DSC630: Assignment 1.2

Scott Breitbach

8/31/2021

1.2 Assignment: R/Python Refresher

1. *Import, Plot, Summarize, and Save Data*

Using the US Bureau of Labor Statistics data, choose a dataset that interests you. Then generate summary statistics for 2 variables, plot some of the features (e.g., histograms, box plots, density plots, etc.) of several variables, and save the data locally as CSV files.

Data Set:

> CPI Average Price Data, U.S. city average (AP)
>
> Series Id: APU000074714
> Series Title: Gasoline, unleaded regular, per gallon/3.785 liters in U.S. city average, average price, not seasonally adjusted
> Area: U.S. city average
> Item: Gasoline, unleaded regular, per gallon/3.785 liters

```
# Import spreadsheet
gasPrices <- read_excel('SeriesReport_20210829184422_d810b7.xlsx',
                        range='A10:M56')

# Preivew data
head(gasPrices)
```

```
## # A tibble: 6 x 13
##    Year   Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  1976 0.605 0.6   0.594 0.592 0.6   0.616 0.623 0.628 0.63  0.629 0.629 0.626
## 2  1977 0.627 0.637 0.643 0.651 0.659 0.665 0.667 0.667 0.666 0.665 0.664 0.665
## 3  1978 0.648 0.647 0.647 0.649 0.655 0.663 0.674 0.682 0.688 0.69  0.695 0.705
## 4  1979 0.716 0.73  0.755 0.802 0.844 0.901 0.949 0.988 1.02  1.03  1.04  1.06
## 5  1980 1.13  1.21  1.25  1.26  1.27  1.27  1.27  1.27  1.26  1.25  1.25  1.26
## 6  1981 1.30  1.38  1.42  1.41  1.4   1.39  1.38  1.38  1.38  1.37  1.37  1.36
```

```
# Assign data to a data frame
gasPricesDF <- as.data.frame(gasPrices)

# Set year column as row names
rownames(gasPricesDF) <- gasPricesDF$Year
```

```r
gasPricesDF$Year <- NULL

# Preview data
head(gasPricesDF)
```

```
##        Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
## 1976 0.605 0.600 0.594 0.592 0.600 0.616 0.623 0.628 0.630 0.629 0.629 0.626
## 1977 0.627 0.637 0.643 0.651 0.659 0.665 0.667 0.667 0.666 0.665 0.664 0.665
## 1978 0.648 0.647 0.647 0.649 0.655 0.663 0.674 0.682 0.688 0.690 0.695 0.705
## 1979 0.716 0.730 0.755 0.802 0.844 0.901 0.949 0.988 1.020 1.028 1.041 1.065
## 1980 1.131 1.207 1.252 1.264 1.266 1.269 1.271 1.267 1.257 1.250 1.250 1.258
## 1981 1.298 1.382 1.417 1.412 1.400 1.391 1.382 1.376 1.376 1.371 1.369 1.365
```

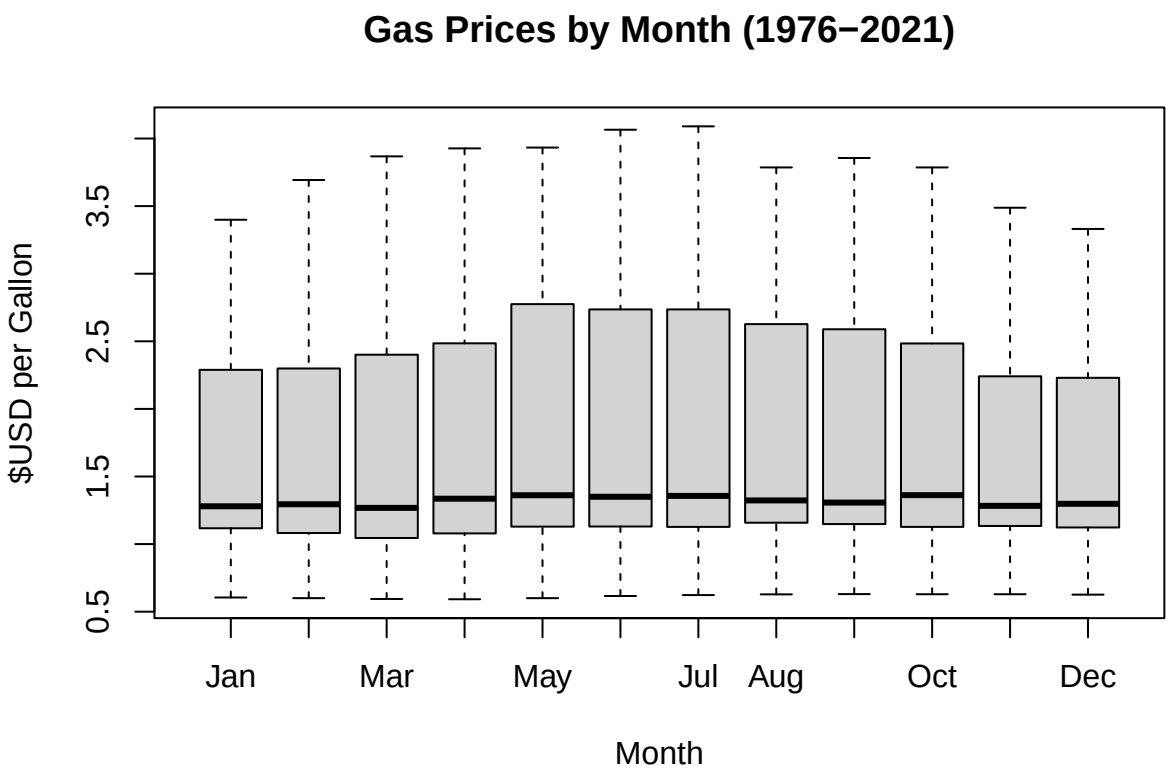
Generate Summary Statistics

```r
# Generate summary stats for variables
summary(gasPrices)
```

