Analysis of Flight Landings

Introduction

In the following project, a flight landings dataset named as FAA1 is being imported and analyzed to study the impact of various factors on the distance of flight landings. The report is divided into three chapters namely:

- Data Importing and Cleaning
- Exploratory Data Analysis
- Statistical Modeling

All the results and code are mentioned in the chapters along with description of each step we are carrying out.

1. Data Importing and Cleaning

Loading required libraries

```
library(readxl)
library(dplyr)
library(ggplot2)
```

Reading the dataset

```
FAA1 <- read_xls('FAA1.xls')
```

Basic summary of the data

```
str(FAA1)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    800 obs. of  8 variables:
##  $ aircraft    : chr  "boeing" "boeing" "boeing" "boeing" ...
##  $ duration    : num  98.5 125.7 112 196.8 90.1 ...
##  $ no_pasg     : num  53 69 61 56 70 55 54 57 61 56 ...
##  $ speed_ground: num  107.9 101.7 71.1 85.8 59.9 ...
##  $ speed_air   : num  109 103 NA NA NA ...
##  $ height      : num  27.4 27.8 18.6 30.7 32.4 ...
##  $ pitch       : num  4.04 4.12 4.43 3.88 4.03 ...
##  $ distance    : num  3370 2988 1145 1664 1050 ...
```

One interesting observation from the dataset is that R considers column ‘Aircraft’ as charcater instead of factor. So we have to convert Aircraft variable to factor.

```
FAA1$aircraft <- as.factor(FAA1$aircraft)
```

Checking the class of Aircraft now

```
class(FAA1$aircraft)
```

```
## [1] "factor"
```

Checking the duplicates in our dataset

```
sum(duplicated.data.frame(FAA1))
```

```
## [1] 0
```

Through above code we can see that there are no duplicates in our dataset that need to be removed.

Now finding Number of Missing values for each data

```
sapply(FAA1, function(x) sum(is.na(x)))
```

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

Number of Missing Values for Speed_Air : 600

Percent of Missing Values

```
nmiss_table <- matrix(c(sum(is.na(FAA1$speed_air)),nrow(FAA1),(sum(is.na(FAA1$speed_air))/nrow(F
AA1))*100),
                ncol=3,byrow=TRUE)
colnames(nmiss_table) <- c("Missing","Total","Percent")
rownames(nmiss_table) <- c("Speed_Air")
nmiss_table <- as.table(nmiss_table)
nmiss_table
```

```
##           Missing Total Percent
## Speed_Air     600   800      75
```

Since almost 75 percent data is missing for Speed_Air column, we cannot remove those observations as it will result in loss of important information.

Checking for Abnomral Values in our dataset

```
abn_duration <- FAA1$duration < 40
abn_speedAir <- (FAA1$speed_air < 30 | FAA1$speed_air > 140)& !is.na(FAA1$speed_air)
abn_speedGround <- FAA1$speed_ground < 30 | FAA1$speed_ground > 140
abn_height <- FAA1$height < 6
abn_distance <- FAA1$distance > 6000



countabn_duration <- sum(abn_duration)
countabn_speedAir <- sum(abn_speedAir)
countabn_speedGround <- sum(abn_speedGround)
countabn_height <- sum(abn_height)
countabn_distance <- sum(abn_distance)

abn_table <- matrix(c(countabn_duration,countabn_speedGround,countabn_speedAir,countabn_height,c
ountabn_distance),
                ncol=1,byrow=TRUE)
colnames(abn_table) <- c("Number of Abnormal Values")
rownames(abn_table) <- c('Duration','Speed_Ground','Speed_Air','Height','Distance')
abn_table <- as.table(abn_table)
abn_table
```

```
##              Number of Abnormal Values
## Duration                             5
## Speed_Ground                         3
## Speed_Air                            1
## Height                              10
## Distance                             2
```

Removing the Abnormal Values

```
FAA_Cleaned <-FAA1[!(abn_duration | abn_speedAir | abn_speedGround | abn_height |
                    abn_distance),]
```

Our finally cleaned dataset contains **781 observations and 8 columns**. Lets have a look at its structure and first few observations.

```
str(FAA_Cleaned)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    781 obs. of  8 variables:
##  $ aircraft    : Factor w/ 2 levels "airbus","boeing": 2 2 2 2 2 2 2 2 2 2 ...
##  $ duration    : num  98.5 125.7 112 196.8 90.1 ...
##  $ no_pasg     : num  53 69 61 56 70 55 54 57 61 56 ...
##  $ speed_ground: num  107.9 101.7 71.1 85.8 59.9 ...
##  $ speed_air   : num  109 103 NA NA NA ...
##  $ height      : num  27.4 27.8 18.6 30.7 32.4 ...
##  $ pitch       : num  4.04 4.12 4.43 3.88 4.03 ...
##  $ distance    : num  3370 2988 1145 1664 1050 ...
```

```
head(FAA_Cleaned)
```

