Project: Statistical Analysis of flight landing

- Identifying the factors affecting landing distance using SAS

Submitted in partial fulfillment of the requirements of the course

BANA 6043 Statistical Computing

By

Adrian Valles

MS Business Analytics Candidate 2018

Department of Operations, Business Analytics & Information Systems

Carl H Lindner School of Business: University of Cincinnati

Project 1-Statistical computing

With the motivation to reduce the risk of flight landing overrun, the following report studies the factors that impact the landing distance of a commercial flight. Landing data from 950 simulated commercial flights (divided in two Excel files 'FAA-1.xls' and 'FAA-2.xls') will be analyzed in the report. Below it is a summary of the variables of each fligh. The report is divided into 3 chapters covering data preparation (Ch.1), a descriptive study of the variables (Ch. 2) and statistical modeling with a brief model diagnostic (Ch. 3). Some of the techniques applied to the study are data cleaning techniques such as combining files from different sources, performing validity and completeness checks of the variables and elimination of duplicates and missing values. Other more advanced techniques that have been applied are the use of plots and correlation studies of the variables, the formulation of linear regression models to fit the data and the diagnostic check of these models. The final result of the study is the creation of a regression model that allows to predict landing overruns based on the factors that affect this landing distance. The software used for this statistical report is SAS and some of the output as well as the code can be found in the report.

Variables:

Aircraft: The make of an aircraft (Boeing or Airbus).

Duration (in minutes): Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min.

No_pasg: The number of passengers in a flight.

Speed_ground (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.

Speed_air (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.

Height (in meters): The height of an aircraft when it is passing over the threshold of the runway. The landing aircraft is required to be at least 6 meters high at the threshold of the runway.

Pitch (in degrees): Pitch angle of an aircraft when it is passing over the threshold of the runway.

Distance (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

CHAPTER 1-Data preparation

Step 1: Importing the files and combining them. I decided to sort the combined dataset by aircraft name.

```
proc import datafile= '/home/vallesan0/BANA 6043 STAT COMP/FAA1.xls'
out=data1
dbms=xls
replace;
run;

proc import datafile= '/home/vallesan0/BANA 6043 STAT COMP/FAA2.xls'
out=data2
dbms=xls
replace;
run;

/********Combine both files by aircraft name*******/
PROC SORT data=data1;
BY Aircraft; /*sorts within the first data set*/
PROC SORT data=data2;
BY Aircraft; /*sorts within the second data set*/
DATA COMBINED2;
SET data1 data2;
BY Aircraft; /*combines in the order of dates*/
RUN;

data Combined3; /* when importing the excel file I got some blank cells, so I am
deleting them here*/
   set Combined2;
   if Aircraft = '' then delete;
run;
proc print data=Combined3;
   run;
```

Step 2: It seems that the two different datasets could have duplicates. Therefore, we are going to check if there are duplicates and if there are, delete them.

```
PROC FREQ data=combined3;
 TABLES aircraft*duration*no_pasg*speed_ground*speed_air*height*pitch*distance/ noprint
out=keylist;
RUN;
PROC PRINT;
 WHERE count ge 2;
RUN;
/*removing the 100 duplicates*/

 Proc sort data=Combined3 out=combined3a nodupkey dupout=Duplicate;
    by distance;   Run;
proc print data=Combined3a;
  run;
```

There are in fact 100 duplicates, so we delete them and get a provisional sample size of 850 observations. Also, it is worth mentioning that we have an extremely high certainty that the repeated values are in fact duplicates. This is due to the high number of decimal values that we have for certain values of the observations.

Examining the number of missing values for each variable will be the next step. The tables that are attached show the variables with missing values (when running the code in SAS, the tables of all the variables are shown, but I have only attached the ones with missing values).

Therefore, only 'duration' and 'speed_Air' have missing values. While the number of missing values for 'duration' represents a 5.8% of the total values, the number of missing values for the Speed air variable represent 75.5% of the total. Therefore we will try to understand the reason for this.

duration	
duration	Frequency
Missing	50
Not Missing	800

speed_air	
speed_air	Frequency
Missing	642
Not Missing	208

```
proc format;
 value $missfmt ' '='Missing' other='Not Missing';
 value  missfmt  . ='Missing' other='Not Missing';
run;

proc freq data=combined3a;
format _CHAR_ $missfmt.;
tables _CHAR_ / missing missprint nocum nopercent;
format _NUMERIC_ missfmt.;
tables _NUMERIC_ / missing missprint nocum nopercent;
run;
```

Step 3: Next action I did is to check the abnormal values considering abnormal conditions in the 'speed ground' 'speed air' 'height' 'duration' and 'distance' variables. I decided to create new variables (as it can be seen below) with outcome 'Normal' or 'Abnormal' depending on the normality of the observation.

ground_abnormal	air_abnormal	height_abnormal	duration_abnormal	distance_abnormal
Normal	Normal	Abnormal	Normal	Normal
Normal	Normal	Normal	Normal	Normal
Normal	Normal	Normal	Normal	Normal
Normal	Normal	Normal	Normal	Normal
Normal	Normal	Normal	Normal	Normal

This procedure allows the reader to easily identify the abnormal values, and it will help during the data cleaning process(These variables will be removed from the final data set after the data cleaning process).

As it can be shown, the number of abnormal values is relatively small to the number of observations (in the case of the variable 'speed_air', missing values are not treated as abnormal values).

ground_abnormal	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Abnormal	3	0.35	3	0.35
Normal	847	99.65	850	100.00

air_abnormal	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Abnormal	1	0.12	1	0.12
Normal	849	99.88	850	100.00

height_abnormal	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Abnormal	10	1.18	10	1.18
Normal	840	98.82	850	100.00

duration_abnormal	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Abnormal	5	0.59	5	0.59
Normal	845	99.41	850	100.00

distance_abnormal	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Abnormal	2	0.24	2	0.24
Normal	848	99.76	850	100.00

```
data combined4;
set combined3a;
if speed_ground<30 or speed_ground>140 then ground_abnormal="Abnormal";
else ground_abnormal="Normal";
if speed_air>140 then air_abnormal="Abnormal";
else air_abnormal="Normal";
if height<6 then height_abnormal="Abnormal";
else height_abnormal="Normal";
if duration <40 and duration>0 then duration_abnormal="Abnormal";
else duration_abnormal="Normal";
if distance>6000 then distance_abnormal="Abnormal";
else distance_abnormal="Normal";

run;
proc print data=combined4;
run;
proc freq data=combined4;
tables ground_abnormal air_abnormal height_abnormal duration_abnormal
distance_abnormal;  /* be careful because the missing values are beng counted as abnormal*/
run;
```

Step 4: Given all the information of missing values and abnormal values, it is time to make some cleaning on the dataset.

For this, I have started by taking the following measures:

1. Eliminate observations with abnormal values. Given to the possibility that these abnormal values are due to an input error, I have decided to eliminate all the

observation that contain abnormal values. The other main reason to eliminate these observations is that training a model with normal values leads to a better performance of the model. Finally, abnormal observations only represent 2.2% of the total. I have created a separate dataset containing only abnormal values with the 19 abnormal observations, just in case we want to use it to test our model in the future.

The following dataset shows all the observations containing abnormal values. It can be a useful for model testing purposes.

