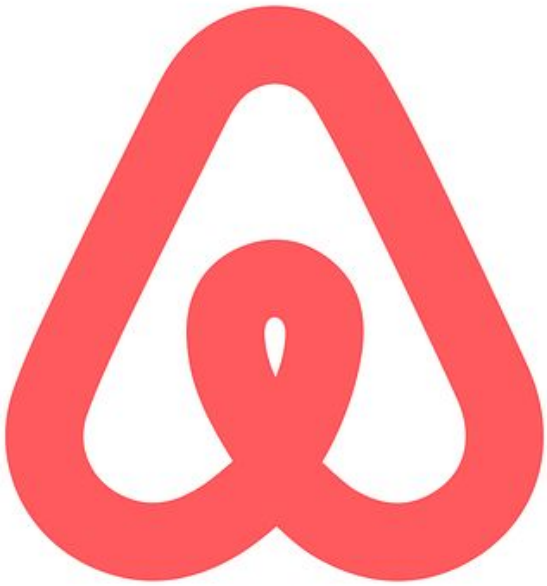

COMP5310 Assignment 1

Sydney AirBnB Analysis

Name: Nahian-Al Hasan
Student ID: 460430331
Unikey: nhas9102

Problem

Airbnb is an online marketplace that connects people who want to rent out their homes with people who are looking for accommodations in that locale. It currently covers more than 81,000 cities and 191 countries worldwide [1]. Since its inception in 2008, AirBnB has seen global success by disrupting the hospitality industry as more and more travelers decide to use AirBnB as their primary accommodation provider. In Australia, no city is more popular than Sydney in terms of the number of AirBnB listings. According to [2], there was a peak of approximately 55,000 AirBnB listings in New South Wales alone at the beginning of 2019. In [3], a report showed that Sydney hosts had an average monthly income of AUD 2,965 in 2019. Therefore, it might not seem like such a bad prospect to rent out one's empty apartment in Sydney.

However, it is quite difficult for hosts to assess the true value and the demand of their homes. This project aims to discover insights regarding the AirBnB rental landscape in Sydney and answer the following questions - **(a)** *Which factors influence how in-demand an AirBnB accommodation is in Sydney?* **(b)** *Can we determine a fairly spot-on daily price for a new accommodation that fits into its specific market environment and competitors in Sydney?* **(c)** *Is it possible to make a recommendation to new hosts about the presentation of their accommodation based on existing listing data and reviews?*

This analysis is expected to help aspiring Sydney AirBnB hosts to evaluate the value and opportunity costs of listing a new accommodation, and to provide recommendations to optimise price matching and profit.

Data

The Sydney AirBnB dataset consists of multiple categories of data. The data has been divided into three different categories: **(a)** *Reviews* **(b)** *Listings* **(c)** *Calendar Entries*. The most relevant and meaningful data are the reviews and listings data. The data that is being used was last updated on 16th March, 2020 and can be located at the URL, Inside Airbnb. Adding data to the debate [4]. It contains information about all listings and reviews of accommodations in Sydney from 2015 to present, and was compiled by Murray Cox, an independent digital storyteller and technologist [5]. In addition to the AirBnB dataset, I have also extracted all NSW transport service locations from OpenData, which can be found at the URL, Public Transport - Location Facilities and Operators [6].

All datasets were formatted as comma separated value (csv) files. The listings data set was *143.2 MiB* in size and contained *39,670* entries and *106* columns. The columns of interest used to perform the exploratory data analysis were *price, cleaning fee, neighbourhood, latitude, longitude, property type, capacity, bathrooms, bedrooms* and *superhost status.* The review data set was *168.5 MiB* in size and contained *642,498* entries and *6* columns. The columns of note were *listing_id* and *review*. Lastly, the transport service location data set was *222.1 KiB* in size and contained *791* entries and *11* columns, of which the *latitude* and *longitude* columns were of most interest.

This analysis was performed using the *python* programming language and compiled in a couple of *Jupyter Notebooks*. A multitude of libraries were used to perform the analysis, and are listed in detail in **Appendix A**. The datasets were cleaned using standard approaches, i.e. dropping duplicates, and selectively dropping rows where nulls would be a problem, e.g. *listing_id*. The *price* and *cleaning* columns in the listing dataset were in string format, and thus converted to *int64*. Empty rows in the *cleaning fee* column were treated as AUD 0. The boxplot of prices (**Appendix B**) showed there is quite a skewed distribution with a long tail of high-priced outliers. However, 75% of all rentals only cost upto AUD 200. For this project, extremely high priced rentals above AUD 500/night were removed to maintain comparability. Similarly, 75% of listings were within 3 km of a transport service. For that reason, listings which did not have transport services within a 10 km radius were also discarded. In order to preserve recency and relevancy, a sample of data from Oct, 2018 to Nov, 2019 was used to perform the analysis. Data from late 2019 and 2020 was excluded due to the travel impacts of the Australian bushfires and COVID-19 travel ban. Owing to sampling only a year of data, the *reviews per month* of each listing was manually feature engineered using the reviews data set, and then merged with the listings data. Additionally, a combination of *pymongo* and *MongoDB Atlas* database was used to engineer proximities of each listing from its nearest transport service using *MongoDB's geospatial indexing* capabilities (**Appendix C**), and merged with the rest of the data as the *proximity* column, which contained the distance in metres of each listing to its nearest train station, bus stop or wharf. A visualisation of this can be viewed in **Appendix D**. Lastly, and most importantly, it was quite essential to calculate the *occupancy*

rate[1] of each listing - as the *occupancy rate* can be used to approximate the *monthly income* of each listing. The methodology for deriving the *occupancy rate* used a variant of the San Francisco model [7] to estimate occupancy rates. **Appendix E** describes and visualises how a modest and optimistic occupancy rate was estimated using my approach.

