MA575 Report 2

C2 Team 4: Zhiquan Shen, Katherine Albrecht, Jiachen Chen, Jiahe Zhang, Dahyun Hong, Xinzhi Zhang.

10/03/2021

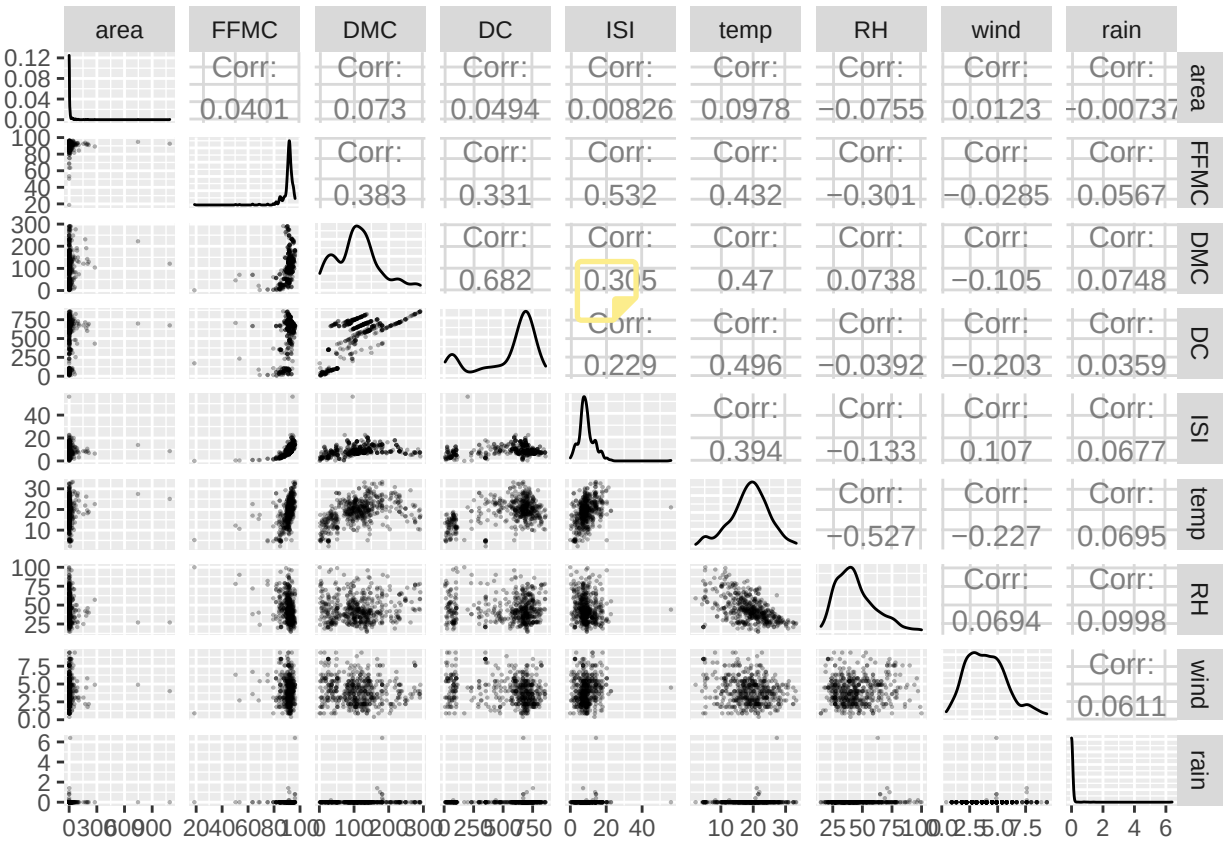
1. Our response variable is log(total area burned across the park + 1) on a given day. Our explanatory variable is the mean temperature across all observations from that day.

We first do some exploratory analysis based on the original data.

