BANA 6043: Statistical Computing

Project: Factors impacting landing distance

Submitted By: Anupreet Gupta (M12823098)

The FAA data set consists of parameters including pitch, ground speed, air speed, duration of flight, no. of passengers, height of plane while landing and landing distance. The task in hand was to determine and predict the landing distance of a flight based on above parameters so that over run could have been avoided. After performing data cleaning and exploratory data analysis procedures, 831/950 (~ 87 %) of observations were used in regression model to determine the impact of parameter and successfully predict the landing distance. The analysis revels that ground speed, height and pitch impacts the landing distance with ground speed having major impact. While if only Boeing Aircrafts are looked into, pitch doesn't have an impact on landing distance. Thorough analysis attached.

Chapter 01: Data Preparation

Goal:

The goal of this chapter is to setup the given data into system and prepare it for data analysis purposes by performing the data sanity check on each variable.

Steps Followed:

1. **Uploading the available data sets (FAA1.xlsx and FAA1.xlsx) into the SAS environment and merging the two data sets.**

```
*Importing Data Set FAA1;
%web_drop_table(BANA6043.FAA1);
FILENAME REFFILE '/folders/myfolders/BANA_6043/FAA1.xls';
PROC IMPORT DATAFILE=REFFILE
      DBMS=XLS
      OUT=BANA6043.FAA1;
      GETNAMES=YES;
RUN;
PROC CONTENTS DATA=BANA6043.FAA1; RUN;

*Importing Data Set FAA2;
%web_drop_table(BANA6043.FAA2);
FILENAME REFFILE '/folders/myfolders/BANA_6043/FAA2.xls';
PROC IMPORT DATAFILE=REFFILE
      DBMS=XLS
      OUT=BANA6043.FAA2;
      GETNAMES=YES;
RUN;
PROC CONTENTS DATA=BANA6043.FAA2; RUN;

*Merging both the data sets;
DATA BANA6043.FLIGHTS_COMBINED;
SET BANA6043.FAA1 BANA6043.FAA2;
RUN;
PROC PRINT DATA=BANA6043.FLIGHTS_COMBINED (OBS=10);
RUN;
```

2. **Removing observations with complete missing values and duplicates.**
   As a part of data preparation, it is imperative to get rid of missing observations and duplicates. Also, while estimating the correlation of any variable w.r.t dependent variable, the variable should not have missing values for more than 25-35 % of total observations to have correct estimation.

Hence, we shall be also analyzing missing values count variable wise to rule out any variable which has a considerable amount of missing values.

```
*Removing missing observations (The data set currently has 50 missing observations);
DATA BANA6043.FLIGHTS_COMBINED;
SET BANA6043.FLIGHTS_COMBINED;
IF MISSING(NO_PASG) AND MISSING(DURATION) AND MISSING(AIRCRAFT) AND MISSING(SPEED_GROUND)
AND MISSING(SPEED_AIR) AND MISSING(HEIGHT) AND MISSING(PITCH) AND MISSING(DISTANCE)
THEN DELETE;
RUN;
PROC PRINT DATA=BANA6043.FLIGHTS_COMBINED;
RUN;

*Removing duplicates (There were 100 observations which were removed followed by this step);

PROC SORT DATA=BANA6043.FLIGHTS_COMBINED  NODUPKEY;
BY AIRCRAFT NO_PASG SPEED_AIR SPEED_GROUND HEIGHT PITCH DISTANCE;

PROC MEANS DATA=BANA6043.FLIGHTS_COMBINED N NMISS MEDIAN STDDEV QRANGE  ;
TITLE Statistical Summary for Each Variable;
```

Statistical Summary For Each Variable

The MEANS Procedure

Variable	Label	N	N Miss	Median	Std Dev	Quartile Range
duration	duration	800	50	153.9480975	49.2592338	69.4432967
no_pasg	no_pasg	850	0	60.0000000	7.4931370	10.0000000
speed_ground	speed_ground	850	0	79.6428041	19.0594903	26.2129610
speed_air	speed_air	208	642	101.1473493	10.2590370	13.1907754
height	height	850	0	30.0931324	10.2877268	13.6923145
pitch	pitch	850	0	4.0082875	0.5288298	0.7371201
distance	distance	850	0	1258.09	928.5600816	1053.66

