Product's Color Can Affect the Product's Price and Customer Behaviors


(Burchfield, 2017)

ZAID ALTUKHI
DR. CHARLES LYNCH

George Mason University
Fall 2021
AIT 664-001
Project Recap & Lessons Learned

November 28, 2021

Table of Contents

Introduction

One of the essential questions that appear when a product owner wants to launch a product or who has markets and wants to buy products to sell is the suitable color that could be used or bought? Can product colors determine the product cost and the customer's decisions? Anyone who works in the marketing or sale department can tell how product color is an essential factor affecting customer behavior. However, some details need some analytics to be answered, such as, if the product has more than one color, will this affect the customer's actions? What are the most men, women, kid's products colors? If there is a product with several colors, will this influence customer reviews? Analyzing how colors can affect consumers' behavior will provide a good idea to those seeking to sell many products and gain profits. Although the designers working on creating the product details know how important to pick the right color for each piece of the product, knowing how the color can influence the product price and people's desires will help the decision-maker make the right decisions.

This project will try to answer this kind of question by examining the hypothesis that says, "product's color can affect the product's price and customer behaviors." In addition, answering these questions needs to consider many factors, such as the target audiences and the product nature.

The following paragraphs will show the data used, project requirements, data acquisition and sources, data analytics, and data visualizations.

Project Requirements

After the hypothesis was determined, the searching for datasets started by looking for the pertinent data to use in multiple data sources that provide the datasets we can analyze. Since the hypothesis focuses on the product's characteristics, the dataset we are looking for a must have list of products with specific kinds of features that can examine to test the hypothesis. To select the dataset, it must contain at least 10,000 products with different categories, colors, prices, and rates done by customers. Moreover, the data type needs to be categorical, numerical, and nominal.

On the other hand, finding the proper analytical tools is essential for the analysis operation. To do this analysis, the tools used are Python and Tableau. Python can be used to explore, clean, prepare and visualize. In addition, it can be used to apply statistical models to find if there is a relationship between the features. At the same time, Tableau can visualize and create dashboards that show the trends in the data, answer the primary questions, and find patterns and insights. Using python and Tableau can help determine if the hypothesis we test is right or wrong using different data analytics methods.

Data Acquisition and Data Sources

Many data sources provide this kind of data—for example, Kaggle.com and data.world.com. After looking at many datasets were provided by those websites. Most of the data were available in CSV files, common file types used in the data analytics field. However, most of these datasets have broken at least one of our data requirements until datasets found in data.world.com fulfill the requirements.

The datasets were found in data.world.com have more than four datasets that show the list of products from an e-commerce website called NewChic.com (Mabilama, J.M.,2020). Each dataset contains around 11,000 records. Also, I found that the data provided are proper to examine the hypothesis. For example, product categories, product name, price, discount percentage, number of likes the product gets, and product colors.

Datasets Overview

There are five datasets that will be used in this project. Each dataset contains specific kinds of products, clothes, bags, and shoes. The datasets were used in this project as follow:

1- Women products
2- Men products
3- Children products
4- Shoes
5- bags

After exploring the data, some variables were found to work on. For example

1- raw price: original product price
2- current price: the price after discount applied
3- colors: there are some products that contain only primary colors, and others have two main colors
4- Likes count: number of likes that product got from customers.
5- Discount: the discount percentages per product.

In addition, the datasets contain different data types as follow:

- categorical features: main product category, product sub-category, primary color, secondary color
- numerical features: product price, product discount percentage, number of likes
- nominal features: product name, product IDs

Although most of the data are in good shape and quality (defined, measurable, unitized, relatable, normalized, and quality). However, there is a country column with multiple values that need to handle, and the color columns have many different forms for the same color.

Prepare Data

Many tools are used to prepare and clean data: Microsoft Visual Studio Code (Visual Studio Code, 2019), python programming language (van Rossum & Drake Jr, 1995). The Python language supports powerful libraries that help the developers perform the analytics processes. This project uses multiple software and packages to complete the data cleaning, preparing, and analyzing the data:

- Anaconda (Anaconda Software Distribution ,2020)
- Pandas (The pandas development team, 2020)
- Numpy (Harris et al., 2020)
- Sklearn (Varoquaux et al., 2015)
- SciPy (Virtanen et al., 2020)

Each dataset contains 22 columns. All datasets were concatenated together to prepare the datasets to perform analyzing models and visualizations. The final concatenated data set consists of 47,193 records and 22 columns, as shown in **Error! Reference source not found.**.

```
Int64Index: 47193 entries, 0 to 11822
Data columns (total 22 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   category              47193 non-null  object
 1   subcategory           47193 non-null  object
 2   name                  47193 non-null  object
 3   current_price         47193 non-null  float64
 4   raw_price             47193 non-null  float64
 5   currency              47193 non-null  object
 6   discount              47193 non-null  int64
 7   likes_count           47193 non-null  int64
 8   is_new                47193 non-null  bool
 9   brand                 9018 non-null   object
 10  brand_url             5608 non-null   object
 11  codCountry            41389 non-null  object
 12  variation_0_color     46752 non-null  object
 13  variation_1_color     38535 non-null  object
 14  variation_0_thumbnail 46752 non-null  object
 15  variation_0_image     46752 non-null  object
 16  variation_1_thumbnail 38535 non-null  object
 17  variation_1_image     38535 non-null  object
 18  image_url             47193 non-null  object
 19  url                   47193 non-null  object
 20  id                    47193 non-null  int64
 21  model                 47193 non-null  object
dtypes: bool(1), float64(2), int64(3), object(16)
```

Figure 1: New Dataset after concatenated [Python]

Feature Engineering

The feature engineering process was done on the data by dropping columns, processing null and unified values, and adding a new column. There are 10 features in the dataset have no relation to our project: currency, brand_url, is_new variation_0_thumbnail, variation_0_image, variation_1_thumbnail, variation_1_image, image_url, url, brand, and codCountry; those columns were removed.

The color columns need to be cleaned. The main color column ‘variation_0_color’ has 559 null values. In addition, after exploring the datasets, some color values could not be used shown in Figure 2 and, these

values were also removed. Moreover, the column Variation_0_color and Variation_1_color contain 433, 460 unique values respectively. Some of these values have typos or were written in different formats, such as the red color written in various formats (RED, Red, red, red1). This issue needs to be fixed because it may affect the analyzing results. Similar colors are aggregated in one color to fix this issue the color columns. For example, there are around 60 values that indicate the brown color, and 21 colors belong to the black color. Moreover, a two_color column was added to indicate whether the items have more than one color, so we need a Boolean column to identify this information.

```
'Couleur de sable de haricot' 'Rouge + Noir' 'la grille' '09'
'Blanc + Violet' 'Blanc 2' 'Gris + Noir' 'Bleu royal' '2 #' '5'
'Noir kaki' 'Navet blanc' 'Rose vif' 'Camouflage Blanc'
'Noir + blanc + rouge' 'Blanc bleu' '23' '# 4' 'Pêche' 'Recoloriée'
'Gris Rose' 'Marron clair' '\xa0Frontière colorée' 'Rouge et gris'
'Rayures bleues' 'Marron profond' 'Café profond' 'Large rhombic gold'
'\xa0Bordure bleue' 'As shown' 'Fleurs' 'Marron 1' 'Rouge-marron'
'Violet 1' 'Arbre' 'Éléphant' 'Coffee1' 'Ombre' 'Beige stitching'
'Noir 3' 'lattice' 'Graffiti' '04' 'Red1' 'Fer à repasser' 'Shell'
```

Figure 2: Sample of the values located in the color columns [Python]

The shape of cleaned and final dataset is 45,425 record and 12 columns as follow:

No.	Column Name	Data Type	Description
1	category	String	Shows the item's category. There are five different categories: women, men, kids, bags, and shoes).
2	subcategory	String	Under each category there are different sub-categories. The dataset contains 178 subcategories.
3	name	String	Items name.
4	Current_price	Integer	Items price after discount.
5	Raw_price	Integer	Items price before discount.
6	Discount	Integer	Discount percentage.
7	Likes_count	Integer	Number of likes the item got from customers.
8	Variation_0_color	String	Item main color.
9	Variation_1_color	String	Item secondary color.
10	Id	String	Item id, as primary key.
11	model	string	Item model.
12	Two_colors	Boolean	Indicates if the item has only one color or more.

