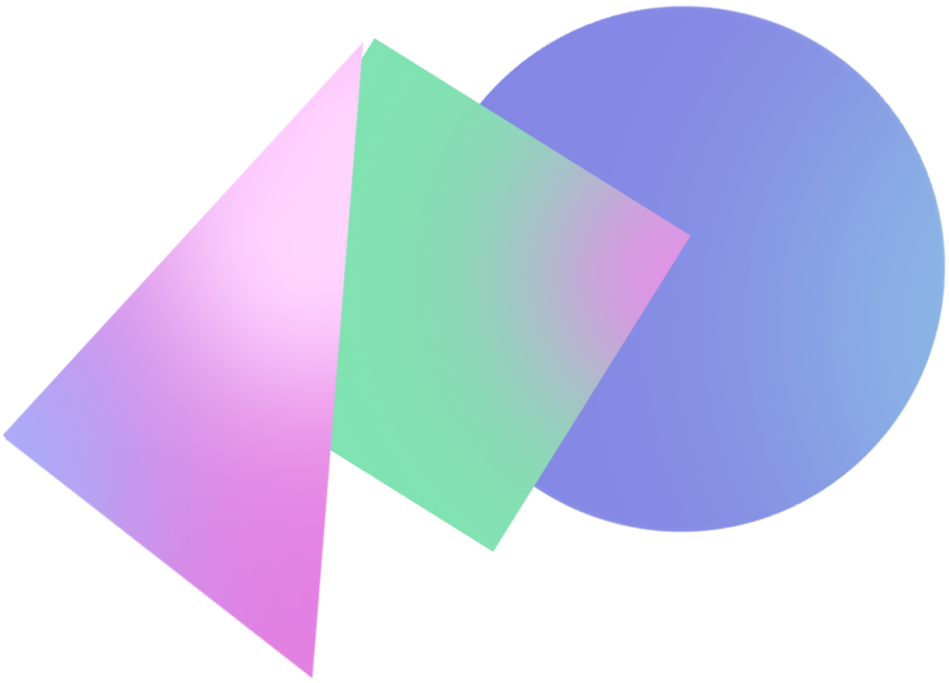


Data Analytics Graduate Capstone

NKM2 Task 2: Data Analytics Report and Executive Summary - D214

Restaurant Revenue Correlation Analysis

Performance Assessment

WGU - MSDA

Thoai Thomas Tran
Student ID # 001462881
July 2024

Research Question 4
A. Summary 4
Research Question 4
Justification 4
Context 4
Data Collection 5
B. Relevant Data Collected 5
Data-Gathering Methodology 5
Challenges and Solutions 6
Data Extraction and Preparation 7
C. Data Extraction 7
Data Preparation 9
Tools and Techniques Justification 11
Advantage and Disadvantage 12
Analysis 12
D. Analysis 12
Analysis Techniques 12
Pearson Correlation 12
Multiple Regression Analysis 14
ANOVA (Analysis of Variance) 15
Hypothesis Testing 17
Data Summary and Implications 19
E. Summary of Analysis 19
Implications 21
Limitation of Analysis 21
Recommendations 21
Future Directions 22

Answering the Research Question 22
Confirming the Hypothesis 23
F. Sources 24
Web Sources: 24
In-text Citations: 24

Research Question

A. Summary

Research Question

To what extent do factors such as location, cuisine type, average meal price, marketing budget, social media followers, and service quality score affect a restaurant's revenue?

Justification

Understanding the key attributes that impact restaurant revenue is crucial for optimizing business operations and strategies. This research will help identify the most significant revenue drivers and provide actionable insights for strategic planning. By analyzing these factors, restaurant owners can make informed decisions to enhance their profitability and sustainability, potentially leading to significant business growth and success.

Context

The restaurant industry is highly competitive, with revenue being influenced by many factors. By understanding the key attributes significantly impacting revenue, restaurant owners and managers can make data-driven decisions to optimize their operations, marketing strategies, and overall customer experience. This study aims to analyze the correlation between various restaurant attributes and their revenue, providing insights that can help improve profitability and business sustainability. This analysis will enhance the understanding of revenue drivers and support strategic planning in the restaurant industry. Forecasting restaurant sales leads to increased revenue and can significantly lower operating costs, all contributing to the restaurant’s overall profitability (Jobalia, 2022). Correlation analysis can assist business leaders in gaining valuable insights into the relationships between business outcomes (Rinehart, 2023).
Hypothesis

- **Null Hypothesis**: Location, cuisine type, average meal price, marketing budget, social media followers, and service quality score do not significantly affect the revenue of a restaurant.

- **Alternate Hypothesis**: Location, cuisine type, average meal price, marketing budget, social media followers, and service quality score significantly affect the revenue of a restaurant.

