**REPORT**

**CASE STUDY #4**

**Group 10**

**Background and Motivation**

1) Our business problem to solve:

We intend to help people and the government of New York City to take more advantages from the current government open data sets. From the abundant raw data, we would like to present interesting and important findings which can help people know more and better about what a neighborhood is like and what is going on around it. In this case, we take a focus on the districts of Manhattan.

2) Why it is important to solve?

In the age of information explosion, people have more access to various data every day, including the government data which used to be closed and confidential in the past. The open data should be able to produce many important insights for the people as it is authoritative, reliable, and covers many dimensions of the people’s daily life. However, for many people, it is not easy to see the insights from the data and to use it effectively due to the big volume, sparsity of information, and lack of inner correlation between the raw data. Therefore, we propose an effective media to address the issue so that people can understand and take as many advantages as possible from the data in a better way.

3) What is our idea to solve the problem?

By collecting selected data from NYC OpenData website ([Link](https://nycopendata.socrata.com/)), we filtered 7 data sets of different categories: 311 Service Requests - on Commercial Noise, Illegal Parking, Industrial Waste, and Air Quality, Park Location, Hospital and Health Corporation, and WIFI Location. With these data sets, we can produce a vivid picture of how a specific region is like, describing it in a more well-rounded way for people to understand and helping them in decision-making regarding their choices of areas to stay or visit.

4) What differences we could make with our data science approach?

With the support of the API of Google Map, we combine all the different facilities or events together into a region based on their location information such as longitude, latitude, and address. All the data that we collect from NYC OpenData and Google Map API are merged to produce new information from which people can learn the 7 attributes of each district in the forms of easy-to-follow maps, graphs, and charts. With this handy tools, people can better figure out and decide which neighborhood they should refer, stay, or visit based on their own interests or needs. For instance, in terms of noise, which district has least complaints and why? Besides, we produce district based-clusters containing all the attributes, using unsupervised clustering algorithms: Scaled PCA and K-Means. The final two clusters divide the 70 districts into 2 clusters, one with significantly higher 4-type complaints than the other.

5) Why do we believe the idea deserves the investment of the "sharks"?

Nowadays, we are surrounded by information stream, bombarding us with fragmented messages. Meanwhile, without an effective presentation of data and information, it is not easy to link them and see what knowledge they show us. Therefore, our product will enable people to get to know a place from a higher level, regardless they are the citizens of NYC or new comers or even visitors. We believe that such multi-dimensional information will be pretty much useful for various needs, especially for those who are moving out or looking for long or short-term accommodations. Furthermore, the current OTA websites like booking.com as well as local host renting sites like Airbnb only provide information about housing conditions without any related/supporting information such as transportation and recreation. Meanwhile, our product, integrating the unique 311 information (contains people’s complaints from noise to potholes happening around them), can definitely help people dig more about a place for better decision making. To make a right decision, sometimes we need more than one things to take into account. That’s why we believe that offering the multi-dimensional data product is an effective solution.

# Algorithms

# We accessed official data from NYC OpenData and selected several data sets. By using Google maps geocoding API and Pandas, we could get readable data frame. We merged all these data sets by combining the features that have same district. Then we got a processed data sets. So the problem we choose in problem 1 can be translated to a math problem of estimating in a certain district whether there has sophisticate public facilities.

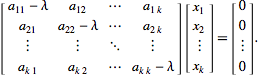
Here is our detailed math solution step by step:

1. We use scale to do the normalization work.

2. By leveraging the normalized data, we do principal component analysis through computing the covariance matrix and eigenvalues:







3. After we completed principal component analysis, we deployed K-means to get our final result.

\sum_{i=0}^{n}\min_{\mu_j \in C}(||x_j - \mu_i||^2)

# Findings

# Part I Clustering

# The following plot and table are what we achieved by using K- Means:

# The red dots represent the cluster with Label 1, which contains 14 districts as follow. Judging by the statistical result table, Label 1 cluster has more complaints in four types. Relatively, Cluster Label 0 receives fewer complaints by people.

