ABHISHEK DHANASETTY

BANA 6043 PROJECT – REPORT

2/21/17

Background: Flight landing

Motivation: To reduce the risk of landing overrun

Goal: To study what factors and how they would impact the landing distance of a commercial flight.

Chapter-1 - Understanding:

Understanding the objectives:

- Based on the information given to us we must analyze the data, so that it can reduce landing overrun.
- Factor's such as wind, speed, pitch, weight, temperature might affect landing overrun.

Understanding the data provided:

- Two files FAA1.xls and FAA2.xls are given for analysis.
- Entire database contains two aircrafts they are "Airbus" and "Boeing".
- Factors that are given for analysis are speed of flight on ground and air, height of flight from ground, duration of flight, and pitch.

Determining the goals:

- Study and analyze the factors that might affect the distance of flights Airbus and Boeing.

Requirements:

- SAS
- MS Excel

Assumptions:

- Assuming, that speed of flight in air and ground is nearly the same.
- As it is a commercial flight I'm assuming that number of passengers does not affect our analysis.

Data Understanding:

Collecting Initial Data:

- Given two excel files FAA1.xls and FAA2.xls
    - FAA1.xls contains:
        - Variables: aircraft, duration, no_pasg, speed_ground, speed_air, height, pitch, and distance.
        - Number of rows: 800
    - FAA2.xls contains:
        - Variables: aircraft, no_pasg, speed_ground, speed_air, height, pitch, and distance.
        - Number of rows: 150

Importing data into SAS:

- I'm using SAS studio, university edition
- I created a folder locally in SAS studio and named it Project
- Uploaded files FAA1.xls and FAA2.xls in Project folder.
- Importing the files in SAS:

```
PROC IMPORT OUT= WORK.File1
    DATAFILE= "~/Project/FAA1.xls"
    DBMS=xls REPLACE;
GETNAMES=YES;
RUN;
```

```
PROC IMPORT OUT= WORK.File2
    DATAFILE= "~/Project/FAA2.xls"
    DBMS=xls REPLACE;
GETNAMES=YES;
RUN;
```

- Renamed the files as FLIGHT_LANDING_FILE1 and FLIGHT_LANDING_FILE2

```
DATA FLIGHT_LANDING_FILE1;
    SET WORK.FILE1;
RUN;
DATA FLIGHT_LANDING_FILE2;
    SET WORK.FILE2;
RUN;
```

Exploring the data:

- **Aircraft:** The make of an aircraft (Airbus or Boeing)
- **Duration:** Flight duration in minutes during takeoff and landing. The duration of a normal flight should always be greater than 40 mins
- **No_pasg:** The number of passengers in the flight
- **Speed_ground:** The ground speed of an aircraft in miles per hour 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:** The air speed of an aircraft in miles per hour 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:** The height of an aircraft in meters 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:** Pitch angle in degrees of an aircraft when it is passing over the threshold of the runway.
- **Distance:** The landing distance of an aircraft in feet. 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 2 – Data Preparation:

Merging the datasets:

- Using the SET statement to concatenate the data sets and named the merged file as FLIGHT_LANDING_FILE

```
DATA FLIGHT_LANDING_FILE;
    SET FLIGHT_LANDING_FILE1 FLIGHT_LANDING_FILE2;
RUN;
```

Handling unknown values:

- Handling unknown numbers in the merged file and named it as FLIGHT_LANDING_UNK_NUMBERS.

```
DATA FLIGHT_LANDING_UNK_NUMBERS;
    SET WORK.FLIGHT_LANDING_FILE;
        ARRAY UNKNOWN_NUMBERS _NUMERIC_;
    DO OVER UNKNOWN_NUMBERS;
        IF UNKNOWN_NUMBERS = . THEN UNKNOWN_NUMBERS
= 0;
    END;
RUN;
```

- Handling unknown characters in the dataset and named it as FLIGHT_LANDING_UNK_CHARACTERS

```
DATA FLIGHT_LANDING_UNK_CHARACTERS;
    SET WORK.FLIGHT_LANDING_UNK_NUMBERS;
        ARRAY UNKNOWN_CHARS _CHARACTER_;
    DO OVER UNKNOWN_CHARS;
        IF UNKNOWN_CHARS = ' ' THEN UNKNOWN_CHARS =
'UNK';
    END;
RUN;
```

<u>Removing duplicates:</u>

- Using the keyword NODUP, I'm removing all the duplicates from the dataset.

```
DATA FLIGHT_LANDING_NODUP;
    SET FLIGHT_LANDING_UNK_CHARACTERS;
RUN;
PROC SORT DATA=FLIGHT_LANDING_NODUP NODUP
OUT=FLIGHT_LANDING_NODUP;
    BY AIRCRAFT;
RUN;
```

<u>Selecting data:</u>

- Assuming as it is a commercial flight the number of passengers does not affect landing overrun.
    - So, dropping no_pasg from the dataset and renaming the dataset as COMPLENESS_CHECK.

```
DATA COMPLETENESS_CHECK;
    SET FLIGHT_LANDING_NODUP;
    DROP NO_PASG;
RUN;
```

Deleting the missing values:

- Deleting the missing values if there are any from the dataset and renamed the dataset as *FILTERED_DATA.*

```
DATA FILTERED_DATA;
    SET COMPLETENESS_CHECK;
    IF CMISS(OF _ALL_) THEN DELETE;
RUN;
```

Find the count:

- Earlier in handling unknown values, we replaced missing integers with 0 and missing characters as ‘UNK’
- Using procedure FREQ, we find the count of missing values.

```
PROC FORMAT;
    VALUE ZEROF
    0 = 'ZERO'
    OTHER = 'NOT ZERO';
QUIT;
PROC FREQ DATA=FILTERED_DATA;
    FORMAT _NUMERIC_ ZEROF.;
    TABLES _NUMERIC_/MISSING;
RUN;
```

Duration:

duration				
duration	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Zero	151	15.88	151	15.88
Not Zero	800	84.12	951	100.00

Speed_ground:

speed_ground				
speed_ground	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Zero	1	0.11	1	0.11
Not Zero	950	99.89	951	100.00

Speed_air:

speed_air				
speed_air	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Zero	712	74.87	712	74.87
Not Zero	239	25.13	951	100.00

Height:

height				
height	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Not Zero	950	99.89	950	99.89
Zero	1	0.11	951	100.00

Pitch:

pitch				
pitch	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Zero	1	0.11	1	0.11
Not Zero	950	99.89	951	100.00

- We see that most of the values in speed_air is missing.
- Finding the correlation value between speed of air and speed of ground.

```
PROC CORR DATA=FILTERED_DATA;
    VAR SPEED_AIR;
    WITH SPEED_GROUND;
    TITLE CO-RELATION COEFFIECIENT OF SPEED OF AIR WITH
RESPECT TO SPEED OF GROUND;
RUN;
```

Pearson Correlation Coefficients, N = 951 Prob > |r| under H0: Rho=0	
	speed_air
speed_ground speed_ground	0.75446 <.0001

- As speed of ground is closely related to speed of ground, I'm replacing the missing values of speed of air with speed of ground.

```
DATA COMPLETENESS_CHECK_2;
    SET FILTERED_DATA;
    /*DROP SPEED_AIR; */
    IF SPEED_AIR = 0.0 THEN SPEED_AIR = SPEED_GROUND;
RUN;
```

Validity Check:

- Based on the condition given to us I'm validating the given dataset.

```
DATA VALIDITY_DATA;
    SET COMPLETENESS_CHECK_2;
    IF DURATION <= 40 THEN DELETE; /* ALWAYS BE */
    IF SPEED_GROUND > 140 THEN DELETE;
```

```
        IF SPEED_GROUND < 30 THEN DELETE;
        IF SPEED_AIR > 140 THEN DELETE;
        IF SPEED_AIR < 30 THEN DELETE;
        IF HEIGHT <= 6 THEN DELETE; /* ATLEAST */
        IF DISTANCE > 6000 THEN DELETE;
RUN;
```

So, after handling the missing vales and cleaning the data, we remain with the following data:

Obs	aircraft	duration	speed_ground	speed_air	height	pitch	distance
1	airbus	132.46942492	100.01055305	100.891677	41.033010684	4.2975016214	2554.8330623
2	airbus	109.19713407	82.483044979	82.483044979	30.140024889	4.0896284195	1321.0000654
3	airbus	93.952926911	96.878686347	98.085883143	29.178095121	3.967524021	2008.2207232
4	airbus	45.635423091	93.793862117	93.793862117	42.830935448	4.271324799	2003.4386496
5	airbus	99.148062915	97.096913917	96.913737767	33.144245658	3.5162975656	2060.1694249
6	airbus	199.43713308	58.10907688	58.10907688	24.20102213	3.6341657268	418.01946274
7	airbus	141.96833358	85.849362338	85.849362338	46.468182053	3.4015527555	1492.6717204
8	airbus	203.13135186	90.261004686	90.261004686	42.318923044	3.2745717105	1446.8557482
9	airbus	199.79840009	61.712054098	61.712054098	34.976545351	4.0805575423	643.85634155
10	airbus	112.87149908	104.45540038	103.6715358	23.783587114	3.9026553246	2488.9984842
11	airbus	148.49500413	99.874522521	98.724063607	39.520425649	3.9041206536	2404.7430929
12	airbus	89.075548734	74.212201979	74.212201979	25.747432194	3.6751924109	852.77611439
13	airbus	109.79101574	85.88079228	85.88079228	33.468314922	3.462927709	1408.5685921
14	airbus	209.19366153	50.812930767	50.812930767	38.841316346	4.0338980996	566.92692802
15	airbus	143.52069259	78.580289917	78.580289917	33.75012614	2.2844801423	1011.337258
16	airbus	147.81402601	84.685046724	84.685046724	45.643491432	3.6631635685	1454.2976548
17	airbus	217.12308376	94.81425838	97.631341718	33.058365517	3.8235547791	2017.6011486
18	airbus	142.46940356	63.918959665	63.918959665	30.153465712	3.0650903012	428.99182821
19	airbus	99.019165671	91.236709563	91.236709563	15.59892741	3.3688761342	1450.5750216
20	airbus	198.8694626	61.127025526	61.127025526	22.51924496	3.4452935393	397.54283343
21	airbus	131.73109556	131.03518222	131.3379485	28.277965541	3.6601936464	4896.2946083
22	airbus	274.21773493	84.136858635	84.136858635	28.401059676	3.593594411	944.66620396
23	airbus	160.39281504	103.27582495	105.18709549	54.198540346	3.95212311	2837.0808498
24	airbus	79.343951801	92.534347718	92.534347718	22.228410408	3.5995757116	1816.9775926
25	airbus	87.345969963	87.926511371	87.926511371	28.790896161	4.011354326	1555.4007483

