Study what factors and how they would impact the landing distance of a commercial flight

Mohammed Nifaullah Sailappai

1/18/2020

1. Introduction

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).

Variable dictionary

Aircraft: The make of an aircraft (Boeing or Airbus). Duration (in minutes): Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min. No_pasg: The number of passengers in a flight. Speed_ground (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal. Speed_air (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal. Height (in meters): The height of an aircraft when it is passing over the threshold of the runway. The landing aircraft is required to be at least 6 meters high at the threshold of the runway. Pitch (in degrees): Pitch angle of an aircraft when it is passing over the threshold of the runway. Distance (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

Library

```
library(ggplot2)
library(MASS)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Reading Data

```
faa1 <- read.csv("C:/Users/nifaullah/Downloads/FAAc1.csv")
faa2 <- read.csv("C:/Users/nifaullah/Downloads/FAAc2.csv")
```

2. Structure of the Data

FAA 1

First dataset has 800 observations and 8 variables. 7, including one count data, of the variables are numerical in nature. Aircraft is a categorical variable with 2 levels of factors, Airbus and Boeing respectively.

```
str(faa1)
```

```
## 'data.frame':    800 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     : int  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 ...
```

FAA 2

Second dataset has 200 observations but only has 7 variables, with duration being the missing variable. Aircraft again is a categorical variable, but there seems to be 3 levels factor as opposed to 2 levels in earlier case, on a closer look it seems 3rd level is actually an empty string with missing data.

```
str(faa2)
```

```
## 'data.frame':    200 obs. of  7 variables:
##  $ aircraft    : Factor w/ 3 levels "","airbus","boeing": 3 3 3 3 3 3 3 3 3 3 ...
##  $ no_pasg     : int  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 ...
```

```
# Removing empty factor data if any
faa2 <- droplevels(subset(faa2, aircraft != ""))
```

3. Merge

Merging the two datasets vertically, also removing duplicate rows and removing the data belonging to missing factor

```
# Creating the missing column duration in faa2 before merging the datasets vertically
faa2$duration <- NA
# Merging Verically
faa <- rbind(faa1, faa2)
# Selecting duplicates minus duration as durtion was not originally present in 2nd Dataset
duplicate_rows <- faa  %>%
  select(-duration) %>%
  duplicated() %>%
  which()
# Number of duplicates
length(duplicate_rows)
```

```
## [1] 100
```

```
#Removing duplicates
faa <- faa[-duplicate_rows,]
# After removing duplicates
dim(faa)
```

```
## [1] 850   8
```

4. Combined data

Structure

850 observation implies that 150 observation were either duplicates or belonged to the missing factor, which were removed in the previous operation, & 8 variables suggest that all the variables ,including the missing data from FAA2 have been merged safely.

```
str(faa)
```

```
## 'data.frame':    850 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     : int  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 ...
```

Summary Statistics

1. Boeing has slightly lower representation compared to Airbus. Speed ground and Pitch also seem fairly alright.
2. Duration has 150 missing values likely coming from the second dataset.
3. Speed Air proportionately has very high number of missing values, most likely this column will be dropped from the analysis.
4. Height has a negative value which suggests presence of bad data in the column.
5. Range for distance is quite huge.

```
summary(faa)
```

```
##    aircraft      duration         no_pasg      speed_ground      speed_air
##  airbus:450   Min.   : 14.76   Min.   :29.0   Min.   : 27.74   Min.   : 90.00
##  boeing:400   1st Qu.:119.49   1st Qu.:55.0   1st Qu.: 65.90   1st Qu.: 96.25
##               Median :153.95   Median :60.0   Median : 79.64   Median :101.15
##               Mean   :154.01   Mean   :60.1   Mean   : 79.45   Mean   :103.80
##               3rd Qu.:188.91   3rd Qu.:65.0   3rd Qu.: 92.06   3rd Qu.:109.40
##               Max.   :305.62   Max.   :87.0   Max.   :141.22   Max.   :141.72
##               NA's   :50                                       NA's   :642
##      height           pitch          distance
##  Min.   :-3.546   Min.   :2.284   Min.   :  34.08
##  1st Qu.:23.314   1st Qu.:3.642   1st Qu.: 883.79
##  Median :30.093   Median :4.008   Median :1258.09
##  Mean   :30.144   Mean   :4.009   Mean   :1526.02
##  3rd Qu.:36.993   3rd Qu.:4.377   3rd Qu.:1936.95
##  Max.   :59.946   Max.   :5.927   Max.   :6533.05
##
```

5. Data Cleansing

Data is cleaned as per the description provided in the variable dictionary above i.e. values which are considered abnormal will be removed. As seen below there're 17 rows which had abnormal values and all of them have been excluded from the new dataframe.

```
# Checking if any missing values are still present in the duration column.
any(is.na(faa$duration))
```

```
## [1] TRUE
```

```
# Imputing groupwise mean for missing values
faa <- faa %>%
  group_by(aircraft) %>%
  mutate(duration=ifelse(is.na(duration),mean(duration,na.rm=TRUE),duration))
# Checking if any missing values are still present in the duration column.
any(is.na(faa$duration))
```

```
## [1] FALSE
```

```
# Remove all abnormal rows as per the dictionary definition
faa_normal <- faa %>%
  filter(duration >40,
         (speed_ground >=30 | speed_ground <=140),
         (is.na(speed_air) | speed_air >=30 | speed_air <=140),
         height >=6, distance < 6000)
# Number of rows removed based on abnormal values
nrow(faa) - nrow(faa_normal)
```

```
## [1] 17
```

```
#Looking structure & summary after cleaning the data
str(faa_normal)
```

```
## Classes 'grouped_df', 'tbl_df', 'tbl' and 'data.frame':  833 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     : int  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 ...
##  - attr(*, "groups")=Classes 'tbl_df', 'tbl' and 'data.frame':   2 obs. of  2 variables:
##   ..$ aircraft: Factor w/ 2 levels "airbus","boeing": 1 2
##   ..$ .rows   :List of 2
##   .. ..$ : int  390 391 392 393 394 395 396 397 398 399 ...
##   .. ..$ : int  1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, ".drop")= logi TRUE
```

```
summary(faa_normal)
```

```
##    aircraft      duration         no_pasg       speed_ground
##  airbus:444   Min.   : 41.95   Min.   :29.00   Min.   : 27.74
##  boeing:389   1st Qu.:122.73   1st Qu.:55.00   1st Qu.: 66.08
##               Median :156.11   Median :60.00   Median : 79.75
##               Mean   :154.91   Mean   :60.04   Mean   : 79.42
##               3rd Qu.:186.51   3rd Qu.:65.00   3rd Qu.: 91.87
```

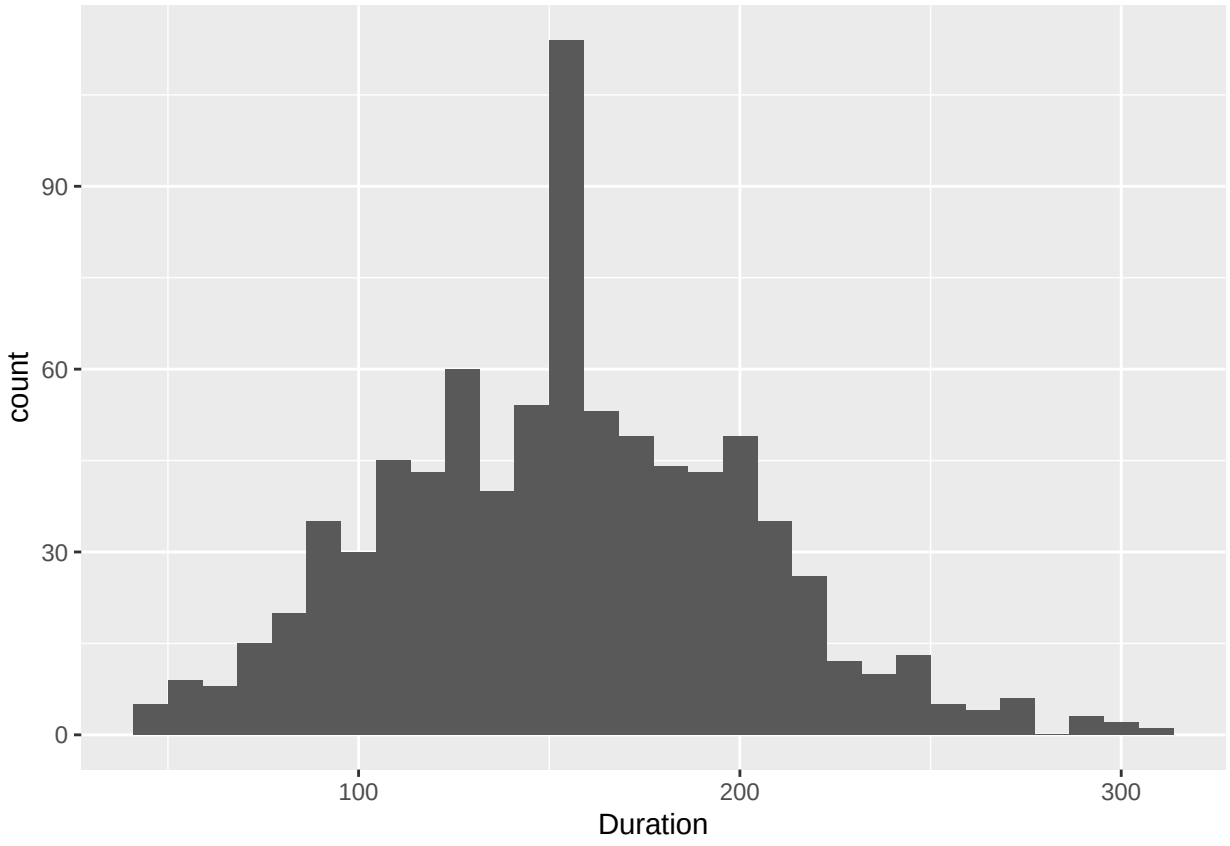
```
##                Max.   :305.62   Max.   :87.00   Max.   :132.78
##
##    speed_air          height           pitch          distance
##  Min.   : 90.00   Min.   : 6.228   Min.   :2.284   Min.   :  41.72
##  1st Qu.: 96.23   1st Qu.:23.530   1st Qu.:3.641   1st Qu.: 893.58
##  Median :101.12   Median :30.159   Median :4.004   Median :1262.15
##  Mean   :103.48   Mean   :30.442   Mean   :4.006   Mean   :1521.71
##  3rd Qu.:109.36   3rd Qu.:36.995   3rd Qu.:4.371   3rd Qu.:1936.01
##  Max.   :132.91   Max.   :59.946   Max.   :5.927   Max.   :5381.96
##  NA's   :630
```

Histogram

Duration

