CAPSTONE (BA-64099-021)

CAPSTONE PROJECT REPORT:

CUSTOMER SENTIMENT & RETAIL SALES FORECASTING

SUBMITTED TO:

Dr. Rouzbeh Razavi

Associate Professor & Director of MS in Business

SUBMITTED BY:

Priscilla Shrestha

811014289

07/13/2025

1. EXECUTIVE SUMMARY

This capstone project is dedicated to the exploration of using data analysis and machine learning to acquire insightful findings about customer attitude and retail sales patterns. The study resorts to the application of sophisticated analytical tools to extract the key insight into consumer preferences, rank the best-performing product groupings, and forecast the future sales trends, based on two complementary datasets: one consists of women fashion products reviews, the other is made of the synthetic retail transactions data.

The major goals comprised determining high-performing categories, predicting your demand, and segmenting the products to get strategic information. Some of the techniques applied included TextBlob sentiment research, Facebook Prophet to estimate the time-series, and KMeans to segment. The Python and Tableau were used to develop visualizations.

The result was an exceptionally positive customer attitude, especially on Tops and Dresses. According to sales data, the categories Electronics and Sports & Outdoors were leading the revenues. The projections were showing that the Fashion category would remain stable and through clustering, there was an evident product segment that would be used in merchandising. These insights were drawn by means of Tableau dashboards to deliver them in an understandable form.

The results can be used to inform practical plans of inventory management, business promotion, and operational effectiveness and demonstrate how technical models can realise practical business benefits.

2. INTRODUCTION AND OBJECTIVES

Retail fashion and product industry faces challenges in predicting demand and understanding customer feedback. This project aims to close the gaps between customer sentiments and sales forecasting. The main objectives are:

- To perform sentiment analysis on customer reviews to determine satisfactions.
- To predict customer demand using time-series modeling.
- Using clustering techniques to improve management of categories. The retail and business environment targets the need of retail companies to assist them in optimizing and understanding customer behaviour through data-driven insights.

3. METHODOLOGY

a. Data Sources:

- Women's Clothing E-commerce Reviews from Kaggle.
- Synthetic Online Retail Dataset from Kaggle.

b. Tools Used:

- Python
- Tableau
- Google Colab
- Kaggle
- Excel

c. Analytical Techniques:

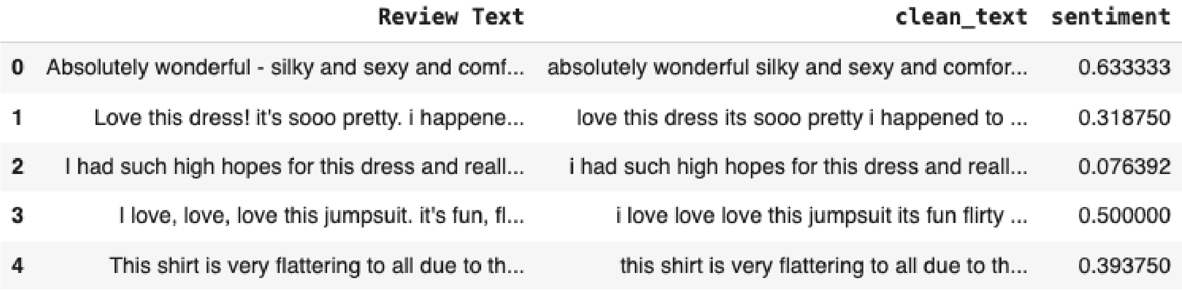
- Text cleaning and sentiment analysis with TextBlob.
- Time-series forecasting with Facebook Prophet.

- KPI aggregation using groupby and summary statistics.
- Product-level aggregation and feature engineering (example; total price, average rating, etc)
- Clustering analysis using K-means.
- Creating SKU-level feature tables.
- Dimensionality reduction using PCA for visualization.
- Visualization using matplotlib, seaborn, Tableau.
- Customer segmentation analysis by different demographic attributes like gender and age.

4. RESULTS AND DISCUSSION

a. Sentiment Analysis

- After cleaning and evaluating a new sentiment score column is generated using TextBlob where to reflect whether a review was positive (closer to 1), neutral (around 0), or negative (closer to -1).



	Review Text	clean_text	sentiment
0	Absolutely wonderful - silky and sexy and comf...	absolutely wonderful silky and sexy and comfor...	0.633333
1	Love this dress! it's sooo pretty. i happene...	love this dress its sooo pretty i happened to ...	0.318750
2	I had such high hopes for this dress and reall...	i had such high hopes for this dress and reall...	0.076392
3	I love, love, love this jumpsuit. it's fun, fl...	i love love love this jumpsuit its fun flirty ...	0.500000
4	This shirt is very flattering to all due to th...	this shirt is very flattering to all due to th...	0.393750

Fig 1 : Sentiment Score

- Reviews in Intimates and Jackets had higher average sentiment scores.