As a starting point, a correlation heatmap was produced to understand how each feature influenced one another (**Appendix F**). The heat map produced some unexpected findings such as - *cleaning fees* being highly correlated with income, but more surprisingly - the *bedroom, bathroom* counts and *capacity* were also highly correlated with monthly income. But unfortunately, the proximity to transport services didn't seem to indicate any correlation. A seasonal demand is to be expected when dealing with AirBnB accommodations. The time-series graph produced for average reviews per month (**Appendix G**) clearly shows that there is a very clear seasonal demand for listings in Sydney. Demand appears to peak once early in the year during March, and then later again during November. This might primarily be due to a climate factor - where temperatures during March and November are much amenable than the rest of the year; as well as New Year's Eve celebration and Christmas. It was observed that 28% of all hosts were *super hosts*[2]. A bar chart produced for monthly income by area and host status (**Appendix H**) showed that super hosts were earning heaps more than normal hosts - and in some neighbourhoods, even 40% higher. A couple of lollipop charts were generated to visualise the average occupancy estimate in % by district (**Appendix I**) and the average monthly income in % by district (**Appendix J**). It was observed that areas such as Auburn, Penrith and Fairfield have relatively high occupancy rates. You would expect Sydney CBD (Sydney) to have a high occupancy rate, but the districts mentioned above are quite surprising to be at the top! However, when measuring income estimates per region, Sydney CBD, and many of the areas close to beaches or attraction points rank higher. A closer analysis on occupancy estimates and income grouped by proximity per district (**Appendix K**) showed a more predictable trend. Incomes from listings closer to transport services in city areas such as Sydney CBD, Hurstville, Paramatta and Campbelltown generate more income. Whereas for listings located near attraction spots, the trend appears to be the opposite, as the more profitable listings tend to be closer to attraction spots and farther away from transport services. And lastly, a bar chart produced for average occupancy by capacity (**Appendix L**) showed that bigger homes seem to sell more often than smaller ones. This might be because a larger group of people tend to be able to afford more than 1-2 people. Therefore, it follows that accommodations with a bigger capacity enjoy greater popularity. It follows that from a prediction point of view, features such as capacity, cleaning fee, neighbourhood, proximity to transport services and being a super host are incredibly valuable indicators. Whereas from a recommendation point of view, seasonal demand, reviews and descriptions would be great assets as well.

Proposal

For stage 2 of the project, the goal is to further refine current findings and attempt possible solutions for the research questions proposed in the Problem section of the report. The next steps include but are not limited to:

a) Performing a deeper analysis on categorical data, e.g. *presence of profile picture, amenities* and *description* to find more insights on occupancy rate influencers.

b) Perform a text-based analysis on descriptions and reviews to extract insights on popularity metrics of listings.

c) Depending on the outcome of further exploratory analysis, conduct a feasibility study of which features to use to produce a statistical or a machine learning model.

d) Produce a statistical or machine learning model to predict prices of a new listing based on the proposed features.

e) Propose recommendations based on popular listing reviews and descriptions on how to present the new accommodation to travellers.

Some proposed models for producing the Machine Learning model to predict price include Random Forests, XGBoost and Neural Networks. A statistical study will be conducted to evaluate the accuracy, precision, recall and f1 scores of each model through a stratified cross-validation approach. Lastly, a statistical significance test will be conducted for each variant to evaluate which model performed best.

[1] **occupancy rate** - ratio of rented or used space to the total amount of available space

[2] **super host -** experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests [8]

Bibliography

[1] “Airbnb: Advantages and Disadvantages,” Investopedia, 2020. [Online]. Available: https://www.investopedia.com/articles/personal-finance/032814/pros-and-cons-using-airbnb.asp. [Accessed: 11-Apr-2020].
[2] Radoslaw Panczak and T. Sigler, “Ever wondered how many Airbnbs Australia has and where they all are? We have the answers,” The Conversation, 12-Feb-2020. [Online]. Available: https://theconversation.com/ever-wondered-how-many-airbnbs-australia-has-and-where-they-all-are-we-have-the-answers-129003. [Accessed: 11-Apr-2020].
[3] The Flawed Consumer, “How much can you earn from Airbnb in Australia? - The Flawed Consumer,” The Flawed Consumer, 15-Mar-2019. [Online]. Available: https://www.theflawedconsumer.com/how-much-money-can-you-earn-from-airbnb-in-australia/. [Accessed: 11-Apr-2020].
[4] “Inside Airbnb. Adding data to the debate.,” Inside Airbnb, 2020. [Online]. Available: http://insideairbnb.com/get-the-data.html. [Accessed: 11-Apr-2020].
[5] “Inside Airbnb. Adding data to the debate.,” Inside Airbnb, 2014. [Online]. Available: http://insideairbnb.com/behind.html. [Accessed: 11-Apr-2020].
[6] “Public Transport - Location Facilities and Operators | TfNSW Open Data Hub and Developer Portal,” Nsw.gov.au, 2020. [Online]. Available: https://opendata.transport.nsw.gov.au/dataset/public-transport-location-facilities-and-operators. [Accessed: 11-Apr-2020].
[7] “Inside Airbnb. Adding data to the debate.,” Inside Airbnb, 2015. [Online]. Available: http://insideairbnb.com/about.html. [Accessed: 11-Apr-2020].
[8] “What is a Superhost? | Airbnb Help Centre,” Airbnb.com.au, 2020. [Online]. Available: https://www.airbnb.com.au/help/article/828/what-is-a-superhost. [Accessed: 11-Apr-2020].