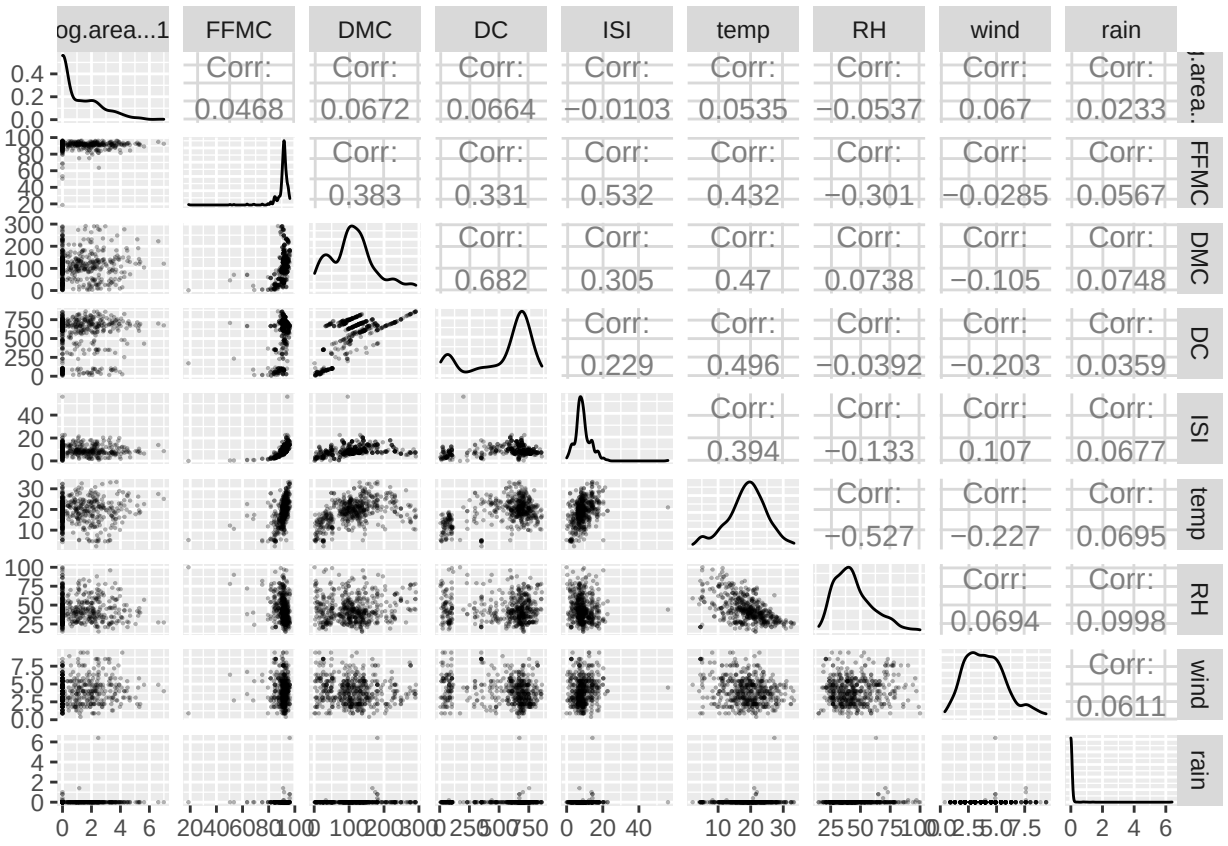
```
# Load visualization packages
library(carData)
library(car)
library(ggplot2)
library(hrbrthemes)
library(GGally)
library(dplyr)
```

```
# Read data from csv file and attach
firedatafull <- read.csv("forestfires.csv",header=TRUE)
attach(firedatafull)
```

```
# Exploratory plots for preparation
# Plot pairwise correlations between continuous variables in full dataset
data <- data.frame(area, FFMC, DMC, DC, ISI, temp, RH, wind, rain)

ggpairs(data, lower = list(continuous = wrap("points", alpha = 0.3,    size=0.1)))
```




```
# Try log-transforming area
logdata <- data.frame(log(area+1), FFMC, DMC, DC, ISI, temp, RH, wind, rain)
ggpairs(logdata, lower = list(continuous = wrap("points", alpha = 0.3,    size=0.1)))
```



