IN-CLASS PROJECT

Goal: the goal of the project is to study the factors and how they would impact the landing distance of a commercial flight

Step 1: Data Preparation

Function: on the given two data sets, we do

Data Exploration: Two files FAA1.xls and FAA2.xls, FAA1 contains aircraft, duration, no_pasg, speed_ground, speed_air, height, pitch and distance and FAA2 contains same fields but missing duration values.
Data Cleaning: Goal to form consolidated data set by combining two data sets and cleaning task irrelevant data
On FAA1.xls, we apply the following cleaning conditions

if duration>40 and height>6;
if speed_air>30 and speed_air<140;
if speed_ground >30 and speed_ground <140;

On FAA2.xls, we apply the following cleaning conditions
if height>6;
if speed_air>30 and speed_air<140;
if speed_ground >30 and speed_ground <140;

if duration values are missing, we should not delete all observations because of missing values in duration.

SAS Code:

```
/* importing data from FAA1.xls*/
proc import out=FAA1
datafile='/folders/myfolders/FAA1.xls'
dbms=xls replace;
getnames=yes;
run;
/* importing data from FAA2.xls*/

proc import out=FAA2
datafile='/folders/myfolders/FAA2.xls'
dbms=xls replace;
getnames=yes;
run;

/* Data Cleaning in FAA1 xls*/
DATA FAA1_New;
set FAA1;
if duration>40 and height>6;
if speed_air>30 and speed_air<140;
if speed_ground >30 and speed_ground <140;
if distance <6000;
```

```
RUN;

/*Data Cleaning in FAA2 xls,duration values are missing we do not apply condition that we need to remove all
observations duration>40  */
DATA FAA2_New;
set FAA2;
if  height>6;
if speed_air>30 and speed_air<140;
if speed_ground >30 and speed_ground <140;
if distance <6000;
RUN;
/*Merge two data sets*/
PROC SORT data=FAA1_New;
BY aircraft; /*sorts within the first data set*/
PROC SORT data=FAA2_New;
BY aircraft; /*sorts within the second data set*/
DATA CombinedFAA;
SET FAA1_New FAA2_New;
BY aircraft;
RUN;
PROC SORT DATA=CombinedFAA;
BY aircraft;
RUN;
PROC print DATA=CombinedFAA;
RUN;
```

Output:

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	airbus	132.46942492	80	100.01055305	100.891677	41.033010684	4.2975016214	2554.8330623
2	airbus	93.952926911	58	96.878686347	98.085883143	29.178095121	3.967524021	2008.2207232
3	airbus	99.148062915	63	97.096913917	96.913737767	33.144245658	3.5162975656	2060.1694249
4	airbus	112.87149908	60	104.45540038	103.6715358	23.783587114	3.9026553246	2488.9984842
5	airbus	148.49500413	53	99.874522521	98.724063607	39.520425649	3.9041206536	2404.7430929
6	airbus	217.12308376	66	94.81425838	97.631341718	33.058365517	3.8235547791	2017.6011486
7	airbus	131.73109556	60	131.03518222	131.3379485	28.277965541	3.6601936464	4896.2946083
8	airbus	160.39281504	55	103.27582495	105.18709549	54.198540346	3.95212311	2837.0808498
9	airbus	122.73330136	68	94.351548777	97.566738197	45.457865547	4.211725648	2376.8009789
10	airbus	123.30242152	41	97.568203986	96.978436701	38.409192953	3.5322719834	2167.7576915
11	airbus	45.502778921	58	107.28766839	105.68511992	21.833564348	3.3998185919	2542.3356261
12	airbus	113.37605039	52	98.891571993	100.52805637	33.573007382	3.5034607776	2169.6725734
13	airbus	247.49599004	66	100.75477196	100.83679396	19.028711072	3.1679780428	2123.1470877
14	airbus	193.12007702	59	96.196564169	98.116074155	42.307731844	3.6139143714	2330.9880709
15	airbus	249.89360363	70	99.069709651	98.871419862	36.195843364	4.491129194	2268.7826989
16	airbus	108.72206895	64	99.817383459	100.30989703	16.215280593	4.314091631	2080.219501
17	airbus	145.1702721	67	109.61058292	107.31454298	27.894764076	4.2154127887	2781.7263302
18	airbus	223.95233137	56	99.325035899	99.397364536	36.782682861	3.1800709689	2059.5377377
19	airbus	123.18315223	63	106.92922516	106.7602536	26.927859592	3.2035138242	2770.4296167

