NYC Airbnb House Price vs Crime

Team member name 1: Keye Chen (keyechen); section (001)
Team member name name 2: Fengyu Zhang (fyzhang); section (001)
Team member name name 3: Boyan Wu (wuboyan); section (001)

1. Motivation

1.1 Abstract

When people choose their accommodations, they weigh numerous factors, with cost and safety topping the list. In this project, we are delving into the dynamics of Airbnb pricing and crime across New York City's vibrant neighborhoods

1.1.1 Demands

- Our inquiry is driven by the need to understand how Airbnb prices are influenced by crime rates, Concurrently, the project aims to equip renters with a knowledge base to make informed accommodation choices, and city planners and law enforcement to understand the relationship between tourism accommodations and crime, effectively bridging the gap between cost and security.

1.1.2 Questions

- The central question guiding this research is: what variations in Airbnb house prices can be partly attributed to the perceived safety or crime statistics of different neighborhoods? What are the driving factors behind our analysis and how can these insights be utilized to optimize the findings and performance around geographic regions in NYC?

1.1.3 Goal

- This study aims to answer questions like the nature of the correlation between Airbnb house prices and crime rates, identify neighborhoods with significant discrepancies, and predict trends in these metrics. The goal is to provide actionable insights and recommendations for various stakeholders.

2. Data Sources

2.1 Source 1: NYC-Airbnb-2023.csv

2.1.1 Location

- The dataset is obtained from Kaggle (https://www.kaggle.com/datasets/godofoutcasts/new-york-city-airbnb-2023-public-data/data)

2.1.2 Format: CSV (6.6 MB)

2.1.3 Important Variable Description

- ‘neighborhood_group’, ‘neighborhood’, ‘latitude’, ‘longitude’: containing the geographical information about each record.
- ‘price’: the amount needed to live in for one night.

- ‘minimun_nights’, … ‘availability_365’: containing the information about the hardware of the room and the viewing statistics of the guests on the website.

2.1.4 Records Num & Time Period

- The dataset contains 42931 rows × 18 columns, describing the Airbnb statistics in NYC for 2023.

2.2 Source 2: NYC_crime.csv

2.2.1 Location

- The dataset is obtained from Kaggle (https://www.kaggle.com/datasets/ajkarella/nyc-crime-stats)

2.2.2 Format: CSV (688.2 MB)

- **Since the dataset is too large, in our zip file, we will only include the first 100 rows of crime.csv and NYC_crime.csv.**

2.2.3 Important Variable Description

- ‘latitude’, ‘longitude’: containing the geographical information about each record.
- ‘pd_desc’: Description of internal classification corresponding with PD code (more granular than Offense Description).
- ‘ofns_desc’: Description of offense corresponding with key code.
- ‘law_cat_cd’: Level of offense: felony, misdemeanor, violation.
- ‘perp_race’: Racial types of the criminals.

2.2.4 Records Num & Time Period

- The dataset contains 3881989 rows × 18 columns, describing the crime statistics in NYC from 2006 to 2019.

2.3 Source 3: fullDownload.geojson

2.3.1 Location

- The dataset is obtained from https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58&text=downloads

2.3.2 Format: GEOJSON (17.8MB)

2.3.3 Important Variable Description

- ‘state’, ‘city’, ‘name’: containing brief geographical information about the neighborhood in America.
- ‘geometry’, ‘coordinates’: containing detailed geometrical information about each neighborhood.
- ‘holc_grade’: Overall description of the dangerous extent of each neighborhood.

2.3.4 Records Num & Time Period

- The dataset contains 8878 records, describing the geographical statistics all over American neighborhoods.

3. Data Manipulation Methods

3.1 Source 1: NYC-airbnb-2023.csv

3.1.1 Initial Process

- Data was loaded from the NYC-Airbnb-2023.csv file. We first checked for data types and null values of each column. We dropped several columns that have little relevance to our main focus.

3.1.2 Locate each Airbnb record and add neighborhood_cd column

- See spec in 3.3.3.
  Handling incorrect and missing values:
  We first replace -1 in the neighborhood code column with NaN and drop these rows, since -1 means these records don't belong to any one of listed New York neighborhoods and will not contribute to our analysis. For the left regional missing values, they will automatically be omitted after we establish the new data frame.

