# STAT COMPUTING

# Project: Flight Landing

SUBMITTED BY: SWAPNIL SHARMA

M NUMBER: M10743959

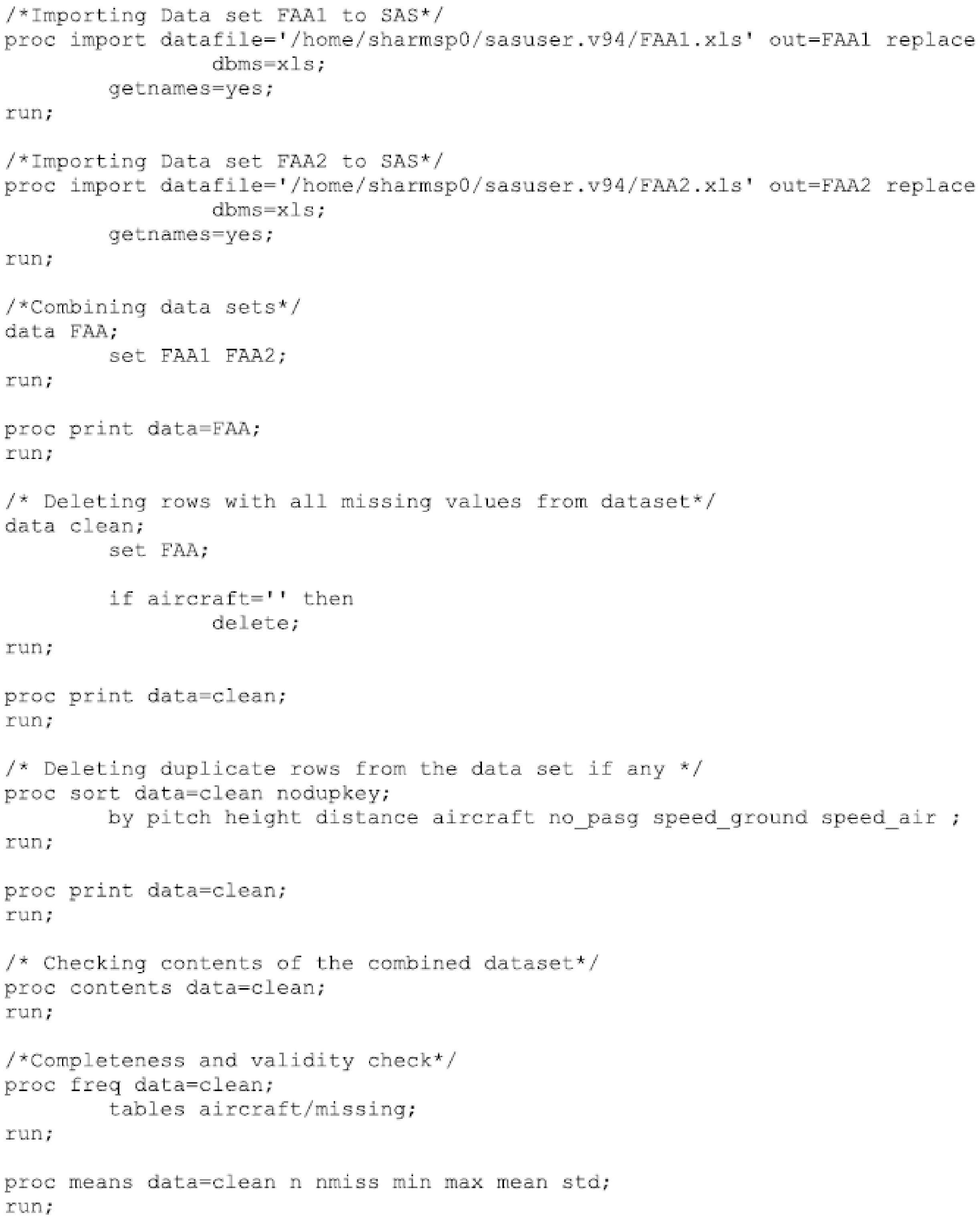
**Summary:**

To reduce the risk of landing overrun, study is done on different factors and how they impact the landing distance of commercial flight. Two raw data sets with over 950 observations are cleaned, explored, modelled and checked using SAS.  Total of 831 observations (flight) are used to fit the final regression model. Regression modelling suggest factors like speed of aircraft on ground, height of an aircraft when it is passing over the threshold runway and make of aircraft impact landing distance of flight.

## Chapter 1: Data Preparation

**Goal:** The goal is to prepare final data set for analysis from raw data set. Raw data set needs to be cleaned before analysis can be done on it as it might contain missing values, outliers, duplicate data and so on. Such data points will have adverse effects on analysis. For this we will first merge data set from different files, remove blank lines, count missing values, remove duplicate values if present, validity and completeness check and remove outlier data points thereby we will prepare dataset for analysis.