3. **Abnormality observation check.**

```
*Validity check on each variable;
DATA BANA6043.FLIGHTS_COMBINED ;
SET BANA6043.FLIGHTS_COMBINED ;
```

```
IF DURATION>=40 OR DURATION='' THEN IS_DURATION_ABNORMAL=0;
ELSE IS_DURATION_ABNORMAL=1;
IF SPEED_GROUND>=30 AND SPEED_GROUND<=140 THEN
IS_SPEED_GROUND_ABNORMAL=0;
ELSE IS_SPEED_GROUND_ABNORMAL=1;
IF (SPEED_AIR>=30 AND SPEED_AIR<=140) OR (SPEED_AIR='') THEN
IS_SPEED_AIR_ABNORMAL=0;
ELSE IS_SPEED_AIR_ABNORMAL=1;
IF HEIGHT>=6 THEN IS_HEIGHT_ABNORMAL=0;
ELSE IS_HEIGHT_ABNORMAL=1;
IF DISTANCE<=6000  THEN IS_DISTANCE_ABNORMAL=0;
ELSE IS_DISTANCE_ABNORMAL=1;
IF IS_DURATION_ABNORMAL + IS_SPEED_GROUND_ABNORMAL +
IS_SPEED_AIR_ABNORMAL + IS_HEIGHT_ABNORMAL + IS_DISTANCE_ABNORMAL > 0
THEN IS_LANDING_ABNORMAL=1;
ELSE IS_LANDING_ABNORMAL=0;
RUN;

*Created 5 new variables to flag the abnormality corresponding to each
variable and another factor which collectively determines the
abnormality;

*Normal Landing Data;
DATA BANA6043.FLIGHTS_COMBINED_NORMAL;
SET BANA6043.FLIGHTS_COMBINED;
IF IS_LANDING_ABNORMAL=0;
PROC MEANS DATA=BANA6043.FLIGHTS_COMBINED_NORMAL;
TITLE NORMAL LANDING DATA;
```

NORMAL LANDING DATA

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
duration	duration	781	154.7757191	48.3499237	41.9493694	305.6217107
no_pasg	no_pasg	831	60.0553550	7.4913166	29.0000000	87.0000000
speed_ground	speed_ground	831	79.5426997	18.7356754	33.5741041	132.7846766
speed_air	speed_air	203	103.4850352	9.7362774	90.0028586	132.9114649
height	height	831	30.4578695	9.7848114	6.2275178	59.9459639
pitch	pitch	831	4.0051609	0.5265690	2.2844801	5.9267842
distance	distance	831	1522.48	896.3381524	41.7223127	5381.96
IS_DURATION_ABNORMAL		831	0	0	0	0
IS_SPEED_GROUND_ABNORMAL		831	0	0	0	0
IS_SPEED_AIR_ABNORMAL		831	0	0	0	0
IS_HEIGHT_ABNORMAL		831	0	0	0	0
IS_DISTANCE_ABNORMAL		831	0	0	0	0
IS_LANDING_ABNORMAL		831	0	0	0	0

Statistical Summary for Normal Data

The MEANS Procedure

Variable	Label	N Miss	Median	Mode	Lower Quartile	Upper Quartile	Quartile Range
duration	duration	50	154.2845505	.	119.6314577	189.6629425	70.0314848
no_pasg	no_pasg	0	60.0000000	61.0000000	55.0000000	65.0000000	10.0000000
speed_ground	speed_ground	0	79.7939604	.	66.1925304	91.9496075	25.7570772
speed_air	speed_air	628	101.1189240	.	96.1964606	109.3823005	13.1858399
height	height	0	30.1670844	9.6883077	23.5298692	37.0143018	13.4844326
pitch	pitch	0	4.0010380	.	3.6403979	4.3710717	0.7306738
distance	distance	0	1262.15	.	892.9839743	1937.26	1044.27
IS_DURATION_ABNORMAL		0	0	0	0	0	0
IS_SPEED_GROUND_ABNORMAL		0	0	0	0	0	0
IS_SPEED_AIR_ABNORMAL		0	0	0	0	0	0
IS_HEIGHT_ABNORMAL		0	0	0	0	0	0
IS_DISTANCE_ABNORMAL		0	0	0	0	0	0
IS_LANDING_ABNORMAL		0	0	0	0	0	0

4. Treating Missing Value.

```
*After removing 19 observations from Combined data because of
abnormality, we have 831 observational data with normal values. Out of
these there are speed_air variable has 628 missing value. Hence we
will be discarding speed_air during our analysis;

*Working on missing values for duration;
DATA BANA6043.FLIGHTS_COMBINED_NORMAL;
SET BANA6043.FLIGHTS_COMBINED_NORMAL;
IF DURATION='' THEN IS_DURATION_MISSING=1;
ELSE IS_DURATION_MISSING=0;
RUN;

*Replacing missing values for duration with mean(duration) ;
PROC STDIZE DATA=bana6043.FLIGHTS_COMBINED_NORMAL REPONLY METHOD=MEAN
out=BANA6043.FLIGHTS_COMBINED_NORMAL ;
VAR DURATION;
RUN;

*Statistical Summary after cleaning data set;
PROC MEANS DATA=BANA6043.FLIGHTS_COMBINED_NORMAL N NMISS MEDIAN MODE
MIN MAX Q1 Q3 QRANGE;
TITLE Statistical Summary for Normal Data : Cleaned;
```