# cid:920F6829-0204-480C-A1C0-378952D78767@WIFI.WPI.EDU

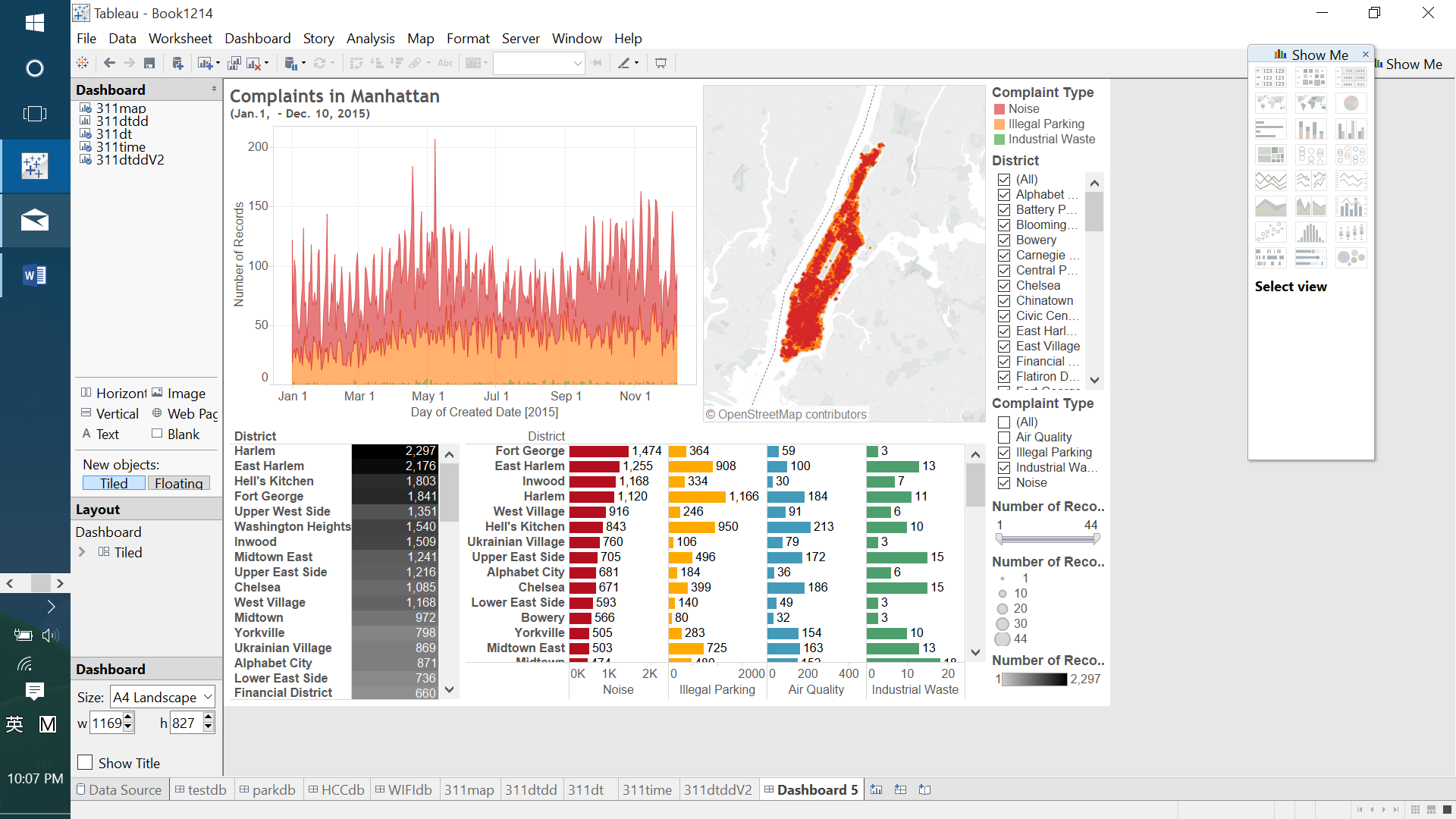
# cid:83DFC83C-A81B-487B-BB75-A7344357442D@WIFI.WPI.EDU

# Part II Story-telling

As our products aim to serve all kinds of data, how can we effectively convey the information is one of our priority concerns. Consumers are expected to get what they want by just clicking the mouse without frequent switching through windows or comprehending numerous numbers and index. Thus, Data visualization, in an interactive way, becomes our choice.

