US Accident Dataset: Apurva Shekhar, ashekhar@scu.edu

The dataset comprise of 2.25 million records of accidents on road across all

state in the USA. I have tried to find answers for the following questions -

1. Accidents trend over Years

2. Accident By Weekday

3. Top States By Accident

4. Top 10 states with most severe accidents

5. Accidents by Severity

6. Number of Accidents By Time of Day In Caifornia

7. Dominant Factors For Accidents In Each County

```
# Load dataset
# Download - https://www.kaggle.com/sobhanmoosavi/us-accidents
setwd("/Users/apurvashekhar/Desktop/SCU/Quarter_1/R")
accident_data <- read.csv("US_Accidents_May19.csv", stringsAsFactors = FALSE)

# Required libraries.
if (!require(ggplot2)) install.packages("ggplot2")

## Loading required package: ggplot2

if (!require(stringr)) install.packages("stringr")

## Loading required package: stringr

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(stringr)
```

Parsed WeatherTimestamp in accident data to convert it to POSIXct time.

Also, extracted date, day of week, hour, month and year as seperate columns from

ParsedDateTime column for future use.

```
accident_data$ParsedDateTime <-
as.POSIXct(strptime(accident_data$Weather_Timestamp, format="%Y-%m-%d
%H:%M:%S"))
accident_data$Date <- format(accident_data$ParsedDateTime, format="%Y-%m-%d")
accident_data$Day <- format(accident_data$ParsedDateTime, format="%A")
accident_data$Hour <- format(accident_data$ParsedDateTime, format="%H")
accident_data$Month <- format(accident_data$ParsedDateTime, format="%m")
accident_data$Year <- format(accident_data$ParsedDateTime, format="%Y")
```

Parsed boolean attribute columns as boolean values.

```
accident_data$Amenity_bool = as.logical(accident_data$Amenity)
accident_data$Bump_bool = as.logical(accident_data$Bump)
accident_data$Crossing_bool = as.logical(accident_data$Crossing)
accident_data$GiveWay_bool = as.logical(accident_data$Give_Way)
accident_data$Junction_bool = as.logical(accident_data$Junction)
accident_data$NoExit_bool = as.logical(accident_data$No_Exit)
accident_data$Railway_bool = as.logical(accident_data$Railway)
accident_data$Roundabout_bool = as.logical(accident_data$Roundabout)
accident_data$Station_bool = as.logical(accident_data$Station)
accident_data$Stop_bool = as.logical(accident_data$Stop)
accident_data$TrafficCalming_bool = as.logical(accident_data$Traffic_Calming)
accident_data$TrafficSignal_bool = as.logical(accident_data$Traffic_Signal)
accident_data$TurningLoop_bool = as.logical(accident_data$Turning_Loop)
```

Accidents trend over Years: This graph shows the trend of accident occurrences.

The count of accidents is grouped on (Month, Year). The dots in the graph represents

