NYC Real Estate Analytics Project- Descriptive Analytics

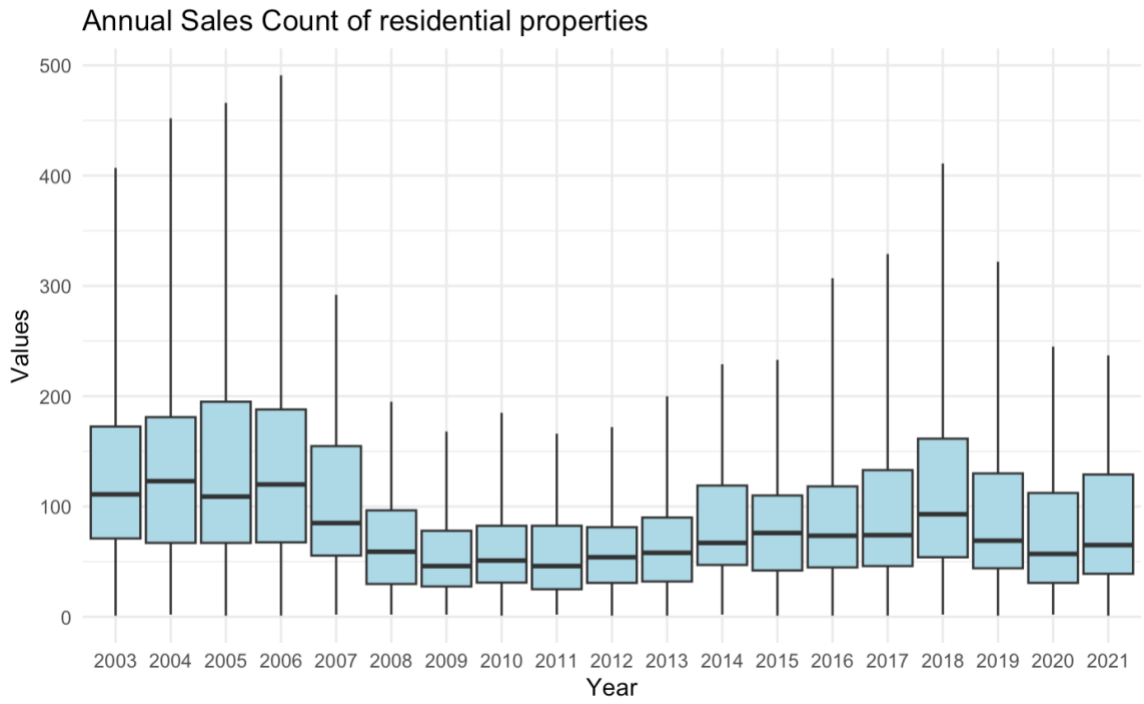
Introduction

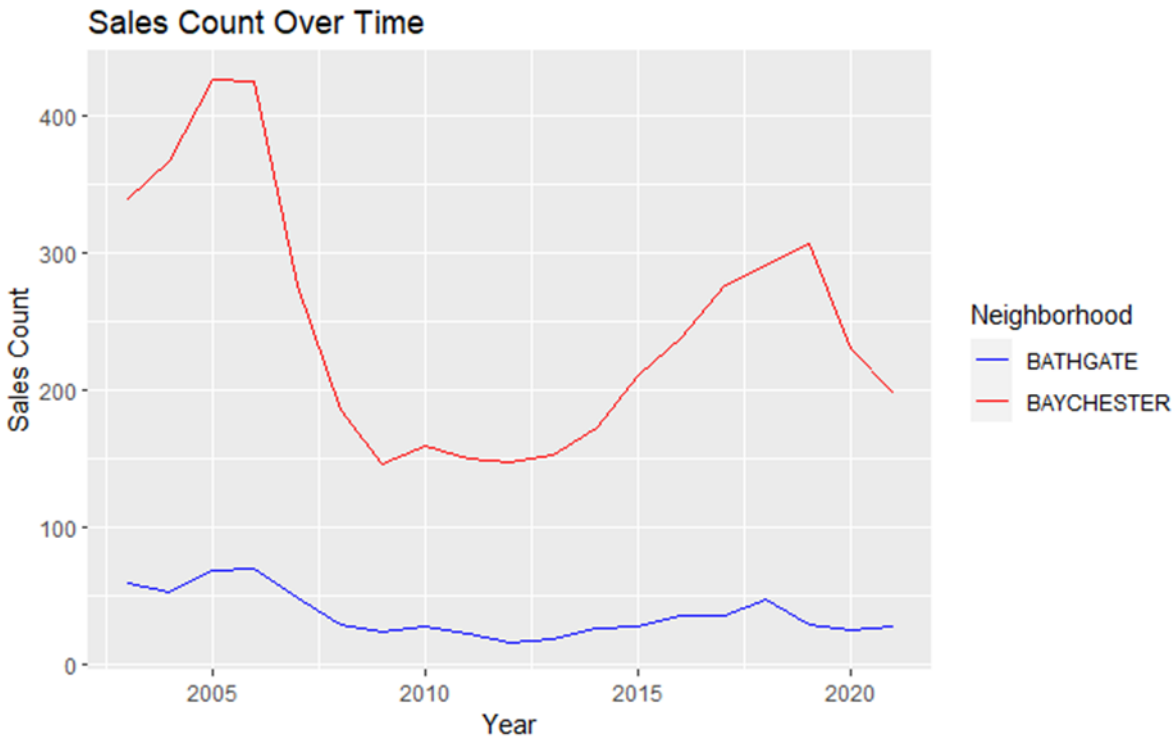
Our goal is to analyze real estate data from 2003 to 2021 using R/RStudio. We'll filter this data to focus on residential properties in specific neighborhoods within each of our assigned boroughs. We'll remove missing data and set a reasonable threshold for sales values to maintain accuracy.We'll save this filtered data to avoid repetitive processing and calculate six Key Performance Indicators (KPIs) to understand long-term market performance.

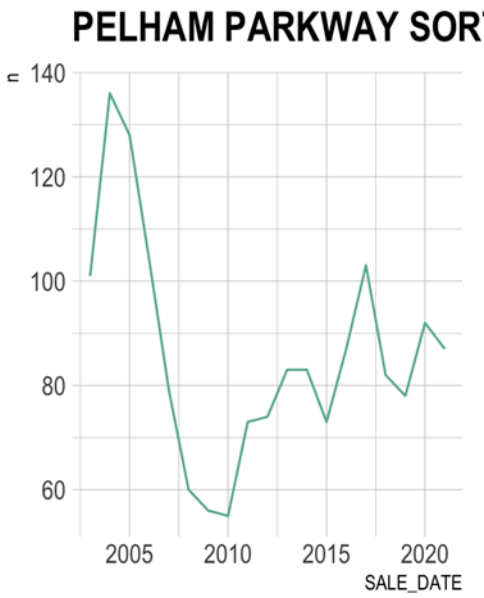
The first aim is to compute all KPIs and potential revenue for each borough of all the years. Our results will be presented in clean and professional tables and plots, allowing us to compare each neighborhood based on KPIs and Potential Revenue.If in a group, we'll combine results to decide the best location to open a branch office. This will help us make informed decisions about our chosen neighborhoods based on comprehensive data analysis. Real estate analytics is a multidisciplinary field that applies data analysis, statistical techniques, and machine learning to gaininsights and make informed decisions in the real estate industry. In the past,real estate analytics focused on traditional metrics. It has gained significant traction in recent years due to advancements in data collection, technology,and a growing recognition of the importance of data-driven decision-making in real estate. Real estate has long been a data-driven industry, with professionals relying on market data, property assessments, and financial models for decision-making. Data analytics should have its own strategic direction with long-term roles and goals beyond just a few pilot projects and use cases. Researchers have developed predictive models for property prices,rental rates, and market trends. These models use historical data and factors such as location, economic indicators, and demographics to forecast future property values. Machine learning algorithms, including regression models,decision trees, and neural networks, are increasingly used to automate property valuation and predict market dynamics.

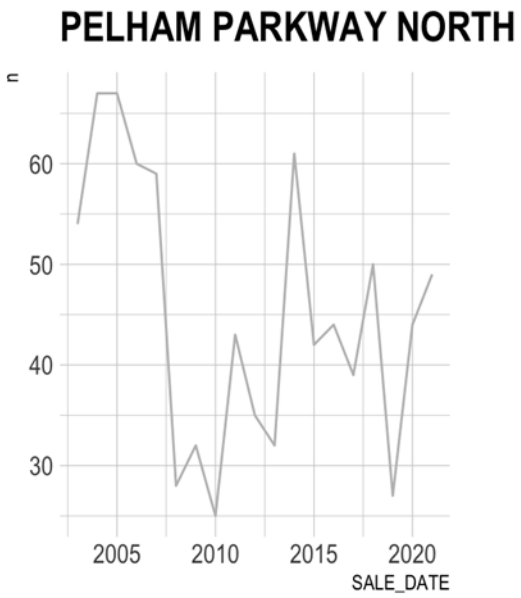
The second aim is to use the heatmap and some other tools to find some clues of the KPIs and Potential Revenue, which is a graphical representation of data where individual values contained in a matrix are represented as colors. This visualization technique helps in understanding the magnitude of relationship between two variables, finding patterns, and variance in the dataset. Heatmaps are quite popular in various fields such as biology, web analytics, and geography, but they are notably prevalent in correlational analysis where the correlation coefficients between pairs of variables are visualized using color gradients. And use the K-Means Algorithm that partitions a dataset into KK clusters by minimizing the within-cluster variances. In R, various functions and packages allow you to perform K-Means clustering to get the different clusters so that we can classify neighborhoods differently.

KPI 1:


Annual Sales Count of residential properties
Values
500
400
300
200
100
0
2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
Year


Sales Count Over Time
Sales Count
400
300
200
100
0
2005
2010
2015
2020
Year
Neighborhood
BATHGATE
BAYCHESTER


PELHAM PARKWAY SORTH
n
140
120
100
80
60
2005
2010
2015
2020
SALE_DATE

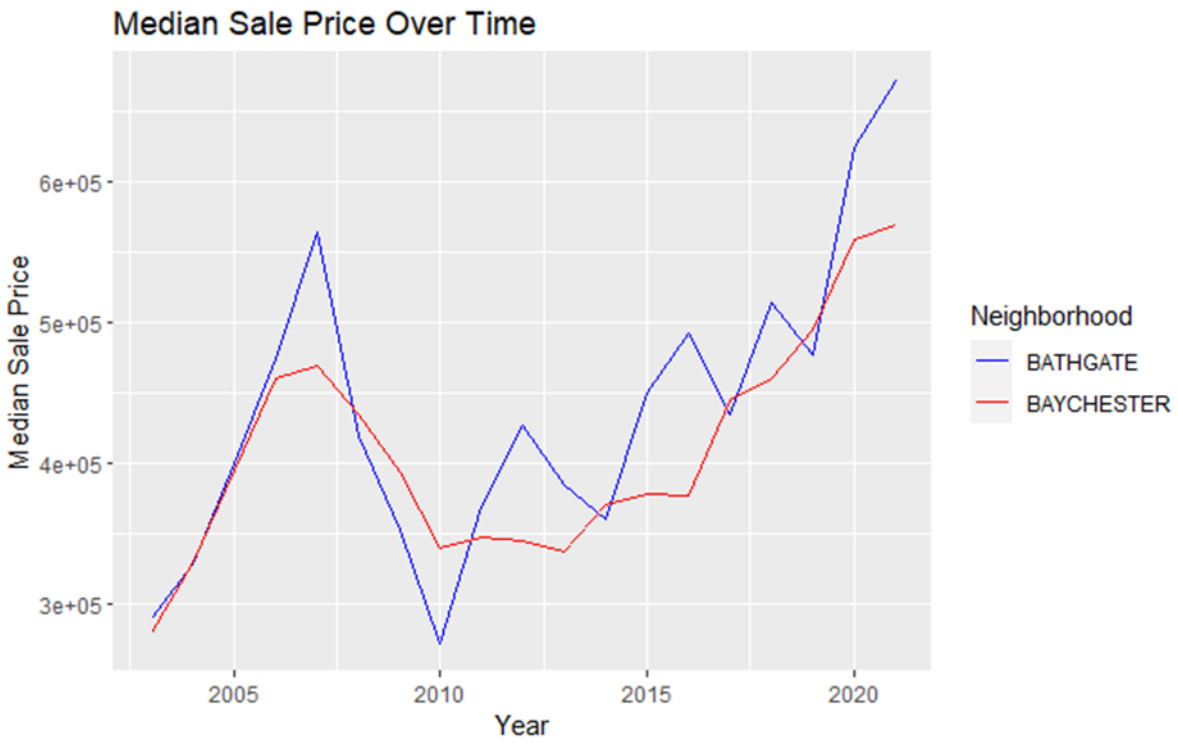

PELHAM PARKWAY NORTH
n
60
50
40
30
2005
2010
2015
2020
SALE_DATE

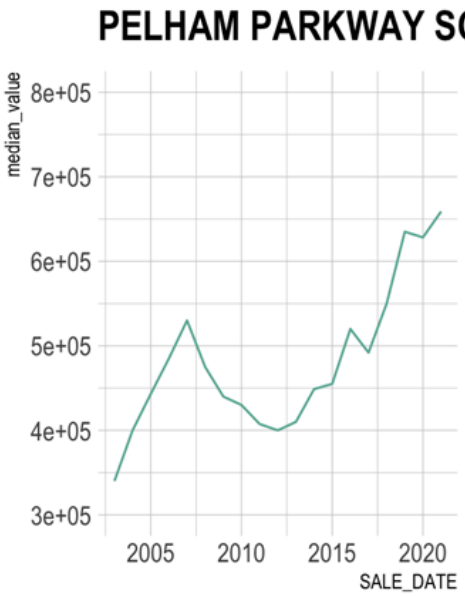

Westchester annual sales count of residential properties
Count
160
120
80
2005
2010
2015
2020
Year

