

Anh Nguyen

Professor Liu

Statistical Computing

10 December 2019

FACTORS THAT AFFECT FLIGHT LANDING AND HOW TO REDUCE THE RISK OF LANDING OVERRUN

Executive Summary: After cleaning and analyzing the data set FAA, we've found that the 2 variables that affect the landing distance significantly are:

- **Speed_ground** (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway.
- **Speed_air** (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway.

According to this equation:

$$\text{Distance} = -5425.49 + 91.63 \text{ Air speed} - 12.32 \text{ Ground speed (feet)}$$

From this formula, we can put a limit of landing distance, then consider the level of air speed and ground speed most optimized for safety and time-efficient. Thus we would be able to use this design policies fittingly.

When we analyze only individual Boeing and Airbus, we found that a similar situation with Airbus, but it can't be disproved that the variable Ground speed is not statistically significant to affect Landing Distance.

$$\text{Airbus: Distance} = -5977.15 + -21.49 \text{ Air speed} - 103.68 \text{ Ground speed (feet)}$$

$$\text{Boeing: Distance} = \text{Distance} = -5378.033 + 80.453 \text{ Air speed (feet)}$$

Step 1: Data Import and Exploration

1. The goal: To import the raw data set we want to operate on, and have a summary of it (Its variables and observations)

2. The code:



#1. Importing the file into R

```
FAA<-read.table("/Users/MACOS/Downloads/FAA1.csv",header = TRUE, sep = ",")
```

#Understanding the data

```
str(FAA)
```

=> **The output:**

 FAA	800 obs. of 8 variables	
---	-------------------------	--

- Contents of the terminal:

```
> str(FAA)
```

```
'data.frame': 800 obs. of 8 variables:
```

```
$ aircraft : Factor w/ 2 levels "airbus","boeing": 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 ...
```

=> **Observations:**

- The combined data set there are 8 variables and 800 observations in total.

Step 2: Data Cleaning, based on the information in the prompt

1. **The goal:** To create a dataset clean enough to make sure there's no abnormality, so it won't skew the result of the analysis

2. **The code:**

#2. Data Cleaning

```
FAA<-subset(FAA, speed_ground>=30 & speed_ground<=140)
```



```
FAA<-subset(FAA, duration > 40)
```

```
FAA<-subset(FAA, height >= 6)
```

```
FAA<-subset(FAA, distance < 6000)
```

```
FAA<-subset(FAA, speed_air>=30 & speed_air <= 140 | is.na(speed_air))
```

=> **The output:**

 FAA	781 obs. of 8 variables	
---	-------------------------	---

=> **The observations:**

- I deleted all the observations that are abnormal and missing the “durations” values. I didn't delete the observations that miss the “speed_air” values, since a lot of the data set miss all “speed_air” values, therefore, it'll make no sense deleting a huge chunk of the dataset.
- The end result of the cleaning: There are 781 observations left

Step 3: Data Visualization using R

1. **The goal:** To create distributions graphs of all variables, to better understand them; and to get a sense of the relationships between them, using xy plot.

For the purpose of this project, I will create graphs with Distance as the dependent variable.

2. **The code:**

#3. Data visualization

Histogram of some of variables

hist(FAA\$duration)

hist(FAA\$no_pasg)

hist(FAA\$distance)

hist(FAA\$speed_ground)

hist(FAA\$pitch)

hist(FAA\$speed_air)

hist(FAA\$height)

#XY plot showing relationships between variables

plot(FAA\$speed_ground, FAA\$distance)

plot(FAA\$duration, FAA\$distance)

