FAA Flight Landing Overrun Analysis and Modeling

Anupam Shukla

M13469377

MS, Business Analytics

Project Overview

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.

Data: Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). See two Excel files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights).

Summary:

The given airline data was analyzed with the purpose of studying the factors that impact the flight landing distance in order to accurately predict landing distances, given other conditions, and reducing the risk of landing overrun.

Before working on data analysis the two .xlsx files were converted to .csv files as part of this study in R.

The project included the following steps in order - data cleaning and analysis, data visualization and descriptive analysis and linear regression modeling.

During data cleaning, 100 duplicate records were found on combining FAA1 and FAA2 datasets.50 records did not have any value for duration column and 18 abnormal values were found. All these records were removed resulting in 782 records on which further analysis and modeling was carried out. Additionally ~80% missing values were found for speed_air column but since this was a huge chunk of the given data, it was decided neither to remove these observations nor to impute these values.

In data visualization and descriptive analysis, plots were drawn and correlation analysis among variables was done. Based on the analysis, it was decided that speed_ground, pitch and height are to be included as part of the regression model.

In the Modeling part, three regression approaches were adopted – first with speed_ground as the only predictor for distance variable, second consisted of speed_ground and pitch as predictors while the third approach consisted of speed_ground, pitch and height as predictors. Best results were found in the third approach.

The detailed description with code and figures/graphs for each of the above steps are provided in the report.

A- Data Cleaning and Analysis

i. Combining data sets from different sources:

In this study, we are given two datasets 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights). These datasets were first converted to .csv files and combined using below R Code.

R Code:

----------X----------

```
#read csv files

faa1 <- read.csv('FAA1.csv')

faa2 <- read.csv('FAA2.csv')

library('plyr')

faa <- rbind.fill(faa1,faa2)


#remove duplicate rows from dataframe

library(tidyverse)

faa <- as_tibble(faa)

faa.no.dup <- distinct(faa,aircraft,no_pasg,speed_ground,speed_air,height,pitch,distance,.keep_all = T)

faa.no.dup <- faa.no.dup[1:850,]
```

----------X----------

Findings – Total 950 records were present after combining the two datasets. 100 duplicate rows were found in first and second dataset which were removed reducing total number of records to 850 flights.

ii. Performing the completeness check of each variable

In this step, we examined if there are any missing values in each of the variable in the combined dataset obtained in step 1.

Output:

aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
0	50	0	0	642	0	0	0

R code:

----------X----------

```
#print missing values of each column

sapply(faa.no.dup, function(x) sum(is.na(x)))
```

----------X----------

Findings - We observed two columns have missing values – duration and speed_air.

First, we will analyze the duration column.

The below table depicts the correlation between the predictors and response variable(distance)

duration	no_pasg	speed_ground	speed_air	height	pitch
-0.06208864	-0.03033129	0.8619638	0.947275	0.1362407	0.1026875

R code:

----------X----------

```
#correlation between the variables

faa.temp <- faa.no.dup[,-1]

corr <- cor(faa.temp$distance,faa.temp[,-7],use = "pairwise.complete.obs")

corr
```

----------X----------

After removing the missing values in duration column and rerunning the correlation analysis, we get the following table–

duration	no_pasg	speed_ground	speed_air	height	pitch
-0.06208864	-0.03022184	0.8632804	0.9482677	0.1426631	0.08627321

R code:

----------X----------

```
#delete rows with missing value of duration column

removerows <- function(data, desiredCols) {

  completeVec <- complete.cases(data[, desiredCols])

  return(data[completeVec, ])

}

faa.no.dup <- removerows(faa.no.dup, "duration")

faa.temp1 <- faa.no.dup[,-1]

corr1 <- cor(faa.temp1$distance,faa.temp1[,-7],use = "pairwise.complete.obs")

corr1
```

----------X----------

We don't see a significant difference in the correlations of the predictor variables with response variable of the two tables and hence proceed with deleting the 50 missing values in the duration column. (analysis was also done with imputing the values with mean and median but was insignificant)

