Abstract

The question I wanted to research was the how different user ratings for 10 categories of travel venues were distributed, the frequency of the distribution, the qualitative skewness of the ratings, and what these results reveal about the various accommodations encountered during the travel venues in East Asia. Overall monumental institutions such as Parks_Picnic_Spots and Religious Institutions were rated more positively than other venues visited during travel.

Motivation

While there are many categories of review from each different user on the 10 categories of travel venues (accommodations), the purpose of the research is to find and compare the relationships between the each users' ratings of each different category, to explore how the ratings are concentrated and distributed according to category, and how the frequency of specific ratings influence how users feel about the labeled categories of data when traveling to a particular region and personal recommendations for other travelers who may consider future itineraries within the region.

Dataset

The dataset used was Travel Reviews. It was downloaded from
UCI Machine Learning repository (https://archive.ics.uci.edu/ml/datasets/Travel+Reviews) and the data were recorded in 2018. This dataset was created originally through TripAdvisor(https://archive.ics.uci.edu/ml/datasets/Travel+Reviews) and in the dataset, users provided their reviews on 10 categories of travel venues while traveling in East Asia.

The data collected in this dataset provided average reviews by each user for each category of travel venues and accommodations and the ratings were provided from 0-4 where 0 was considered Terrible, 1 was considered Poor, 2 was considered Average, 3 was considered Very Good, and 4 was considered Excellent (https://archive.ics.uci.edu/ml/datasets/Travel+Reviews). There are 11 categories total in this dataset. Attribute 1 is the Unique user id, Attributes 2-10 are average user reviews based on the travel setting. Categories 2-5 represent average user feedback on art galleries, dance clubs, juice bars, and restaurants. Categories 6-11 represent average user feedback on museums, resorts, parks/picnic spots, beaches, theaters, and religious institutions. However, the original column names were retained for simplicity of analysis.

In [39]:
```
#Classification

import seaborn as sns
import pandas as pd
import numpy as np

%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()


%matplotlib inline
```

Import the required dataset to analyze

In [40]:
```
#Import dataset
df = pd.read_csv('C:/Users/Cynthia Cho/Documents/tripadvisor_review.csv')

#Run the dataframe
df.head()
```

Out[40]:

	User ID	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8	Category 9	Category 10
0	User 1	0.93	1.8	2.29	0.62	0.80	2.42	3.19	2.79	1.82	2.42
1	User 2	1.02	2.2	2.66	0.64	1.42	3.18	3.21	2.63	1.86	2.32
2	User 3	1.22	0.8	0.54	0.53	0.24	1.54	3.18	2.80	1.31	2.50
3	User 4	0.45	1.8	0.29	0.57	0.46	1.52	3.18	2.96	1.57	2.86
4	User 5	0.51	1.2	1.18	0.57	1.54	2.02	3.18	2.78	1.18	2.54

Data Preparation and Cleaning

After downloading the data from the website and loading it to the hard drive I uploaded the data into the Jupyter notebook. As shown in the code cells I checked for null values in which fortunately there were no null values. I also counted the number of rows and analyzed the shape of the data. Overall, the dataset itself did not appear problematic. It was under 1000 rows and every field was populated. I did however rename the columns to make it more tangible to analyze.

In [44]:
```
df.columns = ['User_id', 'Art_Galleries', 'Dance_Clubs', 'Juice_Bars', 'Restaurants',
              'Museums', 'Resorts', 'Parks_Picnic_Spots', 'Beaches', 'Theatres', 'Religious_institut
ions']
df.shape
```

Out[44]: (980, 11)

In [45]:
```
df.head()
```

Out[45]:

	User_id	Art_Galleries	Dance_Clubs	Juice_Bars	Restaurants	Museums	Resorts	Parks_Picnic_Spots	Beaches	Theatre
0	User 1	0.93	1.8	2.29	0.62	0.80	2.42	3.19	2.79	1.8
1	User 2	1.02	2.2	2.66	0.64	1.42	3.18	3.21	2.63	1.8
2	User 3	1.22	0.8	0.54	0.53	0.24	1.54	3.18	2.80	1.3
3	User 4	0.45	1.8	0.29	0.57	0.46	1.52	3.18	2.96	1.5
4	User 5	0.51	1.2	1.18	0.57	1.54	2.02	3.18	2.78	1.1

In [53]:
```
df.describe().transpose()
```

Out[53]:

	count	mean	std	min	25%	50%	75%	max
Art_Galleries	980.0	0.893194	0.326912	0.34	0.67	0.83	1.0200	3.22
Dance_Clubs	980.0	1.352612	0.478280	0.00	1.08	1.28	1.5600	3.64
Juice_Bars	980.0	1.013306	0.788607	0.13	0.27	0.82	1.5725	3.62
Restaurants	980.0	0.532500	0.279731	0.15	0.41	0.50	0.5800	3.44
Museums	980.0	0.939735	0.437430	0.06	0.64	0.90	1.2000	3.30
Resorts	980.0	1.842898	0.539538	0.14	1.46	1.80	2.2000	3.76
Parks_Picnic_Spots	980.0	3.180939	0.007824	3.16	3.18	3.18	3.1800	3.21
Beaches	980.0	2.835061	0.137505	2.42	2.74	2.82	2.9100	3.39
Theatres	980.0	1.569439	0.364629	0.74	1.31	1.54	1.7600	3.17
Religious_institutions	980.0	2.799224	0.321380	2.14	2.54	2.78	3.0400	3.66

