FAA dataset assignment : Summary

The **goal** of this project is to assess a given dataset that comprises 8 variables, potentially important for flight landing, and investigate the risk factors associated with landing distance of a flight and how they affect it. The **motivation** is to reduce landing over run of flights in the runway. To derive our final conclusions, we followed the following steps:

1. Data processing (Exploration and cleaning) - Chapter 1
2. Descriptive analysis (Investigating correlations between given variables) - Chapter 2
3. Statistical Modeling (Regression model to assess what factors and how they impact the landing distance of a commercial flight) - Chapter 3

We were able to derive our conclusions in a statistically significant way and answered the following questions.

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

Although we were given a total of 950 flights' observations, we used **831 total observations** to fit our final model. Rest of the observations were dropped out of our model because of either being a duplicate observation in the dataset or possessing an abnormal values for at least one of the variables and the abnormal values were only a small percentage (about 1% or less) of the total sample size. I kept all the rows containing missing values for at least one of the variable, as I did not know the reason behind the values being missing.

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

Landing distance is highly dependent on the make of the aircraft, speed_air and height of the aircraft while passing through the threshold of the runway.

Our final model is as follows:

Landing distance = -6390.38 (se=109.84) + 427.44*aircraft make (se =19.17) + 82.14*speed_air (se=0.98) + 13.7*height (se=1)

- **One-unit change (1 mph)** in the value of **speed_air** leads to about **82 ft of increase** (with a standard error of 0.9 ft) in landing **distance** in average if all other variables remain constant.
- **One-unit change (1ft)** in the value of **height** leads to about **14 ft of increase** (with a standard error of 1 ft) in landing **distance** in average if all other variables remain constant.
- The biggest impact comes from the aircraft make. The landing distance for boeing is 427.44 ft more than airbus (confidence interval : 408, 446), if all other variable remain constant.

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

Yes, there are significant differences between boeing and airbus. The means of pitch and landing distance are different between the makes of these commercial aircrafts. Most importantly, as stated above, the average landing distance for boeing is 427.44 ft more than airbus , if all other variable remain constant.

FAA dataset assignment Chapter 1 (revised): Data exploration and cleaning

- **Import the 2 tables** in SAS env

```
PROC IMPORT OUT= WORK.FAA1 DATAFILE= "/home/u49592956/FAA1.xls"
            DBMS=xls REPLACE;
     SHEET="FAA1";
     GETNAMES=YES;
RUN;
PROC CONTENTS data=FAA1;  /*Inspect the data for variables etc.*/
RUN;

PROC IMPORT OUT= WORK.FAA2 DATAFILE= "/home/u49592956/FAA2.xls"
            DBMS=xls REPLACE;
     SHEET="FAA2";
     GETNAMES=YES;
RUN;
PROC CONTENTS data=FAA2;  /*Inspect the data for variables etc.*/
RUN;
```

FAA1: 800 rows, 8 variables

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	12	$12.	$12.	aircraft
8	distance	Num	8	BEST12.		distance
2	duration	Num	8	BEST12.		duration
6	height	Num	8	BEST12.		height
3	no_pasg	Num	8	BEST12.		no_pasg
7	pitch	Num	8	BEST12.		pitch
5	speed_air	Num	8	BEST12.		speed_air
4	speed_ground	Num	8	BEST12.		speed_ground

FAA2: 200 rows, 7 variables (same as FAA1, duration missing)

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	aircraft	Char	12	$12.	$12.	aircraft
7	distance	Num	8	BEST12.		distance
5	height	Num	8	BEST12.		height
2	no_pasg	Num	8	BEST12.		no_pasg
6	pitch	Num	8	BEST12.		pitch
4	speed_air	Num	8	BEST12.		speed_air
3	speed_ground	Num	8	BEST12.		speed_ground

- **Concatenate** two tables

```
Data FAA_merged;
SET FAA1 FAA2;
Run;
PROC PRINT data=FAA_merged;
Run:
```

We saw multiple empty rows (output not included in the report), that is common when excel tables are imported in SAS.

- **Remove empty rows**

```
options missing = ' '; /*deleting all empty rows*/
data FAA;
   set FAA_merged;
   if missing(cats(of _all_)) then delete;
run;
PROC PRINT data=faa;
Run;
```

FAA has 950 observations and 8 variables

- Check for the **summary statistics** in the new table FAA

```
/*Summary statistics*/
Title1 "Summary Statistics";
Title2 "Unprocessed merged data";
options nolabel;
PROC Means DATA=FAA N nmiss MIN MAX MEAN MEDIAN STDDEV RANGE;
  VAR duration no_pasg speed_ground speed_air height pitch distance;
Run;
PROC Means DATA=FAA noprint;
  output out=missings (drop=_type_ _FREQ_) nmiss= ;
RUN ;
proc transpose data=missings
               out= missings;
run;
Data FAA_missing;
 set missings (rename=(_NAME_=Variable));
 percent_missing=(COL1/950)*100;
 drop COL1;
run;
proc print data=faa_missing;
Title1 "Percent of missing values";
Title2 "Unprocessed merged data";
run;
```

Summary Statistics
Unprocessed merged data

The MEANS Procedure

Variable	N	N Miss	Minimum	Maximum	Mean	Median	Std Dev	Range
duration	800	150	14.7642071	305.6217107	154.0065385	153.9480975	49.2592338	290.8575036
no_pasg	950	0	29.0000000	87.0000000	60.1652632	60.0000000	7.4900041	58.0000000
speed_ground	950	0	27.7357153	141.2186354	79.2849940	79.4129094	19.3364178	113.4829200
speed_air	239	711	90.0028586	141.7249357	103.7304174	100.8916770	10.6051134	51.7220771
height	950	0	-3.5462524	59.9459639	30.1392714	29.9044945	10.3593491	63.4922163
pitch	950	0	2.2844801	5.9267842	4.0192472	4.0153874	0.5260322	3.6423041
distance	950	0	34.0807833	6533.05	1548.82	1267.44	948.6812561	6498.97

Percent of missing values
Unprocessed merged data

Obs	Variable	percent_missing
1	duration	15.7895
2	no_pasg	0.0000
3	speed_ground	0.0000
4	speed_air	74.8421
5	height	0.0000
6	pitch	0.0000
7	distance	0.0000

- **Check for Duplicate records** : removing duplicate records are necessary since they may skew our observations for correlation and variance analysis downstream.

