BLACK FRIDAY DATA ANALYSIS

Applied Statistics – STAT 614

December 18, 2023

Done by:

Hasan Mehdi

Basanth Varaganti

Introduction

Black Friday is an international holiday celebrating the opportunity to purchase items at reduced prices. Doesn't sound too exciting on paper, but the data surrounding this yearly rush for discounted toasters is quite astounding. According to a yearly report of sales by Adobe Analytics, the total revenue made on over 1 trillion online transactions grossed over $9.8 billion in the U.S. on November 24th, 2023. This was a near 400% increase in earnings when compared to a regular day, which ranges from a measly $2 to $3 billion in revenue.

E-commerce is rapidly becoming *the* way to shop, rising from 13.8% of total retail sales worldwide in 2019 to 19.5% in 2023 and projected to increase by about 1% per year, according to a figure posted on Statista. With this method of shopping, the collection of valuable demographic data from customers is incredibly easy, possible in a couple clicks when a new user creates their account.

Hand-in-hand with the rise of e-commerce comes the integration of data modeling, artificial intelligence, and machine learning. Retailers worldwide have shifted to consumer-targeted advertising, and the need for customer data has become a priority. This data can be leveraged to specifically target items and coupons towards individuals who frequent the store more often. Alternatively, it encourages strategies that are geared towards retaining demographics who typically spend less time at the store.

In the midst of this evolving retail environment, ABC Private Limited, a retail company, is seeking insights into customer purchasing behavior (specifically, purchase amount). They've provided a purchase summary of selected high-volume products from the previous month. The dataset includes information on customer demographics such as age, gender, marital status, city type, and duration of stay in the current city. Additionally, it contains details about the products, including product_id, product category, and the total purchase amount from the last month.

In response to this data, ABC Private Limited aims to develop a predictive model to estimate the purchase amount made by customers for different products. The goal is to use this model to create personalized offers tailored to individual customers based on their preferences and purchasing patterns.

To help with this model, four statistical analyses have been performed on the various information provided in the purchase summary. With these preliminary analyses, broad correlations between basic customer demographics and purchase behaviors can be drawn. This will serve as a foundational understanding and proof of concept for more complex data modelling endeavors in the future.

Methods

Data Retrieval

The data for ABC Private Limited was collected from Kaggle, a platform owned by Google and an online community of data scientists. Kaggle enables users to find and publish datasets to solve data science challenges.

Data Pre-Processing

The dataset consisted of an extensive Excel file of over 550,000 rows of customer data. Utilizing a program created in Python and the Python package Pandas, a controlled random sample of 1000 records was generated. The sample was kept intentionally balanced, featuring 500 records for both male and female customers. This approach was done to maintain representativeness and to avoid potential biases in downstream analyses.

Data Analysis

The processed dataset was then migrated to JMP Pro, a statistical analysis software. A wide number of tests, including a One-Sample Hypothesis T-Test, a Two-Sample Hypothesis T-Test, a One-Sample ANOVA Test, and a Chi-Squared Test of Independence were performed on the data. These statistical tests were selected based on the types of our data (almost entirely categorical) and to derive relevant conclusions.

