Capstone Project | Battle of the Neighbourhoods

Seattle's Housing Market & Airbnb

See blog post on Medium

Contents

Introduction

Background

Seattle has changed significantly after Amazon moved into its South Lake Union headquarters in 2010. Prices and rents skyrocketed, sending the city into a housing downward spiral. This trend is starting to shift and taper off in the past two years due to a surge in new property building and slow-down in rent growth. Nevertheless, housing remains an issue. A combination of factors shapes the state of the housing market in Seattle and other big cities across the world: starting from land availability (constrained by geography and urban planning), housing subsidies and taxes; increasingly exorbitant rent market prices, housing supply, mortgage interest rates, construction costs, as well as internal and international migration. Supply is also increasingly affected by demand by foreign investors and private buyers (primarily from China). Last but not least, the growth of services like Airbnb affect supply by taking entire properties off the conventional rental market.

Problem

Renting a place to live in major developed cities like Seattle is increasingly difficult. It is out of the scope of this capstone project to analyse all variables affecting rental prices, so at this first stage, I will focus on claims that Airbnb listings exacerbate the housing crisis by "squandering precious long-term rental housing stock". There has not been a proper look comparing entire place listings (which do take properties out of the long-term rental market) and single room listings which people rent out on Airbnb to supplement their income while they still live on the property. A higher number of listings corresponding to higher rental prices in a neighbourhood might be due to other factors besides supply shortage. For example, investors and private owners of properties in expensive neighbourhoods might be more tempted to list one or more of them on Airbnb. It is interesting to check whether single room vs entire place Airbnb clusters would correspond to different rental price clusters across neighbourhoods.

Interest

Housing issues and soaring rental prices are increasingly becoming a problem across developed cities around the world. Open data projects and data insights on the topic can be used to inform public policy or generate productive debate on the topic.

Data

Sources, Acquisition & Cleaning

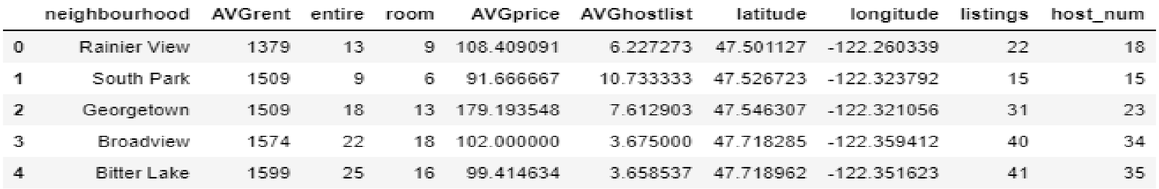
- Airbnb & GeoData: Inside Airbnb is an independent, non-commercial Open Source data tool which provides Airbnb listings data to the public. I used it to download a .csv with current Airbnb listings in Seattle (updated in September 2019). **It also features geo coordinates and neighbourhoods**, which I supplemented with a geojson file from SeattleIO on github.
- The most recent rental data was collected from RentCafe

Data Cleaning & Feature Selection

Since the rent database has a slightly different spelling of columns I had to rename the Neighborhood column, so that I could better append data to the airbnb listings. The Airbnb data included features which are not relevant to the current analysis, so I had to drop the columns (listing) name, host_name, minimum_nights, number_of_reviews, last_review, reviews_per_month and availability_365.

