Tess Newkold
Jan 27th 2019
Project 1 Part 1

R Code – Green Output – Blue **Observation – RED** Conclusion – Black

Initial Exploration of the Data

Step 1.

```
FAA1 <-
read_excel("/Users/Tess/Documents/Tess/Ohio/UniversityOfCincinnati/BuisnessAnalytics/Spring2019/StatisticalModeling/Week1/FAA1.xls")
FAA2 <-
read_excel("/Users/Tess/Documents/Tess/Ohio/UniversityOfCincinnati/BuisnessAnalytics/Spring2019/StatisticalModeling/Week1/FAA2.xls")
```

Output – NA, just two data frames made

Both files can be seen in the "environment" section in R

Both files were imported successfully

Step 2.

```
str(FAA1)
str(FAA2)
```

Classes 'tbl_df', 'tbl' and '**data.frame':** **800 obs. of 8 variables:**

$ aircraft : **chr** "boeing" "boeing" "boeing" "boeing" ...

$ duration : **num** 98.5 125.7 ….etc

Classes 'tbl_df', 'tbl' and 'data.frame': **150 obs. of 7 variables:**

$ aircraft : chr "boeing" "boeing" "boeing" "boeing" ...

$ no_pasg : num 53 69 61 56 ….etc

Both are data frames, and both are all numeric variables except aircraft which is a chr. FAA2 has one less variable than FAA1.

FAA1 has 800 observations and 8 variables

FAA2 has 150 observations and 7 variables

FAA2 has one less variable than FAA1, if you view() them you can see FAA2 is missing the variable duration.

Step 3.

```
combinedFAA <- merge(FAA1, FAA2, all = TRUE)
combinedFAA[duplicated(combinedFAA$speed_ground),]
[1] aircraft    no_pasg     speed_ground
[4] speed_air   height      pitch
[7] distance    duration
<0 rows> (or 0-length row.names)
```

Once merged, there are 8 variables in the new data frame, and there are 0 rows that are duplicated.

The duplications were taken care of in the merge( ) function. But I double checked them in the following line of code and there are 0 duplications.

Step 4.

```
str(combinedFAA)
summary(combinedFAA)
sd(combinedFAA$no_pasg)
sd(combinedFAA$speed_ground)
sd(combinedFAA$speed_air, na.rm = TRUE)
sd(combinedFAA$height)
sd(combinedFAA$pitch)
sd(combinedFAA$distance)
sd(combinedFAA$duration, na.rm = TRUE)
```

```
'data.frame':   850 obs. of  8 variables:
 $ aircraft    : chr  "airbus" "airbus" "airbus" "airbus" ...
 $ no_pasg     : num  36 38 40 ….etc
```

There are 850 observations and 8 variables

Summary Statistics

	Mean	SD	MIN,Max	%Missing
No_pasg	60.1	7.49	29, 87	0
Speed_ground	79.45	19.06	27.74, 141.22	0
Speed_air	103.80	10.26	90, 141.72	75.5
Height	30.144	10.29	-3.55, 59.95	0
Pitch	4.01	0.53	2.28, 5.927	0
Distance	1526.95	928.6	34.08, 6533.1	0
Duration	154.01	49.26	14.76, 305.62	5.88

Conclusion: merge seems successful, summary statistics look good, with some irregularities which will be removed in further steps.

Step 5.

- There are 850 observations and 8 variables in the combined data set.
- There are a significant number of missing values in the Speed_Air variable, there are a few in duration, the rest do not have any missing.
- Some of the variables look a little messy, but they will be cleaned up in further steps

Data Cleaning and Further Exploration

Step 6.

```
filterFAA <- filter(combinedFAA, duration > 40)
filterFAA1 <- filter(filterFAA, speed_ground > 30 | speed_ground < 140)
filterFAA2 <- filter(filterFAA1, speed_air > 30 | speed_ground < 140)
filterFAA3 <- filter(filterFAA2, height > 6)
cleanFAA <- filter(filterFAA3, distance < 6000)
```

67 observations were removed for the final clean data set.

