Accidents Analysis in the U.S.

Statistical Methods and Application I - CU Boulder

Kexin Yu

2021/11/21

Contents

1 Introduction

Accidents always have influence on our lives. A small and not serious accident can cause people to spend money that they don't need to spend, and a big and severe accident can cause people die. Accidents also cause traffic jams which can bring unnecessary trouble to people. So, it's important to know the reason that lead to accidents. Therefore, we can try our best to avoid these.

2 Identify and import the data

2.1 Load the packages

```
library(tidyverse)
library(readr)
library(graphics)
library(lubridate)
library(RColorBrewer)
```

2.2 Read the datasets

The first data set is about the accidents in U.S from February 2016 to December 2020. This dataset we used for this project was found via Kaggle. The site is https://www.kaggle.com/sobhanmoosavi/us-accidents.

```
US_Accidents <- read.csv('US_Accidents_Dec20_updated.csv',
                         header = TRUE, as.is = TRUE)
US_Accidents <- as_tibble(US_Accidents)
```

The description for each variable can be found in Kaggle, we just show the variables we need later.

Variable	Description
Start_Time	Shows start time of the accident in local time zone.
City	Shows the city in address field.
State	Shows the state in address field.
Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.

The second dataset is about the cases and deaths of the COVID-19 in U.S from January 2020 till now. This data will be used when we try to find the relationship between accidents and COVID-19. Data is coming from github site of Johns Hopkins University. The site is https://github.com/CSSEGISandData/COVID-19.

```
url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_d
US_Cases <- read_csv(url, show_col_types = FALSE)
```

2.3 Tidy the data

It's time to tidy all of data.

Firstly, we need to tidy the data about accidents in U.S.

After looking at the dataframe US_Accidents, we could find that there are many variables that we do not need, such as ID, End_Time, etc. So we want to select all of variables we need.

```
US_Accidents <- US_Accidents %>%
  select(Severity, Start_Time, City, State, Weather_Condition, Sunrise_Sunset)
```

And we can see that there are variety of weather conditions, because we do not need so elaborate, so we can divide them into different large categories.

```
US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Blowing Dust" |
  US_Accidents$Weather_Condition == "Blowing Dust / Windy" |
  US_Accidents$Weather_Condition == "Dust Whirls" |
  US_Accidents$Weather_Condition == "Sand" |
  US_Accidents$Weather_Condition == "Sand / Dust Whirls Nearby" |
  US_Accidents$Weather_Condition == "Sand / Dust Whirlwinds" |
  US_Accidents$Weather_Condition == "Volcanic Ash" |
  US_Accidents$Weather_Condition == "Widespread Dust")] <- "Dust"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Blowing Snow" |
  US_Accidents$Weather_Condition == "Blowing Snow / Windy" |
  US_Accidents$Weather_Condition == "Drifting Snow" |
  US_Accidents$Weather_Condition == "Heavy Blowing Snow" |
  US_Accidents$Weather_Condition == "Heavy Snow" |
  US_Accidents$Weather_Condition == "Heavy Snow / Windy" |
  US_Accidents$Weather_Condition == "Heavy Snow with Thunder" |
  US_Accidents$Weather_Condition == "Light Blowing Snow" |
  US_Accidents$Weather_Condition == "Light Snow" |
  US_Accidents$Weather_Condition == "Light Snow / Windy" |
  US_Accidents$Weather_Condition == "Light Snow Shower" |
  US_Accidents$Weather_Condition == "Light Snow Showers" |
  US_Accidents$Weather_Condition == "Light Snow with Thunder" |
  US_Accidents$Weather_Condition == "Low Drifting Snow" |
  US_Accidents$Weather_Condition == "Snow / Windy" |
  US_Accidents$Weather_Condition == "Snow Grains")] <- "Snow"

US_Accidents$Weather_Condition[which(
```

```
  US_Accidents$Weather_Condition == "Cloudy / Windy" |
  US_Accidents$Weather_Condition == "Funnel Cloud" |
  US_Accidents$Weather_Condition == "Mostly Cloudy" |
  US_Accidents$Weather_Condition == "Mostly Cloudy / Windy" |
  US_Accidents$Weather_Condition == "Partly Cloudy" |
  US_Accidents$Weather_Condition == "Partly Cloudy / Windy" |
  US_Accidents$Weather_Condition == "Scattered Clouds")] <- "Cloudy"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Drizzle" |
  US_Accidents$Weather_Condition == "Drizzle / Windy" |
  US_Accidents$Weather_Condition == "Drizzle and Fog" |
  US_Accidents$Weather_Condition == "Freezing Drizzle" |
  US_Accidents$Weather_Condition == "Freezing Rain" |
  US_Accidents$Weather_Condition == "Freezing Rain / Windy" |
  US_Accidents$Weather_Condition == "Heavy Drizzle" |
  US_Accidents$Weather_Condition == "Heavy Freezing Drizzle" |
  US_Accidents$Weather_Condition == "Heavy Rain" |
  US_Accidents$Weather_Condition == "Heavy Rain / Windy" |
  US_Accidents$Weather_Condition == "Heavy Rain Shower" |
  US_Accidents$Weather_Condition == "Heavy Rain Showers" |
  US_Accidents$Weather_Condition == "Light Drizzle" |
  US_Accidents$Weather_Condition == "Light Drizzle / Windy" |
  US_Accidents$Weather_Condition == "Light Freezing Drizzle" |
  US_Accidents$Weather_Condition == "Light Freezing Rain" |
  US_Accidents$Weather_Condition == "Light Freezing Rain / Windy" |
  US_Accidents$Weather_Condition == "Light Rain" |
  US_Accidents$Weather_Condition == "Light Rain / Windy" |
  US_Accidents$Weather_Condition == "Light Rain Shower" |
  US_Accidents$Weather_Condition == "Light Rain Shower / Windy" |
  US_Accidents$Weather_Condition == "Light Rain Showers" |
  US_Accidents$Weather_Condition == "Light Rain with Thunder" |
  US_Accidents$Weather_Condition == "Rain / Windy" |
  US_Accidents$Weather_Condition == "Rain Shower" |
  US_Accidents$Weather_Condition == "Rain Showers" |
  US_Accidents$Weather_Condition == "Showers in the Vicinity")] <- "Rain"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Fair / Windy")] <- "Fair"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Fog / Windy" |
  US_Accidents$Weather_Condition == "Light Fog" |
  US_Accidents$Weather_Condition == "Light Freezing Fog" |
```

```
  US_Accidents$Weather_Condition == "Partial Fog" |
  US_Accidents$Weather_Condition == "Patches of Fog" |
  US_Accidents$Weather_Condition == "Patches of Fog / Windy" |
  US_Accidents$Weather_Condition == "Shallow Fog" |
  US_Accidents$Weather_Condition == "Mist" |
  US_Accidents$Weather_Condition == "Mist / Windy")] <- "Fog"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Haze / Windy" |
  US_Accidents$Weather_Condition == "Light Haze")] <- "Haze"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Hail" |
  US_Accidents$Weather_Condition == "Heavy Ice Pellets" |
  US_Accidents$Weather_Condition == "Light Ice Pellets" |
  US_Accidents$Weather_Condition == "N/A Precipitation" |
  US_Accidents$Weather_Condition == "Small Hail")] <- "Ice Pellets"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Heavy T-Storm" |
  US_Accidents$Weather_Condition == "Heavy T-Storm / Windy" |
  US_Accidents$Weather_Condition == "Heavy Thunderstorms and Rain" |
  US_Accidents$Weather_Condition == "Heavy Thunderstorms and Snow" |
  US_Accidents$Weather_Condition == "Heavy Thunderstorms with Small Hail" |
  US_Accidents$Weather_Condition == "Light Thunderstorms and Rain" |
  US_Accidents$Weather_Condition == "Light Thunderstorms and Snow" |
  US_Accidents$Weather_Condition == "T-Storm" |
  US_Accidents$Weather_Condition == "T-Storm / Windy" |
  US_Accidents$Weather_Condition == "Thunderstorms and Rain")] <- "Thunderstorm"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Thunder / Windy" |
  US_Accidents$Weather_Condition == "Thunder / Wintry Mix / Windy" |
  US_Accidents$Weather_Condition == "Thunder and Hail" |
  US_Accidents$Weather_Condition == "Thunder and Hail / Windy" |
  US_Accidents$Weather_Condition == "Thunder in the Vicinity")] <- "Thunder"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Light Sleet" |
  US_Accidents$Weather_Condition == "Light Sleet / Windy" |
  US_Accidents$Weather_Condition == "Light Snow and Sleet" |
  US_Accidents$Weather_Condition == "Light Snow and Sleet / Windy" |
  US_Accidents$Weather_Condition == "Sleet / Windy" |
  US_Accidents$Weather_Condition == "Snow and Sleet" |
```

```
  US_Accidents$Weather_Condition == "Snow and Sleet / Windy" |
  US_Accidents$Weather_Condition == "Wintry Mix" |
  US_Accidents$Weather_Condition == "Wintry Mix / Windy")] <- "Sleet"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Smoke / Windy")] <- "Smoke"

US_Accidents$Weather_Condition[which(
  US_Accidents$Weather_Condition == "Squalls" |
  US_Accidents$Weather_Condition == "Squalls / Windy" |
  US_Accidents$Weather_Condition == "Tornado")] <- "Windy"
```

Secondly, we need to tidy the data about COVID-19 in U.S.

We just need the number of people who were infected by COVID-19, and change the format of Date, so that we can use this data more convenient.

```
US_Cases <- US_Cases %>%
  select(-UID, -iso2, -iso3, -code3, -FIPS, -Province_State,
         -Country_Region, -Lat, -Long_, -Combined_Key) %>%
  pivot_longer( -Admin2, names_to = "Date", values_to = "Cases") %>%
  select(Date, Cases) %>%
  mutate(Date = mdy(Date))
```

3 Factors affecting the number of accidents

3.1 Weekday and weekend

Firstly, we need to select the columns we need.

```
Time <- US_Accidents %>%
  select(Severity, Start_Time, Sunrise_Sunset) %>%
  glimpse()
```

```
## Rows: 1,516,064
## Columns: 3
## $ Severity       <int> 3, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3~
## $ Start_Time     <chr> "2016-02-08 00:37:08", "2016-02-08 05:56:20", "2016-02-~
## $ Sunrise_Sunset <chr> "Night", "Night", "Night", "Night", "Night", "Day", "Da~
```

According to the data above, we find that Start_Time include two parts date and time. So we need to separate the Start_Time into two columns Date and Time. Date stores the specific year,month and day. Time stores the specific hour and minute.

```
Time <- Time %>%
  separate(col = Start_Time, into = c("Date", "Time"), sep = " ") %>%
  glimpse()
```

