— Report —

Overview of Shopping Behavior Habits and Data

Data Visualization Report for STAT 112

by Burak Şahin Küçük
2023 December
METU//ODTÜ

1. Introduction: The first rule of being better is this: customers must be well-known to sell more and compete with competitors. To match this competitive environment, users' data must be collected. This report used data from user transactions from a nationwide shopping company. The data ranges from some personal characteristics to the different categories of products sold by the market. The main aim of this report is to visualize the data effectively to understand how the shopping company could be better.

1.1 Data Description: These data have 3900 observations. Also, the data have 17 variables. There are four numerical and 13 categorical data. Raw data:

Numerical Variables:	Categorical Variables:
Age (Continuous, but in this case, it is Discrete)	**Gender** (Nominal – Boolean)
Purchase.Amount.USD. (Continuous, but there is no decimal point, so for this data set, it is discrete)	**Item.Purchased** (Nominal)
Review.Rating (Continuous)	**Category** (Nominal)
Previous.Purchases (Discrete)	**Location** (Geographical)
	Size (Ordinal
	Color (Nominal)
	Season (Nominal)
	Subscription.Status (Nominal – Boolean)
	Shipping.Type (Nominal)
	Discount.Applied (Nominal – Boolean)
	Promo.Code.Used (Nominal – Boolean)
	Payment.Method (Nominal)
	Frequency.of.Purchases (Nominal)

2. Data Preprocessing: Cleaning data is one of our most important jobs. First, after loading our data into Tableau, we edited the names of the variables (removed the dots and made them more readable). Since we had more than 3000 observations, Tableau did not make the "data interpreter" feature available, so we scanned the data to see any missing data. At this stage, which was checked many times, no gaps or **null** data were found. Then click the string icon (Abc) on the Location variable in Shopping behavior data and select the state option from Geographical role. After that, we need to merge our second data (country) with our data in Tableau. First, we add our Country data by clicking Add next to connections. After dragging the Country data to the data field, it does the operation automatically, but we still checked it. It combined the Location data with Location (Location). Then changed the string value of the variable Location (Location) from geographical role and selected state.

It could remove duplicate observations in the Country data, but this is not a very necessary operation (just for this case) since they both belong to the same data and follow each other (each in the same order and has 3900 observations). Nevertheless, this could be removed either with Excel or Tableau (Calculated Field). In other words, you could create a Calculated Field for Location (Location) in the second data and use this type of code: {FIXED [Location (Location)]: MIN(Location code)}. However, it is easier to do this in Excel. After opening your data set, click on any cell. Then go to Data. Under Data Tools, click Remove Duplications. As mentioned at the beginning, this is not a very important operation **for this**

data, but duplications will be removed in the rest of the report. Three thousand nine hundred repeat states reduced to 50.

The country in the map data is entered as Turkey. This is incorrect data. The error was fixed by changing the country to "United States" so that all states started to appear. Since no operation can be done with Customer ID, it has been converted from a number to a string value.