Assumptions

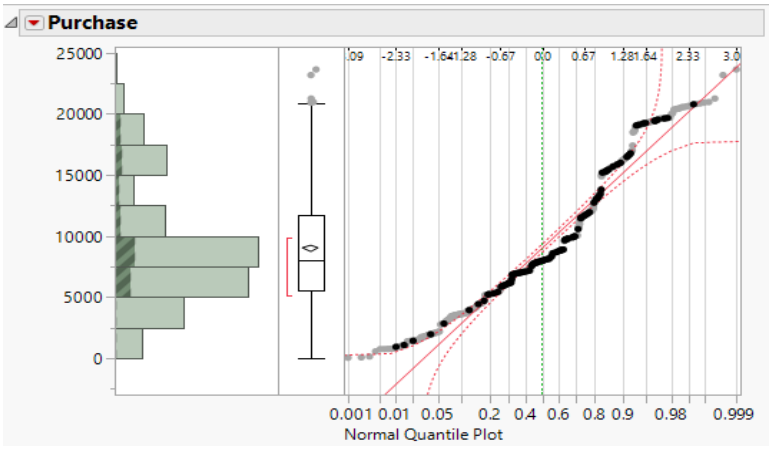

Figure 1: Normal Quantile Plot of Purchase Data

After creating a normal quantile plot of the data, we can see that the data for purchase amount is not normally distributed. However, this is expected and fine for a couple reasons:

- Customers spending is inherently skewed because customers tend to spend within a specific range
- Customer demographics are diverse which leads to variability in spending habits

It is mentioned in each question, but a level of significance $\alpha = 0.05$ has been chosen because of common practice.

Results

Analysis 1

The sales department of ABC Private Limited wants to know if there is evidence to claim that the average customer at their store is spending more money than the average American is spending on Black Friday in 2023. According to a survey conducted by Deloitte, the average amount an American planned to spend on Black Friday 2023 was $138 (equivalent to 11,502.43 Indian Rupees). Given is the level of significance α = 0.05.

$$H_0: \mu = 11{,}502.43$$

$$H_1: \mu > 11{,}502.43$$

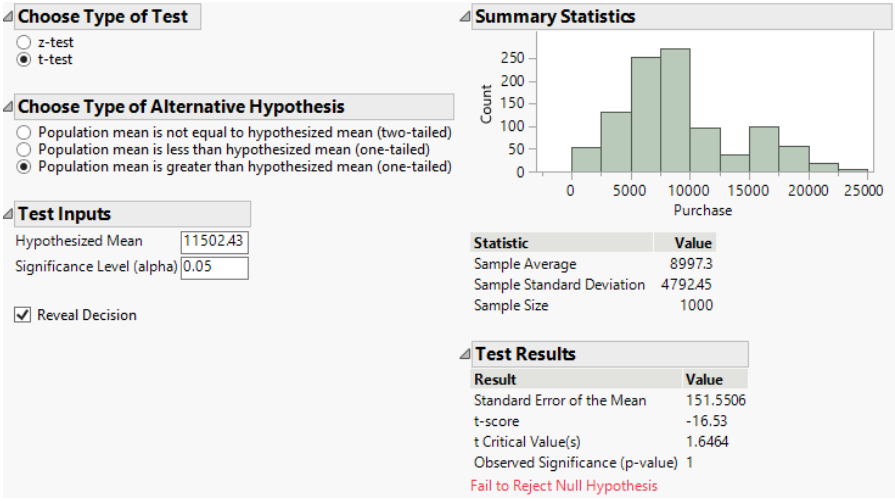

Figure 2: Hypothesis t-Test for One-Sample

$$p - value > \alpha$$

$$1 > 0.05$$

Decision: Fail to Reject Null Hypothesis

Conclusion: With 5% level of significance, we have insufficient evidence to claim that the average a customer spent at ABC Private Limited is more than the average American planned to spend on Black Friday.

