project1

cherry

10/3/2019

## Chapter 1 - Introduction

The crime incidence in the US is one of the most discussed issues in the country. Although the Federal Bureau of Investigation (FBI) in its 2018 report found an overall decline in violent and property crimes in 2018, there has been more media releases of increasing crime incidence in the US, especially mass shootings in recent periods. Increased incidence of crime is a threat to public safety and welfare. At the national level, violent crime and homicide rates increased from 2014 to 2016, but rate remain near historical lows compared to rates in the 90’s.

Washington DC saw an increase in murder rate by 35.6 percent in 2018. (Brennan Center for Justice). Between 2017 and 2018, of all the types of crimes, homicide rates in DC increased the most by about 38%, followed by auto theft which increased by 13%. There was a decrease in other crimes such as sex abuse, assualt and robbery.

This report focuses on all reported crimes in the DC metro police system which includes violent crime, theft, arson, assault, homicide, sex abuse, and burglary. These crimes can be categorized into violent crime and property crime. Violent crime refers to murder, robbery, rape and aggravated assault. Property crime includes burglary, larceny-theft, and motor vehicle theft. Murder includes murder and non-negligent manslaughter. Total crime incidence includes all the above.

The rest of this report contains 7 chapters - chapter 2 includes the description of data (source, definition of variables and geographic coverage), chapter 3 includes the crime types and methods, chapter 4 shows the discticts of DC area, chapter 5 describes the spatial distribution of crime in DC, chapter 6 and 7 analyses the relationship between time and crime and and chapter 8 presents a conclusion of the report.

## Chapter 2 - Description of Data

### 2.1 Source of the data

The source data for our exploratory data analysis is a CSV containing crime incident data in DC for 2018. This data was sourced from OpenDC. The CSV contains 33,783 crime incidence with reported data and time of incidence, method/weapon used, offence classification for the crime, the location of the crime (block, ward, neighbourhood, voting\_precinct, latitude and longitude), start date, end date and record ID.

Links: [Crime\_Incidents\_in\_2018](https://opendata.dc.gov/datasets/crime-incidents-in-2018?page=3376) / [Police\_Districts](https://opendata.dc.gov/datasets/police-districts)

### 2.2 Definition of the variables in the dataset

For context, homicide in this report refers to the killing of a person purposely, or otherwise, with a malicious aforethought. Sexual abuse as engaging in or causing another person to submit to a sexual act by force, threat or fear. Arson refers to malicious burning or attempt to burn a property, structure, vessel or vehicle of another person. Robbery refers to the act of taking anything of value from another person by force, fear or violence. Assault can be defined as purposely or knowingly causing serious bodily injury, threatening to do so or engaging in any act that creates a risk of physical injury to another person. Burglary is the unlawful entry into a property with the intent to commit a criminal offence. The report date is the date the offense was reported to the police which may be later than the date the incident occurred (DC metropolitan police department).

## Chapter 3 - Crime types and methods

In the first part, we will explore the frequency of the main types and methods of metro crime in DC. First, we need to find out what types of crimes often occur on the metro and try to classify these types of crimes. Then, as for the methods of committing crimes, we can divide them into crime with weapon and crime without weapon. So, before we start to explore the relationship of crime and other factors, we need to classify the crime types and crime methods.

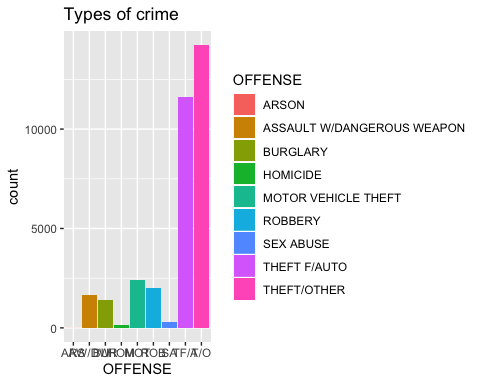
### 3.1 SMART Question

Is there any correlation between types of crime and the use of weapons?

