CS6220

Data Mining Techniques

Final Project Report

Impact of Crime on Housing Rates and forecasting of Crime Data

By:

Mitresh Pandya
Tirth Patel

1. Abstract

The main problem we are trying to solve here is to find the correlation between the crime rate in Boston with the housing prices in those areas. The focus was to find out if there is a relation between the crime rates and the housing prices, that is, is it a case where the higher crimes lowers the housing prices and vice versa. While this is the primary objective, our focus was also on performing various analysis on Boston crime data to try and identify interesting facts out of it. Some of the goals were to find out the streets which are highly affected by crimes, or the least safe ones, then identify the safest streets, see the crime distribution among the districts of Boston, the proportion of most violent crimes, trend of crimes over the years etc. Based on these studies, apply some sort of prediction model and train the model to predict the probability of crime happening in certain areas to use that study for further use. The solution to this was achieved by gathering correct data, enough to serve the purpose, was gathered. Grouping and various preprocessing techniques were applied on those datasets to connect the datasets to derive the outcome. The datasets that were used had to be authentic and trustworthy. So, the crime dataset is something that is made public by the government of Boston is used. Since there was a need to identify the zip codes of the crimes that happen to connect it with the housing data, the mapquest API was used to get zip codes from latitude and longitude. For the further use of this dataset, we wanted to use some sort of prediction model. Since this was a dataset which had values and frequency of something in specific time span, the problem derived to be a time series analysis problem. Hence, use of ARIMA model can be seen in the project as the prediction model. The interesting finding through the analysis came up that during the months of summer, crime rates go up significantly compared to the rest of the year. The reasons to that could be anything, be it weather, tourists during those times, the holidays in high schools causing kids to wander around and commit crimes, anything for that matter. Use of prediction model was important to figure out the safer zones in Boston in upcoming times. Since the entire goal of this project was to find the correlation between the crimes and housing prices, we wanted to identify which could be the safer options in future to property buyers to go for. The initial thoughts were that crime is going to be a decisive factor when it comes to housing prices. That is, higher crimes will cause the lower prices of houses in those areas. The results supported our initial thoughts and it does come up that where the crimes are high the house prices are low and vice versa. There were some exceptions, which were mainly the University areas, where because of the high volume of students, crimes were common since they are the easy targets and crime rates were higher, but the house prices are higher as well because of the same reason. That is, there are so many students who live on or around campus to be close by. Some interesting findings also came up that no matter what day of the week is the crime frequency is not much different. That means the

day of the week doesn't affect the crime rate that much. For Boston, the most common of part one crimes was larceny. Thus, the final outcome and interesting analysis of these datasets supports our initial claim of crime rates affecting housing prices and also bring out some interesting facts about crimes.

2. Introduction

The problem statement can be summarized as "Identifying how crime in the area affects the real estate prices". In other words, it is a problem of correlating the crimes with the housing prices.

Before correlating, we need to analyze the crime data to recognize patterns, predict the outcomes by training a model, etc. This plays a vital role in analyzing the crime data for Boston.

According to us, this problem is vital to solving because there are so many cases in which naive buyers or someone who is unfamiliar with the area ends up buying a house which is not in the safe area. This could be very dangerous as well as prone to a monetary loss.

So focusing only on Boston, it is growing very fast and there are so many opportunities in all fields which attract lots of young professionals. Plus, with the high cost of real estate near the city could easily throw someone off to buy a cheaper house in the unsafe area. If there is a system in place for something like this, then it could prevent such issues.

This is merely a very simple example, but if steps are not taken to prevent the crime from happening, then this could spread and make the entire place unsafe to be in. There are several datasets available for our purpose. For example, https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system is the dataset by the government on the crime reports. There are several datasets for property and real estate based on areas too. This could help, fulfill our evaluation needs.

Based on the literature review we study some of the papers that analyzed housing prices and some that related the crime to housing. For example, Ihlanfeldt, K.et.al. in his paper[5] talks about how different types of crime impact housing prices. Which got us the intuition to focus more on the part one type of crime as categorized by Boston Police dataset.

For analysis of Machine Learning techniques that we would want to implement for this task we researched and learned about the work done by Baldominos, A. et.al. in their paper[4]. Now, this

was more about predicting the house prices while we thought about predicting a regression type model for the crime data.

Further, to support the task we were willing to work on we found a paper by Pope, D. [7] that had similar research area but a bit different implementation to what we had in mind.

3. Methodology

3.1. Datasets used:

Datasets that we have used for this project are as follows.

1) Boston Crime Dataset (https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system)
2) Boston Prices Housing Dataset (https://data.boston.gov/dataset/property-assessment)

The datasets we have used here are from a reliable source. As per the crime dataset, it comes from the Boston government itself and it is updated daily.

3.2. Crime Dataset

Our main aim was to somehow perform some analysis on the data that will identify the patterns in the crime that has happened already and extract as much information as possible from the dataset that will uncover some unknown or interesting facts. Once those analyses are in place, the next step would be to predict the crimes based on the places, that is, predict where the crime rates are going to go down or up. Based on that, the choice of buying a property in an area can easily be made.

Our solution to achieve the outcome was straight forward. Preprocess the datasets and generate the zip code of the locations for crime dataset. Perform the join of two datasets based on those zip codes and create an outcome comparing the values of houses in those zip codes with the crimes there. On top of that, performing time series analysis to implement a prediction model

on crime dataset to predict and find out which areas are going to be more/less affected by crimes in coming time.