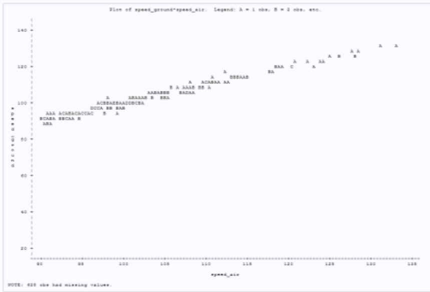
Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	airbus	150.94674427	58	66.421119468	.	-2.915335901	3.1225583646	34.080783293
2	boeing	133.45985625	73	57.045299494	.	1.2538552556	4.7153842391	371.27726086
3	airbus	157.91497689	68	56.497986661	.	-0.067758596	4.6928768405	380.36298195
4	boeing	283.76336844	62	58.889312381	.	4.2644634439	4.7721930401	425.85856098
5	airbus	163.52364053	62	72.028024252	.	0.086105484	3.6220566648	537.91958189
6	boeing	175.08462089	64	52.493139102	.	-3.546252405	4.2132855404	581.38099947
7	boeing	124.37864547	72	60.367043725	.	3.7889195211	3.7060888319	641.59956822
8	boeing	146.04337112	69	71.787305883	.	-1.528129182	4.1994604645	738.65436932
9	boeing	119.64402906	68	70.178463873	.	2.2051944554	3.7397746803	816.20664104
10	airbus	31.7016661	61	76.354176433	.	30.991021813	2.8173796019	948.47376723
11	boeing	17.375513046	63	63.57042961	.	28.406673108	3.9378640453	1032.4646189
12	boeing	212.94303494	61	29.227656382	.	23.349901124	4.3961881217	1076.855217
13	boeing	141.93411511	46	27.735715303	.	24.400127629	4.3682093233	1323.7157777
14	airbus	103.09084673	73	92.994942381	.	-3.332387973	4.8305592948	1567.6657219
15	airbus	16.893454896	54	94.511052223	95.930926862	37.476967053	4.1733221259	2162.92737
16	boeing	31.391008253	51	98.219800666	99.057514589	52.473140903	4.1623371208	2808.3151244
17	boeing	14.764207145	59	108.29169029	109.32758442	46.930873666	4.8096217396	3645.6110025
18	boeing	119.92455279	64	136.65915832	136.42342138	44.286109179	4.1694037368	6309.9459762
19	boeing	180.61655753	54	141.21863535	141.72493569	23.575935009	5.2168022511	6533.0476506

```
data abnormal_height;
set combined4;
if height_abnormal='Abnormal';
run;
proc print data=abnormal_height;
run;

data abnormal_ground;
set combined4;
if ground_abnormal='Abnormal';
run;
proc print data=abnormal_ground;
run;
data abnormal_air;
set combined4;
if air_abnormal='Abnormal';
run;
proc print data=abnormal_air;
run;
data abnormal_distance;
set combined4;
if distance_abnormal='Abnormal';
run;
proc print data=abnormal_distance;
run;
```

```
data abnormal_duration;
set combined4;
if duration_abnormal='Abnormal';
run;
proc print data=abnormal_duration;
run;
/* data set containing all theabnormal values*/
data abnormal;
set abnormal_ground abnormal_height abnormal_duration
abnormal_distance abnormal_air;
drop ground_abnormal    air_abnormal     height_abnormal
        duration_abnormal         distance_abnormal;
by distance;
run;
Proc sort data=abnormal out=abnormal_final nodupkey
dupout=Duplicate;
   by distance;
   Run;
proc print data=abnormal_final;
  run;
```

2. After some research, I have decided to do something with the speed_air variable. At a first glance, it is interesting to see that there are no values below 90. Consequently, it seems that the tool that measures the speed_air has some kind of trouble with values below 90. I have found out that there is a strong correlation between speed ground values and speed_air values (0.9879), as it can also be seen in the graph.

	speed_ground	speed_air
speed_ground speed_ground	1.00000 831	0.98794 <.0001 203
speed_air speed_air	0.98794 <.0001 203	1.00000 203



Therefore, given that we only have 24.5% of the values of the speed_air variable, we are going to predict the missing values given the ground_speed values. When creating a new variable of the difference of these 2 variables, this is the summary statistics that we get for the 'difference' variable.

Analysis Variable : difference				
N	Mean	Std Dev	Minimum	Maximum
203	-0.0738829	1.5321314	-3.4350609	5.3756363

Therefore, by setting the speed_air missing values equal to Speed_ground – 0.07388, we can be 95% confident that the imputed value will be +-3 (2 standards of deviation) to its real value. For a value of speed_air of 90, an error of 3 corresponds to an error of 3.33%, which is pretty small. As a consequence, the prediction seems pretty accurate.

3. Finally, for the missing values of the variable 'duration', no further action is going to be taken, because as we will see in later in chapter 2, it is not highly correlated with any other variable. Therefore, predicting the missing values with other variables will be too risky.

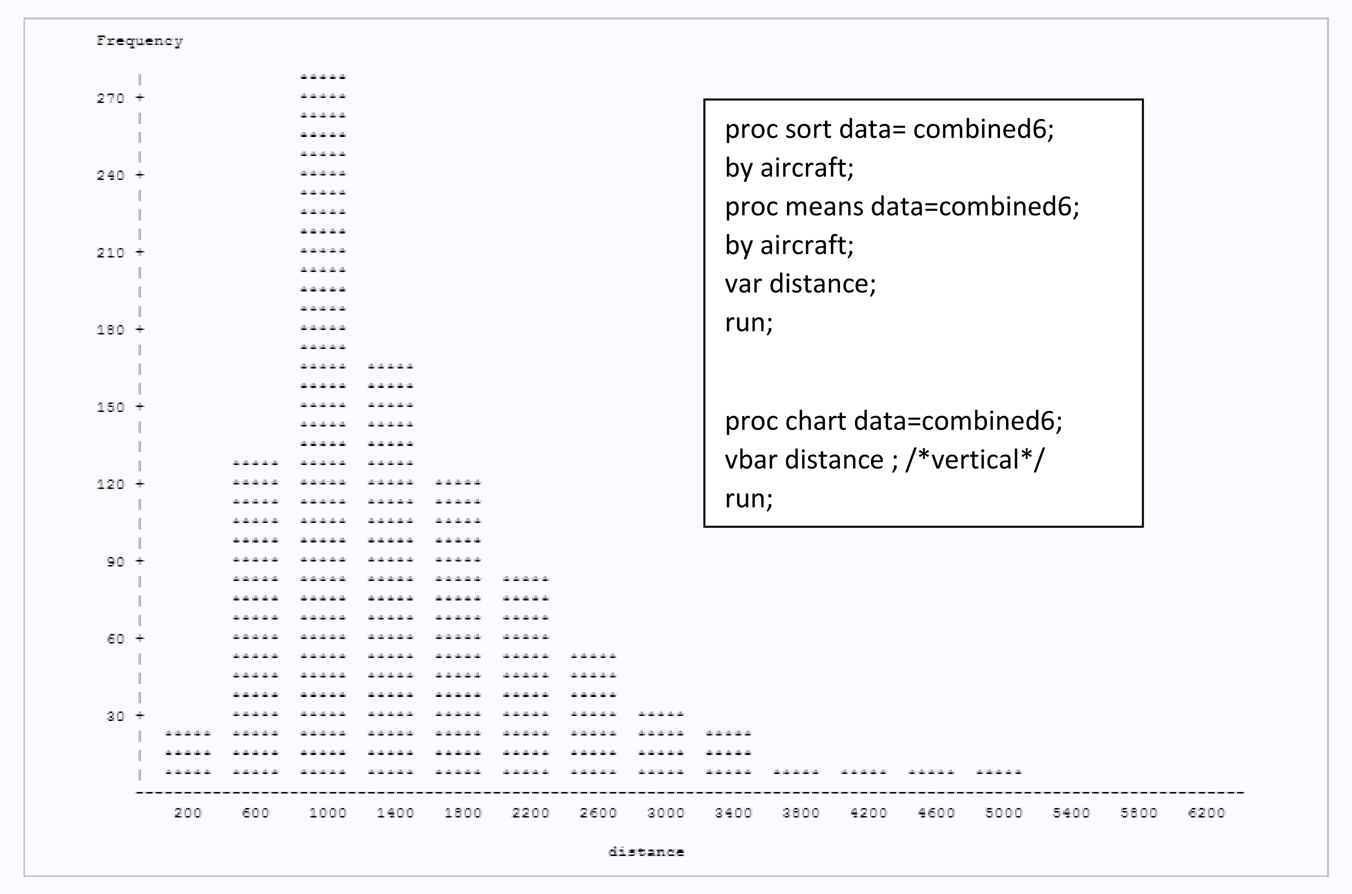
After having done this data cleaning actions, the dataset that I got consist of **831 observations** and **8 variables** as it can be seen below (sample of the first 10 obs.). It has 50 missing values form the duration variable and no abnormal values.

