Institute of Technology of Cambodia

ASSIGNMENT OF STATISTIC

Topic : Customer Behavior

GROUP: I3-AMS-A
Team-1

Name of Students	ID of Students	Score
HENG Seaklong	e20210329	.....
Bun Ratanatepy	e20210320	.....
PEANG Ratatanak	e20210072	.....
CHHIN Visal	e20210742	.....
KHON Khengmeng	e20210176	.....

Lecturer: Mr.Touch Sopheak (TD)
Mr. Phok Ponna (Course)

Department of Applied Mathematics and Statistics

Contents

Abstract

This abstract highlights key findings related to customer behavior analysis, specifically focusing on time spent per week on coffee-related activities and preferred payment methods. The point estimates derived from our analysis serve as pivotal reference points, offering insights into the average time commitment to coffee and prevalent payment behaviors within our customer base. Understanding these patterns enables businesses to tailor strategies, enhancing the overall customer experience and optimizing operational efficiency. These insights are crucial for informed decision-making, ensuring that marketing efforts and payment processes align with the typical behaviors of our diverse customer demographic.

1.Introduction

The purpose of this project is to analyze customer behavior in order to gain valuable insights that can inform business strategies and decision-making. By understanding customer preferences, purchasing patterns, and engagement levels, we aim to enhance the overall customer experience and optimize business performance.

2.Data Collection

Data for this analysis was gathered through a variety of channels, including transaction records, customer surveys, website analytics, and social media interactions. The dataset is comprehensive, covering a diverse range of variables, such as customer demographics, purchase history, and online behavior. Notably, a significant portion of the dataset originated from a survey conducted within our community, providing a substantial source of information.

```
df.head(5)
```

[6]

	Age	Gender	Take_coffe	Type_coffe	Times_per_Week	Total_Payment	choice_shop	environmental	Payment
0	20	Men	With sugar/sweetener	Latte	3	6.75	Variety of menu options	Very important	Cash
1	19	Women	With sugar/sweetener	Latte	2	5.00	Quality of coffee	Moderately important	Cash
2	19	Women	With both cream/milk and sugar/sweetener	Cappuccino	4	7.00	Atmosphere/environment	Very important	Mobile payment apps
3	18	Women	With sugar/sweetener	Espresso	1	2.80	Quality of coffee	Moderately important	Mobile payment apps
4	19	Men	With cream/milk	Latte	5	10.00	Proximity to home/work	Slightly important	Cash

3.Data Visualization

To provide a comprehensive overview of customer behavior, we employed various data visualization techniques. Key visualizations include:

- **Customer Segmentation:** Utilizing clustering algorithms, we identified distinct customer segments based on demographic and behavioral characteristics.
- **Age vs Type of coffee:** We use a bar chart that shows the distribution of different coffee types among different age groups. This visualization would help identify whether certain age groups prefer specific types of coffee such as Americano, Cappuccino, Espresso, and Latte.
- **Age vs Time spent per week on coffee:** A box plot was utilized to compare the time spent on coffee per week across different age groups. Each box in the plot represents an age category, with the length of the box indicating the interquartile range and the line inside denoting the median.

- **Age vs Payment method:** A bar chart was applied to illustrate the distribution of payment methods for coffee purchases. The visualization might uncover whether there are generational preferences for payment methods, such as credit card, mobile payment, or cash.

