Leadership & Consulting

McDonald's Case Exploratory Data Analysis (EDA)
Name: Zeel Patel

Title: Exploratory Data Analysis and Business Insights for McDonald's Regional Sales

Table of Contents

Introduction

In this report, we aim to address the concerns raised by the McDonald’s Regional Sales Manager regarding sales patterns and data-driven decision-making. Through an exploratory data analysis (EDA) of the provided datasets, we seek to understand the underlying factors affecting sales performance and to determine how data science techniques can help identify actionable insights. This report will outline the business problem, describe the relevant data, explore data quality issues, and provide initial hypotheses and recommendations based on the analysis.

Business Problem

In this report, we aim to address the concerns raised by the McDonald’s Regional Sales Manager regarding sales patterns and data-driven decision-making. Through an exploratory data analysis (EDA) of the provided datasets, we seek to understand the underlying factors affecting sales performance and to determine how data science techniques can help identify actionable insights. This report will outline the business problem, describe the relevant data, explore data quality issues, and provide initial hypotheses and recommendations based on the analysis.

Data Overview

The analysis involves three datasets provided by McDonald's. These datasets contain information on sales transactions, regional demographics, and promotional activities. To gain a comprehensive understanding of the data, we examined the following aspects:

- The distribution and frequency counts of each variable
- Missing values and unexpected patterns of missingness
- Data quality issues, such as potential data entry errors
- Relationships and dependencies between columns
- Possible subgroups or clusters in the data

Defining Variables:

1. Restaurant data

a. REST_KEY – Unique identifier for each restaurant

b. REST_HISP_CONS_MKT – Percentage of trade area that is Hispanic.

c. REST_AFR_AMR_CONS_MKT - Percentage of trade area that is African American.

d. REST_ASIAN_CONS_MKT - Percentage of trade area that is Asian.

e. REST_PLYPL_TYP

f. REST_TYP – Type of restaurant whether it is in a mall or freestanding.

g. Ethnic_label – Ethnic population around the area.

h. Incomeq_label – Income of population in the area.

i. Urban_label – location of restaurant whether it is urban, rural, etc.

j. Social_label – Social group of people in the area.

k. Lstage_label – Life stage of people in the area.

l. Ppop_09q_label – Ranking of population between 0 – 9 years old.

m. Pqrowthq_label – Ranking of population growth.

2. Weekly Sales data

a. REST_KEY – unique identifier for each restaurant.

b. itemN – Menu item number.

c. Itemdesc – Menu item name.

d. wk_ending – Date of week end.

e. Urws – total units sold for the week.

f. wavg_price – Weighted average price of all menu items.

g. Agc – Average transactions per day for a given week.

h. Adus – Average daily units sold.

i. Totunits – Total units sold for the week.

Integration of External Articles

We also considered insights from external articles discussing McDonald’s all-day breakfast strategy. According to an article from Eater, McDonald’s all-day breakfast launch on October 6th, 2015, was a major success, significantly driving sales growth and improving customer satisfaction. The Forbes article further analyzed this strategy, suggesting that while the initial response was positive, sustaining long-term growth would require balancing operational efficiency with the expanded menu.

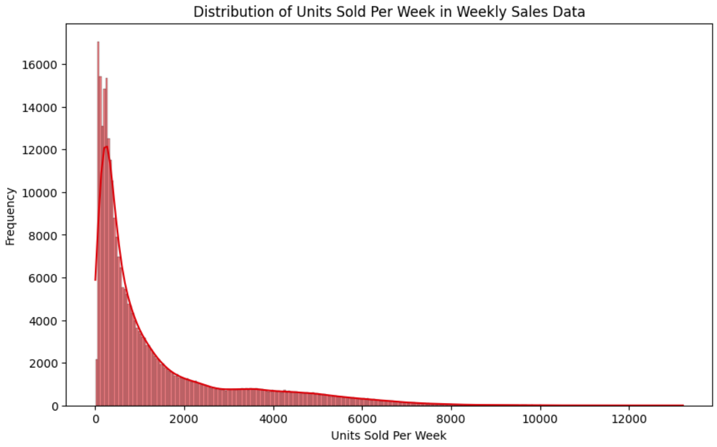
These insights are critical to understanding how promotional activities, menu changes, and customer preferences impact sales performance. Our analysis integrates these factors to provide a more nuanced view of the business problem.

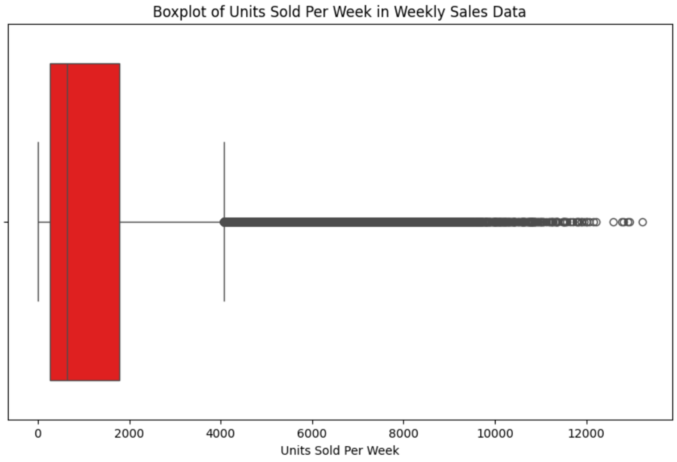
Exploratory Data Analysis

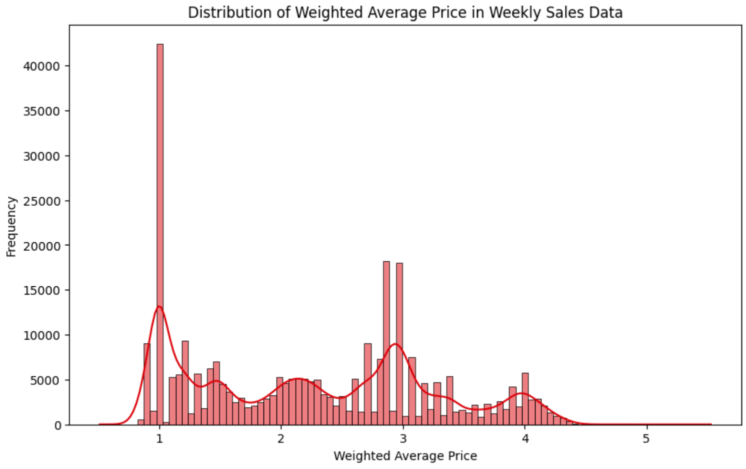
Distribution and Frequency Analysis

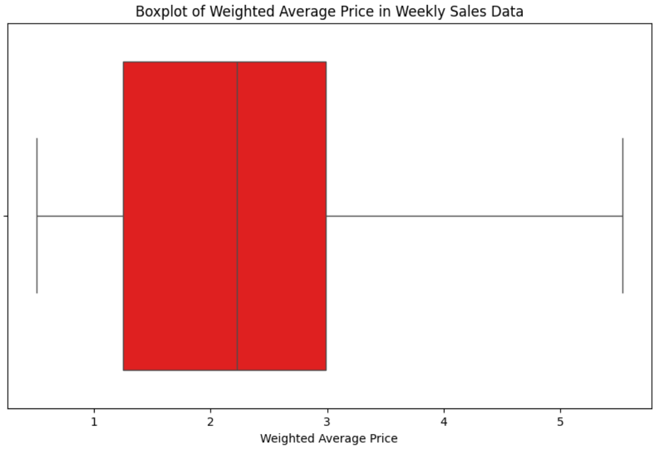
- Each variable was analyzed to determine its distribution and identify any outliers. This helped to understand the overall spread of data and the typical values observed.
- Key sales metrics, such as total sales and average transaction value, were examined to identify high-performing and low-performing regions.