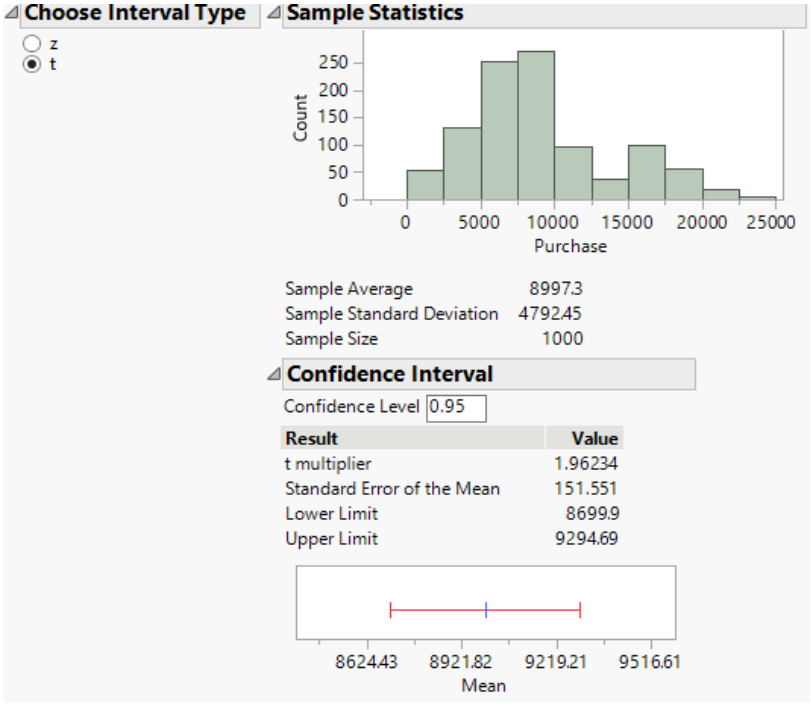

Figure 3: Confidence Interval for True Mean

We can with 95% confidence claim that the average purchase amount per customer is between 8699.9 and 9249.69 Indian Rupees. Even the upper confidence bound is much less than the hypothesized mean.

Analysis 2:

The marketing department over at ABC Private Limited is misogynistic. They thought that women would shop more out of pure instinct which led to the department not spending enough money advertising to that demographic. A quick glance at the spreadsheets exposed that perhaps men were the super-consumers. The CEO himself reached out to the statistics department to find out the real answers. Is there evidence that the amount that men spent, on average, is *greater* than the amount that women spent on Black Friday? Level of significance $\alpha = 0.05$.

