Water Use in NYCHA Buildings

Anil Bhusal

Data Science Scholars,

The City College of New York,

New York

August 11, 2022

Table of Contents

ABSTRACT

Water Use in NYCHA Buildings. ANIL BHUSAL (Data Science Scholars, The City College of NewYork, New York, 10031) NARESH DEVINENI (Department of Civil Engineering,The City College of NewYork, New York, 10031)

In the light of looming future extreme droughts, it is necessary to keep track of water consumption for mega cities like New York (NYC), one of the most densely populated cities in the world. NYC uses nearly one billion gallons of water per day. Thus, it requires a careful analysis to supply the right amount of clean/pure water. We use NYCHA buildings (it is home to 1 in 16 New Yorkers in 2022) data for our analysis. They are easily accessible and public buildings which need more attention for conservation. We analyze water consumption and patterns of consuming water over nine years in each borough and buildings. We also work to find the high consumption areas and discover the relationship between NYCHA and its importance in controlling extreme consumption. Some of the significant methodologies are Data Preparation, Yearly analysis (it will give us overall scenario for each borough from 2013 to 2019), Seasonality analysis (it will inform us about the seasonal behavior of each boroughs), Extreme Percentile (it will give us idea about the extreme consumed areas and buildings that requires more attention for the conservation of the water), and Trend Analysis (which will give detail information about change in consumption). We found out that the most extreme consumption in Manhattan and Brooklyn is during winter months. Overall, water consumption is also increasing. As for Brooklyn, the zip code 11225, consumption is increasing rapidly, which requires more attention. Similarly, Zip code 10030 is consuming excess water in Manhattan. Also, it is important that people should be awarded for conservation of extreme consumption.

INTRODUCTION

Even though water is a basic need to survive, 785 million people lack drinking-water services globally(The Water Crisis), which is more than double the population of the United States. Also, a child dies of waterborne diseases every 15 seconds(Chambers). So, it is necessary to have a clear analysis of water use to ensure its safe and secure access. We choose New York City, which is densely populated and uses nearly one billion gallons of water each day(Chambers) as a case study. In New York City, we will use NYCHA buildings which are home to 1 in 16 New Yorkers in 2022(NYCHA Fact Sheet 2022), for our analysis. In NYCHA buildings, we will mainly analyze the pattern of its consumption in each borough over the nine years, find the high consumption buildings or areas for conservation of the extreme consumption, and discover the relationship between these buildings with New York City and globally.

MATERIALS AND METHODS

We use python and excel for our research. Using these tools, we use different data science techniques.

Data Preparation:

We first prepare data for the analysis. In the data preparation process, we clean duplicate data initially to minimize the error in our calculation and handle the error during the execution of our program. After that, we split the data in terms of boroughs for a detailed analysis of each borough's consumption. In each borough, we find out the unique meter numbers of each building. After that, we set outliers to clean our data from the non-uniform dataset that might have occurred due to unpredicted reasons or data entry errors; doing this will give us a better data set for our analysis. We calculate the outliers using the formula:

$$\frac{X - Mean}{Standard\ deviation}$$

X represents the value of the consumption column and mean is the average value of the consumption column. Finally, we will get the outliers and we set normal values to be in the range of 10 deviation values from the above equation.

we are working with the periodic data set, our data must be continuous and unique in terms of the date for each unique meter number. So, we first fix the duplicate date entry dataset by averaging their value. After that, we also remove the insufficient data set from our original data to minimize the error in our analysis because some buildings have an entry dataset for only one, two, or three years. For this sort of data, it is hard to predict their nature. During this insufficient data cleaning process, we set that the building should have at least 40 months of their data to be involved for the analysis over nine years. Also, we check if each building has all the data for nine years. If a building doesn't have data for nine years, we will randomly select the values for the missing month from its data set. For example, if the missing month of a building is February of any year, we choose the data for that date from the available January, February, and March datasets). Finally, we will have prepared a data set for our further analysis process.

Yearly Analysis:

After the data preparation process, we use these data sets for our different methods of analysis. The first method of analysis we performed is Yearly analysis, where we analyze the pattern of consumption between 2013 to 2019. We sum each month of each year among each building and calculate the overall yearly report.

Seasonality Analysis:

The second method of analysis is Seasonality analysis, where we will look over the average monthly behavior of the database. It will lead us to see how each borough is acting in each season. Their nature might differ in Fall, Spring, summer, and Winter. Hence, seasonality analysis will give us analysis in terms of season. We group each month and average their value to see the seasonality analysis.

Extreme Percentile:

The third method of analysis is the Extreme Percentile method, where we set our percentile to be 95% and find out the 95th percentile of each building. We use the below formula to calculate 95 percentile.

$$N = \frac{95}{100} * \text{ total number values in the column}$$

From the above equation we will get the position of 95th extreme value from the sorted column. And any value greater than the value will be the extreme values.

After finding the percentiles of each building, we plot them in a graph to see if they are in the same month or the same year. It will inform us maximum consumption is occurring in that specific month because most of the extreme consumption is happening in that period. It will also lead us to discuss why it is during that period. Also, we will discuss different factors of extreme consumption months like the temperature in that period and area of consumption. Also, we will visualize the occurrence of extreme building in each zip code. We calculate this using formula as,

$$\text{Probability: } \frac{Total\ number\ of\ occurrence\ in\ extreme\ months}{12}$$

Finally, we will see how many times each building are occurring over each zip code.

Trend Analysis:

Finally, we will also analyze the pattern of each building whether their consumption is increasing or decreasing, or remaining constant. It will help us predict which places require more attention or which area is more responsible for high consumption.

RESULTS

In our original data set, we have a total number of 46,747 rows. After cleaning or removing the duplicate row, we will work with 36,397 data sets of rows for our further analysis. After that, we separated the data sets as the borough of New York City to analyze each data set separately and closely. Even though NYC has five boroughs, our data set has two extra pieces of information about a borough. They are FHA and Non-Development Facility with 17083 and 29 rows, respectively. Similarly, Manhattan, Queens, Brooklyn, Bronx, and Staten Island have 5962, 3054, 6598, 3586, and 85 rows, respectively. Now, let us view each of the boroughs individually.

a) Manhattan

In Manhattan, we have 59 buildings after the data preparation method. We will be working with this set of data for our further analysis. Let's start with a Yearly analysis.

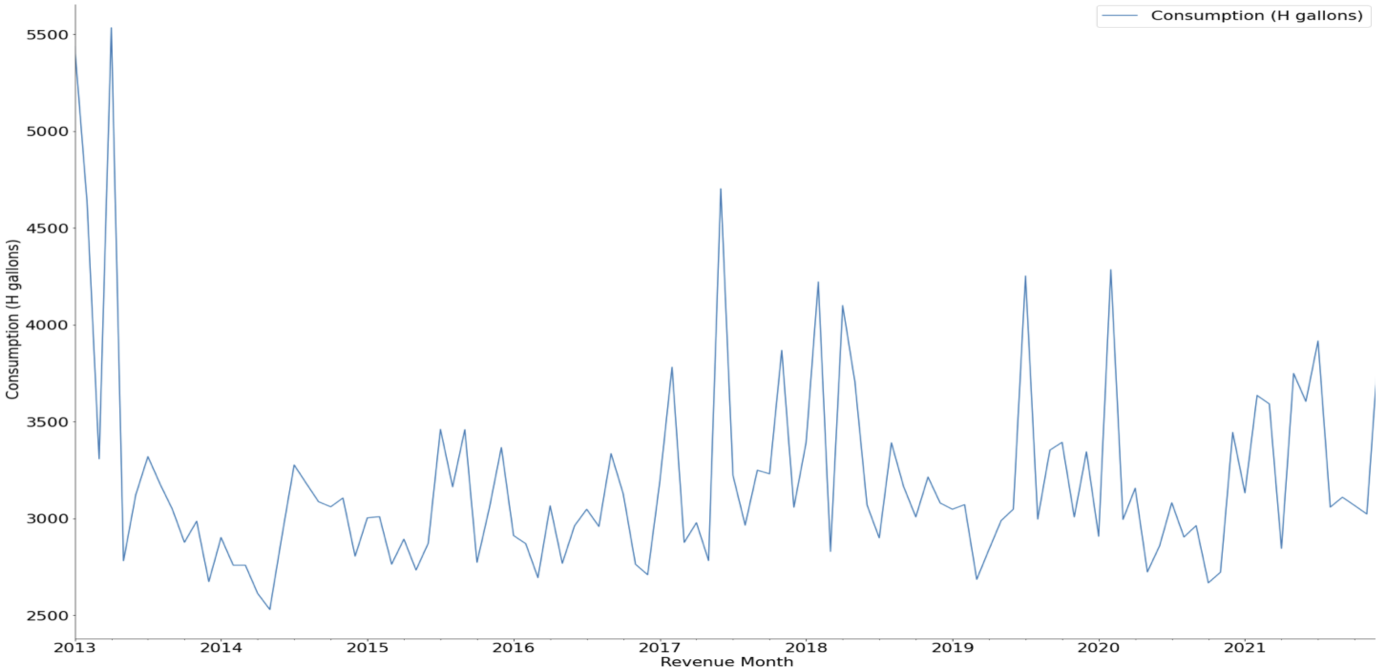

Figure 1: Yearly analysis of Consumption of water in Hundred gallons over nine years

In figure(1), we calculate the average consumption of manhattan over nine years. If we see a graph closely, during 2013, the consumption of manhattan was pretty high compared to the rest of the years. In the starting summer of 2014, the consumption dropped, or it was the lowest in nine years. Also, we can see that the overall consumption has increased from 2017 and remains high, which means that the consumption ratio from 2017 is increased. Now, let's analyze the average monthly behavior of the manhattan database, which is a seasonality analysis.

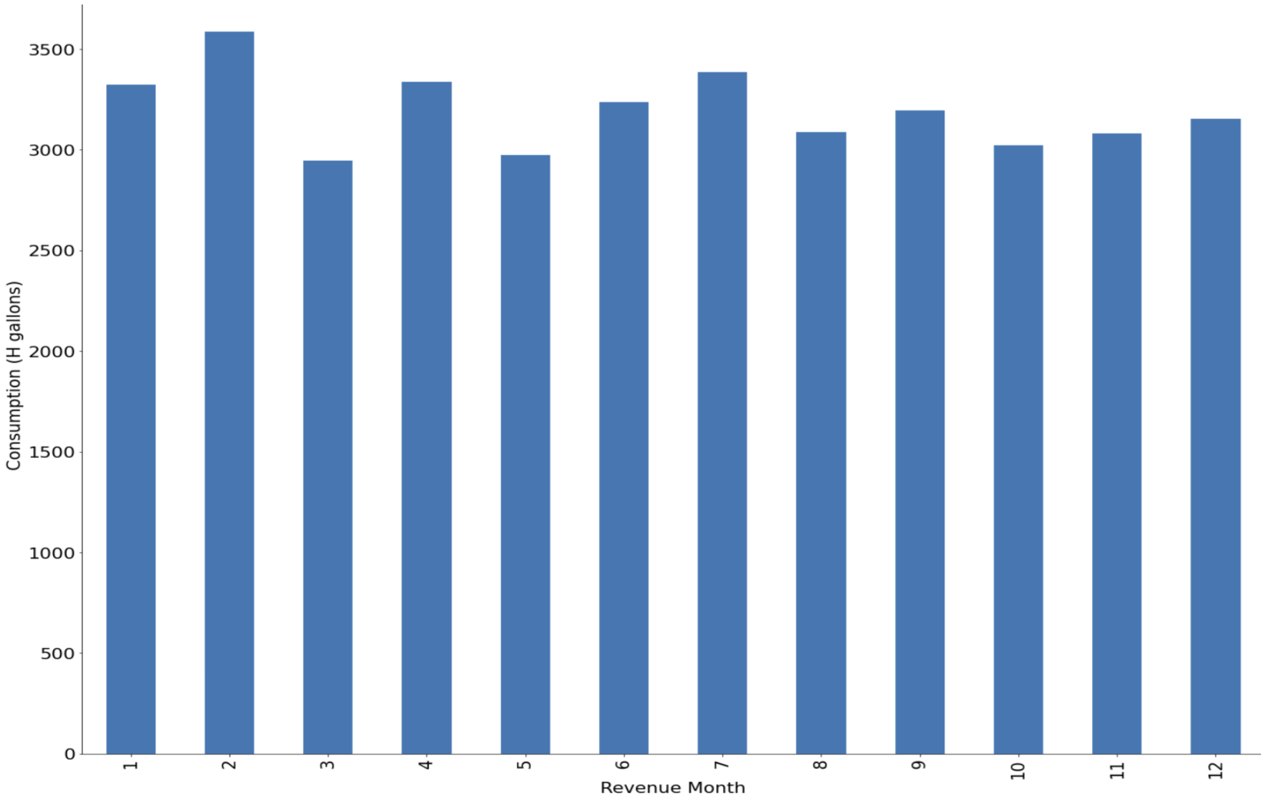

Figure 2: Seasonality analysis of Consumption of water in Hundred gallons over nine years

In Figure (2), we can see that each month has about similar consumption. There is no high difference in the consumption of the water. Looking deep in the value, February has the highest consumption, and March has the sudden drop in consumption. On average, each month is about 3300 Hundreds gallons consumed. We had hypothesized that the winter season might have high consumptions , but each season has a nearly equal average ratio. They are balanced with different factors.