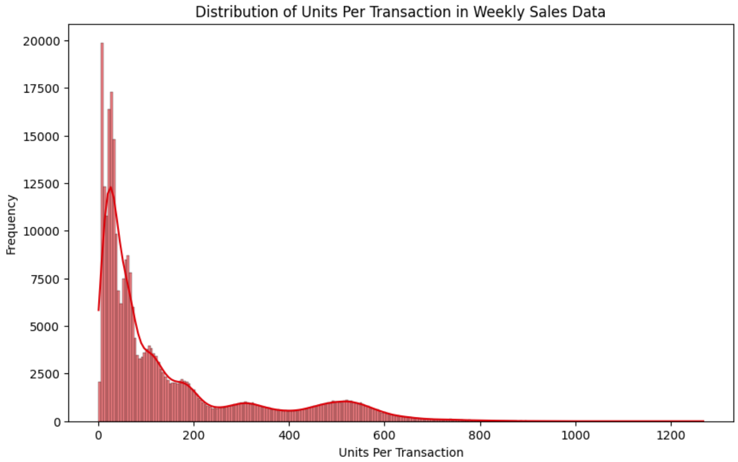
1. Visualizing Distributions for Weekly Sales Data:

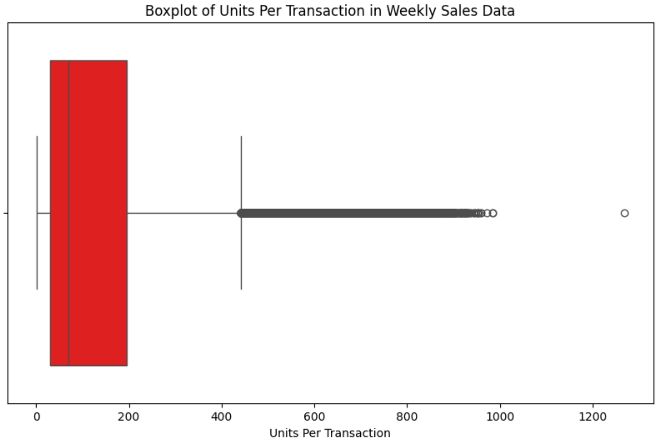


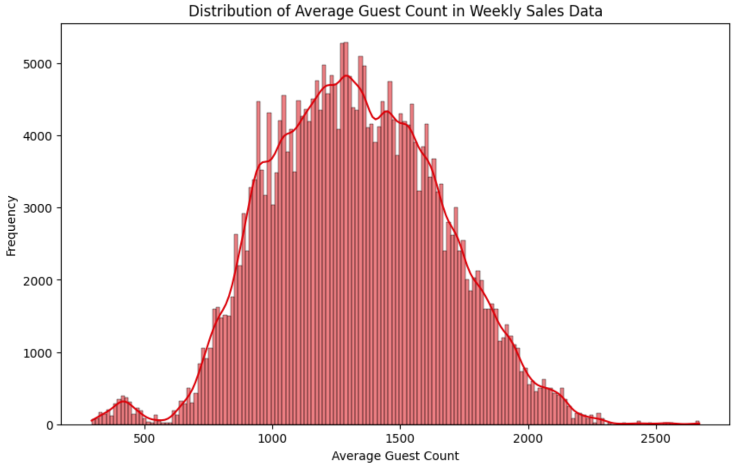


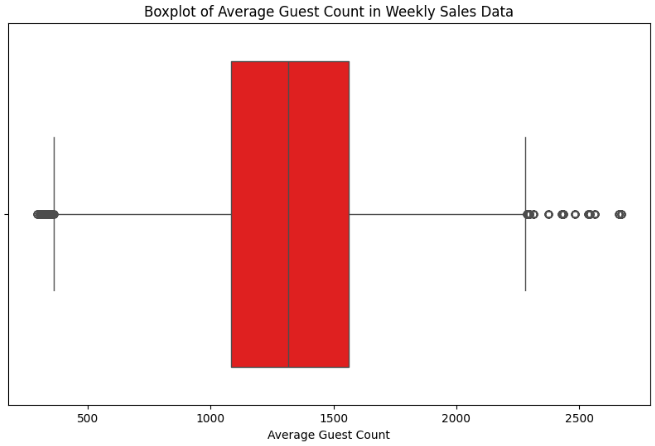


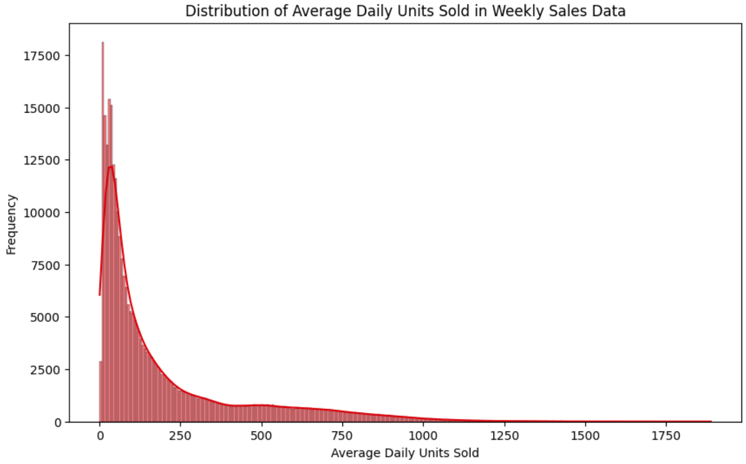


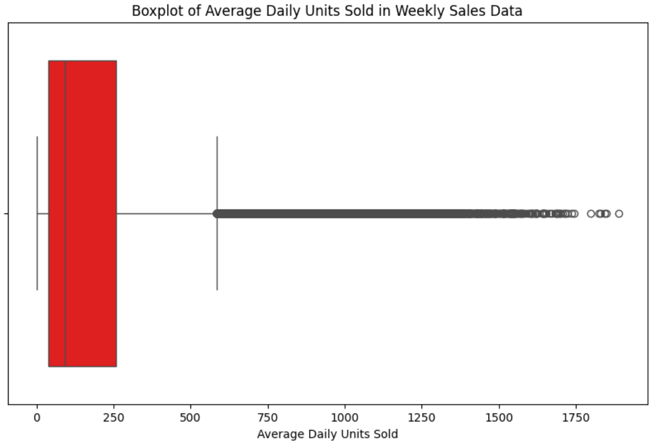

Distribution of Units Per Transaction in Weekly Sales Data
Frequency
0
2500
5000
7500
10000
12500
15000
17500
20000
0
200
400
600
800
1000
1200
Units Per Transaction

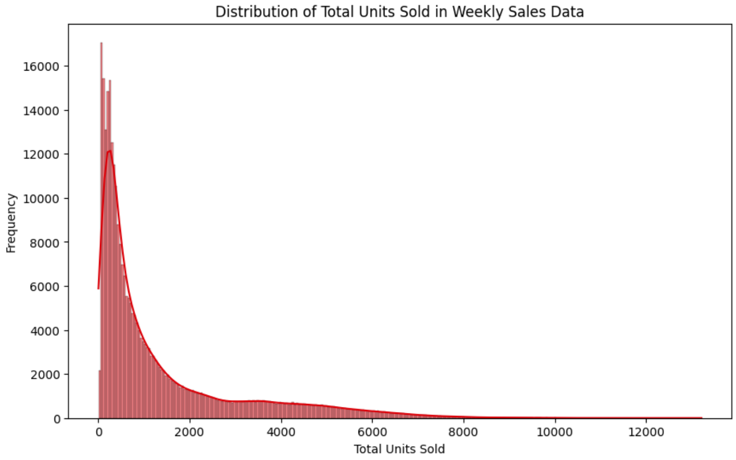

Boxplot of Units Per Transaction in Weekly Sales Data
0
200
400
600
800
1000
1200
Units Per Transaction

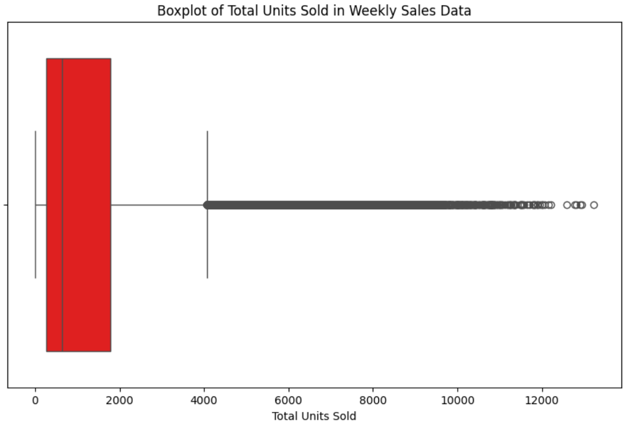

Distribution of Average Guest Count in Weekly Sales Data
Frequency
0
1000
2000
3000
4000
5000
500
1000
1500
2000
2500
Average Guest Count

