STA504 HW2 Jessica Choe

#1. Call libraries

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.2
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.
3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

```
## Warning: package 'purrr' was built under R version 3.6.2
```

```
## Warning: package 'forcats' was built under R version 3.6.2
```

```
## -- Conflicts ------------------------------------------- tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.6.2
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.6.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(stringr)
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 3.6.2
```

#2. Read in data and clean column name

```
#Read in columne name (second row) from the data
col_names <- array(read_excel('C:/Users/user/Desktop/2020spring/504 data visu
alization/hw/hw2/US_Crude_Oil.xlsx', sheet = 'Sheet1', n_max = 1,skip=1, col_
names = FALSE))
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ... and 5 more problems
```

```
#Read in the entire data except column name(from 4th row)
oil <- data.frame(read_excel('C:/Users/user/Desktop/2020spring/504 data visua
lization/hw/hw2/US_Crude_oil.xlsx', sheet = 'Sheet1', skip = 3, col_names = F
ALSE))
```

```
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ... and 6 more problems
```

```
options(pillar.sigfig = 8)

#insert the columne name into oil data.
colnames(oil) <- col_names
head(oil,2)
```

```
##   Year-Month Week 1       NA Week 2       NA Week 3       NA Week 4
## 1   1983-Jan 01/07  8,634    01/14  8,634    01/21  8,634    01/28
## 2   1983-Feb 02/04  8,660    02/11  8,660    02/18  8,660    02/25
##         NA Week 5  NA
## 1 8,634      <NA>
## 2 8,660      <NA>
```

```
#Give names to columns with no name.
names(oil)[3] <-"Wk1"
names(oil)[5] <-"Wk2"
names(oil)[7] <-"Wk3"
names(oil)[9] <-"Wk4"
names(oil)[11] <-"Wk5"
head(oil,2)
```

```
##   Year-Month Week 1      Wk1 Week 2      Wk2 Week 3      Wk3 Week 4
## 1   1983-Jan 01/07  8,634    01/14  8,634    01/21  8,634    01/28
## 2   1983-Feb 02/04  8,660    02/11  8,660    02/18  8,660    02/25
##        Wk4 Week 5 Wk5
## 1 8,634      <NA>
## 2 8,660      <NA>
```

```
#remove any unnecessary hidden dots in Year-Month column
oil$`Year-Month`<-str_trim(oil$`Year-Month`)

head(oil,5)
```

```
##   Year-Month Week 1      Wk1 Week 2      Wk2 Week 3      Wk3 Week 4
## 1   1983-Jan 01/07  8,634    01/14  8,634    01/21  8,634    01/28
## 2   1983-Feb 02/04  8,660    02/11  8,660    02/18  8,660    02/25
## 3   1983-Mar 03/04  8,677    03/11  8,677    03/18  8,677    03/25
## 4   1983-Apr 04/01  8,677    04/08  8,686    04/15  8,686      <NA>
## 5   1983-May   <NA>          05/13  8,682    05/20  8,682      <NA>
##        Wk4 Week 5      Wk5
## 1 8,634      <NA>
## 2 8,660      <NA>
## 3 8,677      <NA>
## 4          04/29  8,686
## 5            <NA>
```

#3. Create a tall table for “date”

```
# From wide to tall table for date.
oil_date <-oil%>%
  select(`Year-Month`,
         `Week 1`,`Week 2`,`Week 3`,`Week 4`, `Week 5`)%>%      gather(key="W
eek",value="Month-Date",2:6)

head(oil_date)
```

```
##   Year-Month   Week Month-Date
## 1   1983-Jan Week 1     01/07
## 2   1983-Feb Week 1     02/04
## 3   1983-Mar Week 1     03/04
## 4   1983-Apr Week 1     04/01
## 5   1983-May Week 1       <NA>
## 6   1983-Jun Week 1     06/03
```

```
# remove any rows with NA from the tall table(oil_date).
oil_date<-oil_date[complete.cases(oil_date), ]
head(oil_date)
```

```
##   Year-Month   Week Month-Date
## 1   1983-Jan Week 1     01/07
## 2   1983-Feb Week 1     02/04
```

```
## 3   1983-Mar Week 1     03/04
## 4   1983-Apr Week 1     04/01
## 6   1983-Jun Week 1     06/03
## 7   1983-Jul Week 1     07/01