```
ggplot(data=faa_normal, aes(faa_normal$duration)) +
  geom_histogram() +
   labs(x="Duration")
```

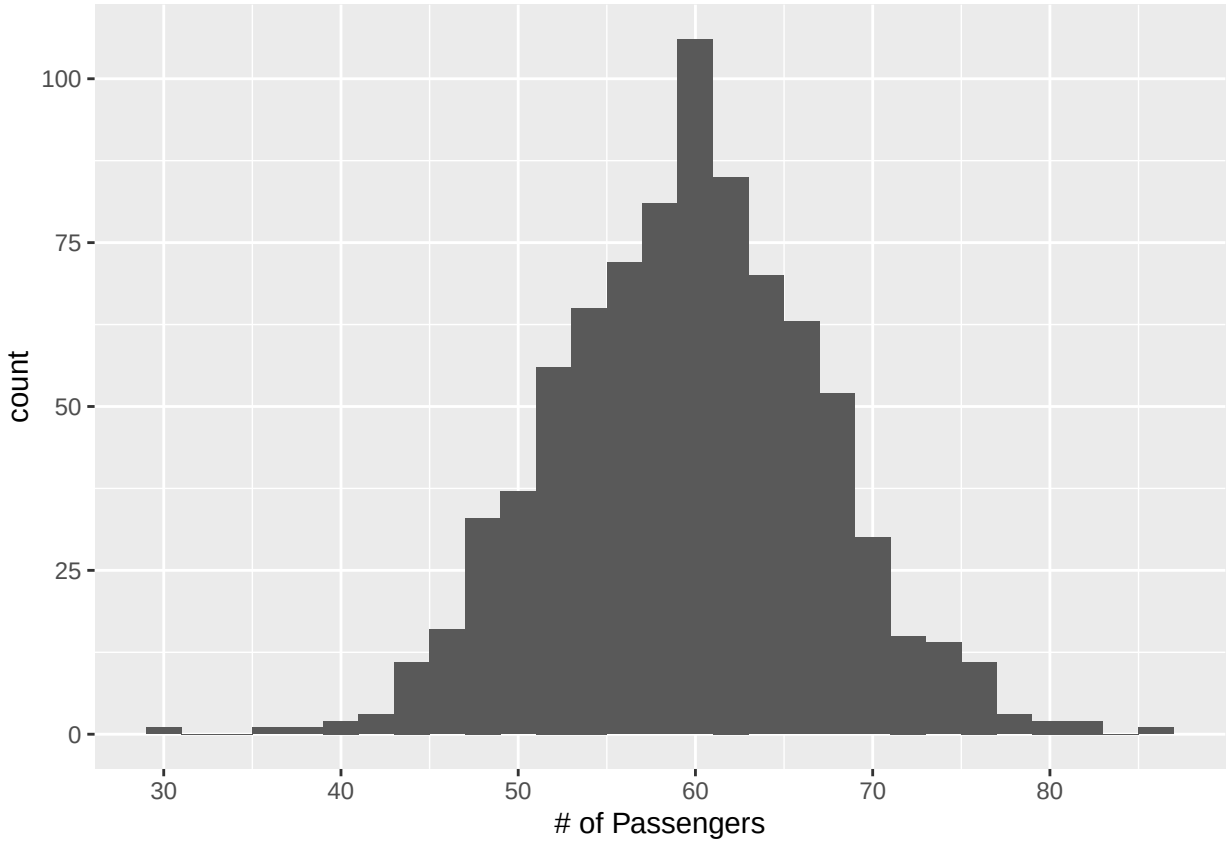
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Number of Passengers

```
ggplot(data=faa_normal, aes(faa_normal$no_pasg)) +
  geom_histogram() +
   labs(x="# of Passengers")
```

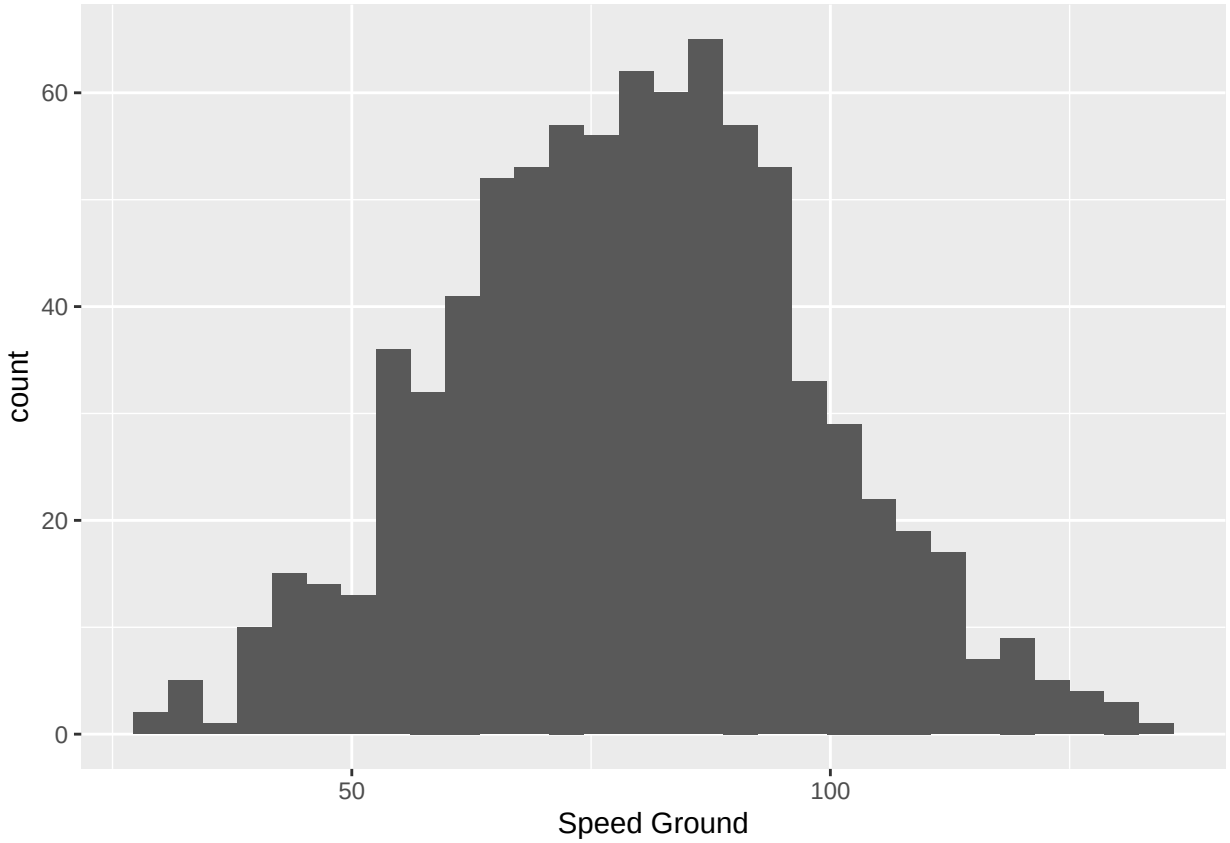
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Speed Ground

```
ggplot(data=faa_normal, aes(faa_normal$speed_ground)) +
  geom_histogram() +
   labs(x="Speed Ground")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