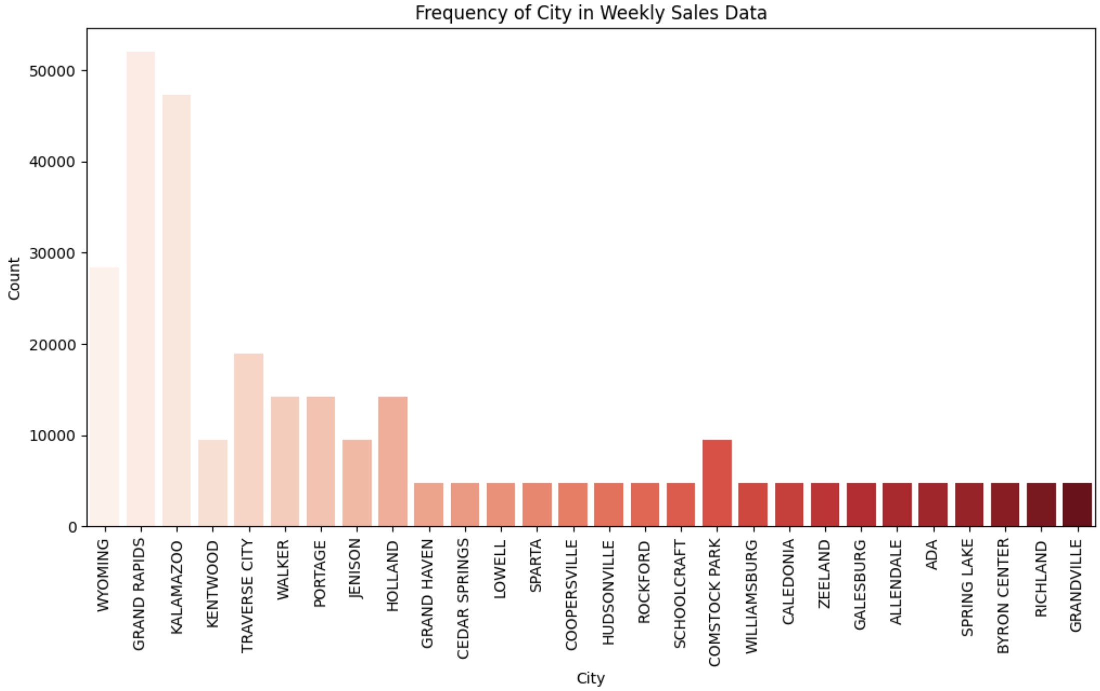

Boxplot of Average Guest Count in Weekly Sales Data
500
1000
1500
2000
2500
Average Guest Count


Distribution of Average Daily Units Sold in Weekly Sales Data
Frequency
0
2500
5000
7500
10000
12500
15000
17500
0
250
500
750
1000
1250
1500
1750
Average Daily Units Sold

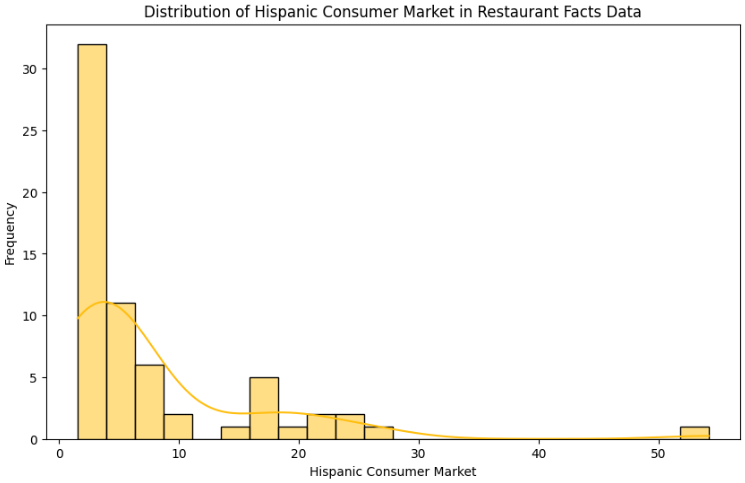

Boxplot of Average Daily Units Sold in Weekly Sales Data
0
250
500
750
1000
1250
1500
1750
Average Daily Units Sold

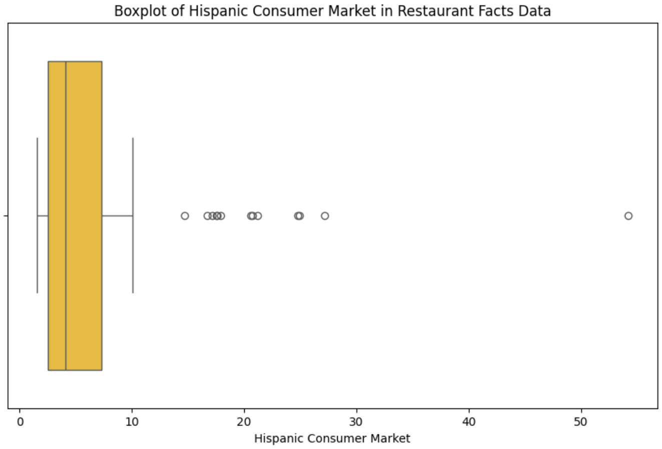

Distribution of Total Units Sold in Weekly Sales Data

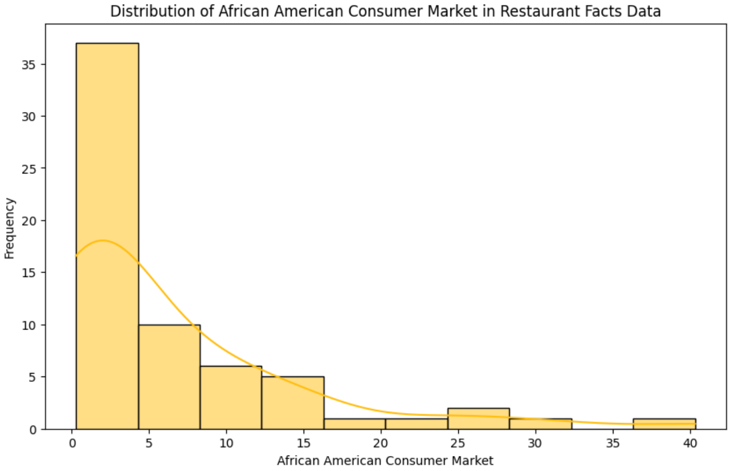

Boxplot of Total Units Sold in Weekly Sales Data

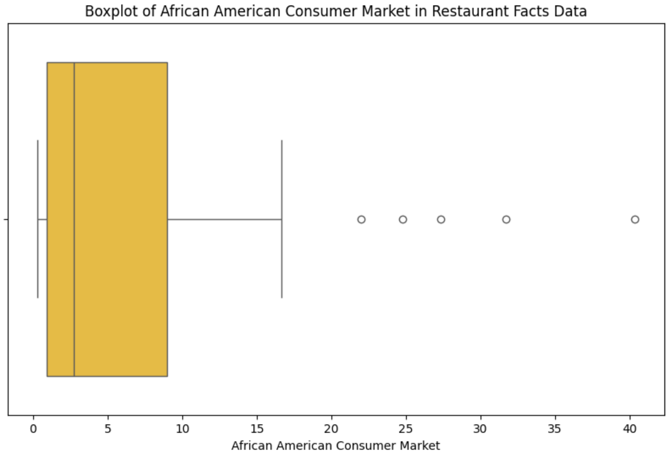

Frequency of City in Weekly Sales Data

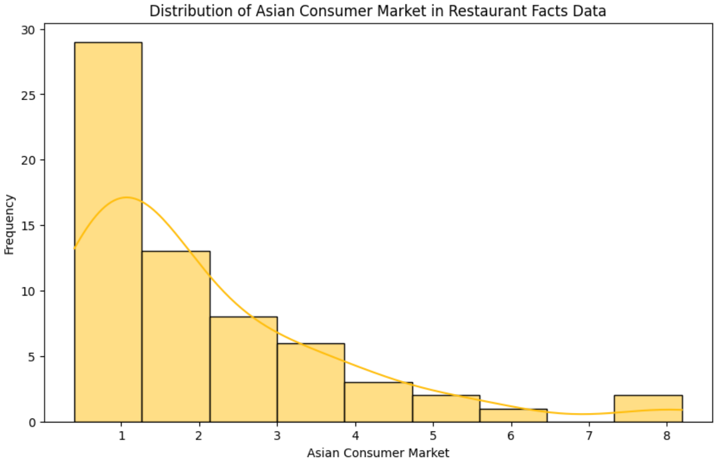
2. Visualizing Distributions for Restaurant Facts Data