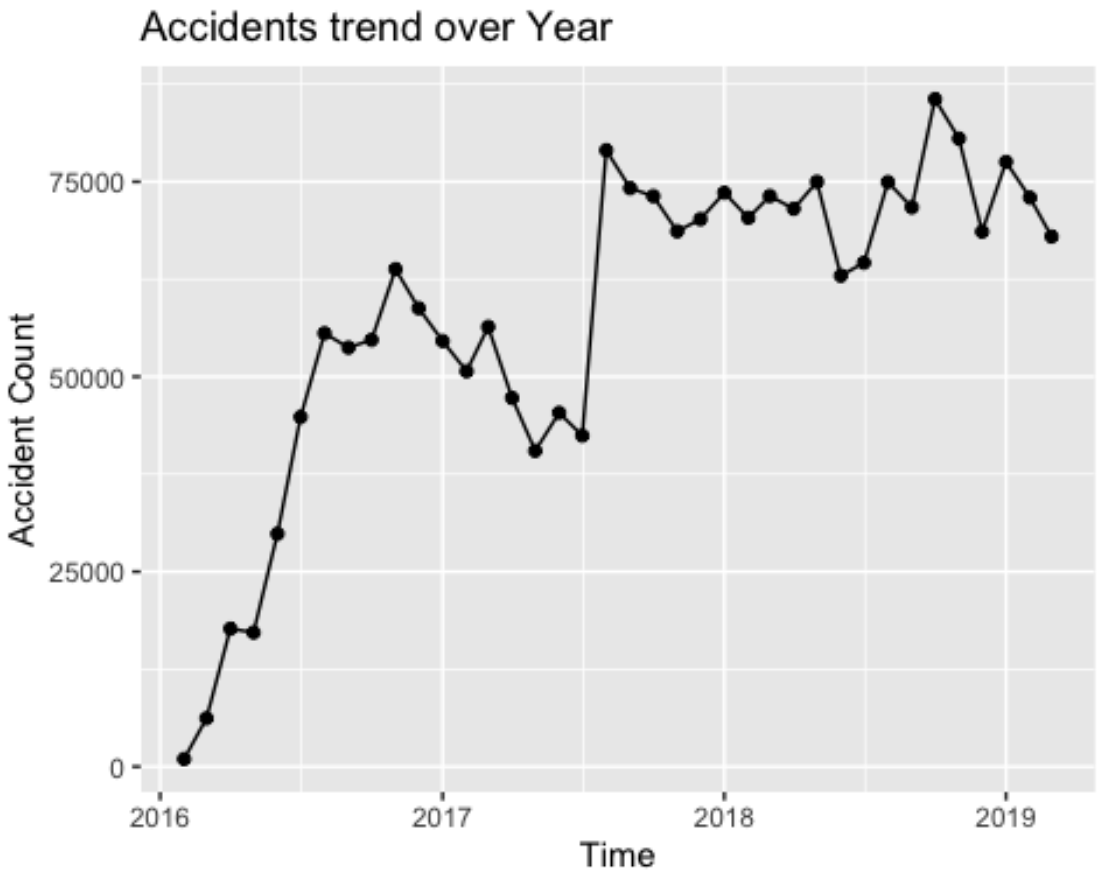
the months of that particular year.I observed that accident rates increased by approximately

1.5x between 2016 and 2017 across all states.

```
# Grouped by state and counted the number of accidents in each state.
accidents_by_date <- accident_data %>%
  filter(!is.na(ParsedDateTime)) %>%
  group_by(Month = as.POSIXct(paste0(strftime(ParsedDateTime,format="%Y-
%m"),"-01")) ) %>%
  summarise(Count = n())

## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(data=(accidents_by_date), aes(x = Month, y = Count, group = 1)) +
  geom_line() +
  geom_point() +
  ggtitle("Accidents trend over Year") +
  labs(x= "Time", y = "Accident Count")
```



Accident By Weekday: The plot counts number of accidents by weekdays.

The count has been calculated by grouping on Day of Weeks of accident.

