URP 535, *Urban Informatics*
University of Michigan, Ann Arbor
Taubman College, Professor Xiaofan Liang
https://github.com/kfukutom/urban-informatics-final

Crime in Neighborhoods of Chicago
And Feasibility of Algorithmic Policing

Ken Fukutomi
Benjamin Spilo

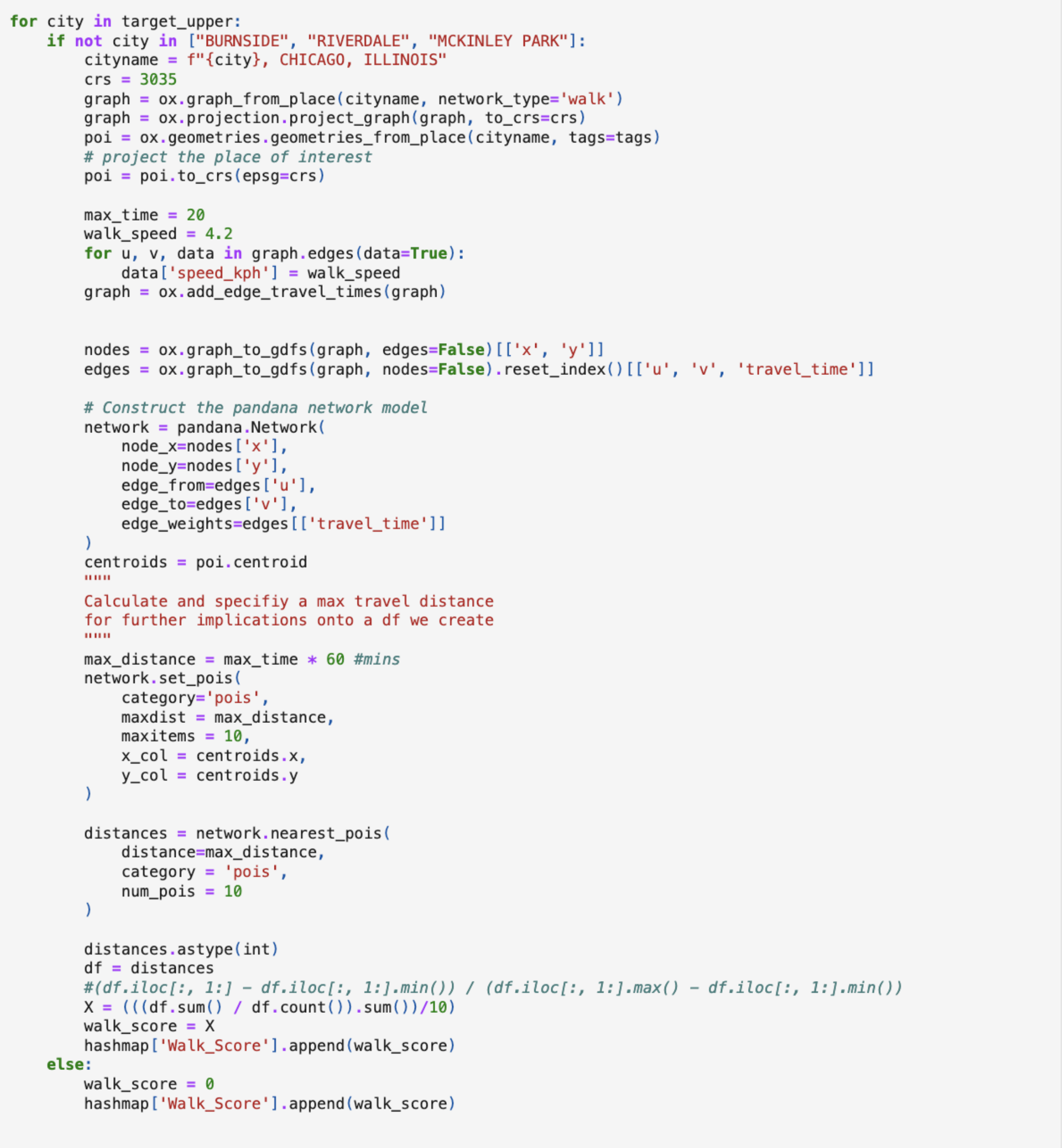
Abstract:

In contemporary society, the application of artificial intelligence and predictive machine learning models is rapidly expanding across various industries globally. This expansion has led to notable advancements in fields such as fintech, commercial real estate, medical diagnostics, and technology, which continue to grow daily. Such developments prompt a reevaluation of the potential benefits of integrating machine learning and artificial intelligence into urban planning and development. A recent initiative by AI researchers and urban planning PhD students from Tsinghua University in Beijing, China has highlighted this potential. The team successfully developed an urban planning system that has demonstrated capabilities surpassing those of traditional human urban planners. The system was designed based on the "15-minute city" model, incorporating previously successful human-designed projects and elements deemed beneficial, such as parks, bike paths, and entertainment areas. This burgeoning comprehension of data-driven applications in research and analysis, including the ability to predict, forecast, and model urban dynamics, has sparked interest in experimental applications. Along these lines, my colleague Ben and I were intrigued by the possibility of developing a scaled-down prototype version of this urban planning system, tailored to utilize specific urban data. To this end, we elected to employ a rich dataset from the city of Chicago's comprehensive database portal, which contains thousands of historical data entries related to various urban concepts and ideas. This dataset provided a foundation for our exploration into the feasibility of using machine learning to identify criminal activity hotspots, thereby aiding in the effective deployment of police patrols in high-crime areas. Our research aims to determine whether such technological applications can enhance urban safety without exacerbating existing urban environment issues. Nevertheless, to align our project with the curriculum of this course, we are concentrating on various categorical variables associated with each specific crime. We aim to identify factors that may increase the likelihood of certain crimes occurring. Additionally, we are incorporating course materials related to smart cities and data governance into our analysis.

POI Score Metric: OSMnx and Pandana

When downloading the initial .csv file from the open data portal of Chicago, Illinois, we were prompted with a very generic dataset, where each row represents a single crime case, with around >25 columns where each can be used to identify key concepts such as "Community Area/Code", "Data/Time", "Location Description", etc. With this dataset, in order to apply some additional content(s) relating to the sphere of the urban contexts of the cities, we initially asked some few key interests we had regarding some other categorical variables that could potentially play a role in impacting how "likely" a crime were to occur in a given area. Additionally, we would be able to feed more information into a feed-forward style machine learning model later on to explore some other quantitative results. But for now, we applied the use of Python's OSMnx, Pandana, and GeoPandas package in order to manually compute a relative walkscore of a given Chicago community area with its given name. To look more closely at the case, by taking a look at our jupyter notebook file in the repository under the name, walkscore_metric.ipynb, we manually compute the walkability index of a given neighborhood using its relative "closeness" to a location that fits under the criteria of GeoPandas' "amenity tags", which all identically serve as a place of interest in a neighborhood

< POI Network Calculator >



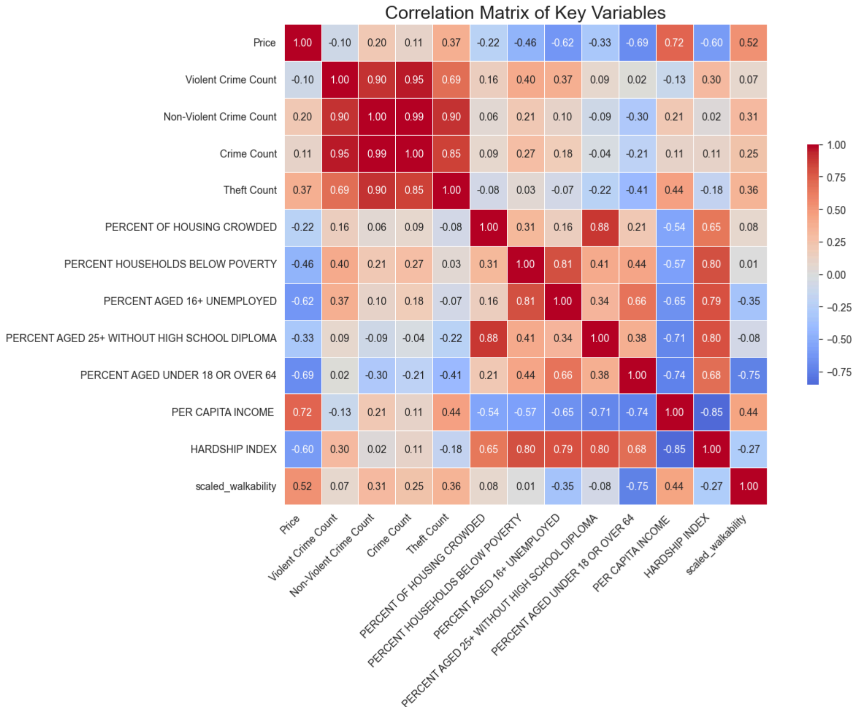
```
for city in target_upper:
    if not city in ["BURNSIDE", "RIVERDALE", "MCKINLEY PARK"]:
        cityname = f"{city}, CHICAGO, ILLINOIS"
        crs = 3035
        graph = ox.graph_from_place(cityname, network_type='walk')
        graph = ox.projection.project_graph(graph, to_crs=crs)
        poi = ox.geometries.geometries_from_place(cityname, tags=tags)
        # project the place of interest
        poi = poi.to_crs(epsg=crs)

        max_time = 20
        walk_speed = 4.2
        for u, v, data in graph.edges(data=True):
            data['speed_kph'] = walk_speed
        graph = ox.add_edge_travel_times(graph)


        nodes = ox.graph_to_gdfs(graph, edges=False)[['x', 'y']]
        edges = ox.graph_to_gdfs(graph, nodes=False).reset_index()[['u', 'v', 'travel_time']]

        # Construct the pandana network model
        network = pandana.Network(
            node_x=nodes['x'],
            node_y=nodes['y'],
            edge_from=edges['u'],
            edge_to=edges['v'],
            edge_weights=edges[['travel_time']]
        )
        centroids = poi.centroid
        """
        Calculate and specifiy a max travel distance
        for further implications onto a df we create
        """
        max_distance = max_time * 60 #mins
        network.set_pois(
            category='pois',
            maxdist = max_distance,
            maxitems = 10,
            x_col = centroids.x,
            y_col = centroids.y
        )

        distances = network.nearest_pois(
            distance=max_distance,
            category = 'pois',
            num_pois = 10
        )

        distances.astype(int)
        df = distances
        #(df.iloc[:, 1:] - df.iloc[:, 1:].min()) / (df.iloc[:, 1:].max() - df.iloc[:, 1:].min())
        X = (((df.sum() / df.count()).sum())/10)
        walk_score = X
        hashmap['Walk_Score'].append(walk_score)
    else:
        walk_score = 0
        hashmap['Walk_Score'].append(walk_score)
```

