Correlation of Crime and Service Requests in the City of Baltimore

Presented to

Dr. Abhijit Dutt of

The University of Maryland, Baltimore County

Introduction to Data Science

CMSC 491

by

Tyler Little, Jaime Orellana, Celestine Wong

20 May 2018

1. Introduction

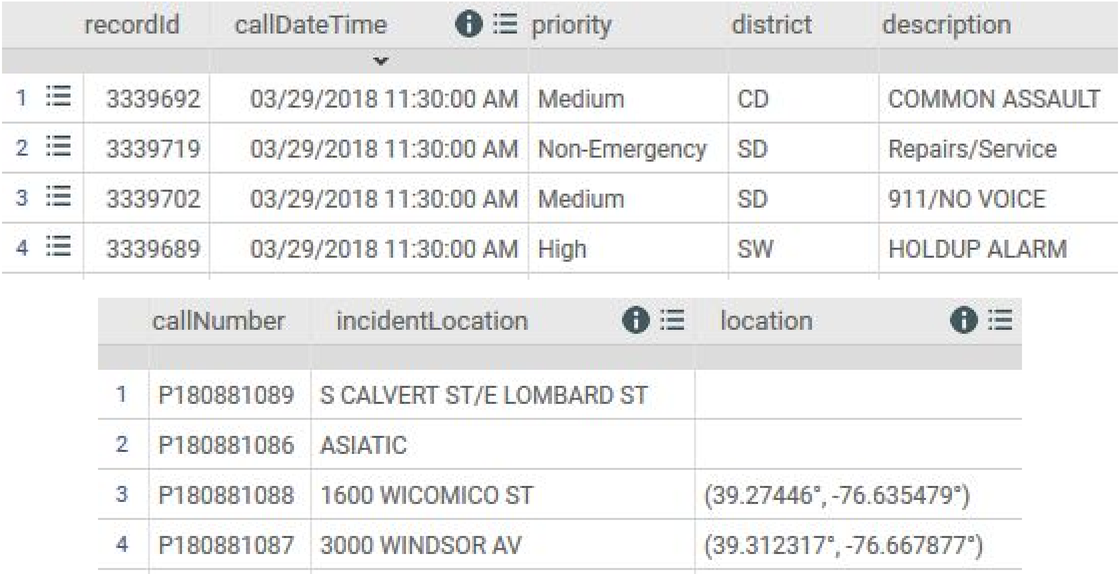
Baltimore City resides on the Chesapeake Bay and is arguably the heart of the state of Maryland. Baltimore city, also known as Charm city, can be called a “city of neighborhoods”; each neighborhood offers its own history, quirkiness, and diversity. The city, as a whole, has a population near 600,000 people encapsulated in ninety-two square miles. This area of dense population requires the need for more upkeep than the surrounding areas of Maryland. This introduces the city’s 311 request service in addition to non-emergency 911 calls. Civilians can make a request by calling, or through an online form. These requests and calls are made available through Open Baltimore and will be used in our study. This study aims to potentially assist Baltimore City in their endeavors to maintain sustainability and well being. To do so, we will analyze any conceivable patterns of crime rates in particular neighborhoods based on patterns of service requests, a neighborhood’s income, and non-emergency 911 calls.

2. Description of Data

Open Baltimore (https://data.baltimorecity.gov/) is one of several open source websites that supply data sheets, maps, documents, reports, and general information related to the City of Baltimore. These accounts are accessible and available for download for the general public. The data sheets that were selected for this study are published from different Baltimore city offices/departments and are updated more than once a day.

2.1 911 Police Calls for Service - Public Safety Data Set

The first data set used in this study is from the Baltimore Police Department. First published on July 13th, 2015, this data set includes entries of 911 emergency and non-emergency calls to the police department with calls starting from June 2015 to the present, March 2018. With over 3.3 million entries, each entry includes numerical values such as the date/time the call was made and its location (latitude/longitude) and categorical values such as its priority, police district, and description. However, this 911 Police Calls for Service data set calls into question how priority is defined and categorized for each call. In addition, then can ask if priority is according to a case chart or table, or if it is subjective to the operator or caller. Other points such as the difference between location and incident location are not explained in the data. Looking at the content of the data, some incident location addresses include house numbers, while others only have the street name. It is also important to note that the location column contains the longitude and latitude; this column is also not a required field since some entries do not have locations. These are some of the observations from our first data set:



	recordId	callDateTime	priority	district	description
1	3339692	03/29/2018 11:30:00 AM	Medium	CD	COMMON ASSAULT
2	3339719	03/29/2018 11:30:00 AM	Non-Emergency	SD	Repairs/Service
3	3339702	03/29/2018 11:30:00 AM	Medium	SD	911/NO VOICE
4	3339689	03/29/2018 11:30:00 AM	High	SW	HOLDUP ALARM

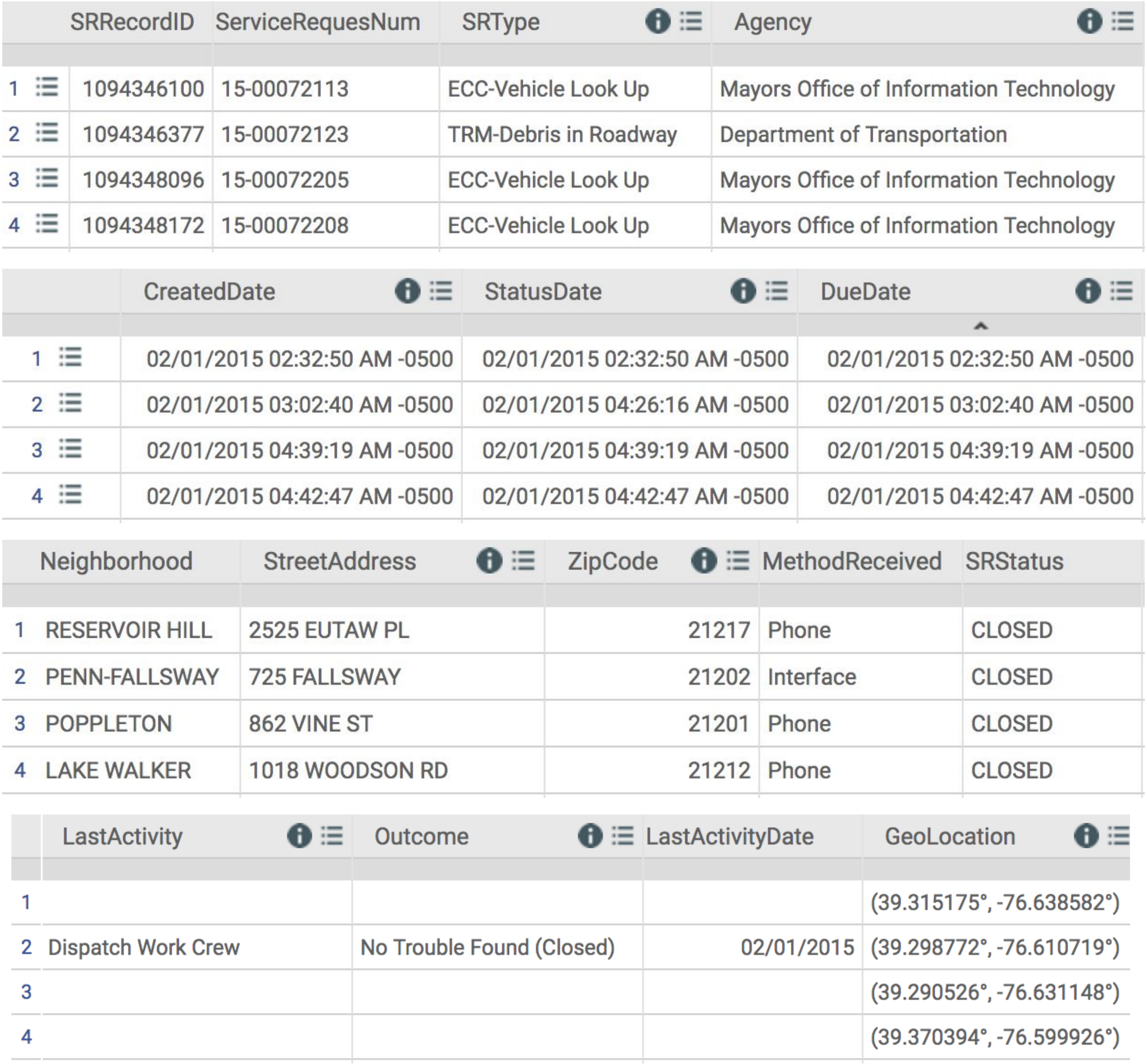
	callNumber	incidentLocation	location
1	P180881089	S CALVERT ST/E LOMBARD ST	
2	P180881086	ASIATIC	
3	P180881088	1600 WICOMICO ST	(39.27446°,-76.635479°)
4	P180881087	3000 WINDSOR AV	(39.312317°,-76.667877°)

Figure 2.1: Snapshot of Police Calls for Service data set

The data set contains both categorical (priority, district, incident location) and quantitative (call datetime and location) data. More on the focus of the data set and dissection in section 3.1

2.2 Customer Service Requests

To analyze patterns in service requests for the city of Baltimore, the 311 Customer Service Requests datasheet contains an updated record of the ongoing and completed requests. According to Open Baltimore, this data set table correlates to Baltimore City's official 311 services request from (https://www.baltimorecity.gov/311-services). This data is provided by the Mayor's Office of Information Technology (MOIT). It contains over 2.5 million entries, since it's creation in January 2011. Each request is given a unique id (numerical, "recordId" column) and service request number ("ServiceRequesNum" column). The other 14 columns note information about the request itself. It is a collection of categorical and numerical data values. For this study, the columns of focus are the service type, neighborhood, service request status, created date, status date, due date, and geolocation. The different dates and geolocation can be considered numeric, while the other columns are more so categorical. Some questions that arise from the data are the categories for service request status and certain acronyms that appear in service request type. It is unclear how an "OPEN" status would differ from "O-Pending", "O-WIP", or the meaning for "HOLDLOCK" and "LOCKED". This data set also includes unfamiliar acronyms for service request type. We speculate the acronym notes the agency responsible for the request. Lastly, it is unclear whether civilians fill out all fields of the service form and if they directly correlate to the columns of the data set, or if personel from a city office categorizes the service request and completes the entry. These few questions are considered minor to the research, but will be noted in the final discussions along with any assumptions made.