3.2.1. Preprocessing:

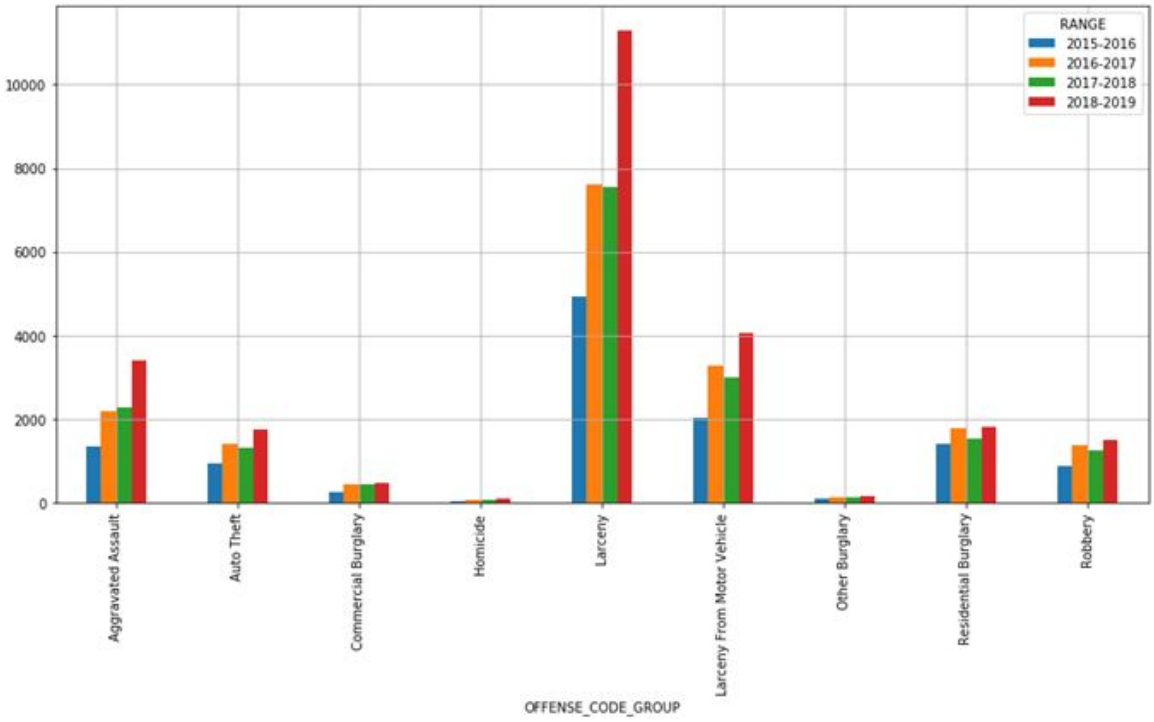
As far as the solution goes for the problem statement, we were planning to first identify a link which can work as a bridge for these two datasets in order for us to perform the studies. While studying the dataset, there was no present feature which could connect them. So, we decided that since we have exact locations of crime happening in the dataset, we can identify the zip code from that and zip codes were the part of the housing dataset. Thus, we found out the connecting link between these two. After the connecting link was found, it was time for preprocessing the data to generate the zip codes. Looking at the different options, we came upon the API called mapquest, which uses reverse geolocation to identify the zip codes from the latitude and longitude of the location. Once that was done, we moved on to the next steps of preprocessing the dataset. Since we wanted to perform the time series analysis on the crime dataset, and the dataset of crime was from June 2015 to June 2019, we decided to add a range column in the dataset that will part the data in the span of June 2015-June 2016 and so on. This way we can have uniform distribution of data in the ranges. Once the data was distributed in the range, the next thing was to perform the time series analysis. We have researched and decided to use ARIMA model for that purpose. Further steps of preprocessing the crime dataset included formatting of the dates of occurrences. The dates were in a different format than what it needed to be for the model to be imposed, so using python's DateTime library, we converted all the dates to the desired format. Before all of these steps were taken, the removal of data with no significance was performed. Since there were many occurrences which had no exact location or time, it created the noise in the data. This could lead to incorrect analysis and hence those were filtered first.

3.2.2. Analysis:

While performing the analysis, we mainly focused on part one type of crimes. These are the most violent crimes which are considered to be more dangerous than others. These are the most significant part of the entire analysis when it comes to connect it to the housing dataset and perform the pricing analysis there. The various analysis which is performed includes statistics of part one types crime over the years, crime trend over the years, distribution of part one crime among itself, safest and unsafest streets, shooting involved incidents and crime according to the districts of Boston.

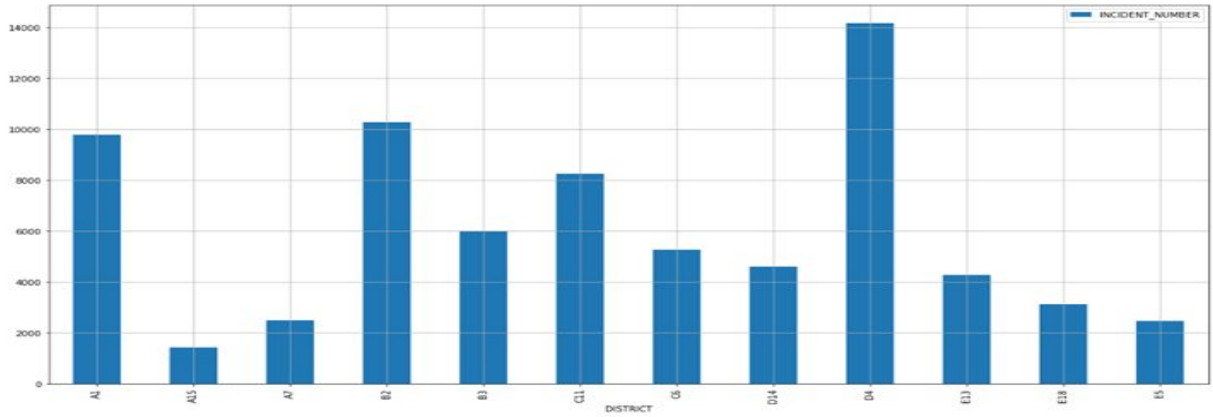
Following are the screenshots and some justification for it.

Part one crime statistics.



