```
/* Importing SAS data */
PROC IMPORT DATAFILE='/folders/myfolders/FAA1.xls'
        DBMS=XLS
        OUT=WORK.FAA1;
        GETNAMES=YES;
RUN;
/*FA2 does not have a duration column*/
PROC IMPORT DATAFILE='/folders/myfolders/FAA2.xls'
        DBMS=XLS
        OUT=WORK.FAA2;
        GETNAMES=YES;
RUN;
/*1. Merging the two data sets*/
data FAA;
        set work.faa2 work.FAA1;
run;


/*We have 50 missing observations in this data set. There are also 711 missing values in the speed_air column and 150 missing in the duration column */
proc means data=FAA N NMISS Mean Std Min Max;
run;
```

Variable	Label	N	N Miss	Mean	Std Dev	Minimum	Maximum
no_pasg	no_pasg	850	0	60.1035294	7.4931370	29.0000000	87.0000000
speed_ground	speed_ground	850	0	79.4523229	19.0594903	27.7357153	141.2186354
speed_air	speed_air	208	642	103.7977237	10.2590370	90.0028586	141.7249357
height	height	850	0	30.1442223	10.2877268	-3.5462524	59.9459639
pitch	pitch	850	0	4.0093577	0.5288298	2.2844801	5.9267842
distance	distance	850	0	1526.02	928.5600816	34.0807833	6533.05
duration	duration	700	150	154.9343769	49.4058373	14.7642071	305.6217107

```
/*2. Missing observations are present. Removing extra observations that are missing*/

OPTIONS missing = ' ';

data FAA;

        set FAA;

        IF missing(cats(of _all_))

   THEN

    DELETE;

run;

/*Here I am removing any duplicate values */

proc sort data=FAA nodupkey;

        by aircraft height speed_ground no_pasg;

run;


/*After removing the missing observations and duplicate values, our sample size is 850 and out
speed_air column is still missing 600 values*/

proc means data=FAA;

run;
```

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
no_pasg	no_pasg	850	60.1035294	7.4931370	29.0000000	87.0000000
speed_ground	speed_ground	850	79.4523229	19.0594903	27.7357153	141.2186354
speed_air	speed_air	208	103.7977237	10.2590370	90.0028586	141.7249357
height	height	850	30.1442223	10.2877268	-3.5462524	59.9459639
pitch	pitch	850	4.0093577	0.5288298	2.2844801	5.9267842
distance	distance	850	1526.02	928.5600816	34.0807833	6533.05
duration	duration	700	154.9343769	49.4058373	14.7642071	305.6217107

```
/*3. Validity Check */

/*Based on this summary, there were a few instances of the ground speed or air speed being in the
abnormal range on the high and low end*/

/*There is also an issue with height at there is at least 1 negative value for height which should not be
possible. Height should be at least 6 meters*/

/*Duration should always be greater than 40 minutes. The data does not reflect this in some cases.*/
```

/*The minimum of the distance column reveals that a plane stopped 34 feet past the start of the runway - this seems fishy.

There are also examples of the distance being greater than 6,000*/

```
title 'Extreme Value Observations';

ods select ExtremeObs;

proc univariate data=FAA;
```

The UNIVARIATE Procedure
Variable: no_pasg (no_pasg)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
29	606	80	380
36	24	80	746
38	339	82	272
40	194	82	549
41	357	87	412

The UNIVARIATE Procedure
Variable: speed_ground (speed_ground)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
27.7357	567	129.307	565
29.2277	554	131.035	181
33.5741	336	132.785	497

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
33.8230	761	136.659	818
34.1178	726	141.219	559

The UNIVARIATE Procedure
Variable: speed_air (speed_air)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
90.0029	825	128.418	708
90.1110	806	131.338	181
90.3674	788	132.911	497
90.4767	780	136.423	818
90.5033	489	141.725	559

The UNIVARIATE Procedure
Variable: height (height)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-3.5462524	451	55.0935	847

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-3.3323880	1	58.0818	848
-2.9153359	2	58.0835	849
-1.5281292	452	58.2278	450
-0.0677586	3	59.9460	850

The UNIVARIATE Procedure
Variable: pitch (pitch)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
2.28448	288	5.31068	578
2.66891	404	5.32475	813
2.67133	204	5.52678	33
2.67599	384	5.55640	827
2.68955	131	5.92678	477