	SRRecordID	ServiceRequesNum	SRType	Agency
1	1094346100	15-00072113	ECC-Vehicle Look Up	Mayors Office of Information Technology
2	1094346377	15-00072123	TRM-Debris in Roadway	Department of Transportation
3	1094348096	15-00072205	ECC-Vehicle Look Up	Mayors Office of Information Technology
4	1094348172	15-00072208	ECC-Vehicle Look Up	Mayors Office of Information Technology

	CreatedDate	StatusDate	DueDate
1	02/01/2015 02:32:50 AM -0500	02/01/2015 02:32:50 AM -0500	02/01/2015 02:32:50 AM -0500
2	02/01/2015 03:02:40 AM -0500	02/01/2015 04:26:16 AM -0500	02/01/2015 03:02:40 AM -0500
3	02/01/2015 04:39:19 AM -0500	02/01/2015 04:39:19 AM -0500	02/01/2015 04:39:19 AM -0500
4	02/01/2015 04:42:47 AM -0500	02/01/2015 04:42:47 AM -0500	02/01/2015 04:42:47 AM -0500

	Neighborhood	StreetAddress	ZipCode	MethodReceived	SRStatus
1	RESERVOIR HILL	2525 EUTAW PL	21217	Phone	CLOSED
2	PENN-FALLSWAY	725 FALLSWAY	21202	Interface	CLOSED
3	POPPLETON	862 VINE ST	21201	Phone	CLOSED
4	LAKE WALKER	1018 WOODSON RD	21212	Phone	CLOSED

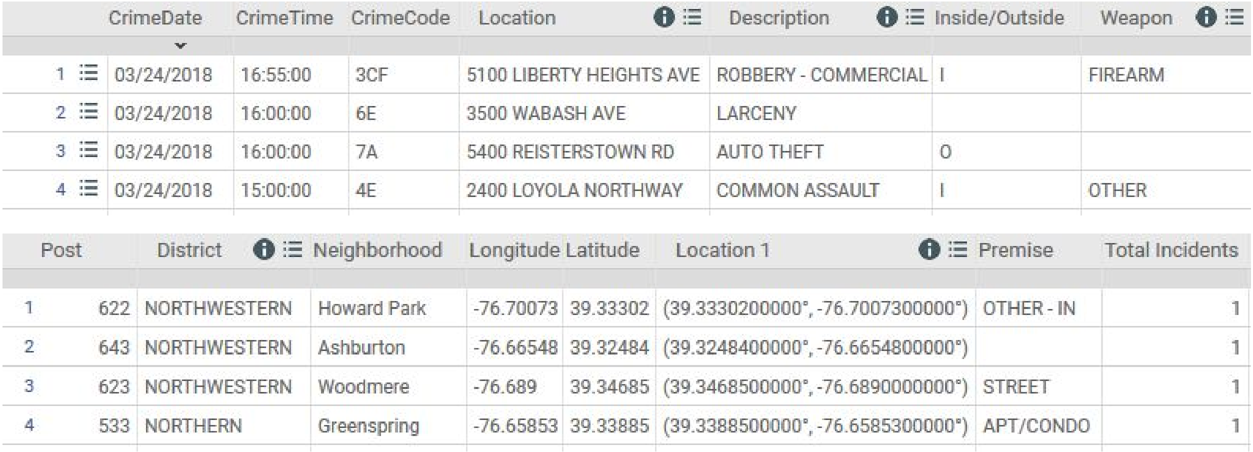
	LastActivity	Outcome	LastActivityDate	GeoLocation
1				(39.315175°, -76.638582°)
2	Dispatch Work Crew	No Trouble Found (Closed)	02/01/2015	(39.298772°, -76.610719°)
3				(39.290526°, -76.631148°)
4				(39.370394°, -76.599926°)

Figure 2.2: Snapshot of Customer Service Request data set

This data set contains the both categorical and quantitative data. Categorical data includes column information from agencies, type of service requested, neighborhoods, and the current status of the ticket. Quantitative data is given by the time stamp of the event.

2.3 Baltimore Police Department - Victim Based Crime

To explore crime in the district, we chose the BPD Part 1 Victim Based Crime Data set from the Baltimore Police Department. With over 303K rows/entries, these records note the date and time of the crime, a crime code along with the description. Although these details outline the context of the crime, this study will highlight the neighborhood, description, longitude and latitude for each entry when focusing on the total crime of an area. Using this data set works well with the 911 Police Calls for Service since they are supplied by the same department. This data set is supplemental to the 911 Police Calls for Service to look into victim based crimes or further drill downs.



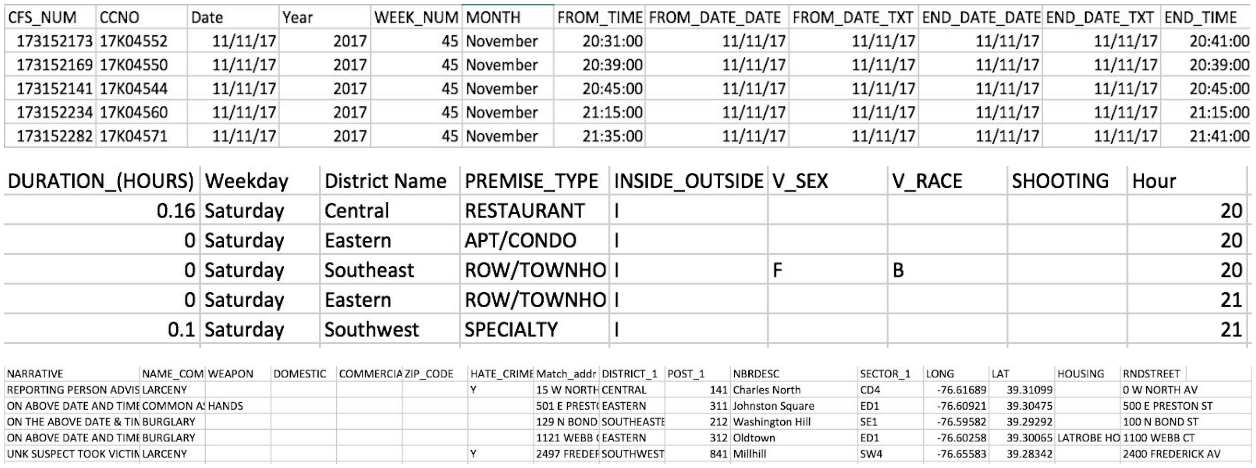
	CrimeDate	CrimeTime	CrimeCode	Location	Description	Inside/Outside	Weapon
1	03/24/2018	16:55:00	3CF	5100 LIBERTY HEIGHTS AVE	ROBBERY - COMMERCIAL	I	FIREARM
2	03/24/2018	16:00:00	6E	3500 WABASH AVE	LARCENY		
3	03/24/2018	16:00:00	7A	5400 REISTERSTOWN RD	AUTO THEFT	O	
4	03/24/2018	15:00:00	4E	2400 LOYOLA NORTHWAY	COMMON ASSAULT	I	OTHER

	Post	District	Neighborhood	Longitude	Latitude	Location 1	Premise	Total Incidents
1	622	NORTHWESTERN	Howard Park	-76.70073	39.33302	(39.3330200000°, -76.7007300000°)	OTHER - IN	1
2	643	NORTHWESTERN	Ashburton	-76.66548	39.32484	(39.3248400000°, -76.6654800000°)		1
3	623	NORTHWESTERN	Woodmere	-76.689	39.34685	(39.3468500000°, -76.6890000000°)	STREET	1
4	533	NORTHERN	Greenspring	-76.65853	39.33885	(39.3388500000°, -76.6585300000°)	APT/CONDO	1

Figure 2.3: Snapshot of Victim Based Crime data set

2.4 Baltimore Police Department - Redacted Year-to-date Master Crime

Another data set the group will utilized is the Redacted YTD Master Crime data set provided by Shiloh. Having a little over 280,000 rows spanning from January 1st 2012, to November 11th 2017, this data set provides information regarding crimes occurring in Baltimore City. One notable difference between this data set and the Victim Based Crime data set is that there are more specific column values regarding these crimes, such as the sex and race of the victim, or if the crime was a shooting or not. One downside is that there are a lot of columns that are not useful to analysis, such as CFS_NUM or CCNO.



CFS_NUM	CCNO	Date	Year	WEEK_NUM	MONTH	FROM_TIME	FROM_DATE_DATE	FROM_DATE_TXT	END_DATE_DATE	END_DATE_TXT	END_TIME
173152173	17K04552	11/11/17	2017	45	November	20:31:00	11/11/17	11/11/17	11/11/17	11/11/17	20:41:00
173152169	17K04550	11/11/17	2017	45	November	20:39:00	11/11/17	11/11/17	11/11/17	11/11/17	20:39:00
173152141	17K04544	11/11/17	2017	45	November	20:45:00	11/11/17	11/11/17	11/11/17	11/11/17	20:45:00
173152234	17K04560	11/11/17	2017	45	November	21:15:00	11/11/17	11/11/17	11/11/17	11/11/17	21:15:00
173152282	17K04571	11/11/17	2017	45	November	21:35:00	11/11/17	11/11/17	11/11/17	11/11/17	21:41:00

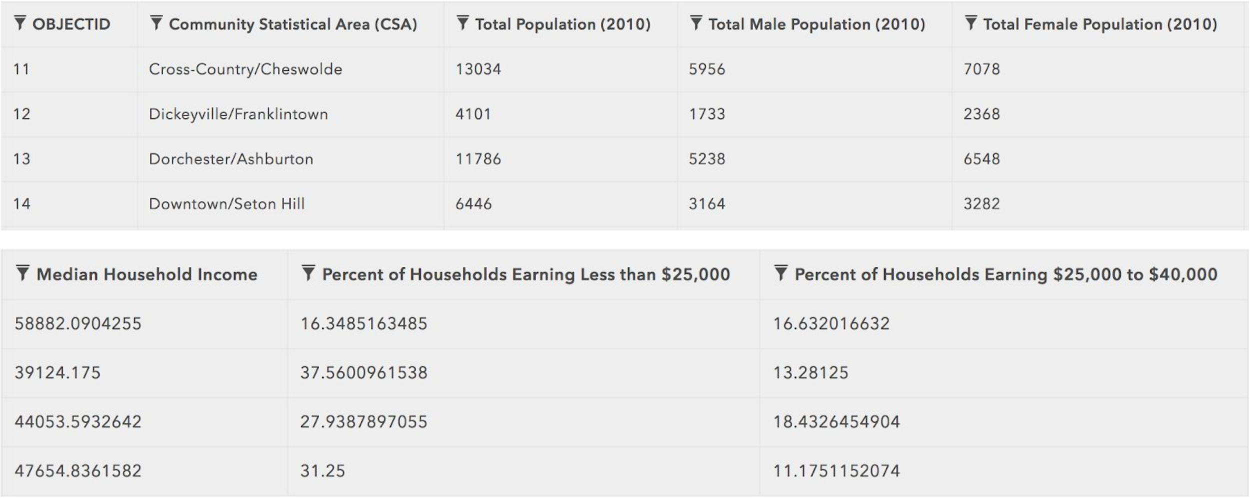
DURATION_(HOURS)	Weekday	District Name	PREMISE_TYPE	INSIDE_OUTSIDE	V_SEX	V_RACE	SHOOTING	Hour
0.16	Saturday	Central	RESTAURANT	I				20
0	Saturday	Eastern	APT/CONDO	I				20
0	Saturday	Southeast	ROW/TOWNHO	I	F	B		20
0	Saturday	Eastern	ROW/TOWNHO	I				21
0.1	Saturday	Southwest	SPECIALTY	I				21

NARRATIVE	NAME_COM	WEAPON	DOMESTIC	COMMERCIA	ZIP_CODE	HATE_CRIME	Match_addr	DISTRICT_1	POST_1	NBRDESC	SECTOR_1	LONG	LAT	HOUSING	RNDSTREET
REPORTING PERSON ADVIS	LARCENY					Y	15 W NORTH	CENTRAL	141	Charles North	CD4	-76.61689	39.31099		0 W NORTH AV
ON ABOVE DATE AND TIME	COMMON A:	HANDS					501 E PRESTO	EASTERN	311	Johnston Square	ED1	-76.60921	39.30475		500 E PRESTON ST
ON THE ABOVE DATE & TIN	BURGLARY						129 N BOND	SOUTHEASTE	212	Washington Hill	SE1	-76.59582	39.29292		100 N BOND ST
ON ABOVE DATE AND TIME	BURGLARY						1121 WEBB (	EASTERN	312	Oldtown	ED1	-76.60258	39.30065	LATROBE HO	1100 WEBB CT
UNK SUSPECT TOOK VICTIN	LARCENY					Y	2497 FREDER	SOUTHWEST	841	Millhill	SW4	-76.65583	39.28342		2400 FREDERICK AV