Data Collection

B. Relevant Data Collected

The dataset collected for this study is the Restaurant Revenue Prediction Dataset from Kaggle (Therrien, 2023). It includes various attributes that may influence a restaurant's revenue. The dataset is designed to help predict a restaurant's revenue based on these features. The specific columns in the dataset include:

- **Name**: The name of the restaurant.
- **Location**: The restaurant's location (e.g., Rural, Downtown).
- **Cuisine**: The type of cuisine offered (e.g., Japanese, Mexican, Italian).
- **Rating**: The average rating of the restaurant.
- **Seating Capacity**: The number of seats available in the restaurant.
- **Average Meal Price**: The average price of a meal at the restaurant.
- **Marketing Budget**: The marketing budget allocated for the restaurant.
- **Social Media Followers**: The number of social media followers.
- **Chef Experience Years**: The number of years of experience of the head chef.
- **Number of Reviews**: The total number of reviews the restaurant has received.
- **Avg Review Length**: The average length of reviews.
- **Ambiance Score**: A score representing the ambiance of the restaurant.
- **Service Quality Score**: A score representing the quality of service.
- **Parking Availability**: Indicates if parking is available (Yes/No).
- **Weekend Reservations**: The number of reservations made on weekends.
- **Weekday Reservations**: The number of reservations made on weekdays.
- **Revenue**: The total revenue generated by the restaurant.

The dataset consists of 8,368 unique entries, each representing a different restaurant.

Data-Gathering Methodology

The data was collected from Kaggle, a well-known platform for data science competitions and datasets. The dataset is publicly available under the CC BY-SA 4.0 license, allowing sharing and adaptation with proper attribution. This makes it an excellent resource for academic and research purposes.

Advantage: One significant advantage of using this dataset is its comprehensiveness and various attributes. It includes a wide range of factors that can influence restaurant revenue, providing a rich data source for analysis. Additionally, the dataset's availability on Kaggle ensures it is easily accessible and ready for use without needing primary data collection.

Disadvantage: A potential disadvantage of using this dataset is the lack of control over the data collection process. As the data is pre-collected and provided by an external source, inherent biases or inaccuracies might affect the analysis. Moreover, some attributes may have missing values or inconsistencies that must be addressed during the data preparation phase.

Challenges and Solutions

One of the challenges encountered during the data collection process was dealing with missing values and ensuring consistency in data formatting. The following steps were taken to overcome this:

- **Handling Missing Values**: Various techniques, such as imputation (filling in missing values with mean, median, or mode) and removing rows or columns with excessive missing data, were considered. The final approach depended on the extent and pattern of missingness observed in the dataset.

- **Ensuring Consistency**: Data preprocessing steps were implemented to standardize the data format. This included converting categorical variables to a consistent format (e.g., Yes/No to binary 1/0), ensuring numerical values were within expected ranges, and normalizing continuous variables to facilitate analysis.

Addressing these challenges prepared the dataset for further analysis, ensuring that the subsequent steps in the research process could be carried out effectively and accurately.

Data Extraction and Preparation

The data extraction and preparation process is crucial for ensuring the quality and consistency of the dataset before performing any analysis. Just as a chef needs fresh, quality ingredients to prepare a delicious meal, it is essential to clean and preprocess data to provide accurate and reliable results. Dirty or inconsistent data can lead to incorrect conclusions and poor decision-making (Ortega, 2023). Below are the detailed steps and justifications for each stage of this process.

C. Data Extraction

The dataset was downloaded from Kaggle and contains information about various restaurants and their corresponding revenue. It includes several attributes: location, cuisine type, average meal price, marketing budget, social media followers, and more.

Loading the Dataset

The dataset was loaded into a Pandas DataFrame for easy manipulation and analysis.

```
# Load the dataset
dataset = pd.read_csv('restaurant_data.csv')
```

