Diwali Delight: Unveiling Insights from the Festive Sales

Andrea Labra Orozco, Vamshi Reddy Madem, Siddhartha Vaddempudi

California State University East Bay

STATS 630 - Statistical Methods

Wendy Rummerfield

December 12, 2023

Abstract

Diwali signifies the victory of light over darkness and evil..It is celebrated around mid-September and mid-November and it lasts for around five or six days. According to the Confederation of All India Traders Diwali-related sales added up to an estimated $ **3.75 trillion** Rupees in 2023 , which has a big impact on the Indian Economy . Is there an association between age and amount spent by customers for the Diwali celebration. After conducting statistical methods and hypothesis testing using Linear regression we came to the conclusion that we do not have enough evidence that age and amount spent in Diwali is linearly associated.

Keywords: Diwali sales, Linear Regression, Age

Background/Introduction:

Diwali is a winter festival celebrated by millions of people worldwide. Diwali, one of the most important Indian holidays, represents a new beginning for every family, business, and society, which makes it an excellent time to explore internet business ideas. The main reason Diwali is special is because people like celebrating and exchanging gifts. Businesses give their employees Diwali bonuses as a token of appreciation. So diwali sales include sale of gold, silver, cars, bikes, clothes, and much more. It is a trillion rupees market thriving in India. Different age groups of people do shopping during these festivals. So, a study was conducted to see if there is a relationship between age and amount spent by customers during diwali sale.

So, our null hypothesis is, there is no association between age and the amount spent by customers.and the alternative hypothesis is, there is a positive association between age and the amount spent by customers.

Methods:

a. **Source of the Data :**

In this case, The data was collected from a retail store in India on the account of the festival diwali for the date of 11 November 2023. This dataset was collected from https://www.kaggle.com/datasets/saadharoon27/diwali-sales-dataset Website of Week 46 2023 from TidyTuesday. The data was collected by author SAAD HAROON who is a student at Queen's University Belfast, Northern Ireland, United Kingdom from a retail store in India.

Variables in the dataset:

Variables	Attributes
User_Id	Unique numeric ID of customer
Cust_name	Customer Name
Gender	Gender of the Customer (Male OR Female)
Age_group	Age group to which customer belongs (ex : 0- 17, 18-25, 26-35)
Age	Age of the Customer(in Years)
Marital_status	Married or Unmarried(0 or 1)
State	State to which the customer belongs (ex : Andhra, Maharashtra)
Zone	Zone to which the customer belongs (ex :Southern or Northern)
Occupation	Occupation of the customer
Orders	No of orders by the customer
Amount	Amount spent by the customer for buying diwali items

This dataset seems to cover various aspects of sales transactions, including invoice details, customer demographics, product information, pricing, and ratings.

b. **Statistical Methods**

The sampling method or technique was not mentioned. But according to our understanding of the dataset, the sampling method might be Stratified sampling as it involves dividing the population into subgroups or strata based on certain characteristics and then sampling from each stratum. In this case the data is divided first into zones such as Central, Northern, Southern, Eastern and Western and further the respective states of the zones are mentioned.

Results:

For descriptive analyses, we have to start by looking at the variables that we are going to be taking into account this data set, which are going to be age and amount of money spent. Our data had 1273 observations overall. The following is the summary statistics of the amount spent(in rupees), this analysis did contain 3 N/As

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
569	5397	8100	9467	12783	23952

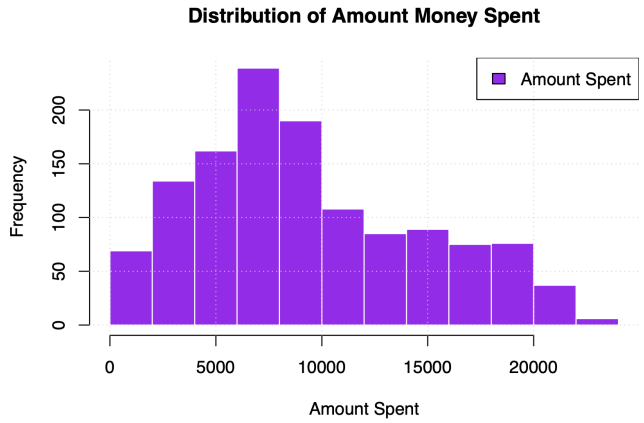


Figure 1

In Figure 1 we can see that the distribution of the amount spent had a slightly right skewed distribution, with the median at 8100 rupees and the max at around 23,952 rupees. Now let's look at the summary statistics of age from our data set.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
12	26	33	35.88	44	91

