Exploration of Vancouver Open Crime data

Reading in the data

```
crime <- read_csv("../../../mds/3_term/22_workflows/DSCI_522_Vancouver_Bike_Theft_Analysis/data/crime_cs
```

```
## Parsed with column specification:
## cols(
##   TYPE = col_character(),
##   YEAR = col_integer(),
##   MONTH = col_character(),
##   DAY = col_character(),
##   HOUR = col_character(),
##   MINUTE = col_character(),
##   HUNDRED_BLOCK = col_character(),
##   NEIGHBOURHOOD = col_character(),
##   X = col_double(),
##   Y = col_double()
## )
```

```
head(crime)
```

```
## # A tibble: 6 x 10
##   TYPE   YEAR MONTH DAY   HOUR  MINUTE HUNDRED_BLOCK NEIGHBOURHOOD      X
##   <chr> <int> <chr> <chr> <chr> <chr>  <chr>         <chr>          <dbl>
## 1 Brea~  2003 08    09    00    40     15XX MARINER~ Fairview      4.90e5
## 2 Thef~  2003 02    05    22    00     47XX JOYCE ST Renfrew-Coll~ 4.98e5
## 3 Thef~  2003 12    22    08    47     47XX KILLARN~ Renfrew-Coll~ 4.97e5
## 4 Thef~  2003 03    24    22    00     47XX LANARK ~ Kensington-C~ 4.95e5
## 5 Othe~  2003 12    24    12    15     2X W HASTING~ Central Busi~ 4.92e5
## 6 Thef~  2003 11    17    22    30     47XX LITTLE ~ Kensington-C~ 4.95e5
## # ... with 1 more variable: Y <dbl>
```

```
summary(crime)
```

```
##      TYPE                YEAR         MONTH               DAY
##  Length:584053      Min.   :2003   Length:584053      Length:584053
##  Class :character   1st Qu.:2006   Class :character   Class :character
##  Mode  :character   Median :2010   Mode  :character   Mode  :character
##                     Mean   :2010
##                     3rd Qu.:2014
##                     Max.   :2018
##      HOUR              MINUTE          HUNDRED_BLOCK
##  Length:584053      Length:584053      Length:584053
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
##  NEIGHBOURHOOD            X                Y
##  Length:584053      Min.   :     0   Min.   :      0
##  Class :character   1st Qu.:489986   1st Qu.:5453726
##  Mode  :character   Median :491521   Median :5456880
##                     Mean   :442658   Mean   :4907171
##                     3rd Qu.:493510   3rd Qu.:5458662
##                     Max.   :511303   Max.   :5512579
```

What types of crime are present in the data?

```
crime %>%
  distinct(TYPE)
```

```
## # A tibble: 11 x 1
##    TYPE
##    <chr>
##  1 Break and Enter Residential/Other
##  2 Theft of Vehicle
##  3 Other Theft
##  4 Offence Against a Person
##  5 Theft from Vehicle
##  6 Mischief
##  7 Break and Enter Commercial
##  8 Theft of Bicycle
##  9 Vehicle Collision or Pedestrian Struck (with Fatality)
## 10 Vehicle Collision or Pedestrian Struck (with Injury)
## 11 Homicide
```

How many of each type of crime do we have in our dataset?