$$H_0: \mu_{men} = \mu_{women}$$

$$H_1: \mu_{men} > \mu_{women}$$

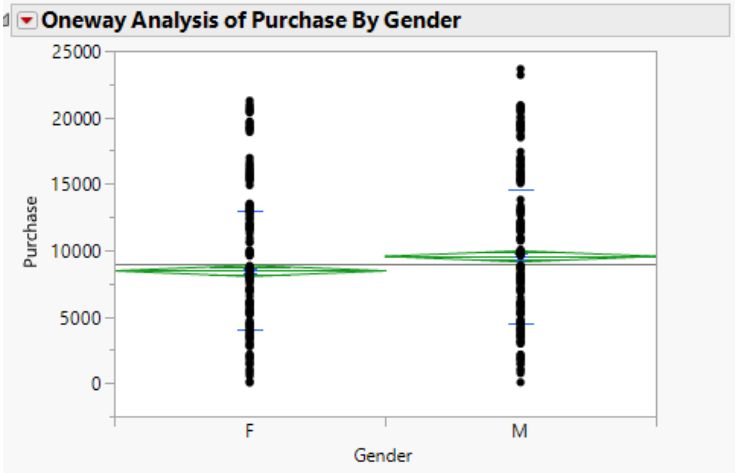

Figure 5: One-Way Analysis of Purchase vs Gender

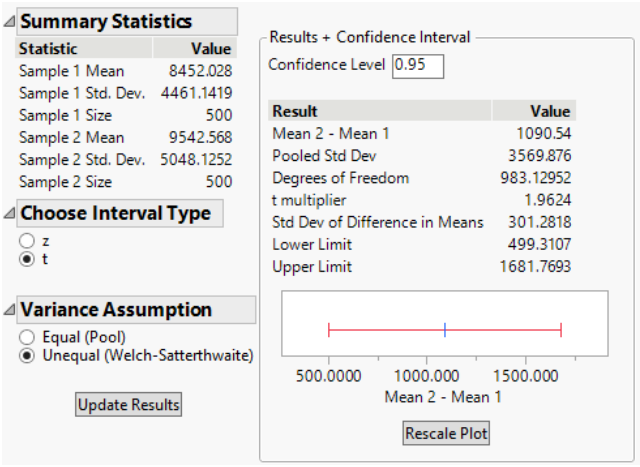

Figure 4: Confidence Interval for Difference

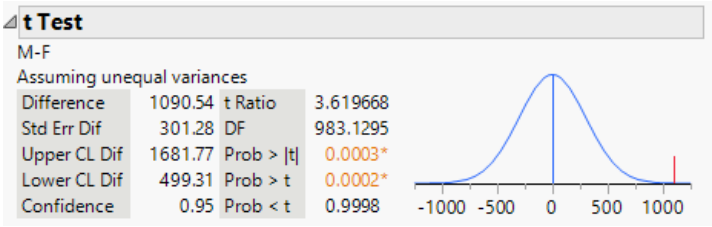

Figure 6: Hypothesis t-Test for Two Samples

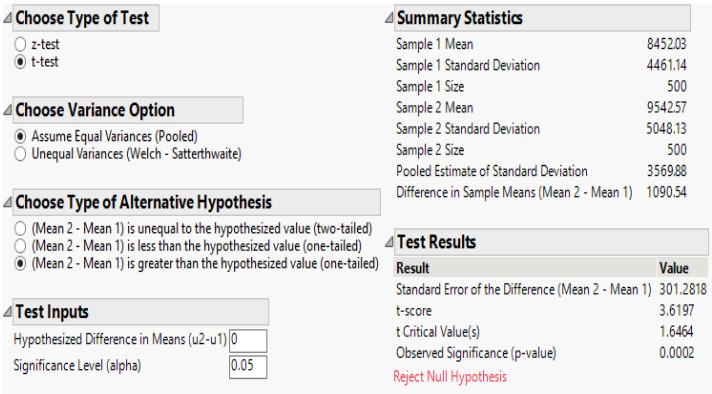

Figure 7: Hypothesis t-Test for Two Samples (pt.2)

$$p - value < \alpha$$

$$0.0002 < 0.05$$

Decision: Reject Null Hypothesis

Conclusion: With 5% level of significance, we have sufficient evidence to claim that the average spent by men is greater than the average spent by women.

Since the 95% confidence interval does not include 0, we arrive at the same conclusion as the hypothesis test, that we cannot claim that the average amount spent by men is the same as the average amount spent by women.

Analysis 3

The sales department at ABC Private Limited is furious upon discovering that profits are low for this year's Black Friday. They scramble together an ANOVA department (which is separate from the statistics department, for some reason) to find out if there are any significant variations in purchase amounts among different age groups. They rush to find the suitable level of significance α = 0.05.

$$H_0: \mu_{0-17} = \mu_{18-25} = \mu_{26-35} = \mu_{36-45} = \mu_{46-50} = \mu_{51-55} = \mu_{55+}$$

