Golf Course Manager SDA

Thomas Kulch

Load Data

```
df <- read.csv("../data/processed/golf_cleaned.csv")
summary(df)
```

```
##    player_id         name               round            score      
##  Min.   :  1.0   Length:6434        Min.   :  1.00   Min.   :60.00  
##  1st Qu.: 86.0   Class :character   1st Qu.:  4.00   1st Qu.:70.00  
##  Median :201.0   Mode  :character   Median : 10.00   Median :72.00  
##  Mean   :205.1                      Mean   : 15.73   Mean   :72.56  
##  3rd Qu.:315.0                      3rd Qu.: 21.00   3rd Qu.:75.00  
##  Max.   :466.0                      Max.   :114.00   Max.   :92.00  
##      date              handicap         avg_temp     precipitation    
##  Length:6434        Min.   :-0.900   Min.   :34.34   Min.   : 0.000  
##  Class :character   1st Qu.: 3.800   1st Qu.:55.94   1st Qu.: 0.000  
##  Mode  :character   Median : 4.300   Median :66.02   Median : 0.000  
##                     Mean   : 4.453   Mean   :64.35   Mean   : 3.146  
##                     3rd Qu.: 5.100   3rd Qu.:72.68   3rd Qu.: 1.800  
##                     Max.   :14.800   Max.   :90.14   Max.   :68.100  
##    wind_speed    day_of_week        day_of_week_int 
##  Min.   : 3.11   Length:6434        Min.   :0.000   
##  1st Qu.: 8.26   Class :character   1st Qu.:3.000   
##  Median : 9.82   Mode  :character   Median :5.000   
##  Mean   :10.54                      Mean   :4.158   
##  3rd Qu.:12.30                      3rd Qu.:6.000   
##  Max.   :26.59                      Max.   :6.000
```

```
head(df)
```

```
##   player_id            name round score       date handicap avg_temp
## 1         1 Collin Morikawa     1    72 2017-12-18     -0.9    71.42
## 2         2   Jordan Spieth    24    68 2021-10-11      2.2    78.44
## 3         3 Dylan Frittelli     5    64 2019-12-01      3.5    57.02
## 4         4    Kevin Kisner    14    71 2019-10-11      4.2    51.62
## 5         5  Phil Mickelson    83    77 2022-10-22      3.8    70.52
## 6         6       Adam Long     1    72 2018-04-11      6.6    66.20
##   precipitation wind_speed day_of_week day_of_week_int
## 1           1.3      12.30      Monday               0
## 2           4.8      11.62      Monday               0
## 3           0.0       6.46      Sunday               6
## 4           0.0      15.22      Friday               4
```

```
## 5           0.0       4.72    Saturday               5
## 6           0.0       9.82   Wednesday               2
```