Figure 2.4: Snapshot of Redacted Year-to-date Master Crime data set

2.5 Vital Signs 16 Census Demographics - Income

Provided by ArcGIS Hub, this data set covers demographic information of residents in Baltimore City. This data set was referenced to the group by Shiloh, a data scientist working for the city. Columns include demographic information regarding income, such as median income and percentage of income in households from differing ranges, and other information regarding race and age. For analytic purposes, the group focused on utilizing the median household income of unique neighborhoods in Baltimore City. Below are screenshots of several columns of the data set. For more information regarding the rest of the columns, this data set can be found at https://data-bniajfi.opendata.arcgis.com/datasets/vital-signs-16-census-demographics/



OBJECTID	Community Statistical Area (CSA)	Total Population (2010)	Total Male Population (2010)	Total Female Population (2010)
11	Cross-Country/Cheswolde	13034	5956	7078
12	Dickeyville/Franklintown	4101	1733	2368
13	Dorchester/Ashburton	11786	5238	6548
14	Downtown/Seton Hill	6446	3164	3282

Median Household Income	Percent of Households Earning Less than $25,000	Percent of Households Earning $25,000 to $40,000
58882.0904255	16.3485163485	16.632016632
39124.175	37.5600961538	13.28125
44053.5932642	27.9387897055	18.4326454904
47654.8361582	31.25	11.1751152074

Figure 2.5: Snapshot of Vital Signs 16 Census Demographics data set

3. Business Challenges

This sections enumerates on obstacles the team has faced during the process of the analysis of data to answer the research questions later posed in this report. The first challenge the team faced was deciding a plan of action to facilitate the analysis of service requests and crime in Baltimore City. This includes the creation of research questions to guide analysis. The next challenge was aggregating the data sets to answer the research questions created. This includes sifting the data sets of relevant column data. Another challenge the team has faced was the cultivation of statistical models to decide if there existed patterns in crime rate and service requests in Baltimore City. These challenges guide the rest of the report and will be further elaborated.

4. Research Questions

In the following sections, the three main focuses of our project will be visited, in terms of goals that will be accomplishable with the data sets. Each question will then follow with a paragraph explaining it in more depth.

4.1 Question One

Is there a correlation between service time to completion and crime rate?

Focusing on the Customer Service Requests and Baltimore Police Department Crime datasheets, we hope to analyze the 2016 data to see if there is a correlation between the amount of service requests, and the time it takes to complete them with the existing crime rate of that neighborhood. This question can probe the distribution of resources across the Baltimore area, particularly with the Baltimore Police Department, since this department answers to both calls. The US City Open Data Census organization publishes other city information for Service Requests (311), from this initial question, it can further explore how the Baltimore area compares to other cities in America.

4.2 Question Two

Is there a correlation between crime rate and the median income of the area?

This second question, takes a look at a larger picture of the Baltimore area. It is common to associate low income areas to having higher crime rates. John Hipp, from the Department of

Criminology, Law and Society and Department of Sociology University of California Irvine, did a recent study and literature review that suggests, "studies have tested how the distribution of economic resources across neighborhoods, as measured by income or poverty, affects neighborhood crime rates or the how the distribution of racial/ethnic minority members across neighborhoods, as measured by the percent nonwhite, and so on, affects neighborhood crime rates" (Hipp, 2007, p. 666). He further explains that a large obstacle in these studies is finding an appropriate reference group. Hipp explores different theories such as relative deprivation, social disorganization, social distance, group threat and consolidated inequality theory. His data is composed of 19 cities selected out of convenience from the US. Census Bureau. Similar to Hipp's study, we hope to focus on the Baltimore area to see if their median household incomes and crime rates play to this ideal and what conclusions we can draw from these results.

4.3 Question Three

Can the crime count be predicted based on the day of the year, the number of service requests of that day, and the median income of the area?

This research question is our focus for exploring a model that could potentially predict the crime count of a certain neighborhood, given the day of the year and the amount of service requests. Because neighborhoods are considered categorical data, we hope to map neighborhoods to their median household incomes. If our model is effective, it could create weight to show the influence of each of our variables. This would also help officers to see how crime numbers may change across the year on a day to day basis and from year to year as median household incomes change from each census.

5. Exploratory Data Analysis

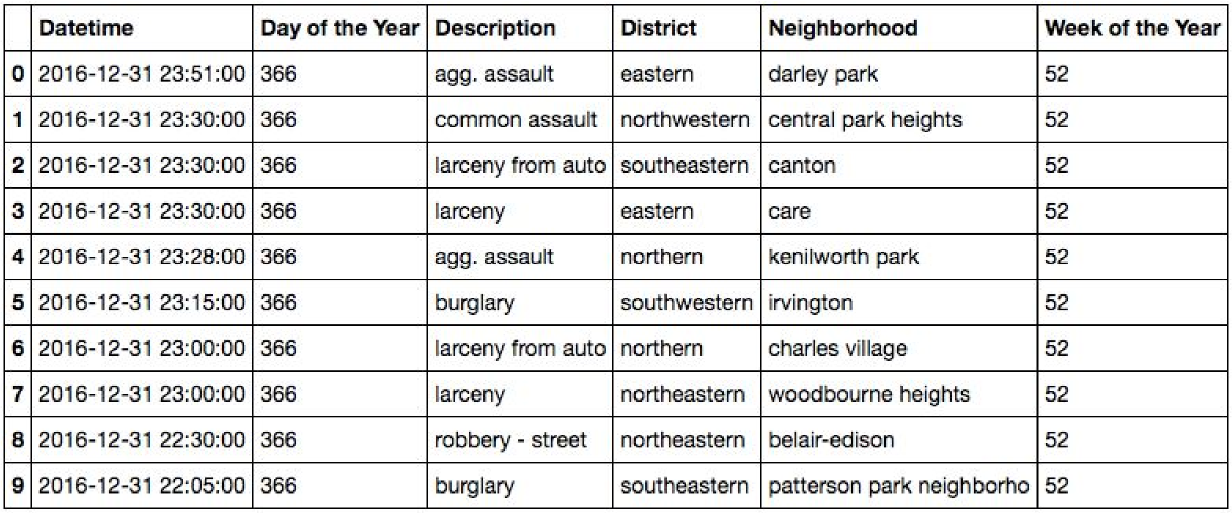
This section delves into the physical implementation of the research project to uncover patterns in crime rates and service requests in Baltimore City. The group accomplishes this by addressing the research questions posed in section four. First, the data sets used are preprocessed in Python, and then loaded into R to create statistical models. The following sections will further discuss and justify the methods used in these processes. Then, there will be a reflection of insights made from analysis.

5.1 Data Processing in Python

This subsection introduces each data set and the expected processes that will be performed to make each data set useful for the creation of our analysis. Note that training and testing data will be split into 70-30 sets for the group's sample data to test any models created in the future.

5.1.1 Data Set: 911, BPD Victim Based Crime

This data set is extremely large, and some data is incomplete. We began our processing by creating start and end dates (Jan 1, 2016 - Dec 31, 2016). Next we created lists for attributes of interest, that is, the date and time of crime, day of the week/year, description, district and the neighborhood. We chose these attributes because we believed a correlation existed among them. The added columns to this sheet are day of the year and week in the year; by adding these columns we added numerical values for days (1-365) and weeks (1-52). Once these lists are created, we filtered through each row for complete entries to create our selected data sheet for later. Below is a snapshot of our data in table form after processing and adding columns:

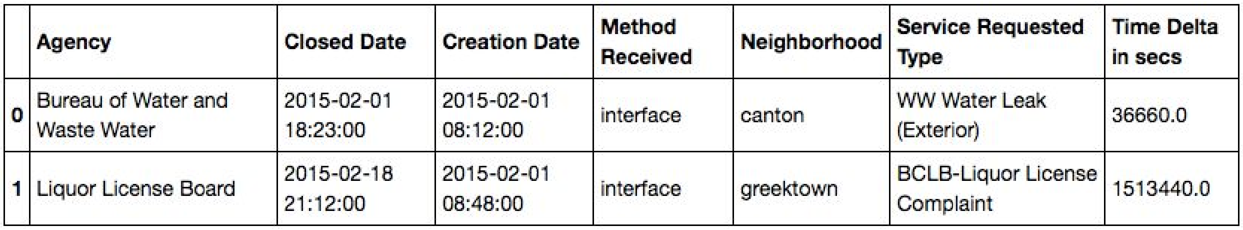


	Datetime	Day of the Year	Description	District	Neighborhood	Week of the Year
0	2016-12-31 23:51:00	366	agg. assault	eastern	darley park	52
1	2016-12-31 23:30:00	366	common assault	northwestern	central park heights	52
2	2016-12-31 23:30:00	366	larceny from auto	southeastern	canton	52
3	2016-12-31 23:30:00	366	larceny	eastern	care	52
4	2016-12-31 23:28:00	366	agg. assault	northern	kenilworth park	52
5	2016-12-31 23:15:00	366	burglary	southwestern	irvington	52
6	2016-12-31 23:00:00	366	larceny from auto	northern	charles village	52
7	2016-12-31 23:00:00	366	larceny	northeastern	woodbourne heights	52
8	2016-12-31 22:30:00	366	robbery - street	northeastern	belair-edison	52
9	2016-12-31 22:05:00	366	burglary	southeastern	patterson park neighborho	52

Figure 5.1.1: Snapshot of 911 data set after preprocessing

5.1.2 Data Set: 311

Similar to 911, Victim Based Crime Data, we filtered the millions of 311 requests also to fit for the year 2016. One difference in our processing is casting the CreatedDate and StatusDate to datetime objects. The attributes of interest for this sheet include: service type, agency, method received, creation date, completion date and difference in time. When iterating over the rows in this dataset, we looked specifically for closed requests. Once we know the request is closed, we take the difference between the created date and status datel this is used to populate the difference in time attribute. This was one of the components we learned from Shiloh—he mentioned in class that the best time to gauge requests time is to look at the status date. Below is a snapshot of our data in table form after processing and adding columns:

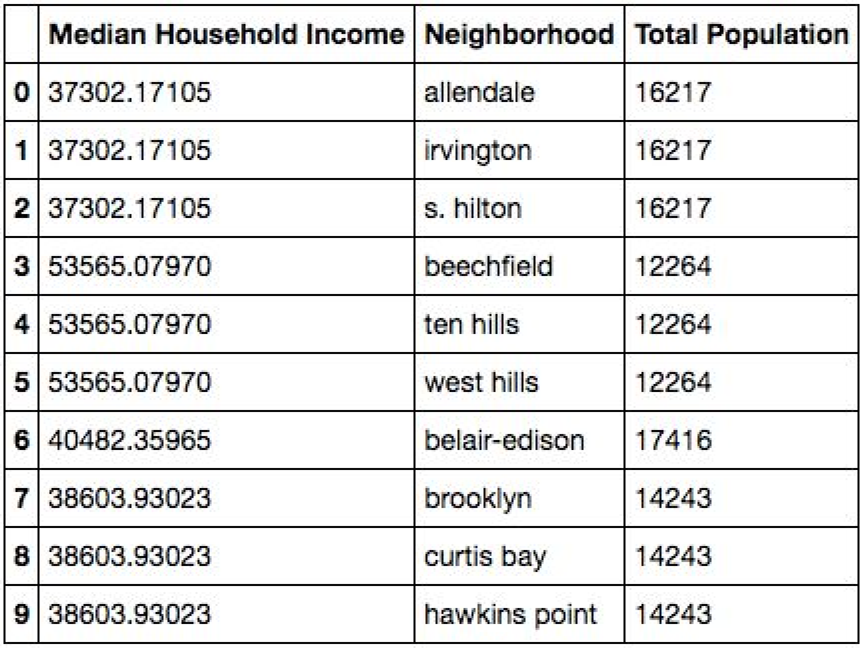


	Agency	Closed Date	Creation Date	Method Received	Neighborhood	Service Requested Type	Time Delta in secs
0	Bureau of Water and Waste Water	2015-02-01 18:23:00	2015-02-01 08:12:00	interface	canton	WW Water Leak (Exterior)	36660.0
1	Liquor License Board	2015-02-18 21:12:00	2015-02-01 08:48:00	interface	greektown	BCLB-Liquor License Complaint	1513440.0

Figure 5.1.2: Snapshot of 311 data set after preprocessing

5.1.3 Incomes, Vital Signs 16 Census Demographics

Since neighborhoods is considered to be categorical data, our group decided on representing these areas based on their median household income. This information is found from the 2016 Census Demographics sheet. We looked on a CSA2010 level, as suggested by Shiloh. From this data sheet we filtered out the columns neighborhood, total population, and median household income. Having neighborhoods is the way we can join our datasheets. The median household gives us another numerical indicator. Lastly, the total population will help in identifying the crime rate of a certain area (e.g. total crime/total population). Below is a snapshot of our data in table form

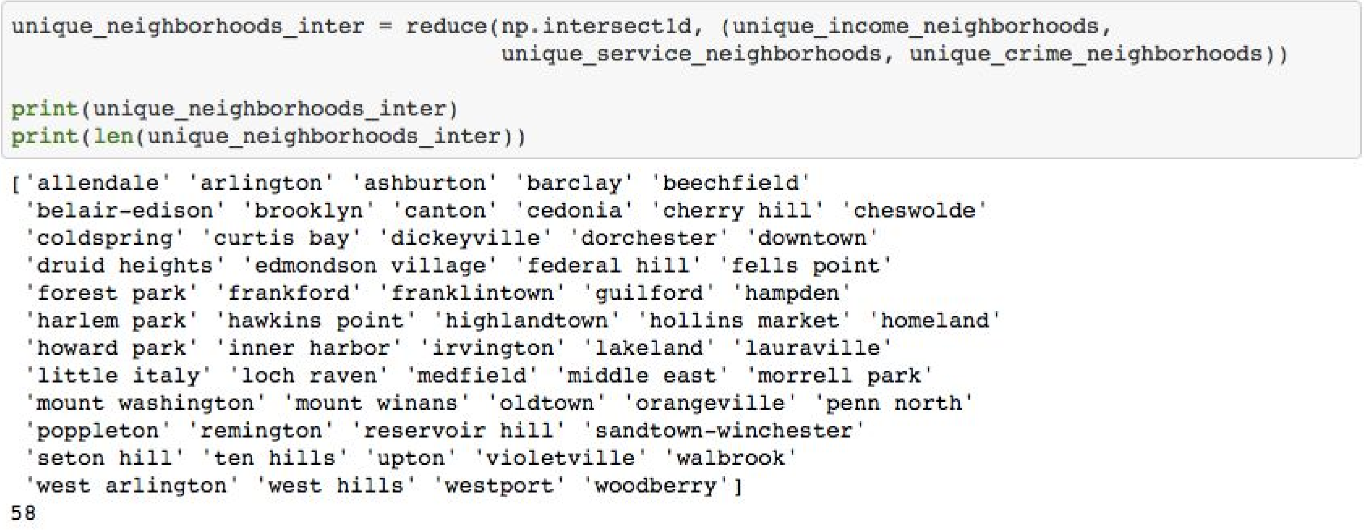


	Median Household Income	Neighborhood	Total Population
0	37302.17105	allendale	16217
1	37302.17105	irvington	16217
2	37302.17105	s. hilton	16217
3	53565.07970	beechfield	12264
4	53565.07970	ten hills	12264
5	53565.07970	west hills	12264
6	40482.35965	belair-edison	17416
7	38603.93023	brooklyn	14243
8	38603.93023	curtis bay	14243
9	38603.93023	hawkins point	14243

Figure 5.1.3: Snapshot of income data set after preprocessing

5.1.4 Data Fusion for Model

To effectively facilitate the observation of a correlation between service request and total crime, the data set of 911 non-emergency calls will be appended to the data set of 311 calls for service. This can be done essentially by having a ‘join’ on neighborhoods. We first took a look at the neighborhoods from each of the three datasheets and created a subset to serve as the intersect for all three. From this we have 58 neighborhoods to filter through across all the datasheets.



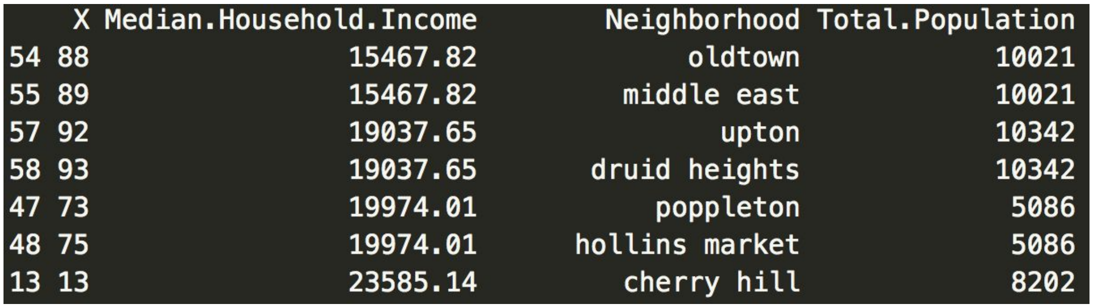
```
unique_neighborhoods_inter = reduce(np.intersect1d, (unique_income_neighborhoods,
                                    unique_service_neighborhoods, unique_crime_neighborhoods))

print(unique_neighborhoods_inter)
print(len(unique_neighborhoods_inter))
```

```
['allendale' 'arlington' 'ashburton' 'barclay' 'beechfield'
 'belair-edison' 'brooklyn' 'canton' 'cedonia' 'cherry hill' 'cheswolde'
 'coldspring' 'curtis bay' 'dickeyville' 'dorchester' 'downtown'
 'druid heights' 'edmondson village' 'federal hill' 'fells point'
 'forest park' 'frankford' 'franklintown' 'guilford' 'hampden'
 'harlem park' 'hawkins point' 'highlandtown' 'hollins market' 'homeland'
 'howard park' 'inner harbor' 'irvington' 'lakeland' 'lauraville'
 'little italy' 'loch raven' 'medfield' 'middle east' 'morrell park'
 'mount washington' 'mount winans' 'oldtown' 'orangeville' 'penn north'
 'poppleton' 'remington' 'reservoir hill' 'sandtown-winchester'
 'seton hill' 'ten hills' 'upton' 'violetville' 'walbrook'
 'west arlington' 'west hills' 'westport' 'woodberry']
58
```

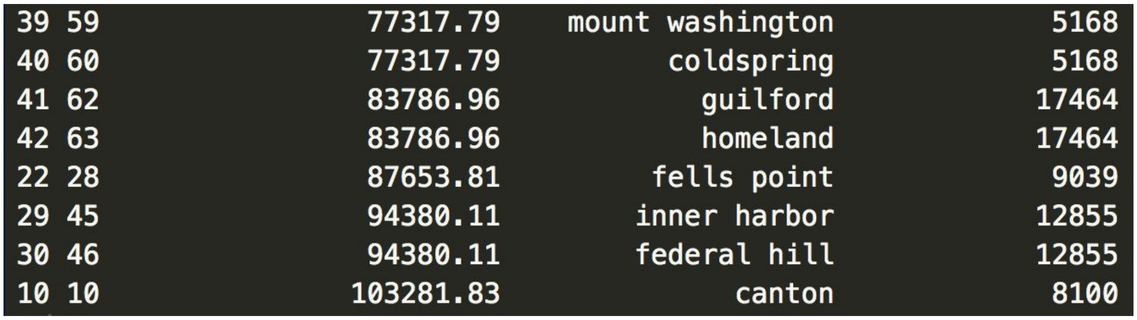
Figure 5.1.4.a: List of unique neighborhoods in Baltimore City among all data sets

We went back to the previous filtered datasets to include those 58 neighborhoods. From this we exported three csv files (1) income, (2) crime, and (3) service. This was further processed in R. Before processing the models, we explored our filtered datasets.

From the filtered income data set, it was possible to assess to the low and high income areas:



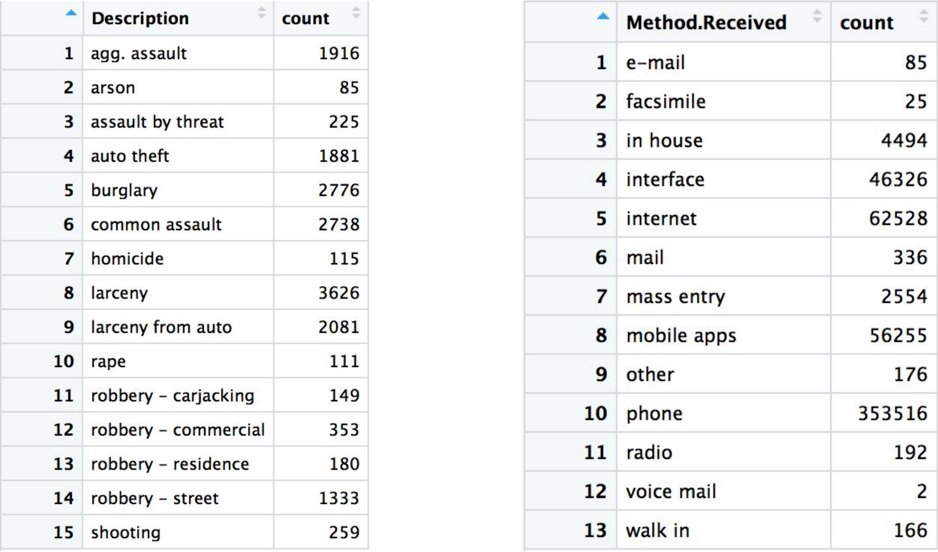
```
    X Median.Household.Income        Neighborhood Total.Population
54 88                15467.82             oldtown            10021
55 89                15467.82         middle east            10021
57 92                19037.65               upton            10342
58 93                19037.65       druid heights            10342
47 73                19974.01           poppleton             5086
48 75                19974.01      hollins market             5086
13 13                23585.14         cherry hill             8202
```



```
39 59              77317.79    mount washington         5168
40 60              77317.79          coldspring         5168
41 62              83786.96            guilford        17464
42 63              83786.96            homeland        17464
22 28              87653.81          fells point        9039
29 45              94380.11        inner harbor        12855
30 46              94380.11        federal hill        12855
10 10             103281.83              canton         8100
```

