**Travelezy Case Study**

Table of Contents

[Objective: 1](#_Toc1454091173)

[Data cleaning 2](#_Toc1470984849)

[Dashboards 2](#_Toc111388114)

[Dashboard 1 3](#_Toc1064581800)

[Dashboard 2 3](#_Toc1112842555)

[Purvi Jain 4](#_Toc1507685401)

[Using Tabpy, Clustering of different areas based on zip codes using K Means clustering Algorithm 5](#_Toc1898057240)

[Which category does mostly in top 100 hosts by review scores rating belong to? 6](#_Toc738586245)

[Insights 7](#_Toc1777422657)

[Daniyar Kurmanbayev 7](#_Toc1876567340)

[Price vs Review Score Correlation between Property Types and Years 8](#_Toc2039405404)

[Price vs Review Score Correlation between Room Types and Years 8](#_Toc1548215755)

[Insights 9](#_Toc1960722107)

[Shanka Attanayake 9](#_Toc862969170)

[Insight: 10](#_Toc500795314)

[Vrinda Parameswaran: 10](#_Toc503882449)

[Insights: 11](#_Toc1691108064)

[Price vs Beds 11](#_Toc94751417)

[Number of reviews vs Price 11](#_Toc960847111)

[Review Score Rating vs Price 12](#_Toc1161147771)

[Tanvir Singh Ahuja: 13](#_Toc1133107286)

[Insights: 15](#_Toc509745507)

# Objective:

Use more advanced data Visualization methods to:

* Explore, find insights, and facilitate the decision-making process
* Explore and discover patterns, recognize, and analyze trends
* Discoveries and corresponding analysis, identifying the problem(s) or issue(s) you found.
* Use any root-cause analysis techniques to solve the problem and/or improve performance

# Data cleaning

Before creating the operational dashboards, upon looking carefully at the chosen dataset using Tableau Prep Builder, we found many significant issues with the raw data including but not limited to:

* Replacement and removal of the null values in fields including beds, zip codes, neighbourhood, ratings, etc.

* Correction of the correct format of the zip codes

# Dashboards

## Dashboard 1

Chart

Description automatically generated

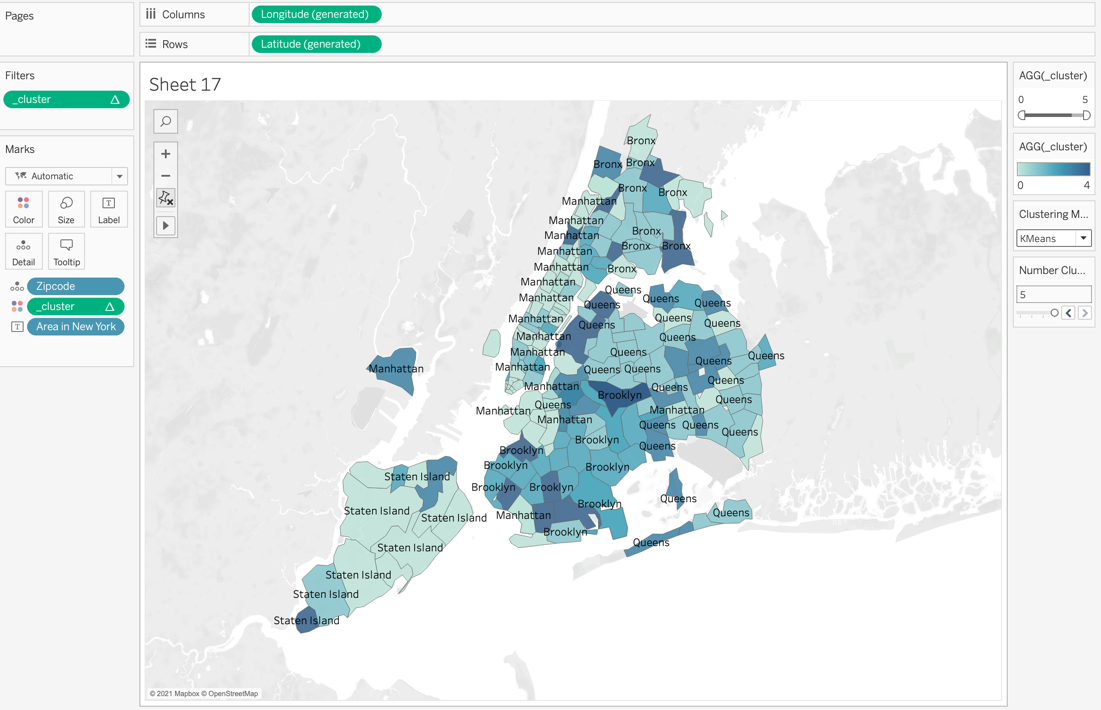
## Dashboard 2

Chart, bubble chart

Description automatically generated

# Purvi Jain

## Using Tabpy, Clustering of different areas based on zip codes using K Means clustering Algorithm



In this Visualisation, K Means clustering algorithm is used to cluster the areas based on their zipcodes. Two parameters are created and then a calculated field is used to connect to Tabpy. And the following script is used. Using this, it was able to identify some areas correctly.