The UNIVARIATE Procedure
Variable: distance (distance)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
34.0808	2	5147.41	787
41.7223	32	5343.20	497
133.0869	78	5381.96	708
180.5652	28	6309.95	818
241.1610	180	6533.05	559

The UNIVARIATE Procedure
Variable: duration (duration)

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
14.7642	831	289.320	18
16.8935	346	293.230	722
31.3910	845	298.522	590
31.7017	234	302.967	345
41.9494	768	305.622	56

```
/* 4. Data Cleaning */

/* Here we are filtering out all values that would be considered abnormal */

data FAA_clean;

        set FAA;
```

```
        if duration < 40 and duration NE ' ' then delete;
        if Speed_ground < 30 or Speed_ground > 140 then delete;
        if (Speed_air < 30 or Speed_air > 140) and Speed_air NE ' ' then delete;
        if height < 6 then delete;
        if distance > 6000 then delete;
run;

/* 5. Summarizing the variable distributions */
/* There are now no more abnormal values and we now have 832 observations*/
/* Proc means will present the mean standard deviation, minimum, maximum, and the number of each variable*/
/* There are 146 values missing in the duration column */
/* there are 629 values missing in the speed_air column */
proc means data=FAA_clean;
run;
```

The MEANS Procedure

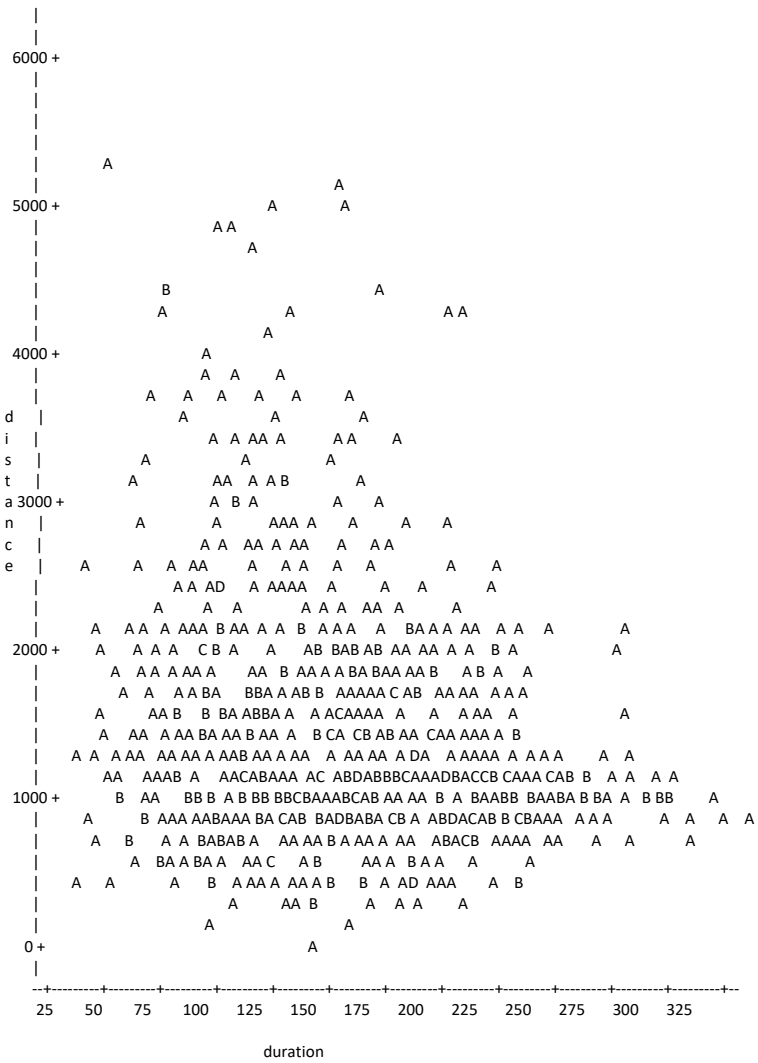
Variable	Label	N	Mean	Std Dev	Minimum	Maximum
no_pasg	no_pasg	832	60.0588942	7.4875038	29.0000000	87.0000000
speed_ground	speed_ground	832	79.5235023	18.7325852	33.5741041	132.7846766
speed_air	speed_air	203	103.4850352	9.7362774	90.0028586	132.9114649
height	height	832	30.4554041	9.7791808	6.2275178	59.9459639
pitch	pitch	832	4.0050800	0.5262573	2.2844801	5.9267842
distance	distance	832	1521.89	895.9597497	41.7223127	5381.96
duration	duration	686	155.9744209	48.7720790	41.9493694	305.6217107

```
/*Our plots for duration, height, number of passengers, and pitch did not show very significant correlation
```