When exploring the dataset, we found that there were many observations where the values for month, day of week, DC, DMC, FFMC, and ISI matched perfectly. Because this is extremely unlikely to occur by chance, we conclude that all these fires occurred on the same day. Fire size often varies greatly between observations on the same day, as shown in the example below. As a result of the wide variation in area outcome for the same combination of predictor values, even variables like FFMC and ISI that were developed and tested for the purposes of predicting fire risk show no significant correlation with area in a scatterplot matrix based on the original data (as shown in plots above).

```
# Exmaples for observations on the same day
firedatafull[c(374,416,429,433),]
```

```
##     X Y month day FFMC   DMC    DC  ISI temp RH wind rain   area
## 374 5 4   aug thu 94.8 222.4 698.6 13.9 20.3 42  2.7    0   0.00
## 416 8 6   aug thu 94.8 222.4 698.6 13.9 27.5 27  4.9    0 746.28
## 429 1 3   aug thu 94.8 222.4 698.6 13.9 26.2 34  5.8    0   0.00
## 433 8 6   aug thu 94.8 222.4 698.6 13.9 23.9 38  6.7    0   0.00
```

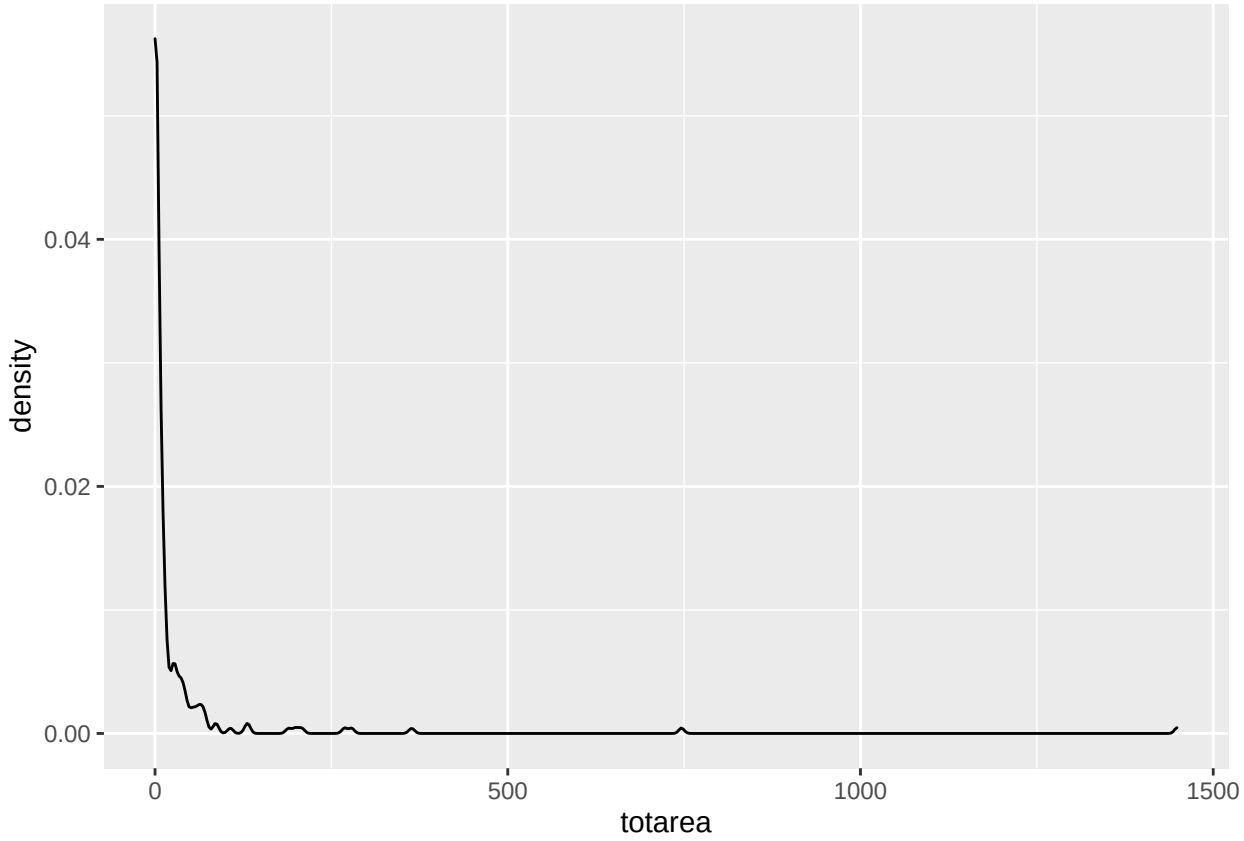
When trying to predict fire damage, if multiple fires occur on the same day, it is more practically meaningful to consider the damage in terms of the total area burned rather than the average size of one fire. To consider this as a response variable, we created the variable 'totarea' by summing the area values across each set of observations with matching month, day, FFMC, DMC, DC, and ISI values.

```
# create new dataset summing area in responses with matching month, day, DC, DMC, FFMC, and ISI values

firedatafull$dateFWIinds <- as.factor(paste(firedatafull$month, firedatafull$day,
firedatafull$DMC, firedatafull$ISI, firedatafull$FFMC, firedatafull$DC))
data2 <- group_by(firedatafull, dateFWIinds)
data3 <- summarise(data2, DC=mean(DC), DMC=mean(DMC), FFMC= mean(FFMC), ISI = mean(ISI),
temp= mean(temp), wind=mean(wind), RH = mean(RH), totarea = sum(area))
```

