PALASH SIDDAMSETTIWAR
M10676557

PROJECT
STAT COMPUTING- SEC 001

Project Objective: To reduce the risk of landing overrun by studying which factors and how they would impact the landing distance of a commercial flight.

The project involves multiple steps such as data collection, data cleaning, EDA (exploratory data analysis), modeling, model verification, finalization of model, building inferences and reporting. We will complete these steps in subsequent chapters. Since the data is already collected, we will start with data cleaning and exploration

Data: The data is present in two files- FAA1.xlsx and FAA2.xlsx.

Dataset	**FAA1**	**FAA2**
Observations	800	150
Variables	8 (aircraft, duration, no_pasg, speed_ground, speed_air, height, pitch, distance)	7 (aircraft, no_pasg, speed_ground, speed_air, height, pitch, distance)

CHAPTER 1: DATA PREPARATION- CLEANING AND PREPARATION

Chapter Objectives:

1. Importing the different datasets
2. Combination of different data sources
3. Performing completeness check of each variable – checking for missing values (NULLs) and cleaning the dataset
4. Performing validity check of each variable- checking for abnormal values (outliers) and cleaning the dataset
5. Summarizing the distribution of each variable through tables and figures

1. Importing the two datasets involve using the PROC IMPORT procedure using the “xls” DBMS option since both our datasets, FAA1 and FAA2 are .xls excel spreadsheets.

   SAS Code

```
/*Importing Dataset 1*/
PROC IMPORT OUT=FAA1 DATAFILE="/folders/myfolders/sasuser.v94/FAA1" DBMS="xls"
                    REPLACE;
          GETNAMES=YES;
RUN;
PROC PRINT;
RUN;

/*Importing Dataset 2*/
PROC IMPORT OUT=FAA2 DATAFILE="/folders/myfolders/sasuser.v94/FAA2" DBMS="xls"
                    REPLACE;
          GETNAMES=YES;
RUN;
PROC PRINT;
RUN;
```

Outputs:

Note: Data had 800 rows, data truncated to avoid superfluous text

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	Distance
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584
5	boeing	90.095381357	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976
6	boeing	137.59581722	55	75.014343744	.	41.21496259	4.203853398	1627.0681991
7	boeing	73.023794916	54	54.4298029	.	24.03532163	3.8376457299	805.30399317
8	boeing	52.903187872	57	57.101661737	.	19.388837508	4.6436717769	573.62178606
9	boeing	155.51861605	61	85.443624251	.	35.375389749	4.2287278648	1698.9927548
10	boeing	176.86203205	56	61.796710514	.	36.748816124	4.1843990127	1137.7457579
…	…	…	…	…	…	…	…	…
791	airbus	264.59337482	55	102.74650485	102.41247143	38.534020829	4.1697182481	2623.6512568
792	airbus	182.74887203	71	84.333339769	.	28.932963149	2.9914135514	1190.0228225
793	airbus	80.477362793	67	60.789195406	.	33.797453043	3.7235699727	563.10266864
794	airbus	194.99198538	68	91.15047186	.	15.378493757	3.5127744454	1445.1634341
795	airbus	149.36036239	67	98.02641239	99.421688766	40.993052838	4.7268210237	2440.381218
796	airbus	98.461455246	50	73.939616162	.	42.353383637	3.9571042618	1027.2134659
797	airbus	114.23485681	57	79.982703338	.	42.244751261	3.785954219	1162.404395
798	airbus	118.57607182	63	75.368171615	.	31.340776135	3.5580199527	960.25559642
799	Airbus	200.62136624	63	77.148459304	.	23.602422529	3.020177825	899.43055864
800	Airbus	124.14010259	59	66.464640399	.	48.067790297	4.1656597705	853.86453785

Note:

1. We can also see that the first few observations of both the datasets are the same.
2. FAA2 data has only 150 rows, but as can be observed here, the imported data has 200 rows with bottom 50 being rows with only missing values.

We will treat these after combining the two datasets.

Obs	aircraft	no_pasg	speed_ground	speed_air	height	pitch	distance
1	Boeing	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	Boeing	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	Boeing	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	Boeing	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584
5	boeing	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976
6	boeing	55	75.014343744	.	41.21496259	4.203853398	1627.0681991
7	boeing	54	54.4298029	.	24.03532163	3.8376457299	805.30399317
8	boeing	57	57.101661737	.	19.388837508	4.6436717769	573.62178606
9	boeing	61	85.443624251	.	35.375389749	4.2287278648	1698.9927548
10	boeing	56	61.796710514	.	36.748816124	4.1843990127	1137.7457579
…	…	…	…	…	…	…	…
150	airbus	70	100.9894669	102.46163794	19.703120709	4.3621150325	2348.6175889
151		.	.	.	.	.	.
152		.	.	.	.	.	.
153		.	.	.	.	.	.
…	…	…	…	…	…	…	…
196		.	.	.	.	.	.
197		.	.	.	.	.	.
198		.	.	.	.	.	.
199		.	.	.	.	.	.
200		.	.	.	.	.	.

2. Combining the two datasets involve taking union of data-source FAA1 and FAA2, one below the other, since the variables in both are the same (with the exception of duration which is not present in FAA2- which will result in missing values on combination)

SAS Code

```
/* Combining the two datasets */
DATA COMBINED;
        SET FAA1 FAA2;
RUN;
PROC PRINT;
RUN;
```

OUTPUT:

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	98.4790912	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
2	boeing	125.73329732	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
3	boeing	112.0170008	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
4	boeing	196.82569105	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584
5	boeing	90.095381357	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976
…	…	…	…	…	…	…	…	…
796	airbus	98.461455246	50	73.939616162	.	42.353383637	3.9571042618	1027.2134659
797	airbus	114.23485681	57	79.982703338	.	42.244751261	3.785954219	1162.404395
798	airbus	118.57607182	63	75.368171615	.	31.340776135	3.5580199527	960.25559642
799	airbus	200.62136624	63	77.148459304	.	23.602422529	3.020177825	899.43055864
800	airbus	124.14010259	59	66.464640399	.	48.067790297	4.1656597705	853.86453785
801	boeing	.	53	107.91568005	109.32837648	27.418924252	4.0435145715	3369.8363638
802	boeing	.	69	101.65558863	102.8514051	27.804716181	4.1174316991	2987.8039235
803	boeing	.	61	71.051960883	.	18.589385734	4.4340431286	1144.922426
804	boeing	.	56	85.813327679	.	30.744597235	3.8842361245	1664.2181584
805	boeing	.	70	59.888528183	.	32.397688062	4.0260964152	1050.2644976
…	…	…	…	…	…	…	…	…
996		.	.	.	.	.	.	.
997		.	.	.	.	.	.	.
998		.	.	.	.	.	.	.
999		.	.	.	.	.	.	.
1000		.	.	.	.	.	.	.

SAS Code

```
/* Remove duplicate rows */

PROC SORT DATA=COMBINED;
BY AIRCRAFT NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DISTANCE DESCENDING DURATION;
RUN;
PROC SORT DATA=COMBINED NODUPKEY;
BY NO_PASG SPEED_GROUND SPEED_AIR HEIGHT PITCH DISTANCE;
RUN;
PROC PRINT;
RUN;
```

OUTPUT:

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1		.	.	.	.	.	.	.
2	boeing	159.22116836	29	62.699670618	.	27.145647213	3.6351738156	1016.9505364
3	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162	250.68976141
4	airbus	188.01797726	38	85.180842251	.	37.028793691	4.1216901717	1257.0092519
5	airbus	93.540807771	40	80.627416679	.	28.60255713	3.6234201886	1021.0888117
6	airbus	123.30242152	41	97.568203986	96.978436701	38.409192953	3.5322719834	2167.7576915
7	boeing	163.96367943	42	78.282488619	.	15.730271196	4.0579920873	1328.9854598
8	airbus	109.19713407	43	82.483044979	.	30.140024889	4.0896284195	1321.0000654
9	boeing	155.00792518	43	92.130909422	93.675816291	35.186988882	3.6908539437	2370.007204
10	boeing	206.06572604	44	61.847975974	.	26.939627352	3.9372737398	896.58091126
…	…	…	…	…	…	…	…	…
842	airbus	206.55378907	77	76.457173777	.	31.918146708	3.8023203942	877.018871
843	boeing	172.56012205	77	82.29713755	.	44.758716354	4.2293090445	1809.27205
844	boeing	228.17710591	78	61.220375598	.	21.772286622	4.5955283685	970.04651856
845	boeing	107.11331938	78	86.807962025	.	25.477015381	4.4142187986	1910.8768699
846	boeing	128.93810992	79	106.93389135	108.42651323	30.457709156	4.8421492	3203.3188407
847	boeing	161.82569155	80	82.509055403	.	36.680194026	4.685310032	1590.3719225
848	airbus	132.46942492	80	100.01055305	100.891677	41.033010684	4.2975016214	2554.8330623
849	boeing	194.4671661	82	40.815188666	.	22.618444074	4.8765952309	761.4850777
850	airbus	96.765375204	82	90.744313306	.	33.024489327	3.545556835	1811.98402
851	airbus	185.4025176	87	92.12527293	.	45.093607704	3.4182713652	1826.8010013

We can see that all but one of the rows with missing data is removed and all the 100 observations from FAA2 with duplicate data are also removed.

3. The third step involves checking and treating for missing values.
    - Here, we will first remove all rows with only missing values.
    - Then we will check for the ratio of missing values to total rows in each column
    - For the missing values, we will replace them with mean of that column; unless the number of missing values are too high (>40%). In that case, we remove those variables from consideration
    - Observations with variables having lower than the threshold decided in step c will be removed from data since they can be considered as bad data

SAS CODE:

```
/* a. Removing the rows with all NULL values */
DATA COMBINED;
        SET COMBINED;
        IF COMPRESS(CATS(OF _ALL_), '.')=' ' THEN
                DELETE;
RUN;
PROC PRINT;
RUN;
```

OUTPUT:

Obs	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
1	boeing	159.22116836	29	62.699670618	.	27.145647213	3.6351738156	1016.9505364
2	airbus	172.04931209	36	47.486765029	.	13.984809941	4.2990197162	250.68976141
3	airbus	188.01797726	38	85.180842251	.	37.028793691	4.1216901717	1257.0092519
4	airbus	93.540807771	40	80.627416679	.	28.60255713	3.6234201886	1021.0888117
5	airbus	123.30242152	41	97.568203986	96.978436701	38.409192953	3.5322719834	2167.7576915
…	…	…	…	…	…	…	…	…
846	boeing	161.82569155	80	82.509055403	.	36.680194026	4.685310032	1590.3719225
847	airbus	132.46942492	80	100.01055305	100.891677	41.033010684	4.2975016214	2554.8330623
848	boeing	194.4671661	82	40.815188666	.	22.618444074	4.8765952309	761.4850777
849	airbus	96.765375204	82	90.744313306	.	33.024489327	3.545556835	1811.98402
850	airbus	185.4025176	87	92.12527293	.	45.093607704	3.4182713652	1826.8010013

SAS CODE:

```
/* b. Missing values to total observations in each column */
/*For the continuous variables */
PROC MEANS DATA=COMBINED N NMISS;
VAR _NUMERIC_;
RUN;
/* For the discrete variable- aircraft */
PROC FREQ DATA = COMBINED;
TABLES AIRCRAFT;
RUN;
```

We can see that the missing values are present only in two variables, speed_air and duration

OUTPUT:

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

aircraft				
aircraft	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
airbus	450	52.94	450	52.94
boeing	400	47.06	850	100.00

As we can see from the outputs of step b, for the speed_air variable, a large number of values are missing 642/ (642+208) = 75.52%.

So it will be best to **remove the speed_air variable from consideration** due to the high number of missing values present.

