ASSESSMENTCOVERSHEET

Attach this coversheet as the cover of your submission. All sections must be completed.

Section A: Submission Details

Programme : BIOT

Course Code & Name : IIB30104- DATA ANALYTIC

Course Lecturer(s) : MADAM ADIDAH

Submission Title : GROUP PROJECT

Deadline : Day 8 Month 6 Year 2024 Time 3:16pm

Penalties :
- 5% will be deducted per day to a maximum of four (4) working days, after which the submission will **not** be accepted.
- Plagiarised work is an Academic Offence in University Rules & Regulations and will be penalised accordingly.

Section B: Academic Integrity

Tick (√) each box below if you agree:

- [/] I have read and understood the UniKL's policy on Plagiarism in University Rules & Regulations.
- [/] This submission is my own, unless indicated with proper referencing.
- [/] This submission has not been previously submitted or published.
- [/] This submission follows the requirements stated in the course.

Section C: Submission Receipt

(Must be filled in manually)

Office Receipt of Submission

Date & Time of Submission (stamp)	Student Name(s)	Student ID(s)
Friday 3:16pm	Muhamad Haziq Azfar bin Abdul Rahim Afiq Hazim bin Azaddin	52224122273 52224122132

Student Receipt of Submission

This is your submission receipt, the only accepted evidence that you have submitted your work. After this is stamped by the appointed staff & filled in, cut along the dotted lines above & retain this for your record.

Date & Time of Submission (stamp)	Course Code	Submission Title	Student ID(s) & Signature(s)
Friday 3:16pm	IIB30104	GROUP PROJECT	52224122273 52224122132

UNIVERSITI KUALA LUMPUR

MALAYSIAN INSTITUTE OF INFORMATION TECHNOLOGY

Name of Course	**DATA ANALYTIC**
Course Code	**IIB30104**
Lecturer	**AP DR ADIDAH LAJIS**
Semester / Year	**MARCH 2024 (1/2024)**
Date	**18 MAY 2024**

Assessment	**PROJECT**
Weightage	**20%**
Course Outcome to achieve: 1. Develop a customer behavior model based on affinity-based marketing use cases. (C3, PLO4)	

Scenario

BOOK.COM employs you. You are assigned to study the data for one of the holiday stays belonging to them to answer the questions stated in Task 2. The details of the task are as shown below. BOOK.COM has numbers of holiday stays throughout the world especially in Europe.

Task 1

You are given a set of 3 data namely as stated below.

No	Dataset	Description
1	Listings	Details on listing features like location, amenities, reviews and others
2	Calendar	Availability calendar of listing over a one-year period
3	Reviews	Reviews posted by customers after their stay

Task 2

You are required to apply the CRISP-DM methodology. You are required to analyze the dataset to answer the following questions.

a. How do listing prices change across locations and time?
b. What are the main drivers of listing prices?
c. What features of listings influence customer satisfaction the most?

Task 3

Describe the process in Phase 1: Business Problem Understanding.

The first phase of the CRISP-DM methodology is critical as it lays the foundation for the entire data. This phase involves comprehending the project objectives and requirements from a business perspective and converting this knowledge into a data mining problem definition and a preliminary plan. This includes process for this phase:

a) Determine Business Objectives

- **Background**: Understand the overall context and background of the business problem.
- **Business Objectives**: Clearly define the business objectives. These are the goals that the business wants to achieve, such as improving pricing strategies or increasing customer satisfaction.
- **Business Success Criteria**: Identify the criteria that will determine whether the business objectives have been met. This might include specific metrics like increased revenue, higher customer satisfaction scores, etc.

b) Assess Situation

- **Inventory of Resources:** List all available resources, including data, technology, expertise, and any other relevant assets.
- **Requirements, Assumptions & Constraints:** Document any specific requirements (both business and data), assumptions (e.g., data availability, quality), and constraints (e.g., budget, time, regulations).
- **Risks and Contingencies:** Identify potential risks that could impact the project (e.g., data privacy issues, changing market conditions) and plan contingencies to mitigate these risks.

- **Terminology:** Define any specific terminology that will be used in the project to ensure a common understanding among all stakeholders.
- **Costs and Benefits:** Estimate the costs involved in the project and the expected benefits. This helps in assessing the feasibility and value of the project.

c) Determine Data Mining Goals

- **Data Mining Goals:** Translate the business objectives into specific data mining goals. These goals should be clear and measurable, such as identifying factors influencing listing prices or determining features affecting customer satisfaction.
- **Data Mining Success Criteria:** Define the criteria for success from a data mining perspective. This might include accuracy of predictive models, insights derived, or the ability to generalize findings to new data.

d) Produce Project Plan

- **Project Plan:** Develop a detailed project plan that outlines the steps required to achieve the data mining goals. This plan should include tasks, timelines, milestones, and responsibilities.
- **Initial Assessment of Tools and Techniques:** Identify the tools and techniques that will be used in the project. This could include data analysis software, statistical methods, machine learning algorithms, etc.

Task 4

Describe the process in Phase 2: Data Understanding.

In this phase, you can form a better understanding of the data if you able to link the attribute meanings to the data. Provide the summary report of the data such as;

- Number of data and attribute
- The data type for each attribute
- Number of unique values for each attribute
- Number of missing values for each attribute
- Summary statistics like mean, median, and standard deviation for numeric attribute
- The date range for the date attribute

The data type for each attribute

Calendar Austin

No.	Attributes	Data type
1	listing_id	Integer
2	date	Date-time
3	available	Nominal
4	price	Nominal
5	adjusted_price	Nominal
6	minimum_nights	Integer
7	maximum_nights	Integer

Listing Austin

No.	Attributes	Data type
1	id	Nominal
2	Listing_url	Nominal
3	Scrape_id	Nominal
4	Last_scraped	Nominal
5	Source	Nominal
6	Name	Nominal
7	description	Nominal
8	Neighborhood_overview	Nominal
9	Picture_url	Nominal
10	Host_id	Integer

