FLIGHT LANDING DISTANCE ANALYSIS

NAME: Renganathan Lalgudi Venkatesan **UCID:** M12366827

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.

DATA: Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). See two Excel files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights).

SUMMARY:

I split the project into three Chapters: Data Cleaning, Analysis of relationship between variables and Regression modelling. I have summarized the methods used at each step and the results I got for all three parts below:

- Data Cleaning and Preparation: Here I tried identifying the structure of the data and got rid of abnormal observations. There were several missing values but I didn't want to get rid of those observations at the very outset of the analysis. Finally, at the end of data cleaning I had a data set with 831 observations
- Analysis of relationship between variables: I used x-y plots and correlation to understand the relationship between variables. I found that the duration variable did not have any significant correlation with the landing distance so eliminated it. Also found that the Speed variables had a high correlation.
- Regression Modelling: Developed a linear regression model with the selected variables to predict the landing distance. Found that the r-squared value was inflated due to the correlated speed variables. Identified the magnitude of correlation by using the VIF and Tolerance metrics and thus eliminated Speed_air from the analysis as it had a large correlation with the Speed_ground and also because the variable had close to 641 missing values. Modelled the landing distance again by compensating for the multi collinearity and got a decent predicting model with an r-squared close to .85.

Finally, I can see from the regression analysis that the factors like Speed_ground, No_pasg, pitch, Height and type of aircraft impact the landing distance.

Out of all the predictors, I find that Speed_gound, pitch, height and type of aircraft are the factors majorly affecting the landing distance. Factors like no_pasg has comparatively lesser impact on the landing distance.

CHAPTER: 1 DATA PREPARATION AND CLEANING

1.1. DATA PREPARATION

Since we have the data files in the xlsx files, we first create a library called "Proj1SC" on the SAS Server (SAS on Demand for Analytics). Then import the xls files as SAS datafiles.

1.1.1. IMPORTING THE DATA

CODE:

```
FILENAME REFFILE '/home/lalgudrn0/StatsComputing_Project/FAA1.xls';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLS
	OUT=PROJ1SC.FAA1;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=PROJ1SC.FAA1; RUN;
```

Output Description for FAA1 file:

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	7	$7.	$7.	aircraft
8	distance	Num	8	BEST12.		distance
2	duration	Num	8	BEST12.		duration
6	height	Num	8	BEST12.		height
3	no_pasg	Num	8	BEST8.		no_pasg
7	pitch	Num	8	BEST12.		pitch
5	speed_air	Num	8	BEST12.		speed_air
4	speed_ground	Num	8	BEST13.		speed_ground

```
FILENAME REFFILE '/home/lalgudrn0/StatsComputing_Project/FAA2.xls';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLS
	OUT=PROJ1SC.faa2;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=PROJ1SC.faa2; RUN;
```

Output Description for FAA2 file:

Alphabetic List of Variables and Attributes						
#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	7	$7.	$7.	aircraft
7	distance	Num	8	BEST12.		distance
5	height	Num	8	BEST12.		height
2	no_pasg	Num	8	BEST8.		no_pasg
6	pitch	Num	8	BEST12.		pitch
4	speed_air	Num	8	BEST12.		speed_air
3	speed_ground	Num	8	BEST13.		speed_ground

1.1.2. MERGING THE DATAFILES

The first step here is to merge the data files we have:

CODE:

```
data Proj1SC.merged;
	set Proj1SC.faa1 Proj1SC.faa2;
run;
```

We would like to see how the data looks. Mainly with respect to the missing values.

CODE:

```
proc means data = proj1sc.merged N Nmiss;
run;
```

Output:

The MEANS Procedure

Variable	Label	N	N Miss
duration	duration	800	200
no_pasg	no_pasg	950	50
speed_ground	speed_ground	950	50
speed_air	speed_air	239	761
height	height	950	50
pitch	pitch	950	50
distance	distance	950	50

As we can see, there are 1000 rows of data if we simply merge the files using concatenation technique. We want to remove the observations where there are no values at all. There are 50 such observations. We can do that by using one of the variables that is present in all rows. Using the variable "aircraft" to delete the observations, we can get this done.

