Exploratory Data Analysis on Crime in Austin, Texas

Maya Joiner

2022-12-22

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.4.0      ✔ purrr   0.3.5
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Introduction

In this notebook, I will analyze different aspects of data on crime in Austin, Texas. To get the csv files I import in this notebook, I used BigQuery (SQL) to filter/group data/make new columns. Here are the code snippets I used to make the dataframes:

1) Analyzing incident distribution by district

SELECT district, COUNT(unique_key) AS number_crimes FROM `bigquery-public-data.austin_crime.crime` GROUP BY district;

2) Type of incident by district

SELECT district, latitude, longitude, primary_type FROM `bigquery-public-data.austin_crime.crime` WHERE latitude IS NOT NULL;

3) How long it took for a case to be cleared by district

SELECT district, EXTRACT(HOUR from clearance_date-timestamp) AS difference_hours, timestamp, clearance_date FROM `bigquery-public-data.austin_crime.crime`

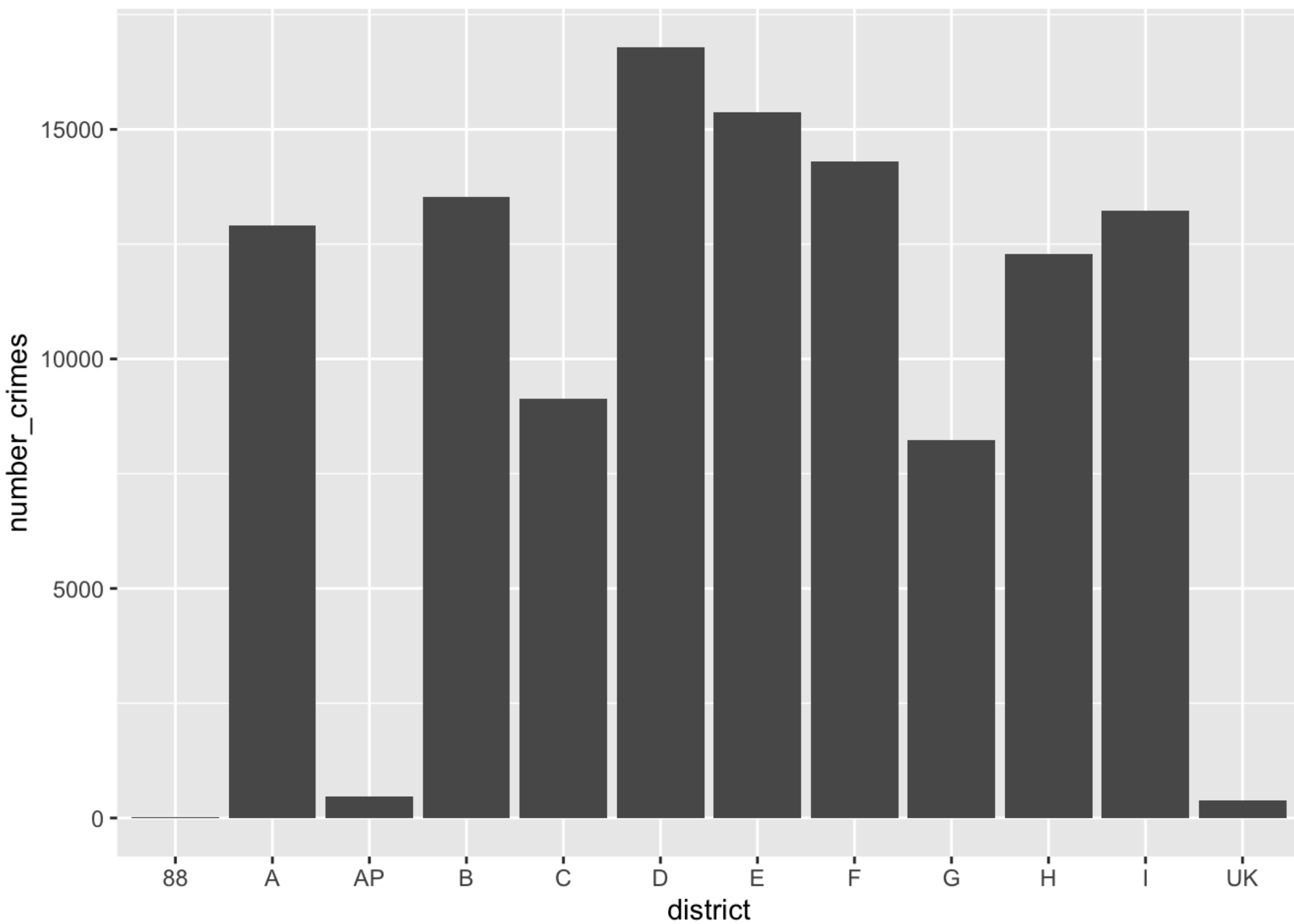
4) Incident count over time by district

SELECT district, EXTRACT(YEAR from timestamp) AS crime_year, COUNT(unique_key) AS number_crimes FROM `bigquery-public-data.austin_crime.crime` GROUP BY district, crime_year ORDER BY district, crime_year;

Part 1: Analyzing incident distribution by district

```
df1 = read.csv("crime_counts_by_district.csv")
df1
```

```
##    district number_crimes
## 1        UK           378
## 2         I         13222
## 3         H         12294
## 4         A         12915
## 5        88            26
## 6         B         13520
## 7         D         16794
## 8         F         14311
## 9         C          9142
## 10        G          8229
## 11        E         15383
## 12       AP           460
```



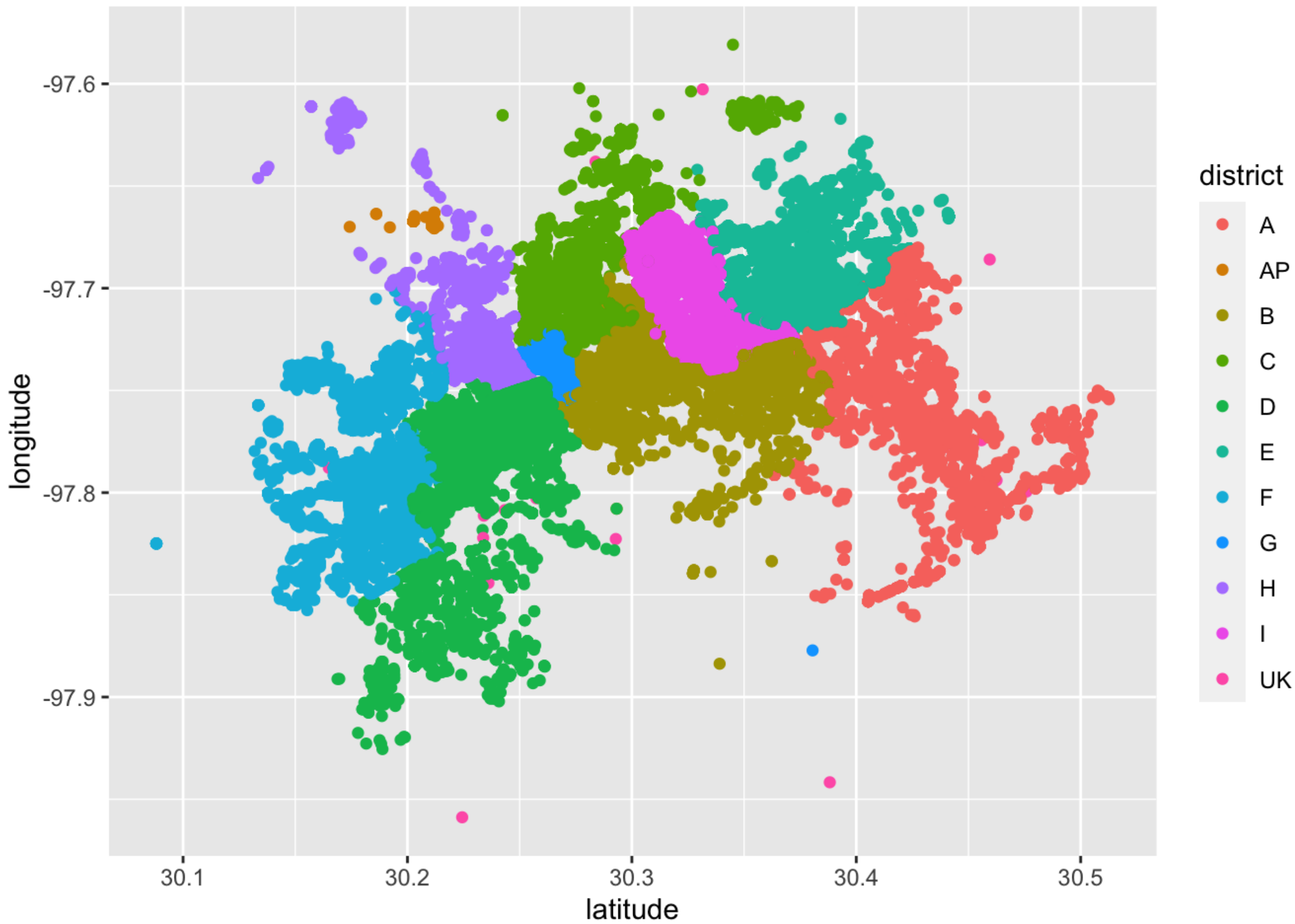
We can see that 88, AP, and UK have a very low number of crimes, followed by C and G. The other districts have about the same number of crimes, D being the highest.