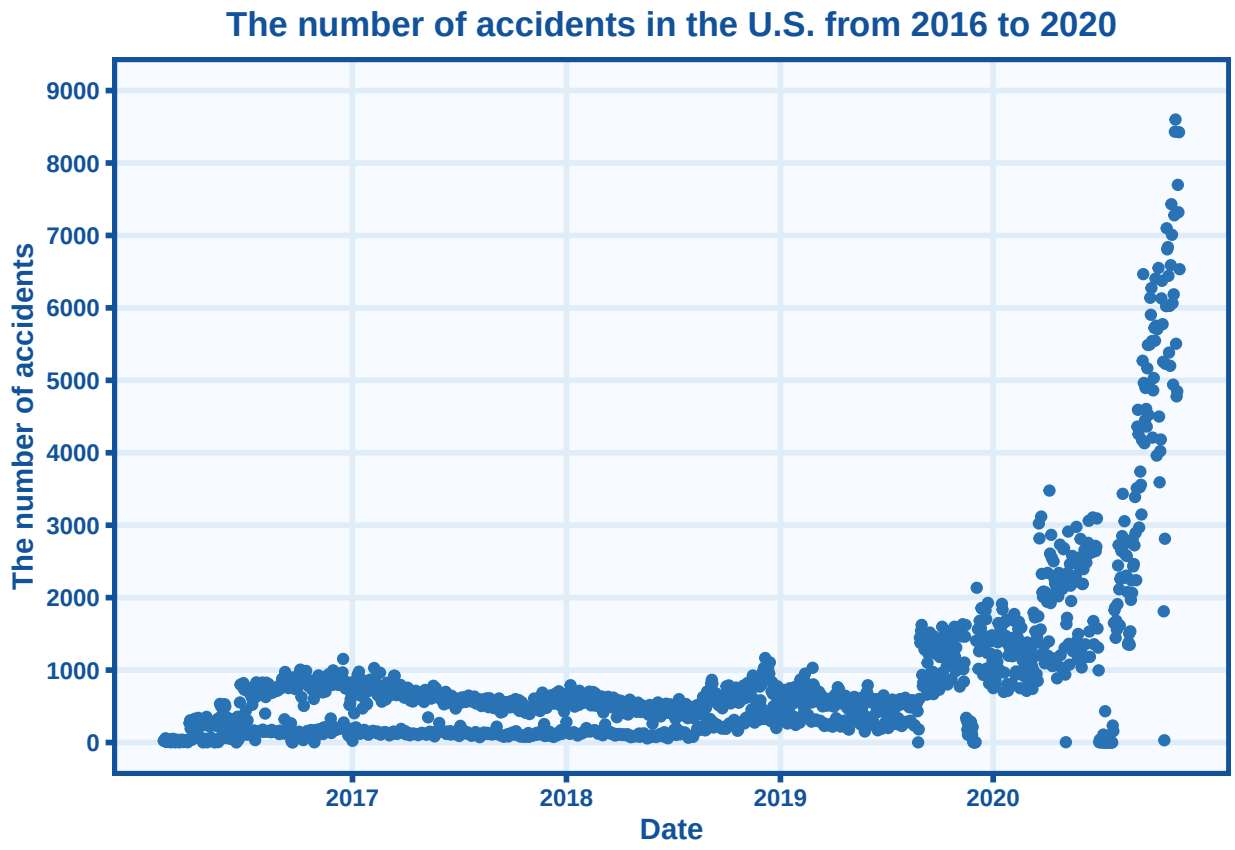
```
## Rows: 1,516,064
## Columns: 4
## $ Severity       <int> 3, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3~
## $ Date           <chr> "2016-02-08", "2016-02-08", "2016-02-08", "2016-02-08",~
## $ Time           <chr> "00:37:08", "05:56:20", "06:15:39", "06:15:39", "06:51:~
## $ Sunrise_Sunset <chr> "Night", "Night", "Night", "Night", "Night", "Day", "Da~
```

Let's calculate the number of accidents occur per day.

```
Accidents_per_day <- Time %>%
  group_by(Date) %>%
  summarize(number = table(Date))
```

Then we use scatter plot to show the number of accidents from 2016-02-08 to 2020-12-31. According to this plot, because of so many days, we can only see an roughly increase tendency. And also because the difference between the number of accidents in different years is great, we cannot see the specific tendency except 2020.

```r
Accidents_per_day %>%
  ggplot(aes(x = Date, y = number)) +
  geom_point(color = "#2973B5") +
  scale_x_discrete(expand = c(0.05,0), breaks = c("2017-01-01", "2018-01-01",
                                                  "2019-01-01", "2020-01-01"),
                   labels = c("2017", "2018", "2019", "2020")) +
  scale_y_continuous(limits = c(0, 9000), breaks = seq(0, 9000, 1000)) +
  theme_bw() +
  labs(title = "The number of accidents in the U.S. from 2016 to 2020") +
  ylab("The number of accidents") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1))
```

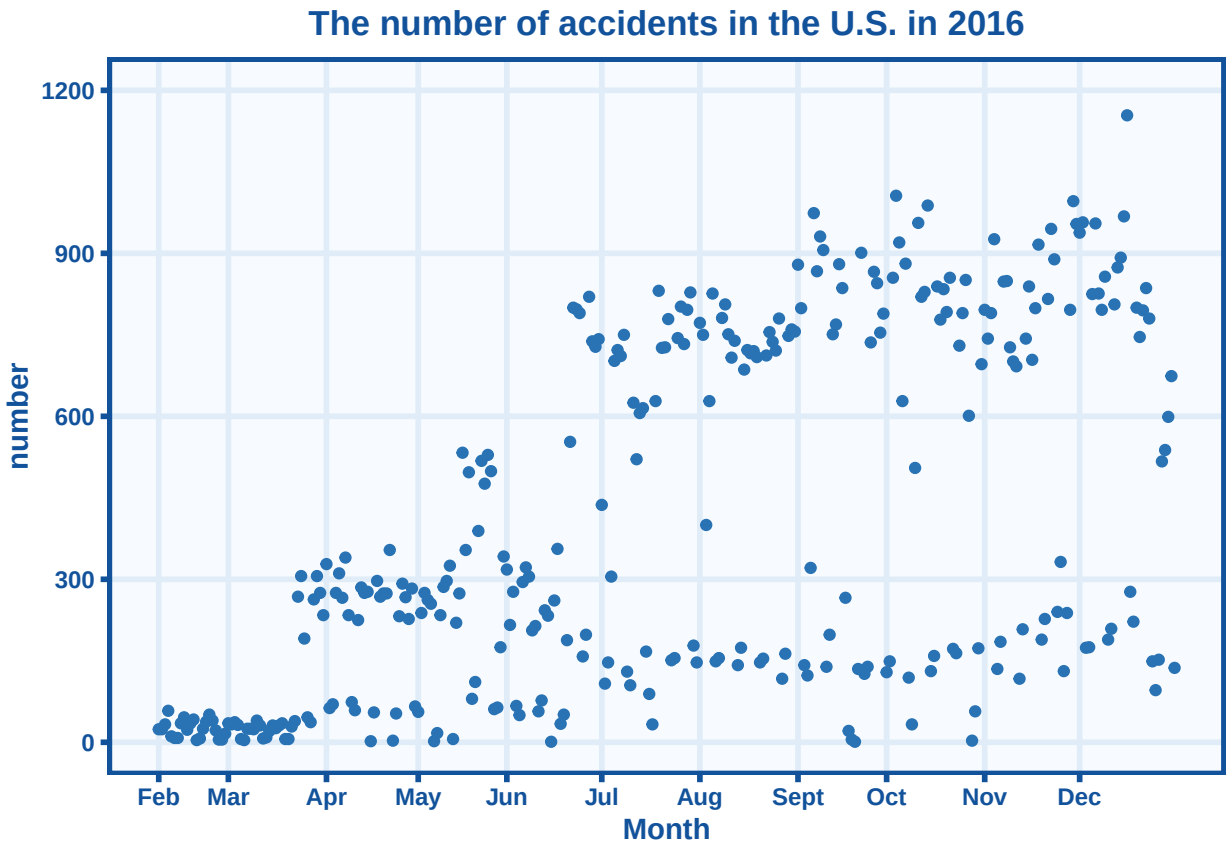


It's necessary for us to show the data in five different graphs. Therefore, we can see the tendency clearly.

```
Accidents_per_day <- Accidents_per_day %>%
  separate(col = Date, into = c("Year", "Month", "Day"), sep = "-") %>%
  unite(col = "Month_Day", Month, Day, sep = "")
```

In 2016, we can see that except the February and March when the record began, almost all of other months have two parts: the number of accidents is high and the number of accidents is very low.

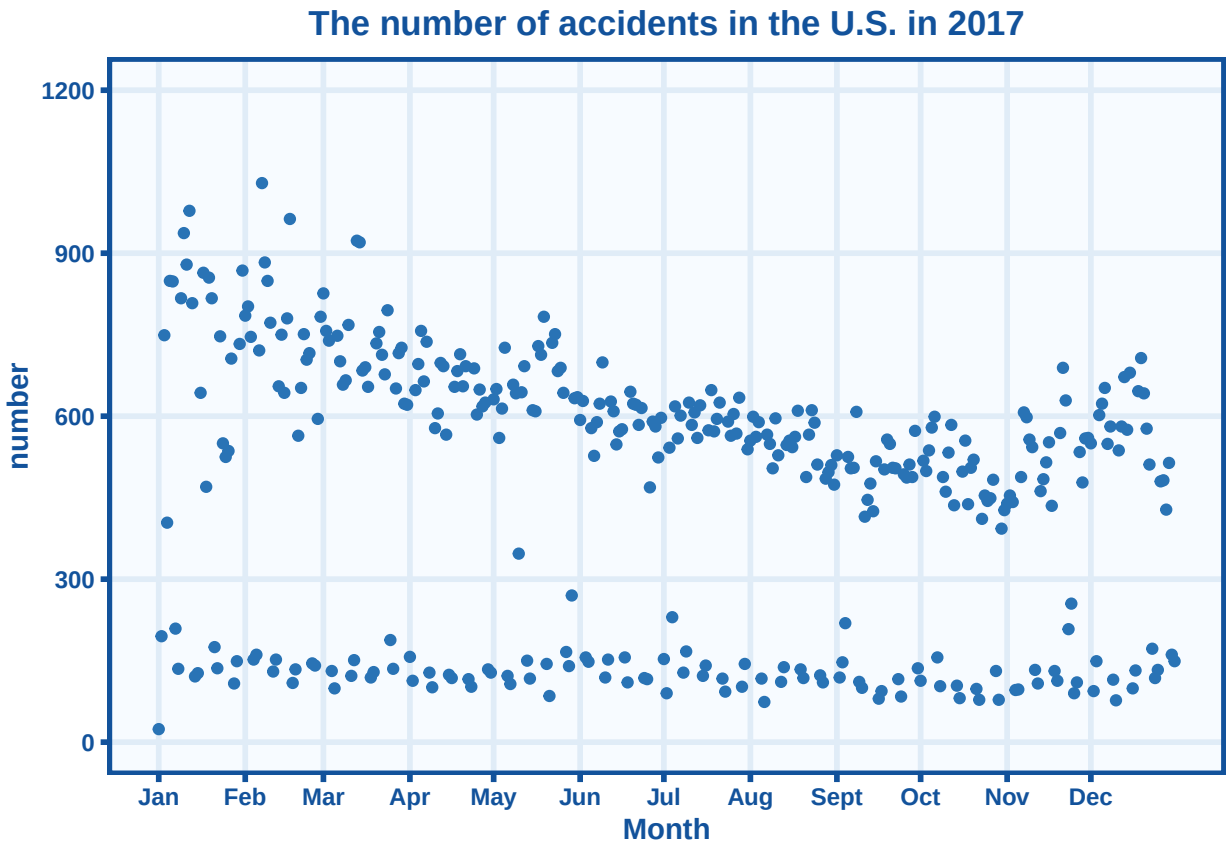
```
Accidents_per_day %>%
  filter(Year == "2016") %>%
  ggplot(aes(x = Month_Day, y = number)) +
  geom_point(color = "#2973B5") +
  scale_x_discrete(expand = c(0.05,0), breaks = c("0208", "0301", "0401",
                                                  "0501", "0601", "0701",
                                                  "0801", "0901", "1001",
                                                  "1101", "1201"),
                   labels = c("Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug",
                              "Sept", "Oct", "Nov", "Dec")) +
  scale_y_continuous(limits = c(0, 1200), breaks = seq(0, 1200, 300)) +
  theme_bw() +
  labs(title = "The number of accidents in the U.S. in 2016", x = "Month") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1))
```



In 2017 and 2018, it is easier for us to find that these patterns of accidents are the same as the pattern of accidents in 2016.

