2353_project

Ivy Zhao

2024-04-14

```
bike <- read.csv("~/Downloads/bike+sharing+dataset/day.csv")
bikeDf<- subset(bike, select = c("yr", "season", "holiday", "workingday","weathersit", "temp", "atemp",
summary(bikeDf)
```

```
##        yr             season         holiday          workingday
##  Min.   :0.0000   Min.   :1.000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.00000   1st Qu.:0.000
##  Median :1.0000   Median :3.000   Median :0.00000   Median :1.000
##  Mean   :0.5007   Mean   :2.497   Mean   :0.02873   Mean   :0.684
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:0.00000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :4.000   Max.   :1.00000   Max.   :1.000
##    weathersit         temp             atemp              hum
##  Min.   :1.000   Min.   :0.05913   Min.   :0.07907   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.33708   1st Qu.:0.33784   1st Qu.:0.5200
##  Median :1.000   Median :0.49833   Median :0.48673   Median :0.6267
##  Mean   :1.395   Mean   :0.49538   Mean   :0.47435   Mean   :0.6279
##  3rd Qu.:2.000   3rd Qu.:0.65542   3rd Qu.:0.60860   3rd Qu.:0.7302
##  Max.   :3.000   Max.   :0.86167   Max.   :0.84090   Max.   :0.9725
##    windspeed           cnt
##  Min.   :0.02239   Min.   :  22
##  1st Qu.:0.13495   1st Qu.:3152
##  Median :0.18097   Median :4548
##  Mean   :0.19049   Mean   :4504
##  3rd Qu.:0.23321   3rd Qu.:5956
##  Max.   :0.50746   Max.   :8714
```

```
nrow(bikeDf)
```

```
## [1] 731
```

```
#SD for continuous variables
sd_values <- sapply(bikeDf[, c("cnt", "temp", "atemp", "hum", "windspeed")], sd)
print(sd_values)
```

```
##          cnt         temp        atemp          hum    windspeed
## 1.937211e+03 1.830510e-01 1.629612e-01 1.424291e-01 7.749787e-02
```

#Univariate Analysis on Categorical Data

```
yr_counts <- table(bikeDf$yr)
yr_percentages <- prop.table(yr_counts) * 100
yr_summary <- data.frame(yr = names(yr_counts),
                         Count = as.numeric(yr_counts),
                         Percentage = yr_percentages)
print(yr_summary)
```

```
##   yr Count Percentage.Var1 Percentage.Freq
## 1  0   365              0         49.9316
## 2  1   366              1         50.0684
```

```
season_counts <- table(bikeDf$season)
season_percentages <- prop.table(season_counts) * 100
season_summary <- data.frame(Season = names(season_counts),
                             Count = as.numeric(season_counts),
                             Percentage = season_percentages)
print(season_summary)
```

```
##   Season Count Percentage.Var1 Percentage.Freq
## 1      1   181              1        24.76060
## 2      2   184              2        25.17100
## 3      3   188              3        25.71819
## 4      4   178              4        24.35021
```

```
holiday_counts <- table(bikeDf$holiday)
holiday_percentages <- prop.table(holiday_counts) * 100
holiday_summary <- data.frame(holiday = names(holiday_counts),
                             Count = as.numeric(holiday_counts),
                             Percentage = holiday_percentages)
print(holiday_summary)
```

```
##   holiday Count Percentage.Var1 Percentage.Freq
## 1       0   710              0       97.127223
## 2       1    21              1        2.872777
```

```
workingday_counts <- table(bikeDf$workingday)
workingday_percentages <- prop.table(workingday_counts) * 100
workingday_summary <- data.frame(workingday = names(workingday_counts),
                             Count = as.numeric(workingday_counts),
                             Percentage = workingday_percentages)
print(workingday_summary)
```

```
##   workingday Count Percentage.Var1 Percentage.Freq
## 1          0   231              0        31.60055
## 2          1   500              1        68.39945
```

```
weathersit_counts <- table(bikeDf$weathersit)
weathersit_percentages <- prop.table(weathersit_counts) * 100
weathersit_summary <- data.frame(weathersit = names(weathersit_counts),
                             Count = as.numeric(weathersit_counts),
                             Percentage = weathersit_percentages)
print(weathersit_summary)
```

```
##   weathersit Count Percentage.Var1 Percentage.Freq
## 1          1   463              1       63.337893
## 2          2   247              2       33.789330
## 3          3    21              3        2.872777
```

```
#VIF for initial model
require(faraway)
```

```
## Loading required package: faraway
```

```
model1 <- lm(cnt ~ ., data = bikeDf)
vif(model1)
```

