Executive Summary –

In this project we are trying to understand the impact of various factors – aircraft, duration, no_pasg, speed_ground, speed_air, height, pitch and duration on the landing distance of an aircraft. We started with **950 rows in the master dataset** of 2 aircraft makes Airbus and Boeing, and were left with **850 rows post duplicates** removal and finally had **831 rows post outliers, abnormalities removal.** We observed that **75% of the values are missing in the speed_air column** which we haven't replaced with mean as that would reduce the variability of the data and predicting 75% of the data using 25% of the data points would also be incorrect.

Mean landing distance for Boeing is higher than the mean landing distance of Airbus. **Pitch** is also higher for Boeing as compared to Airbus. Pitch maybe driving this difference in the landing distance between Boeing and Airbus or there may be some other variable not in the data which can be driving it. Conducting **t – test,** we conclude that there is **statistical evidence** to indicate that the **mean distance for Airbus is significantly different from mean distance for Boeing.**

Speed_ground and Distance were found to be highly positively correlated with a **correlation coefficient** of **0.87**. **Speed Air and Distance** are also highly correlated with a **correlation coefficient of 0.94. Speed_ground and speed_air** are highly correlated (multicollinearity) with a correlation coefficient of 0.98. It would make sense to **remove the speed_air** variable otherwise we would get incorrect estimates during the modelling phase since it is highly correlated with speed_ground and it has only 25% of the values

From the XY plots of distance with various independent variables, variables **height, pitch, no_pasg, duration do not** show any **significant relationship** with distance. Relationship between distance and speed_ground is not exactly linear, hence we modelled on 568 data points (speed > 70) basically the portion of the plot which is linear.

Speed_ground, height and pitch turned out to be **significant** with parameter estimates of 63.25, 12.83 and 182.25 respectively. **All parameter estimates** are **positive** indicating landing distance increases with increase in any of these factors. These parameters explain the **change in distance with change in 1 unit of these factors.** For eg: Landing distance would change by 63.25 feet with 1 miles/hour increase in speed_ground. The observed model R square was .**8987** which basically means **89.87% of the variability in landing distance can be explained by these 3 variables.**

Modelling for both types of aircraft we observe that the variable **pitch** is coming to be **significant** with a positive relationship with distance for **Airbus**, whereas it is coming to be **insignificant** for **Boeing**. In the overall model pitch came to be significant. We should ideally be using an **interaction** variable for pitch in the next phase of the project.

Next steps -

Due to the non linear relationship between speed_ground and landing distance, other methods like **transformation (using a quadratic term for speed ground**) can be tested too. The model should be tested with **test data** before providing it to the clients. **Interaction variable for pitch** should be introduced as this it is showing different behaviour for 2 makes of the aircraft. **Other techniques** which could explain relationship between distance and the independent variables should be explored as well.

1. Data Understanding, Cleaning and exploration -

a) Data Load

Specific Goal:

First step should be to import the 2 files into SAS.

```
/* Importing data */

DATA FAA_1;

INFILE '/home/u40911506/sasuser.v94/FAA1.csv' dlm=',' firstobs=2 DSD;

INPUT aircraft $ duration no_pasg speed_ground speed_air height pitch distance;

RUN;

PROC PRINT data=FAA_1;

RUN;


DATA FAA_2;

INFILE '/home/u40911506/sasuser.v94/FAA2.csv' dlm=',' FIRSTOBS=2 DSD;

INPUT aircraft $ no_pasg speed_ground speed_air height pitch distance;

RUN;

PROC PRINT data=FAA_2;

RUN;
```

Output:

We observe that both datasets have been uploaded with names FAA_1 and FAA_2.

b) File Content exploration:

Specific Goal:

Now let us look at the contents of the datasets we loaded into SAS.

Code:

```
/*Identify the contents of DS1*/

proc contents data = WORK.FAA_1;

RUN;


/*Identify the contents of DS2*/

proc contents data = WORK.FAA_2;

RUN;
```

Output:

FAA_1:

1. There are 800 observations in this dataset.
2. There are 8 columns in the dataset – 1 character column (aircraft make) and 7 numeric columns.

FAA_2:

1. There are 150 observations in this dataset.
2. There are 7 columns in the dataset – they are the same columns as FAA_1 dataset, with duration column not present in this dataset.
3. These columns might just be a repetition of the FAA_1 dataset hence we should check for duplicates post appending the 2 datasets.

c) Data Append:

```
/* Appending data */

data FAA_combined;

set FAA_1 FAA_2;

run;

PROC PRINT data=FAA_combined;

RUN
```

Finding/Observations:

We notice that there are in total 950 rows in the combined dataset out of which 150 rows from the FAA_2 data that do not contain duration column.

d) Removal of empty rows:

Post append we notice that there are 50 empty rows in the dataset. Let's remove them

Findings/Observations:

We are left with (1000 – 50) = 950 rows post removal of the empty rows.

e) Duplicates Identification and Removal:

Specific Goal:

As we have combined the 2 datasets now, let us sort the data and remove any duplicate rows if there are any in the combined dataset

Code:

```
proc sort data=FAA_combined nodupkey

out = FAA_combined_2;

by _all_;

run;
```

```
proc print data=FAA_combined_2;

run;
```

Findings/Observations:

We observe there are 951 rows excluding 1 blank row at the top which we can ignore, total rows= 950 rows in the dataset post duplicate removal. Hence we can say there are no duplicates at an overall (all columns) level.

