

Project Assignment 4: Data Analytics Research Project

Puneeth Velivela

G#: G01462306

Data Analytics Engineering, George Mason University

AIT-580-DL2: Prof. Alla Webb

December 04, 2023

Index

1 Abstract - Page 3
2 Introduction - Page 3
3 Data Ingestion and Exploration (using SQL) - Page 4
SQL Schema – Page 4
Simple SQL Queries - Page 4
4 Univariate Analyses - Page 5
Numerical Columns Summary Statistics - Page 5
Ordinal Column Analysis - Page 5
Interval and Ratio Columns Analysis - Page 6
5 Research Questions - Page 6
Research Question 1 - Page 6
Research Question 2 - Page 7
Research Question 3 - Page 8
Research Question 4 - Page 9
6 Literature Search - Page 11
Research Paper 1 - Page 11
Research Paper 2 - Page 11
Research Paper 3 - Page 11
7 Results and Discussion - Page 11
Research Question 1 - Page 12
Research Question 2 - Page 12
Research Question 3 - Page 12
Research Question 4 - Page 12
8 Limitations - Page 12
9 Conclusion - Page 12
10 Appendix (List of Technical Terms) - Page 13
11 References - Page 14

Exploring Property Data in New York City: A Data Analytics Research Project

1. ABSTRACT:

Using the analytical capabilities of R, Python, and SQL, this Data Analytics Research Project investigates a large **dataset** ^[1] that lists properties in New York City. The investigation is centered on specific research questions, guiding the exploration, analysis, and interpretation of the dataset. The dataset is ready for informative analyses using both univariate and multivariate methods, along with statistical summaries and visualizations, after thorough data cleaning and preprocessing.

The research paper's structure demonstrates careful planning and adherence to best practices for explanatory clarity and graphic design. A review of the literature gives the project a contextual foundation and links it to previous studies in the topic. The study demonstrates abilities in a variety of data analytics techniques and methods by providing answers to the research questions as well as advancing an advanced knowledge of property dynamics in New York City.

2. INTRODUCTION:

In this Data Analytics Research Project, we explore a comprehensive **dataset** ^[1] that intricately captures the diverse landscape of properties in New York City. The dataset encompasses a wealth of information, ranging from tax classifications and building characteristics to geographic coordinates and community-related attributes. With New York City serving as a dynamic urban hub, this **dataset** ^[1] becomes a valuable lens through which we aim to gain insights into the multifaceted aspects of the city's real estate dynamics.

The relevance of this exploration lies in the potential to uncover patterns, correlations, and trends that can inform our understanding of property-related phenomena. By using the analytical capabilities of R, Python, and SQL, we start on a journey to extract valuable insights from this rich **dataset** ^[1]. The exploration is guided by specific research questions that dive into the relationships between tax classifications and building attributes, the predominant tax classes in different boroughs, the distribution of square footage across the city, and the interplay between street names, zip codes, and the number of buildings on a tax lot.

Below are the Research Questions that are intended to answer in this Research project.

Research Questions:

- Is there a correlation between the tax class of properties and the number of buildings located on the same tax lot, and if so, what patterns emerge?
- What is the predominant tax class for properties in each borough, and are there significant differences in the distribution of tax classes among the boroughs?
- How does the gross square footage of properties differ across the five boroughs, and are there any correlations between square footage and building class?

- Can we identify any relationships between street names, zip codes, and the number of buildings on a tax lot, and are there any interesting patterns or clusters of properties in specific areas?

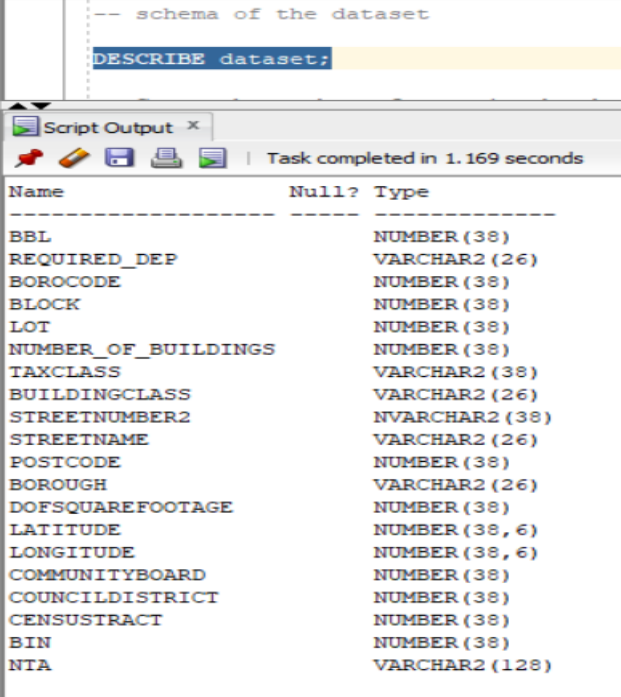
3. Data Ingestion and Exploration (using SQL ^[2]):

The dataset consists of property-related information in New York City, having columns such as "BBL," "Tax Class," "Building Class," "Street Name," "Postcode," and more. These columns capture diverse attributes, providing an overall view of properties across the city.