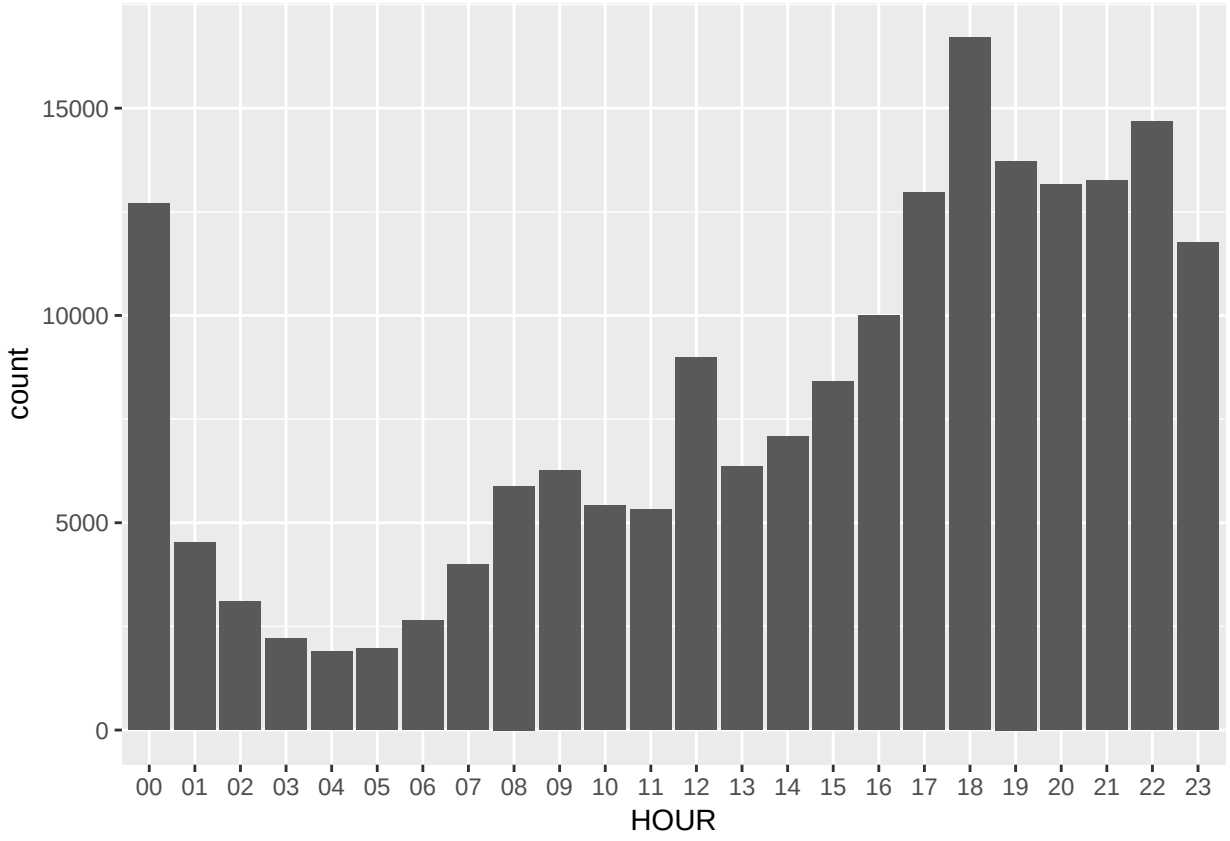
```
crime %>%
  group_by(TYPE) %>%
  summarize(counts = n()) %>%
  arrange(desc(counts))
```

```
## # A tibble: 11 x 2
##    TYPE                                                counts
##    <chr>                                                <int>
##  1 Theft from Vehicle                                  193009
##  2 Mischief                                             78418
##  3 Break and Enter Residential/Other                    64213
##  4 Other Theft                                          59376
##  5 Offence Against a Person                             58578
##  6 Theft of Vehicle                                     40236
##  7 Break and Enter Commercial                           36722
##  8 Theft of Bicycle                                     28970
##  9 Vehicle Collision or Pedestrian Struck (with Injury)    24015
## 10 Vehicle Collision or Pedestrian Struck (with Fatality)    276
## 11 Homicide                                               240
```