Step 2: Explorative Analysis

Function: our goal in this step is to find correlation among the variables and how they impact landing distance

The following code gives the understanding of the data
SAS Code :

```
PROC MEANS DATA=CombinedFAA N NMISS MEAN STD MIN MAX RANGE;
TITLE SUMMARY STATISTICS FOR Flights landing;
RUN;
```

Output:

SUMMARY STATISTICS FOR Flights landing

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Std Dev	Minimum	Maximum	Range
duration	duration	195	38	150.8830087	48.1531032	45.5027789	287.0025157	241.4997368
no_pasg	no_pasg	233	0	60.0000000	7.0202908	41.0000000	80.0000000	39.0000000
speed_ground	speed_ground	233	0	103.2535315	9.9894832	88.6875803	132.7846766	44.0970963
speed_air	speed_air	233	0	103.2934783	9.8814105	90.0028586	132.9114649	42.9086063
height	height	233	0	30.3742911	9.4795768	9.6972160	58.2277997	48.5305837
pitch	pitch	233	0	4.0455453	0.5502469	2.7019237	5.3106775	2.6087538
distance	distance	233	0	2787.38	834.2412876	1740.90	5381.96	3641.06

We observed there are 38 missing values in duration

We want to check the correlation among the data variables, Using Proc Corr we can find the correlation

SUMMARY STATISTICS FOR Flights landing

The CORR Procedure

7 Variables:	distance duration no_pasg speed_ground speed_air height pitch

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	233	2787	834.24129	649459	1741	5382	distance
duration	195	150.88301	48.15310	29422	45.50278	287.00252	duration
no_pasg	233	60.00000	7.02029	13980	41.00000	80.00000	no_pasg
speed_ground	233	103.25353	9.98948	24058	88.68758	132.78468	speed_ground
speed_air	233	103.29348	9.88141	24067	90.00286	132.91146	speed_air
height	233	30.37429	9.47958	7077	9.69722	58.22780	height
pitch	233	4.04555	0.55025	942.61206	2.70192	5.31068	pitch

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	distance	duration	no_pasg	speed_ground	speed_air	height	pitch
distance distance	1.00000 233	0.05242 0.4668 195	-0.01395 0.8323 233	0.93009 <.0001 233	0.94379 <.0001 233	0.07280 0.2684 233	0.01989 0.7626 233
duration duration	0.05242 0.4668 195	1.00000 195	-0.06918 0.3366 195	0.02389 0.7403 195	0.04454 0.5364 195	0.07378 0.3054 195	-0.05628 0.4346 195
no_pasg no_pasg	-0.01395 0.8323 233	-0.06918 0.3366 195	1.00000 233	0.00866 0.8954 233	0.00056 0.9932 233	0.02359 0.7202 233	-0.01597 0.8084 233
speed_ground speed_ground	0.93009 <.0001 233	0.02389 0.7403 195	0.00866 0.8954 233	1.00000 233	0.98818 <.0001 233	-0.08264 0.2088 233	-0.06141 0.3507 233
speed_air speed_air	0.94379 <.0001 233	0.04454 0.5364 195	0.00056 0.9932 233	0.98818 <.0001 233	1.00000 233	-0.07582 0.2490 233	-0.04525 0.4919 233
height height	0.07280 0.2684 233	0.07378 0.3054 195	0.02359 0.7202 233	-0.08264 0.2088 233	-0.07582 0.2490 233	1.00000 233	-0.03698 0.5744 233
pitch pitch	0.01989 0.7626 233	-0.05628 0.4346 195	-0.01597 0.8084 233	-0.06141 0.3507 233	-0.04525 0.4919 233	-0.03698 0.5744 233	1.00000 233

From the correlation matrix, we came to know that only **speed_ground, speed_air** and **distance** are strongly correlated.