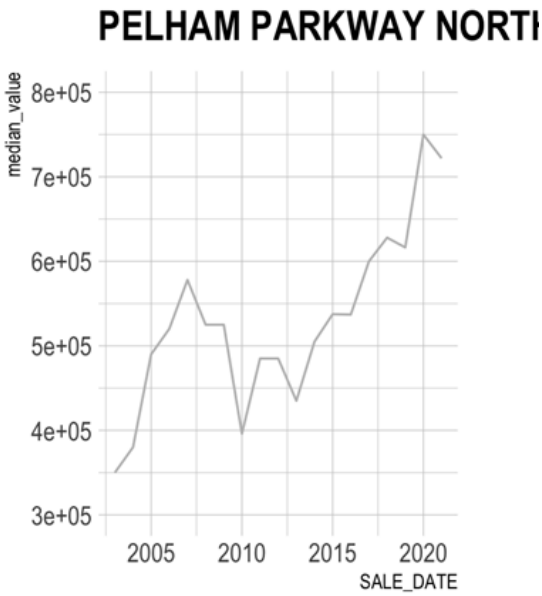

Williamsbridge annual sales count of residential properties
Count
500
400
300
200
2005
2010
2015
2020
Year

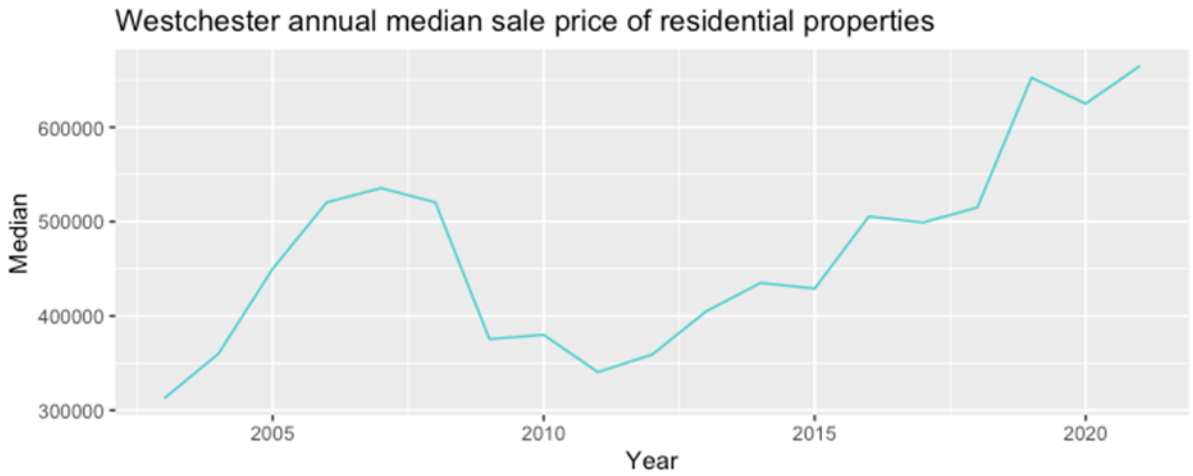
Comparing the boxplot of all neighborhoods of our borough and the line graphs of each neighborhood we chose before, it is obvious that the median sales count (represented by the horizontal line inside each box) has been relatively stable over the years, fluctuating around the 200-mark. And there is considerable variability in sales count across the neighborhoods within the borough, as evidenced by the heights of the boxes and the lengths of the whiskers. Some years, like 2005 and 2019, exhibit larger variability than others, like 2010 or 2012. The vertical lines (or whiskers) extending above and below each box represent the range within which the majority of values lie, and any point outside of these whiskers is typically considered an outlier. For instance, there is a significant upward spike in 2018, indicating that at least one neighborhood had a notably high sales count that year. The years 2018 and 2021 seem to have neighborhoods with notably high sales counts when compared to other years, as their boxes (particularly their upper quartiles) are much higher. The years 2014-2017 show a denser clustering (boxes are shorter), suggesting that the neighborhoods had more similar sales counts in those years. In the more recent years, from 2018-2021, there seems to be a slight upward trend in the lower quartile of sales, indicating that even neighborhoods with lower sales are performing better than in previous years. In general, the total trends of neighborhoods we chose are similar to the trend of all the neighborhoods of the borough. But only Pelham Parkway South and West Chester are also quantitatively similar to the total region average which are around 50 to 150. Differences in other areas are quite significant; they either greatly exceed the average or are much below the average level.

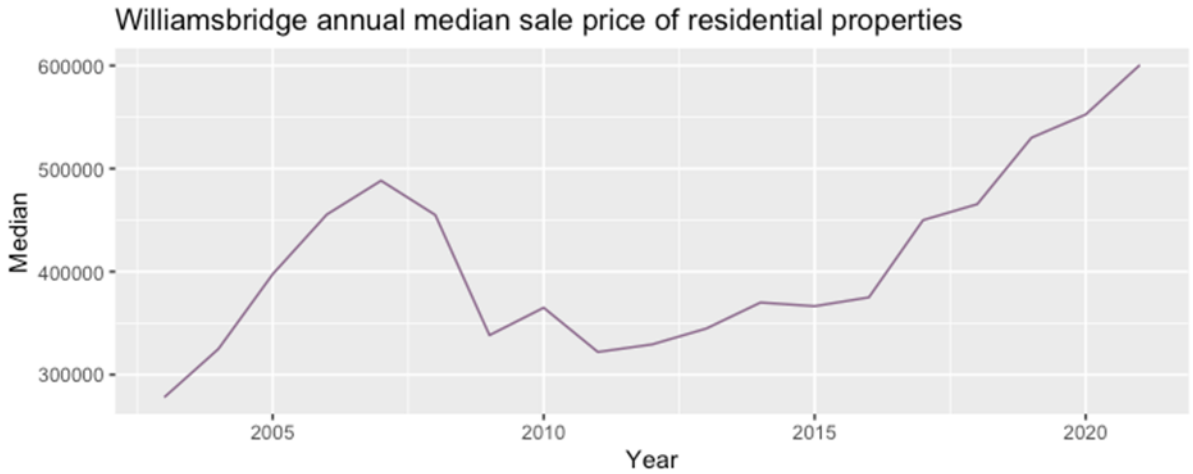
KPI 2：


Median Sale Price Over Time
Median Sale Price
6e+05
5e+05
4e+05
3e+05
2005
2010
2015
2020
Year
Neighborhood
BATHGATE
BAYCHESTER


PELHAM PARKWAY SORTH
median_value
8e+05
7e+05
6e+05
5e+05
4e+05
3e+05
2005
2010
2015
2020
SALE_DATE


PELHAM PARKWAY NORTI
median_value
8e+05
7e+05
6e+05
5e+05
4e+05
3e+05
2005
2010
2015
2020
SALE_DATE

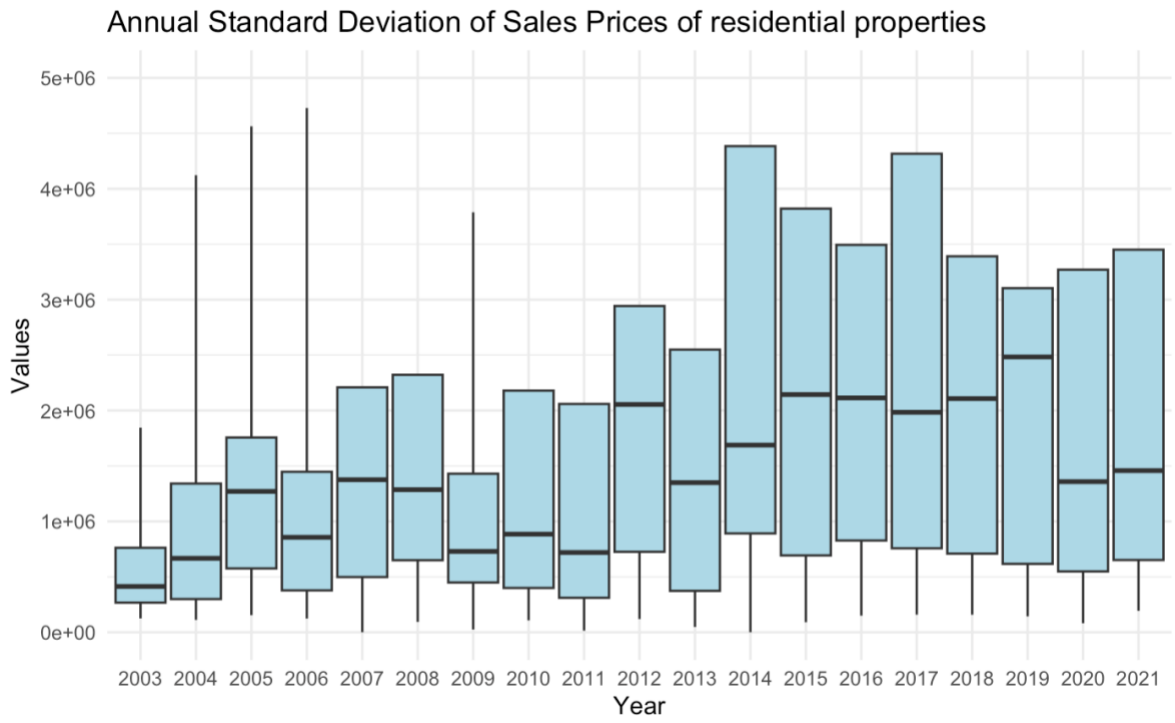


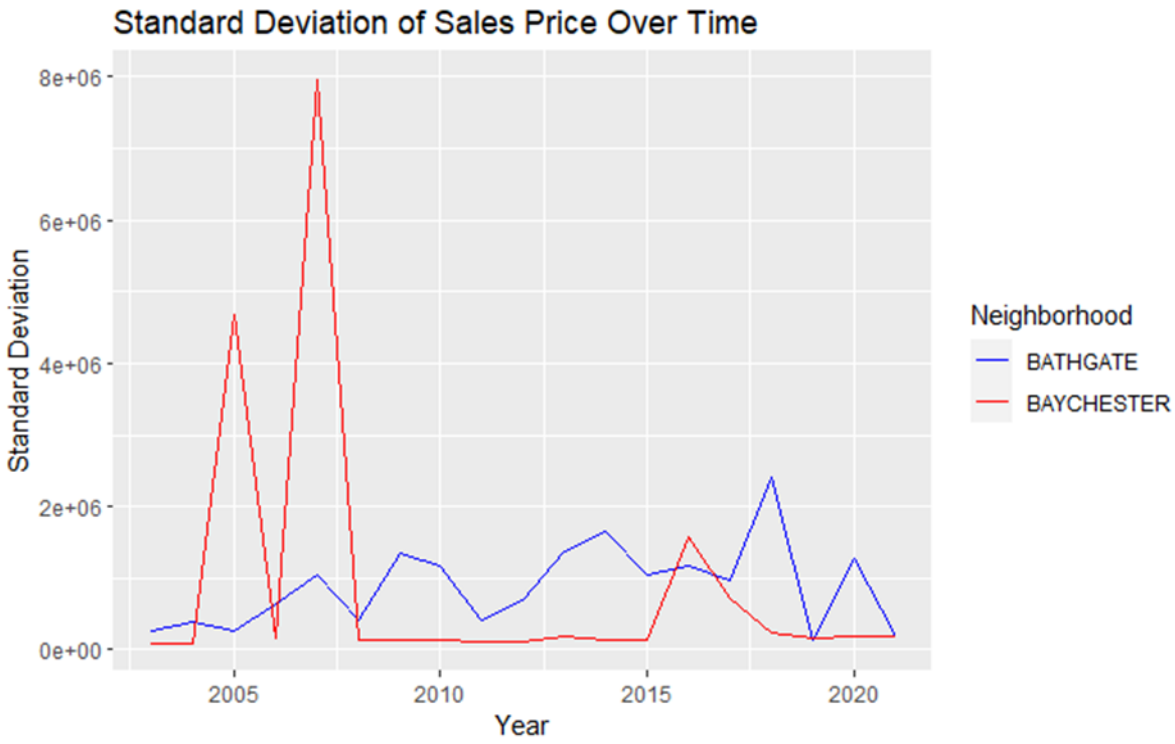


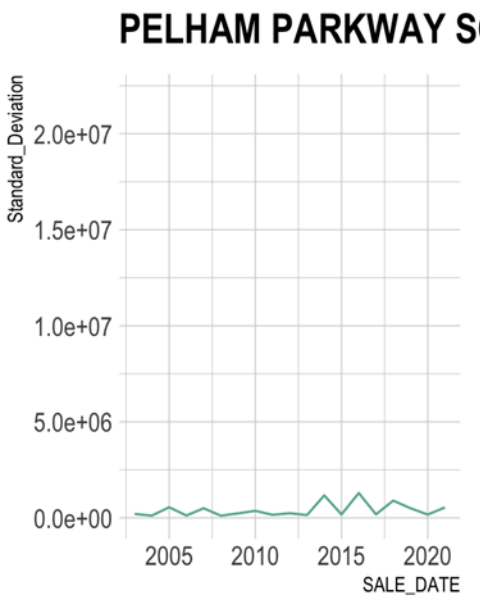
For KPI 2, Annual median sale price of residential properties, the provided boxplot offers an insightful visual representation of the annual median sale price of residential properties across various neighborhoods of a borough, spanning from 2003 to 2021. The analysis of this boxplot will endeavor to Over the 18-year period, there has been a clear upward trajectory in the median sale price of residential properties. This indicates a growing demand or appreciation in property values in the borough.2003 to 2008: The first few years (2003 to 2008) display a relatively larger spread in values, which suggests that during this period, the difference in median sale prices among neighborhoods was more pronounced.2009 to 2021: Post-2008, the boxes become less tall, showing less variability among the neighborhoods. This could indicate a stabilization or homogenization of property prices across the neighborhoods.There are few noticeable outliers, especially in the years 2015, 2016, and 2018. These anomalies could be attributed to specific events or developments in certain neighborhoods that did not impact others to the same extent. It would be beneficial to investigate these particular years more closely to determine the cause.While the bottom boundary of the boxes shows modest growth, it's the upper boundary and the median which exhibit a more pronounced rise, especially after 2014. This implies that the upper tier neighborhoods in the borough saw a more significant appreciation in their median residential property values compared to the lower-tier neighborhoods. The year 2020 stands out with a relatively taller box, suggesting an increase in variability among neighborhoods in their median sale prices. Considering global events, this might be a reflection of the economic repercussions of the COVID-19 pandemic, where some neighborhoods might have been affected differently due to various socioeconomic factors. For the neighborhoods we chose, their trends of annual median sale price are similar to the boxplot trend for the total, which means that what we chose are all the typical ones for

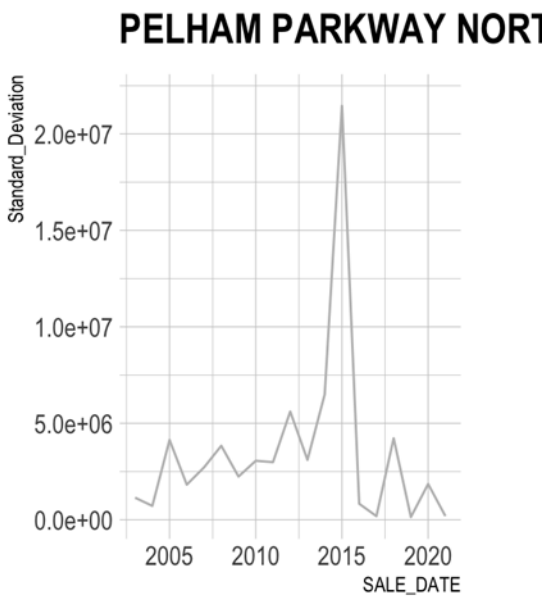
the median sale price. The difference lies in whether their degree of change is gentle or intense. And in our opinions, Bathgate is the most similar to the average level of median sale price, which means this area can be the most representative among 6 neighborhoods.