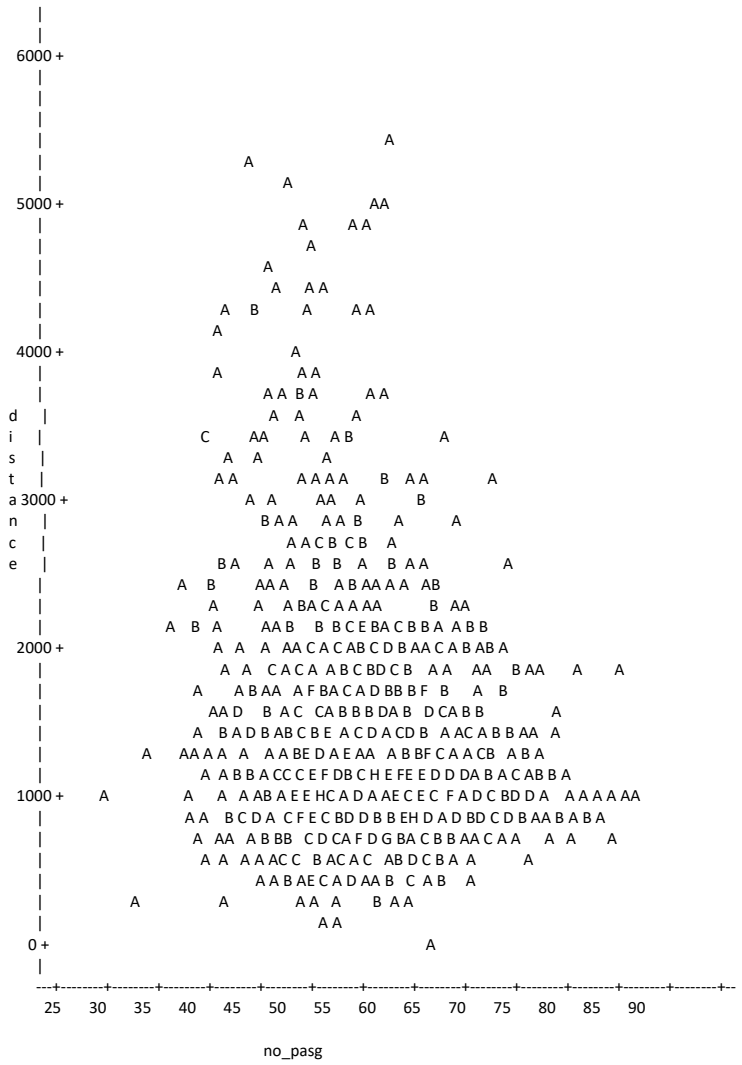
On the other hand, there was a very distinct linear relationship between speed_air/speed_ground and distance*/

```
proc plot data=FAA_clean;
        plot distance*duration;
        plot distance*no_pasg;
        plot distance*speed_air;
        plot distance*speed_ground;
        plot distance*height;
        plot distance*pitch;
run;
```

