**Crime Data Analysis**

**A Major Project Report**

*Submitted in Partial Fulfillment of the  
 Requirements   
for the Degree of*

**MASTER OF SCIENCE**

**IN**

**ENGINEERING MANAGEMENT**

By

Usha Khetan - <khetan.u@husky.neu>

Jonathan Ajit Solomon - [solomon.jo@husky.neu.edu](mailto:solomon.jo@husky.neu.edu)

Under the Guidance of

Prof. Kathleen Durant  
  


Data Analytics DA 5020  
334 Snell Engineering Center  
360 Huntington Avenue  
Boston, MA 02115

CONTENTS

Overview ………………………………………………………………………………………………………………………………………….. 2

**Project Objective ………………………………………………………………………………………………………………………………. 3**

**Packages …………………………………………………………………………………………………………………………………………… 3**

**Phase 1 Scraping (Collecting of Data)** …………………………………………………………………………………………………….4

**Phase 2 Storing**……………………………………………………………………………………………………………………………………….7

**Phase 3 Retrieving and Plotting** ……………………………………………………………………………………………………………. 9

**Insights** ………………………………………………………………………………………………………………………………………………. 14

**[Bibliography](#_Toc469667919)** [……………….………………………………………………………………………………………………………………………..15](#_Toc469667919)

**Overview**

*What do we do when we want to move into a new neighborhood in the United States?* Selecting the right neighborhood is essential for kids to go to school in a safe place, for business to thrive in a safe environment and that we all as a society feel secure however late in the night it might be in the locality.

Crime data plays an essential role in the business, social, intellectual and artistic life of a thriving society. Los Angeles being the gang capital has around 700 crime gangs with 68 only in Compton. Crime rate however is high in LA compared to many other states as per statistics from 1980 to 2010. LA being divided into 21 divisions have police stationed here, but are unable to curb crime rate completely. Efforts however have been made to certain extent to shunt these gangs and individuals from extending in to more territories and creating havoc among society.

Today, in the advent of the internet, information is quickly spread around making police and society aware of the happenings of the surroundings. This fact, coupled with the rapid increase in crime rate, has made crime data analysis one of the most powerful factors influencing an individual or business decisions.

Los Angeles County Department of police is an American government body. It consists of individuals adroit in curbing crime prevailing across LA. My project focuses on the Exploratory Crime Data Analysis for LA county and to provide near perfect analysis on parameters such as:

* Types of crimes committed and their description
* Locations of the crime scene

1. City
2. State
3. Street
4. Latitude
5. Longitude

* Data and time of the crime?
* Which week of the month has the highest crime rate?
* Are crimes higher during weekends or weekdays?

Further we would like to analyze these contributing factors and draw useful conclusions about these factors which might help the LA crime department even more in tackling criminals.

**Project Objective**

For the above study, we are using data provided by the open source LA county department government website and creating analysis and scatter plots using GGplot, ggmap packages in the Los Angeles area as they contribute to almost 60% of the Countries crime. This data has been employed to correlate the demography of the crime with the neighborhood and the types of the crime being committed. Another correlation links time at which crime occurs to the address of the crime scene. Moreover, based on this data, a relational scale can be introduced in the website which prevents individuals from starting business ventures or schools in these localities.

