EDA

Exploratory Data Analysis

Useful links

- Data set

Imports

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

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

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

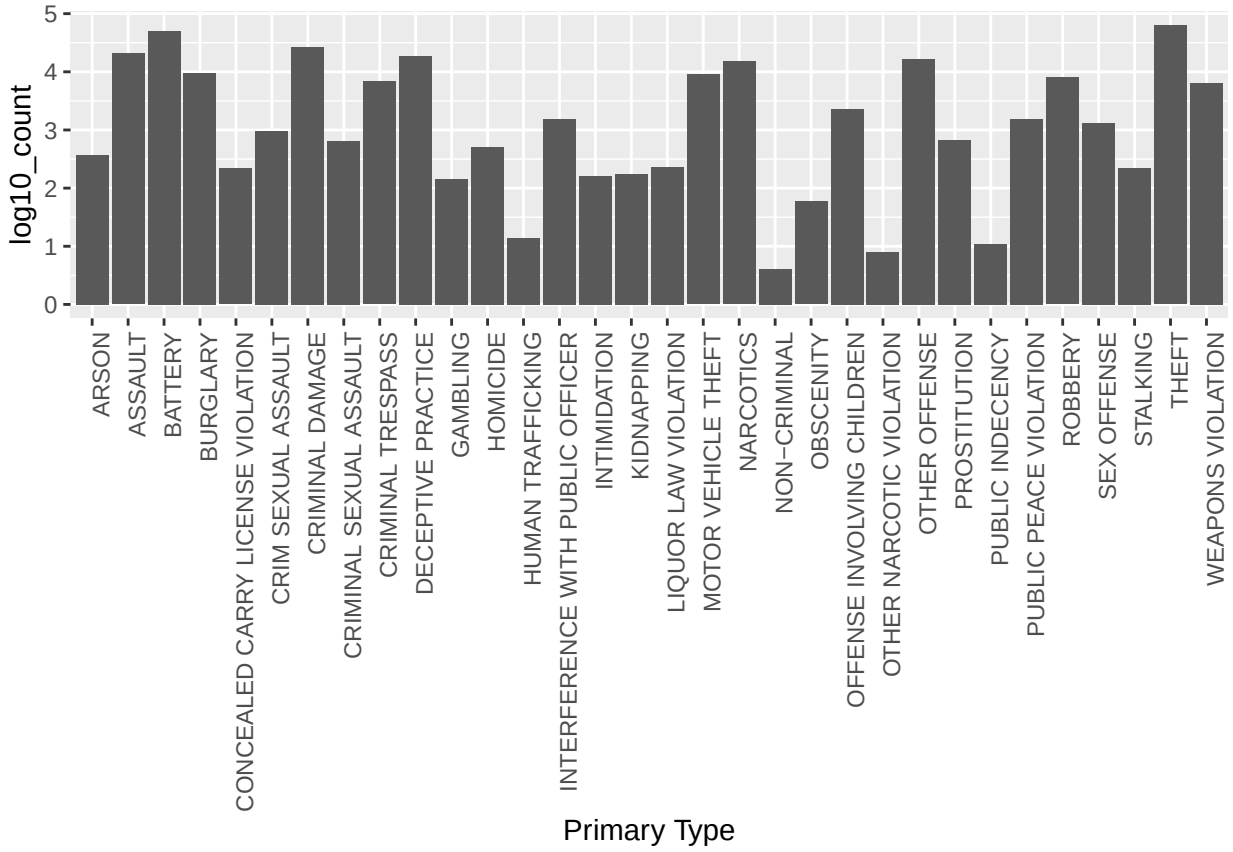
```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Load data