aircraft <fctr>	duration <dbl>	no_pasg <dbl>	speed_ground <dbl>	speed_air <dbl>	height <dbl>	pitch <dbl>	distance <dbl>
boeing	98.47909	53	107.91568	109.3284	27.41892	4.043515	3369.836
boeing	125.73330	69	101.65559	102.8514	27.80472	4.117432	2987.804
boeing	112.01700	61	71.05196	NA	18.58939	4.434043	1144.922
boeing	196.82569	56	85.81333	NA	30.74460	3.884236	1664.218
boeing	90.09538	70	59.88853	NA	32.39769	4.026096	1050.264
boeing	137.59582	55	75.01434	NA	41.21496	4.203853	1627.068

6 rows

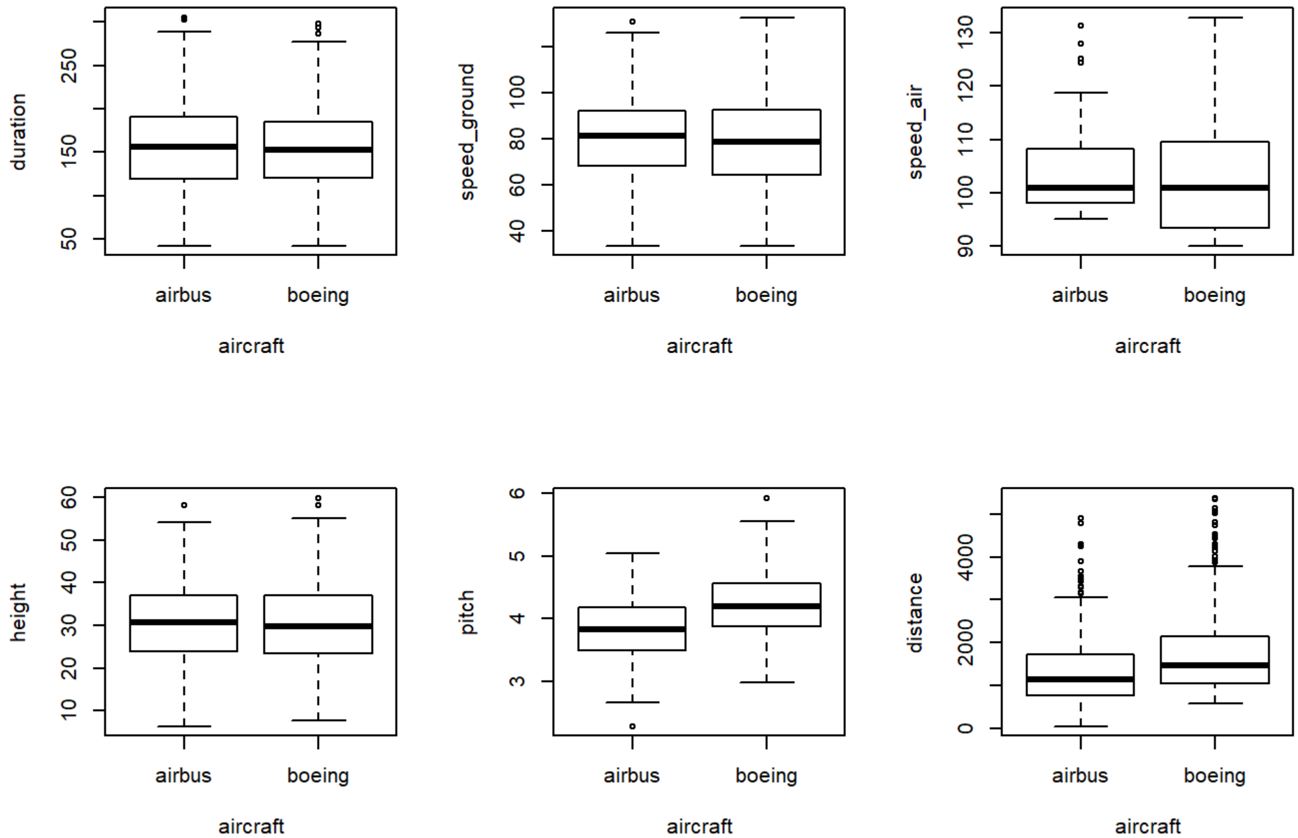
2. Exploratory Data Analysis

```
summary(FAA_Cleaned)
```

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

```
attach(FAA_Cleaned)
```

```
par(mfrow = c(2,3))
boxplot(duration ~ aircraft)
boxplot(speed_ground ~ aircraft)
boxplot(speed_air ~ aircraft)
boxplot(height ~ aircraft)
boxplot(pitch ~ aircraft)
boxplot(distance ~ aircraft)
```