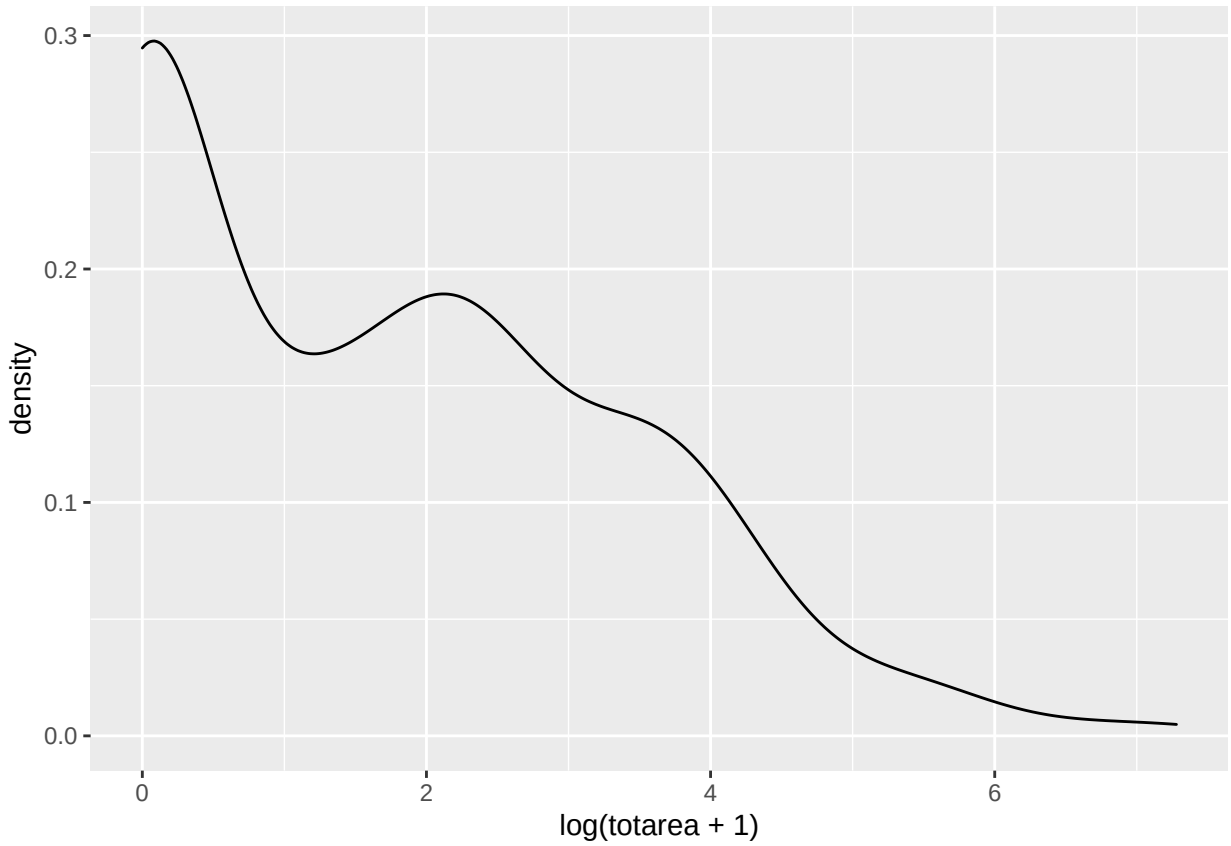
```
# firedatafull is no longer in use
detach(firedatafull)

attach(data3)
```

