DATA SCIENCE
FINAL REPORT

Title:

Neighborhood Analysis

Team participants:

Akash Reddy Gurram (HR69338), Vineeth Reddy Anugu (FD78097)

Abstract:

Every building has a property tax associated with it. This property tax varies from neighborhood to neighborhood. Some neighborhoods have a high property tax associated with its buildings or properties and some have relatively low values. The reason for this is that there might be some issues like the neighborhood has a lot of dirty streets, clogged pipes etc. Additionally, crime in that neighborhood might play a key role. Neighborhoods with high crime might have a less property tax because people usually don't prefer to live in such neighborhoods. So, we are comparing the Property taxes, sustainability and crime datasets. We found relations of property taxes to different police districts and council districts. How it is being affected with the characteristics described above.

We ranked Council districts, police districts and neighborhood based on a unique metric called Quality of Life. We calculated this metric based on attributes of sustainability and the crime & safety datasets. We also found insights and negated a few insights which we thought were elementary. This is what we accomplished.

Details of the datasets:

We took 3 datasets from the Open Baltimore website. They are

- Real Property Taxes
- Sustainability
- Crime and safety

Real property taxes dataset contains information about various properties in the Baltimore city neighborhoods. Every property, be it a building or an open space has a property address associated with it. For every property there is a lot size which gives information about the size of the property. Every property is imposed with a property tax. This property tax is the described in the dataset as State tax and City tax. Also, information about what Police district and the Council district does the property belong to is given. Location of the property is given as latitudes and longitudes. Finally, information about the property being a principal residence or a non-principal residence is given. A principal residence is a house hold where as a non-principal residence can be considered as a company space or anything other than a house hold.

The sustainability dataset contains the most recent information available regarding various metrics of different neighborhoods in the Baltimore city area. These metrics include the reports of dirty streets, clogged drains over the years, the number of community-managed open spaces, amount of tree in the neighborhood. It also contains a unique metric called Walk Score which tells us how far away the basic amenities such as grocery stores, schools and hospitals etc.

The Crime and Safety dataset contains details regarding various crime rates in Baltimore city neighborhoods. Some values include the crime rate, the rate of property related crimes, narcotic crimes, auto theft crime and similar data for the crimes committed by juveniles in these neighborhoods. This will give is a good insight into what neighborhoods are safe or dangerous to live in or own property.

Exploring the datasets:

Taxes Dataset:

```
In [9]: taxes.head(50)
```

	PropertyID	Block	Lot	Ward	Sect	PropertyAddress	LotSize	CityTax	StateTax	ResCode	AmountDue	AsOfDate	Neighborhood	PoliceDistrict	CouncilDistrict
0	0001002	0001	002	15	370	2043 W NORTH AVE	14X83-10	1112.76	55.44	NOT A PRINCIPAL RESIDENCE	NaN	06/04/2018	Easterwood	Western	7.0
1	5918064	5918	064	26	380	3429 SHANNON DR	18X91-9	2801.01	139.55	PRINCIPAL RESIDENCE	1356.50	09/29/2018	Belair-Edison	Notheastern	13.0
2	0001004	0001	004	15	370	2039 W NORTH AVE	14X83-10	472.08	23.52	NOT A PRINCIPAL RESIDENCE	NaN	06/04/2018	Easterwood	Western	7.0
3	0001005	0001	005	15	370	2037 W NORTH AVE	14X83-10	472.08	23.52	NOT A PRINCIPAL RESIDENCE	NaN	06/04/2018	Easterwood	Western	7.0
4	0001006	0001	006	15	370	2035 W NORTH AVE	14X83-10	247.28	12.32	NOT A PRINCIPAL RESIDENCE	NaN	06/04/2018	Easterwood	Western	7.0
5	4179P033	4179P	033	26	340	3811 LYNDALE	14X100	2032.19	101.25	NOT A PRINCIPAL	2123.28	07/01/2018	Belair-Edison	Notheastern	13.0

Fig 1: Taxes dataset

The above is a snippet of the taxes dataset. There are few columns in this that we think do not give insights for our analysis. So, we are dropping those columns. By further exploring it we found that the LotSize column has values in different formats like aXb, aXbXc, Acres etc. We converted all these values to a single format of square feet. Coming to noisy data, the LotSize column has a lot of noisy data. For example, different formats for ACRES like ACERS, ACCERS, ACRESS, 'O' instead of 0, SQ FT etc. Rather than dropping such rows, we converted them to a square feet format. We wanted a total property tax for each property. So, we combined the state tax and the city tax columns and stored them as a new column called TotalTax. We split the location column to a latitude and longitude column

Finally, there were some missing values in the dataset, so we dropped them.

