Capstone Project

Oliver Yu

```
#Oliver Yu
#CDS 492: Data Science Capstone (Summer 2022)
#George Mason University

#Importing libraries

suppressPackageStartupMessages(library(tidyverse))
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3

## Warning: package 'ggplot2' was built under R version 4.1.3

## Warning: package 'tibble' was built under R version 4.1.3

## Warning: package 'tidyr' was built under R version 4.1.3

## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'dplyr' was built under R version 4.1.3
```

```
suppressPackageStartupMessages(library(broom))
```

```
## Warning: package 'broom' was built under R version 4.1.3
```

```
suppressPackageStartupMessages(library(modelr))

suppressPackageStartupMessages(library(plotly))
```

```
#Importing Dataset

forestfires <- read.csv("forestfires.csv")
```

```
forest_var <- forestfires %>%
  select(X, FFMC, DMC, DC, ISI, temp, RH, wind, rain, area) %>%
  rename("Xcoord" = "X" )
```

```
#FFMC Summary Statistics

forest_FFMC <- forest_var %>%
  group_by(Xcoord) %>%
```

```
  summarize(
    s_obs = n(),
    s_mean = mean(FFMC, na.rm = TRUE),
    s_median = median(FFMC, na.rm = TRUE),
    std_dev = sd(FFMC, na.rm = TRUE),
    IQR = IQR(FFMC, na.rm = TRUE),
    min_val = min(FFMC, na.rm = TRUE),
    max_val = max(FFMC, na.rm = TRUE)
  )
```

```
#DMC Summary Statistics

forest_DMC <- forest_var %>%
  group_by(Xcoord) %>%
  summarize(
    s_obs = n(),
    s_mean = mean(DMC, na.rm = TRUE),
    s_median = median(DMC, na.rm = TRUE),
    std_dev = sd(DMC, na.rm = TRUE),
    IQR = IQR(DMC, na.rm = TRUE),
    min_val = min(DMC, na.rm = TRUE),
    max_val = max(DMC, na.rm = TRUE)
  )

forest_DMC
```

```
## # A tibble: 9 x 8
##   Xcoord s_obs s_mean s_median std_dev   IQR min_val max_val
##    <int> <int>  <dbl>    <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1      1    48  137.     130.     51.9  58.2    51.3    276.
## 2      2    73  125.     118.     60.7  53       3.6    290
## 3      3    55   98.6     99      56.6  99.3     2.4    248.
## 4      4    91   97.3     99.6    59.0  85.4     1.1    290
## 5      5    30  108.     100.     70.0  92.9     4.9    290
## 6      6    86   91.7     94.3    61.7  93.2     3      291.
## 7      7    60  116.     104.     71.7  93.2     3      287.
## 8      8    61  133.     130.     66.3 106.     27.8    274.
## 9      9    13   89.0     68.6    73.1  54.7     6.8    248.
```

```
#DC Summary Statistics

forest_DC <- forest_var %>%
  group_by(Xcoord) %>%
  summarize(
    s_obs = n(),
    s_mean = mean(DC, na.rm = TRUE),
    s_median = median(DC, na.rm = TRUE),
    std_dev = sd(DC, na.rm = TRUE),
    IQR = IQR(DC, na.rm = TRUE),
    min_val = min(DC, na.rm = TRUE),
    max_val = max(DC, na.rm = TRUE)
  )
```

```
forest_DC
```

```
## # A tibble: 9 x 8
##   Xcoord s_obs s_mean s_median std_dev   IQR min_val max_val
##    <int> <int>  <dbl>    <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1      1    48   660.     693.    134.  102.   104.     825.