CODE:

```
data proj1sc.remove_empty_rows;
	set proj1sc.merged;
	if aircraft = "" then delete;
run;

proc print data = proj1sc.remove_empty_rows;
run;
```

1.1.3. REMOVING DUPLICATES:

We find that there are a lot of duplicates in the dataset after merging. We need to remove those duplicates.

```
proc sort data = proj1sc.remove_empty_rows out= proj1sc.final nodupkey;
```

```
	by aircraft no_pasg speed_ground speed_air height pitch distance;
run;

proc print data = proj1sc.final;
run;
```

1.1.4. COMPUTING MISSING VALUES IN THE MERGED DATA SET:

CODE:

```
proc means data = proj1sc.final N NMISS;
run;
```

The MEANS Procedure

Variable	Label	N	N Miss
duration	duration	800	50
no_pasg	no_pasg	850	0
speed_ground	speed_ground	850	0
speed_air	speed_air	208	642
height	height	850	0
pitch	pitch	850	0
distance	distance	850	0

We have now removed the duplicate rows from the data files and we find that the merged dataset has 850 observations.

1.2. DATA EXPLORATION

1.2.1. COMPUTING FOR THE ABNORMAL AND MISSING VALUES IN VARIABLES:

1. Aircraft type

```
proc freq data = proj1sc.final ;
table aircraft /nocum nopercent nofreq;
where aircraft is missing;
run;
```

Inference: We find that there are no missing values in the aircraft column.

2. Duration:

```
proc freq data = proj1sc.final;
table duration /nocum nopercent nofreq;
where duration < 40 or duration is missing;
run;
```

The FREQ Procedure

duration	
duration	Frequency
14.764207145	1
16.893454896	1
17.375513046	1
31.391008253	1
31.7016661	1
Frequency Missing = 50	

Inference: We find that there are 5 abnormal values and 50 missing values with respect to the Duration variable.

3. **Speed Ground:**

```
proc freq data = proj1sc.final ;
table speed_ground /nocum nopercent nofreq;
where speed_ground < 30 or speed_ground >140 or speed_ground is missing;
run;
```

The FREQ Procedure

speed_ground	
speed_ground	Frequency
27.7357153033	1
29.2276563817	1
141.218635352	1

Inference: There are no missing values in Speed_ground column but there are 3 observations with abnormal data values in it.

4. **Speed Air:**

```
proc freq data = proj1sc.final ;
table speed_air /nocum nopercent nofreq;
where speed_air < 30 or speed_ground >140 or speed_air is missing;
run;
```

The FREQ Procedure

speed_air	
speed_air	Frequency
141.72493569	1
Frequency Missing = 642	

Inference: There are 642 missing values in Speed_Air column and there is 1 observation with abnormal data values in it.

5. **Height**:

```
proc freq data = proj1sc.final ;
table height /nocum nopercent nofreq;
where height < 6 or height is missing;
run;
```

The FREQ Procedure

height	
height	Frequency
-3.546252405	1
-3.332387973	1
-2.915335901	1
-1.528129182	1
-0.067758596	1
0.086105484	1
1.2538552556	1
2.2051944554	1
3.7889195211	1
4.2644634439	1

Inference: There are no missing values in height column but there are 10 observations with abnormal data values in it.

6. **Distance**:

```
proc freq data = proj1sc.final ;
table distance /nocum nopercent nofreq;
where distance >6000 or distance is missing or distance < 0;
run;
```

The FREQ Procedure

distance	
distance	Frequency
6309.9459762	1
6533.0476506	1

Inference: There are no missing values in distance column but there are 2 observations with abnormal data values in it.

1.3. DATA CLEANING

1.3.1. DEALING WITH MISSING VALUES

Since we can see that for the variable *Speed*_air there are 642 missing values, we cannot just delete all the observations with missing values in the dataset. We would like to explore options to find out if there is any way we can compensate for the missing observations. The possible approximations we tried and their consequences are summarized below:

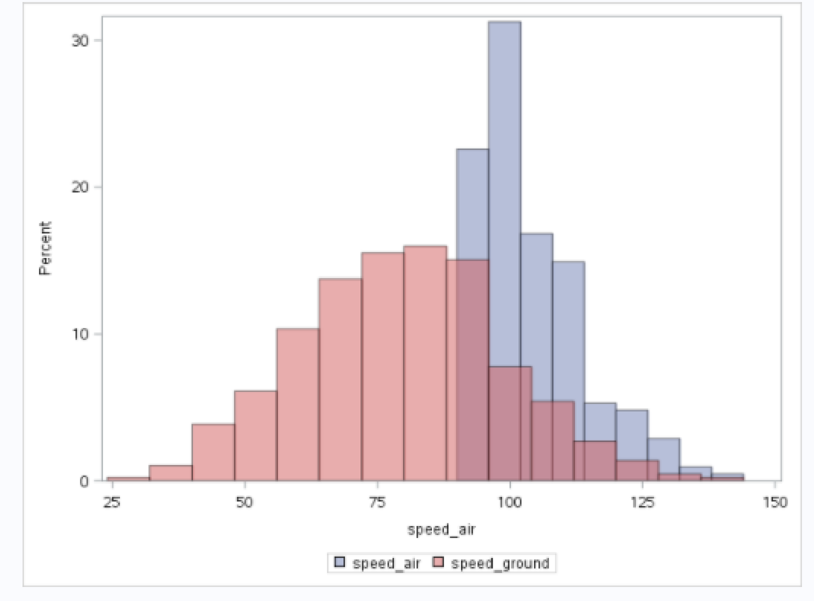
1.3.1.1. EXPLORING THE DISTRIBUTION OF SPEED_AIR

We want to understand the distribution of Speed_air and see if we use Speed_ground to make approximations for the Speed_air values in the missing observations:

When we try plotting the records for Speed_air and Speed_ground, we get the following distribution:

CODE:

```
proc sgplot data = proj1sc.final;
histogram speed_air /transparency=0.5;
histogram speed_ground /transparency=0.5;
run;
```



As we can see, the values of Speed_Air are present only above a value of 90. But after the values appear, we can see that the distribution of Speed_air overlaps with the distribution of Speed_ground. This might be because the sensors measuring speed_air are calibrated to operate only of the value is greater than 90. So, we cannot delete the observations with a value of NULL for Speed_Air as we might be missing out on a lot of details by doing that.

1.3.1.2. USING AVERAGES FOR MISSING OBSERVATIONS

We can maybe replace the missing values with the averages, but as we can see in the distribution, we do not want to make assumptions for variables at the beginning of the project.

1.3.2. DEALING WITH ABNORMAL VALUES

We find that the data has very few abnormal values which might skew our predictions if we use them for building a predictive model using them. But we just cannot get rid of the abnormal values. So we can store them in a Separate Dataset called “Abnormal_Flight”. We might want to use them later in our analysis.

```
NOTE: There were 21 observations read from the data set PROJ1SC.ABNORMAL.
NOTE: 2 observations with duplicate key values were deleted.
```

But for our analysis for now, we use a data set that does not contain any abnormal values in it. By cleaning the data set of the abnormal value, we have a data set with 831 observations.

```
NOTE: There were 831 observations read from the data set PROJ1SC.CLEAN_5.
NOTE: The data set PROJ1SC.FLIGHTCLEANED has 831 observations and 8 variables.
```

1.4. OBSERVATIONS

In the cleaned data set that we have now, we would like to understand how that variables are distributed.

1. Let us understand how many missing values we have in the dataset:

CODE:
```
proc means data= proj1sc.FLIGHTCLEANED N NMISS;
RUN;
```

The MEANS Procedure

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

We find that in all we have 831 observations for the 8 variables.

2. Now let us analyse at the variable level:

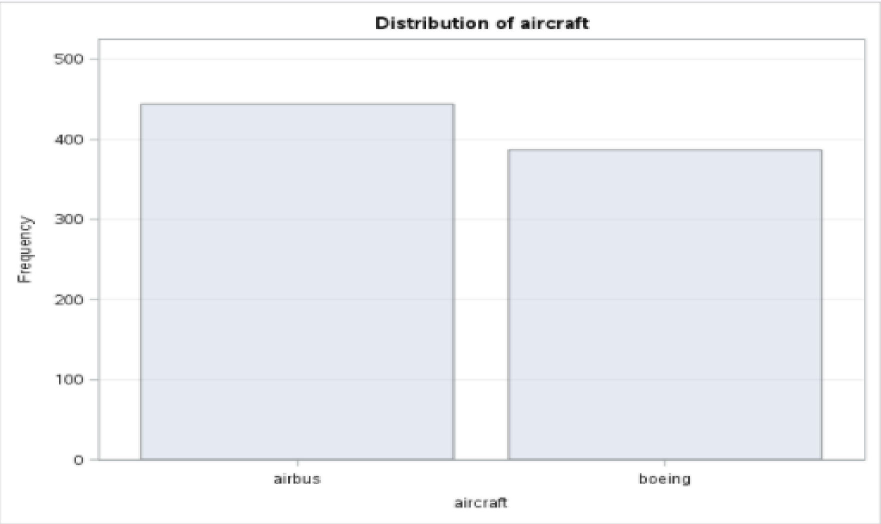
The aircraft column is of categorical type, so we use the FREQ function to estimate a frequency plot to understand the distribution of the variable.