# Create date variable in POSIX format.
date_long <- oil_date %>%
  mutate(Year=str_sub(`Year-Month`,1,4),
         Date.string = paste0(Year, "/",`Month-Date`))
head(date_long)

##   Year-Month   Week Month-Date Year Date.string
## 1   1983-Jan Week 1     01/07  1983 1983/01/07
## 2   1983-Feb Week 1     02/04  1983 1983/02/04
## 3   1983-Mar Week 1     03/04  1983 1983/03/04
## 4   1983-Apr Week 1     04/01  1983 1983/04/01
## 5   1983-Jun Week 1     06/03  1983 1983/06/03
## 6   1983-Jul Week 1     07/01  1983 1983/07/01

#change the date in string format into POSIX format.
date_long2 <- date_long %>%
  mutate(Date = ymd(Date.string)) %>%
  arrange(Date) %>%
  select(`Year-Month`,Week,Date)

head(date_long2,8)

##   Year-Month   Week       Date
## 1   1983-Jan Week 1 1983-01-07
## 2   1983-Jan Week 2 1983-01-14
## 3   1983-Jan Week 3 1983-01-21
## 4   1983-Jan Week 4 1983-01-28
## 5   1983-Feb Week 1 1983-02-04
## 6   1983-Feb Week 2 1983-02-11
## 7   1983-Feb Week 3 1983-02-18
## 8   1983-Feb Week 4 1983-02-25
```

#4. Create a second tall table for “Production” and combine 2 tables together into one table

```
#Select weekly production amount variables together with Year-Month variable.
production_tall<-oil%>%
select(`Year-Month`,
       `Wk1`,`Wk2`,`Wk3`,`Wk4`,`Wk5`)

#Change the column names so that later we can combine the two tables with sam
e value.
colnames(production_tall)<-c("yearmonth","Week 1","Week 2","Week 3","Week 4",
"Week 5")
head(production_tall)
```

```
##   yearmonth   Week 1   Week 2   Week 3   Week 4   Week 5
## 1  1983-Jan 8,634    8,634    8,634    8,634
## 2  1983-Feb 8,660    8,660    8,660    8,660
## 3  1983-Mar 8,677    8,677    8,677    8,677
## 4  1983-Apr 8,677    8,686    8,686             8,686
## 5  1983-May          8,682    8,682
## 6  1983-Jun 8,676    8,676    8,676    8,676
```

```
#From Wide to Tall format
production_tall<-production_tall%>%
  gather(key="production_wk",value="Production",2:6)
head(production_tall,3)
```

```
##   yearmonth production_wk Production
## 1  1983-Jan        Week 1   8,634
## 2  1983-Feb        Week 1   8,660
## 3  1983-Mar        Week 1   8,677
```

```
#Find white cells(empty-looking cells) with hidden character"   ", change it
into NA, and then erase the row if it contains NA
hidden_dots=production_tall[5,3]
production_tall<-production_tall %>%
  mutate_all(~ifelse(. %in% c("null",hidden_dots),NA,.)) %>%
  na.omit()
head(production_tall)
```

```
##   yearmonth production_wk Production
## 1  1983-Jan        Week 1   8,634
## 2  1983-Feb        Week 1   8,660
## 3  1983-Mar        Week 1   8,677
## 4  1983-Apr        Week 1   8,677
## 6  1983-Jun        Week 1   8,676
## 7  1983-Jul        Week 1   8,676
```

```
#Combine two tables together.
