STA 404/504 Homework 2

Due Friday Feb 28th, by 5:00pm, submit through canvas

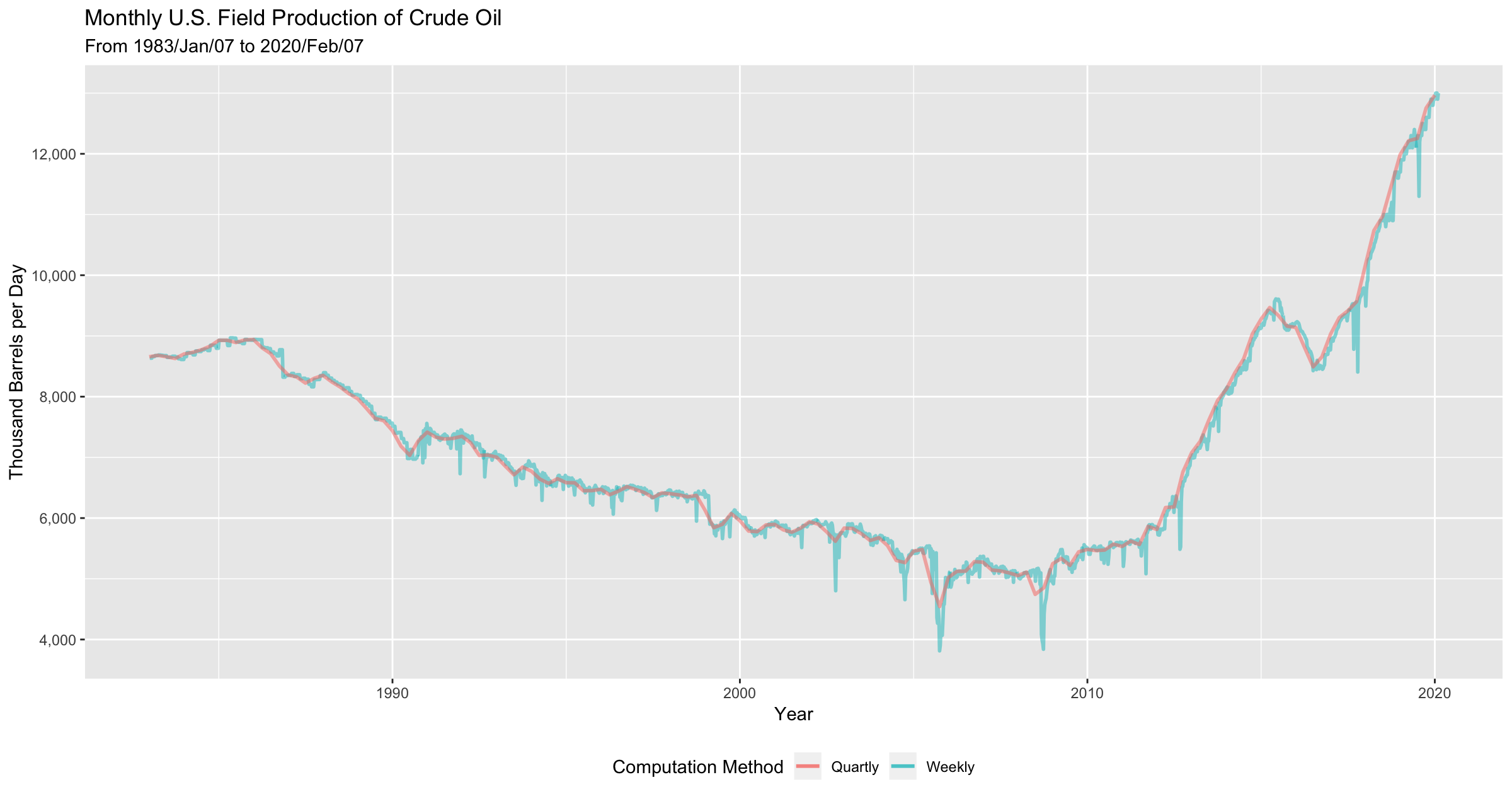
**Learning Objectives:**

* Forming a data cleaning plan, break it into smaller tasks and complete it step by step
* Reading in data with poorly structured heading
* Restructuring specific columns from wide format to tall format
* Work with time formatting variables and strings
* Documenting a fully reproducible data cleaning/visualization process in a scripted language