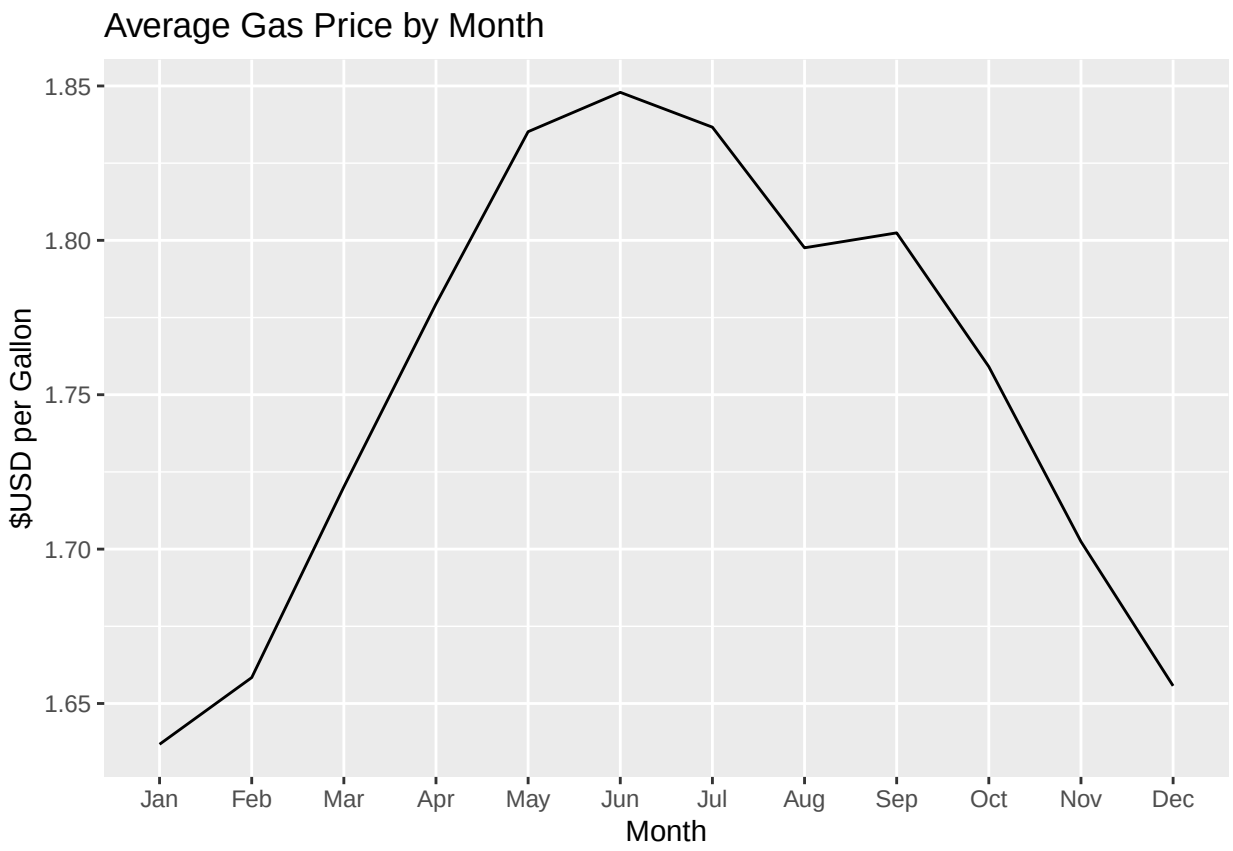
```
##       Year           Jan             Feb             Mar             Apr
##  Min.   :1976   Min.   :0.605   Min.   :0.600   Min.   :0.594   Min.   :0.592
##  1st Qu.:1987   1st Qu.:1.120   1st Qu.:1.089   1st Qu.:1.048   1st Qu.:1.085
##  Median :1998   Median :1.280   Median :1.294   Median :1.268   Median :1.335
##  Mean   :1998   Mean   :1.637   Mean   :1.658   Mean   :1.720   Mean   :1.779
##  3rd Qu.:2010   3rd Qu.:2.285   3rd Qu.:2.296   3rd Qu.:2.381   3rd Qu.:2.468
##  Max.   :2021   Max.   :3.399   Max.   :3.693   Max.   :3.868   Max.   :3.927
##
##       May             Jun             Jul             Aug
##  Min.   :0.600   Min.   :0.616   Min.   :0.623   Min.   :0.628
##  1st Qu.:1.131   1st Qu.:1.135   1st Qu.:1.129   1st Qu.:1.158
##  Median :1.361   Median :1.350   Median :1.357   Median :1.323
##  Mean   :1.835   Mean   :1.848   Mean   :1.837   Mean   :1.798
##  3rd Qu.:2.678   3rd Qu.:2.710   3rd Qu.:2.688   3rd Qu.:2.627
##  Max.   :3.933   Max.   :4.065   Max.   :4.090   Max.   :3.786
##                                                  NA's   :1
##       Sep             Oct             Nov             Dec
##  Min.   :0.630   Min.   :0.629   Min.   :0.629   Min.   :0.626
##  1st Qu.:1.148   1st Qu.:1.127   1st Qu.:1.134   1st Qu.:1.123
##  Median :1.307   Median :1.362   Median :1.283   Median :1.298
##  Mean   :1.802   Mean   :1.759   Mean   :1.702   Mean   :1.656
##  3rd Qu.:2.589   3rd Qu.:2.484   3rd Qu.:2.241   3rd Qu.:2.230
##  Max.   :3.856   Max.   :3.786   Max.   :3.488   Max.   :3.331
##  NA's   :1       NA's   :1       NA's   :1       NA's   :1
```