### 3.2 Basic analysis

#### 3.2.1 Which types of crime occurs most in DC metro? Which occurs least?

## ARSON ASSAULT W/DANGEROUS WEAPON   
## 5 1672   
## BURGLARY HOMICIDE   
## 1419 160   
## MOTOR VEHICLE THEFT ROBBERY   
## 2393 2024   
## SEX ABUSE THEFT F/AUTO   
## 274 11609   
## THEFT/OTHER   
## 14227



There are 9 different types of crimes occurred in DC metro. The above bar plot shows the frequencies of different types of crimes. From the plot, we can see there are 11609 times of THEFT(F/AUTO) and 14227 times of OTHER THEFT. So, the THEFT crime is the most frequent type compared to others and the least frequent type is ARSON, which only happened 5 times.

#### 3.2.3 What percentage of crimes are committed with weapons？

## GUN KNIFE OTHERS   
## 1598 772 31413

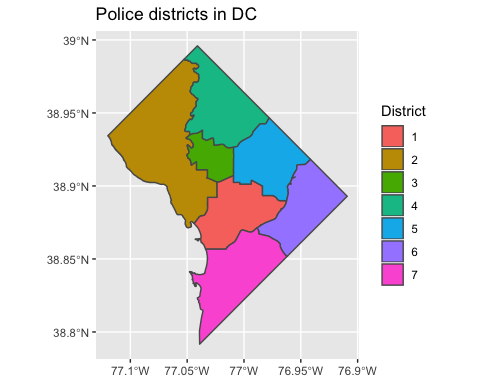
By analyzing the method of crimes, we found that some criminals carried weapons, but others were not. Based on whether carrying weapons or not, we can preliminarily judge the risk factor of the type of crime. We found that approximately 7.02% of the crimes are committed with weapons. There are 1598 crimes in which the criminals used a gun, 772 crimes in which the criminal used knife.

#### 3.2.4 In what types of crimes will the offender use a weapon? Or whether the use or non-use of a weapon varies among the types of crimes?

##   
## GUN KNIFE OTHERS  
## ARSON 0 0 5  
## ASSAULT W/DANGEROUS WEAPON 639 597 436  
## BURGLARY 6 2 1411  
## HOMICIDE 122 10 28  
## MOTOR VEHICLE THEFT 0 0 2393  
## ROBBERY 818 141 1065  
## SEX ABUSE 12 19 243  
## THEFT F/AUTO 0 0 11609  
## THEFT/OTHER 1 3 14223

To further explore the relationship between types of crime and method of crime, we used frequency tables to show in which types of crime the criminals were more likely to use weapons. The result reveals guns were used most in ROBBERY, ASSAULT and HOMICIDE, while knife were frequently used in ASSULT and ROBBERY.

## Chapter 4 - Police districts in DC



There are seven police districts in Washington, DC, and each police district is divided into three sectors with a sector being an informal grouping of Police Service Areas (PSAs). In the following analysis, we will look at the crimes happened in each police district.

## Chapter 5 - Crime location

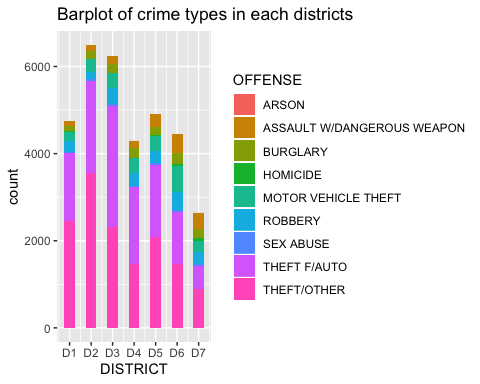
In this session, we seek to investigate the spatial distribution of the crimes in DC metro. We sought to analyze locations of crimes in DC area by category in order to derive insights into the crime frequencies of different area. This dataset only covers basic geographic information about the location of the crime.

