Airbnb Investment

January 2021

Team

Raph Serrano
Swobabika Jena
Jason Sutton
Linda Levy

Introduction

As potential investors, our project is to investigate investment opportunities for an Airbnb property in Melbourne.
The questions of most interest to us as we began were:

- Which suburbs (or neighbourhoods) have the highest occupancy rates, and have the greatest return financially, either because of the level of bookings or because of the price charged.
- Is there a correlation between highest occupancy and highest return?
- What neighbourhoods have the highest ratings from reviews?
- What is the proximity to train stations as a significant form of transport?

Our hypothesis was that:

- Proximity to the CBD will increase earnings
- Train proximity will increase earnings
- Neighbourhoods with the highest earnings are the most popular areas
- The popularity of property types will be reflected in higher earnings

Data Sources

We decided not to use 2020 data because of the impact of COVID.
There were subsets of data from Airbnb readily available in Kaggle. However, we were concerned that each subset was created for a particular purpose by the user who created it, and did not necessarily indicate the full wealth of data that might be available elsewhere.
This proved to be the case when we dug further on Airbbnb, and found we were able to download the raw data from a website called InsideAirbnb. InsideAirbnb is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world.

The data stored by Inside Airbnb was broken down into several files, stored as either a CSV file or in a .gz format, which required unzipping with Winzip prior to saving as a csv file. Some of the data sets were extremely large. For example, the calendar file contained over 4million rows.

Train data proved easier to source and download. The Victoria Department of Environment, Land, Water and Planning website (DELWP) provides access to land registration, surveying, valuations, place naming, maps and spatial data. The process of identifying, ordering and then downloading PTV Melbourne Metro train station data proved very straight forward.

Data Clean up

Irrelevant records

Based on the data needs for our questions, the following records are not required and will be dropped from the data.

Room type: Drop private rooms / shared rooms / hotel rooms
The question is around investing in an Airbnb property so we are only interested in entire homes

Neighbourhood: To focus on purely metropolitan properties we further drooped listings to within 15km of CBD - achieved by using latitude and longitude in Airbnb data to determine the distance

Incorrect fields

Following the guidance of the following article:
https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

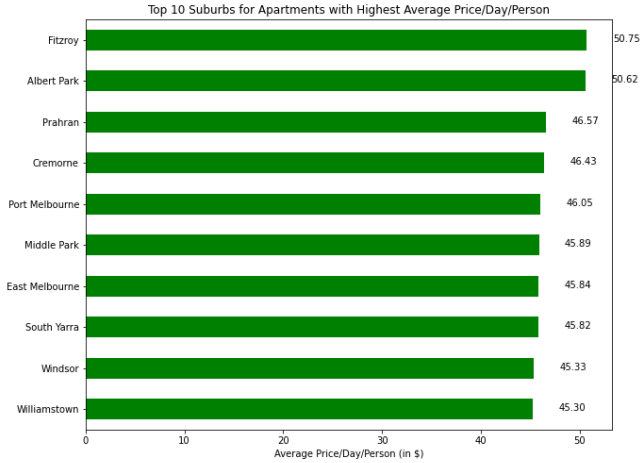
Removing the following data issues:

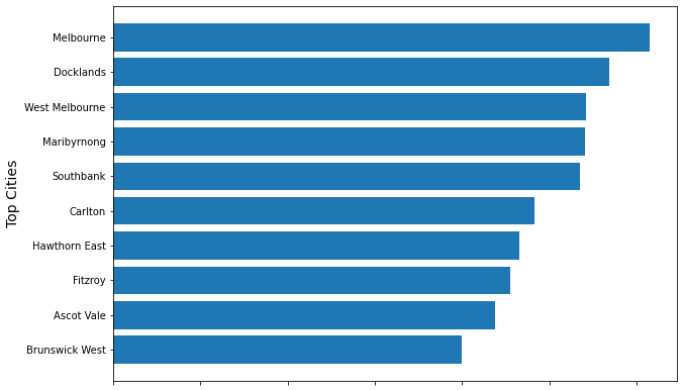
- Duplicates - none identified
- Type conversion - done
- Syntax errors - done
  i.e. Remove white spaces
- Pad strings - not required
  i.e. 313 => 000313 (6 digits)
- Fix typos - done
  i.e. check unique values for consistency
- Standardize - done
- Missing values - done
- Outliers - completed as part of further data exploration
- In-record & cross-datasets errors -not required
  i.e. ensuring total column equalled the individual breakdown columns.
- Verifying - completed as part of further data exploration
- Reporting - completed as part of further data exploration
  Reporting how healthy the data is

