HP-Exploratory analysis

preethi

11/23/2019

Loaded the required packages

```r
library(googledrive)
```

```
## Warning: package 'googledrive' was built under R version 3.6.1
```

```r
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.6.1
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.1
```

```r
library(scales)
```

```
## Warning: package 'scales' was built under R version 3.6.1
```

```r
library(reshape2)
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 3.6.1
```

Download file and read it

```r
temp <- tempfile(fileext = ".zip")
dl <- drive_download(
  as_id("https://drive.google.com/open?id=1dyKTsCDegBCDDBDVJRJAf_JIUxd8YcET"), path = temp, overwrite =
out <- unzip(temp, exdir = tempdir())
expenditure <- read.csv(out, header = TRUE, stringsAsFactors = FALSE)
unlink(temp)

budget <- read.csv("hoa_wise_prep_data.csv", header = TRUE, stringsAsFactors = FALSE)
```

I cannot seem to download and extract the budget file from the drive through R so I have downloaded it and saved it locally

##temp <- tempfile(fileext = “.zip”) ##dl <- drive_download(as_id(“https://drive.google.com/open?id=1LQNEV3vQDI3nkofVwADHrTZdPIh29oAg”), path = temp, overwrite = TRUE) ##out1 <- unzip(temp, exdir = tempdir()) ##budget <- read.csv(out1, header = TRUE, stringsAsFactors = FALSE)

Convert the format of the date from factor to date format

```r
## change date format

expenditure$TRANSDATE <- as.Date(expenditure$TRANSDATE, format = "%Y-%m-%d")
budget$date <- as.Date(budget$date, format = "%Y-%m-%d")
```

Add a month_year column to sort by month_year

I am going to assume that the revised and sanctioned data is in 100000's and do the following.

```
budget$Revisedlakh <- budget$REVISED *100000
budget$SanctionedLakh <- budget$SANCTION * 100000
```

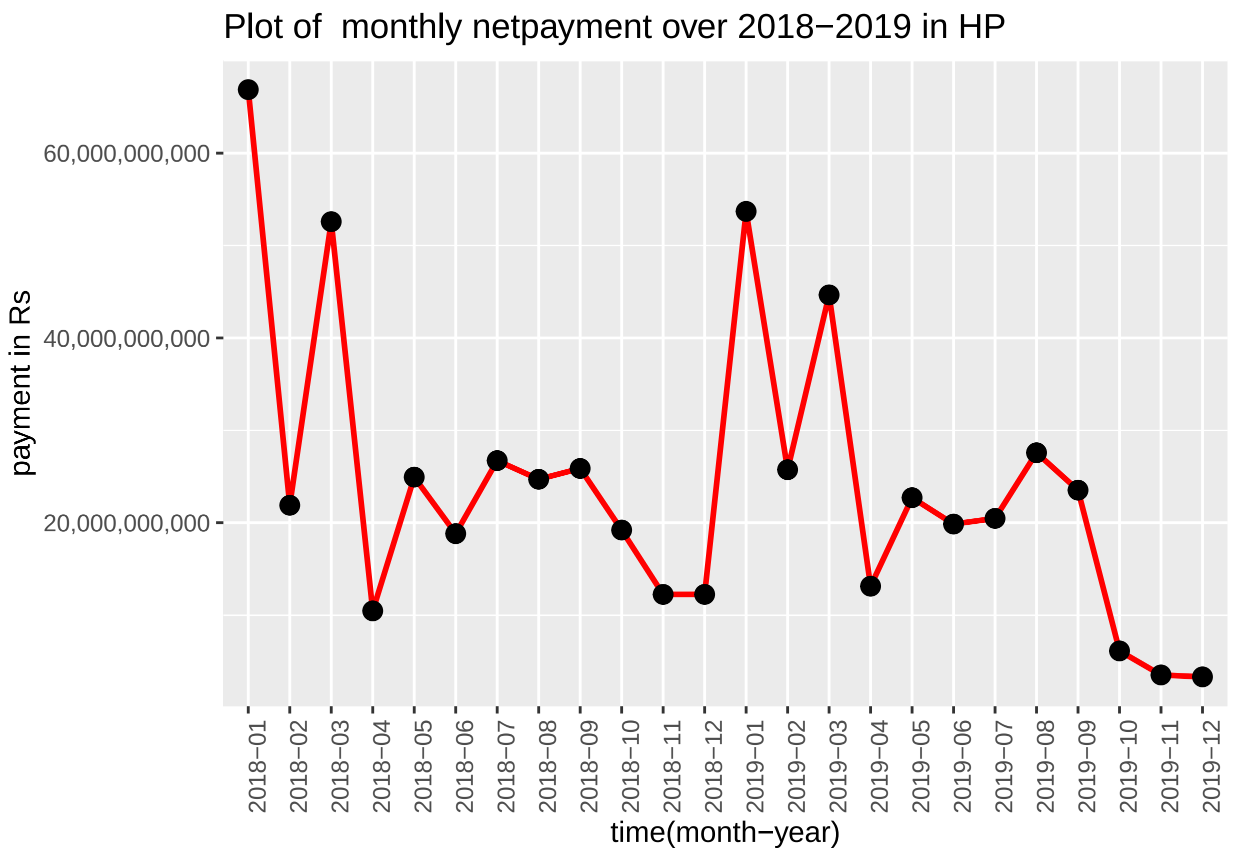
##budget and expenditure data is then grouped by month_year

```
## expenditure data grouped by month and netpayment

expenditure <- expenditure %>% mutate(month_year = format(TRANSDATE, "%Y-%m"))

Sumpayment <- expenditure %>% group_by(month_year) %>% summarise(total = sum(NETPAYMENT, na.rm = TRUE))

BudgetSum <- budget %>% mutate(month_year = format(date, "%Y-%m")) %>%
group_by(month_year) %>% summarise_at(c("Revisedlakh","SanctionedLakh"), sum, na.rm = TRUE)

## budget data grouped by medical budget
medical_budget <- budget %>% mutate(month_year = format(date, "%Y-%m")) %>%
group_by(month_year) %>% filter(major == 2210)

medical_budget_month <- budget %>% mutate(month_year = format(date, "%Y-%m")) %>%
group_by(month_year) %>% filter(major == 2210) %>% summarise_at(c("Revisedlakh","SanctionedLakh"), sum,

medical_exp_distric <- expenditure %>% mutate(month_year = format(TRANSDATE, "%Y-%m")) %>%
group_by(month_year) %>% filter(major == 2210)

medical_expenditure <- expenditure %>% mutate(month_year = format(TRANSDATE, "%Y-%m")) %>% group_by(mon
filter(major == 2210) %>% summarise(total = sum(NETPAYMENT, na.rm = TRUE))
```