Initial Inspection

The first few rows of the dataset were inspected to understand its structure and content.

```
# Display the first few rows of the dataset
print(dataset.head())
```

```
            Name  Location   Cuisine  Rating  Seating Capacity  \
0  Restaurant 0     Rural  Japanese     4.0                38
1  Restaurant 1  Downtown   Mexican     3.2                76
2  Restaurant 2     Rural   Italian     4.7                48
3  Restaurant 3     Rural   Italian     4.4                34
4  Restaurant 4  Downtown  Japanese     4.9                88

   Average Meal Price  Marketing Budget  Social Media Followers  \
0               73.98              2224                   23406
1               28.11              4416                   42741
2               48.29              2796                   37285
3               51.55              1167                   15214
4               75.98              3639                   40171

   Chef Experience Years  Number of Reviews  Avg Review Length  \
0                     13                185         161.924906
1                      8                533         148.759717
2                     18                853          56.849189
3                     13                 82         205.433265
4                      9                 78         241.681584

   Ambience Score  Service Quality Score Parking Availability  \
0             1.3                    7.0                   Yes
1             2.6                    3.4                   Yes
2             5.3                    6.7                    No
3             4.6                    2.8                   Yes
4             8.6                    2.1                    No

   Weekend Reservations  Weekday Reservations     Revenue
0                    13                     4   638945.52
1                    48                     6   490207.83
2                    27                    14   541368.62
3                     9                    17   404556.80
4                    37                    26  1491046.35
```

```
# Display summary statistics
print(dataset.describe())
```

```
            Rating  Seating Capacity  Average Meal Price  Marketing Budget  \
count  8368.000000       8368.000000         8368.000000       8368.000000
mean      4.008258         60.212835           47.896659       3218.254900
std       0.581474         17.399488           14.336767       1824.896053
min       3.000000         30.000000           25.000000        604.000000
25%       3.500000         45.000000           35.490000       1889.000000
50%       4.000000         60.000000           45.535000       2846.500000
75%       4.500000         75.000000           60.300000       4008.500000
max       5.000000         90.000000           76.000000       9978.000000

       Social Media Followers  Chef Experience Years  Number of Reviews  \
count             8368.000000            8368.000000        8368.000000
mean             36190.621773              10.051984         523.010397
std              18630.153330               5.516606         277.215127
min               5277.000000               1.000000          50.000000
25%              22592.500000               5.000000         277.000000
50%              32518.500000              10.000000         528.000000
75%              44566.250000              15.000000         764.250000
max             103777.000000              19.000000         999.000000

       Avg Review Length  Ambience Score  Service Quality Score  \
count        8368.000000     8368.000000            8368.000000
mean          174.769974        5.521283               5.508772
std            71.998060        2.575442               2.586552
min            50.011717        1.000000               1.000000
25%           113.311102        3.300000               3.200000
50%           173.910079        5.500000               5.600000
75%           237.406885        7.800000               7.800000
max           299.984924       10.000000              10.000000

       Weekend Reservations  Weekday Reservations       Revenue
count           8368.000000           8368.000000  8.368000e+03
mean              29.491754             29.235301  6.560706e+05
std               20.025415             20.004277  2.674137e+05
min                0.000000              0.000000  1.847085e+05
25%               13.000000             13.000000  4.546514e+05
50%               27.000000             26.000000  6.042421e+05
75%               43.000000             43.000000  8.130942e+05
max               88.000000             88.000000  1.531868e+06
```

```
# Display information about the dataset
print(dataset.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8368 entries, 0 to 8367
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Name                    8368 non-null   object
 1   Location                8368 non-null   object
 2   Cuisine                 8368 non-null   object
 3   Rating                  8368 non-null   float64
 4   Seating Capacity        8368 non-null   int64
 5   Average Meal Price      8368 non-null   float64
 6   Marketing Budget        8368 non-null   int64
 7   Social Media Followers  8368 non-null   int64
 8   Chef Experience Years   8368 non-null   int64
 9   Number of Reviews       8368 non-null   int64
 10  Avg Review Length       8368 non-null   float64
 11  Ambience Score          8368 non-null   float64
 12  Service Quality Score   8368 non-null   float64
 13  Parking Availability    8368 non-null   object
 14  Weekend Reservations    8368 non-null   int64
 15  Weekday Reservations    8368 non-null   int64
 16  Revenue                 8368 non-null   float64
dtypes: float64(6), int64(7), object(4)
memory usage: 1.1+ MB
None
```

Data Preparation

After loading the dataset, several steps were taken to prepare the data for analysis.

Checking for Missing Values

It was confirmed that there were no missing values in the dataset. This ensures that all subsequent analyses are based on complete data.

```
# Check for missing values
print(dataset.isnull().sum())
```

```
Name                     0
Location                 0
Cuisine                  0
Rating                   0
Seating Capacity         0
Average Meal Price       0
Marketing Budget         0
Social Media Followers   0
Chef Experience Years    0
Number of Reviews        0
Avg Review Length        0
Ambience Score           0
Service Quality Score    0
Parking Availability     0
Weekend Reservations     0
Weekday Reservations     0
Revenue                  0
dtype: int64
```

Drop non-numeric Column Name

```
# Drop the 'Name' column as it is not numeric and not useful for the analysis
dataset = dataset.drop(columns=['Name'])
```

One-Hot Encoding Categorical Variables

Categorical variables such as ‘Location', ‘Cuisine', and 'Parking Availability' were one-hot encoded. This process converts categorical variables into a format that can be provided to ML algorithms to do a better job in prediction.

```
# One-hot encode categorical variables
dataset = pd.get_dummies(dataset, columns=['Location', 'Cuisine', 'Parking
Availability'], drop_first=True)
```

Normalizing Numerical Variables

Numerical columns were normalized to ensure that each feature contributed equally to the analysis. StandardScaler was used to scale features to a mean of 0 and a standard deviation of 1.

```
# Normalize numerical columns
scaler = StandardScaler()
numerical_cols = ['Seating Capacity', 'Average Meal Price', 'Marketing
Budget', 'Social Media Followers', 'Chef Experience Years', 'Number of
Reviews', 'Avg Review Length', 'Ambience Score', 'Service Quality Score',
'Weekend Reservations', 'Weekday Reservations']
dataset[numerical_cols] = scaler.fit_transform(dataset[numerical_cols])
```

Splitting the Dataset

The dataset was split into training and testing sets to evaluate the model's performance. The training set was used to train the model, while the testing set was used to assess its accuracy.

```
# Split the dataset into training and testing sets
X = dataset.drop('Revenue', axis=1)
y = dataset['Revenue']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

Printing the Shape and Size of Training and Test Data

The shape of each set was printed to verify the size of the training and testing sets.