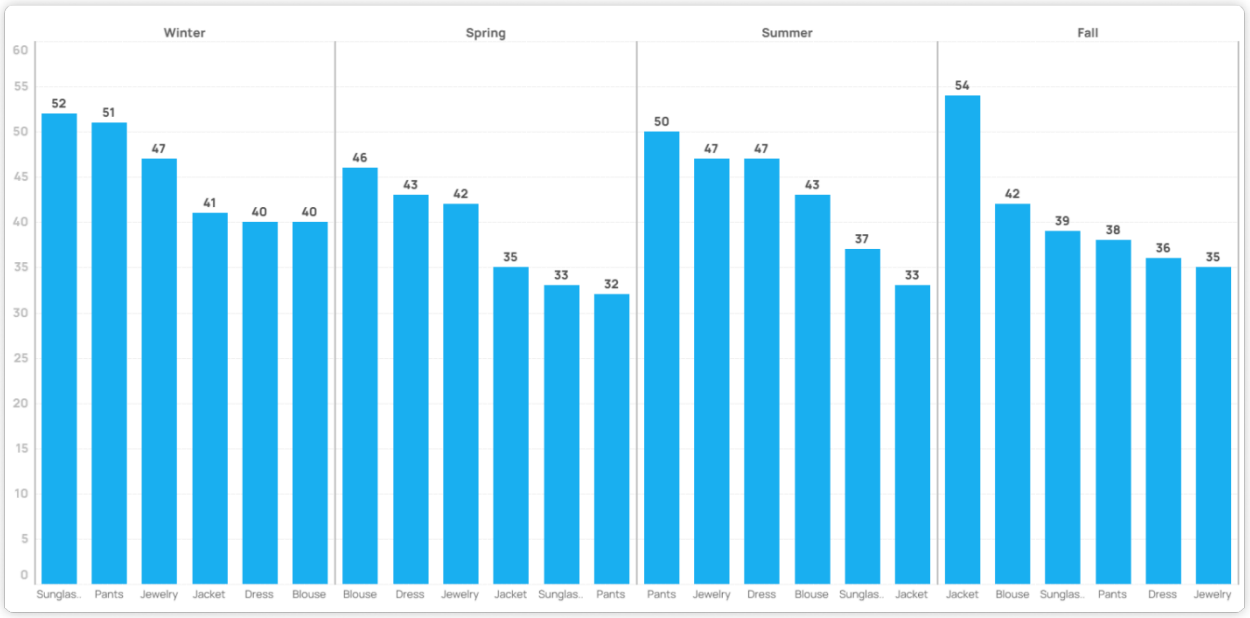
3. Exploratory Data Analysis: Before moving on to the questions and explanations, it is essential to mention some skeptical points. All questions in the dataset with YES and NO answers have only NO values for women. These variables are Subscription Status, Discount Applied, and Promo Code Used. Although the data set has been checked several times, we have determined that we did not cause this problem. As a result of these findings, gender interpretation cannot be made for the three variables mentioned above. In addition, any visualization made with these three variables will only include male members. Women cannot be included in these visualizations as they do not have any answer other than no. This may be caused by the subscription system's input on the company website providing the data, or the company may have collected the data incorrectly. In the rest of the report, care has been taken to avoid using inaccurate data like this. Since this data can be used to interpret the status of men, we did not wholly exclude the variables from the dataset but made the necessary filtering.

And finally, it should be noted. Promo-code Used variable and Discount Applied variable here has the same value for each observation. One of these variables should be removed from the dataset. Also, a product with a promo code is already in the discounted product category. For our specific research, we need to have observations where these two values are different from each other. But this dataset has the same result for all observations, so discount = promo code. So, we decided to remove the promo code variable.

3.1 Research Questions: We will analyze the company with at least seven questions and try to draw some conclusions. Since we don't have enough data on what can be done, we will only talk about possible ways forward.

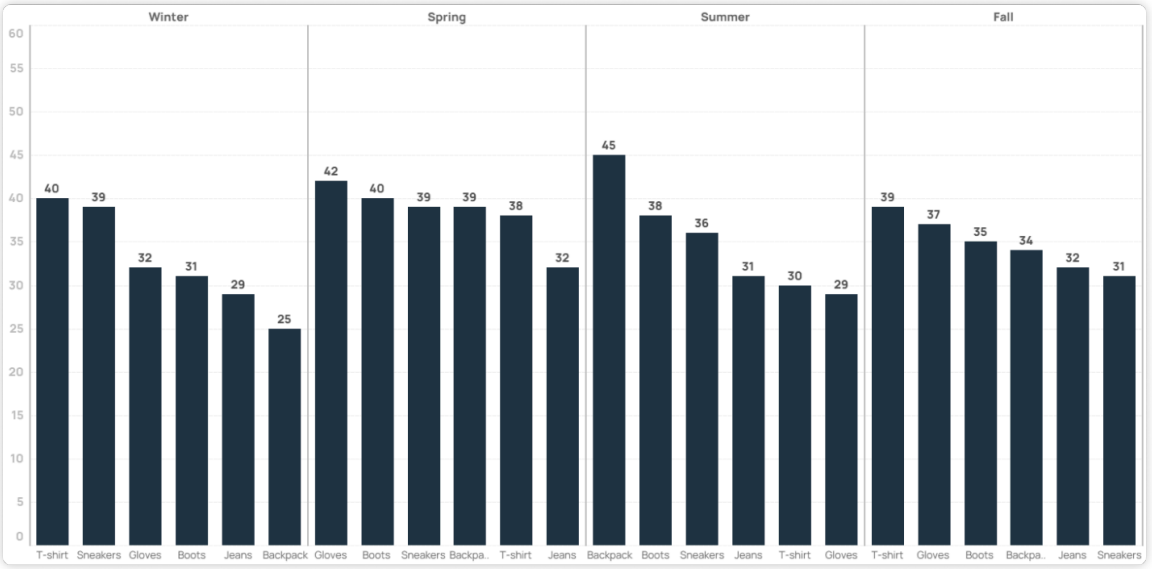
3.1.1 How does the number of sold products vary according to the season?