```
df = read_csv("./data/crime-2019.csv", col_types = cols())
df$Timestamp = ymd_hms(mdy_hms(df$Date))
df$Date = date(df$Timestamp)
```

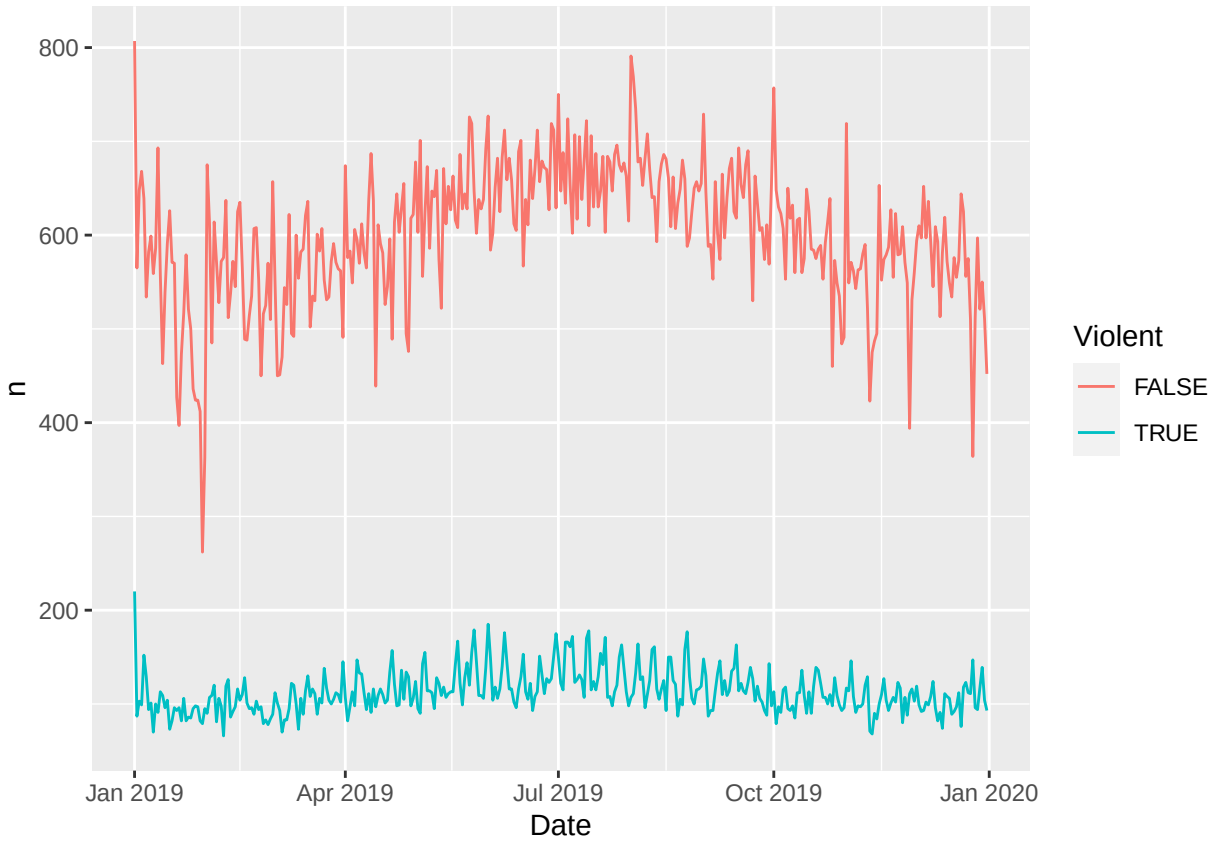
Overall:

```
df %>%
  group_by(`Primary Type`) %>%
  tally(name = "count") %>%
  mutate(log10_count = log10(count)) %>%
  ggplot() +
  geom_col(aes(x=`Primary Type`, y=log10_count)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



