FINAL PROJECT FOR IBM'S APPLIED DATA SCIENCE CAPSTONE ON COURSERA

The Battle of the Neighborhoods:

Amsterdam and its Boroughs

Robert B.

11/13/2020

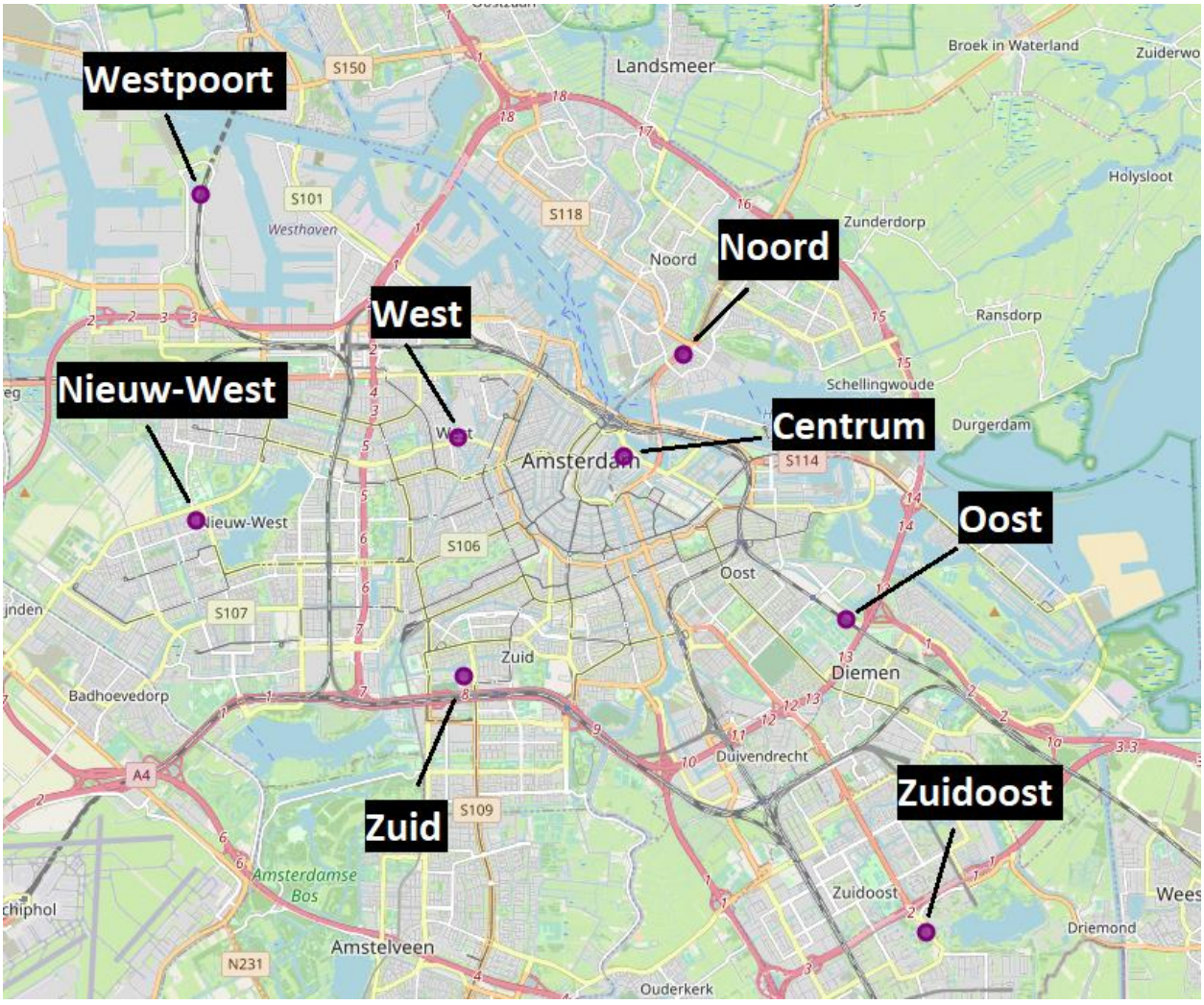
This paper looks to examine the relationship between Safety, Population Density, and Number of Venues, between the eight districts in Amsterdam. K Nearest Neighbor, as well as a Multi Linear Regression will be used to examine the data found through Foursquare and the website of the Government of Amsterdam. Ultimately, the results are, statistically speaking, entirely insignificant, which likely stems from a lack of data points. Future research would benefit from dividing the eight districts up further into smaller neighborhoods (42 of them) which might lead to a more meaningful analysis. Regardless of the insignificance though, the regression did seem to suggest that Population Density is positively (barely) correlated with crime rates, while Number of Venues in a district is negatively correlated with the rate of crime. K Nearest Neighbor also was able to identify the difference between industrial work zones and more residential, commercial, districts in Amsterdam.

1. Introduction

Background

In this paper, I will be looking at the eight boroughs in Amsterdam and analyzing them based on a few key variables. Amsterdam has been around since the 13th century, yet the current eight boroughs are rather recent. It wasn't until May 1st of 2010 that the 15 boroughs in Amsterdam were condensed down into the 8 that they are now (see Figure 1 below). Amsterdam is a huge tourist attraction for many reasons: the canals, museums, famous painters, architecture… and of course the coffee shops and Red Light District… but it is also a very densely populated city with over 800,000 residents living in these eight boroughs. That averages out to about 100,000 residents per borough, which is especially impressive given the relatively small size of each borough (the biggest borough is 49km^2 while the smallest is only 8km^2!)

