 0.5780 203	0.94210 <.0001 203
height height	0.00908 0.7937 831	0.04699 0.1760 831	-0.05761 0.0970 831	-0.07933 0.2606 203	1.00000 831	0.02298 0.5082 831	0.09941 0.0041 831
pitch pitch	-0.04722 0.1738 831	-0.01793 0.6057 831	-0.03912 0.2599 831	-0.03927 0.5780 203	0.02298 0.5082 831	1.00000 831	0.08703 0.0121 831
distance distance	-0.05106 0.1414 831	-0.01776 0.6093 831	0.86624 <.0001 831	0.94210 <.0001 203	0.09941 0.0041 831	0.08703 0.0121 831	1.00000 831

From above table, we can conclude that speed air is most correlated with distance (0.94) followed by speed ground(0.86). However, speed air and speed ground have correlation coefficient almost 1 (0.987). Since speed air has 75% missing values and highly correlated with speed ground, we can avoid using speed_air variable while building the model.

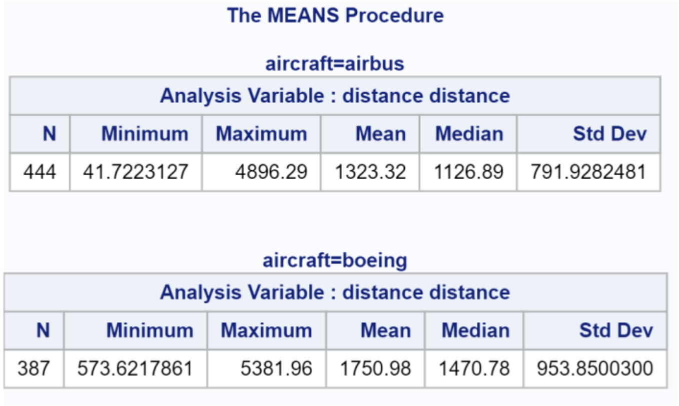
Duration, no_pasg and pitch these variables have very correlation with distance.

Comparison between two aircrafts

```
proc sgplot data=WORK.COMBINED_CLEANED;
       title height=14pt "Boxplot of distance X aircraft";
       vbox distance / category=aircraft;
       yaxis grid;
run;
```



```
/* basic stat comparison*/
proc means data=combined_cleaned N min max mean median std;
var distance;
by aircraft;
```



The MEANS Procedure

aircraft=airbus

Analysis Variable : distance distance

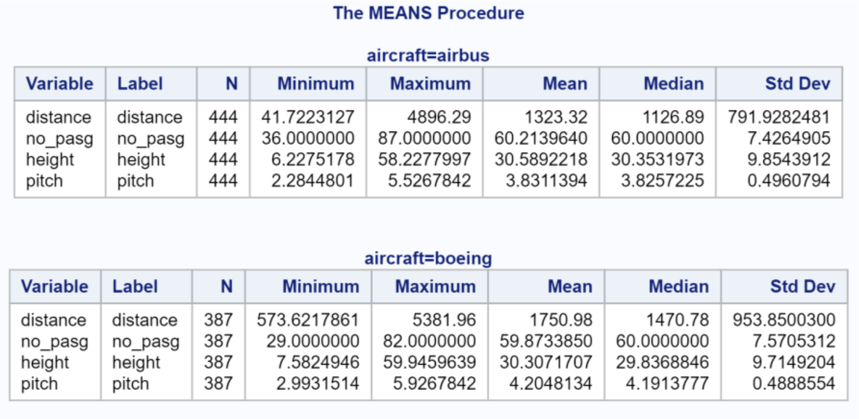
N	Minimum	Maximum	Mean	Median	Std Dev
444	41.7223127	4896.29	1323.32	1126.89	791.9282481

aircraft=boeing

Analysis Variable : distance distance

N	Minimum	Maximum	Mean	Median	Std Dev
387	573.6217861	5381.96	1750.98	1470.78	953.8500300

```
/* basic stat comparison*/
proc means data=combined_cleaned N min max mean median std;
var distance no_pasg height pitch;
by aircraft;
```