Table 1: Final dataset features description

There are five operations performed on the datasets. First, aggregate the datasets into one data frame. Second, drop the unrelated columns and not understandable values. The third process is to unify the columns that have the item's color. The number of unique color values after unified be 12 unique values. Lastly, add a new column to indicate if the item has more than one color or not. Figure 3 shows the final dataset information after clean and prepare processes were performed.

```
Int64Index: 45425 entries, 0 to 11822
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   category          45425 non-null  int64
 1   subcategory       45425 non-null  int64
 2   name              45425 non-null  object
 3   current_price     45425 non-null  float64
 4   raw_price         45425 non-null  float64
 5   discount          45425 non-null  int64
 6   likes_count       45425 non-null  int64
 7   variation_0_color 45425 non-null  int64
 8   variation_1_color 37383 non-null  float64
 9   id                45425 non-null  int64
 10  model             45425 non-null  int64
 11  two_colors        45425 non-null  bool
dtypes: bool(1), float64(3), int64(7), object(1)
```

Figure 3: Final dataset shape [Python]

```
fdata.head(10)
✓ 1.6s                                                                                                                    Python
```

	category	subcategory	name	current_price	raw_price	discount	likes_count	variation_0_color	variation_1_color	id	model	two_colors
0	women	T-shirts	T-shirt boutonné à manches courtes et imprimé ...	23.99	46.99	49	313	White	Blue	1690810	SKUF40137	True
1	women	Soutiens-gorge	Plus Soutiens-gorge avec fermeture à l'avant	15.99	40.36	60	4603	White	Gray	1533303	SKUC91583	True
2	women	Pantalons & Shorts	Pantalon décontracté à taille élastique de cou...	25.99	50.99	49	5564	Brown	Black	1661710	SKUE94621	True
3	women	Robes imprimées	Robe midi décontractée à imprimé floral	23.99	46.99	49	1262	Brown	White	1691484	SKUF41372	True
4	women	T-shirts	T-shirt brodé de fleurs	15.99	38.88	59	4485	Green	Blue	1655044	SKUE83526	True
5	women	Robes imprimées	Robe longue à imprimé floral vintage	33.99	70.99	52	2419	Red	Yellow	1527134	SKUC88481	True
6	women	Combinaisons & Grenouillères	Combinaison à bretelles à imprimé floral	32.99	64.98	49	33	Red	Blue	1715194	SKUF79731	True
7	women	Combinaisons & Grenouillères	Combinaison à bretelles à imprimé floral	29.99	63.98	53	4132	Black	Brown	1660117	SKUE91643	True
8	women	Pantalons & Shorts	Pantalon décontracté à fleurs	21.99	47.99	54	3095	Blue	Yellow	1663956	SKUE99093	True
9	women	Pantalons & Shorts	Pantalon décontracté à cordon de couleur unie	23.51	45.99	49	4367	Gray	Green	1617649	SKUE10410	True

Figure 4: First 10 rows from the final dataset [Python]

Labeling the data is a useful technique to analyze the categorical data. The variables need to convert from string to numerical values. The statistical models cannot understand the string values. Five columns were converted: category, subcategory, variation_0_color, variation_1_color, and model. The table below shows the categorical values with the numerical ones.

No	Color Name	Color number
1	Black	0
2	Blue	1
3	Brown	2
4	Gray	3
5	Green	4
6	Orange	5
7	Pink	6
8	Purple	7
9	Red	8
10	White	9
11	Yellow	10
12	other	11

Table 2: color code in the labeled data

```
data1.head(10)
✓ 0.4s                                                                                                                                                    Python
```

	category	subcategory	name	current_price	raw_price	discount	likes_count	variation_0_color	variation_1_color	id	model	two_colors
0	4	153	T-shirt boutonné à manches courtes et imprimé ...	23.99	46.99	49	313	10	12.0	1690810	41192	True
1	4	140	Plus Soutiens-gorge avec fermeture à l'avant	15.99	40.36	60	4603	10	1.0	1533303	24854	True
2	4	89	Pantalon décontracté à taille élastique de cou...	25.99	50.99	49	5564	2	11.0	1661710	36933	True
3	4	109	Robe midi décontractée à imprimé floral	23.99	46.99	49	1262	2	8.0	1691484	41386	True
4	4	153	T-shirt brodé de fleurs	15.99	38.88	59	4485	4	12.0	1655044	36075	True
5	4	109	Robe longue à imprimé floral vintage	33.99	70.99	52	2419	9	9.0	1527134	24704	True
6	4	41	Combinaison à bretelles à imprimé floral	32.99	64.98	49	33	9	12.0	1715194	44737	True
7	4	41	Combinaison à bretelles à imprimé floral	29.99	63.98	53	4132	0	0.0	1660117	36709	True
8	4	89	Pantalon décontracté à fleurs	21.99	47.99	54	3095	1	9.0	1663956	37332	True
9	4	89	Pantalon décontracté à cordon de couleur unie	23.51	45.99	49	4367	3	2.0	1617649	31210	True

Figure 5: Sample of the data after has been labeled [Python]

Conduct Exploratory Analysis

We need to find the relationships between different features to understand the data. Correlations are useful statistical models used to find the relationships between the features, either positive or negative, and how this relationship is strong (Wilson & Wilson, n.d.). Figure 6shows the correlations between the different variables in the dataset. It can be noticed that there are many strong correlations between most variables. According to Wilson and Wilson, the strong correlations are between -1.0 to -0.5 or 1.0 to 0.5, and the correlations -0.5 to -0.3 or 0.3 to 0.5 are considered a moderate correlation. The variation_0_color groups the correlations results shown in Figure 6 from the final dataset as an independent variable.