I observed that number of accidents are much less on weekends. This may

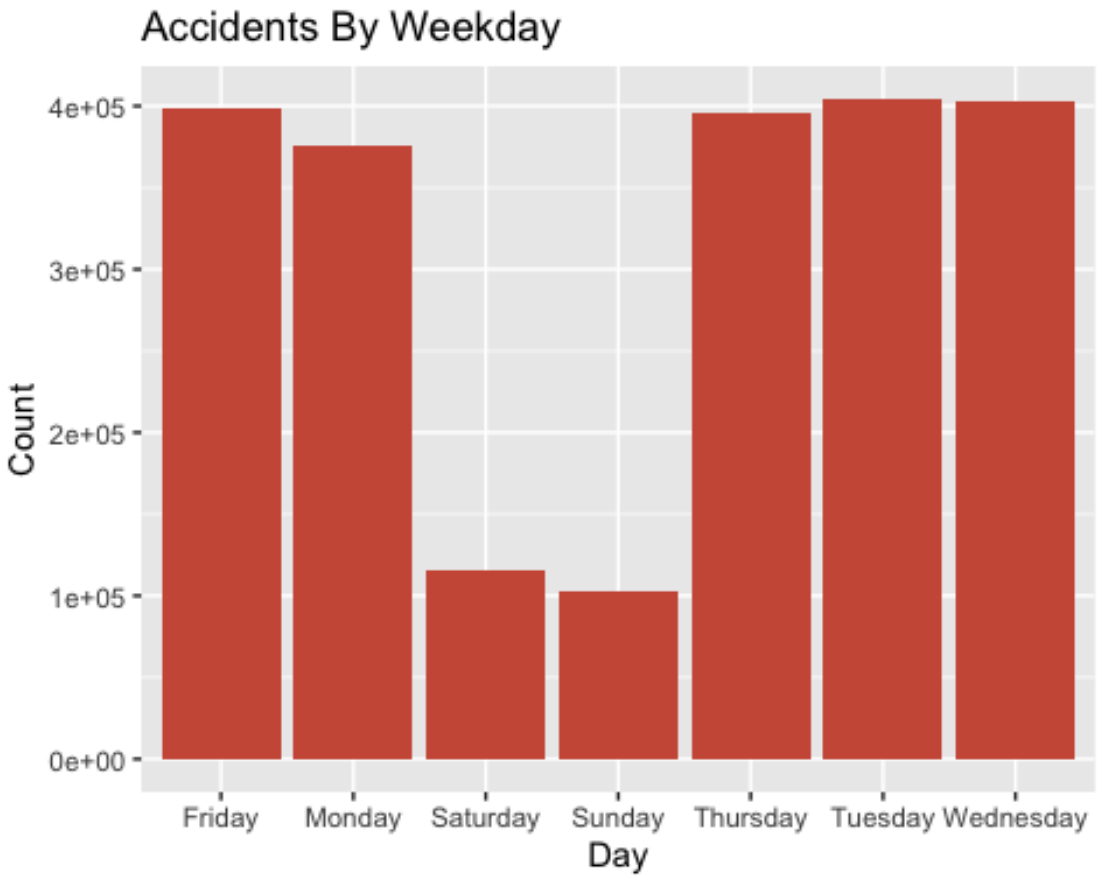
be attributed to fewer number of cars on the roads due to offices being

closed.

```
# Grouped by state and counted the number of accidents in each state.
accidents_by_weekday <- accident_data %>%
  filter(!is.na(Day)) %>%
  group_by(Day) %>%
  summarise(Count = n())

## `summarise()` ungrouping output (override with `.groups` argument)

# Plot the result.
ggplot(data=(accidents_by_weekday), aes(x = Day, y=Count)) +
  geom_bar(stat = "identity", fill="coral3") +
  ggtitle("Accidents By Weekday")
```