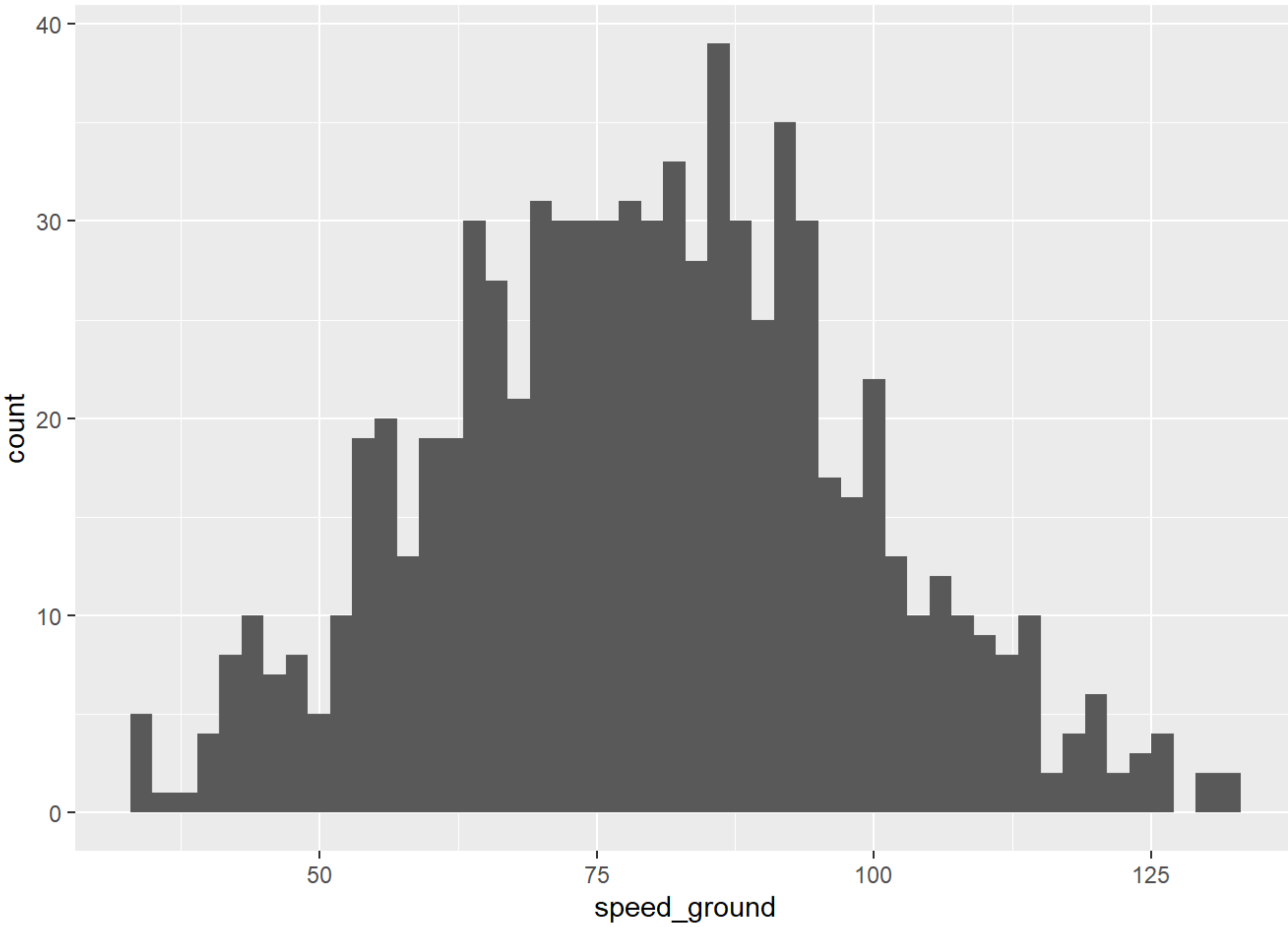


Key Inferences from Boxplot :