Figure 1: Amsterdam's Districts



Problem

This leaves us with a problem though: how can a tourist best spend their time in Amsterdam? Surely, you want a place that has many interesting venues in the vicinity, but you also want to try and avoid crowds as much as possible (especially now that COVID-19 is such a serious threat), so where do you go? Given this problem, I think it would be interesting to look at Amsterdam and see if its high population density and number of venues is in any way correlated to the overall safety ratings of each borough, so as to hopefully find the best place to stay that has 1) many venues and attractions nearby 2) not too many people around and 3) a low crime rating.

Interest

But who would be interested in this? Well, literally anyone thinking of visiting the city, of course! Granted, it might not be happening anytime soon, given the pandemic, but once travelling becomes safer again, a lot of people will be hopping on planes again, jetting off to all corners of the world. Where will the people visiting Amsterdam go to stay? Perhaps this paper might shed some light on that for them. I think this is of interest even to people currently living in Amsterdam already, too. Imagine you are living in the city's center: it is way too busy for you and crime is higher than you might like, but would moving to a district be the right choice? Would you be giving up access to too many venues? Perhaps the other neighborhoods are even less safe… This right here is an excellent way to find out those answers!

2. Data

In order for there to be any meaningful analysis later on in the paper, data will have to be collected. The data that needs to be found includes the locations of the boroughs in Amsterdam, their population densities, safety ratings, and list of venues in and around each borough. The name of the boroughs, as well as their population densities, are easily found on Wikipedia and do not need to be converted further. The locations (coordinates) of each borough were found on the City of Amsterdam's very own government website. As you can see, the government website refers to the boroughs as "districts" while Wikipedia lists them as boroughs. In this paper, I will use the terms "borough", "district", and "neighborhood" interchangeably. The table below (Table 1) shows the names, coordinates, and densities of the eight districts in Amsterdam after the data has been turned into a dataframe on Jupyter Notebooks.

Table 1: Amsterdam's Eight Districts

	Neighborhood	Latitude	Longitude	Population Density (people/km^2)
0	Centrum	52.37321	4.903712	13748
1	Westpoort	52.41095	4.803871	10
2	Zuidoost	52.30465	4.974994	4391
3	West	52.37611	4.86452	15252
4	Nieuw-West	52.36407	4.802676	4478
5	Oost	52.34981	4.956049	7635
6	Zuid	52.34172	4.86605	9349
7	Noord	52.388	4.917663	2269

The next variable that needs to be found is each boroughs' safety index. Fortunately, I speak Dutch, so I was able to, again, go to the Government of Amsterdam's website and read through their 2019 publication aptly titled "City District in Numbers 2019". This comprehensive publication lists and outlines many different variables and statistics for the eight districts in Amsterdam, including the safety ratings. These ratings go from low (safe) to high (dangerous). The report uses a simple, binary, way of indexing whether or not a neighborhood is considered "safe": a Safety Index under 100 is considered safe, while a Safety Index over 100 is considered dangerous. We will not use a binary system though, as I believe it will be more meaningful to have further distinctions in safety (80 vs 85, for example, as opposed to just 0 or 1).

This brings us to our final variable: the number of venues in each district. Much like in the lab with New York, or the peer reviewed assignment with Toronto; here too we will be using Foursquare and its API function via Python to retrieve a list of venues nearby. I will not get into the specifics of this, as I am sure everyone taking this course has had plenty of experience by now doing this, but the code (with explanatory markdowns) can be found in the Jupyter Notebook that was shared along with this report on my GitHub repository. The important things to note are that I used a radius of 500 meters (given Amsterdam's relatively small size) and a limit of 200 venues per district.

Table 2: Most Common Venues by Neighborhood

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Centrum	Bar	Hotel	Coffee Shop
1	Nieuw-West	Soccer Field	Theater	Gym / Fitness Center
2	Noord	Plaza	Yoga Studio	Bed & Breakfast
3	Oost	Soccer Field	Hockey Field	Playground
4	West	Nightclub	Snack Place	Supermarket
5	Westpoort	Heliport	Harbor / Marina	Yoga Studio
6	Zuid	Bagel Shop	Office	Supermarket
7	Zuidoost	Farm	Harbor / Marina	Park

Before we go ahead and add the number of venues per district to our data table, I thought it would be interesting to take a quick look at the most common venues found in each district, according to Foursquare. As can be seen in the table above (Table 2), the Centrum (Dutch for center, yes!) has bars, hotels, and coffee shops as the top three most common types of venues. This makes a lot of sense, given that the Centrum of Amsterdam is the biggest tourist hotspot, catering largely to people looking to drink, smoke, and stay the night. Taking quick looks at the data like this can be helpful, I find, to verify that our assumptions regarding the data are, so far, correct.