Part 2: Type of incident by district

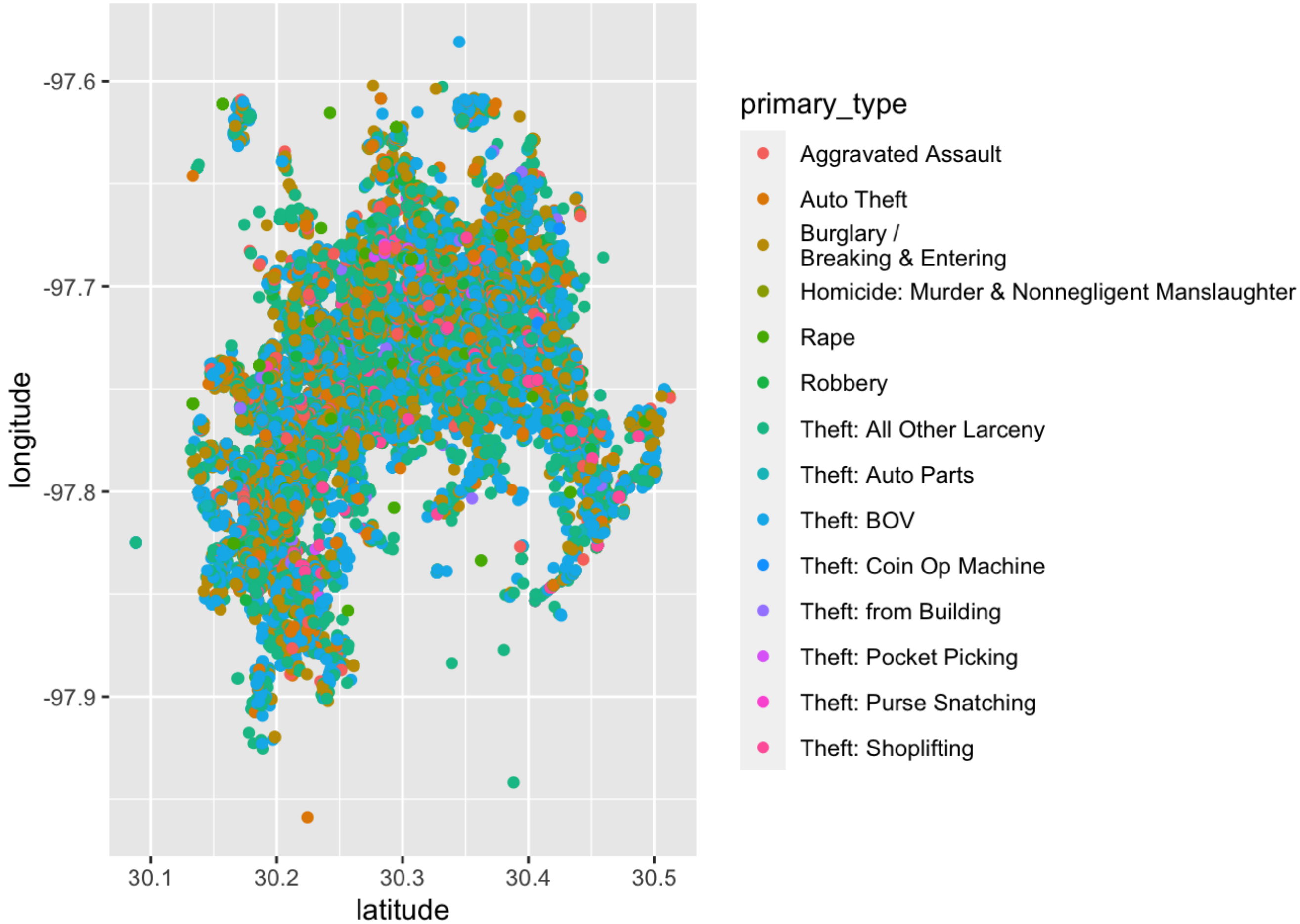
```
df2 = read.csv("latlong.csv")
head(df2)
```

```
##   district latitude longitude primary_type
## 1        G 30.26498  -97.7466         Rape
## 2       UK 30.26498  -97.7466         Rape
## 3        G 30.26498  -97.7466         Rape
## 4        A 30.26498  -97.7466         Rape
## 5        B 30.26498  -97.7466         Rape
## 6        G 30.26498  -97.7466         Rape
```

```
ggplot(data = df2, mapping=aes(x = latitude, y = longitude, color=district) ) + geom_poi
nt()
```

