Hoo Fang Yu - Singapore Tourism Board Data Assessment

Question 1

a) Describe the approach you will take and data fields you would look into when it comes to data preparation

Step 1: Preliminary Inspection

The first step involves conducting a preliminary inspection of the dataset. This step provides an initial understanding of the dataset in terms of its size and the various features it contains. It also helps to identify any potential issues at a first glance.

From this preliminary inspection, we are working with 22,974 entries of survey data with 51 features from 2018. Initially, we noticed that there are two Occupation columns which need to be fixed. Upon further analysis, we determined that the first Occupation column represents the designation, while the second represents the sector. These will be renamed to Occupation and Sector respectively.

Step 2: Data Cleaning

The second step of data preparation is data cleaning. This involves:

1. Ensuring Correct Data Types:
   - Some feature classes, such as the shopping features and the travel companion features, are not in the correct data types. These will be converted to doubles.
2. Handling Missing Values:
   - City of Residence: Geographical information is covered by Country of Residence, so missing values in City of Residence can be ignored.
   - Sector, Other Designation, and Designation (free text): These columns provide additional details about Occupation, so missing values in these columns can be ignored.
   - case: There are two entries without a case number. These will be fixed by assigning new unique case numbers: max(case) + row_number().
   - shopping_xxx Features: The shopping_any feature is affected by missing values in other shopping_xxx features. These will be filled with 0, and shopping_any will be corrected as the sum of the shopping_xxx features.
   - XXX_Terminal Features: These represent the point of entry/exit for each tourist. We must ensure that at least one terminal for each tourist is not missing. For cases where all three terminals are missing, we will fill them with the most frequent mode of entry for tourists from the same country of residence:
     - Case 699 (Vietnam): Air_Terminal = 2
     - Case 811 (Thailand): Air_Terminal = 3
     - Case 824 (Italy): Air_Terminal = 1
   - MainHotel: This column has 6,690 missing entries, which is more than 25% of the dataset. To handle this, we will fill the missing values with a unique ID (max(MainHotel) + 1) to represent 'not specified'. In addition there are a few hotel IDs that represent unknown values (991,992,993,994,999,9996 and 9999). These would also be filled with the not specified id.
3. Handling Duplicates:
   - Fortunately, there are no duplicate entries in this dataset.
4. Data Inconsistencies/ Illogical Data:
   - Will be addressed in part b of Question 1

Step 3: Data Transformation and Feature Engineering

For this dataset, we are addressing a visitor segmentation problem. The data must be transformed into an appropriate format for segmentation tasks, and features must be created to support these tasks. This will be covered in more detail in Question 2.

Step 4: Exploratory Data Analysis

The final step of the Data Preparation phase involves conducting Exploratory Data Analysis (EDA). This includes:

- Obtaining Descriptive Statistics: Understanding the overall distribution of the data.
- Creating Visualizations: Understanding visitor demographics, visit purposes, spending patterns, and visitor trends over the months, given the time-series nature of the dataset.

Data Fields for Analysis

1. Demographics: Country of Residence, City of Residence
    - Identifies the geographical origin of visitors, helping to tailor marketing strategies to specific regions and cities.
2. Visit Information: Purpose of Visit, Main Purpose of Visit
    - Provides insights into why visitors are coming to Singapore, enabling targeted marketing based on visit purposes such as leisure, business, or family visits.
3. Point of Entry/Exit: Land, Sea, and Air Terminals
    - Reveals the preferred modes of entry and exit, which can assist in improving transportation infrastructure and services for tourists.
4. Occupation: Occupation, Sector, Designation
    - Helps in understanding the professional background of visitors, which can be useful for segmenting the market and offering tailored experiences or services.
5. Weights_QTR
    - Adjusts the data to be representative of the broader population, ensuring accurate and unbiased analysis of visitor trends and behaviours.
    - (Personally, I am not very sure about how to use this feature so I will omit it)
6. Travel Companions: Various Travel Companion columns
    - Highlights travel group dynamics, aiding in the creation of travel packages and marketing campaigns targeting families, solo travellers, or business groups.
7. Visitor Spending: Various Spending columns
    - Tracks spending patterns across different categories (e.g., accommodation, food, shopping), helping to identify high-spending segments and optimize revenue streams.

b) Highlight the data idiosyncrasies / issues you found in this dataset and how would you deal with it.

Below are the first data idiosyncrasies/issues that I have found. Note that points 1-4 have been mentioned in the previous part.

1. Duplicate Columns:

- **Issue:** Two columns named Occupation.
- **Solution:** Renamed to Designation and Sector to clearly distinguish between the two.

2. Missing Values:

- **City of Residence:**
  - **Issue:** Some entries are missing this information.
  - **Solution:** Since geographical information is covered by Country of Residence, the missing values in City of Residence can be ignored for the analysis.
- **Designation, Sector, Other Designation, Designation (free text):**
  - **Issue:** Some entries are missing.
  - **Solution:** These columns provide additional details about Occupation, so missing values in these columns can be ignored as long as Occupation is complete.
- **Case:**
  - **Issue:** Two entries have missing case numbers.
  - **Solution:** Assigned new unique case numbers using max(case) + row_number() to ensure all entries have a unique identifier.
- **Shopping Features:**
  - **Issue:** Missing values in various shopping_xxx features affect the shopping_any feature.
  - **Solution:** Filled missing values with 0 and corrected shopping_any as the sum of the shopping_xxx features.
- **XXX_Terminal Features:**
  - **Issue:** Significant number of missing values.
  - **Solution:** Ensured at least one terminal is not missing for each tourist by filling missing values with the most frequent mode of entry for tourists from the same country:
    - Case 699 (Vietnam): Air_Terminal = 2
    - Case 811 (Thailand): Air_Terminal = 3
    - Case 824 (Italy): Air_Terminal = 1
- **MainHotel:**
  - **Issue:** 6,690 missing entries.
  - **Solution:** Filled missing values with a unique ID (max(MainHotel) + 1) to represent 'not specified'.

