Code ▾

# Group Project 29

Chloe Hu, Eric Jones, Yanfeng Shao, Jolie Tran

r Sys.Date()

Hide

```
library(tidyverse)
library(openintro)
library(ggpubr)
library(infer)
```

## Read in Data

Hide

```
Birds <- read_csv("Birds.csv", show_col_types=FALSE)
Birds$Season <- factor(Birds$Season)
Birds$WindDir <- factor(Birds$WindDir)
BirdsC <- Birds[complete.cases(Birds),]
```

## Problem 1: Group Work

We are working online and we used divide and conquer.

Problem 2: by Yanfeng Shao

Problem 3: by Yanfeng Shao, Jolie Tran

Problem 4: by Eric Jones

Problem 5: by Eric Jones

Problem 6: by Jolie Tran

Problem 7: by Chloe Hu

## Problem 2: Population (Yanfeng Shao)

The population of the data set would be the birds that migrates along the newly installed transmission line around New Jersey during spring or fall.

# Problem 3: Explore Season and WindDir (Yanfeng Shao, Jolie Tran)

## Part a)

Hide

```
# Raw Counts - Table
BirdsC %>%
  count(Season, WindDir) %>%
  pivot_wider(id_cols = c(Season, WindDir),
              names_from = WindDir, values_from = n)
```

```
## # A tibble: 2 x 4
##   Season   `1`   `2`   `3`
##   <fct>  <int> <int> <int>
## 1 Fall     278    16   288
## 2 Spring   228    14    97
```

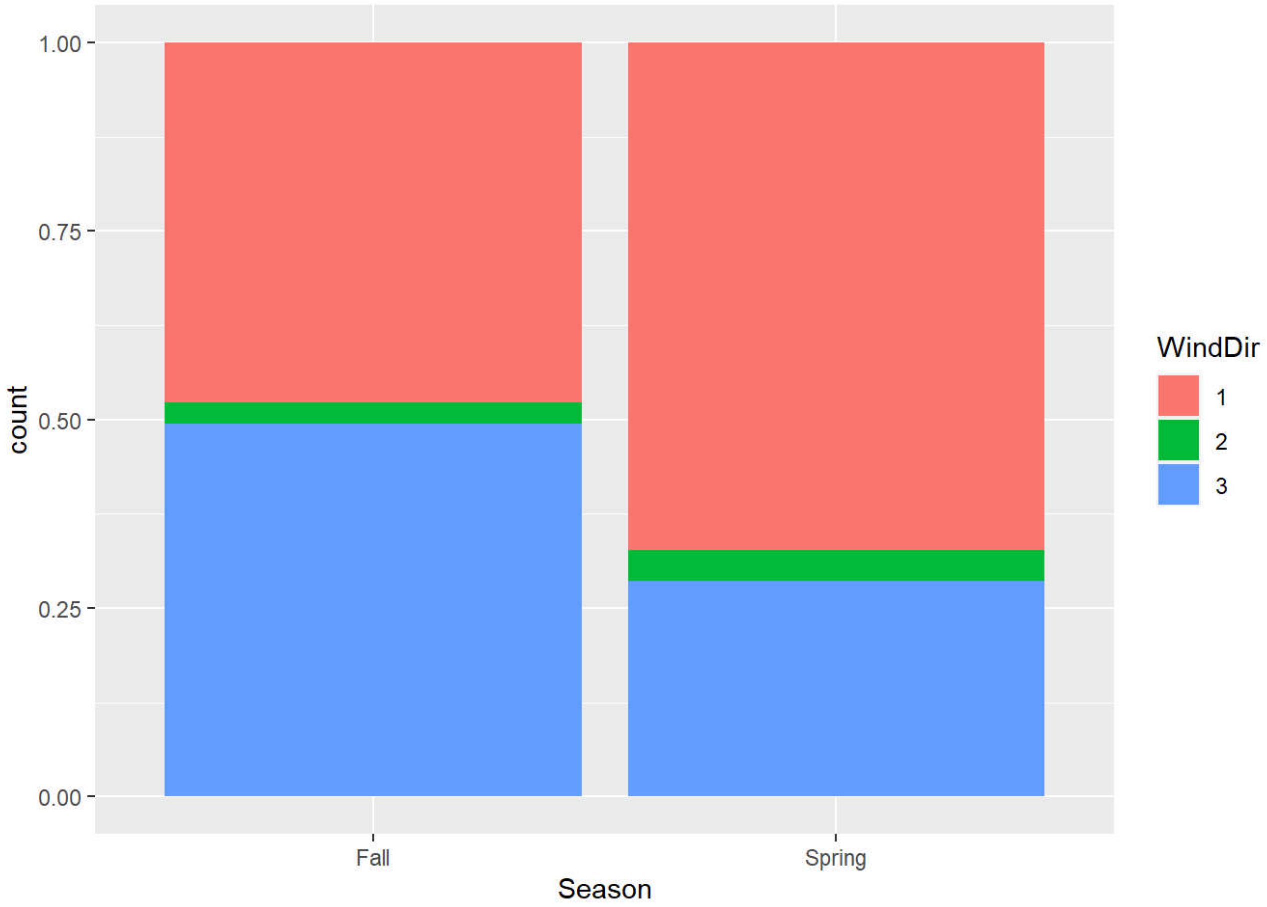
Hide

```
# Percentages Conditioned on Season - Table
BirdsC %>%
  count(Season, WindDir) %>%
  group_by(Season) %>%
  mutate(Prop = n / sum (n) * 100) %>%
  pivot_wider(id_cols = c(Season, WindDir),
              names_from = WindDir, values_from = Prop)
```

```
## # A tibble: 2 x 4
## # Groups:   Season [2]
##   Season   `1`   `2`   `3`
##   <fct>  <dbl> <dbl> <dbl>
## 1 Fall    47.8  2.75  49.5
## 2 Spring  67.3  4.13  28.6
```