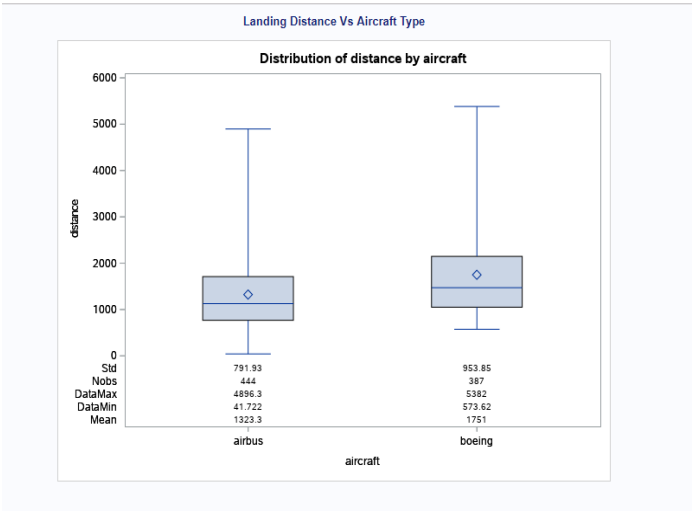
Statistical Summary for Normal Data : Cleaned

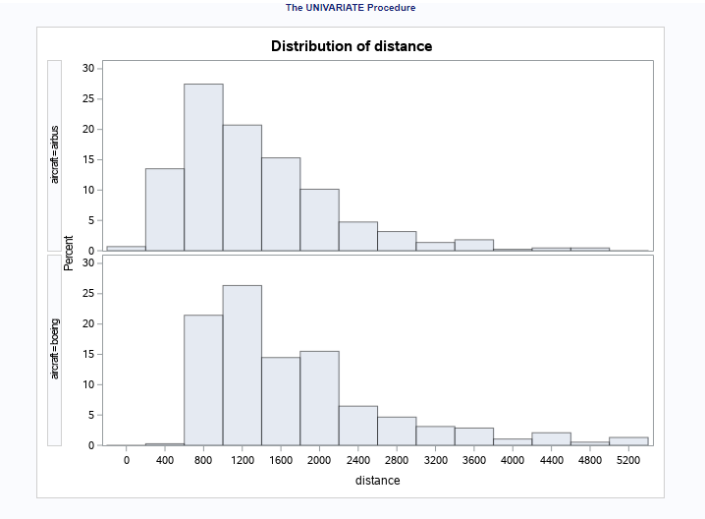
The MEANS Procedure

Variable	Label	N	N Miss	Median	Mode	Minimum	Maximum	Lower Quartile	Upper Quartile	Quartile Range
duration	duration	831	0	154.7757191	154.7757191	41.9493694	305.6217107	122.6022169	186.5109312	63.9087143
no_pasg	no_pasg	831	0	60.0000000	61.0000000	29.0000000	87.0000000	55.0000000	65.0000000	10.0000000
speed_ground	speed_ground	831	0	79.7939604	.	33.5741041	132.7846766	66.1925304	91.9496075	25.7570772
speed_air	speed_air	203	628	101.1189240	.	90.0028586	132.9114649	96.1964606	109.3823005	13.1858399
height	height	831	0	30.1670844	9.6883077	6.2275178	59.9459639	23.5298692	37.0143018	13.4844326
pitch	pitch	831	0	4.0010380	.	2.2844801	5.9267842	3.6403979	4.3710717	0.7306738
distance	distance	831	0	1262.15	.	41.7223127	5381.96	892.9839743	1937.26	1044.27
IS_DURATION_ABNORMAL		831	0	0	0	0	0	0	0	0
IS_SPEED_GROUND_ABNORMAL		831	0	0	0	0	0	0	0	0
IS_SPEED_AIR_ABNORMAL		831	0	0	0	0	0	0	0	0
IS_HEIGHT_ABNORMAL		831	0	0	0	0	0	0	0	0
IS_DISTANCE_ABNORMAL		831	0	0	0	0	0	0	0	0
IS_LANDING_ABNORMAL		831	0	0	0	0	0	0	0	0
IS_DURATION_MISSING		831	0	0	0	0	1.0000000	0	0	0