3. Incorrect Data Types:

- **Issue:** Some feature classes (shopping features, travel companion features) are not in the correct data types.
- **Solution:** Converted these features to numeric to ensure correct data analysis.

4. Weights:

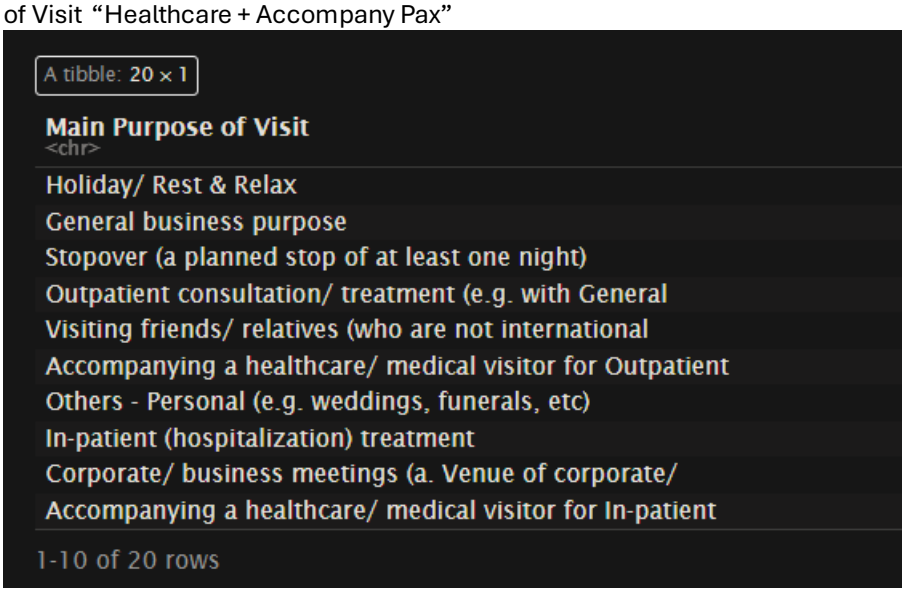
- **Issue:** The Weights_QTR column needs to be applied during analysis to adjust the data to be representative of the broader population.
- **Solution:** Ensure that weights are applied during statistical calculations for accurate and unbiased results.

5. Negative Values for `tototh` column

- **Issue:** Negative values present in `tototh` column when viewing summary statistics. This could possibly represent forms of profits made from Singapore (gambling, work etc.). If we are going to approach the problem from the perspective of tourist spending habits, then earnings won't be as interesting for our problem.
- **Solution:** Set negative values to zero and correct total expenditure column

6. Main Purpose of Visit does not match Purpose of Visit

- **Issue:** Main Purpose of Visit does not correspond to the correct category e.g. for the Purpose of Visit Healthcare, we have respondents that indicated their Main Purpose of Visit was for a 'General Business Purpose'. Below is a screenshot for the values for Main Purpose of Visit for the Purpose of Visit "Healthcare + Accompany Pax"



A tibble: 20 × 1

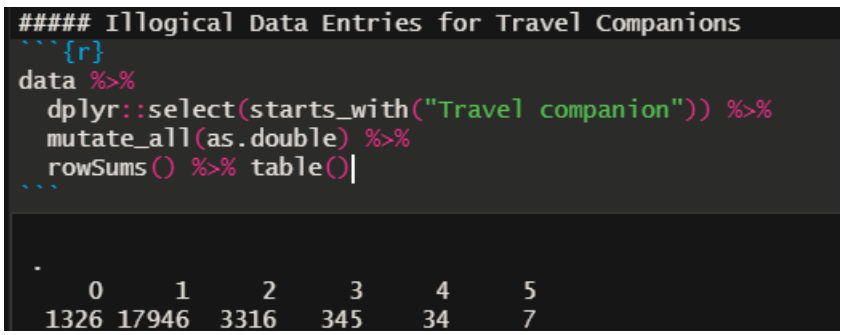
Main Purpose of Visit <chr>
Holiday/ Rest & Relax
General business purpose
Stopover (a planned stop of at least one night)
Outpatient consultation/ treatment (e.g. with General
Visiting friends/ relatives (who are not international
Accompanying a healthcare/ medical visitor for Outpatient
Others - Personal (e.g. weddings, funerals, etc)
In-patient (hospitalization) treatment
Corporate/ business meetings (a. Venue of corporate/
Accompanying a healthcare/ medical visitor for In-patient

1-10 of 20 rows

- **Solution:** Gather all the Purpose of Visits and reclassify them to the correct Purpose of Visit. This can be done through a prompt via ChatGPT and validating that the outputs of ChatGPT correspond to the correct categories

7. Illogical Data Entries for Travel Companion

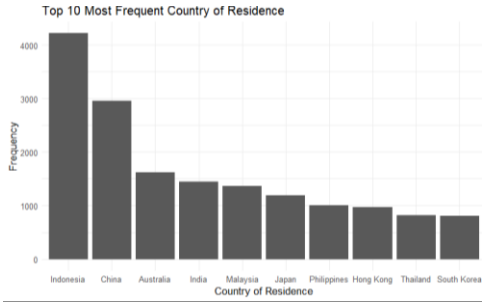
- **Issue**: Some survey entries reflect that a particular tourist is neither travelling by themselves nor with someone else. This is just not possible. (Fortunately, those who have indicated that they are travelling alone did not input inconsistencies that they are travelling with others)
- **Solution**: Remove illogical entries during particular analysis for this feature



