Analysis of Safe Landing Distance for Flights

Stat Computing (BANA 6043) – Final Project

Author: Debasmita Basak (UCID: M08606277)

Summary

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

Process followed:
We are provided with a dataset by FAA, containing 8 variables – *aircraft*, *no_pasg*, *duration*, *height*, *pitch*, *distance*, *speed_ground* and *speed_air*. In order to study the impact of each of these variables on the landing distance of the flights, we follow the CRISP-DM methodology. We start by understanding business requirement (the goal), studying the data sets and moving onto preparing the data by removing abnormalities. After Data Cleansing, we end up with 781 good rows for our future analysis. Thereafter, we proceed with studying the interactions between these 8 variables. We find out, that *speed_air* can be ignored since it shows high correlation with *speed_ground*. So for our modeling process we proceed with 7 variables. By running Model fitting we find out the best model having the highest Adj R-Square value. We finally reach the conclusion that *speed_ground*, $(speed\_ground)^2$, *height*, *no_pasg* and *aircraft* are the only factors impacting the landing distance of the flights. After we finalize the model we proceed to validate our assumptions about residuals (or error).

Datasets:

FAA1.xls

Chapter 1 – Data Preparation

```
/*Reading Dataset */
FAA <- read.csv("C:/Users/debas/Documents/SASLabs/Labs/FAA1.csv",header=TRUE)
FAA
> nrow(FAA)
[1] 800

/* Produce a summary of the data set*/
summary(FAA)

> summary(FAA)
   aircraft      duration         no_pasg       speed_ground      speed_air
 airbus:400   Min.   : 14.76   Min.   :29.00   Min.   : 27.74   Min.   : 90.00
 boeing:400   1st Qu.:119.49   1st Qu.:55.00   1st Qu.: 65.87   1st Qu.: 96.16
              Median :153.95   Median :60.00   Median : 79.64   Median :100.99
              Mean   :154.01   Mean   :60.13   Mean   : 79.54   Mean   :103.83
              3rd Qu.:188.91   3rd Qu.:65.00   3rd Qu.: 92.33   3rd Qu.:109.48
              Max.   :305.62   Max.   :87.00   Max.   :141.22   Max.   :141.72
                                                                NA's   :600
     height           pitch           distance
 Min.   :-3.546   Min.   :2.284   Min.   :  34.08
 1st Qu.:23.338   1st Qu.:3.658   1st Qu.: 900.95
 Median :30.147   Median :4.020   Median :1267.44
 Mean   :30.122   Mean   :4.018   Mean   :1544.52
 3rd Qu.:36.981   3rd Qu.:4.388   3rd Qu.:1960.44
 Max.   :59.946   Max.   :5.927   Max.   :6533.05
```

Figure 1: Summary of FAA data

Observation: We can see that **speed_air** has 600 missing values (75%); excluding these rows would mean loss of relevant data from other variables. Hence, we will ignore **speed_air** NULL values during our data cleaning process.

/* Data Cleansing*/

```
> data<-subset(FAA, duration>40)
> data<-subset(data,speed_ground>30 & speed_ground<140)
> data<-subset(data, height>6)
> data<-subset(data, distance<6000)
> summary(data)
   aircraft      duration         no_pasg       speed_ground      speed_air
 airbus:394   Min.   : 41.95   Min.   :29.00   Min.   : 33.57   Min.   : 90.00
 boeing:387   1st Qu.:119.63   1st Qu.:55.00   1st Qu.: 66.19   1st Qu.: 96.15
              Median :154.28   Median :60.00   Median : 79.79   Median :100.89
              Mean   :154.78   Mean   :60.08   Mean   : 79.64   Mean   :103.50
              3rd Qu.:189.66   3rd Qu.:65.00   3rd Qu.: 92.13   3rd Qu.:109.42
              Max.   :305.62   Max.   :87.00   Max.   :132.78   Max.   :132.91
                                                                NA's   :586
     height           pitch          distance
 Min.   : 6.228   Min.   :2.284   Min.   :  41.72
 1st Qu.:23.594   1st Qu.:3.653   1st Qu.: 919.05
 Median :30.217   Median :4.014   Median :1273.66
 Mean   :30.455   Mean   :4.014   Mean   :1541.20
 3rd Qu.:36.988   3rd Qu.:4.382   3rd Qu.:1960.43
 Max.   :59.946   Max.   :5.927   Max.   :5381.96
```