### 5.1 SMART Question

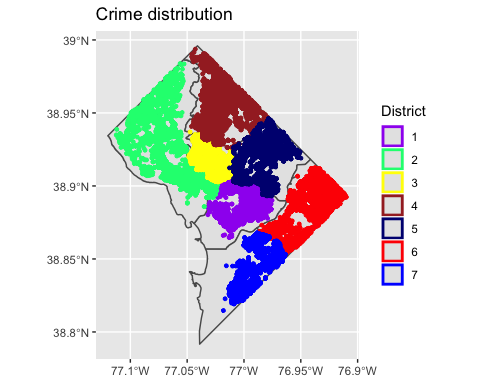
What is the distribution of crime types in each police districts?

### 5.2 Basic analysis

#### 5.2.1 Which police district has the highest crime frequency?

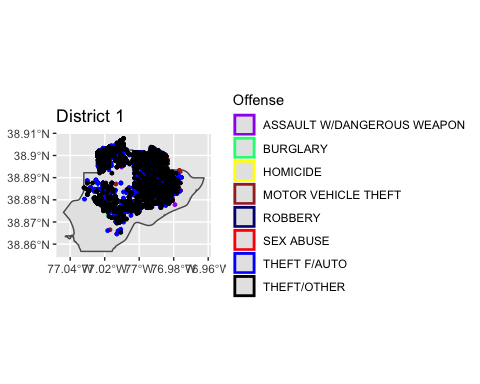


From this plot, we can see district 2 and district 3 have the highest number of crimes, which are above 6000. District 7 has the lowest number of crimes, which is under 3000. The number of crimes committed in district 1, 4, 5 and 6 range from 4000 to 5000.

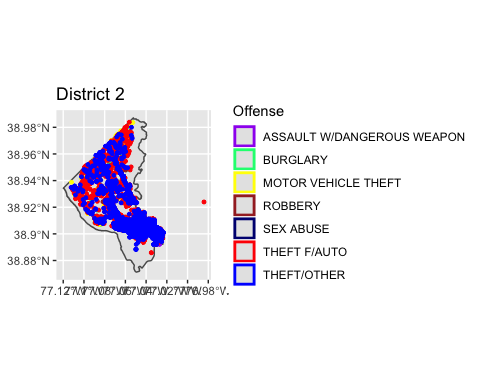


To see the distribution of crimes, first we divide the DC area into seven police districts and map the latitude and longitude of each crime. Only from the map, we can see crimes are concentrated in every area. Since lacking the data about the population of each district, we can only analyze the crime frequency rather than crime rate.

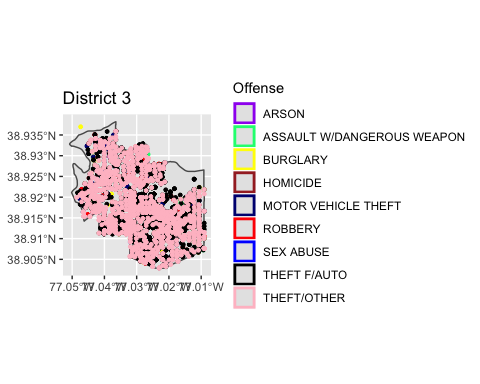
#### 5.2.2 Are there differences in the spatial distribution of different types of crime?



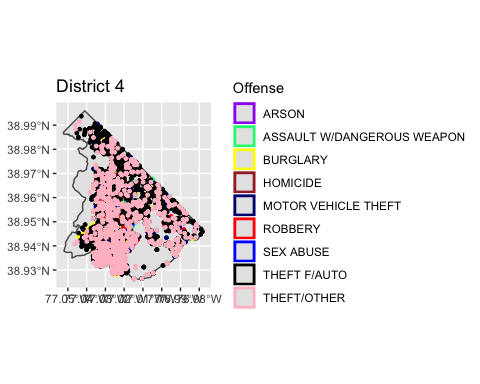
By dividing crime types and zooming each district in the map, we can clearly see the distribution of different types of crime in each area. In District 1, most of the metro crimes were happened in the north and central area. Obviously, THEFTS happened most often in District 1.