11	Host_url	Nominal
12	Host_name	Nominal
13	Host_since	Nominal
14	Host_location	Nominal
15	Host_about	Nominal
16	Host_reponse_time	Nominal
17	Host_reponse_rate	Nominal
18	Host_acceptance_rate	Nominal
19	Host_is_superhost	Nominal
20	Host_thumbnail_url	Nominal
21	Host_picture_url	Nominal
22	Host_neighbourhood	Nominal
23	Host_listings_count	Integer
24	Host_total_listings_count	Integer
25	Host_verications	Nominal
26	Host_has_profile_pic	Nominal
27	Host_identity_verified	Nominal
28	neighbourhood	Nominal
29	Neighbourhood_cleansed	Integer
30	Neighbourhood_group_cleansed	Nominal
31	latitude	Real
32	Longitude	Real
33	Property_type	Nominal
34	Room_type	Nominal
35	Accommodates	Integer
36	Bathrooms	Nominal
37	Bathrooms_text	Nominal
38	Bedrooms	Integer
39	beds	Integer
40	Amenities	Nominal
41	Price	Nominal
42	Minimum_nights	Integer
43	Maximum_nights	Integer
44	Minimum_minimum_nights	Integer
45	Maximum_minimum_nights	Integer
46	Minimum_maximum_nights	Integer
47	Maximum_maximum_nights	Integer
48	Minimum_nights_avg_ntm	Real
49	Maximum_nights_avg_ntm	Real
50	Calendar_updated	Nominal
51	Has_availability	Nominal
52	Availability_30	Integer
53	Availability_60	Integer
54	Availability_90	Integer
55	Availability_365	Integer
56	Calendar_last_scraped	Nominal
57	Number_of_reviews	Integer
58	Number_of_reviews_ltm	Integer
59	Number_of_reviews_l30d	Integer
60	First_review	Nominal

61	Last_review	Nominal
62	Review_scores_rating	Real
63	Review_scores_accuracy	Real
64	Review_scores_cleanliness	Real
65	Review_scores_checkin	Real
66	Review_scores_communication	Real
67	Review_scores_location	Real
68	Review_scores_value	Real
69	License	Nominal
70	Instant_bookable	Nominal
71	Calculated_host_listings_count	Integer
72	Calculated_host_listings_count_entire_homes	Integer
73	Calculated_host_listings_count_private_rooms	Integer
74	Calculated_host_listings_count_shared_rooms	Integer
75	Reviews_per_month	Real

Review Austin

No	Attributes	Data Type
1	Listing_id	Nominal
2	Id	Nominal
3	Date	Nominal
4	Reviewer_id	Integer
5	Reviewer_name	Nominal
6	comments	Nominal

Number of unique values

Unique values review

No	Attributes	Number of unique values
1	Listing_id	19,643
2	Id	476,945
3	Date	6243
4	Reviewer_id	409,991
5	Reviewer_name	44,097
6	comments	403,294

Unique values Calendar

No.	Attributes	Number of unique values
1	listing_id	14,861
2	date	366
3	available	2
4	price	4809
5	adjusted_price	4806
6	minimum_nights	78
7	maximum_nights	186

Unique values Listing

No.	Attributes	Number of unique values
1	id	10,784
2	Listing_url	9915
3	Scrape_id	641
4	Last_scraped	406
5	Source	276
6	Name	3697
7	description	8265
8	Neighborhood_overview	4064
9	Picture_url	8794
10	Host_id	5984
11	Host_url	5999
12	Host_name	2550
13	Host_since	3071
14	Host_location	413
15	Host_about	2595
16	Host_reponse_time	9
17	Host_reponse_rate	53
18	Host_acceptance_rate	94
19	Host_is_superhost	3
20	Host_thumbnail_url	5864
21	Host_picture_url	5864
22	Host_neighbourhood	520

23	Host_listings_count	91
24	Host_total_listings_count	107
25	Host_verications	7
26	Host_has_profile_pic	2
27	Host_identity_verified	2
28	neighbourhood	15
29	Neighbourhood_cleansed	45
30	Neighbourhood_group_cleansed	0
31	latitude	7580
32	Longitude	7386
33	Property_type	66
34	Room_type	4
35	Accommodates	17
36	Bathrooms	0
37	Bathrooms_text	34
38	Bedrooms	16
39	beds	30
40	Amenities	8142
41	Price	850
42	Minimum_nights	55
43	Maximum_nights	55
44	Minimum_minimum_nights	58
45	Maximum_minimum_nights	58
46	Minimum_maximum_nights	110
47	Maximum_maximum_nights	110
48	Minimum_nights_avg_ntm	195
49	Maximum_nights_avg_ntm	207
50	Calendar_updated	0
51	Has_availability	2
52	Availability_30	30
53	Availability_60	60
54	Availability_90	90
55	Availability_365	365
56	Calendar_last_scraped	2
57	Number_of_reviews	0
58	Number_of_reviews_ltm	237
59	Number_of_reviews_l30d	19
60	First_review	2155
61	Last_review	1246
62	Review_scores_rating	133
63	Review_scores_accuracy	132
64	Review_scores_cleanliness	132
65	Review_scores_checkin	132
66	Review_scores_communication	133
67	Review_scores_location	132
68	Review_scores_value	132
69	License	0
70	Instant_bookable	2
71	Calculated_host_listings_count	38
72	Calculated_host_listings_count_entire_homes	36

73	Calculated_host_listings_count_private_rooms	17
74	Calculated_host_listings_count_shared_rooms	6
75	Reviews_per_month	663

Missing values

Calendar Austin

No	Attributes	Missing values
1	listing_id	0
2	date	0
3	available	0
4	price	8
5	adjusted_price	8
6	minimum_nights	2
7	maximum_nights	2

Listing Austin

No.	Attributes	Missing values
1	id	0
2	Listing_url	2468
3	Scrape_id	3191
4	Last_scraped	3747
5	Source	4307
6	Name	4859
7	description	5122
8	Neighborhood_overview	9163
9	Picture_url	5228
10	Host_id	5470
11	Host_url	5423
12	Host_name	5429
13	Host_since	5462
14	Host_location	7023
15	Host_about	10028
16	Host_reponse_time	5471
17	Host_reponse_rate	5471
18	Host_acceptance_rate	5471
19	Host_is_superhost	5660
20	Host_thumbnail_url	5475
21	Host_picture_url	5475
22	Host_neighbourhood	6731
23	Host_listings_count	5475
24	Host_total_listings_count	5475
25	Host_verications	5473
26	Host_has_profile_pic	5475
27	Host_identity_verified	5475
28	neighbourhood	9487
29	Neighbourhood_cleansed	5473
30	Neighbourhood_group_cleansed	14428
31	latitude	5473
32	Longitude	5473