Plot Some Features of Variables

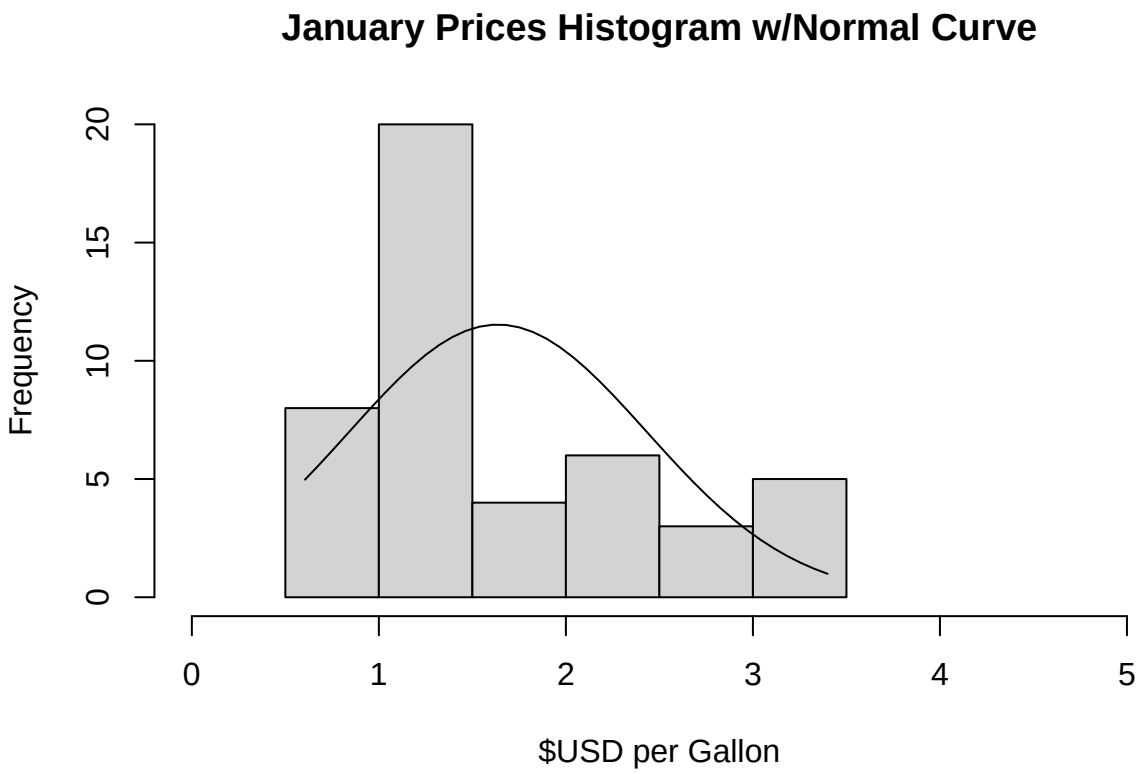
```r
# Plot boxplots
boxplot(gasPricesDF, ylab="$USD per Gallon", xlab="Month",
        main="Gas Prices by Month (1976-2021)")
```