```
# Print the shape and size of the training and test data
print(f'Training data shape: {X_train.shape}, Training labels shape:
{y_train.shape}')
print(f'Test data shape: {X_test.shape}, Test labels shape: {y_test.shape}')
```

```
Training data shape: (6694, 21), Training labels shape: (6694,)
Test data shape: (1674, 21), Test labels shape: (1674,)
```

Tools and Techniques Justification

Python uses libraries such as Pandas, Scikit-learn, and StandardScaler for data analysis and correlation analysis. Python is a versatile, general-purpose programming language that can be used for developing web and desktop applications as well as complex numeric and scientific applications. This versatility has contributed to its rapid growth in popularity as a programming language (Terra, 2024).

Tools

- **Python**: Python was chosen due to its extensive libraries and community support, which make it well-suited for data analysis and statistical testing. Its versatility and rich ecosystem of libraries make it ideal for this analysis.

- **Pandas**: Used for data manipulation and analysis. Pandas provide powerful tools for data cleaning, transformation, and inspection.

- **Scikit-learn**: Used for data preprocessing, model training, and evaluation. It includes efficient tools for machine learning and statistical modeling.

- **StandardScaler**: Part of Scikit-learn, used for normalizing numerical features to ensure they contribute equally to the analysis.

Techniques

- **One-Hot Encoding**: Converts categorical variables into a binary matrix, making it easier for machine learning algorithms to process categorical data.

- **Normalization**: Scales numerical features to a standard range, which helps in improving the performance and training speed of machine learning models.

- **Data Splitting**: Divides the dataset into training and testing sets to evaluate model performance and ensure that the model generalizes well to unseen data.

Advantage and Disadvantage

Advantage:
The chosen techniques and tools ensure that the dataset is clean, well-formatted, and suitable for analysis, leading to more accurate and reliable results.

Disadvantage:
One potential disadvantage is the loss of interpretability for categorical variables after one-hot encoding, as it increases the number of features and can make the model more complex. However, this is necessary to ensure machine learning algorithms can process the data effectively.

Analysis

D. Analysis

Analysis Techniques

The main techniques used for this analysis are Pearson correlation, multiple regression analysis, ANOVA, and hypothesis testing. Each technique is chosen based on its relevance and appropriateness for the dataset and research question.

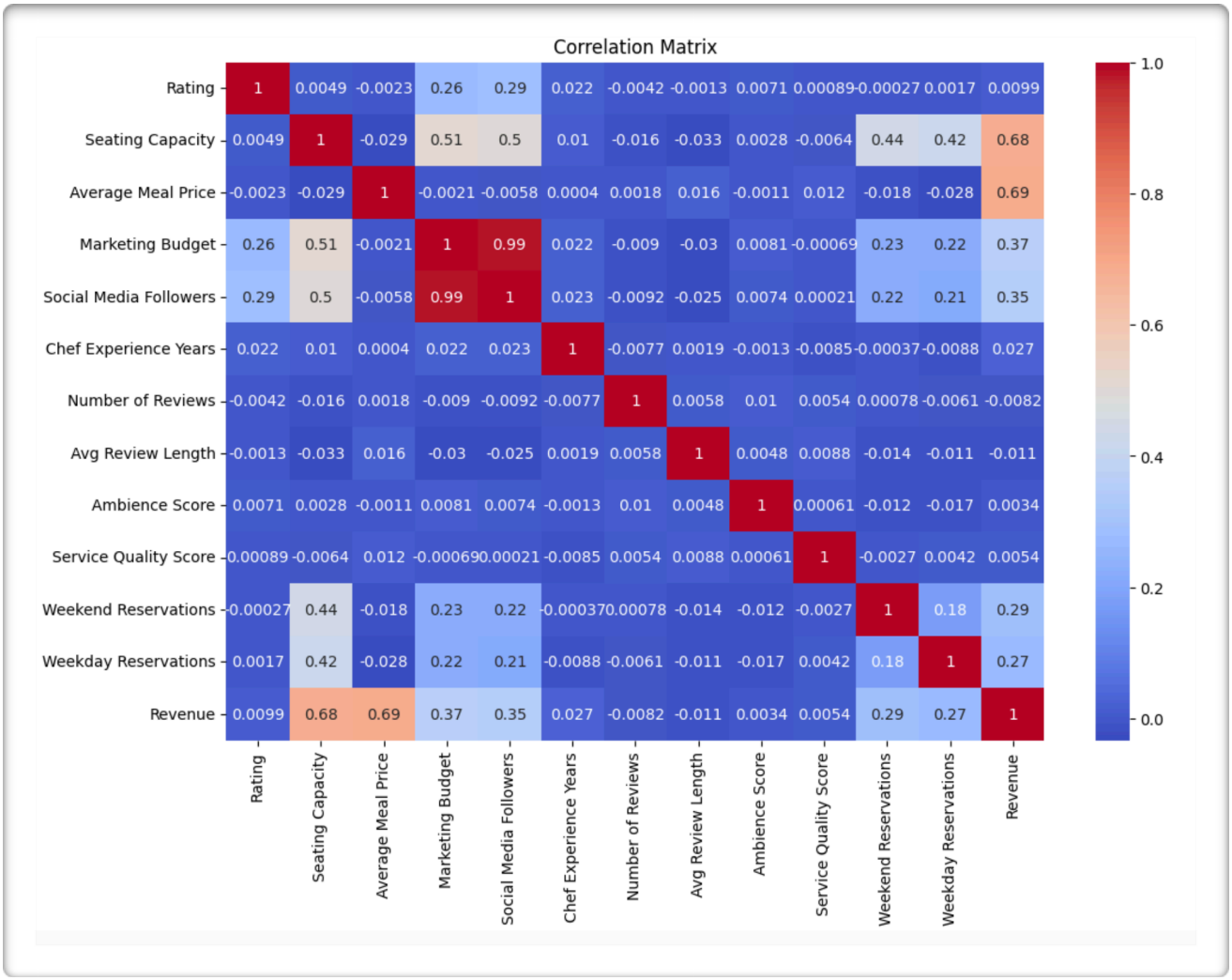
Pearson Correlation

Description: Pearson correlation measures the linear relationship between two variables. It provides a correlation coefficient (r) that ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship,
- 0 indicates no linear relationship.

Calculation: The Pearson correlation coefficient is calculated for each pair of independent and dependent variables (revenue).