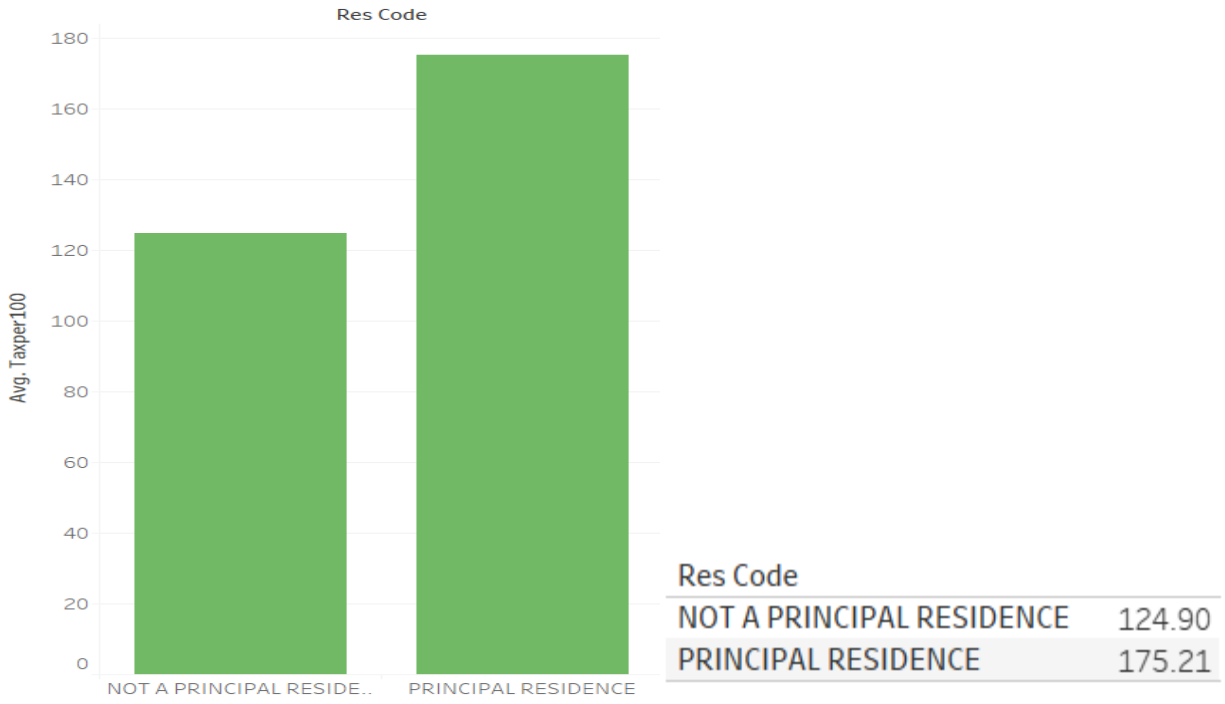

Fig 2: Average tax in different residential codes

There is a Residential Code associated with each property. This tells us if the property is a Principal residence or not. We plotted a bar graph for Residential code against Average Tax per 100 square feet. It is seen that Principal residences have relatively high property tax. The average value can be seen in the above figure.

Further exploring into the dataset, we plotted another bar graph. This time we used the police district column and the average tax per 100 square feet column. We observed that the neighborhoods in the SouthEastern Police district have a high property tax compared to the other. The reason for this may be that these are the neighborhoods with less crime or the ones with less dirty streets. This can be compared using the heatmaps provided in the notebook.

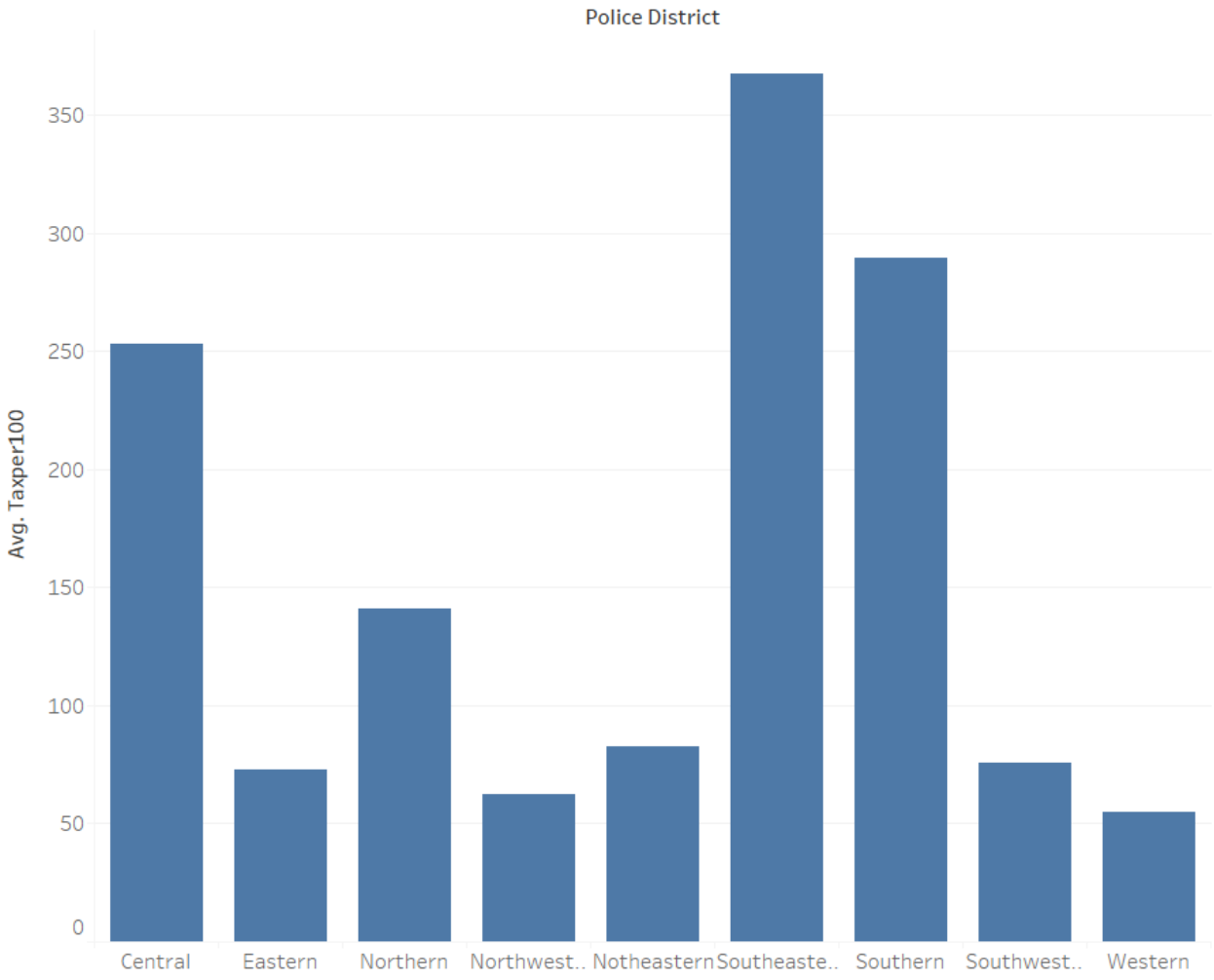

Fig 3: Average Tax in Police Districts

Sustainability Dataset:

By exploring the sustainability dataset, we noticed that it provides a lot of metrics for various neighborhoods. Some metrics are useful while a lot more are not relevant to our analysis. So, we removed these unnecessary columns such as the number of registered voters in a neighborhood. We decided to use 5 columns from this dataset namely, rates of dirty streets and clogged drains, number of trees in the neighborhood, Walk Score and number of community-managed open spaces. We chose these metrics as the dirty streets and clogged drains give us an idea of poorly maintained neighborhoods and on the contrary, the other three metrics tell us about neighborhoods that provide great value for investment by providing great facilities and being a good spot in general.