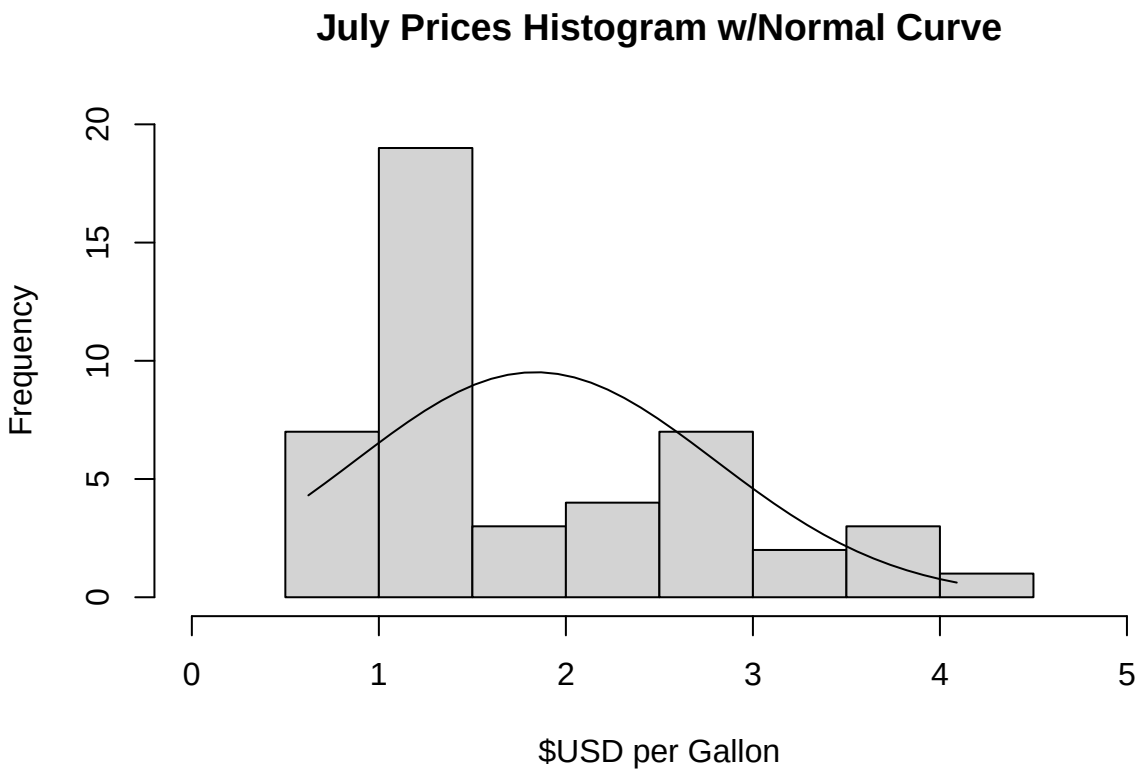

```
# Plot mean values by month
df <- gasPricesDF
df.prices <- df %>% select(Jan:Dec) %>% gather(month, price) # gather prices
df.avg <- df.prices %>% group_by(month) %>%
  summarize(average=mean(price, na.rm=TRUE)) # get mean by month
df.avg$month <- factor(df.avg$month, levels=names(df)) # order by month
ggplot() + geom_line(data=df.avg, aes(x=month, y=average, group=NA)) + # plot
  labs(title="Average Gas Price by Month", x="Month", y="$USD per Gallon")
```