colnames(date_long2)
```

```
## [1] "Year-Month" "Week"       "Date"
```

```
colnames(production_tall)
```

```
## [1] "yearmonth"     "production_wk" "Production"
```

```
date_production<-date_long2 %>% inner_join(production_tall, by =c("Year-Month
"="yearmonth","Week"="production_wk"))
head(date_production)
```

```
##   Year-Month   Week       Date Production
## 1   1983-Jan Week 1 1983-01-07   8,634
## 2   1983-Jan Week 2 1983-01-14   8,634
## 3   1983-Jan Week 3 1983-01-21   8,634
```

```
## 4   1983-Jan Week 4 1983-01-28   8,634
## 5   1983-Feb Week 1 1983-02-04   8,660
## 6   1983-Feb Week 2 1983-02-11   8,660
```

```
#remove whitespace and "," in Production column and then change its type in n
umeric
head(date_production,2)
```

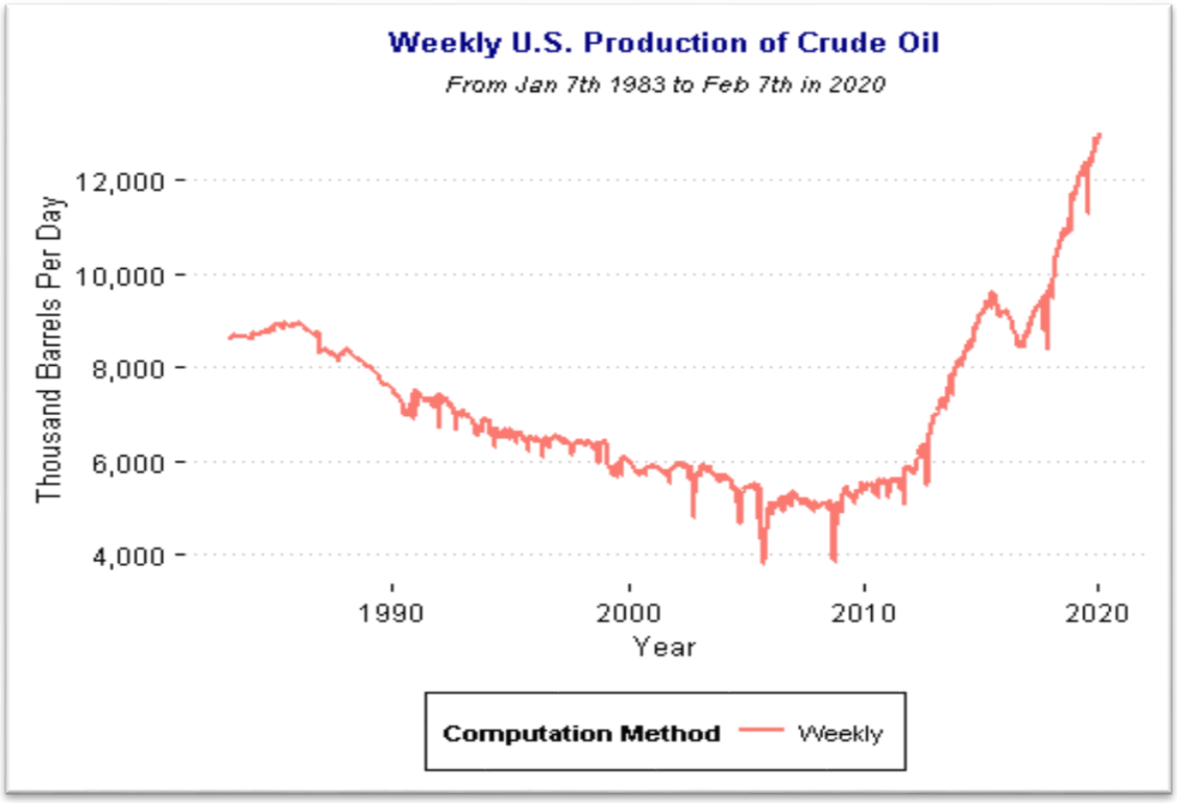
```
##   Year-Month   Week       Date Production
## 1   1983-Jan Week 1 1983-01-07   8,634
## 2   1983-Jan Week 2 1983-01-14   8,634
```

```
date_production$Production<-str_trim(date_production$Production)
date_production$Production<-as.numeric(gsub(",","",date_production$Production
))

#Show two columns only as final table.