In this dataset we have information regarding the dirty streets and clogged drains for 4 years (2010 - 2013). We plotted a graph (see Fig 5) to see if the reports on number of dirty streets were increasing or decreasing in Police Districts. We can observe that the average reports for dirty streets follow a similar pattern in all the police districts. They increased during the years 2010 to 2012. However, in 2013, we can see that these reports were reduced. Hence, we can consider

this data to be relevant as they follow a trend. Also, it might be possible that in the year 2013, the council men in those police districts took these complaints seriously. There might be several reasons for this. Maybe the elections were approaching.

	Dirty Streets'10	Dirty Streets'11	Dirty Streets'12	Dirty Streets'13	CloggedDrains'10	CloggedDrains'11	CloggedDrains'12
0	41.7	41.0	48.1	37.0	4.2	4.6	6.2
1	41.7	41.0	48.1	37.0	4.2	4.6	6.2
2	41.7	41.0	48.1	37.0	4.2	4.6	6.2
3	10.5	12.2	13.9	10.2	3.4	3.3	4.8
4	10.5	12.2	13.9	10.2	3.4	3.3	4.8
5	10.5	12.2	13.9	10.2	3.4	3.3	4.8
6	90.2	66.3	79.1	65.9	4.1	6.1	3.8
7	91.5	112.1	82.8	66.0	4.2	5.8	5.5
8	91.5	112.1	82.8	66.0	4.2	5.8	5.5
9	91.5	112.1	82.8	66.0	4.2	5.8	5.5

Fig 4: Snippet of the dataset

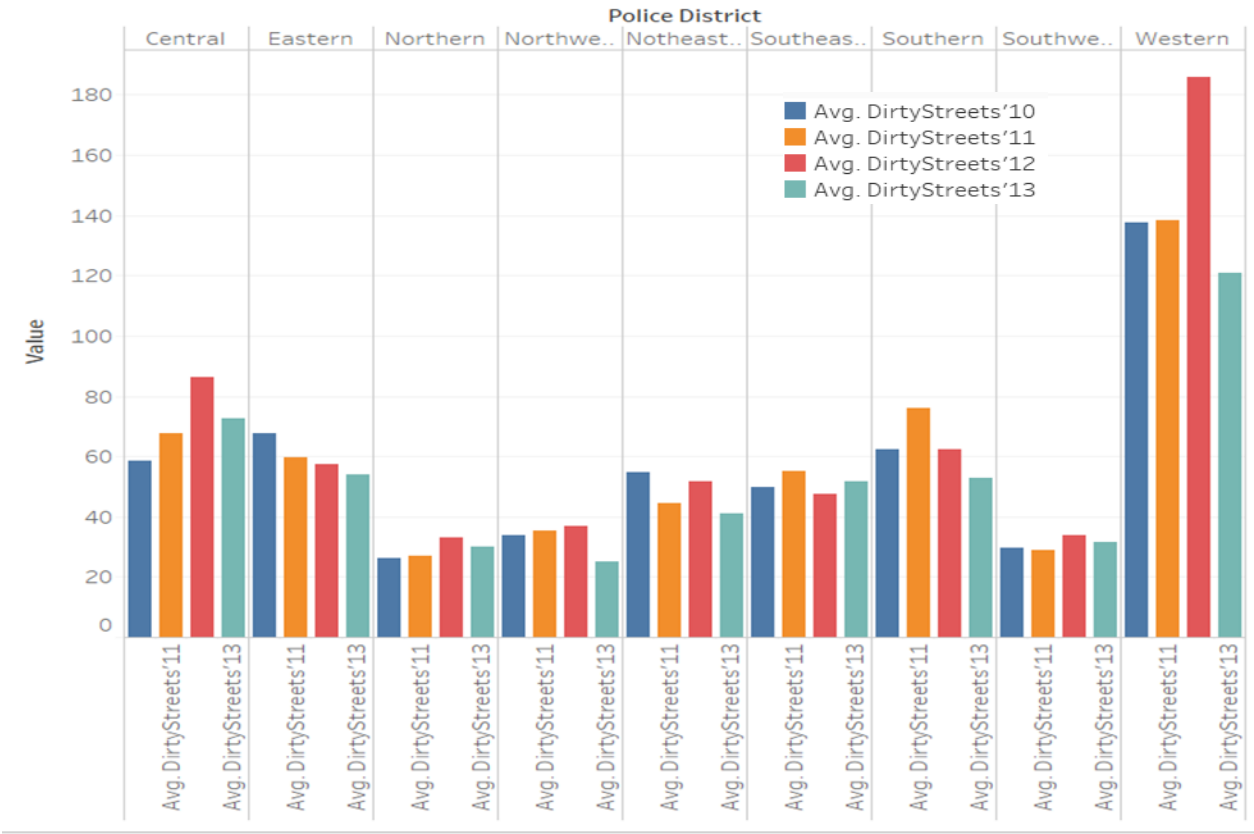

Fig 5: Dirty streets over the years

Crime and Safety Dataset:

Coming to the Crime and Safety Dataset, it provides crime statistics regarding various neighborhoods. This was a very clean dataset which didn't need any preprocessing other than splitting combined rows of neighborhoods into their own individual rows to compare across datasets. All the metrics provided in this dataset are values of rate of specific crime per 1000 residents except for rate of homicides which is provided for 10,000 residents.