Step 7.

```
str(cleanFAA)
summary(cleanFAA)
sd(cleanFAA$no_pasg)
sd(cleanFAA$speed_ground)
sd(cleanFAA$speed_air, na.rm = TRUE)
sd(cleanFAA$height)
sd(cleanFAA$pitch)
sd(cleanFAA$distance)
sd(cleanFAA$duration, na.rm = TRUE)
'data.frame':   783 obs. of  8 variables:
 $ aircraft    : chr  "airbus" "airbus" "airbus" "airbus" ...
 $ no_pasg     : num  36 38 .....etc
```

There are now 783 observations and 8 variables

Summary Statistics

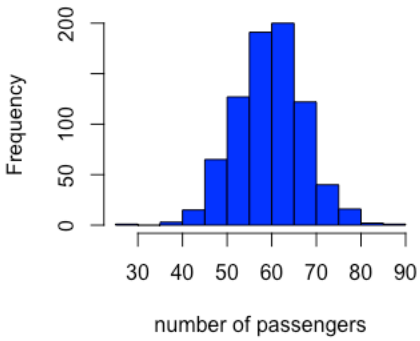
	Mean	SD	MIN,Max	%Missing	% Poor Data
No_pasg	60.07	7.53	29, 87	0	6.47
Speed_ground	79.51	19.05	27.74, 132.78	0	0
Speed_air	103.50	9.88	90, 132.91	75.1	0
Height	30.438	9.73	6.23, 59.95	0	1.2
Pitch	4.015	0.52	2.28, 5.927	0	0
Distance	1540.33	903.6	41.72, 5381.9	0	0
Duration	154.83	48.33	41.95, 305.62	0	0
					7.67 – Total % of Poor Data

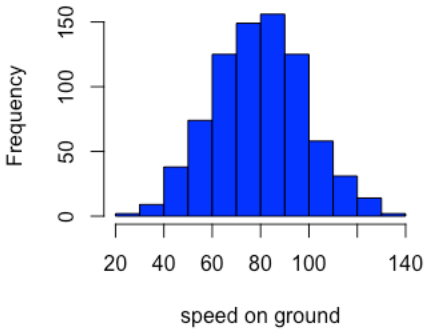
Conclusion: clean up looks like it went well. Irregularities seem to be all gone, all of the height observations are positive. All variables are in an acceptable range. SD are pretty high, im sure standardizing variables later will help.

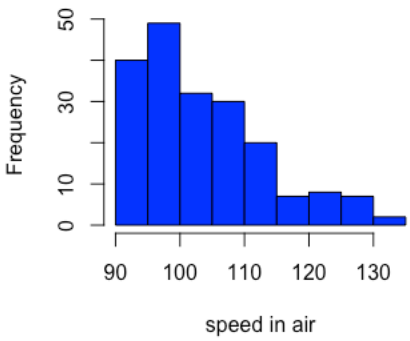
Step 8.

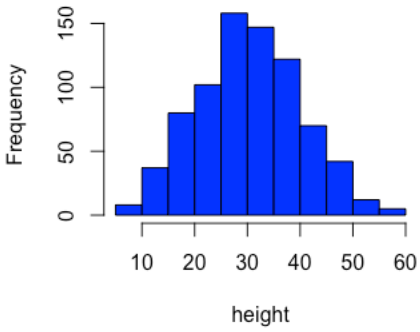
```
hist(cleanFAA$no_pasg, main = " ", xlab = "number of passengers", col = "blue")
hist(cleanFAA$speed_ground, main = " ", xlab = "speed on ground", col = "blue")
hist(cleanFAA$speed_air, main = " ", xlab = "speed in air", col = "blue")
hist(cleanFAA$height, main = " ", xlab = "height", col = "blue")
hist(cleanFAA$pitch, main = " ", xlab = "pitch", col = "blue")
hist(cleanFAA$distance, main = " ", xlab = "distance", col = "blue")
hist(cleanFAA$duration, main = " ", xlab = "duration", col = "blue")
```