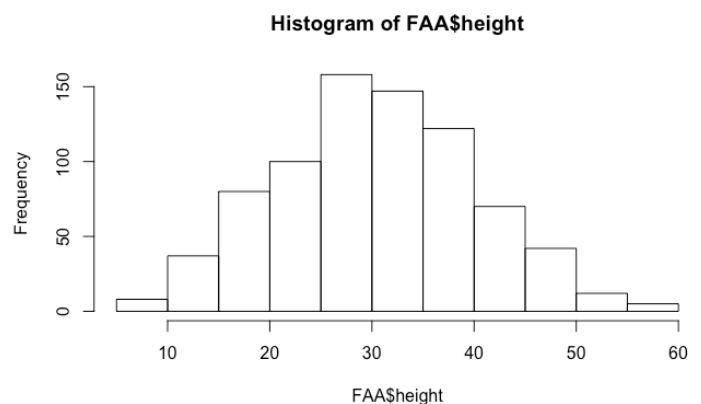
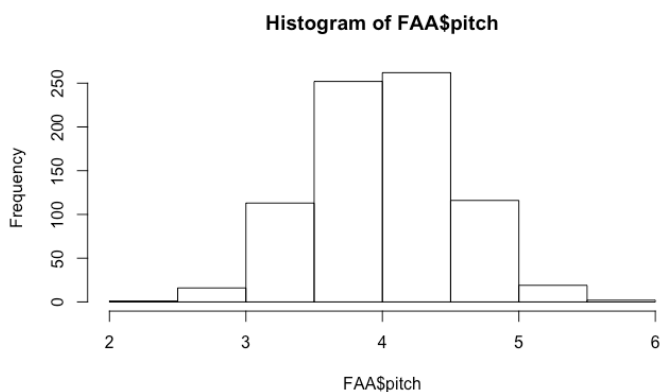
plot(FAA\$no_pasg, FAA\$distance)

plot(FAA\$speed_air, FAA\$distance)

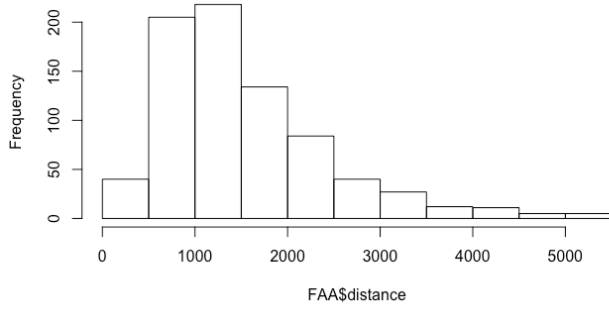
plot(FAA\$height, FAA\$distance)

plot(FAA\$pitch, FAA\$distance)

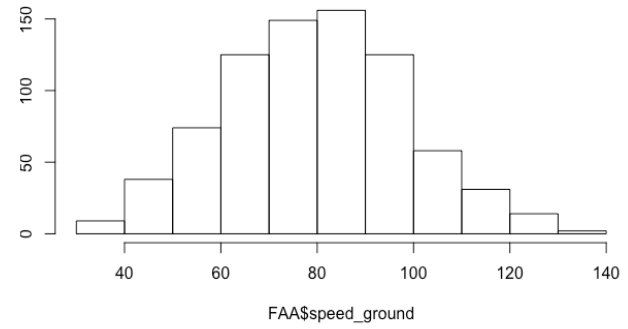
=> The output:



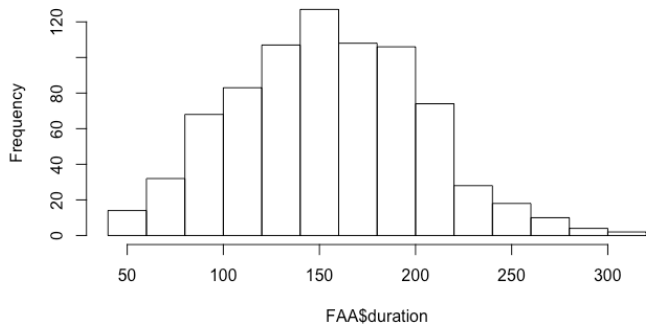
Histogram of FAA\$distance



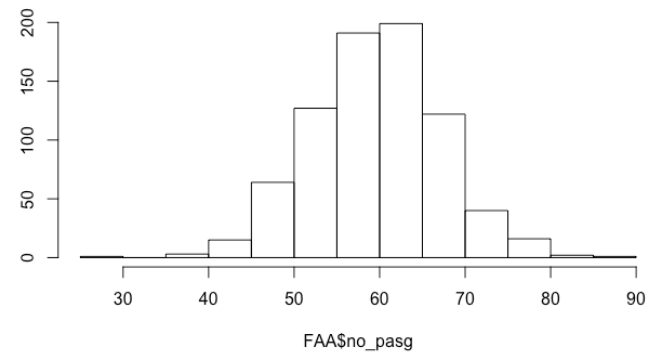
Histogram of FAA\$speed_ground



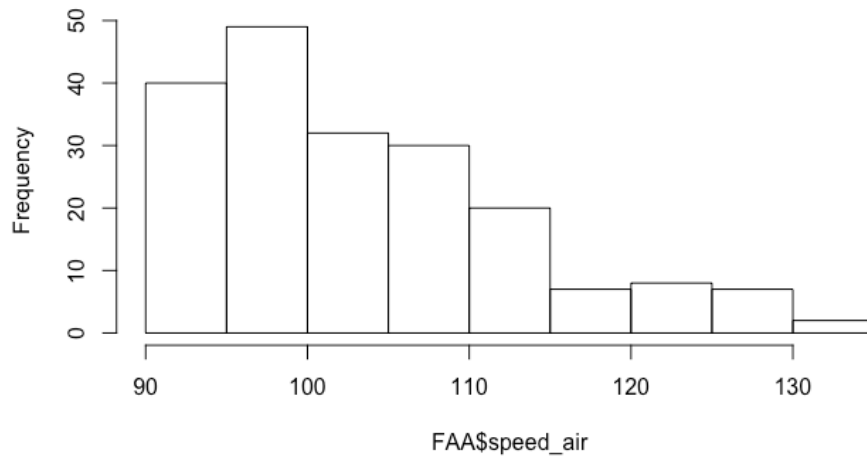
Histogram of FAA\$duration

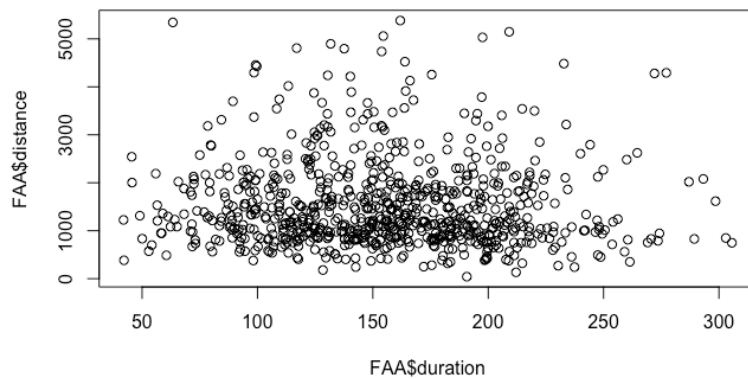
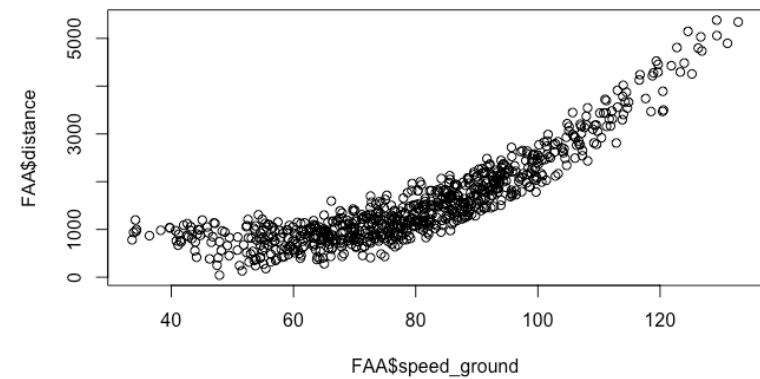
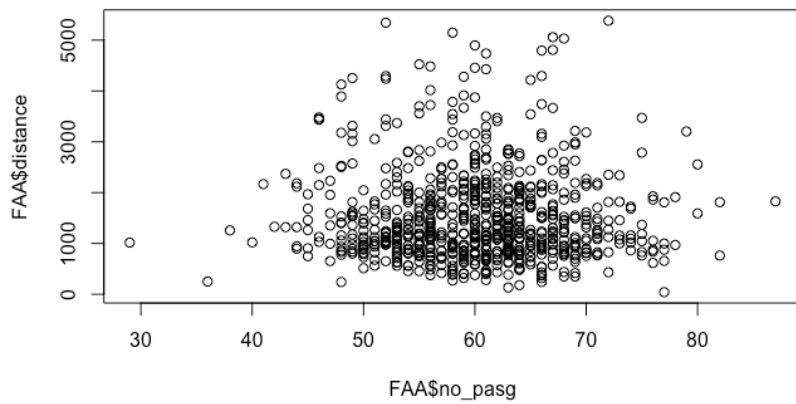
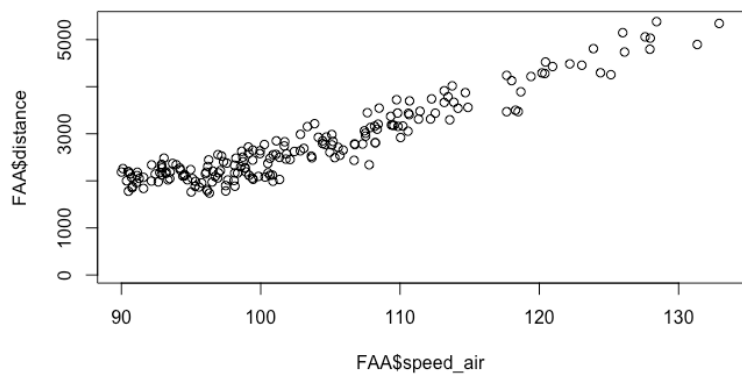
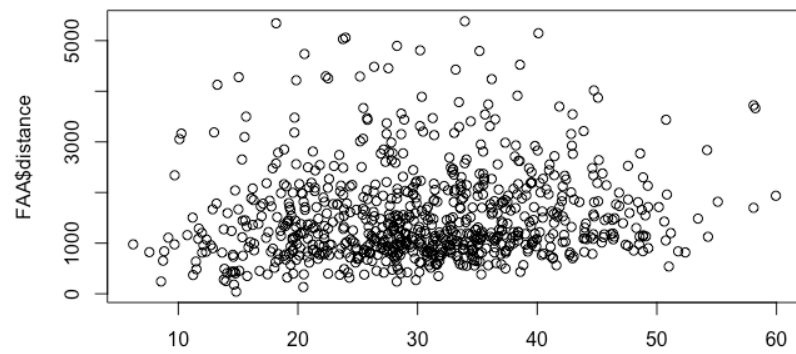
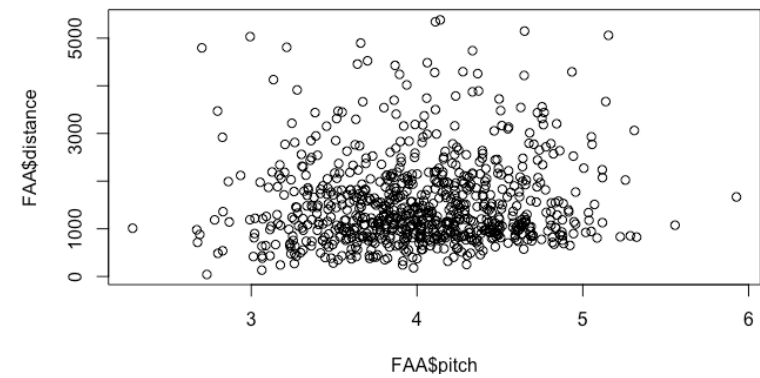


Histogram of FAA\$no_pasg



Histogram of FAA\$speed_air





Step 4: Model fitting

The goal: To decide what variable we should keep in the model

1. Correlations through the figures

The code:

```
#Correlation Matrix
cor(FAA$speed_ground,FAA$distance)
CorrFAA<-subset( FAA, select = -c(2,3,6,7))
CorrFAA<-subset( CorrFAA, speed_air>=30 & speed_air <= 140)
cor(CorrFAA$distance,CorrFAA$speed_air)
```

=> The output:

```
> cor(CorrFAA$distance,CorrFAA$speed_air)
[1] 0.943219
> cor(FAA$speed_ground,FAA$distance)
[1] 0.8677115
```

2. Linear Regression:

The code:

```
#Regression Model
mode<-lm(FAA$distance~FAA$speed_ground+FAA$speed_air)
summary(mode)
```