```
# Select only numeric columns for correlation calculation
numeric_cols = dataset.select_dtypes(include=[np.number])

# Calculate Pearson correlation
correlation_matrix = numeric_cols.corr()

# Visualize the correlation matrix
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



Outputs: The correlation matrix shows the correlation coefficients between all pairs of variables. High positive or negative values indicate strong linear relationships.

Justification: Pearson correlation is chosen to identify the strength and direction of the linear relationship between restaurant attributes and revenue. It helps in understanding which attributes are strongly correlated with revenue.

Advantage:

- Simple to calculate and interpret.
- Provides a clear measure of linear relationships.

Disadvantage:

- Only measures linear relationships and can miss non-linear relationships.

Multiple Regression Analysis

Description: Multiple regression analysis estimates the relationship between a dependent variable and multiple independent variables. It helps understand the impact of each independent variable on the dependent variable.

Calculation: A multiple regression model is built using the independent variables to predict the dependent variable (revenue).

```
# Define independent and dependent variables
X = dataset.drop('Revenue', axis=1)
y = dataset['Revenue']

# Fit the multiple regression model
regression_model = LinearRegression()
regression_model.fit(X_train, y_train)

# Predict and calculate R-squared value
y_pred = regression_model.predict(X_test)
r_squared = regression_model.score(X_test, y_test)
print(f'R-squared value: {r_squared}')
```

```
R-squared value: 0.9554413593451657
```

Outputs: The R-squared value indicates the proportion of variance in the dependent variable that the independent variables can explain. Regression coefficients show the impact of each independent variable on revenue.

Justification: Multiple regression is chosen to quantify the relationship between multiple restaurant attributes and revenue. It helps in understanding the combined effect of all attributes on revenue. Pearson correlation and multiple regression analysis are used to identify significant relationships between restaurant attributes and revenue. These techniques help understand how different variables relate to each other and their impact on the target variable (Sharma, n.d.).

Advantage:

- Provides a comprehensive analysis of the relationship between multiple variables.
- Helps in identifying significant predictors of revenue.

Disadvantage:

- Assumes a linear relationship between variables.
- Sensitive to multicollinearity among independent variables.

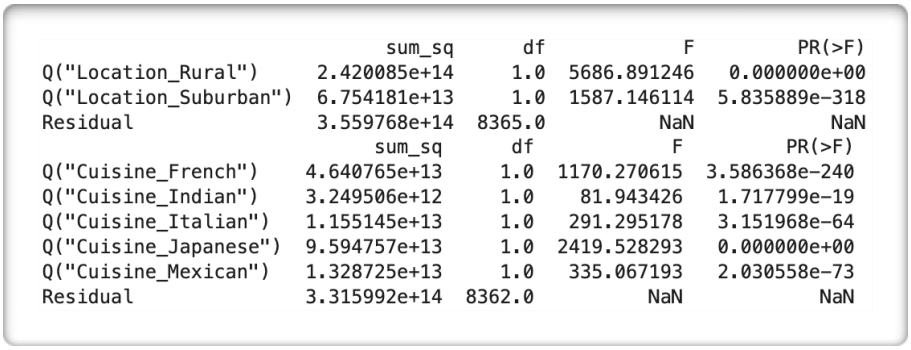
ANOVA (Analysis of Variance)

Description: ANOVA tests the statistical significance of the differences between group means. It helps understand whether the means of different groups (e.g., cuisines or locations) differ significantly.