Another part of the data table that stands out is the Westpoort (West Haven) district, given that the first most common type of venue is listed as “heliport”. Having lived in Amsterdam for over a decade, this does not come as a surprise to me, as Westpoort is almost entirely a business district / industrial park that serves as the corporate headquarters for many, many companies operating in Amsterdam. As could be seen in Table 1, the population density for Westpoort is a mere 10 residents per square kilometer, so having heliports and harbors as the most common types of venues is not entirely surprising. What will be interesting in the analysis and results section of this paper is to see whether Westpoort will have a significant impact, given how different it is from the other seven districts.

Table 3: Complete Data Table

	Neighborhood	Latitude	Longitude	Venues	Population Density	Safety Index	Safe
0	Centrum	52.37321	4.903712	88	13748	69	1
1	Westpoort	52.41095	4.803871	2	10	0	1
2	Zuidoost	52.30465	4.974994	4	4391	132	0
3	West	52.37611	4.86452	17	15252	79	1
4	Nieuw-West	52.36407	4.802676	10	4478	130	0
5	Oost	52.34981	4.956049	6	7635	80	1
6	Zuid	52.34172	4.86605	9	9349	56	1
7	Noord	52.388	4.917663	13	2269	112	0

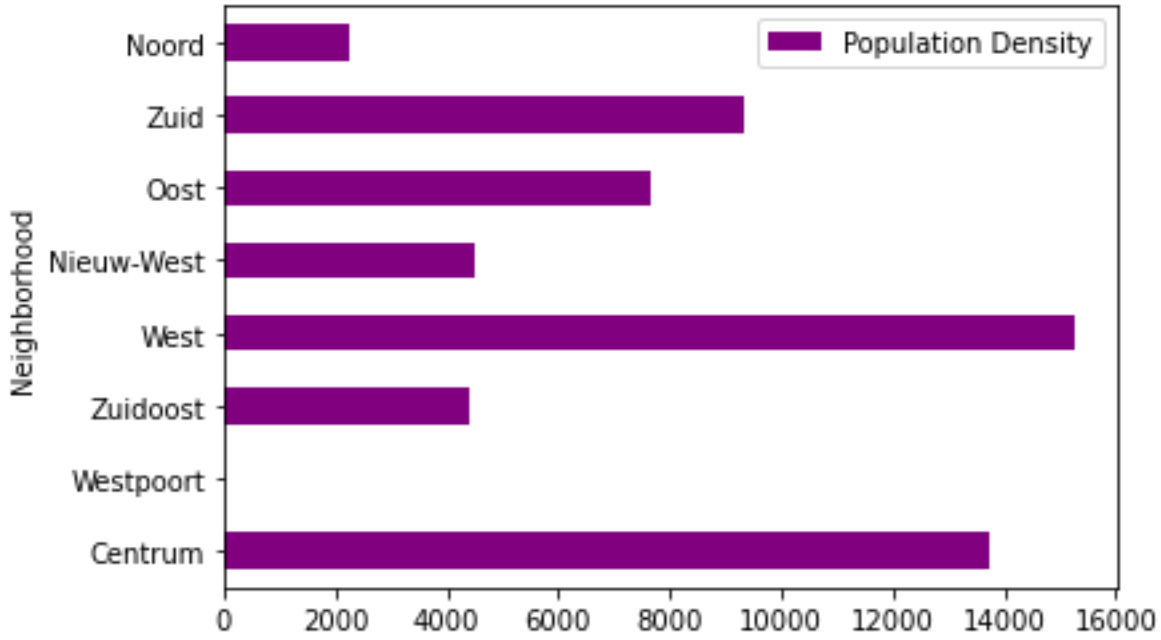
Now that all of the data has been collected, we can gather it all together into one data table (Table 3). New additions are the “Venues” column, which shows the number of venues found via Foursquare in each district, as well as the “Safety Index” and “Safe” binary variable. Unsurprisingly, the number of venues is greatest in Centrum, as that is the tourist hotspot. West, Nieuw-West (New-West), and Noord (North) are all in the double digits, while very few venues were found for Westpoort, Zuidoost (Southeast), Oost (East), and Zuid (South). As stated previously, any neighborhood with a safety index value below 100 is considered “safe”, while anything above is considered “unsafe” by the government’s guidelines. Only three of the eight neighborhoods are considered unsafe: Zuidoost, Nieuw-West, and Noord.