.

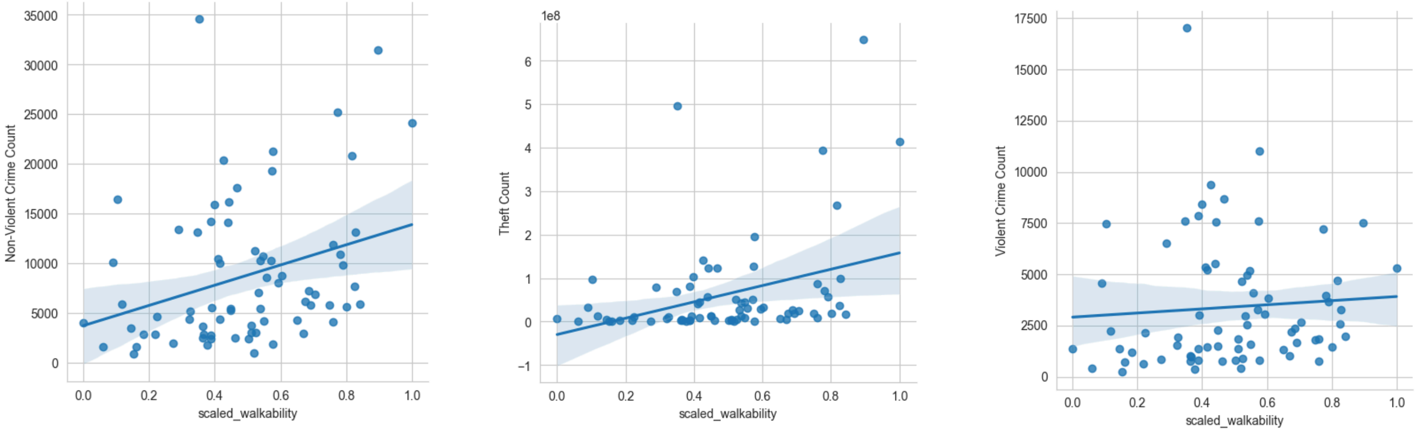
For every Chicago neighborhood NOT named like the ones listed in the second line of the function, we essentially skipped, since the location did not have an adequate amount of information regarding the node. To explain what we did in this code: We began by iterating through a list of cities, excluding "BURNSIDE," "RIVERDALE," and "MCKINLEY PARK." For the remaining cities, we appended ", CHICAGO, ILLINOIS" to each city name and used this to retrieve pedestrian network data and points of interest (POI) from OpenStreetMap, setting the walking speed to 4.2 km/h for the network. We then projected the network and POI data to coordinate reference system 3035, calculating and assigning travel times for enhanced network analysis. Using the pandana library, we constructed a network model to analyze accessibility, calculating distances to the nearest POIs within a 20-minute walk. We summarized the accessibility through a 'Walk Score,' which we stored for each city, setting a score of zero for excluded cities. To further refine our analysis, we adjusted the initially calculated walk scores to the median value from our dataset and researched the walk score metric online to align our scores with established standards. During this process, we noticed that O'Hare Airport had an unreasonably high walk score due to the numerous amenities it houses, despite the general area being relatively unwalkable. Consequently, we revised the walk score for O'Hare Airport to better reflect the actual walkability of its surrounding community area. This adjustment was also incorporated later in the code. Finally, we applied each individual score with a scaled version, as for additional computation, and then processed the result into a .csv file and utilized it back in the main jupyter notebook file. Ultimately though, this metric better represents a city's relatedness to a POI, how close one is to a place with amenities. Hence, we decided later on that "scaled_walk" should be more of a representation of how close each Chicago Community is to a place of interest. We would then later on merge the scaled walkability to our main dataframe, back in the main jupyter notebook file to conduct some further applications. The main reason for analyzing this case was to understand how places of interest and social spaces, referred to as 'third urban spaces' within our urban environment, could experience an increase in criminal activities such as sexual assault, theft, and assault. These areas are likely to attract larger gatherings of people engaging in conversation, having drinks, and socializing, which could potentially make them targets for criminal behavior or locations where violent activities occur due to the increased social interaction.

