Crime Information Report on the District Of Columbia in 2017(Washington DC)

Logan Salinas

August 14, 2025

1 Background

The research that I will be conducting will revolve around the crimes committed as well as the crimes reported in the United States Capital, Washington DC. I will be responsible on identifying any trends or patterns that is within data set containing crime reports in Washington DC. Specifically, I look into the type of offense that was committed and when the crime was reported. I'm hoping by the end of my research, I'll be able to answer the to the points mentioned above and give people a good understanding of where in D.C. the most crimes are committed and hopefully help them avoid those areas if they were ever to come visit Washington DC one day.

1.1 Description of Data

The data that I'll be using today is a csv file called `Crime_incidents_in_2017.csv`. I got this data from data.gov which is a United States open government website. This website allows the public to access data regarding anything that has to something or is associated with the U.S. government. The data that I'm using is information regarding crime reports in Washington D.C. in the year 2017. Some important things on the csv file are the variables and observations. The data that I'm handling here has around 33,000 rows of data/observations and 27 variables/columns. So I have to be extra careful when I utilize data manipulation.

1.1.1 Data Dictionary

- **REPORT_DAT:** The Date in time in which the crime was reported.
- **SHIFT:** The police shift during which the crime occurred.
- **OFFENSE:** Type of crime committed based on the report.
- **LONGITUDE/LATITUDE:** The coordinates of where the crime happened.
- **Month:** The month in which the crime was reported.
- **Hour:** The hour in which the crime was reported.

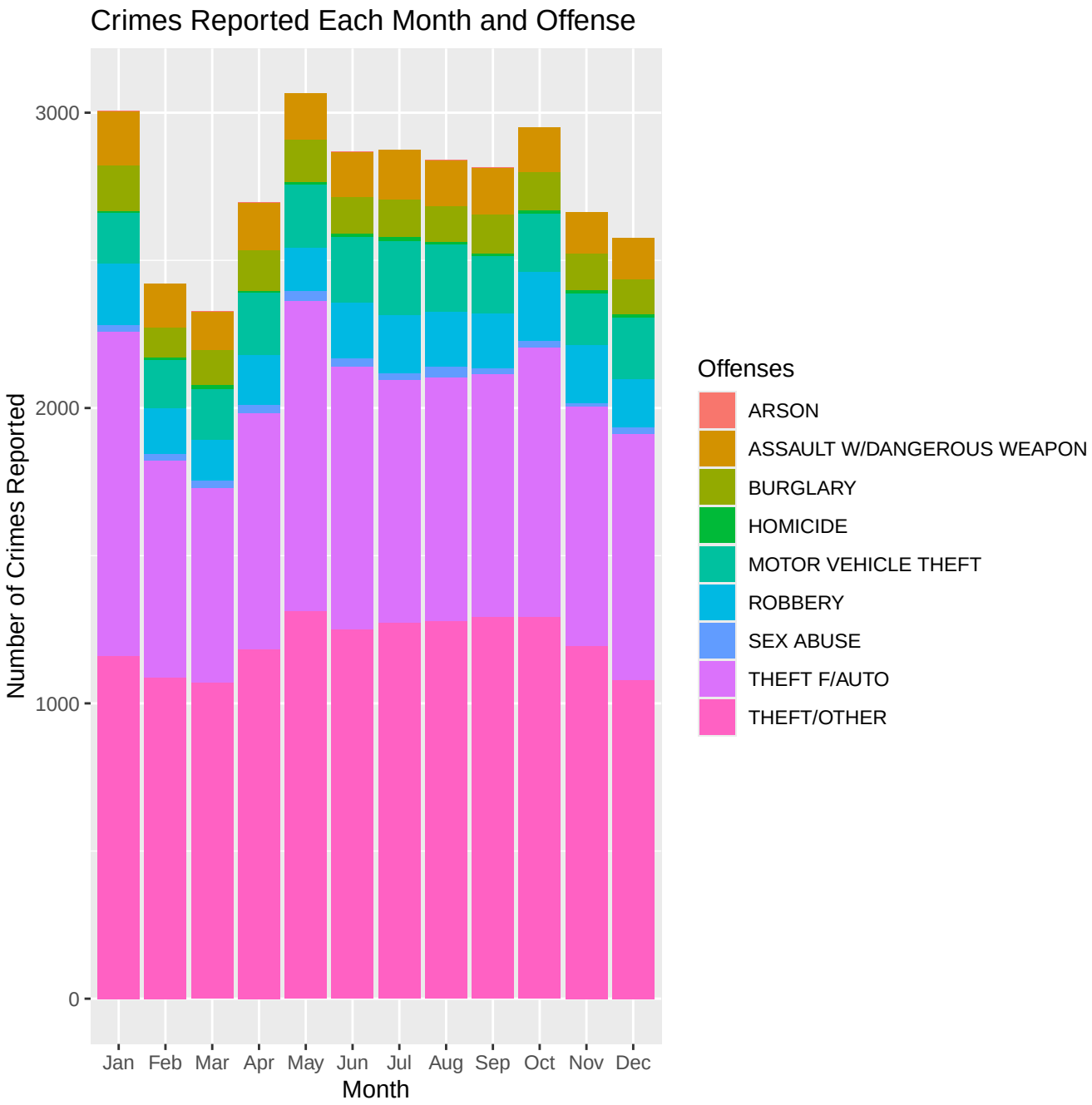
1.2 Summary Table of Crimes Reported Each Month per Offense (Bivariate Aggregation)

We are exploring the the possible relationship between Month in which the crime was reported and the type of offense in the report itself, the two variables `Month` and `OFFENSE` are categorical. By using the `Crime_Incidents_in_2017.csv` data set, we are able to display a table of the type of offenses that was reported in each month and identify a pattern if certain type of offenses are being committed more than others on a certain month.