Appendix A - List of libraries

List of libraries used to perform the analysis:

```
folium==0.10.1
ipykernel==5.1.4
ipython==7.13.0
ipython-genutils==0.2.0
ipywidgets==7.5.1
jupyter==1.0.0
jupyter-client==6.0.0
jupyter-console==6.1.0
jupyter-core==4.6.3
matplotlib==3.2.0
numpy==1.18.1
pandas==1.0.3
pymongo==3.10.1
python-dotenv==0.12.0
scipy==1.4.1
seaborn==0.10.0
```

Appendix B - Boxplot of prices

Figure 1: Boxplot of prices in $

Appendix C - Code to calculate proximities of each listing to nearest transport

```python
def find_distance_to_nearest_transport(long, lat):
    print('long', long, 'lat', lat)
    pipeline = [
        {'$geoNear':{
            'near': { 'type': "Point", 'coordinates': [ long, lat ] },
            'distanceField': "distance",
            'spherical' : True }
        },
        { "$sort": { "distance": 1 } },
        { '$limit': 1 }
    ]

    result = list(transports.aggregate(pipeline))
    if len(result):
        return float(result[0]['distance'])
    else:
        raise Error('DB did not return result')
```

```python
# Make new column called proximity

df_1['proximity'] = df_1.apply(
    lambda row: find_distance_to_nearest_transport(
        row.longitude,
        row.latitude
    ),
    axis=1
)

df_1.head(10)
```

Figure 2: Python and MongoDB query to calculate proximity

Appendix D - Visualisation of accommodations by proximity and occupancy

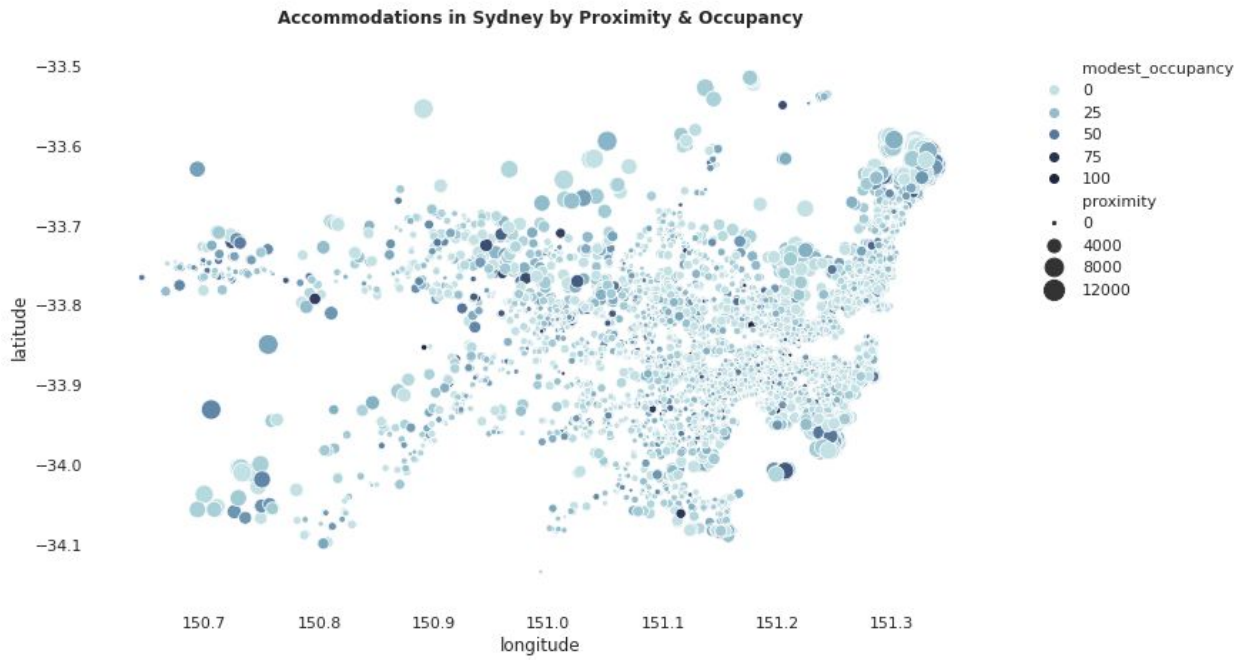

Figure 3: Accommodations in Sydney by Proximity in metres & Occupancy Rate in %

Appendix E - Occupancy Rate & The San Francisco Model

One of the biggest issues with Airbnb is getting the occupancy rate for each host or for a market. Inside Airbnb, the website I sourced the data from, uses an occupancy model which they call the "San Francisco Model" with the following methodology:

1. A **Review Rate** of 50% is used to convert reviews to estimated bookings. Other administrative authorities are said to use a review rate of 72% (however this may be attributed to an unreliable source: Airbnb's CEO and co-founder Brian Chesky) - or one of 30.5% (based on comparing public data of reviews to the The New York Attorney General's report on Airbnb released in October 2014.) Inside Airbnb chose 50% as it sits almost exactly between 72% and 30.5%. It basically means that only 50% of all visitors write a review. With that said, the number of reviews per month divided by the review rate equals an estimate of actual visitors.
2. An **average length** of stay for each city is usually published by Airbnb. This number multiplied by the estimated bookings for each listing over a period of time gives the occupancy rate.
3. Finally, the **income** can be calculated by multiplying the occupancy rate by the price and the time period of interest - here, 12 months:

> ***Monthly Occupancy Rate = Average Length of Stay * (No. of reviews per Month/Review Rate)***
>
> *According to the latest Business Insider Update, the average length of stay in Sydney is 5 nights*
>
> ***Yearly Income = Monthly Occupancy Rate * Price * 12 Months***

4. For the modest estimate, I use a review rate of **50%**, and for an optimistic estimate, I use a review rate of **40%**. This gives us an estimate for the average yearly income, which can be viewed in *Fig 4*.

