Shopping Trends Data Analysis





Dani Luna

Shauntel Phillips

Jon Unger

Ivette Reese

**Introduction**

We found a dataset of a decent sample size, with quantitative and qualitative data, with no null values and did not need any cleaning so we

Believe we could generate and conduct various analyses to provide insights into consumer patterns in the U.S.

**Research questions**

* Are male or female customers making more purchases?
* Which season had the most purchases?
* How frequent are the purchases?
* How consumers with a subscription are represented by gender?
* Is there a correlation between Age and Purchase Amount (USD) Average? Age and Review Rating Average?
* Are Gender means in Purchase Amount (USD) different? In Review Rating different?
* What percent of each payment option was used?
* Do certain states have a higher usage of discount and promo codes?

A graph showing a number of men and women

Description automatically generated

Figure 1. More male than female customers.

Total customers: 3,900 Male: 2,652 Female: 1,248

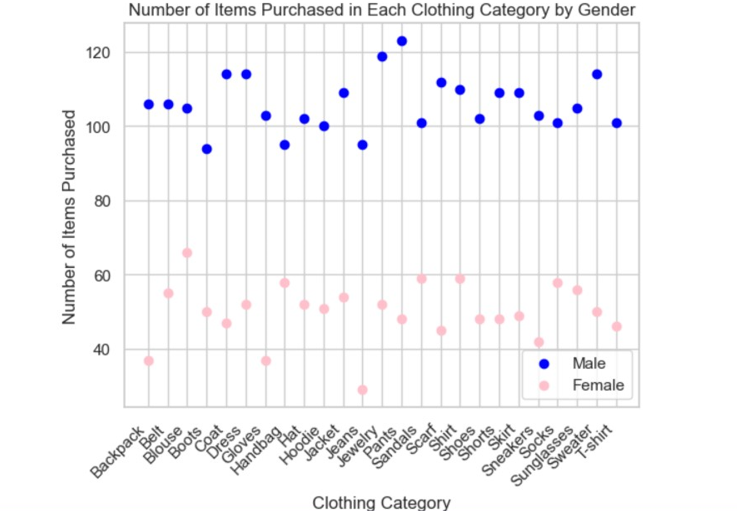


Figure 2. Male customers had purchased significantly more in each clothing category than female customers.

A graph showing a number of purchases

Description automatically generated

Figure 3. Male customers had more previous purchases than female customers.

A blue and pink pie chart

Description automatically generated

Figure 4 above. The Average of previous purchases made by male and female.

Although there are more male customers, female customers are proving to be repeat customers.

A pie chart with numbers and a red triangle

Description automatically generated

Figure 5. Subscription status (27.1% of customers subscribed.

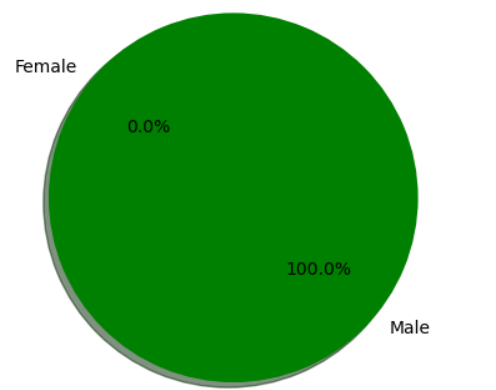


Figure 6. Subscription services by gender. There were 1053 subscriptions, and all subscribers were males.

A graph of different seasons

Description automatically generated

Figure 7. How many purchases by season.

Spring 999

Fall 975

Winter 971

Summer 955

Spring had the most purchase orders. However, Fall brought in the larger profit.

A graph with green bars

Description automatically generated

Figure 8. The frequency of purchases.

My exploratory journey started with correlation questions using a regression line to determine the r-squared that would tell us the strength of the relationship between two quantitative variables and resolve if the model could be used for prediction in consumer trends.

Initially, we wanted to answer my questions with three columns: Age, Purchase Amount (USD), and Review Rating from the original DataFrame. First, I was curious to find the average age (44 years old) in the dataset. Then, I calculated how many purchases were made by age and how many review ratings were made by age. However, I needed averages to plot the regression lines, not the counts.

**Data engineering**. I had to opt for data engineering by adding two new columns to the original DataFrame to plot the regression lines. The new columns were Purchase Amount (USD)\_avg by Age (Figure 9) and Review Rating\_avg by Age; then, I merged the two new means into the DataFrame

(Figure 10)

A screenshot of a computer

Description automatically generated

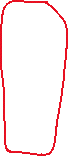


Figure 9



A screenshot of a computer

Description automatically generated

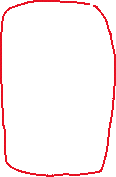


Figure 10



**Linear regression**. With a similar hypothesis (age being the independent variable), I plotted the two regression lines, and both had an r-squared equal to zero and correlation coefficients of zero, as shown in Figures 11 and 12.

