Gas Data

Yulia Tyukhova

February 3, 2023

Introduction

The data set contains expenses and other records for one car from April 2016 to January 2023. I visualize a number of patterns related to the expenses and gas consumption.

Loading and preparing the data

```
## Loading necessary libraries
library(dplyr)
library(ggplot2)
## For working with dates
library(lubridate)
## Table manipulations
library(kableExtra)

## Load data
df<-read.csv("2016_2023gas.csv",skip=8, header=TRUE,na.strings=",")
## Choose columns
data=df[,c(1,2,3,4,6,7)]
## Read Date column as date
Date<-as.Date(df$Date,"%m/%d/%Y")
## Assigning new data column to the data instead of the old format date
data$Date<-Date
## Rename columns
names(data)<-c("expense","cost","Date","location","gallon_price","gallons")
## Read gallon_price column as a number
data$gallon_price<-as.numeric(data$gallon_price)
##str(data)

## Add weekdays column from the Date
data<-mutate(data,weekday=weekdays(data$Date))
## Split Date into 3 columns - year, month, day
data2<-data.frame(date=data$Date,
                  year=as.numeric(format(data$Date,format="%Y")),
                  month=as.numeric(format(data$Date,format="%m")),
                  day=as.numeric(format(data$Date,format="%d")))
## Combine 2 data sets and get rid of the extra date column
data3<-cbind(data,data2)
data3<-mutate(data3,Date=NULL)
str(data3)
```

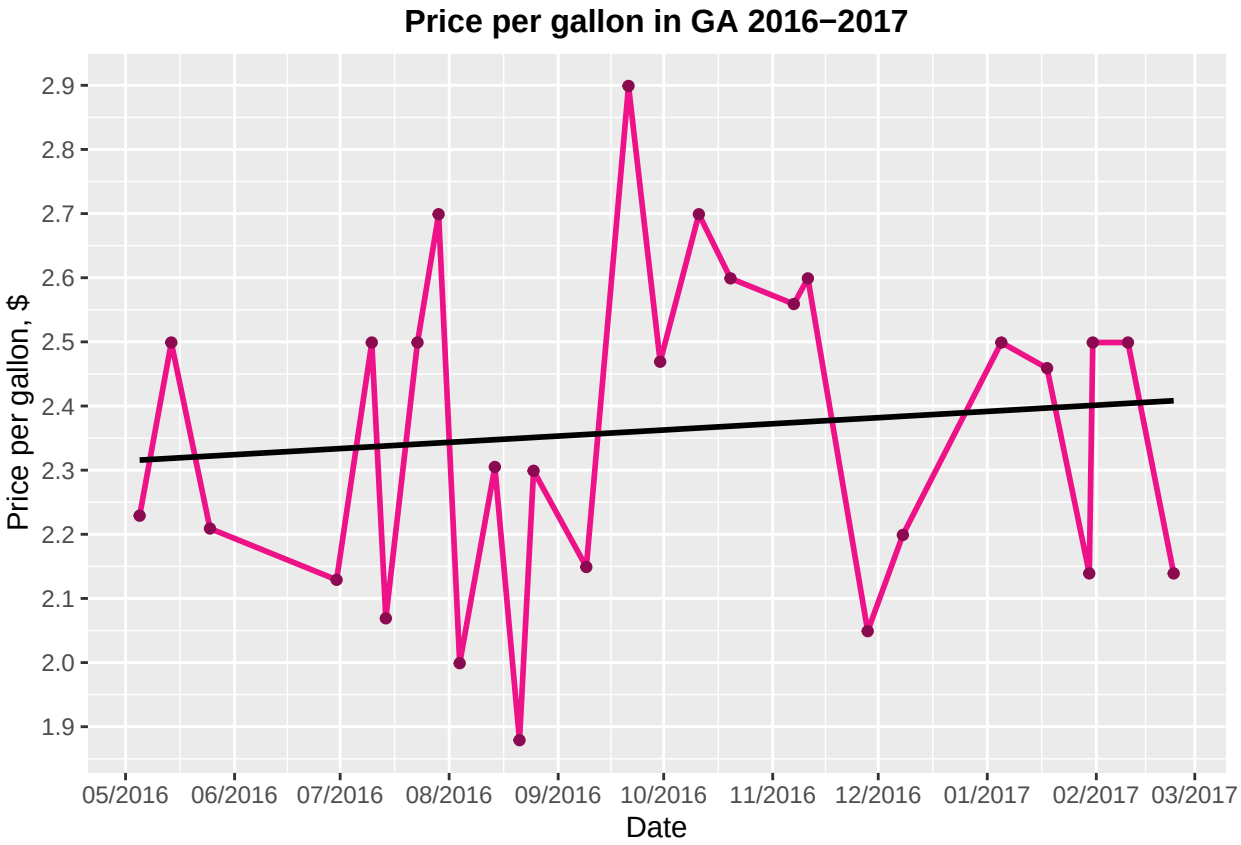
```
## 'data.frame':    319 obs. of  10 variables:
##  $ expense     : chr  "purchase" "Gas" "Gas" "Insurance" ...
##  $ cost        : num  NA 23.9 27.7 393.3 23.2 ...
##  $ location    : chr  "Marietta, GA" "Decatur, GA" "Atlanta, GA" "" ...
##  $ gallon_price: num  NA 2.23 2.5 NA 2.21 ...
##  $ gallons     : num  NA 10.7 11.1 NA 10.5 ...
##  $ weekday     : chr  "Tuesday" "Thursday" "Saturday" "Sunday" ...
##  $ date        : Date, format: "2016-04-26" "2016-05-05" ...
##  $ year        : num  2016 2016 2016 2016 2016 ...
##  $ month       : num  4 5 5 5 5 5 6 6 6 6 ...
##  $ day         : num  26 5 14 15 25 26 4 4 8 24 ...
```

