San-Francisco Crime Analysis- Exploratory Analysis and Classification

Vivek Yadav, PhD

January 30, 2016

Synopsis

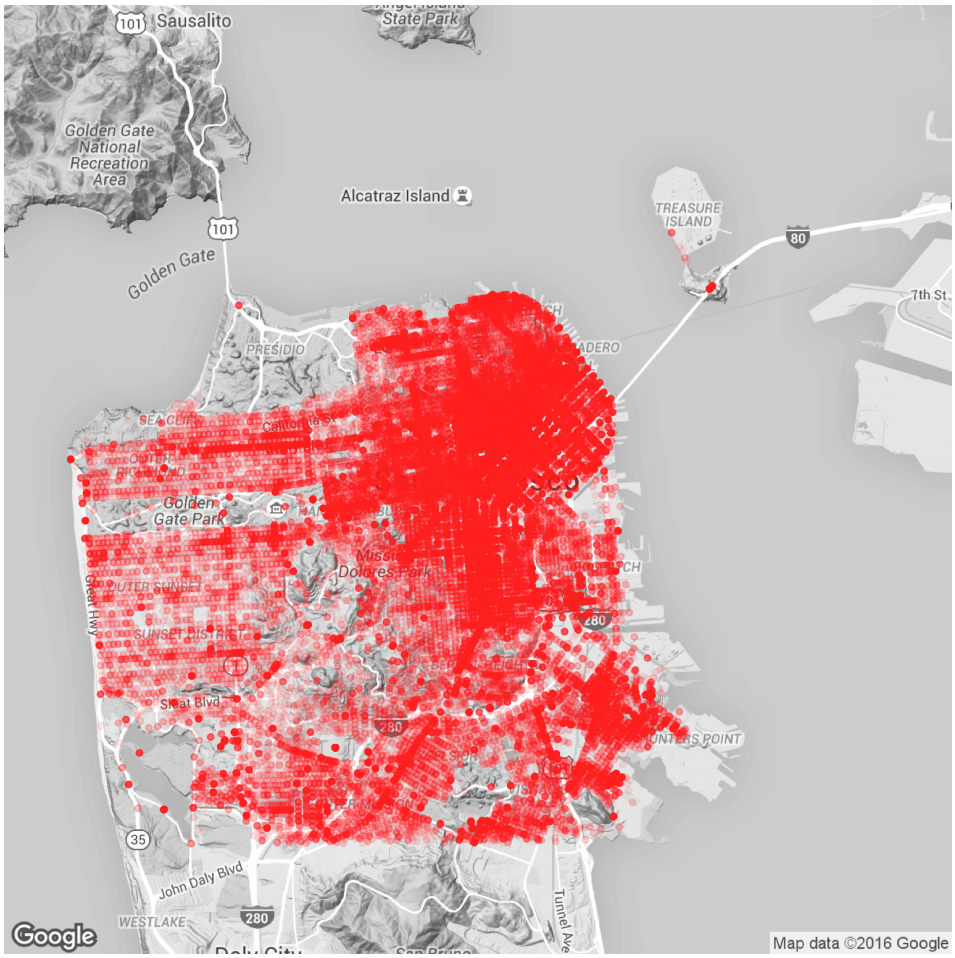
Here I analyzed the crime incidents in the city of San Francisco, and built a linear classifier to predict the probability of a crime belonging to certain category. I downloaded the data set from Kaggle's San Francisco crime classification competition (https://www.kaggle.com/c/sf-crime (https://www.kaggle.com/c/sf-crime)). I first loaded the data into R. The data set has information on the location of the crime and the time at which the crime occured starting from 2003. The data set had more than 800,000 rows and 9 columns. I did exploratory analysis and made several variables to quantify various aspects of the crime incident. For example, from time information, I got information about data, hour of the day, month, year, etc. After exploratory analysis, I identified that the day of week, hour of the day, month, year and location were the main factors that affected crime rates. I therefore used these to predict the probability of a crime belonging to given category. I used R's Liblinear package to implement a L2-regularized logistic regression model to predict probability of each crime. To validate my model, I split into a 50% training set and 50% validation set. I then fitted the model on training data and improved based on its performance on the validation set. My final model had day of week, hour of the day, month, year, location and interaction between location and year. I then trained this model on full data. I then uploaded this on Kaggle and my best submission got a score of 401/1173. This is a work in progress, and I will make more refinements to the model in future.

```
##                 Dates       Category                     Descript
## 1 2015-05-13 23:53:00       WARRANTS               WARRANT ARREST
## 2 2015-05-13 23:53:00 OTHER OFFENSES       TRAFFIC VIOLATION ARREST
## 3 2015-05-13 23:33:00 OTHER OFFENSES       TRAFFIC VIOLATION ARREST
## 4 2015-05-13 23:30:00  LARCENY/THEFT   GRAND THEFT FROM LOCKED AUTO
## 5 2015-05-13 23:30:00  LARCENY/THEFT   GRAND THEFT FROM LOCKED AUTO
## 6 2015-05-13 23:30:00  LARCENY/THEFT GRAND THEFT FROM UNLOCKED AUTO
##   DayOfWeek PdDistrict     Resolution                    Address         X
## 1 Wednesday   NORTHERN ARREST, BOOKED        OAK ST / LAGUNA ST -122.4259
## 2 Wednesday   NORTHERN ARREST, BOOKED        OAK ST / LAGUNA ST -122.4259
## 3 Wednesday   NORTHERN ARREST, BOOKED VANNESS AV / GREENWICH ST -122.4244
## 4 Wednesday   NORTHERN           NONE  1500 Block of LOMBARD ST -122.4270
## 5 Wednesday       PARK           NONE 100 Block of BRODERICK ST -122.4387
## 6 Wednesday  INGLESIDE           NONE       0 Block of TEDDY AV -122.4033
##          Y
## 1 37.77460
## 2 37.77460
## 3 37.80041
## 4 37.80087
## 5 37.77154
## 6 37.71343
```

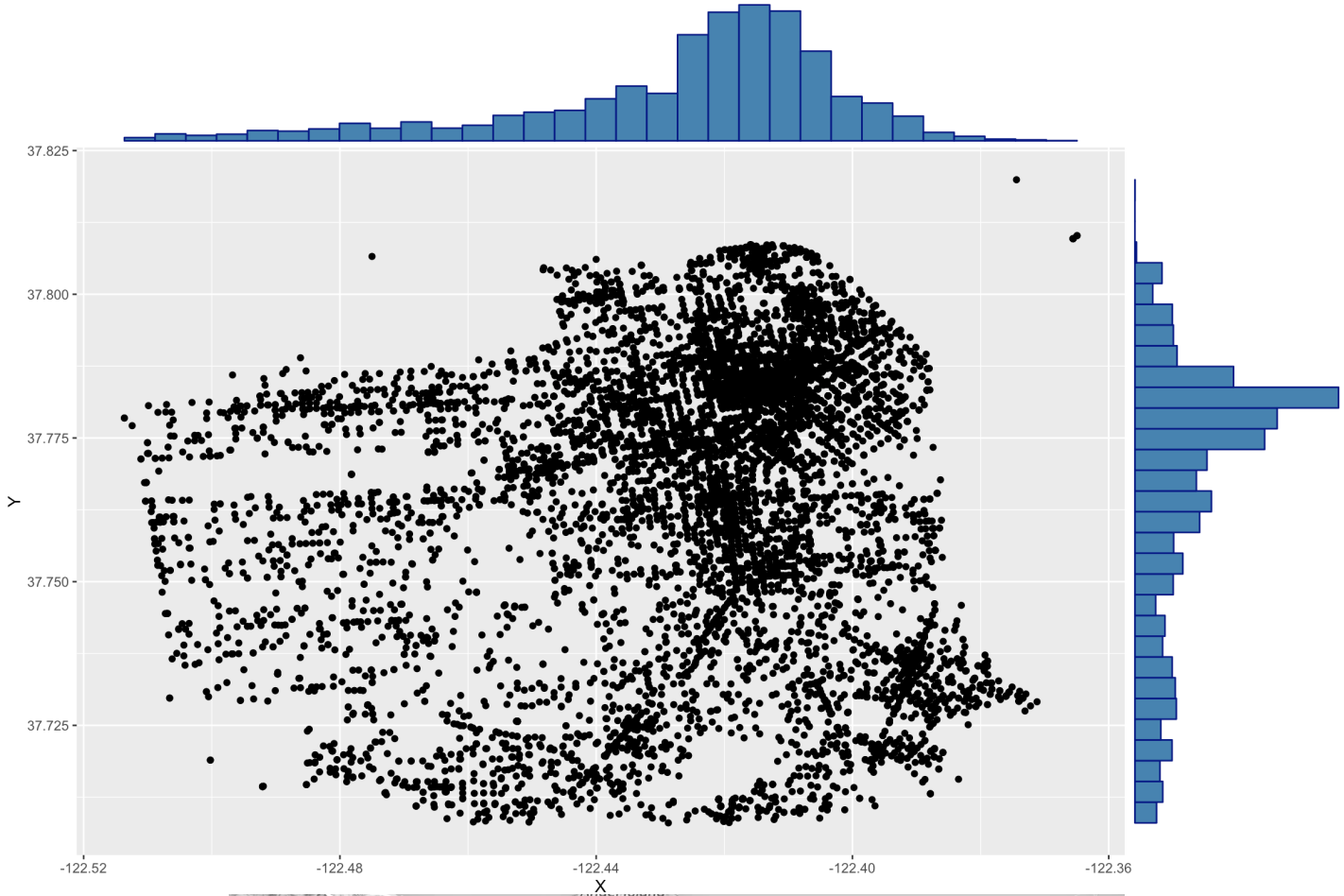
Univariate Plot

Crime distribution

I downloaded San Francisco crime data set from Kaggle. https://www.kaggle.com/c/sf-crime (https://www.kaggle.com/c/sf-crime). After loading the data sets, I checked for distribution of crime across San Francisco. I first plotted the map of San Francisco with crime in red. Plot of map of 100000 randomly sampled crime location shows that the incidences of crime are higher in the eastern San Francisco area.

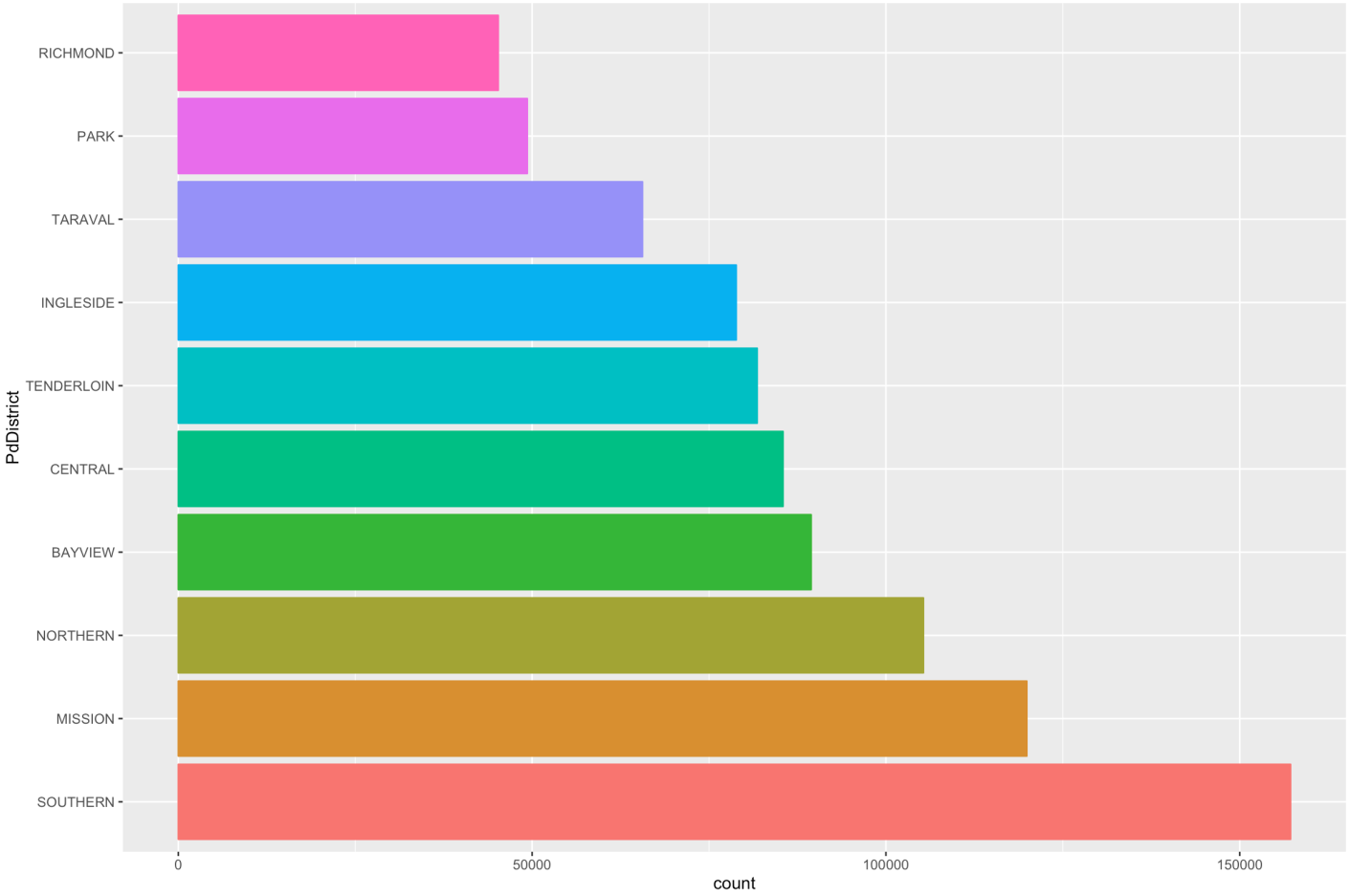


Distribution of locations of crime plotted for a subsample of 100000 points shows that most of the crime is concentrated in the north-eastern region of the map. However, more detailed contour maps show that these crimes are concentrated in 2 specific areas.



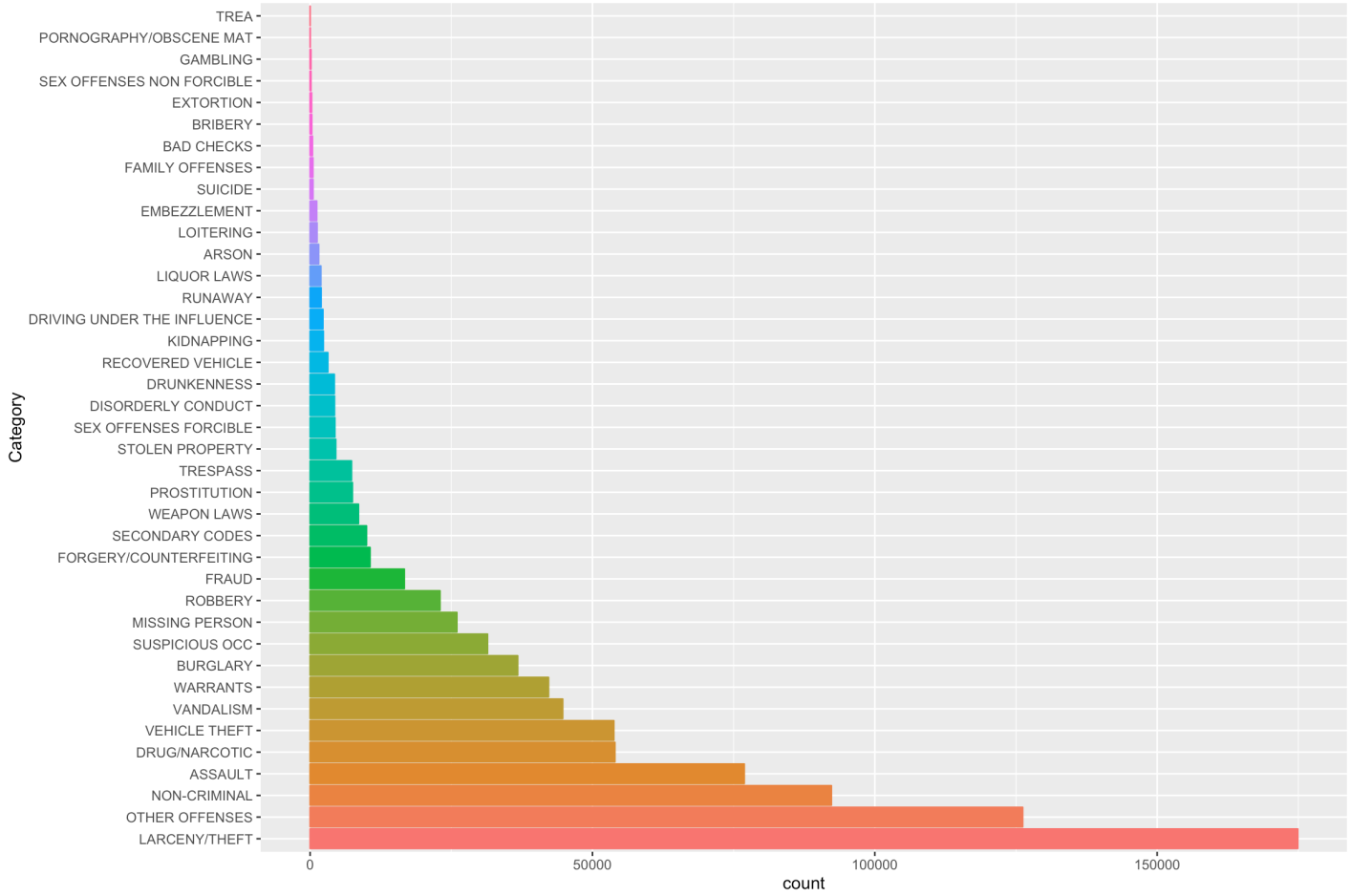
Location/police districts.

I plotted crime vs police district. Figure below shows a greater incidence of crime in Southern and Mission districts of San Francisco.



Types of crime.

I next draw a bar plot to show number of crimes in each category. Bar plots indicate that the top crime categroy is Larceny/Theft. Further, top 20 crime types account for 97% of the crimes.




```
## [1] "Top 10 crimes"
```

```
##                Category  count
## 17          LARCENY/THEFT 174900
## 22         OTHER OFFENSES 126182
## 21           NON-CRIMINAL  92304
## 2                ASSAULT  76876
## 8           DRUG/NARCOTIC  53971
## 37          VEHICLE THEFT  53781
## 36              VANDALISM  44725
## 38               WARRANTS  42214
## 5                BURGLARY  36755
## 33         SUSPICIOUS OCC  31414
## 20         MISSING PERSON  25989
## 26                ROBBERY  23000
## 14                  FRAUD  16679
## 13 FORGERY/COUNTERFEITING  10609
## 28        SECONDARY CODES   9985
## 39             WEAPON LAWS   8555
## 24           PROSTITUTION   7484
## 35                TRESPASS   7326
## 31        STOLEN PROPERTY   4540
## 29  SEX OFFENSES FORCIBLE   4388
```

```
## [1] "Percentage of crimes in top 20 categories =  0.969965229730915"
```

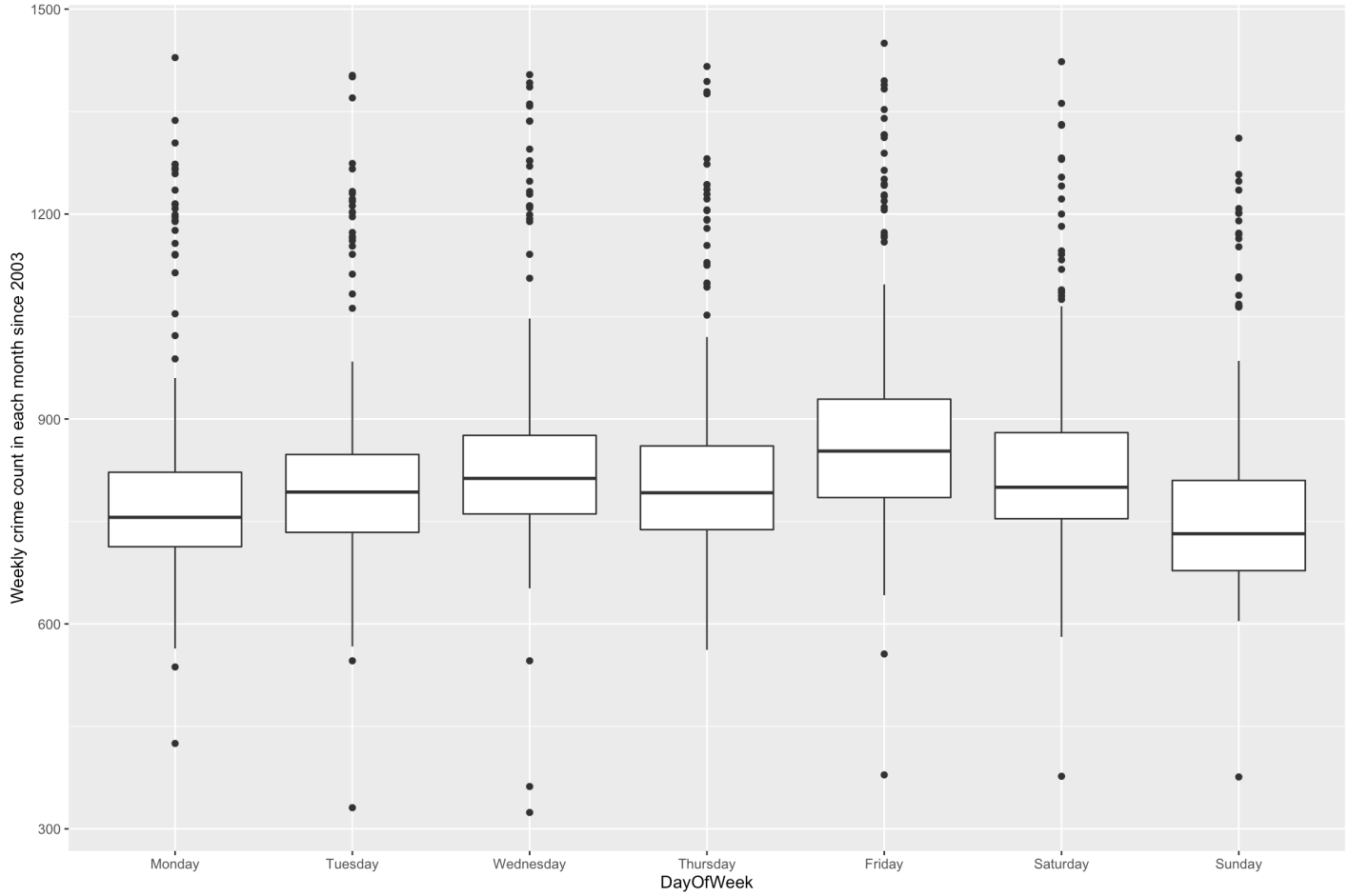
Extracting time information from date tag, and plotting crimes as functions of time.

In addition to crime category, time of the incident was also proivided in the data set. I used strptime function to convert time string to a datetime object and then used strftime to obtain more time-variables for the crime, for example, hour, month, year, day of week and day of month, etc.

```
##                 Dates       Category                     Descript
## 1 2015-05-13 23:53:00       WARRANTS                WARRANT ARREST
## 2 2015-05-13 23:53:00 OTHER OFFENSES       TRAFFIC VIOLATION ARREST
## 3 2015-05-13 23:33:00 OTHER OFFENSES       TRAFFIC VIOLATION ARREST
## 4 2015-05-13 23:30:00  LARCENY/THEFT   GRAND THEFT FROM LOCKED AUTO
## 5 2015-05-13 23:30:00  LARCENY/THEFT   GRAND THEFT FROM LOCKED AUTO
## 6 2015-05-13 23:30:00  LARCENY/THEFT GRAND THEFT FROM UNLOCKED AUTO
##   DayOfWeek PdDistrict     Resolution                    Address         X
## 1 Wednesday   NORTHERN ARREST, BOOKED        OAK ST / LAGUNA ST -122.4259
## 2 Wednesday   NORTHERN ARREST, BOOKED        OAK ST / LAGUNA ST -122.4259
## 3 Wednesday   NORTHERN ARREST, BOOKED VANNESS AV / GREENWICH ST -122.4244
## 4 Wednesday   NORTHERN           NONE  1500 Block of LOMBARD ST -122.4270
## 5 Wednesday       PARK           NONE 100 Block of BRODERICK ST -122.4387
## 6 Wednesday  INGLESIDE           NONE       0 Block of TEDDY AV -122.4033
##          Y Years Month DayOfMonth Hour YearsMo weekday      AddressType
## 1 37.77460  2015    05         13   23 2015-05 Weekday     Intersection
## 2 37.77460  2015    05         13   23 2015-05 Weekday     Intersection
## 3 37.80041  2015    05         13   23 2015-05 Weekday     Intersection
## 4 37.80087  2015    05         13   23 2015-05 Weekday Non-Intersection
## 5 37.77154  2015    05         13   23 2015-05 Weekday Non-Intersection
## 6 37.71343  2015    05         13   23 2015-05 Weekday Non-Intersection
```

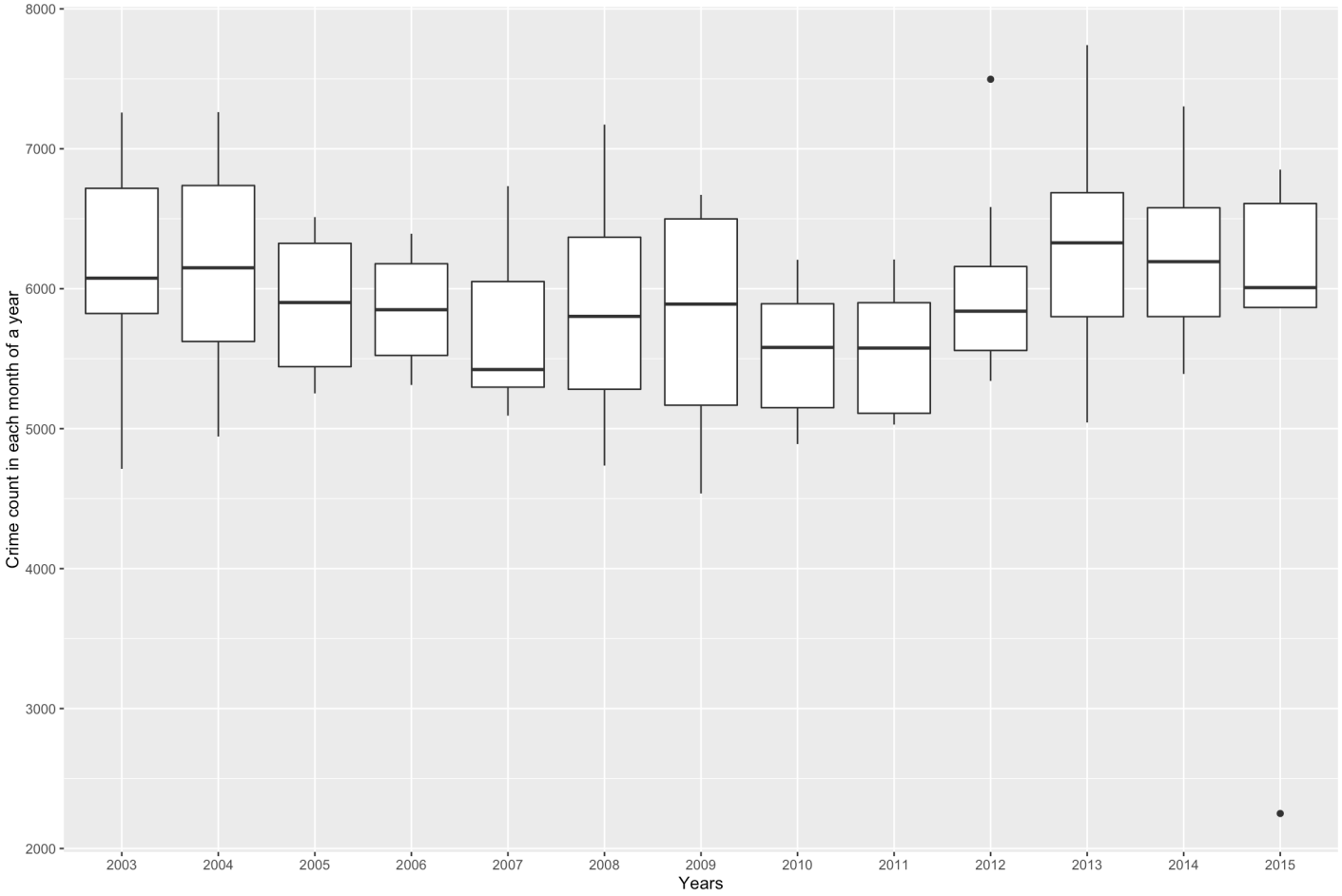
Day of week.