Exploratory Data Analysis

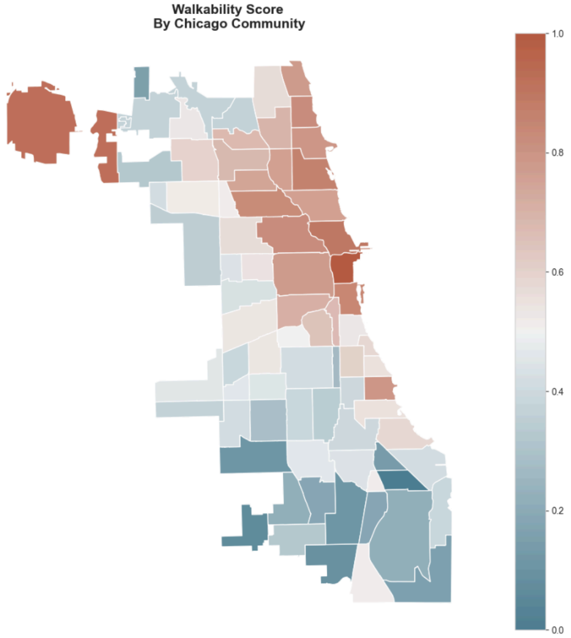
We sourced our data from a variety of locations, then merged it and did preliminary analysis to discover what areas of interest there might be. We used the Geopandas Chicago dataset to get a basic overview of the layout of the city, along with population values and neighborhood names. We then merged this dataset with a number of crime related datasets from the Chicago Open Data Portal, our POI data, and housing price data from Zillow divided by Chicago neighborhoods. Based on this correlation matrix above, which we had found to be extremely useful, we found that many variables in regards to Chicago community's demographics tended to correlate positively with other variables in regards to demographics /income/poverty. We then did a preliminary analysis to determine what data points correlated with each other in significant ways. While there were a number of predictable outcomes, such as higher income people living in more expensive neighborhoods, there were some outcomes that merited further investigation. These included negative correlations between crime and the ratio of children and elderly in a neighborhood and between walkability and the number of elderly and children living in a neighborhood. We also decided to split up the types of crimes between violent, non-violent, and theft specifically to get more specific insights into who and where is affected by which types of crimes. In the preliminary analysis, we found a correlation between how expensive a neighborhood is, and the amount of theft committed there. However, price had much less of an impact on the amount of violent and overall crime. We also compared walkability and various types of crime. As shown below, there was a visible correlation between walkability and non-violent crimes or theft, but significantly less of one between walkability and the amount of violent crime in a neighborhood.



	Price	Violent Crime Count	Non-Violent Crime Count	Crime Count	Theft Count	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER CAPITA INCOME	HARDSHIP INDEX	scaled_walkability
Price	1.00	-0.10	0.20	0.11	0.37	-0.22	-0.46	-0.62	-0.33	-0.69	0.72	-0.60	0.52
Violent Crime Count	-0.10	1.00	0.90	0.95	0.69	0.16	0.40	0.37	0.09	0.02	-0.13	0.30	0.07
Non-Violent Crime Count	0.20	0.90	1.00	0.99	0.90	0.06	0.21	0.10	-0.09	-0.30	0.21	0.02	0.31
Crime Count	0.11	0.95	0.99	1.00	0.85	0.09	0.27	0.18	-0.04	-0.21	0.11	0.11	0.25
Theft Count	0.37	0.69	0.90	0.85	1.00	-0.08	0.03	-0.07	-0.22	-0.41	0.44	-0.18	0.36
PERCENT OF HOUSING CROWDED	-0.22	0.16	0.06	0.09	-0.08	1.00	0.31	0.16	0.88	0.21	-0.54	0.65	0.08
PERCENT HOUSEHOLDS BELOW POVERTY	-0.46	0.40	0.21	0.27	0.03	0.31	1.00	0.81	0.41	0.44	-0.57	0.80	0.01
PERCENT AGED 16+ UNEMPLOYED	-0.62	0.37	0.10	0.18	-0.07	0.16	0.81	1.00	0.34	0.66	-0.65	0.79	-0.35
PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	-0.33	0.09	-0.09	-0.04	-0.22	0.88	0.41	0.34	1.00	0.38	-0.71	0.80	-0.08
PERCENT AGED UNDER 18 OR OVER 64	-0.69	0.02	-0.30	-0.21	-0.41	0.21	0.44	0.66	0.38	1.00	-0.74	0.68	-0.75
PER CAPITA INCOME	0.72	-0.13	0.21	0.11	0.44	-0.54	-0.57	-0.65	-0.71	-0.74	1.00	-0.85	0.44
HARDSHIP INDEX	-0.60	0.30	0.02	0.11	-0.18	0.65	0.80	0.79	0.80	0.68	-0.85	1.00	-0.27
scaled_walkability	0.52	0.07	0.31	0.25	0.36	0.08	0.01	-0.35	-0.08	-0.75	0.44	-0.27	1.00