Table 1: Total Amount of Crimes Reported Each Month per Offense

Month	ARSON	ASSAULT W/DANGEROUS WEAPON	BURGLARY	HOMICIDE	MOTOR VEHICLE THEFT	ROBBERY	SEX ABUSE	THEFT F/AUTO	THEFT/OTHER
Jan	1	183	154	8	170	209	22	1098	1160
Feb	0	148	104	6	165	155	22	734	1087
Mar	0	131	117	14	173	138	24	661	1068
Apr	2	160	138	7	209	170	28	799	1183
May	0	158	141	8	214	148	34	1051	1310
Jun	1	154	124	11	222	188	29	889	1250
Jul	0	169	128	12	250	197	24	823	1271
Aug	1	157	120	8	228	187	35	826	1278
Sep	0	160	130	10	193	187	20	821	1292
Oct	0	152	130	11	198	231	24	912	1292
Nov	0	141	125	9	176	195	14	809	1194
Dec	0	138	119	11	209	164	21	833	1079

The bar plot that is going to be displayed below will give us an idea of how many times a certain offense was reported at a certain month. I believe that using a bar plot here is appropriate because it will give us a plot of results for the number of crimes reported each month, and the amount of offenses in each of those reports.



1.2.1 Discussion

The table gave us an idea of how many type of offenses were reported for every month of the year 2017. By the looks of it, one of the type of crimes, arson, is rarely reported. The results of the stacked bar plot show that the majority of crimes that were reported in Washington DC were crimes related to theft, and auto theft. There isn't a month that stands out by a huge margin where I can say that a particular month is known for having the highest amount of reports on a certain offense. But what I can say is that theft and auto theft offenses increased during the middle of the year (May, June, July). All other offenses are around similar amounts for each month.

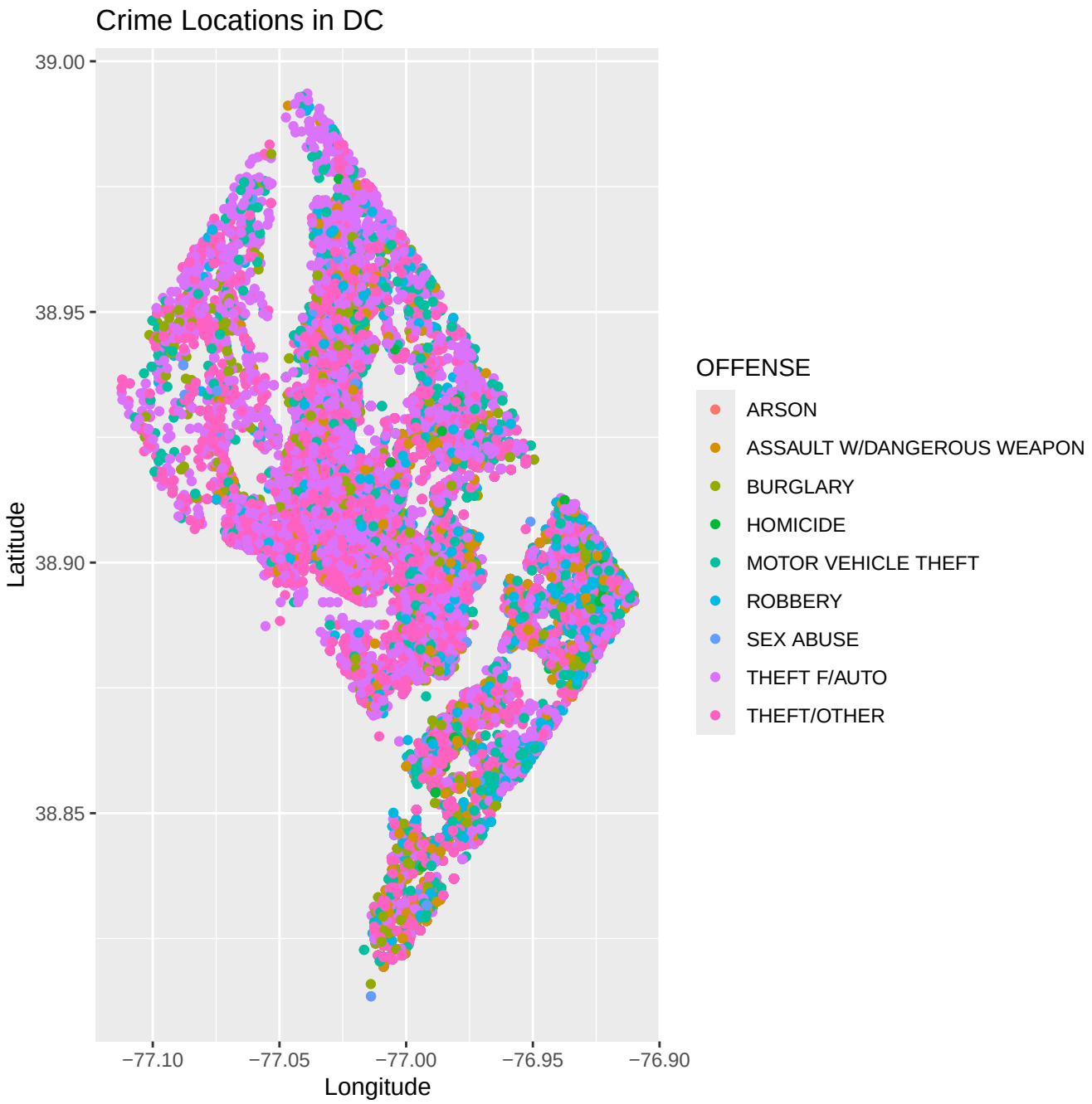
2 Geographic Summary of Crime Coordinates (Univariate Aggregation)