Just to be double sure, we can remove duplicates at a few column levels and check the results to check if FAA_2 has some values same as the FAA_1 dataset. Because duration column is null at the overall dataset would never remove duplicates at the overall level if there were common values between FAA_1 and FAA_2.

Let us remove duplicates at aircraft, no_pasg and pitch level

Code:

```
proc sort data=FAA_combined nodupkey

out = FAA_combined_2;

by aircraft no_pasg pitch;

run;

proc print data=FAA_combined_2;

run;
```

We observe there are 850 observations in the dataset now. This may be due to the fact that in the 2nd dataset FAA_2 there are 100 rows which have the same values as the first dataset except the duration column, thus these duplicates don't show up when we are removing duplicates at all the levels. 50 rows are different. This can be said because FAA_1 has 800 unique rows so the duplicates would come only from FAA_2 dataset. Let us proceed with 850 observations dataset.

f) Reporting the missing values and missing values percentage:

Code:

```
proc means data = FAA_combined_3 stackods N MIN MAX MEAN STD NMISS RANGE MEDIAN;

ods output summary = FAA_combined_summary;

run;

proc print data=FAA_combined_summary;

run;

/* Reporting the percentage of missing values to see which variable has a higher percentage of missing values*/

data percent_missing;

set FAA_combined_summary;
```

```
if NMiss > 0 then percentage_missing = (NMiss/(NMiss + N))*100;

else percentage_missing = 0;

run;

proc print data=percent_missing;

run;
```

Output:

The MEANS Procedure

Variable	N	Minimum	Maximum	Mean	Std Dev	N Miss	Range	Median
duration	800	14.7642072	305.6217107	154.0065385	49.2592338	51	290.8575036	153.9480976
no_pasg	850	29.0000000	87.0000000	60.1035294	7.4931370	1	58.0000000	60.0000000
speed_ground	850	27.7357153	141.2186354	79.4523229	19.0594903	1	113.4829201	79.6428041
speed_air	208	90.0028586	141.7249357	103.7977237	10.2590370	643	51.7220771	101.1473493
height	850	-3.5462524	59.9459639	30.1442223	10.2877268	1	63.4922163	30.0931324
pitch	850	2.2844801	5.9267842	4.0093577	0.5288298	1	3.6423041	4.0082875
distance	850	34.0807833	6533.05	1526.02	928.5600816	1	6498.97	1258.09

Obs	Variable	N	Min	Max	Mean	StdDev	NMiss	Range	Median	percentage_missing
1	duration	800	14.764207	305.621711	154.006538	49.259234	51	290.857504	153.948098	5.9929
2	no_pasg	850	29.000000	87.000000	60.103529	7.493137	1	58.000000	60.000000	0.1175
3	speed_ground	850	27.735715	141.218635	79.452323	19.059490	1	113.482920	79.642804	0.1175
4	speed_air	208	90.002859	141.724936	103.797724	10.259037	643	51.722077	101.147349	75.5582
5	height	850	-3.546252	59.945964	30.144222	10.287727	1	63.492216	30.093132	0.1175
6	pitch	850	2.284480	5.926784	4.009358	0.528830	1	3.642304	4.008288	0.1175
7	distance	850	34.080783	6533.047651	1526.023095	928.560082	1	6498.966868	1258.091506	0.1175

Findings/Observations:

1. Number of missing values are minimal in all the columns except duration and speed_air. 75% of the entries of speed_air are missing. We could go ahead with replacing the missing values with the mean of rest of the values of speed_air but that would not be accurate since 75% of the data would have the same value, hence this column would not provide us with any additional insight. Hence, we would go ahead with not replacing them. This would also reduce the overall variability of this column.
2. For duration, 6% of the data is missing which Is not a very high number, we can go ahead with it for now.
3. Duration's range is extremely wide (14.76, 305.62) – this column might contain outliers, let us check that during outlier detection stage.
4. Minimum value of height is -3.54. Since height is the height of the aircraft when it is passing over the threshold of the runaway, this value cannot be -ve. May be this is a wrong data entry and would be removed in the data cleaning step.

g) Outlier Detection and Removal:

Specific Goal:

In this stage, we would be deleting the rows which are beyond the prescribed business rules.

We would be first counting 3 things – number of missing values, number of data points not in the business rules and the correct data points. We would be encoding number of missing values as “Variable is missing”, number of data points not in the business rules as “DEL” and other data points as “VALID”. Post this we would be deleting data based on the below conditions. (if values not satisfying the business rules are minimal).

a. Duration is less than 40 mins and not equal to blank.
b. Speed ground is less than 30 or greater than 140 and not equal to blank.
c. Speed air is less than 30 or greater than 140 and not equal to blank.
d. Height = landing height less than 6 metres at the threshold of the runway, and not equal to blank.
e. Landing Distance < 6000 since the length of the airport is typically less than 6000 feet and not equal to blank.

