Anh Nguyen

Professor Liu

Statistical Computing

19 November 2019

FACTORS THAT AFFECT FLIGHT LANDING

AND HOW TO REDUCE THE RISK OF LANDING OVERRUN

Executive Summary: After cleaning and analyzing the 2 data set FAA1 and FAA2, we've found that the 2 variables that affect the landing distance significantly are:

- **Speed_ground** (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway.
- Speed_air (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway.

According to this equation:

$$\text{Distance} = -5425.5 + 91.63\ \text{Air speed} - 12.319\ \text{Ground speed (feet)}$$

From this formula, we can put a limit of landing distance, then consider the level of air speed and ground speed most optimized for safety and time-efficient. Thus we would be able to use this design policies fittingly.

Chapter 1: Data Understanding and Exploration

1. The goal: First, the objective is to wrangle data so it'll become clean enough for further operations. Next, it's to have an deep understanding of the datasets we're given in every aspect. We try to understand every single variables in the dataset, its mean, median,…

2. The code:

- Step 1: Import data and Combine data sets from different sources:

```
TITLE "CHAPTER 1: ";
*1 Combining data sets from different sources;

proc import datafile='/home/u42905097/Class 2/FAA1.xls' out=FAA1
dbms = xls
replace;
getnames=yes;
run;

proc import datafile='/home/u42905097/Class 2/FAA2.xls' out=FAA2
dbms = xls
replace;
getnames=yes;
run;

proc sort data=faa1;
by Aircraft;
run;

proc sort data = faa2;
by Aircraft;
run;

data FAACombined;
set FAA1 FAA2;
by Aircraft;
run;

options missing = '';
data Faacombined1;
  set Faacombined;
  if missing(cats(of _all_)) then delete;
run;

proc sort data=faacombined1 nodupkey;
by speed_ground height;
run;
```

proc sort data=faacombined1;
by aircraft;
run;

title '1. Combination';
proc print data=FAACombined1;
run;

proc contents data= faacombined1;
run;

=> The output:

- A piece of combined FAA data set sorted by Aircraft:

284	airbus	128.5872	62	86.2395		10.75378	4.113208	1158.589
285	airbus	91.30575	53	86.35629		27.529	3.862561	1381.56
286	airbus	94.42954	61	86.41954		25.40328	4.719329	1561.03
287	airbus		54	86.42505		14.74857	3.541838	1476.178
288	airbus	94.82857	49	86.48323		29.28925	3.781126	1555.526
289	airbus		56	86.52884		40.94902	3.727026	1437.634
290	airbus	164.6229	75	86.60301		27.66436	4.579674	1366.383
291	airbus	171.4373	55	86.77431		25.91501	3.980676	1188.164
292	airbus	115.8692	45	86.87588		34.83807	3.799768	1262.154
293	airbus	178.337	55	87.01596		38.19466	4.355066	1693.403
294	airbus	130.9023	62	87.06464		36.29385	3.864921	1606.312
295	airbus	156.9548	55	87.50801		30.62009	3.59257	1548.474
296	airbus	160.8416	48	87.55147		28.54558	4.422983	1410.189
297	airbus	150.0304	67	87.57164		17.53825	3.022595	1219.624
298	airbus	132.2714	50	87.91721		37.78704	3.742716	1644.089
299	airbus	87.34597	59	87.92651		28.7909	4.011354	1555.401
300	airbus	179.1942	52	87.93797		18.54648	3.874064	1243.035
301	airbus	147.2498	70	87.99743		34.08079	4.401252	1467.7
302	airbus	176.0849	58	88.01342		16.6981	4.069899	1380.274
303	airbus	153.0251	56	88.10262		21.24791	3.155941	1296.224
304	airbus	117.9727	62	88.11053		28.39527	3.837378	1539.761
305	airbus	144.8292	53	88.20592		27.99095	3.09096	1259.174
306	airbus	191.3616	62	88.39424		28.34876	4.362478	1465.259
307	airbus		57	88.4181		45.02439	3.703694	1616.336
308	airbus	99.34654	51	88.51343		29.68572	3.923984	1446.573
309	airbus	168.6394	67	88.52277		36.18514	3.82789	1574.196

- Contents of the file:

1. Combination

The CONTENTS Procedure

Data Set Name	WORK.FAACOMBINED1	Observations	850
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	18:06:35 22/11/2019	Observation Length	64
Last Modified	18:06:35 22/11/2019	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	YES
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	2043
Obs in First Data Page	850
Number of Data Set Repairs	0
Filename	/saswork/SAS_work67090000BF68_odaws02-prod-us.oda.sas.com/SAS_workFFA30000BF68_odaws02-prod-us.oda.sas.com/faacombined1.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	760269
Access Permission	rw-r--r--
Owner Name	u42905097
File Size	256KB

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	8	$8.	$8.	aircraft
8	distance	Num	8	BEST8.		distance
2	duration	Num	8	BEST8.		duration
6	height	Num	8	BEST8.		height
3	no_pasg	Num	8	BEST8.		no_pasg
7	pitch	Num	8	BEST8.		pitch
5	speed_air	Num	8	BEST8.		speed_air
4	speed_ground	Num	8	BEST8.		speed_ground

Sort Information	
Sortedby	aircraft
Validated	YES
Character Set	ASCII

=> Observations:

- The combined data set there are 8 variables and 1000 observations in total. After deleting all empty observations and all duplicating observations, we're down to 850 unique observations.