3. Data Analysis and Methodology

Bar Charts

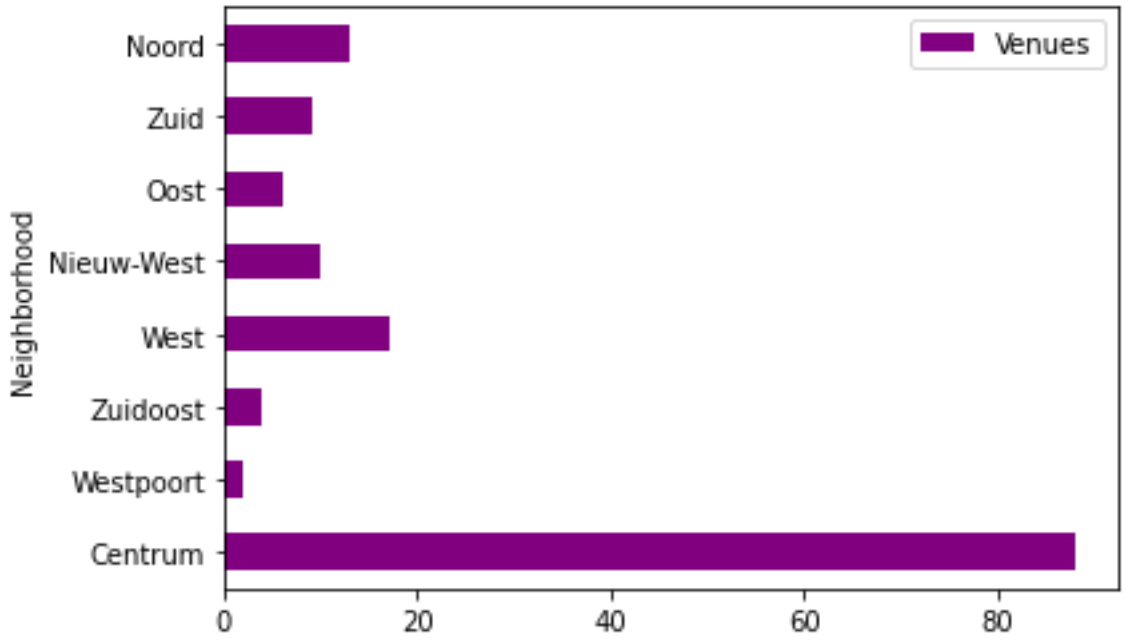
With the data collecting and scrubbing out of the way, it is time for us to take a closer look at what we’ve collected. To start, let’s visualize the data by creating some bar charts, this will give us a good idea of what the data looks like. Looking at the population density (Figure 2) below, we can see that there are some stark differences between neighborhoods. West and Centrum have the most citizens per square kilometer; while Westpoort has such a small number that is isn’t even visible on the chart. As mentioned previously, this does not come as a big surprise, per se, but it certainly highlights the stark difference between heavily populated places such as Centrum and industrial work zones like Westpoort.

Figure 2: Population Density Bar Chart



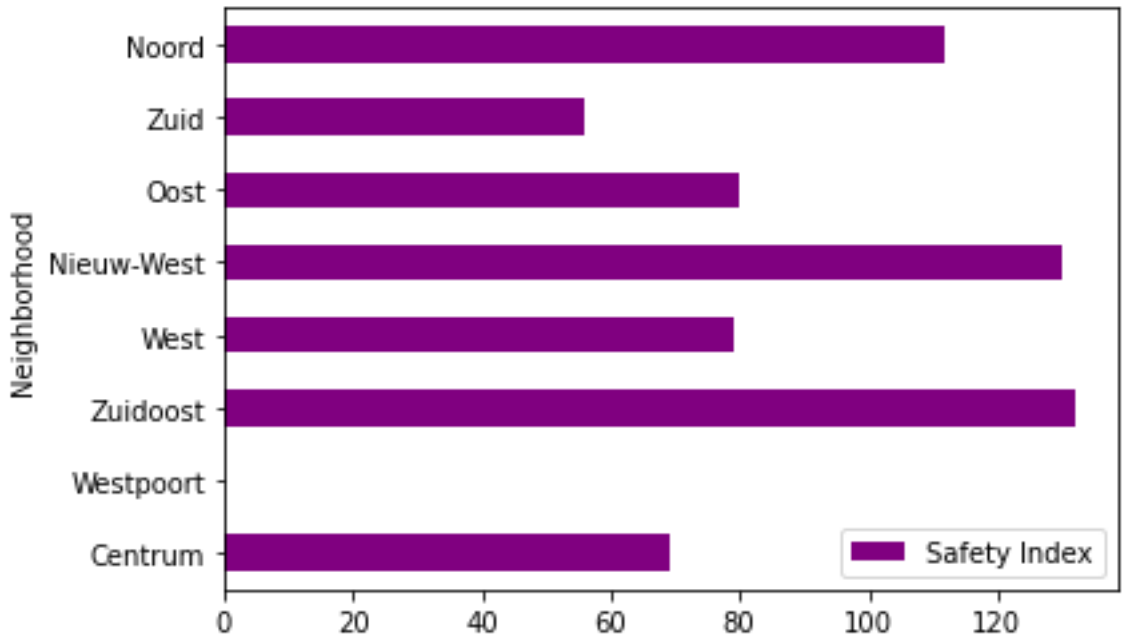
The second variable we can visualize are the number of venues in each neighborhood. Looking below (Figure 3), we can see that Centrum is by far the most popular place to set up shop. Despite the high population density in West and Zuid, they have relatively few venues found by Foursquare.

Figure 3: Venues Bar Chart



The final variable, the safety index, shows us that three districts (Noord, Nieuw-West, and Zuidoost) are above the 100 point threshold, making them “less safe” as defined by the government of Amsterdam. Zuid, Oost, West, and Centrum have middling values, allowing them to be marked as “safer” neighborhoods according to the Dutch report. Finally, we have Westpoort, the industrial work zone, with no rating. I find this the most surprising, as crime still seems like something that would occur in a work zone, even if it is less residential. Perhaps security is a lot stricter there, or there is nothing noteworthy going on? Either way, that is beyond the scope of this paper.