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	airbus	192.28287317	64	33.574104065	33.500221165	36.970689868	4.3584643374	782.7174172
2	boeing	143.33261546	68	33.822953314	33.749070414	37.680495945	4.0703468804	936.57807216
3	boeing	84.296505847	65	34.11776613	34.04388323	35.266706534	4.2902689148	1198.0892035
4	boeing	233.43123856	68	34.222063657	34.148180757	28.629155926	4.7888425657	955.9096666
5	boeing	192.181773	67	34.30363513	34.22975223	30.138274703	4.4995089168	1001.0805665
6	airbus	126.07843541	54	36.421388861	36.347505961	33.799699892	4.8661106393	869.03373396
7	boeing	136.32776148	52	38.259020081	38.185137181	28.338283561	3.9376315891	981.14796314
8	boeing	222.70208536	52	39.725711308	39.651828408	33.265348033	4.4522817052	1037.914549
9	boeing	142.15534911	46	39.769294325	39.695411425	39.655921061	4.5992872267	1030.457488
10	boeing	71.877046469	50	40.676738571	40.602855671	39.550442157	4.1852995438	974.57210835

```
/* data cleaning***/
data combined5;
set combined4;

if height_abnormal='Normal'; /* eliminate all the observations where height is abnormal*/
if ground_abnormal='Normal'; /* eliminate all the observations where ground is abnormal*/
if air_abnormal='Normal';
if duration_abnormal='Normal';
if distance_abnormal='Normal';
run;
proc print data= combined5;
run;

data combined6;
set combined5 ;
keep aircraft     duration no_pasg speed_ground speed_air height pitch       distance;
run;
proc print data=combined6;
run;
/************relationship between speed ground and speed air********/
proc sort data= combined6;
by speed_ground;
proc print data= combined6;
run;

data difference;
set combined6;
difference= speed_ground-speed_air;
run;
proc print data=difference;
run;

proc means data=difference;
var difference;
run;

PROC PLOT data=difference;
PLOT speed_ground*speed_air;
run;

proc corr data=difference;
var speed_ground speed_air ;
title correlation coefficients;
run;

/* predict values for the missing values of speed_air***/
data combined7;
set combined6;
if speed_air='.' then speed_air=speed_ground-0.0738829;
run;
proc print data=combined7;
run;
/* see if duration is highly correlated with any other variable inorder ot imput their values*/
proc corr data=combined7;
var duration ;
with    no_pasg speed_ground   speed_air      height   pitch    distance;
run;
```

Step 5: Finally, I have proceeded to summarize the distributions of the variables. I have started with the variable 'distance' (response variable) as it is probably the most important in our study. I have started my study on the 'distance' variable with a histogram that shows the 'distance' values of all 831 observations of the study. As we can see, most of the values are concentrated around 1000 ft. It almost seems that the distance variable follows an exponential distribution. It is at least clear that the distribution is skewed to the right.

Then, I decided to study the distance based on the 'aircraft' variable to see how the airbus differs from the Boeing in the breaking distance. At a first glance, It seems like the Boeing aircrafts need more distance than the Airbus as it can be seen in the tables below



aircraft=airbus

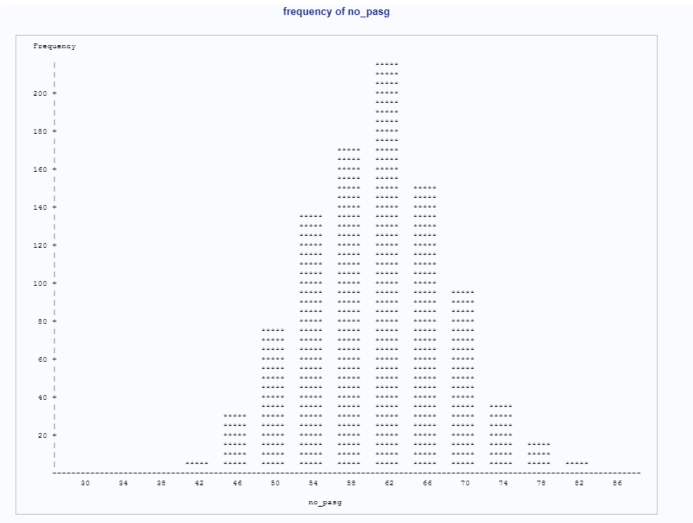
Analysis Variable : distance distance				
N	Mean	Std Dev	Minimum	Maximum
444	1323.32	791.9282481	41.7223127	4896.29

