Henley
Business School
UNIVERSITY OF READING

Informatics MSc Programme Area
Henley Business School
University of Reading

Assessed Coursework Set Front Page

Module code: INMR77
Module name: Business Intelligence and Data Mining
Lecturer responsible: Dr Yin Leng Tan

Work to be handed in by:
Full time student: 9th May 2022 @2pm
Part time student: 3rd June 2022 @2pm

Assignment Specification:

The module is assessed ***100% through this coursework assignment***.

The aim of this coursework is to assess your understanding of business intelligence and data mining and ability to perform data mining tasks by applying suitable concepts, methods and techniques learned during the lectures and practical sessions for business intelligence.

The coursework is carried out ***individually***. You are required to produce an **individual report** for the tasks as set out below. The complete report should not exceed **20 pages of A4** with a 10% of variation, including tables and diagrams but excluding references and appendices.

An appendix can be used to include support materials to back up main body points where necessary. In addition, **you are also required to submit the supplementary materials** of your work ***using SAS Enterprise Miner*** on blackboard by the specified deadline.

Opportunities and Challenges of Sharing Economy: A case of Airbnb and Inside Airbnb

Airbnb - Holiday Lets, Homes, Experiences & Places (airbnb.co.uk)

Airbnb is an online marketplace for arranging or *offering short-term rental/lodging i.e. temporary accommodation, primarily homestays, or tourism experiences*. It was founded in August 2008 by Brain Chesky and friends, and it currently has 6,300 employees.

On December 10, 2020, Airbnb went public with a valuation of over $100 billion, making it one of the largest IPOs (Initial Public Offerings) of 2020. It is reported that Airbnb capital market was more than the top three largest hotel chains (Marriott, Hilton, and Intercontinental) combined[1]. Though some are calling over evaluation, the company lacks traditional mortgages, employee fees, and maintenance fees which burden hotels. Airbnb hosts pay their own

[1] https://www.businessinsider.com/airbnb-ipo-valuation-tops-three-hotel-chains-combined-opening-day-2020-12?r=US&IR=T

mortgage and clean their apartments, leaving the company much freer of debt, thus making it far more valuable.

Airbnb service overview

Airbnb ***provides a platform*** for hosts to accommodate guests with *short-term lodging and tourism-related activities*. Guest can search for accommodation using filters such as location, price, specific types of home. Before booking, users must provide personal and payment information. Some hosts also require a scan of government-issues identification before accepting a reservation. Hosts provide prices and other details for their rental or listing e.g. number of guests included in the price, type of property, type of room, number of bathrooms, number of bedrooms, number of beds and type of bed, and minimum number of nights for a reservation, and amenities.

In addition, Airbnb provides a ***guest review system*** where hosts and guests can leave reviews about their experience, and rate each other after a stay. However, the truthfulness and impartiality of reviews may be adversely affected by concerns of future stays because prospective hosts may refuse to host a user who generally leaves negative reviews. Besides, the company's policy requires users to forego anonymity, which may also detract from users' willingness to leave negative reviews.

Criticism of Airbnb

Airbnb has attracted criticism for increasing rent prices in cities where it operates, and creating nuisances and security issues etc for those living near leased properties and has negatively affects the quality of life in residential areas. The company has attracted regulatory attention from cities such as San Francisco, New York City, and the European Union over the past number years. It has also faced challenges from the hotel industry and other, similar companies.

Airbnb has made a quarter (25%) of its global workforce redundant in 2020 due to the global pandemic[2]. But the news was welcome by some campaigners who were fighting for soaring rents in cities with large number of Airbnb hosts. The number of longer-term rental properties in central Dublin was up 71% on comparable period last year, as landlords abandoned short-term lets through Airbnb[3].

Inside Airbnb – Adding Data to the Debate (insideairbnb.com)

Inside Airbnb is an independent, non-commercial set of tools and data that allows individual to explore how Airbnb is really used in cities around the world. It was set up by Murray Cox and John Morries in 2016.

The following description (excerpt) is taken from Inside Airbnb. For more information, see http://insideairbnb.com/about.html

Airbnb claims to be part of the "sharing economy" and disrupting the hotel industry. However, data shows that most Airbnb listings in most cities are entire homes, many of which are rented all year round – disrupting housing and communities.

