Middle East Technical University

Shopping Behavior in United States: Influence of Gender and Average Review Rating

Course Code and Name: STAT112: Introduction to Dataset and Visualizations

From: Ozan Mert İBİŞ

To: Prof. Dr. Ceylan YOZGATLIGİL

Date: 10.12.2023

**INTRODUCTION**

The research is about profiles, characteristics, and behaviors of 3900 different customers in the United States, and how those can affect the marketing strategies, trends, purchasing patterns and the preferences. 2 datasets are combined by their common “location” variable.

In the final combined dataset, we have 21 different variables.

* **Customer ID(:** A unique identifier that is assigned to each individual customer.
* **Age:** The age of the customer
* **Gender:** The gender of the customer
* **Item Purchased:** The specific product or item selected by the customer.
* **Category:** The broad classification or group of the purchased item
* **Purchase Amount (USD):** The value of the purchase, in terms of US Dollars
* **Location:** The geographical location where the purchase was made
* **Size:** The size specification (if applicable) of the purchased item, relevant for apparel, footwear, and certain consumer goods.
* **Color:** The color of purchased item
* **Season:** The season in which the purchase transaction takes place.
* **Review Rating:** A numerical or qualitative assessment provided by the customer regarding their satisfaction with the purchased item.
* **Subscription Status:** Indicates whether the customer has opted for a subscription service.
* **Shipping Type:** Specifies the method used to deliver the purchased item.
* **Discount Applied:** Indicates if any promotional discounts were applied to the purchase.
* **Promo Code Used:** Notes whether a promotional code or coupon was utilized during the transaction.
* **Previous Purchases:** Provides information on the number or frequency of prior purchases made by the customer.
* **Payment Method:** Specifies the mode of payment preferred by the customer.
* **Frequency of Purchases:** Indicates how often the customer engages in purchasing activities.
* **Latitude:** Latitude of the location numerically.
* **Longitude:** Longitude of the location numerically.
* **Location Code:** Two letter abbreviation of the locations (the states of USA).

**DATA CLEANING**

In the “Frequency of Purchases” variable, the sub-strings of “Quarterly” and “Every three months” indicate the same amount of duration. Thus, all “Quarterly" strings were combined with “Every three months”. In the same way, sub-strings of “Fortnightly”, and “Bi-Weekly” were combined since they are also the same time interval.

Variables “Discount Applied” and “Promo Code Used” display the same statistics. Therefore, they are the same. Instead of using the same variable twice, we can eliminate one of those variables.

Variable Location(Location) in the dataset does not have a geographical role. By assigning the State/Province geographical role to the variable, and editing the locations by matching location names with US states, we can convert the Location(Location) variable into a geospatial variable.

