Assignment A2: Estimation				
Student Name	Yash Mistry		Student No	218612723
Problem attempted	Complex Model 80-100%	Simple Model 40-79%	Student Id	Ydmistry@deakin.edu.au
Place "Yes" in one only	?	yes	*Do not attempt a complex model unless you can complete a simple model first!*	

Partial Submission	Exceptional	Very Good	Good	Acceptable	Improve	Unaccept.
Exec Problem	5	4	3	2	1	0
Data Exploration	10	8	6	4	2	0

Final Submission	Exceptional	Very Good	Good	Acceptable	Improve	Unaccept.
Exec Solution	5	4	3	2	1	0
Data Preparation	20	16	12	8	4	0
Model Development	20	16	12	8	4	0
Model Evaluation	20	16	12	8	4	0
Model Application	20	16	12	8	4	0

Brief Comments

Read these notes

These and the following notes are trying to help you!
Read the rubric on how the report content is going to be assessed!
Your partial submission may not be perfect but has to reflect a genuine effort.
We expect your partial submission sections to be improved for the final submission.
We will not look at your partial submission until we mark the final submission.
We will assess the final submission and its mark stands.
However, we will deduct marks if the quality of the partial submission is poor.

Do not attempt a complex model unless you can complete a simple model first!
If you cannot formulate a complex problem, you will not get extra points for other complex criteria.
Use the font already used in the template, i.e. Arial 10 (and not MyTiniestFont 2).
If any submission aspects could only be determined by running the process, the marks will be severely reduced.

Note: If it is not in this report, it does not exist and does not get marked!

So, we will not check your RapidMiner scripts to check anything that was missing from the report.
Any part which carries points but is missing in the report gets zero marks.
We expect consistency between the report and RapidMiner scripts, so...

Note: Anything reported that cannot be substantiated
by RapidMiner scripts will be marked as zero.

It means that we will check the RapidMiner scripts when in doubt or even just curious.

Total

0 to 100

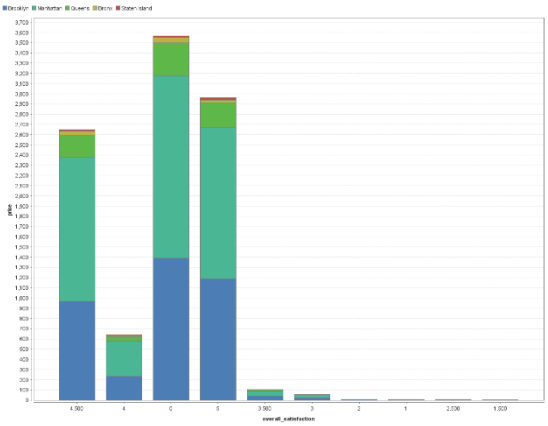
Executive problem statement (one page)

Aim
To figure out if there is any relation in overall experience and price in the Airbnb Database.

Simple Model
Airbnb approached us with their database to figure out some of their questions and wanted some insights with the help of database. In this analysis we have to check if there is any trend related to the attribute's overall satisfaction and price of the Airbnb listings.

Airbnb has listed across 882000 listing in their database. After filtering and making sure of no duplicate value is present, we analyse the data and figure out if there is any relation between price of the listing and the satisfaction value. The data consists of different value and attributes. Attributes such as room id, geo location, min nights to stay, the number of occupants allowed etc. the data is quite large so for the prediction purposes we will have to rationalise the data into a smaller sample size data and get through with the ans. Airbnb listings have a satisfactory attribute which will help us in analysing the data more insightful.

To check if any relation is present regarding the price the listings and the satisfactory rating provided by the occupants. There can be a lot of factors that can be playing a role in overall satisfaction related data analysis.



From the different type of image and illustrations, we can see that there are a lot of missing information in the database. To get corrected we can remove any duplicate, or even remove all the value with no overall satisfaction values. In future we would try to put a specific value in the database for a much better prediction in the database.

In the data base we can even see a lot of missing value and a lot of over all satisfaction rating of 5. Rating of 3 and lower are very scarce and even very less to even predict.From the data we can see that there seems of be a lot of data that is missing. A lot of the attributes that are missing are overall_satisfation and even reviews. From the data analysis we can even figure out that price and overall_satisfaction have a trend which result in a better rating of the listing overall. the data analysis showed me that the less the min stay of a listing and the better pricing of the Airbnb listing is a very good factor which plays in have a overall good satisfaction rating. The neighbourhood attributes played some role in the data analysis too. If suburb such as Staten island has a listing which is very expensive might not get any overall rating whatsoever. But if a suburb such as Brooklyn has a listing which has a good affordable prize will have a good rating, review and a better overall satisfaction.