Figure 5.1.4.b and 5.1.4.c: Snapshots of low and high income neighborhoods in Baltimore City

We then assessed the number reports from each crime from the filtered crime data set. And finally, from the filtered service dataset, we assess the number of reports from each service request:



	Description	count
1	agg. assault	1916
2	arson	85
3	assault by threat	225
4	auto theft	1881
5	burglary	2776
6	common assault	2738
7	homicide	115
8	larceny	3626
9	larceny from auto	2081
10	rape	111
11	robbery – carjacking	149
12	robbery – commercial	353
13	robbery – residence	180
14	robbery – street	1333
15	shooting	259

	Method.Received	count
1	e–mail	85
2	facsimile	25
3	in house	4494
4	interface	46326
5	internet	62528
6	mail	336
7	mass entry	2554
8	mobile apps	56255
9	other	176
10	phone	353516
11	radio	192
12	voice mail	2
13	walk in	166

Figure 5.1.4.d: Summary table of preprocessed crime data set (left)
Summary table of preprocessed service data set (right)

5.2 Merging the Data Sets

Using the preprocessed data sets allows for the cultivation of statistical model building in R. To do this, once the data is loaded, we were able to create a joint dataset with the neighborhoods, its crime ratio, the average wait time for a service request, the median service wait time and the incomes. The crime ratio is calculated by the crime count divided by the total population. The average and median service wait times were calculated to also show the disparity between the two. This shows us that there are outliers in our data that may skew our data. We then decided to create a second joint datasheet, but this second datasheet would focus more on grouping neighborhoods with day of the year by crime counts and service counts. At the end of our data processing we have two datasheets to help answer our three research questions; the first (1) contains the neighborhoods names, incomes, crime rate, average and median service wait time, the second (2) contains service requests and crime counts based on neighborhood and day of the year.

5.3 Primitive Analysis

With this new data sheet we created, we can do some primitive analyses into the values of this data frame for every neighborhood in Baltimore City:

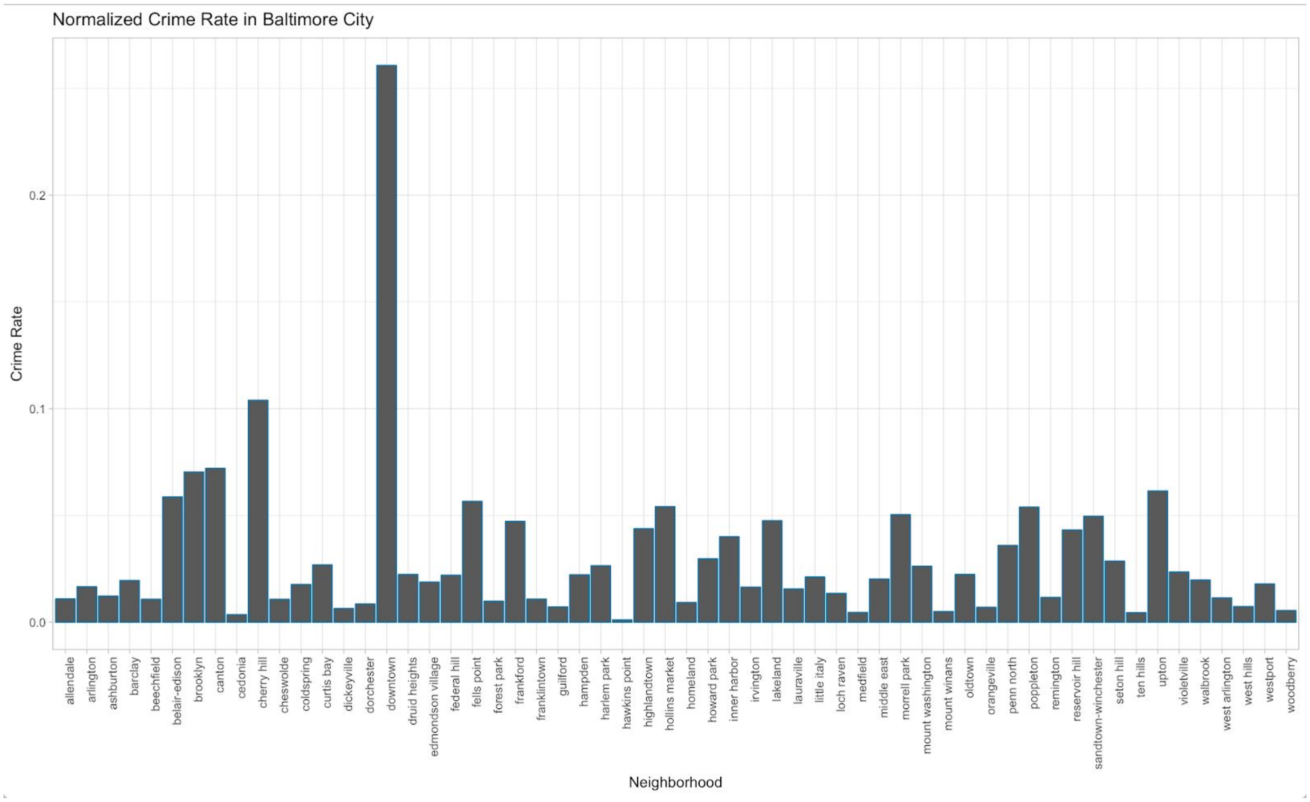


Figure 5.3.a: Bar graph of normalized crime rates

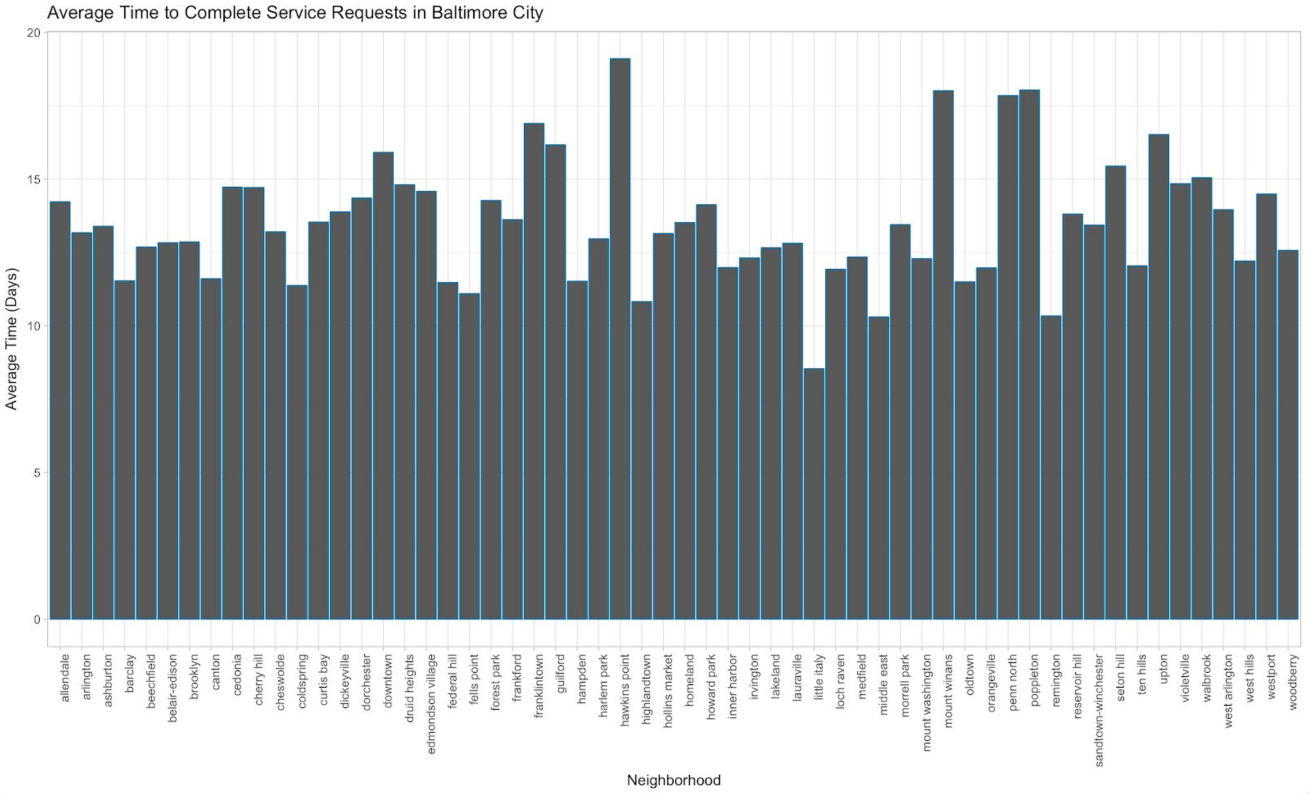


Figure 5.3.b: Bar graph of average time to complete service requests

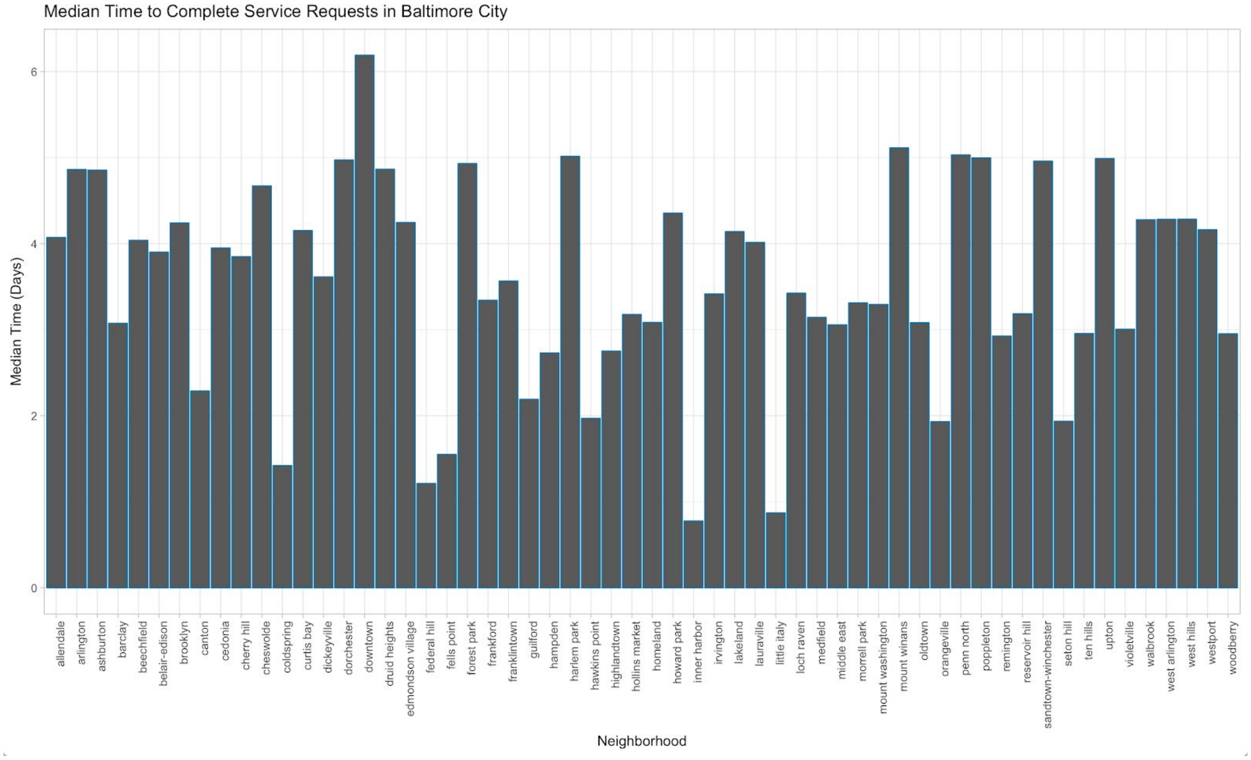


Figure 5.3.c: Bar graph of median time to complete service requests

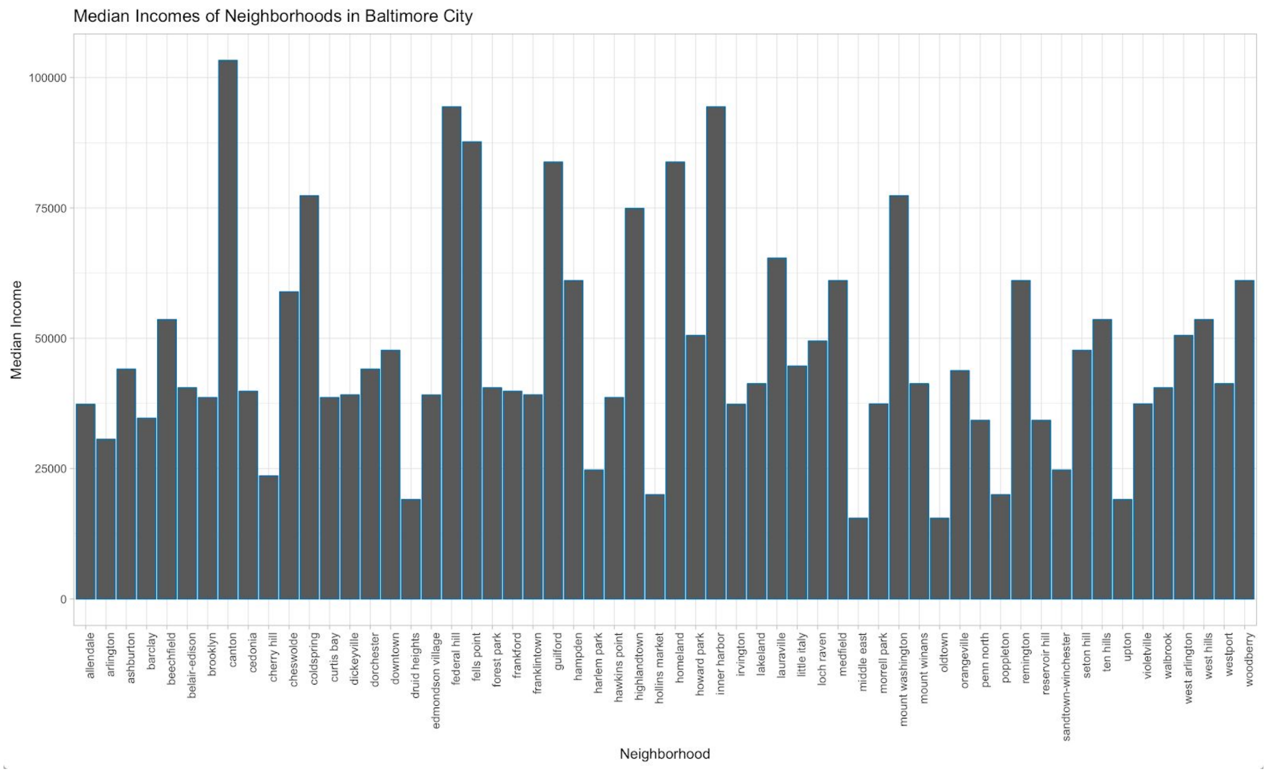


Figure 5.3.d: Bar graph of median income

6. Predictive Modeling Results

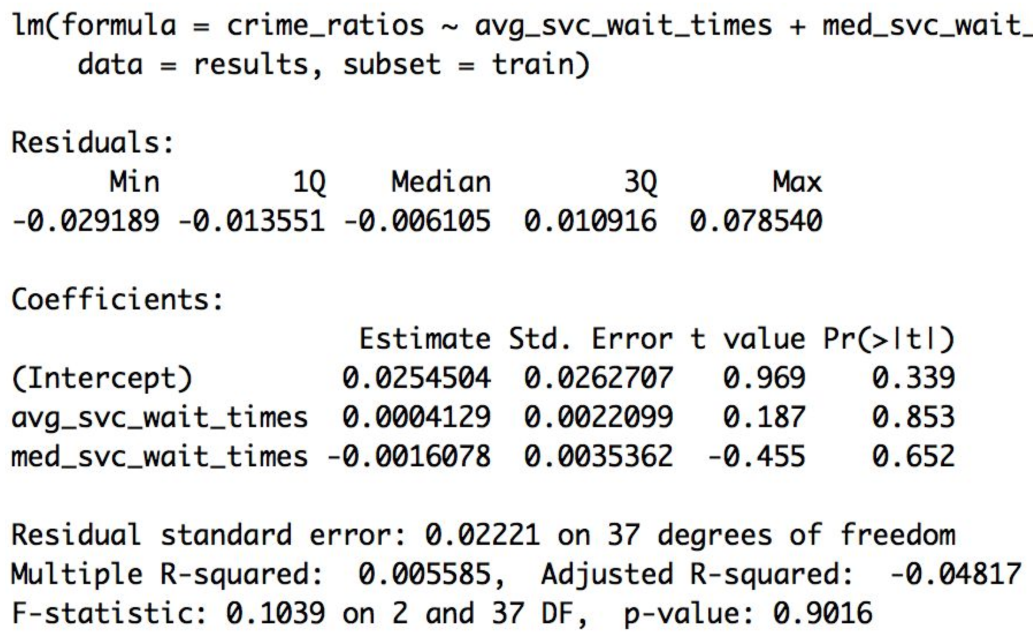
The focus of this section is to analyze the models we created using the algorithms of Multiple Regression and Hierarchical Clustering.

6.1: Multiple Regression

6.1.1 Multiple Regression for Crime Ratio

To see if there is a correlation between crime rate and the time taken to complete a service request, we create a multiple regression model using the crime rate as the response variable with average and median times to complete service requests as the predictor variables. We chose this model because it is a simplistic approach for looking at the correlation of all attributes regarding time of service requests.

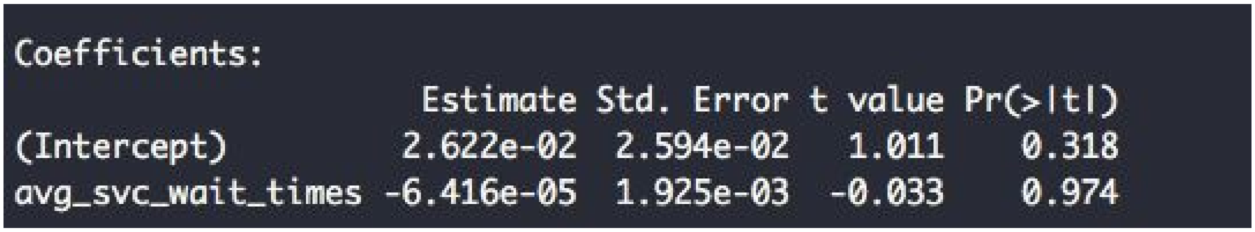
A summary of the initial model is available on the next page.



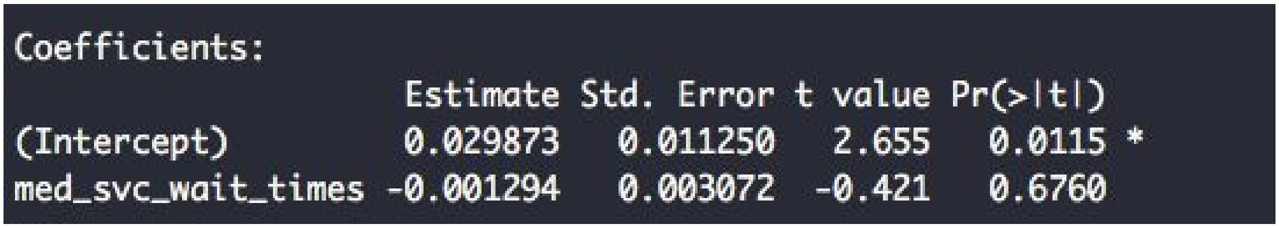
```
lm(formula = crime_ratios ~ avg_svc_wait_times + med_svc_wait_times,
    data = results, subset = train)

Residuals:
      Min        1Q    Median        3Q       Max
-0.029189 -0.013551 -0.006105  0.010916  0.078540

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.0254504  0.0262707   0.969    0.339
avg_svc_wait_times  0.0004129  0.0022099   0.187    0.853
med_svc_wait_times -0.0016078  0.0035362  -0.455    0.652

Residual standard error: 0.02221 on 37 degrees of freedom
Multiple R-squared:  0.005585,  Adjusted R-squared:  -0.04817
F-statistic: 0.1039 on 2 and 37 DF,  p-value: 0.9016
```

Figure 6.1.1.a: Multiple regression model of crime rate

From first glance at these summary results, the Multiple R-squared is significantly low as to suggest the ineffectiveness of the model. In addition to this, our p-values for each coefficient are too high to consider each variable important. Another observation we made from these results show the Adjusted R-squared to be -0.04817. With a negative value, the model suggests that there is a variable that is detrimental to our model. Therefore, we created linear regression models for each of the variables individually and get the following results:



```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.622e-02  2.594e-02   1.011    0.318
avg_svc_wait_times -6.416e-05  1.925e-03  -0.033    0.974
```



```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.029873   0.011250   2.655   0.0115 *
med_svc_wait_times -0.001294   0.003072  -0.421   0.6760
```

Figure 6.1.1.b: Results from individual response variables for average service wait time, and median service wait time.