	CrimeRate	ViolentCrimeRate	PropertyCrimeRate	JuvenileArrestRate	JuvenileViolentOffenses	JuvenileDrugOffenses	911Shootings
0	48.965453	11.328527	37.262429	17.316017	2.597403	4.329004	3.183223
1	47.531149	9.460083	37.148131	8.666667	0.666667	0.000000	1.461314
2	42.300066	10.514931	30.583428	16.818500	4.905396	0.000000	1.021450
3	72.435553	26.114393	44.978518	53.972104	10.309278	14.554275	9.667025
4	72.435553	26.114393	44.978518	53.972104	10.309278	14.554275	9.667025
5	50.313932	12.557271	36.738503	10.933558	2.523129	2.523129	2.630239
6	50.313932	12.557271	36.738503	10.933558	2.523129	2.523129	2.630239
7	48.468684	10.944542	37.064288	13.011152	3.717472	1.858736	1.655477
8	48.468684	10.944542	37.064288	13.011152	3.717472	1.858736	1.655477
9	66.538384	23.698602	41.219364	29.668412	1.745201	6.980803	6.380393

Fig 6: Snippet of Crime and safety dataset

We could not infer a lot from the above 3 datasets individually. But after merging all the 3 and exploring them, the inferences became clear.

Merging the Datasets:

All our datasets have a common attribute called Neighborhoods, which was used as the basis for merging them together.

Neighborhood	PoliceDistrict	CouncilDistrict	Latitude	Longitude	TotalTax	Taxper100	...	PropertyCrimeRate	JuvenileArrestRate	JuvenileViolentOffenses
Belair-Edison	Notheastern	13.0	39.322915	-76.562356	2940.56	177.763269	...	39.389068	24.112231	9.206488
Belair-Edison	Notheastern	13.0	39.316802	-76.565057	2133.44	152.388571	...	39.389068	24.112231	9.206488
Belair-Edison	Notheastern	13.0	39.323463	-76.564369	2423.72	154.771392	...	39.389068	24.112231	9.206488
Belair-Edison	Notheastern	13.0	39.323428	-76.564324	2475.64	158.086845	...	39.389068	24.112231	9.206488
Belair-Edison	Notheastern	13.0	39.323364	-76.564234	2423.72	154.771392	...	39.389068	24.112231	9.206488
Belair-Edison	Notheastern	13.0	39.316841	-76.564766	2145.24	153.231429	...	39.389068	24.112231	9.206488

Fig 7: Snippet of the merged dataset

Visualizations/Results:

We used Tableau to help us for visualization purposes. We also used matplotlib and seaborn plots for other visualizations in the notebook.

1.Average tax per 100 in each council district (Circle Plot)

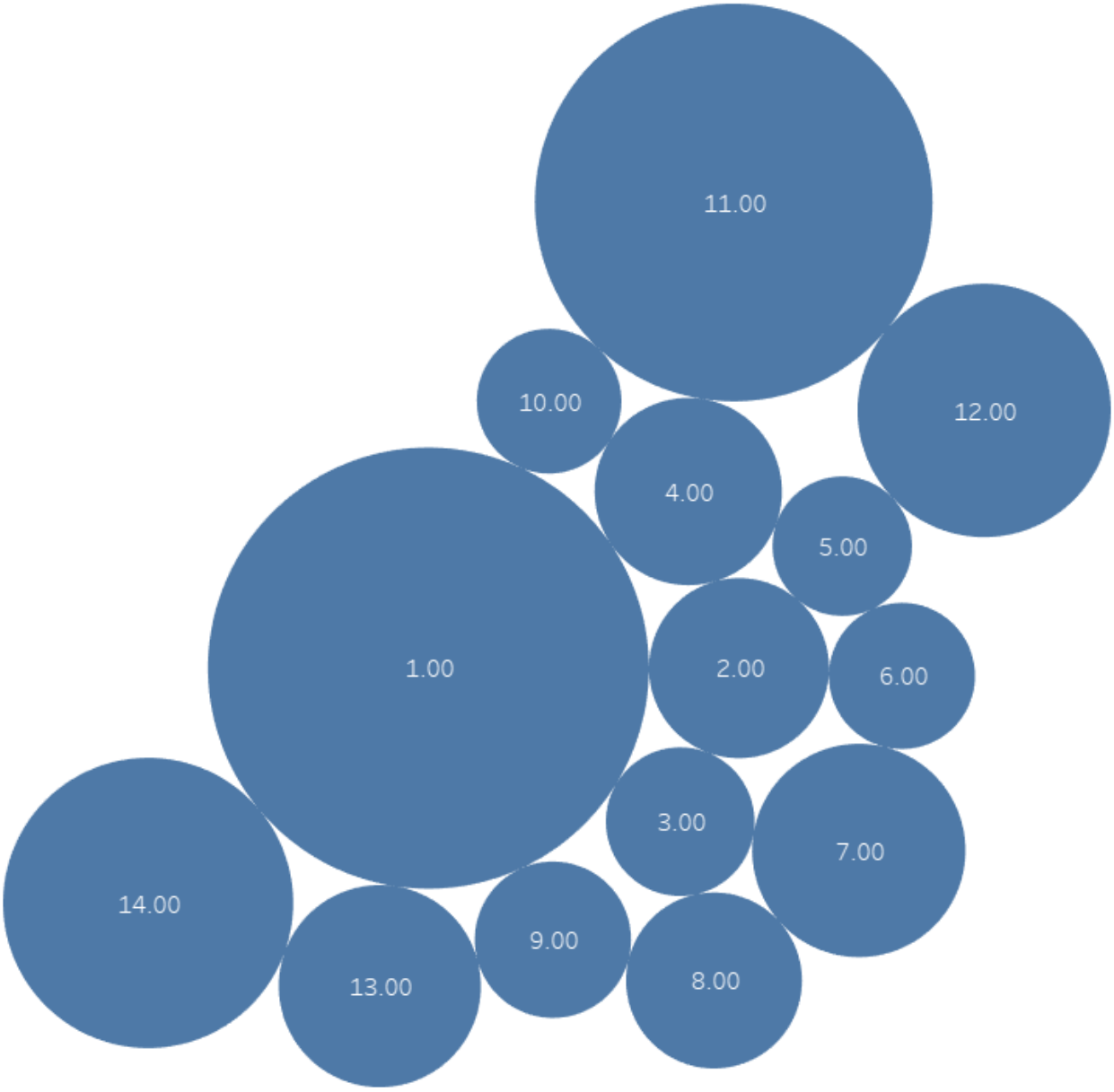

Fig 8: Bubble chart showing the council district with highest property tax

This plot shows that the council district 1 has the highest property tax. The smaller bubbles have small property tax and the bigger bubbles have relatively high property tax per 100 square feet. Why does council district 1 and 11 have high property tax? Maybe they have a smaller number of dirty streets, less crime and other such factors. Let's see if this is true by further analysis.