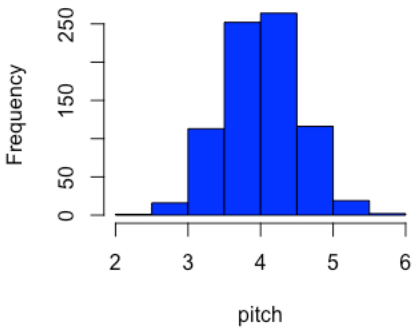
Histograms for all variables

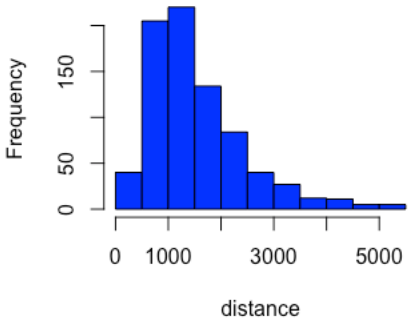


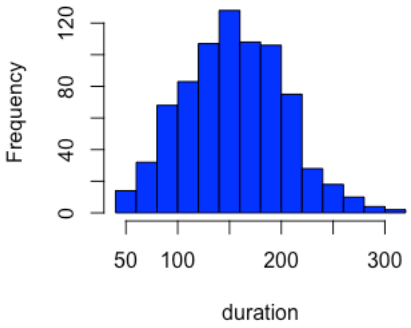












Step 9.

- About 7.67% of the data was considered ‘poor’ and thus removed, that was 67 observations removed resulting in a clean data set of 783 observations and 8 variables
- High percentage of missing variables in Speed in Air variable
- Histograms look good with speed_in_air and distance being skewed to the right.
- Standard deviation of variables looks high, so will probably want to standardize variables for the model

Initial Analysis for Identifying Important Factors that Impact the Response Variable "Landing Distance"

Step 10.

```
dummyFAA <- dummy_cols(cleanFAA, select_columns = "aircraft")
corData <- dummyFAA[,2:length(dummyFAA)]
round(cor(corData),2)
cor(cleanFAA$distance, cleanFAA$speed_air, use = "complete.obs")
```

Output – Table 1

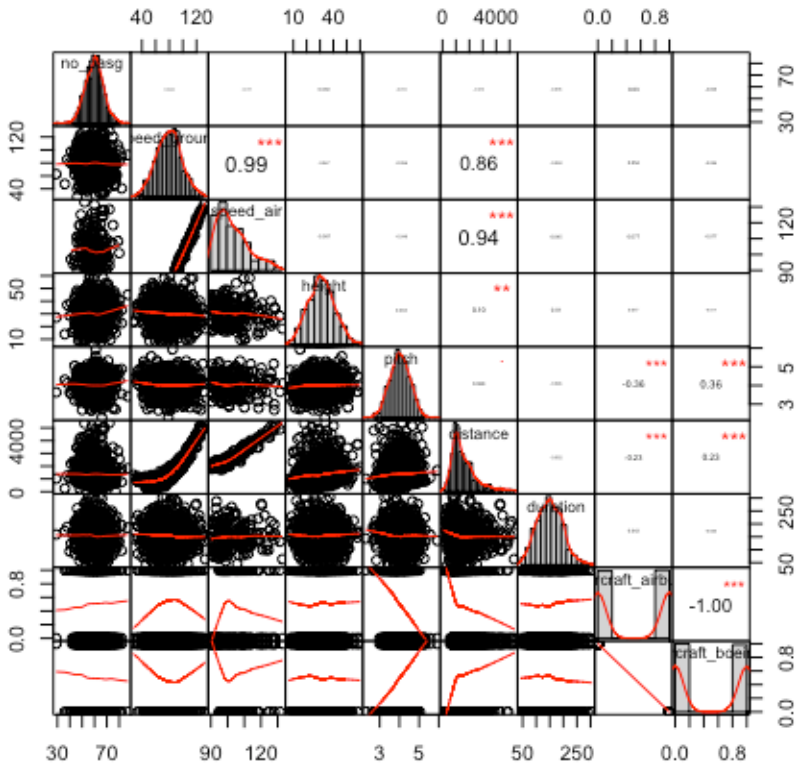
Table 1.