What we are doing is using the `longitude` and `latitude` variables to find where exactly these crimes are being reported in Washington DC as well as the type of crime that took place in that area. For the univariate summary table, I display the minimum, maximum, mean, median, and standard deviation values of the longitude and latitude. These values will help is in finding where exactly in Washington DC does reported crime happen the most in.

Table 2: Univariate Summary Table of Longitude and Latitude

name	value
longitude_min	-77.1123165
longitude_max	-76.9100205
longitude_mean	-77.0069839
longitude_median	-77.0106764
longitude_sd	0.0363280
latitude_min	38.8134705
latitude_max	38.9935598
latitude_mean	38.9064802
latitude_median	38.9060616
latitude_sd	0.0302385

For this plot, I ended up doing a scatter plot, the reasoning beheind this is that we are dealing with longitude and latitude values. We can treat the scatter plot grid as like a map and use the variables stated previous to pin point the where exactly what type of crime was report. What's very interesting about the scatter plot is that it outline the boarders of Washington D.C..



2.0.1 Discusion on Scatter plot

From the table, we were able to obtain the summary statistics of the longitude and latitude variables. Nothing really stood out in the table. If we took the mean of the longitude and latitude and put them together, then we would get the average coordinates of where crimes are reported.

As stated before, the points on the scatter plot outline the boarder of Washington DC. And based off the amount of color points plotted on the scatter plot, we can say that the majority crimes that happen in the central part of D.C. is Theft, and behind Theft is auto Theft and Robbery. This information helps further prove our question above about certain offenses being reported more in certain parts of D.C. compared to others.

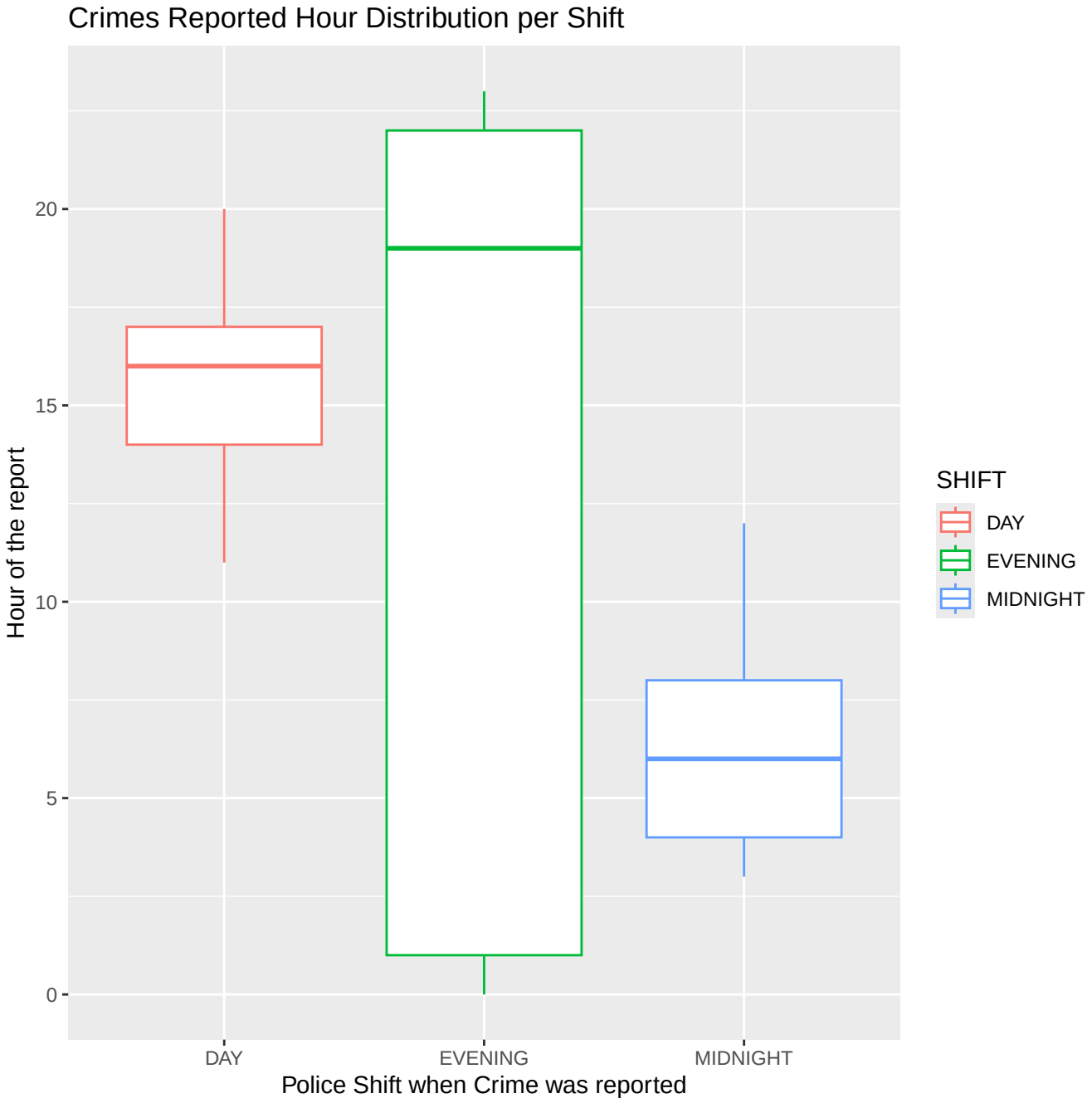
3 Summary Table of Hours For Each Shift (Bivariate Aggregation)