By analysing publicly available information about a city's Airbnb's listings, Inside Airbnb provides filters and key metrics so users can see how Airbnb is being used to compete with the residential housing market. With Inside Airbnb, user can ask fundamental questions about Ainbnb in any neighbourhood, or across the city as a whole, such as:

[2] https://www.theguardian.com/technology/2020/may/06/airbnb-to-make-quarter-of-its-global-workforce-redundant

[3] https://www.rte.ie/news/ireland/2020/0417/1132149-rent-sale-property-ireland/

- how many listings are in my neighbourhood and where are they?
- how many houses and apartments are being rented out frequently to tourists and not to long-term residents?
- how much are hosts making from renting to tourists (compare that to long-term rentals)?
- which host are running a business with a multiple listings and where are they?

These questions (and the answers) get to the core of the debate for many cities around the world, with ***Airbnb claiming that their hosts only occasionally rent the homes in which they live***. In addition, many cities or state legislation or ordinances that address residential housing, short term or vacation rentals, and zoning usually make reference to allowed use, including:

- how many nights a dwelling is rented per year
- minimum nights stay
- whether the host is present
- how many rooms are being rented in a building
- the number of occupants allowed in a rental
- whether the listing is licensed

The Inside Airbnb tool or data can be used to answer some of these questions. Some understanding of how the Airbnb platform is being used will help clear up the laws as they change.

Source: Wikipedia, Airbnb.co.uk, insideairbnb.com
Further information of Airbnb, please visit: https://www.airbnb.co.uk/
Further information of Inside Airbnb, please visit: http://insideairbnb.com/index.html

---

Coursework requirements and what to deliver:

The sharing economy has brought opportunities and challenges to homeowners, society, residents, communities and governments. **One of the biggest issues with Airbnb is whether hosts are renting out residential properties permanently as hotels, as opposed to sharing the primary residence in which they live "occasionally".**

Airbnb could easily answer this question but instead it is up to us to shape our communities and solve our urgent need to house tourists, and to address the nuisances, security and safety issues etc for those living near leased properties by Airbnb.

In this assignment, you are required to carry out data mining tasks **using datasets** of ***Airbnb listings of Greater Manchester from Inside Airbnb – (as below and attached)***, and to report your findings as a result your data mining/analysis effort as to address issues as stated above.

DATA SETS – two sets of data sets will be used for this assignment.

1. **Listings as on and up to 23rd December 2021**
   - **GM_listings_2021.csv** contains detailed listing data for Greater Manchester. The data was compiled on 23 December 2021. Each row of the data represents a single listing and contains information about the host of the property, the property's characteristics and overall rating of the property and its features by guests. There are 3,447 listings and 67 variables in the data set.

   - **GM_reviews_2021.csv** contains the reviews for each listing. The data was used for a number of derived variables in the listing dataset i.e. number_of_reviews, number_of_review_ltm, first_review, last_review, and reviews_per_months.

- **GM_calendar_2021.csv** contains detailed calendar data i.e. the availability calendar for 365 days in the future for each listing.

2. **Listing as on and up to 23 December 2020**

   - **GM_listings_2020.csv** contains detailed listing data for Greater Manchester. The data was compiled on 23 December 2020. Each row of the data represents a single listing and contains information about the host of the property, the property's characteristics and overall rating of the property and its features by guests. There are 3,516 listings and 65 variables in the data set.

   - **GM_reviews_2020.csv** contains the detailed reviews for each listing. The data was used for a number of derived variables for the listings 2020 dataset i.e number_of_reviews, number_of_review_ltm, first_review, last_review, and reviews_per_months

   - **GM_calendar_2020.csv** contains the detailed calendar data (the availability calendar for 356 days in 2020) for each listing.

3. **A data dictionary for the variables used.**

You are required to use the ***listings data sets*** for the purpose of this assignment. You may want to refer to the ***calendar and review data sets*** and derive further new variables where necessary although it is not compulsory. Listings can be deleted in the Airbnb platform. The data presented is a snapshot of listings available at a particular of time i.e. 23 December 2020 and 23 December 2021.