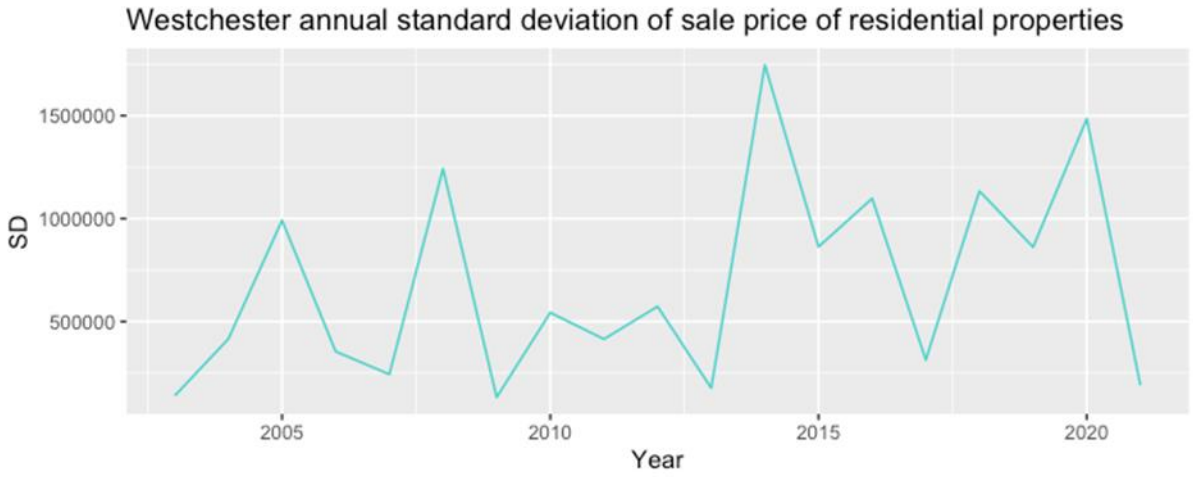
KPI 3:






PELHAM PARKWAY SORTH
Standard_Deviation
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
2005
2010
2015
2020
SALE_DATE


PELHAM PARKWAY NORT
Standard_Deviation
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
2005
2010
2015
2020
SALE_DATE


Westchester annual standard deviation of sale price of residential properties
SD
1500000
1000000
500000
2005
2010
2015
2020
Year


Williamsbridge annual standard deviation of sale price of residential properties
SD
1500000
1000000
500000
2005
2010
2015
2020
Year

The boxplot presented above visualizes the annual standard deviation of sales prices of residential properties across various neighborhoods of a specific borough from 2003 to 2021. This analysis aims to highlight key patterns, trends, and outliers that emerge from the data.

Boxplots, also known as whisker plots, depict the distribution of data based on the median, quartiles, and potential outliers. In this context, the central rectangle spans the first quartile to the third quartile, the segment inside the rectangle shows the median, and "whiskers" above and below the box demonstrate the range within which the bulk of the values fall. Points outside of these whiskers typically represent outliers.

Observing the progression from 2003 to 2021, there seems to be a noticeable increase in the standard deviation of sales prices. This suggests that the variability or disparity in property prices among different neighborhoods in the borough has expanded over the years.

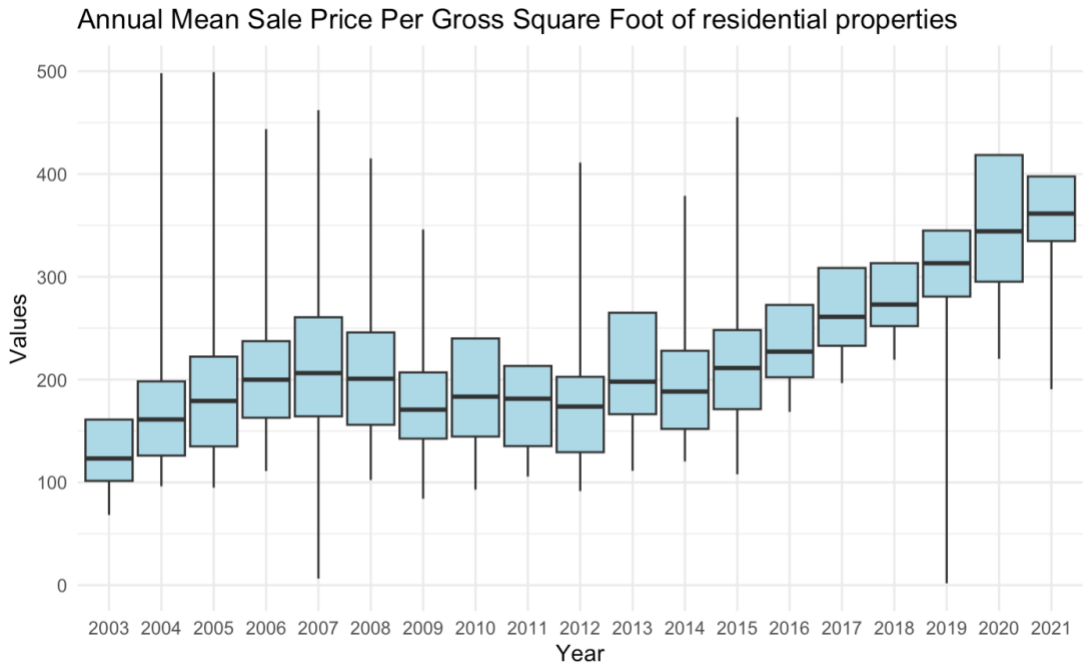
Notably, the years 2017 to 2019 witnessed a significant surge in standard deviation, with 2018 presenting the most pronounced variability. This could imply that during this period, certain neighborhoods in the borough saw exceptionally high property sales prices, while others might have experienced stagnation or even a decrease.In contrast, the years between 2004 to 2007 and again in 2021 show relatively smaller boxes, suggesting a more uniform pricing structure across the borough's neighborhoods during these years.

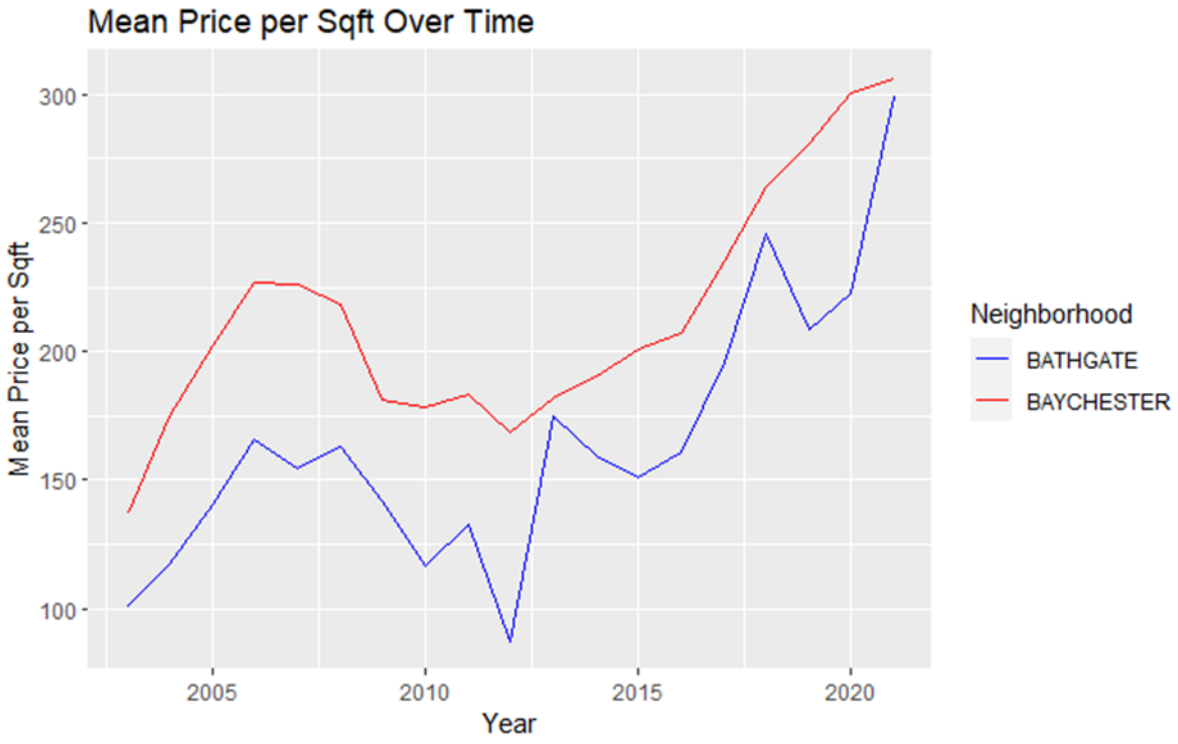
The years 2003, 2005, and 2013 display individual points below the main box, which might indicate neighborhoods that had significantly lower standard deviations in property prices compared to the overall trend for those years.

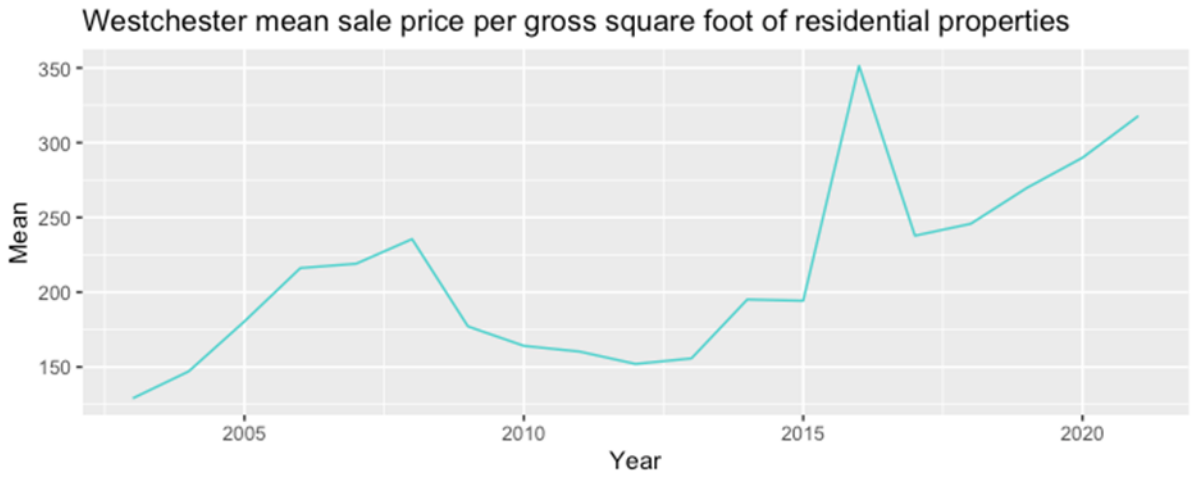
The rising standard deviation over the years, especially around 2017 to 2019, indicates growing economic disparity among the neighborhoods in terms of property prices. Such a trend could be attributed to various factors, including economic booms, infrastructural developments, or increasing demand in specific neighborhoods. On the flip side, the periods of reduced variability suggest a more cohesive and less disparate borough in terms of property pricing.

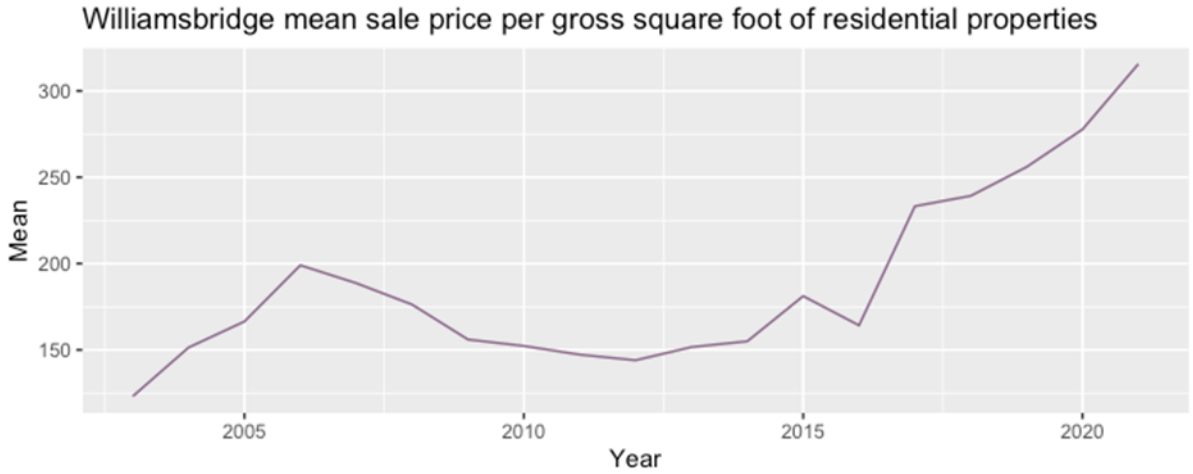
Compared with the neighborhoods we chose, Baychester has the largest variation in standard deviation and its value is significantly higher than other areas and the overall average from 2003 to 2007. In general, most neighborhoods we chose have the similar standard deviation trends and averages, especially Westchester and Williamsbridge. Their sd is representative for this borough.