Variable	Correlation	Direction of Correlation (+/-)
Speed of Air	0.94	+
Speed on Ground	0.86	+
Aircraft – Airbus	0.23	-
Aircraft – Boeing	0.23	+
Height	0.10	+
Pitch	0.07	+
Duration	0.05	-
Number of Passengers	0.02	-

Conclusions – all correlations correspond with expected outcome. Order of size of correlation is in a logical order from top to bottom.

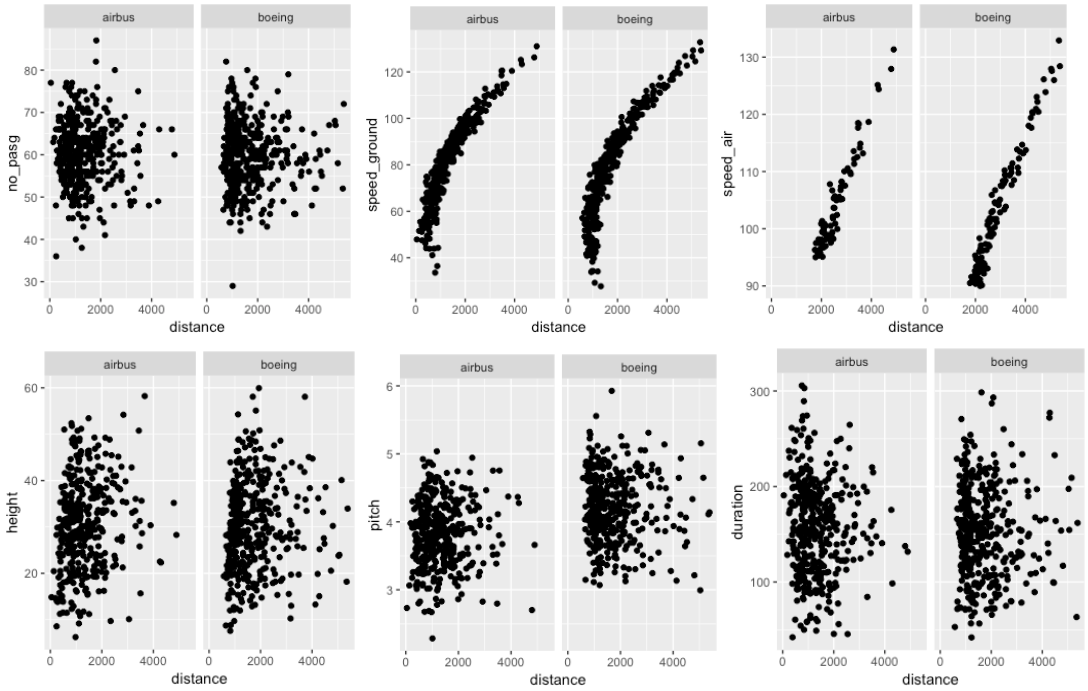
Step 11.

```
myData <- dummyFAA[, c(2,3,4,5,6,7,8,9,10)]
chart.Correlation(myData, histogram = TRUE, pch = 19, na.rm = TRUE)
```



The correlation strength observed in scatter plots are consistent with the values in Table 1.

Step 12.
Yes, I included the airplane make in the above analysis, in step 10 and 11.
Scatter plots below are also a nice way to visualize any differences between aircraft makes.



Regression Using a Single Factor Each Time
Step 13.

```
#model for number of passengers
modelNoPasg <- lm(cleanFAA$distance ~ no_pasg, data = cleanFAA)
summary(modelNoPasg)
#model for speed on ground
modelSpeedOnGround <- lm(cleanFAA$distance ~ speed_ground, data = cleanFAA)
summary(modelSpeedOnGround)
#model for speed in air
modelSpeedInAir <- lm(cleanFAA$distance ~ speed_air, data = cleanFAA)
summary(modelSpeedInAir)
#model for height
modelHeight <- lm(cleanFAA$distance ~ height, data = cleanFAA)
summary(modelHeight)
#model for pitch
modelPitch <- lm(cleanFAA$distance ~ pitch, data = cleanFAA)
summary(modelPitch)
#model for duration
modelDuration <- lm(cleanFAA$distance ~ duration, data = cleanFAA)
summary(modelDuration)
#model for aircraft
dummy <- as.numeric(cleanFAA$aircraft == "airbus")
modelAircraft <- lm(cleanFAA$distance ~ dummy, data = cleanFAA)
summary(modelAircraft)
```