I plotted crime as function of the Day of the week. Plots indicate that the crimes peak on Wednesday and Fridays, and sunday seems to have lower crime.



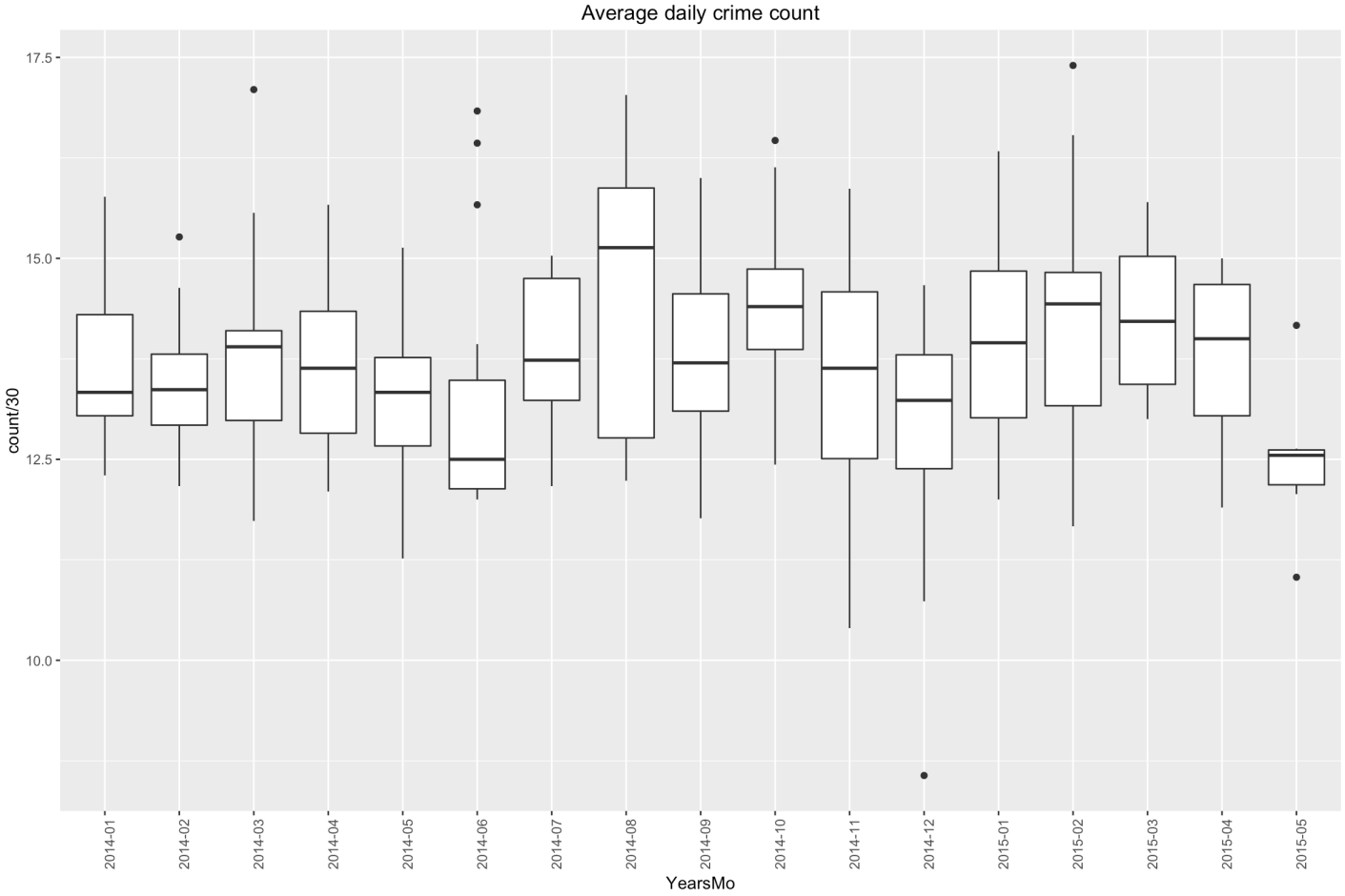
Years

I next plotted yearly trend of crime. I plotted total crime over all the years. It appears that there is a big dip in 2015. This because the data set has data only until May of 2015. Other than that, from 2003 to 2010 there was overall decrease in crime. However, since 2010, the number of crimes has increased.



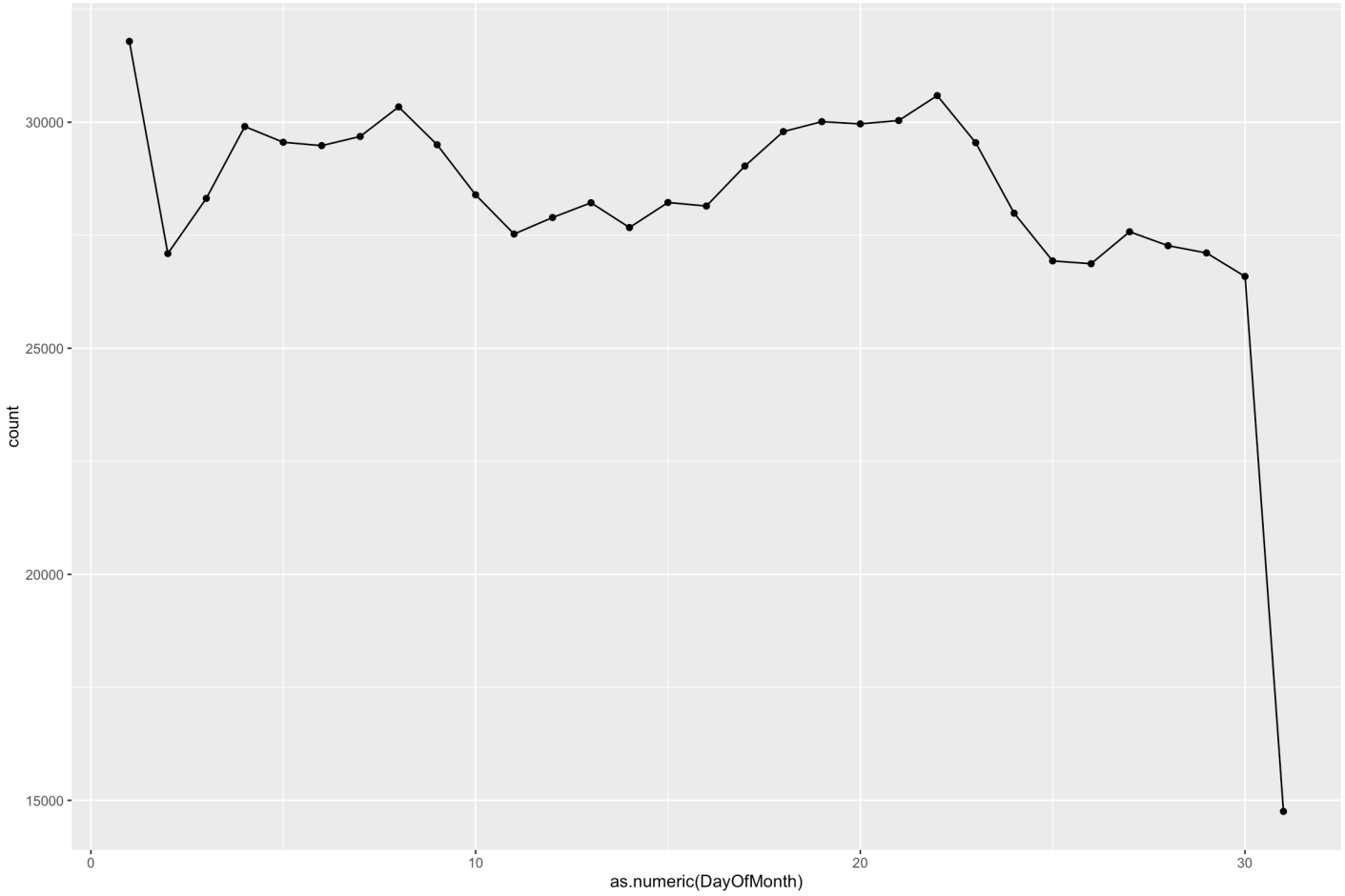
By Year-month combo

To check if there is complete data for 2015, I created a year-month variable and plotted it. As it can be seen from the graph below, the data is available only until May of 2015. A lower number in May indicates that this count is not complete. Therefore, may of 2015 will not be included in further analysis.



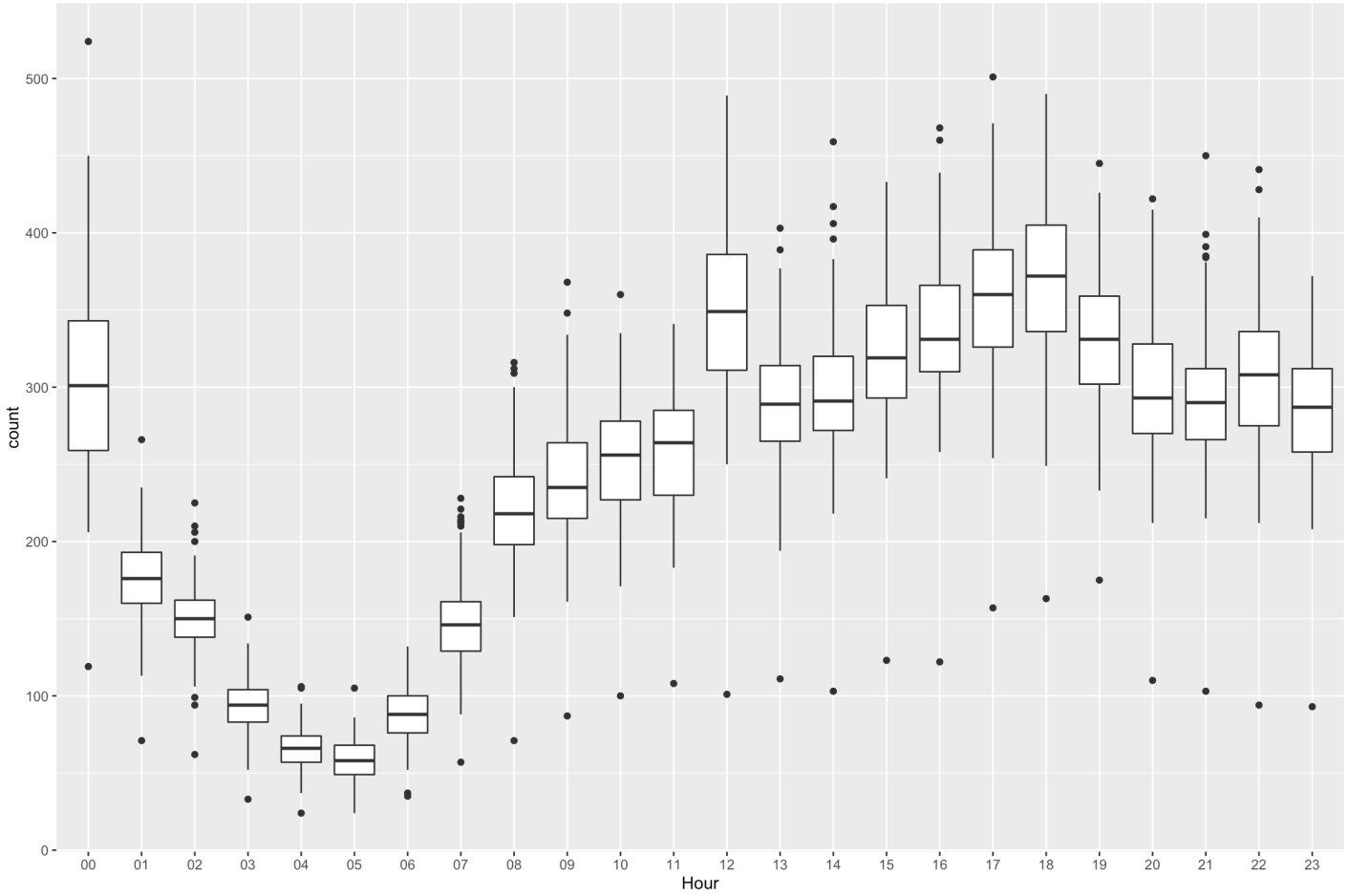
By Day of the month

I next plotted crime by the day of the month. First thing to note is that on 31st, the numbers of crimes is much lower than on other days. This is because in a year, there are very few days whose date is 31, 7 vs 12 (or 11) for other dates. In addition, number of crimes on 1st are higher than any other date. This may be because of 1 being default for the date in cases where exact date is not available. Excluding 1 and 31, there is cyclicity in the number of crimes vs date. In particular, there is higher incidence of crime between 4th and 8th of each month and 18 to 22 of each month. I will investigate this further in bivariate and multivariate plots section.



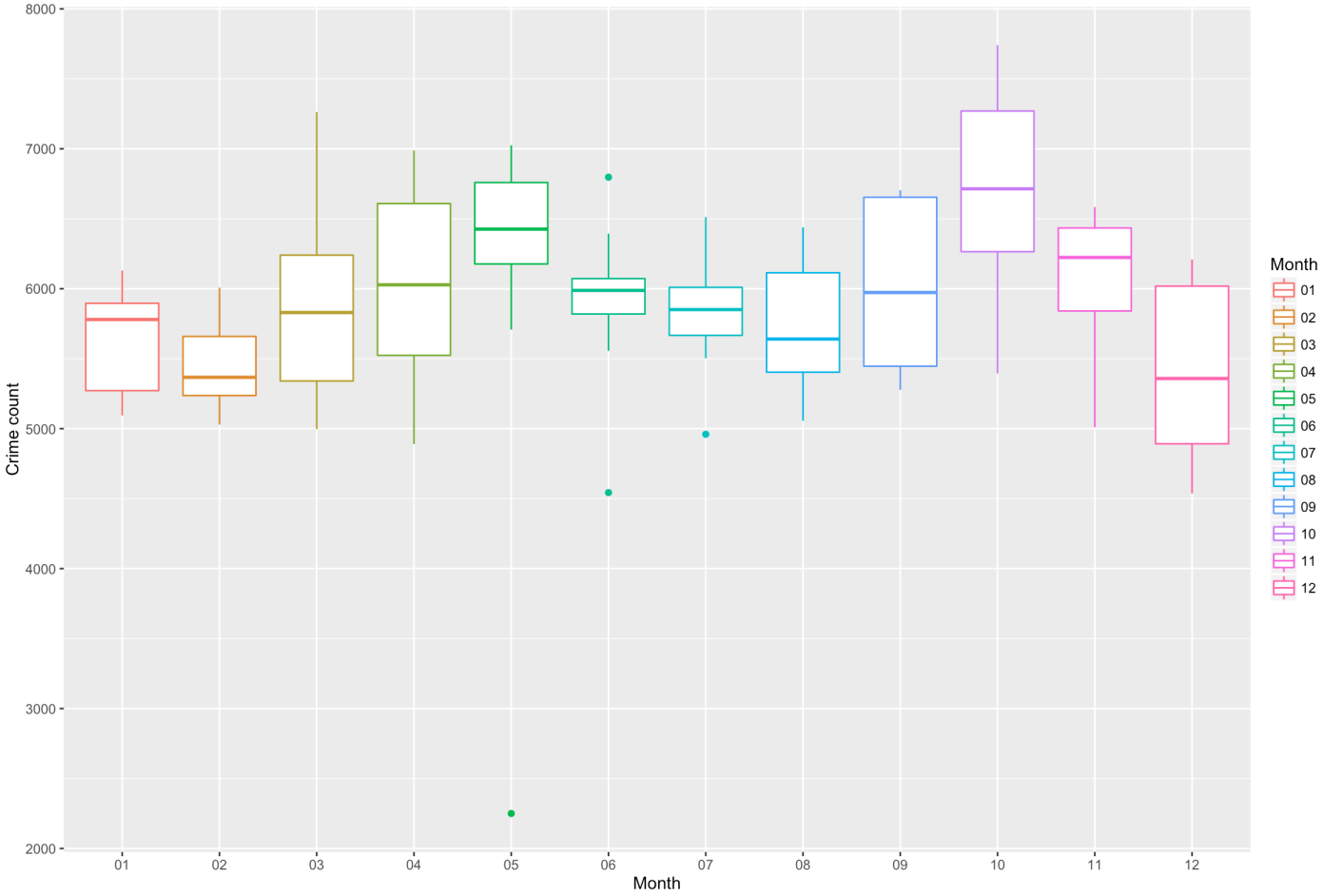
By hour of the day

I next investigated crime by the hour of the day. Figure below shows a clear dip in crime from midnight to 5 am. The number of crimes then increase steadily until 10 am, and remain at high levels until midnight.



Crimes by Month

Crime indicences vs month indicates a drop in crime rates in december and peak in crime in october and may. I didnt expect to see such a trend. I will investigate these trends more in bivariate section.



What is the structure of your dataset?

There are a 878049 rows of data in 17 columns. These columns correspond to,

1. Date: Date on which crime occured
2. Category: Category of the crime, Category has 36 levels corresponding to the type of crime.
3. Descript: Descript corresponds to description of the incident.
4. DayOfWeek: “Monday” to “Sunday” factor variable indicating day of the week.
5. PdDistrict: Police district.
6. Resolution: Result of crime.
7. Address: Address where crime occured
8. X: Longitude of location of crime
9. Y: Latitude of location of crime
10. Years: Year in which crime occured
11. Month: Month in which crime occured 1-January, 2-February, . . . , 12- December.
12. DayOfMonth: 1 to 31 indicating date
13. Hour: Hour of the day
14. YearsMo: Years-month combination to investigate time-trend
15. HourZn: Coarser descritication of hours in a day.
16. weekday: Factor variable indicating if a day falls on weekend or on weekday,
17. AddressType: Some addresses were entered as intersections and others as full addresses so I made this variable from address.

```
## [1] "Number of rows in data :  878049"
```

```
## [1] "Number of columns in data :  16"
```

```
##                 Dates       Category                     Descript
## 1 2015-05-13 23:53:00       WARRANTS               WARRANT ARREST
## 2 2015-05-13 23:53:00 OTHER OFFENSES       TRAFFIC VIOLATION ARREST
## 3 2015-05-13 23:33:00 OTHER OFFENSES       TRAFFIC VIOLATION ARREST
## 4 2015-05-13 23:30:00  LARCENY/THEFT   GRAND THEFT FROM LOCKED AUTO
## 5 2015-05-13 23:30:00  LARCENY/THEFT   GRAND THEFT FROM LOCKED AUTO
## 6 2015-05-13 23:30:00  LARCENY/THEFT GRAND THEFT FROM UNLOCKED AUTO
##   DayOfWeek PdDistrict     Resolution                    Address         X
## 1 Wednesday   NORTHERN ARREST, BOOKED        OAK ST / LAGUNA ST -122.4259
## 2 Wednesday   NORTHERN ARREST, BOOKED        OAK ST / LAGUNA ST -122.4259
## 3 Wednesday   NORTHERN ARREST, BOOKED VANNESS AV / GREENWICH ST -122.4244
## 4 Wednesday   NORTHERN           NONE  1500 Block of LOMBARD ST -122.4270
## 5 Wednesday       PARK           NONE 100 Block of BRODERICK ST -122.4387
## 6 Wednesday  INGLESIDE           NONE       0 Block of TEDDY AV -122.4033
##          Y Years Month DayOfMonth Hour YearsMo weekday      AddressType
## 1 37.77460  2015    05         13   23 2015-05 Weekday     Intersection
## 2 37.77460  2015    05         13   23 2015-05 Weekday     Intersection
## 3 37.80041  2015    05         13   23 2015-05 Weekday     Intersection
## 4 37.80087  2015    05         13   23 2015-05 Weekday Non-Intersection
## 5 37.77154  2015    05         13   23 2015-05 Weekday Non-Intersection
## 6 37.71343  2015    05         13   23 2015-05 Weekday Non-Intersection
```

What is/are the main feature(s) of interest in your dataset?

The goal of this analysis is to predict the probability of a given category based on location and the time of the crime. Therefore, most important feature of the data set is Category. Other important features are the location of the crime and time at which the crime occured. Based on these, I will develop a model to predict the probability of the category given location and time of the incident.

```
## [1] "Number of rows in test data :  884262"
```

```
## [1] "Number of columns in test data :  14"
```

```
##   Id               Dates DayOfWeek PdDistrict                  Address
## 1  0 2015-05-10 23:59:00    Sunday    BAYVIEW  2000 Block of THOMAS AV
## 2  1 2015-05-10 23:51:00    Sunday    BAYVIEW       3RD ST / REVERE AV
## 3  2 2015-05-10 23:50:00    Sunday   NORTHERN   2000 Block of GOUGH ST
## 4  3 2015-05-10 23:45:00    Sunday  INGLESIDE 4700 Block of MISSION ST
## 5  4 2015-05-10 23:45:00    Sunday  INGLESIDE 4700 Block of MISSION ST
## 6  5 2015-05-10 23:40:00    Sunday    TARAVAL    BROAD ST / CAPITOL AV
##           X        Y Years Month DayOfMonth Hour YearsMo weekday
## 1 -122.3996 37.73505  2015    05         10   23 2015-05 Weekend
## 2 -122.3915 37.73243  2015    05         10   23 2015-05 Weekend
## 3 -122.4260 37.79221  2015    05         10   23 2015-05 Weekend
## 4 -122.4374 37.72141  2015    05         10   23 2015-05 Weekend
## 5 -122.4374 37.72141  2015    05         10   23 2015-05 Weekend
## 6 -122.4590 37.71317  2015    05         10   23 2015-05 Weekend
##        AddressType
## 1 Non-Intersection
## 2     Intersection
## 3 Non-Intersection
## 4 Non-Intersection
## 5 Non-Intersection
## 6     Intersection
```

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

Other important features are the variables that I extracted from the time of the crime. Some specific patterns to note are,

1. Crime is low and drops between midnight and 6 am.
2. Crime is low during December and highest in October.
3. Most crimes were of Larceny/Theft.
4. Southern and Mission districts had the highest crime incidents.

Did you create any new variables from existing variables in the dataset?

I created 8 additional variables,

1. Years: Year in which crime occured
2. Month: Month in which crime occured 1-January, 2-February, . . . , 12- December.
3. DayOfMonth: 1 to 31 indicating date
4. Hour: Hour of the day
5. YearsMo: Years-month combination to investigate time-trend
6. HourZn: Coarser descritication of hours in a day.
7. weekday: Factor variable indicating if a day falls on weekend or on weekday,
8. AddressType: Some addresses were entered as intersections and others as full addresses so I made this variable from address.

Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

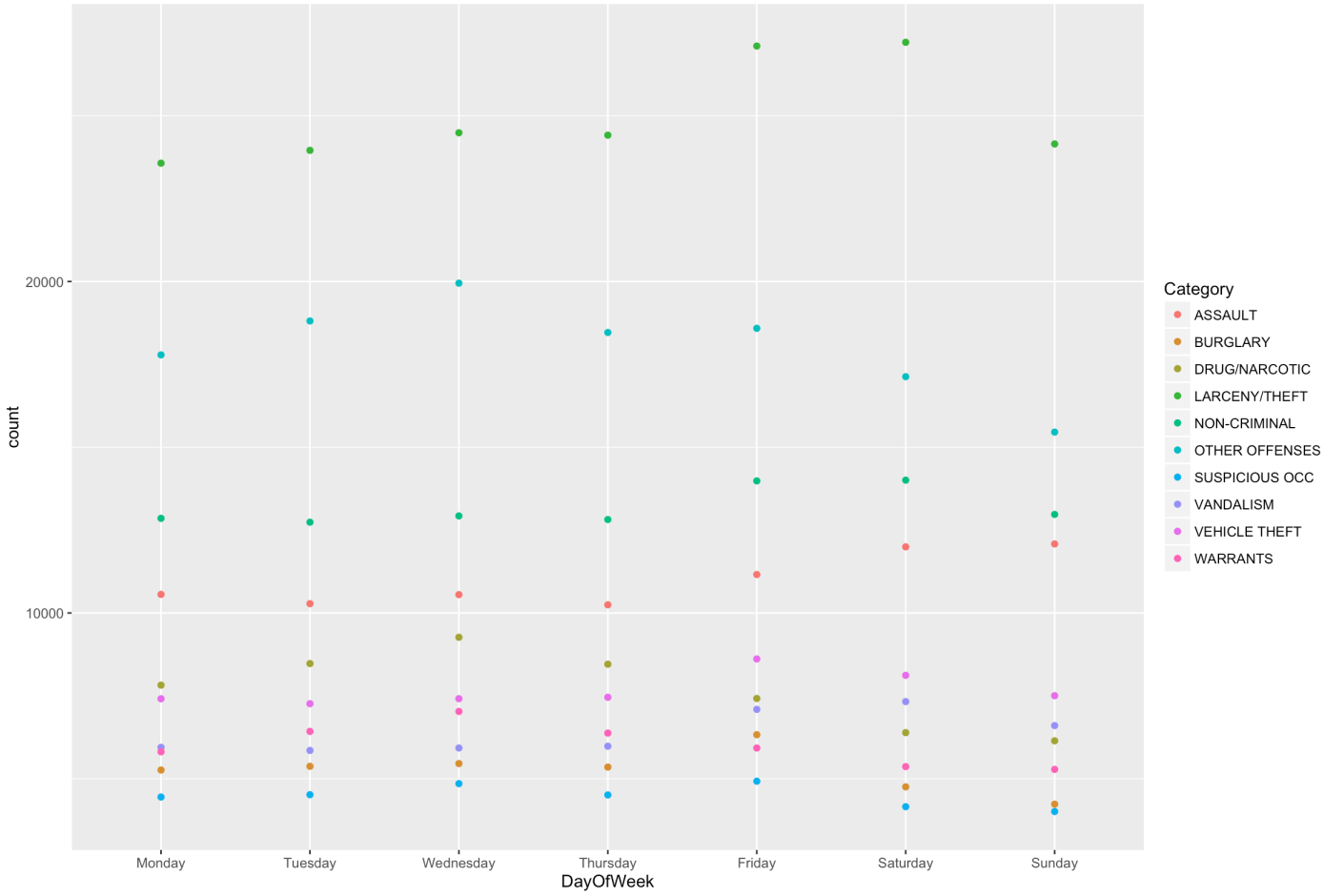
I extracted several features from time of the crime data. I got several additional variables to include effects of time and seasonal trends in crime. I also included variables for address type.