The MEANS Procedure

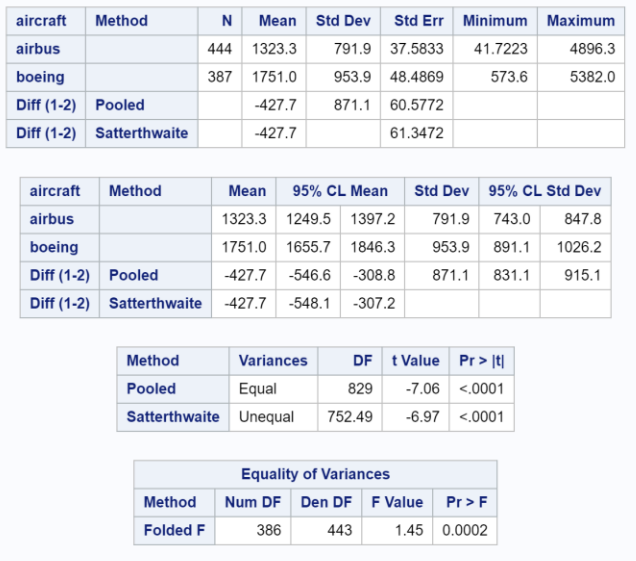
aircraft=airbus

Variable	Label	N	Minimum	Maximum	Mean	Median	Std Dev
distance	distance	444	41.7223127	4896.29	1323.32	1126.89	791.9282481
no_pasg	no_pasg	444	36.0000000	87.0000000	60.2139640	60.0000000	7.4264905
height	height	444	6.2275178	58.2277997	30.5892218	30.3531973	9.8543912
pitch	pitch	444	2.2844801	5.5267842	3.8311394	3.8257225	0.4960794

aircraft=boeing

Variable	Label	N	Minimum	Maximum	Mean	Median	Std Dev
distance	distance	387	573.6217861	5381.96	1750.98	1470.78	953.8500300
no_pasg	no_pasg	387	29.0000000	82.0000000	59.8733850	60.0000000	7.5705312
height	height	387	7.5824946	59.9459639	30.3071707	29.8368846	9.7149204
pitch	pitch	387	2.9931514	5.9267842	4.2048134	4.1913777	0.4888554

```
/* t test between two aircrafts */
proc ttest data=combined_cleaned;
class aircraft;
var distance;
run;
```

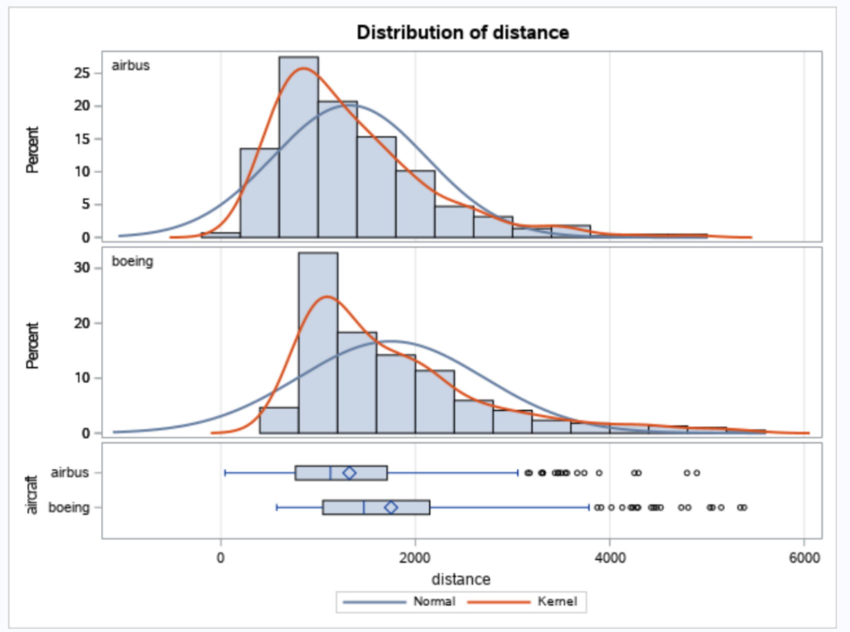


aircraft	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus		444	1323.3	791.9	37.5833	41.7223	4896.3
boeing		387	1751.0	953.9	48.4869	573.6	5382.0
Diff (1-2)	Pooled		-427.7	871.1	60.5772		
Diff (1-2)	Satterthwaite		-427.7		61.3472		

aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		1323.3	1249.5	1397.2	791.9	743.0	847.8
boeing		1751.0	1655.7	1846.3	953.9	891.1	1026.2
Diff (1-2)	Pooled	-427.7	-546.6	-308.8	871.1	831.1	915.1
Diff (1-2)	Satterthwaite	-427.7	-548.1	-307.2			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	829	-7.06	<.0001
Satterthwaite	Unequal	752.49	-6.97	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	386	443	1.45	0.0002



By looking at the basic statistic comparison between two aircrafts and results of t-test we can say that there is a significant difference between distance of Airbus and Boeing aircrafts.