**Packages**

**R studio Packages used:**

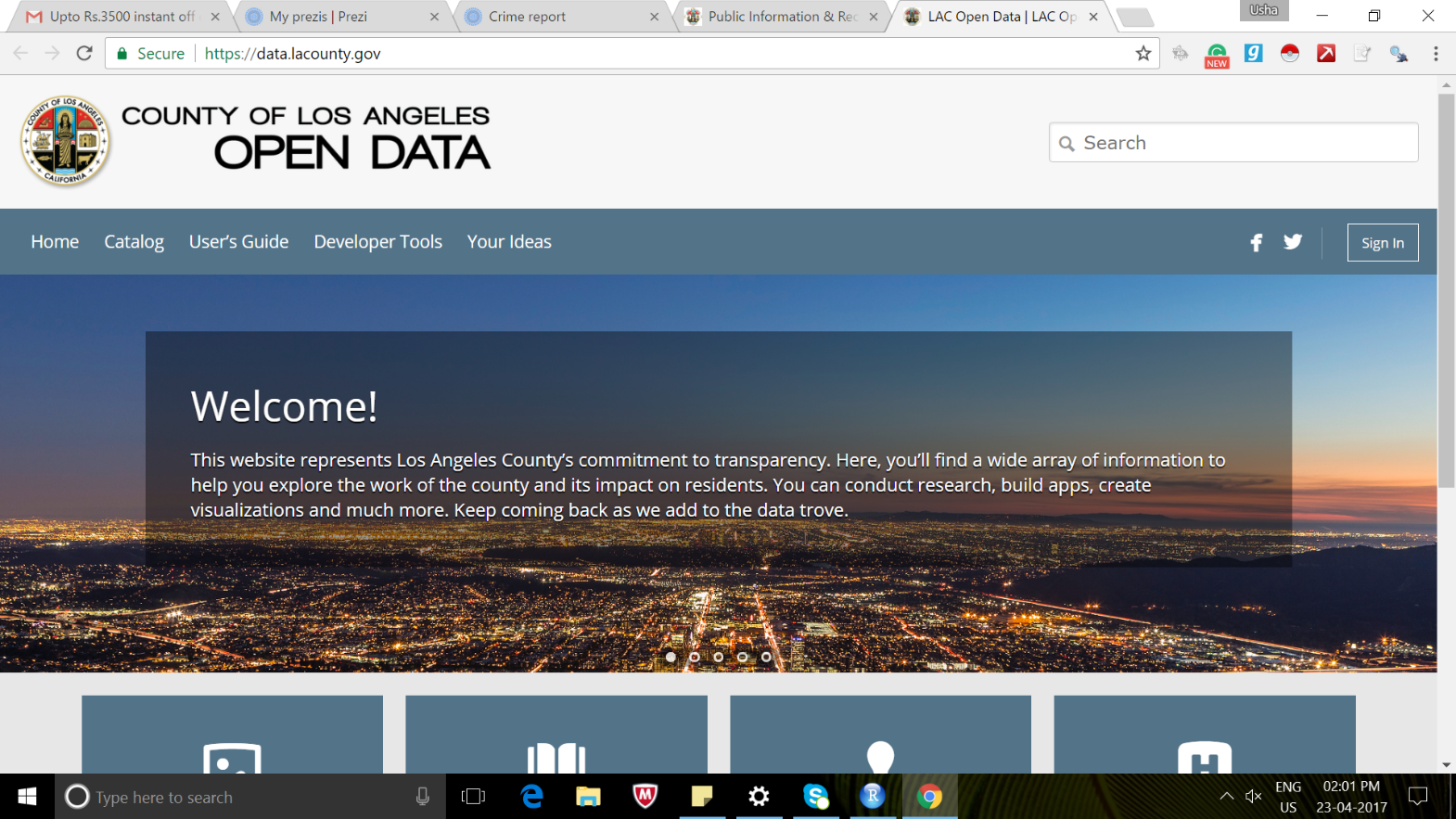
*The packages that we have incorporated in this project includes the following:*

1. **'httr'**
2. **'dplyr'**
3. **'stringr'**
4. **'jsonlite'**
5. **'lubridate'**
6. **'RSQLite'**
7. **'ggplot2'**
8. **'ggmap'**
9. **'plotrix'**