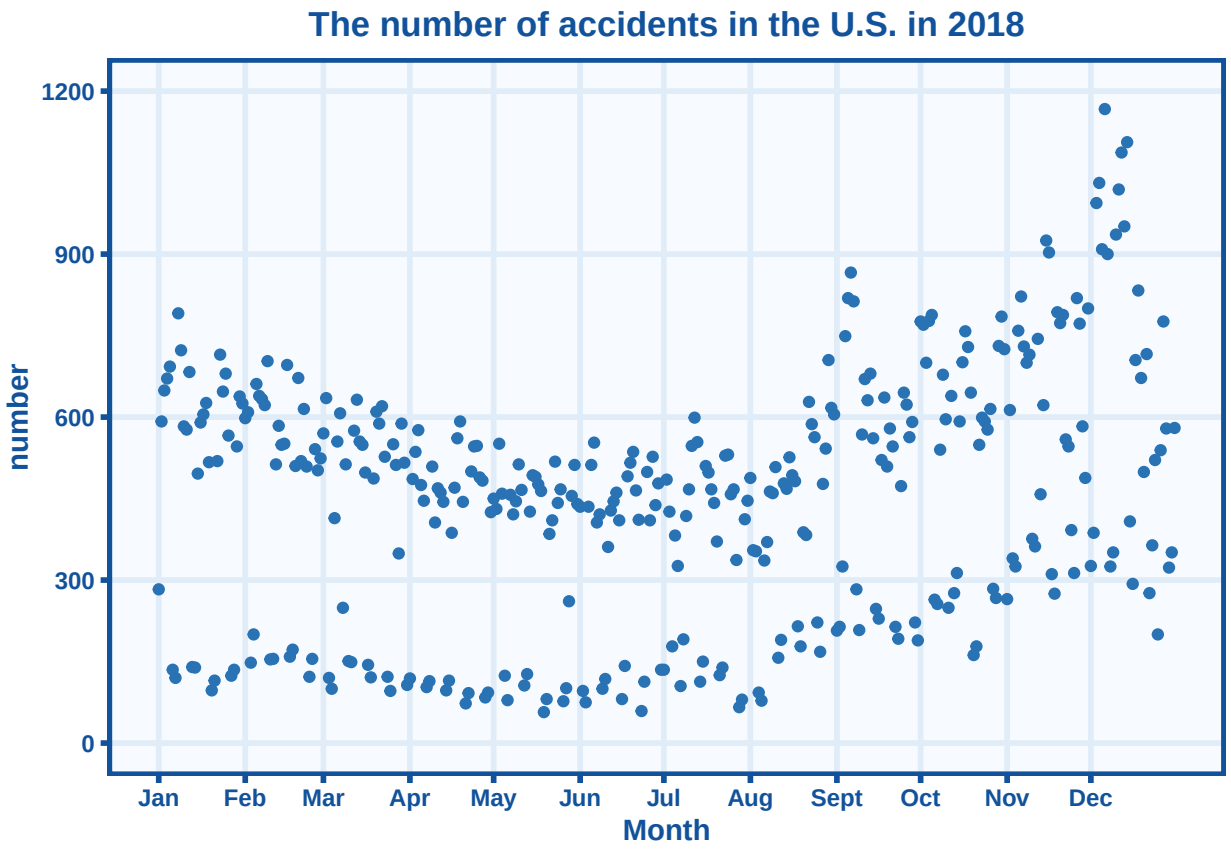
```
Accidents_per_day %>%
  filter(Year == "2017") %>%
  ggplot(aes(x = Month_Day, y = number)) +
  geom_point(color = "#2973B5") +
  scale_x_discrete(expand = c(0.05,0), breaks = c("0101", "0201", "0301",
                                                  "0401", "0501", "0601",
                                                  "0701", "0801", "0901",
                                                  "1001", "1101", "1201"),
                   labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
                              "Aug", "Sept", "Oct", "Nov", "Dec")) +
  scale_y_continuous(limits = c(0, 1200), breaks = seq(0, 1200, 300)) +
  theme_bw() +
  labs(title = "The number of accidents in the U.S. in 2017", x = "Month") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
```

```
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1))
```




```
Accidents_per_day %>%
  filter(Year == "2018") %>%
  ggplot(aes(x = Month_Day, y = number)) +
  geom_point(color = "#2973B5") +
  scale_x_discrete(expand = c(0.05,0), breaks = c("0101", "0201", "0301",
                                                  "0401", "0501", "0601",
                                                  "0701", "0801", "0901",
                                                  "1001", "1101", "1201"),
                   labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
                              "Aug", "Sept", "Oct", "Nov", "Dec")) +
  scale_y_continuous(limits = c(0, 1200), breaks = seq(0, 1200, 300)) +
  theme_bw() +
  labs(title = "The number of accidents in the U.S. in 2018", x = "Month") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
```

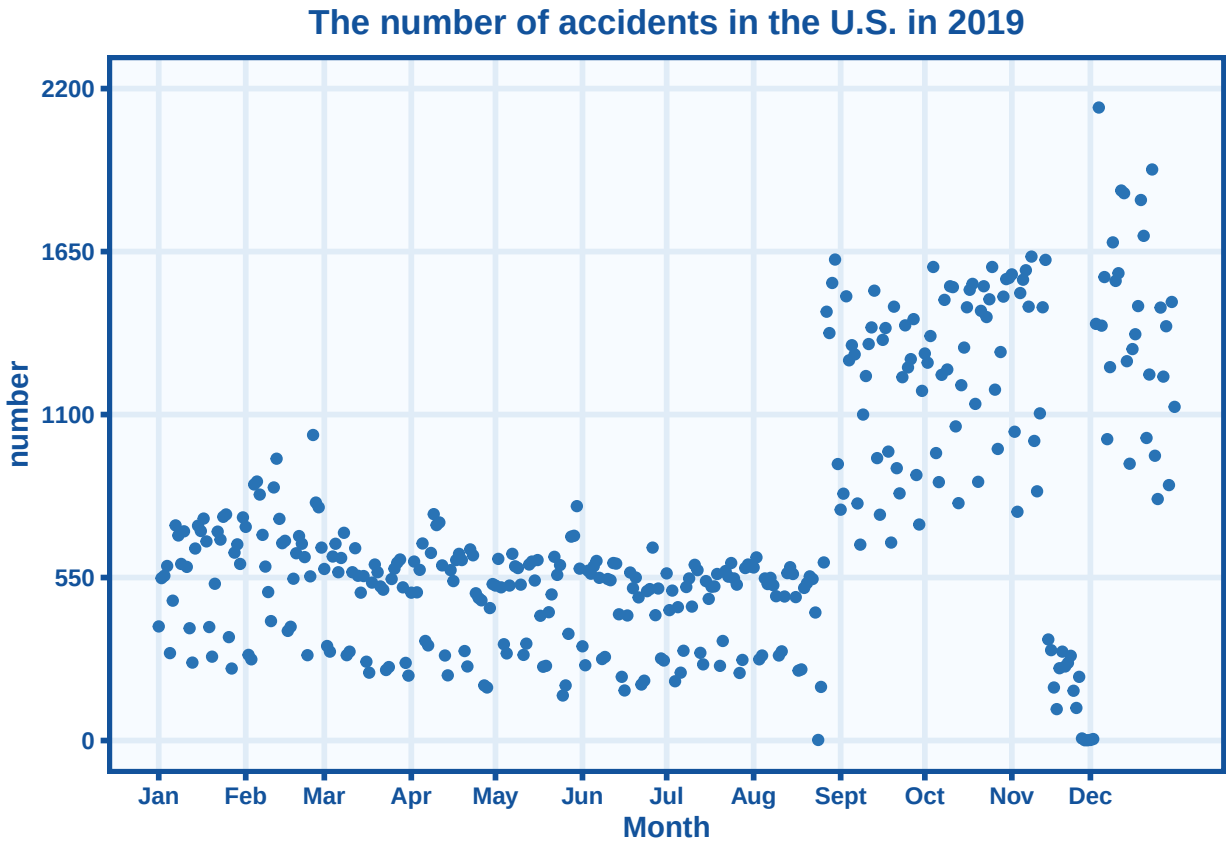
```
        panel.border = element_rect(color = "#12529B", size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1))
```



In 2019, we find that there also are two parts between January and August, and the other months are not as clearly as these 8 months, but we can find that the two parts also exist. And we also find that the number of accidents is doubled or even tripled than before between September and December.

```
Accidents_per_day %>%
  filter(Year == "2019") %>%
  ggplot(aes(x = Month_Day, y = number)) +
  geom_point(color = "#2973B5") +
  scale_x_discrete(expand = c(0.05,0), breaks = c("0101", "0201", "0301",
                                                  "0401", "0501", "0601",
                                                  "0701", "0801", "0901",
                                                  "1001", "1101", "1201"),
                   labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
                              "Aug", "Sept", "Oct", "Nov", "Dec")) +
  scale_y_continuous(limits = c(0, 2200), breaks = seq(0, 2200, 550)) +
```

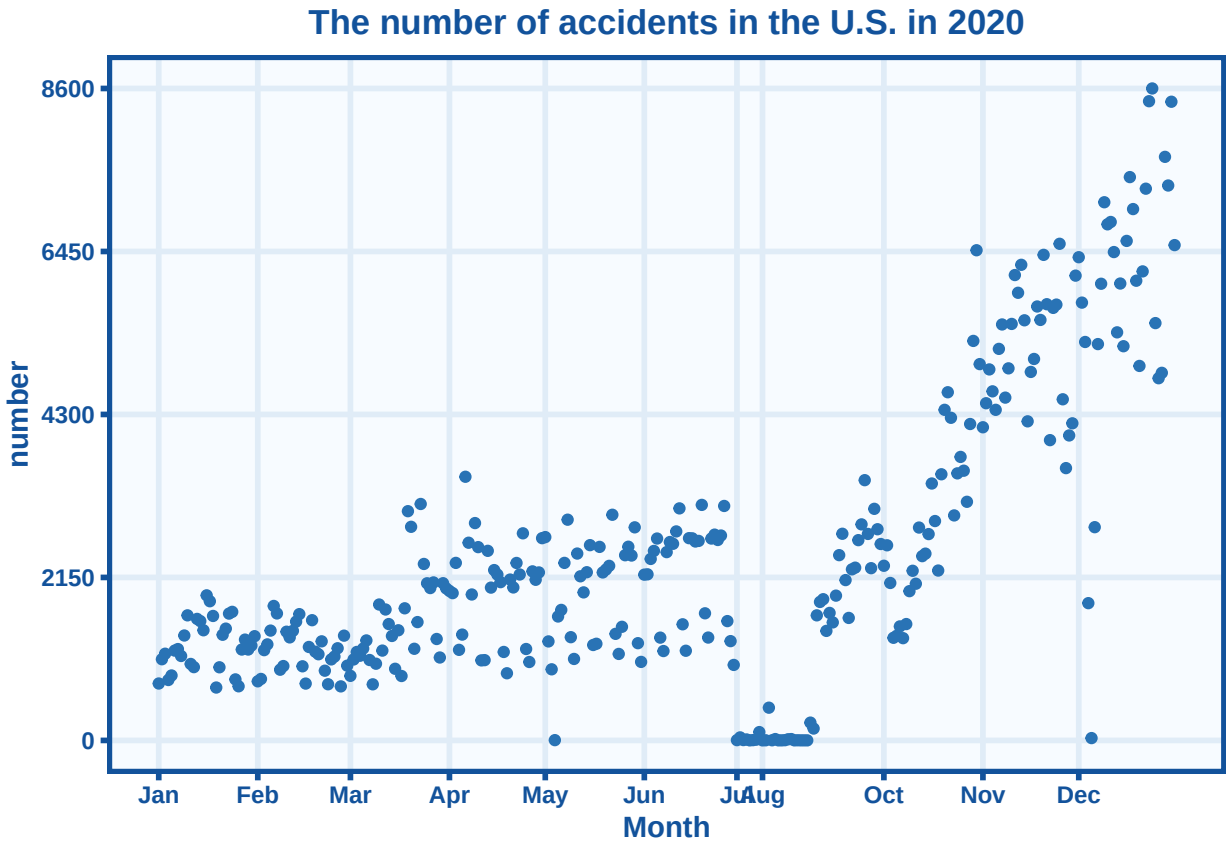
```
  theme_bw() +
  labs(title = "The number of accidents in the U.S. in 2019", x = "Month") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1))
```