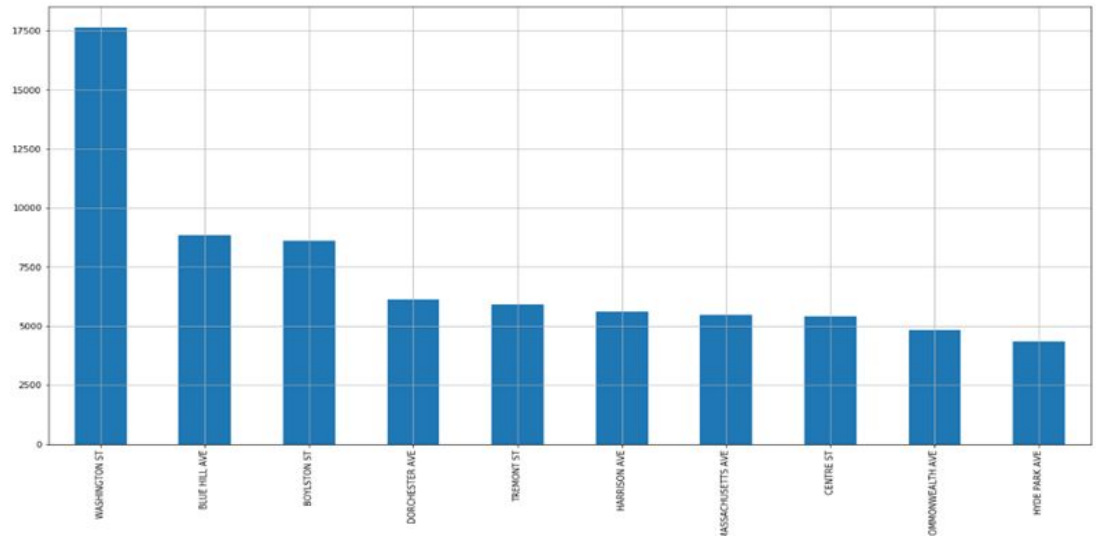
From the above screenshot, it is obvious that the part one crimes are rising as the time goes by. From 2015 to 2019, the crime rates are up for each type of crime.

Following is the graph of total crimes according to the districts.



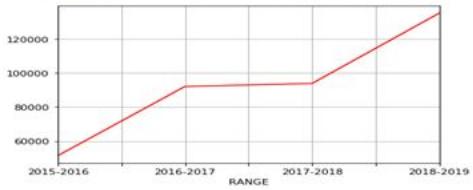
As you can see, the crimes in district D4 are on the top, which is the south end of Boston.

Top 10 unsafe streets, with Washington street on the top.

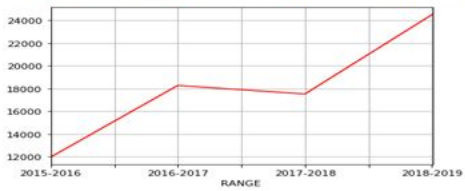


Crime trends over the years.

```
In [24]:  # Overall crime rate over the years from 2015 to 2019.
          crime_data.groupby(['RANGE']).count()['INCIDENT_NUMBER'].plot.line(c='r', grid='true')
Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x1d6bfa6e5f8>
```

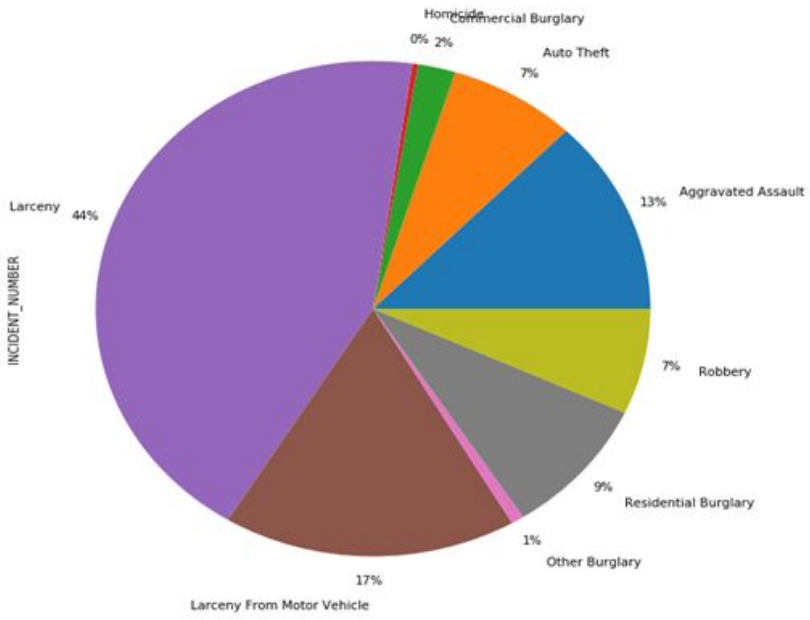



```
In [25]:  # Only part one crime rate over the years from 2015 to 2019.
          part_one_crime.groupby(['RANGE']).count()['INCIDENT_NUMBER'].plot.line(c='r', grid='true')
Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0x1d6bf721860>
```