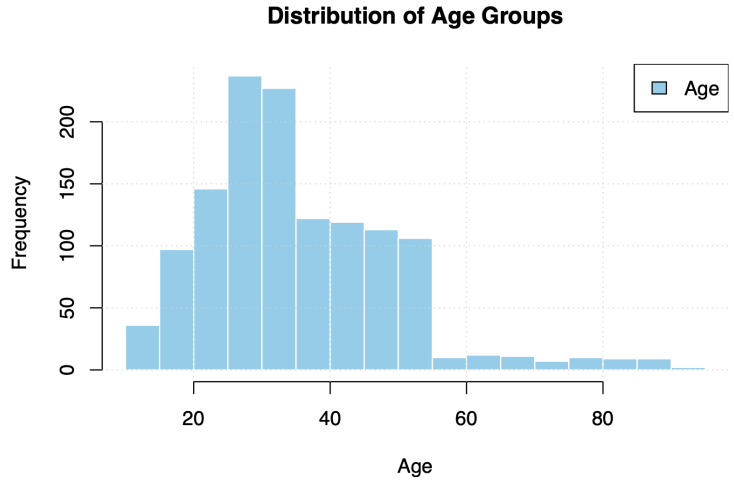


Figure 2

In Figure 2 we can see that the youngest sample is 12 years old and the oldest is 91 years old, with the median at 33 years old. Additionally age has a slightly right skewed distribution. With this in mind, our group decided that the best way to see if there is an association between the age and the amount spent by customers in the diwali celebration is to do a linear regression. In which have assigned HO: $\beta = 0$ - meaning there is no association between age and the amount spent by customers. Making HA: $\beta > 0$ - meaning there is a positive association between age and the amount spent by customers.

Since we are doing a linear regression test, the data must pass a check for certain conditions. Firstly, the data meaning each observation must be independent from each other - we can check for this using the 10% rule - 1,273 is less than 10% of all of the diwali customers. Next we will check plots for normality, linearity and residuals.

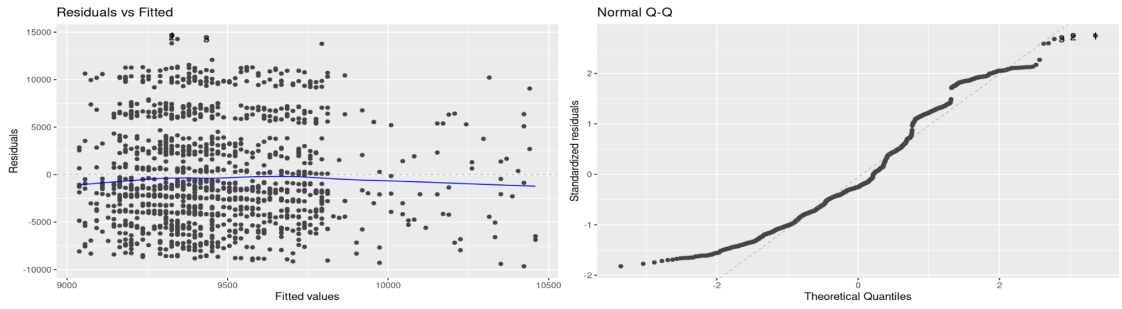


Figure 3

The residuals vs fitted plot should ideally show a random scatter of points with no clear pattern. The normal QQ plot is checking for the normality of the residuals and should ideally fall along the line.

