**Sampling Plan**

**INFO-MATIX Systems**

**Data link**: <https://data.cityofnewyork.us/City-Government/NYC-Wi-Fi-Hotspot-Locations/yjub-udmw>

The NYC WIFI Hotspot Locations would like to expand their WIFI connectivity in Brooklyn and have requested the help of INFO-MATIX Systems to analyze the connectivity in regards to Brooklyn and give appropriate measures on how to conduct these operations.

Data was provided by the Department of Information Technology & Telecommunications (DoITT), but accessed from the NYC Open Data portal. A direct link to the page is provided as data link at the top of the document.

Since from the dataset provided we had different cities like New York, Brooklyn, Queens, Bronx, Staten Island among other cities, we locked through (.loc) the whole population according to Brooklyn city so that it can fit our target audience for the study.

**SAMPLING OBJECTIVES AND RELIABILITY**

Using the data of NYC Wi-Fi Hotspot Location, evaluate the relationship between various trends in the data. A step of calculating the different statistical variables i.e. mean, median, mode, quartiles and the range function provides an in-depth analysis to familiarize with certain parameter and nearly all columns and rows in the data set.

The research question to be examined is; Does Location T affect the persons connected to the Wi-Fi network?

This will require one to perform various hypothesis to at least try and evaluate the relationship between different locations in the data set. This will aid in the identification of which location is the preferred amongst the ones present and also know by what degree is the null hypothesis being rejected or accepted.

Performance of various sampling techniques will be done so as to choose the appropriate method by which a sample will be randomly selected.

**POPULATION**

The data had 29 columns and we dropped some so as to fit our defined target population. An example of these columns are name, SSID, remarks, BBL among others. The data set parameters are outlined and explained below. Brooklyn city has a total of 823106 persons. This is obtained from a total sum of the Census Tract column.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Type** |
| Objectid | Identification number automatically generated | Number/ int64 |
| Type | Type of Wi-Fi provided by franchise | Plain Text/ object |
| Provider | Franchise who is providing the Wi-Fi connection | Plain Text/ object |
| Location | A brief description of where the Wi-Fi point is | Plain Text/ object |
| Latitude | Latitude: Points that fall North or South of the equator | Number/ float64 |
| Longitude | Longitude: Points that fall East or West of the equator | Number/ float64 |
| X | A horizontal value where a point is located | Number/ float64 |
| Y | A vertical value where a point is located | Number/ float64 |
| Location\_T | he type of location that a Wi-Fi hotspot is present in | Plain Text/ object |
| City | The city in which a hotspot is located | Plain Text/ object |
| Activated | The approximated date that this Wi-Fi site went active | Date & Time/ object |
| BoroCode | The New York City borough where the hotspots are located. | Number/ int64 |
| Census\_Tract | 2010 U.S. Census data based on location | Number/ int64 |
| BCTCB2010 | Combined value of Borough CD and Census 2010 | Number/ int64 |
| BIN | Building Identification Number | Number/ int64 |

**SAMPLING APPROACH**

Sampling is the process of selecting certain members or a subset of the population to make statistical inferences from them and to estimate characteristics of the whole population. Different methods of probability sampling were used to try and identify the degree of effectiveness of each sampling method.

Among the ones used are:

* Simple random sampling
* Stratified sampling
* Cluster sampling
* Systematic sampling

**Simple Random Sampling**

This is a technique used to pick the desired sample size and for selecting observations from a population in such a way that the observation has an equal chance of selection until the desired sample size is archived.

Since the whole population had been locked to only Brooklyn city, a sample of 100 was obtained randomly from the population to describe the dataset. The sample had a census mean of 194.23 persons and a standard deviation of 230.695.

**Stratified Sampling**

It’s a sampling technique that involves breaking a population into key subgroups and obtaining a simple random sample from each group. The steps involved are;

* Define the population
* Choosing the relevant stratification
* Listing the population according to the stratification
* Choosing your sample size
* Calculating a proportionate stratification

Each of these steps are conducted in the colaboratory notebook. A stratified sample of 58 was obtained according to location t, the mean of the census was 225.967 and a standard deviation of 226.977.

**Cluster Sampling**

This is a sampling method in which the entire population of study is divided into externally homogeneous but internally heterogeneous groups called clusters. We defined our cluster according to location, i.e. latitude and longitude and this was achieved with the pandas method. iloc to slice the columns we want.

The next step was creating an object from the KMeans method we imported from sklearn library, this indicates the number of clusters we are aiming for. We clustered the results into a new copy of data set as either 0 or 1. Then finally plotted the new copy as per its longitude and latitude to identify the grouping of the clusters.