/* Check record count after cleansing*/

```
> nrow(data)
[1] 781
>
```

Chapter 2 – Data Exploration

```
#Exploratory Analysis using visualization
>  par(mfrow=c(2,3))
>  boxplot(duration~aircraft, data=data)
>  boxplot(no_pasg~aircraft, data=data)
>  boxplot(speed_ground~aircraft, data=data)
>  boxplot(speed_air~aircraft, data=data)
>  boxplot(pitch~aircraft, data=data)
>  boxplot(distance~aircraft, data=data)
```

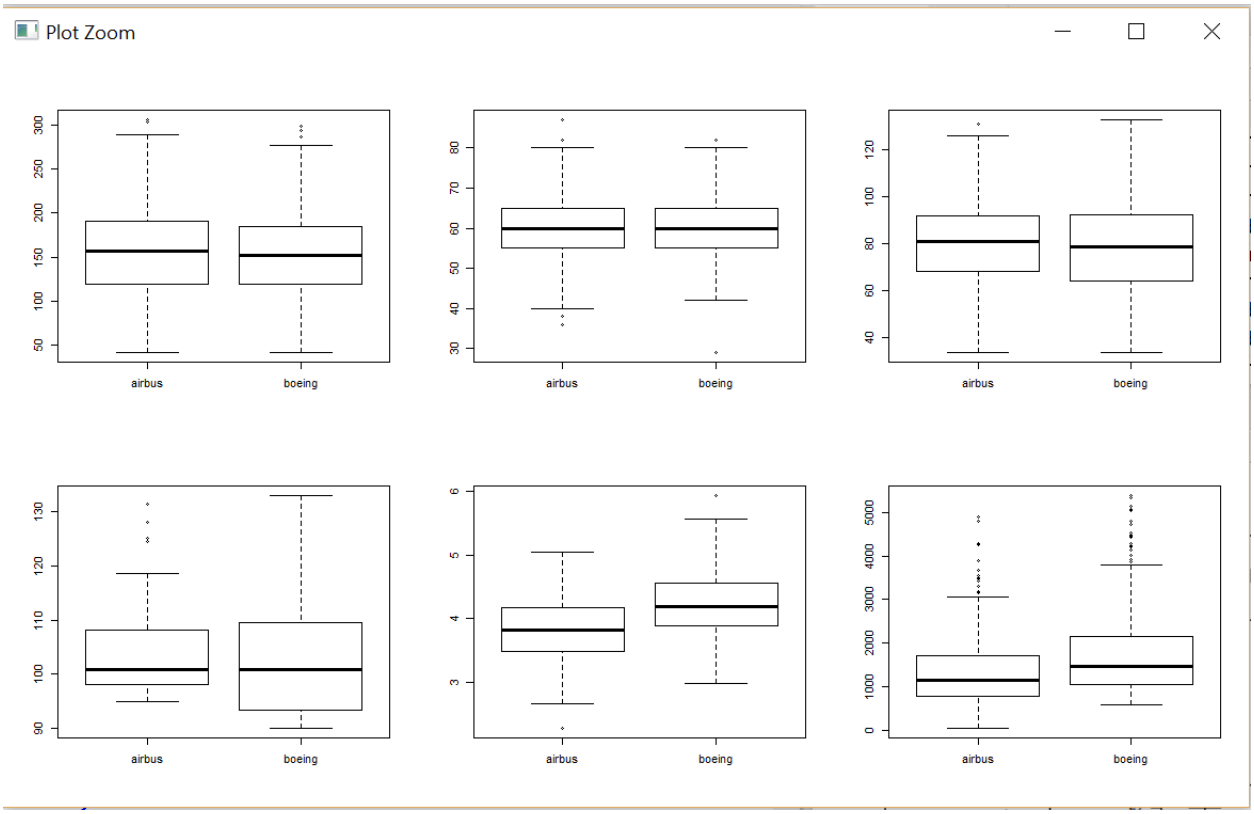
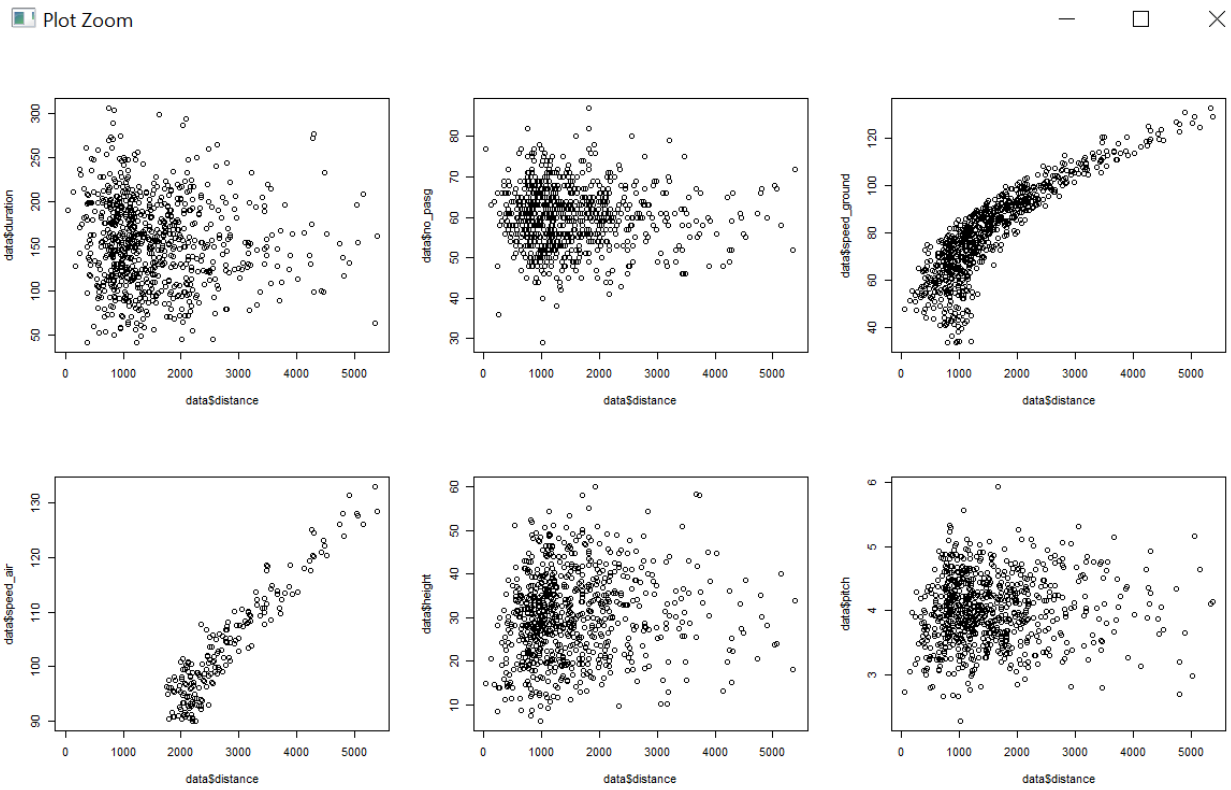