Sample output for
Call:
lm(formula = cleanFAA$distance ~ no_pasg, data = cleanFAA)

Residuals:
   Min   1Q Median   3Q   Max
-1465.5 -629.0 -263.6   411.8 3865.0
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1657.903   259.783   6.382   3e-10 ***
no_pasg   -1.957   4.291 -0.456   0.648

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 904.1 on 781 degrees of freedom
Multiple R-squared: 0.0002663,	Adjusted R-squared: -0.001014
F-statistic: 0.208 on 1 and 781 DF, p-value: 0.6484

Table 2.

Variable	Size of p-value	Direction of Regression Coef
Speed of Air	2.2e(-16)	+
Speed on Ground	2.2e(-16)	+
Aircraft – Airbus	9.52e(-11)	-
Aircraft – Boeing	9.52e(-11)	+
Height	0.003477	+
Pitch	0.05742	+
Duration	0.1458	-
Number of Passengers	0.6484	-

Step 14.

```
#standarize variables
scaledFAA <- cleanFAA
scaledFAA$no_pasg_scaled <- scale(scaledFAA$no_pasg)[, 1]
scaledFAA$distance_scaled <- scale(scaledFAA$distance)[, 1]
scaledFAA$speed_air_scaled <- scale(scaledFAA$speed_air)[, 1]
scaledFAA$speed_ground_scaled <- scale(scaledFAA$speed_ground)[, 1]
scaledFAA$duration_scaled <- scale(scaledFAA$duration)[, 1]
scaledFAA$height_scaled <- scale(scaledFAA$height)[, 1]
scaledFAA$pitch_scaled <- scale(scaledFAA$pitch)[, 1]
#models with standardized variables
modelStandardNoPasg <- lm(scaledFAA$distance_scaled ~ scaledFAA$no_pasg_scaled)
modelStandardSpeedAir <- lm(scaledFAA$distance_scaled ~ scaledFAA$speed_air_scaled)
modelStandardDuration <- lm(scaledFAA$distance_scaled ~ scaledFAA$duration_scaled)
modelStandardHeight <- lm(scaledFAA$distance_scaled ~ scaledFAA$height_scaled)
modelStandardPitch <- lm(scaledFAA$distance_scaled ~ scaledFAA$pitch_scaled)
modelStandardSpeedGround <- lm(scaledFAA$distance_scaled ~ scaledFAA$speed_ground_scaled)
dummy_standard <- as.numeric(scaledFAA$aircraft == "airbus")
modelStandardAircraft <- lm(scaledFAA$distance ~ dummy_standard, data = scaledFAA)
```

Table 3.

Variable (Standardized)	Regression Coef	Direction of Regression Coef
Speed of Air	0.87	+
Speed on Ground	0.86	+
Height	0.1043	
Pitch	0.06794	+
Duration	0.05203	-
Number of Passengers	0.01632	-

Conclusion: the level of significance correlated to the models that are not standardized, as well as the direction of the coefficients.

Step 15.

Tables 1, 2, and 3 all have the variable in the same order of significance. This is obviously a good thing, and will impact the addition of variables in model development.

FAA Meeting

- Speed of the plane in the air has the strongest correlation to landing distance with the highest level of significance.
- There could be more than one variable that has a significant impact on the model and thus landing distance.
- There may be a difference in the make of the plane with boeing having a larger correlation to landing distance.
- Below in Table 0 the variables impacting landing distance are listed in order of significance from top to bottom.