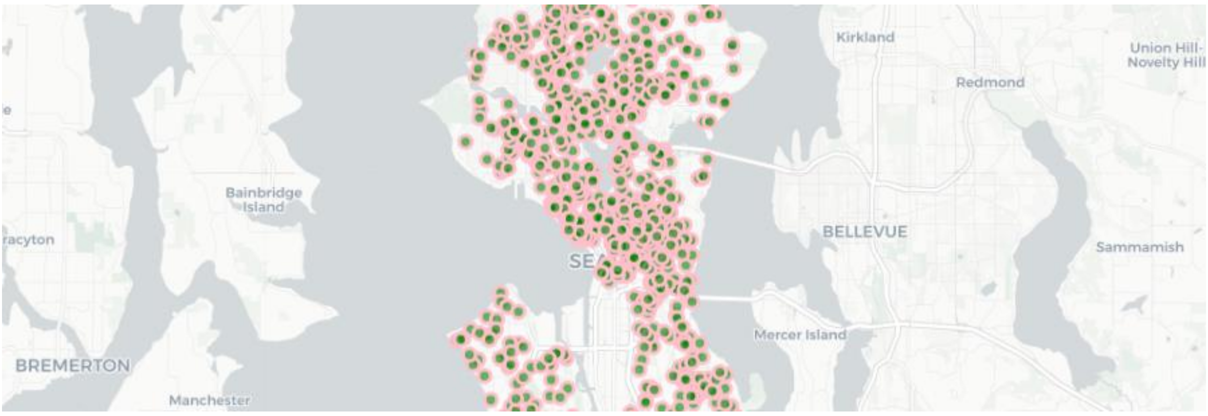
The remaining features neighbourhood_group, neighbourhood, latitude, longitude, room_type and price are directly relevant to analysing whether entire place or single room listings are clustered in neighbourhoods with higher rental prices. What is more, the columns Host_id and calculated_host_listings_count indicate property owners with multiple listed properties: a variable related to the notion that the relationship between Airbnb listings and neighbourhood prices going up goes two ways (and owners with higher income and more properties are likely list property in neighbourhoods where they can get more money out of them).

In the end I had data on the number of entire place and single room listings per neighbourhood, as well as the average listing price, average number of hosts, and average rent and number of entire place or room (single/shared) only listings. After removing data for 7 neighbourhoods for which there was no available average rent I was able to look at information about 81 out of 88 neighbourhoods in the Airbnb listing database for Seattle.



	neighbourhood	AVGrent	entire	room	AVGprice	AVGhostlist	latitude	longitude	listings	host_num
0	Rainier View	1379	13	9	108.409091	6.227273	47.501127	-122.260339	22	18
1	South Park	1509	9	6	91.666667	10.733333	47.526723	-122.323792	15	15
2	Georgetown	1509	18	13	179.193548	7.612903	47.546307	-122.321056	31	23
3	Broadview	1574	22	18	102.000000	3.675000	47.718285	-122.359412	40	34
4	Bitter Lake	1599	25	16	99.414634	3.658537	47.718962	-122.351623	41	35