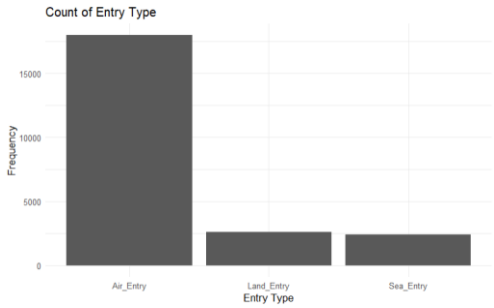
````
##### Illogical Data Entries for Travel Companions
```{r}
data %>%
  dplyr::select(starts_with("Travel companion")) %>%
  mutate_all(as.double) %>%
  rowSums() %>% table()
```
````

```
.
    0     1     2     3     4     5
 1326 17946  3316   345    34     7
```

Sources of Potential Bias



- Most of the tourists are predominantly from Indonesia and China



- Most of the tourists enter Singapore by Air

c) What are the considerations that you will take when analysing survey data. Use this data set to **illustrate and explain** your approach and considerations.

When analysing survey data, it's crucial to ensure data quality from the data preparation step, as discussed in previous parts of this question. This includes addressing missing values, removing duplicates, and ensuring correct data types. Applying survey weights (Weights_QTR) is essential to ensure the sample accurately represents the broader population.

Identifying outliers is also important, as they can skew the results and need to be handled appropriately. For instance, certain categories of expenditures show excessively high maximums compared to the third quartile, which could distort the analysis. This will be addressed in subsequent questions when segmenting respondents into groups.

Ensuring that survey entries are logical and free from errors is essential for accurate analysis. In this dataset, some discrepancies were noted between the "Main Purpose of Visit" and the "Purpose of Visit," likely due to respondents skimming through the survey. These inconsistencies need to be identified and corrected.

During the Exploratory Data Analysis (EDA) step, it's important to recognize and display the demographics of survey respondents to acknowledge any biases in the data collection process. For example, if the survey under-represents or over-represents certain demographics, this needs to be noted. From this particular dataset, majority of the respondents are from Indonesia and China which may skew the trends observed towards for these two countries and underrepresenting the other countries.

Additionally, the sample size and potential response biases, such as collecting a disproportionate amount of data from various points of entry, should be considered. For instance, as shown below if most participants arrive via air travel, this could lead to an over-representation of more affluent respondents.

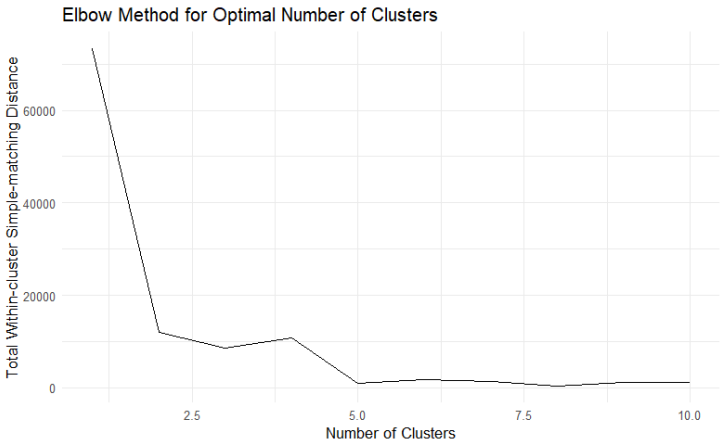
Question 2: Part 1

Based on the provided dataset, I have identified three groups of tourists who have visited Singapore. My approach involved several steps:

1. **Feature Engineering:** The mode of entry for each tourist was given in terms of the exact location of the checkpoint that they entered Singapore by. Abstraction was needed to simplify this into how they entered the country. This was done by creating new features 'AirEntry', 'LandEntry', 'SeaEntry' which are binary values on how a particular tourist has entered Singapore

2. **Selection of Important Categorical Variables**: I selected key categorical variables, including 'Country of Residence', 'AirEntry', 'LandEntry', 'SeaEntry', 'Gender', 'Marital Status', and 'Travel companion - Alone'. These variables were chosen because they are more general (having fewer categories) compared to other categorical variables, allowing us to form clusters that are more generalizable while still capturing essential information about the respondents' travel characteristics (e.g., 'Country of Residence' effectively captures geographical information).

3. **Clustering Process**: Using these selected variables, I performed K-Modes clustering to group the tourists into four distinct clusters. This method is particularly suitable for categorical data, as it uses a matching dissimilarity measure to form clusters. I performed K-Modes for up to 10 clusters and used the elbow method to determine the optimal number of clusters (5 for this case)



4. **Analysis of Clusters**: After clustering, I analysed the clusters to understand the characteristics and patterns within each group. This involved determining the most common values (modes) for the categorical features within each cluster as well as the cluster size.