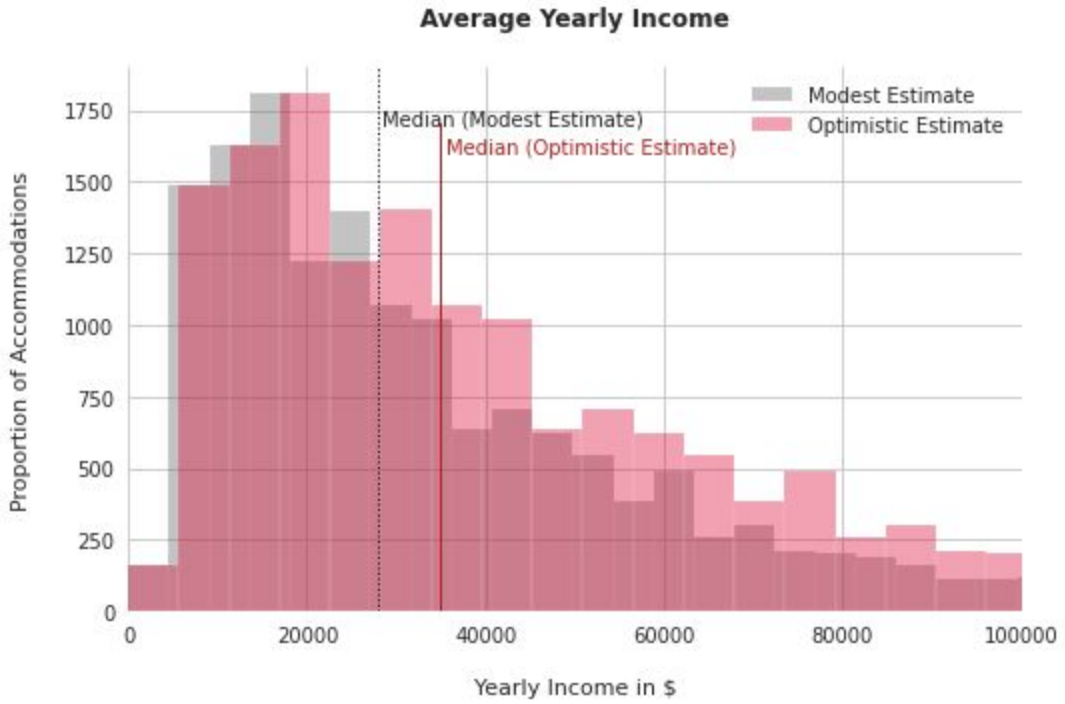

Figure 4: Average yearly income in $ of hosts in Sydney

Appendix F - Correlation Heat Map of metrics

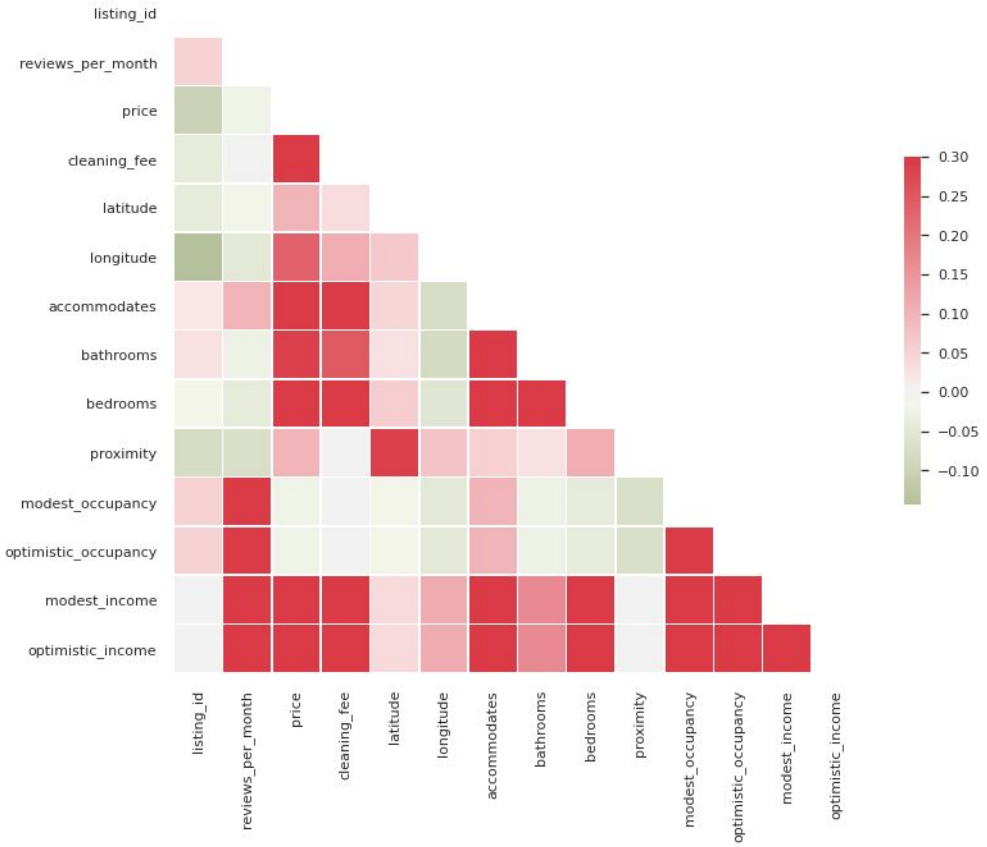

Figure 5: Correlation Heat Map

Appendix G - Seasonal Demand