Statistical Inferences with Our Observations:

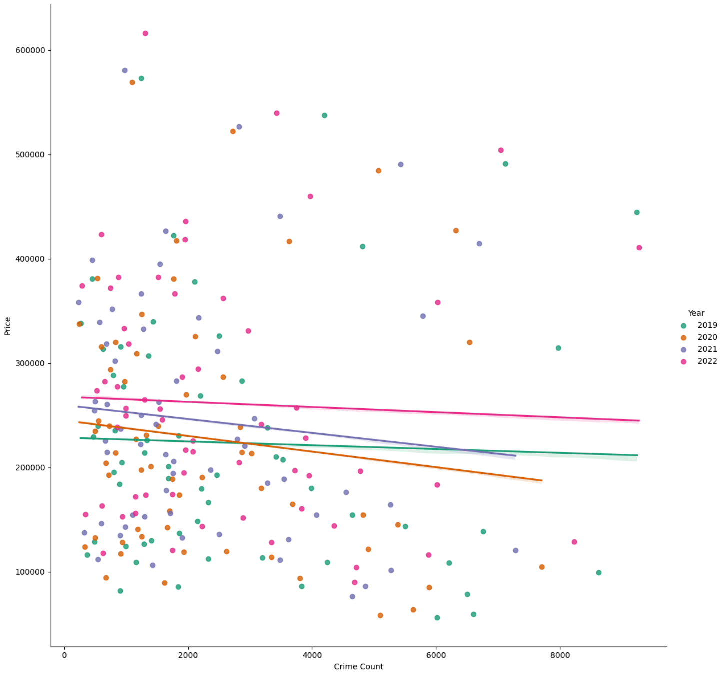


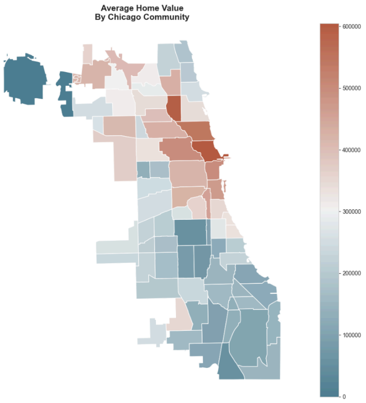
In our `eda.ipynb` notebook, we have provided several interpretations, but here we want to highlight some particularly interesting observations. Our analysis revealed that social places with crime cases in Chicago are not predominantly violent. Specifically, using our computed POI (Point of Interest) closeness data for each Chicago community, we found that locations with higher walkability scores tend to have more non-violent crimes. This suggests that government officials in Chicago could use this data to strategically patrol areas with frequent gatherings, anticipating higher crime rates. However, it is important to note that violent crimes often occur in more isolated buildings and areas with a history of criminal activity. Thus, we should avoid assuming that criminal activity is exclusively concentrated in areas with high POI closeness. Our analysis also showed that affluent areas, particularly those near the shore and in the northern parts of the neighborhood, are where theft-related crimes are most common. These areas, characterized by better public infrastructure and funding, attract more businesses and projects due to potential revenue and interest. Given this, it would be prudent to extend patrolling in these neighborhoods beyond dusk, carefully balancing the need to prevent crime with the risk of over policing. Furthermore, we acknowledge that some crimes may be driven by unpredictable factors, such as someone committing assault under the influence at a bar, which are beyond our ability to predict or control. Finally, we were pleased to calculate neighborhood walkability scores using OpenStreetMap data. These scores not only reflect the physical attributes of the neighborhoods but also help us understand patterns of crime and safety during peak hours.