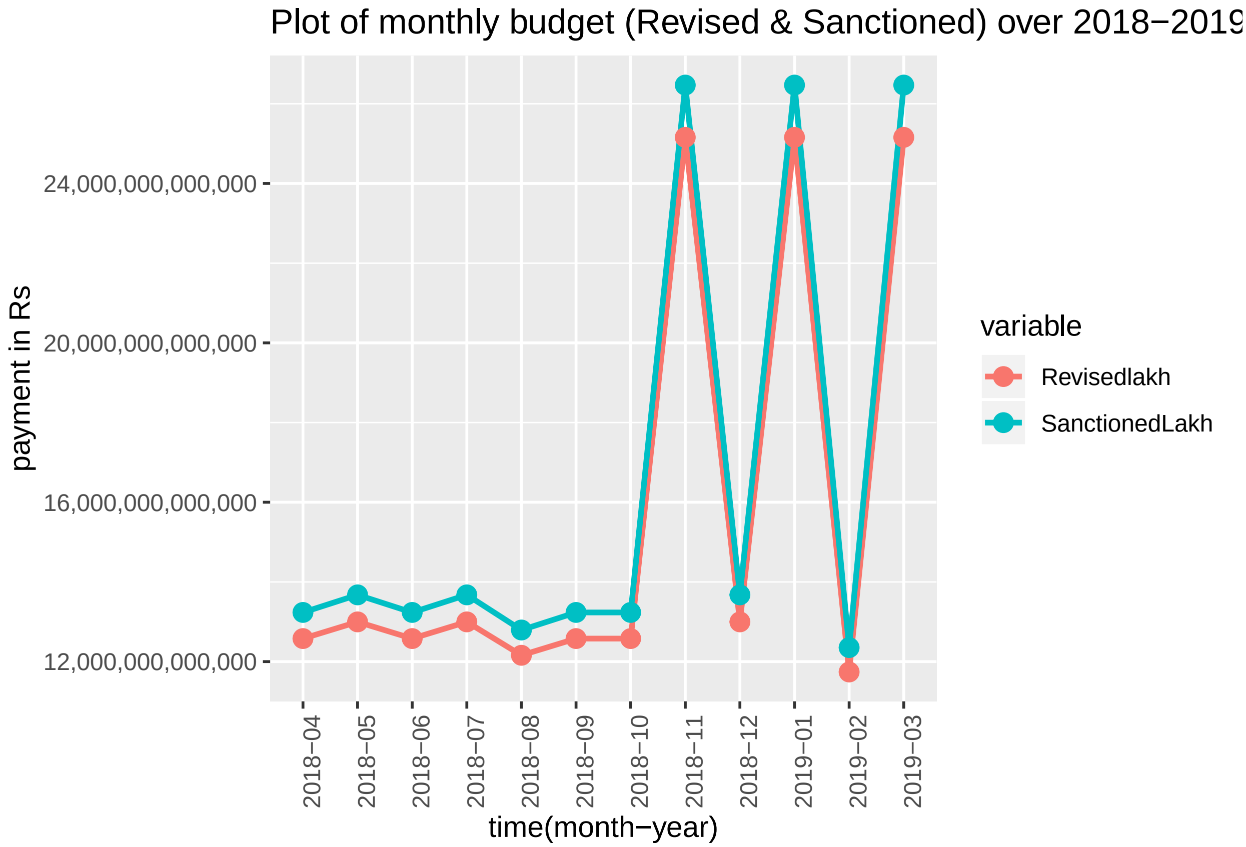
Melt the data so we can visualize the revised and sanctioned estimates as variables

```
BudgetSum.long <- melt(BudgetSum, id = "month_year", measure = c("Revisedlakh","SanctionedLakh"))

medical_budget.long <- melt(medical_budget_month, id = "month_year", measure = c("Revisedlakh","Sanctio
```

