Title: Assignment 2 – Practical Data Science with Python (COSC2670)
Student ID: s3958556
Student Name: Athul Varghese Thampan
Email (contact info): s3958556@student.rmit.edu.au
Affiliations: RMIT University.
Date of Report: 24/05/2024

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

Table of Contents

Abstract

This project investigates patterns in travellers' preferences for various destination types in South India, utilising the BuddyMove Data Set, which comprises reviews from 249 reviewers on holidayiq.com. Through data exploration and modelling, including classification techniques like K Nearest Neighbours (KNN) and Decision Trees, the study identifies relationships between interests in different destination categories. Notably, a strong linear relationship was observed between religious and shopping destinations, and a moderate relationship between sports and picnic destinations. The classification models achieved accuracy rates between 79% and 84%, providing insights that can enhance tourism strategies and visitor experiences by targeting specific traveller interests effectively.

Introduction

The goal of this project is to explore individuals' interests in various types of destinations across South India, inferred from their preferences for other destinations. By examining the number of reviews provided by travellers, we aim to uncover patterns and relationships between different destination types using data modelling techniques, specifically classification. These insights can inform tourism strategies, enhancing the appeal of various destinations and improving the visitor experience.

Methodology

For this project, I have used the BuddyMove Data Set. This comprehensive dataset was compiled from destination reviews submitted by 249 reviewers on the travel website

holidayiq.com, covering the period up until October 2014. The reviews are categorised into six different types of destinations across South India, and the dataset records the number of reviews in each category for each reviewer (traveler).

The BuddyMove Data Set includes several attributes that provide a detailed snapshot of reviewer activities. The first attribute is a unique user ID, which identifies each reviewer individually. The subsequent attributes capture the number of reviews each user has written in various categories. Specifically, Attribute 2 records the number of reviews on sports destinations such as stadiums, sports complexes, and similar venues. Attribute 3 contains reviews of religious institutions, while Attribute 4 counts reviews of nature destinations such as beaches, lakes, and rivers. Attribute 5 logs reviews of theatres and exhibitions, and Attribute 6 tracks reviews of malls and shopping places. Finally, Attribute 7 details the number of reviews on parks and picnic spots.

First, I read the data set from the CSV file. Next, I pre-process the data by initially verifying the data types to ensure that the reviews are in integer format. Then, I check for any null values, and it indicates that there are none. Finally, I drop the 'User Id' column as it is not a relevant feature.

Afterward, I explored the data using general descriptive statistics of the dataset, including counts, means, standard deviations, minimum and maximum values, as well as percentiles (25th, 50th, and 75th). Following that, I delved into each column of the dataset—Sports, Religious, Nature, Theatre, Shopping, and Picnic. Employing a loop, I iterated through each column, showcasing individual descriptive statistics alongside the distribution of reviews for each column. The distribution of reviews for each column was visualised using individual histograms.

Next, I explored each pair of attributes using a scatterplot matrix. From this, I noticed that the columns "religious" and "shopping," as well as the columns "picnic" and "nature," exhibited a linear relationship. Hence, I plotted individual scatterplots for each of these pairs of attributes.

After the data exploration, I conducted data modelling using classification, specifically employing K Nearest Neighbours and Decision Trees. I performed this analysis for both pairs of attributes mentioned above (religious vs. shopping, as well as picnic vs. nature).

To begin, I generated the train-test set using the train_test_split function from sklearn. I introduced a new column named "shopping interest," which stores values 1 and 0, where 1 indicates an interest (i.e., number of reviews) higher than average, and 0 indicates an interest lower than average. In this setup, the variable X represents religious, and the target Y represents shopping interest. I then split the data 50/50, employing a random state of 4. A 50/50 split ensures that the model is trained on half of the available data and tested on the other half, providing a balanced evaluation of the model's performance. By setting the random_state parameter, I ensure consistent results across multiple runs of the code.

Next, I applied the K Nearest Neighbours classification method, setting k=3. A small k (like 1) can make the classifier sensitive to noise, while a large k can make the classification boundaries less distinct. I then generated a confusion matrix and a classification report for this model. Following that, I implemented a Decision Tree and again produced a confusion matrix and a classification report.
Similarly, I repeated the same process for the "picnic vs. sports" comparison. Here, I introduced a new column named "picnic interest," following the same methodology as for the "shopping interest" column. The variable X represents sports, and the target Y represents picnic interest. I generated the train/test set in a similar fashion and applied both K Nearest Neighbours and Decision Tree algorithms, followed by confusion matrix and classification report analysis.