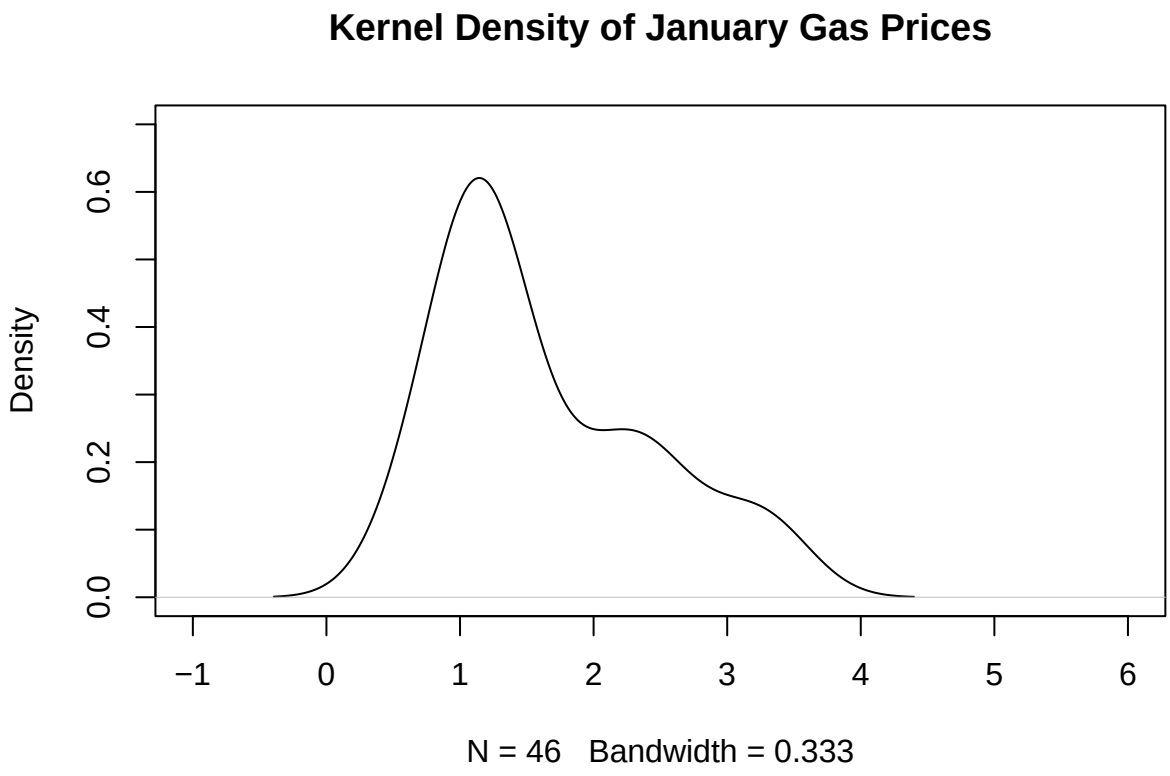

```
# Plot histograms (with normal curve)
x <- gasPrices$Jan
h <- hist(x, xlab="$USD per Gallon", xlim=c(0,5), ylim=c(0,20),
          main="January Prices Histogram w/Normal Curve")
xfit <- seq(min(x), max(x), length=40)
yfit <- dnorm(xfit, mean=mean(x), sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit)
```



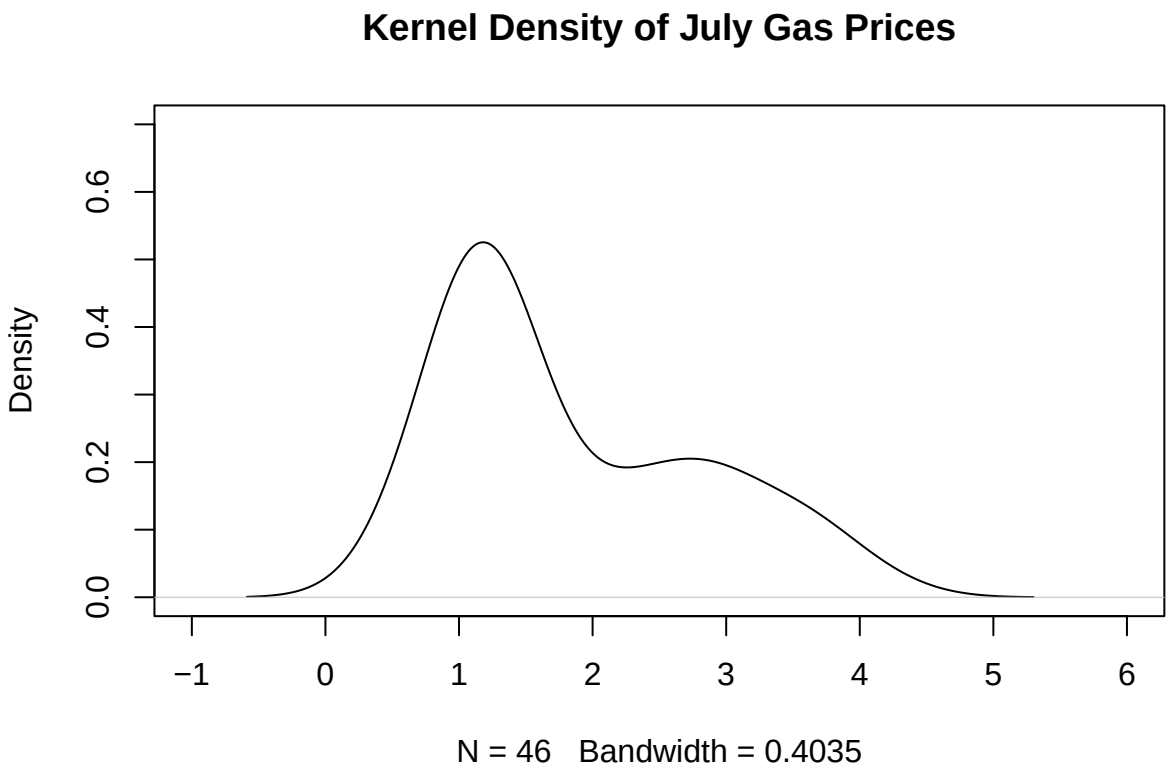

```
x <- gasPrices$Jul
h <- hist(x, xlab="$USD per Gallon", xlim=c(0,5), ylim=c(0,20),
          main="July Prices Histogram w/Normal Curve")
xfit <- seq(min(x), max(x), length=40)
yfit <- dnorm(xfit, mean=mean(x), sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit)
```




```
# Plot kernel density plots
d <- density(gasPrices$Jan)
plot(d, xlim=c(-1, 6), ylim=c(0, .7),
     main="Kernel Density of January Gas Prices")
```




```
d <- density(gasPrices$Jul)
plot(d, xlim=c(-1, 6), ylim=c(0, .7),
     main="Kernel Density of July Gas Prices")
```