We are leaving the missing entries as it is, as they are anyway not very high in number, except the speed_air column which we would be looking into in the upcoming parts

Code:

```
/* Flagging and Counting the number of values where 1. Variable is missing 2. Which are outlying
based on the given business rules 3. Valid */

DATA quality_check;
SET FAA_combined_3;
IF duration =. then qual_duration = "Variable is missing";
else IF duration < 40 and duration <> . THEN qual_duration="DEL";
ELSE qual_duration="Valid";

IF speed_ground =. then qual_speed_ground = "Variable is missing";
else if (speed_ground < 30 or speed_ground > 140) THEN qual_speed_ground="DEL";
ELSE qual_speed_ground="Valid";

if speed_air =. then qual_speed_air = "Variable is missing";
else IF (speed_air < 30 or speed_air > 140) and speed_air <> . THEN qual_speed_air="DEL";
ELSE qual_speed_air ="Valid";

if height =. then qual_height = "Variable is missing";
else IF (height < 6) and height <> . then qual_height = "DEL";
ELSE qual_height ="Valid";

if distance =. then qual_distance = "Variable is missing";
else IF distance > 6000 and distance <> . THEN qual_distance ="DEL";
ELSE qual_distance ="Valid";
run;

proc print data = quality_check;
run;

proc freq data =quality_check;
Tables qual_duration qual_speed_ground qual_speed_air qual_height qual_distance ;
```

```
run;

data outlying_1;
set quality_check;
if qual_duration = 'DEL' then outlying_duration = 1;
else outlying_duration = 0;
if qual_speed_ground = 'DEL' then outlying_speed_ground = 1;
else outlying_speed_ground = 0;
if qual_speed_air = 'DEL' then outlying_speed_air = 1;
else outlying_speed_air = 0;
if qual_height = 'DEL' then outlying_height = 1;
else outlying_height = 0;
if qual_distance = 'DEL' then outlying_distance = 1;
else outlying_distance = 0;
run;

proc print data=outlying_1;
run;

PROC SQL;
CREATE TABLE outlying_count AS
SELECT
sum(outlying_duration) AS count_outlying_duration,
sum(outlying_speed_ground) AS count_outlying_speed_ground,
sum(outlying_speed_air) AS count_outlying_speed_air,
sum(outlying_height) AS count_outlying_height,
sum(outlying_distance) AS count_outlying_distance
FROM outlying_1
;
QUIT;
proc print data=outlying_count;
run;
```

Output:

Obs	count_outlying_duration	count_outlying_speed_ground	count_outlying_speed_air	count_outlying_height	count_outlying_distance
1	5	3	1	10	2

Findings/Observations:

1. In total, there are **21 values** under DEL (which are beyond the prescribed business rules) hence we would be deleting them in the next step.

Let us remove the values under the DEL flag -

```
data FAA_combined_4;
```

```
set FAA_combined_3;

if duration <>  ' ' and duration < 40 then delete;

if Distance <> ' ' and distance > 6000 then delete;

if speed_ground <> ' ' and (speed_ground < 30 or speed_ground > 140) then delete;

if speed_air <> ' ' and (speed_air < 30 or speed_air > 140) then delete;

if height <> ' ' and (height < 6) then delete;

run;
```

Output:

817	boeing	236.193	73	52.360	.	44.1211	4.49709	1078.10
818	boeing	168.230	74	86.853	.	16.8945	3.83090	1725.38
819	boeing	112.317	74	79.258	.	37.1972	4.33700	1158.84
820	boeing	118.264	75	70.168	.	17.7433	4.26698	830.71
821	boeing	124.544	75	69.880	.	31.3114	4.68792	1045.03
822	boeing	79.706	75	106.746	106.733	18.3462	4.80740	2785.86
823	boeing	147.032	76	63.598	.	36.4890	4.49177	1051.94
824	boeing	219.721	76	88.103	.	42.0855	4.65401	1927.05
825	boeing	130.950	76	44.733	.	32.7830	4.86188	874.80
826	boeing	130.169	77	55.087	.	38.0328	4.09712	998.10
827	boeing	172.560	77	82.297	.	44.7587	4.22931	1809.27
828	boeing	107.113	78	86.808	.	25.4770	4.41422	1910.88
829	boeing	228.177	78	61.220	.	21.7723	4.59553	970.05
830	boeing	128.938	79	106.934	108.427	30.4577	4.84215	3203.32
831	boeing	161.826	80	82.509	.	36.6802	4.68531	1590.37
832	boeing	194.467	82	40.815	.	22.6184	4.87660	761.49

Finding / Observations:

1. We are left with 832 rows in the dataset post applying the above business rules out of which there is 1 row with all blank values so 831 rows in total.

h) Distribution of the variables:

Specific Goal:

Investigating the variables by looking at the **summary statistics** and the **distribution** of each of the variables in the **clean** dataset, to find any interesting patterns that can be used while modelling

Code:

```
proc means data=FAA_combined_4 stackods n min max mean std nmiss range median q1 q3 qrange;

ods output summary=summary_stats;

run;
```

Output:

The MEANS Procedure

Variable	N	Minimum	Maximum	Mean	Std Dev	N Miss	Range	Median	Lower Quartile	Upper Quartile	Quartile Range
duration	781	41.949369	305.621711	154.775719	48.349924	51	263.672341	154.284551	119.631458	189.662943	70.031485
no_pasg	831	29.000000	87.000000	60.055355	7.491317	1	58.000000	60.000000	55.000000	65.000000	10.000000
speed_ground	831	33.574104	132.784677	79.542700	18.735675	1	99.210573	79.793960	66.192530	91.949608	25.757077
speed_air	203	90.002859	132.911465	103.485035	9.736277	629	42.908606	101.118924	96.196461	109.382301	13.185840
height	831	6.227518	59.945964	30.457870	9.784811	1	53.718446	30.167084	23.529869	37.014302	13.484433
pitch	831	2.284480	5.926784	4.005161	0.526569	1	3.642304	4.001038	3.640398	4.371072	0.730674
distance	831	41.722313	5381.958862	1522.482873	896.338152	1	5340.236549	1262.153891	892.983974	1937.256256	1044.272282

Findings/Observations:

1. The duration column's range is extremely wide from minimum value of 41.94 to maximum value of 305.62. One hypothesis can be that flights with lower duration have lower distance. We will check this when do a correlation plot in the further sections.
2. We observe that the mean and median of each of the variables are very close to each other. This can be an indication of no skewness in the data. We would investigate this in the distribution plots below.