Among speed_air,speed_ground, these two are also strongly correlated.

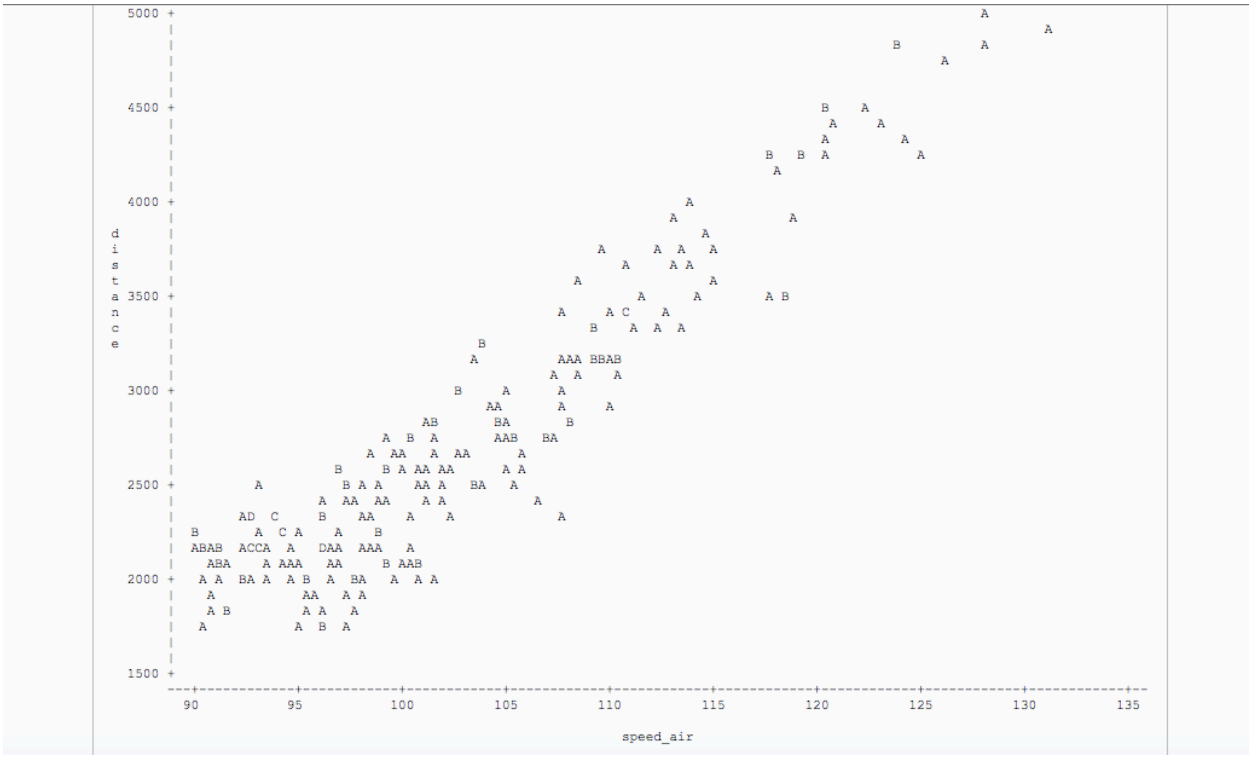
So distance has impact with only **speed_air**

Step 3: Modeling

Define model the distance depends on speed_air, for modeling we do
Step 1. Do the plots

Plots:

The plots shows it is linear modal and slope is positive



Step 2: Calculate the correlation matrix

SUMMARY STATISTICS FOR Flights landing

The CORR Procedure

2 Variables:	distance speed_air

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	233	2787	834.24129	649459	1741	5382	distance
speed_air	233	103.29348	9.88141	24067	90.00286	132.91146	speed_air

Pearson Correlation Coefficients, N = 233
Prob > |r| under H0: Rho=0

	distance	speed_air
distance distance	1.00000	0.94379 <.0001
speed_air speed_air	0.94379 <.0001	1.00000

Correlation of distance and speed_air is 0.94379. Shows that these two are strongly correlated.

Step 3. Do regression analysis
Regression is concerned with finding a model that describes the relationship between a response variable and several predictor (explanatory) variables

SAS Code:

```
proc reg data=CombinedFAA;
model distance=speed_air /r;
output out=diagnostics r=residual;
title Regression analysis of the simulated data set;
run;
```

Regression analysis of the simulated data set

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	233
Number of Observations Used	233

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	143821003	143821003	1883.22	<.0001
Error	231	17641375	76370		
Corrected Total	232	161462378			

Root MSE	276.35048	R-Square	0.8907
Dependent Mean	2787.37914	Adj R-Sq	0.8903
Coeff Var	9.91435		

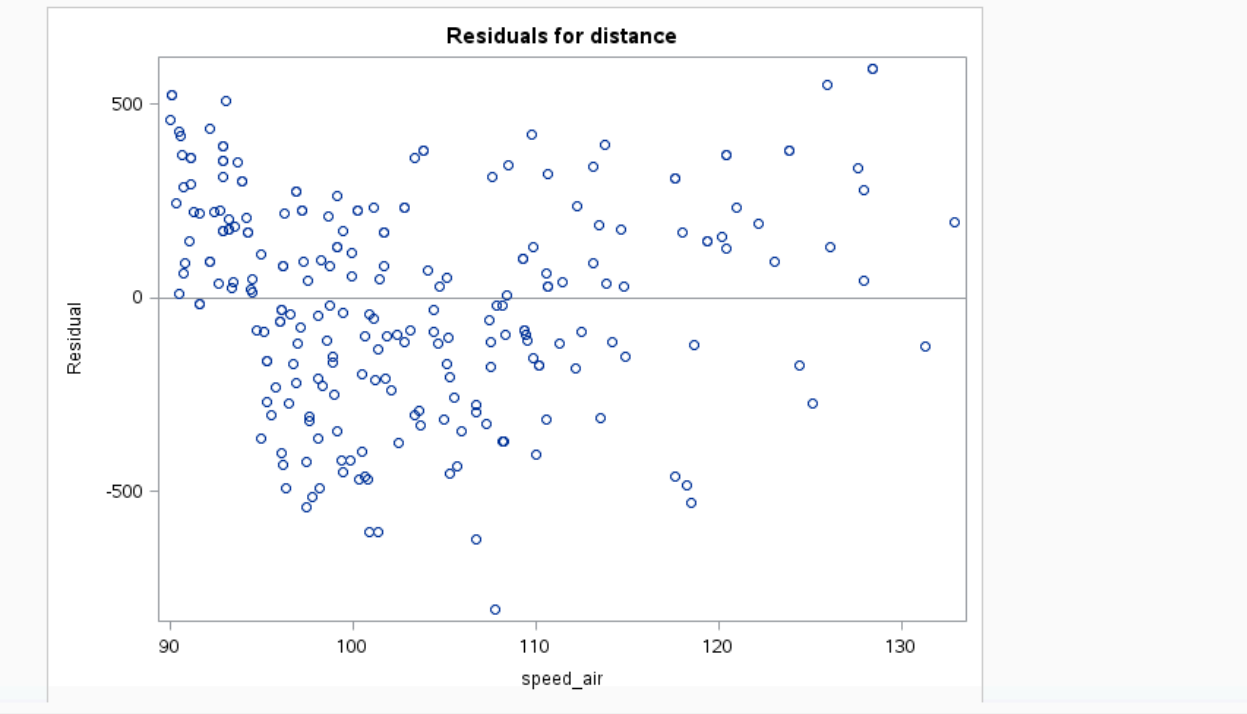
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5443.02432	190.51972	-28.57	<.0001
speed_air	speed_air	1	79.67980	1.83610	43.40	<.0001