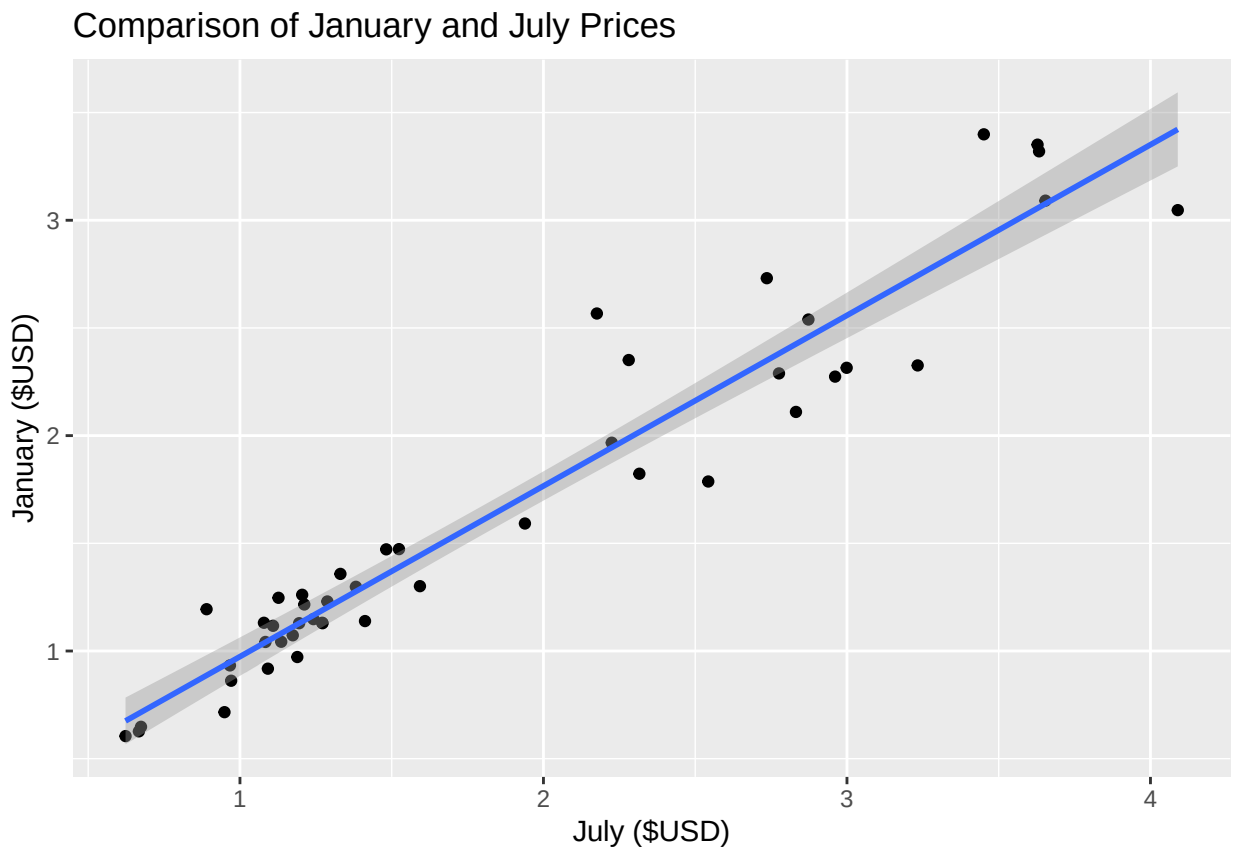
Save CSV File

```
# Write to CSV file
write.csv(gasPricesDF, "gasPrices.csv", row.names=TRUE)
```