INT(SCRIPT\_REAL("

import numpy as np

import numpy.ma as ma

from sklearn.preprocessing import StandardScaler

from sklearn.cluster import KMeans, AffinityPropagation, MiniBatchKMeans

print('Start')

sc= StandardScaler()

avg\_price = sc.fit\_transform(np.array(\_arg1).reshape(-1,1))

med\_review = sc.fit\_transform(np.array(\_arg2).reshape(-1,1))

med\_beds = sc.fit\_transform(np.array(\_arg3).reshape(-1,1))

sum\_review = sc.fit\_transform(np.array(\_arg4).reshape(-1,1))

n\_cl = \_arg5[0]

X\_comb = np.column\_stack([avg\_price, med\_review, med\_beds, sum\_review])

X = np.where(np.isnan(X\_comb), ma.array(X\_comb, mask=np.isnan(X\_comb)).mean(axis=0), X\_comb)

result = []

if \_arg6[0]==1:

kmeans = KMeans(n\_clusters = n\_cl, random\_state=134)

result = kmeans.fit\_predict(X).tolist()

elif \_arg6[0]==2:

minib = MiniBatchKMeans(n\_clusters = n\_cl, random\_state=134)

result = minib.fit\_predict(X).tolist()

else:

aff = AffinityPropagation().fit(X)

result = aff.predict(X).tolist()

return result

",

AVG([Price]),

MEDIAN([Review Scores Rating]),

MEDIAN([Beds]),

sum([Number Of Reviews]),

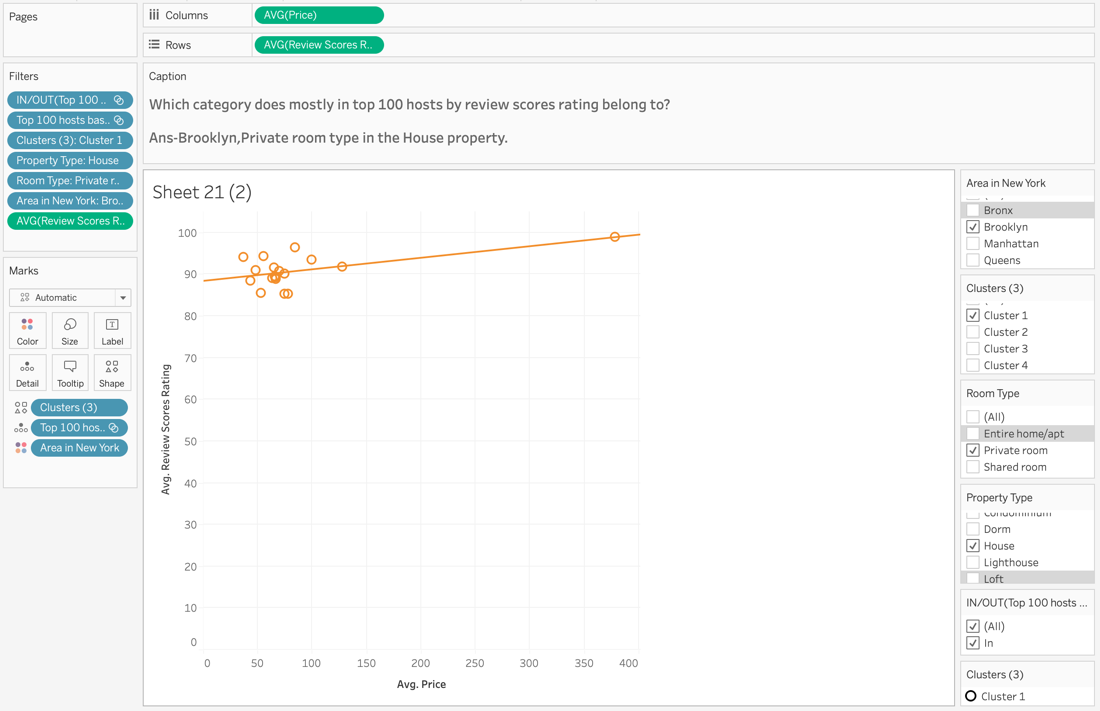
[Number Clusters],

[Clustering Method]

))

Which category does mostly in top 100 hosts by review scores rating belong to?

