



Analyzing clickstream data for online shopping

Agenda

- ▶ Dataset Introduction
- ▶ Distribution of variables
- ▶ Relationships between the variables
- ▶ Predictive Analysis
- ▶ Support, Confidence, Lift
- ▶ Clustering
- ▶ References
- ▶ Contribution


Dataset

resource

UC Irvine
Machine Learning
Repository

Datasets Contribute Dataset About Us

Search datasets...



Clickstream Data for Online Shopping

Donated on 12/8/2019

The dataset contains information on clickstream from online store offering clothing for pregnant women.

Dataset Characteristics
Multivariate, Sequential

Subject Area
Business

Associated Tasks
Classification, Regression, Clustering

Feature Type
Integer, Real

Instances
165474

Features
14

Dataset Information

Additional Information

The dataset contains information on clickstream from online store offering clothing for pregnant women. Data are from five months of 2008 and include, among others, product category, location of the photo on the page, country of origin of the IP address and product price in US dollars.

Has Missing Values?
No

[Introductory Paper](#)

DOWNLOAD

CITE

1 citations
21887 views

Keywords
retail

DOI
10.24432/C5QK7X

License
This dataset is licensed under Commons Attribution 4.0 International License.
This allows for the sharing of the datasets for any purpose as long as the appropriate credit is given.

LICENSE

GNU GENERAL PUBLIC LICENSE
Version 3, 29 June 2007

Copyright (C) 2007 Free Software Foundation, Inc. <<https://fsf.org/>>
Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

Preamble

The GNU General Public License is a free, copyleft license for software and other kinds of works.

The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program—to make sure it remains free software for all its users. We, the Free Software Foundation, use the GNU General Public License for most of our software; it applies also to

Data description

- Dataset Schema:
- Data is already indexed as integer. The only **string type** is page 2.

```
root
|-- year: integer (nullable = true)
|-- month: integer (nullable = true)
|-- day: integer (nullable = true)
|-- order: integer (nullable = true)
|-- country: integer (nullable = true)
|-- session ID: integer (nullable = true)
|-- page 1 (main category): integer (nullable = true)
|-- page 2 (clothing model): string (nullable = true)
|-- colour: integer (nullable = true)
|-- location: integer (nullable = true)
|-- model photography: integer (nullable = true)
|-- price: integer (nullable = true)
|-- price 2: integer (nullable = true)
|-- page: integer (nullable = true)
```

from April (4) to
August (8)

1 - trousers
2 - skirts
3 - blouses
4 - sale

US \$

1 - Australia
2 - Austria
3 - Belgium
.
.
.
46 - net (*.net)
47 - org (*.org)

1- top left
2- top in the middle
.
.
.
6- bottom right

No missing values!

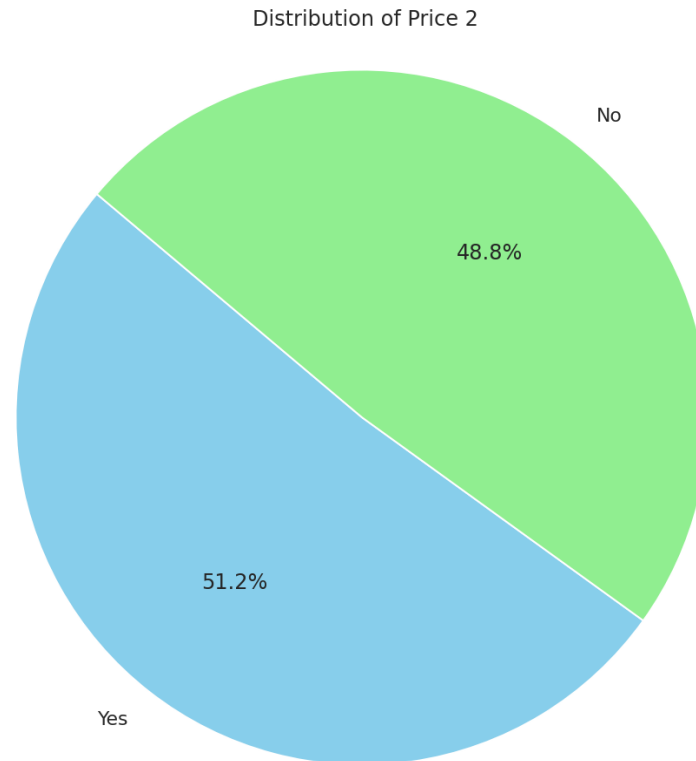
- Good news for us

[illegible]

Distribution of Variables

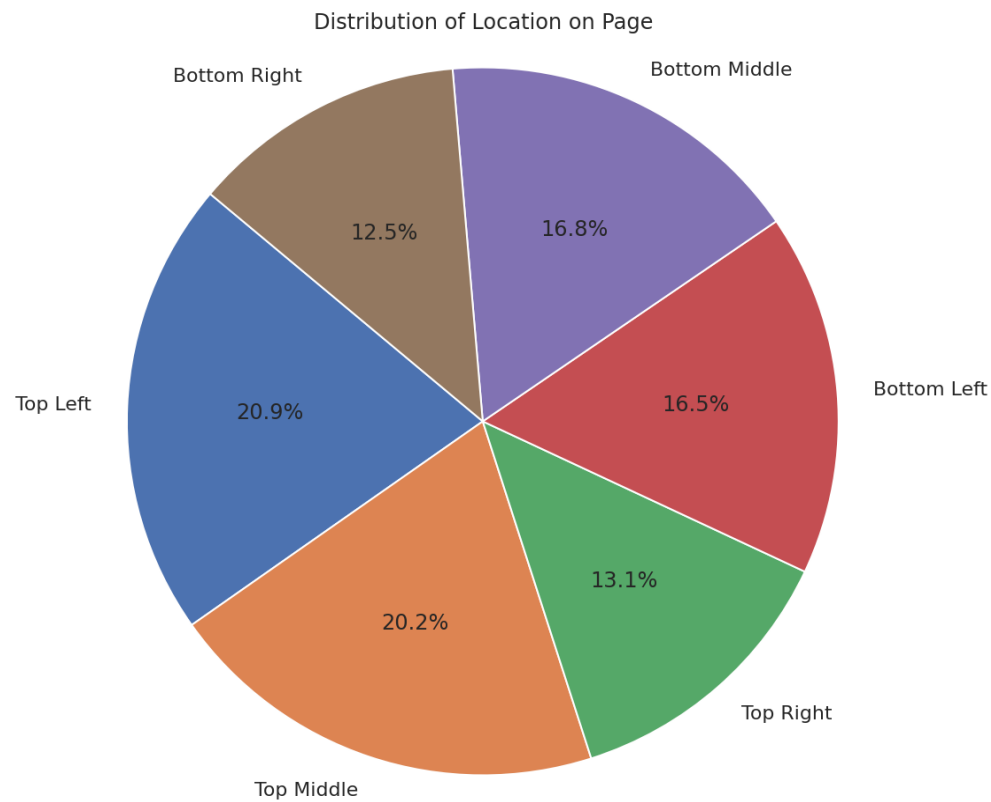
Price 2

- The price of a particular product is higher than the average price for the entire product category.



Location

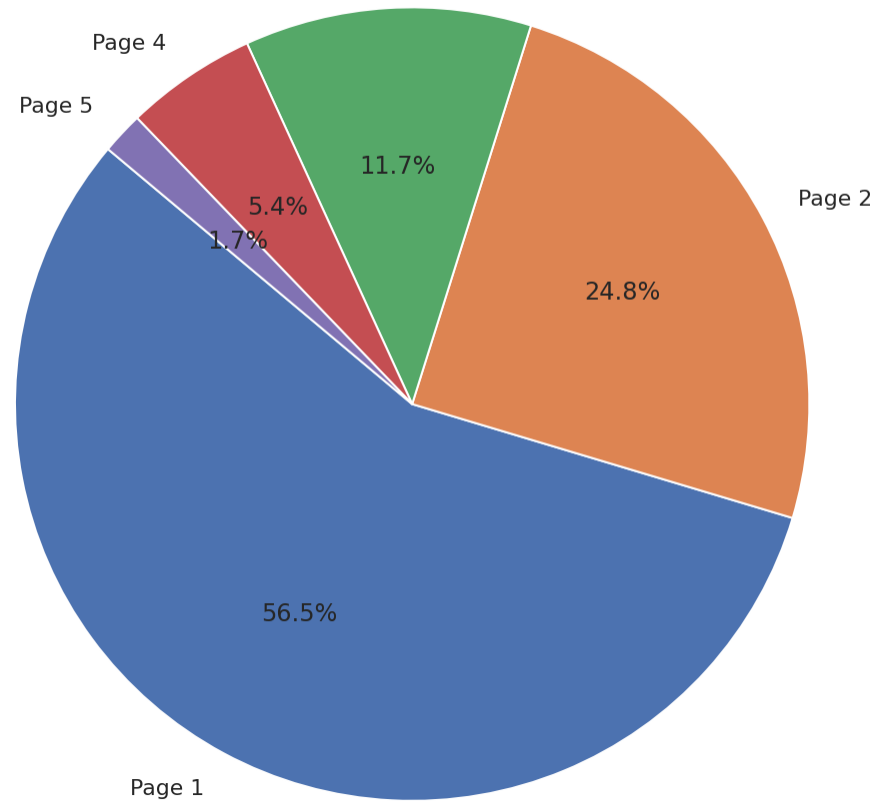
- Photo location on the page, the screen has been divided into six parts