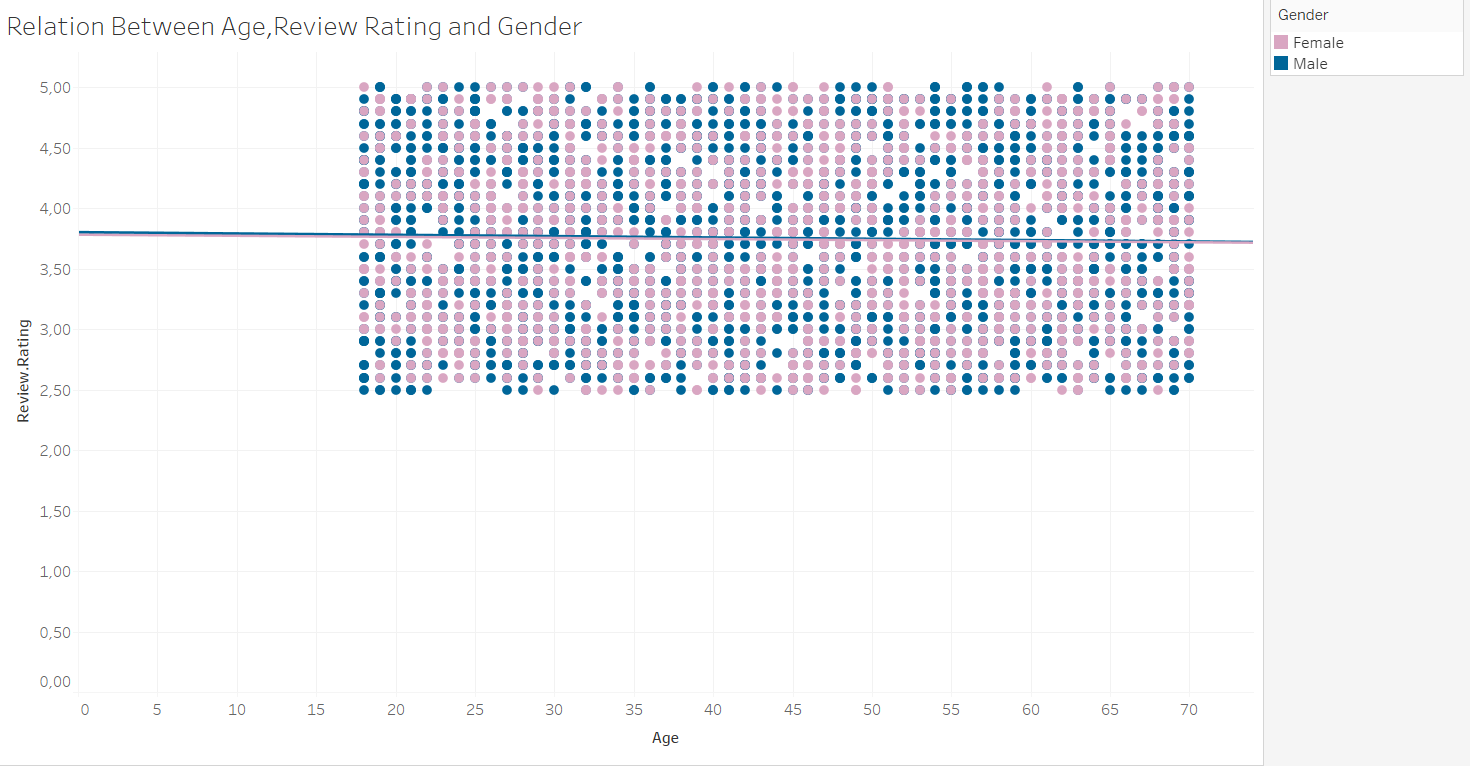
**EXPLORATORY DATA ANALYSIS**

**Summary Statistics of the Numeric Variables**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | x̄ (Arithmetic Mean) | s (Standard Dev) | Minimum  Value  (Min) | Median  Value (Q2) | First Quartile (Q1) | Third Quartile (Q3) | Maximum  Value  (Max) |
| AGE: | 44.07 | 15.21 | 18 | 44 | 43.19 | 45.27 | 70 |
| REVIEW RATING: | 3.75 | 0.72 | 2.5 | 3.7 | 3.10 | 4.40 | 5.0 |
| AVERAGE PURCHASE AMOUNT($) | 59.76 | 23.69 | 20 | 60 | 39 | 81 | 100 |