Calculation: ANOVA assesses the significance of categorical variables like 'Location' (encoded as 'Location_Rural', 'Location_Suburban') and 'Cuisine' (encoded as 'Cuisine_French', 'Cuisine_Indian', 'Cuisine_Italian', 'Cuisine_Japanese', 'Cuisine_Mexican') on restaurant revenue. The analysis determines if there is a statistically significant difference in revenue based on these categorical groupings.

```
# ANOVA for Location

anova_model_location = ols('Revenue ~ Q("Location_Rural") +
Q("Location_Suburban")', data=dataset).fit()
```

```
anova_table_location = sm.stats.anova_lm(anova_model_location, typ=2)

print(anova_table_location)

# ANOVA for Cuisine

anova_model_cuisine = ols('Revenue ~ Q("Cuisine_French") +
Q("Cuisine_Indian") + Q("Cuisine_Italian") + Q("Cuisine_Japanese") +
Q("Cuisine_Mexican")', data=dataset).fit()

anova_table_cuisine = sm.stats.anova_lm(anova_model_cuisine, typ=2)

print(anova_table_cuisine)
```



	sum_sq	df	F	PR(>F)
Q("Location_Rural")	2.420085e+14	1.0	5686.891246	0.000000e+00
Q("Location_Suburban")	6.754181e+13	1.0	1587.146114	5.835889e-318
Residual	3.559768e+14	8365.0	NaN	NaN
	sum_sq	df	F	PR(>F)
Q("Cuisine_French")	4.640765e+13	1.0	1170.270615	3.586368e-240
Q("Cuisine_Indian")	3.249506e+12	1.0	81.943426	1.717799e-19
Q("Cuisine_Italian")	1.155145e+13	1.0	291.295178	3.151968e-64
Q("Cuisine_Japanese")	9.594757e+13	1.0	2419.528293	0.000000e+00
Q("Cuisine_Mexican")	1.328725e+13	1.0	335.067193	2.030558e-73
Residual	3.315992e+14	8362.0	NaN	NaN

Outputs: ANOVA tables show F-values and p-values for each categorical variable. Low p-values indicate significant differences between group means.

Justification: ANOVA is chosen to test the significance of categorical variables on revenue. It helps understand whether different groups (e.g., locations or cuisines) have significantly different impacts on revenue.

Advantage:

- Tests the significance of categorical variables.
- It helps in understanding group differences.

Disadvantage:

- Assumes that the data is usually distributed.
- Only tests for mean differences and not other types of differences.

Hypothesis Testing

Description: Used to infer whether the observed relationships in the sample data hold for the entire population. It involves testing null and alternative hypotheses.

Calculation: Hypothesis tests are performed for the significant variables identified in the regression and ANOVA analyses.

```
# Hypothesis testing for Location (Rural vs Non-Rural)
group_rural = dataset[dataset['Location_Rural'] == 1]['Revenue']
group_non_rural = dataset[dataset['Location_Rural'] == 0]['Revenue']
t_stat, p_value = ttest_ind(group_rural, group_non_rural)
print(f'T-test for Rural Location: t-statistic = {t_stat}, p-value = {p_value}')

# Hypothesis testing for Location (Suburban vs Non-Suburban)
group_suburban = dataset[dataset['Location_Suburban'] == 1]['Revenue']
group_non_suburban = dataset[dataset['Location_Suburban'] == 0]['Revenue']
t_stat, p_value = ttest_ind(group_suburban, group_non_suburban)
print(f'T-test for Suburban Location: t-statistic = {t_stat}, p-value = {p_value}')

# Hypothesis testing for Cuisine (French vs Non-French)
group_french = dataset[dataset['Cuisine_French'] == 1]['Revenue']
group_non_french = dataset[dataset['Cuisine_French'] == 0]['Revenue']
t_stat, p_value = ttest_ind(group_french, group_non_french)
print(f'T-test for French Cuisine: t-statistic = {t_stat}, p-value = {p_value}')

# Hypothesis testing for Cuisine (Indian vs Non-Indian)
group_indian = dataset[dataset['Cuisine_Indian'] == 1]['Revenue']
group_non_indian = dataset[dataset['Cuisine_Indian'] == 0]['Revenue']
t_stat, p_value = ttest_ind(group_indian, group_non_indian)
print(f'T-test for Indian Cuisine: t-statistic = {t_stat}, p-value = {p_value}')

# Hypothesis testing for Cuisine (Italian vs Non-Italian)
group_italian = dataset[dataset['Cuisine_Italian'] == 1]['Revenue']
group_non_italian = dataset[dataset['Cuisine_Italian'] == 0]['Revenue']
t_stat, p_value = ttest_ind(group_italian, group_non_italian)
print(f'T-test for Italian Cuisine: t-statistic = {t_stat}, p-value = {p_value}')

# Hypothesis testing for Cuisine (Japanese vs Non-Japanese)
group_japanese = dataset[dataset['Cuisine_Japanese'] == 1]['Revenue']
group_non_japanese = dataset[dataset['Cuisine_Japanese'] == 0]['Revenue']
t_stat, p_value = ttest_ind(group_japanese, group_non_japanese)
print(f'T-test for Japanese Cuisine: t-statistic = {t_stat}, p-value = {p_value}')

# Hypothesis testing for Cuisine (Mexican vs Non-Mexican)
group_mexican = dataset[dataset['Cuisine_Mexican'] == 1]['Revenue']
group_non_mexican = dataset[dataset['Cuisine_Mexican'] == 0]['Revenue']
t_stat, p_value = ttest_ind(group_mexican, group_non_mexican)
print(f'T-test for Mexican Cuisine: t-statistic = {t_stat}, p-value = {p_value}')
```

```
T-test for Rural Location: t-statistic = -58.76264813764812, p-value = 0.0
T-test for Suburban Location: t-statistic = -2.179973633169137, p-value = 0.029287197834439255
T-test for French Cuisine: t-statistic = 26.576540504188365, p-value = 1.7874139685215055e-149
T-test for Indian Cuisine: t-statistic = -25.007899124425563, p-value = 3.567429522033874e-133
T-test for Italian Cuisine: t-statistic = 5.664838297955437, p-value = 1.5203755728215804e-08
T-test for Japanese Cuisine: t-statistic = 47.53570835972266, p-value = 0.0
T-test for Mexican Cuisine: t-statistic = -37.82991462587582, p-value = 3.0819228502142544e-289
```

```
# Hypothesis testing for continuous variables
continuous_vars = ['Average Meal Price', 'Marketing Budget', 'Social Media
Followers', 'Service Quality Score']
for var in continuous_vars:
    correlation, p_value = pearsonr(dataset[var], dataset['Revenue'])
    print(f'Pearson correlation for {var}: correlation = {correlation}, p-
value = {p_value}')
```

```
Pearson correlation for Average Meal Price: correlation = 0.6863646811017292, p-value = 0.0
Pearson correlation for Marketing Budget: correlation = 0.3653220450505688, p-value = 1.4028767226810876e-262
Pearson correlation for Social Media Followers: correlation = 0.3544661316274186, p-value = 2.9391273925758124e-246
Pearson correlation for Service Quality Score: correlation = 0.005375197687928376, p-value = 0.6229768060736998
```