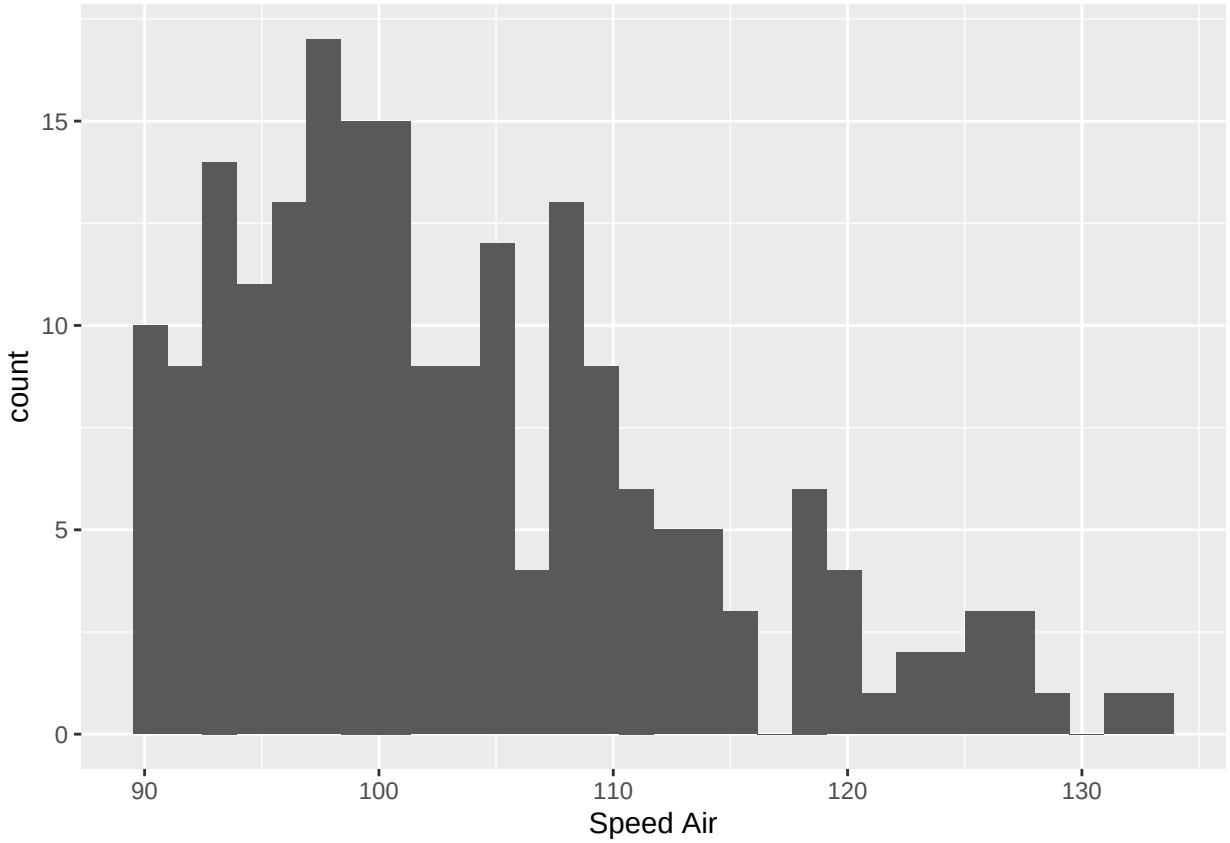


Speed Air

```
ggplot(data=faa_normal, aes(faa_normal$speed_air)) +
  geom_histogram() +
   labs(x="Speed Air")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

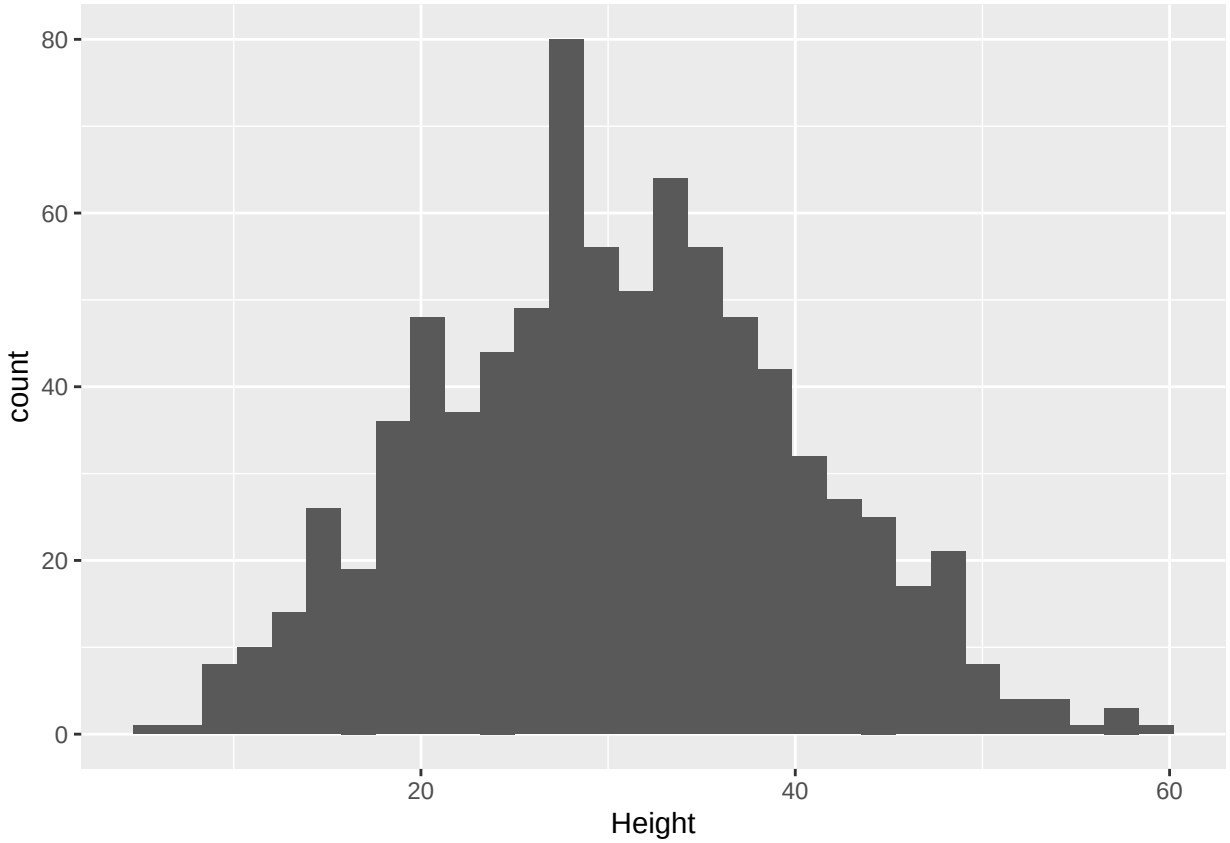
```
## Warning: Removed 630 rows containing non-finite values (stat_bin).
```



Height

```
ggplot(data=faa_normal, aes(faa_normal$height)) +
  geom_histogram() +
   labs(x="Height")
```

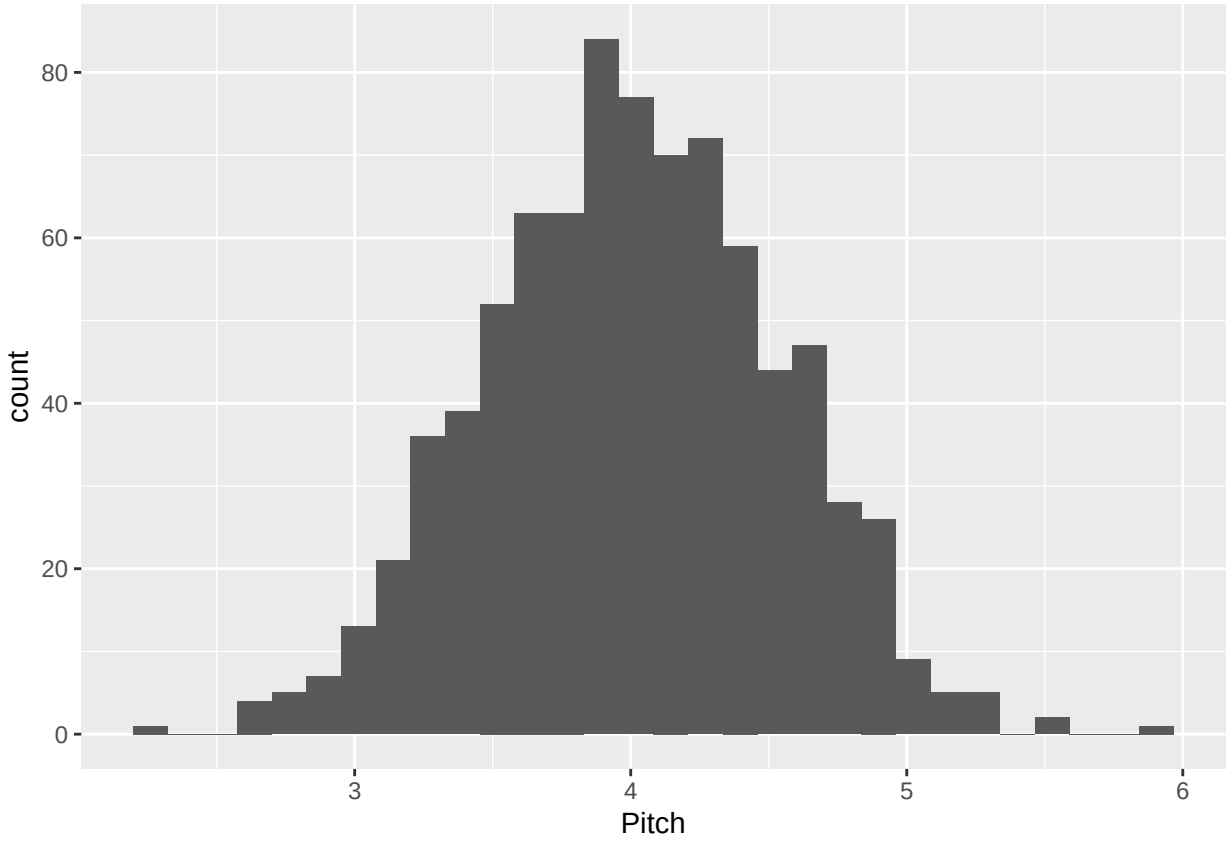
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Pitch