2. Lets us see if number of trees affect the property tax: Negative Result

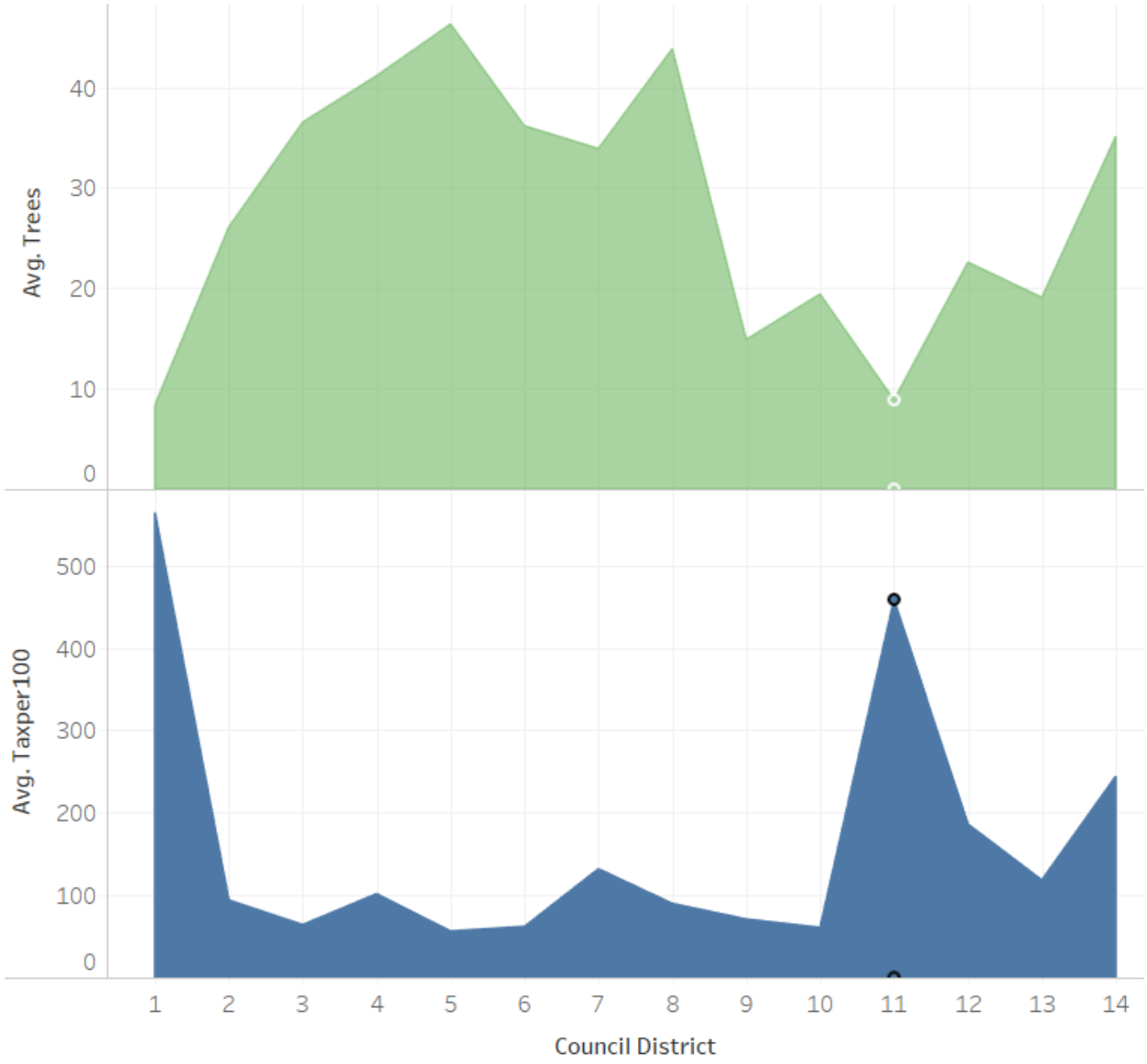

Fig 9: Trees vs Tax in council districts

In council district 1 and 11, the property tax is extremely high. We first thought that since it has a high tax, maybe it is a good neighborhood for families and households. Generally, such areas have a lot of trees. So, we plotted a graph to see the number of trees in those council districts. We observed that council district 1 and 11 have the least number of planted trees. What we inferred is that maybe these are the neighborhoods that are urbanized. Hence, less trees in those council districts. In fact, Downtown belongs to the council district 11. Hence, this inference is acceptable.

