Crash_Patterns_Q3

Jack Francis

3/19/2020

The identification of crash patterns is important for policy makers to assist drivers in avoiding dangerous behavior and driving times. Crash patterns can include the time of the crash (i.e. month, week, day, hour), weather conditions, type of road, and type of vehicle. We first investigate individual factors and then hypothesize some potential combinations of factors that may be related in identifying crash patterns

```
require(tidyverse)
```

```
## Loading required package: tidyverse

## -- Attaching packages -----------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts -------------------------------------------------------------------------------- tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
require(lubridate)
```

```
## Loading required package: lubridate

##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```
require(nycflights13)
```

```
## Loading required package: nycflights13
```

```
require(hms)
```

```
## Loading required package: hms

##
## Attaching package: 'hms'

## The following object is masked from 'package:lubridate':
##
##     hms
```

```
require(stringr)
require(forcats)
require(fs)
```

```
## Loading required package: fs
```

```r
library(ggplot2)
library(tidyverse)
library(sf)
```

```
## Linking to GEOS 3.7.2, GDAL 2.4.2, PROJ 5.2.0
```

```r
library(readxl)
```

Initial Thoughts

- There is probably a strong correlation between the time of day and crashes. Most miles are driven around rush hour during the week, while the most dangerous miles are driven at night on the Weekends, due to drunk/drugged driving.
- In general, it will be interesting to investigate the most likely times for crashes to occur during the day and year
- We examine the effect of body type, but only in relation to other factors, since the direct effect is asked about in Question 5
- Does the type of road have an impact on fatal crashes? Initial guess is that more lanes would lead to more fatal crashes, because these roads generally have higher speeds
- How does weather affect the US? Generally we expect more crashes in snow/rain/fog, but would be interesting to look at how this effect varies by state (i.e. are Southern drivers significantly worse at driving in snow)
- Curious if there is a relationship between the number of accidents on a given type of road and the time of day
- Expect to see spikes in accidents near holidays

First, read in the data including our external data sources giving information about the total miles of road and state population/area. Next, take a few columns from the Accident table and join with the vehicle table.

```r
ACC_df <- read_csv("../FARS_Data/FARS2018NationalCSV/ACCIDENT.csv")
VEH_df <- read_csv("../FARS_Data/FARS2018NationalCSV/VEHICLE.csv")
```

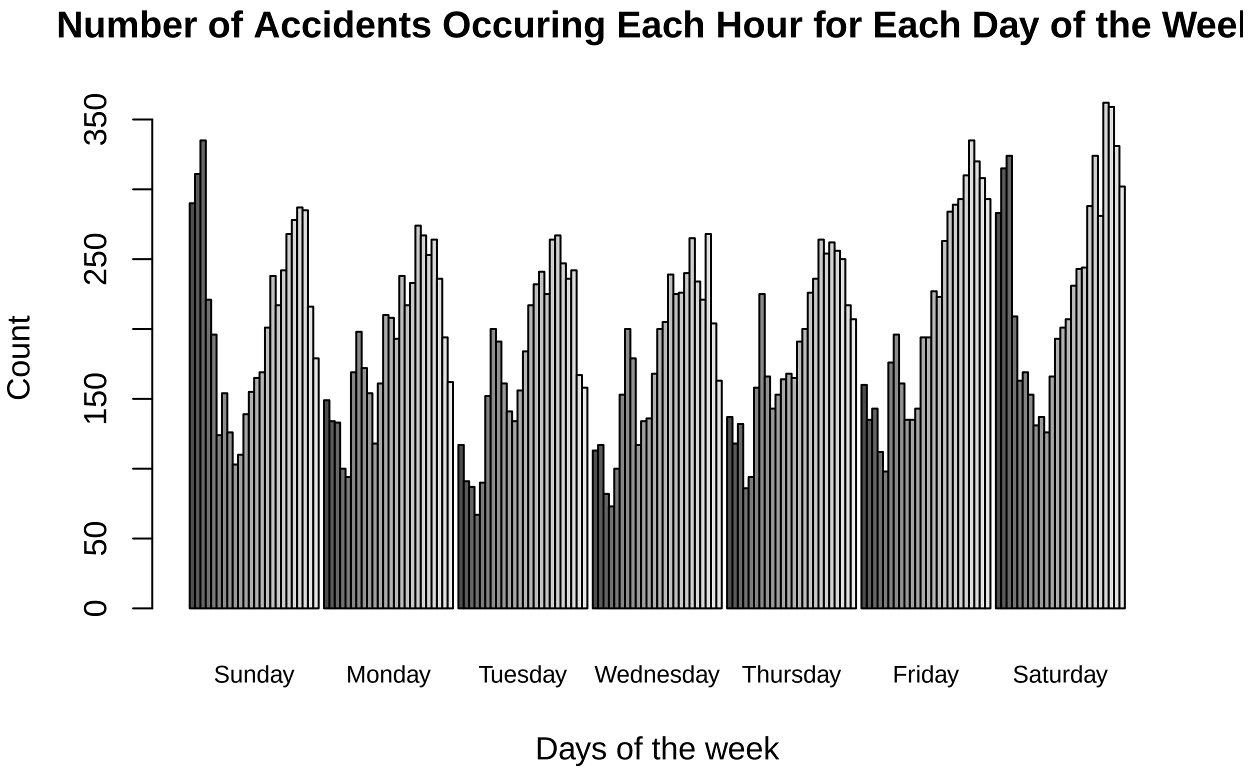
```
## Warning: 46 parsing failures.
##  row      col               expected      actual                                        file
## 1371 TRLR2VIN no trailing characters BN1T274XJP2 '../FARS_Data/FARS2018NationalCSV/VEHICLE.csv'
## 2185 TRLR2VIN no trailing characters JJV281D0JL0 '../FARS_Data/FARS2018NationalCSV/VEHICLE.csv'
## 6130 TRLR2VIN no trailing characters T9KA1N25A10 '../FARS_Data/FARS2018NationalCSV/VEHICLE.csv'
## 6817 TRLR2VIN no trailing characters G9B21209F10 '../FARS_Data/FARS2018NationalCSV/VEHICLE.csv'
## 6915 TRLR2VIN no trailing characters C9F302P2DF0 '../FARS_Data/FARS2018NationalCSV/VEHICLE.csv'
## .... ........ ...................... ........... ..............................................
## See problems(...) for more details.
```

```r
miles_of_road <- read_excel("../Background_Information/2013 miles of road per state.xlsx")
state_population <- read_excel("../Background_Information/2014 state population and total area.xlsx")
a <- select(ACC_df, ST_CASE, DAY_WEEK, RUR_URB, FUNC_SYS)
# Get the day of the week for each accident and road information
VEH_df <- left_join(VEH_df, a, by = c(ST_CASE = "ST_CASE"))
VEH_df <- VEH_df %>% mutate(time_of_day = if_else(HOUR < 12, "Morning", "Afternoon"))
```