Missing values after rows with missing values of duration are removed –

aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
0	0	0	0	600	0	0	0

R code:

----------X----------

#output after missing values of duration are removed

sapply(faa.no.dup, function(x) sum(is.na(x)))

----------X----------

Now, we examine the speed_air column which has almost 600 missing values. Values on running correlation analysis with speed_air as the dependent variable and all other variables as independent variables, it was found that the variables speed_air and speed_ground are highly correlated. So, including both these variables in the model will lead to multicollinearity issue. Also, the percentage of missing values are too high, hence we neither delete or impute these values.

duration	no_pasg	speed_ground	height	pitch	distance
0.04911196	-5.605937e-05	0.989658	-0.07251648	-0.0006224996	0.9482677

R Code:

----------X----------

#checking correlation of speed_air variable with other variables

corr2 <- cor(faa.no.dup$speed_air,faa.temp1[,-4],use = "pairwise.complete.obs")

corr2

----------X----------

iii. Performing the validity check of each variable – examine if abnormal values are present –

R code:

----------X----------

#remove abnormal values

height_abnormal <- sum(faa.no.dup$height<6)

duration_abnormal <- sum(faa.no.dup$duration<40)

speed_ground_abnormal <- sum(faa.no.dup$speed_ground<30 | faa.no.dup$speed_ground > 140)

height_abnormal;duration_abnormal;speed_ground_abnormal

----------X----------

On running the above we find that there are 10 records where height < 6m, 5 records where duration of flight < 40 mins and 3 records for which speed_ground < 30mph or <140mph. We will remove these values from the dataset in the next step.

iv. Cleaning the data

R code:

----------X----------

```
#remove abnormal values

faa.no.dup <- faa.no.dup[faa.no.dup$height > 6,]

faa.no.dup <- faa.no.dup[faa.no.dup$duration > 40,]

faa.no.dup <- faa.no.dup[faa.no.dup$speed_ground > 30,]

faa.no.dup <- faa.no.dup[faa.no.dup$speed_ground < 140,]
```

----------X----------

v. Summarizing the distribution of each variable

R code:

----------X----------

```
#summary of variables

summary(faa.no.dup$duration)

summary(faa.no.dup$no_pasg)

summary(faa.no.dup$speed_ground)

summary(faa.no.dup$speed_air)

summary(faa.no.dup$height)

summary(faa.no.dup$pitch)

summary(faa.no.dup$distance)
```

----------X----------

Output:

Duration:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
41.95	119.65	154.26	154.73	189.64	305.62

No_pasg:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
29.00	55.00	60.00	60.09	65.00	87.00

Speed_ground

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
33.57	66.20	79.83	79.71	92.26	136.66

Speed_air:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
90.00	96.16	100.99	103.67	109.48	136.42	586

Height:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.228	23.596	30.240	30.473	36.993	59.946

Pitch:

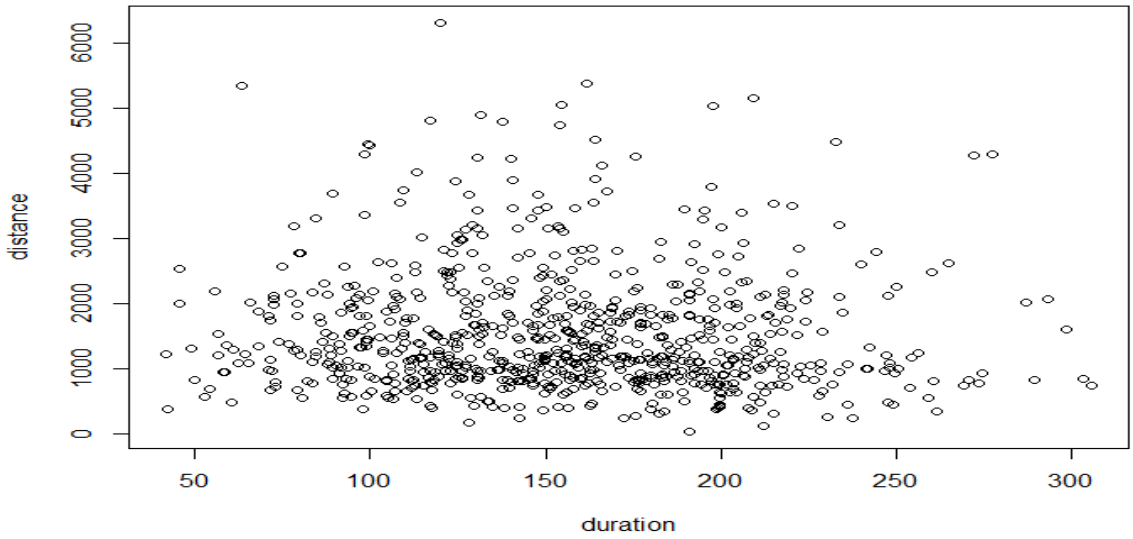
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.284	3.654	4.015	4.014	4.382	5.927