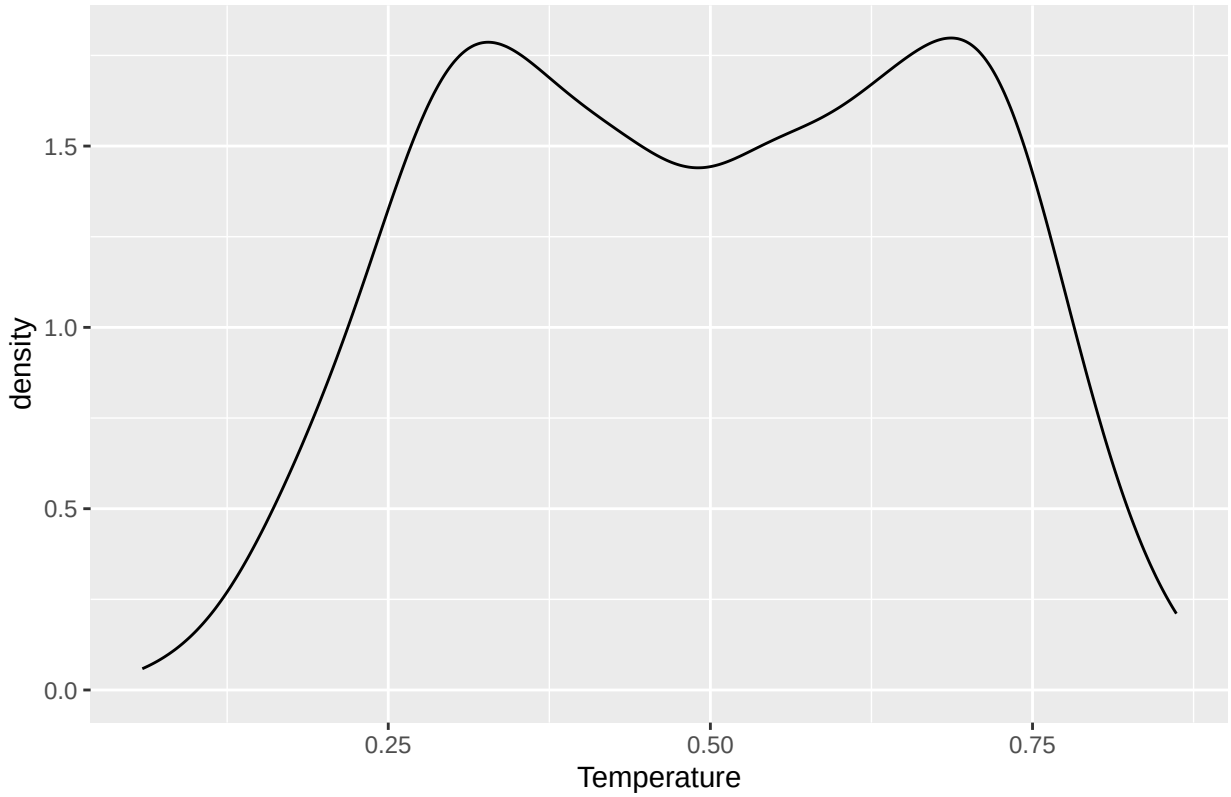
```
##         yr     season    holiday workingday weathersit       temp      atemp
##   1.019702   1.193101   1.070683   1.076189   1.736811  63.139620  64.144333
##        hum  windspeed
##   1.885224   1.198316
```

```
#VIF with atemp removed
model2 <- lm(cnt ~ yr + season + holiday + workingday + weathersit +temp + hum + windspeed, data = bikeI
vif(model2)
```