Lets look at when during the day and week an accident is most likely to happen.

```r
ACC_df$DAY_WEEK <- as.character(ACC_df$DAY_WEEK)
Acc_per_hour_day <- ACC_df %>% filter(HOUR < 24) %>%
  mutate(DAY_WEEK = fct_recode(DAY_WEEK, Sunday = "1", Monday = "2", Tuesday = "3",
```

```
                                             Wednesday = "4", Thursday = "5", Friday = "6", Saturday = "7"))
h <- table(Acc_per_hour_day$HOUR, Acc_per_hour_day$DAY_WEEK)
barplot(h, beside = T, cex.names = 0.75, ylab = "Count", xlab = "Days of the week", main = "Number of Ac
```

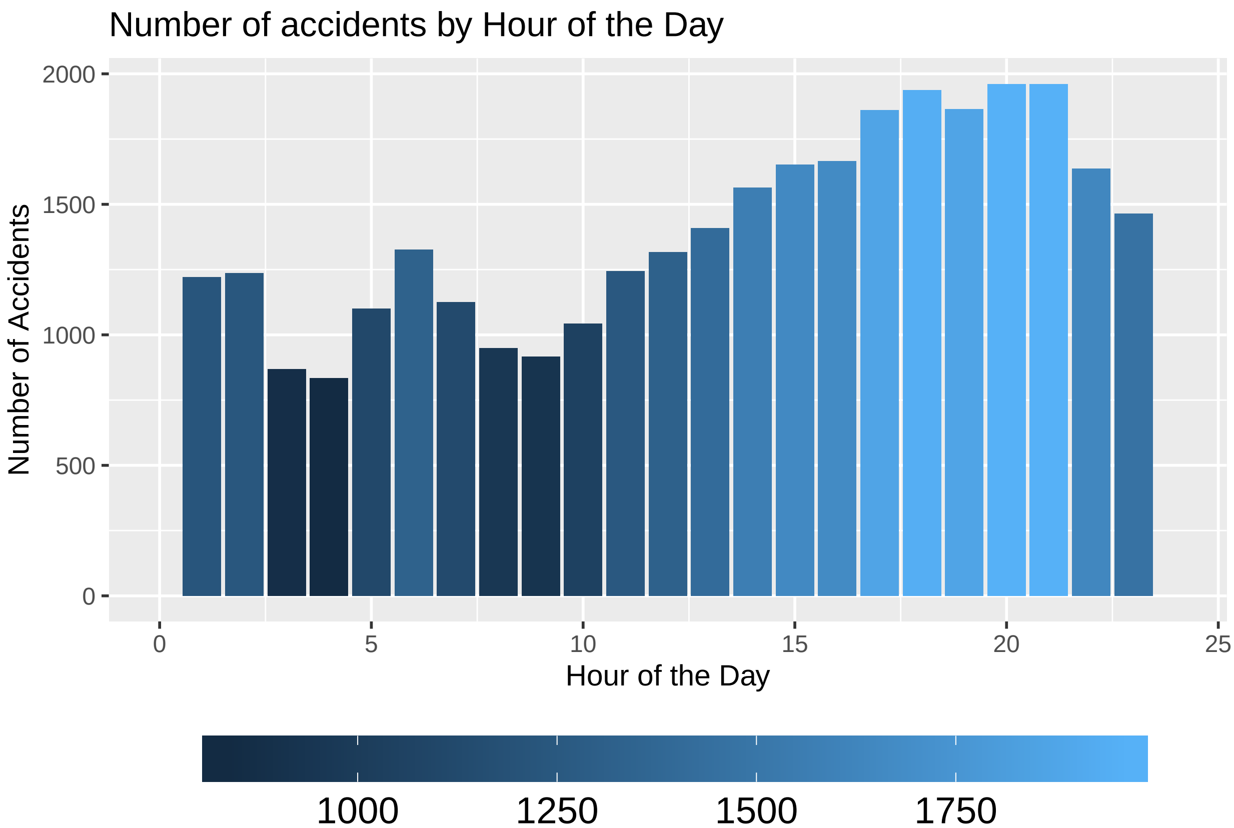


For weekdays, there is a spike at 6 AM, followed by a decrease until about 10 AM. After 10 AM, there is a slow increase until rush hour (5 - 7 PM), then a reduction in crashes until the next day at 6 AM. There are large peaks on weekend nights (Friday and Saturday Night) from roughly 8 PM to 2 AM. Our initial hypothesis is that this is strongly correlated with drunk drivers and will be tested shortly.

Next, let's combine all days together and see how accidents change by hour of the day

```
ggplot(data = Acc_per_hour_day, aes(x = HOUR, y = ..count.., fill = ..count..)) +
  geom_bar() +
  xlim(c(0, 24)) +
  xlab("Hour of the Day") +
  ylab("Number of Accidents") +
  ggtitle("Number of accidents by Hour of the Day") +
  theme(legend.position="bottom", legend.direction="horizontal", legend.text = element_text(size=14), le
```

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



Similar to the previous plot, there is a spike around 6 AM, a decrease until 10 AM and then a slow increase the remainder of the day. This plot shows that most accidents occur between 4 and 9 PM. Thus, in general it is more dangerous to drive in rush hour traffic than late night weekend traffic on average.

