Problem Set 2 Intructions

Version 1.0

2022-09-22

Notification:

1. For Problem Set 2, we will implement peer grading using Kritik.
2. Please use **R Markdown** to build your submission. Be sure that for each R chunk in your markdown file has **echo = TRUE** so your code is included in your submission. You can do this for all the R Chunks by using this setup at the top of the markdown file:

Important Notes

For each of your submission plots be sure to include the following:

- Title
- Labels on your axes, or a key
  - Be sure the labels are meaningful outside of the context of your code.
- Make sure the elements in your plot are appropriately scaled (sized to increase interpret-ability).
- A one sentence summary about your plot and its contents.

Fuel Economy

You have been dealing with fuel economy data since bootcamp, using the mtcars and mpg built-in data sets. In the US, the home of fuel economy data is the Department of Energy website which has an extensive downloads page. The two files for this problem are **vehicles.csv** and **index.html**.

To get started, import **vehicles.csv** with the **import > From Text(base)** function in the Environment panel in RStudio. **index.html** will display in your default browser when you double-click it in your file browser.

When you have loaded **vehicles.csv** into R, you will see that it has 45,471 rows of data and 83 variables. It covers cars from 1984 through 2023. Before you can do anything with this data you must explore it. The data dictionary is in **index.html**. In your browser, scroll down to get to the data dictionary.

Now use what you learned about sub-setting to explore. (We'll talk about more modern approaches to this exploration next week.) But for this assignment, I'll start you with some simple exploration.

Start by looking at vehicles made by Mazda:

```
m2 <- vehicles[vehicles$make == "Mazda",c("year", "make", "model", "mpgData")]
```

You can see in the **mpgData** variable that not every row will have mpgData. Don't be too fast to get ride of the rows that don't report mpg, however.

This data set include electric and hybrid vehicles – note the variables **phevBlended** and **highwayE**. Look them up in the data dictionary.

So now make another, more useful, subset if the vehicles data.
It includes phevBlended which I can see takes on logical values TRUE and FALSE.

```
m3 <- vehicles[vehicles$year == "2022",c("year", "make", "model", "mpgData", "phevBlended")]
```

When I search, however, I don't get anything. Are there really no hybrid cars in the data?

```
m3[m3$phevBlended == TRUE,]  ## nothing is returned
```

```
## [1] year        make        model       mpgData     phevBlended
## <0 rows> (or 0-length row.names)
```

```
## and for further evidence, I try
try(sum(m3$phevBlended))
```

```
## Error in sum(m3$phevBlended) : invalid 'type' (character) of argument
```

```
## it throws a type error  --  OH!! Of course. The csv file had strings

## confirm the problem
typeof(m3$phevBlended)
```

```
## [1] "character"
```

```
## fix it
m3$phevBlended <- as.logical(m3$phevBlended)
sum(m3$phevBlended)
```

```
## [1] 44
```

OK, for the year 2022 there are 44 hybrid vehicles in the dataset. Now look at eletric vehicles. Build another subset with variable you might have ignored at first. Again, look them up.

```
m6 <- vehicles[vehicles$year == "2022",
               c("year", "make", "model", "mpgData",
                 "phevBlended", "highwayE",
                 "fuelType", "fuelType1", "fuelType2",
                 "barrels08", "barrelsA08", "charge120", "charge240")]

## don't forget to fix this problem again
m6$phevBlended <- as.logical(m6$phevBlended)
sum(m6$phevBlended)               ## reconfirm the fix
```

```
## [1] 44
## take a look at the hybrid cars
head(m6[m6$phevBlended == TRUE, ])
```

```
##       year    make                     model mpgData phevBlended highwayE
## 37319 2022    MINI Cooper SE Countryman All4       N        TRUE       47
## 37386 2022 Hyundai   Santa Fe Plug-in Hybrid       N        TRUE       47
## 37534 2022    Audi         A7 TFSI e quattro       N        TRUE       44
## 37535 2022    Audi         Q5 TFSI e quattro       N        TRUE       55
## 37537 2022 Hyundai      Ioniq Plug-in Hybrid       Y        TRUE       28
## 37594 2022     Kia       Niro Plug-in Hybrid       Y        TRUE       34
##                           fuelType        fuelType1   fuelType2 barrels08
## 37319     Premium and Electricity Premium Gasoline Electricity  5.839945
## 37386 Regular Gas and Electricity Regular Gasoline Electricity  3.772597
## 37534     Premium and Electricity Premium Gasoline Electricity  4.256578
## 37535     Premium and Electricity Premium Gasoline Electricity  5.016688
## 37537 Regular Gas and Electricity Regular Gasoline Electricity  2.494817
## 37594 Regular Gas and Electricity Regular Gasoline Electricity  3.016750
##       barrelsA08 charge120 charge240
## 37319   4.075479         0      2.00
## 37386   3.914605         0      3.40
## 37534   4.250143         0      3.00
## 37535   4.877213         0      3.00
## 37537   2.500084         0      2.25
## 37594   2.833429         0      2.25
```

Your Assignment

1. Do an analysis of Fuel economy over the 40 year span 1984 through 2023, inclusive. You may want to do an analysis by type of fuel which will ignore hybrids and electric vehicles for most the the years under analysis.

2. Now, examine vehicle makers. Which ones have made the most progress? Make at least two plots that address the questions above. As you do your work, you may make many plots. If you include plots in addition to the two that described above, make sure that they address different issues and are not simply intermediate steps you took as you made the to plots you're submitting for questions 1 and 2.

