**1. Description of Data**

The chosen dataset for this report contains information about Airbnb listings in New York City. For a bit of background, Airbnb is popular accommodation booking service that allows property owners to rent out real estate to others. Airbnb doesn’t own any of the property itself, but rather acts as a trusted middle-man to facilitate and manage the rental process.

Since almost anyone with extra living space can signup as a “host” and provide lodging to others, the website became quite popular. Airbnb was able to offer travelers a variety of accommodation options, generally at a lower price than conventional hotels.

Airbnb listings are public, which allows data to be easily extracted, compiled, and analyzed by other interested parties with little alteration and approximation of the original data. The data for this report comes from one such project called Airbnb Open Data and was hosted on the Kaggle platform. The link to the data can be found here: <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>. The data appears to be mostly raw information scraped from Airbnb’s public listings with a few exceptions.

1. To protect the privacy and safety of Airbnb hosts, the platform does not publicly share the exact addresses of listings but rather gives a general vicinity so guests will know roughly where they’re staying. Thus, the latitude and longitude given for each listing are taken from the center of the publicly shown vicinity and are approximations.
2. From the approximate latitude and longitude, each listing’s neighborhood and borough seem to have been extrapolated and included in the dataset. Luckily, deriving a neighborhood from given coordinates is a relatively simple task, so there is very little potential for error that needs to be considered.

With this data, the aim of subsequent reports exploring more advanced machine learning techniques will be to look for factors that affect the price of listings in New York City. Regression seems to be the most apt technique to be used here. The primary input variable will most likely be some sort of location-based information, like the listing neighborhood, borough, and/or exact coordinates. If none of these location-based attributes are sufficient, a possible solution would be to add new attributes that use the given latitude and longitude for each listing to calculate distance away from popular landmarks. These calculated distances could then be fed into the regression to provide more useful and varied results.

Listing popularity could be another useful input variable into a hypothetical future regression. Each listing has attributes pertaining to reviews by guests. Counting the number of reviews over a span of time should provide a decent analog to listing popularity – the more reviews the more people are staying at the listing and vice versa.

Exploring how these variables, and potentially others, effect the price of a listing would be a useful topic for further analysis through machine learning techniques.

**2. Explanation of Attributes**

Each observation has the following attributes:

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Attribute Name** | **Explanation** | **Type** | **Issues** |
| id | Listing unique identifier | Nominal | No |
| name | Name of listing | Nominal | No |
| host\_id | Host unique identifier | Nominal | No |
| host\_name | Name of host | Nominal | No |
| neighborhood \_group | Listing borough (Manhattan, Brooklyn, etc) | Nominal | No |
| neighborhood | Specific neighborhood of listing | Nominal | No |
| latitude | Approximated latitude of listing | Interval | No |
| longitude | Approximated longitude of listing | Interval | No |
| room\_type | Room type (Private, shared, etc.) | Nominal | No |
| price | Price per night | Ratio | No |
| minimum\_nights | Minimum number of nights in a stay | Discrete | No |
| number\_of\_reviews | Number of guests who’ve left reviews | Discrete | No |
| last\_review | Date of the last review | Interval | Yes |
| reviews\_per\_month | Average number of reviews per month | Ratio | Yes |
| host\_listings\_count | Number of listings the host owns | Discrete | No |
| availability\_365 | Number of days in the year listing is available | Discrete | No |

The attributes relating to the listing location, accommodation type, reviews, and price are of most interest. Neighborhood and neighborhood group will be useful as classifiers of varying specificity. Latitude and longitude can be used for very specific location-based regressions and potentially generating new attributes representing distance away from various points of interest.

Accommodation type should clearly affect price as shared rooms, though less common, are almost always cheaper than private rooms / apartments.

Review related attributes (Both reviews\_per\_month and last\_review) give a glimpse into apartment popularity. More frequent and recent reviews are indicative of more popularity and potentially higher prices. The average listing gets about 1.4 reviews per month.

Price is arguably the most import attribute. For most of our machine learning to be done later, the price will be the output variable that is used for training models.

Minimum\_nights, host\_listings\_count, and availability\_365 don’t seem to be of much interest now, but may prove to affect the price or other attributes of a listing.

The other attributes (IDs, names, etc) can be disregarded as they don’t offer much useful information.

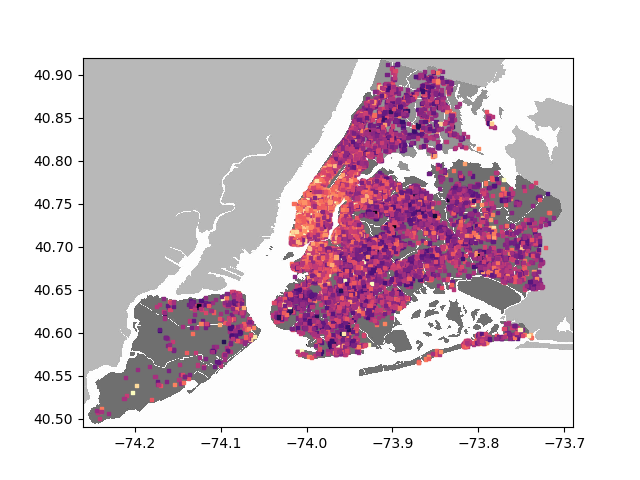
**3. Data Visualization**

The goal of the data visualization was to look for correlations between attributes that could potentially be explored further through further machine learning and statistical techniques. The first and most obvious potential correlation was the location of a neighborhood in relation to the price of a listing.