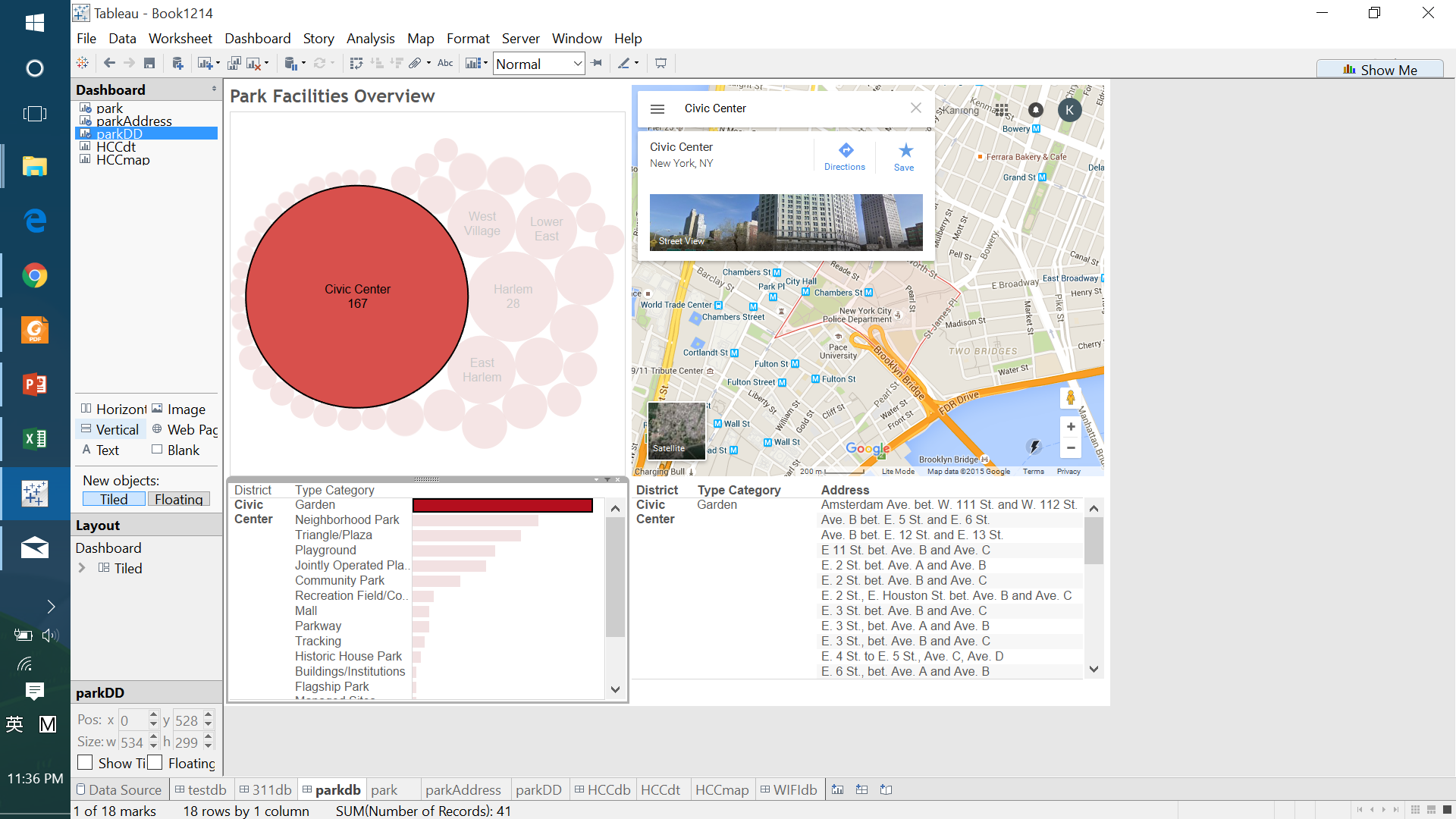
We will tell some simple stories by casting an eye on different dimensions of our data.