Bivariate section

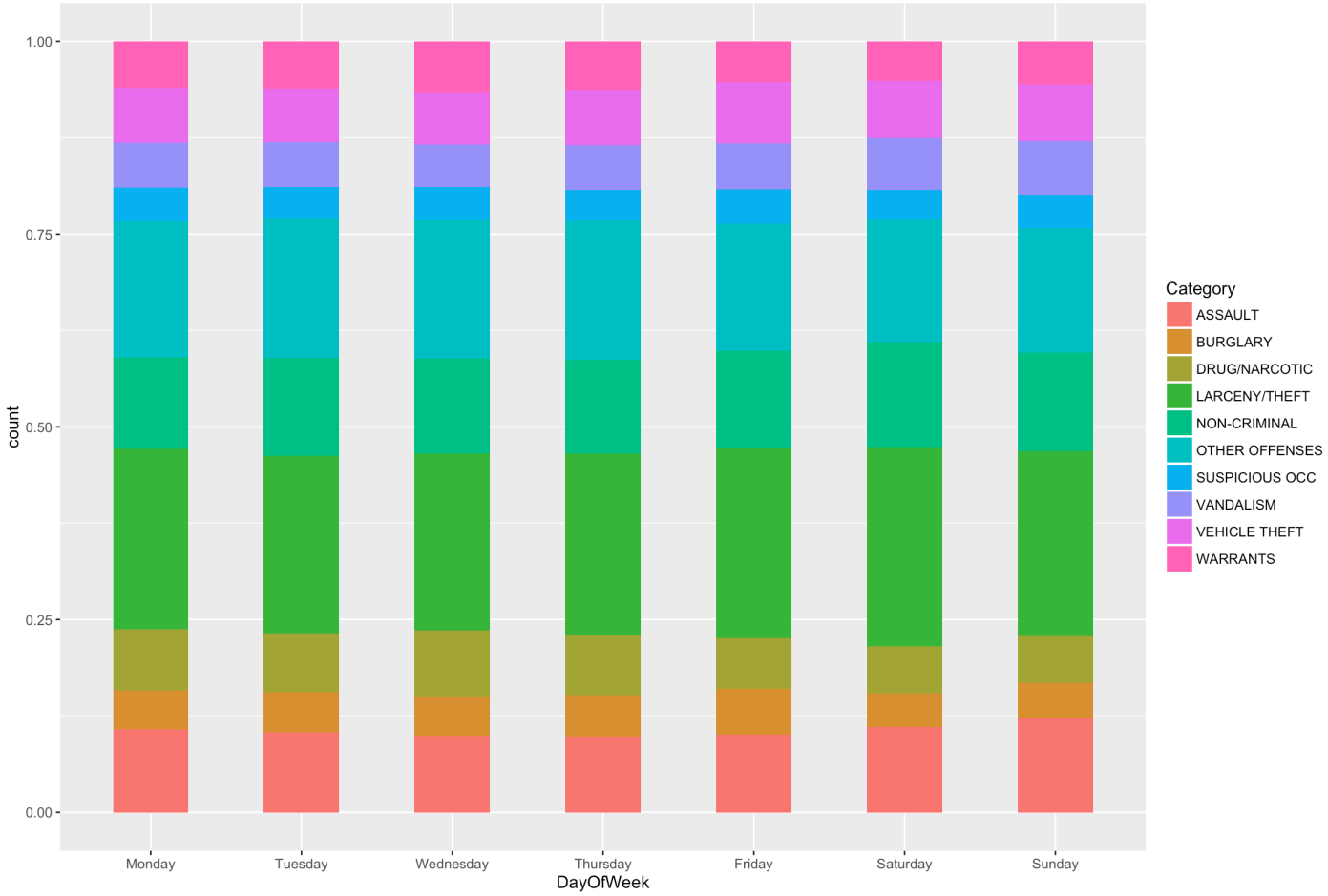
In this section, I performed bivariate analysis where I plotted 2 or more variables against 1. I specifically fraction of each crime category as a function of different variables.

Types of crime vs Day of the week

Here I plot category of the crime vs day of the week. First I plotted total count.

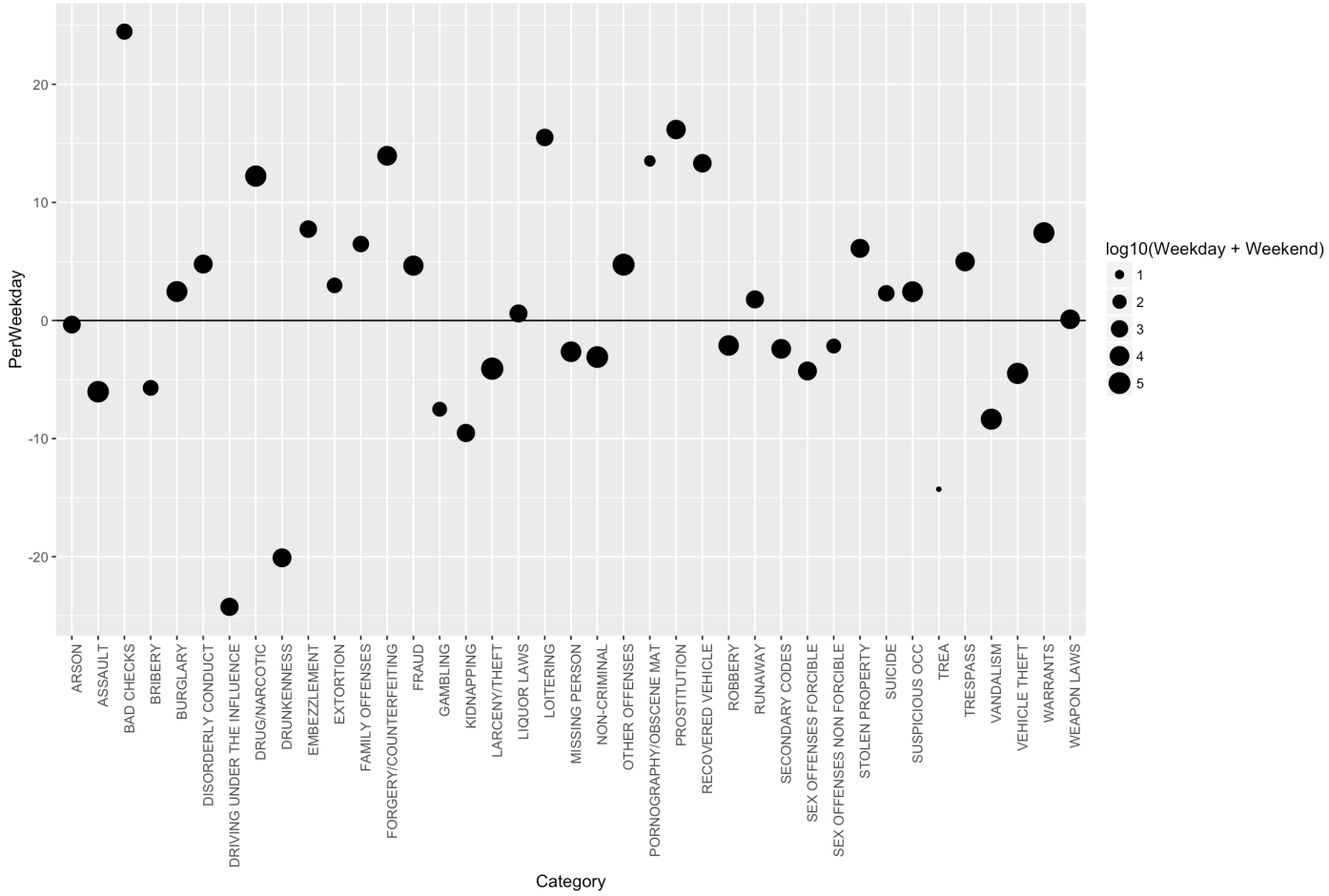


As I intend to use these results to predict probability of a crime category, I plotted fraction of category vs day of crime. As Figure below shows that crime trends are different on different days. Saturday and Sundays seem to have higher larceny rates.



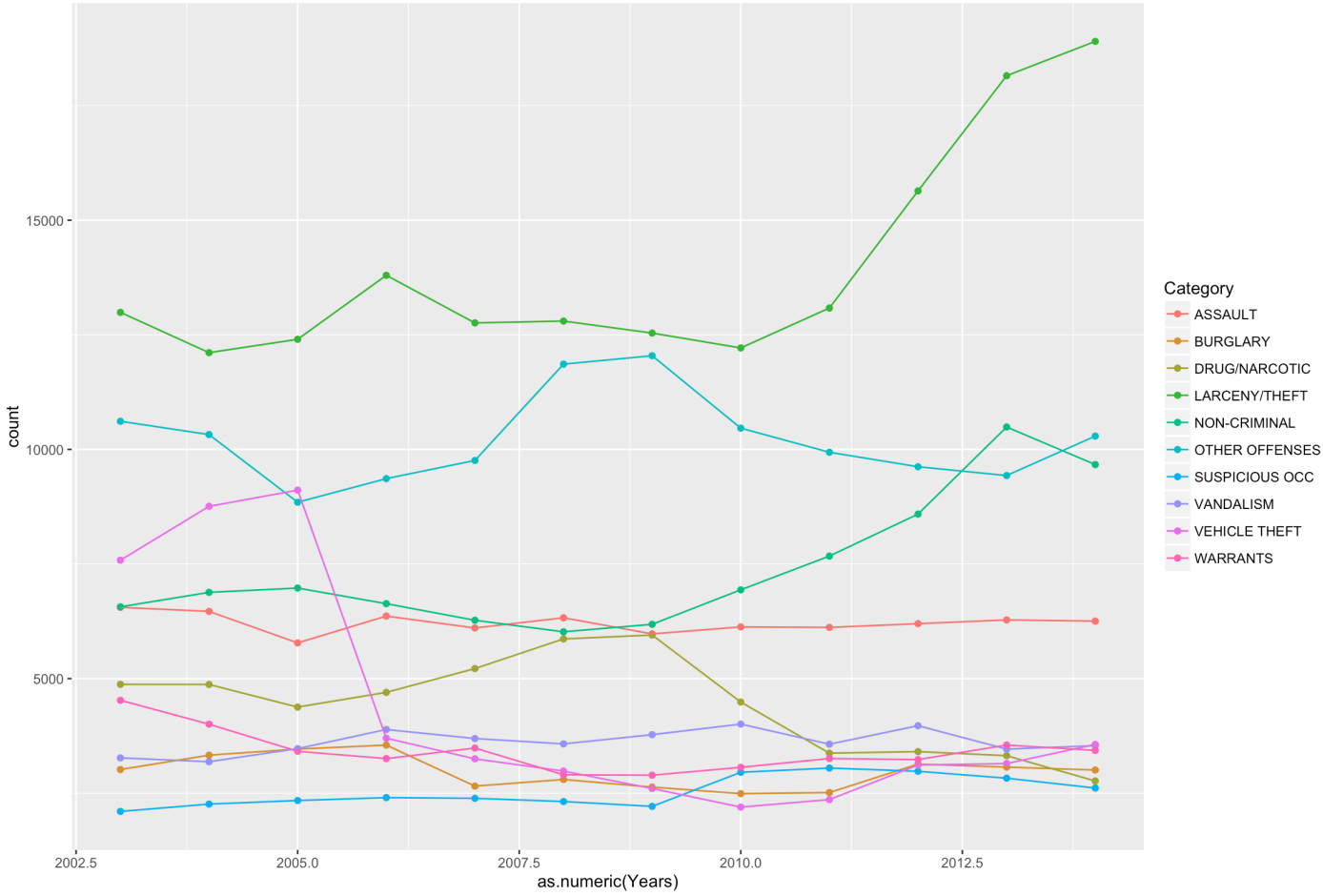
Previous plots indicates a higher crime rate on weekends. So I divided days into 2 zones, crimes during 4 weekdays (Mon-Thu) and on 3 days of weekends (Friday-Sunday). I computed average crime count by dividing total crime by the number of days in the weekday variable (4 for weekdays, and 3 for weekends). I then calculated percentage difference in average number of crimes, a positive value indicates higher crime on weekdays, and a negative value indicates lower crime on weekdays. As evident, there is a strong affect of weekday (weekend or during week) variable on crime rates. I was surprised to see that average DUI incidents are higher during weekdays, I had expected it to be higher during weekends.

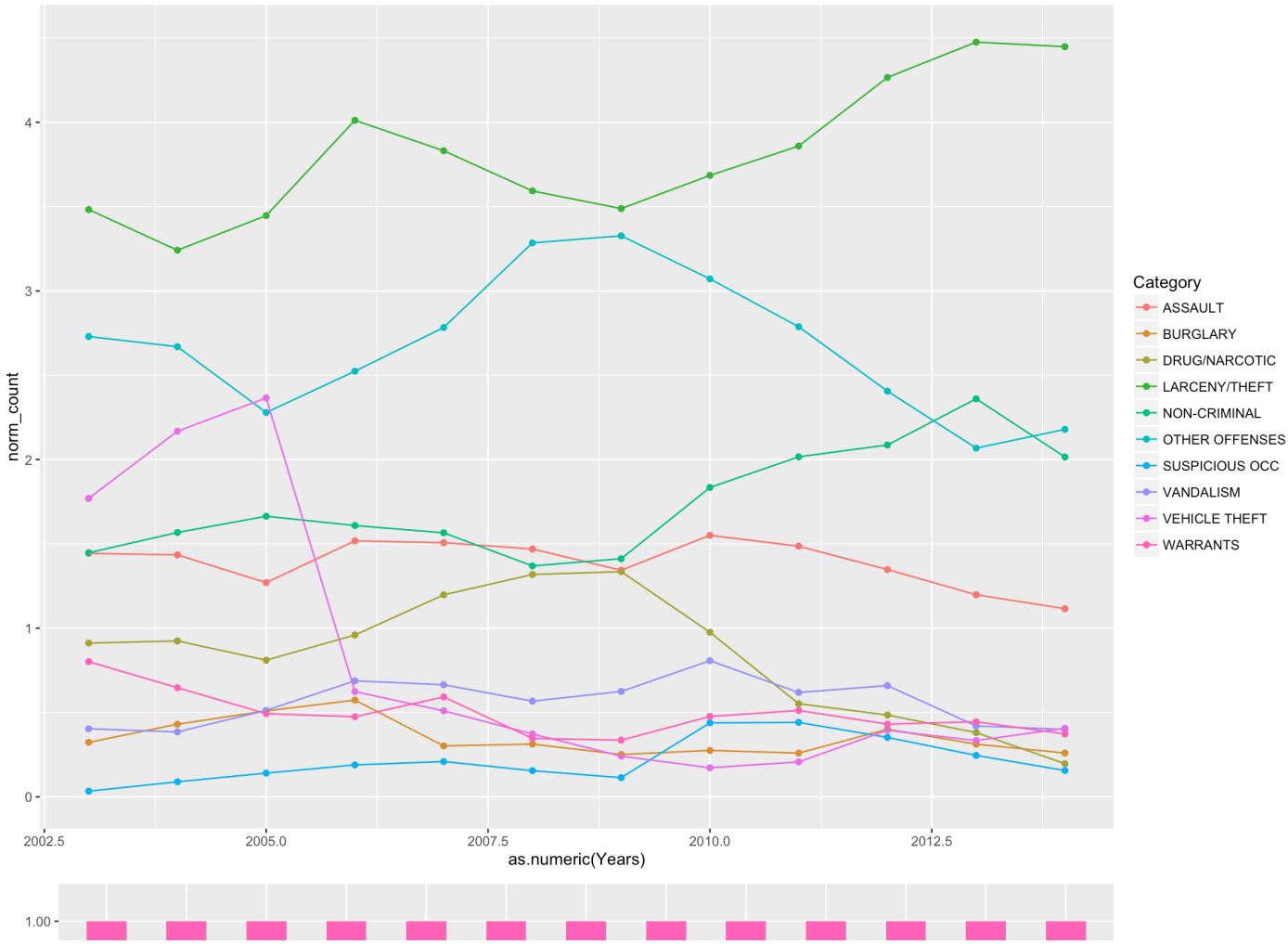
```
##                      Category Weekday Weekend  PerWeekday
## 1                       ARSON     862     651  -0.3468208
## 2                     ASSAULT   41639   35237  -6.0297519
## 3                  BAD CHECKS     279     127  24.4609665
## 4                     BRIBERY     157     132  -5.7057057
## 5                     BURGLARY   21443   15312   2.4534748
## 6           DISORDERLY CONDUCT    2569    1751   4.7787370
## 7  DRIVING UNDER THE INFLUENCE    1017    1251 -24.2458101
## 8                DRUG/NARCOTIC   34018   19953  12.2298835
## 9                  DRUNKENNESS    2012    2268 -20.0953137
## 10               EMBEZZLEMENT     710     456   7.7389985
## 11                  EXTORTION     150     106   2.9748284
## 12            FAMILY OFFENSES     296     195   6.4748201
## 13     FORGERY/COUNTERFEITING    6773    3836  13.9500322
## 14                     FRAUD    9908    6771   4.6472328
## 15                   GAMBLING      78      68  -7.5098814
## 16                  KIDNAPPING    1227    1114  -9.5243947
## 17               LARCENY/THEFT   96429   78471  -4.0779480
## 18                 LIQUOR LAWS    1093     810   0.5982513
## 19                   LOITERING     791     434  15.5025554
## 20              MISSING PERSON   14513   11476  -2.6441421
## 21                NON-CRIMINAL   51340   40964  -3.0942883
## 22              OTHER OFFENSES   75008   51174   4.7305222
## 23    PORNOGRAPHY/OBSCENE MAT      14       8  13.5135135
## 24                PROSTITUTION    4856    2628  16.1722488
## 25          RECOVERED VEHICLE    1994    1144  13.3169161
## 26                     ROBBERY   12904   10096  -2.1138869
## 27                     RUNAWAY    1129     817   1.7881292
## 28              SECONDARY CODES    5588    4397  -2.3986959
## 29       SEX OFFENSES FORCIBLE    2415    1973  -4.2742948
## 30   SEX OFFENSES NON FORCIBLE      83      65  -2.1611002
## 31              STOLEN PROPERTY    2729    1811   6.1110751
## 32                      SUICIDE     296     212   2.3041475
## 33               SUSPICIOUS OCC   18325   13089   2.4401152
## 34                         TREA       3       3 -14.2857143
## 35                    TRESPASS    4364    2962   4.9879711
## 36                    VANDALISM   23705   21020  -8.3540063
## 37                VEHICLE THEFT   29545   24236  -4.4773385
## 38                     WARRANTS   25643   16571   7.4329844
## 39                 WEAPON LAWS    4893    3662   0.1057046
```



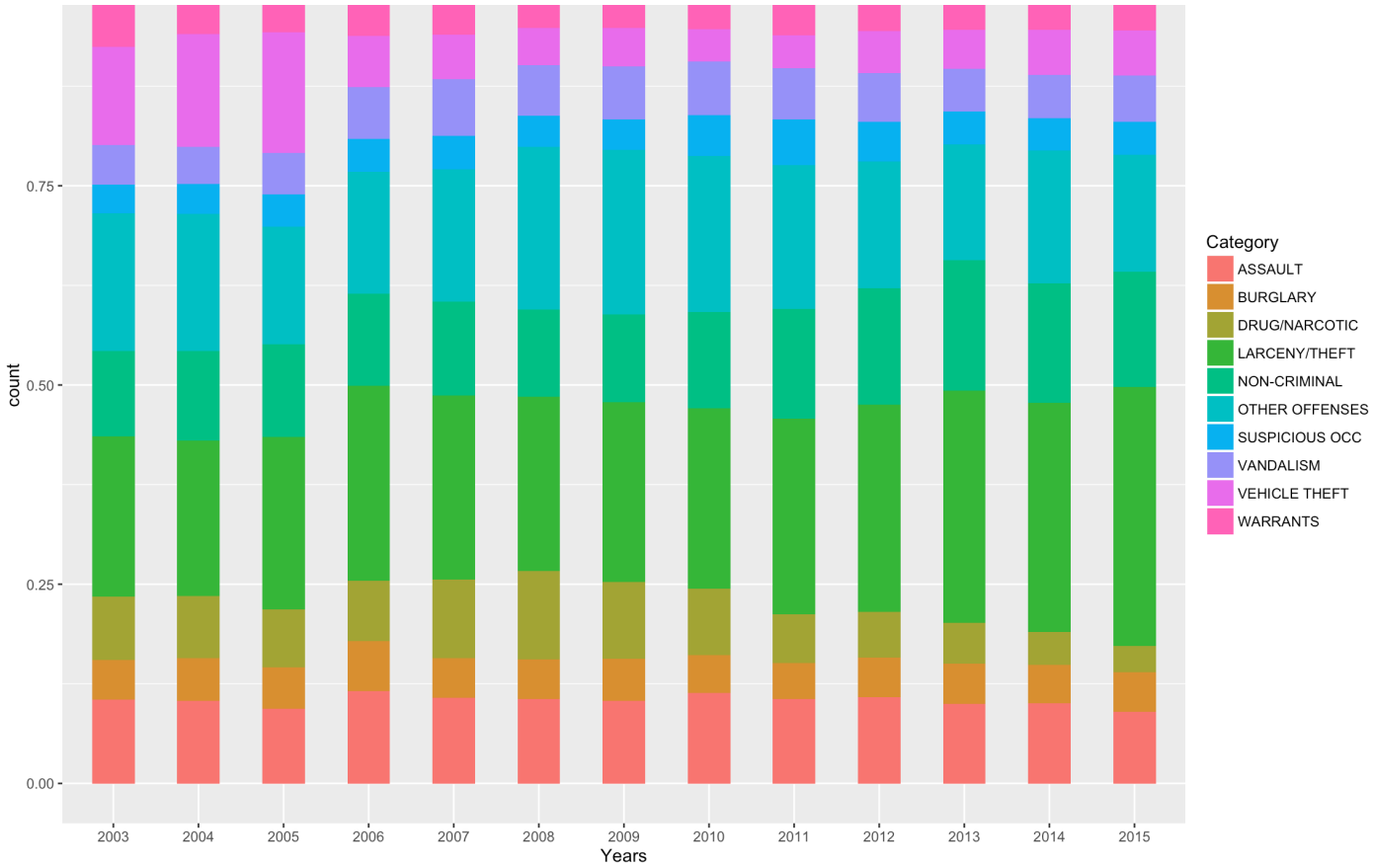
Types of crime vs Year

I next plotted crime variation across Years starting from 2003. In this part, I did not include the year 2015 because full data for 2015 was not available. Plots show a clear rise in Larceny/Theft and non-criminal incidences. In addition the plots show a sharp decline in vehicle theft in 2006. To obtain a better understanding of the crime rates, I normalized the crime count by subtracting the mean for each year and dividing by standard deviation. This allowed for a fairer comparison.