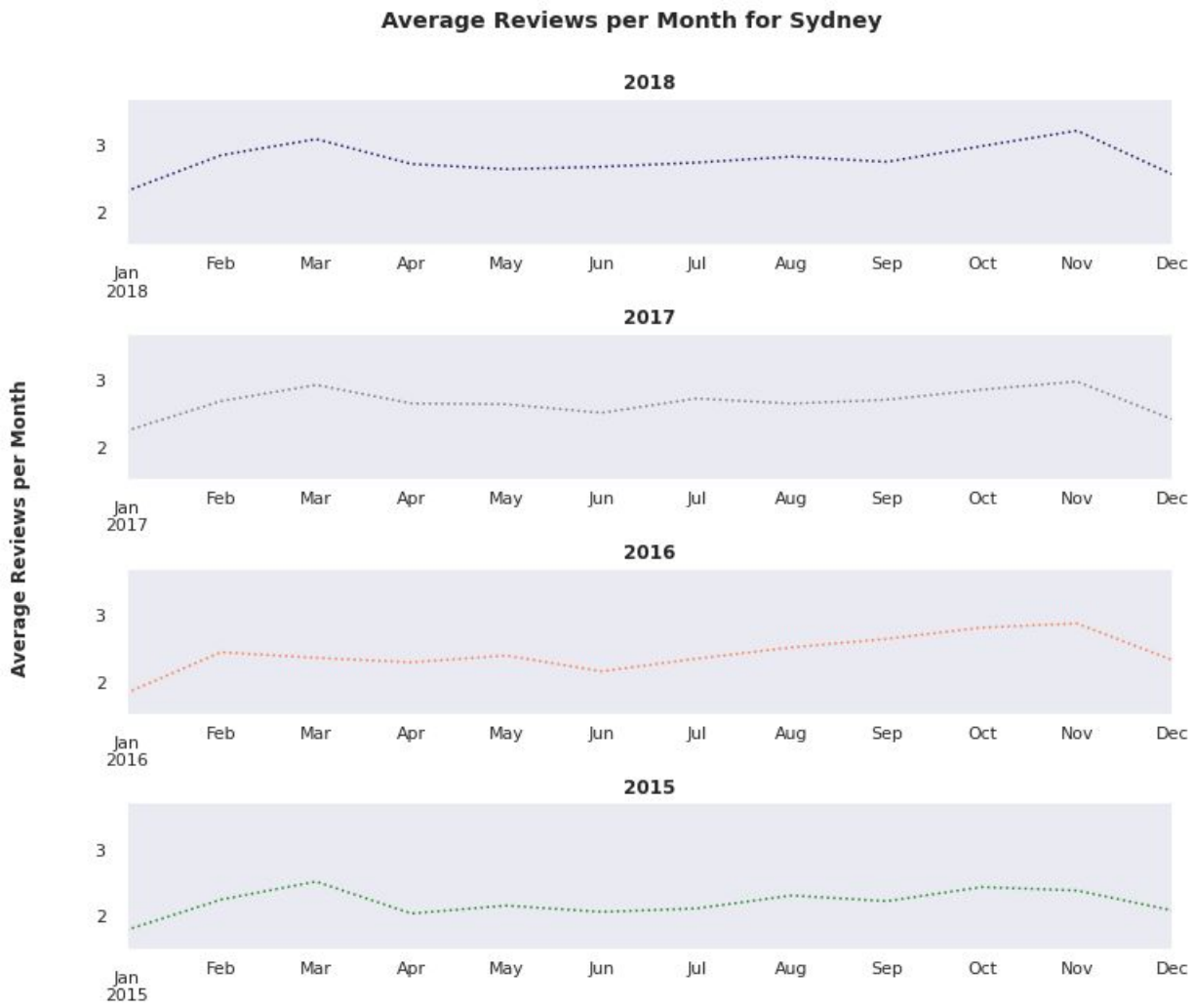

Figure 6: Average Reviews Per Month from 2015 to 2018

Appendix H - Super Hosts & Income

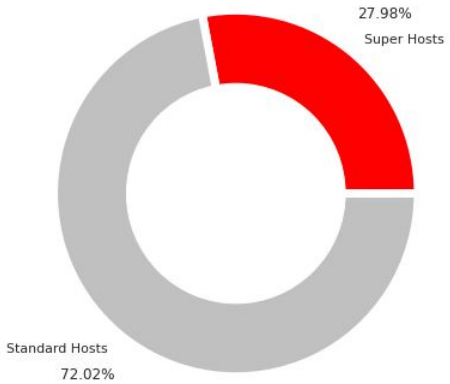

Figure 7: Pie Chart of super hosts vs standard hosts

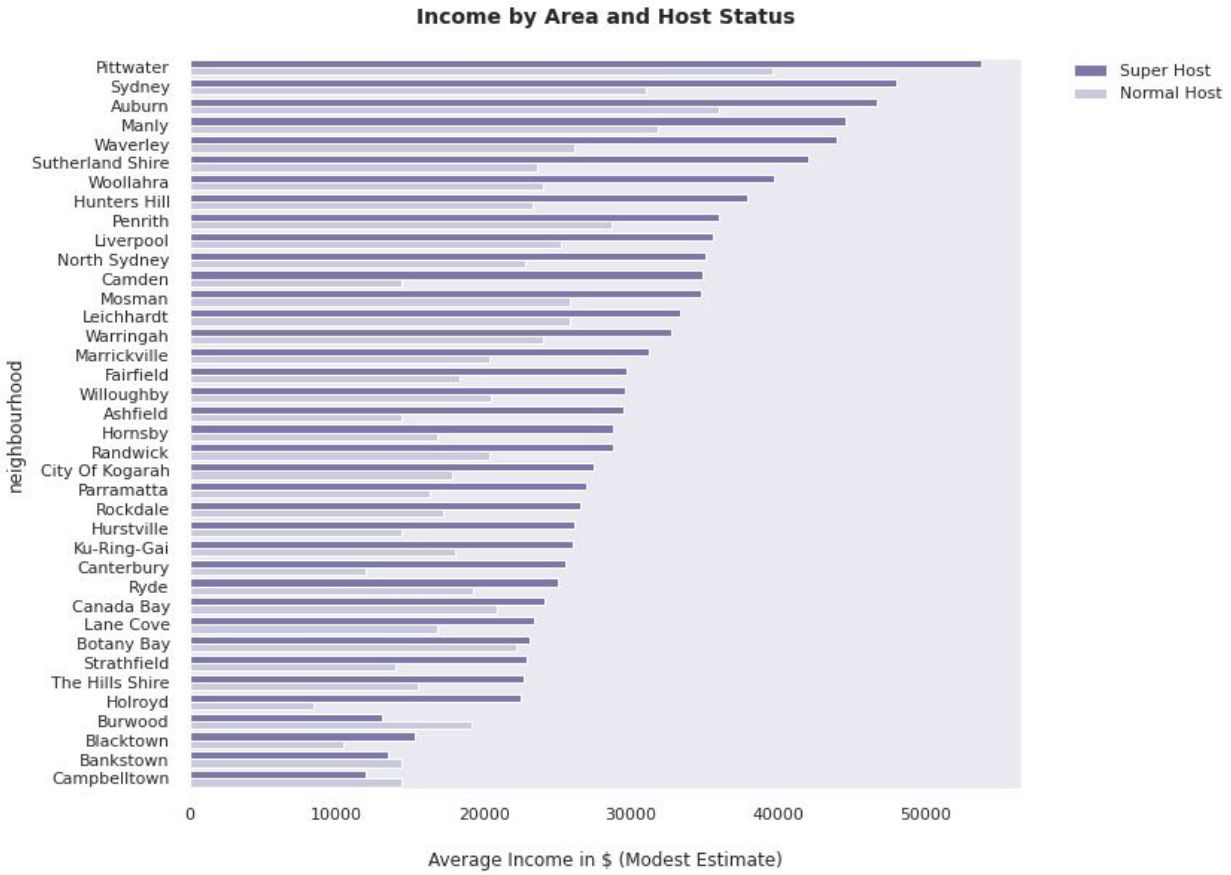