3. Comparing 911 for narcotics and 911 for shooting:

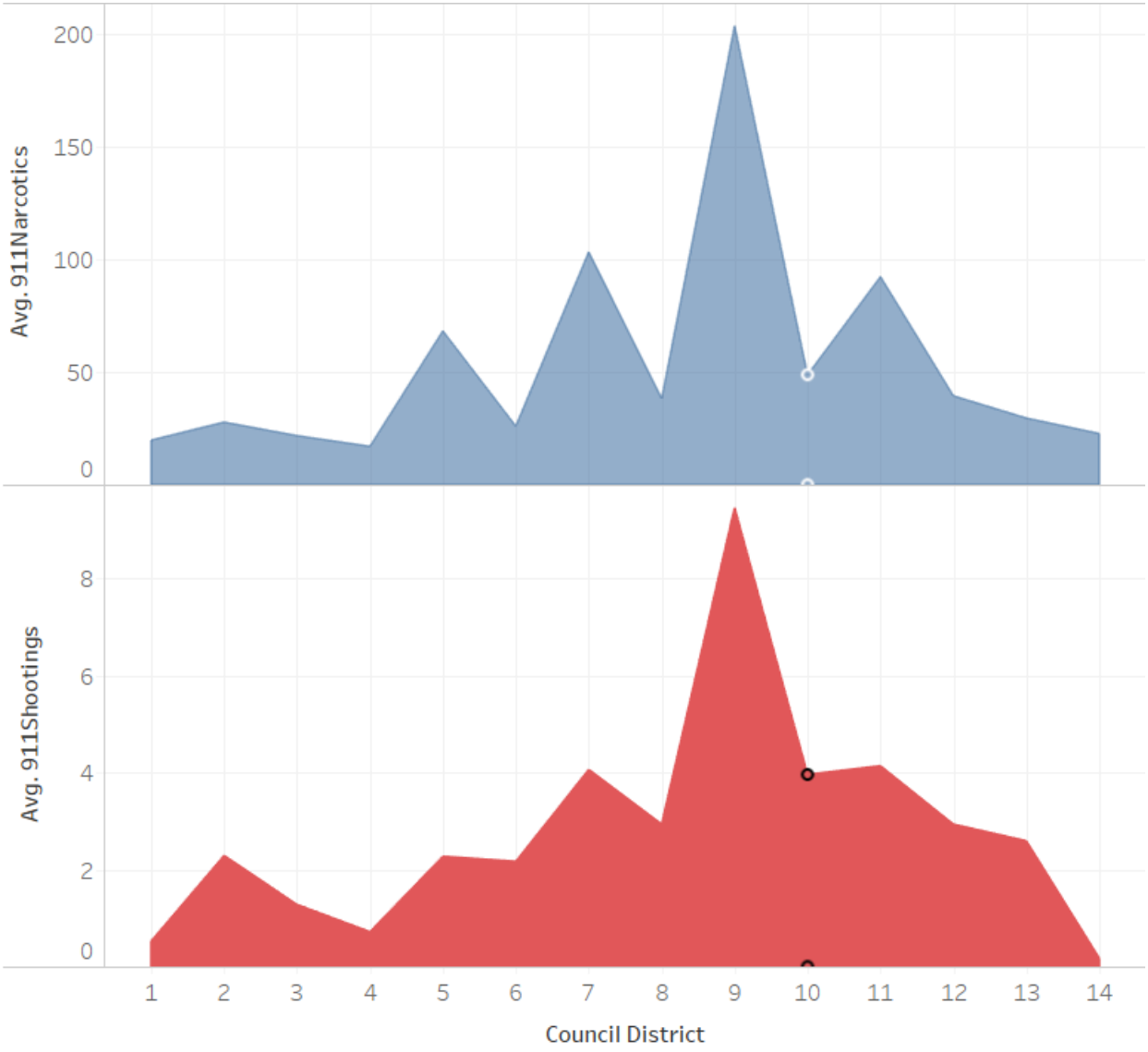

Fig 10: 911 calls narcotics vs 911calls shooting

We compared the average number of 911 calls for narcotics to that of shootings. In almost all the council districts they both follow the same pattern. From the above graphs, we can see that council district 9 has the highest calls for narcotics and shooting. We can infer that these shooting are directly related to the narcotics related incidents. Also, when we ranked the council districts at then end, we found that council district 9 has a low rank thus supporting this result.