Since duration column is not present in FAA2, I am ignoring duration for this purpose.

```
/*Check duplicate data*/
Proc sort data=FAA dupout=duplicate nodupkey;
by aircraft no_pasg speed_ground speed_air height pitch distance;
run;

/*Remove duplicate data*/
Proc sort data=FAA out=FAA_unique nodupkey;
by aircraft no_pasg speed_ground speed_air height pitch distance; /* Since there is no duration
column in FAA2, I am ignoring that column here*/
run;
```

FAA_unique now has 850 observations (after removing 100 duplicate records) and 8 variables

- Check for **summary statistics in the new table FAA_unique** that has no duplicate records or entirely empty rows.

```
/*Check for summary statistics again*/
Title1 "Summary Statistics";
Title2 "Merged data : Unique records";
options nolabel;
PROC Means DATA=FAA_unique N nmiss MIN MAX MEAN MEDIAN STDDEV RANGE;
  VAR duration no_pasg speed_ground speed_air height pitch distance;
RUN ;
PROC Means DATA=FAA_unique noprint;
  output out=missings_unique (drop=_type_ _FREQ_) nmiss= ;
RUN ;
proc transpose data=missings_unique
               out= missings_unique;
run;
Data FAA_missing_unique;
 set missings_unique (rename=(_NAME_=Variable));
 percent_missing=(COL1/850)*100;
 drop COL1;
run;
```

```
proc print data=faa_missing_unique;
Title1 "Percent of missing values";
Title2 "Merged data : Unique records";
run;
```

Summary Statistics
Merged data : Unique records

The MEANS Procedure

Variable	N	N Miss	Minimum	Maximum	Mean	Median	Std Dev	Range
duration	800	50	14.7642071	305.6217107	154.0065385	153.9480975	49.2592338	290.8575036
no_pasg	850	0	29.0000000	87.0000000	60.1035294	60.0000000	7.4931370	58.0000000
speed_ground	850	0	27.7357153	141.2186354	79.4523229	79.6428041	19.0594903	113.4829200
speed_air	208	642	90.0028586	141.7249357	103.7977237	101.1473493	10.2590370	51.7220771
height	850	0	-3.5462524	59.9459639	30.1442223	30.0931324	10.2877268	63.4922163
pitch	850	0	2.2844801	5.9267842	4.0093577	4.0082875	0.5288298	3.6423041
distance	850	0	34.0807833	6533.05	1526.02	1258.09	928.5600816	6498.97

Percent of missing values
Merged data : Unique records

Obs	Variable	percent_missing
1	duration	5.8824
2	no_pasg	0.0000
3	speed_ground	0.0000
4	speed_air	75.5294
5	height	0.0000
6	pitch	0.0000
7	distance	0.0000

We see many **abnormal or unexpected values** in the summary statistics such as negative height , very low duration time, very low or high speeds and longer landing distance than usual. Also, about 75% of the speed_Air values and 6% or the duration values are missing. So we want to carefully inspect the data.

```
PROC univariate data=faa;
Run;
```

Few interesting observations for these variables:

<u>Duration:</u>

Quantiles (Definition 5)	
Level	Quantile
100% Max	305.6217
99%	275.6969
95%	234.1229
90%	214.4738
75% Q3	188.9179
50% Median	153.9481
25% Q1	119.4746
10%	92.0313
5%	74.4080

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
14.7642	629	289.320	110
16.8935	97	293.230	674
17.3755	706	298.522	770

1%	45.5691
0% Min	14.7642

31.3910	505	302.967	315
31.7017	242	305.622	106

For duration, less than 40 min for a flight, is considered unusual. Here we see bottom 5 are all bellow 40 but total number of abnormal observations are bellow 1%. Also, for this variable ~6% data were missing initially. So I decided to remove these abnormal values.

Speed_ground:

Quantiles (Definition 5)	
Level	Quantile
100% Max	141.2186
99%	126.2443
95%	111.6739
90%	104.1367
75% Q3	92.0983
50% Median	79.6428
25% Q1	65.8853
10%	54.9904
5%	47.8821
1%	38.2590
0% Min	27.7357

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
27.7357	459	129.307	788
29.2277	658	131.035	229
33.5741	306	132.785	521
33.8230	789	136.659	743
34.1178	744	141.219	547

If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal. There are only 3 values of this kind and are out of 99%, so we will remove these observations.

Speed_air:

Quantiles (Definition 5)	
Level	Quantile
100% Max	141.7249
99%	132.9115
95%	125.1385
90%	118.6726
75% Q3	109.4202
50% Median	101.1473
25% Q1	96.2294
10%	92.7526
5%	91.0725
1%	90.3674
0% Min	90.0029

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
90.0029	591	128.418	832
90.1110	676	131.338	229
90.3674	560	132.911	521
90.4767	697	136.423	743
90.5033	592	141.725	547

If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal. Since there is only 1 value of this kind and are out of 99%, so we will remove this observation.

Height:

Quantiles (Definition 5)	
Level	Quantile
100% Max	59.94596
99%	53.43862
95%	47.38932
90%	43.91020
75% Q3	36.99458
50% Median	30.09313
25% Q1	23.30227

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-3.5462524	726	55.0935	586
3.3333880	434	58.0818	577

10%	17.21471
5%	13.80759
1%	3.78892
0% Min	-3.54625

-2.9153359	164	58.0835	616
-1.5281292	806	58.2278	376
-0.0677586	377	59.9460	571

The landing aircraft is required to be at least 6 meters high at the threshold of the runway and also, it can never be negative. The negative values must be removed.

Distance:

Quantiles (Definition 5)	
Level	Quantile
100% Max	6533.0477
99%	4807.8798
95%	3439.7255
90%	2741.1321
75% Q3	1937.2563
50% Median	1258.0915
25% Q1	883.5989
10%	645.4543
5%	492.7502
1%	280.8044
0% Min	34.0808

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
34.0808	164	5147.41	610
41.7223	445	5343.20	521
133.0869	281	5381.96	832
180.5652	308	6309.95	743
241.1610	16	6533.05	547

The length of the airport runway is typically less than 6000 feet and so a landing distance greater than that indicates a flight overrun. There are 2 data points that are abnormal with this criteria.

- check **percentage of abnormal values** for each variable.

```
/*Check for abnormal values and delete values*/

Data FAA_unique_clean;
Set FAA_unique;
IF duration < 40 and duration ^=.  Then duration_abnormal = 1;
IF duration >= 40 Then duration_abnormal = 0;
IF height < 6 and height ^=.  THEN height_abnormal = 1;
ELSE height_abnormal = 0;
IF 0 < speed_ground < 30 or speed_ground > 140 Then speed_ground_abnormal = 1;
ELSE speed_ground_abnormal = 0;
IF 0 < speed_air < 30 or speed_air > 140 then speed_air_abnormal = 1;
ELSE speed_air_abnormal = 0;
IF distance > 6000 then OverRun = 1;
ELSE OverRun = 0;
Run;
PROC print data=FAA_unique_clean;
Run;
proc summary data= faa_unique_clean;
var duration_abnormal speed_ground_abnormal speed_air_abnormal height_abnormal OverRun;
output out=total (drop=_type_ _FREQ_) sum=;
run;
proc transpose data = total
               out = total;
run;
```

```
proc print data = total;
run;
Data total_abPCN;
 set total;
 rename _NAME_=Variable;
 percent_abnormal = (COL1/850)*100;
 drop COL1;
run;
proc print data = total_abPCN;
Title "Percentage of abnormal values";
run;
```

Percentage of abnormal values

Obs	Variable	percent_abnormal
1	duration_abnormal	0.58824
2	speed_ground_abnormal	0.35294
3	speed_air_abnormal	0.11765
4	height_abnormal	1.17647
5	OverRun	0.23529