Figure 8: Modest Average Income in $ by area and host status

Appendix I - Average Occupancy Rate in % by District

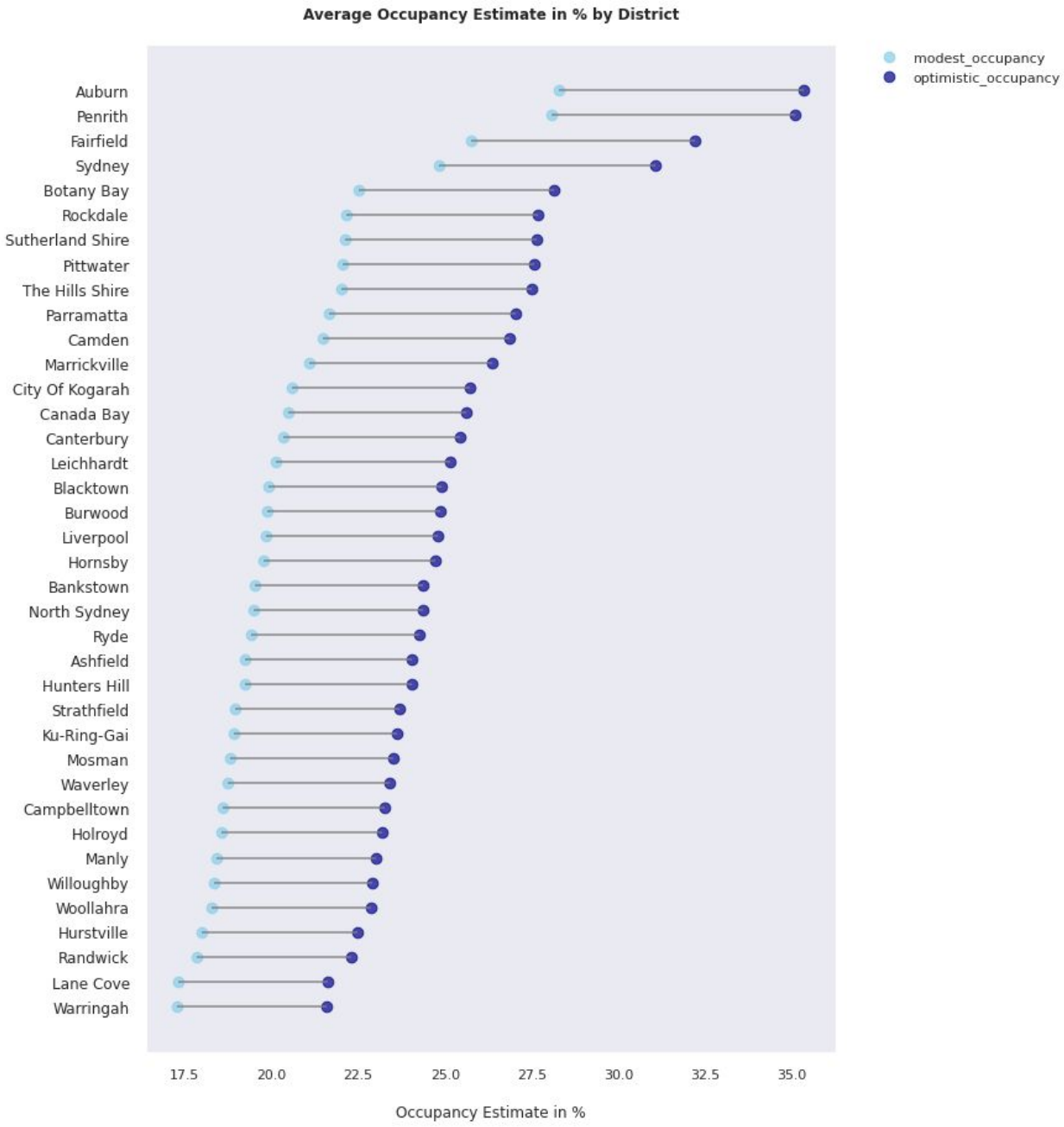

Figure 9: Average Occupancy Rate in % by District

Appendix J - Average Income Estimate in $ by District