Purpose of Visit <chr>	AirEntry <chr>	LandEntry <chr>	SeaEntry <chr>	Gender <chr>	Marital Status <chr>	Travel companion - Alone <chr>	Count <int>
Leisure	1	0	0	Female	Single	0	4464
Business + Accompanying Pax	1	0	0	Male	Married	1	3856
Leisure	1	0	0	Male	Married	0	6822
Leisure	0	0	1	Male	Married	0	2096
Leisure	1	0	0	Female	Married	0	5736

From the K-Modes Clustering Approach that I took, we could see certain demographics that are dominant. The 5 clusters include:

1. Married businessmen traveling for work alone
2. Married men traveling for leisure (arriving via air) with someone else
3. Married men traveling for leisure (arriving via sea) with someone else
4. Married women travelling for leisure with someone else

5. Single women travelling for leisure with someone else

These groups can then be further condensed into three groups

1. Married businessmen traveling for work alone
2. Married couples traveling for leisure together
3. Single women travelling for leisure with someone else (possibly with other female friends)

Note that the country of residence was omitted due to uninteresting characteristics (mode was all Indonesia/China which were the top 2 country of residences)

Choice of Metric

For this analysis, the within-cluster simple matching distance was used as the metric. This metric is particularly suitable for categorical data, which is the primary nature of the dataset. Simple matching distance counts the number of mismatches (or matches) between data points across all features. This choice is ideal because it directly addresses the categorical nature of the data, ensuring that similar categories are grouped together. The metric is straightforward and easy to interpret, making it clear how clusters are formed based on matching or mismatching attributes. Additionally, it is computationally efficient, which is crucial when dealing with large datasets.

Assumptions Made

Several assumptions were made during the development of this solution. First, it was assumed that each category is equally important. Second, it was assumed that each feature contributes independently to the clustering process. The homogeneity within clusters was another assumption, meaning that the data points within each cluster are more similar to each other than to those in other clusters. It was also assumed that the dataset has a sufficient sample size to form meaningful clusters and that there are no significant outliers that could skew the clustering results.

Approaches Considered

Several approaches were considered for clustering the data.

- **K-Modes Clustering** was chosen for its specific design to handle categorical data, making it suitable for the dataset. The advantages of K-Modes include its ability to handle categorical data directly, its efficiency and scalability for large datasets, and its simple and interpretable clustering results. However, it requires specifying the number of clusters beforehand and may not handle mixed data types as effectively without preprocessing.
- **K-Prototypes Clustering** was also considered, which can handle both categorical and numerical data. Its advantages include suitability for mixed data types and combining the strengths of K-Means and K-Modes, which suits the context of the dataset since this dataset is rich in both categorical and numerical data. However, it is more complex and computationally intensive than K-Modes and requires careful tuning of the weighting parameter for numerical and categorical features.
- **Hierarchical Clustering** was another option considered. It can create a hierarchy of clusters without specifying the number of clusters initially. The advantages include not needing to specify the number of clusters in advance and producing a dendrogram useful for visualizing the clustering process. However, it is computationally expensive, especially for large datasets, and less effective with categorical data compared to K-Modes.

Ultimately, K-Modes clustering was chosen due to its suitability for categorical data, simplicity, and efficiency.

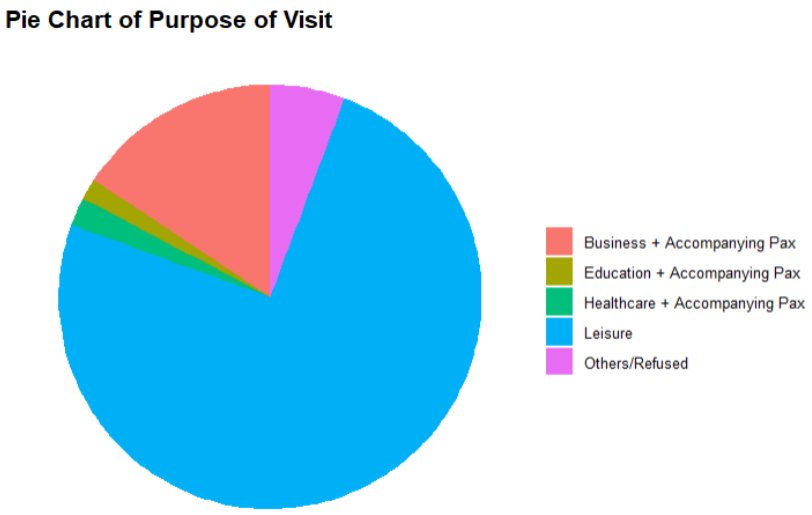
Performance of the Analytical Approach

The analytical approach using K-Modes clustering performed well for several reasons. It effectively handled the categorical nature of the data, grouping similar categories together based on the simple matching distance. The resulting clusters were easy to interpret, providing clear insights into the characteristics of each group of tourists. Additionally, the approach was computationally efficient, making it feasible to apply to large datasets without significant performance issues.

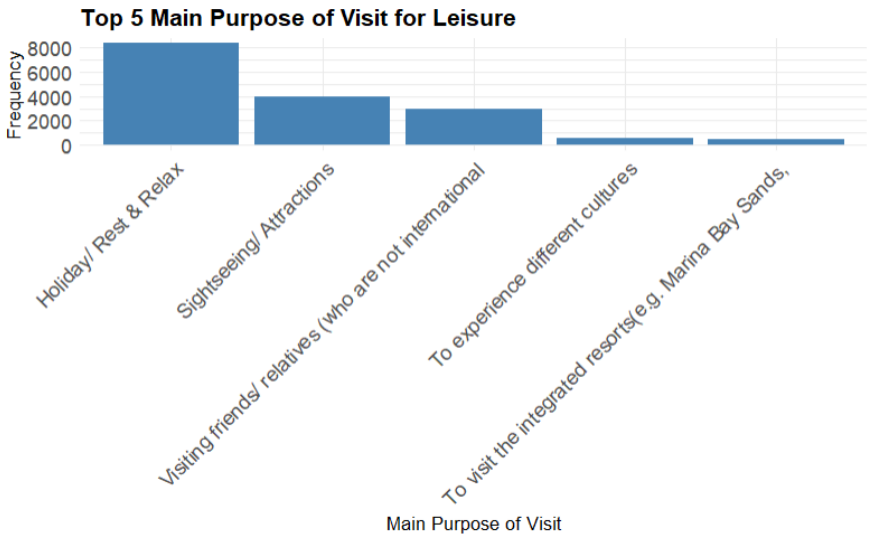
However, there are areas for potential improvement. The main issue with this approach is that this does not consider the numerical data within the dataset which could possibly help generate more informative clusters. The workaround this would be to calculate the median values for each numerical feature for each cluster. However, we may still lose a lot of information that could be useful in clustering (spending groups, interests etc.)