33	Property_type	5473
34	Room_type	5473
35	Accommodates	5473
36	Bathrooms	14428
37	Bathrooms_text	5481
38	Bedrooms	7468
39	beds	5548
40	Amenities	5473
41	Price	5473
42	Minimum_nights	5473
43	Maximum_nights	5473
44	Minimum_minimum_nights	5473
45	Maximum_minimum_nights	5473
46	Minimum_maximum_nights	5473
47	Maximum_maximum_nights	5473
48	Minimum_nights_avg_ntm	5473
49	Maximum_nights_avg_ntm	5473
50	Calendar_updated	14428
51	Has_availability	5473
52	Availability_30	5473
53	Availability_60	5473
54	Availability_90	5473
55	Availability_365	5473
56	Calendar_last_scraped	5473
57	Number_of_reviews	5473
58	Number_of_reviews_ltm	5473
59	Number_of_reviews_l30d	5473
60	First_review	7849
61	Last_review	7849
62	Review_scores_rating	7849
63	Review_scores_accuracy	7903
64	Review_scores_cleanliness	7903
65	Review_scores_checkin	7904
66	Review_scores_communication	7903
67	Review_scores_location	7904
68	Review_scores_value	7904
69	License	14428
70	Instant_bookable	5473
71	Calculated_host_listings_count	5473
72	Calculated_host_listings_count_entire_homes	5473
73	Calculated_host_listings_count_private_rooms	5473
74	Calculated_host_listings_count_shared_rooms	5473
75	Reviews_per_month	7849

Review Austin

No	Attributes	Data Type
1	Listing_id	0

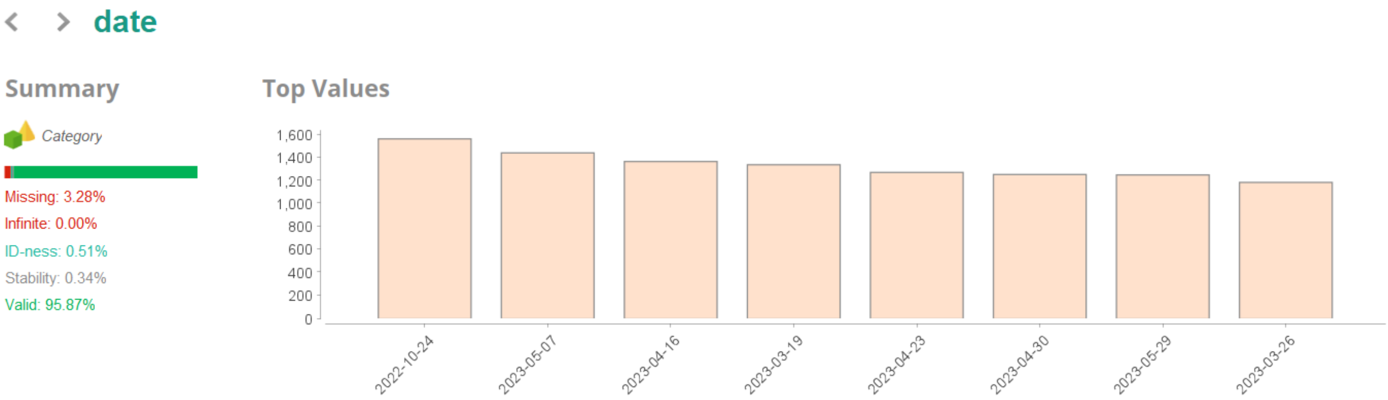
2	Id	14311
3	Date	16139
4	Reviewer_id	18139
5	Reviewer_name	17583
6	comments	17871

<u>Summary statistics like mean, median, and standard deviation for numeric attribute</u>

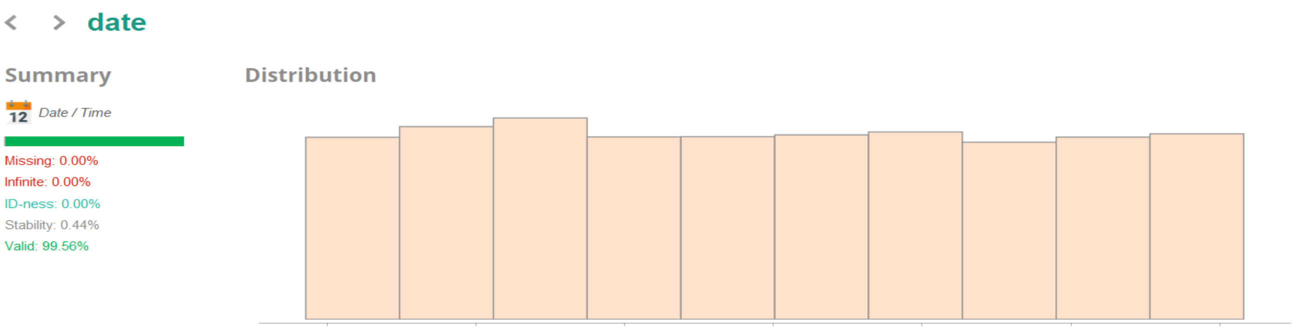
No.	Dataset	Numeric attribute	Mean	Median	Standard Deviation
1.	Review AUSTIN	Reviewer_id	158,545,831.803	114,071,541	142,467,882.816
2.	Calendar AUSTIN	Listing_id	341,896,273,454,300,220	517,68641	395,191,700,179,944,770
3.	Listings AUSTIN	Host_id	152,211,772.222	747,76943	164,772,739.260

<u>The date range for the date attribute</u>

Review AUSTIN DATASET:



Calendar AUSTIN DATASET:



Statistics

Name	Value
From	Sep 10, 2023
Until	Sep 9, 2024
Duration	365 days

Task 5

Describe the process in Phase 3: Data Preparation

Data preparation is the whole gamut of manipulations and transformations performed on data to tackle inconsistencies like missing values or outliers, convert columns to the right format, or enrich data with new features.

The summary report created in Task 4 helps to identify data inconsistencies like incorrect data types and missing values.

Below are the key data preparation steps that are usually applied to any dataset

- To convert columns to correct data types.
- drop identical duplicate rows.
- drop columns with constant values.
- impute missing values.
- encode categorical features.

Finally, perform Exploratory Data Analysis on the prepared data and report your findings accordingly.

