Crime Prediction as a factor of Demographic, Economic and Psychographic variables

Manuel Garrido
Data Science Program
General Assembly
https://github.com/manugarri/Data_Science_Homework/tree/master/Data%20Science%20Final%20Project

0.Abstract

This is the Final Project for General Assembly's Data Science Program n 4.

The goal of this Research Project will be to get an accurate Risk Crime Index for each street block in the US, by using zip code level Demographic, Economic and Psycho graphic information as well as Crime Risk.

I will use the knowledge learned on the Data Science Program to retrieve and prepare a functional Dataset, and perform Machine Learning Modeling to find the most optimal Crime Risk Values.

1.Data Sources

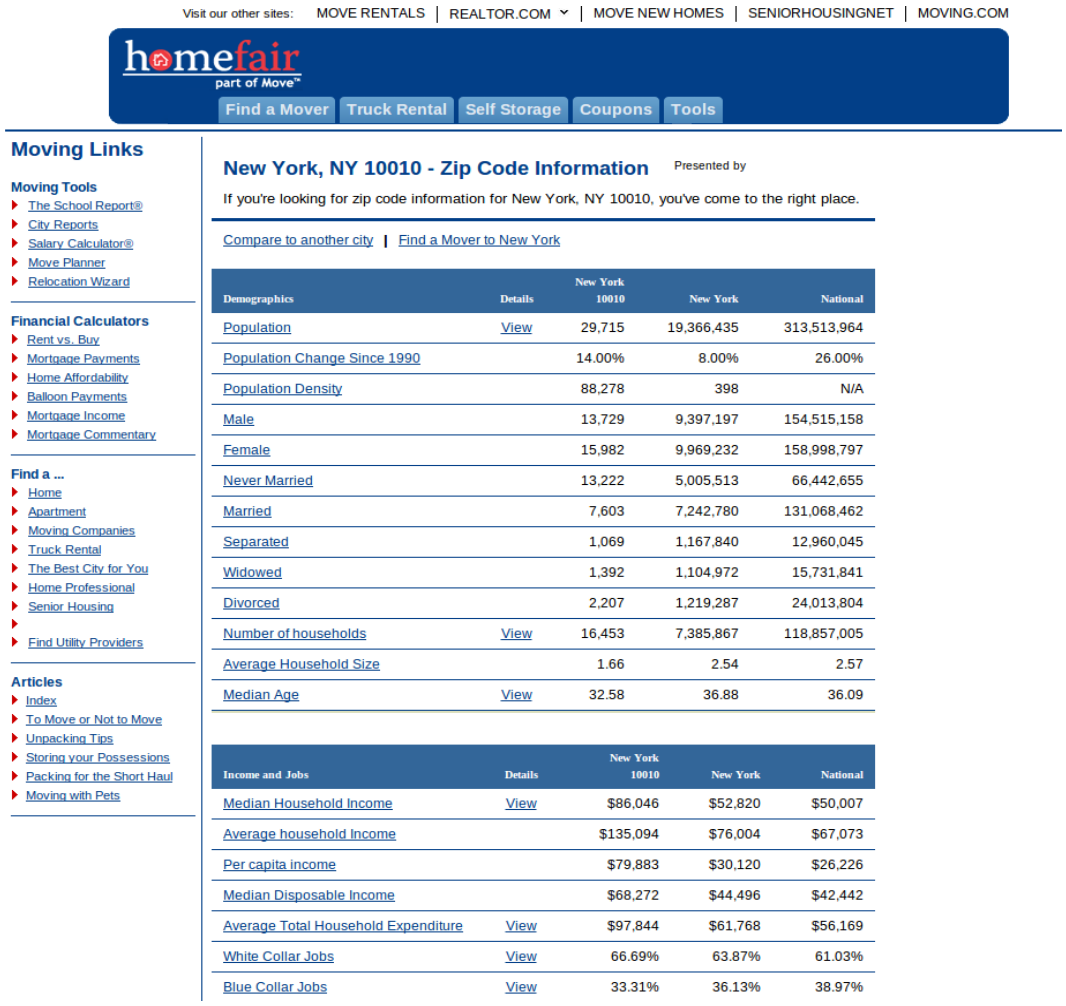
To achieve the goal described above, two different sources of information were used:

-The Real Estate Website HomeFair.com. A website that provides information to help users make better decisions regarding their moving needs.