Data exploring

- Isolated relevant columns into new dataframe as per relevant questions to be explored.
- Determined outliers and removed values at the extreme range but kept some outliers as felt it was important to accept that the nature or our data has a large spread
- Verifying our early findings we re-inspected data to determine comparing overall rates for properties was not comparing like for like. Decision was made to only analyse apartments and to convert rate to price per person per night.
- Further data exploration indicated that suburbs with low data points were skewing finding. Removed all low quantity data points per suburb

The explorations were starting to indicate that the hypothesises were not being sustained.

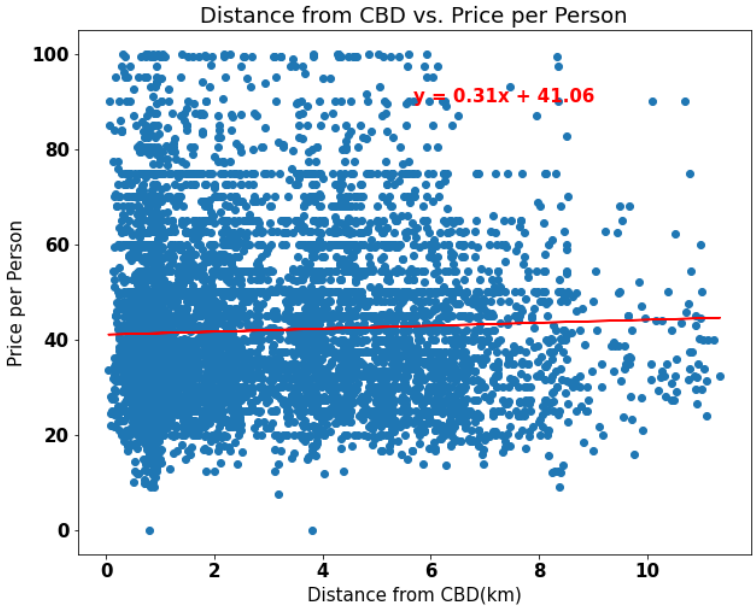




Data Analysis of the key questions

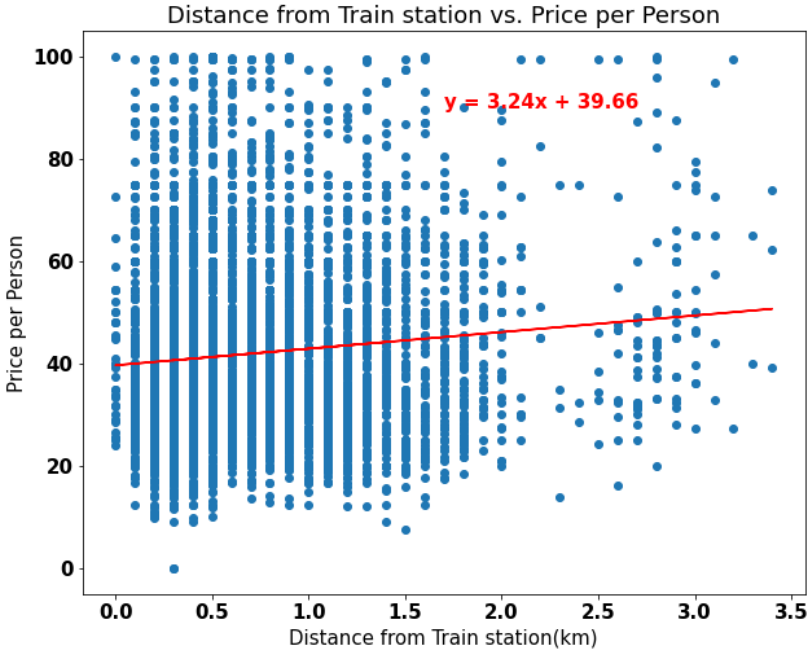
Proximity to the CBD will increase earnings