CODE:
```
proc freq data = proj1sc.clean_5 NLevels;
	table aircraft /plots = freqplot;
run;
```

The FREQ Procedure

Number of Variable Levels		
Variable	Label	Levels
aircraft	aircraft	2

aircraft				
aircraft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
airbus	444	53.43	444	53.43
boeing	387	46.57	831	100.00



All other variables are continuous, so we would like to perform a univariate analysis to understand the variables characteristics. But since a univariate analysis might give us a lot of information about the measure of central tendency and spread of the variable, we can use a MEANS procedure to get a basic understanding of all the variables.

CODE:

```
proc means data= proj1sc.FLIGHTCLEANED mean stddev median q1 q3 min max;
var no_pasg speed_ground speed_air height pitch distance duration;
run;
```

The MEANS Procedure

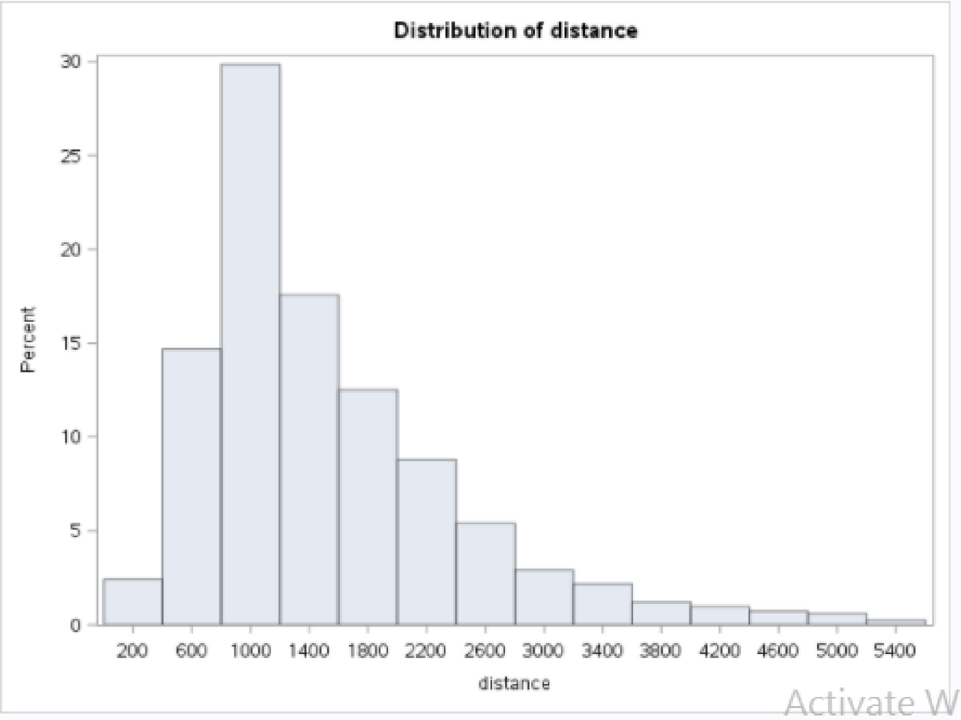
Variable	Label	Mean	Std Dev	Median	Lower Quartile	Upper Quartile	Minimum	Maximum
no_pasg	no_pasg	60.0553550	7.4913166	60.0000000	55.0000000	65.0000000	29.0000000	87.0000000
speed_ground	speed_ground	79.5426997	18.7356754	79.7939604	66.1925304	91.9496075	33.5741041	132.7846766
speed_air	speed_air	103.4850352	9.7362774	101.1189240	96.1964606	109.3823005	90.0028586	132.9114649
height	height	30.4578695	9.7848114	30.1670844	23.5298692	37.0143018	6.2275178	59.9459639
pitch	pitch	4.0051609	0.5265690	4.0010380	3.6403979	4.3710717	2.2844801	5.9267842
distance	distance	1522.48	896.3381524	1262.15	892.9839743	1937.26	41.7223127	5381.96
duration	duration	154.7757191	48.3499237	154.2845505	119.6314577	189.6629425	41.9493694	305.6217107

3. Since the major problem statement focuses on the landing Distance, lets focus on understanding the distribution of the landing distance which can be useful in the analysis going forward.