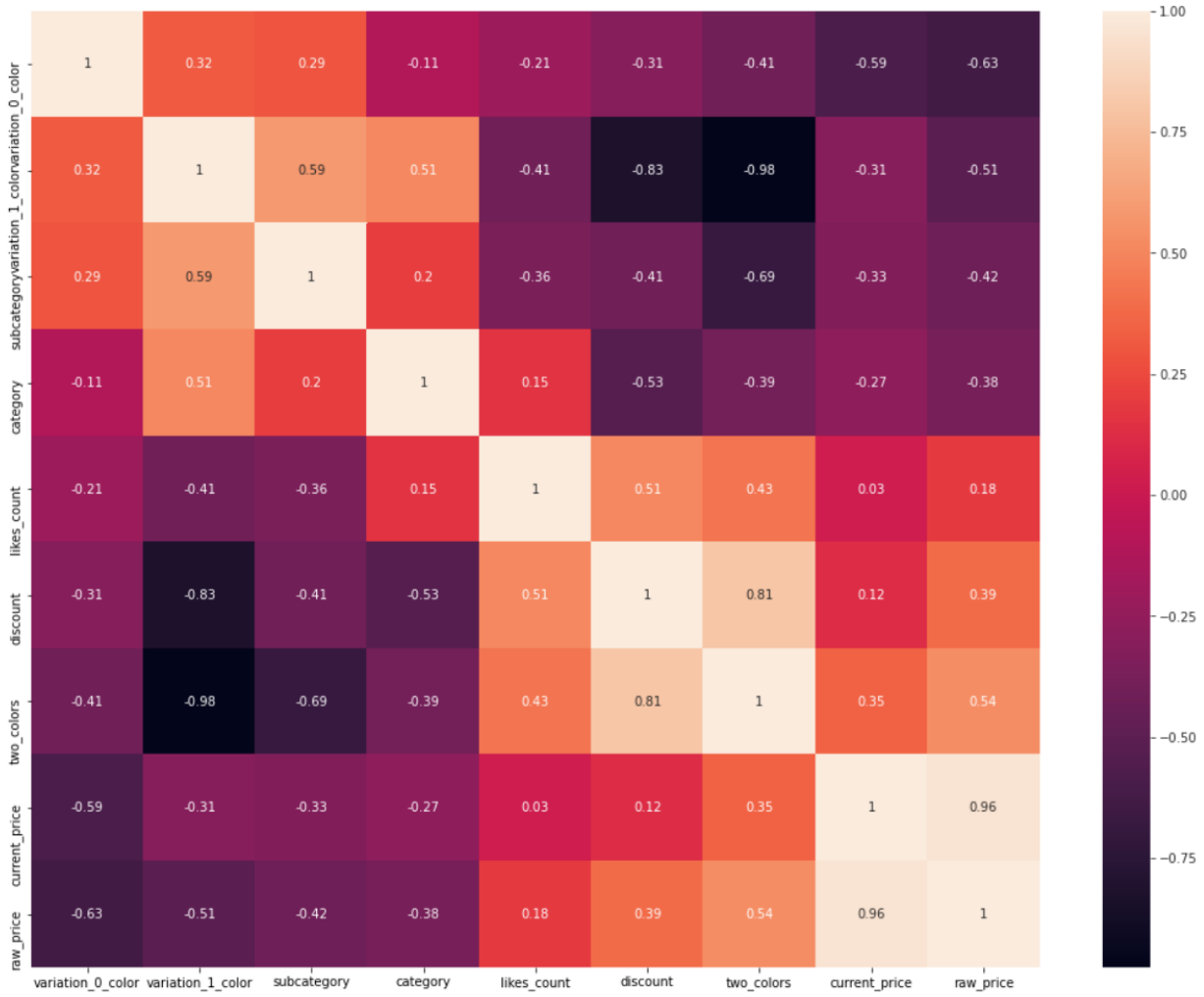

Figure 6: Data correlation [Python]

Figure 7 and Figure 8 show the values in a statistical format. This kind of data can provide useful information about how the data are distributed and give an idea.

	current_price	raw_price	discount	likes_count	id
count	45425.000000	45425.000000	45425.000000	45425.000000	4.542500e+04
mean	28.738523	60.719223	52.222609	224.296599	1.465872e+06
std	16.025378	39.936230	10.408630	631.966251	2.045275e+05
min	0.140000	0.000000	0.000000	0.000000	2.792800e+04
25%	18.040000	39.270000	47.000000	29.000000	1.311385e+06
50%	24.990000	53.040000	52.000000	75.000000	1.506630e+06
75%	35.690000	73.990000	59.000000	189.000000	1.657185e+06
max	314.590000	5089.000000	100.000000	21547.000000	1.724666e+06

Figure 7: Final dataset quantitative data description [Python]

	category	subcategory	current_price	raw_price	discount	likes_count	variation_0_color	variation_1_color	id	model
count	45425.000000	45425.000000	45425.000000	45425.000000	45425.000000	45425.000000	45425.000000	37383.000000	4.542500e+04	45425.000000
mean	2.561431	87.277248	28.738523	60.719223	52.222609	224.296599	4.512669	3.961747	1.465872e+06	22648.031657
std	1.346878	48.888451	16.025378	39.936230	10.408630	631.966251	4.128551	3.666722	2.045275e+05	13063.923894
min	0.000000	0.000000	0.140000	0.000000	0.000000	0.000000	0.000000	0.000000	2.792800e+04	0.000000
25%	2.000000	42.000000	18.040000	39.270000	47.000000	29.000000	0.000000	1.000000	1.311385e+06	11341.000000
50%	3.000000	99.000000	24.990000	53.040000	52.000000	75.000000	3.000000	3.000000	1.506630e+06	22649.000000
75%	4.000000	128.000000	35.690000	73.990000	59.000000	189.000000	9.000000	7.000000	1.657185e+06	33976.000000
max	4.000000	176.000000	314.590000	5089.000000	100.000000	21547.000000	12.000000	12.000000	1.724666e+06	45257.000000

Figure 8:Labeled dataset description [python]

Exploratory Visualizations

In order to explore the data, we need to create some visualizations to understand the values. The first chart shows the different colors with the number of items in Figure 9. The second graph represents the number of items for each category Figure 10. Lastly, Figure 11 shows the current price, which is the selling price per primary color. All these visualizations give us an idea of how the information is distributed in the dataset, allowing us to understand the data content profoundly. By the way, all the visualizations shown in this section were produced using the Altair library (VanderPlas et al., 2018).

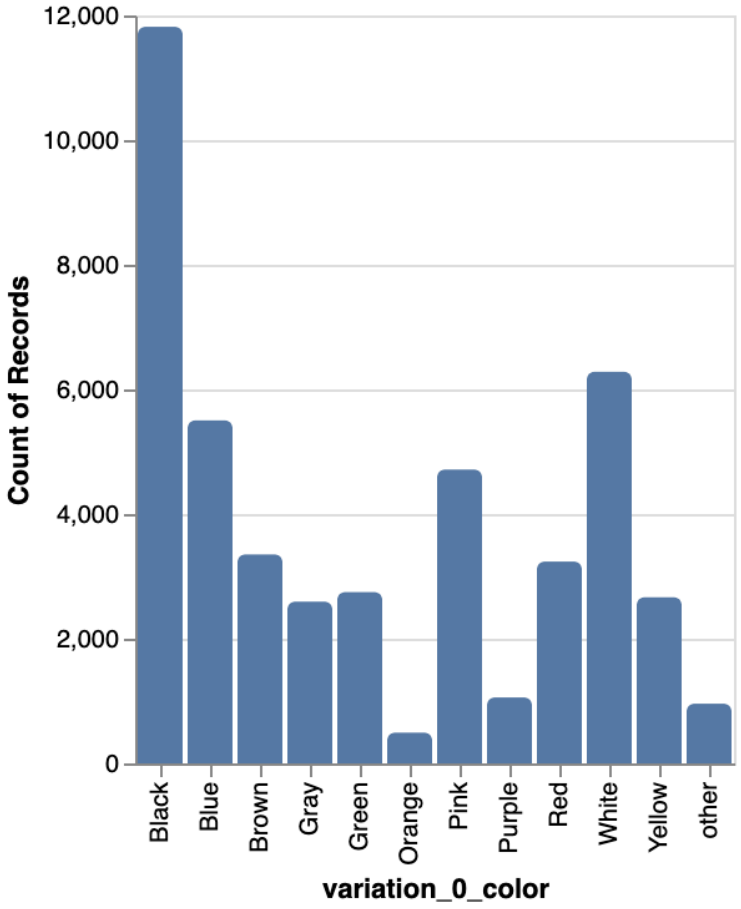

Figure 9: number of items pre color [Python by Altair]