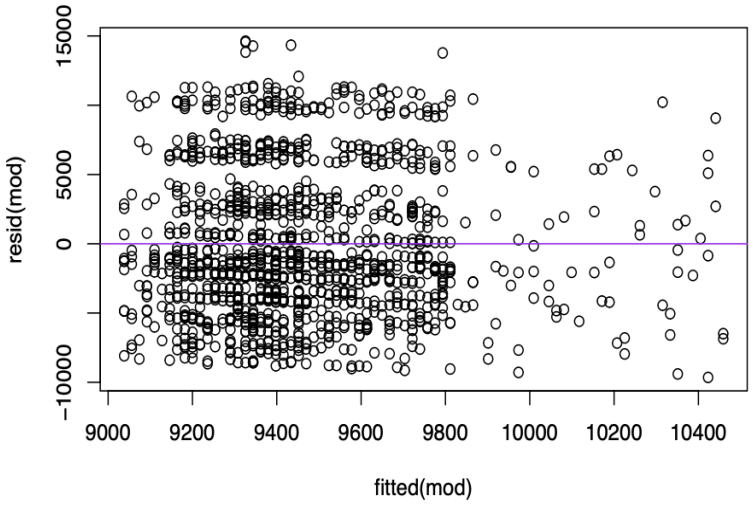


Figure 4

This figure should be points randomly distributed around the zero line at zero. It indicates that the variance of the residuals is roughly consistent across different levels of fitted values. Note about the conditions, conditions may have not been satisfied, we removed duplicate rows, looked at only those who made one order, and removed rows with duplicated IDs. It seems that there is something wrong with the data set that we chose, when it gets imported. We received approval to move on with this data set for this case. So we will move forward with the testing.

After performing the testing we got a result of 17.99(.21 cents) rupee increase, meaning as a person increases one year of age, the amount spent on diwali increases by 17.99 rupees. We therefore obtain a p value of 0.0934. Which leads us to the conclusion to fail to reject the null hypothesis HO, we do not have enough statistical evidence that age and the amount of money spent in diwali is linearly associated.

Discussion :

a. **Limitation :**

The Retail store size is one of the limitations, because if we are looking at a bigger store like Costco or Pitco their sales record will be higher when compared to a local retail store. Another factor that affects the most is the day on which the dataset was recorded. If the sales record was for the day before diwali then the sales record will be higher when compared to the 5-6 days before diwali.

b. Future Works :

Will try to implement new confounding variables like Region and Occupation so that we can target specific regions or audiences.

c. Conclusion

Analyzing the correlation between customer age and festival spending informs targeted marketing, optimizes product offerings, and enhances strategic planning for long-term business success. Analyzing the correlation between customer age and festival spending informs targeted marketing, optimizes product offerings, and enhances strategic planning for long-term business success.and enhances strategic planning for long-term business success.

References:

Haroon, S. (2023, November 14). *Diwali sales data*. GitHub. https://github.com/rfordatascience/tidytuesday/blob/master/data/2023/2023-11-14/readme.md

Shivaji Sarkar (2023, November 26). *The Sunday Guardian*

https://sundayguardianlive.com/business/diwali-sales-soared-to-rs-3-75-trillion-in-spite-of-economic-struggles

Appendix :

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
569	5397	8100	9467	12783	23952

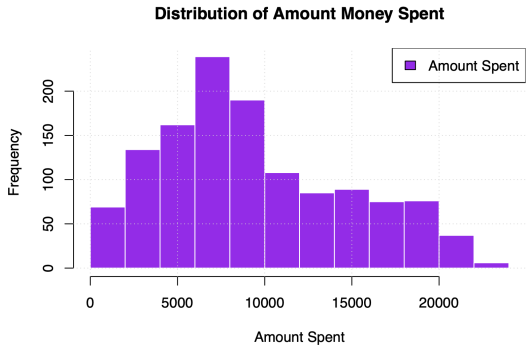
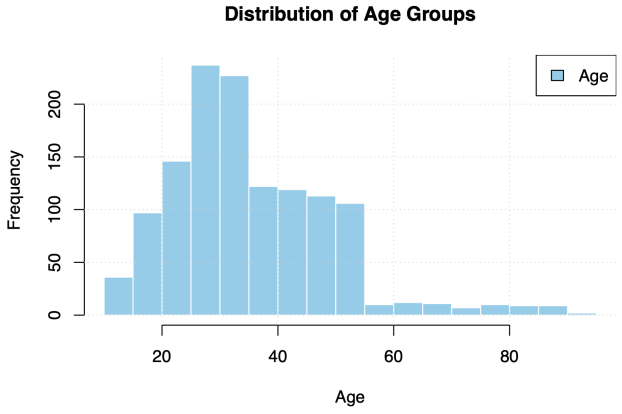