Table 0. Variables listed in order of importance

Variable
Speed of Air
Speed on Ground
Height
Pitch
Duration
Number of Passengers

Check Collinerity

Step 16.

```
#Model 1
summary(modelStandardSpeedGround)
#Model 2
summary(modelStandardSpeedAir)
#Model 3
modelStandardSpeedGroundAndAir <- lm(scaledFAA$distance_scaled ~ scaledFAA$speed_ground_scaled +
scaledFAA$speed_air_scaled)
summary(modelStandardSpeedGroundAndAir)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.513e-16	1.812e-02	0.00	1	
scaledFAA$speed_ground_scaled	8.621e-01	1.813e-02	47.54	<2e-16	***

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.37686	0.02191	62.85	<2e-16	***
scaledFAA$speed_air_scaled	0.86647	0.02196	39.45	<2e-16	***

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.7030	0.3442	4.948	1.64e-06	***
scaledFAA$speed_ground_scaled	-0.2597	0.2735	-0.949	0.344	
scaledFAA$speed_air_scaled	1.0019	0.1443	6.941	5.82e-11	***

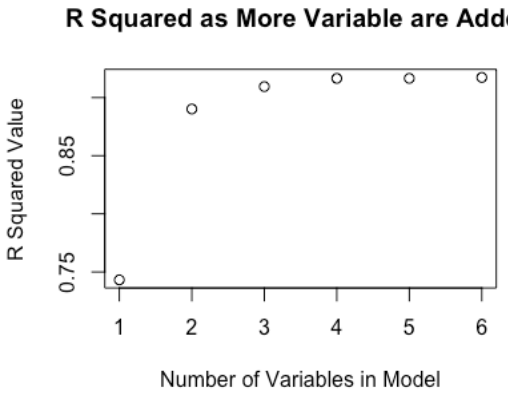
Regression coefficients:
Model 1 – 0.8621
Model 2 – 0.86647
Model 3 - -0.2597 and 1.0019

Conclusion: there is a significant change to spped_ground in model 3 when speed_air is added as the sign of the coefficient for speed_ground goes from positive to negative. The variable speed_air in model 3 doesn't change sign and it also continues to have a significant impact on landing distance.
I would pick speed_air because the variable coefficient remains high even when additional variables are added to the model. It also has the highest significant impact on landing distance.

Variable Selection Based on our Ranking in Table 0
Step 17.

```
#Model 1
modelStandardSpeedGround <- lm(scaledFAA$distance_scaled ~ scaledFAA$speed_ground_scaled)
#Model 2
modelStandardSpeedAir <- lm(scaledFAA$distance_scaled ~ scaledFAA$speed_air_scaled)
#Model 3
model3 <- lm(scaledFAA$distance_scaled ~ speed_ground_scaled + speed_air_scaled + height_scaled, data =
scaledFAA)
#Model 4
model4 <- lm(scaledFAA$distance_scaled ~ speed_ground_scaled + speed_air_scaled + height_scaled +
pitch_scaled, data = scaledFAA)
#Model 5
model5 <- lm(scaledFAA$distance_scaled ~ speed_ground_scaled + speed_air_scaled + height_scaled +
pitch_scaled + duration_scaled, data = scaledFAA)
#Model 6
model6 <- lm(scaledFAA$distance_scaled ~ speed_ground_scaled + speed_air_scaled + height_scaled +
pitch_scaled + duration_scaled + no_pasg_scaled, data = scaledFAA)
```