SAS CODE:

```
/* c. Removing columns with >40% missing values, i.e. speed_air */
DATA COMBINED;
SET COMBINED;
DROP SPEED_AIR;
RUN;
PROC PRINT;
RUN;
```

OUTPUT:

Obs	aircraft	duration	no_pasg	speed_ground	height	pitch	distance
1	boeing	159.22116836	29	62.699670618	27.145647213	3.6351738156	1016.9505364
2	airbus	172.04931209	36	47.486765029	13.984809941	4.2990197162	250.68976141
3	airbus	188.01797726	38	85.180842251	37.028793691	4.1216901717	1257.0092519
4	airbus	93.540807771	40	80.627416679	28.60255713	3.6234201886	1021.0888117
5	airbus	123.30242152	41	97.568203986	38.409192953	3.5322719834	2167.7576915
...	...	...	...	...	...	...	...
847	airbus	132.46942492	80	100.01055305	41.033010684	4.2975016214	2554.8330623
848	boeing	194.4671661	82	40.815188666	22.618444074	4.8765952309	761.4850777
849	airbus	96.765375204	82	90.744313306	33.024489327	3.545556835	1811.98402
850	airbus	185.4025176	87	92.12527293	45.093607704	3.4182713652	1826.8010013

As we can see from the outputs of step b, for the duration variable, a small but significant number of values are missing
150/(800+150) = 15.8% which is lower than our threshold of 40%

Here, we can **replace the missing values with a central tendency (mean in this case** since mean and median are very similar i.e., no significant effect of outliers on the distribution)

SAS CODE:

```
/* d. Treating missing observations in columns with <20% missing values */

/* To check if mean can be used to replace values */
PROC MEANS DATA=COMBINED MEAN MEDIAN;
RUN;
/*Replacing missing values in duration column with mean of the column */
DATA COMBINED;
SET COMBINED;
IF DURATION='.' THEN DURATION=154.0065385;
RUN;
PROC PRINT;
RUN;
```

OUTPUT:

Variable	Label	Mean	Median
duration	duration	154.0065385	153.9480975
no_pasg	no_pasg	60.1035294	60.0000000
speed_ground	speed_ground	79.4523229	79.6428041
height	height	30.1442223	30.0931324
pitch	pitch	4.0093577	4.0082875
distance	distance	1526.02	1258.09

Obs	aircraft	duration	no_pasg	speed_ground	height	pitch	distance
1	boeing	159.22116836	29	62.699670618	27.145647213	3.6351738156	1016.9505364
2	airbus	172.04931209	36	47.486765029	13.984809941	4.2990197162	250.68976141
3	airbus	188.01797726	38	85.180842251	37.028793691	4.1216901717	1257.0092519
4	airbus	93.540807771	40	80.627416679	28.60255713	3.6234201886	1021.0888117
5	airbus	123.30242152	41	97.568203986	38.409192953	3.5322719834	2167.7576915
…	…	…	…	…	…	…	…
847	airbus	132.46942492	80	100.01055305	41.033010684	4.2975016214	2554.8330623
848	boeing	194.4671661	82	40.815188666	22.618444074	4.8765952309	761.4850777
849	airbus	96.765375204	82	90.744313306	33.024489327	3.545556835	1811.98402
850	airbus	185.4025176	87	92.12527293	45.093607704	3.4182713652	1826.8010013

4. The fourth step involves looking for abnormal values (extreme values in each column).
   For variables which have stipulations mentioned, they are
   - DURATION>40 min
   - SPEED_GROUND>30mph and SPEED_GROUND<140mph
   - HEIGHT>6 m
   - DISTANCE<6000 feet

   We will remove the observations where this case is present

   SAS CODE:

```
/*Remove observations using the conditions mentioned */
DATA COMBINED;
SET COMBINED;
IF DURATION<40 THEN GOOD_DATA="NO";
IF SPEED_GROUND<30 THEN GOOD_DATA="NO";
IF SPEED_GROUND>140 THEN GOOD_DATA="NO";
IF HEIGHT<6 THEN GOOD_DATA="NO";
IF DISTANCE>6000 THEN GOOD_DATA="NO";
RUN;
PROC PRINT;
RUN;

DATA COMBINED;
SET COMBINED;
IF GOOD_DATA="NO" THEN DELETE;
DROP GOOD_DATA;
RUN;
PROC PRINT;
RUN;
```

   OUTPUT:

Obs	aircraft	duration	no_pasg	speed_ground	height	pitch	distance
1	boeing	159.22116836	29	62.699670618	27.145647213	3.6351738156	1016.9505364
2	airbus	172.04931209	36	47.486765029	13.984809941	4.2990197162	250.68976141
3	airbus	188.01797726	38	85.180842251	37.028793691	4.1216901717	1257.0092519
4	airbus	93.540807771	40	80.627416679	28.60255713	3.6234201886	1021.0888117
5	airbus	123.30242152	41	97.568203986	38.409192953	3.5322719834	2167.7576915
…	…	…	…	…	…	…	…
828	airbus	132.46942492	80	100.01055305	41.033010684	4.2975016214	2554.8330623
829	boeing	194.4671661	82	40.815188666	22.618444074	4.8765952309	761.4850777
830	airbus	96.765375204	82	90.744313306	33.024489327	3.545556835	1811.98402
831	airbus	185.4025176	87	92.12527293	45.093607704	3.4182713652	1826.8010013

Hence, 19 observations were removed since they had at least one variable with an abnormal value.

5. Now that all the variables are treated for missing values and all the abnormal data is removed, we will look at how each variable is distributed using the PROC MEANS and PROC UNIVARIATE procedures to look at the distributions statistics such as mean, range, maximum and minimum values along with the normalcy of the distribution, the kurtosis and the skewness of the distribution

SAS CODE:

```
PROC FREQ DATA=COMBINED;
VAR AIRCRAFT;
RUN;

PROC UNIVARIATE DATA=COMBINED NORMAL PLOT;
RUN;
```

OUTPUT:

As we can see from these tables and charts,

- Aircraft discrete variable has 2 levels with almost equal number of observations in each (444 and 387)
- Duration variable has high range and variability but is not skewed
- Number of passengers, ground speed and pitch variables have low deviation, range and skewness
- Pitch variable has low deviation and skewness but high range
- Distance variable has high standard deviation, kurtosis and is highly skewed
- Other than Duration, all other variables follow normal distribution very well

aircraft				
aircraft	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
airbus	444	53.43	444	53.43
boeing	387	46.57	831	100.00

The UNIVARIATE Procedure
Variable: duration
(duration)

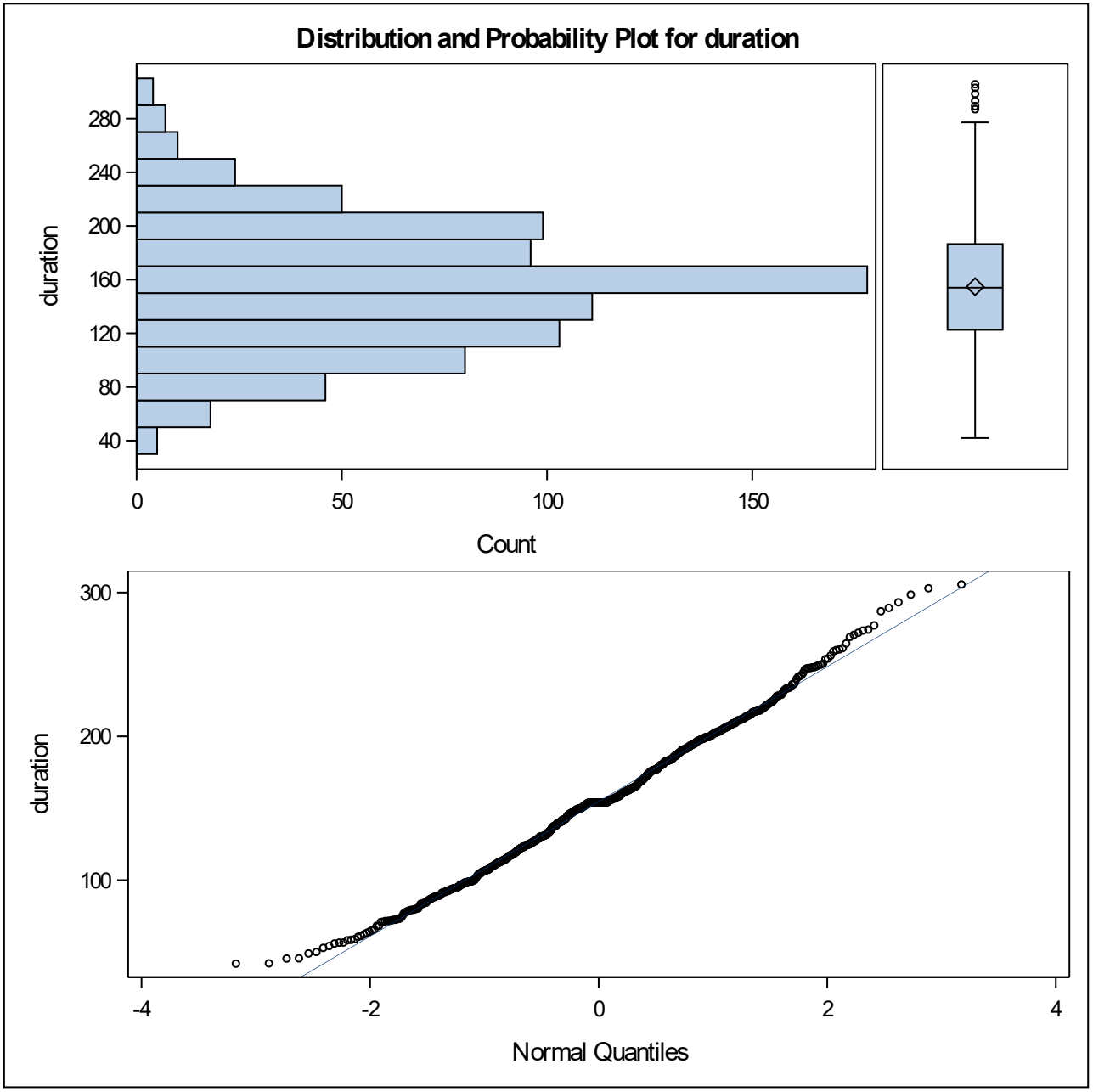
Moments			
N	831	**Sum Weights**	831
Mean	154.729439	**Sum Observations**	128580.164
Std Deviation	46.8713388	**Variance**	2196.9224
Skewness	0.19879091	**Kurtosis**	-0.0153812
Uncorrected SS	21718582.1	**Corrected SS**	1823445.6
Coeff Variation	30.2924506	**Std Error Mean**	1.62594873

Basic Statistical Measures			
Location		**Variability**	
Mean	154.7294	**Std Deviation**	46.87134
Median	154.0065	**Variance**	2197
Mode	154.0065	**Range**	263.67234
		Interquartile Range	63.90871

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	95.16256	**Pr > |t|**	<.0001
Sign	**M**	415.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	172848	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.994235	**Pr < W**	0.0029
Kolmogorov-Smirnov	**D**	0.040206	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.212362	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	1.086294	**Pr > A-Sq**	0.0079

The UNIVARIATE Procedure



The UNIVARIATE Procedure
Variable: no_pasg
(no_pasg)