Ans-Brooklyn,Private room type in the House property. This is found using the clustering analytics by drilling down in top 100 hosts and selecting each area and then drilling down into all categories of room types and property types.



Insights:

Though the data shows that Manhattan is very popular but the data analysis shows that Brooklyn has large number of hosts in top 100, who get more average review scores rating.

# Daniyar Kurmanbayev

## Price vs Review Score Correlation between Property Types and Years

Chart, scatter chart

Description automatically generated

﻿SCRIPT\_REAL("

import numpy as np

mean1 = np.mean([val for val in \_arg1 if val is not None])

mean2 = np.mean([val for val in \_arg2 if val is not None])

\_arg1 = [val if val is not None else mean1 for val in \_arg1]

\_arg2 = [val if val is not None else mean2 for val in \_arg2]

corrs = np.corrcoef(\_arg1, \_arg2)[0, 1]

return corrs

",

AVG([Price]), AVG([Review Scores Rating]))

Above code is used in calculated field to calculate Pearson correlation, also None values were replaced with mean value of argument.

## Price vs Review Score Correlation between Room Types and Years

Timeline

Description automatically generated

The heat map above illustrates the Pearson correlation between mean price and mean review score between Room Types in different years.

The code bellow was used to calculate the correlation.

﻿CORR(

{ INCLUDE [Name] : AVG([Price]) },

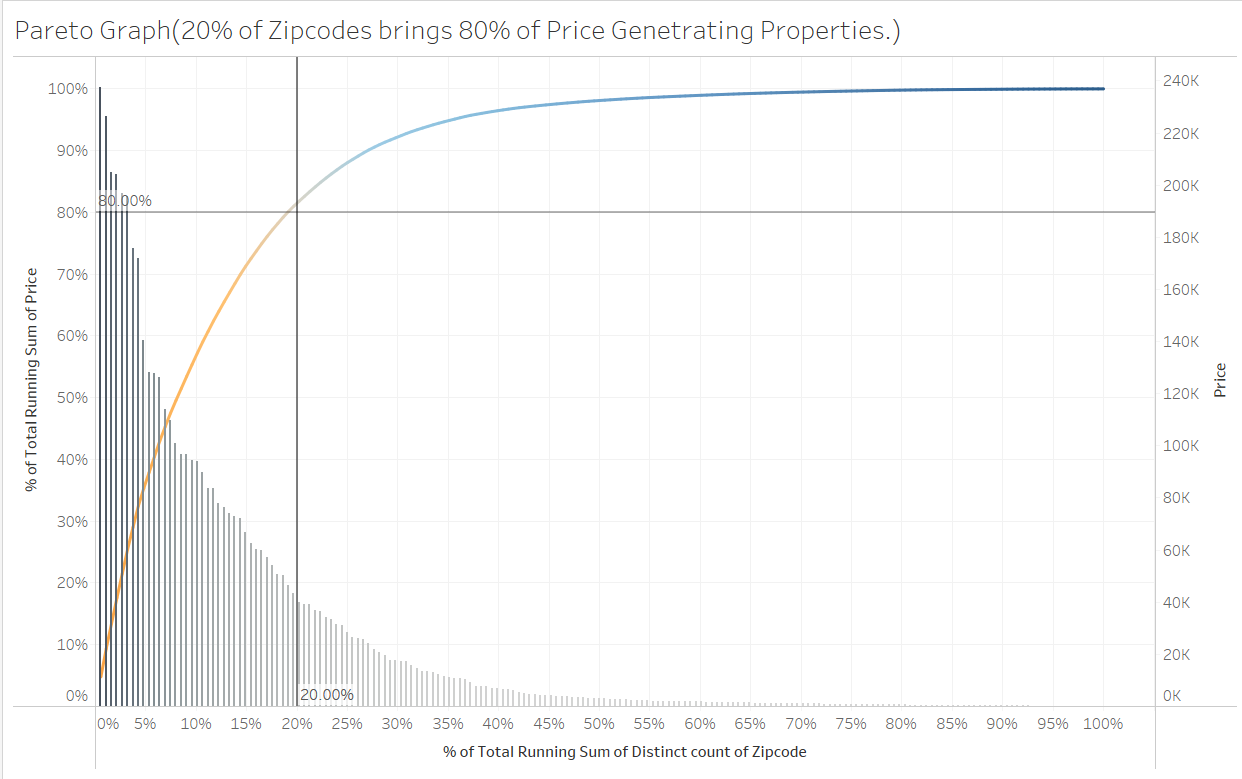
{ INCLUDE [Name] : AVG([Review Scores Rating]) }

)

## Insights

From previous visualizations, Price and Review Score is positively correlated, although in some cases are not, for example for Entire room/house rent in 2008. It means that the higher the price the better review the host gets. It might be because the properties with higher prices offer better conditions and quality. It is seen especially for private rooms every year shows positive correlation, the entire room/apt and shared room though in some years show zero or negative correlation. Outliers might have also affected the correlation in this way.