In 2020, we can notice that the pattern is not the same as before, only a few months have two parts. And also, in this year, the number of accidents increase rapidly.

```
Accidents_per_day %>%
  filter(Year == "2020") %>%
  ggplot(aes(x = Month_Day, y = number)) +
  geom_point(color = "#2973B5") +
  scale_x_discrete(expand = c(0.05,0), breaks = c("0101", "0201", "0301",
                                                  "0401", "0501", "0601",
                                                  "0701", "0801", "0901",
```

```
                                                    "1001", "1101", "1201"),
                   labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
                              "Aug", "Sept", "Oct", "Nov", "Dec")) +
scale_y_continuous(limits = c(0, 8600), breaks = seq(0, 8600, 2150)) +
theme_bw() +
labs(title = "The number of accidents in the U.S. in 2020", x = "Month") +
theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
      panel.background = element_rect(fill = "#F7FBFF"),
      panel.grid = element_blank(),
      panel.grid.major = element_line(color = "#E0ECF7", size = 1),
      panel.border = element_rect(color = "#12529B", size = 1.5),
      axis.title = element_text(color = "#12529B", face = "bold"),
      axis.text = element_text(color = "#12529B", face = "bold"),
      axis.ticks = element_line(color = "#12529B", size = 1))
```



According to the five graphs above, in these years, over three quarters of month has the pattern that there are some days which cause a large number of accidents, and there are also some days which cause not as many as that. So, let's see if there are some rules? Maybe the days which have less accidents are weekends, because of less people to go to work by cars?

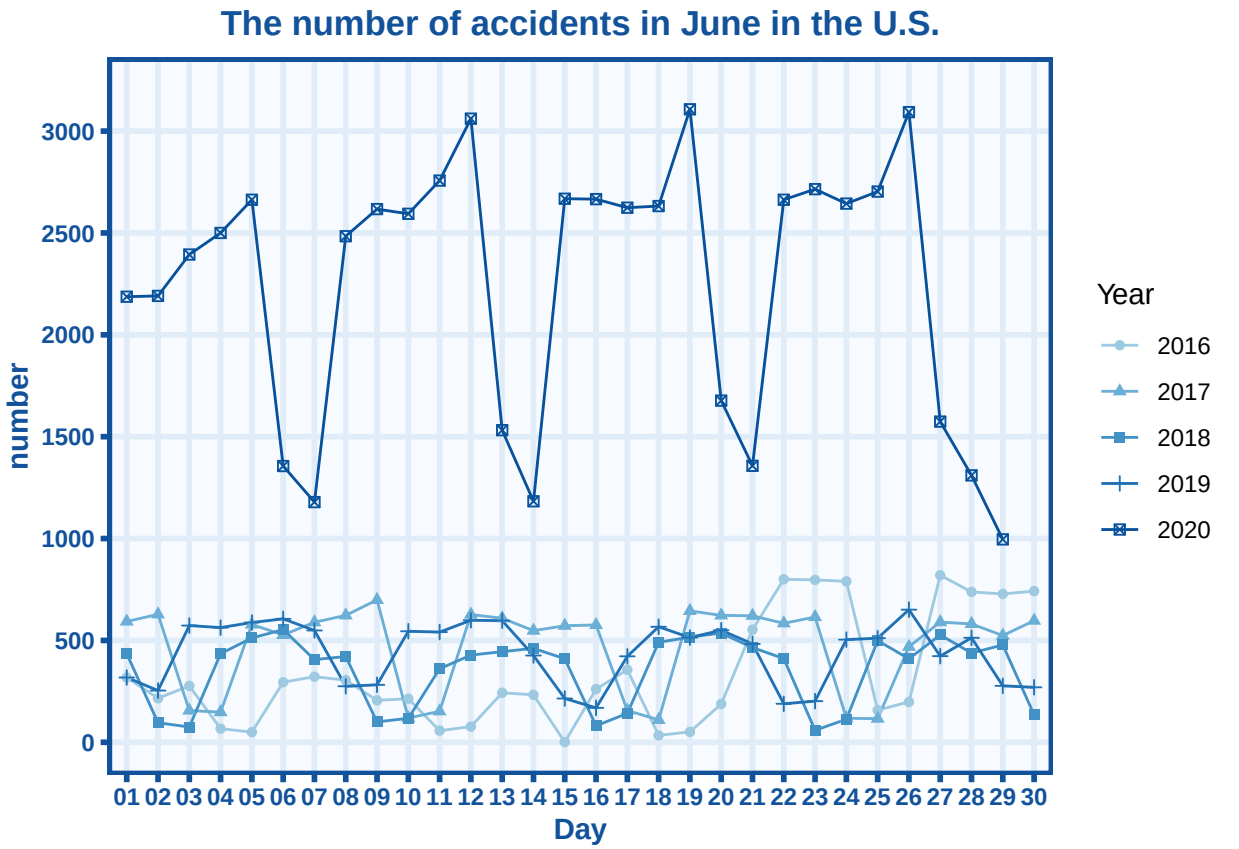
Let's choose some months in every year as our examples to prove if this speculation is right.

Firstly, we choose June. According to this graph, there are two successive days in seven successive days have less accidents. Although, the number of accidents do not decrease and increase in the same day, the weekends in every year are not always in the same day, so let's look at the calendar to check if our speculation is true.

In June 2020, 6th and 7th, 13th and 14th, 20th and 21st, 27th and 28th are the weekends. We know that the number of accidents in these days is less than other days. So, these results are the same as our speculation. In June 2019, 1st and 2nd, 8th and 9th, 15th and 16th, 22nd and 23rd, 29th and 30th are the weekends. And the number of accidents in these days is still less than other days. This rule still holds true in other years(2016, 2017 and 2018).

```
Accidents_Jun <- Accidents_per_day %>%
  separate(col = "Month_Day", into = c("Month", "Day"), sep = 2) %>%
  filter(Month == "06")

Accidents_Jun %>%
  ggplot(aes(x = Day, y = number, group = Year, color = Year, pch = Year)) +
  geom_line() +
  geom_point() +
  scale_y_continuous(limits = c(0, 3200), breaks = seq(0, 3200, 500)) +
  theme_bw() +
  scale_color_manual(values = brewer.pal(9, "Blues")[4:8]) +
  labs(title = "The number of accidents in June in the U.S.") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1))
```

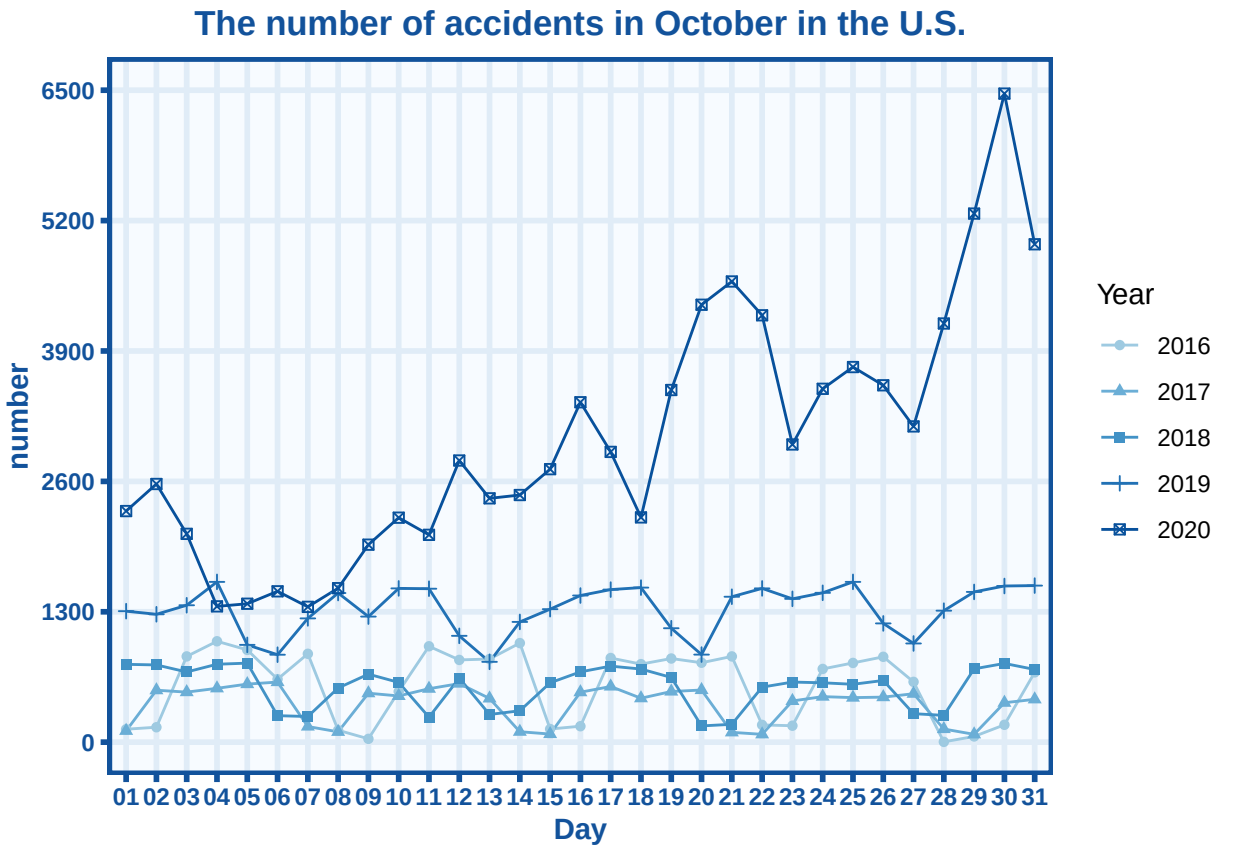


Secondly, we choose October. Indeed, except the 2020, other years all have the same rule as we speculate. As we know, from October 2020, the number of accidents increase rapidly. So maybe this is the reason why 2020 does not follow this rule.

```
Accidents_Oct <- Accidents_per_day %>%
  separate(col = "Month_Day", into = c("Month", "Day"), sep = 2) %>%
  filter(Month == "10")

Accidents_Oct %>%
  ggplot(aes(x = Day, y = number, group = Year, color = Year, pch = Year)) +
  geom_line() +
  geom_point() +
  scale_y_continuous(limits = c(0, 6500), breaks = seq(0, 6500, 1300)) +
  theme_bw() +
  scale_color_manual(values = brewer.pal(9, "Blues")[4:8]) +
  labs(title = "The number of accidents in October in the U.S.") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
```

```
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1))
```



Therefore, we can confirm that the number of accidents is related to the days whether weekdays or not.