Price per gallon in GA and CA over time

There are two places of residence in this data set - Georgia (GA) and California (CA). The following split in two data sets does not account for the trips made outside of these two states.

```
## Data for GA
dataGA<-data3[grep("GA", data3$location), ]
## Data for CA
dataCA<-data3[grep("CA", data3$location), ]

## Gas data for GA and CA
dataGA_gas<-dataGA[grep("gas", dataGA$expense,ignore.case=T), ]
dataCA_gas<-dataCA[grep("gas", dataCA$expense,ignore.case=T), ]
## Ommi NAs
dataGA_gas<-na.omit(dataGA_gas)
dataCA_gas<-na.omit(dataCA_gas)

## Price per gallon over time in each state
## Georgia
g1<-ggplot(data=dataGA_gas, aes(x=date,y=gallon_price))+
  geom_line(color="deeppink2",linewidth=1)+
  ggtitle(label="Price per gallon in GA 2016-2017")+
  labs(x="Date",y="Price per gallon, $")+
  theme(plot.title = element_text(size = 12,hjust=0.5,face="bold"))+
  geom_smooth(method=lm,se=FALSE, col='black', linewidth=1)+
  #add linear trend line without confidence interval
  geom_point(color="deeppink4")+
  ## geom_smooth() this adds a curved loess line
  scale_x_date(date_breaks = "month" , date_labels = "%m/%Y")+
  scale_y_continuous(breaks=seq(1.5, 3, by = 0.1))
g1
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

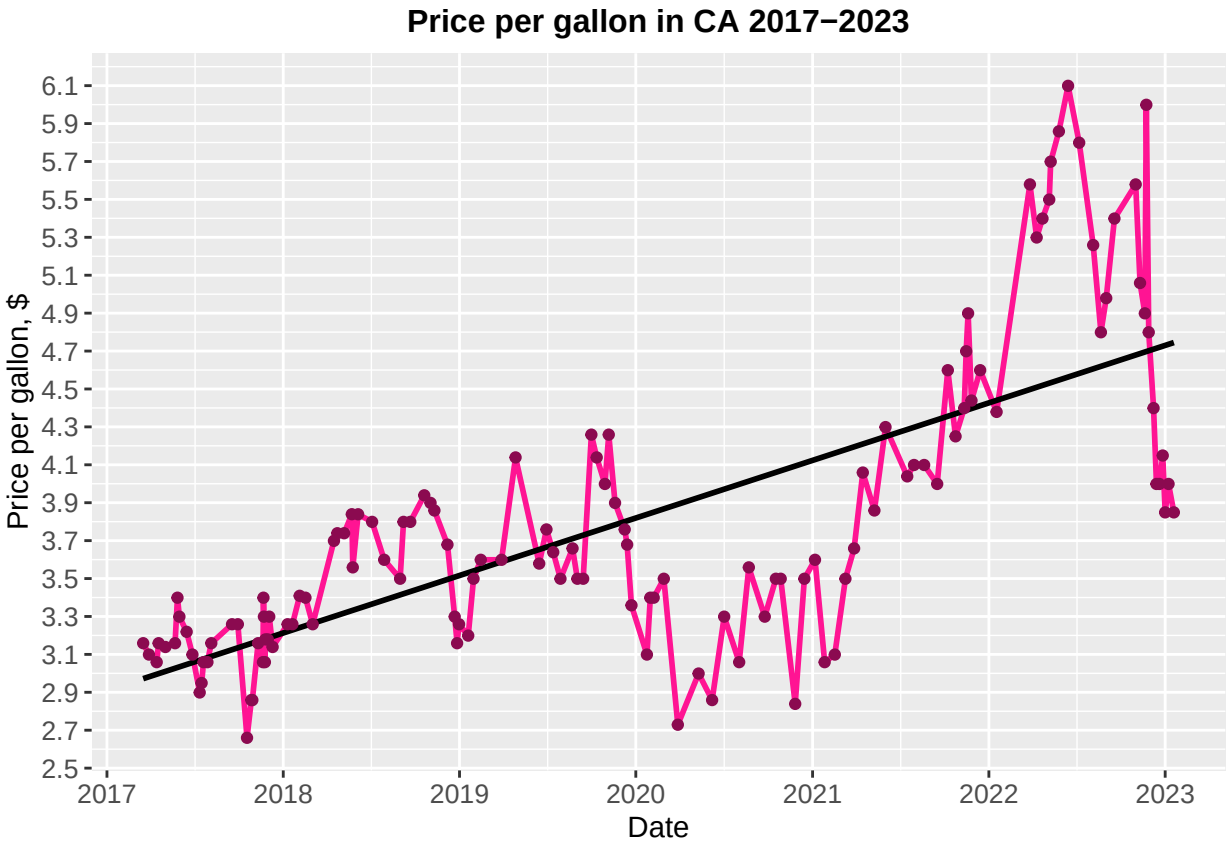



