Shopping Trends Data Analysis

Dani Luna
Shauntel Phillips
Jon Unger
Ivette Reese

Introduction

The dataset of choice was selected because of sample size, ample quantitative and qualitative data, with no null values. The assumption was that the data could generate and conduct various analyses to provide insights into consumer patterns in the U.S.

Research Questions

1. Gender demographics
    a. Which gender is responsible for the most purchases
    b. How consumers with a subscription are represented by gender?
    c. Does gender play a factor in average purchase dollars?
    d. Will there be a difference in review rating by gender
2. Seasonality
    a. When are majority purchases occurring?
    b. What is the frequency of said purchases?
3. Age- Correlations and relationships
    a. Is there a correlation between Age and Purchase Amount (USD) Average?
    b. Is there a correlation between Age and Review Rating Average?
4. Payment/Discount usage?
    a. What percent of each payment option was used?
    b. Do certain states have a higher usage of discount and promo codes?

Data Descriptives,Describes, and Info

```
In [4]:  1 shopping_trends.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             3900 non-null   int64
 1   Age                     3900 non-null   int64
 2   Gender                  3900 non-null   object
 3   Item Purchased          3900 non-null   object
 4   Category                3900 non-null   object
 5   Purchase Amount (USD)   3900 non-null   int64
 6   Location                3900 non-null   object
 7   Size                    3900 non-null   object
 8   Color                   3900 non-null   object
 9   Season                  3900 non-null   object
 10  Review Rating           3900 non-null   float64
 11  Subscription Status     3900 non-null   object
 12  Shipping Type           3900 non-null   object
 13  Discount Applied        3900 non-null   object
 14  Promo Code Used         3900 non-null   object
 15  Previous Purchases      3900 non-null   int64
 16  Payment Method          3900 non-null   object
 17  Frequency of Purchases  3900 non-null   object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

Figure 1:

```
In [7]:
#how many male customers vs female customers
female_count = shopping_trends[shopping_trends["Gender"] == "Female"].shape[0]
print(f"Number of females: {female_count}")

Number of females: 1248
```

```
In [8]:
male_count = shopping_trends[shopping_trends["Gender"] == "Male"].shape[0]
print(f"Number of males: {male_count}")

Number of males: 2652
```

Figure 1a:

Research Question 1

Upon reviewing the dataset, the focus rested on determining any correlations between purchases and customer gender. The primary goal was to ascertain if a specific gender exhibited a propensity for higher purchase frequency or spending levels at the retail store. To address these inquiries, they commenced by calculating the overall customer count and then segregating them into two groups: male and female.

Upon the completion of calculating the total of male and female customers, a distinct disparity in customer gender became evident. The dataset reflected 2,652 male customers and 1,248 female customers, highlighting a substantial difference of 1,404 between the two groups. This realization hinted at potential bias in research queries centered on customer gender, leaning towards the male customer group due to the significant numerical gap. Nonetheless, the decision was made to persist and delve deeper into analyzing purchases across both genders. The investigation extended by calculating the total quantity of items bought in each clothing category by male and female customers. As anticipated from the earlier gender segmentation, the data showed that male customers had made more purchases across every clothing category in the dataset. These findings led to the conclusion that, across all clothing categories, male customers consistently outpaced female customers in total purchases.

Figure 2:

```
Gender
Female    30696
Male      68175
Name: Previous Purchases, dtype: int64
```

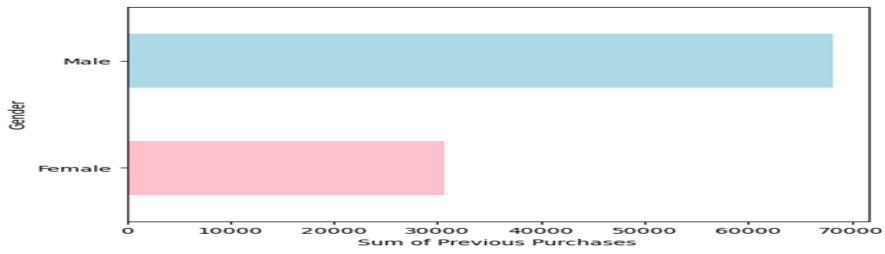


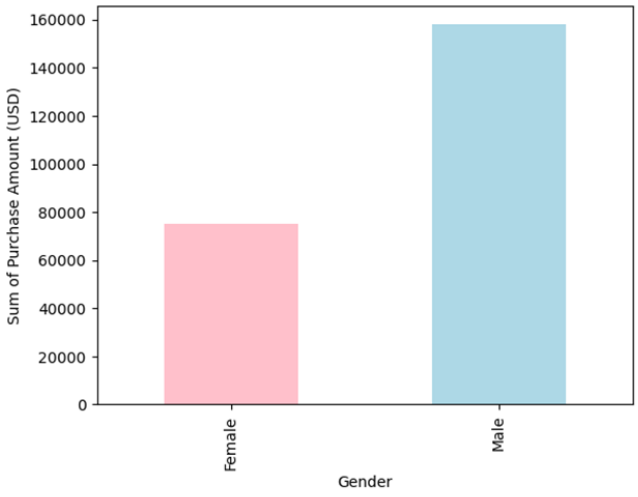
Figure 2a

Following an examination of purchases across genders, the focus shifted to assessing the history of prior purchases made by both groups. The inquiry aimed to discern whether the observed pattern of male customers outpacing female customers in purchases persisted or represented a recent trend. Sorting the data into male and female segments allowed for an evaluation of the previous purchase count for each customer. The findings revealed a substantial contrast: male customers had a total of 68,175 previous purchases, while female customers had recorded only 30,696. Notably, this difference prompted an investigation into the monetary value of these purchases to determine the consistency of the ratio. Grouping the data by 'Gender' and 'Purchase Amount (USD),' the total sum of previous purchases for each gender was calculated. Considering the male customer base accounted for over 50% of the existing customers, there was an expectation that their cumulative previous purchase amount would at least double in comparison. The analysis confirmed this speculation, revealing that male customers had spent $157,890, while female customers had spent a total of $75,191.

Figure 3:

```
Gender
Female     75191
Male      157890
Name: Purchase Amount (USD), dtype: int64
```

Figure 3a.

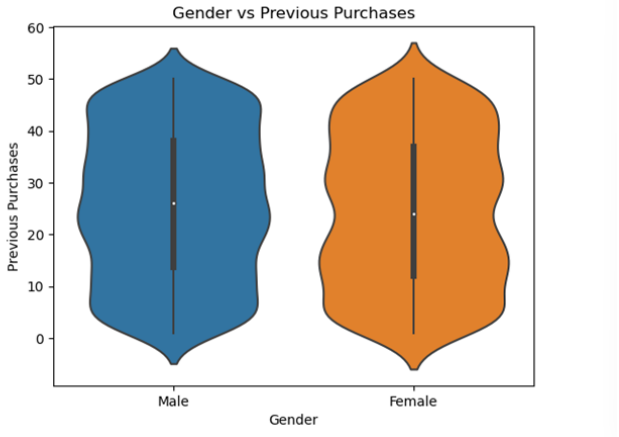


Returning to the dataset encompassing previous customer purchases, the decision was made to explore the average previous purchase count within both customer groups. The intention was to validate whether this metric aligned with the prevailing ratio. Despite the noticeable discrepancy of 37,479 more purchases by male customers compared to female customers, the calculated averages for previous purchases were nearly equal. Male customers averaged 26 previous purchases, while female customers averaged 25. This insight suggested that although male customers accounted for a majority of the purchases, existing female customers displayed a similar trend to continue to make purchases.

Figure 4:

```
Gender
Female    24.596154
Male      25.707014
Name: Previous Purchases, dtype: float64
```

Figure 4a:



In summary, when we looked at what customers bought, men made more purchases and spent more money overall. But when we checked the average, both men and women had bought roughly the same amount on average. Even though men made most of the purchases, women who were already customers tended to buy just as regularly. This tells us that while men made more purchases in total, women who were already customers kept up with buying regularly, balancing things out when we looked at averages.

Research Question 1(cont)

Initially the goal was to discover any trends that can be concluded based on consumers demographics, such as gender and age,that would best enable retailers to target a specific market segmentation. Figure 5, answers how many of the consumer samples are in possession of a subscription. It is demonstrated that the majority of the populus does not retain subscription status. Next, was to discover of consumers that possess a subscription,how many were represented by gender,see Figure 5a. It was discovered that all subscribers were male. It can be

concluded that a limitation/bias lies within the data because this would state that women do not purchase

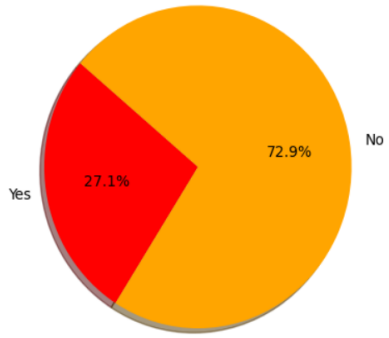


Figure 5: Consumer Subscription Status

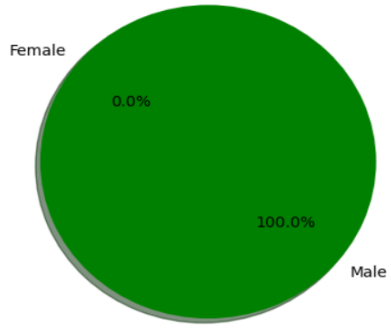


Figure 5a: Subscription services by gender.

Research Question 2

Research question 2 seeks to find opportunities based on seasonal purchases. Secondary questions include:

- Which season are consumers likely to spend the most?
- Based on seasonality,can a retailer expect more revenue?
- What potential product offerings can a retailer introduce?

Figure 6,shows that the amount of purchases across season are relatively the same,however Spring was responsible for the majority of purchases made.
Figure 6a,shows the purchase amount by season.

It is important to note that while Spring has the most purchases, it's Fall that is responsible for the majority purchase dollars. This could be related to common seasonal sales such as Back Friday or Cyber Monday,which explains why there is a larger dollar amount.

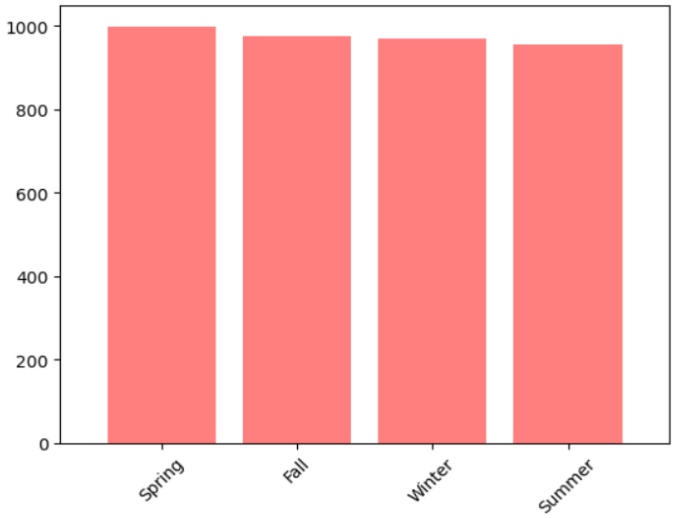


Figure 6. PurchaseCount by Season

Season	
Fall	60018
Spring	58679
Summer	55777
Winter	58607

Figure 6a. Purchase Amount (USD) by Season

Research Question 3

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

Data engineering. 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 7) and Review Rating_avg by Age; then, merged the two new means into the DataFrame (Figure 8).

```
average_purchase_amount_by_age = shopping_trends_df.groupby("Age")["Purchase Amount (USD)"].mean().reset_index()
average_purchase_amount_by_age.head()
```

	Age	Purchase Amount (USD)
0	18	59.985507
1	19	61.000000
2	20	55.225806
3	21	64.101449
4	22	60.772727

```
# Merge the average purchase amount back into the original DataFrame
shopping_trends_df = pd.merge(shopping_trends_df, average_purchase_amount_by_age, on="Age", how="left", suffixes=('', '_
shopping_trends_df.head()
```

nder	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	...	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases	Purchase Amount (USD)_avg
Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	...	Express	Yes	Yes	14	Venmo	Fortnightly	59.369863
Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	...	Express	Yes	Yes	2	Cash	Fortnightly	61.000000
Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	...	Free Shipping	Yes	Yes	23	Credit Card	Weekly	59.397590
Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	...	Next Day Air	Yes	Yes	49	PayPal	Weekly	64.101449
Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	...	Free Shipping	Yes	Yes	31	PayPal	Annually	56.736111

Figure 7. Purchase Amount (USD)_avg