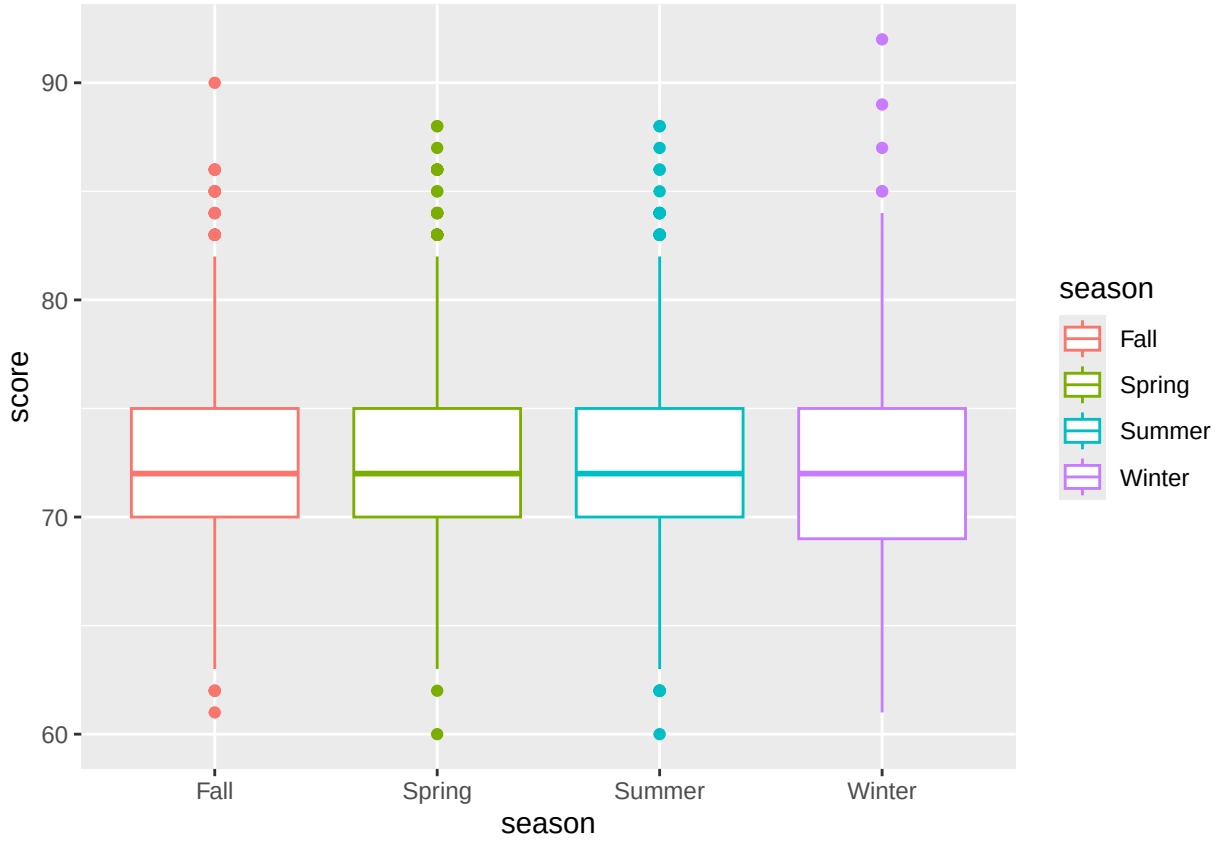
Add season to df for plotting

```
df$date <- as.Date(df$date)
df <- df %>%
  mutate(season = case_when(
    month(date) %in% c(12, 1, 2) ~ "Winter",
    month(date) %in% c(3, 4, 5) ~ "Spring",
    month(date) %in% c(6, 7, 8) ~ "Summer",
    month(date) %in% c(9, 10, 11) ~ "Fall"
  ))
head(df)
```

```
##   player_id            name round score       date handicap avg_temp
## 1         1 Collin Morikawa     1    72 2017-12-18     -0.9    71.42
## 2         2   Jordan Spieth    24    68 2021-10-11      2.2    78.44
## 3         3 Dylan Frittelli     5    64 2019-12-01      3.5    57.02
## 4         4    Kevin Kisner    14    71 2019-10-11      4.2    51.62
## 5         5  Phil Mickelson    83    77 2022-10-22      3.8    70.52
## 6         6       Adam Long     1    72 2018-04-11      6.6    66.20
##   precipitation wind_speed day_of_week day_of_week_int season
## 1           1.3      12.30      Monday               0 Winter
## 2           4.8      11.62      Monday               0   Fall
## 3           0.0       6.46      Sunday               6 Winter
## 4           0.0      15.22      Friday               4   Fall
## 5           0.0       4.72    Saturday               5   Fall
## 6           0.0       9.82   Wednesday               2 Spring
```