I have imported the **dataset** ^[1] into **SQL** ^[2].

Schema for the Dataset in SQL: Here is the SQL schema for the "dataset" table

```
-- schema of the dataset
DESCRIBE dataset;
```



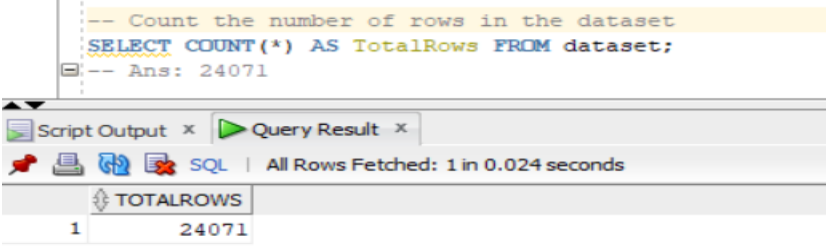
Name	Null?	Type
BBL		NUMBER (38)
REQUIRED_DEP		VARCHAR2 (26)
BOROCODE		NUMBER (38)
BLOCK		NUMBER (38)
LOT		NUMBER (38)
NUMBER_OF_BUILDINGS		NUMBER (38)
TAXCLASS		VARCHAR2 (38)
BUILDINGCLASS		VARCHAR2 (26)
STREETNUMBER2		NVARCHAR2 (38)
STREETNAME		VARCHAR2 (26)
POSTCODE		NUMBER (38)
BOROUGH		VARCHAR2 (26)
DOFSQUAREFOOTAGE		NUMBER (38)
LATITUDE		NUMBER (38, 6)
LONGITUDE		NUMBER (38, 6)
COMMUNITYBOARD		NUMBER (38)
COUNCILDISTRICT		NUMBER (38)
CENSUSTRACT		NUMBER (38)
BIN		NUMBER (38)
NTA		VARCHAR2 (128)

The above query gives the column names along with their data types in the dataset.

Let's execute a few simple SQL queries to understand the dataset.

```
-- Count the number of rows in the dataset
SELECT COUNT(*) AS TotalRows FROM dataset;
```

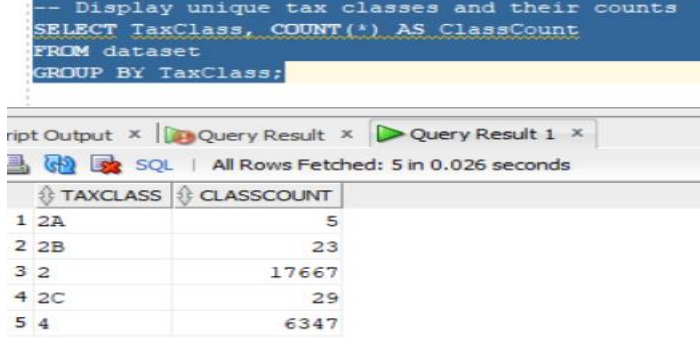
```
-- Ans: 24071
```



TOTALROWS
1 24071

The above query gives the total number of rows in the dataset.

```
-- Display unique tax classes and their counts
SELECT TaxClass, COUNT(*) AS ClassCount
FROM dataset
GROUP BY TaxClass;
```



TAXCLASS	CLASSCOUNT
1 2A	5
2 2B	23
3 2	17667
4 2C	29
5 4	6347

Now here we can see the total number of classes in the dataset and the count of observations in each class.

4. Univariate Analyses:

Now let's perform univariate analyses for each type of **NOIR** ^[3] data (Numerical, Ordinal, Interval, and Ratio) on the dataset and use appropriate statistical summaries and visualizations in **python** ^[4].

Numerical columns summary statistics:

Output:

```
In [50]: numerical_summary
Out[50]:
```

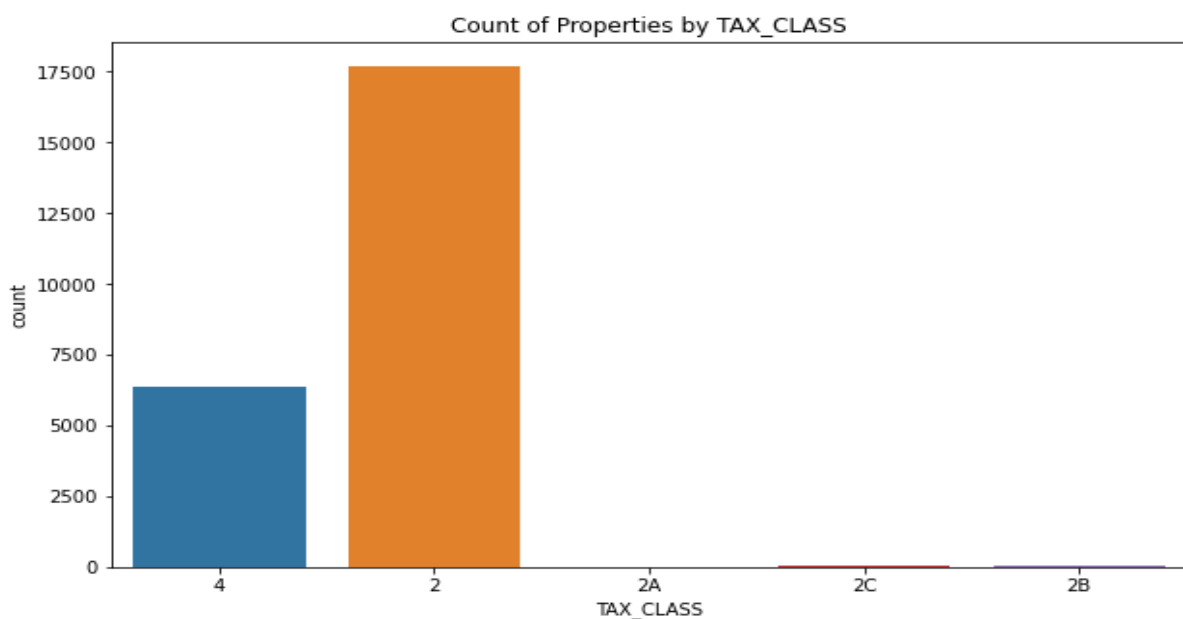
	NUMBER_OF_BUILDINGS	DOF_SQ_FOOTAGE	LATITUDE	LONGITUDE
count	24071	24071	23943	23943
unique	42	17267	22721	22201
top	1	30,000	40.744821	-73.970914
freq	22936	104	5	4

Ordinary column analysis:

In the dataset there is only one ordinary class i.e Tax Class columns. Let's plot a bar plot of the tax class of how many observations are there for each class.

```
In [57]: ordinal_counts
Out[57]:
```

TAX_CLASS	
2	17667
4	6347
2C	29
2B	23
2A	5



Interval and Ratio Columns Analysis:

```
In [65]: interval_ratio_summary
Out[65]:
```

	BLOCK	LOT	POSTCODE	COUNCIL_DISTRICT	CENSUS_TRACT	BIN
count	24071	24071	24071	23943	23943	23520
unique	5884	599	193	51	1136	22995
top	2180	1	0	4	33	3000000
freq	68	2753	1854	1800	204	138

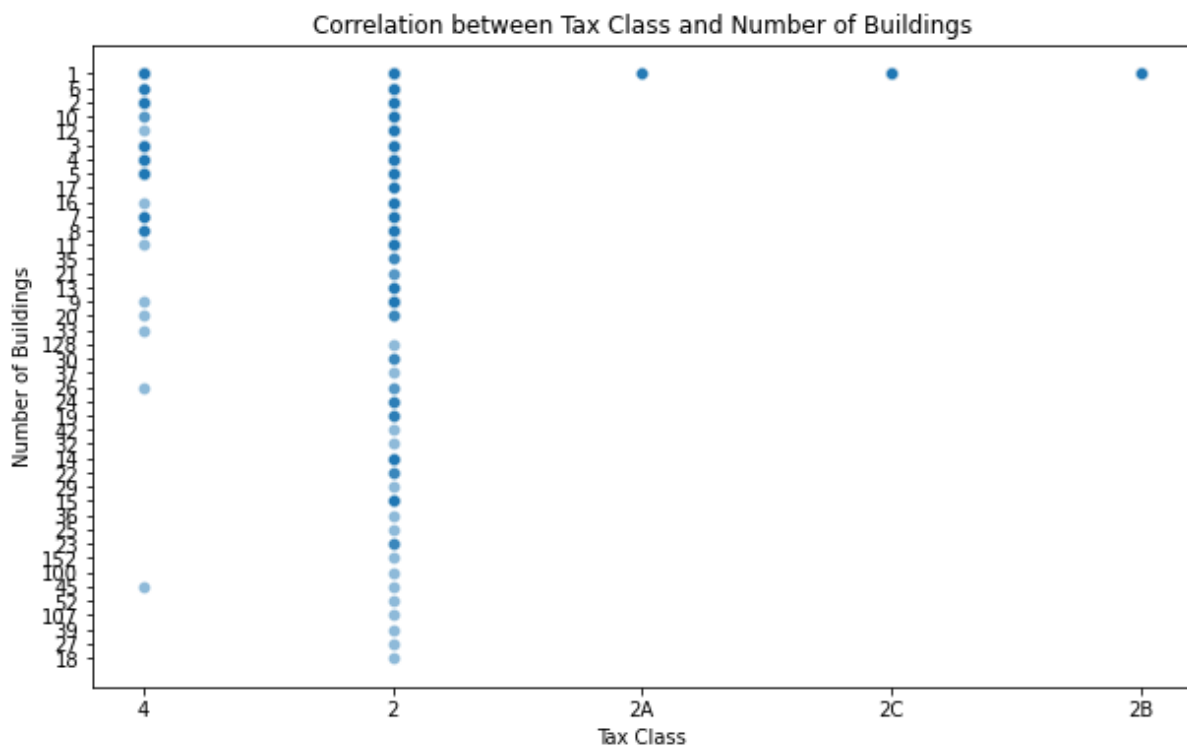
5. Research Questions:

Now let's start answering our research questions using **Python** ^[4] and **R** ^[5]

Research Question 1:

“Is there a correlation between the tax class of properties and the number of buildings located on the same tax lot, and if so, what patterns emerge?”

I have generated a scatter plot between tax class and number of buildings and below is the graph.



Now, let's calculate correlation coefficient in python to check the relation between tax class and number of buildings.

Code:

```
# Calculate correlation coefficient
correlation_coefficient = selected_data['TAX_CLASS'].corr(selected_data['NUMBER_OF_BUILDINGS'])
print(f"Correlation Coefficient: {correlation_coefficient}")
```

Output:

```
In [82]: print(f"Correlation Coefficient: {correlation_coefficient}")
Correlation Coefficient: -0.032831971558564745
```

Upon investigating the dataset, we observed a correlation coefficient of approximately -0.033 between the tax class of properties and the number of buildings on the same tax lot. This

indicates a weak negative correlation. The scatter plot illustrates scattered data points without a clear linear trend, supporting the modest correlation coefficient.

The negative sign implies that as the tax class decreases, there is a slight tendency for the number of buildings on the tax lot to increase, and vice versa. However, it's crucial to note that the correlation is not strong, suggesting that other factors contribute significantly to the variation in the number of buildings.

This finding implies that, on average, there is a subtle association between tax class and the density of buildings on a tax lot.

Research Question 2:

“What is the predominant tax class for properties in each borough, and are there significant differences in the distribution of tax classes among the boroughs?”

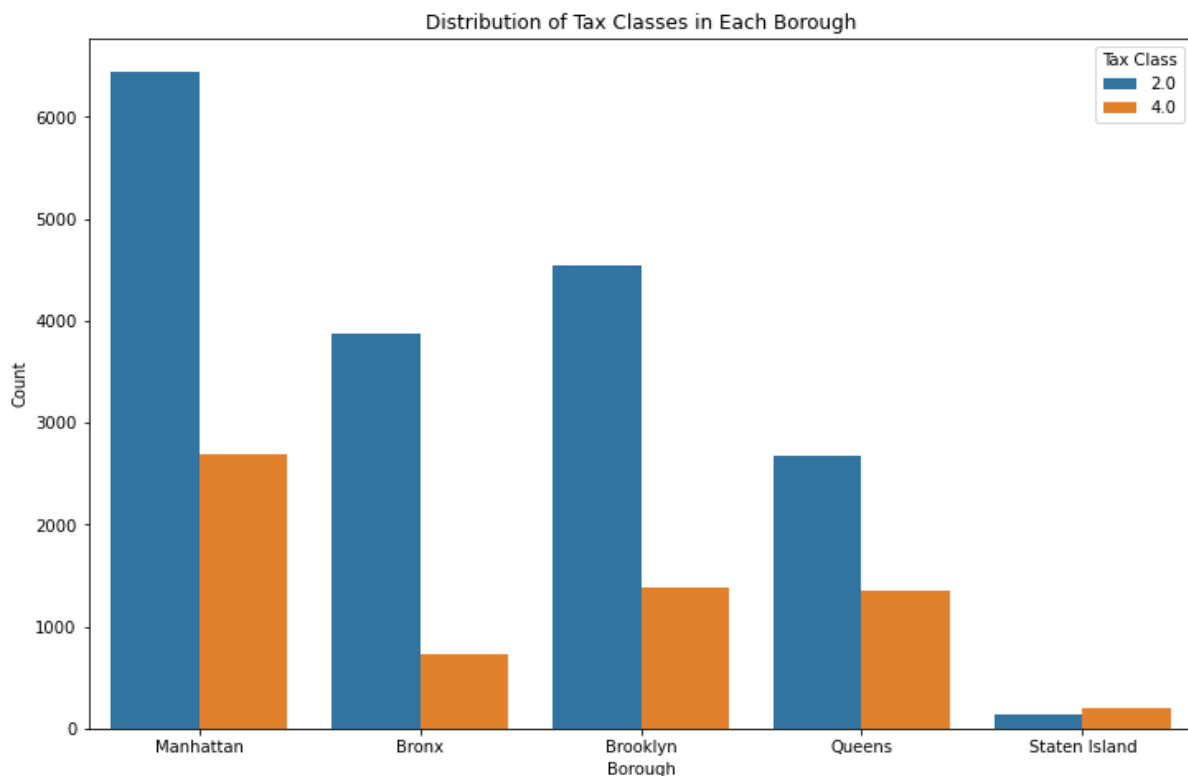
Python ^[4] code:

```
#Research Question2
# Select relevant columns
selected_columns2 = ['BOROUGH', 'TAX_CLASS']

# Drop rows with missing values in selected columns
selected_data2 = df[selected_columns2].dropna()

# Plotting the distribution of Tax Classes in each Borough
plt.figure(figsize=(12, 8))
sns.countplot(data=selected_data2, x='BOROUGH', hue='TAX_CLASS')
plt.title('Distribution of Tax Classes in Each Borough')
plt.xlabel('Borough')
plt.ylabel('Count')
plt.legend(title='Tax Class')
plt.show()
```