Page

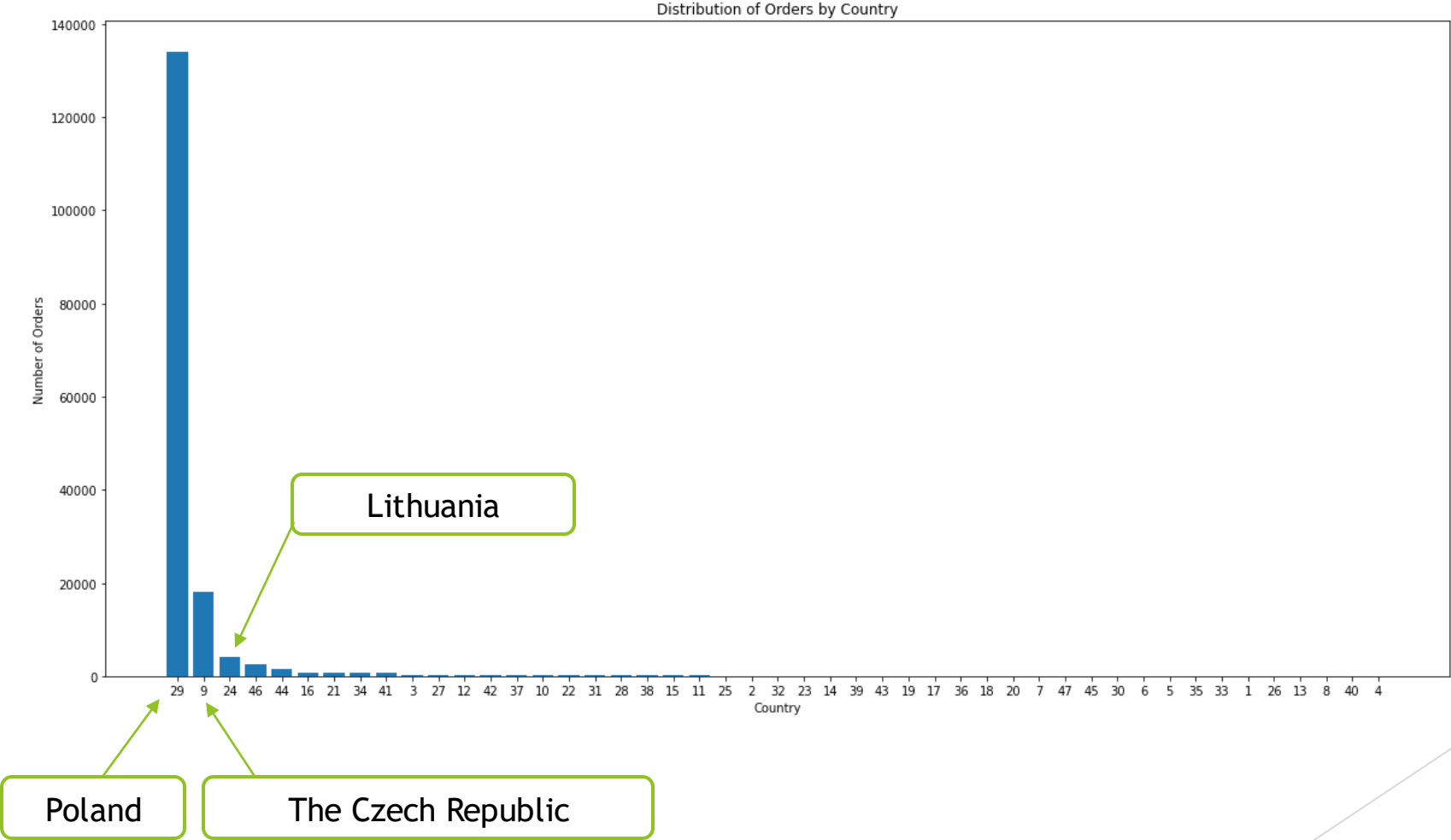
- Page number within the e-store website (from 1 to 5)

Distribution of Page Numbers

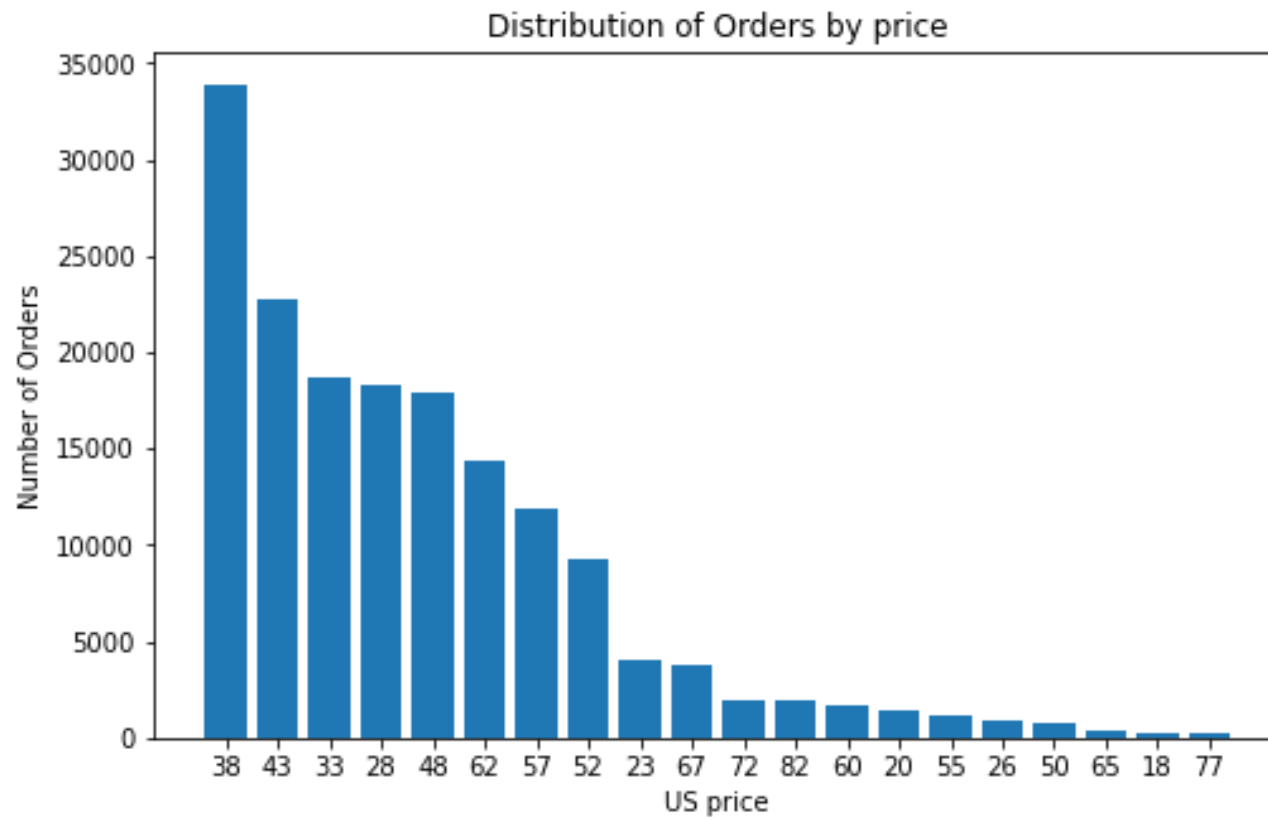


Expected!