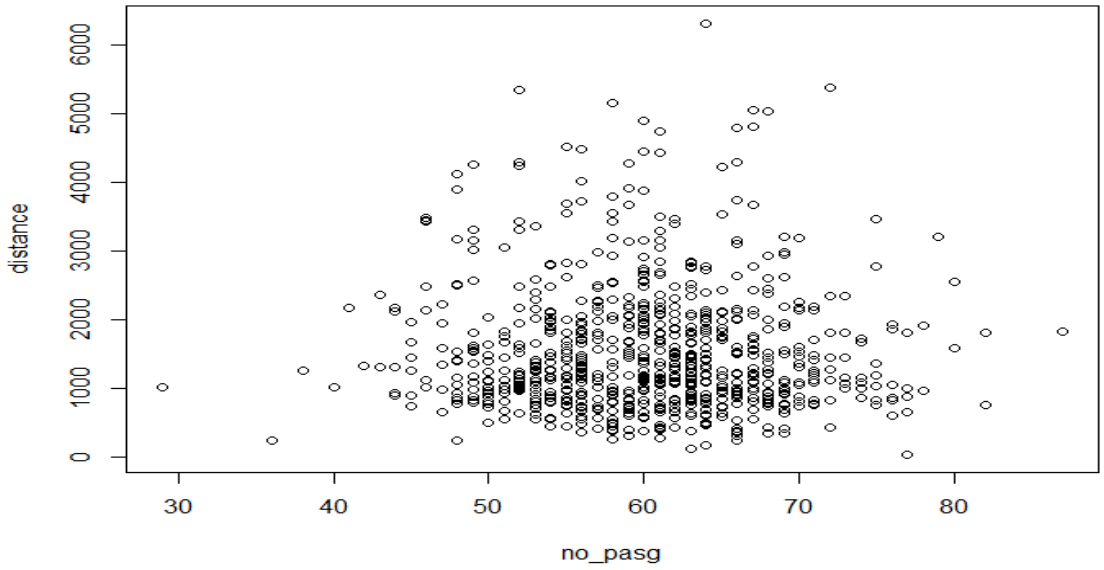
Distance:

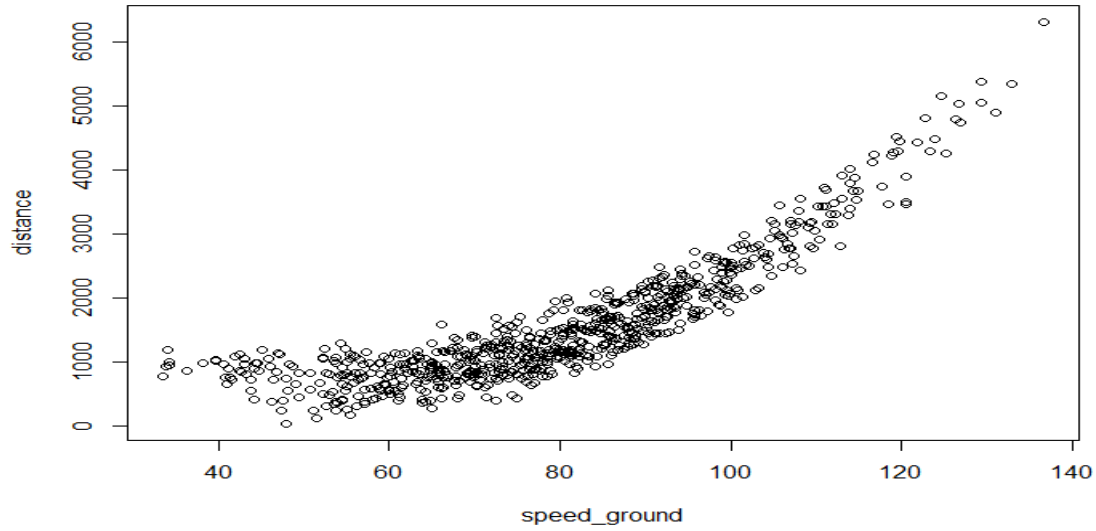
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
41.72	919.36	1277.47	1547.30	1960.46	6309.95

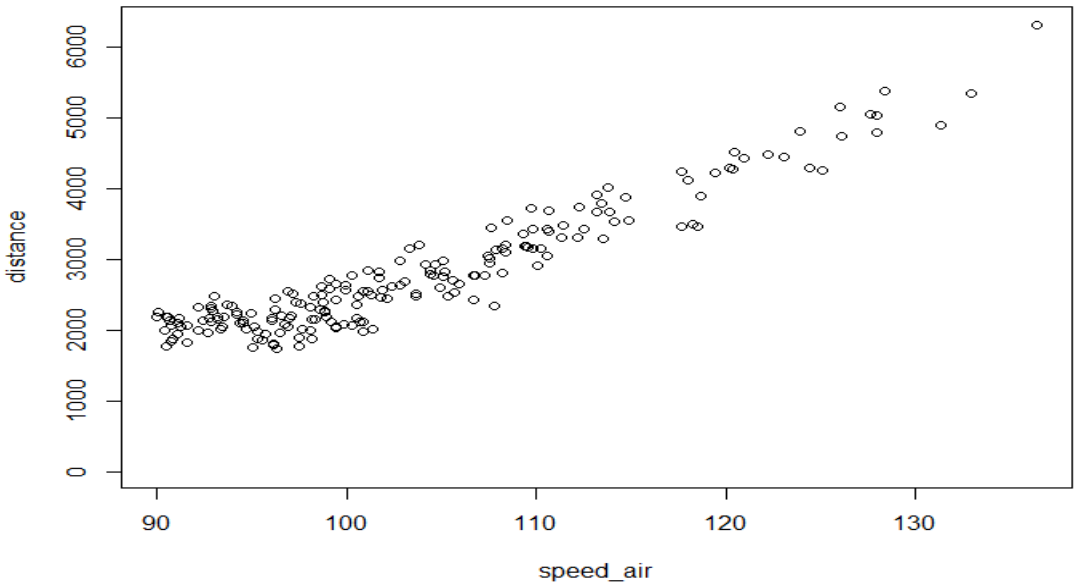
B- Data visualization and descriptive analysis