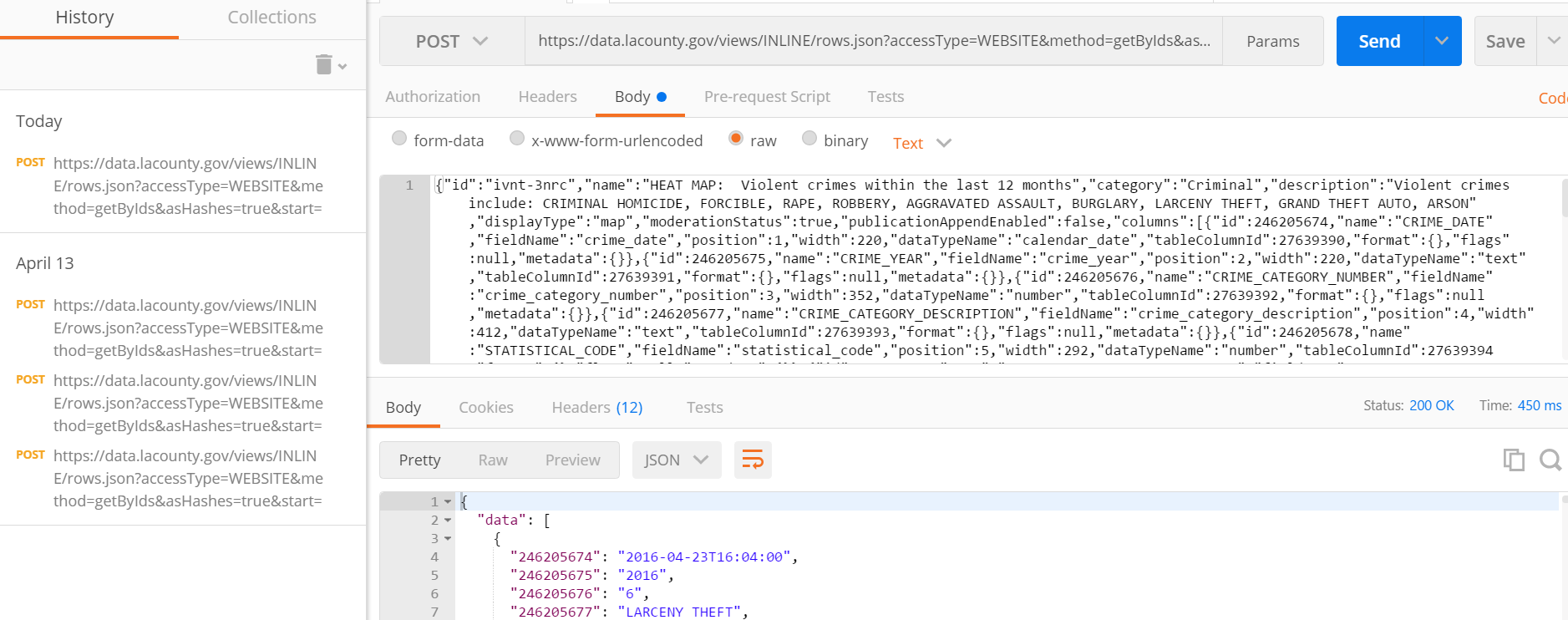
**Phase 1**

**Scraping**

**Website**: https://data.lacounty.gov/browse?category=Public+Safety&utf8=%E2%9C%93



We have used LA count open Data source for extracting the data. Since it is an official website we checked the data scraping using the Chrome POSTMAN extension to check the data coming from the LA count server.



Here, we have used HTTR package to handle the http post request as data is loading dynamically using the server request. Since it is dynamic data being loaded, traditional R packages like RVest and RCurl cannot be used.

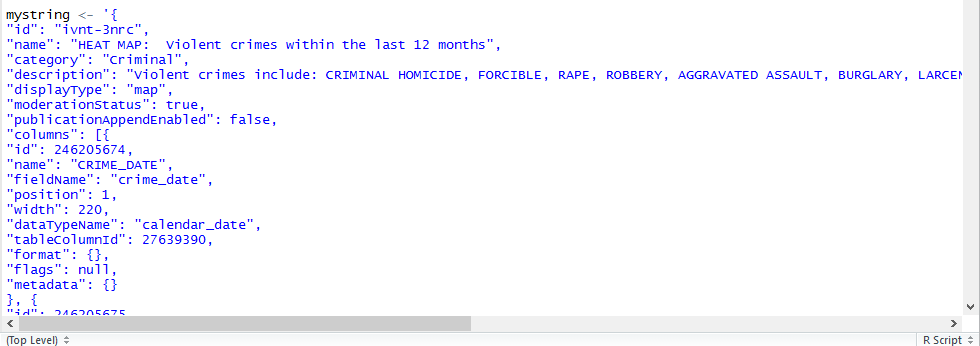
We have used HTTR package to send the request to the LA county server. Our extracted data contains 50,000 records approximately.

**R code :**

***Step 1:*** Get the url of the LA county site



***Step 2:*** Prepare the query string for post request to the server. Query string contains the metadata and templating requirement for dynamic loading the data

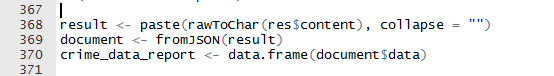
****

***STEP 3:*** POST request to the server.

****

***STEP 4:*** Handling the JSON response and storing it in data frame

Since the data received from the post request is in **JSON** format**,** we used **“jsonlite”**  R package to handle the json data



**Cleaning Data**

On analyzing the dataset received from extraction, we figure out that it contains data from different cities and countries in US. So, we narrowed our cleaning for Los Angeles, California State. Columns containing Date and Time also were cleaned as well as handling of NA values.

**Phase 2**

**Storing- Database**

At first it seemed like a NoSQL databases capable of handling JSON objects would be the appropriate database for this. However, the requirement of structuring the data for analysis, the availability of libraries to structure data and a necessity to have a relational model led to choosing a relational DB. Also, notwithstanding the retrieval functions and schema definition requirements, SQLite was chosen.