Question 2: Part 2

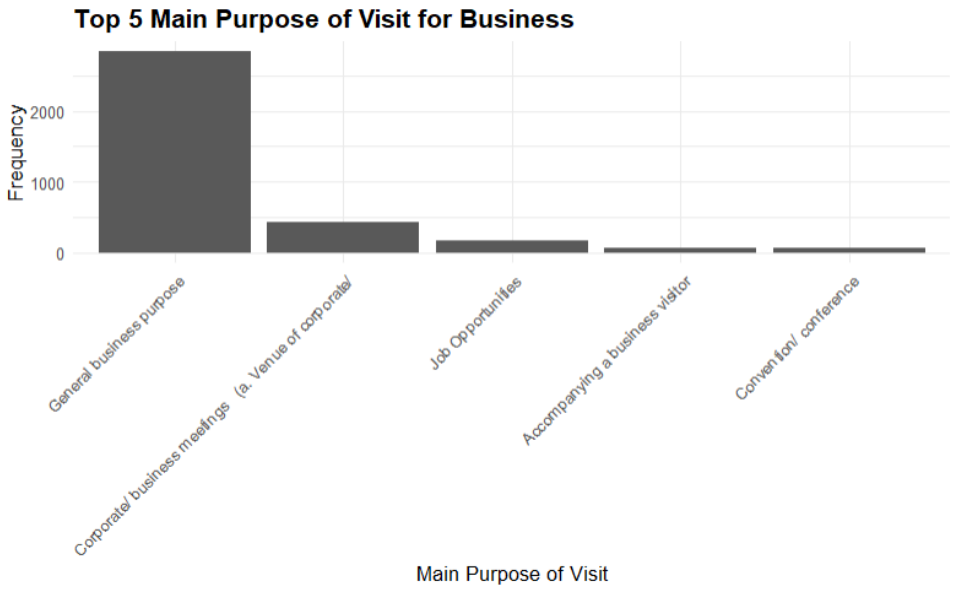
a) What can we learn about our visitors from the survey data that will help the Marketing team better reach out / market Singapore as a tourist destination them?



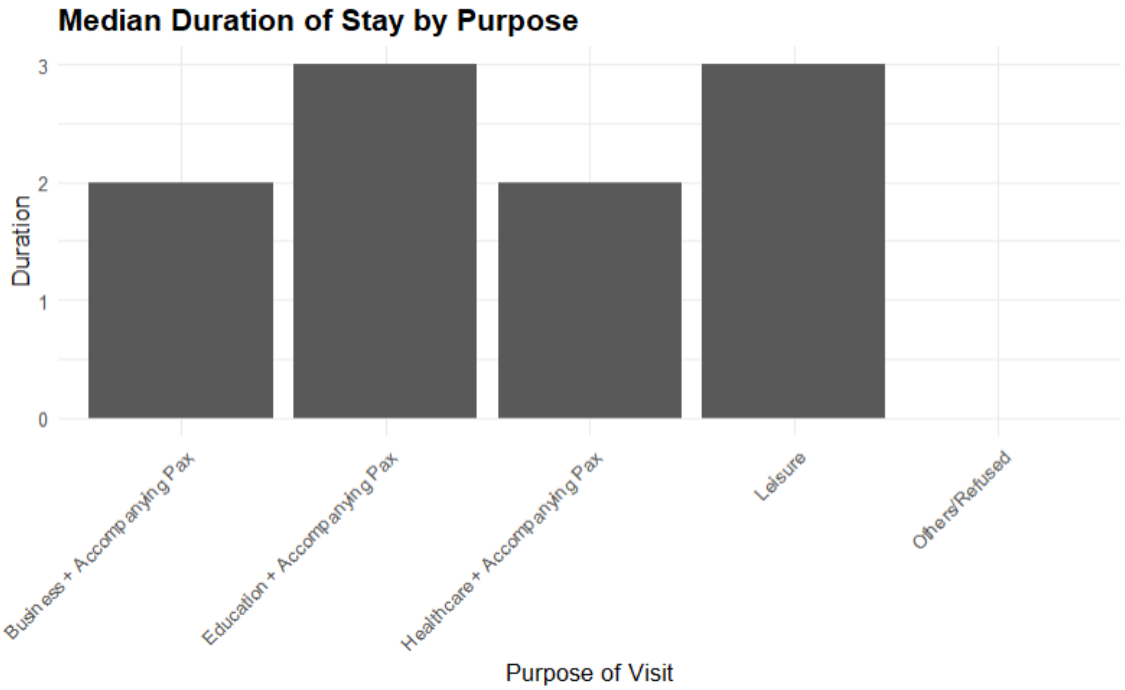
- The survey data reveals that majority of the visitors come to Singapore for leisure, with a small but significant proportion of visitors are here for Business. From our clustering analysis, we have observed that Male Businessmen are a dominant demographic.