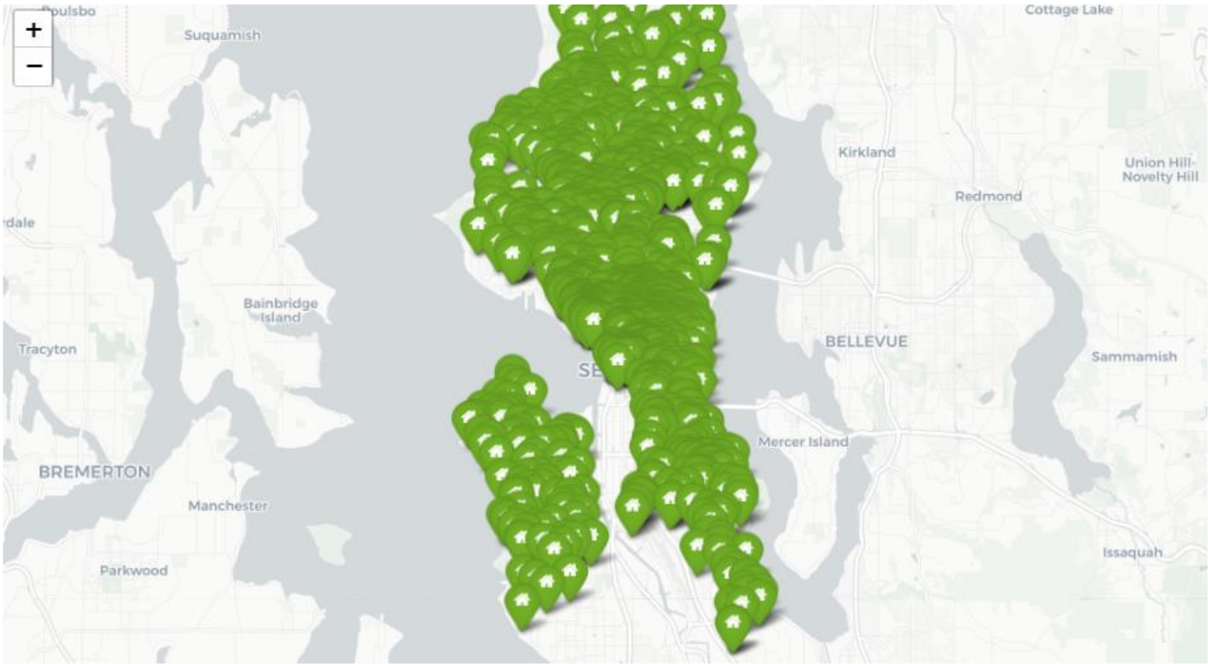
Geo Data | Map of Seattle



To map the analysed data we will first need a Folium map of the city, based on Geographic coordinates.

Geo Data | Mapping Listings

Using the longitude and latitude data in the Airbnb listings database and a geojson file I can map listings by room type, placing icons on the map markers