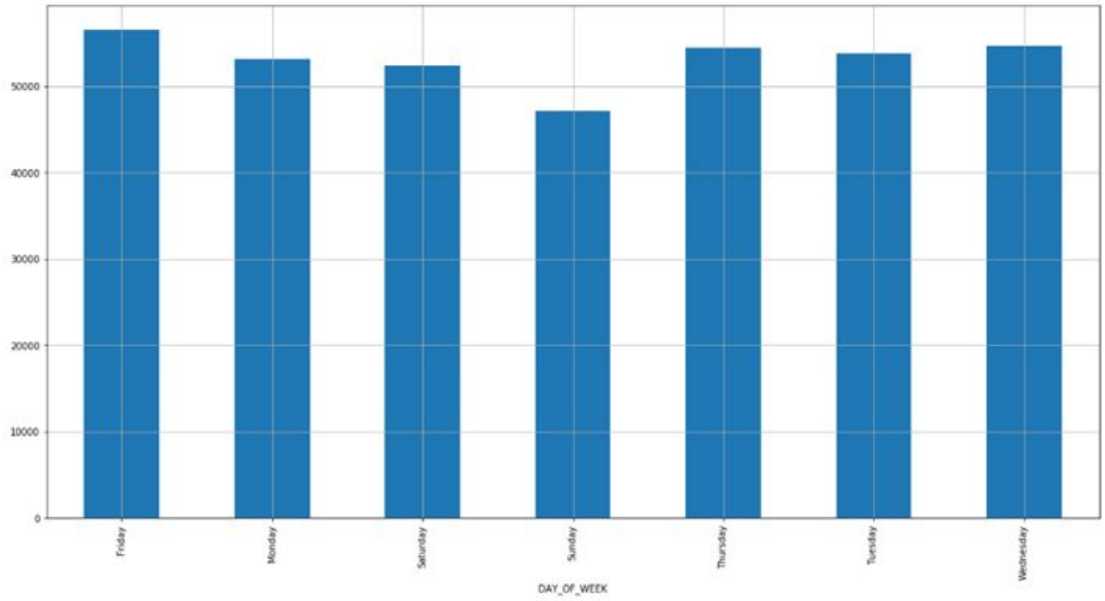


As we can see from the graph itself, both the overall crimes and part one crime are rising since 2015-2016, only being little stable during 2016-2017 and 2017-2018 and then shooting up again in 2018-2019.

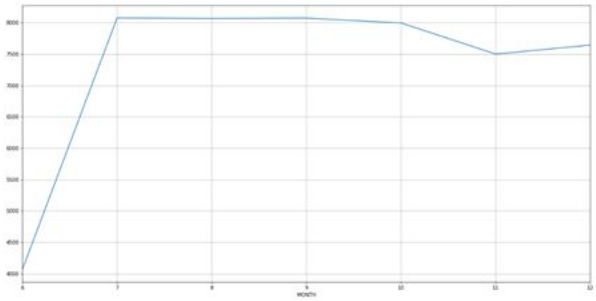
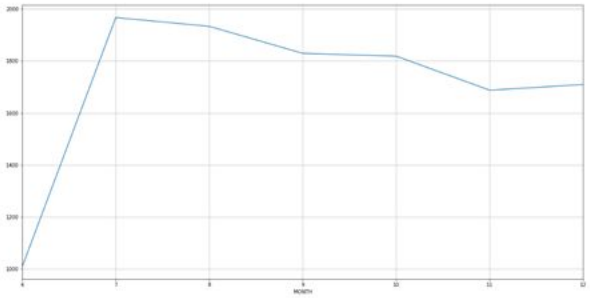
Following is the distribution of the part one crimes among itself. Which shows that Larceny is the commonly occurring offense in this group.



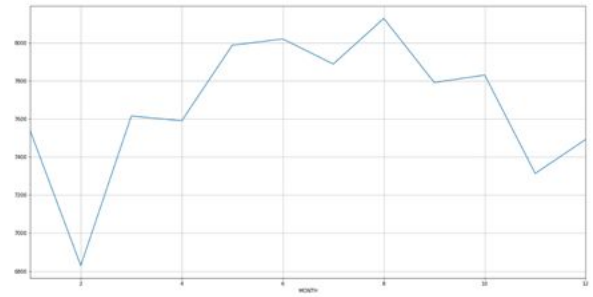
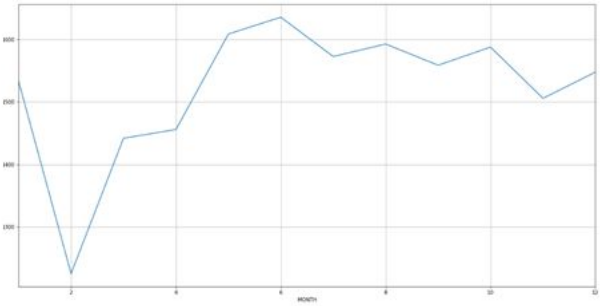
The following analysis shows that the day of the week doesn't matter when it comes to crimes.



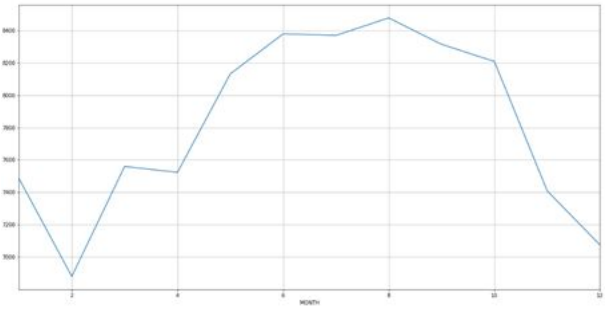
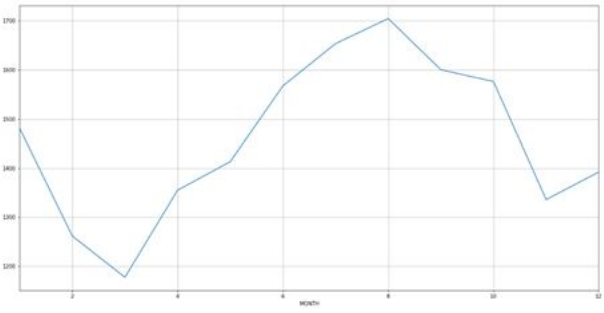
Some screenshots that show that the crimes during summertime top the other months in a year.