date_production_only<-select(date_production,Date,Production)
head(date_production_only)
```

```
##         Date Production
## 1 1983-01-07       8634
## 2 1983-01-14       8634
## 3 1983-01-21       8634
## 4 1983-01-28       8634
## 5 1983-02-04       8660
## 6 1983-02-11       8660
```

#5. Create 3 plots (Question 2) # Create first 2 plots using Weekly(Plot1) and Monthly(Plot2) data

```
#Plot 1. Weekly average production amount
ggplot() +
  geom_line(aes(x=Date,y=Production,color="Weekly"),
            data=date_production,size=1)+
  scale_y_continuous(label = scales::comma)+  labs(x="Year",y="Thousand Barre
ls Per Day",color="Computation Method",size=2)+
  ggtitle("Weekly U.S. Production of Crude Oil",
      subtitle="From Jan 7th 1983 to Feb 7th in 2020")+
  theme_clean()+
  theme(plot.title = element_text(size = 10,
        hjust=0.5,face = "bold", color="navy"),
        plot.subtitle=element_text(size=8, hjust=0.5, face="italic"),
      legend.text=element_text(size=8),
        legend.title=element_text(size=8))+
  theme(legend.position= "bottom")
```




```
# Plot 2. Quarterly average production amount

#Calculate average monthly production and create month format.(The first date
 of each month will represent each month itself.(example: 1983-01-01 means Ja
nuary 1983))
monthly_production <-date_production %>%
  group_by(`Year-Month`) %>%   summarise(monthly_production=mean(Production))

head(monthly_production)
```

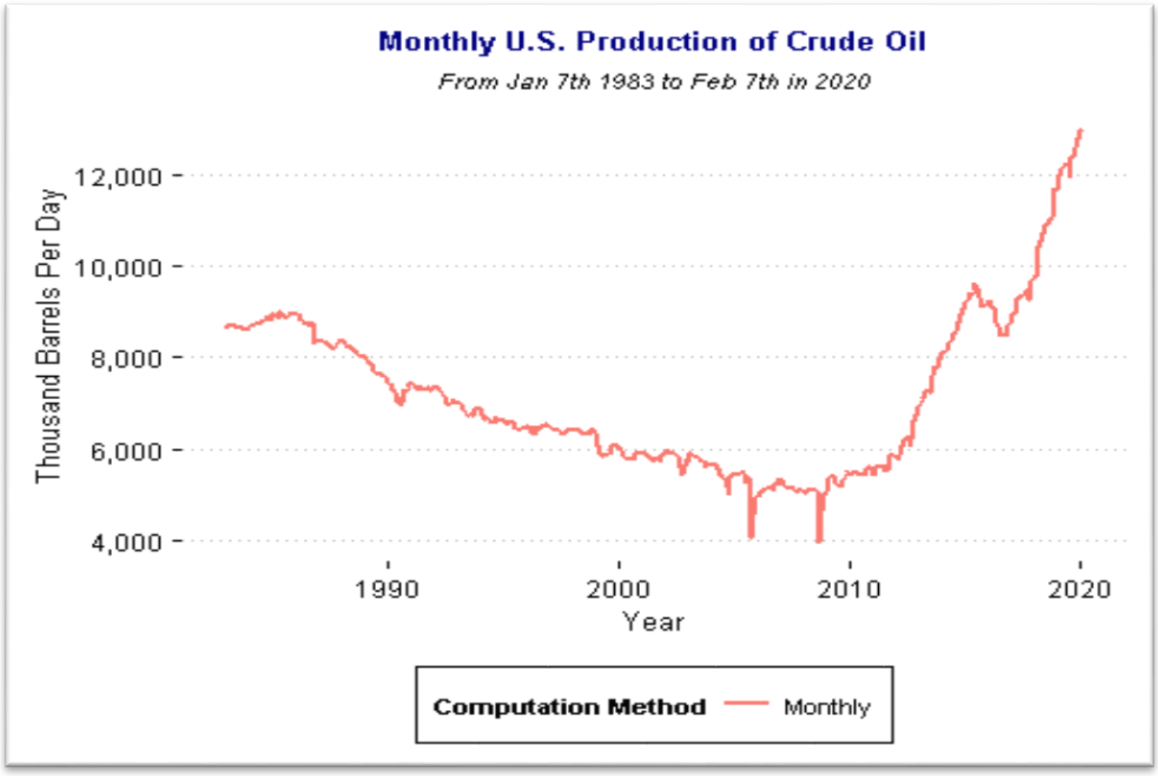
```
## # A tibble: 6 x 2
##   `Year-Month` monthly_production
##   <chr>                     <dbl>
## 1 1983-Apr                8683.75
## 2 1983-Aug                8653
## 3 1983-Dec                8612
## 4 1983-Feb                8660
## 5 1983-Jan                8634
## 6 1983-Jul                8652.800
```

```
monthly_string <- monthly_production %>%
  mutate(month.string = paste0(`Year-Month`,"-01"))
```

```r
monthly_production2 <-monthly_string %>% mutate(Month = ymd(`month.string`))
head(monthly_production2)