```
## California
g2<-ggplot(data=dataCA_gas, aes(x=date,y=gallon_price))+
  geom_line(color="deeppink",size=1)+
  ggtitle(label="Price per gallon in CA 2017-2023")+
  labs(x="Date",y="Price per gallon, $")+
  theme(plot.title = element_text(size = 12,hjust=0.5,face="bold"),
        axis.text=element_text(size=10))+
  geom_smooth(method=lm,se=FALSE, col='black', size=1)+
  geom_point(color="deeppink4")+
  scale_x_date(date_breaks = "years" , date_labels = "%Y")+
  scale_y_continuous(breaks=seq(2.5, 6.5, by = 0.2))
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
```

```
g2
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



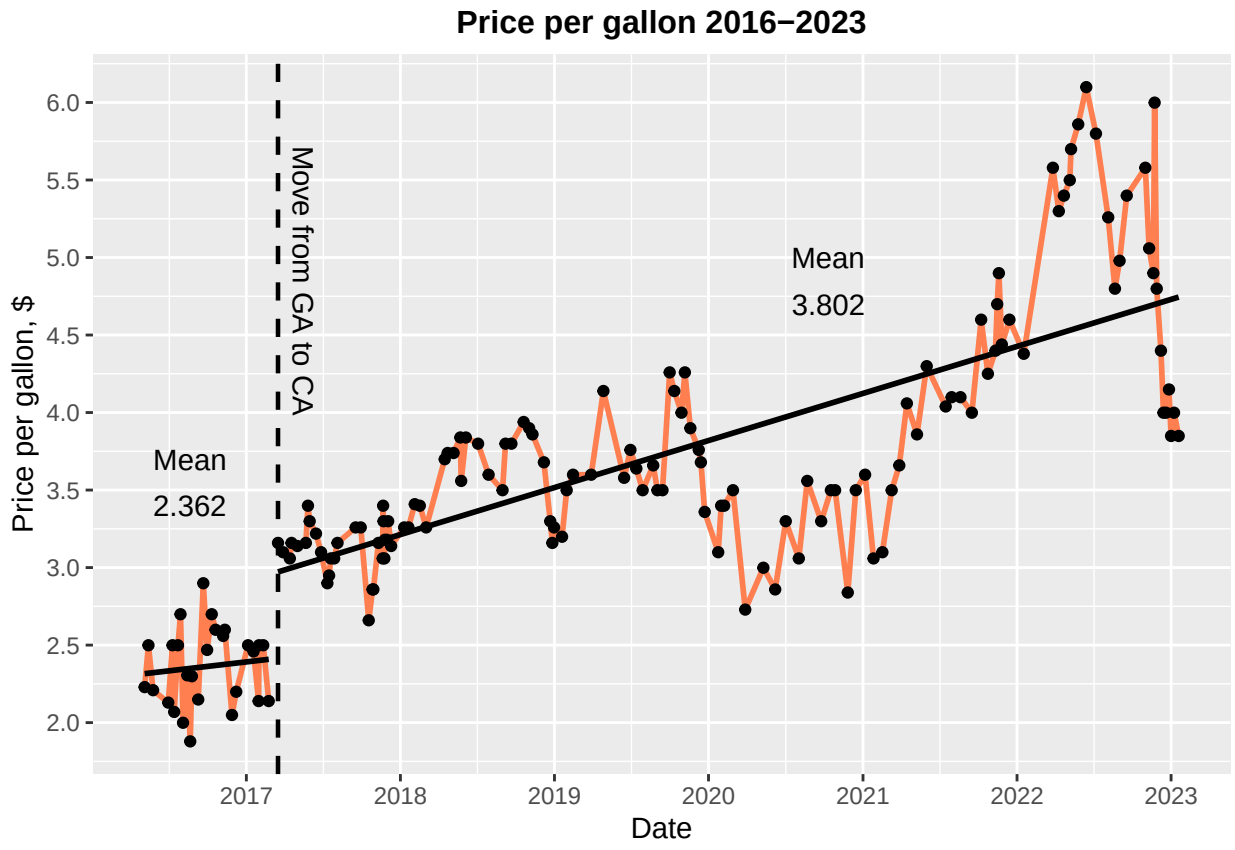

```
## Display price per gallon for both states - GA and CA - on one graph
## Show vertical line of the move from GA to CA on the following graph
data_gas<-data3[grep("gas", data3$expense,ignore.case=TRUE), ]
data_gas<-na.omit(data_gas)
## Gives only the date column from the data set with "CA" in the location
date_CA<-data_gas$date[grep("CA", data_gas$location)]
first_CA<-date_CA[1]
## This date will be used in the following graph to show the move from GA
## to CA with the vertical line geom_vline
first_CA
```

```
## [1] "2017-03-17"
```

```
## Combine two graphs and two trend lines for 2 states
dataGA_gas$group<-1