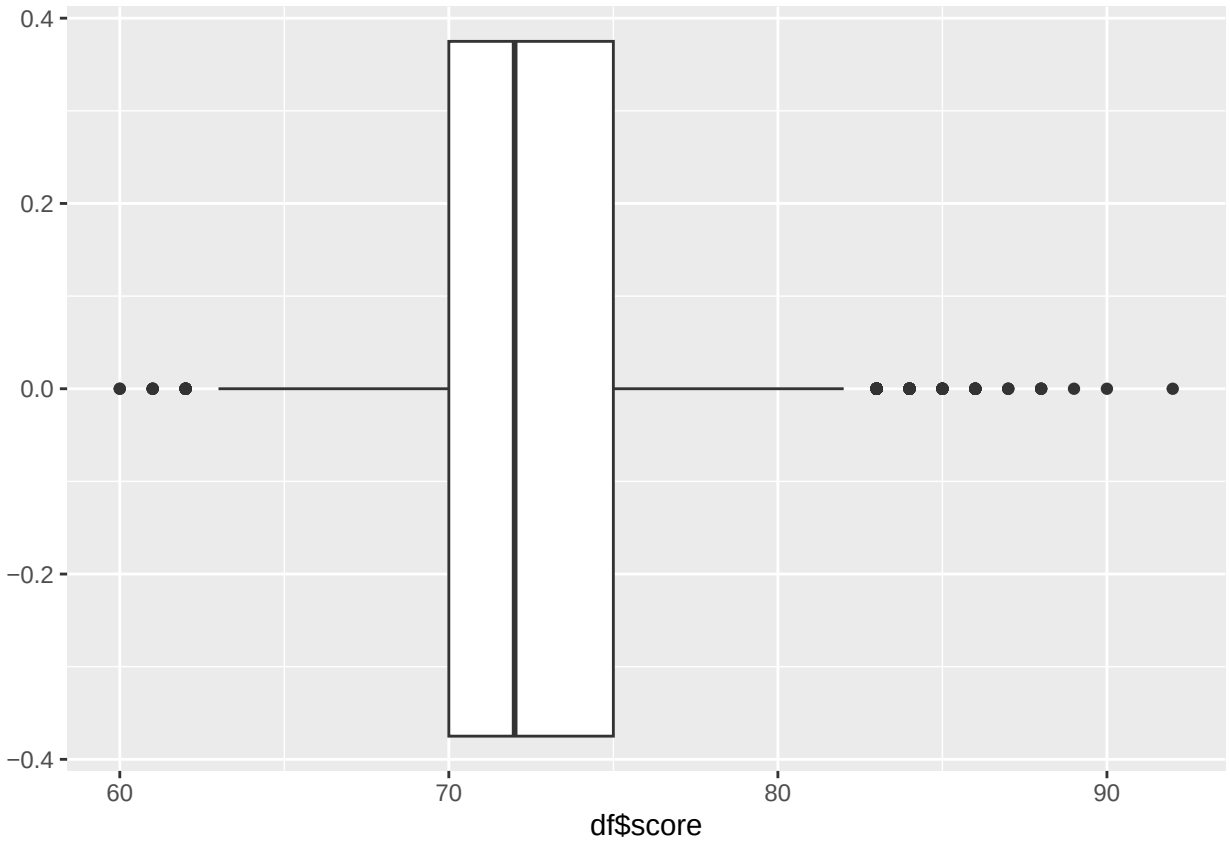
Conditional Boxplots

```
df |>
  ggplot(aes(x=season,y=score,color=season)) + geom_boxplot(orientation="x")
```



The medians are around the same for all seasons. Spring and Summer have virtually identical distributions of scores with more outliers on the lower end. Winter has some higher score outliers and no outliers on the lower end.