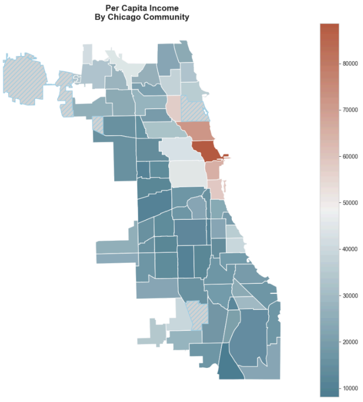
—---

Housing Price Indices: {Rent, Home Ownership in Chicago}

Going back to the idea of our main dataset not having some variables beyond our interest, we had to manually scrape and extract additional .csv file data from Zillow/Redfin in order to compute some basic analysis that went into more of the detail behind the relationship between property price indices and relative crime frequency, using a merged dataset of crime spanning from the years 2019 through 2022. After collecting this data, we did preliminary analysis on the relationship between crime and housing price in a neighborhood. We found that there was a slight negative correlation between the crime in a neighborhood and the price of houses in that neighborhood. This indicates that as the price of housing in a neighborhood rises, the amount of crime in that neighborhood tends to decrease. This held true during the pandemic years, which included large spikes in crimes committed during the 2020 riots. This indicates that the riots may have affected higher income or more expensive areas less. However, Zillow's data may not be entirely accurate, and some of these correlations may be self reinforcing. Zillow's Zestimate calculations are designed in such a way that they may not be fully accurate. Zillow's design often doesn't represent a number of other factors about a neighborhood, beyond comp house prices and sales that are publicly available. Zillow's design pushes users into a certain way of thinking about houses, and the way they decide on their prices is a significant factor in that design. As Loukissas explains, Zillow attempts to draw people into using their service by pushing ease of use over accuracy. In our investigation, we found that Zillow had the most accessible, comprehensive estimate of house price by neighborhood. We generally determined this to be accurate, with most neighborhoods in the city center, with higher population and population density having higher housing prices, while houses in the suburbs far from amenities tended to be less expensive. Home prices definitely did see a surge in prices as it came closer to the North, and in contrast the prices decreased as it went down South. Interestingly, most of the home prices correlated well with the walkability scoring metric. We also calculated that there could be that instance regarding public wealth and disparity of taxpayer income that could afford to fund certain community amenities such as parks, compared to others.

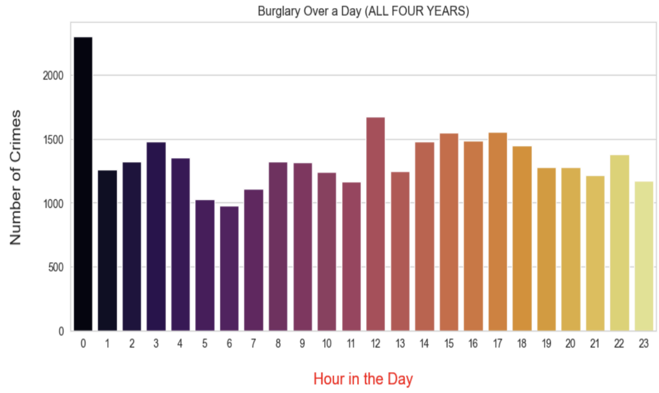


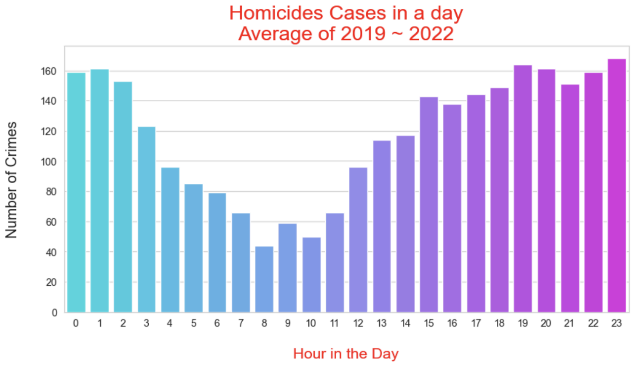




Time of Day

Another component we were interested in investigating was the significance of time in relation to the frequency of criminal activity in a given area. Specifically, we explored the importance of identifying which hour of the day exhibits more signs of criminal activity. For instance, in our first time-series analysis, we examined burglary cases within Chicago, Cook County. We discovered that most burglary-related cases occurred at midnight, with trends changing over a span of four years, yet consistently indicating that burglaries were more frequent at night. This finding aligned with our previous observations that criminal activity occurred more often across all areas of Cook County compared to violent crimes, which tended to concentrate in close-knit residential neighborhoods, suggesting a potential connection to gang-related activities. For now, our output appeared as follows:



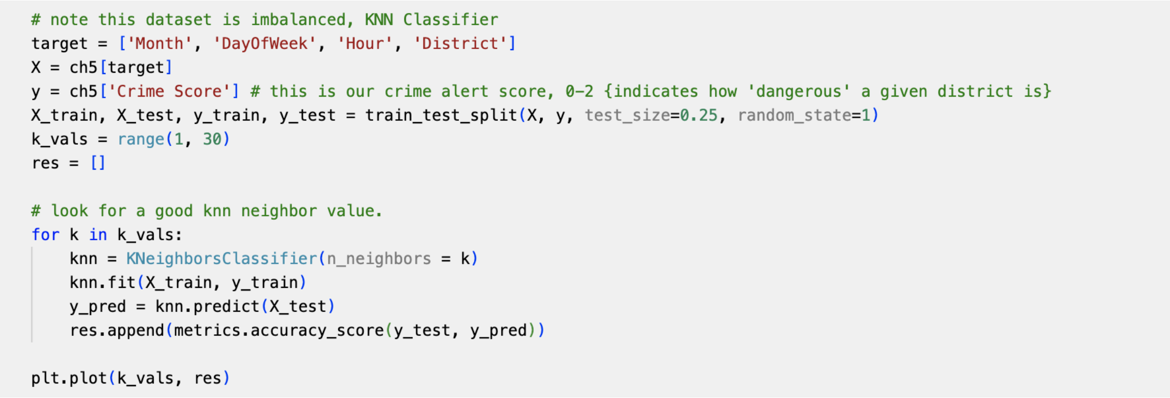


We found that, based on our time-series data, criminal activities such as homicides tend to occur after sundown, during the darker hours in cities. For example, we investigated numerous criminal activities related to narcotics (drug-related activities) and gun violence to determine which hours show a trend or spike in criminal activity. From this analysis, we observed that categories of crime related to homicide and violent crime tend to have a higher occurrence during nighttime. Hence, would it be reasonable to increase policing activity at dusk, compared to the brighter hours of the day, in order to mitigate crimes such as theft that may occur during these hours? We have conducted extensive time-based analysis, detailed in the main Jupyter notebook file. For now, however, we want to shift our focus to the broader concept of time itself and examine a four-year period to see whether we can forecast crime counts and use machine learning to assist government officials in gaining novel insights into crime patterns. Initially, we observed that the main characteristics of the crime data highlight periodicity and variance, with the graph showing a fluctuating trend over all four years and a sudden spike in 2020. This increase in reported crime could be attributed to a sharp rise in riot-related activities following the George Floyd protests, particularly noting the timing of the violent surge, which began in the last week of May, just days after George Floyd's death. Another point to mark is the nature of the data not being stationary all throughout the time-series data's period. To fix this, we moved forward with a decision to use an ARIMA model. We first analyzed the time series data for stationarity, as well as stationarity in the residuals of the time series. Then, we went onto identifying noticeable patterns and seasonality within the data, by performing some seasonal decomposition of the time series data. Within multiple trials of the process, stationarity was hard to achieve, hence leading us to learn more about how to create an ARIMA model which could be better to make predictions and forecast data, minimizing our errors, with the RMSE score. To explain simply, **the lower our RMSE score, meant that our model performance was better.** We aimed to evaluate the validity of a model and explore its construction based on data patterns and seasonality to potentially enhance policing at specific times in future years. To achieve this, we utilized resources from YouTube to deepen our understanding of these concepts and apply them to the fields of urban informatics and data science. Though we found this process to be the most challenging, in the end our model came out to be a very stationary version of the time series data, which was based on a Date-Month index. Results are in the main notebook.

Machine Learning and Predictive Policing

We found that crime is, generally, hard to predict using existing data. While there were some factors that had correlation with crime, they often were based on economic factors. The most prominent correlations we found in our exploratory analysis, which held true through our machine learning application, was that economic factors had the largest impact on crime. Although we had some models performing at a baseline accuracy score beyond what we would consider to be good ~70%, much of the dependent variables that we fed into the model helped to most likely identify some of the interests that arose. One type of classification model we used was the KNN Neighbors Classification model, which we thought to be perfect to find variables that are similar, knowing that some crimes could tend to cluster in certain regions compared to other(s). With that, we constructed some models of our own.

```
< Classifier Accuracy/F1 Score ~ 0.7064178835406526 >
```



```
# note this dataset is imbalanced, KNN Classifier
target = ['Month', 'DayOfWeek', 'Hour', 'District']
X = ch5[target]
y = ch5['Crime Score'] # this is our crime alert score, 0-2 {indicates how 'dangerous' a given district is}
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=1)
k_vals = range(1, 30)
res = []

# look for a good knn neighbor value.
for k in k_vals:
    knn = KNeighborsClassifier(n_neighbors = k)
    knn.fit(X_train, y_train)
    y_pred = knn.predict(X_test)
    res.append(metrics.accuracy_score(y_test, y_pred))

plt.plot(k_vals, res)
```

We wanted to figure out the best nearest neighbors value for the most optimal prediction and accuracy metric, so we first plotted it as an independent matplotlib graph, from there we would compute the score like anybody else and make the graphs and use it as a basis for our best n_neighbors variable later on.

To evaluate the overall performance of our models, we encountered challenges when processing raw, unbalanced data formats. Specifically, we attempted to predict the 'Primary Type' of crime using geolocation details such as longitude, latitude, and Community Name/Area. The resulting accuracy score was approximately 0.4, indicating that the models did not perform as well as anticipated. This project has been invaluable in deepening our understanding of these techniques, particularly within the context of urban environments. It has highlighted the limitations of relying solely on machine learning algorithms for policing decisions. The use of data and trends cannot definitively dictate policing strategies in specific areas. Instead, it is crucial for human analysts to interpret and contextualize the data, recognizing the inherent biases and limitations that exist within purely numerical analyses. This approach ensures a more nuanced and effective application of predictive policing, where human judgment plays a central role alongside algorithmic insights. Policing based on datasets, in theory, seems feasible, but it risks becoming a self-fulfilling prophecy. This occurs when the data continually highlights high-crime areas, leading to increased policing in those locations. As a result, more crimes are reported in these areas, which the models then input as new data. This cycle perpetuates itself, causing the model to continuously validate its own predictions, effectively creating a never-ending loop of heightened surveillance and crime reporting. This feedback loop demonstrates the need for cautious and critical use of data in predictive policing.