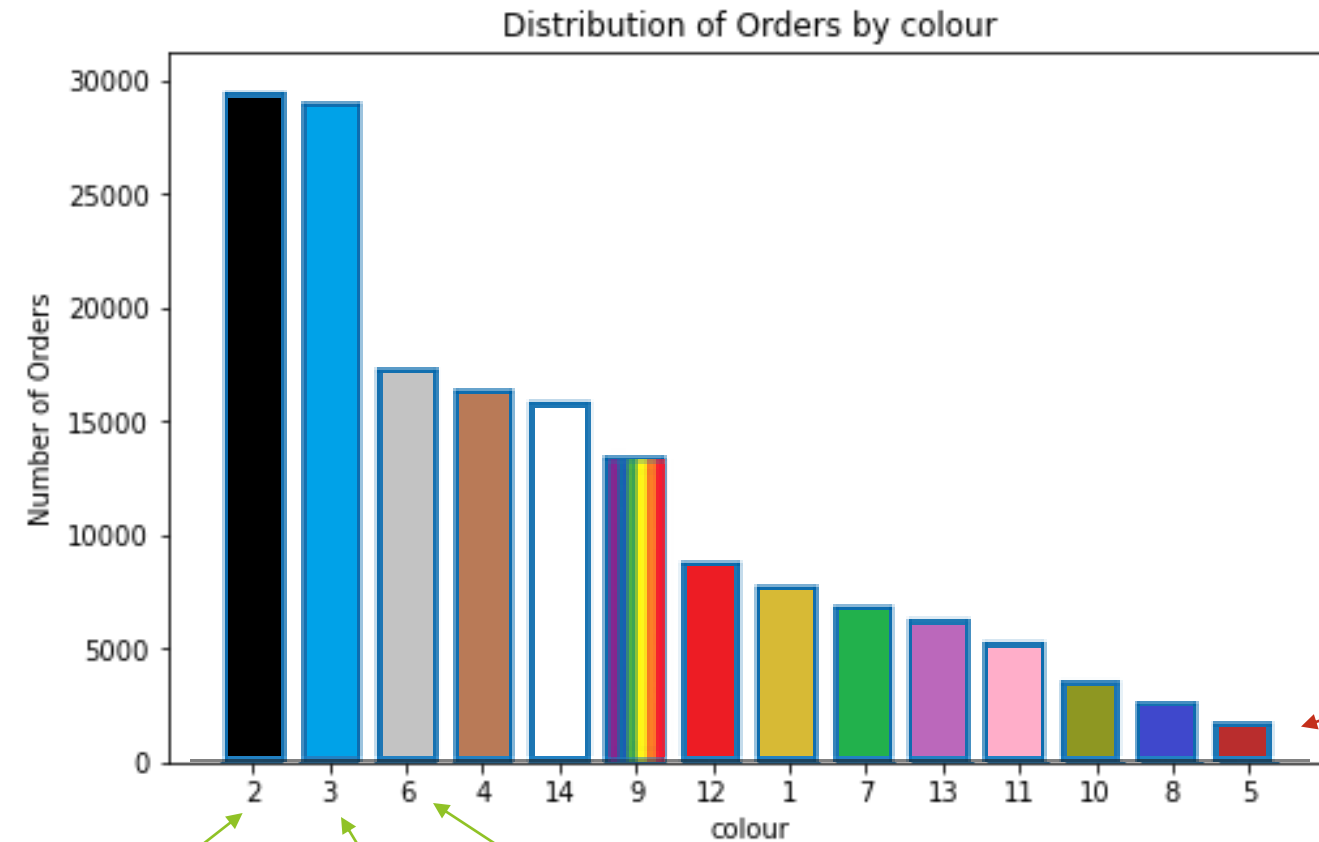
Distribution of orders by country



Distribution of orders by price



Distribution of orders by colour



black

blue

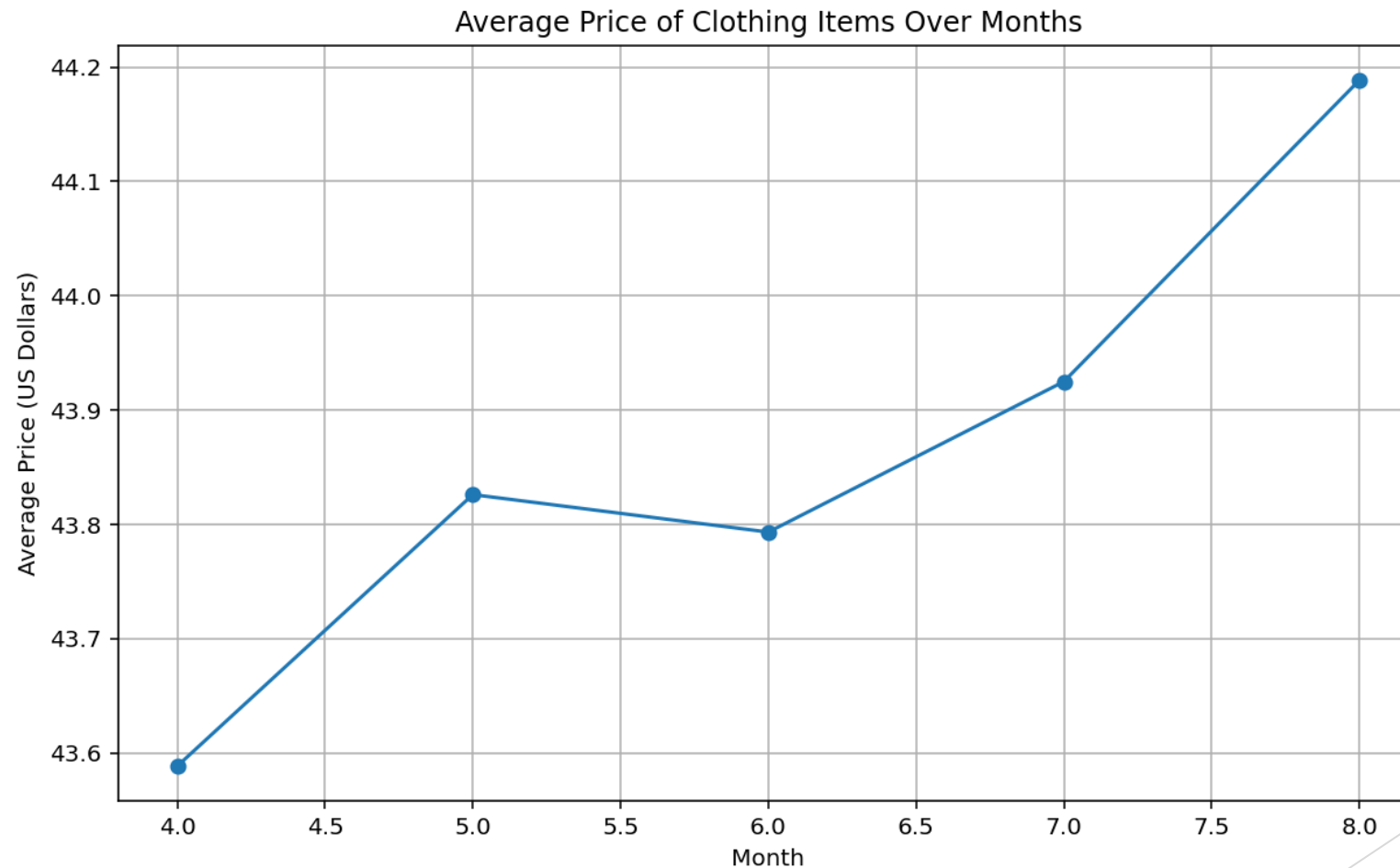
grey

Burgundy

HEX #800020
RGB(128, 0, 32)

Price Over Months

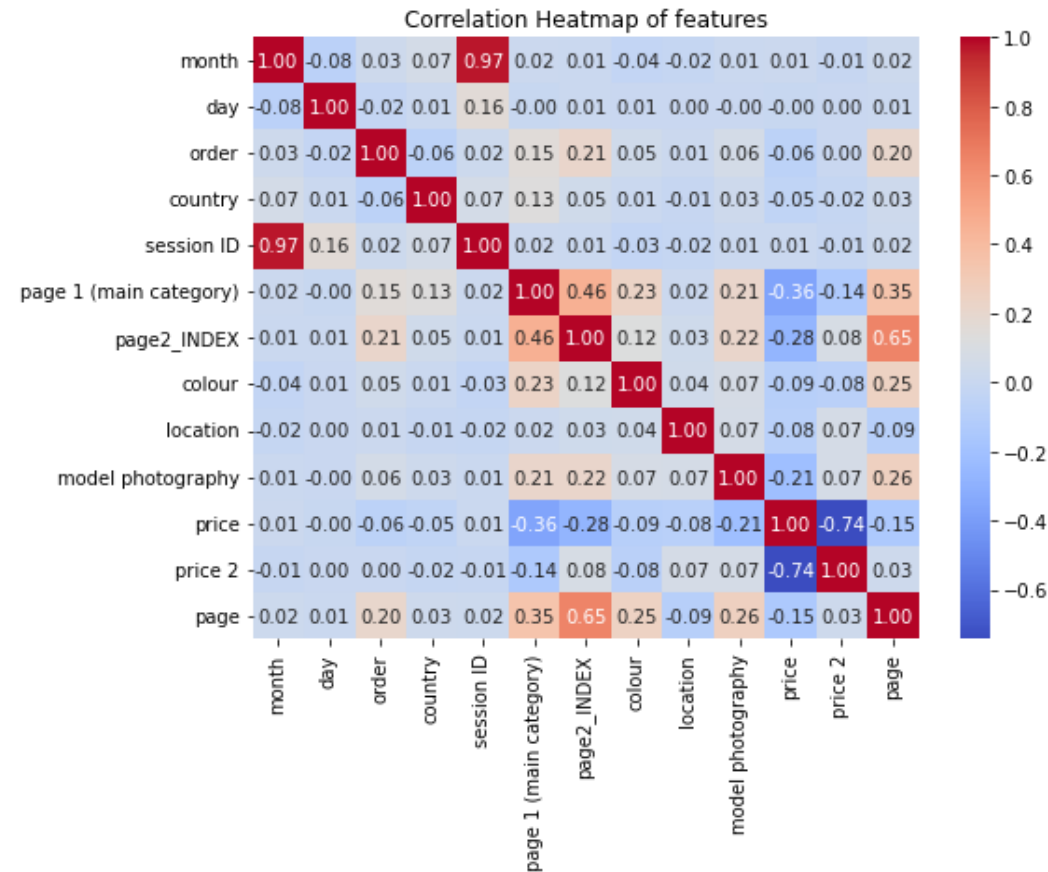
- The average price of clothes is higher in August than in other months.



Relationship Between the Variables

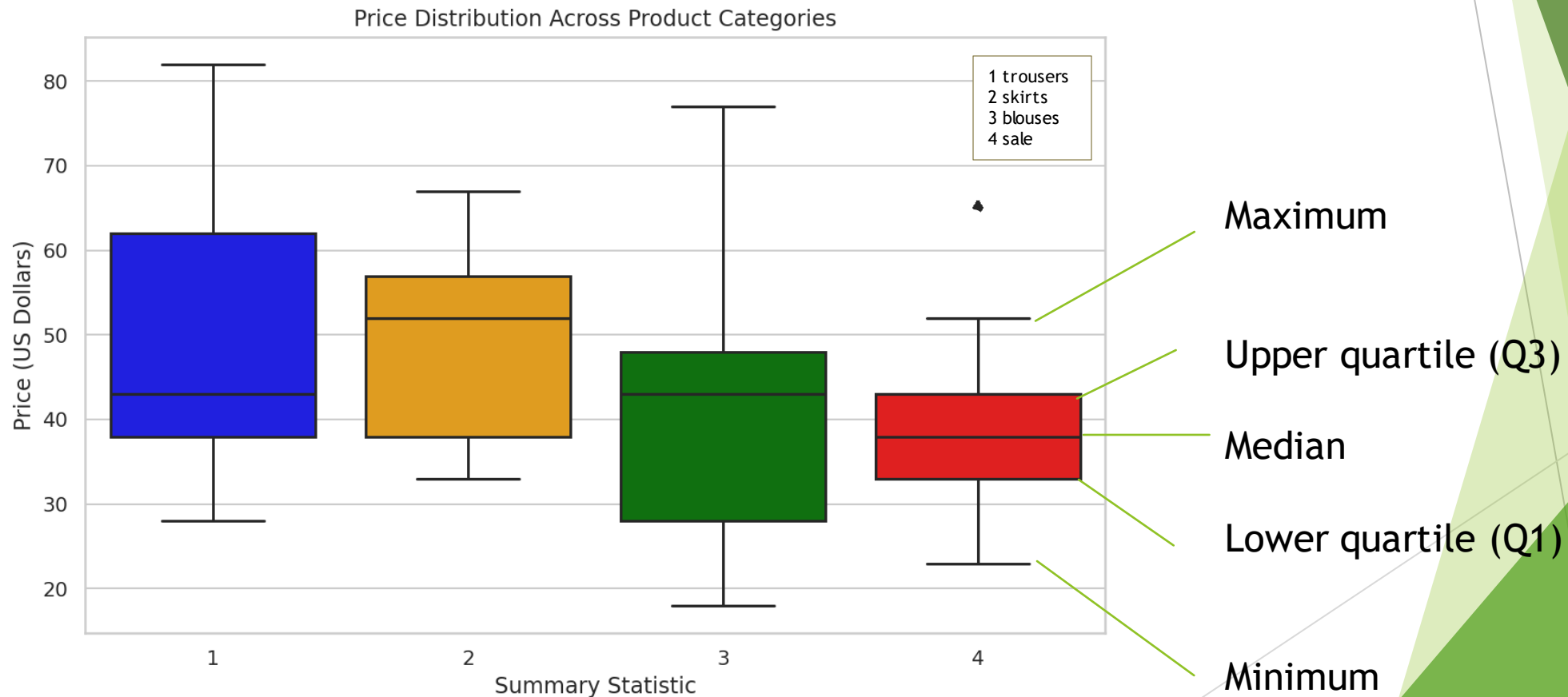
Correlation

- ▶ Session ID and month are strongly correlated
- ▶ Model Photography and price are negatively correlated