dataCA_gas$group<-2
two_states<-rbind(dataGA_gas,dataCA_gas)

g3<-ggplot(data=two_states, aes(x=date,y=gallon_price,group=group))+
  geom_line(col='coral',size=1)+
  ggtitle(label="Price per gallon 2016-2023")+
  labs(x="Date",y="Price per gallon, $")+
  theme(plot.title = element_text(size = 12,hjust=0.5,face="bold"))+
  geom_smooth(method=lm,se=FALSE, col='black', size=1)+
  geom_point(col='black')+
```

```
  scale_x_date(date_breaks = "years" , date_labels = "%Y")+
  geom_vline(xintercept=first_CA, linetype='dashed', color='black', size=0.8)+
  annotate("text",x=first_CA,y=4.85, label="Move from GA to CA", vjust = -0.7,
           angle='-90')+ ## had to use date format for x
  annotate("text",x=first_CA,y=c(3.7,3.4),
           label=c("Mean",round(mean(dataGA_gas$gallon_price),3)),hjust = 1.7)+
  annotate("text",x=first_CA,y=c(5,4.7),
           label=c("Mean",round(mean(dataCA_gas$gallon_price),3)),hjust = -7)+
  scale_y_continuous(breaks=seq(1.5, 6.5, by = 0.5))

g3
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Average price per gallon per year & separately per month/year in CA

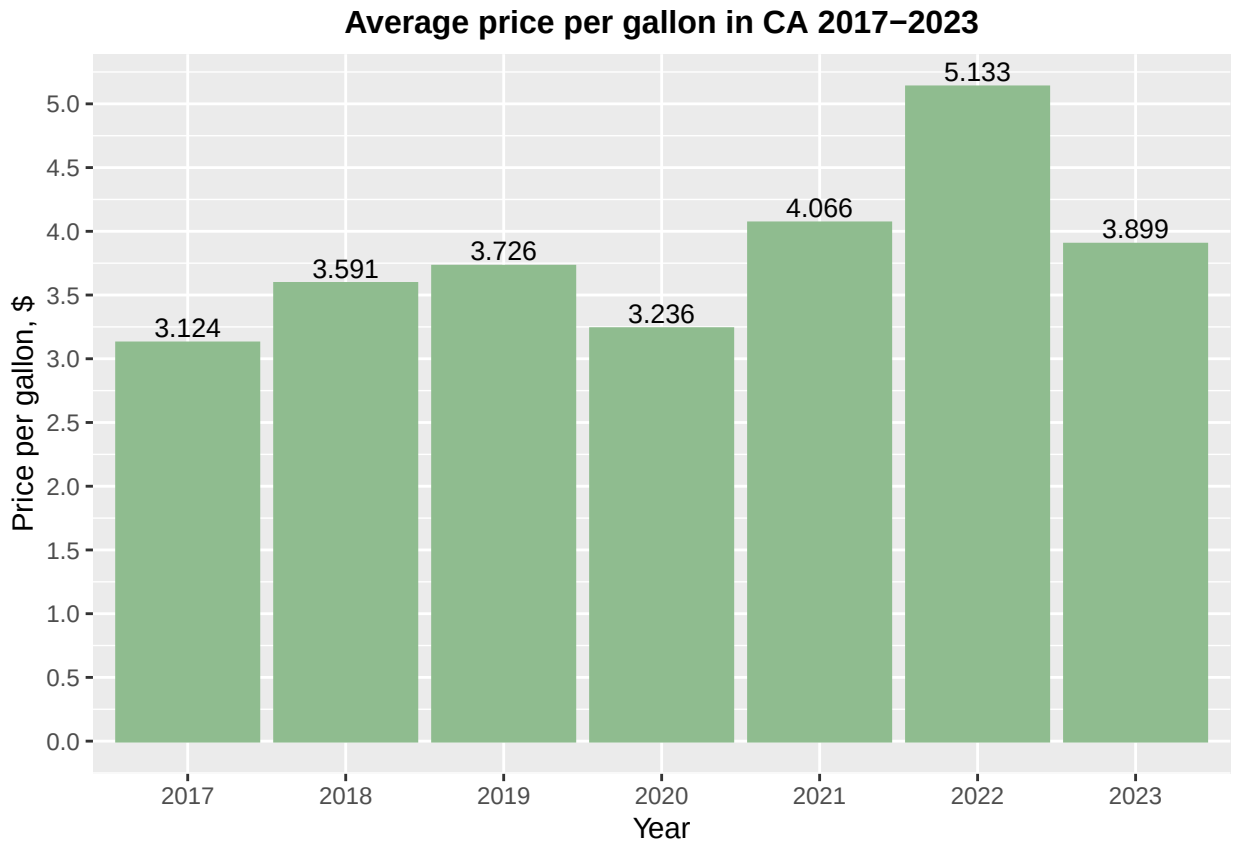
Below I explore how the average price per gallon changed in CA over the period of 2017-2023 based on data from one car.