```
ggplot(data=faa_normal, aes(faa_normal$pitch)) +
  geom_histogram() +
   labs(x="Pitch")
```

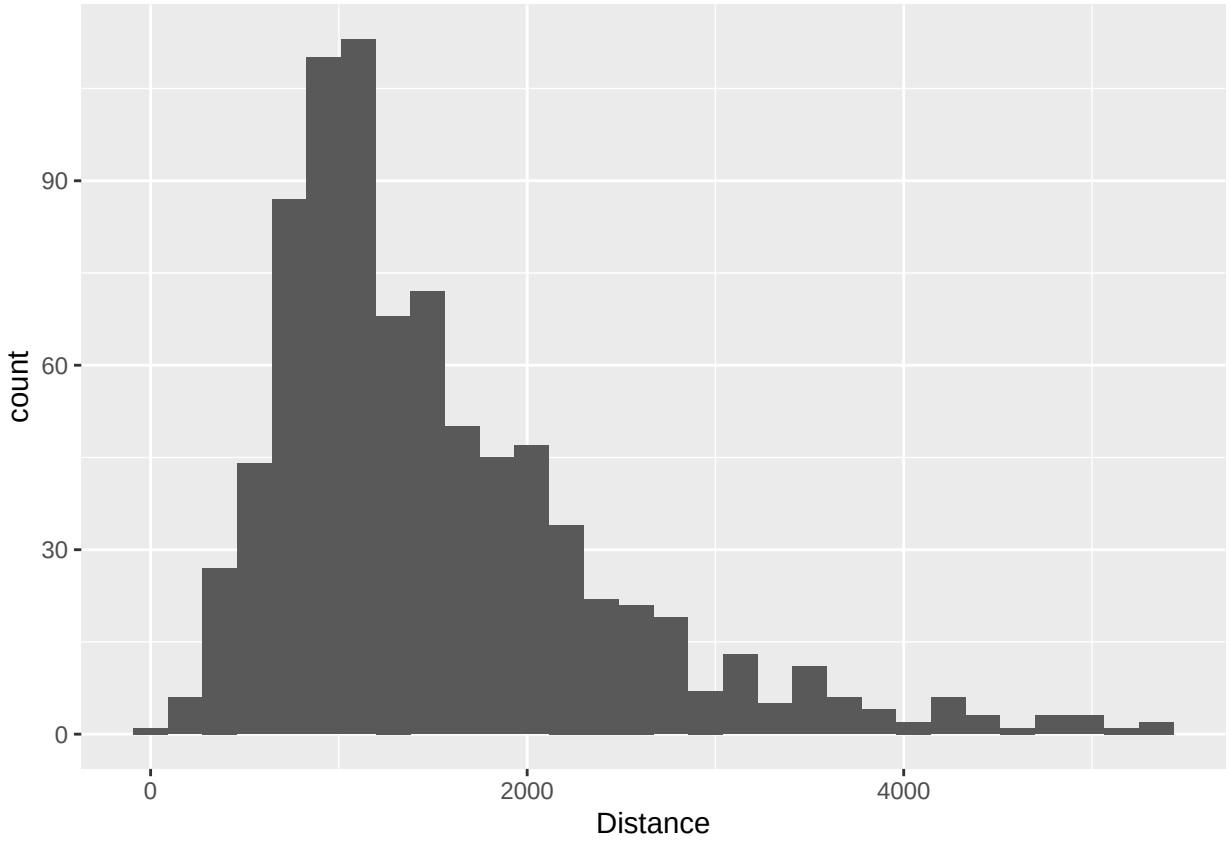
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Distance

```
ggplot(data=faa_normal, aes(faa_normal$distance)) +
  geom_histogram() +
   labs(x="Distance")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Cleaned Data Summary

1. Missing values in duration column have been dealt with. Still the range seems to be relatively wide and also the distribution is close to Normal.
2. Speed Ground is well spread out with a thick centre suggesting it to be Normal.
3. Speed Air proportionately still has very high number of missing values and also has a big tail on the right suggesting some right skew
4. Height is relatively stable now after removing all the abnormal values and tending towards normal distribution.
5. Range for distance is still quite big and has a very big right tail.

6. Identification of features impacting the target variaible - (Landing) Distance

First we create a table which ranks the pairwise correlation with the target variable. From initial analysis it seems Speed Air, inspite of large number of missing values, and Speed Ground seem to have the highest correlation and most correlations are positive.

```
# Converting Aircraft columns to Binary  type. airbus = 0 boeing = 1
faa_normal$aircraft <- ifelse(faa_normal$aircraft == "airbus", 0, 1)
# Creating an ordered correlation table against response variable with sign and magnitude
table1 <-
  cor(faa_normal, use="complete.obs") %>%
  as.data.frame() %>%
  mutate(variable = rownames(.)) %>%
  select(variable, correlation = distance) %>%
  filter(variable != "distance") %>%
```

```
  mutate(sign = ifelse(correlation > 0, "positive", "negative"),
         correlation = abs(correlation)) %>%
  arrange(desc(correlation))

table1
```

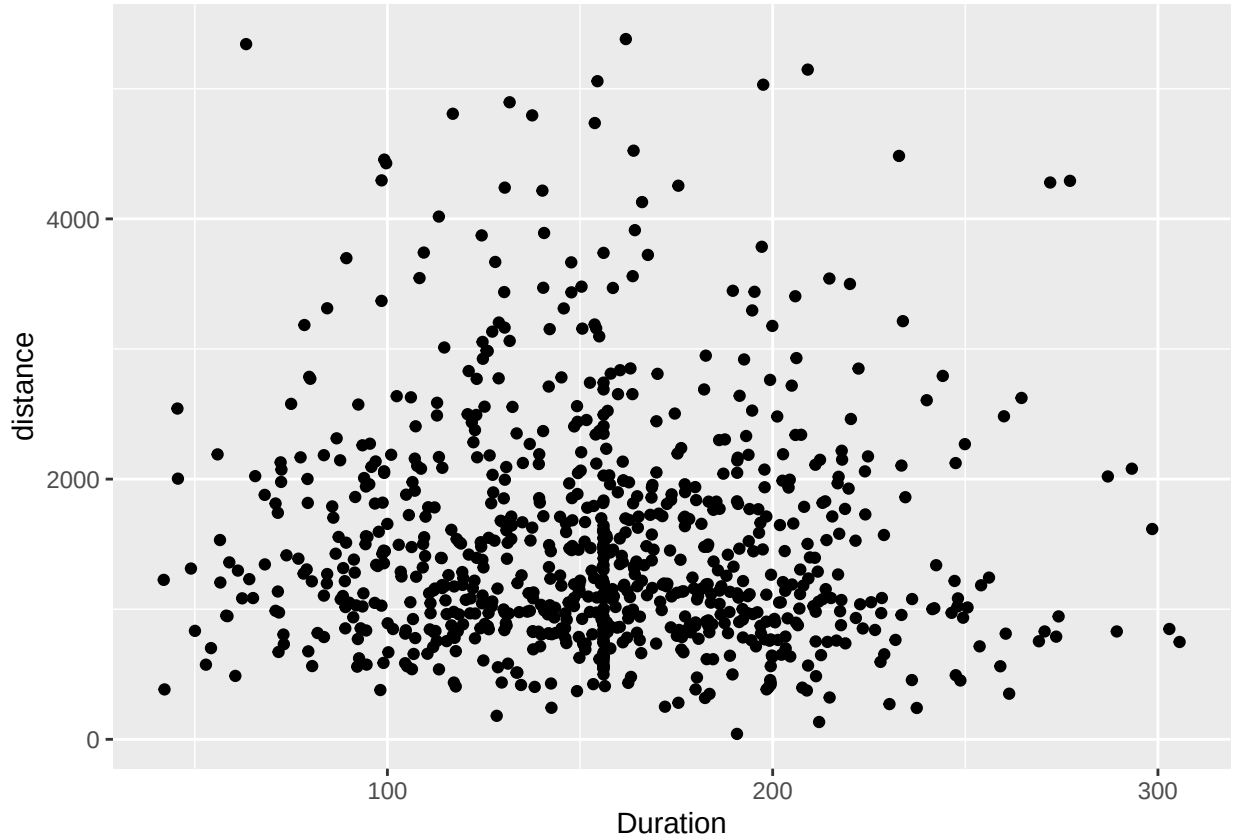
```
##       variable correlation     sign
## 1    speed_air  0.94209714 positive
## 2 speed_ground  0.92658618 positive
## 3     aircraft  0.17992145 positive
## 4       height  0.06814248 positive
## 5     duration  0.05058886 positive
## 6        pitch  0.04234550 positive
## 7      no_pasg  0.03994273 negative
```

Pairwise Scatter Plots

Below pairwise scatter plots confirms the pairwise correlation observed above.