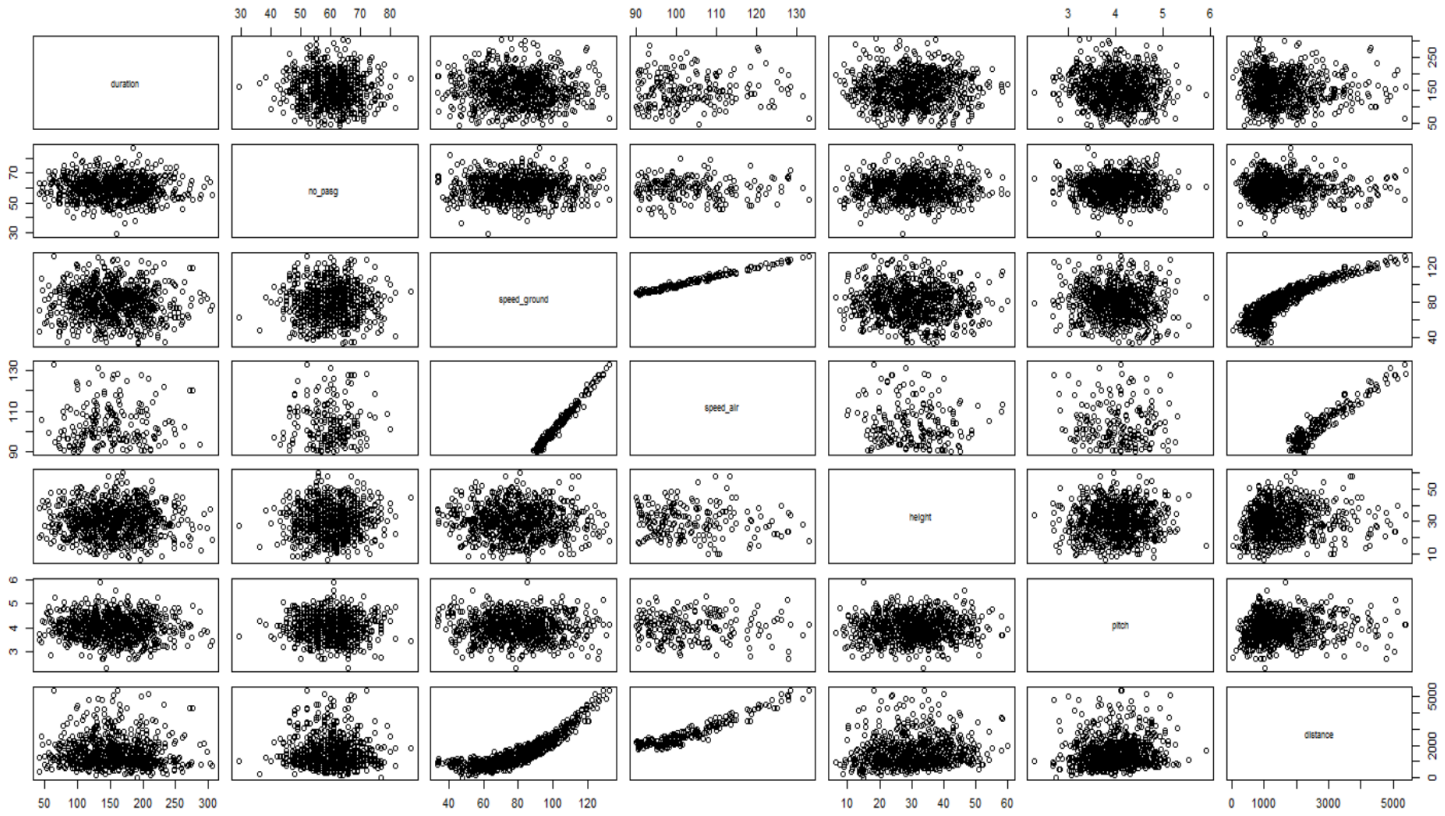
Figure 2: Exploratory Analysis – Box Plot

```
#To check relationship between independent and dependent variables
>  par(mfrow=c(2,3))
>  plot(data$distance,data$duration)
>  plot(data$distance,data$no_pasg)
>  plot(data$distance,data$speed_ground)
>  plot(data$distance,data$speed_air)
>  plot(data$distance,data$height)
>  plot(data$distance,data$pitch)
>  data_measure<-subset(data, select=-c(aircraft))
```



