3MDL

Research Proposal

*AirBnB uptake by Business Sector  
- Drivers for policy change -*

|  |  |
| --- | --- |
| **Version:** | 0.2 |
| **Date:** | 29-April-2018 |
| **Authors:** | 3MDL – Data Science Team |

Contents

[Introduction 2](#_Toc512335271)

[Document Purpose 2](#_Toc512335272)

[Rationale 2](#_Toc512335273)

[Research Questions 2](#_Toc512335274)

[Method 2](#_Toc512335275)

[Data selection 3](#_Toc512335276)

[Assumptions 3](#_Toc512335277)

[Techniques 3](#_Toc512335278)

[Challenges 3](#_Toc512335279)

[Timeline 4](#_Toc512335280)

[Appendix A – Data Acquisition 5](#_Toc512335281)

[Appendix B – Data merging 6](#_Toc512335282)

[Appendix D – Proposed Google Directions Locations 7](#_Toc512335283)

[Appendix E – Complete list of research questions 8](#_Toc512335284)

[Appendix F – Research questions with merit for future investigation 9](#_Toc512335285)

[Appendix G – Something Really Cool 10](#_Toc512335286)

\*\* note: lots of appendices b/c trying to separate information of interested to technical audiences from main section of interest to non-technical audiences.

Document Control

|  |  |  |  |
| --- | --- | --- | --- |
| Version | Purpose | Authors | Date |
| 0.1 | Draft Template | 3MDL – DS team | 23-04-2018 |
| 0.2 | Initial Draft – with stakeholder feedback | 3MDL – DS team | 24-04-2018 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| 1.0 | Final version | 3MDL – DS team | 29-4-2018 |

# Introduction

The 3MDL Data Science Team is proposing a research project to identify the factors driving increased use of AirBnB for business travel in Australia and its impact on the local communities. This project aims to deliver a foundation for identifying the value to property owners and local-residents in regulating or incentivising AirBnB for the business sector. The findings from this research project will be used to make recommendations for changes to existing planning and zoning policy in alignment with 3MDL’s strategy vision statement “To hear the community voice”.

Identifying factors that influence…”alternative hotel market: identify the factors that influence price, and occupancy rates in AirBnB properties in Sydney. To answer a range of questions from different stakeholders … govt, hotel and community sectors

## Document Purpose

This proposal document presents an overview of the project rational and articulates the research questions to be investigated. Also presented are findings from the data selection process, proposed modelling techniques and anticipated challenges.

# Rationale

*{Introduce and define our broad research area and outline the purpose of our investigation.}.*

The area of interest for this investigation is business sector opportunities for property owners using AirBnB and the impact on local communities. (or are we looking at housing, /business sector)

The proposed research aims to progress the following objectives:

* Improve an understanding of the value in listing a property with ‘Business Ready’ badge
* Identify the factors that distinguish BR listings and non-BR listings based on attributes of the property
* Identify appropriate ‘Business Ready’ badge qualifications for Australian properties and business sector

# Research Questions

*{introduce our questions. List them. Explain how these relate to the area of research}*

This investigation will attempt to answer the following questions:

* What are the factors for high BR listings?
* What factors result in a good review?
* Do Australian ‘Business Ready (BR)’ listings adhere to BR badge rules?
* Can we predict the increase in BR listings based on projected increase in visitors traveling on business?
* Can we Predict the increase in ‘Business Ready’ badged listings
* How can we use price and occupancy rates to see what drives BR?. Anticipate lower top-end (travel caps)
* {link to making use of under occupied.
* {link into price … factors that influence price, but are these factors for BR? Eg: BR tend to be 2bedroom houses … what are the recommendations around relaxing BR properties to make it more suitable to Aus. BR push up prices?

Explanation of how these relate to the rationale … (??)

# Method

*{identify the question(s) that we will use a predictive regression model to answer, and discuss}*