```
## Group CA data by year to find average price per gallon for each year
year_group<-group_by(dataCA_gas,year)
by_year<-summarize(year_group,round(mean(gallon_price),3))
data_year<-as.data.frame(by_year)
```

```
names(data_year)<-c("year","average_gallon_price")
data_year
```

```
##   year average_gallon_price
## 1 2017                3.124
## 2 2018                3.591
## 3 2019                3.726
## 4 2020                3.236
## 5 2021                4.066
## 6 2022                5.133
## 7 2023                3.899
```

```
g4<-ggplot(data=data_year, aes(x=factor(year),y=average_gallon_price))+
  geom_bar(stat="identity",color="darkseagreen",fill="darkseagreen")+
  ggtitle(label="Average price per gallon in CA 2017-2023")+
  labs(x="Year",y="Price per gallon, $")+
  theme(plot.title = element_text(size = 12,hjust=0.5,face="bold"))+
  geom_text(aes(label=average_gallon_price), vjust=-0.3, size=3.5)+
  scale_y_continuous(breaks=seq(0, 6, by = 0.5))
g4
```




```
## Average gas price per gallon per month in CA across 2017-2023
## Group by year and month to find average price per gallon
group_CAym<-group_by(dataCA_gas,year,month)
by_yearmonth<-summarize(group_CAym,round(mean(gallon_price),3))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

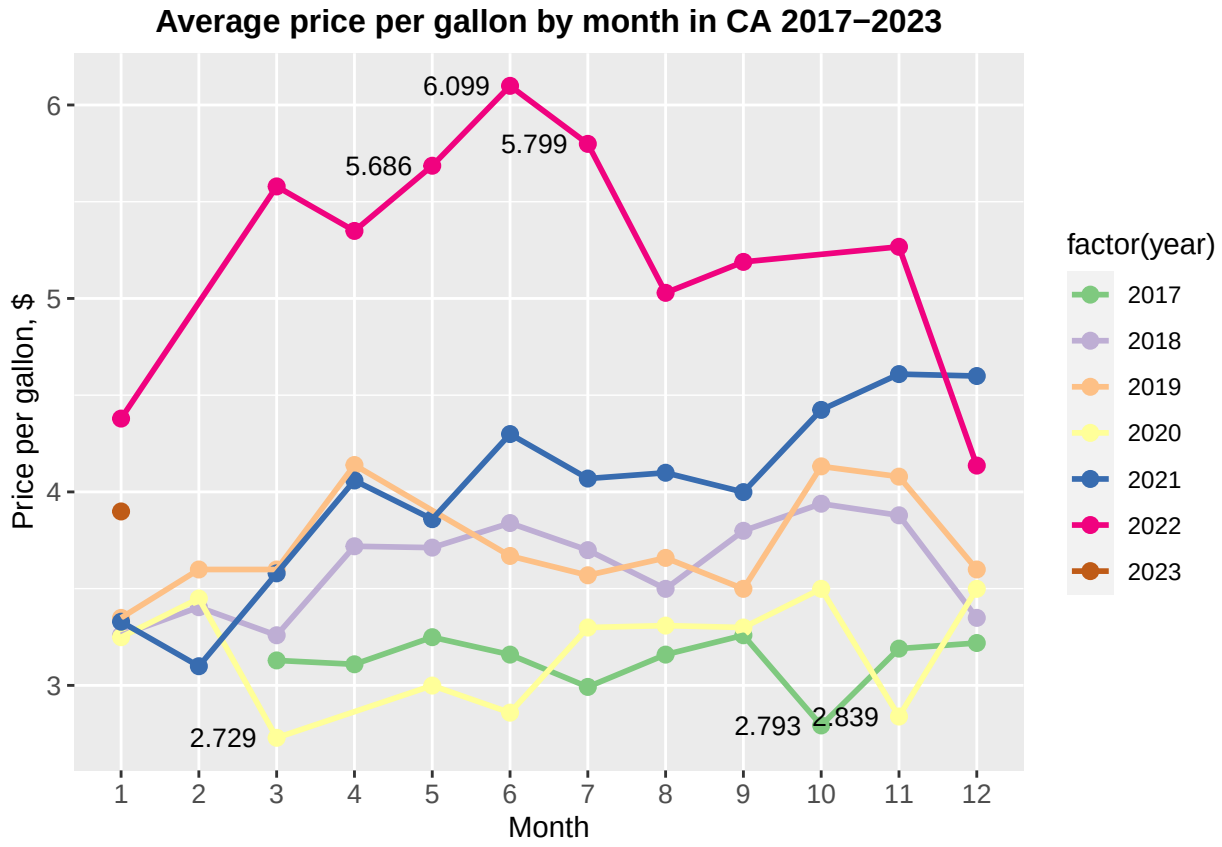
```
data_ym<-as.data.frame(by_yearmonth)
names(data_ym)<-c("year","month","gallon_price")
## data_ym

##Three most expensive gas prices for labeling on the graph below
high3price<-slice_max(data_ym,gallon_price,n=3)
high3price
```

```
##   year month gallon_price
## 1 2022     6        6.099
## 2 2022     7        5.799
## 3 2022     5        5.686
```

```
##Three least expensive gas prices
low3price<-slice_min(data_ym,gallon_price,n=3)
low3price
```

```
##   year month gallon_price
## 1 2020     3        2.729
## 2 2017    10        2.793
## 3 2020    11        2.839
```

```
g5<-ggplot(data=data_ym, aes(x=factor(month),y=gallon_price,color=factor(year),group=year))+
  geom_point(size=2.5)+
  geom_line(size=1)+
  ggtitle(label="Average price per gallon by month in CA 2017-2023")+
  labs(x="Month",y="Price per gallon, $")+
  theme(plot.title = element_text(size = 12,hjust=0.5,face="bold"),
        axis.text=element_text(size=10))+
    scale_color_brewer(palette = "Accent")+
  geom_text(data=high3price,aes(x=month,y=gallon_price),label=high3price$gallon_price,
            hjust=1.3, size=3.5,color="black")+
  geom_text(data=low3price,aes(x=month,y=gallon_price),label=low3price$gallon_price,
            hjust=1.3, size=3.5,color="black")
g5
```




```
## The biggest change of price per gallon within a year in CA
## Group gas data in CA by year and get max and min
group_CAy<-group_by(dataCA_gas,year)
by_yearmax<-summarize(group_CAy,round(max(gallon_price),3))
data_maxCA<-as.data.frame(by_yearmax)
names(data_maxCA)<-c("year","max_gallon_price")

by_yearmin<-summarize(group_CAy,round(min(gallon_price),3))
data_minCA<-as.data.frame(by_yearmin)
names(data_minCA)<-c("year","min_gallon_price")

## Biggest change in cost over the course of the year
rangeCA<-cbind(data_maxCA,data_minCA)
## Delete only one "year" column - third
rangeCA<-select(rangeCA,-(3))
rangeCA<-mutate(rangeCA,range=(max_gallon_price-min_gallon_price))
## rangeCA
kable_classic(kbl(rangeCA,caption = "Biggest change of price per gallon within each year in CA"),
              font_size = 10,full_width = F, html_font = "Cambria",latex_options = "HOLD_position")
```