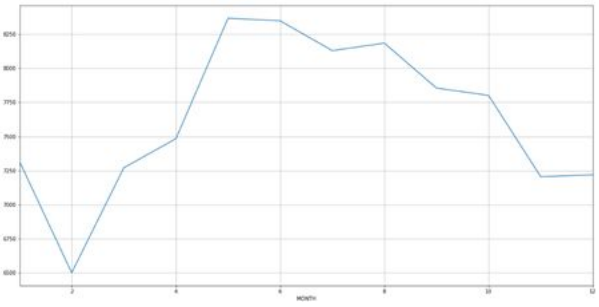
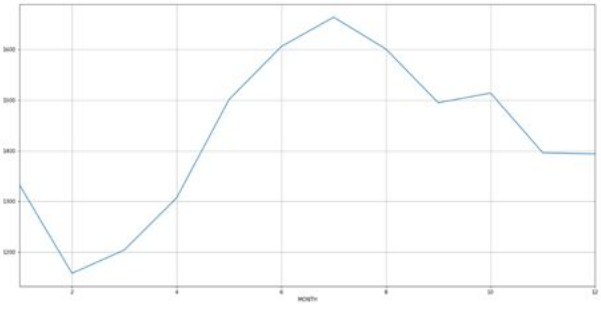
2015

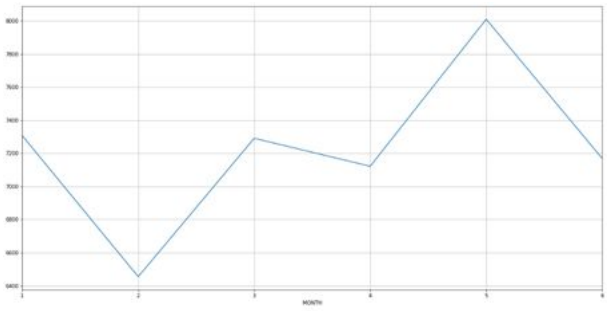
2016

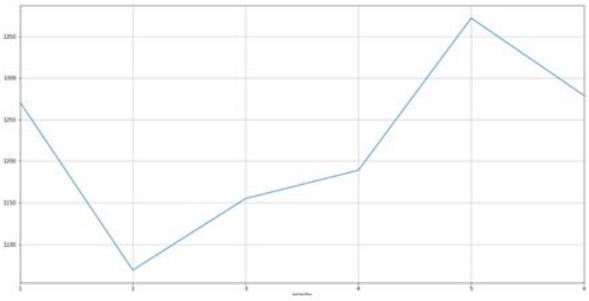



2017

2018





2019

3.3. Housing Dataset analysis:

3.3.1. Pre-processing:

So, for pre-processing, we first choose a subset of data to be extracted from the full data, got the zip code and average total of the house. Further, we filtered only the residential housing for getting the correlation for housing data.

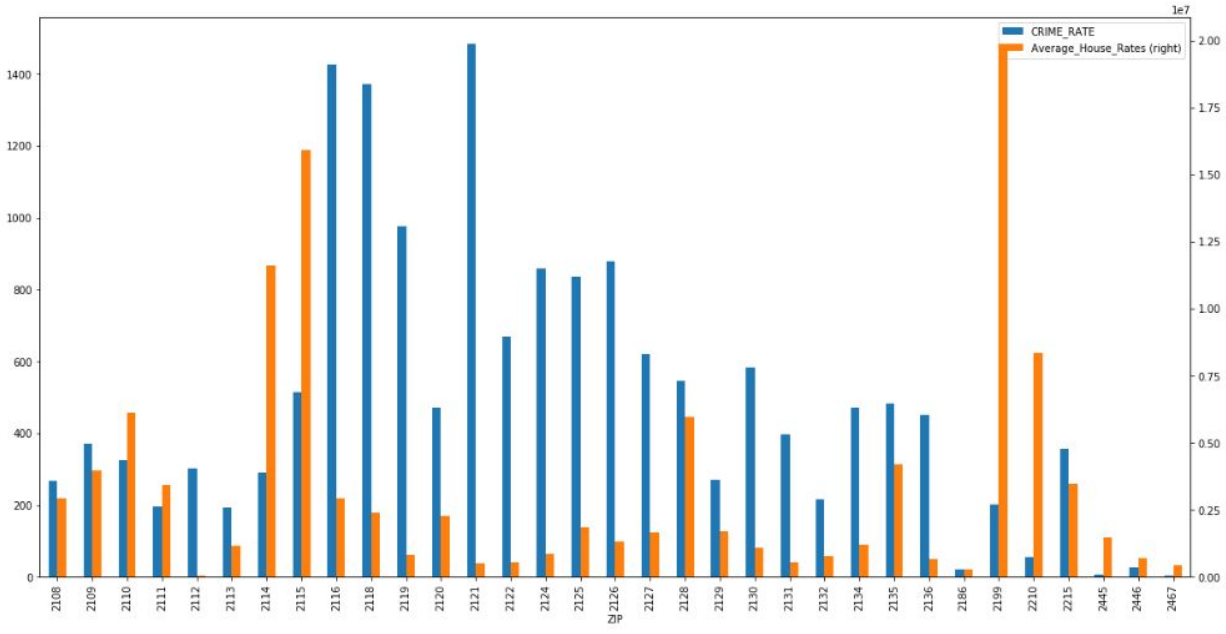
3.3.2. Analysis:

So firstly, we analyzed the dataset by getting the prices of house in a particular region and getting the ones with the highest prices on a particular street in a particular region.

2108	HAMILTON PL	3.177460e+05
	MT VERNON ST	6.554000e+05
	WILLOW ST	7.603886e+05
	RIVER ST	1.000343e+06
	CHESTNUT ST	1.426300e+06
	RIVER STREET PL	1.509900e+06
	CEDAR LANE WY	1.560694e+06
	BEACON ST	1.563050e+06
	SPRUCE CT	1.771525e+06
	MT VERNON ST	1.786410e+06
2109	ARCHWAY ST	5.079000e+05
	HANOVER ST	5.095909e+05
	SALUTATION ST	5.580667e+05
	CLARK ST	5.809655e+05
	FOSTER ST	5.868107e+05
	HOLDEN CT	6.563250e+05
	FAIRFIELD PL	6.723500e+05
	HANOVER AV	7.411850e+05
	FLEET ST	7.644445e+05
	FULTON ST	8.587540e+05
2110	HAWLEY ST	7.510000e+04
	TREMONT ST	5.628000e+05
	PEARL ST	7.638600e+05
	LEWIS WH	8.214000e+05
	E INDIA RO	8.306241e+05
	BROAD ST	8.829762e+05
	DEVONSHIRE ST	9.327500e+05
	SUMMER ST	1.084841e+06
	ATLANTIC AV	1.176731e+06
		...
2210	CHANNEL CENTER ST	1.159386e+06
	ATLANTIC AV	2.318621e+06
	LIBERTY DR	3.842513e+06
	CONGRESS ST	3.999097e+06
2215	BACK ST	6.610000e+04
	FENWAY	2.417500e+05
	PARK DR	3.166750e+05

So, from this, we can co-relate to the ones we got from crime data.