We see **about or less than 1% of abnormal values** for each variable and so we delete these observations.

```
/*Delete abnormal observations*/

Data FAA_unique_clean;
Set FAA_unique;
IF duration < 40 and duration ^=.  then delete;
IF height < 6 and height ^=.  then delete;
IF speed_ground < 30 and speed_ground ^=.  then delete;
IF speed_ground > 140 then delete;
IF speed_air < 30 and speed_air ^=.  then delete;
IF speed_air > 140 then delete;
IF distance > 6000 then delete;
Run;
PROC print data = FAA_unique_clean;
Run;
proc summary data = faa_unique_clean;
var duration no_pasg speed_ground speed_air height pitch distance;
run;
```

- Run summary statistics on the clean data

```
Title1 "Summary Statistics";
Title2 "Clean data_Final";
options nolabel;
PROC Means DATA=FAA_unique_clean N nmiss MIN MAX MEAN MEDIAN STDDEV RANGE;
  VAR duration no_pasg speed_ground speed_air height pitch distance;
RUN ;
```

```
RUN ;
PROC Means DATA=FAA_unique_clean noprint;
  output out=missings_unique_clean (drop=_type_ _FREQ_) nmiss= ;
RUN ;
proc transpose data=missings_unique_clean
               out= missings_unique_clean;
run;
Data FAA_missing_unique_clean;
 set missings_unique_clean (rename=(_NAME_=Variable));
 percent_missing=(COL1/831)*100;
 drop COL1;
run;
proc print data=faa_missing_unique;
Title1 "Percent of missing values : data_Final";
run;
```

FAA_unique_clean now has 831 observations (after removing rows containing abnormal values for at least one of the variables) and 8 variables

Summary Statistics
Clean data_Final

The MEANS Procedure

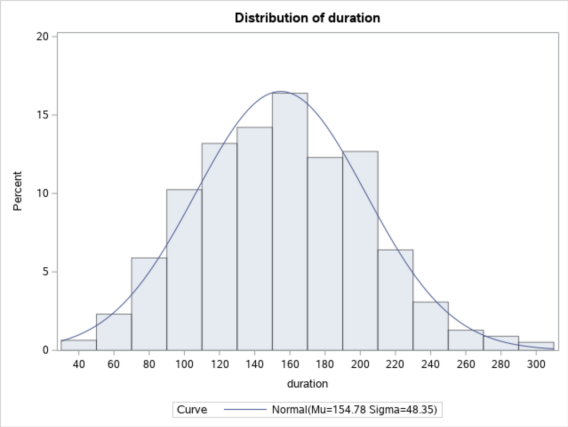
Variable	N	N Miss	Minimum	Maximum	Mean	Median	Std Dev	Range
duration	781	50	41.9493694	305.6217107	154.7757191	154.2845505	48.3499237	263.6723414
no_pasg	831	0	29.0000000	87.0000000	60.0553550	60.0000000	7.4913166	58.0000000
speed_ground	831	0	33.5741041	132.7846766	79.5426997	79.7939604	18.7356754	99.2105726
speed_air	203	628	90.0028586	132.9114649	103.4850352	101.1189240	9.7362774	42.9086063
height	831	0	6.2275178	59.9459639	30.4578695	30.1670844	9.7848114	53.7184462
pitch	831	0	2.2844801	5.9267842	4.0051609	4.0010380	0.5265690	3.6423041
distance	831	0	41.7223127	5381.96	1522.48	1262.15	896.3381524	5340.24

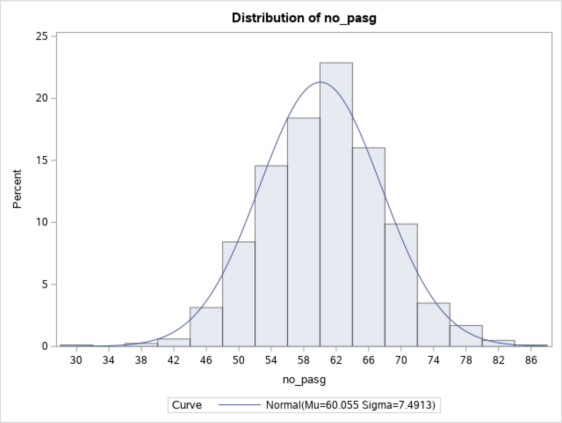
Percent of missing values : data_Final

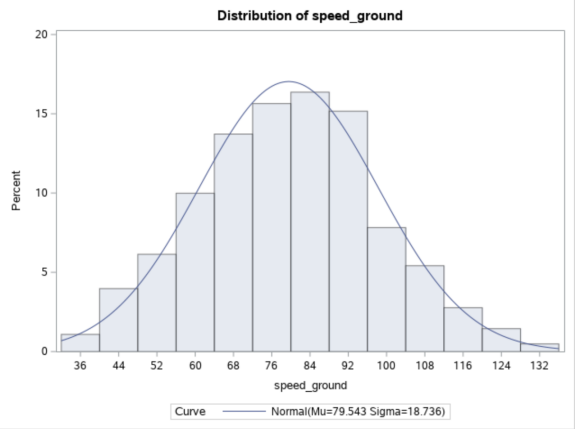
Obs	Variable	percent_missing
1	duration	5.8824
2	no_pasg	0.0000
3	speed_ground	0.0000
4	speed_air	75.5294
5	height	0.0000
6	pitch	0.0000
7	distance	0.0000

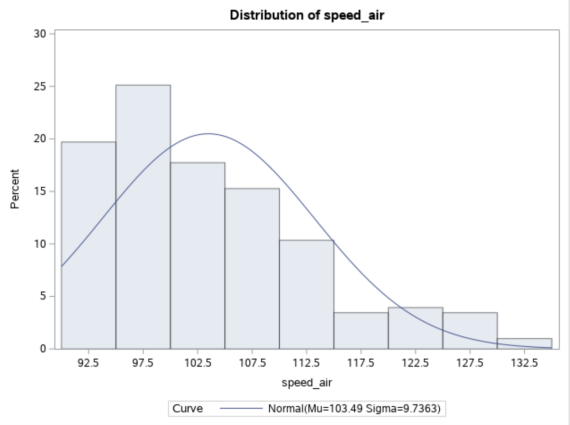
- Check the normalcy of distribution each variable
-

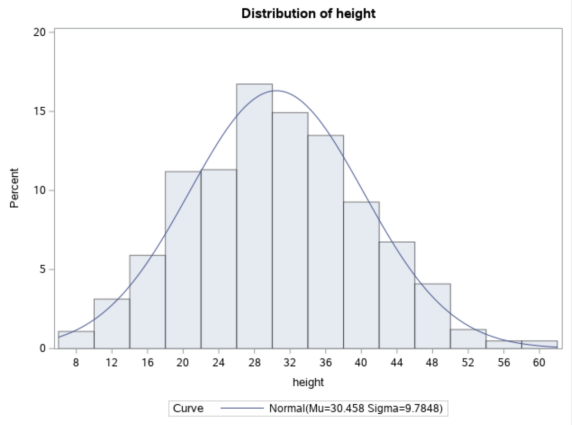
```
/*Plot each variable to check distribution*/
PROC UNIVARIATE DATA = FAA_unique_clean;
VAR _numeric_;
options nolabel;
HISTOGRAM _numeric_ / NORMAL; /* HISTOGRAM variables / <OPTIONS>*/
PROBPLOT _numeric_ / NORMAL; /* PROBPLOT variables / <OPTIONS>*/
Run;
```