```
average_review_rating_by_age = shopping_trends_df.groupby("Age")["Review Rating"].mean().reset_index()
average_review_rating_by_age.head()
```

	Age	Review Rating
0	18	3.744928
1	19	3.820988
2	20	3.841935
3	21	3.815942
4	22	3.730303

```
# Merge the average review rating back into the original DataFrame
shopping_trends_df = pd.merge(shopping_trends_df, average_review_rating_by_age, on="Age", how="left", suffixes=('', '_av
shopping_trends_df.head()
```

Purchase Amount (USD)	Location	Size	Color	Season	...	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases	Purchase Amount (USD)_avg	Review Rating_avg
53	Kentucky	L	Gray	Winter	...	Yes	Express	Yes	Yes	14	Venmo	Fortnightly	59.369863	3.663014
64	Maine	L	Maroon	Winter	...	Yes	Express	Yes	Yes	2	Cash	Fortnightly	61.000000	3.820988
73	Massachusetts	S	Maroon	Spring	...	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly	59.397590	3.825301
90	Rhode Island	M	Maroon	Spring	...	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly	64.101449	3.815942
49	Oregon	M	Turquoise	Spring	...	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually	56.736111	3.626389

Figure 8. Purchase Amount (USD)_Avg and Review Rating_avg Merged

Linear regression. With a similar hypothesis (age being the independent variable), the plotted two regression lines had an r-squared equal to zero and correlation coefficients of zero, as shown in Figures 9 and 10.

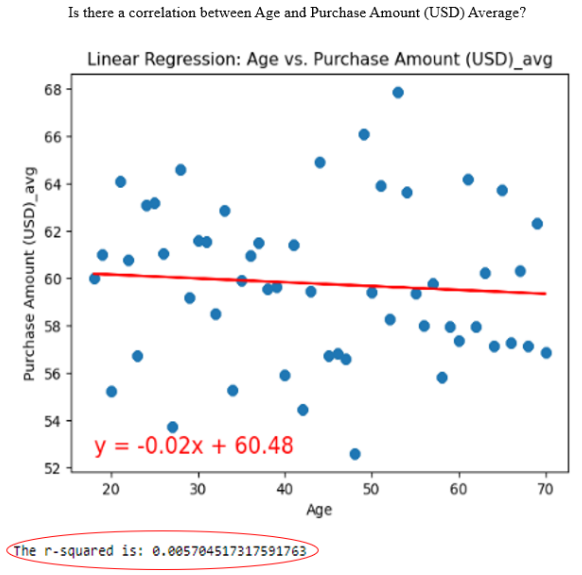


Figure 9. Linear Regression

Hypothesis. Age impacts Purchase Amount (USD)

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

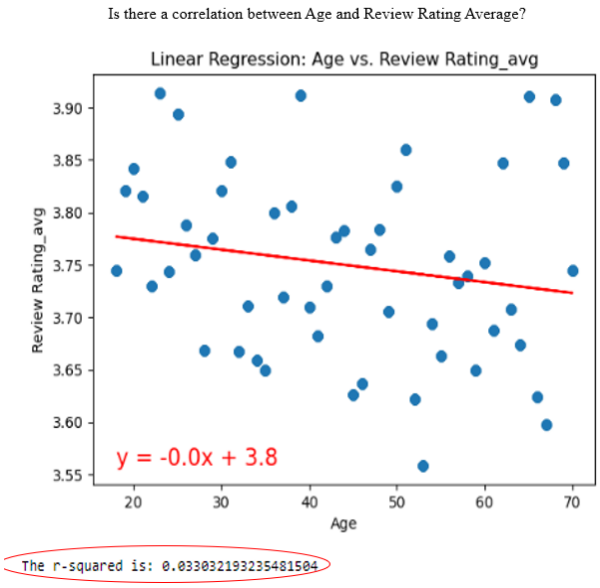


Figure 10. Linear Regression

Hypothesis. Age impacts Review Rating.

The scattered plots indicate no relationship between the variables; no relationship between Review Rating_avg and Age.

Neither line could not be used as a model for prediction. Age had no effects in Amount purchase or Review Rating. To verify results, calculated the mean and median for both datasets. 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.

Test. Afterward, a Test to compare the two means of the two new columns. However, 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 the regression lines for the two groups were different. Had to find two groups (not averages) to determine whether the two groups' means were different. The group of males and the group of females. First, Gender against Purchase Amount (USD) -Figure 11 below.

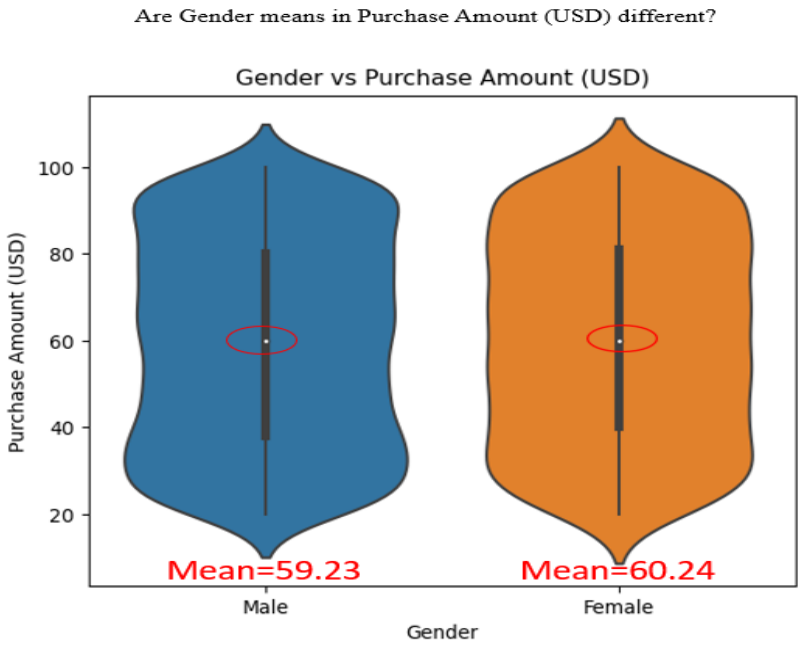


Figure 11. Ttest

Hypothesis. 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.