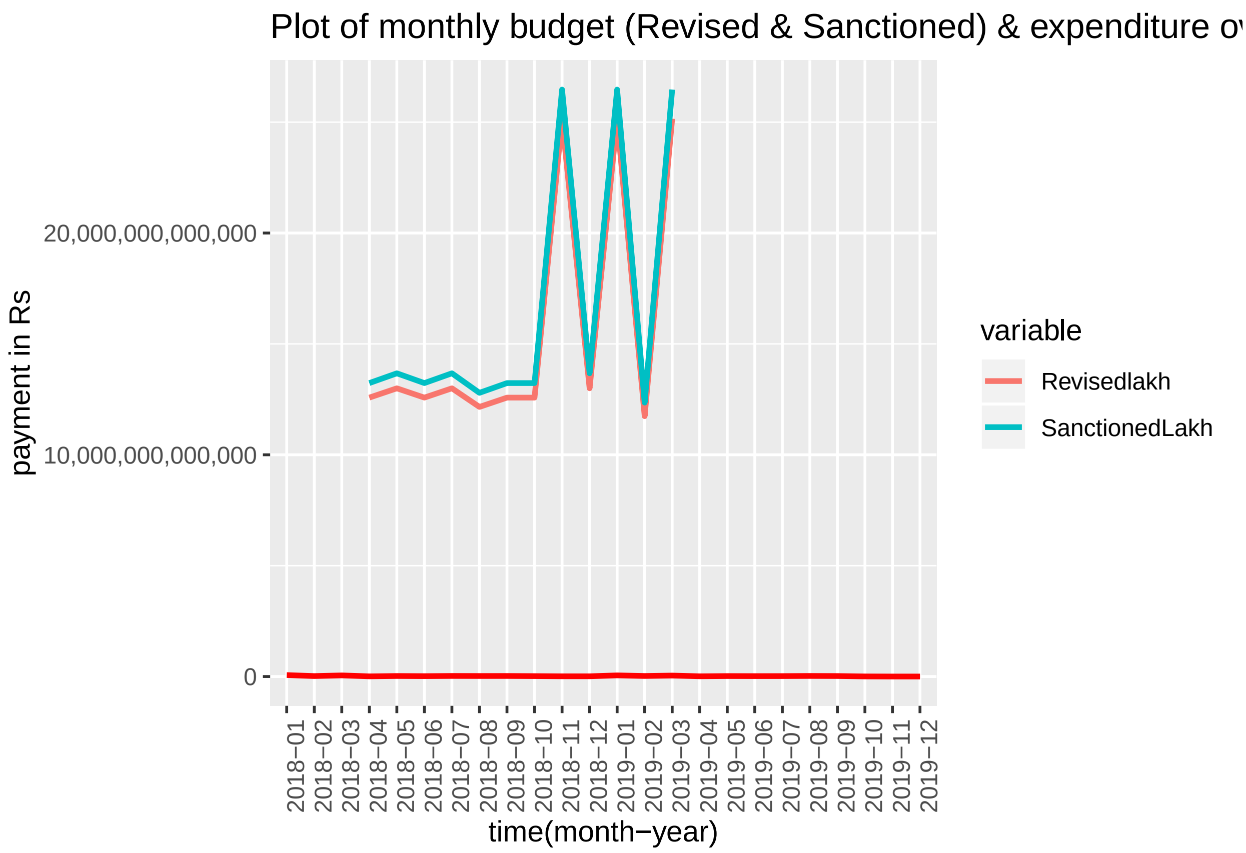
```
## plot of total expenditure monthly for state

plot1 <- ggplot (data = Sumpayment, aes(x = month_year, y=total, group = 1)) + geom_line(color = "red",

print (plot1)
```



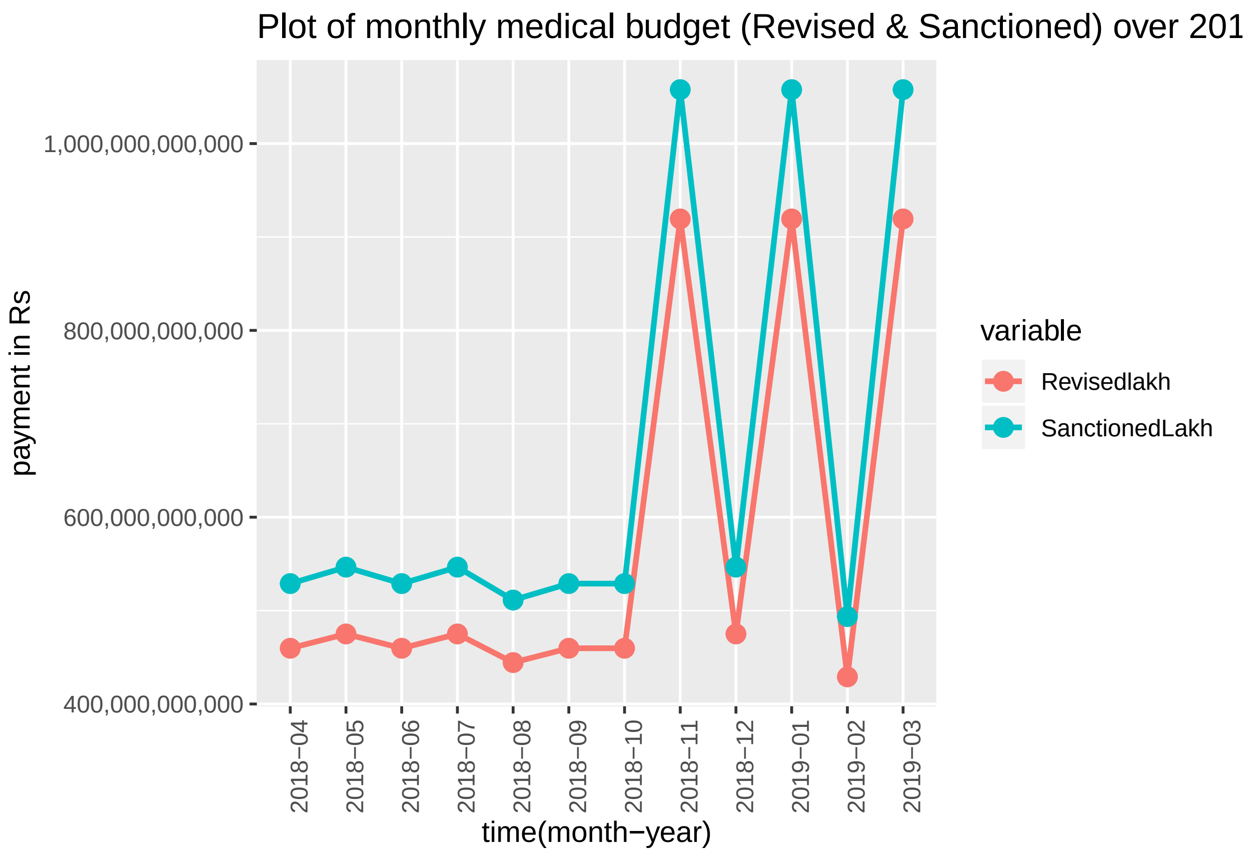
```r
##plot of monthly budget data

plot2a <- ggplot (data = BudgetSum.long,aes(x = month_year, y=value, color = variable, group = variable)

print (plot2a)
```