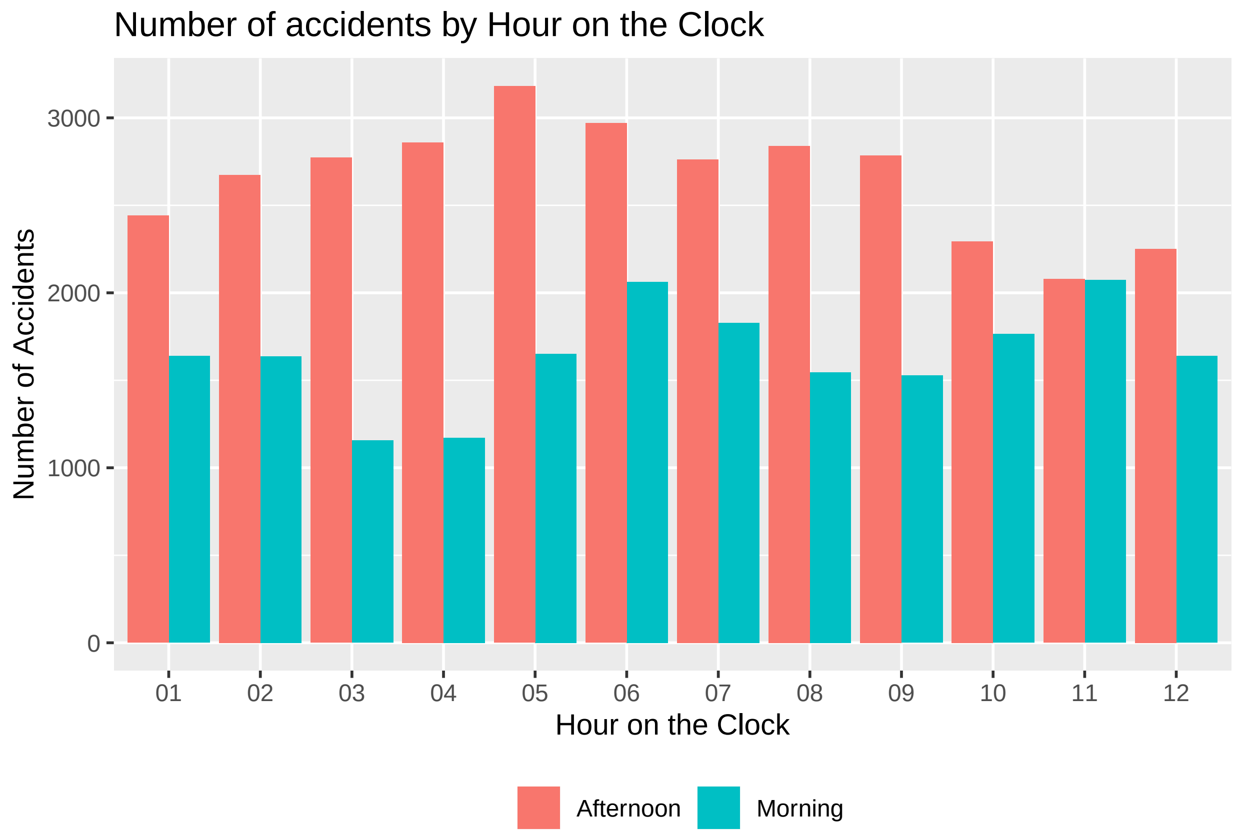
Another visualization is to see how the times of AM and PM compare to each other.

```
VEH_df$HOUR2 <- format(strptime(VEH_df$HOUR, "%H"), "%I")
hourly_accident_plot <- VEH_df %>% filter(HOUR2 < 13) %>%
  ggplot(aes(x = HOUR2, y = ..count.., shade = time_of_day, fill = ..count..)) +
  geom_bar(mapping = aes(x = HOUR2, y = ..count.., fill = time_of_day), position = "dodge") +
  xlab("Hour on the Clock") +
  ylab("Number of Accidents") +
  ggtitle("Number of accidents by Hour on the Clock") +
  theme(legend.position="bottom", legend.direction="horizontal", legend.title = element_blank())#+
  coord_polar()
```

```
## <ggproto object: Class CoordPolar, Coord, gg>
##     aspect: function
##     backtransform_range: function
##     clip: on
##     default: FALSE
##     direction: 1
##     