3.2 Source 2: NYC_crime.csv

3.2.1 Initial Process

- Data was loaded from the nyc_crime.csv file. We first checked for data types and null values of each column. Then, we rename the last four ambiguous regional columns' titles to community district, borough boundaries, city council districts and police precincts.

3.2.2 Create new neighborhood-centered crime dataframe

- **Workflow:** New York city has 300+ neighborhoods, and we will assign each crime to its corresponding neighborhoods coded from 1-300+. Then, we will calculate crime statistics in each neighborhood and compress the 700MB giant dataset into 300+ rows. The new dataframe will be indexed by neighborhood code, followed by various columns statistically evaluating crime features in that neighborhood area. Finally, we will right merge the new dataframe with airbnb by the shared neighborhood code column.
- **Challenge1:** Locate each crime and add neighborhood_cd column
  See spec in 3.3.3.
  Handling incorrect and missing values:
  We first replace -1 in the neighborhood code column with NaN and drop these rows, since -1 means these crimes don't belong to any one of listed New York neighborhoods and will not contribute to our analysis. For the left regional missing values, they will automatically be omitted after we establish the new data frame.
- **Challenge2:** Adding crime features for each coded neighborhood
  column 1. annual crime rate
  We first use datetime and add a year-wise column for each crime. Then, we group the data frame by neighborhood code and year, counting the number of crimes, unstacking the data frame and calculating the annual crime rate by mean method.
  column 2-5. misdemeanor_rate, felony_rate, violation_rate, infraction_rate

Similar to 1, this time we group the data frame by neighborhood code and crime level including felony, misdemeanor, violation and infraction, calculating their annual happening frequency for each neighborhood.

column 6. dominant_ofns

We group the data frame by neighborhood code and specific crime category, counting the frequency, sorting the value to find out and keep the most frequent crime type in each neighborhood.

column 7. dominant_perp_race

Same as column 6, we seek to find out and keep the most frequent criminal's race in each neighborhood.

3.3 Source 3: fullDownload.geojson

3.3.1 Initial Process

- Data was loaded from the 'fullDownload.json'. Based on the features of the geojson file, we initialized the data frame using 'pd.read_json', and extracted the 'features' series. Then based on the 'state' and 'city'(see 2.3.3), we got the filtered series which only contains data in NYC.

3.3.2 Create class list

- We then defined a class called 'NYCDistrict' to store the geographical information of each district. Then we created a list called 'Districts' to store all the NYC districts. Now we had all the shape information of NYC districts.
- **Challenge:** However, we found that there are a great number of districts that don't have the name in the geojson file. We decided to use the order in the 'Districts' to distinguish the NYC district. The 'neighborhood_cd' column's process in the above two sources is based on this idea.

3.3.3 'neighborhood_cd' column process

- Based on the preexisting 'longitude' and 'latitude' value in 'NYC-airbnb-2023.csv', 'NYC_crime.csv', and the 'coordinates' information in 'NYCDistricts' which could form polygons, we assign the order of the district stored in 'Districts' to each record if the 'longitude' and 'latitude' of the record falls into the polygon. We assign -1 to records that don't fall into any polygon.

4. Analysis

After bringing together two different data resources, we are able to investigate further and find a new insight that could not have been answered with either data resource alone. Our analysis originated from three sections: EDA, a significant correlation between crime and house price and segmentation & clustering.

4.1 EDA(Exploratory Data Analysis)

4.1.1 Data Overview of Airbnb Crime

- In the process of data analysis, our data has merged both the crime and Airbnb datasets and conducted the initial cleaning of the missing value(See Part 3.2.2). In the Airbnb Crime dataset, values type consisted of int, object, and float. Values such as outliers(price, annual crime rate, etc) have been dropped by setting the parameters in the foreseeable figure or changing the percentile of the data to meet and grasp the critical aspects of the outcome.

4.1.2 Dominant Offenses & Price Range

- Given the Airbnb crime dataset, the visualization of all types of crime has been collected with respect to price segmentation. The 0-100 price range seems to have the highest overall count of offenses across all price ranges, suggesting a potential hotspot for various criminal activities. Higher price ranges (301-500 and above) have fewer total offenses, which could imply a correlation where more expensive areas have lower crime rates, or fewer listings in those price ranges are in high-crime areas. Specific offenses like "controlled substance, possession" and "marijuana, possession" are prevalent across multiple price ranges, which may reflect city-wide patterns in drug-related offenses.