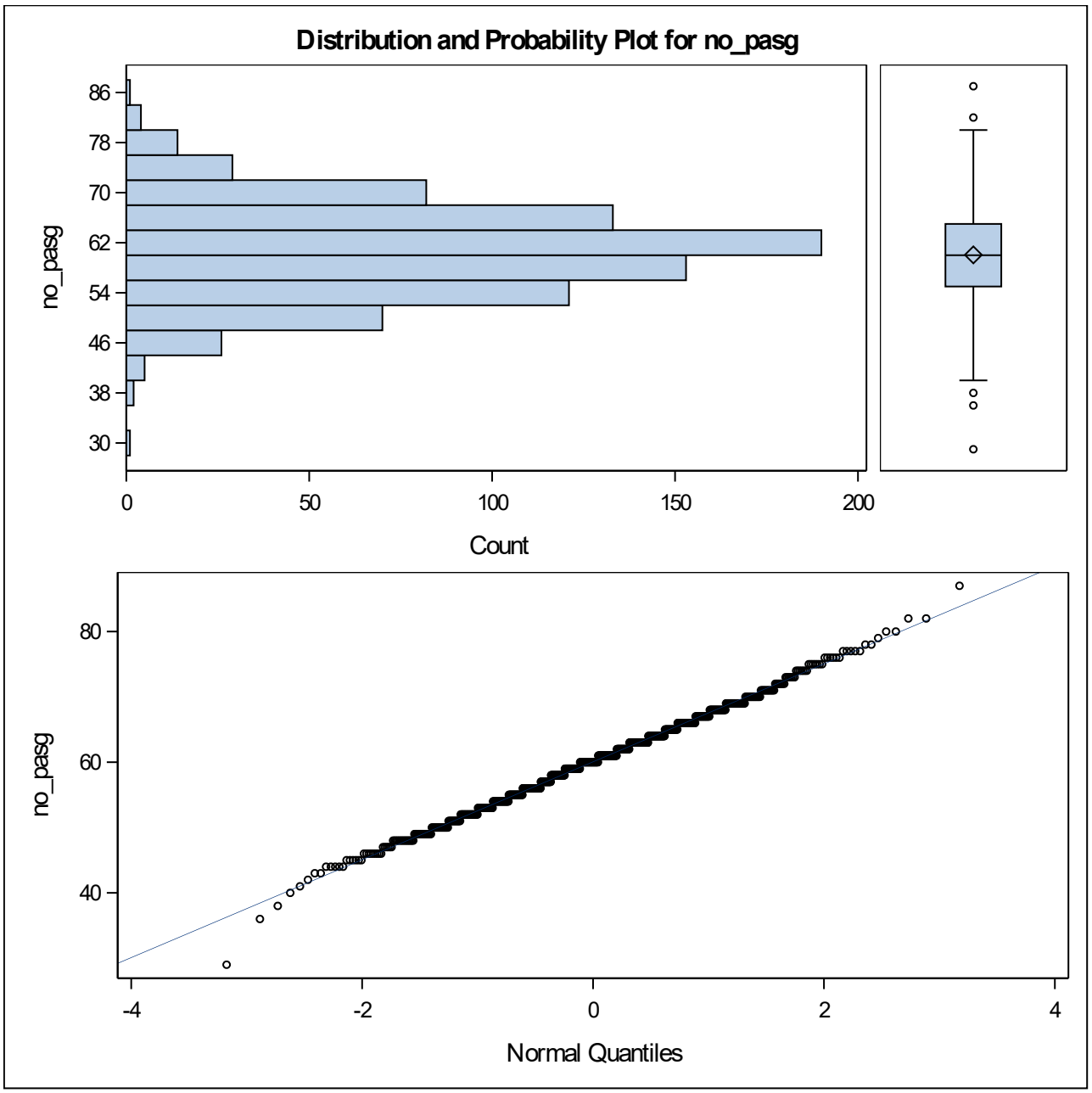
Moments			
N	831	**Sum Weights**	831
Mean	60.055355	**Sum Observations**	49906
Std Deviation	7.49131655	**Variance**	56.1198237
Skewness	-0.0135746	**Kurtosis**	0.30027454
Uncorrected SS	3043702	**Corrected SS**	46579.4537
Coeff Variation	12.4740193	**Std Error Mean**	0.25987089

Basic Statistical Measures			
Location		**Variability**	
Mean	60.05535	**Std Deviation**	7.49132
Median	60.00000	**Variance**	56.11982
Mode	61.00000	**Range**	58.00000
		Interquartile Range	10.00000

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	231.0969	**Pr > |t|**	<.0001
Sign	**M**	415.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	172848	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.996773	**Pr < W**	0.0914
Kolmogorov-Smirnov	**D**	0.042179	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.156592	**Pr > W-Sq**	0.0207
Anderson-Darling	**A-Sq**	0.847795	**Pr > A-Sq**	0.0300

The UNIVARIATE Procedure



The UNIVARIATE Procedure
Variable: speed_ground
(speed_ground)

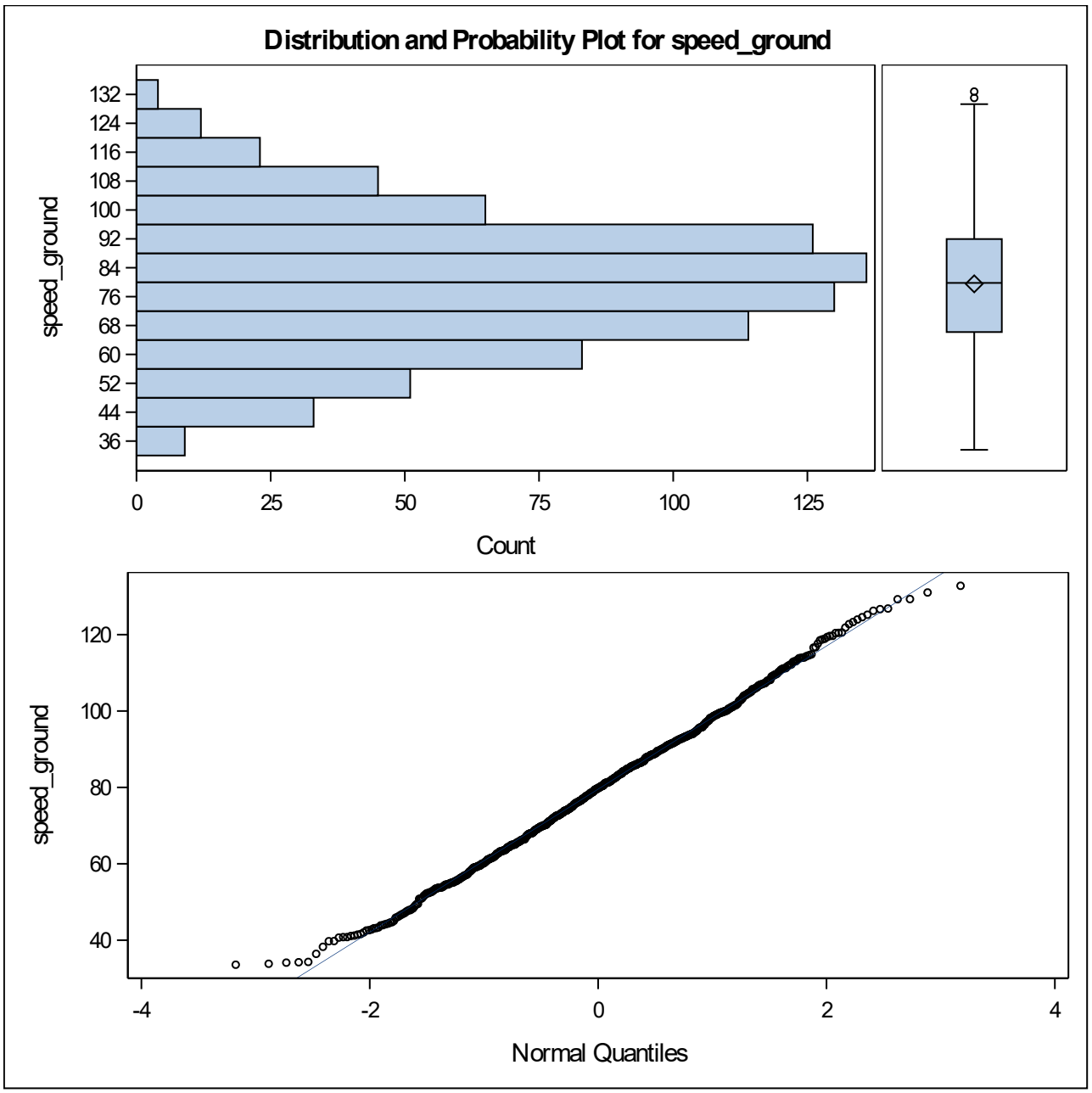
Moments			
N	831	**Sum Weights**	831
Mean	79.5426997	**Sum Observations**	66099.9835
Std Deviation	18.7356754	**Variance**	351.025533
Skewness	0.08890294	**Kurtosis**	-0.2324866
Uncorrected SS	5549122.33	**Corrected SS**	291351.193
Coeff Variation	23.5542363	**Std Error Mean**	0.64993338

Basic Statistical Measures			
Location		**Variability**	
Mean	79.54270	**Std Deviation**	18.73568
Median	79.79396	**Variance**	351.02553
Mode	.	**Range**	99.21057
		Interquartile Range	25.75708

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	122.3859	**Pr > |t|**	<.0001
Sign	**M**	415.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	172848	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.996876	**Pr < W**	0.1051
Kolmogorov-Smirnov	**D**	0.018552	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.034826	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.285505	**Pr > A-Sq**	>0.2500

The UNIVARIATE Procedure



The UNIVARIATE Procedure
Variable: height
(height)

Moments			
N	831	**Sum Weights**	831
Mean	30.4578695	**Sum Observations**	25310.4896
Std Deviation	9.78481143	**Variance**	95.7425347
Skewness	0.12714447	**Kurtosis**	-0.3338733
Uncorrected SS	850369.892	**Corrected SS**	79466.3038
Coeff Variation	32.1257251	**Std Error Mean**	0.33943135

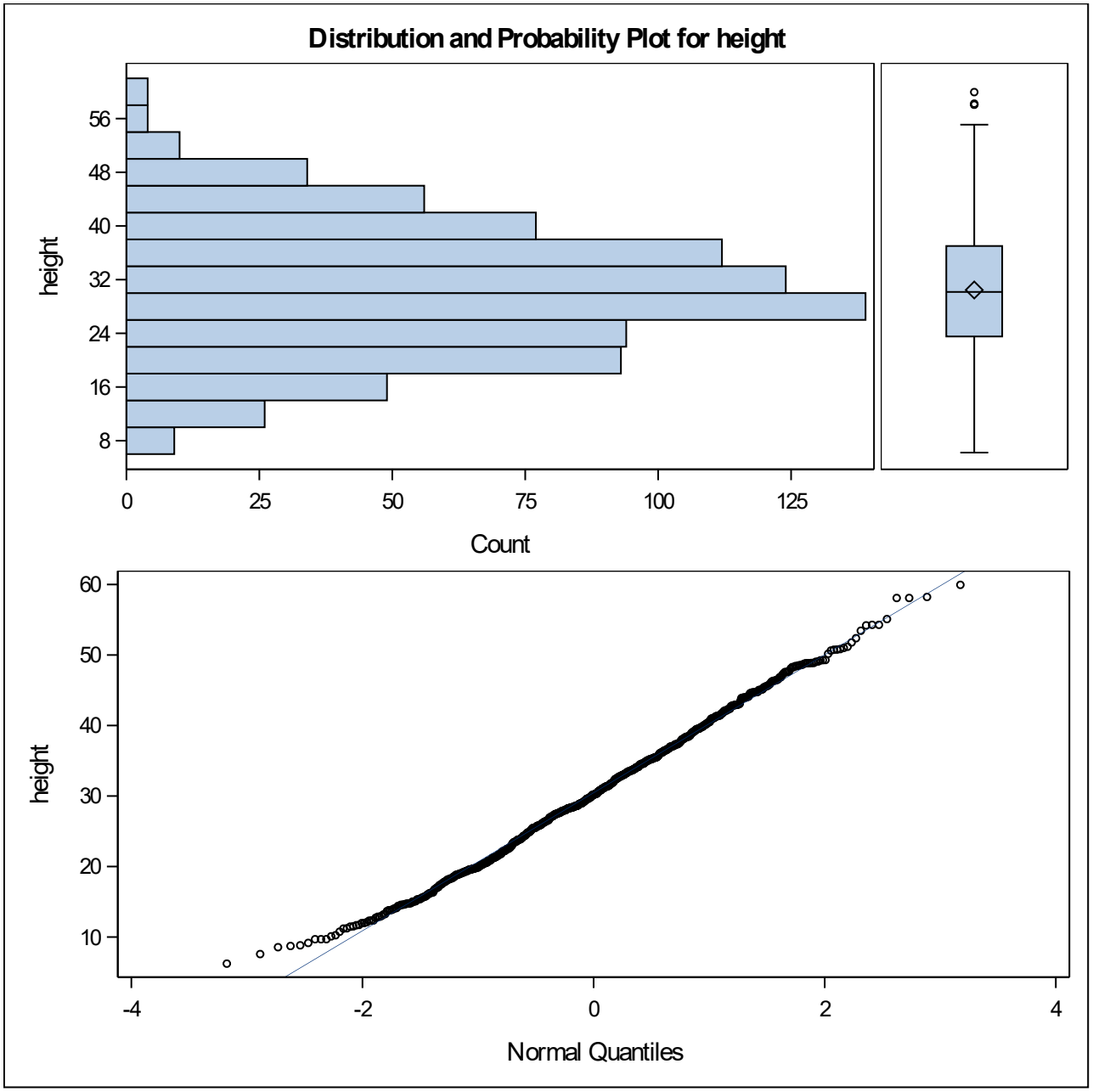
Basic Statistical Measures			
Location		**Variability**	
Mean	30.45787	**Std Deviation**	9.78481
Median	30.16708	**Variance**	95.74253
Mode	9.68831	**Range**	53.71845
		Interquartile Range	13.48443

Note: The mode displayed is the smallest of 49 modes with a count of 2.