For this table and plot we are exploring the relationship between the hours that a report was made and the police shift during which the crime was reported. We use the `crime_incidents_df` data set to find the minimum, lower quantile, median, upper quantile, maximum, mean, and standard deviation of the hour of the report for every shift. This table will give us a better understanding of crimes being reported at certain hours of the day and the shift that documents these crimes.

Table 3: Bivariate Summary Table of Hours per Shift

SHIFT	min_hour	lower_quantile_hour	median_hour	upper_quantile_hour	max_hour	mean_hour	sd_hour
DAY	11	14	16	17	20	15.543717	2.181546
EVENING	0	1	19	22	23	12.454643	9.990945
MIDNIGHT	3	4	6	8	12	6.181396	2.357499

For this part, I decided to use a bar plot because the use of a bar plot would help us find the hours crimes that are reported during a certain police shift.



3.0.1 Discussion on Box plot

The table gives us information regarding when crimes are reported during certain hours of the day for every shift. One issue that I have with the table is the min_hour for the evening shift. I asked myself how is crime reported at the early hours of the morning considered a report done during the evening shift?

As we can see in the box plot, crimes reported during the day shift are usually reported during the afternoon hours, crimes reported during the evening shift are usually reported almost all hours of the day, and crimes reported during the midnight

shift are usually reported during the early hours of the morning. With this information, we know have an idea of what hours of the day are reported to each shift.

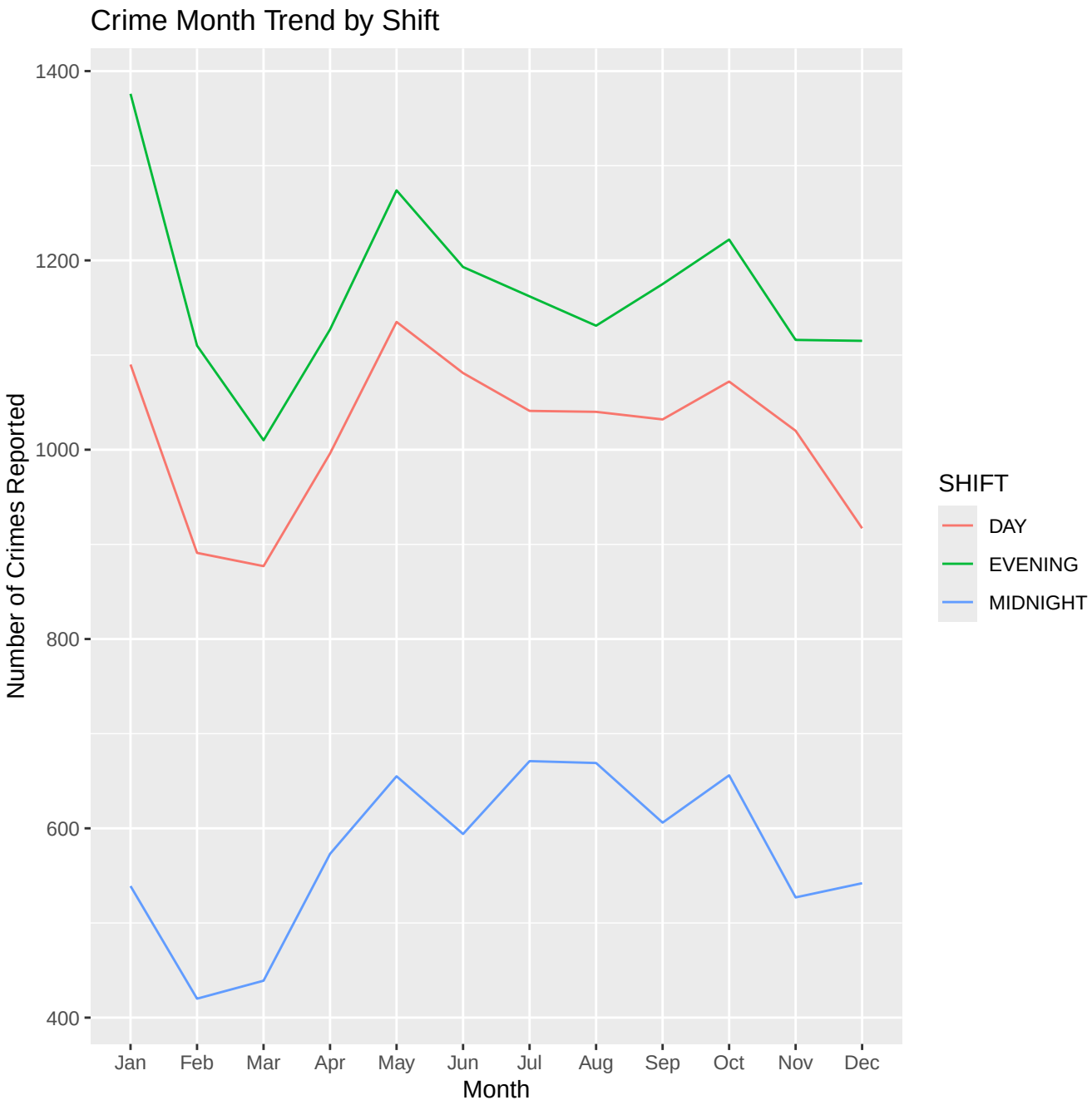
4 Summary Table of Number of Reports for Each Shift (Bivariate Aggregation)