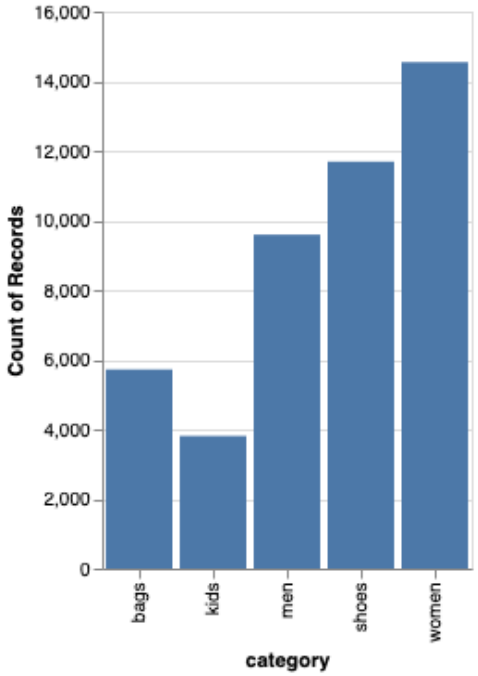

Figure 10: number of items pre category [Python by Altair]

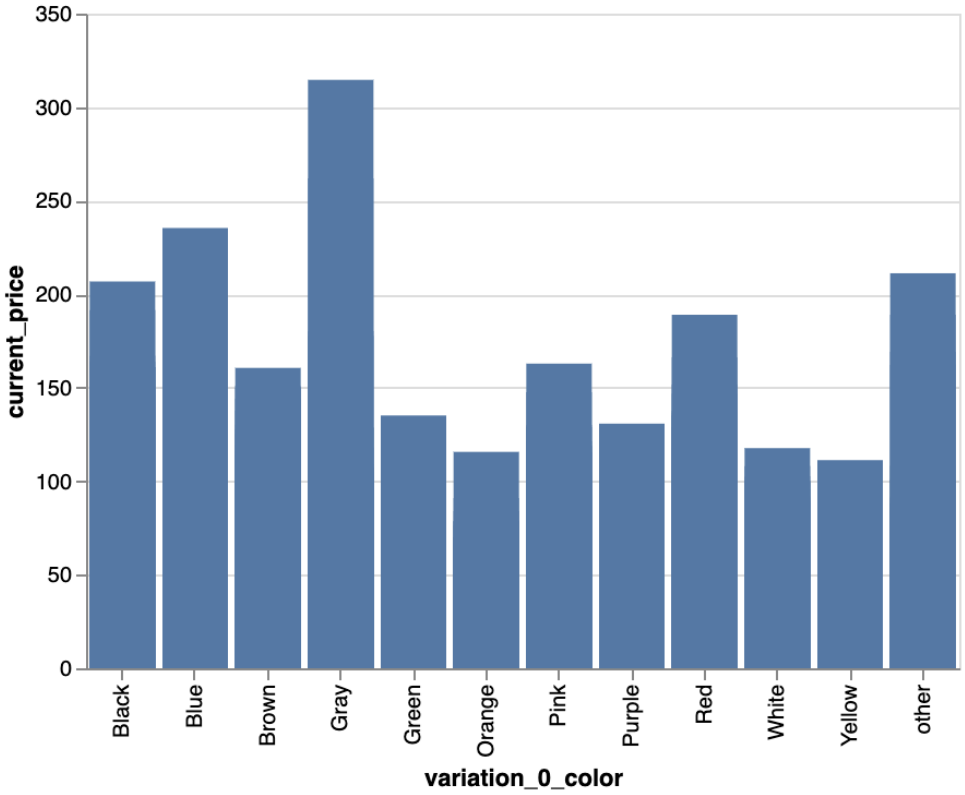

Figure 11: sum of the current price per color [Python by Altair]

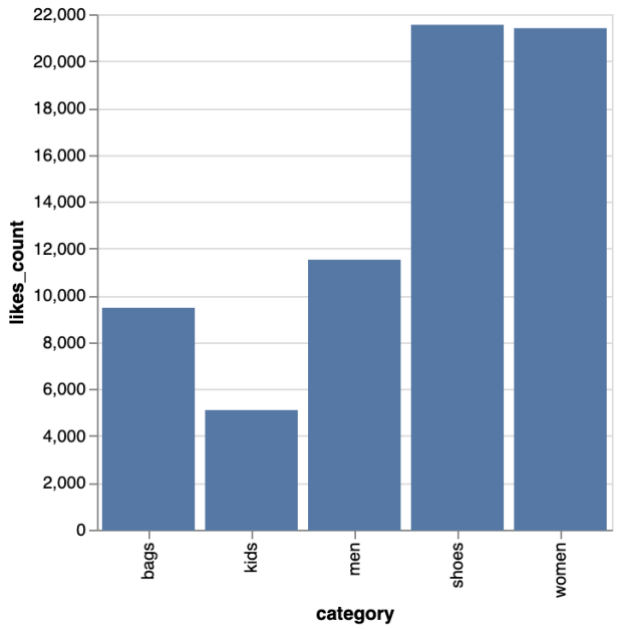

Figure 12: sum of the like counts per category [Python by Altair]

Exploratory statistical model

Confidence Intervals

A confidence interval (CI) is a set of numbers expected to encompass a population figure with a high degree of certainty. When a population means falls between two intervals, it is commonly stated as a percentage (Mcleod, 2019). The table below represents the different confidence interval values for the different colors with likes count, current price, raw price, and discount percentages. Also, we highlighted the highest values for each variable.

color	Likes_count			Current_price			Raw_price			discount		
	mean	ci95_hi	ci95_lo	mean	ci95_hi	ci95_lo	mean	ci95_hi	ci95_lo	mean	ci95_hi	ci95_lo
Black	193.04	202.86	183.22	31.83	32.14	31.52	66.74	67.77	65.71	51.82	52.00	51.63
Blue	253.53	273.21	233.86	28.03	28.42	27.64	59.79	60.63	58.95	52.42	52.69	52.15
Brown	222.44	242.65	202.24	33.44	34.09	32.79	70.78	72.06	69.51	52.79	53.14	52.43
Gray	221.10	241.25	200.95	29.81	30.49	29.12	61.96	63.28	60.64	51.59	52.01	51.17
Green	259.10	287.87	230.33	26.99	27.50	26.47	57.47	58.55	56.38	52.26	52.62	51.91
Orange	284.63	359.77	209.50	26.64	27.84	25.44	57.56	60.28	54.83	52.61	53.46	51.75
Pink	272.48	294.59	250.37	25.54	25.95	25.12	55.95	56.78	55.12	54.29	54.59	53.99
Purple	188.42	218.01	158.83	25.20	26.10	24.31	56.19	58.25	54.12	54.54	55.16	53.92
Red	231.14	253.65	208.63	31.12	31.71	30.53	66.56	67.81	65.32	52.52	52.87	52.17
White	201.80	214.51	189.09	24.09	24.38	23.79	49.61	50.17	49.05	51.22	51.48	50.95
Yellow	255.93	282.78	229.07	28.27	28.83	27.72	59.94	61.10	58.79	52.35	52.73	51.97
other	165.16	193.95	136.37	24.82	25.89	23.76	47.93	49.90	45.96	48.00	48.64	47.36

Table 3: Confidence Intervals summary

Modeling & Algorithms

From Figure 6, we can see the how relations between the main color column and others features using the correlation analysis. It can be seen that there are many strongest positive correlations between the primary color and multiple variables. However, the most substantial negative relationship is -0.52 between the main color column and the raw price, which is the price before the discount. Also, it can be noticed that the strongest correlations for the number of likes and the products with two colors is 0.57.