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	89.73205	**Pr > |t|**	<.0001
Sign	**M**	415.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	172848	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.995678	**Pr < W**	0.0202
Kolmogorov-Smirnov	**D**	0.024062	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.068733	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.561139	**Pr > A-Sq**	0.1495

The UNIVARIATE Procedure



The UNIVARIATE Procedure
Variable: pitch
(pitch)

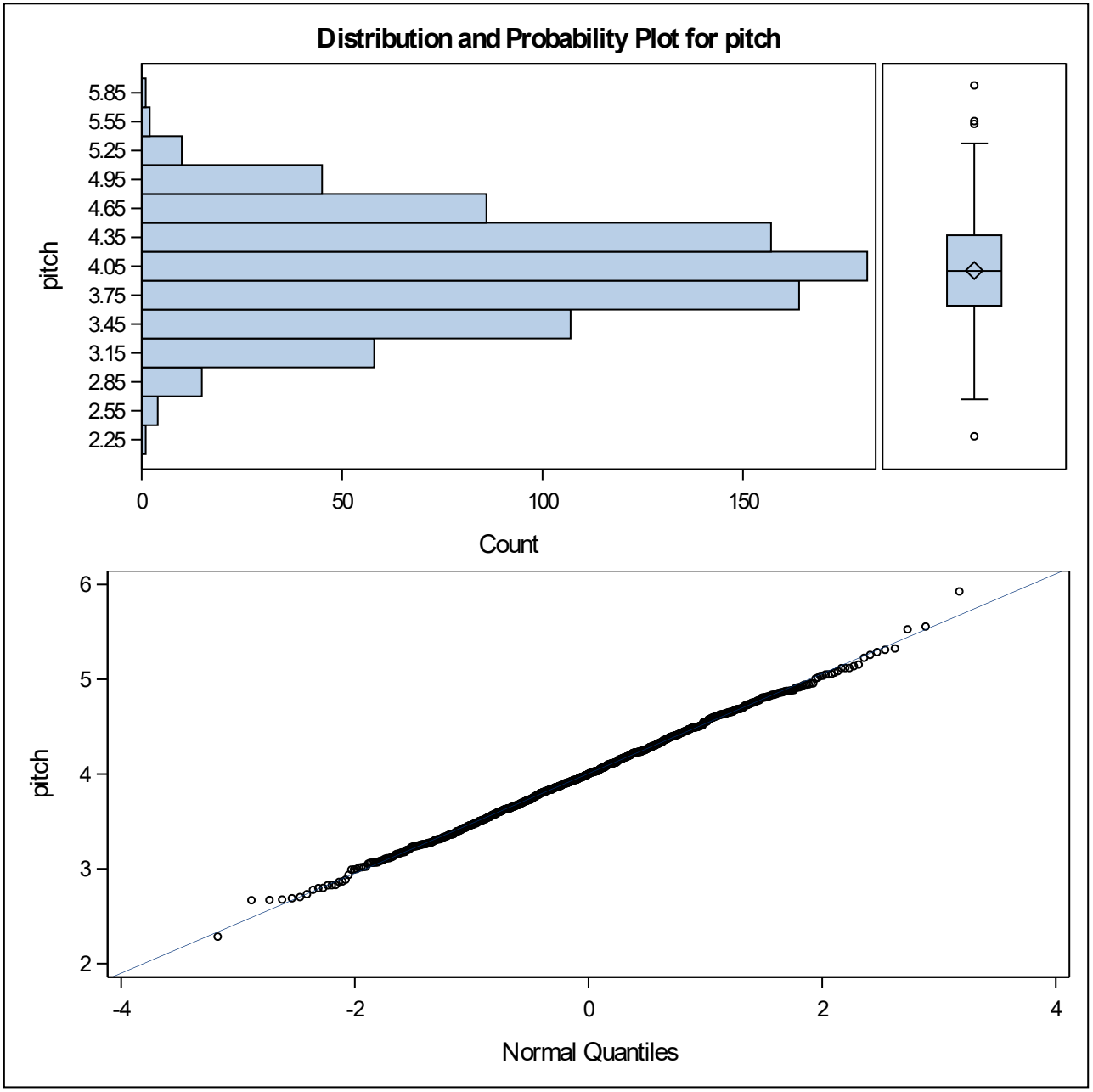
Moments			
N	831	**Sum Weights**	831
Mean	4.00516086	**Sum Observations**	3328.28868
Std Deviation	0.52656905	**Variance**	0.27727496
Skewness	0.01730511	**Kurtosis**	-0.0907921
Uncorrected SS	13560.4698	**Corrected SS**	230.138218
Coeff Variation	13.1472634	**Std Error Mean**	0.01826648

Basic Statistical Measures			
Location		**Variability**	
Mean	4.005161	**Std Deviation**	0.52657
Median	4.001038	**Variance**	0.27727
Mode	.	**Range**	3.64230
		Interquartile Range	0.73067

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	219.2629	**Pr > |t|**	<.0001
Sign	**M**	415.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	172848	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.998917	**Pr < W**	0.9117
Kolmogorov-Smirnov	**D**	0.014278	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.021245	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.183351	**Pr > A-Sq**	>0.2500

The UNIVARIATE Procedure



The UNIVARIATE Procedure
Variable: distance
(distance)

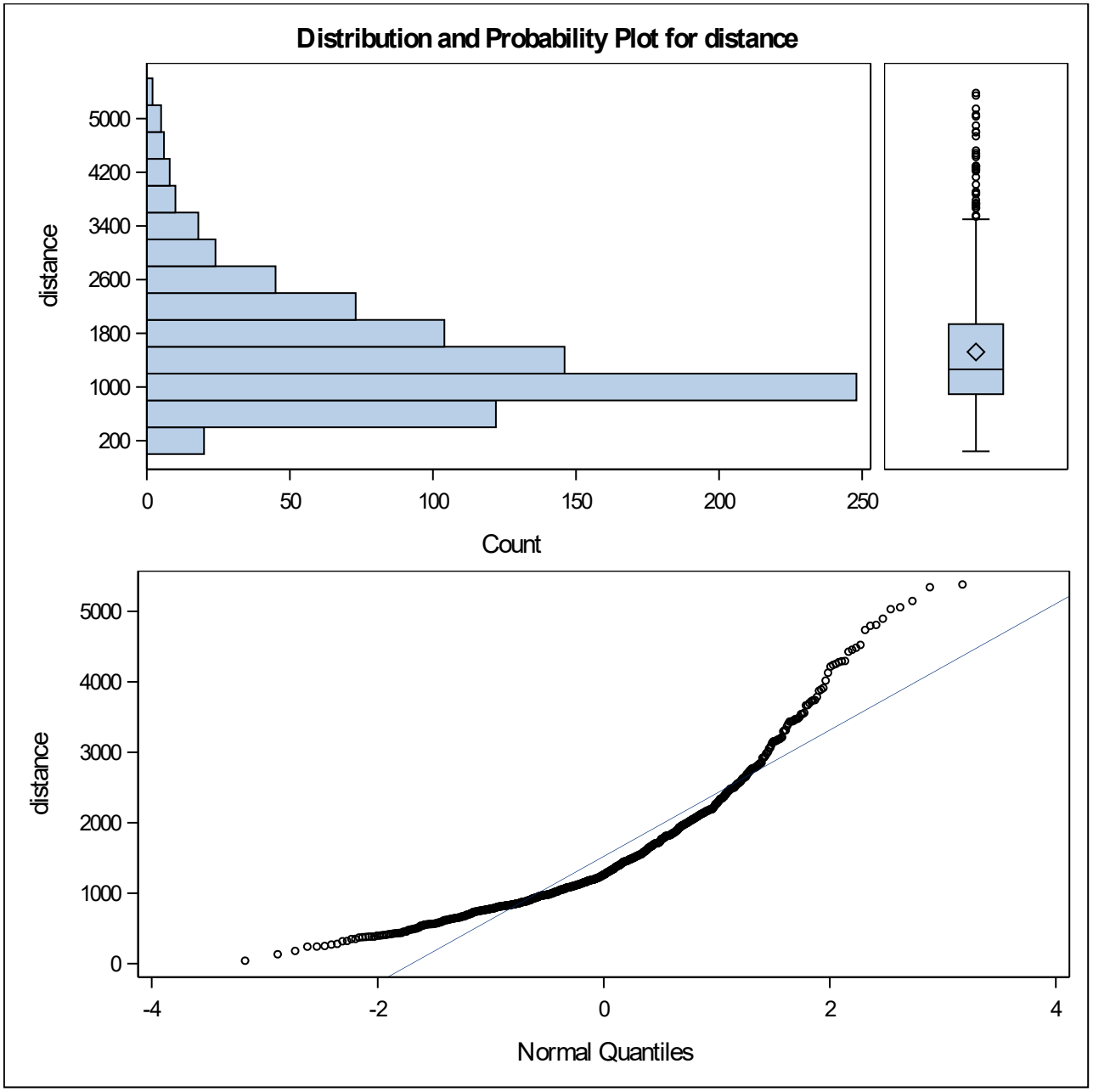
Moments			
N	831	**Sum Weights**	831
Mean	1522.48287	**Sum Observations**	1265183.27
Std Deviation	896.338152	**Variance**	803422.083
Skewness	1.47639585	**Kurtosis**	2.54813164
Uncorrected SS	2593060185	**Corrected SS**	666840329
Coeff Variation	58.8734473	**Std Error Mean**	31.093626

Basic Statistical Measures			
Location		**Variability**	
Mean	1522.483	**Std Deviation**	896.33815
Median	1262.154	**Variance**	803422
Mode	.	**Range**	5340
		Interquartile Range	1044

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	48.96447	**Pr > |t|**	<.0001
Sign	**M**	415.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	172848	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.882121	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.117065	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	4.376585	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	26.08147	**Pr > A-Sq**	<0.0050

The UNIVARIATE Procedure



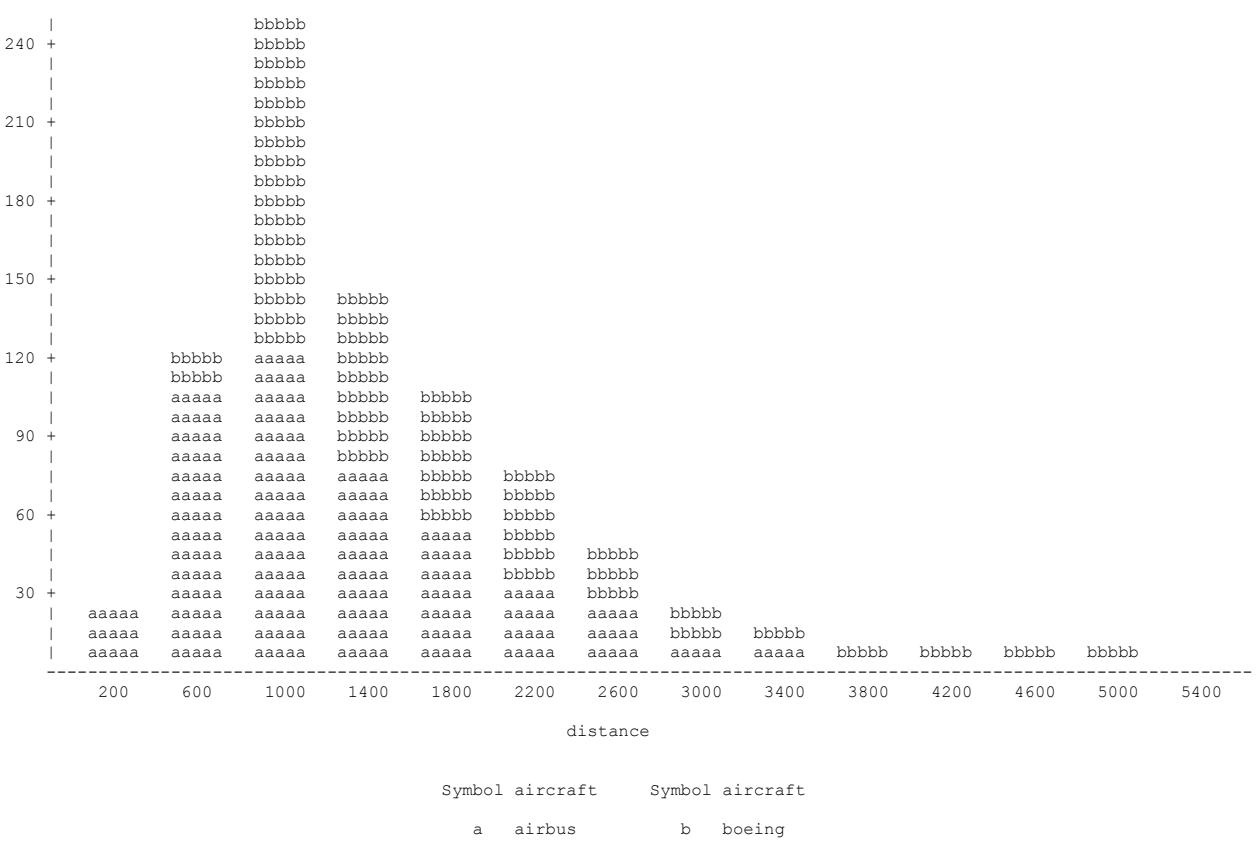
CHAPTER 2: DATA EXPLORATION: VISUALIZATION AND CORRELATION ANALYSIS

Chapter Objectives:

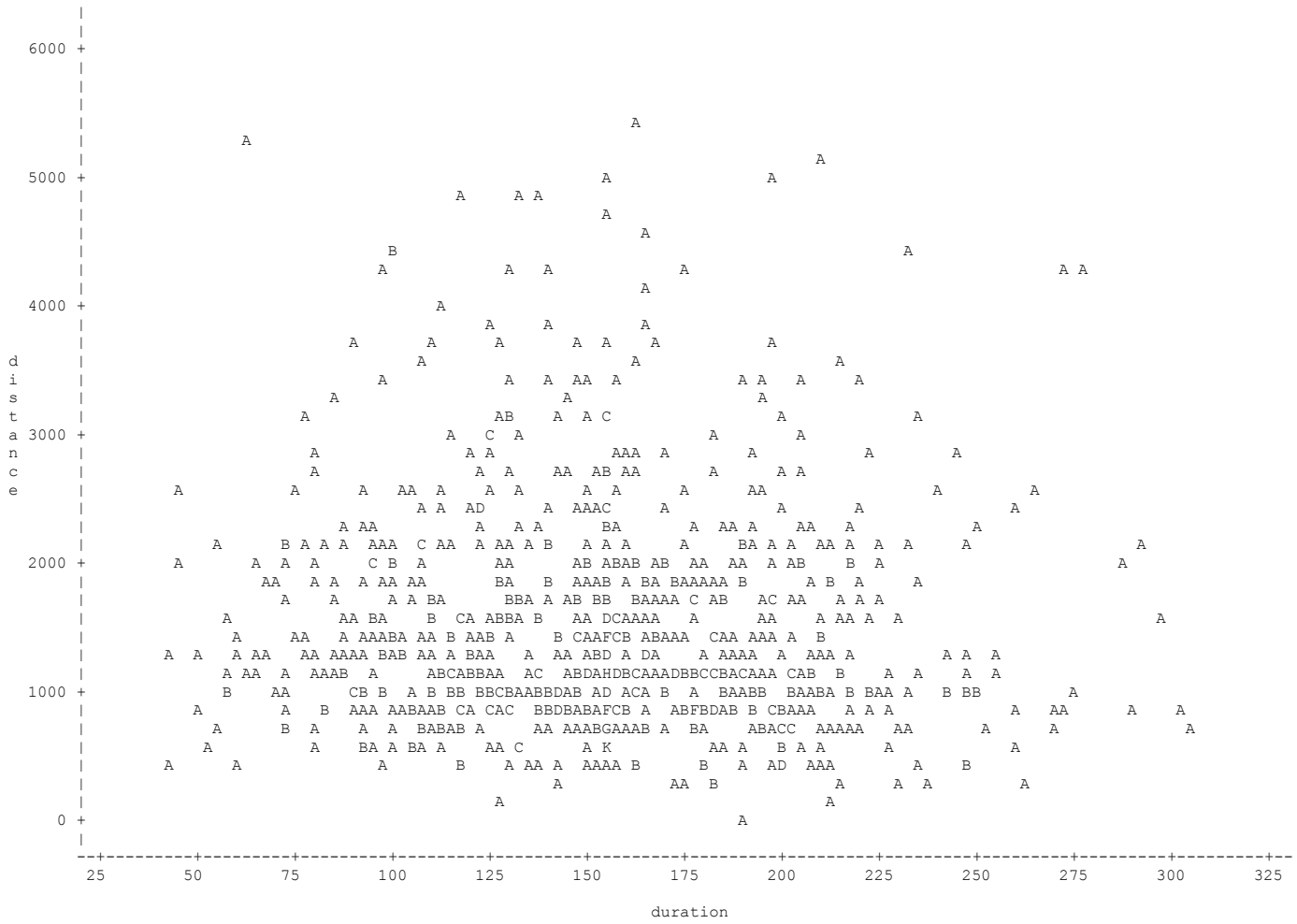
1. To plot X-Y plots for and X-Y plots between distance and each other variable
2. To create correlation values for distance and each other variable

1.

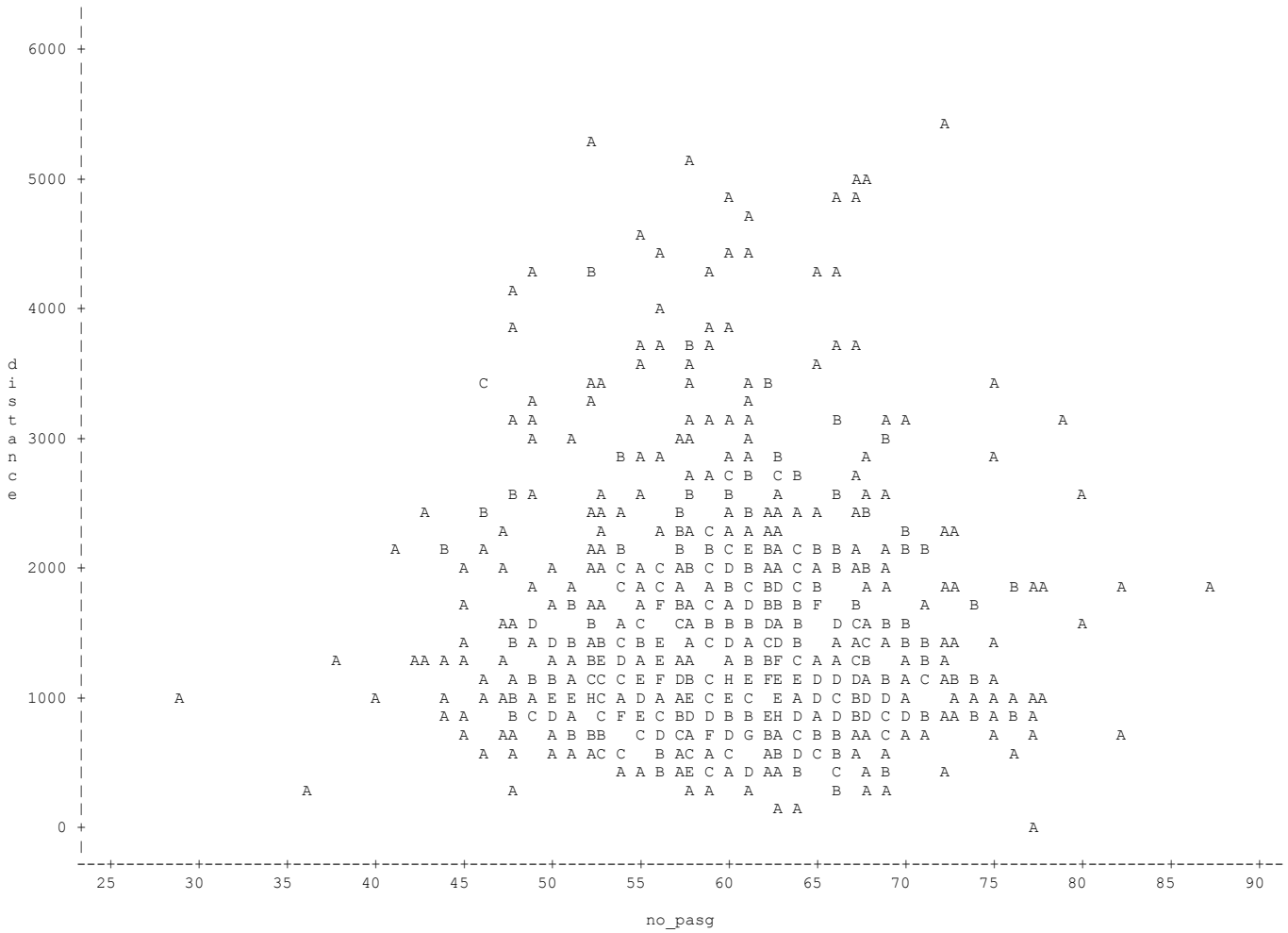
```
/* Create X-Y plots for each variable with distance*/
PROC CHART DATA=COMBINED;
VBAR DISTANCE / SUBGROUP=AIRCRAFT;
RUN;

PROC PLOT DATA=COMBINED;
PLOT DISTANCE*DURATION;
RUN;

PROC PLOT DATA=COMBINED;
PLOT DISTANCE*NO_PASG;
RUN;

PROC PLOT DATA=COMBINED;
PLOT DISTANCE*SPEED_GROUND;
RUN;

PROC PLOT DATA=COMBINED;
PLOT DISTANCE*HEIGHT;
RUN;

PROC PLOT DATA=COMBINED;
PLOT DISTANCE*PITCH;
RUN;
```


Frequency
240
210
180
150
120
90
60
30
200
600
1000
1400
1800
2200
2600
3000
3400
3800
4200
4600
5000
5400
distance
Symbol aircraft
a airbus
Symbol aircraft
b boeing