```
1 stats.ttest_ind(males, females, equal_var=True)
```

```
TtestResult(statistic=-0.8769152065030424, pvalue=0.38058673555268097, df=3898.0)
```

Then calculated the tTest for Gender means in Review Rating (Figure 12 below)

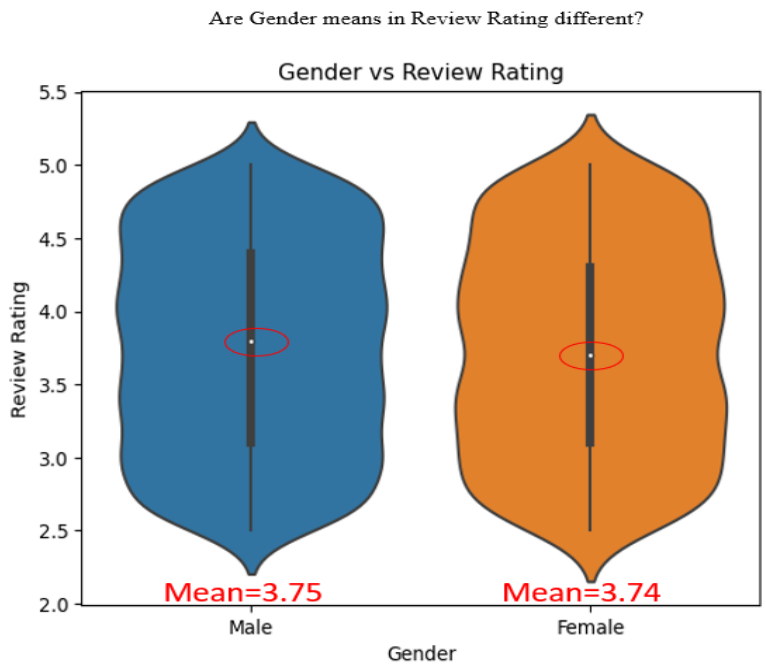


Figure 12. Ttest

Hypothesis. The 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.