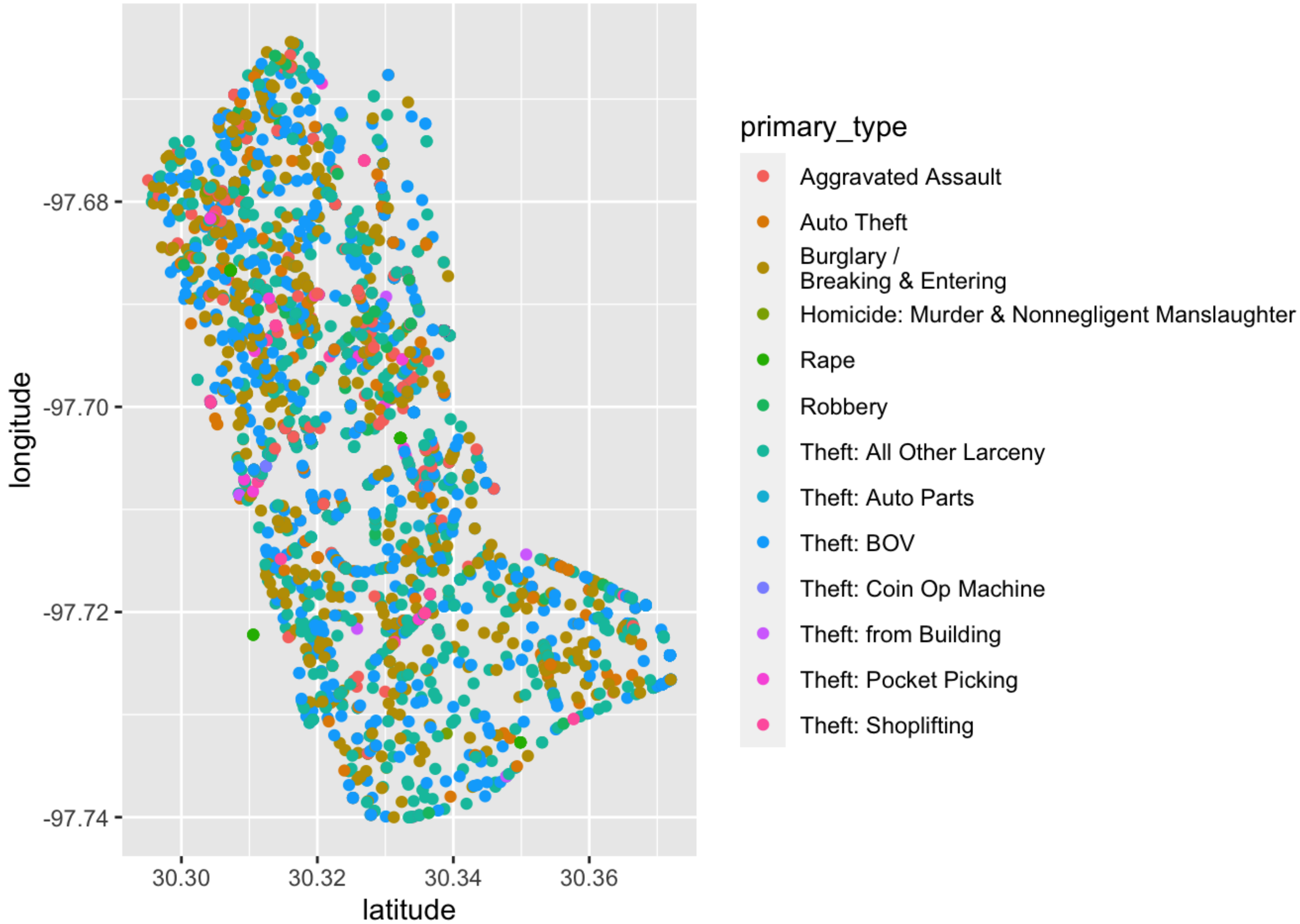



```
ggplot(data = df2, mapping=aes(x = latitude, y = longitude, color=primary_type) ) + geom
_point()
```