Observation: We can see that distance*speed_ground is an upward curve; Hence we will convert speed_ground to speed_ground^2

```
# To check correlation between independent variables

>  pairs(data_measure)
>  round(cor(data_measure,use="complete.obs"),2)
```



	duration	no_pasg	speed_ground	speed_air	height	pitch	distance
duration	1.00	-0.07	0.02	0.04	0.07	-0.06	0.05
no_pasg	-0.07	1.00	0.00	0.00	-0.01	-0.04	-0.03
speed_ground	0.02	0.00	1.00	0.99	-0.10	-0.06	0.93
speed_air	0.04	0.00	0.99	1.00	-0.09	-0.05	0.94
height	0.07	-0.01	-0.10	-0.09	1.00	-0.03	0.06
pitch	-0.06	-0.04	-0.06	-0.05	-0.03	1.00	0.03
distance	0.05	-0.03	0.93	0.94	0.06	0.03	1.00

Figure 3: Correlation between independent variables

Observation: We can see that **speed_ground** and **speed_air** are highly correlated (**0.99**). Hence, to avoid discrepancy due to multi-collinearity, we will drop speed_air from our future analysis and modeling procedures.

Chapter 3 – Modeling

#Data Wrangling:

Adding *speed_g* to the existing dataset *data* which will contain *(speed_ground)^2*

```
> data$speed_g=data$speed_ground*data$speed_ground
>  data
    aircraft  duration no_pasg speed_ground speed_air    height    pitch   distance   speed_g
1     boeing  98.47909      53    107.91568 109.32838 27.418924 4.043515 3369.83636 11645.794
2     boeing 125.73330      69    101.65559 102.85141 27.804716 4.117432 2987.80392 10333.859
3     boeing 112.01700      61     71.05196        NA 18.589386 4.434043 1144.92243  5048.381
4     boeing 196.82569      56     85.81333        NA 30.744597 3.884236 1664.21816  7363.927
5     boeing  90.09538      70     59.88853        NA 32.397688 4.026096 1050.26450  3586.636
6     boeing 137.59582      55     75.01434        NA 41.214963 4.203853 1627.06820  5627.152
7     boeing  73.02379      54     54.42980        NA 24.035322 3.837646  805.30399  2962.603
8     boeing  52.90319      57     57.10166        NA 19.388838 4.643672  573.62179  3260.600
9     boeing 155.51862      61     85.44362        NA 35.375390 4.228728 1698.99275  7300.613
10    boeing 176.86203      56     61.79671        NA 36.748816 4.184399 1137.74576  3818.833
```