=> The output:

```
Call:
lm(formula = FAA$distance ~ FAA$speed_ground + FAA$speed_air)

Residuals:
    Min       1Q   Median       3Q      Max
-820.6  -182.0    7.7   204.2   633.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5425.49     209.08  -25.950  < 2e-16 ***
FAA$speed_ground  -12.32      12.98   -0.949   0.344
FAA$speed_air     91.63      13.20    6.941 5.82e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 276.5 on 192 degrees of freedom
(586 observations deleted due to missingness)
Multiple R-squared:  0.8902,    Adjusted R-squared:  0.889
F-statistic: 778.1 on 2 and 192 DF,  p-value: < 2.2e-16
```

=> Observations/Conclusion:

- R failed to run correlation with missing speed_air values, so I created another dataset only with available speed_air value to run correlation then.
- We found the correlation values to be 0.94 and 0.87, very closed to 1, therefore suggesting a very likely positive linear relationship. We can also see this in the graphs in the previous section.
- After processing Linear regression, we got the formula:

$$\text{Distance} = -5425.49 + 91.63 \text{ Air speed} - 12.32 \text{ Ground speed (feet)}$$

- We found that the P-value are less than 0.001 for intercept and speed_air correlations, however P-value for speed_ground 0.34 > 0.05, so we may not be able to reject the hypothesis that this variable is not statistically significant, and therefore does not affect the change in landing distance. We will have to model check after this step.
- R-squared = 0.8902, thus the model can explain roughly 90% of all observations, which is pretty accurate

=> It's fair to assume that the variables speed_air and speed_ground will fit our model

***5. Aircraft Makers Examination**

The goal: Consider if once we do model fitting of dataset based on individual aircraft makers, we could learn any new insight on how factors would impact landing distance

The code:

#4. Consider individual Aircraft makers

```
FAA_Boeing<-subset(FAA, aircraft == "boeing")
```

```
FAA_Airbus<-subset(FAA, aircraft == "airbus")
```

#Correlation Matrix

```
cor(FAA_Boeing$speed_ground,FAA_Boeing$distance)
```

```
cor(FAA_Boeing$height,FAA_Boeing$distance)
```

```
cor(FAA_Boeing$duration,FAA_Boeing$distance)
```

```
cor(FAA_Boeing$no_pasg,FAA_Boeing$distance)
```

```
cor(FAA_Boeing$pitch,FAA_Boeing$distance)
```

```
CorrFAA_Boeing<-subset(FAA_Boeing, speed_air>=30 & speed_air <= 140)
```

```
cor(CorrFAA_Boeing$distance,CorrFAA_Boeing$speed_air)
```

```
cor(FAA_Airbus$speed_ground,FAA_Airbus$distance)
```

```
cor(FAA_Airbus$height,FAA_Airbus$distance)
```

```
cor(FAA_Airbus$duration,FAA_Airbus$distance)
```



```
cor(FAA_Airbus$no_pasg,FAA_Airbus$distance)
cor(FAA_Airbus$pitch,FAA_Airbus$distance)
```

```
CorrFAA_Airbus<-subset(FAA_Airbus, speed_air>=30 & speed_air <= 140)
cor(CorrFAA_Airbus$distance,CorrFAA_Airbus$speed_air)
```

#Regression Model

```
modelBoeing<-
```

```
lm(FAA_Boeing$distance~FAA_Boeing$speed_ground+FAA_Boeing$speed_air)
```

```
summary(modelBoeing)
```

```
modelAirbus<-lm(FAA_Airbus$distance~FAA_Airbus$speed_ground+FAA_Airbus$speed_air)
```

```
summary(modelAirbus)
```

```
modelBoeing<-lm(FAA_Boeing$distance~FAA_Boeing$speed_air)
```

```
summary(modelBoeing)
```