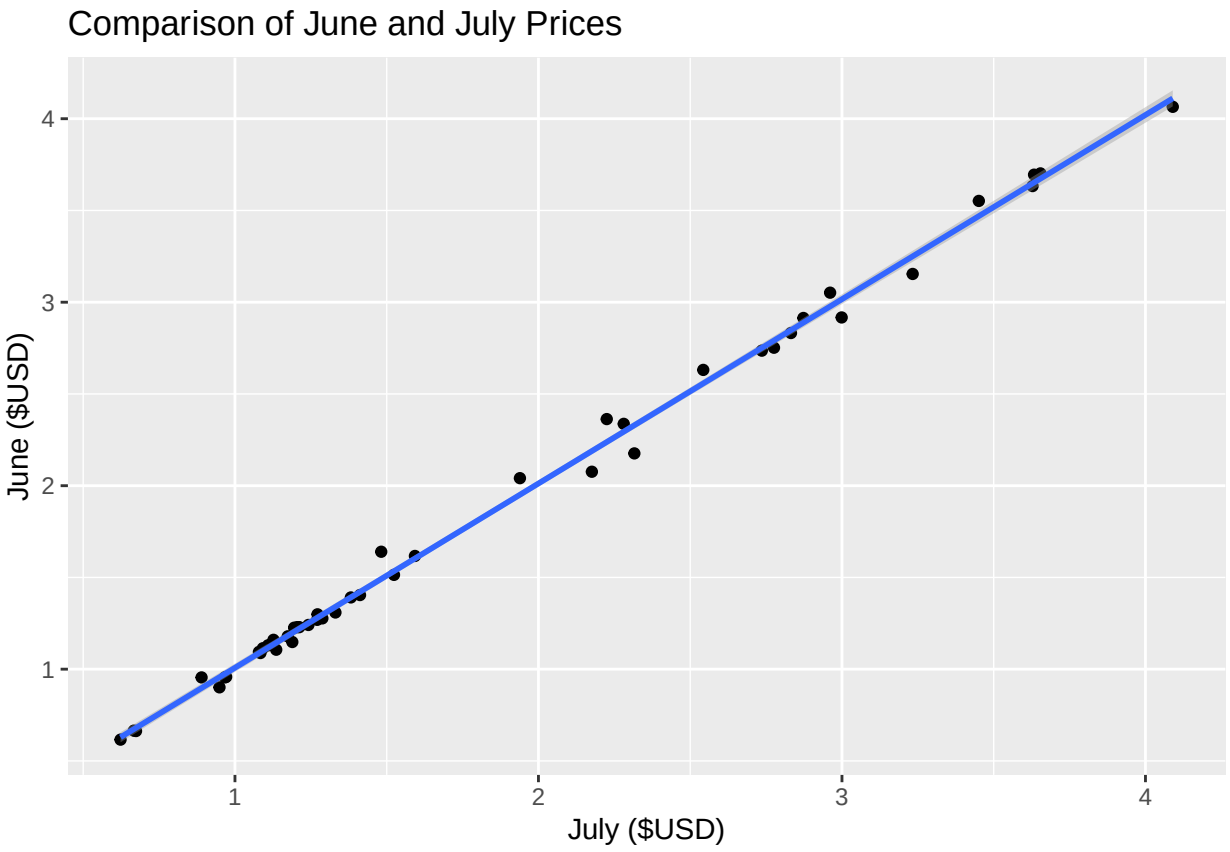
2. *Explore Some Bivariate Relations*

Use the same dataset within the same website to explore some bivariate relations (e.g. bivariate plot, correlation, table cross table etc.).