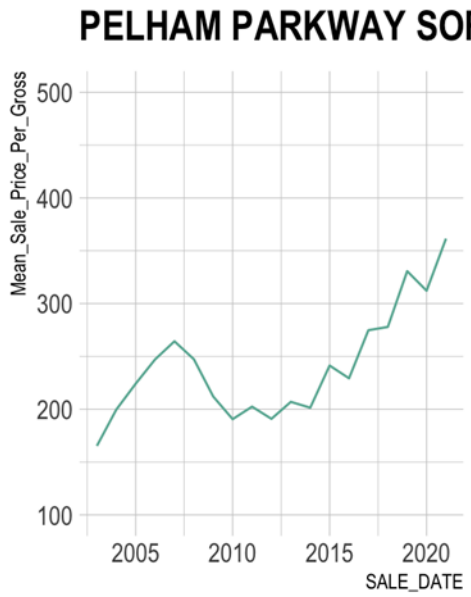
KPI 4 :

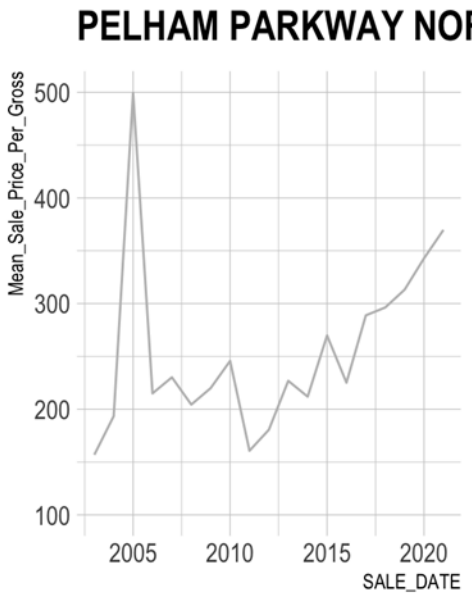

Annual Mean Sale Price Per Gross Square Foot of residential properties
Values
500
400
300
200
100
0
2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
Year


Mean Price per Sqft Over Time
Mean Price per Sqft
300
250
200
150
100
2005
2010
2015
2020
Year
Neighborhood
BATHGATE
BAYCHESTER


Westchester mean sale price per gross square foot of residential properties
Mean
350
300
250
200
150
2005
2010
2015
2020
Year


Williamsbridge mean sale price per gross square foot of residential properties
Mean
300
250
200
150
2005
2010
2015
2020
Year


PELHAM PARKWAY SORTH
Mean_Sale_Price_Per_Gross
500
400
300
200
100
2005
2010
2015
2020
SALE_DATE


PELHAM PARKWAY NORTH
Mean_Sale_Price_Per_Gross
500
400
300
200
100
2005
2010
2015
2020
SALE_DATE

The given boxplot illustrates the annual mean sale price per gross square foot of residential properties across various neighborhoods within a specific borough, spanning from 2003 to 2021. The purpose of this analysis is to identify notable trends, anomalies, and insights that may be pertinent to understanding the real estate market dynamics within the borough over these years.

From an overarching perspective, it's evident that the mean sale price per gross square foot has experienced a general upward trajectory over the years. Starting from just above 100 in 2003, the value has seen a steady climb, reaching close to 500 by 2021. This indicates increasing demand, appreciation in property value, or both, over this period.

While the overall trend shows an increase, there are years with more pronounced variations in sale price. The years 2006, 2008, 2010, 2013, and 2017 display broader boxplots, implying greater disparities in prices among the different neighborhoods within the borough. Such disparities might be attributed to various factors such as socio-economic developments, infrastructural advancements, or specific neighborhood-based events.

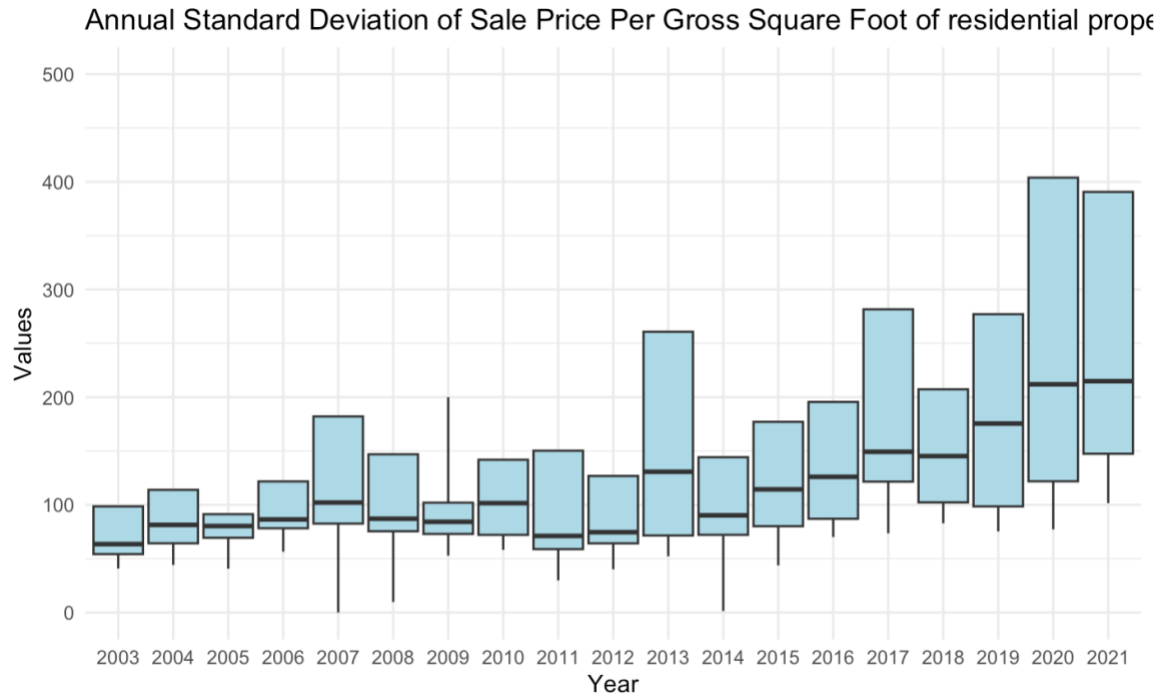
From 2015 onwards, the sale price appears to escalate at a faster rate. This could be indicative of either a borough-wide development boom, increasing attractiveness of the borough to potential buyers, or a general property market surge in the larger region.

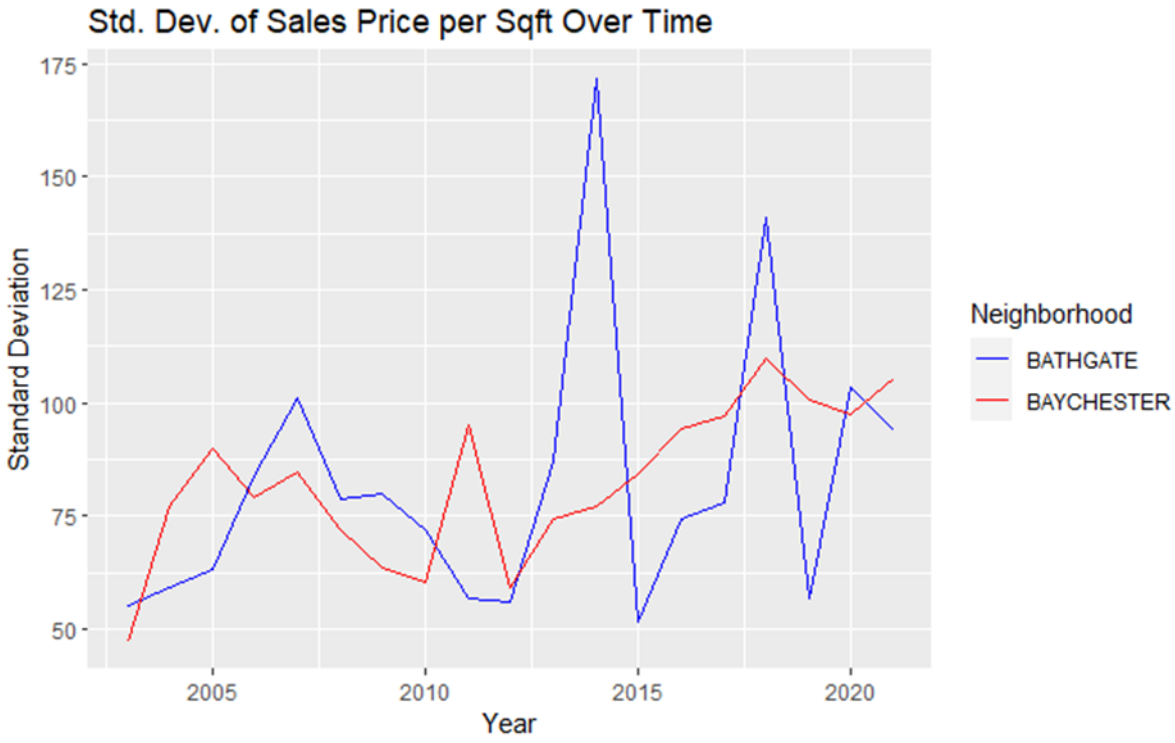
There are no significant outliers in the data, which means that all neighborhoods, on average, adhered reasonably well to the broader borough trends. This might imply that the borough has relatively homogenous characteristics, with no particular neighborhood dramatically outperforming or underperforming the others.

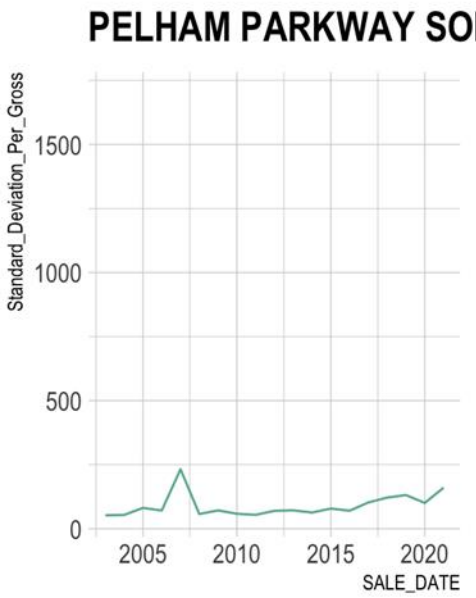
Between 2010 and 2014, there seems to be relative stability in the sale price, with limited growth and less variance among neighborhoods. This could indicate a period of economic stability or stagnation, depending on broader economic indicators.

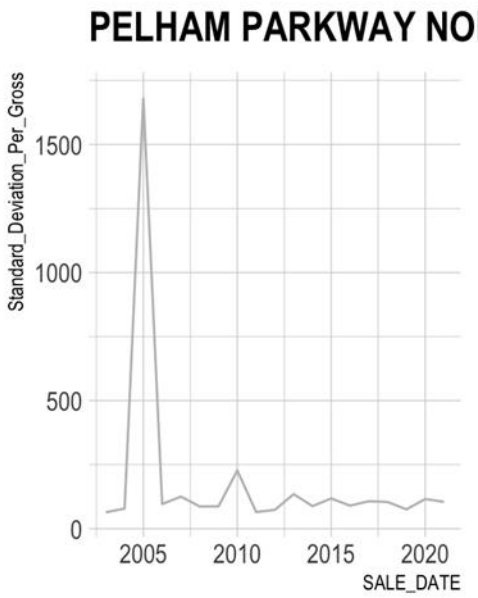
Compared with the neighborhoods we chose, it is obvious that those trends of neighborhoods are similar to the trend of total. In particular, Pelham Park North reached a significant peak in 2005, which greatly differs from the data of other areas and the overall average. Additionally, Westchester saw a rapid upward trend in 2016. These are minor discrepancies, but overall, in terms of the trend in average prices, the areas we selected are all fairly representative.