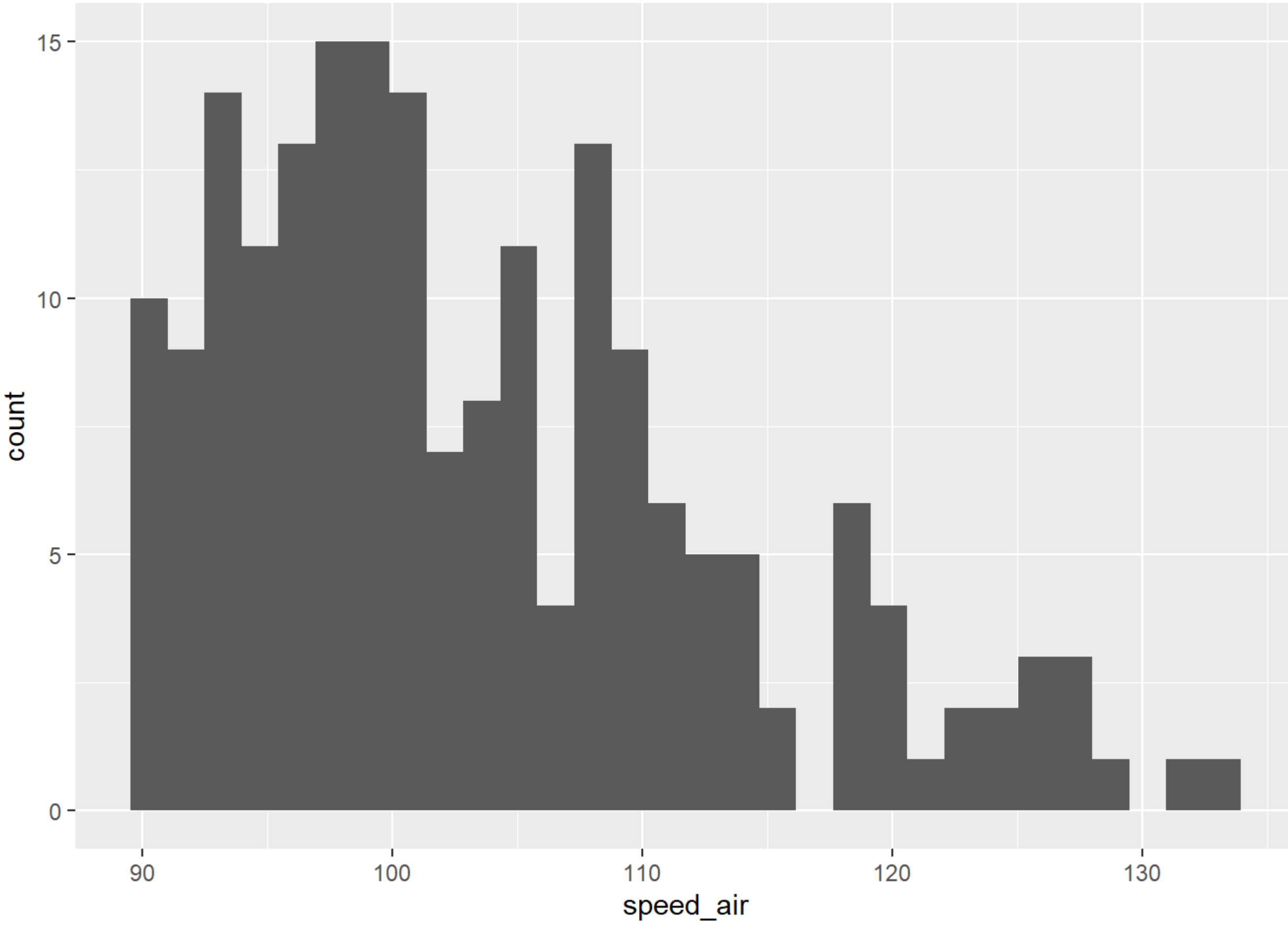
- We can see that certain variables like ‘speed_air’,‘pitch’ and ‘distance’ have different distribution for each type of aircraft.
- There are a number of outliers in ‘distance’ variable for each type of aircraft whereas for ‘speed_air’ outliers exist only for airbus.