3.2 Day and night

Firstly, we find that there are three kinds of value: null, Day and Night.

```
Time %>%
  group_by(Sunrise_Sunset) %>%
  summarize(test = table(Sunrise_Sunset))
```

```
## # A tibble: 3 x 2
##   Sunrise_Sunset test
##   <chr>          <table>
## 1 ""                 83
## 2 "Day"          909838
## 3 "Night"        606143
```

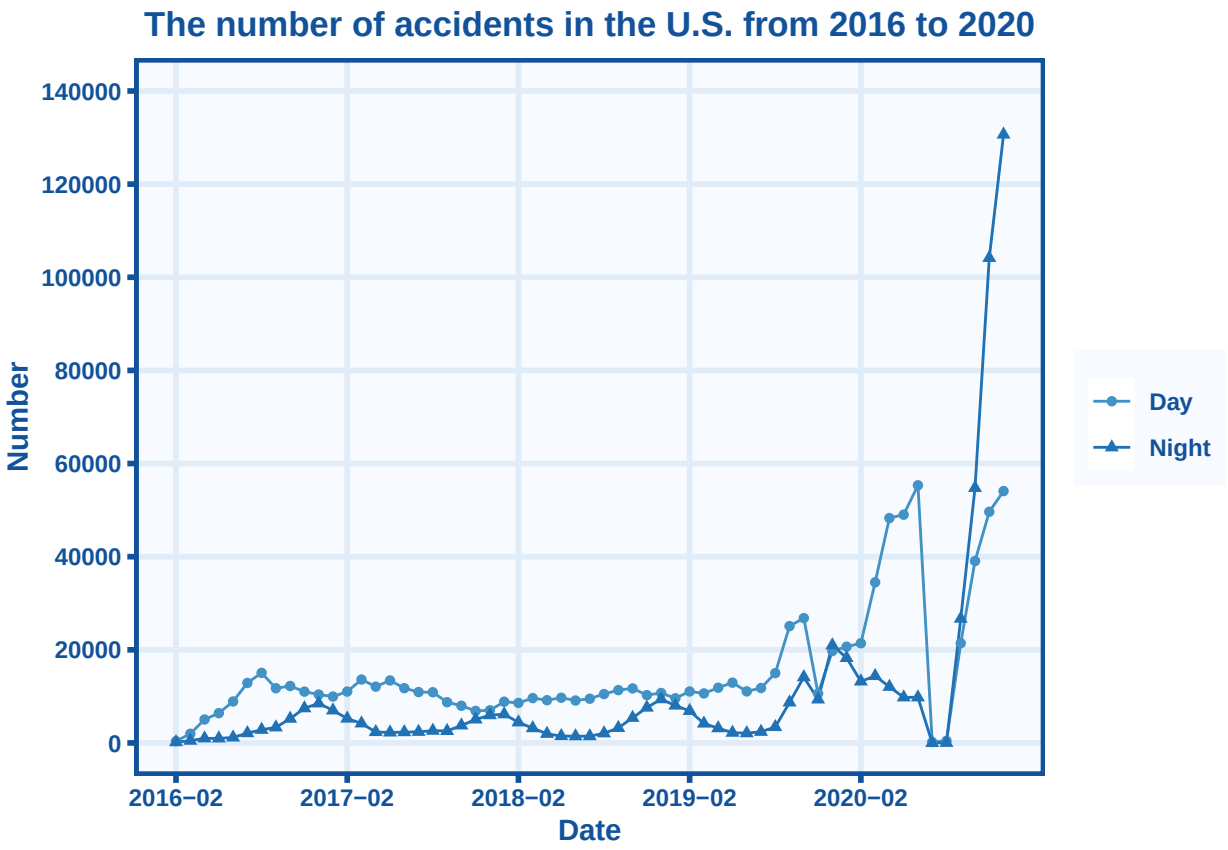
Because day is no use, so we can separate the Date into two parts and then delete the day part. Therefore, we just need to select the columns Date(year and month) and Sunrise_Sunset. It's necessary to delete the null value. And then, calculate the number of accidents occurred during the day and night every month.

```
Accidents_time <- Time %>%
  separate(col = "Date", into = c("Date", "Day"), sep = 7) %>%
  select(Date, Sunrise_Sunset) %>%
  filter(Sunrise_Sunset != "") %>%
  mutate(DN = ifelse(Sunrise_Sunset == "Day", "Day", "Night")) %>%
  unite(col = "Date", Date, DN, sep = " ") %>%
  group_by(Date) %>%
  summarize(Number = table(Sunrise_Sunset)) %>%
  separate(col = "Date", into = c("Date", "Sunrise_Sunset"), sep = " ")
```

Let's use line chart to show the number of accidents that occurred during the day and night each month.

```
Accidents_time %>%
  ggplot(aes(x = Date, y = Number, group = Sunrise_Sunset,
             color = Sunrise_Sunset, pch = Sunrise_Sunset)) +
  geom_line() +
  geom_point() +
  scale_x_discrete(expand = c(0.05,0), breaks = c("2016-02", "2017-02",
                                                  "2018-02", "2019-02",
                                                  "2020-02"),
                   labels = c("2016-02", "2017-02", "2018-02", "2019-02",
                              "2020-02")) +
```

```
  scale_y_continuous(limits = c(0, 140000), breaks = seq(0, 140000, 20000)) +
  theme_bw() +
  scale_color_manual(values = brewer.pal(9, "Blues")[6:7]) +
  labs(title = "The number of accidents in the U.S. from 2016 to 2020") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        legend.title = element_blank(),
        legend.text = element_text(color = "#12529B", face = "bold"),
        legend.background = element_rect(fill = "#F7FBFF"),
        axis.ticks = element_line(color = "#12529B", size = 1))
```



In this graph, we can find that except the last four months, almost every month the number of accidents occurred during the day is more than the number of accidents occurred during the night. So, we cannot affirm that accidents must be more likely during day, but most of the time it is.

3.3 Location

Firstly, we need to select the columns we need.

```
Location <- US_Accidents %>%
  select(Severity, City, State) %>%
  glimpse()
```

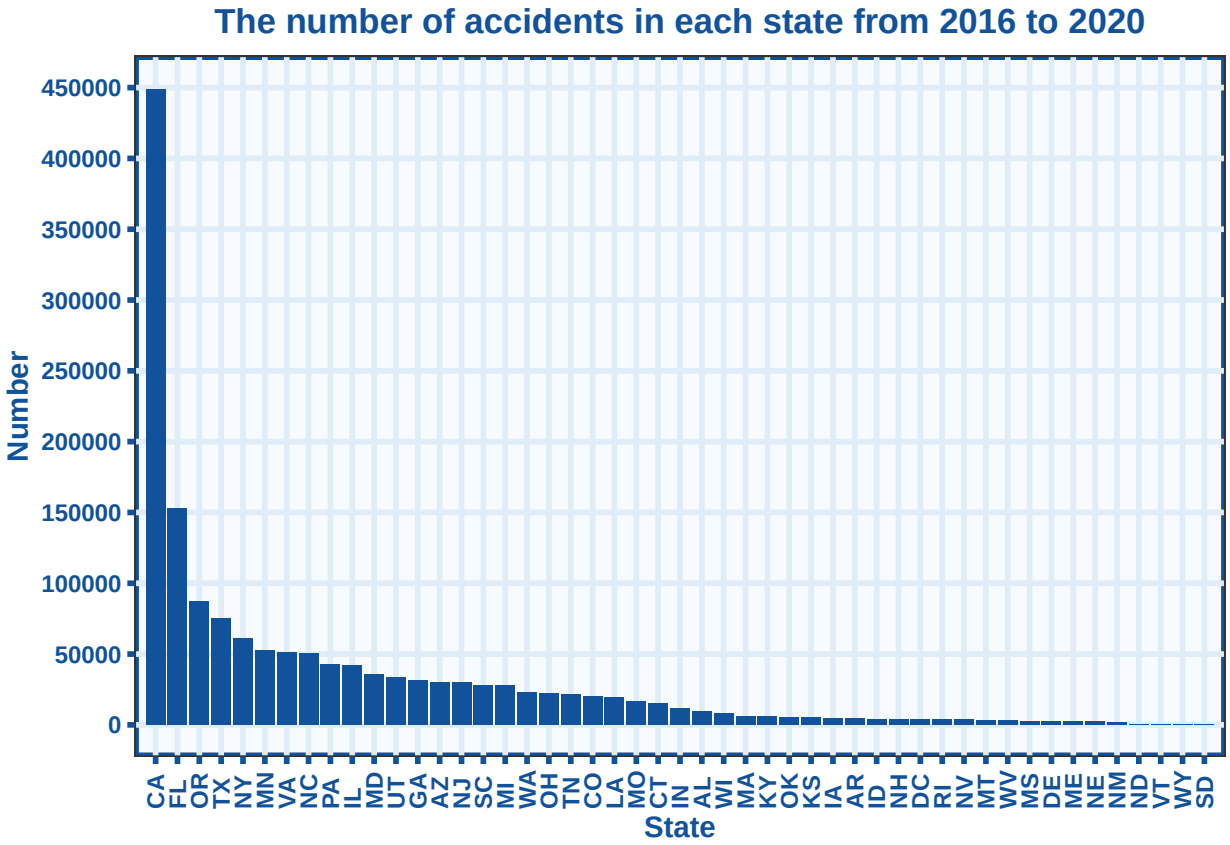
```
## Rows: 1,516,064
## Columns: 3
## $ Severity <int> 3, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 2, 3~
## $ City     <chr> "Dublin", "Dayton", "Cincinnati", "Cincinnati", "Akron", "Cin~
## $ State    <chr> "OH", "OH", "OH", "OH", "OH", "OH", "OH", "OH", "OH", "OH", "~
```

Then, in order to know the number of accidents that occurred in every state, we decide to figure them up.

```
Accidents_state <- Location %>%
  group_by(State) %>%
  summarize(Number = table(State))
```

Let's show this data in a bar graph. We notice that CA is the state with the most accidents. There were three times as many accidents as the second state. It was followed by FL, OR, TX and NY.

```
Accidents_state %>%
  mutate(State = fct_reorder(State, desc(Number))) %>%
  ggplot(aes(x = State, y = Number)) +
  geom_bar(stat = "identity", fill = "#12529B") +
  scale_x_discrete(expand = c(0.02,0)) +
  scale_y_continuous(limits = c(0, 450000), breaks = seq(0, 450000, 50000)) +
  theme_bw() +
  labs(title = "The number of accidents in each state from 2016 to 2020") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF", color = "#12529B",
                                        size = 1.5),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 0),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1))
```



We can find a state as an example to learn more details. So we need to calculate the number of accidents which occurred in every city from 2016 to 2020.

```
Accidents_city <- Location %>%
  group_by(State, City) %>%
  summarize(Number = table(City))
```