Phase 3: Data Preparation

Data preparation is the next phase in the CRISP-DM process which is interconnected and critical for preparing the data to be ready for modelling. Proper data preparation ensures that the subsequent steps, such as modelling, evaluation, and deployment, are based on high-quality, well-understood, and appropriately formatted data. Steps involved are:

1. **Dataset**

Dataset Description: Provide a comprehensive description of the dataset, including its source, contents, and structure.

2. **Select Data**

Rationale for Inclusion/Exclusion: Decide which data will be included or excluded from the analysis and document the reasons for these decisions.

3. Clean Data

Data Cleaning Report: Perform data cleaning tasks such as handling missing values, correcting errors, and removing duplicates. Document the cleaning steps and results in a report.

4. Construct Data

Derived Attributes: Create new attributes (features) from the existing data that may help in the analysis.

Generated Records: Generate new records, if necessary, through techniques such as data augmentation or synthetic data generation.

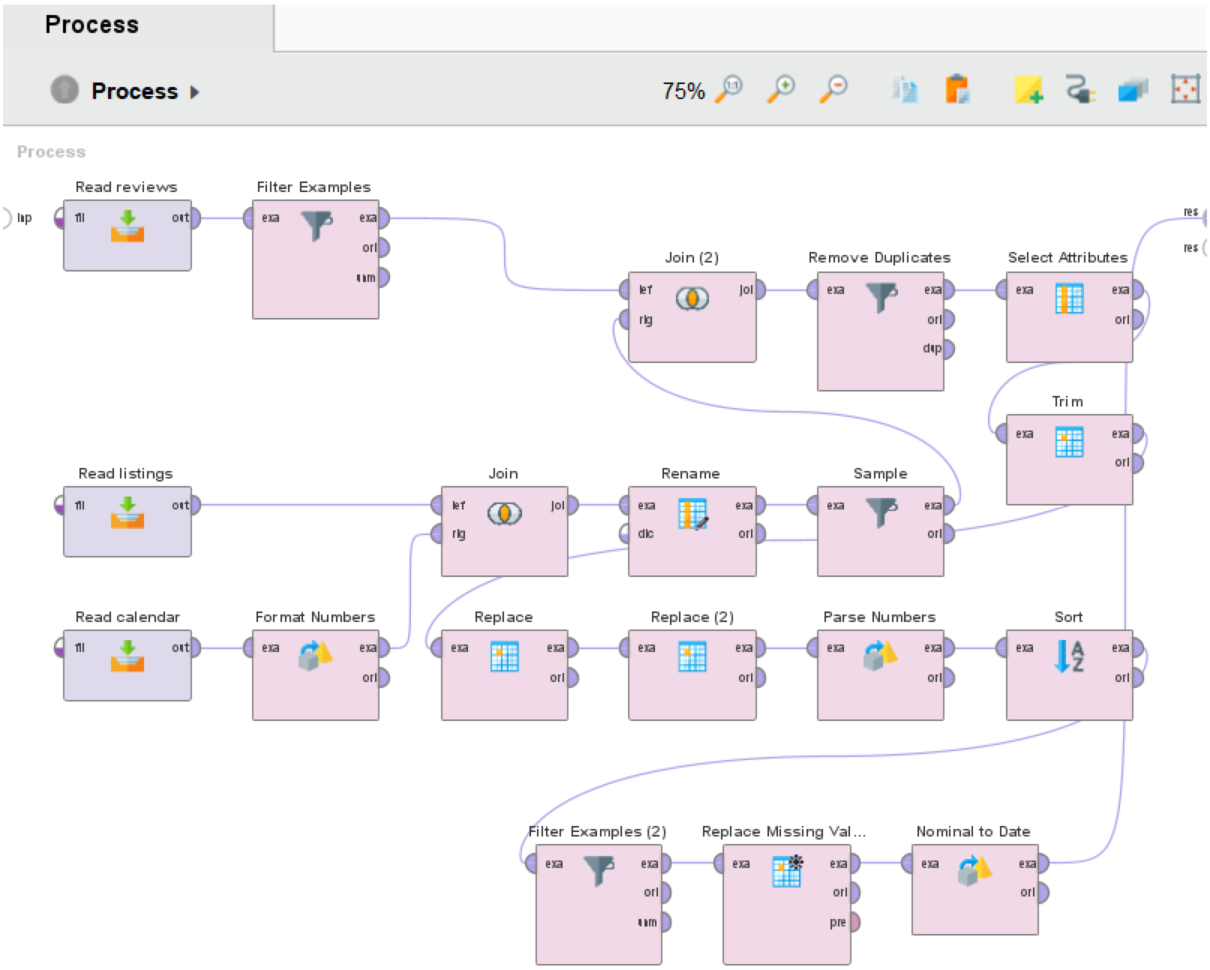
5. Integrate Data

Merged Data: Combine data from different sources or tables to create a cohesive dataset for analysis.

6. Format Data

Reformatted Data: Ensure that the data is in the appropriate format required for analysis or modeling. This may involve transformations such as scaling, normalization, or encoding categorical variables.

Data preparation process in RapidMiner:



Exploratory Data Analysis:

After accomplishing the data preparation process, we managed to handle all missing data which includes the duplicate ones. Furthermore, we also removed all unnecessary symbols within the dataset to as it could provide misinterpretation. Data preparation is a crucial step in the data analysis and machine learning process for several reasons which includes:

1) Handling Missing Values: Missing data can lead to incorrect analyses and models. Data preparation involves identifying and handling missing values appropriately.
2) Removing Duplicates: Duplicate records can skew results and lead to biases in your models.

Correcting Errors: Data entry errors, inconsistencies, and anomalies need to be corrected to ensure the data accurately represents the real-world scenario.

