Date: 06-Dec-23

Author: Haritha Injam

CS 5764: Information Visualization

Professor :
Dr. Reza Jafari, Ph.D
Collegiate Associate Professor

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Used Dataset after Preprocessing for fast app load: App Link1: https://dashapp-zqsxzaj23a-nn.a.run.app/
Used Full Dataset: App Link2: https://myntraanalytics9999-j5dzbxl6fa-uc.a.run.app/

ABSTRACT

This project explores a comprehensive dataset of fashion product listings from the Indian online shopping platform Myntra. Using various data analysis techniques, the study aims to uncover patterns and insights into consumer preferences, pricing strategies and product features in the online fashion retail industry. Key methods include feature engineering, data preprocessing, outlier detection, principal component analysis (PCA), normality testing and correlation analysis to understand the dynamics of online fashion retail.

Keywords: Preprocessing, PCA, Normality Test, Correlation Analysis.

INTRODUCTION

Objective:

The Final Term Project (FTP) involves exploring and visualizing a dataset through static graphs and tables. The primary focus is to unearth underlying patterns and information using various plotting techniques.

Dataset:

The dataset pertains to fashion clothing, with a specific focus on aspects like pricing, discounts, ratings, reviews, Original and Final Prices and categorical attributes like gender, size options, Brand name, Discount Offer, Individual Category Colour and Individual categories.

Approach:

Data Preparation:
Imported necessary libraries for data manipulation and visualization.
Loaded the dataset and performed initial data cleaning including renaming columns for clarity, extracting color from descriptions and handling missing values.

Outlier Detection and Treatment:
Identified and handled outliers in key numerical features like prices, discount, ratings and reviews to ensure accurate analysis.

Feature Standardization and PCA Analysis:
Performed standardization of numerical features and Principal Component Analysis (PCA) to understand feature importance and reduce dimensionality.

Normality Tests and Data Transformation:
Conducted normality tests like the Kolmogorov–Smirnov Test (K-S Test) or Shapiro-Wilk test or DAgostino's K-squared Test and apply transformations (e.g., Box-Cox) to show data distribution with assumptions of statistical tests.

Visualization Techniques:

Categorical Features: Used pie charts, bar plots (stacked and grouped), and count plots to visualize categorical data.

Numerical Features: Employed line plots, distribution plots (KDE, histogram), QQ plots, and joint plots to explore numerical data.
Multivariate Analysis: Created pair plots, heatmap, 3D plots, and cluster maps to analyze relationships between multiple variables.
Specialized Plots: Implemented violin plots, area plots, swarm plots, and hexbin plots for detailed insights.

Subplots: Incorporated subplots to create a comprehensive storyboard that highlights various aspects of the dataset.

Statistical Analysis Tables:
Utilized PrettyTable to present statistical analyses like mean, variance and standard deviation of key features.

Customizations and Aesthetics:
Enhanced plots with customizations such as color palettes, line widths, font sizes, and appropriate labels and legends for clarity and aesthetics.

Phase II:

This report also presents a comprehensive overview of the development and functionalities of an interactive web-based dashboard designed using the Dash framework. The dashboard is structured to provide dynamic data visualization, enabling users to interact with the data in real time without the need for re-executing the underlying Python code. The primary objective of this dashboard is to offer a user-friendly interface for data analysis and visualization, facilitating efficient and insightful exploration of myntra dataset.

The dashboard integrates various core and HTML components of Dash, such as Checklists, Dropdowns, Graphs, Loaders, Download options, Radio items, Range Sliders, Sliders, Tabs, Textareas, Tooltips and more. These components enhance the dashboard's interactivity, allowing users to select, filter, and manipulate data visualizations according to specific parameters or interests. The layout is meticulously organized, utilizing components like Div, Figure, Headers, Images, Labels and different heading levels (H1-H6) to ensure clarity and a cohesive structure.

A key feature of the dashboard is its ability to update graphs and visualizations based on user inputs without re-running the entire Python script. This is achieved through callback functions in Dash, which listen for changes in the user interface and update specific parts

of the dashboard accordingly. This approach significantly improves the performance and responsiveness of the dashboard, providing a seamless user experience.

In addition to its interactivity, the dashboard is designed with a focus on aesthetic appeal and usability. The use of external stylesheets and careful consideration of layout design ensures that the dashboard is not only functional but also visually engaging.

In summary, this report outlines the procedures followed to achieve the FTP objectives through the development of the interactive web-based dashboard. It details the integration of various Dash component and the implementation of callback functions for dynamic data visualization.

DESCRIPTION OF THE DATASET

The Myntra dataset consists of over 50,000 observations i.e., 526565 rows, each detailing a fashion product. It includes

Numerical features like

- Final Price: The price of the product after the discount has been applied.
- Original Price: The price of the product before any discounts
- Ratings: The average customer rating for the product on a scale, usually from 1 to 5
- Reviews: The number of reviews the product has received
- Discount Amount: The amount of money that is deducted from the original price, calculated as the difference between the original and final prices.

Categorical features like

- Brand Name: The name of the manufacturer or designer of the product, categorizing items by the company that produced them.
- Category: A broad classification that groups products into general types such as 'Bottom Wear' or 'Topwear,' reflecting the overarching segment each item belongs to.
- Individual Category: A more detailed categorization that specifies the exact type of product, such as 'jeans' or 'shirts,' providing a finer classification within the broader category.
- Category by Gender: A designation of the intended gender demographic for the product, such as 'Men' or 'Women,' used to sort items based on gender suitability.
- DiscountOffer: Promotional information displayed as a percentage or other value, indicating the price reduction from the original cost.
- SizeOption: The range of available sizes for the product, represented categorically (e.g., S, M, L, XL) and not intended to be quantitatively analyzed.
- Individual Category Colour: The primary color noted in the product description, acting as a categorical attribute to describe the dominant color of the item.

'Discount Amount', 'Individual Category Colour' serves as the dependent variables, with others acting as independent variables. This dataset is crucial for analyzing market trends, consumer behavior and pricing strategies in the fashion industry.

PRE-PROCESSING DATASET

The provided code performs data cleaning in the following ways:

Color Extraction: A custom function extract_first_color is defined to parse the 'Description' column for colors and create a new column 'Individual_Category_Colour'. The extraction of color information from the 'Description' column can be considered a form of feature engineering, as it creates a new feature that categorizes products by color.
For rows where no color is found (i.e., None) are dropped, which cleans out entries without identifiable color information.

Dropping Unnecessary Columns: Columns such as 'URL', 'Product_id', and 'Description' are dropped since they are not needed for the analysis.

Checking for Missing Values: The code checks for missing values using both isna() and isnull() methods. Both methods are essentially the same in pandas and return the number of missing values in the DataFrame.

Removing Rows with Missing Values: The code uses dropna(inplace=True) to remove any rows with missing values across the entire DataFrame, ensuring that all remaining data is complete.

Confirmation of Cleaning: After the cleanup process, a check confirms that there are no missing values left in the DataFrame isna().sum().sum() and isnull().sum().sum() both return 0.

Renaming Columns: The 'DiscountPrice (in Rs)' column is renamed to 'Final Price (in Rs)' for clarity.

Calculating Discount Amount: A new feature 'Discount Amount' is created by subtracting 'Final Price (in Rs)' from 'OriginalPrice (in Rs)', which is used for analyzing the extent of discounts on products.

Observations:

- The code successfully cleans the dataset by removing rows with missing values, which totaled 868,594 before cleaning.
- After cleaning, the dataset contains 108,861 entries, indicating a significant reduction in size due to the removal of incomplete records.

- The first five entries of the cleaned dataset show a variety of products, categories and a new 'Individual_Category_Colour' feature, as well as the calculated 'Discount Amount'.
- The preprocessing steps have enhanced the dataset with additional features that may be relevant for further analysis, such as color-based categorization and quantitative discount information.

Final Display of Cleaned Data:

The final output of the code section is a display of the first five records in the cleaned dataset, which includes the newly engineered features and the relevant statistics such as brand names, categories, final prices, original prices, discount offers, size options, ratings, reviews, colors, and discount amounts.

```
myntra = df.copy()

myntra.rename(columns={'DiscountPrice (in Rs)': 'Final Price (in Rs)'}, inplace=True)

colors = ['red', 'green', 'blue', 'black', 'white', 'yellow', 'pink', 'navy',
          'olive', 'maroon', 'khaki', 'burgundy', 'grey', 'beige', 'orange',
          'purple', 'lavender', 'brown', 'mauve','peach','violet','magenta']


1 usage  Haritha Injam +1
def extract_first_color(desc):
    for color in colors:
        if color in desc:
            return color
    return None  # Return None if no color is found

    # Apply the function to create a new color column
myntra['Individual_Category_Colour'] = myntra['Description'].apply(extract_first_color)

    # Drop rows where 'Individual_Category_Colour' is None
myntra = myntra.dropna(subset=['Individual_Category_Colour'])
```

```python
    # Assuming myntra is your DataFrame
myntra = myntra.drop(labels: ['URL', 'Product_id', 'Description'], axis=1)

    #print(myntra.head())

    # Check for missing values using isna() and isnull()

missing_values_na = myntra.isna().sum().sum()
missing_values_null = myntra.isnull().sum().sum()

print(f'Missing values using isna(): \n{missing_values_na}')
print(f'Missing values using isnull():\n{missing_values_null}')

myntra['Discount Amount']=myntra['OriginalPrice (in Rs)']-myntra['Final Price (in Rs)']
    # myntra['Discount Amount'].sort_values (ascending=False)

    #CleanUp
myntra.dropna(inplace=True)
```

```python
    # Confirming that dataset is clean
missing_values_na_after = myntra.isna().sum().sum()
print(f'Missing values after dropping using isna(): {missing_values_na_after}')

missing_values_null_after = myntra.isnull().sum().sum()
print(f'Missing values after dropping using isnull(): {missing_values_null_after}')

print('After Data Cleaning!')
print(myntra.shape[0])
print(myntra.head(5))
```

```
Missing values using isna():
868594
Missing values using isnull():
868594
Missing values after dropping using isna(): 0
Missing values after dropping using isnull(): 0
After Data Cleaning!
108861
```

OUTLIER DETECTION & REMOVAL

The code uses the Interquartile Range (IQR) method to detect outliers. Here's how it works:

- **Calculate Q1 and Q3:** These are the 25th and 75th percentiles of the data. Q1 is the median of the first half of the data, and Q3 is the median of the second half.
- **Compute IQR:** This is the range between the first and third quartile, $IQR = Q3 - Q1$. It measures the middle spread of the data.
- **Determine Boundaries:** The lower boundary is calculated as $Q1 - 1.5 \times IQR$, and the upper boundary is $Q3 + 1.5 \times IQR$. Data points outside these boundaries are typically considered outliers. Below is the function defined to define boundaries for performing Outlier Analysis.

```
def find_outliers_IQR(data):
    q1 = data.quantile(0.25)
    q3 = data.quantile(0.75)
    iqr = q3 - q1
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    outliers = data[(data < lower_bound) | (data > upper_bound)]
    return q1, q3, iqr, lower_bound, upper_bound, outliers
```

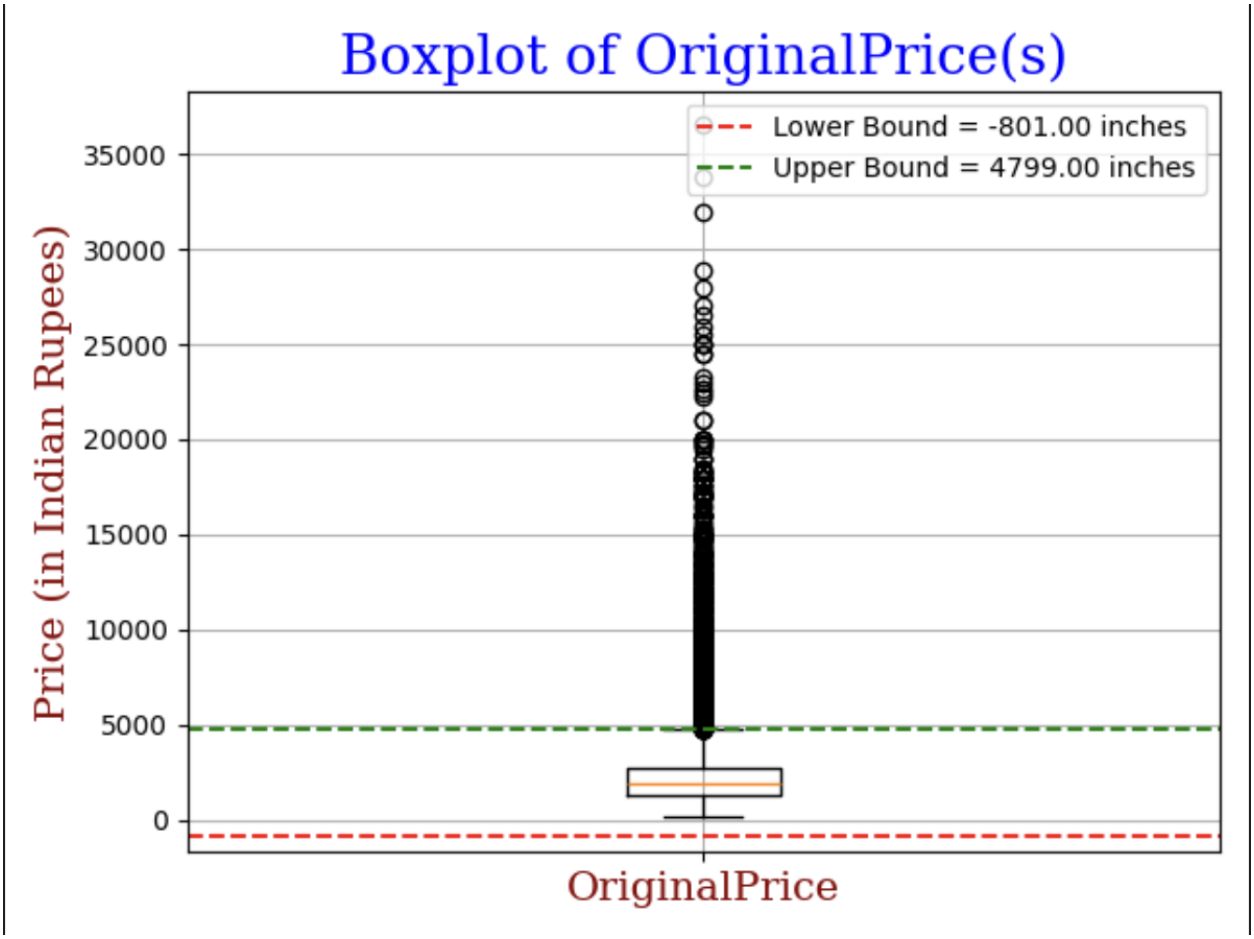
- **Filter Outliers:** The dataset is then filtered to exclude any data points that fall outside the lower and upper bounds.
- **Data Visualization:** Boxplots are generated before and after removing the outliers to visualize the effect of the outlier removal process.

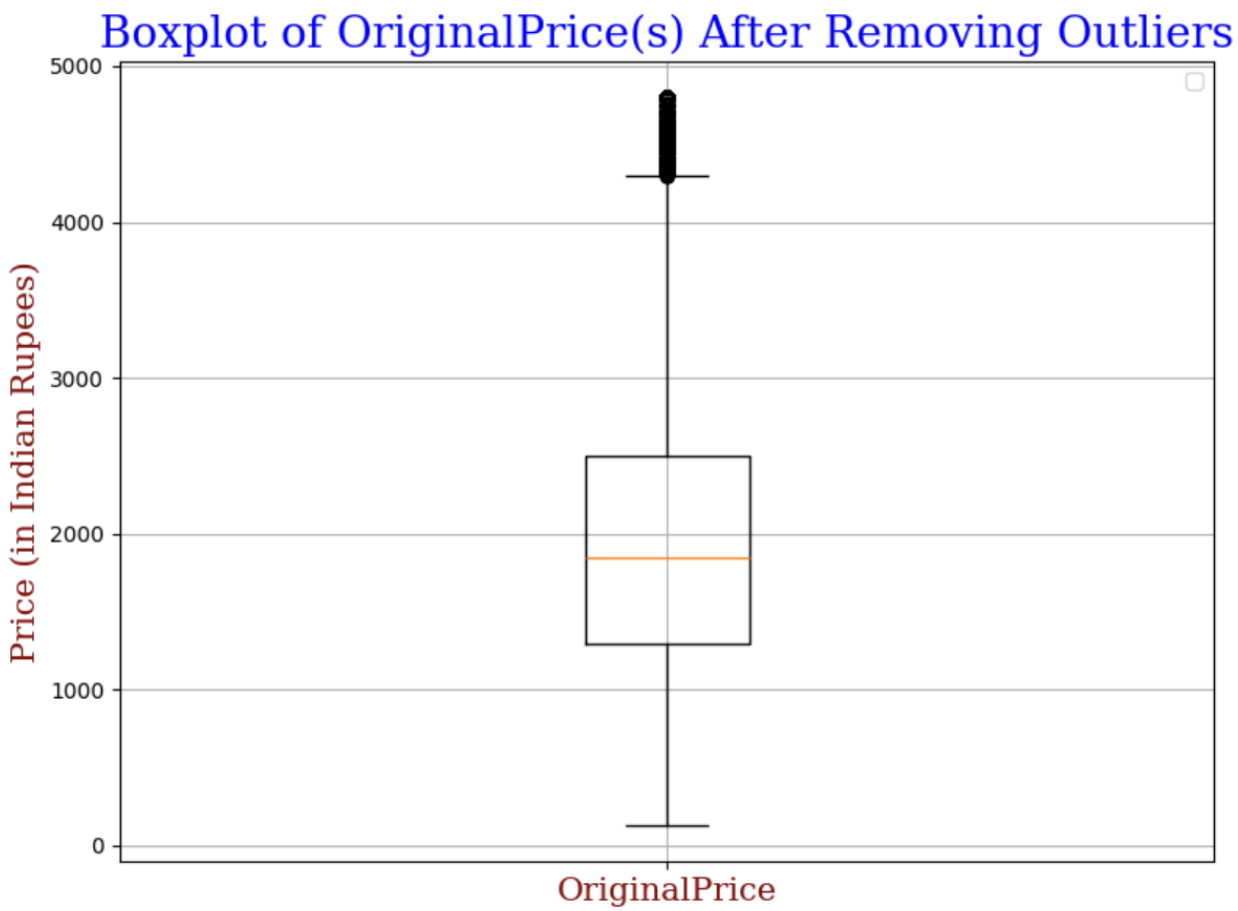
Observations:

Data Cleaning: The dataset is cleaned of NaN values, ensuring that the analysis is not affected by missing data.
Initial Data Analysis: The boxplot of the original prices shows a significant number of outliers above the upper bound. This indicates a long tail in the distribution of the clothing prices.
Thresholds: The calculated lower bound is -801, which doesn't make sense in the context of prices and indicates that there are no lower outliers. The upper bound is 4799, suggesting that prices above this value are considered unusually high for this dataset.
Outlier Impact: The number of outliers, as well as the maximum and minimum outlier values, provided insight into the range and impact of these unusual data points. The large difference between the maximum price and the upper bound suggests some items are priced much higher than typical.

Post-Removal Analysis: After removing the outliers, the boxplot shows a more compact distribution, indicating that the extreme values have been removed. This could lead to better model performance if the outliers were due to data entry errors or were not representative of the dataset as a whole.
Final Dataset: After cleaning and removing outliers, the dataset presumably has a more standardized range of prices, which could be more representative of the typical items sold.

```
Q1, Q3, IQR, Lower Bound, Upper Bound:
1299.0 2699.0 1400.0 -801.0 4799.0
Prices lower than -801.0 and higher than 4799.0 are considered outliers.
```





NOTE: The Explanation for the graphs used in Outlier Analysis section of dashboard can be found here.

Principal Component Analysis (PCA):

Standardization of the Feature Space:

- Standardizing the dataset is a crucial preprocessing step before applying PCA. This is because PCA is sensitive to the variances of each feature and assumes that the features have been centered around zero and scaled to have similar variances.
- To achieve this, the StandardScaler from the sklearn.preprocessing package was utilized, which transformed the feature space such that each feature had a mean of zero and a standard deviation of one. Specifically, the features 'Final Price (in Rs)', 'OriginalPrice (in Rs)', 'Ratings', 'Reviews' and 'Discount Amount' were normalized.

Singular Value Decomposition (SVD):

- SVD was applied to the standardized data to decompose it into its eigenvalues and eigenvectors. This step is fundamental in PCA as it identifies the directions (principal components) that maximize the variance in the data.
- It was observed that the condition number was exceedingly high (approximately 2.89×10^13) which can lead to numerical instability during computations due to the presence of very small singular values.

Correlation Matrix:

- The correlation matrix is an essential tool to examine before PCA because it provides insight into the linear relationships between features. It helps in identifying features that can be combined to reduce redundancy.
- In the original feature space, there was a discernible correlation between 'Final Price' and 'Original Price', as well as between 'Original Price' and 'Discount Amount'. PCA aims to transform these correlated features into a set of linearly uncorrelated principal components.

PCA Analysis:

Explained Variance:

- In PCA, explained variance refers to the proportion of the dataset's total variance that is attributed to each principal component. It effectively measures the importance of each component in capturing the variability in the data.
- The cumulative explained variance plot indicated that around 95% of the total variance was captured by the first four principal components. This informed the decision on the number of components to retain.

Optimal Number of Components:

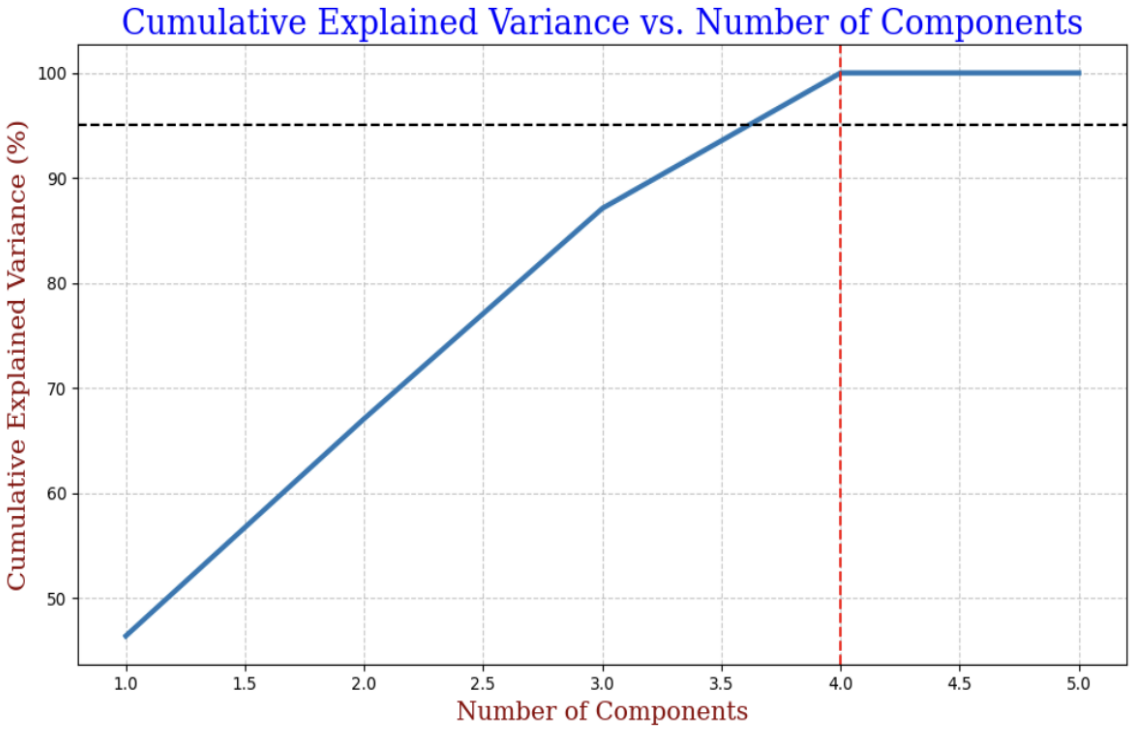
- The choice of retaining four principal components was made after evaluating the explained variance ratio, which suggested that these components accounted for at least 95% of the total variance.
- The "elbow method" used in the scree plot visualizes the point where the marginal gain in explained variance decreases significantly, indicating that additional components do not contribute much to capturing further data variance.