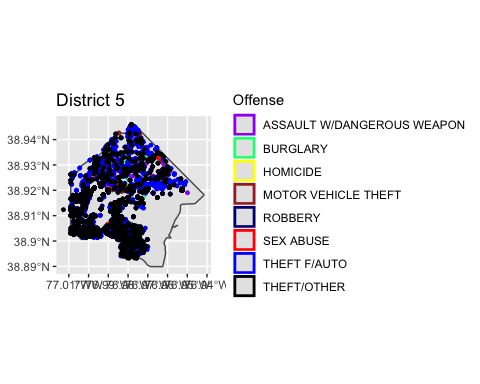
In District 2, most of the metro crimes were happened in the northwest and southeast area. Similarly, THEFTS occurred most frequently in District 2. However, no homicide occured in District 2 compared to other districts.

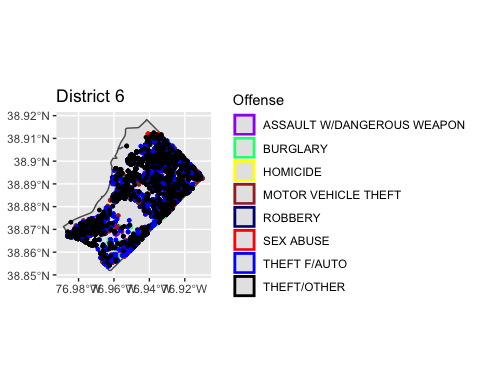


Crimes are evenly distributed in District 3. THEFTS are also the most frequent crime types occurred in this area.

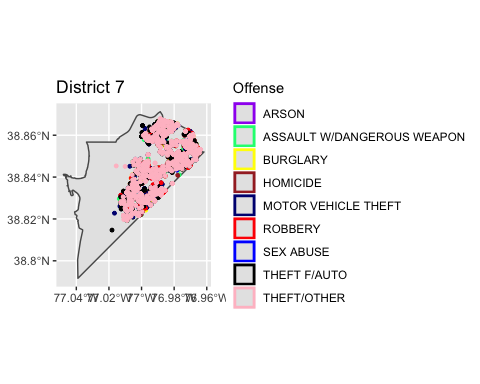


Different types of crimes occurred in District 4, even the THEFTS are the most, other kinds of crimes including burglaries, robberies and sex abuses happened frequently. It is noteworthy that there had been several ASSAULTS WITH WEAPONS in this area.



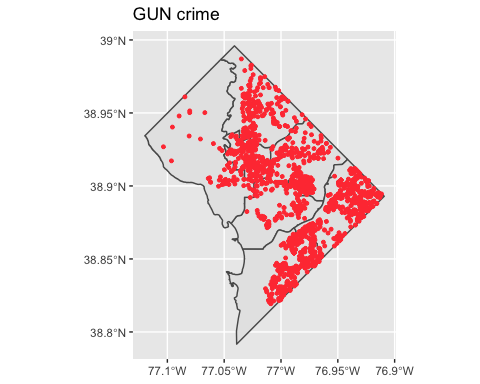


Most crimes occurred in southwest and north area of District 5. Not surprisingly, THEFTS are also the most frequent crime types in both District 5 and District 6.



In District 7, metro crimes distributed in the northeast area. THEFTS are the most, and ASSAULTS WITH WEAPON also happened frequently in the area.

#### 5.2.3 Where are the gun crimes taking place?



Through the above analysis, we observe that there are some crimes with weapons which we consider as dangerous crimes in certain areas. So in order to figure out what is the location distribution of this specific types of crime, we subset the data by selecting crimes method which is gun and map the gun shooting crimes in DC. From the map, we can see most gun crimes are distributed in the east area of DC. It is obvious that crimes with gun occurred least in district 2.