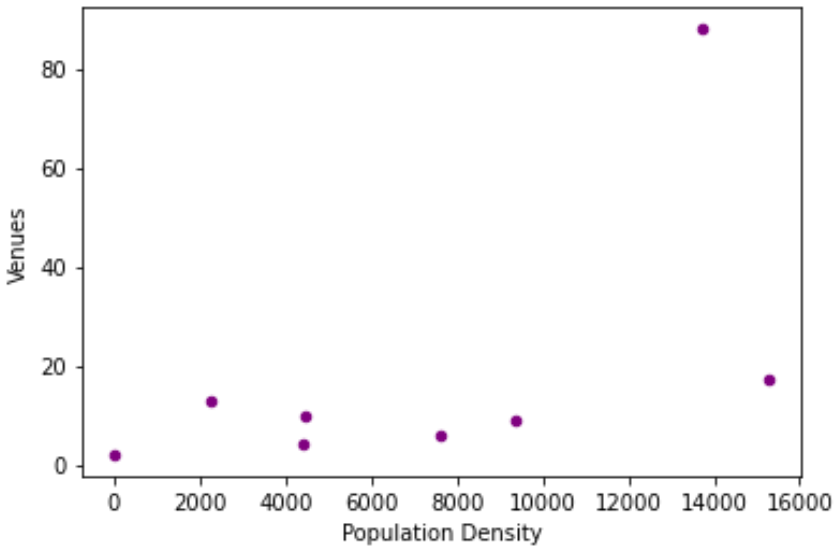
Figure 4: Safety Index Bar Chart



Scatter Plots

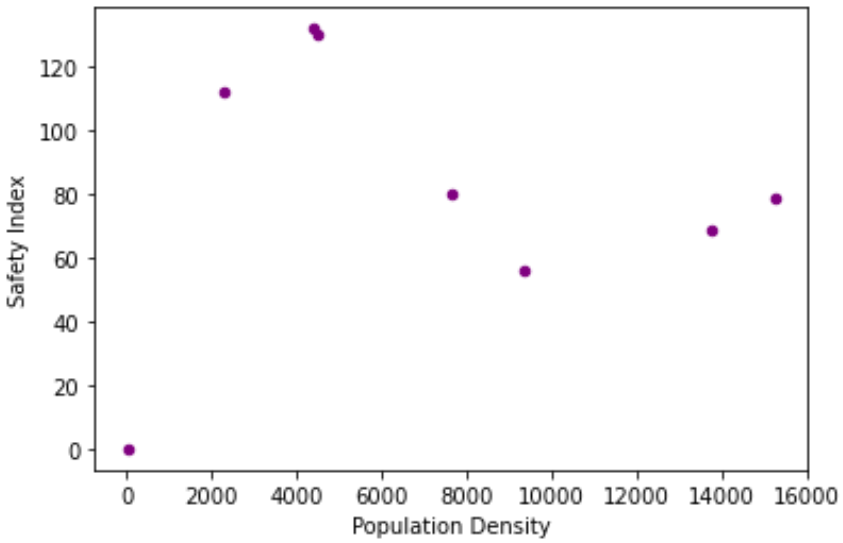
Having seen each variable visually represented on its own, it is now time to see how they might look plotted against each other. To do this, we will be using a couple of scatter plots to see if there are any trends or correlations visible when graphing the variables. First up, we’ll look at the Population Density versus the number of Venues in each district. As we can see (Figure 5), there really does not seem to be much of a correlation at all. Centrum is, obviously, a huge outlier, while the other points appear to, more or less, be a horizontal straight line. Without Centrum, one could argue that there is a very slight increase in number of venues as population density goes up, but it is so insignificant that it is best ignored.

Figure 5: Population Density vs. Venues Scatterplot



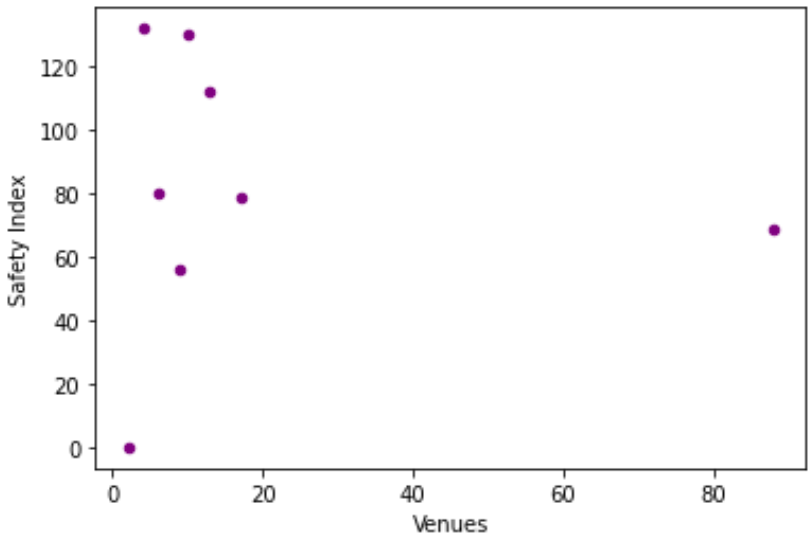
Instead, let's look at the interaction between population density and the safety index. From the graph (Figure 6), we can see that the relationship appears to almost be a wave function in nature, going from one maximum, to a minimum, back up to a maximum. While it might look like a wave with the inflection points, we have to be realistic and admit that there are only 8 data points on this graph. Truly, they are too few, so this might just be a rare coincidence.