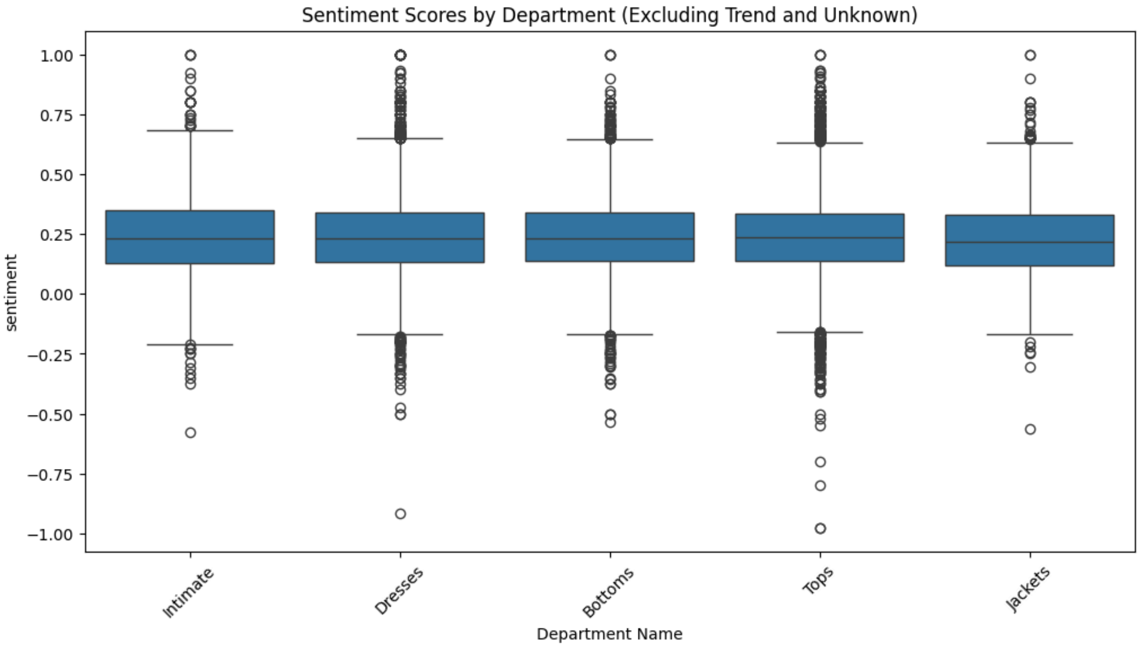


Fig 2 : Sentiment Scores by Department

b. Top Reviewed Products

- Dresses and tops were the most frequently reviewed.
- Python confirmed this with its multi-panel bar charts.

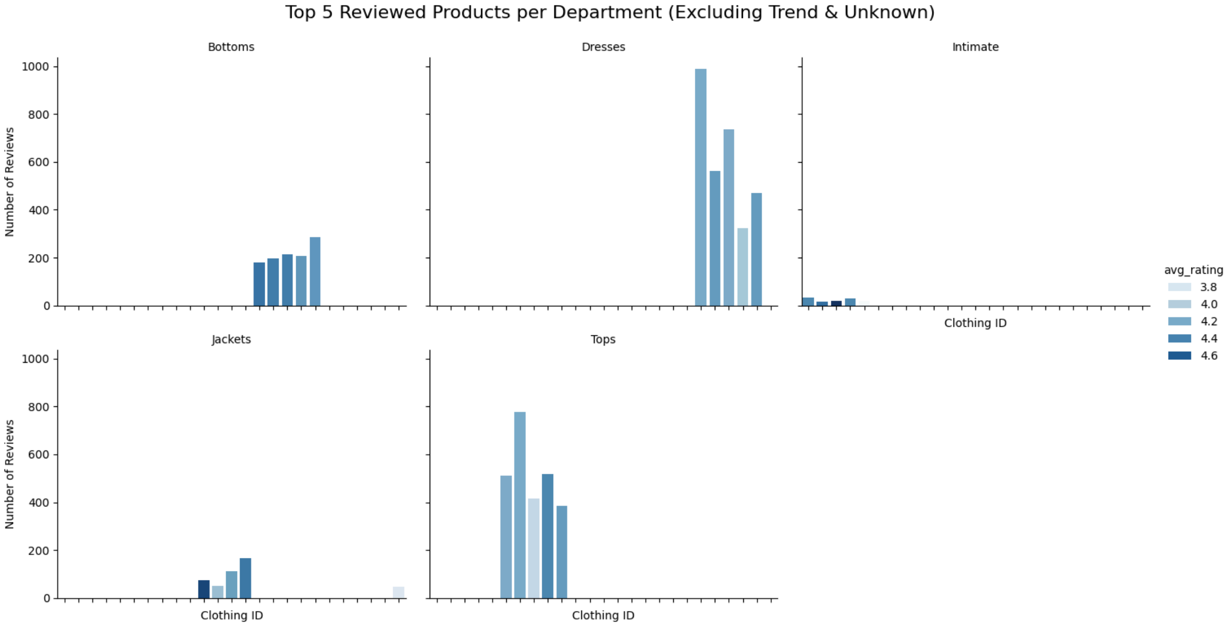
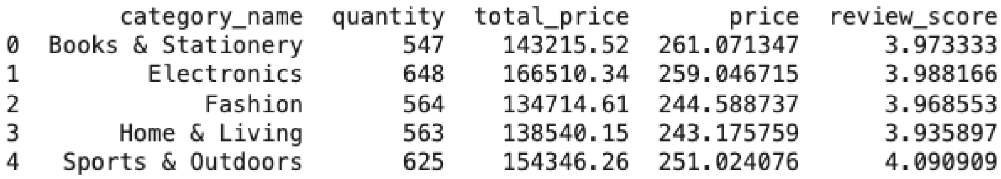