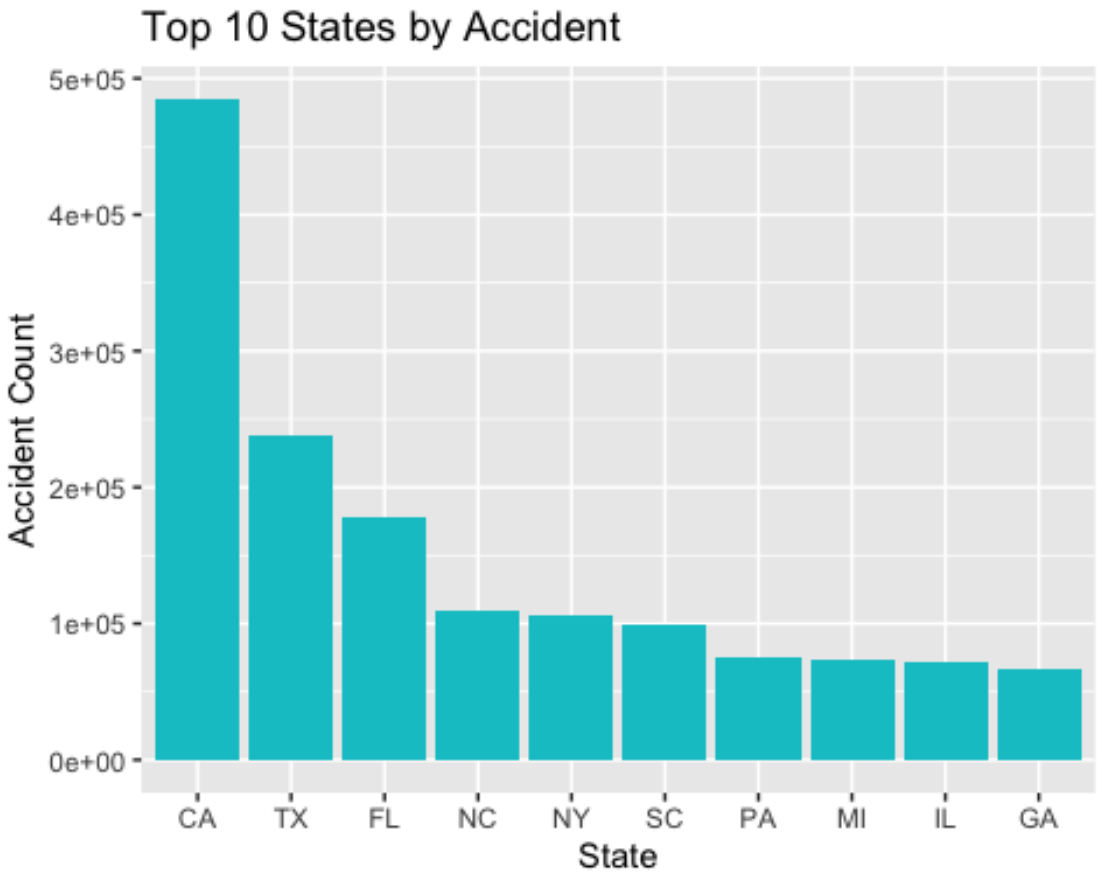
Top States By Accident: This table and graph shows the top 10 states that have the most number of accidents. The table points out that California leads the pack with almost double the number of accidents of Texas, that stands second. The interesting fact is that even though Texas has higher speed limit i.e 85 mph on highways, it is still behind California in the count of accidents. This could be attributed to the fact that California has much higher density of cars.

```
# Grouped by state and counted the number of accidents in each state.
top_states_by_accident <- accident_data %>%
                          group_by(State) %>%
                          summarise(Count = n()) %>%
                          ungroup() %>%
                          arrange(desc(Count)) %>%
                          top_n(10, Count)

## `summarise()` ungrouping output (override with `.groups` argument)

# Bar Graph Plot.
ggplot(data=top_states_by_accident, aes(x= reorder(State, -Count), y=Count))
+
```

```
  geom_bar(stat = "identity", fill="turquoise3") +
  ggtitle("Top 10 States by Accident") +
  labs(x= "State", y = "Accident Count")
```




```
# Table View.
top_states_by_accident
```