### 5.3 Conclusion

In conclusion, this chapter analyzes the relationship between crime spots and crime types. There are several insights we can derieved from this chapter. First, THEFT are the most commom crime types in every police district. Second, crimes types are significant different among each polic districts. Third, gun crimes are more frequent in the east of DC than in the west of DC.

## Chapter 6 - Crime Time

The dataset includes specific crime time including date and time in 2018 year, therefore we can derive our insights into it. Firstly, we explore the crime occurrence distribution in different crime time. Secondly, we take a look at the relationship between crime offense and crime time.

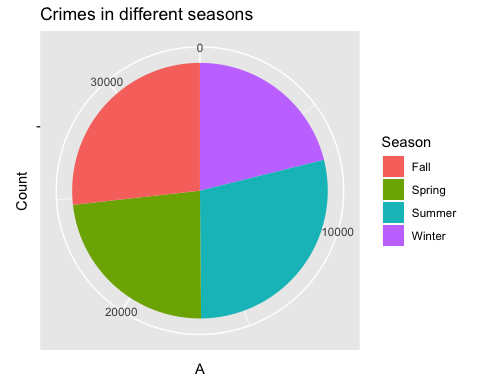
### 6.1 SMART Question

Does crime occurrence frequency and crime offense differ by crime time including seasons in one year and time in one day?

### 6.2 Basic Data Processing of crime time

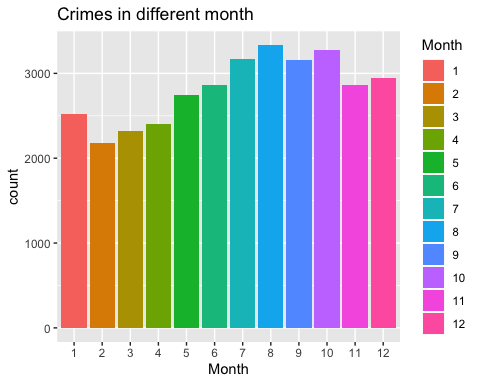
Firstly, we add two new columns into the dataset according to the inforamation in “start\_date” and group them into “crimemonth” and “crimeseason”(We select Months Dec, Jan and Feb as winter). Later we will use these two variables with frequency of crime occurrence to analyze more.

### 6.3 Crime Occurrence Frequency



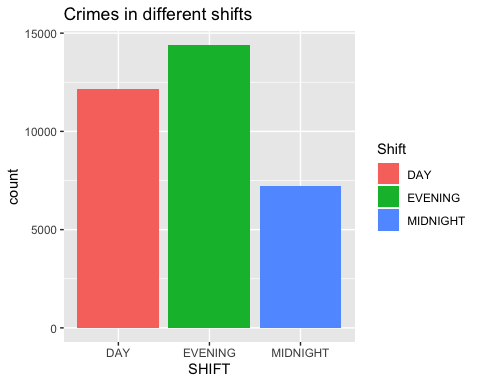
## Winter Spring Summer Fall   
## 7136 7892 9709 9046

The pie chart above shows the crime frequency in different seasons. We can find that the frequency in Autumn and Summer accounts most.



## 1 2 3 4 5 6 7 8 9 10 11 12   
## 2521 2180 2321 2407 2742 2866 3166 3334 3158 3279 2868 2941

We take a look at the first bar plot and it shows the crime occurrence in different months in 2018. The tendency of frequency distribution shows central high and two edges low. It appears that during the summer, there are more likely to occur crime. The most crime incidence month is August and the frequency is 3334.The fewest is in February belonging to winter and the frequency is 2321.



## DAY EVENING MIDNIGHT   
## 12150 14394 7239

This bar plot shows the crime frequency in different shift in one day. We divide one day into three parts, which are Day, Evening and Midnight. From the results, we can see that most crime occur at evening and the frequency is 14394. The fewest crime occur in the day and the frequency is 12150, which is lower obviously than another two shift.