Figure 6: Population Density vs. Safety Index Scatterplot



Finally, let's take a look at how the venues and safety index ratings look when graphed against each other. Again, we can see from the graph (Figure 7) that Centrum is a rather big outlier. The other data points seem to indicate a potentially positive, linear, relationship, but again, without more data points, this becomes hard to ascertain.

Figure 7: Venues vs. Safety Index Scatterplot



Regression

With all of the visualizing out of the way, it is time to look at what (if any) correlations we can really find between the dependent variable (Safety Index) and the two independent variables (Population Density and Venues). In the figure below (Figure 8) we can see the results from the OLS regression done on the data. First off, before even looking at any of the coefficients, we have to acknowledge the incredibly low R^2 value of 0.007, combined with the P values that are so high; they essentially show that none of these results are even remotely statistically significant. Let's keep that for the results and discussion sections of the paper though, and for now end with the equation found by having run the regression:

$$Safety\ Index = (0.0005 \times Population\ Density) - (0.157\ \times Venues) + 81.5631$$

Figure 8: Multiple Linear Regression - Results

OLS Regression Results

Dep. Variable:	Safety Index	R-squared:	0.007
Model:	OLS	Adj. R-squared:	-0.390
Method:	Least Squares	F-statistic:	0.01747
Date:	Sat, 14 Nov 2020	Prob (F-statistic):	0.983
Time:	07:47:51	Log-Likelihood:	-40.984
No. Observations:	8	AIC:	87.97
Df Residuals:	5	BIC:	88.21
Df Model:	2		
Covariance Type:	nonrobust		

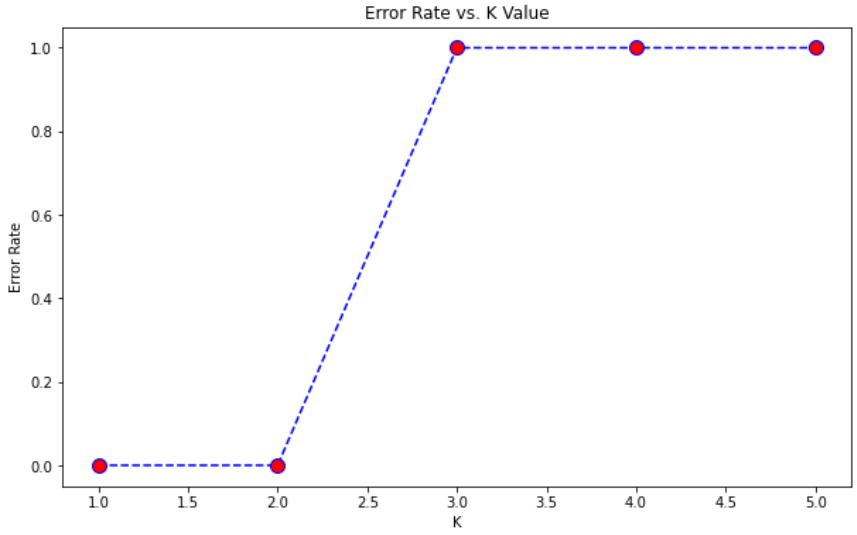
	coef	std err	t	P>|t|	[0.025	0.975]
const	81.5631	31.592	2.582	0.049	0.354	162.772
Population Density	0.0005	0.004	0.114	0.914	-0.011	0.012
Venues	-0.1570	0.840	-0.187	0.859	-2.317	2.003

Omnibus:	1.181	Durbin-Watson:	2.716
Prob(Omnibus):	0.554	Jarque-Bera (JB):	0.491
Skew:	-0.570	Prob(JB):	0.782
Kurtosis:	2.580	Cond. No.	1.52e+04

KNN

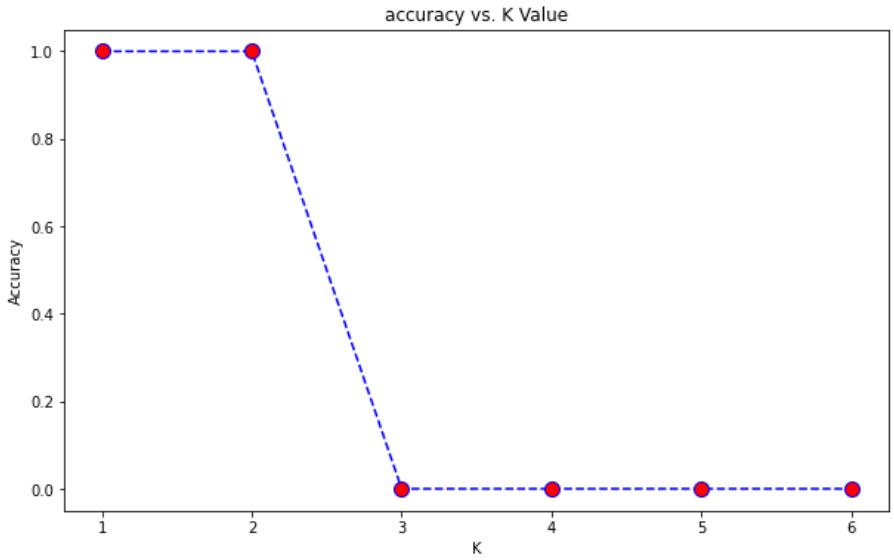
Last, but not least, let's see what happens when we try and cluster the districts using K Nearest Neighbor (KNN). When it comes to using KNN, I like to graph both the Error Rates as well as the Accuracy for each value of K. I think this is an easy way to help us determine what value of K to use. Obviously, we want to use the value of K so that our error is minimized, while our accuracy is maximized. Below, I have graphed the error rate vs. the values of K (figure 9).