Hide

```
# Bar Plot
ggplot(BirdsC, aes(x = Season, fill = WindDir)) +
  geom_bar(position = "fill")
```



The Season and WindDir appear to be dependent.

According to the tables and the graph, the proportion of tailwind (1) and crosswind (3) are roughly equal and they also appear to be the main wind directions in Fall while in Spring, tailwind (1) is the most common wind direction, followed by crosswind (3), then headwind (2). In either season, headwind (2) is the least common wind direction.

Generally, the tailwind (1) occurs more often in Spring than in Fall, and the crosswind (3) occurs more often in Fall than in Spring.

## Part b)

Hide

```
#From the contingency table, we know that the counts for crosswind (3) in Fall and Spri
         ng are 288 and 97 respectively, and the total counts for different wind direct
         ions for Fall and Spring are 582 and 339 respectively.
prop.test (x = c(288, 97), n = c(582, 339), conf.level = 0.95, correct = FALSE)$conf.in
         t
```

```
## [1] 0.1457448 0.2716746
## attr(,"conf.level")
## [1] 0.95
```

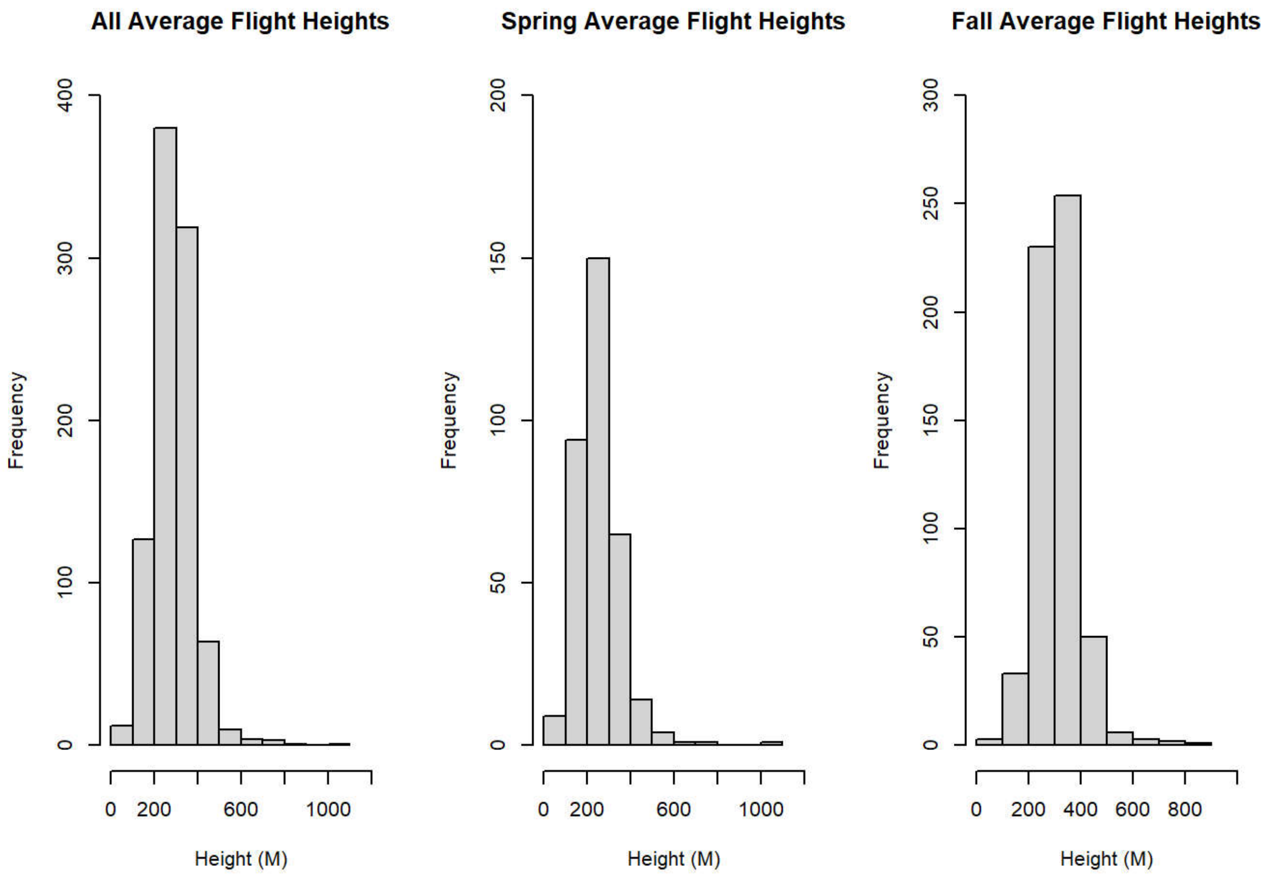
We are 95% confident that the population difference between the proportion of crosswinds in fall and spring falls between about 14.6% and 27.2%.