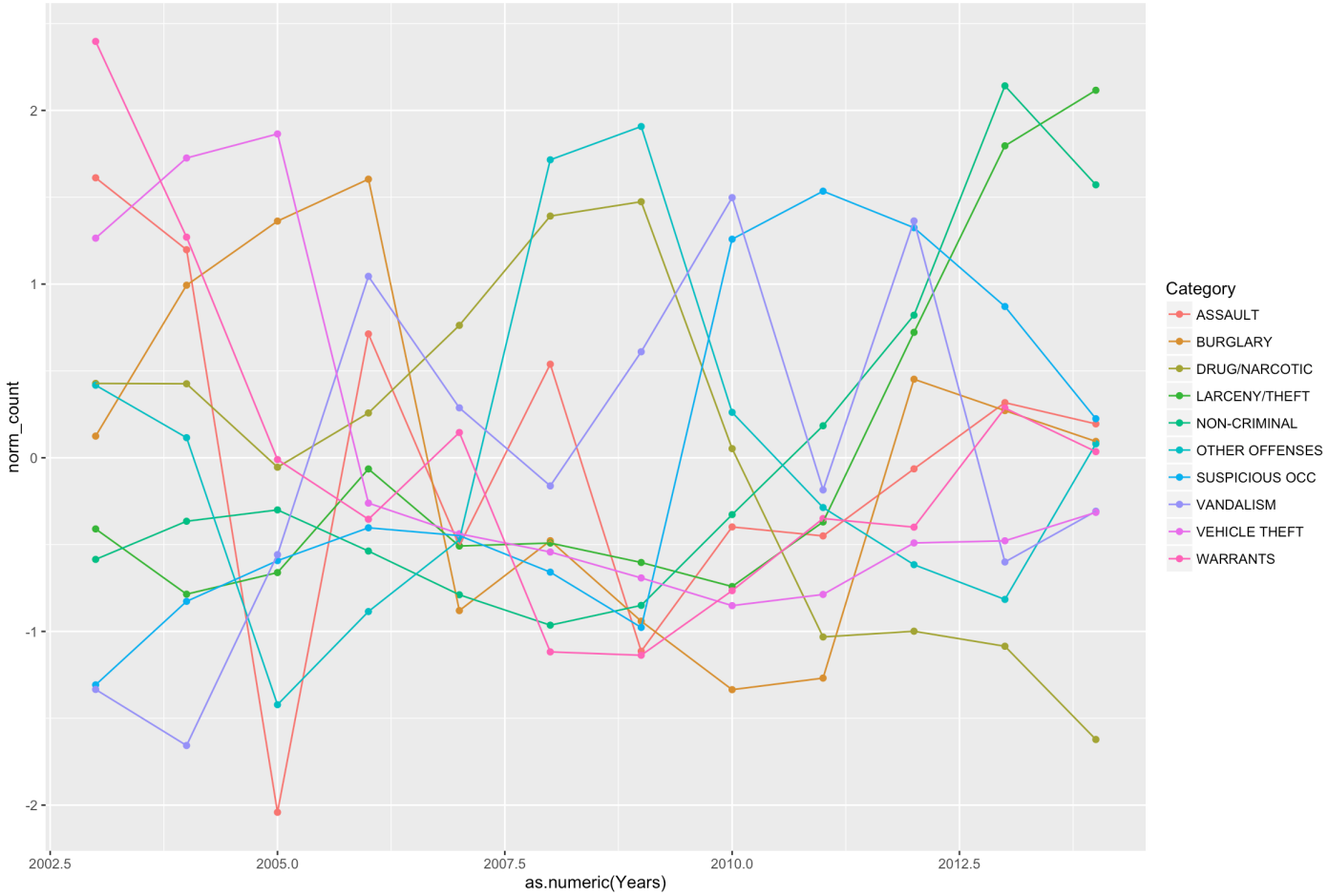

15000
10000
5000
count
2002.5
2005.0
2007.5
2010.0
2012.5
as.numeric(Years)
Category
ASSAULT
BURGLARY
DRUG/NARCOTIC
LARCENY/THEFT
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS


4
3
2
1
0
norm_count
2002.5
2005.0
2007.5
2010.0
2012.5
as.numeric(Years)
Category
ASSAULT
BURGLARY
DRUG/NARCOTIC
LARCENY/THEFT
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS

1.00

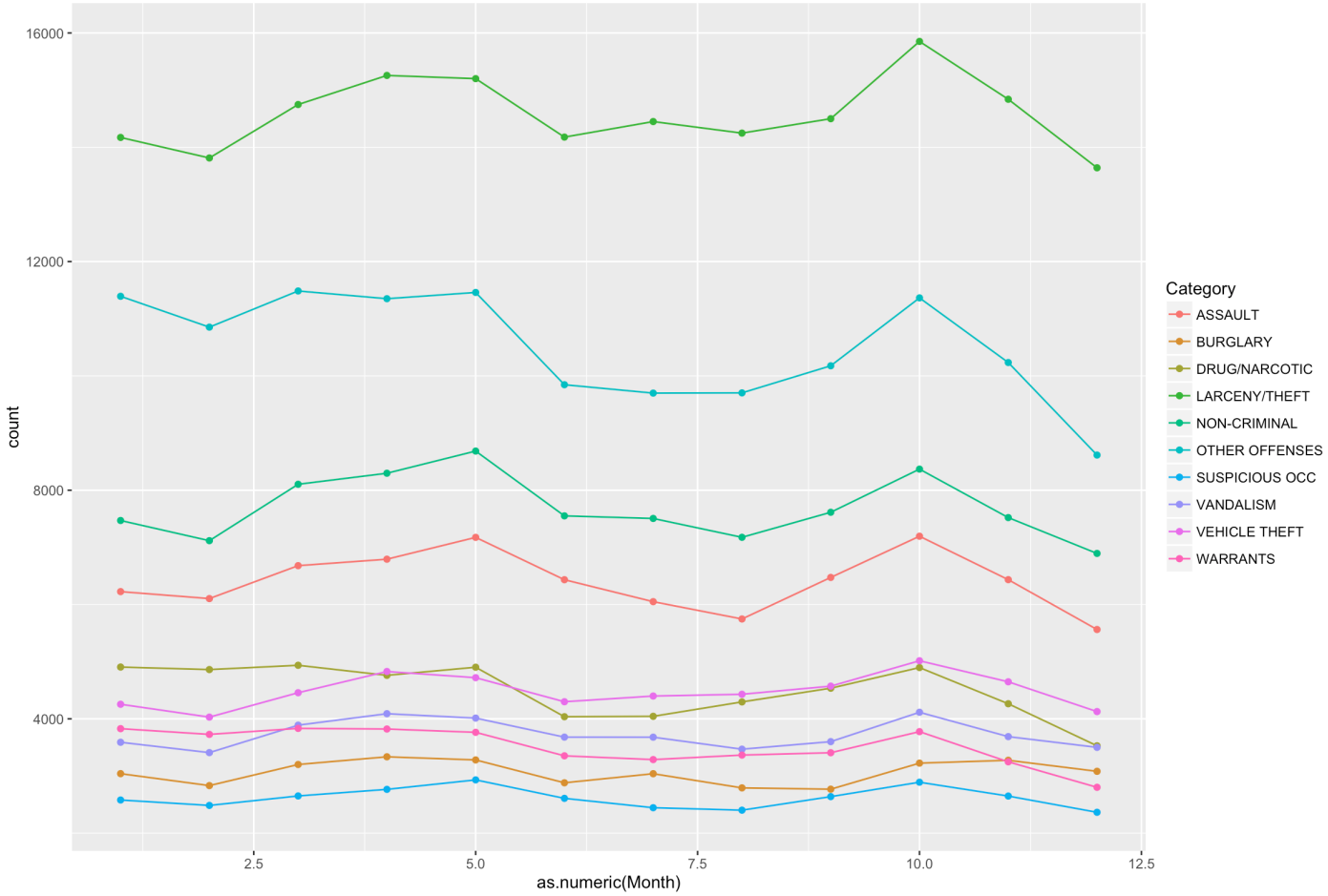


Plots above show rise in Larceny and non-criminal crimes since 2010. Further, vehicle thefts and drug/narcotics related offence are on decline. These plots show that the distribution of crimes is different for different years. Therefore for predicting the current Category of crime, I will use data from 2012 onwards alone. I wanted to test if there is an overall year-dependent pattern in number of crime. I therefore normalized the number of crime in each year by subtracting the mean and dividing by the standard deviation. The crime data shows no year-dependent trend that is consistent across different crime categories.

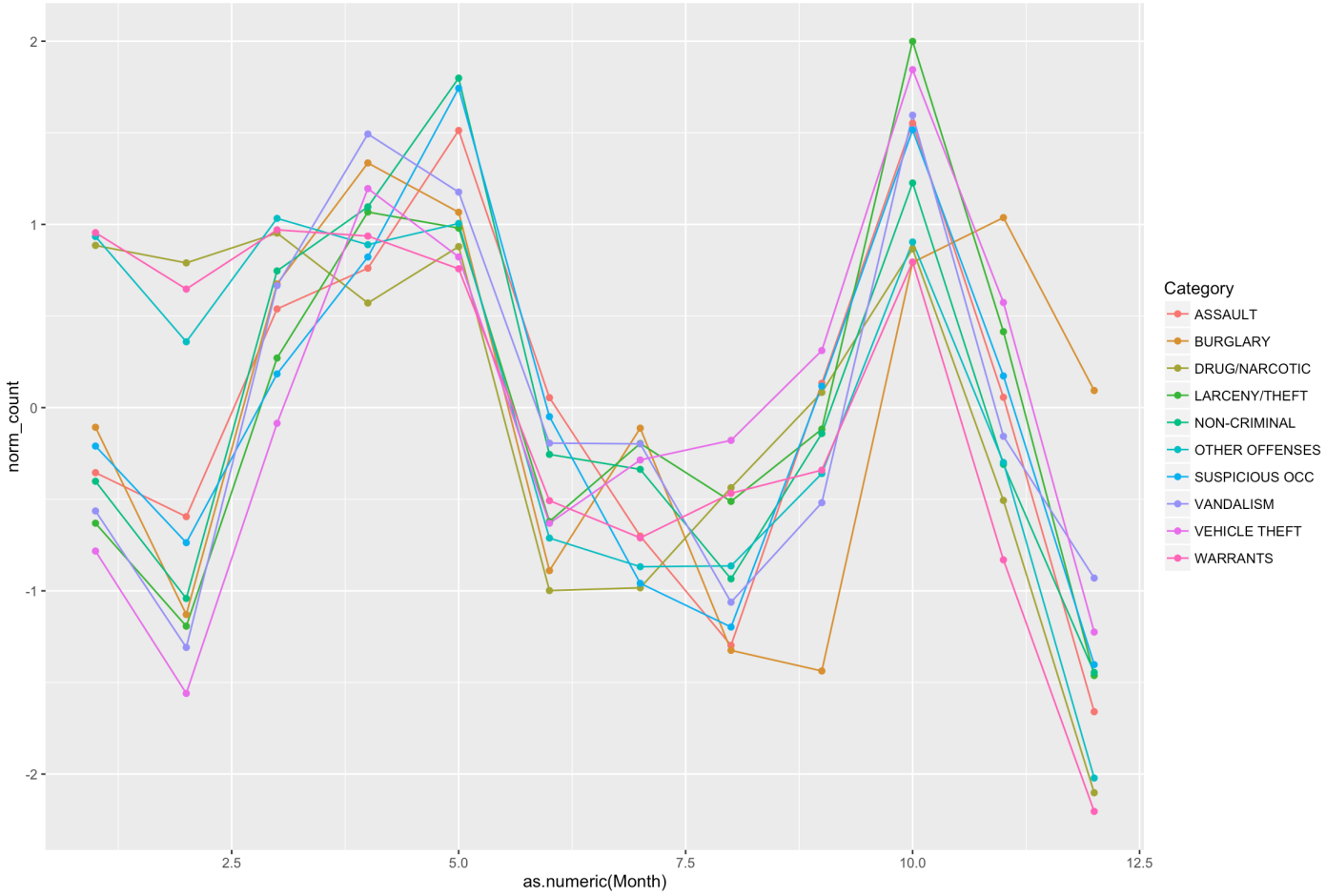


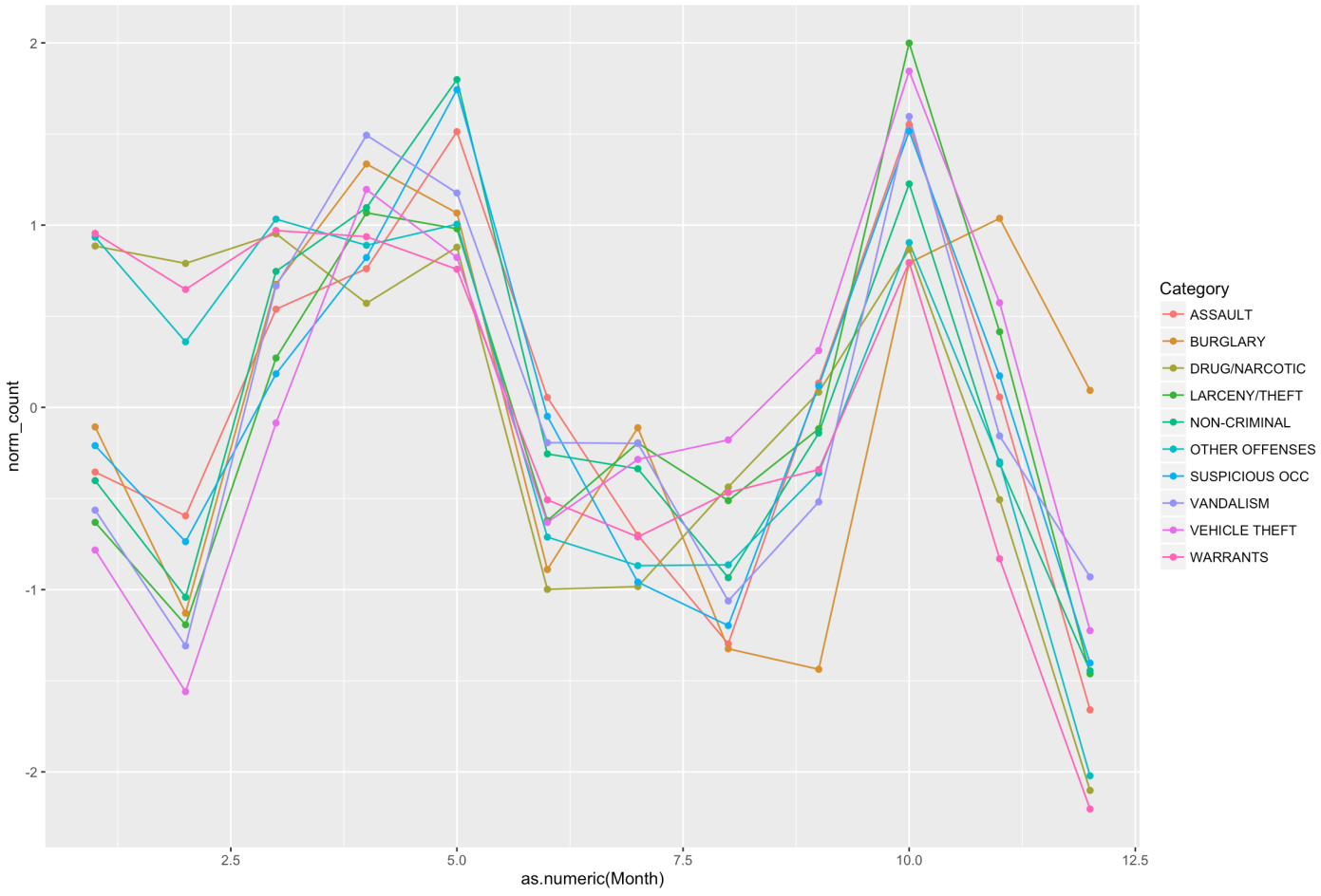
Types of crime vs Month

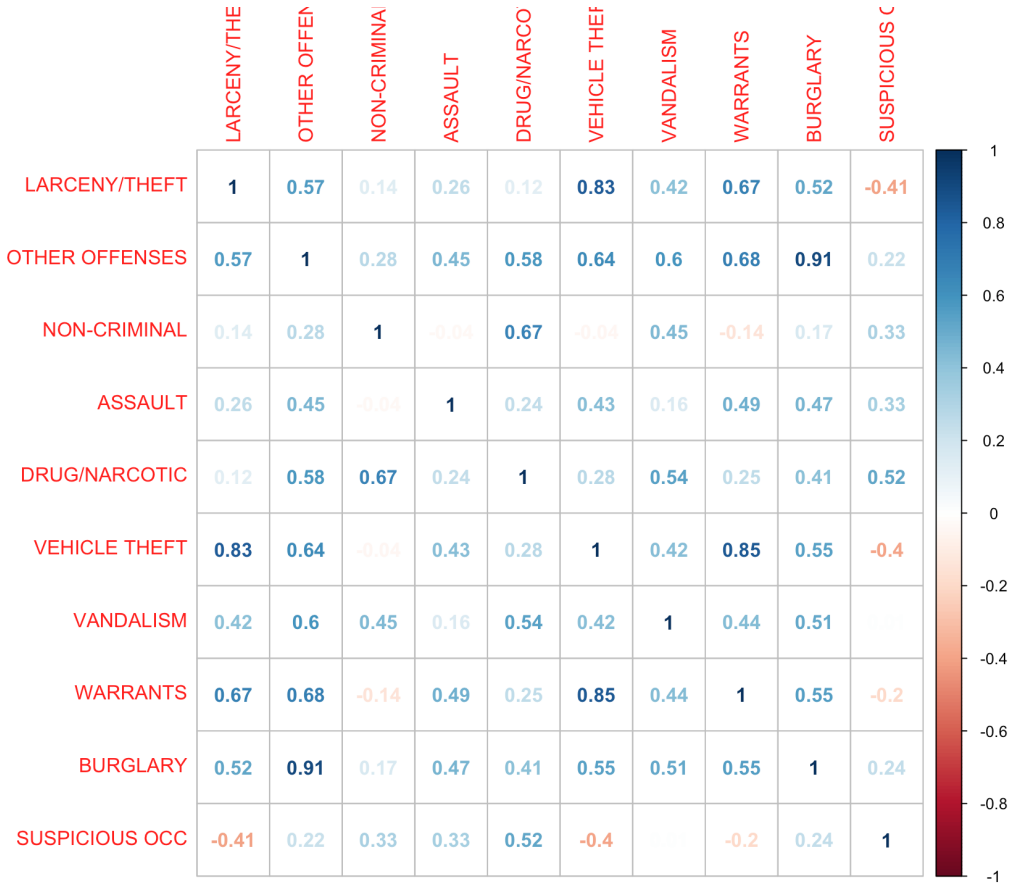
I next plotted crime variation across different months to check if there are any seasonal patterns in crime rates. First plot indicates that there may be a time-trend in number of crime vs month. For assault, Larceny, other offenses and non-criminal offenses, the number of crimes show a strong correlation. I therefore, normalized the number of crimes in each month by subtracting the mean and dividing by standard deviation. After doing this, a clear month-dependent trend emerged.



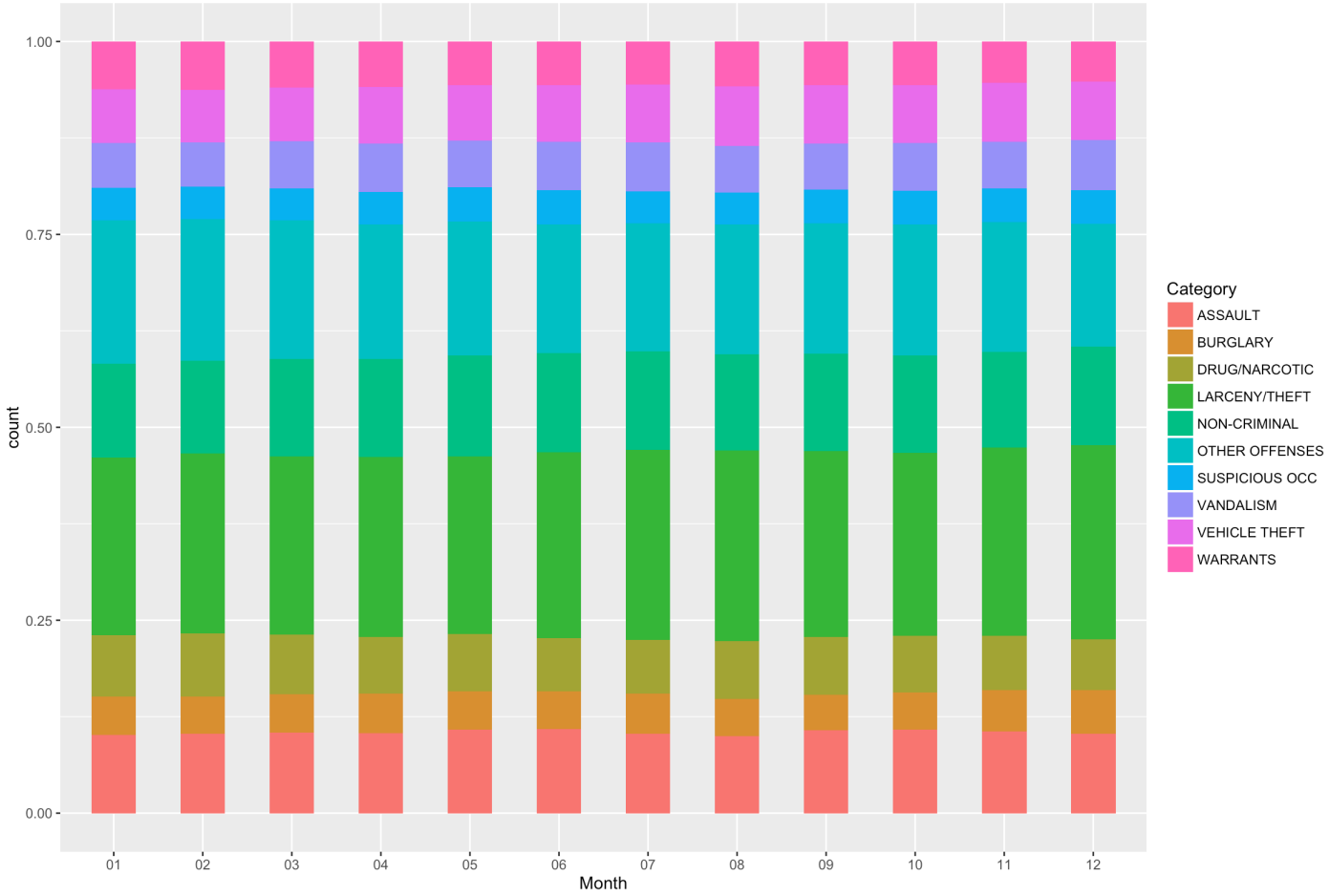
After normalizing, its clear that there is a strong month-dependent trend in crime incidents. These plots show a high correlation in monthly crime incidents across different categories.


norm_count
as.numeric(Month)
Category
ASSAULT
BURGLARY
DRUG/NARCOTIC
LARCENY/THEFT
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS


norm_count
as.numeric(Month)
Category
ASSAULT
BURGLARY
DRUG/NARCOTIC
LARCENY/THEFT
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS

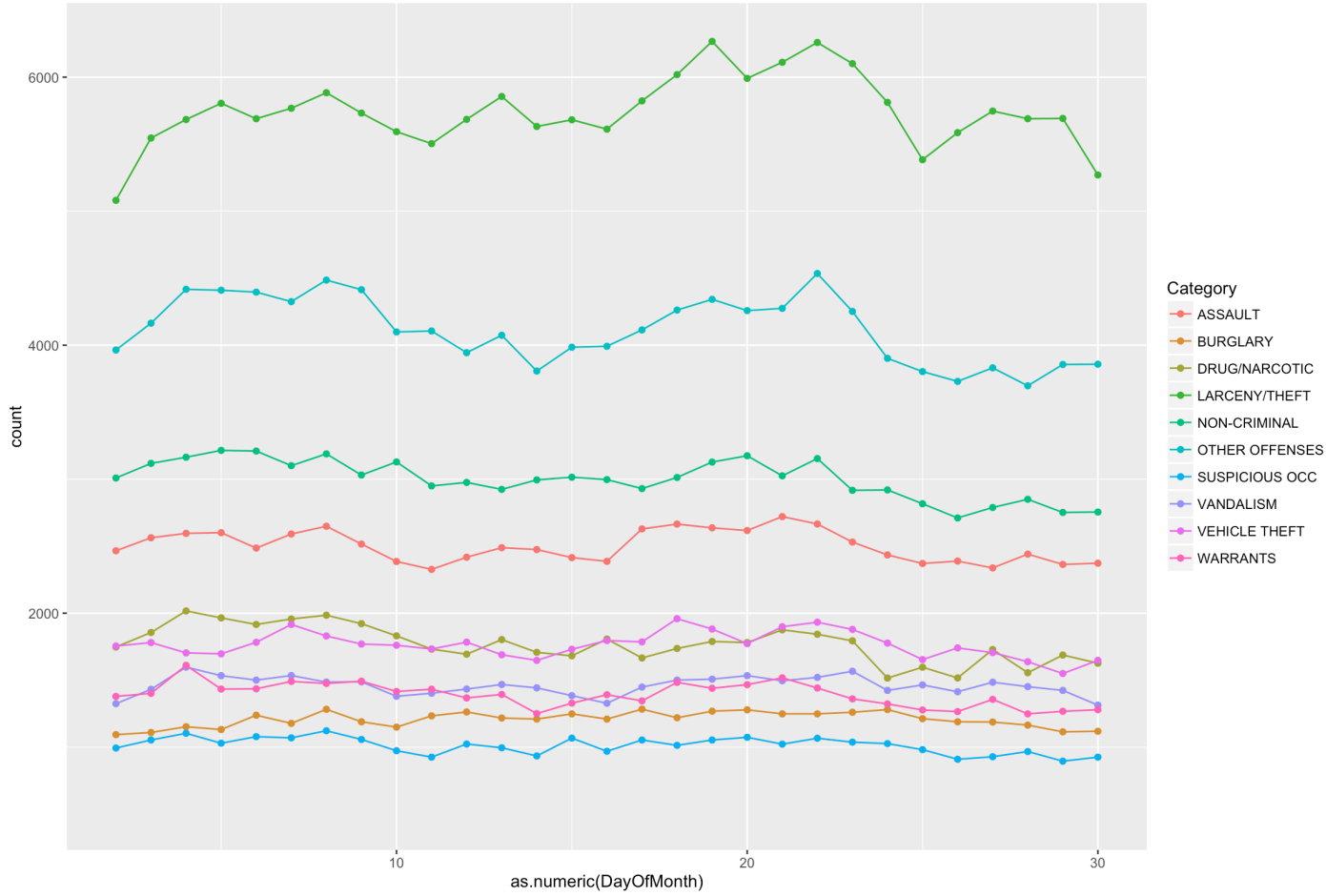


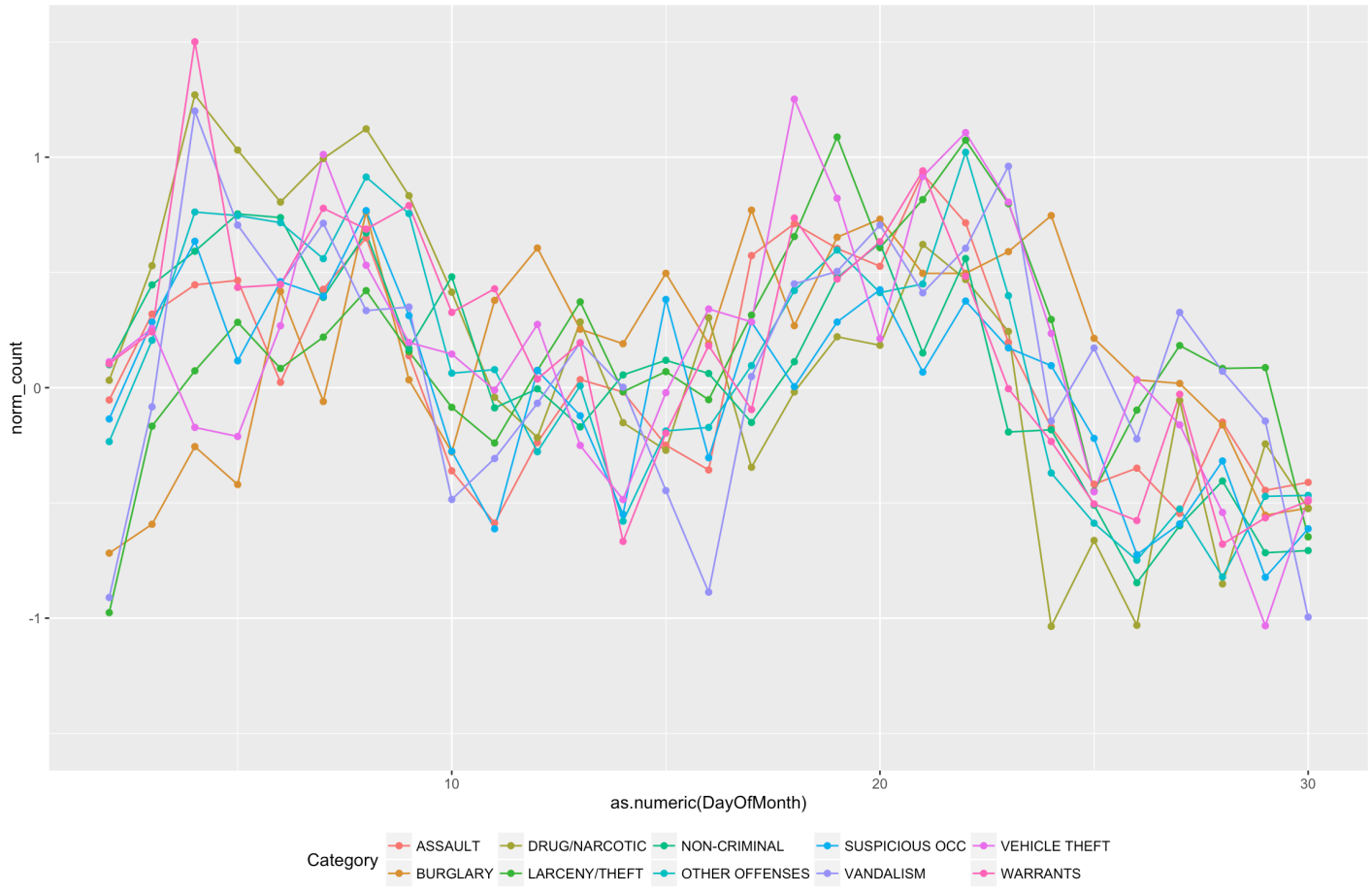
Plots above show that there is a strong correalion between crime categories across months. Therefore, it may be possible to drop the month from prediction of crime category. Plot below shows that the relative ratios of crime remains relatively unchanges. Therefore, there is weak effect of month on crime category, and month can be dropped from final prediction model.