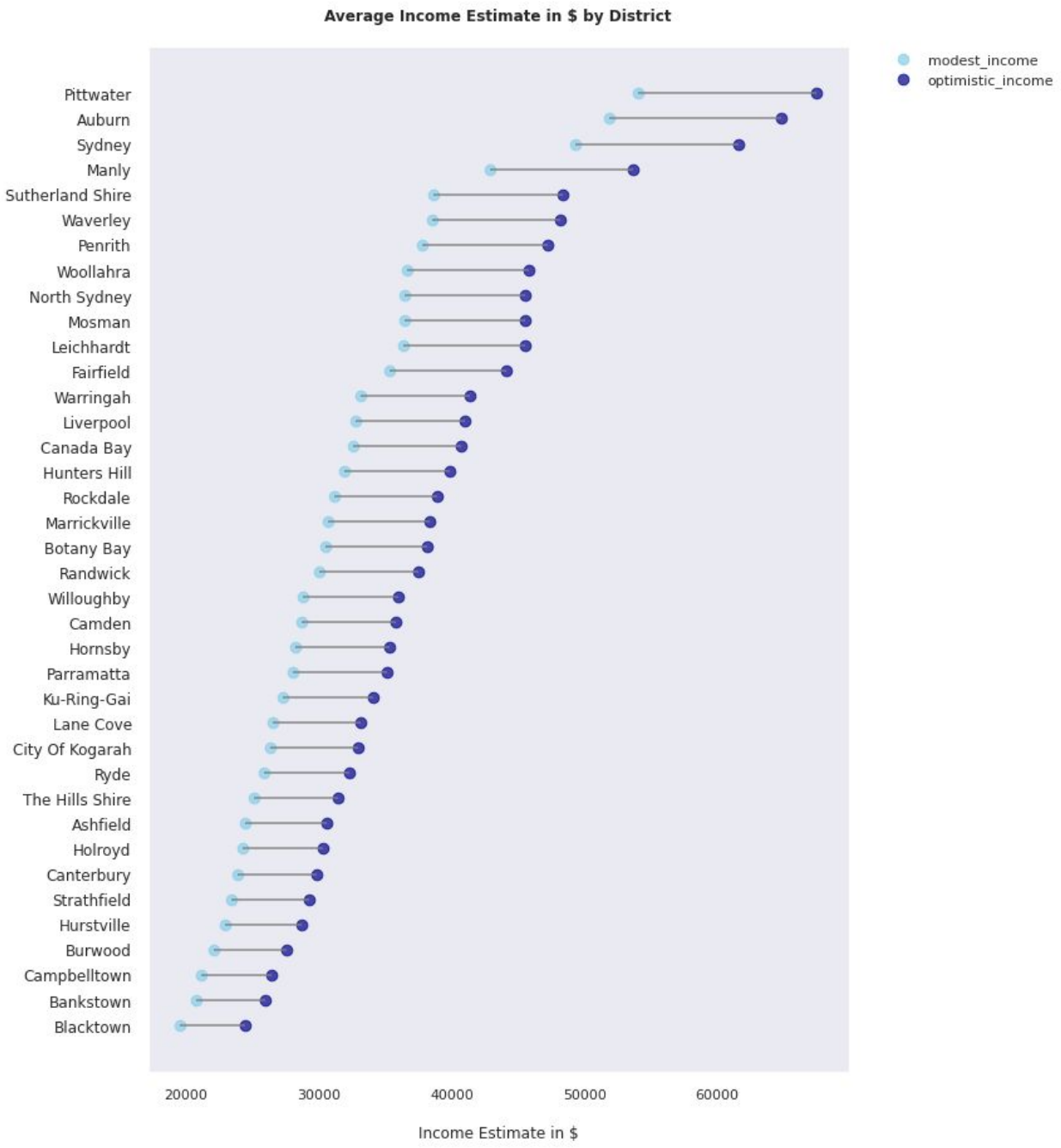

Figure 10: Average Income Estimate in $ by District

Appendix K - Proximity Transport Services and Income in $ by District

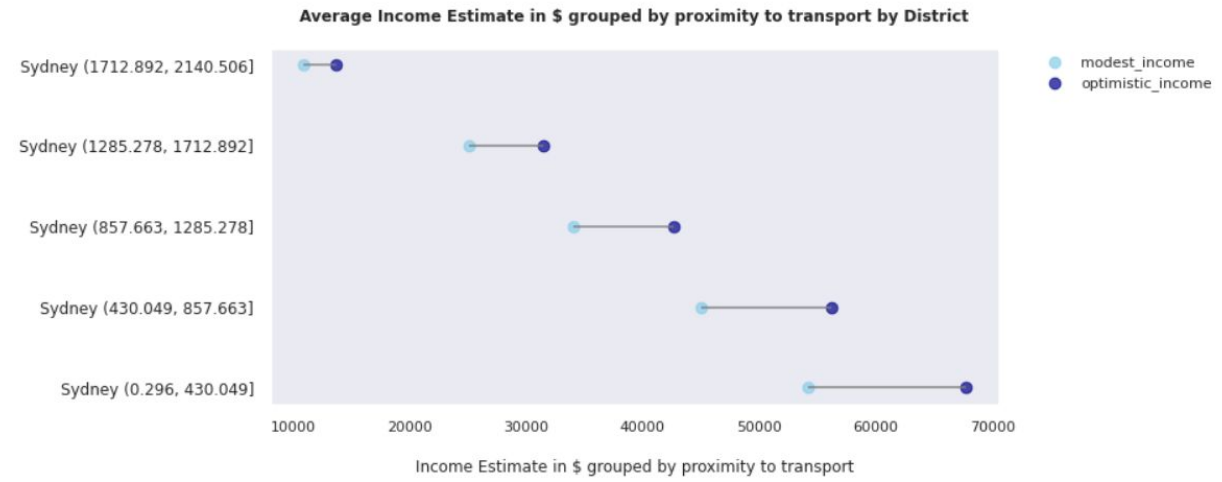

Figure 11: Income varying with distance for Sydney CBD district