CODE:

```
proc univariate data = proj1sc.FLIGHTCLEANED;
var distance;
Histogram distance;
run;
```

The UNIVARIATE Procedure



CHAPTER: 2 RELATIONSHIP BETWEEN VARIABLES

Now that we have cleaned the dataset, we would now like to understand the impact of every factor on the Distance variable using linear regression modelling.

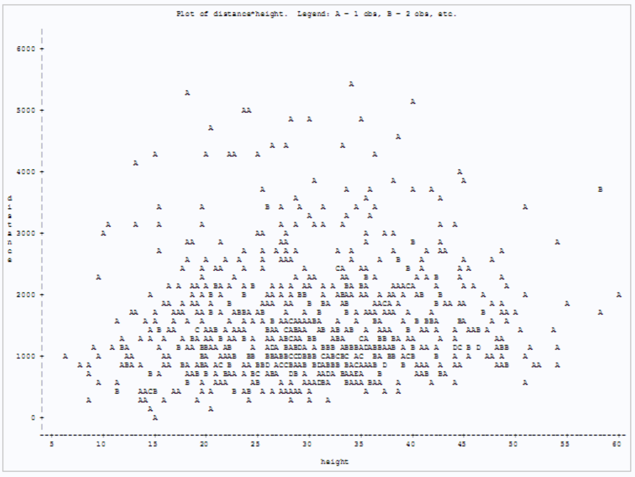
2.1. X-Y PLOTS:

First, let us try plotting the distribution of Distance with respect to some of the variables we have in our data set.

2.1.1. DISTANCE VS HEIGHT

CODE

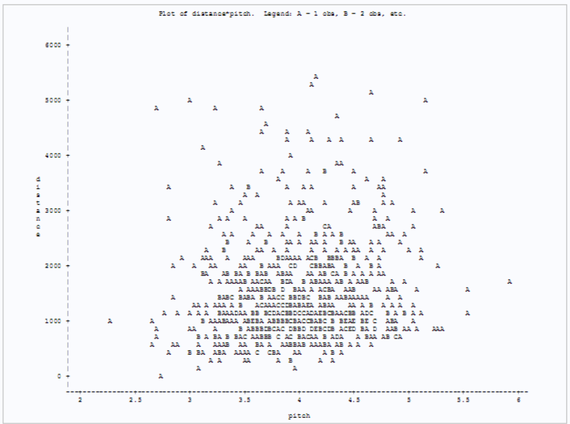
```
proc plot data = proj1sc.FLIGHTCLEANED;
plot distance*height;
run;
```



Looking at the graph, we are not able to make any inference about the relationship between distance and height.

2.1.2. DISTANCE VS PITCH

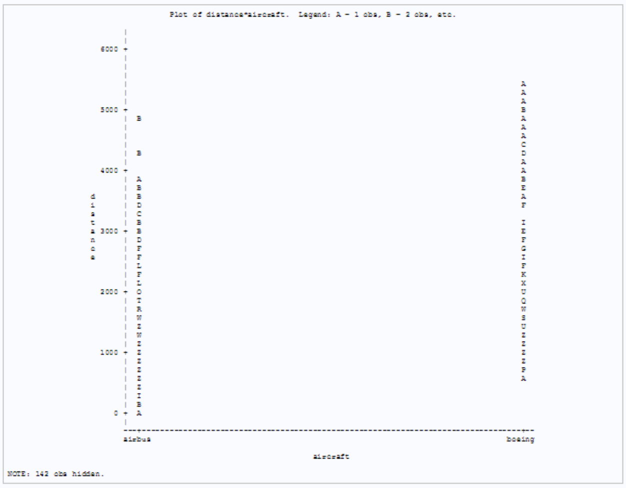
```
proc plot data = proj1sc.FLIGHTCLEANED;
plot distance*pitch;
run;
```



Looking at the graph, we are not able to make any inference about the relationship between distance and pitch.