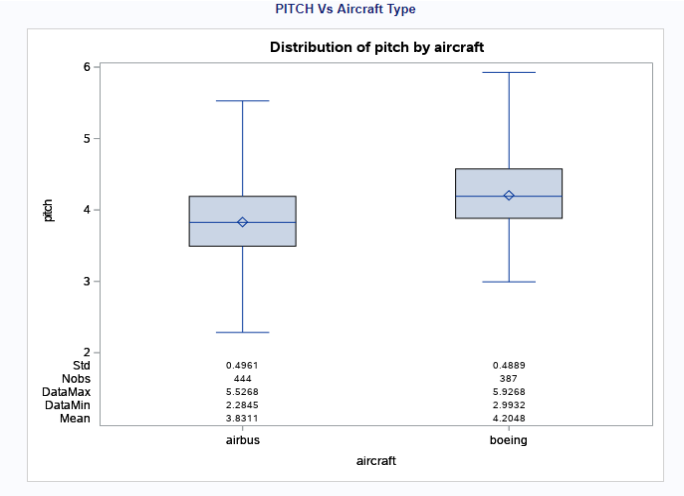
5. **Checking for distribution of parameter against Aircraft type**

```
TITLE 'Landing Distance Vs Aircraft Type';
PROC BOXPLOT DATA=BANA6043.FLIGHTS_COMBINED_NORMAL;
PLOT DISTANCE*AIRCRAFT;
INSETGROUP MEAN MIN MAX N STDDEV /
      HEADER = 'EXTREMES BY AIRCRAFT'
      POS=TM;
RUN;

PROC UNIVARIATE DATA=BANA6043.FLIGHTS_COMBINED_NORMAL NOPRINT;

CLASS AIRCRAFT;
HISTOGRAM DISTANCE;
RUN;
```

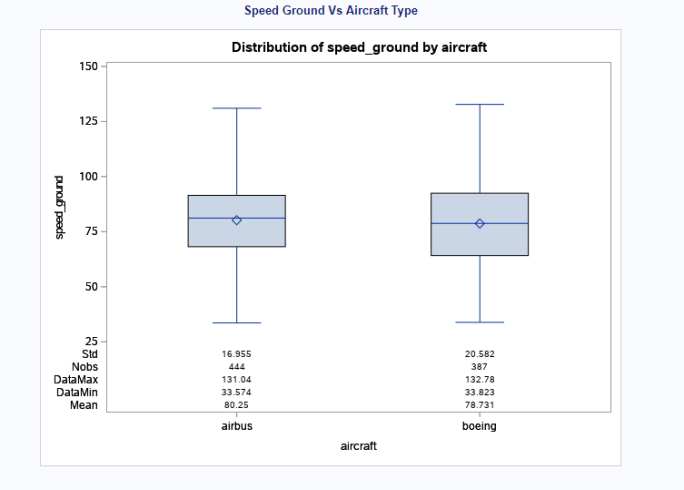




Histogram depicts that distribution is normal for Airbus while slightly skewed for Boeing. Boeing has significantly higher mean and std for distance in comparison to airbus. Meaning that it requires more landing distance in comparison to Airbus. Not enough data to support the evident data but a general intuition says that the stark difference could be possibly because of the difference in size of these two types of Aircrafts.



Boxplot here depicts that in average the pitch for Boeing is 10 % higher than that of Airbus.



Boxplot here depicts that variability of speed ground for Boeing is 21 % higher than that of Airbus.

Boxplot and histograms for height and no_pasg at an aircraft level had no significance difference.

Chapter 02: Modeling and Validation

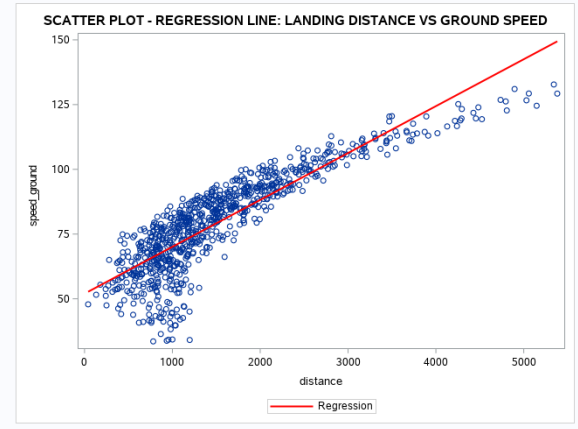
Goal:

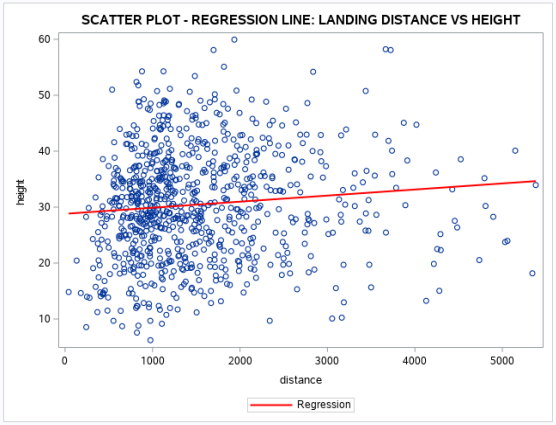
The goal of this chapter is to establish a relationship between parameters and landing distance. So, as it can be used to predict and help avoid of landing overrun in future.

