Project Phase 1

# Project Phase 1

## Olivia Fowler

library(ggplot2)  
library(lubridate)

##   
## Attaching package: 'lubridate'

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

library(zoo)

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

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

library(knitr)  
library(tidyverse)

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

## v tibble 2.1.3 v purrr 0.3.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts --------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x lubridate::as.difftime() masks base::as.difftime()  
## x lubridate::date() masks base::date()  
## x dplyr::filter() masks stats::filter()  
## x lubridate::intersect() masks base::intersect()  
## x dplyr::lag() masks stats::lag()  
## x lubridate::setdiff() masks base::setdiff()  
## x lubridate::union() masks base::union()

chicago <- read\_csv("~/Spring 2020/BAN 502/Project/chicago.csv")

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## ID = col\_double(),  
## Arrest = col\_logical(),  
## Domestic = col\_logical(),  
## Ward = col\_double(),  
## `Community Area` = col\_double(),  
## `X Coordinate` = col\_double(),  
## `Y Coordinate` = col\_double(),  
## Year = col\_double(),  
## Latitude = col\_double(),  
## Longitude = col\_double()  
## )

## See spec(...) for full column specifications.

chicago=chicago%>%mutate(Date = mdy\_hms(Date))  
chicago = chicago %>% mutate(Hour = hour(Date)) #creates new variable in dataset with the  
summary(chicago$Hour) #shows summary of this new variable

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 9.00 14.00 13.08 18.00 23.00

chicago = chicago %>% dplyr::select (Date, Hour, Block, IUCR, "Primary Type", Description, Arrest, Domestic, Beat, District, Ward, "Community Area", "FBI Code", "Location Description",Year, Latitude, Longitude)  
chicago = chicago %>% mutate(Arrest = as.factor(Arrest)) %>%  
 mutate(Arrest = fct\_recode(Arrest, "False" = "No", "True" = "Yes" )) %>%  
 mutate(Hour = as.factor(Hour)) %>% mutate('Primary Type' = as\_factor(as.character('Primary Type')))%>%  
 mutate(Description = as.factor(Description))%>%mutate('Location Description' = as\_factor(as.character('Location Description'))) %>%  
 mutate(Domestic = as.factor(Domestic))%>%  
 mutate(District = as.factor(District))%>%  
 mutate(Year = as.factor(Year))%>%  
 mutate(Ward = as.factor(Ward))%>%mutate('Community Area' = as\_factor(as.character('Community Area')))%>%mutate(Block = as.factor(Block))%>%mutate(IUCR = as.factor(IUCR))

## Warning: Unknown levels in `f`: No, Yes

str(chicago)

## Classes 'spec\_tbl\_df', 'tbl\_df', 'tbl' and 'data.frame': 267185 obs. of 17 variables:  
## $ Date : POSIXct, format: "2018-01-01 00:00:00" "2018-01-01 00:00:00" ...  
## $ Hour : Factor w/ 24 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Block : Factor w/ 27909 levels "0000X E 100TH PL",..: 22645 22811 23264 16957 18229 5703 7469 19762 4264 18504 ...  
## $ IUCR : Factor w/ 322 levels "0110","0142",..: 200 109 115 199 199 7 13 199 199 254 ...  
## $ Primary Type : Factor w/ 1 level "Primary Type": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Description : Factor w/ 301 levels "$500 AND UNDER",..: 241 127 119 3 3 33 178 3 3 135 ...  
## $ Arrest : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 2 1 1 ...  
## $ Domestic : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 2 1 1 1 1 1 1 ...  
## $ Beat : chr "2431" "2411" "1613" "1911" ...  
## $ District : Factor w/ 23 levels "001","002","003",..: 21 21 15 18 9 20 12 7 9 15 ...  
## $ Ward : Factor w/ 50 levels "1","2","3","4",..: 49 50 41 40 16 21 27 15 20 41 ...  
## $ Community Area : Factor w/ 1 level "Community Area": 1 1 1 1 1 1 1 1 1 1 ...  
## $ FBI Code : chr "02" "11" "11" "17" ...  
## $ Location Description: Factor w/ 1 level "Location Description": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Year : Factor w/ 1 level "2018": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Latitude : num NA NA NA NA NA ...  
## $ Longitude : num NA NA NA NA NA ...

chicago$day <- as.factor(wday(chicago$Date))  
chicago$month <- as.factor(month(chicago$Date))

library(tidytext)  
library(ggplot2)  
library(tidyr)  
library(corrplot)

## corrplot 0.84 loaded

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

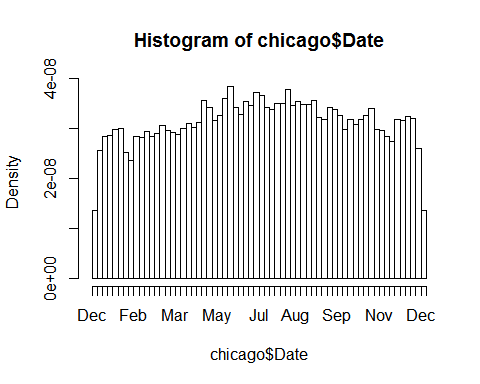
##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

library(ggcorrplot)

hist(chicago$Date, breaks=100)

## Warning in breaks[-1L] + breaks[-nB]: NAs produced by integer overflow



table(chicago$month)