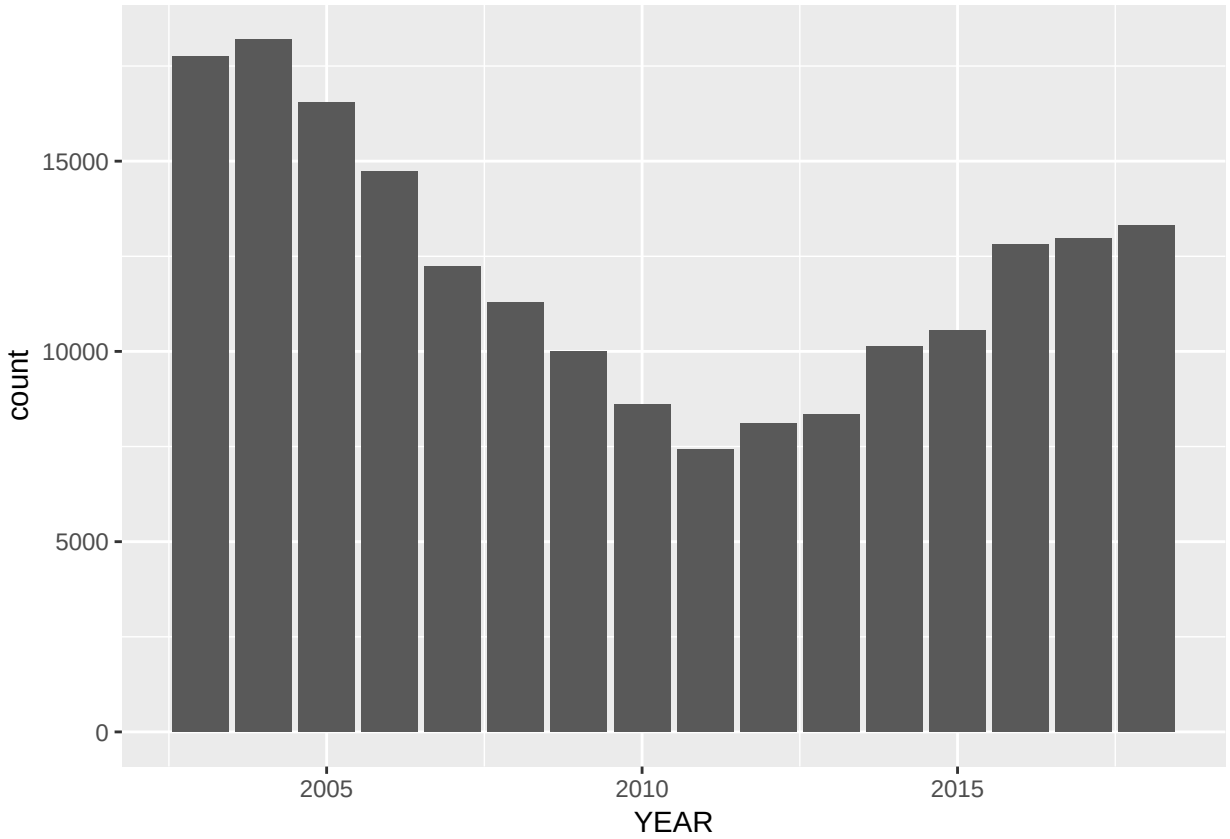
```
# Note: Theft of vehicle is by far the most.
```

Counts of car theft reported at each hour of the day

```
car_theft <- crime %>%
  filter(TYPE == "Theft from Vehicle")

car_theft %>%
  ggplot(aes(x=HOUR)) +
  geom_bar()
```




```
# Counts of thefts from cars for each year between 2003 and 2018*

car_theft %>%
  ggplot(aes(x=YEAR)) +
  geom_bar()
```



How many neighborhoods do we have in the dataset and how many thefts from cars happened in each?

```
car_theft %>%
  distinct(NEIGHBOURHOOD)
```

```
## # A tibble: 25 x 1
##    NEIGHBOURHOOD
##    <chr>
##  1 Riley Park
##  2 Grandview-Woodland
##  3 Sunset
##  4 Mount Pleasant
##  5 Kensington-Cedar Cottage
##  6 Central Business District
##  7 Hastings-Sunrise
##  8 Kitsilano
##  9 Strathcona
## 10 Renfrew-Collingwood
## # ... with 15 more rows
```