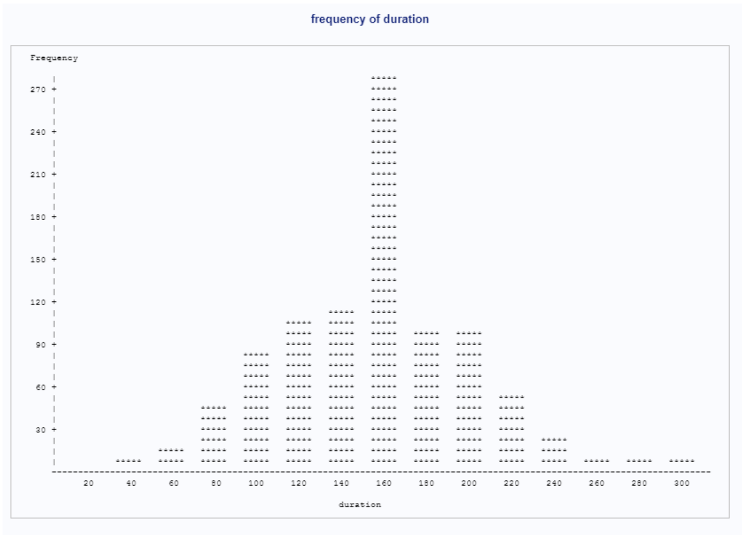
aircraft=boeing

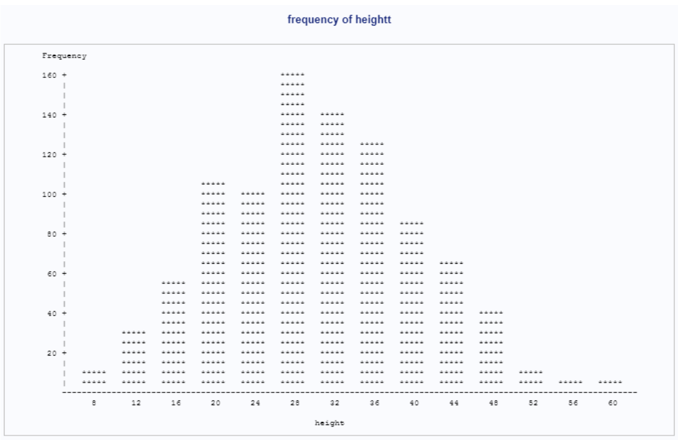
Analysis Variable : distance distance				
N	Mean	Std Dev	Minimum	Maximum
387	1750.98	953.8500300	573.6217861	5381.96

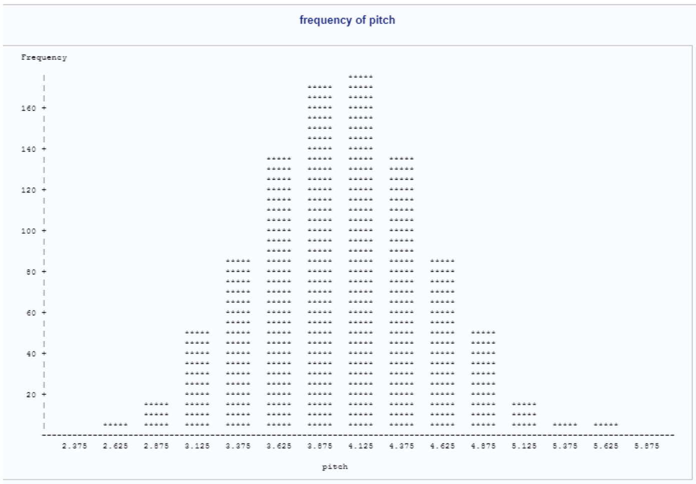
Once I have got some insights on the 'distance' variable, I decided to study the distribution of the other variables. Below, there is a simple summary statistic of all the other variables and all of them seem to follow normal distribution as the bell shape of their histograms shows For the 'speed_air' variable, I have decided to study the distribution of all the values, including the ones that I predicted based on the 'speed_ground'.

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
pitch	pitch	831	4.0051609	0.5265690	2.2844801	5.9267842
height	height	831	30.4578695	9.7848114	6.2275178	59.9459639
speed_ground	speed_ground	831	79.5426997	18.7356754	33.5741041	132.7846766
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_air	speed_air	831	79.5049137	18.7619959	33.5002212	132.9114649

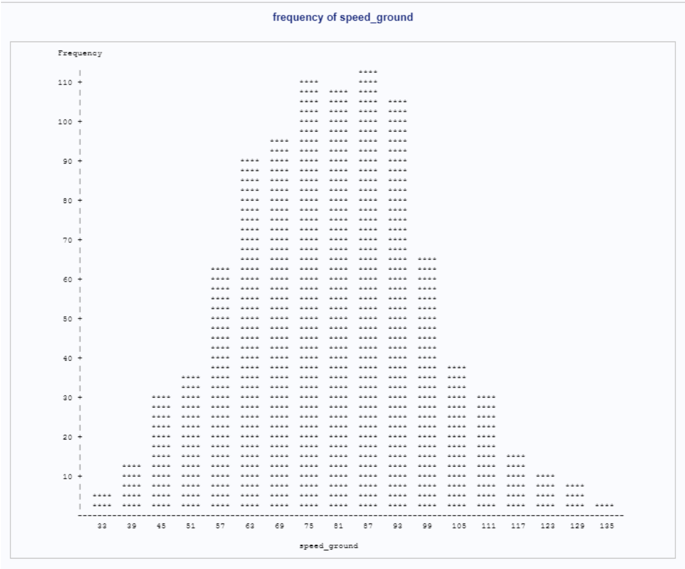


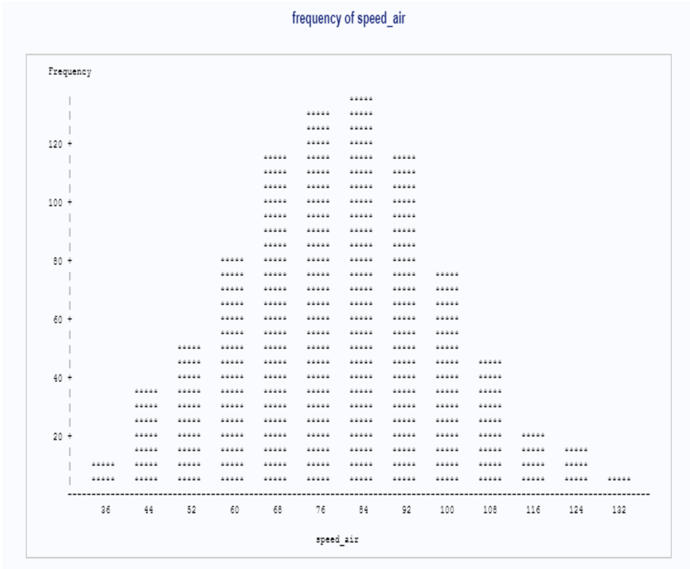






Adrian Valles M07593734






```
proc chart data=combined7;
vbar no_pasg ; /*vertical*/
TITLE frequency of no_pasg;
run;


proc chart data=combined7;
vbar duration ; /*vertical*/
TITLE frequency of duration;
run;
proc chart data=combined7;
vbar speed_ground ; /*vertical*/
TITLE frequency of speed_ground;
run;
proc chart data=combined7;
vbar speed_air ; /*vertical*/
TITLE frequency of speed_air;
run;
/* histogram for duration...normal dist*/

proc chart data=combined7;
vbar pitch ; /*vertical*/
TITLE frequency of pitch;
run;

proc chart data=combined7;
vbar height ; /*vertical*/
TITLE frequency of heightt;
run;


proc means data=combined7;
var pitch height speed_ground duration no_pasg speed_air;
run;
```

Therefore, based on the histograms, all the variables but 'distance' seem to follow a clear normal distribution

Additional research will be conducted in the next chapters.