##   
## 1 2 3 4 5 6 7 8 9 10 11 12   
## 20419 17310 21197 21098 24682 24184 25201 25356 23033 22789 20622 21294

table(chicago$Description)[which.max(table(chicago$Description))]

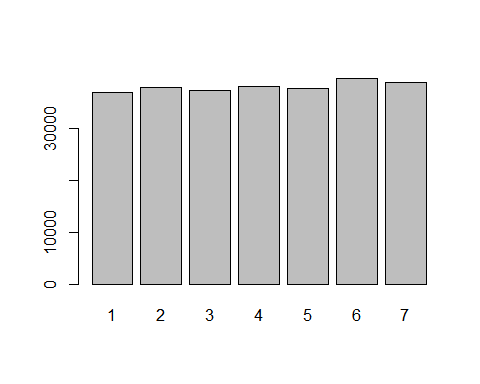
## SIMPLE   
## 29660

table(chicago$Description)[which.min(table(chicago$Description))]

## AGG PRO.EMP: OTHER FIREARM   
## 1

Days <- table(chicago$day)

barplot(Days)



table(chicago$"FBI Code")

##   
## 01A 01B 02 03 04A 04B 05 06 07 08A 08B 09 10   
## 585 12 1853 9640 5980 7718 11716 65088 9959 14622 42047 372 1542   
## 11 12 13 14 15 16 17 18 19 20 22 24 26   
## 17639 34 83 27727 5584 729 1579 13432 202 1732 266 2286 24758

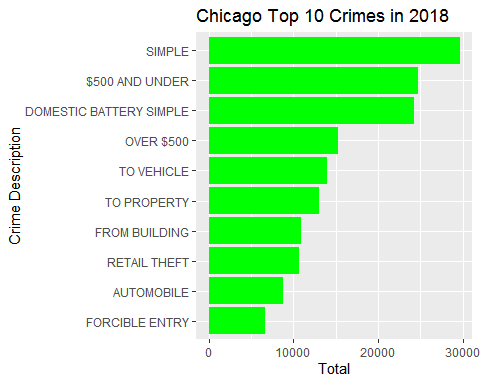
table(chicago$"Location Description")

##   
## Location Description   
## 267185

topcrime = chicago %>%  
 group\_by(Description) %>%  
 summarise(total = n ()) %>%  
 distinct() %>%  
 top\_n(10)

## Selecting by total

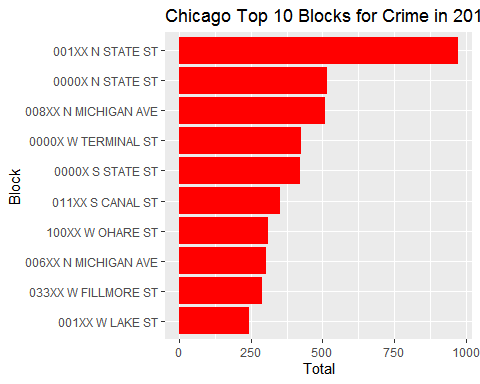
topcrime %>%  
 ggplot(aes(reorder(Description, total), y=total)) +  
 geom\_col(fill = "green") +  
 coord\_flip() +  
 labs(title = "Chicago Top 10 Crimes in 2018",  
 x= "Crime Description",  
 y= "Total")



topblock = chicago %>%  
 group\_by(Block) %>%  
 summarise(total = n ()) %>%  
 distinct() %>%  
 top\_n(10)

## Selecting by total

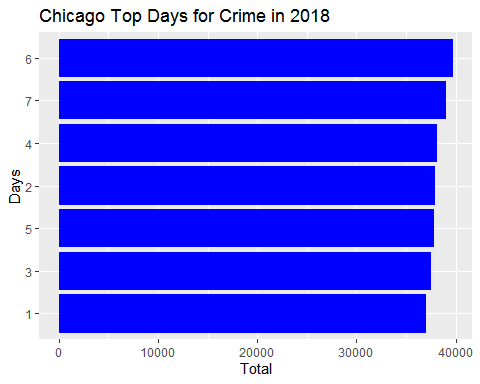
topblock %>%  
 ggplot(aes(reorder(Block, total), y=total)) +  
 geom\_col(fill = "red") +  
 coord\_flip() +  
 labs(title = "Chicago Top 10 Blocks for Crime in 2018",  
 x= "Block",  
 y= "Total")



topday = chicago %>%  
 group\_by(day) %>%  
 summarise(total = n ()) %>%  
 distinct() %>%  
 top\_n(7)

## Selecting by total

topday %>%  
 ggplot(aes(reorder(day, total), y=total)) +  
 geom\_col(fill = "blue") +  
 coord\_flip() +  
 labs(title = "Chicago Top Days for Crime in 2018",  
 x= "Days",  
 y= "Total")



topDistrict = chicago %>%  
 group\_by(District) %>%  
 summarise(total = n ()) %>%  
 distinct() %>%  
 top\_n(10)

## Selecting by total

topDistrict %>%  
 ggplot(aes(reorder(District, total), y=total)) +  
 geom\_col(fill = "purple") +  
 coord\_flip() +  
 labs(title = "Chicago Top Districts for Crime in 2018",  
 x= "District",  
 y= "Total")