Finally, we are exploring the the relationship between the months and the number of crimes reported for each police shift. We first use `crime_incidents_df` data set to find the summary statistics of the number of crimes reported per shift. This information will give us an idea of which shifts had the most or least number of crimes reported. It will also help us visualize the trend for crimes reported per shift based on the months.

Table 4: Bivariate Summary Table of Reports per Shift

SHIFT	min_reports	max_reports	mean_reports	median_reports	sd_reports
DAY	877	1135	1016.000	1036.0	81.80798
EVENING	1010	1376	1167.583	1146.5	92.93933
MIDNIGHT	420	671	574.250	583.5	85.03596

I decided to use a plot trend line for this part because I believed that using the plot trend line method would help me answer the question of "what is the monthly crime trend by shift?" based on the number of crimes reported. The plot trend line method would be able to utilize all three variables to answer that question.



4.0.1 Discussion

From the information of the summary table, Evening shifts receive the majority amount of crime reported on average, and have received the highest amount of crimes reported for one month compared to all months for other shifts. The trend line graph backs up the table information. As we can see in the plot trend line, the line, that represents the trend of crimes reported per shift, that is green(Evening shift) is higher on the graph then the red(Day Shift) or blue(Midnight Shift) line.

5 Proposed Research Method and Questions

1. One question that I want to find the answer to is `if we can predict the type of offenses based on the month, hour, and coordinates of the crime report given to us?` Since we are given a data of crime reports in Washington DC, we could use variables in the data frame and the entries associated with those variables as training data. We train the model to understand the relationship between pair variables and try to predict the type of offenses from new crime reports, which would contain the month, hour, longitude and latitude. Since we are trying to predict the type of offense that is going to be committed based on a couple of variables, the statistical learning technique I am going to use will be supervised learning. Utilizing this technique will allow me to learn more about the relationship between the time/location and the type of offense in each report in the data set. This will help me create and train a

model that can make predictions on the type of offenses based on the time and location mentioned in the crime report, assuming the report had a time and location written, but not the type of offense committed.

```
Call:
multinom(formula = OFFENSE ~ Hour + LONGITUDE + LATITUDE, data = crime_train,
    family = binomial)

Coefficients:
                          (Intercept)        Hour LONGITUDE    LATITUDE
ASSAULT W/DANGEROUS WEAPON   750.08633 -0.08536744  6.024605  -7.1736749
BURGLARY                     -15.06984 -0.05356940 -2.310996  -4.0155403
HOMICIDE                     109.02404 -0.22044783 -6.493457 -15.5213464
MOTOR VEHICLE THEFT          353.08625 -0.04567103  6.203280   3.3815174
ROBBERY                      161.39056 -0.08961609  2.052821   0.1048833
SEX ABUSE                     71.42175 -0.06241796 -1.727118  -5.1217929
THEFT F/AUTO                -756.90273 -0.03793543 -6.322689   7.1551910
THEFT/OTHER                 -677.68220 -0.04521840 -9.190826  -0.5468664

Std. Errors:
                           (Intercept)       Hour  LONGITUDE    LATITUDE
ASSAULT W/DANGEROUS WEAPON 0.0004868488 0.08346952 0.36926995 0.730480036
BURGLARY                   0.0004162205 0.08348162 0.06944330 0.134826264
HOMICIDE                   0.0001840228 0.08606662 0.01433694 0.008677754
MOTOR VEHICLE THEFT        0.0006291977 0.08345131 0.36655093 0.724939860
ROBBERY                    0.0010927775 0.08345633 0.36904888 0.729975132
SEX ABUSE                  0.0001895192 0.08383249 0.01625894 0.017563178
THEFT F/AUTO               0.0009366610 0.08340620 0.23853626 0.471347030
THEFT/OTHER                0.0012285455 0.08340203 0.21405571 0.422864320

Residual Deviance: 71548.95
AIC: 71612.95
```