One-Way-ANOVA

After looking the strongest relations ether positive or negative, it can perform the statistical models to look deeply into the data. In this project we will examine the data using One-Way-ANOVA because we test 13 groups of data. The library was using to perform the model is statsmodels (Seabold & Perktold, 2010).

Independent column	Dependent column	sum_sq	F	PR(>F)
variation_0_color	Current_price	12966.597452	773.670567	**7.499581e-169**
	Raw_price	9375.489549	556.77557	**2.310318e-122**
	Discount	5.757827	0.337798	0.561106
	Likes_count	85.970344	5.044197	**0.024713**
variation_1_color	Current_price	2055.451694	153.503942	**3.485611e-35**
	Raw_price	1023.543114	76.282336	**2.558763e-18**
	Discount	404.791298	30.131003	**4.064432e-08**
	Likes_count	14.179106	1.054614	0.304453

Table 4: One-Way ANOVA results

From the One-Way-ANOVA analysis, it can be seen that there are significant relationships between the product's primary color and current price, raw price, and likes count. Also, there are significant relationships between the secondary color and current price, raw price, and discount percentages. We can say that the One-Way-ANOVA analysis results support the hypothesis we test.

Ordinary Least Squares regression (OLS) Regression

In an article in xlstat.com, the author introduces the ordinary least squares regression (OLS) (2017) it is a widely used method for calculating the coefficients of linear regression equations that represent the connection between one or more independent variables and a dependent variable (simple or multiple linear regression). The least-squares mistake is referred to as the least-squares error (SSE) (XLSTAT, 2017). This analysis was performed using statsmodels (Seabold & Perktold, 2010) in Python. The OLS analysis results show in the table below:

Independent column	Dependent column	R-squared	Coefficient	Std error	Prob (F-statistic)
variation_0_color	Current_price	0.346	25.2131	8.624	**0.0444**
	Likes_count	0.042	10.0836	6.984	0.522
	Raw_price	0.397	25.4963	7.838	**0.0281**

Independent column	Dependent column	R-squared	Coefficient	Std error	Prob (F-statistic)
	Discount	0.097	41.1466	34.483	0.325
variation_1_color	Current_price	0.095	10.8319	5.366	0.331
	Raw_price	0.256	13.8619	4.607	0.0936
	Discount	0.686	55.6206	10.756	**0.000876**
	Likes_count	0.167	10.2022	3.445	0.187

Table 5: OLS results

If the prob F-statistic less than 0.05 this indicates that there is a significant liner regression relationship between the independent variable and the dependent variables. From the table above we can find that there is a significant liner regression relationship between the main color and current price and raw price. Also, there is a single significant liner regression relationship between the secondary color and the discount percentage.

Chi-Square

One technique to illustrate a link between two category variables is to use a chi-square statistic. Numerical (countable) variables and non-numerical (categorical) variables are the two categories of variables in statistics (Statistics How To, 2021). The following table shows the chi-square test results for the main colors with the current price, likes count, raw price, and the discount percentage. From the table we can find there are strong relationships between the independent variables the main colors because the p-values for all variables are under 0.05 which is the value that we can reject the null hypothesis in it.

Independent variable	Dependent variable	Statistic	p-value
variation_0_color	Current_price	222.95358163016857	1.2515423949015698e-
	Likes_count	2619.180846695888	0.0
	Raw_price	587.8100941280893	5.667774466868258e-119
	Discount	504.24785343349635	3.9825957848225824e-

Table 6: Chi-Square summary

Data Visualizations

Visualizations considered one of the most important ways to explore and understand the data. In addition, visualizations could be used to see the data distribution and take a general idea about it. Visualizing a variable's distribution is critical to quickly comprehend properties like frequencies, peaks, skewness, center, modality, and how variables and outliers behave in the data range (Iyim, 2020).

There are six variables were used to create the charts bellow as follow:

Variable Name	Data Type	Data Range	Describe
Category	String	Women, men, kids, shoes, and bags	The product category.
Main color	String	Black, blue, brown, gray, green, orange, pink, purple, red, white, yellow, and other[1].	The product main color. There is 0 Null value in this column.
Secondary color	String	Black, blue, brown, gray, green, orange, pink, purple, red, white, yellow, and other.	The product additional color. There are some Null values in this column.
Current price	Number	0.14 - 314.59	The product price.
Discount	Number	0% – 100%	The discount percentage.
Likes count	Number	0 - 21547	Number of likes.
Two colors	Boolean	True – False	Indicates if the product has just main color or more than one color.

Table 7:Variables are used in the visualizations.

All charts were generated by Tableau software (Chabot et al., 2021), which makes creating charts more flexible and more manageable. In addition, all charts were sorted descending from high to low. For more data understanding, here are the numerical variables describe:

Variable name	count	mean	std	min	25%	50%	75%	max
Current price	45425	28.73852	16.02538	0.14	18.04	24.99	35.69	314.59
Discount	45425	52.22261	10.40863	0	47	52	59	100
Likes count	45425	224.2966	631.9663	0	29	75	189	21547

Table 8: Numerical variables describe

[1] The other in the product color means the products have pictures or patterns.