The distribution of "totarea" is shown in the first plot below, indicating a skewed trait that most of the fires are small in size. In order to reduce skewness, we applied a logarithm transformation y=log(x+1) to the "totarea" and made the data closer to normal distribution (the second plot below). This transformed response variable also showed a stronger correlation with other covariates than the log(area+1) variable from the original dataset.

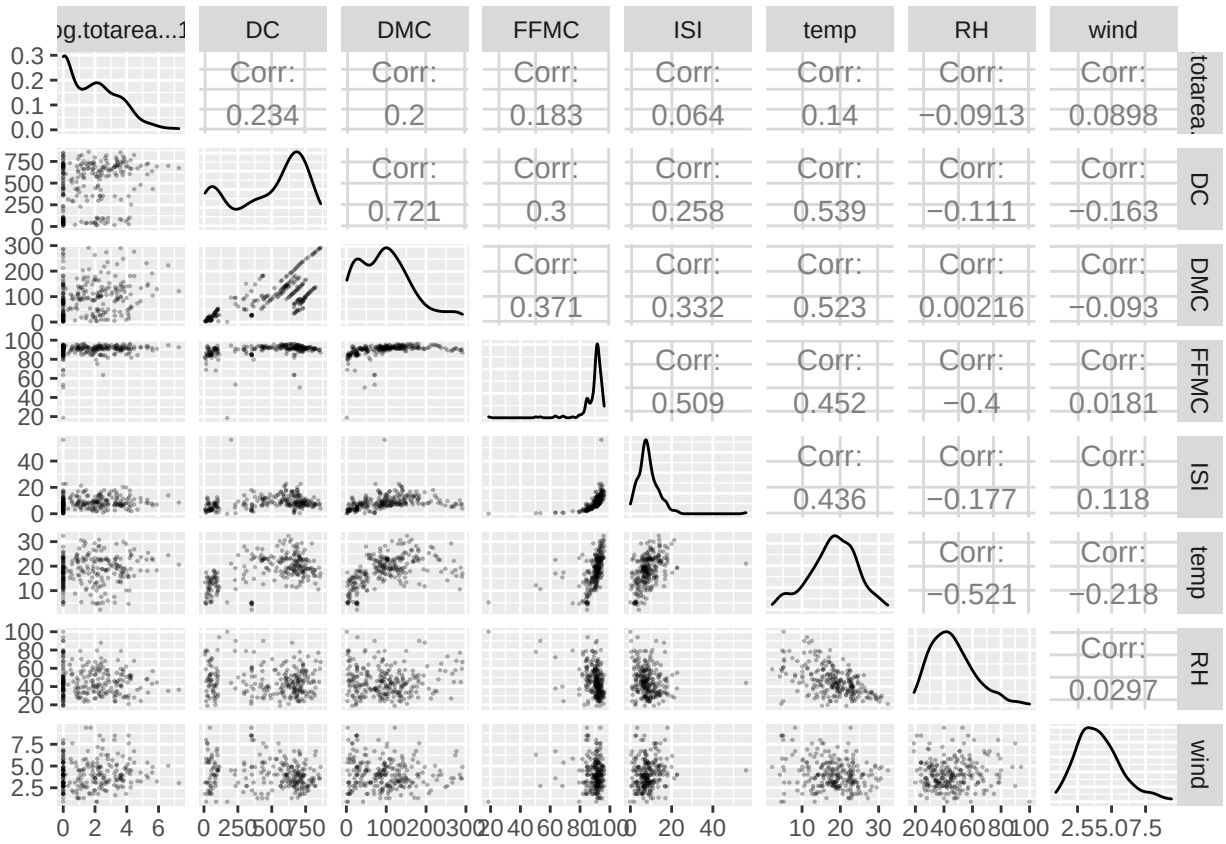
```
#the density of the totarea and the logarithm transformed totarea
ggplot(data3, aes(x=totarea)) + geom_density()
```




```
ggplot(data3, aes(x=log(totarea+1))) + geom_density()
```




```
# exploratory pairwise correlation plots for the total area burned dataset
plotdata <- data.frame(log(totarea+1), DC, DMC, FFMC, ISI, temp, RH, wind)
ggpairs(plotdata, lower = list(continuous = wrap("points", alpha = 0.3, size=0.1)))
```



