Airbnb Invesment

Work plan

**15 January 2021**

**Team**

Linda Levy

Raph Serrano

Swobabika Jena

Jason Sutton

**OVERVIEW**

Our project is to uncover investment opportunities for Airbnb in Melbourne metropolitan area. We will examine relationships between neighbourhoods and occupancy rate, revenue and occupancy and whether ratings have an impact on profitability.

**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:** Drop Greater Dandenong / Manningham / Melton/ Monash/ Nillumbik/ Yarra Ranges  
  
To focus on purely metropolitan properties we further drooped listings to within 15km of CBD  
 - achieved by using latitude and longitude in data to determine the distance from the CBD

**Task: Raph**

Incorrect fields

https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Removing the following data issues:

Duplicates

* Irrelevant data
* Duplicates
* Type conversion  
  Make sure numbers are stored as numerical data types. A date should be stored as a date object, or a Unix timestamp (number of seconds), and so on.
* A word of caution is that the values that can’t be converted to the specified type should be converted to NA value (or any), with a warning being displayed. This indicates the value is incorrect and must be fixed.
* Syntax errors  
  Remove white spaces: Extra white spaces at the beginning or the end of a string should be removed.
* Pad strings: ie 313 => 000313 (6 digits)'
* Fix typos: check unique values

Gender

m

Male

fem.

FemalE

Femle  
To fix: dataframe['gender'].map({'m': 'male', fem.': 'female', ...})

* Standardize

⁃ For strings, make sure all values are either in lower or upper case

⁃ For numerical values, make sure all values have a certain measurement unit.

⁃ Dates

* Scaling / Transformation???
* Normalization
* Missing values

⁃ One. Drop. = drop columns or rows

⁃ Two. Impute ie use a linear regression

* Outliers - worth investigating before removing
* in-record & cross-datasets errors ie   
  For example, if we have a dataset about the cost of living in cities. The total column must be equivalent to the sum of rent, transport, and food.
* Verifying - re-inspecting data - sometimes even just a common sense check
* Reporting - Reporting how healthy the data is???

**Task: Jason**

**QUESTIONS**

From our discussion this morning, I think we want to head towards a final table. It might be

helpful to keep that in mind.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Neighbourhood (not individual suburb) | Average distance to train station | occupancy rate | overall rating | revenue |
| **Top 5 or 10 ?** |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| **Bottom 5 or 10?** |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**The story we are trying to tell**

* The highest occupancy rates are in the following neighbourhoods.  
  The highest earners are in the following neighbourhoods  
  Is there a correlation? Are high occupancies the highest earners.
* What neighbourhoods have the highest ratings?  
  Is there a correlation with the highest earners?  
  If there is no correlation with overall rating, is there a particulate category that rates highly with high earners.
* What is the average distance to a train station in each neighbourhood?  
  If occupancy rates and revenue are the same neighbourhoods, is close proximity to a train station a contributing factor?  
  If occupancy rates and revenue are in different neighbourhoods, which one has a close proximity to a train station a contributing factor?
* Ultimately, can we pick a recipe for a successful Airbnb – yes or nor – or more info needed.

**Data familiarity**

**2018\calendar\_dec\_2018.csv**

* **columns** = listing\_id, date, available
* available is true or false for each date for each listing ID
* Date range from 07/12/2018 - 07/12/2019

**2018\listings\_2018.csv** \*saved as calendar\_dec\_2018\_prelimClean.csv

* calendar\_last\_scraped: 7/12/18
* **Records**: 22895
* **Columns**
* D - I Listing profiles - Free text - not useful
* J - T Host profiles - incl superhost
* U - AH Accommodation location
* AI - AQ Accommodation characteristics
* AR - AX Price
* AY - BF Availability
* BG - BQ Reviews - including reviews oer month!!!
* Columns Removed:
  + equires\_license / license / jurisdiction\_names / instant\_bookable / is\_business\_travel\_ready / cancellation\_policy / require\_guest\_profile\_picture / require\_guest\_phone\_verification / calculated\_host\_listings\_count (duplicate info)

**2018\listings\_summary\_dec18.csv**

* calendar\_last\_scraped: 7/12/18
* **Records**: 22895 - same as calendar\_dec\_2018\_prelimClean.csv
* **Colums**:
  + id name / host\_id / host\_name / neighbourhood\_group / neighbourhood / latitude / longitude / room\_type / price / minimum\_nights / number\_of\_reviews / last\_review / reviews\_per\_month / calculated\_host\_listings\_count / availability\_365

**ANALYTICS**

**Which neighbourhoods indicate highest occupancy?**

Considerations

* Distance from CBD?
* Listing
* areas
* occupancy
* Average occupancy for that neighbou
* Top 10
* Bottom 10
* Bar chart – top 10 bottom10
* Average stay overall

**Task: Jason**

**What properties are getting the highest revenue? (price versus occupancy**) -

Considerations

* minumum stay
* Price
* number of reviews
* occupancy rates
* Does Price stays the same for night / week / month etc  
  Narrow down cost parameter - User average overall stay to determine which pricing column to use
* Ultimate goals - What is the revenue of highest occupancy and what neighbourhoods
* What is the top ten highest price - occupancy rate – revenue – any correlation to location
* Plot final table example

**Task: Swobabika**

**Do reviews impact overall revenue?**

Considerations

* Top 10 revenue earners what TOTAL review do they get?
* If no correlation with total review: Is there one category that they all have a high rating ?

**Task: UNASSIGNED**

**Does proximity to train station impact occupancy?**

Considerations

* Both listings and training data have long lat
* Top 5 earners – what average distance are stations?
* Bottom 5 earners - what average distance are stations?
* Aiming to see if every top earners trains stations are in walking distance

**Task: Linda**

Other areas of interest:

* most popular – rooms or whole house / flat
* Explore factors of becoming a superhosts?
  + (are top earners super host)

**Data Source**

* Melbourne Airbnb Open Data from Kaggle
* API gmaps nearest train station