Conclusion

Cutting into the reading topic from Yanni Alexander Loukissas' "Thinking Critically in a Data-driven Society" (2019), we wanted to integrate the finds of that reading into our understanding of predictive policing, especially in the context of our earlier findings on crime and POI closeness in Chicago, emphasizes the importance of local context in interpreting data-driven predictions. Loukissas argues that data is inherently local and deeply contextual, a perspective that is crucial when considering predictive policing strategies. When applying predictive policing techniques that use data on non-violent crimes in locations with high walkability or POI closeness, it's important to critically assess the local factors influencing these data points. For example, affluent neighborhoods might show different crime patterns not just because of their economic status but also due to their specific local characteristics, such as public infrastructure, historical crime rates, and community engagement levels. Understanding that data reflects local circumstances helps in tailoring policing strategies that are not only responsive but also respectful of community dynamics. For instance, while data might suggest increased patrols in areas with high foot traffic and affluent populations to mitigate theft, a nuanced approach is required to avoid over policing, which can strain community-police relations. Furthermore, recognizing that certain crime spikes might be driven by unpredictable local variables (like social events or temporary gatherings) underscores the necessity for flexible policing that can adapt to real-time assessments of local conditions. Thus, integrating Loukissas's perspective on the locality of data with our findings about crime and POI closeness in Chicago can enhance the effectiveness and appropriateness of predictive policing. It highlights the need for police strategies that not only use data-driven insights but also **incorporate an understanding of the specific social and environmental contexts of each neighborhood**. This approach ensures that policing strategies are more equitable and contextually informed, potentially leading to safer and more harmonious community interactions. In conclusion, our research highlights the complexity of factors influencing crime occurrence. While statistical analysis reveals various correlations, it is crucial to remember that correlation does not imply causation. Assumptions about crime in specific areas based solely on data can be misleading. The volatile nature of the data underscores the necessity of a nuanced approach. While machine learning and data analytics can simplify some aspects of analysis, they should not be the sole basis for making judgments. This project has been highly educational, enhancing our understanding of both the potential and the limitations of these technological tools.

Bibliography:

Artyushina, A. (2020). Is Civic Data Governance the Key to Democratic Smart Cities? The Role of the Urban Data Trust in Sidewalk Toronto. Telematics and Informatics, 55, 101456.

Latham, A., & Layton, J. (2019). Social Infrastructure and the Public Life of Cities: Studying Urban Sociality and Public Spaces. Geography Compass, 13(7), e12444. https://compass.onlinelibrary.wiley.com/doi/full/10.1111/gec3.12444 (open-access)

Liang, X., Baker, J., DellaPosta, D., & Andris, C. (2023). Is Your neighbor Your Friend? Scan Methods for Spatial Social Network Hotspot Detection. Transactions in GIS, 27(3), pp.607-625.

Loukissas, Y. A. (2019). All Data Are Local: Thinking Critically in a Data-driven Society (Chapter 5, p.130-140). MIT press.

Mattern, S. (2021). A City Is Not a Computer: Other Urban Intelligences (Chapter 1 and 2, p.18-72). Routledge.

Meijer, Albert, and Martijn Wessels. "Predictive policing: Review of benefits and drawbacks." International Journal of Public Administration, vol. 42, no. 12, 12 Feb. 2019, pp. 1031–1039, https://doi.org/10.1080/01900692.2019.1575664.

Psyllidis, A., Gao, S., Hu, Y., Kim, E. K., McKenzie, G., Purves, R., Yuan, M. & Andris, C. (2022). Points of Interest (POI): A Commentary on the State of the Art, Challenges, and Prospects for the Future. Computational Urban Science, 2(1), 20. https://link.springer.com/article/10.1007/s43762-022-00047-w (open-access)

Scott, J. C. (1998). Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed (Chapter 2, p.53-63). Yale University Press.

Group Contribution:
Benjamin: Report, Exploratory Analysis, Data Collection
Ken: Report, Exploratory Analysis, Machine Learning, Data Collection

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data (Chicago Open Data Portal for Crime, Census Data)

https://www.zillow.com/research/data/ (Zillow for Scraping Property Indices, Getting Listing Prices)

https://www.openstreetmap.org/ (OSMnx Amenities Tag References)

https://osmnx.readthedocs.io/en/stable/ ← OSMnx Documentation, alongside Pandana