```
#r squared values for each model
r.squared.1<-summary(modelStandardSpeedGround)$r.squared;
r.squared.2<-summary(modelStandardSpeedGroundAndAir)$r.squared;
r.squared.3<-summary(model3)$r.squared;
r.squared.4<-summary(model4)$r.squared;
r.squared.5<-summary(model5)$r.squared;
r.squared.6<-summary(model6)$r.squared;
#plot r squared values against number of variables added
plot(1:6, c(r.squared.1,r.squared.2,r.squared.3,r.squared.4,r.squared.5,r.squared.6), xlab = "Number of
Variables in Model", ylab = "R Squared Value", main = "R Squared as More Variable are Added")
```

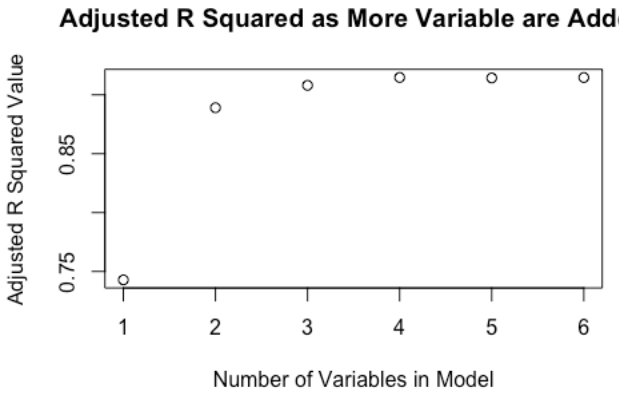


Conclusion: You can observe that the patter of R-squared increases as more variables are added to the model.

Step 18.

```
adjr.squared.1<-summary(modelStandardSpeedGround)$adj.r.squared;
adjr.squared.2<-summary(modelStandardSpeedGroundAndAir)$adj.r.squared;
adjr.squared.3<-summary(model3)$adj.r.squared;
adjr.squared.4<-summary(model4)$adj.r.squared;
adjr.squared.5<-summary(model5)$adj.r.squared;
adjr.squared.6<-summary(model6)$adj.r.squared;

plot(1:6, c(adjr.squared.1,adjr.squared.2,adjr.squared.3,adjr.squared.4,adjr.squared.5,adjr.squared.6), xlab =
"Number of Variables in Model", ylab = "Adjusted R Squared Value", main = "Adjusted R Squared as More
Variable are Added")
```

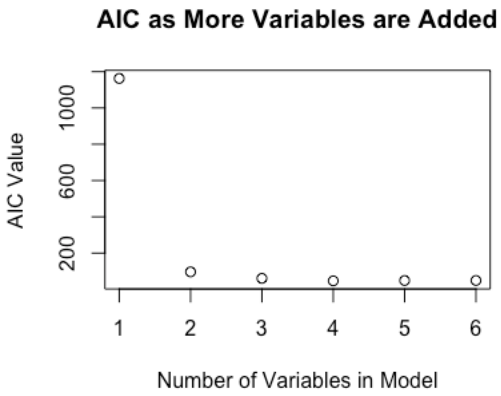


Conlcusion: Adjusted R squared increases and then you can see a slight decrease around the addition of variable 5 and 6, because adjusted r-squared penalizes the addition of variables.

Step 19.

```
AIC1 <- AIC(modelStandardSpeedGround)
AIC2 <- AIC(modelStandardSpeedGroundAndAir)
AIC3 <- AIC(model3)
AIC4 <- AIC(model4)
AIC5 <- AIC(model5)
AIC6 <- AIC(model6)

plot(1:6, c(AIC1, AIC2, AIC3, AIC4, AIC5, AIC6), xlab = "Number of Variables in Model", ylab = "AIC Value", main = "AIC as More Variables are Added")
```



Conclusion: AIC decreases as you continue to add variables to the model.

Step 20.

Comparing the results between adjusted r-squared and AIC, I would choose to add 3 variables to build a predictive model for landing distance. Three is the last variable added that has any significant impact on the adjusted r-squared value, after three variables it tappers off. Same is said for AIC, it tapers off after the addition of the third variable.

Step 21.

```
stepAIC(model6)
Call:
lm(formula = scaledFAA$distance_scaled ~ speed_air_scaled + height_scaled +
   pitch_scaled, data = scaledFAA)

Coefficients:
    (Intercept)  speed_air_scaled     height_scaled
        1.37407          0.88170          0.13694
   pitch_scaled
        0.07272
```

Conclusion: The final model contains three variables, speed_air, height, and pitch. This corresponds to how many variables I thought would be included in the model from the adjusted r-squared and AIC values. However, I would have thought that speed_ground would be included in the model, but perhaps that variable does not impact landing distance as much as previously thought.