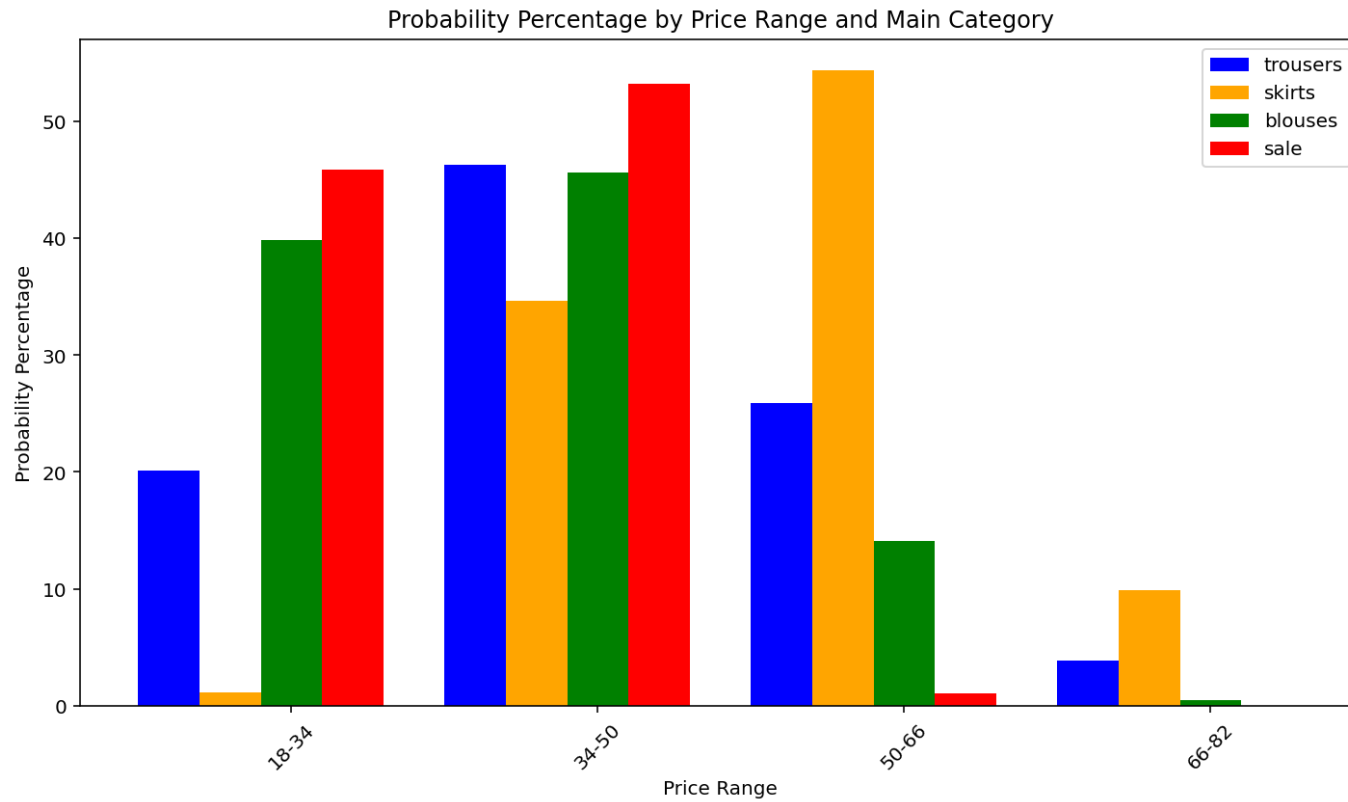


Price distribution in relation to Page 1 (product category)

- ▶ Clothes that are on sale have the lowest median, and skirts the highest
- ▶ The price range for blouses varies the most compared to the other categories



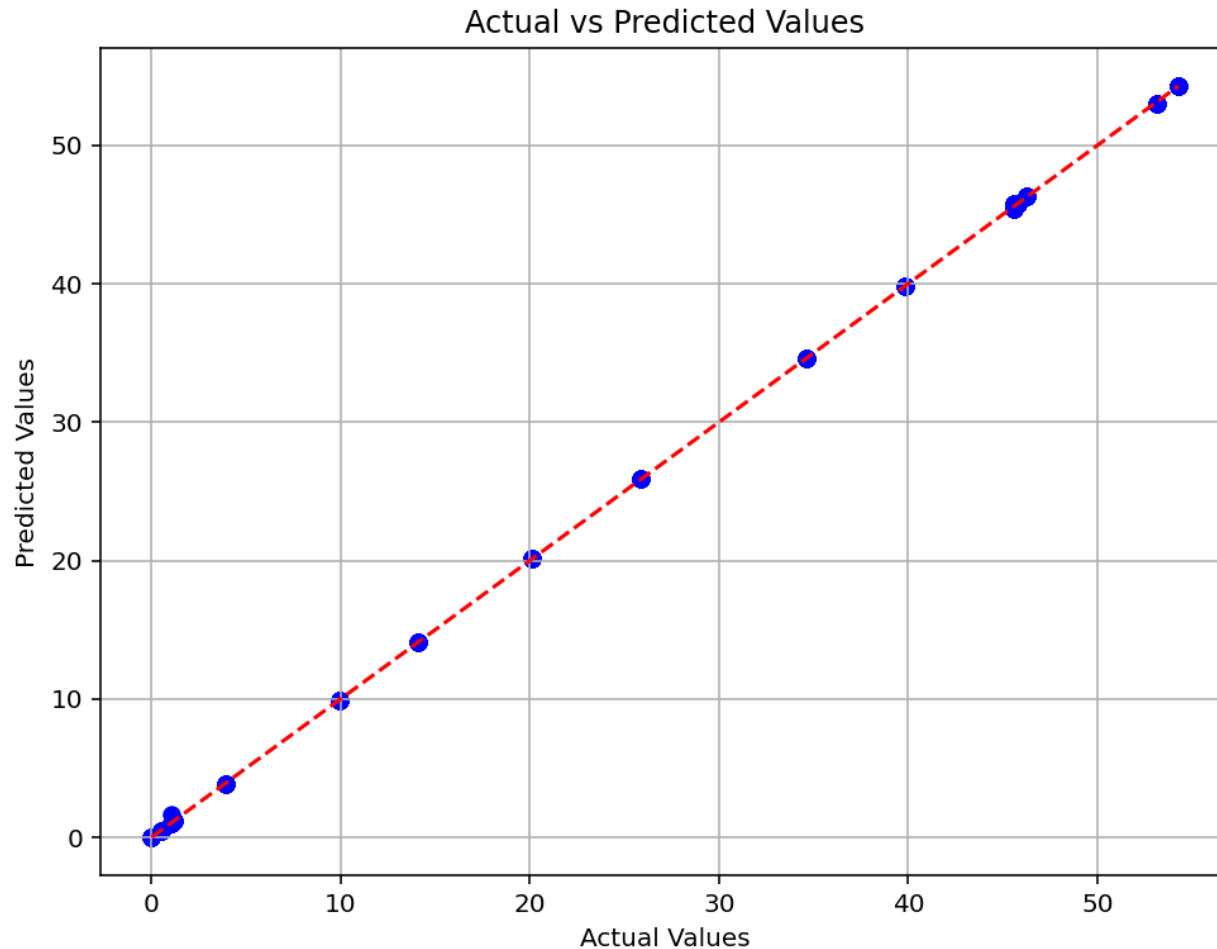
Percentage of Buying From a category in 4 Price Ranges



Predictive Analysis

Model Evaluation on Test Data

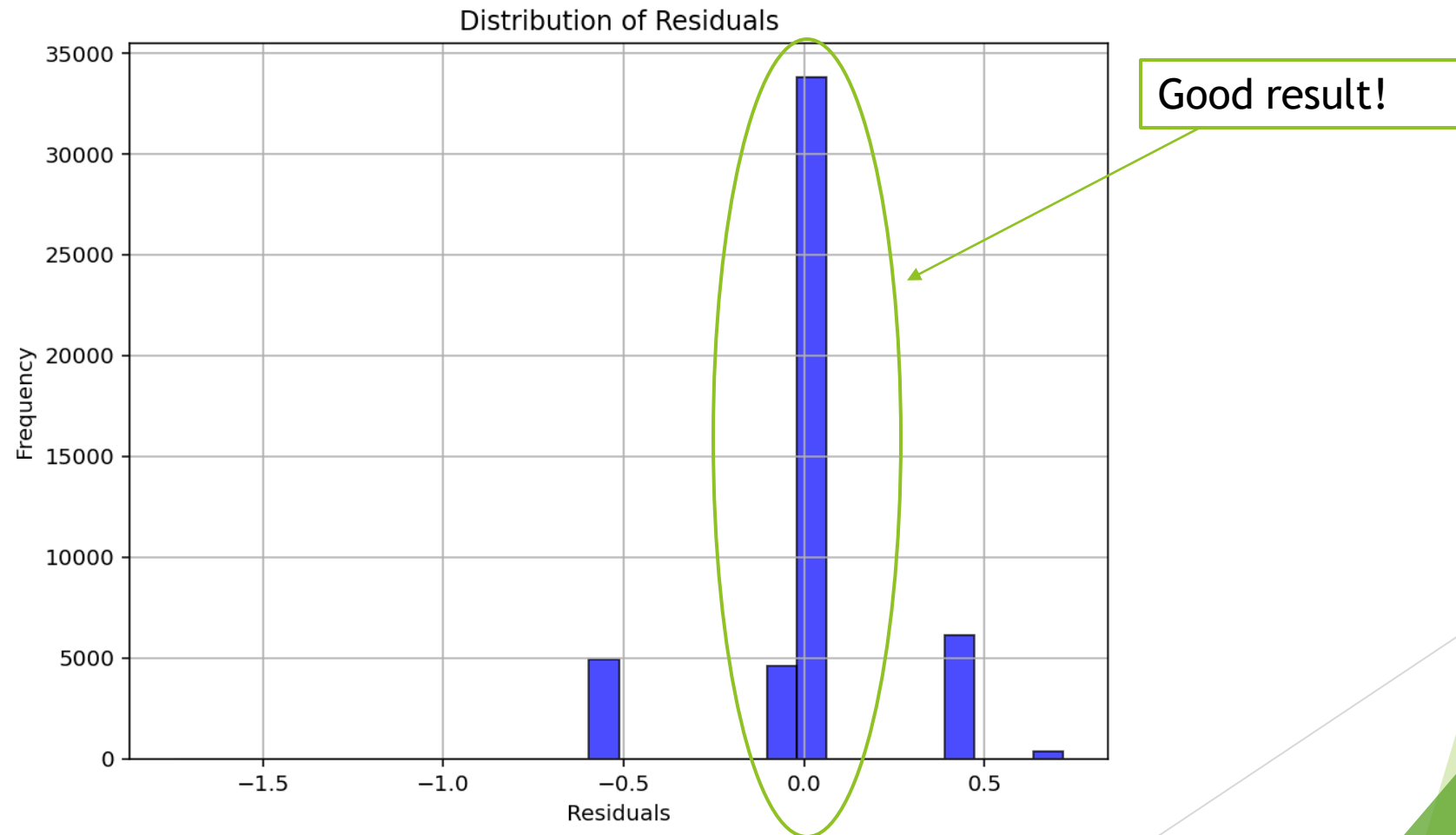
- ▶ Predicting the probability percentage
- ▶ Created a Gradient Boosted Tree (GBT) regression model



- ▶ Mean Absolute Error (MAE):
0.03974245497554473
- ▶ Mean Squared Error (MSE):
0.0064731610206736605
- ▶ R-squared (R2):
0.9999654853058938
- ▶ Accuracy: 0.89

Residual

- The difference between the observed value and the predicted value



Association Rules



Market Basket Analysis Using Association Rules

- ▶ The association rules provide valuable insights into customer purchasing behavior and item relationships.
- ▶ Association rule mining techniques to identify frequently co-occurring products in orders. This can help in understanding which products are likely to be purchased together.
- ▶ Overall, understanding these association rules enables businesses to tailor their marketing strategies, optimize product placement, and enhance customer satisfaction by offering relevant product recommendations based on observed purchasing patterns.