2) Descriptive Study:

a) Difference in values across the make of the aircraft (Boeing or Airbus)

Specific Goal:

We want to understand if there are any key difference between the 2 variables in terms of the distance variable and if there is it due to the difference from 1 predictor variable. For eg: if there is a difference between the mean distance between boeing and airbus and there is a difference between the mean speed_ground but no difference between the other variables, we can get an idea that speed_ground may be driving this difference. Let us check this hypothesis:

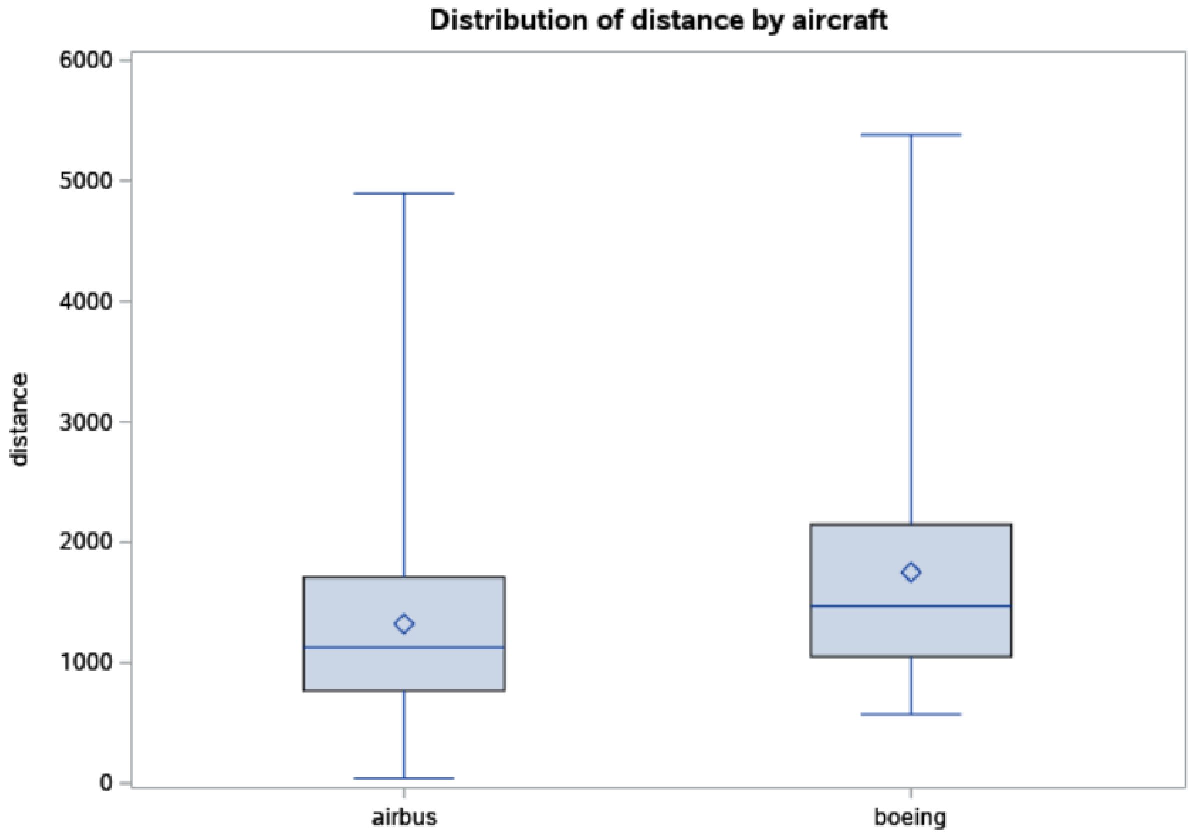
Code:

Using SAS Macro

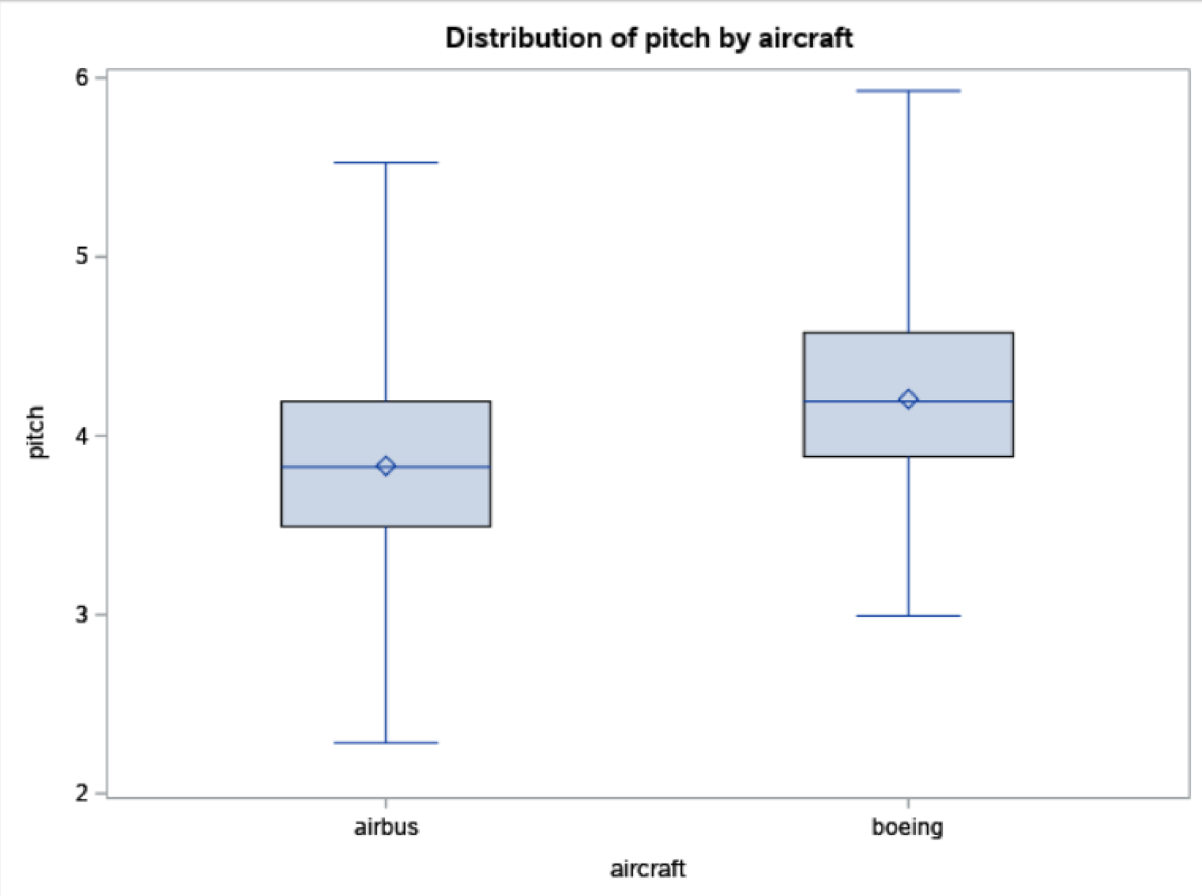
```
%macro box(variable, aircraft);

proc boxplot data=FAA_combined_4;

plot &variable * aircraft;

title "Summary statistics for Aircraft vs &variable";

run;

%mend box;

%box(distance,aircraft);

%box(pitch,aircraft);
```

Output and Findings/Observations:

We observe a large difference between the mean distance for boeing (1750.98) vs airbus (1323.31).



Observing the boxplots of the other variables we find that pitch has a similar pattern with boeing's pitch > airbus's pitch.



We can thus conclude that difference in pitch between the 2 makes can be driving the difference in distance between the 2 makes or there could be another factor which is not included in the data given to us. Let us validate this difference using a t – test.

Code:

```
proc ttest data=FAA_combined_4;

class aircraft;

var distance;

title T-Test to compare means of distance between Airbus and Boeing;

run;
```

Output:

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

Findings/Insights:

Fooled F shows us that the variances of the 2 aircraft types are unequal hence we should be looking at the Satterthwaite section which shows a p value of less than the significance level of alpha = 0.05 thus we can conclude that there is statistical evidence to indicate that the mean distance of Airbus is different from mean distance of Boeing.

b) Histogram Plot

Specific Goal:

We would like to look at the histograms of speed_ground and distance, speed_air and distance to understand their distribution better and see if we can find any more interesting patterns.

Code:

```
proc chart data=FAA_combined_4;

hbar aircraft/ type=freq;

vbar speed_ground distance/ type=pct;

run;

proc chart data=FAA_combined_4;

hbar aircraft/ type=freq;

vbar speed_air distance/ type=pct;

run;
```

Findings/Observations:

We observe that speed_ground is normally distributed, speed_air is skewed to the right and distance is also skewed to the right. This chart has not provided us with many useful insights. We would now be exploring the relationship between the independent and dependent variables using a correlation plot and XY plot.

j) Correlation Plot:

Specific Goal:

We would like to look at the correlations among the predictor variables to assess any multicollinearity, also if any variable is correlated with the distance variable using scatter plot matrix