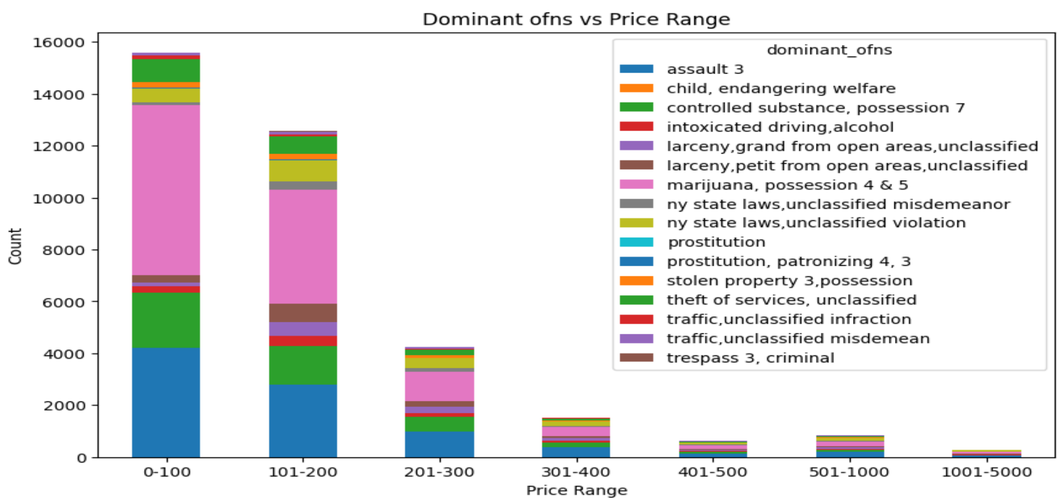

Fig4.1.2-1. Dominant ofns vs Price Range

4.2 Significant Correlation between Crime and House Price

4.2.1 Dominant Crime Type & Price

- Firstly, we want to explore whether the price distributions within each dominant crime type category will vary with each other. In order to better understand the significant difference, we first parse the airbnb_crime dataset into four subsequent datasets by room_type, the aim of which is to eliminate the prime influence by room types. However, since the number of items in shared room (465) and hotel room (107) is too small compared with private room (15244) and entire room (19891), we will only show two box plots where x-axis is different dominant crimes and y-axis is the price:

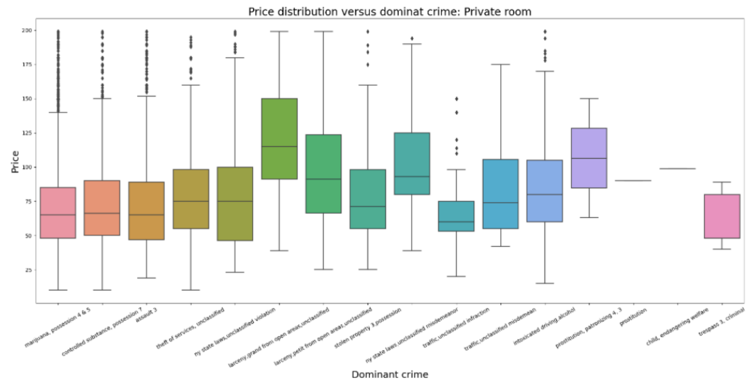

Fig4.2.1-1. Price distribution versus dominant crime: Private room

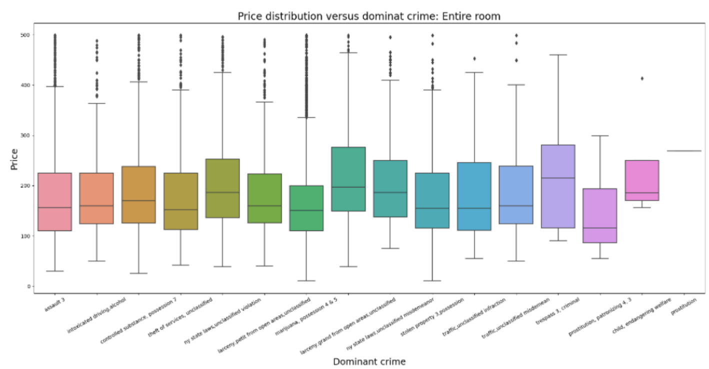

Fig4.2.1-2. Price distribution versus dominant crime: Entire room

According to Fig4.2.1-1 and Fig4.2.1-2, we can see that for some rooms lying in certain types of dominant crime neighborhoods, such as larceny and ny state law violation, the price distribution and their mean prices are evidently different from others. For further statistical proof, we conduct pairwise Tukeyhsd analysis as follows (Since the output is too long, we will present a part of the screenshot. The whole output can be referred in the combined_analysis.ipynb):