Distribution of each Variable

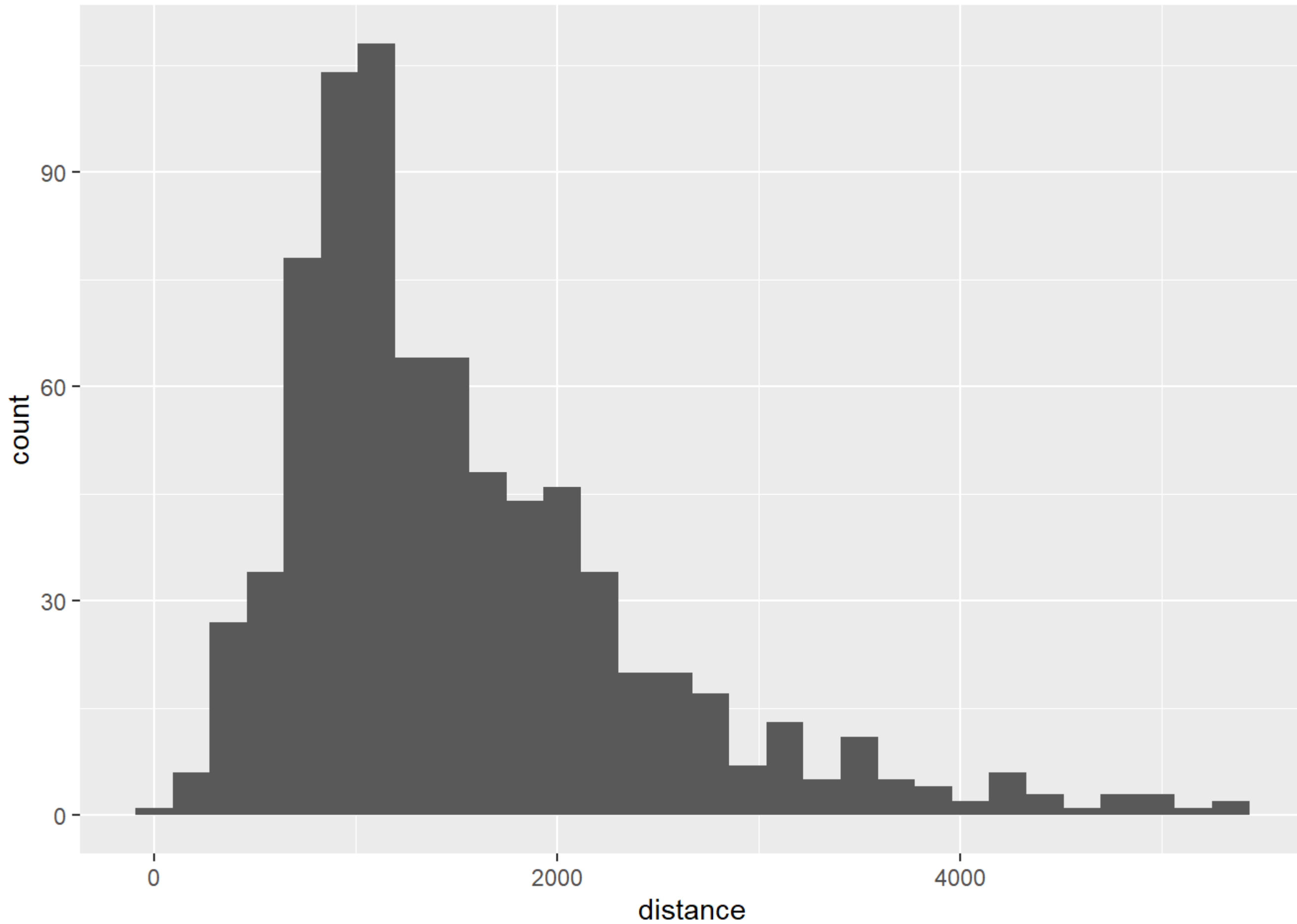
```
par(mfrow = c(2,3))
p <- ggplot(FAA_Cleaned)
p + geom_histogram(aes(x=speed_ground),binwidth = 2)
```



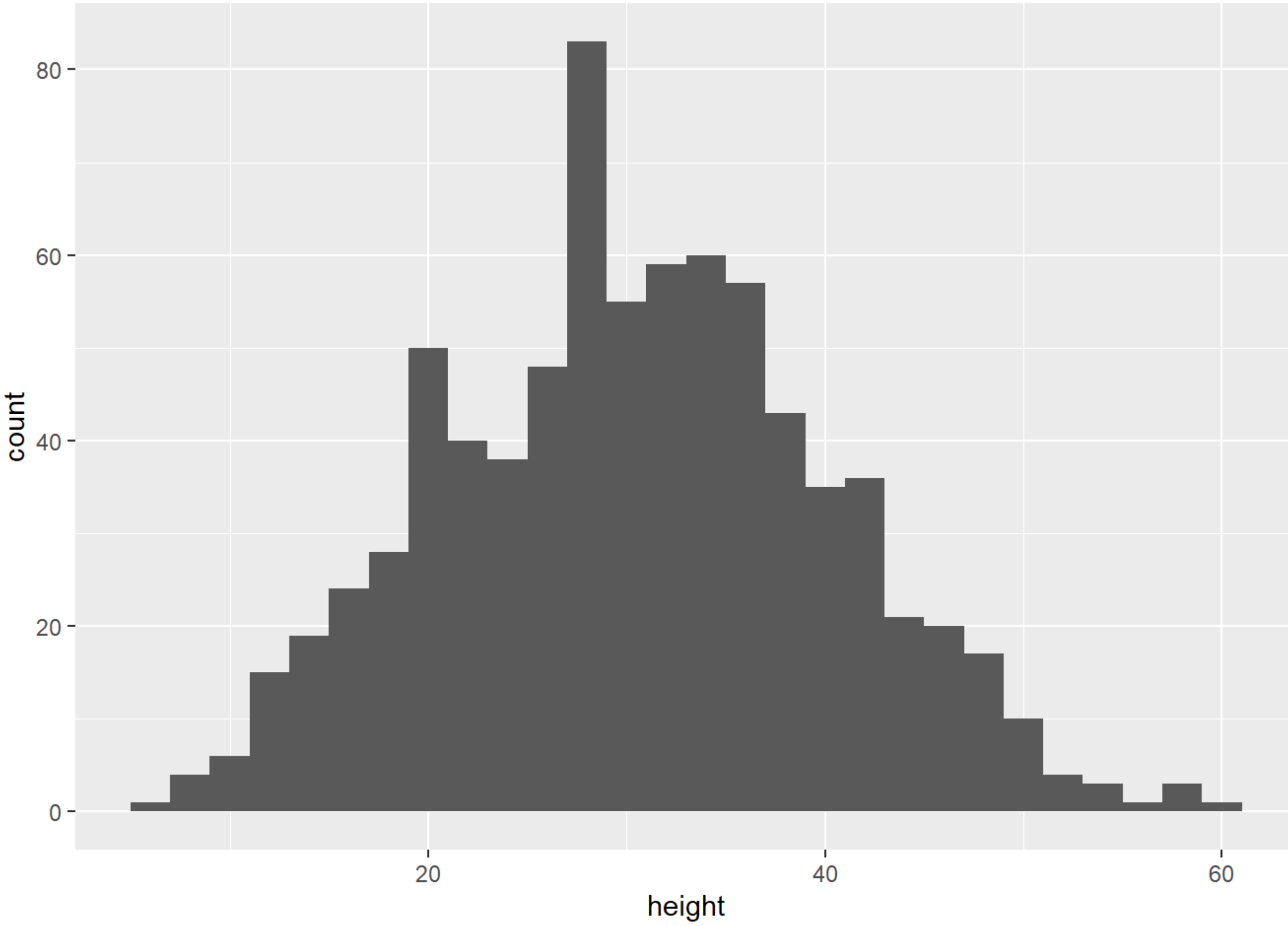

```
p + geom_histogram(aes(x=speed_air))
```




```
p + geom_histogram(aes(x=distance))
```