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

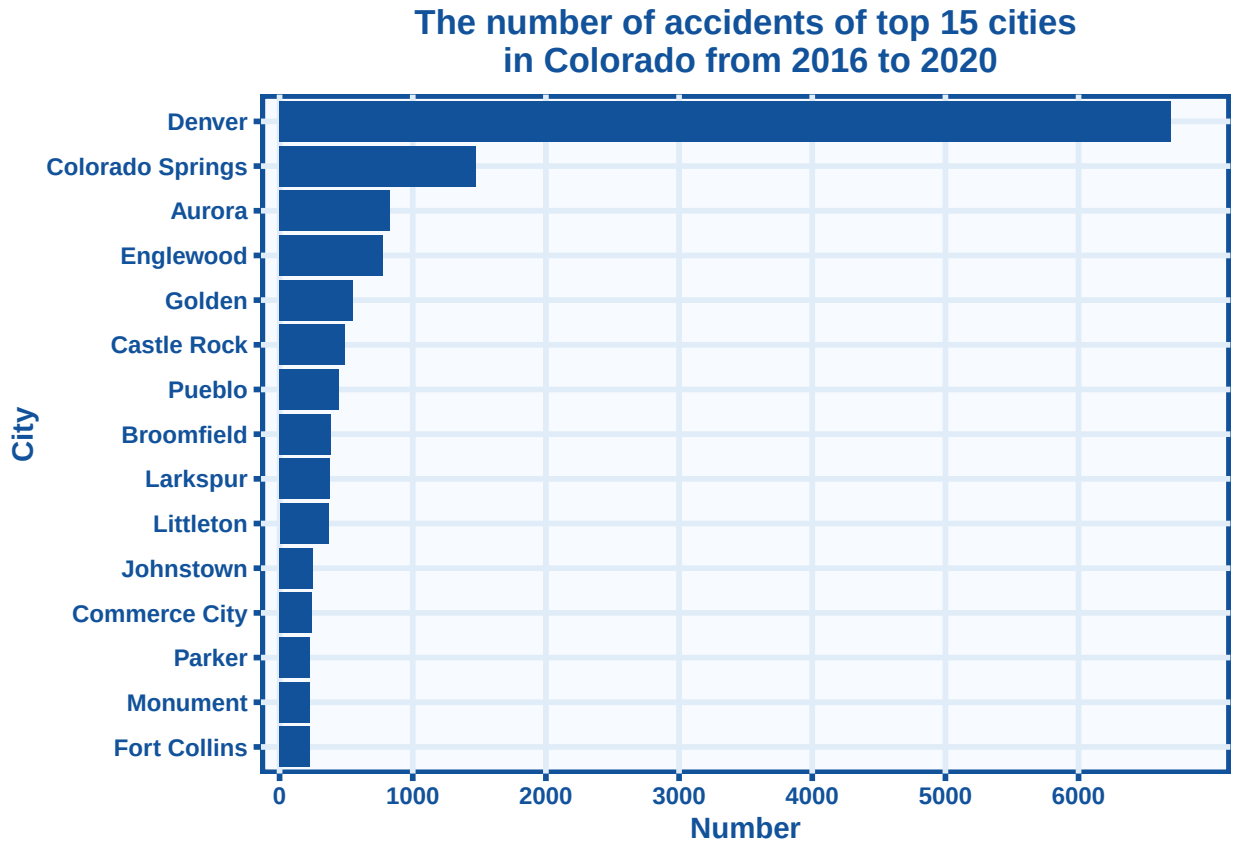
We choose Colorado as an example. Because there are so many cities, so we choose the top 15, and show them in a bar graph. Denver is the city with the most accidents. Then the city with the second and third most accidents are Colorado Springs and Aurora.

```
Accidents_colorado <- Accidents_city %>%
  filter(State == "CO") %>%
  arrange(desc(Number))

Accidents_colorado <- Accidents_colorado[1:15,]

Accidents_colorado %>%
  mutate(City = fct_reorder(City, Number)) %>%
```

```
ggplot(aes(x = City, y = Number)) +
geom_bar(stat = "identity", fill = "#12529B") +
scale_y_continuous(expand = c(0.02,0), limits = c(0, 7000),
                   breaks = seq(0, 6000, 1000)) +
labs(title = "The number of accidents of top 15 cities\n in Colorado from 2016 to 2020
theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
      panel.background = element_rect(fill = "#F7FBFF", color = "#12529B",
                                      size = 1.5),
      panel.grid = element_blank(),
      panel.grid.major = element_line(color = "#E0ECF7", size = 1),
      axis.title = element_text(color = "#12529B", face = "bold"),
      axis.text = element_text(color = "#12529B", face = "bold"),
      axis.ticks = element_line(color = "#12529B", size = 1)) +
coord_flip()
```



3.4 Weather

Let's choose the columns we need in order to see the relationship between weather and the number of accidents.

```
Weather <- US_Accidents %>%
  select(Severity, Weather_Condition) %>%
  glimpse()
```

```
## Rows: 1,516,064
## Columns: 2
## $ Severity          <int> 3, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3~
## $ Weather_Condition <chr> "Rain", "Rain", "Overcast", "Overcast", "Overcast", ~
```

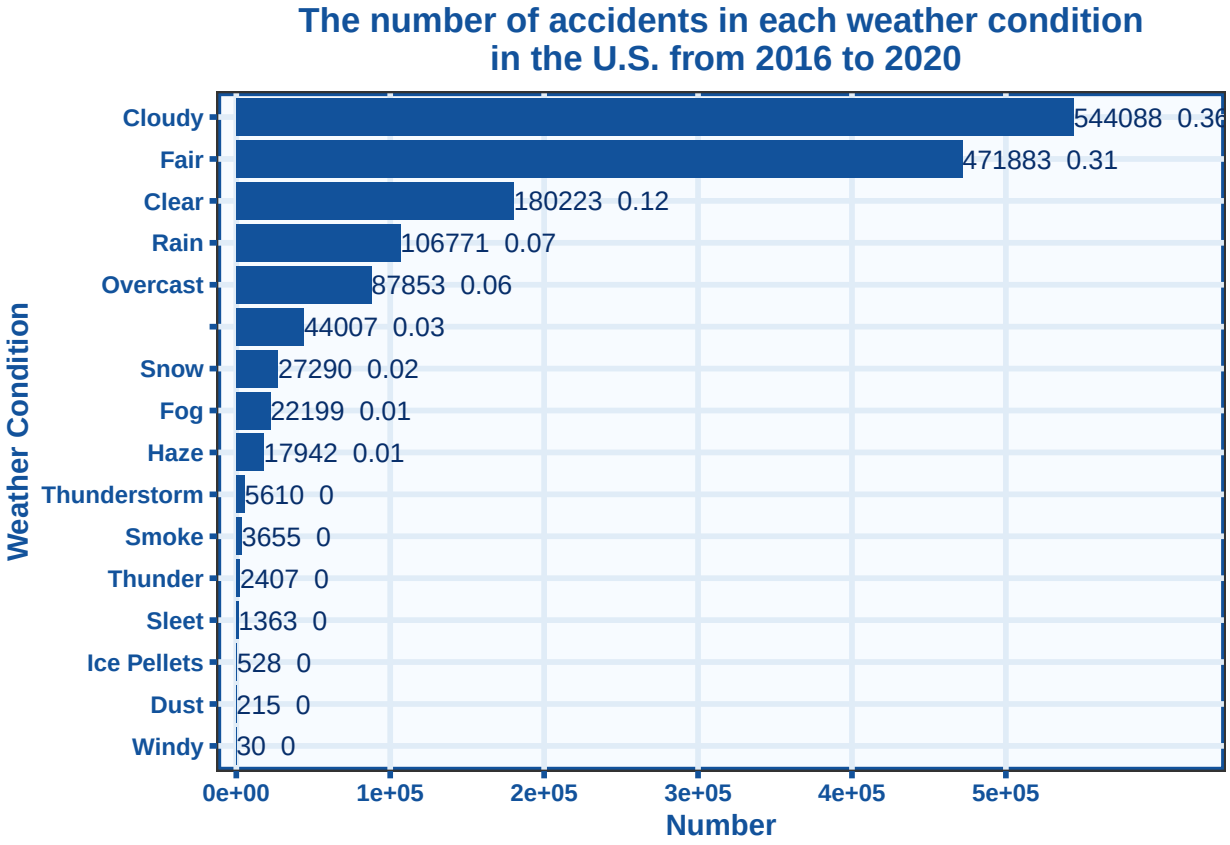
We want to see if the weather condition will effect the number of accidents, so we need to count the number of accidents occurred in every kind of weather condition.

```
Accidents_weather_condition <- Weather %>%
  select(Weather_Condition) %>%
  group_by(Weather_Condition) %>%
  summarize(Number = table(Weather_Condition))
```

Then it is time to show this data in a bar graph. We notice that almost 80% accidents happened in weather that does not affect the car. But it doesn't mean weather does not affect the occur of accidents.

```
Accidents_weather_condition %>%
  mutate(Weather_Condition = fct_reorder(Weather_Condition, Number)) %>%
  ggplot(aes(x = Weather_Condition, y = Number)) +
  geom_bar(stat = "identity", fill = "#12529B") +
  geom_text(aes(label = paste(Number, round(Number/sum(Number), digit = 2),
                              sep = "  "),
                hjust = 0, vjust = 0.5),
            color = "#072E6A", size = 3.5) +
  scale_y_continuous(expand = c(0.02,0), limits = c(0, 630000),
                     breaks = seq(0, 500000, 100000)) +
  theme_bw() +
  labs(title = "The number of accidents in each weather condition\n in the U.S. from 201
  xlab("Weather Condition") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF", color = "#12529B",
                                        size = 1.5),
        panel.grid = element_blank(),
```

```
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1)) +
  coord_flip()
```



If we want to know the specific relationship between weather condition and the number of accidents, we would better to choose a specific city and find the data about the weather condition of this city each year. And the calculate the probability of accidents happened in weather that do not affect the car and the probability of accidents happened in bad weather.

Because the weather in each day is not changeless, although we can find the dataset about everyday weather condition, we must separate the weather condition of the time accidents happened and the weather condition of the time there is nothing happened. It's a big and complex project. In my opinion, we can select a state or city that the number of accidents is very big. More accidents means more accurate about the calculation.

Therefore, we prefer to choose CA which is the state that happened the largest number of accidents in the past five years.