Support, Confidence, and Lift

Support

- Support is a measure of how frequently an itemset (combination of items) appears in the dataset. It indicates the popularity or occurrence of the itemset relative to the total number of transactions.
- Mathematically, support is calculated as the number of transactions containing the itemset divided by the total number of transactions.
- High support values indicate that the item appears frequently, making it a strong candidate for association rule mining.

Confidence

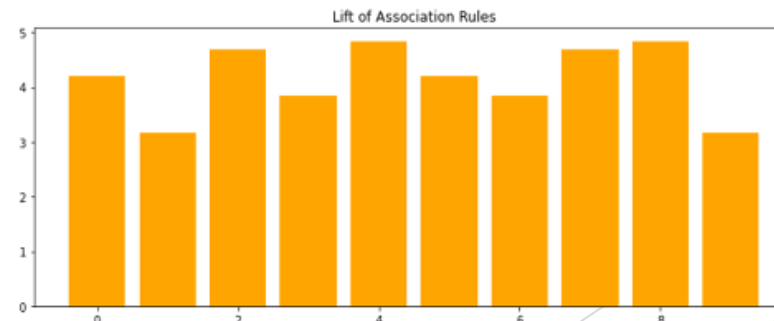
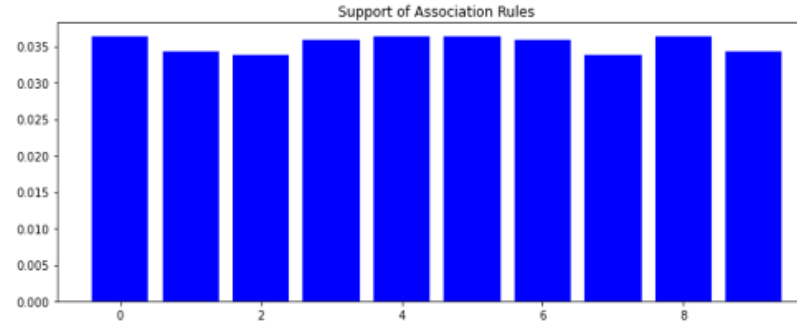
- Confidence is a measure of the reliability or certainty of an association rule. It indicates the likelihood that the presence of the antecedent will result in the presence of the consequent.
- Mathematically, confidence is calculated as the number of transactions containing both the antecedent and consequent divided by the number of transactions containing the antecedent.
- High confidence values suggest a strong association between the antecedent and consequent, implying a higher probability of the consequent item being purchased when the antecedent item is present.

Lift

- Lift measures the strength of association between the antecedent and consequent in an association rule, relative to their individual occurrence frequencies.
- It compares the observed support of the itemset with the expected support if the antecedent and consequent were independent of each other.
- A lift value greater than 1 indicates that the antecedent and consequent appear together more often than expected by chance, suggesting a positive association. A lift value less than 1 indicates a negative or unlikely association, while a lift value equal to 1 indicates independence.

Association Rules For Different Products

Antecedent	Consequent	Confidence	Lift	Support
[A2]	[A5]	0.34	4.21	0.0363
[A2]	[A11]	0.32	3.16	0.0343
[A2]	[A3]	0.32	4.69	0.0338
[A2]	[A1]	0.34	3.84	0.0358
[B10]	[B13]	0.36	4.85	0.0363
[A5]	[A2]	0.45	4.21	0.0363
[A1]	[A2]	0.41	3.84	0.0358
[A3]	[A2]	0.50	4.69	0.0338
[B13]	[B10]	0.49	4.85	0.0363
[A11]	[A2]	0.34	3.15	0.0343



Observations

► For instance, consider the rule $[A2] \rightarrow [A5]$:

- This rule has a confidence of 0.34, indicating that 34% of transactions containing item A2 also contain item A5.
- The relatively moderate confidence suggests a moderate level of predictability in the occurrence of item A5 given the presence of item A2.
- The lift value of 4.21 indicates a significant positive association between A2 and A5, implying that the occurrence of A2 increases the likelihood of A5 being purchased more than four times compared to random chance.

► Conversely, examining the rule $[A5] \rightarrow [A2]$:

- This rule exhibits a higher confidence of 0.45 compared to the previous rule, indicating a stronger association between A5 and A2.
- The lift value of 4.21 further confirms the positive association, suggesting that the occurrence of A5 increases the likelihood of A2 being purchased by approximately 4.21 times compared to random chance.

Observations

► Considering the rule [B10] -> [B13]:

- With a confidence of 0.36, it implies that 36% of transactions containing item B10 also contain item B13.
- The lift value of 4.85 indicates a substantial positive association between B10 and B13, suggesting that the occurrence of B10 increases the likelihood of B13 being purchased by approximately 4.85 times compared to random chance.

► Exploring the rule [B13] -> [B10]:

- This rule has a similar confidence of 0.49, indicating a strong association between B13 and B10.
- The lift value of 4.85 confirms the positive association, suggesting that the occurrence of B13 increases the likelihood of B10 being purchased by approximately 4.85 times compared to random chance.

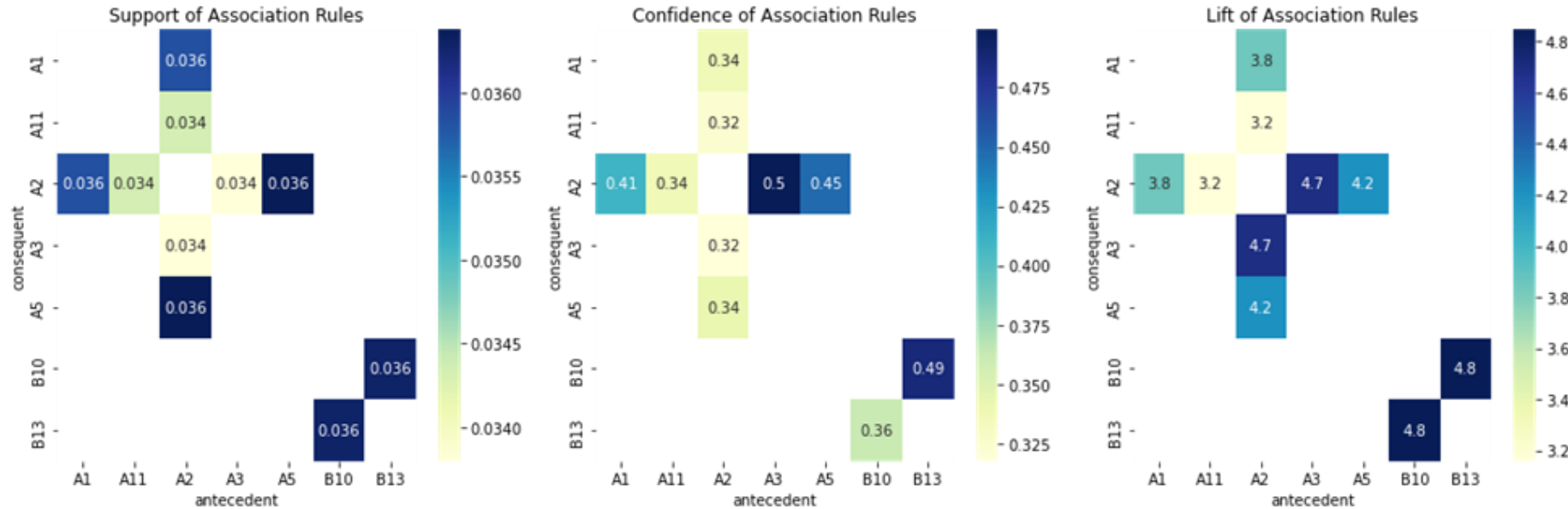
► Examining the association rule [A1] -> [A2]:

- With a confidence of 0.41, it suggests that 41% of transactions containing item A1 also contain item A2.
- The lift value of 3.85 indicates a positive association, implying that the occurrence of A1 increases the likelihood of A2 being purchased by approximately 3.85 times compared to random chance.

► Analyzing the association rule [A3] -> [A2]:

- This rule exhibits a confidence of 0.50, indicating that 50% of transactions containing item A3 also contain item A2.
- The lift value of 4.69 confirms a strong positive association between A3 and A2, suggesting that the occurrence of A3 increases the likelihood of A2 being purchased by approximately 4.69 times compared to random chance.