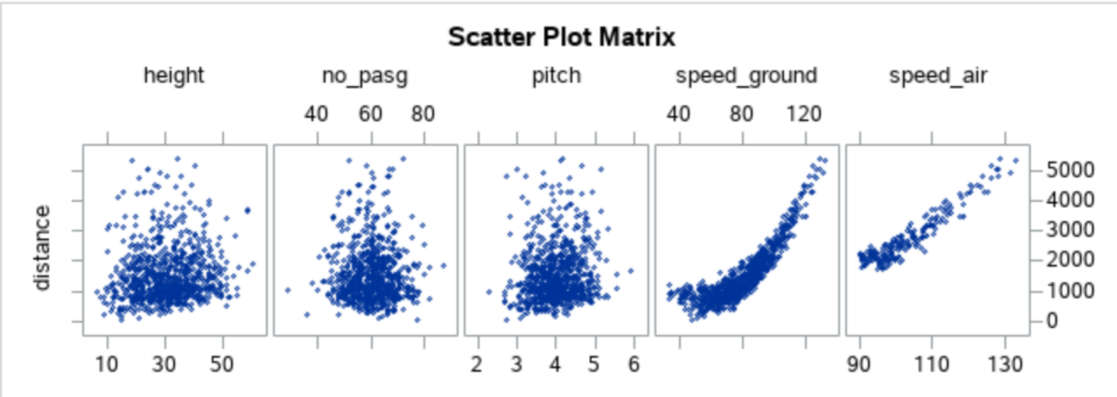
Code:

```
proc corr data = FAA_combined_4;

VAR duration no_pasg speed_ground speed_air height pitch distance;

run;

ods graphics on;
```

```
proc corr data=FAA_combined_4
plots = matrix(histogram);
var height no_pasg pitch speed_ground speed_air duration;
with distance;
title Correlation coefficients of all factors with Distance;
run;
ods graphics off;
```

Output:

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	height	no_pasg	pitch	speed_ground	speed_air	duration
distance	0.09941 0.0041 831	-0.01776 0.6093 831	0.08703 0.0121 831	0.86624 <.0001 831	0.94210 <.0001 203	-0.05138 0.1514 781



Findings/Observations:

We need to look at the variable combinations which are significant (where p < 0.001 for a 2 tailed test). For ex: correlation of 0.98794 between speed_ground and speed_air for 203 common combinations, p value being significant.

1. Speed Ground and Speed Air are highly correlated with a pearson correlation coefficient of 0.98794. It would make sense to remove the speed air variable since it has only 25% of observations, and it would give us incorrect parameter estimates.
2. From the scatter plot matrix, Speed Ground and Distance are highly correlated with a pearson correlation coefficient of 0.86624. Basically distance increases with increase in ground speed or vice versa. This can be an important factor impacting distance and we would analyse this further during the modelling phase.
3. From the scatter plot matrix, Speed air and Distance are highly correlated with a pearson correlation coefficient of 0.94210. Basically distance increases with increase in speed air or vice versa. We would still need to remove the speed air variable because as stated it is highly correlated with speed ground (multicollinearity) and it has only 25% of the entries.

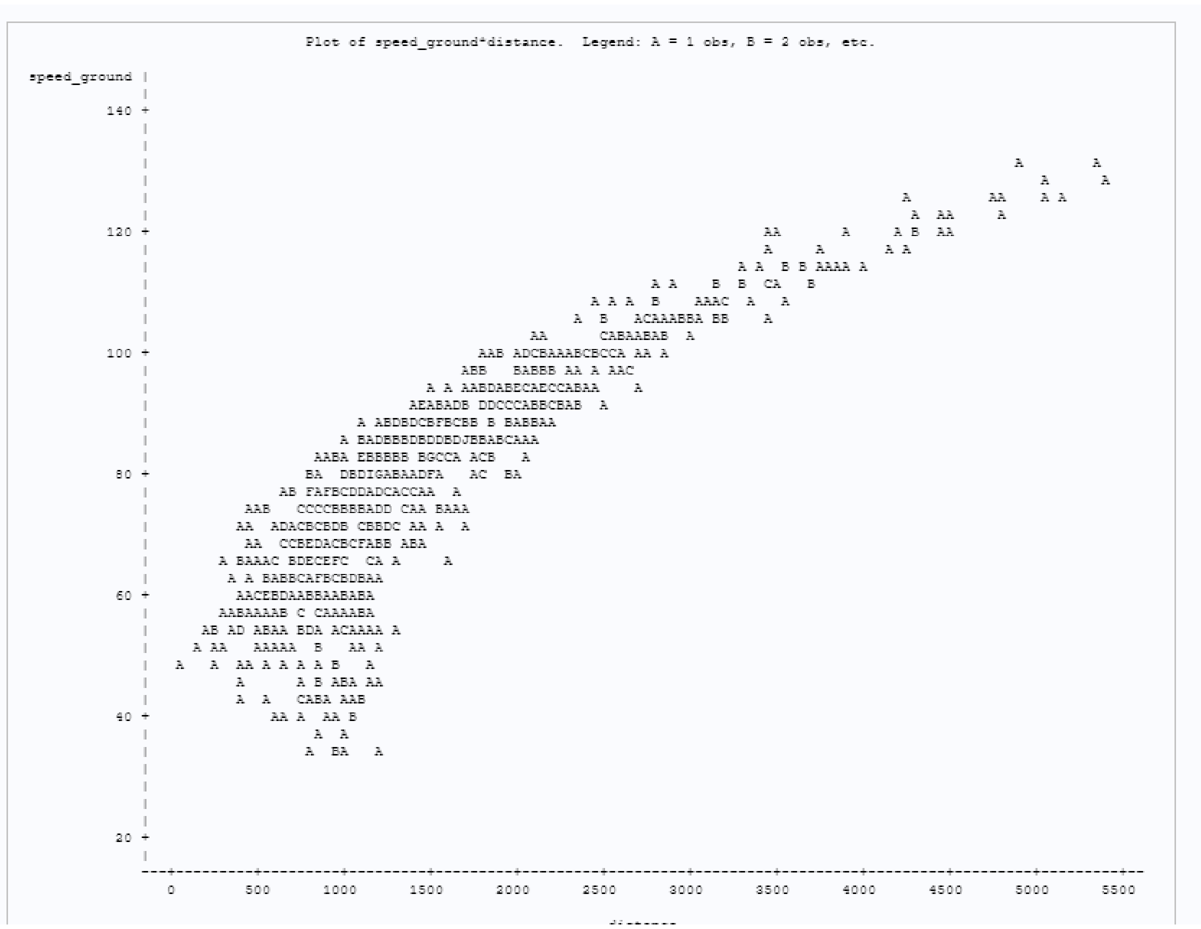
c) XY Plots of independent variables with Distance:

Specific Goal:
The only two variables that seem to show a relationship with distance are speed_ground and speed_air. Let us validate this using a XY plot.