Filter (9 / 9 attributes): Search for Attributes

Name	Type	Missing	Statistics		
price	Numeric	0	Min 1	Max 19107	Average 231.809
listing_id	Polynominal	0	Least ì§'ì—□ ë [...] ˆë‹¤. (0)	Most 44334720 (818)	Values 44334720 (818), 4974255 (783), ...[1(
id	Polynominal	0	Least ì□ ´ì›ƒ ì [...] ‹¤ë§Œ (0)	Most 206052641 (37219)	Values 206052641 (37219), 100010969 (1), .
host_location	Polynominal	0	Least the Mediterranean (0)	Most Austin, TX (191591)	Values Austin, TX (191591), San Francisco, (
room_type	Polynominal	0	Least Hotel room (86)	Most Entire home/apt (186626)	Values Entire home/apt (186626), Private roo
amenities	Polynominal	0	Least ["Wirele [...] ron"] (0)	Most ["Air co [...] "] (1504)	Values ["Air co [...] , "Pool"] (1504), ["Air co [..
available	Polynominal	0	Least t (73027)	Most f (147891)	Values f (147891), t (73027)
review_scores_rating	Real	0	Min 0	Max 5	Average 4.843
date	Date time	0	Earliest date Mar 8, 2009 12:00 AM	Latest date Sep 10, 2023 12:00 AM	Duration 5299d 0h 0m 0s

Data visualisation examples:

We are also able to analyse the data visualization in line graph to observe the number review scores rating in each average price range. Data visualization is important in data analysis and decision-making. This is because:

1) Simplifies Complex Data

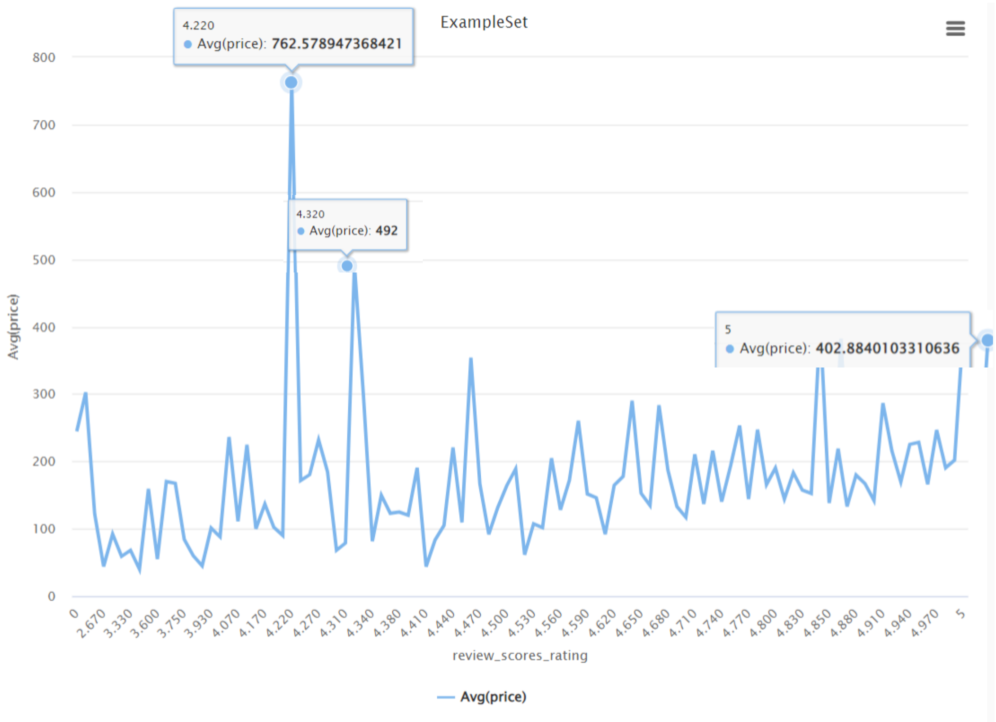
   Understanding Trends and Patterns: Visualization helps in quickly identifying trends, patterns, and anomalies within large and complex datasets.

   Making Data Accessible: Converts complex data into a visual format that is easier to understand, even for non-technical stakeholders.

2) Enhances Data Interpretation

   Clarity: Visual representations can make data more intuitive and straightforward, allowing for easier interpretation.

   Context: Provides context by showing data in relation to other data points, which can be critical for understanding the bigger picture.



As we can observe in the above figure, we can analyse that rating 4.22 is the highest score rating among other rooms in 762.57 price range. Followed by rating 4.32 in price range of 492 while for 5 score rating is on 402.88 price range from the rest. From this analysis, we can learn from other consumers based on their ratings that not all expensive rooms has higher score ratings than the cheaper ones.

Task 6

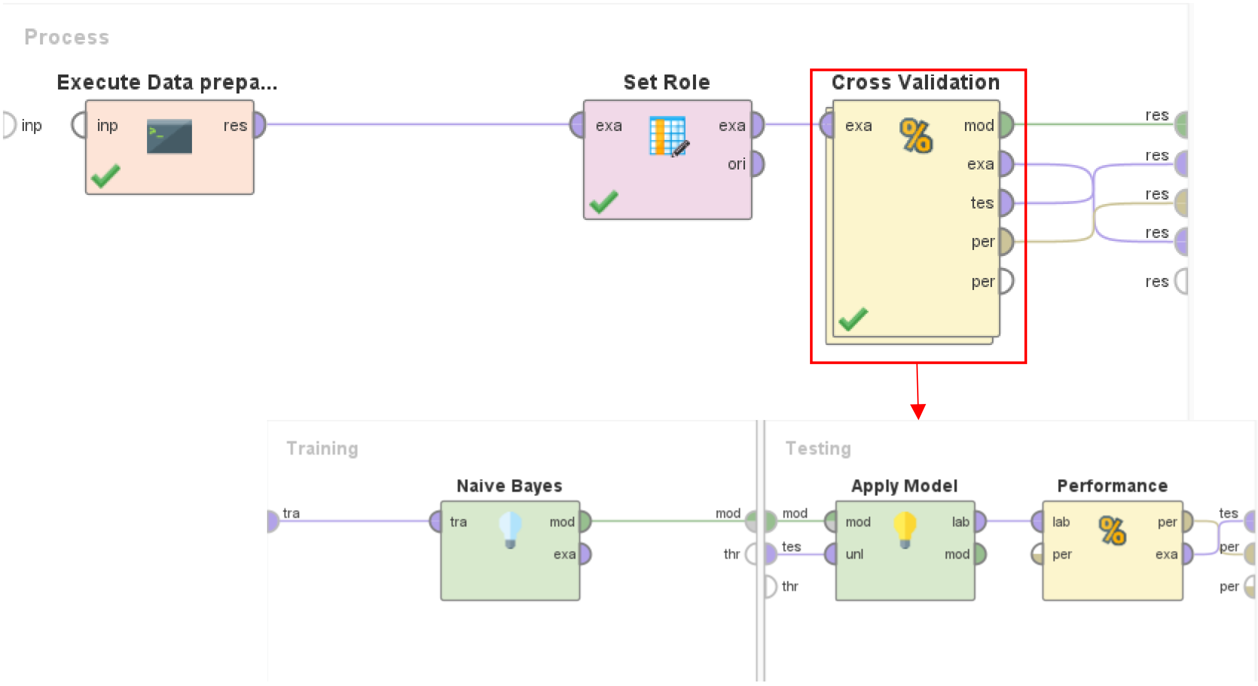
Describe the process in Phase 4: Modelling and Evaluation

Depending on the problem you are solving the solution could be framed as a machine learning model. The quantity of interest is usually treated as a target variable or other information as a predictor variable.

