Assignment 2 : Visualization of Utah Crash Data for 2014-2018

Vinod Krishnan

April 24, 2019

##Project Objective & Target Audience

#The objective of this project is to use Data Visualization Techniques to make better sense of data which then can be used to make informed decisions. I chose to use Crash data for Utah for the period of 2014-2018. This was obtained from Utah.gov website. They were provided this data by Utah Department of Transportation. Link for the data is as below:- (https://opendata.utah.gov/Transportation/Crash-data-for-Utah-2014-2018/a64b-mcum (https://opendata.utah.gov/Transportation/Crash-data-for-Utah-2014-2018/a64b-mcum))

#My target audience would be the Utah Transportation department who can use the visualizations to identify areas that are highly prone to accidents. It will help them to mobilize units where needed the most. They could also use it to identify target areas where they would need to put up signages and boards cautioning drivers and pedestrains in those areas to be careful to avoid possible mishaps.

```
knitr::opts_chunk$set(echo = FALSE,tidy.opts=list(width.cutoff=50),tidy=TRUE)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------- tidyverse
1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

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

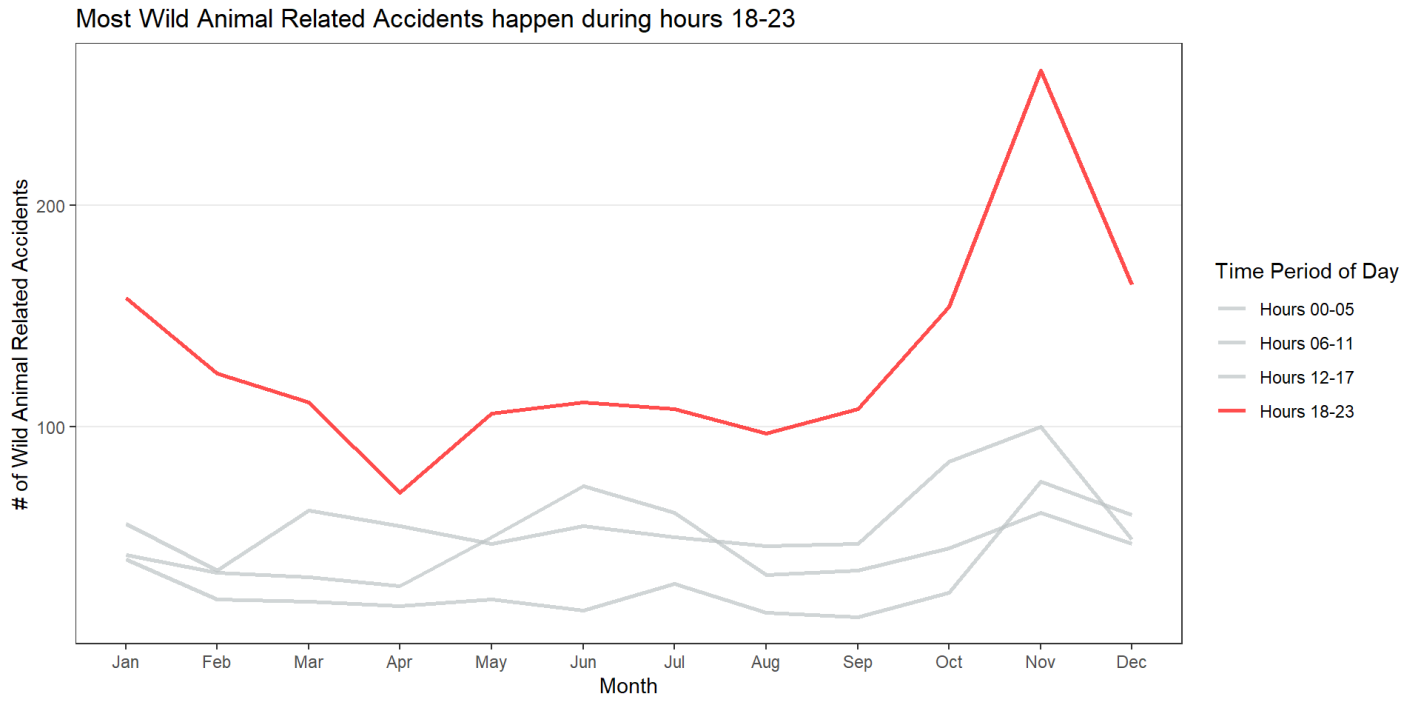
```
library(ggplot2)
library(stringr)
```

#Here I am reading in the data and formatting/cleaning data so that it can be used for Visualization