distance: function
##     is_free: function
##     is_linear: function
##     labels: function
##     modify_scales: function
##     r: y
##     range: function
```

```
##      render_axis_h: function
##      render_axis_v: function
##      render_bg: function
##      render_fg: function
##      setup_data: function
##      setup_layout: function
##      setup_panel_params: function
##      setup_params: function
##      start: 0
##      theta: x
##      transform: function
##      super:  <ggproto object: Class CoordPolar, Coord, gg>
```

```
#+ geom_text(aes(y=..count..,label= ..count..), color= 'white', size =2)
hourly_accident_plot
```



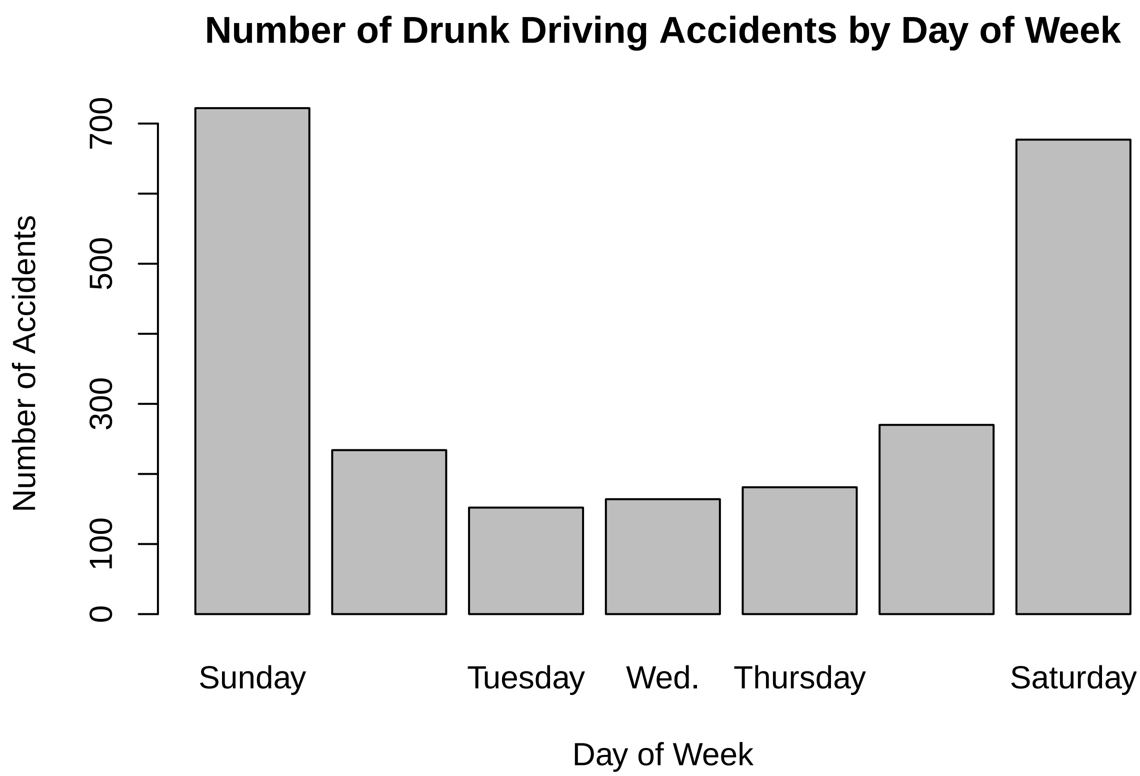
We find that the afternoon hours (i.e. 12 PM to 12 AM) are more dangerous. Interestingly, 11 AM and 11 PM have roughly the same amount of accidents.