Plot:



The above analysis reveals the predominant tax class for properties in each borough:

Bronx: The predominant tax class in the Bronx is Class 2.0.

Brooklyn: Brooklyn also exhibits a predominant tax class of 2.0.

Manhattan: Similar to the Bronx and Brooklyn, Manhattan predominantly has properties classified under Tax Class 2.0.

Queens: The predominant tax class in Queens is Class 2.0.

Staten Island: Staten Island, on the other hand, stands out with a predominant tax class of 4.0.

Significant Differences in Tax Class Distribution:

While the predominant tax class is consistent (Class 2.0) across Bronx, Brooklyn, Manhattan, and Queens, Staten Island deviates with a predominant tax class of 4.0. This discrepancy suggests that Staten Island may have a distinct property tax classification pattern compared to the other boroughs.

The distribution plot visually highlights the prevalence of Tax Class 2.0 across multiple boroughs, emphasizing its dominance in the dataset. However, the unique prevalence of Tax Class 4.0 in Staten Island indicates a potential area of interest for further investigation into borough-specific property tax policies or characteristics. Further statistical tests could be employed to formally assess the significance of these differences.

In summary, the analysis provides insights into the predominant tax classes in each borough and suggests noteworthy distinctions, particularly in Staten Island, warranting further exploration into the factors contributing to these differences.

Research Question 3:

“How does the gross square footage of properties differ across the five boroughs, and are there any correlations between square footage and building class?”

To address this question let's build a linear regression model in R [5].

Code:

```
# Convert necessary columns to numeric and factor
newYork$DOFSquareFootage <- as.numeric(newYork$DOFSquareFootage)
newYork$BuildingClass <- as.factor(newYork$BuildingClass)
class(newYork$BuildingClass)
# Linear regression model
model <- lm(DOFSquareFootage ~ Borough + BuildingClass, data = newYork)
# Summary of the regression model
summary(model)
```

The linear regression model provides valuable insights into the relationships between gross square footage, boroughs, and building classes in the dataset. The coefficients for each variable indicate the estimated change in square footage associated with a unit change in the corresponding predictor, holding other variables constant.

The intercept represents the estimated square footage for the reference categories (intercept, Brooklyn, and BuildingClassA) when all other predictors are zero. The coefficients for each borough indicate the estimated change in square footage compared to the reference borough (Bronx). For example, properties in Brooklyn have an estimated decrease of 17,651.8 square feet compared to those in the Bronx.

Similarly, the coefficients for different building classes show how square footage varies with each class, relative to the reference class (BuildingClassA). For instance, BuildingClassD3 has a substantial positive coefficient of 142,747.4, suggesting a significant increase in square footage compared to BuildingClassA.

The overall model is statistically significant ($p\text{-value} < 2.2e-16$), indicating that at least one predictor variable significantly contributes to explaining the variance in square footage. However, the relatively low adjusted R-squared (0.1115) suggests that the model explains only a modest proportion of the variability in square footage, highlighting the complexity of predicting square footage based on boroughs and building classes alone.

Let's perform ANOVA test significance on the model.