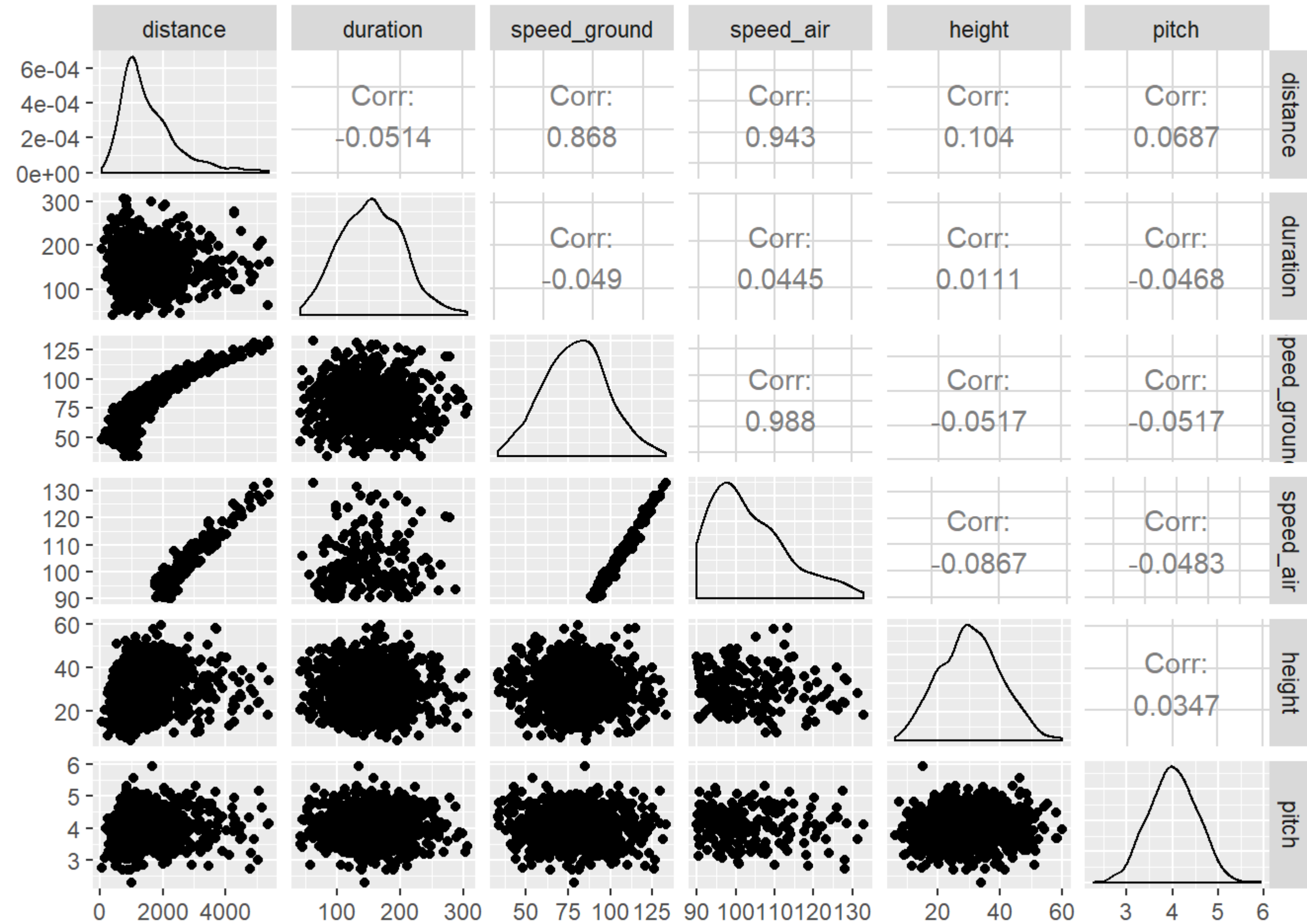



```
p + geom_histogram(aes(x=height),binwidth = 2)
```