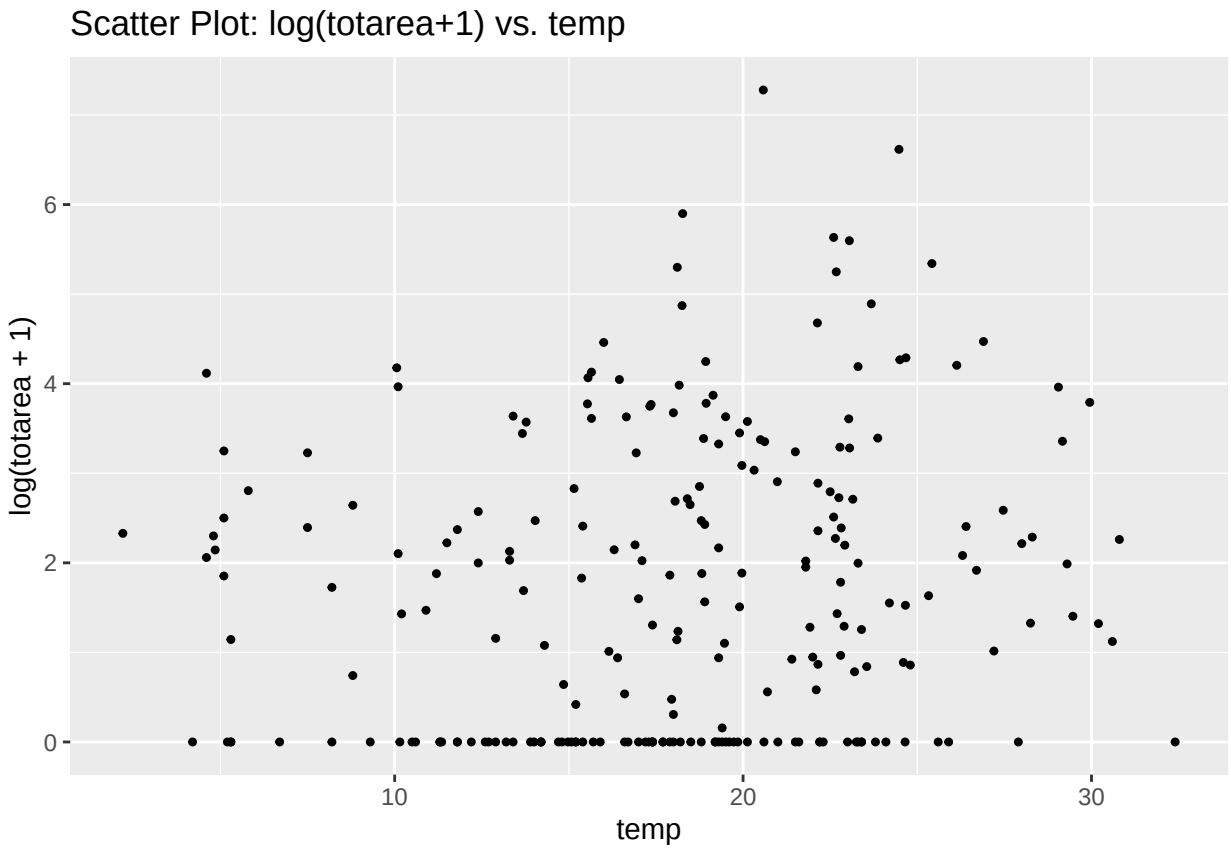

```
cor.test(log(totarea+1),temp)
```

```
##
##  Pearson's product-moment correlation
##
## data:  log(totarea + 1) and temp
## t = 2.168, df = 234, p-value = 0.03117
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01285422 0.26330653
## sample estimates:
##       cor
## 0.1403244
```