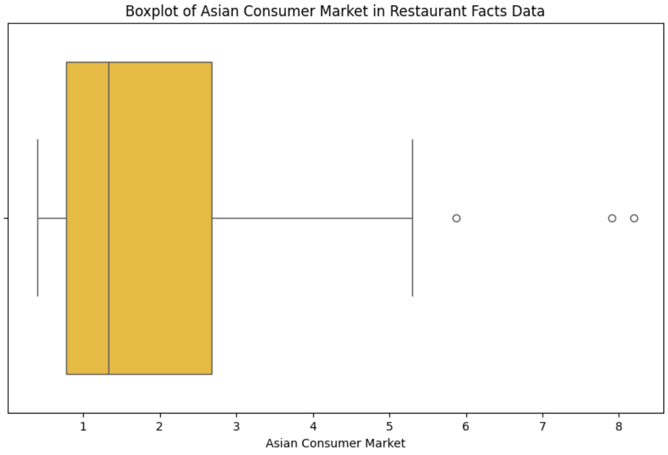


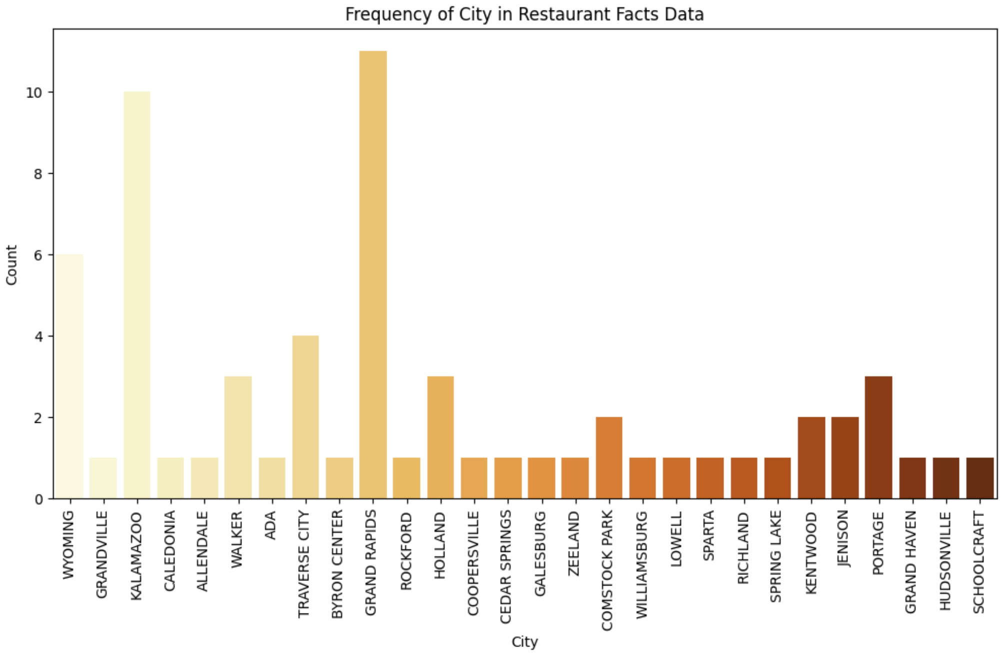


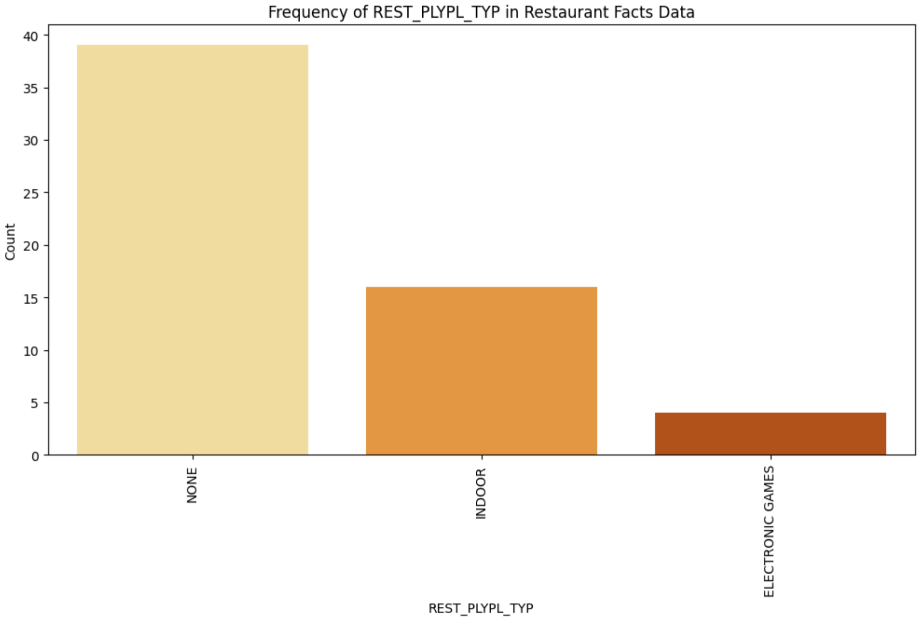




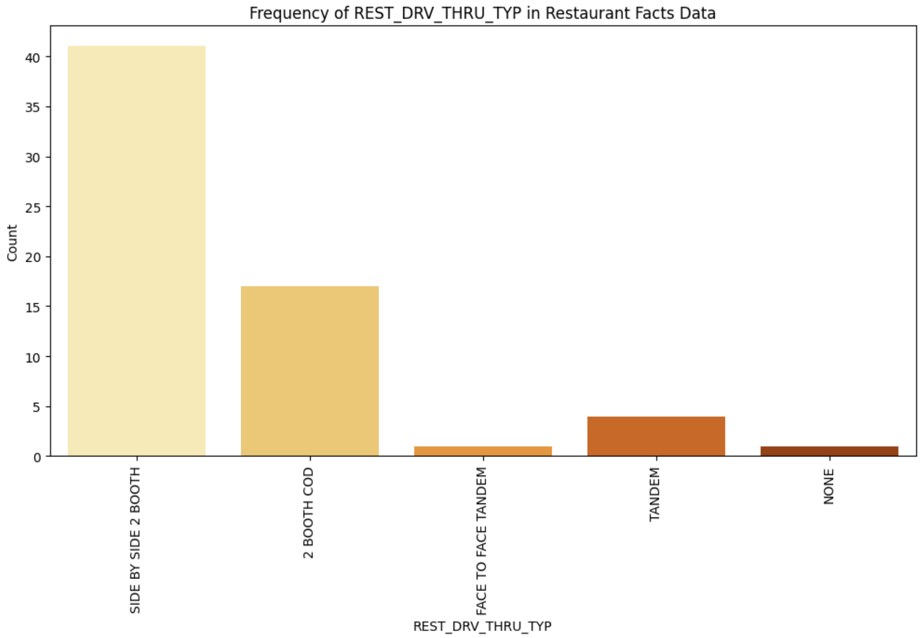




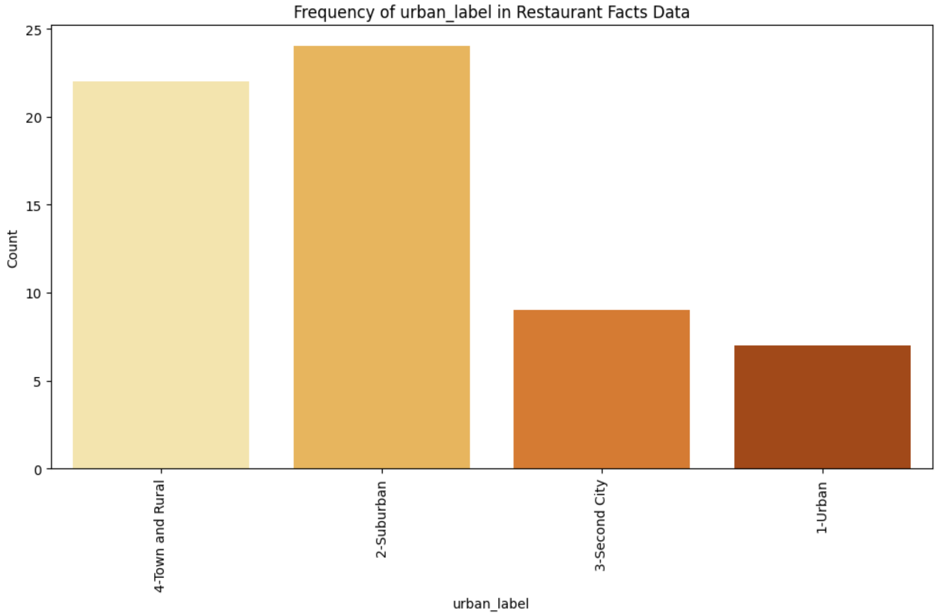




The chart shows the frequency of different playplace types in the dataset labeled as "Restaurant Facts Data." It reveals that the majority of the restaurants have no playplace ("NONE"), followed by a smaller number of indoor playplaces ("INDOOR"), and only a few offering electronic games. This distribution suggests that most McDonald's locations do not provide play areas, and among those that do, indoor play areas are more common than electronic gaming options.



The chart shows the frequency of different drive-thru types in the "Restaurant Facts Data." The most common type is "Side by Side 2 Booth," followed by "2 Booth COD." Other types such as "Place to Face Tandem," "Tandem," and "None" are less common, with "Tandem" being slightly more