NASDAQ Composite

The Nasdaq Composite (ticker symbol ^IXIC) is a stock market index that includes almost all stocks listed on the Nasdaq stock exchange. Along with the Dow Jones Industrial Average and S&P 500, it is one of the three most-followed stock market indices in the United States.

Your Assignment

Use ggplot to create a Candlestick chart with the Nasdaq Composite data from September 20, 2021 to September 20, 2022, using file `IXIC21-22.csv`. If you don't know what the Candlestick chart is, the following link might give some information: **Candlestick**.

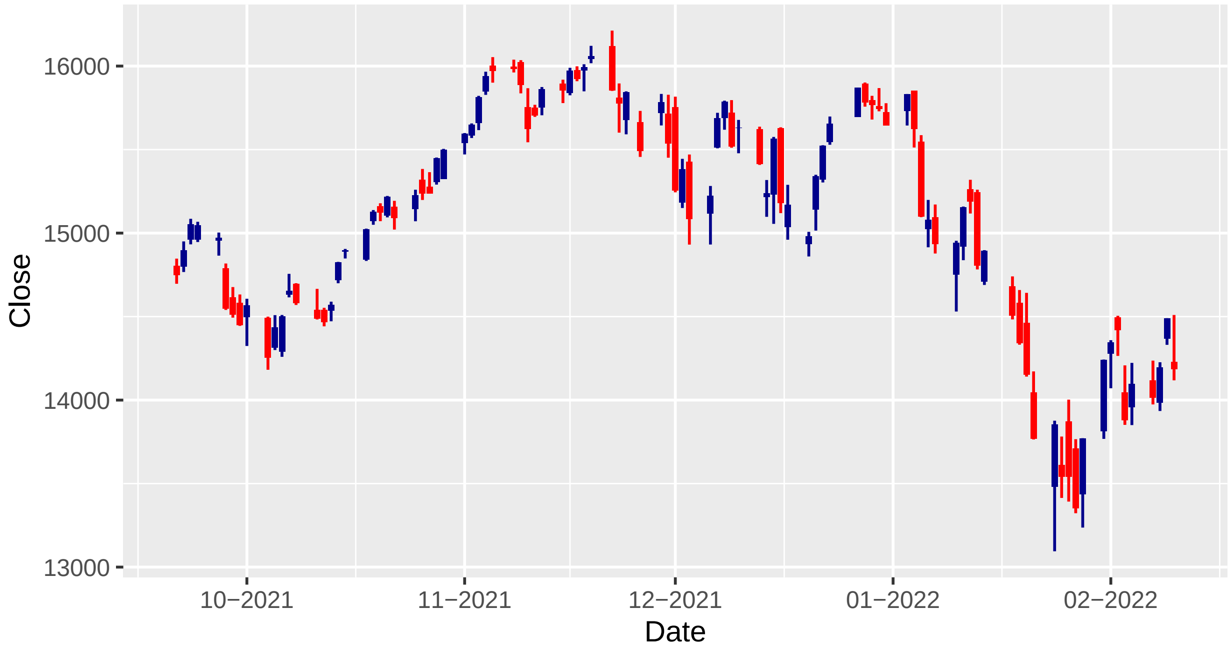
For your convenience, the following links provide some examples might help you to create the chart:

- **Candlestick Example 1**.
- **Candlestick Example 2**

Be aware of the following when you create your chart:

- Make sure the X-axis is **Date** and the Y-axis is **Adjust Close**.
- Include an appropriate title with your graph.
- Make sure your data is clean before you create the plot.
- The format for X label will be **month-year or year-month** eg, Jul-2022, 07-2022, 2022-07
- You might need to use `tidyquant` package to create this plot.
- In a sentence or two, what does this plot show?

Here is an example of the candlestick plot:



Rural Capacity Index

The Rural Capacity Index (RCI) was created by Headwaters Economics to assess municipal "capacity". Capacity being "the staffing, resources, and expertise—to apply for funding, fulfill onerous reporting requirements, and design, build, and maintain infrastructure projects over the long term." (Hernandez, 2022). The RCI consolidates various data from the American Community Survey, conducted by the US Census: education levels, populations, health insurance, and even broadband access. `ruralCapacityData.csv` contains the RCI and its constituents about the counties in New Mexico.

New Mexican urban and rural communities have very disparate characteristics. Bernillio County, containing New Mexico's most populous urban setting, has over 1500 times the population of Harding County. During your exploration of the *Rural* Capacity Index you may want to omit Berniliio County to better represent the more numerous and varied small communities.

Your Assignment

1. Create a plot that emphasizes rural capacity indexes. Choose your other variables to reflect their contribution to the rural capacity index.
2. Create a plot that demonstrates the relationship between the number of houses with broadband and the percent of adults, 25 and older, with bachelor degrees. Include information about the rural capacity indexes.
3. Explore different sizes of communities and their capacity indexes. Create three plots that describe communities with total population < 16000, 16000 < total population < 55000, and total population > 55000. What facets of each population subsection stand out to you, demonstrate them in your plots.

Column descriptions about the RCI data:

- cap_index – The Rural Capacity Index, between 0 and 100.
- pop_over_25 – Population 25 years and older.
- pop_bachelors – Population with a bachelors degree.
- per_over_25_with_bach – Percent of population 25 and over with a bachelors degree.
- num_fam – Number of families.
- per_fam_below_pov – Percentage of families below the poverty level.
- tot_house – Total households.
- house_broadband – Total households with broadband connection.
- pop_noninst – Non-institutionalized population.
- pop_insured – Population with health insurance.
- per_insured – Percentage of population with health insurance.
- pop_uninsured – Population without health insurance.
- per_uninsured – Percent of population without health insurance.
- pop_total – Total population.

Problem 4

Coming soon...