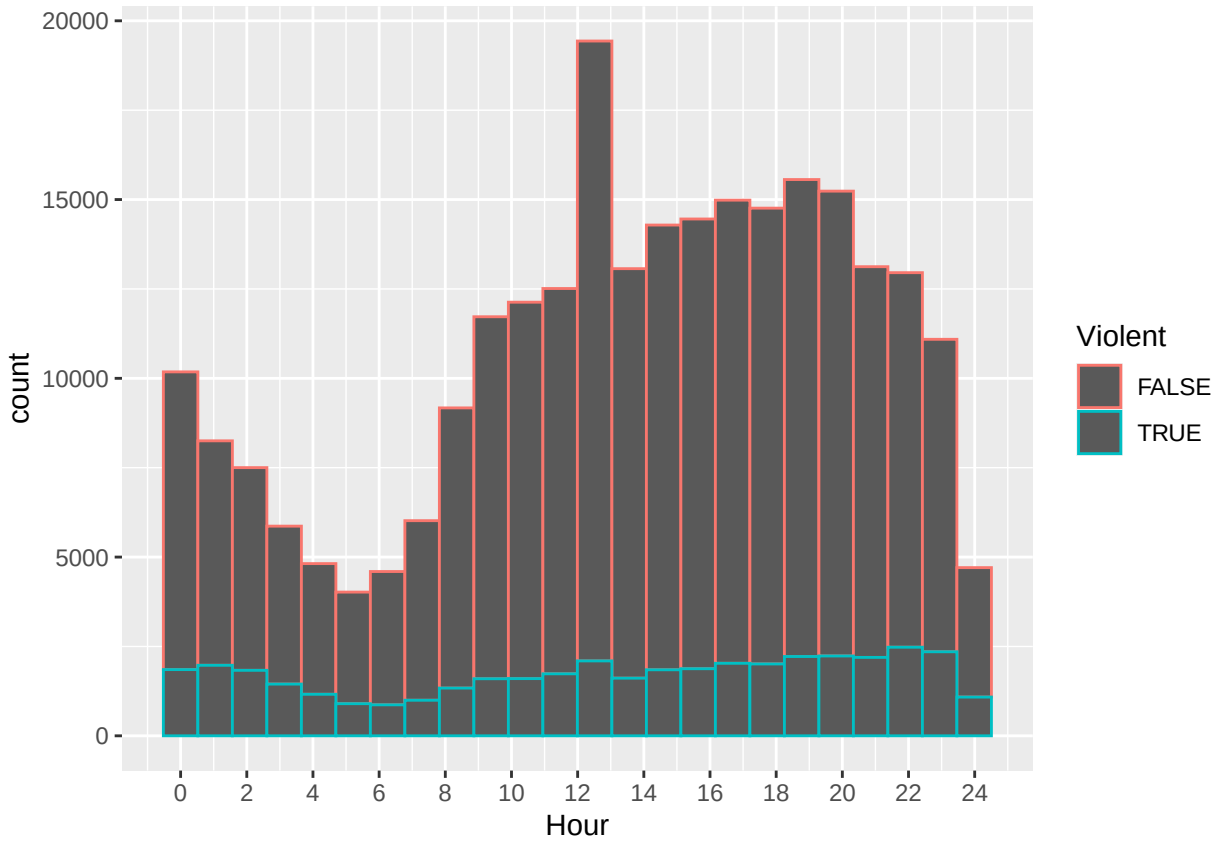
Date and time Over the year:

```
descs = unique(df$Description)

# This might not be perfect but looks pretty good
violent = descs[str_detect(descs, "SEX") |
                    str_detect(descs, "ABUSE") |
                    str_detect(descs, "HOMICIDE") |
                    str_detect(descs, "VIOLENT") |
                    str_detect(descs, "BATTERY") |
                    (str_detect(descs, "AGGRAVATED") & !str_detect(descs, "NON-AGGRAVATED"))
                ]

df$Violent = df$Description %in% violent

df %>%
  group_by(Date, Violent) %>%
  tally() %>%
  ggplot(aes(x=Date, y=n, col=Violent)) +
    geom_line()
```