Clearly set the target variable and justify the selection.

Perform a Model Validation based on the requirement and justify the selection of the model accordingly.

Model Validation process:



Based on the above diagram, we used Naïve Bayes as our model validation because Naive Bayes is a simple yet effective classification algorithm. By using Naive Bayes, we can quickly get an idea of how well our data can be modelled and classified. This helps in setting expectations for more complex models. Naive Bayes can also handle both categorical and continuous data, making it versatile for various types of datasets which is suitable for our datasets that must process numerous amounts of data. Furthermore, Naive Bayes can produce probabilistic outputs that can be easily interpreted. This helps in understanding the likelihood of different outcomes and the relative importance of features.

Table View Plot View

accuracy: 96.66% +/- 6.26% (micro average: 96.66%)

	true Entire home/apt	true Private room	true Shared room	true Hotel room	class precision
pred. Entire home/apt	181223	12	1	0	99.99%
pred. Private room	336	31136	2	2	98.92%
pred. Shared room	439	166	1090	0	64.31%
pred. Hotel room	4628	1715	84	84	1.29%
class recall	97.10%	94.27%	92.61%	97.67%	

The confusion matrix and performance metrics shown in the above diagram provide a detailed evaluation of a classification model using Naïve Bayes. The breakdown of the information is:

• Accuracy: 96.66% ± 6.26%

- Indicates the overall correctness of the model. A high accuracy suggests that the model is performing well in general.

• Micro Average: 96.66%

- The micro-average metric aggregates the contributions of all classes to compute the average metric. It is often used when there is an imbalance in the class distribution.

The confusion matrix from the diagram shows the true versus predicted classifications for four different classes: Entire home/apt, Private room, Shared room, and Hotel room.

Precision indicates how many of the predictions for a given class were correct. High precision for "Entire home/apt" and "Private room" suggests the model rarely misclassifies other categories as these. Lower precision for "Shared room" and especially "Hotel room" indicates more false positives in these categories.

Recall measures the ability of the model to identify all relevant instances of a class. High recall across all classes suggests the model is effective at capturing true positives.

Interpretation

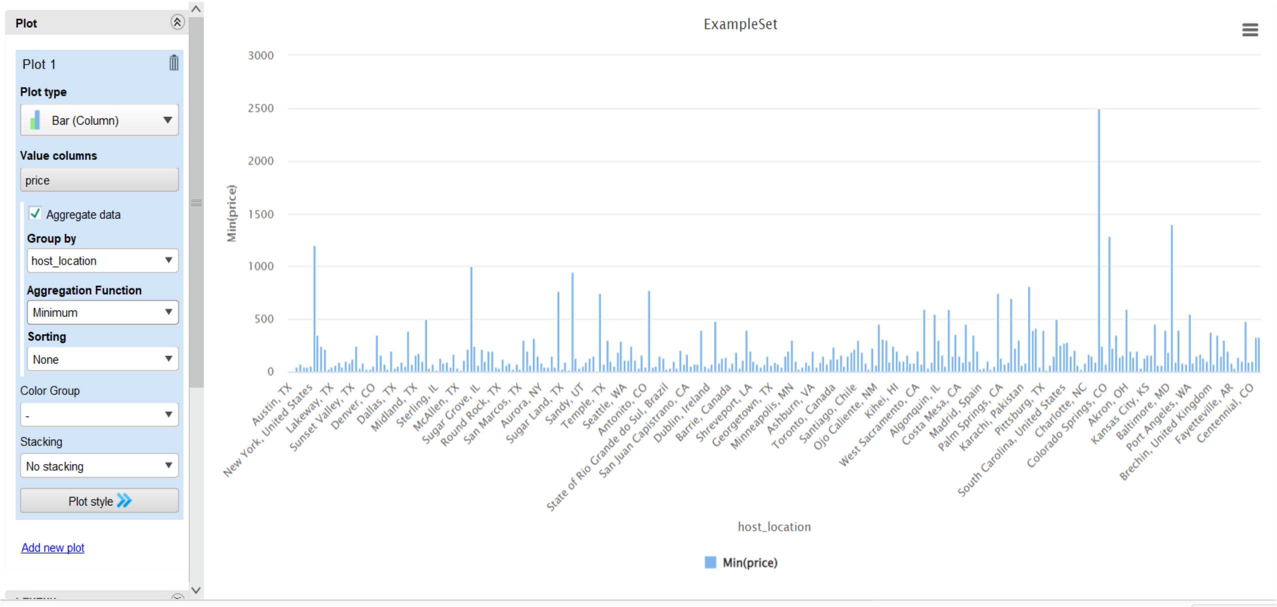
- **Entire home/apt**:
    - High precision (99.99%) and recall (97.10%) indicate that the model performs exceptionally well in identifying listings that are Entire home/apartments.
- **Private room**:
    - Also performs well with high precision (98.92%) and recall (94.27%), meaning predictions for Private rooms are usually correct and most true Private rooms are identified.
- **Shared room**:
    - Moderate recall (92.61%) but lower precision (64.31%) suggests the model captures most Shared rooms but has many false positives.
- **Hotel room**:
    - The very low precision (1.29%) despite high recall (97.67%) indicates that while the model finds almost all Hotel rooms, it frequently misclassifies other types as hotel rooms.

The confusion matrix and associated metrics highlight the strengths and weaknesses of the classification model. The high overall accuracy and micro-average suggest that the model is performing well in general. However, precision and recall for specific classes, especially "Shared room" and "Hotel room," indicate areas for improvement.