752	boeing	92.450287837	101.5435039	101.84980019	18.169394071	4.8226833856	2573.0532475
753	boeing	106.10804784	97.174871746	98.637718197	38.441153122	3.5773087522	2628.5363159
754	boeing	158.4190187	65.642893609	65.642893609	26.96473354	3.4595436201	1154.2199175
755	boeing	146.54974486	58.288973718	58.288973718	27.968682076	4.3118895338	919.04747904
756	boeing	63.32952055	132.78467664	132.9114649	18.177030219	4.1106642414	5343.2009539
757	boeing	213.98450886	80.394057703	80.394057703	16.962413199	4.0980200281	1531.2870582
758	boeing	173.75152907	68.462817944	68.462817944	35.027645688	4.9566852937	1098.2848748
759	boeing	141.64645173	64.99673603	64.99673603	12.720887595	4.3681656285	960.18473411
760	boeing	84.368450612	75.116022418	75.116022418	34.879568015	3.9307575129	1270.3031785
761	boeing	146.18126345	65.334357647	65.334357647	32.666905195	4.4053438458	968.43681985
762	boeing	147.99216651	80.992495294	80.992495294	21.705147966	3.3636279418	1485.5412586
763	boeing	177.62133285	65.570869498	65.570869498	8.8251728965	3.6673684977	839.26516938
764	boeing	192.27232801	71.753544367	71.753544367	15.870447594	4.4654794397	920.71562998
765	boeing	136.64928054	61.202449961	61.202449961	28.405156035	3.9999589106	1123.4040619
766	boeing	160.33576673	78.270298206	78.270298206	32.732115723	3.9023897774	1346.5410738
767	boeing	233.43123856	34.222063657	34.222063657	28.629155926	4.7888425657	955.9096666
768	boeing	168.63061318	77.535257271	77.535257271	21.047219906	4.2192924831	1375.0200626
769	boeing	94.319896766	83.635110177	83.635110177	23.466406143	3.5487071321	1499.7242663
770	boeing	164.77436392	53.113984349	53.113984349	27.920433284	4.3227229045	819.23600163
771	boeing	130.75025897	65.760129141	65.760129141	28.666938655	4.4815795592	893.57795417
772	boeing	142.42032129	80.433449799	80.433449799	19.787504785	4.016768468	1445.8852618
773	boeing	183.95750281	45.857193744	45.857193744	39.804470778	3.96076458	1050.550585
774	boeing	147.90533612	90.499523559	90.499523559	31.401174198	4.3570430756	1853.9479702
775	boeing	134.21929344	43.124453971	43.124453971	36.98798361	3.8493425417	977.3980544
776	boeing	137.92287323	47.148822575	47.148822575	46.42619482	5.1183234022	1128.968076
777	boeing	99.681502958	121.83713667	120.95340518	33.184596582	3.8674761307	4427.670764
778	boeing	68.201536698	67.955636835	67.955636835	39.479483513	4.7795649622	1344.3403491
779	boeing	83.512569699	78.625488084	78.625488084	17.342190802	3.8634086267	1104.404226
780	boeing	212.29018	89.533713205	90.626181428	35.494742904	4.0010380484	2148.1079287
781	boeing	153.8344532	126.83927854	126.11864818	20.547833848	4.3345575101	4736.6045811

Completeness check:

```
proc means data=validity_data n nmiss min mean p1 p5 p10 p25 p50 p75 p95 p99;
    var duration speed_ground speed_air height distance;
run;
proc freq data=validity_data;
    tables aircraft/missing list;
run;
```

The MEANS Procedure

Variable	Label	N	N Miss	Minimum	Mean	1st Pctl	5th Pctl	10th Pctl	25th Pctl	50th Pctl	75th Pctl	95th Pctl	99th Pctl
duration	duration	781	0	41.9493694	154.7757191	54.2182399	78.4556628	92.6481219	119.6314577	154.2845505	189.6629425	233.8024979	274.2177349
speed_ground	speed_ground	781	0	33.5741041	79.6397499	39.7257113	47.8821171	55.0275560	66.1925304	79.7939604	92.1314349	111.6739179	125.2123041
speed_air	speed_air	781	0	33.5741041	79.6585394	39.7257113	47.8821171	55.0275560	66.1925304	79.7939604	92.3863588	110.6599156	125.9869146
height	height	781	0	6.2275178	30.4549525	9.6972160	14.6030868	17.9047677	23.5944766	30.2165682	36.9879836	46.8587929	53.4386181
distance	distance	781	0	41.7223127	1541.20	280.8044030	498.7680133	677.5833959	919.0474790	1273.66	1960.43	3437.66	4795.64

The FREQ Procedure

aircraft				
aircraft	Frequency	Percent	Cumulative Frequency	Cumulative Percent
airbus	394	50.45	394	50.45
boeing	387	49.55	781	100.00

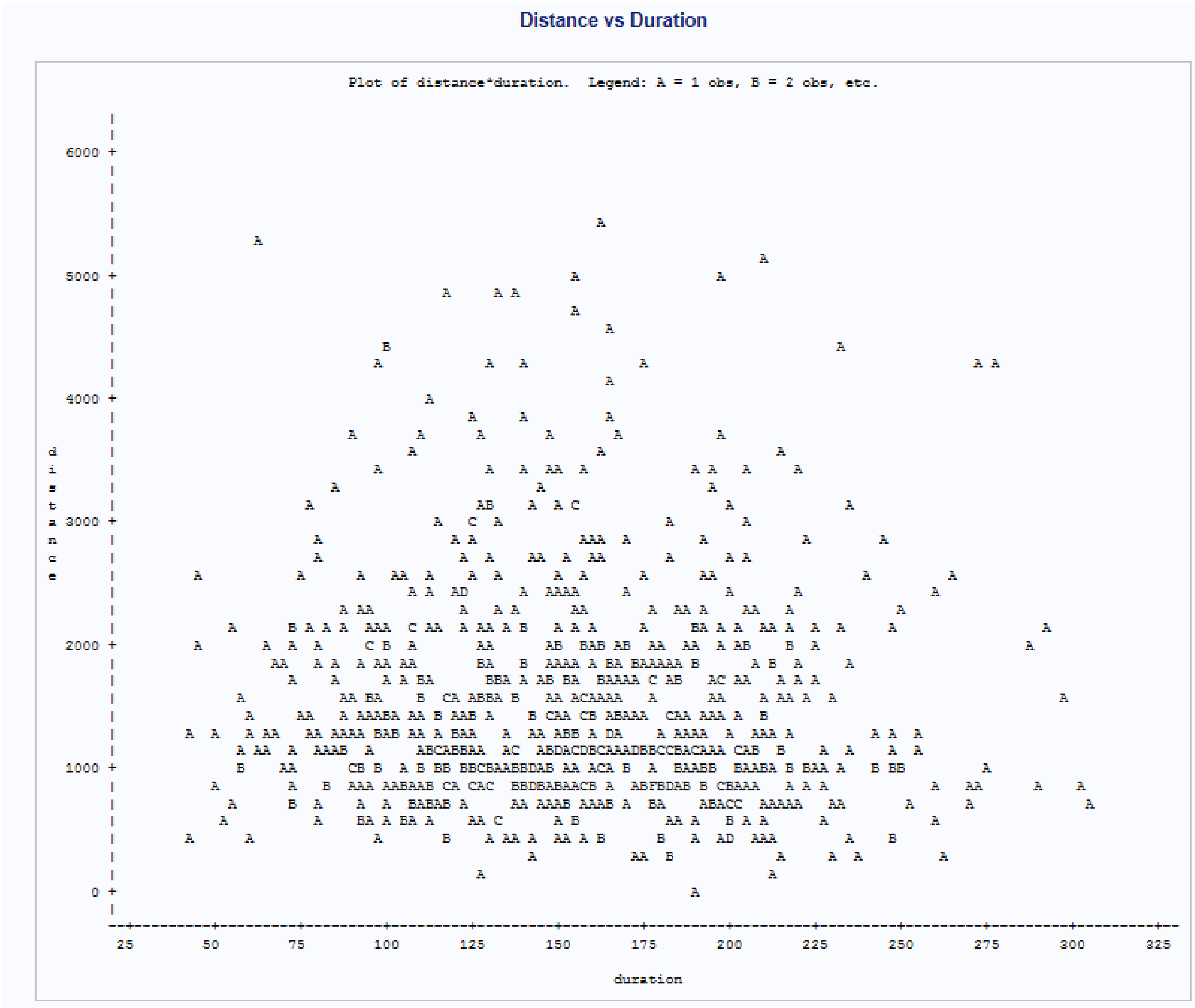
The final number of observations remaining are 781, after cleaning / pruning and the abnormal values are also removed.

