MAT 4860: Project

Ricky Trujillo

2024-04-29

Data Pre-Processing

```r
weather_2022 = read.table("CA_LA_Weather_22.txt")
weather_2022 = weather_2022[,-c(6,9,11,12,13,14,16,17,18)]
colnames(weather_2022)<- c("Year", "Month", "Day", "Hour", "Temperature",
                           "Pressure","Sky_Cover", "Wind_Speed", "Rainfall")

weather_2022$Temperature<-gsub("L","",as.character(weather_2022$Temperature))
weather_2022$Temperature<-gsub("R","",as.character(weather_2022$Temperature))
weather_2022$Temperature<-gsub("F","",as.character(weather_2022$Temperature))

weather_2022$Sky_Cover<-gsub("L","",as.character(weather_2022$Sky_Cover))
weather_2022$Sky_Cover<-gsub("R","",as.character(weather_2022$Sky_Cover))
weather_2022$Sky_Cover<-gsub("F","",as.character(weather_2022$Sky_Cover))

weather_2022$Rainfall<-gsub("L","",as.character(weather_2022$Rainfall))
weather_2022$Rainfall<-gsub("R","",as.character(weather_2022$Rainfall))
weather_2022$Rainfall<-gsub("F","",as.character(weather_2022$Rainfall))
```

```r
weather_2023 = read.table("CA_LA_Weather_23.txt")
weather_2023 = weather_2023[,-c(6,9,11,12,13,14,16,17,18)]
colnames(weather_2023)<- c("Year", "Month", "Day", "Hour", "Temperature",
                           "Pressure","Sky_Cover", "Wind_Speed", "Rainfall")

weather_2023$Temperature<-gsub("L","",as.character(weather_2023$Temperature))
weather_2023$Temperature<-gsub("R","",as.character(weather_2023$Temperature))
weather_2023$Temperature<-gsub("F","",as.character(weather_2023$Temperature))

weather_2023$Sky_Cover<-gsub("L","",as.character(weather_2023$Sky_Cover))
weather_2023$Sky_Cover<-gsub("R","",as.character(weather_2023$Sky_Cover))
weather_2023$Sky_Cover<-gsub("F","",as.character(weather_2023$Sky_Cover))

weather_2023$Rainfall<-gsub("L","",as.character(weather_2023$Rainfall))
weather_2023$Rainfall<-gsub("R","",as.character(weather_2023$Rainfall))
weather_2023$Rainfall<-gsub("F","",as.character(weather_2023$Rainfall))
```

```r
weather_2024 = read.table("CA_LA_Weather_24.txt")
weather_2024 = weather_2024[,-c(6,9,11,12,13,14,16,17,18)]
colnames(weather_2024)<- c("Year", "Month", "Day", "Hour", "Temperature",
                           "Pressure","Sky_Cover", "Wind_Speed", "Rainfall")

weather_2024$Temperature<-gsub("L","",as.character(weather_2024$Temperature))
```

```
weather_2024$Temperature<-gsub("R","",as.character(weather_2024$Temperature))
weather_2024$Temperature<-gsub("F","",as.character(weather_2024$Temperature))

weather_2024$Sky_Cover<-gsub("L","",as.character(weather_2024$Sky_Cover))
weather_2024$Sky_Cover<-gsub("R","",as.character(weather_2024$Sky_Cover))
weather_2024$Sky_Cover<-gsub("F","",as.character(weather_2024$Sky_Cover))

weather_2024$Rainfall<-gsub("L","",as.character(weather_2024$Rainfall))
weather_2024$Rainfall<-gsub("R","",as.character(weather_2024$Rainfall))
weather_2024$Rainfall<-gsub("F","",as.character(weather_2024$Rainfall))
```