Results

The following are the relevant findings from the exploration of each column.

```
Descriptive statistics for column 'Sports':
count    249.000000
mean      11.987952
std        6.616501
min        2.000000
25%        6.000000
50%       12.000000
75%       18.000000
max       25.000000
Name: Sports, dtype: float64
```

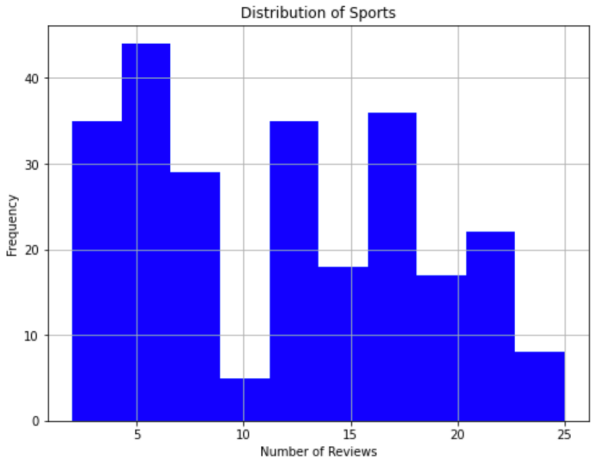

Figure 1

From Figure 1, the mean of the sports column is only 11.98, whereas the rest of the columns have means above 100. This is because the sports column's minimum is 2 and its maximum is 25, which are much lower than those of the other columns. This could indicate that sports destinations were the least popular among travellers.

```
Descriptive statistics for column 'Nature':
count    249.000000
mean     124.518072
std       45.639372
min       52.000000
25%       89.000000
50%      119.000000
75%      153.000000
max      318.000000
Name: Nature, dtype: float64
```

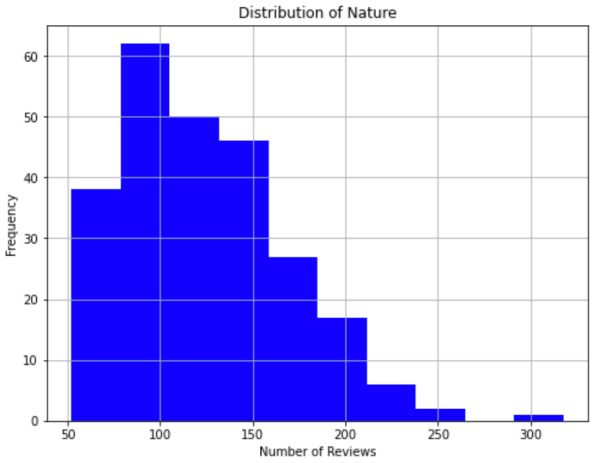

Figure 2

From Figure 2, we can see that the nature column has a mean of 124.5 reviews, the highest mean among all the columns. This suggests that nature destinations, such as beaches, rivers, lakes, etc., were the most popular destinations for travellers.