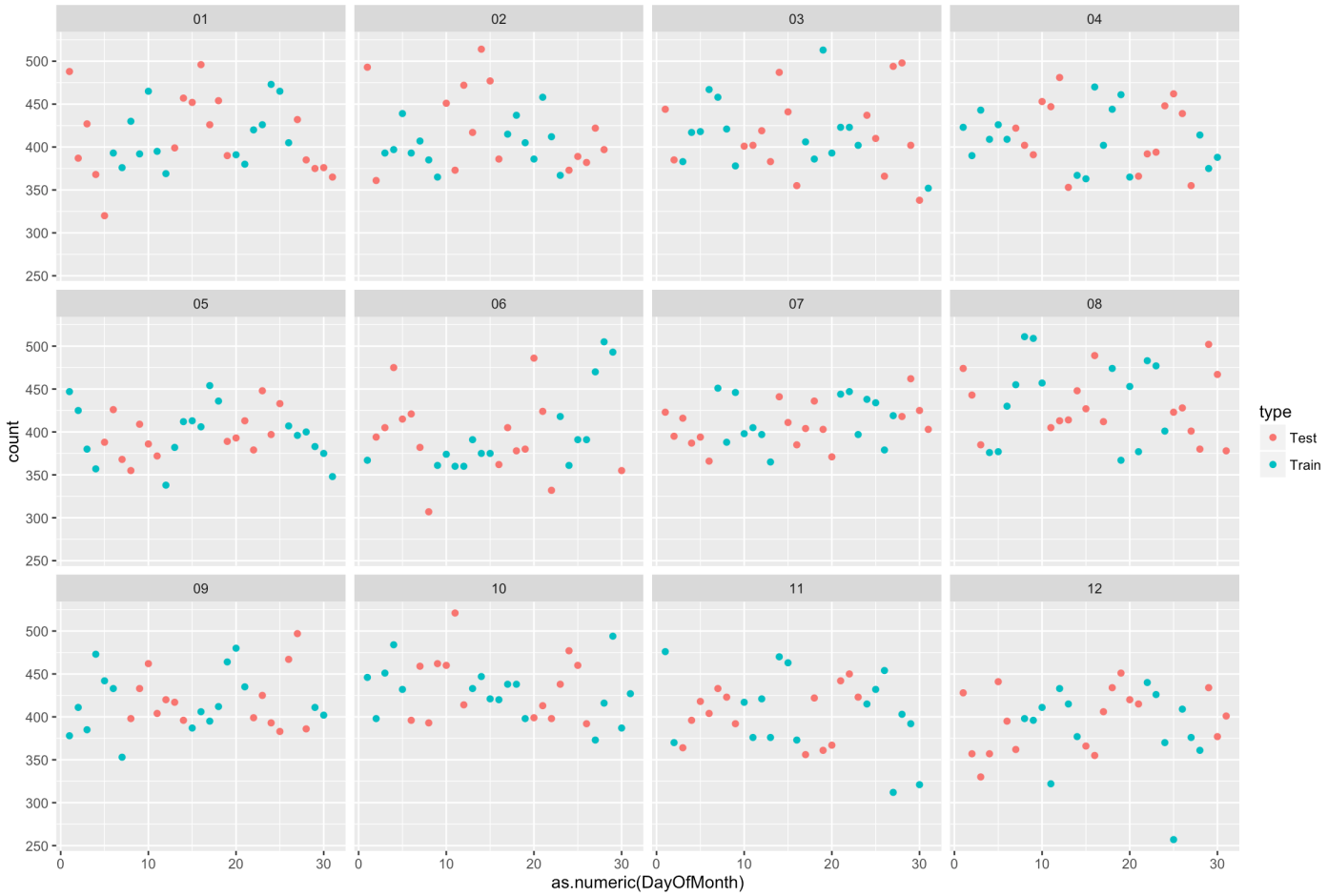
Types of crime vs Day of the month

I next plotted crime variation across hours of the day. I did not include the dates with value 31 and 1. I removed 31 because the number of months with 31 days is much less than the number of days with 30 days. I removed 1 from exploratory analysis because in most cases 1 may have been used as the default date. Plots below indicate a strong day-dependent trend in the crime number. This trend is clear when I normalized the crime count by subtracting the mean and dividing by the standard deviation. Days 10-16 and 25-30 have lower crime incidents than days between 5 and 10, and 15 to 20. Plot of fraction of crime vs day of the month also shows that the fraction of crime varies across the day of the month.


6000
4000
2000
count
10
20
30
as.numeric(DayOfMonth)
Category
ASSAULT
BURGLARY
DRUG/NARCOTIC
LARCENY/THEFT
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS

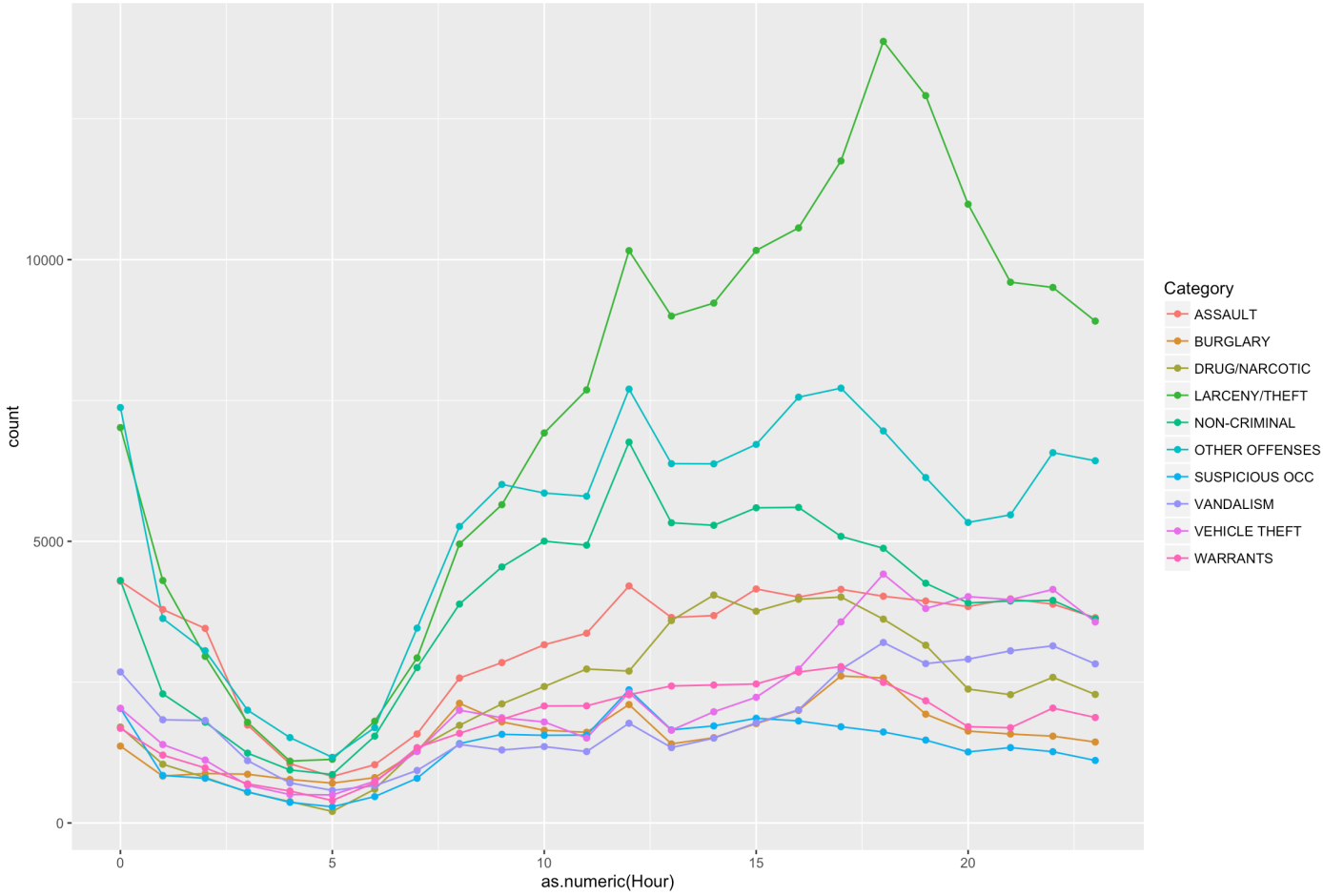

1
0
-1
norm_count
10
20
30
as.numeric(DayOfMonth)
Category
ASSAULT
BURGLARY
DRUG/NARCOTIC
LARCENY/THEFT
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS

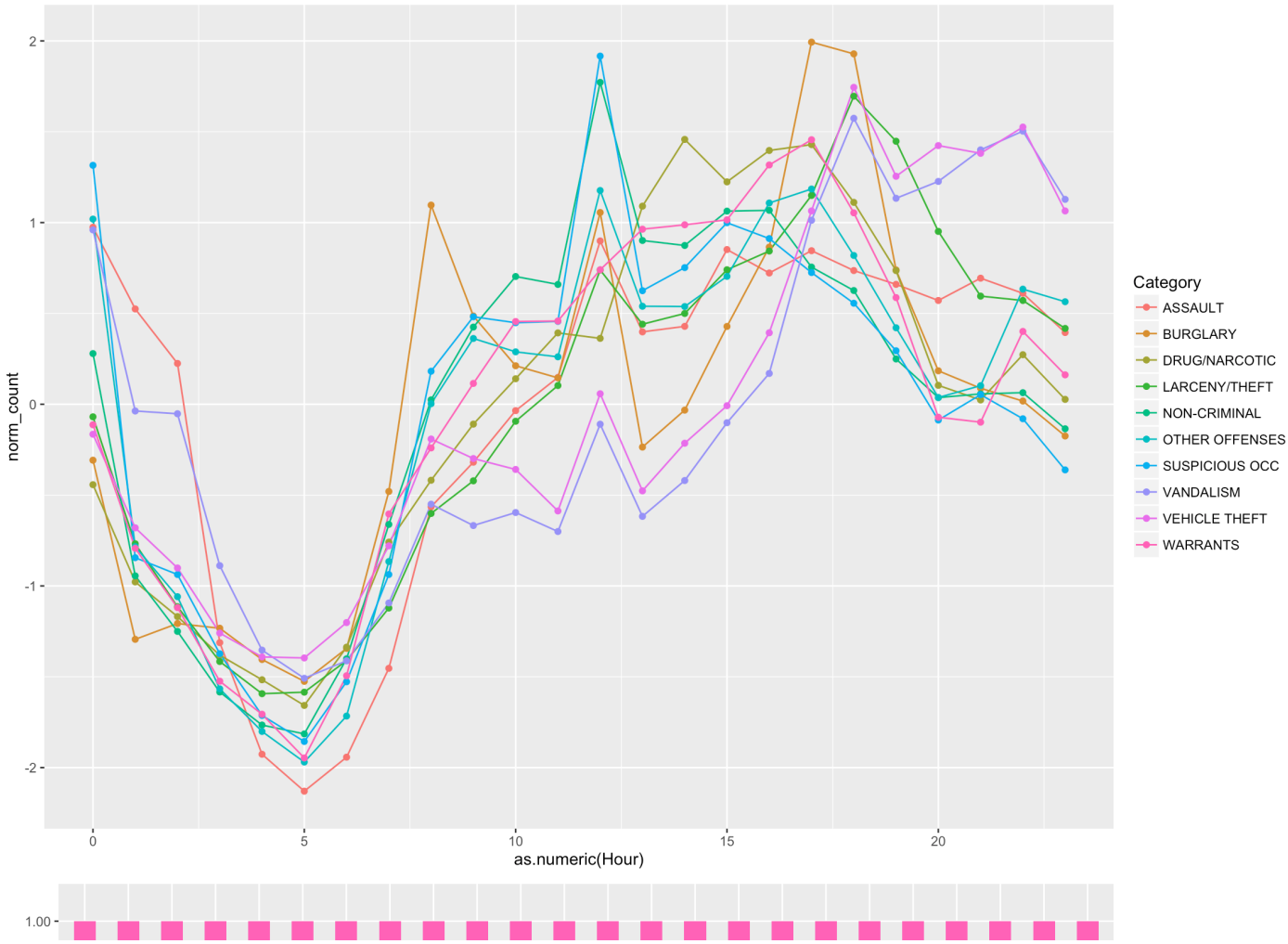
To check the reason for this pattern, I plotted test and training data from 2014. Plots below show that the data was split in such a way that the data for every other week was assigned to train or test data sets. Therefore, I will not use day of the month for building the model.



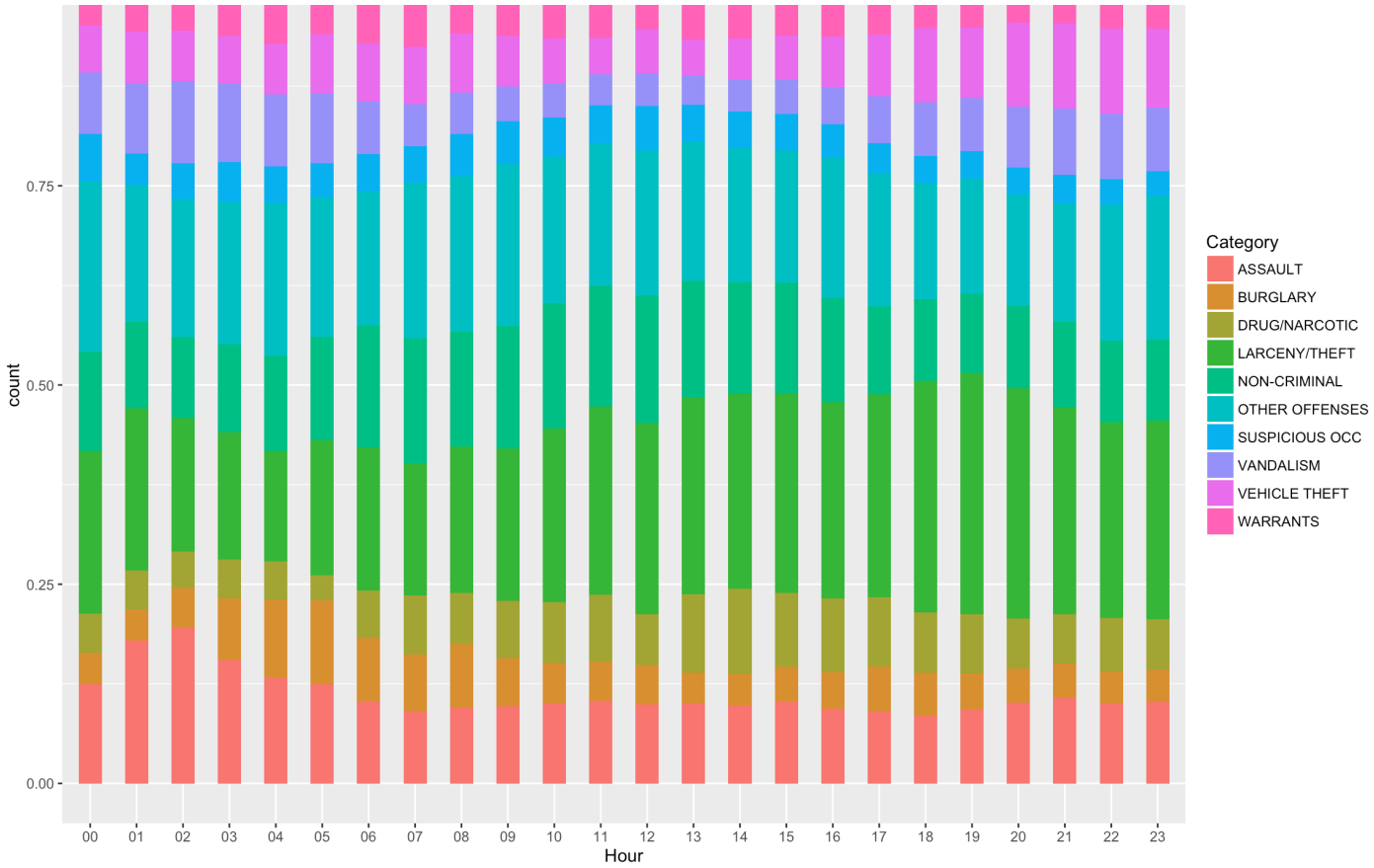
Types of crime vs Hour

I next plotted crime variation across different hours of the day. Plots show a greater criminal activity between 10 am and midnight, and a sharp drop around 5 pm. Similar trend was observed across all crime catergories after normalizing by substracting the mean and dividing by standard deviation. As hour affects crime incidents, it will be included in the model to predict probability of the crime category.

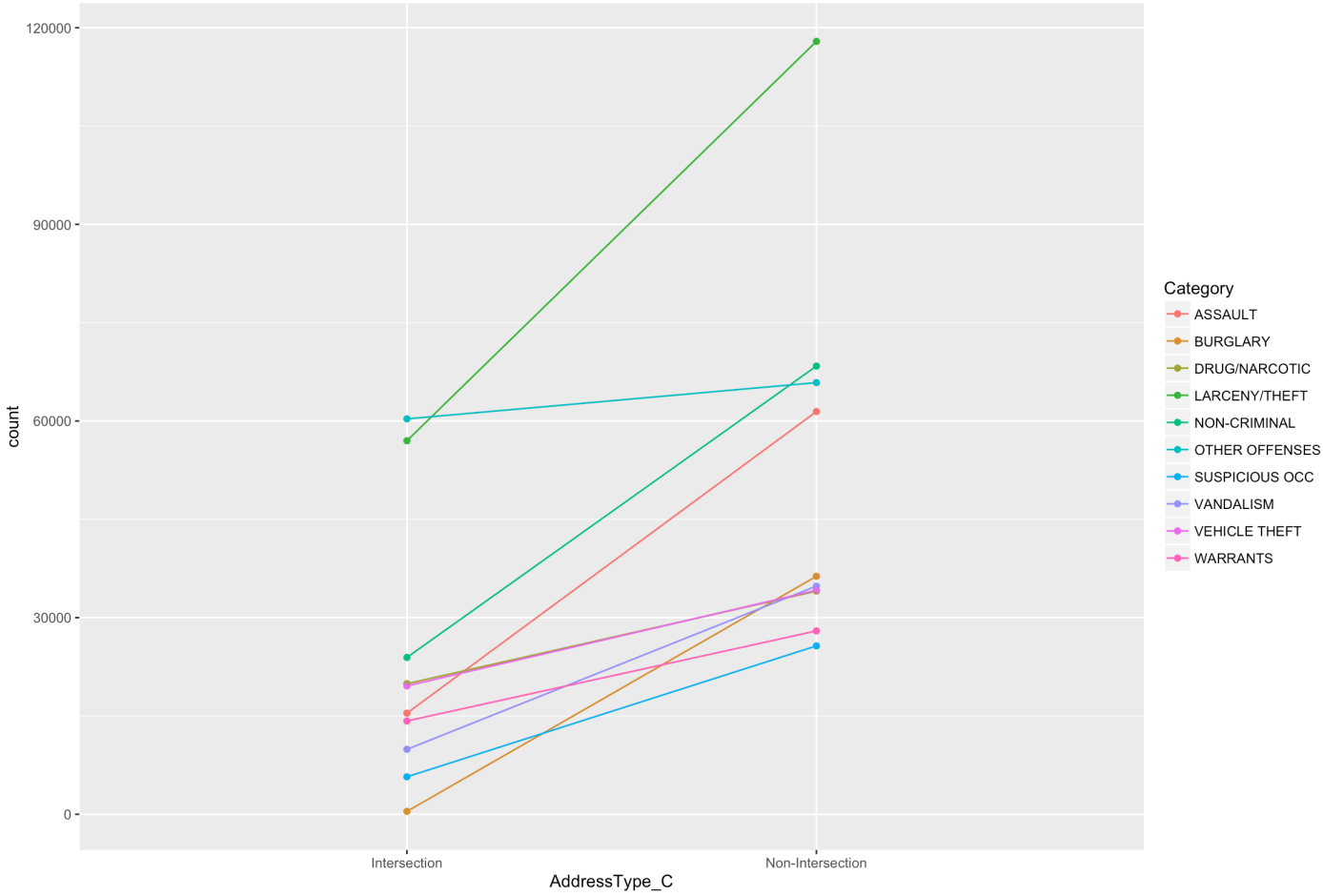

count
10000
5000
0
0
5
10
15
20
as.numeric(Hour)
Category
ASSAULT
BURGLARY
DRUG/NARCOTIC
LARCENY/THEFT
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS

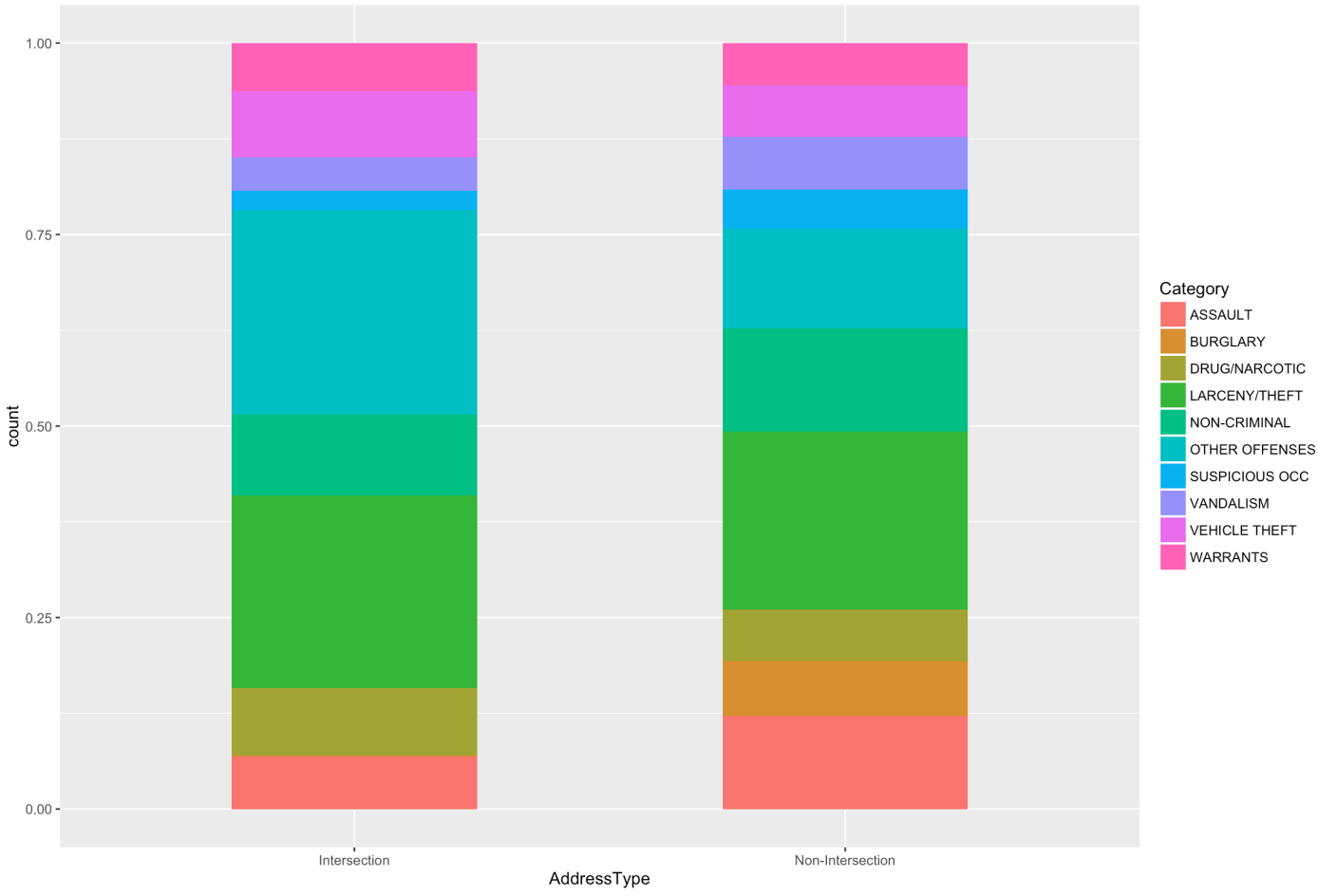

norm_count
2
1
0
-1
-2
0
5
10
15
20
as.numeric(Hour)
Category
ASSAULT
BURGLARY
DRUG/NARCOTIC
LARCENY/THEFT
NON-CRIMINAL
OTHER OFFENSES
SUSPICIOUS OCC
VANDALISM
VEHICLE THEFT
WARRANTS

1.00



Types of crime vs Addresstype

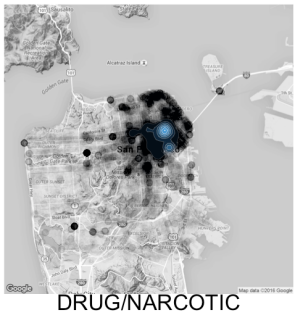




Crime category by location

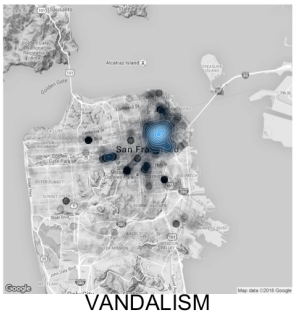
I next plot contours of crime distributions across San Francisco in year 2014 only. I chose 2014 because crime trends were affected by the years. Further, including only 2014 data signifianctly reduced the number of data points, which helped in faster plotting. Plots below show that the type of crime heavily depends on the location on the map. For example, larceny was more concentrated in the north east area of the map, where as vehicle theft is more evenly spread across the eastern region of the map.

LARCENY/THEFT



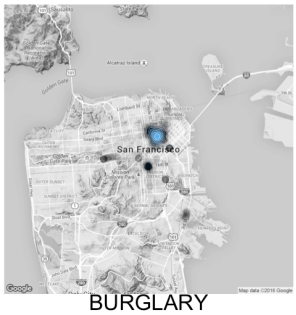
OTHER OFFENSES

NON-CRIMINAL

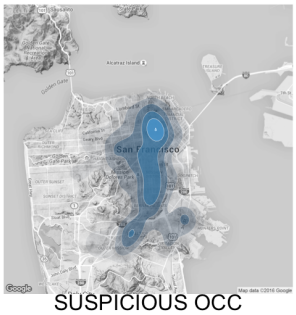


ASSAULT

DRUG/NARCOTIC



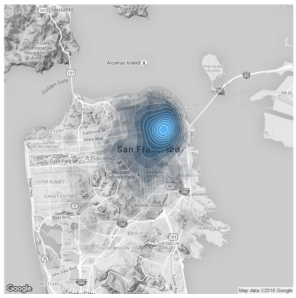
VEHICLE THEFT



VANDALISM

WARRANTS

BURGLARY



SUSPICIOUS OCC

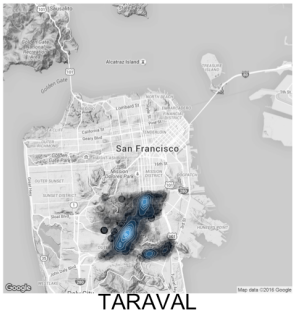
Crime category by Police District

I next plot contours of crime distributions across San Francisco in year 2014 only classified by police district. Plots below confirm that the type of crime heavily depends on the location on the map. These crimes however indicate the regions of police districts, and the crime rates may be different in each one.