Although it seems like all districts have a lot of burgulary/BOV theft, we are not 100% sure because of overplotting. To resolve this, we could jitter the data or make the points smaller/have a different shape. However, since I think this graph will still be overplot even when jittered and the points are already small, I will focus on one district of the map.

```
ggplot(data = df2[df2$district == "I", ], mapping=aes(x = latitude, y = longitude, color
=primary_type) ) + geom_point()
```



At least for district I, theft/burglary are the most common crimes.

Part 3: How long it took for a case to be cleared by district

```
df3 = read.csv("time_taken_to_clear.csv")
sum(is.na(df3)) / nrow(df3[complete.cases(df3),])
```

```
## [1] 0.05145646
```

Since the number of missing rows are not that significant (only 5% of all rows are missing), we will just ignore these rows. However, we still need to analyze if there is a pattern between the data that is missing.

Missing Analysis

```
cases <- df3[ , c("district", "difference_hours")]
grouped <- cases %>% group_by(district) %>% summarize(count=n())
all_count <- grouped$count
```

```
df3notna <- df3[complete.cases(df3),]
head(df3notna)
```

```
##   district difference_hours                       timestamp
## 1       UK             1320 2016-01-19 12:00:00.000000 UTC
## 2       UK             1584 2016-01-25 12:00:00.000000 UTC
## 3       UK             5640 2016-01-25 12:00:00.000000 UTC
## 4       UK             1200 2016-02-01 12:00:00.000000 UTC
## 5       UK             1128 2016-02-25 12:00:00.000000 UTC
## 6       UK             1320 2016-03-10 12:00:00.000000 UTC
##                    clearance_date
## 1 2016-03-14 12:00:00.000000 UTC
## 2 2016-03-31 12:00:00.000000 UTC
## 3 2016-09-16 12:00:00.000000 UTC
## 4 2016-03-22 12:00:00.000000 UTC
## 5 2016-04-12 12:00:00.000000 UTC
## 6 2016-05-04 12:00:00.000000 UTC
```