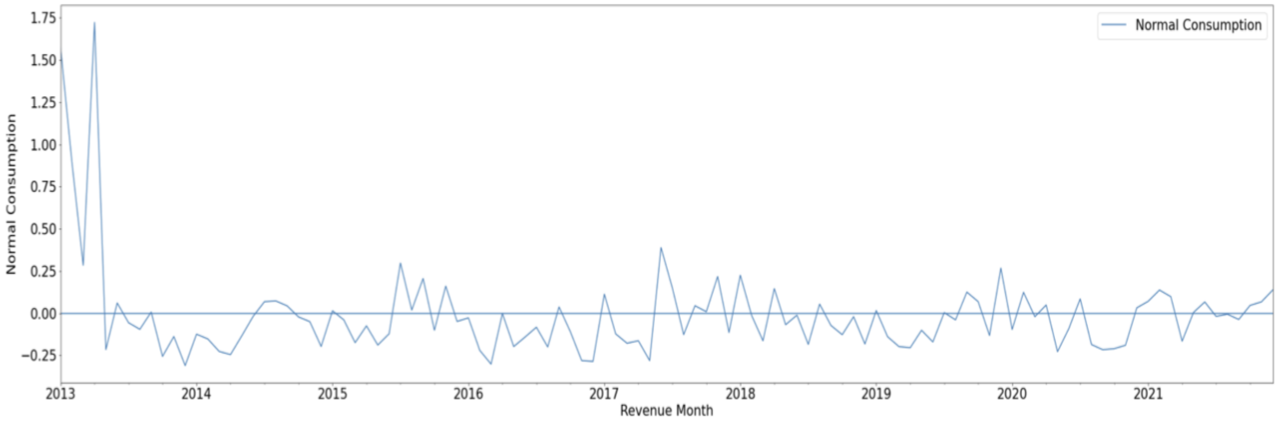

Figure 3: Normal Consumption of Manhattan Building over Nine years (Yearly normal)

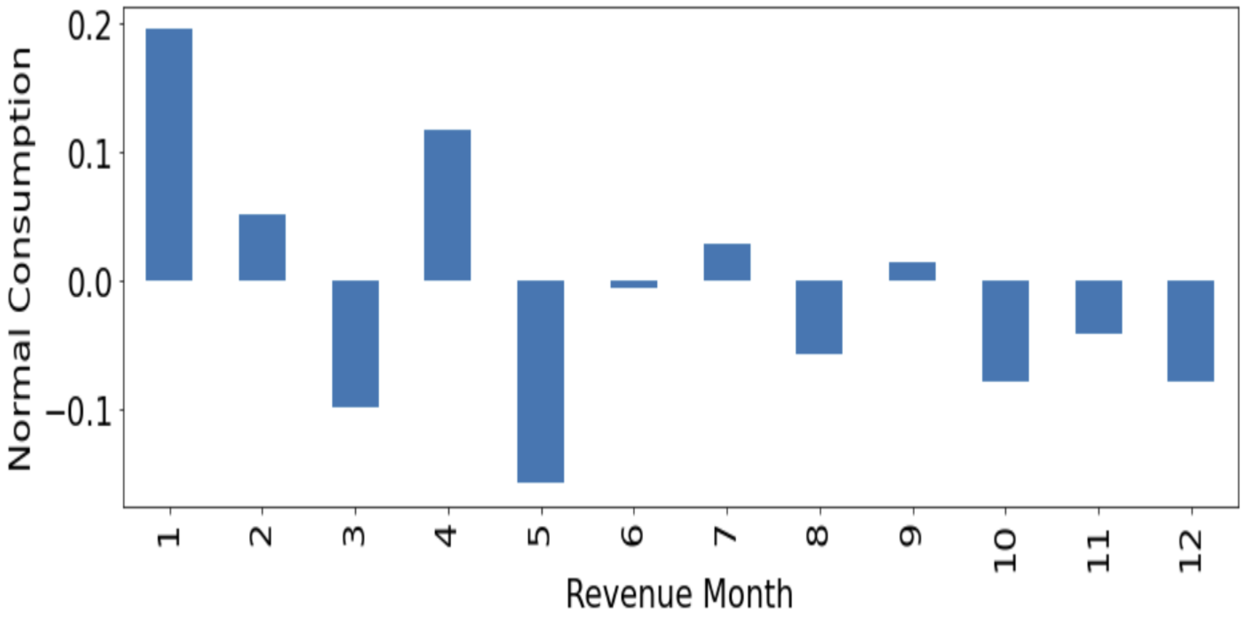

Figure 4: Normal consumption of Seasonality analysis.

In figure (3) and figure (4), we calculated the deviation consumption of manhattan over nine years and the deviation consumption each month of manhattan. These two figures will give us an idea of how each of the values deviated from the normal. we can see that each of the values are not deviated very high or very low. For the further analysis of our data we will have the least error in our analysis.

We calculate the 95th percentile of consumption of each building. After that we plot them graphically shown below to represent orientation of each month. Here, the y axis represents the total number of buildings in a specific month. X axis represents the months over nine years which is shown below in figure 5.

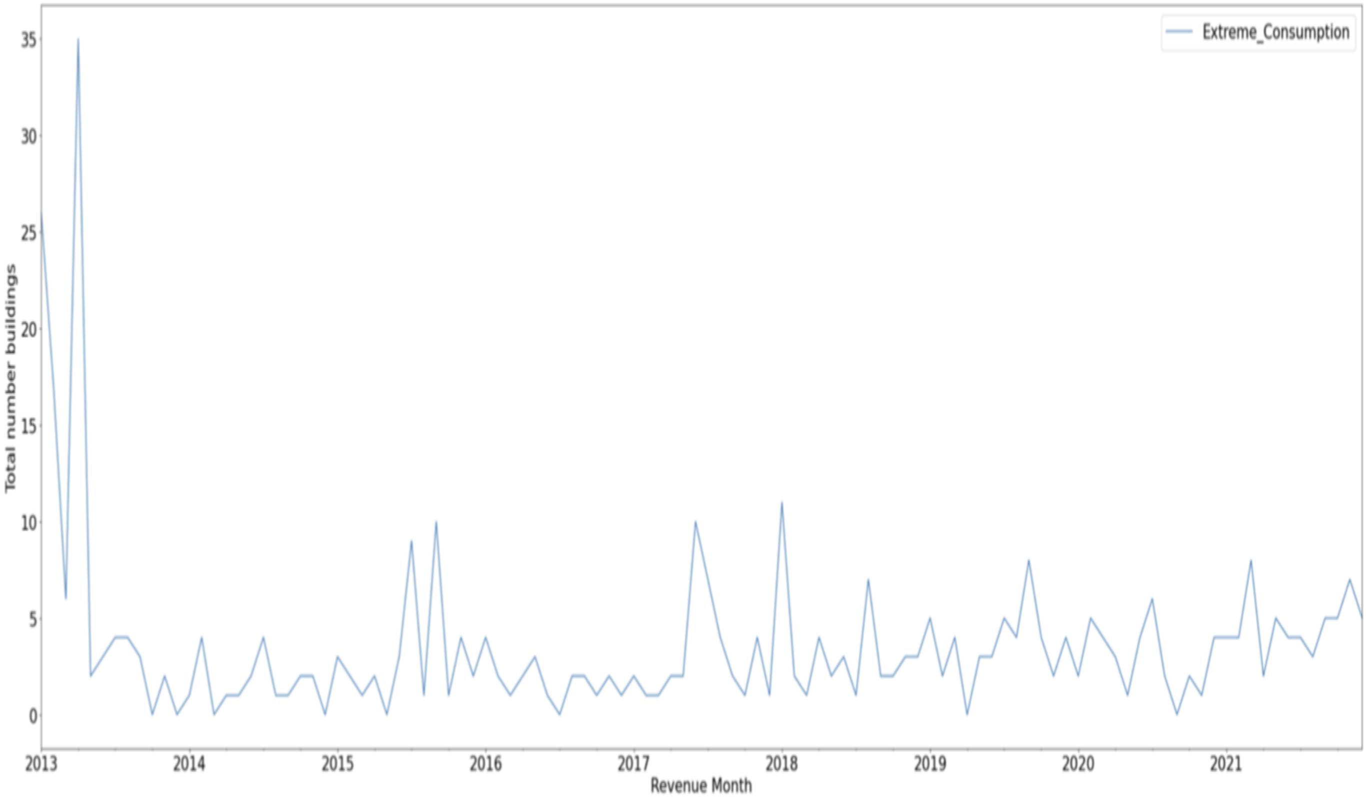

Figure 5: 95th Extreme Percentile of Manhattan

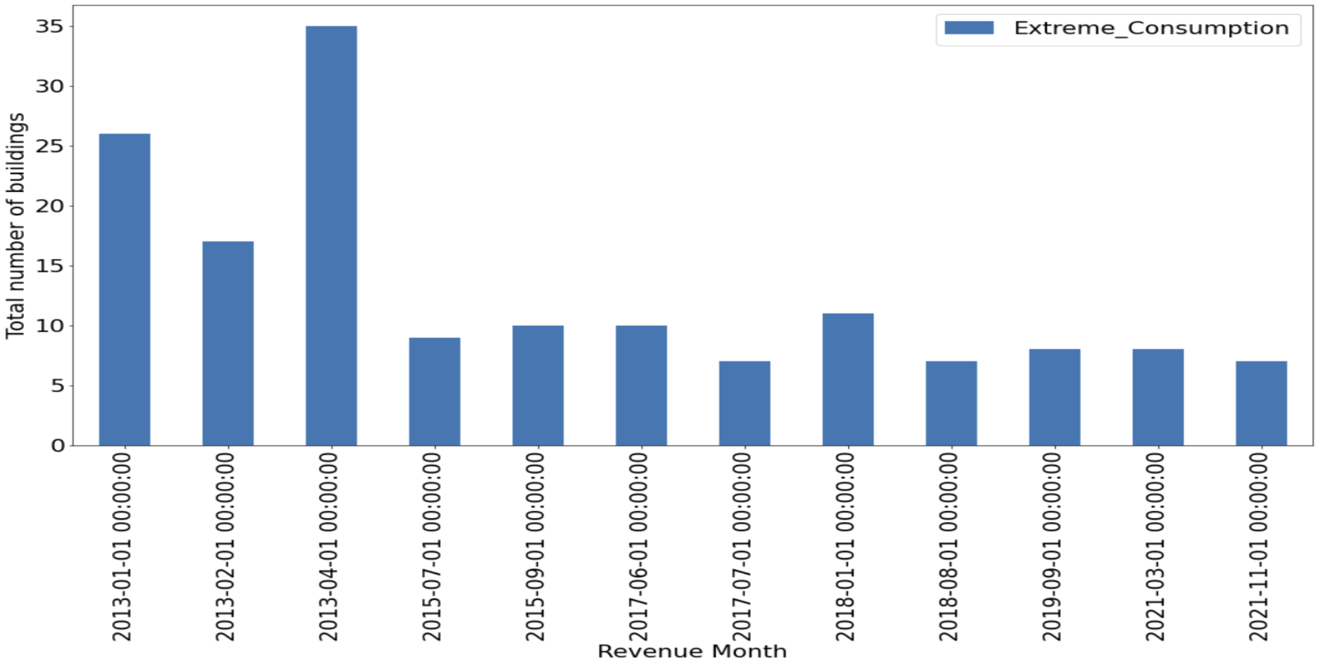

Figure 6: Top twelve extreme consumption months of Manhattan.

In figure 6, we calculate the top twelve extreme months. The y-axis in the figure represents the total number of buildings, and the x-axis represents the top twelve extreme months. April 2013 has the highest number of extreme buildings, which means this month has the most extreme consumption over nine years. Similarly January and February of 2013 have second and third highest extreme consumption.