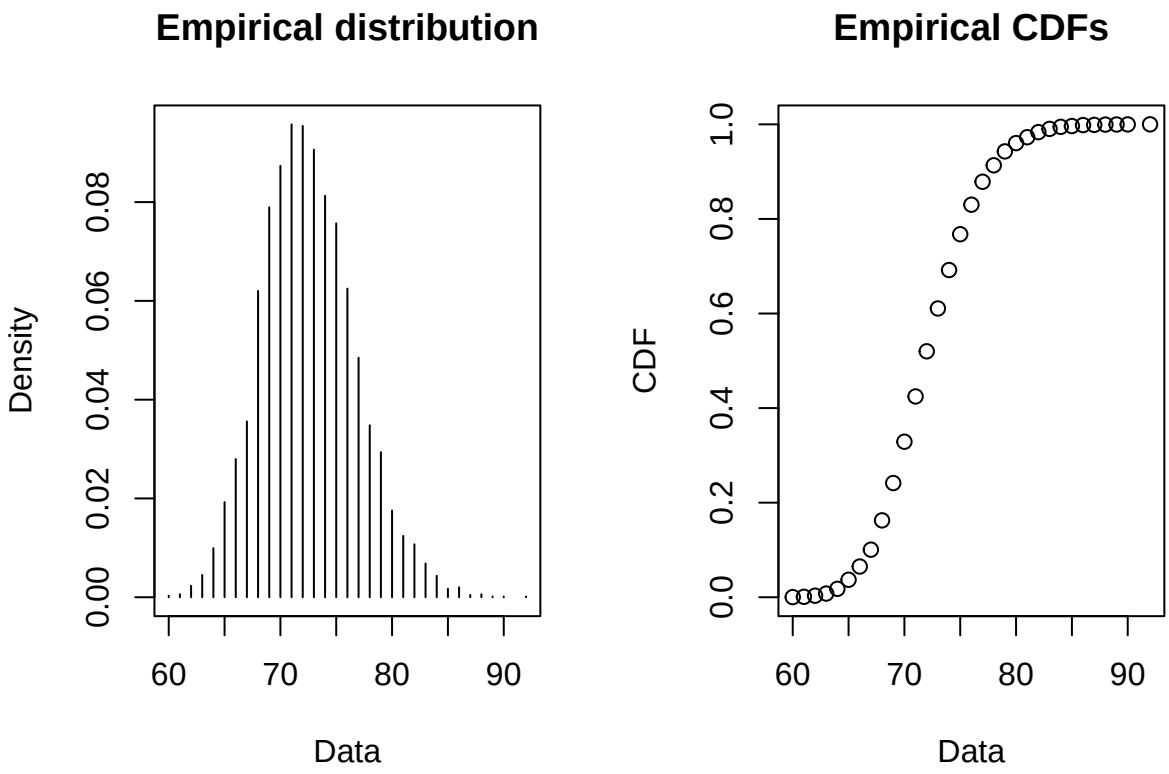
```
ggplot(tibble(measurement=df$score)) +
  geom_boxplot(aes(x=df$score))
```



There are quite a few outliers on the high end, but it's likely that weather or other conditions caused this and we should leave these records in

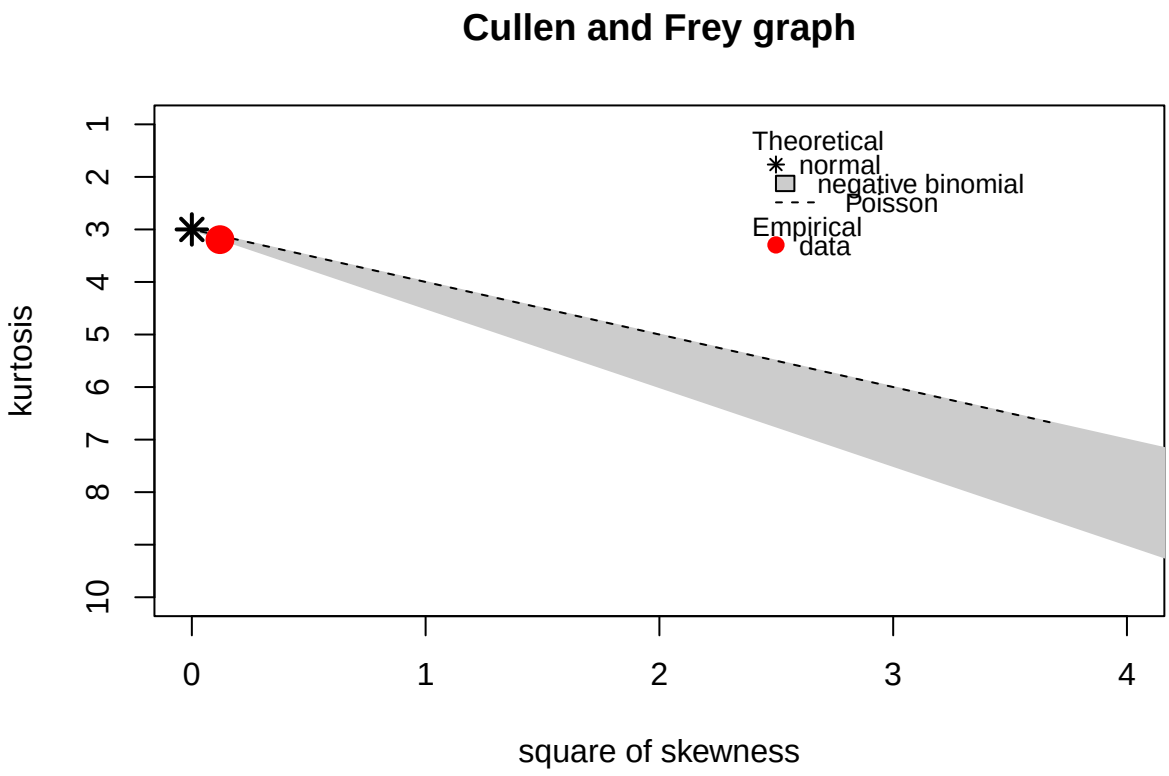
Plot the histogram and empirical CDF

```
plotdist(data=df$score,
         histo=TRUE,
         demp=FALSE,
         discrete=TRUE)
```