```
1 stats.ttest_ind(males, females, equal_var=True)
```

```
TtestResult(statistic=0.5097147504896427, pvalue=0.6102801734916257, df=3898.0)
```

When comparing both p-values to a significance level of 0.05, both p-values were higher, which indicated results were not statistically significant for both datasets. Next, calculated the mean for both groups to verify the 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 the 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.

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. There is likely no difference in male and female groups in both Purchase Amount (USD) and Review Rating.

Research Question 4

What percent of each payment option was used?
The analysis began by importing python libraries and reading in the CSV file to display the study dataset. It was determined that payment source data could be analyzed to determine the favored payment method.

Hypothesis: This study hypothesizes that credit cards will be the favored category, with Cash 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 differentiating between a Bank Transfer, a Debit Card purchase, and Cash. The following are our results for payment method analysis.

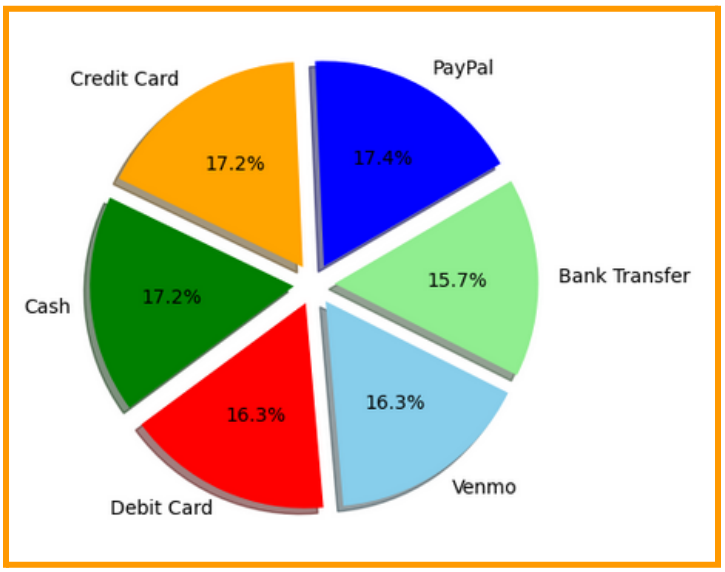