4. Crime rate across various neighborhoods:

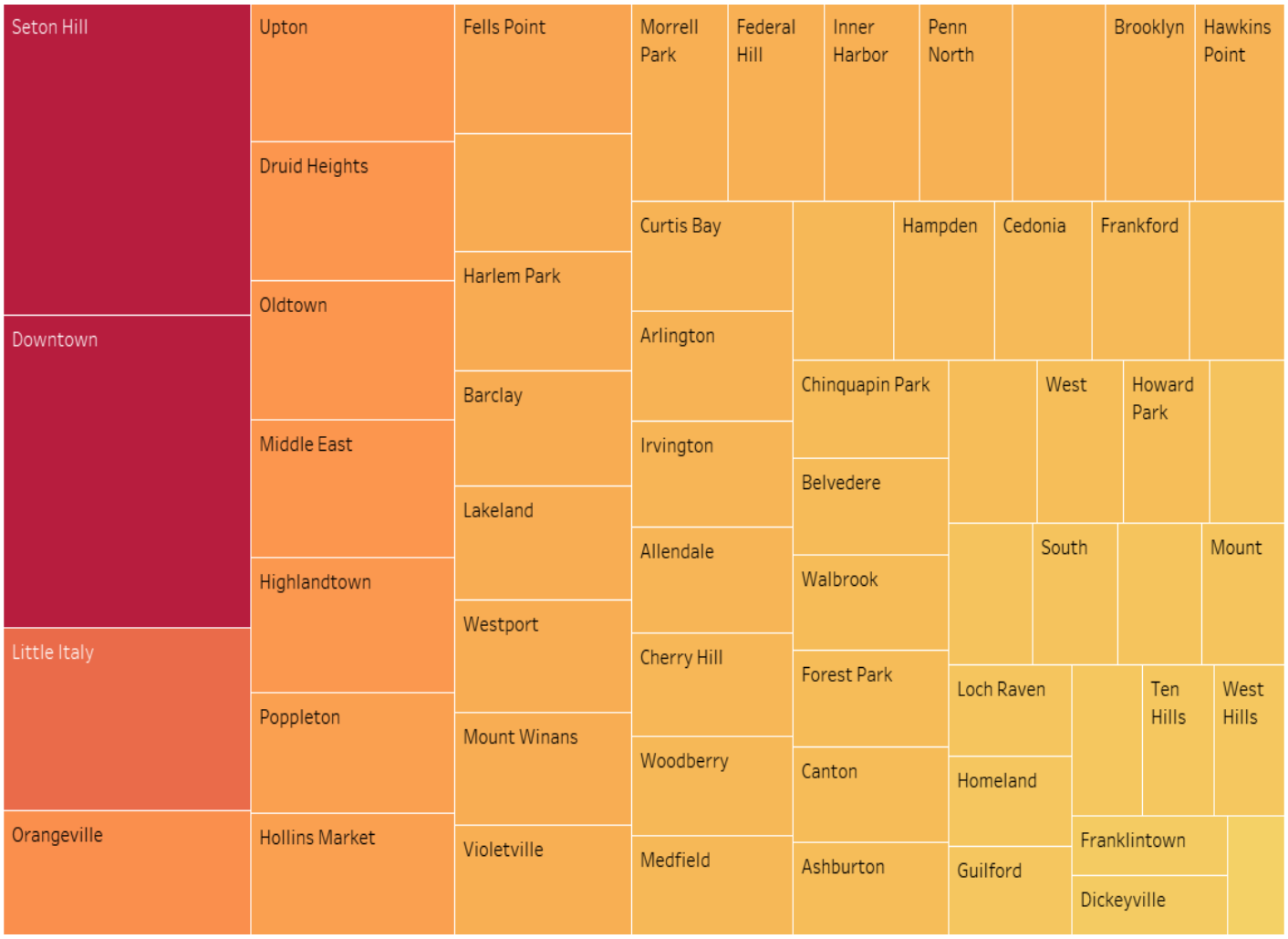

Fig 11: Tree map showing crime rate in neighborhoods

The above is a tree map which shows the neighborhoods and their crime rate. Red indicates high and lighter shades of it indicate low crime rate. We can see that Seton Hill is the neighborhood with the highest crime. Downtown also has the same crime rate. The reason for those equal values is that Downtown and Seton Hill are geographically close separated by 0.7 miles. If anyone were to buy a property and the property is a house, then I would suggest not to buy one in downtown because of this crime rate. Moreover because of high non-principal residencies in downtown, their property tax will be high.

.

5. Relation between property tax and property crime: Negative Result

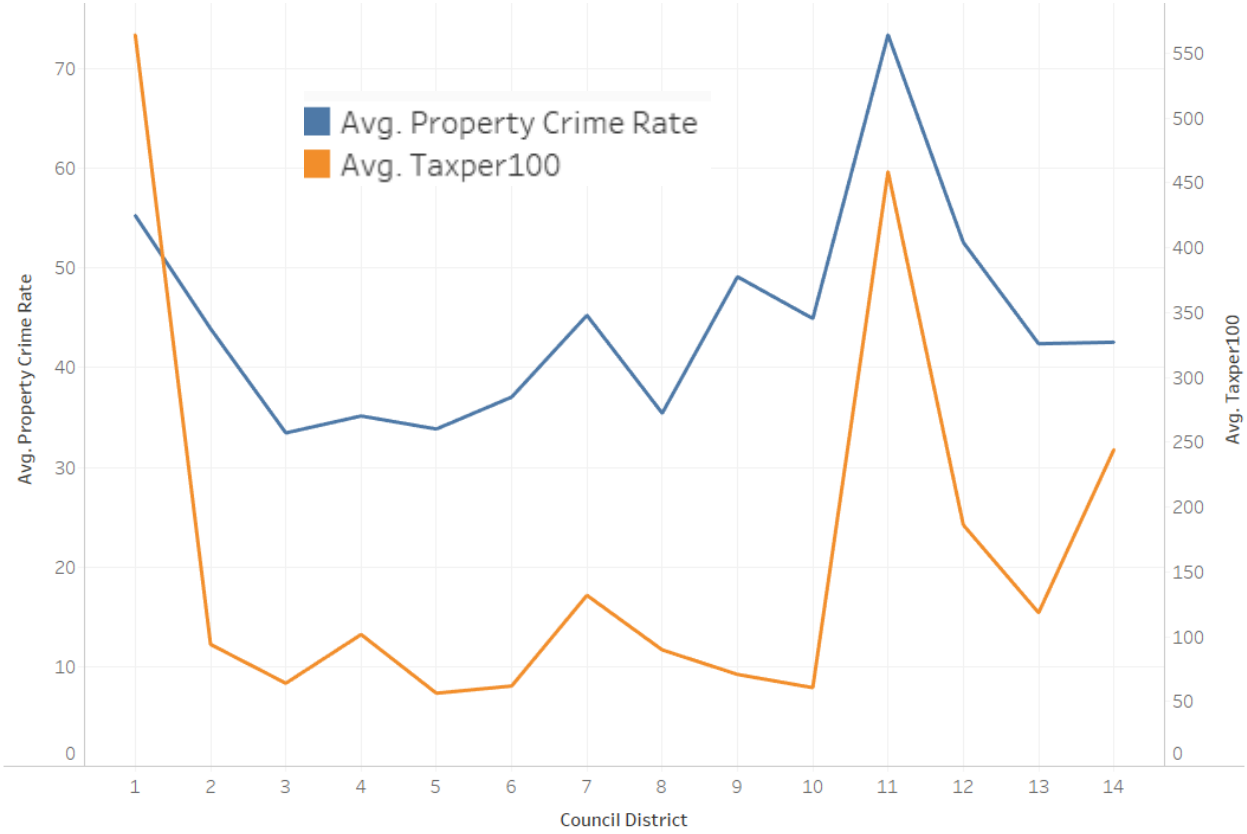

Fig 12: Property tax and property crime in a council district

This is a negative result because we were expecting the property tax to be less in areas with high property crime. Property crime includes crime like burglary, larceny, theft etc. But the above visualization shows that they are relative. For example, the council district 11 has the highest property crime rate as well as a high property tax per 100 square feet. So, we inferred that the neighborhoods with high property tax have expensive life style or richer life style households, jeweler shops etc. Thus, property crimes may be more in such areas. That may be the reason for the above result.