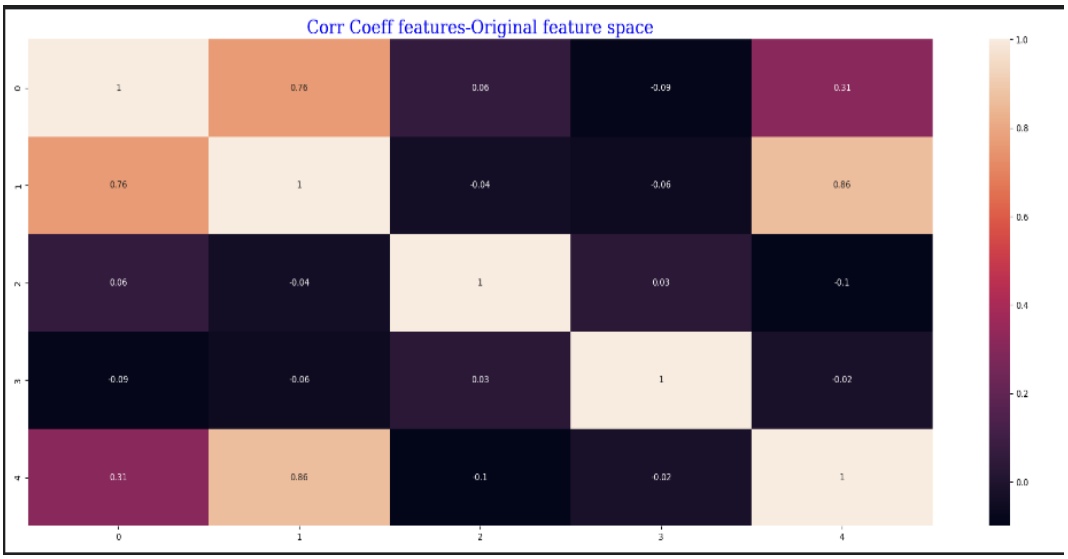
Condition Number of Reduced Feature Space:

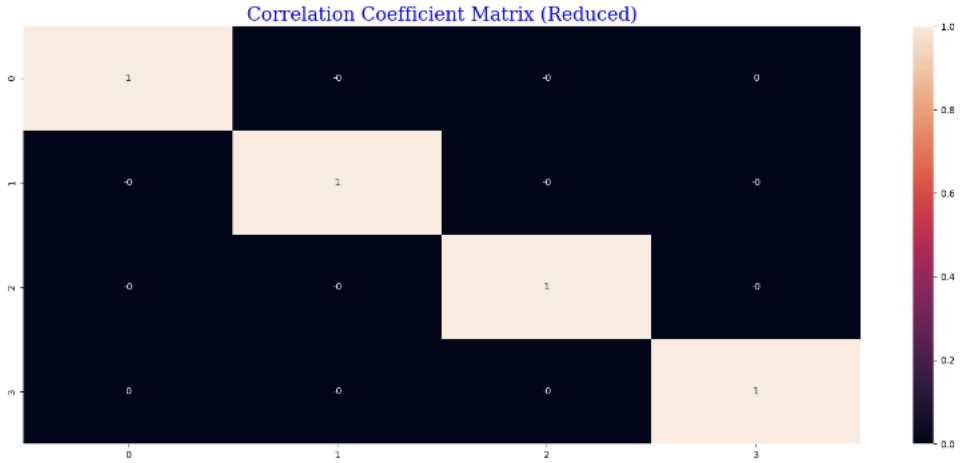
- The condition number for the reduced feature space improved drastically to 1.897, compared to the original feature space. This reduction confirms that the PCA-transformed feature space is more numerically stable.

Correlation Coefficient Matrix for Reduced Feature Space:

- The correlation coefficient matrix for the reduced feature space displayed near-zero off-diagonal elements, validating that the principal components are indeed uncorrelated, which is a desired outcome in PCA.



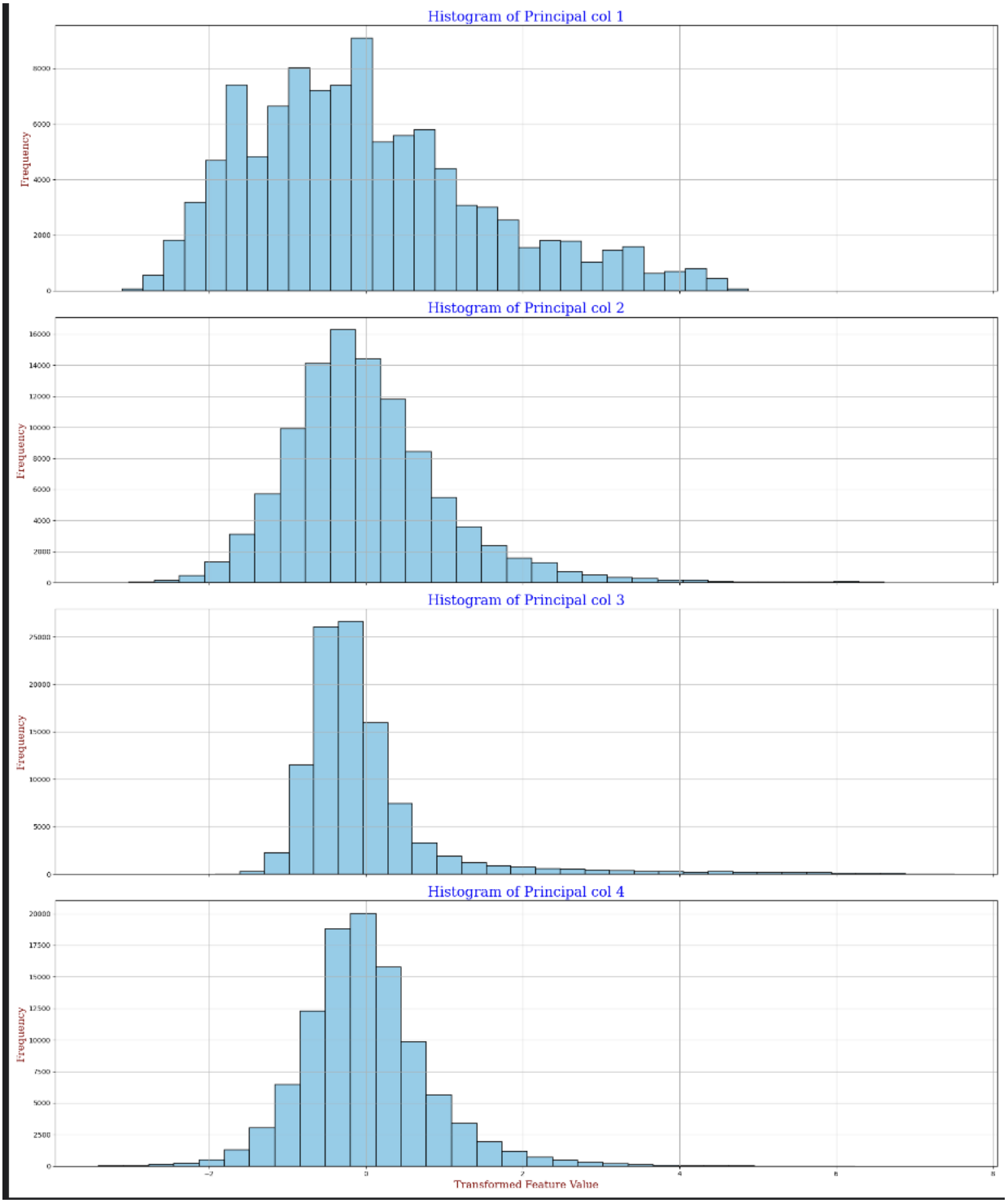





```
Singular Values for Original Feature Space: [4.87545794e+02 3.25408459e+02 3.20278520e+02 2.57002147e+02
 1.68862674e-11]
Condition Number for Original Feature Space: 28872324538739.746
Correlation Coefficient between features-Original feature space
[[ 1.          0.75868838  0.05532217 -0.09182834  0.31406445]
 [ 0.75868838  1.         -0.03602859 -0.06098921  0.85676829]
 [ 0.05532217 -0.03602859  1.          0.03040568 -0.09630052]
 [-0.09182834 -0.06098921  0.03040568  1.         -0.01619032]
 [ 0.31406445  0.85676829 -0.09630052 -0.01619032  1.        ]]
[4.64060171e-01 2.06728876e-01 2.00262252e-01 1.28948701e-01
 5.57583803e-28]
[0.46406017 0.67078905 0.8710513  1.         1.        ]
```

```
Explained Variance Ratio (Original Feature Space): [4.64060171e-01 2.06728876e-01 2.00262252e-01 1.28948701e-01
 5.57583803e-28]
Cumulative Explained Variance Ratio (Original Feature Space): [0.46406017 0.67078905 0.8710513  1.         1.        ]
Number of features to be removed: 1
```

```
Explained Variance Ratio (Reduced Feature Space): [0.46406017 0.20672888 0.20026225 0.1289487 ]
Singular Values for Reduced Feature Space: [487.54579375 325.4084587  320.27852017 257.00214692]
Condition Number for Reduced Feature Space: 1.8970494977771044
Correlation Coefficient Matrix (Reduced)
[[ 1.00000000e+00 -2.00127375e-14 -1.93611081e-14  5.58265301e-14]
 [-2.00127375e-14  1.00000000e+00 -4.11307736e-15 -1.37179716e-14]
 [-1.93611081e-14 -4.11307736e-15  1.00000000e+00 -3.52827278e-15]
 [ 5.58265301e-14 -1.37179716e-14 -3.52827278e-15  1.00000000e+00]]
```



NOTE: The Explanation for the graphs used in PCA Analysis section of dashboard can be found here.

NORMALITY TEST & DATA TRANSFORMATION

The purpose of the normality test is to determine whether the data distribution deviates significantly from a normal distribution. Normality is a common assumption for many statistical tests because normal data has predictable behavior and a known distribution.
The Kolmogorov-Smirnov test compares the cumulative distribution of the sample data with the expected cumulative normal distribution. The closer the sample distribution is to the normal distribution, the smaller the test statistic will be. In my case, the K-S test for the original data yielded a statistic of 0.10 with a p-value of 0.00, indicating a significant deviation from normality.
Similarly, the D'Agostino's K-squared test measures skewness and kurtosis to assess the normality of the data. A high K-squared statistic, such as 7212.11 for my original data, suggests that the data does not follow a normal distribution.

```
================================================
K-S test: Myntra OriginalPrice Normality Test KSTest dataset: statistics= 0.10 p-value = 0.00
================================================
da_k_squared test: Myntra OriginalPrice DAK Normality Test dataset: statistics= 7212.11 p-value = 0.00
================================================
Shapiro test : statistics = 0.96 p-value of =0.00
Myntra OriginalPrice Normality Test Shapiro Test dataset looks Not Normal with 99% accuracy
```

Data Transformation:

Data transformation method like the Box-Cox transformation is used to stabilize variance and make the data as "normal" (Gaussian) as possible, as many statistical methods require normally distributed data. The Box-Cox transformation identified a lambda value that maximized the log-likelihood function, indicating the power to which all data should be raised.

After applying the Box-Cox transformation, the K-S test and D'Agostino's K-squared test statistics decreased, with the K-S statistic dropping to 0.04. Although the p-values remained significant, indicating that the data still deviates from normality, the improvement in the test statistics suggests that the transformation made the distribution closer to normal.

Observations from Histograms and QQ-Plots:

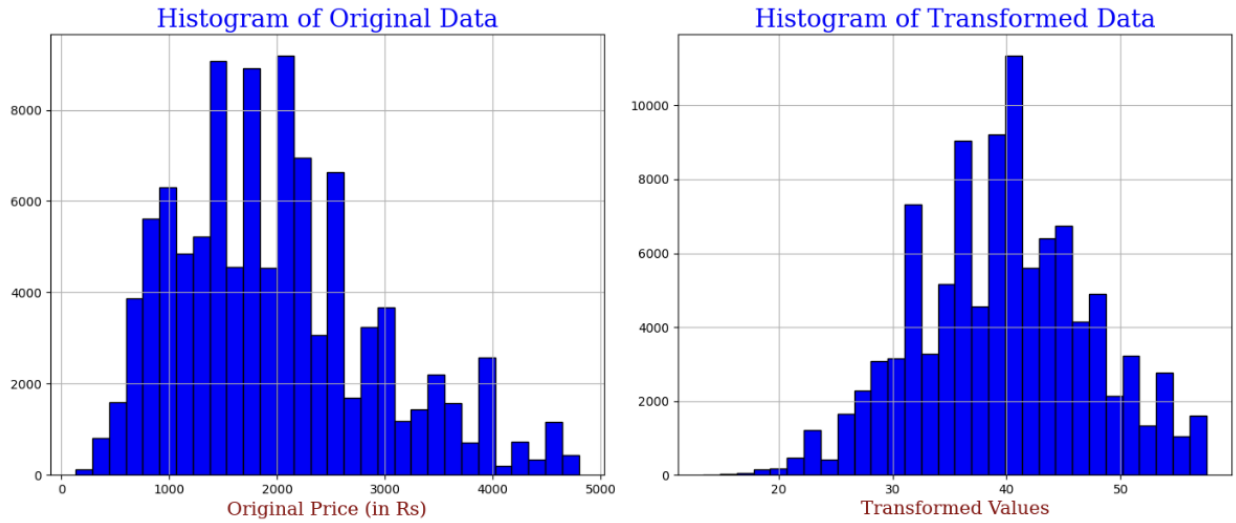
The histograms before and after transformation show a clear difference in distribution shape. The original data's histogram displays a right-skewed distribution, common in

pricing data where a few high values stretch the tail to the right. The transformed data's histogram appears more symmetrical, a characteristic of a normal distribution.

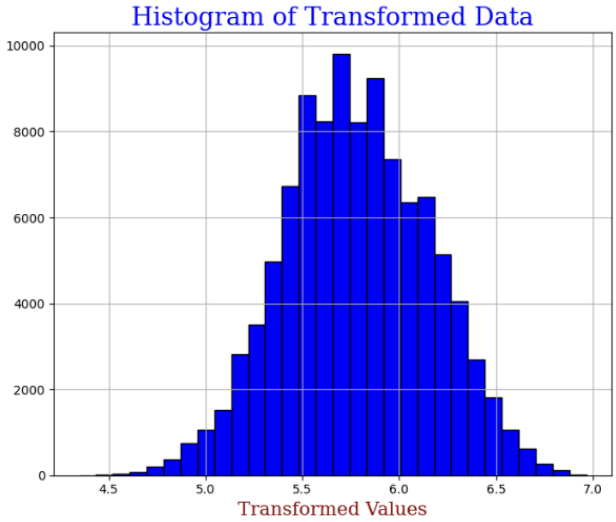
The QQ-plots visually reinforce these findings. In the QQ-plot of the original data, the points deviate significantly from the line, especially in the tails, indicating that the data is not normally distributed. The QQ-plot of the transformed data shows points that lie closer to the line, particularly in the center of the distribution, suggesting a better fit to a normal distribution.

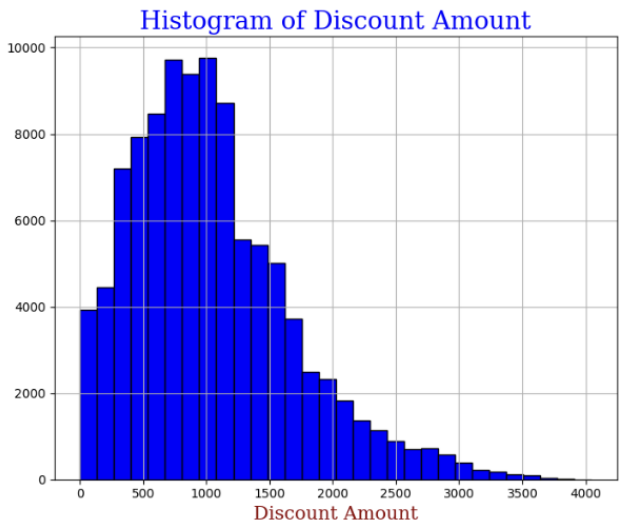
Conclusions:

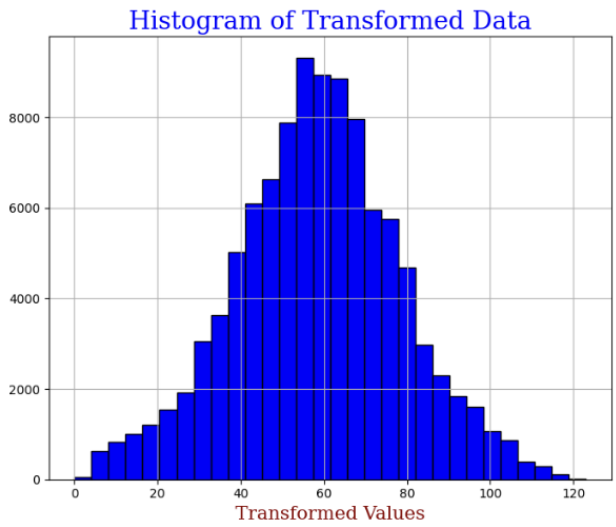
- The original pricing data from the "Myntra Fashion Clothing" dataset does not follow a normal distribution, as confirmed by the K-S and D'Agostino's K-squared tests, along with visual inspection of the histogram and QQ-plot.
- The Box-Cox transformation improved the normality of the data, as shown by the reduction in the test statistics, although p-values indicate that the transformed data still does not perfectly fit a normal distribution.
- This improvement is visually evident in the transformed data's histogram, which looks more symmetric, and the QQ-plot, where points align more closely with the theoretical line.
- The process of normalizing the data, while not yielding a perfect Gaussian distribution, has nonetheless made the distribution more suitable for parametric statistical tests and techniques that assume normality.

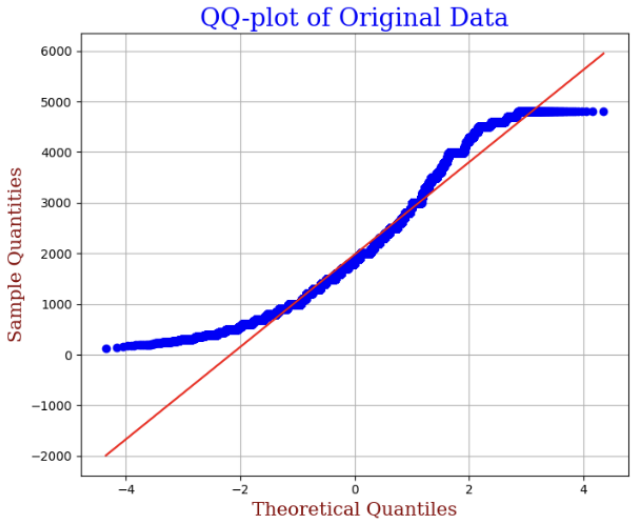


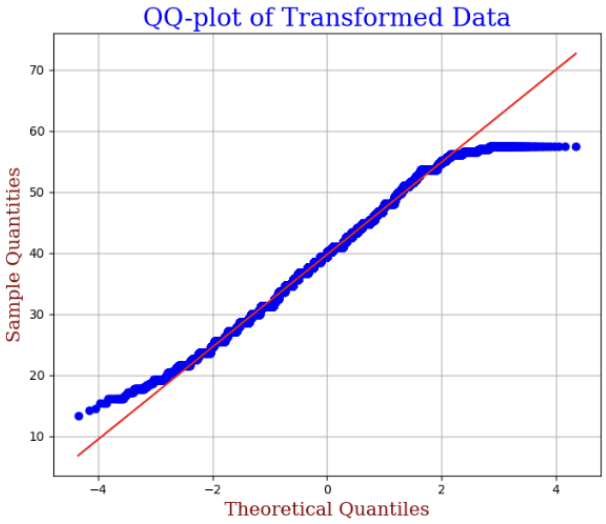

Histogram of Final Price
Final Price (in Rs)