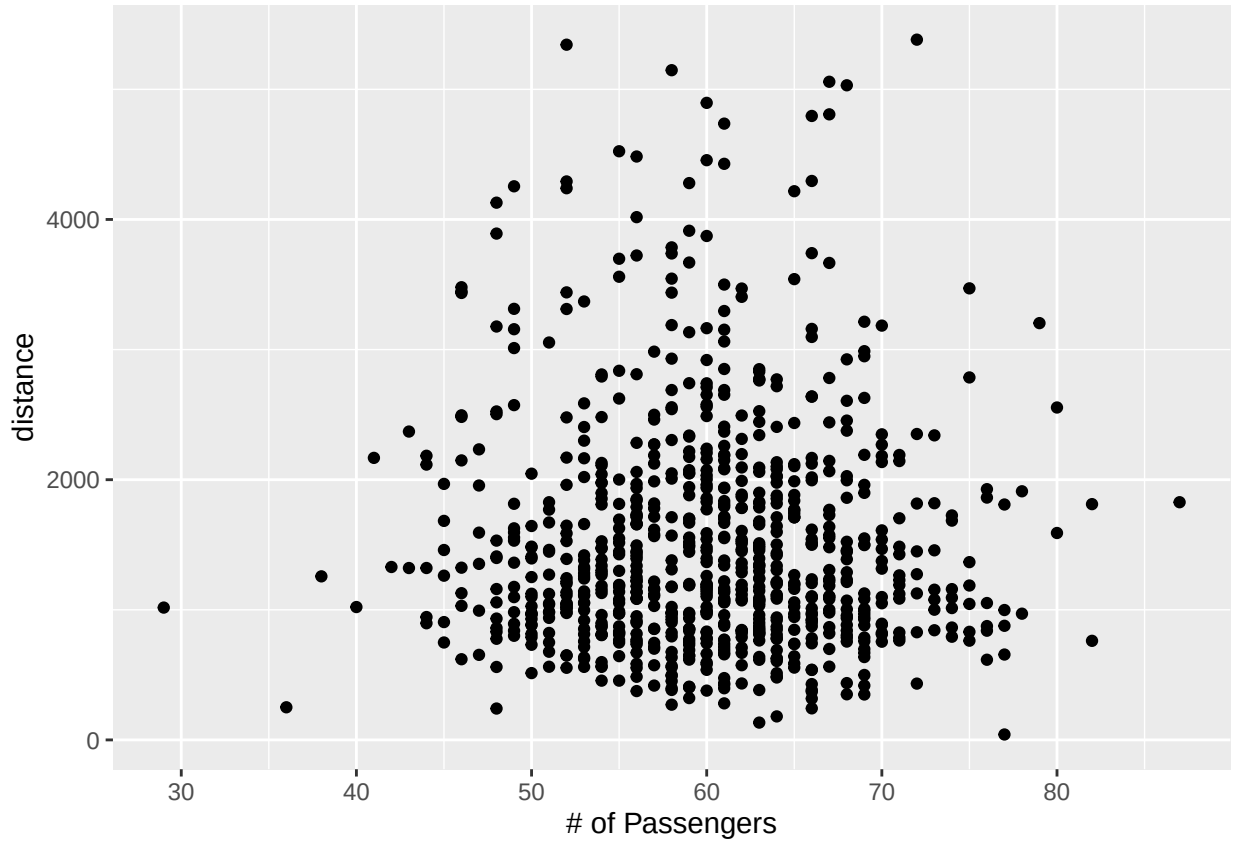
Duration

```
ggplot(faa_normal, aes(x=duration, y=distance)) +
  geom_point() +
  labs(x="Duration", y="distance")
```



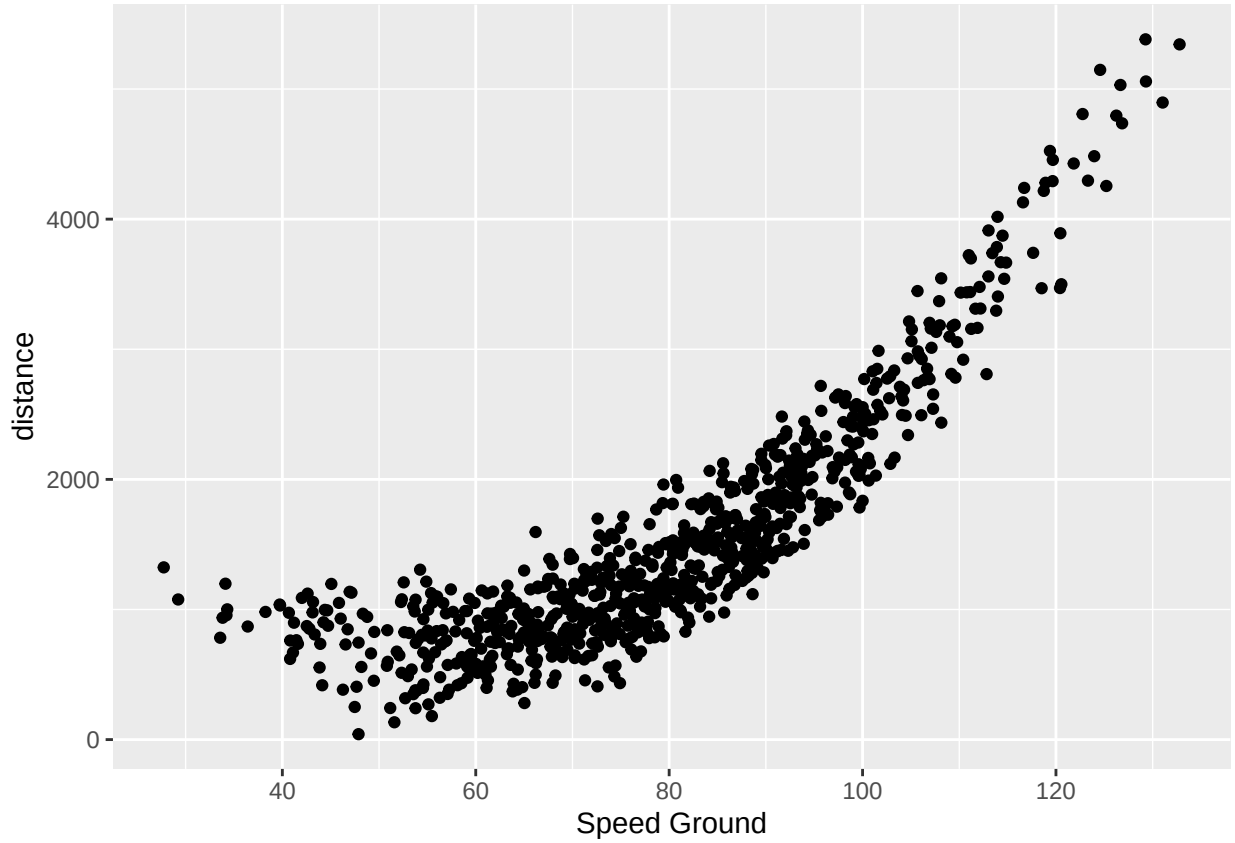
Number of Passengers

```
ggplot(faa_normal, aes(x=no_pasg, y=distance)) +
  geom_point() +
  labs(x="# of Passengers", y="distance")
```



Speed Ground

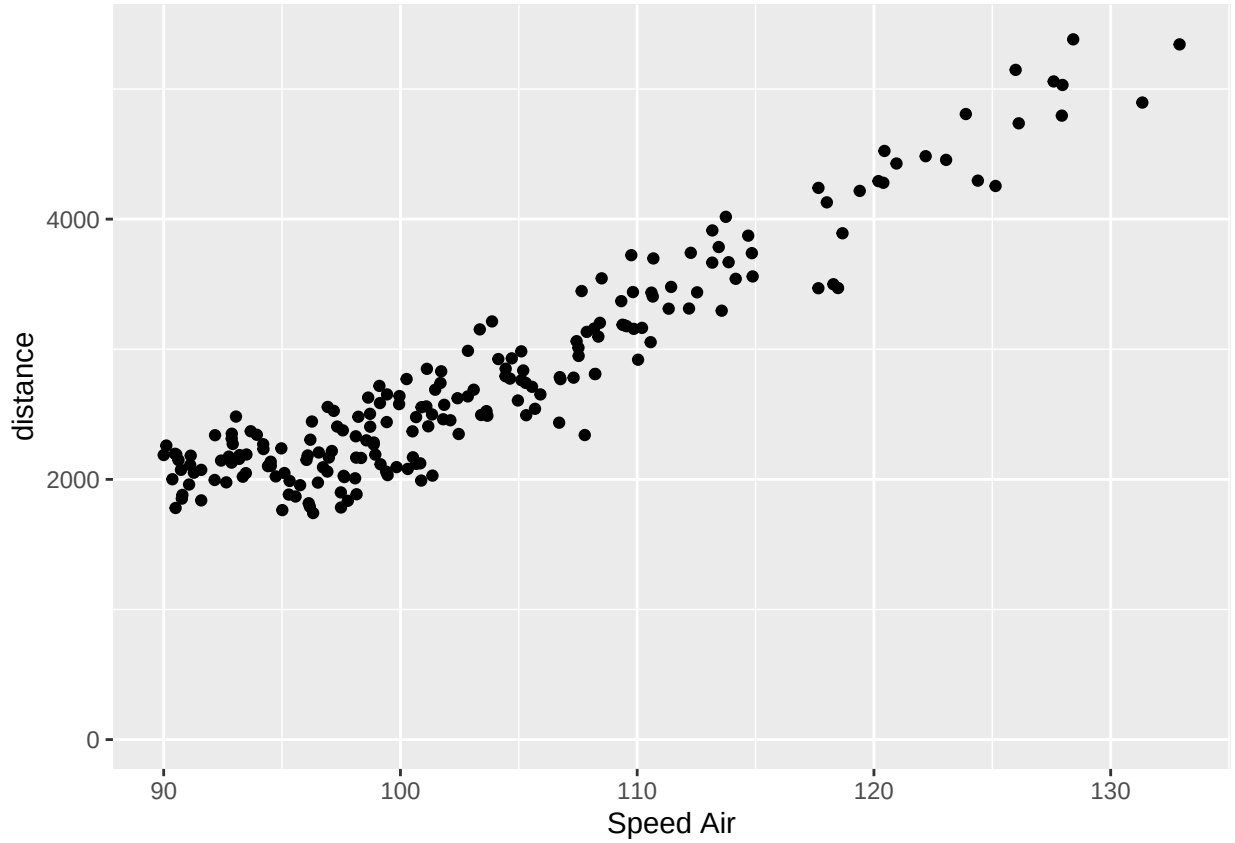
```
ggplot(faa_normal, aes(x=speed_ground, y=distance)) +
  geom_point() +
  labs(x="Speed Ground", y="distance")
```