*The schema has two tables:*

**CRIME INCIDENT**

**PK: Crime\_ID**

**Street**

**City**

**Date**

**Time**

**FK: CategoryID**

**PK: CategoryID**

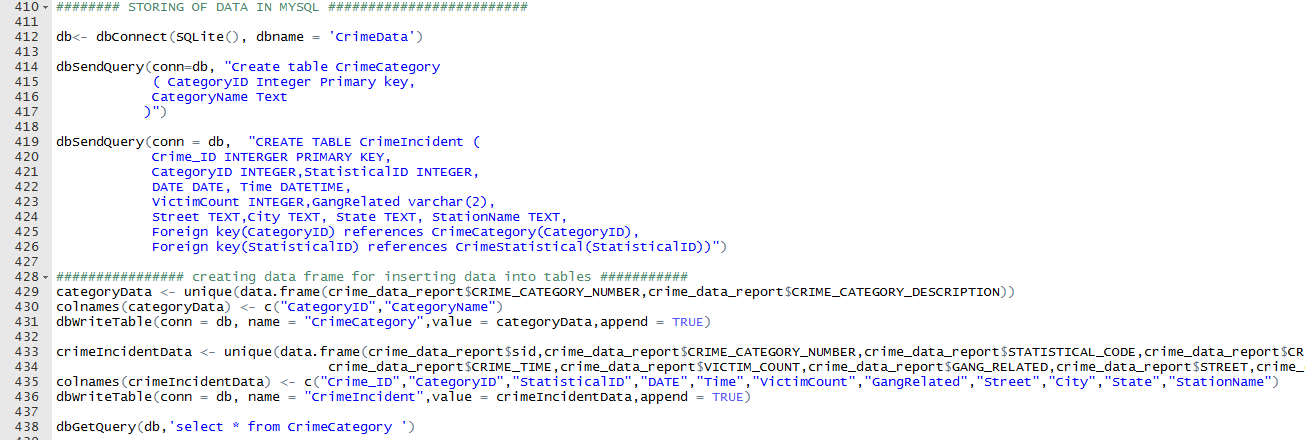
**CategoryName**

**CRIME CATEGORY**

**STEP 1:** Create DB Connection to connect to the database

**STEP 2:** Defining the structure of the data to be stored in database

**STEP 3:** Sub-setting the data frame and inserting data into the tables

****

**Phase 3**

**Retrieval and Analysis**

Now having stored the data making use of RSQLite package, phase 3 consists of retrieving and making valid analysis.

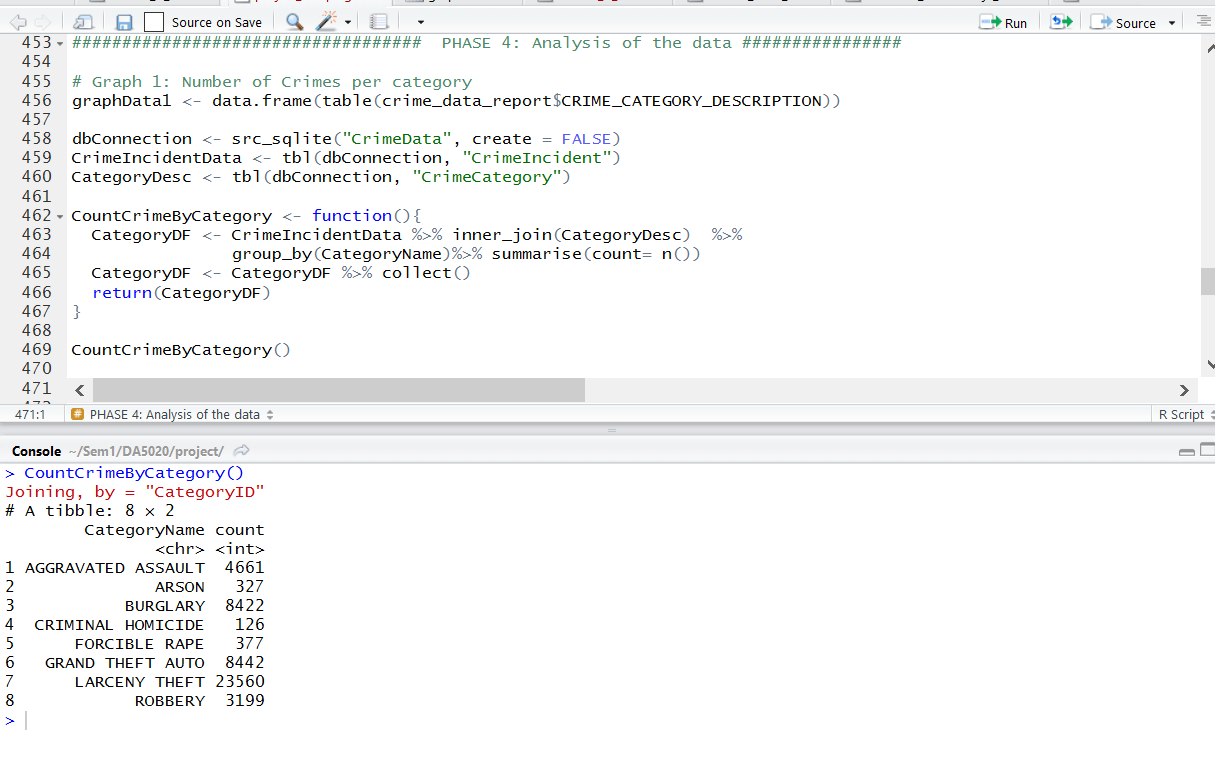
*Our analysis consists of the following:*