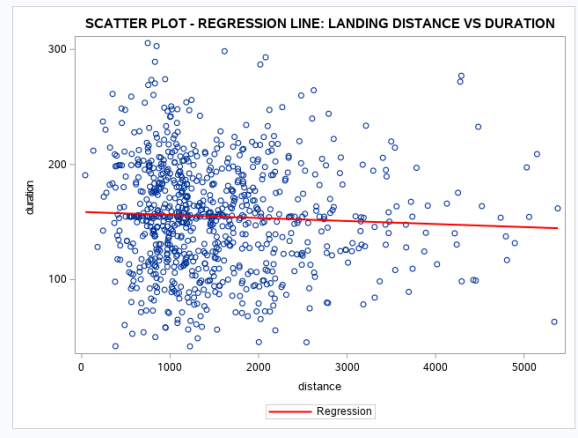
Steps Followed:

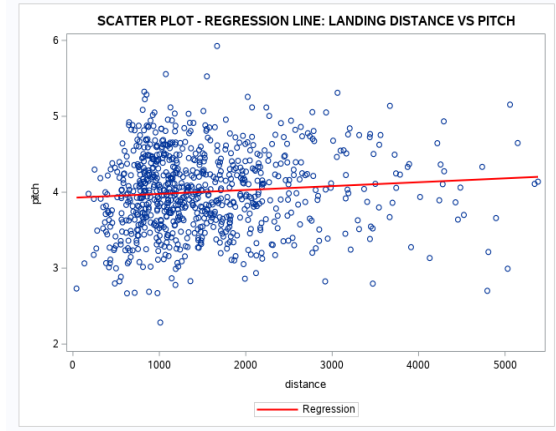
1. **Calculating Correlation between independent variable (parameters) and dependent variable (landing distance).**
   Using XY plots and correlation values to establish the relationship strength.

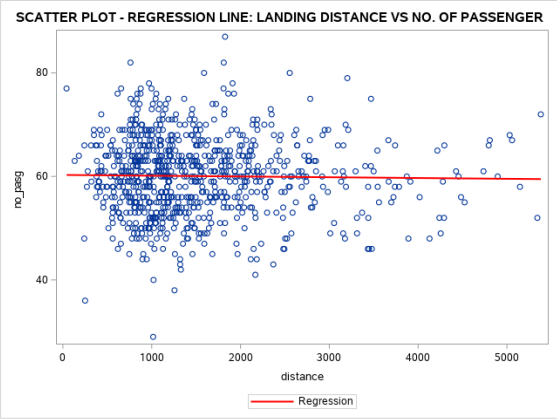
```
/* Plotting Landing Distance Vs. The Ground Speed*/
PROC SGPLOT DATA=BANA6043.FLIGHTS_COMBINED_NORMAL ;
	REG X=DISTANCE Y=SPEED_GROUND / LINEATTRS=(COLOR=RED
THICKNESS=2);
	TITLE "SCATTER PLOT - REGRESSION LINE: LANDING DISTANCE VS GROUND
SPEED";
RUN;
```











Looking at the graphs above, ground speed have a strong positive correlation with landing distance while pitch and height have slightly positive relation with landing distance. On the other hand, no. of passengers and height have either no or slightly negative relation with landing distance.

2. **Calculating Correlation value to determine the strength of relationship between independent and dependent variable. Also this could help in identifying any relationship among independent variables.**

```
*CALCULATING THE CORRELATIONS;
PROC CORR DATA=BANA6043.FLIGHTS_COMBINED_NORMAL;
      VAR DISTANCE SPEED_GROUND DURATION NO_PASG HEIGHT PITCH;
RUN;
```

Pearson Correlation Coefficients, N = 831
Prob > |r| under H0: Rho=0

	distance	speed_ground	duration	no_pasg	height	pitch
distance	1.00000	0.86624	-0.05027	-0.01776	0.09941	0.08703
distance		<.0001	0.1477	0.6093	0.0041	0.0121
speed_ground	0.86624	1.00000	-0.04788	-0.00013	-0.05761	-0.03912
speed_ground	<.0001		0.1679	0.9969	0.0970	0.2599
duration	-0.05027	-0.04788	1.00000	-0.03544	0.01073	-0.04496
duration	0.1477	0.1679		0.3075	0.7574	0.1954
no_pasg	-0.01776	-0.00013	-0.03544	1.00000	0.04699	-0.01793
no_pasg	0.6093	0.9969	0.3075		0.1760	0.6057
height	0.09941	-0.05761	0.01073	0.04699	1.00000	0.02298
height	0.0041	0.0970	0.7574	0.1760		0.5082
pitch	0.08703	-0.03912	-0.04496	-0.01793	0.02298	1.00000
pitch	0.0121	0.2599	0.1954	0.6057	0.5082	

The above matrix depicts the Pearson Correlation Coefficients among parameters and with landing distance. In accordance to our results in step 1; height and pitch does have a positive correlation with landing distance. While the p-value for any other pair is > 0.05. Hence, at 95 % confidence, we could say that there is no correlation among any other parameters.