Chapter 3 (Statistical Modeling):

- As our file goal is related to the distance I'm going to compare each and every variable with respect to distance.

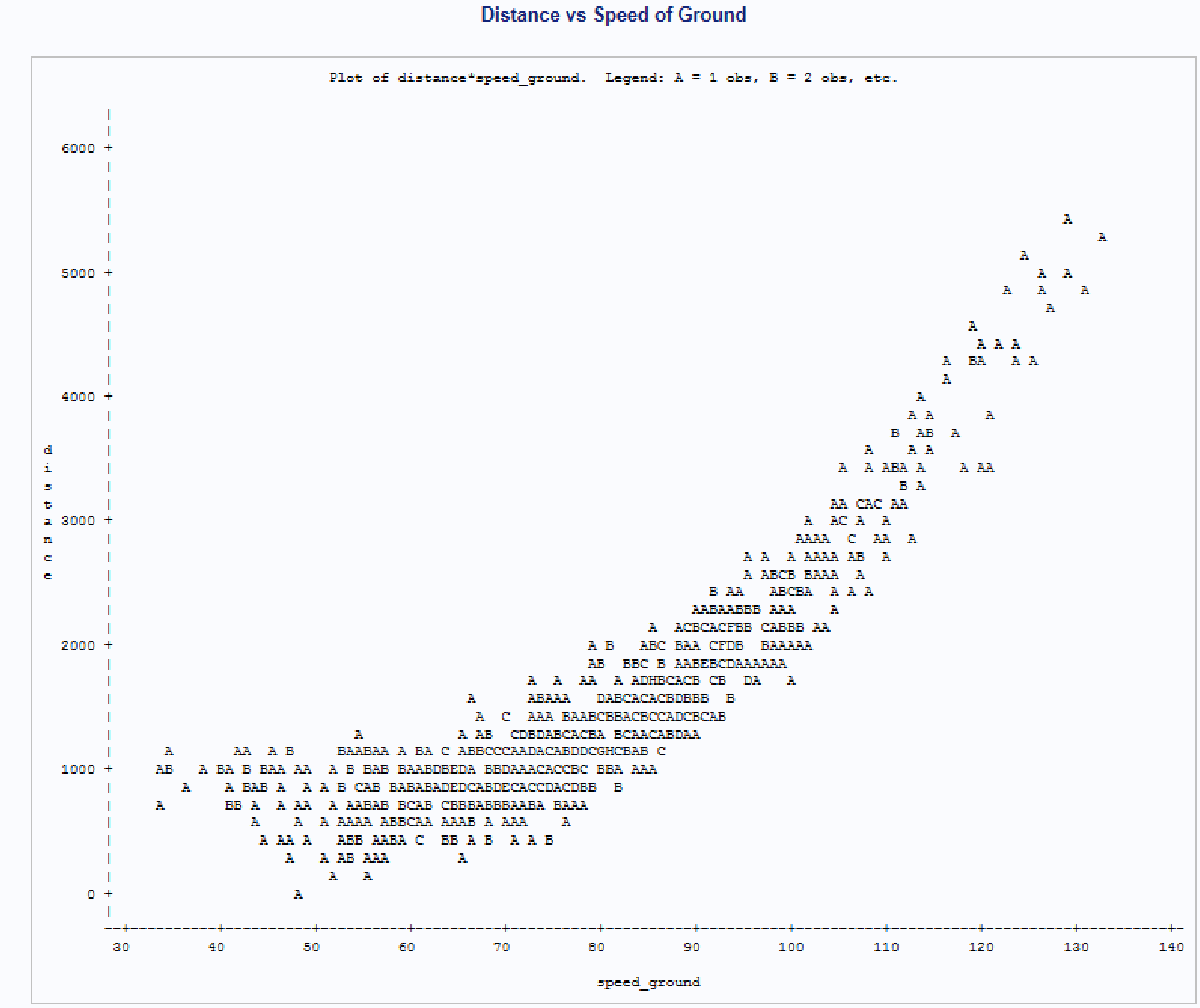
Plotting distance vs duration:

```
proc plot data= validity_data;
    plot distance * duration;
    title Distance vs Duration;
run;
```



Plotting distance vs speed of flight on ground:

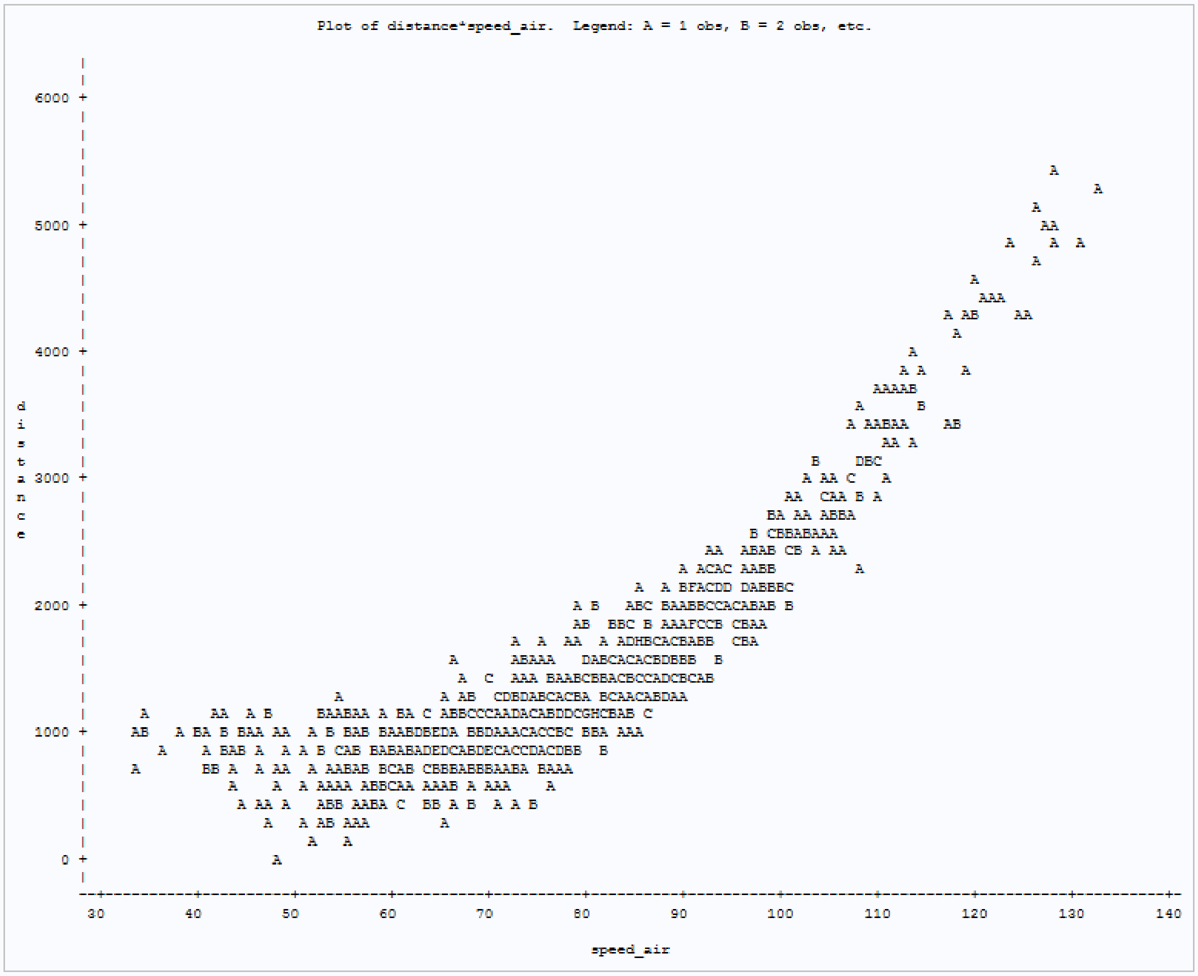
```
proc plot data= validity_data;
    plot distance * speed_ground;
    title Distance vs Speed of Ground;
run;
```



<u>Plotting distance vs speed of flight in air:</u>

```
proc plot data= validity_data;
    plot distance * speed_air;
    title Distance vs Speed of air;
run;
```