Further, on merging both crime and housing dataset we got to this result:



We can see that housing prices on the right y-axis are lesser where there are higher comparative crime rates we observe on the left y-axis. So, there definitely seems to be a

correlation between them. Crime might be one of the significant factors that would affect housing rates.

So, now to exploit this rule we wanted to try a time series model-based prediction. This would impact the way people invest in real-estate, rents of the houses in a particular region, it also has the potential to help police with analyzing the prediction trends and taking further actions.

3.4. Time series analysis:

From the analysis of the crime dataset, we found various trends in the occurrence of crime. There are seasons or even sometimes months that have higher crime rates, while in others there are not. Keeping these trends in mind, we thought to implement a time series model for the task.

3.1.1. Data Analysis and Stationarity:

For any time series model, it is important that the data is stationary and there isn't much variance or covariance. Observing the time series data of the crime, we observed that it was a bit stationary in nature while still having a good amount of variance or co-variance.

3.1.2. ARIMA model:

Establishing the stationarity, the next thing to implement was to make an ARIMA Model. So what is the ARIMA model? ARMA models are commonly used in time series modeling. In the ARMA model, AR stands for auto-regression and MA stands for moving average. I in ARIMA stands for Integrated, which induces a difference term.

3.1.2.1. AR Model: Auto-regressor model

Let's understanding AR models using the case below:

The current GDP of a country say x(t) is dependent on the last year's GDP i.e. x(t – 1). The hypothesis being that the total cost of production of products & services in a country in a fiscal year (known as GDP) is dependent on the set up of manufacturing plants / services in the previous year and the newly set up industries / plants / services in the current year. But the primary component of the GDP is the former one.

Hence, we can formally write the equation of GDP as:

$$x(t) = \text{alpha} * x(t - 1) + \text{error}(t)$$

This equation is known as *AR(1) formulation*. The numeral one (1) denotes that the next instance is solely dependent on the previous instance. The alpha is a coefficient which we seek so as to minimize the error function. Notice that x(t- 1) is indeed linked to x(t-2) in the same fashion. Hence, any shock to x(t) will gradually fade off in future.

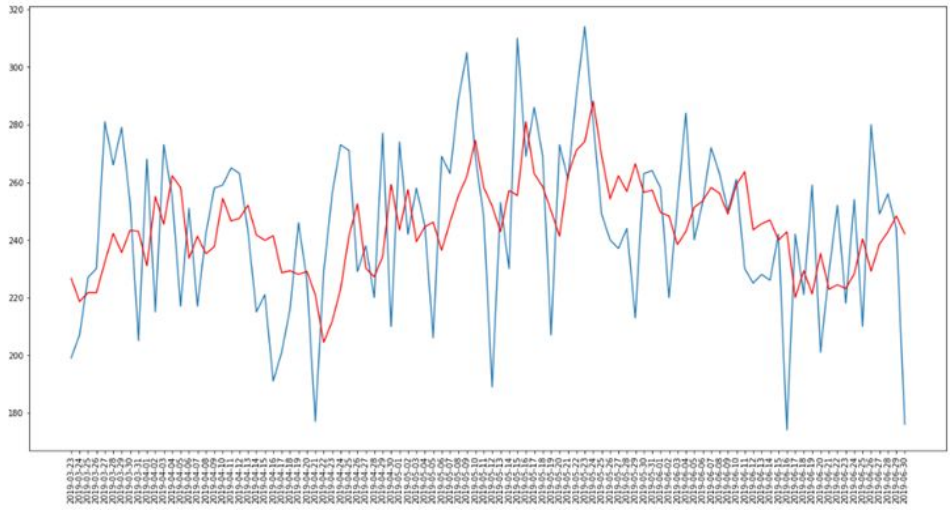
3.1.2.2. MA Model: Moving Average Model

Let's take another case to understand Moving average time series model.

A manufacturer produces a certain type of bag, which was readily available in the market. Being a competitive market, the sale of the bag stood at zero for many days. So, one day he did some experiment with the design and produced a different type of bag. This type of bag was not available anywhere in the market. Thus, he was able to sell the entire stock of 1000 bags (lets call this as x(t)). The demand got so high that the bag ran out of stock. As a result, some 100 odd customers couldn't purchase this bag. Lets call this gap as the error at that time point. With time, the bag had lost its woo factor. But still few customers were left who went empty handed the previous day. Following is a simple formulation to depict the scenario :