Table 1: Biggest change of price per gallon within each year in CA

year	max_gallon_price	min_gallon_price	range
2017	3.399	2.660	0.739
2018	3.939	3.159	0.780
2019	4.259	3.199	1.060
2020	3.559	2.729	0.830
2021	4.899	3.059	1.840
2022	6.099	3.999	2.100
2023	3.999	3.849	0.150

Gas consumption

Below I explore gas consumption in gallons by month/year over the period of 2016-2023.

```
## Most gas consumption by month/year
## Group by year and month and sum up total consumption in gallons
group_ym<-group_by(data3,year,month)
by_yearmonthGAL<-summarize(group_ym,round(sum((gallons),na.rm=TRUE),1))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

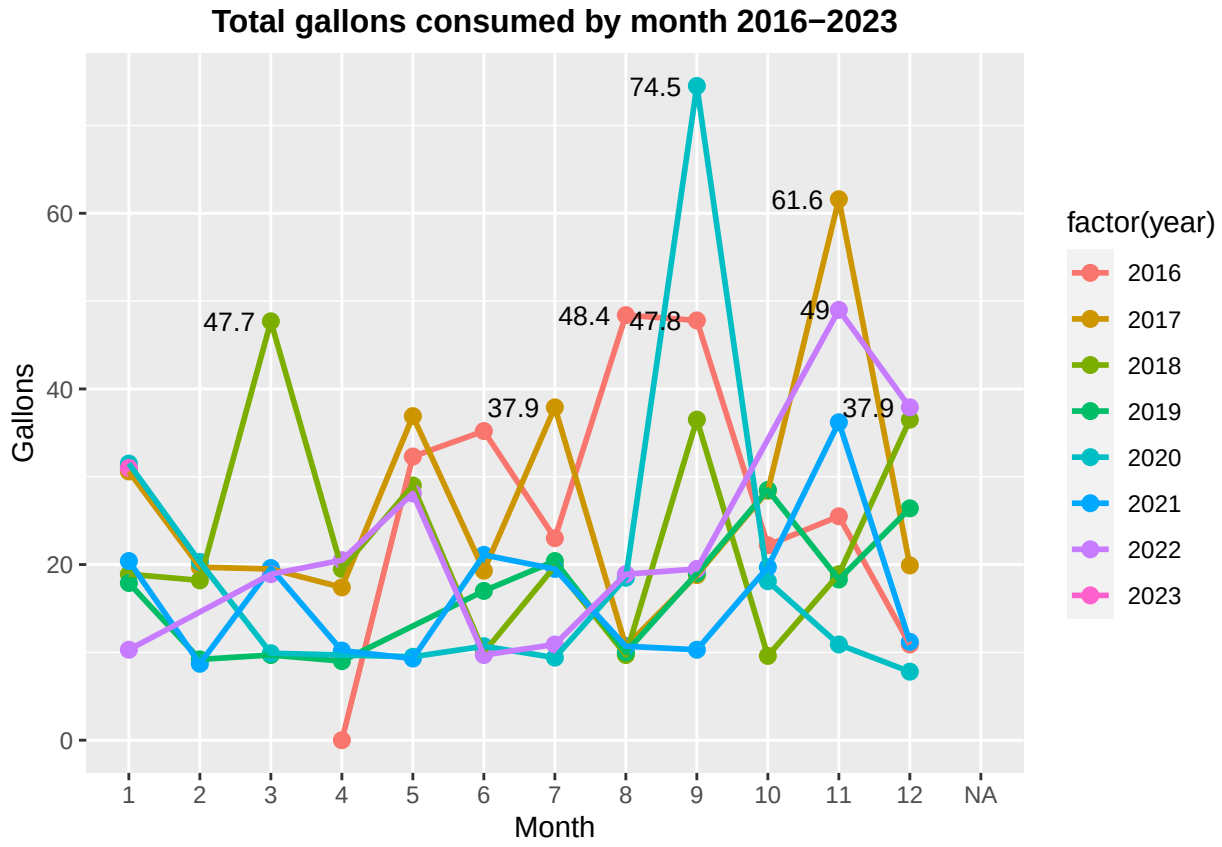
```
data_ymGAL<-as.data.frame(by_yearmonthGAL)
names(data_ymGAL)<-c("year","month","gallons")