Figure 1

Minimum	1st Quartile	Median	Mean	3rd Quartile	Max
12	26	33	35.88	44	91



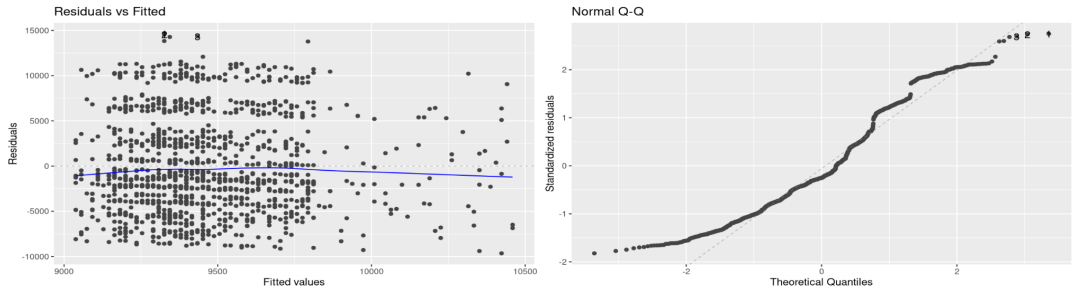


Figure 3

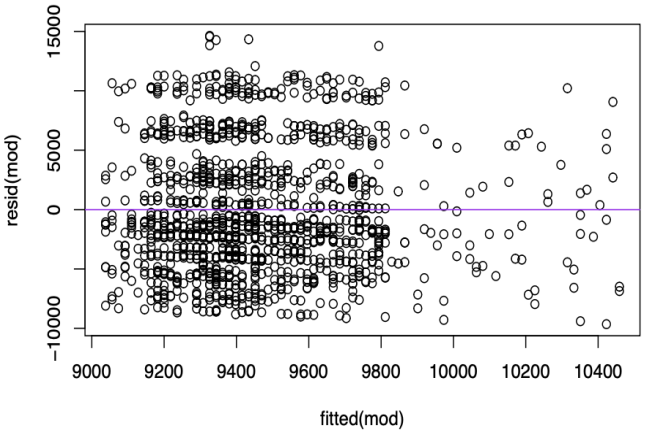


Figure 4

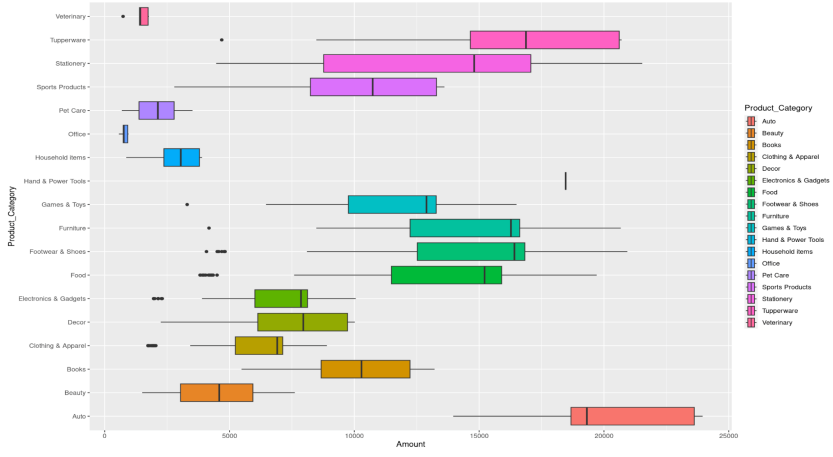


Figure 5

This figure shows the product category and the amount spent, we can see that the highest amount is between auto, stationary and tupperware.