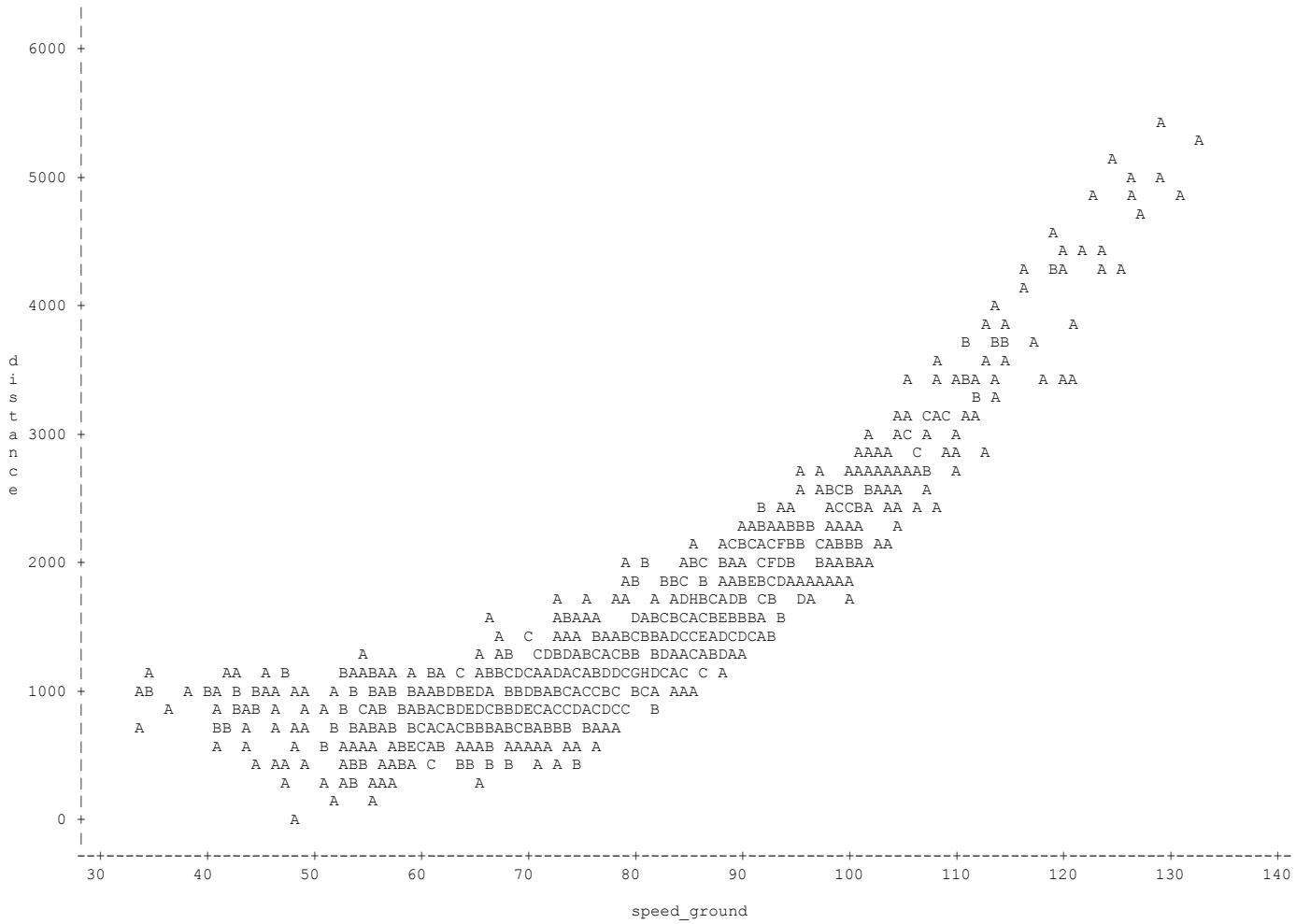
Plot of distance*duration. Legend: A = 1 obs, B = 2 obs, etc.



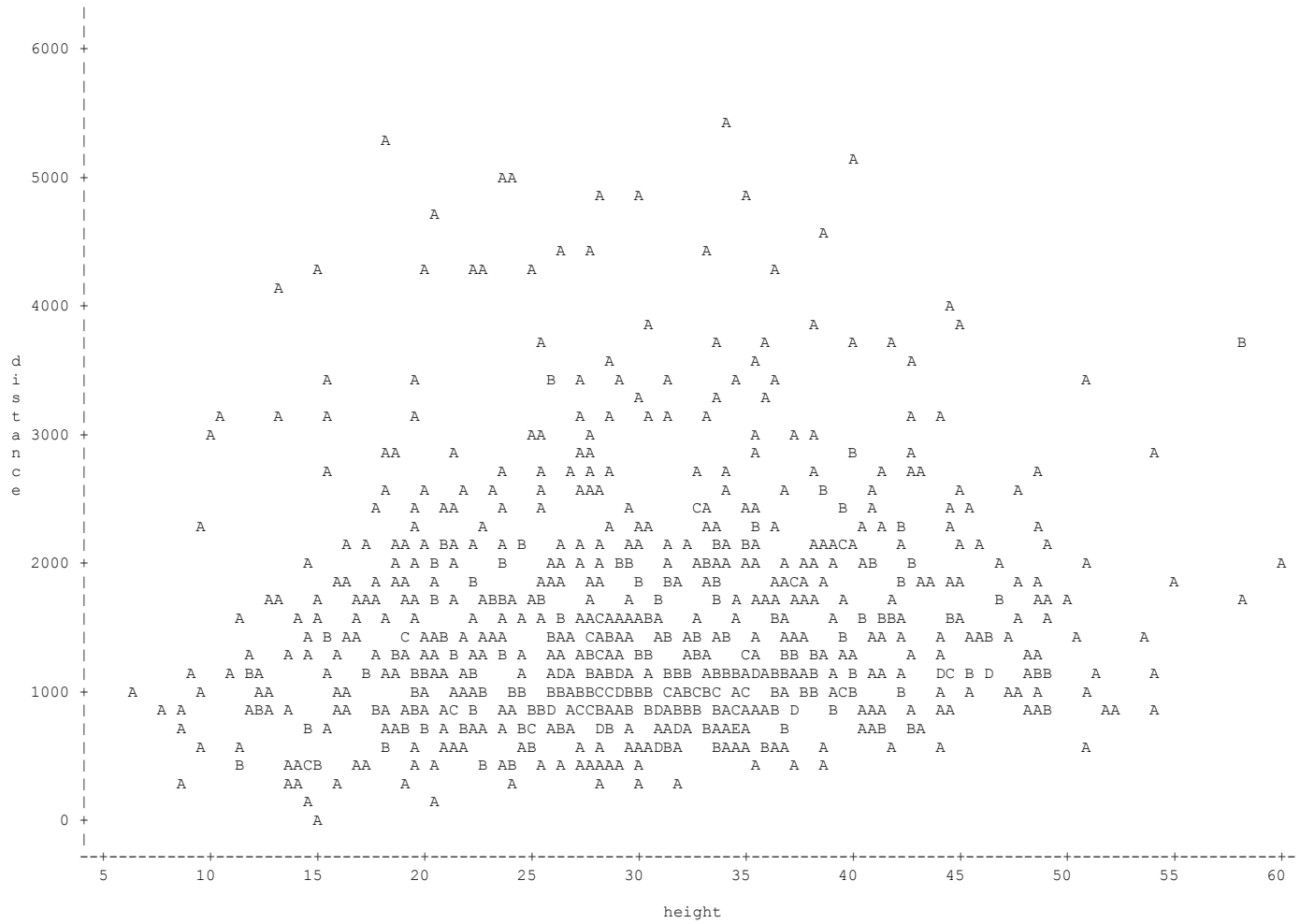
Plot of distance*no pasg. Legend: A = 1 obs, B = 2 obs, etc.



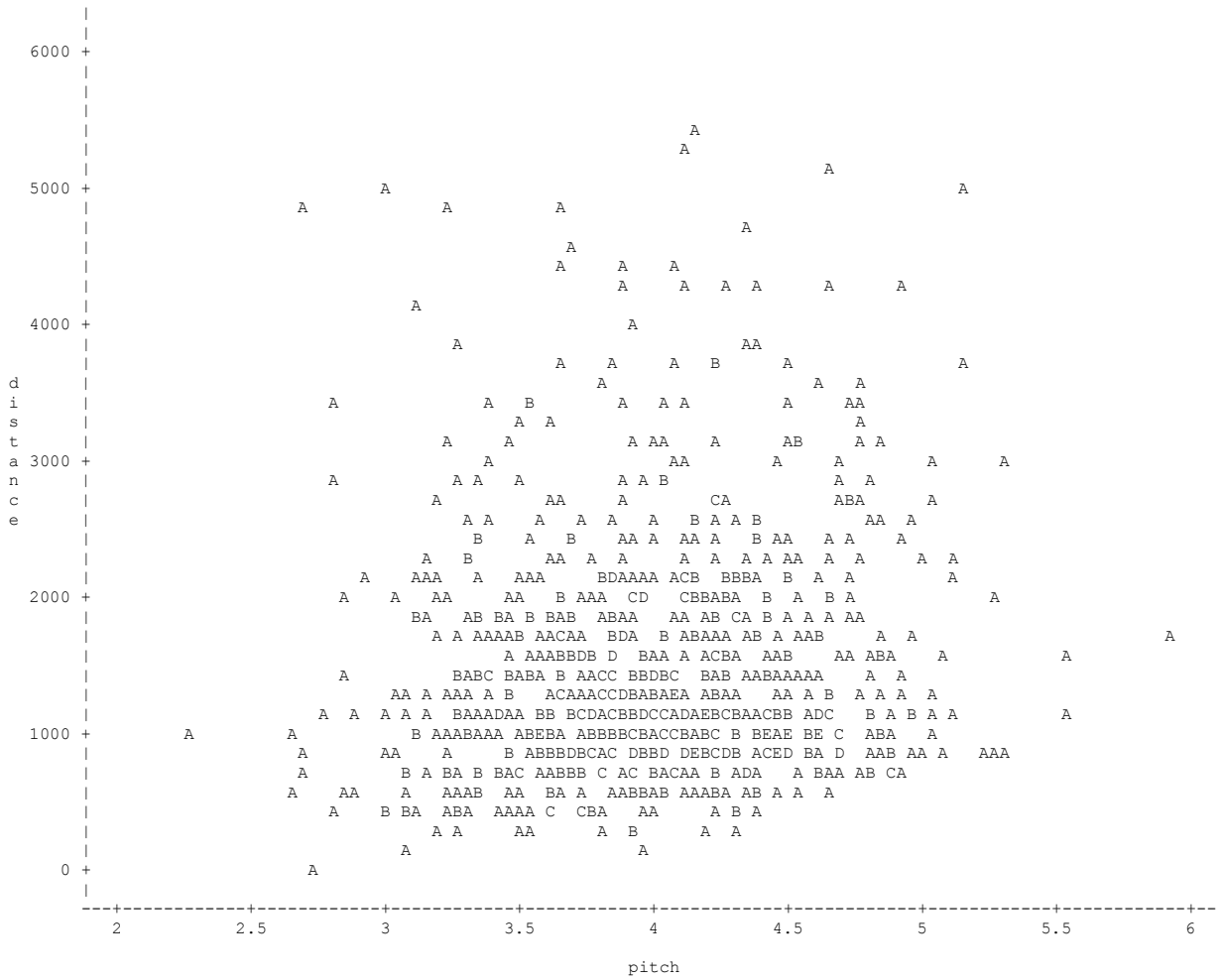
Plot of distance*speed ground. Legend: A = 1 obs, B = 2 obs, etc.



Plot of distance*height. Legend: A = 1 obs, B = 2 obs, etc.



Plot of distance*pitch. Legend: A = 1 obs, B = 2 obs, etc.



We can see from these plots that for aircraft type, there seems to be some pattern for the distribution of distance across the two types, with Boeing flights recording longer distances than Airbus flights.

For speed_ground, there seems to be a direct correlation with distance as the plot is along the x=y line. For all the other variables though, there seems to be little or no pattern with the distance variable.

2.

```
PROC CORR DATA=COMBINED;
VAR DURATION NO_PASG SPEED_GROUND HEIGHT PITCH;
WITH DISTANCE;
RUN;

PROC CORR DATA=COMBINED;
VAR DURATION NO_PASG SPEED_GROUND HEIGHT PITCH;
RUN;
```

1 With Variables:	distance
5 Variables:	duration no_pasg speed_ground height pitch

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	831	1522	896.33815	1265183	41.72231	5382	distance
duration	831	154.72944	46.87134	128580	41.94937	305.62171	duration
no_pasg	831	60.05535	7.49132	49906	29.00000	87.00000	no_pasg
speed_ground	831	79.54270	18.73568	66100	33.57410	132.78468	speed_ground
height	831	30.45787	9.78481	25310	6.22752	59.94596	height
pitch	831	4.00516	0.52657	3328	2.28448	5.92678	pitch

Pearson Correlation Coefficients, N = 831
Prob > |r| under H0: Rho=0

	duration	no_pasg	speed_ground	height	pitch
distance	-0.04995	-0.01776	0.86624	0.09941	0.08703
distance	0.1503	0.6093	<.0001	0.0041	0.0121

From this table, we can see the correlation between our dependent variable, distance, and our independent variables.

We can see that there is a high correlation between speed_ground and distance with a very low p-value too. For all the other variables, the correlation is not that stark as we can see that the values are pretty close to 0. This was evident though the X-Y plots too.