**RESEARCH QUESTIONS**

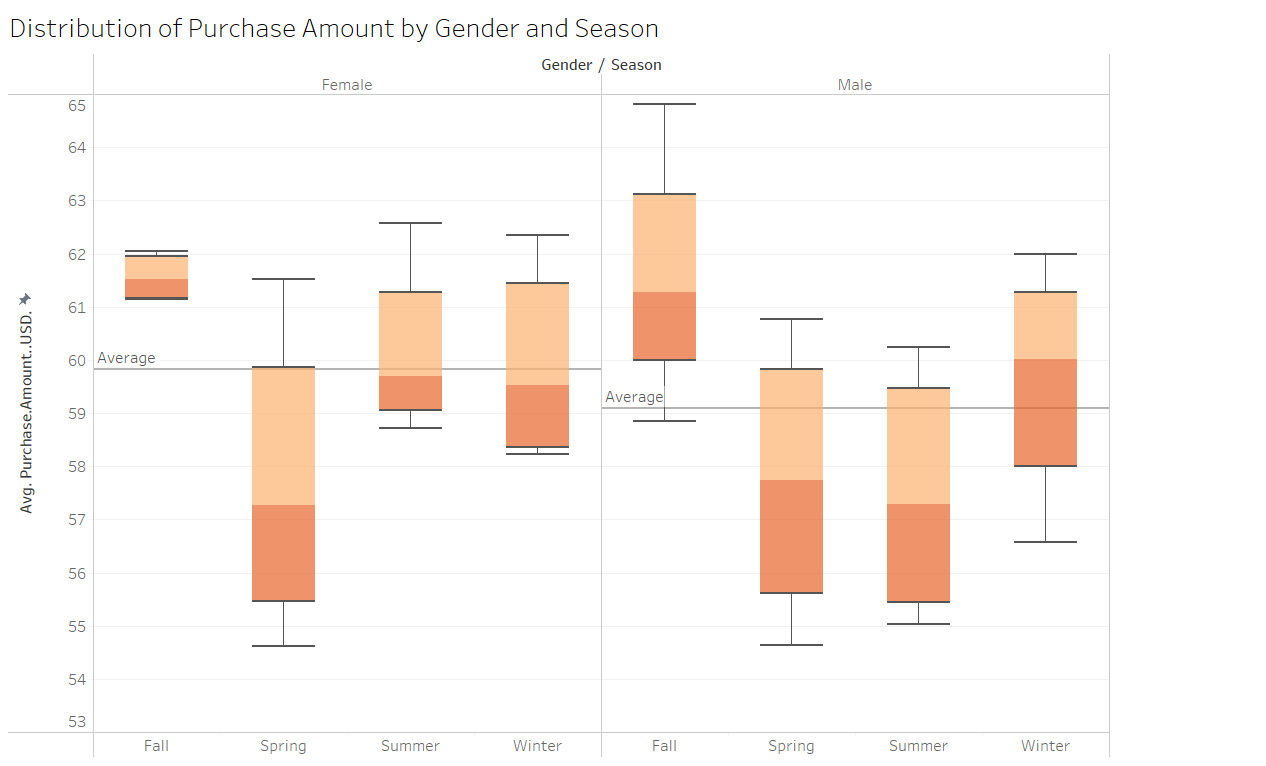
**1.Are people with higher average rating older or younger people? Is gender also influential on average ratings?**

By the help of scatter plot, we can say that neither the gender nor the age has an impact on average ratings as the correlation coefficients (r), which are displayed by the trend lines in this scatter plot, are very close to zero for both genders, implying no meaningful relation.

**2.How does the average rating change according to the category of the purchased product?**

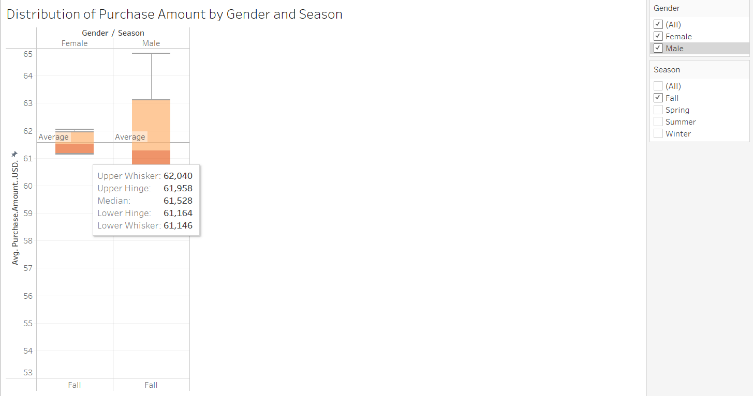
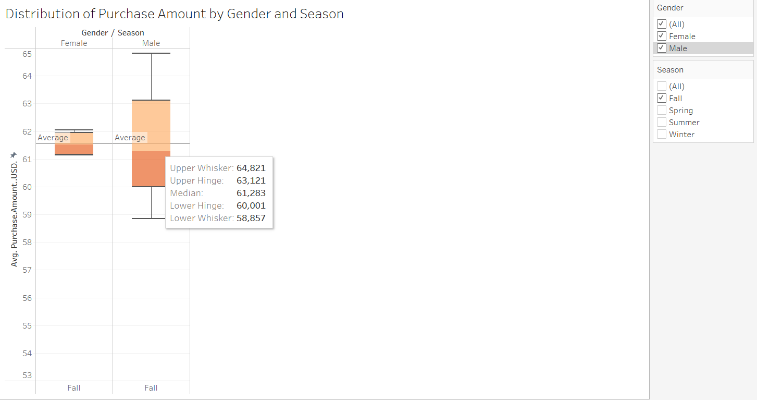
**A graph with numbers and points

Description automatically generated with medium confidence**By looking at both line and lollipop charts, we can say that clothing has the lowest average rating among the categories. Moreover, we have the highest amount of data for clothing, meaning that our data is more reliable than the other categories, especially footwear and outerwear categories. For the highest, footwear has the highest average rating.