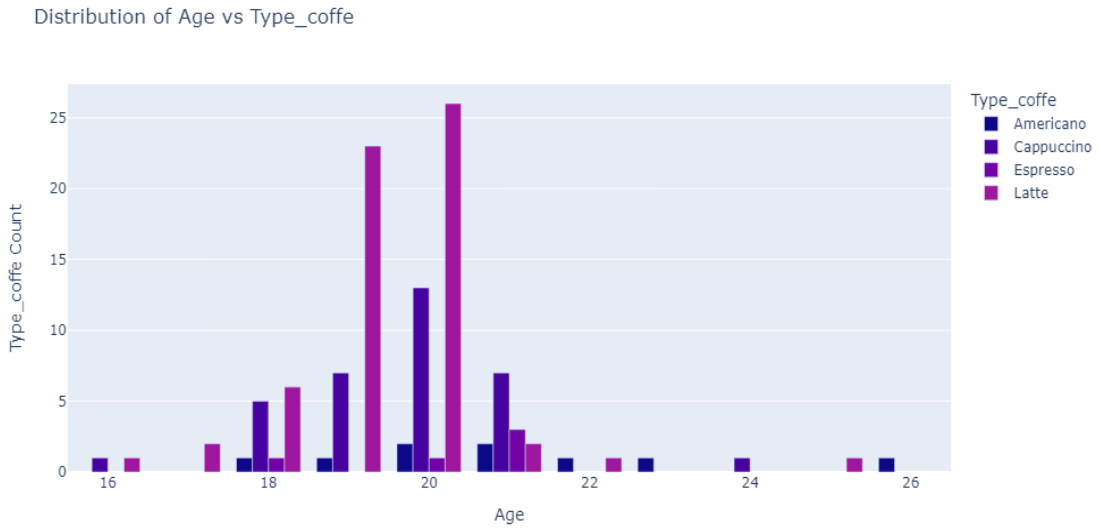


Figure 1.1: Age Vs Type Coffee

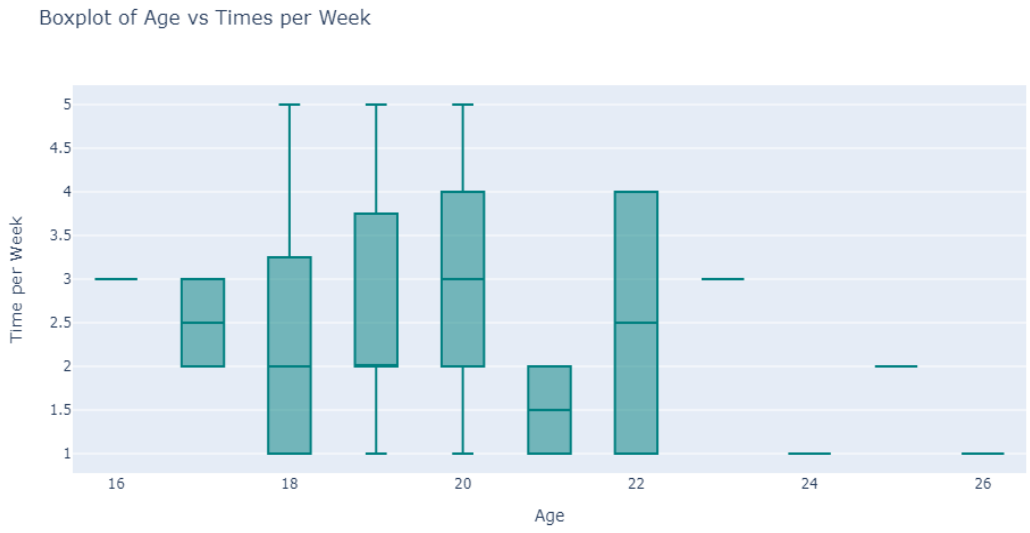


Figure 1.2: Age Vs Time per Week

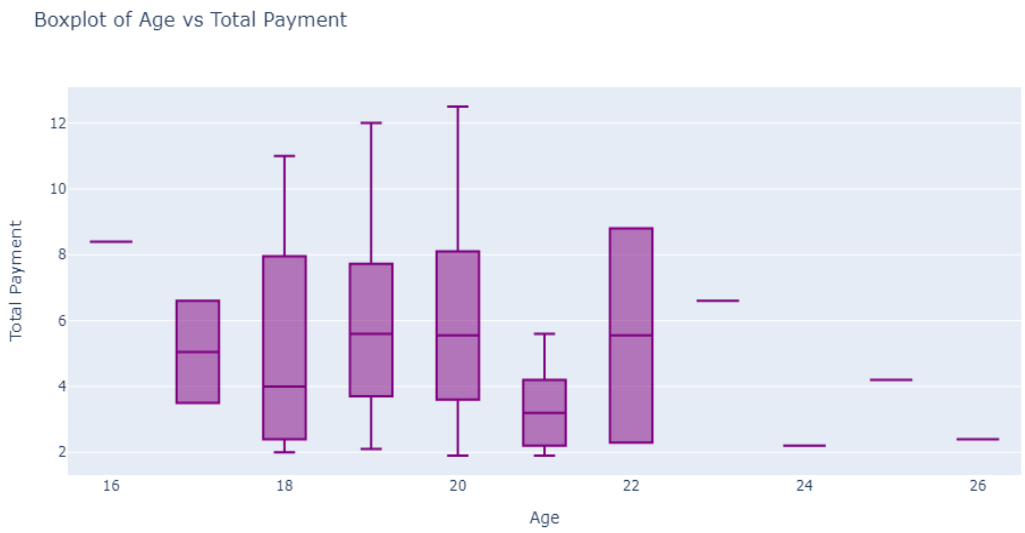


Figure 1.3: Age Vs Total Payment

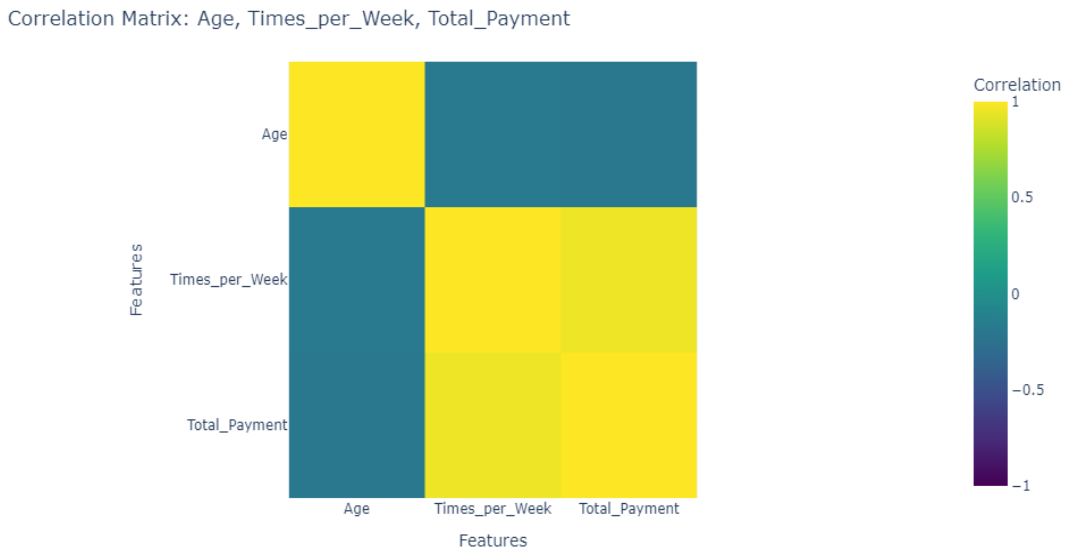
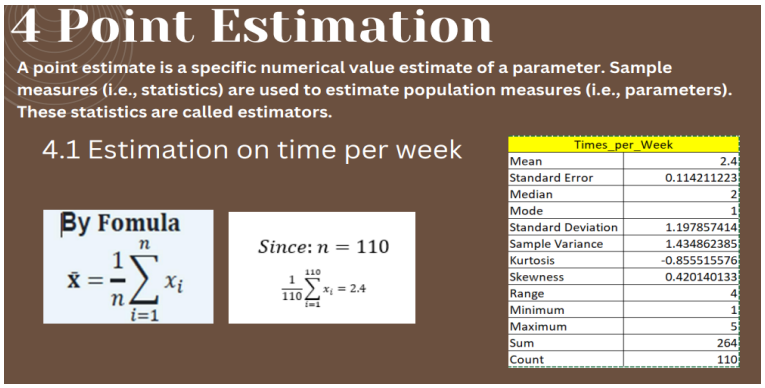


Figure 1.4: Correlation Each Variable

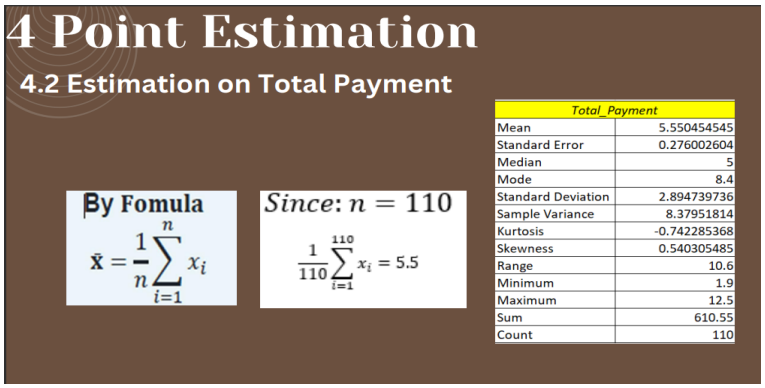
4.Point Estimate

The point estimate derived from our analysis serves as a central reference point, indicating the average time spent per week on coffee-related activities and the predominant payment method within our customer base. This single value allows us to gauge the typical behavior of customers in terms of their time commitment to coffee and their preferred mode of payment, providing valuable insights for strategic decision-making For example, the average weekly spending per customer is estimated at $5.55, providing a baseline for assessing changes in spending patterns over time.