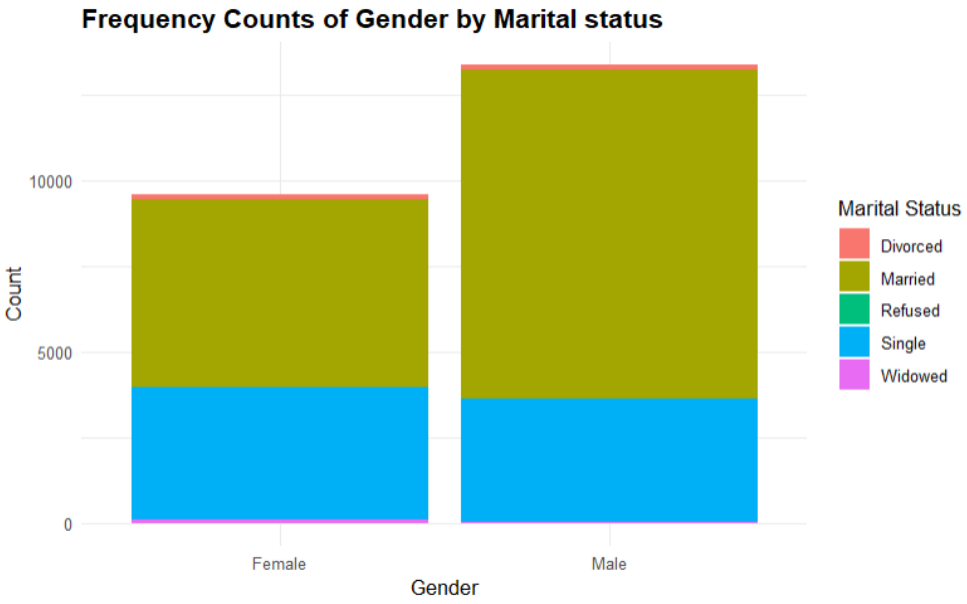
- When analysing the Main Purpose of Visit for the Leisure Category, we can observe that majority of visitors who are here for Leisure are primarily here for Holiday/Rest and Relax but also for Sightseeing and the Local Attraction.



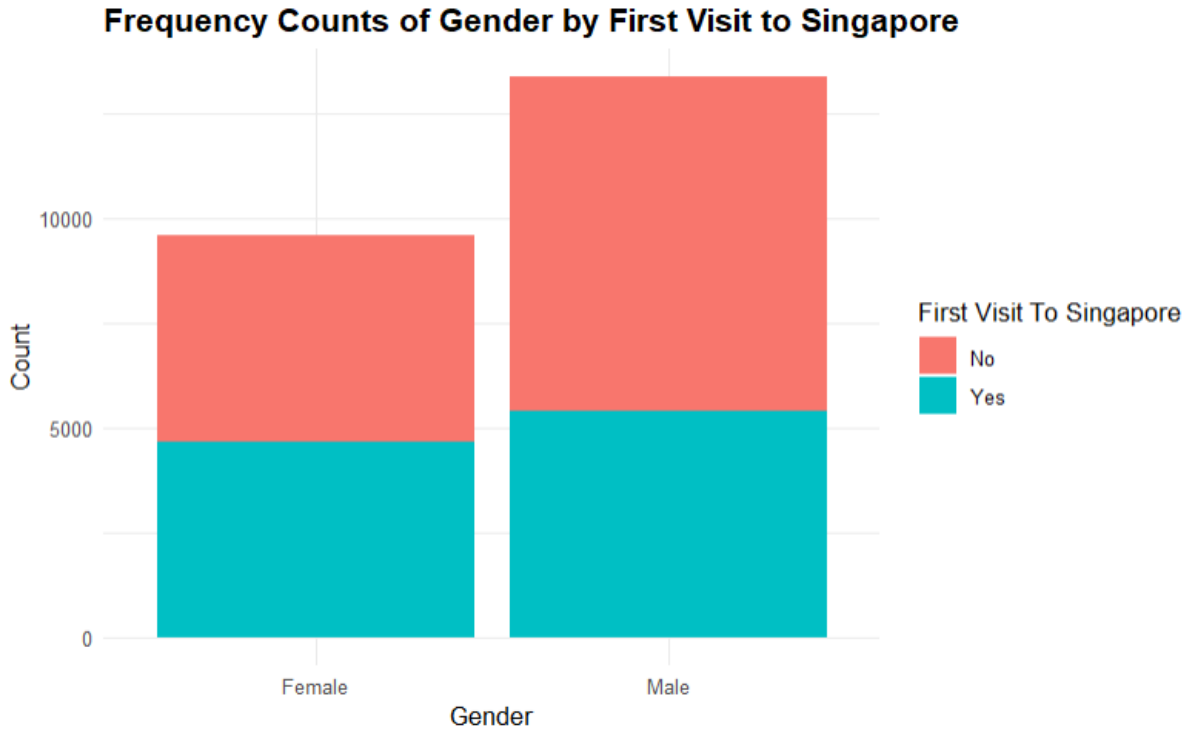
- Analysis of the Main Purpose of Visit for Business gives uninteresting results, they are mostly here for general business purposes which are very vague.