Step 2: Performing the completeness check of each variable

title '2. Completeness check';

proc means data = FAACombined1 n nmiss;
var duration no_pasg Speed_ground Speed_air Height Pitch Distance;
run;

=> The output:

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

=> The observations:
- We found out the statistics on how many observations have got which variables. For example, all 850 observations have their speed_ground values.
- However, for durations and speed_air, there are many missing data points:

Variable	Number of Obs with value	Number of Obs w/o value	% of obs with variable	% of obs w/o variable
Duration	800	50	94.12%	5.88%
Speed_air	208	642	24.47%	75.53$

Step 3: Validity check according to presented conditions, examine if abnormal values are presented

**3. Performing the validity check of each variable – examine if abnormal values are present;*
title '3. Validity check';
data FAAValidity;
set faacombined1;
if Duration>40 then ValidDuration = 1;
else ValidDuration = 0;
if Speed_ground <30 or Speed_ground>140 then ValidSpeed_ground = 0;
else ValidSpeed_ground = 1;
if (Speed_air >=30 and Speed_air<=140) or Speed_air=. then ValidSpeed_air = 1;
**if Speed_air=. then ValidSpeed_air = 1;*
else ValidSpeed_air = 0;
if Height>6 then ValidHeight = 1;
else ValidHeight = 0;
if Distance<6000 then ValidDistance = 1;

else ValidDistance = 0;
run;

proc print data = FAAValidity;
run;

=> The output: Part of the code output

3. Validity check

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance	ValidDuration	ValidSpeed_ground	ValidSpeed_air	ValidHeight	ValidDistance
1	airbus	192.2829	64	33.5741		36.97069	4.358464	782.7174	1	1	1	1	1
2	airbus	126.0784	54	36.42139		33.7997	4.866111	869.0337	1	1	1	1	1
3	airbus		46	40.80179		24.40013	3.968209	620.0905	0	1	1	1	1
4	airbus	100.2531	61	41.10099		34.56394	3.490939	668.9332	1	1	1	1	1
5	airbus	128.6879	56	43.85281		34.11456	3.312578	554.0662	1	1	1	1	1
6	airbus	169.7494	65	43.92356		43.05077	3.365144	735.8234	1	1	1	1	1
7	airbus	134.6856	57	44.12614		27.18238	3.012829	417.5431	1	1	1	1	1
8	airbus	191.3109	66	44.25848		49.28774	3.69714	901.4067	1	1	1	1	1
9	airbus	42.14623	63	46.26472		20.49071	3.481912	383.5585	1	1	1	1	1
10	airbus	172.0493	36	47.48677		13.98481	4.29902	250.6898	1	1	1	1	1
11	airbus	117.7406	59	47.6798		28.60649	3.752047	406.0893	1	1	1	1	1
12	airbus	190.7394	77	47.88212		14.83596	2.732284	41.72231	1	1	1	1	1
13	airbus	92.17284	58	48.1728		34.75702	3.948344	558.3651	1	1	1	1	1
14	airbus	248.7291	58	49.46213		16.79618	3.761689	452.1313	1	1	1	1	1
15	airbus	209.1937	54	50.81293		38.84132	4.033898	566.9269	1	1	1	1	1
16	airbus		54	50.90311		35.72948	4.54404	597.9855	0	1	1	1	1
17	airbus	142.5876	66	51.15823		8.559069	3.913448	242.5959	1	1	1	1	1
18	airbus	212.054	63	51.58704		20.45129	3.063686	133.0869	1	1	1	1	1
19	airbus		59	52.09981		34.55297	4.88663	647.4882	0	1	1	1	1
20	airbus	133.7396	50	52.33337		18.0381	3.935193	513.4958	1	1	1	1	1
21	airbus	182.4478	66	52.70784		24.30264	4.185967	317.8127	1	1	1	1	1
22	airbus	60.53364	64	52.98413		23.865	4.322738	487.4687	1	1	1	1	1
23	airbus	106.3736	66	53.37241		51.00335	2.827557	538.9741	1	1	1	1	1
24	airbus	183.6185	69	53.53924		31.73942	3.523775	349.1585	1	1	1	1	1
25	airbus	98.1763	60	53.74913		25.54579	3.714201	378.8258	1	1	1	1	1

=> The observations:
- I encoded each variables with a new variable as 1 or 0 according to the conditions to consider which on is valid or invalid.

Step 4: Clean Data according to step 2 and step 3:

**4. Cleaning the data based on the results of Steps 2 and 3;*
title '4. Data Cleaning';
data FAACleaned;
set FAAValidity;
if ValidDuration=1 and ValidSpeed_ground=1 and ValidSpeed_air=1 and ValidHeight=1 and ValidDistance = 1;
drop ValidDuration ValidSpeed_ground ValidSpeed_air ValidHeight ValidDistance;
run;

proc print data = FAACleaned;
run;

proc contents data = FAACleaned;
run;

=> The output:

4. Data Cleaning

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	airbus	192.2829	64	33.5741		36.97069	4.358464	782.7174
2	airbus	126.0784	54	36.42139		33.7997	4.866111	869.0337
3	airbus	100.2531	61	41.10099		34.56394	3.490939	668.9332
4	airbus	128.6879	56	43.85281		34.11456	3.312578	554.0662
5	airbus	169.7494	65	43.92356		43.05077	3.365144	735.8234
6	airbus	134.6856	57	44.12614		27.18238	3.012829	417.5431
7	airbus	191.3109	66	44.25848		49.28774	3.69714	901.4067
8	airbus	42.14623	63	46.26472		20.49071	3.481912	383.5585
9	airbus	172.0493	36	47.48677		13.98481	4.29902	250.6898
10	airbus	117.7406	59	47.6798		28.60649	3.752047	406.0893
11	airbus	190.7394	77	47.88212		14.83596	2.732284	41.72231
12	airbus	92.17284	58	48.1728		34.75702	3.948344	558.3651
13	airbus	248.7291	58	49.46213		16.79618	3.761689	452.1313
14	airbus	209.1937	54	50.81293		38.84132	4.033898	566.9269
15	airbus	142.5876	66	51.15823		8.559069	3.913448	242.5959
16	airbus	212.054	63	51.58704		20.45129	3.063686	133.0869
17	airbus	133.7396	50	52.33337		18.0381	3.935193	513.4958
18	airbus	182.4478	66	52.70784		24.30264	4.185967	317.8127
19	airbus	60.53364	64	52.98413		23.865	4.322738	487.4687
20	airbus	106.3736	66	53.37241		51.00335	2.827557	538.9741
21	airbus	183.6185	69	53.53924		31.73942	3.523775	349.1585
22	airbus	98.1763	60	53.74913		25.54579	3.714201	378.8258
23	airbus	237.4053	48	53.77401		28.2608	3.17553	241.161

=> The observations:

- I deleted all the observations that are abnormal and missing the “durations” values. I didn’t delete the observations that miss the “speed_air” values, since one of the dataset misses all “speed_air” values, therefore, it’ll make no sense deleting entire dataset.
- The end result of the cleaning: There are 781 observations left

4. Data Cleaning

The CONTENTS Procedure

Data Set Name	WORK.FAACLEANED	Observations	781
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	14:46:52 24/11/2019	Observation Length	64
Last Modified	14:46:52 24/11/2019	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Step 5: Summarizing distributions of each variables:

```
*5. Summarizing the distribution of each variable (what tables and figures will you present?);
title '5. Distributions of variable';
proc sort data = FAACleaned;
by aircraft;
run;

proc means data = FAACleaned noprint;
by aircraft;
output out=FAASummaryMean
mean(duration speed_ground) = MeanDuration MeanSpeed_Ground
mean(no_pasg Speed_air Height Pitch Distance) = MeanNo_Pasg MeanSpeed_Air MeanHeight
MeanPitch MeanDistance;
run;

proc means data = FAACleaned noprint;
by aircraft;
output out=FAASummaryStddev
std(duration speed_ground) = StdDuration StdSpeed_Ground
std(no_pasg Speed_air Height Pitch Distance) = StdNo_Pasg StdSpeed_Air StdHeight StdPitch
StdDistance;
run;

proc print data = FAASummaryMean;
title Summary Mean for variables;
run;

proc print data = FAASummaryStddev;
title Summary Standard Deviation for variables;
run;

%macro summary(dataset, variable);
proc means data = &dataset n nmiss mean median std min max;
var &variable;
title Summary Statistics for &variable in &dataset;
run;
%mend summary;

%summary(FAACleaned, duration);
%summary(FAACleaned, speed_ground);
%summary(FAACleaned, speed_air);
%summary(FAACleaned, height);
%summary(FAACleaned, pitch);
%summary(FAACleaned, distance);
```

=> The output:

Summary Mean for variables

Obs	aircraft	_TYPE_	_FREQ_	MeanDuration	MeanSpeed_Ground	MeanNo_Pasg	MeanSpeed_Air	MeanHeight	MeanPitch	MeanDistance
1	airbus	0	394	156.9033	80.53199	60.2868	104.4454	30.60011	3.826832	1335.152
2	boeing	0	387	152.6096	78.73137	59.87339	102.891	30.30717	4.204813	1750.983

Summary Standard Deviation for variables

Obs	aircraft	_TYPE_	_FREQ_	StdDuration	StdSpeed_Ground	StdNo_Pasg	StdSpeed_Air	StdHeight	StdPitch	StdDistance
1	airbus	0	394	49.18829	17.06018	7.486475	8.330329	9.774034	0.485669	802.8381
2	boeing	0	387	47.44672	20.5825	7.570531	10.76242	9.71492	0.488855	953.85

Summary Statistics for duration in FAACleaned

The MEANS Procedure

Analysis Variable : duration duration						
N	N Miss	Mean	Median	Std Dev	Minimum	Maximum
781	0	154.7757191	154.2845505	48.3499237	41.9493694	305.6217107

Summary Statistics for speed_ground in FAACleaned

The MEANS Procedure

Analysis Variable : speed_ground speed_ground						
N	N Miss	Mean	Median	Std Dev	Minimum	Maximum
781	0	79.6397499	79.7939604	18.8971690	33.5741041	132.7846766

Summary Statistics for speed_air in FAACleaned

The MEANS Procedure

Analysis Variable : speed_air speed_air						
N	N Miss	Mean	Median	Std Dev	Minimum	Maximum
195	586	103.5047686	100.8916770	9.8803757	90.0028586	132.9114649

Summary Statistics for height in FAACleaned

The MEANS Procedure

Analysis Variable : height height						
N	N Miss	Mean	Median	Std Dev	Minimum	Maximum
781	0	30.4549525	30.2165682	9.7396415	6.2275178	59.9459639

Summary Statistics for pitch in FAACleaned

The MEANS Procedure

Analysis Variable : pitch pitch						
N	N Miss	Mean	Median	Std Dev	Minimum	Maximum
781	0	4.0141289	4.0140064	0.5223688	2.2844801	5.9267842

Summary Statistics for distance in FAACleaned

The MEANS Procedure

Analysis Variable : distance distance						
N	N Miss	Mean	Median	Std Dev	Minimum	Maximum
781	0	1541.20	1273.66	904.5903306	41.7223127	5381.96

=> The observations:

- I produce the statistics mean and standard deviations for each variables in Boeing and Airbus
- Then I produce the statistics number of obs, missing obs, mean, median, std dev, minimum and maximum for each variables in general.