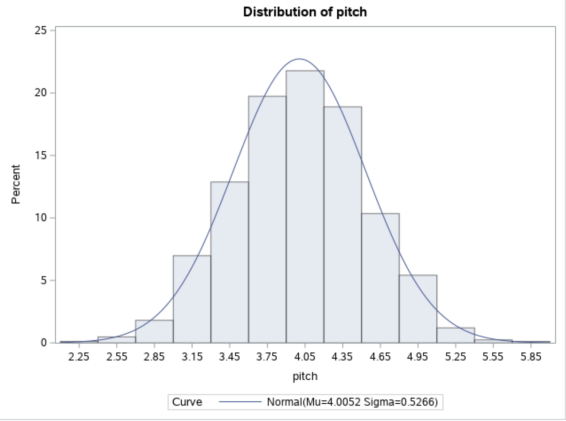

Distribution of duration
Percent
duration
Curve Normal(Mu=154.78 Sigma=48.35)

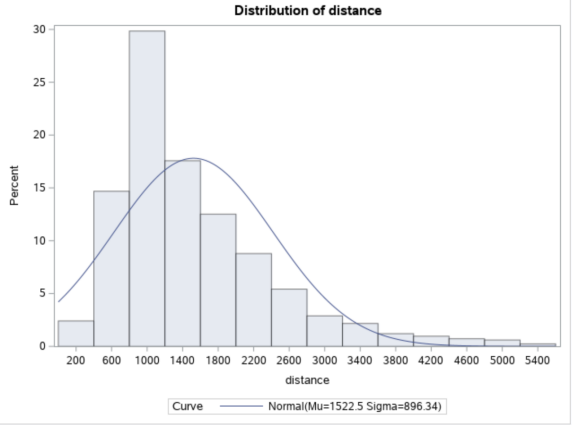

Distribution of no_pasg
Percent
no_pasg
Curve Normal(Mu=60.055 Sigma=7.4913)


Distribution of speed_ground
Percent
speed_ground
Curve Normal(Mu=79.543 Sigma=18.736)

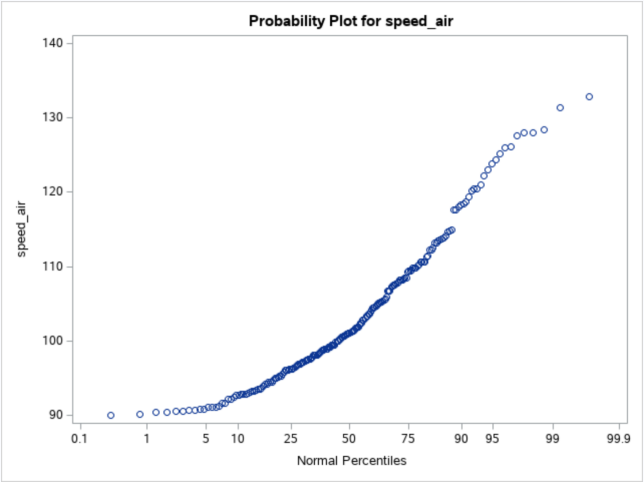

Distribution of speed_air
Percent
speed_air
Curve Normal(Mu=103.49 Sigma=9.7363)

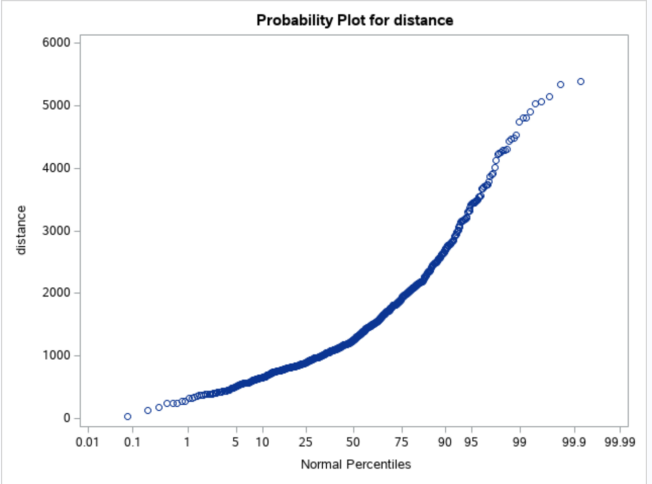

Distribution of height
Percent
height
Curve Normal(Mu=30.458 Sigma=9.7848)


Distribution of pitch
Percent
pitch
Curve Normal(Mu=4.0052 Sigma=0.5266)


Distribution of distance
Percent
distance
Curve Normal(Mu=1522.5 Sigma=896.34)

We checked the distribution of each variable. While all other variable looks **normally distributed**, **speed_Air and distance looks right-skewed**. We also checked the probability plot (below) which indicated that the data may not be **normally distributed.**





- TTest to check if the means of speed-air and distance are 0 (mu=0) to confirm if they are normally distributed.

```
PROC MEANS data = faa_unique_clean T PRT ;
VAR speed_air distance;
Run;
```

The MEANS Procedure

Variable	t Value	Pr > |t|
speed_air	145.92	<.0001
distance	47.91	<.0001

The p-value from the t-test is significantly low and we reject the the null hypothesis that these two variables are normally distributed.

- Check the difference in summary statistics between aircraft models

```
/*Check for differences between aircraft models*/

Title1 "Summary Statistics";
Title2 "Clean data_Final";
options nolabel;
PROC Means DATA=FAA_unique_clean N nmiss MIN MAX MEAN MEDIAN STDDEV RANGE;
  VAR duration no_pasg speed_ground speed_air height pitch distance;
  By aircraft;
RUN ;
```

Summary Statistics
Clean data_Final

The MEANS Procedure

aircraft=airbus

Variable	N	N Miss	Minimum	Maximum	Mean	Median	Std Dev	Range
duration	398	52	42.1462262	305.6217107	156.7723322	156.4468009	49.0162163	263.4754846
no_pasg	450	0	36.0000000	87.0000000	60.2466667	60.0000000	7.4174927	51.0000000
speed_ground	450	0	33.5741041	131.0351822	80.1994492	81.1131950	16.9206507	97.4610782
speed_air	86	364	95.0113646	131.3379485	104.2123333	101.2648043	8.0924561	36.3265839
height	447	3	0.0861055	58.2277997	30.5372899	30.3548877	9.9320715	58.1416943
pitch	450	0	2.2844801	5.5267842	3.8317436	3.8257225	0.5004493	3.2423041
distance	450	0	34.0807833	4896.29	1318.19	1124.13	792.3479576	4862.21

aircraft=boeing

Variable	N	N Miss	Minimum	Maximum	Mean	Median	Std Dev	Range
duration	397	3	41.9493694	298.5223339	152.8909742	152.5763119	47.5281704	256.5729646
no_pasg	400	0	29.0000000	82.0000000	59.9425000	60.0000000	7.5834005	53.0000000
speed_ground	400	0	27.7357153	141.2186354	78.6118058	78.4539884	21.1999092	113.4829200
speed_air	122	278	90.0028586	141.7249357	103.5054579	101.1070213	11.5689208	51.7220771
height	398	2	1.2538553	59.9459639	30.1100760	29.6167123	10.1034520	58.6921087
pitch	400	0	2.9931514	5.9267842	4.2091735	4.1963624	0.4874715	2.9336328
distance	400	0	371.2772609	6533.05	1759.84	1453.42	1012.25	6161.77

The major difference between these two aircrafts seems to be in the minimum landing distance. There are more missing data in airbus.