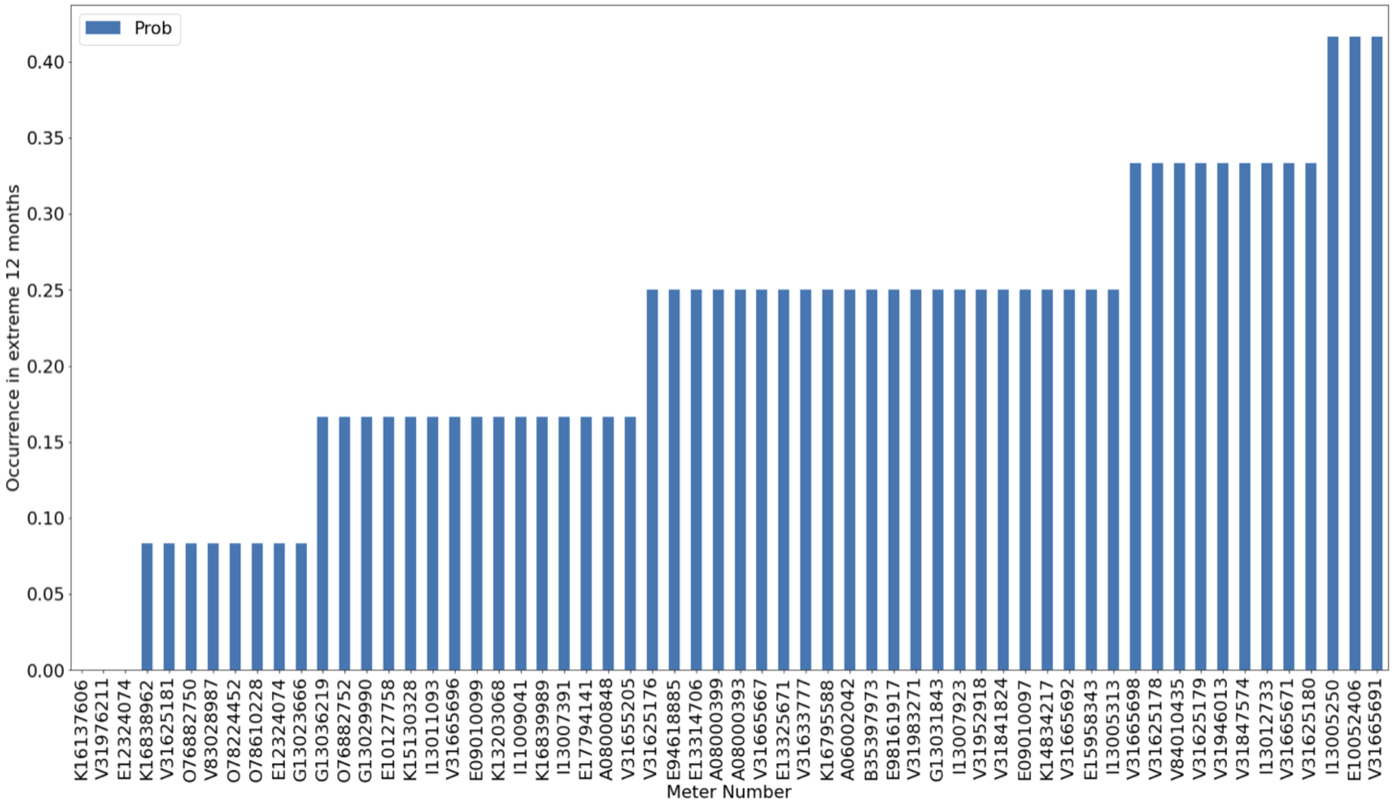

Figure 7: Extreme Occurrence Analysis of each building

In figure 7, we calculate the extreme occurrence of each building. The y-axis represents the probability of occurring in extreme twelve months, and the x-axis represents all the buildings of Manhattan. Most of the buildings have a probability of 0.25 in extreme twelve months, which means that most of the buildings are occurring three times in the top twelve months. Also, none of the buildings are occurring in the whole twelve months. The maximum number of times each building occurs is five times, whereas three buildings are not part of extreme months.

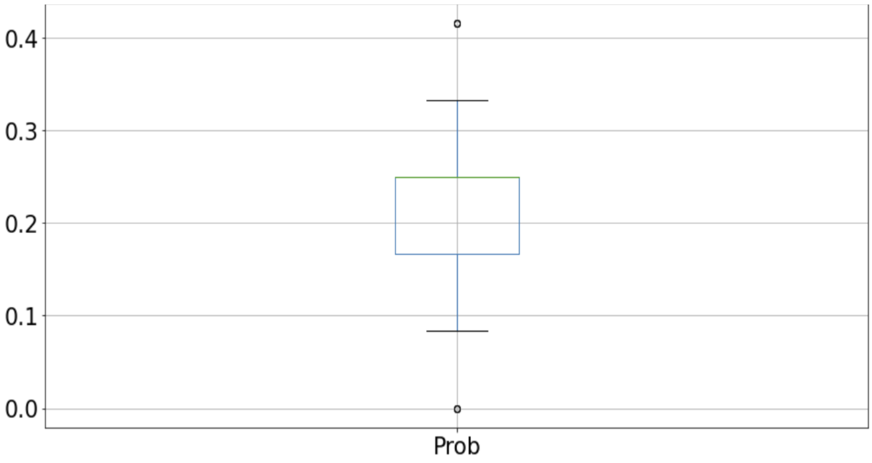

Figure 8: Box plot of occurrence of extreme twelve months.

In figure 8, we calculate the box plot to visualize how each building is occurring. We can see that most of the occurrences are between 1.67 and 0.25. The majority of the building is occurring at this range. Also, we have a minimum probability of a building as 0 and maximum in Manhattan as 0.42. This figure is a similar illustration as figure 7 with different visualization.

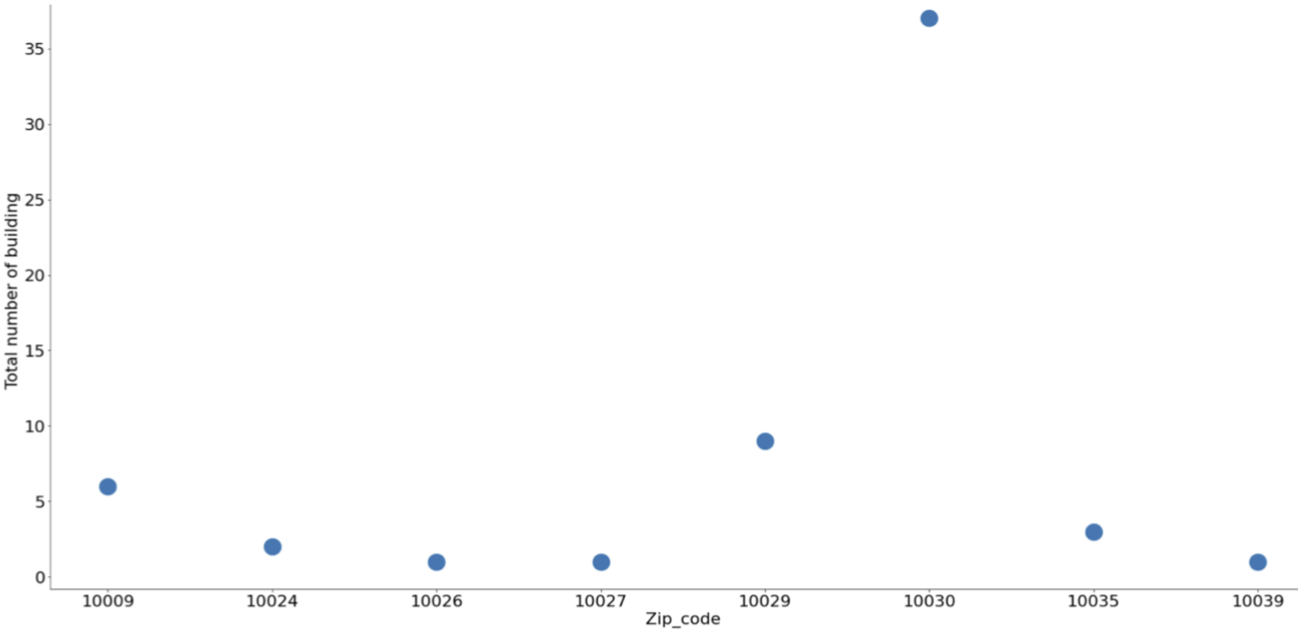

Figure 9: Total number of buildings in each zip code.

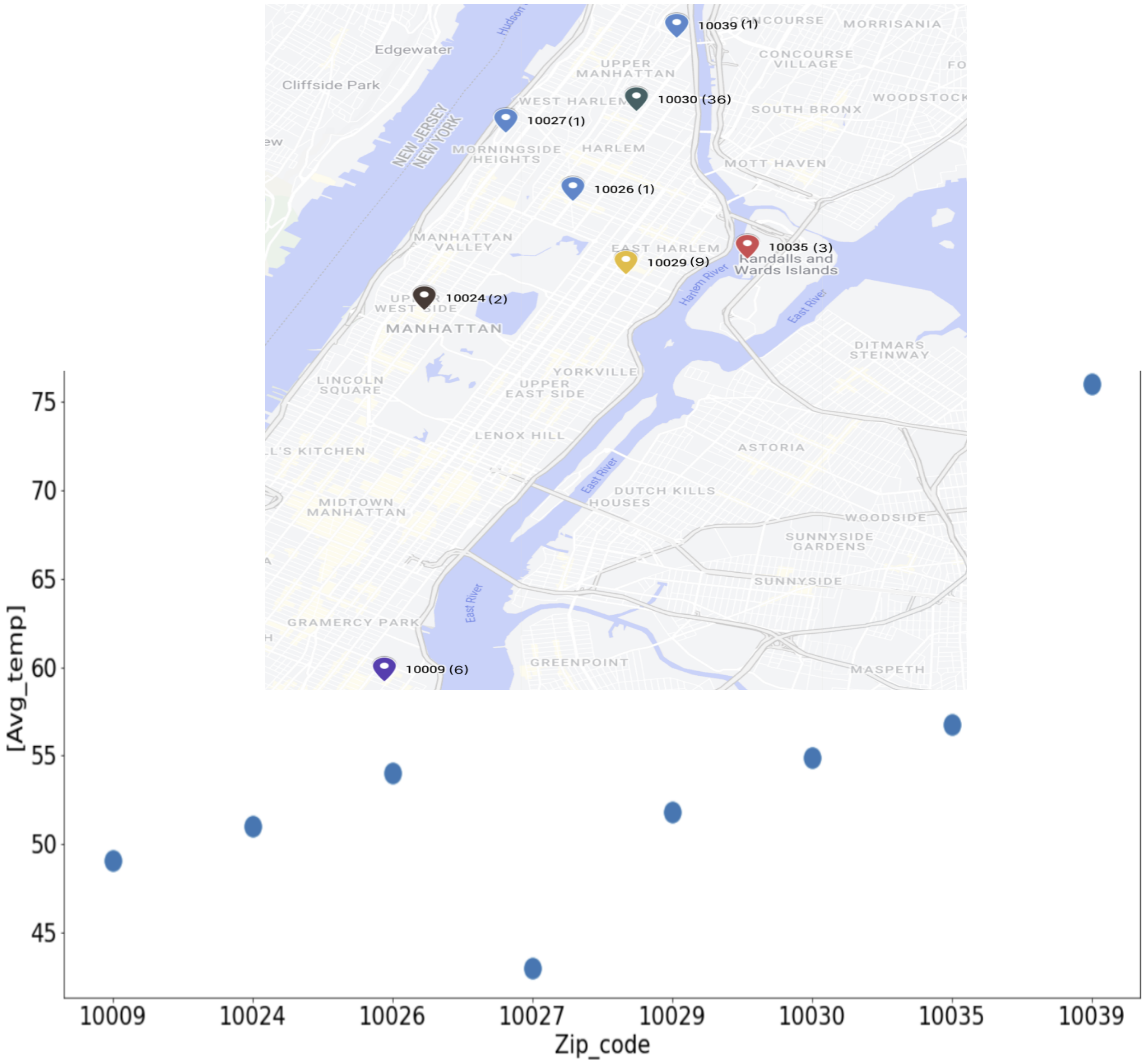

Figure 10: Spatial plot of Manhattan with Average temperature

In figure (9) and figure(10), we get a spatial plot for the Manhattan zip code, Temperature, and the total number of buildings. Figure(9) shows how many buildings are in each zip code. We can see that zip code 1030 is a densely populated zip code, and there is high consumption. This zip code requires more attention for the conservation of the consumption of water. Figure(10) represents the plot of each zip code on the map to visualize where they lie. It also represents the average temperature of each zip code. Most of the zip code lies within the range of 50 to 55. Thus, most of the consumption is during cold weather.