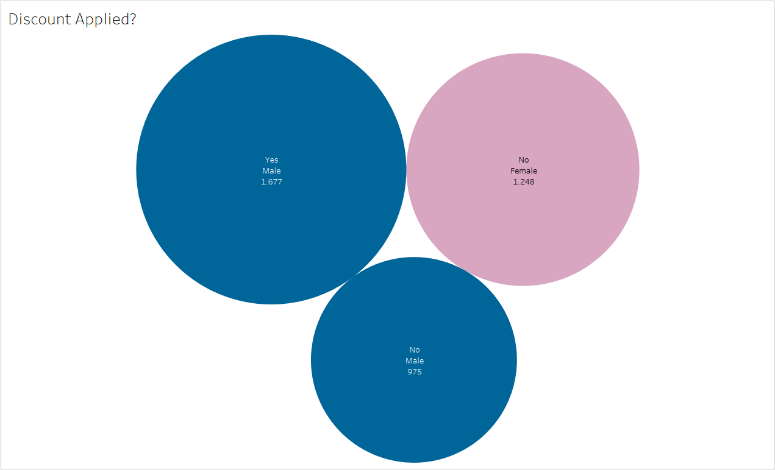
**3.How is the distribution of average purchase Amount (In USD) by Gender and Season?**

Looking at the box plot; for summer, fall and winter seasons, Distribution of purchase amount in males are more various than in females. In other words, females are more consistently spending more money than males as we can also interpret from the average lines.

To comment on the skewness of the distributions, all box plots are very close to a symmetric distribution, but some of box plots have their median close to first quartile such as in summer season and female gender. For this case, there is a right skewed distribution.



For instance, in the fall season, all purchase amount values in females are between 61.15$ and 62.04$. On the other hand, they are between 58.86$ and 64.82$ for males, showing greater variation. However, for spring season, the box plots for each gender are very similar. In other words, males and females are doing much closer amounts of purchase.

**4.Comparison of genders by whether a discount is used or not.**

In these packed bubbles, the gender can be seen as color, the application status of discount can be seen in the label, and size of the bubbles represent the data count. In the light of this, we can state that the number of people who applied for discounts is less than those who didn’t apply. Besides, there is something unusual about this data. Surprisingly, there doesn’t exist a single female that applied discounts.

**5. How does the purchase frequency of people change? Does gender have an impact on it?**

A screenshot of a computer screen

Description automatically generatedIn the tree map, we can see that People tend to make more purchases either bi-weekly or every three months, having the most data points in total, 1147 and 1089 respectively. Least preferred is purchasing weekly, having only 539 data points in total.

To see a relationship between genders and purchase frequency, we shouldn’t look at the numbers because in our dataset we don’t have close number of people in each gender. Instead of the number of people, we can look at the percentages. In all of the frequency of purchases, the percentages of males and females are very close to each other. Hence, there is no impact of gender on the purchase frequency.

**6. Which purchased items’ value is the most and least in total? Is it also the case for both genders?**

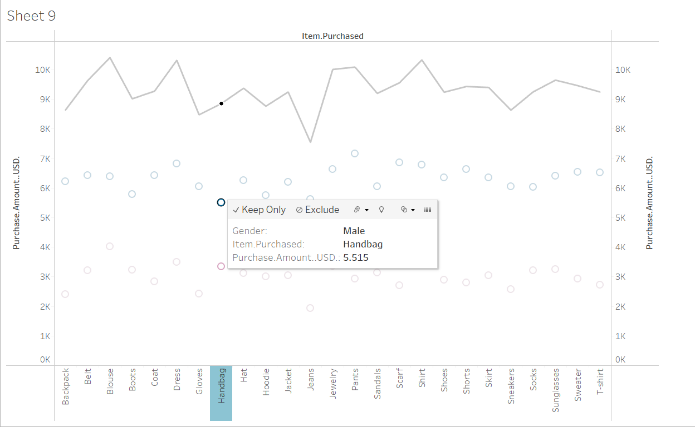
**A graph of a graph

Description automatically generated with medium confidence** Looking at the combined graph of both purchased items separately for each gender as circles, and their total as a line graph, we can say that Blouse is the item with the most purchase value with a value of $10.410, whereas Jeans have the least purchase value with $7.548.