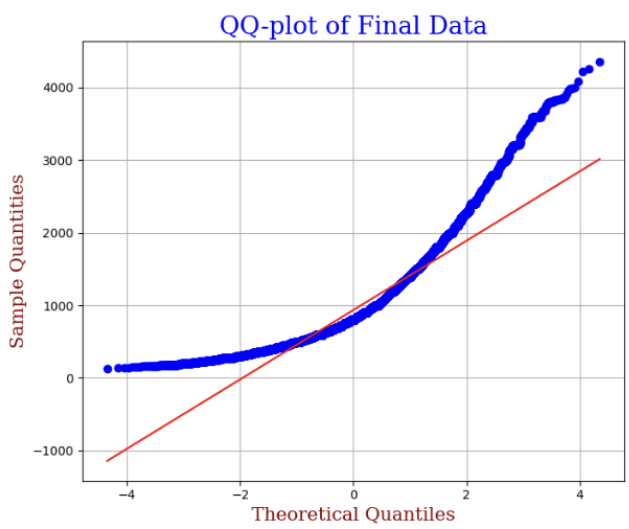

Histogram of Transformed Data
Transformed Values

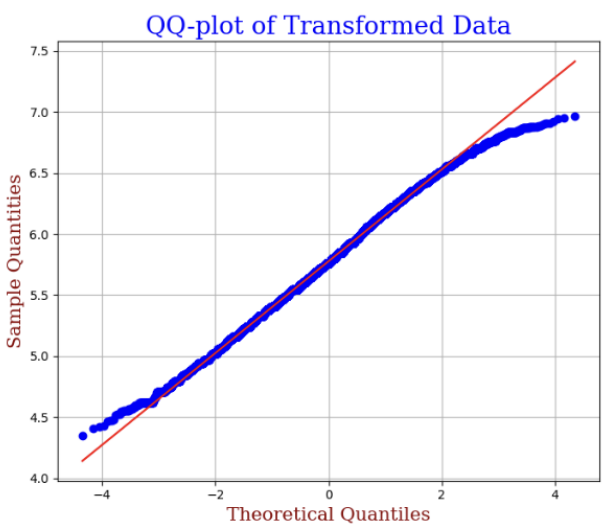

Histogram of Discount Amount
Discount Amount

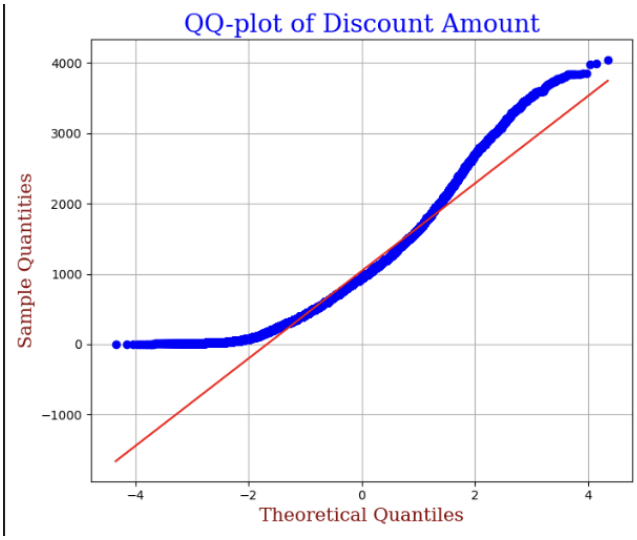

Histogram of Transformed Data
Transformed Values

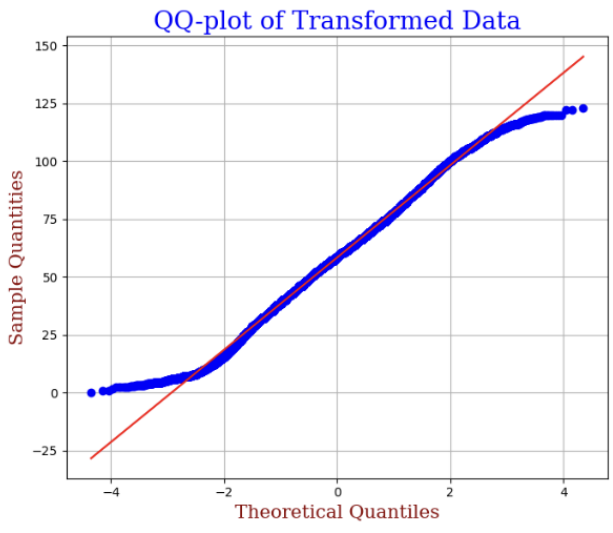

QQ-plot of Original Data
Sample Quantities
Theoretical Quantiles


QQ-plot of Transformed Data
Sample Quantities
Theoretical Quantiles


QQ-plot of Final Data
Sample Quantities
Theoretical Quantiles


QQ-plot of Transformed Data
Sample Quantities
Theoretical Quantiles

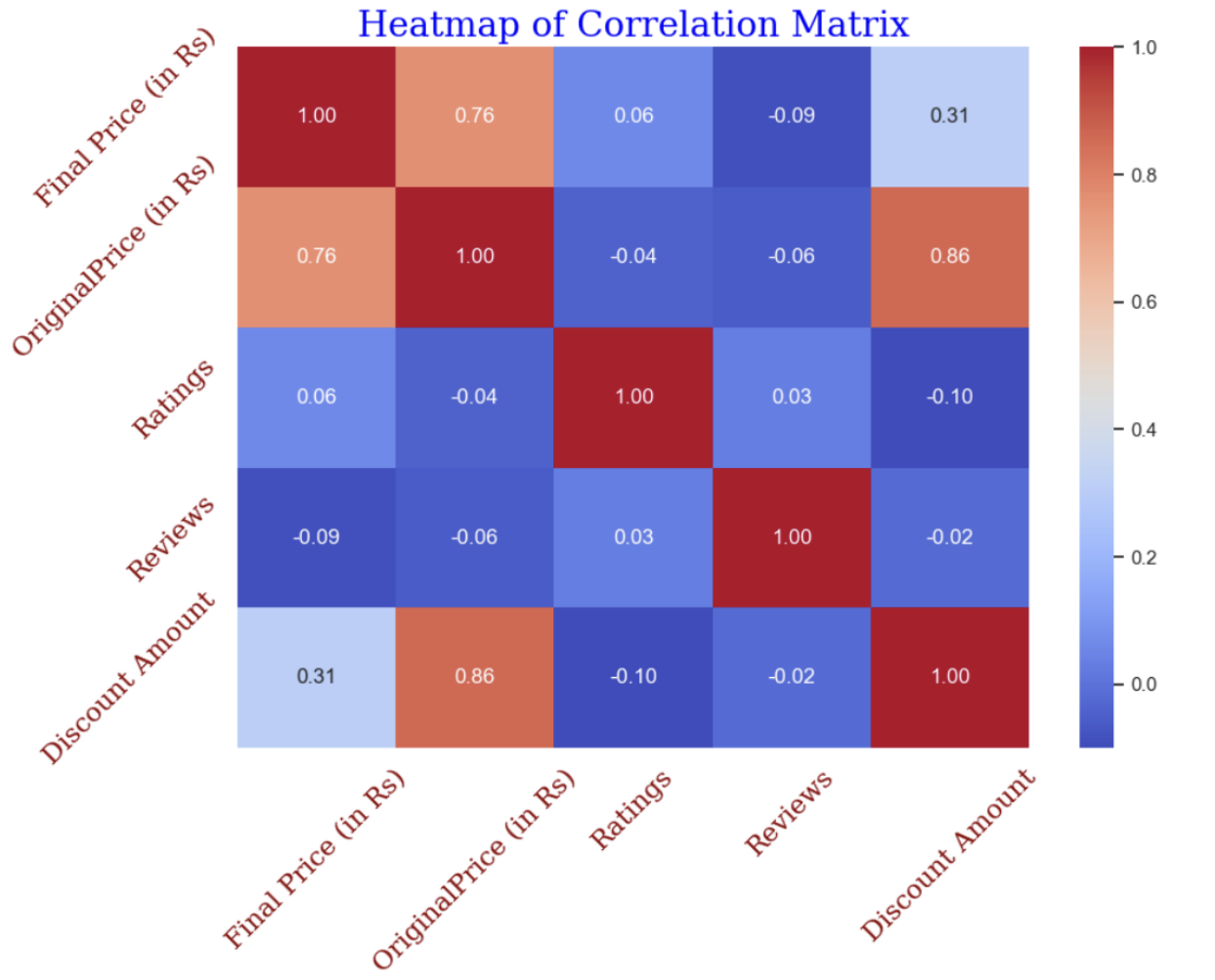

QQ-plot of Discount Amount
Sample Quantities
Theoretical Quantiles

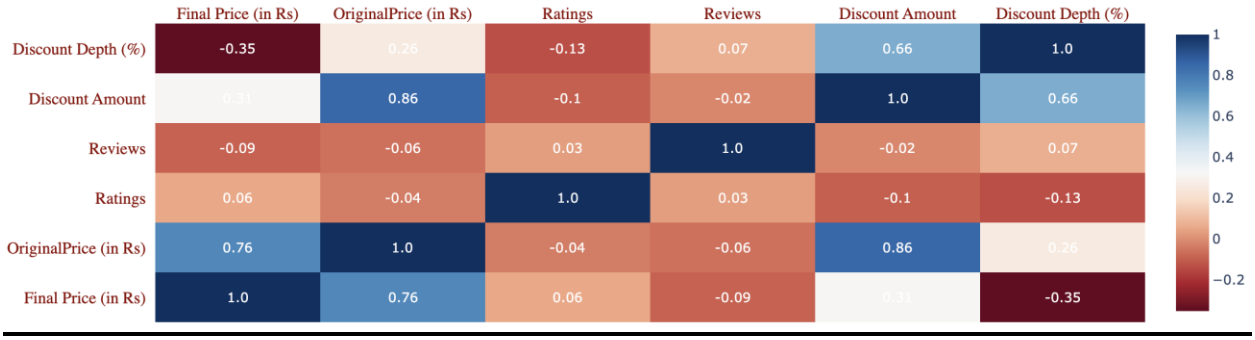

QQ-plot of Transformed Data
Sample Quantities
Theoretical Quantiles

NOTE: The Explanation for the graphs used in Normality Test & Transformation section of dashboard can be found here.

Heatmap & Pearson correlation coefficient matrix

	Final Price (in Rs)	OriginalPrice (in Rs)	Ratings	Reviews	Discount Amount
Final Price (in Rs)	1.00	0.76	0.06	-0.09	0.31
OriginalPrice (in Rs)	0.76	1.00	-0.04	-0.06	0.86
Ratings	0.06	-0.04	1.00	0.03	-0.10
Reviews	-0.09	-0.06	0.03	1.00	-0.02
Discount Amount	0.31	0.86	-0.10	-0.02	1.00





Observations:

- There is a strong positive correlation (0.76) between the Final Price and Original Price, suggesting that items with higher original prices tend to have higher final prices after discounts are applied.
- The Original Price shows a very strong positive correlation (0.86) with the Discount Amount, which indicates that higher-priced items tend to have larger absolute discounts.
- The correlations between Ratings and other variables are very weak, close to 0, implying no significant linear relationship.
- Reviews also show very weak correlations with other variables, suggesting that the number of reviews is independent of prices and discount amounts.
- A moderate positive correlation (0.31) between Final Price and Discount Amount could be due to the fact that after a certain amount of discount is applied, the final price reflects the reduced price from the original, which still relates to the magnitude of the original price.

Heatmap Visualization:

The heatmap provides a visual representation of the correlation matrix, with color intensity indicating the strength of the correlation between variables. Blue indicates a positive correlation, while red indicates a negative correlation. The stronger the color, the stronger the correlation.

- The correlation of 1.00 along the diagonal represents the perfect positive correlation of each variable with itself.
- The shades of blue between Final Price and Original Price, as well as between Original Price and Discount Amount, visually reinforce the strong positive correlations identified in the table.
- Red shades in the rest of the heatmap suggest weaker correlations, with some variables showing almost no relationship as indicated by the lighter colors.

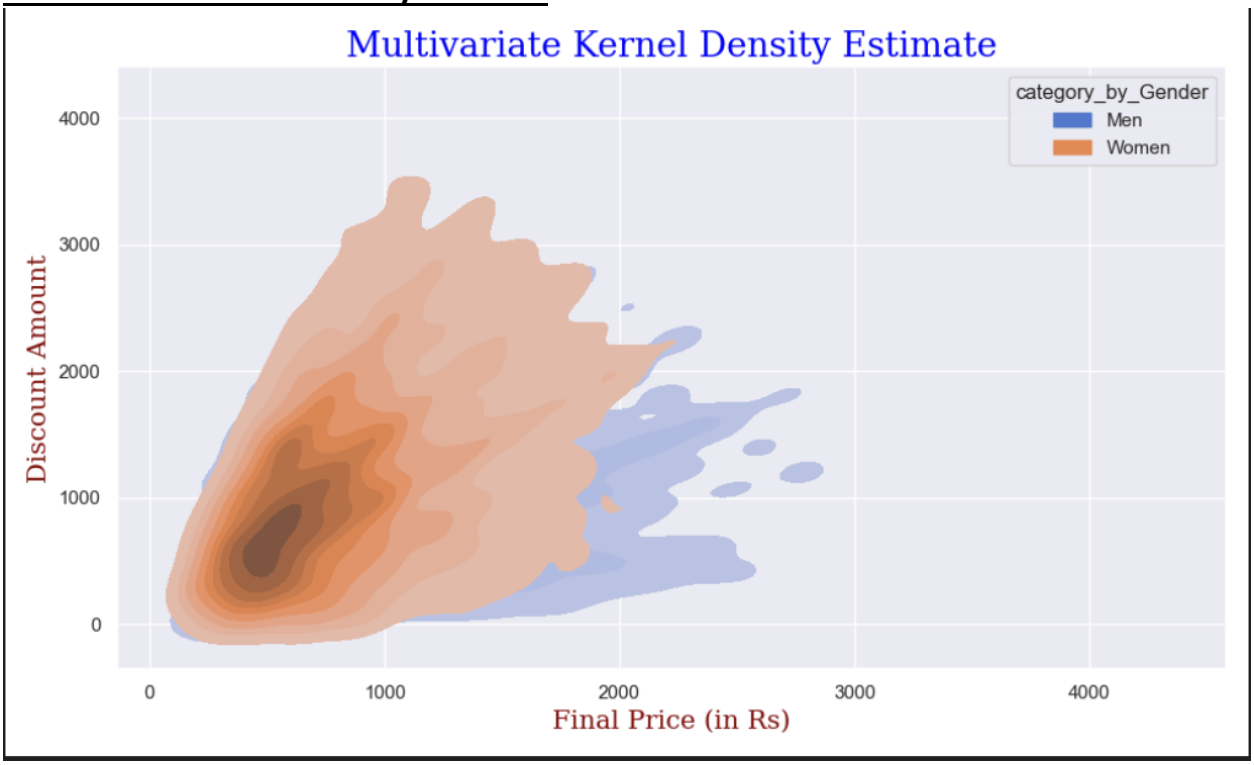
NOTE: The Explanation for the graphs used in Correlation Coefficient section of dashboard can be found here.

STATISTICS

Statistical Description: The statistical description of the cleaned dataset is obtained using myntra.**describe**().

	Product_id	DiscountPrice (in Rs)	OriginalPrice (in Rs)	Ratings	Reviews
count	526,564.00	333,406.00	526,564.00	190,412.00	190,412.00
mean	15,069,387.01	1,237.44	2,414.07	4.09	61.99
std	3,225,709.70	1,052.06	1,916.96	0.49	125.71
min	27,399.00	127.00	99.00	1.00	0.00
25%	13,880,530.00	659.00	1,299.00	3.90	8.00
50%	15,971,057.00	952.00	1,999.00	4.20	18.00
75%	17,347,414.50	1,469.00	2,899.00	4.40	52.00
max	18,464,352.00	27,996.00	90,000.00	5.00	999.00

Multi-Variate Kernel Density Estimate:



Observations:

There is a significant concentration of density for men's items around the lower range of final prices and discount amounts. This suggests that most of the men's items are clustered in a lower price range with moderate discounts.

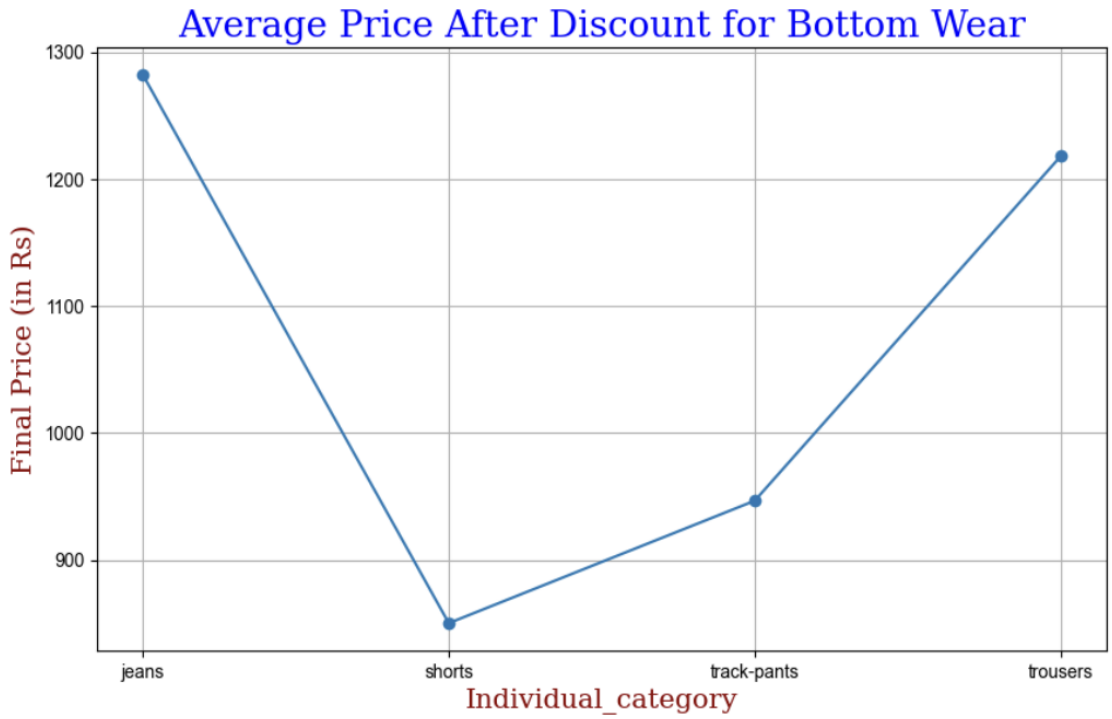
The density for women's items is spread over a larger range of final prices and discount amounts, with the highest density noted at higher values than for men. This indicates that women's items are generally listed at higher prices and with higher discounts compared to men's items.

There is an area of overlap between the two densities, indicating that there are some items for both men and women that have similar price and discount characteristics.

There are sparse regions (areas with light color) in the plot, indicating fewer items with those price and discount characteristics.

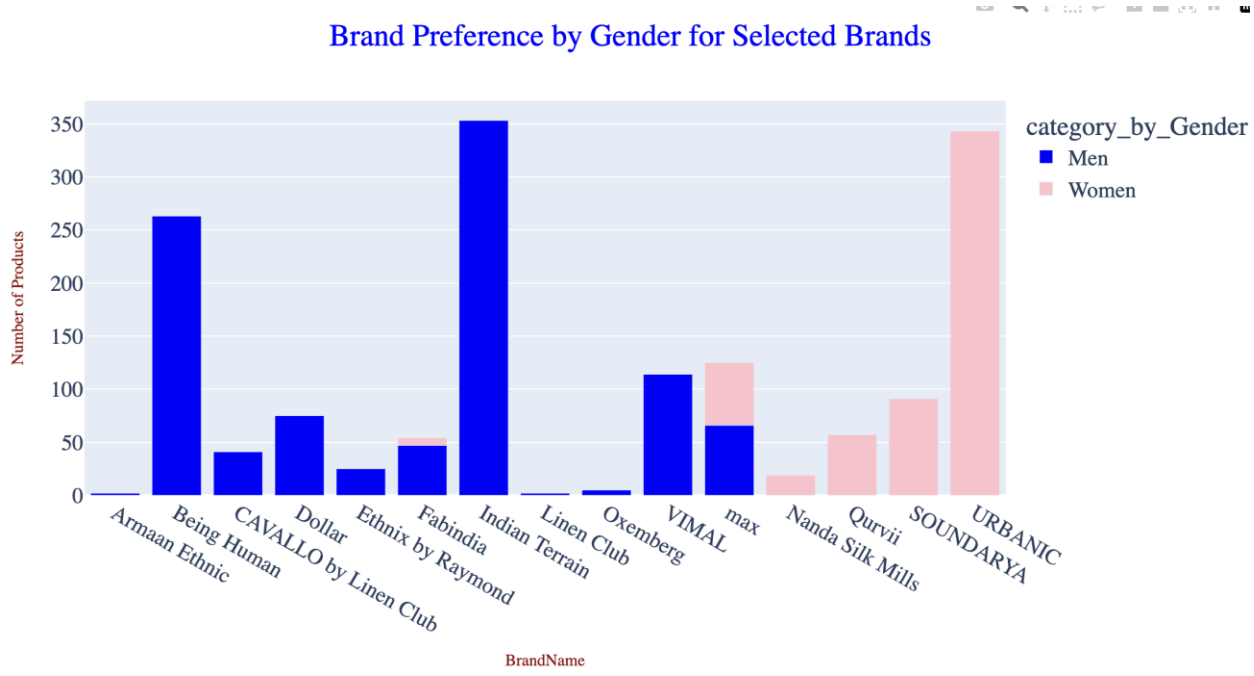
DATA VISUALIZATION

Line Plot:

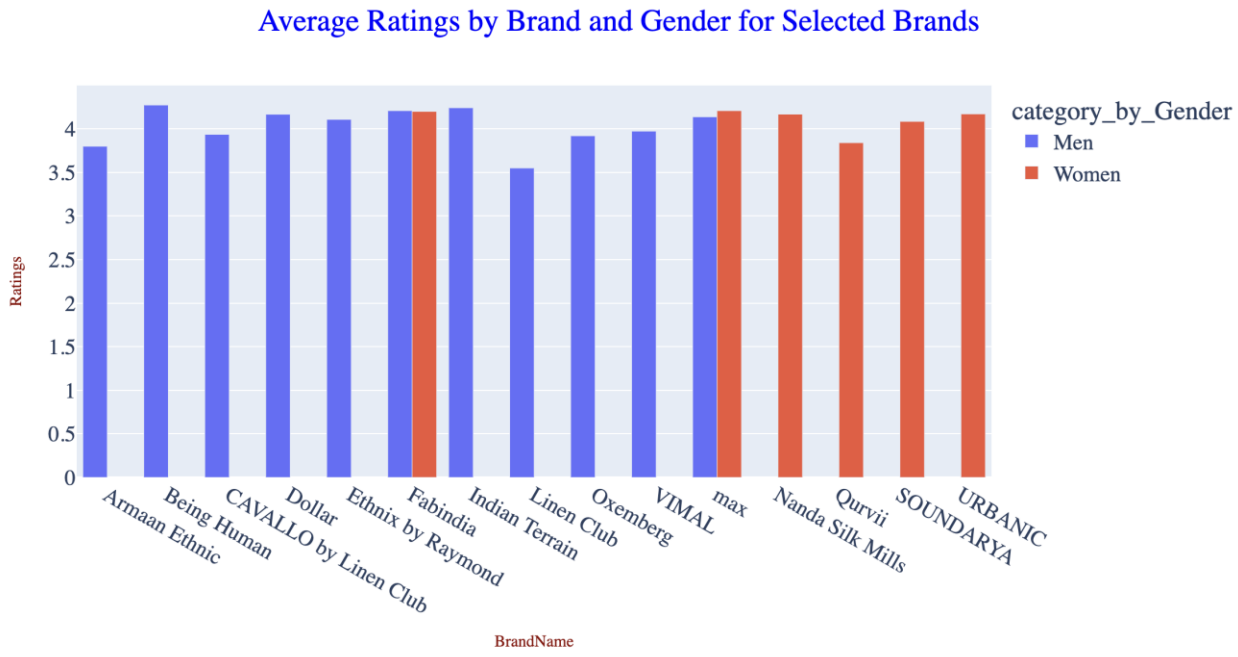


Bar Plot:

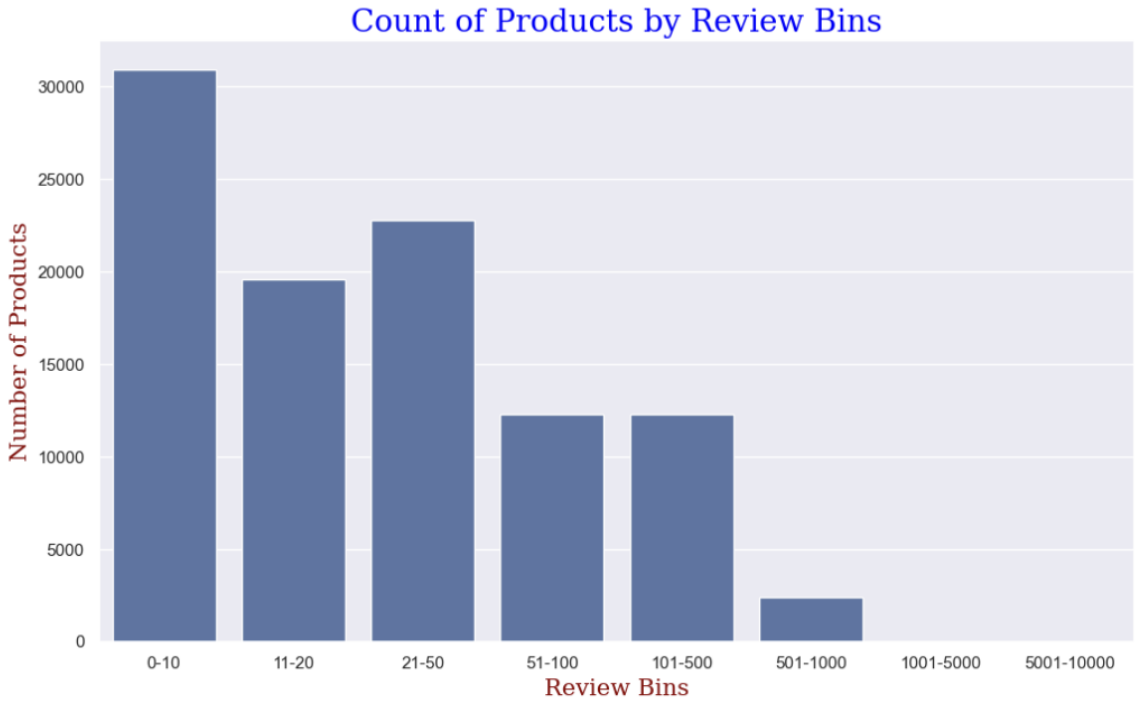
Bar Plot (Stacked):



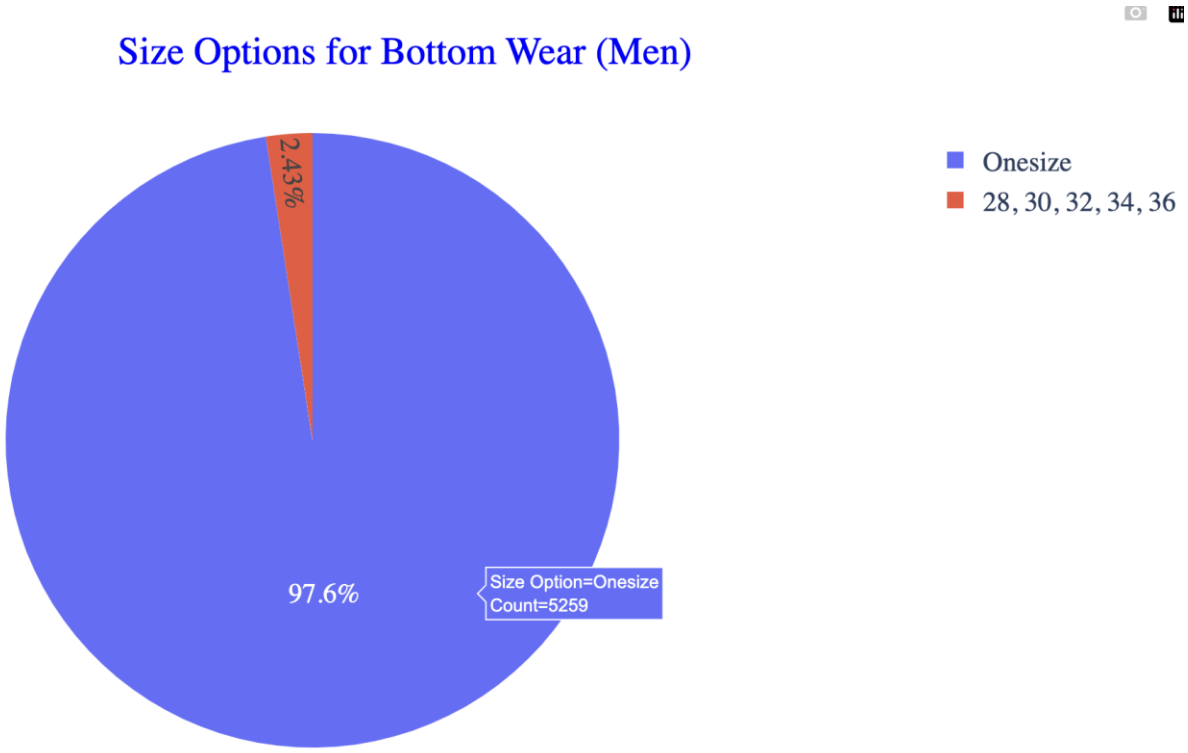
Bar Plot(Grouped) :



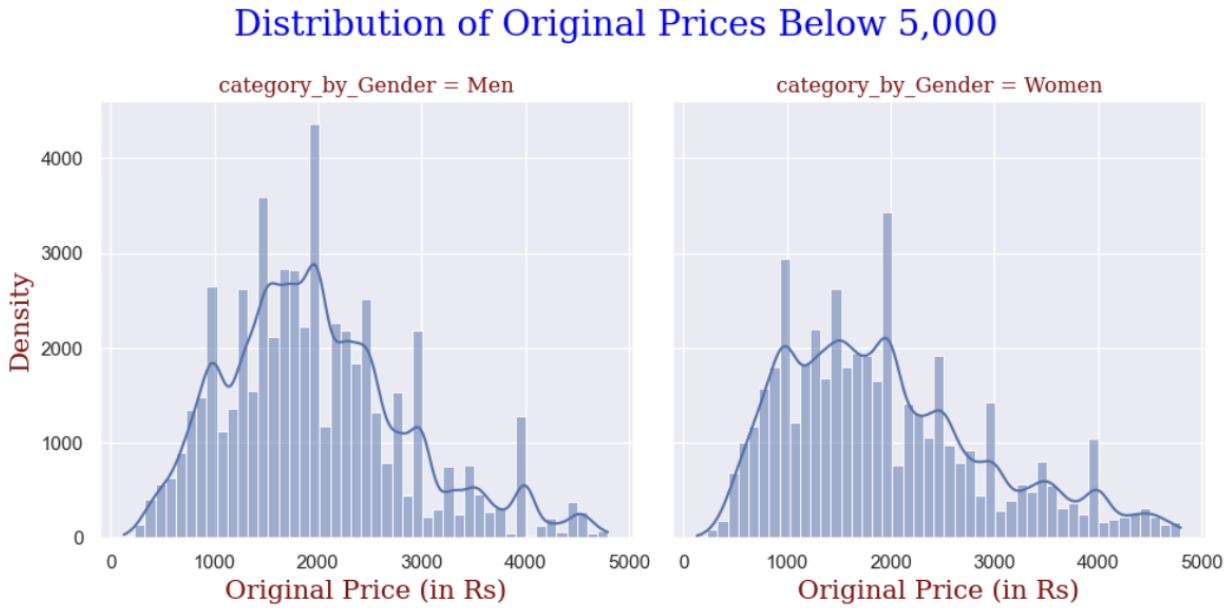
Count plot:



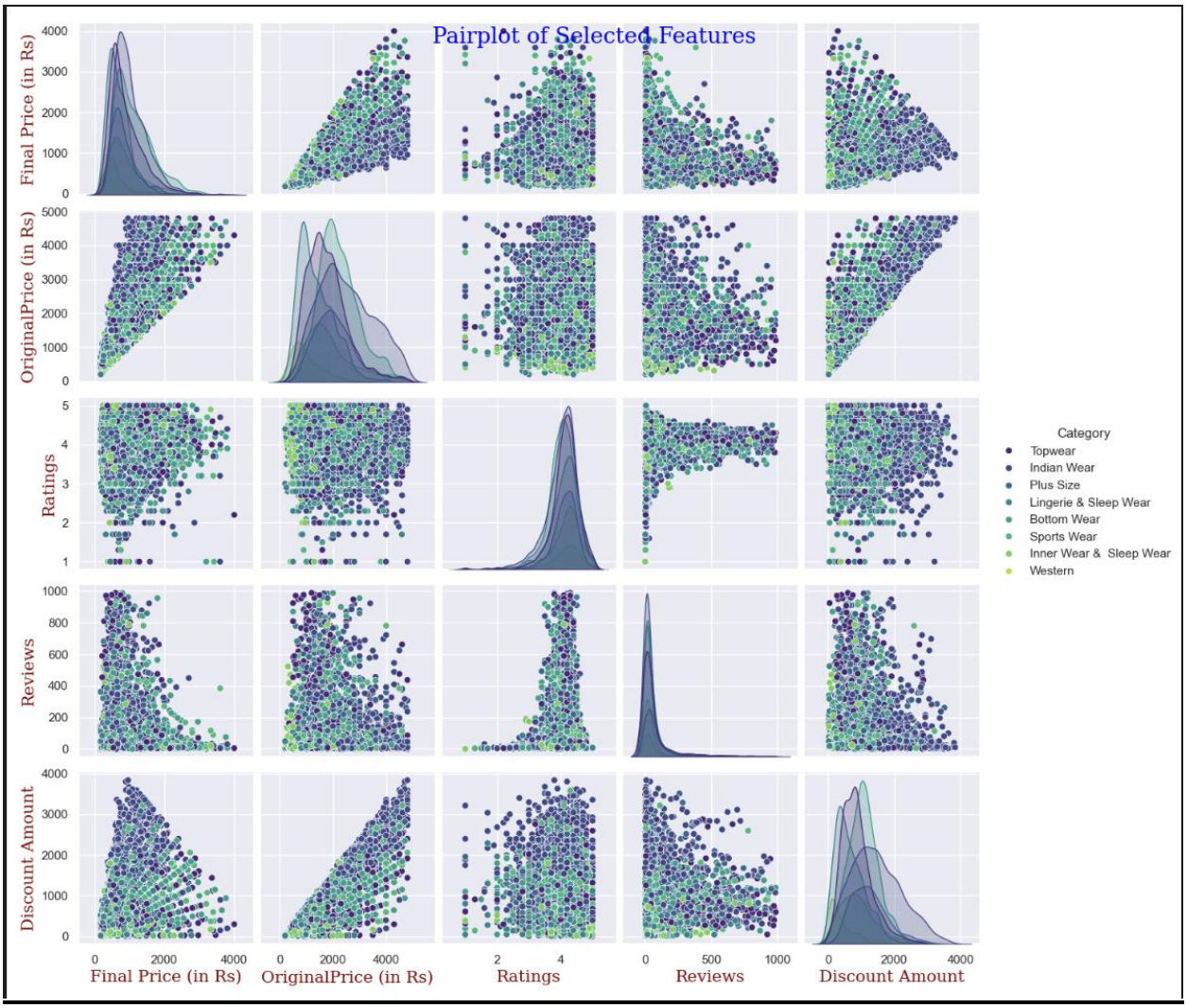
Pie chart:



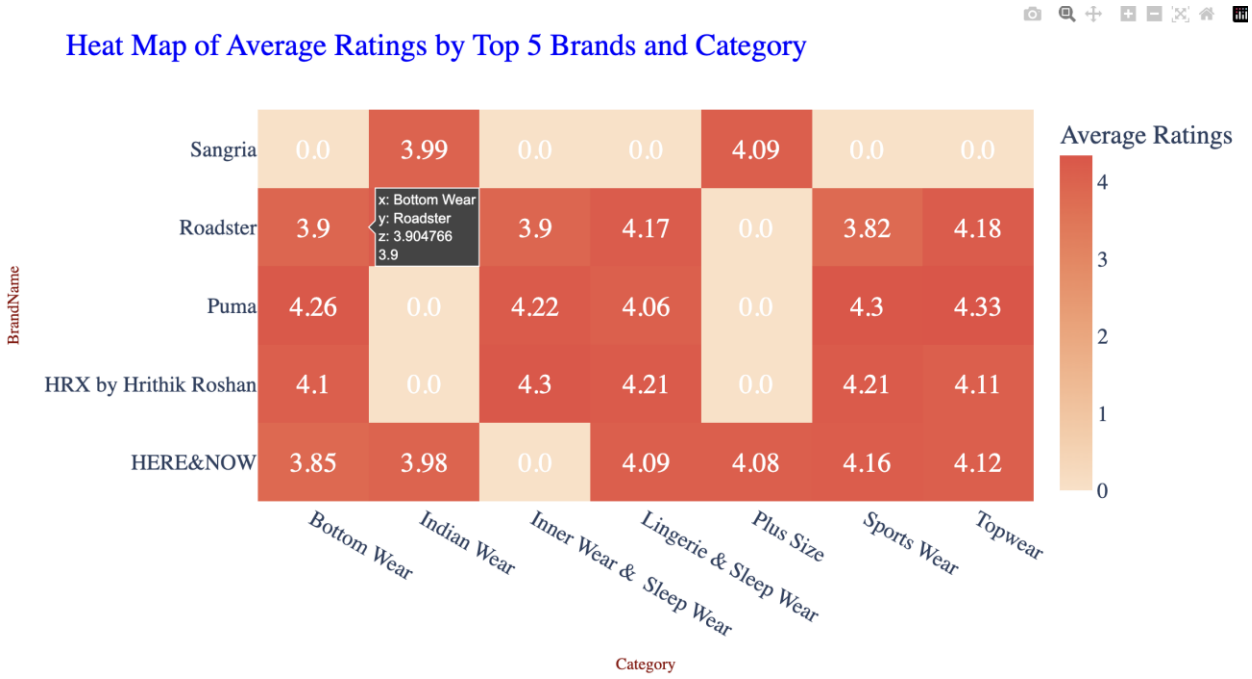
Dist Plot:



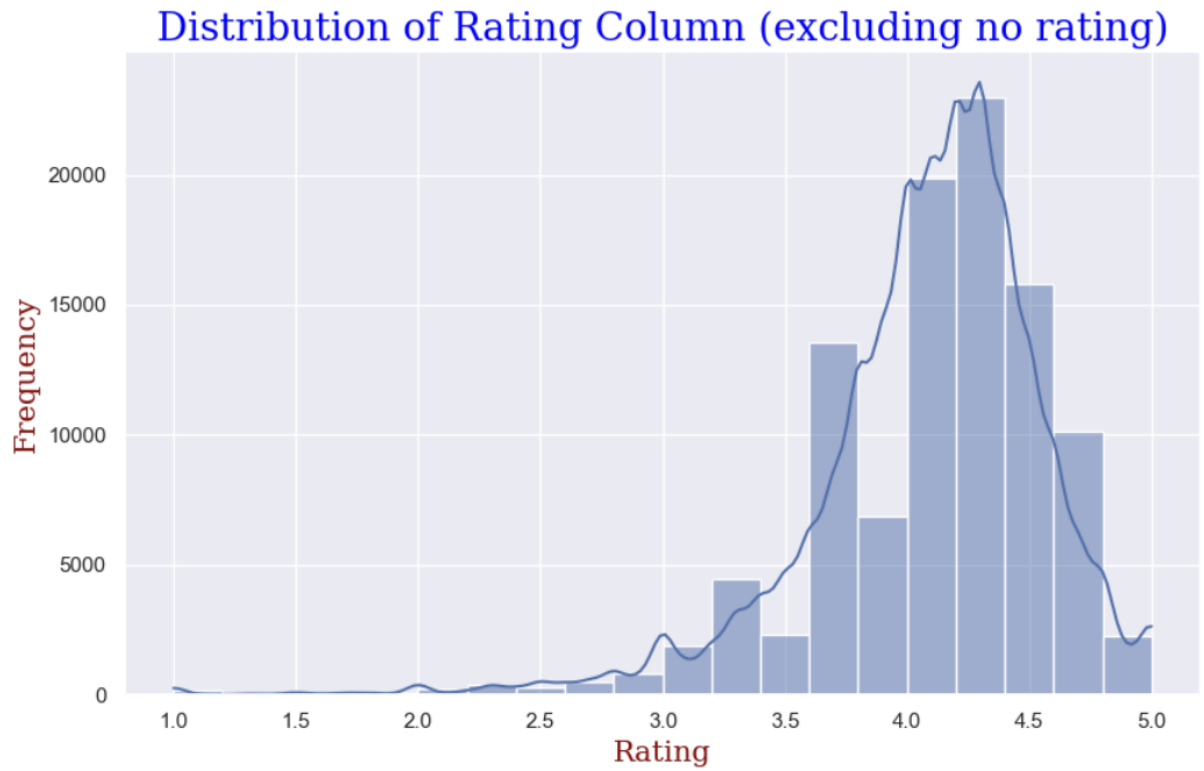
Pair Plot:



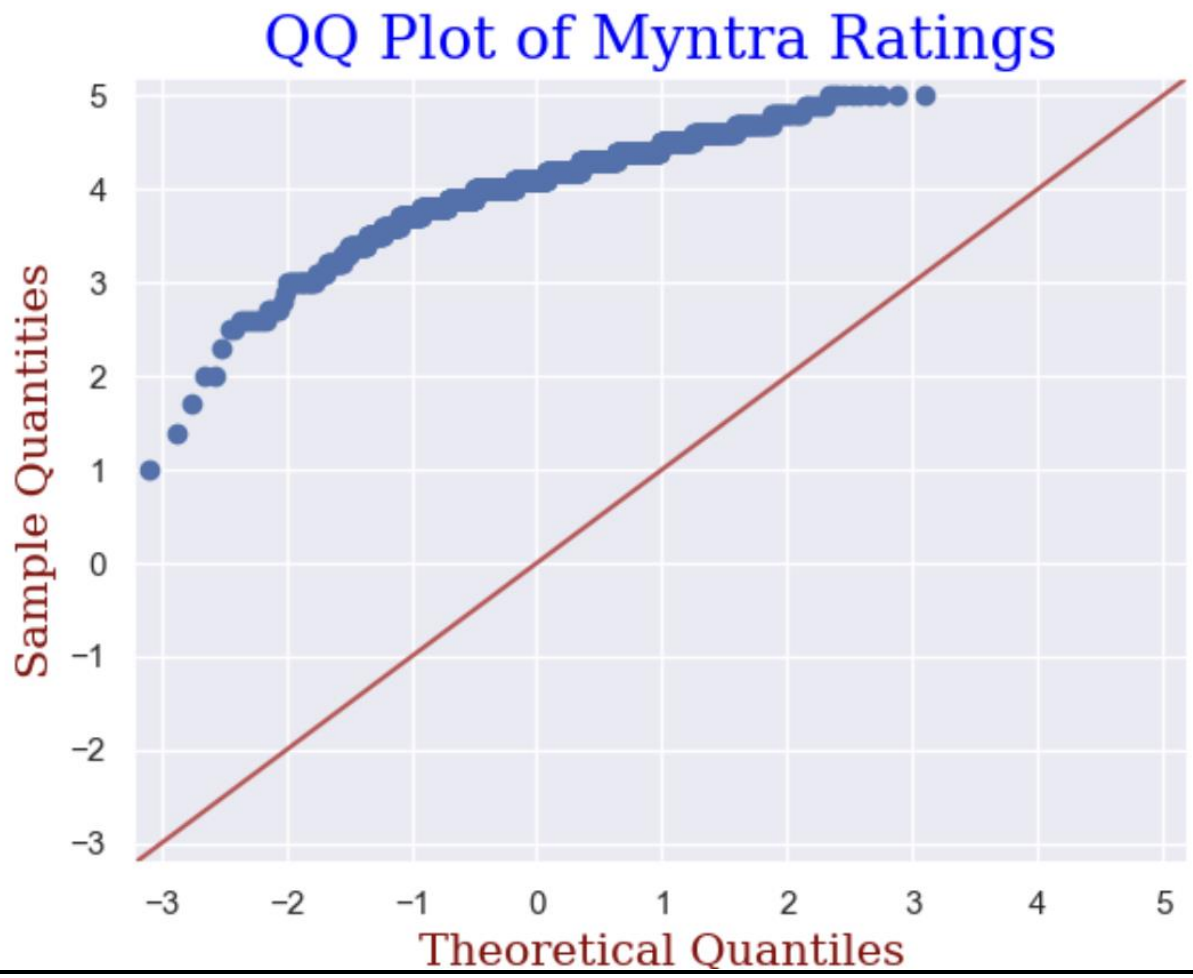
Heatmap with cbar:



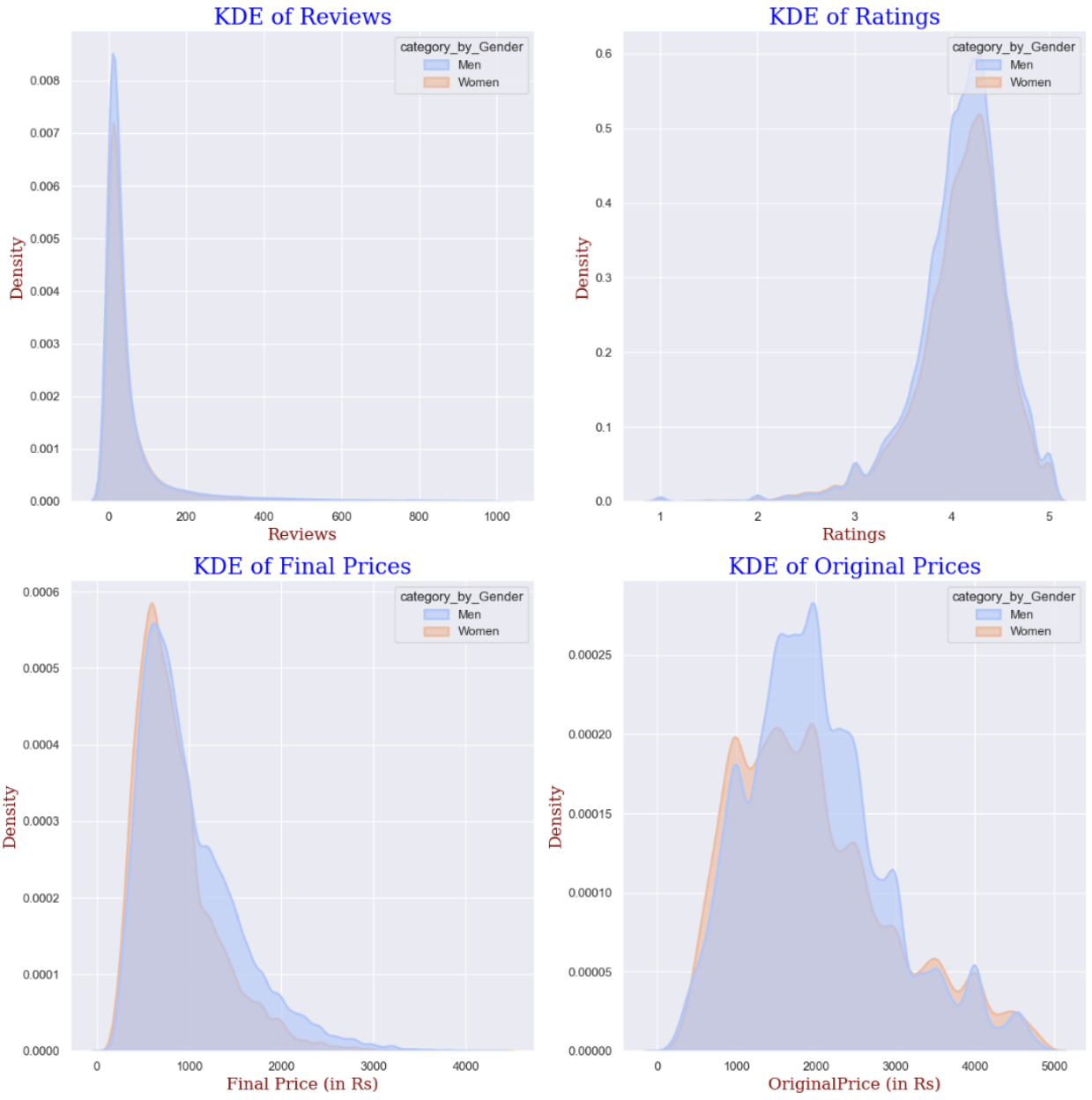
Histogram with KDE:



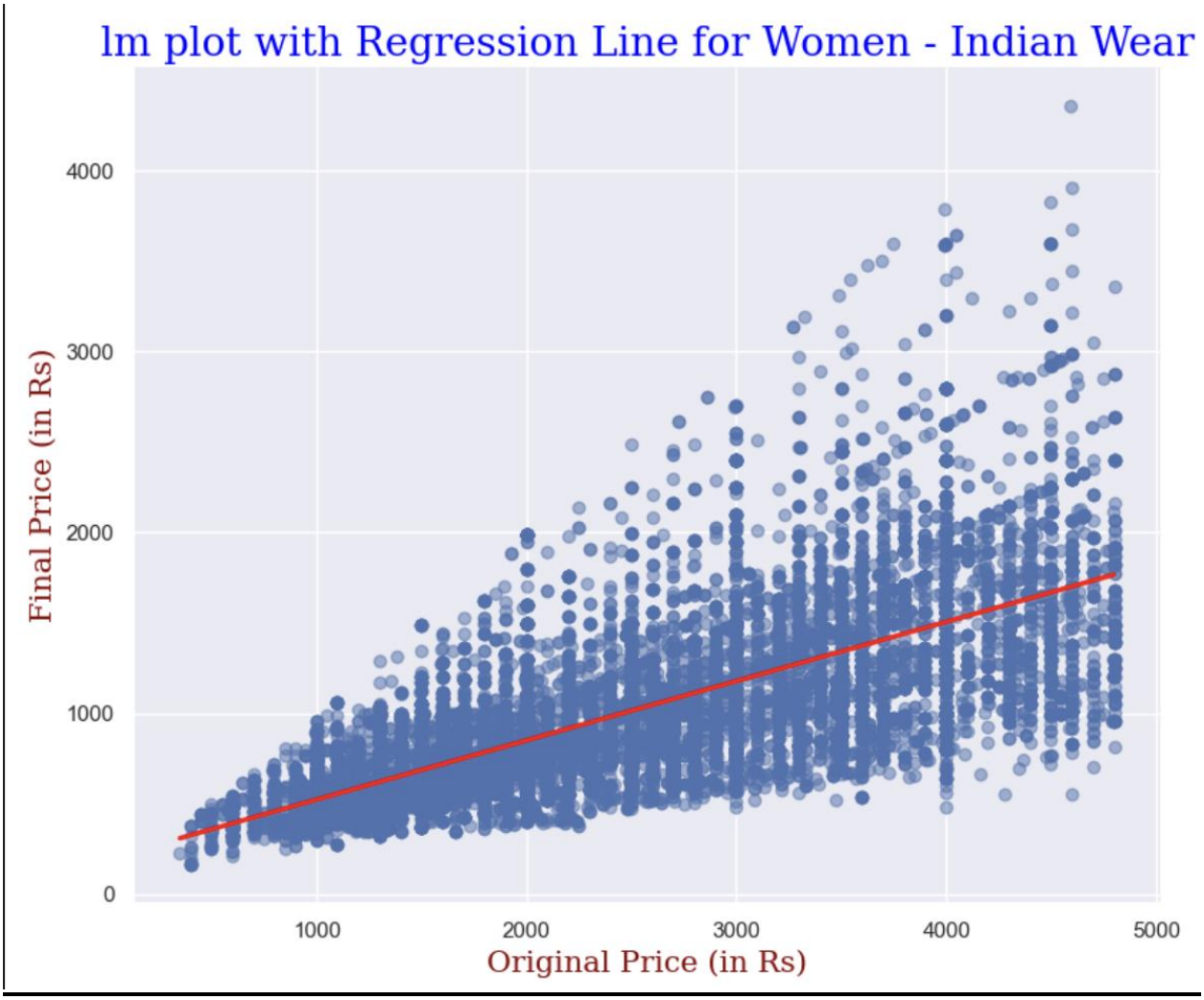
QQ-plot:



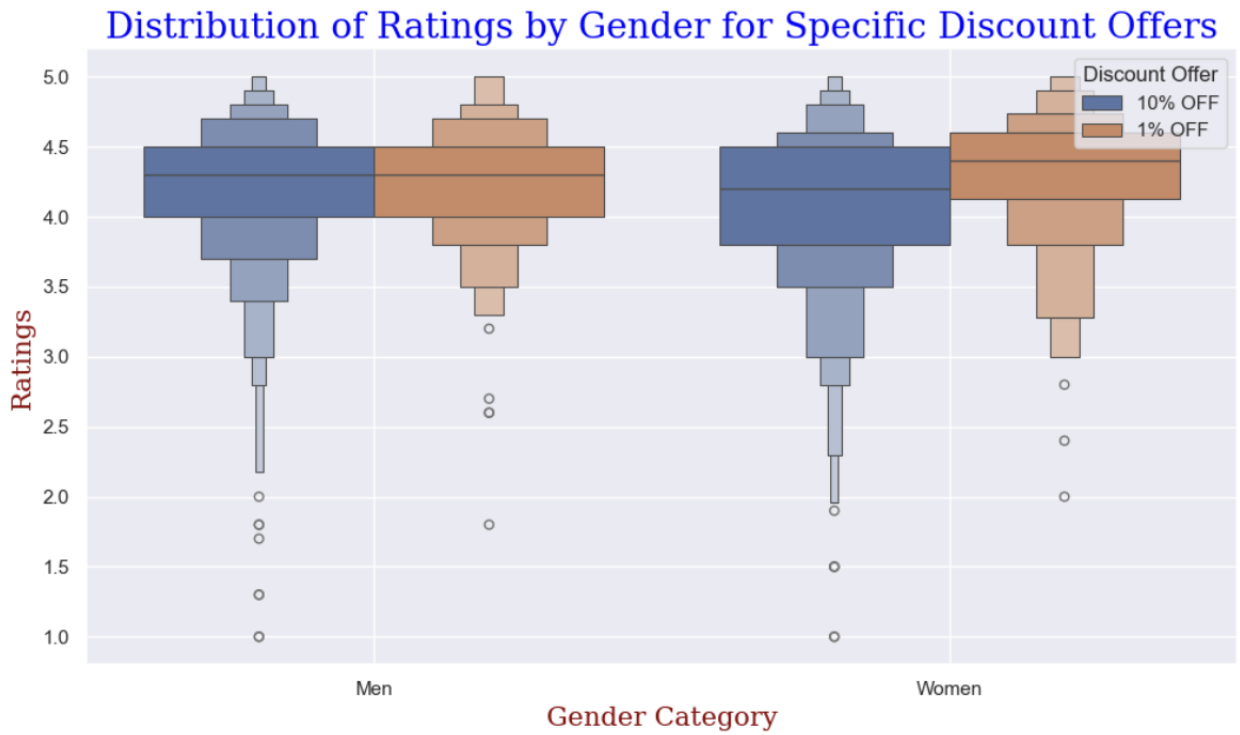
KDE plot will fill, alpha = 0.6, pick a palette, pick a linewidth:



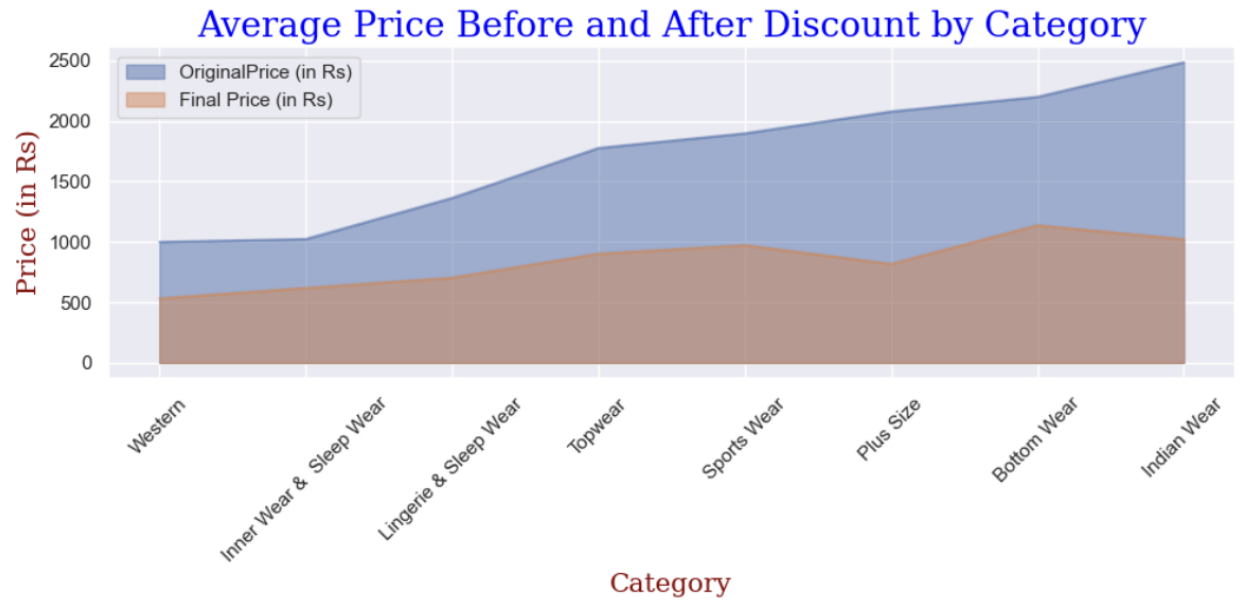
Im or reg plot with scatter representation and regression line:



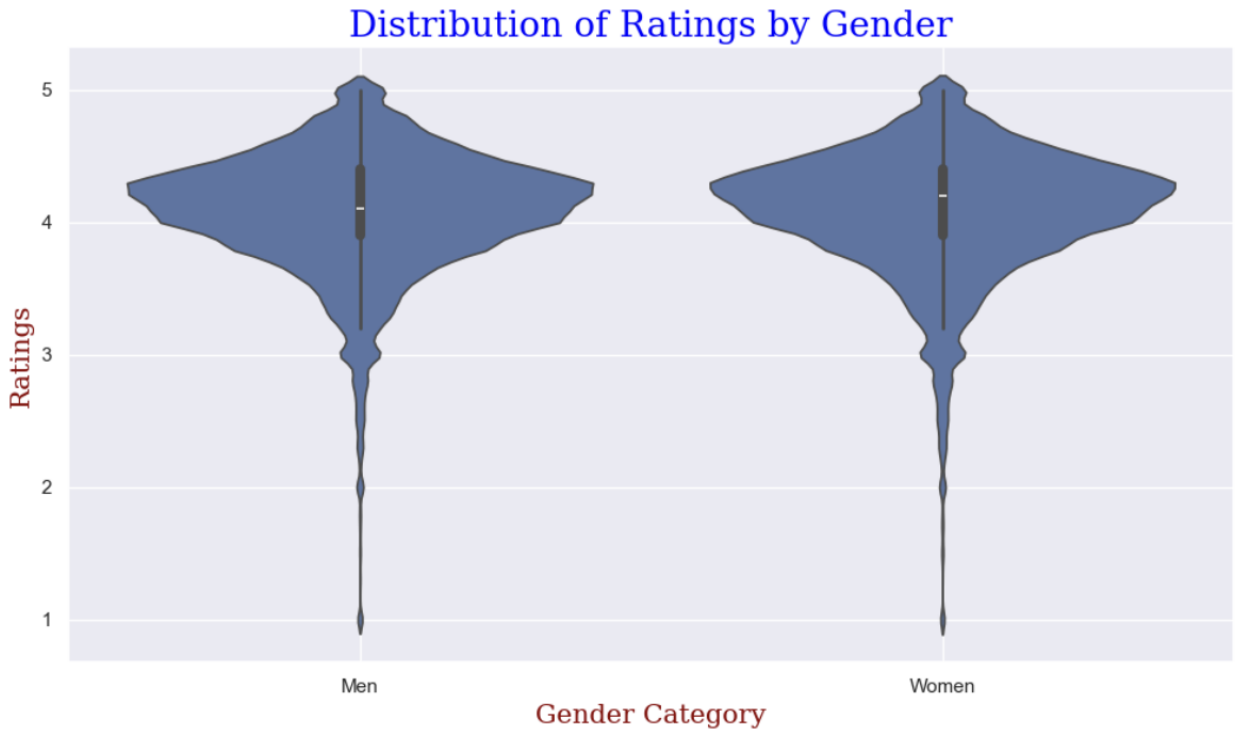
Multivariate Box or Boxen plot:



Area plot:



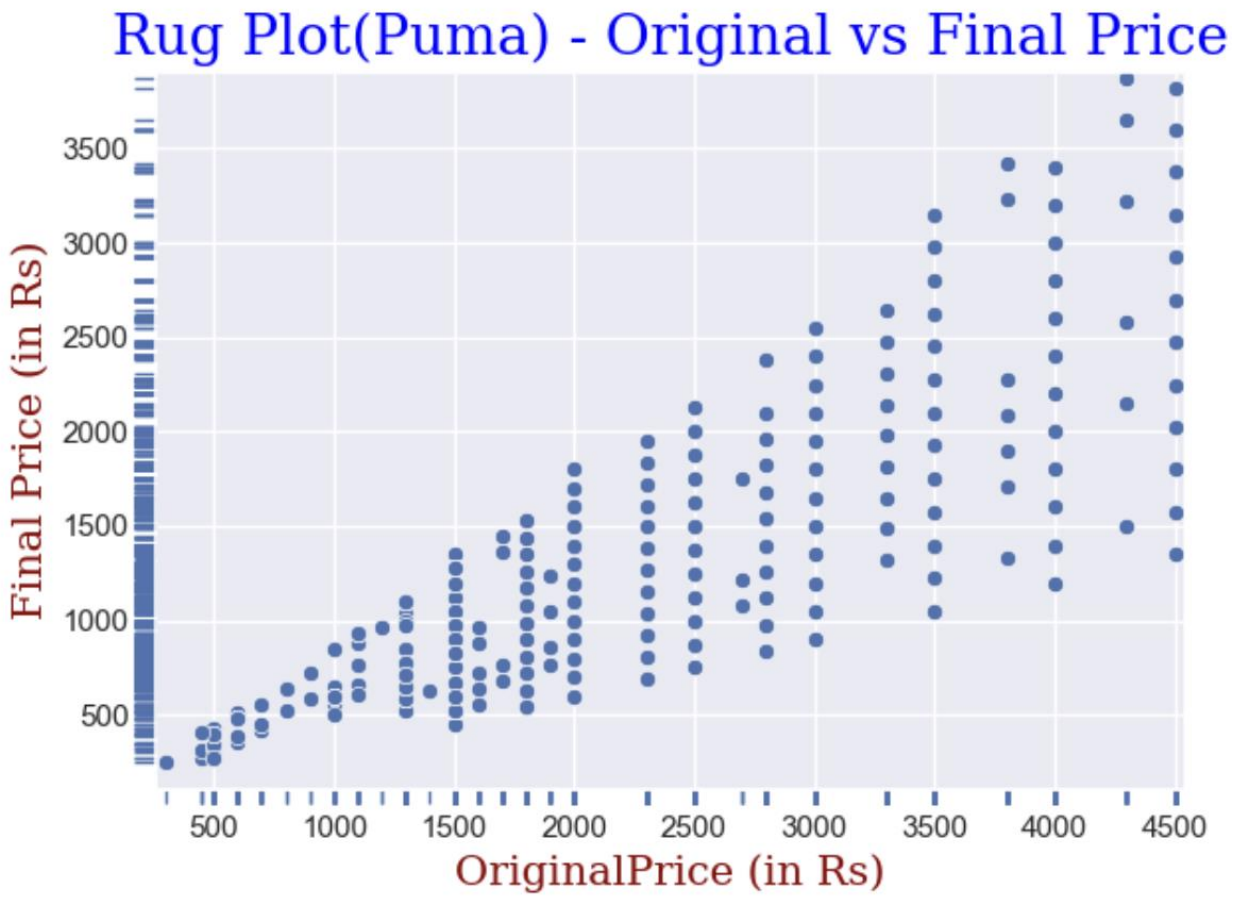
Violin plot:



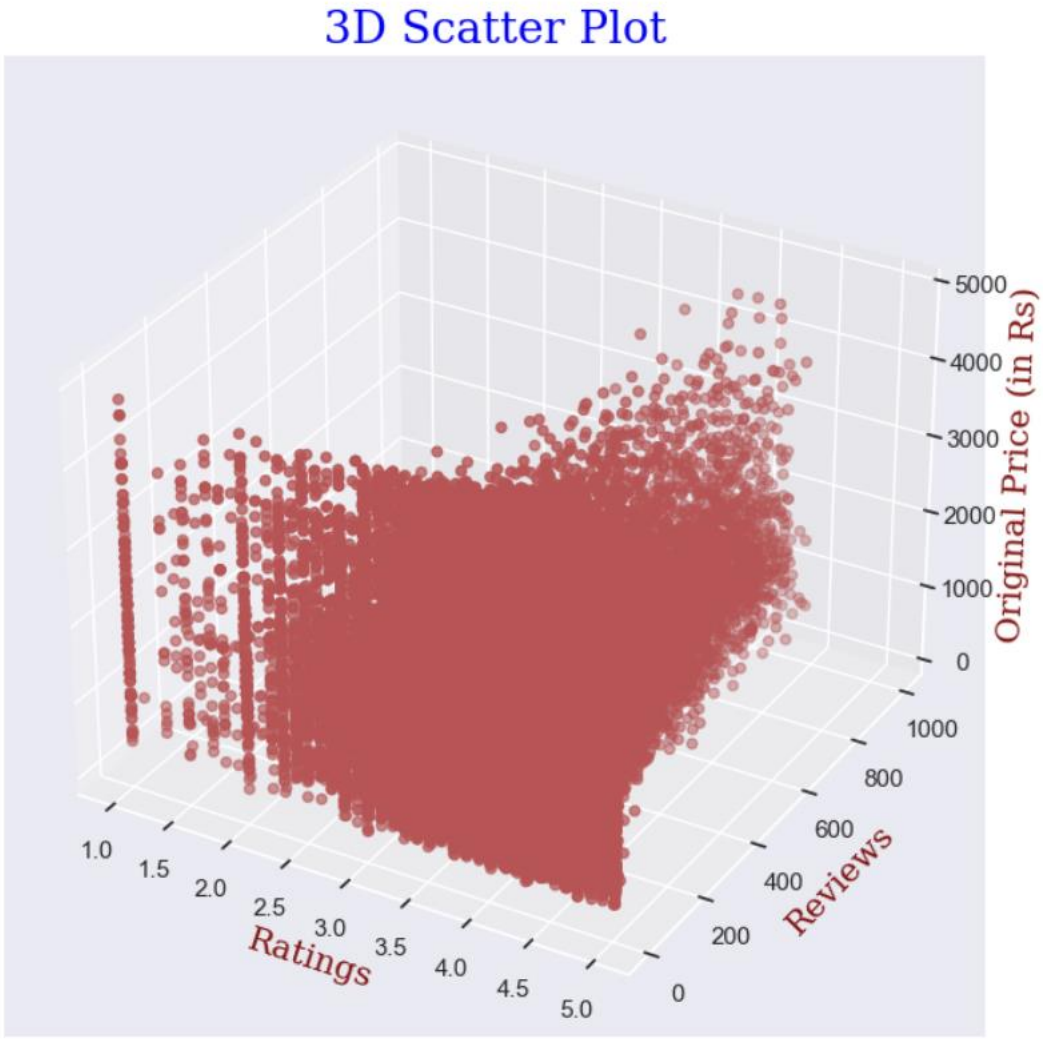
Joint plot with KDE and scatter representation:


Joint Plot - Ratings vs Final Price
Final Price (in Rs)
4000
3000
2000
1000
0
1
2
3
4
5
Ratings