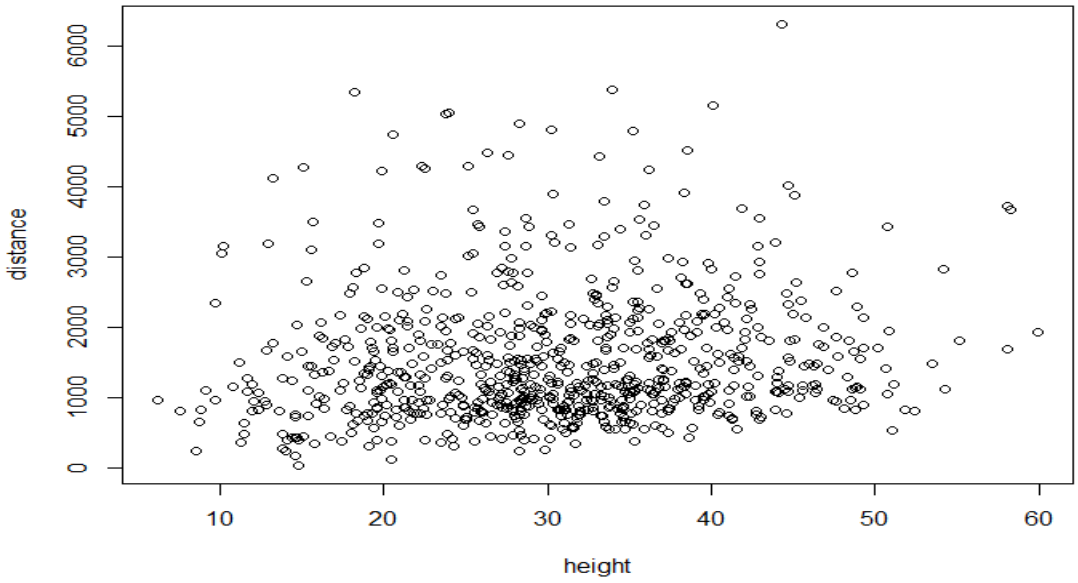
i. X-Y Plots

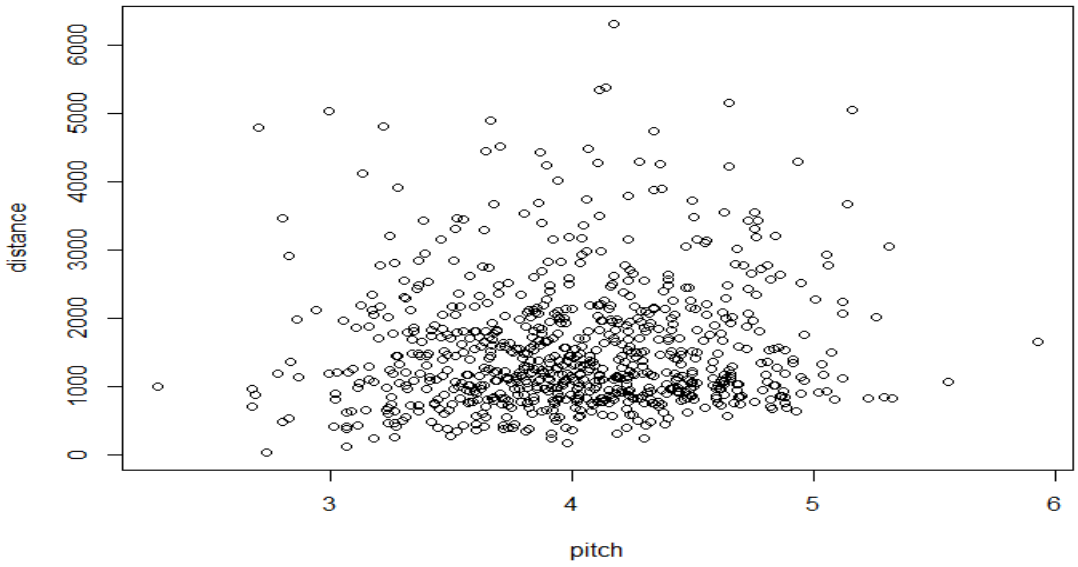

distance
0
1000
2000
3000
4000
5000
6000
50
100
150
200
250
300
duration


distance
0
1000
2000
3000
4000
5000
6000
30
40
50
60
70
80
no_pasg


distance
0
1000
2000
3000
4000
5000
6000
40
60
80
100
120
140
speed_ground


distance
0
1000
2000
3000
4000
5000
6000
90
100
110
120
130
speed_air


distance
0
1000
2000
3000
4000
5000
6000
10
20
30
40
50
60
height


distance
0
1000
2000
3000
4000
5000
6000
3
4
5
6
pitch

Observations:

1. From the above plots, it can be seen that duration variable does not have a relationship with distance variable (this will also be clear in the correlation analysis in the following section)

2. The distance variable is highly correlated with speed_ground and speed_air variables. However, as we have already established earlier, these variables are highly correlated themselves and since there are almost 80% missing values in the speed_air column we will choose speed_ground as one of the predictors.