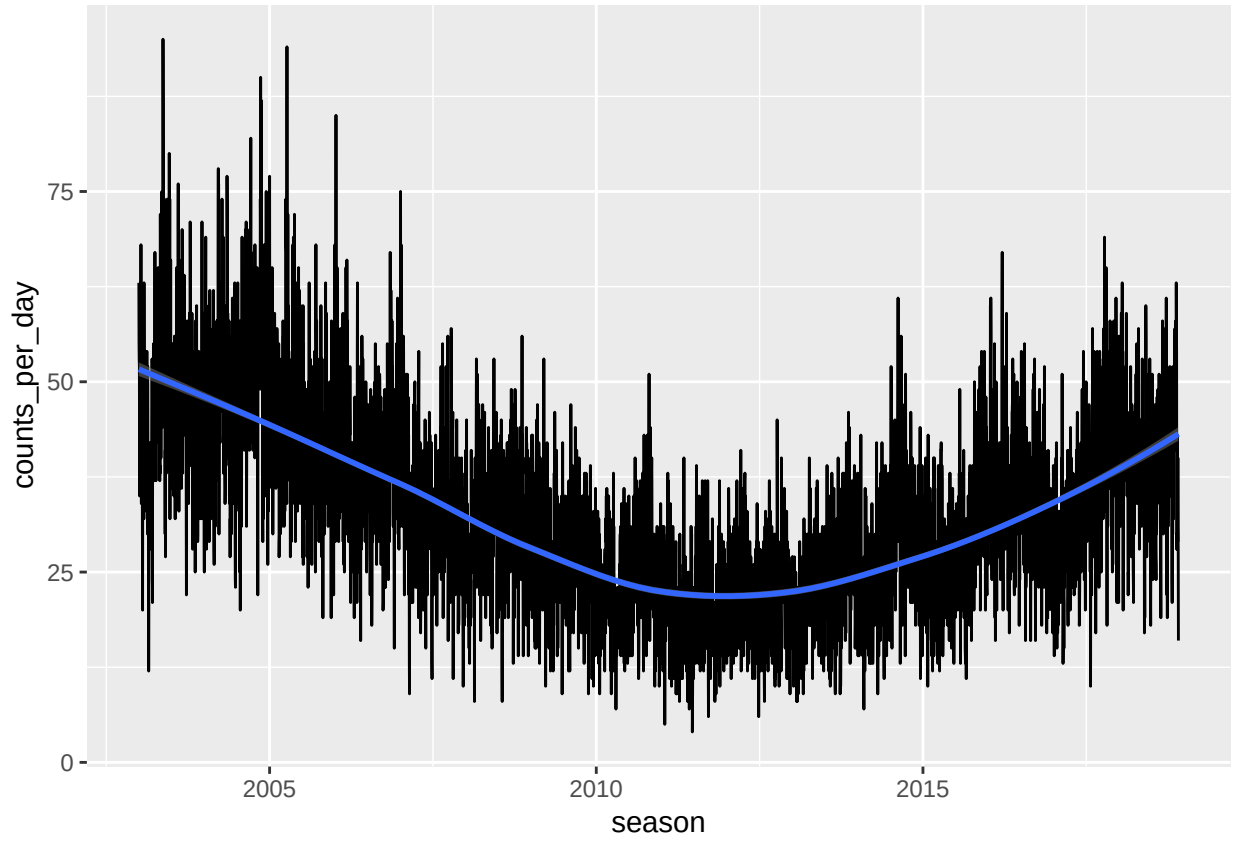
```
car_theft %>%
  group_by(NEIGHBOURHOOD) %>%
  summarize(count = n()) %>%
  drop_na(NEIGHBOURHOOD)
```

```
## # A tibble: 24 x 2
```