5 Variables:	duration	no_pasg	speed_ground	height	pitch

Simple Statistics							
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**	**Label**
duration	831	154.72944	46.87134	128580	41.94937	305.62171	duration
no_pasg	831	60.05535	7.49132	49906	29.00000	87.00000	no_pasg
speed_ground	831	79.54270	18.73568	66100	33.57410	132.78468	speed_ground
height	831	30.45787	9.78481	25310	6.22752	59.94596	height
pitch	831	4.00516	0.52657	3328	2.28448	5.92678	pitch

Pearson Correlation Coefficients, N = 831 Prob > |r| under H0: Rho=0					
	duration	**no_pasg**	**speed_ground**	**height**	**pitch**
duration duration	1.00000	-0.03539 0.3083	-0.04780 0.1686	0.01072 0.7575	-0.04470 0.1980
no_pasg no_pasg	-0.03539 0.3083	1.00000	-0.00013 0.9969	0.04699 0.1760	-0.01793 0.6057
speed_ground speed_ground	-0.04780 0.1686	-0.00013 0.9969	1.00000	-0.05761 0.0970	-0.03912 0.2599
height height	0.01072 0.7575	0.04699 0.1760	-0.05761 0.0970	1.00000	0.02298 0.5082
pitch pitch	-0.04470 0.1980	-0.01793 0.6057	-0.03912 0.2599	0.02298 0.5082	1.00000

From this table, we can study the correlations between the various independent variables.
We can see that there are no significantly large correlations between the variables, thus making the variables pretty independent from each other.

Hence, it makes sense to keep all the variables in consideration. The X-Y plots and correlation studies makes it clear that our current set of variables are robust enough for the linear model we plan to build. Speed_ground seems like an important variable due to its high correlation with the distance variable.

CHAPTER 3: MODELING

Objective: To build a robust data using the data we have at hand to establish the relationship between distance and other variables.

As we established in the previous chapter, we have 831 observations and 7 variables (including distance to build this model.

We will be building a linear model using the data we have at hand using the PROC REG procedure.

```
/*For AIRCRAFT variable, we need to convert it into a binary variable called BOEING with value 1 if it is BOING or 0 if it is AIRBUS*/
DATA MODEL_DATA;
SET COMBINED;
IF AIRCRAFT="boeing" THEN boeing=1;
ELSE boeing=0;
DROP AIRCRAFT;
RUN;

PROC PRINT;
RUN;
```

Obs	duration	no_pasg	speed_ground	height	pitch	distance	boeing
1	159.22116836	29	62.699670618	27.145647213	3.6351738156	1016.9505364	1
2	172.04931209	36	47.486765029	13.984809941	4.2990197162	250.68976141	0
3	188.01797726	38	85.180842251	37.028793691	4.1216901717	1257.0092519	0
4	93.540807771	40	80.627416679	28.60255713	3.6234201886	1021.0888117	0
5	123.30242152	41	97.568203986	38.409192953	3.5322719834	2167.7576915	0
…	…	…	…	…	…	…	…
827	161.82569155	80	82.509055403	36.680194026	4.685310032	1590.3719225	1
828	132.46942492	80	100.01055305	41.033010684	4.2975016214	2554.8330623	0
829	194.4671661	82	40.815188666	22.618444074	4.8765952309	761.4850777	1
830	96.765375204	82	90.744313306	33.024489327	3.545556835	1811.98402	0
831	185.4025176	87	92.12527293	45.093607704	3.4182713652	1826.8010013	0

```
PROC REG DATA=MODEL_DATA;
MODEL DISTANCE= BOEING DURATION NO_PASG SPEED_GROUND HEIGHT PITCH;
OUTPUT OUT=DIAGNOSTICS R=RESIDUAL;
TITLE Regression Analysis of the Cleaned Dataset;
RUN;
```

Regression Analysis of the Cleaned Dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	831
Number of Observations Used	831

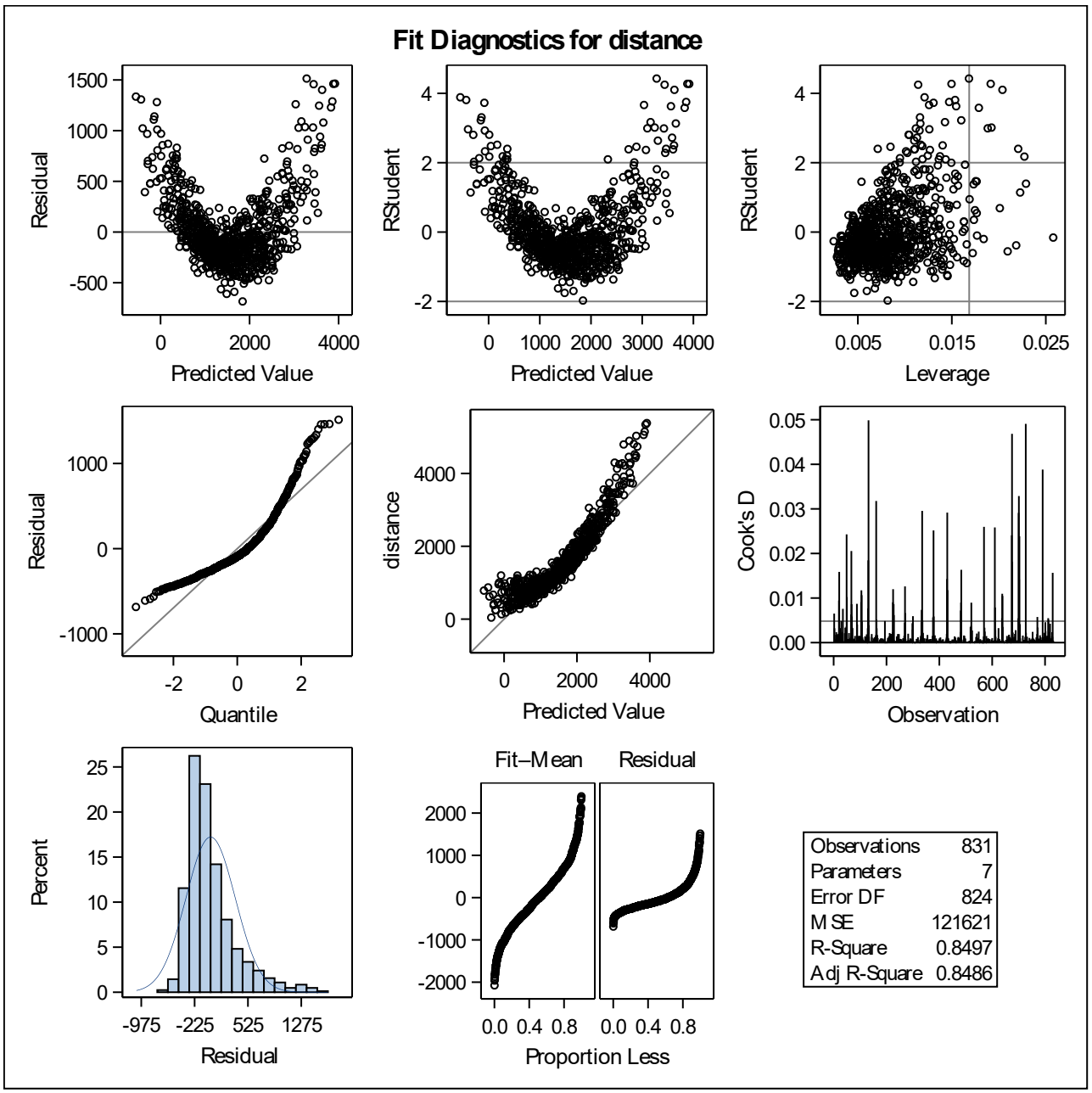
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	566625033	94437505	776.49	<.0001
Error	824	100215296	121621		
Corrected Total	830	666840329			

Root MSE	348.74132	**R-Square**	0.8497
Dependent Mean	1522.48287	**Adj R-Sq**	0.8486
Coeff Var	22.90609		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	-2541.73377	159.31120	-15.95	<.0001
boeing		1	480.83562	25.96817	18.52	<.0001
duration	duration	1	0.04766	0.25914	0.18	0.8541
no_pasg	no_pasg	1	-2.19282	1.61929	-1.35	0.1760
speed_ground	speed_ground	1	42.43550	0.64873	65.41	<.0001
height	height	1	14.16807	1.24129	11.41	<.0001
pitch	pitch	1	39.35996	24.61663	1.60	0.1102

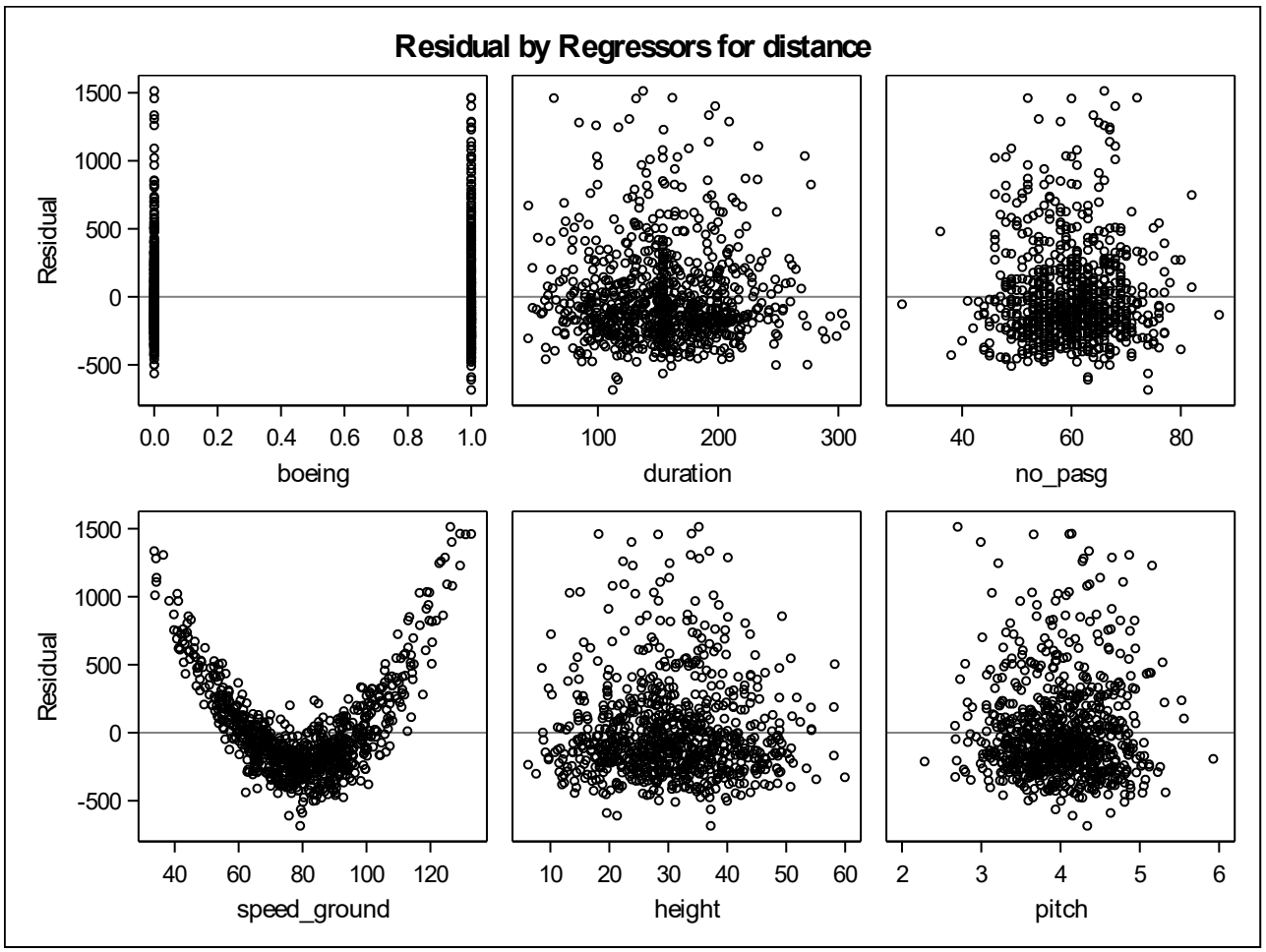
Regression Analysis of the Cleaned Dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance



Regression Analysis of the Cleaned Dataset

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance



CHAPTER IV: MODEL ANALYSIS AND VERIFICATION

Objectives:

1. To look at various model parameters to judge accuracy of model
2. Try to build a better model by removing the ones with high p-values
3. Check for the distribution of the residuals
4. Answer questions and make conclusions.

1. To check for the validity of the model, we can look at the Adjusted R-square and the p-values of the parameter estimates. We can see that the Adjusted R-square is pretty good (>.80), which means that our model is robust.