We observed from above results that "temp" has a significant correlation with "log(totarea+1)". Because temperature is an easily measurable variable that can intuitively be expected to increase fire risk, we selected it as our predictor. As weather variables are observed in real-time and can vary across observations on the same day, it was necessary to compute a summary statistic for temperature. For simplicity, we used mean(temp) in this analysis. However, a weighted mean temperature based on the proportional contribution of each observation to the total area burned is arguably more appropriate and will be considered for future models.

2. Scatterplot for log(totarea+1) vs. mean(temperature)

```
# single scatterplot for selected variables
ggplot(data3, aes(x= temp, y= log(totarea+1))) + geom_point(size=0.9) +
ggtitle("Scatter Plot: log(totarea+1) vs. temp")
```



3 & 4. Performing OLS

```
# linear model
m.ols <- lm(log(totarea+1) ~ temp, data = data3)
summary(m.ols)
```

```
##
## Call:
## lm(formula = log(totarea + 1) ~ temp, data = data3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.267 -1.541 -0.176  1.214  5.460
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.03806    0.33428   3.105  0.00213 **
## temp         0.03793    0.01750   2.168  0.03117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.632 on 234 degrees of freedom
## Multiple R-squared:  0.01969,    Adjusted R-squared:  0.0155
## F-statistic:   4.7 on 1 and 234 DF,  p-value: 0.03117
```

```
# confidence interval 95%
round(confint(m.ols), digits=4)
```

```
##                2.5 % 97.5 %
## (Intercept) 0.3795 1.6966
## temp        0.0035 0.0724
```