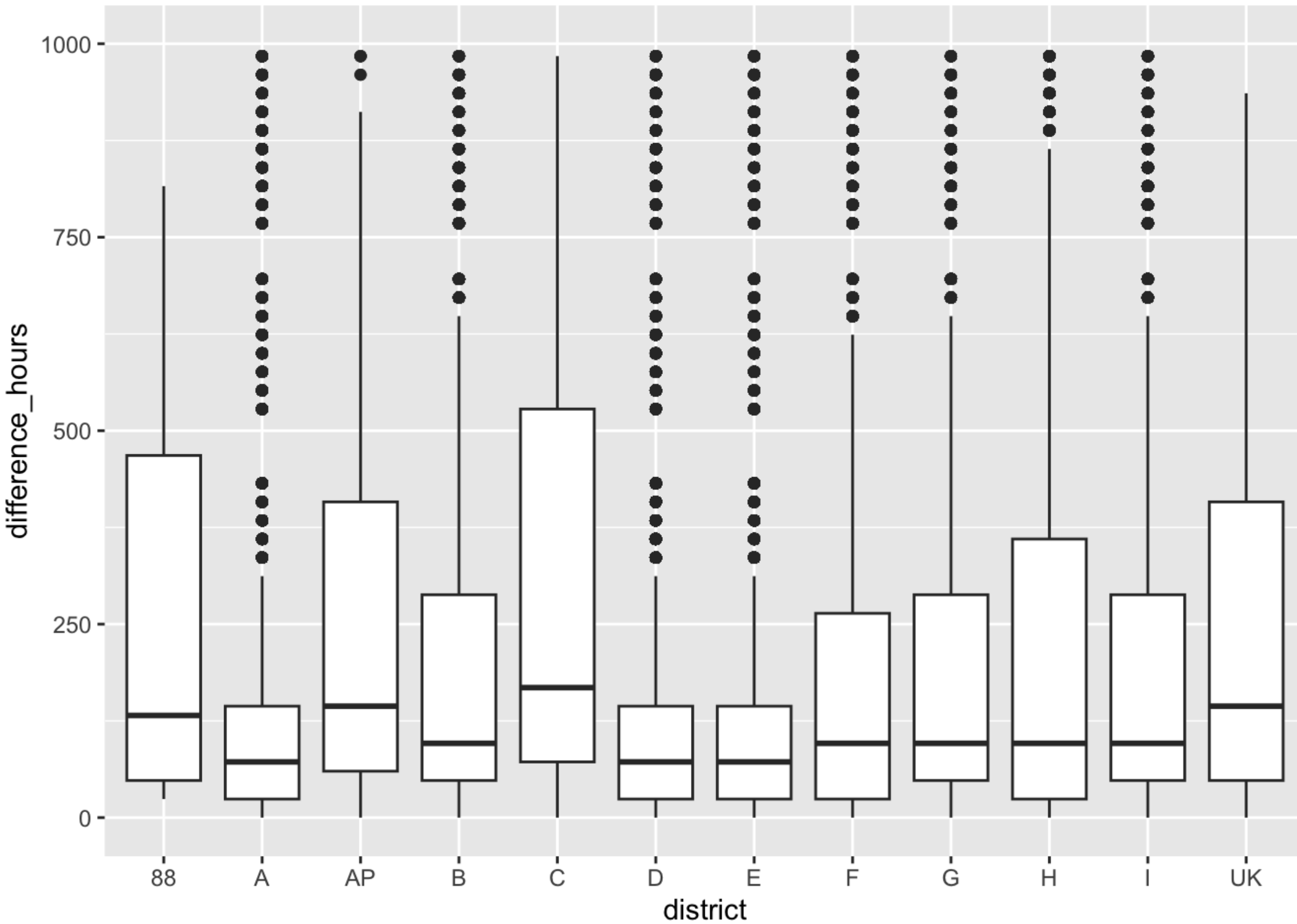
```
missingdf <- df3[!complete.cases(df3), c("district", "difference_hours")]
missing <- missingdf %>% group_by(district) %>% summarize(count=n())
missing_count <- missing$count
district <- missing$district
```

```
percent_missing <- (missing_count / all_count)*100
counts <- data.frame(district, missing_count, all_count, percent_missing)
```