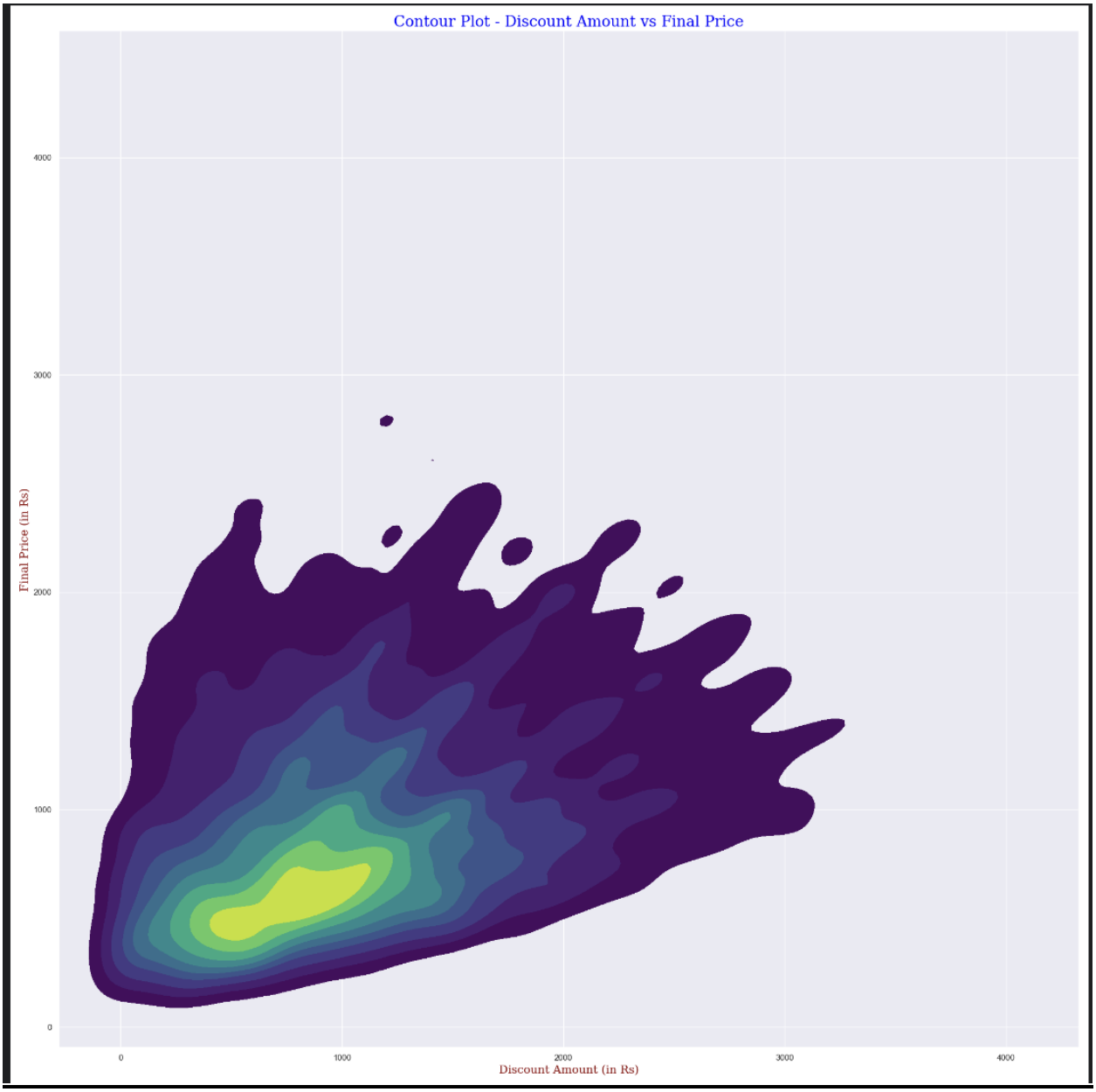
Rug plot:



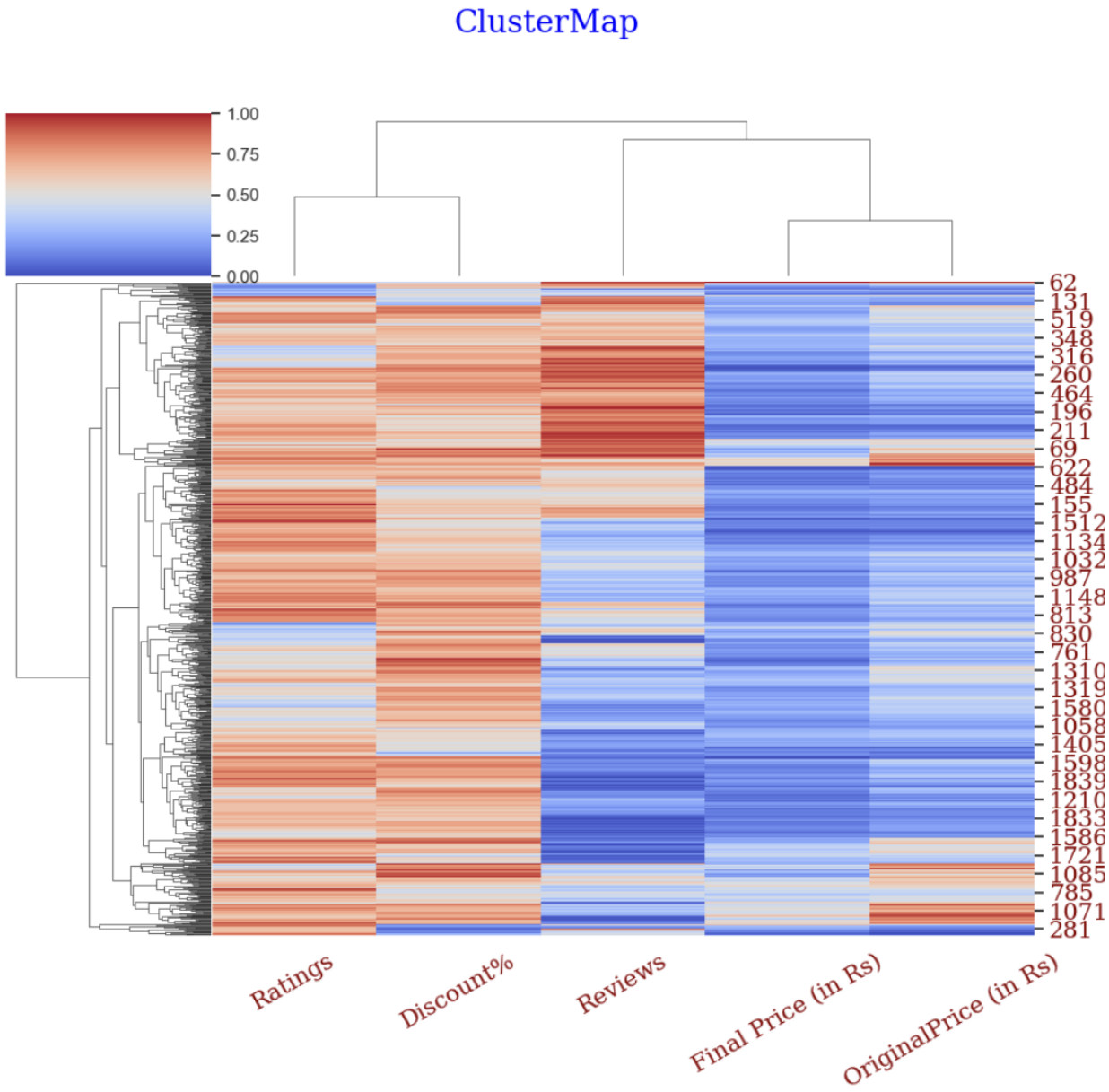
3D plot and contour plot:



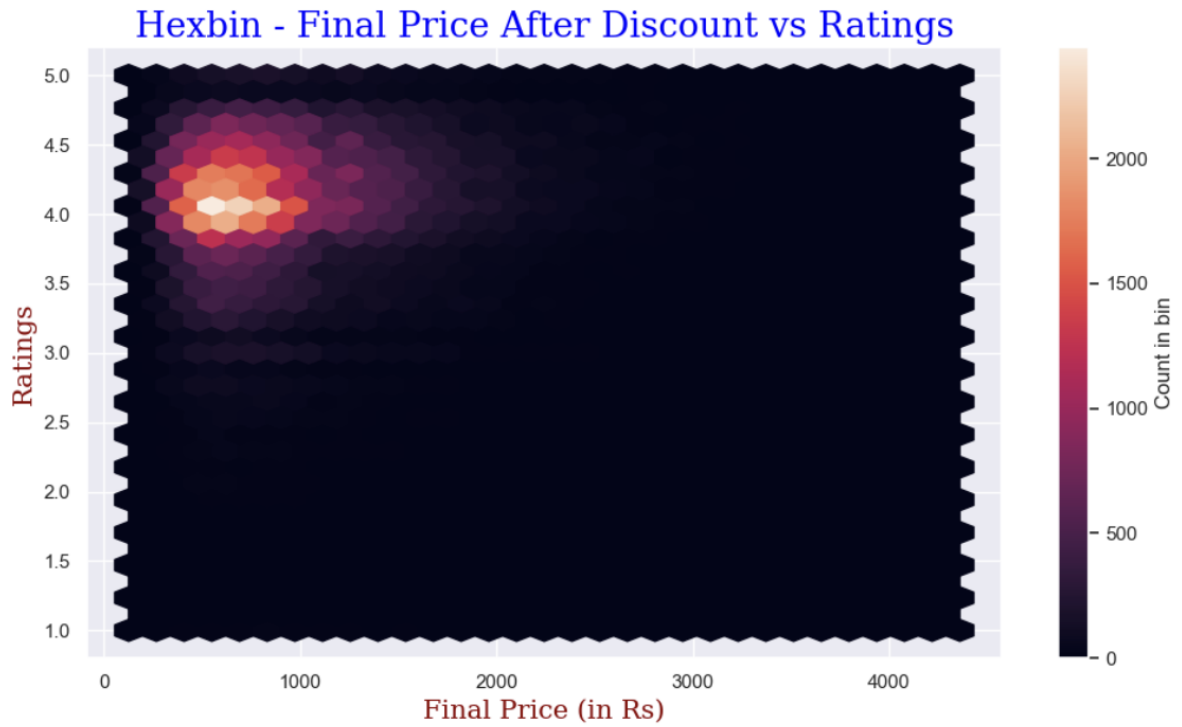
Contour Plot:


Contour Plot - Discount Amount vs Final Price
Final Price (in Rs)
Discount Amount (in Rs)
0
1000
2000
3000
4000

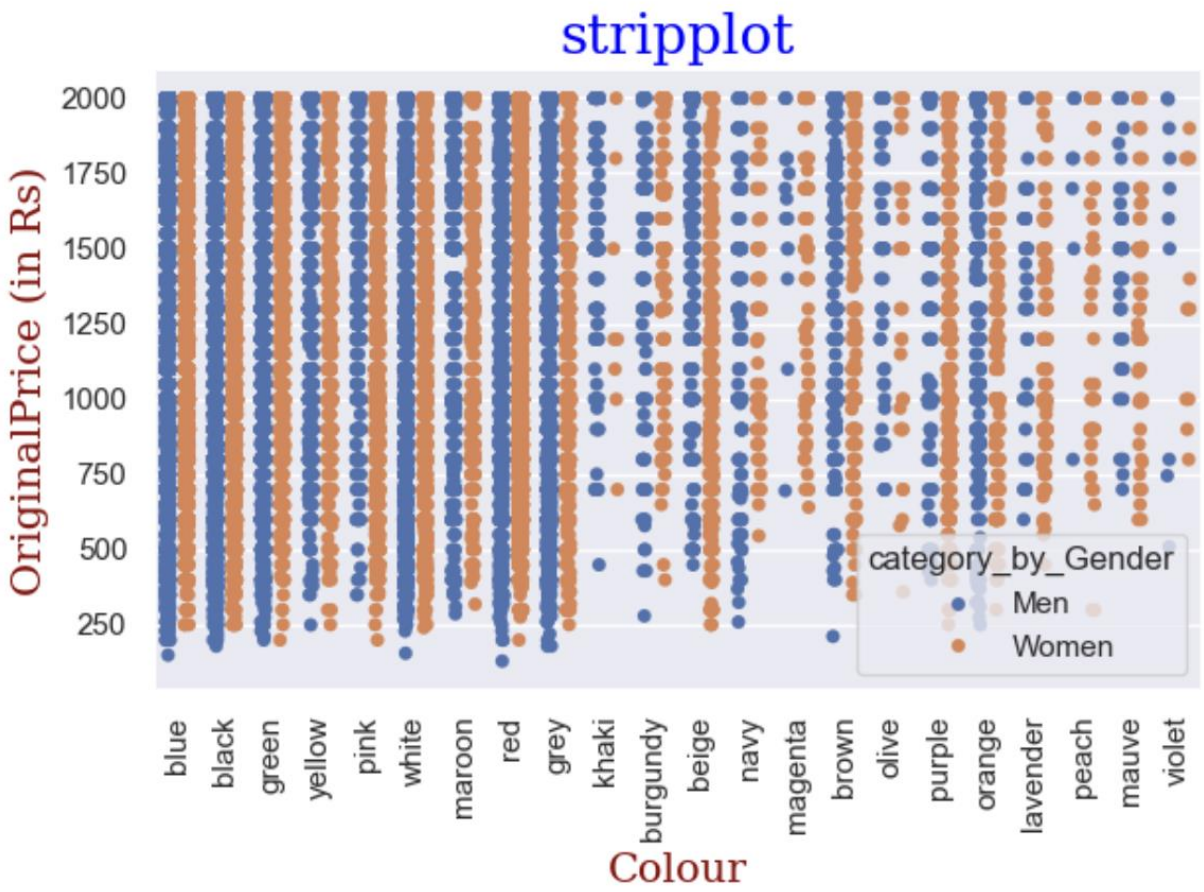
Cluster Map:


ClusterMap
1.00
0.75
0.50
0.25
0.00
62
131
519
348
316
260
464
196
211
69
622
484
155
1512
1134
1032
987
1148
813
830
761
1310
1319
1580
1058
1405
1598
1839
1210
1833
1586
1721
1085
785
1071
281
Ratings
Discount%
Reviews
Final Price (in Rs)
OriginalPrice (in Rs)

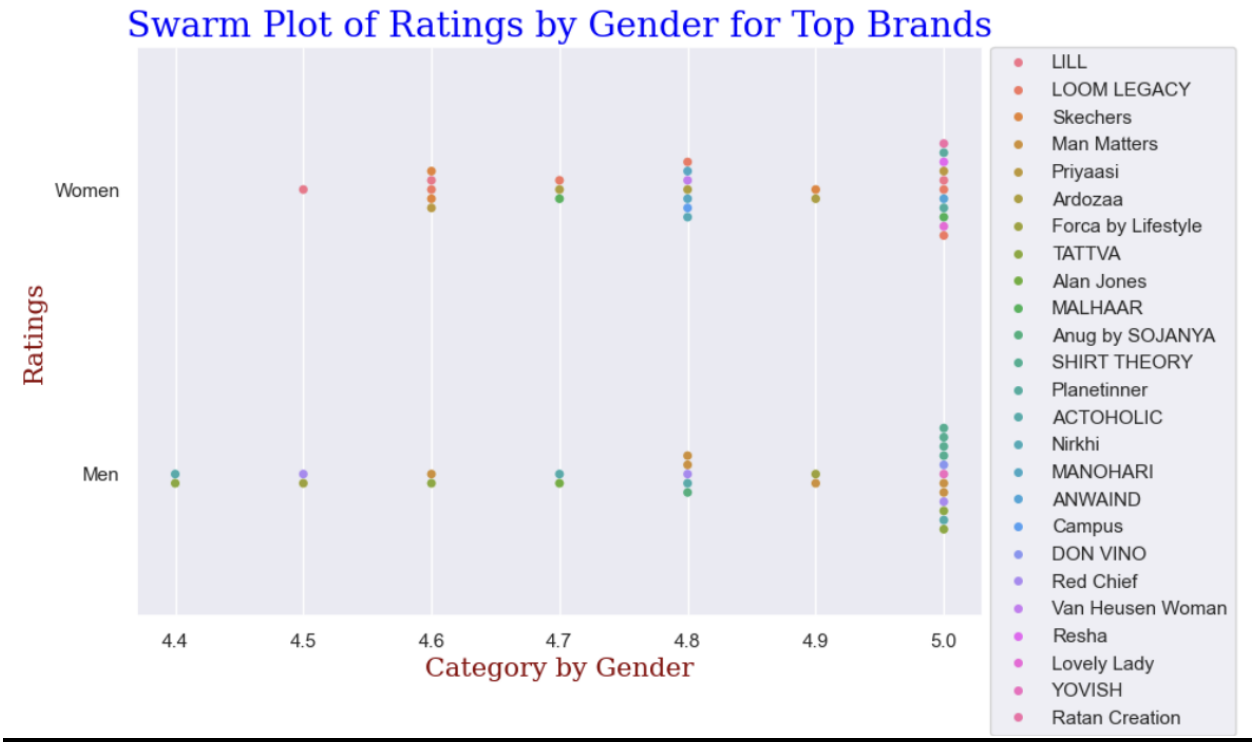
Hexbin:


Hexbin - Final Price After Discount vs Ratings
Ratings
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
1000
2000
3000
4000
Final Price (in Rs)
2000
1500
1000
500
0
Count in bin

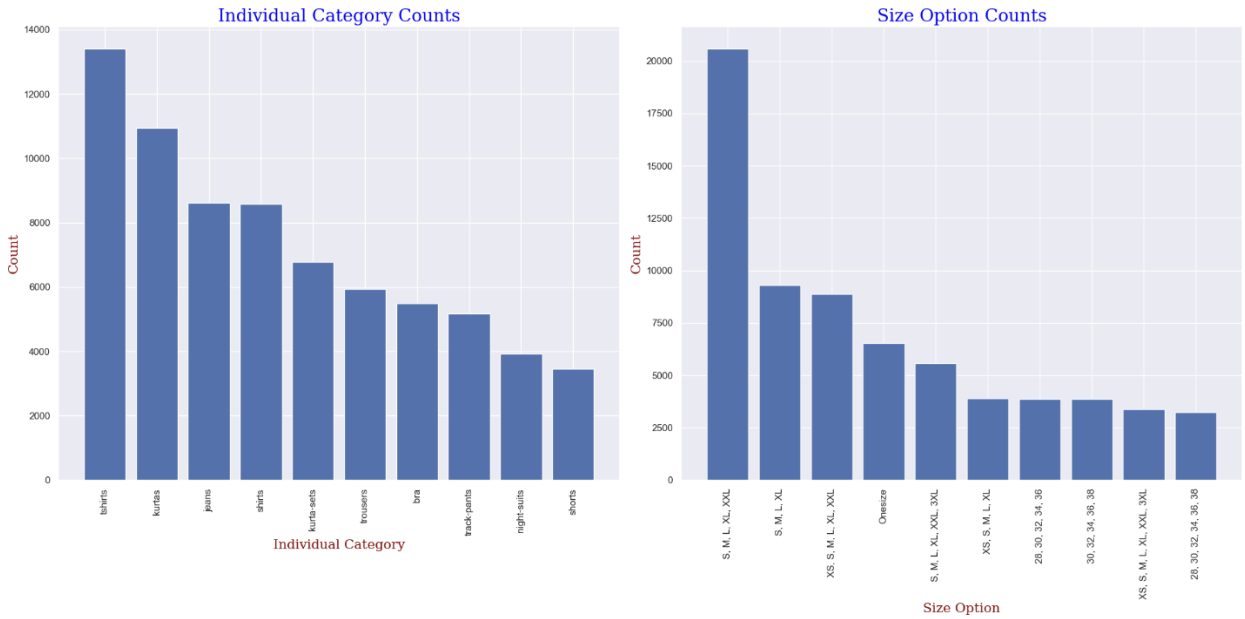
Strip Plot:


stripplot
OriginalPrice (in Rs)
2000
1750
1500
1250
1000
750
500
250
category_by_Gender
Men
Women
blue
black
green
yellow
pink
white
maroon
red
grey
khaki
burgundy
beige
navy
magenta
brown
olive
purple
orange
lavender
peach
mauve
violet
Colour

Swarm Plot:



SUBPLOTS:



OBSERVATION:

Individual Category Counts:

- This chart shows the count of items within individual clothing categories.
- tShirts appear to be the most common category, followed by kurtas and then jeans.
- Shirts, kurta-sets, trousers have a moderate count
- The least common categories are track pants, night suits and shorts, with shorts being the least common among the listed categories.

Size Option Counts:

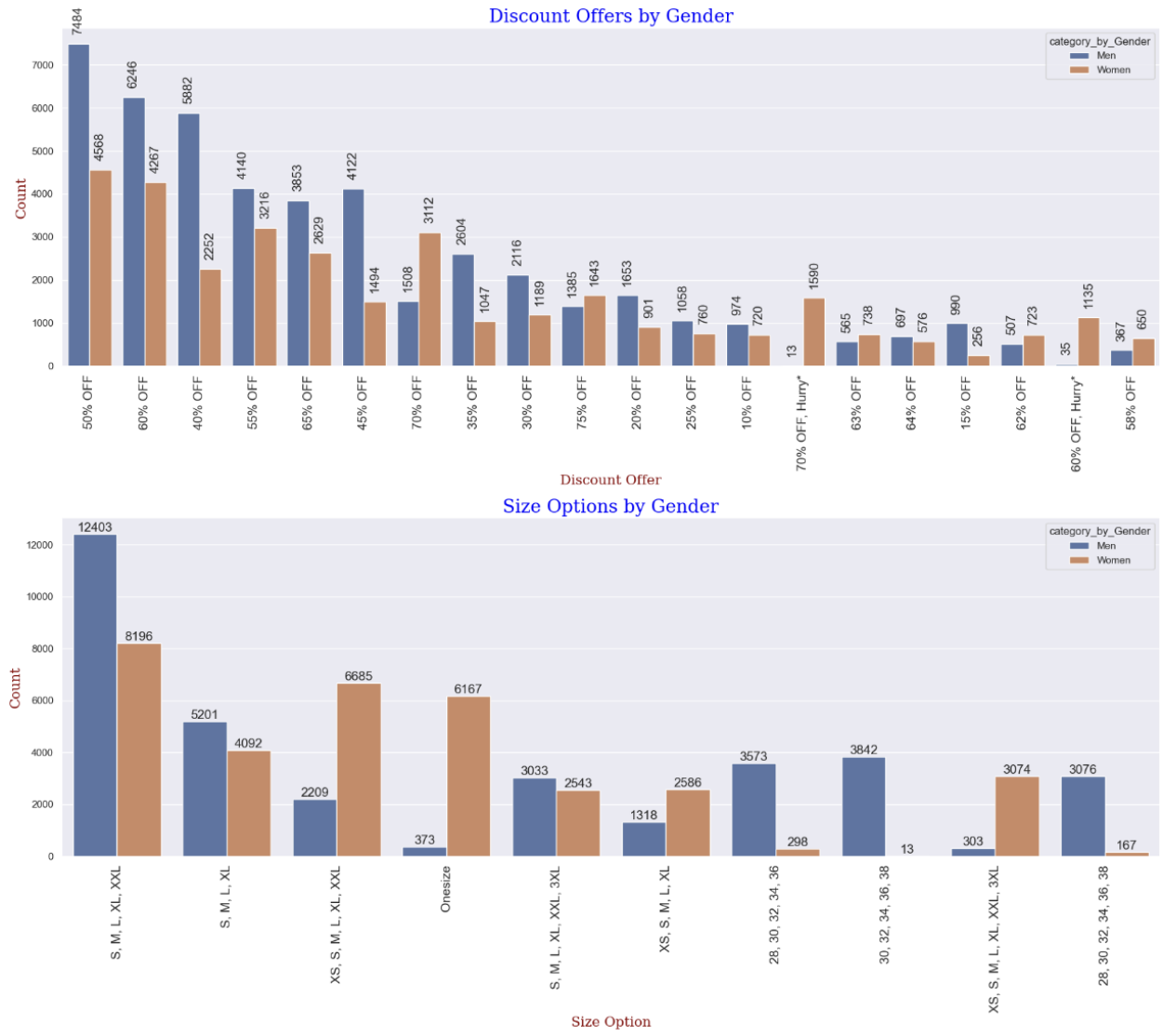
- This chart presents the count of different size options available.
- “S,M,L,XL,XXL is the most common size option, significantly outnumbering others, suggesting a large stock of items available in this combination of sizes.
- Specialized sizes such as XS,S,M,L,XL, XXL, and 3XL are less common, and there are even fewer items available in size-specific numbers like “28,30,32,34,36,38” etc.