It is important to know which product sells better in which season. A company that sells online and offline should pay attention to this data to increase its sales. Data is used instead of intuition to find this out. If intuition is trusted, mistakes can be made. When you think about it, it is quite normal to expect that the most sold product in the **winter** season will be a **sweater,** but you can see from the bar chart that this is highly inaccurate. Visualizing this data is extremely important. A bar graph was used to answer this question. According to this, while the seasons were the main column, the category was the secondary column. Thus, two charts were obtained. Then, these graphs were filtered to give the first six and last six products.



Visualization 1: Most Purchased 6 Items by Seasons

Interpreting the Visualization 1: All these inferences are among the six best-selling products. The six best-selling products include Sunglasses, pants, jewelry, jackets, dresses, and blouses. In all four seasons, although the order of the six most sold products changes, their names do not change. Sunglasses are the best-selling product in winter. As slightly mentioned above, there are shopping habits that are beyond human foresight. Pants are the most sold product in spring. Also, Blouse sales are highest in spring and lowest in winter among the six products. Jewelry sales are decreasing every two periods. As expected, sales of jackets come last among the six products in the summer.



Visualization 2: The 4 Least Purchased Products by Seasons

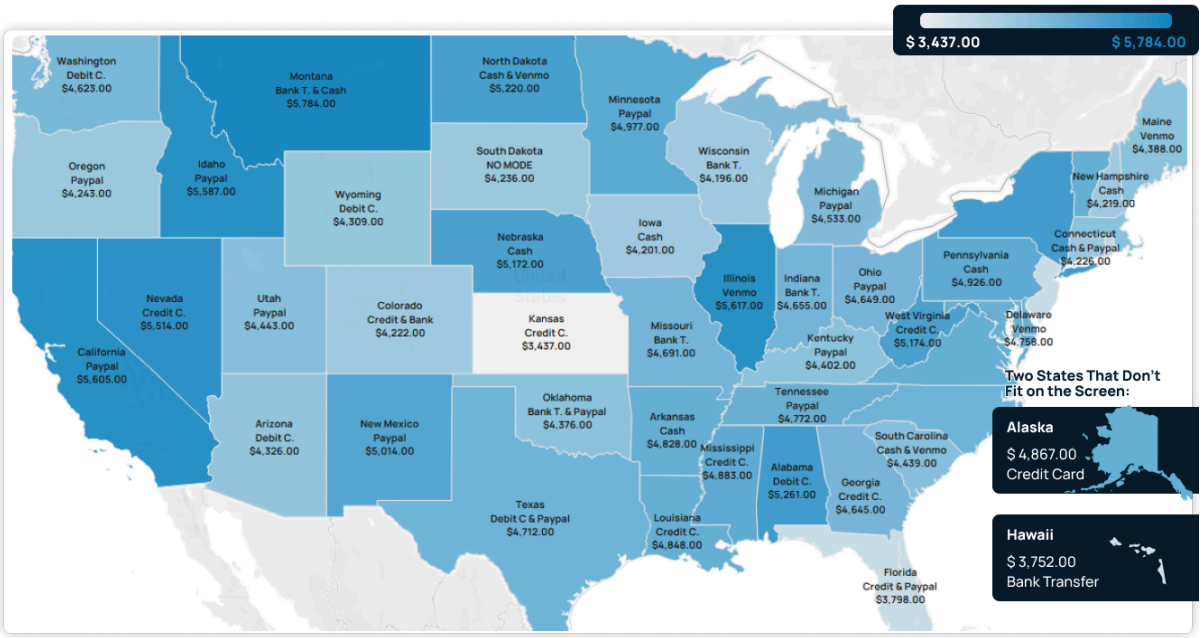
Interpreting the Visualization 2: All these inferences are about the six least sold products. The six least sold products include T-shirts, Sneakers, Gloves, Boots, Jeans, and Backpacks. In all four seasons, although the order of the four least sold products changes, their names do not change. Contrary to expectations, the gloves were purchased in spring, not

winter. And it was the best-selling product among the last six products in the spring. It seems that even writing bots haven't become the least-selling product.