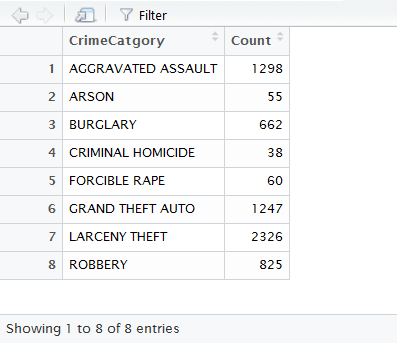
* **To find out the number of crimes per category?**
* **What kind of crime committed is highest during which month?**
* **Percentage of crimes committed during the days of a week?**
* **What is the count of crimes committed during the different interval of a day?**

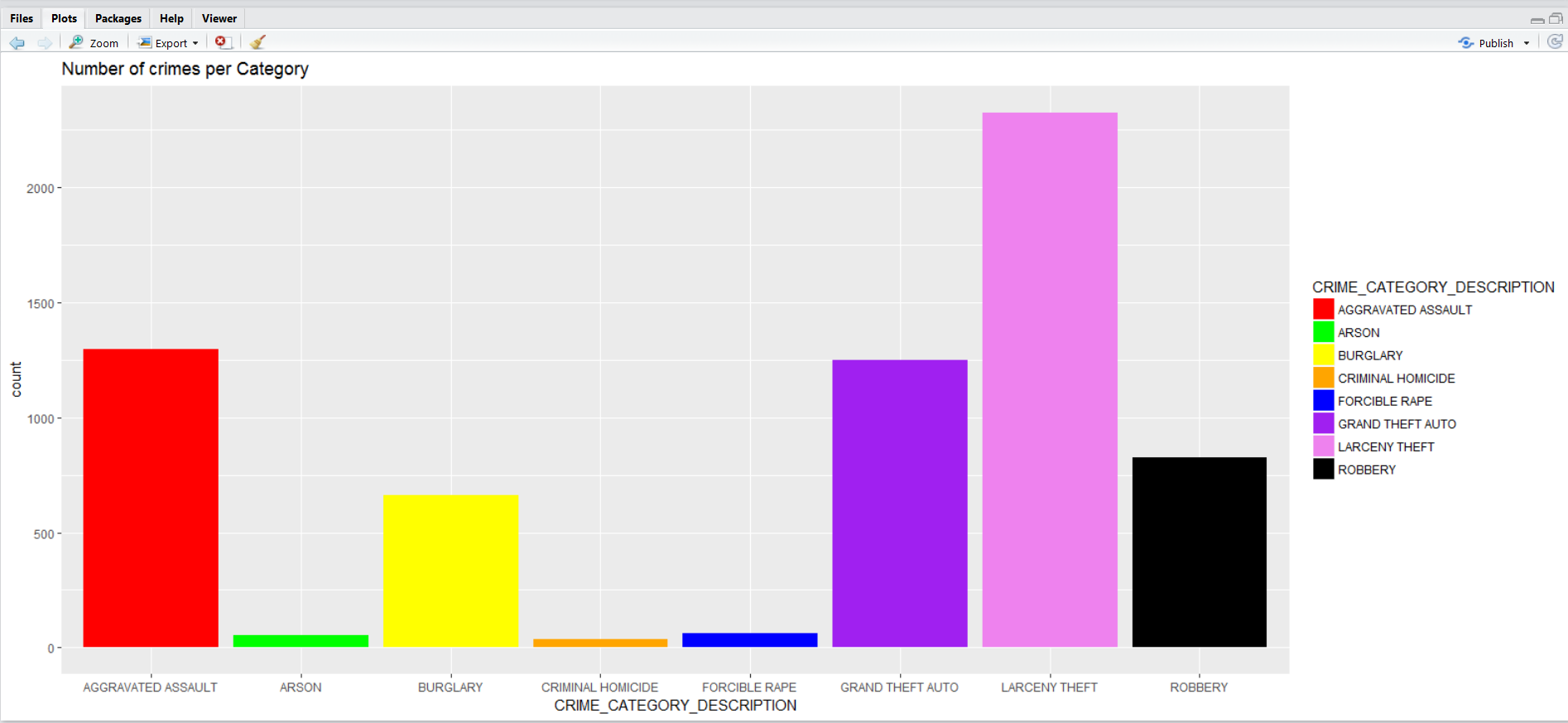
To illustrate these above questions, packages like GGplot2, ggmap and plotrix we used by us.