Our data looks to be a bit right skewed, but could be normal. The CDF indicates a normal distribution with low standard deviation. We will run the Cullen Frey still.

```
descdist(df$score,
         discrete=TRUE
         )
```

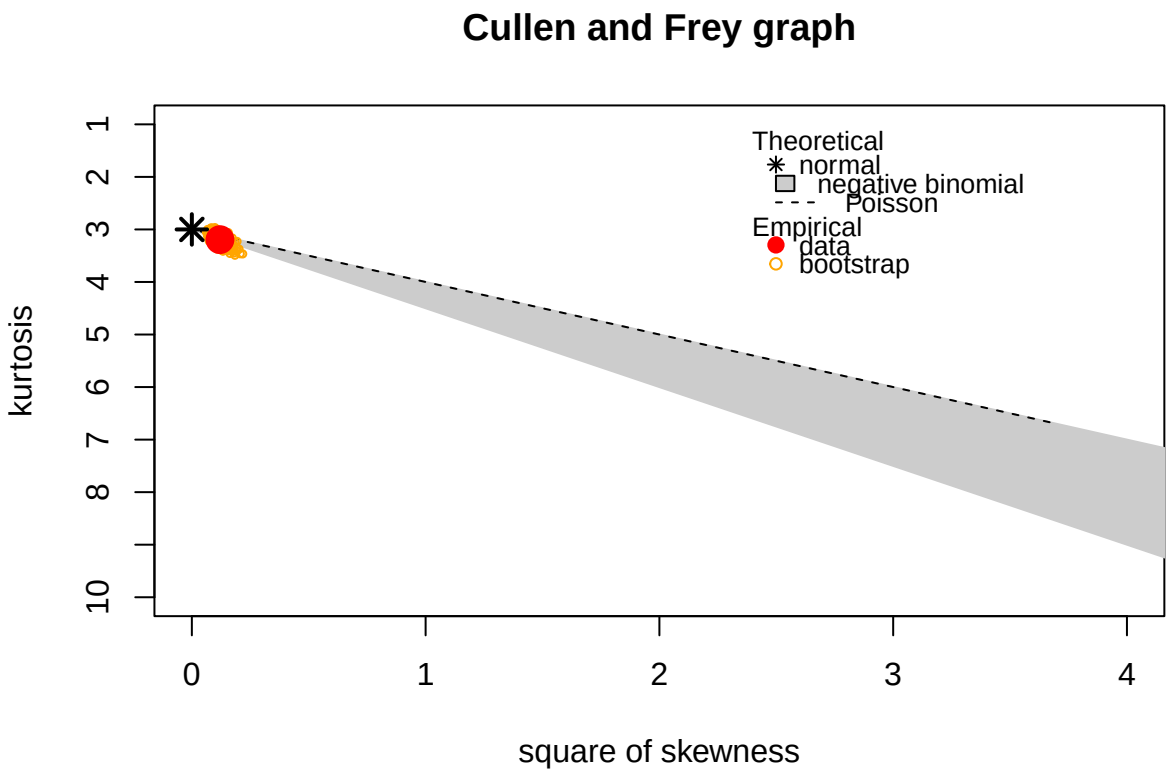



```
## summary statistics
## ------
## min:  60   max:  92
## median:  72
## mean:  72.56015
## estimated sd:  4.218891
## estimated skewness:  0.3467211
## estimated kurtosis:  3.194477
```

The data is very close to the normal distribution, but could also potentially be Poisson.