Interpreting the Results: According to the visualizations above, although the order of the first six products sold and the last six products sold changes according to the seasons, no new product comes between them. The company can make period-specific offers. They can put the product that sells the least in winter next to the product that sells the most in winter at a discount. For example, a 30% discount on backpacks if you buy glasses in winter. Thus, a way to dispose of less-selling products, even to some extent, is opened. Additionally, with this data, in order to prevent inventory issues, more of the products that are in high demand during specific periods should be procured from suppliers.

3.1.2 What is the most frequently used payment method and total sales by state?

The payment method used by people may vary from region to region. Determining the payment methods that people constantly use according to some categories benefits the company in many ways. They can create regional co-operation with banks and FinTech companies and create campaigns for users. In this research question, we will look at the total amount of money coming from the states and which payment method they frequently use through the map graph created. While creating this map, since it is very difficult to find the mode of categorical data in Tableau, the most used payment methods in the states were revealed by using different tools due to the importance of this problem. To solve this research question, the map feature of Tableau and map coloring with numeric data were used.



Visualization 3: Most Used Payment Methods and Total Revenue for Each State with Tableau Map Feature

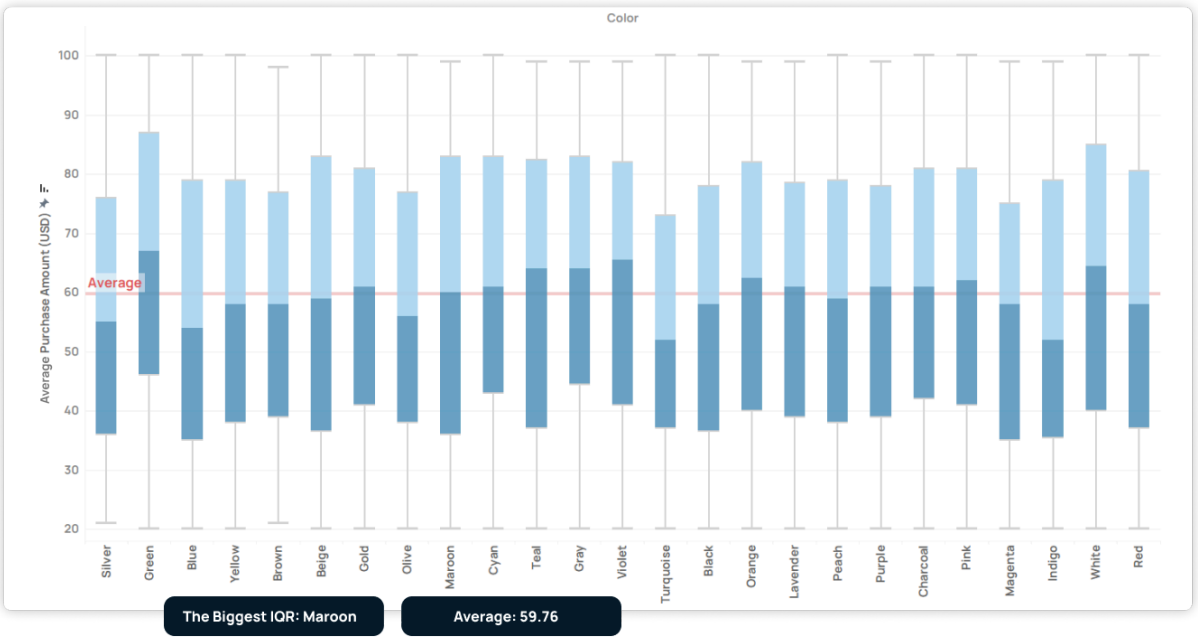
Interpreting the Visualization 3: Montana ($5,784.00), Illinois ($5,617.00), and California ($5,605.00) are the states with the most purchases from the company that provided this data. All states except South Dakota have at least one frequently used payment method. The payment methods of the states can be grouped as follows:

- **PayPal (18):** Oregon, California, Idaho, New Mexico, Minnesota, Kentucky, Tennessee, Michigan, Ohio, Rhode Islan Utah, Vermont, Texas, Florida, Connecticut, Oklahoma
- **Credit Card (14):** Nevada, Kansas, Louisiana, Mississippi, Georgia, Alaska, Maryland, Virginia, West Virginia, Florida, Colorado, New York
- **Bank Transfer (9):** Missouri, Wisconsin, Colorado, Oklahoma, Montana, Hawaii, Indiana, New York,
- **Venmo (9):** Illinois, Delaware, Maine, New Jersey, North Dakota, South Carolina
- **Debit Card (8):** Washington, Wyoming, Alabama, Arizona, North Carolina, Texas, Massachusetts,
- **Cash (14):** Nebraska, Arkansas, Iowa, North Dakota, South Carolina, Connecticut, Massachusetts, Nebraska, New Hampshire, Pennsylvania, Montana

PayPal is the most preferred payment method in most cities. The debit card payment method is the least preferred among them.

Interpreting the Results: Thanks to this map graph prepared to offer pinpoint campaigns, bilateral agreements with PayPal can be tried to be made in the states where PayPal is used the most. Thanks to this map graph prepared to offer region-specific campaigns, bilateral agreements with PayPal can be tried to be made in the states where PayPal is used the most (California, New Mexico, Kentucky, and so on). It should be remembered that strategies may change. Contrary to what is said, if the least used payment method has a high power in the region, this indicates that there are many potential customers, but they do not prefer this company. It is very significant to try to include in the system these users who do not prefer us. So, with this data, it can also be seen which payment method is used less where.

3.1.3 Does Product Color Make Customers Spend More Money?

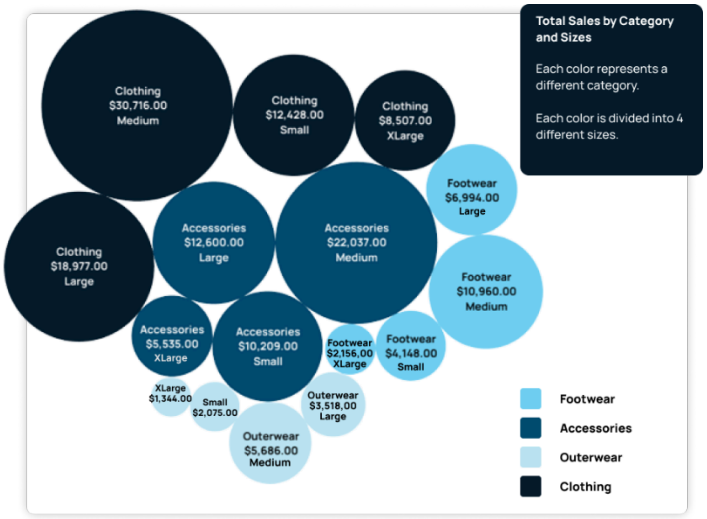


Visualization 4: Relationship Between Average Expenditure and Color (Box Plot)

A client can change the price they pay according to the color of the products. In this box plot, we see the average expenditures by color preference. The average line was also added so that

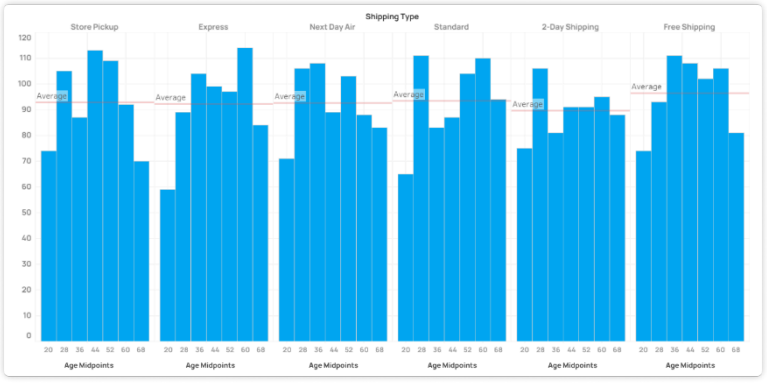
the data could be easily distinguished. This visualization becomes more meaningful as the product category and product become more specific, but we will look at this graph in general. Green, white, violet, gray, and teal were able to increase people's spending. But the spread in teal is much greater than in the others. Indigo, turquoise, blue, and olive are the colors that generally pull the average downward, meaning that the amount people spend on these colors has not increased. According to this box plot, the following comments can be made: Green, white, purple, gray, and teal are more expensive than Indigo, turquoise, blue, and olive. In the future period, it can be examined to increase the prices of blue-colored products with lower average prices. This is because the prices of products below the average remain more competitive compared to other products.