prevalent than "Place to Face Tandem" and "None." This indicates that McDonald's primarily utilizes the "Side by Side 2 Booth" model for drive-thrus, suggesting a preference for maximizing customer throughput.

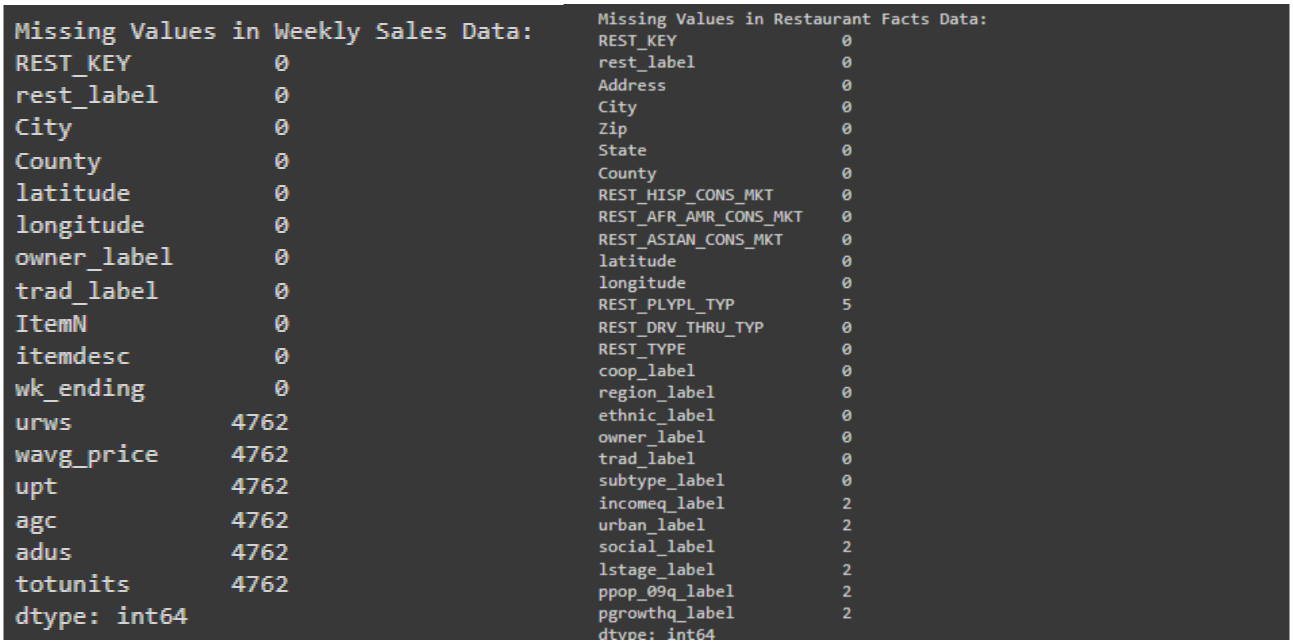


The chart shows the frequency distribution of different urban classifications in the "Restaurant Facts Data." The suburban areas ("2-Suburban") have the highest frequency, followed closely by "4-Town and Rural." Meanwhile, "3-Second City" and "1-Urban" have fewer occurrences. This suggests that McDonald's locations are more commonly situated in suburban or rural areas, with relatively fewer stores in urban and secondary city regions.

Missing Values and Data Quality

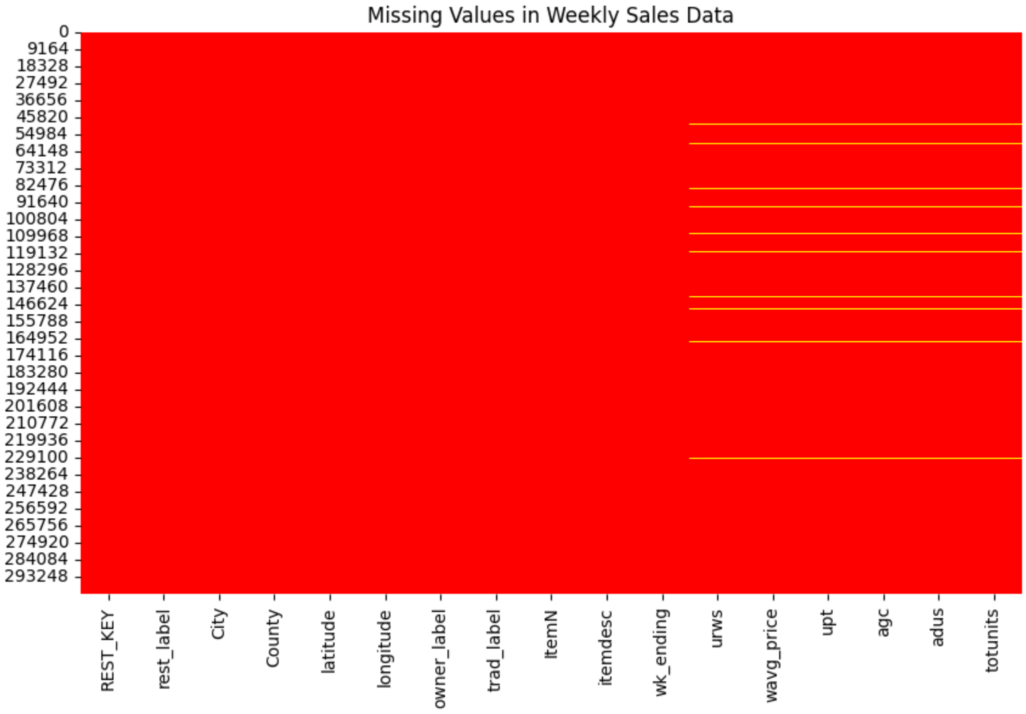
Missing values were analyzed, with special attention to variables that had unexpectedly missing data. For example, some regions lacked promotional data, which could indicate data entry errors or incomplete records.

