J. Sharp – Feb 28th 2024

**Achievement 6 – Project Management Plan**

**Data Source:**

Airbnb Amsterdam December 2018 Open-Source Data:

<https://www.kaggle.com/datasets/erikbruin/airbnb-amsterdam>, Erik Bruin

The Data chosen contains details of all properties listed as part of the ‘Amsterdam’ AirBnB product offering as of December 2018.

The listing details contain a mixture of multiple Continuous Variables (e.g Price, N\* of reviews, minimum nights, etc.) & Categorical (e,g Availability [T/F], Room\_type, Neighbourhood). Since the data set also contains Long-Lat coordinates, we can usual spatial analysis to try and make sense of prices/popularity across the city as well as textual analysis on columns like listing descriptions to try and find key words that customers are likely to respond to. Because the data was scraped from the web, it doesn’t have any particular biases in and of itself, but there may be general collection biases stemming from AirBnB themselves *(e.g review samples only being indicative of overwhelming positive/negative reviews, because people with middling experiences failed to participate in the review collection).*

I chose this dataset because it offers a wide variety of opportunities for different visualizations that will make up a dashboard deliverable to the hypothetical stakeholder group.

**Data Profile**

6 CSV files:

* **Calendar (173.83 Mb):**
  + Listing\_id > Identifier value, No numerical significance.
  + Date: Date value, each listing appears up to 365 days (depending on when they were added to the platform)
  + Available: Binary (T/F) indicating whether a property was available to rent or not (True values indicate vacant and can be booked) – First thought would be to perhaps use a COUNT function to indicate days per year a particular property was rented (with the caveat that F may indicate it’s not on the market, rather than rented, e.g Landlord is doing repairs and had to make it unavailable)
  + Price: Quantitative, Continuous variable – Only present IF property is available on a given row.
* **Listings (3.22 Mb):**
  + Listing\_id > Identifier value, No numerical significance.
  + Name: String, No structure
  + Host\_id > Identifier value, No numerical significance.
  + Host\_name > String, No significance
  + Neighbourhood\_group > No values in column – **To be deleted**
  + Neighbourhood > String, Catagorical
  + **Latitude > Continuous – Use to Generate Point Map?**
  + **Longitude > Continuous – Use to Generate Point Map?**
  + Room\_type > categorical
  + Price > Quantitative, continuous
  + Minimum\_nights > Quantitative, discrete
  + Number\_of\_reviews > > Quantitative, discrete
  + Last\_review > Date
  + Reviews\_per\_month > Quantitative, continuous – Values don’t have *much* significance, in that they don’t indicate average stay duration or anything like that.
  + Calculated\_host\_listings > Quantitative, discrete – Unlikely to hold much significance in future analysis
  + Availability\_365 > review whether this is the same as Count Of Calendar *Availability* value
* **Listings\_details (78.15Mb):**
  + Listing\_id > Identifier value, No numerical significance.
  + Listing\_url > Self-explanatory, No significance
  + Scrape\_Id . Identifier value, no significance - **to be deleted**
  + Last\_scraped > Date Value of web scraping – No significance **to be deleted**
  + Name > String, No Structure
  + Summary > String, Possible textual analysis?
  + Space > String, Possible textual analysis?
  + description > String, Possible textual analysis?
  + Experiences\_offered > 1 value [None] – No significance, **to be deleted**
  + Neighbourhood\_overview > string, No Structure, appears to be text field inputted by host so no apparent consistency between listings
* Neighbourhoods **[Do Not Use, no Significance**] (420B):
  + Neighbourhood\_group > No values
  + Neighbourhood > 22 different items, no identifier value to be joined into prior sheets
* Reviews **[Do Not Use, no Significance**] (8.32 Mb):
  + Listing\_id > Identifier value, No numerical significance.
  + Date > date value, No additional values on the sheet, so no significance.
* **Reviews\_details (151.55Mb) [Do not Use, no significance – Review aggregates already added to listings\_details]:**
  + Listing\_id > Identifier value, No numerical significance.
  + Id > No indication as to what this pertains to – Review\_id? There are no links to other sheets, so this doesn’t have significance to the analysis.
  + Date > Date value, 9 years worth of review data – indicates date of review, not *when* reviewer stayed at a particular listings.
  + Reviewer\_id > identifier value, no numerical Significance
  + Reviewer\_name > self explanatory, string value – No significance
  + Comments > string, no structure / consistency – textual Analysis perhaps?

|  |
| --- |
| Removed from Original ‘Listings\_details’ sheet |
| access |
| calendar\_last\_scraped |
| city |
| cleaning\_fee |
| country\_code |
| experiences\_offered |
| first\_review |
| host\_about |
| host\_acceptance\_rate |
| host\_listings\_count |
| host\_location |
| host\_name |
| host\_neighbourhood |
| host\_picture\_url |
| host\_response\_rate |
| host\_response\_time |
| host\_since |
| host\_thumbnail\_url |
| host\_url |
| host\_verifications |
| house\_rules |
| interaction |
| is\_business\_travel\_ready |
| is\_location\_exact |
| jurisdiction\_names |
| last\_review |
| last\_scraped |
| license |
| listing\_url |
| market |
| medium\_url |
| monthly\_price |
| neighborhood\_overview |
| neighbourhood\_group\_cleansed |
| notes |
| picture\_url |
| require\_guest\_phone\_verification |
| require\_guest\_profile\_picture |
| scrape\_id |
| security\_deposit |
| smart\_location |
| square\_feet |
| state |
| street |
| thumbnail\_url |
| transit |
| weekly\_price |
| xl\_picture\_url |
| zipcode |