Conclusion:

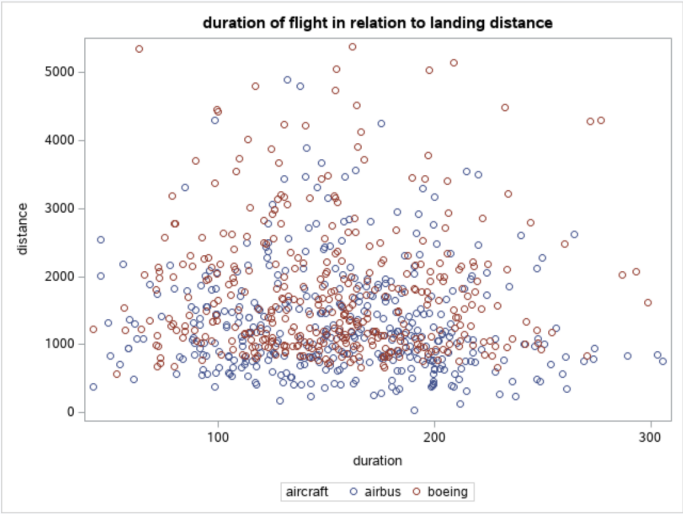
- The combined dataset had 950 observations for two aircraft models in total and 8 variables.
- After data cleaning (removing duplicate and abnormal observations) our final dataset contains 831 observations for 8 variables.
- We need to do run correlation analysis as next step to predict risk factor for landing overrun and also perform an unpaired t-test between aircraft models to check if there is a risk factor associated with the make of the aircraft.

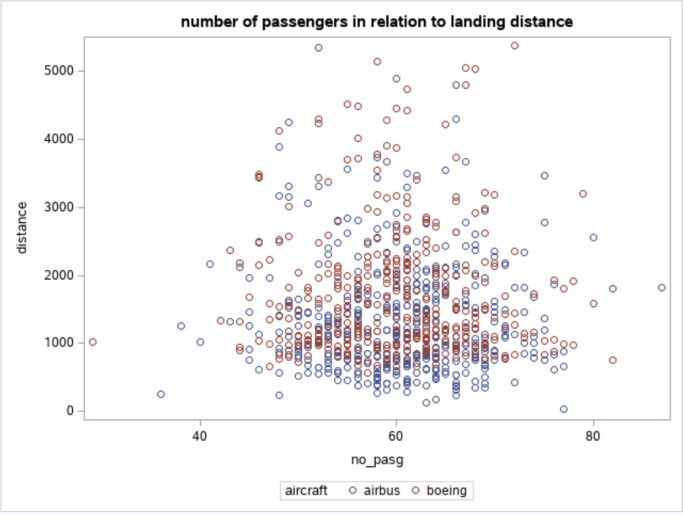
FAA dataset assignment Chapter 2 : Descriptive Analysis

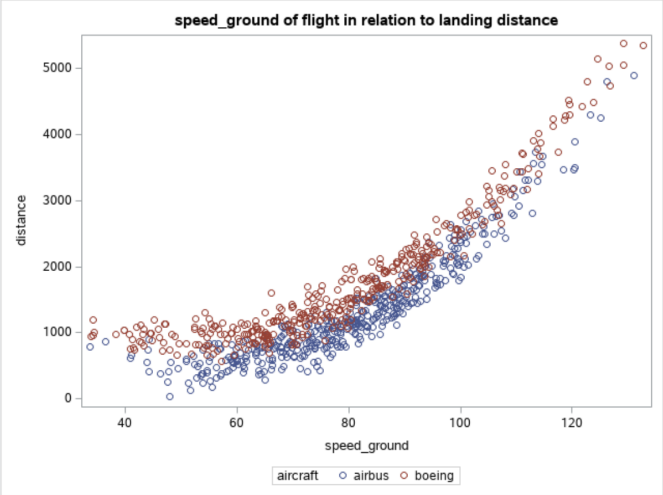
Our **goal** for this part is to assess the relation of each variable to landing distance and investigate which ones are potentially impacting the landing distance that may have influence in runway over-run.

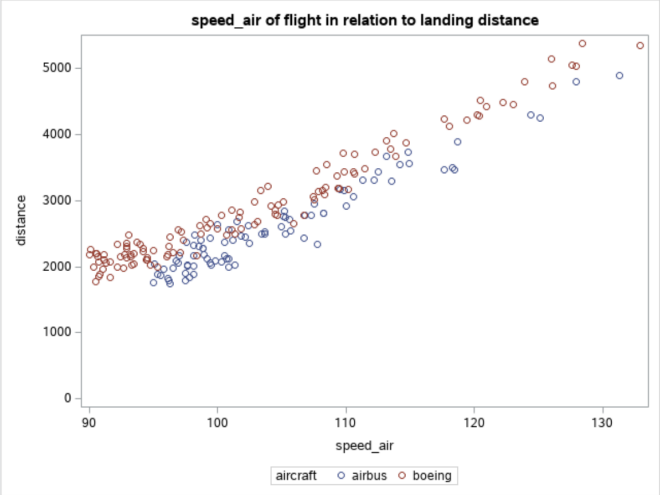
- **X-Y Plot and assess relation** of each independent variable to distance

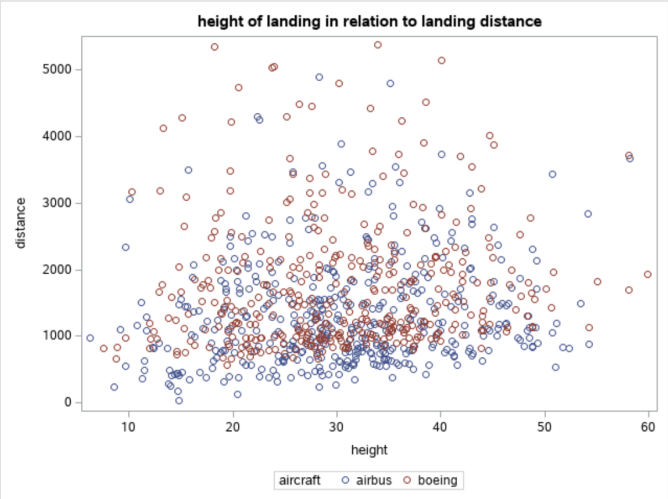
```
/*XY plots for each variable to distance color coded by aircraft*/
proc sgplot data = FAA_unique_clean;
 scatter x=duration y=distance / group=aircraft;
run;
proc sgplot data = FAA_unique_clean;
 scatter x=no_pasg y=distance / group=aircraft;
run;
proc sgplot data = FAA_unique_clean;
 scatter x=speed_ground y=distance / group=aircraft;
run;
proc sgplot data = FAA_unique_clean;
 scatter x=speed_air y=distance / group=aircraft;
run;
proc sgplot data = FAA_unique_clean;
 scatter x=height y=distance / group=aircraft;
run;
proc sgplot data = FAA_unique_clean;
 scatter x=pitch y=distance / group=aircraft;
run;
```