CHAPTER 2-Descriptive study

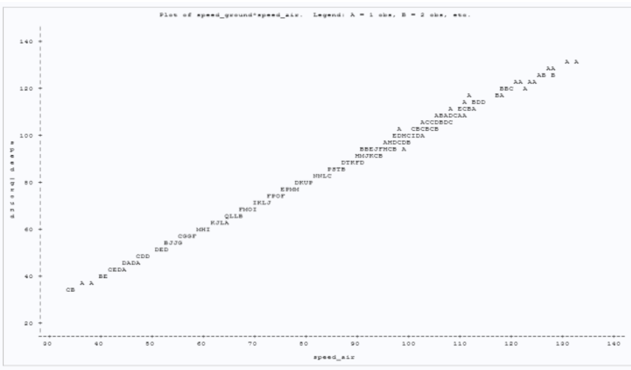
Next, I will conduct a descriptive study in order to see the relationship and correlation between the variables. We are particularly interested in the relationship between the response variable 'distance' and the other variables.

We will start by studying the correlation between the independent variables. Attached are the correlation coefficients between all the independent variables.

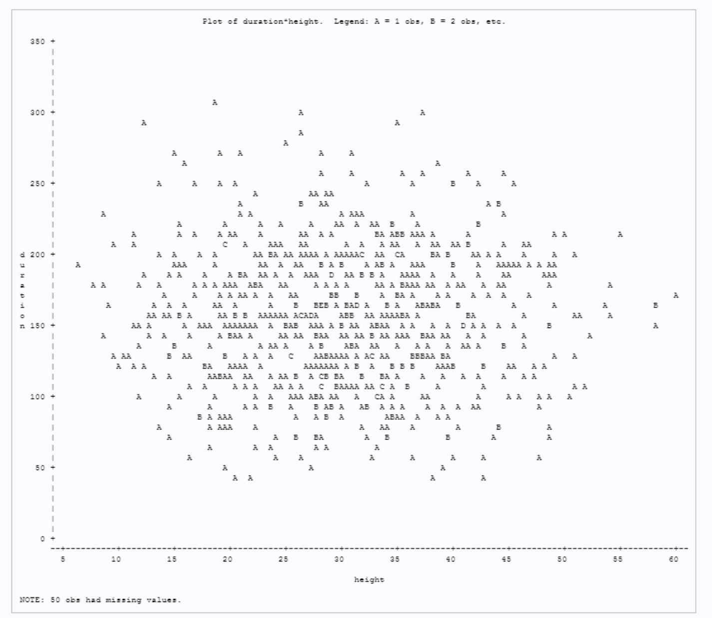
Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations						
	duration	no_pasg	speed_ground	speed_air	height	pitch
duration duration	1.00000 781	-0.03639 0.3098 781	-0.04897 0.1716 781	-0.04645 0.1948 781	0.01112 0.7564 781	-0.04675 0.1918 781
no_pasg no_pasg	-0.03639 0.3098 781	1.00000 831	-0.00013 0.9969 831	-0.00056 0.9871 831	0.04699 0.1760 831	-0.01793 0.6057 831
speed_ground speed_ground	-0.04897 0.1716 781	-0.00013 0.9969 831	1.00000 831	0.99918 <.0001 831	-0.05761 0.0970 831	-0.03912 0.2599 831
speed_air speed_air	-0.04645 0.1948 781	-0.00056 0.9871 831	0.99918 <.0001 831	1.00000 831	-0.05631 0.1048 831	-0.03616 0.2978 831
height height	0.01112 0.7564 781	0.04699 0.1760 831	-0.05761 0.0970 831	-0.05631 0.1048 831	1.00000 831	0.02298 0.5082 831
pitch pitch	-0.04675 0.1918 781	-0.01793 0.6057 831	-0.03912 0.2599 831	-0.03616 0.2978 831	0.02298 0.5082 831	1.00000 831

At a first glance, most of the variables doesn't seem to have high levels of correlation. However, there is a case that has a high level of correlation and therefore needs to be studied.

1. This is the case of the extremely high correlation between speed_Air and speed_ground. We already studied this relationship in chapter 1 when we decided to predict the speed_air missing values based on the speed_ground ones. It is worth mentioning that logically the correlation score is now higher (0.999) than when we had missing values. This is obviously due to the fact that we have used speed_ground values to predict speed_air ones. Below it is the speed_air*speed_ground plot and as we can see, the higher the speed_ground values, the higher the speed_Air values. There is therefore a positive relationship between these 2 variables



2. All the other variables are independent of each other as they don't present high levels of correlation. A clear example of this is the relationship between duration and height. As we can see in the plot, any height value can almost take any duration value and vice-versa.

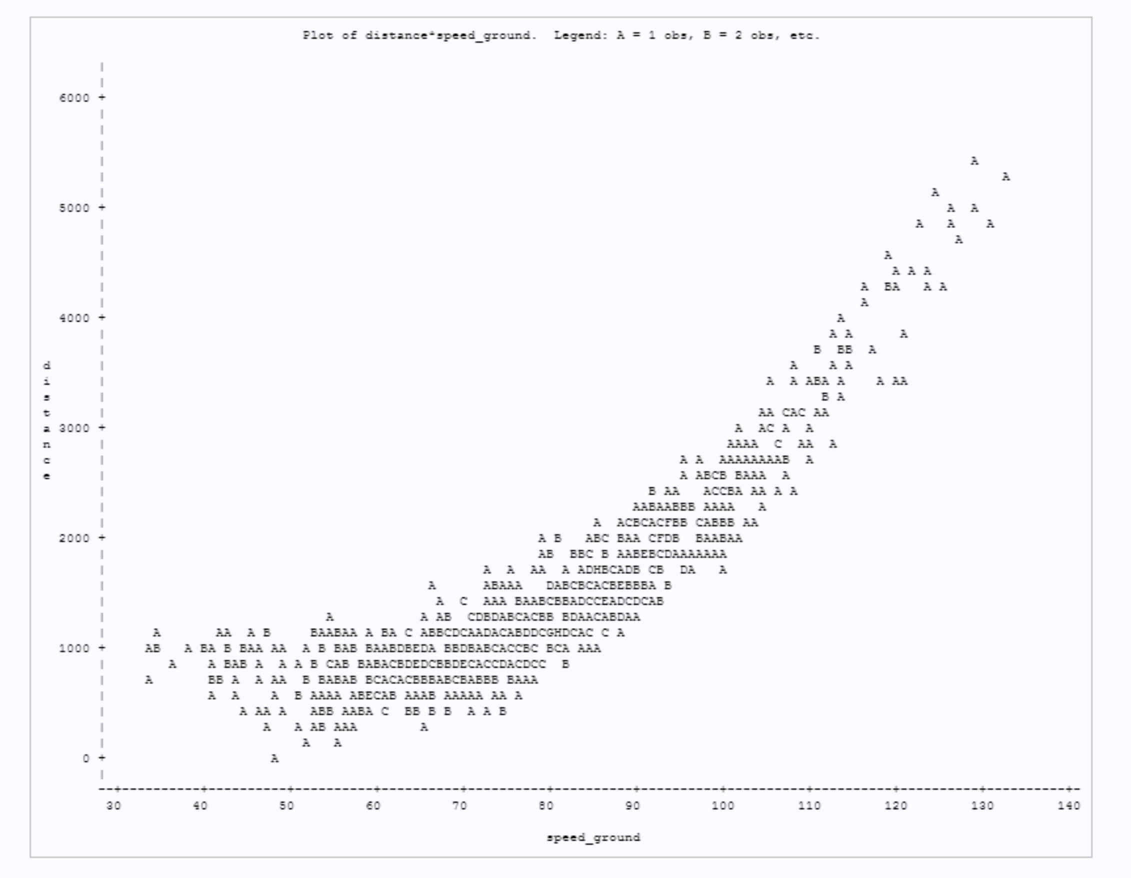



```
proc corr data=Combined7;
var     duration        no_pasg
        speed_ground  speed_air
        height   pitch    ;
title  correlation coeficients
run;

/*plots of highly correlated
values*/
proc plot data=combined7;
plot speed_ground*speed_air;
plot height*duration;
run;
```