For Blouse, both answers are clear from the graph. it also has the most purchase value for females, i.e. Money is spent the most for blouses by females, with value of $4.016. However, it is not the same case for males, as they spend money for the Pants the most with value of $7.154, compared to $6.394 for Blouses.

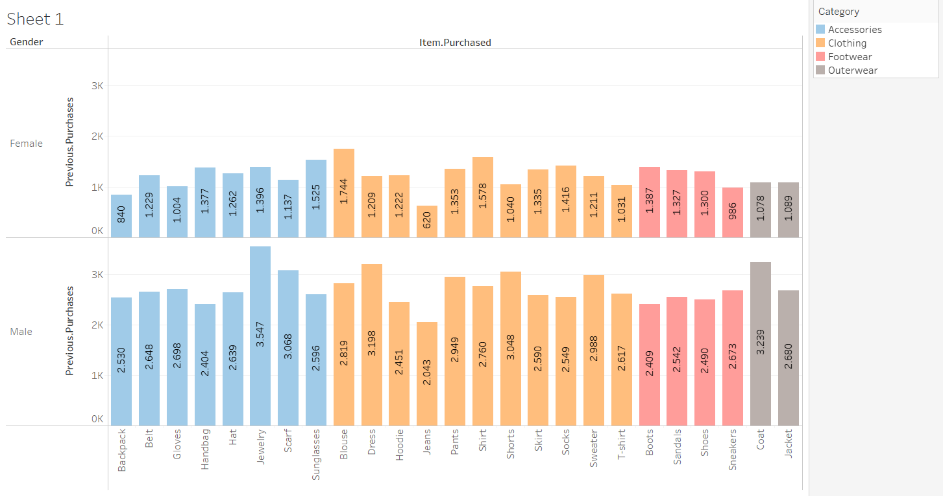
A screenshot of a graph

Description automatically generatedFor the Jeans, it is clear from the graph that Jeans also has the least value for female purchases. However, it is hard to determine for males as Handbag and Jeans have very similar locations in the graph.



By hovering on each data point, we can see that Handbags actually have the least purchased item value at $5.515, while it is at $5.614 for Jeans.

**7.** **Which items were bought more by each gender previously?**

****It looks like for males, jewelry, coats, and dresses are most prevalent purchases with the numbers of 3547, 3239 and 3198 respectively. Number of dresses purchased is surprising since very few or none of the males are wearing one.

For females, the most bought items are blouses, shirts, and sunglasses with the numbers of 1744, 1578 and 1525. Unlike data for males, there is nothing surprising for the most previous purchases.

**8.Which states have the highest and lowest average review rating?**

**A map of the united states

Description automatically generated**By looking at the darker color shades and legend applied to the map, we can say that the state with “TX” code (Texas) has the highest average review rating among all US States with a 3,91 rating. For the lowest average review rating, it is the state with “WV” code (West Virginia), with a 3,58 rating.

**CONCLUSION**

Throughout the project, we focused on the gender and average rating variable by comparing them with other variables. Firstly, we started with comparing the two focused variables, and saw that there is no correlation between gender, average rating, and age, which is not surprising since it is how review rating works, some people like the item and rate high, some dislike and do the opposite.

Average ratings by categories were somewhat similar, meaning that there was not much we can conclude.

Distributions of purchase amounts in USD by gender and season were quite interesting, we saw that males make purchases in more variable amounts of money, while females make it more consistently. We also saw that in almost all seasons, purchase amounts are very symmetric.

Application of discount data is quite surprising and looks missing as there is no female shopper that applies discount for her purchase. It might be really missing, but it is how the data set is.

We looked at the purchase frequencies by gender by percentage, as we do not have the close number of men and women. We concluded that gender doesn’t affect the frequency of purchase, which is not a surprise again.

Then we examined the most valuable item purchases in total and previously purchased items by gender.

Lastly, we analyzed the average ratings by each state.

Overall, average ratings are not much of an explanatory variable, while the influence of gender on other variables is much higher and can be considered explanatory variable.

**DASHBOARD LINK:** <https://public.tableau.com/app/profile/ozan.i.bi./viz/TableauProject_17018923677710/Dashboard1?publish=yes>