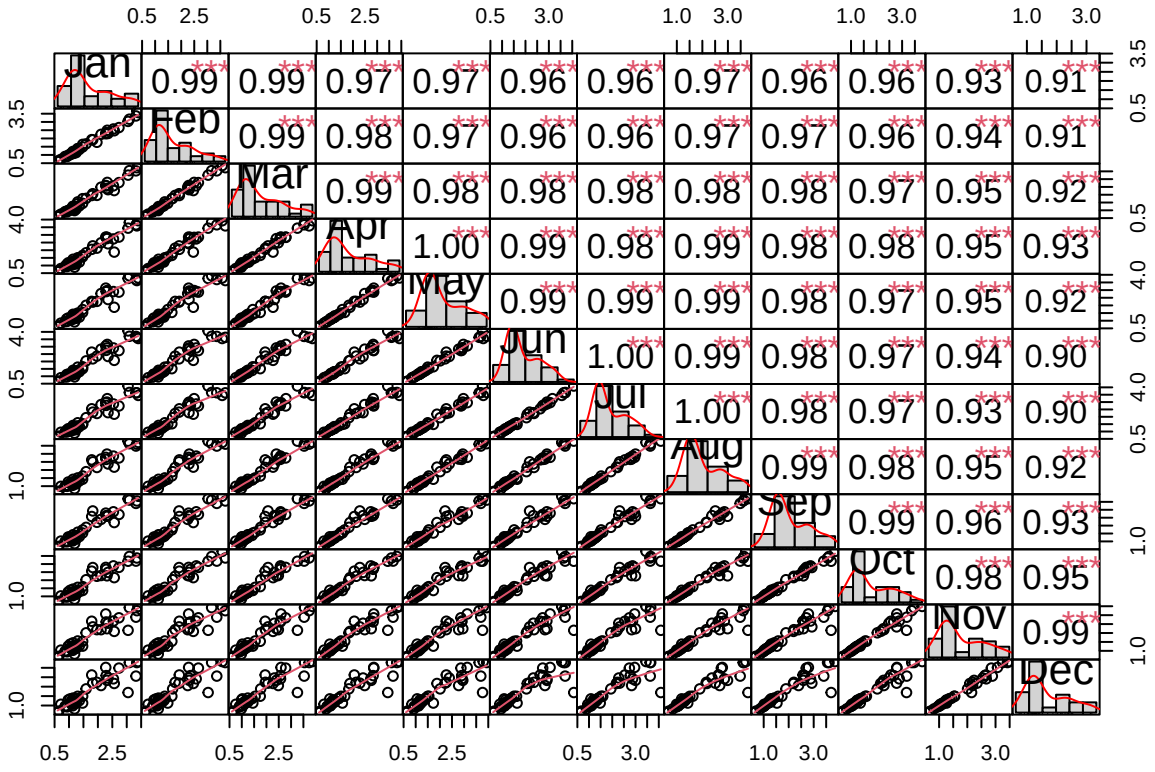
```
# Plot different months with regression line
ggplot(gasPricesDF, aes(x=Jul, y=Jan)) + geom_point() +
  geom_smooth(method="lm", formula='y ~ x') +
  labs(title="Comparison of January and July Prices",
       x="July ($USD)", y="January ($USD)")
```




```
ggplot(gasPricesDF, aes(x=Jul, y=Jun)) + geom_point() +
  geom_smooth(method="lm", formula='y ~ x') +
  labs(title="Comparison of June and July Prices",
       x="July ($USD)", y="June ($USD)")
```




```
# Chart correlation matrix
chart.Correlation(df, histogram=TRUE, pch=19)
```



3. *Organize a Data Report*

Generate a summary report. Make sure to include: summary for every variable, structure and type of data elements, discuss four results of your data.

Summary Report:

If the Facebook posts of uncles across the US are any indication, gas prices are an important indicator of how we're doing as a society. In an effort to gain a deeper understanding of the topic, I selected the data series: "Gasoline, unleaded regular, per gallon/3.785 liters in U.S. city average, average price, not seasonally adjusted", or the monthly average gas price for short.

Since this is just tracking gas prices, there isn't a great deal of variety in the data, with the main variables being the Year or the Month. I selected the entire available data set, which goes back to the beginning of 1976 and up through the month of July 2021 (I imagine August will be added soon). Because the data for the current year only goes through July, I was left with Null values for August through December 2021 that I had to deal with in some instances.

All of the gas price values are in US Dollars ($), to the thousandth of a decimal and thus were loaded into R as Doubles. The 'Year' column also loaded as a double and I played around a bit with converting the years to characters, but ultimately ended up converting the tibble to a data frame and setting the years as the row names. This removed them from some of the calculations and charts, which is useful because they aren't really relevant in the comparisons.

You can see in the box plots and the *Average Gas Price by Month* chart that gas prices tend to increase in the summer months. You also see that there is more variation in prices in the summer months, with an

increase starting around May. This makes sense because May is typically when school ends and summer vacations begin, which often consist of a significant amount of driving.

You can see by looking at the Histograms that, while they are fairly right-skewed (this is typical in charts dealing with money), there is still quite a bit of variability and they do not fit a normal curve very well. I selected January and July for comparisons because they are halfway points for the year. I thought that they may be more representative overall and they won't be too similar because they are not too close to each other. I added kernel density curves in the following charts to better represent the distribution for these months.

When looking at bivariate distributions, you can see by looking at the two scatter plots that January and July have a pretty good correlation, but June and July is much tighter (as reflected by the linear regression line). I expect this is because there isn't typically a lot of change from one month to the next. To investigate that further, I made a Correlation Chart showing all of the months with scatter plots, correlations, and p-values for each. You can see that with the exception of December to January, the closer months are in proximity, the higher their R-squared values. Note: the p-values are represented by asterisks; 3, 2, and 1 stars represent p-values of 0.001, 0.01, and 0.1, respectively.