Data preparation (one page)

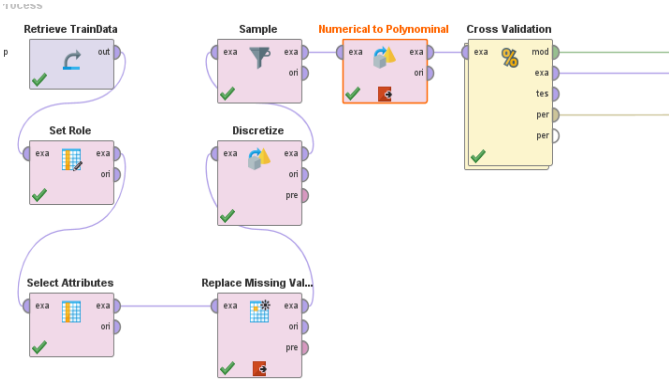
Aim
To demonstrate my understanding of the data analysis and my report of the Airbnb data.

Simple Model
In the part of data exploration, we have to check all the proper data exploration and data analysis part of the finding. We first need to label and use specific attributes into prediction.

Name	Type	Missing	Statistics
Label **overall_satisfaction**	Real	250359	Min 0 Max 5
room_id	Integer	0	Min 105 Max 17732388
host_id	Integer	280	Min 43 Max 120885348

As we can see from the following figure, there are a lot of missing value in the attributes such as review and even overall satisfaction. To solve that we have/ can replace the missing value with zero or even impute some random and appropriate values. We could also eliminate all the attributes which are essential but would not be correctly predicted. So just to solve the problem with as much dtaa as possible we can impute a number which will not alter our data and its prediction.

From the following above illustrations, we can see the data analysis I have conducted. In my analysis I have taken the overall_satistifaction attributes as my label. From my analysis we can see that price and even min stay can be said to be playing a very important role in the overall satisfactory attributes. Price is another big attributes which plays a big role in it.



Row No.	overall_sati...	room_id	reviews	latitude	longitude	price	borough	neighborhood
1	5	9361	4	40.684	-73.964	low	Brooklyn	Clinton Hill
2	4	15796	19	40.682	-73.959	low	Brooklyn	Clinton Hill
3	5	18152	55	40.770	-73.951	very low	Manhattan	Upper East S...
4	5	22589	23	40.778	-73.985	medium	Manhattan	Upper West ...
5	5	44973	40	40.707	-73.953	low	Brooklyn	Williamsburg
6	1	45940	0	40.716	-73.959	low	Brooklyn	Williamsburg
7	5	60680	18	40.727	-73.980	low	Manhattan	East Village
8	5	64982	19	40.730	-73.982	medium	Manhattan	East Village
9	5	71010	2	40.835	-73.939	medium	Manhattan	Washington ...
10	5	72815	73	40.716	-73.996	high	Manhattan	Chinatown
11	5	82946	15	40.785	-73.979	very low	Manhattan	Upper West ...
12	5	83378	32	40.745	-74.002	medium high	Manhattan	Chelsea
13	5	83760	41	40.720	-73.997	low	Manhattan	Little Italy
14	5	84744	39	40.712	-73.965	low	Brooklyn	Williamsburg
15	1	85429	0	40.686	-73.959	very low	Brooklyn	Bedford-Stuyv...

The following illustration shows us the accuracy and even the kappa of our data analysis. We have a respectable accuracy of 88.96 and kappa just shy of 0.79. from this we can visualize all the attributes and even figure out the result we would like to seek. Which is does the price and overall satisfaction have any trend.

kappa: 0.783 +/- 0.007 (micro average: 0.783)

	true 1	true 2	true 3	true 4	true 5	tr
pred. 1	3454	2	7	31	135	0
pred. 2	0	0	0	1	0	0
pred. 3	1	0	0	1	8	0

accuracy: 88.76% +/- 0.36% (micro average: 88.76%)

	true 1	true 2	true 3	true 4
pred. 1	3454	2	7	31
pred. 2	0	0	0	1
pred. 3	1	0	0	1
pred. 4	12	3	2	21

Executive solution statement (one page)

Aim
To clearly articulate your understanding of the business solution to management.

Simple Model
The business solution is succinctly described for executives and justified. Cross-references with the technical sections of the report must be provided for support, e.g. to tables and plots.
Answers to business questions (B) and (C) are given and justified.

Complex Model
In addition, business decisions and actions that are supported by the complex solution must be explained.

Hints
Ensure that whatever problem you describe can be solved using the provided data.
Make sure the exec summary describes the solution from the business perspective and not a technical perspective.
Use business language and not computer / mathematical / statistical / data science language.
The solution statement should describe the high level benefit and not the methods of their delivery.
Think and state who the company clients are and what the likely benefits of this project are for them.
Ensure that your solution clearly matches the problem statement.
Ensure that the solution is formulated in terms of achieving the high-level aim.
Do not include any charts or tables in the solution statement section.
However, cross-reference your problem statement with tables or charts from the following section, e.g. you can refer to them as "… (see Figure 1)" or "As shown in Table 4…".