Speed Air

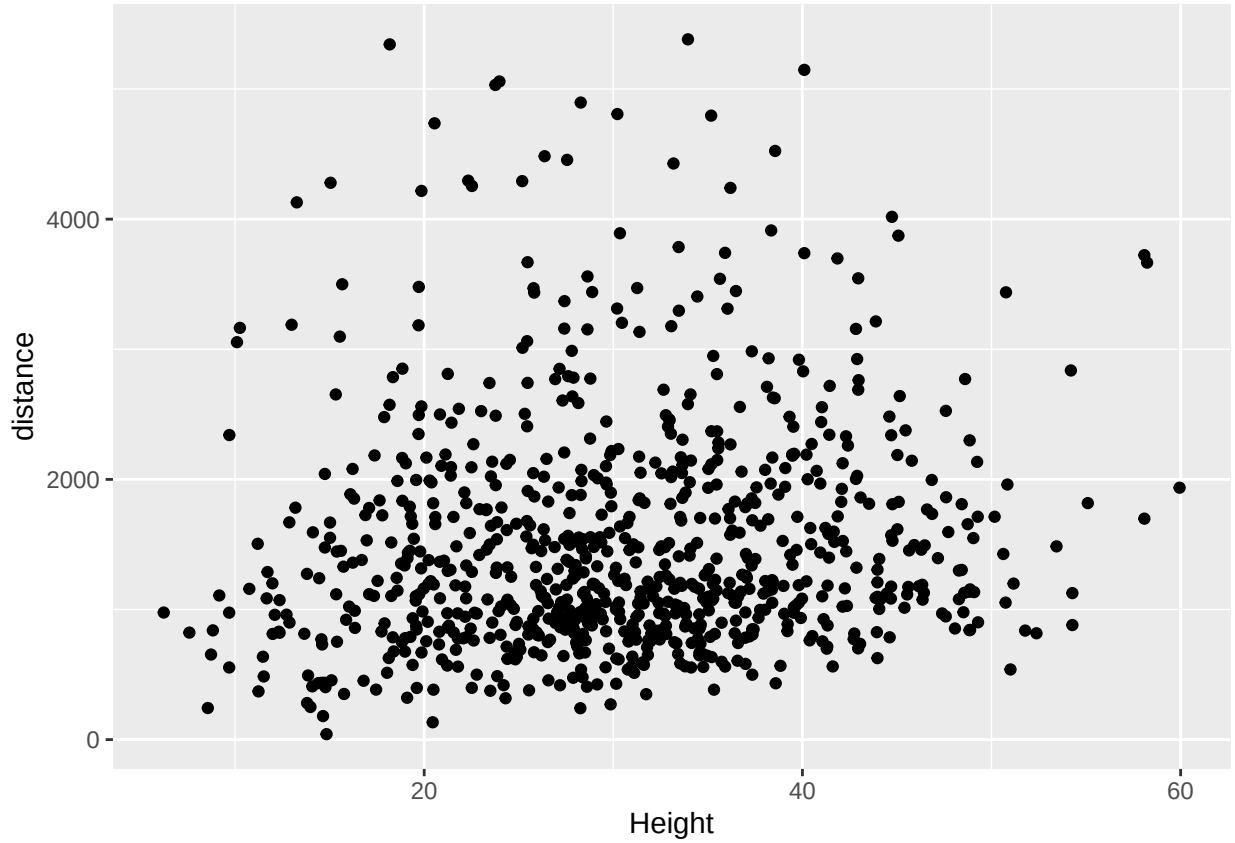
```
ggplot(faa_normal, aes(x=speed_air, y=distance)) +
  geom_point() +
  labs(x="Speed Air", y="distance")
```

```
## Warning: Removed 630 rows containing missing values (geom_point).
```



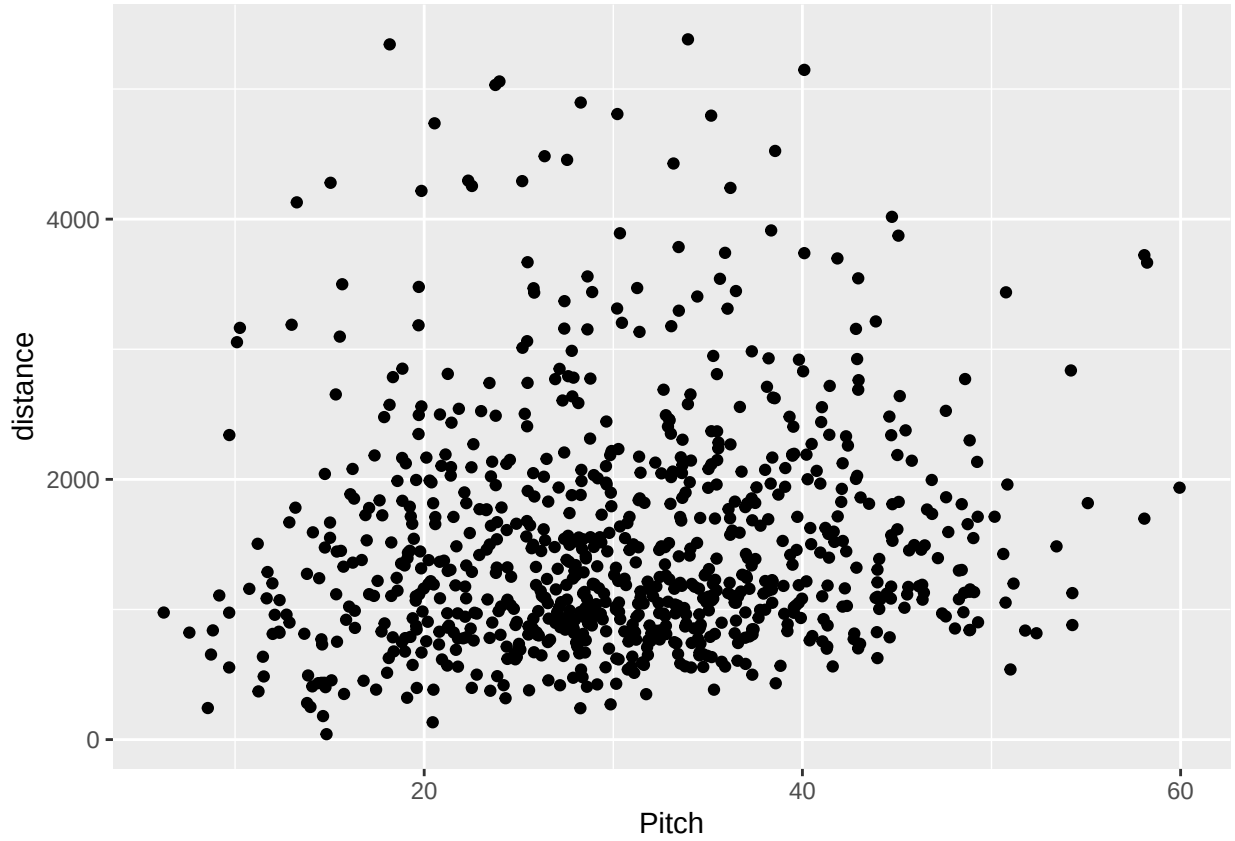
Height

```
ggplot(faa_normal, aes(x=height, y=distance)) +
  geom_point() +
  labs(x="Height", y="distance")
```



Pitch

```
ggplot(faa_normal, aes(x=height, y=distance)) +
  geom_point() +
  labs(x="Pitch", y="distance")
```



7. Regression using a single feature each time

Regressing Y (landing distance) on each of the X variables. Based on below results it seems aircraft is the most significant variable followed by speed_air and height

```
# Regressing target variables on all other predictor variables
model <- lm(distance ~ ., data=faa_normal)

# calculating the p-value and sign of correlation of all variables in the above regression model
# Creating an ordered p-value table for each variable regressed against response variable
table2 <-
  summary(model)$coefficients[-1,c(1,4)] %>%
    as.data.frame() %>%
    mutate(variable = rownames(.),
           coef_direction = ifelse(Estimate >0, "positive", "negative")) %>%
    select(variable, 'Pr(>|t|)', coef_direction) %>%
    arrange(.[,2])

table2
```

```
##       variable     Pr(>|t|) coef_direction
## 1     aircraft 6.978583e-52       positive
## 2    speed_air 2.085074e-30       positive
## 3       height 1.426938e-29       positive
## 4      no_pasg 9.816591e-02       negative
## 5 speed_ground 3.920247e-01       negative
## 6     duration 5.418284e-01       positive
```

```
## 7        pitch 8.233101e-01       negative
```

Below each variable is standardized and now, the standardized target varaible is regressed on all other variables. It is found that Speed Air, Aircraft, Height and Speed Ground to be significant variables based on the magnitude of coeeficients.

```
# Standardizinng numerical variables
faa_scaled <- scale(faa_normal)%>%
  as.data.frame()
# Normalized dataframe
head(faa_scaled)
```

```
##   aircraft   duration    no_pasg speed_ground   speed_air      height
## 1 1.067716 -1.2042214 -0.9388392    1.5093398  0.60016176 -0.30916533
## 2 1.067716 -0.6226258  1.1950027    1.1777583 -0.06507929 -0.26971182
## 3 1.067716 -0.9153269  0.1280817   -0.4432398          NA -1.21212953
## 4 1.067716  0.8944622 -0.5387439    0.3386332          NA  0.03093892
## 5 1.067716 -1.3831270  1.3283678   -1.0345390          NA  0.19999439
## 6 1.067716 -0.3694835 -0.6721090   -0.2333622          NA  1.10170439
##         pitch   distance
## 1  0.07115958  2.0639787
## 2  0.21161695  1.6373268
## 3  0.81324220 -0.4207936
## 4 -0.23150143  0.1591532
## 5  0.03806159 -0.5265071
## 6  0.37583552  0.1176643
```

```
# Regressing target variable on all normalized variables
model <- lm(distance ~ ., data=faa_scaled)
# calculating the coefficient and sign of coefficient in the above regression model
table3 <-
  summary(model)$coefficients[-1,c(1), drop=FALSE] %>%
  as.data.frame() %>%
  mutate(variable = rownames(.),
         coef_direction = ifelse(Estimate >0, "positive", "negative"),
         coefficient = abs(Estimate)) %>%
  select(variable, coefficient, coef_direction) %>%
  arrange(desc(coefficient))

table3
```

```
##       variable coefficient coef_direction
## 1    speed_air 0.951489519       positive
## 2     aircraft 0.238213257       positive
## 3       height 0.148397171       positive
## 4 speed_ground 0.113695588       negative
## 5      no_pasg 0.018772280       negative
## 6     duration 0.006504340       positive
## 7        pitch 0.002374167       negative
```

8. Comparing Results

Results from above 3 analysis is compared with each other to check if the significant variables are consistent in each analysis. From the below table it is seen that all 3 analysis are reasonably consistent if not entirely. Speed Air, Aircraft, Height & Speed Ground are orderly listed as the most influencing factors for the target variable (Landing) Distance.