Result:

```
> # ANOVA to test significance of the model
> anova(model)
Analysis of Variance Table

Response: DOFSquareFootage
          Df      Sum Sq   Mean Sq F value    Pr(>F)
Borough      4 2.2358e+13  5.5894e+12 106.676 < 2.2e-16 ***
BuildingClass 115 1.4219e+14 1.2364e+12  23.598 < 2.2e-16 ***
Residuals 23951 1.2549e+15  5.2396e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

The analysis of variance (ANOVA) results provides valuable insights into the significance of the regression model used to explore the relationship between square footage (DOFSquareFootage) and the predictor variables (Borough and BuildingClass).

The Borough variable is highly significant in explaining the variability in square footage. The low p-value ($\text{Pr}(>F)$) suggests that there are significant differences in square footage among different boroughs. The BuildingClass variable is also highly significant. The low p-value indicates that different building classes significantly contribute to explaining the variance in square footage.

The residuals represent the unexplained variability in square footage not accounted for by the model. The F-value for residuals is not calculated as it serves as a baseline for comparison against the other variables.

In conclusion, both Borough and BuildingClass are crucial factors in predicting square footage, and the overall model is highly significant. The results support the idea that the chosen predictor variables are important in understanding the variations in square footage across the dataset.

Research Question 4:

“Can we identify any relationships between street names, zip codes, and the number of buildings on a tax lot, and are there any interesting patterns or clusters of properties in specific areas?”

Using **R** ^[5] let's do cluster analysis.

Scatter plots for number of building vs street names and zip code. This visually represents how many buildings are there for each street and each zip code.

Using clustering algorithms like hierarchical clustering to identify patterns.

Code:

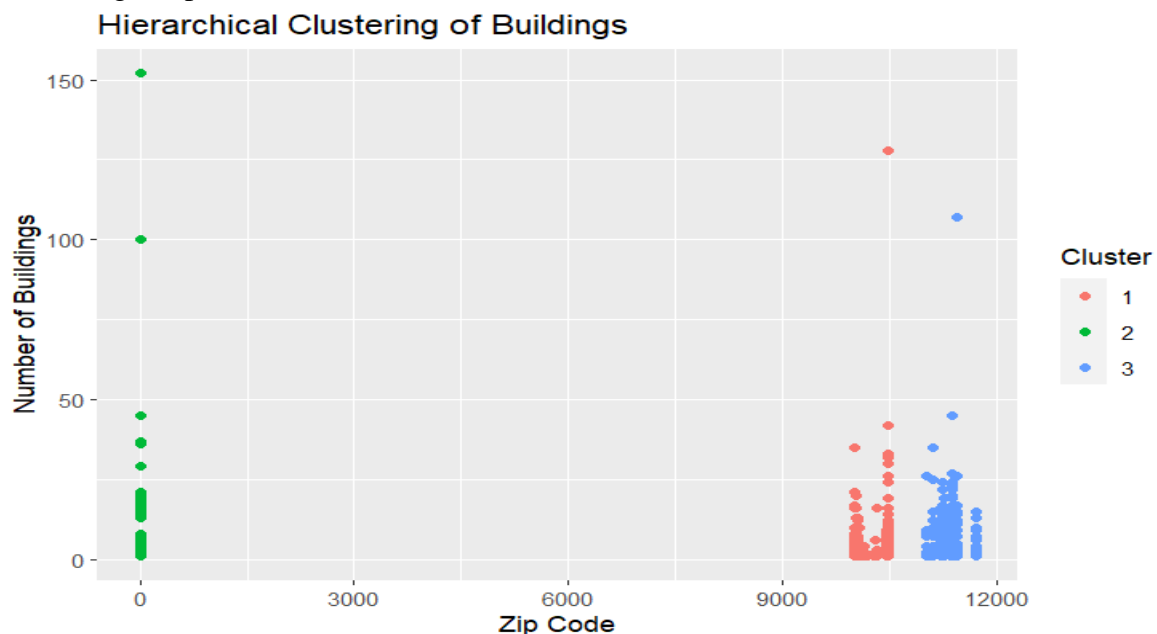
(cont....)