- Price per person has a weak positive correlation with distance from CBD
- This outcome indicates that distance from CBD would not factor strongly into an investment decision.



Train proximity will increase earnings

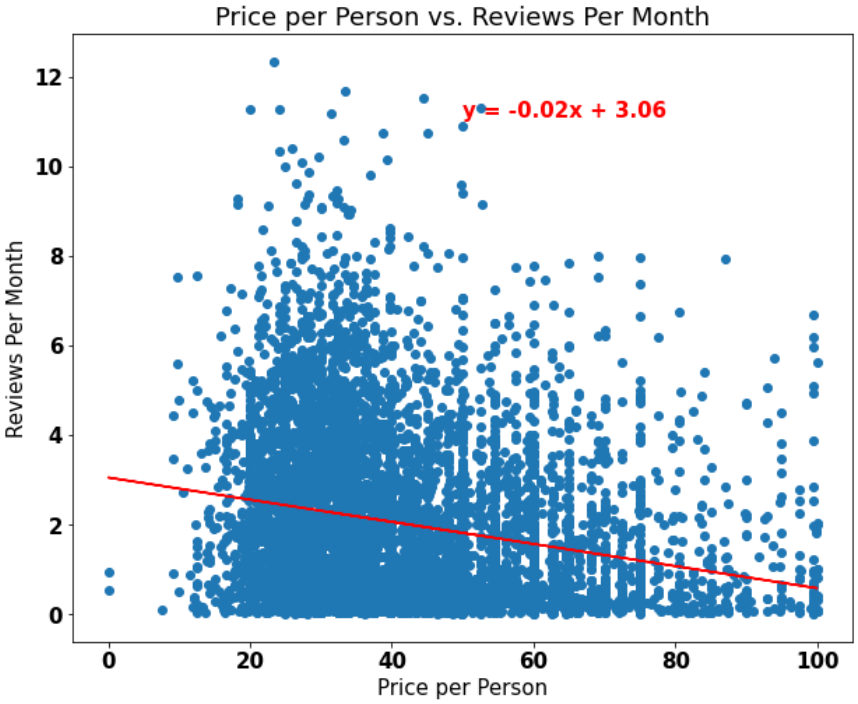
- Earnings has a slight positive correlation with distance from train station up to 3.5km
- Our initial hypothesis was that properties closest to train stations would be able to command a high rate than similar properties further away. The findings began to indicate that this may need to be broken down further. The plot here shows that within a 3.5km radius, being too close to a train station has a slight negative impact on the rate you can charge. What isn't clear from the analysis done, is whether as you move beyond 3.5km away from a station the rate goes down again.



Neighbourhoods with the highest earnings are the most popular areas

- Popularity has a weak negative correlation with prospective earnings
- This hypothesis was disproven early with our initial data exploration.
- This plot looks further into whether popularity has a correlation with the ability to charge a higher rate per person, which is also disproven.

- From an investment perspective, more analysis would be required into initial and ongoing cost of properties to determine whether the more popular low-rate properties are more profitable than the properties that charge higher rates.



The popularity of property types will be reflected in higher earnings

- This question became redundant as we realized that apartment property types far outnumbered any other property types and we focused our analyses on apartments.