There are large spikes of accidents from 8 PM to 3 AM for Friday Night and Saturday Night. First, let's do a preliminary check and make sure drinking is correlated by only examining accidents from 12 AM to 3 AM.

```
drunk_early_morning<-VEH_df%>%
  select(ST_CASE, HOUR, DAY_WEEK, DR_DRINK)%>%
  count(DAY_WEEK, HOUR, DR_DRINK)%>%
  filter(HOUR<= 3, DR_DRINK==1)

Late_drunk_acc<-aggregate(drunk_early_morning$n, by= list(Category=drunk_early_morning$DAY_WEEK),FUN= su
barplot(Late_drunk_acc$x,
```

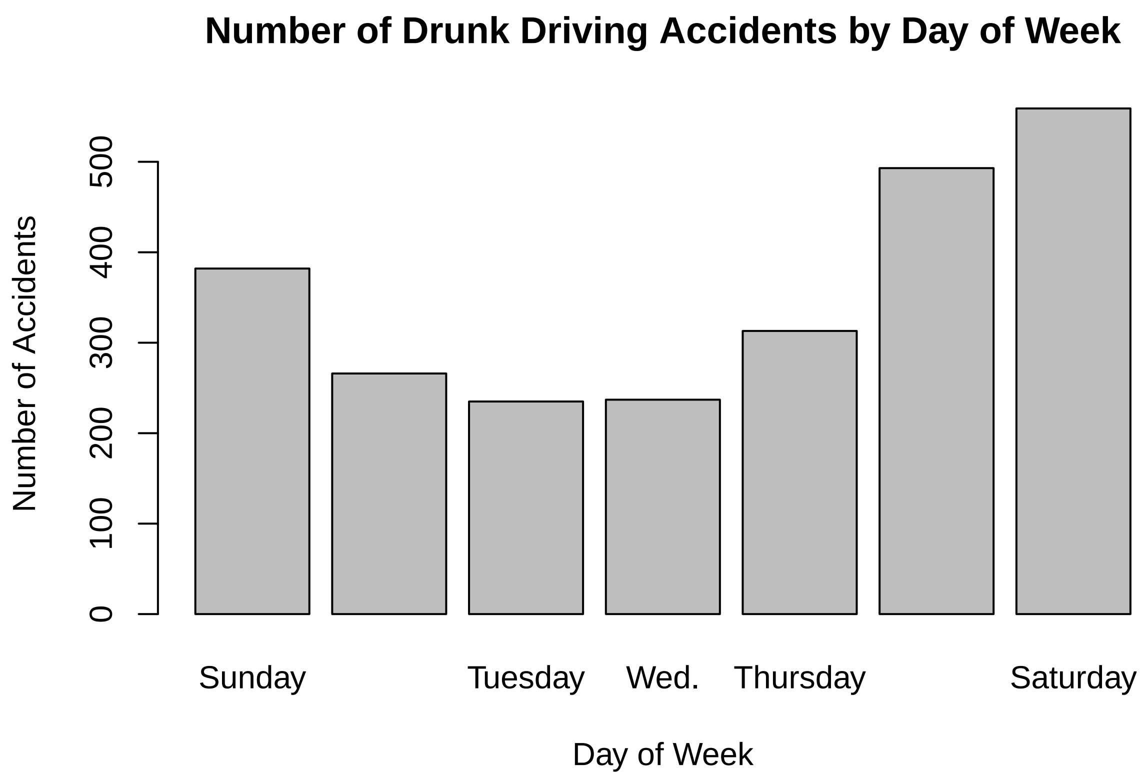
```
        xlab = "Day of Week",
        names = c("Sunday", "Monday", "Tuesday", "Wed.", "Thursday", "Friday", "Saturday"),
        ylab = "Number of Accidents",
        main = "Number of Drunk Driving Accidents by Day of Week")
```