3.1.4 How are the revenues generated from sales based on category and size?



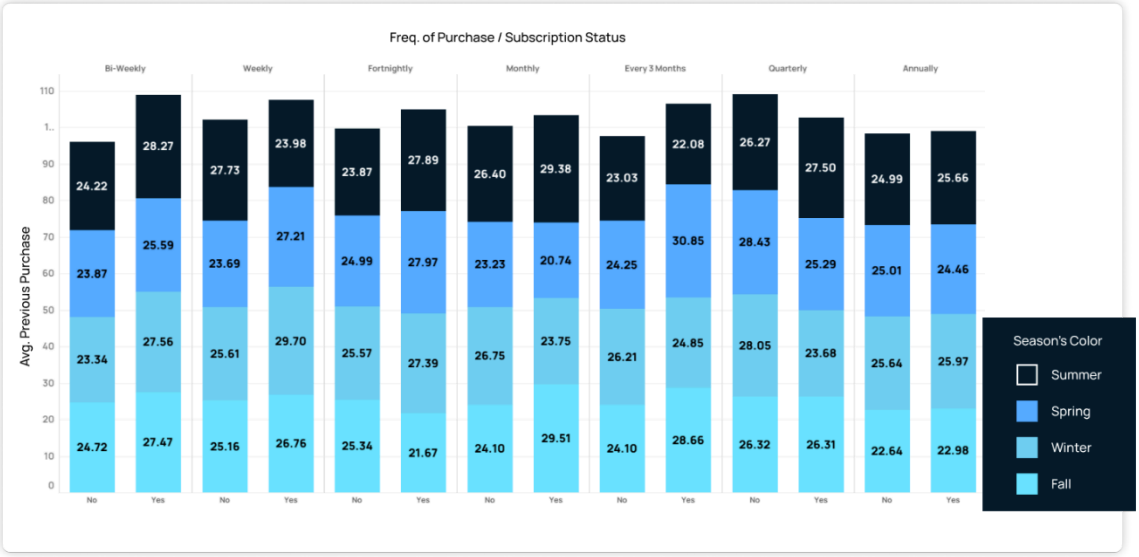
Segments were created from category and size with a bubble graph. The total sales amount of each was used for the bubbles' size. When each category was evaluated separately, the medium was the size that generated the most revenue. On the other hand, X Large was the least monetized body size. Based on this, a large proportion of customers bought medium products, while a very small number opted for X Large products. We interpret the categories; the most sales were obtained from the clothing category. On the other hand, the lowest revenue came from the outerwear category. Based on the findings, the number of stocks of medium products should be well controlled. Some special offers can be created. For example, someone who buys products from the clothing category can be given a discount coupon to use on some of the outerwear products. In this way, more income can be generated from low-income categories.

3.1.5 What does the distribution of age histograms by shipping type signify?



I dragged the shipping type to the column and created an age histogram. The age range of this age histogram is 16-72 and increases by 8. People were then dispersed in these intervals. This graph shows us which shipping type is used by people in a certain age range. This histogram illustrates the preferred shipping methods among individuals within a specific age range. According to the histogram, older people want products to arrive quickly. This is reflected in the express service. Young people, on the other hand, do not have such a tendency compared to older people. Free shipping almost has a bell-shaped distribution. Older people, as well as middle-aged people prefer express delivery and free shipping. Young people are among those who prefer 2-day shipping the most. We see this more clearly when young people are evaluated among themselves. Likewise, when the elderly are evaluated among themselves, it seems clearer that they prefer express delivery. From all this, we might surprise the older people. For example, if we send express shipping to an elderly person who has no money on their standard delivery orders several times and cut it off, they will probably start paying express shipping.