2.1.3. DISTANCE VS TYPE OF AIRCRAFT

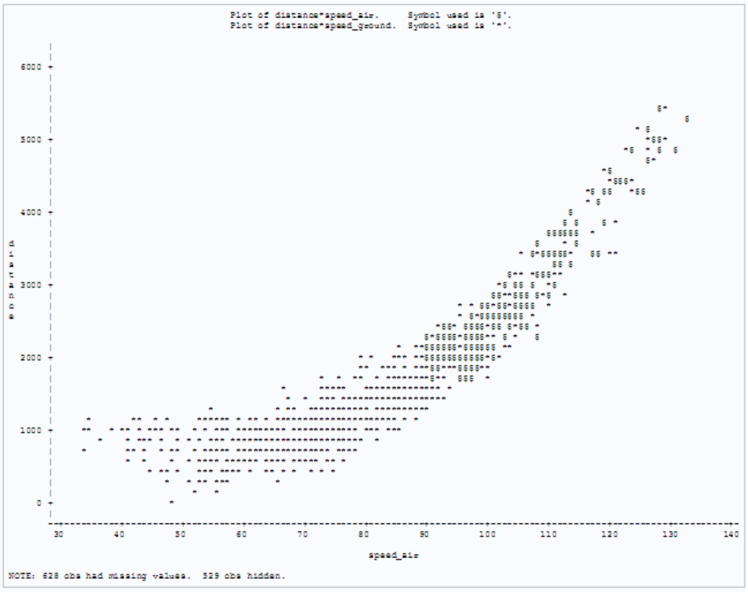
```
proc plot data = proj1sc.FLIGHTCLEANED;
plot distance*aircraft;
run;
```



Looking at the graph, we can say that the distribution of distance between the two types of aircrafts is slightly different. So, this variable might have an impact on the prediction of landing distance.

2.1.4. DISTANCE VS SPEED VARIABLES

```
proc plot data = proj1sc.FLIGHTCLEANED;
plot distance*speed_air  = "$" distance*speed_ground = "*" / overlay;
run;
```



Looking at the graph, we can see that the Speed_air and Speed_ground almost has the same relationship with the distance variable but the value of speed_air starts only after a value of

90. One take away from the graph is that both the speed variables might be useful in predicting the landing distance.

2.2. CORRELATION ANALYSIS:

Next step would be to run a correlation analysis on the data set to identify the significant variable to be considered for the prediction of distance.

2.2.1. CODING THE CATEGORICAL VARIABLE

Now to run a correlation analysis, firstly let us convert the categorical variable aircraft into numeric values. Only then we will be able to run a regression analysis.

```
/* Convert the categorical variable into a numerical condition */
data proj1sc.Flight;
set proj1sc.flightcleaned;
if (aircraft = "boeing") then type = 0;
else type = 1;
drop aircraft;
run;
```

2.2.2. PAIRWISE CORRELATION BETWEEN ALL VARIABLES

```
COMPUTING PAIRWISE CORRELATION:
proc corr data = proj1sc.flight;
var distance type no_pasg Speed_air Speed_ground height pitch duration;
title "Pairwise Correlation";
run;
```

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

	distance	type	no_pasg	speed_air	speed_ground	height	pitch	duration
distance distance	1.00000 831	-0.23814 <.0001 831	-0.01776 0.6093 831	0.94210 <.0001 203	0.86624 <.0001 831	0.09941 0.0041 831	0.08703 0.0121 831	-0.05138 0.1514 781
type	-0.23814 <.0001 831	1.00000 831	0.02269 0.5136 831	0.07207 0.3069 203	0.04045 0.2441 831	0.01439 0.6788 831	-0.35420 <.0001 831	0.04443 0.2149 781
no_pasg no_pasg	-0.01776 0.6093 831	0.02269 0.5136 831	1.00000 831	-0.00616 0.9305 203	-0.00013 0.9969 831	0.04699 0.1760 831	-0.01793 0.6057 831	-0.03639 0.3098 781
speed_air speed_air	0.94210 <.0001 203	0.07207 0.3069 203	-0.00616 0.9305 203	1.00000 203	0.98794 <.0001 203	-0.07933 0.2606 203	-0.03927 0.5780 203	0.04454 0.5364 195
speed_ground speed_ground	0.86624 <.0001 831	0.04045 0.2441 831	-0.00013 0.9969 831	0.98794 <.0001 203	1.00000 831	-0.05761 0.0970 831	-0.03912 0.2599 831	-0.04897 0.1716 781
height height	0.09941 0.0041 831	0.01439 0.6788 831	0.04699 0.1760 831	-0.07933 0.2606 203	-0.05761 0.0970 831	1.00000 831	0.02298 0.5082 831	0.01112 0.7564 781
pitch pitch	0.08703 0.0121 831	-0.35420 <.0001 831	-0.01793 0.6057 831	-0.03927 0.5780 203	-0.03912 0.2599 831	0.02298 0.5082 831	1.00000 831	-0.04675 0.1918 781
duration duration	-0.05138 0.1514 781	0.04443 0.2149 781	-0.03639 0.3098 781	0.04454 0.5364 195	-0.04897 0.1716 781	0.01112 0.7564 781	-0.04675 0.1918 781	1.00000 781

We find that the variables Speed_air and Speed_ground are highly correlated by the order of 98.7%. While building the model we surely need to consider their impact on inflating the predictions due to this multicollinear relationship.