Code:

```
proc plot data=FAA_combined_4;
plot duration*distance;
plot speed_ground*distance;
plot speed_air*distance;
plot no_pasg*distance;
plot height*distance;
plot pitch*distance;
run;
```

Output:



Findings/Observations:

1. Variables height, pitch, no_pasg, duration do not show a positive or a negative relationship with the distance variable. Mostly they would have to be removed from the model due to insignificant p values. We will check the model results for that.

2. As observed in the scatter plot matrix above, speed_air is showing an evident positive correlation with distance but as stated above because 75% of the values are missing in this column, we would not be able to use this column for any useful insights. Another reason is that it is highly correlated with speed_ground variable thus it would make sense to keep the speed_ground variable since it has more data points compared to speed_air.

3. Another interesting thing, is that the relationship between speed_ground and distance might not be exactly linear. It is looking more like an exponential relationship which would be exploring more in the model diagnostics section.

Major Findings of Descriptive Study:

1. Landing Distance of the Boeing aircraft is significantly higher than the landing distance for Airbus. We proved that the mean distance for Boeing is statistically different from the landing distance of Airbus. Pitch for Boeing aircraft can be driving this difference or some other variable which is not included in the data can be driving this difference.
2. Speed_air and speed_distance show a strong positive correlation with distance. There is multicollinearity in the data. Speed_distance is highly correlated with speed_air. It would make sense to drop the speed_air column since 75% of the data is missing for this column and it would not add any valuable insights and due to multicollinearity
3. Variables height, pitch, no_pasg, duration do not show any evident relationship with distance. We will explore this during the modelling phase see if any of these variables have a significant p value.
4. Speed ground and distance do not show an exact linear relationship. Relationship is more exponential in nature which would be exploring in the model diagnostic section – modelling only for linear data points, transformation of the variable or modelling exponential, linear separately can be options we can explore.

3 Modelling:

a. Fitting the model to all the variables:

Code:

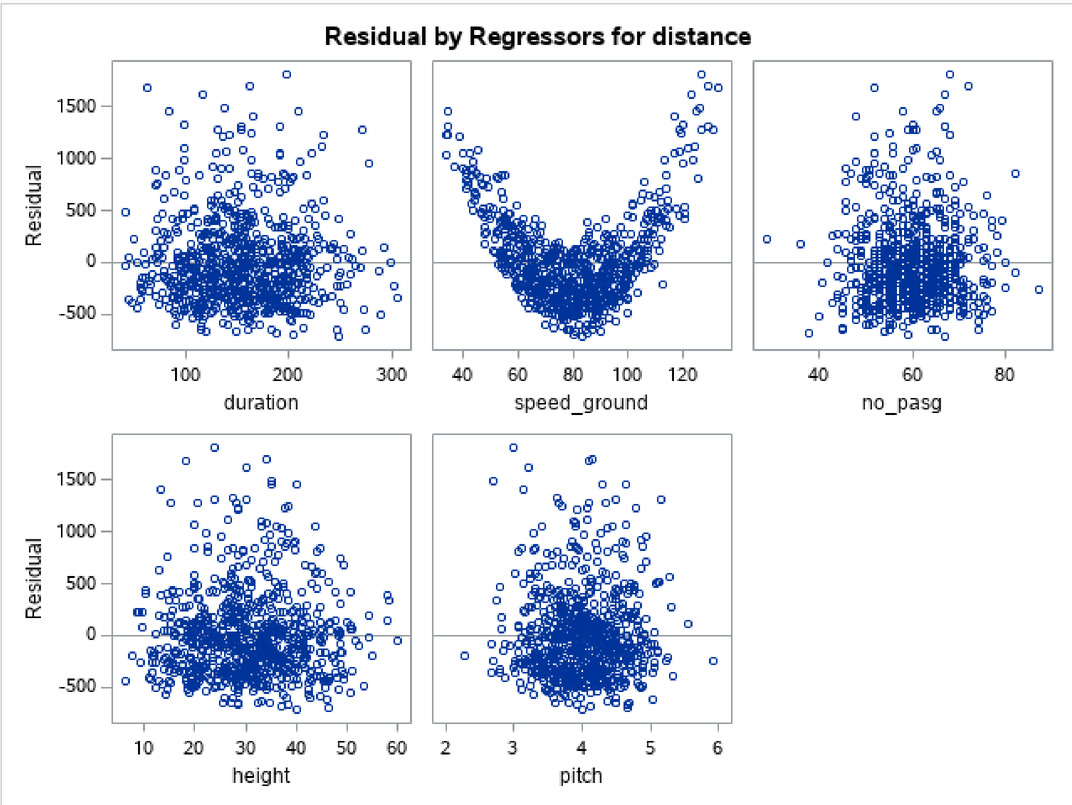
```
proc reg data=FAA_combined_4;

model distance=duration speed_ground speed_air no_pasg height pitch;

run;
```