$$H_1 = at\ least\ two\ means\ differ$$

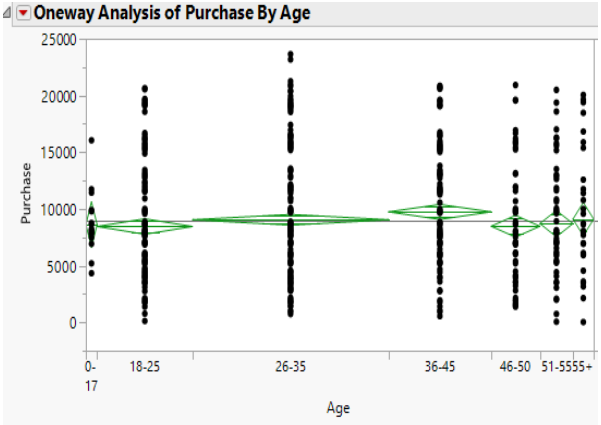

Figure 10: One-Way Analysis of Purchase vs Age

Oneway Anova

Summary of Fit

Rsquare	0.008899
Adj Rsquare	0.00291
Root Mean Square Error	4785.472
Mean of Response	8997.298
Observations (or Sum Wgts)	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Age	6	204177596	34029599	1.4860	0.1796
Error	993	2.274e+10	22900739		
C. Total	999	2.2945e+10			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
0-17	22	8625.27	1020.3	6623.1	10627
18-25	188	8457.35	349.0	7772.5	9142
26-35	386	9052.61	243.6	8574.6	9531
36-45	202	9747.54	336.7	9086.8	10408
46-50	96	8480.03	488.4	7521.6	9438
51-55	65	8714.74	593.6	7550.0	9880
55+	41	9114.85	747.4	7648.3	10581

Std Error uses a pooled estimate of error variance

Figure 9: One-Way ANOVA Test

$$p - value > \alpha$$

$$0.1796 > 0.05$$

Decision: Fail to Reject Null Hypothesis

Conclusion: With 5% level of significance, we have insufficient evidence to claim that at least one age group spends significantly differently than the rest.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
36-45	18-25	1290.194	484.956	-142.56	2722.949	0.1096
36-45	46-50	1267.508	593.228	-485.12	3020.141	0.3319
36-45	0-17	1122.267	1074.389	-2051.91	4296.444	0.9435
36-45	51-55	1032.801	682.414	-983.32	3048.926	0.7368
36-45	26-35	694.928	415.570	-532.83	1922.688	0.6349
55+	18-25	657.508	824.844	-1779.41	3094.427	0.9853
55+	46-50	634.822	892.807	-2002.89	3272.531	0.9920
36-45	55+	632.686	819.710	-1789.07	3054.438	0.9876
26-35	18-25	595.266	425.606	-662.15	1852.677	0.8027
26-35	46-50	572.580	545.782	-1039.88	2185.038	0.9423
55+	0-17	489.581	1264.712	-3246.89	4226.048	0.9997
26-35	0-17	427.339	1048.938	-2671.64	3526.322	0.9996
55+	51-55	400.115	954.397	-2419.56	3219.786	0.9996
26-35	51-55	337.873	641.598	-1557.66	2233.409	0.9985
51-55	18-25	257.393	688.572	-1776.93	2291.711	0.9998
51-55	46-50	234.707	768.680	-2036.28	2505.695	0.9999
0-17	18-25	167.927	1078.311	-3017.84	3353.690	1.0000
0-17	46-50	145.241	1131.146	-3196.62	3487.101	1.0000
51-55	0-17	89.466	1180.365	-3397.81	3576.737	1.0000
55+	26-35	62.242	786.055	-2260.08	2384.565	1.0000
46-50	18-25	22.686	600.301	-1750.85	1796.216	1.0000

Figure 8: Pairs of Means Table

By comparing all pairs of means, there is not a single pair that has a p-value less than the significance level, which means that age group does not have a significant impact on purchase amount.

Analysis 4

ABC Private Limited has a human resources department that is disinterested in economics and more interested in correlations. They have asked the statistics department to examine the relationship between age and marital status in the sample. The company aims to identify the age group with the highest rate of marriage. They have provided a level of significance $\alpha = 0.05$.

$$H_0: Marital\ status\ is\ independent\ of\ age$$

$$H_1: Marital\ status\ is\ dependent\ on\ age$$

Tests

N	DF	-LogLike	RSquare (U)
1000	6	63.979079	0.0941

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	127.958	<.0001*
Pearson	118.047	<.0001*