Blah blah … the regression model question … blah blah

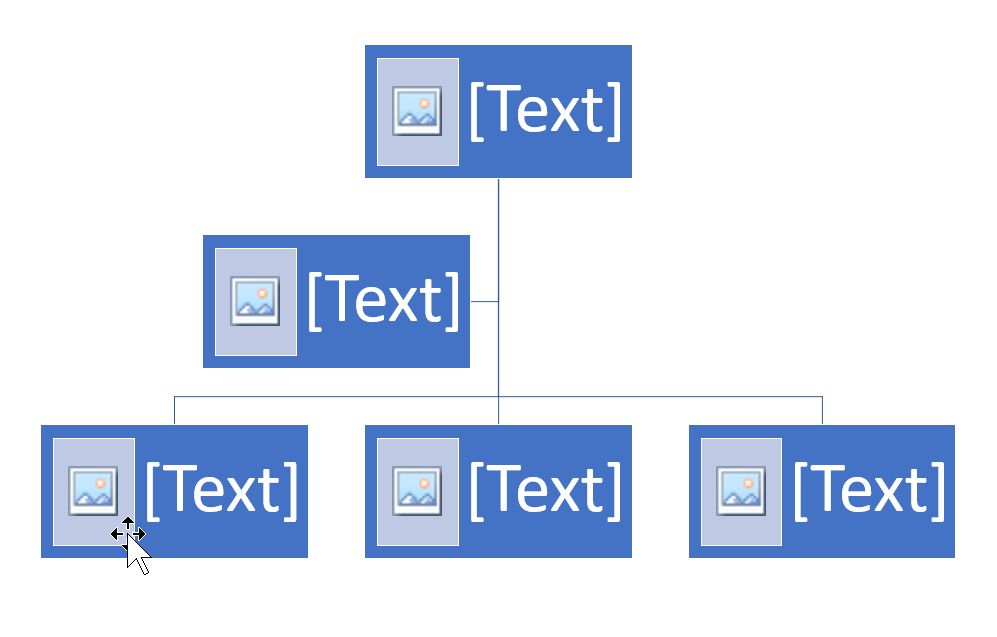


Figure - module map outlining our approach (picture tells a 1000 words but needs a better caption!!)

Refer to Appendices for data acquisition and data merger code samples.

## Data selection

*{Discuss range of datasets we considered. List those we have selected and selection process.   
Discuss how these contribute to answering the question, commonality, how they can be merged – possibly touch on privacy/ethics/legality}*

A range of datasets were identified during the ideation stage of this project. Each was assessed based on it quality, validity, accessibility and applicability to the research questions.

The following data was rejected b/c …. (insufficient granularity, could not be linked to primary dataset, volume too high so reduced for EDA trials, high level of sparsity)

The xx, xxx and xxx datasets were selected based on commonality

The final stage of selection was to address the P/E/L considerations such as … (assumption that we are working with identifiable data -exact location, identifiable property, hosts names – Inside AirBnB claims this is not private b/c it’s publicly avavalbe …. Aust privacy act refers to personal rather than private. We know this is about a person, our analysis may derive inferences about that person that may or may not be true. Eg: female guests only

Selected datasets:

* Scrapings from AirBnB, collected and made available by Inside AirBnB
* TRA – visitor numbers (rolled up listings – regression not of value?)
* Distance (distance, time, transport mode. grain: geo)
* Image analysis (separate acquisition - feature engineering)
* Sentiment review (constructed dataset – feature engineering)
* ~~Rental (to be located - )~~
* ~~Obike data (impact on community)~~

Refer to Appendix XXX for complete list considered.

## Assumptions

Project assumptions and the basis for forming them, refer to Table 1.

Table

|  |  |
| --- | --- |
|  | **Assumption** |
| Proxy for ‘vacancy’ – number of reviews per month. | A stay results in a review. |
| Accuracy of location data | ?? |
| Inside AirBnB dataset shows apparent discrepancies (eg: is\_business\_ready) | Scrape is an accurate representation of AirBnB listings |
| Alternate items listed in Image analysis | Take the item with highest level of confidence |
| Google distances API | Validation of locations against free-text query Google Maps Locations is sufficient. |

## Techniques

*{Discuss techniques used in the data acquisition and merger process and EDA modelling. – highlight our creativity and mastery of R}*

custom functions – price clean (converts 3 lines of code into 2 lines of code)

**Data Acquisition:**

Google distance API, Azure API – sentiment analysis, AWS Rekognition and S3 Rest API for Image Recoganition, Excel API – file extraction

**Data merging:**

Sourcing from external sites and caching on GDrive to provide centralised datasets for use by the team.

R – various munging techniques, refer to Appendix B – Data merging

**EDA modelling:**

Linear regression, logistic regression – for inference tasks, more interpretable results

R functions – explorer(), lrfit(), glm()

Visualisations – histogram, scatterplot, wordcloud

# Challenges

*{Discuss the challenges we anticipate based on the issues we encountered during EDA. Outline our approach for addressing these challenges, propose potential mitigations for any risks.}*

Data volumes – prohibitive expense during EDA (modelling phase requires appropriate resources and funding)

Merging datasets – publicly available data with little commonality to match on

Sourcing assumptions – cite some of the issues from insideairbnb.com (?)

# Timeline

Refer to Figure 1 for proposed timeline and next steps.



Figure - Project timeline

# Appendix A – Data Acquisition

*{code samples – data acquisition}*

# Appendix B – Data merging

*{code samples – data merging\_*

# Appendix D – Proposed Google Directions Locations

*{Full list of locations we’d like to include … obviously we can’t do all in our code but this gives our proposal substance}*

# Appendix E – Complete list of research questions

*{Use this if our word count gets too high!}*

# Appendix F – Research questions with merit for future investigation

*{The questions we think would have been really cool but are just too big for this assessment task. Highlight our creativity!}*

# Appendix G – Something Really Cool

*{Not quite sure what but this is the place to put anything we think is great but doesn’t fit within our word count 😉}*