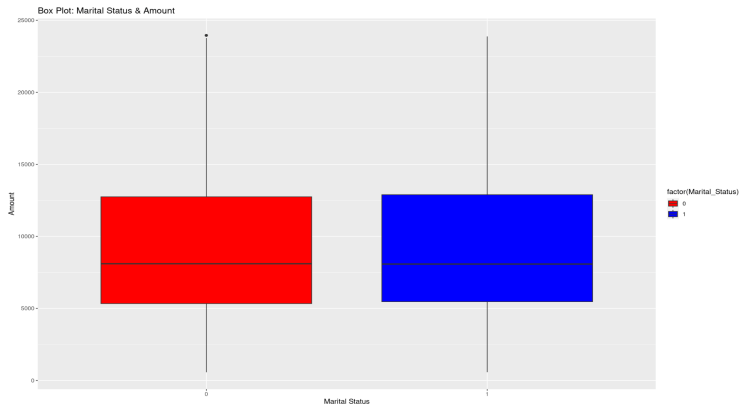


Figure 6

This figure shows a boxplot of married and unmarried individuals sampled and we can see that the plots are almost identical, showing the median at around the same mark of 7500 rupees.

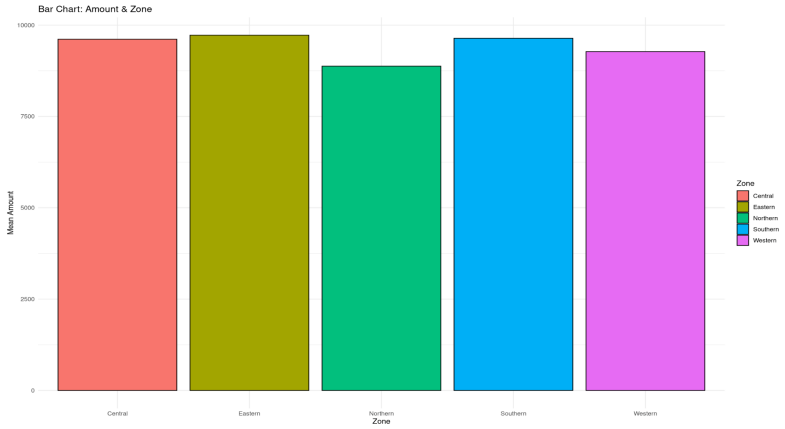


Figure 7

This figure shows the amount of money spent by each zone. We can see that the Eastern zone spent the most money and the western zone spent the least amount of money.

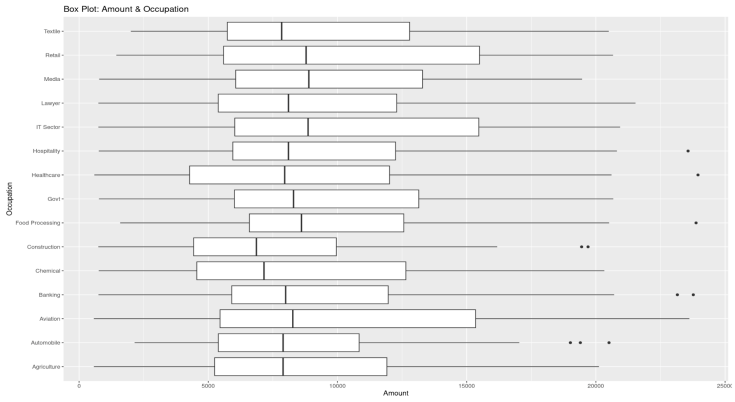


Figure 8

This figure shows a boxplot of all of the occupations sampled and the amount of money spent. We can see that aviation, IT sector, lawyers and healthcare tend to spend the most money.