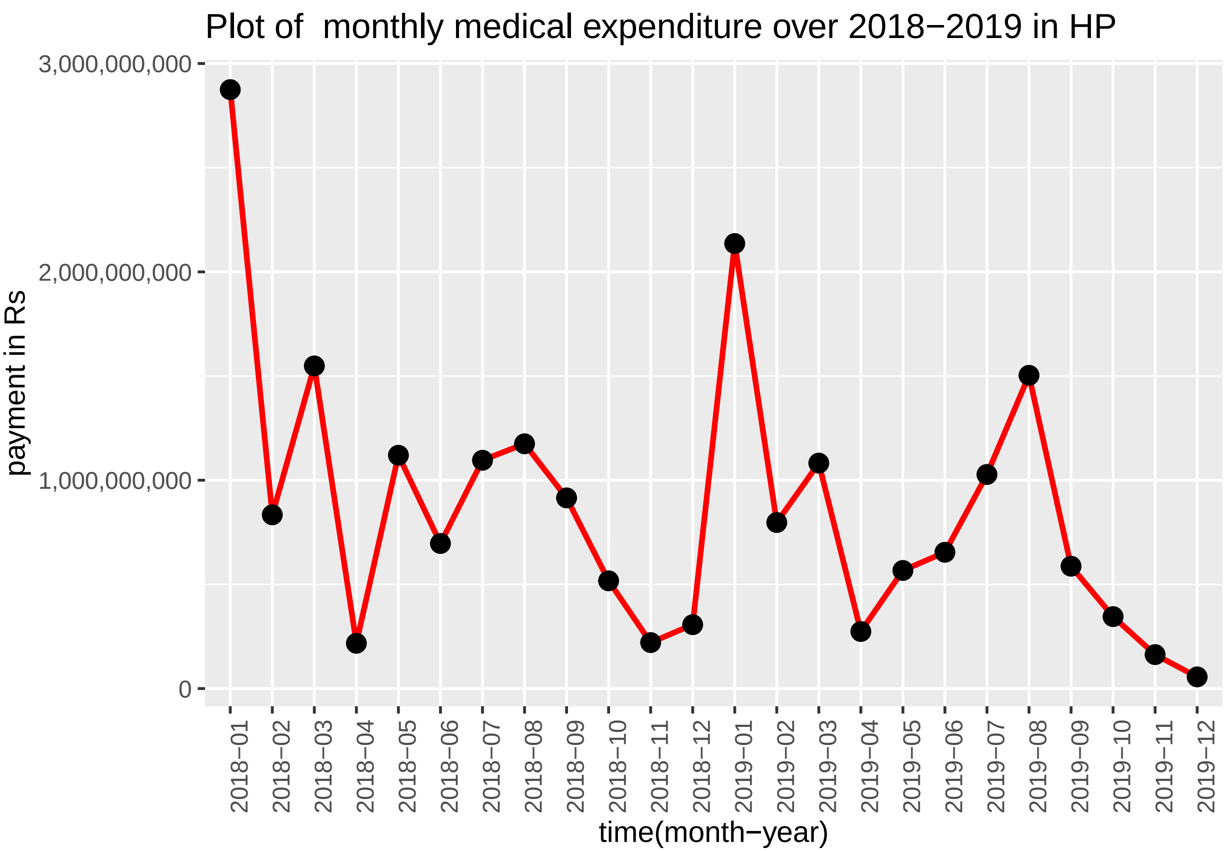
```r
##plot of budget and expenditure data

plot3 <- ggplot() + geom_line(data = Sumpayment, aes(x = month_year, y=total, group = 1), color = "red"
  geom_line(data = BudgetSum.long, aes(x= month_year,y=value,colour = variable, group = variable),size
geom_point(size=3)+
  scale_y_continuous(labels = scales::comma)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggtitle("Plot of monthly budget (Revised & Sanctioned) & expenditure over 2018-2019 in HP") + xlab("t

print(plot3)
```



```
##plot of monthly medical budget estimates (revised and sanctioned)
plot4a <- ggplot(data = medical_budget.long, aes(x= month_year, y=value, color=variable, group = variabl

print(plot4a)
```