=> Conclusions: After cleaning for abnormal and missing values, there are 781 observations left.

Chapter 2: Descriptive Study

The goal: The objective is to understand the correlations between each variables with the landing distance of the planes. Then we would be able to choose which variables to run regression with

The code:

Step 1: Plot

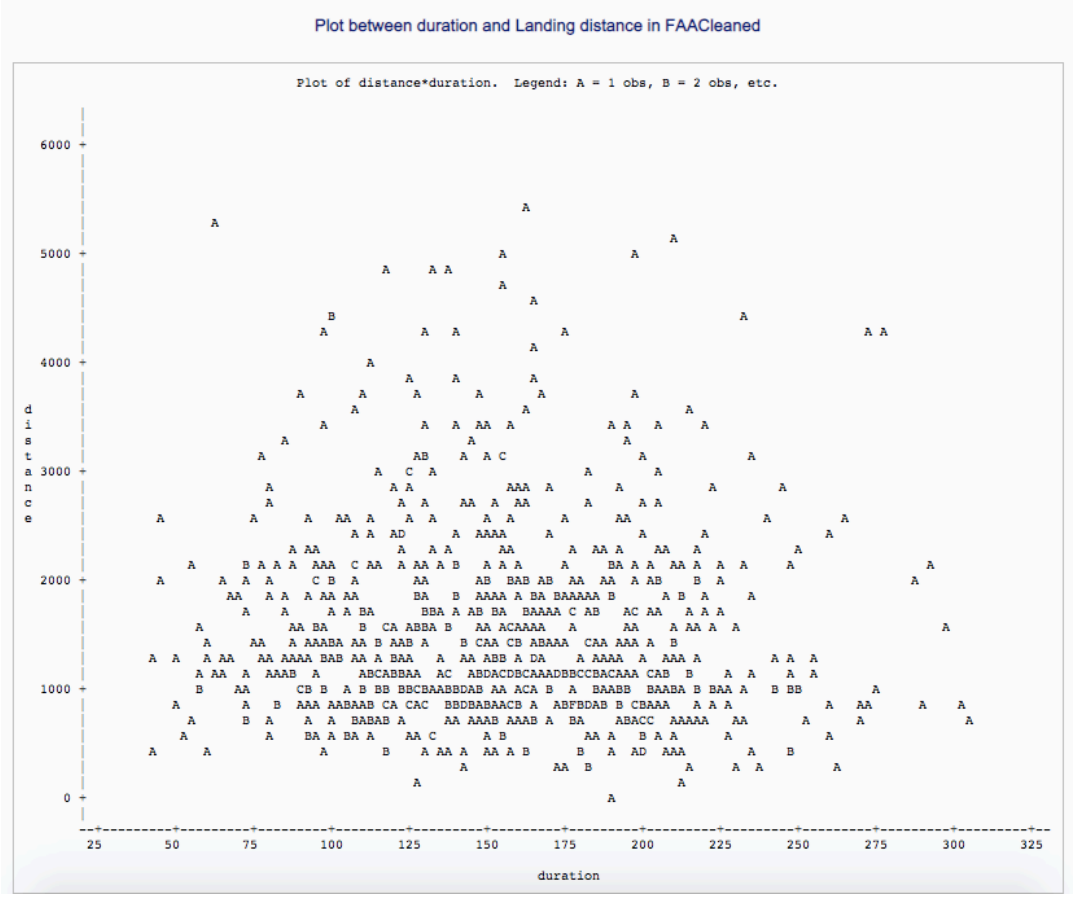
```
TITLE "CHAPTER 2: ";

%macro xyplot(dataset, variable);
proc plot data = &dataset;
plot distance*&variable;
title Plot between &variable and Landing distance in &dataset;
```