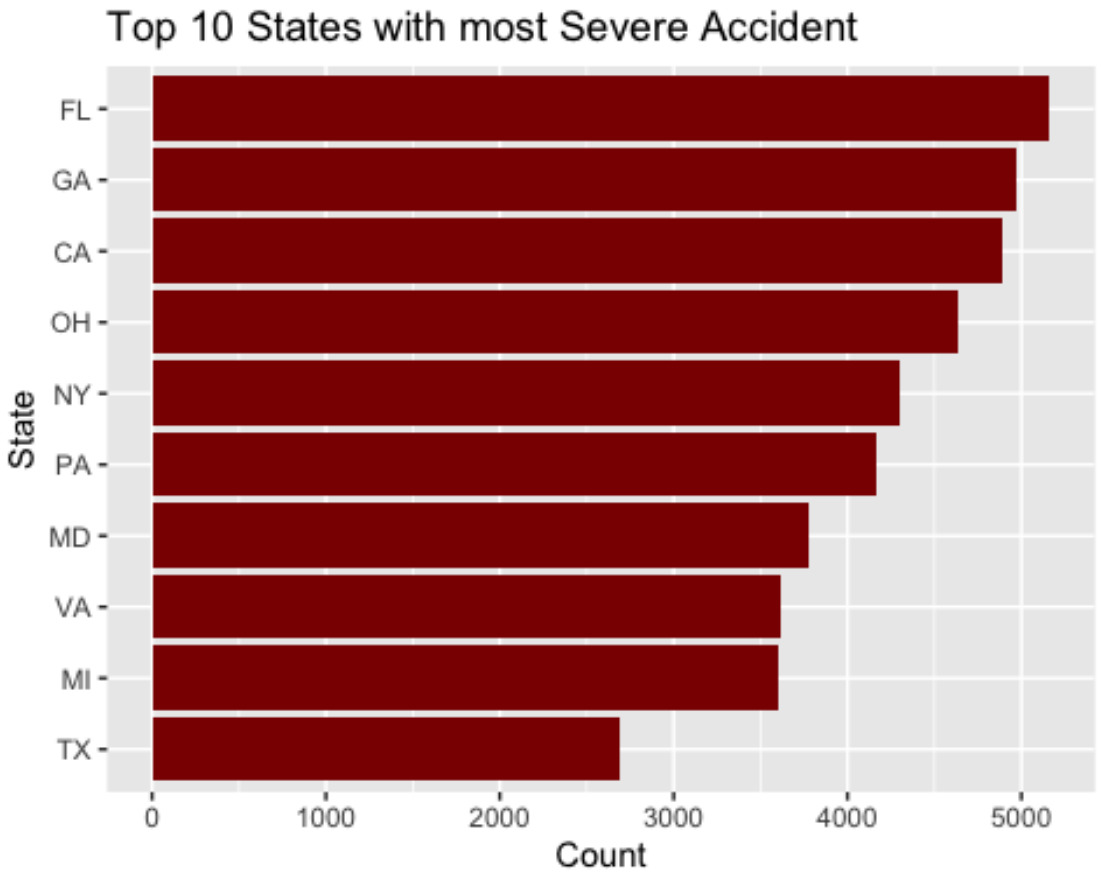
```
## # A tibble: 10 x 2
##    State  Count
##    <chr>  <int>
##  1 CA    484706
##  2 TX    237637
##  3 FL    177490
##  4 NC    108916
##  5 NY    105523
##  6 SC     99890
##  7 PA     75814
##  8 MI     74045
##  9 IL     71701
## 10 GA     66637
```

Top 10 states with most severe accidents: The table and plot below shows the top ten states that have the most severe accidents. Even though California has almost 3x number of accidents than Florida, Florida beats California in occurences of severe accidents.

```
# Filtered the dataset to keep most severe accidents i.e accidents with
severity 4.
severe_accidents <- filter(accident_data, Severity == max(Severity))
top_states_by_severe_accidents <- severe_accidents %>%
                        group_by(State) %>%
                        summarise(Count = n()) %>%
                        ungroup() %>%
                        arrange(desc(Count)) %>%
                        top_n(10, Count)

## `summarise()` ungrouping output (override with `.groups` argument)

# Bar graph plot.
ggplot(data=top_states_by_severe_accidents,
       aes(x= reorder(State, Count), y=Count)) +
  geom_bar(stat = "identity", fill="darkred") +
  ggtitle("Top 10 States with most Severe Accident") +
  coord_flip() +
  labs(x="State", y= "Count")
```




```
# Table format.
top_states_by_severe_accidents
```

```
## # A tibble: 10 x 2
##    State Count
##    <chr> <int>
##  1 FL     5165
##  2 GA     4968
##  3 CA     4891
##  4 OH     4634
##  5 NY     4300
##  6 PA     4165
##  7 MD     3771
##  8 VA     3621
##  9 MI     3607
## 10 TX     2689
```

Accidents by Severity: The plot below shows the distribution of accident severity