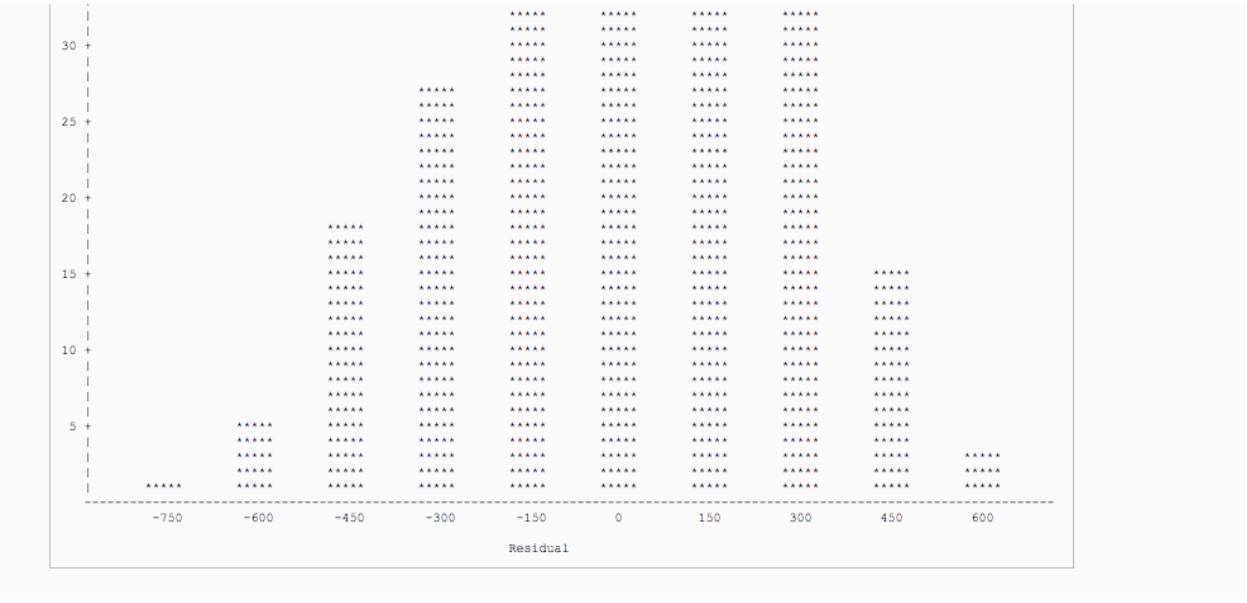
Step 4. Model Checking
In regression analysis, assumptions are made about the error (noise) terms. They are assumed to be 1. Independent. 2. Normally distributed. 3. Mean 0. 4. Constant variance.

Independent:

Plot the residuals vs Air_speed



2.Normally distributed.
Check the distribution is normal or not
Residuals are normal observed from the graph



Test for normality from univariate

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.988108	Pr < W	0.0506
Kolmogorov-Smirnov	D	0.055312	Pr > D	0.0817
Cramer-von Mises	W-Sq	0.112172	Pr > W-Sq	0.0809
Anderson-Darling	A-Sq	0.749136	Pr > A-Sq	0.0502

3. Mean 0

Regression analysis of the simulated data

The MEANS Procedure

Analysis Variable : residual Residual	
t Value	Pr > |t|
-0.00	1.0000

pr value gives that 1, So that, hypothesis is strong pr value. We can not reject the hypothesis.

4.Constant variance.

After checking the residual plots, it has constant variance