## 2      2    73   590.     658.    204.  114      9.3    855.
## 3      3    55   498.     608.    272.  460.    15.5    823.
## 4      4    91   531.     662.    261.  356.    55      855.
## 5      5    30   506.     689.    303.  624.    15.8    855.
## 6      6    86   488.     658.    300.  630.    16.2    861.
## 7      7    60   572.     671.    234.  238.     7.9    849.
## 8      8    61   593.     664.    178.  111.    77.5    819.
## 9      9    13   400.     355.    232.  262.    26.6    754.
```

```
#ISI Summary Statistics

forest_ISI <- forest_var %>%
  group_by(Xcoord) %>%
  summarize(
    s_obs = n(),
    s_mean = mean(ISI, na.rm = TRUE),
    s_median = median(ISI, na.rm = TRUE),
    std_dev = sd(ISI, na.rm = TRUE),
    IQR = IQR(ISI, na.rm = TRUE),
    min_val = min(ISI, na.rm = TRUE),
    max_val = max(ISI, na.rm = TRUE)
  )

forest_ISI
```

```
## # A tibble: 9 x 8
##   Xcoord s_obs s_mean s_median std_dev   IQR min_val max_val
##    <int> <int>  <dbl>    <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1      1    48   9.10      8.4    2.80  3.03     5      17
## 2      2    73   9.60      8.6    4.69  6        0.4    21.3
## 3      3    55   8.60      8.1    3.93  4.35     0.7    20.3
## 4      4    91   8.54      8.5    4.42  3.55     0      20
## 5      5    30  10.1       9.2    3.54  4.55     3.9    17.7
## 6      6    86   7.88      7.5    3.85  3.2      0.4    22.7
## 7      7    60  10.2       8.85   7.39  5        1.9    56.1
## 8      8    61   9.39      8.5    3.86  5.9      1.9    18
## 9      9    13   8.56      9      3.45  3.5      3.2    14
```

```
#Temp Summary Statistics

forest_temp <- forest_var %>%
  group_by(Xcoord) %>%
  summarize(
    s_obs = n(),
    s_mean = mean(temp, na.rm = TRUE),
    s_median = median(temp, na.rm = TRUE),
```

```
    std_dev = sd(temp, na.rm = TRUE),
    IQR = IQR(temp, na.rm = TRUE),
    min_val = min(temp, na.rm = TRUE),
    max_val = max(temp, na.rm = TRUE)
  )

forest_temp
```

```
## # A tibble: 9 x 8
##   Xcoord s_obs s_mean s_median std_dev   IQR min_val max_val
##    <int> <int>  <dbl>    <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1      1    48   21.0     21.4    4.82  5.57     8.3    32.4
## 2      2    73   20.6     20.9    5.47  5.5      4.6    33.1
## 3      3    55   17.5     17.6    5.16  6.3      4.6    32.3
## 4      4    91   18.2     18      6.32  6.05     2.2    32.6
## 5      5    30   18.4     18.4    5.50  8.75     7.5    27.6
## 6      6    86   17.3     18.2    6.33  7.88     4.2    33.3
## 7      7    60   18.1     19.2    5.06  6.05     5.1    27.3
## 8      8    61   20.1     20.4    5.41  7.2      5.1    31
## 9      9    13   22.6     24.5    6.23  4.8      6.7    30.2
```

```
#Relative Humidity Summary Statistics

forest_RH <- forest_var %>%
  group_by(Xcoord) %>%
  summarize(
    s_obs = n(),
    s_mean = mean(RH, na.rm = TRUE),
    s_median = median(RH, na.rm = TRUE),
    std_dev = sd(RH, na.rm = TRUE),
    IQR = IQR(RH, na.rm = TRUE),
    min_val = min(RH, na.rm = TRUE),
    max_val = max(RH, na.rm = TRUE)
  )

forest_RH
```

```
## # A tibble: 9 x 8
##   Xcoord s_obs s_mean s_median std_dev   IQR min_val max_val
##    <int> <int>  <dbl>    <dbl>   <dbl> <dbl>   <int>   <int>
## 1      1    48   43.2       40    16.1  18.2      15      88
## 2      2    73   43.0       41    14.2  18        19      79
## 3      3    55   43.2       40    14.7  14.5      18      87
## 4      4    91   41.8       41    16.5  26        15     100
## 5      5    30   45.8       42.5  16.8  23.5      24      80
## 6      6    86   43.7       39    16.8  22        21      94
## 7      7    60   50.4       47    16.8  23.8      27      96
## 8      8    61   46         43    18.1  26        22      99
## 9      9    13   40.9       36    14.6   9        25      79
```