```

18 #research question 4
19 newYork <- na.omit(newYork)
20 # Check for infinite values in the 'Street name' variable
21 sum(!is.finite(newYork$StreetName))
22 # Convert 'Street' to a factor if it's a character variable
23 newYork$StreetName <- as.factor(newYork$StreetName)
24
25 cluster_data <- newYork[, c("Postcode", "NumberofBuildings")]
26
27 # Perform hierarchical clustering
28 hc <- hclust(dist(cluster_data), method = "ward.D2")
29 num_clusters <- 3
30 clusters <- cutree(hc, num_clusters)
31 newYork$Cluster <- as.factor(clusters)
32 library(ggplot2)
33 # Visualize the clusters
34 ggplot(newYork, aes(x = Postcode, y = NumberofBuildings, color = Cluster)) +
35   geom_point() +
36   labs(title = "Hierarchical Clustering of Buildings",
37        x = "Zip Code",
38        y = "Number of Buildings",
39        color = "Cluster")
40
41
42

```

Clustering Graph:



The hierarchical clustering analysis has identified three distinct clusters in our dataset, represented by the colors green, red, and blue in the graph. The placement and grouping of points within each cluster provide insights into the relationships between 'Zip Code' and 'Number of Buildings' for the properties in your dataset.

Green Cluster: Positioned on the total left of the graph, the green cluster consists of points that are more distant from each other. This indicates that properties in this cluster have significantly different characteristics regarding the number of buildings and their associated zip codes.

Red and Blue Clusters: The red and blue clusters, positioned towards the right side, are closer together. This suggests that properties within these clusters share more similarities in terms of the number of buildings and zip codes.

The proximity of the red and blue clusters might indicate that they share some common characteristics, possibly similar building structures or distribution across zip codes.

The grouped nature of points within the red and blue clusters implies a higher degree of similarity among properties in these clusters compared to the green cluster.

In summary, the hierarchical clustering has helped identify distinct groups of properties based on the specified variables. The differences in the placement and grouping of points provide valuable insights into the patterns and relationships within your dataset. Further analysis or domain knowledge may be applied to understand the specific characteristics that differentiate the green cluster from the red and blue clusters.

6. Literature Search:

Research Paper 1: "Why You Should Pay Attention to Properties in Tax Classes 2A and 2B" [6]

This paper delves into the significance of property allocation for tax purposes, specifically focusing on properties falling within Tax Classes 2A and 2B. The report highlights that the city imposes limitations on tax increases for these properties, providing insights into tax breakdowns and benefits for real estate investors in New York City. Despite not directly addressing the research questions outlined, the paper contributes valuable information about tax classification and its economic implications for investors, particularly in the context of property conversions into condominiums.

Research Paper 2: "Taxation of New York City Real Property" [7]

This report comprehensively explores various aspects of real estate taxation in New York City, covering the appraisal system, challenges in property appraisal, tax benefit programs, and the importance of real estate tax literacy. It emphasizes the complexity of real estate transactions, comprising more than 40% of the city's revenue. The paper includes a lesson plan outlining the role of the New York City Department of Finance and key dates in the valuation schedule. While providing information on different tax classes, it primarily focuses on the assessment and taxation process and does not directly address the correlation between tax class and the number of buildings.

Research Paper 3: "An Old, Unfair System: New York City's Property Tax Conundrum - Part II - Deep and Complex Inequities" [8]

This report delves into the complexities and inequities of the property tax system in New York City, emphasizing challenges related to complexity, inequality, and disparity. It provides insights into the historical development of the city's property tax system, efforts to address disparities, and a property tax classification system with different assessment methods. While not directly addressing all research questions, it sheds light on challenges influencing tax burdens in different neighborhoods and disparities in property taxes and square footage. The paper is valuable in understanding broader property tax issues and inequalities, although specific data connections may require additional analysis or exploration from other sources.

7. Results and Discussion:

In addressing the first research question regarding the correlation between the tax class of properties and the number of buildings on the same tax lot, the analysis revealed a weak negative correlation of approximately -0.033. This indicates that as the tax class decreases, there is a slight tendency for the number of buildings on the tax lot to increase, and vice versa. However, it's crucial to note that the correlation is not strong, suggesting that other factors significantly contribute to the variation in the number of buildings. This subtle association between tax class and building density aligns with existing literature, particularly in Research Paper 1, which emphasizes the economic implications of tax classification for real estate investors.

For the second research question, exploring the predominant tax class for properties in each borough, the analysis identified Class 2.0 as the predominant tax class across multiple boroughs, including the Bronx, Brooklyn, and Manhattan. However, Staten Island deviates with a predominant tax class of 4.0. This distinction echoes findings in Research Paper 1, which highlights specific tax benefits for properties falling within certain classes. The unique prevalence of Tax Class 4.0 in Staten Island suggests potential variations in property tax policies or characteristics specific to this borough, warranting further investigation.

Addressing the third research question regarding the gross square footage of properties across the five boroughs, the linear regression model demonstrated significant associations with both boroughs and building classes. The analysis indicates that Borough and BuildingClass are crucial factors in predicting square footage. However, the model's relatively low adjusted R-squared suggests the complexity of predicting square footage based solely on boroughs and building classes. This complexity aligns with the emphasis on real estate tax literacy in Research Paper 2, which underscores the intricate nature of real estate transactions in New York City.