3. The relationship does not appear much significant between distance with pitch and height variables, however this will become more clear once we run the correlation analysis.

R Code:

----------X----------

```
#plot scatterplots between response(distance) and predictor variables

plot(faa.no.dup$duration,faa.no.dup$distance, xlab = 'duration', ylab = 'distance')

plot(faa.no.dup$no_pasg,faa.no.dup$distance, xlab = 'no_pasg', ylab = 'distance')

plot(faa.no.dup$speed_ground,faa.no.dup$distance, xlab = 'speed_ground', ylab = 'distance')

plot(faa.no.dup$speed_air,faa.no.dup$distance, xlab = 'speed_air', ylab = 'distance')

plot(faa.no.dup$height,faa.no.dup$distance, xlab = 'height', ylab = 'distance')

plot(faa.no.dup$pitch,faa.no.dup$distance, xlab = 'pitch', ylab = 'distance')
```

----------X----------

ii. Correlation Analysis-

Correlation with distance -

```
   duration     no_pasg speed_ground speed_air    height      pitch
-0.05525478 -0.01310574    0.8676448 0.9452968 0.1111992 0.06945846
```

As we can see from above table that the correlation between distance and speed_ground is high hence it can be taken to be one of the predictors.

Considering $\alpha = 0.05$, p-value for pitch and height variables are also close to α, we will add these variables in a forward selection approach and determine if they improve the model.

R code:

----------X----------

```
faa.temp2 <- faa.no.dup[,-1]
```

```
corr3 <- cor(faa.no.dup$distance,faa.temp2[,-7],use = "pairwise.complete.obs")

corr3
```

----------X----------

C - Modeling

i. Regression analysis with speed_ground as the only predictor

```
Call:
lm(formula = distance ~ speed_ground, data = faa.no.dup)

Residuals:
    Min      1Q  Median      3Q     Max
-912.09 -317.00  -77.63  212.19 2369.66

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1802.3554    70.6491  -25.51   <2e-16 ***
speed_ground    42.0216     0.8622   48.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.7 on 780 degrees of freedom
Multiple R-squared:  0.7528,    Adjusted R-squared:  0.7525
F-statistic:  2375 on 1 and 780 DF,  p-value: < 2.2e-16
```

Observations:

a. Equation for the estimated linear regression is :

$\hat{y} = 42.02x - 1802.355$,where y = distance , x = speed_ground

b. R-square value = 0.7528

c. Adjusted R- Square value = 0.7525

R Code:

----------X----------

```
fit <- lm(distance ~speed_ground, data = faa.no.dup)

summary(fit)
```

----------X----------

ii. Regression analysis with speed_ground and pitch as predictors

```
Call:
lm(formula = distance ~ speed_ground + pitch, data = faa.no.dup)

Residuals:
    Min      1Q  Median      3Q     Max
-781.37 -297.46  -83.59  207.00 2322.99

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2625.9640   143.7751 -18.264  < 2e-16 ***
speed_ground    42.2973     0.8412  50.285  < 2e-16 ***
pitch          199.6933    30.6045   6.525 1.22e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 445.9 on 779 degrees of freedom
Multiple R-squared:  0.7656,    Adjusted R-squared:  0.765
F-statistic:  1272 on 2 and 779 DF,  p-value: < 2.2e-16
```

Observations:

a. Equation for the estimated linear regression is :