3. **Fitting into model.**

   After we have establish the correlation between parameters and landing distance, we would want to come to an equation which best fits these data point and in future could predict the landing distance given the input parameters.

   Since, we are also interested in understanding the difference in model because of the aircraft type, we would be coming up with three different relation equations:
   1) Overall equation fitting the data 2) Equation corresponding to Airbus 3) for Boeing

```
* Performing the regression analysis on complete data set;

PROC REG DATA=BANA6043.FLIGHTS_COMBINED_NORMAL;
      MODEL DISTANCE=DURATION NO_PASG SPEED_GROUND HEIGHT PITCH;
      OUTPUT OUT=temp_reg r=resid;
      RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	524928360	104985672	610.33	<.0001
Error	825	141911969	172015		
Corrected Total	830	666840329			

Root MSE	414.74632	R-Square	0.7872
Dependent Mean	1522.48287	Adj R-Sq	0.7859
Coeff Var	27.24144		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2858.64768	188.37914	-15.17	<.0001
duration	duration	1	-0.10127	0.30806	-0.33	0.7424
no_pasg	no_pasg	1	-2.71593	1.92548	-1.41	0.1588
speed_ground	speed_ground	1	42.05881	0.77114	54.54	<.0001
height	height	1	13.60170	1.47578	9.22	<.0001
pitch	pitch	1	199.78261	27.40248	7.29	<.0001

The model used 831 observations to present the equation which could be denoted as:

Distance= -2858.64768 + 42.05881*speed_ground + 13.60170*height + 199.78261*pitch

P-value corresponding to parameters no. of passengers and duration > 0.05. Hence they are not part of the linear regression here.

From ANOVA table, p-value < 0.05 and F value is quite high, indicating that landing distance could be well predicted by the input data provided for parameters here.
Parameter estimators table used here to determine the value of coefficients and noise.

```
PROC UNIVARIATE DATA=temp_reg NORMAL;

      VAR RESID;
      QQPLOT RESID /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
RUN;
```

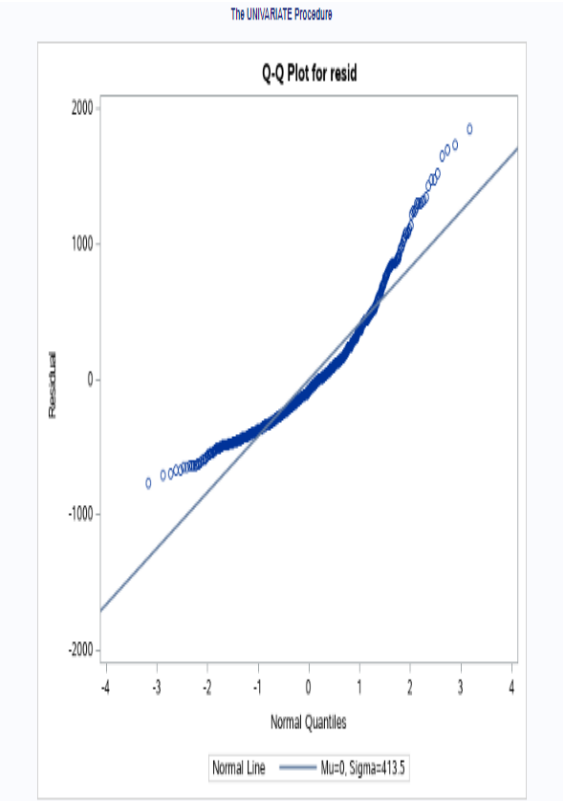
The UNIVARIATE Procedure
Variable: resid (Residual)

Moments			
N	831	Sum Weights	831
Mean	0	Sum Observations	0
Std Deviation	413.495195	Variance	170978.276
Skewness	1.31502135	Kurtosis	2.21339647
Uncorrected SS	141911969	Corrected SS	141911969
Coeff Variation	.	Std Error Mean	14.3439894

Basic Statistical Measures			
Location		Variability	
Mean	0.0000	Std Deviation	413.49519
Median	-78.8545	Variance	170978
Mode	.	Range	2611
		Interquartile Range	460.37979

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-63.5	Pr >= |M|	<.0001
Signed Rank	S	-24471	Pr >= |S|	0.0004

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.909968	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.098573	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.93149	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	18.30492	Pr > A-Sq	<0.0050



The test for normality indicated that the residual is not normally depicted since the p-value is less than 0.05.

Similarly, running the regression model only for Airbus and Boeing observations results in the below equations:

```
* Performing the regression analysis on Airbus data set;