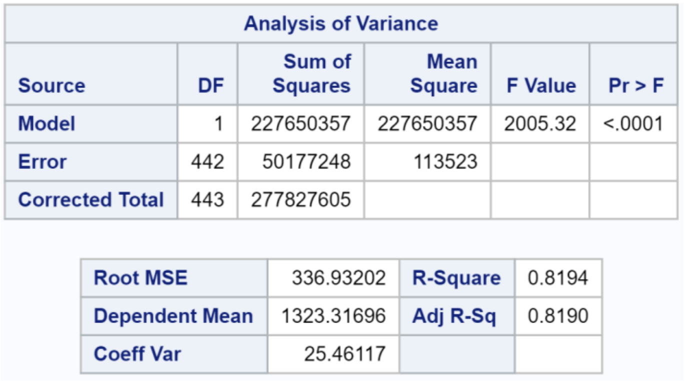
Boeing aircraft's mean distance is greater than Airbus's mean distance by almost 427 units. So, modelling should be done for these two aircrafts separately to predict the distance more accurately.

Chapter3: Linear Regression Model

Airbus aircraft

Variables are selected according to the descending order of correlation coefficient with distance variable for each iteration.

```
/* airbus */
/* iteration 1*/
proc reg data=combined_cleaned;
where aircraft='airbus';
model distance = speed_ground;
title Airbus: regression analysis model;
run;
/* 0.8194 */
```

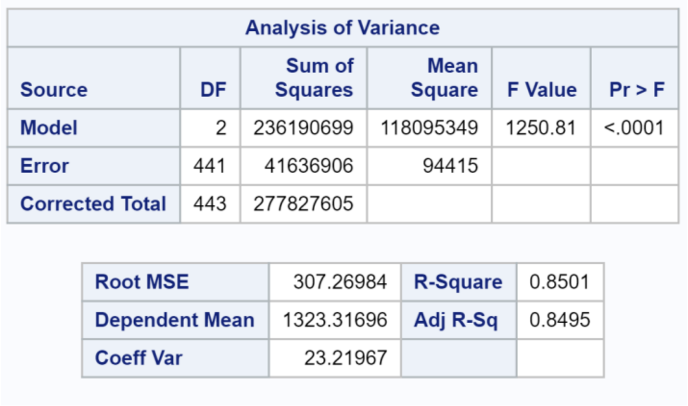


Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	227650357	227650357	2005.32	<.0001
Error	442	50177248	113523		
Corrected Total	443	277827605			

Root MSE	336.93202	R-Square	0.8194
Dependent Mean	1323.31696	Adj R-Sq	0.8190
Coeff Var	25.46117		

```
/* iteration 2*/
proc reg data=combined_cleaned;
where aircraft='airbus';
model distance = speed_ground height;
title Airbus: regression analysis model;
run;
/* 0.8501*/
```



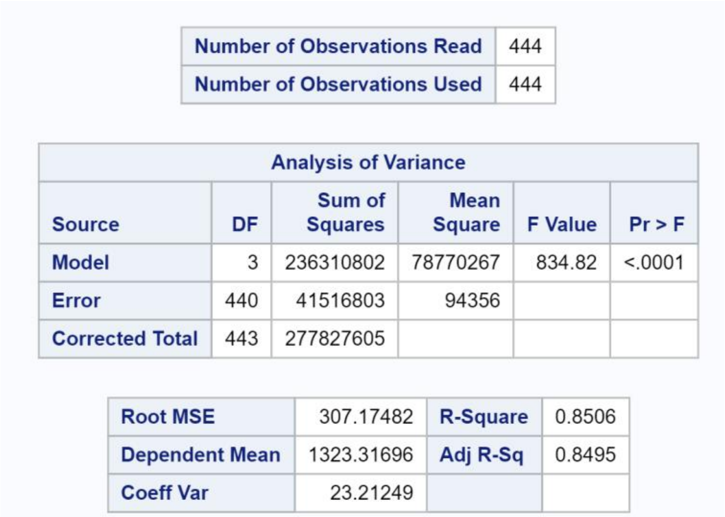
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	236190699	118095349	1250.81	<.0001
Error	441	41636906	94415		
Corrected Total	443	277827605			