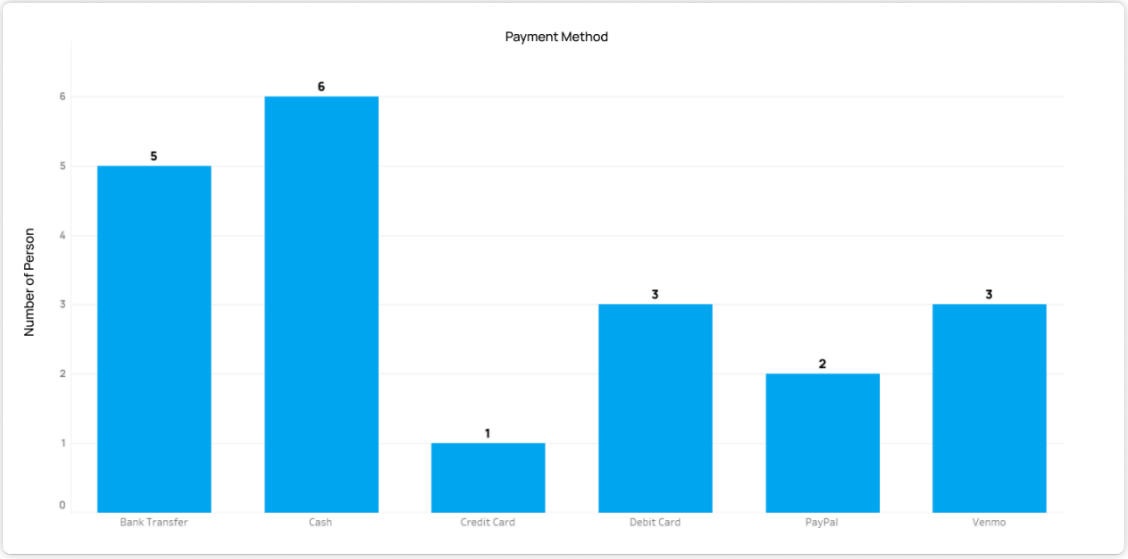
3.1.6 How does the subscription status among men change with shopping frequency based on past order numbers?



Visualization 7: Stacked Bar Chart – (Freq. of Purchase / Subscription) / Avg. Previous Purchase

The stacked bar was created to answer this question. Average orders per person made in the past were organized by subscription and shopping frequency. What is clear from this data is that in all periods except quarterly, male subscribers made more purchases in this company. Based on this, it can be interpreted that subscribed users are more loyal to the company. The biggest difference between subscribers and non-subscribers is seen in bi-weekly. On the other hand, there is not much difference in the past purchases of people whose purchase frequency is annual. What can be said in general is that as the frequency of using the shopping company increases, the loyalty of subscribers increases. And the difference between them is also greater when the frequency of using the shopping company increases. Customers who are most loyal to the shopping company use the spring period every three months.

3.1.7 What payment methods do people using medium-sized clothing in Montana prefer?



A filtered bar chart was created to find out which payment method was used by people buying clothes in medium sizes in Montana. According to this graph, most people prefer to pay with cash, followed by bank transfers. The least used payment method is the credit card payment option. Based on this, people in this stratified sample do not like the option of paying by credit and do not want to go into debt for clothes. There is also no payment method that is never used.

3.1.8 If the company were to downsize in Indiana, which products should it first consider discontinuing?

Location

Item Pur..	Alaba..	Alaska	Arizona	Arkans..	Califor..	Colora..	Conne..	Delaw..	Florida	Georgia	Hawaii	Idaho	Illinoi	Indi..	Iowa	Kansas	Kentuc..	Louisi..	Maine
Socks	187	297	183	121	83		122	400	67	416	63	125	19	336	69	268	29	202	376
Belt	53	307	169	220	185	137	143	166	94	62	132	488	33	317	298	149	128	260	34
Skirt	412	262	274	89	224	206	216	132	267	99	225	193	9	315	21		86	60	111
Jacket	110	135	260	54	193	320	131	52	145		388	138	54	305	316	111	235	99	227
Sweater	81	151	352	286	168	174		105	168	94	216	193	37	284	451		447	267	396
Jeans	265	143	119	226	522		67	156	152	39	109	98	20	275	138		82	199	173
Sandals	208	257	227	232	266	63	125	310	49	140	21	89	12	270	245	342	63	274	128
Shorts	147	230	92	285	380	95	238	86	179	59	103	376	17	248	191	170	131	96	147
Shirt	241	238	160	295	214	258	187	136	199	136	144	249	60	227	68	174	359	388	76
Hat	228	158	87	70	147	184	22	287	324	141	215	169	13	202	390	136	224	163	161
Pants	253		158	274	286	325	235	442	95	317	202	226	18	200	36	144	154	231	180
T-shirt	105	227	124	104	207	118	117	200	168	100	163	262	15	198	132	169		297	152
Gloves	170	199	87	386	106	258	126	361	310	62	61	207	29	187	175	50	176	94	123
Boots	221	56		50	250		257	89	103	85	148	181	24	187	148	287	138	55	278
Scarf	226	135	107	135	218	136	273	216	117	301		72	39	179	58	113	351	81	203
Shoes	326	160	87	238		270	207	291	26	300	81	201	14	177	291		21	153	425
Jewelry	464	206	203	262	294	181	286	164	135	343	67	247	17	176	270	309	289	328	52
Coat	327	91	196	217	165	46	267	115	307	93	209	522	4	146	122	186	353	44	177
Sunglass..	188	249	198	112	347	68	43	91	59	209	124	191		109	36	184	134	75	221
Blouse	208	270	183	198	79	110	337	211	31	454	253	325	22	100	119	293	232	135	293
Backpack	74	373	243	270	287	285	120	99	59	179	96	213	18	89	102	20	214	270	
Handbag	173	92	176	100	261	205	321	102	134	92	322	182	33	69	161		382	224	136
Sneakers	159	220	157	69	146	321		79	256	302	115	215	14	36	179	125		166	182
Hoodie	159	113	200	221	168	303	126	238	257	64	253	330	11	23		69		503	56
Dress	276	298	284	314	409	159	260	230	97	558	42	95	18		185	138	174	184	81