DATA BANA6043.FLIGHTS_COMBINED_NORMAL_AIRBUS;
SET BANA6043.FLIGHTS_COMBINED_NORMAL (where=(aircraft='airbus'));
RUN;

PROC REG DATA=BANA6043.FLIGHTS_COMBINED_NORMAL_AIRBUS;
      MODEL DISTANCE=DURATION NO_PASG SPEED_GROUND HEIGHT PITCH;
      OUTPUT OUT=temp_reg r=resid;
      RUN;

PROC UNIVARIATE DATA=temp_reg NORMAL;
      VAR RESID;
      QQPLOT RESID /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

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	5	237637306	47527461	517.96	<.0001
Error	438	40190299	91759		
Corrected Total	443	277827605			

Root MSE	302.91694	R-Square	0.8553
Dependent Mean	1323.31696	Adj R-Sq	0.8537
Coeff Var	22.89073		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2793.49269	198.43149	-14.08	<.0001
duration	duration	1	-0.31047	0.31152	-1.00	0.3195
no_pasg	no_pasg	1	-1.29891	1.95328	-0.66	0.5064
speed_ground	speed_ground	1	42.52132	0.85074	49.98	<.0001
height	height	1	13.82587	1.46396	9.44	<.0001
pitch	pitch	1	106.60235	29.28711	3.64	0.0003

The model used 444 observations to present the equation which could be denoted as:

Distance= - 2793.49269 + 42.52132*speed_ground + 13.82587*height + 106.60235*pitch

P-value corresponding to parameters no. of passengers and duration > 0.05. Hence they are not part of the linear regression here.

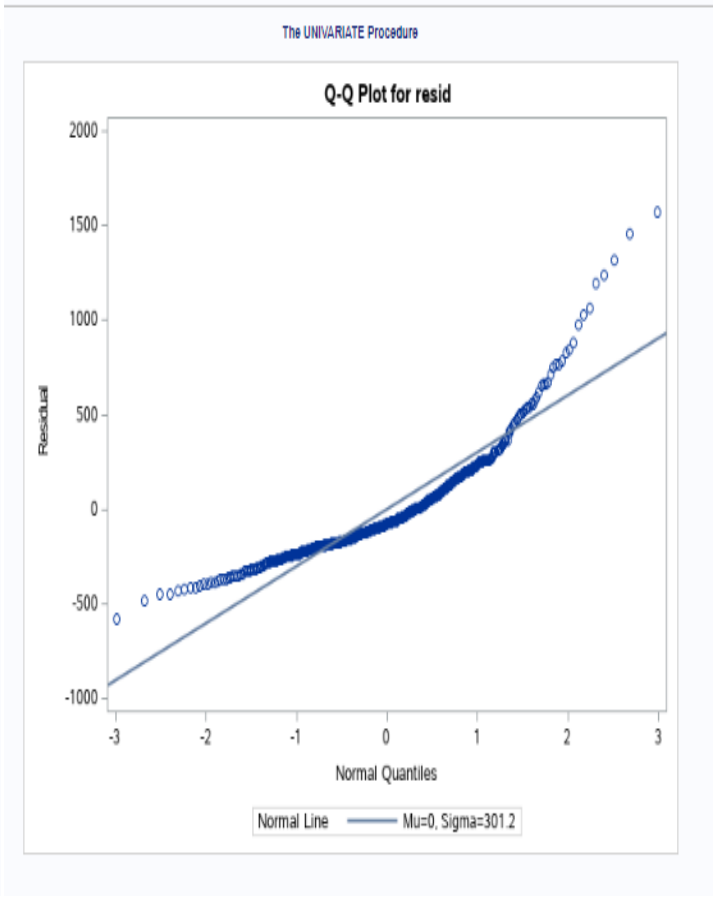
The UNIVARIATE Procedure
Variable: resid (Residual)

Moments			
N	444	Sum Weights	444
Mean	0	Sum Observations	0
Std Deviation	301.202628	Variance	90723.0232
Skewness	1.80217045	Kurtosis	4.80670913
Uncorrected SS	40190299.3	Corrected SS	40190299.3
Coeff Variation	.	Std Error Mean	14.2944441

Basic Statistical Measures			
Location		Variability	
Mean	0.0000	Std Deviation	301.20263
Median	-75.8442	Variance	90723
Mode	.	Range	2156
		Interquartile Range	312.25222

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-53	Pr >= |M|	<.0001
Signed Rank	S	-8617	Pr >= |S|	0.0014

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.862469	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.134063	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.588933	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	14.94942	Pr > A-Sq	<0.0050