Findings / Observations:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6249.84344	291.58960	-21.43	<.0001
duration	1	0.02246	0.36763	0.06	0.9514
speed_ground	1	-2.27284	11.56846	-0.20	0.8445
speed_air	1	82.90693	11.75796	7.05	<.0001
no_pasg	1	-3.34026	2.48183	-1.35	0.1800
height	1	12.65927	1.87055	6.77	<.0001
pitch	1	123.73575	31.32815	3.95	0.0001

1. We are seeing the impact of interaction / multicollinearity between speed_air and speed_distance here –
   a. Speed_ground shows a negative estimate here whereas we observed it to have a positive correlation with distance.
   b. Duration is showing a positive estimate here whereas we observed it to have a negative correlation with distance.

Let us remove the speed_air column and rerun the model with the rest of the variables –



These coefficients seem better as they are capturing the correct relationship between duration and speed_ground but looking at the model diagnostics we notice that there is a clear quadratic relationship between distance and speed_ground.

The present model R square is 78.7 that means the selected variables are explaining 78.7% of the variability in distance is explained by duration, speed_ground, no_pasg, height, pitch

b) Fitting the model to the linear part of the data:
To avoid the quadratic relationship, Let us now take only those values which have a linear Relationship with distance and run the model on that.

From the XY plot between speed_ground and distance we observe that speed_ground above 70 shows a linear relationship with distance. Hence, we would be taking values of speed_ground above 70 and re-runnning the model. Note we have removed speed_air due to multicollinearity.

Code:

```
data FAA_combined_linear;
set FAA_combined_4;
if speed_ground>70;
run;

proc plot data=FAA_combined_linear;
plot distance*speed_ground;
run;

proc reg data=FAA_combined_linear;
model distance=duration speed_ground no_pasg height pitch;
run;
```

Output:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	390839681	78167936	948.69	<.0001
Error	529	43587525	82396		
Corrected Total	534	434427206			

Root MSE	287.04717	R-Square	0.8997
Dependent Mean	1872.42251	Adj R-Sq	0.8987
Coeff Var	15.33026		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4838.80899	173.77493	-27.85	<.0001
duration	1	0.17862	0.25983	0.69	0.4921
speed_ground	1	63.38224	0.93722	67.63	<.0001
no_pasg	1	-1.34352	1.66553	-0.81	0.4202
height	1	12.77463	1.25613	10.17	<.0001
pitch	1	177.44093	23.83264	7.45	<.0001

1. ANOVA table tells us that the model is overall significant basically atleast one of the coefficients are non – zero as we can notice in the output.
2. The R square is 89.87 which is better than the previous R square we got so we know that this model is better at explaining variability of distance.
3. From this model – we notice that speed_ground, height and pitch have an impact on the landing distance. They have p values lesser than alpha (0.05) significance level.
4. Duration variable is insignificant and can be removed also because from a logical stand point duration of a flight should ideally not have an impact on the landing distance.
5. No_pasg should have had an impact but it is coming as insignificant for this subset of data. Maybe if we take more volume of data, this can come to be significant.

c) Fitting the model by considering only the 3 significant variables -

Now let us build a model considering only the 3 significant variables we got – speed_ground, pitch and height.

Code:

```
proc reg data=FAA_combined_linear;
model distance=speed_ground height pitch;
run;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	408990433	136330144	1671.00	<.0001
Error	564	46014383	81586		
Corrected Total	567	455004816			

Root MSE	285.63226	R-Square	0.8989
Dependent Mean	1851.03042	Adj R-Sq	0.8983
Coeff Var	15.43099		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4912.27904	126.51028	-38.83	<.0001
speed_ground	1	63.25880	0.91257	69.32	<.0001
height	1	12.83493	1.20575	10.64	<.0001
pitch	1	182.25331	22.83922	7.98	<.0001

When we are including only the significant variables we obtain similar values of parameter estimates as we obtained before, with all the variables considered to be significant.

R square is also high, we can finalize this as our final model.

Model interpretation –

Variable	Estimates	Significance	Positive/Negative relationship
speed_ground	63.25	Yes	Landing distance increases with increase in speed_ground
height	12.83	Yes	Landing distance increases with increase in height
pitch	182.25	Yes	Landing distance increases with increase in pitch

d) Modelling for the 2 aircraft makes separately

Output –

Airbus –

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4601.60210	114.27582	-40.27	<.0001
speed_ground	1	60.44637	0.87054	69.44	<.0001
height	1	12.58052	1.06493	11.81	<.0001
pitch	1	125.62500	21.02914	5.97	<.0001

Boeing –

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3741.80335	160.74169	-23.28	<.0001
speed_ground	1	63.77357	0.98425	64.79	<.0001
height	1	12.77314	1.40397	9.10	<.0001
pitch	1	-59.35957	29.08827	-2.04	0.0423

The variable pitch is coming to be significant with a positive relationship with distance for Airbus, whereas it is coming to be insignificant for Boeing. In the overall Model the variable Pitch is significant. We should ideally model using an interaction variable to identify the correct impact of this variable on distance.