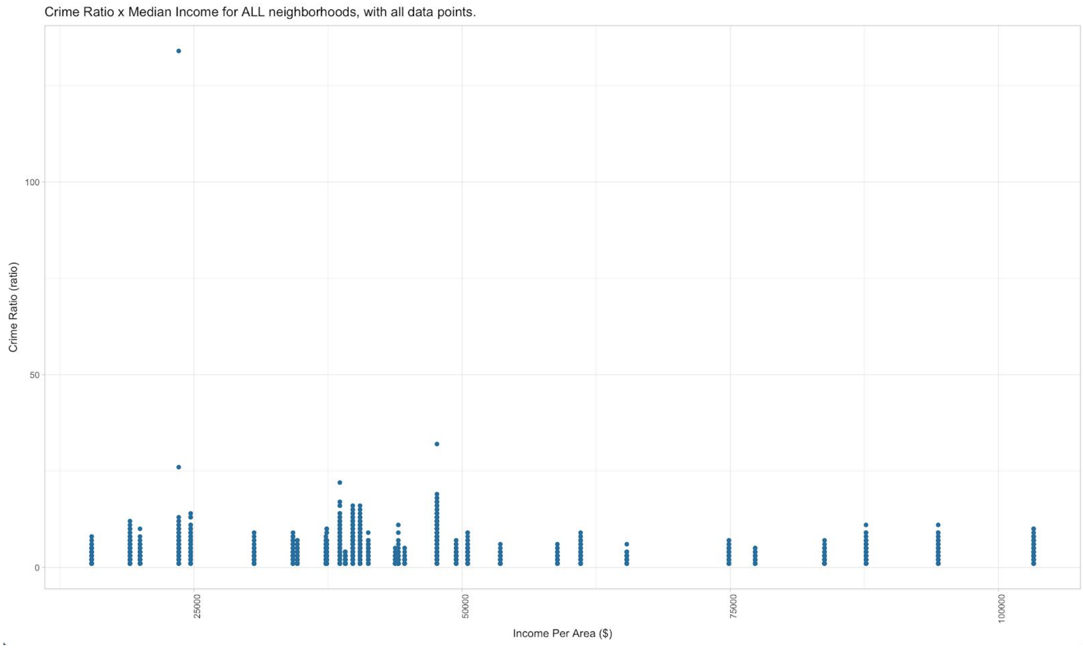
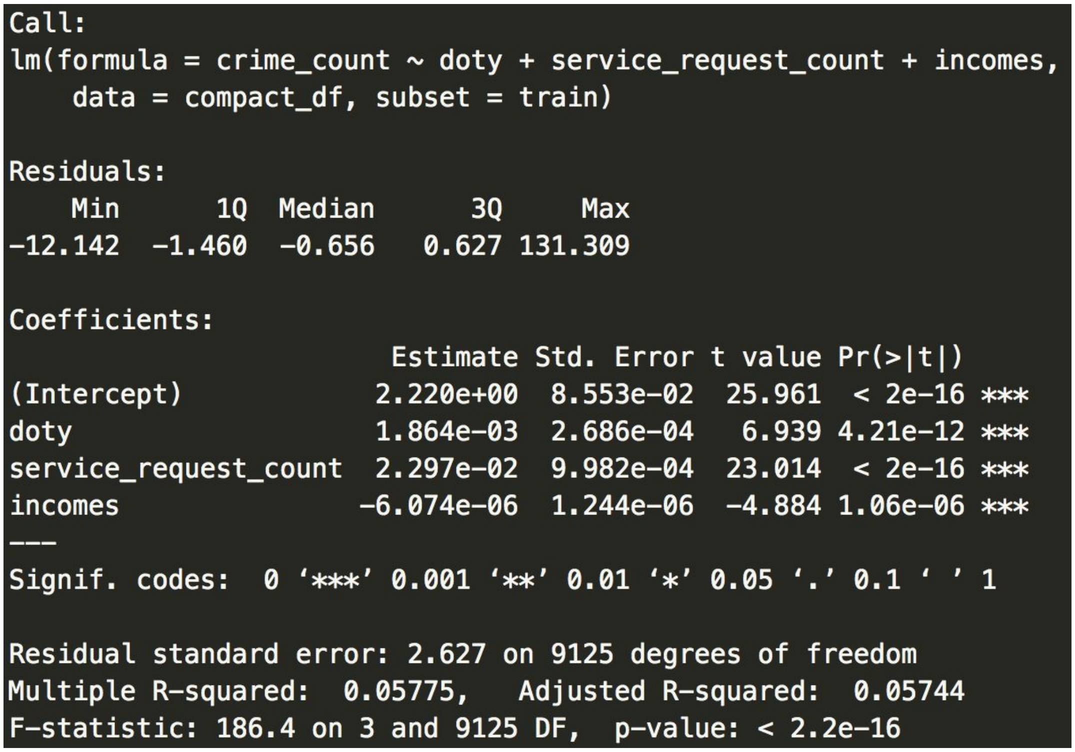


Figure 6.1.1.c: Scatter Plot of Income of all neighborhoods, and their respective crime ratio.

These results show that the median service request time to completion is a better predictor variable than the average service request time to completion because of its smaller p-value. This can be explained when looking at our data. When looking at the average and median calculations, the two values had a high difference due to outliers in the dataset. Knowing this, the median would be a better tell for this model. However the coefficient and the p value are still considered low for an acceptable model. Because of these results, we decided looking at these summary values are sufficient to bypass any accuracy checks, rather than approaching the model with cross validation.

6.1.2 Multiple Regression for Crime Count

After looking specifically at service request completion times and seeing our models yield unacceptable results, we tried a different approach in relating crime, season, neighborhoods and service requests. This approach still used Multiple Regression in handling these different variables. In this approach, our response variable is crime count, rather than a ratio. The predictor variables for this then became day of the year, service request counts for that day, and the income of the area (to represent neighborhoods). This model used the second data sheet we created that list specifically crime counts, service counts, grouped by income (to represent neighborhoods) and day of the year (doty). When using this datasheet in R, we get the following summary results for our Multiple Regression model:



```
Call:
lm(formula = crime_count ~ doty + service_request_count + incomes,
    data = compact_df, subset = train)

Residuals:
    Min      1Q  Median      3Q     Max
-12.142  -1.460  -0.656   0.627 131.309

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.220e+00  8.553e-02  25.961  < 2e-16 ***
doty                   1.864e-03  2.686e-04   6.939 4.21e-12 ***
service_request_count  2.297e-02  9.982e-04  23.014  < 2e-16 ***
incomes               -6.074e-06  1.244e-06  -4.884 1.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.627 on 9125 degrees of freedom
Multiple R-squared:  0.05775,   Adjusted R-squared:  0.05744
F-statistic: 186.4 on 3 and 9125 DF,  p-value: < 2.2e-16
```

Figure 6.1.2.a: Multiple regression model of crime count based on service request count, neighborhood incomes and day of the years

The first observation we made from these results is the improvement of overall p values for each of the coefficients. However, when looking at the estimates we also see low coefficients to suggest low influence of that variable to the model. Secondly, we can note the positive Adjusted R-squared value, comparative to our prior negative value. If we are to use the model, we can get the following equation using each of the coefficient estimates.

$$\begin{aligned} \textit{crime count} \;=\; 2.220 \quad &+ \;.001864\, \textit{doty} \\ &+ \;.02297\, \textit{service request count} \\ &- \;.000006074\, \textit{incomes} \end{aligned}$$

Although the p values show smaller values than our previous multiple regression models, the multiple r-squared value is still considered low. Therefore, the variation of the response variable, in this case the crime count, from the data is not explained well by the model and is hard to make conclusions from it.

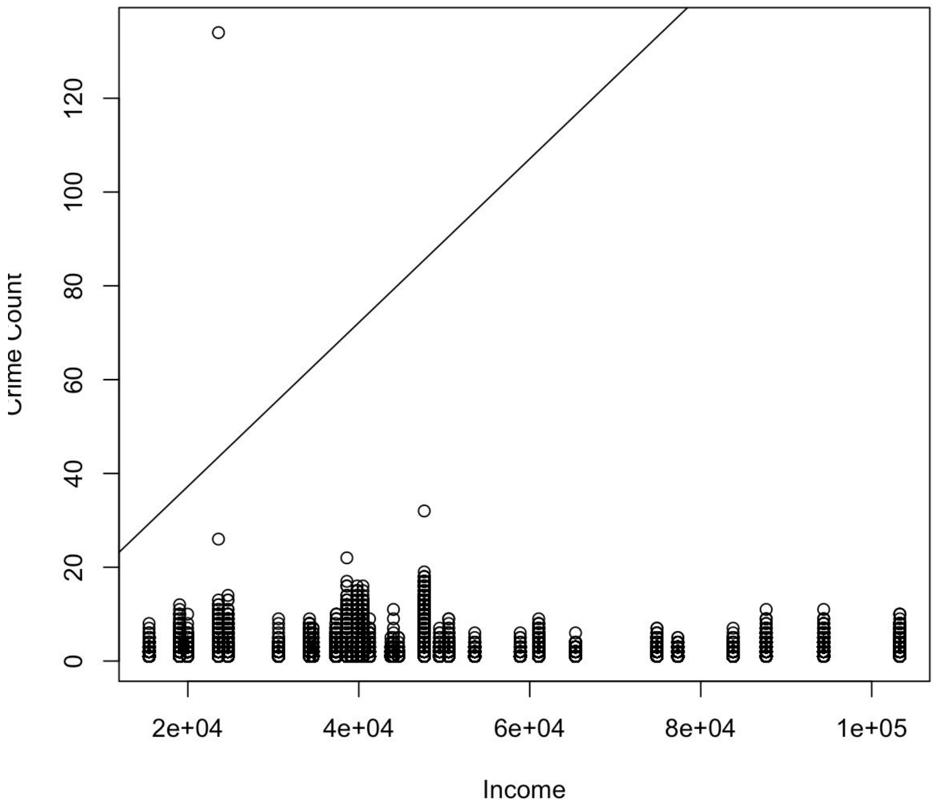


Figure 6.1.2.b: Linear fit model for income and crime count prediction. No direct correlation is visible.

6.2 Hierarchical Clustering

To try another learning algorithm, we decided on a way of clustering the data in hopes to see some pattern or structure in the data. We decided on using hierarchical clustering as our method of clustering. Different from the clustering methods we learned in class, the hierarchical clustering method can be done by looking at plot points, calculating euclidean distances from points and grouping close points together, ultimately these pairings turn to groupings/clusterings as clusters get closer to one another -- this then creates a hierarchy. The merging of clusters are the visual splittings of the tree where the next branch is the next closest point/cluster. Our linkage is depicted on the next page. We used this algorithm to attempt at seeing whether areas of high crime rate are close to one another, and how they distribute across baltimore. In addition, it was speculated that there must be an underlying connection between crime and income ratios. Below depicted is a graph of crime ratios and the average median household income.