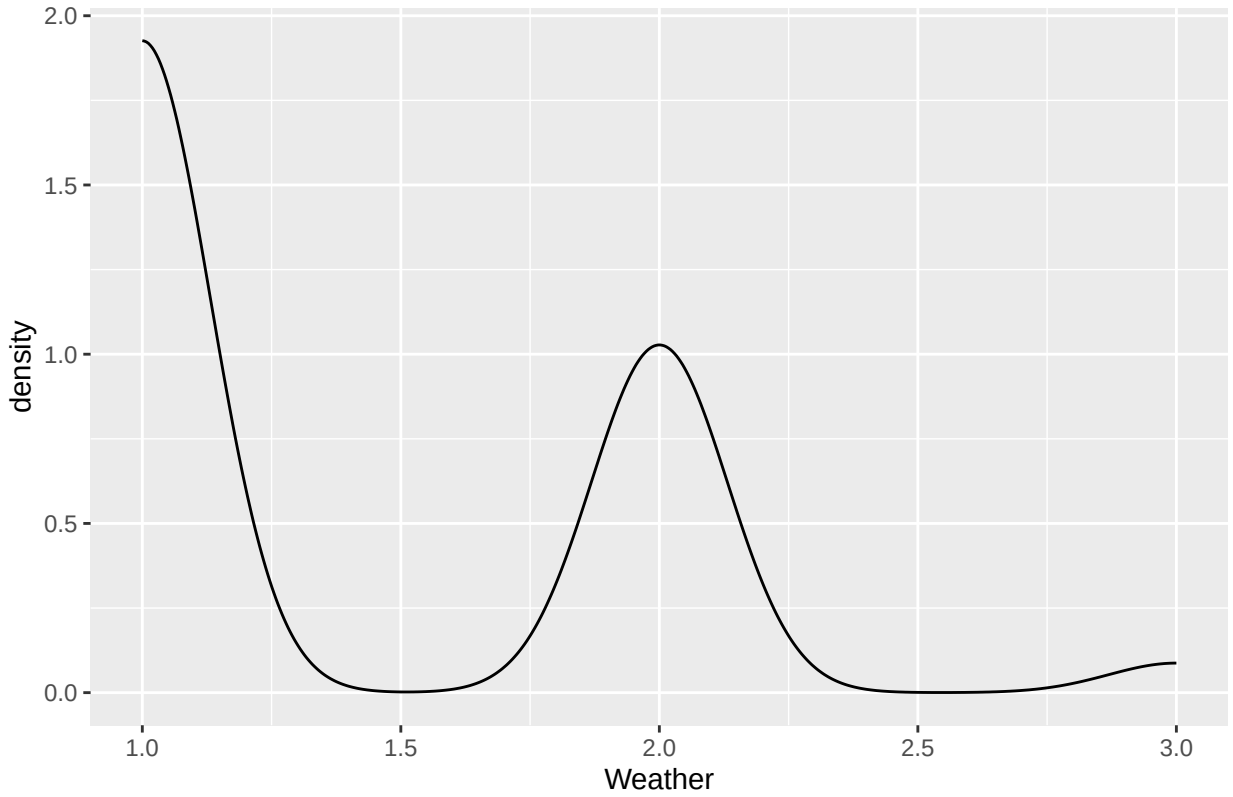
```
##         yr     season    holiday workingday weathersit       temp        hum
##   1.019699   1.190599   1.069313   1.076135   1.728863   1.197241   1.871160
##  windspeed
##   1.164768
```

```
#Density Plot: Bike Counts
library(ggplot2)
ggplot(bikeDf, aes(x = cnt)) +
  geom_density(color = "black") +
  labs(title = "",
       x = "Bike counts",
       y = "density")
```


```
#Density Plot: Temperature
library(ggplot2)
ggplot(bikeDf, aes(x = temp)) +
  geom_density(color = "black") +
  labs(title = "",
       x = "Temperature",
       y = "density")
```