The number of products based on color

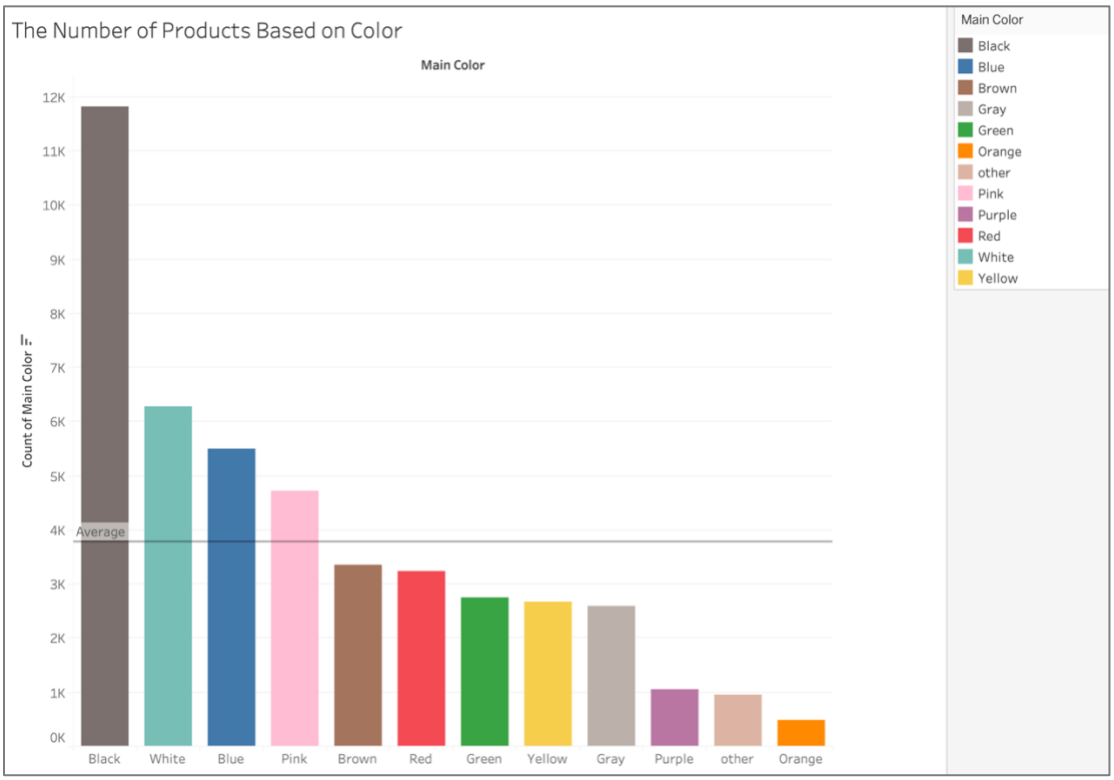

Figure 13: Product counts based on color [generated by Tableau]

This bar chart demonstrates the color distribution for the entire dataset. The bar graph is a useful tool to represent the categorical data frequency. Also, it is a visual tool that compares data across categories using bars (Bar Graph - Learn About Bar Charts and Bar Diagrams, 2015). It can be noticed that the black color has the highest frequency with 11,818 products. Conversely, the lowest number of product's colors is the orange ones with 493 products. However, the average number of products grouped by color is 3,785 products per color. The average helps us to find what is the products' colors above and under the mean. There are four colors above the average: black, white, blue, and pink. Understanding the distribution of the product's color can help the decision-maker to make the right decisions about the color of their products if they want the product to be within the average or not. For example, if they want to show that the product is different from most products, they can pick the less used colors. Nonetheless, if they want to use standard colors, they can pick colors above the average.

However, this distribution could not be generalized to the entire market because this sample was collected from one e-commerce website and for a few products. Nevertheless, the sample could give an idea about how the population would be.

The average number of likes based on product's color

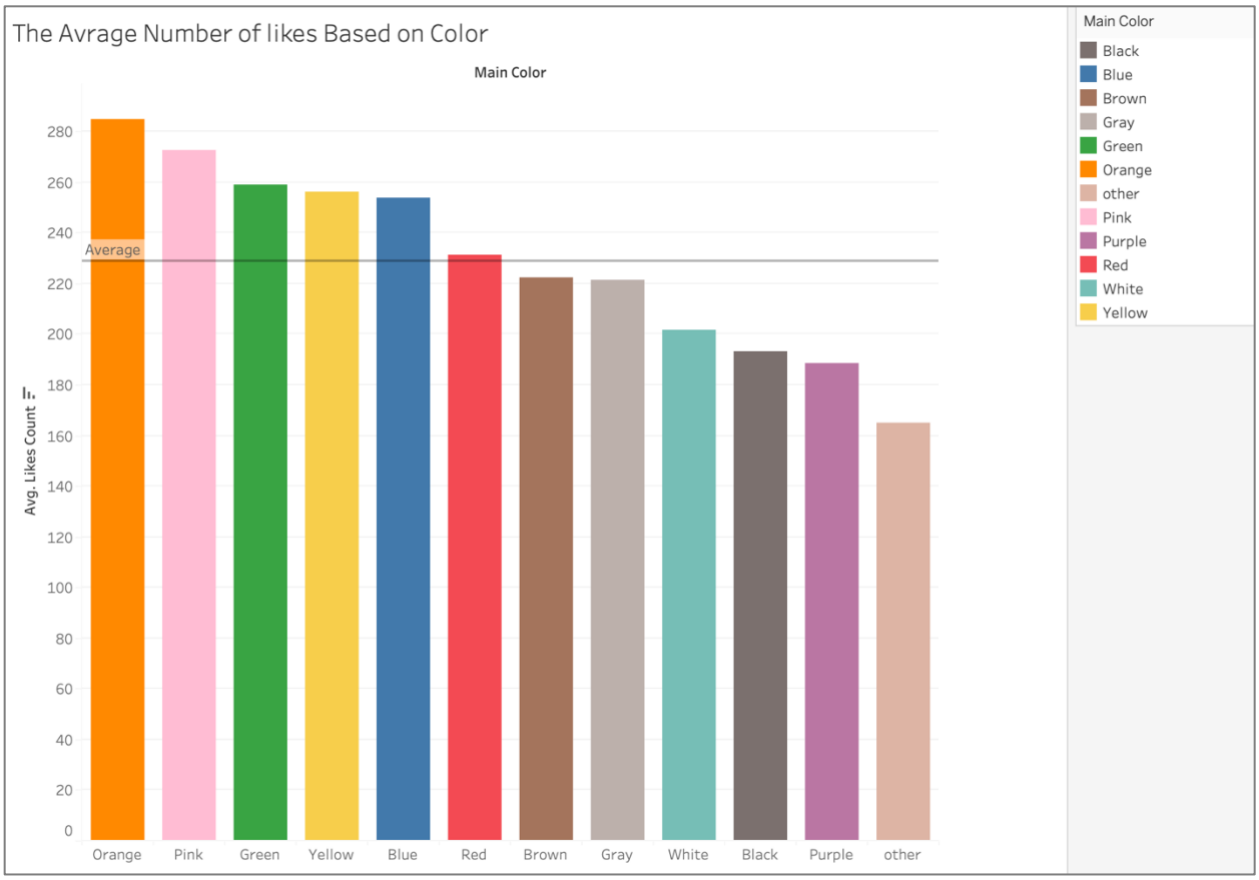

Figure 14: Mean of likes count based on product's color [generated by Tableau]

The bar graph demonstrations the average number of likes per color. This visualization surprised me because the color that gets the highest average number of likes was the lowest color used, which is orange. The orange color showed in the dataset 493 times and got 284.6 average number of likes. While the most color shown in the dataset, which is black, got 193 average number of likes which is less than the overall average likes count. Also, the average number of likes for all colors is 229.1 likes. In addition, it can be noticed that the lowest average number of likes was to the product that does not have a specific color. The average number of likes can help marketers, and decision-makers understand how the color can impact the product likes rate and what customers like the colors most.

This chart supports the project hypothesis by showing how the product color can influence customer behaviors. Although the orange color shows a small amount of product, it got the highest average number of likes. Also, the pink color is used most in the kid's product but got the average number of likes more than the mean.