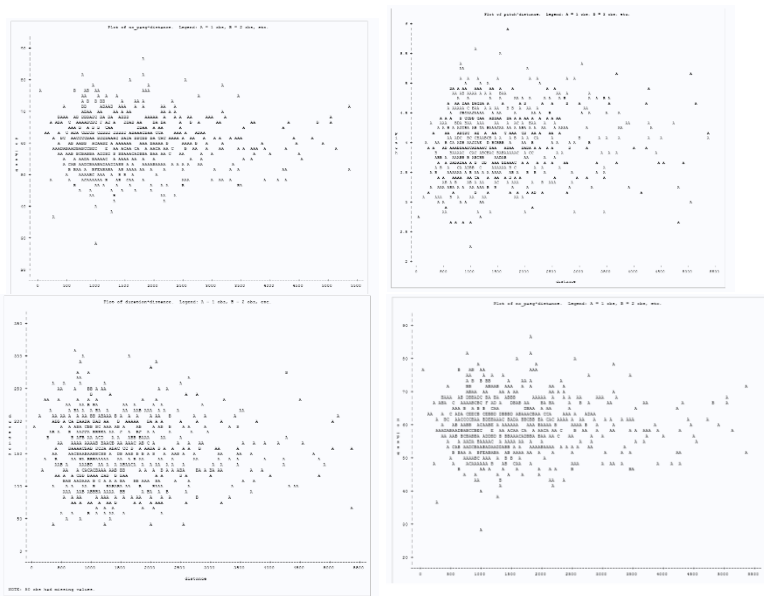
After having studied the relationship between the independent variables we can conclude that almost none of them are correlated. The only strong correlation that we have observed is the one between speed_air and speed_ground. Our next action is to observe the correlation of the response variable with the independent variables. The next table shows this correlation scores. At a first glance, only speed_ground (and obviously speed_air) seem to have a high level of correlation with the response variable. All the other variables aren't correlated with the 'distance' variable.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	
	distance
duration duration	-0.05138 0.1514 781
no_pasg no_pasg	-0.01776 0.6093 831
speed_ground speed_ground	0.86624 <.0001 831
speed_air speed_air	0.86780 <.0001 831
height height	0.09941 0.0041 831
pitch pitch	0.08703 0.0121 831

Therefore, we will plot the speed_ground vs distance plot (attached below) in order to see how this relationship is. There is in fact a positive relation between the 2 variables. In other terms, the higher the ‘speed_ground’ is, the more distance an aircraft needs to land. It is also worth mentioning that ‘speed_air’ will follow the same pattern as the graph below.



As we can observe, all the other variables (duration, no_pasg, pitch and height) aren’t correlated with distance. I have decided to show small plot of all of them just to show the random and uncorrelated pattern that they follow.



CHAPTER 3-Statistical modeling

Once we have finished the descriptive study and analyzed the relationship between the variables, it is time for statistical modeling actions.

For this, we will fit a linear regression model. We will start our analysis by summarizing in the next table the relationships between the independent variables and the response variable. It is important to notice that the variable make has been introduced for the regression analysis. This is a dummy variable with value 0 if he aircraft is an airbus and value 1 if the aircraft is a Boeing. As we saw in page 8 the landing distance is different for Airbus than for Boeing. But is this difference big enough to be significant in a regression model?

```
/**********introducing the dummy variable aircratft******/
data combined8;
set combined7;
if  aircraft="airbus" then make=0;
else make=1;
run;
proc print data=combined8;
run;
```

	Direction	Corr	β	Significance
Duration	-	-0.05138	-0.96133	
No.pasg		-0.01776	-2.12459	
S_ground	+	0.86624	41.44219	*
S_air	+	0.8678	41.45855	*
height	+	0.09941	9.10657	*
pitch	+	0.08703	148.1419	*
make	+	0.23814	427.6663	*

Assuming a significance level of 0.05, we get that the variables speed_ground, speed_air, height, pitch and make are significant. These values have been obtained from running individual regression analysis for each independent variable and with the variable distance as the response variable. Therefore, the only purpose of this table is to get an initial regression model containing the variables that are likely to be significant. By doing this we can clearly drop Duration and No_pasg from our final model because they don’t seem to have any correlation with distance.

Therefore, the initial model that we get is the following:

Distance= -2634.18 --9.08733*Speed_ground +51.47616*speed_air +13.97350*height +34.23338*height +481.36930*make

The following ANOVA table and parameter estimates table summarize this model regression analysis of the simulated 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	567652644	113530529	944.30	<.0001
Error	825	99187686	120227		
Corrected Total	830	666840329			

Root MSE	346.73837	R-Square	0.8513
Dependent Mean	1522.48287	Adj R-Sq	0.8504
Coeff Var	22.77453		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2634.18236	116.17392	-22.67	<.0001
speed_ground	speed_ground	1	-9.08733	15.94304	-0.57	0.5688
speed_air	speed_air	1	51.47616	15.91780	3.23	0.0013
height	height	1	13.97350	1.23327	11.33	<.0001
pitch	pitch	1	34.23338	24.51583	1.40	0.1630
make		1	481.36930	25.80388	18.65	<.0001

These tables provide us some useful information. For instance, considering a significance level of 5%, we can see that only 3 of the variables (speed_air, height and make) as well as the intercept are significant. Therefore, it seems that a better model needs to be built. Other important information is the R_sq which has a value of 0.8513. It is a fairly high value that means that the model fits the data fairly well. However, we will try to find a model that fits the data in a more accurate way.

It is interesting to highlight that while the correlation of both speed_ground and speed_air was highly positive with the response variable 'distance', the parameter estimate of speed_ground is negative. This occurs since speed_ground and speed_air are dependent of each other as we saw earlier. Therefore, we will need to check if a regression model with only one of these two variables fits the data better. We will try to keep speed_ground over speed_air because most of the values of the 'speed_air' variable are just predicted values (vs the real values from the 'speed_ground' variable).

Therefore, we are going to proceed to run a new model using the stepwise selection method. An important assumption that we have made is to remove the variable speed_air from the potential candidates to build the model. This is done because of the high dependency between speed_ground and speed_air.

```
/********************individual models***********/
proc reg data=combined8;
model distance=duration;
title regression analysis of the simulated dataset;
run;

proc reg data=combined8;
model distance=no_pasg;
title regression analysis of the simulated dataset;
run;
proc reg data=combined8;
model distance=speed_ground;
title regression analysis of the simulated dataset;
run;
proc reg data=combined8;
model distance=speed_air;
title regression analysis of the simulated dataset;
run;
proc reg data=combined8;
model distance=height;
title regression analysis of the simulated dataset;
run;
proc reg data=combined8;
model distance=pitch;
title regression analysis of the simulated dataset;
run;
proc reg data=combined8;
model distance=make;
title regression analysis of the simulated dataset;
run;
  /* correlation coefficients*/
 proc corr data=Combined8;
var distance;
with duration    no_pasg         speed_ground   speed_air       height   pitch
make;
title  correlation coeficients
run;

/* initial model*/
proc reg data=combined8;
model distance= speed_ground speed_air height pitch make;
title regression analysis of the simulated dataset;
run;
```

```
proc reg data=combined7;
model distance=   duration           no_pasg speed_ground     speed_air         height    pitch;
title regression analysis of the simulated dataset;
run;
```

A summary of the steps of the Stepwise selection method can be observed below.