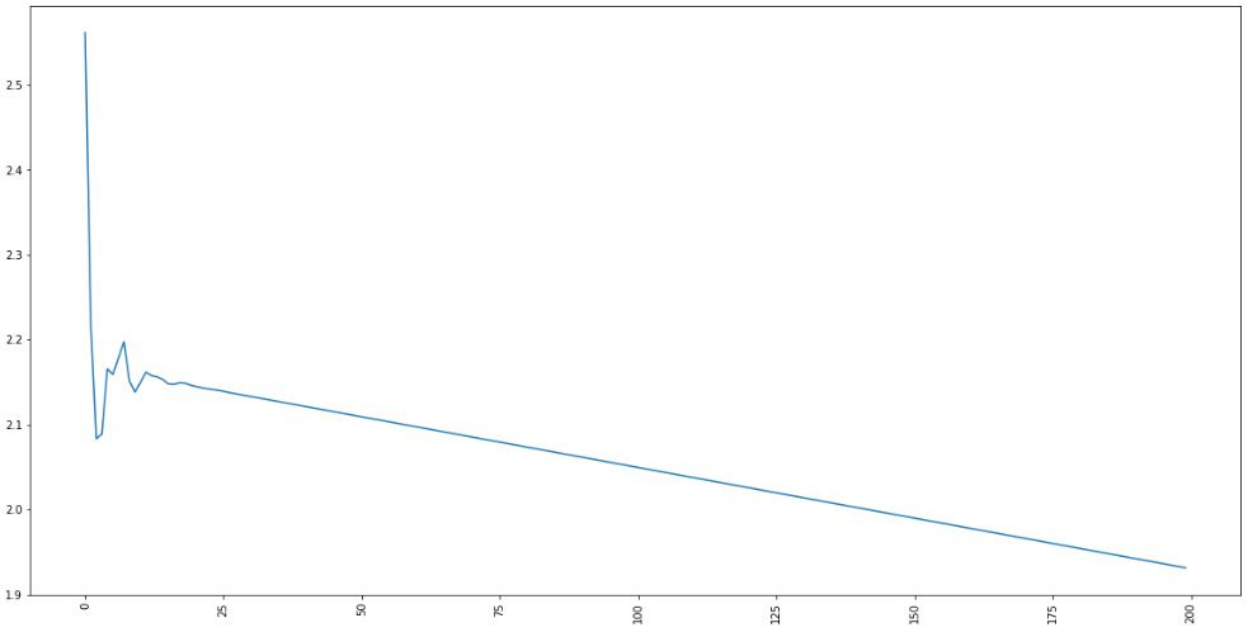
$$x(t) = \text{beta} * \text{error}(t-1) + \text{error}(t)$$

So, we created a model using the ARIMA model and tested it on the data giving us following graph:



Based on the graph above we can see that the graph doesn't overfit much and follows the trend of crime data as per the seasonality of the crime.

Further, as a case study we wanted to predict the crime rates in a region of Dorchester near University of Massachusetts, Boston. We used the model trained above to predict the crime rate over the period not included in the training data and we got a graph looking something like:



So, seems like a downward trend, which might indicate a decrease in crime rate and potential increase in housing prices, which has potential impact on housing rents, real-estate rates and police activity in the region.

4. Code

4.1. Crime Dataset:

4.1.1. Preprocessing:

Removing unnecessary data. Following is the screenshot for that, which shows the code snippet of how the noisy data rows of data are removed from the dataset.

```
crime_data = crime_data.dropna(how="any")

indexNames = crime_data[ crime_data['DISTRICT'] == '0' ].index
# Delete these row indexes from dataFrame
crime_data.drop(indexNames , inplace=True)

indexNames2 = crime_data[ crime_data['DISTRICT'] == 0 ].index
# Delete these row indexes from dataFrame
crime_data.drop(indexNames2 , inplace=True)
```

Here we have utilized python utilities and functions to achieve that, where the district is 0, which means unknown.
Following is the code snippet which shows the occurrence date format conversion using python's DateTime library.

```
def date_time_conversion(dt):
    datetime_object = datetime.strptime(dt, '%m/%d/%Y').strftime('%Y-%m-%d')
    return datetime_object
crime_data['OCCURRED_ON_DATE'] = crime_data.OCCURRED_ON_DATE.apply(lambda x: date_time_conversion(x))
```

Following is the snippet which shows the use of mapquest API to achieve the zip codes from the locations.

```
location_to_zip = dict()
```

```
def find_zip_codes(location):
    location = location.replace('(','')
    location = location.replace(')','')
    location = location.replace(' ','')
    # api-endpoint
    URL = "http://open.mapquestapi.com/geocoding/v1/reverse"

    # defining a params dict for the parameters to be sent to the API
    PARAMS = {'key':"0UivSSWlMrR71m5haZ8pJJxWAW4vE3KL",'location':location}

    if location in location_to_zip:
        return location_to_zip[location]

    # sending get request and saving the response as response object
    r = requests.get(url = URL, params = PARAMS)

    # extracting data in json format
    data = r.json()
    location_to_zip[location] = data['results'][0]['locations'][0]['postalCode']
    return data['results'][0]['locations'][0]['postalCode']

crime_data['ZIP'] = crime_data.Location.apply(lambda x: find_zip_codes(x))
```

Here we are storing the location and the corresponding zip code in a dictionary and if upcoming location is not there in the dict, then only we will make an API call to get the zip code, since there were around 19000 unique locations and API would allow 15000 calls only, so we utilized two different keys to achieve it.

Following is the code snippet showing how the data was sorted into the range of dates.

```
# Added a new column named range for easier further use.

crime_data['RANGE'] = '0'
crime_data.loc[((crime_data.YEAR == 2015) & (crime_data.MONTH >= 1)) | ((crime_data.YEAR == 2016) & (crime
_data.MONTH <= 6)),'RANGE'] = '2015-2016'
crime_data.loc[((crime_data.YEAR == 2016) & (crime_data.MONTH >= 1)) | ((crime_data.YEAR == 2017) & (crime
_data.MONTH <= 6)),'RANGE'] = '2016-2017'
crime_data.loc[((crime_data.YEAR == 2017) & (crime_data.MONTH >= 1)) | ((crime_data.YEAR == 2018) & (crime
_data.MONTH <= 6)),'RANGE'] = '2017-2018'
crime_data.loc[((crime_data.YEAR == 2018) & (crime_data.MONTH >= 1)) | ((crime_data.YEAR == 2019) & (crime
_data.MONTH <= 6)),'RANGE'] = '2018-2019'
```

Here the pandas inbuilt loc is used to filter data and apply the value of range to it.

The coding part itself was not difficult, the most difficult part was what was needed, how to preprocess the data, how and what to use to group the data in such a way that will get the helpful information out of it. Utilizing it later on for various purposes was also a challenge. Then came the code that was used to plot data to make it more interactively readable.