Since both the lower and upper bounds of the CI are strictly greater than 0, this suggest that the two proportions are in fact different and that the percentage of crosswind in Fall is in fact greater than the percentage of that in Spring.

# Problem 4: Explore Avg Flight Height (Eric Jones)

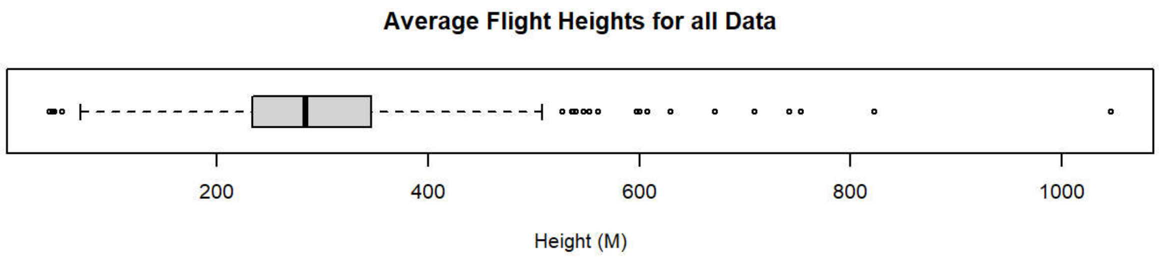
Hide

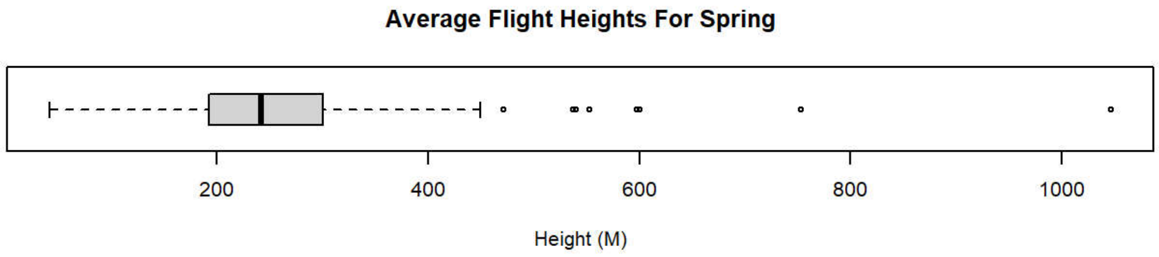
```
par(mfrow=c(1,3))
Spring <- filter(BirdsC, Season == "Spring")
Fall <- filter(BirdsC, Season == "Fall")
hist(BirdsC$Height, main = "All Average Flight Heights", xlab = "Height (M)", xlim = c(
         0, 1200), ylim = c(0, 400))
hist(Spring$Height, main = "Spring Average Flight Heights", xlab = "Height (M)", xlim =
         c(0, 1200), ylim = c(0, 200))
hist(Fall$Height, main = "Fall Average Flight Heights", xlab = "Height (M)", xlim = c(0
         , 1000), ylim = c(0, 300))
```