• private_room

```
res2 = pairwise_tukeyhsd(private_room['price'], private_room['dominant_ofns'])
res2.summary()
```

✓ 1.7s Python

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
assault 3	child, endangering welfare	19.7362	1.0	-158.1272	197.5995	False
assault 3	controlled substance, possession 7	-1.456	0.9997	-6.2526	3.3405	False
assault 3	intoxicated driving,alcohol	25.3322	0.0	13.7296	36.9347	True
assault 3	larceny,grand from open areas,unclassified	76.3815	0.0	66.379	86.384	True
assault 3	larceny,petit from open areas,unclassified	37.0718	0.0	26.1235	48.0202	True
assault 3	marijuana, possession 4 & 5	-3.4301	0.0927	-7.0772	0.2169	False
assault 3	ny state laws,unclassified misdemeanor	43.407	0.0	29.1048	57.7091	True
assault 3	ny state laws,unclassified violation	20.5674	0.0	12.5119	28.6229	True
assault 3	prostitution	10.7362	1.0	-115.0483	136.5206	False
assault 3	prostitution, patronizing 4, 3	27.2362	1.0	-98.5483	153.0206	False
assault 3	stolen property 3,possession	12.9436	0.0655	-0.3357	26.2229	False
assault 3	theft of services, unclassified	13.8395	0.0	7.2923	20.3866	True
assault 3	traffic,unclassified infraction	-9.2986	0.8793	-26.1251	7.5278	False
assault 3	traffic,unclassified misdemean	33.7568	0.0	15.4766	52.0369	True
assault 3	trespass 3, criminal	-18.2638	1.0	-97.8475	61.3198	False
child, endangering welfare	controlled substance, possession 7	-21.1922	1.0	-199.0746	156.6902	False
child, endangering welfare	intoxicated driving,alcohol	5.596	1.0	-172.5998	183.7918	False
child, endangering welfare	larceny,grand from open areas,unclassified	56.6453	0.9995	-121.4535	234.7442	False
child, endangering welfare	larceny,petit from open areas,unclassified	17.3357	1.0	-160.8188	195.4901	False
child, endangering welfare	marijuana, possession 4 & 5	-23.1663	1.0	-201.0214	154.6888	False
child, endangering welfare	ny state laws,unclassified misdemeanor	23.6708	1.0	-154.7212	202.0628	False
child, endangering welfare	ny state laws,unclassified violation	0.8312	1.0	-177.1688	178.8313	False
child, endangering welfare	prostitution	-9.0	1.0	-226.8093	208.8093	False
child, endangering welfare	prostitution, patronizing 4, 3	7.5	1.0	-210.3093	225.3093	False
child, endangering welfare	stolen property 3,possession	-6.7926	1.0	-185.1054	171.5203	False
child, endangering welfare	theft of services, unclassified	-5.8967	1.0	-183.8349	172.0415	False

```
res2 = pairwise_tukeyhsd(entire_room['price'], entire_room['dominant_ofns'])
res2.summary()
```