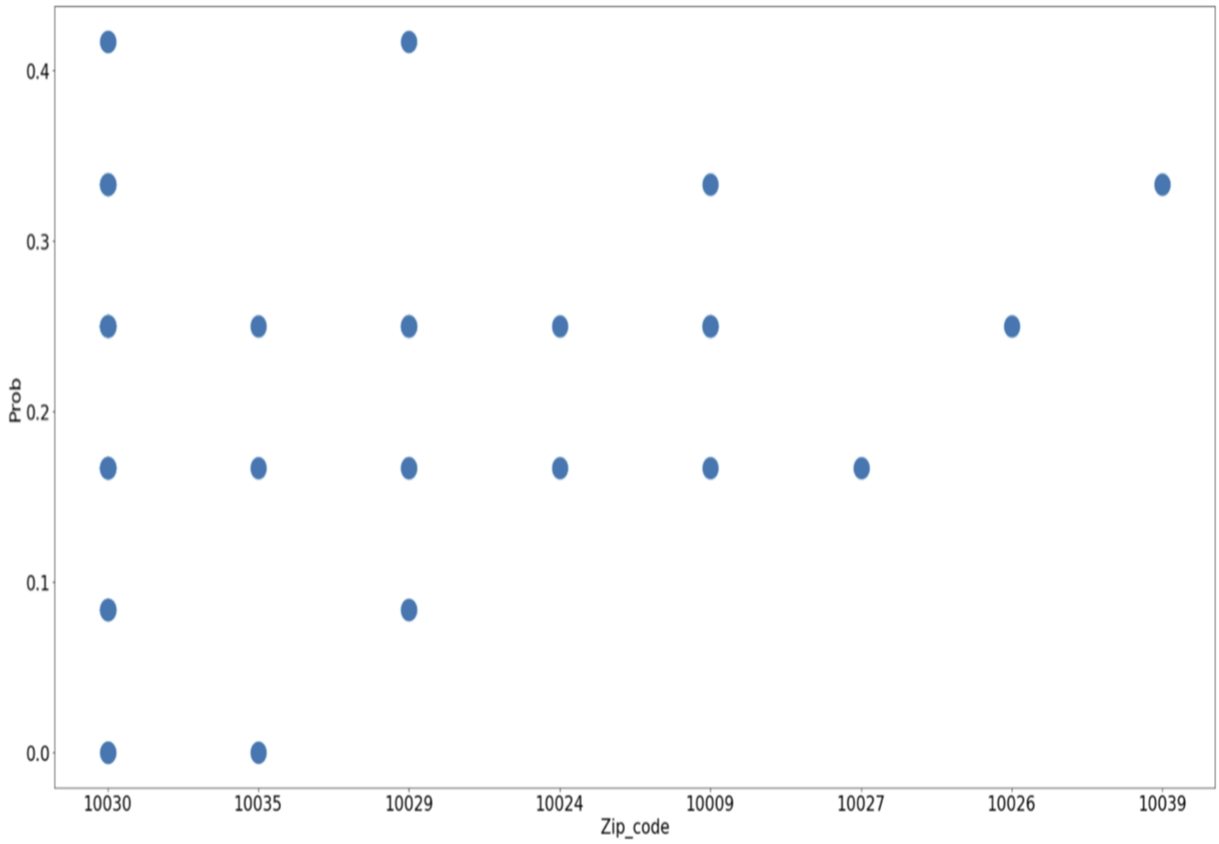

Figure 11: visualization of occurrence of extreme building in each zip code

In figure 11, the y axis represents the probability of occurrence in extreme 12 months and x axis represents its corresponding zip code. This graph mainly represents how each of the Zipcodes have variable distribution in their occurrence. For example, Zip Code 10039 has only occurrence once that means it has only single nature occurrence in its zip code.

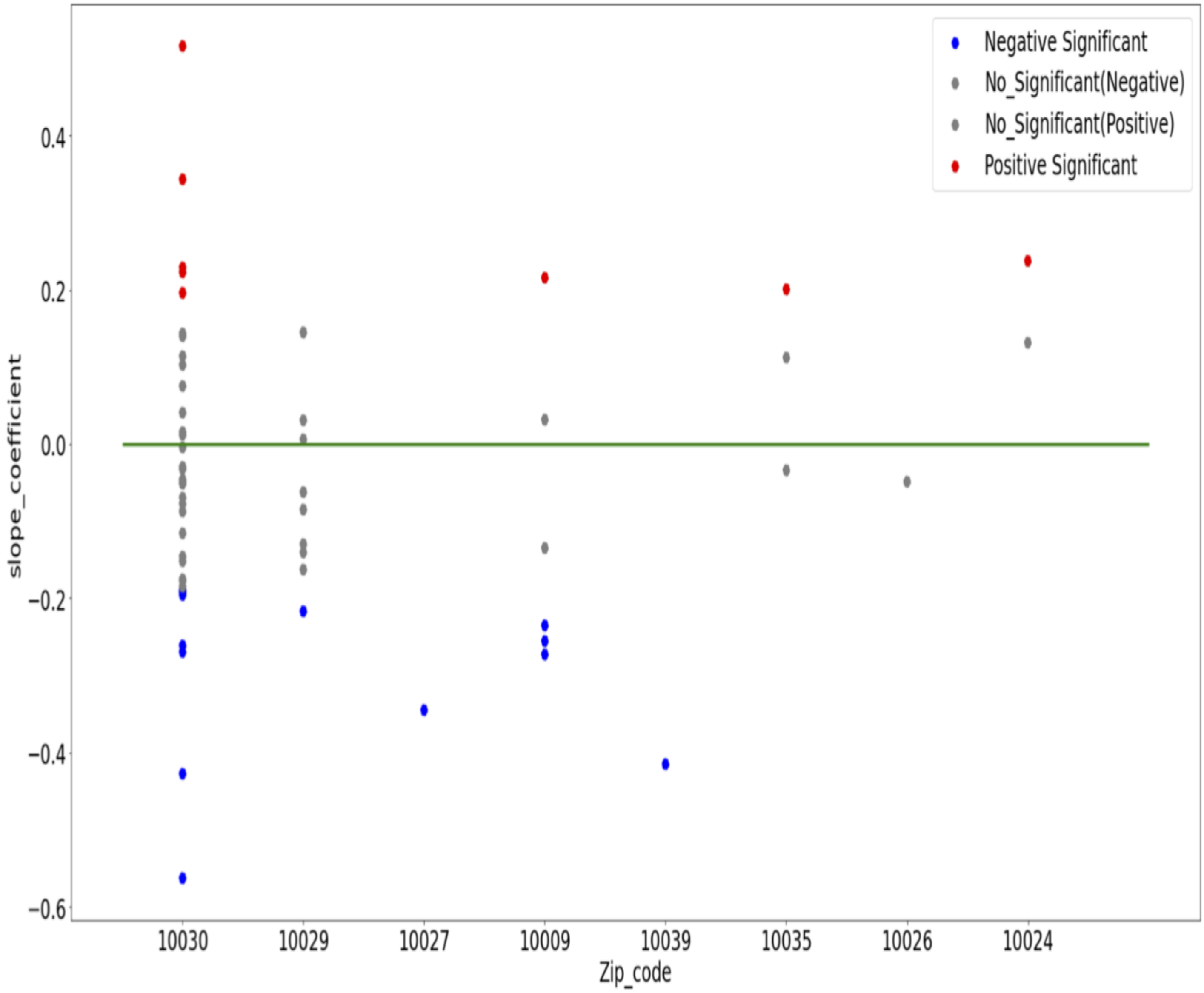

Figure 12: Trend Analysis of Manhattan building

Figure (12) represents the trend analysis of each building in Manhattan with its zip code. The y-axis in the figure represents the slope coefficient, and the x-axis represents the zip code. Each dot with a different color represents its significance in each zip code. Here, Red and blue colors represent the Positive and Negative Significance of each building, whereas gray color represents the no significant buildings in the zip code. Figure (13) represents the total count of each significance in each zip code.

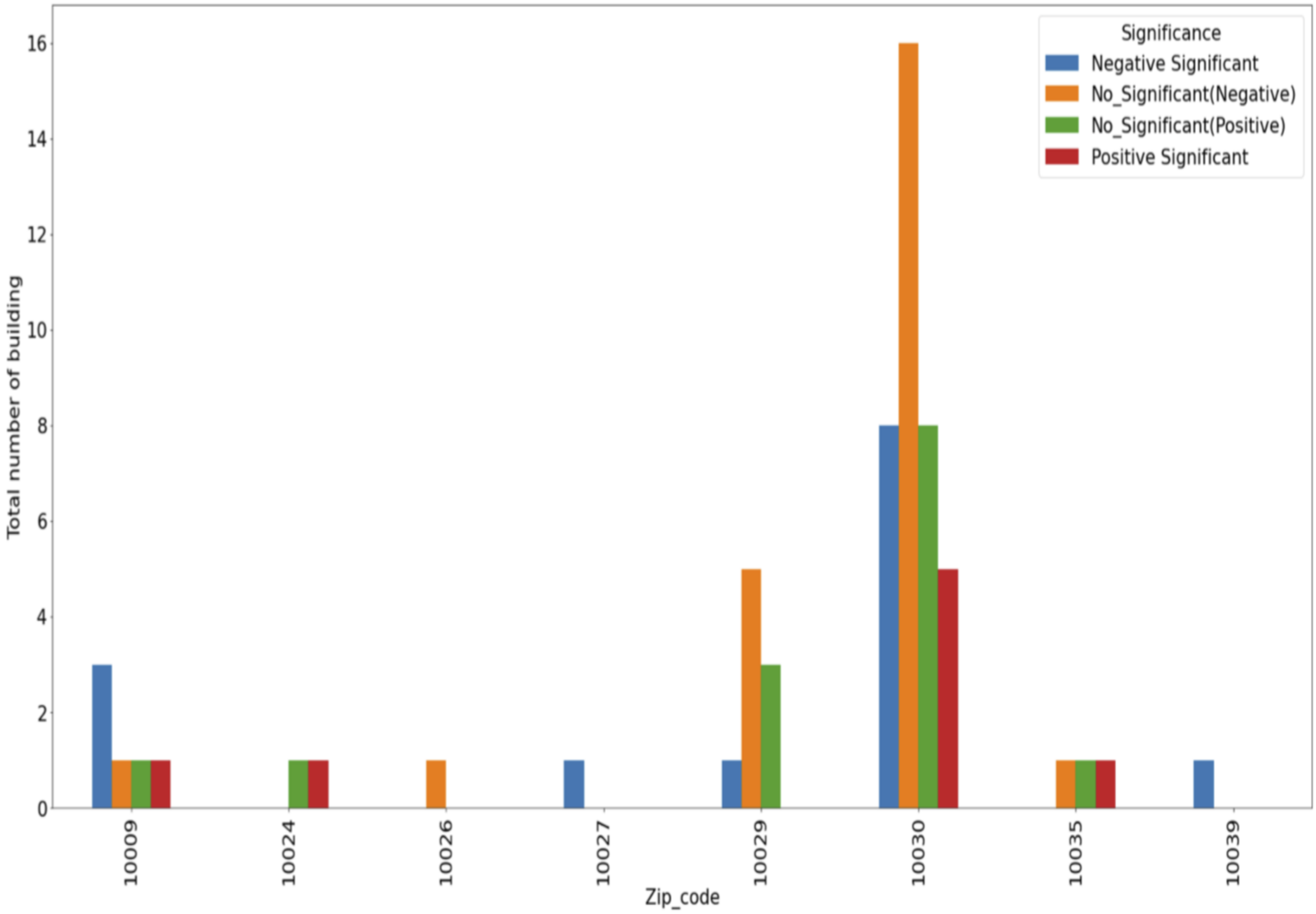

Figure 13: Trend Analysis of Manhattan building and count of each building

Figure (13) represents the trend analysis of each building in Manhattan with its zip code and the total number of buildings. The y-axis in the figure represents the total number building, and the x-axis represents the zip code. Here, Red and blue colors represent the Positive and Negative Significant of each building. Also, green and orange color shows the positive no significant, and negative no significant buildings in each zip code. We need to account more on positive significant values in each code because they are increasing rapidly over time. Zip code 10030 requires more attention to minimize the consumption of water. Also, 10009, 10024, and 10035 requires equal attention for conservation of consumption.

b) BROOKLYN

In brooklyn, we have 63 buildings after the data preparation method. We will be working with this set of data for our further analysis. Let's start with a Yearly analysis.