OBSERVATION:

The image shows a collection of four pie charts, each representing the distribution of Individual product categories within four different clothing segments or Categories: Indian Wear, Bottom Wear, Topwear and Lingerie & Sleep Wear. Here are some observations based on the pie charts:

Indian Wear:
Kurtas constitute the majority, with 48.5% of the segment.
kurti-sets are the next significant category at 32.6%.
The remainder is made up of Sarees at 19.0%.
Bottom Wear:
Jeans are the most prevalent, taking up 45.6% of the segment.
Track pants represent 22.8%.
Trousers also takes a notable portion of 31.6%,
Topwear:
Shirts make up nearly half of the segment at 45.5%.
T-shirts are also a large category at 46.3%.
The smallest portion is sweatshirts at 8.2%.
Lingerie & Sleep Wear:
Bras have the largest share at 47.5% of the segment.
Nightdresses are the next significant category at 20.1%.
Night-suits are the smallest group at 32.4%.

Overall, each clothing segment displays a dominant category (kurtas, jeans, T-shirts, and bras respectively). These dominant categories could indicate the most popular or widely available items within each segment. The charts also show that for the Bottom Wear and Topwear segments, two categories (jeans and trousers for Bottom Wear, shirts and t-shirts for Topwear) occupy a almost similar proportion of the segment, suggesting a balanced variety within those segments. The distribution in each pie chart can help retailers and analysts understand consumer preferences or inventory distribution within these clothing segments.



OBSERVATIONS:

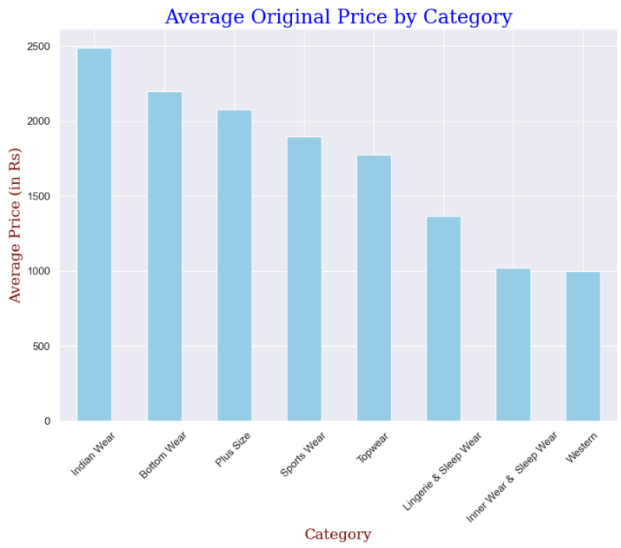
Discount Offers by Gender:

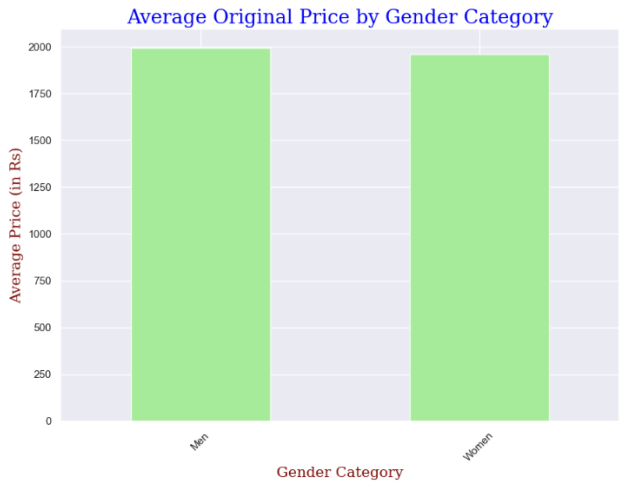
The chart compares the count of different discount offers received by men and women.
Men seem to receive more offers overall, with the highest counts for 50% OFF, 40% OFF, and 60% OFF discounts.
Women receive fewer discounts than men, with their counts peaking at the 50% OFF and 60% OFF as well but with lower overall numbers.
Men received least amount of offers for 70%OFF, Hurry and 60%OFF, Hurry whereas women received least amount of offers for 15%OFF,64%OFF.

Size Options by Gender:

This chart compares the availability of different clothing sizes for men and women.

Men have a significantly higher count of “S,M,L,XL, XXL” size category compared to women.
Options like Onesize are more available for women than men. On the other hand, size categories like “28,30,32,34,36,38” and “30,32,34,36,38” are more available to men.
Both genders have very few options in the extreme sizes (such as XS,S,M,L,XL,XXL 3XL).





OBSERVATIONS:

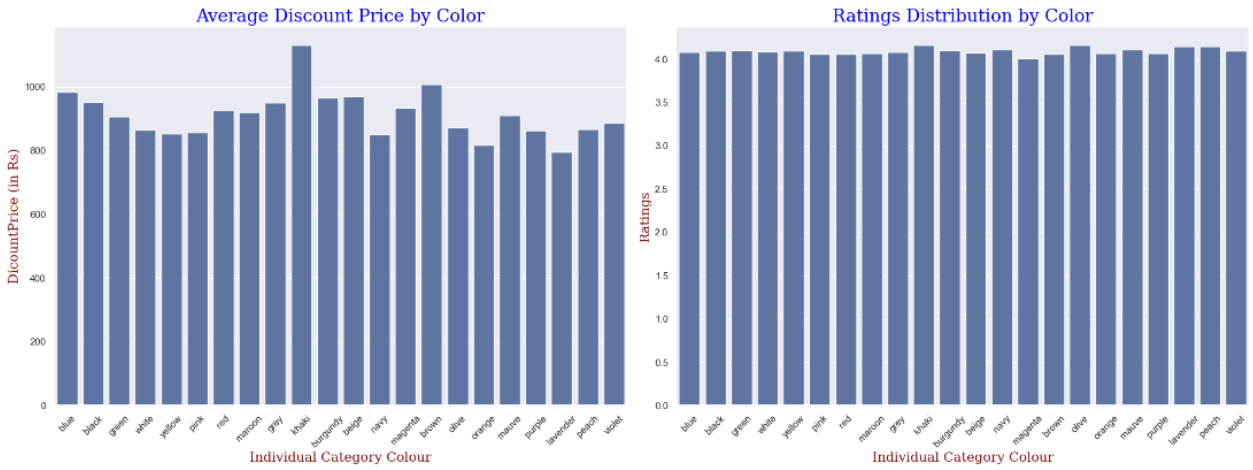
Average Original Price by Category:

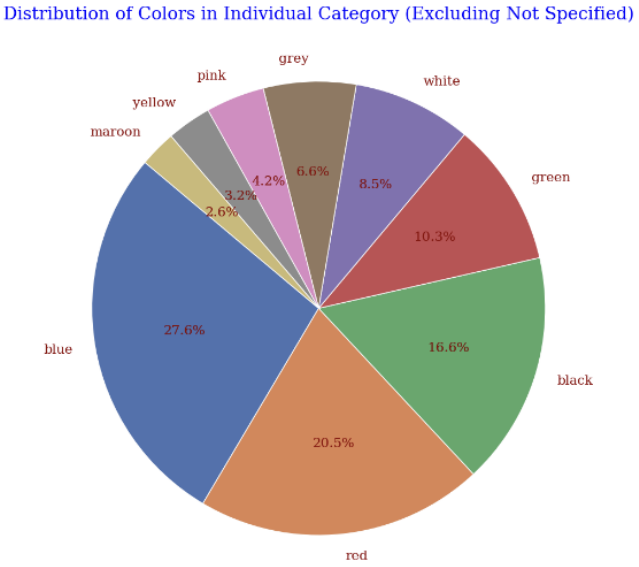
- The chart shows various product categories along the x-axis and their corresponding average original prices along the y-axis.
- 'Indian Wear' has the highest average original price among the categories displayed.
- The categories follow with descending average prices: 'Bottom Wear', 'Plus Size', 'Sports Wear', 'Topwear', 'Lingerie & Sleep Wear', 'Inner Wear & Sleep Wear', and 'Western'.
- 'Western' wear has the lowest average original price among the listed categories

Average Original Price by Gender Category:

- This chart compares the average original price of products for men and women.
- The average original price for men's products is slightly higher than for women's products.

Overall, from these charts, it can be inferred that 'Indian Wear' tends to be the most expensive category on average, while 'Western' wear is the least expensive. Additionally, men's products are, on average, priced higher than women's products.





OBSERVATIONS:

Average Discount Price by Color:

- The chart shows the average discount price of products categorized by color.
- ‘Khaki has the highest average discount price, followed by ‘brown’ and ‘Blue’.
- ‘Orange’ and 'Lavender' have the lowest average discount prices among the colors shown.
- The prices seem to range approximately between Rs. 800 to Rs. 1100.

Ratings Distribution by Color:

- This chart shows the distribution of ratings for products, also categorized by color.
- All colors have very similar ratings, with minimal variation.
- The ratings are consistently high, all appearing to be above 4 out of 5.
- There is no color with a distinctively higher or lower rating than the others.

The pie chart titled "Distribution of Colors in Individual Category (Excluding Not Specified)" shows the percentage distribution of colors for a certain category of items:

- Blue is the most prevalent color, making up 27.6% of the category.
- Red comes next, representing 20.5% of the items.
- Black is also a common color at 16.6%.
- Green and white have a significant share with 10.3% and 8.5%, respectively.
- Grey has a smaller portion at 6.6%.
- Pink, yellow, and maroon have even smaller shares, at 4.2%, 3.2%, and 2.6% respectively.

The pie chart indicates that blue, red, and black are the most dominant colors in this particular category, while pink, yellow, and maroon are the least common. This distribution can provide insights into color preferences or inventory decisions within this category.

Note: This plot above showing sub plot(s) including two bar plots and one pie chart is included in last section of the dashboard (in Figures & Graphs)

TABLES

Category-wise Financial Analysis of Fashion Products			
Category	OriginalPrice	Final Price	Discount Amount
Bottom Wear	2200.64	1138.25	1062.39
Indian Wear	2486.97	1022.04	1464.94
Inner Wear & Sleep Wear	1023.88	618.86	405.02
Lingerie & Sleep Wear	1364.29	703.70	660.59
Plus Size	2079.57	817.62	1261.95
Sports Wear	1898.55	972.99	925.57
Topwear	1777.41	901.97	875.44
Western	1000.00	532.00	468.00
Maximum Value	2486.97	1138.25	1464.94
Minimum Value	1000.00	532.00	405.02
Category Max Value	Indian Wear	Bottom Wear	Indian Wear
Category Min Value	Western	Western	Inner Wear & Sleep Wear

statistics for each numerical feature in the dataset					
Statistic	Final Price (in Rs)	OriginalPrice (in Rs)	Ratings	Reviews	Discount Amount
Mean	935.66	1977.78	4.08	63.06	1042.12
Variance	256022.07	867721.77	0.23	15878.91	408552.29
Standard Deviation	505.99	931.52	0.48	126.01	639.18
Median	799.0	1850.0	4.1	19.0	950.0

DASHBOARD

TABLE IN DASHBOARD:

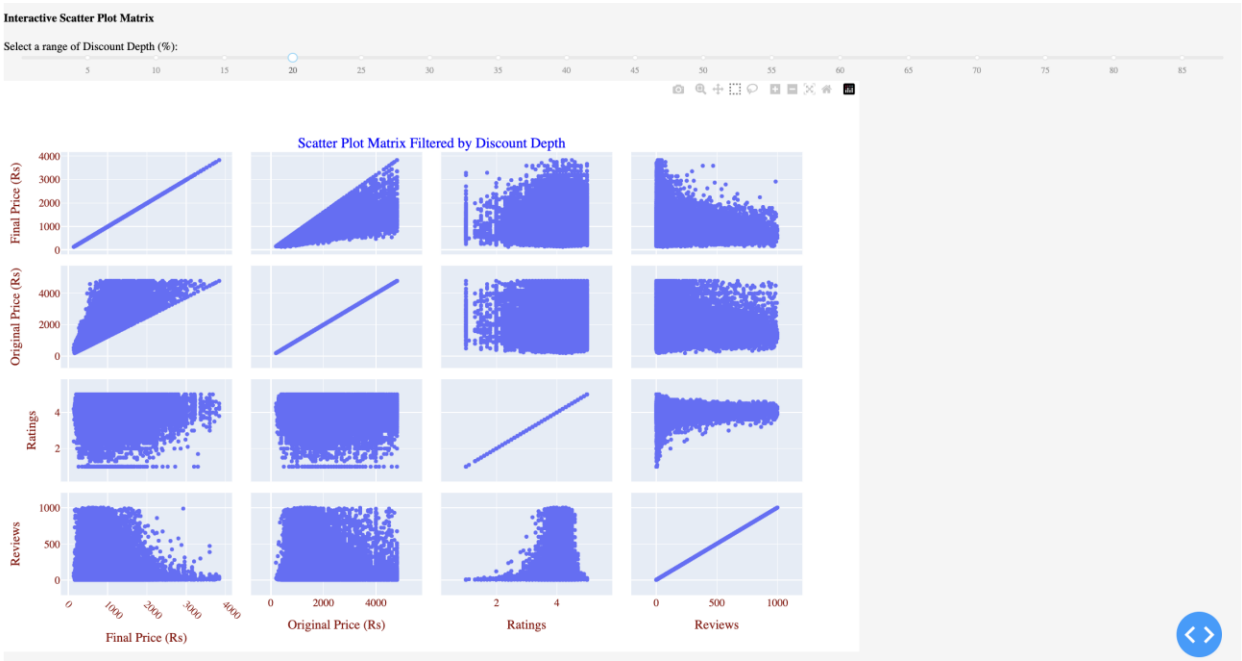
Category	Individual_category	Final Price (in Rs)
Bottom Wear	jeans	1282.35
Bottom Wear	shorts	850.25
Bottom Wear	track-pants	946.69
Bottom Wear	trousers	1218.08
Indian Wear	blazers	2253
Indian Wear	bra	617.8
Indian Wear	burqas	1710.25
Indian Wear	churidar	541.52
Indian Wear	clothing-set	1525.5
Indian Wear	co-ords	1411.56
Indian Wear	dhotis	931.46

« ‹ 1 / 15 › »

Table (from Dashboard – Figures and Graphs Section):

The table depicts the average Final Price (in Indian Rupees) for each Individual Category within Category.

Interactive Scatter Plot Matrix (from Dashboard – "Figures & Graphs")Section:



OBSERVATIONS:

'Discount Depth (%)' is calculated. It represents the percentage reduction from the original price to the final price. A scatter matrix (scatter plot matrix) is created using Plotly Express (px.scatter_matrix). This matrix shows the relationships between 'Final Price (in Rs)',

'OriginalPrice (in Rs)', 'Ratings', and 'Reviews'. The DataFrame is filtered to only include rows where the 'Discount Depth (%)' is greater than or equal to the selected value from the slider.

Final Price vs. Original Price: There is a strong, positive linear correlation between the final price and the original price. This indicates that as the original price increases, the final price tends to also increase in a linear fashion. The cluster of points along the diagonal suggests that for a large number of items, the final price is a proportionate reduction from the original price.

Diagonal Histograms: The diagonal plots, which are histograms, show the distribution of each variable. The histograms for both final and original prices show that a significant number of products are priced at the lower end of the scale, with fewer products as the price increases. The ratings are mostly high, indicating that most products have good ratings. The reviews histogram shows that most products have a lower number of reviews.

Final Price vs. Ratings and Reviews: The scatter plots do not show a clear relationship between the final price and ratings or reviews. The points are spread out, indicating that the final price does not significantly affect the ratings or number of reviews a product receives.

Original Price vs. Ratings and Reviews: Similar to the final price, the original price does not display any apparent correlation with ratings or reviews, with points dispersed across the plot.

Ratings vs. Reviews: There seems to be a slight positive trend between ratings and reviews, suggesting that items with higher ratings may have a marginally higher number of reviews. However, the relationship is not as pronounced, and there's a lot of variability.

Interactive Range Selection: The top of the image has an interactive component that suggests users can select a range of discount depths. This implies that the scatter plots will be able to filter based on the discount depth selected.

Scatter Plot Matrix (from Dashboard – "Correlation Coefficient")Section:

OBSERVATIONS:

Positive Correlation: There seems to be a strong positive correlation between 'FinalPrice (Rs)' and 'OriginalPrice (in Rs)', indicating that as the original price increases, the final price also tends to increase.

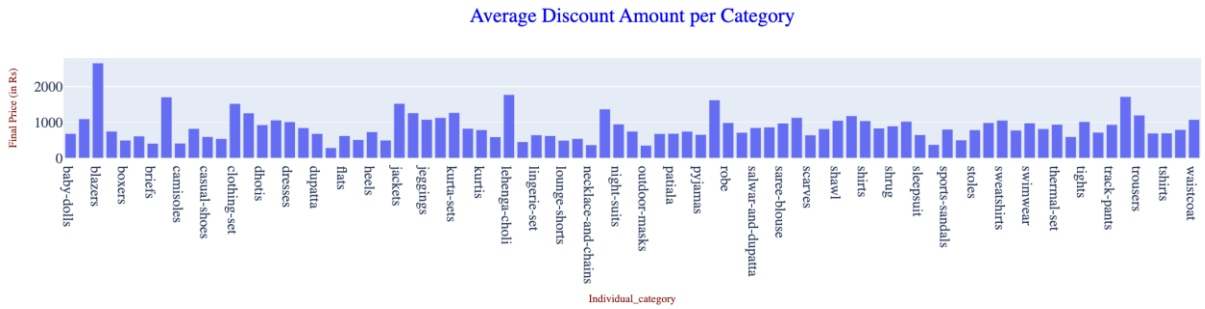
Rating Distributions: As the concentration of data points at the higher end of the scale, around 4 to 5. This could imply that many products have high ratings.

Discount Depth (%): Looking from the plot most products have a lower percentage of discount, and few products have a very high discount percentage.

For the '**Discount Amount**' it seems like most discounts are on the lower side. There are fewer instances of high discount amounts.

Review Distributions: For the 'Reviews' variable most of the data points close to 0, suggesting that many products have few reviews.

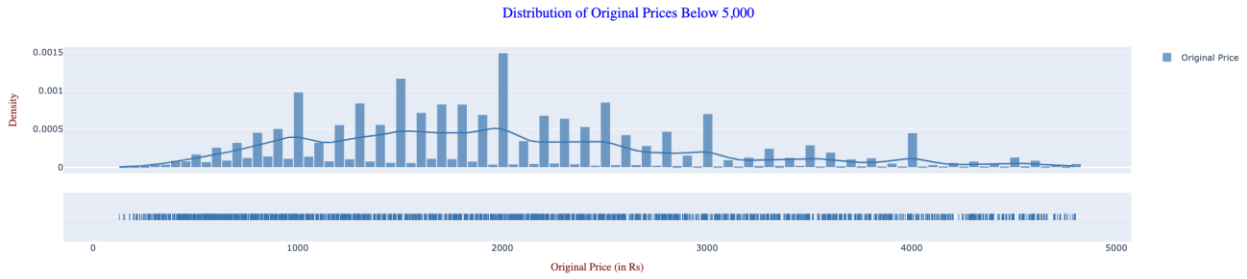
Bar Plot Average Discount Amount Per Category (from Dashboard – "Figures & Graphs")Section:



OBSERVATIONS:

The average final price is highest for blazers which is 2662 Indian Rupees followed by burqas & lehenga-choli for approximately 1775 Indian Rupees. The average final prices for items like earrings and socks, outdoor masks have much lower final prices when compared to clothing available in Individual Category.

Distribution of Original Prices Below 5,000 (from Dashboard – "Figures & Graphs")Section:



OBSERVATIONS:

The histogram aims to show the distribution of original prices for items that cost below 5,000 in currency (likely Rupees, indicated by "Rs").
The horizontal axis (x-axis) represents the original price range, segmented into bins that span from 0 to 5,000 Rs.
The vertical axis (y-axis) measures the probability density of the prices falling within each bin.
There are several peaks in the histogram, indicating that certain price ranges are more common.
The highest peak is around the 1,000Rs and around 2000 Rs mark, suggesting that a significant number of items are priced close to this amount.

OBSERVATIONS

Table 1: Category-wise financial Analysis of Fashion Products

Average Pricing and Discounts by Category:
Indian Wear: Has the highest average original price at ₹2486.97, indicating that it's a premium category. It also has the highest average discount amount of ₹1464.94, suggesting significant price reductions, possibly to attract buyers despite the high initial prices.
Bottom Wear: Shows the highest average final price at ₹1138.25, despite not having the highest original price, which might indicate less discounting compared to other categories.
Plus Size: Exhibits a considerable average original price of ₹2079.57 and a significant average discount of ₹1261.95, reflecting a strong discounting strategy, possibly to encourage sales in this category.
Comparing Maximum and Minimum Values:
Maximum Original Price and Discount: Both are observed in the 'Indian Wear' category, reinforcing its position as a high-value category with substantial discounts.
Maximum Final Price: Found in 'Bottom Wear', indicating its popularity or lower discount rates compared to 'Indian Wear'.
Minimum Values: All the minimum values (original price, final price, and discount amount) are found in different categories, with 'Western' having the lowest original and final prices. This might suggest a more affordable range or smaller inventory in the 'Western' category.
Category-Specific Insights:
Inner Wear & Sleep Wear: While not the lowest in any category, it shows relatively lower average prices and discount amounts, possibly due to the nature of the products being more essential and less fashion-driven.
Sports Wear and Topwear: These categories show a balanced approach with moderate pricing and discounting, indicating steady demand and pricing strategies that don't rely heavily on discounts.
Strategic Implications:
The data suggests a varied pricing and discounting strategy across different categories. Categories like 'Indian Wear' and 'Plus Size' might be using discounts as a key strategy to drive sales, while 'Bottom Wear' and 'Topwear' might rely more on their popularity or less on aggressive discounting.
The lower average prices in 'Western' and 'Inner Wear & Sleep Wear' could be due to these being more regularly purchased items, thus priced more affordably.

Table 2: Statistics for each Numerical Feature in the Dataset

- **Mean (Average) Values**:

Final Price (in Rs): The average final price is ₹935.66, suggesting this is the typical price customers pay after discounts.
Original Price (in Rs): The average original price is significantly higher at ₹1977.78, indicating a substantial markup before discounts are applied.
Ratings: The average rating is 4.08, suggesting generally favorable customer reviews.
Reviews: An average of 63.06 reviews per product, indicating a moderate level of customer feedback engagement.
Discount Amount: The average discount amount is ₹1042.12, showing a significant reduction from the original price.

- **Variance**:

High variance in both the original and final prices (867721.77 and 256022.07, respectively) indicates a wide range of prices across different products. Similarly, the discount amount's high variance (408552.29) suggests significant variability in the discounts offered.
Ratings and reviews have lower variances (0.23 and 15878.91), indicating more consistency in customer ratings and the number of reviews per product.

- **Standard Deviation**:

Standard deviations for final price, original price, and discount amount (505.99, 931.52, and 639.18 respectively) are relatively high, further confirming a wide spread in these values.
A standard deviation of 0.48 in ratings implies that most ratings cluster around the average, with fewer extremes.