If you need to support your statements / analysis / argument with references to any published materials, use Harvard citation style as described in: http://www.deakin.edu.au/students/studying/study-support/referencing/harvard. As the executive summary should not take even one page, we suggest to include your bibliographic references at the bottom of this page, immediately below the executive summary (or problem description).

All comments, such as this, which are not part of your submission should be deleted to save space.

Data Preparation and Exploration (first page of a two page section)

Aim
To demonstrate your understanding of data by describing complex relationships between attributes.
Depending on the selected model some attributes may need to be transformed or new attributes created.

Simple Model
Use clustering and segmentation analysis to investigate customer satisfaction. Investigate different number of clusters. All charts and tables must have captions and be annotated (with text and arrows) to highlight important insights. ***Provide support for answer to question B.***

Complex Model
In addition, in a complex solution your clustering must be optimised, cluster performance evaluated, anomalies visualised and removed. PCA used for cluster and anomaly visualisation. Selected nominal attributes are dummy (or as unique integer) encoded, dimensionality of data reduced for predictive modelling.

Hints
Many hints are identical to those in the section on “Data Exploration” so read them!
Some preliminary data exploration has already been conducted in the previous sections.
Focus on depicting attributes relationships as emerging from cluster and segmentation analysis and not from other investigation of attribute characteristics,
Include here the text of your analysis with tables and charts.

Avoid indiscriminate “dumping” of tables, charts or code into this section – all content must have some purpose.
All included charts, tables or RM processes (or their parts) have to be described or used in the discussion.
Make sure that all charts, tables and important results are labelled for cross-referencing, e.g. “Figure 1 - Histogram of Overall Rating” or “Table 4 – Comparison of model performance”.

All comments, such as this, which are not part of your submission should be deleted to save space.

Data Preparation and Exploration (Second page of a two page section)

Continuation of the previous page.

Model Development (one page)

Aim
To explain details of developed predictive models and selected methods for data preparation and reporting.

Simple Model
Create two estimation models, i.e. linear regression and decision tree, to support answering question C.
Operator aims/roles and their main parameters must be explained (with text annotations and arrows).

Complex Model
In addition, in a complex solution, an ensemble should be included as the third estimation model. Other models may be included. Cluster attributes need to be used in model development, e.g. as predictors.

Hints
Your textbook will be extremely helpful in this task.
Include here screenshots of all or parts of the RM process or multiple processes.
If your process is very large, consider splitting it into sub-processes or separate processes.
If your process does not fit into this page, include only the most important parts.
Focus on the key components of your process(es).
By including arrows and text boxes (e.g. with numbers to refer to) annotate each operator and its properties.
Note that some of your justifications may utilise cross-referencing with tables or charts from other sections.
Avoid indiscriminate “dumping” of RM processes/models into this section – all content must have some purpose.
You may include a brief description of the operators and what they did but this is NOT the aim of this section.
Do not include definition of terms or a “textbook” description of operations – we already know this!

All comments, such as this, which are not part of your submission should be deleted to save space.

Model Evaluation and Optimisation (first page of a two page section)

Aim
To report and explain the performance of developed predictive models.

Simple Model
Cross-validate predictive models in terms of RMSE, MAE and r2. Include honest testing. Experiment with model parameters. Tabulate and compare performance of different models with different parameters and select the best.
Provide support for question C.

Complex Model
In addition, in a complex solution, optimise the best predictive model, e.g. based on ridge regression or decision tree height (and other parameters). Compare performance of all models.

Hints
Your textbook will be extremely helpful in this task.
If you have few results to report, include here screenshots of your results.
If you have many results to report, include here a table of all results.
You need to describe and explain your results.
It is the most important that you include here the detailed analysis of your results –
explain the impact of the obtained results on the future use of the model to support decision making.
Avoid indiscriminate "dumping" of performance results – all content must have some purpose.

All comments, such as this, which are not part of your submission should be deleted to save space.

Model Evaluation and Optimisation (second page of a two page section)

Continuation of the previous page.

Model Application (one page)

Aim
To demonstrate and explain how the developed analytic processes can be applied to new data.

Simple Model
Apply the best model to new data and investigate results. All results must be analysed, interpreted and reported. A statement is to be included with justification as to what degree the model advice can actually be trusted.

Complex Model
In addition, in a complex solution ensure that all pre-processing and predictive models from training are correctly applied to new data. Also apply your best model to a single example and describe the result.

Hints
Your textbook will be extremely helpful in this task.
If you have few results to report, include here screenshots of your results.
If you have many results to report, include here a table of all results.
You need to describe and explain your results.
It is the most important that you include here the detailed analysis of your results –
explain the impact of the obtained results on the future use of the model to support decision making.
Avoid indiscriminate "dumping" of performance results – all content must have some purpose.

All comments, such as this, which are not part of your submission should be deleted to save space.

Any materials, analysis or reports that do not fit into 10 pages (including the front page) will not be looked at or marked.