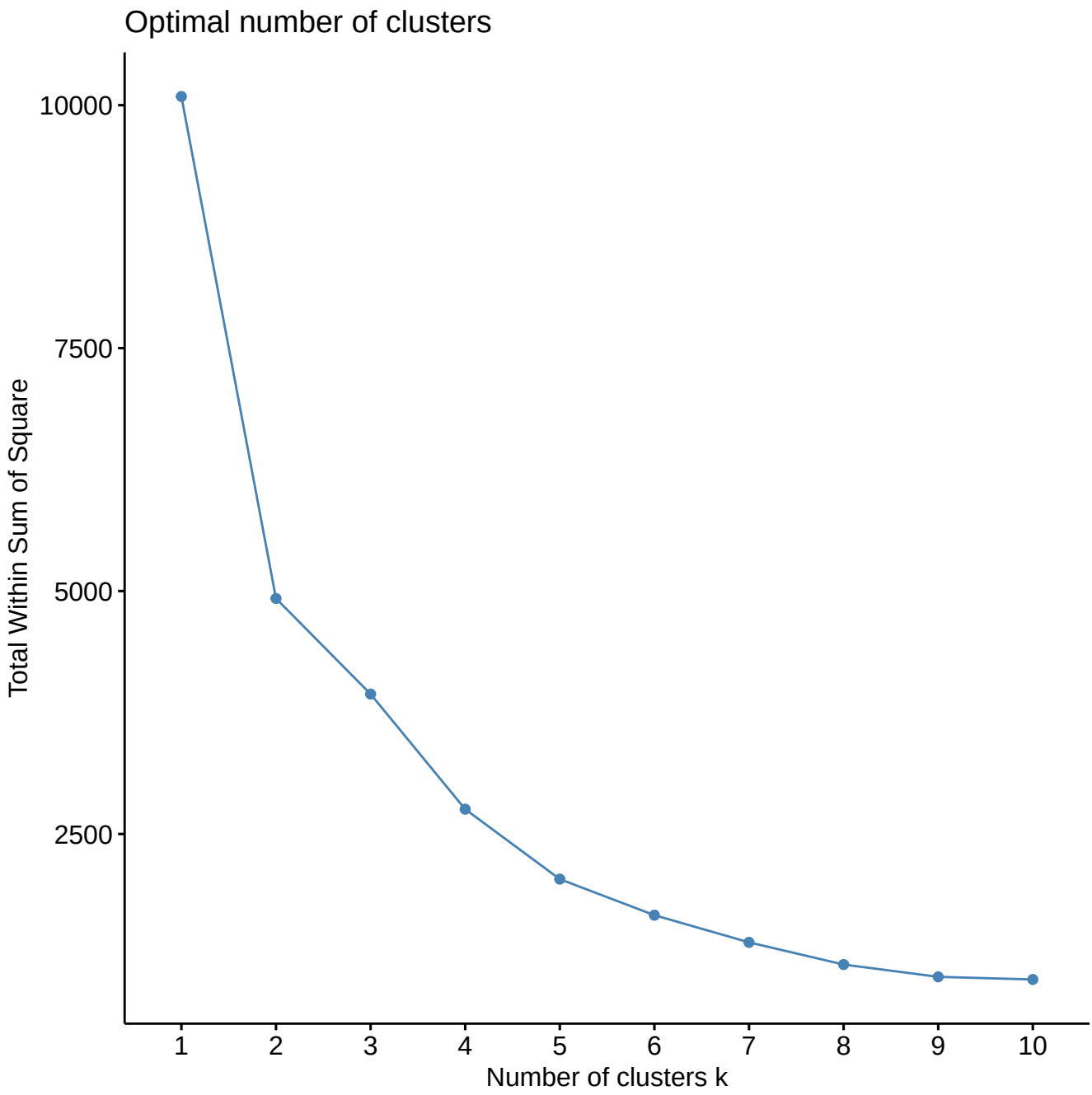
2. Another research question I could look into is `does an offense's frequency vary across months?` From the trend line plot shown earlier in the report, we have already answered a third of this question, that being the frequency of crimes being reported for each shift by month. The statistical learning technique I am going to use to answer this research question will be Chi-square test. The null hypotheses will be that the type of offense and the month are independent, and the alternative hypotheses will be that the type of offense and the month are dependent. The results that I get for carrying out the chi-square test will determine whether we reject or fail to reject the null hypothesis. If we reject the null hypothesis, then both the type of offense and month are dependent, which means that there is a pattern/trend. If we fail to reject the null hypotheses, then both the type of offense and month are independent, which means that there is no pattern/trend.