### 6.4 crime offense during different time

This bar plot shows the frequency of different offenses in 4 seasons. Matching with the former part, the offense THEFT F/AUTO and THEFT/OTHER have the most two frequency. And almost all of offenses occur in the four seasons.

To see if the offense the same across different crime time including seasons in one year and time in one day? We do furthuer statiscal inference to explore more.

#### 6.4.1 offense ~ seasons

H0: Offense and crime time for seasons in one year are independent.

H1: They are not independent.

We do the chi-squared to test this hypothesis and calculate the p-value. The outputs below are the summary of Chi-squared test and p-value.

Since the p-value is small, we accept the alternative hypothesis that offense and crime time for seasons in one year are not independent.

#### 6.4.2 offense ~ time in one day

H0: Offense and crime time in one day are independent. H1: They are not independent.

The same as the offense ~ seasons, we use the chi-squared to test the hypothesis and calculate the p-value. The outputs below are the summary of Chi-squared test and p-value.

## Warning in chisq.test(contable1): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: contable1  
## X-squared = 2245.6, df = 16, p-value < 2.2e-16

The p-value here is still small, we reject the null hypothesis that offense and crime time in one day are not independent.

### 6.5 Crime Time Analysis Conclusion

The crime time is affected by crime time including different seasons in one year and different time in one day.

## Chapter 7 - Report Time

From the dataset, we also found that there exists different time gap between the date of report and crime. We found it interesting for the reason that there may exists some relationship between the time difference with other variables including offense and crime time in one day. Therefore, In this part, we will explore more on this.

### 7.1 SMART Question

Does the time difference between report date and crime date have a correlation with offense and crime time in one day?

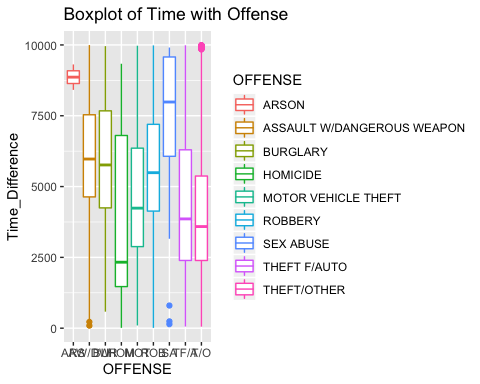
### 7.2 Data processing of time difference

First, as for the raw dataset only have columns report date and crime date, we need to acquire the time difference and created a new column named ‘time difference’ calculated by using column ‘report\_date’ subtracting column ‘start\_date’.

### 7.3 Basic analysis

### 7.3.1 Is time difference affected by offense?

As there are some ineffective data in the column time\_difference, which the value is less than 0, we only select the value difference between 0 and 1000.



Above are the graph of the time difference statistics grouped by offense in the box-plot, we can find that they are different obviously. The SEX ABUSE offense has the largest time difference. It might because for some specific offense such as SEX ABUSE, people may not struggle inside and hesitate to report it due to self-esteem. However, for the homicide offense, people are more willing to report them in a hurry, so the time difference in it is relatively low. Besides, we found the range of ARSON offense is extremely small, whereas others are relatively large.

To find if there are relationship between time difference an offense. We use statiscal inference further.

H0: The mean value of the time difference is the same across offense. H1: They are different.

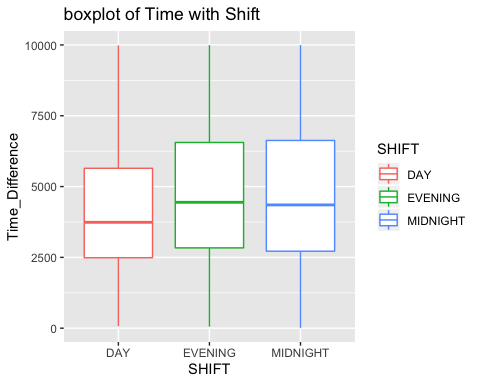
We use ANOVA to test the hypothesis and calculate the p-value. The outputs below are the summary of ANOVA test and p-value.