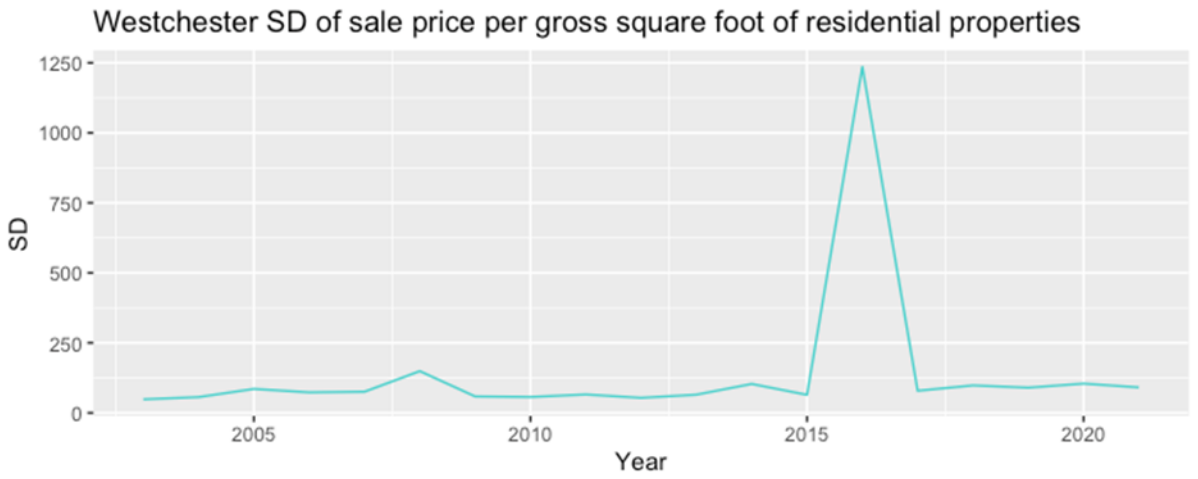
KPI5：

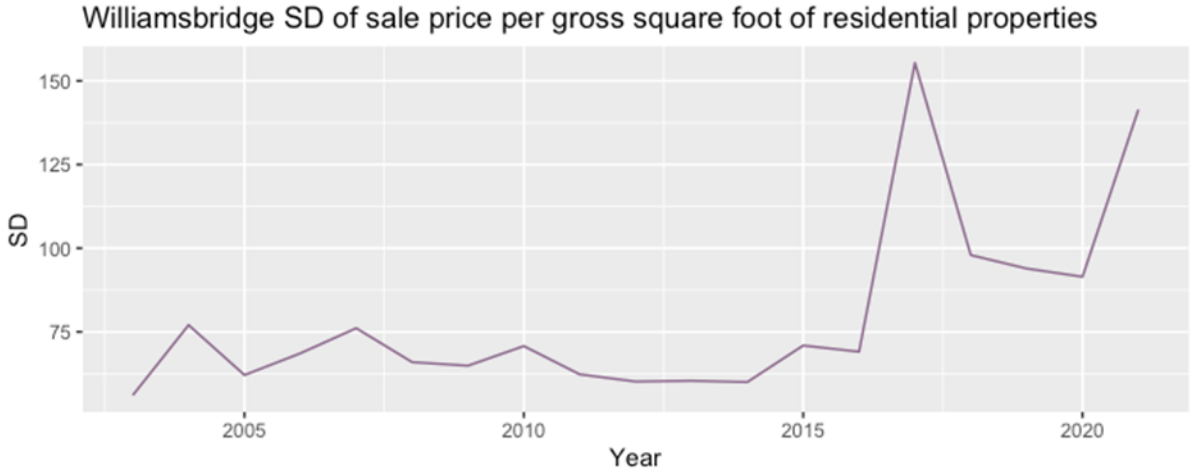

Annual Standard Deviation of Sale Price Per Gross Square Foot of residential prope
Values
500
400
300
200
100
0
2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
Year


Std. Dev. of Sales Price per Sqft Over Time
Standard Deviation
175
150
125
100
75
50
2005
2010
2015
2020
Year
Neighborhood
BATHGATE
BAYCHESTER


PELHAM PARKWAY SORTH
Standard_Deviation_Per_Gross
1500
1000
500
0
2005
2010
2015
2020
SALE_DATE


PELHAM PARKWAY NORTH
Standard_Deviation_Per_Gross
1500
1000
500
0
2005
2010
2015
2020
SALE_DATE


Westchester SD of sale price per gross square foot of residential properties
SD
1250
1000
750
500
250
0
2005
2010
2015
2020
Year


Williamsbridge SD of sale price per gross square foot of residential properties
SD
150
125
100
75
2005
2010
2015
2020
Year

The given boxplot illustrates the annual mean sale price per gross square foot of residential properties across various neighborhoods within a specific borough, spanning from 2003 to 2021. The purpose of this analysis is to identify notable trends, anomalies, and insights that may be pertinent to understanding the real estate market dynamics within the borough over these years.

From an overarching perspective, it's evident that the mean sale price per gross square foot has experienced a general upward trajectory over the years. Starting from just above 100 in 2003, the value has seen a steady climb, reaching close to 500 by 2021. This indicates increasing demand, appreciation in property value, or both, over this period.

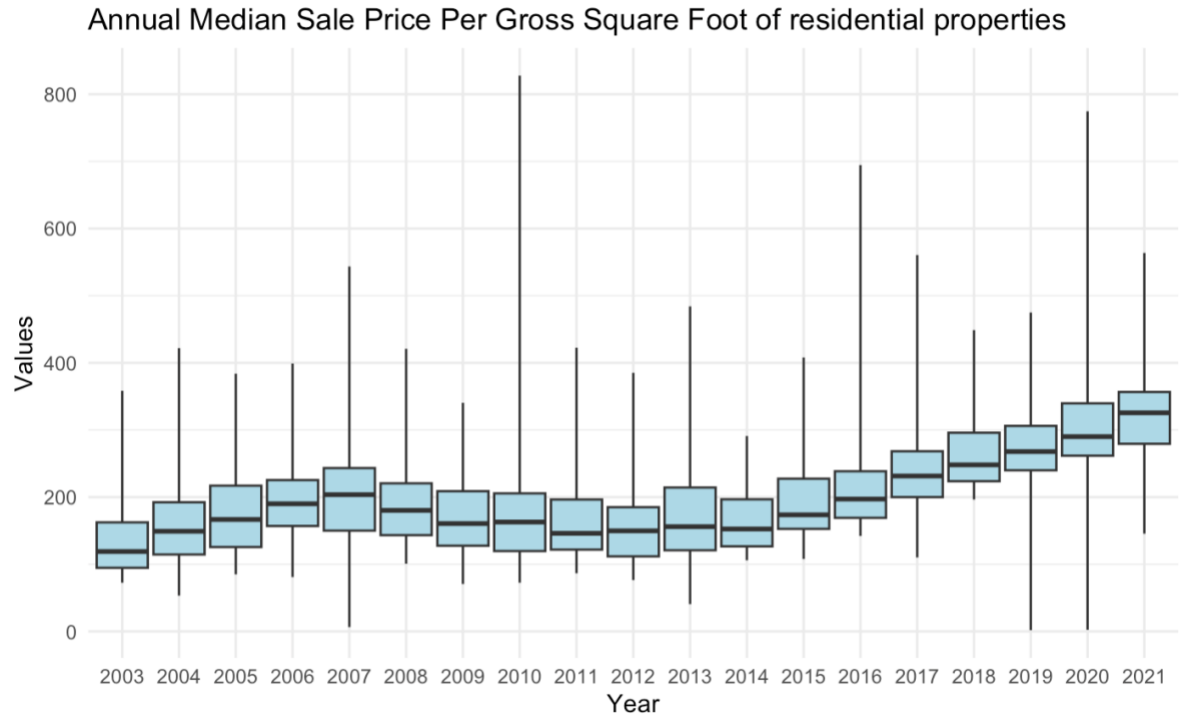
While the overall trend shows an increase, there are years with more pronounced variations in sale price. The years 2006, 2008, 2010, 2013, and 2017 display broader boxplots, implying greater disparities in prices among the different neighborhoods within the borough. Such disparities might be attributed to various factors such as socio-economic developments, infrastructural advancements, or specific neighborhood-based events.

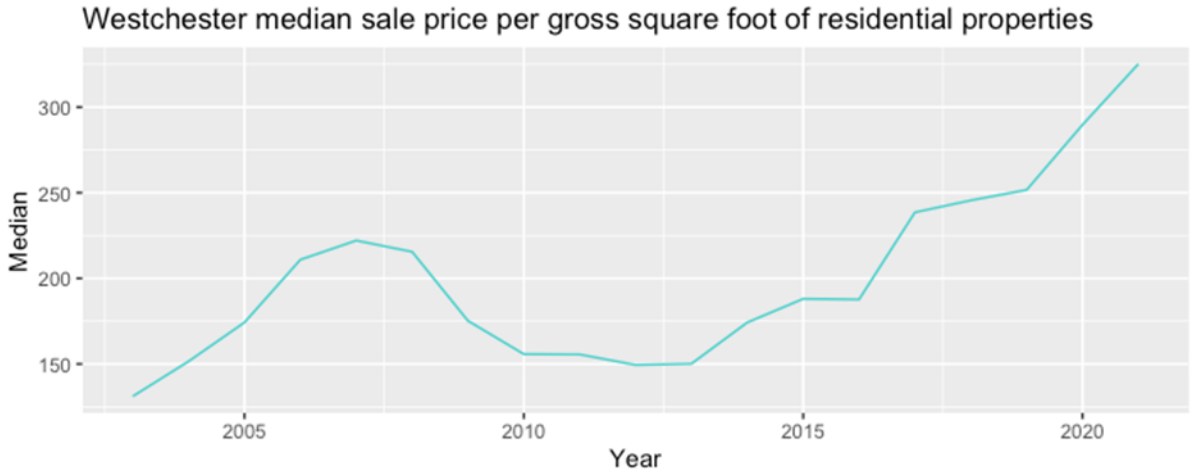
From 2015 onwards, the sale price appears to escalate at a faster rate. This could be indicative of either a borough-wide development boom, increasing attractiveness of the borough to potential buyers, or a general property market surge in the larger region.

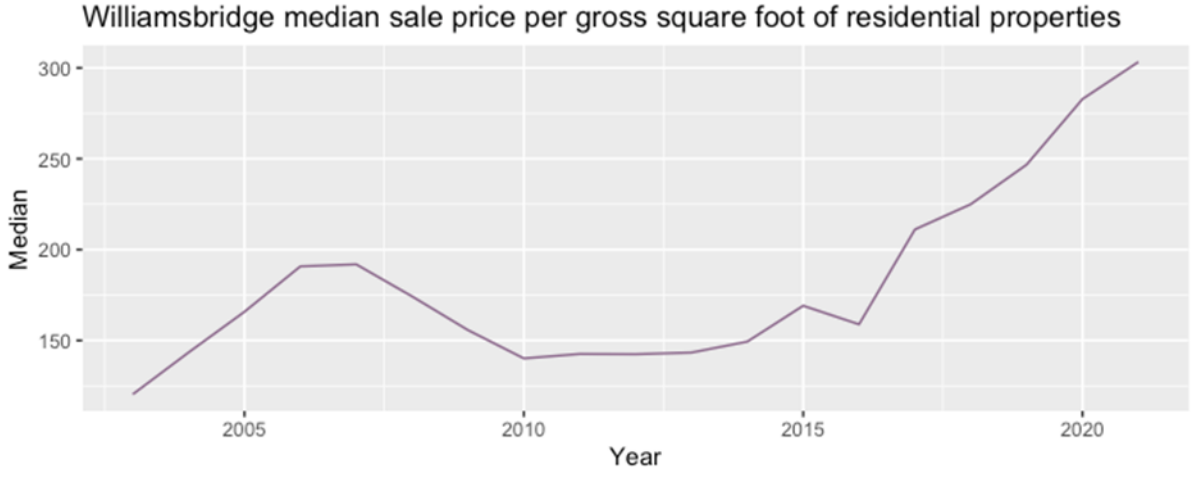
There are no significant outliers in the data, which means that all neighborhoods, on average, adhered reasonably well to the broader borough trends. This might imply that the borough has relatively homogenous characteristics, with no particular neighborhood dramatically outperforming or underperforming the others.

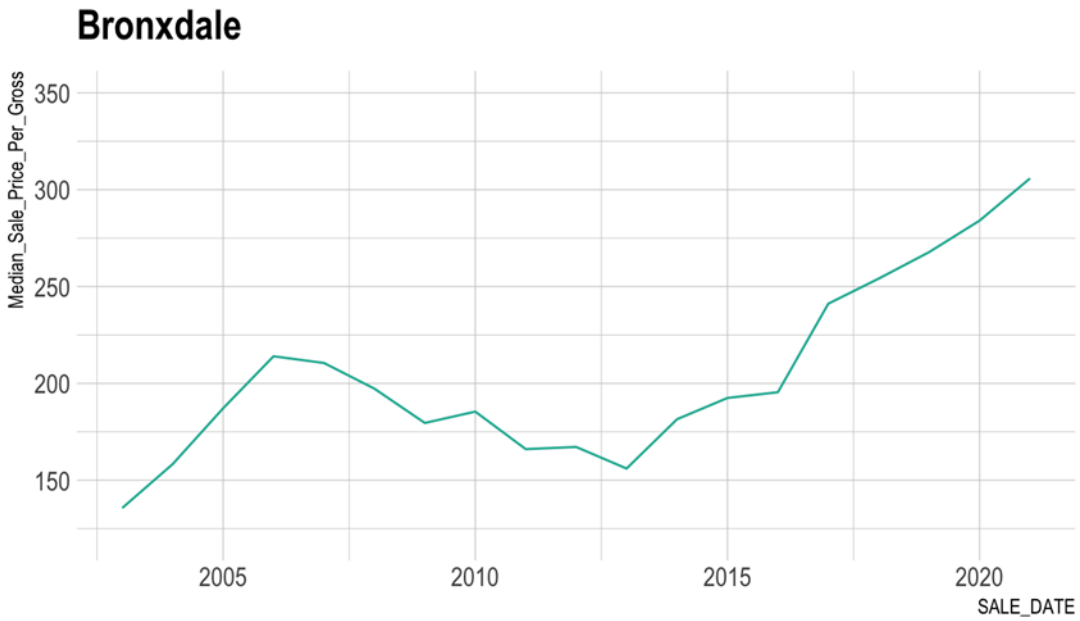
Between 2010 and 2014, there seems to be relative stability in the sale price, with limited growth and less variance among neighborhoods. This could indicate a period of economic stability or stagnation, depending on broader economic indicators.