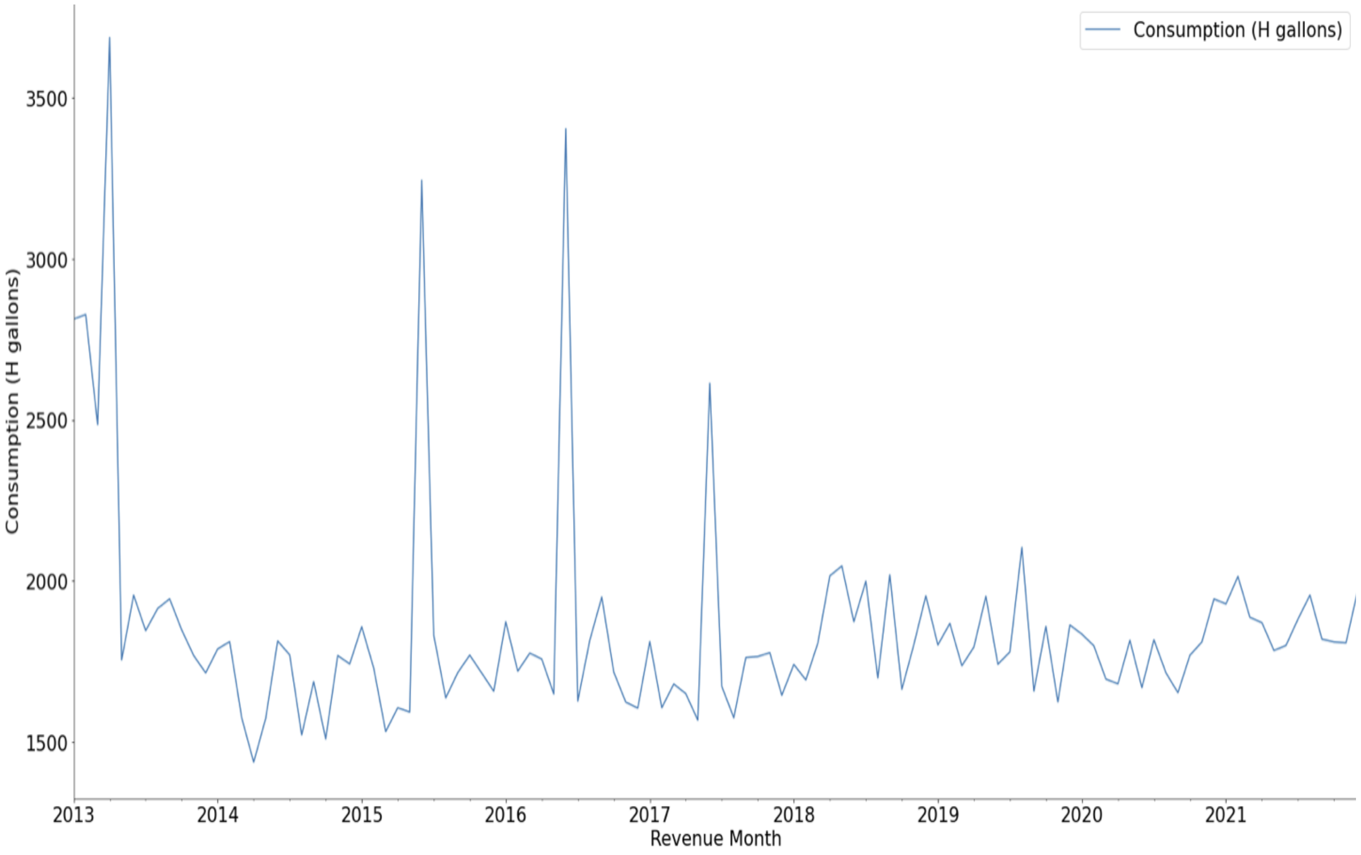

Figure 14: Yearly analysis of Consumption of water in Hundred gallons over nine years

In figure(14), we calculate the average consumption of brooklyn over nine years. In the graph, we can see that in early 2013, and mid of 2015 and 2016, the consumption of brooklyn was pretty high compared to the rest of the years. In the starting summer of 2014, the consumption dropped, or it was the lowest in nine years. Also, we can see that the overall consumption has increased from 2018 and remains high, which means that the consumption ratio from 2018 is increased.

Now, let's analyze the average monthly behavior of the manhattan database, which is a seasonality analysis.

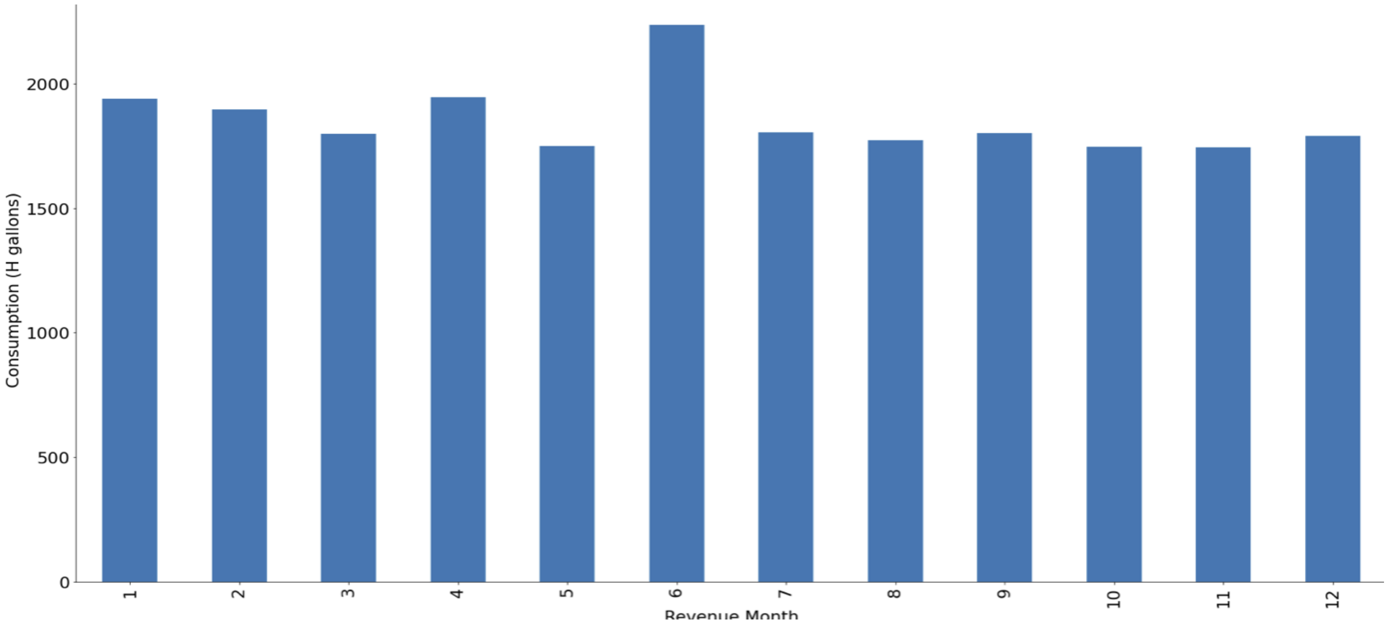

Figure 15: Seasonality analysis of Consumption of water in Hundred gallons over nine years

In Figure (15), we can see that June has the highest consumption over all seasons and the rest of the month has about similar consumption. There is no high difference in the consumption of the water except June. On average, each month is about 2800 Hundreds gallons consumed.

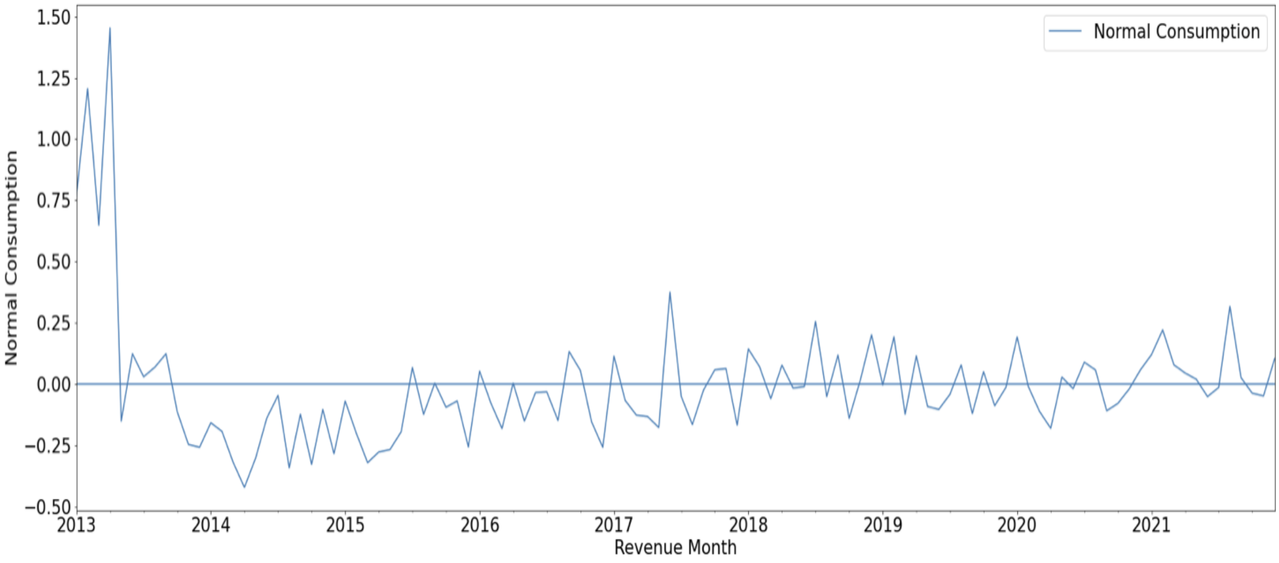

Figure 16: Normal Consumption of Brooklyn Building over Nine years (Yearly normal)

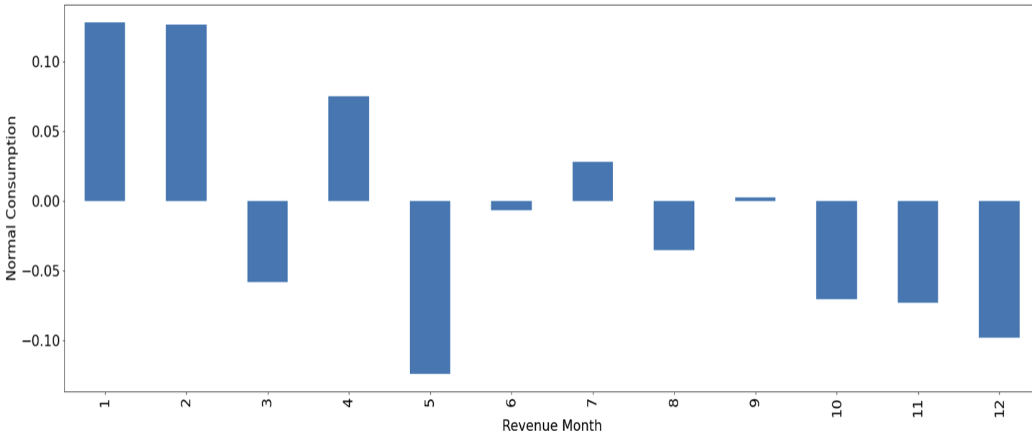

Figure 17: Normal consumption of Seasonality analysis.

In figure (16) and figure (17), we calculated the deviation consumption of brooklyn over nine years and the deviation consumption each month of brooklyn. These two figures will give us an idea of how each of the values deviated from the normal. we can see that each of the values are not deviated very high or very low. For the further analysis of our data we will have the least error in our analysis.