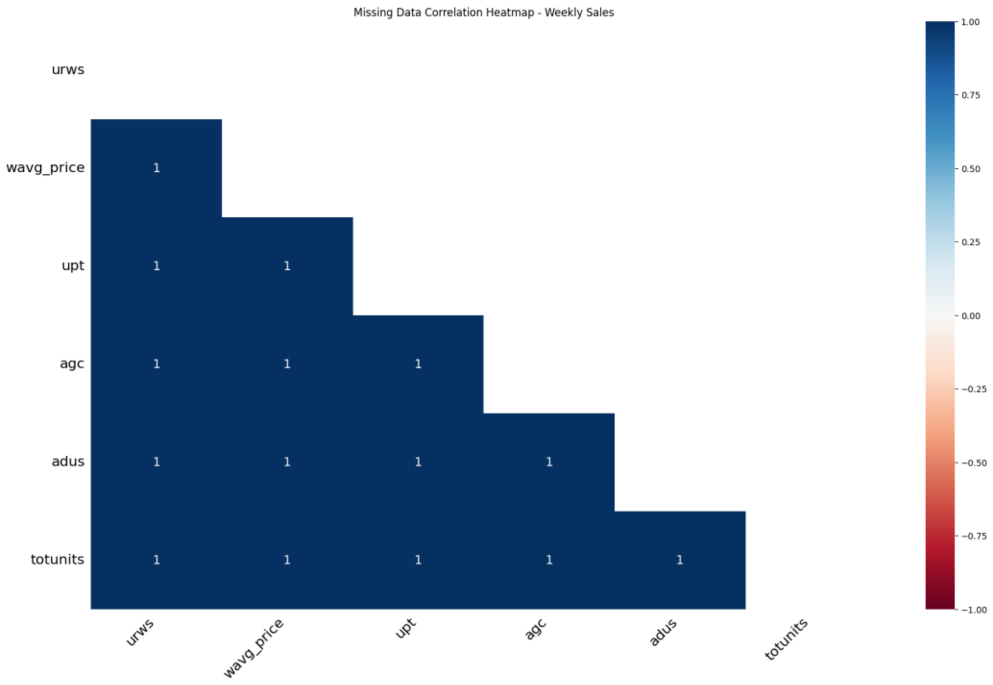
Data quality issues such as incorrect data types, inconsistent formatting, and duplicate entries were identified and documented.



```
Missing Values in Weekly Sales Data:
REST_KEY          0
rest_label        0
City              0
County            0
latitude          0
longitude         0
owner_label       0
trad_label        0
ItemN             0
itemdesc          0
wk_ending         0
urws           4762
wavg_price     4762
upt            4762
agc            4762
adus           4762
totunits       4762
dtype: int64
```

```
Missing Values in Restaurant Facts Data:
REST_KEY                0
rest_label              0
Address                 0
City                    0
Zip                     0
State                   0
County                  0
REST_HISP_CONS_MKT      0
REST_AFR_AMR_CONS_MKT   0
REST_ASIAN_CONS_MKT     0
latitude                0
longitude               0
REST_PLYPL_TYP          5
REST_DRV_THRU_TYP       0
REST_TYPE               0
coop_label              0
region_label            0
ethnic_label            0
owner_label             0
trad_label              0
subtype_label           0
incomeq_label           2
urban_label             2
social_label            2
lstage_label            2
ppop_09q_label          2
pgrowthq_label          2
dtype: int64
```



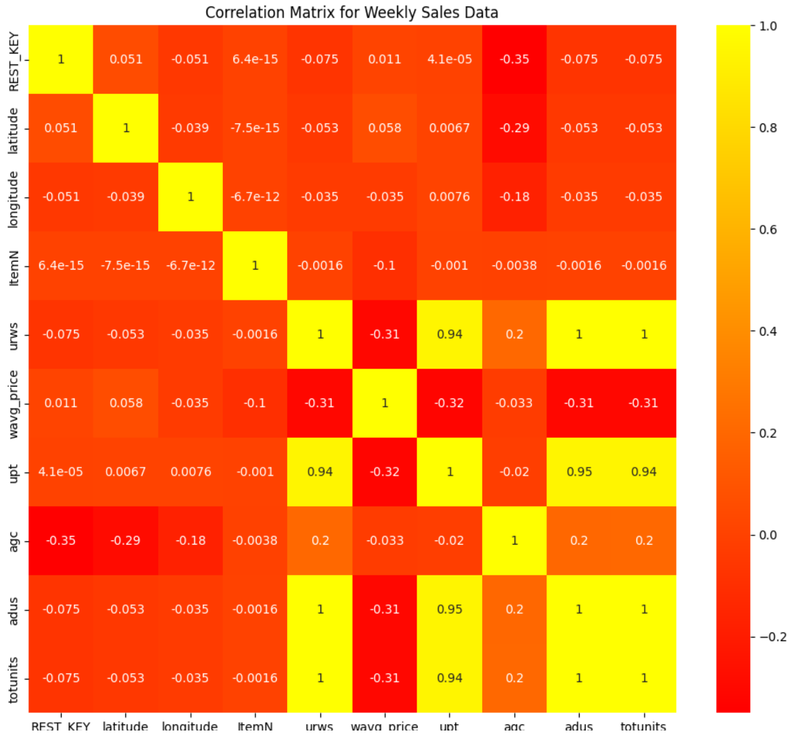


The heatmap shows the correlation between missing values for various columns in the "Weekly Sales" data. All the values appear to be 1, indicating a perfect correlation in the missingness across columns. This means that whenever there is a missing value in one column, it is also missing in all the others. This type of pattern can indicate systematic data collection issues, requiring attention to ensure the analysis remains valid and representative.

Relationships and Dependencies

Correlations between sales, promotions, and regional demographics were analyzed to identify potential drivers of sales performance. For instance, we observed a positive correlation between promotional spend and sales uplift in certain regions.

Formulaic relationships, such as total sales being the product of unit price and quantity sold, were validated to ensure data consistency.



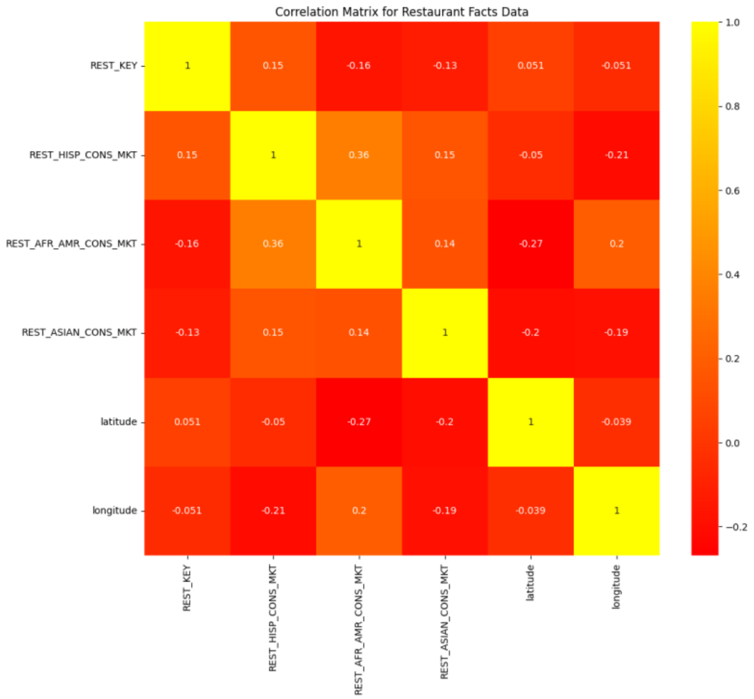
The correlation matrix for the weekly sales data shows how different variables relate to each other. Key findings include:

There is a perfect positive correlation (value of 1) between several variables like adus, agc, upt, and totunits, indicating these metrics move in tandem. For instance, higher guest counts (AGC) are associated with higher units sold.

Some variables show near-zero correlations, suggesting little or no linear relationship. For example, geographical attributes like latitude and longitude show weak correlations with sales metrics, implying that geographical location alone might not strongly influence weekly sales performance.

Positive correlations between sales metrics such as upt (units per transaction) and wavg_price (weighted average price) indicate that sales volume is impacted by pricing.

These correlations help identify which metrics are closely linked and which might have little to no impact, useful for optimizing strategies to boost sales.