✓ 1.9s Python

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
assault 3	child, endangering welfare	55.4513	0.9965	-92.6516	203.5542	False
assault 3	controlled substance, possession 7	10.5085	0.0003	2.8542	18.1629	True
assault 3	intoxicated driving,alcohol	4.9135	0.9967	-8.2675	18.0945	False
assault 3	larceny,grand from open areas,unclassified	38.0768	0.0	26.076	50.0777	True
assault 3	larceny,petit from open areas,unclassified	5.1462	0.9439	-5.1475	15.4398	False
assault 3	marijuana, possession 4 & 5	-12.4684	0.0	-18.2647	-6.6721	True
assault 3	ny state laws,unclassified misdemeanor	27.0263	0.0	12.2308	41.8218	True
assault 3	ny state laws,unclassified violation	26.7435	0.0	17.7206	35.7664	True
assault 3	prostitution	89.4513	0.9997	-206.6548	385.5574	False
assault 3	prostitution, patronizing 4, 3	-37.7362	0.9352	-111.8873	36.4148	False
assault 3	stolen property 3,possession	0.1786	1.0	-15.347	15.7041	False
assault 3	theft of services, unclassified	-0.539	1.0	-11.2275	10.1494	False
assault 3	traffic,unclassified infraction	-0.9764	1.0	-24.8719	22.9191	False
assault 3	traffic,unclassified misdemean	10.7213	0.9429	-10.6791	32.1217	False
assault 3	trespass 3, criminal	43.9513	0.9675	-49.7802	137.6828	False
child, endangering welfare	controlled substance, possession 7	-44.9427	0.9997	-193.1105	103.225	False
child, endangering welfare	intoxicated driving,alcohol	-50.5378	0.9988	-199.0936	98.0181	False
child, endangering welfare	larceny,grand from open areas,unclassified	-17.3745	1.0	-165.8303	131.0813	False
child, endangering welfare	larceny,petit from open areas,unclassified	-50.3051	0.9988	-198.6327	98.0224	False
child, endangering welfare	marijuana, possession 4 & 5	-67.9197	0.9734	-216.0031	80.1638	False
child, endangering welfare	ny state laws,unclassified misdemeanor	-28.425	1.0	-177.1328	120.2828	False
child, endangering welfare	ny state laws,unclassified violation	-28.7077	1.0	-176.9525	119.537	False
child, endangering welfare	prostitution	34.0	1.0	-297.0195	365.0195	False
child, endangering welfare	prostitution, patronizing 4, 3	-93.1875	0.8627	-258.6973	72.3223	False
child, endangering welfare	stolen property 3,possession	-55.2727	0.9968	-204.0549	93.5095	False
child, endangering welfare	theft of services, unclassified	-55.9903	0.9962	-204.3458	92.3652	False
child, endangering welfare	traffic,unclassified infraction	-56.4277	0.9963	-206.3146	93.4593	False
child, endangering welfare	traffic,unclassified misdemean	-44.73	0.9997	-194.2395	104.7795	False

Fig4.2.2-1 Multiple Comparison of Means - Tukey HSD (Private room)

Fig4.2.2-2 Multiple Comparison of Means - Tukey HSD (Entire room)

- By Tukey HSD, from entire room and private room groups, room prices lying in larceny and ny state law violation dominant crime neighborhoods tend to be obviously significantly different from room prices lying in other types of dominant crime neighborhoods. It is because their pairwise p-values are lower than the chosen significance level (commonly 0.05), which suggests the difference in means between the two groups is statistically significant.
- For our understanding, a possible explanation can be given that these neighborhoods' airbnb room prices are normally set higher than other neighborhoods, which means the owner or customers who choose these rooms could be richer. As a result, the house owners or customers are easier to be chosen as a larceny target.