Before fitting any model, let us explore the relationship between our response variable ‘distance’ and other predictor variables.

```
GGally :: ggpairs(data = FAA_Cleaned[,c(8,2,4,5,6,7)])
```



Looking at the above graphs, the relationship between predictor variable and some response variables like speed_air and speed_ground seem to be clear. Similar cannot be said about other predictor variables. Coming to the correlation values, the correlation of distance is high with speed_air and speed_ground and also with height although not as high as the other two.

However, before jumping to any conclusions we have to check whether any two predictor variables are correlated among themselves resulting in Multicollinearity. We see a very high value of 0.988 betwen speed_air and speed_ground and it may create issues with our model if don't consider this aspect.

3. Statistical Modeling

Including Aircraft type as dummy variable

After performing the above steps, we will try to fit the model that best explains our data. Generalized form of linear regression is: **$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3$** Where

- $\beta 0$ = Intercept
- $\beta 1$ = Coefficient of Speed_Air, X1 = Speed_Air
- $\beta 2$ = Coefficient of Height, X2= Height
- $\beta 3$ = Coefficient of Aircraft Type, X3= Aircraft

```
model1 <- lm(distance ~ speed_air + height + aircraft)
summary(model1)
```

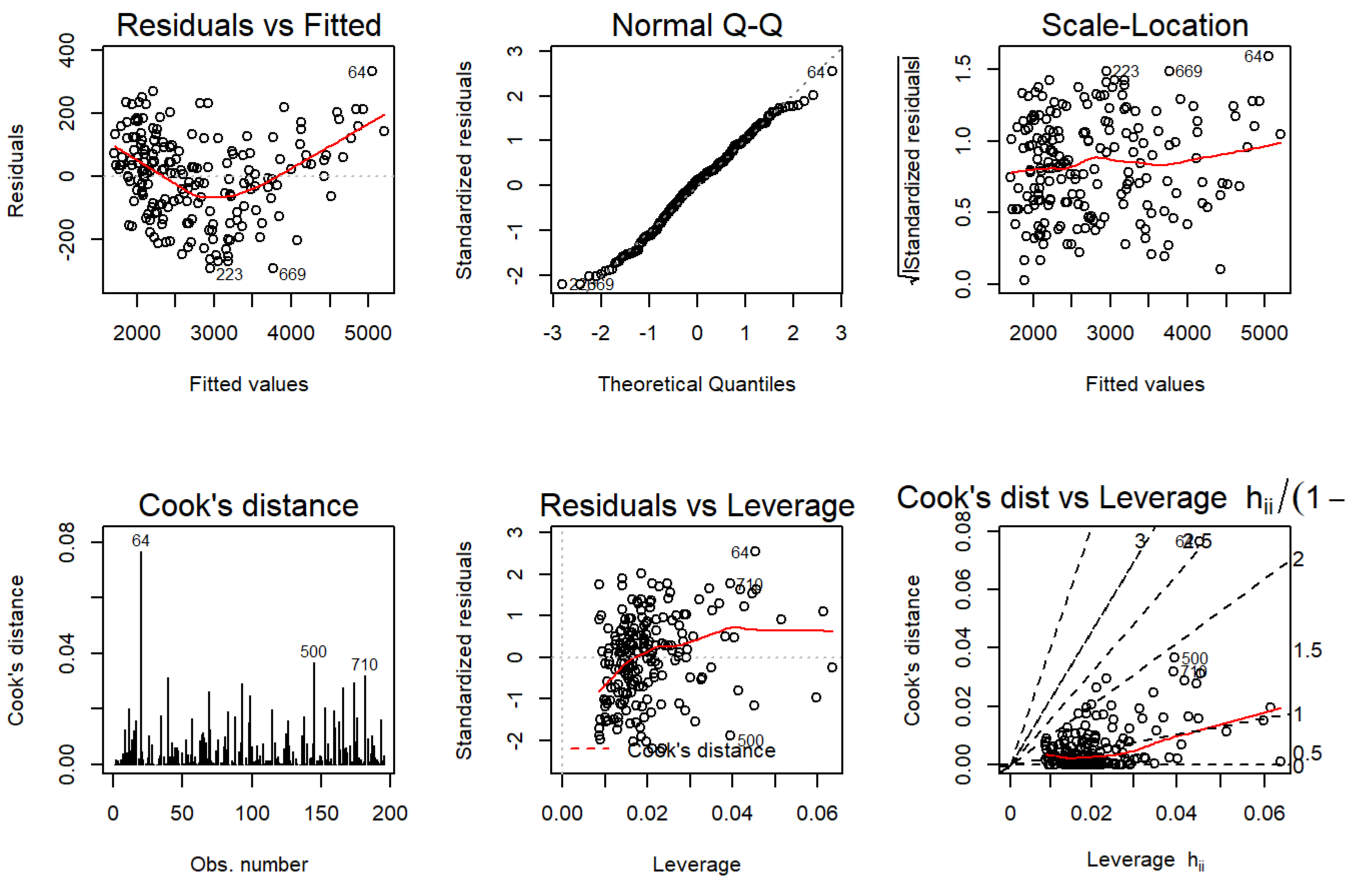
```
##
## Call:
## lm(formula = distance ~ speed_air + height + aircraft)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -293.22  -93.83   15.35   90.05  332.84
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6388.1241   111.3135  -57.39   <2e-16 ***
## speed_air         82.0393     0.9827   83.48   <2e-16 ***
## height            13.7913     1.0324   13.36   <2e-16 ***
## aircraftboeing   433.7406    19.7657   21.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 134.3 on 191 degrees of freedom
##   (586 observations deleted due to missingness)
## Multiple R-squared:  0.9742, Adjusted R-squared:  0.9738
## F-statistic:  2408 on 3 and 191 DF,  p-value: < 2.2e-16
```