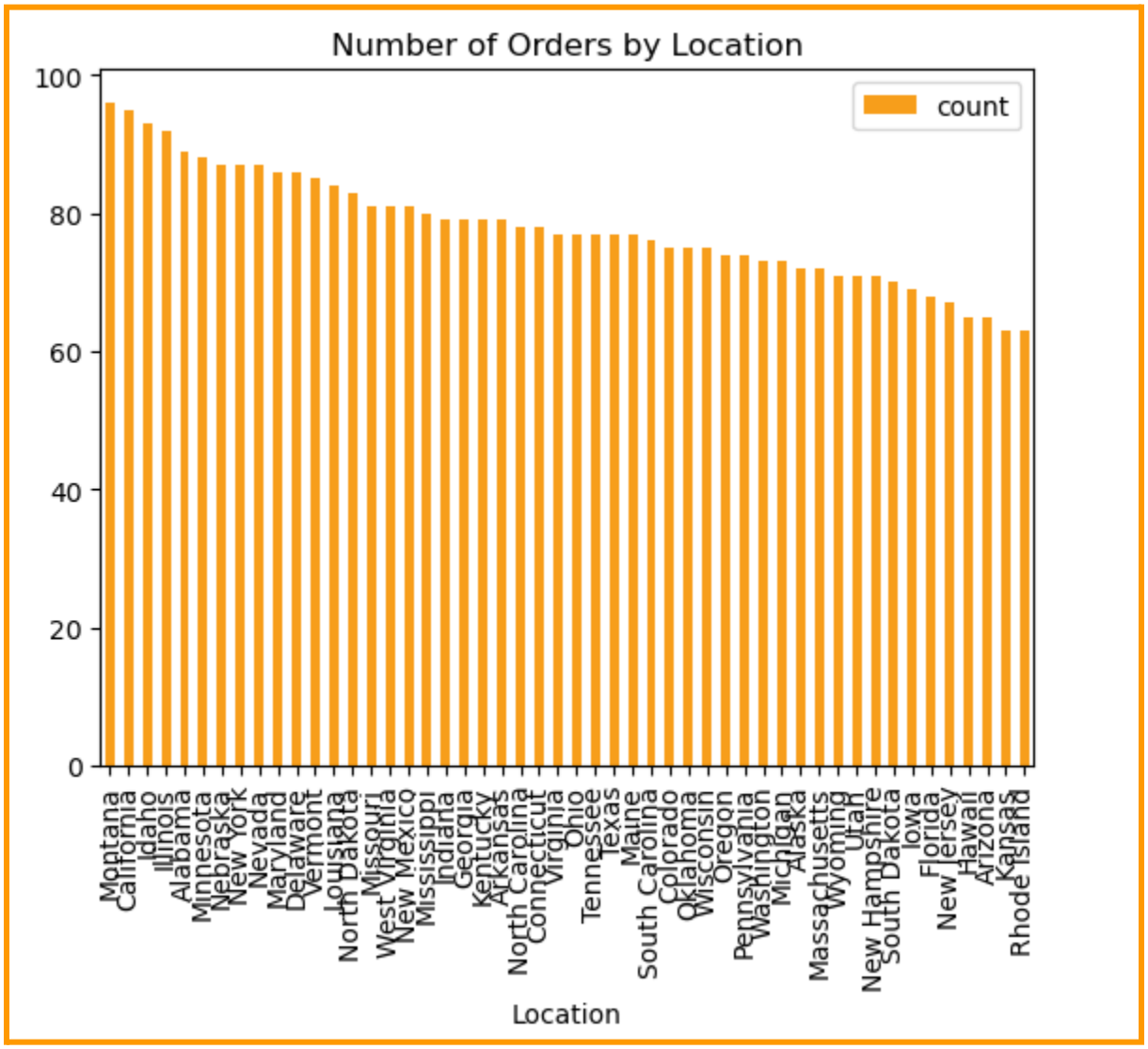


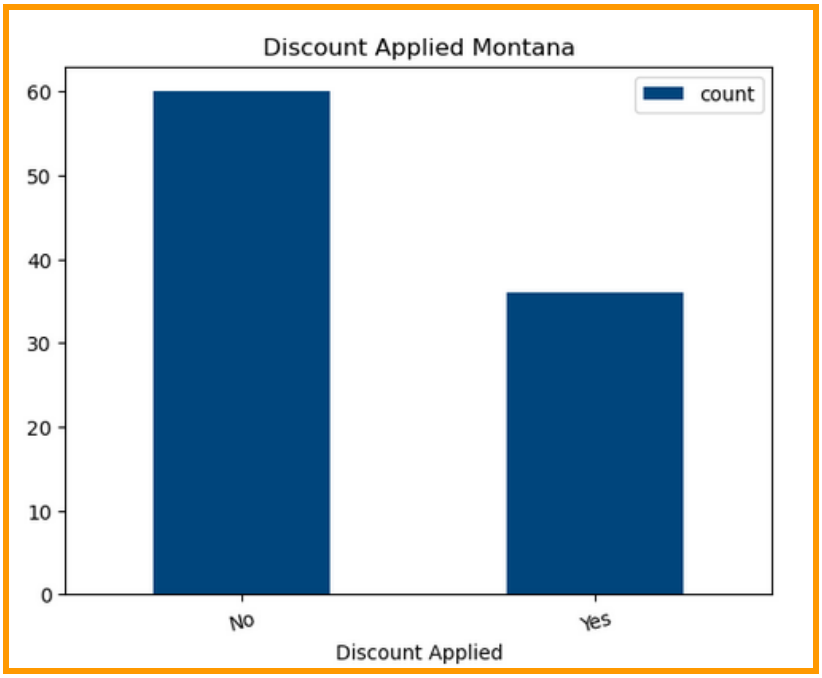
Figure 13

Do certain states have a higher usage of discount and promo codes?
The next question regarded whether certain locations had a higher usage of discounts or promo codes. Once the analysis dug into the Discount and Promo Code columns, it was discovered that these columns were likely the same dataset repeated. Further analysis should operate on this assumption for other locations in this dataset.

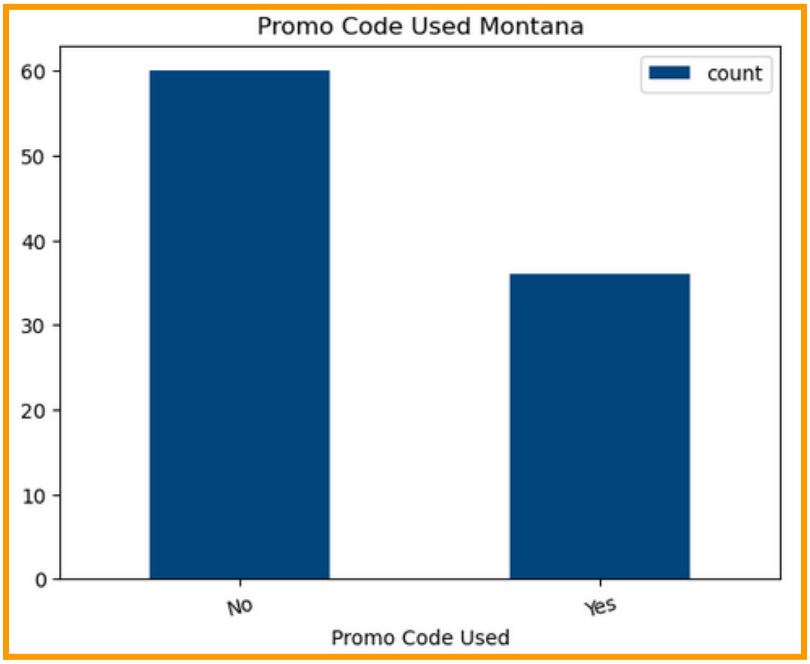
Hypothesis: The number of transactions using discounts and promo codes is the same among the states. After analysis, this study found roughly 60-65% of all transactions did not utilize any discount or promo code across all locations provided. This study holds that the hypothesis is true, and there was not a significant difference between the locations. Below is a sample of one location's data- we did not feel the need to provide the graphs obtained from every area in this report. However, the code to do so is included in this project.

Figure 14,15,16,17,18,19,





Discount Applied	count
No	60
Yes	36
Discount Applied	count
No	62.5
Yes	37.5



Promo Code Used	count
No	60
Yes	36
Promo Code Used	count
No	62.5
Yes	37.5