Next, lets look at drinking accidents from 8 PM to 12 AM.

```
drunk_late_night<-VEH_df%>%
  select(ST_CASE, HOUR, DAY_WEEK, DR_DRINK)%>%
  count(DAY_WEEK, HOUR, DR_DRINK)%>%
  filter(HOUR>= 20, HOUR < 24, DR_DRINK==1)

Late_drunk_acc_night<-aggregate(drunk_late_night$n, by= list(Category=drunk_late_night$DAY_WEEK),FUN= su
barplot(Late_drunk_acc_night$x,
        xlab = "Day of Week",
        names = c("Sunday", "Monday", "Tuesday", "Wed.", "Thursday", "Friday", "Saturday"),
        ylab = "Number of Accidents",
        main = "Number of Drunk Driving Accidents by Day of Week")
```



There is an increase in drunk driving accidents from 8 PM to 12 AM on Friday and Saturday night, but it seems that the majority of drunk driving accidents occur after 12 AM as shown by the previous figure.

This needs to be updated, because rush hour is not a thing on Saturday and Sunday. Also it may be interesting to look at the 6-8 AM rush hour combined with this. Number of accidents during rush hour compared to the number of accidents not during rush hour

```
Rush_hour_acc<-VEH_df%>%
  select(ST_CASE, HOUR, DAY_WEEK)%>%
  count(DAY_WEEK, HOUR)%>%
  filter(16<=HOUR, HOUR< 20)

rush_hour_accidents <- aggregate(Rush_hour_acc$n, by= list(Category=Rush_hour_acc$DAY_WEEK),FUN= sum)

Not_rush_hour_acc<-VEH_df%>%
  select(ST_CASE, HOUR, DAY_WEEK)%>%
  count(DAY_WEEK, HOUR)%>%
  subset(HOUR >=20 | HOUR<16)

not_rush_hour_accidents <- aggregate(Not_rush_hour_acc$n, by= list(Category=Not_rush_hour_acc$DAY_WEEK)

rush_hour_acc_percent <- sum(rush_hour_accidents[,2]) / (sum(rush_hour_accidents[,2]) +
                                                          sum(not_rush_hour_accidents[,2]))

print(sprintf("Percentage of Accidents that Occur during Rush Hour is %s%%",
              round(rush_hour_acc_percent*100, digits = 3)))
```

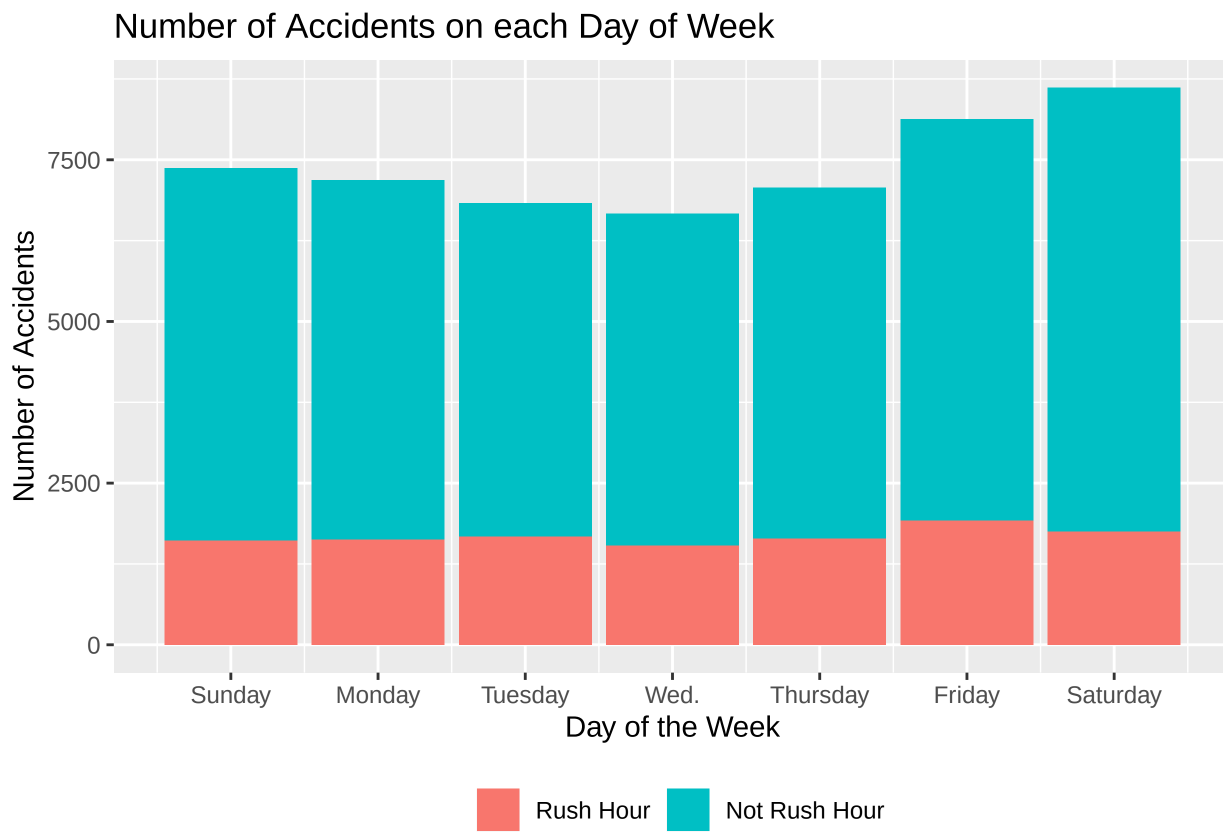
```
## [1] "Percentage of Accidents that Occur during Rush Hour is 22.698%"
```

```
rush_hour_accidents = cbind(rush_hour_accidents, "Rush Hour")
colnames(rush_hour_accidents) <- c("Day", "Number", "Category")
```

```
not_rush_hour_accidents = cbind(not_rush_hour_accidents, "Not Rush Hour")
colnames(not_rush_hour_accidents) <- c("Day", "Number", "Category")

all_accidents = rbind(rush_hour_accidents, not_rush_hour_accidents)

ggplot(all_accidents, aes(fill=all_accidents$Category, y=all_accidents$Number, x=all_accidents$Day)) +
  geom_bar(position=position_stack(reverse = TRUE), stat="identity") +
  xlab("Day of the Week") +
  # names(c("Sunday", "Monday", "Tuesday", "Wed.", "Thursday", "Friday", "Saturday")) +
  ylab("Number of Accidents") +
  ggtitle("Number of Accidents on each Day of Week ") +
  scale_x_continuous(breaks=1:7, labels=c("Sunday", "Monday", "Tuesday", "Wed.", "Thursday", "Friday",
  theme(legend.position="bottom", legend.direction="horizontal", legend.title = element_blank())
```