Firstly, we need to select the data we need.

```
Accidents_CA <- US_Accidents %>%
  filter(State == "CA") %>%
  separate(col = Start_Time, into = c("Date", "Time"), sep = " ") %>%
  separate(col = Date, into = c("Year", "Month", "Day"), sep = "-") %>%
  select(Year, Month, Day, Weather_Condition) %>%
  glimpse()
```

```
## Rows: 448,833
## Columns: 4
## $ Year              <chr> "2016", "2016", "2016", "2016", "2016", "2016", "201~
## $ Month             <chr> "03", "03", "03", "03", "03", "03", "03", "03", "03"~
## $ Day               <chr> "22", "22", "22", "22", "22", "22", "23", "23", "23"~
## $ Weather_Condition <chr> "Clear", "Cloudy", "Cloudy", "Clear", "Clear", "Clea~
```

Then, we assume that there is no repetition about the weather condition in a day. Let's count the number of accidents in each weather condition in each day.

```
Accidents_CA <- Accidents_CA %>%
  group_by(Year, Month, Day, Weather_Condition) %>%
  summarize(Number = table(Weather_Condition)) %>%
  glimpse()
```

```
## `summarise()` has grouped output by 'Year', 'Month', 'Day'. You can override using th
```

```
## Rows: 7,997
## Columns: 5
## Groups: Year, Month, Day [1,645]
## $ Year              <chr> "2016", "2016", "2016", "2016", "2016", "2016", "201~
## $ Month             <chr> "03", "03", "03", "03", "03", "03", "03", "03", "03"~
## $ Day               <chr> "22", "22", "23", "23", "23", "23", "24", "24", "24"~
## $ Weather_Condition <chr> "Clear", "Cloudy", "", "Clear", "Cloudy", "Fair", ""~
## $ Number            <table> <table[26]>
```

Thirdly, it is time to calculate the number of weather conditions appear in these days.

```
Weather_Condition_summary <- Accidents_CA %>%
  group_by(Weather_Condition) %>%
  summarize(Number = table(Weather_Condition)) %>%
  glimpse()
```

```
## Rows: 16
## Columns: 2
## $ Weather_Condition <chr> "", "Clear", "Cloudy", "Dust", "Fair", "Fog", "Haze"~
## $ Number            <table> <table[16]>
```

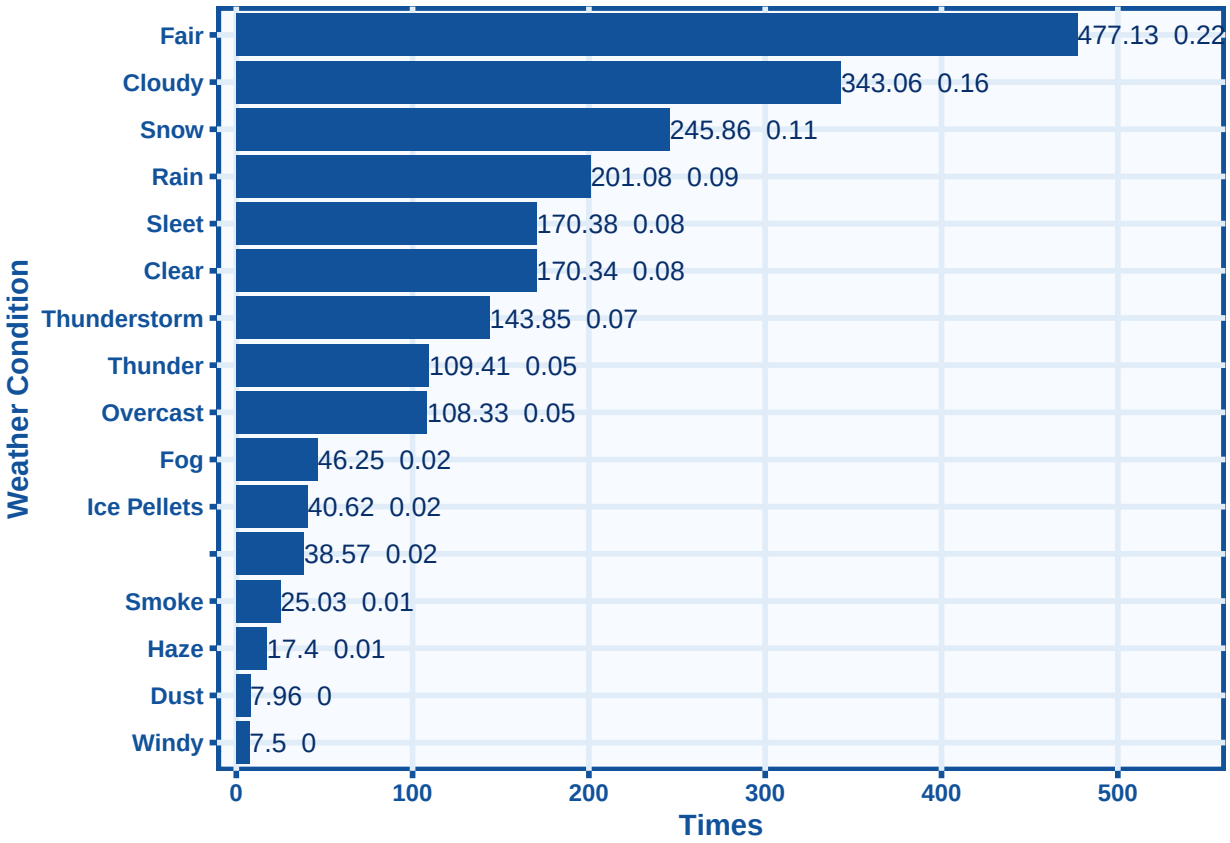
Fourthly, because we have already acquire the total number of accidents in each weather condition before(in Accidents_weather_condition). So, we can try to compute the number of accidents in each weather condition(the total number of accidents in each weather condition is divided by the number of weather conditions appear in these days).

```
Weather_probability <- merge(x = Weather_Condition_summary,
                             y = Accidents_weather_condition,
                             by = "Weather_Condition", all = TRUE)

Weather_probability <- Weather_probability %>%
  group_by(Weather_Condition) %>%
  summarize(Times = round(Number.y / Number.x, 2)) %>%
  mutate(Probability = round(Times / sum(Times), 2))
```

According to the result we get, we can use bar graph to show the number of accidents occurred in each weather condition every time.

```
Weather_probability %>%
  mutate(Weather_Condition = fct_reorder(Weather_Condition, Times)) %>%
  ggplot(aes(x = Weather_Condition, y = Times)) +
  geom_bar(stat = "identity", fill = "#12529B") +
  geom_text(aes(label = paste(Times, round(Probability / sum(Probability),
                                           digit = 2)
                              , sep = "  "),
                hjust = 0, vjust = 0.5),
            color = "#072E6A", size = 3.5) +
  scale_y_continuous(expand = c(0.02,0), limits = c(0, 550),
                     breaks = seq(0, 500, 100)) +
  labs(x = "Weather Condition") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF", color = "#12529B",
                                        size = 1.5),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1)) +
  coord_flip()
```



We also need the probability of the number of weather conditions appeared.

```
Weather_Condition_summary <- Weather_Condition_summary %>%
  mutate(Probability = round(Number / sum(Number), 2))
```

Then, we can merge this two dataframes into one dataframe. Probability.x means the probability of the number of accidents in each weather condition. Probability.y means the probability of the number of each weather condition.

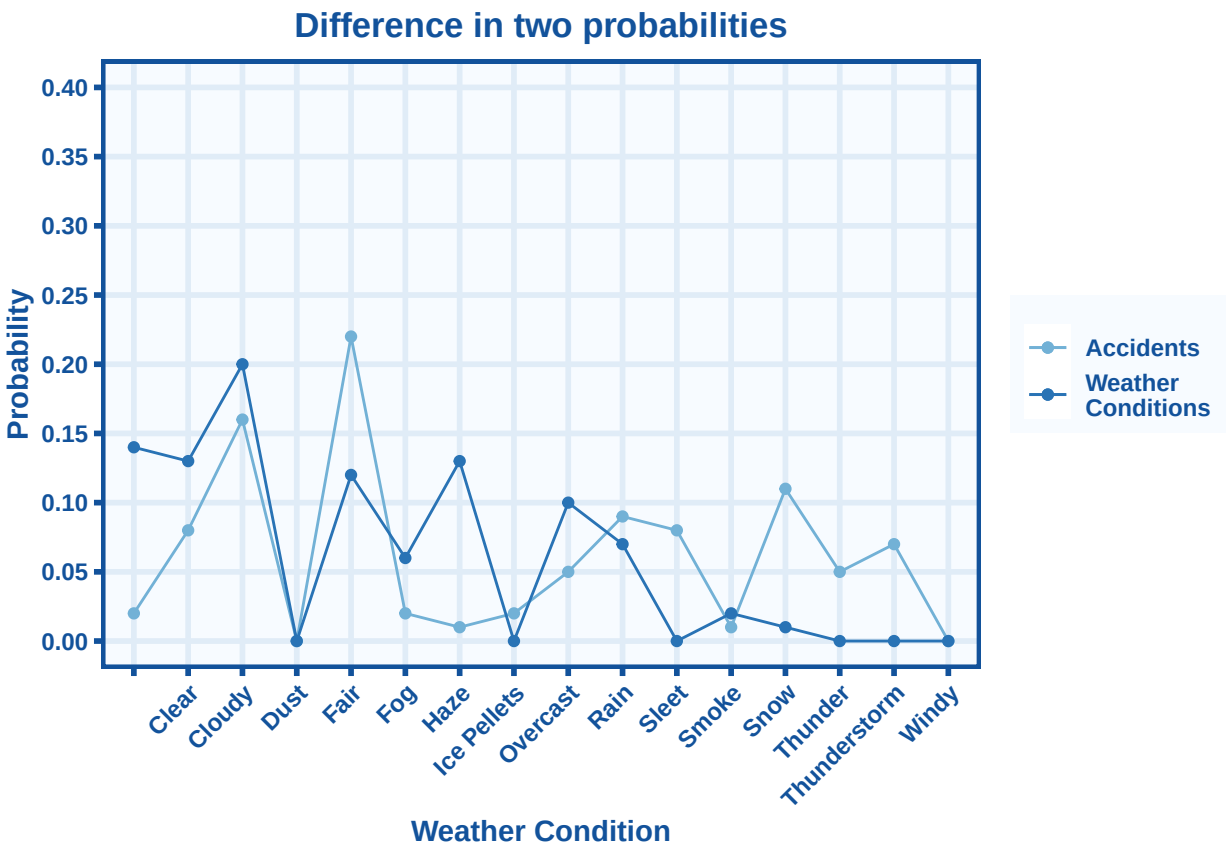
```
Weather_difference <- merge(x = Weather_probability,
                            y = Weather_Condition_summary,
                            by = "Weather_Condition", all = TRUE)

Weather_difference <- Weather_difference %>%
  select(Weather_Condition, Probability.x, Probability.y)
```

Let's use these two probabilities to see the difference, and show them in line chart.

```
Weather_difference %>%
  ggplot(aes(x = Weather_Condition, y = Probability.x, group = 1)) +
  geom_line(aes(color = "Probability.x")) +
```

```
  geom_point(aes(color = "Probability.x")) +
  geom_line(aes(y = Probability.y, color = "Probability.y")) +
  geom_point(aes(y = Probability.y, color = "Probability.y")) +
  scale_y_continuous(limits = c(0, 0.4), breaks = seq(0, 0.4, 0.05)) +
  scale_colour_manual(labels = c("Accidents", "Weather\nConditions"),
                      values = c("#72B1D6", "#2973B5")) +
  theme_bw() +
  labs(x = "Weather Condition", y = "Probability",
       title = "Difference in two probabilities") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.ticks = element_line(color = "#12529B", size = 1),
        legend.title = element_blank(),
        legend.text = element_text(color = "#12529B", face = "bold"),
        legend.background = element_rect(fill = "#F7FBFF"))
```



If there are some weather conditions that the probability of the number of accidents in each weather condition is larger than the probability of the number of each weather condition, we can say that these weather conditions are easier to cause the accidents to occur. According to the graph above, we find that fair, ice pellets, rain, sleet, snow, thunder and thunderstorm are the weather conditions that the probability of the number of accidents in each weather condition is larger than the probability of the number of each weather condition.

In the conclusion, we can say that weather condition is a kind of factor that could cause accident occur.

Indeed, there are something that confuse me, it is why the probability of the number of accidents in fog and haze is less than the probability of the number of fog, haze and smoke. If we want to obtain an accurate conclusion we still need to find the relationship between visibility and the probability of accidents occurred in fog, haze and smoke.