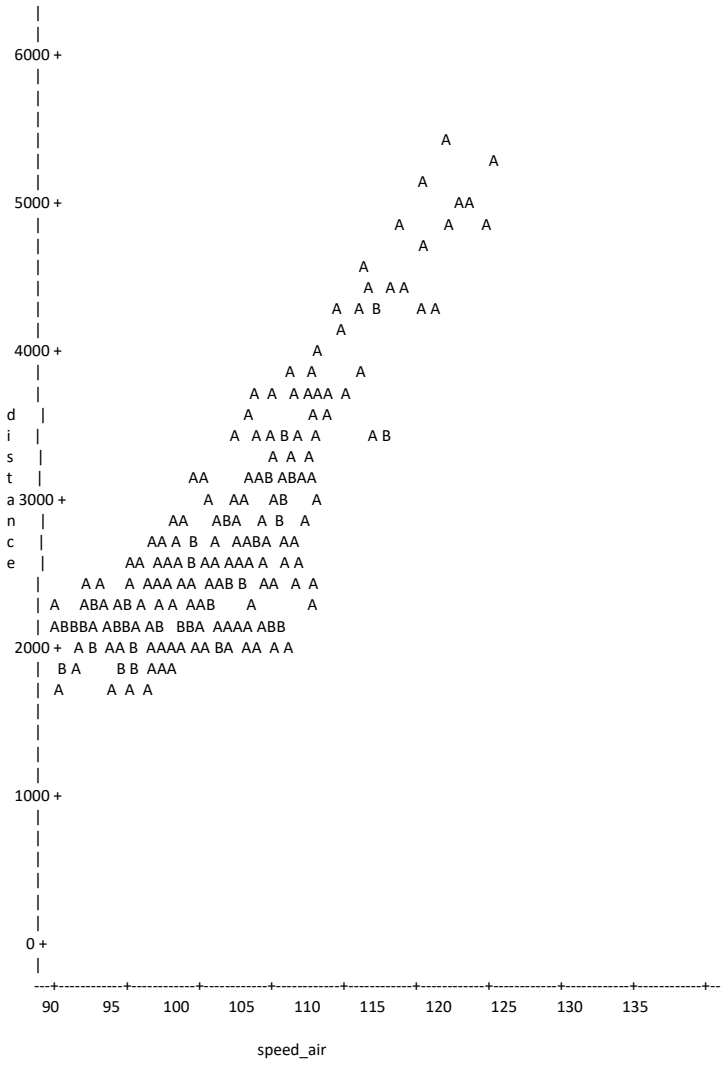
Plot of distance*duration. Legend: A = 1 obs, B = 2 obs, etc.



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

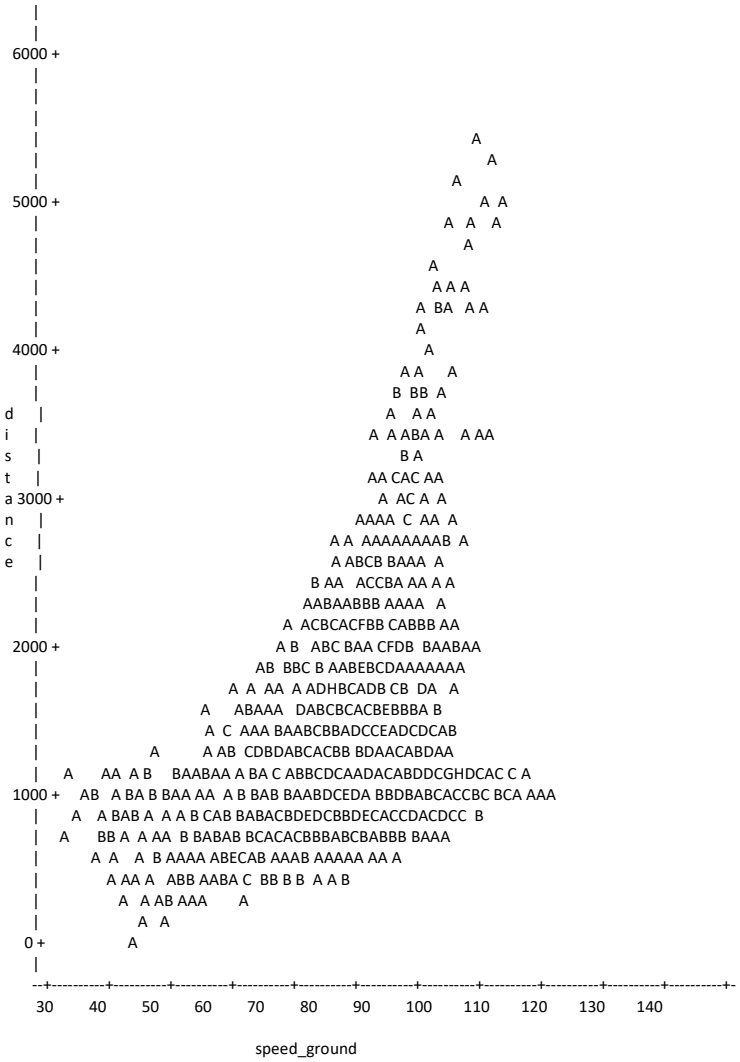


Plot of distance*speed_air. Legend: A = 1 obs, B = 2 obs, etc.