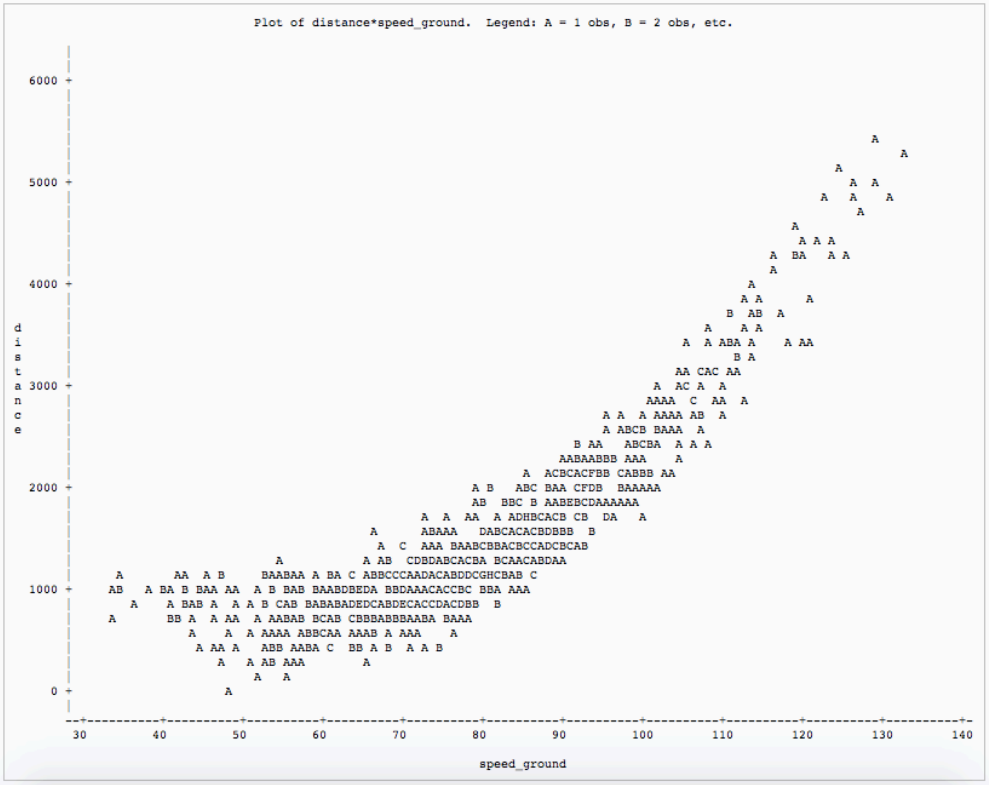
run;
%mend xyplot;

%xyplot(FAACleaned, duration);
%xyplot(FAACleaned, speed_ground);
%xyplot(FAACleaned, speed_air);
%xyplot(FAACleaned, height);
%xyplot(FAACleaned, pitch);
%xyplot(FAACleaned, no_pasg);

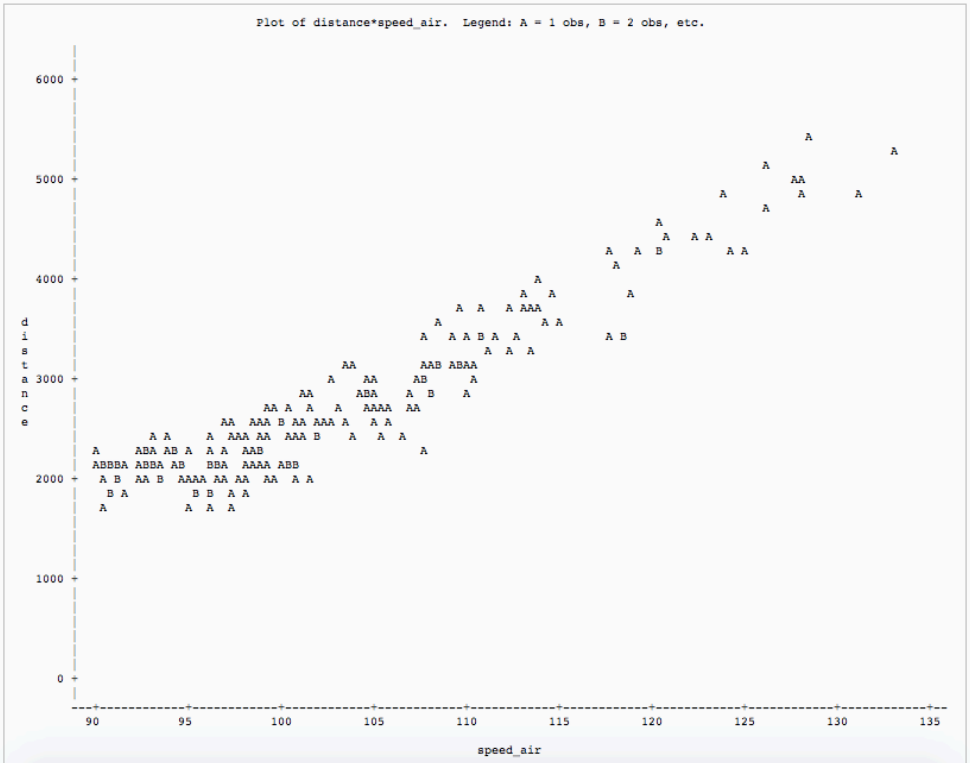
=>The output:



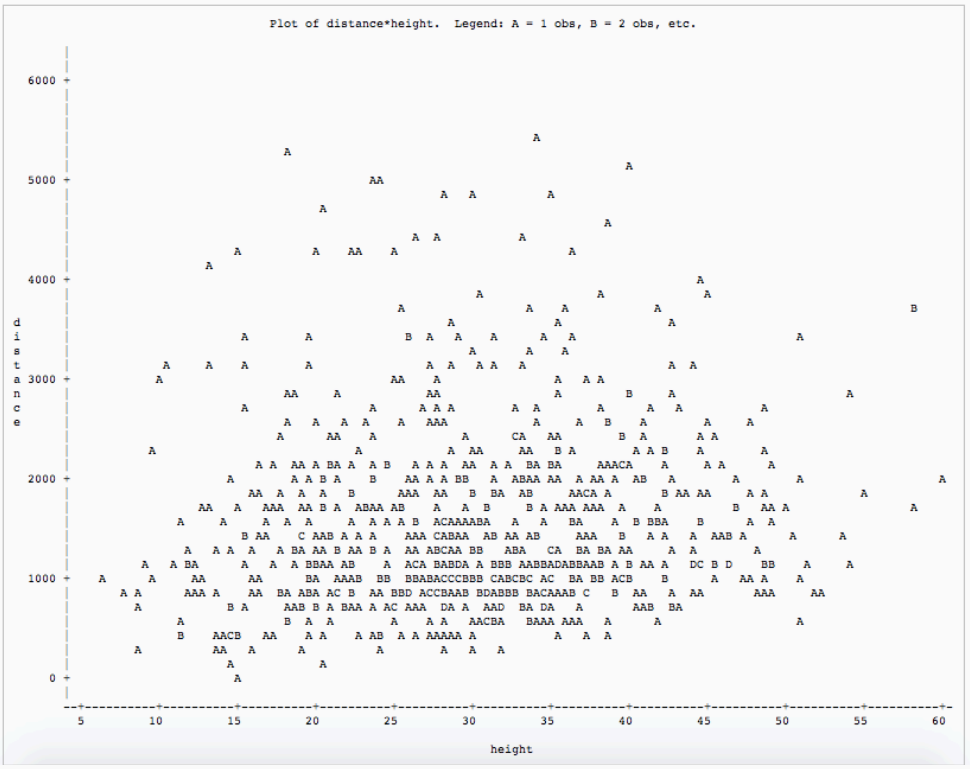
Plot between speed_ground and Landing distance in FAACleaned



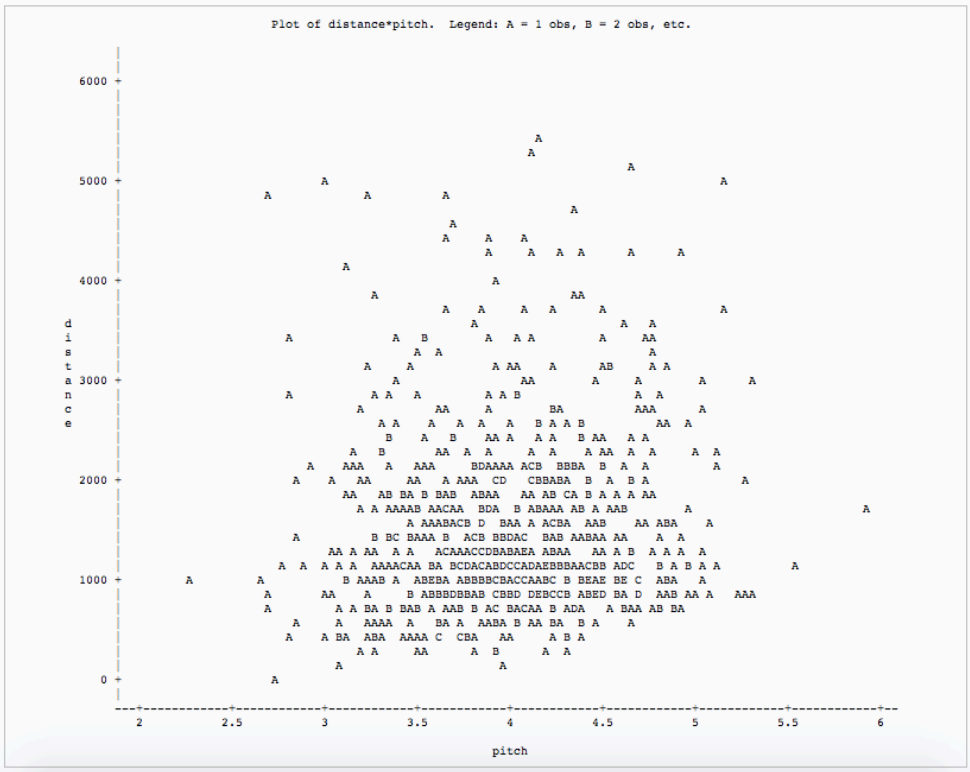
Plot between speed_air and Landing distance in FAACleaned