4.2. Housing Dataset:

4.2.1. Pre-processing:

For preprocessing of the data, we used much of the Pandas functionality to reduce the dimensions of the dataset.

```
def clean_zip_code(zip_code):

    if type(zip_code) is float:
        zip_code = int(zip_code)
        zip_code = str(zip_code)
    else:
        if len(zip_code) > 0:
            zip_code = zip_code[:-1]
            zip_code = zip_code[1:]

    return zip_code

def pre_process_data(housing_data_all_cols):
    housing_data = housing_data_all_cols[['ST_NAME',
                                          'ST_NAME_SUF',
                                          'ZIPCODE',
                                          'LU',
                                          'AV_TOTAL',
                                          'YR_BUILT',
                                          'LIVING_AREA']]
    housing_data['ST_NAME'] = housing_data['ST_NAME'] + " " + housing_data['ST_NAME_SUF']
    housing_data = housing_data.drop(['ST_NAME_SUF'], axis=1)
    housing_data = housing_data.dropna(how='any')
    housing_data = housing_data[(housing_data['LU']=='R1')
                                | (housing_data['LU']=='R2')
                                | (housing_data['LU']=='R3')
                                | (housing_data['LU']=='R4')
                                | (housing_data['LU']=='RC')
                                | (housing_data['LU']=='CD')
                                | (housing_data['LU']=='RL')]
    housing_data['ZIPCODE'] = housing_data.ZIPCODE.map(lambda x:clean_zip_code(x))

    indexNames = housing_data[ housing_data['AV_TOTAL'] == 0 ].index

    # Delete these row indexes from dataFrame
    housing_data.drop(indexNames , inplace=True)

    return housing_data
```

So, we here get a subset of columns, merge the street names, filter only residential data and clean the zip code so that the format is similar to the crime dataset.

For the rest, we use database type queries to generate the visualizations as seen above in methodology.

4.3. ARIMA Model

With this, we would train the ARIMA model, here we can see that in every iteration, a new model gets created so as to generate only t-1 dependency.

```
# Time series analysis for crimes of part one.

part_one_date_count_data = pd.DataFrame(part_one_crime.groupby(['OCCURRED_ON_DATE']).count()['INCIDENT_NUMBER'])

train, test = part_one_date_count_data[:"2018-04-13"], part_one_date_count_data["2018-04-13":]
predictions = list()

for index, row in test.iterrows():
    model = ARIMA(train, order=(6,1,0), freq='D')
    model_fit = model.fit(disp=0)
    output = model_fit.forecast()
    yhat = output[0]
    predictions.append(yhat)
    train.loc[index] = row.INCIDENT_NUMBER
#     print('predicted=%f, expected=%f' % (yhat, row.INCIDENT_NUMBER))
error = mean_squared_error(test, predictions)
print('Test MSE: %.3f' % error)
# plot
pyplot.plot(test)
pyplot.plot(predictions, color='red')
pyplot.figure(figsize=(15,7))
pyplot.show()
```

Now to further to predict the new dates, we used:

```
forcasted = model_fit.forecast(steps=200)[0]
```

```
pyplot.figure(figsize=(20,10))
pyplot.plot(forcasted)
pyplot.xticks(rotation='vertical')
pyplot.show()
```

5. Results

From the datasets, the results we found were surprising and expected. Why surprising because we didn't expect to see that a month in a year can affect the crime. To support, we have the analysis that shows that crime rates during the months of May through August has the higher crimes from 2015 to 2019 compared to any other month of the year. We came to the conclusion that this is possible because of the weather conditions during this time of year in Boston are comfortable and more people are out which can cause this. Also, there are cases in which travelers in summer could be the victims of a crime, this can also be the case. Then, there is a possibility where because of the summer break, young people are out doing what they are not supposed to be doing and committing crimes. The results have shown that, if there is an incident of Aggravated Assault, it is a higher probability that shooting can happen compared to any other crimes. When we join the crime dataset with the housing dataset and then generate the outcome, it supports our initial claim that higher crimes in the area do mean lower prices of houses in those areas and vice versa.

6. Discussions

The results give us some important insights into the data. Most crime affected street in Boston is Washington street. It has the highest amount of crimes compared to any other streets. It doesn't matter to the criminals what day of the week it is, they will commit the crimes. Probably it s the weather during the summertime which increases the crime rates in Boston. Maybe it is possible that the crimes have gone up since 2015 because of some underlying change in situations which is yet to be found. May be further research can dig it out why it has happened. The outcome, however, also shows that there are some places where the crime rates are high, but the housing prices are high too. Upon further reading, it came to the notice that it is the zip code of the areas where there are prime locations such as university area or business places which supports those results since the crimes do happen more in those areas, but because of the locality, the house prices are high too.

7. Future work

Although the project shows the outcome of what we have claimed to be, there is still a lot of scopes to advance this. One significant change we can think of is root cause analysis on the crime. That is, if we can get the dataset which can include the specifics of the person who commits the crime with their mental condition status, we can find the root cause of some of the crimes. This can show the city of what can be done to prevent those crimes and maybe we can help to decrease the crime rate. Then there is another scope that we can think of is, we can combine the data and start predicting the prices of houses in certain regions.

8. Conclusion

Thus from the above results and analysis, we conclude that crime data has a potentially major impact on the housing data and a time series model might potentially help in various applications like it might help real estate investors, people looking to rent houses in a particular region or police resources can be saved by focusing less on a particular region and less on others.

9. References

1) https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/
2) https://data.boston.gov/dataset/property-assessment
3) https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
4) Baldominos, A., Blanco, I., Moreno, A., Iturrarte, R., Bernárdez, Ó. and Afonso, C. (2019). *Identifying Real Estate Opportunities Using Machine Learning*.
5) Ihlanfeldt, K. and Mayock, T. (2010). Panel data estimates of the effects of different types of crime on housing prices. *Regional Science and Urban Economics*, 40(2-3), pp.161-172.
6) McNulty, T. (2000) et. al.. *Race, Crime, and Public Housing in Atlanta: Testing a Conditional Effect Hypothesis*.
7) Pope, D. and Pope, J. (2012). Crime and property values: Evidence from the 1990s crime drop. *Regional Science and Urban Economics*, 42(1-2), pp.177-188.