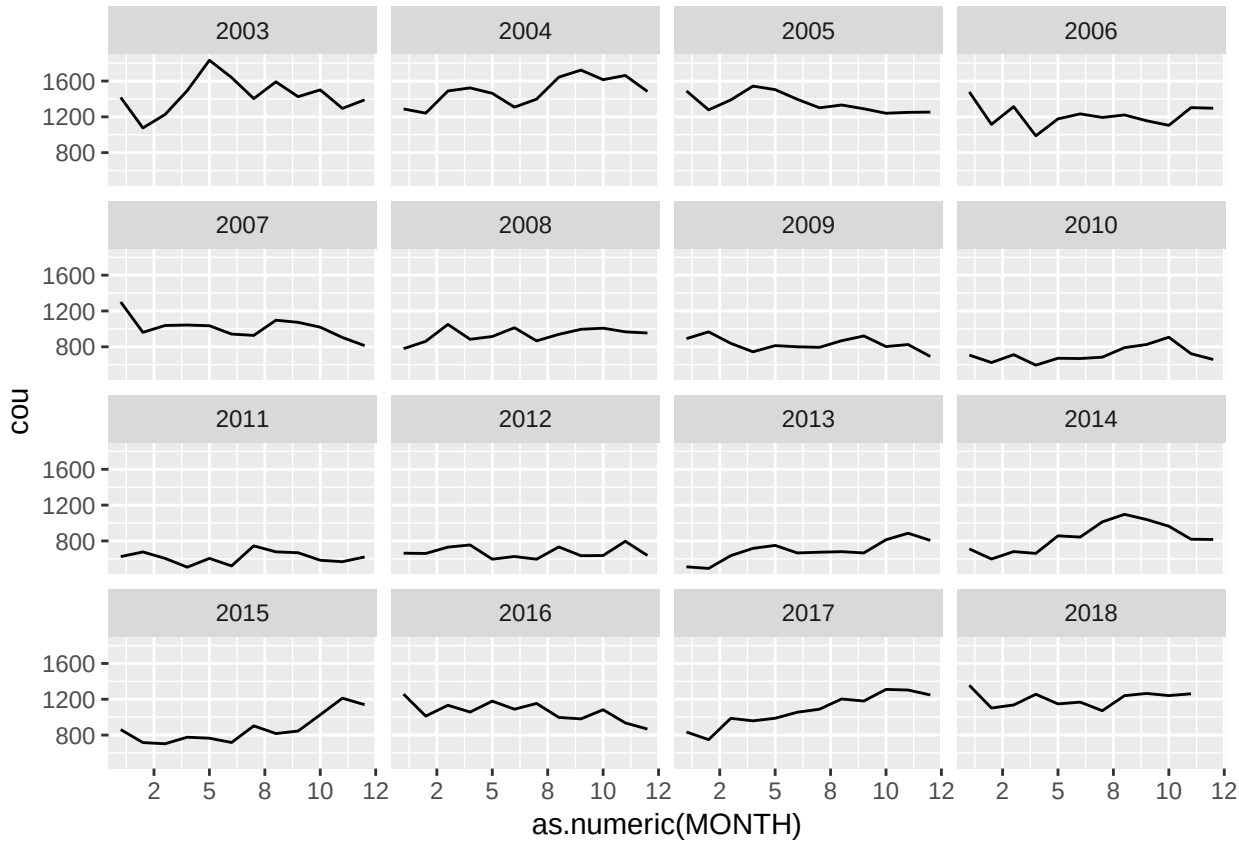
```
##    NEIGHBOURHOOD             count
##    <chr>                     <int>
##  1 Arbutus Ridge              2021
##  2 Central Business District 54892
##  3 Dunbar-Southlands          3182
##  4 Fairview                  12884
##  5 Grandview-Woodland         8300
##  6 Hastings-Sunrise           6459
##  7 Kensington-Cedar Cottage   8203
##  8 Kerrisdale                 3044
##  9 Killarney                  4343
## 10 Kitsilano                  9923
## # ... with 14 more rows
```

```r
car_theft <- car_theft %>%
  drop_na(NEIGHBOURHOOD)
```

```r
#class(car_theft$YEAR)

# Making date time column to look at seasonality

car_theft %>%
  mutate(season = lubridate::make_date(YEAR,MONTH,DAY)) %>%
  group_by(season) %>%
  summarize(counts_per_day = n()) %>%
  ggplot(aes(x=season,counts_per_day)) +
  geom_line() +
  geom_smooth(method="loess")
```