**To find out the number of crimes per category?**

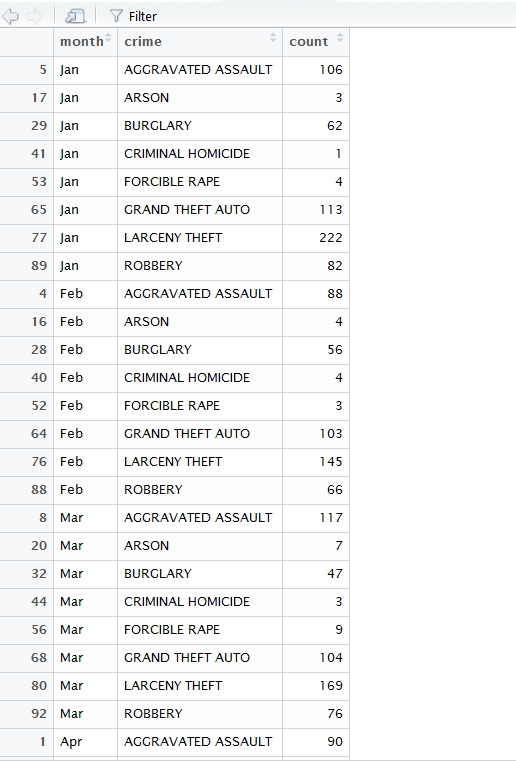
**Retrieval R Code querying RSQLite** (Have included only for one of them in documentation)