- **Median Values**:

The median final price is ₹799.0 and the median original price is ₹1850.0, slightly lower than their respective means, suggesting a slight skew towards cheaper products.
The median rating is 4.1, aligning closely with the mean, reinforcing the overall positive feedback from customers.
The median number of reviews is significantly lower (19.0) than the mean, indicating that a few products might have a very high number of reviews, skewing the average upwards.
The median discount amount of ₹950.0 is also close to the mean, indicating a fairly consistent discounting strategy across products.

Line Plot:

There's a noticeable variation in the average final prices in Indian rupees among the different bottom wear categories. This suggests that certain types of bottom wear are, on average, priced higher than others even after discounts are applied.

Jeans have the highest average price after the discount among the categories displayed, while shorts have the lowest. This could imply that jeans is costliest in the market compared to other categories.

The plot shows a downward trend moving from jeans to shorts, indicating a decrease in the average price after discount. However, there's an upward trend from shorts to track-pants and then to trousers, suggesting that the average price after discount for trousers is nearly as high as for jeans.

Note: This same plot used as Line Plot in dashboard for various Individual Categories vs Avg Final Price.

Bar plot:

The bar chart visualizes the average discount amount across different product categories from the Myntra dataset. This suggests that some categories tend to have higher average discounted prices than others.

Indian Wear and **Plus Size** categories have the highest average discount amounts, indicating that these categories might be more heavily discounted than others.

The **Inner Wear & Sleep Wear**, Western Wear and Lingerie & Sleep Wear category shows a significantly lower average discount amount compared to Indian Wear and Plus Size, which might suggest less aggressive discounting strategies for these products or possibly higher base prices.

Bottom Wear, **Topwear** and Sports Wear categories have moderate average discount amounts, suggesting a balanced discount strategy across these commonly purchased items.

Bar plot(Stacked):

The stacked bar chart presents brand preference by gender across selected Indian brands from the Myntra dataset. The chart effectively uses blue and pink to traditionally represent men and women, making it easy to distinguish between the two categories at a glance.

Indian Terrain and **URBANIC** are the most preferred brands among men and women, respectively, with Being Human having a significantly second higher preference among men compared to all other brands in the dataset.

In **FABINDIA** & **MAX**, the number of products preferred by men and women is relatively similar, suggesting a more gender-neutral positioning or product offering.

Armaan Ethnic, Lenin Club & Oxemberg show a low level of preference for both genders, with a slight inclination towards men's preference for all three.

Nanda Silk Mills shows a low level of preference for women, among other women centric brands.

Note: This same plot used as Bar Plot(Stacked) in dashboard for brand preference by gender for selected brands

Bar plot(Grouped):

The graph titled "Average Ratings by Brand and Gender for Selected Brands" shows the mean ratings for different clothing brands, split by gender categories (Men and Women). Here are some observations:

Ratings range from around 3.5 to nearly 4.5, indicating generally favorable reviews for all brands.
For the brand Fabindia, men have rated it slightly higher than women whereas for max, women's ratings are slightly higher than men's.

BEING HUMAN has the highest average ratings among all the selected brands for men whereas max, Nanda silk mills and Urbanic received equally similar highest ratings among all the selected brands for women.

Linen club, Qurvii received low average rating of around 3.7 among all the selected brands for men and women respectively.

Note: This same plot used as Bar Plot(Grouped) in dashboard for Average Ratings by brand and Gender for selected Brands.

Count Plot :

The graph titled "Count of Products by Review Bins" shows the distribution of product counts across various review count ranges. Here are some observations:

The x-axis represents different bins of review counts, which are categorically organized as 0-10, 11-20, 21-50, 51-100, 101-500, 501-1000, and 5001-10000 reviews.
The y-axis indicates the number of products falling into each review bin, with counts ranging from 0 up to approximately 30,000 products.
The majority of products have a low number of reviews, with the 0-10 review bin containing the highest number of products, suggesting that many products receive only a few reviews.

As the number of reviews per product increases, the number of products in those bins generally decreases. This indicates that fewer products receive a large number of reviews.

The 11-20 and 21-50 review bins contain a substantial number of products, although these counts are notably lower than the 0-10 review bin.

There is a significant drop in product counts as the review bins increase to 51-100 and further to 101-500 reviews, suggesting that it's less common for products to reach such high numbers of reviews.

The bins for 501-1000 reviews have much lower counts, indicating a very small proportion of products accumulate this many reviews.

Note: This same plot used as Count Plot in dashboard for Count of Product by Review Bins.

Pie Chart:

The pie chart titled "Size Options for Bottom Wear (Men)" illustrates the distribution of size options for men's bottom wear. Here are some observations:

The chart is dominated by a single category labeled "Onesize," which occupies the vast majority of the chart at 97.6%. This suggests that a significant portion of the bottom wear comes in a one-size-fits-all option.

The remaining 2.4% is represented by individual sizes: 28, 30, 32, 34, 36. These sizes combined make up a very small portion of the overall size options available.

The specific focus on "Onesize" indicate a trend or a specific market segment where one-size-fits-all garments are predominant.

Note: The same plot is used in Dashboard as Size Options for Categories and Gender

HeatMap with cbar :

The heat map titled "Heat Map of Average Ratings by Top 5 Brands and Category" shows the average ratings for various clothing categories across five brands. Here are some observations:

The brands included are Sangria, Roadster, Puma, HRX by Hrithik Roshan and HERE&NOW.

The categories represented are Bottom Wear, Indian Wear, Inner & Sleep Wear, Lingerie & Sleep Wear, Plus Size, Sports Wear and Topwear.

The color gradient represents average ratings, with darker red indicating higher ratings and lighter shades indicating lower ratings. The scale ranges from 0 to 4.

Several categories for certain brands have a rating of 0.0, indicates that there are no products from that brand in that category.

Puma and HRX by Hrithik Roshan have high ratings across most categories they are present in, suggesting customer satisfaction in these categories.

The highest average rating on the map is 4.33 for Puma in the Plus Size category, indicating a very positive reception.

Sangria has the broadest variation in ratings, with a high of 4.09 in Topwear and a low of 0.0 in several other categories.

Roadster has consistent ratings across Indian Wear and Topwear (4.18) but has a lower rating in the Bottom Wear, Inner wear & sleep wear, Sports wear category (3.9).

HERE&NOW has ratings in all categories except Inner wear & Sleep wear, with the highest being 4.16 in Sports Wear.

Pair Plot:

The diagonal plots show the distribution of each variable for different categories. The distributions for final price and original price are right-skewed, indicating that most of the data points are at the lower end of the price scale with fewer items at the higher price end. Ratings have a peak around the higher values, suggesting that higher ratings are more common. Reviews and discount amount distributions are highly right-skewed, indicating that fewer products have a high number of reviews or a high discount amount.

Final Price vs. Original Price: There is a strong positive correlation between final price and original price. Products with higher original prices tend to have higher final prices after discount. This relationship is linear and dense, with less variation as the price increases.

Ratings: The ratings appear to be less variable and mostly concentrated in the higher score range across all categories. There is no clear trend between ratings and prices, suggesting that the price does not significantly affect the ratings given by customers.

Reviews : Most products have a lower number of reviews, with only a few products having a very high number of reviews, regardless of the category. There doesn't appear to be a strong relationship between the number of reviews and the prices or discount amounts.

Discount Amount: There seems to be a trend where higher original prices have higher absolute discount amounts, which is expected since a fixed percentage discount on a higher original price would result in a higher discount amount in absolute terms.

Multivariate box plot:

The plot compares the distribution of ratings for two gender categories: Men and Women. These categories are further divided by two discount offers: 10% OFF and 1% OFF.

For both Men and Women, the 10% OFF discount appears to have a higher median rating than the 1% OFF discount, suggesting that customers might be more satisfied with their purchases or perceive greater value when the discount is higher.

The ratings for Men have a broader interquartile range (the box portion of the plot) for the 10% OFF discount compared to the 1% OFF discount, indicating more variability in how men rated their purchases at the higher discount level.

Women's ratings have a slightly broader range at the 10% OFF discount than at the 1% OFF discount, but the difference is not as pronounced as with the Men's ratings.

Both Men and Women have outliers in their ratings, as indicated by the individual points beyond the whiskers of the boxes. This suggests there are some exceptionally high or low ratings that fall outside the typical distribution range.

The median ratings for both Men and Women are above 4.0 for the 10% OFF discount, which is within the upper half of the possible rating scale, implying general satisfaction with the products purchased with this discount.

The 1% OFF discount has a narrower interquartile range and fewer outliers for both genders, indicating more consistency in the ratings, albeit slightly lower than the 10% OFF discount.

Distplot:

The plot shows two histograms with a kernel density estimate (KDE) overlay, showing the distribution of original prices below 5,000 Rupees for products categorized by gender into Men and Women.

Both distributions appear to be right-skewed, meaning there are a greater number of products at the lower price range and fewer as the price increases. For both Men and Women, the majority of products are concentrated in the lower price range, with the highest density observed around 2,000 Rupees.

The distribution for Men has a pronounced peak around 2,000 Rupees, indicating a significant concentration of products at this price point. After this peak, the frequency gradually decreases but shows smaller spikes, suggesting that there are certain price points at which products are commonly priced.

The distribution for Women also shows a peak around 2,000 Rupees, which is less pronounced than for Men. Similar to the Men's category, the frequency decreases as the price increases, with less pronounced spikes at higher price points.

QQ - Plot:

The Q-Q (quantile-quantile) plot compares the quantiles of Myntra ratings data to the quantiles of a theoretical normal distribution.

There is a noticeable deviation from the line in the lower tail (bottom left of the plot), where the sample quantiles are higher than the theoretical quantiles. This indicates that there are more low ratings than what would be expected in a normal distribution.

There is a slight deviation in the upper tail (top right of the plot) where the sample quantiles slightly exceed the theoretical quantiles. This suggests that there are more high ratings than would be expected if the data were perfectly normally distributed.

The middle range of the data, which corresponds to the bulk of the ratings, adheres quite closely to the theoretical line, indicating that the ratings in this range are well-modeled by a normal distribution.

Histogram plot with kde:

The most common ratings are clustered around the 4.0 to 4.5 range, indicating that the majority of the ratings are high.
The frequency of ratings peaks at just above 4.0. There's a notable decrease in frequency as the ratings approach 5.0, suggesting that perfect scores are less common.
There are significantly fewer ratings in the 1.0 to 2.0 range, indicating that customers are less likely to give very low ratings.
The distribution is left-skewed, with the tail extending more towards the lower ratings, indicating that lower ratings are less frequent but still present.
The title indicates that the histogram excludes products that have not been rated. This suggests that the dataset only includes items that have received an active rating from customers.

Lm plot or reg plot:

The plot depicts a lm plot with a regression line, showing the relationship between the original price and the final price for women's Indian wear. Here are the observations from the plot :

There is a positive correlation between the original price and the final price, as indicated by the upward slope of the regression line. This means that as the original price increases, the final price tends to increase as well.

There is considerable variability in the final prices at almost all levels of original prices. The scatter of points around the regression line suggests that the final price is influenced by factors other than just the original price.

Both original and final prices range from the lower end near 0 to about 5000 Rupees. The majority of data points are concentrated below 3000 Rupees for the original price.

The regression line seems to fit the central trend of the data well, especially in the middle range of original prices. However, the fit is less accurate at the extremes, particularly at the high original price end, where the data points are more spread out.

Strip plot:

The strip plot depicts the distribution of original prices for items categorized by color and gender. Here are some observations:

There is a wide range of colors represented on the x-axis, showing that the data includes a variety of color options, from blue to violet.

The data points are color-coded by gender, with one color for Men and another for Women, allowing for a comparison between the two within each color category.

For most color categories, there are items available across the entire price range for both genders. However, some colors, such as peach ,mauve and violet show a more limited range of prices. I also observed that some rare colors are available at higher prices than easily available colors.

The concentration of data points varies across colors. Some, like blue and black, have a dense collection of data points across the price range, while others, like olive and mauve, have fewer data points, indicating fewer items or less price variation.

There is a noticeable overlap in the price ranges for Men's and Women's items within each color category, suggesting that items for both genders are offered at similar price points.

Hexbin plot:

The image shows a hexbin plot, which is a two-dimensional histogram used to represent the density of data points. Here are some observations:

The ratings on the y-axis range from 1 to 5, and the highest density of data points appears to be around the 4.0 rating mark and The final prices on the x-axis range from 0 to around 4000 Rs. The highest density of data points occurs at lower price points, indicating that most products are priced lower rather than higher.

The hexagons with the lightest color (indicating the highest density) appear to be clustered around certain final price points, which might suggest common price points at which a large number of products are rated.

The distribution of the densest areas doesn't show a clear directional trend from the bottom-left to the top-right or vice versa. This suggests that there is not a strong linear correlation between final price and ratings.

The most common price and rating combinations are where the hexbins are brightest. This seems to be for products with a final price of around 1000 Rs and ratings close to 4.0.

The plot suggests that the majority of products fall within a moderate price range and receive fairly high ratings. There are fewer products with very low ratings and also fewer products with very high ratings.

Rug plot:

There is a positive correlation between the original price and the final price. Items with higher original prices tend to also have higher final prices.
The density of points is greater at the lower end of the original price scale, suggesting that there are more Puma products for men in the lower price range.
The rug plot (the small lines along the x and y axes) provides a one-dimensional representation of the distribution of the original and final prices. The concentration of lines along the bottom x-axis indicates that a significant number of products have lower original prices.
The range of final prices is broad, but most items seem to fall below the 2000 Rs mark.
This plot shows that while there are products across a wide range of prices, most purchases occur at the lower end of the price spectrum.

Kde plot:

KDE of Reviews:

The density of reviews is skewed toward lower numbers for both genders, suggesting that most products have fewer reviews.
There's a sharp peak near the beginning of the Reviews axis, indicating a large number of products with very few reviews.
The tail of the distribution extends towards higher review counts, but with significantly lower density.

KDE of Ratings:

The ratings density for both genders shows a pronounced peak at the higher end, likely around 4 or 5, which suggests that most products receive high ratings.
The distribution is skewed left, indicating fewer products with lower ratings.

KDE of Final Price in Rs:

There's a peak in the lower price range for both genders, showing that a large number of products have lower final prices.
The distribution tails off as the price increases, with very few products in the highest price range.
The peak for Women's products appears to be slightly higher in price than for Men's.

KDE of Original Prices:

Similar to the final price, the original price distribution has a peak at the lower end, indicating that most products are originally priced in the lower range.
The distribution for Women's products at the original price has a broader base than Men's, suggesting a wider range of original prices for Women's products.

Joint Plot:

- There's a high concentration of data points at higher ratings, suggesting that products tend to receive favorable ratings.
- The final price has a broad range but tends to cluster at lower price points.
- There isn't a clear linear relationship between ratings and final price, as high densities of points are spread across various price levels for highly-rated products.
- The marginal histograms or ode provide additional insight, confirming that high ratings are common and that there's a wide distribution of final prices.

Contour Plot:

The brightest area (likely yellow or light green) within the contours indicates the highest concentration of data points. This suggests that most products have a discount amount and final price within this region.
The data points spread outwards from the high-density region to lower-density areas, creating a fan-like shape. This indicates a wide range of both discount amounts and final prices.
The horizontal spread of the data shows that discount amounts vary widely, with a concentration of discounts in the lower to mid-range as indicated by the center of the high-density area.
The vertical spread of the data shows that final prices also vary, with a tendency to cluster in the lower to mid-range of prices.
The shape of the contour lines suggests that there may be a non-linear relationship between discount amount and final price. Higher discounts don't necessarily correspond to lower final prices in a linear fashion.

3D Plot:

The plot extends along the ratings axis from 1 to 5, with most data points clustering towards the higher end of the scale, which suggests that the products generally receive good ratings.

The reviews are concentrated at the lower end of their axis, indicating that most products have fewer reviews. There is a long tail stretching towards higher review counts, but these are less frequent.

The original price is spread along its axis, showing a range of prices from low to high. There seems to be a concentration of products at lower original prices, with fewer products at the higher price end.

There appears to be a dense concentration of points where the ratings are high, reviews are low, and the original price is moderate. This could indicate that most products within the dataset are priced moderately, have high ratings, and fewer reviews.

The 3D aspect of the plot allows for observation of the interaction between all three variables at once. It seems there is no clear pattern indicating that higher prices lead to better ratings or more reviews, or vice versa.

Cluster Map:

The cluster map, combines a heatmap with hierarchical clustering to group similar rows together based on the similarity of their data values.

The dendrogram on the left side of the heatmap shows how the rows are clustered together. Rows that are more similar in terms of their data values are grouped closer together in the dendrogram.

The dendrogram along the top, which would show the clustering of columns based on their similarity across all rows.

The colors in the heatmap represent the magnitude of the variables, with a color gradient typically going from red (higher values) to blue (lower values).

The pattern of colors across the rows can indicate relationships between the variables. For instance, a row with a red color in the 'Discount%' column and blue in the 'Final Price (in Rs)' column might suggest that higher discounts correspond to lower final prices for those particular data points.

You can also look for clusters that show distinct patterns. For example, a cluster with uniformly high values across all variables represents a group of products that are rated highly, have high discounts, many reviews and high original and final prices.

Area Plot:

The plot titled "Average Price Before and After Discount by Category" compares the original and final prices across various clothing categories. Here are some observations:

The categories shown on the x-axis are Western, Inner Wear & Sleep Wear, Lingerie & Sleep Wear, Topwear, Sports Wear, Plus Size, Bottom Wear, and Indian Wear.The y-axis represents the price in Indian Rupees (Rs), ranging from Rs 0 to Rs 2500.

For each category, there are two areas represented – the original price in a lighter blue shade and the final price after discount in a lighter orange shade.

The difference between the original and final prices appears significant across all categories, indicating that discounts have a substantial impact on the final price paid by consumers that is The plot suggests that the discounts are consistent across categories, as the final price lines remain parallel to the original price lines.

The Western category has the lowest original and final prices, while the Indian Wear category has the highest original and final prices.

Inner Wear & Sleep Wear and Lingerie & Sleep Wear have similar pricing structures, with a moderate average price and discount.

The Topwear, Sports Wear, and Plus Size categories show a gradual increase in both original and final prices, with Topwear being the least expensive and Plus Size being more expensive in these three categories.

Bottom Wear has a similar average original price as Plus Size and ends up with a similar final price post-discount.

Violin Plot:

The plot titled "Distribution of Ratings by Gender" is a violin plot that shows the distribution of ratings for products, segmented by gender category: Men and Women. Here are some observations:

The ratings range from 1 to 5 for both Men and Women, which is the typical range for product ratings.
The violin plot for both genders is wide around the rating of 4, indicating that a large number of ratings cluster around this value. This suggests a generally high level of satisfaction with the products.
Both distributions have a median rating of around 4, indicated by the white dot in the center of the violin.
The plot for both genders shows a slightly wider distribution around the higher ratings (4 and above) than the lower ones, implying that fewer products receive lower ratings.
The distributions are quite symmetrical for both Men and Women, suggesting a similar satisfaction level across genders.
There are thin tails near the bottom of the violins for both genders, indicating that very few ratings are at the lower end of the scale (1-2).
No extreme outliers are visible in the plot, which would be represented by individual points outside the main body of the violins.
The consistent width from the top to the bottom of the violins suggests that there is variability in the ratings, but no rating is extremely rare or unusually common.

Swarm Plot:

The plot titled "Swarm Plot of Ratings by Gender for Top Brands" displays the distribution of product ratings for the top brands by gender category.

The x-axis represents the ratings, ranging from 4.4 to 5.0, suggesting that only products with high ratings are being shown.
Each dot on the swarm plot represents a product, colored differently to represent various brands. There are 25 unique brands shown in the legend.

The distribution of ratings among Women is wider than Men, which might suggest a broader range of product satisfaction or a larger number of products rated within the top brands for Women.

Both distributions have ratings that go up to the maximum of 5.0, indicating some products are rated as perfect.

The brands with the highest frequency of high ratings (dots clustered towards the right) could be inferred as having higher customer satisfaction.

No single brand dominates the highest ratings (5.0) for either gender, suggesting that there is a competitive range of products with high satisfaction levels.

Some brands have ratings spread across the range for Women, while others are more clustered, which might indicate consistency in product quality or perception.

CONCLUSION

a. **Insights from Created Graphs:** The graphical analysis of the Myntra dataset provided a multifaceted view of the online fashion retail landscape. The pair plots and 3D scatter plots revealed complex relationships between product features such as ratings, reviews, and prices, highlighting trends and consumer preferences. Histograms and QQ-plots, especially post Box-Cox transformation, offered insights into the distribution characteristics of product prices, aiding in understanding pricing strategies. The PCA graphs, including the scree plot and cumulative explained variance graph, were pivotal in understanding the data's dimensionality, emphasizing the most informative features. The heatmap for the Pearson correlation coefficient matrix illuminated the interdependencies between different product attributes, allowing for a more nuanced understanding of factors influencing consumer decisions.

b. **Python Dashboard Utility:** The Python dashboard created for this project serves as an interactive tool, enabling users to explore the Myntra dataset intuitively. Its design allows for real-time data manipulation and visualization making complex data more accessible and understandable. Users can interact with various elements like sliders and dropdowns to filter and view specific subsets of data, which is particularly useful for identifying trends, making comparisons and conducting a thorough market analysis.

c. **User-Friendliness of the App:** To assess the user-friendliness of the app, feedback was solicited via LinkedIn. Users appreciated the app's intuitive design and the ease with which they could navigate through different features. Many commented on the clarity of visualizations and how they facilitated a deeper understanding of the data.

d. **Functionality of the App:** The app demonstrates a high level of functionality, integrating advanced data analysis techniques into a user-friendly interface. It successfully bridges the gap between complex data analytics and practical business applications. The app’s ability to handle large datasets efficiently, provide real-time data insights and its adaptability to different user requirements, positions it can serve as powerful tool for data-driven decision-

making in the fashion retail industry. The integration of features like outlier detection, PCA and correlation analysis in an interactive format is particularly noteworthy, showcasing the app's capability to perform sophisticated data analyses in an accessible manner.

This project not only enhanced my understanding of the Myntra fashion dataset but also demonstrated the power of data visualization and interactive tools in extracting actionable insights from complex datasets. The successful application of these tools in this project underscores their potential in various industry settings particularly in data-driven sectors like online retail.

REFERENCES:

1. https://www.kaggle.com/datasets/manishmathias/myntra-fashion-dataset
2. https://canvas.vt.edu/courses/176323/files?preview=29332417 (Cover Page Template)
3. https://seaborn.pydata.org/generated/seaborn.kdeplot.html
4. https://matplotlib.org/stable/plot_types/index.html
5. https://canvas.vt.edu/courses/176323/files?preview=30644261
6. https://canvas.vt.edu/courses/176323/files?preview=29733908
7. https://canvas.vt.edu/courses/176323/files?preview=29887156
8. https://canvas.vt.edu/courses/176323/files?preview=29733905
9. https://canvas.vt.edu/courses/176323/files?preview=30505680
10. https://www.kaggle.com/code/babuninayak820/exploring-myntra-fashion-trends-insights