Regarding the fourth research question on relationships between street names, zip codes, and the number of buildings on a tax lot, hierarchical clustering identified three distinct clusters. The green cluster on the left represents properties with significantly different characteristics, while the red and blue clusters on the right suggest similarities in terms of the number of buildings and zip codes. This clustering analysis aligns with the discussion in Research Paper 3, which addresses the complexity and differences in the New York City property tax system, emphasizing challenges related to inequality and disparity.

8. Limitations:

Despite the valuable insights gained, this study has some limitations. The dataset used may have constraints, such as missing or incomplete information, which could impact the accuracy and comprehensiveness of the analyses. Additionally, the analytical approach, while informative, may not capture all nuances of the complex real estate dynamics in New York City. These limitations should be considered when interpreting the results.

9. Conclusion:

In conclusion, this data analytics research project provides valuable insights into property dynamics in New York City, addressing key research questions and uncovering patterns,

correlations, and trends. The findings contribute to existing literature on real estate taxation and property dynamics, highlighting the nuanced relationships between tax classes, building attributes, square footage, and geographical factors. Despite limitations, the study emphasizes the importance of leveraging data analytics techniques to inform our understanding of multifaceted urban real estate phenomena. The implications extend to policymakers, real estate investors, and researchers seeking a deeper understanding of the intricacies of property dynamics in a dynamic urban environment like New York City.

10. APPENDIX (List of Technical terms):

BBL (Borough, Block, Lot):

Definition: A unique identifier for a tax lot in New York City, composed of three components - Borough, Block, and Lot.

Explanation: It's a standardized code used to uniquely identify and locate properties in the city. The Borough is represented by a numerical code (1-5), the Block is a unique number within the Borough, and the Lot is a unique number within the Block.

Tax Class:

Definition: A classification assigned to a property that determines the amount of property tax.

Explanation: Different types of properties have different tax obligations. Tax Class helps categorize properties based on their use, structure, or other criteria, impacting the property tax calculation.

Gross Square Footage:

Definition: The total floor area of a building, including all interior and exterior spaces.

Explanation: It quantifies the size of a property, providing insights into its overall capacity and scale. It's a key metric in real estate to understand the spatial extent of a building.

Hierarchical Clustering:

Definition: A method of cluster analysis that builds a hierarchy of clusters.

Explanation: Properties are grouped based on similarities, forming a tree-like structure (dendrogram). It helps identify patterns and relationships in the dataset by revealing clusters of properties with similar characteristics.

ANOVA (Analysis of Variance):

Definition: A statistical method used to analyze the differences among group means in a sample.

Explanation: In the context of this analysis, ANOVA is applied to test the significance of the regression model. It helps determine if there are significant differences in the mean square footage among different boroughs and building classes.

Dendrogram:

Definition: A tree diagram that shows the arrangement of clusters produced by hierarchical clustering.

Explanation: It visually represents the hierarchy of clusters, helping to interpret the relationships and similarities between different groups of properties.

Adjusted R-squared:

Definition: A modified version of R-squared that adjusts for the number of predictors in a model.

Explanation: It provides a measure of the proportion of the variability in the response variable (square footage) that is explained by the predictors (boroughs and building classes) while considering the number of predictors in the model.

Cluster Analysis:

Definition: A statistical technique used to identify groups (clusters) of similar items within a dataset.

Explanation: In the project, it's applied to identify patterns and relationships between street names, zip codes, and the number of buildings. The clusters represent groups of properties sharing similar characteristics.

11. References:

[1]: Dataset Reference: NYC Covered Building's List (2020). (2022, May 9) www.data.gov ; [online] Available:

<https://catalog.data.gov/dataset/nyc-covered-buildings-list-2020>

[2]: w3schools. (2019). SQL Tutorial. W3schools.com. [Online]: Available <https://www.w3schools.com/sql/>

[3]: Python String Methods. (n.d.). Wwww.w3schools.com. [online]: Available https://www.w3schools.com/python/python_ref_string.asp

[4]: Data Types | Noir Documentation. (n.d.). Noir-Lang.org. Retrieved December 5, 2023, [Online]: Available https://noir-lang.org/dev/language_concepts/data_types/

[5]: rstudio/cheatsheets. (2023, October 30). GitHub. [Online]: Available <https://github.com/rstudio/cheatsheets>

[6]: Richard Velotta, (2018, July 19). Why You Should Pay Attention to Properties In Tax Classes 2A and 2B. Commercial Observer. [Online]: Available <https://commercialobserver.com/2018/07/why-you-should-pay-attention-to-properties-in-tax-classes-2a-and-2b>

[7]: Tishco, S. (n.d.). Taxation of New York City Real Property. American Property Tax Counsel (APTC). Retrieved November 6, 2023, [Online]: Available: <https://www.aptcnet.com/property-tax-resources/published-property-tax-reports/taxation-of-new-york-city-real-property>

[8]: Geringer-Sameth, E. (n.d.). An Old, Unfair System: New York City's Property Tax Conundrum - Part II - Deep and Complex Inequities. Gotham Gazette. Retrieved November 6, 2023, [Online]: Available <https://www.gothamgazette.com/state/8713-old-unfair-system-new-york-city-property-tax-conundrum-part-ii-classes>

[9]: "Blackboard Learn," mymasonportal.gmu.edu. [Online] Available: https://mymasonportal.gmu.edu/ultra/courses/_499970_1/cl/outline