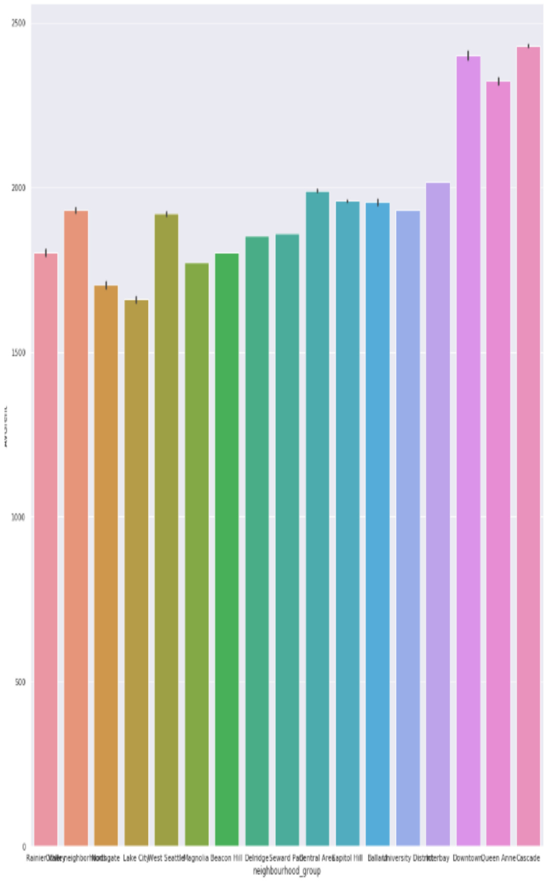


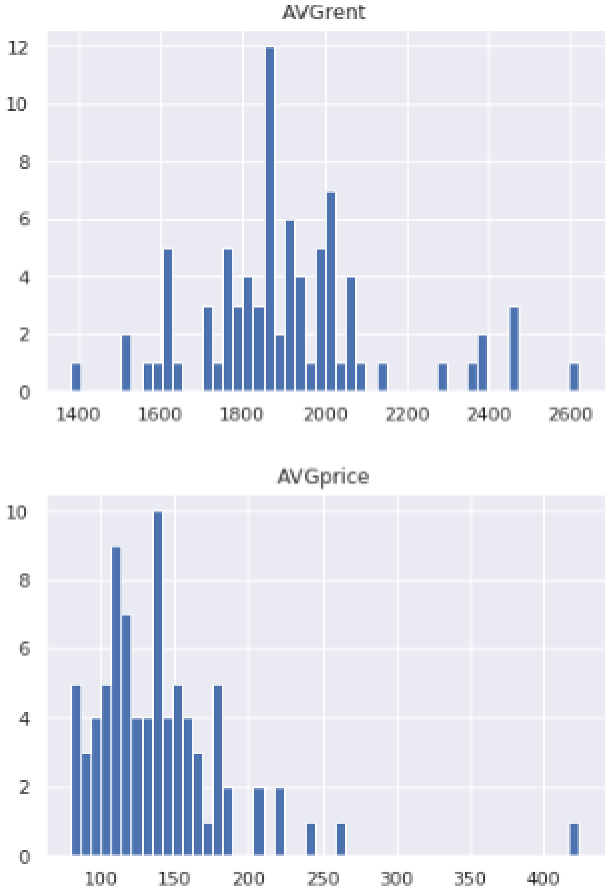
Methodology & Data Analysis

Exploratory Data Analysis (EDA)

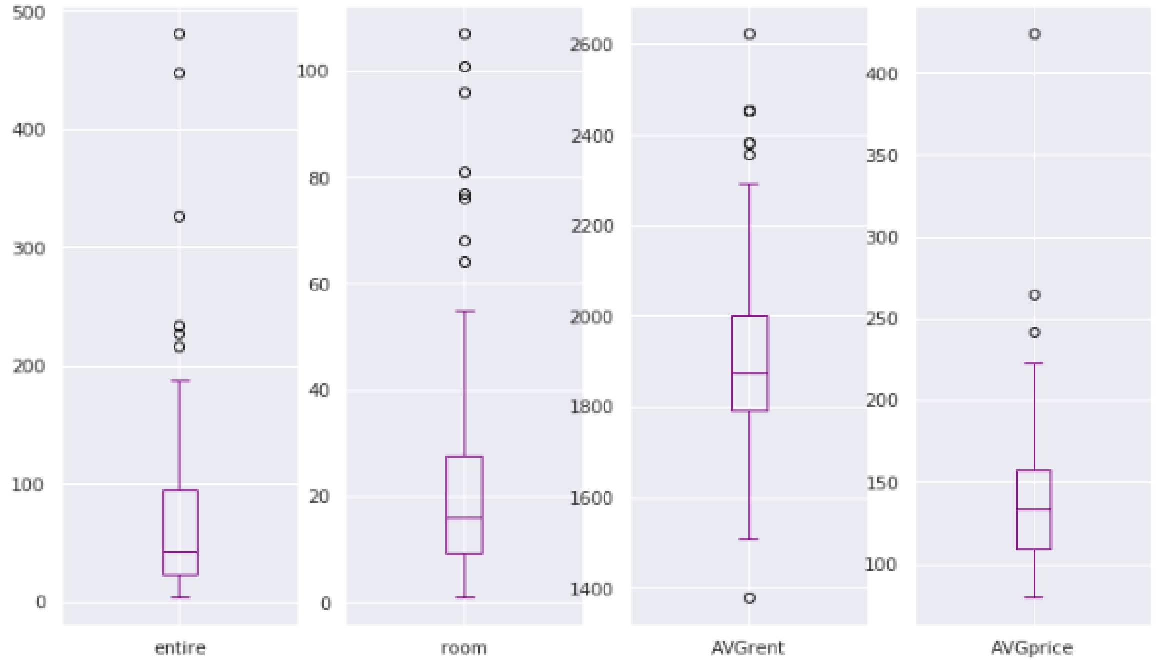
First I looked at my data and visualised some of it to get an initial idea about the data through EDA. After this initial step was completed the planned analysis looked at several types of clustering techniques: KNN (supervised), CBSCAN (unsupervised) to see whether reliable clusters can be formed based on Airbnb listings characteristics and neighbourhood average rent prices.

Bar charts & Histograms for getting a basic idea of frequency and distribution within data categories.