Among the services provided, HomeFair.com provides a section named "City Reports". In this section, HomeFair.com users can search for information about a specific zip code. As per the website Glossary, the data is updated at least once a year.[1]. The available list of zip code level Variables is available in the file *'Glossary of Terms.txt'*.

1 http://www.HomeFair.com/real-estate/compare-cities/index.asp

As shown in Figure 1, HomeFair provides a variety of information for a zip code, as well as the average values for the US as well as the state the selected zip code is located. One set of variables provided include Crime information, more specifically Personal, Property and Total Crime Risk. The later will be the target variable used in this Project.



Visit our other sites: MOVE RENTALS | REALTOR.COM | MOVE NEW HOMES | SENIORHOUSINGNET | MOVING.COM

homefair part of Move™

Find a Mover | Truck Rental | Self Storage | Coupons | Tools

Moving Links

Moving Tools
- The School Report®
- City Reports
- Salary Calculator®
- Move Planner
- Relocation Wizard

Financial Calculators
- Rent vs. Buy
- Mortgage Payments
- Home Affordability
- Balloon Payments
- Mortgage Income
- Mortgage Commentary

Find a ...
- Home
- Apartment
- Moving Companies
- Truck Rental
- The Best City for You
- Home Professional
- Senior Housing
- Find Utility Providers

Articles
- Index
- To Move or Not to Move
- Unpacking Tips
- Storing your Possessions
- Packing for the Short Haul
- Moving with Pets

New York, NY 10010 - Zip Code Information Presented by

If you're looking for zip code information for New York, NY 10010, you've come to the right place.

Compare to another city | Find a Mover to New York

Demographics	Details	New York 10010	New York	National
Population	View	29,715	19,366,435	313,513,964
Population Change Since 1990		14.00%	8.00%	26.00%
Population Density		88,278	398	N/A
Male		13,729	9,397,197	154,515,158
Female		15,982	9,969,232	158,998,797
Never Married		13,222	5,005,513	66,442,655
Married		7,603	7,242,780	131,068,462
Separated		1,069	1,167,840	12,960,045
Widowed		1,392	1,104,972	15,731,841
Divorced		2,207	1,219,287	24,013,804
Number of households	View	16,453	7,385,867	118,857,005
Average Household Size		1.66	2.54	2.57
Median Age	View	32.58	36.88	36.09

Income and Jobs	Details	New York 10010	New York	National
Median Household Income	View	$86,046	$52,820	$50,007
Average household Income		$135,094	$76,004	$67,073
Per capita income		$79,883	$30,120	$26,226
Median Disposable Income		$68,272	$44,496	$42,442
Average Total Household Expenditure	View	$97,844	$61,768	$56,169
White Collar Jobs	View	66.69%	63.87%	61.03%
Blue Collar Jobs	View	33.31%	36.13%	38.97%

Figure 1: The GUI of the HomeFair.com City Report Zip Code Information

-Claritas Prime Location. Claritas, a Nielsen Company, is a Market Research Company focused on Market Segmentation. Their main product, PrimeLocation, is a mapping and demographic research tool
that provides easy to pull demographic information down to the block level.

I will use Prime Location to pull all the block level and zip code level information used to predict the block level Total Crime Risk.

2.Cleaning the data

The first step once the initial idea and goal has been set, the next step is to obtain the dataset used to create the model.

2.1 Obtaining zip code and Block level information from Claritas PrimeLocation

The PrimeLocation demographic data is easily pulled by using the 'Batch Data Pull' section on the PrimeLocation GUI. About 4,000 features were pulled and stored in two files, one, named 'Variables 2 zip code.txt' contain the information aggregated on a zip code level. The other file, named *'Variables 2 blocks2.txt'* containing the same information but on a Block level.