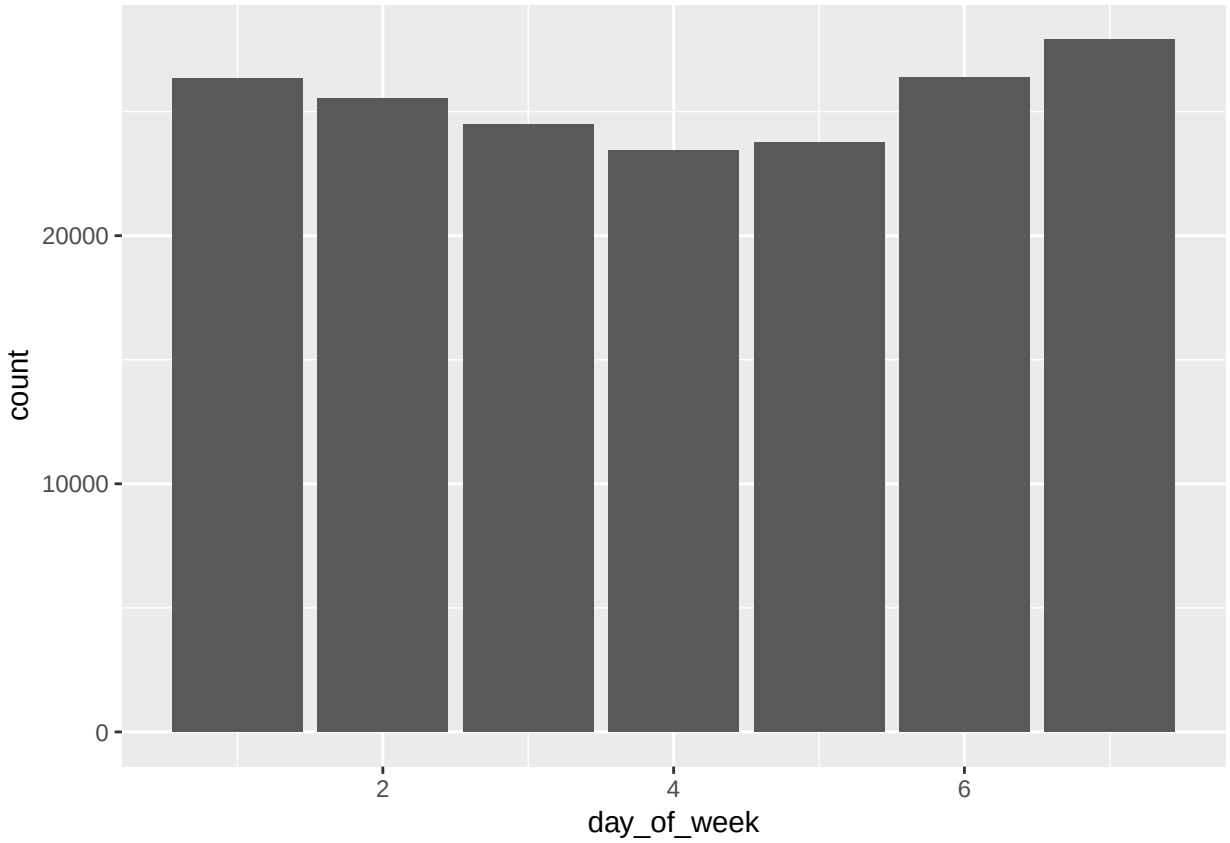



```
  #select(HUNDRED_BLOCK,NEIGHBOURHOOD,X,Y,dt)

  #car_dt

# Plotting counts of thefts from cars by month across all years

car_theft %>%
  group_by(YEAR,MONTH) %>%
  summarize(cou = n()) %>%
  ggplot(aes(as.numeric(MONTH),cou)) +
  geom_line() +
  facet_wrap(~YEAR) +
  scale_x_continuous(labels = scales::number_format(accuracy = 1))
```



Looking at data between 2003 and 2017 and looking at difference in the days of the week

```
car_theft %>%
  filter(!YEAR==2018) %>%
  mutate(datetime = lubridate::make_date(YEAR,MONTH,DAY)) %>%
  select(NEIGHBOURHOOD,datetime) %>%
  mutate(day_of_week = lubridate::wday(datetime)) %>%
  group_by(day_of_week) %>%
  summarize(count = n()) %>%
  ggplot(aes(day_of_week,count)) +
  geom_col()
```