**SAS Code:**

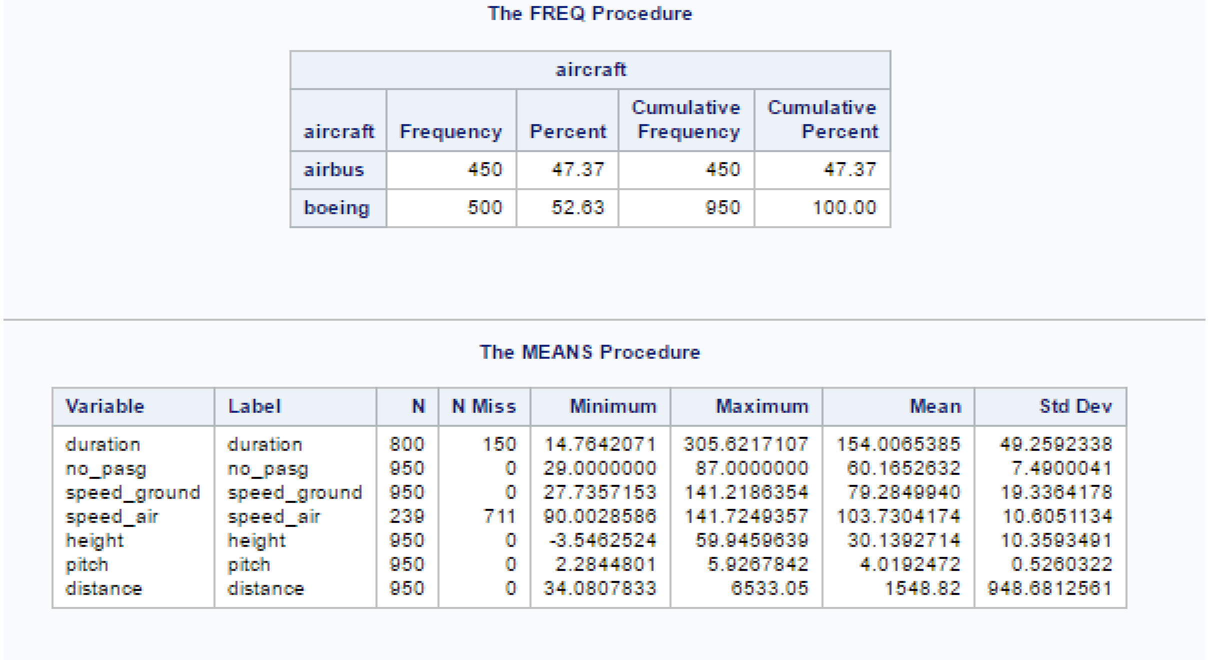


```
/*Importing Data set FAA1 to SAS*/
proc import datafile='/home/sharmsp0/sasuser.v94/FAA1.xls' out=FAA1 replace
                dbms=xls;
        getnames=yes;
run;

/*Importing Data set FAA2 to SAS*/
proc import datafile='/home/sharmsp0/sasuser.v94/FAA2.xls' out=FAA2 replace
                dbms=xls;
        getnames=yes;
run;

/*Combining data sets*/
data FAA;
        set FAA1 FAA2;
run;

proc print data=FAA;
run;

/* Deleting rows with all missing values from dataset*/
data clean;
        set FAA;

        if aircraft='' then
                delete;
run;

proc print data=clean;
run;

/* Deleting duplicate rows from the data set if any */
proc sort data=clean nodupkey;
        by pitch height distance aircraft no_pasg speed_ground speed_air ;
run;

proc print data=clean;
run;

/* Checking contents of the combined dataset*/
proc contents data=clean;
run;

/*Completeness and validity check*/
proc freq data=clean;
        tables aircraft/missing;
run;

proc means data=clean n nmiss min max mean std;
run;
```

```
        set clean;

        if height <6 then
                delete;

        if distance > 6000 then
                delete;

        if duration <=40 and duration ~=. then
                delete;

        if (speed_ground < 30 or speed_ground > 140) then
                delete;

        if (speed_air <30 or speed_air > 140) and speed_air ~=. then
                delete;
run;

/*Cross checking  distribution of each variable after removing outliers*/
proc means data=clean1 n nmiss min max mean std var median;
run;

proc freq data=clean1;
        tables aircraft;
run;

proc univariate data=clean1;
run;

/*printing histogram of each variable*/
proc chart data=clean1;
        vbar height;
        vbar distance;
        vbar duration;
        vbar speed_ground;
        vbar speed_air;
        vbar pitch;
        vbar no_pasg;
run;
```

**Output:**

Before removing outliers following was the content of dataset:



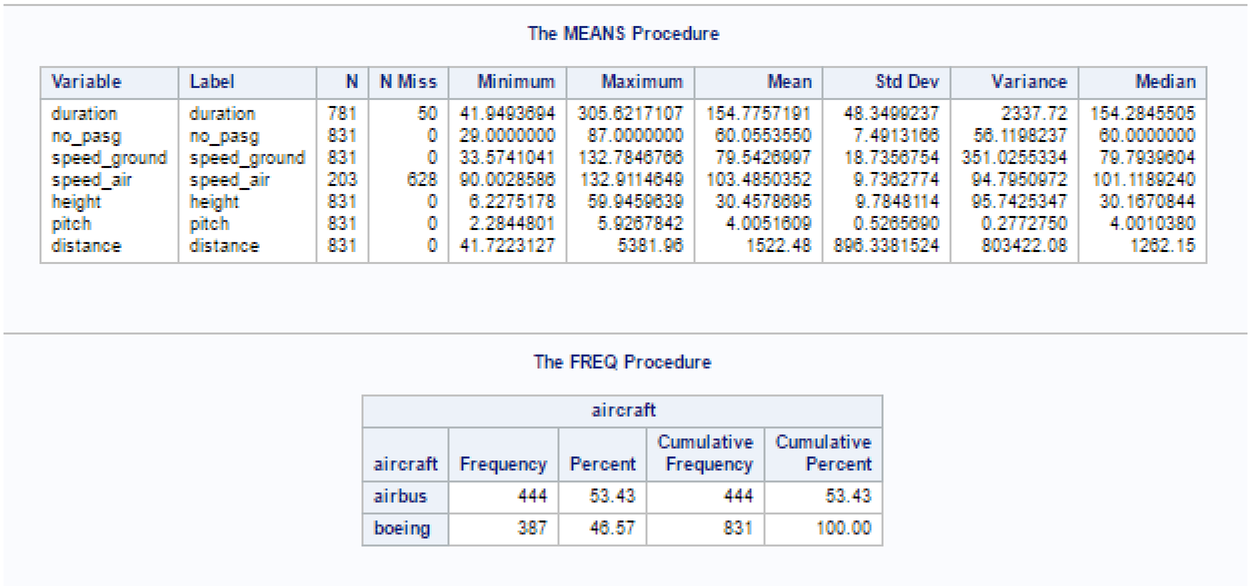
The FREQ Procedure

| aircraft | | | | |
|---|---|---|---|---|
| aircraft | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| airbus | 450 | 47.37 | 450 | 47.37 |
| boeing | 500 | 52.63 | 950 | 100.00 |

The MEANS Procedure

| Variable | Label | N | N Miss | Minimum | Maximum | Mean | Std Dev |
|---|---|---|---|---|---|---|---|
| duration | duration | 800 | 150 | 14.7642071 | 305.6217107 | 154.0065385 | 49.2592338 |
| no_pasg | no_pasg | 950 | 0 | 29.0000000 | 87.0000000 | 60.1652632 | 7.4900041 |
| speed_ground | speed_ground | 950 | 0 | 27.7357153 | 141.2186354 | 79.2849940 | 19.3364178 |
| speed_air | speed_air | 239 | 711 | 90.0028586 | 141.7249357 | 103.7304174 | 10.6051134 |
| height | height | 950 | 0 | -3.5462524 | 59.9459639 | 30.1392714 | 10.3593491 |
| pitch | pitch | 950 | 0 | 2.2844801 | 5.9267842 | 4.0192472 | 0.5260322 |
| distance | distance | 950 | 0 | 34.0807833 | 6533.05 | 1548.82 | 948.6812561 |

After removing the outliers following was the content of dataset:



The MEANS Procedure

| Variable | Label | N | N Miss | Minimum | Maximum | Mean | Std Dev | Variance | Median |
|---|---|---|---|---|---|---|---|---|---|
| duration | duration | 781 | 50 | 41.9493694 | 305.6217107 | 154.7757191 | 48.3499237 | 2337.72 | 154.2845505 |
| no_pasg | no_pasg | 831 | 0 | 29.0000000 | 87.0000000 | 60.0553550 | 7.4913166 | 56.1198237 | 60.0000000 |
| speed_ground | speed_ground | 831 | 0 | 33.5741041 | 132.7846766 | 79.5426997 | 18.7356754 | 351.0255334 | 79.7939604 |
| speed_air | speed_air | 203 | 628 | 90.0028586 | 132.9114649 | 103.4850352 | 9.7362774 | 94.7950972 | 101.1189240 |
| height | height | 831 | 0 | 6.2275178 | 59.9459639 | 30.4578695 | 9.7848114 | 95.7425347 | 30.1670844 |
| pitch | pitch | 831 | 0 | 2.2844801 | 5.9267842 | 4.0051609 | 0.5265690 | 0.2772750 | 4.0010380 |
| distance | distance | 831 | 0 | 41.7223127 | 5381.96 | 1522.48 | 896.3381524 | 803422.08 | 1262.15 |

The FREQ Procedure

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

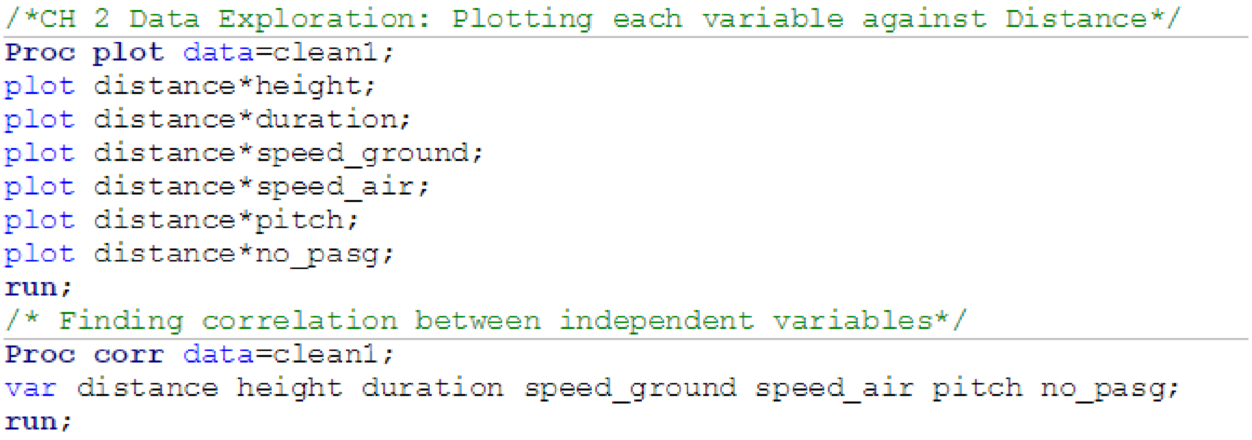
**Observation:** On merging the two data sets it was observed that there were 950 rows with 7 variables in given dataset. On viewing data set it was found 50 rows were missing in values for all variables. They were then deleted. While doing the validity and completeness check it was observed that 150 values were missing in duration and 711 values were missing in speed_air and many observations were outliers (not in the range as it should be as per given conditions). These outliers were then removed and completeness and validity check was done again. We were finally left with 831 rows of data for 7 variables.