2.2 Obtaining zip code Total Crime Risk from HomeFair.com

Since the target variable is located on a website, the first step is to retrieve the data so it is usable. To do so, I built a 'Web parser' a program that automatically retrieves data from a Website.

The package used to perform the actual web parsing was the Ruby gem *'Nokogiri'*, an *HTML/XML* parser.

The way to obtain the data was very simple. For each zip code searched, HomeFair.com directs the user to a URL containing the zip code.

http://www.HomeFair.com/real-estate/city-profile/results.asp?Zip=10020

This URL directs to the General information for the zip code 10020

http://www.HomeFair.com/real-estate/city-profile/details.asp?format=popup&Zip=10020&SectionID=2&SectionName=Population#Population

This URL directs to the Population Section for the zip code 10020

HomeFair zip code Information containing one main table, and 4 expanded Sections that provide more detail about specific areas. These specific Sections can be accessed by clicking 'View' on the desired category.

The Ruby script can be found on the file *'hfparser.rb'*. It follows this sequence:
1. Load the required packages. *Nokogiri*, for the actual parsing, *sqlite*, to perform SQL queries to save the data and *open-uri*, part of the standard Ruby library that can open urls.
2. Create an empty *sqlite* database.
3. For each of the tables, create the columns in its specific table.
4. For each zip code, change a URL string and load the specific zip code Site. To find the list of

existing zip codes in the US, I used the zip codes available in the PrimeLocation zip code dataset.
5. Use *Nokogiri* to find the CSS selector *DataTbl and retrieve all the <td> and the values within.*
5. Use *.gsub* method to remove non numeric characters, such as '$', ','.
6. Use *sqlite* package to perform an *INSERT* query to insert the zip code data into the proper table.

Once the HomeFair data was accessible locally, I wrote a python script that use the sqilite3 python package to connect to the HomeFair database, and perform 4 queries for each zip code (one for each Section). Then I used the *pandas* package to create a data frame and export to csv.

Some extra cleanup was done in R, to remove the zip codes where some of the values weren't available ('N/A'). The cleaned file was then saved as 'HomeFair.csv'.

An issue I found is that there were about 28175 zips available in HomeFair.com vs 30241 in PrimeLocation. HomeFair.com doesn't provide information for Alaska.

2.3 Merging the PrimeLocation data

Once I had functional datasets for zip codes, one including Total Crime Risk, and the other including all demographic information, the next step was to merge both datasets.

To do so, I used the Python package '*pandas*' to perform an inner merge function.

```
import pandas as pd
import numpy as np


#=============================================================================
#CREATING THE TRAINING SET
data=pd.read_table('Variables 2 zipcode.txt')
#remove index column created by PrimeLocation
data=data.drop(data.columns[-1],axis=1)
#We will only use the zipcode and the Total Crime Risk from Homefair.com
crime=pd.read_csv('homefair.csv',usecols=[2,45])
#Merge both datasets
train = pd.merge(data, crime, left_on='GEOGRAPHY_ID', right_on='zip', how= 'inner')
#Export the new datasets to a csv file
train.to_csv('train.csv',encoding='utf-8')
```

Figure 2: Python code used to merge HomeFair and PrimeLocation datasets

2.4. Target Variable: Total Crime Risk

As per HomeFair Glossary, the Total Crime Risk is *"a score that represents the combined risks of rape, murder, assault, robbery, burglary, larceny and vehicle theft compared to the national average*

of 100. A score of 200 indicates twice the national average total crime risk, while 50 indicates half the national risk. The different types of crime are given equal weight in this score, so murder, for example, does not count more than vehicle theft. Scores are based on demographic and geographic analyses of crime over seven years."
Sources of this information include the Federal Bureau of Investigation, local police departments and municipalities for crime information, and is updated at least once a year.