Task 7

Based on the model answer the questions stated in TASK 2

a. How do listing prices change across locations and time?



Based on the bar graph above, the listing prices change depending on the location. Nevertheless, not all location has the same listing price as the others. In some locations, it is cheaper to purchase probably because of high populated area factors. Based on our gathered analysis:

- **Price Variation by Location**:
  - The minimum listing prices vary significantly across different locations.
  - Some locations exhibit much higher minimum prices compared to others, indicating a higher baseline cost for listings in these areas.
- **Notable Locations with High Minimum Prices**:
  - Locations such as Austin, TX, Seattle, WA, and New York, NY, appear to have relatively higher minimum listing prices.
  - The spikes in the chart suggest that these cities have a higher cost of entry for listings, possibly due to higher demand, popularity, or cost of living.
- **Locations with Lower Minimum Prices**:
  - On the other hand, there are many locations with lower minimum prices, suggesting more affordable entry-level listings.
  - These might be smaller cities or locations with less demand.

b. What are the main drivers of listing prices?

Performance | Description | Annotations

Criterion: accuracy, kappa

Table View | Plot View

accuracy: 93.72% +/- 1.51% (micro average: 93.72%)

	true Austi...	true Hous...	true Euge...	true Faye...	true Segu...	true Palm...	true Hono...	true Bost...	true Bastr...	true Was...	true Roun...
pred. Aus...	179301	0	0	0	0	0	1	0	0	0	0
pred. Hou...	0	1384	0	0	0	0	0	0	0	0	0
pred. Eug...	4	0	24	0	0	0	0	0	0	0	0
pred. Fay...	0	0	0	3	0	0	0	0	0	0	0
pred. Seg...	6	0	0	0	157	0	0	0	0	0	0
pred. Pal...	0	0	0	0	0	38	0	0	0	0	0
pred. Hon...	7	0	0	0	0	0	34	0	0	0	0
pred. Bos...	4	0	0	0	0	0	0	153	0	0	0
pred. Bas...	125	1	0	0	0	0	0	0	5	0	0
pred. Wa...	0	0	0	0	0	0	0	0	0	237	0
pred. Rou...	0	0	0	0	0	0	0	0	0	0	413
pred. Tex...	0	0	0	0	0	0	0	0	0	0	0
pred. Mari...	0	0	0	0	0	0	0	0	0	0	0
pred. Ne...	0	0	0	0	0	0	0	0	0	0	0
pred. San...	0	0	0	0	0	0	0	0	0	0	0
pred. Lim...	0	0	0	0	0	0	0	0	0	0	0

PerformanceVector (Performance (2)) | ExampleSet (Cross Validation) | SimpleDistribution (Naive Bayes)

Result History | ExampleSet (Set Role)

Performance | Description | Annotations

PerformanceVector

```
PerformanceVector:
accuracy: 93.72% +/- 1.51% (micro average: 93.72%)
ConfusionMatrix:
True:   Austin, TX      Houston, TX     Eugene, OR      Fayetteville, AR        Seguin, TX      Palmer, AK      Honolulu, HI    Boston,
Austin, TX:     179301  0       0       0       0       0       1       0       0       0       0       0       0       0       0
Houston, TX:    0       1384    0       0       0       0       0       0       0       0       0       0       0       0       0
Eugene, OR:     4       0       24      0       0       0       0       0       0       0       0       0       0       0       0
Fayetteville, AR:       0       0       0       3       0       0       0       0       0       0       0       0       0       0
Seguin, TX:     6       0       0       0       157     0       0       0       0       0       0       0       0       0       0
Palmer, AK:     0       0       0       0       0       38      0       0       0       0       0       0       0       0       0
Honolulu, HI:   7       0       0       0       0       0       34      0       0       0       0       0       0       0       0
Boston, MA:     4       0       0       0       0       0       0       153     0       0       0       0       0       0       0
Bastrop, TX:    125     1       0       0       0       0       0       0       5       0       0       2       2       1       0
Washington, DC: 0       0       0       0       0       0       0       0       0       237     0       0       0       0       0
Round Rock, TX: 0       0       0       0       0       0       0       0       0       0       413     0       0       0       0
Texas, United States:   0       0       0       0       0       0       0       0       0       0       0       500     0       0
Marina del Rey, CA:     0       0       0       0       0       0       0       0       0       0       0       0       0       0
New York, NY:   0       0       0       0       0       0       0       0       0       0       0       0       0       1424    0
Santa Rosa, CA: 0       0       0       0       0       0       0       0       0       0       0       0       0       0       13
Lima, Peru:     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Oakland, CA:    291     0       0       0       0       0       0       0       0       0       0       0       0       0       0
San Antonio, TX:        168     5       0       0       0       0       0       0       0       0       0       0       0       0
Thun, Switzerland:      0       0       0       0       0       0       0       0       0       0       0       0       0       0
Denver, CO:     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Minneapolis, MN:        0       0       0       0       0       0       0       0       0       0       0       0       0       0
New York, United States:        2       0       0       0       0       0       0       0       0       0       0       0       0
Edinburg, TX:   183     0       0       0       0       0       0       0       0       0       0       0       0       0       0
Longview, TX:   0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
Manchaca, TX:   138     0       0       0       0       0       0       0       0       0       0       0       0       0       0
New Braunfels, TX:      0       0       0       0       0       0       0       0       0       0       0       0       0       0
```

Observing the data modelling table above, population can significantly affect listing prices and serve as one of the main drivers. Based on our gathered analysis population influences listing prices with:

1) **Demand and Supply Dynamics:**

- **Higher Population**: Cities with larger populations generally have higher demand for housing and accommodations. This increased demand can drive up listing prices.
- **Lower Population**: Smaller towns or cities with lower populations may have less demand, resulting in lower listing prices.

2) **Cost of Living:**
- **Higher Cost of Living:** Populous cities often have a higher cost of living, which can be reflected in higher listing prices.
- **Lower Cost of Living:** Areas with smaller populations usually have a lower cost of living, leading to more affordable listing prices.