**Assignment Description:**

The dataset “US\_Crude\_Oil.xlsx” on canvas contains the average weekly U.S. field production of crude oil (thousand barrels per day). The data is from the first week of 1983 to the first week of February in 2020. The goal of this assignment is to use this data to create time series line plots. The assignment for 404 and 504 is different:

* **For students who take 404**: Create two time series line plots, one if for the weekly U.S. field production of crude oil, the other one is for the monthly U.S. field production of crude oil. The format of the picture should look similar like the one in this website: <https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=PET&s=WCRFPUS2&f=W>
* **For students who take 504:** Apart from the two time series line plots described, please provide a third time series line plot that looks like the image below. Basically, include both the weekly and quarterly line plots in one picture and color mark them. (You don’t need to follow exactly the colors or line shapes, but the audience should be able to distinguish the two lines and see the overlap clearly.



In addition, **for all students,** when creating the plots, make sure you add at least one additional aesthetic feature, that has not been used in class, to the plot. For example, you may change the shape of the line, color, background, etc.

**General Problem Solving Process:**

The problem is that the data is very messy and improperly structured to create this visualization. You may complete the assignment follow the following steps:

Step 1: Clean and restructure the data in R using the dplyr, tidyr, stringr, and lubridate packages so that the cleaned data is saved as data frame that has **only two columns: date and production.**

The date column should be converted from a character to a POSIX format (POSIX dates consists of the year, followed by the month and day, separated by slashes or dashes) using a function from the lubridate package. You may open the data file in a spreadsheet editor like excel to take a look at the structural issues, but all the work about data cleaning must be done in R. Others will be able to run your code and get the cleaned data, without editing the code, except for changing the working directory.

Step 2: Use appropriate *ggplot()* code to make plots. You will need to utilize some appropriate *ggplot()* options that has not been discussed in class.

**Some Hints:**

1. To read in the excel file, load the package readxl, and use the function read\_excel(). The header of this data is two rows of poorly formatted labels. When you read in the data you may want to use the skip= option.
2. The first column has the year/month combined followed by five pairs of columns for the dates and productions associated with weeks 1 through 5 of each month. Each of these five column pairs will need to be moved from wide format to tall format using functions in the tidyr package.
3. Consider separating the data into two data sets (one for dates and the other for the productions). Then combine the two together after cleanup. You may search for “Join two tbls together” to find a useful way that combines these two datasets.
4. You will need to do some string handling to get things working.
5. To get the y-axis unemployment claims in a format such as 10,000, try the option labels=scales::comma in scale\_y\_continuous().

**Submission:**

1. **An R code file** that reads in data from the original file “US\_Crude\_Oil.xlsx”, loads the necessary packages, cleans the data and creates plots as instructed. Please be sure to properly document and organize your code so that the data cleaning and plot creation processes are clear. The instructor will be able to replicate your results by running the code you provided.

The entire data cleaning process must be conducted in R, no cleaning “by hand” may be done in excel before loading to R.

1. **A pdf file** that is converted from word containing the code, the outputs (the plots).

Note: Alternatively, you may use Rmarkdown to generate a pdf, or html file, including all your answers and discussions. In this case, please upload the .Rmd file, as well as the knitted html/pdf file.

**Grading Rubric:**

* **For students who take 404**:

10 points - Successful data cleaning process (header, wide-to-tall, string handling, date formatting)

4 points - Clarity of documentation, structure, and comments in code; meaningful variable names; code readability

6 points – Plots created similar to the figures in the reference website (structure, labels, formats). Make sure at least one additional aesthetic feature, that have not been used in class, is added to the plot.

* **For students who take 504**:

7 points - Successful data cleaning process (header, wide-to-tall, string handling, date formatting)

4 points - Clarity of documentation, structure, and comments in code; meaningful variable names; code readability

9 points – Plots created similar to the figures in the reference website, as well as the provided plot above (structure, labels, formats). Make sure at least one additional aesthetic feature, that have not been used in class, is added to the plot.