Create a bootstrap cluster on the Cullen-Frey chart

```
descdist(df$score,
         discrete=TRUE,
         boot=1000
         )
```




```
## summary statistics
## ------
## min:  60   max:  92
## median:  72
## mean:  72.56015
## estimated sd:  4.218891
## estimated skewness:  0.3467211
## estimated kurtosis:  3.194477
```

Our bootstrap cluster is very tight and very close to the normal distribution, so it is safe to say our data is normally distributed and that is what I will consider as the family

#Normal

```
fit_normal <- fitdist(df$score, "norm")
summary(fit_normal)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##       estimate Std. Error
## mean 72.560149 0.05259252
## sd    4.218563 0.03718852
## Loglikelihood:  -18391.16   AIC:  36786.32   BIC:  36799.86
## Correlation matrix:
##               mean            sd
## mean  1.000000e+00 -6.985447e-10
## sd   -6.985447e-10  1.000000e+00
```

Poisson

```
fit_poisson <- fitdist(df$score, "pois")
summary(fit_poisson)
```

```
## Fitting of the distribution ' pois ' by maximum likelihood
## Parameters :
##        estimate Std. Error
## lambda 72.56015  0.1061961
## Loglikelihood:  -20482.52   AIC:  40967.05   BIC:  40973.82
```

The loglikelihood of Poisson is much worse than Normal so we will stick with Normal Distribution as our family. This confirms that we can use Linear Regression as our ML model later on and we can begin feature validation using TTests.

TTest #1: Rain vs. Shine

NULL HYPOTHESIS: Rainy conditions do not negatively affect players golf scores ALTERNATIVE HYPOTHESIS: Rainy conditions do negatively affect players golf scores

```
dry_days <- df[df$precipitation == 0,]
rainy_days <- df[df$precipitation > 0,]

t.test(dry_days$score, rainy_days$score)
```

```
##
##  Welch Two Sample t-test
##
## data:  dry_days$score and rainy_days$score
## t = -28.371, df = 4892.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.134723 -2.729496
## sample estimates:
## mean of x mean of y
##  71.47006  74.40217
```

This ttest shows that the rainy conditions do heavily affect player scores and we can reject the null hypothesis

TTest #2: Windy vs. Not windy

NULL HYPOTHESIS: The wind does not negatively affect players golf scores ALTERNATIVE HYPOTHESIS: The wind does negatively affect players golf scores

```
calm_days <- df[df$wind_speed <= 15,]
windy_days <- df[df$wind_speed > 15,]

t.test(calm_days$score, windy_days$score)
```

```
##
##  Welch Two Sample t-test
##
## data:  calm_days$score and windy_days$score
## t = -18.537, df = 706.89, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.643985 -2.946012
## sample estimates:
## mean of x mean of y
##  72.26056  75.55556
```

This ttest shows that the windy conditions do heavily affect player scores and we can reject the null hypothesis. But which wind levels affect scores the most?

TTest #2.1

```
calm_days <- df[df$wind_speed <= 10,]
windy_days <- df[df$wind_speed > 10,]

t.test(calm_days$score, windy_days$score)
```

```
##
##  Welch Two Sample t-test
##
## data:  calm_days$score and windy_days$score
## t = -27.764, df = 6394.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.957303 -2.567231
## sample estimates:
## mean of x mean of y
##  71.20263  73.96490
```

The difference here is still pretty significant and our p value is still really low.