$\hat{y} = 42.2973x_1 + 199.63x_2 - 2625.964$,where y = distance , x_1 = speed_ground, x_2 = pitch

b. R-square value = 0.7656

c. Adjusted R- square value = 0.765

We can conclude that there is an improvement in the model prediction when pitch is included in the predictor variables as there is an increase in the adjusted R-square value

R Code:

----------X----------

```
fit2 <- lm(distance ~speed_ground + pitch, data = faa.no.dup)

summary(fit2)
```

----------X----------

iii. Regression analysis with speed_ground, pitch and height as predictors

```
Call:
lm(formula = distance ~ speed_ground + pitch + height, data = faa.no.dup)

Residuals:
    Min      1Q  Median      3Q     Max
-749.78 -304.05  -85.78  172.09 2113.69

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3041.6031   144.7313 -21.016  < 2e-16 ***
speed_ground    42.6136     0.8025  53.103  < 2e-16 ***
pitch          191.0995    29.1843   6.548 1.06e-10 ***
height          13.9446     1.5630   8.922  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 425 on 778 degrees of freedom
Multiple R-squared:  0.7874,    Adjusted R-squared:  0.7866
F-statistic: 960.3 on 3 and 778 DF,  p-value: < 2.2e-16
```

Observations:

a. Equation for the estimated linear regression is :

$$\hat{y} = 42.358x_1 + 191.0995x_2 + 13.9446x_3 - 3041.603$$

where y = distance , x_1 = speed_ground, x_2 = pitch , x_3 = height

b. R-square value = 0.7874

c. Adjusted R-square value = 0.7866

Conclusion: This is a again an improvement in the adjusted R-square value on addition of the Hecne, we can conclude that this model using speed_ground, pitch and height as the predictors is the best model of all the three.

R Code:

----------X----------

```
fit3 <- lm(distance ~speed_ground + pitch + height, data = faa.no.dup)

summary(fit3)
```

----------X----------

****The below analysis is not part of the conclusion of this report and should be seen only as an potential improvement in the model if missing values for the speed_air column are provided. Since ~80% values are missing from the dataset, this is not included in the final model****

Regression analysis with speed_ground, speed_air, height and pitch as predictors-

(For carrying out this analysis, speed_air values were imputed using regression from speed_ground values since there was high correlation among this variables)

The REG Procedure
Model: MODEL1
Dependent Variable: distance distance

Number of Observations Read	782
Number of Observations Used	782

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	553439353	138359838	999.74	<.0001
Error	777	107533727	138396		
Corrected Total	781	660973080			

Root MSE	372.01619	R-Square	0.8373
Dependent Mean	1547.30208	Adj R-Sq	0.8365
Coeff Var	24.04289		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-7199.12435	297.52399	-24.20	<.0001
speed_ground	speed_ground	1	39.61536	0.72872	54.36	<.0001
speed_air	speed_air	1	42.15848	2.72984	15.44	<.0001
height	height	1	14.29749	1.36825	10.45	<.0001
pitch	pitch	1	194.85559	25.54569	7.63	<.0001

Observations:

a. Equation for the estimated linear regression is:

$$\hat{y} = 39.61x_1 + 42.158x_2 + 14.29x_3 + 194.855x_4 - 7199.12$$

where y = distance, x_1 = speed_ground, x_2 = speed_air, x_3 = height, x_4 = pitch

b. R-square value = 0.8373 and RMSE = 372.06179

This is a significant improvement over all the models discussed above(there is a significant change in the intercept term from the above models).

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

Ans. 782 flights were used for the final model.

All 950 flights were not used because -

a. 100 records were duplicate

b. 50 records had missing values for duration column and while data cleaning we found that removing these records do not cause significant changes in the linear relationships of the variables

c. 18 records were abnormal values. We found that there are 10 records where height < 6m, 5 records where duration of flight < 40 mins and 3 records for which speed_ground < 30mph or <140

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

Ans. In the modeling section, we found that speed_ground, height and pitch are the three factors which impact the landing distance of a flight.