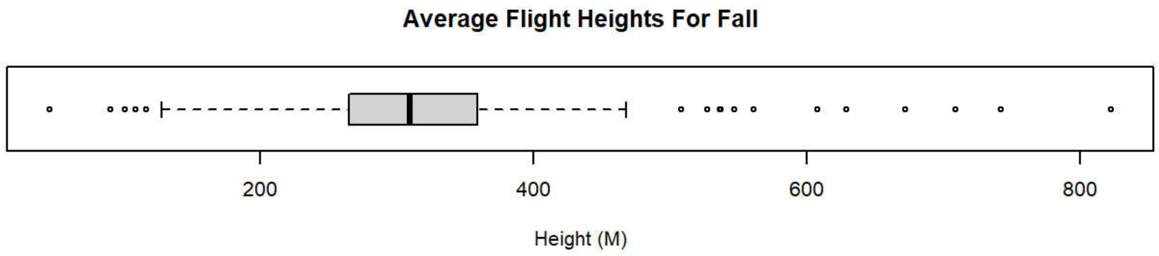


Hide

```
par(mfrow=c(3,1))
boxplot(BirdsC$Height, main = "Average Flight Heights for all Data", xlab = "Height
          (M)", horizontal = TRUE)
boxplot(Spring$Height, main = "Average Flight Heights For Spring", xlab = "Height (M)",
         horizontal = TRUE)
boxplot(Fall$Height, main = "Average Flight Heights For Fall", xlab = "Height (M)", hor
         izontal = TRUE)
```


Average Flight Heights for all Data


Average Flight Heights For Spring


Average Flight Heights For Fall

Hide

```
summary(BirdsC$Height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    42.0   233.9   284.4   291.6   346.1  1047.0
```

Hide

```
summary(Spring$Height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    42.0   192.1   241.5   255.3   300.5  1047.0
```

Hide

```
summary(Fall$Height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    46.0   264.7   309.3   312.8   359.1   823.0
```

The sample distributions for all data displays a unimodal form with a notable right skew. This is a similar case for both spring and fall, however, the fall distribution appears to have slightly less of a skew than the others. In all three distributions, there are apparent outliers. When comparing the fall and spring subsets, we can see that spring featured the absolute maximum height value of 1047m as well as more variability about the mean.

# Problem 5. Estimate Average Flight Height (Eric Jones)

## Part a)

Hide

```
Spring2 <- filter(BirdsC, Season == "Spring", WindDir == "2")
Fall2 <- filter(BirdsC, Season == "Fall", WindDir == "2")
t.test(Spring2$Height, Fall2$Height, conf.level = 0.95)$conf.int
```

```
## [1] -128.02209   23.40745
## attr(,"conf.level")
## [1] 0.95
```

We are 95% confident that the difference in population mean flight height for birds flying into a headwind between Spring and Fall seasons falls within about -128.02m and 23.41m. Since zero is contained in the interval, we cannot rule out that the population mean flight heights for for birds flying into a headwind between Spring and Fall are the same.

## Part b)

Hide

```
Spring1 <- filter(BirdsC, Season == "Spring", WindDir == "1")
Fall1 <- filter(BirdsC, Season == "Fall", WindDir == "1")
t.test(Spring1$Height, Fall1$Height, conf.level = 0.95)$conf.int
```

```
## [1] -67.24056 -34.32893
## attr(,"conf.level")
## [1] 0.95
```

We are 95% confident that the difference in population mean flight height for birds flying with a tailwind between Spring and Fall seasons falls within about -67.24m and -34.33m. Since zero is not contained in this interval, it appears that the mean flight height for birds flying with a tailwind in fall is higher than in spring.

## Part c)

Hide

```
Spring3 <- filter(BirdsC, Season == "Spring", WindDir == "3")
Fall3 <- filter(BirdsC, Season == "Fall", WindDir == "3")
t.test(Spring3$Height, Fall3$Height, conf.level = 0.95)$conf.int
```

```
## [1] -86.90083 -41.94535
## attr(,"conf.level")
## [1] 0.95
```

We are 95% confident that the difference in population mean flight height for birds flying with a crosswind between Spring and Fall seasons falls within about -86.90m and -41.95m. Since zero is not contained in this interval, it appears that the mean flight height for birds flying with a crosswind in fall is indeed higher than in spring.

## Part d)

These three intervals tell us, that in fall, birds tend to fly higher with a tailwind or crosswind than they do in the spring. In regards to flying into a headwind, there may or may not be a difference in mean flight heights between the two seasons.