Fig 3: Top 5 Reviewed Products per Department

c. Retail KPIs

- According to the KPI Sports & Outdoors had the highest average review score.



```
       category_name  quantity  total_price       price  review_score
0  Books & Stationery       547    143215.52  261.071347      3.973333
1         Electronics       648    166510.34  259.046715      3.988166
2             Fashion       564    134714.61  244.588737      3.968553
3       Home & Living       563    138540.15  243.175759      3.935897
4   Sports & Outdoors       625    154346.26  251.024076      4.090909
```

Fig 4 : KPI Calculation by Category

- The largest contribution in terms of revenues was electronics

d. Weekly Sales Trends

- The stable sale during the week was observed in the Fashion and Sports categories.

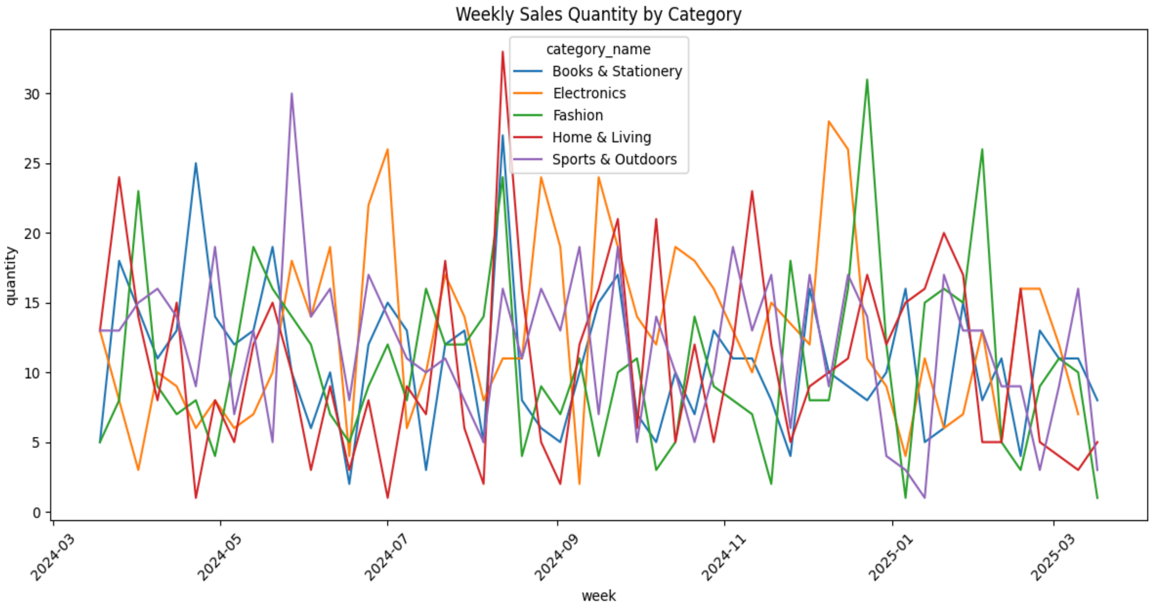


Fig 5: Weekly Sales Quantity by Category

- At Prophet, forecasting for Fashion had steadiness in minor fluctuations.