The final regression equation that we derived was-

$$\hat{y} = 42.358x_1 + 191.0995x_2 + 13.9446x_3 - 3041.603$$

where y = distance , x_1 = speed_ground, x_2 = pitch , x_3 = height

Thus, for every increase of 1 unit in speed_ground, average landing distance increases by 42.358 units keeping all other variables constant

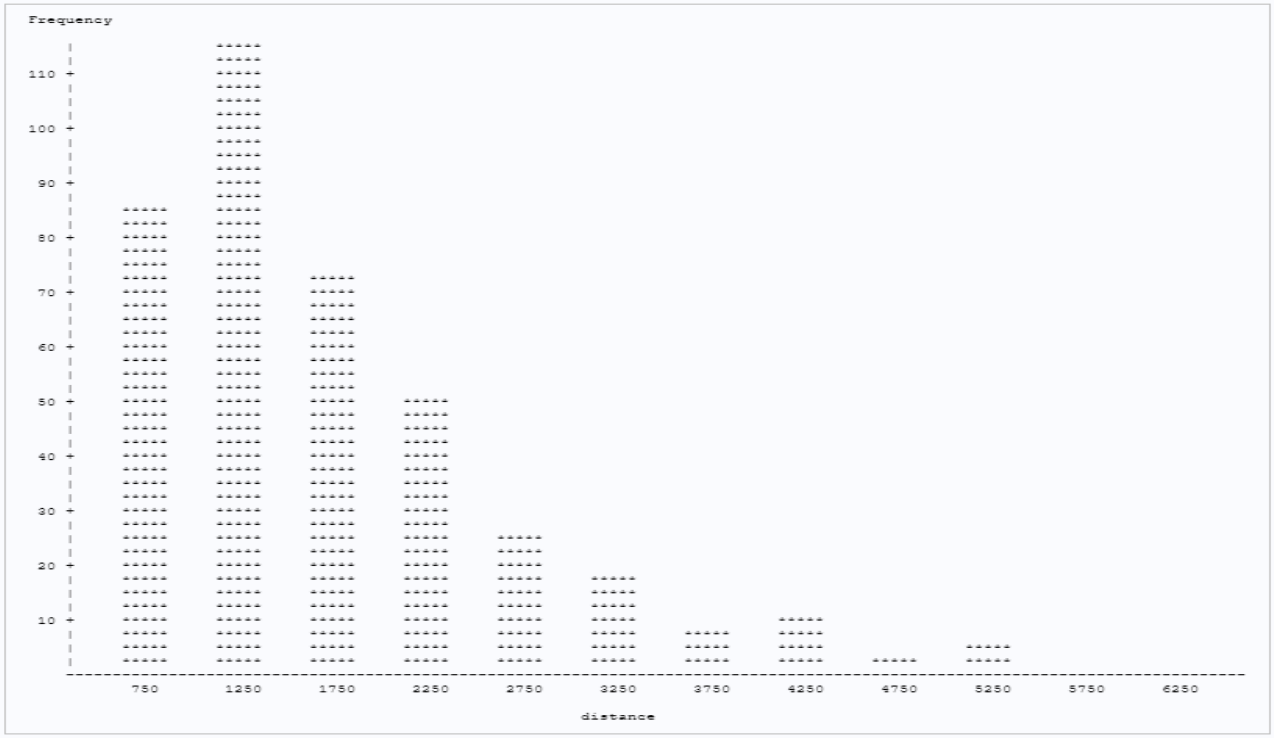
For every increase of 1 unit in pitch, average landing distance increases by 191 units keeping all other variables constant

for every increase of 1 unit in height, average landing distance increases by 14.28 units keeping all other variables constant

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

There is not much significant difference in the Boeing and Airbus planes although from the below figure we can see that, the minimum landing distance for airbus planes appears to be 400 feet while for Boeing planes it is 750 feet. Also, the maximum landing distance for Boeing planes goes till 5250 feet while for airbus it is 4800 feet.

Histogram of Landing distance for Boeing planes



Histogram of Landing distance for Airbus planes