6. How are the juvenile drug offences in different council districts?

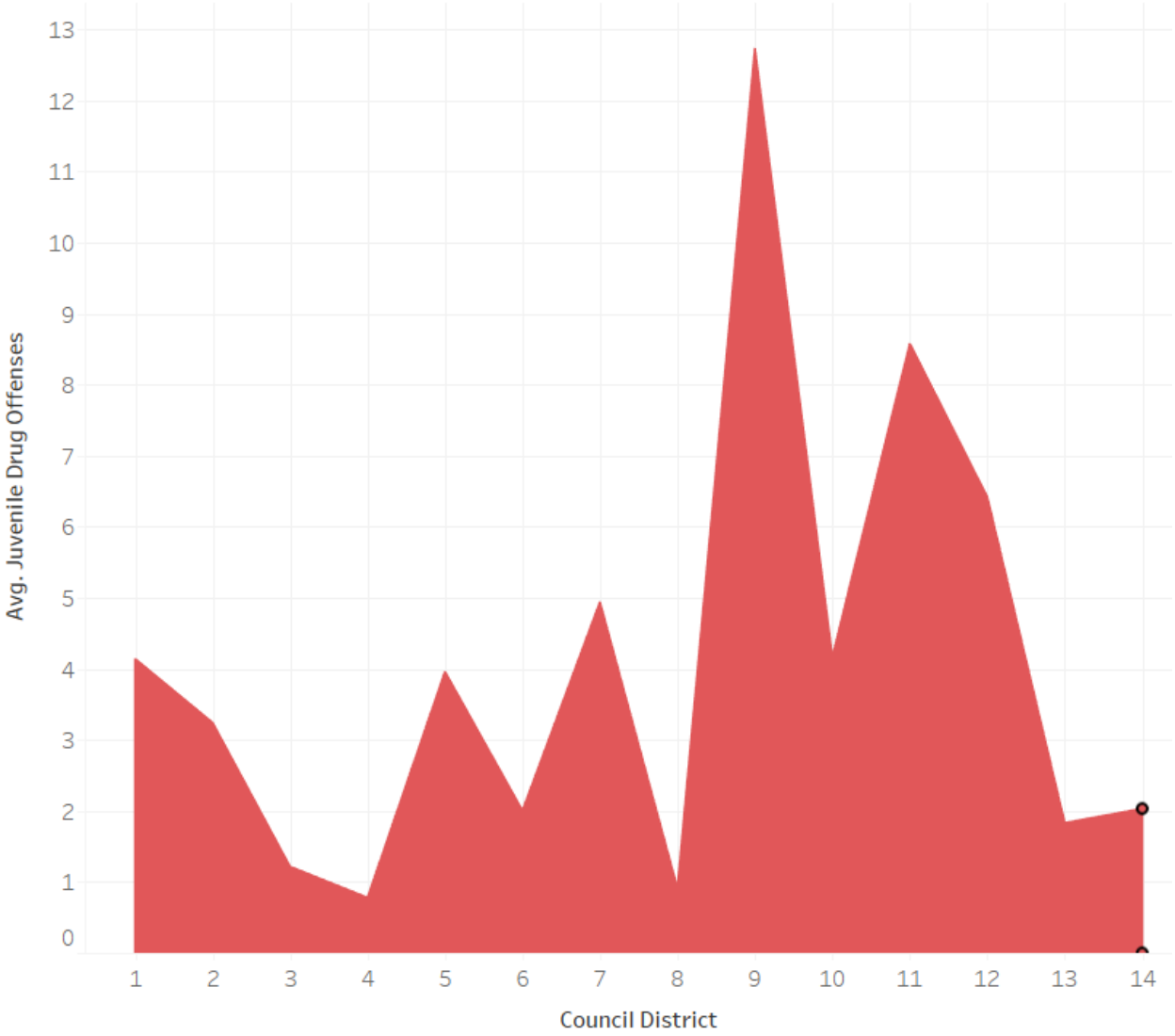

Fig 13: Avg Juvenile drug offences in a council district

We can observer that in Council district 9, the average juvenile drug offences are high compared to the others. People would be interested in buying a property in council district 9 because of its low property tax (evident from the 1st visualization). However, the juvenile drug cases here are high. So, if someone with teenage children were to buy a new house, our analysis would tell them that it is a bad choice.

Apart from the above results, there are some other results like Heat maps, 3-dimensional scatter plot in the notebook.

Insights gained from the data:

We created a quality of life factor using the datasets. Using this we found that, Council district 4 has the highest quality of life factor. This would be the best council district to invest in property. Council districts 9 and 11 are ranked the least. I would suggest someone not to buy a house in these neighborhoods because they have high crime rates, dirty streets etc.

We learned the following from our analysis and our work on the datasets:

- We learned that narcotics and shooting incidents are very relative to each other. We inferred that shootings mostly happened due to narcotics fueled incidents. Thus, preventing narcotics might also reduce the shooting related incidents.
- Council districts 11 and 9 should be avoided (in general as well) especially when looking for houses/properties to raise a family due to the high juvenile crime rate in those council districts.
- We were able to find out the best neighborhoods, police districts, council districts to own a property or invest in a property based on the quality of life which we calculated using all the attributes from the given data.
- From the quality of life metric, we also noticed that it directly corelates to the tax in that area.

Why is this useful?

This analysis of neighborhoods, police districts and council districts data can be useful in many ways. They are:

- For a real estate agency to figure out which properties make for good investments based on our quality of life metrics.
- Similarly, this helps individuals in making a calculated decision regarding which property to buy/invest in.
- This data can be used by the government agencies to focus on areas to improve in the Baltimore city area.
- This can also aid tourists in avoiding certain areas in the Baltimore city based on our easily interpretable visualizations of the Baltimore city area data.

Future works:

Are crime and the sustainability data the only factors that affect the property taxes? In the future we would like to check this with other data sets like 311 requests etc. We would do the same with similar datasets from other cities like NYC, Boston etc. and see if their property taxes are affected by crime or not. We would like to use datasets from yelp to find the number of restaurants in a council district and see if they have any relationship to the property taxes. Further, we would like to add data regarding famous hotspots/attractions in the city and how they would affect the areas due to more care from the government to preserve these locations and leading to higher property taxes in said areas.