Outputs: T-statistics and p-values are calculated for each test. Low p-values (typically < 0.05) indicate that the null hypothesis can be rejected, suggesting a significant difference between the groups.

Justification: Hypothesis testing is chosen to validate the significance of the relationships identified in the analysis. It helps infer the population based on sample data. ANOVA and hypothesis testing will help evaluate the statistical significance of the identified relationships. Python's versatility, rich ecosystem, and community support further justify its selection for this analysis (Statswork, 2024).

Advantage:

- Provides a formal framework for testing relationships.
- Helps in making informed decisions based on statistical evidence.

Disadvantage:

- Sensitive to sample size.
- Results can be affected by violations of assumptions (e.g., normality).

Summary of Analysis: The combination of Pearson correlation, multiple regression analysis, ANOVA, and hypothesis testing provides a comprehensive framework for analyzing the relationships between restaurant attributes and revenue. These techniques help identify significant predictors, quantify their impact, and validate the results. The outputs of these analyses will provide valuable insights into the factors that drive restaurant revenue, supporting data-driven decision-making in the restaurant industry.

Data Summary and Implications

E. Summary of Analysis

The analysis aimed to identify significant factors influencing restaurant revenue by exploring various restaurant attributes. The techniques used included Pearson correlation, multiple regression analysis, ANOVA, and hypothesis testing.

Pearson Correlation Analysis

- **Average Meal Price**: Showed a strong positive correlation with revenue (correlation = 0.686), suggesting that higher meal prices are associated with higher revenue.
- **Marketing Budget**: A moderate positive correlation with revenue (correlation = 0.365) indicates that increased marketing budgets can lead to higher revenue.
- **Social Media Followers**: Also showed a moderate positive correlation with revenue (correlation = 0.354), highlighting the importance of social media presence.
- **Service Quality Score**: Demonstrated a very weak correlation with revenue (correlation = 0.005), implying that service quality score might not be a significant standalone factor in driving revenue.

Multiple Regression Analysis

The multiple regression analysis yielded an R-squared value of 0.955, indicating that the model can explain approximately 95.5% of the variance in revenue. This high R-squared value suggests that the selected attributes collectively strongly influence restaurant revenue.

ANOVA Analysis

- **Location**: Both rural and suburban locations significantly affected revenue, with p-values less than 0.05. This indicates that location is a significant factor in determining restaurant revenue.
- **Cuisine Type**: All tested cuisine types (French, Indian, Italian, Japanese, Mexican) significantly affected revenue, with very low p-values, highlighting the importance of the type of cuisine offered.