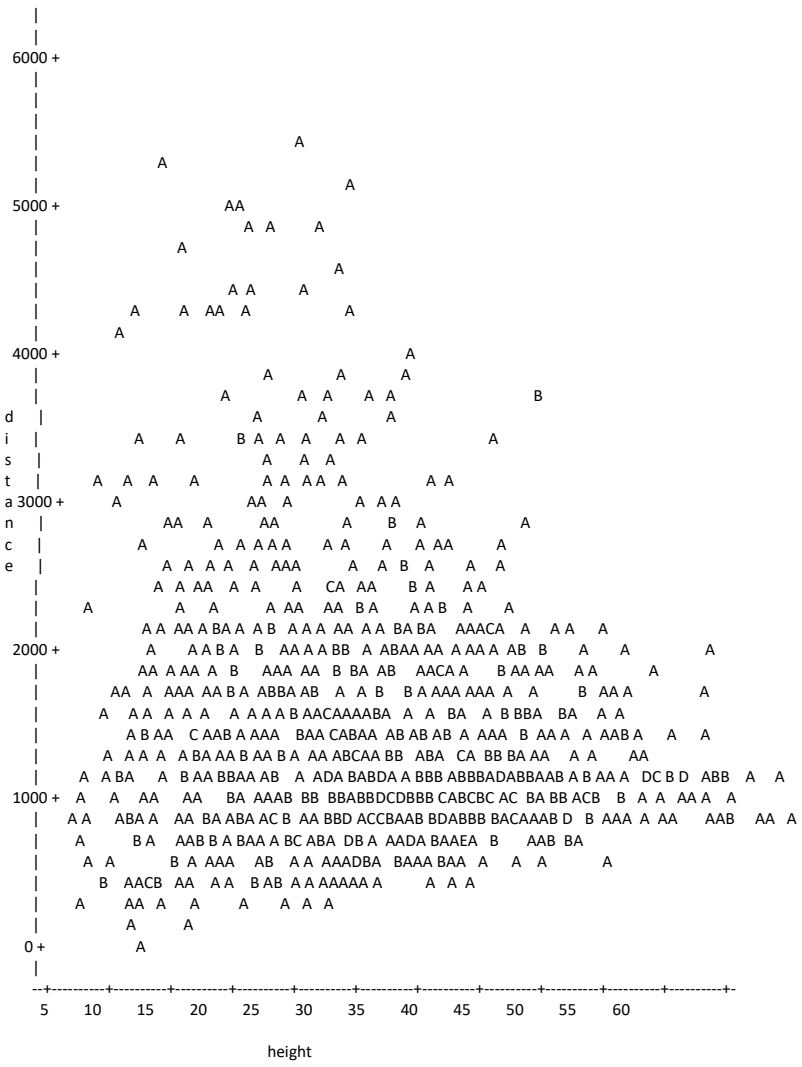


NOTE: 629 obs had missing values.

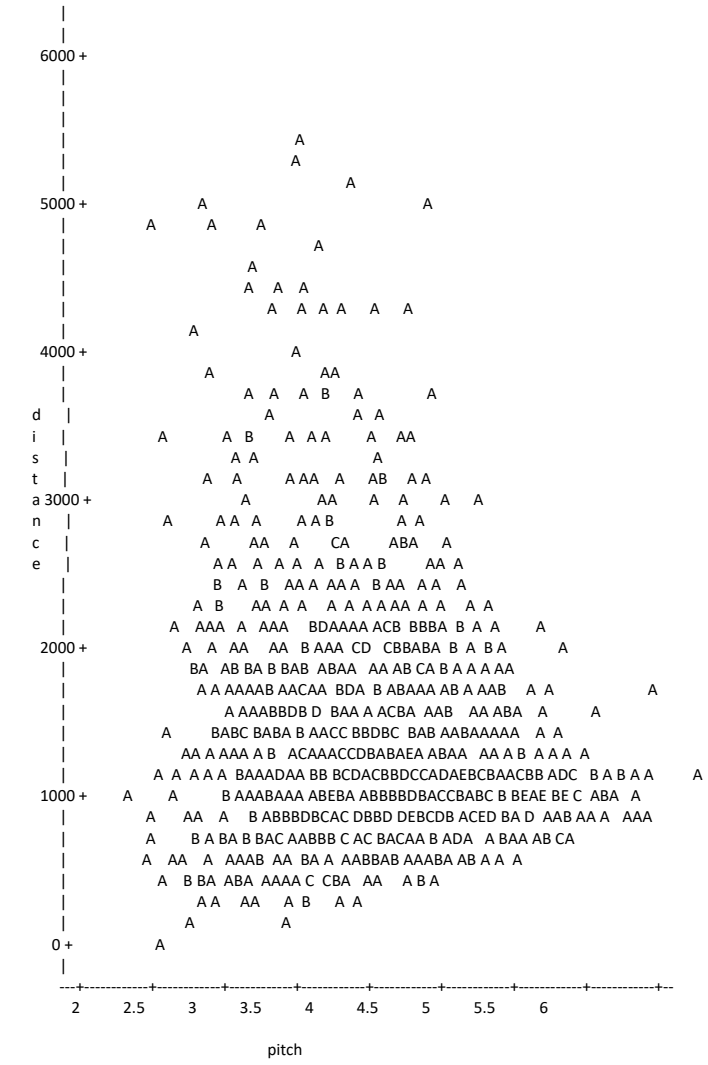
Plot of distance*speed_ground. Legend: A = 1 obs, B = 2 obs, etc.