```
#Wind Speed Summary Statistics
```

```r
forest_wind <- forest_var %>%
  group_by(Xcoord) %>%
  summarize(
    s_obs = n(),
    s_mean = mean(wind, na.rm = TRUE),
    s_median = median(wind, na.rm = TRUE),
    std_dev = sd(wind, na.rm = TRUE),
    IQR = IQR(wind, na.rm = TRUE),
    min_val = min(wind, na.rm = TRUE),
    max_val = max(wind, na.rm = TRUE)
  )

forest_wind
```

```
## # A tibble: 9 x 8
##   Xcoord s_obs s_mean s_median std_dev   IQR min_val max_val
##    <int> <int>  <dbl>    <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1      1    48   3.77     3.6     1.83   2.4     0.9     8.5
## 2      2    73   3.86     3.6     1.76   2.7     0.9     9.4
## 3      3    55   4.21     4       1.79   2.3     0.4     8.5
## 4      4    91   3.97     4       1.92   2.2     0.9     8.5
## 5      5    30   4.47     4       1.68   1.8     1.3     8
## 6      6    86   4.08     4       1.81   2.2     0.9     9.4
## 7      7    60   4.28     4.25    1.95   2.8     0.9     9.4
## 8      8    61   3.89     4       1.57   2.2     1.3     8.9
## 9      9    13   3.3      3.1     1.08   1.8     0.9     4.5
```

```r
#Rain Summary Statistics

forest_rain <- forest_var %>%
  group_by(Xcoord) %>%
  summarize(
    s_obs = n(),
    s_mean = mean(rain, na.rm = TRUE),
    s_median = median(rain, na.rm = TRUE),
    std_dev = sd(rain, na.rm = TRUE),
    IQR = IQR(rain, na.rm = TRUE),
    min_val = min(rain, na.rm = TRUE),
    max_val = max(rain, na.rm = TRUE)
  )

forest_rain
```

```
## # A tibble: 9 x 8
##   Xcoord s_obs  s_mean s_median std_dev   IQR min_val max_val
##    <int> <int>   <dbl>    <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1      1    48 0              0  0          0       0     0
## 2      2    73 0              0  0          0       0     0
## 3      3    55 0              0  0          0       0     0
## 4      4    91 0.00440        0  0.0419     0       0     0.4
## 5      5    30 0.0467         0  0.256      0       0     1.4
## 6      6    86 0              0  0          0       0     0
## 7      7    60 0.14           0  0.838      0       0     6.4
## 8      8    61 0.0164         0  0.105      0       0     0.8
## 9      9    13 0              0  0          0       0     0
```

```r
forest_area <- forest_var %>%
  group_by(Xcoord) %>%
  summarize(
    s_obs = n(),
    s_mean = mean(area, na.rm = TRUE),
    s_median = median(area, na.rm = TRUE),
    std_dev = sd(area, na.rm = TRUE),
    IQR = IQR(area, na.rm = TRUE),
    min_val = min(area, na.rm = TRUE),
    max_val = max(area, na.rm = TRUE)
  )

forest_area
```

```
## # A tibble: 9 x 8
##   Xcoord s_obs s_mean s_median std_dev   IQR min_val max_val
##    <int> <int>  <dbl>    <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1      1    48  13.4     0.38    36.3   6.57       0   213.
## 2      2    73   9.57    1.47    31.4   6.43       0   201.
## 3      3    55   2.46    0        6.35  1.36       0    35.9
## 4      4    91  10.4     0.79    22.4   9.34       0   155.
## 5      5    30   3.05    0.045    5.70  3.36       0    24.2
## 6      6    86  20.1     0.955  118.    6.92       0  1091.
## 7      7    60  11.1     0.205   38.0   7.18       0   279.
## 8      8    61  24.5     1.23   100.    7.19       0   746.
## 9      9    13  18.5     1.63    33.7   8.16       0   106.
```