NORTHERN

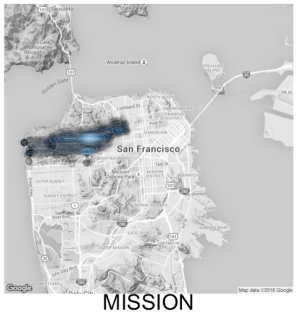
PARK

INGLESIDE



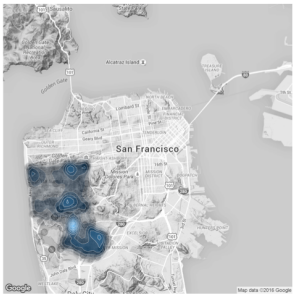
BAYVIEW

RICHMOND



CENTRAL

TARAVAL



TENDERLOIN

MISSION

SOUTHERN

Bivariate Analysis

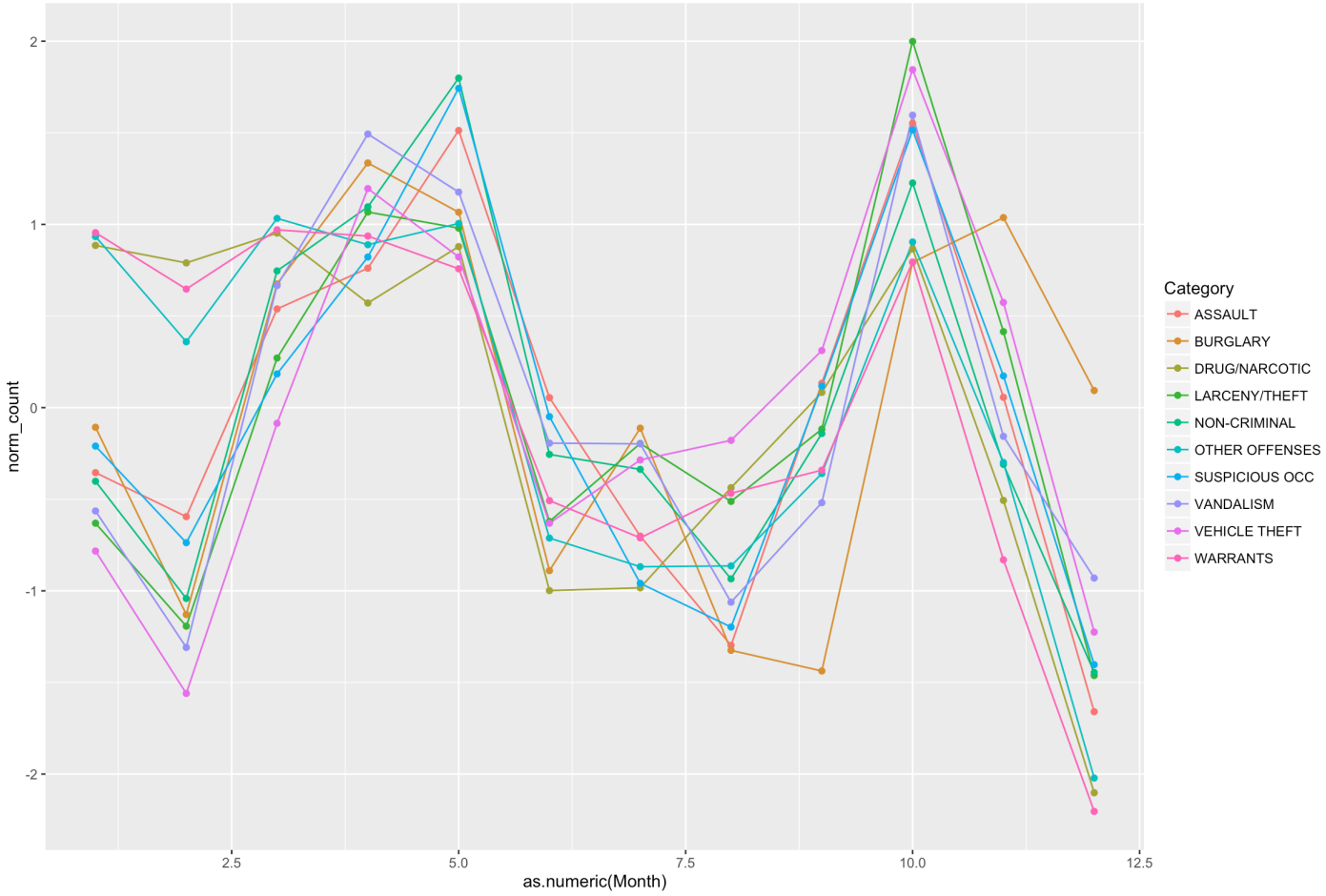
Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

From the analysis above, the main factors that affect crime rates are, - Month: Crime numbers showed a seasonal pattern where crime was less during February, and higher from march to may, and in october. - Hour of day: Crime incidents varied with the hour of the day. The numbers dropped gradually from midnight to 5 am, and rose after that until midnight - Day of the month: There was minor variation in crime numbers with the day of month. Crime rates were higher from 5th to 10th and from 18th to 22. - Day of the week: - Geographic location: Of all the variables, the crime rate was most affected by the geographic location of the crime. Crimes were higher in eastern region of them map. Further investigation revealed that these crimes were localized around certain areas. Some crimes, like vehicle theft were spread more evenly over larger area than others.

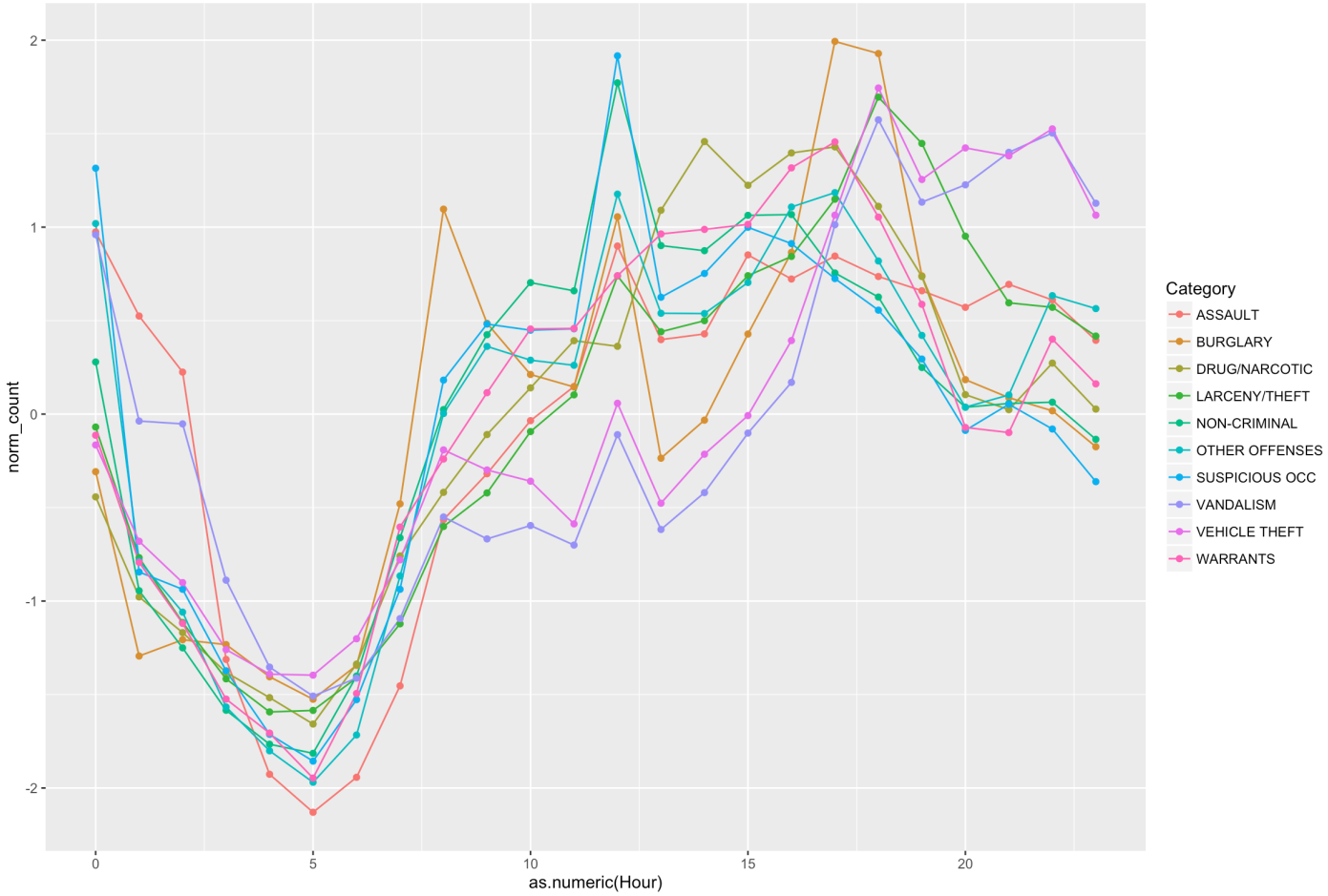
What was the strongest relationship you found?

Strongest relations in bi-variate section were,

1- Hour of the day, 2- Month, 3- Day of week, 4- Location 5- Years



Another significant factor that influences crime incidences is the hour of the day. The plot below shows a steady decline from midnight to 5 am and rise of the



Multivariate plots

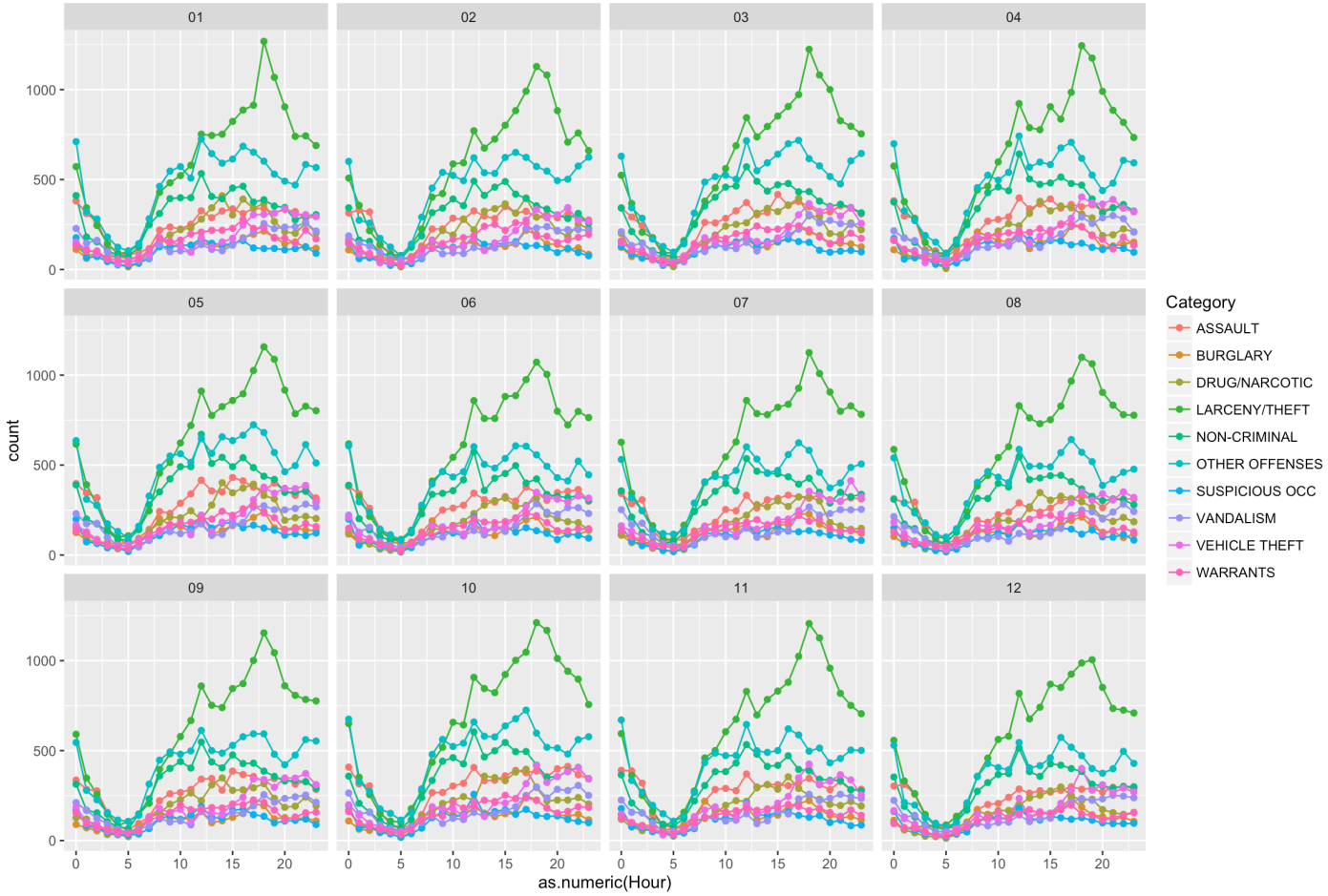
From the analysis above, the main factors that affect crime rates are,

- Month: Crime numbers showed a seasonal pattern where crime was less during February, and higher from march to may, and in october.
- Hour of day: Crime incidents varied with the hour of the day. The numbers dropped gradually from midnight to 5 am, and rose after that until midnight
- Day of the month: There was minor variation in crime numbers with the day of month. Crime rates were higher from 5th to 10th and from 18th to 22. - Day of the week:
- Geographic location: Of all the variables, the crime rate was most affected by the geographic location of the crime. Crimes were higher in eastern region of them map. Further investigation revealed that these crimes were localized around certain areas. Some crimes, like vehicle theft were spread more evenly over larger area than others.

In this section, I investigate variations in crime rate based on different combinations of the main factors identified from previous exploratory analysis.

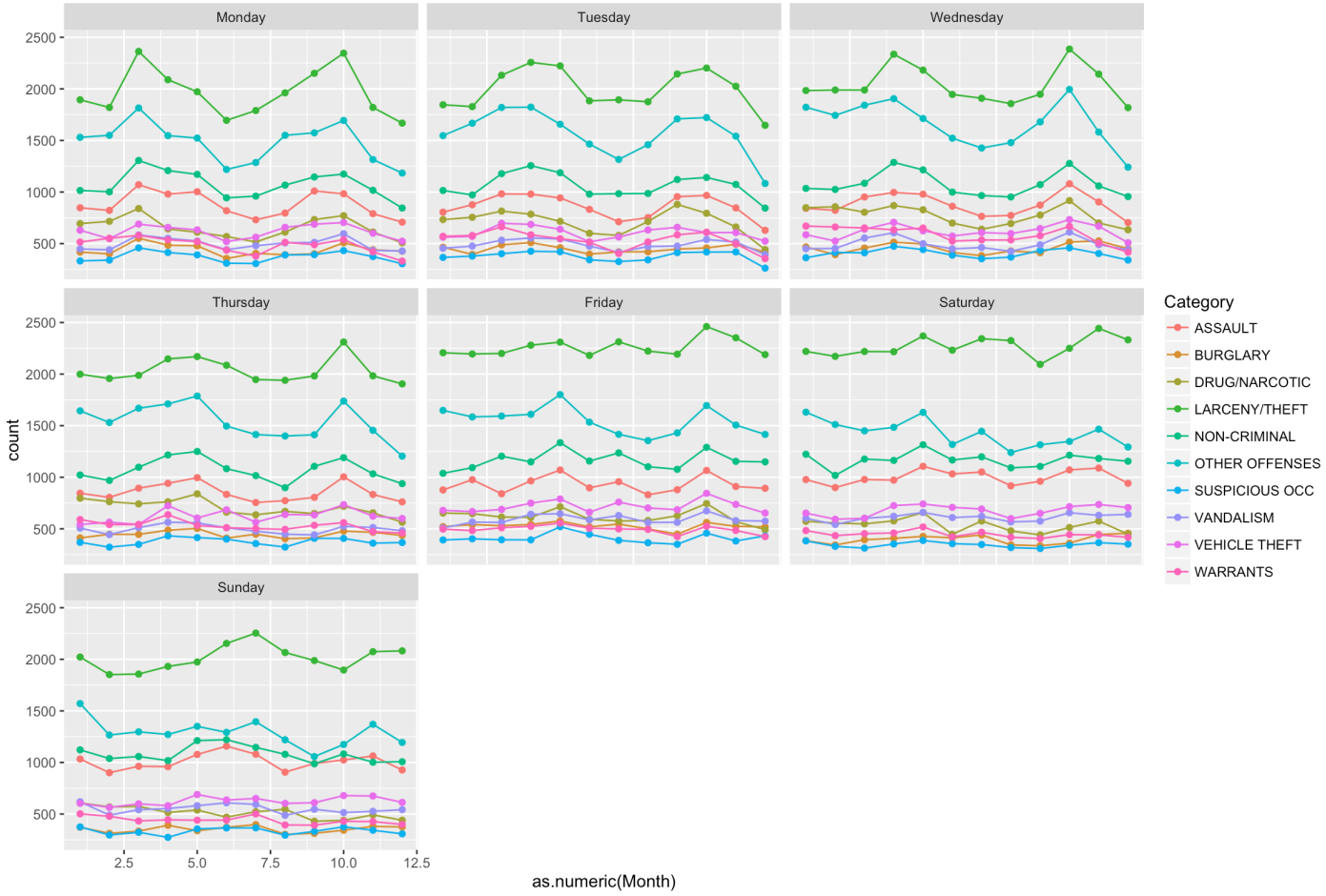
Crime rate vs Month and Hour of day,

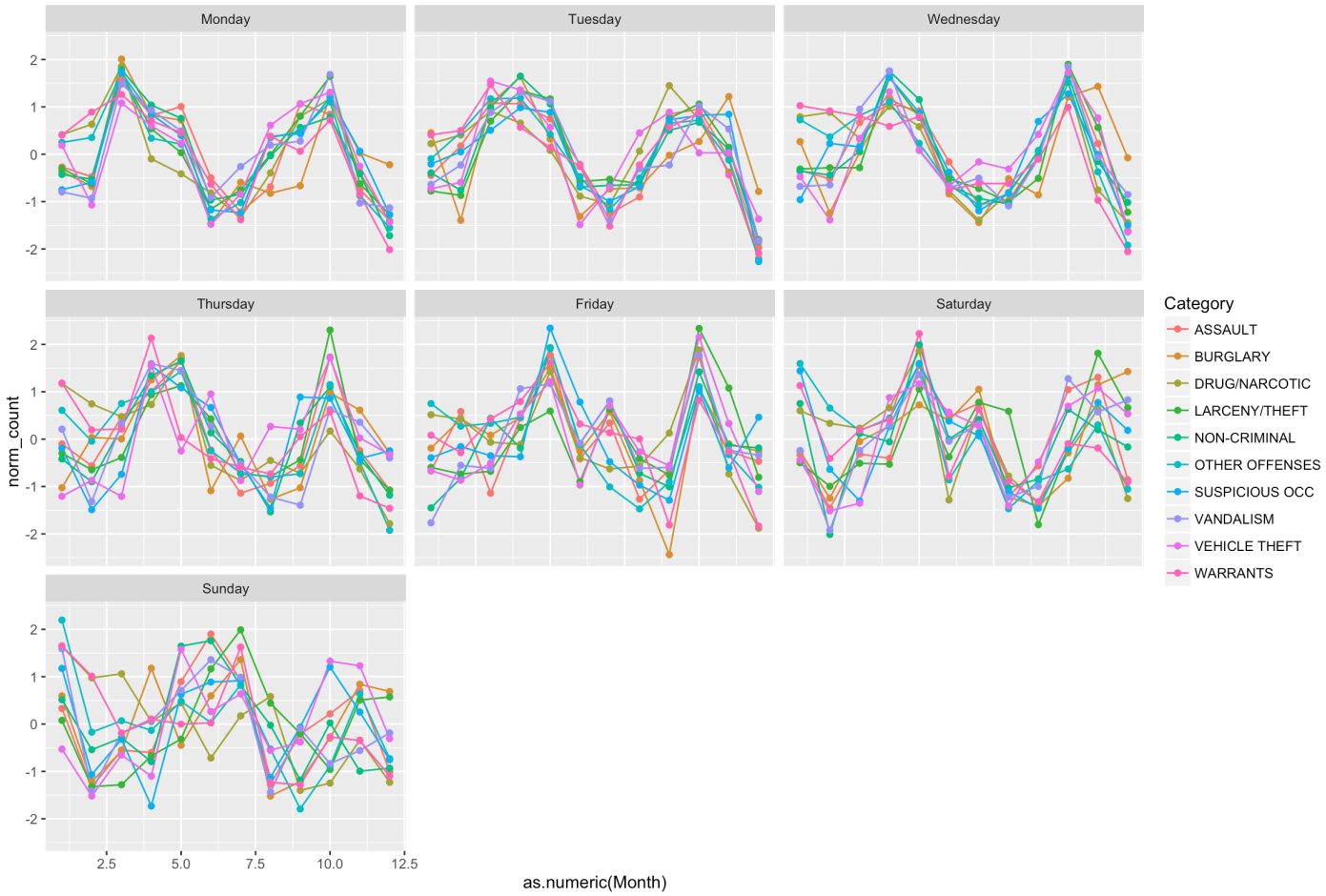
Crime vs hour faceted by month indicates that the hourly patterns are maintened for all the months, and the patterns do not appear to be different in different months.



Crime rate vs Month and DayOfWeek,

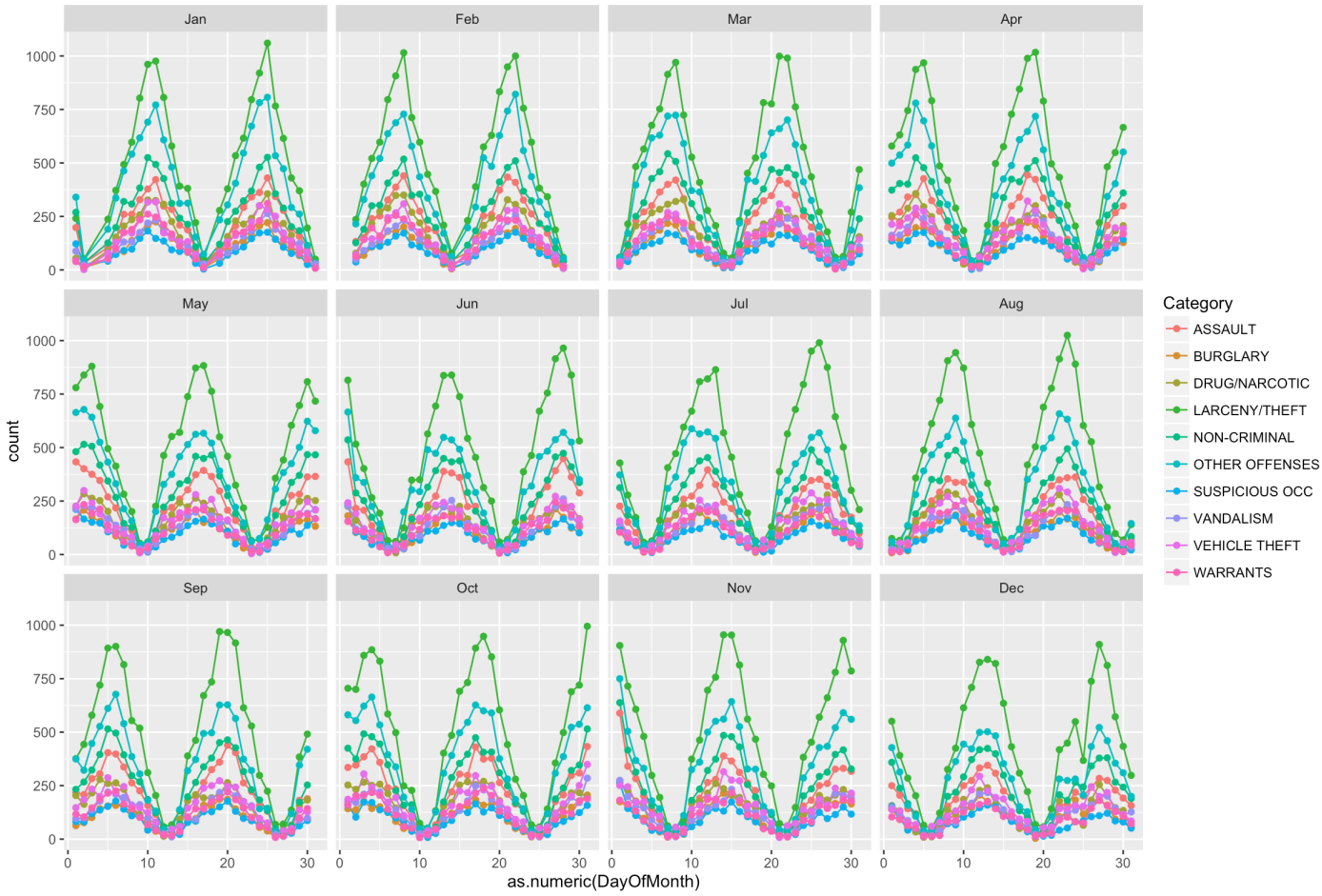
Crime vs Month faceted by the day of week indicates that the patterns of crime in each month is different for different days of the week. Monday, tuesday and wednesday have a bimodal pattern where crimes rise around march and in october. For thursdays and fridays, the crimes peak in october. However, this peak is not observed for saturday and sunday. Further, all days have high Larceny/theft incidences and about even distribution of other crime types. However, on sunday and saturday, assult, non-criminal and other offeneses are concentrated around 1000, and other crimes are lower. Except larceny, all other crimes are lower on weekends.