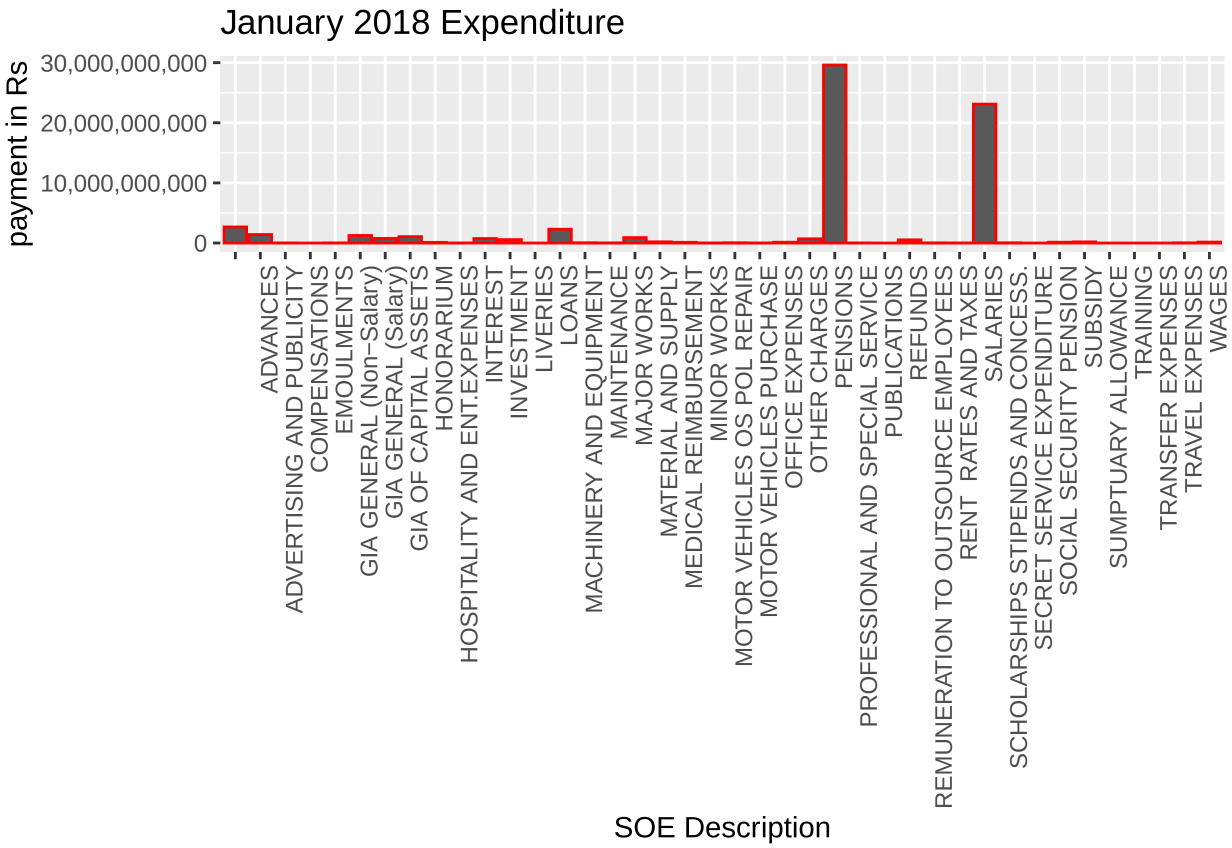


```
##plot of monthly medical expenditures
plot5 <- ggplot (data = medical_expenditure, aes(x = month_year, y=total, group = 1)) + geom_line(color

print (plot5)
```