```
c<-read_csv("Crash_data_for_Utah_2014-2018.csv")
c<-na.omit(c)

c$CRASH_DATETIME<-as.POSIXct(c$CRASH_DATETIME, format="%m/%d/%Y %I:%M:%S %p")

Hours <- format(as.POSIXct(strptime(c$CRASH_DATETIME,"%Y-%m-%d %H:%M:%S",tz="")) ,format = "%H:%
M:%S")

Dates <- format(as.POSIXct(strptime(c$CRASH_DATETIME,"%Y-%m-%d %H:%M:S%",tz="")) ,format = "%Y-%
m-%d")

c<-separate(c, CRASH_DATETIME, c("Date", "Time"), sep = "\\b\\s\\b", remove = FALSE)

c<-separate(c, Date, c("Year", "Month", "Day"), sep = "-", remove = FALSE)

c$Year <- as.integer(c$Year)
c$Month <- as.integer(c$Month)
c$Day <- as.integer(c$Day)

c$Date<-as.Date(c$Date)

c$Time1<-substr(c$Time,1,2) #Extracting just the hours from time

q1 <- c('00','01','02','03','04','05')
q2 <- c('06','07','08','09','10','11')
q3 <- c('12','13','14','15','16','17')
q4 <- c('18','19','20','21','22','23')

c$Timesplit <- if_else(c$Time1 %in% q1,'Hours 00-05',ifelse(c$Time1 %in% q2,'Hours 06-11',ifelse
(c$Time1 %in% q3,'Hours 12-17','Hours 18-23')))

month_list <- c('Jan','Feb','Mar','Apr','May','Jun','Jul','Aug','Sep','Oct','Nov','Dec')
```

Visualization 1 - Trying to identify Wild Animal related accidents Monthly across time periods

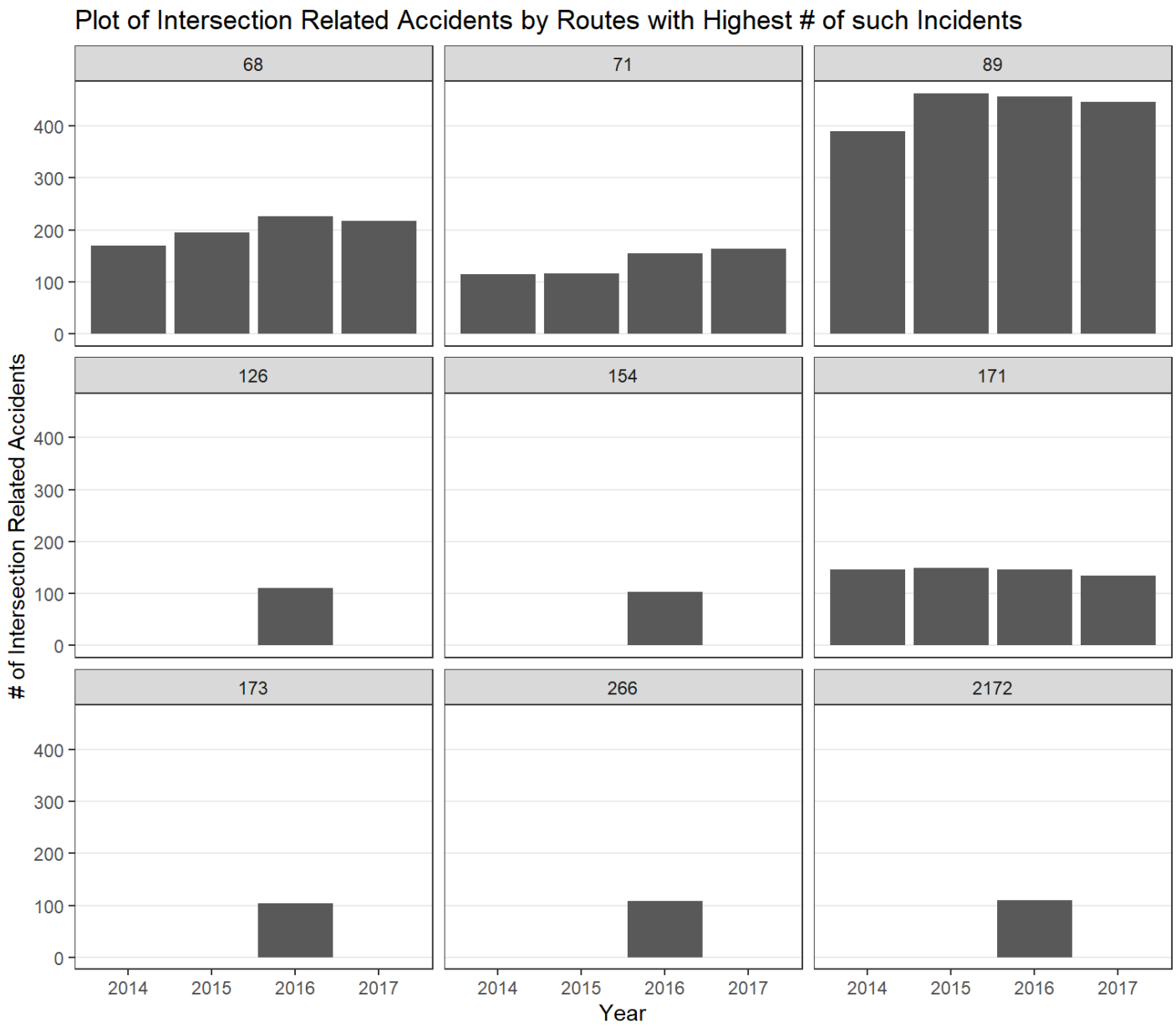
```
bbbb <- c %>% group_by(WILD_ANIMAL_RELATED,Month,Timesplit) %>% summarise(n=n()) %>% filter(WILD
_ANIMAL_RELATED=='Y')