We need to take into account the counts of the missing values for each district when we make the following barplot:

```
ggplot(data = df3notna, mapping=aes(x = district, y = difference_hours) ) + geom_boxplot
() + ylim(0, 1000)
```

```
## Warning: Removed 14571 rows containing non-finite values (`stat_boxplot()`).
```



The medians are all about the same. Although the bars for some districts are larger than others, this might be due to a lack of data. Districts like A, D, and E seem to have a small gap between the time the crime occurred and when the case was cleared while the time for 88, AP, C, and UK are relatively high. However, for these four districts, either a large percent of the data is missing or there isn't much data to start with – therefore, the lengths of the bars may not be as accurate and may just be fluctuations.

Part 4: Incident count over time by district

```
df4 = read.csv("crime_count_over_time.csv")
df4
```

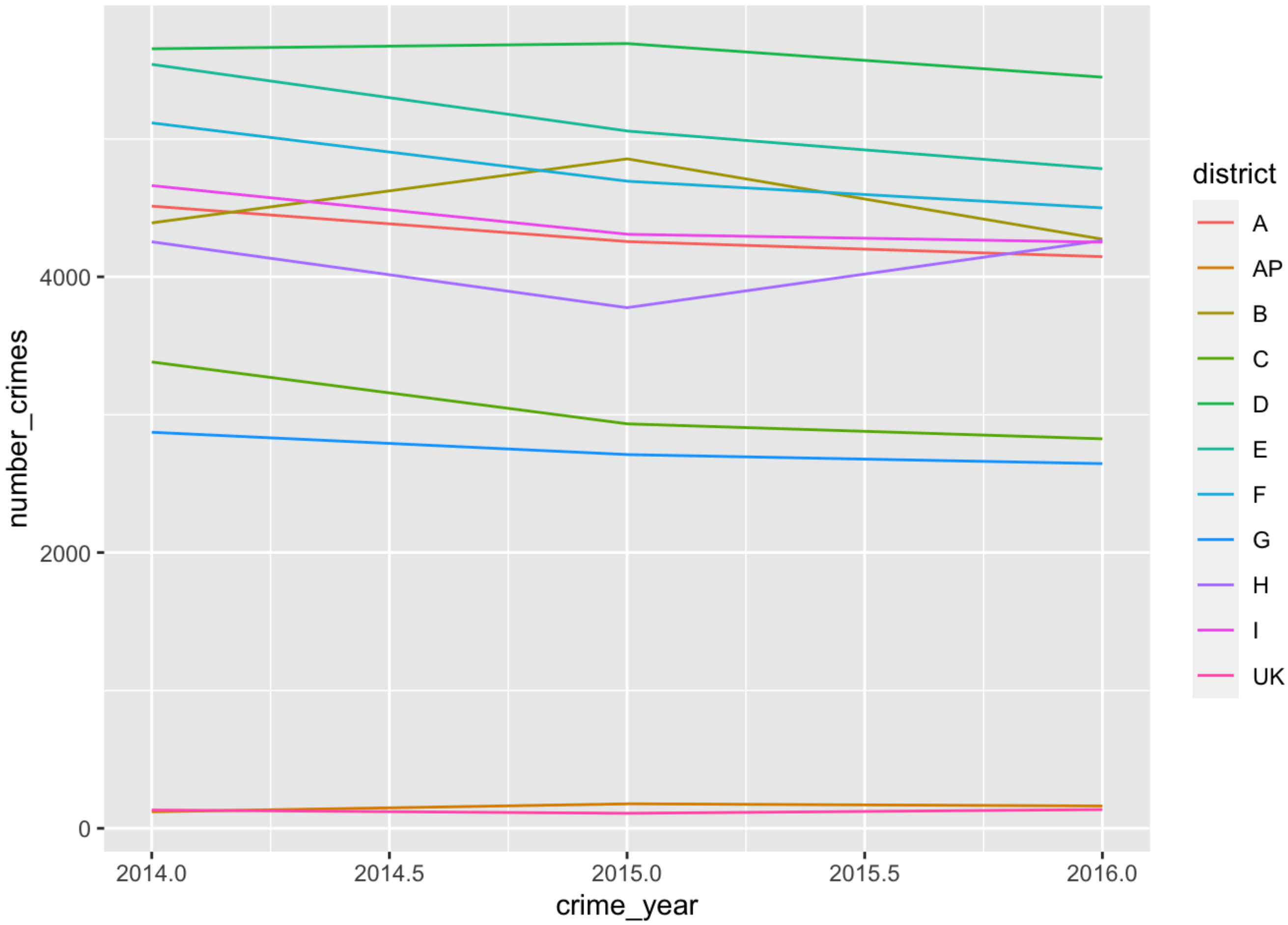
```
##    district crime_year number_crimes
## 1        88       2016            26
## 2         A       2014          4513
## 3         A       2015          4256
## 4         A       2016          4146
## 5        AP       2014           120
## 6        AP       2015           178
## 7        AP       2016           162
## 8         B       2014          4391
## 9         B       2015          4856
## 10        B       2016          4273
## 11        C       2014          3383
## 12        C       2015          2934
## 13        C       2016          2825
## 14        D       2014          5654
## 15        D       2015          5692
## 16        D       2016          5448
## 17        E       2014          5541
## 18        E       2015          5058
## 19        E       2016          4784
## 20        F       2014          5117
## 21        F       2015          4694
## 22        F       2016          4500
## 23        G       2014          2873
## 24        G       2015          2711
## 25        G       2016          2645
## 26        H       2014          4254
## 27        H       2015          3776
## 28        H       2016          4264
## 29        I       2014          4662
## 30        I       2015          4309
## 31        I       2016          4251
## 32       UK       2014           133
## 33       UK       2015           109
## 34       UK       2016           136
```