Root MSE	307.26984	R-Square	0.8501
Dependent Mean	1323.31696	Adj R-Sq	0.8495
Coeff Var	23.21967		

```
/* iteration 3*/
proc reg data=combined_cleaned;
where aircraft='airbus';
```

```
model distance = speed_ground height duration;
title Airbus: regression analysis model;
run;
/* 0.8506 */
```

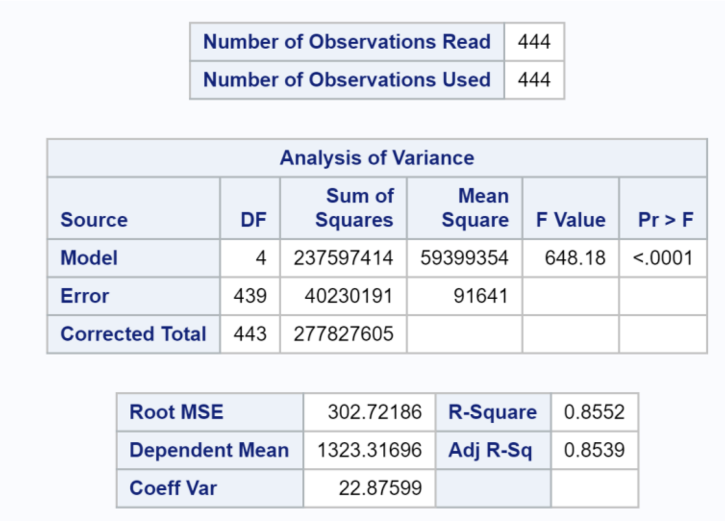


Number of Observations Read	444
Number of Observations Used	444

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	236310802	78770267	834.82	<.0001
Error	440	41516803	94356		
Corrected Total	443	277827605			

Root MSE	307.17482	R-Square	0.8506
Dependent Mean	1323.31696	Adj R-Sq	0.8495
Coeff Var	23.21249		

```
/* iteration 4*/
proc reg data=combined_cleaned;
where aircraft='airbus';
model distance = speed_ground height duration pitch;
title Airbus: regression analysis model;
run;
/* 0.8552 */
```

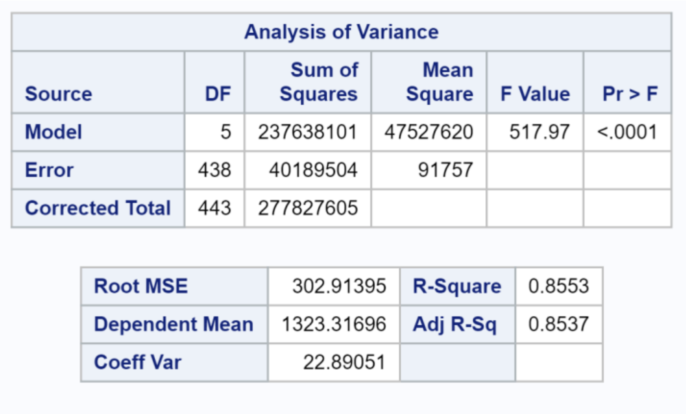


Number of Observations Read	444
Number of Observations Used	444

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	237597414	59399354	648.18	<.0001
Error	439	40230191	91641		
Corrected Total	443	277827605			

Root MSE	302.72186	R-Square	0.8552
Dependent Mean	1323.31696	Adj R-Sq	0.8539
Coeff Var	22.87599		

```
/* iteration 5*/
proc reg data=combined_cleaned;
where aircraft='airbus';
model distance = speed_ground height duration pitch no_pasg;
title Airbus: regression analysis model;
run;
/* 0.8553 */
```



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	237638101	47527620	517.97	<.0001
Error	438	40189504	91757		
Corrected Total	443	277827605			

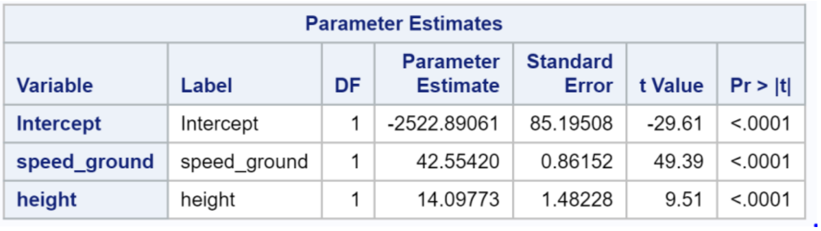
Root MSE	302.91395	R-Square	0.8553
Dependent Mean	1323.31696	Adj R-Sq	0.8537
Coeff Var	22.89051		