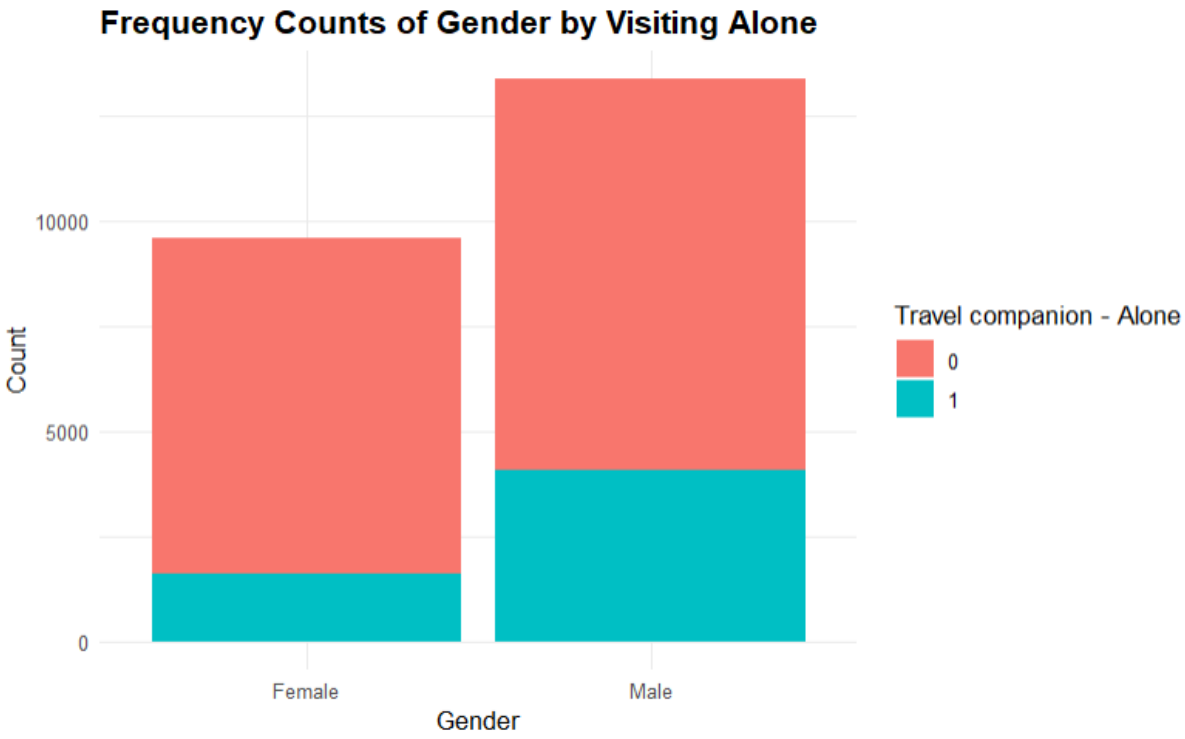
- Median Duration of Stay across all purposes are very short between 2 to 3 days. Tourists are spending very little time in Singapore.



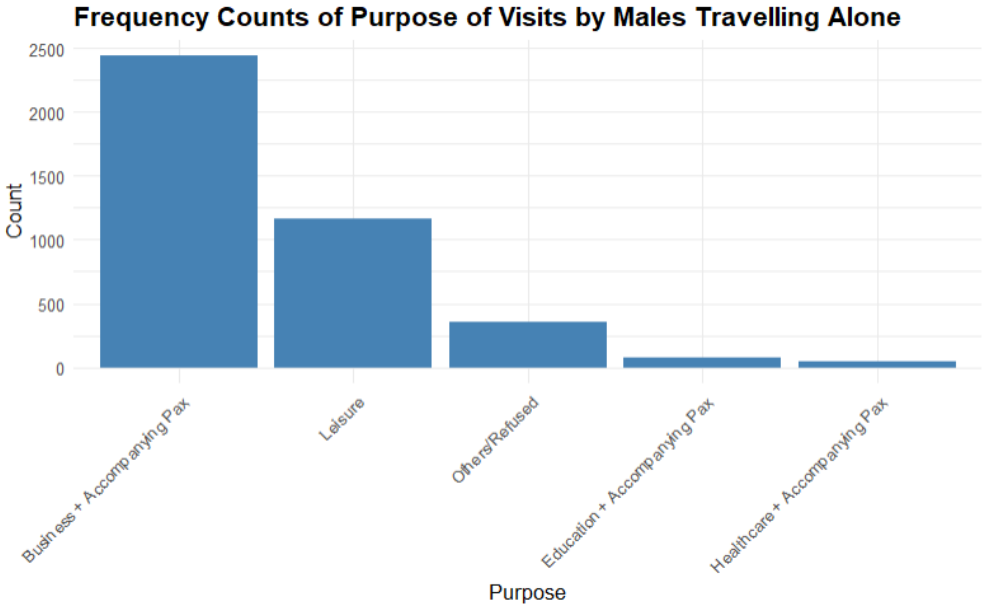
- Majority of the tourists are Male. Gender is going to be very useful in identifying key demographics for this particular problem as different genders are likely to have different preferences when touring Singapore.
- There is a higher proportion of Female tourists who are single as compared to Male tourists who are single so this could be another possible demographic to look at.