**Conclusion:** As per our goal we have cleaned our data set without deleting the valuable information and hence data is now ready for analysis.

## Chapter 2 Data Exploration

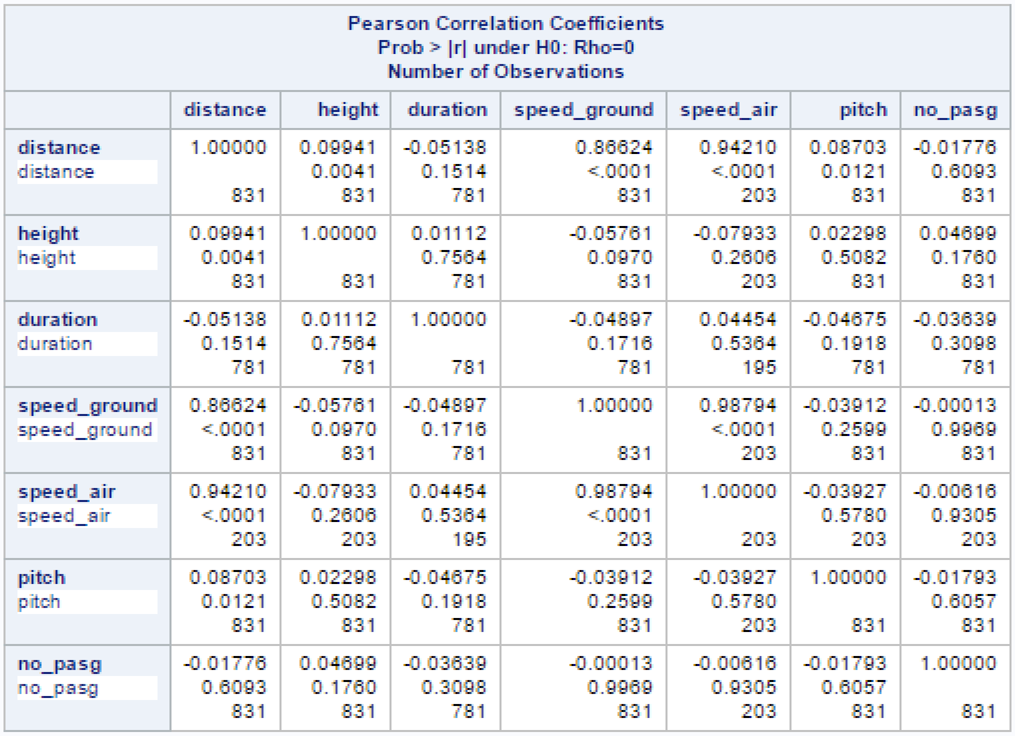
**Goal:** To study the co-relation between independent and dependent variables. Here we will study how distance (dependent variable) is co-related with other variables.

**SAS Code:**



```
/*CH 2 Data Exploration: Plotting each variable against Distance*/
Proc plot data=clean1;
plot distance*height;
plot distance*duration;
plot distance*speed_ground;
plot distance*speed_air;
plot distance*pitch;
plot distance*no_pasg;
run;
/* Finding correlation between independent variables*/
Proc corr data=clean1;
var distance height duration speed_ground speed_air pitch no_pasg;
run;
```

**Output:**



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

| | distance | height | duration | speed_ground | speed_air | pitch | no_pasg |
|---|---|---|---|---|---|---|---|
| **distance**<br>distance | 1.00000<br><br>831 | 0.09941<br>0.0041<br>831 | -0.05138<br>0.1514<br>781 | 0.86624<br><.0001<br>831 | 0.94210<br><.0001<br>203 | 0.08703<br>0.0121<br>831 | -0.01776<br>0.6093<br>831 |
| **height**<br>height | 0.09941<br>0.0041<br>831 | 1.00000<br><br>831 | 0.01112<br>0.7564<br>781 | -0.05761<br>0.0970<br>831 | -0.07933<br>0.2606<br>203 | 0.02298<br>0.5082<br>831 | 0.04699<br>0.1760<br>831 |
| **duration**<br>duration | -0.05138<br>0.1514<br>781 | 0.01112<br>0.7564<br>781 | 1.00000<br><br>781 | -0.04897<br>0.1716<br>781 | 0.04454<br>0.5364<br>195 | -0.04675<br>0.1918<br>781 | -0.03639<br>0.3098<br>781 |
| **speed_ground**<br>speed_ground | 0.86624<br><.0001<br>831 | -0.05761<br>0.0970<br>831 | -0.04897<br>0.1716<br>781 | 1.00000<br><br>831 | 0.98794<br><.0001<br>203 | -0.03912<br>0.2599<br>831 | -0.00013<br>0.9969<br>831 |
| **speed_air**<br>speed_air | 0.94210<br><.0001<br>203 | -0.07933<br>0.2606<br>203 | 0.04454<br>0.5364<br>195 | 0.98794<br><.0001<br>203 | 1.00000<br><br>203 | -0.03927<br>0.5780<br>203 | -0.00616<br>0.9305<br>203 |
| **pitch**<br>pitch | 0.08703<br>0.0121<br>831 | 0.02298<br>0.5082<br>831 | -0.04675<br>0.1918<br>781 | -0.03912<br>0.2599<br>831 | -0.03927<br>0.5780<br>203 | 1.00000<br><br>831 | -0.01793<br>0.6057<br>831 |
| **no_pasg**<br>no_pasg | -0.01776<br>0.6093<br>831 | 0.04699<br>0.1760<br>831 | -0.03639<br>0.3098<br>781 | -0.00013<br>0.9969<br>831 | -0.00616<br>0.9305<br>203 | -0.01793<br>0.6057<br>831 | 1.00000<br><br>831 |

**Observation:**

It is observed that Pearson Correlation Coefficient is 0.86 for distance and speed_ground. Thus they are strongly positively related. Moreover, it is observed that speed_air and speed_ground have high Pearson Correlation Coefficient (0.98). As 75% values of speed_air is missing we will not use that in model and thereby reduce multi collinearity as well.