A graph with numbers and a red line

Description automatically generated

**Hypothesis**. Age impacts Purchase Amount (USD).

The scatter plots indicate no relationship between the variables; no relationship between Purchase Amount (USD)\_avg and age.

A graph with blue dots and red line

Description automatically generated

**Hypothesis**. Age impacts Review Rating.

The scattered plots indicate no relationship between the variables; no relationship between Review Rating\_avg and Age.

We could not use either line model for prediction. I could not establish that Age had any effects in Amount purchase or Review Rating. I calculated the mean and median for both datasets to verify my results. For Age vs. Purchase Amount (USD)\_avg, the mean = 59.76 and the median =59.66, while the mean and the median for Age vs. Review Rating\_avg were mean =3.748 and the median = 3.744. In both cases, the mean and the median were almost the same. Both graphs have a normal distribution. In a normal distribution, the means lie on the line, the points centered on the line or closer to the line. These results proved that the regression lines were accurate at establishing that the variables had no impact on each other. Furthermore, the points on the scattered plots did not show any pattern; there is no relationship between them.

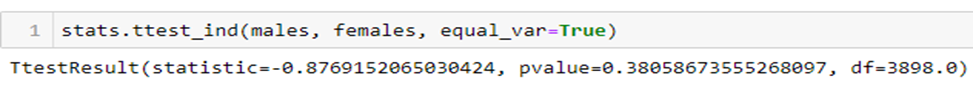
**Ttest**. Afterward, I used a Ttest to compare the two means of the two new columns. However, I could not compare the two averages to determine a difference between the variables because the mean and median of both averages were very close, and we knew by the regression lines that the two groups were different. I had to find two groups (not averages) to determine whether the two groups' means were different. I decided on the group of males and the group of females. First, Gender against Purchase Amount (USD).

A diagram of a person and person

Description automatically generated

**Hypothesis**. My null hypothesis was Gender means in Purchase Amount (USD) are equal. The alternative hypothesis was Gender means in Purchase Amount (USD) are not equal.

Using the t-test to compare the two means resulted in Gender against Purchase Amount with a p-value = 0.38.



Then I calculated the tTest for Gender means in Review Rating.

A blue and orange leaf with red circles

Description automatically generated

**Hypothesis**. My null hypothesis was Gender means in Review Rating are equal. The alternative hypothesis was Gender means in Review Rating are not equal.

Using the t-test to compare the two means resulted in Gender vs Review Rating p-value = 0.61.

A screenshot of a computer code

Description automatically generated

When I compared both p-values to a significance level of 0.05,

both p-values were higher, which indicates results were not statistically significant for both datasets. I calculated the mean for both groups to verify my results. The two group means for either the male group mean (59.53) or the female group mean (60.24) in the Purchase Amount (USD) are very close. The means for the male group (3.75) and female group (3.74) in Review Rating are also very close.

No correlation with a regression line supports my findings, the mean and median numbers are very close, and the tTest results p-values higher than 0.05. There is no correlation or predictive power between Purchase Amount (USD)\_avg and Age, neither by Review Rating\_avg and Age as shown by the two flat regression lines. There is no ability to predict whatsoever. Neither model fits the data well.

I failed to reject both null hypotheses that the Gender means in Purchase Amount are equal, and Gender means in Review Rating are equal. Both male and female group means in Purchase Amount, and male and female group means in Review Rating are almost the same. The difference between the two groups is not [statistically significant](https://www.scribbr.com/statistics/statistical-significance/). There is likely no difference in male and female groups in both Purchase Amount (USD) and Review Rating.

We concluded that there is no correlation between Age and either Purchase Amount (USD)\_avg or Review Rating\_avg. The regression lines are not a good fit to predict the relationship between age, the independent variable, and either Purchase Amount (USD)\_avg or Review Rating\_avg (dependent variables).

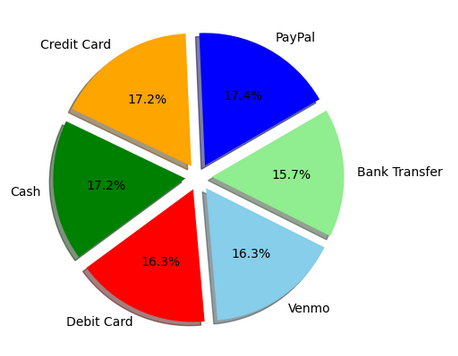
The tTests resulted in p-values greater than 0.05, which are not [statistically significant](https://www.scribbr.com/statistics/statistical-significance/) and indicate strong evidence for the null hypotheses. In other words, I failed to reject the null hypotheses, which suggests I retained the null hypotheses and rejected the alternative hypotheses. The sample data does not provide sufficient data to conclude that there is a difference between the two groups of means. The male and female groups mean in Amount Purchase (USD) or Review Rating are likely equal.

I have learned from analyzing the data set that Age and Gender are independent variables with no significant effect on other variables in the dataset. Therefore, we would not recommend retailers base their market strategy on consumer demographics such as age or gender.