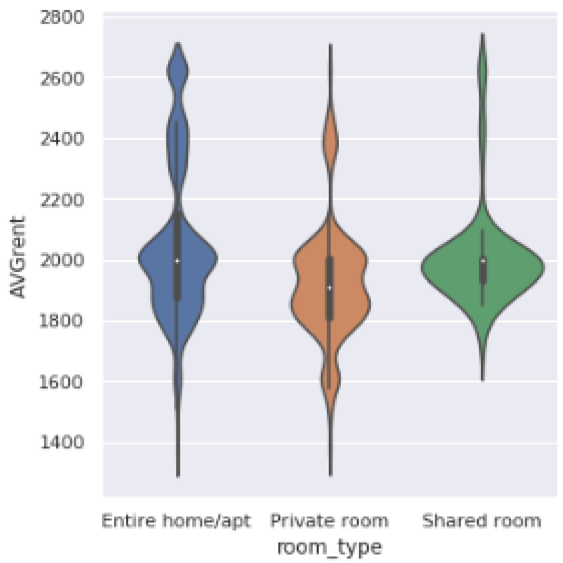




Box plots are useful for understanding variability in the distribution across the categories I was interested in. There are a lot of outliers across the categories, as well as a wide range between max and min values and mostly skewed rather than symmetric distributions.

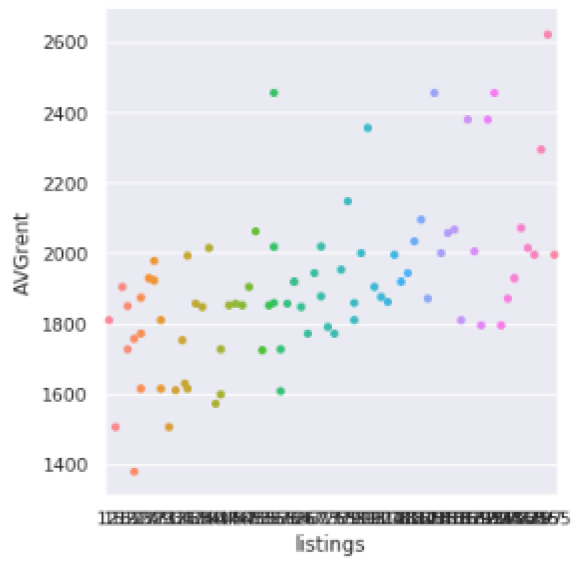


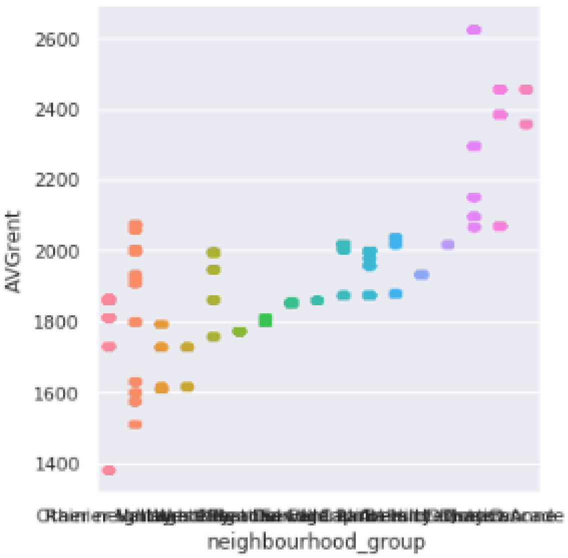
Violin plots are a variation of box plots using the kernel density estimate to provide a richer description of the distribution.