Overall, rush hour accidents comprise 22% of accidents, while only being 16% of total time during the week.

Next, let's look at how the type of vehicle is related to crash patterns.

```
Truck <- VEH_df %>% count(MODEL) %>% filter(MODEL > 400, MODEL < 500)
Truck <- length(Truck)

Automobile <- VEH_df %>% count(MODEL) %>% filter(MODEL < 400)
Automobile <- length(Automobile)

Motorcycles <- VEH_df %>% count(MODEL) %>% filter(MODEL > 700, MODEL < 710)
Motorcycles <- sum(Motorcycles)

Heavy_Truck <- VEH_df %>% count(MODEL) %>% filter(MODEL > 880, MODEL < 900)
```

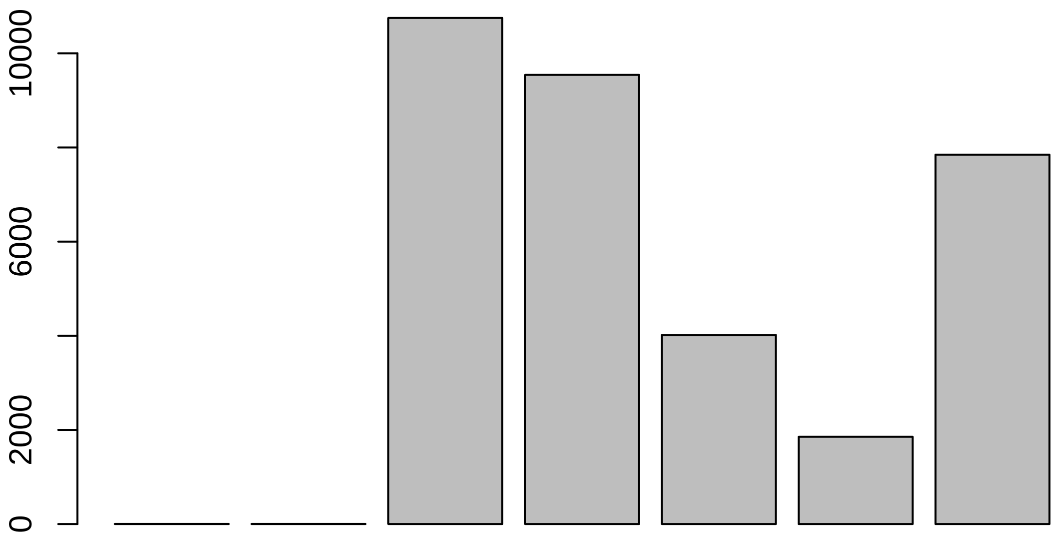
```r
Heavy_Truck <- sum(Heavy_Truck)

ATV <- VEH_df %>% count(MODEL) %>% filter(MODEL > 730, MODEL < 740)
ATV <- sum(ATV)

MotorHome_Van <- VEH_df %>% count(MODEL) %>% filter(MODEL > 849, MODEL < 871)
MotorHome_Van <- sum(MotorHome_Van)

Bus <- VEH_df %>% count(MODEL) %>% filter(MODEL > 900, MODEL < 990)
Bus <- sum(Bus)

k <- rbind(Automobile, Truck, Motorcycles, Heavy_Truck, ATV, MotorHome_Van, Bus)
k <- as.data.frame(k)
barplot(k$V1)
```




```r
# normailized which car is more frequently droven
```