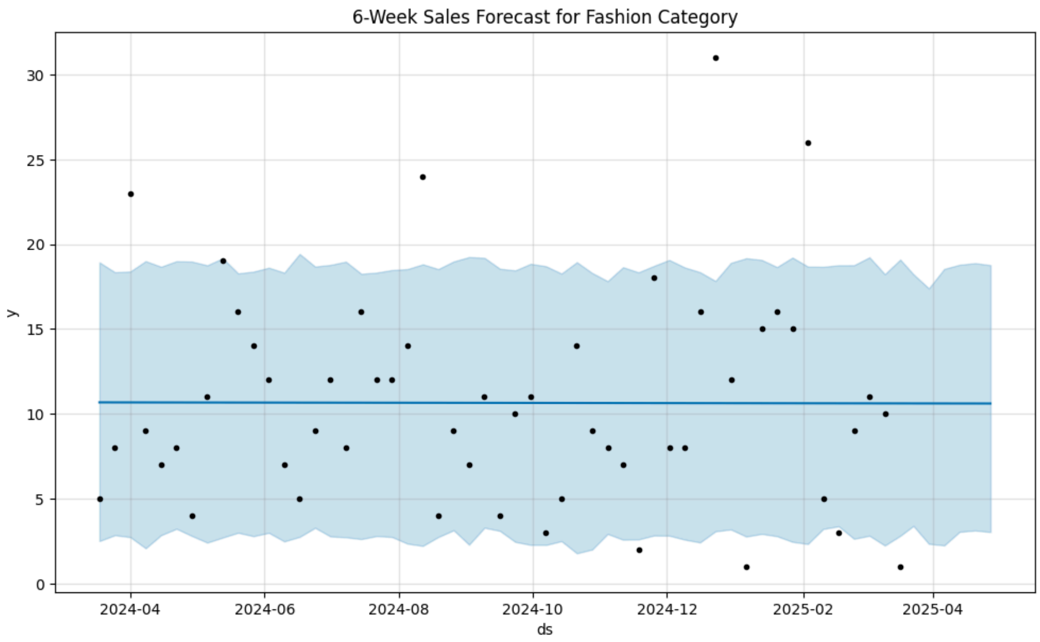


Fig 6 : 6-Week Sales Forecast for Fashion Category

e. Monthly Sentiment Tools

- There were averagely high scores and irregularity with seasons.

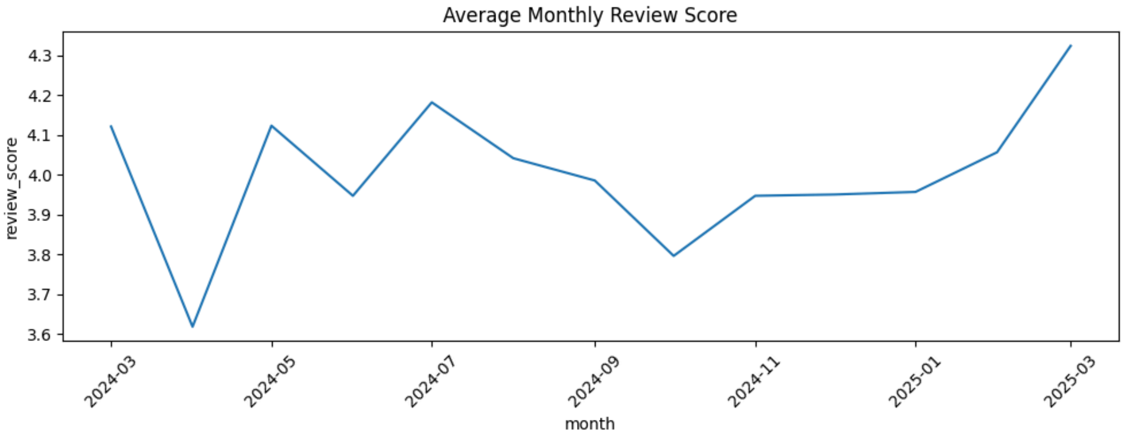
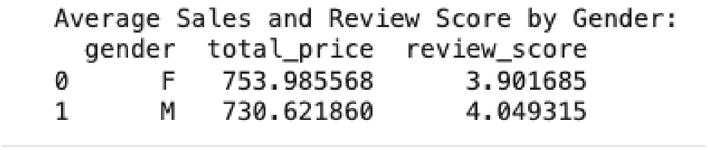


Fig 7 : Average Monthly Review Score

- Small downturns could indicate late deliveries or time-of-the-year performance problems.

f. Customer Demographics

- Even proportionally among males and females, women ended up spending a little bit more per order.



```
Average Sales and Review Score by Gender:
  gender  total_price  review_score
0      F   753.985568      3.901685
1      M   730.621860      4.049315
```

Fig 8 : Average Sales & Review Score by Gender

- Men gave higher review scores on average.