You are to conduct cluster analysis and identify cluster/segment of similar listings based on the information of the host of the property, the property's characteristics and so on, and to build a model that could differentiate hosts/listings that are for "occasionally-short-term let" vs "long-term let".

To do so, you are also expected to create new derive variable(s) to calculate the occupancy model. Note, Inside Airbnb Greater Manchester uses the following parameters for the occupancy model:

- A high availability metric and filter of 60 days per year
- A frequently rented filter of 60 days per year
- A review rate of 50% for the number of guests making a booking who leave a review
- An average booking of 3 nights unless and higher minimum nights is configured for a listing
- A maximum occupancy rate of 70% to ensure the occupancy model does not produce artificially high results based on the available data.

Detailed information of *disclaimers and occupancy model* of the data is available on: http://insideairbnb.com/about.html and http://insideairbnb.com/greater-manchester/. However, you are welcome to define your own parameters/measurements for the occupancy model, make sure you provide the reasons of the approaches taken, backed up by relevant sources/literature/

You may also want to compare and analyse the characteristics of deleted Airbnb listing and/or compares the listings and any changes of the listings between the 2021 and 2022.

The total of 100 marks will be allocated to the following aspects of the report, which should also be used as a guideline to structure the report.

1. **Introduction - identifying the problem and opportunity (20%)**
   The introduction section should provide an analysis of the specific problem situation/context to be addressed as in the case study, including a discussion on the impact of Airbnb to key stakeholders, and how data mining could be used to address the issues identified. You are also expected to conduct further research in the problem domain and justify your statements using relevant literature/sources. You should also provide a clear statement of the data mining goal.

2. **Data understanding and data preparation (30%)**
   This section should detail your data understanding and data preparation process in relation to the data mining goal.
   a) Data understanding: in this section, you are required to conduct exploratory data analysis using suitable and relevant techniques and methods e.g. summaries statistics and data visualisation techniques, and report your key finding, including variables and measurement identified for your data preparation tasks.
   b) Data preparation/Data pre-processing: in this section, you are expected to take the data identified in the previous step and prepare them for analysis by data mining methods. This should include data cleaning/missing data handling, data transformation (e.g construct new attributes/derived variables e.g. length of host etc), and data reduction (e.g. reduce number of variables, correlation analysis etc)

   You are also expected to justify the approaches taken by using literature/sources. Make sure to include figures and tables (screenshot) to support your analyses and findings.

3. **Model building, results interpretation and evaluation (30%)**
   In this section, you are expected to conduct both unsupervised (cluster analysis) and supervised e.g classification learning.

   a) For unsupervised learning, you are expected to conduct cluster analysis to identify clusters/segments of listings based on a different or a combination set of variables e.g. host's characteristic, listings/property's characteristics and availability, and reviews from guests, as so on, and to find any patterns for the listings of Greater Manchester (eg occasionally vs long term let). Make sure you identify and justify the variables/measurements and techniques used for your clustering tasks.

      You are also required to interpret the results obtained and comments on the characteristics of the clusters/segments obtained.

   b) For supervised learning, you are expected to build a classification model based on the results obtained from your clustering tasks e.g. to predict listings/hosts who are likely to be long term let or vice versa. Since this information would most likely to be used in identifying those listings/hosts are likely to be long term let as opposed to short term let, it would be more meaningful to select a segment/cluster that would be defined as "long-term let" in your classification model.
      You are also required to justify the variables used for your classification tasks and evaluate the performance of your model.

Make sure to include figures and tables (screenshots) to support your model buildings, analyses and findings. Supplement materials can be provided at the appendix section.

4. **Conclusion, critical evaluation, and further improvements (20 marks)**
   In this section, you should conclude the outcomes of your finding in relation to your data mining goal. And discuss the limitations of your data mining results, this might include the assessment of the suitability e.g. data and variables, methods and techniques used, impact and potential risks for the assumptions made, and provide suggestion for further improvements for your model building.