# Problem 6: Testing a Claim About Heights (Jolie Tran)

## Part a) HT

Hide

```
t.test(Fall$Height, Spring$Height, mu = 0, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  Fall$Height and Spring$Height
## t = 8.8149, df = 600.85, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  46.70606      Inf
## sample estimates:
## mean of x mean of y
##  312.7886  255.3476
```

Null hypothesis: The average flight heights in Fall for migratory birds are equal to that in Spring ( $H_0 : \mu_{h(F)} = \mu_{h(S)} => \mu_{h(F)} - \mu_{h(S)} = 0$)

Alternative hypothesis: The average flight heights in Fall are higher than in Spring for migratory birds ($H_A : \mu_{h(F)} > \mu_{h(S)} => \mu_{h(F)} - \mu_{h(S)} > 0$)

According to the alternative hypothesis, we want to see if the difference between average flight heights in Fall and Spring is greater than the null hypothesized value. So this will be a right-tailed test.

And for the ROT, we will choose the significance level $\alpha = 0.05$ for hypothesis testing.

Running the t-test we got the degree of freedom 600.85 and a p-value < 2.2e-16.

Since the p-value < 0.05, we reject the null. There is sufficient evidence to conclude that the average flight heights in Fall are higher than in Spring.

## Part b) Assumptions

We assume that the difference between the average flight height in Fall and Spring has unimodal, normal distribution, and its significance level is 5%.

We also assume that the data is randomly selected from a representative of the total population.

## Part c) Permutation Test

Hide

```
# Permutation Test
set.seed(25)

PermsOut <- BirdsC %>%
  rep_sample_n(size = nrow(BirdsC), reps = 1000, replace = FALSE) %>%
  mutate(Height_perm = sample(Height)) %>%
  group_by(replicate, Season) %>%
  summarize(prop_Height_perm = mean(Height_perm), prop_Height = mean(Height)) %>%
  # Since the default order of subtraction is Spring - Fall (we know it by looking at t
         he boxplot in Problem 4, we put a negative sign in front of the diff() fuction
         to get the Fall - Spring order of subtraction)
  summarize(diff_perm = -diff(prop_Height_perm), diff_orig = -diff(prop_Height))
```

```
## `summarise()` has grouped output by 'replicate'. You can override using the `.groups
` argument.
```

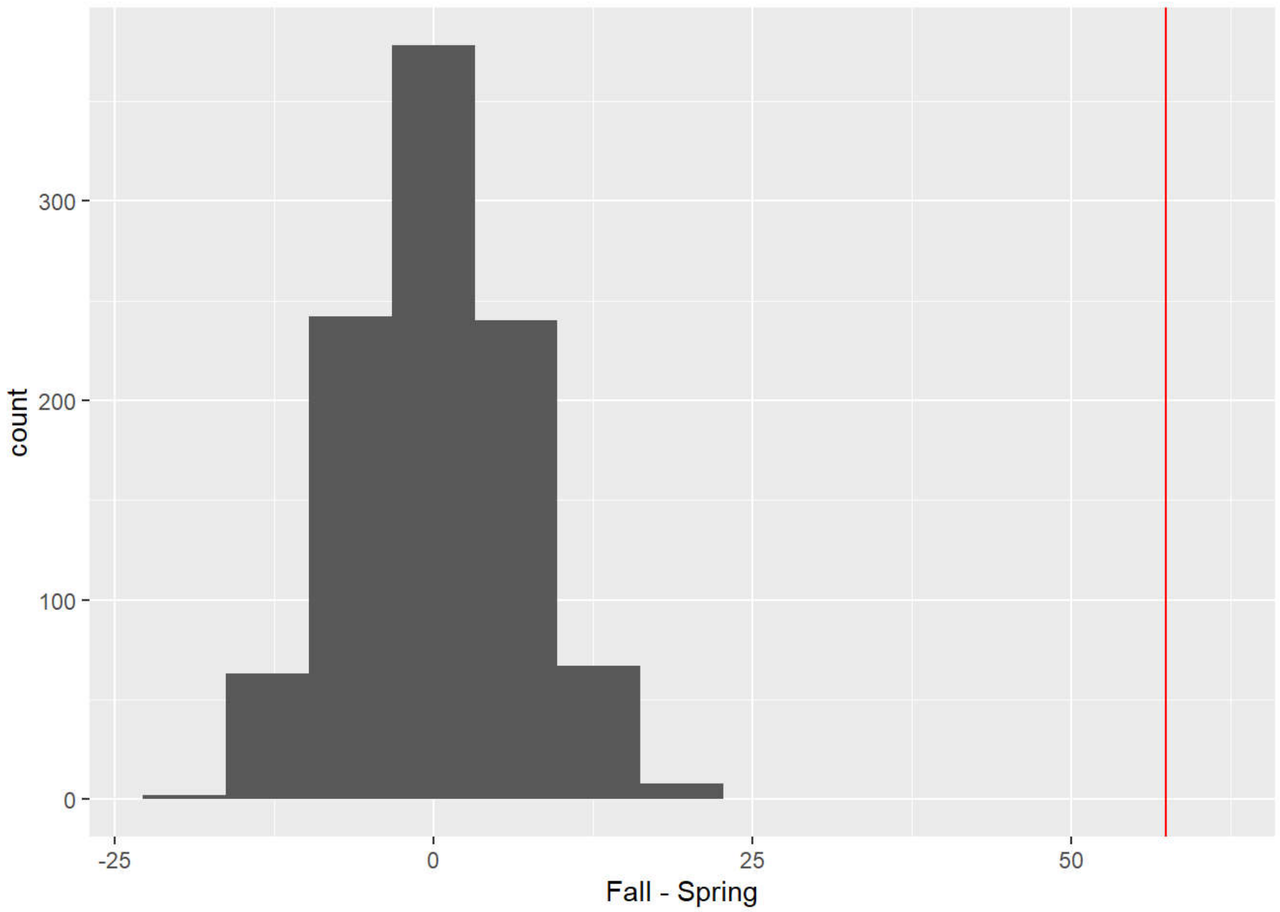
Hide

```
PermsOut
```

```
## # A tibble: 1,000 x 3
##    replicate diff_perm diff_orig
##        <int>     <dbl>     <dbl>
##  1         1     -5.14      57.4
##  2         2     -6.09      57.4
##  3         3     -3.71      57.4
##  4         4     -2.96      57.4
##  5         5     -1.37      57.4
##  6         6      7.77      57.4
##  7         7      1.03      57.4
##  8         8     13.6       57.4
##  9         9      4.22      57.4
## 10        10      1.38      57.4
## # ... with 990 more rows
```