Summary of all iterations for aircraft Airbus

Airbus				
# Iteration	Independent variables	List of variables	R square	Adj R square
1	1	speed_ground	0.8194	0.819
2	2	speed_ground height	0.8501	0.8495
3	3	speed_ground height duration	0.8506	0.8495
4	4	speed_ground height duration pitch	0.8552	0.8539
5	5	speed_ground height duration pitch no_pasg	0.8553	0.8537

By looking at the summary table, we can say that after including height along with speed_ground in iteration 2, R square values increased and did not change much in further iterations. So, iteration 2 can chosen as the final one for modelling.

Iteration 2 results are as following:



Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2522.89061	85.19508	-29.61	<.0001
speed_ground	speed_ground	1	42.55420	0.86152	49.39	<.0001
height	height	1	14.09773	1.48228	9.51	<.0001

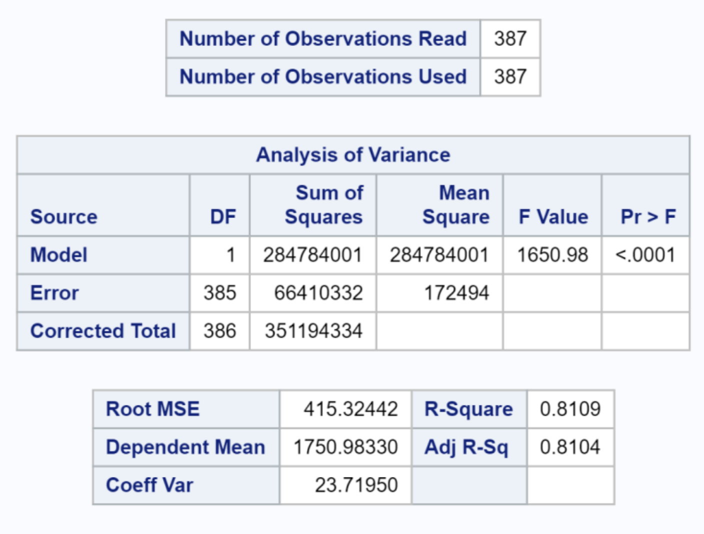
So, Final equation for Airbus aircraft is

Distance = -2522.89061 + 42.55420 * speed_ground + 14.09773 * height

Boeing aircraft

Variables are selected according to the descending order of correlation coefficient with distance variable for each iteration.

```
/* boeing */
/* iteration 1*/
proc reg data=combined_cleaned;
where aircraft='boeing';
model distance = speed_ground;
title Airbus: regression analysis model;
run;
/* 0.8109*/
```

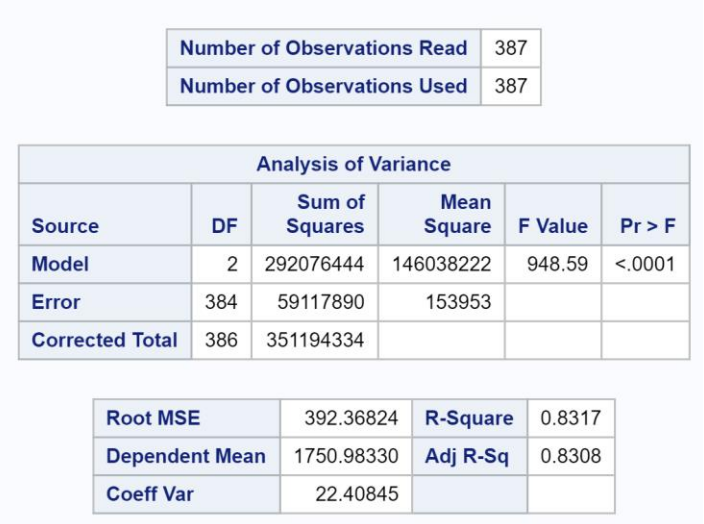


Number of Observations Read	387
Number of Observations Used	387

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	284784001	284784001	1650.98	<.0001
Error	385	66410332	172494		
Corrected Total	386	351194334			