Figure 12: Chi-Squared Test of Independence

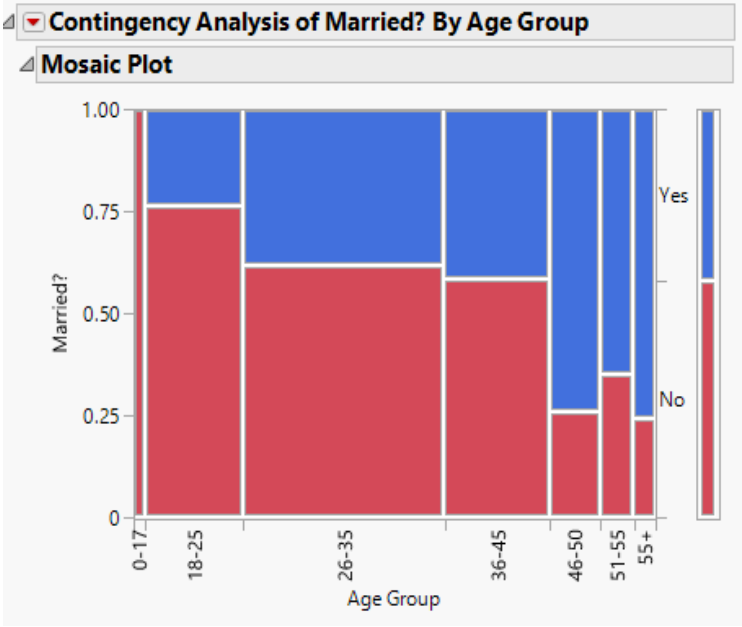

Figure 11: Mosaic Plot of Age Group vs Marital Status

$$p - value < \alpha$$

$$.0001 < 0.05$$

Decision: Reject Null Hypothesis

Conclusion: At 5% level of significance, we have sufficient evidence to conclude that there is an association between age and marital status. Additionally, after analyzing the mosaic plot, it was observed that the age groups 46-50 and 55+ exhibit the highest proportion of married individuals.

Conclusions:

The statistical analyses provided several valuable insights regarding customer purchasing behavior and demographics that can guide marketing initiatives for ABC Private Limited:

- Analysis 1 showed that the company's customers did not spend more than the average American on Black Friday. In fact, the company's customers actually spent much less. Although perhaps both American customers and customers at ABC Private Limited spent a significantly larger amount on Black Friday than they would on a normal day, the American still spent much more.
- Analysis 2 proved that men and women do not have the same spending habits. Men spend on average a significantly higher amount on Black Friday than women.
- Analysis 3 identified that there was no age group that spent a significantly different amount than the rest. However, just for the company's records, the 36-45 age group spent the highest amount on average.
- Age and marital status were shown to have a strong correlation by Analysis 4. The age group with the highest rate of married status is 46-50 and 55+.

Recommendations:

Some recommendations for ABC Private Limited:

- The company should consider expanding locations to the United States of America, since on average, Americans spend more on Black Friday.
- Since women are spending less than men, the marketing apartment should aim to increase their advertising towards female customers to increase company profits.
- In order to maximize profits, the company should tailor promotions to professionals in their peak earning years (36-45), since they are more likely to spend more.
- Although marital status is not typically a data point that ABC Private Limited collects from its users, transitioning from single-people items to more couple-related items as the customer ages is recommended to increase profits.

References

"2023 Holiday Shopping Trends & Insights | Adobe for Business." *Adobe Experience Cloud*, business.adobe.com/resources/holiday-shopping-report.html. Accessed 13 Dec. 2023.

Coppola, Daniela. " E-Commerce as Percentage of Total Retail Sales Worldwide from 2015 to 2027 ." *Statista*, 29 Aug. 2023, www.statista.com/statistics/534123/e-commerce-share-of-retail-sales-worldwide/.

Pietryga, Sabrina. "Black Friday-Cyber Monday Embraced as Consumers Seek Deals to Stretch Their Budgets." *Deloitte United States*, 17 Nov. 2023, www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/black-friday-cyber-monday-embraced-as-consumers-seek-deals-to-stretch-their-budgets.html#.

Vardhan, Ishan. "Black Friday Sales: Analysis and Prediction." *Kaggle*, Kaggle, 11 June 2019, www.kaggle.com/code/ishanvardhan/black-friday-sales-analysis-and-prediction.