Times_per_Week	
Mean	2.4
Standard Error	0.114211223
Median	2
Mode	1
Standard Deviation	1.197857414
Sample Variance	1.434862385
Kurtosis	-0.855515576
Skewness	0.420140133
Range	4
Minimum	1
Maximum	5
Sum	264
Count	110

Figure 1.3: Point Estimate on Time per week



Total_Payment	
Mean	5.550454545
Standard Error	0.276002604
Median	5
Mode	8.4
Standard Deviation	2.894739736
Sample Variance	8.37951814
Kurtosis	-0.742285368
Skewness	0.540305485
Range	10.6
Minimum	1.9
Maximum	12.5
Sum	610.55
Count	110

Figure 1.4: Point Estimate on Total Payment

5.Confidence Interval

To quantify the uncertainty associated with our point estimates, we calculated confidence intervals. These intervals provide a range within which we can be reasonably confident that the true population parameter lies. For instance, the 95% confidence interval for the average customer spending is [$5.00, $6.09], signifying that we are 95% confident that the true average spending falls within this range.

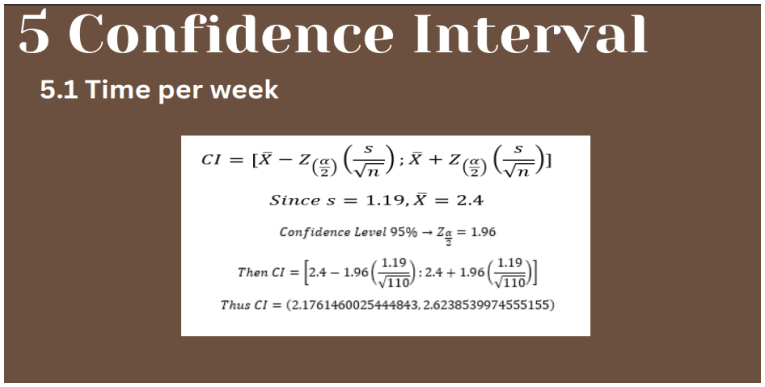


Figure 1.3: Confident interval of Time per week

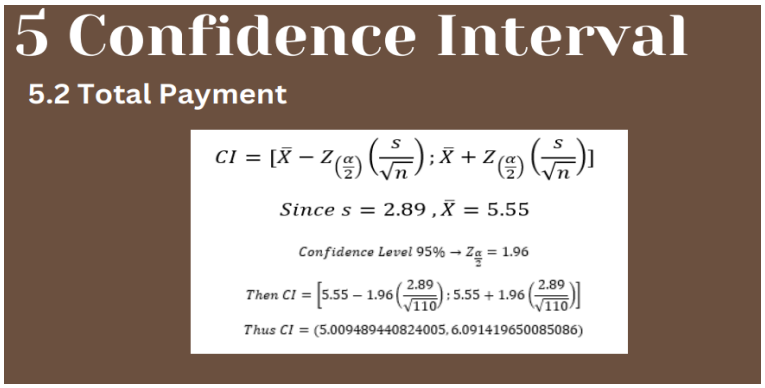


Figure 1.4: Confident interval Total payment

6.Test of statistics Hypothesis

This comprehensive analysis aims to perform statistical hypothesis tests on three distinct datasets: 'Age', 'Total Payment', and 'Times per Week'. Each test evaluates whether the respective dataset's average significantly differs from specified values.Each dataset was subjected to a Z-test following these steps:

A. **Data Selection:** Relevant data for each test was extracted from `csv('data_v5cvs.csv')` file.

B. **Hypothesis Formulation:**

- **Null Hypothesis (H_0):** The mean of the dataset equals a specified value.
- **Alternative Hypothesis (H_A):** The mean of the dataset does not equal the specified value.
- **Significance Level (α):** 0.05 for all tests.

C. **Statistical Calculations:**

- Sample Mean and Standard Deviation: Computed for each dataset.
- Z-Statistic: Calculated for each test as $\frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$.
- P-value: Derived for two-sided tests.