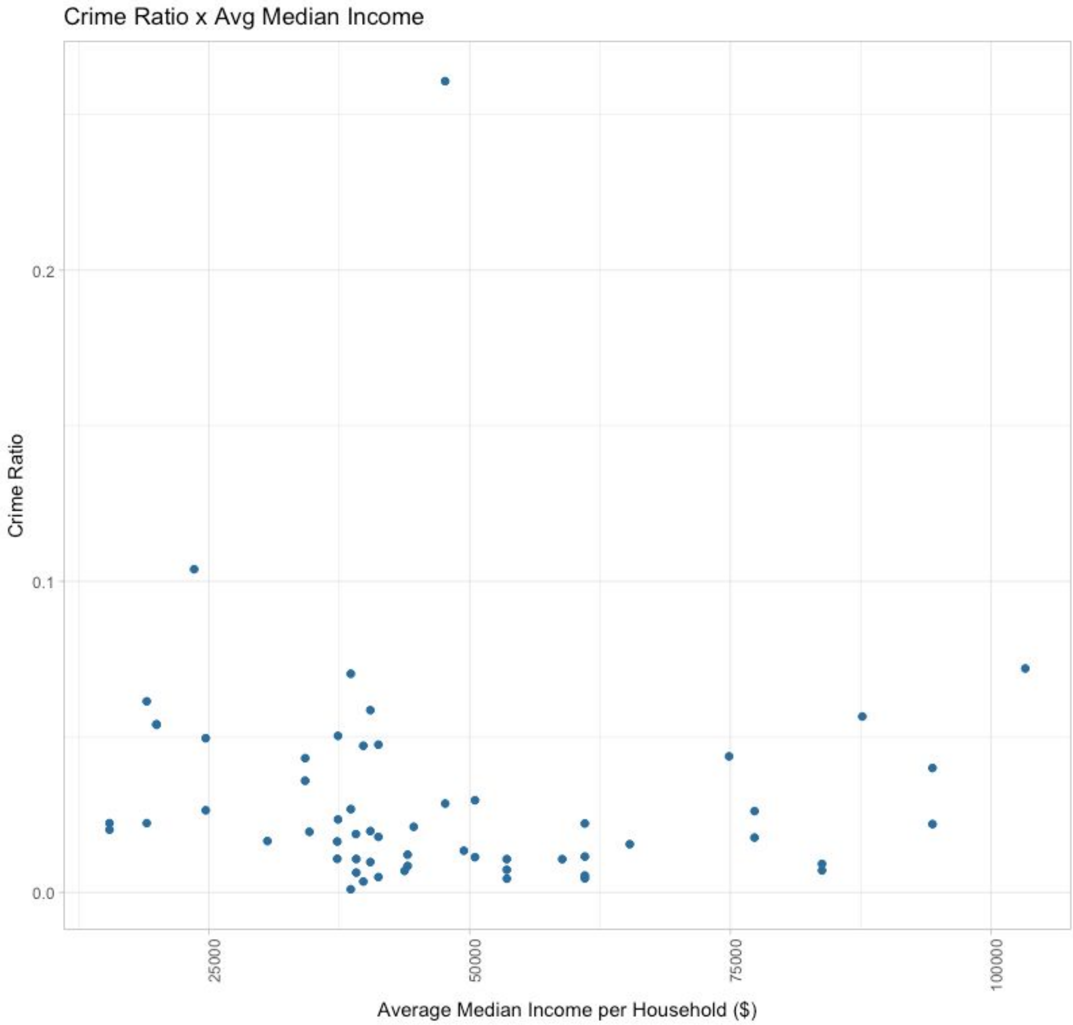
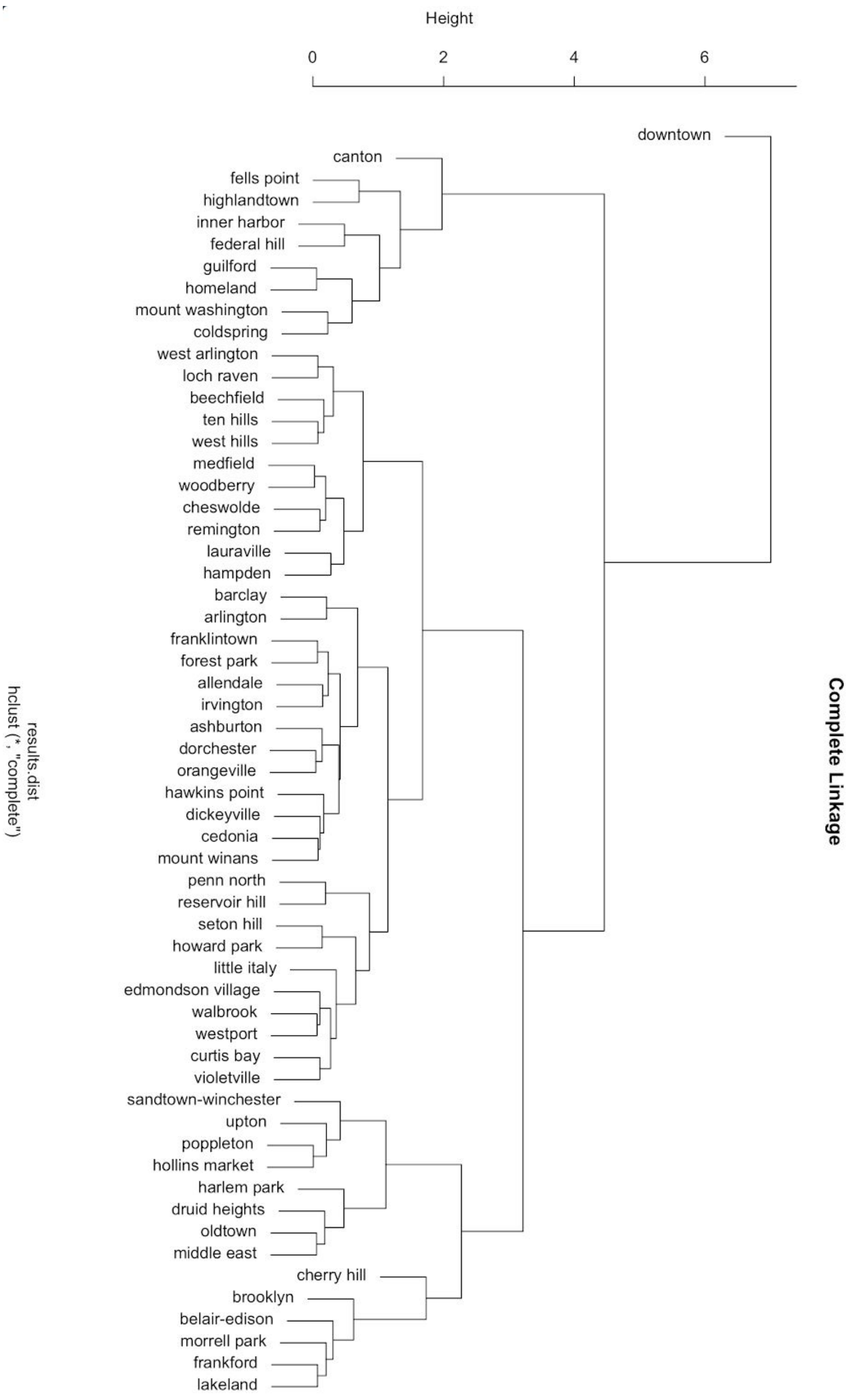


Figure 6.2.1 Scatter Plot for yearly annual income, and and yearly crime ratio averages. Crime ratio defined as occurring once a year to an individual in the neighborhood.

Looking at the graph there didn't seem to be a direct correlation, and so in attempt to build a generalization and prove that income was related to crime we built a hierarchical cluster. This is depicted in the next page. From the hierarchical tree, it is possible to identify correct clusters for high crime activity, between Cherry Hill, Brooklyn, Belair-Edison, Morrell Park, Frankford, and Lakeland. There are areas well known to the public to not only known for high crime, but generally dangerous to step foot in. Alongside to the left of the tree it is possible to notice below height two the growth of a node that is generally known to be very safe to visit. Examples like Canton, Fells Point, Federal Hill, and Inner Harbor tend to be extremely safe to visit. It is apparent across the tree that the clusters identified high and low crime areas not only by the crime count and income, but in addition to it the severity of those crimes.


Complete Linkage
results.dist
hclust (*, "complete")
Height
0
2
4
6
downtown
canton
fells point
highlandtown
inner harbor
federal hill
guilford
homeland
mount washington
coldspring
west arlington
loch raven
beechfield
ten hills
west hills
medfield
woodberry
cheswolde
remington
lauraville
hampden
barclay
arlington
franklintown
forest park
allendale
irvington
ashburton
dorchester
orangeville
hawkins point
dickeyville
cedonia
mount winans
penn north
reservoir hill
seton hill
howard park
little italy
edmondson village
walbrook
westport
curtis bay
violetville
sandtown-winchester
upton
poppleton
hollins market
harlem park
druid heights
oldtown
middle east
cherry hill
brooklyn
belair-edison
morrell park
frankford
lakeland

7. Results and Discussion

Given the data on Baltimore City's crime, service requests, and neighborhood information, we were able to process the data to be useful for creating models in attempt to find variables for prediction; some of our models were proven more effective than others. In this section, we revisit the researched questions posed earlier in this report, and finally conclude our analysis.

7.1 Question One

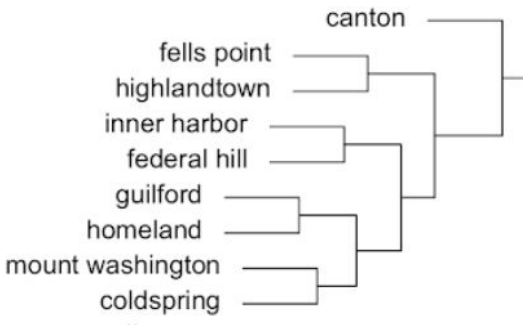
Is there a correlation between service time to completion and crime rate?

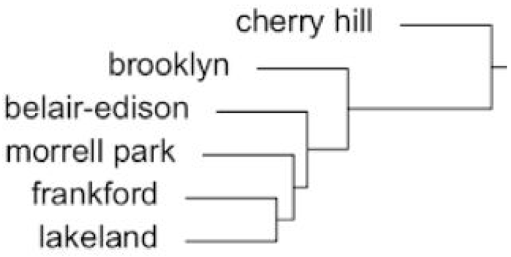
With regards to the multiple regression model we created: given that the p value is 0.9, the multiple r-squared value is low, and the t values for the response variables are fairly low, we are unable to deduce a correlation among service time to completion and the rate of crime.

7.2 Question Two

Is there a correlation between crime rate and the median income of the area?

By answering the question whether or not there existed a relationship between crime rate and income, it was possible to determine clusters. Below are pictures depicting a hierarchical cluster and their corresponding high and low crime areas that are defined based on the severity of crimes committed in those areas. Basically, it was possible to cluster where the extremely dangerous places were located in Baltimore, regardless if there was no direct correlation in the graphs of income and crime.





7.3 Question Three

Can the crime count be predicted based on the day of the year, the number of service requests of that day, and the median income of the area?

In our study, we attempt to answer this question using our second multiple regression model that tried to predict crime count using the day of the year, the median income of the neighborhood, and the amount of service requests made that day of the year. Although this model showed lower p values than the prior regression model, the severely low multiple r-squared value makes our model inconclusive. By our model, we are unable to make confident conclusions about the correlation between service requests, median income and the day of the year. In a further study, there can deeper analysis of one of the variables, such as, for service requests there are different attributes like service types and service statuses that may have more information to give compared to service requests volume.

7.4 Conclusion

With our analysis of data sets provided by Baltimore City and fellow colleagues, we have attempted to provide a means of aid to citizens of the surrounding area by the prediction of crime rates. Throughout this process, we have learned many new techniques involving the extraction, aggregation, and modeling of extremely large, time series logged and processed data.

With the given data and its many attributes, it is difficult to create a generalized model to correlate service requests with neighborhoods, and then with seasons and neighborhoods. Further analysis with this data should focus on one of the four attributes, such as the disparity of income among adjacent neighborhoods.

From this project, we have learned that data science is most certainly an art form. We have faced numerous challenges and obstacles during the course of our analysis, and we have tried our best to approach the problem from a logical perspective. However, we were unable to build a statistical model that predicted the rate of crime in Baltimore City given the allotted time period of one semester. Although our efforts were unsuccessful, we hope to give insight to others cattempting to decipher this ongoing dilemma that occurs not only in Baltimore City, but all over the world.

Works Cited

John R. Hipp. 2007. “Income Inequality, Race, and Place: Does the Distribution of Race and Class within Neighborhoods Affect Crime Rates?” *Criminology* 45:3, 665–97.