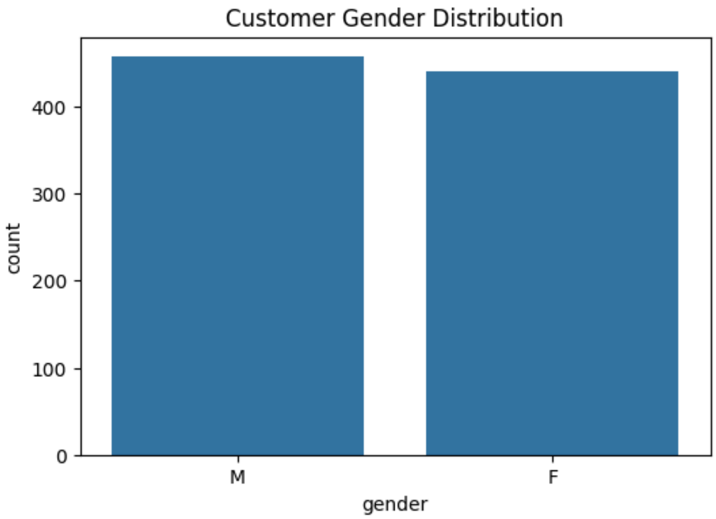


Fig 9 : Gender Distribution

g. Product Clustering

- The products fell in 3 product clusters (high-performing, low-performing, niche).
- The clusters were distinct to view in the PCA scatterplots.

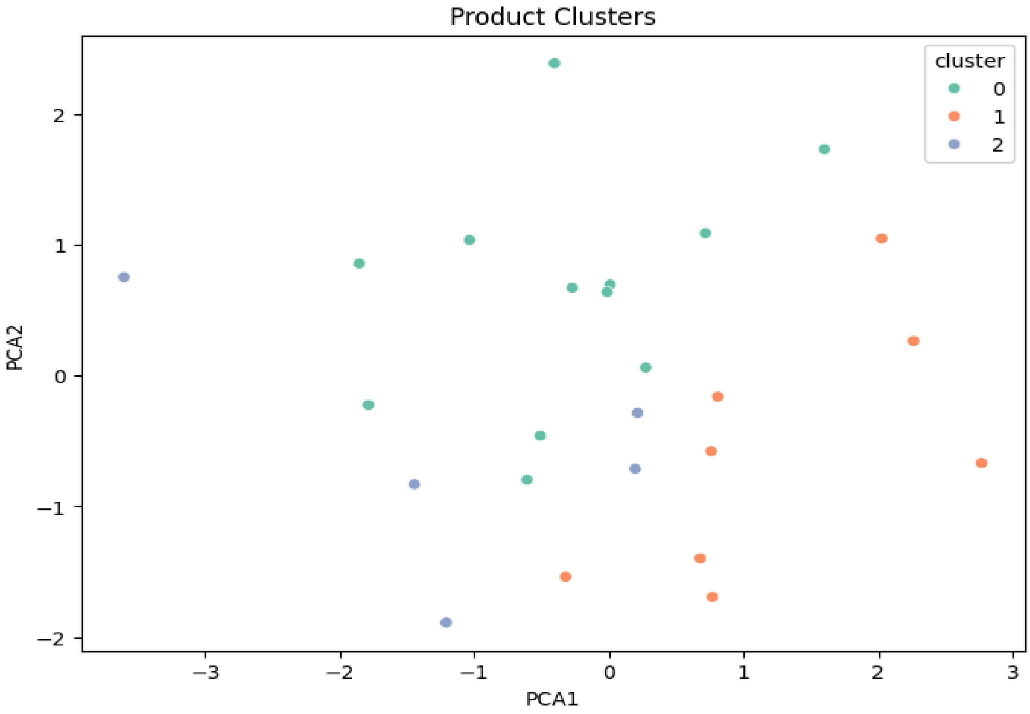


Fig 10 : Product Clusters

5. CONCLUSION

   - The sentiment that customers have is fairly positive, with Tops and Dresses combining in that regard.
   - The Prophet was able to predict the dynamics of weekly sales.
   - Clustering assists in controlling the SKUs based on performance.