```

```
## # A tibble: 6 x 4
##   `Year-Month` monthly_production month.string Month
##   <chr>                     <dbl> <chr>        <date>
## 1 1983-Apr                8683.75  1983-Apr-01  1983-04-01
## 2 1983-Aug                8653     1983-Aug-01  1983-08-01
## 3 1983-Dec                8612     1983-Dec-01  1983-12-01
## 4 1983-Feb                8660     1983-Feb-01  1983-02-01
## 5 1983-Jan                8634     1983-Jan-01  1983-01-01
## 6 1983-Jul                8652.800 1983-Jul-01  1983-07-01
```

```r
#Second plot (Monthly plot)

ggplot()+
  geom_line(aes(x=Month,y=monthly_production,
                color="Monthly"),size=1,
            data=monthly_production2)+
  scale_y_continuous(label = scales::comma)+
  labs(x="Year",y="Thousand Barrels Per Day",color="Computation Method",size=
1)+
  ggtitle("Monthly U.S. Production of Crude Oil",
          subtitle="From Jan 7th 1983 to Feb 7th in 2020")+
  theme_clean()+
  theme(plot.title = element_text(size = 10,
        hjust=0.5,face = "bold", color="navy"),
        plot.subtitle=element_text(size=8, hjust=0.5, face="italic"),
      legend.text=element_text(size=8),
        legend.title=element_text(size=8))+
  theme(legend.position= "bottom")
```

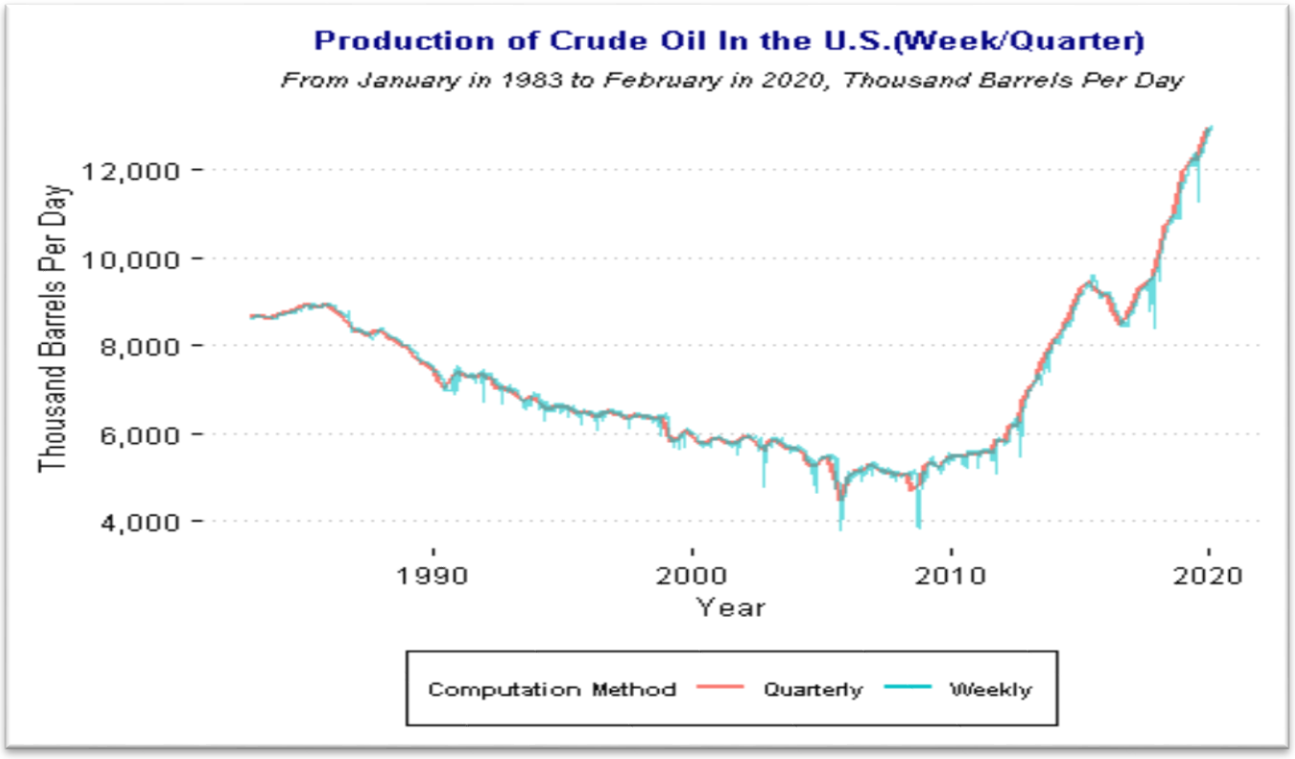


#6. Plot 3. Combined plot (Quarterly, Yearly)

```
#Categorize each month into the combination of year and quarter
quarterly_production <- monthly_production2 %>%
  mutate(Quarter=paste(year(Month),"-",quarter(Month)),
         quarter.num=quarter(Month))%>%group_by(Quarter)

# Change the order of quarter(1st/2nd/3rd/4th qt) into date format.
# Example:3rd quarter->"-10-01"(oct 1st:First day of each Quarter)
quarterly_production$quarter.num <-
case_when(
  quarterly_production$quarter.num==1 ~ "-01-01",
  quarterly_production$quarter.num==2 ~ "-04-01",
  quarterly_production$quarter.num==3 ~ "-07-01",
  TRUE ~ "-10-01"
)
# Create a string format for Quarterly first date and  change it into date fo
rmat. Select necessary columns only.