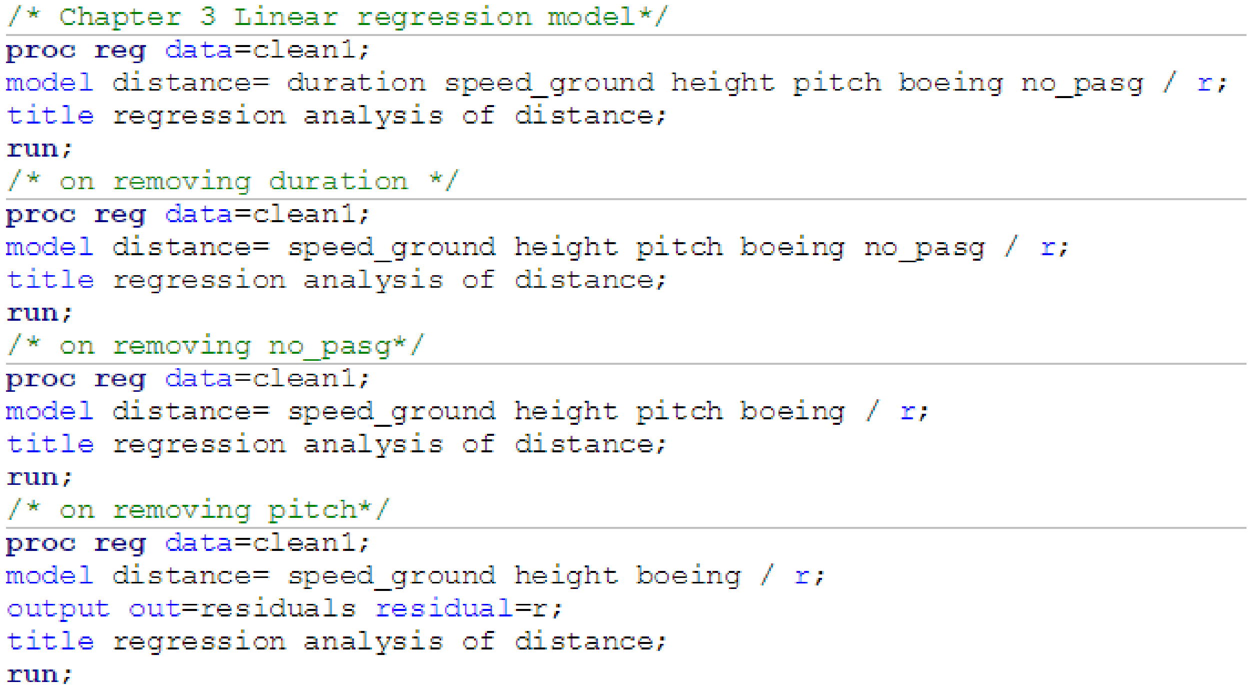
**Conclusion:**

We have successfully studied the relationship between independent and dependent variables as well as relationship in between independent variables.

**Chapter 3: Modelling**

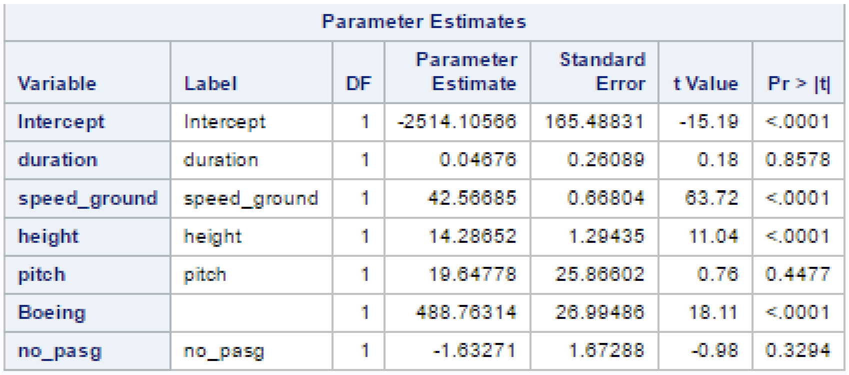
**Goal:** To build a linear regression model that describes the relationship between a response variable (Read Landing distance) and several predictor variables (read height duration speed_ground etc).