Hide

```
# Histogram of Null Distribution
origdiff <- PermsOut$diff_orig[1]
pPerm <- ggplot(data = PermsOut, aes(x = diff_perm)) +
  geom_histogram(bins = 13) +
  xlab("Fall - Spring") +
  geom_vline(xintercept = origdiff, col="Red")
pPerm
```



Hide

```
# p-value
PermsOut %>%
  summarize(count = sum(diff_orig <= diff_perm),
            proportion = mean(diff_orig <= diff_perm))
```

```
## # A tibble: 1 x 2
##   count proportion
##   <int>      <dbl>
## 1     0          0
```

The shape of the null distribution is fairly normally distributed and the observed sample difference is very far away from the right of the histogram. It appears that there are no values in the null distribution that are more extreme than the observed value and in particular, the p-value from this permutation test is 0.

Since p-value = 0 is smaller than 0.05, we can reject the null hypothesis in favor of the alternative hypothesis.

This result aligns with the result we got in part (a). In both circumstances, the p-value is smaller than the level of significance. Thus, it is sufficient evidence to conclude that the average flight heights in Fall are higher than in Spring.

# Problem 7: Simple Linear Regression (Chloe Hu)

## Part a)

We believe that spring and fall data should be separate in different models. In question 3, we conclude that season and other predictors are dependent and some wind factors are significantly different in spring and in fall. In question 5, we conclude that the flight height varies when with a crosswind and tailwind between spring and fall. All those conclusions show that season and wind factors have correlations. Additionally, the correlations and scatterplots calculated above show clear differences between wind factors and each of their correlations. By using separate linear regression, the model will be best in predicting flight height.