First, we created a density chart. We put products in rows and states in columns. After that, dark blue is the color that makes the most money, and white is the color that makes no

money. Indiana is one of the least profitable states, so the company wants to leave slowly. In this process, the products that should be discontinued first are the ones generating the least revenue. The company never sold a dress in Indiana. Therefore, there is no need to ship this product. Additionally, hoodies, sneakers, handbags, and backpacks haven't been profitable for the company in Indiana. Therefore, there is no need to keep these products in storage or proceed with their shipment. The products that should not stop shipping in the first phase are socks, belts and skirts.

5. Conclusion and Discussion: Many questions were answered in these eight carefully selected questions. This report explored various aspects of shopping behavior using data from a nationwide shopping company. The analysis covered seasonal product preferences, payment method trends, the impact of product color on spending, revenue generation based on category and size, age-related shipping preferences, the relationship between subscription status and shopping frequency among men, and also, payment preferences for medium-sized clothing purchases in Montana.

The insights gained from these analyses can guide the company in optimizing its strategies for marketing, pricing, inventory management, and customer loyalty with subscription. Understanding seasonal trends allows for targeted promotions, while insights into payment preferences and shipping choices enable the company to tailor its services to meet customer expectations.

The findings also highlight opportunities for cross-selling and pricing adjustments based on product color. Effective stock management strategies, especially for medium-sized products, can contribute to maximizing revenue. The analysis of subscription status and shopping frequency emphasizes the importance of customer loyalty.

Overall, this report provides actionable insights for the company to enhance customer experience, increase sales, and stay competitive. The detailed visualizations and observations offer a roadmap for strategic decision-making in the dynamic retail landscape.

Dashboard Link: Before sharing the link, we would like to state that we believe that dashboards should be plain and simple. For this reason, the charts used in the report have been changed to be more descriptive. Also, two dashboards were prepared. In other words, a separate layout was created for phones, while one of them works only for the computer version. The computer version is the recommended dashboard. It provides more functionality and filtering options. On the other hand, the phone version can only handle some graphics and some integrations. If you see "you are in phone mode" as a note when you open the link below, please try the computer version as well.

- Main dashboard: https://public.tableau.com/app/profile/burak.k.k/viz/OverviewofShoppingBehaviourHabitsandDataFinalRevised1_0/OverviewofShoppingBehaviourHabitsandData#1

Note: This report and dashboard were started to be prepared 18 days before the first submission date. Unfortunately, there is no possibility of changing anything in the data after submission. Used Tools: Figma, Tableau Public, Tableau Desktop, Word, Grammarly, Deepl
Color Palette: *McKinsey New Look* Fonts: Open Sauce Sans, Manrope, Bower

Submission Date: 12.04.2023

References

McKinsey & Company. (2019, February 24). *Meet the new McKinsey look*. https://www.mckinsey.com/about-us/new-at-mckinsey-blog/meet-the-new-mckinsey-look