## Df Sum Sq Mean Sq F value Pr(>F)   
## OFFENSE 8 7.785e+09 973181387 187.8 <2e-16 \*\*\*  
## Residuals 17040 8.831e+10 5182504   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since the p-value is small, we reject the hypothesis that the mean value of the time difference is the same across offense. The report time after being attacked have relationsHip with the crime offense. Victims seem to reluctant to report some specific offense such as SEX ABUSE as soon as possible.

#### 7.3.2 Is time difference affected by crime time in one day?

Similar with above, firstly we take a look at the time\_difference distribution in the different shift of one day. The results are as follows. The range and average of time\_difference in midnight is larger than another two. However, in the day, people are tend to report crime as soon as possible, in which the average of difference is the smallest.

 To find if there are some relationship between time difference and shift, we continue to do hypothesis test.

H0: The mean value of the time difference is the same across different time in one day. H1: They are different.

We use ANOVA to test the hypothesis and calculate the p-value. The outputs are the summary of anova test, p-value and TukeyHSD.

## Df Sum Sq Mean Sq F value Pr(>F)   
## SHIFT 2 1.116e+09 558002878 100.1 <2e-16 \*\*\*  
## Residuals 17046 9.498e+10 5571941   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Time\_Difference ~ SHIFT, data = crimeTimeAndShift)  
##   
## $SHIFT  
## diff lwr upr p adj  
## EVENING-DAY 552.53910 455.3952 649.6830 0.0000000  
## MIDNIGHT-DAY 512.86239 398.3016 627.4232 0.0000000  
## MIDNIGHT-EVENING -39.67671 -149.3451 69.9917 0.6731052

We find the p-value is very small. We reject the null hypothesis that the mean value of the time difference is the same across different time in one day. Since the trend here is not clear, we want to know which pairs are different using TukeyHSD. The p-value for evening-day pair and midnight-day pair are extremely small, so we think they are significant. However, for the midnight-evening, the p-value is 0.67 which is larger than 0.05, we fail to reject the null hypothesis and think that the mean value of time differences are the same between this.

### 7.4 Report Time Analysis Conclusion

In this section, we do the EDA and ANOVA tests to explore the relationship between time difference and other facors. Firstly, we found that crime offense affects the time difference between report time and crime time. Secondly, the time differences among offense are differently. The SEX ABUSE offense has the largest time difference. Thirdly, the time difference is also related with shift in one day,

## 8 Summary

### 8.1 conclusion

There are lots of points we can find from examining the reported crimes in the DC metro police system data.

First, we take a look at the lists of crime offense and crime method. Then we analyze the relationship between crime spots and crime types. We conclude that THEFT are the most commom crime types in every police districts and crimes types are significant different among each polic districts.

Next, for the crime time variable, not only we found the time for crime incidence occur most in summer, especially in August, but also observed there exists relationship between crime offense and crime time, including different time in one day and different seasons in one year.

Finally we uncovered the time difference between report date and crime date, and found that both the crime offense and crime time have relationship with the time difference between crime time and report time.

### 8.2 future analysis

First, the dataset itself has some limitations. Most of variables in this dataset are categorical or discrete rather than continuous, so our statistical analysis methods are limited. In the final project we can use other models, such as logistic regression, to analyze these categorical variables.

Second, the information in this dataset are also limited. Crime is the core concept in this report, but we lack the information about the population, so we can only use crime frequency instead of crime rate to evaluate the risk in each regions. But crime frequency is actually not an perfectly objective indicators, so we can’t simply conclude that this area is dangerous because of its high crime frequency. In future analysis, we will try to find more objective indicators.

Third, this dataset only contains some basic information about crimes including times and locations. But if we want to further explore what kinds of social factors may lead to these crimes, we may add some other dataset to supplement this dataset. We speculate that other social factors associated with urban crimes include populations, housing price, unemployment, inequality, the rapid pace of urbanization and so on. So in the final project, we will append other data set, such as General Social Survey, to do further exploration.