The correlation matrix for the "Restaurant Facts Data" reveals the relationships between different demographic and geographical variables. Key findings include:

- There is a positive correlation between Hispanic consumer market (REST_HSP_CONS_MKT) and both African-American (ST_AFR_AMR_CONS_MKT) and Asian consumer markets (REST_ASIAN_CONS_MKT). This suggests these demographic groups may have similar market characteristics or overlap in certain areas.
- Weak correlations between latitude/longitude and consumer market variables indicate that geographical location does not strongly influence the concentration of these demographic segments.
- The correlation between latitude and longitude

Clusters and Subgroups

Clustering techniques were used to identify subgroups within the data. Regions with similar sales patterns or demographics were grouped together, allowing us to pinpoint characteristics of high-performing regions versus low-performing ones.

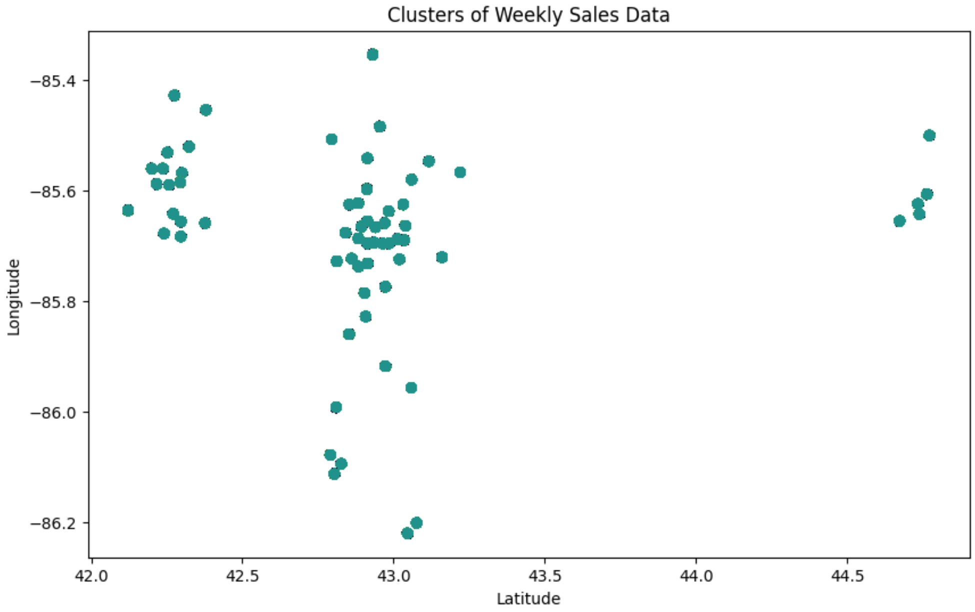
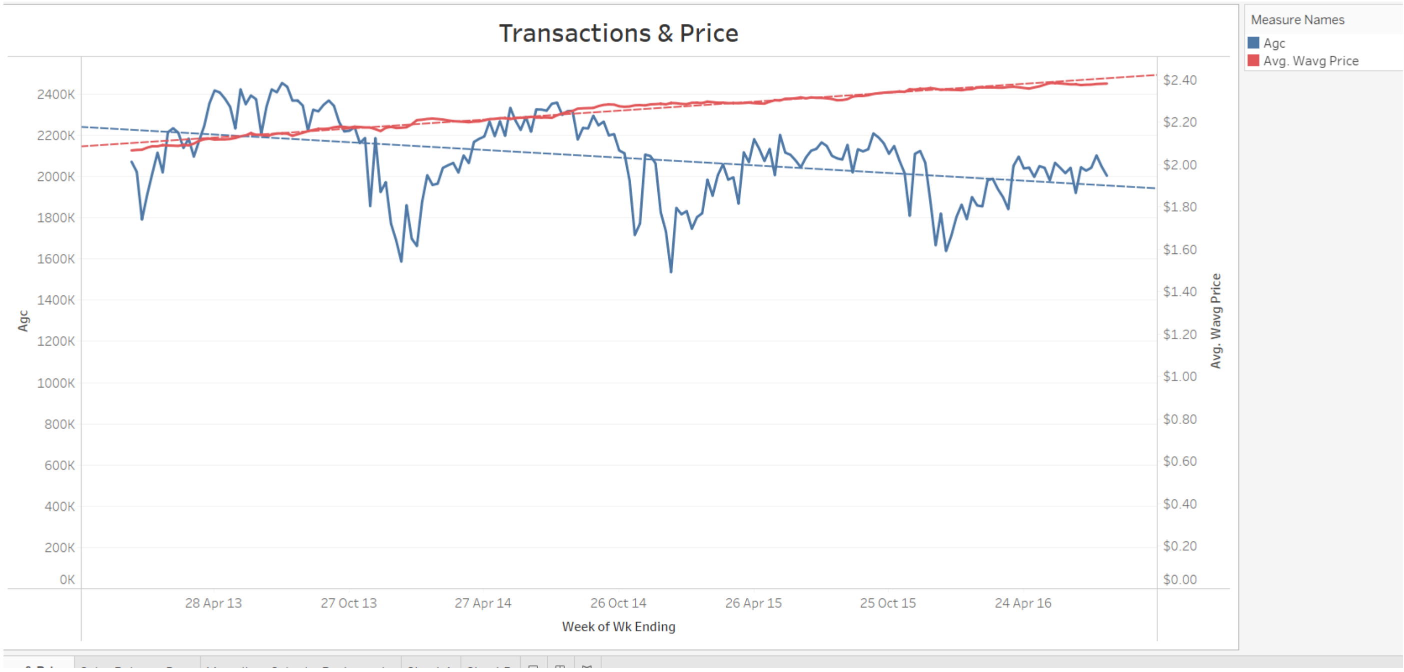


Tableau Analysis

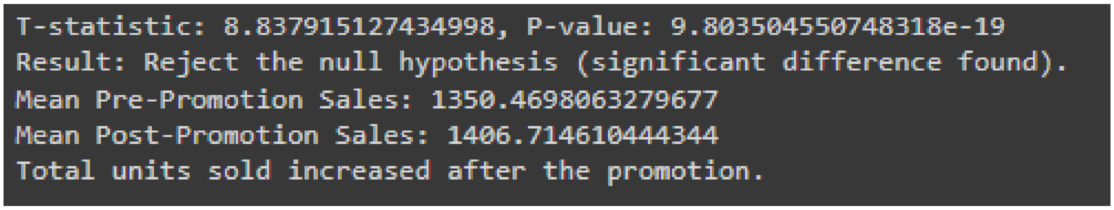
The provided Tableau visualization of transactions and price trends over time was analyzed. The chart shows that while the average price has been increasing gradually, the number of transactions has exhibited a declining trend. This divergence suggests potential price sensitivity among customers, which may be contributing to declining sales volumes despite higher prices.

The declining trend in transactions, despite the promotion of the all-day breakfast menu, indicates that while the promotion initially drove some sales growth, the overall effect may not have been sustained. The data suggests that price increases could have offset the benefits of the promotion, leading to a net decline in customer transactions over time.



Hypothesis Testing

Hypothesis Test 1: Impact of Promotion on Total Sales



```
T-statistic: 8.837915127434998, P-value: 9.803504550748318e-19
Result: Reject the null hypothesis (significant difference found).
Mean Pre-Promotion Sales: 1350.4698063279677
Mean Post-Promotion Sales: 1406.714610444344
Total units sold increased after the promotion.
```

Based on the provided output:

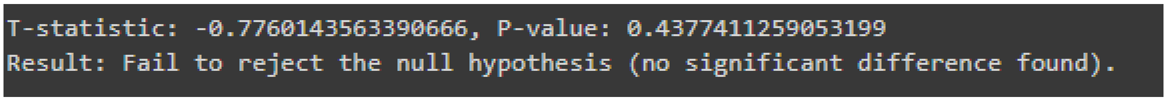
- **T-statistic**: 8.84
- **P-value**: 9.8e-19 (very small)

This result indicates that there is a statistically significant difference between pre-promotion and post-promotion sales, allowing us to **reject the null hypothesis**.