Story 1



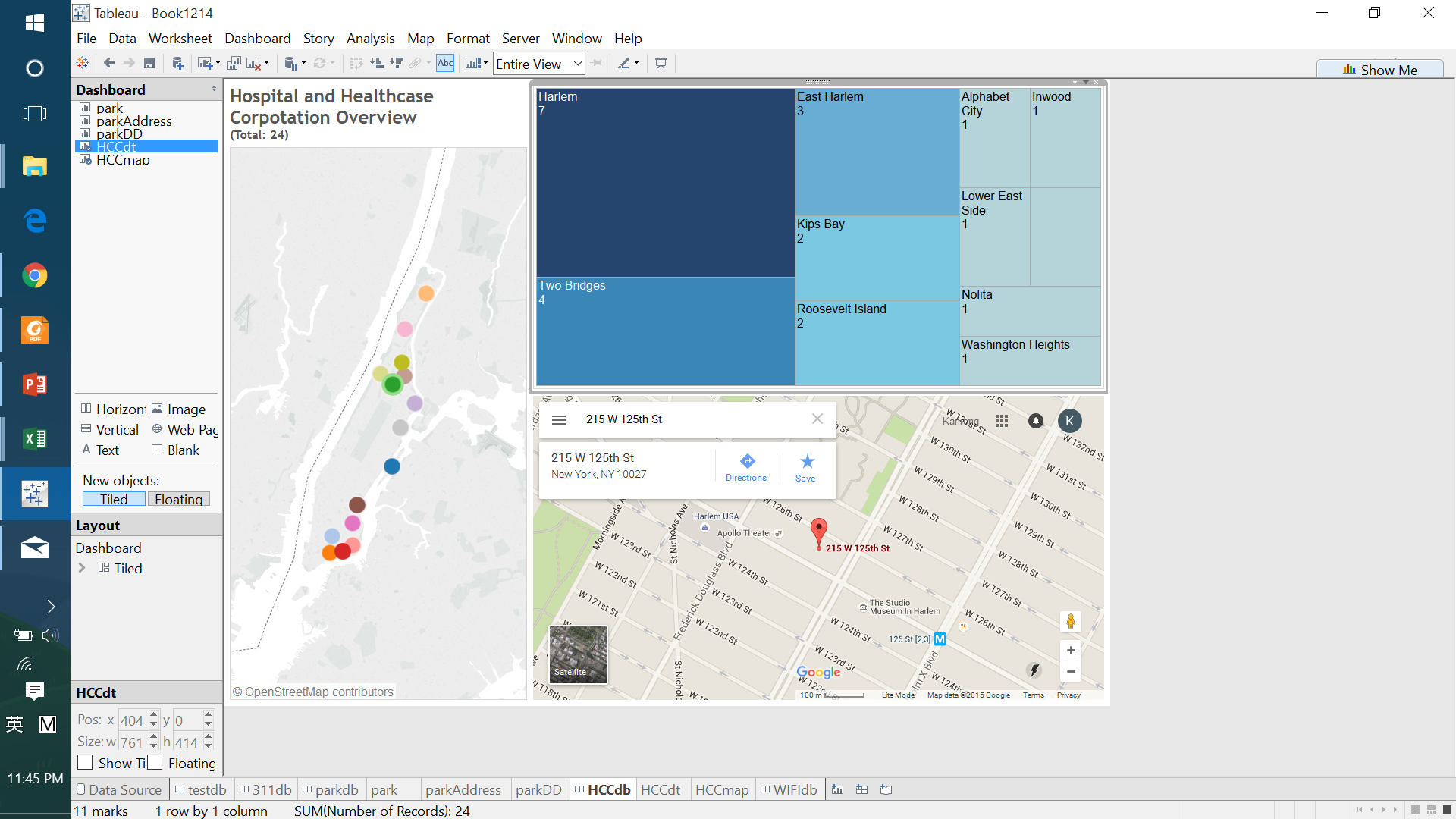
Struggling to find a quiet place in this city that never sleeps? Then you should probably avoid Fort George, or even Washington Heights! As the former receives the most noise complaint, while the latter is troubled by illegal parking. On the other hand, since World Trade Center and Upper Manhattan suffered least from these two, consider your budget and check the housing there.