## Seven highest consumption months/years
most_consumption<-slice_max(data_ymGAL,gallons,n=7)

g6<-ggplot(data=data_ymGAL, aes(x=factor(month),y=gallons,color=factor(year),group=year))+
  geom_point(size=2.5)+
  geom_line(size=1)+
  ggtitle(label="Total gallons consumed by month 2016-2023")+
  labs(x="Month",y="Gallons")+
  theme(plot.title = element_text(size = 12,hjust=0.5,face="bold"))+
  scale_color_brewer(palette = "Set3")+
  geom_text(data=most_consumption,aes(x=month,y=gallons),label=most_consumption$gallons,
            hjust=1.3, size=3.5,color="black")+
  scale_colour_discrete(na.translate = F) ## removes NA as a factor
```

```
## Scale for colour is already present.
## Adding another scale for colour, which will replace the existing scale.
```

```
g6
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 1 row containing missing values (`geom_line()`).
```




```
most_consumption$index<-1:nrow(most_consumption)
most_consumption<-most_consumption[,c(4,1,2,3)]
most_consumption
```

```
##   index year month gallons
## 1     1 2020     9    74.5
## 2     2 2017    11    61.6
## 3     3 2022    11    49.0
## 4     4 2016     8    48.4
## 5     5 2016     9    47.8
## 6     6 2018     3    47.7
## 7     7 2017     7    37.9
## 8     8 2022    12    37.9
```

Now I explore total cost of gas by month/year over the period of 2016-2023.

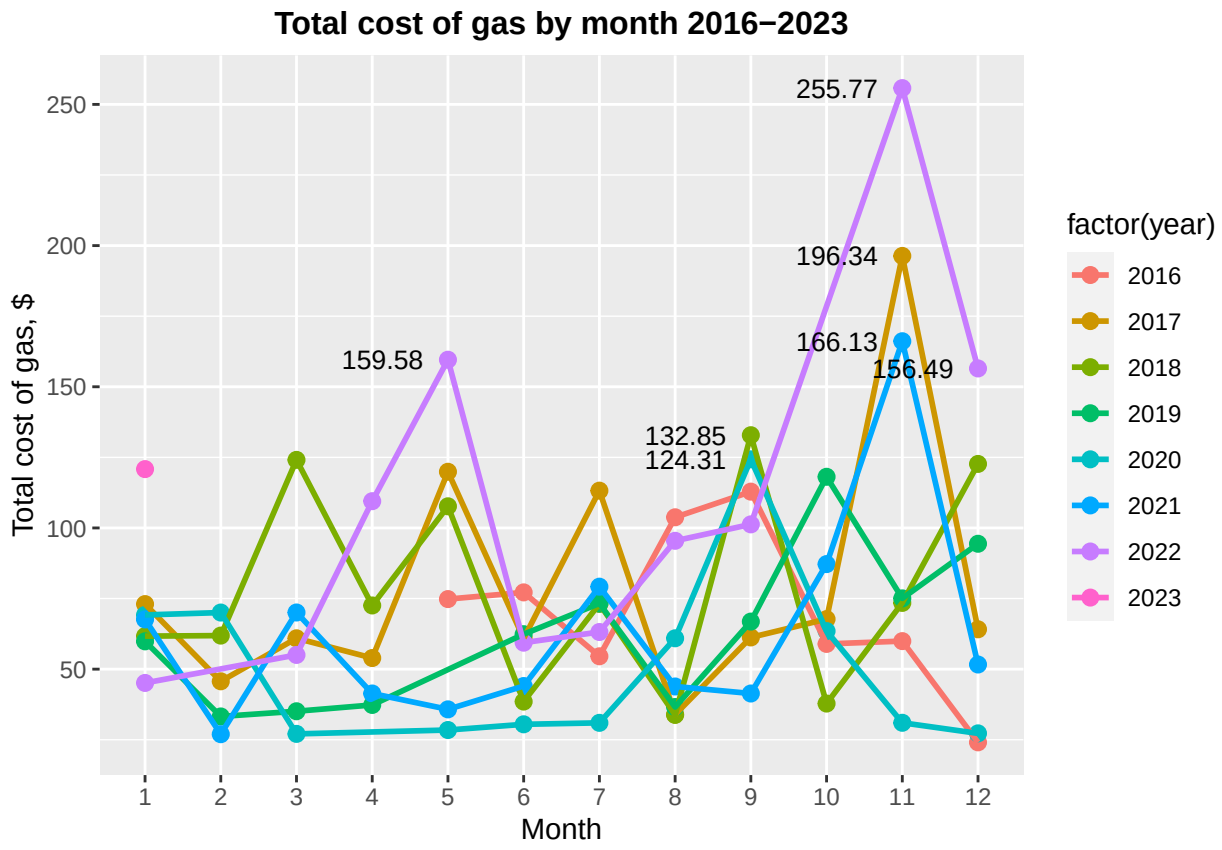
```
## Cost of gas by month/year
group_cost<-group_by(data_gas,year,month)
by_yearmonthcost<-summarize(group_cost,round(sum((cost),na.rm=TRUE),3))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
data_cost<-as.data.frame(by_yearmonthcost)
names(data_cost)<-c("year","month","cost")