## Part b) Best Predictor Using SLR (ADP, WindSpd, ABP)

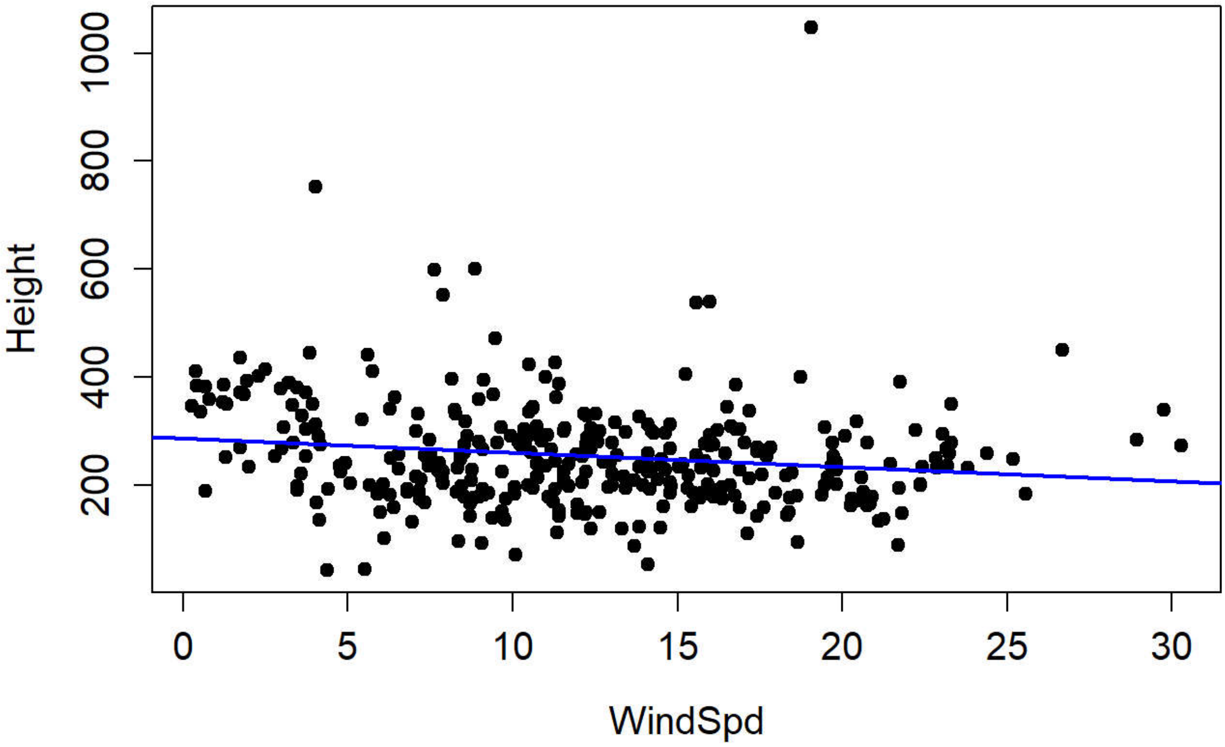
Hide

```
lm.out <- lm(Spring$Height ~ Spring$WindSpd)
summary(lm.out)
```

```
##
## Call:
## lm(formula = Spring$Height ~ Spring$WindSpd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -233.85  -61.11  -13.38   48.34  810.28
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    287.4849    12.1288  23.703  < 2e-16 ***
## Spring$WindSpd  -2.6636     0.8976  -2.968  0.00322 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 100.6 on 337 degrees of freedom
## Multiple R-squared:  0.02547,    Adjusted R-squared:  0.02257
## F-statistic: 8.806 on 1 and 337 DF,  p-value: 0.003217
```

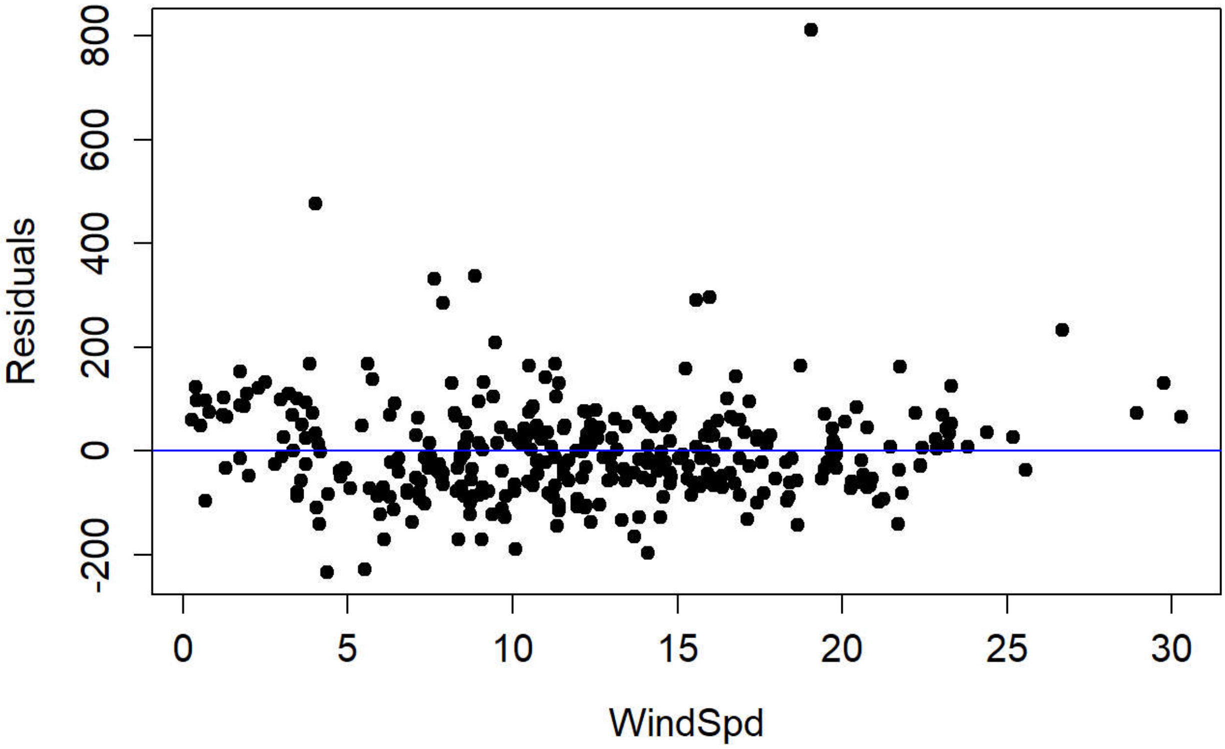
Hide

```
plot(Spring$WindSpd, Spring$Height, xlab="WindSpd", pch=19,
     ylab="Height", cex.lab=1.25, cex.axis=1.25)
abline(lm.out, col="blue", lwd=2)
```