Analysis of Effects Eligible for Entry				
Effect	DF	Score Chi-Square	Pr > ChiSq	Effect Label
speed_ground	1	1096.5774	<.0001	speed_ground
height	1	5.5349	0.0186	height
pitch	1	4.3103	0.0379	pitch
make	1	42.1931	<.0001	

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
1318.8381	4	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
speed_ground	1	-0.08425	0.00259	1055.2439	<.0001	0.919	speed_ground

The study starts with speed_ ground as it has the lowest p-value. It concludes that the variable speed_ground is significant and will be included in the final model.

Analysis of Effects Eligible for Entry				
Effect	DF	Score Chi-Square	Pr > ChiSq	Effect Label
height	1	107.5073	<.0001	height
pitch	1	36.9403	<.0001	pitch
make	1	350.6950	<.0001	

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
514.0598	3	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
speed_ground	1	-0.10466	0.00288	1319.5474	<.0001	0.901	speed_ground
make	1	-1.42745	0.08042	315.0892	<.0001	0.240	

The dummy variable make is studied next and the study also concludes that it is significant with speed_ground.

Analysis of Effects Eligible for Entry				
Effect	DF	Score Chi-Square	Pr > ChiSq	Effect Label
height	1	157.8905	<.0001	height
pitch	1	10.2180	0.0014	pitch

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
174.2663	2	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
speed_ground	1	-0.11555	0.00300	1482.7334	<.0001	0.891	speed_ground
height	1	-0.05058	0.00405	155.7761	<.0001	0.951	height
make	1	-1.55612	0.08135	365.8894	<.0001	0.211	

The variable height is studied next and the study also concludes that it is significant with speed_ground

Analysis of Effects Eligible for Entry				
Effect	DF	Score Chi-Square	Pr > ChiSq	Effect Label
pitch	1	15.6678	<.0001	pitch

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
15.6678	1	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
speed_ground	1	-0.11775	0.00309	1450.8141	<.0001	0.889	speed_ground
height	1	-0.05147	0.00405	161.2078	<.0001	0.950	height
pitch	1	-0.27499	0.06953	15.6427	<.0001	0.760	pitch
make	1	-1.52556	0.08260	341.1212	<.0001	0.217	

Finally, the variable Pitch is also included in the stepwise model as it is found to be significant. No other variables are included in the model and therefore the stepwise process is over. The conclusion is that all the potential variables that have been studied have been considered significant

```
proc phreg data=combined8;
    model distance=                 speed_ground              height    pitch make
                    / selection=stepwise slentry=0.05
                     slstay=0.05 details;
  run;
```

The model obtained using the stepwise method is the following:

Distance= -2664.32233+ 42.42833*speed_ground + 14.09086*height + 39.60761*Pitch + 481.26818*make

It is interesting to observe that all the variables of the model have a positive impact in the response variable. As a consequence, the greater the values of the independent values, the higher the landing distance and therefore the more risk of a landing overrun. It is also interesting to consider the impact of the variable make. Whenever make=1, in other terms, whenever the aircraft is a Boeing, the landing distance will be in average 481 more ft. than for Airbus. So a landing overrun in a Boeing is more likely than in an Airbus.

Below, there is a summary of the ANOVA table as well as the parameter estimates table of this improved model.

regression analysis of the simulated 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	4	566395312	141598828	1164.42	<.0001
Error	826	100445017	121604		
Corrected Total	830	666840329			

Root MSE	348.71785	R-Square	0.8494
Dependent Mean	1522.48287	Adj R-Sq	0.8486
Coeff Var	22.90455		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2664.32233	116.46055	-22.88	<.0001
speed_ground	speed_ground	1	42.42833	0.64788	65.49	<.0001
height	height	1	14.09086	1.23977	11.37	<.0001
pitch	pitch	1	39.60761	24.59908	1.61	0.1078
make		1	481.26818	25.95117	18.55	<.0001

```
proc reg data=combined8;
model distance=          speed_ground          height    pitch make;
title regression analysis of the simulated dataset;
run;
```

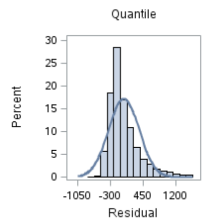
The most important facts to consider from the parameter estimates table is the significance value of the Pitch variable. While it is not below 0.05 as it is required to be significant, it still has a fairly big impact in the response variable (even the stepwise method found it significant). If we wouldn't include it, the goodness of fit of this model will be reduced

The R-square also needs in fact to be considered. The value is almost the same as in the full model. We can therefore conclude that the elimination of speed_air hasn't impacted on how well the model fits the data.

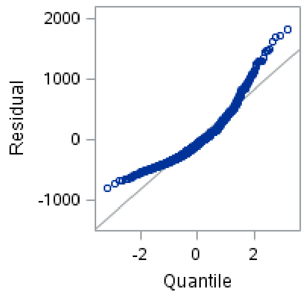
Consequently, we can conclude that the improved model is valid through the model diagnostics, the regression analysis study will be over.

Two brief conclusions that we can derivate from model diagnostics plots are the following:

- The residuals don't quite follow a normal distribution. In fact, this distribution is a little skewed to the right

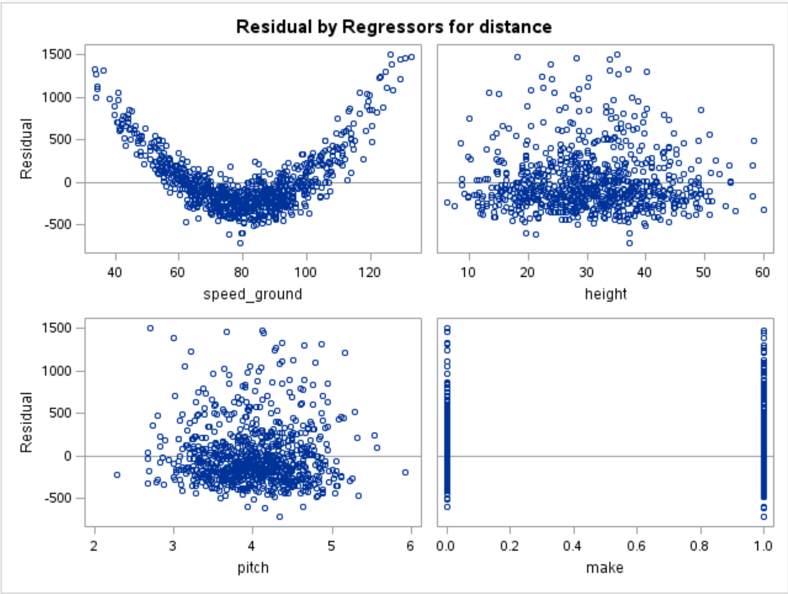


- The QQ plot also demonstrates the normal distribution of the residuals is not quite met. Therefore the normality assumption is violated




```
proc reg data=combined8;
model distance=          speed_ground   height   pitch make/r;
title regression analysis of the simulated dataset;
output out=diagnostics r=residual;
run;
```

Therefore, we will study the model diagnostics of the individual variables, to see if we can build a model that better meets the model diagnostics requirements



As seen in the figure above, the residuals of both height, pitch and make don't look to follow any pattern. However, speed_ground seems to follow a pattern. By introducing a quadratic term in this variable, we can probably get rid of this pattern.