Distance vs Speed of air



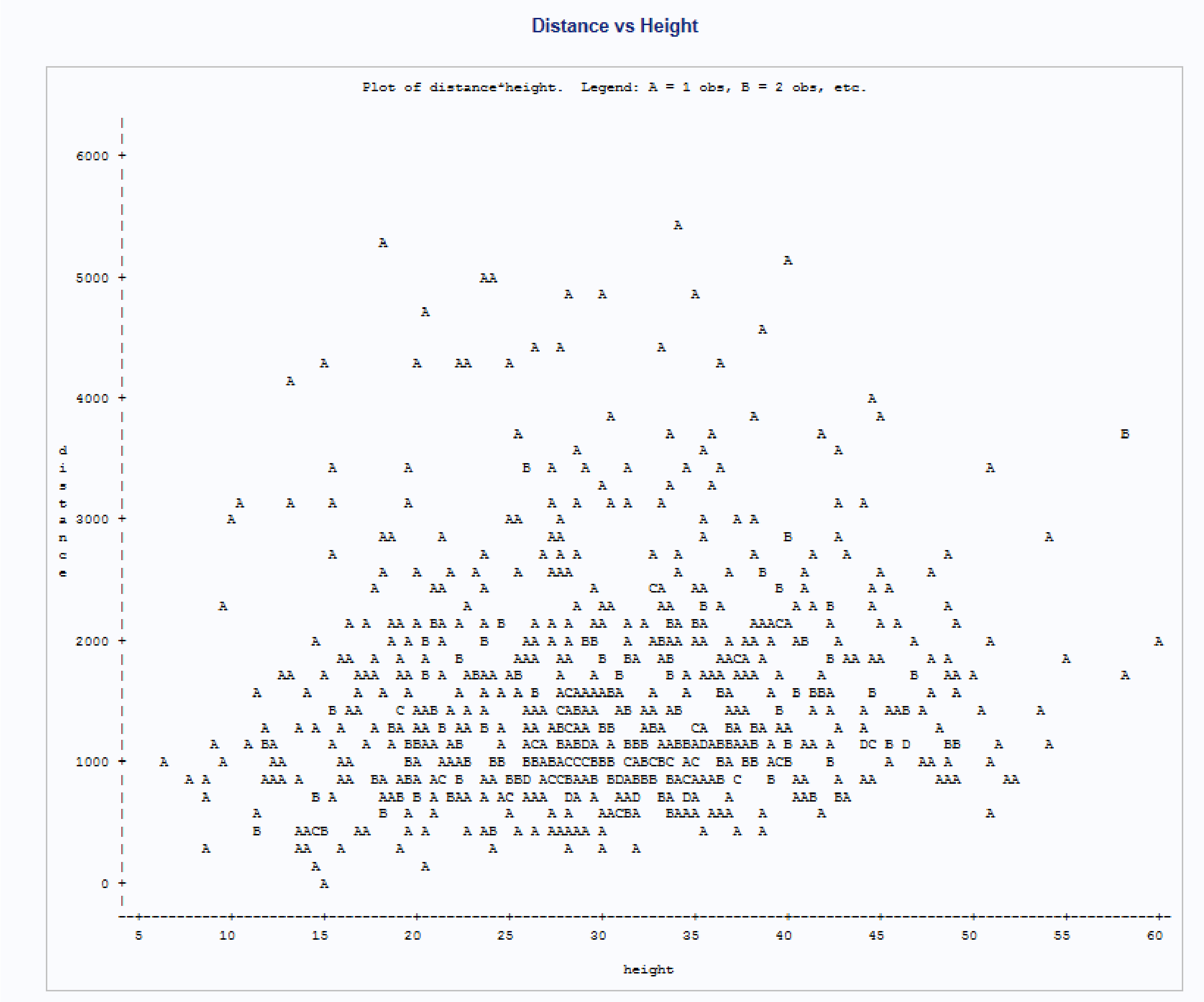
__Plotting distance vs height from the sea level:__

```
proc plot data= validity_data;

plot distance * height;

title Distance vs Height;
```

run;



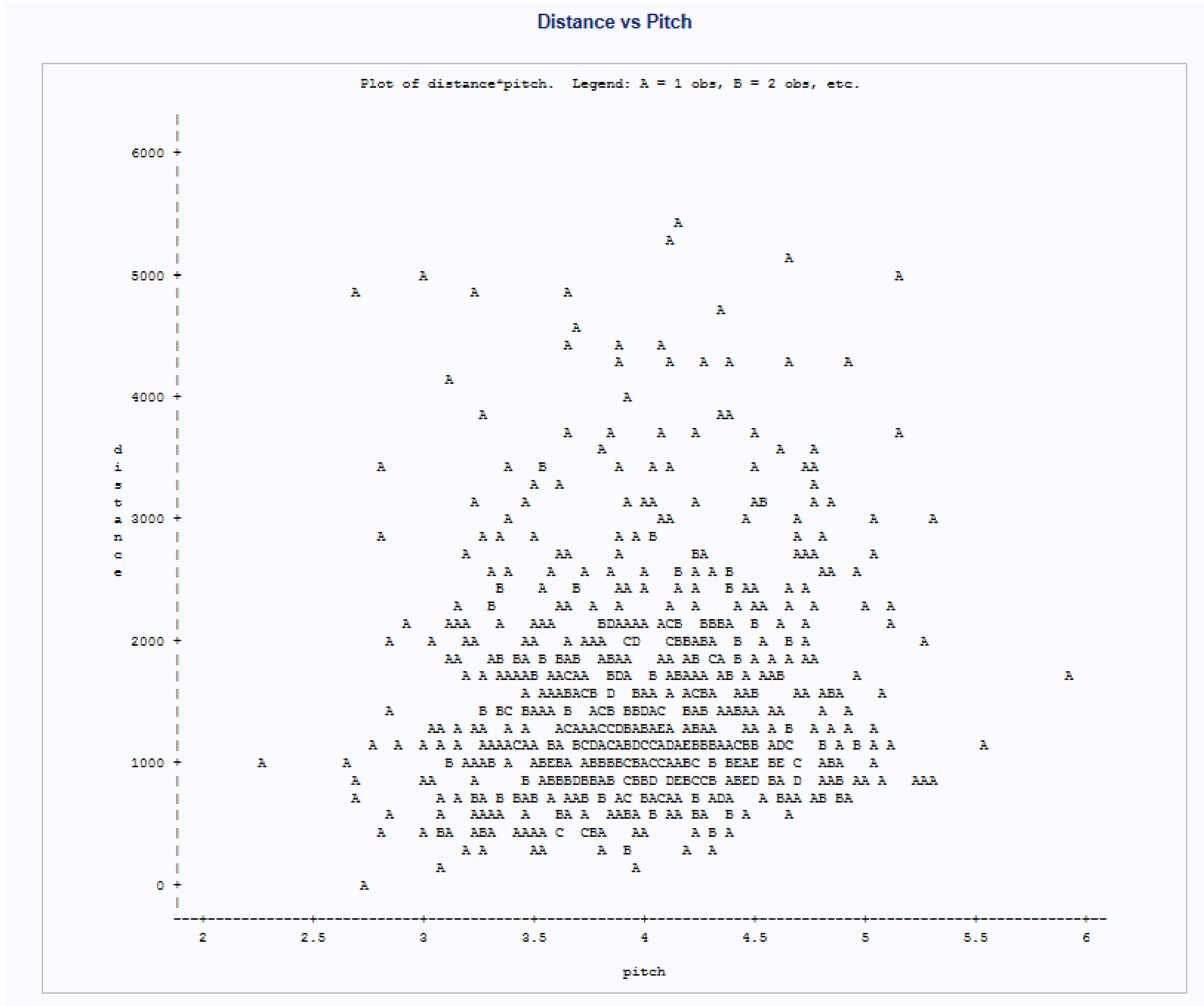
<u>Plotting distance with respect to pitch:</u>

proc plot data= validity_data;

plot distance * pitch;

title Distance vs Pitch;

```
run;
```



__Univariate analysis:__

```
proc sort data=validity_data;
    by descending aircraft speed_air speed_ground height pitch distance;
run;
proc univariate data=validity_data normal;
    histogram/normal;
    qqplot;
run;
```

Using the above univariate statement, we have performed a univariate analysis for each numeric variable in the data set.

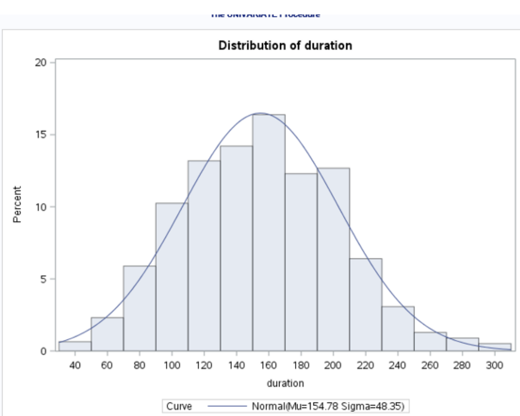
__Duration:__

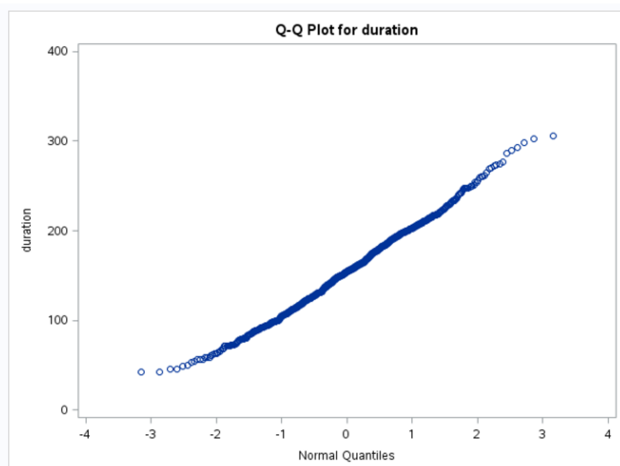
Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.02877447	Pr > D	0.118
Cramer-von Mises	W-Sq	0.09426269	Pr > W-Sq	0.137
Anderson-Darling	A-Sq	0.67659005	Pr > A-Sq	0.081

Moments			
N	781	Sum Weights	781
Mean	154.775719	Sum Observations	120879.837
Std Deviation	48.3499237	Variance	2337.71512
Skewness	0.18986566	Kurtosis	-0.1958773
Uncorrected SS	20532681.4	Corrected SS	1823417.79
Coeff Variation	31.2387007	Std Error Mean	1.73009629

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	54.2182	42.2970
5.0	78.4557	75.2472
10.0	92.6481	92.8128
25.0	119.6315	122.1642
50.0	154.2846	154.7757
75.0	189.6629	187.3872
90.0	214.7272	216.7386
95.0	233.8025	234.3043
99.0	274.2177	267.2545

Basic Statistical Measures			
Location		Variability	
Mean	154.7757	Std Deviation	48.34992
Median	154.2846	Variance	2338
Mode	.	Range	263.67234
		Interquartile Range	70.03148





We can see that it is positively skewed, and we can see that few outliers are present as well.