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



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



The duration of the flight and number of passengers did not appear to have any significant effect on the landing distance of the aircraft They both had low correlation coefficients and a p-value < 0.05. Height and pitch did display a statistically significant correlation, because the p-values were < 0.05. However, they both only displayed a slight correlation of less than 0.01. On the other hand, speed_air and speed_ground both displayed large statistically significant correlations. They each has correlation coefficients of 0.94 and 0.86 respectively, and both had p-values < 0.05. */

```
proc corr data=FAA_clean;
        var duration no_pasg speed_air speed_ground height pitch;
        with distance;
        title Correlation of independents with distance;
run;
```

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

	duration	no_pasg	speed_air	speed_ground	height	pitch
distance	-0.04556	-0.01801	0.94210	0.86627	0.09953	0.08710
distance	0.2334	0.6039	<.0001	<.0001	0.0041	0.0120
	686	832	203	832	832	832

/* 1.2 to fit the regression and t-test models, I am only using 831 of the 950 observations. The reason I do not use all the observations from the data set is that there were observations that were duplicated and some observations that had values in abnormal ranges. We do not want to keep duplicates, because this will give extra weight to specific observations and reduce the generalizability of our model. We remove the abnormal values, because these values could mean that it a not a reliable data point because of how the data was processed or collected.

/* 2.2 One significant difference between our correlation and linear regression is that while both speed_air and speed_ground were highly correlated with distance and considered statistically significant in the correlation, only speed_air was considered significant with a high coefficient when compared with landing distance in the linear regression. This is likely due to issues of multi-collinearity which would not to be addressed in further analysis. Height and number of passengers were also considered statistically

significant but had smaller coefficient values. Finally, pitch was the column with the largest coefficient and was statistically significant. This is interesting because this coefficient is much larger than what seems to be represented by the correlation. The statistically significant variables with positive coefficients were height(12), speed_air (87), and pitch (147). Number of passengers had a negative coefficient of about 6. Positive coefficients indicate a larger landing distance per unit increase for the respective variable and negative coefficients indicate a decrease */

```
proc reg data=FAA_clean;
        model distance=duration no_pasg speed_air speed_ground height pitch;
        title Regrssion analysis of the flight data with distance;
        output out=reg r=residual;
run;
```

Regression analysis of the flight data with distance

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-6264.16991	312.32095	-20.06	<.0001
duration	duration	1	0.30756	0.38124	0.81	0.4210
no_pasg	no_pasg	1	-5.59392	2.60439	-2.15	0.0332
speed_air	speed_air	1	87.29059	12.50693	6.98	<.0001
speed_ground	speed_ground	1	-6.70232	12.22216	-0.55	0.5842
height	height	1	11.92614	1.97216	6.05	<.0001
pitch	pitch	1	146.69582	31.98994	4.59	<.0001

/* 3.2 Our t-test indicates to us that there is a statistically significant affect on landing distance when using one of the different aircraft makes. The means for Boeing and Airbus are 1759 and 1318 respectively. The t-test tells us the difference between these mean distances is statistically significant, and we can reject our null hypothesis that the mean for distance for each make equal. */

```
proc ttest data=FAA;
        class aircraft;
        var distance;
run;
```

The TTEST Procedure

Variable: distance (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)	Satterthwaite		-441.7		62.9027		

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