Figure 9: Error Rate vs. K Value



As you can see, the error rate is at zero, until we increase the value of K to 3 or higher, at which point our error rate shoots up to 100%. This would heavily suggest that K=2 is the right value of K to use, but let's plot the accuracy vs. the values of K as well, just to be safe.

Figure 10: Accuracy vs. K Value

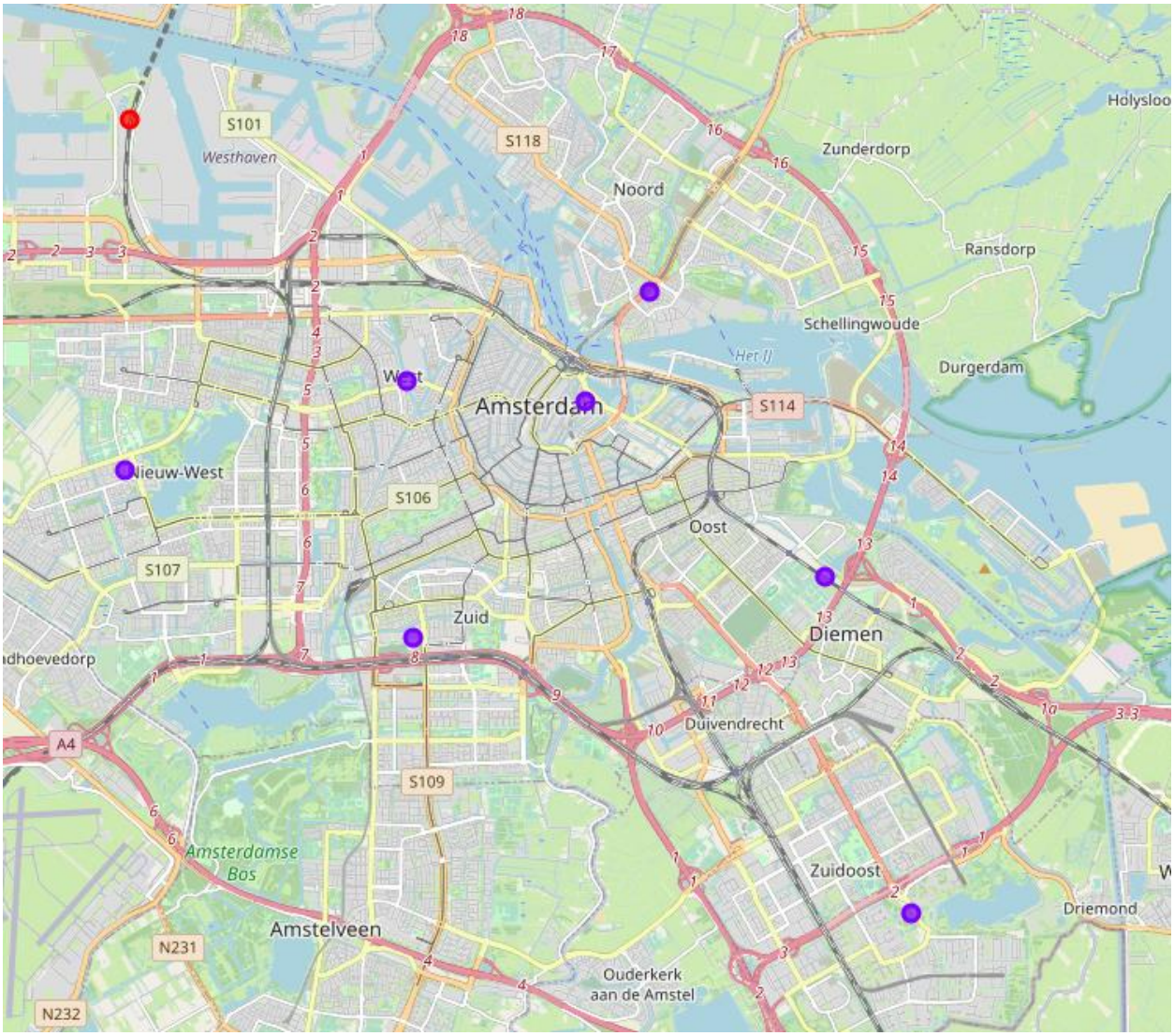


This graph (Figure 10) shows us exactly what we'd expect, having seen the error rate graph above: accuracy is at 100% until K=3, at which point it plummets down to 0%. As such, it seems to me incredibly obvious that K=2 is the ideal number to use here. Running the code to cluster the neighborhoods, we find the following results shown in the table below (Table 4): all districts fall into the same cluster, with the exception of Westpoort. Figure 11 shows the clustering on the map.