```
# Select _Location_,
montana_df = reduced_shopping_trends_df.loc[reduced_shopping_trends_df["Location"] == "Montana"]
print(montana_df)

# Discount df
count_montana_discount = montana_df["Discount Applied"].value_counts()
count_montana_discount_df = pd.DataFrame(count_montana_discount)
print(count_montana_discount_df)
print(count_montana_discount_df / len(montana_df)*100)

# Plot data
count_montana_discount_df.plot.bar(rot=15, color= "#00457c", title="Discount Applied Montana");
plt.show(block=True);

# Promo Code df
count_montana_promo = montana_df["Promo Code Used"].value_counts()
count_montana_promo_df = pd.DataFrame(count_montana_promo)
print(count_montana_promo_df)
print(count_montana_promo_df / len(montana_df)*100)

# Plot data
count_montana_promo_df.plot.bar(rot=15, color= "#00457c", title="Promo Code Used Montana");
plt.show(block=True);
```

Conclusion

The regression lines are not a good fit to predict relationships. 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). Age and gender are independent variables with no significant effect on other variables in the Dataset. The tTests resulted in p-values greater than 0.05, which are not statistically significant and indicate strong evidence for the null hypotheses. In other words, failed to reject the null hypotheses, which suggests retaining the null hypotheses and rejecting 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. There is no preferred payment method in the Dataset. However, we find that the data collected is insufficient.

Learned from analyzing the data set that Age and Gender are independent variables with no significant effect on other variables in the Dataset. Therefore, would not recommend retailers base their market strategy on consumer demographics such as age or gender. The limitations and biases found included that the dataset was prefabricated, men shop more often than women and only men were interested in subscription services. Holiday sales play a role in consumers' shopping habits. The initial thought was that the Dataset was a decent size, but after starting various analyses, seeing additional data would have been helpful, such as a timeframe for this data. 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 have been more likely to obtain statistically significant results to generalize to the population. Realized that the future should include a real dataset of shopping trends, in-person vs online sales, advertising types and social media impacts, specific bank and credit card data, and a larger sample size.

Works Cited

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