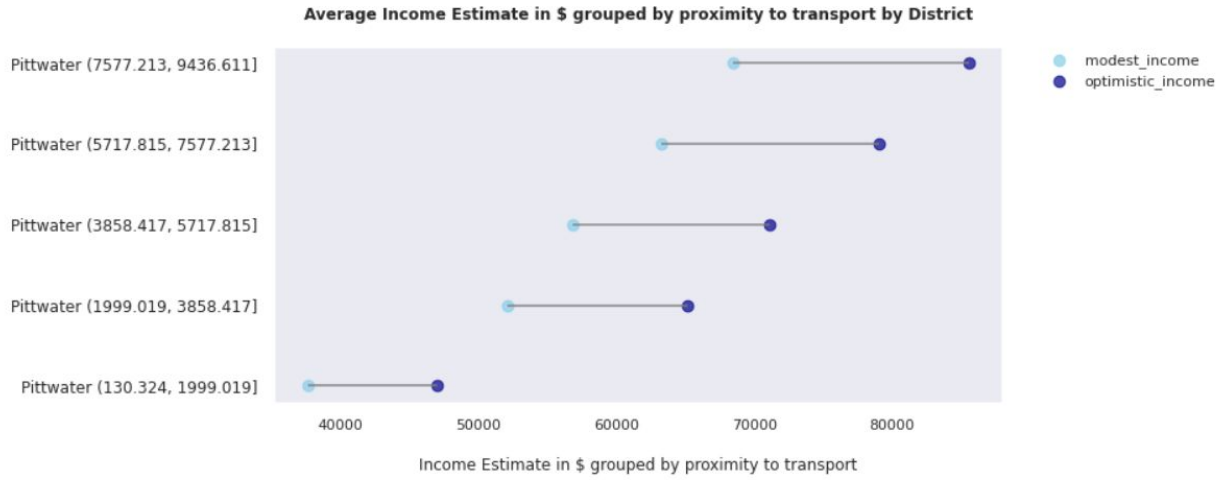

Figure 12: Income varying with distance for Pittwater district

Appendix L - Occupancy by Capacity

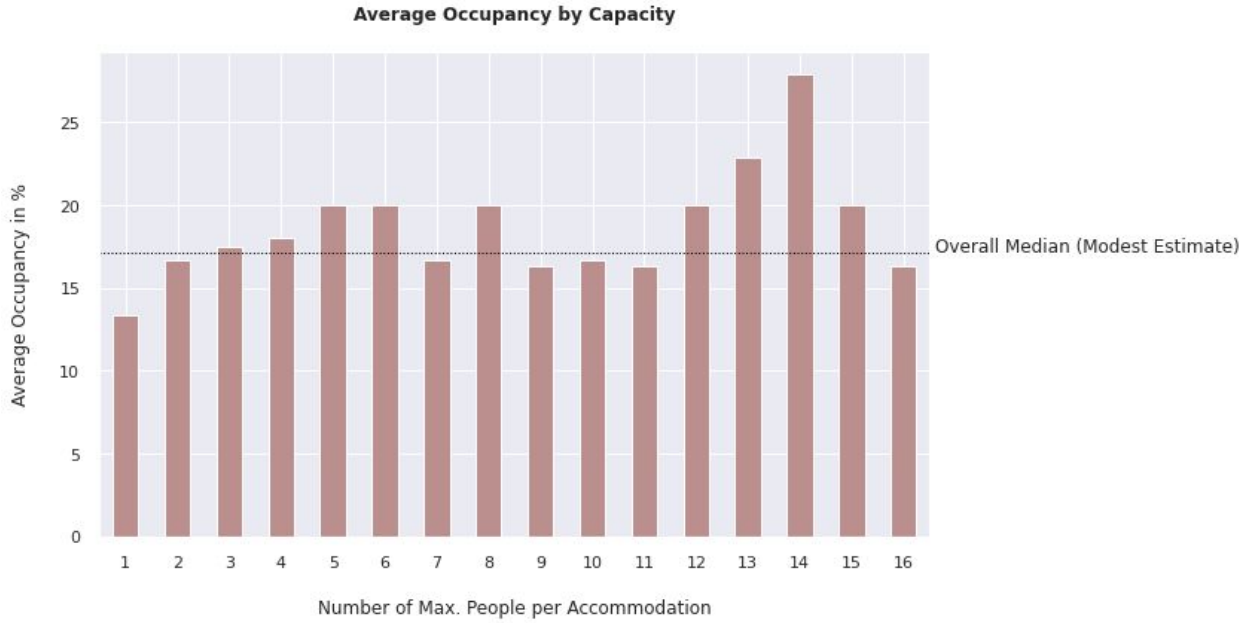

Figure 13: Average Occupancy by Capacity