4.2.2 Annual Crime Rate & Price

- Secondly, aside from categorical crime data, we also look into quantitative features of annual crime rate as continuous variables in detail. We want to answer whether these additional features can help us explain the price better. Thus, we perform the OLS (ordinary least squared) model where room_type and four level annual_crime_rate (felony, misdemeanor, violation and infraction) as independent variables and room price as dependent variable:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Q('price')   R-squared:                       0.312
Model:                            OLS   Adj. R-squared:                  0.312
Method:                 Least Squares   F-statistic:                     2233.
Date:                Wed, 06 Dec 2023   Prob (F-statistic):               0.00
Time:                        02:59:46   Log-Likelihood:            -1.9715e+05
No. Observations:               34448   AIC:                         3.943e+05
Df Residuals:                   34440   BIC:                         3.944e+05
Df Model:                           7
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                        185.4820      0.689    269.083      0.000     184.131     186.833
Q('room_type')[T.Hotel room]      26.4795      7.579      3.494      0.000      11.624      41.335
Q('room_type')[T.Private room]   -97.3435      0.812   -119.831      0.000     -98.936     -95.751
Q('room_type')[T.Shared room]   -103.5223      3.519    -29.416      0.000    -110.420     -96.624
Q('annual_felony_rate')           -0.0074      0.003     -2.182      0.029      -0.014      -0.001
Q('annual_misdemeanor_rate')      -0.0085      0.001     -5.799      0.000      -0.011      -0.006
Q('annual_violation_rate')         0.1225      0.006     19.944      0.000       0.110       0.135
Q('annual_infraction_rate')       -0.9852      0.088    -11.160      0.000      -1.158      -0.812
==============================================================================
Omnibus:                     9704.058   Durbin-Watson:                   1.885
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            26872.036
Skew:                           1.503   Prob(JB):                         0.00
Kurtosis:                       6.113   Cond. No.                     4.09e+04
==============================================================================
```

Fig4.2.2-1 OLS Regression Results

- According to the OLS regression result:

1. The model indicates that annual crime rates are significant predictors (p-value < 0.05) of Airbnb prices. The negative coefficients for annual felony, misdemeanor, and infraction rates suggest that higher crime rates in these categories are associated with lower prices.