Looking at data between 2003 and 2017 (omitting 2018 because it is incomplete), do we see any variation between summer and winter months. Looking at the plot above it does not appear as though there is any significant difference.

```
car_theft %>%
  filter(!YEAR==2018) %>%
  mutate(is_summer = if_else(MONTH %in% c("06","07","08"),TRUE,FALSE),is_winter = if_else(MONTH %in% c(
         is_spring = if_else(MONTH %in% c("03","04","05"),TRUE,FALSE),is_fall = if_else(MONTH %in% c("0
  count(is_summer,is_winter,is_fall,is_spring)
```

```
## # A tibble: 4 x 5
##   is_summer is_winter is_fall is_spring     n
##   <lgl>     <lgl>     <lgl>   <lgl>     <int>
## 1 FALSE     FALSE     FALSE   TRUE      43928
## 2 FALSE     FALSE     TRUE    FALSE     46486
## 3 FALSE     TRUE      FALSE   FALSE     42512
## 4 TRUE      FALSE     FALSE   FALSE     44943
```

Incredible. The number of car thefts is nearly the same across all the seasons. The lowest being in winter at 43000 and the highest being Fall at 47000. Over 14 years, that difference is nearly negligible.

Mapping theft from cars for 2004 in Vancouver using Leaflet

```
theft_04 <- car_theft %>%
  filter(YEAR == 2004) %>%
  select(NEIGHBOURHOOD,X,Y)

van_map <- leaflet() %>% setView(lat = 49.25,lng = -123.1,zoom=12) %>% addTiles()

#van_map

# Converting UTM coordinates to Lat and Long (Resource: http://rstudio-pubs-static.s3.amazonaws.com/200

utms <- SpatialPoints(theft_04[, c("X", "Y")],
                      proj4string=CRS("+proj=utm +zone=10"))

longlats <- spTransform(utms, CRS("+proj=longlat"))

# Plugging them back into the dataset

theft_04$X <- longlats$X

theft_04$Y <- longlats$Y

theft_04 %>%
  summary()
```

```
##  NEIGHBOURHOOD            X                Y
##  Length:17835       Min.   :-123.2   Min.   :49.20
##  Class :character   1st Qu.:-123.1   1st Qu.:49.25
##  Mode  :character   Median :-123.1   Median :49.27
##                     Mean   :-123.1   Mean   :49.26
##                     3rd Qu.:-123.1   3rd Qu.:49.28
##                     Max.   :-123.0   Max.   :49.31
```

```
van_map <- van_map
  #addMarkers(data=theft_04,~X,~Y,clusterOptions = markerClusterOptions())

leaflet(options = leafletOptions(preferCanvas = TRUE)) %>% setView(lat = 49.25,lng = -123.1,zoom=12) %>%
  addTiles() %>%
  #addTiles('http://{s}.basemaps.cartocdn.com/dark_all/{z}/{x}/{y}.png',
    #attribution='Map tiles by <a href="http://stamen.com">Stamen Design</a>, <a href="http://creativec
  addCircles(data=theft_04,~X,~Y,color = "red",radius=0.01)
```

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

Mapping theft from cars in Vancouver using ggplot2

```
#Reading in Vancouver neighborhood boundary data
vancouver <- readOGR("cov_localareas.kml",layer = "local_areas_region")
```

```
## OGR data source with driver: KML
```

```
## Source: "/Users/mohamadmakkaoui/Desktop/Code/van_car_theft_vis/cov_localareas.kml", layer: "local_ar
## with 22 features
## It has 2 fields
```