2.2.3. CORRELATION OF VARIABLES WITH THE DISTANCE VARIABLE

The major area of interest is in understanding the correlation of all the variables to the distance variable.

CORRLATION WITH DISTANCE:

```
proc corr data = proj1sc.flight;
var type no_pasg Speed_air Speed_ground height pitch duration;
with distance;
title "Correlation with Distance";
run;
```

Correlation with Distance

The CORR Procedure

1 With Variables:	distance
7 Variables:	type no_pasg speed_air speed_ground height pitch duration

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	831	1522	896.33815	1265183	41.72231	5382	distance
type	831	0.53430	0.49912	444.00000	0	1.00000	
no_pasg	831	60.05535	7.49132	49906	29.00000	87.00000	no_pasg
speed_air	203	103.48504	9.73628	21007	90.00286	132.91146	speed_air
speed_ground	831	79.54270	18.73568	66100	33.57410	132.78468	speed_ground
height	831	30.45787	9.78481	25310	6.22752	59.94596	height
pitch	831	4.00516	0.52657	3328	2.28448	5.92678	pitch
duration	781	154.77572	48.34992	120880	41.94937	305.62171	duration

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations							
	type	no_pasg	speed_air	speed_ground	height	pitch	duration
distance distance	-0.23814 <.0001 831	-0.01776 0.6093 831	0.94210 <.0001 203	0.86624 <.0001 831	0.09941 0.0041 831	0.08703 0.0121 831	-0.05138 0.1514 781

We find that excepting the duration variable, all other variables have a significant correlation with the distance variable at 95% confidence level. Since the NULL hypothesis that rho = 0 for the distance variable cannot be rejected, we can leave the duration variable from our analysis.

CHAPTER: 3 REGRESSION MODELLING

3.1. CONSTRUCTING A MODEL WITH ALL VARIABLES

```
proc reg data = proj1sc.flight;
model distance = type no_pasg Speed_air Speed_ground height pitch;
title "Regression Analysis of the Flight Dataset";
run;
```

Regression Analysis of the Flight Dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	203
Number of Observations with Missing Values	628

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	132951503	22158584	1235.07	<.0001
Error	196	3516465	17941		
Corrected Total	202	136467968			

Root MSE	133.94458	R-Square	0.9742
Dependent Mean	2774.67289	Adj R-Sq	0.9734
Coeff Var	4.82740		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5804.26967	156.63260	-37.06	<.0001
type		1	-426.89121	20.32755	-21.00	<.0001
no_pasg	no_pasg	1	-2.29922	1.34357	-1.71	0.0886
speed_air	speed_air	1	88.03951	6.31100	13.95	<.0001
speed_ground	speed_ground	1	-5.89752	6.22131	-0.95	0.3443
height	height	1	13.62837	1.00747	13.53	<.0001
pitch	pitch	1	-4.65754	18.00975	-0.26	0.7962

3.2. ANALYSIS OF RESULTS

When we try interpreting the results, we find the following two conditions:

1. The model has used only 203 observations out of the 831 obseravations in the data set. This is due to the missing values in Speed_air variable.
2. We also find that despite both the speed variables having a large positive correlation with the distance, we are seeing that one of the speed variables is having a negative coefficient in the regression model which is counter intuitve. This might be due to the correlation between the speed variables. We need to explore the effect of multi collinearity on the regression model.

DISTANCE VS SPEED VARIABLES

```
proc corr data = proj1sc.flight;
var speed_air speed_ground;
with distance;
run;
```

The CORR Procedure

1 With Variables:	distance
2 Variables:	speed_air speed_ground

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	831	1522	896.33815	1265183	41.72231	5382	distance
speed_air	203	103.48504	9.73628	21007	90.00286	132.91146	speed_air
speed_ground	831	79.54270	18.73568	66100	33.57410	132.78468	speed_ground

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	speed_air	speed_ground
distance distance	0.94210 <.0001 203	0.86624 <.0001 831