Pairwise Correlation (Table 1)

Regression (Table 2)

Standardized Regression (Table 3)

Speed Air

Speed Ground

Aircraft

Speed Air

Height

Speed Air

Aircraft

Height

Speed Ground

Table 0

An ordered table is created as 'Table 0' to list factors affecting the target variable based on above analysis.

Rank

Variable

1

Speed Air

2

Aircraft

3

Height

4

Speed Ground

5

No. of Passengers

6

Pitch

7

Duration

9. Checking for Collinearity

1. Coefficients in both the model1 & model2 for respective variables Speed Ground & Speed Air are positive.
2. Surprisingly coefficient for Speed Ground in model 3 is negative, suggesting multi-collinearity.
3. Very high positive correlation of 0.988 between Speed Air & Speed Ground confirms the above assumption.

4. Speed Air is preferrred to Speed Ground as the adjusted R-Square for model2 is significantly higher than model 2.

```
model1 <- lm(distance ~ speed_ground, data=faa_normal)
summary(model1)
```

```
##
## Call:
## lm(formula = distance ~ speed_ground, data = faa_normal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -904.18 -319.13  -75.69  213.51 1912.03
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1720.6284    68.3579  -25.17   <2e-16 ***
## speed_ground    40.8252     0.8374   48.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 456 on 831 degrees of freedom
## Multiple R-squared:  0.7409, Adjusted R-squared:  0.7406
## F-statistic:  2377 on 1 and 831 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(distance ~ speed_air, data=faa_normal)
summary(model2)
```

```
##
## Call:
## lm(formula = distance ~ speed_air, data = faa_normal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -776.21 -196.39    8.72  209.17  624.34
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5455.709    207.547  -26.29   <2e-16 ***
## speed_air      79.532      1.997   39.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 276.3 on 201 degrees of freedom
##   (630 observations deleted due to missingness)
## Multiple R-squared:  0.8875, Adjusted R-squared:  0.887
## F-statistic:  1586 on 1 and 201 DF,  p-value: < 2.2e-16
```

```
model3 <- lm(distance ~ speed_ground + speed_air, data=faa_normal)
summary(model3)
```

```
##
## Call:
## lm(formula = distance ~ speed_ground + speed_air, data = faa_normal)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -819.74 -202.02    3.52  211.25  636.25
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5462.28     207.48 -26.327  < 2e-16 ***
## speed_ground   -14.37      12.68  -1.133    0.258
## speed_air       93.96      12.89   7.291 6.99e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 276.1 on 200 degrees of freedom
##   (630 observations deleted due to missingness)
## Multiple R-squared:  0.8883, Adjusted R-squared:  0.8871
## F-statistic:   795 on 2 and 200 DF,  p-value: < 2.2e-16
```

```
# Correlation between speed_air & speed_ground
cor(faa_normal$speed_ground, faa_normal$speed_air, use = "complete.obs")
```

```
## [1] 0.9879383
```

10. Variable Selection Based on Ranking in Table 0.

Six different models are built based on ranking listed in table 0 leaving out speed_ground (for multi-collineraity reasons) and compare R-square, Adjusted R-square and AIC for these varaiables.

From the below analysis it is found that 1. R-Square increases with increase in the number of varaiables. It increases very slowly after model 3. 2. Adjusted increases with increase in the number of varaiables up until model 4 afterwards there is a slow decrease. Similar to R-Square Criteria it remains almost constant after model 3. 3. AIC decreases with increase in the number of varaiables up until model 4 afterwards there is a slow increase in AIC. Similarly to above criterias AIC is almost constant after model 3.

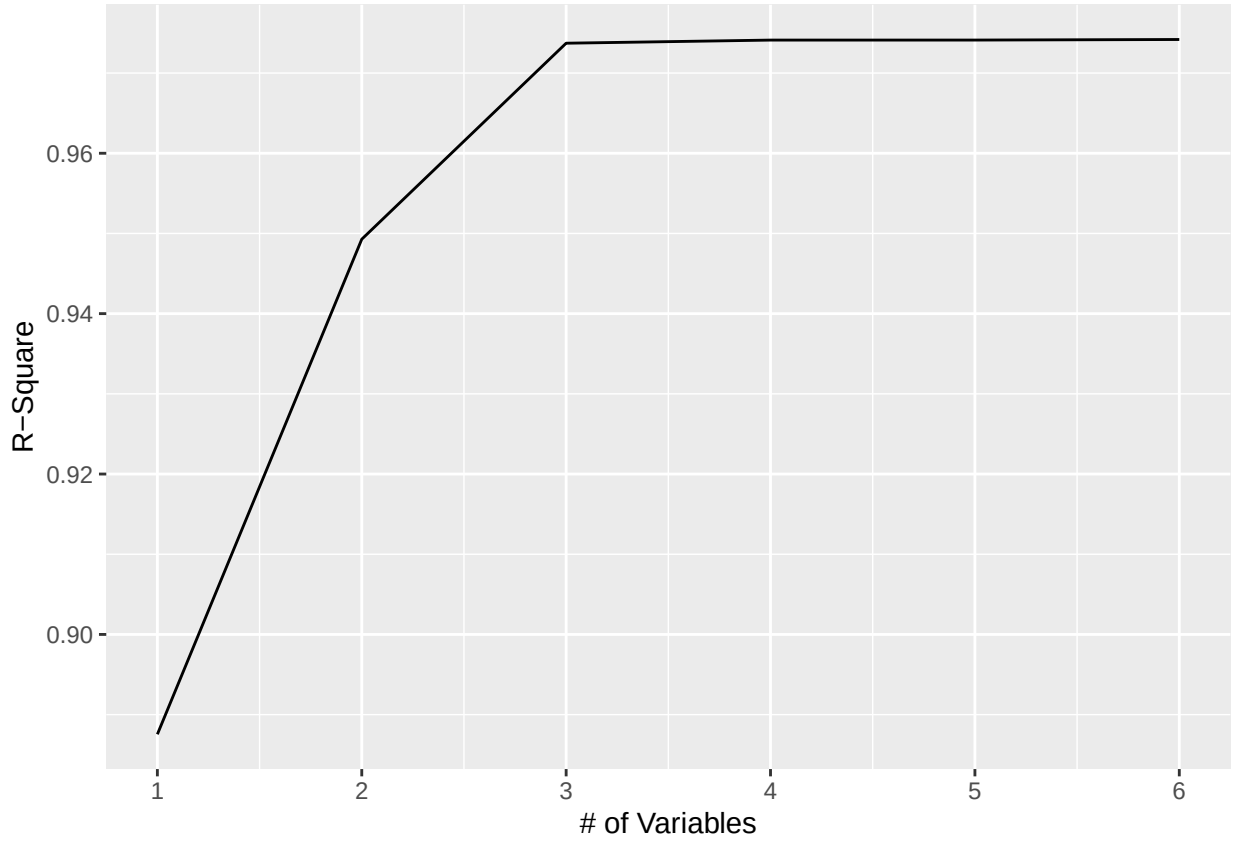
```
# running 6 separate regression models by adding variables in each model based on Table 0
model1 <- lm(distance ~ speed_air, data=faa_normal)
model2 <- lm(distance ~ speed_air+aircraft, data=faa_normal)
model3 <- lm(distance ~ speed_air+aircraft+height, data=faa_normal)
model4 <- lm(distance ~ speed_air+aircraft+height+no_pasg, data=faa_normal)
model5 <- lm(distance ~ speed_air+aircraft+height+no_pasg+pitch, data=faa_normal)
model6 <- lm(distance ~ speed_air+aircraft+height+no_pasg+pitch+duration, data=faa_normal)

# Tabulating R-Square, Adjusted R-Square & AIC for each model
manual_model_analysis <-
rbind(c(summary(model1)$r.squared,summary(model1)$adj.r.squared,AIC(model1),1),
      c(summary(model2)$r.squared,summary(model2)$adj.r.squared,AIC(model2),2),
      c(summary(model3)$r.squared,summary(model3)$adj.r.squared,AIC(model3),3),
      c(summary(model4)$r.squared,summary(model4)$adj.r.squared,AIC(model4),4),
      c(summary(model5)$r.squared,summary(model5)$adj.r.squared,AIC(model5),5),
      c(summary(model6)$r.squared,summary(model6)$adj.r.squared,AIC(model6),6)) %>%
  as.data.frame() %>%
  select(r_square = V1,adj_r_square = V2, AIC = V3, model = V4)
manual_model_analysis
```