Table 4: Amsterdam District Variables - Clustered

	Neighborhood	Latitude	Longitude	Venues	Population Density	Safety Index	Safe	Cluster labels
0	Centrum	52.37321	4.903712	88	13748	69	1	1
1	Westpoort	52.41095	4.803871	2	10	0	1	0
2	Zuidoost	52.30465	4.974994	4	4391	132	0	1
3	West	52.37611	4.86452	17	15252	79	1	1
4	Nieuw-West	52.36407	4.802676	10	4478	130	0	1
5	Oost	52.34981	4.956049	6	7635	80	1	1
6	Zuid	52.34172	4.86605	9	9349	56	1	1
7	Noord	52.388	4.917663	13	2269	112	0	1

Figure 11: Clustered Districts in Amsterdam



4. Results

So, what have we found? Well, honestly, not that much. While the bar charts did show an interesting distribution of the variables, the scatter plots quickly showed us that we did not have enough data points and that there didn't seem to be any patterns immediately emerging. The OLS regression further proved that the findings were so statistically insignificant, that they really have no merit at all. The result that we did obtain from running the regression can be expressed in the following equation:

$$Safety\ Index = (0.0005 \times Population\ Density) - (0.157\ \times Venues) + 81.5631$$

Unfortunately, the R^2 value and P values for the regression showed us that, statistically speaking, these results are entirely insignificant. In fact, with the regression we have run, we are probably looking at around a <1% confidence interval, which I need not tell you is bad.

If we ignore the poor R^2 and P values, however, we do find that there is a potential negative effect of additional venues on crime, while population density has a minute positive effect. This seems to make sense, theoretically, as higher population density would almost certainly increase crime. More surprising is that the number of venues would decrease crime, perhaps because when there are things to do, people are less likely to loot or vandalize? Either way, I think that if the analysis was done with bigger, better, data we would potentially be able to confirm these correlations between key variables.

The clustering did show that Westpoort is uniquely different from the other districts in Amsterdam, and the graphs did show us that Centrum, while clearly the busiest and most popular district, was not put in its own cluster. As such, I think KNN clustering would be extremely interesting if we were to use the 42 smaller neighborhoods in Amsterdam, to see if all of the industrial neighborhoods in Westpoort would be grouped together, and what it would do to the remaining 7 districts' neighborhood. As of now, they have all been clustered together, but I imagine that if you were to increase K (which we did not do because our accuracy rates plummeted to zero past K=2) we would be able to further segment the smaller neighborhoods, which could lead us to finding distinct differences between more residential zones like Oost, and tourist hotspots like Centrum.

5. Discussion & Recommendations

When I initially chose Amsterdam to analyze for this project, I figured that segmenting the city into the eight common districts would be the best way to go. Having plotted the graphs and run the regression, however, I can see that that was a mistake. Each district in Amsterdam can be further segmented into smaller neighborhoods, which would give us 42 data points as opposed to the 8 that we were working with in this paper. I think that, to find any meaningful results, we need to have a much larger data set. Eight districts equate to eight data points, which is way too few for any rigorous analysis. Going up to 42 data points would significantly increase the amount of data we'd be working with, and would allow for us to find any potential correlations/patterns between variables. Granted, with Amsterdam being a relatively small city, and there being 42 separate neighborhoods, it would be very difficult to use the Foursquare API in such a way that fetching the data wouldn't have us overlapping in each tiny neighborhood. I was afraid we wouldn't find enough venues in each neighborhood if we divided it up into the 42 separate, smaller, neighborhoods, which is why I went with the 8 bigger districts instead. I still,

somewhat, wonder how effective the analysis would be with the 42 neighborhoods, as we'd need enough venues in each place for it to work.

It would be remiss of me not to also mention my concerns with Foursquare here, as part of the discussion. Having lived in Amsterdam for the better part of my life, I couldn't help but notice that a lot of the venues and venue locations returned by Foursquare simply were not correct. Venues such as the Kalverstraat, a busy shopping street in the Centrum part of Amsterdam, were listed as being in Oost, for example. Furthermore, some of the venues found by Foursquare were questionable at best. In Oost it found the Jaap Eden ice rink, which is indeed a popular ice skating rink located in the Oost district of Amsterdam, but it also found the Jaap Eden ice rink watchtower. This "venue" is simply a structure where security monitors the entire ice rink from, so it really should not have been included in the data fetch from Foursquare.

6. Conclusions

There is one very, very, obvious conclusion that can be drawn from this research. In fact, it is so obvious that none of this data gathering and analyzing was even necessary, but here it is anyway: an industrial work zone like Westpoort is NOT the best place to visit as a tourist, nor is it a place residents should look to live in. Using KNN to cluster the districts showed that Westpoort is uniquely distinct from the other districts in Amsterdam, which we had correctly assumed at the start of this paper, given its classification as an industrial work zone.

The regression showed us that there is a potential negative effect of additional venues on crime, while population positively affects the crime rate (only barely), but all of the values found were so insignificant that they really do not mean much of anything. The real conclusion that can be drawn here is that we needed far more data points to have any true meaningful analysis.

Ultimately, there are just far too few data points to really make any meaningful conclusions from the data. I do think that the linear regression we did has some merit, despite the large P values, as it does, hypothetically, make sense. As the population density in a place increases, crime is almost certain to increase. We found a positive coefficient so small that it was practically zero, but I believe that with more observations and better data, we would find a bigger, more significant, positive correlation between population density and the safety index.