Not surprising, Manhattan is the most expensive borough, though more drastic differences in average price were expected. To explore this further, the number of listings in each borough was examined.

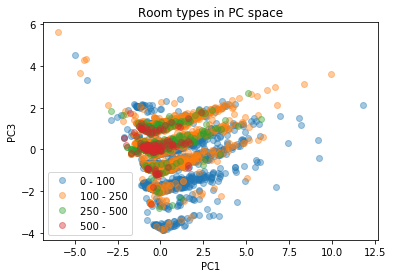
The above chart reveals useful information and introduces some potential problems. Since the majority of listings are concentrated in Brooklyn and Manhattan, the average prices for those boroughs are more representative and less likely to be affected by outliers. Also, because more data exists for Brooklyn and Manhattan, any analysis that examines correlations between neighborhood group and other variables is likely to be skewed.

Since borough seems to be too general of a classifier, an experiment with latitude and longitude was conducted, resulting in the following visualization.



In the above visualization, each dot represents the position of a listing in New York City. The color of the dot corresponds to its price on a logarithmic scale. Any price over $1000 per night is treated as $1000 to prevent outliers from disrupting the scale. Brighter, more orange colors correspond to higher prices.

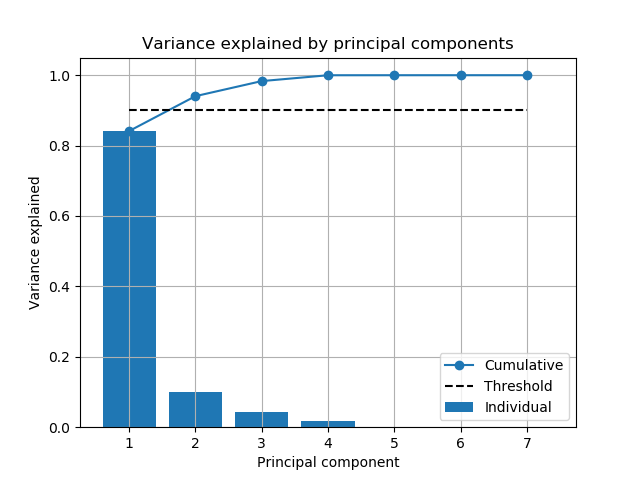
The visualization is better at conveying the relationship between location and price. Certain niche neighborhoods may contradict the average trend of their borough. For example, some of the smaller islands in Queens and the Bronx are significantly more valuable than the average property. By applying clustering and regression techniques on the coordinates of each listing, it should be possibly to derive more useful and specific correlations.

Lastly, to examine other possible correlations, Principle Component Analysis was done.

We can see in the figure that the points are generally clustered around 0 for both PC1 and PC3. While it appears in the figure that higher prices are more clustered around 0 for PC3, this is likely due to the paucity of higher-price data points; after all, generally speaking, the more points you have in a distribution, the greater the spread will be. A similar statement can ostensibly be made for PC1 and price.

The striking thing about this figure, though, is the "stripes" in the data, as well as the curve. The "curve" in this case refers to a kink in the graph around PC1 = 0, where each "stripe" switches from a negative PC1-PC3 correlation for PC1 < 0 to a positive PC1-PC3 correlation for PC1 > 0. This looks to be some function of an absolute value of PC1, implying that only the magnitude, and not the sign, of PC1 affects PC3.

The "stripes" refer to the clustering of data around a set of lines. In particular, at PC1 = 0, at each integer value of PC3 from -4 to 2, there is a cluster of data points, with more density near -4 to 2 and less density near -3.5, -2.5, ..., .5, 1.5. The PC3-value of these points then follows the aforementioned clustering around some function of the absolute value of PC1 as PC1 varies, and the points remain clustered as PC1 varies. This suggests the presence in the data of a quantized variable, as the data representing such a variable would tend to be clustered around discrete values. This is thus likely attributable to our labeling of neighborhoods using integers.



In the above figure, we see our seven principal components (PCs) sorted by how much variance in the data they explain. In a linear regression, the r^2 value of a term is the proportion of variance it explains, and the values displayed in this chart are the result of running a linear regression on our data using each of our seven PCs as independent variables. We see that PC1 is by far the most important, explaining over 80% of the variance in our data, whereas PCs 5-7 contribute only negligibly.

**4. Conclusions**

In conclusion, we found that there seems to be a strong and further examinable correlation between location and price. Further exploration using more detailed techniques should be possible in further projects.

In future PCA, we should find a better way to categorize boroughs rather than by assigning each a integer. This introduces a sense of scale into a nominal value which seems to have lead to the stripes exhibited in the first PCA diagram.

**5. Collaboration**

|  |  |  |  |
| --- | --- | --- | --- |
| Section # | Luca’s Contribution | Weston’s Contribution | Michael’s Contribution |
| 1 | 25% -- Discussed possible goals of the data analysis. | 50% -- Lead discussions of possible goals of the data analysis and typed up our ideas. | 25% -- Discussed possible goals of the data analysis. |
| 2 | 20% -- Discussed use of various attributes. | 40% -- Wrote the majority of the section. | 40% -- Discussed use of various attributes and assisted with producing the table |
| 3 | 50% -- Wrote the majority of the PCA code | 10% -- Created the first two charts and helped write the section. | 40% -- Wrote the code for the map diagram and helped write the section. |
| 4 | 33% -- Discussed and wrote up conclusions as a group. | 33% -- Discussed and wrote up conclusions as a group. | 33% -- Discussed and wrote up conclusions as a group. |