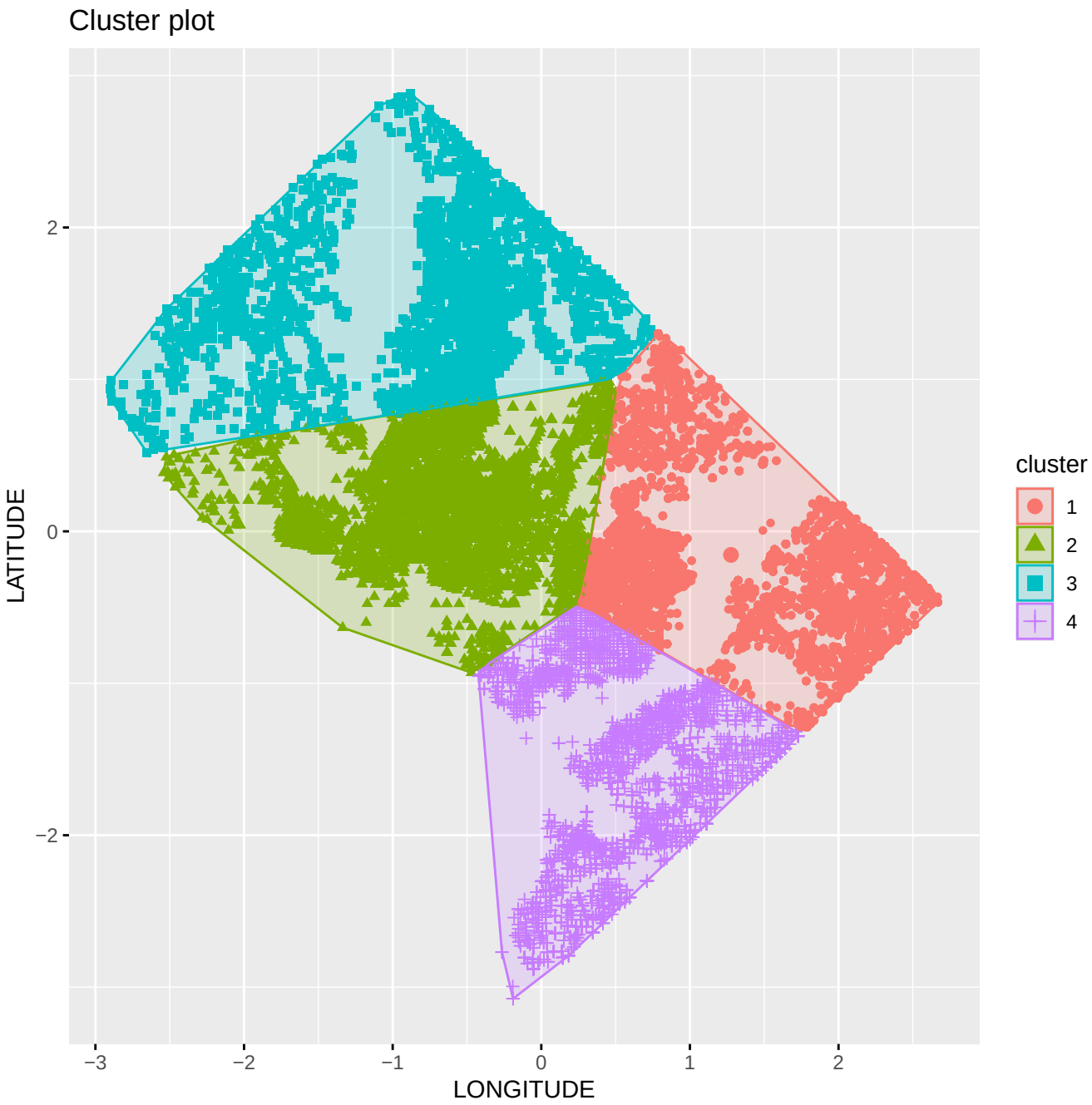
```
    Pearson's Chi-squared test

data:  table_crime
X-squared = 192.8, df = 88, p-value = 8.289e-10
```

Answer to Question 2: Based on the p value we got after carrying out the chi-square test, we confidently reject the null hypothesis. Which means that the type of offense and the month are dependent on each other which implies that there is a pattern between the two variables.

3. And the last research question I could look into is `Are there certain areas in Washington DC where certain type of offenses are reported a lot more than others?` In the data set, there is a latitude and longitude entry in each crime report. By using the unsupervised learning method, we can find points that are in the same cluster, which indicates that the points are similar to each other in that they are very close to one another. With this information, there is good chance that the type of offense in the cluster is the same.


Optimal number of clusters
Total Within Sum of Square
10000
7500
5000
2500
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Cluster plot
LATITUDE
LONGITUDE
2
0
−2
−3
−2
−1
0
1
2
cluster
1
2
3
4

clusters	OFFENSE	count
1	THEFT/OTHER	2852
1	THEFT F/AUTO	1943
1	MOTOR VEHICLE THEFT	873
1	ASSAULT W/DANGEROUS WEAPON	669
1	ROBBERY	644
1	BURGLARY	402
1	SEX ABUSE	86
1	HOMICIDE	42
1	ARSON	2
2	THEFT/OTHER	6949
2	THEFT F/AUTO	4855
2	ROBBERY	646
2	MOTOR VEHICLE THEFT	608
2	BURGLARY	451
2	ASSAULT W/DANGEROUS WEAPON	420
2	SEX ABUSE	109
2	HOMICIDE	10
3	THEFT/OTHER	2132
3	THEFT F/AUTO	2034
3	MOTOR VEHICLE THEFT	404
3	ROBBERY	344
3	BURGLARY	283
3	ASSAULT W/DANGEROUS WEAPON	206
3	SEX ABUSE	35
3	HOMICIDE	13
3	ARSON	1
4	THEFT/OTHER	2531
4	THEFT F/AUTO	1424
4	ASSAULT W/DANGEROUS WEAPON	556
4	ROBBERY	535
4	MOTOR VEHICLE THEFT	522
4	BURGLARY	394
4	SEX ABUSE	67
4	HOMICIDE	50
4	ARSON	2

6 Conclusion

This research project so far was set out to investigate the crime report patterns in Washington D.C. based on the type of offenses, location, and time. Based on the analysis, we were able to conclude for now that the majority of offenses reported in Washington D.C. involved theft, and that the majority of offenses reported were done during the Evening Shift. I do believe that we would've had slightly better results on our tables and plots if the data set didn't have so many missing values. In the future, if I were to approach a similar task of investigating something with data, I would prioritize handling missing values with care.