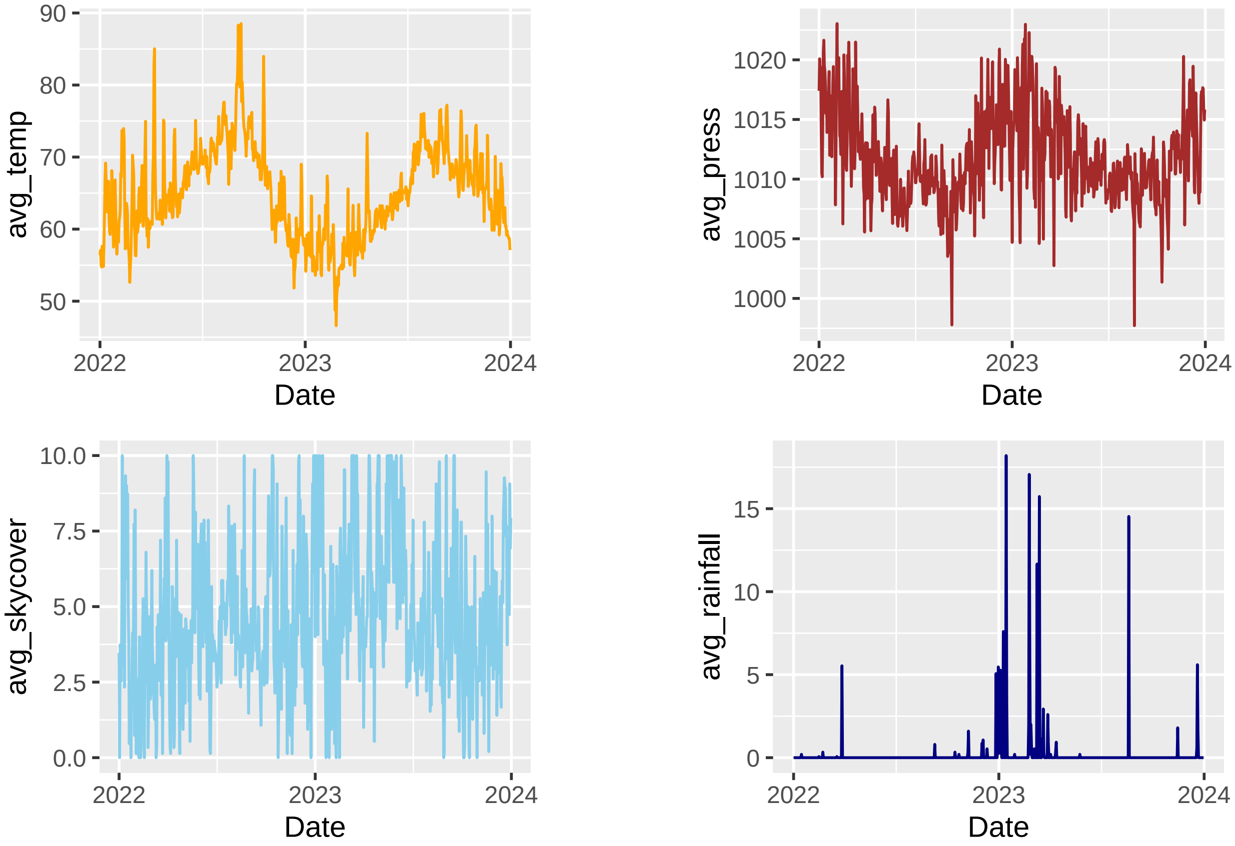
```
weather = rbind(weather_2022, weather_2023, weather_2024)

suppressWarnings(weather %<>% mutate(Temperature = as.numeric(Temperature),
                                          Pressure = as.numeric(Pressure),
                                          Sky_Cover = as.integer(Sky_Cover),
                                          Rainfall = as.integer(Rainfall)))
#write.csv(weather_2023, "weather_22.csv")
```

```
weather = weather %>%
  mutate(Temperature = na.approx(Temperature, na.rm="FALSE"),
         Pressure = na.approx(Pressure,na.rm="FALSE"),
         Sky_Cover = na.approx(Sky_Cover, na.rm="FALSE"),
         Rainfall = na.approx(Rainfall, na.rm="FALSE"))
```

```
weather_summary = weather %>%
  filter(Hour %in% c(7:21)) %>%
  group_by(Year, Month, Day)%>%
  summarise(avg_temp = mean(Temperature)*(9/5)+32,
         avg_press = mean(Pressure),
         avg_skycover = mean(Sky_Cover),
         avg_rainfall = mean(Rainfall),
         .groups = 'drop')
```

#EDA

```
time_plot_df = weather_summary %>%  filter(Year=="2022" | Year=="2023") %>%
  mutate(Date= as.Date(paste(Year, Month,Day, sep="-"), "%Y-%m-%d") )

p1<-time_plot_df %>% ggplot(aes(x=Date, y= avg_temp,))+
  geom_line(col="orange")+
  scale_x_date(date_breaks = "1 year",
               date_labels = "%Y")

p2<-time_plot_df %>% ggplot(aes(x=Date, y= avg_press))+
  geom_line(col="brown")+
  scale_x_date(date_breaks = "1 year",
               date_labels = "%Y")

p3<-time_plot_df %>% ggplot(aes(x=Date, y= avg_skycover))+
  geom_line(col="skyblue")+
  scale_x_date(date_breaks = "1 year",
               date_labels = "%Y")

p4<- time_plot_df %>% ggplot(aes(x=Date, y= avg_rainfall))+
  geom_line(col="navy")+
  scale_x_date(date_breaks = "1 year",
               date_labels = "%Y")

grid.arrange(p1,p2,p3,p4,
         nrow=2, ncol=3,
         widths = c(2,0.5,2),
         layout_matrix = rbind(c(1, NA, 2),
                               c(3,NA, 4)))
```