Speed_ground:

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	79.63975
Std Dev	Sigma	18.89717

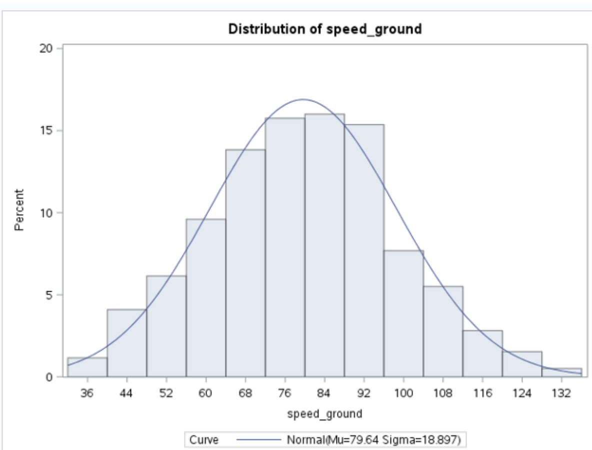
Moments			
N	781	Sum Weights	781
Mean	79.6397499	Sum Observations	62198.6447
Std Deviation	18.897169	Variance	357.102994
Skewness	0.08770841	Kurtosis	-0.2378247
Uncorrected SS	5232024.84	Corrected SS	278540.336
Coeff Variation	23.7283128	Std Error Mean	0.67619387

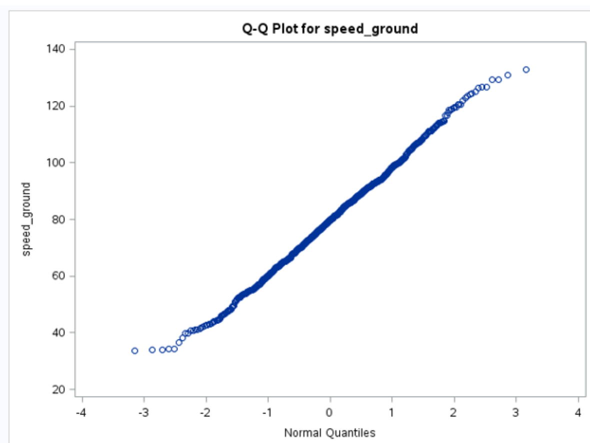
Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.01836482	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.03255509	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.28344119	Pr > A-Sq	>0.250

Basic Statistical Measures			
Location		Variability	
Mean	79.63975	Std Deviation	18.89717
Median	79.79396	Variance	357.10299
Mode	.	Range	99.21057
		Interquartile Range	25.93890

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	39.7257	35.6784
5.0	47.8821	48.5567
10.0	55.0276	55.4221
25.0	66.1925	66.8938
50.0	79.7940	79.6397
75.0	92.1314	92.3857
90.0	104.2951	103.8574
95.0	111.6739	110.7228
99.0	125.2123	123.6011

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	117.7765	Pr > |t|	<.0001
Sign	M	390.5	Pr >= |M|	<.0001
Signed Rank	S	152685.5	Pr >= |S|	<.0001





We see that it is also positively skewed, and we also see that there are not many outliers. We can also observe that speed of ground passes all the normaility test as p value is greater that .05

Speed_air:

Moments			
N	781	Sum Weights	781
Mean	79.6585394	Sum Observations	62213.3193
Std Deviation	18.898451	Variance	357.151449
Skewness	0.08169985	Kurtosis	-0.2478637
Uncorrected SS	5234400.28	Corrected SS	278578.13
Coeff Variation	23.7243252	Std Error Mean	0.67623974

Basic Statistical Measures			
Location		Variability	
Mean	79.65854	Std Deviation	18.89845
Median	79.79396	Variance	357.15145
Mode	.	Range	99.33736
		Interquartile Range	26.19383

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	117.7963	Pr > |t|	<.0001
Sign	M	390.5	Pr >= |M|	<.0001
Signed Rank	S	152685.5	Pr >= |S|	<.0001

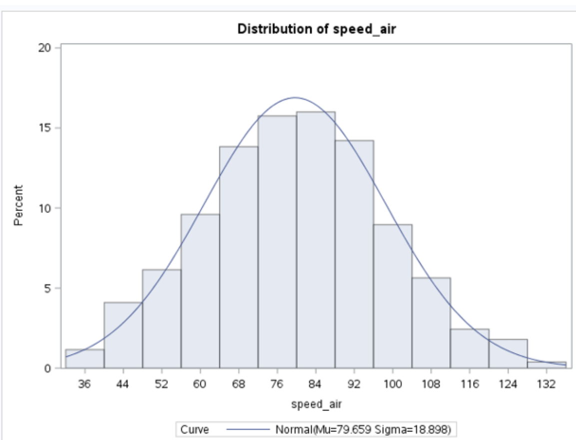
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.996815	Pr < W	0.1234
Kolmogorov-Smirnov	D	0.018661	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.027459	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.243803	Pr > A-Sq	>0.2500

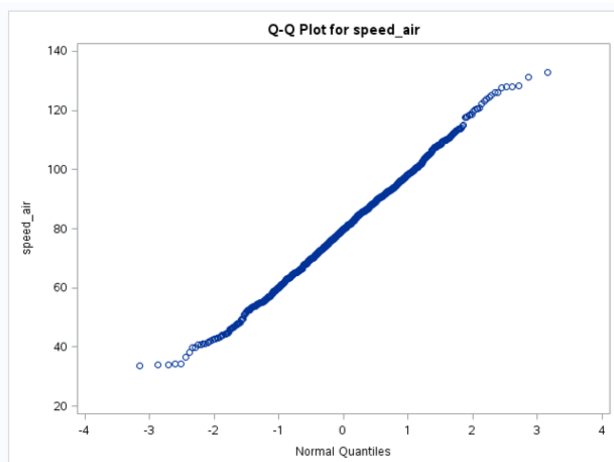
The UNIVARIATE Procedure
Fitted Normal Distribution for speed_air (speed_air)

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	79.65854
Std Dev	Sigma	18.89845

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.01866099	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.02745896	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.24380263	Pr > A-Sq	>0.250

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	39.7257	35.6942
5.0	47.8821	48.5734
10.0	55.0276	55.4392
25.0	66.1925	66.9117
50.0	79.7940	79.6585
75.0	92.3864	92.4054
90.0	104.4377	103.8779
95.0	110.6599	110.7437
99.0	125.9869	123.6229





This parameter is almost similar to speed of ground.

Height:

Moments			
N	781	Sum Weights	781
Mean	30.4549525	Sum Observations	23785.3179
Std Deviation	9.73964153	Variance	94.8606171
Skewness	0.1194969	Kurtosis	-0.3067498
Uncorrected SS	798372.01	Corrected SS	73991.2814
Coeff Variation	31.980485	Std Error Mean	0.34851177

Basic Statistical Measures			
Location		Variability	
Mean	30.45495	Std Deviation	9.73964
Median	30.21657	Variance	94.86062
Mode	.	Range	53.71845
		Interquartile Range	13.39351

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	87.38572	Pr > |t|	<.0001
Sign	M	390.5	Pr >= |M|	<.0001
Signed Rank	S	152685.5	Pr >= |S|	<.0001

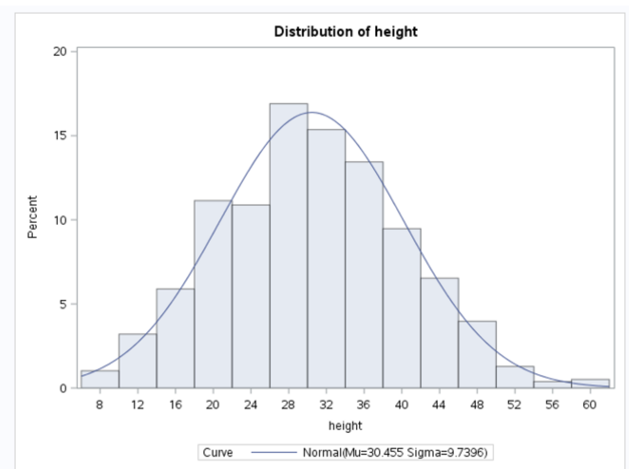
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.995908	Pr < W	0.0381
Kolmogorov-Smirnov	D	0.024857	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.062035	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.503241	Pr > A-Sq	0.2129

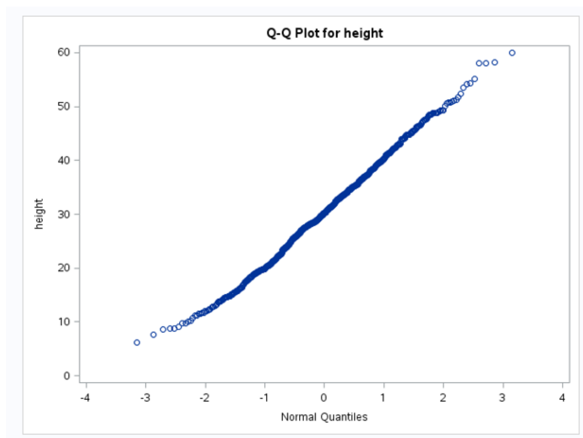
The UNIVARIATE Procedure
Fitted Normal Distribution for height (height)

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	30.45495
Std Dev	Sigma	9.739642

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.02485705	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.06203530	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.50324082	Pr > A-Sq	0.213

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	9.69722	7.79716
5.0	14.60309	14.43467
10.0	17.90477	17.97310
25.0	23.59448	23.88566
50.0	30.21657	30.45495
75.0	36.98798	37.02424
90.0	43.07741	42.93681
95.0	46.85879	46.47524
99.0	53.43862	53.11275





It is positively skewed and it follows normal distribution as p value is greater than .05.