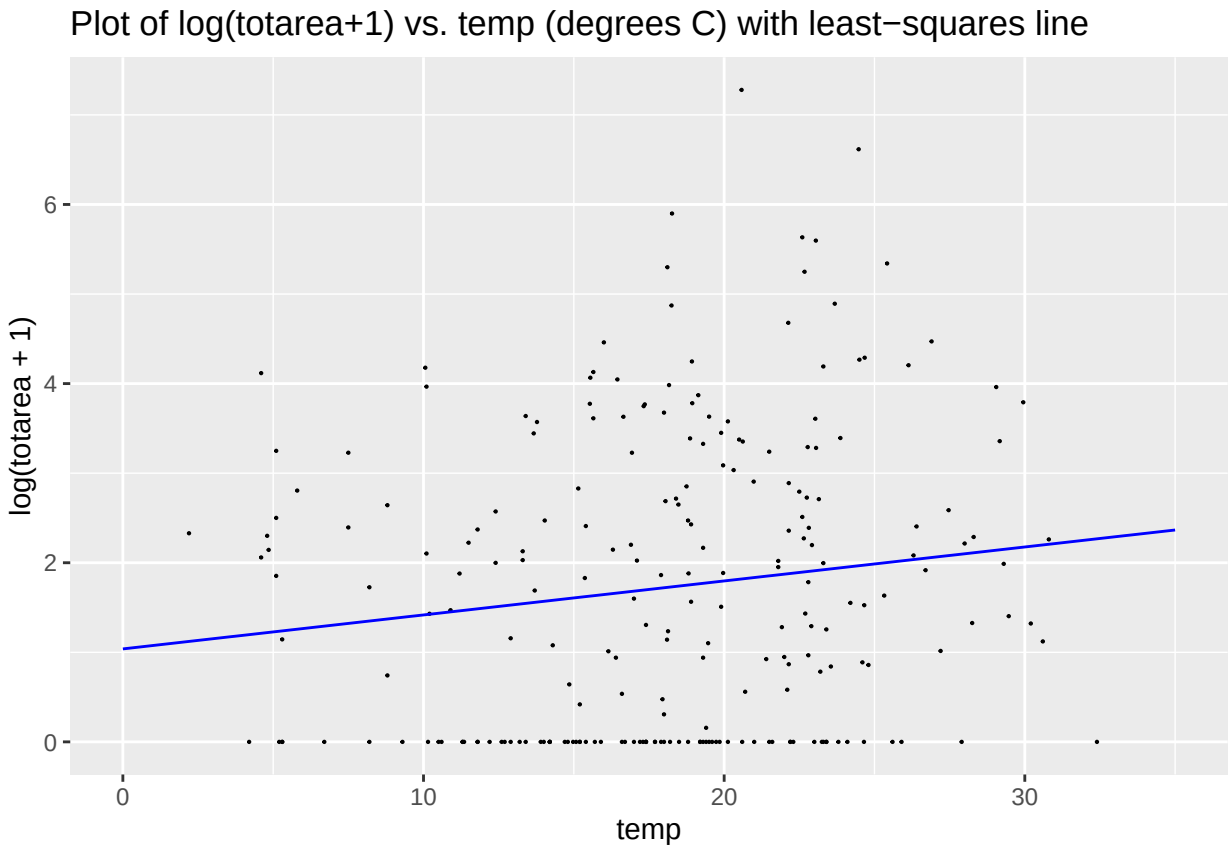
After performing OLS for the linear model log(totarea+1)=β_0 + β_1*temp, we notice that R-squared is 0.01969, which is small, indicating that the proportion of variability in log(totarea+1) explained by temperature is low (slightly under 2%).

We also notice that the p-value for β_1 is 0.031(<0.05) with a t-statistic on 234 degrees of freedom of 2.168 and the F-statistic on 1 numerator and 234 denominator degrees of freedom is 4.7, which all indicate that if we perform a hypothesis test with H0:$\beta_1 = 0$ Ha:$\beta_1 \neq 0$, we will reject the null hypothesis and conclude that $\beta_1 \neq 0$, at significance of of 0.05.

Further, the coefficient estimate is 0.0379 (95% CI: [0.0035, 0.0724]) for β_1 and the estimated intercept (β_0) is 1.038 (95% CI: [0.3795, 1.6966]). For each 1 degree increase in temperature, we would expect the total area burned in the park to increase by approximately 3.9% (re-calculated based on the increase of the logarithm-transformed total area burned in the park). Although the linear relationship between temp and log(totarea+1) is weak, we see a relationship that log(totarea+1)=1.038+0.0379*temp.

5. Scatterplot along with the linear regression fit

```
# plot the graph
temp_new<-seq(0,35,len=length(data3$temp))
predLinear = predict(m.ols,newdata=data.frame(temp=temp_new));
ggplot(data3, aes(x= temp, y= log(totarea+1))) + geom_point(size = 0.1) +
  geom_line(mapping = aes(x = temp_new, y = predLinear), color="blue")+
  ggtitle("Plot of log(totarea+1) vs. temp (degrees C) with least-squares line")
```



```
#cleanup
detach(data3)
```

From the above graph, we can obviously see that there is a weak linear relationship since each point is relatively far from the least-square line. It also appears that the variance of the residuals may be greater at higher temperatures, rather than being constant as would occur in a well-fitting model. Therefore, it is not a good linear regression model.