**SAS Code:**



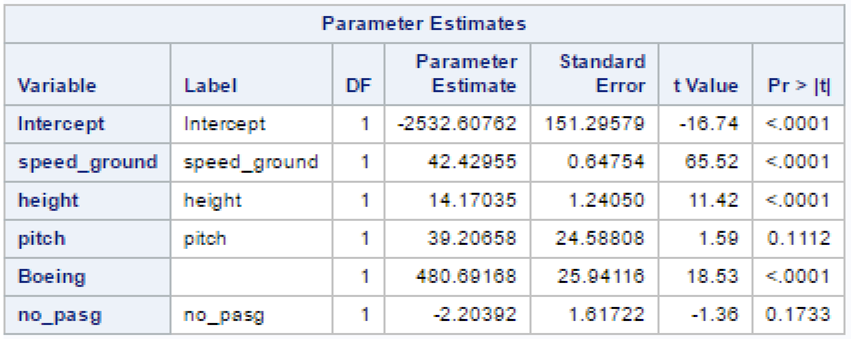
```
/* Chapter 3 Linear regression model*/
proc reg data=clean1;
model distance= duration speed_ground height pitch boeing no_pasg / r;
title regression analysis of distance;
run;
/* on removing duration */
proc reg data=clean1;
model distance= speed_ground height pitch boeing no_pasg / r;
title regression analysis of distance;
run;
/* on removing no_pasg*/
proc reg data=clean1;
model distance= speed_ground height pitch boeing / r;
title regression analysis of distance;
run;
/* on removing pitch*/
proc reg data=clean1;
model distance= speed_ground height boeing / r;
output out=residuals residual=r;
title regression analysis of distance;
run;
```

**Output:**



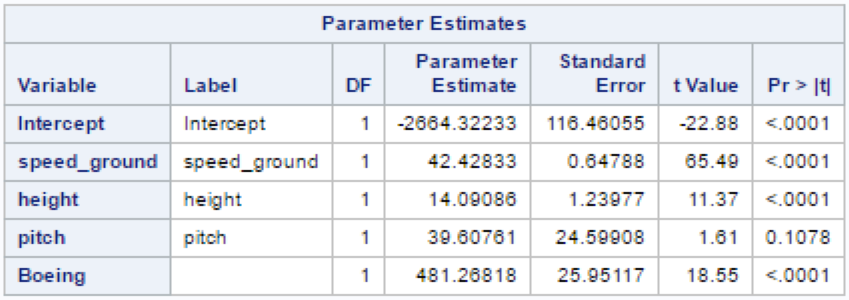
Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | -2514.10566 | 165.48831 | -15.19 | <.0001 |
| duration | duration | 1 | 0.04676 | 0.26089 | 0.18 | 0.8578 |
| speed_ground | speed_ground | 1 | 42.56685 | 0.66804 | 63.72 | <.0001 |
| height | height | 1 | 14.28652 | 1.29435 | 11.04 | <.0001 |
| pitch | pitch | 1 | 19.64778 | 25.86602 | 0.76 | 0.4477 |
| Boeing | | 1 | 488.76314 | 26.99486 | 18.11 | <.0001 |
| no_pasg | no_pasg | 1 | -1.63271 | 1.67288 | -0.98 | 0.3294 |

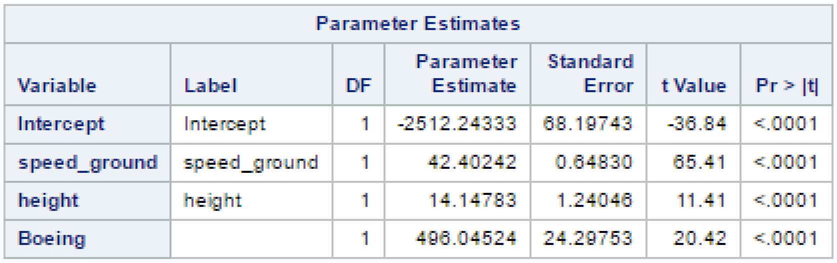


Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | -2532.60762 | 151.29579 | -16.74 | <.0001 |
| speed_ground | speed_ground | 1 | 42.42955 | 0.64754 | 65.52 | <.0001 |
| height | height | 1 | 14.17035 | 1.24050 | 11.42 | <.0001 |
| pitch | pitch | 1 | 39.20658 | 24.58808 | 1.59 | 0.1112 |
| Boeing | | 1 | 480.69168 | 25.94116 | 18.53 | <.0001 |
| no_pasg | no_pasg | 1 | -2.20392 | 1.61722 | -1.36 | 0.1733 |



| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | -2664.32233 | 116.46055 | -22.88 | <.0001 |
| speed_ground | speed_ground | 1 | 42.42833 | 0.64788 | 65.49 | <.0001 |
| height | height | 1 | 14.09086 | 1.23977 | 11.37 | <.0001 |
| pitch | pitch | 1 | 39.60761 | 24.59908 | 1.61 | 0.1078 |
| Boeing | | 1 | 481.26818 | 25.95117 | 18.55 | <.0001 |



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

**Observation:**

To build the regression model we kept on iterating equation by removing independent variables whose P value was greater than 0.05. This was done as null hypothesis (Coefficient of variable is zero) is rejected because of lower P value(P<0.05). Make of aircraft Boeing is coded as 1 for running model.

**Conclusion:**

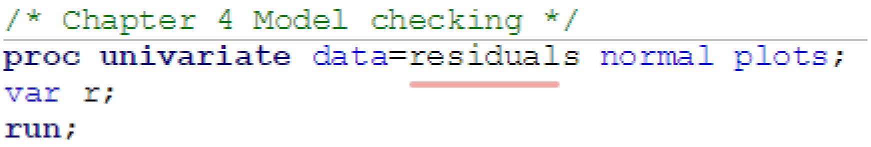
After iteration, our final equation shows that speed_ground, Height and make of aircraft are the variables that affect the landing distance.