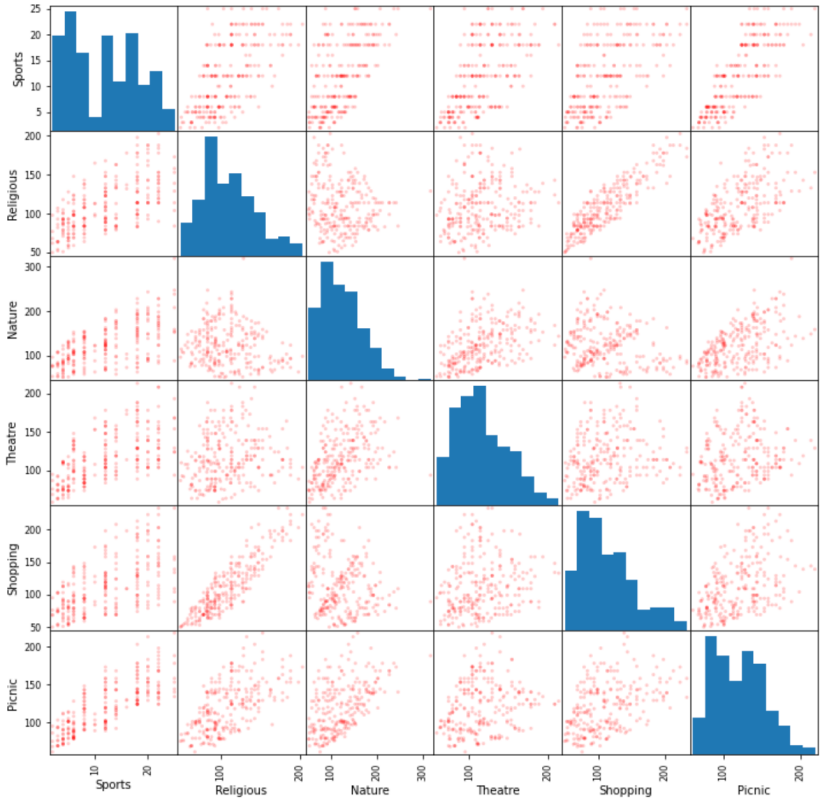

Figure 3

From the scatterplot matrix plotted (Figure 3) between each pair of attributes, we can see that the pairs picnic vs. sports and shopping vs. religious have linear relationships. Therefore, individual scatterplots for each pair were plotted.

Figure 4

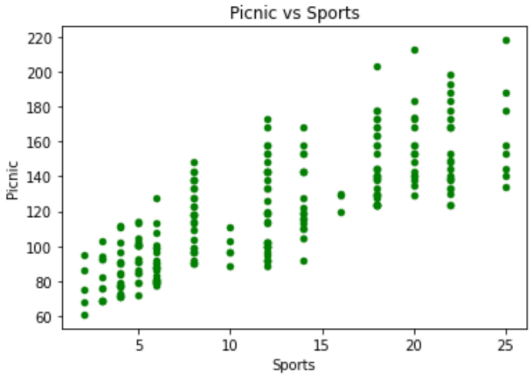

Figure 5

From Figure 4 and Figure 5, we can observe that each of these pairs exhibits a linear relationship, with religious vs. shopping showing a stronger linear relationship than picnic vs. sports. Hence, we can propose two plausible hypotheses for each of these pairs:

1. Individuals interested in religious destinations are also interested in shopping destinations. This could be attributed to the fact that visitors to religious sites often seek souvenirs and religious artifacts as tangible reminders of their spiritual journey. Local markets and shops typically offer a variety of items such as prayer beads, icons, religious books, and other memorabilia that hold spiritual significance.
2. Individuals interested in sports destinations are also interested in picnic destinations. This correlation may arise from the shared characteristic of outdoor activities associated with both sports and picnicking. Those who enjoy spending time outdoors for sports may also appreciate the leisurely, relaxed atmosphere of a picnic.

For the classification section, I first started with the analysis of religious vs. shopping interests. I generated the train/test set with shopping interest as the target variable. Using this train/test set, I applied the K Nearest Neighbours (KNN) algorithm to make predictions, obtaining 'y_predicted' from 'X_test'.

```
Confusion Matrix:
 [[62  8]
 [18 37]]
```

Figure 6

I then generated the confusion matrix for 'y_test' and 'y_predicted', as shown in Figure 6. From Figure 6, we can see the following results:

- Correctly predicted 62 instances where travellers interested in religious destinations had below-average interest in shopping destinations.
- Incorrectly predicted 8 instances where travellers interested in religious destinations had below-average interest in shopping destinations as above average.
- Incorrectly predicted 18 instances where travellers interested in religious destinations had above-average interest in shopping destinations as below average.
- Correctly predicted 37 instances where travellers interested in religious destinations had above-average interest in shopping destinations.

	precision	recall	f1-score	support
0	0.78	0.89	0.83	70
1	0.82	0.67	0.74	55
accuracy			0.79	125
macro avg	0.80	0.78	0.78	125
weighted avg	0.80	0.79	0.79	125

Figure 7

The classification report (Figure 7) for KNN in the religious vs. shopping analysis showed an accuracy of 79%.

```
Confusion Matrix:
 [[63  7]
 [17 38]]
```

Figure 8

Next, I conducted a decision tree analysis using the same train/test set. The confusion matrix (Figure 8) for this model revealed the following:

- Correctly predicted 63 instances where travellers interested in religious destinations had below-average interest in shopping destinations.
- Incorrectly predicted 7 instances where travellers interested in religious destinations had below-average interest in shopping destinations as above average.
- Incorrectly predicted 17 instances where travellers interested in religious destinations had above-average interest in shopping destinations as below average.
- Correctly predicted 38 instances where travellers interested in religious destinations had above-average interest in shopping destinations.

	precision	recall	f1-score	support
0	0.79	0.90	0.84	70
1	0.84	0.69	0.76	55
accuracy			0.81	125
macro avg	0.82	0.80	0.80	125
weighted avg	0.81	0.81	0.80	125

Figure 9

The classification report (Figure 9) for the decision tree model indicated an accuracy of 81%.