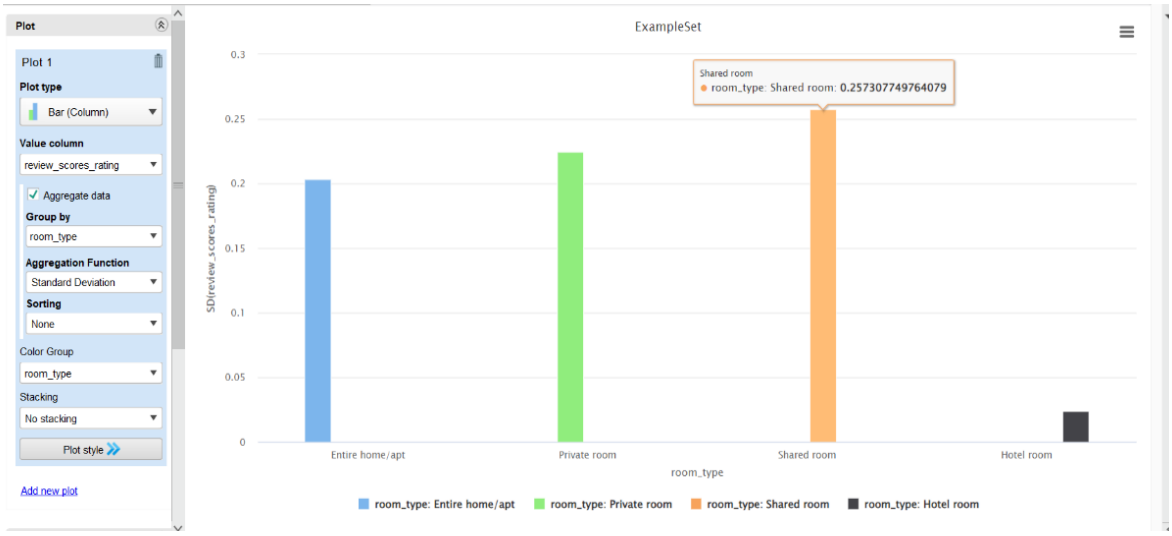
3) **Economic Activity:**
- **Urban Centres:** Large cities often have more economic activities, job opportunities, and tourism, which can lead to higher demand for short-term rentals and consequently higher prices.
- **Rural Areas:** Areas with lower economic activity may not attract as many visitors or business travellers, leading to lower demand and prices.

4) **Tourism**:

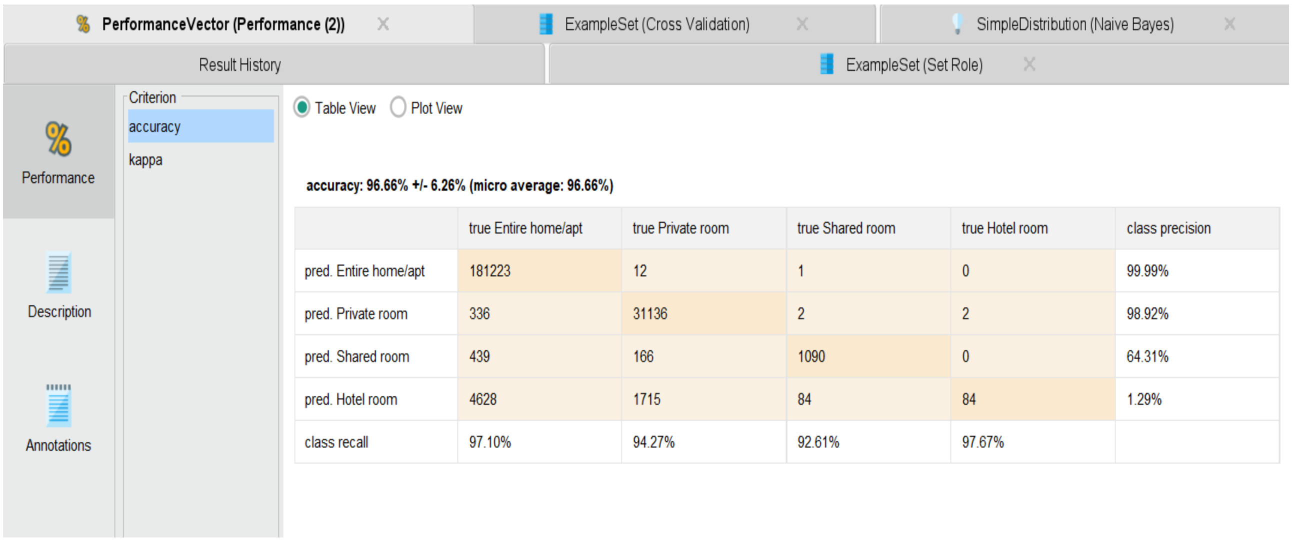
- **Popular Destinations**: Cities with large populations often have more attractions, events, and activities, making them popular tourist destinations. This popularity can lead to higher listing prices due to increased demand.
- **Less Touristic Areas**: Places with smaller populations might not have as many attractions, resulting in lower demand and prices.

Hence, Population is indeed a significant driver of listing prices. By including population data in our analysis, we can gain a more comprehensive understanding of the factors influencing listing prices and make more informed decisions based on this insight.

c. What features of listings influence customer satisfaction the most?



Based on the bar graph above, the standard deviation (SD) for shared room is higher than the standard deviation of other rooms. In our analysis, high standard deviation is good for catering to diverse budgets, bad for a standardized pricing strategy. This means that a high standard deviation suggests that there is a wide range of prices within the dataset. This variability can be beneficial for businesses because it means that they can cater to customers with diverse budgets. Some customers may be willing to pay higher prices for premium products or services, while others may prefer more budget-friendly options. By offering a range of prices, businesses can attract a broader customer base and maximize their revenue potential.



accuracy: 96.66% +/- 6.26% (micro average: 96.66%)

	true Entire home/apt	true Private room	true Shared room	true Hotel room	class precision
pred. Entire home/apt	181223	12	1	0	99.99%
pred. Private room	336	31136	2	2	98.92%
pred. Shared room	439	166	1090	0	64.31%
pred. Hotel room	4628	1715	84	84	1.29%
class recall	97.10%	94.27%	92.61%	97.67%	

Regarding the confusion matrix table above, it predicts that the number of people that will be booking for entire home/apartment is higher than any other room with 99.99% class precision and 97.10%. The lowest is the prediction of shared room which stands for 92.61% with 1090 people will be most likely to book. To mention the lowest class precision is at 1.29% for prediction of hotel room with 84 people to book stating that the category is most likely to be false while the high-class recall indicates the high overall accuracy and micro-average that the model is performing well in general.