   However, we can see that the significance of our parameter estimates for variables- duration, no_pasg, pitch are very low.

2. Hence, let us build a model without these 3 variables and find how good our model is

```
PROC REG DATA=MODEL_DATA;
MODEL DISTANCE= BOEING SPEED_GROUND HEIGHT;
OUTPUT OUT=DIAGNOSTICS1 R=RESIDUAL;
TITLE Regression Analysis of the Cleaned Dataset;
RUN;
```

Root MSE	349.05344	**R-Square**	0.8489
Dependent Mean	1522.48287	**Adj R-Sq**	0.8484
Coeff Var	22.92659		

Parameter Estimates

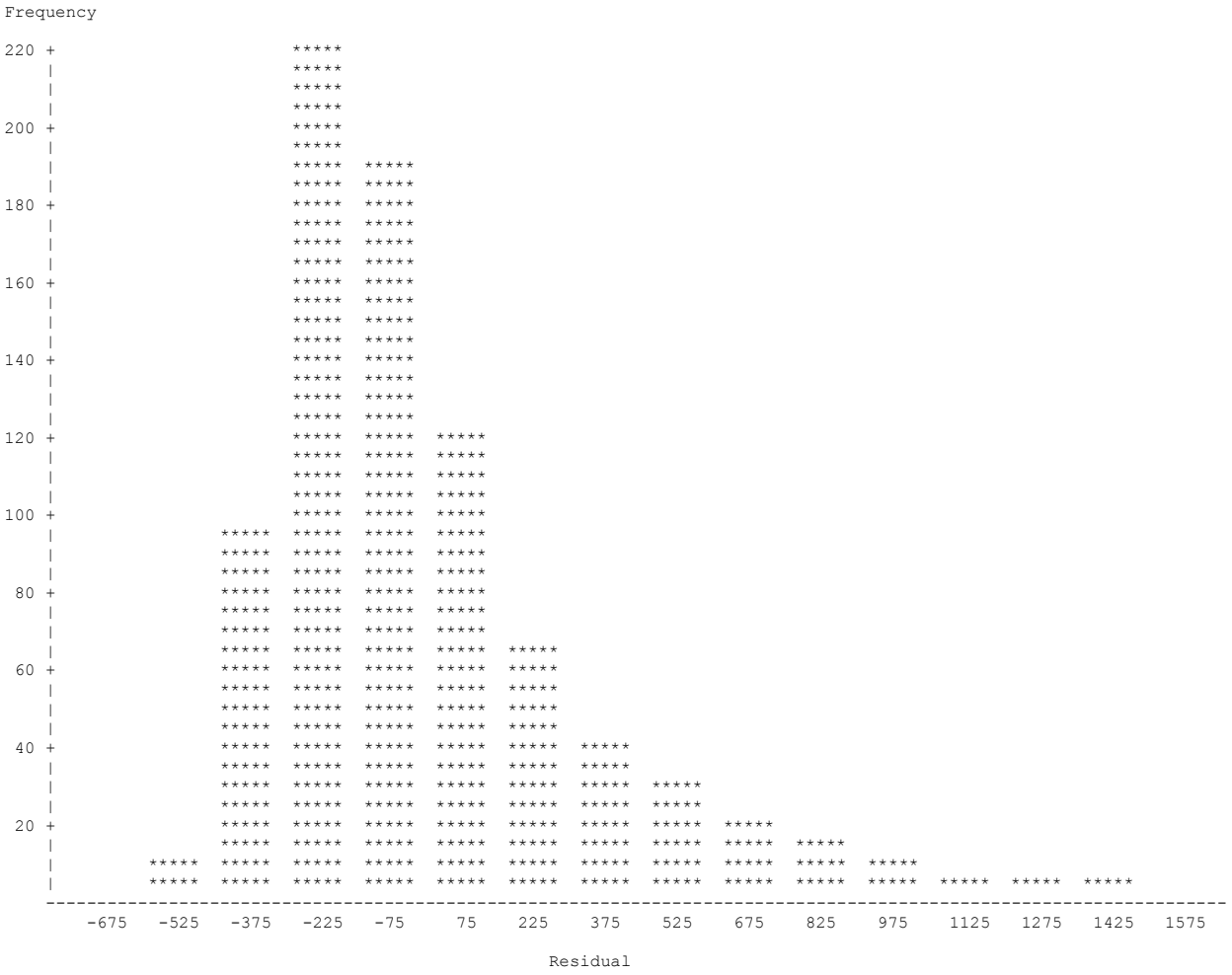
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2512.24333	68.19743	-36.84	<.0001
boeing		1	496.04524	24.29753	20.42	<.0001
speed_ground	speed_ground	1	42.40242	0.64830	65.41	<.0001
height	height	1	14.14783	1.24046	11.41	<.0001

We can see that even though we removed the ones with high p-values, our model's R-square values are worse now. Hence, let us stick with the earlier version where we consider all the variables (expect speed_air)

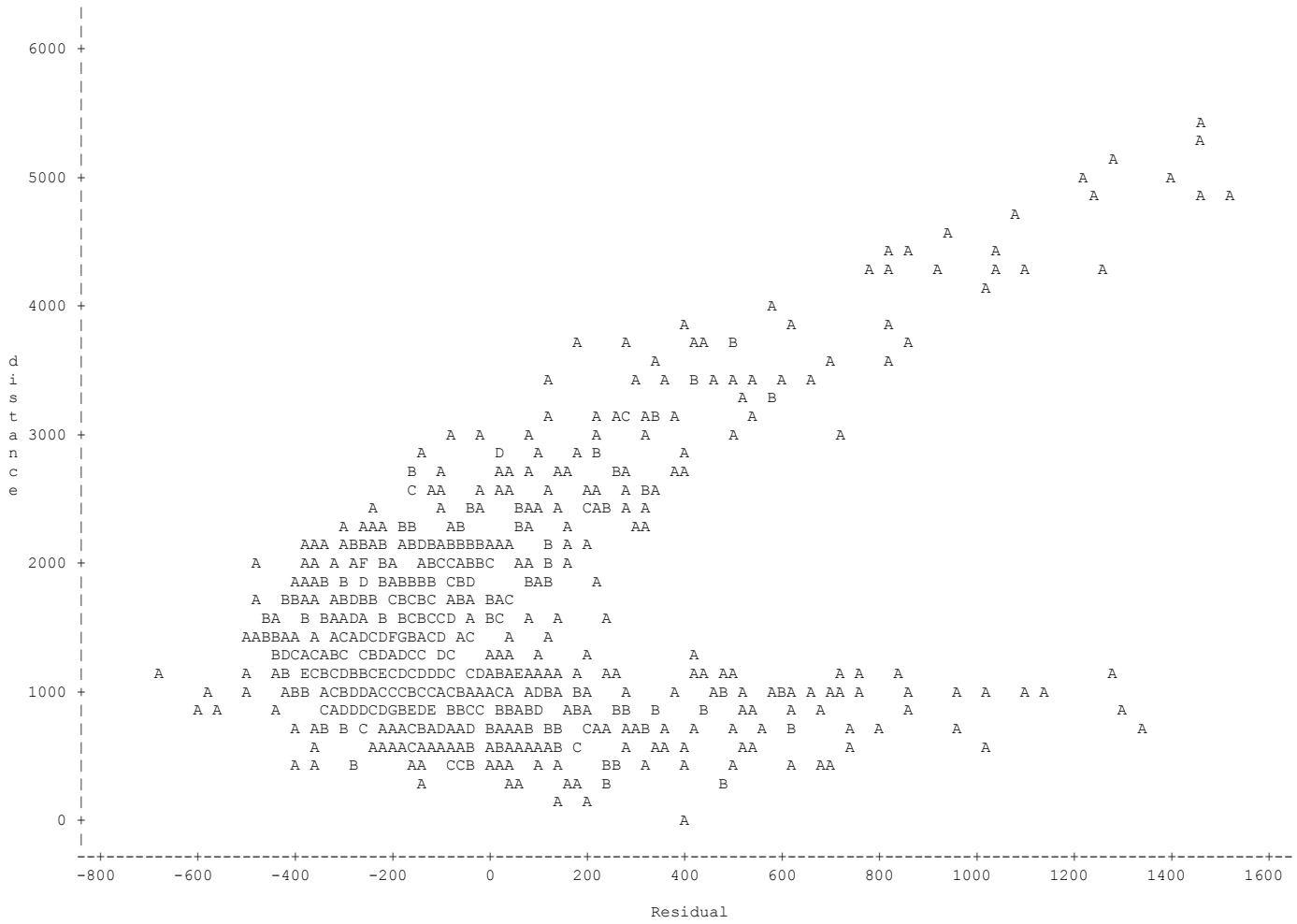
3. To check the distribution of these residuals, we look at the X_Y plot between the residuals and the distance variable, the histogram of the residuals and the results from the univariate procedure. WE

can see that the mean is close to 0, the variances are uniformly distributed. The residuals do follow a bell-shaped curve, but are very skewed to be called normally distributed.

```
PROC CHART DATA=DIAGNOSTICS;
VBAR RESIDUAL;

PROC PLOT DATA=DIAGNOSTICS;
PLOT DISTANCE*RESIDUAL;
RUN;

PROC UNIVARIATE DATA=DIAGNOSTICS NORMAL;
VAR RESIDUAL;
RUN;
```


Frequency
220
200
180
160
140
120
100
80
60
40
20
-675
-525
-375
-225
-75
75
225
375
525
675
825
975
1125
1275
1425
1575
Residual

Plot of distance*RESIDUAL. Legend: A = 1 obs, B = 2 obs, etc.



The UNIVARIATE Procedure
Variable: RESIDUAL
(Residual)

Moments			
N	831	**Sum Weights**	831
Mean	0	**Sum Observations**	0
Std Deviation	347.478519	**Variance**	120741.321
Skewness	1.58531354	**Kurtosis**	3.07047656
Uncorrected SS	100215296	**Corrected SS**	100215296
Coeff Variation	.	**Std Error Mean**	12.0538963

Basic Statistical Measures			
Location		**Variability**	
Mean	0.0000	**Std Deviation**	347.47852
Median	-91.0671	**Variance**	120741
Mode	.	**Range**	2198
		Interquartile Range	350.08724

Tests for Location: Mu0=0				
Test	**Statistic**		**p Value**	
Student's t	**t**	0	**Pr > |t|**	1.0000
Sign	**M**	-104.5	**Pr >= |M|**	<.0001
Signed Rank	**S**	-31564	**Pr >= |S|**	<.0001

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.872118	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.136329	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	5.086849	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	29.24084	**Pr > A-Sq**	<0.0050

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	1513.5794
99%	1281.1730
95%	724.7385
90%	475.9190
75% Q3	125.5888
50% Median	-91.0671

The UNIVARIATE Procedure
Variable: RESIDUAL
(Residual)

Quantiles (Definition 5)	
Level	**Quantile**
25% Q1	-224.4984
10%	-332.3025
5%	-390.0523
1%	-475.7007
0% Min	-684.4166

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
-684.417	801	1402.21	727
-609.759	542	1458.07	430
-589.690	549	1460.99	132
-563.547	802	1463.66	791
-508.907	58	1513.58	675

4.

Hence, we can see that the model is best fit when all the variables except speed_air. Speed_air is excluded from the analysis since a lot of its values are missing (~75%). Also, the values that are present are highly correlated to the speed_ground values (from observation).

1. How many observations (flights) do you use to fit your final model? If not all 950 flights, why?
Ans. 831 observations were used to fit the final model. All 950 flights weren't used to avoid duplicate data (100 rows) and data with abnormal values (19 rows).

2. What factors and how they impact the landing distance of a flight?
Ans. The factors that most significantly affect the landing distance of a flight are spped_ground, height and make (Boeing or Airbus).

3. Is there any difference between the two makes Boeing and Airbus?
Ans. While at first look, there is no significant difference between the two in terms of distribution within themselves, their impact on other variables is significant. As we saw, the aircraft make was a significant factor in deciding landing distance.

Hence, to conclude, we should look at the aircraft_make, the speed_ground and by extension, speed_air (since the two are correlated), and the height of the plane to judge the possibility of landing overrun and inimize it. Keeping these three factors tuned while keeping an eye on the number of passengers, pitch and duration of the flight will help analysts to reduce the risk of landing overrun in long run.