2. Despite the statistical significance of the predictors, the moderate R-squared value implies that other factors not included in the model also play a significant role in determining Airbnb prices, which means the regression power of annual crime rates is very limited. Thus, we have some reasons to believe that room prices are less likely to be determined by features of crime rates.

4.3 Segmentation & Clustering

4.3.1 Clustering (exclude price)

- Firstly, we want to find the potential patterns inside the records without the influence of price. Thus we dropped the 'id' and 'price' columns. Then we used the pipeline to first scale each remaining column, then reduce the dimension to 5, and finally use K-means to cluster the records.
- We plotted the cluster visualization and the average silhouette score with the K-means clustering number varying from 2 to 20. We found when the cluster number is 7, the average silhouette score is the highest (0.318).

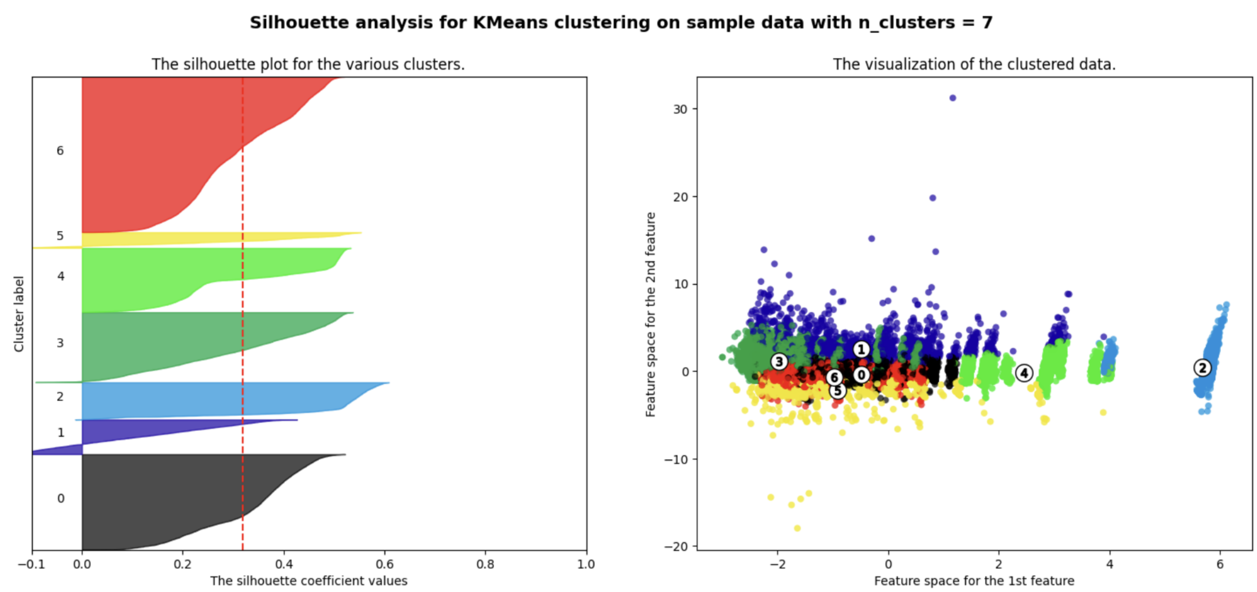

Fig4.3.1 Optimal Clustering Result

4.3.2 Clustering Result Analysis

- Based on the optimal clustering result, we assign the segments to the original data (including 'id' and 'price'). An intuitive idea is that each cluster's price distribution should be in line with each crime rate if there exist certain relationships. We decided to compare them using box plots.

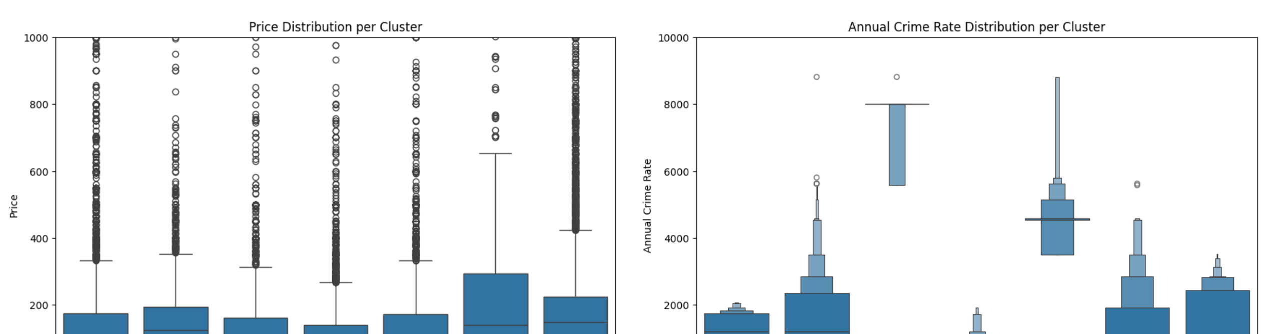

Fig4.3.2 Price/Annual Crime Rate per Cluster

- From the above two boxplots, we cannot find a direct relationship between the clusters and the crime rate. For example, Cluster 2 has the highest crime rate while Cluster 3 has the lowest crime rate, but there is no significant difference in prices between the two clusters.

4.3.3 Comparison of Visualization (clusters exclude price, clusters only with price, clusters from geojson)

- Now we attempt to visualize the clusters on the map. We first assign the neighborhood to the most frequent cluster in the neighborhood. Then we plot the clusters on the map. We decided to have three map visualizations based on clusters excluding price, clusters only with price, and clusters from geojson.

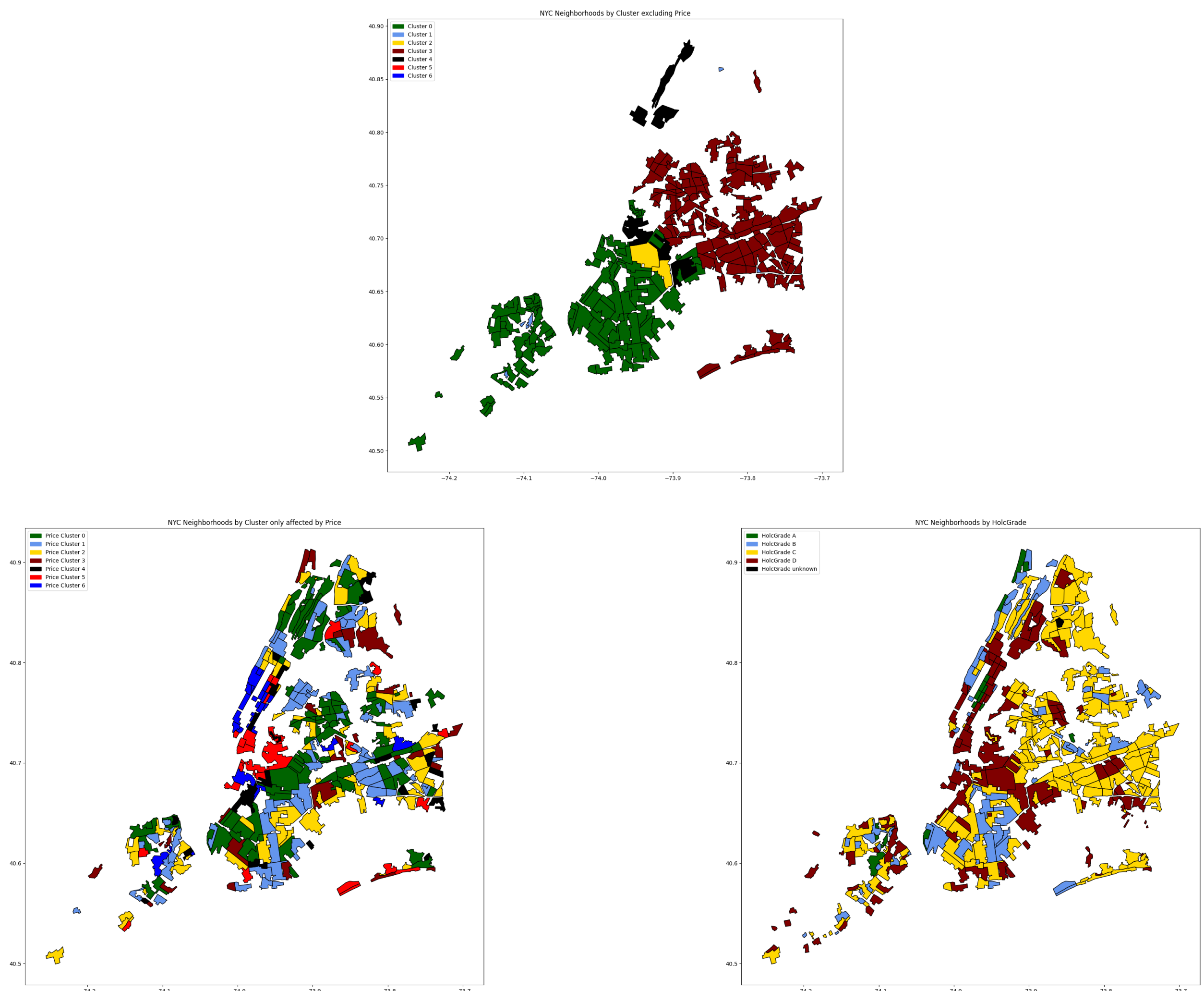

Fig4.3.3 Map Visualizations (top: clusters without price; left bottom: clusters by price; right bottom: clusters by 3.3)

- We find that the clusters from K-means do not correlate with the price, but are determined only by the region of the NYC city. Additionally, the general dangerous extent has no obvious correlations with price from the comparison of the bottom two. Thus, the crime rate may not be toppest main factor affecting the price of Airbnb.

5. Conclusion