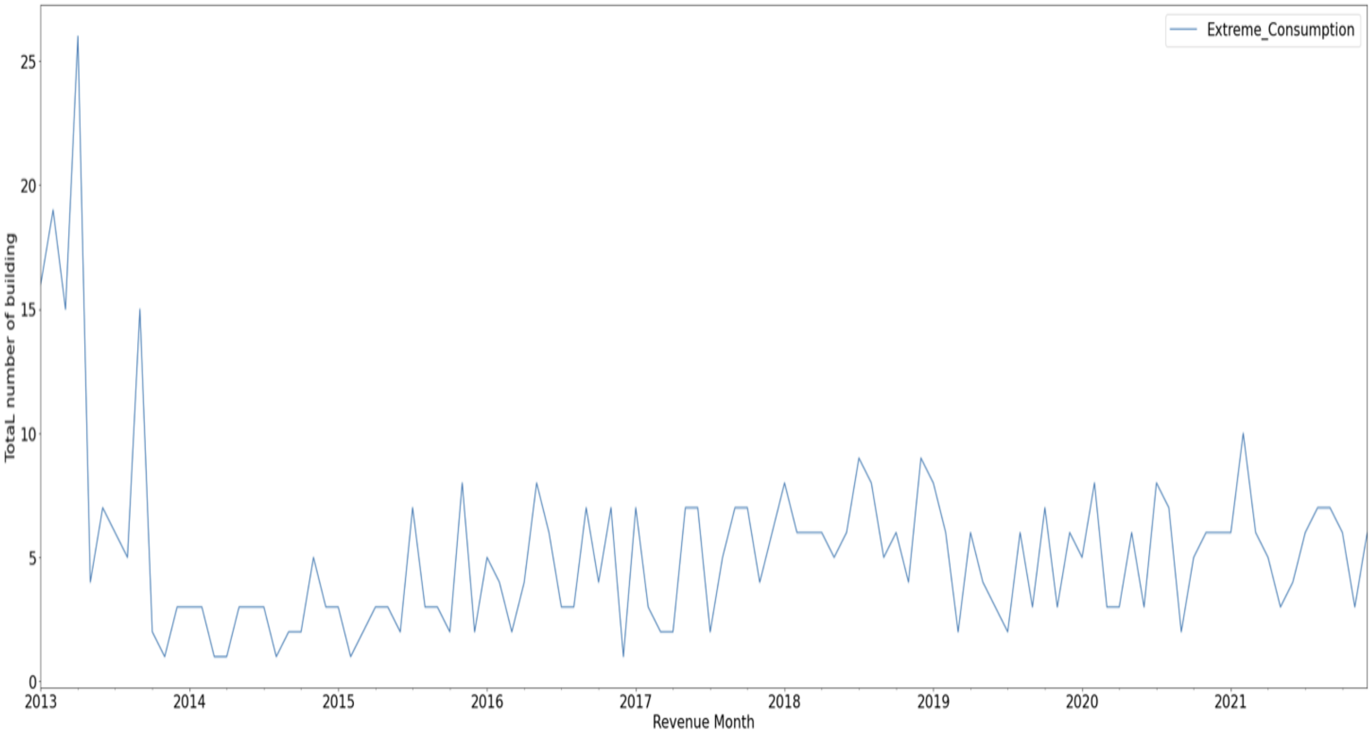

Figure 18: 95th Extreme Percentile of Brooklyn

We calculate the 95th percentile of consumption of each building. After that we plot them graphically shown below to represent orientation of each month. Here, the y axis represents the total number of buildings in a specific month. X axis represents the months over nine years which is shown below in figure 18. Most of the extreme months are in early 2013.

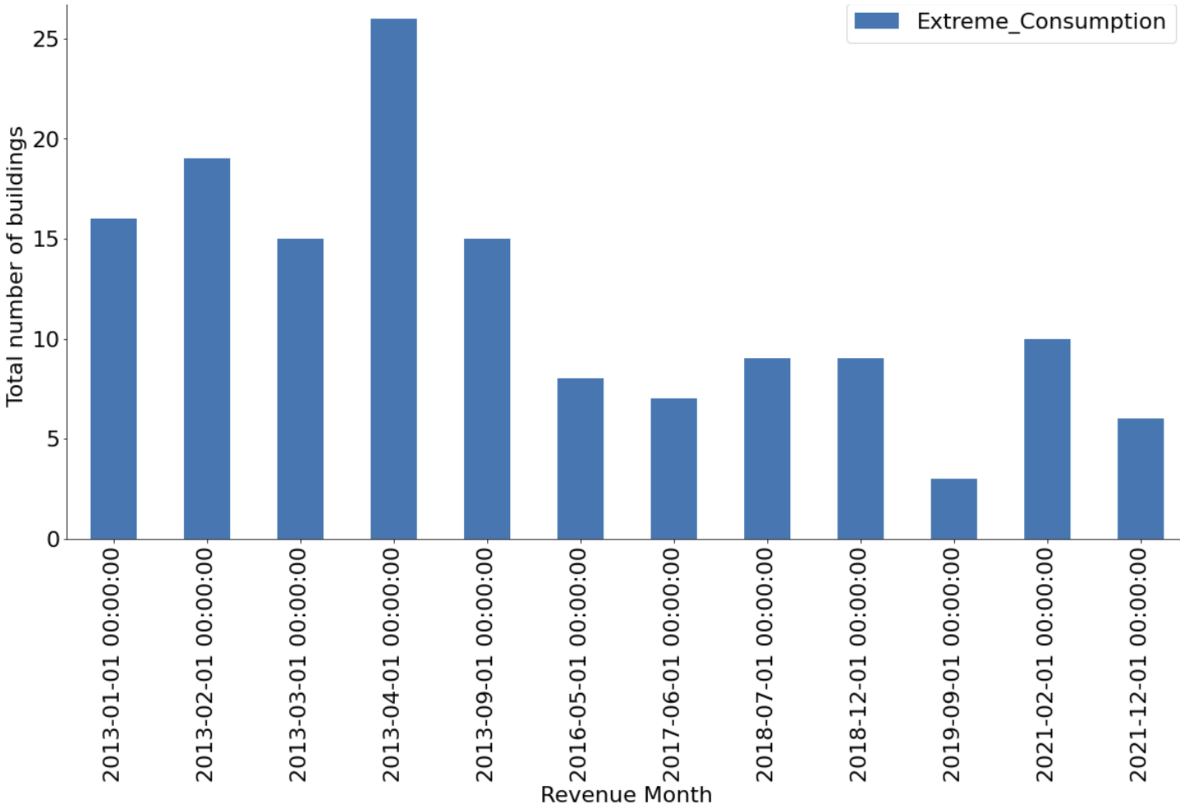

Figure 19: Top twelve extreme consumption months of Manhattan.

In figure 19, we calculate the top twelve extreme months. The y-axis in the figure represents the total number of buildings, and the x-axis represents the top twelve extreme months. April 2013 has the highest number of extreme buildings, which means this month has the most extreme consumption over nine years. There is extreme consumption in early 2013

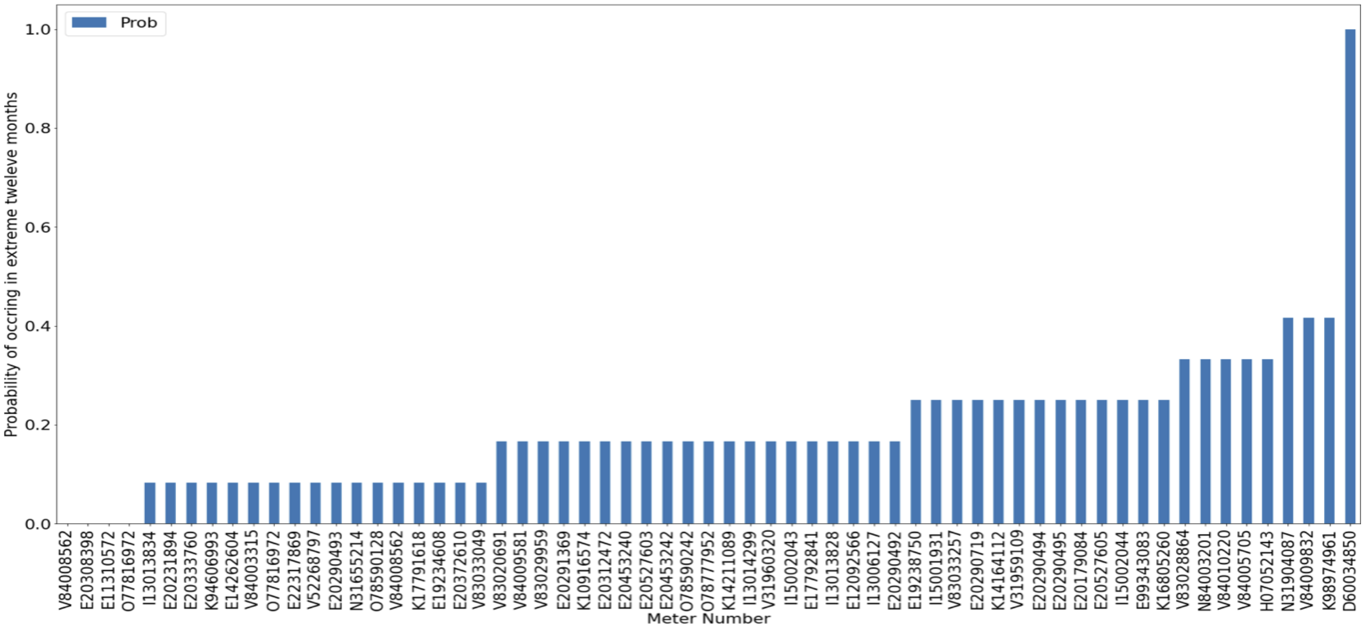

Figure 20: Extreme Occurrence Analysis of each building

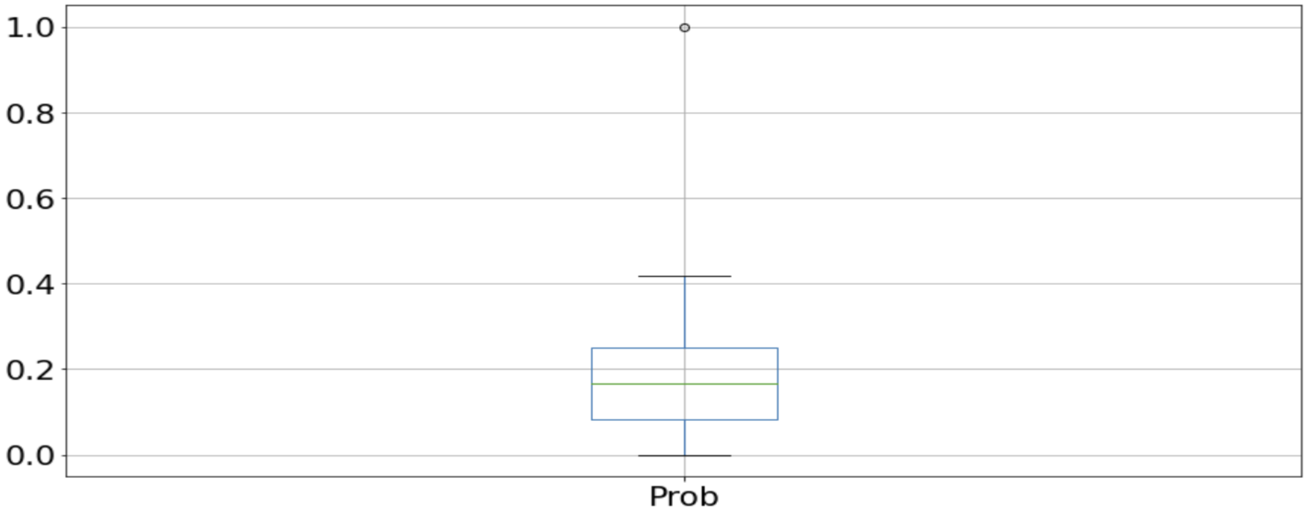

Figure 21: Box plot of occurrence of extreme twelve months.

In figure 20 and figure 21, we calculate the extreme occurrence of each building. In figure 20, The y-axis represents the probability of occurring in extreme twelve months, and the x-axis represents all the buildings of Brooklyn. Most of the buildings have a probability of 0.167 in extreme twelve months, which means that most of the buildings are occurring two times in the top twelve months. Also,one of the buildings is occurring in the whole twelve months. Four buildings are not part of extreme months.