The mean for the census from the clusters is 273.5597 with a standard deviation 259.9518.

**Systematic Sampling**

The technique is used where the sample members from a population are selected according to a random starting point but with a fixed periodic interval. The idea in systematic sampling is that, given the population units numbered from 1 to N, we compute for the sampling interval, given by k=N/n, where n is the number of units needed for the sample.

A conclusion of pursuing with Simple Random Sampling method was done. This is because the data set had an equal chance of selection and hence no bias and also its simple to interpret the results from the sampling method.

**SAMPLE SIZE**

The data set had 3319 columns before being broken down to the scope of study i.e. Brooklyn City. Once broken down to Brooklyn the data had 702 columns and at least some missing values that were resolved by dropping the columns with the missing values. After checking for outliers and dropping the outliers the data had 586 rows. But since we’re using simple random sampling technique our data has 100 rows that are generated with an equal chance.

No duplicates were found in the data set though standardization was done in order for the row titles to have all an uppercase, with spaces and also have white spaces all around. A brief description of each element in the data set is presented below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Objectid | latitude | Longitude | X | Y | Borocode | Cencus\_Tract | BCTCB2010 | BIN |
| Count | 100.00 | 100.00 | 100.00 | 100.00+02 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00+02 |
| Mean | 10859.69 | 40.68 | -73.97 | 9.91+05 | 186741.93 | 3.0 | 194.23 | 194.23 | 2.27+06 |
| Std | 1001.00 | 0.02 | 0.02 | 6.67+03 | 7283.19 | 0.0 | 230.69 | 230.69 | 1.43+06 |
| Min | 9672.00 | 40.61 | -74.03 | 9.77+05 | 162605.44 | 3.0 | 5.00 | 5.00 | 0.00+00 |
| 25% | 9888.50 | 40.67 | -73.99 | 9.87+05 | 181738.12 | 3.0 | 17.25 | 17.25 | 0.00+00 |
| 50% | 10434.00 | 40.68 | -73.98 | 9.89+05 | 18816.86 | 3.0 | 137.50 | 137.50 | 3.01+06 |
| 75% | 11941.75 | 40.69 | -73.93 | 9.96+05 | 191627.88 | 3.0 | 279.00 | 279.00 | 3.26+06 |
| max | 12908.00 | 40.72 | -73.90 | 1.01+06 | 202444.11 | 3.0 | 1098.00 | 1098.00 | 3.42+06 |

The interquartile (iqr) range is the difference between the third quartile (q3) and the first quartile (q1). The iqr is shown below.

* OBJECTID 2.053250e+03
* LATITUDE 2.715288e-02
* LONGITUDE 3.032110e-02
* X 8.413264e+03
* Y 9.889764e+03
* BOROCODE 0.000000e+00
* POSTCODE 2.500000e+01
* CENSUS\_TRACT 2.617500e+02
* BCTCB2010 2.617500e+02
* BIN 3.257036e+06

We have a range of 1093 between the maximum and minimum of Census Tract. This shows the spread of our data in the data set between the minimum value and the maximum value. The variance of the whole dataset is presented below.

* OBJECTID 1.002005e+06
* LATITUDE 3.996417e-04
* LONGITUDE 5.774149e-04
* X 4.442818e+07
* Y 5.304492e+07
* BOROCODE 0.000000e+00
* POSTCODE 1.699661e+02
* CENSUS\_TRACT 5.322010e+04
* BCTCB2010 5.322010e+04
* BIN 2.046203e+12

Skewness is the asymmetry in a symmetrical bell curve in a data set. The skewness is shown below.

* OBJECTID 0.489437
* LATITUDE -0.650053
* LONGITUDE 0.843328
* X 0.843345
* Y -0.649987
* BOROCODE 0.000000
* POSTCODE 0.420394
* CENSUS\_TRACT 1.746759
* BCTCB2010 1.746759
* BIN -0.962752

The kurtosis is a statistical measure used to describe the distribution of observed data around the mean.

* OBJECTID -1.289693
* LATITUDE 0.698452
* LONGITUDE 0.493481
* X 0.494749
* Y 0.699498
* BOROCODE 0.000000
* POSTCODE -1.227337
* CENSUS\_TRACT 2.923925
* BCTCB2010 2.923925
* BIN -1.045453

The latitude, longitude, x and y all are platykurtic since they are less than 0 (Kurt < 0). Borocode is mesorkutic since its equal to 0 (Kurt = 0) and objectid, postcode, census tract, bctcb2010 and bin is leptokurtic since its greater than 0 (Kurt > 0).