#Model fitting

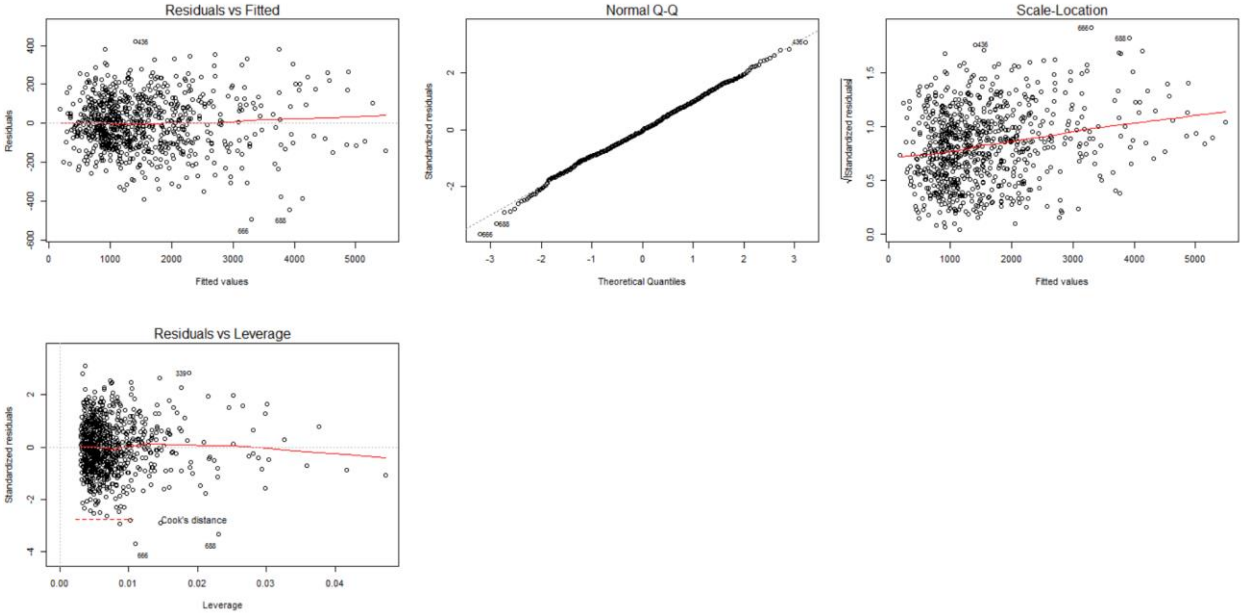
```
> model<-lm(distance~aircraft+no_pasg+speed_ground+speed_g+height-1, data=data)
> summary(model)

Call:
lm(formula = distance ~ aircraft + no_pasg + speed_ground + speed_g +
    height - 1, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-496.25  -91.79   -5.91   91.88  419.23

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
aircraftairbus 1869.23035   79.89033  23.397   <2e-16 ***
aircraftboeing 2271.33212   78.47745  28.942   <2e-16 ***
no_pasg          -1.59746    0.64485  -2.477   0.0135 *
speed_ground    -68.81084    1.69265 -40.653   <2e-16 ***
speed_g           0.69110    0.01038  66.560   <2e-16 ***
height           13.75862    0.49893  27.576   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 135.4 on 775 degrees of freedom
Multiple R-squared:  0.9943,    Adjusted R-squared:  0.9943
F-statistic: 2.254e+04 on 6 and 775 DF,  p-value: < 2.2e-16
```

Figure 7. Summary of Model

Model Diagnostic

```
> coefficients(model)
aircraftairbus aircraftboeing        no_pasg   speed_ground        speed_g         height
  1869.2303480   2271.3321197     -1.5974577    -68.8108429      0.6910986     13.7586205
```

```
> residuals<-residuals(model) # residuals
> plot(model)
```



Using the estimates returned by lm(), the equation for the Landing Distance comes down to:

$D_{airbus} = 1869.2304 - 1.5975*no\_pasg + 13.7586*height + 0.6911*speed\_ground^2 - 68.8108*speed\_ground$

$D_{boeing} = 2271.3321 - 1.5975*no\_pasg + 13.7586*height + 0.6911*speed\_ground^2 - 68.8108*speed\_ground$