**Chapter 4: Model checking**

**Goal:** We need to check the assumptions that we made while building regression model. The assumptions for residuals were as follow:

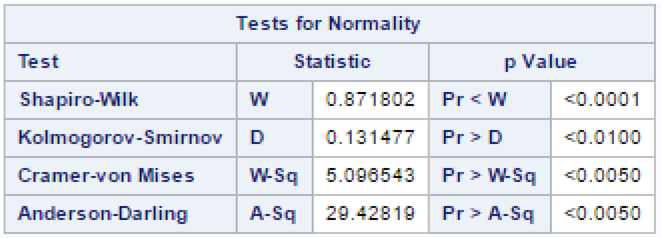
**a)** Mean is zero
**b)** Variance is constant
**c)** Normally distributed
**d)** Independent

**SAS Code:**

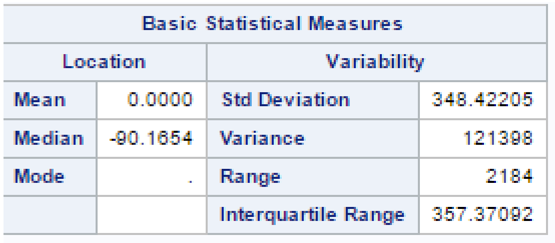


```
/* Chapter 4 Model checking */
proc univariate data=residuals normal plots;
var r;
run;
```