7 Appendix A: R Code

7.1 Code for table 1: appendix-load-data-n-table

```
# We first read in the crime_incidents csv file and store it as a data frame
crime_incidents_df <- read.csv("Crime_Incidents_in_2017.csv")

# We make a new column called "Month", and extract the month from the REPORT_DAT variable
# and place in the Month column
crime_incidents_df$Month <- month(crime_incidents_df$REPORT_DAT, label = TRUE)

# We create an entirely new table that summarizes the months and offenses together
```

```
wide_summary_table <- pivot_wider(summarize(group_by(crime_incidents_df, Month, OFFENSE),
                                        Total = n()),
                              names_from = OFFENSE, values_from = Total, values_fill = 0)

# We use kable() to give the table a better format
kable(wide_summary_table, caption = "Total Amount of Crimes Reported Each Month per Offense")
```

7.2 Code for bar-graph: appendix-plot-bar-graph

```
# We use pivot_longer() to create a new data frame with the variables month, Offenses,
# and Number of Crimes reported for those offenses.
long_summary_table <- pivot_longer(wide_summary_table,
                                   col = -Month,
                                   names_to = "Offenses",
                                   values_to = "Number_of_Times_Crimes_Reported")

# We use ggplot() and geom_bar() to create a bar graph of the number of crimes reported
# for each crime in each month.
# We use aes() to give the plot object x, y and fill values and geom_bar() to make
# the plot object a bar plot object.
ggplot(data = long_summary_table,
       aes(x = Month, y = Number_of_Times_Crimes_Reported, fill = Offenses)) +
  geom_bar(stat = "identity") +
  labs(title = "Crimes Reported Each Month and Offense",
       x = "Month",
       y = "Number of Crimes Reported")
```

7.3 Code for table 2: appendix-load-univariate-table

```
# We use summarize() to get the summary statistic of the longitude and latitude
# variables, then we store those values onto a table.
univariate_summary_table <- summarize(crime_incidents_df,
                                      longitude_min = min(LONGITUDE),
                                      longitude_max = max(LONGITUDE),
                                      longitude_mean = mean(LONGITUDE),
                                      longitude_median = median(LONGITUDE),
                                      longitude_sd = sd(LONGITUDE),
                                      latitude_min = min(LATITUDE),
                                      latitude_max = max(LATITUDE),
                                      latitude_mean = mean(LATITUDE),
                                      latitude_median = median(LATITUDE),
                                      latitude_sd = sd(LATITUDE))

# Next we use pivot_longer to make the table horizontal, make it more readable.
univariate_summary_table <- pivot_longer(univariate_summary_table, everything())

# After we use kable() to properly format the table and give the table a caption.
kable(univariate_summary_table, caption = "Univariate Summary Table of Longitude and Latitude")
```

7.4 Code for scatter plot: appendix-plot-scatterplot

```
# We use ggplot() and geom_point() to create the plot object and make the plot object a scatter plot.
# We use aes() to give the plot object data to plot and labs() to give the plot labels.
ggplot(crime_incidents_df, aes(x = LONGITUDE, y = LATITUDE)) + geom_point(aes(color = OFFENSE)) +
  labs(title = "Crime Locations in DC",
       x = "Longitude",
       y = "Latitude")
```

7.5 Code for Table 3: appendix-load-bivariate-summary-table

```
# We extract the hour from the entries in the variable "REPORT_DAT" and store those entries
# in a new Hour variable.
crime_incidents_df$Hour <- hour(crime_incidents_df$REPORT_DAT)

# We calculate the summary statistics of the Hour per Shift, this is by using summarize(),
# group_by(), which groups the data set itself and the SHIFT variable, then summarize()
# produces the results for the data set and SHIFT variable. We calculate the the summary
# statistics for the SHIFT variable.
summary_table <- summarize(group_by(crime_incidents_df, SHIFT),
                           min_hour = min(Hour),
                           lower_quantile_hour = quantile(Hour, 0.25),
                           median_hour = median(Hour),
                           upper_quantile_hour = quantile(Hour, 0.75),
                           max_hour = max(Hour),
                           mean_hour = mean(Hour),
                           sd_hour = sd(Hour))

# We use kable() to properly format the table we created and give it a caption
kable(summary_table, caption = "Bivariate Summary Table of Hours per Shift")
```

7.6 Code for boxplot: appendix-box-plot

```
# We use ggplot() and geom_boxplot() to create a plot object with the SHIFT and hour variables
# from the data set.
# Then we use geom_boxplot() to make the plot object a boxplot object with its labels included.
ggplot(crime_incidents_df,
       aes(x = SHIFT,
           y = Hour,
           color = SHIFT)) +
  geom_boxplot() +
  labs(title = "Crimes Reported Hour Distribution per Shift",
       x = "Police Shift when Crime was reported",
       y = "Hour of the report")
```

7.7 Code for Table 4: appendix-load-bivariate-summary-table-2

```
# We use summarize() and group() and include the crime data set, variable Month and Shift
# to create a table that contains
# the information of the amount of crimes reported for each type of offense for each month.
month_shift_trend <- summarize(group_by(crime_incidents_df, Month, SHIFT), n = n())

# We use summarize() and group() and include the crime data set, variable n and Shift to
# find the summary statistics of the number of crimes reported per shift.
summary_table <- summarize(group_by(month_shift_trend, SHIFT),
                           min_n = min(n),
                           max_n = max(n),
                           mean_n = mean(n),
                           median_n = median(n),
                           sd_n = sd(n))

# We use kable() to properly format the table we created and give it a caption
kable(summary_table, caption = "Bivariate Summary Table of Hours per Shift")
```

7.8 Code for plot-trend-line: appendix-plot-trend-line

```
# We use ggplot() along with the month_shift_trend data set, variables Month and
# n (number of crimes committed for each month) to create a trend line graph.
# labs() is used to give the graph its labels etc.
ggplot(month_shift_trend,
```

```
     aes(x = Month,
         y = n,
         color = SHIFT,
         group = SHIFT)) +
geom_line() +
labs(title = "Crime Month Trend by Shift",
     x = "Month",
     y = "Number of Crimes Reported")
```