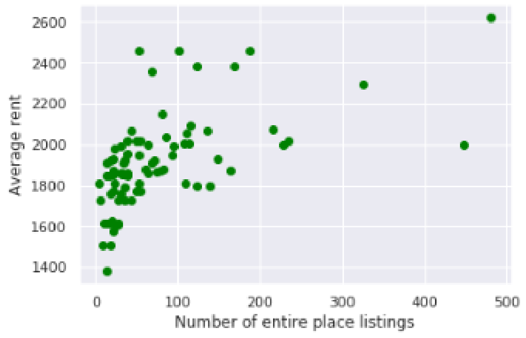
Scatterplots: before running any regression analysis it is useful to look at the relationship between the sets of dependent and independent variables I am interested in through scatterplots

Since I hypothesised a possible relationship between the different types of listings and the average rent in each neighbourhood, I created scatterplots looking at different pairs of data (rent by type of room (entire/single) and by average listing price for the neighbourhood. The scatterplots below contain information about the Pearson r correlation test as well.

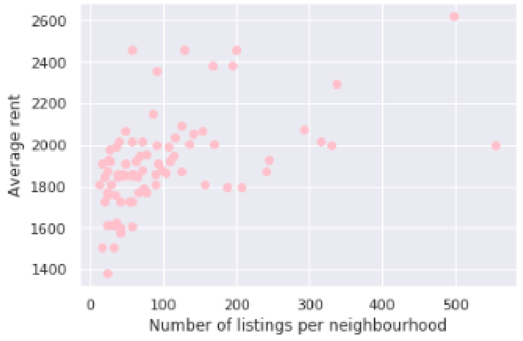




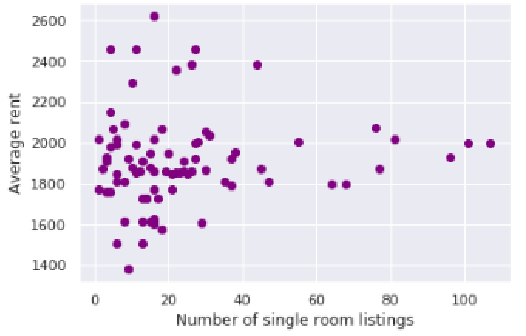
There is a moderate correlation between the number of listings per neighbourhood and the average rent price (.51) and between entire place listings and average rent price (.53). As predicted, the correlation between entire place listings and rent is higher than the one between the number of single room listings and neighbourhood rent.



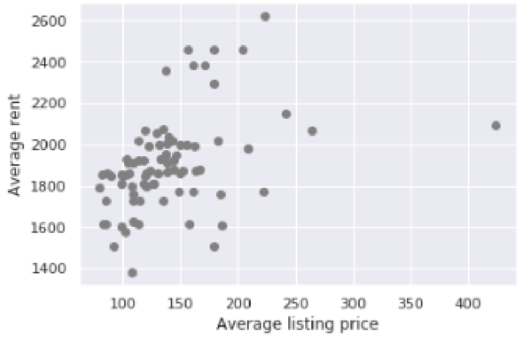
```
0.5733944895604494
```



```
0.511366820678075
```



```
0.10044246358025356
```



```
0.4247357629275412
```

I then wanted to build a regression model to investigate the relationship between a the dependent variable (rent) and the predictor (listing type/density). The dataset was divided into two mutually exclusive sets for training (80%) and testing (20%) to provide a more accurate evaluation of out-of-sample accuracy. Based on the scatterplots above, a polynomial regression model would be more suitable for the current data.

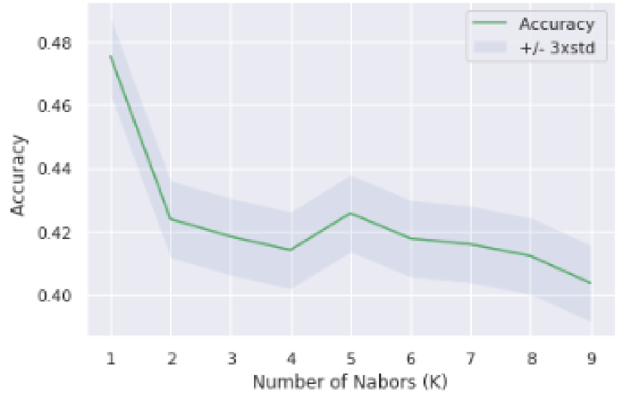
```
Coefficients:  [[0.          2.74319388 -0.00344486]]
Intercept:  [1706.04574651]
Mean absolute error: 145.79
Residual sum of squares (MSE): 35359.04
R2-score: -1.81

Train set: (6520, 2) (6520,)
Test set: (1630, 2) (1630,)

The Jaccard Similarity Score is
0.06993865030674846
```