<u>Pitch:</u>

The UNIVARIATE Procedure
Variable: pitch (pitch)

Moments			
N	781	Sum Weights	781
Mean	4.01412892	Sum Observations	3135.03468
Std Deviation	0.52236881	Variance	0.27286917
Skewness	-0.0088482	Kurtosis	-0.0903272
Uncorrected SS	12797.2713	Corrected SS	212.837956
Coeff Variation	13.0132545	Std Error Mean	0.01869183

Basic Statistical Measures			
Location		Variability	
Mean	4.014129	Std Deviation	0.52237
Median	4.014006	Variance	0.27287
Mode	.	Range	3.64230
		Interquartile Range	0.72900

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	214.7532	Pr > |t|	<.0001
Sign	M	390.5	Pr >= |M|	<.0001
Signed Rank	S	152685.5	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.998833	Pr < W	0.9037
Kolmogorov-Smirnov	D	0.015507	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.022015	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.193091	Pr > A-Sq	>0.2500

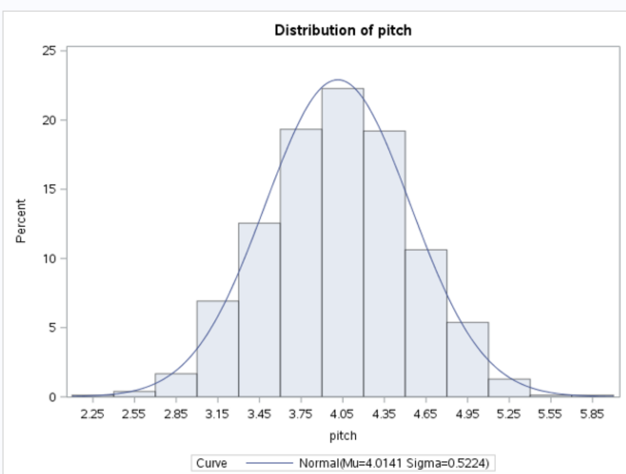
The UNIVARIATE Procedure
Fitted Normal Distribution for pitch (pitch)

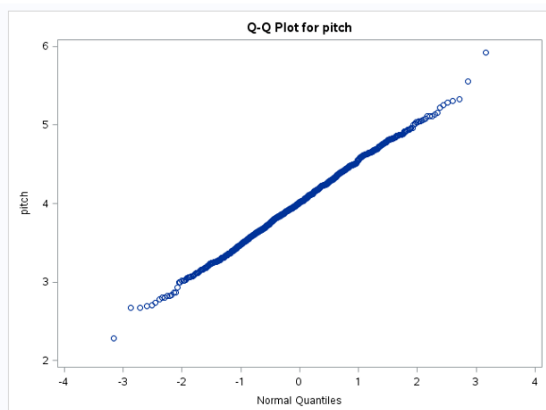
Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	4.014129
Std Dev	Sigma	0.522369

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.01550720	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.02201534	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.19309144	Pr > A-Sq	>0.250

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	2.79673	2.79892
5.0	3.16798	3.15491
10.0	3.31906	3.34469
25.0	3.65330	3.66180
50.0	4.01401	4.01413
75.0	4.38229	4.36646
90.0	4.68531	4.68357
95.0	4.85324	4.87335
99.0	5.15470	5.22934

The UNIVARIATE Procedure





Finding the correlation of each of the other variables with distance:

With speed of flight on ground level:

```
proc corr data=validity_data;
    var speed_ground;
    with distance;
    title Co-relation coeffiecient of speed of ground with respect to distance;
run;
```

Co-relation coeffiecient of speed of ground with respect to distance

The CORR Procedure

1 With Variables:	distance
1 Variables:	speed_ground

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	781	1541	904.59033	1203680	41.72231	5382	distance
speed_ground	781	79.63975	18.89717	62199	33.57410	132.78468	speed_ground

Pearson Correlation Coefficients, N = 781 Prob > |r| under H0: Rho=0	
	speed_ground
distance distance	0.86771 <.0001

With respect to speed of flight in air:

```
proc corr data=validity_data;
    var speed_air;
    with distance;
    title Co-relation coeffiecient of speed of air with respect to distance;
run;
```

Co-relation coeffiecient of speed of air with respect to distance

The CORR Procedure

1 With Variables:	distance
1 Variables:	speed_air

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	781	1541	904.59033	1203680	41.72231	5382	distance
speed_air	781	79.65854	18.89845	62213	33.57410	132.91146	speed_air

Pearson Correlation Coefficients, N = 781
Prob > |r| under H0: Rho=0

	speed_air
distance distance	0.86883 <.0001

With respect to duration of the flight:

```
proc corr data=validity_data;
        var duration;
        with distance;
        title Co-relation coeffiecient of duration with respect to distance;
run;
```

Co-relation coeffiecient of speed of air with respect to distance

The CORR Procedure

1 With Variables:	distance
1 Variables:	speed_air

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	781	1541	904.59033	1203680	41.72231	5382	distance
speed_air	781	79.65854	18.89845	62213	33.57410	132.91146	speed_air

Pearson Correlation Coefficients, N = 781 Prob > |r| under H0: Rho=0	
	speed_air
distance distance	0.86883 <.0001

With respect to height of the aircraft:

```
proc corr data=validity_data;
    var height;
    with distance;
    title Co-relation coeffiecient of height with respect to distance;
run;
```

Co-relation coeffiecient of height with respect to distance

The CORR Procedure

1 With Variables:	distance
1 Variables:	height

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	781	1541	904.59033	1203680	41.72231	5382	distance
height	781	30.45495	9.73964	23785	6.22752	59.94596	height

Pearson Correlation Coefficients, N = 781 Prob > |r| under H0: Rho=0	
	height
distance distance	0.10372 0.0037

With respect to pitch:

```
proc corr data=validity_data;

var pitch;

with distance;

title Co-relation coeffiecient of pitch with respect to distance;

run;
```

Co-relation coeffiecient of pitch with respect to distance

The CORR Procedure

1 With Variables:	distance
1 Variables:	pitch

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	781	1541	904.59033	1203680	41.72231	5382	distance
pitch	781	4.01413	0.52237	3135	2.28448	5.92678	pitch

Pearson Correlation Coefficients, N = 781 Prob > |r| under H0: Rho=0	
	pitch
distance distance	0.06868 0.0550

<u>To check which variables are highly correlated:</u>

```
proc corr data=validity_data;
run;
```

Pearson Correlation Coefficients, N = 781
Prob > |r| under H0: Rho=0

	duration	speed_ground	speed_air	height	pitch	distance
duration duration	1.00000	-0.04897 0.1716	-0.04643 0.1950	0.01112 0.7564	-0.04675 0.1918	-0.05138 0.1514
speed_ground speed_ground	-0.04897 0.1716	1.00000	0.99918 <.0001	-0.05167 0.1491	-0.05167 0.1491	0.86771 <.0001
speed_air speed_air	-0.04643 0.1950	0.99918 <.0001	1.00000	-0.05037 0.1596	-0.04938 0.1680	0.86883 <.0001
height height	0.01112 0.7564	-0.05167 0.1491	-0.05037 0.1596	1.00000	0.03474 0.3323	0.10372 0.0037
pitch pitch	-0.04675 0.1918	-0.05167 0.1491	-0.04938 0.1680	0.03474 0.3323	1.00000	0.06868 0.0550
distance distance	-0.05138 0.1514	0.86771 <.0001	0.86883 <.0001	0.10372 0.0037	0.06868 0.0550	1.00000

From the above histograms and this table we see that there exists a strong correlation between speed of ground and distance as well as speed of air and distance.
Pitch and height also have a significant correlation with distance, but there linear relationship is week.

<u>Chapter 4: (Data Modelling)</u>

<u>Regression:</u>

```
data final;
    set validity_data;
    if aircraft = 'boeing' then aircraft_type = 1;
    else aircraft_type = 0;
run;
proc reg data=final;
    model distance = aircraft_type speed_ground speed_air duration pitch height /vif;
```

output out=regression_output residual=residual_output;

run;