From this comparison, we can see that the decision tree results were slightly more accurate than the KNN results.

Similarly, I conducted the analysis for picnic vs. sports, where picnic interest was the target variable. I generated the train/test set and applied KNN to obtain ‘y_predicted’.

```
Confusion Matrix:
 [[68  0]
 [20 37]]
```

Figure 10

By comparing ‘y_predicted’ and ‘y_test’ using a confusion matrix (Figure 10), the results were as follows:

- Correctly predicted 68 instances where travellers interested in sports destinations had below-average interest in picnic destinations.
- Correctly predicted 0 instances where travellers interested in sports destinations had below-average interest in picnic destinations as above average.
- Incorrectly predicted 20 instances where travellers interested in sports destinations had above-average interest in picnic destinations as below average.
- Correctly predicted 37 instances where travellers interested in sports destinations had above-average interest in picnic destinations.

	precision	recall	f1-score	support
0	0.77	1.00	0.87	68
1	1.00	0.65	0.79	57
accuracy			0.84	125
macro avg	0.89	0.82	0.83	125
weighted avg	0.88	0.84	0.83	125

Figure 11

From the classification report (Figure 11), we can see that KNN achieved an accuracy of 84%.

Next, I performed the same analysis using a decision tree for Picnic vs. Sports.

```
Confusion Matrix:
 [[67  1]
 [19 38]]
```

Figure 12

The confusion matrix (Figure 12) for this model showed:

- Correctly predicted 67 instances where travellers interested in sports destinations had below-average interest in picnic destinations.
- Incorrectly predicted 1 instance where travellers interested in sports destinations had below-average interest in picnic destinations as above average.
- Incorrectly predicted 19 instances where travellers interested in sports destinations had above-average interest in picnic destinations as below average.
- Correctly predicted 38 instances where travellers interested in sports destinations had above-average interest in picnic destinations.

	precision	recall	f1-score	support
0	0.78	0.99	0.87	68
1	0.97	0.67	0.79	57
accuracy			0.84	125
macro avg	0.88	0.83	0.83	125
weighted avg	0.87	0.84	0.83	125

Figure 13

The classification report (Figure 13) for the decision tree indicated an accuracy of 84% as well.

Comparing these two, we see that although both models have the same overall accuracy, the KNN model correctly predicted all instances of travellers interested in sports destinations having below-average interest in picnic destinations.

Discussion

The primary objective of this project was to explore and classify individual interests in different types of destinations across South India based on travellers' preferences, as indicated by the number of reviews they provided. By employing classification techniques such as K Nearest Neighbours (KNN) and Decision Trees, the project aimed to identify patterns and relationships between interests in various destination categories, specifically religious vs. shopping and sports vs. picnic destinations.

This project addresses a significant knowledge gap in understanding the correlation between different types of destination interests among travellers in South India. By analysing the BuddyMove Data Set, this project provides a detailed exploration of these relationships and offers insights that can inform tourism marketing strategies and destination management.

The scatterplot analysis indicated a strong linear relationship between interests in religious and shopping destinations. This finding aligns with the hypothesis that visitors to religious sites often engage in shopping for souvenirs and religious artifacts. The scatterplot analysis also revealed a linear relationship between interests in sports and picnic destinations, though it was less pronounced than the relationship between religious and shopping interests. This supports the hypothesis that individuals who enjoy outdoor sports activities are also likely to enjoy picnicking.

Understanding the relationships between different destination interests can help tailor marketing strategies to target specific traveller segments more effectively.

Conclusion

This study successfully identified significant patterns in travellers' destination preferences across South India by analysing the BuddyMove Data Set. The findings revealed a strong correlation between interests in religious and shopping destinations, and a moderate correlation between interests in sports and picnic destinations. These insights are crucial for tailoring tourism strategies to better cater to specific traveller segments. The results have broader applicability in enhancing marketing efforts and improving visitor experiences in the tourism industry.

References

1. archive.ics.uci.edu. UCI Machine Learning Repository. [online] Available at: https://archive.ics.uci.edu/dataset/476/buddymove+data+set