In [54]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 980 entries, 0 to 979
Data columns (total 11 columns):
User_id                 980 non-null object
Art_Galleries           980 non-null float64
Dance_Clubs             980 non-null float64
Juice_Bars              980 non-null float64
Restaurants             980 non-null float64
Museums                 980 non-null float64
Resorts                 980 non-null float64
Parks_Picnic_Spots      980 non-null float64
Beaches                 980 non-null float64
Theatres                980 non-null float64
Religious_institutions  980 non-null float64
dtypes: float64(10), object(1)
memory usage: 91.9+ KB
```

In [47]:
```
df[df.isnull().any(axis=1)]
```

Out[47]:

User_id	Art_Galleries	Dance_Clubs	Juice_Bars	Restaurants	Museums	Resorts	Parks_Picnic_Spots	Beaches	Theatres

In [48]:
```
before_rows = df.shape[0]
print(before_rows)
```

```
980
```

In [49]:
```
df = df.dropna()
```

In [50]:
```
after_rows = df.shape[0]
print(after_rows)
```

```
980
```

In [51]:
```
before_rows - after_rows
```

Out[51]: 0

Research Question(s)

What is the relationship between user ratings of a category and a category?

How do these ratings compare against one another?

What does it reveal about the user preferences of that category in their travel venues?

Methods

After downloading dataset and analyzing the statistics, a histogram was plotted in order to quickly and very broadly assess the statistical distribution of user ratings across 10 categories. Furthermore, it is a convenient way to accurately assess different average user ratings across each different venue. The reason for this rationale is there are multiple categories of data in this dataset with different users and it is difficult to use general Classification, Supervised Learning, general Machine Learning methods, and traditional Python libraries to accurately reflect these inputs and outputs accurately without actions such as deleting a column such as User_id. It turns out to be more constructive to apply this attribute rather than segregate it and focus on classifying the 10 other categories.

Limitations

Some of the limitations of this analysis is that it occurs 1 year after the data was taken. For that reason, it is not possible to accurately conduct a Twitter sentiment analysis. If a Twitter sentiment analysis was performed more updated and exact count along with detailed user demographics would be easier to obtain. This factor would have allowed for clarity of measurements. However, as far as the plot is concerned, although it is accurate, others may find it inadequate in more detailed statistical values such as errors were not properly accounted for in this data analysis and by not applying and comparing a wide variety of machine learning algorithms (i.e. Supervised Learning and other Classification algorithms) valuable statistics which can be obtained through generating a Classification and Confusion Matrix make it difficult to compare and account for errors using a wide variety of methods.

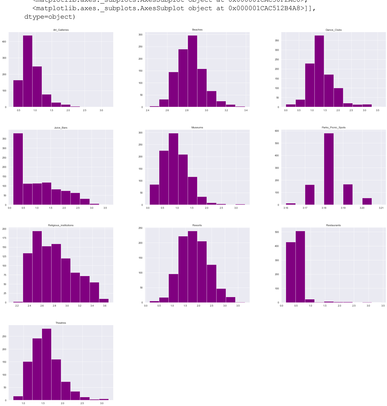
Findings

According to the results Art Galleries, Juice Bars, Dance Clubs, and Restaurants all skew to the lower ratings towards the left. Restaurants tends to have more user ratings concentrated rated in the range from Terrible to Poor. Museums and Theatres were skewed slightly to the left below Average with Museums rated lower than Theatres. The Beaches and Resorts had user ratings which were more towards Average. Religious institutions were ranked from Average ratings to some Good ratings in which the ratings range skews slightly to the right. Parks_Picnic_Spots had a rating over 3 with disparate rather than concentrated results.

Findings

In [52]:
```
df.hist(figsize=(30,30), color = 'purple')
```

Out[52]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC4CF53C8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC4F37B70>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC4F60400>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC4FB9D30>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC4FC47D0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC4FFE080>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC5023C88>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC5050DA0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC5050DD8>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC50C6128>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC50F2AC8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x000001CAC5128AA8>]],
      dtype=object)
```



Conclusions:

As noted above most users' traveler ratings for different travel venues in East Asia, some of the places users had traveled to tended to be rated mostly Terrible, mostly Low, mostly Average, or Average with some Good user ratings. Ratings varied depending on the place users traveled to. However the data seems to conclude users had more positive ratings of major public visiting area, average ratings of leisurely areas (beaches and resorts), and lower ratings of other areas.