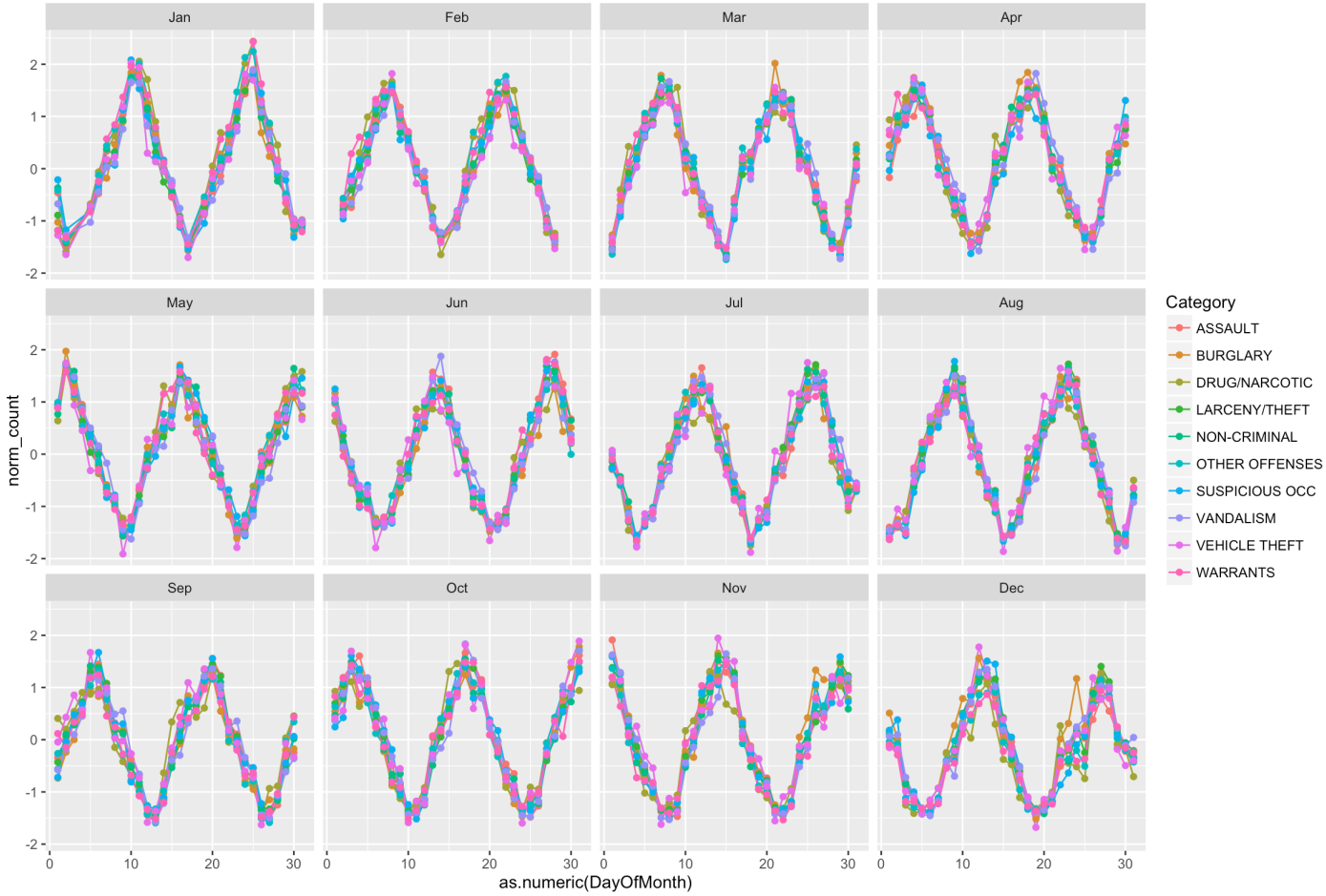


Crime rate vs Month and day of the month.

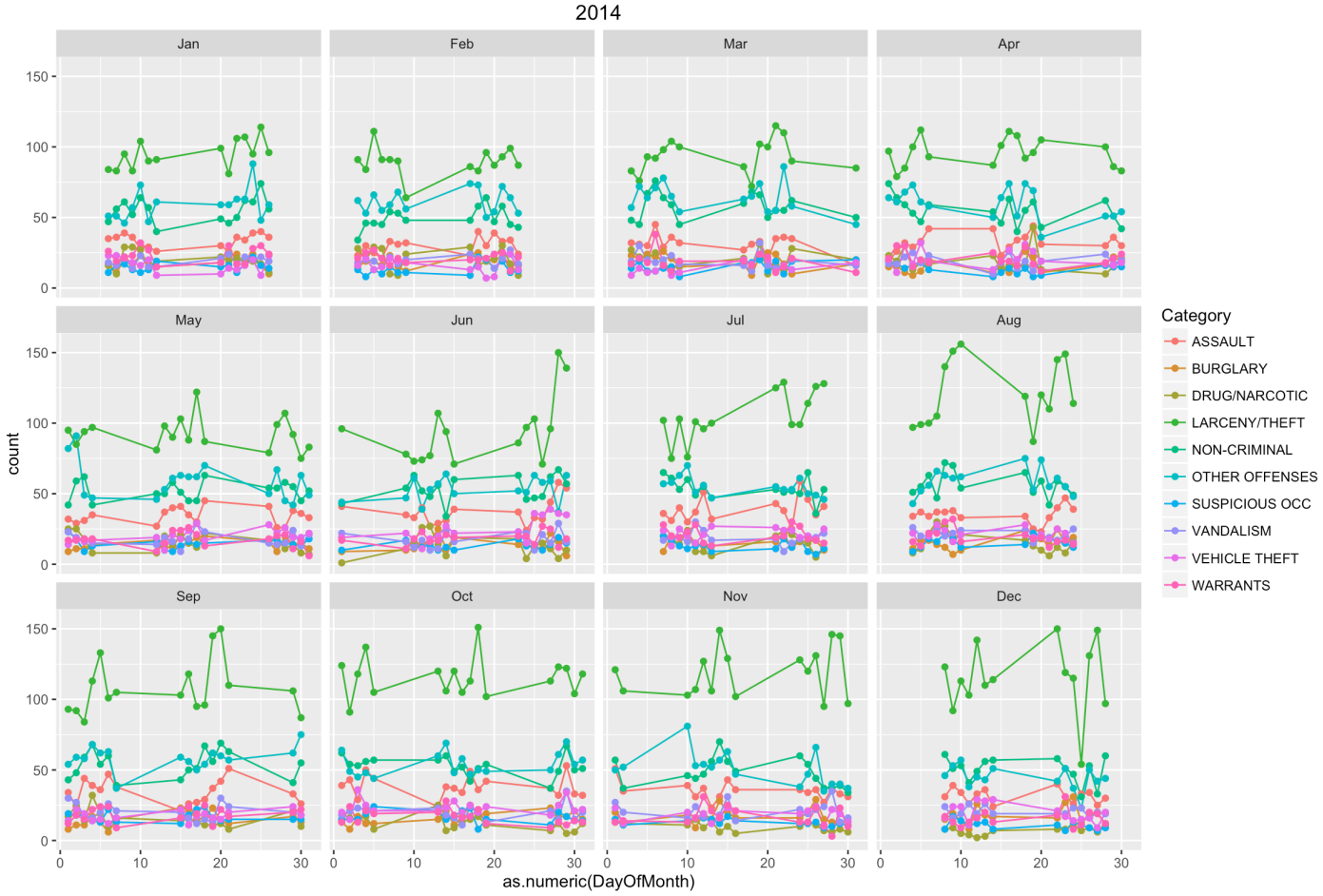
Crime vs day of the month faceted by month shows a strong seasonal pattern in crime incidents. The crime incidences peak about every 2 weeks. Infact, there are 26 times in a year that the crimes peaks.

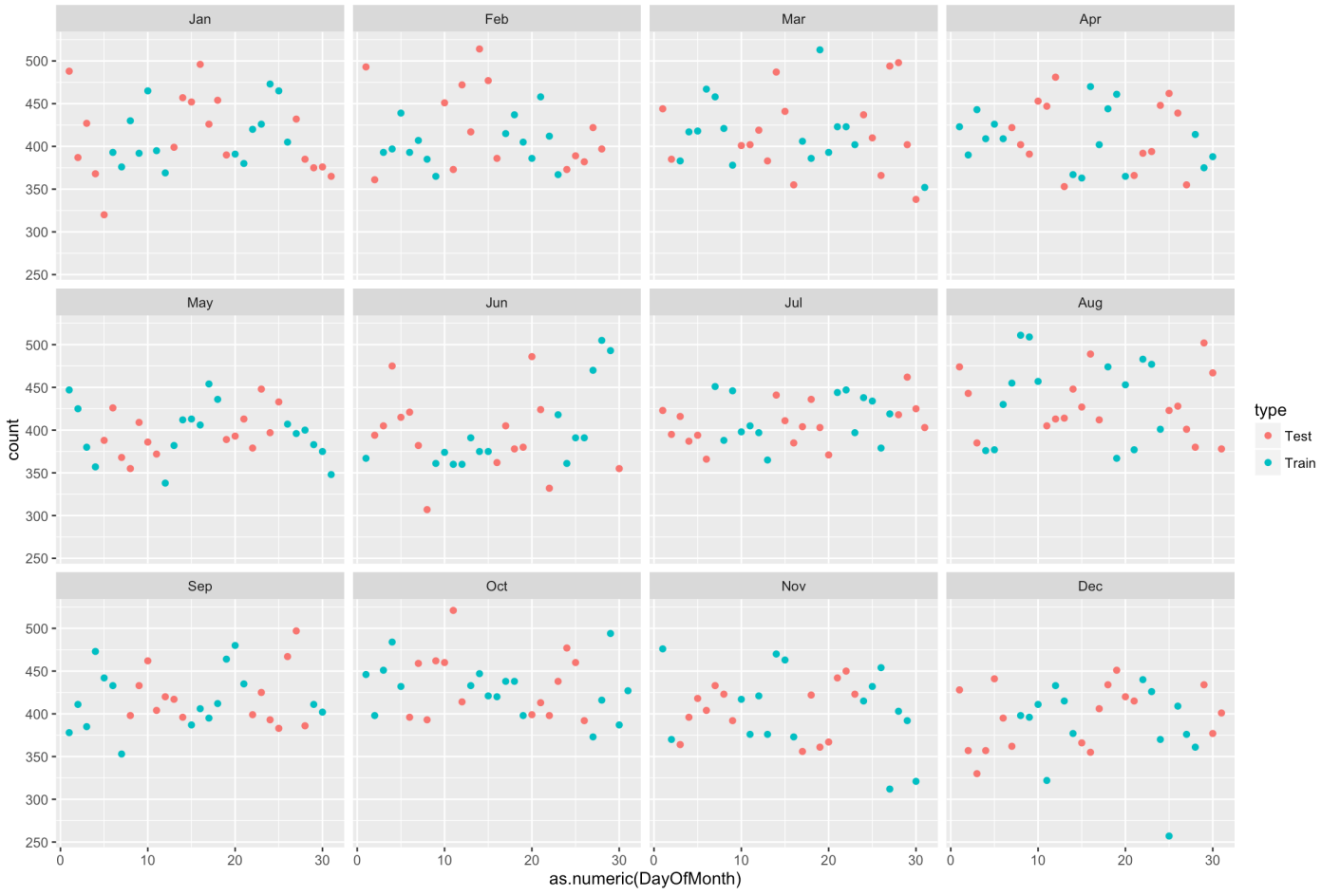


To further investigate crime rate's seasonal pattern. I normalized the crime by subtracting mean and dividng by standard deviation in each category. The normalized z-score shows strong seasonality in crime trends. Its surprising that this trend occurs in more than 800000 data points collected over past 12 years.



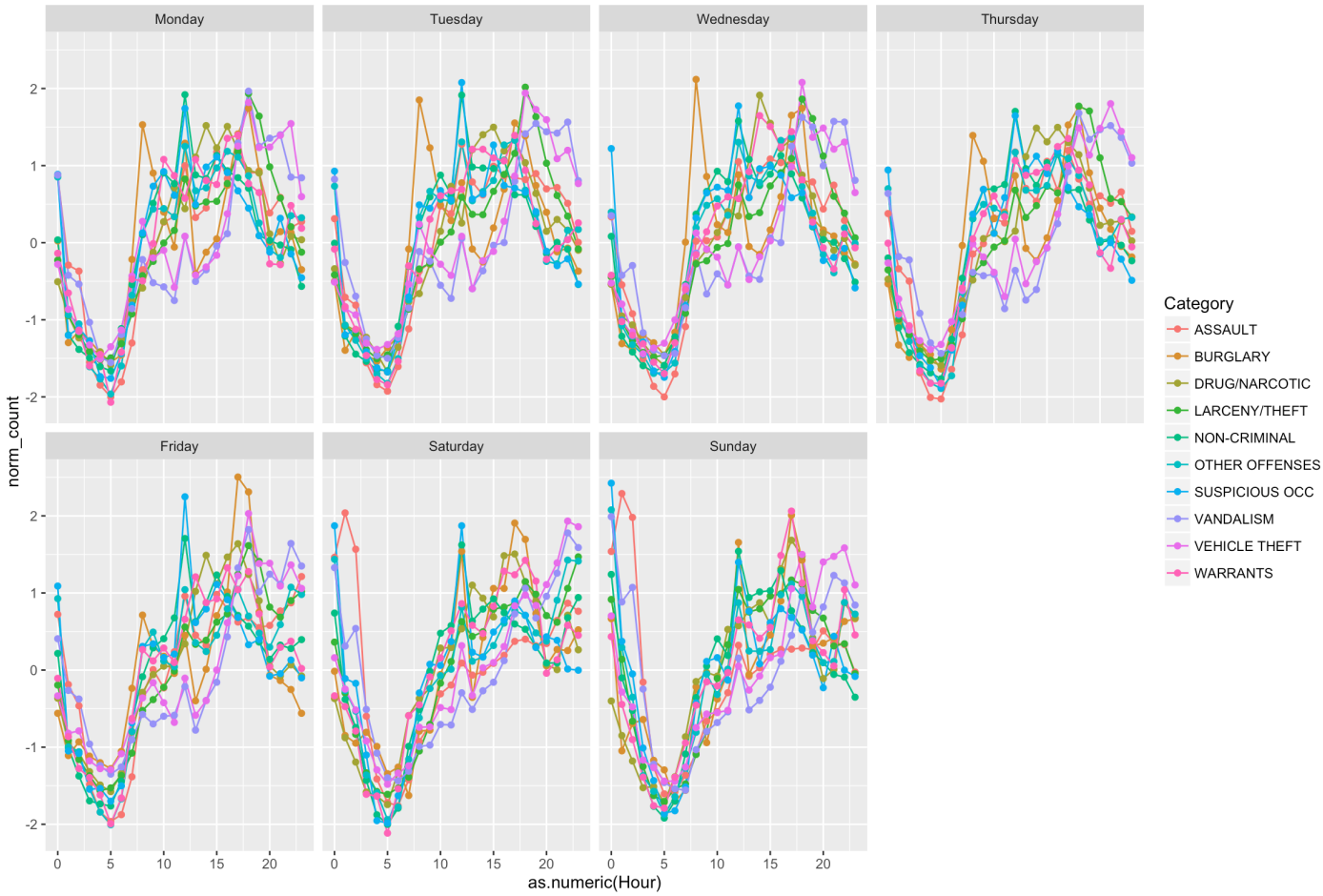
To investigate the seasonal pattern, I investigated crime variation in 2014 year. These patterns further confirm the previous finding that the observed 2 week periodicity is due to how test and train data were separated. Therefore, day of month will not be used for modeling.





Crime rate vs day of week and Hour of day

Crime vs hour of the day faceted by the day of week reveals a similar trend across all days of the week. The crime numbers drop around 5 pm and rise sharply and remain at high levels until midnight.

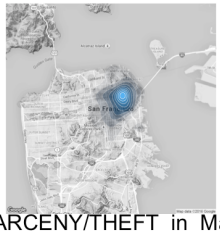


Crime vs Month and Location

Crime across San Francisco plotted for each month reveals that the incidents of crime varies with georgaphic location, and these incidents are different for different month. For a better comparision and brevity, I checked variations for one crime category only.

LARCENY/THEFT in Jan

LARCENY/THEFT in Feb



LARCENY/THEFT in Mar

LARCENY/THEFT in Apr

LARCENY/THEFT in May

LARCENY/THEFT in Jun

LARCENY/THEFT in Jul

LARCENY/THEFT in Aug

LARCENY/THEFT in Sep

LARCENY/THEFT in Oct

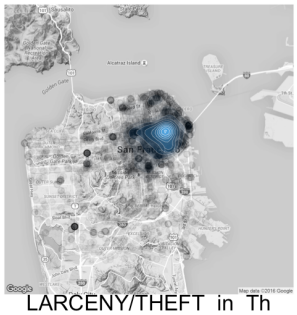
LARCENY/THEFT in Nov

LARCENY/THEFT in Dec

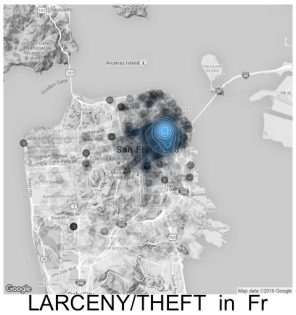
Crime vs Day of week and location

Crime across San Francisco plotted for each month reveals that the incidents of crime varies with georgaphic location, and these incidents are different for different month. For a better comparision and brevity, I checked variations for larsony only.

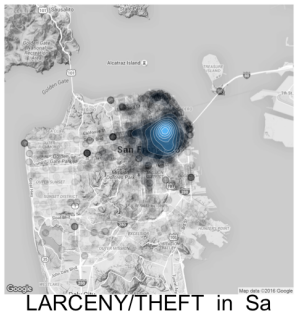
LARCENY/THEFT in Mo



LARCENY/THEFT in Tu



LARCENY/THEFT in We



LARCENY/THEFT in Th

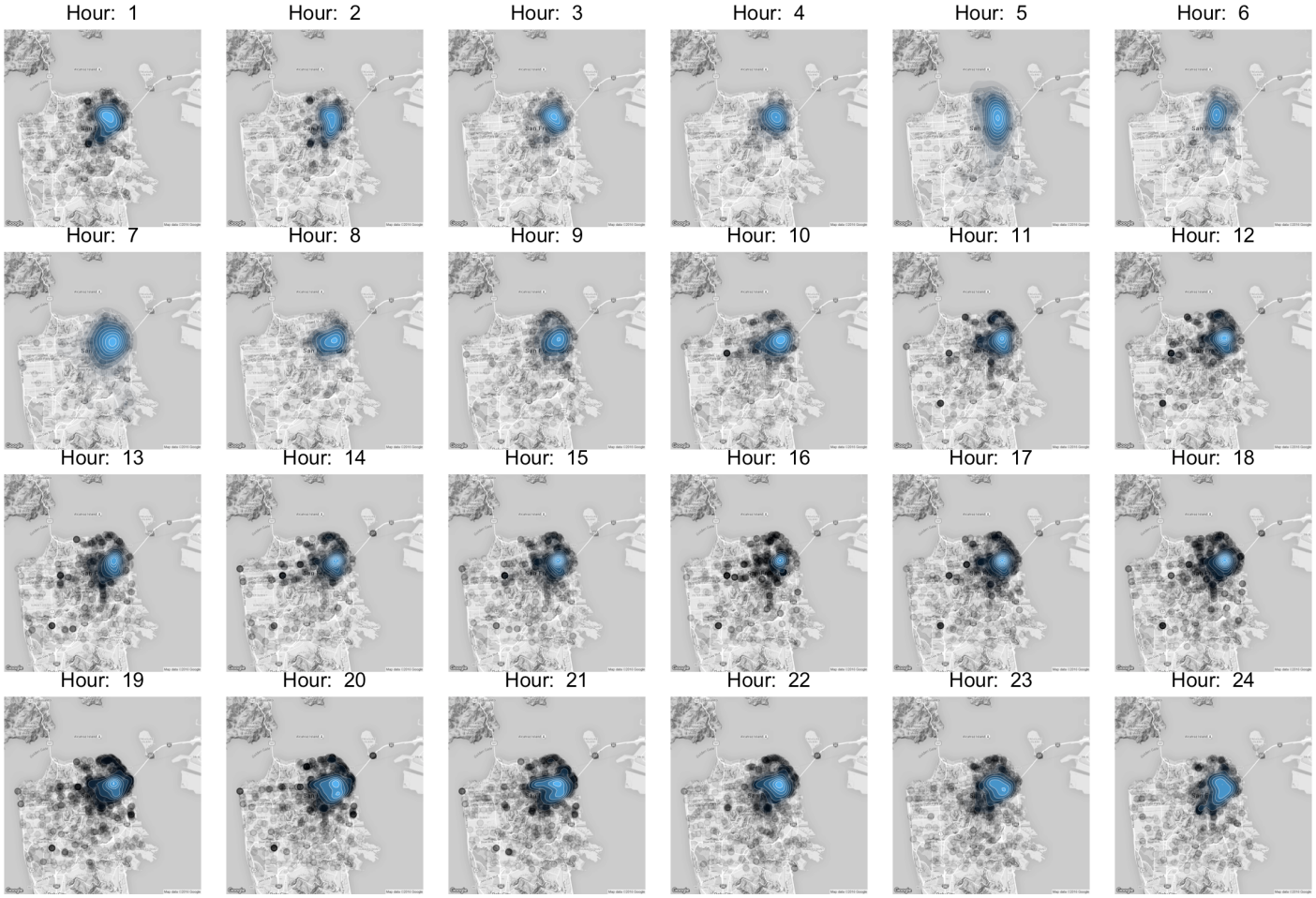
LARCENY/THEFT in Fr

LARCENY/THEFT in Sa

LARCENY/THEFT in Su

Crime vs Hour and location

Crime across San Francisco plotted for each month reveals that the incidents of crime varies with georgaphic location, and these incidents are different for different month. For a better comparision and brevity, I checked variations for larceny only.



Final model building

I wanted to predict the probability of a crime belonging to a given category, given the time and location of the crime.

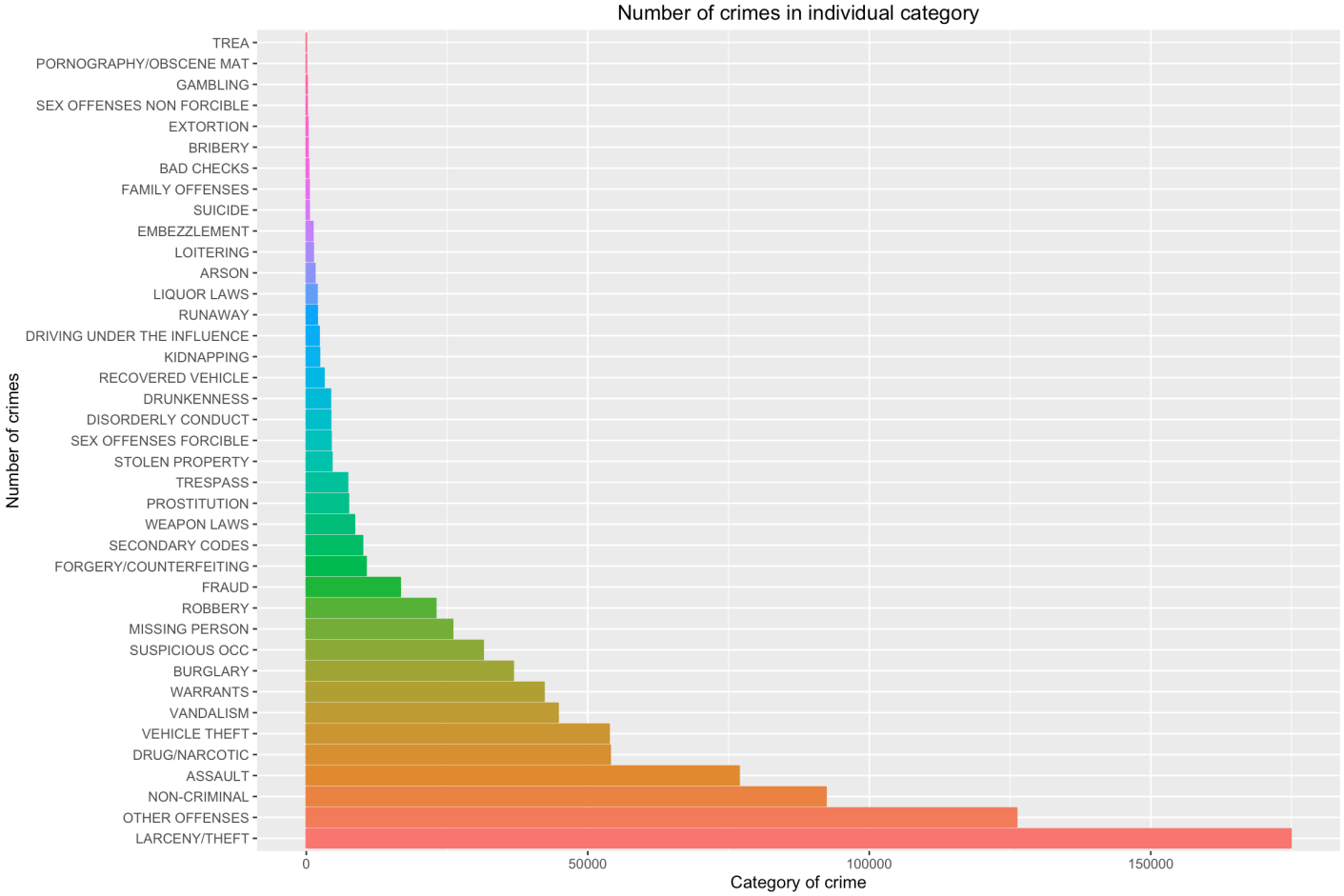
1- I divided the data into a 50% training and 50% validation sets. I then fit several models with hour, day of week, month, year, police district and all combined as independent variables. 2- I fit the data for each model on the training data set, and tested its accuracy on a validation set. 3- I used LibLinear package of R to perform multi-class classification. The final model had the following varaibles,

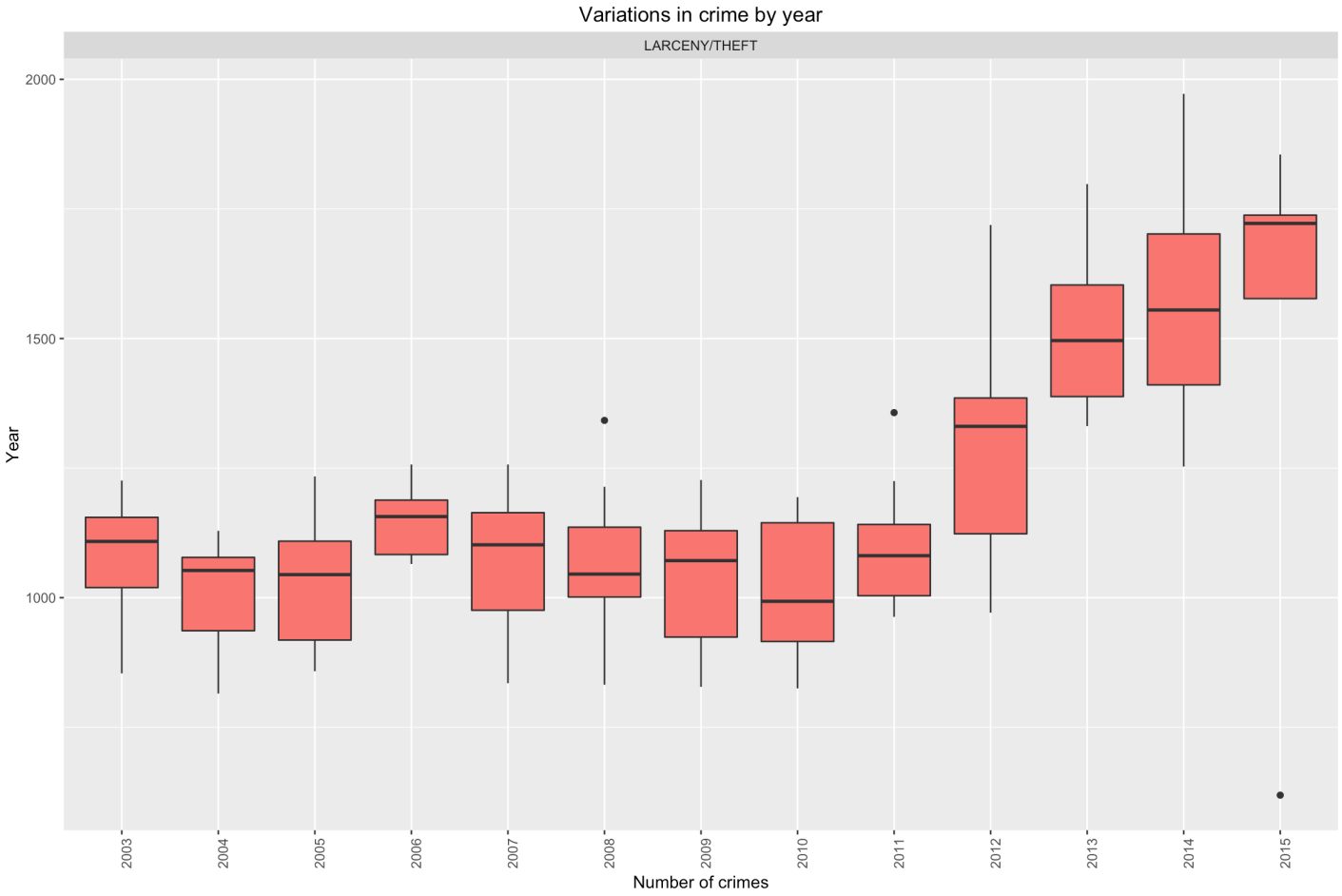
```
    - Hour of the day
    - Day of the week
    - Month
    - Police District
```

```
## [1] "Model:  h , Multi-logloss:  3.5925016832666"
## [1] "Model:  hz , Multi-logloss:  3.59934752088348"
## [1] "Model:  m , Multi-logloss:  3.6281685008858"
## [1] "Model:  Pd , Multi-logloss:  3.55743014810473"
## [1] "Model:  y , Multi-logloss:  3.60680095377129"
## [1] "Model:  dow , Multi-logloss:  3.62502962269724"
## [1] "Model:  all , Multi-logloss:  3.4977670267742"
```