Root MSE	415.32442	R-Square	0.8109
Dependent Mean	1750.98330	Adj R-Sq	0.8104
Coeff Var	23.71950		

```
/* iteration 2*/
proc reg data=combined_cleaned;
where aircraft='boeing';
model distance = speed_ground height;
title Airbus: regression analysis model;
run;
/* 0.8317*/
```

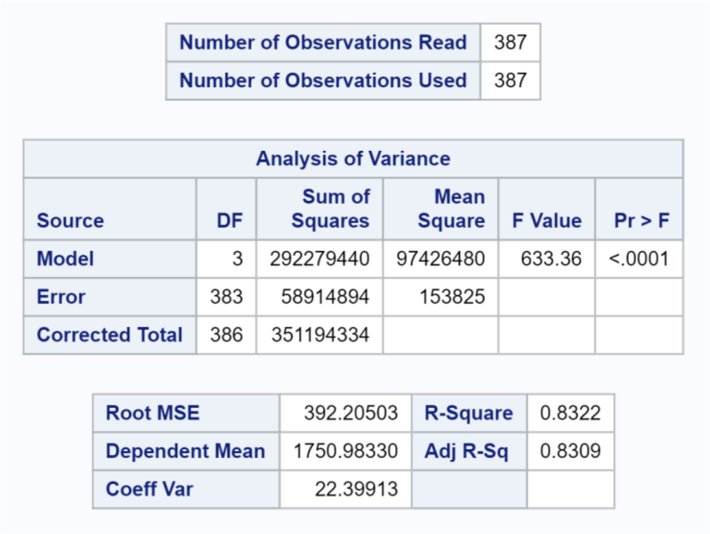


Number of Observations Read	387
Number of Observations Used	387

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	292076444	146038222	948.59	<.0001
Error	384	59117890	153953		
Corrected Total	386	351194334			

Root MSE	392.36824	R-Square	0.8317
Dependent Mean	1750.98330	Adj R-Sq	0.8308
Coeff Var	22.40845		

```
/* iteration 3*/
proc reg data=combined_cleaned;
where aircraft='boeing';
model distance = speed_ground height duration;
title Airbus: regression analysis model;
run;
/* 0.8322 */
```

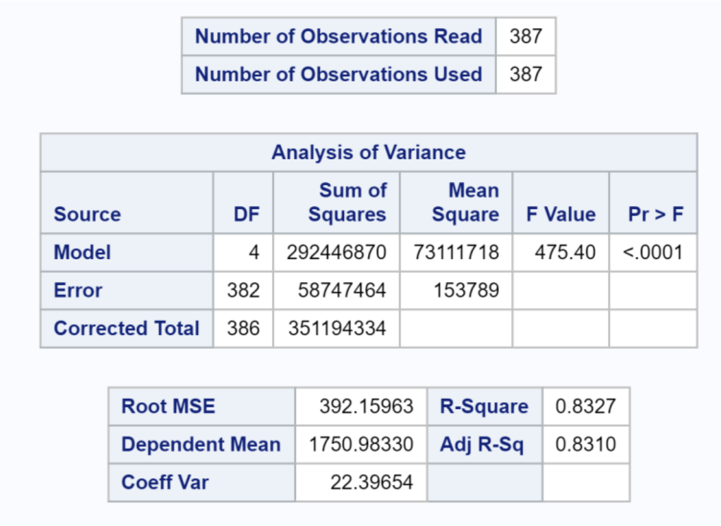


Number of Observations Read	387
Number of Observations Used	387

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	292279440	97426480	633.36	<.0001
Error	383	58914894	153825		
Corrected Total	386	351194334			

Root MSE	392.20503	R-Square	0.8322
Dependent Mean	1750.98330	Adj R-Sq	0.8309
Coeff Var	22.39913		

```
/* iteration 4*/
proc reg data=combined_cleaned;
where aircraft='boeing';
model distance = speed_ground height duration pitch;
title Airbus: regression analysis model;
run;
/* 0.8327 */
```