3.3. COMPUTING VIF AND TOLERANCE:

```
proc reg data = proj1sc.flight;
model distance = type no_pasg Speed_air Speed_ground height pitch /vif tol;
title "Regression Analysis of the Flight Dataset";
run;
```

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-5804.26967	156.63260	-37.06	<.0001	.	0
type		1	-426.89121	20.32755	-21.00	<.0001	0.87877	1.13795
no_pasg	no_pasg	1	-2.29922	1.34357	-1.71	0.0886	0.99609	1.00392
speed_air	speed_air	1	88.03951	6.31100	13.95	<.0001	0.02352	42.50919
speed_ground	speed_ground	1	-5.89752	6.22131	-0.95	0.3443	0.02345	42.63875
height	height	1	13.62837	1.00747	13.53	<.0001	0.98610	1.01409
pitch	pitch	1	-4.65754	18.00975	-0.26	0.7962	0.86989	1.14956

We know that, tolerance (requested by the tol option) is the proportion of variance in a given predictor that is NOT explained by all of the other predictors, while the VIF (or Variance Inflation Factor) is simply 1 / tolerance. The VIF represents a factor by which the variance of the estimated coefficient is multiplied due to the multicollinearity in the model

A good “global” check for a multicollinearity problem is to see if the largest condition index is greater than 30.

Here we find that the VIF is 42 for both the speed variables. So, we need to eliminate one of those variables to get a proper fit to the model. We also know that the speed_air variable has 641 missing values. It is always better to fit a model with more data than less. So we can eliminate Speed_air from our model and construct the linear regression equation.

3.4. MODEL COMPENSATED FOR MULTICOLLINEARITY

```
/* Final Model */

proc reg data = proj1sc.flight;
```

```
model distance = type no_pasg Speed_ground height pitch /vif tol;
title "Regression Analysis of the Flight Dataset";
run;
```

Regression Analysis of the Flight Dataset

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	566620920	113324184	932.88	<.0001
Error	825	100219409	121478		
Corrected Total	830	666840329			

Root MSE	348.53705	R-Square	0.8497
Dependent Mean	1522.48287	Adj R-Sq	0.8488
Coeff Var	22.89267		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-2051.91594	156.62442	-13.10	<.0001	.	0
type		1	-480.69168	25.94116	-18.53	<.0001	0.87302	1.14544
no_pasg	no_pasg	1	-2.20392	1.61722	-1.36	0.1733	0.99716	1.00285
speed_ground	speed_ground	1	42.42955	0.64754	65.52	<.0001	0.99436	1.00567
height	height	1	14.17035	1.24050	11.42	<.0001	0.99339	1.00666
pitch	pitch	1	39.20658	24.58808	1.59	0.1112	0.87309	1.14536

We have the following observations:

- The model is based on all the 831 observations present in the cleaned dataset.
- Type of aircraft has a major impact on the landing distance with a coefficient of the order of -480. Which means the type of aircraft affects the landing distance by a factor of 480.
- Factors like Speed_Ground, Height and Pitch have a positive impact on the landing distance
- The no_pasg has a negative impact on the landing distance but the magnitude of impact is small compared to other factors

3.5. JUSTIFICATION FOR VARIABLE SELECTIONS

As per the results we got in our process of understanding the variables, we made two choices with respect to the variable:

1. Eliminated duration variable from the list as we found that the NULL Hypothesis for rho=0 couldn't be rejected for the relationship between duration and distance
2. We eliminated speed_air variable to compensate for the multicollinearity issue we had on the model due to the correlation between speed_air and speed_ground. We chose to eliminate speed_air among the two variables because it has a lot of missing values and we would always want to use more data for developing the model.

QUESTIONS:

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

We used 831 observations from the data set after getting rid of the abnormal observations. Also, we have made variable selections to make sure our model is robust and is not overfitting the given dataset.

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

We can see from the regression analysis that the factors like Speed_ground, No_pasg, pitch, Height and type of aircraft impact the landing distance.

Out of all the predictors, we find that Speed_gound, pitch, height and type of aircraft are the factors majorly affecting the landing distance. Factors like no_pasg has comparatively lesser impact on the landing distance.

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

Yes, there is a difference in the landing distance between the two types of aircrafts. We tried understanding this by plotting the distributions of landing distance and the type of aircraft. Also in the regression model, we get a coefficient of the order of 480 which means that the type of aircraft would affect the landing distance by a factor of 480.

```
proc univariate data = proj1sc.FLIGHTCLEANED;
class aircraft;
histogram distance /overlay;
run;
```