# Shanka Attanayake



## Insight:

In this visualization, it proves the pareto chart to be true, where 80% of all the price generating properties are within 20% of the zip Codes. This was completed by finding the Running Total of all the Zip Codes and finding the Percent of Total. With is information we can course more on how we can maximize those areas because they seem to be the most in demand areas within the dataset.

# Vrinda Parameswaran:

## Insights:

Using Pearson Correlation, I was able to check the correlation between different entities in the data. It was seen number of reviews vs price had correlation in compared to Review score rating vs price.

## Price vs Beds

Chart

Description automatically generated with medium confidence

## Number of reviews vs Price

Chart

Description automatically generated

Insights:

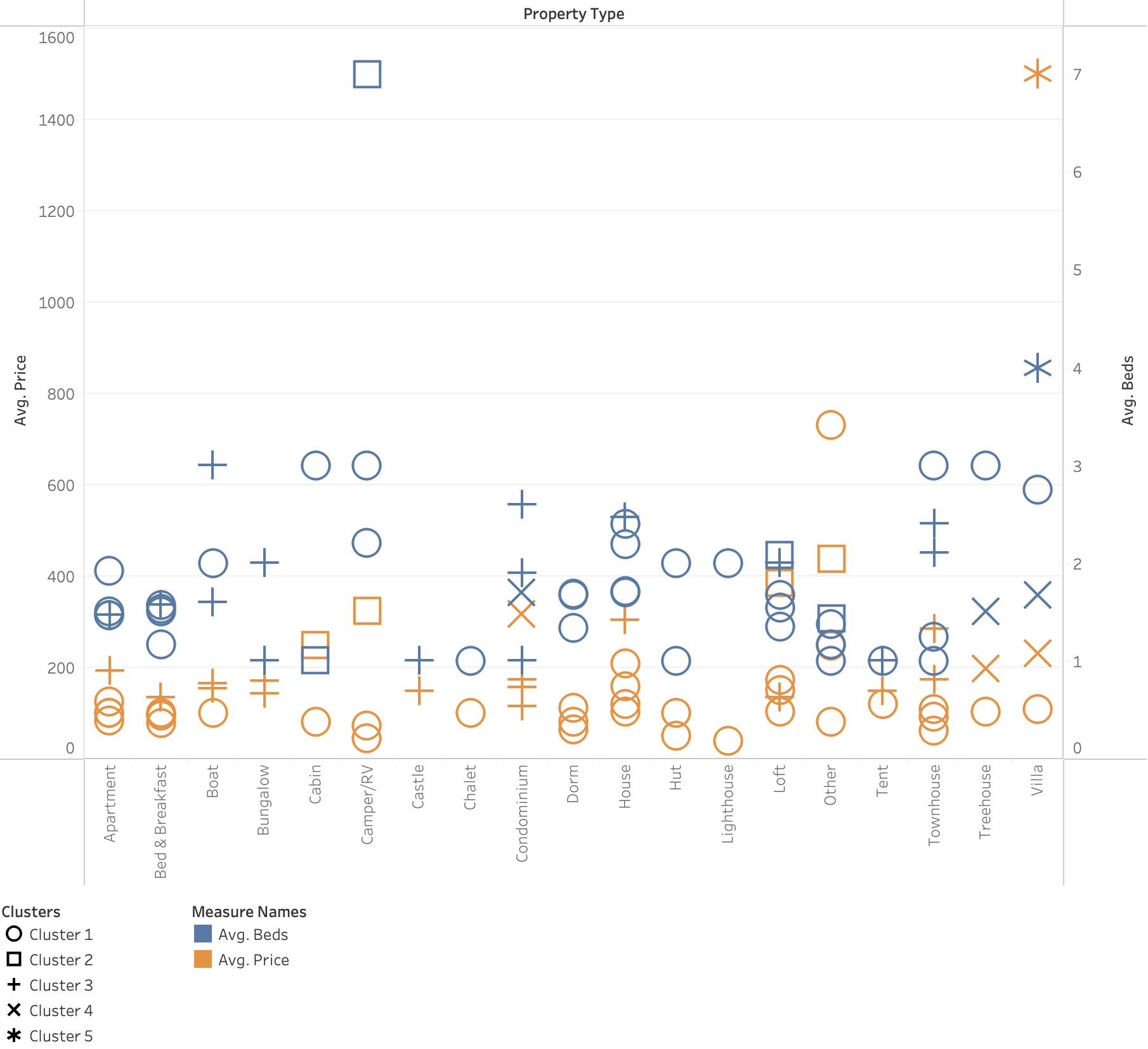
Manhattan had the greatest number of reviews. The Pearson correlation between Number of process vs price gave a particularly substantial value showing high correlation between the two.