Average likes count for the products that have more than one color

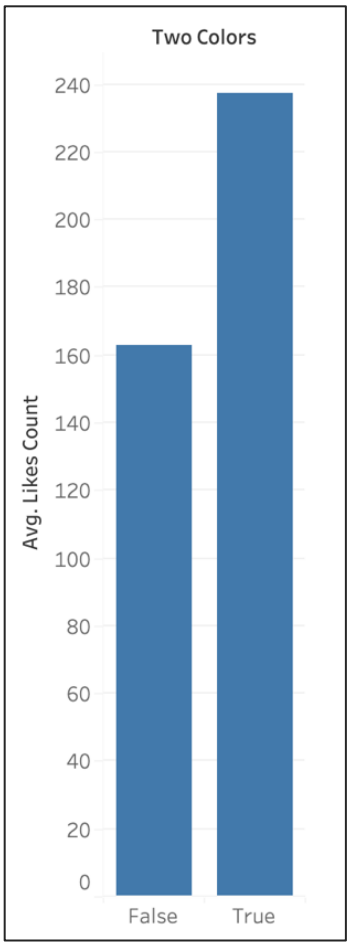

Figure 15: Average number of likes based on the product's color counts [generated by Tableau]

This chart depicts the average number of likes for the product containing multiple colors compared to those with only one primary color. Based on the graph, the products that contain more than one color got a higher average number of likes. However, the difference between the number of likes between the products containing one color and the more than one color is not too significant.

This plot can help the marketing department present more than one color to get customer attention. In addition, the result helps the decision-maker buy or produce products that contain more than one color more than those that have one color because the customers like the products containing multiple colors.

The average number of likes for products that have more than one color based on product's category

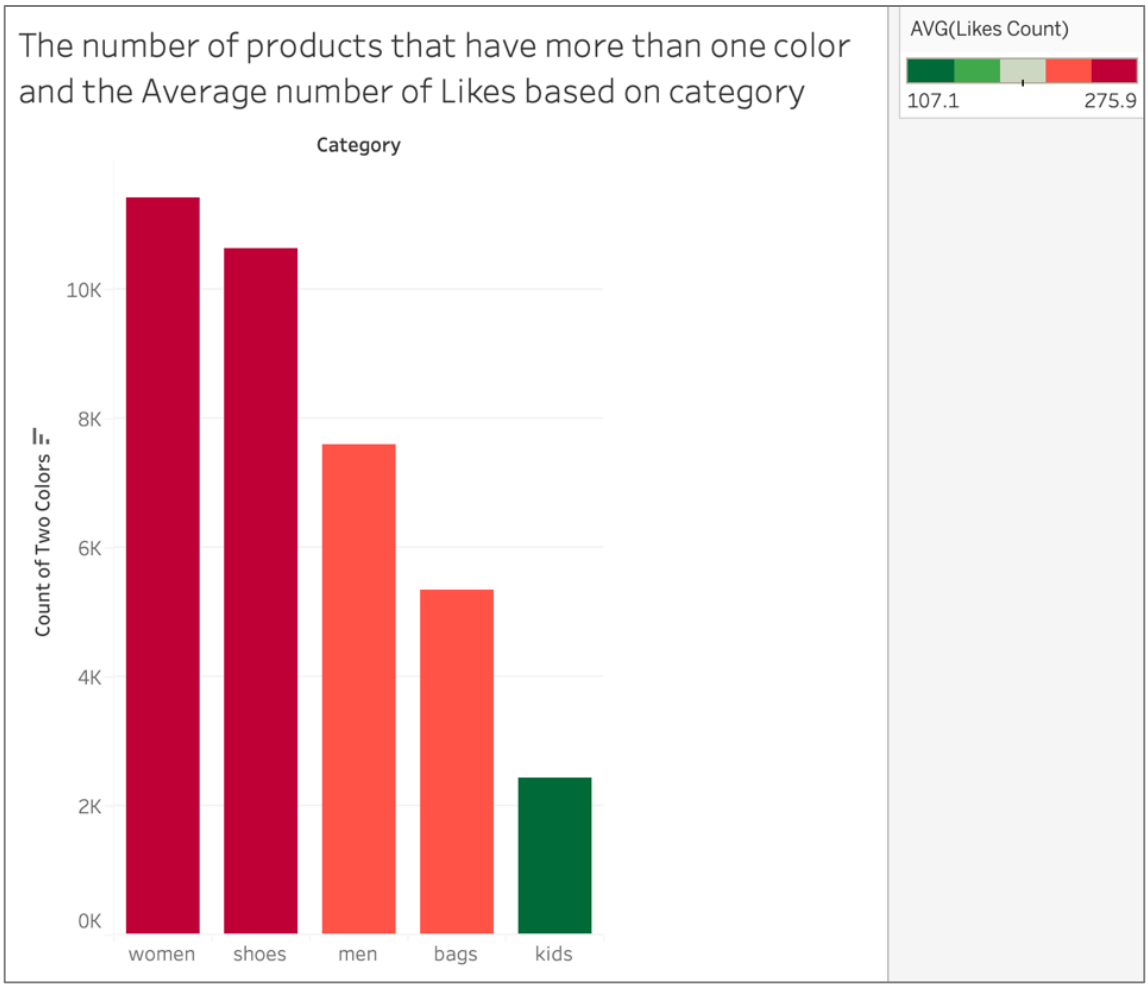

Figure 16: Products count based on category and the average likes count [generated by Tableau]

This bar chart shows the number of products that contain more than one color and the average number of likes divided by product category. The most significant number of likes went to the women and shoes products, followed by the men and bags products. The lowest average number of likes was for the kid's products. This chart indications that the people like the kid's products be in one color—however, women like the products with multiple colors. The chart can help the decision-maker make the right decisions about the product design and understand the customer's desires.

In addition, how much the number of colors can impact customer behavior, which supports the project hypothesis.