gas_expenditure<-slice_max(data_cost,cost,n=7)

g7<-ggplot(data=data_cost, aes(x=factor(month),y=cost,color=factor(year),group=year))+
  geom_point(size=2.5)+
  geom_line(size=1)+
  ggtitle(label="Total cost of gas by month 2016-2023")+
  labs(x="Month",y="Total cost of gas, $")+
  theme(plot.title = element_text(size = 12,hjust=0.5,face="bold"))+
  scale_color_brewer(palette = "Set3")+
  geom_text(data=gas_expenditure,aes(x=month,y=cost),label=gas_expenditure$cost,
            hjust=1.3, size=3.5,color="black")+
  scale_colour_discrete(na.translate = F) ## removes NA as a factor
```

```
## Scale for colour is already present.
## Adding another scale for colour, which will replace the existing scale.
```

```
g7
```




```
gas_expenditure$index<-1:nrow(gas_expenditure)
gas_expenditure<-gas_expenditure[,c(4,1,2,3)]
gas_expenditure
```

```
##   index year month   cost
## 1     1 2022    11 255.77
## 2     2 2017    11 196.34
## 3     3 2021    11 166.13
## 4     4 2022     5 159.58
## 5     5 2022    12 156.49
## 6     6 2018     9 132.85
## 7     7 2020     9 124.31
```

```
kable_classic(kbl(gas_expenditure,
                  caption = "Top 7 years/months with the highest total cost of gas"),
              font_size = 10,full_width = F, html_font = "Cambria",latex_options = "HOLD_position")
```

Table 2: Top 7 years/months with the highest total cost of gas

index	year	month	cost
1	2022	11	255.77
2	2017	11	196.34
3	2021	11	166.13
4	2022	5	159.58
5	2022	12	156.49
6	2018	9	132.85
7	2020	9	124.31

```
kable_classic(kbl(most_consumption,
                  caption = "Top 7 years/months with the highest gas consumption"),
              font_size = 10,full_width = F, html_font = "Cambria",latex_options = "HOLD_position")
```

Table 3: Top 7 years/months with the highest gas consumption

index	year	month	gallons
1	2020	9	74.5
2	2017	11	61.6
3	2022	11	49.0
4	2016	8	48.4
5	2016	9	47.8
6	2018	3	47.7
7	2017	7	37.9
8	2022	12	37.9

Consumption and expenditure for gas summary

- The most gallons were consumed in **September 2020 with 74.5 gallons**, then **November 2017 with 61.6 gallons**, and **November 2022 with 49 gallons**.
- The highest expenditure for gas occurred in **November 2022 with $255.77**, then **November 2017 with $196.34**, and finally **November 2021 with $166.13**.

- Despite of November 2022 being the third in gas consumption, the total cost of gas was the highest during this month due to the most expensive price per gallon in CA in 2022 in the period of 2017-2023.

- In terms of having the highest consumption of gas in September 2020, but only the seventh spot in total cost of gas is explained by the average price per gallon being almost as low as in 2017.