Ttest #2.2

```
calm_days <- df[df$wind_speed <= 5,]
windy_days <- df[df$wind_speed > 5,]

t.test(calm_days$score, windy_days$score)
```

```
##
##  Welch Two Sample t-test
##
## data:  calm_days$score and windy_days$score
```

```
## t = -4.1455, df = 80.309, p-value = 8.349e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.6981467 -0.9479328
## sample estimates:
## mean of x mean of y
##  70.75949  72.58253
```

We still have a pretty significant difference in scores. It seems that any amount of wind is going to affect scores pretty significantly.

Ttest #3: Combined weather conditions and temp

NULL HYPOTHESIS: Poor combined weather conditions do not negatively affect golf scores ALTERNATIVE HYPOTHESIS: Poor combined weather conditions do negatively affect golf scores

```
nice_days <- df[df$wind_speed <= 5 & df$precipitation == 0 & df$avg_temp < 60,]
poor_days <- df[df$wind_speed > 5 & df$precipitation > 0 & df$avg_temp >= 60,]

t.test(nice_days$score, poor_days$score)
```

```
##
##  Welch Two Sample t-test
##
## data:  nice_days$score and poor_days$score
## t = -4.5237, df = 17.55, p-value = 0.0002787
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.222400 -1.905805
## sample estimates:
## mean of x mean of y
##  70.05556  73.61966
```

Our p value is pretty low and the means of each are pretty different meaning we can reject our null hypothesis.

Ttest #4: Temperative alone

NULL HYPOTHESIS: Low temperature does not negatively affect golf scores ALTERNATIVE HYPOTHESIS: Low temperature does negatively affect golf scores

```
cold_days <- df[df$avg_temp < 60,]
warm_days <- df[df$avg_temp >= 60,]

t.test(warm_days$score, cold_days$score)
```

```
##
##  Welch Two Sample t-test
##
## data:  warm_days$score and cold_days$score
```

```
## t = -17.05, df = 3938.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.149092 -1.705810
## sample estimates:
## mean of x mean of y
##  71.90588  73.83333
```

With a low p value and significant difference in means, we can reject our null hypothesis that temperature alone does not negatively affect golf scores

Ttest #5: Day of week

NULL HYPOTHESIS: Day of the week does not negatively affect golf scores ALTERNATIVE HYPOTHESIS: Day of the week does negatively affect golf scores

```
weekends <- df[df$day_of_week_int >= 5,]
weekdays <- df[df$day_of_week_int < 5,]

t.test(weekends$score, weekdays$score)
```

```
##
##  Welch Two Sample t-test
##
## data:  weekends$score and weekdays$score
## t = -13.104, df = 5151.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.612652 -1.192913
## sample estimates:
## mean of x mean of y
##  72.02707  73.42986
```

We can reject our null hypothesis here as scores on the weekends appear to be 1.4 strokes better than on weekdays

Ttest #6: Amount of rounds played

NULL HYPOTHESIS: High rounds played at the course does not positively affect golf scores ALTERNATIVE HYPOTHESIS: High rounds played at the course does positively affect golf scores

```
many_rounds <- df[df$round >= 10,]
low_rounds <- df[df$round < 10,]

t.test(many_rounds$score, low_rounds$score)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  many_rounds$score and low_rounds$score
## t = -14.962, df = 6336.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.753189 -1.346995
## sample estimates:
## mean of x mean of y
##  71.83353  73.38362
```

Those who play many rounds look to shoot much better than those who play less rounds at the course on average. We can reject our null hypothesis here as well.

Now we can run a correlation test to see if handicap is correlated with your scores at the course. # Correlation: Score and Handicap

```
cor.test(df$handicap, df$score)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$handicap and df$score
## t = 25.048, df = 6432, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2756935 0.3202235
## sample estimates:
##       cor
## 0.2981207
```

Based on these results, score and handicap do move in the same direction, but not strongly. This shows that things like conditions and course difficulty matter more at this course than your handicap. Better handicaps, however do indicate better scores. We can still use it in our ML model.

Conclusion

In this SDA, we have been able to determine the relationships between things like weather conditions, temperature, day of the week and handicap to our players scores. We also figured out that our data is normally distributed and the best model to use for our predictions is Linear Regression. The data is in good shape and is a good starting point for our course manager and platform that will be used for predicting your score for the day.