Supervised clustering methods | KNN Clustering

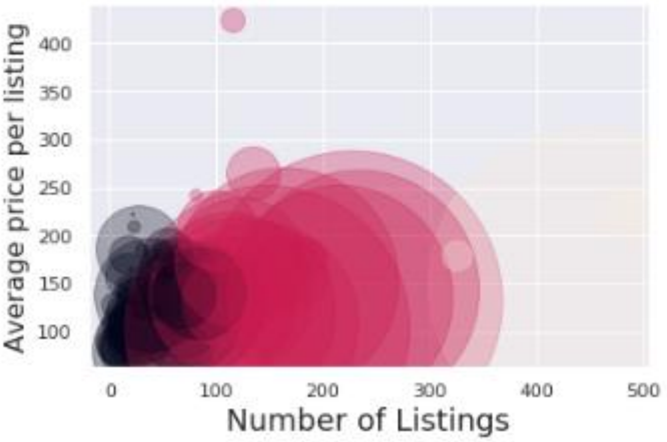
The K-Nearest Neighbors algorithm is a classification algorithm that takes several labeled points and uses them to learn how to label other points.

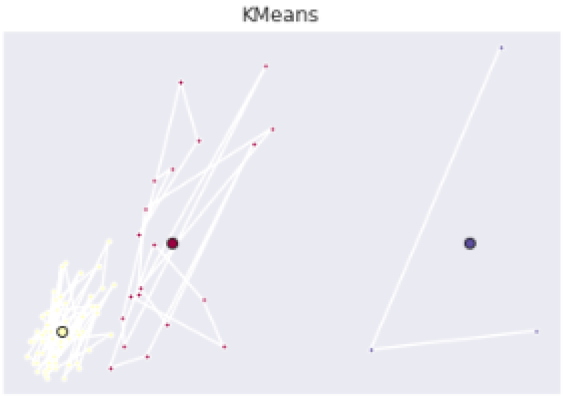
- cases are classified based on their similarity to other cases.
- data points that are near each other are said to be neighbors
- similar cases with the same class labels are near each other.