- **Mean Pre-Promotion Sales**: 1350.47 units
- **Mean Post-Promotion Sales**: 1406.71 units

The average sales **increased** after the promotion from **1350.47** to **1406.71** units, suggesting that the promotion had a positive impact on sales.

Hypothesis Test 2: Average Price per Unit Change



```
T-statistic: -0.7760143563390666, P-value: 0.4377411259053199
Result: Fail to reject the null hypothesis (no significant difference found).
```

Based on the provided output:

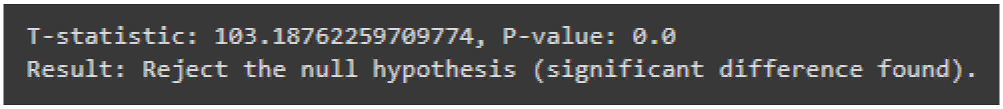
- **T-statistic**: -0.78
- **P-value**: 0.44

Summary:

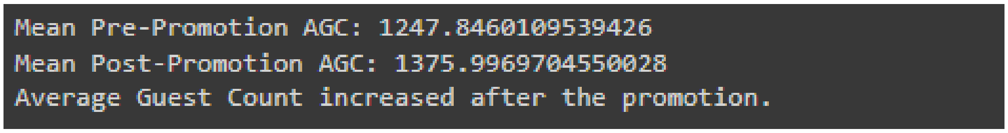
The result indicates that there is **no statistically significant difference** between pre-promotion and post-promotion average prices, as the **P-value (0.44)** is greater than the significance level of 0.05. Therefore, we **fail to reject the null hypothesis**.

This means that the average price (wavg_price) **did not change significantly** after the promotion. The promotion did not have a statistically significant effect on the weighted average price of the products.

Hypothesis Test 3: Effect of Promotion on Average Guest Count (AGC):



```
T-statistic: 103.18762259709774, P-value: 0.0
Result: Reject the null hypothesis (significant difference found).
```



```
Mean Pre-Promotion AGC: 1247.8460109539426
Mean Post-Promotion AGC: 1375.9969704550028
Average Guest Count increased after the promotion.
```

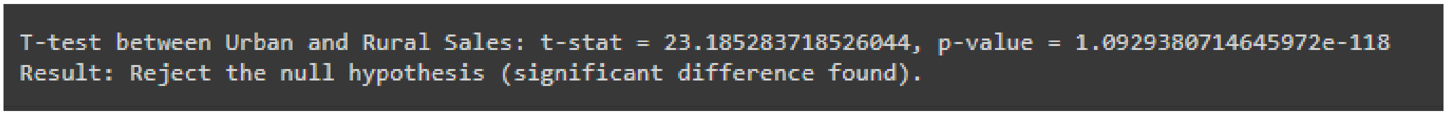
Based on the calculated means:

- Mean Pre-Promotion AGC: 1247.85
- Mean Post-Promotion AGC: 1376.00

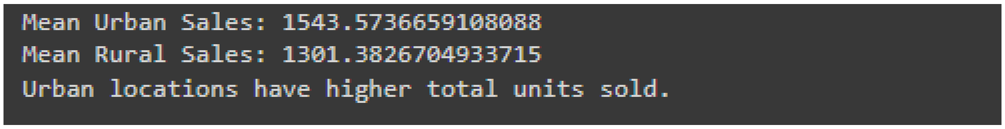
Summary:

The Average Guest Count (AGC) increased after the promotion, rising from 1247.85 to 1376.00. This suggests that the promotion had a positive impact, leading to a higher number of guests visiting after the promotion period. The increase was statistically significant, as indicated by the very low P-value (0.0).

Hypothesis Test 4: Effect on Urban Locations on Sales



```
T-test between Urban and Rural Sales: t-stat = 23.185283718526044, p-value = 1.0929380714645972e-118
Result: Reject the null hypothesis (significant difference found).
```



```
Mean Urban Sales: 1543.5736659108088
Mean Rural Sales: 1301.3826704933715
Urban locations have higher total units sold.
```

Based on the calculated means:

- Mean Urban Sales: 1543.57
- Mean Rural Sales: 1301.38

Summary:

The total units sold in urban locations is higher compared to rural locations, with an average of 1543.57 units in urban areas versus 1301.38 units in rural areas. This difference is also statistically significant, as indicated by the very low P-value (1.09e-118).

Therefore, we conclude that urban locations have significantly higher total sales compared to rural locations.

Overview and Insights

Hypothesis Insights:

- **Effectiveness of Promotion**: The promotion had a **positive impact** on both sales volume (increased total units sold) and customer footfall (increased Average Guest Count). This suggests that the promotion successfully attracted more customers and boosted overall sales, without significantly altering product pricing.
- **Urban vs. Rural Performance**: **Urban locations** outperformed rural locations in terms of total units sold, highlighting a significant difference in performance. Urban locations likely benefit from a larger customer base, higher demand, and possibly better marketing reach.

Overall Insights:

Based on the EDA, we identified the following insights and hypotheses:

- Promotional Impact: The promotion of the all-day breakfast menu on October 6th, 2015, appears to have had an initial positive impact on sales, particularly in regions where breakfast items were in high demand. However, the decline in transactions over time, as seen in the Tableau chart, suggests that the positive impact was not sustained. Regions with higher promotional activity initially showed better sales performance, but price increases may have counteracted these gains.
- Demographic Influence: Certain demographic factors, such as population density and average income, appear to influence sales performance. Regions with higher population density generally show better sales figures.
- All-Day Breakfast Impact: The introduction of all-day breakfast had a significant positive impact on sales initially, particularly in regions where customer demand for breakfast items was previously unmet. However, the subsequent decline in transactions suggests that the effect was short-lived, potentially due to increased prices or operational challenges.
- Price Sensitivity: The Tableau chart indicates that increasing prices may be leading to a decline in transactions. This suggests that customers could be sensitive to price changes, and maintaining competitive pricing may be essential to sustaining sales.
- Data Quality Concerns: Missing promotional data in some regions raises concerns about the completeness of the dataset. Additional data collection may be required to accurately assess the impact of promotions across all regions.

Data Sufficiency and Recommendations

- **Data Completeness**: While the datasets provide a good foundation for analysis, the missing promotional data limits our ability to draw definitive conclusions for all regions. Collecting complete promotional data would enhance the reliability of our insights.
- **Methodological Assumptions**: The analysis assumes that the provided data is representative of all regions. If this assumption does not hold, additional data collection may be necessary to account for regional differences.
- **Recommendations**: To confirm or reject our hypotheses, we recommend additional data collection focused on promotional activities and regional demographics. Furthermore, implementing a controlled experiment to test the impact of different promotional strategies, including the all-day breakfast offering, could provide valuable insights for optimizing marketing efforts. Additionally, maintaining competitive pricing should be prioritized to prevent loss of customer transactions due to price sensitivity.

Executive Summary

This report addresses the concerns raised by McDonald's Regional Sales Manager regarding fluctuating sales performance and the potential impact of data-driven decision-making. Through exploratory data analysis (EDA) of sales, regional demographics, and promotional data, we aimed to understand the underlying factors influencing sales and to provide actionable insights.

The analysis focused primarily on the impact of McDonald's all-day breakfast promotion launched on October 6th, 2015. While the promotion initially led to a sales boost and increased customer footfall, a decline in transactions over time suggests that the positive impact was not sustained. This trend may be attributable to increased pricing, which could have offset the promotional gains, as suggested by both the data analysis and external articles from Eater and Forbes.

Key insights indicate that promotional activity initially drove increased sales, especially in high-demand regions, but sustaining this growth required a balance between operational efficiency and competitive pricing. The data also showed that demographic factors like population density and regional income levels significantly influenced sales performance, with urban locations outperforming rural ones.

The EDA revealed several data quality concerns, including missing promotional data in certain regions, limiting the ability to draw definitive conclusions across all markets. Addressing these data gaps is essential for enhancing the reliability of future analyses.

Recommendations include collecting complete promotional data, conducting controlled experiments to assess different promotional strategies, and maintaining competitive pricing to avoid deterring customers. By addressing these areas, McDonald's can develop more effective data-driven strategies to boost sales performance and better allocate resources across regions.