For the calculation for each test we have:

- **Test of Statistics Hypothesis for Age:**
  - Z-statistic: -2.2979988887342895
  - P-value: 0.021561852375283852
  - There is evidence to support the claim (reject the null hypothesis).
- **Test of Statistics Hypothesis for Total Payment:**
  - Z-statistic: -1.6287725092930991
  - P-value: 0.10336118636854796
  - There is not enough evidence to support the claim (fail to reject the null hypothesis).
- **Test of Statistics Hypothesis for Time Per week:**
  - Z-statistic: 3.5022827776653687
  - P-value: 0.0004612897547240369
  - There is evidence to support the claim (reject the null hypothesis).

7.Inferences Based on two Sample

Introduction: In the evolving landscape of data science, the ability to make informed decisions based on data is crucial. Central to this process is the concept of inference using two variables. This report explores the multifaceted role of two-variable inference in data science, highlighting its relevance and application across various domains.

Data Analysis

Subsets Creation: The data was segmented based on the 'New Payment' categories to focus on each payment group separately and we defined to categorize payment methods into '0' for traditional methods (Cash or Credit/Debit Card) and '1' for other methods. This categorization was applied to the 'Payment' column of the Data Frame, resulting in a new column 'New Payment' for streamlined analysis.
Descriptive Statistics: Summary statistics were generated for these subsets to understand the central tendencies and variabilities within each payment group.

Statistical Testing

T-Test Application: Independent sample t-tests were performed to compare the means of 'Age', 'Times per Week', and 'Total Payment' between the two payment groups.
Results Interpretation: The Z-statistic and p-value were calculated for each test, with an alpha level () of 0.05, to ascertain the statistical significance of the differences. Results Interpretation: The Z-statistic $Z = \dfrac{\bar{x} - \bar{y} - (u_1 - u_2)}{\sqrt{\frac{S_1^2}{m}} - \sqrt{\frac{S_2^2}{n}}}$ and p-value were calculated for each test, with an alpha level (α) of 0.05, to ascertain the statistical significance of the differences.

Results and Visualization

The t-tests provided a quantitative measure of the differences in age, visit frequency, and total payment between the two payment categories. The p-values and Z-statistics indicated whether these differences were statistically significant.