Hide

```
plot(Spring$WindSpd, residuals(lm.out), xlab="WindSpd", pch=19,
     ylab="Residuals", cex.lab=1.25, cex.axis=1.25)
abline(h=0,col="blue")
```



We believe that the model using WindSpd is the best model. First we look at the residual standard error. All three models have a residual standard error around 100. Then we look at the coefficients of determination, squared R, and the model of WindSpd tends to have the highest squared R value. This indicates that the variation of height is explained the most by knowing WindSpd. However, the squared R for ABP comes very close. We then look at the residual plot of the two models and discover that model using WindSpd has fewer outliers and dots bounce around the line relatively more evenly. All together, We believe the model using WindSpd is the best model.

## Part c)

Equation: y = -2.66x + 287.5

Interpretation: The slope -2.66 indicates that on average with 1km/h increase in WindSpd, the flight height will decrease by 2.66 m.

## Part d)

Hide

```
qt(0.975, 919)
```

```
## [1] 1.962549
```

Hide

```
-2.66 + 1.96 * (-2.33)
```

```
## [1] -7.2268
```

Hide

```
-2.66 - 1.96 * (-2.33)
```

```
## [1] 1.9068
```

We are 95% confident that on average, for each 1 degree Celsius increase in ADP, the height will increase between about -7.22 and 1.90 m.

## Part e)

Hide

```
summary(Spring$WindSpd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.270   7.875  11.560  12.065  16.100  30.280
```

Hide

```
lm2.out <- lm(Height ~ WindSpd, data=data.frame(Spring))
new.WT <- data.frame(WindSpd = 16.1)
(CI.WT15 <- predict(lm2.out, newdata = new.WT, interval = "confidence", level = 0.95))
```

```
##        fit      lwr      upr
## 1 244.6013 231.7115 257.4912
```

We are 95% confident that the mean flight height of a bird that has a 16.1 degree Celsius ADP will fall between about 231.71 and 257.49 m.

## Part (f)

Hide

```
lm2.out <- lm(Height ~ WindSpd, data=data.frame(Spring))
new.WT <- data.frame(WindSpd = 16.1)
(PI.WT15 <- predict(lm2.out, newdata = new.WT, interval = "prediction", level = 0.95))
```

```
##        fit      lwr      upr
## 1 244.6013 46.38435 442.8183
```

There is a 95% probability that a randomly selected bird that has a 16.1 degree Celsius ADP will have a height that falls between about 46.38 and 442.8183 m.

## Part g)

Hide

```
Spring2 <- Spring %>% filter(Height < 500)
lm3.out <- lm(Spring2$Height ~ Spring2$WindSpd)
summary(lm3.out)
```

```
##
## Call:
## lm(formula = Spring2$Height ~ Spring2$WindSpd)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -225.471  -53.840   -5.482   53.988  241.969
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     279.1906     9.5325  29.288  < 2e-16 ***
## Spring2$WindSpd  -2.6818     0.7043  -3.808 0.000167 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.3 on 330 degrees of freedom
## Multiple R-squared:  0.04209,    Adjusted R-squared:  0.03919
## F-statistic:  14.5 on 1 and 330 DF,  p-value: 0.0001671
```

Yes. The linear regression model on the modified dataset demonstrate a much larger squared R value, which indicates that there is more variation of height explained by knowing ADP, and also a decreased residual standard error, which indicates smaller average distance from the line. Overall, the model is better at predicting flight height after removing the outliers.

…