Looking at the above summary, the Adjusted R-squared comes as pretty good with a high value of 0.9738. All the coeffcients are significant with very less p-values. The coefficient values are: $\beta0$: -6388.1241 $\beta1$: 82.0393 $\beta2$: 13.7913 $\beta3$: 433.7406 (Coefficient for Boeing Aircraft)

The final equation comes as: **Y = -6388.12 + 82.04X1 + 13.79X2 + 433.74X3 (X3 = 1 for Boeing,0 for Airbus)**

Residual Plots Analysis

```
par(mfrow = c(2,3))
plot(model1,which = 1:6)
```



Residual plot analysis shows that normality of resiual is followed. There is however slight indication of non-linear structure by looking at the first plot.

Fitting the Model for each Aircraft type

Subsetting the data

```
FAA1_boeing <- filter(FAA_Cleaned,aircraft == 'boeing')
FAA1_airbus <- filter(FAA_Cleaned,aircraft == 'airbus')
```

Fitting separate models for each dataset and analysing the model parameters

```
model_boeing <- lm(distance ~ speed_air + height,FAA1_boeing)
summary(model_boeing)
```

```
##
## Call:
## lm(formula = distance ~ speed_air + height, data = FAA1_boeing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -290.08  -97.83   11.19   99.75  333.11
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5956.139    134.839  -44.17   <2e-16 ***
## speed_air      81.996      1.189   68.98   <2e-16 ***
## height         13.997      1.393   10.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 137.2 on 115 degrees of freedom
##   (269 observations deleted due to missingness)
## Multiple R-squared:  0.9764, Adjusted R-squared:  0.976
## F-statistic:  2380 on 2 and 115 DF,  p-value: < 2.2e-16
```

The model parameters are quited good with high Adjusted R-squared and all the coefficients coming as significant.

The equation comes as : **Y = -5956.14 + 81.99X1 + 13.99X2**

Now fitting the model for Airbus type:

```
model_airbus <- lm(distance ~ speed_air + height,FAA1_airbus)
summary(model_airbus)
```

```
##
## Call:
## lm(formula = distance ~ speed_air + height, data = FAA1_airbus)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -295.03  -88.49   20.85   81.00  272.46
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6396.003    196.602 -32.533  < 2e-16 ***
## speed_air      82.200      1.809  45.431  < 2e-16 ***
## height         13.503      1.549   8.716 5.72e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 131.4 on 74 degrees of freedom
##   (317 observations deleted due to missingness)
## Multiple R-squared:  0.9663, Adjusted R-squared:  0.9654
## F-statistic:  1062 on 2 and 74 DF,  p-value: < 2.2e-16
```

Here too the coefficients are almost similar as above. The equation comes as : **Y = -6396 + 82.20X1 + 13.50X2**

Conclusion

> From both the model fitting, we find the slope coefficents are almost similar which is expected. There is a positive difference of 433.706 for boeing meaning the distance for the boeing aircrafts are relatively more as compared to airbus which can also be seen in the boxplot. The model is still not quite accurate because there are a lot of missing values for speed_air and it would be better to consult the customers and ask for the missing data.