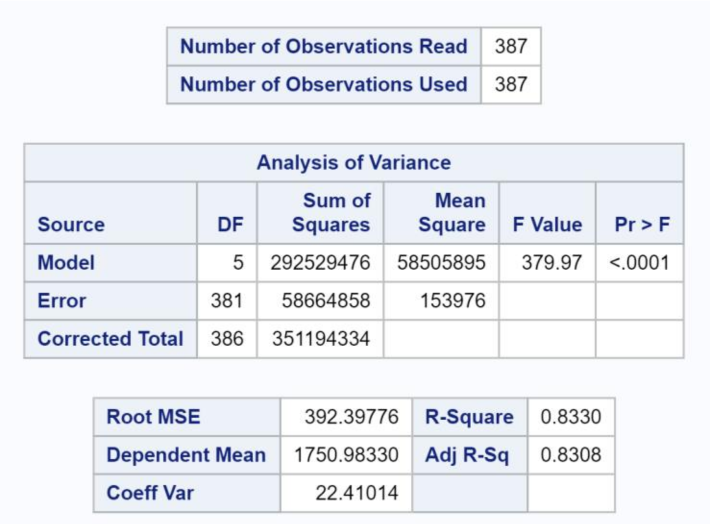


Number of Observations Read	387
Number of Observations Used	387

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	292446870	73111718	475.40	<.0001
Error	382	58747464	153789		
Corrected Total	386	351194334			

Root MSE	392.15963	R-Square	0.8327
Dependent Mean	1750.98330	Adj R-Sq	0.8310
Coeff Var	22.39654		

```
/* iteration 5*/
proc reg data=combined_cleaned;
where aircraft='boeing';
model distance = speed_ground height duration pitch no_pasg;
title Airbus: regression analysis model;
run;
/* 0.8330 */
```



Number of Observations Read	387
Number of Observations Used	387

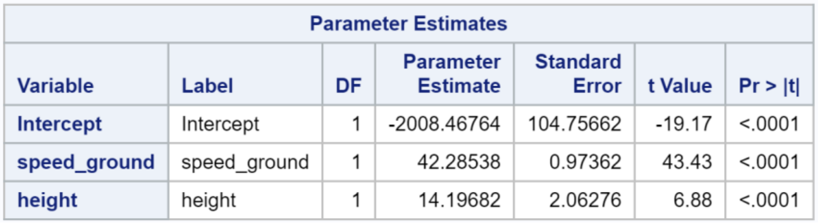
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	292529476	58505895	379.97	<.0001
Error	381	58664858	153976		
Corrected Total	386	351194334			

Root MSE	392.39776	R-Square	0.8330
Dependent Mean	1750.98330	Adj R-Sq	0.8308
Coeff Var	22.41014		

Summary of all iterations for aircraft Airbus

Boeing				
# Iteration	Independent variables	List of variables	R square	Adj R square
1	1	speed_ground	0.8109	0.8104
2	2	speed_ground height	0.8317	0.8308
3	3	speed_ground height duration	0.8322	0.8309
4	4	speed_ground height duration pitch	0.8327	0.831
5	5	speed_ground height duration pitch no_pasg	0.833	0.8308

By looking at the summary table, we can say that after including height along with speed_ground in iteration 2, R square values increased and did not change much in further iterations. So, iteration 2 can chosen as the final one for modelling.
Iteration 2 results are as following:



Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2008.46764	104.75662	-19.17	<.0001
speed_ground	speed_ground	1	42.28538	0.97362	43.43	<.0001
height	height	1	14.19682	2.06276	6.88	<.0001

So, Final equation for Boeing aircraft is

Distance = -2008.46764 + 42.28538 * speed_ground + 14.19682 * height

Chapter4: Questions

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

   I have used 831 observations in my final model.

   Initially there were 950 observations after combining two data sets and removing empty rows. Then 100 observations were removed while deleting the duplicate records from data. Later, 19 invalid observations were removed based on the given definition for validity of each variable. There were few invalid entries, so removal of these rows will not affect the model results much.

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

From Correlation coefficients table, we can conclude that speed air is most correlated with distance (0.94) followed by speed ground(0.86). However, speed air and speed ground have correlation coefficient almost 1 (0.987). Since speed air has 75% missing values and highly correlated with speed ground, we can avoid using speed_air variable while building the model.

Duration, no_pasg and pitch these variables have very correlation with distance. These two variables have negative correlation coefficient with distance. That means, distance will decrease if we increase duration and no_pasg. Other variables have positive correlation coefficients with distance. So, distance will increase if speed_air, speed_ground, height and pitch are increased

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

Yes, there is difference between the two makes Boeing and Airbus.

By looking at the basic statistic comparison between two aircrafts and results of t-test we can say that there is a significant difference between distance of Airbus and Boeing aircrafts. Boeing aircraft’s mean distance is greater than Airbus’s mean distance by almost 427 units.