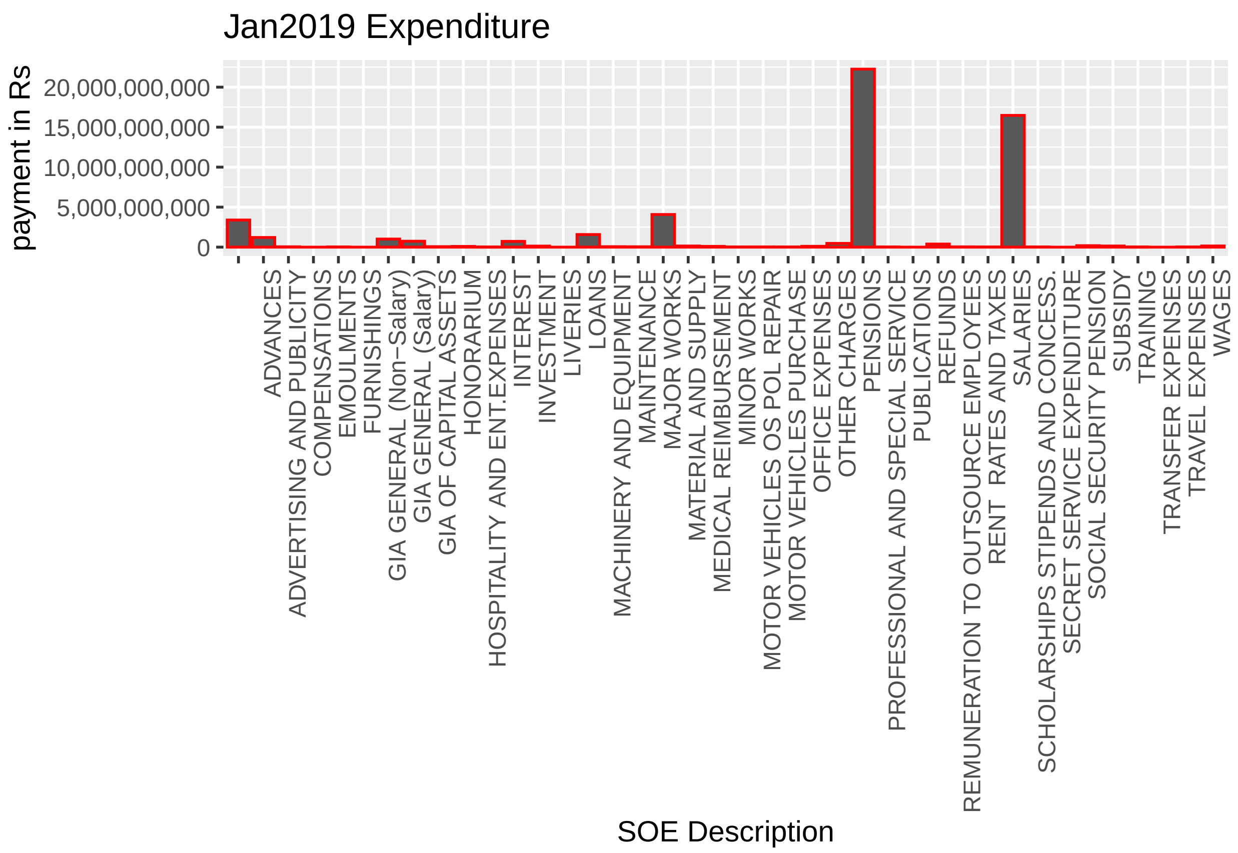


For plot 1: January and march seem to have higher values than the other months, we can subset that data to see why

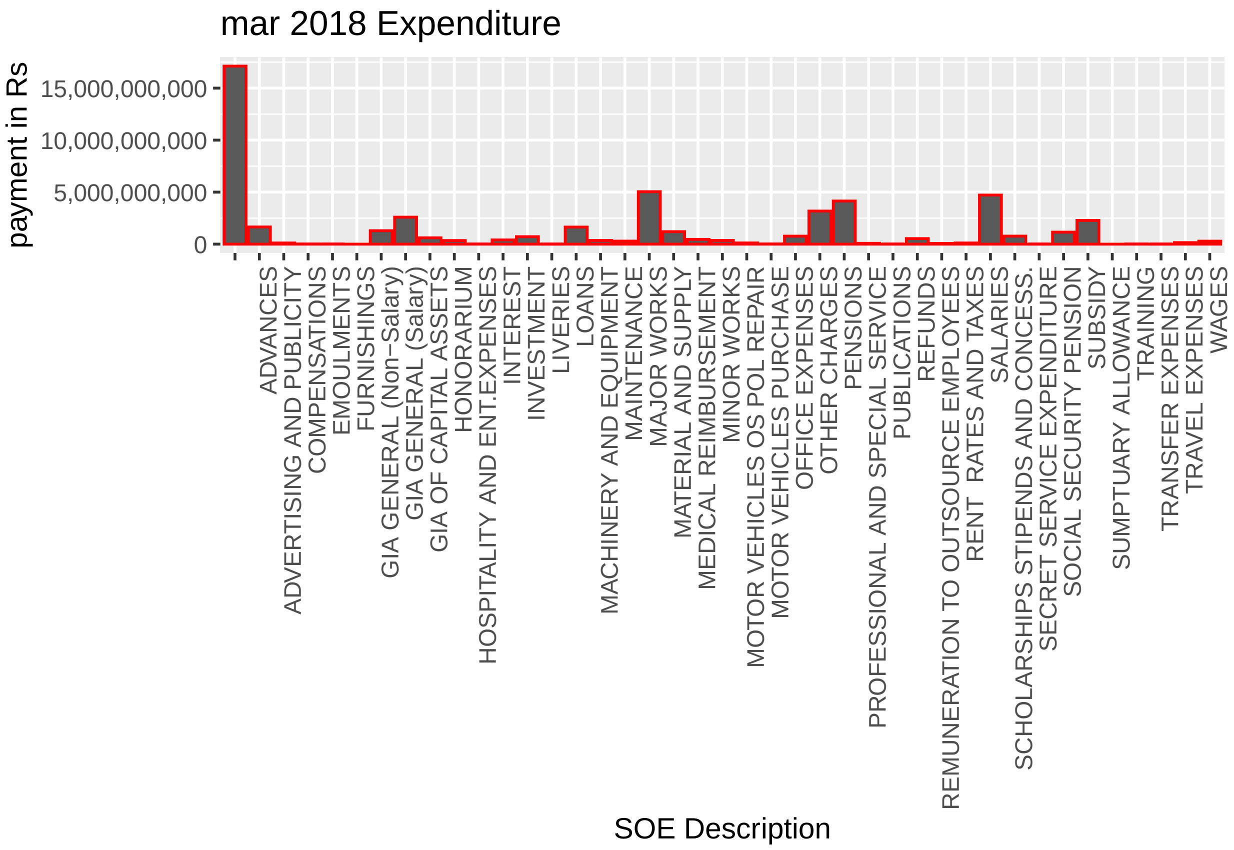
```
expenditure_jan_2018 <- filter(expenditure,expenditure$TRANSDATE >= "2018-01-01" & expenditure$TRANSDAT
expenditure_jan_2019 <- filter(expenditure,expenditure$TRANSDATE >= "2019-01-01" & expenditure$TRANSDAT
expenditure_mar_2018<- filter(expenditure,expenditure$TRANSDATE >= "2018-03-01" & expenditure$TRANSDATE
expenditure_mar_2019 <- filter(expenditure,expenditure$TRANSDATE >= "2019-03-01" & expenditure$TRANSDAT

expenditure_jan_2018 <-  expenditure_jan_2018 %>% group_by(SOE_description) %>% summarise(total = sum(N
expenditure_jan_2019 <-  expenditure_jan_2019 %>% group_by(SOE_description) %>% summarise(total = sum(N

expenditure_mar_2018 <-  expenditure_mar_2018 %>% group_by(SOE_description) %>% summarise(total = sum(N
expenditure_mar_2019 <-  expenditure_mar_2019 %>% group_by(SOE_description) %>% summarise(total = sum(N
##expenditure_mar <- expenditure[expenditure$TRANSDATE >= "2018-03-01" & expenditure$TRANSDATE >= "2018

hist1 <- ggplot(data = expenditure_jan_2018, aes(x = SOE_description,y=total))+
  geom_bar(stat = "identity", color = "red")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_y_continuous(labels = scales::comma)+
  ggtitle("January 2018 Expenditure") +
  xlab("SOE Description") + ylab("payment in Rs")



hist2 <- ggplot(data = expenditure_jan_2019, aes(x = SOE_description,y=total))+
  geom_bar(stat = "identity", color = "red")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_y_continuous(labels = scales::comma)+
  ggtitle("Jan2019 Expenditure") +
  xlab("SOE Description") + ylab("payment in Rs")

print (hist1)
```