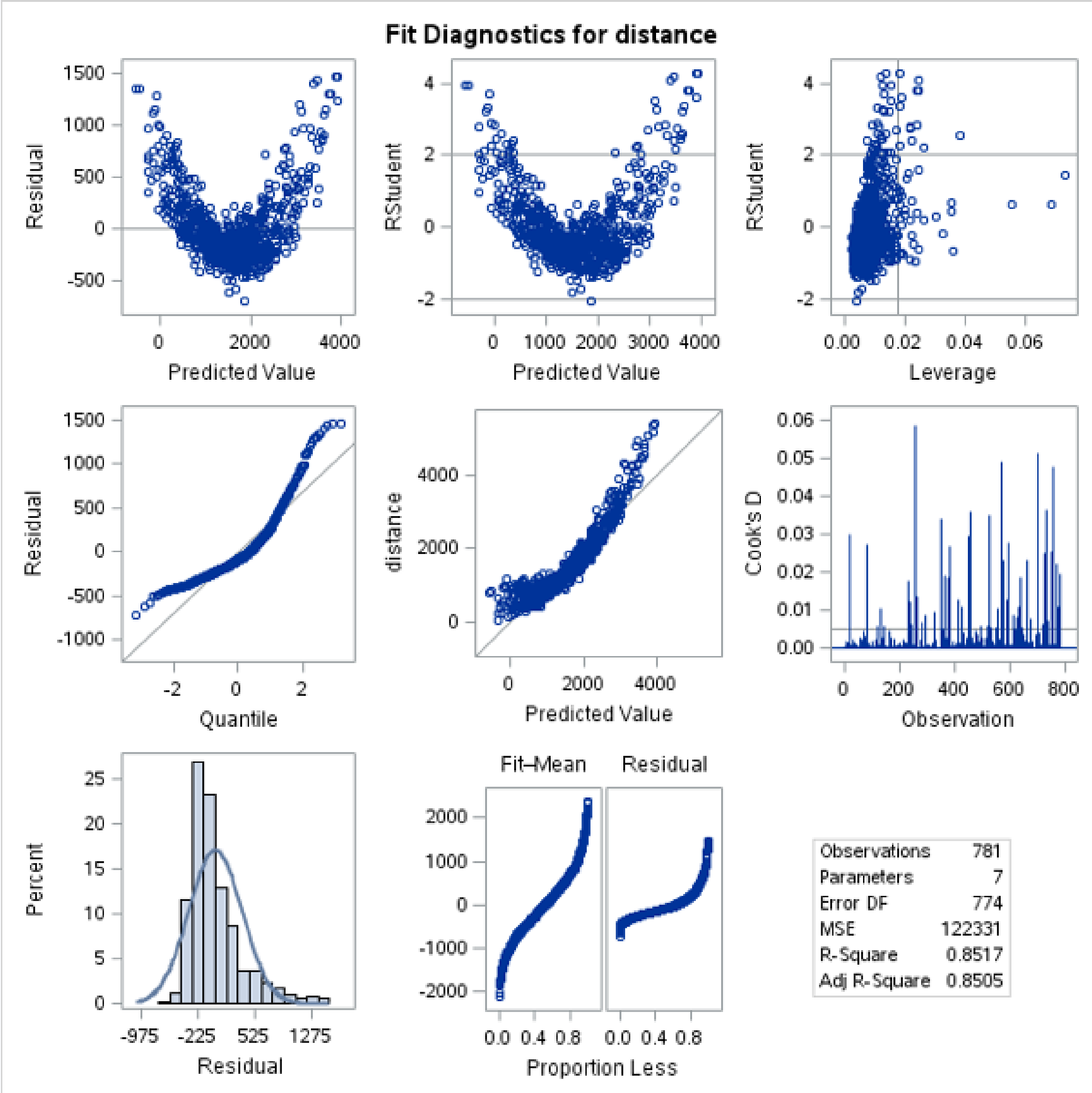
The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	781
Number of Observations Used	781

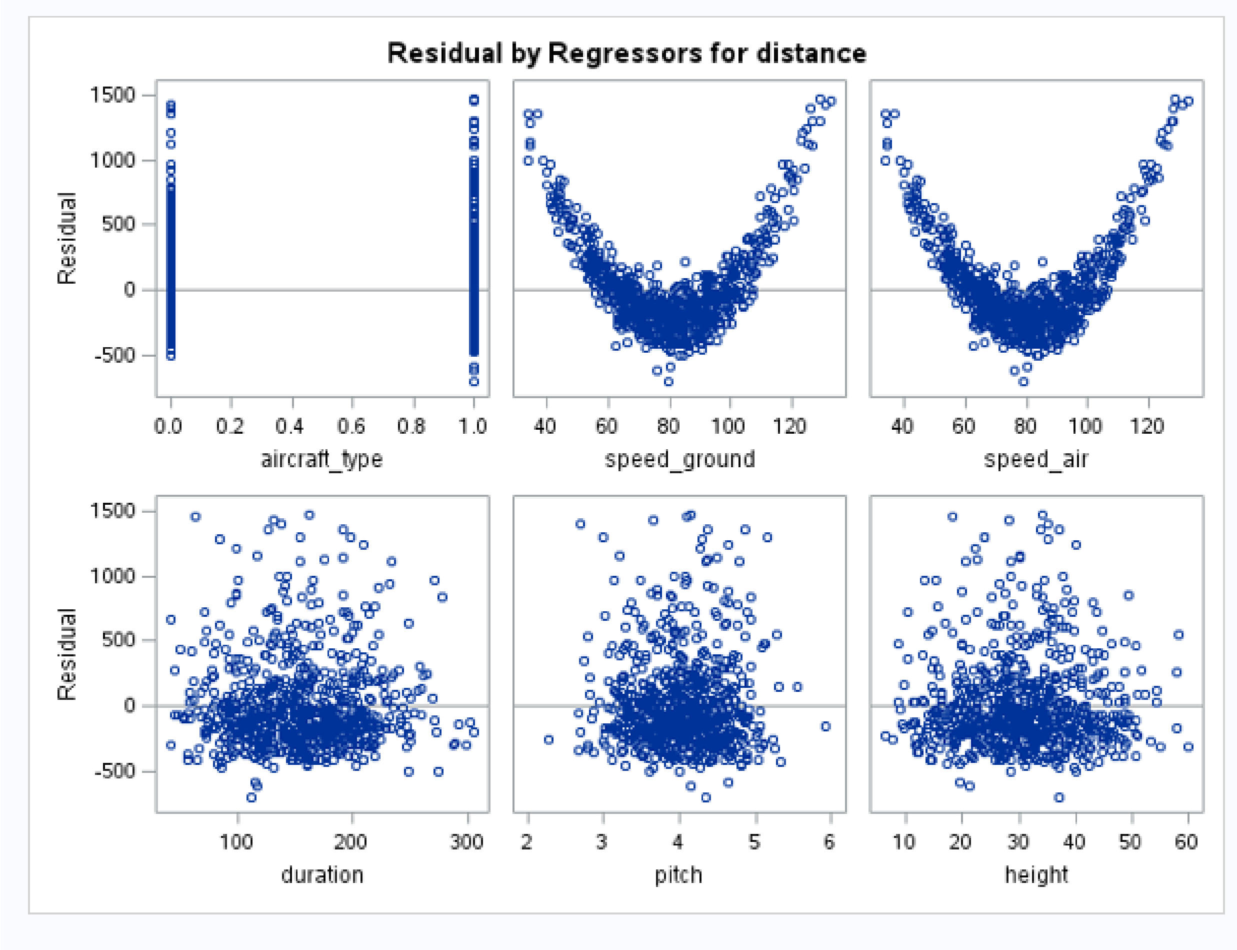
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	543577419	90596237	740.59	<.0001
Error	774	94683840	122331		
Corrected Total	780	638261260			

Root MSE	349.75783	R-Square	0.8517
Dependent Mean	1541.20394	Adj R-Sq	0.8505
Coeff Var	22.69381		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-2592.03191	130.22439	-19.90	<.0001	0
aircraft_type		1	489.43776	26.89198	18.20	<.0001	1.15416
speed_ground	speed_ground	1	0.81012	16.46389	0.05	0.9608	617.18855
speed_air	speed_air	1	41.77718	16.45768	2.54	0.0113	616.80727
duration	duration	1	0.01382	0.26032	0.05	0.9577	1.01014
pitch	pitch	1	16.28106	25.81229	0.63	0.5284	1.15922
height	height	1	14.14514	1.28942	10.97	<.0001	1.00561



Fit Diagnostics for distance
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
distance
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 781
Parameters 7
Error DF 774
MSE 122331
R-Square 0.8517
Adj R-Square 0.8505



From the above tables and figures we observe that:

- From the analysis of variance table, we know that there is some dependability between the independent and dependent variable.
- P value <0.05 implies rejection of null hypothesis.
- R and Adj. R square values are greater than 85 % which indicates significant variance in the dependent variable.
- From estimate table we can drop number of passengers, pitch and duration as they are not significant.
- The plot between standardized residuals and predicted value are not identically distributed.
- QQ plot shows us that the residuals are not following normality assumption

- Speed of ground and air is causing the nonlinearity as we see it in our residual vs predicted value.
- As speed of ground and speed of air are almost similar we observe a high variance inflation.

As speed of ground and speed of air are almost similar I am going to consider only speed of ground for my further analysis.

So, speed of ground, height might affect landing distance.

The equation is as follows:

Y = a0+a1*x1+a2*x2+ a3*x3+error

Y is distance.

a0 = -2592.03

a1 = 489.43

a2=0.81

a3=14.14

x1 = aircraft_type

x2 = speed_ground

x3 = height

distance = -2592.03 + 489.43(aircraft_type) + 0.81(speed_ground) + 14.14(height)

As per the residual plots:

Squaring speed of ground:

```
data final_dataset;
    set final;
    speed_ground_squared = speed_ground * speed_ground;
run;
proc reg data=final_dataset;
    model distance = aircraft_type speed_ground_squared height / vif;
    output out=regression residual=residual_output2;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	781
Number of Observations Used	781

Analysis of Variance

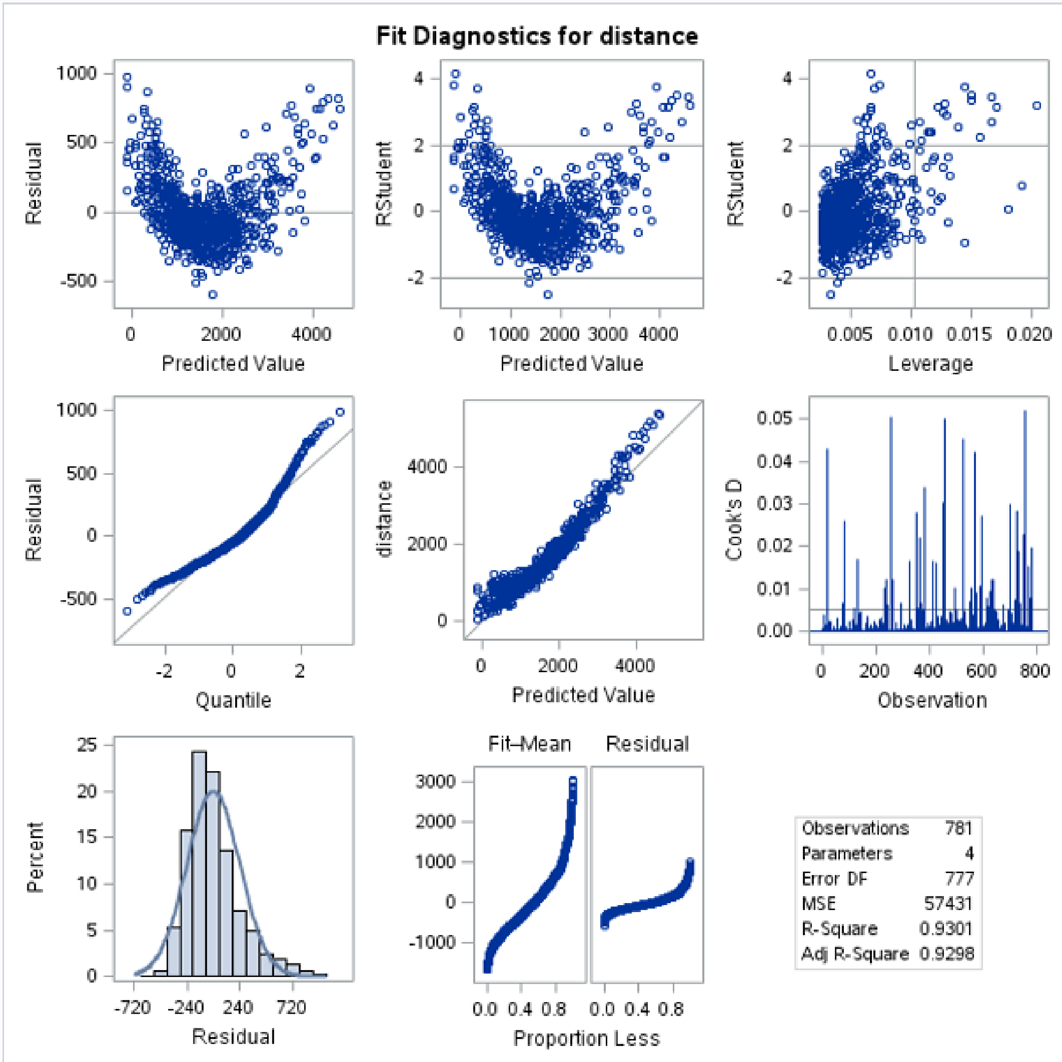
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	593637596	197879199	3445.53	<.0001
Error	777	44623663	57431		
Corrected Total	780	638261260			

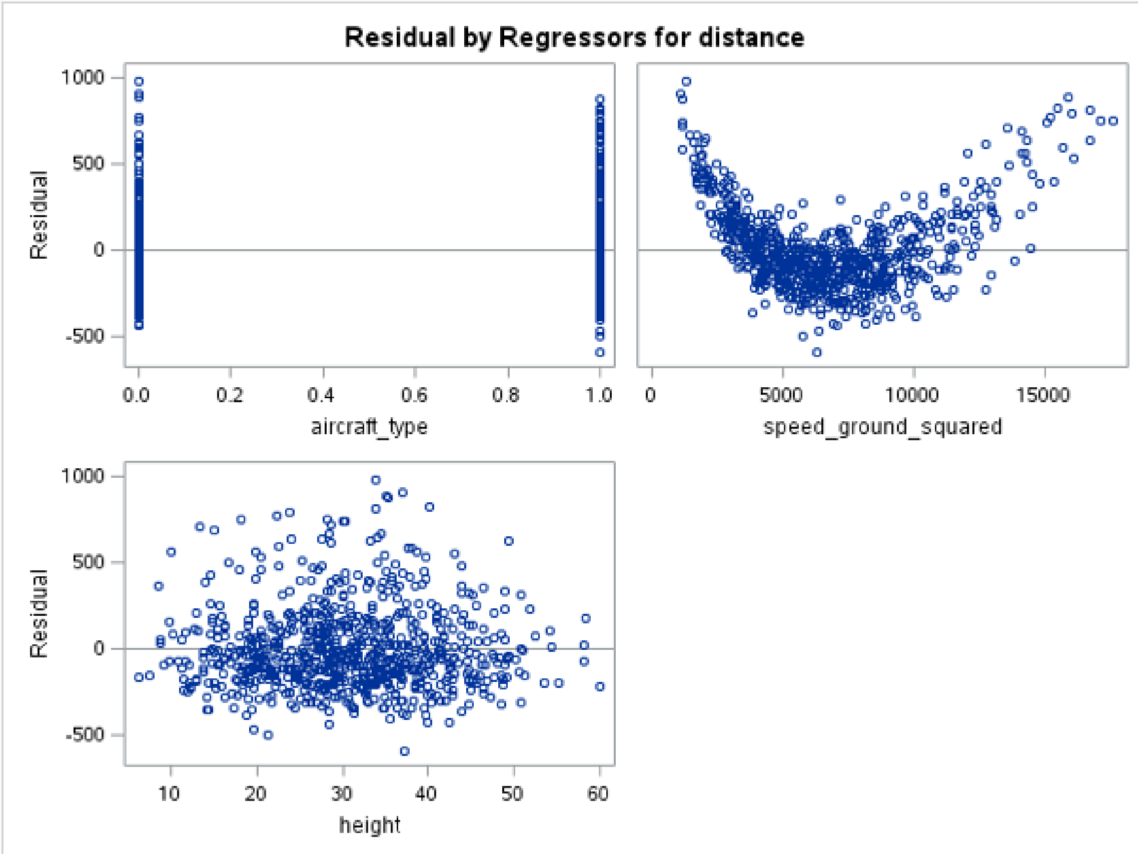
Root MSE	239.64706	R-Square	0.9301
Dependent Mean	1541.20394	Adj R-Sq	0.9298
Coeff Var	15.54934		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-955.80518	35.81600	-26.69	<.0001	0
aircraft_type		1	462.32562	17.15888	26.94	<.0001	1.00090
speed_ground_squared		1	0.27390	0.00279	98.09	<.0001	1.00306
height	height	1	14.21832	0.88218	16.12	<.0001	1.00266

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance





Observations:

- The R and Adj. R value is now 93 %
- All the variables are significant and have an intuitive sign.
- Variable inflation is almost equal to one which implies there no multi-collinearity.
- The standardized residual plot and predicted value are almost identical which implies the randomness in variance is reduced slightly.

The equation is as follows:

Y = a0 + a1*x1+a2*x2^2+a3*x3

Y is distance

a0 = -955.80

a1 = 462.32

a2 = 0.27

a3 = 14.21

Distance = -955.80 + 462.32*aircraft_type + 0.27*speed_ground^2 + 14.21 * height.

Plotting normal plot:

```
proc univariate data=regression normal plot;

run;
```

Variable: residual_output2 (Residual)

Moments			
N	781	Sum Weights	781
Mean	0	Sum Observations	0
Std Deviation	239.185754	Variance	57209.825
Skewness	1.10901002	Kurtosis	1.6133627
Uncorrected SS	44623663.5	Corrected SS	44623663.5
Coeff Variation	.	Std Error Mean	8.5587392

Basic Statistical Measures			
Location		Variability	
Mean	0.0000	Std Deviation	239.18575
Median	-45.5627	Variance	57210
Mode	.	Range	1578
		Interquartile Range	281.22080

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-69.5	Pr >= |M|	<.0001
Signed Rank	S	-19246.5	Pr >= |S|	0.0022

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.933154	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.095111	Pr > D	<0.0100
Cramer-von Mises	W-Sq	2.275496	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	13.81485	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Level	Quantile
100% Max	980.9296
99%	769.4526
95%	475.0715
90%	331.8965
75% Q3	115.6057
50% Median	-45.5627
25% Q1	-165.6151
10%	-253.9316
5%	-308.2868
1%	-388.3894
0% Min	-597.1629

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-597.163	499	820.419	528
-504.617	618	871.309	748
-466.516	562	885.970	256
-442.292	22	904.115	84
-432.395	59	980.930	349

From the above data:

- Residuals do not have zero mean
- P value < .05 which means we reject the null hypothesis.
- Residuals are not normally distributed.

Conclusion:

Landing Distance is highly dependent on speed of aircraft on ground, aircraft type and height from the sea level.

Final equation is:

Distance = -955.80 + 462.32*aircraft_type + 0.27*speed_ground^2 + 14.21 * height

- For a particular aircraft type, the landing distance would be 462.32 times greater than the other.
    - In this case ‘Boeing’ is 462.32 points greater than ‘Airbus’.
- For every unit increase in speed of flight on the ground level, there will be 0.27 unit increase in the landing distance.
- For every unit increase in height, there will be 14.21 unit increase in the landing distance.

Questions:

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

   After removing all the abnormal row values and after handling the missing values, I was left with 781 observations. I deleted almost 169 observations.

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

- For a particular aircraft type, the landing distance would be 462.32 times greater than the other.
  - In this case 'Boeing' is 462.32 points greater than 'Airbus'.
- For every unit increase in speed of flight on the ground level, there will be 0.27 unit increase in the landing distance.
- For every unit increase in height, there will be 14.21 unit increase in the landing distance.

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

   Yes, there is a significant difference between the two aircrafts, in our case study, for 'Boeing' the predicted landing distance is 462.32 times greater than 'Airbus"