across the top three states by number of accidents. This graph demonstrates the

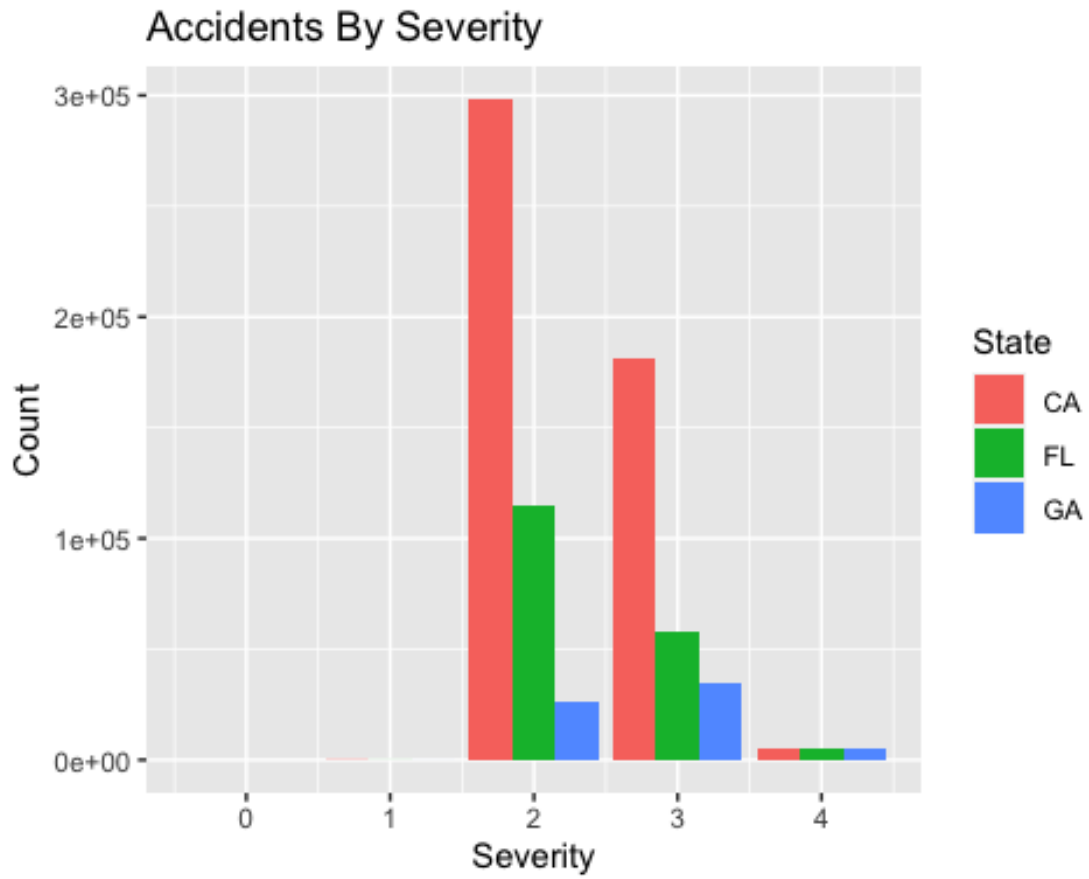
how common are severe accidents in California, Florida and Georgia.

```
# Group accidents by severity to understand frequency of each severity -
Compared three
```

```r
# categories i.e States.
accidents_by_severity <- accident_data %>%
                         filter(State == "CA" | State == "FL" | State == "GA")
%>%
                         group_by(Severity, State) %>%
                         summarise(Count = n());

## `summarise()` regrouping output by 'Severity' (override with `.groups`
argument)

# Plotting the result.
ggplot(data=accidents_by_severity, aes(x= Severity, y=Count, fill = State)) +
  geom_bar(stat = "identity", position = 'dodge') +
  ggtitle("Accidents By Severity") +
  labs(x="Severity", y= "Count")
```



Number of Accidents By Time of Day In Caifornia: I dived deep into the state that has the highest

number of accidents. Dividing the data by the hour of the day, it is indicated that most number of