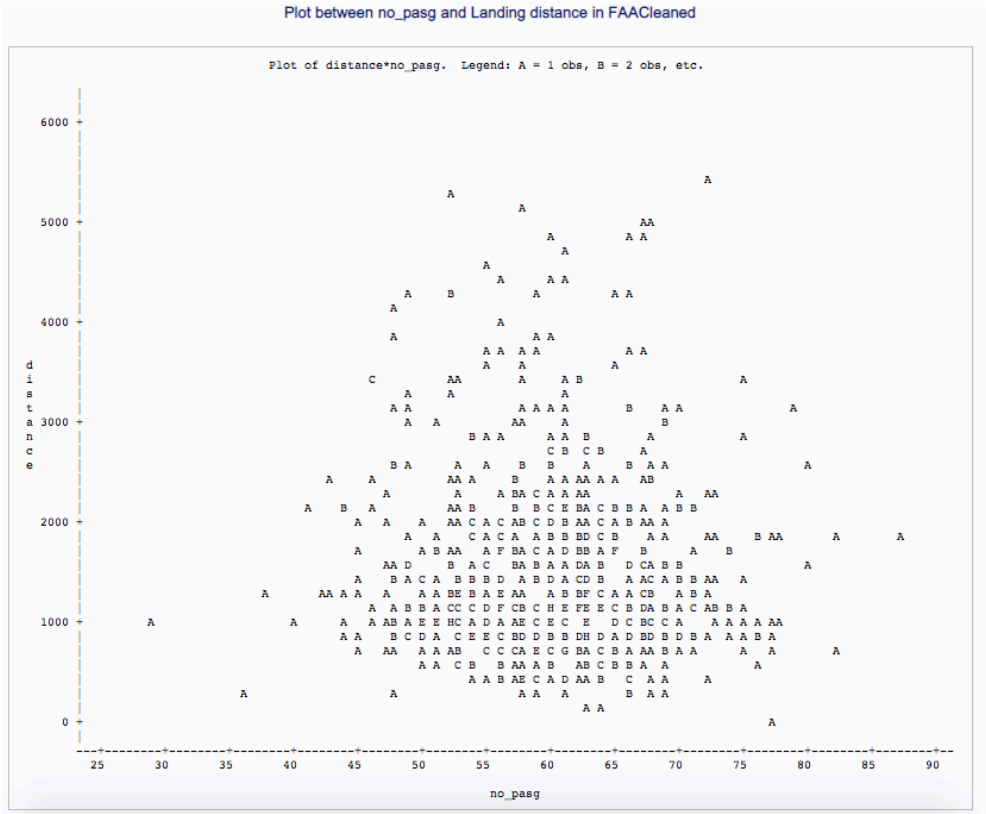


Plot between height and Landing distance in FAACleaned



Plot between pitch and Landing distance in FAACleaned





=> **Observations:** Preliminary observations show that only speed_air and speed-ground have any correlation with the landing distance

Step 2: Correlations through the figures

proc corr data=faacleaned;
var duration speed_ground no_pasg Speed_air Height;
with distance;
title Correlation Coefficient with Distance;
run;

=> **The output:**

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

	duration	speed_ground	no_pasg	speed_air	height
distance	-0.05138	0.86771	-0.01685	0.94322	0.10372
distance	0.1514	<.0001	0.6382	<.0001	0.0037
	781	781	781	195	781

=> **Observations/Conclusion:**
- We found out that the P value of correlations for each variables with Distance, with the Null hypothesis being: “There is no correlations between the 2 variables.” If we take a 95% significant level to consider the hypothesis, there are 3 variables with p-value<0.05, which are speed_air, height and speed_ground.

- For height, the correlation coefficient is 0.104, very closed to 0, so there's barely any correlation at all.
- For speed_air and speed_ground, their corr coeff are positive and closed to 1. This suggests a linear relationship between the variables and Distance.
=> We will conduct linear regression on these 2 variables.

Chapter 3: Statistical Modeling
The goal: To study the specific formula describing how the speed of the airplane would affect the landing distance.
The code:

TITLE "CHAPTER 3: ";
proc reg data = faacleaned;
model distance = speed_air speed_ground;
*plot distance*speed_air="*" distance*speed_ground="+"/overlay;*
run;

=> The output:

CHAPTER 3:

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	781
Number of Observations Used	195
Number of Observations with Missing Values	586