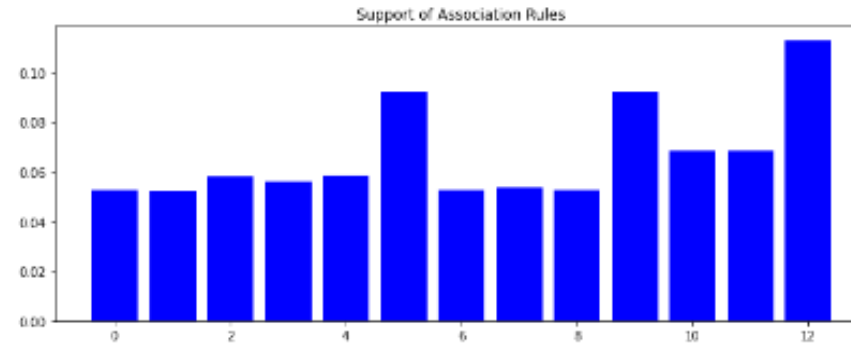
Heatmap for Association Rules



Maximum support will never exceed the percentage of buyers. Confidence is a more important measure for the online retailer because the ability to successfully predict a small percentage of buying sessions as a key to implementing a service strategy focused on most profitable customers [6]

Association Between Colour & Main Category

Antecedent	Consequent	Confidence	Lift	Support
[4,6]	[4,2]	0.59	3.30	0.053
[4,6]	[4,4]	0.58	3.14	0.052
[1,4]	[1,3]	0.61	1.66	0.058
[3,2]	[3,14]	0.55	2.95	0.056
[1,14]	[1,3]	0.56	1.53	0.059
[1,1]	[1,3]	0.59	1.61	0.092
[4,9]	[4,2]	0.55	3.11	0.053
[4,9]	[4,4]	0.56	3.05	0.054
[4,14]	[4,2]	0.55	3.11	0.052
[2,12]	[2,2]	0.55	2.48	0.092
[1,8]	[1,3]	0.78	2.13	0.068
[2,9]	[2,2]	0.56	2.50	0.068
[1, 2]	[1,3]	0.62	1.69	0.11



Clustering

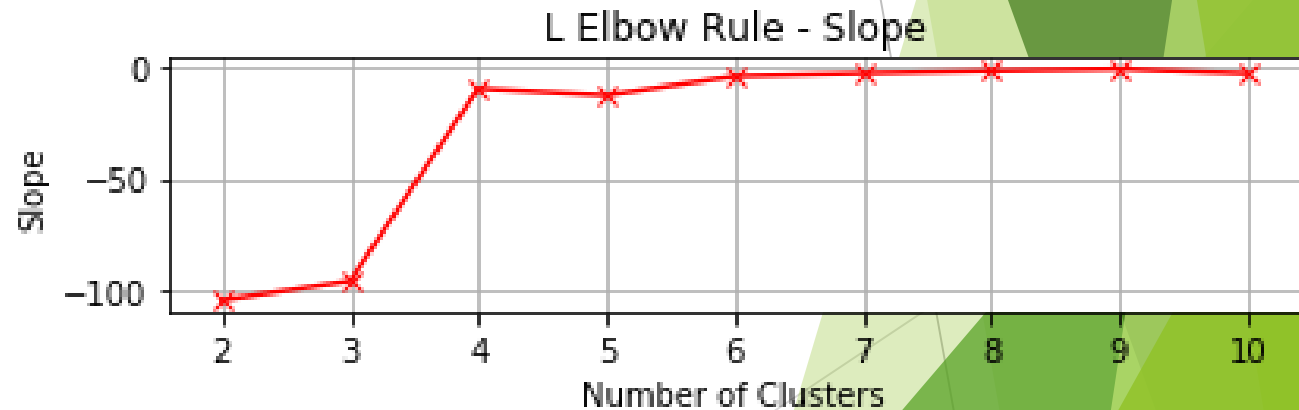
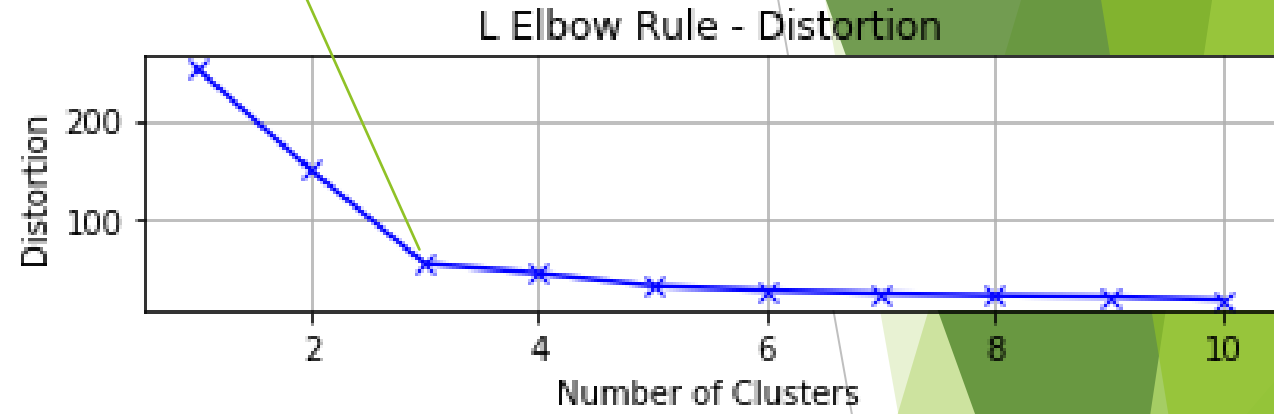
Importance of Clustering Analysis in Online Sales

- ▶ Online sales data often contains a wide range of price values.
- ▶ Clustering helps identify distinct groups or segments within these price values. Clustering allows us to group similar price points together.
- ▶ This helps in understanding customer behavior and preferences based on their purchasing habits.
- ▶ By clustering price values, we can analyze how different customer segments respond to changes in prices.
- ▶ This enables targeted pricing strategies to maximize revenue.
- ▶ Clustering analysis is essential for understanding the complexities of price dynamics in online sales.
- ▶ It enables data-driven decision-making for pricing strategies and enhances customer satisfaction, help to identify potential customers and determine recommendations for customers.

Elbow Method

- ▶ To find the optimal value of clusters
- ▶ It executes the K-means clustering algorithm on the dataset for different values of K (in range 1- 10).
- ▶ For every value of K, it calculates the WCSS value. (Within Cluster Sum of Squares, which gives the total variations within a certain cluster)
- ▶ Plots a curve between calculated WCSS values and the number of clusters K.
- ▶ The point of bend or a point of the plot appears to be like an arm, then that point is considered the best value of K. [4]

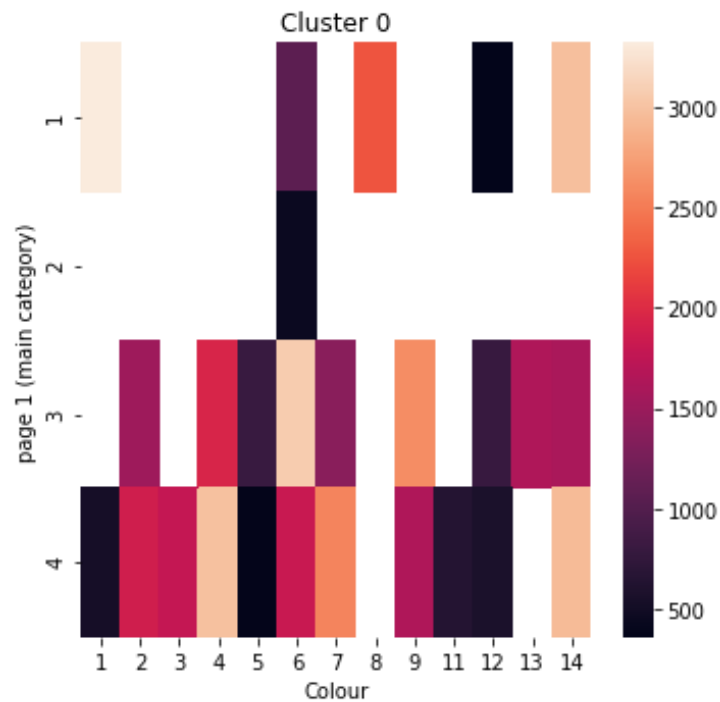
Optimal Number of clusters: 3



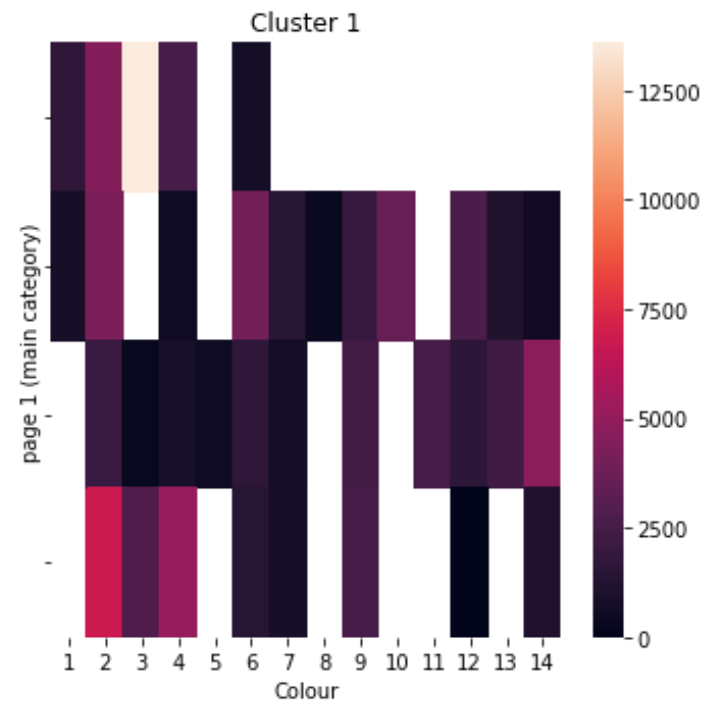
Clustering on price range:

- ▶ Data is clustered based on price feature, into 3 clusters:
 - Cluster 0 Center: Feature price: 33.12 --- > lower prices
 - Cluster 1 Center: Feature price: 62.25 --- > higher prices
 - Cluster 2 Center: Feature price: 46.51 --- > average prices
- ▶ after clustering each feature within the clusters can be analyzed
- ▶ For instance, for each price levels:
 - we can find out the demand level of each product category in different colours.
 - we can analyze the interest of each country in different colours
 - estimate number of interested customers in each color separated by country
 - pricing strategy and demand level of each product category and colour in different countries
- ▶ The Silhouette Score is used to assess the clustering algorithm. it measures of how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1. here the Silhouette Score is equal 0.594 indicates that the clusters are well separated, and the objects within each cluster are relatively cohesive. This is generally considered a good score, indicating strong cluster structure in the data. [1]