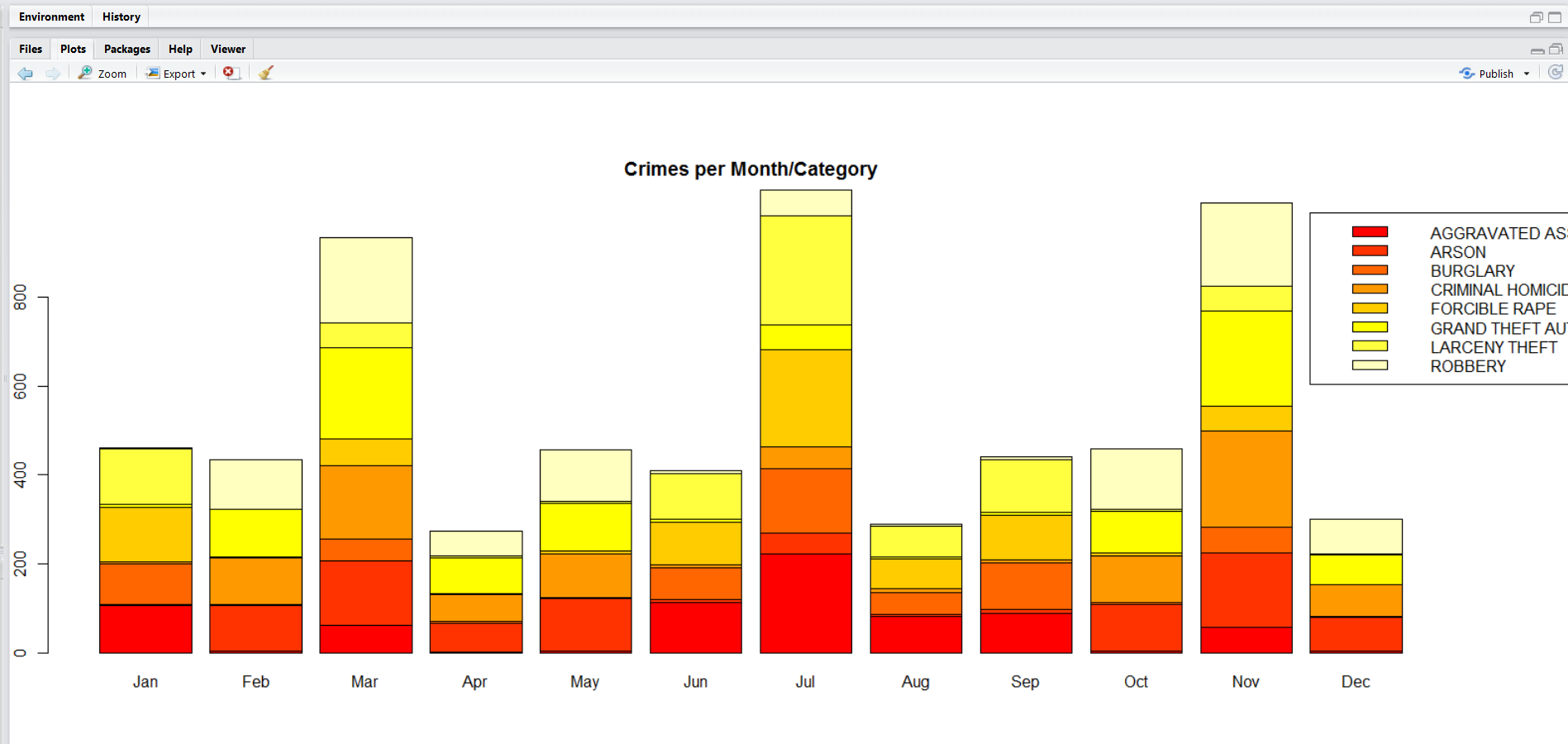
****

****

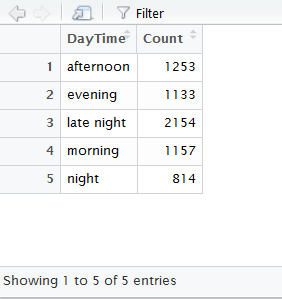


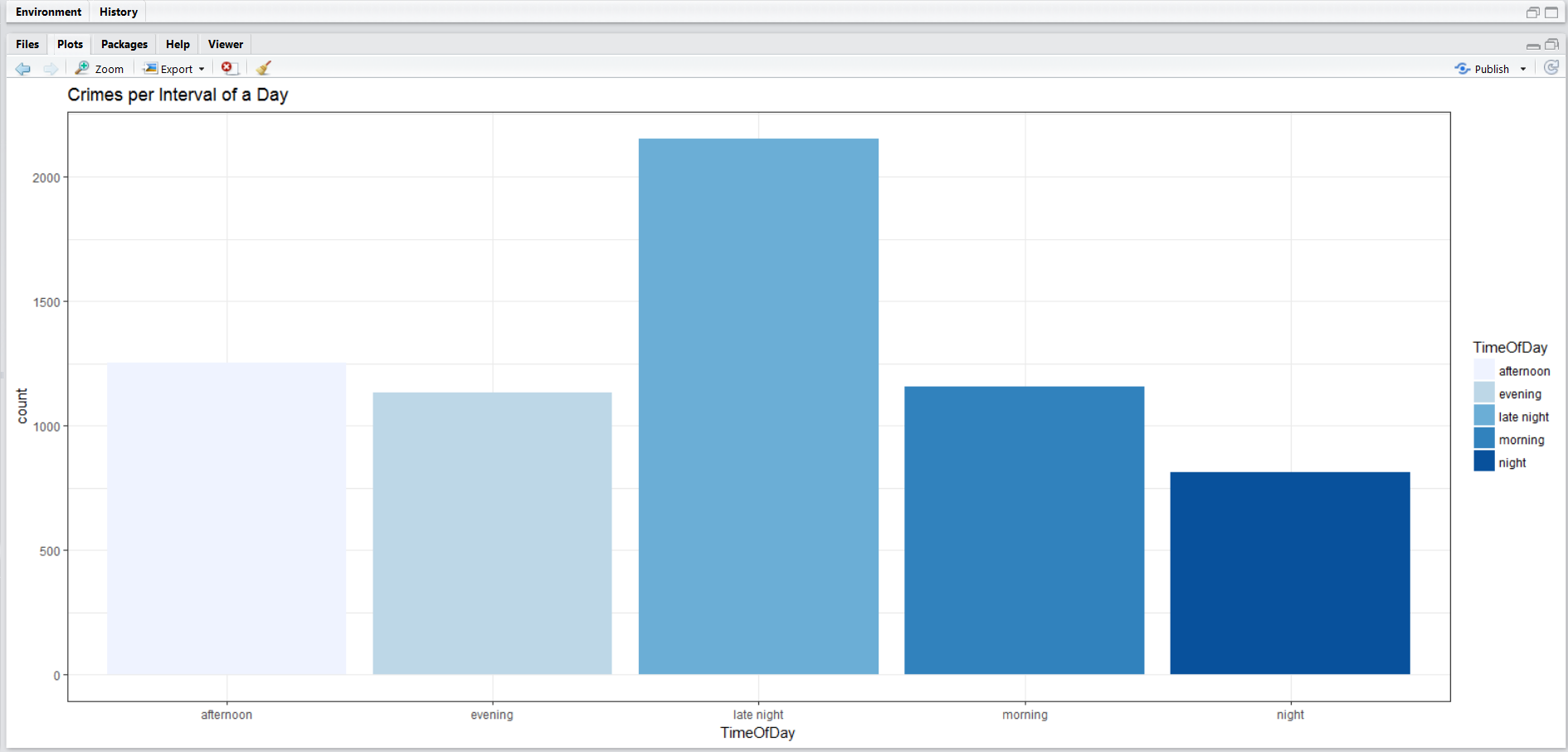
**What kind of crime committed is highest during which month?**