The test for normality indicated that the residual is not normally depicted since the p-value is less than 0.05.

```
* Performing the regression analysis on Boeing data set;

DATA BANA6043.FLIGHTS_COMBINED_NORMAL_BOEING;
SET BANA6043.FLIGHTS_COMBINED_NORMAL (where=(aircraft='boeing'));
RUN;

PROC REG DATA=BANA6043.FLIGHTS_COMBINED_NORMAL_ BOEING;
      MODEL DISTANCE=DURATION NO_PASG SPEED_GROUND HEIGHT PITCH;
      OUTPUT OUT=temp_reg r=resid;
      RUN;

PROC UNIVARIATE DATA=temp_reg NORMAL;
      VAR RESID;
      QQPLOT RESID /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

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		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1796.98903	256.90886	-6.99	<.0001
duration	duration	1	0.45775	0.42224	1.08	0.2790
no_pasg	no_pasg	1	-1.95240	2.66558	-0.73	0.4643
speed_ground	speed_ground	1	42.28107	0.97564	43.34	<.0001
height	height	1	14.23727	2.06972	6.88	<.0001
pitch	pitch	1	-39.31818	41.16737	-0.96	0.3401

The model used 387 observations to present the equation which could be denoted as:

Distance= - 1796.98903 + 42.28107*speed_ground + 14.23727*height

P-value corresponding to parameters pitch, no. of passengers and duration > 0.05. Hence they are not part of the linear regression here.

The UNIVARIATE Procedure
Variable: resid (Residual)

Moments

N	387	Sum Weights	387
Mean	0	Sum Observations	0
Std Deviation	389.848044	Variance	151981.497
Skewness	1.45106982	Kurtosis	1.94302713
Uncorrected SS	58664857.9	Corrected SS	58664857.9
Coeff Variation	.	Std Error Mean	19.8170898

Basic Statistical Measures

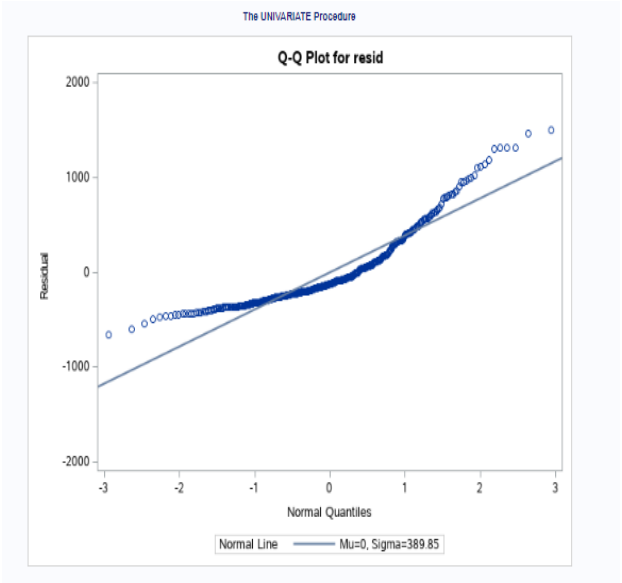
Location		Variability	
Mean	0.000	Std Deviation	389.84804
Median	-123.308	Variance	151981
Mode	.	Range	2163
		Interquartile Range	395.78107

Tests for Location: Mu0=0

Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-56.5	Pr >= |M|	<.0001
Signed Rank	S	-6730	Pr >= |S|	0.0021

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.86457	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.160986	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.997788	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	16.79602	Pr > A-Sq	<0.0050



The test for normality indicated that the residual is not normally depicted since the p-value is less than 0.05.

A stark difference for Boeing Aircraft is that here landing distance is not influenced by the pitch unlike for Airbus.

Final Answers:

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

   831 observations were used to fit the final model. Out of 950 observations, 100 were ruled out because of duplicity while another 19 didn't fit into the criteria mentioned for a normal landing.

   While these 831 observations had 50 observations where duration was missing, but that has been replaced with the mean value for other observations. Since, removing those would have led to a loss of information and duration (intuitively also suggests that should not have an impact on landing distance which was later proved with the correlation matrix).

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

   Model Equation for overall data:
   Distance= -2858.64768 + 42.05881*speed_ground + 13.60170*height + 199.78261*pitch

   Ground Speed, height and pitch are the factors that impact the landing distance of a flight. The equation could be interpreted as:

   - **For Every unit increase (miles/hour) in ground speed, landing distance increases by ~ 42 ft.**
   - **For Every unit increase (meters) in height, landing distance increases by ~ 13.6 ft.**
   - **For Every unit increase (degrees) in pitch, landing distance increases by ~ 200 ft.**

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

   Yes. Pitch which on average is 10 % higher for Boeing than Airbus, is not a deciding factor while evaluating landing distance incase of Boeing. While height and ground speed affect the landing distance for both the aircrafts in same proportion.

Airbus Distance= - 2793.49269 + 42.52132*speed_ground + 13.82587*height + 106.60235*pitch

Boeing Distance= - 1796.98903 + 42.28107*speed_ground + 14.23727*height