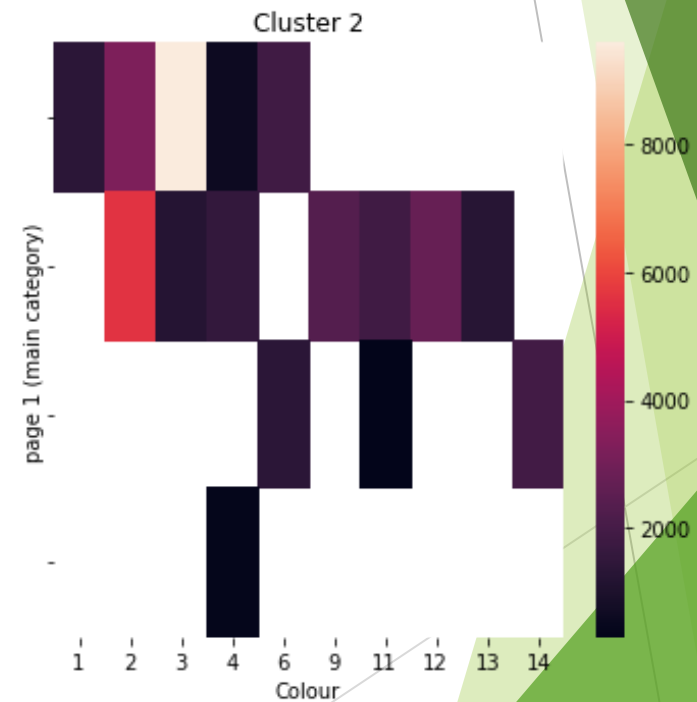
Order frequency in each cluster grouped by colour and main category



○ Cluster 0 --- > lower prices

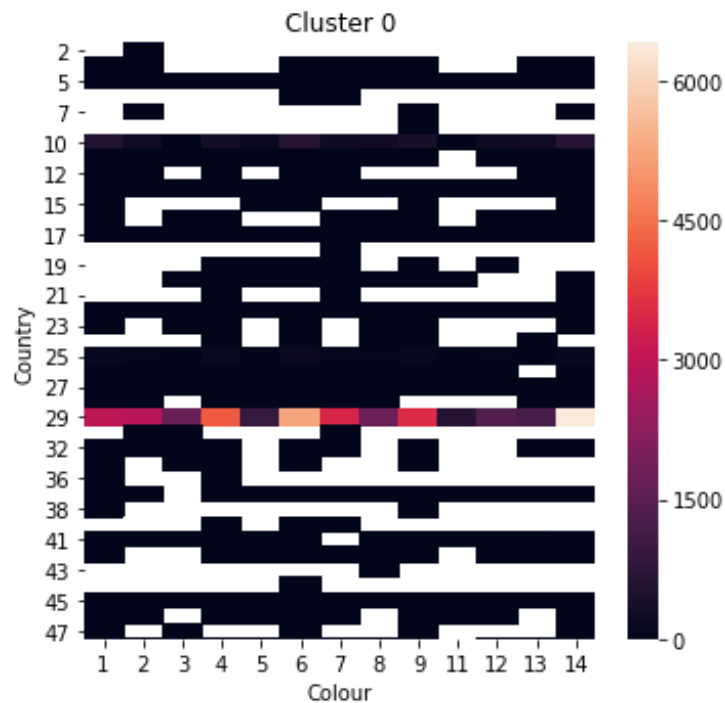


Cluster 1 --- > higher prices

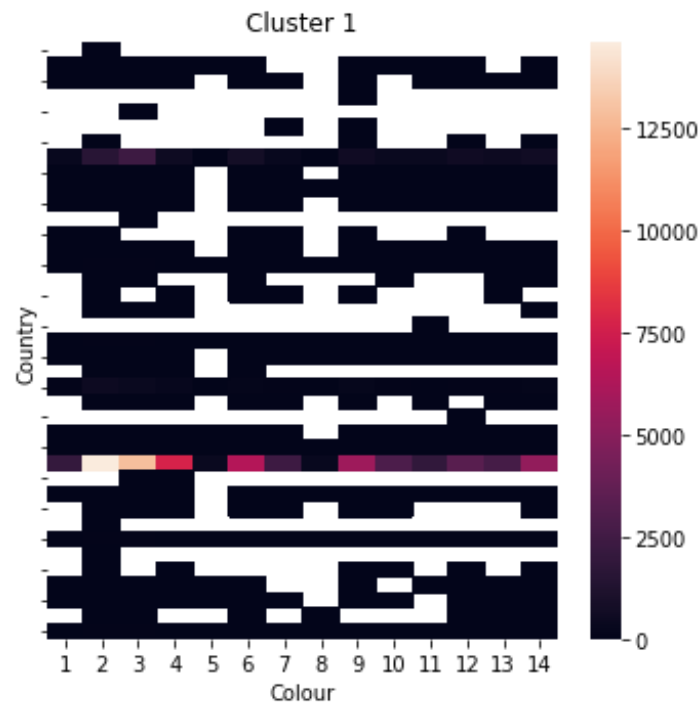


Cluster 2 --- > average prices

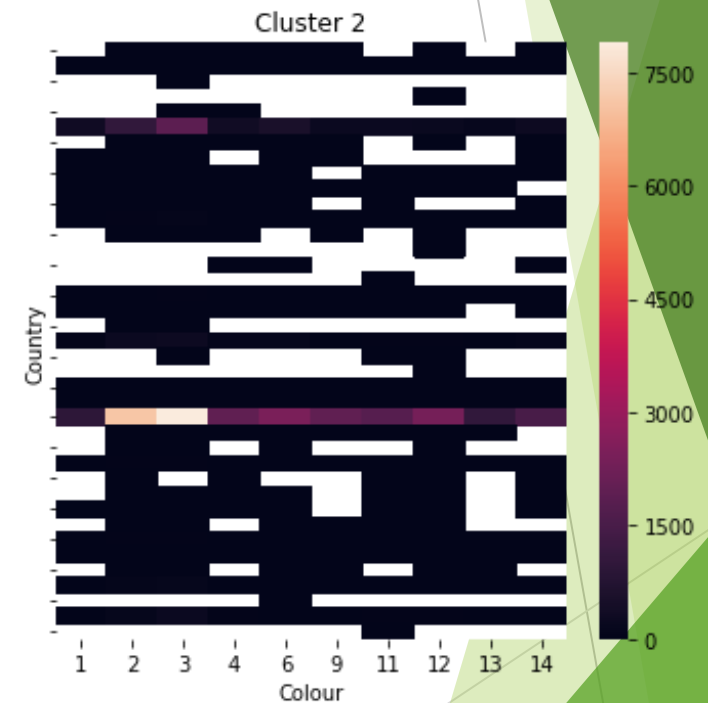
Order frequency in each cluster grouped by colour and country