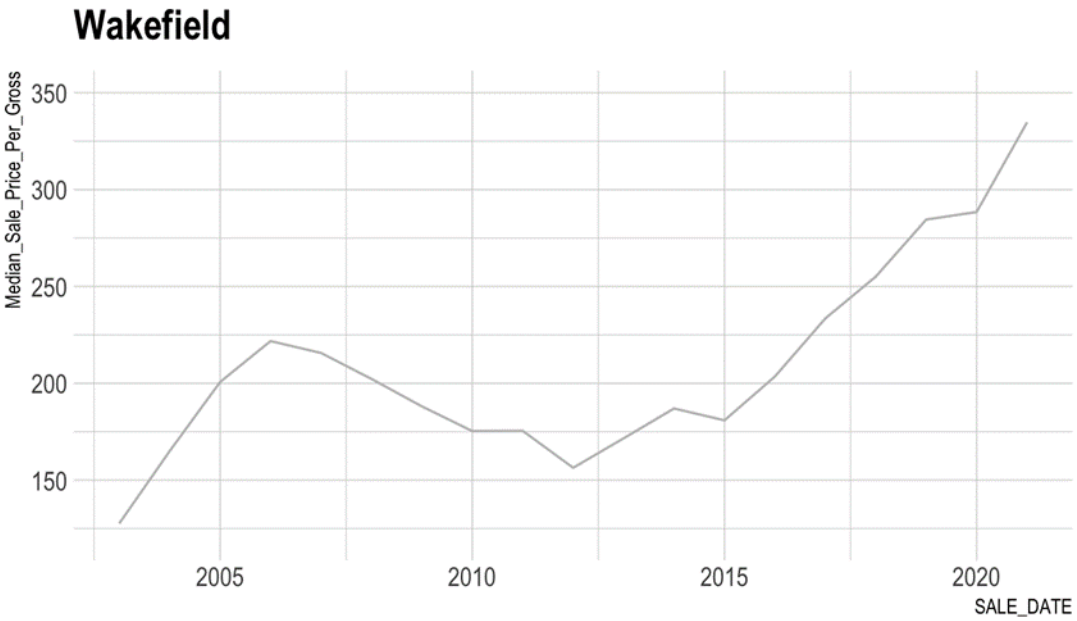
KPI 6：

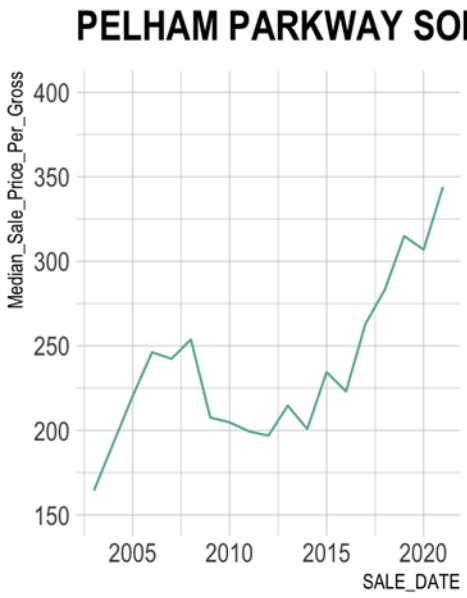

Annual Median Sale Price Per Gross Square Foot of residential properties
Values
800
600
400
200
0
2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
Year

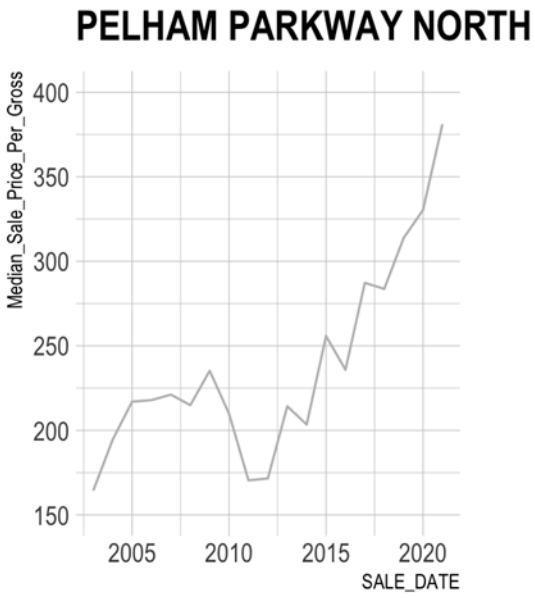

Westchester median sale price per gross square foot of residential properties
Median
300
250
200
150
2005
2010
2015
2020
Year


Williamsbridge median sale price per gross square foot of residential properties
Median
300
250
200
150
2005
2010
2015
2020
Year


Bronxdale
Median_Sale_Price_Per_Gross
350
300
250
200
150
2005
2010
2015
2020
SALE_DATE







The provided boxplot showcases the annual median sale price per gross square foot of residential properties spanning various neighborhoods in a specific borough from the years 2003 to 2021. By closely observing the visual representation, we can deduce several key trends and insights about the property market within this borough over the analyzed period.

One of the most noticeable trends is the consistent and gradual increase in the median sale price per gross square foot from 2003 to 2021. The median value, which represents the middle point of data, seems to have risen from a range below 200 in the early 2000s to a range approaching 400 by 2021. This indicates an appreciation in property values across the borough's neighborhoods over the years.

Between 2003 and 2010, the boxplot displays narrower boxes, suggesting that there was relatively less variability in the median sale prices across the neighborhoods. This might hint at a more stable and predictable property market during these years.

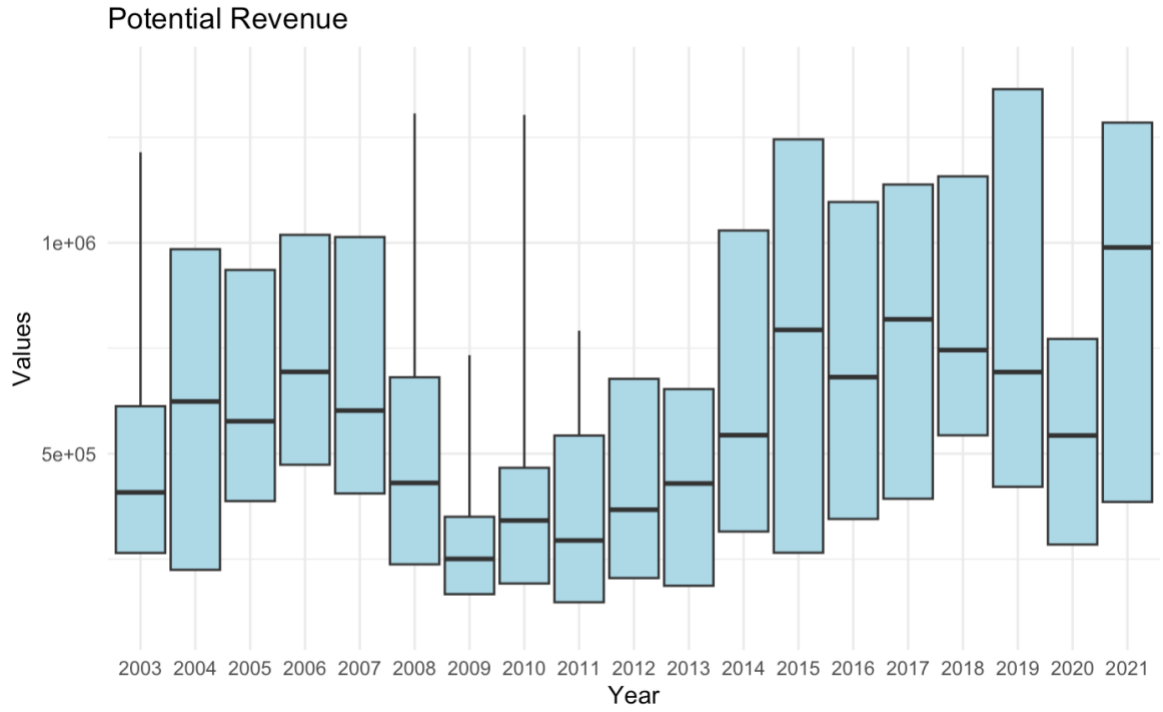
Starting from around 2011, the height of the boxes, which signifies the interquartile range, becomes more varied. This increased variability indicates that while some neighborhoods witnessed significant appreciation in property values, others might have seen lesser growth or even stagnation.

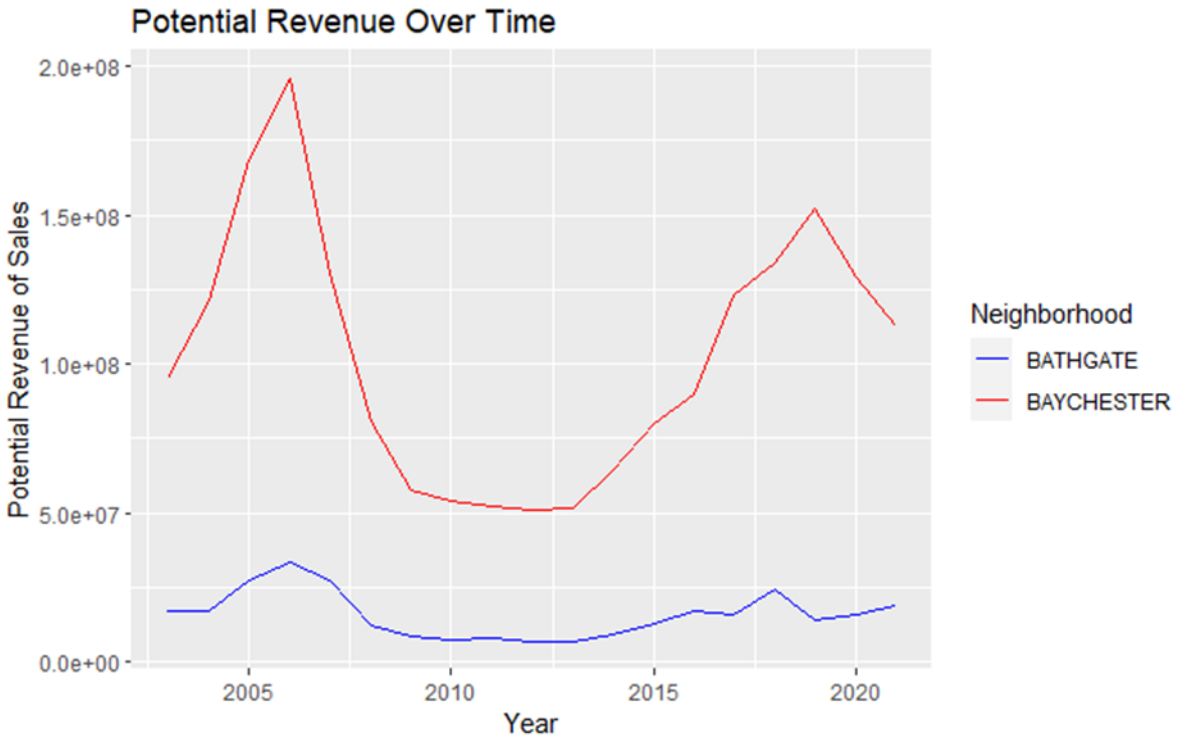
While not prominent, there seem to be a few years where outliers (data points that lie beyond the general trend) are present. These outliers suggest that in those specific years, certain neighborhoods had property sale prices that were exceptionally high or low compared to the general trend. Further research would be necessary to understand the reasons behind these anomalies.

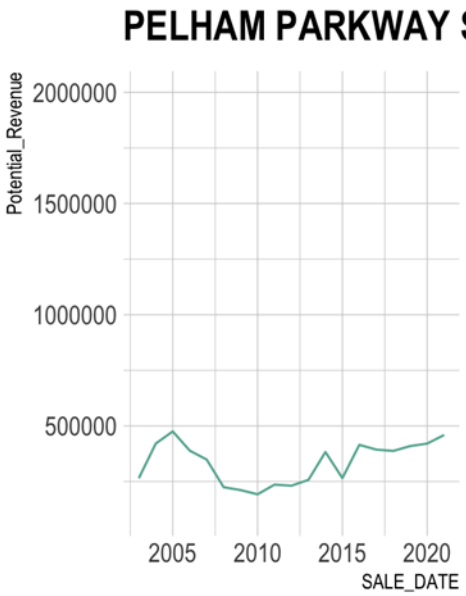
In conclusion, the borough's property market, as visualized through the boxplot, has witnessed a steady appreciation in the median sale price per gross square foot of residential properties over the past two decades. However, post-2010, there has been a more pronounced variability in these prices across different neighborhoods. This could be indicative of socio-economic changes, infrastructural developments, or other external factors influencing certain areas more than others. To gain a comprehensive understanding, it would be beneficial to delve deeper into local developments, policies, and events that might have impacted the property market during the years of increased variability.

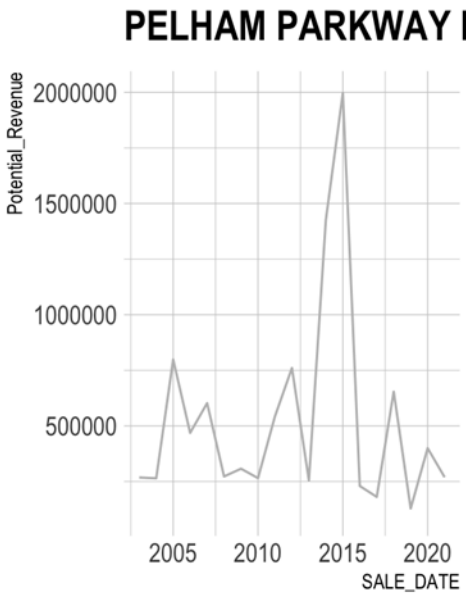
Compared with the neighborhoods we chose, the trends and data of the annual median sale price are all similar to the total trend and data, which means they are representative for our borough.