Crime	Details	New York 10010	New York	National
Total Crime Risk		110	67	100
Personal Crime Risk	View	202	96	100
Property Crime Risk	View	63	62	100

3. Modeling

The first issue was to decide which data selection from the Primelocation Dataset to use. With 4,000 available features, probably most of them won't have any impact in how dangerous an area is.

To find the best data selection, different data mixes were selected and performed exploratory linear models with.

The final dataset features includes demographic (i.e. Population), economic (i.e., average Household Income) and psychographic features (e.g. PRIZM[2] Segments).

The list of existing features in the final datasets can be found in the file *'Variables 2 blocks2_layout.txt'*.

After having one functional training set, I will use sklearn, a python set of tools that provide Machine Learning tools and method for data Analysis.

The code used to do the modeling is available on the file 'model.py'

3 different approaches were taken.

3.1. Linear Regression

I used the available Linear Regression models in *sklearn*, and the results weren't satisfactory.

The chosen measure of accuracy was Mean Absolute Error (MAE).

2 http://www.claritas.com/MyBestSegments/Default.jsp?ID=70

Here are the results of the top performer linear models:

LinearRegression()	MAE:38.0
linear_model.Ridge()	MAE: 36.2
linear_model.LassoLars()	MAE: 48.6
linear_model.Lasso(alpha=50)	MAE: 35.9

One interesting fact of linear models is that they provide weights of the most relevant features. These coefficients (commonly known as *betas*) give an idea of how relevant the different features are to make a zip code safer (negative betas) and riskier (positive betas).

Coefficients for the most accurate model can be observed in the file 'Coefficients.csv'.

3.2.Ensemble Regression techniques

After performing linear regression models, I used Ensemble Regressors to analyze if the accuracy improved.

Even though ensemble techniques require more processing time, they didn't provide an increase in the accuracy in the model.

GradientBoostingRegressor()	MAE: 40.4
RandomForestRegressor()	MAE: 41.9

3.3 Classification + Regression Model

The final approach take to predict the zip code level Total Crime Risk consisted on a 2 - tier model:

1.Use a classifier to classify the zip codes into highcrime and lowcrime zip codes. Highcrime zip codes will be those whose Total Crime Risk value is higher than the median value of all zip codes. That way the split between low and high crime zip codes would be of 50/50.

2.Use different regressors to predict the Total Crime Risk for the highcrime and the lowcrime zip codes separately.

Since the *sklearn* available *cross_val_score* doesn't work with multiple step models, I wrote my own, selecting a random sample (75% of the total training set).

Here are the results of the different classifiers/regressions used:

regressor_high = RandomForestRegressor() regressor_low = linear_model.Lasso(alpha=50) classifier=GradientBoostingClassifier()	Time: 110s MAE:18
regressor_high = RandomForestRegressor() regressor_low = RandomForestRegressor() classifier=GradientBoostingClassifier()	Time: 143.3s MAE: 12.1
regressor_high = linear_model.Lasso(alpha=50) regressor_low = RandomForestRegressor() classifier=RandomForestClassifier()	Time: 45s MAE: 23.2
regressor_high = RandomForestRegressor() regressor_low = RandomForestRegressor() classifier=RandomForestClassifier()	**Time: 77.1s MAE:9.72**

Figure 4: Summary of Classification + Regression Models

The best performing model in terms of MAE uses a RandomForestClassifier as an initial classifier and two different RandomForestRegressors to predict the high_crime and low_crime Total Crime Risk values.

In terms of accuracy, the model performs on a MAE of 9.72 (after 10 fold cross validation). Compared to the average Total Crime Risk Index, it translates to an error of less than 15%. Compared to the total range of Total Crime Risk Index values (0-1636), it translates to an error of 0.6%.

Once the best performing model is selected, the final step was to load the file containing the same set of features but for the 217,000 blocks in the United States.