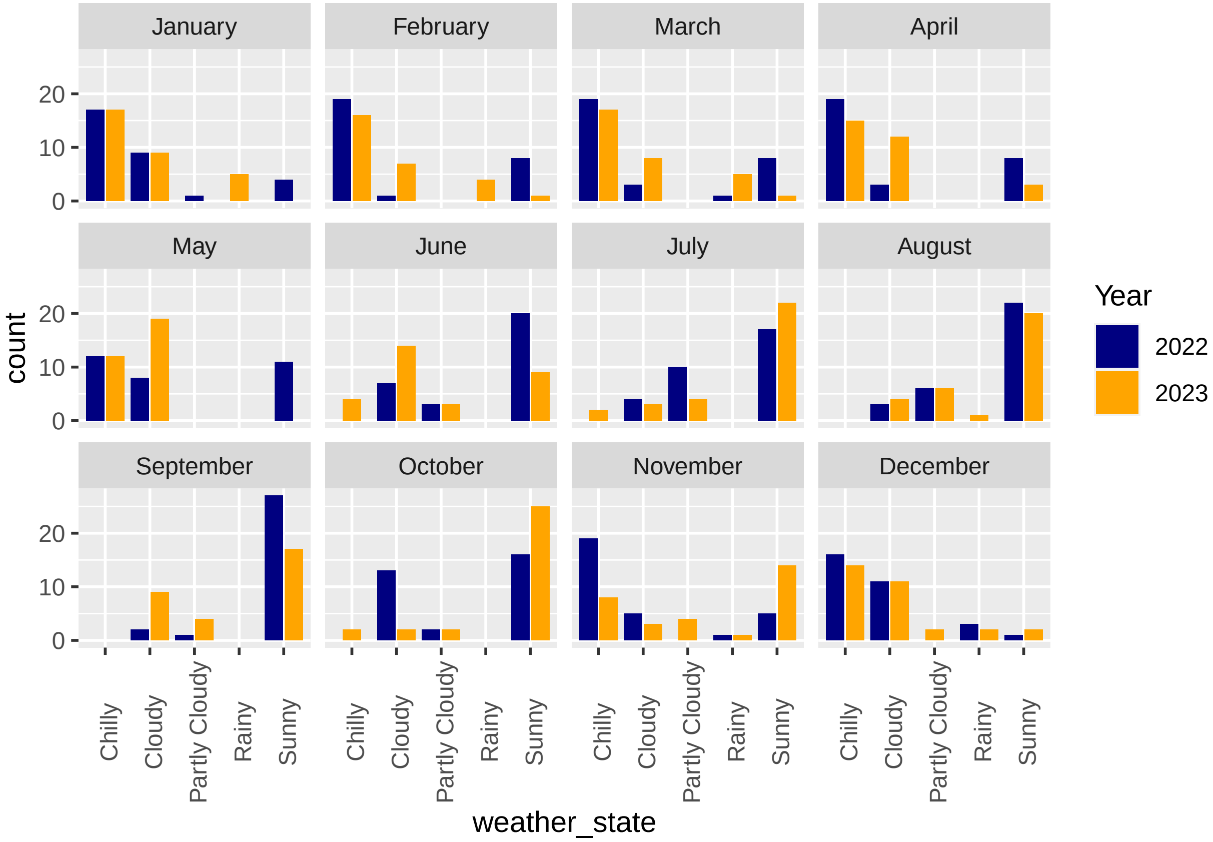


Categorical Variable Rules

```
weather_summary = weather_summary %>%
  mutate(Date= as.Date(paste(Year, Month,Day, sep="-"), "%Y-%m-%d") ) %>%
  filter(Date < as.Date("2024-05-1","%Y-%m-%d")) %>%
  mutate(weather_state = NA,
         weather_state = case_when(avg_rainfall>=1.0 ~ "Rainy",
                                   weather_state %in% NA & avg_skycover>6 ~ "Cloudy",
                                   weather_state %in% NA & avg_temp<65 ~ "Chilly" ,
                                   weather_state %in% NA & avg_temp>=65 & avg_skycover >5 ~ "Partly Clou
                                   weather_state %in% NA & avg_temp>=65 & avg_skycover<=5 | avg_press>=
         weather_state = as.factor(weather_state))

weather_summary[is.na(weather_summary$weather_state),]
```

```
## # A tibble: 0 x 9
## # i 9 variables: Year <int>, Month <int>, Day <int>, avg_temp <dbl>,
## #   avg_press <dbl>, avg_skycover <dbl>, avg_rainfall <dbl>, Date <date>,
## #   weather_state <fct>
```

```
temp_df = weather_summary %>% filter(Date < as.Date("2024-01-01","%Y-%m-%d"))
temp_df$Month = as.factor(temp_df$Month)
levels(temp_df$Month) <- c("January", "February", "March", "April", "May", "June", "July",
                "August", "September", "October", "November", "December")

temp_df %>% mutate(Year = year(Date),
                   Year = as.factor(Year))%>%
  ggplot(aes(x=weather_state, fill=Year)) +
  geom_bar(position = position_dodge2(preserve = "single")) +
  facet_wrap(~Month)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.7)) +
  scale_fill_manual(values=c("navy", "orange"))
```



```
write.csv(weather_summary, "weather.csv")
```