Story 2



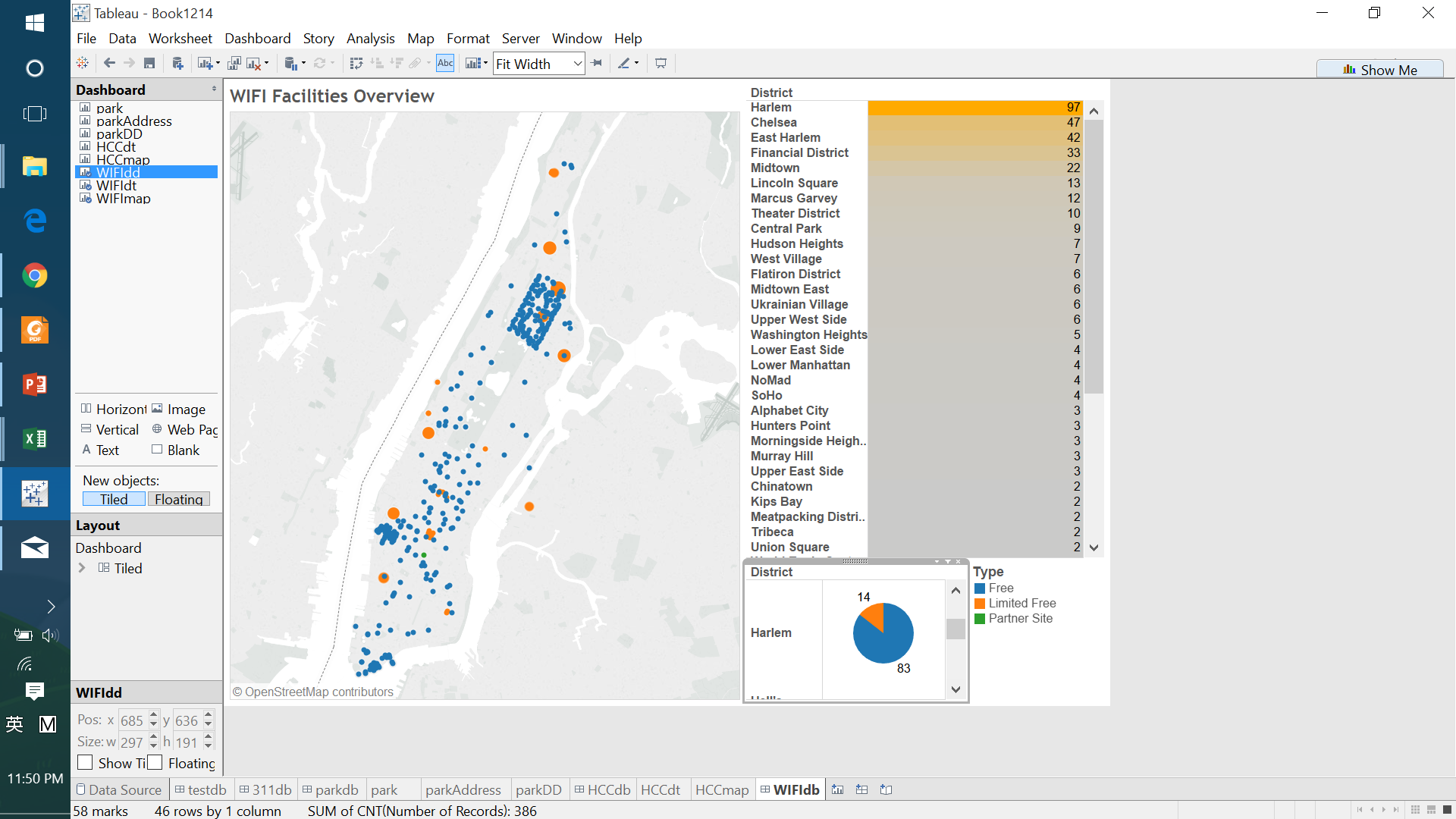
Civic Center owns the most parks, and majority of them are gardens. Go and enjoy them along the Brooklyn Bridge in this cozy district.

Story 3



There are 24 Hospital and Healthcare Corporation in Manhattan. Harlem, adjacent to Columbia University, gathers 7 of them, may be a good choice for senior citizens.

Story 4



Luckily, there’s plentiful free WIFI spots in Manhattan, and Harlem, again, ranks the first. Checking the district on the map which is crowded by blue dots, that’s Harlem! When we dug deeper to learn why there are so many outdoor free WIFI spots in this neighborhood, we found that in 2013, it launched country's largest free public WIFI network in Harlem.

# Appendix

Our presentation script for 90-second pitch:

Hi, do you know that the New York City government has been publishing open data to the public since many years ago? Well, the data has been scaling up with abundant data sets related to various sectors of daily life. But, unfortunately, we haven’t seen any integrated data presentation and analysis provided by the website so that people can learn and better understand what a neighborhood is like and what’s going on around it. We believe that public needs to have this kind of knowledge so that they can make more accurate decisions where to go or visit or stay based on their needs or interests. Therefore, we want to help the people not to make wrong decisions by developing a data product of 7 attributes available in the website, which are people complaints on noise, air quality, industrial waste, and illegal parking, as well as the locations of health centers, schools, parks, and subway stations within the district of Manhattan. All these data sets, coupled with support data from Google Map API and Pandas, are filtered and merged to produce a more handy visualization of the data analysis. That way, it’s easier for people to learn which neighborhood is the most quiet one to get good sleep and rest, which neighborhood having the closest access to health centers so that it would be better for elderly people to stay, or which neighborhood having the cleanest environment which is good for your health, which neighborhood having most of garden parks to enjoy your time with your loved ones, and so forth. On the other hand, the data sets are processed using Scale PCA and K-Means algorithms to produce neighborhood clusters in 2 labels: those having more complaints and those having less complaints. So, if you don't want to lose anything because of wrong decisions, why not trying our accurate data product!