**SAMPLE REPRESENTATIVENESS**

With the use of simple random sampling the population has an equal occurrence of being selected in the sample representative. This shows no biasness in the selection of the sample and the equal distribution in the occurrence of selection of samples. This is shown by the correlation of the whole sample that’s a representation of the population.

A strong correlation exists between census tract and longitude, census tract and x, bctcb2010 and longitude, bctcb2010 and x all have appositive strong correlation. This is because there is a correlation of 0.7 and greater (corr of 0.7 to 0.9).

A moderate correlation doesn’t exist in the sample.

A weak correlation is evident between latitude and objectid, latitude and longitude, latitude and x, latitude and postcode, latitude and census tract, latitude and bctcb2010, latitude and bin have a correlation of between 0.1 and 0.4 hence having a weak correlation.

The median of the sample is shown below;

* OBJECTID 1.043400e+04
* LATITUDE 4.068492e+01
* LONGITUDE -7.398244e+01
* X 9.891185e+05
* Y 1.888169e+05
* BOROCODE 3.000000e+00
* POSTCODE 1.121500e+04
* CENSUS\_TRACT 1.375000e+02
* BCTCB2010 1.375000e+02
* BIN 3.010212e+06

A step further was done using the Principal component analysis to try and identify a 2d graph plot of the sample to be able to have a visual of certain things. The explained variance ratio is 1.00, 2.13e-33.

**DATA COLLECTION**

Data obtained was in secondary form, this is because the data was downloaded from the NYC Open Data site. “Locations of street light poles, traffic light poles and utility poles approved for the potential installation of mobile telecommunications equipment by companies authorized by the New York City Department of Information Technology and Telecommunications.”

The above is a direct citation from the NYC Open Data site, the data was collected by individuals who visited each location that already has a Wi-Fi router in the Borough of Brooklyn and recorded in an excel document that can be accessed by the general public.

The parameters required for our analysis are the following rows; Objectid, Latitude, Longitude, X, Y, Borocode, Postcode, Census Tract, BCTCB2010 and BIN. These rows provided a clear analysis for the research question in the state of Brooklyn and the insights were able to be visually seen and also mathematically calculated.

**QUALITY ASSUARANCE**

A test of the hypothesis was performed to critic the evaluation of the null hypothesis. A hypothesis is a precise, testable statement of what the research predict will be the outcome of the study. Various sub groups of hypothesis were formed to try and assert the difference between different Location T groups.

* **Hypothesis 1**

Null hypothesis - there is no significant difference in the average Location T of Outdoor and Outdoor Kiosk

Alternate hypothesis - there is significant difference in the average Location T of Outdoor and Outdoor Kiosk

The hypothesis provided the following result. Ttest\_indResult (statistic=-1.8791780056303962, p-value=0.06424905686773383)

Since the p-value is greater than 0.05, fail to reject the null hypothesis.

* **Hypothesis 2**

Test of the Hypothesis - Is the average Location T of Indoor and Library

Null hypothesis - there is no significant difference in the average Location T of Indoor and Library

Alternate hypothesis - there is significant difference in the average Location T of Indoor and Library

The hypothesis provided the following result. Ttest\_indResult (statistic=-0.024801632331812228, p-value=0.9812769442296752)

Since the p-value is greater than 0.05, fail to reject the null hypothesis.

* **Hypothesis 3**

Test of the Hypothesis - Is the average Location T of Outdoor TWC Aerial and Outdoor Kiosk

Null hypothesis - there is no significant difference in the average Location T of Outdoor TWC Aerial and Outdoor Kiosk

Alternate hypothesis - there is significant difference in the average Location T of Outdoor TWC Aerial and Outdoor Kiosk

The hypothesis provided the following result. Ttest\_indResult (statistic=-2.381752740359555, p-value=0.025027782418806134)

Since the p-value is less than 0.05, fail to accept the null hypothesis.

**IMPLEMENTATION PLAN**

As observed form the analysis it’s clear that there is linear relationship between Indoor and Library Wi-Fi usage despite the library having a higher mean compared to the indoor location. This can be due to safety of devices i.e. the laptop, ipad / tablet or even mobile device used to access the Wi-Fi network.

At 95% confidence interval this is also observed from the confidence interval relationship between all parameters in the location, as observed in order of descending; library, indoor, outdoor kiosk, subway station, outdoor and outdoor twc aerial is the relationship from the highest confidence interval (CI) to the one with the lowest. CI is the range of values that you can be 95% certain contains the true mean of the population.

Its further expressed by the 98.01% acceptance of the relationship between the indoor and library usage of the Wi-Fi. This is noted by the p- value between indoor and library location. It’s clear that outdoor has the highest value count and despite it being preferred by most people, it doesn’t have a large std and mean.