The criteria used for grading assignment:

Aspects/Criteria	**% Range**	**Descriptors**
Introduction - identifying the problem and opportunity (LO1, LO3, LO5)	80% and above	An outstanding, highly effective introduction and original analysis of the specific problem situation/context to be addressed as stated in the case study. Wide background reading; outstanding examples and wide use of relevance literature/sources in supporting the arguments/viewpoints. The data mining goal is clear and well defined.
	70%-79%	A highly effective introduction and original analysis of the specific problem situation/context to be addressed as stated in the case study. Wide background reading; excellent examples and use of relevance literature/sources in supporting the arguments/viewpoints. The data mining goal is clearly given and well defined.
	60-69%	A very good introduction and analysis of the problem situation/context to be addressed as stated in the case study. Very good background reading; generally a very good use of examples and relevant sources/literature in supporting the arguments/viewpoints. The data mining goal is well given and defined.
	50-59%	Adequate introduction incorporating one or more of the above, yet lacking in clarity in some area(s). Some use examples and sources/literature in supporting the arguments/viewpoints. The data mining goal is reasonably given and defined.
	49% and below	A basic introduction with a narrow or limited reference to defining the area and problem to be addressed in the case study. Little evidence of appropriate reading or ability to synthesise information. No or little examples given. The data mining goal is unclear.

Data understanding and data preparation (ILO2, ILO3, ILO4, ILO6)	80% and above	Outstanding and original. An outstanding, comprehensive, well-focused, original analysis, entirely relevant to the tasks with outstanding support and justifications for the variables and techniques used.
	70%-79%	Comprehensive and original. A comprehensive, coherent, well-focused, original analysis, entirely relevant to the tasks with excellent support and justifications for the variables and techniques used.
	60-69%	A generally clear and coherent analysis and insights with good focus , support or justification of the variables and techniques, which is directly relevant to the tasks. Clear rationale for the approaches taken.
	50-59%	Reasonable analysis but prone to being descriptive with little critical thoughts; little rationale for the variable used and approaches taken.
	49% and below	Weak rationale for approach and little relevant analysis. Failure to understand the purpose of the assignment. Unsubstantiated assertions and factual inaccuracies.
Model Building, Results and Evaluation (ILO2, ILO3, ILO4, ILO6)	80% and above	Outstanding work. An outstanding , coherent, well focused, original approaches in the model building, entirely relevant to the tasks with outstanding support and justifications for the variables, techniques and models used for the modelling. Outstanding discussion and interpretation of the results/analysis obtained. Outstanding model evaluations and comparisons provided with clear evidence of critical analysis of findings.
	70-79%	Excellent work. An excellent, coherent, well focused, original approaches in the model building, entirely relevant to the tasks with excellent support and justifications for the variables, techniques and models used for the modelling. Excellent discussion and interpretation of the results/analysis obtained. Excellent model evaluations and comparisons provided with clear evidence of critical analysis of findings.
	60-69%	A generally clear and coherent discussion with good support or justification for the model building, which is directly relevant to the tasks. Clear rationale for the approaches taken. Very good discussion and interpretation of the results/analysis obtained. Very good model evaluations and comparisons provided with some critical analysis of findings.

	50-59%	Reasonable attempt of the modelling but prone to being descriptive or narrative; little rationale for the approaches taken or justification of the variable used. Generally relevant to the stated tasks. Reasonable discussion and interpretation of the results/analysis obtained. Reasonable discussion of model evaluations and comparisons though with little evidence of critical analysis of findings.
	49% and below	Little discussion and evidence of model building. Failure to understand the purpose of the task. Little discussion and interpretation of the results/analysis obtained. Little or no discussion of model evaluations and comparisons
Conclusion, critical evaluation and future improvements (ILO1, ILO5 and ILO6)	80% and above	Outstanding, comprehensive, and extremely well discussed with original insights that drawing from the analyses conducted and suggestion for future improvements.
	70-79%	Comprehensive and extremely well discussed with original insights that drawing from the analyses conducted and suggestion for future improvements.
	69-69%	Very well discussed with interesting insights that drawing from the results/analyses conducted. Very good critical evaluation and suggestion for future improvement.
	50-59%	Reasonable discussed but prone to being descriptive with little critical analysis based on the results/analyses conducted. Generally relevant to the stated tasks. Some critical analysis but prone to being descriptive or narrative; evidence supports the conclusion, but not always very directly /clearly. The question is not fully addressed.
	49% and below	Largely descriptive. The discussion is limited in scope and/or relevance. The question is only partly addressed.