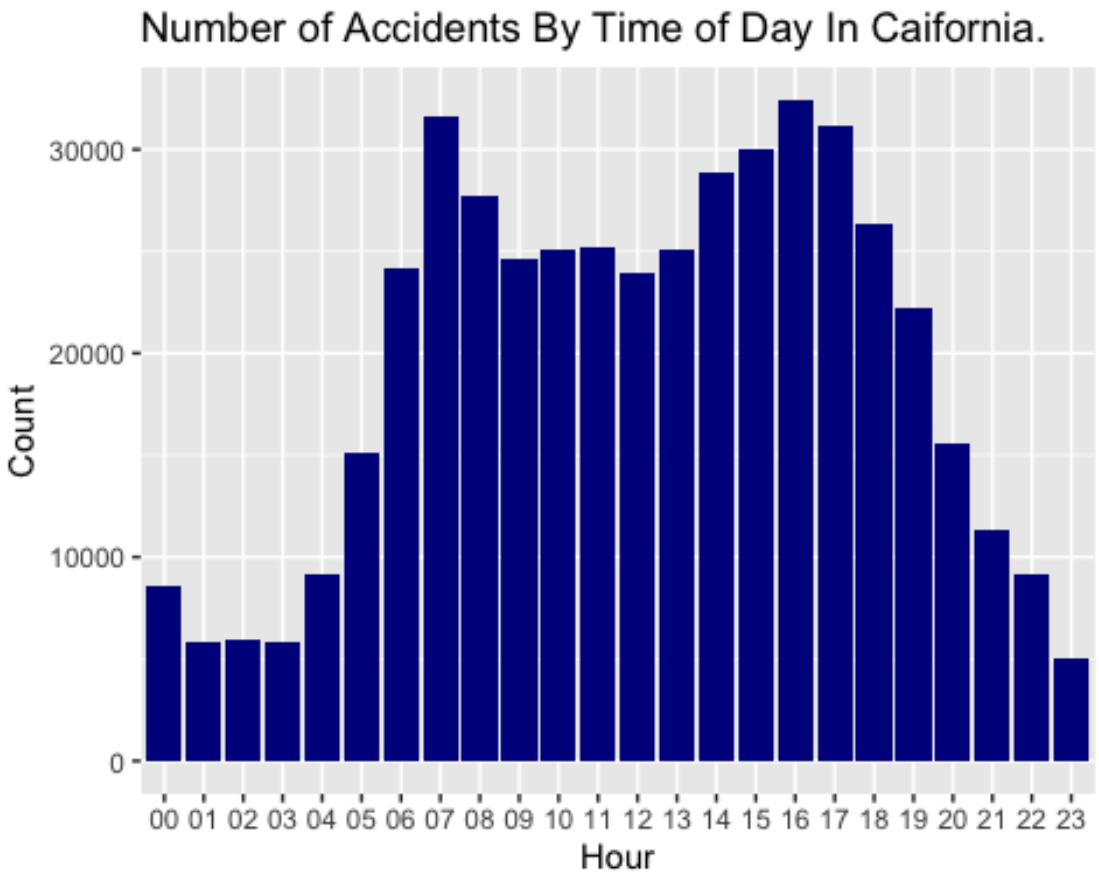
accidents occur at 7am and 4pm. This reason behind this trend could be that those hours are

when people travel to/off from work. Due to higher number of cars, the count of accident soars.

```
california_accidents <- filter(accident_data, State == "CA")
# Grouped by state and counted the number of accidents in each state.
california_accident_by_hour <- california_accidents %>%
  filter(!is.na(Hour)) %>%
  group_by(Hour) %>%
  summarise(Count= n())

## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(data=(california_accident_by_hour), aes(x = Hour, y = Count)) +
  geom_bar(stat = "identity", fill = "darkblue") +
  ggtitle("Number of Accidents By Time of Day In Caifornia.")
```



Dominant Factors For Accidents In Each County: The heatmap below shows the dominant factors

for accidents across counties of California. For example, San Mateo sees most number of accidents

near roundabouts, whereas San Francisco has most number of accident near traffic calming locations.