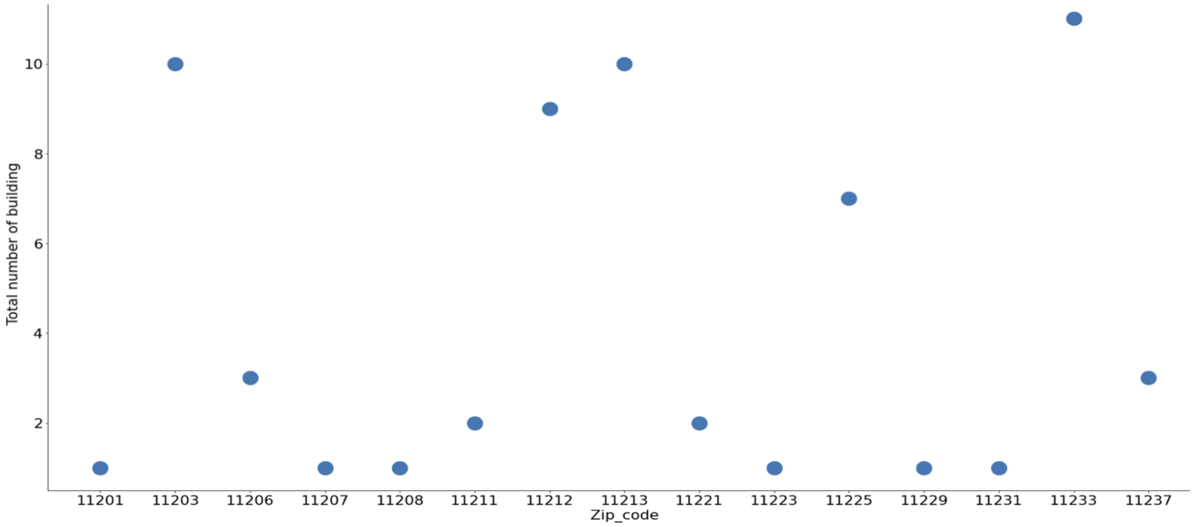

Figure 22: Total number of buildings in each zip code

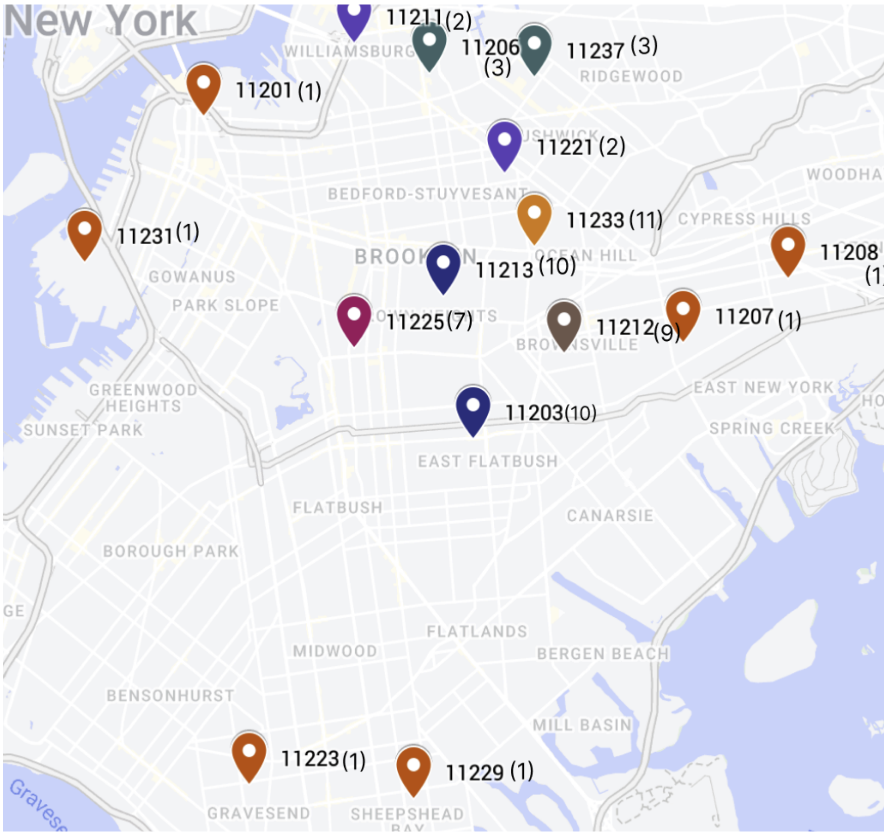

Figure 23: Spatial plot of Manhattan with zip code(count building)

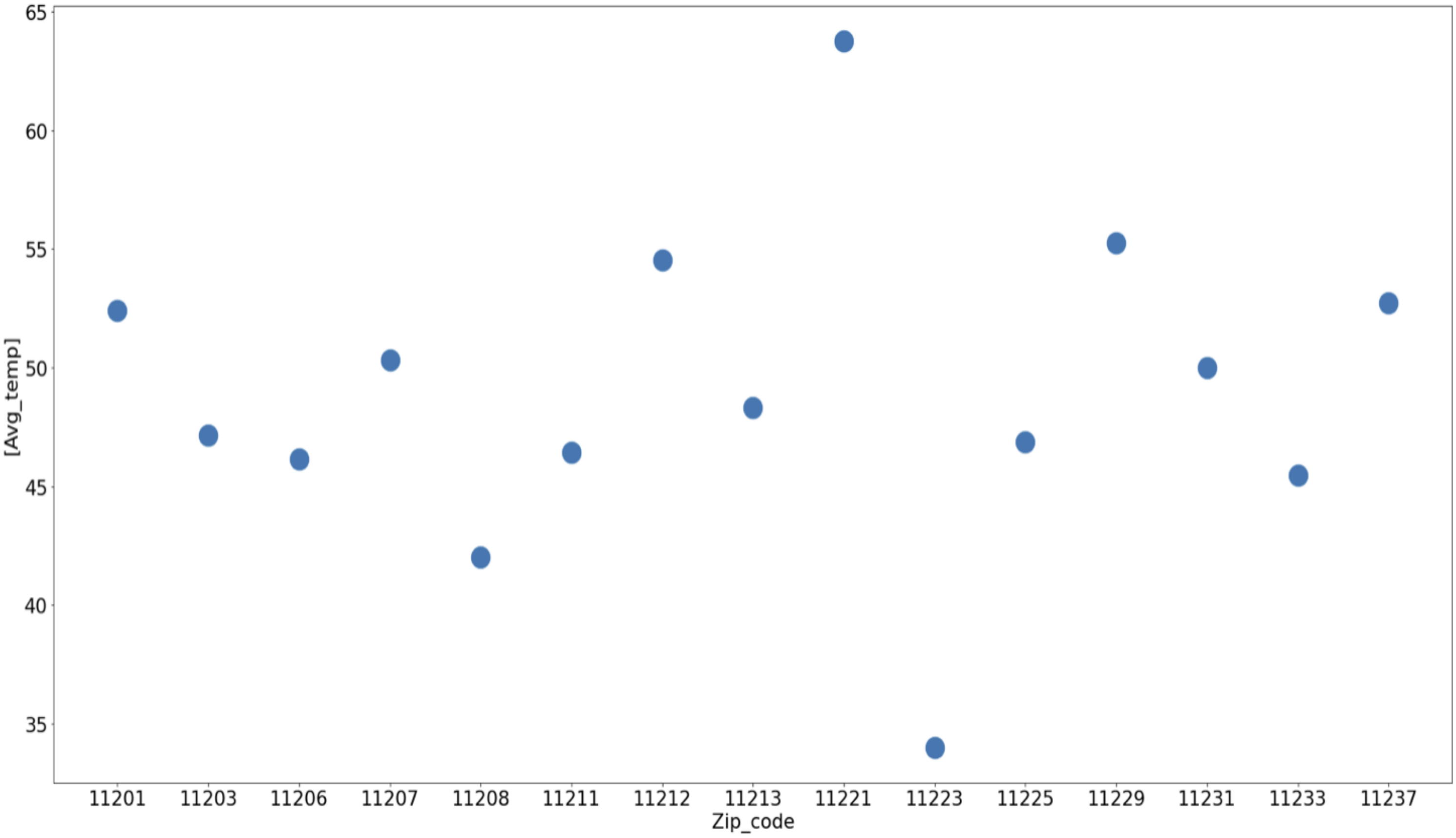

Figure 24: Average temperature in each zip code

In figure (22), we calculate the total number of buildings in each zip code. Figure (23), we get a spatial plot for the Brooklyn zip code, In figure 24, we calculate the average Temperature, in each zip code. We can see that zip code 11233 is a densely populated zip code, and there is high consumption. This zip code requires more attention for the conservation of the consumption of water. Figure(23) represents the plot of each zip code on the map to visualize where they lie. Most of the zip code lies within the range of 45 to 55 average temperature. Thus, most of the consumption is during cold weather. The lowest temperature is 34 and highest average is 64 in the Zip Code 11223 and 11221 , respectively.

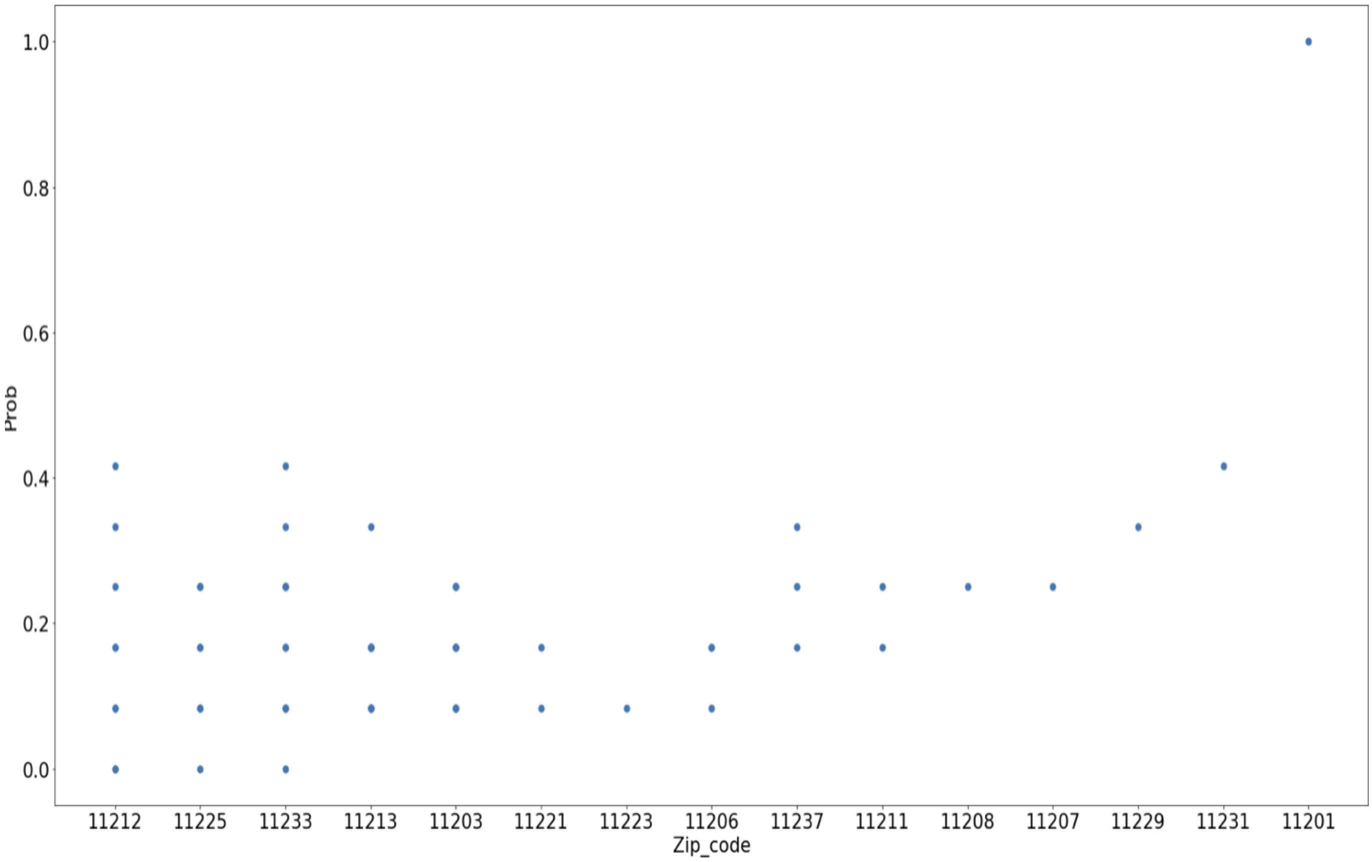

Figure 25: visualization of occurrence of extreme building in each zip code

In figure 25, the y axis represents the probability of occurrence in extreme 12 months and x axis represents its corresponding zip code. This graph mainly represents how each of the Zipcodes have variable distribution in their occurrence. For example, Zip Code 11231 has only occurrence once that means it has only single nature occurrence in its zip code. Whereas, zip code 11212 and 11213 have multiple nature occurrences in their zip code. They are occurring differently. We can also see that zip code 11201 has probability of 1 which means that it has occurred in all twelve extreme months.