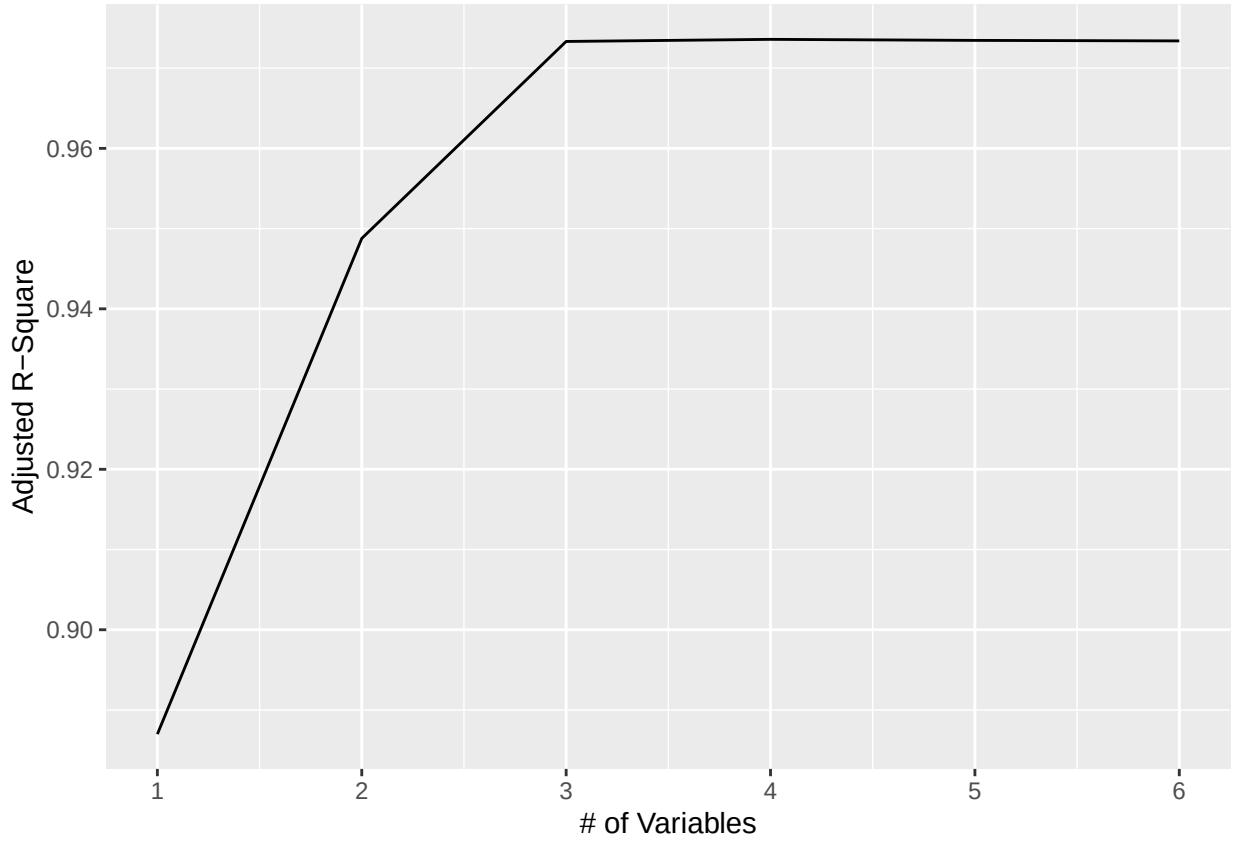
```
##    r_square adj_r_square      AIC model
## 1 0.8875470    0.8869875 2862.423     1
## 2 0.9492819    0.9487747 2702.784     2
## 3 0.9737209    0.9733248 2571.310     3
```

```
## 4 0.9741117    0.9735887 2570.268     4
## 5 0.9741142    0.9734572 2572.249     5
## 6 0.9741845    0.9733942 2573.697     6
```

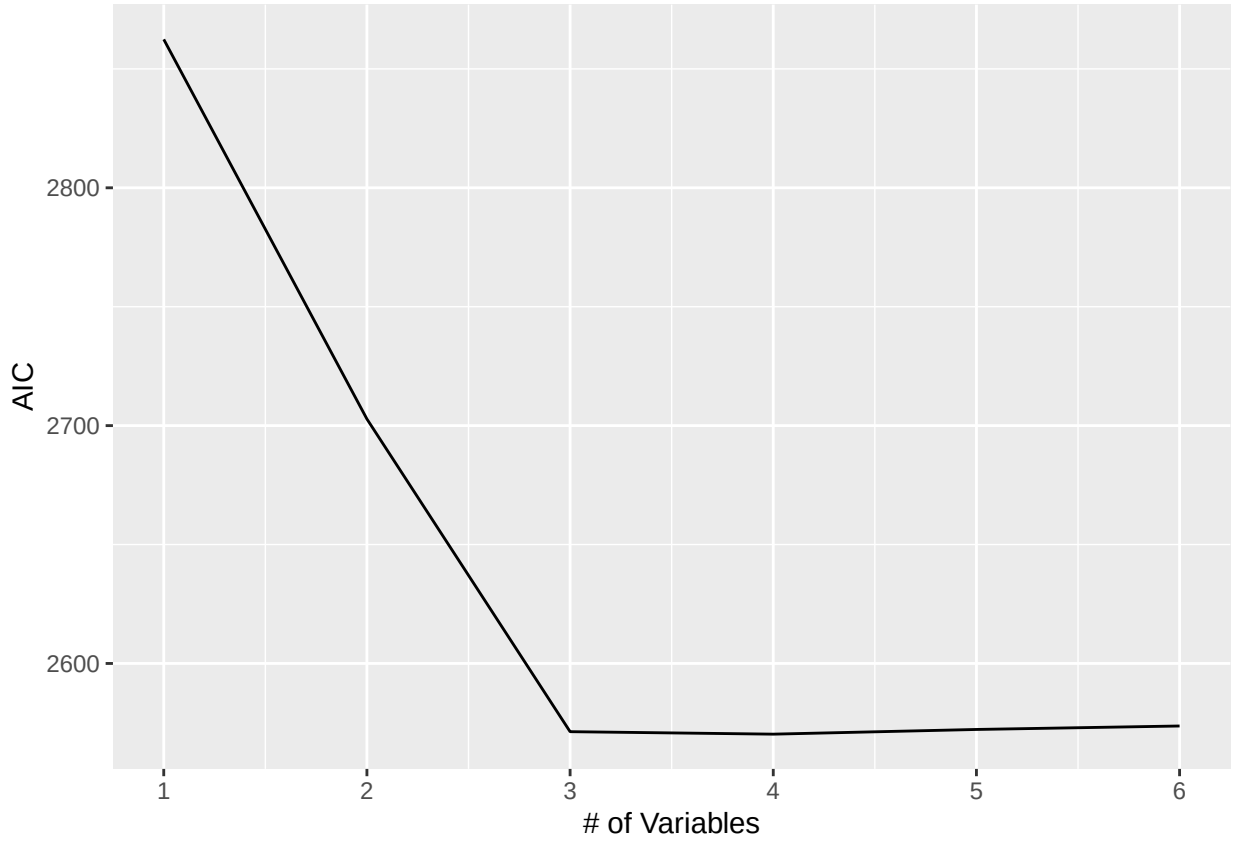
```r
# R-Square vs. No.of Variables included in the model
ggplot(data=manual_model_analysis, aes(x=model, y=r_square)) +
  geom_line() +
  scale_x_continuous(breaks=1:6) +
  labs(x="# of Variables", y="R-Square")
```




```r
# Adjusted R-Square vs. No.of Variables included in the model
ggplot(data=manual_model_analysis, aes(x=model, y=adj_r_square)) +
  geom_line() +
  scale_x_continuous(breaks=1:6) +
  labs(x="# of Variables", y="Adjusted R-Square")
```




```
# AIC vs. No.of Variables included in the model
ggplot(data=manual_model_analysis, aes(x=model, y=AIC)) +
  geom_line() +
  scale_x_continuous(breaks=1:6) +
  labs(x="# of Variables", y="AIC")
```



Final List of Variables

Based on the above analysis & since all 3 criterias almost remain constant after model 3, following 3 variables are selected for the final model. 1. Speed Air 2. Aircraft 3. Height

11. Comparing Manual Analysis with Automated Function

In-built "StepAIC" function is used to select the list of # of variables to be included in the model and the results are consistent with the above conclusions.

```
stepAIC(model6,k=6)
```

```
## Start:  AIC=2023.61
## distance ~ speed_air + aircraft + height + no_pasg + pitch +
##     duration
##
##             Df Sum of Sq       RSS    AIC
## - pitch      1       209   3523198 2017.6
## - duration   1      9599   3532588 2018.2
## - no_pasg    1     50048   3573037 2020.5
## <none>                     3522989 2023.6
## - height     1   3286346   6809335 2151.4
## - aircraft   1   7911209  11434198 2256.6
## - speed_air  1 127190141 130713130 2751.2
##
## Step:  AIC=2017.62
## distance ~ speed_air + aircraft + height + no_pasg + duration
```

```
##
##             Df Sum of Sq       RSS    AIC
## - duration   1      9723   3532921 2012.2
## - no_pasg    1     49919   3573117 2014.5
## <none>                     3523198 2017.6
## - height     1   3288325   6811523 2145.4
## - aircraft   1   8858940  12382138 2266.8
## - speed_air  1 127227956 130751154 2745.2
##
## Step:  AIC=2012.18
## distance ~ speed_air + aircraft + height + no_pasg
##
##             Df Sum of Sq       RSS    AIC
## - no_pasg    1     53331   3586252 2009.2
## <none>                     3532921 2012.2
## - height     1   3332283   6865204 2141.0
## - aircraft   1   8851457  12384379 2260.8
## - speed_air  1 127599057 131131978 2739.8
##
## Step:  AIC=2009.22
## distance ~ speed_air + aircraft + height
##
##             Df Sum of Sq       RSS    AIC
## <none>                     3586252 2009.2
## - height     1   3335147   6921399 2136.7
## - aircraft   1   8956602  12542854 2257.4
## - speed_air  1 127667330 131253582 2734.0

##
## Call:
## lm(formula = distance ~ speed_air + aircraft + height, data = faa_normal)
##
## Coefficients:
## (Intercept)    speed_air     aircraft       height
##    -6390.38        82.15       427.44        13.70
```