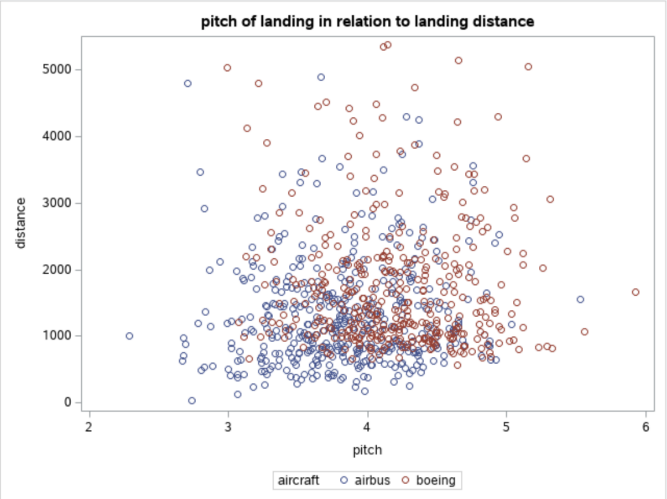












The speed_ground and speed_air seems to be correlated with landing distance for both make of aircraft. While the distribution of values for the two makes of the aircraft look a bit different visually, the relational patterns are very similar for all variables.

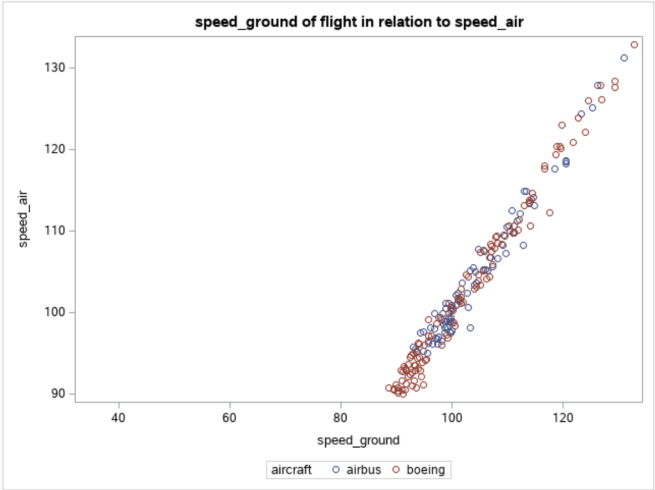
- **Correlation analysis** (Pearson) to find out correlation/ dependency between variables

```
/*Correlation analysis on FAA_unique_clean data*/
proc corr data = faa_unique_clean;
var duration no_pasg speed_ground speed_air height pitch distance;
title Pairwise correlation coefficients for all variables;
run;
```

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
duration	1.00000 800	-0.03391 0.3382 800	-0.06063 0.0866 800	0.04911 0.4898 200	-0.00678 0.8481 800	-0.03896 0.2710 800	-0.06209 0.0792 800
no_pasg	-0.03391 0.3382 800	1.00000 850	-0.00517 0.8803 850	-0.00589 0.9327 208	0.01098 0.7492 850	-0.01490 0.6643 850	-0.03033 0.3771 850
speed_ground	-0.06063 0.0866 800	-0.00517 0.8803 850	1.00000 850	0.98929 <.0001 208	-0.01607 0.6399 850	-0.03062 0.3727 850	0.86196 <.0001 850
speed_air	0.04911 0.4898 200	-0.00589 0.9327 208	0.98929 <.0001 208	1.00000 208	-0.06588 0.3444 208	0.00639 0.9270 208	0.94728 <.0001 208
height	-0.00678 0.8481 800	0.01098 0.7492 850	-0.01607 0.6399 850	-0.06588 0.3444 208	1.00000 850	0.01284 0.7085 850	0.13624 <.0001 850
pitch	-0.03896 0.2710 800	-0.01490 0.6643 850	-0.03062 0.3727 850	0.00639 0.9270 208	0.01284 0.7085 850	1.00000 850	0.10269 0.0027 850
distance	-0.06209 0.0792 800	-0.03033 0.3771 850	0.86196 <.0001 850	0.94728 <.0001 208	0.13624 <.0001 850	0.10269 0.0027 850	1.00000 850

We saw the two variable that shows **high correlation with landing distance** are **speed_ground and speed_air**. They both are positively correlated to distance (correlation coefficients are 0.86 and 0.95 respectively) and very low p-value (<0.0001 for both), indicating that the correlation if statistically significant. This confidently agrees with the visual results from the X-Y plots as well. Interestingly, **speed ground and speed_air** are also highly **correlated** (correlation coefficient 0.99) **with each other** in a statistically significant parameter (p-value <0.0001). We confirmed this by a X-Y plot as well (see below). Interestingly, while duration and number of passengers show no correlation with landing distance, **height and pitch** shows mild correlation in a statistically significant manner with landing distance although a weak linear relationship as seen in the X-Y plot. So, we believe, these 4 variables are potential factors affecting landing distance.

```
proc sgplot data = FAA_unique_clean;
 title speed_ground of flight in relation to speed_air;
 scatter x=speed_ground y=speed_air / group=aircraft;
run;
```



We also wanted to check, if the pattern of correlation is similar in different makes of aircraft separately.

```
/*Correlation analysis on FAA_unique_clean data by aircraft*/
proc corr data = faa_unique_clean;
by aircraft;
var duration no_pasg speed_ground speed_ground height pitch distance;
title Pairwise correlation coefficients for all variables;
run;
```

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

	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
duration	1.00000 400	-0.02334 0.6416 400	-0.06533 0.1922 400	0.05251 0.6480 78	-0.00890 0.8591 400	-0.03954 0.4303 400	-0.08140 0.1040 400
no_pasg	-0.02334 0.6416 400	1.00000 450	0.00724 0.8782 450	-0.05277 0.6294 86	0.00223 0.9624 450	-0.10492 0.0260 450	-0.01034 0.8269 450
speed_ground	-0.06533 0.1922 400	0.00724 0.8782 450	1.00000 450	0.98184 <.0001 86	-0.01767 0.7085 450	-0.00150 0.9746 450	0.90493 <.0001 450
speed_air	0.05251 0.6480 78	-0.05277 0.6294 86	0.98184 <.0001 86	1.00000 86	-0.01363 0.9008 86	-0.00777 0.9434 86	0.96347 <.0001 86
height	-0.00890 0.8591 400	0.00223 0.9624 450	-0.01767 0.7085 450	-0.01363 0.9008 86	1.00000 450	0.03598 0.4464 450	0.16285 0.0005 450
pitch	-0.03954 0.4303 400	-0.10492 0.0260 450	-0.00150 0.9746 450	-0.00777 0.9434 86	0.03598 0.4464 450	1.00000 450	0.07827 0.0973 450
distance	-0.08140 0.1040 400	-0.01034 0.8269 450	0.90493 <.0001 450	0.96347 <.0001 86	0.16285 0.0005 450	0.07827 0.0973 450	1.00000 450