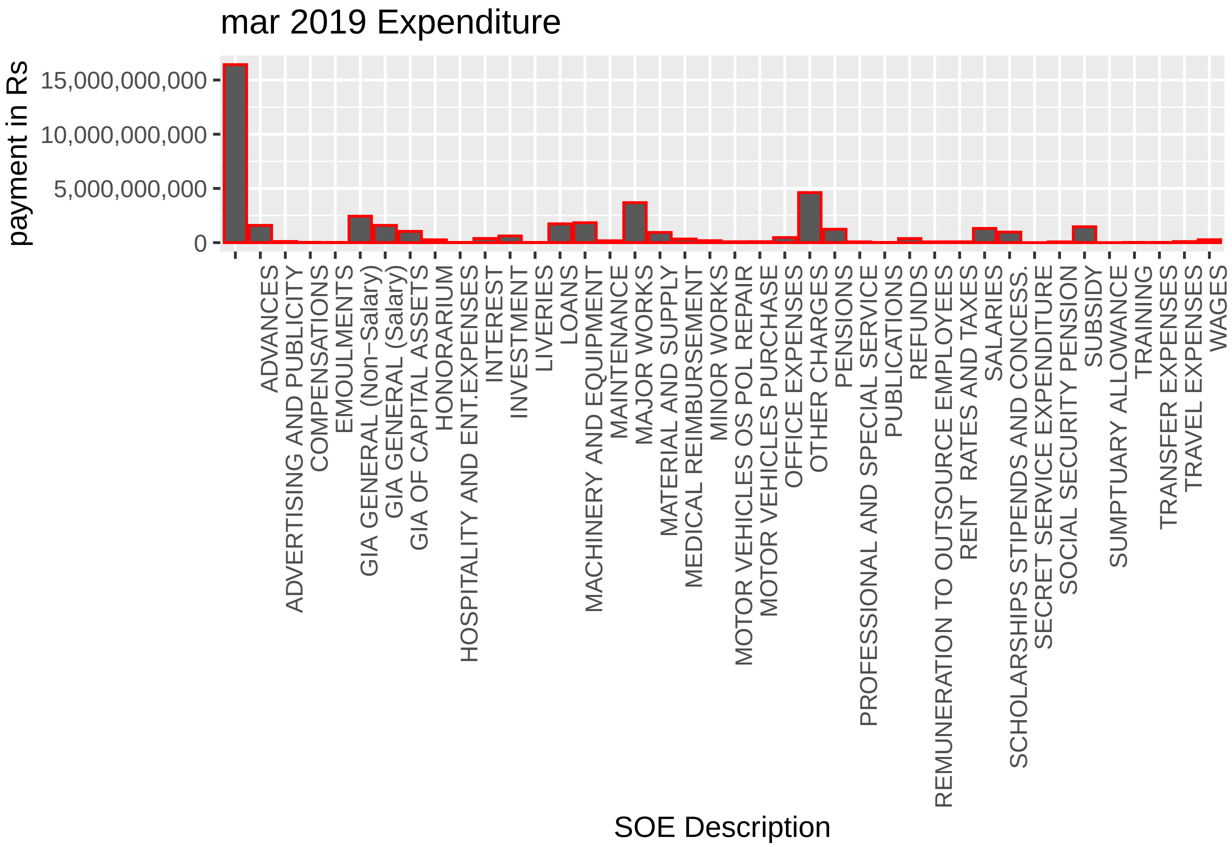
```
print(hist2)
```



```
hist3 <- ggplot(data = expenditure_mar_2018, aes(x = SOE_description,y=total))+
  geom_bar(stat = "identity", color = "red")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_y_continuous(labels = scales::comma)+
  ggtitle("mar 2018 Expenditure") +
  xlab("SOE Description") + ylab("payment in Rs")



hist4 <- ggplot(data = expenditure_mar_2019, aes(x = SOE_description,y=total))+
  geom_bar(stat = "identity", color = "red")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_y_continuous(labels = scales::comma)+
  ggtitle("mar 2019 Expenditure") +
  xlab("SOE Description") + ylab("payment in Rs")

print (hist3)
```



```r
print(hist4)
```