Next, lets look at the road type and how that is relatead to crashes. Need to bring in the FUNC SYS and RURURB tables to identify what these are.

```r
VEH_df %>% count(FUNC_SYS)
```

```
## # A tibble: 10 x 2
##    FUNC_SYS     n
##       <dbl> <int>
##  1        1  7365
##  2        2  2322
##  3        3 16727
##  4        4 11082
##  5        5  6640
##  6        6  1381
##  7        7  5414
##  8       96    66
##  9       98   854
## 10       99    21
```

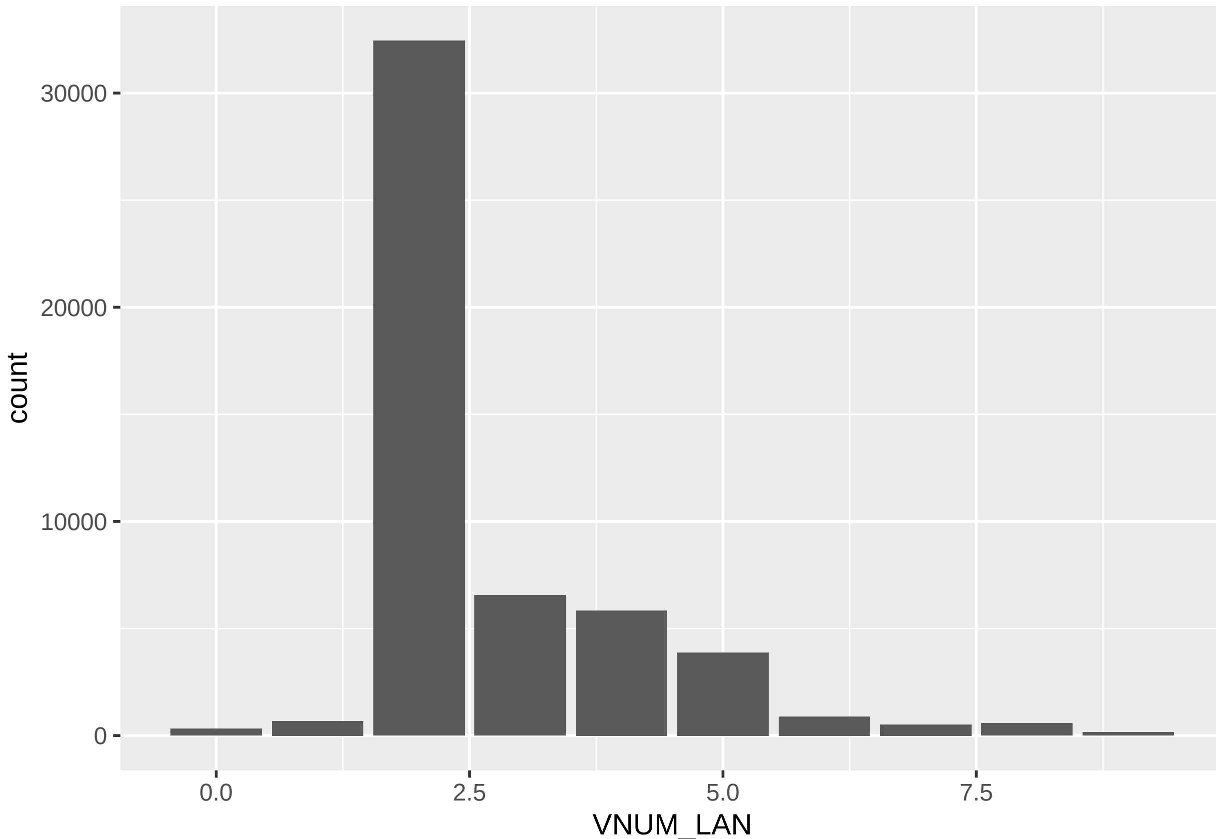
```r
VEH_df %>% count(RUR_URB) %>% filter(RUR_URB<=2)
```

```
## # A tibble: 2 x 2
```

```
##   RUR_URB     n
##     <dbl> <int>
## 1       1 22285
## 2       2 28670
```

Next, lets look at how crashes are related to the number of lanes on a road.

```
VEH_df %>% select(VNUM_LAN, VPROFILE) %>%
  ggplot(aes(x=VNUM_LAN, y=..count.. ))+ geom_bar()
```




```
#%>%count(VPROFILE)
# Majority of roads are flat in the united states but it is interesting to note many accidents occur on
```

RelJct1 - if in a interchange area RelJct2 - Where in the interchange area