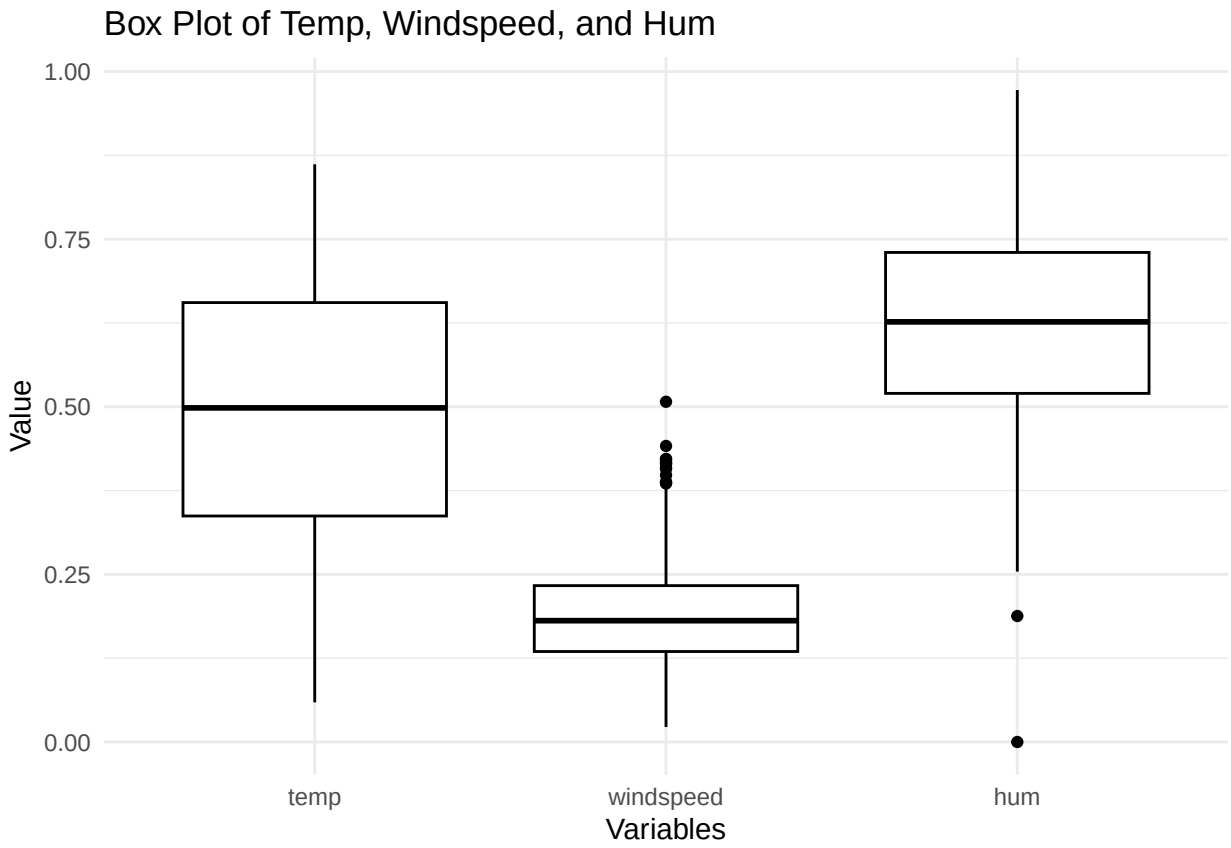

```
#Density Plot: Weather
library(ggplot2)
ggplot(bikeDf, aes(x = weathersit)) +
  geom_density(color = "black") +
  labs(title = "",
       x = "Weather",
       y = "density")
```




```
#Boxplot for outliers
library(ggplot2)

box <- bikeDf[, c("temp", "windspeed", "hum")]
box <- stack(box)

ggplot(box, aes(x = ind, y = values)) +
  geom_boxplot(color = "black") +
  labs(title = "Box Plot of Temp, Windspeed, and Hum",
       x = "Variables",
       y = "Value") +
  theme_minimal()
```


Box Plot of Temp, Windspeed, and Hum
1.00
0.75
0.50
0.25
0.00
Value
temp
windspeed
hum
Variables