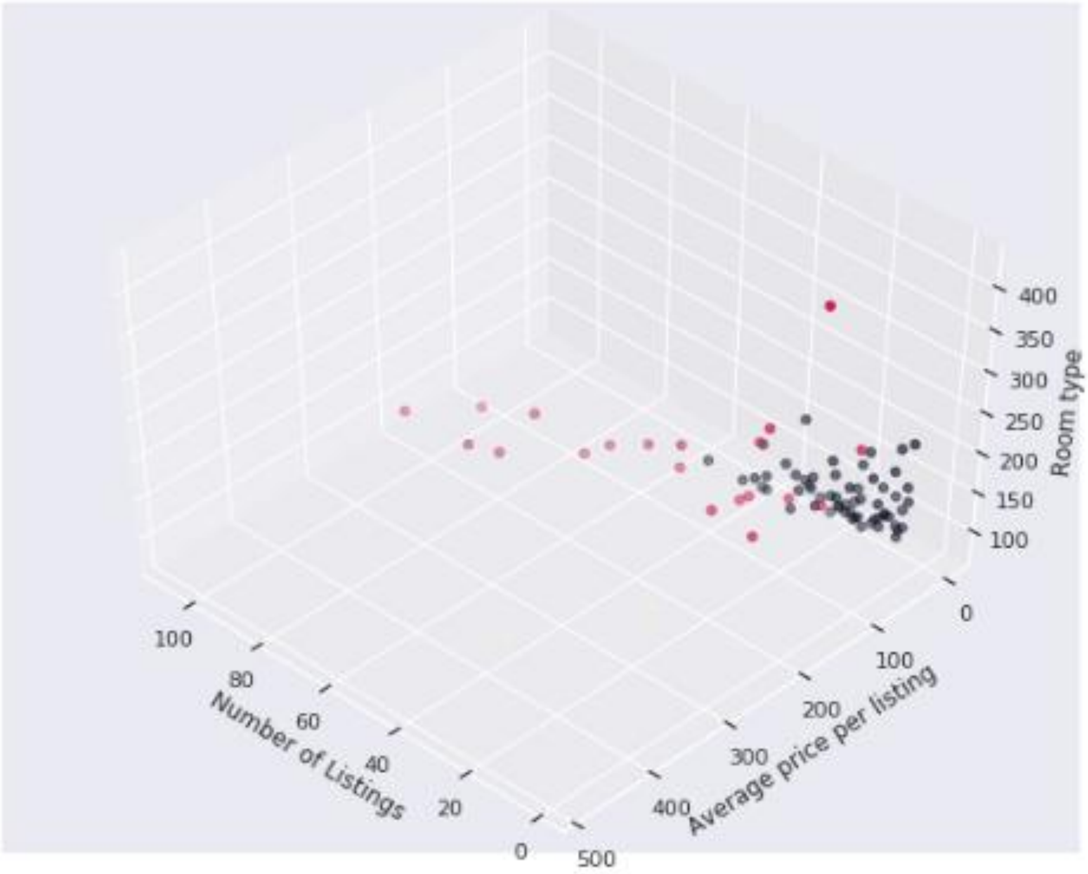
```
Train set Accuracy:  0.6041411042944785
Test set Accuracy:  0.42576687116564416
Array ([0.47546012, 0.42392638, 0.41840491, 0.41411043, 0.42576687,
       0.41779141, 0.41595092, 0.41226994, 0.40368098])

The best accuracy was with 0.4754601226993865 with k= 1
```



The best accuracy was with 0.4754601226993865 with k= 1







Unsupervised clustering methods | k- Means

K-means can group unsupervised data; type of partitioning clustering producing sphere-like clusters relatively efficient in Medium to Large size datasets

Objectives of K-means:

- To form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters.
- To minimize the “intra cluster” distances and maximize the “inter-cluster” distances.
- To divide the data into non-overlapping clusters without any cluster-internal structure

Results & Discussion

The claim that Airbnb listings have an impact on increasing property rents is increasingly popular in media coverage in recent years. However, testing the available datasets for Seattle Airbnb listings and average rent showed there is only a moderate correlation between the number of entire houses/apartments taken off the market and listed on Airbnb and the likelihood of that neighbourhood having higher average rent. Attempts to build regression models and cluster neighbourhoods by number of listings, type of listing or average listing price and average rent in the neighbourhood did not indicate a robust detect a robust relationship between the predictor variables and the dependent variable (average rent per neighbourhood).

Finding no evidence that Airbnb listings impact rents **is not the same thing as finding evidence that they DO NOT impact rent**. It is, however, a good reminder that all claims should be empirically tested before deciding decide whether they are reliable or not. At this point, Seattle data seems to indicate that the impact of Airbnb listings is not as serious as some pundits claim.

Future research

Further analysis of similar Airbnb listing datasets can continue to examine the relationship between listings and rent property availability and average rent. Using datasets from other cities and comparing country-level and regional data can shed further light on the topic. What is more, further investigation should also take into account other factors influencing the market prices of properties and check what is their relative influence compared to that of Airbnb listings.

Limitations This is a debut Python DataSci project, so it is possible that many components of its method and code would have to undergo further development.

Conclusion

Intuitions about causal relationships must always be checked against empirical data. Open databases and the use of data science techniques and python libraries are especially helpful in providing tools to analyse data and test various hypotheses and predictions. The current report is an example of how interesting everyday issues (high rental prices and demand for both long term rentals and holiday stays on Airbnb) can be examined through openly available data.