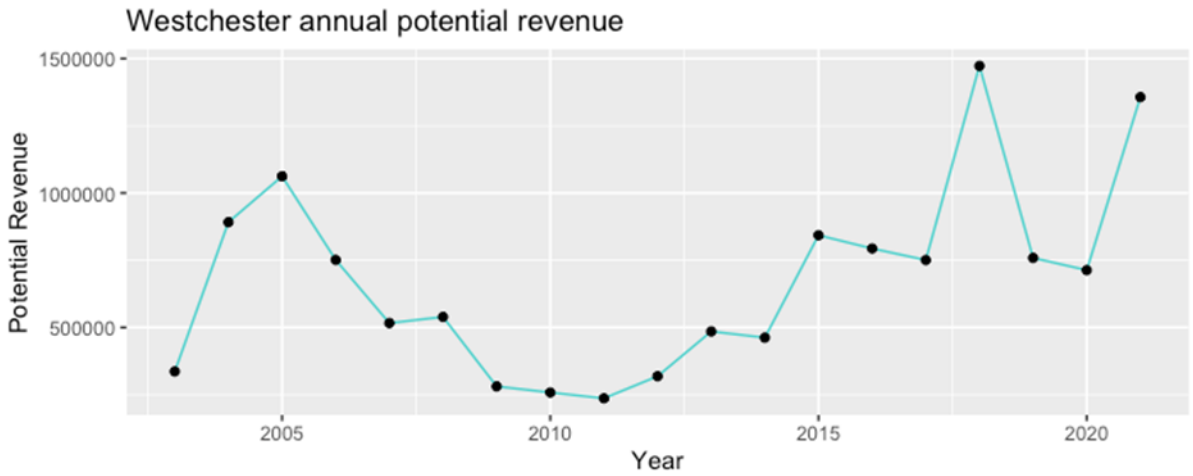
Potential revenue:

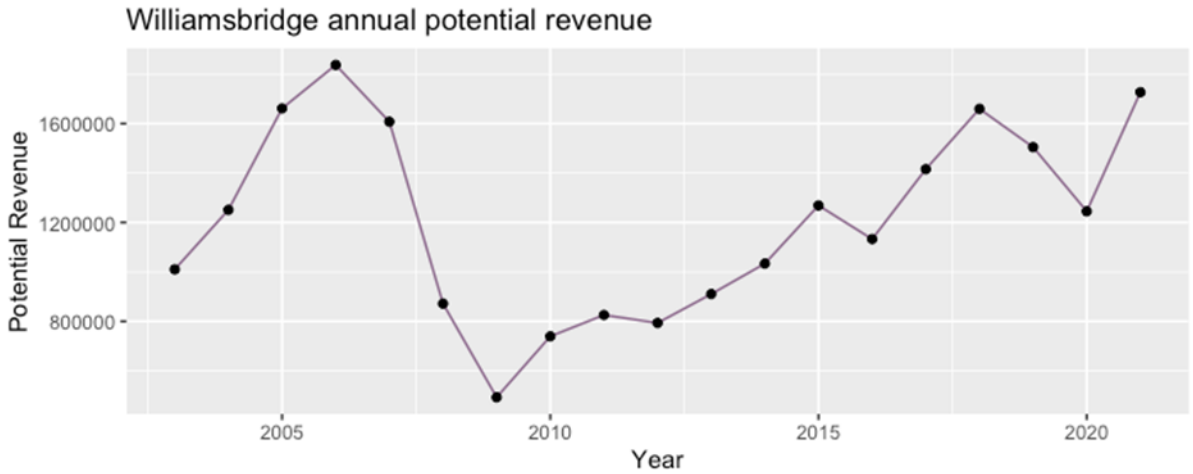



Potential Revenue Over Time
Potential Revenue of Sales
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
2005
2010
2015
2020
Year
Neighborhood
BATHGATE
BAYCHESTER


PELHAM PARKWAY SORTH
Potential_Revenue
2000000
1500000
1000000
500000
2005
2010
2015
2020
SALE_DATE


PELHAM PARKWAY NOR
Potential_Revenue
2000000
1500000
1000000
500000
2005
2010
2015
2020
SALE_DATE





The boxplot presented offers a visualization of the potential revenue across various neighborhoods of a borough from 2003 to 2021. By dissecting the patterns and deviations within the chart, we can gather a series of noteworthy insights and trends that have emerged over the nearly two-decade span.

Initially, from 2003 to around 2007, there's a steady upward trajectory in the potential revenue. The median values, as indicated by the horizontal line within the boxes, rise consecutively, highlighting an environment of economic growth or increased property values in the borough.

Post-2007, until roughly 2014, there appears to be a period of fluctuation. The potential revenue displays both peaks and troughs, with a noticeable dip around 2011. This could be indicative of economic instability or changing dynamics in the property market during these years.

From 2014 onward, there is a resurgence in potential revenue, reaching its peak around 2019. Additionally, the years from 2017 to 2019 demonstrate a relatively reduced variability in revenue across neighborhoods, possibly suggesting a period of economic stability or consistent growth across the borough.

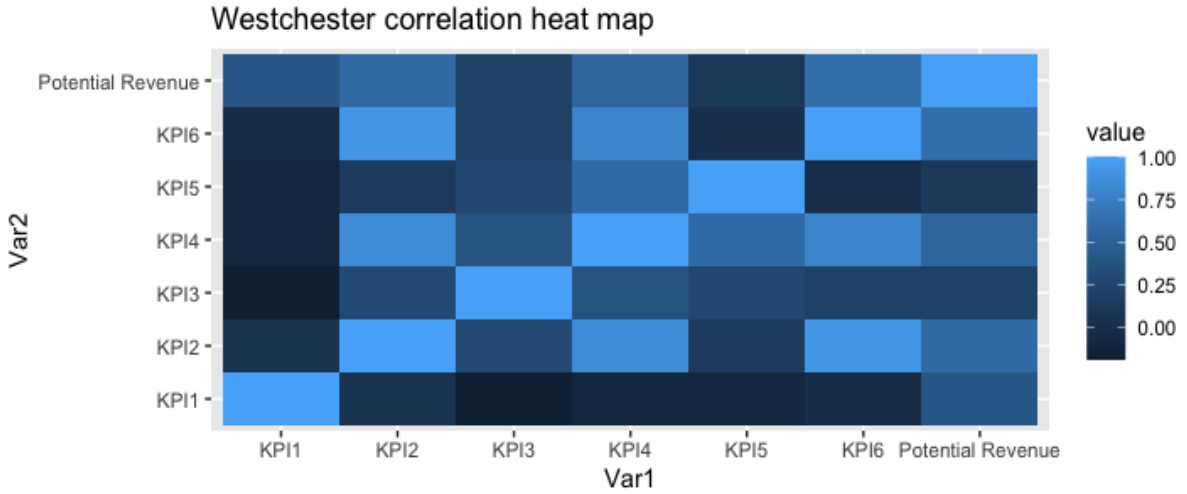
The years 2020 and 2021, however, show a pronounced variability in potential revenue. While the upper quartile indicates neighborhoods with high potential revenue, the lower quartile suggests that some neighborhoods are still lagging behind, indicating disparities in economic growth or property valuation within the borough.

Throughout the years, particularly noticeable between 2007 and 2014, there are several outliers. These data points represent neighborhoods where the potential revenue drastically deviates from the general trend. Such anomalies could be due to specific local events, infrastructural developments, or other unique factors impacting these areas.

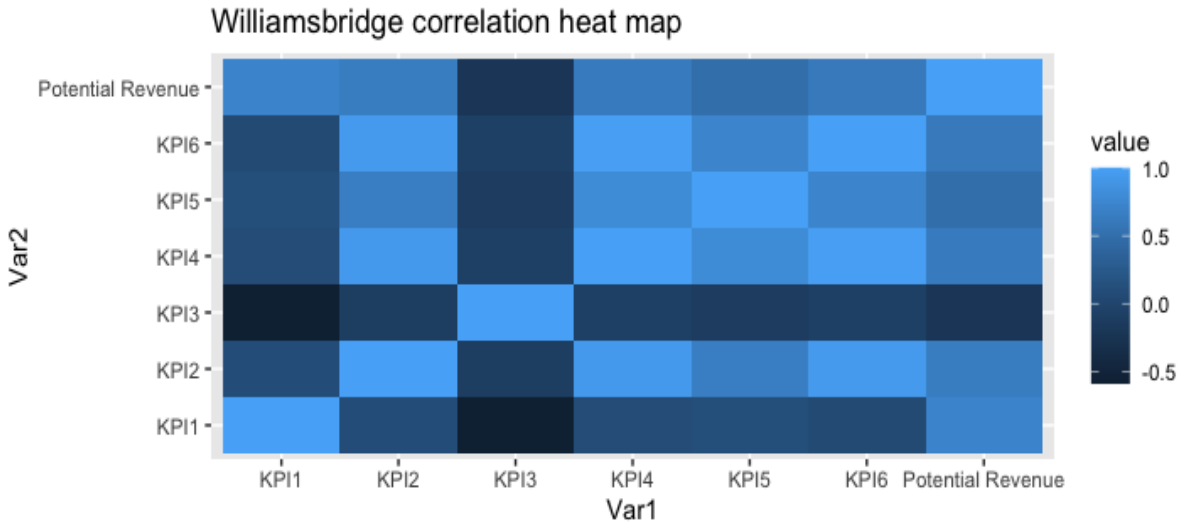
In conclusion, the potential revenue across neighborhoods in the borough has experienced phases of steady growth, fluctuations, and eventual resurgence. The recent years' increased variability suggests that while some neighborhoods are thriving, others may be facing challenges. For a more holistic understanding, it would be imperative to explore underlying factors, such as policy changes, infrastructural projects, and local events, which might have contributed to these revenue patterns over the years.

Compared with the neighborhoods we chose, Westchester has the most similar trends and data for the potential revenues over the years. In this method of calculating the potential revenues, West Chester is the most similar neighborhood for the representative.

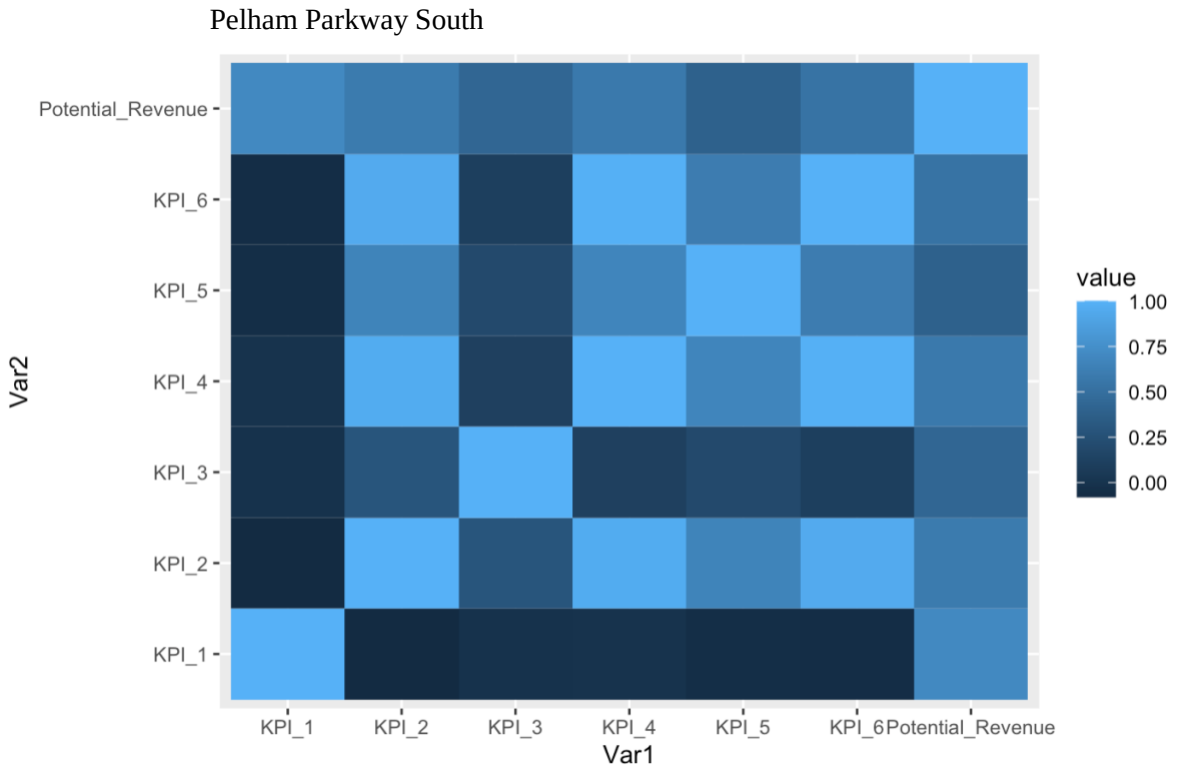
Heating map:



Most KPIs appear to have a moderate correlation with each other, as evidenced by the consistent blue tones.
The darkest shades (indicating higher correlation) are seen among adjacent KPIs, such as between KPI2 & KPI3 or KPI5 & KPI6.
'Potential Revenue' has varying levels of correlation with different KPIs. It seems to have a stronger correlation with KPI2, KPI4, and KPI6.



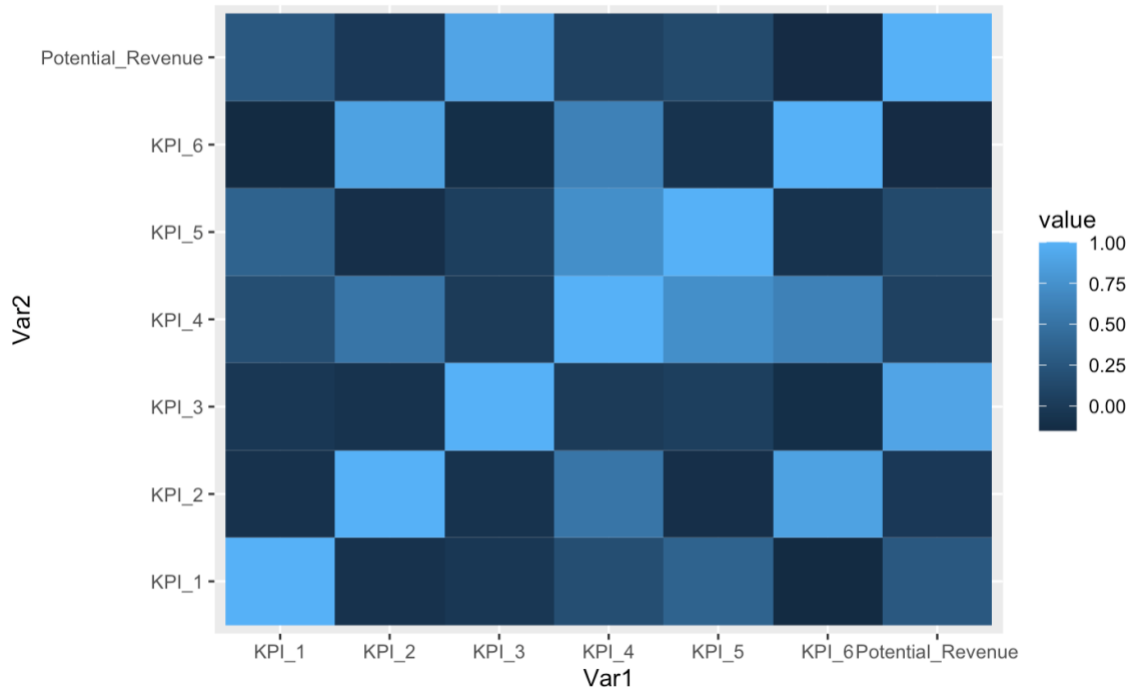
There are clear differences in correlation strengths compared to the Westchester heatmap.
Darker shades (indicative of stronger correlations) are observed between 'Potential Revenue' and KPI2, KPI3, and KPI4.
The correlation between 'Potential Revenue' and the other KPIs (especially KPI5 & KPI6) seems weaker here than in the Westchester heatmap.