6. RECOMMENDATIONS

   - Make sure the most reviewed products are on the top of inventory and the promotions.
   - To enhance the planning of the logistics and production, use sales forecasting.
   - Divide products to concentrate on top-seller or update underperforming products.

7. ETHICAL CONSIDERATION

- Transparency: Users must be clearly informed when review or data are under analysis.
- Bias Mitigation: Sentiment tools might embody some biases; there is to be constant validation.
- Data Privacy: The data used in the retail business should be anonymized so that it does not end up falling in the wrong hands.

Best practices such as ethical AI audits, bias checks and consent notifications are advised.

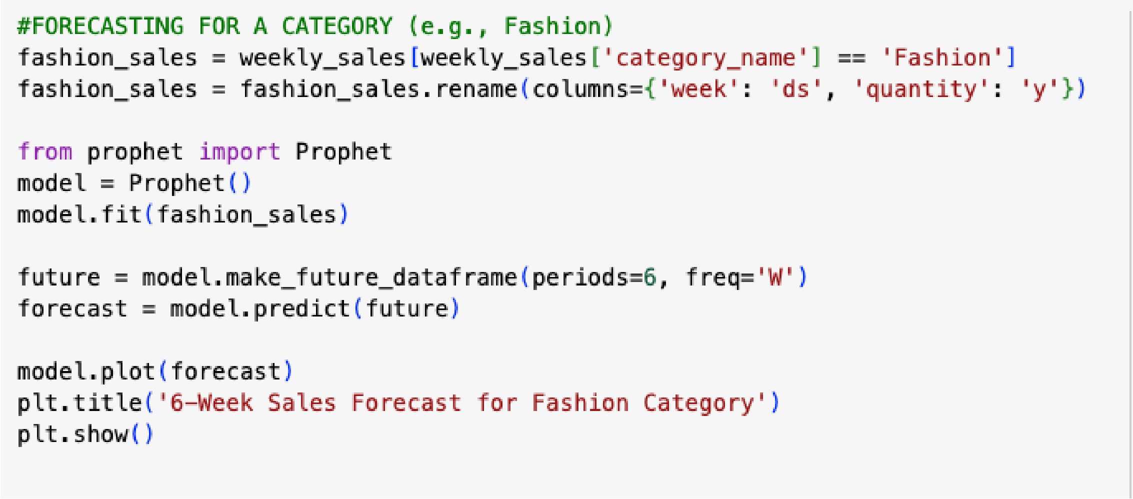
8. REFERENCES

- Brownlee, J. (2019). *Visualizing high-dimensional data using PCA in Python*. Machine Learning Mastery. https://machinelearningmastery.com/principal-components-analysis-for-dimensionality-reduction-in-python/
- Brownlee, J. (2020). *How to cluster data using KMeans in Python*. Machine Learning Mastery. https://machinelearningmastery.com/clustering-algorithms-with-python/
- Facebook Prophet Documentation. (n.d.). *Prophet: Forecasting at scale* [Software]. GitHub. https://github.com/facebook/prophet
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

- Loria, S. (n.d.). *TextBlob: Simplified Text Processing*. https://textblob.readthedocs.io
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281–297. https://projecteuclid.org/euclid.bsmsp/1200512992
- Parmenter, D. (2015). *Key performance indicators: Developing, implementing, and using winning KPIs* (3rd ed.). John Wiley & Sons.
- Scikit-learn developers. (n.d.). *sklearn.cluster.KMeans – Scikit-learn documentation*. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- Tableau Software. (n.d.). *Create KPI visualizations in Tableau*. https://help.tableau.com/current/pro/desktop/en-us/kpi.htm
- Taylor, S. J., & Letham, B. (2018). *Forecasting at scale*. *The American Statistician, 72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