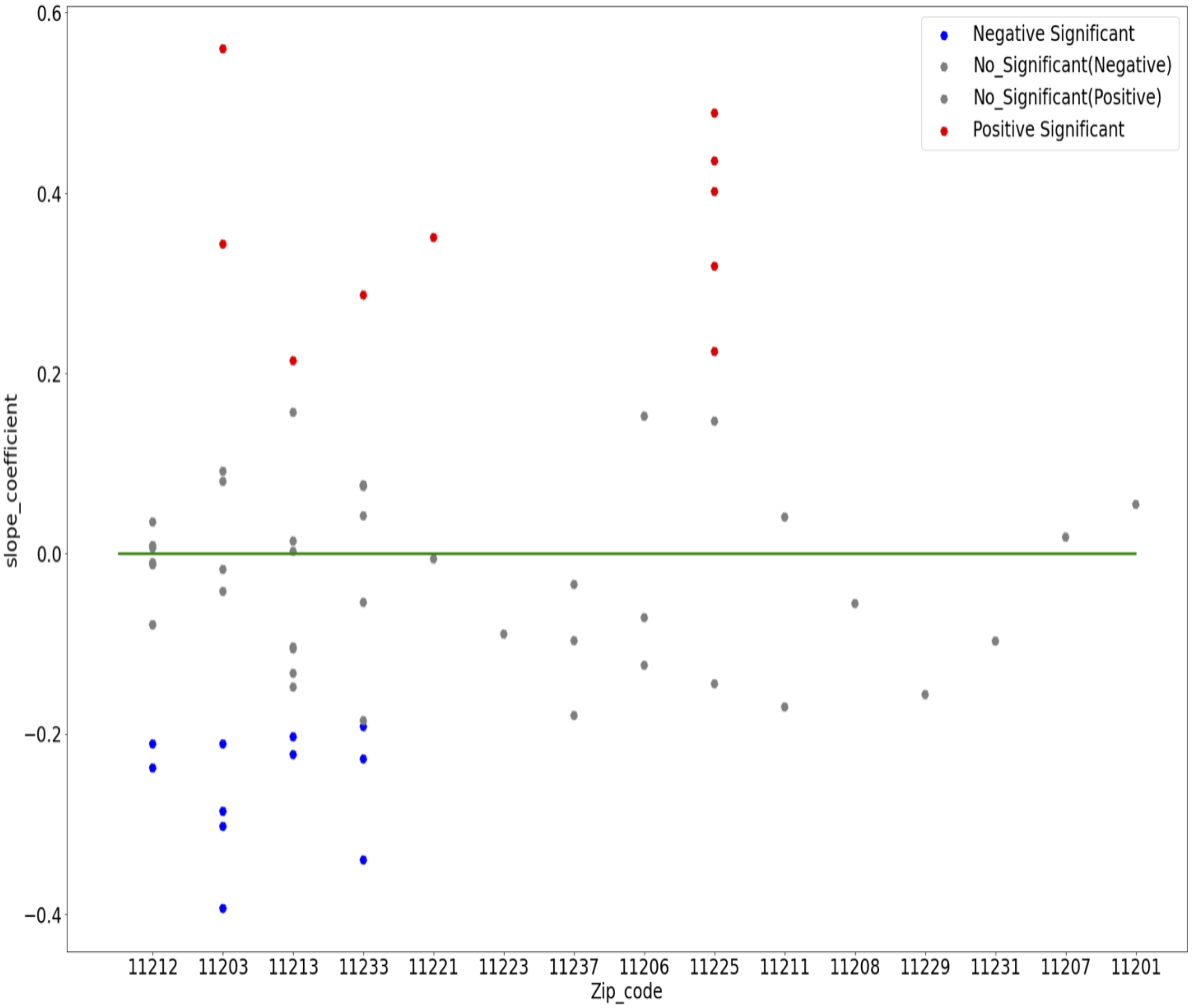

Figure 26: Trend Analysis of Manhattan building

Figure (26) represents the trend analysis of each building in Brooklyn with its zip code. The y-axis in the figure represents the slope coefficient, and the x-axis represents the zip code. Each dot with a different color represents its significance in each zip code. Here, Red and blue colors represent the Positive and Negative Significance of each building, whereas gray color represents the no significant buildings in the zip code. Figure (27) represents the total count of each significance in each zip code.

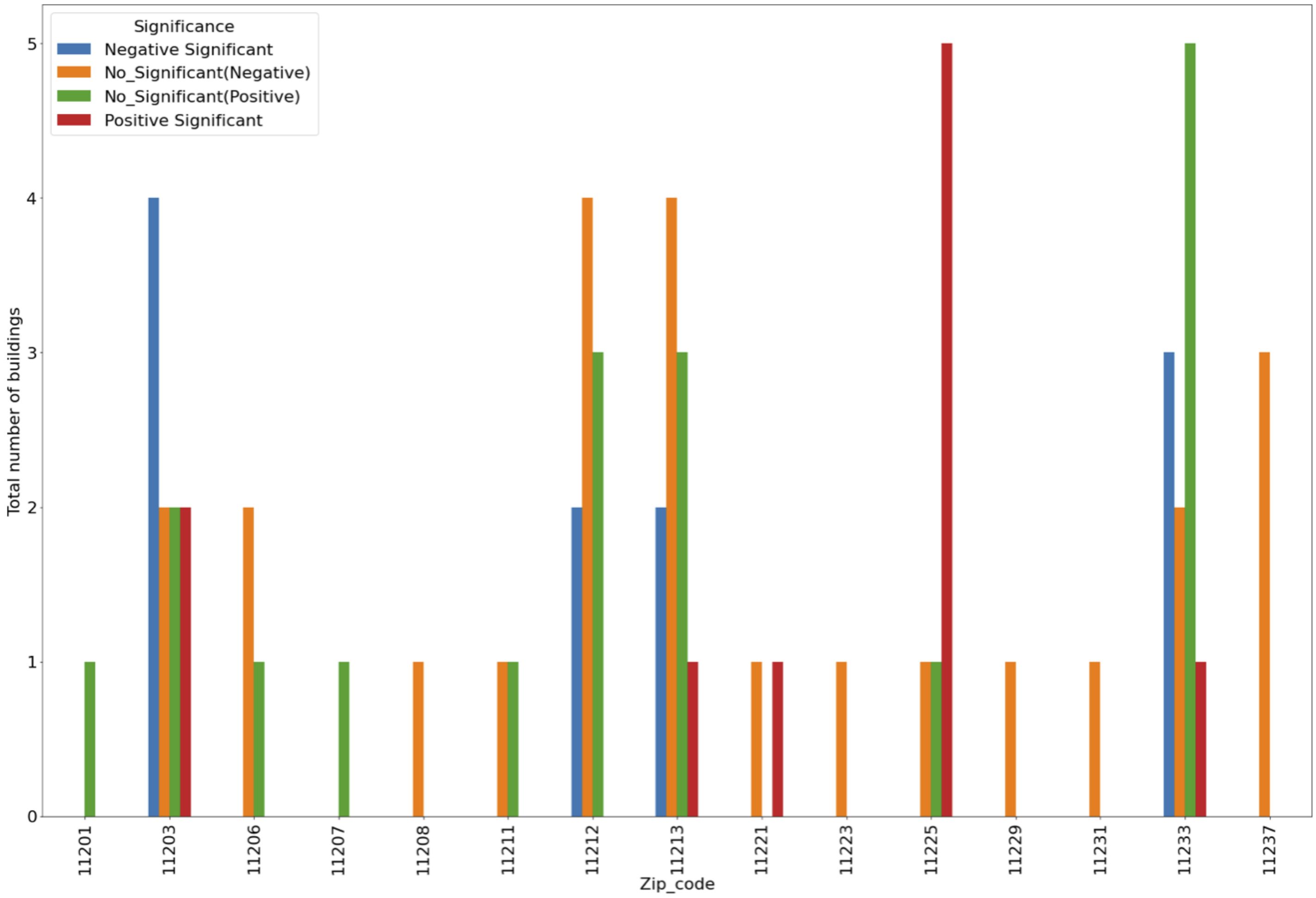

Figure 27: Trend Analysis of Manhattan building and count of each building

The Figure (27) represents the trend analysis of each building in Manhattan with its zip code and the total number of buildings. The y-axis in the figure represents the total number building, and the x-axis represents the zip code. Here, Red and blue colors represent the Positive and Negative Significant of each building. Also, green and orange color shows the positive no significant, and negative no significant buildings in each zip code. We need to account more on positive significant values in each code because they are increasing rapidly over time. Zip code 11225 requires more attention to minimize the consumption of water. Also, 11203 has the second highest positive significance. They have a signal of high consumption, which is necessary to minimize.

COMPARISON AND DISCUSSION

We are comparing only two boroughs of NYC because we didn't complete all the boroughs due to time constraints.

Comparing the Yearly analysis of Manhattan and Brooklyn, we found that both boroughs were high during early 2013. Brooklyn had few repetitive high consumptions in mid-2015, 2016, and 2017. If we view it closely, Manhattan's overall consumption has increased since 2017, and also Brooklyn's Consumption slightly increased since around 2018. We can say that overall consumption of water is increasing after a certain period of time and it is remaining constant with few irregular high and low consumptions. Comparing the Seasonality analysis, we had hypothesized that consumption of water in the winter season will be high because of the steam heating system, people staying at home, water waste initially to get hot tap water and so on. Looking at the seasonality figure for Manhattan in Figure 2, there is a high consumption in February and lowest in March. Overall their consumption is close to each month. For Brookllyn, June has the highest consumption, which contradicts our hypothesis, But looking at overall or the rest of the months, they are very close to each other as well. Thus, we can say that their water consumption does not gradually change over the season. It differs from person to person or community who uses high water in which month.

After Comparing the extreme percentile of Manhattan and Brooklyn. We found out the top 12 months from each of the boroughs. Early 2013 has had similar extreme months for both, and the rest of the extreme months differ. For both of the boroughs April 2013 is an extreme month. During the extreme month only one building from Brooklyn has occurred for all twelve extreme peaks. This is the building to be taken into account for further analysis. In Manhattan none of the buildings appeared to be for all top twelve peaks. only three buildings in Manhattan have occurred rapidly for a maximum of 5 occurrences during these top twelve peaks. Similarly for brooklyn three buildings are occurring for 5 times of extreme peaks. There are only three buildings for Manhattan and four buildings for Brooklyn, which didn't have any extreme months. Comparing the overall extreme occurrence,Manhattan buildings are occurring three times in top extreme months whereas Brooklyn buildings mostly occur at twice. Hence, one building from figure 20, whose probability is one is in extreme high consumption.

Comparing the total number of buildings with zip codes in each borough, we found that 10030 is a densely populated area in Manhattan, and zip code 11233 is a densely populated zip code in Brooklyn. Brooklyn has the most buildings among all the boroughs. Also, there are more zip codes where people reside in Brooklyn. But, looking at Manhattan, most of the buildings are in the same Zip Code 10030. There is more spread out in extreme consumption in each zip code in Brooklyn. Comparing the Temperature of Manhattan and Brooklyn, we found that most buildings have lower temperatures in high consumption months. It means people use more water in cold weather, which proves our initial hypothesis.

Comparing the trend analysis of each building, we calculate the Significance of each building in each zip code. In Figures 13 and 27, Red, blue, green, and orange represent the Positive significance, Negative Significance, Positive non-significance, and Negative non-Significance, respectively. We need to be more focused on Positive Significance, because they are rapidly increasing over time. For Manhattan, ZipCode 10030 has the highest significant value, and zip codes 10009, 10024, and 10035 have also Positive Significance value. They need more analysis for conservation of extreme consumption of water. Similarly, For Brooklyn, Zip Code 11225 has the highest Positive Significance. Also, 11203 has the second highest number of positive significance buildings. They need more attention for high consumption of water. Hence, we should keep an eye on these extreme zip codes for conservation of extreme consumption of water.

CONCLUSION

We found out that the most extreme consumption in Manhattan and Brooklyn is during cold weather. Also, the ratio of consumption of water is increasing periodically. Zip code 10030 is consuming excess water in Manhattan. Among them, five buildings are contributing to extreme consumption. For Brooklyn, zip code 11225, consumption is increasing rapidly, which requires more attention for the conservation of extreme water. The Brooklyn data set is more distributed compared to the Manhattan data set. In future, using alternative methods for heating systems and more awareness programs for water conservation should be conducted timely. Identifying the problems of each high consumed area around the world, and solving them to save water from high consumption will make water easily accessible for all the people around the world. It will finally lead us to a water problem free world. Hence, we should control the excess use of water all around.

ACKNOWLEDGEMENTS

This research was conducted online during Summer 2022. I thank Dr. Naresh Devineni for giving me an opportunity to work in this research and guiding in every step. I also thank the Data Science Scholars Program and Dr. Claude Brathwaite for providing a great platform to work on this project.

REFERENCES

Chambers, Aaron. "How Much Water Does NYC Use Daily?" *Hudson Reed*, 23 Apr. 2018, usa.hudsonreed.com/info/blog/the-truth-about-new-yorks-water-usage/#:%7E:text=New%20York%20City%20uses%20nearly,all%20of%20the%20facts%20below%E2%80%A6

"NYCHA Fact Sheet 2022." *NYCHA*, 2022, www1.nyc.gov/assets/nycha/downloads/pdf/NYCHA_Fact_Sheet_2022.pdf.

"Water Consumption And Cost (2013 - Feb 2022) | NYC Open Data." *NYC Open Data*, 13 May 2022, data.cityofnewyork.us/Housing-Development/Water-Consumption-And-Cost-2013-Feb-2022-/66be-66yr.

"Water Consumption in the City of New York | State of New York." *Open NYC*, data.ny.gov/widgets/ia2d-e54m. Accessed 5 Aug. 2022.

"The Water Crisis." *The Last Well*, thelastwell.org/the-water-crisis/?gclid=Cj0KCQjwuaiXBhCCARIsAKZLt3nSmGXI4cVjFaIKOSGB8GU2eZ6-6vhIZeG34X4PuRPG8saZ5o1CrVgaAjG2EALw_wcB. Accessed 3 Aug. 2022.