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

	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
duration	1.00000 400	-0.04682 0.3503 400	-0.06160 0.2189 400	0.04552 0.6186 122	-0.00612 0.9028 400	-0.01037 0.8363 400	-0.03211 0.5219 400
no_pasg	-0.04682 0.3503 400	1.00000 400	-0.01773 0.7237 400	0.01608 0.8604 122	0.01978 0.6933 400	0.10104 0.0434 400	-0.04030 0.4215 400
speed_ground	-0.06160 0.2189 400	-0.01773 0.7237 400	1.00000 400	0.99195 <.0001 122	-0.01629 0.7454 400	-0.03138 0.5315 400	0.89373 <.0001 400
speed_air	0.04552 0.6186 122	0.01608 0.8604 122	0.99195 <.0001 122	1.00000 122	-0.09547 0.2956 122	0.03320 0.7166 122	0.97925 <.0001 122
height	-0.00612 0.9028 400	0.01978 0.6933 400	-0.01629 0.7454 400	-0.09547 0.2956 122	1.00000 400	0.00310 0.9508 400	0.13075 0.0088 400
pitch	-0.01037 0.8363 400	0.10104 0.0434 400	-0.03138 0.5315 400	0.03320 0.7166 122	0.00310 0.9508 400	1.00000 400	-0.03263 0.5153 400
distance	-0.03211 0.5219 400	-0.04030 0.4215 400	0.89373 <.0001 400	0.97925 <.0001 122	0.13075 0.0088 400	-0.03263 0.5153 400	1.00000 400

We observed similar pattern in correlation for speed_ground, speed_air and height to distance when grouped by aircraft as well. Pitch seems to be not correlated to distance when grouped by aircraft make.

Since we see some variability in the relational observations grouped by aircraft, we wanted to check if there is any significant difference between the predictor variable or the landing distance observations between these two groups.

- Testing if there is significant difference in **landing distance by make of aircraft** (**unpaired T-test**)

```
/*ttest by aircraft*/
PROC TTEST DATA=faa_unique_clean;
CLASS aircraft;
VAR distance;
TITLE TTEST FOR COMPARING THE MEANS OF landing distance by aircraft make;
RUN;
```

TTEST FOR COMPARING THE MEANS OF landing distance by aircraft make

The TTEST Procedure

Variable: distance

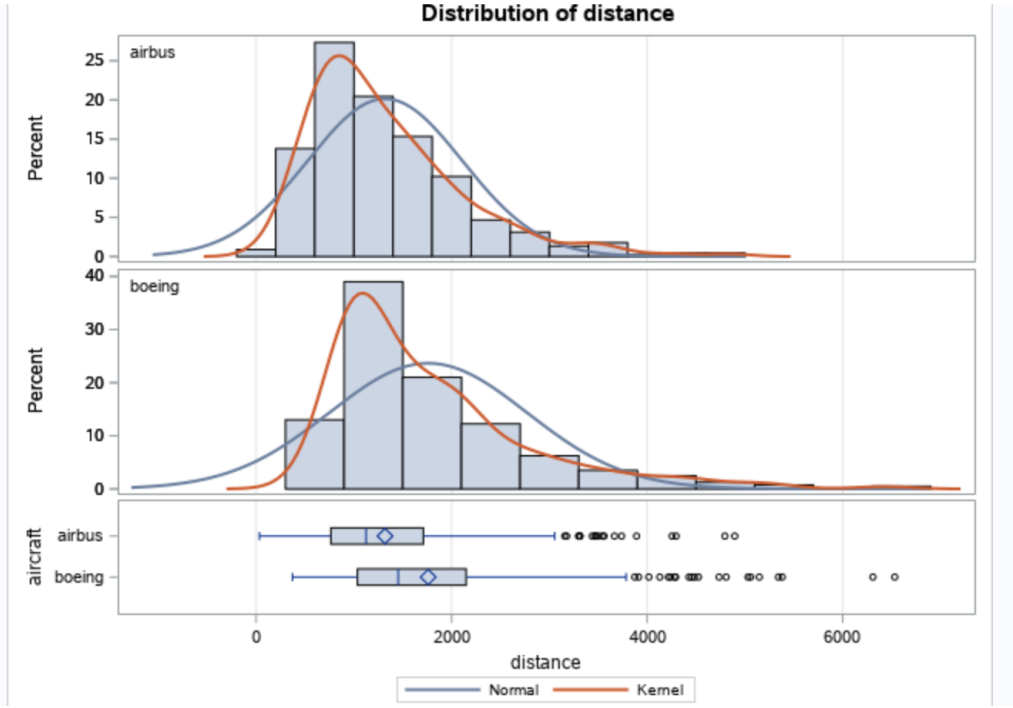
aircraft	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
airbus		450	1318.2	792.3	37.3516	34.0808	4896.3
boeing		400	1759.8	1012.2	50.6123	371.3	6533.0
Diff (1-2)	Pooled		441.7	902.5	62.0193		

Diff (1-2)	Pooled		-441.7	902.5	62.0193		
Diff (1-2)	Satterthwaite		-441.7		62.9027		

aircraft	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
airbus		1318.2	1244.8	1391.6	792.3	743.7	847.8
boeing		1759.8	1660.3	1859.3	1012.2	946.6	1087.7
Diff (1-2)	Pooled	-441.7	-563.4	-319.9	902.5	861.5	947.6
Diff (1-2)	Satterthwaite	-441.7	-565.1	-318.2			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	848	-7.12	<.0001
Satterthwaite	Unequal	753.38	-7.02	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	399	449	1.63	<.0001



The t-test indicated that the **means** of the average **landing distance of two aircrafts are not equal** (rejecting the null hypothesis)in a statistically significant manner and the aircraft make could also potentially be influencing the landing distance.

```
PROC TTEST DATA=faa_unique_clean;
CLASS aircraft;
VAR pitch height speed_ground speed_air; /*checking if other variables/risk fators are different*/
TITLE TTEST FOR COMPARING THE MEANS OF landing distance by aircraft make;
RUN;
```

Pitch also showed significant difference for the means between the two aircraft makes while all other variables looked the same.

Inference:

- Landing distance is highly correlated with speed_ground and speed_air (major risk factors). This assessment is confirmed by x-y plot as well as correlation analysis.
- Speed_ground and speed_air are highly correlated with each other and this observation has also been confirmed by both x-y plot and correlation analysis.
- Since there are more data points or observations in speed_ground compared to speed_air (~75% are missing); in the overlay plot, we see most of the points overlaid with speed_ground.
- Although this correlation pattern is similar between aircraft makes, there is a significant difference between the means of these two aircraft makes, boeing and airbus.
- We did not transform the skewed variables since ttest and anova does not require absolute normal dataset as long as they are normal according to central limit theorem (ref:https://data.library.virginia.edu/normality-assumption/#:~:text=No%2C%20you%20don't%20have,or%20an%20outcome%20(DV)).