****

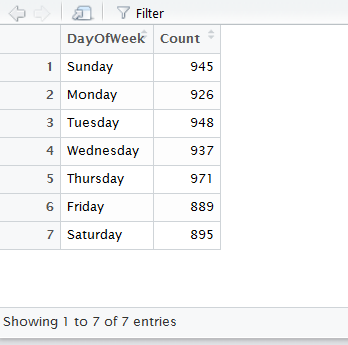
****

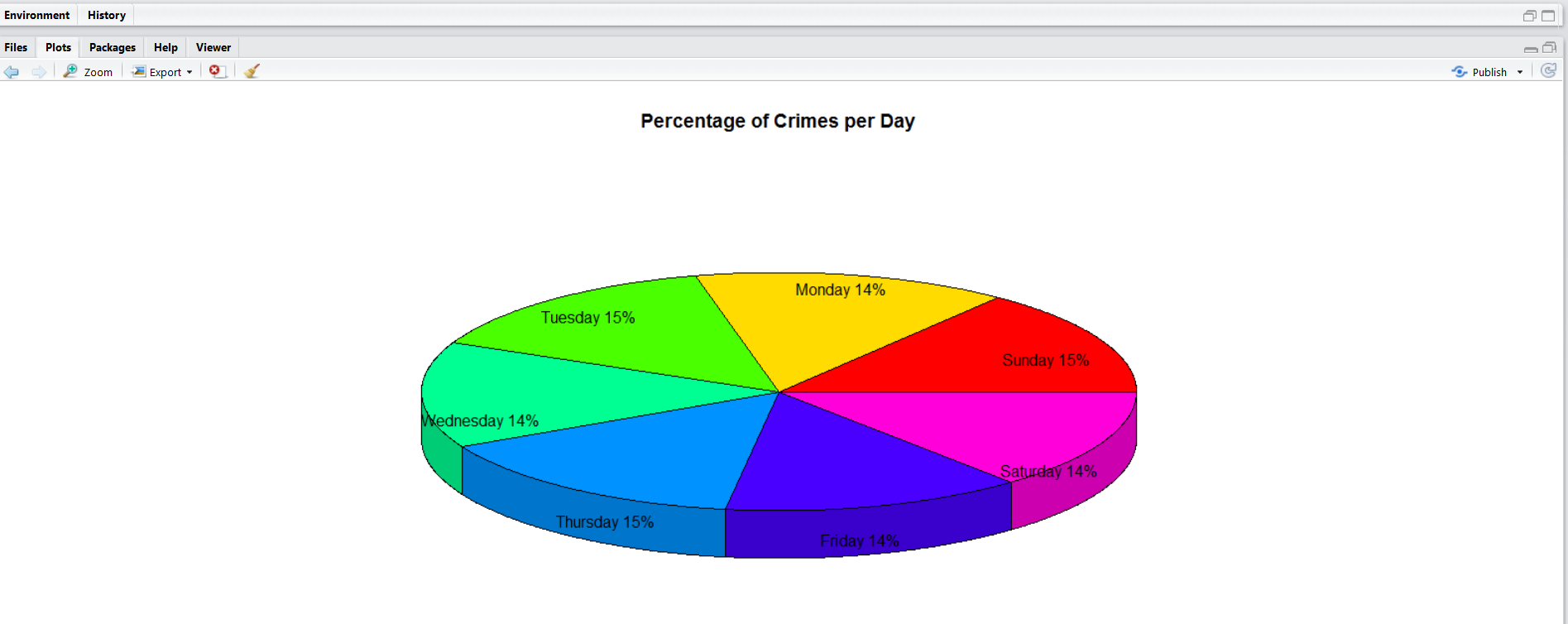
**What is the count of crimes committed during the different interval of a day?**

****

****

**Percentage of crimes committed during the days of a week?**

****

****

**Insights**

1. **Learnings:** ​ While the project is a lot more streamlined now, it ran into some very interesting problems. Data is an important key in any Business decision process, therefore selection of an appropriate information repository is important for data analysis to get into some interesting hidden facts inside the data. While working on this project, we came across various hurdles in scraping the data. We learnt that in today’s web 2.0 world, where consumer is also the data producer in addition to the security reasons of data, many websites have encoded their server request. Traditional scraping mechanism like RVest and RCurl won’t be able to scrape data from such websites. Therefore, we explored various new scraping mechanism to extract the data from a dynamic webpage like RSelenium, HTTR.

We incorporated various cleaning techniques that we practiced in our class. There was huge dilemma in selecting a proper storage structure for the data. Since we have various relations between different crime categories and crime committed, we came up with RSQLite database and retrieved data on various analysis parameters.

2. **Future Work:**

We targeted at one city i.e Los Angeles, California, US. However, we can further expand on venturing into other cities to extract crime data and provide valuable analysis to help curb felons.

Moreover, including temperature would be one factor that would help in analysis of crime reports for the crime department and individuals/society to start business. We plan to incorporate this in future. Ethnicity also would be something we would want to include in future, to analyze which demographic location contains a prevalent crime committing ethnic race. For this perhaps the data should be more readily available to fetch this data.

**Bibliography**

<https://data.lacounty.gov/Criminal/HEAT-MAP-Violent-crimes-within-the-last-12-months/ivnt-3nrc>

https://plot.ly/ggplot2/histograms/