By introducing the quadratic term to the equation, the new model that we get is the following:

Distance= -1689.16+ -69.54*speed_ground + 0.6957* speed_ground ^2 + 13.42*height + 32.36*Pitch + 387.30*make

As shown in the figure below, this model results in a extremely high r^2 (0.9765) where all its variables are significant (p value lower than 0.05). As a consequence, we have found a better model than the previous ones and if we can validate it through the model diagnostics, we would have finished our regression analysis.

Analysis of Variance

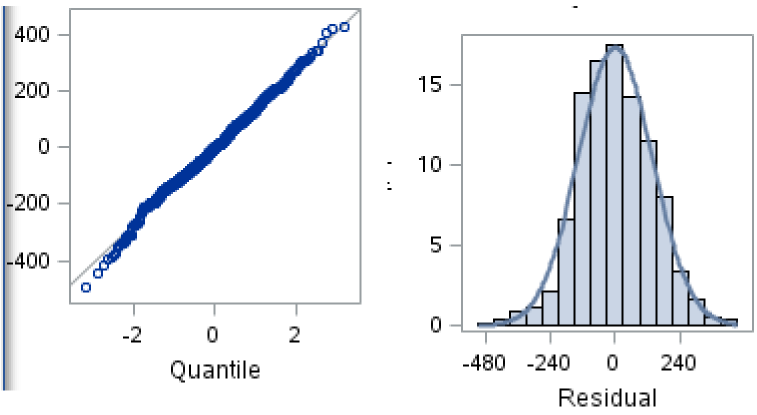
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	651142224	130228445	6844.04	<.0001
Error	825	15698105	19028		
Corrected Total	830	666840329			

Root MSE	137.94204	R-Square	0.9765
Dependent Mean	1522.48287	Adj R-Sq	0.9763
Coeff Var	9.06033		

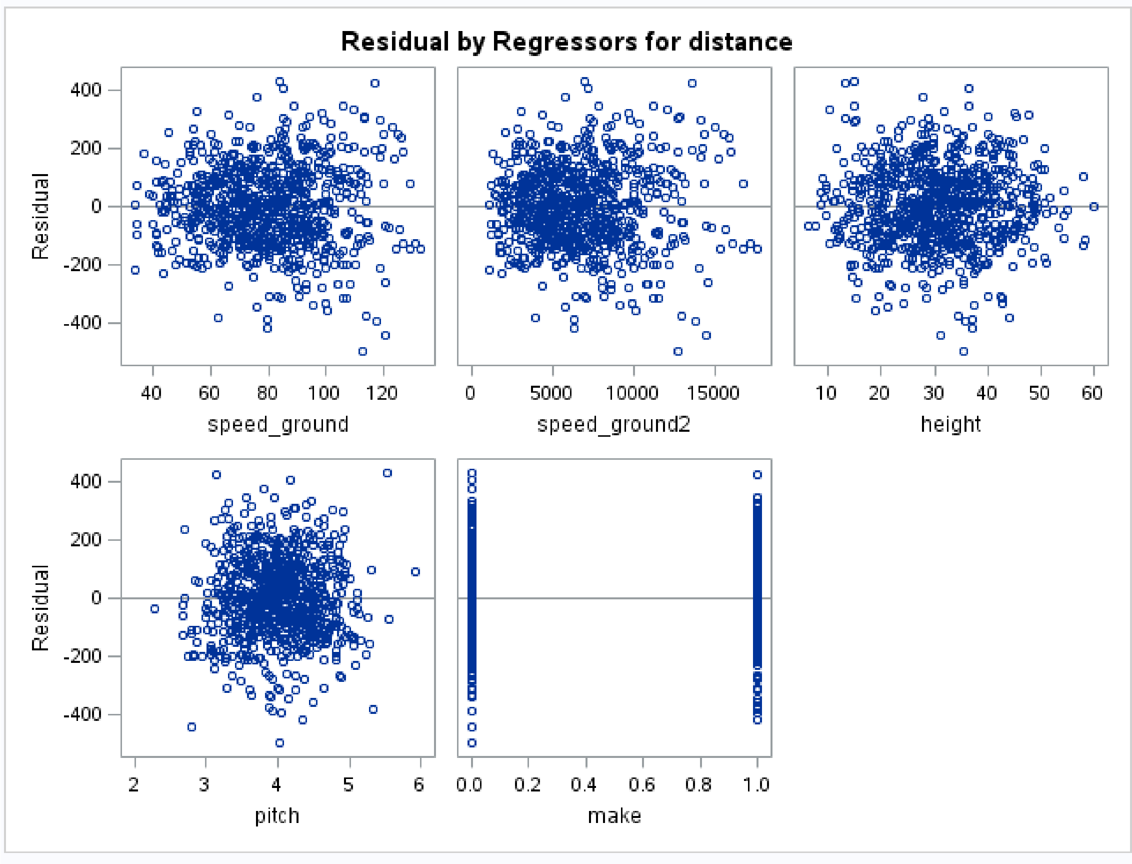
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1689.15718	79.86050	21.15	<.0001
speed_ground	speed_ground	1	-69.53675	1.69717	-40.97	<.0001
speed_ground2		1	0.69572	0.01042	66.74	<.0001
height	height	1	13.41839	0.49052	27.36	<.0001
pitch	pitch	1	32.35946	9.73124	3.33	0.0009
make		1	387.30041	10.36160	37.38	<.0001

The model diagnostics for this model looks the following:



Consequently, we can conclude that the residuals are normally distributed around 0 and the normality assumption is met as we cannot see any pattern in the QQ plot. The figure below also shows that no patterns are found when we study the residuals by individual regressors. We can therefore conclude that we have found a valid model.



```
/********* QUADRATIC MODEL*********/

DATA combined9;

SET combined8;

FORMAT speed_ground2;

speed_ground2 = speed_ground**2;

RUN;


proc reg data=combined9;

model distance=     speed_ground  speed_ground2     height  pitch make/r;

title regression analysis of the simulated dataset;

output out=diagnostics r=residual;

run;
```

CONCLUSIONS

We can finally conclude that the factors that have the most impact on landing overrun are the speed ground, speed ground^2, the height, the pitch angle and the type of aircraft ('make' variable). So, to reduce the risk of landing overrun we must make sure that when substituting the values for these 5 independent variables, the distance is not greater than 6000.

The final model that allows to calculate this distance is the following:

Distance= -1689.16+ -69.54*speed_ground + 0.6957* speed_ground ^2 + 13.42*height + 32.36*Pitch + 387.30*make

By creating a system that instantaneously alerts the pilot when the 'distance' value is expected to be>6000, landing overruns could be drastically reduced.