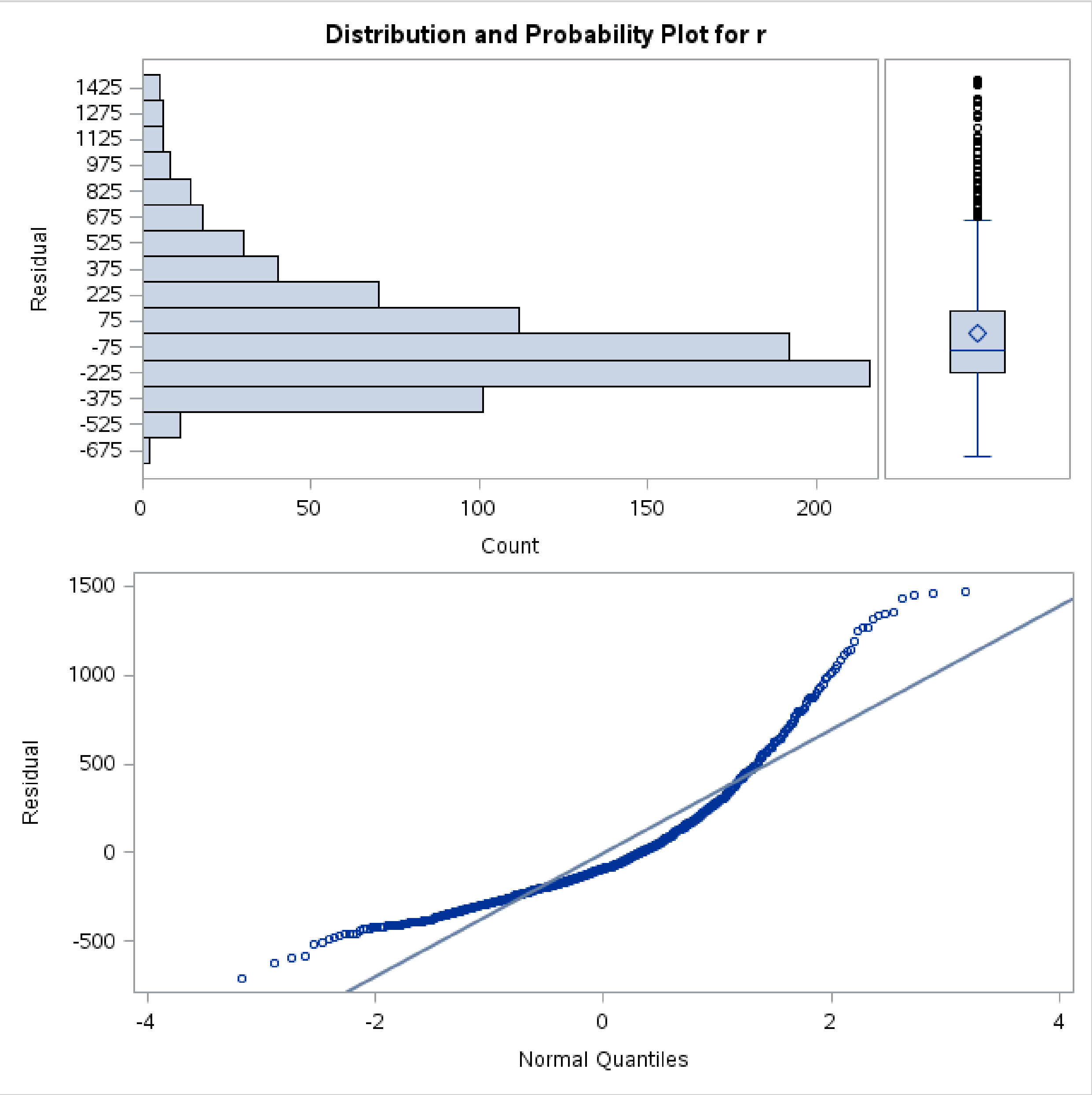
**Output:**

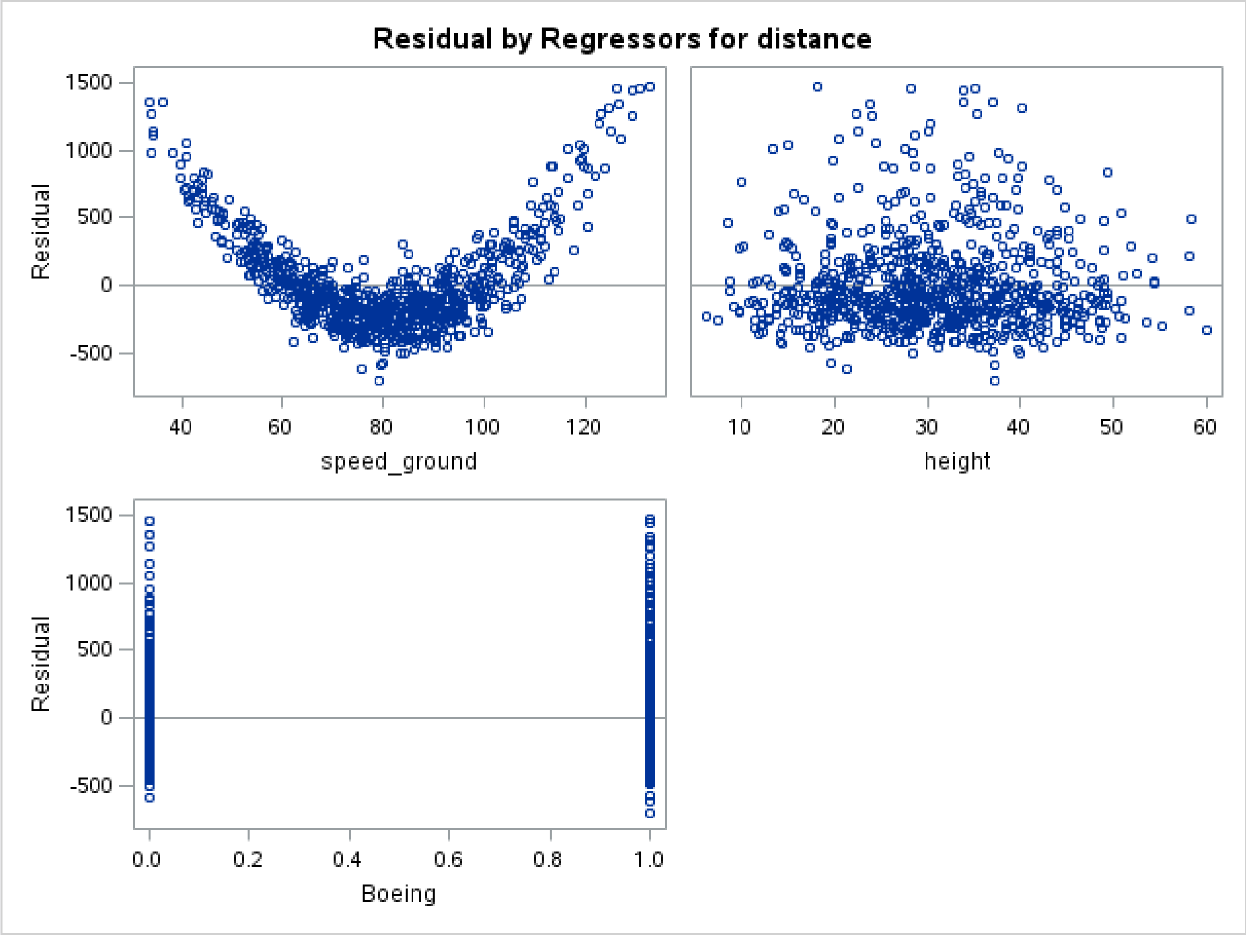


| Tests for Normality | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.871802 | Pr < W | <0.0001 |
| Kolmogorov-Smirnov | D | 0.131477 | Pr > D | <0.0100 |
| Cramer-von Mises | W-Sq | 5.096543 | Pr > W-Sq | <0.0050 |
| Anderson-Darling | A-Sq | 29.42819 | Pr > A-Sq | <0.0050 |



| Basic Statistical Measures | | | |
|---|---|---|---|
| Location | | Variability | |
| Mean | 0.0000 | Std Deviation | 348.42205 |
| Median | -90.1654 | Variance | 121398 |
| Mode | . | Range | 2184 |
| | | Interquartile Range | 357.37092 |

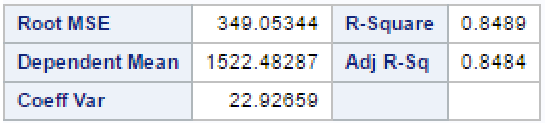

Distribution and Probability Plot for r
Residual
1425
1275
1125
975
825
675
525
375
225
75
-75
-225
-375
-525
-675
0
50
100
150
200
Count
Residual
1500
1000
500
0
-500
-4
-2
0
2
4
Normal Quantiles



**Observation:** We observe that residuals are normally distributed (confirmed using test of Normality) and Mean is zero (Basic Statistical measure table). Moreover, above plots show the independence of independent variables with residuals.

**Conclusion:** All the assumptions made while model building is verified and residual diagnosis is done successfully.

**Result:** Thus the final model contains variables **of speed_ground, height and Boeing (aircraft type**). The resultant R- Square value is **0.8489** i.e. We can predict our distance (dependent variable) with **~85%** variability rest 16% is noise.



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

Write your short answers to these questions:

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

   Answer: We used 831 observations to fit the model. The other observations were removed based on abnormal values present for given variables.

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

   Answer: Three factors impact the landing distance of flight. They are as follow:
   1.speed_ground: It is positively co-related. i.e. Distance increases(decreases) on increase(decrease) in speed_ground.
   2.Height: It is positively co-related. i.e. Distance increases(decreases) on increase(decrease) in Height.
   3.Boeing (Make of Aircraft): It is positively co-related. i.e. Distance increases when aircraft is Boeing keeping all other variables constant.

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

   Answer: Model shows that there is a significant difference between two makes of Boeing and Airbus. Boeing is positively co-related while Airbus is negatively.