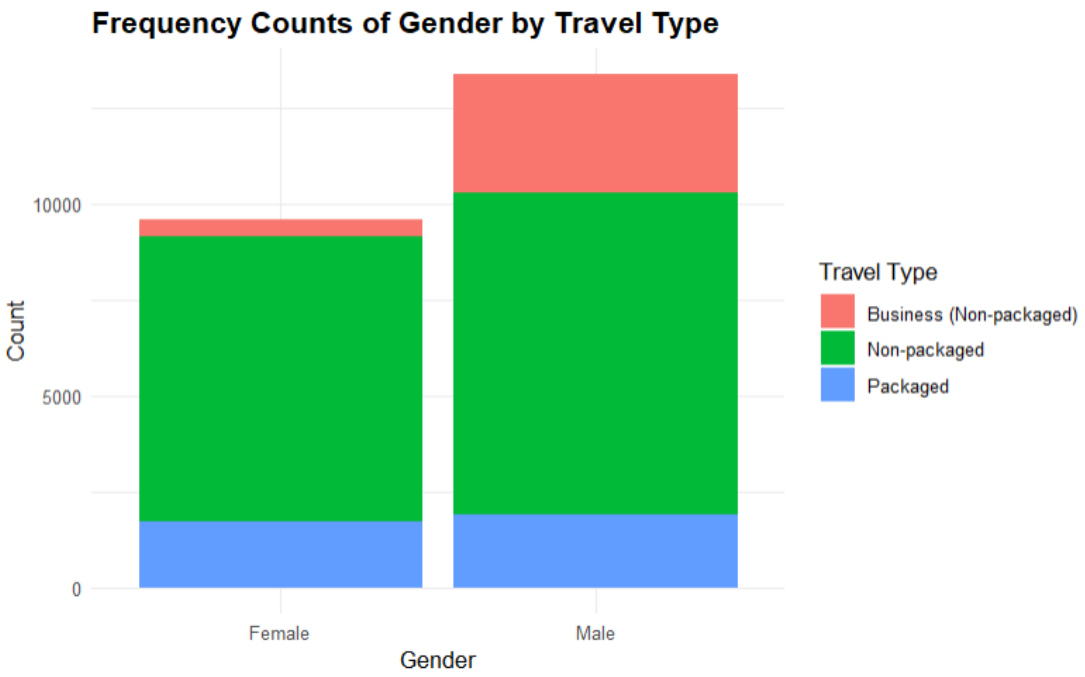
- Most of the tourists are returning visitors. Higher proportion of Male are returning visitors as compared to females. There is a roughly an equal distribution of Females who are first timers and returning tourists



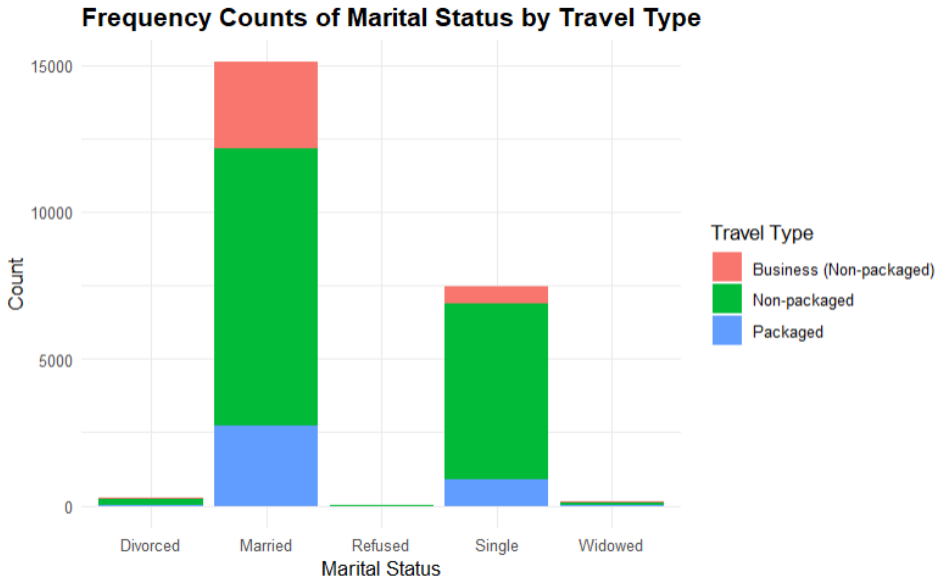
- Most of the tourists are travelling with companions. Interestingly, there is a higher proportion of Male tourists that are travelling alone as compared to female tourists. This leads to the next question that we can ask: "Why is this so?"



- Looking into the Male Tourists travelling alone, we can see that a huge majority of them are here for Business as compared to the general majority who are here for Leisure. Males Travelling alone is potentially another demographic to cater to as they are primarily here for Business.



- Excluding those that came to Singapore for Business, we can see that majority of Tourists that visit Singapore are on non-packaged tours and this distribution is similar across both genders.



- When trying to analyse the travel type according to marital status, we can see that likewise the distribution of Travel Types remained similar across the marital status.

b) Is there a correlation between travel companions and choice of hotel?

We are conducting a statistical analysis to determine if there is a correlation between travel companions and the choice of hotel among survey respondents. To achieve this, we first ensure that the hotel choices are categorized as factors and identify all columns that represent different types of travel companions. A for loop will iterate through each travel companion column, creating a contingency table with the hotel choice and performing a Chi-Square test of independence. This test evaluates whether there is a statistically significant association between the type of travel companion and the choice of hotel. The results are summarized to identify which, if any, travel companion types have a significant impact on hotel selection. Using a significance level of 0.01, if the p-value falls below this significance level, we would determine that there is a correlation between that particular travel companion and the choice of hotel. The threshold of 0.01 was used based on the p-values we have obtained from the analysis as there is a fair number of travel companions that fell below this threshold as compared to the standard 0.05.

Based on the results that we have obtained, there is a significant correlation between these forms of travel companionship and hotel choice: **Alone, Spouse, Children, Parents/Parents-in-law, Siblings, Other relatives, Friends, Business associates/Colleagues and Others.**