Average product's price based on color

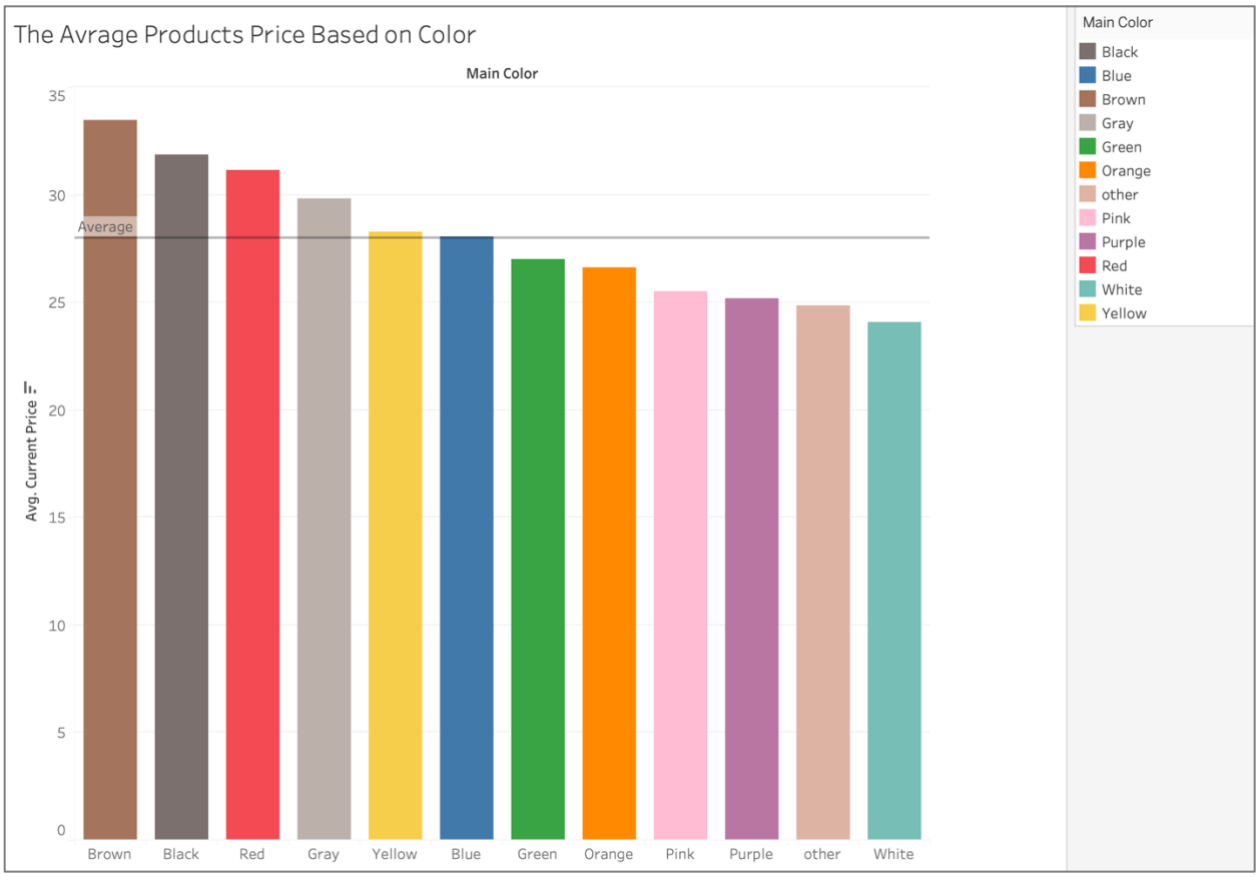

Figure 17: Average product's price based on product's color [generated by Tableau]

This chart depicts the average price per color. The products with brown color got the maximum average price of 33.44 USD, followed by the black products 31.83 USD. At the same time, the lowest average price is for the product that has white colors with 24 USD. Moreover, it can be observed that the blue color has almost the average price. This chart supports the hypothesis used in this research by showing how color can affect the product price. It is obvious that there are differences in the price distribution based on color.

The chart emphasizes the hypothesis correctness by representing the differences in prices based on the product's color.

Challenges

We faced some challenges while examining the hypothesis, but we passed them. The first challenge was the different formats for each color, and this issue required manual work to unify the colors. The second challenge we faced was that the data was not normally distributed. The normality is essential to perform the statistical models. To make the data in normal shape, we take the mean values per category to perform the statistical models.

Also, finding the proper statistical model to examine the data and understand the results were beneficial challenges. This made me read more about the statistical models and understand the purpose of each model. In addition, finding ways to apply these models to the data was another challenge that required searching on the libraries and packages of each model to perform these models.

Moreover, I have tried to collect data from different resources, but I could not found available datasets from different source.

Conclusion

In conclusion, in this project, we examined the hypothesis that "product colors can affect the product price and customer behaviors." To test this hypothesis, we need to set some data requirements, explain how, explore the data, perform statistical and analytical models, and create multiple visualizations. The project outcomes will be beneficial for me by learning new analytical tools and techniques and for those who can use the results to improve their work and make right decisions about the product's colors.

However, after studying all the statistical results and visualizations, we can say that the hypothesis we examine is correct. Because we found significant relationships between the colors and prices and likes number for the products.

The summary of the results got from the visualizations and models are as follow:

- There is a high correlation and significant relationship between the colors and the product's prices.
- Products with orange color have higher confidence intervals to get more likes than other colors.
- Brown products have the high mean prices than other products.
- Purple products have a higher mean discount percentage than other products.
- The black color is the most used in the products.
- Most products have more than two colors.
- The highest average number of likes went for the products with orange color.
- The products with multiple colors get a high average number of likes.
- Women and shoe products with more than one color got more likes than others.
- The highest average price was for the brown products followed by the black ones.
- There is a slightly different discount rate between different colors.
- There is no significant difference in the average price and discount between products with only one color or multiple ones.

On the other hand, it needs to be considered that this distribution is for one e-commerce website and for specific kinds of products, which are clothes, shoes, and bags. We cannot stereotype these results unless we have more sources with a more significant number of products.

Future Work

To enhance the results showed in this research we need to get data from different sources. There are many e-commerce websites that have the same kind of products such as Shein.com, Gap.com, Zara.com to name a few. However, these websites need to use web scrape to get the products data because I could not find any available data comes from these sites. Also, to collect these data we may need permission from the owners to use their data.

On the other hand, I preferer to make the rating out of five rather than collect the sum of likes as we have in this project. Because many e-commerce websites use this rating format.

References:

Anaconda Software Distribution. (2020). Anaconda Documentation. Anaconda Inc. Retrieved from https://docs.anaconda.com/

Bar Graph - Learn About Bar Charts and Bar Diagrams. (2015, November 22). Smartdraw. Retrieved November 7, 2021, from https://www.smartdraw.com/bar-graph/

Burchfield, P. (2017, March 20). [Throttle Roll - Swap Meat Market]. Unsplash. https://unsplash.com/photos/tvG4WvjgsEY

Chabot, C., Beers, A., & Hanrahan, P. (2021). *Tableau* (2021.3.3) [Computer software]. Tableau. https://www.tableau.com

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Iyim, C. (2020, April 19). *Recipes for the Visualizations of Data Distributions*. Medium. Retrieved November 6, 2021, from https://towardsdatascience.com/recipes-for-the-visualizations-of-data-distributions-a1527a0faf77

Mabilama, J. M. (2020, September 7). Products Catalog - from E-commerce Retail site NewChic.com - dataset by jfreex[Catalog of products and the popularity of items]. data.world. https://data.world/jfreex/products-catalog-from-newchiccom

Mcleod, S. (2019, June 10). Z-Score: What are Confidence Intervals in Statistics? Simply Psychology. Retrieved November 27, 2021, from https://www.simplypsychology.org/confidence-interval.html

Statistics How To. (2021, November 20). Chi-Square Statistic: How to Calculate It / Distribution. Retrieved November 27, 2021, from https://www.statisticshowto.com/probability-and-statistics/chi-square/

The pandas development team. (2020). pandas-dev/pandas: Pandas (3.8.8) [Python library]. Zenodo. https://doi.org/10.5281/zenodo.3509134

van Rossum, G., & Drake Jr, F. L. (1995). Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

VanderPlas, J., Granger, B., Heer, J., Moritz, D., Wongsuphasawat, K., Satyanarayan, A., Lees, E., Timofeev, I., Welsh, B., & Sievert, S. (2018). Altair: Interactive Statistical Visualizations for Python. Journal of Open Source Software, 3(32), 1057. https://doi.org/10.21105/joss.01057

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. GetMobile: Mobile Computing and Communications, 19(1), 29–33. https://doi.org/10.1145/2786984.2786995

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . van Mulbregt, P. (2020). Author Correction: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 352. https://doi.org/10.1038/s41592-020-0772-5

Visual Studio Code (1.61.0). (2019). [Software]. Microsoft. https://code.visualstudio.com

Wilson, L. T., & Wilson, L. T. (n.d.). Statistical Correlation. Explorable. Retrieved October 15, 2021, from https://explorable.com/statistical-correlation

XLSTAT. (2017, July 3). Ordinary Least Squares regression (OLS). Retrieved November 27, 2021, from https://www.xlstat.com/en/solutions/features/ordinary-least-squares-regression-ols

Table of Tables

Table of Figures