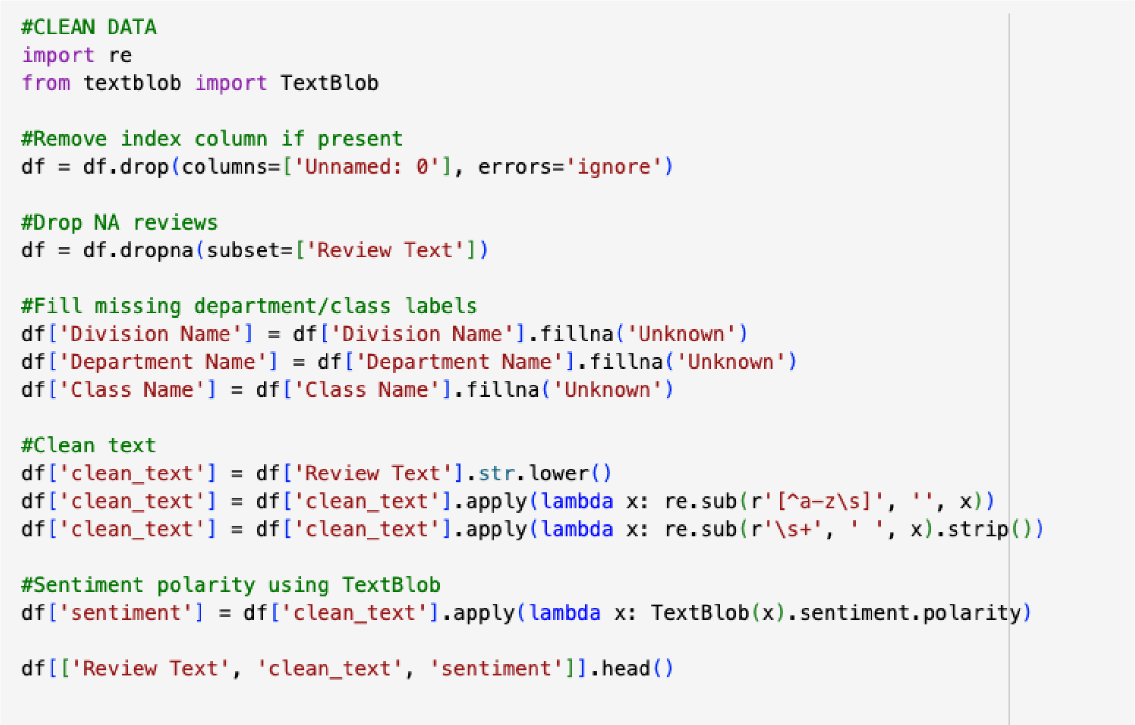
9. APPENDICES

a. Appendix A - Python Code



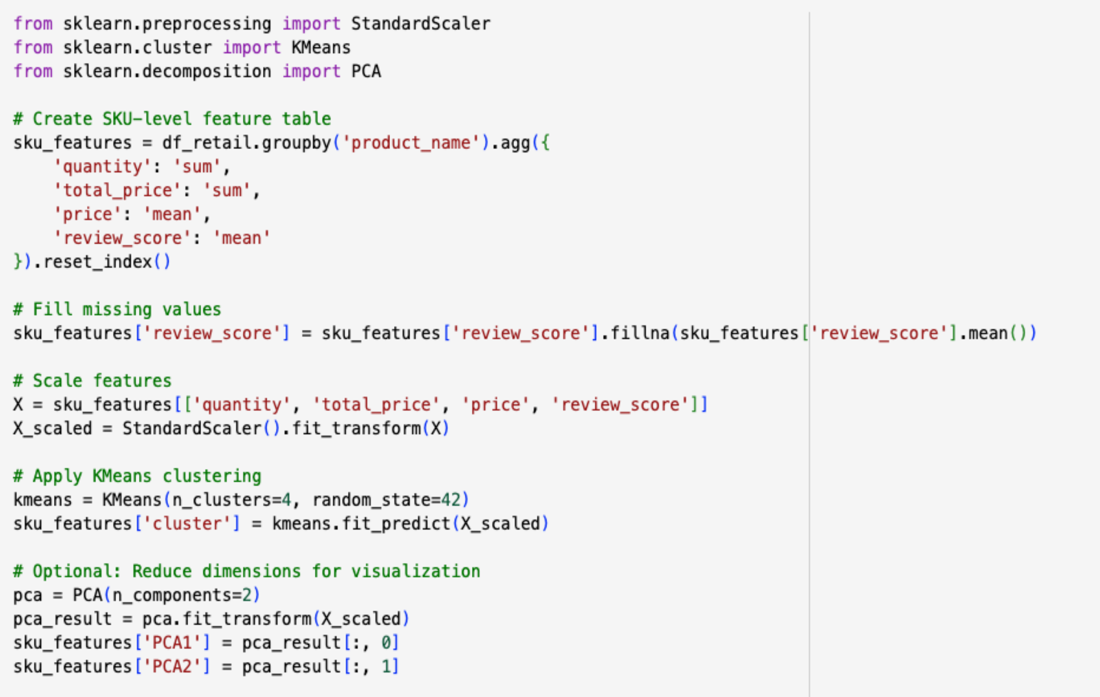
```python
#FORECASTING FOR A CATEGORY (e.g., Fashion)
fashion_sales = weekly_sales[weekly_sales['category_name'] == 'Fashion']
fashion_sales = fashion_sales.rename(columns={'week': 'ds', 'quantity': 'y'})

from prophet import Prophet
model = Prophet()
model.fit(fashion_sales)

future = model.make_future_dataframe(periods=6, freq='W')
forecast = model.predict(future)

model.plot(forecast)
plt.title('6-Week Sales Forecast for Fashion Category')
plt.show()
```