Then using the zip codes as a fitting model, I used the fit, predict to get the predicted Block Level Total Crime Risk Values, and write the Predicted Crime Risk, Block ID and Centroid Latitude/Longitude values into a csv.

4. Crime Visualization

The final step in the project was to plot a point map using the predicted crime rate as a color/size scale, with green being the color for the "safer" areas (lower crime) and 'riskier' areas (high Total Crime Risk). The points would be geolocated using the Block Centroid latitude and longitude.

To do so, I used R and its packages *ggmap* and *ggplot*. Code is available on the file *'plot_map'*.

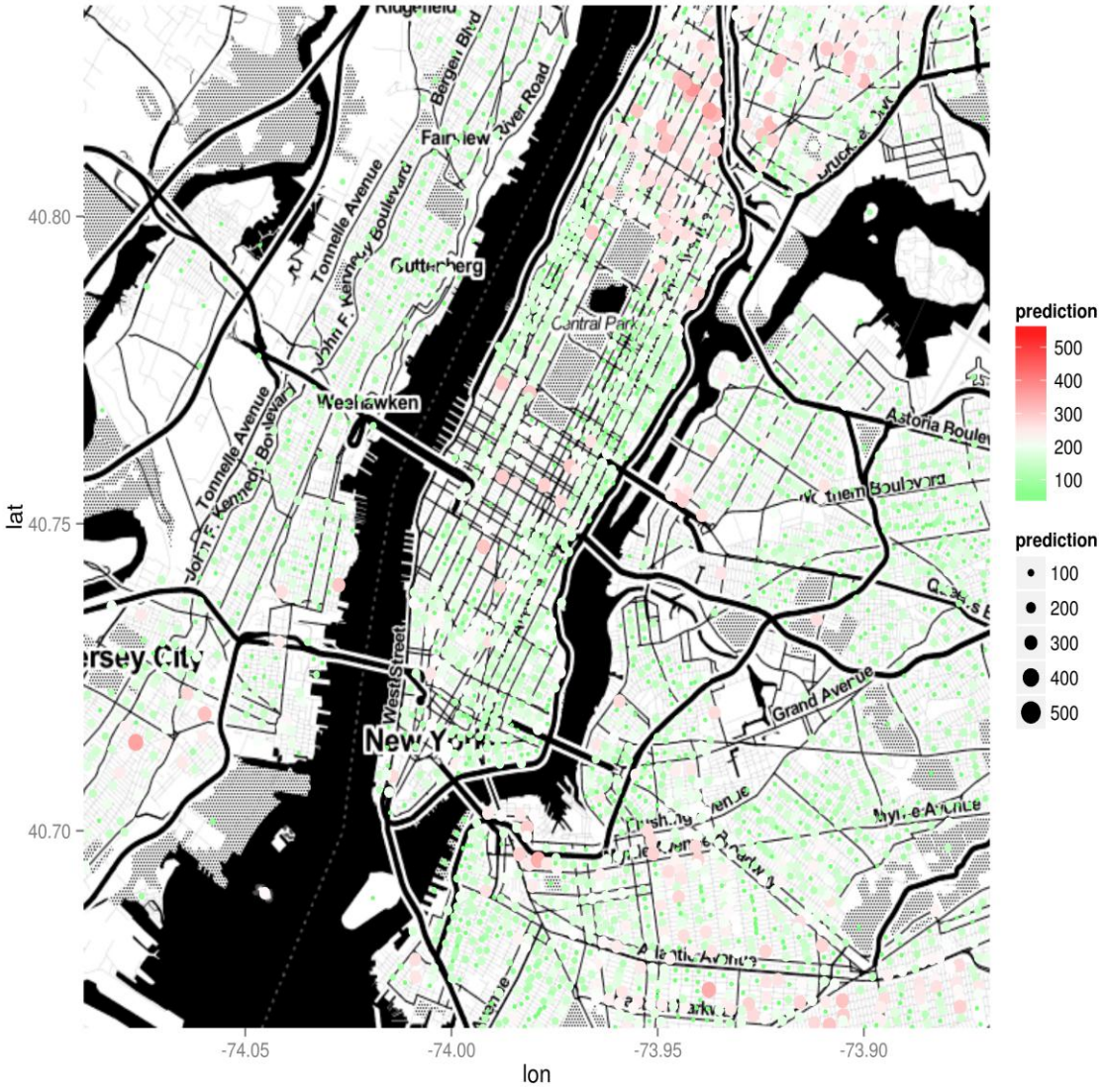


Figure 5: Visualization of the Predicted Block Level Total Crime Risk for the City of New York