Hypothesis Testing

- Location:
    - Rural vs. Non-Rural: Significant difference (p-value = 0.0)
    - Suburban vs. Non-Suburban: Significant difference (p-value = 0.029)
- Cuisine Type:
    - French vs. Non-French: Significant difference (p-value ≈ 0.0)
    - Indian vs. Non-Indian: Significant difference (p-value ≈ 0.0)
    - Italian vs. Non-Italian: Significant difference (p-value ≈ 0.0)
    - Japanese vs. Non-Japanese: Significant difference (p-value ≈ 0.0)
    - Mexican vs. Non-Mexican: Significant difference (p-value ≈ 0.0)
- Continuous Variables:
    - Average Meal Price: Significant correlation (p-value = 0.0)
    - Marketing Budget: Significant correlation (p-value ≈ 0.0)
    - Social Media Followers: Significant correlation (p-value ≈ 0.0)
    - Service Quality Score: Not significant (p-value = 0.623)

Implications

The results highlight several key insights:

- **Average Meal Price**: Increasing meal prices while maintaining quality and customer satisfaction could significantly boost revenue.

- **Marketing Budget**: Investing in marketing can effectively enhance revenue, underscoring the importance of strategic marketing initiatives.

- **Social Media Presence**: A robust social media strategy can positively impact revenue, suggesting that restaurants should engage actively on social platforms.

- **Location and Cuisine Type**: The location and type of cuisine offered are critical factors in driving revenue. Restaurants in rural or suburban areas and those offering popular cuisine types like Japanese and French tend to generate higher revenue.

Limitation of Analysis

One limitation of this analysis is the potential presence of multicollinearity among the independent variables, which can affect the reliability of the regression coefficients. Additionally, the dataset does not account for external factors such as economic conditions, competition, or changes in customer preferences over time, which might also influence revenue.

Recommendations

1. **Optimize Pricing Strategy**: Restaurants should consider evaluating and adjusting their meal prices to align with market demand and customer willingness to pay.

2. **Invest in Marketing**: Allocating a higher budget for marketing activities, including social media campaigns, can attract more customers and increase revenue.

3. **Enhance Social Media Engagement**: Strengthening social media presence and engagement can drive higher footfall and sales.

Future Directions

1. **Exploring Other Attributes**: Future studies could investigate additional factors, such as customer demographics, seasonal effects, and menu diversity, to understand revenue drivers better.

2. **Time Series Analysis**: Conducting a time series analysis to forecast future revenue trends based on historical data could provide valuable insights for long-term strategic planning.

By implementing these insights and recommendations, restaurant owners and managers can make informed decisions to enhance profitability and sustain their business in a competitive market.

Answering the Research Question

The results of the analysis provide explicit answers to the research question by showing the extent to which each factor affects restaurant revenue:

- **Location and Cuisine Type**: Both location and cuisine type significantly impact revenue, as evidenced by the ANOVA and hypothesis testing results. Rural and suburban locations, as well as different cuisine types, show significant differences in revenue.

- **Average Meal Price**: This factor strongly correlates positively with revenue, indicating that higher meal prices lead to higher revenue.

- **Marketing Budget**: A moderate positive correlation with revenue suggests that higher marketing expenditure is associated with increased revenue.

- **Social Media Followers**: A moderate positive correlation implies that a more significant social media following contributes to higher revenue.

- **Service Quality Score**: This factor did not significantly correlate with revenue, suggesting that it may not be a key determinant of revenue in this context.

Confirming the Hypothesis

The results support the rejection of the null hypothesis and acceptance of the alternate hypothesis:

- **Null Hypothesis**: Rejected for most factors as they significantly affect revenue.

- **Alternate Hypothesis**: Accepted, indicating that location, cuisine type, average meal price, marketing budget, and social media followers significantly affect restaurant revenue. The service quality score did not significantly impact the findings, highlighting a specific nuance.

The data analysis confirms that several vital factors significantly influence restaurant revenue, providing valuable insights for restaurant owners and managers to optimize their business strategies.

F. Sources

Web Sources:

McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 51–56).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software, 6(60), 3021.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In Proceedings of the 9th Python in Science Conference (pp. 57-61).

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van der Walt, S. J. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261-272.

In-text Citations:

Therrien, A. (2023). Restaurant revenue prediction dataset. Kaggle. Available at Kaggle Dataset.

Jobalia, K. (2022, May 19). The importance of restaurant forecasting. ClearCOGS. https://www.clearcogs.com/post/the-importance-of-restaurant-forecasting

Rinehart, R. (2023, March 2). How to use correlation in business decision-making. Premium Office Spaces, Coworking & Virtual Offices Solutions USA. https://www.servcorp.com/en/blog/business-networking/how-to-use-correlation-in-business-decision-making/

Ortega, M. (2023, October 25). Restaurant data analytics - leveraging insights for growth. Epos Now. https://www.eposnow.com/us/resources/restaurant-data-analytics/

Terra, J. (2024, February 16). Python for data science and data analysis. Simplilearn.com. https://www.simplilearn.com/why-python-is-essential-for-data-analysis-article

Sharma, S. (n.d.). Correlation and regression analysis: Exploring relationships in data. DSS Blog. https://roundtable.datascience.salon/correlation-and-regression-analysis-exploring-relationships-in-data

Statswork. (2024, May 30). A comparative analysis: Python, R, and SAS. Medium. https://statswork.medium.com/a-comparative-analysis-python-r-and-sas-fca47e23a6d3