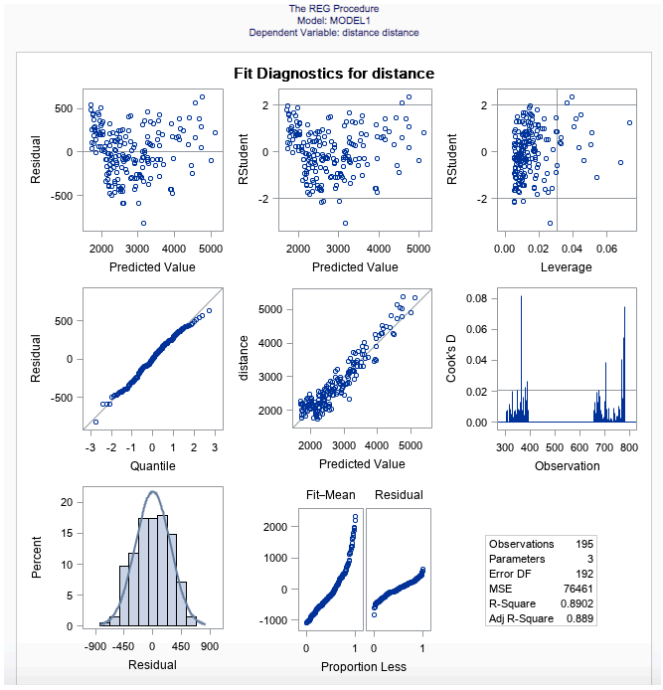
Analysis of Variance

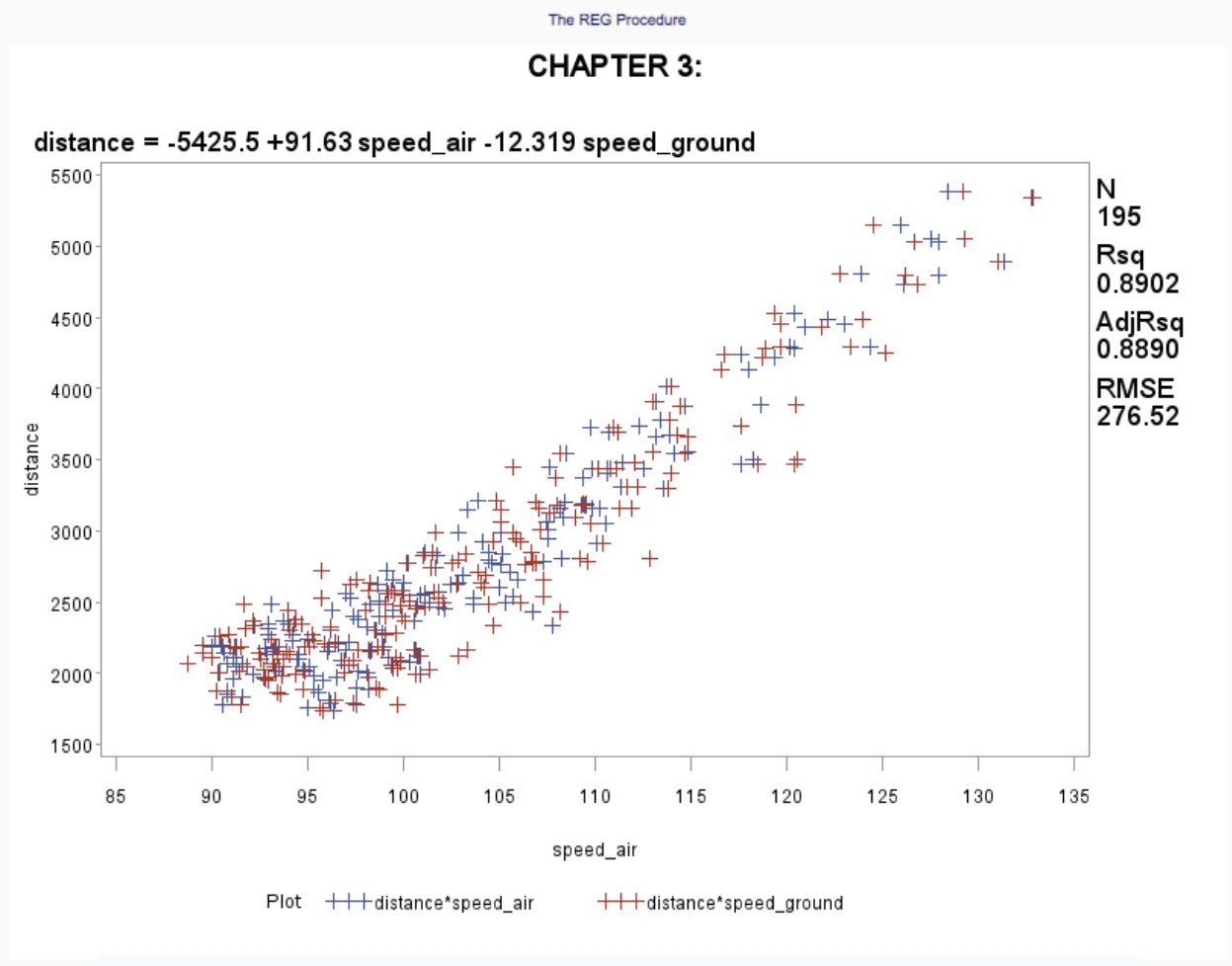
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	118995206	59497603	778.14	<.0001
Error	192	14680603	76461		
Corrected Total	194	133675809			

Root MSE	276.51668	R-Square	0.8902
Dependent Mean	2784.49158	Adj R-Sq	0.8890
Coeff Var	9.93060		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5425.48760	209.07862	-25.95	<.0001
speed_air	speed_air	1	91.62992	13.20051	6.94	<.0001
speed_ground	speed_ground	1	-12.31906	12.97593	-0.95	0.3436





=> Observations:

- We found that the P-value are less than 0.001 for intercept and speed_air correlations, however P-value for speed_ground 0.34> 0.05, so we may not be able to reject the hypothesis that this variable is not statistically significant, and therefore does not affect the change in landing distance. We will have to model check after this step.
- R-squared = 0.8902, thus the model can explain roughly 90% of all observations, which is pretty accurate
- The coefficients for speed_air and speed_ground respectively are 91.63 and -12.31.
- There doesn't seem to be any concerning pattern with residual plot that begs the refitting of the model.

=> Conclusion: Landing distance of planes are determined by the speed of the aircraft on the air and its speed on the ground, according to this linear formula:

Distance = -5425.5 + 91.63 Air speed – 12.319 Ground speed (feet)

This can be understood as, holding all else equal, if you increase the plane speed on the air when crossing the threshold of the runway by 1 mile/hour, the landing distance increase by 91.63 feet. If you increase the plane speed on the ground after crossing the threshold of the runway by 1mile/hour, the distance decrease by 12.319 feet.

**** Disclaimer: It maybe that I don't quite understand the prompt of the problem, but the second statement makes no sense. Is it also possible that it's because the variable is statistically insignificant? Either way, I'm going to keep the report the same for the sake of the submission.

Short answers to the questions:
1. How many flight do you use to fit your models?
- As shown, only 195 observations were used, not all 781 obs. This is due to the fact that a large portion of the dataset miss their values on speed_air. Also, some more observations were cleansed due to containing abnormal values.

2. What factors and how do they impact the landing distance of the flight?
- (See Chapter 3/Conclusion)

3. Is there any difference between the 2 makers Boeing and Airbus?
- (See the summary of Means and Std Dev of statistics on Boeing and Airbus planes at the end of the Chapter 1)
- We found there are roughly 400 feet difference between the sample mean of Airbus and Boeing's landing distance. We need to run a hypothesis on this, but it may be very clear already that Boeing aircrafts can land in significantly shorter distance compared to Airbus aircraft.