Data Visualization

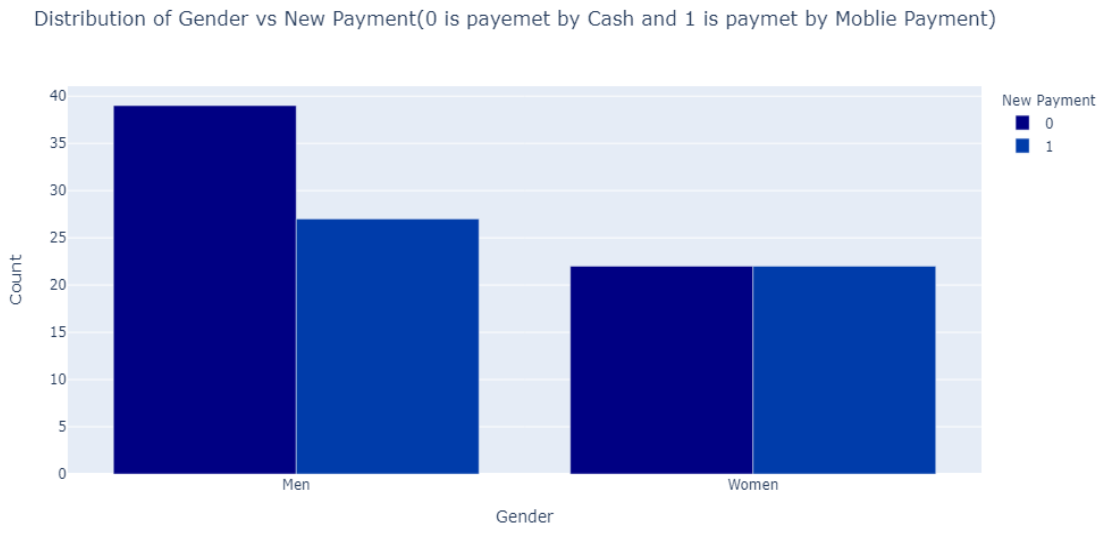


Figure 1.5:Gender Between Pay by cash and Pay by mobile app

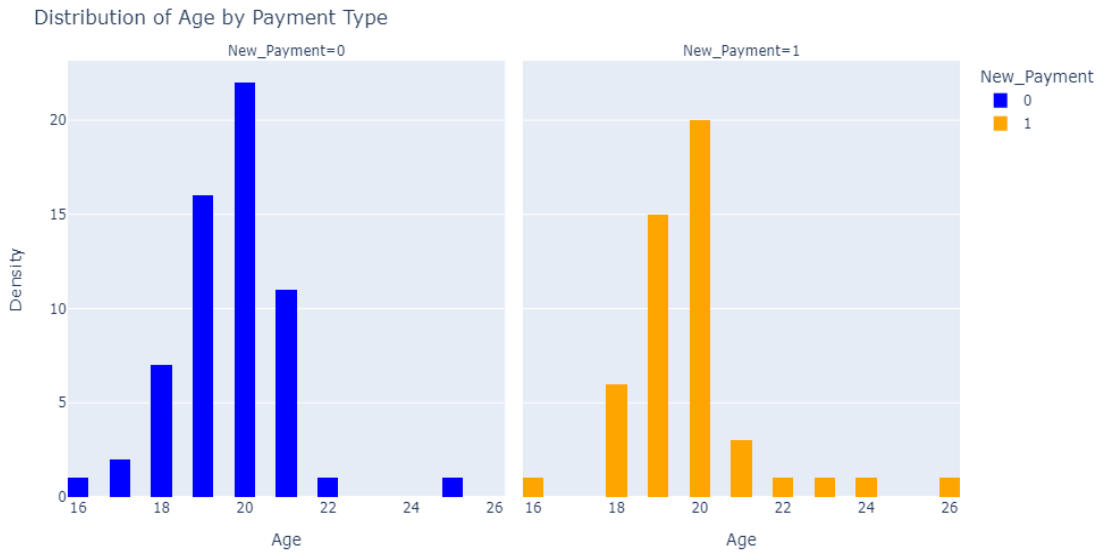


Figure 1.6:Age Between Pay by cash and Pay by mobile app

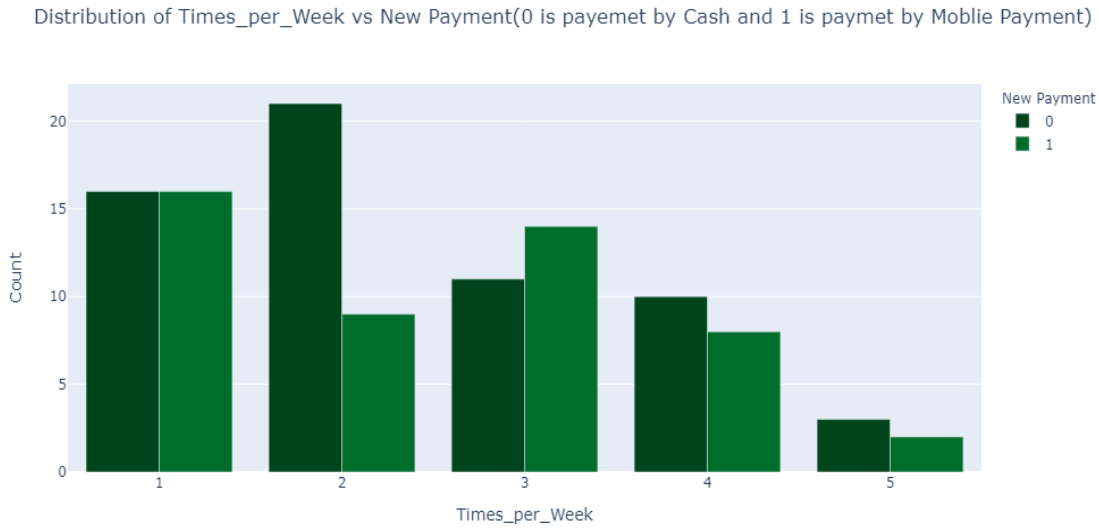


Figure 1.5:Time per week Between Pay by cash and Pay by mobile app

8.Conclusion

In conclusion, this project has offered valuable insights into customer behavior, enabling a more informed approach to business strategies. The combination of data collection, visualization, point estimates, and confidence intervals equips stakeholders with a robust understanding of customer dynamics, facilitating data-driven decision-making and continuous improvement in customer satisfaction and business performance.