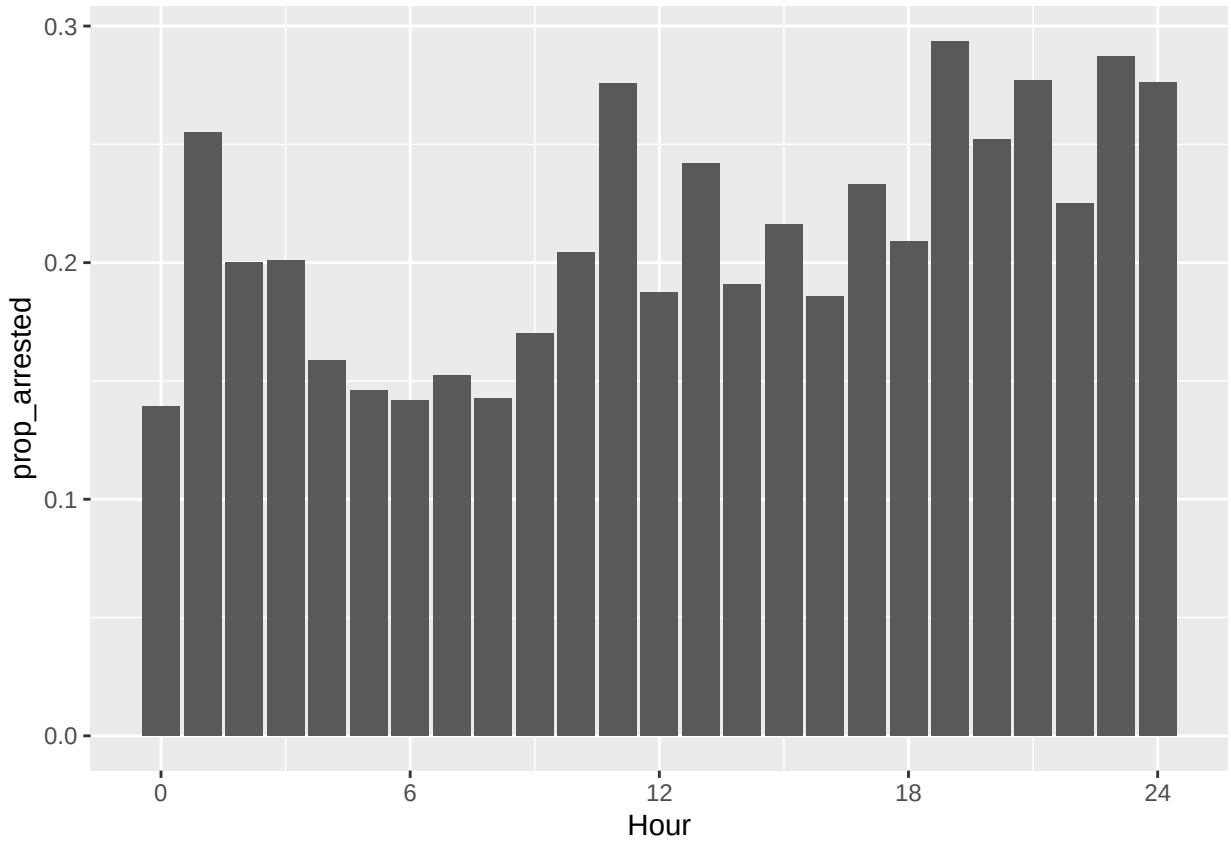
Over 24 hour period:

```
df$Hour = hour(df$Timestamp) + minute(df$Timestamp)/60

df %>%
  ggplot(aes(x=Hour, col=Violent)) +
    geom_histogram(bins=24) +
    scale_x_continuous(breaks = seq(0,24, 2))
```



Proportion of incidents leading to arrests by time of day:

```
df %>%
  mutate(Hour = round(Hour)) %>%
  group_by(Hour, Arrest) %>%
  tally() %>%
  pivot_wider(names_from = Arrest, values_from = n) %>%
  mutate(prop_arrested = `TRUE`/(`TRUE`+`FALSE`)) %>%
  ggplot(aes(x=Hour, y=prop_arrested)) +
  geom_col() +
  scale_x_continuous(breaks = seq(0,24,6))
```



Are there any duplicate case numbers?

```
sum(duplicated(df$`Case Number`))
```

```
## [1] 21
```

Duplicated IDs?

```
sum(duplicated(df$ID))
```

```
## [1] 0
```

Sp

```
library(sf)
```

```
## Linking to GEOS 3.8.1, GDAL 3.0.4, PROJ 6.3.1
## WARNING: different compile-time and runtime versions for GEOS found:
## Linked against: 3.8.1-CAPI-1.13.3 compiled against: 3.8.0-CAPI-1.13.1
## It is probably a good idea to reinstall sf, and maybe rgeos and rgdal too
```

```
library(raster)
```

```
## Loading required package: sp
##
## Attaching package: 'raster'
## The following object is masked from 'package:dplyr':
```

```
##
##     select
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
# Import neighbourhood boundaries
bounds <- st_read("data/nbd_bounds.shp")
```

```
## Reading layer `nbd_bounds' from data source `/home/dw16200/Documents/compass/group_project/chicago-c
## Simple feature collection with 98 features and 4 fields
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -87.94011 ymin: 41.64454 xmax: -87.52414 ymax: 42.02304
## geographic CRS: WGS84(DD)
```