KPI_1 has strong values across the board, especially for itself, indicating a potentially self-reinforcing metric or a strong correlation with itself.

Potential_Revenue has a medium to high correlation with most KPIs. This suggests that as these KPIs improve or increase, there might be a potential increase in revenue.
KPI_4 and KPI_5 seem to have a strong correlation, as indicated by the dark shade.
KPI_6 appears to be less correlated with other KPIs, with many cells showing lighter colors.
There is a distinct dark cell for KPI_2 with KPI_5 indicating a strong relationship or correlation between these two metrics.
The columns for KPI_3 and KPI_5 have a mix of light and dark cells, suggesting they have varied relationships with the other KPIs.

Pelham Parkway North



Potential_Revenue has decent positive correlations with all KPIs, suggesting that all KPIs have some level of influence on the potential revenue.

KPI_1 demonstrates a strong correlation with itself, as expected. It also has notable correlations with other KPIs.

KPI_4 and KPI_5 exhibit a strong positive correlation, indicating that they move in the same direction.

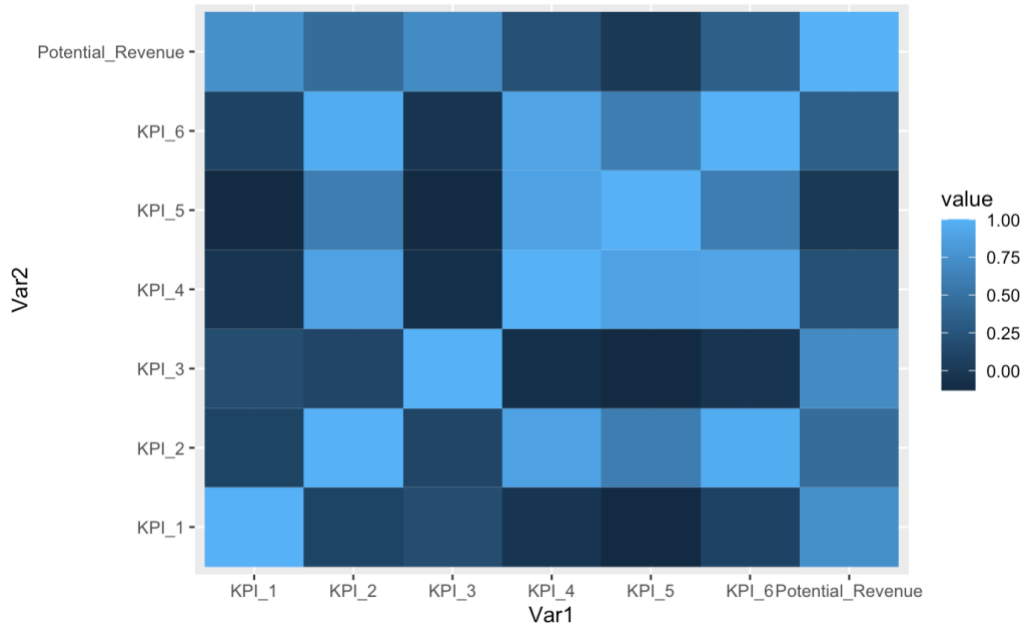
KPI_6 seems to have varied relationships with other KPIs, with both positive and negative correlations observed.

KPI_2 and KPI_5 also have a strong correlation, which means changes in one might be associated with changes in the other.

KPI_3 displays mixed correlations, with both positive and negative correlations with other KPIs.

To conclude, it's evident that all KPIs have some influence on Potential_Revenue, with some KPIs having a stronger relationship with each other. When interpreting this heatmap, it's crucial to remember that correlation doesn't imply causation. So while two variables might move together, it doesn't necessarily mean one causes the other to change. Further analysis, perhaps through regression modeling, would be needed to uncover any causal relationships.

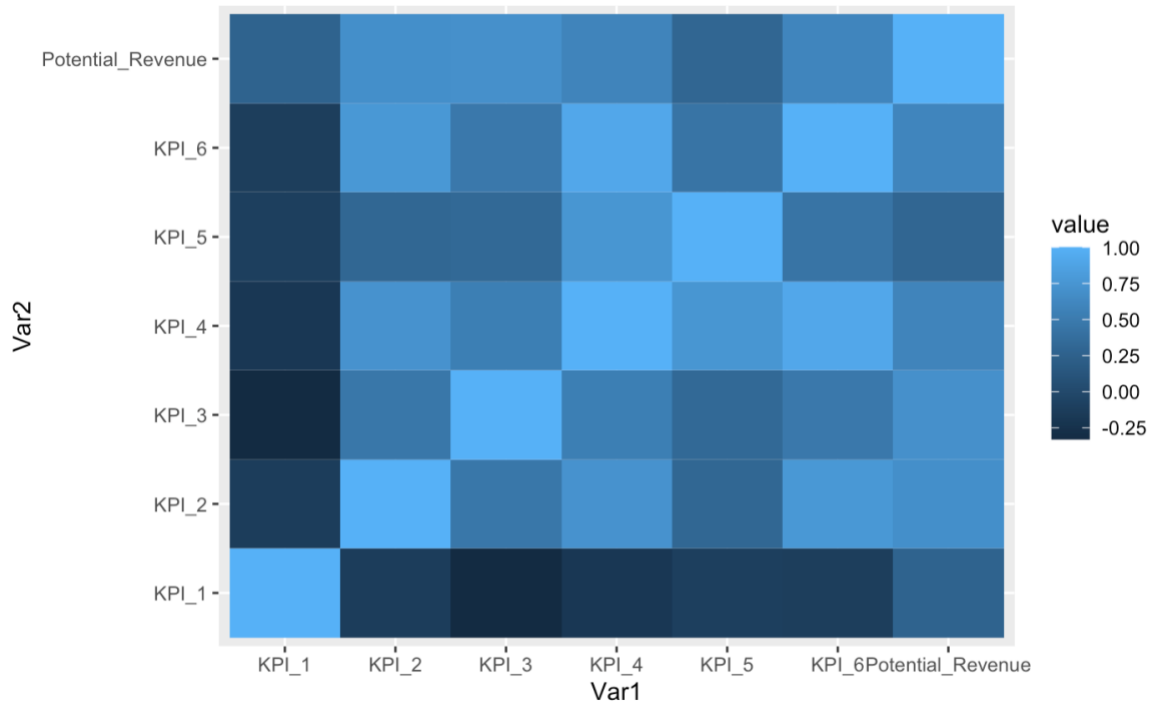
Bathgate



While some KPIs like KPI_1, KPI_4, and KPI_6 have positive correlations with Potential_Revenue, others don't show a strong direct relationship. Additionally, some KPIs are correlated with each other, like KPI_1 and KPI_2 or KPI_3 and KPI_5.

However, it's essential to understand that correlation does not imply causation. While these variables may move together, it doesn't necessarily mean one causes the other to change. Further in-depth analysis, potentially using regression modeling or causal inference methods, would help provide more concrete insights.

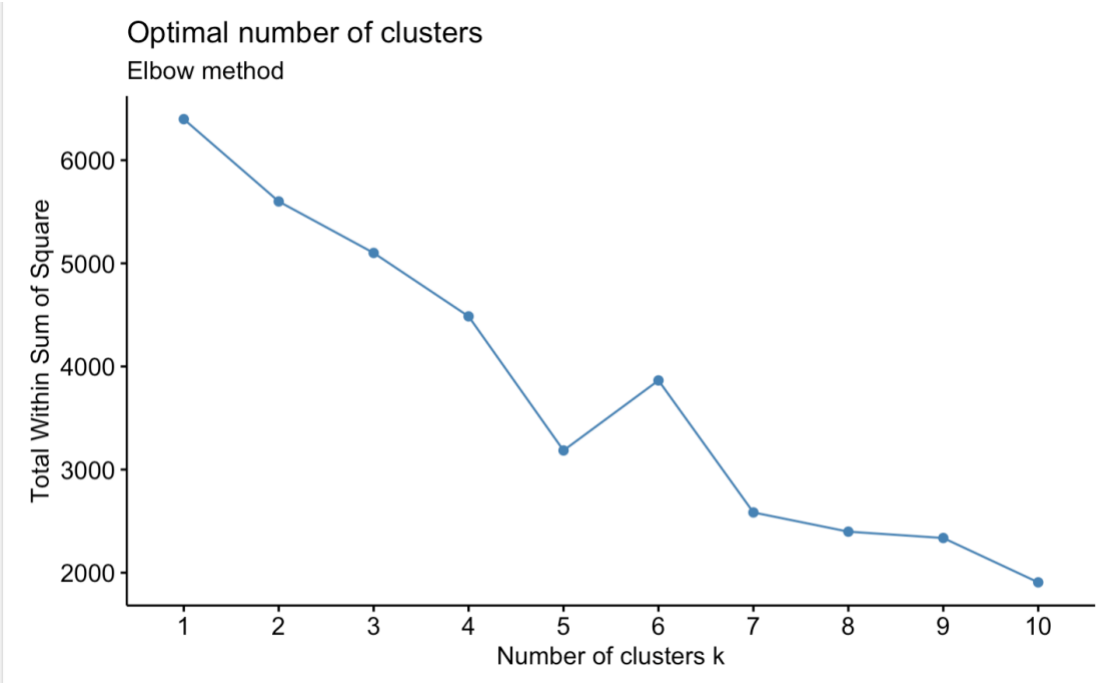
Baychester

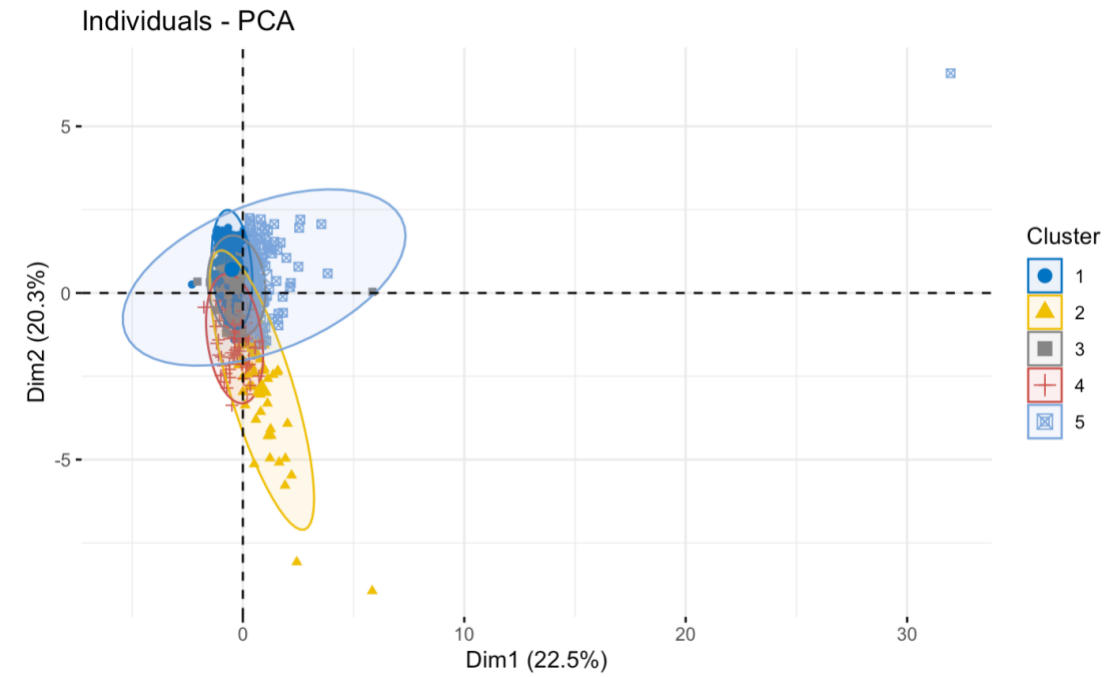


Potential_Revenue:Appears to have a moderately strong positive correlation with KPI_6 (light blue shade).Shows a mild positive correlation with KPI_4 and KPI_1.Has minimal or no correlation with KPI_2, KPI_3, and KPI_5 (darker shades).
KPI_1:Exhibits a strong positive correlation with itself.Shows little to no correlation with other KPIs except a slight positive correlation with Potential_Revenue.
KPI_2:Generally has minimal correlations with other variables, including Potential_Revenue.
KPI_3:Displays minimal or no correlations with other KPIs and Potential_Revenue.
KPI_4:Shows a positive correlation with Potential_Revenue.Displays minimal or no correlations with other KPIs.
KPI_5:Appears to have little to no correlation with all other variables, including Potential_Revenue.
KPI_6:Has a moderately strong positive correlation with Potential_Revenue.Displays minimal or no correlations with other KPIs.

In summary, KPI_6 seems to have the most pronounced positive correlation with Potential_Revenue. KPI_4 and KPI_1 also have mild positive correlations with Potential_Revenue. Other KPIs, namely KPI_2, KPI_3, and KPI_5, don't exhibit a significant relationship with Potential_Revenue based on this heatmap.

Cluster Analysis





We used the “elbow method” to test the optimal number of clusters we would use. In the resulting plot, the curve started to level off at k=5 and 6. We selected k=5 as our optimal number of clusters. Based on the clustering results, it can be seen that this still has a lot of points that overlap, indicating that the data does not compute a good clustering result.