Final Plots and Summary

Plot One


Number of crimes in individual category
TREA
PORNOGRAPHY/OBSCENE MAT
GAMBLING
SEX OFFENSES NON FORCIBLE
EXTORTION
BRIBERY
BAD CHECKS
FAMILY OFFENSES
SUICIDE
EMBEZZLEMENT
LOITERING
ARSON
LIQUOR LAWS
RUNAWAY
DRIVING UNDER THE INFLUENCE
KIDNAPPING
RECOVERED VEHICLE
DRUNKENNESS
DISORDERLY CONDUCT
SEX OFFENSES FORCIBLE
STOLEN PROPERTY
TRESPASS
PROSTITUTION
WEAPON LAWS
SECONDARY CODES
FORGERY/COUNTERFEITING
FRAUD
ROBBERY
MISSING PERSON
SUSPICIOUS OCC
BURGLARY
WARRANTS
VANDALISM
VEHICLE THEFT
DRUG/NARCOTIC
ASSAULT
NON-CRIMINAL
OTHER OFFENSES
LARCENY/THEFT
Number of crimes
0
50000
100000
150000
Category of crime

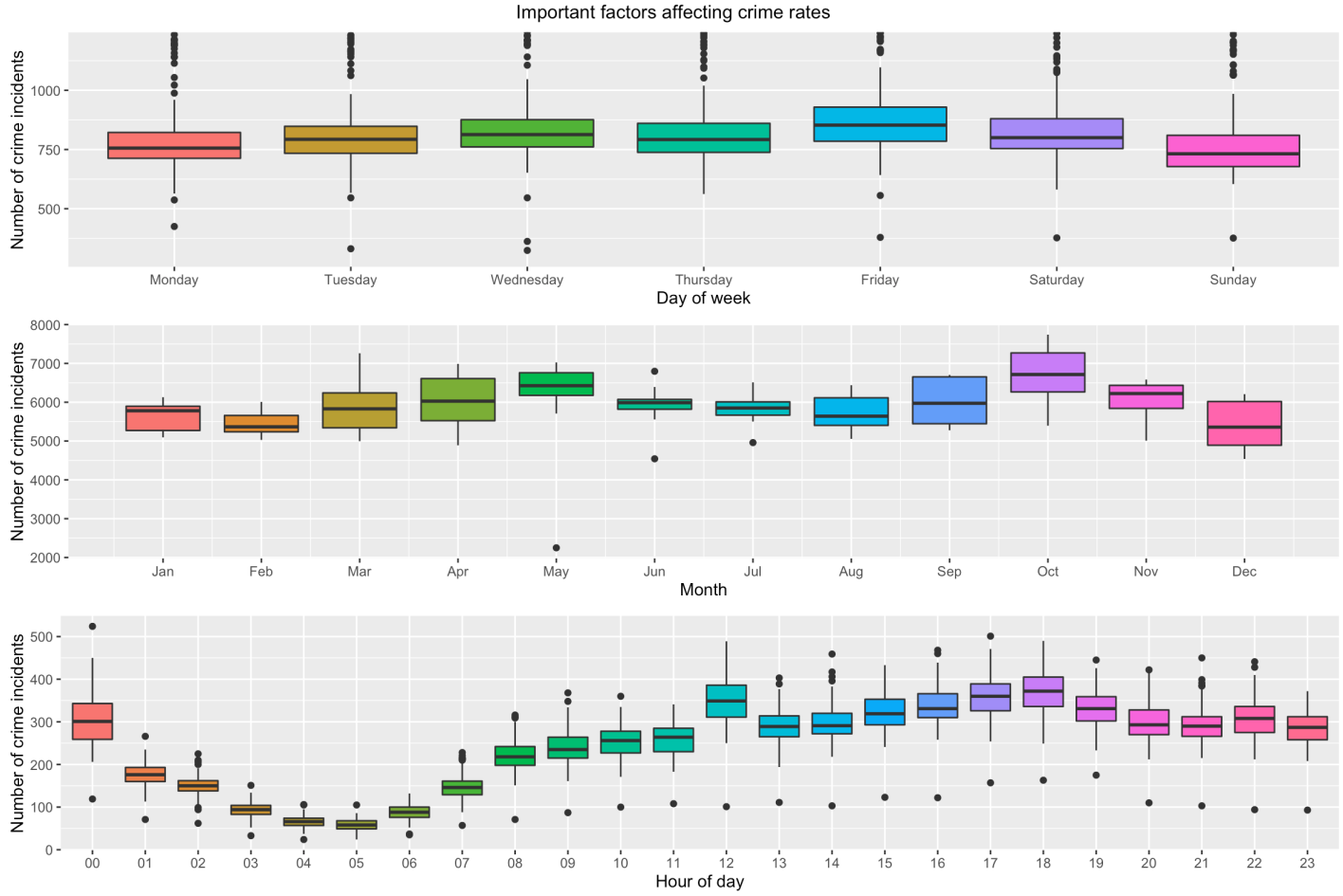

Variations in crime by year
LARCENY/THEFT
2000
1500
1000
Year
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Number of crimes

Description One

Figures above show distribution of crime and change in type of crime since 2003. From the first plot, Larceny/theft is the most common type of crime. Further, there appears to be a skewness in the type of crimes. For example, there were 174900 incidents of LARCENY/THEFT where as only 6 of TREA since 2003. Crimes belonged to the top 10 categories 83% of the time. And top 20 categories had 97% of the crimes. Therefore, a classifier that classifies crime in top 20 categories may be sufficient for most crime categories. For now, I used a model where I was predicting probability for all 39 classes, but in future I will use fewer predictors to see if I can increase accuracy of the model for them.

The second plot shows median total crimes per month from 2003 to 2015. Plot indicates that larceny/theft rates are on rise. Most interesting trend is reduction in number of vehicle thefts from 2006 to 2007.

Plot Two



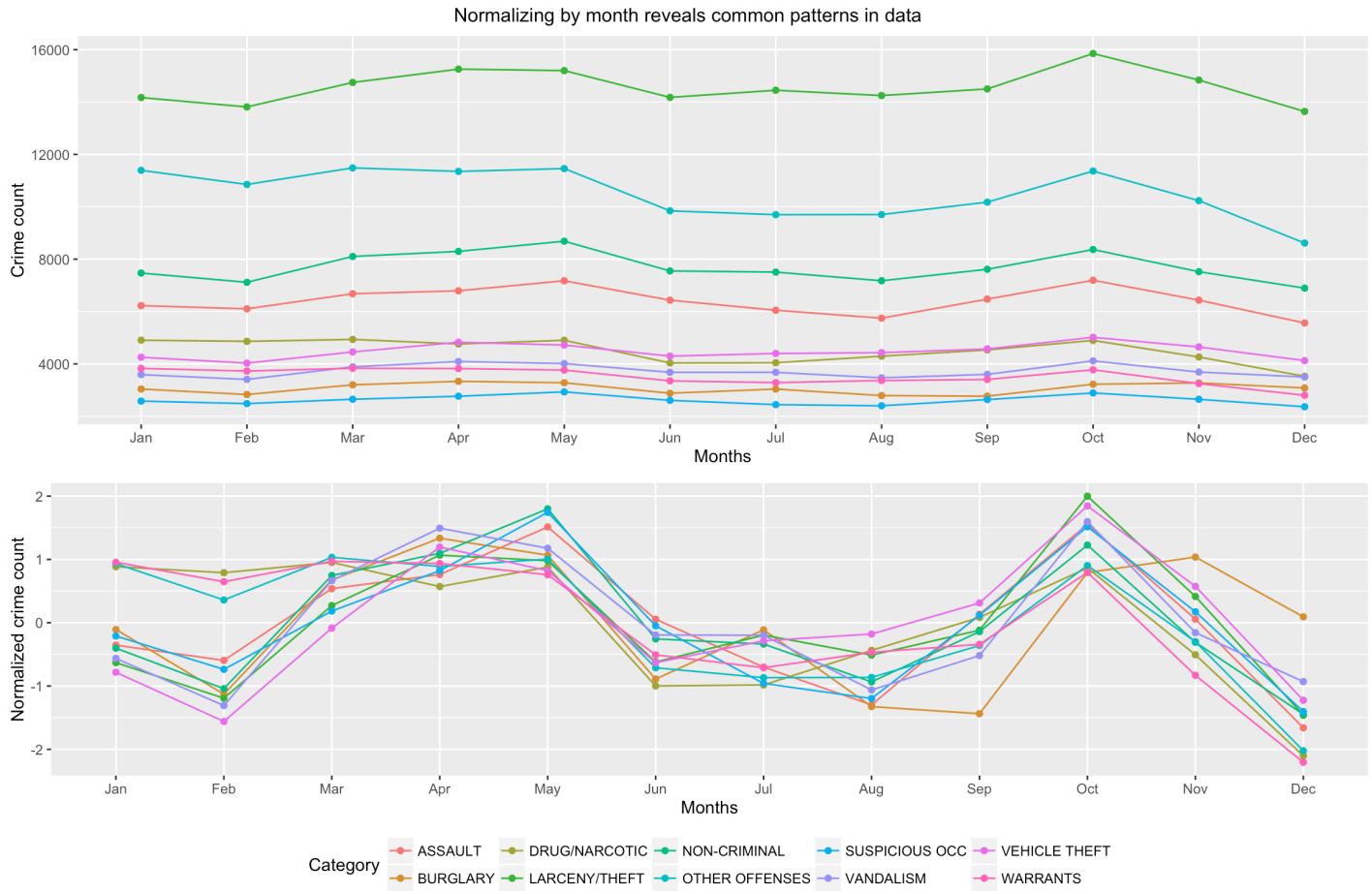
Description Two

Plots above show trends in crime for 3 main date factors,

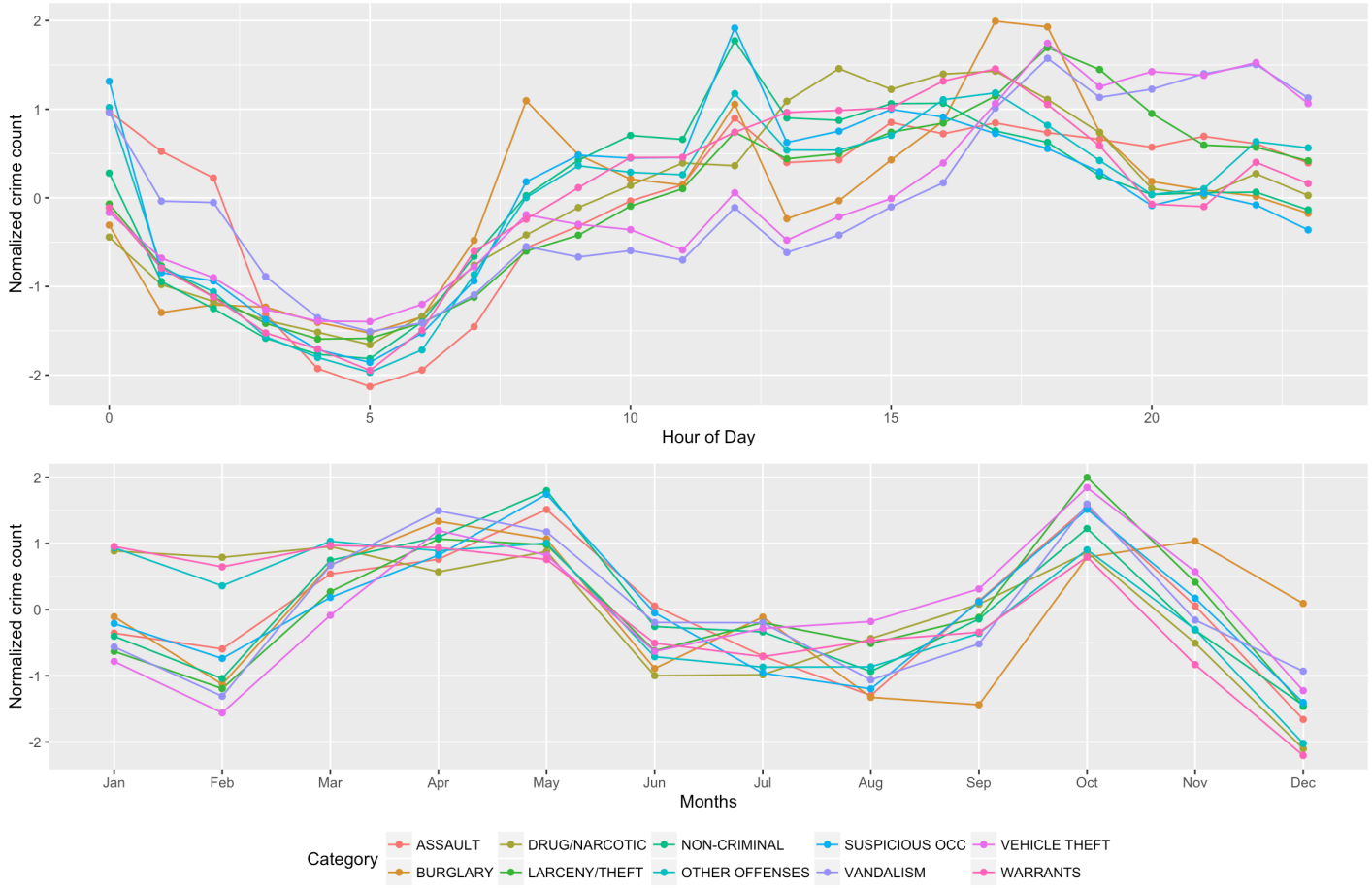
- Day of week: Crime rates change during the week. Crime incidents are higher on Friday, and lowest on Sundays
- Month: Crime rate is highest during october, and lowest in december. Crime seems to follow a bimodal pattern with peaks in May and October and valleys in December and August.
- Hour: Crime vs Hour of day shows a gruadual reduction in crime from midnight to 5 am, after which

it rises from 5 am to 10 am, and remains at sustained high level until midnight.

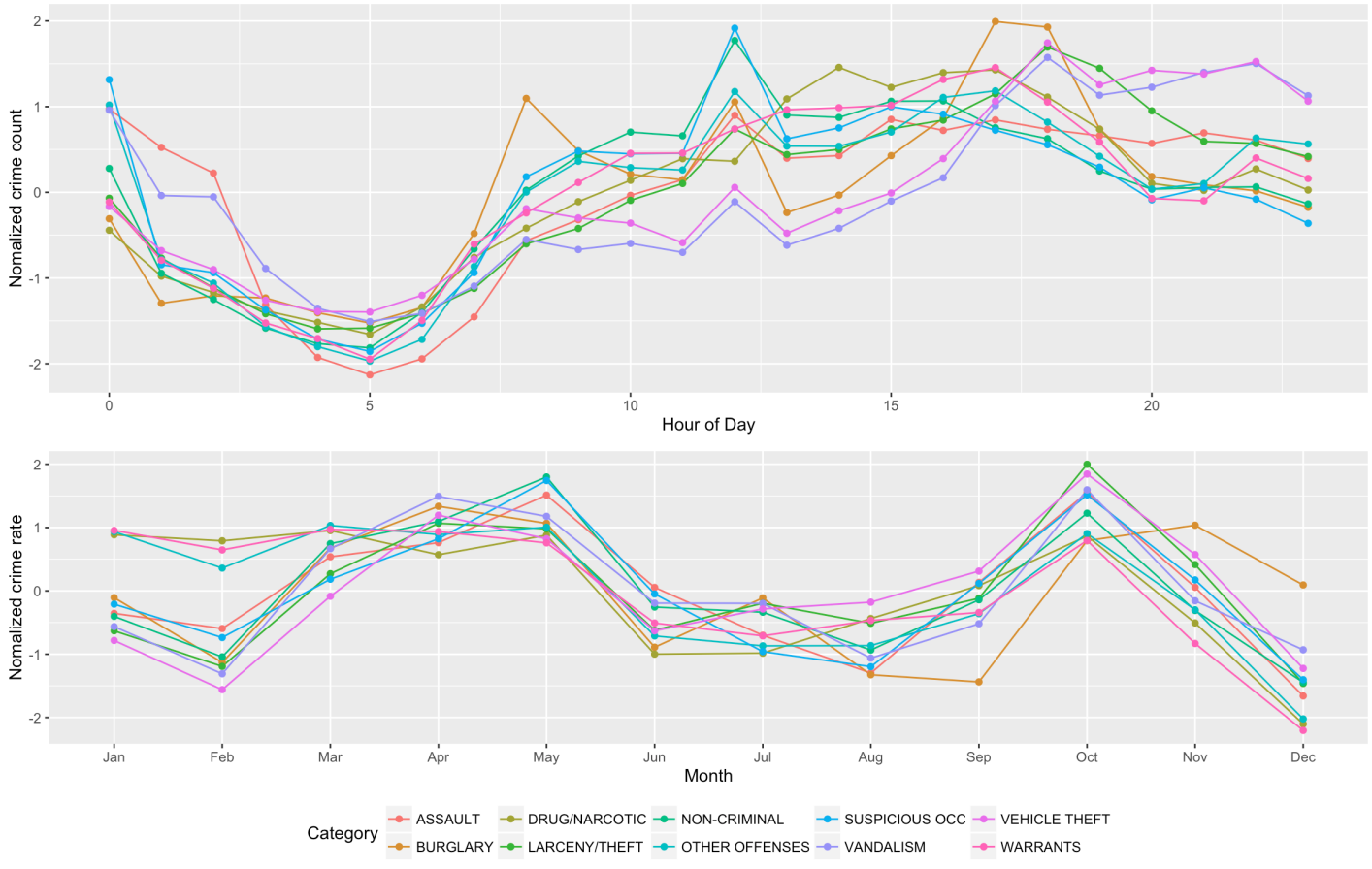
Plot Three



Therefore, normalizing can reveal patterns that are otherwise hidden.



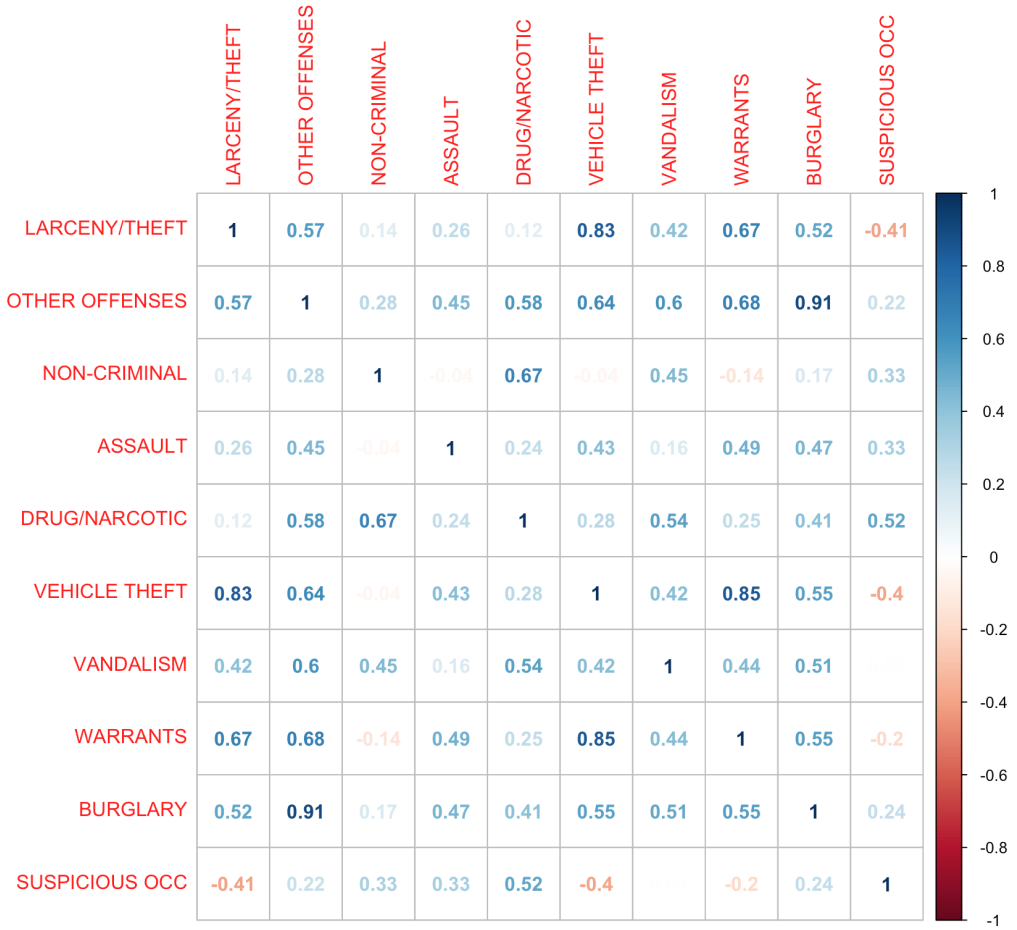
Normalized crime rate



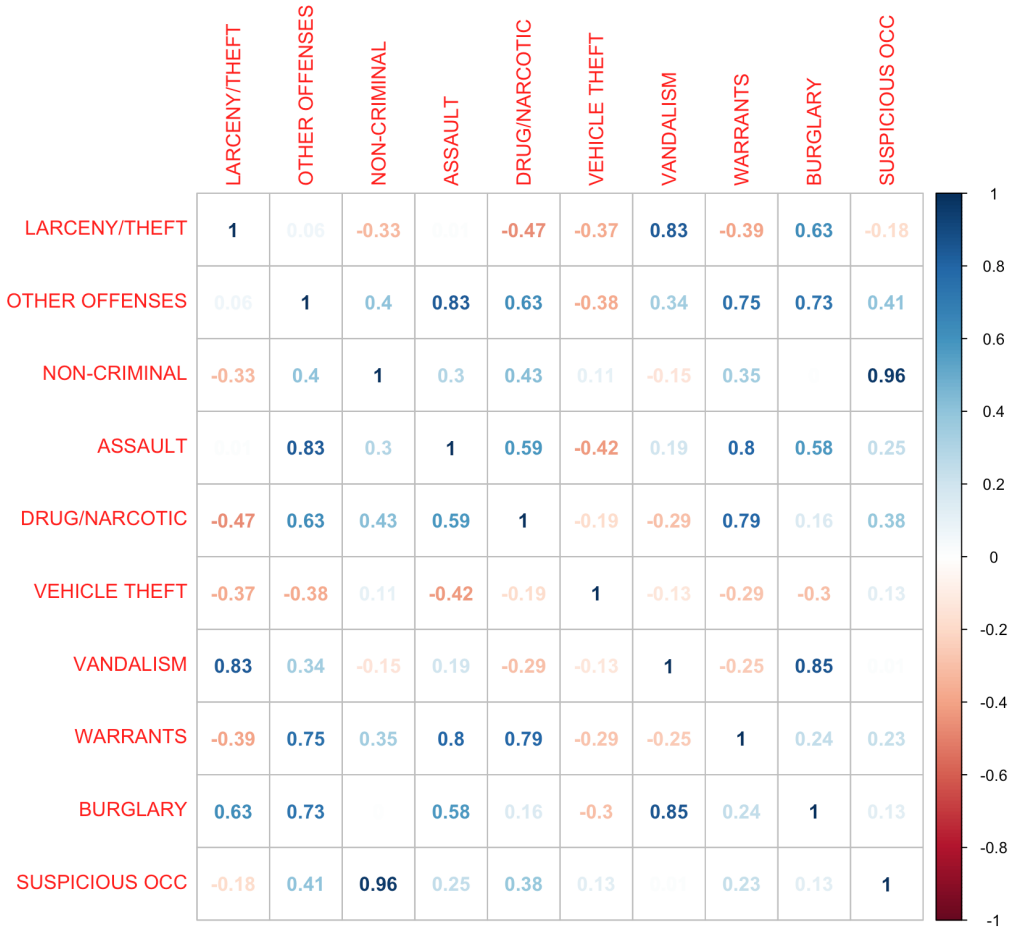
Description Three

Plots above show that although the number of crimes is different, the relative change with hour or month covary. This is especially clear in the first plot where I plotted number of crime in each category and normalized crime count. The next plot highlights this pattern for both the hour of the day and month. Correlation between crime counts per month and hour also highlights the correlation between variations in crime over month or hour of the day.

```
## [1] "Correlation by Month"
```




```
## [1] "Correlation by Hour of day"
```



Reflection

For this project, I downloaded San Francisco's crime data from Kaggle and performed exploratory analysis and built a linear regression model to predict the category of a given crime. During my exploratory analysis, I found that crime incidents although may appear random, when normalized by subtracting mean and dividing by standard deviation, follow similar trends across different crime categories. This trend was observed across crime categories for month, days of week and hours of day. To build the model, I used year, hour, day of week, month and police district as independent variables. I did model fitting using the following steps,

1- I divided the data into a 50% training and 50% validation sets. I then fit several models with hour, day of week, month, year, police district and all combined as independent variables. 2- I fit the data for each model on the training data set, and tested its accuracy on a validation set. 3- I used LibLinear package of R to perform multi-class classification. The table below gives result of fitting.

```
- Hour of the day, 3.59
- Day of the week, 3.63
- Month, 3.63
- Police District, 3.55
- Years, 3.60
- All, 3.5
```

From above, its clear that these variables are important for classifying. Further, year and location had the most affect on crime class, therefore I included an interaction term also. My final model had day of week, hour of the day, month, year, location and interaction between location and year. I then trained this model on full data. I got a log loss value of 2.56 which places my submission on 401 out of 1183.

This is a work in progress, and I will work on improving predictions as I learn more of machine learning. In the current model, I am not utilizing address or location information fully. One idea I want to test is to divide the crime into different groups where each group corresponds to a combination of hour, week day and month, and then obtain a 2-D kernel density estimate of each crime. From these density values, I can get the probability of a crime occuring at a given location given the crime category. I can use this to perform naive bayes calcuation to compute probability of a crime belonging to a category, given time and location of the crime. I also did not include interaction terms to account for the fact that the crime rates may be different across different combination of factors. For example, crime rate may be different during individual days of the week in different month. The simple model that I made does not account for these variations.