Las Vegas Strip Data Set Analysis

The purpose of this analysis is to make insights on how Las Vegas Strip hotel scores are related to their facilities.

Description

The data comes from UCI Machine Learning Repository. As stated in the source: "This dataset includes quantitative and categorical features from online reviews from 21 hotels located in Las Vegas Strip, extracted from TripAdvisor". The dataset contains 504 records and 20 tuned features, 24 per hotel (two per each month, randomly selected), collected between January and August of 2015. It was used in this study: *Mora, S., Rita, P., & Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: The case of Las Vegas Strip. Tourism Management Perspectives, 23, 41-52.*

Let's read the Las Vegas Strip database and view its basic characteristics:

In [1]:
```
import pandas as pd
import seaborn as sns
import numpy as np
from statsmodels.stats.weightstats import ttest_ind
from scipy.stats import shapiro, levene
```

In [2]:
```
df = pd.read_csv(r"C:\Users\lenovo\Desktop\ibm_machine_learning\1_exploratory_data_analysis\week2\assignment\Lab
```

The data frame shape that we get below is the same as in the description. This way, we make sure that we read all the entries:

In [3]:
```
df.shape
```

Out[3]:
```
(504, 20)
```

The summary below shows that there are no null values in the dataset, which is good. Variables are of either integer or object type:

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504 entries, 0 to 503
Data columns (total 20 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   User country       504 non-null    object
 1   Nr. reviews        504 non-null    int64 
 2   Nr. hotel reviews  504 non-null    int64 
 3   Helpful votes      504 non-null    int64 
 4   Score              504 non-null    int64 
 5   Period of stay     504 non-null    object
 6   Traveler type      504 non-null    object
 7   Pool               504 non-null    object
 8   Gym                504 non-null    object
 9   Tennis court       504 non-null    object
 10  Spa                504 non-null    object
 11  Casino             504 non-null    object
 12  Free internet      504 non-null    object
 13  Hotel name         504 non-null    object
 14  Hotel stars        504 non-null    object
 15  Nr. rooms          504 non-null    int64 
 16  User continent     504 non-null    object
 17  Member years       504 non-null    int64 
 18  Review month       504 non-null    object
 19  Review weekday     504 non-null    object
dtypes: int64(6), object(14)
memory usage: 78.9+ KB
```

The first five lines are printed below:

In [5]:
```
df.head()
```

Out[5]:

	User country	Nr. reviews	Nr. hotel reviews	Helpful votes	Score	Period of stay	Traveler type	Pool	Gym	Tennis court	Spa	Casino	Free internet	Hotel name	Hotel stars	Nr. rooms	User continent	M
0	USA	11	4	13	5	Dec-Feb	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	
1	USA	119	21	75	3	Dec-Feb	Business	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	
2	USA	36	9	25	5	Mar-May	Families	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	
3	UK	14	7	14	4	Mar-May	Friends	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	Europe	
4	Canada	5	5	2	4	Mar-May	Solo	NO	YES	NO	NO	YES	YES	Circus Circus Hotel & Casino Las Vegas	3	3773	North America	

Most of these variables have self-explanatory names. For this analysis, we will use review scores that have integer values from 1 to 5, and qualitative variables that have either 'YES' or 'NO' values, depending on whether the hotel provides a particular facility (pool, gym, tennis court, spa, casino, or free internet).

Data Exploration Plan

In this analysis, we are interested in how hotel scores are related to free internet availability. Hotels that provide free internet will be named Group 1, and those that do not - Group 2.

To investigate score differences, we will perform an exploratory data analysis at first - summarize and visualize score data. Score histograms for both groups will demonstrate the differences in score distribution. Also, basic statistical characteristics of score data (both complete and for the subgroups) will show if means and standard deviations for the scores are very different.

Actions With Data

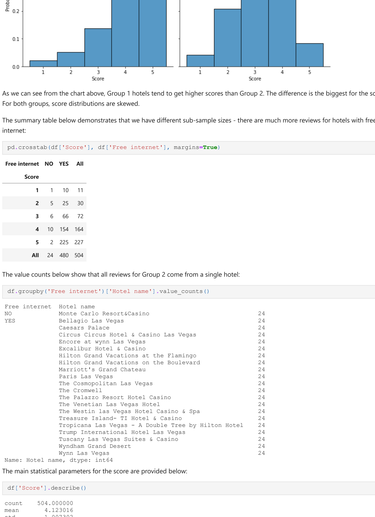
For this analysis, we will use scores and free internet variables. They have no null values (as all the other variables in this dataset). Scores are integer, free internet variable is categorical, so both types are suitable. So no additional actions related to data cleaning or feature engineering will be taken.

Exploratory Data Analysis

Score histograms for both groups are provided below:

In [6]:
```
sns.displot(x='Score', col='Free internet', discrete=True, stat='probability', common_norm=False, data=df)
```

Out[6]:
```
<seaborn.axisgrid.FacetGrid at 0x1aa2eb384a0>
```



As we can see from the chart above, Group 1 hotels tend to get higher scores than Group 2. The difference is the biggest for the score of 5. For both groups, score distributions are skewed.

The summary table below demonstrates that we have different sub-sample sizes - there are much more reviews for hotels with free internet:

In [7]:
```
pd.crosstab(df['Score'], df['Free internet'], margins=True)
```

Out[7]:

Free internet	NO	YES	All
Score			
1	1	10	11
2	5	25	30
3	6	66	72
4	10	154	164
5	2	225	227
All	24	480	504

The value counts below show that all reviews for Group 2 come from a single hotel:

In [8]:
```
df.groupby('Free internet')['Hotel name'].value_counts()
```

Out[8]:
```
Free internet  Hotel name                                                
NO             Monte Carlo Resort&Casino                                     24
YES            Bellagio Las Vegas                                            24
               Caesars Palace                                                24
               Circus Circus Hotel & Casino Las Vegas                        24
               Encore at wynn Las Vegas                                      24
               Excalibur Hotel & Casino                                      24
               Hilton Grand Vacations at the Flamingo                        24
               Hilton Grand Vacations on the Boulevard                       24
               Marriott's Grand Chateau                                      24
               Paris Las Vegas                                               24
               The Cosmopolitan Las Vegas                                    24
               The Cromwell                                                  24
               The Palazzo Resort Hotel Casino                               24
               The Venetian Las Vegas Hotel                                  24
               The Westin Las Vegas Hotel Casino & Spa                       24
               Treasure Island- TI Hotel & Casino                            24
               Tropicana Las Vegas - A Double Tree by Hilton Hotel           24
               Trump International Hotel Las Vegas                           24
               Tuscany Las Vegas Suites & Casino                             24
               Wyndham Grand Desert                                          24
               Wynn Las Vegas                                                24
Name: Hotel name, dtype: int64
```

The main statistical parameters for the score are provided below:

In [9]:
```
df['Score'].describe()
```

Out[9]:
```
count    504.000000
mean       4.123016
std        1.207392
min        1.000000
25%        4.000000
50%        4.000000
75%        5.000000
max        5.000000
Name: Score, dtype: float64
```

Average scores are significantly higher for hotels with free internet than for those without. Standard deviations are similar in both groups, however:

In [10]:
```
df.groupby('Free internet')['Score'].agg([np.mean, np.std])
```

Out[10]:

	mean	std
Free internet		
NO	3.291667	1.041703
YES	4.164583	0.988449

Key Findings

Scores have skewed distributions for both groups, which shows they are not distributed normally (an assumption that would be useful for statistical testing). However, the median score is similar to the mean (4 vs 4.12). Groups differ in size - there are much more reviews for hotels with free internet access, and their average scores differ much (but not standard deviations). The most visible difference - a much lower percentage of maximal scores of 5 for hotels without free internet access.

Hypotheses formulation

Below are the examples of hypotheses that we can formulate about Las Vegas Strip data:

- The average score for all stays is 4 (as the closest integer value to the actual sample data, and equal to the median)
- The average score for hotels with and without free internet is the same
- For hotels with and without free internet, the proportion of stays with score 5 is the same (5 is selected due to a visible difference between the groups)

We will test the second hypothesis. Our null hypothesis is that average scores for Groups 1 and 2 are the same. The alternative hypothesis is that they differ. In other words, the alternative is two-sided.

Hypothesis test

For testing purposes, the scores will be split into two subsets, by whether the hotel provides free internet or not.

In [11]:
```
free_score = df.loc[df['Free internet'] == 'YES', 'Score']
no_free_score = df.loc[df['Free internet'] == 'NO', 'Score']
```

For comparing means between two groups, the t-test for independent samples can be used. Its efficiency depends on whether the two groups of samples are normally distributed. Shapiro-Wilk test can be used to check this.

For both groups, Shapiro-Wilk test gives a low p-value, which allows us to reject the null hypothesis of normally distributed data with a significance level of 0.05:

In [12]:
```
_, p_value = shapiro(free_score)
p_value
```

Out[12]:
```
1.1659724096954562e-24
```

In [13]:
```
_, p_value = shapiro(no_free_score)
p_value
```

Out[13]:
```
0.017601846137717247
```

As we can see, scores for both groups do not follow the normal distribution, which we have observed in the chart above. However, with large enough samples, the normality assumption can be ignored due to the Central Limit Theorem. Our samples are of 480 and 24 items, so we will assume they are large enough.

Also, variances of both groups should be checked for equality because the type of the test depends on that. Levene's test is used for this below:

In [14]:
```
_, p_value = levene(free_score, no_free_score)
p_value
```

Out[14]:
```
0.4411849033706577
```

In this case, p-value is much higher than the significance level of 0.05, so we can assume equal variances.

Now we will perform t-test on two independent samples. Again, we choose 0.05 as our significance level. Based on Levene's test performed above and exploratory data analysis, we assume that score variance for both groups is the same:

In [15]:
```
_, p_value, _ = ttest_ind(free_score, no_free_score, alternative='two-sided', usevar='pooled')
p_value
```

Out[15]:
```
3.0078433823580579e-05
```

The p-value is much lower than 0.05, so we reject the null hypothesis of equal score means for Groups 1 and 2. In other words, the mean score difference between the groups is statistically significant.

Suggestions

This analysis could be expanded for other variables, such as comparing scores (parameters or distributions) for other groups of hotels as well. For example, we could test whether mean scores differ for hotels with or without pools (or any other facility). Also, we could compare scores for the reviews grouped by a feature with multiple categorical values, such as 'Traveler type'. In this case, we could use ANOVA test.

Also, we assumed our samples are large enough to ignore whether the data is distributed normally. In general, having a sample size of more than 30 is viewed as enough to ignore the normality assumption, but our smaller sample has only 24 items. In order to stay on the safe side, alternative approaches to testing, such as non-parametric tests can be used to check if scores for both groups have the same distribution.

Furthermore, all 24 reviews for stays with no free internet come from one hotel - "Monte Carlo Resort&Casino". Its lower scores might be related to some other reasons, apart from free internet availability. So it might be worth including other variables into our analysis.

Summary

In general, data quality is good - we have no null values, thus no data cleaning or imputation is needed. However, since the data was collected between January and August of 2015, it might be useful to include data from other time periods. Alternatively, we could select more reviews from the same period, or do both. Larger sample sizes could improve the inference quality.