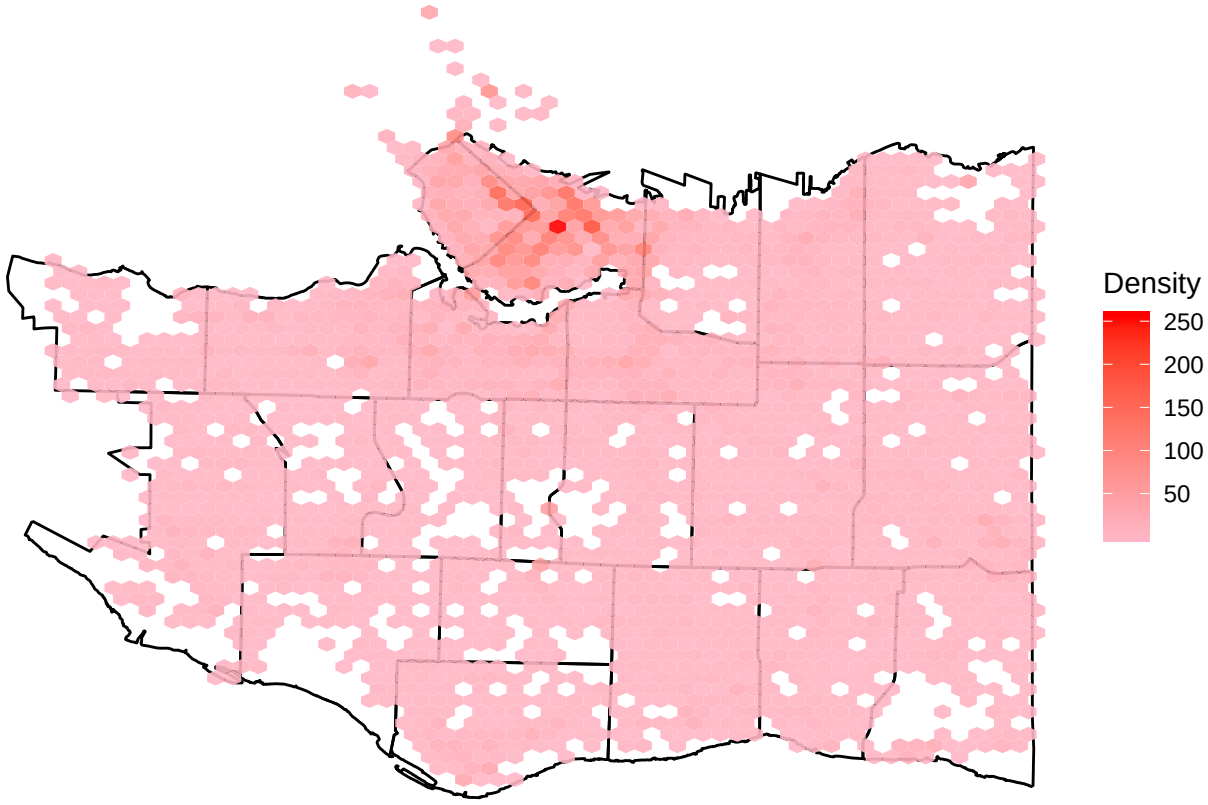
```
# Converting map object into dataframe
van_area <- fortify(vancouver)
```

```
## Regions defined for each Polygons
```

```
head(van_area)
```

```
##        long      lat order  hole piece id group
## 1 -123.1641 49.25748     1 FALSE     1  0   0.1
## 2 -123.1639 49.25746     2 FALSE     1  0   0.1
## 3 -123.1636 49.25745     3 FALSE     1  0   0.1
## 4 -123.1626 49.25743     4 FALSE     1  0   0.1
## 5 -123.1603 49.25740     5 FALSE     1  0   0.1
## 6 -123.1579 49.25736     6 FALSE     1  0   0.1
```

```
ggplot() +
  geom_path(data = van_area,aes(long,lat,group=group)) +
  geom_hex(data = theft_04,aes(X,Y),bins=60,alpha=0.9) +
  scale_fill_gradient(low="pink1", high="red", name="Density") +
  theme_void()
```



Making a choropleth map of Vancouver

Creating our dataset with counts per neighborhood

```
theft_counts <- theft_04 %>%
  group_by(NEIGHBOURHOOD) %>%
  summarize(n=n())

head(theft_counts)
```

```
## # A tibble: 6 x 2
##   NEIGHBOURHOOD                n
##   <chr>                    <int>
## 1 Arbutus Ridge              207
## 2 Central Business District 4418
## 3 Dunbar-Southlands          419
## 4 Fairview                  1295
## 5 Grandview-Woodland         794
## 6 Hastings-Sunrise           602
```

It appears as thought there is some descrepancy between the polygon dataset and the crime dataset when it comes to neighborhood names. Luckily, most of them are correct and will be joinable. The ones that aren't will be merged using the `aggregate` function.