For this part, we will remove district 88 because it only has information on one year.

```
df4 <- df4[-1, ]
df4
```

```
##    district crime_year number_crimes
## 2         A       2014          4513
## 3         A       2015          4256
## 4         A       2016          4146
## 5        AP       2014           120
## 6        AP       2015           178
## 7        AP       2016           162
## 8         B       2014          4391
## 9         B       2015          4856
## 10        B       2016          4273
## 11        C       2014          3383
## 12        C       2015          2934
## 13        C       2016          2825
## 14        D       2014          5654
## 15        D       2015          5692
## 16        D       2016          5448
## 17        E       2014          5541
## 18        E       2015          5058
## 19        E       2016          4784
## 20        F       2014          5117
## 21        F       2015          4694
## 22        F       2016          4500
## 23        G       2014          2873
## 24        G       2015          2711
## 25        G       2016          2645
## 26        H       2014          4254
## 27        H       2015          3776
## 28        H       2016          4264
## 29        I       2014          4662
## 30        I       2015          4309
## 31        I       2016          4251
## 32       UK       2014           133
## 33       UK       2015           109
## 34       UK       2016           136
```

```
ggplot(data = df4, mapping=aes(x = crime_year, y = number_crimes, color=district) ) + ge
om_line()
```



Overall, there seems to be not much of a change in number of crimes over the three years (2014-2016). Most districts' lines have a slope of 0 indicating no change in number of crimes.