ggplot(bbbb) +
  geom_line(aes(x=Month,y=n,color=Timesplit),size=1) +
  scale_color_manual(values=alpha(c('#bcc3c4','#bcc3c4','#bcc3c4','#FF0000'),0.7)) +
  scale_x_continuous(breaks=c(1:12),labels=month_list) +
  labs(x='Month',y='# of Wild Animal Related Accidents',color='Time Period of Day',title='Most W
ild Animal Related Accidents happen during hours 18-23') +
  theme_bw() +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.minor.y = element_blank())
```



#The time of the day between Hours 18-23 has the highest number of Wild Animal related accidents across all months which is significant. Another thing that I noted in the plot is that excepting for Hours 00-05, the highest number of Wild Animals Related Accidents seem to happen in month of November over the period of last 5 years. All of this is interesting and it might make sense to further drill this down to see which routes if any in particular where these are high and what measures need to be taken to avoid this.

Visualization 2 - Identify routes with Highest Intersection related accidents

```
cccc <- c %>% group_by(INTERSECTION_RELATED,Year,ROUTE) %>% summarise(n=n()) %>% filter(INTERSEC
TION_RELATED=='Y' & n>=100 & Year != 2018)


plot1<-ggplot(data=cccc,aes(x=Year,y=n)) +
  geom_bar(stat = "identity", position = "dodge", size=1) +
  facet_wrap(~ROUTE)+
  labs(x='Year',y='# of Intersection Related Accidents',color='Time Period of Day',title='Plot o
f Intersection Related Accidents by Routes with Highest # of such Incidents') +
  theme_bw() +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.minor.y = element_blank())

plot1
```



#From above plot we see that on Routes 68,71,89 & 171 seem to have increasing/same number of Intersection Related Accidents. One must check out these routes and identify these spots and ensure proper warnings/signage must be put up to help predestrians. You could also inform Neighborhood communities along these routes to keep a vigilant eye out for incidents and report back to the Police. In addition polcie could also increase their checks to identify violators. You would also want to strategise solutions to reduce these cases.

#In addition we see that on the other routes on this chart namely 126,266,173 and 2172 seem to have had such cases only in 2016 and then again back to 0 which could mean some actions might have been taken on those routes to avoid such incidents. These can be used in other locations post checking the feasibility.