Our study delved into how crime rates and Airbnb pricing interact within NYC's neighborhoods, revealing that prices aren't solely driven by crime but also by a tapestry of regional traits. Our spatial analysis with K-means clustering, Tukey HSD and geographic visualization, offer a guide for the Airbnb community and a strategic tool for urban development, integrating economic and safety considerations.

5.1 Limitation

While our analysis has yielded significant observations, the dataset still has some flaws. One notable constraint is the reliance on available data, which may not fully capture the real-time fluctuations in crime and economic factors. Furthermore, our clustering analysis indicated that the correlation between Airbnb pricing and crime is not as robust as initially hypothesized, underscoring the need for a more granular approach that considers additional variables and market forces.

Our visualizations, though comprehensive, also underline the absence of a clear-cut relationship between the perceived danger in various regions and the pricing of Airbnb listings. This highlights the complexity of the market and suggests that additional factors must be accounted for to accurately predict and strategize pricing models in correlation with crime statistics.

5.2 Next Step

As we look to the future, we aim to expand our dataset, incorporate real-time analytics, and explore a wider array of contributing factors. This will not only enhance the precision of our insights but also provide a more actionable framework for all stakeholders involved.

6. Statement of Work

6.1 Contribution Report

6.1.1 Our team collaborated on the project based on the following aspects:

Keye Chen	Data sources, geojson file data manipulation, clustering analysis
Fengyu Zhang	Data manipulation of aggregation and merge, correlation analysis
Boyan Wu	Motivation, EDA, Data Visualization, Conclusion and Statement of Work,

Assessment and Improvement:

Our team has distributed work and divided it to leverage individual strengths. Each member contributes to a different facet of the project. However, we can improve our collaboration in listed below in the future:

Regular Check-Ins: Schedule periodic meetings to ensure everyone is on track and to discuss any challenges or findings, making it easier for team members to collaborate and pick up where others left off.

Cross Reviews: Occasionally, swap roles or review each other's work would reduce the risk of siloed information, and increase team members' skills across different areas.

Feedback Improvement: Each team member can raise the feedback on the overall performance to reflect on continuous improvement in team dynamics and project outcomes.