5. Further Steps

The next step to improve the accuracy of the model, would be to use the logarithm function of the Total Crime Risk as the target variable. As shown on figure 6, the natural Zip code Total Crime Risk distribution is heavily skewed to the left. Thus, using the log() function would normalize the data.

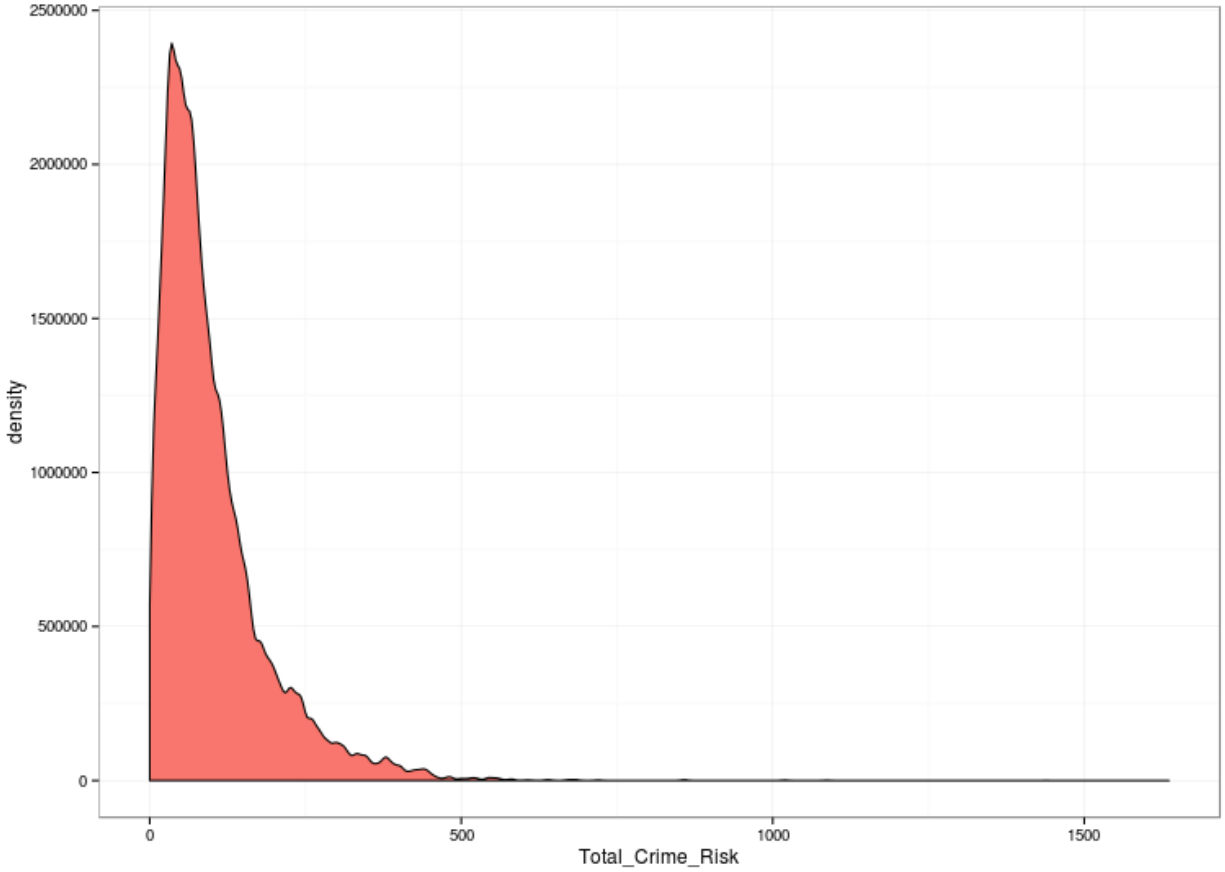


Figure 6: Density Plot of the zip code Total Crime Risk index

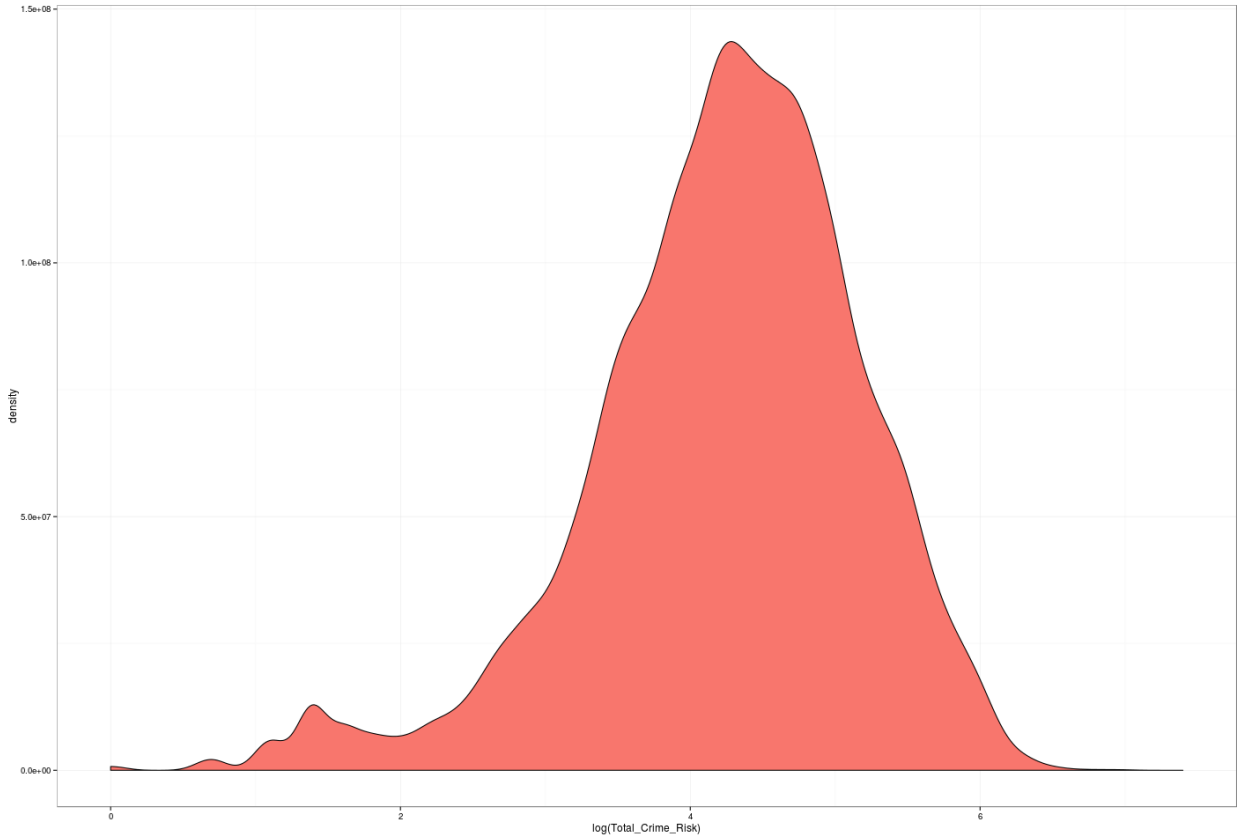


Figure 7: Density Plot of the zip code Log(Total Crime Risk index)