It looks it is mainly accounted by salaries and pensions for the month of jannuary and by misc? for March

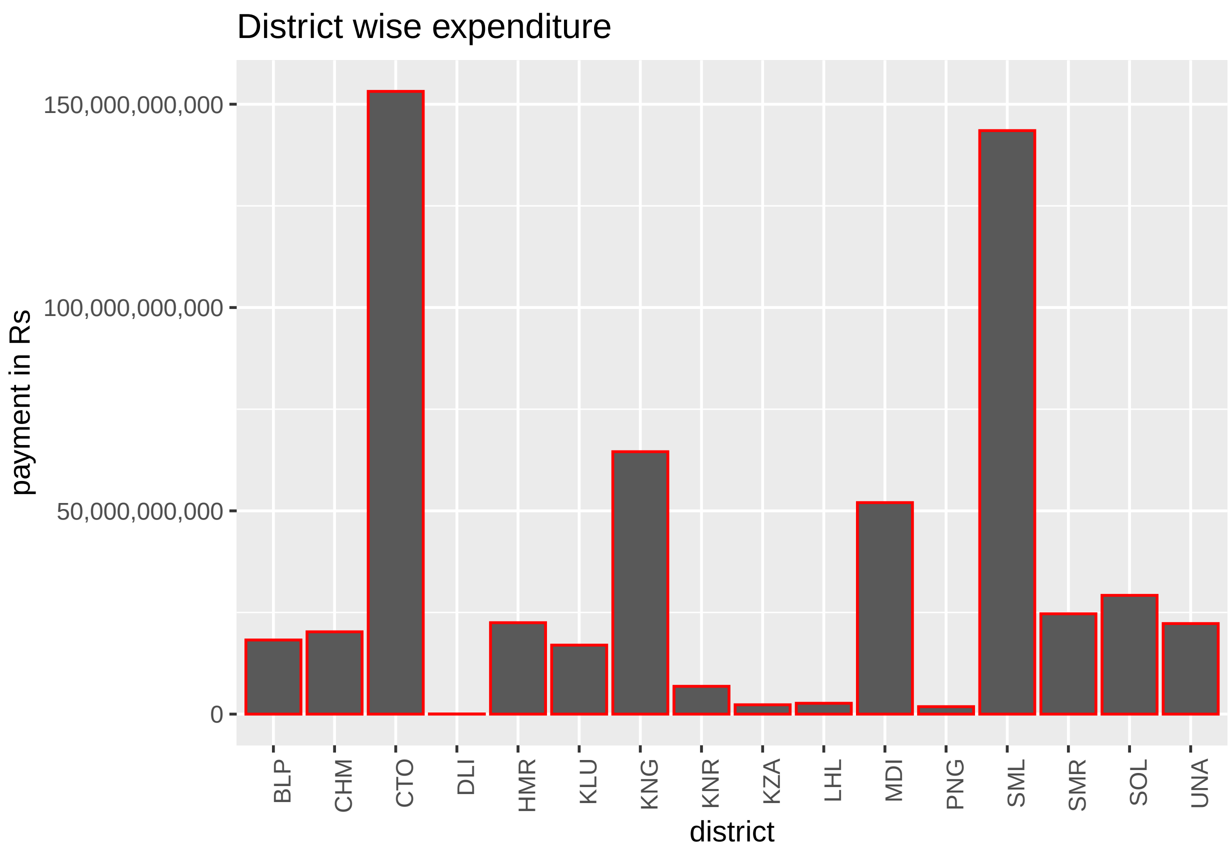
Now we can look at districtwise spending

Districtwise Spending

There are supposed to be 12 districts in Himachal. There is a lot more data here.

Maybe by subsetting the treasury code letters of the alphabet we may come across some semblance of district wise spend

```
expenditure$District_code <- substr(expenditure$Treasury_Code,1,3)

expenditure_district <-  expenditure %>% group_by(District_code) %>% summarise(total = sum(NETPAYMENT, 

hist_dist <- ggplot(data = expenditure_district , aes(x = District_code,y=total))+
  geom_bar(stat = "identity", color = "red")+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  scale_y_continuous(labels = scales::comma)+
  ggtitle("District wise expenditure") +
  xlab("district") + ylab("payment in Rs")

print(hist_dist)
```



data that is not working yet

Districtspend <- expenditure %>% group_by(District)%>% summarise(Total = sum(NETPAYMENT, na.rm = TRUE))

District_spending_month <- expenditure %>% group_by(month,District,SOE_description) %>% summarise(Total = sum(NETPAYMENT, na.rm = TRUE))

##total expenditure monthly per district

district_plot <- ggplot(data = District_spending_month, aes(x=month,y=Total, group =1))+ geom_line(color = “darkorchid4”) + facet_wrap(~ District, ncol = 7) + labs(title = “Total Expenditure by district”, subtitle = “Data plotted by month”, y = “total expenditure”, x = “month”) + theme_bw(base_size = 15) + scale_y_continuous(labels = scales::comma)

print(district_plot)

##pie chart with district wise spending

barplot <- ggplot(subset(District_spending_month, District == “AMB”), aes(x="“,y = Total, fill = SOE_description))+geom_bar(stat=”identity", width = 1)

pie <- barplot + coord_polar(“y”, start=0)

pie

##not clean