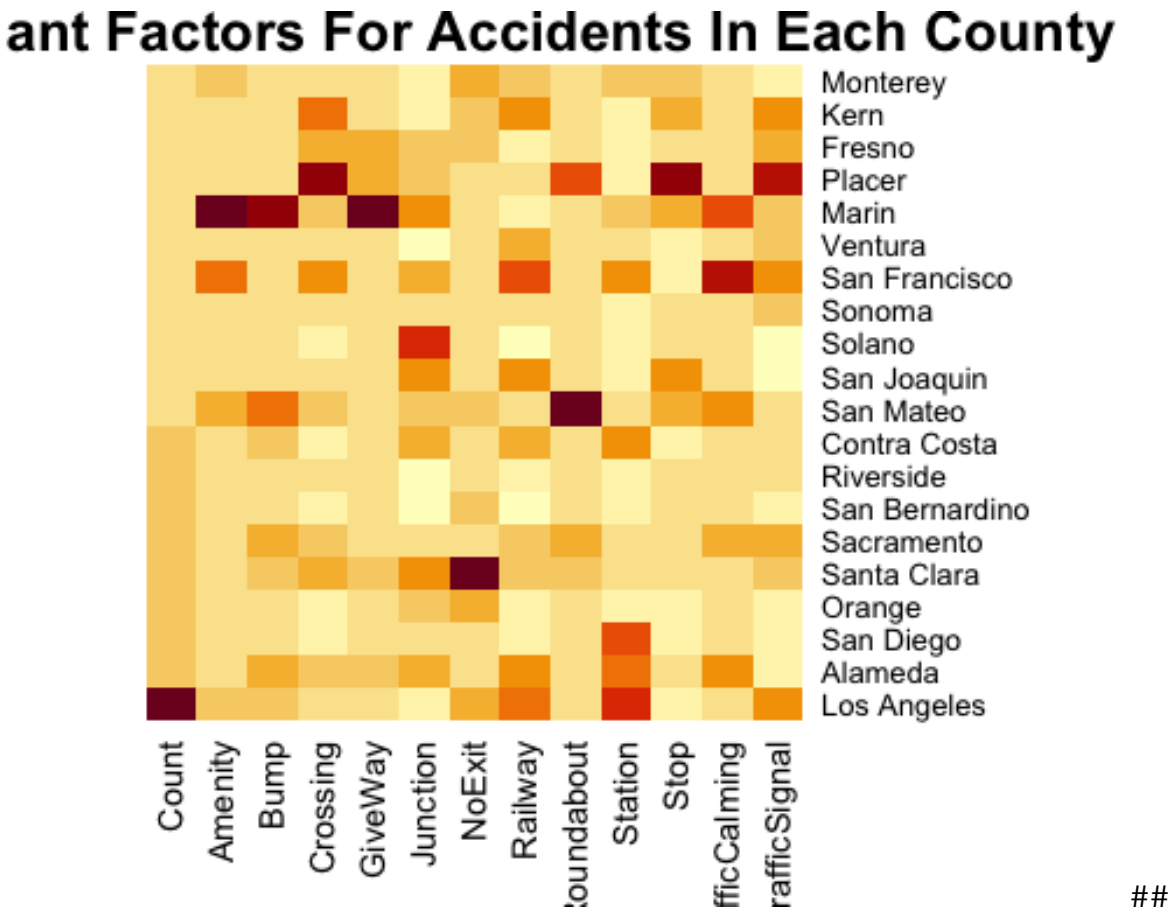
We can conclude that some of the regions for example, Marin has most number of accidents and can

induce traffic measures to make it safer to drive.

```
california_accident_by_county <- california_accidents %>%
  group_by(County) %>%
  summarise(Count = n(), Amenity = mean(Amenity_bool), Bump =
mean(Bump_bool), Crossing = mean(Crossing_bool),
            GiveWay = mean(GiveWay_bool), Junction = mean(Junction_bool),
NoExit = mean(NoExit_bool),
            Railway = mean(Railway_bool), Roundabout = mean(Roundabout_bool),
Station = mean(Station_bool),
            Stop = mean(Stop_bool), TrafficCalming =
mean(TrafficCalming_bool), TrafficSignal = mean(TrafficSignal_bool)) %>%
  ungroup() %>%
  arrange(desc(Count)) %>%
  top_n(20, Count)

## `summarise()` ungrouping output (override with `.groups` argument)

heat_matrix = as.matrix(california_accident_by_county[2:14])
row.names(heat_matrix) <- t(california_accident_by_county[1])
heatmap(heat_matrix, Rowv = NA, Colv = NA, scale = "column", main = "Dominant
Factors For Accidents In Each County")
```



##

Conclusion : ### 1. We can see that there has been a increasing trend in number of accidents in the past three years. ### 2. The number of accidents are more on weekdays than on weekends. This may be due to the less number of cars on ### raods during weekdays due to office hours. ### 3. California has higher number of accidents than Texas. Even though Texas has high probabaility of accident due to ### higher speed limit. California might have higher number of accidents due to higher density of cars. ### 4. Number of Accidents increases in California during the office hours i.e. 7am and 4-5pm. ### 5. High number of accidents occur in Marin County of California at several bumps and give way. May be some measures could be taken ### to make the county safer.