FAA dataset assignment Chapter 3 : Statistical modeling

From our previous analysis, we found out that both **speed_ground** and **speed_air** are highly **correlated with landing distance.** Next, to understand the relation between distance and its risk factors, we would perform a regression analysis. Because speed_ground and speed_air are also highly correlated with each other, taking both of them in our modeling would give rise to multicollinearity, which we want to avoid since this can cause problems in estimating the regression coefficients. To choose only one of them, we decided to take **speed_air** into account because of **higher correlation coefficient** to distance compared to that of speed_ground to distance. Also, ***likely that speed_air has an effect on speed_ground during landing process*** and not vice versa. Although, speed_air has more missing values, we see from the plots that most of the values for higher speed_air (right half of the plot) is retained, which is going to affect the prediction of landing over run (our over all goal of the project).

Other variables that we want to include in our modeling are **height** and **pitch**, since they showed low to moderate correlation with landing distance. We would also want to take into account the aircraft. Instead of modeling in a stratified approach by aircraft groups, we create a new table with assigned values for aircraft and take **aircraft** as an independent variable in our model as well.

- Creating new table

```
/*create new table with binary values for aircraft*/
data FAA_final;
set faa_unique_clean;
IF aircraft = 'airbus' then aircraft_mk = 0;
ELSE aircraft_mk = 1;  /* assigning 1 to boeing */
Run;
PROC print data=FAA_final;
Run;
```

Since landing **distance** (response variable or the dependent variable) is **continuous and we aim for a linear relationship prediction,** we use **linear regression** to model its relationship with the speed_air, height, pitch and aircraft_mk (independent variables or the risk-factors).

- **Regression** modeling

```
/*regression model*/
proc reg data = FAA_final;
model distance = aircraft_mk speed_air height pitch / r ;
title Regression analysis of distance to risk factors;
output out=diagnostics r=residual; /*model checking*/
run;
```

Regression analysis of distance to risk factors

The REG Procedure
Model: MODEL1
Dependent Variable: 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	4	132881890	33220473	1834.22	<.0001
Error	198	3586078	18112		
Corrected Total	202	136467968			

Root MSE	134.57899	R-Square	0.9737
Dependent Mean	2774.67289	Adj R-Sq	0.9732
Coeff Var	4.85027		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6383.41479	131.04432	-48.71	<.0001
aircraft_mk	1	428.11078	20.39889	20.99	<.0001
speed_air	1	82.14684	0.97860	83.94	<.0001
height	1	13.69957	1.00991	13.57	<.0001
pitch	1	-1.75734	17.93681	-0.10	0.9221

We see that while all the other variables has a significant p-value, pitch has a p-value that is not significant along with a negative parameter estimate and that indicates it is not contributing to the model. This is what we expected from our correlation analysis grouped by aircraft. So, we take out pitch and re-run the model.

- Re-iterate regression model after dropping out pitch

```
proc reg data = faa_final;
model distance = aircraft_mk speed_air height / r ;
title Regression analysis of distance to risk factors;
run;
proc ttest data = diagnostics; /*model checking*/
run;
```

Regression analysis of distance to risk factors

The REG Procedure
Model: MODEL1
Dependent Variable: 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	3	132881716	44293905	2457.85	<.0001
Error	199	3586252	18021		
Corrected Total	202	136467968			

Root MSE	134.24368	R-Square	0.9737
Dependent Mean	2774.67289	Adj R-Sq	0.9733
Coeff Var	4.83818		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-6390.37552	109.83914	-58.18	<.0001	0
aircraft_mk	1	427.44156	19.17339	22.29	<.0001	1.00789
speed_air	1	82.14852	0.97601	84.17	<.0001	1.01218
height	1	13.70161	1.00718	13.60	<.0001	1.00901

In our final regression model, the **p-value of the F-test is <0.0001** which indicates the overall model is statistically significant. The R-squared is 0.97 indicates that approximately **97% of the variability** of landing **distance** is accounted for by **aircraft_mk**, **speed_air** and **height** in the regression model and all of them has a positive correlation with distance. The variance inflation factor is ~1 for all variables indicate that they are all independent and no multicollinearity were observed.

The coefficient or the parameter estimate (beta) for speed_air indicates that **one-unit change** in the value of **speed_air** leads to about **82 ft of increase** (with a very small standard error, 0.9 ft) in landing **distance** in average if all other variables remain constant.

The coefficient or the parameter estimate (beta) for height indicates that **one-unit change** in the value of **height** leads to about **14 ft of increase** (with a very small standard error, 1 ft) in landing **distance** in average if all other variables remain constant.

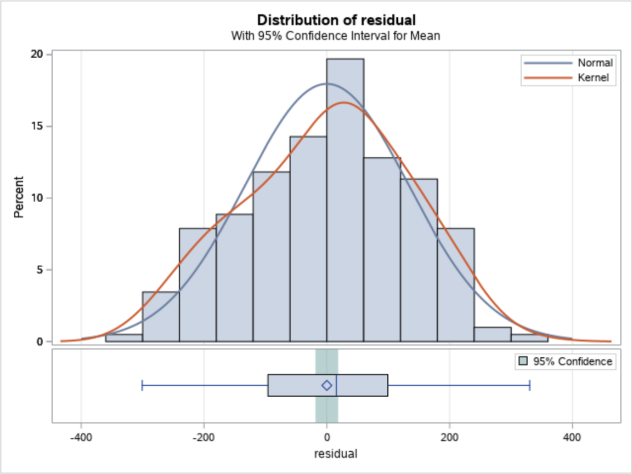
The biggest impact comes from the aircraft make. The average landing distance for boeing is 427.44 ft more than airbus (confidence interval : 408, 446), if all other variable remain constant.

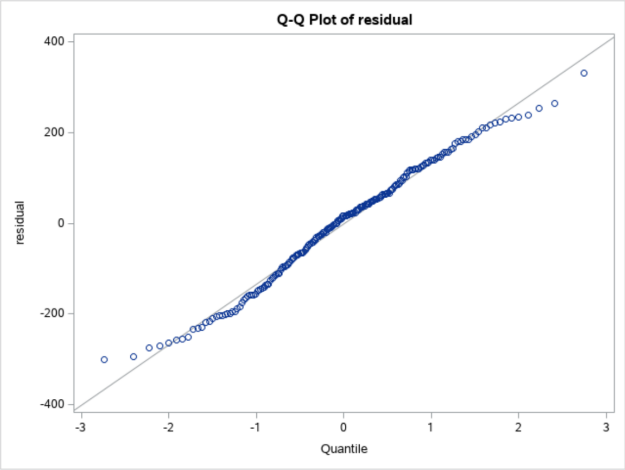
Variable: residual

N	Mean	Std Dev	Std Err	Minimum	Maximum
203	2.27E-12	133.2	9.3518	-300.7	330.4

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
2.27E-12	-18.4397	18.4397	133.2	121.4	147.6

DF	t Value	Pr > |t|
202	0.00	1.0000





The residuals are normally distributes as well (p-value 1 for t-test, can not reject null hypothesis, mu=0), confirming that our modeling is accurate.

Conclusions:

Landing distance is highly dependent on the make of the aircraft, speed_air and height of the aircraft while passing through the threshold of the runway.

Our final model is as follows:

Landing distance = -6390.38 (se=109.84) + 427.44*aircraft make (se =19.17) + 82.14*speed_air (se=0.98) + 13.7*height (se=1)

The residuals from the model follows a normal distribution and hence our model is statistically confidant.