○ Cluster 0 --- > lower prices

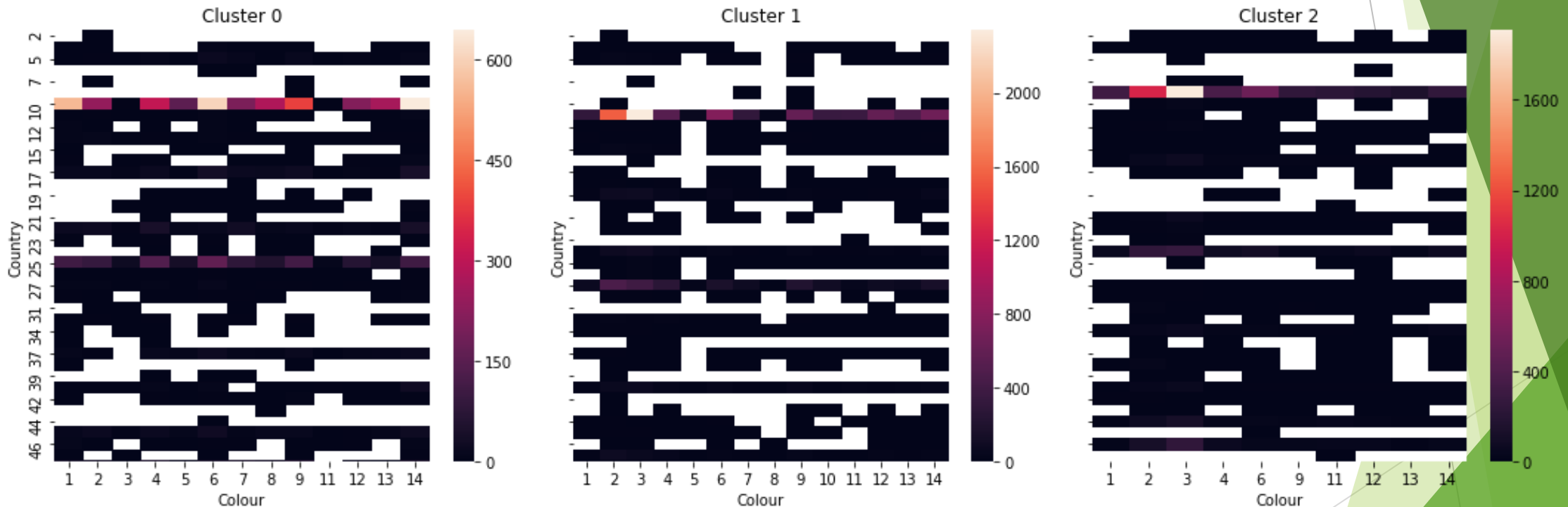


Cluster 1 --- > higher prices



Cluster 2 --- > average prices

Order frequency in each cluster grouped by colour and country



○ Cluster 0 --- > lower prices

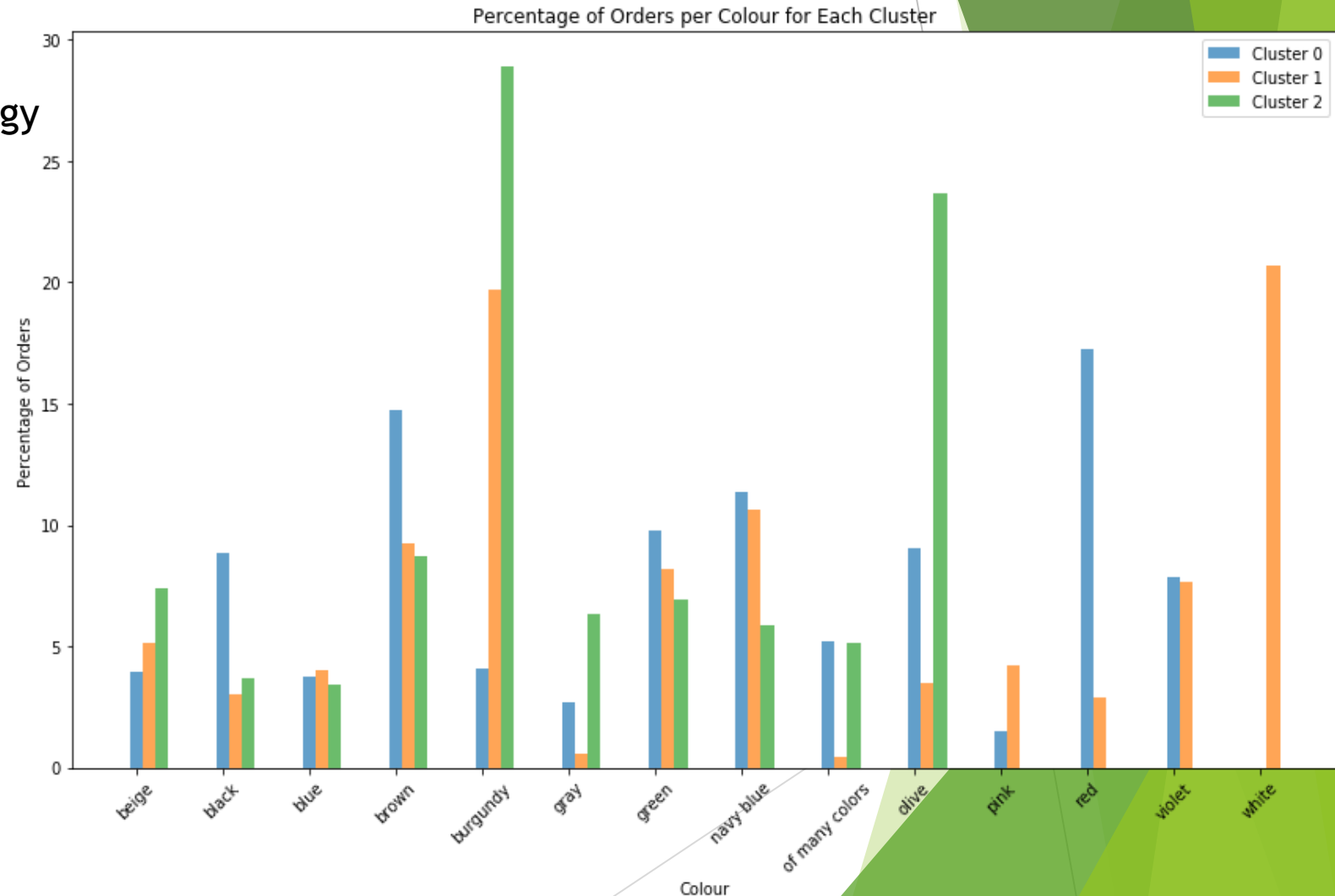
Cluster 1 --- > higher prices

Cluster 2 --- > average prices

- Here the Poland Country has been filtered out, as most of purchases were from this country
- So, we can visualize better other countries data

Order percentage in each cluster grouped by colour

- For product design and pricing strategy
- which color is more likely to be purchased in each price level:
 - higher prices white and burgundy
 - Lower prices: brown and red
 - Average prices: olive and burgundy



○ Cluster 0 --- > lower prices

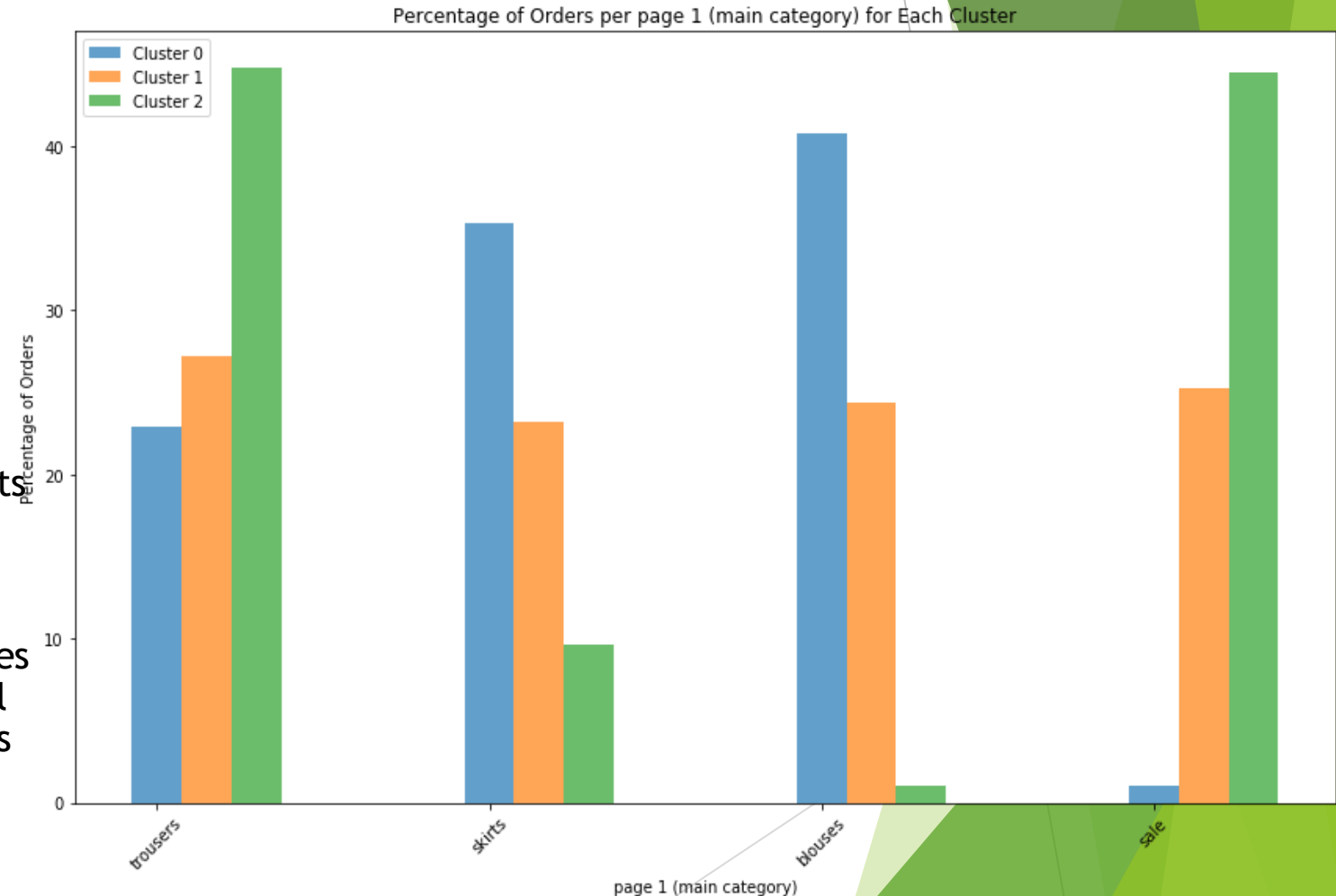
Cluster 1 --- > higher prices

Cluster 2 --- > average prices

Order percentage in each cluster grouped by main category

- For production and pricing strategy
- which product is more likely to be purchased in each price level:
 - higher prices: even distribution
 - Lower prices: blouses and skirts
 - Average prices: trousers and sales products

In the sales category, it is interesting to say customers are more likely to buy higher prices which has a discount on the lower price level than buying average products on lower prices



○ Cluster 0 --- > lower prices

Cluster 1 --- > higher prices

Cluster 2 --- > average prices

References and Related Work

Researchers have explored various methodologies to predict user behavior and enhance the online shopping experience, in the field of clickstream analysis for e-shops.

- ▶ Statistical review and variable correlations [3][4]
- ▶ Machine learning (ML) and Predictive modelling techniques have been widely employed to predict user preferences, anticipate future actions, and personalize recommendations based on clickstream data. also to forecast purchase likelihood, enhancing marketing strategy and sales management.[1][2][9]
- ▶ Association rule mining has been utilized to uncover patterns in user navigation paths, identifying frequently co-occurring pages or products to facilitate cross-selling and upselling strategies. [6][7][10]
- ▶ Clustering algorithms have also found application in segmenting users based on their browsing behavior, allowing for targeted marketing campaigns and tailored product offerings. Overall, the integration of ML, association rule mining, prediction, and clustering techniques has enabled e-shops to optimize marketing strategies, enhance user experiences, and drive business growth.[3][4]

References and Related Work

- ▶ [1] Necula SC. Exploring the Impact of Time Spent Reading Product Information on E-Commerce Websites: A Machine Learning Approach to Analyze Consumer Behavior. Behav Sci (Basel). 2023 May 23;13(6):439. doi: 10.3390/bs13060439. PMID: 37366691; PMCID: PMC10294865.
- ▶ [2] Koehn, Dennis, Stefan Lessmann, and Markus Schaal. "Predicting online shopping behaviour from clickstream data using deep learning." Expert Systems with Applications 150 (2020): 113342.
- ▶ [3] Sakalauskas, V.; Kriksčiuniene, D. Personalized Advertising in E-Commerce: Using Clickstream Data to Target High-Value Customers. Algorithms 2024, 17, 27. <https://doi.org/10.3390/a17010027>
- ▶ [4] Jain, Animesh, and Ashish Kumawat. "ANALYSIS OF CLICKSTREAM DATA." (2022).
- ▶ [5] Yilmazcan Ozyurt, Tobias Hatt, Ce Zhang, and Stefan Feuerriegel. 2022. A Deep Markov Model for Clickstream Analytics in Online Shopping. In Proceedings of the ACM Web Conference 2022 (WWW '22). Association for Computing Machinery, New York, NY, USA, 3071-3081. <https://doi.org/10.1145/3485447.3512027>
- ▶ [6] Suchacka, Grażyna, and Grzegorz Chodak. "Using association rules to assess purchase probability in online stores." Information Systems and e-Business Management 15 (2017): 751-780.
- ▶ [7] Raphaeli, Orit, Anat Goldstein, and Lior Fink. "Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach." Electronic commerce research and applications 26 (2017): 1-12.
- ▶ [8] Triandini, Evi, I. Gede Suardika, and I. Ketut Putu Suniantara. "Database Click Stream of E-commerce Functional." MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer 21.1 (2021): 75-86.
- ▶ [9] Fernandes, Ricardo Filipe, and Costa Magalhães Teixeira. "Using clickstream data to analyze online purchase intentions." (2015).
- ▶ [10] Dinucă, C. E. "An application for clickstream analysis." (2011). INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS Issue 1, Volume 6, 2012 68

Contribution

- ▶ Equal contribution
- ▶ Face-to-face meetings

Thank you for your time and attention!