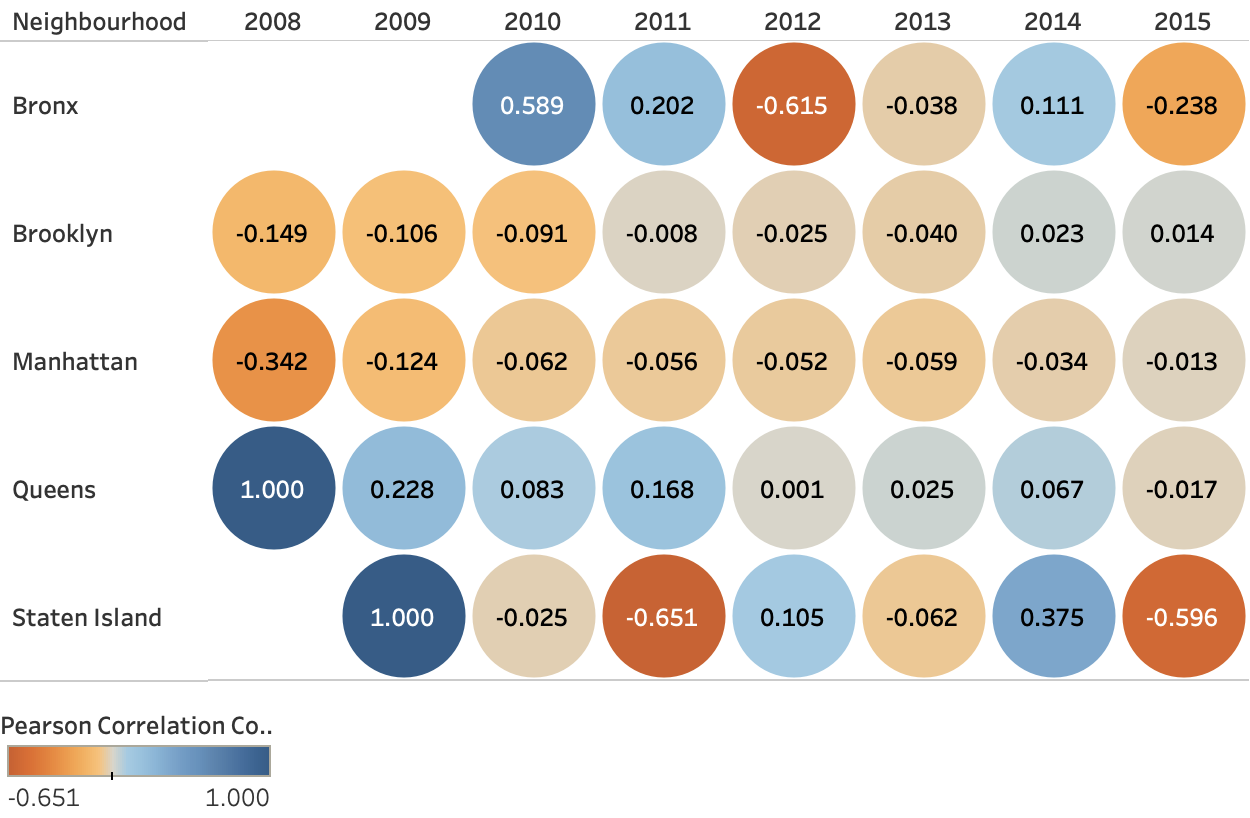
## Review Score Rating vs Price

Chart, scatter chart

Description automatically generated

# Tanvir Singh Ahuja:

Insights: Using clustering of various neighbourhoods, the data is used to determine which listings can generate more review based on the average number of bed and the average price of the property using dual axis function.



Insights: Using the pearson correlation matrix, the sum of number of reviews and rating score rating comparison, we can see which neighbourhoods are positively and negatively correlated. Queens is the only neighbourhood that is consistently being positively correlated.

Code:

CORR (

{INCLUDE ([Host Id]): AVG ([Number of Reviews])},

{INCLUDE ([Host Id]): AVG ([Review Scores Rating])}

)