The output:

```
> cor(FAA_Boeing$speed_ground,FAA_Boeing$distance)
[1] 0.9005007
> cor(FAA_Boeing$height,FAA_Boeing$distance)
[1] 0.06920058
> cor(FAA_Boeing$duration,FAA_Boeing$distance)
[1] -0.01064306
> cor(FAA_Boeing$no_pasg,FAA_Boeing$distance)
[1] -0.01785459
> cor(FAA_Boeing$pitch,FAA_Boeing$distance)
[1] -0.06504457
>
> CorrFAA_Boeing<-subset(FAA_Boeing, speed_air>=30 & speed_air <= 140)
> cor(CorrFAA_Boeing$distance,CorrFAA_Boeing$speed_air)
[1] 0.9775984
~
```

```

>
> cor(FAA_Airbus$speed_ground,FAA_Airbus$distance)
[1] 0.9087595
> cor(FAA_Airbus$height,FAA_Airbus$distance)
[1] 0.15858
> cor(FAA_Airbus$duration,FAA_Airbus$distance)
[1] -0.07850646
> cor(FAA_Airbus$no_pasg,FAA_Airbus$distance)
[1] -0.002610453
> cor(FAA_Airbus$pitch,FAA_Airbus$distance)
[1] 0.04134194
>
> CorrFAA_Airbus<-subset(FAA_Airbus, speed_air>=30 & speed_air <= 140)
> cor(CorrFAA_Airbus$distance,CorrFAA_Airbus$speed_air)
[1] 0.9652658
>
> #Regression Model
> modelBoeing<-lm(FAA_Boeing$distance~FAA_Boeing$speed_ground+FAA_Boeing$speed_air)
> summary(modelBoeing)

```

Call:

```
lm(formula = FAA_Boeing$distance ~ FAA_Boeing$speed_ground +
    FAA_Boeing$speed_air)
```

Residuals:

Min	1Q	Median	3Q	Max
-489.00	-120.86	5.78	133.74	429.13

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5378.1945	167.1072	-32.184	< 2e-16 ***
FAA_Boeing\$speed_ground	-0.8152	11.6095	-0.070	0.944
FAA_Boeing\$speed_air	81.2692	11.7317	6.927	2.62e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 188 on 115 degrees of freedom

(269 observations deleted due to missingness)

Multiple R-squared: 0.9557, Adjusted R-squared: 0.9549

F-statistic: 1240 on 2 and 115 DF, p-value: < 2.2e-16

```
> modelAirbus<-lm(FAA_Airbus$distance~FAA_Airbus$speed_ground+FAA_Airbus$speed_air)
> summary(modelAirbus)
```

Call:

```
lm(formula = FAA_Airbus$distance ~ FAA_Airbus$speed_ground +
    FAA_Airbus$speed_air)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-607.43 -129.92   -8.86  134.61  402.78
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5977.15     266.16  -22.457  < 2e-16 ***
FAA_Airbus$speed_ground    -21.49      13.31   -1.615    0.111
FAA_Airbus$speed_air     103.68      13.76    7.534 9.81e-11 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 183.8 on 74 degrees of freedom

(317 observations deleted due to missingness)

Multiple R-squared: 0.9341, Adjusted R-squared: 0.9323

F-statistic: 524.1 on 2 and 74 DF, p-value: < 2.2e-16

=> Observation/ Conclusion:

- Once we ran individual correlation on each dataset, we found out the same thing: For each Airbus and Boeing, only Speed_air and Speed_ground affect the Landing distance. We run Linear regression for these variables.
- We found equally high R-square values for both models, thus these models explain much of the correlation of data in these datasets.
- For Boeing, we found:

$$\text{Distance} = -5378.19 + 81.27 \text{ Air speed} - 0.8152 \text{ Ground speed (feet)}$$

- For Airbus, we found:

$$\text{Distance} = -5977.15 + -21.49 \text{ Air speed} - 103.68 \text{ Ground speed (feet)}$$

The results are as expected and basically very closed to the model resulting from the combined datasets. P-value for speed_ground for Boeing 0.944, very closed to 1, so we aren't be able to reject the hypothesis that this variable is not statistically significant, and therefore does not affect the change in landing distance.

We remodel the Boeing regression only using Airspeed:

$$\text{Distance} = -5378.033 + 80.453 \text{ Air speed (feet)}$$

=> Output:

```
>
> modelBoeing<-lm(FAA_Boeing$distance~FAA_Boeing$speed_air)
> summary(modelBoeing)
```

Call:

```
lm(formula = FAA_Boeing$distance ~ FAA_Boeing$speed_air)
```

Residuals:

Min	1Q	Median	3Q	Max
-490.22	-121.89	5.85	132.98	428.37

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5378.033	166.373	-32.33	<2e-16 ***
FAA_Boeing\$speed_air	80.453	1.608	50.02	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 187.2 on 116 degrees of freedom
(269 observations deleted due to missingness)

Multiple R-squared: 0.9557, Adjusted R-squared: 0.9553

F-statistic: 2502 on 1 and 116 DF, p-value: < 2.2e-16