The limitation of the study was the size of the data. I would have preferred a larger size of data; more data would have been more certain data to refute the null hypothesis. A larger sample size would have identified any outliers in the dataset and would be more likely to obtain statistically significant results to generalize to the population.

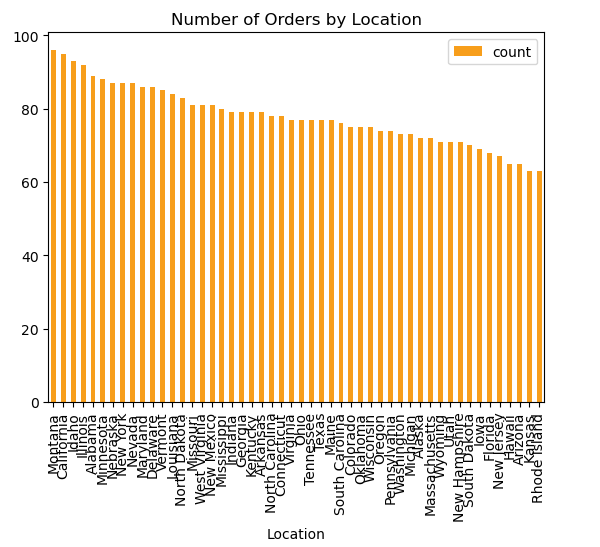
We began the b) questions by importing our libraries and reading in the csv file to display our dataset. We determined that we could get a percentage of the payment source data to determine what if any was the favored method of payment.

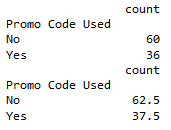
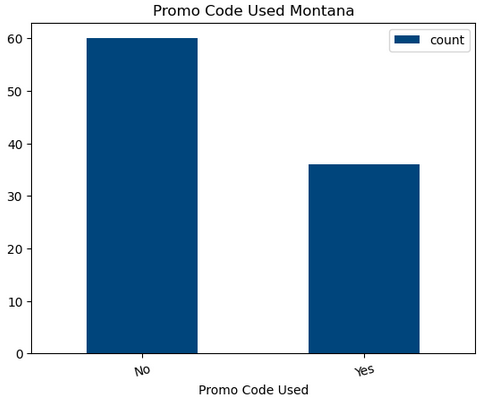
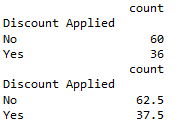
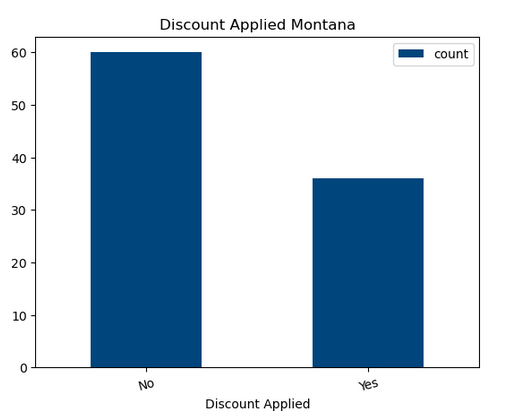
**Hypothesis:** Our hypothesis is that Credit Card was going to be the favored category, with cash being in the last place. This is because shoppers typically do not carry cash with them, and online orders require credit card or bank information, not physical currency. Limitations of this dataset include: not listing specific credit card companies, Venmo being owned by Paypal, Venmo not typically being an accepted form of payment at retailers, Paypal and Venmo being transaction agents and not final payment sources, and what if any was the difference between a Bank Transfer, a Debit Card purchase and Cash? The following is our results for payment method analysis.



Our next question regarded whether certain states had a higher usage of discounts or promo codes. Once we dug into the Discount and Promo Code columns, we discovered that they were likely the exact same dataset repeated. Further analysis should operate on this assumption for other locations in this dataset.

**Hypothesis:** That there is not a significant difference in the number of transactions using discounts and promo codes amongst the states. After our analysis we found that roughly 60-65% of all transactions did not utilize any sort of discount or promo code across all locations provided. Below is a sample of one location’s data- we did not feel the need to provide the graphs obtained from every location, however the code to do so is included in this project.







**Conclusions**

* The regression lines are not a good fit to predict relationships.
* Age and gender are independent variables with no significant effect on other variables in the dataset.
* There is not a preferred payment method among the dataset. However, we find that the data collected is insufficient.

**Call to Action**

* We would not recommend retailers base their market strategy on consumer demographics such as age or gender.

**Limitations/Bias**

* Dataset is prefabricated.
* Size of the data. We originally thought the dataset was a decent size but after starting our various analysis we saw additional data would have been helpful such as a timeframe of this data.
* Men shop more often than females.
* Only men interested in subscriptions services.
* Holiday sales play a role in consumers’ shopping habits.

**Future Work**

* A real dataset of shopping trends
* In person vs online sales
* Types of advertising and the impacts of social media
* Specific bank and credit card data
* A larger sample size

**Works Cited**

https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset/data