3.5 COVID-19

Because we want to find if the accidents increase in 2020 is related to COVID-19, we need to select all of the data in 2020. And then, we also need to calculate the number of cases in each day in 2020.

```
US_Cases_2020 <- US_Cases %>%
  separate(col = "Date", into = c("Year", "Month", "Day"), sep = "-") %>%
  group_by(Year, Month, Day) %>%
  summarize(Cases = sum(Cases)) %>%
  filter(Year == "2020") %>%
  unite(col = "Month_Day", Month, Day, sep = "") %>%
  ungroup() %>%
  select(-Year)
```

```
## `summarise()` has grouped output by 'Year', 'Month'. You can override using the `.gro
```

Then, we need to select the rows which Year are equal to 2020 in dataframe Accidents_per_day which store the data the number of accidents occurred every day from 2016 to 2020.

```
Accidents_per_day_2020 <- Accidents_per_day %>%
  filter(Year == "2020") %>%
  select(-Year)
```

Let's join US_Cases_2020 and Accidents_per_day_2020 together. Of course, there are some null values, because the first case of COVID-19 was appeared in January 22th.

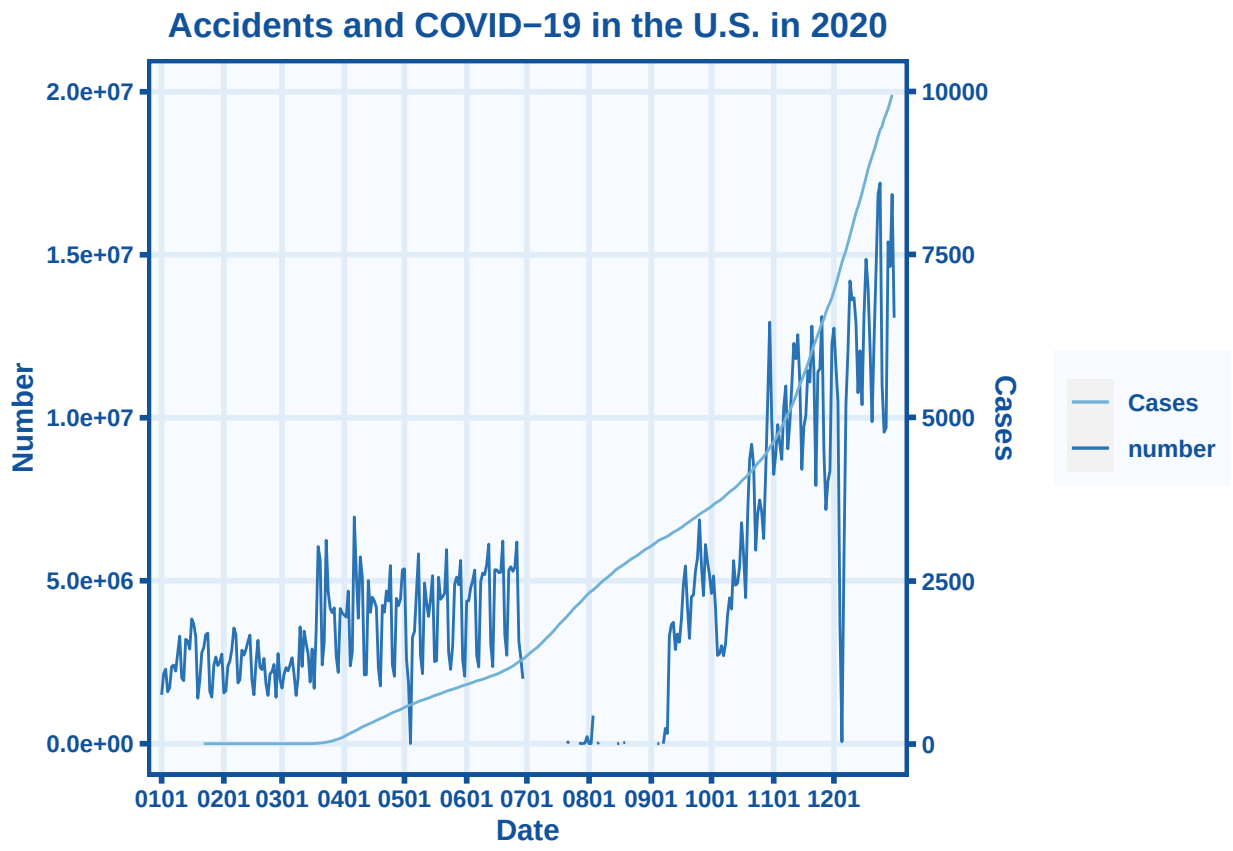
```
Accidents_Cases<- merge(x = Accidents_per_day_2020, y = US_Cases_2020,
                        by = "Month_Day", all = TRUE)
```

It's time to create a graph to test if these two statistics have the same tendency.

```
Accidents_Cases %>%
  ggplot(aes(x = Month_Day, y = number * 2000, group = 1)) +
  geom_line(aes(color = "number")) +
  geom_line(aes(y = Cases, color = "Cases")) +
  scale_x_discrete(expand = c(0.02, 0), breaks = c("0101", "0201", "0301",
                                                   "0401", "0501", "0601",
                                                   "0701", "0801", "0901",
                                                   "1001", "1101", "1201")) +
  scale_y_continuous(limits = c(0, 20000000),
                     breaks = seq(0, 20000000, 5000000),
```

```
                    sec.axis = sec_axis(~./2000, name = "Cases")) +
  scale_colour_manual(values = c("#72B1D6", "#2973B5")) +
  labs(x = "Date", y = "Number", title = "Accidents and COVID-19 in the U.S. in 2020") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", fill = "transparent",
                                    size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1),
        legend.title = element_blank(),
        legend.text = element_text(color = "#12529B", face = "bold"),
        legend.background = element_rect(fill = "#F7FBFF"))
```

```
## Warning: Removed 22 row(s) containing missing values (geom_path).
```



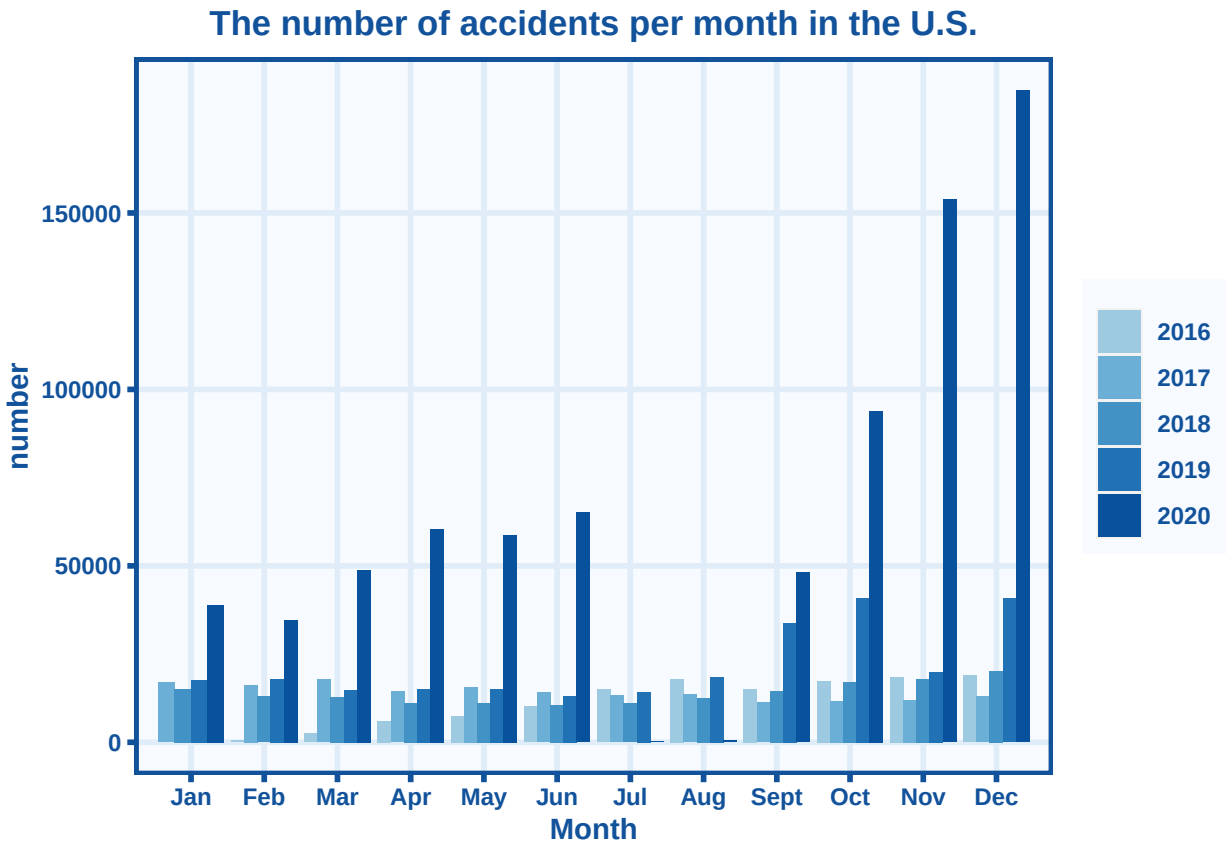
According to the plot above, we can see that the number of accidents does not increase dramatically in the months that the cases of COVID-19 did not increase rapidly. But when

the number of accidents increase dramatically, in the same time(in October, November and December), the cases of COVID-19 also increase rapidly.

In order to make a conclusion, we also need to know the tendency of the number of accidents in these five years(2016-2020).

```
Accidents_per_month <- Accidents_per_day %>%
  separate(col = "Month_Day", into = c("Month", "Day"), sep = 2) %>%
  unite(col = "Year_Month", Year, Month, sep = "") %>%
  select(-Day) %>%
  group_by(Year_Month) %>%
  summarize(number = sum(number)) %>%
  separate(col = "Year_Month", into = c("Year", "Month"), sep = 4)

Accidents_per_month %>%
  ggplot(aes(x = Month, y = number, fill = Year)) +
  geom_bar(stat = "identity", position = "dodge") +
  #scale_fill_brewer(palette = "Blues") +
  scale_fill_manual(values = brewer.pal(9, "Blues")[4:8]) +
  scale_x_discrete(expand = c(0.07, 0), labels = c("Jan", "Feb", "Mar", "Apr",
                                                   "May", "Jun", "Jul", "Aug",
                                                   "Sept", "Oct", "Nov", "Dec")) +
  labs(title = "The number of accidents per month in the U.S.") +
  theme(plot.title = element_text(hjust = 0.5, color = "#12529B", face = "bold"),
        panel.background = element_rect(fill = "#F7FBFF"),
        panel.grid = element_blank(),
        panel.grid.major = element_line(color = "#E0ECF7", size = 1),
        panel.border = element_rect(color = "#12529B", fill = "transparent",
                                    size = 1.5),
        axis.title = element_text(color = "#12529B", face = "bold"),
        axis.text = element_text(color = "#12529B", face = "bold"),
        axis.ticks = element_line(color = "#12529B", size = 1),
        legend.title = element_blank(),
        legend.text = element_text(color = "#12529B", face = "bold"),
        legend.background = element_rect(fill = "#F7FBFF"))
```



We can know that the number of accidents did not increase so fast between 2016 to 2019. And except the months(July, August and September) that do not have the whole data, we have reason to make a conclusion that COVID-19 had affect the number of accidents.