quarterly_production2<-quarterly_production %>%  mutate(year.str=str_sub(Mont
h,1,4)) %>%        mutate(qt.str = paste0(year.str,quarter.num)) %>% mutate(Q
uarter.date = ymd(qt.str)) %>% select(Quarter.date, monthly_production)

## Adding missing grouping variables: `Quarter`

head(quarterly_production2,3)
```

```
## # A tibble: 3 x 3
## # Groups:   Quarter [3]
##   Quarter  Quarter.date monthly_production
##   <chr>    <date>                    <dbl>
## 1 1983 - 2 1983-04-01              8683.75
## 2 1983 - 3 1983-07-01              8653
## 3 1983 - 4 1983-10-01              8612
```

```
#Calculate average of each quarter
quarterly_mean<- quarterly_production2%>%group_by(Quarter.date) %>%    summar
ise(Quarterly_Production=mean(monthly_production))

date_production <-date_production%>%
  select(Date,Production)
names(quarterly_mean) <- c("Year","Production")
names(date_production) <- names(quarterly_mean)
Quarterly<-quarterly_mean
Weekly<-date_production

#Third plot for quartly and weekly
ggplot()+
  geom_line(aes(x=Year,y=Production,color="Quarterly"),
            data=Quarterly,size=1,
            alpha=1)+
  geom_line(aes(x=Year,y=Production,color="Weekly"),
            data=Weekly,size=0.3,
            alpha=0.6)+
  labs(x="Year",y="Thousand Barrels Per Day",color="Computation Method")+
  theme(legend.position= "bottom")+
  scale_y_continuous(label = scales::comma)+
  ggtitle("Production of Crude Oil In the U.S.(Week/Quarter)",subtitle="From
January in 1983 to February in 2020, Thousand Barrels Per Day")+
  theme_clean()+
  theme(plot.title = element_text(size = 10,hjust=0.5, color="navy"),
        plot.subtitle=element_text(size=8, hjust=0.5,
          face="italic", color="black"),
           legend.text=element_text(size=7),
        legend.title=element_text(size=7))+
            theme(legend.position= "bottom")
```




```
#Additional aesthetic feature: Letter face type(Italic), title location(cente
r), a special theme(clean) added
```