Fig 11 : Prophet Model Code



```python
#CLEAN DATA
import re
from textblob import TextBlob

#Remove index column if present
df = df.drop(columns=['Unnamed: 0'], errors='ignore')

#Drop NA reviews
df = df.dropna(subset=['Review Text'])

#Fill missing department/class labels
df['Division Name'] = df['Division Name'].fillna('Unknown')
df['Department Name'] = df['Department Name'].fillna('Unknown')
df['Class Name'] = df['Class Name'].fillna('Unknown')

#Clean text
df['clean_text'] = df['Review Text'].str.lower()
df['clean_text'] = df['clean_text'].apply(lambda x: re.sub(r'[^a-z\s]', '', x))
df['clean_text'] = df['clean_text'].apply(lambda x: re.sub(r'\s+', ' ', x).strip())

#Sentiment polarity using TextBlob
df['sentiment'] = df['clean_text'].apply(lambda x: TextBlob(x).sentiment.polarity)

df[['Review Text', 'clean_text', 'sentiment']].head()
```

Fig 12 : Sentiment Analysis Code



```
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA

# Create SKU-level feature table
sku_features = df_retail.groupby('product_name').agg({
    'quantity': 'sum',
    'total_price': 'sum',
    'price': 'mean',
    'review_score': 'mean'
}).reset_index()

# Fill missing values
sku_features['review_score'] = sku_features['review_score'].fillna(sku_features['review_score'].mean())

# Scale features
X = sku_features[['quantity', 'total_price', 'price', 'review_score']]
X_scaled = StandardScaler().fit_transform(X)

# Apply KMeans clustering
kmeans = KMeans(n_clusters=4, random_state=42)
sku_features['cluster'] = kmeans.fit_predict(X_scaled)

# Optional: Reduce dimensions for visualization
pca = PCA(n_components=2)
pca_result = pca.fit_transform(X_scaled)
sku_features['PCA1'] = pca_result[:, 0]
sku_features['PCA2'] = pca_result[:, 1]
```

Fig 13 : Clustering Code (KMeans + PCA)

b. Appendix B - Additional Visuals

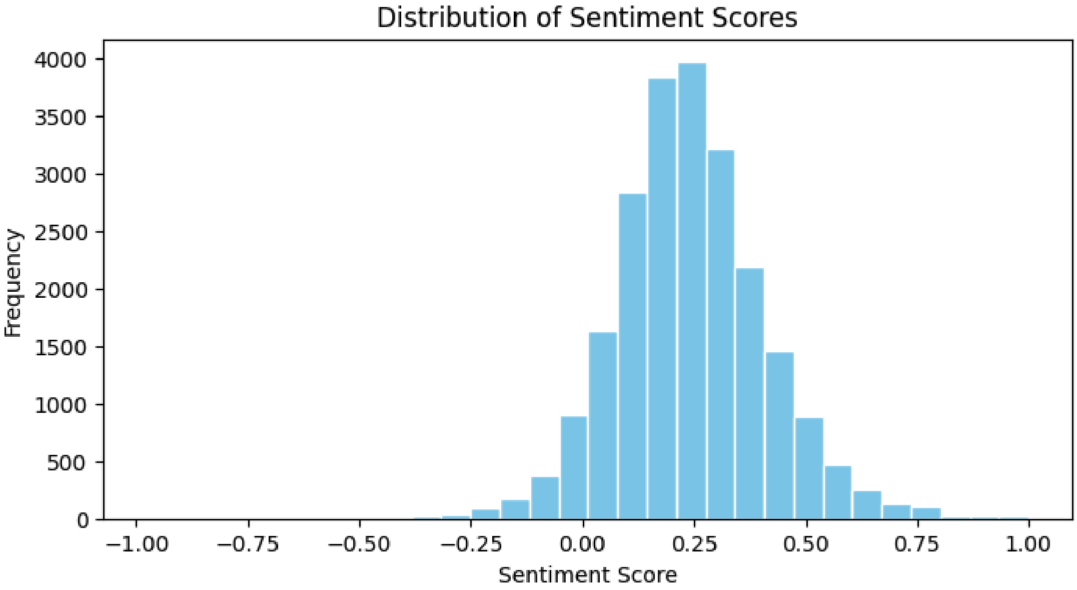


Fig 14 : Distribution of Sentiment Scores

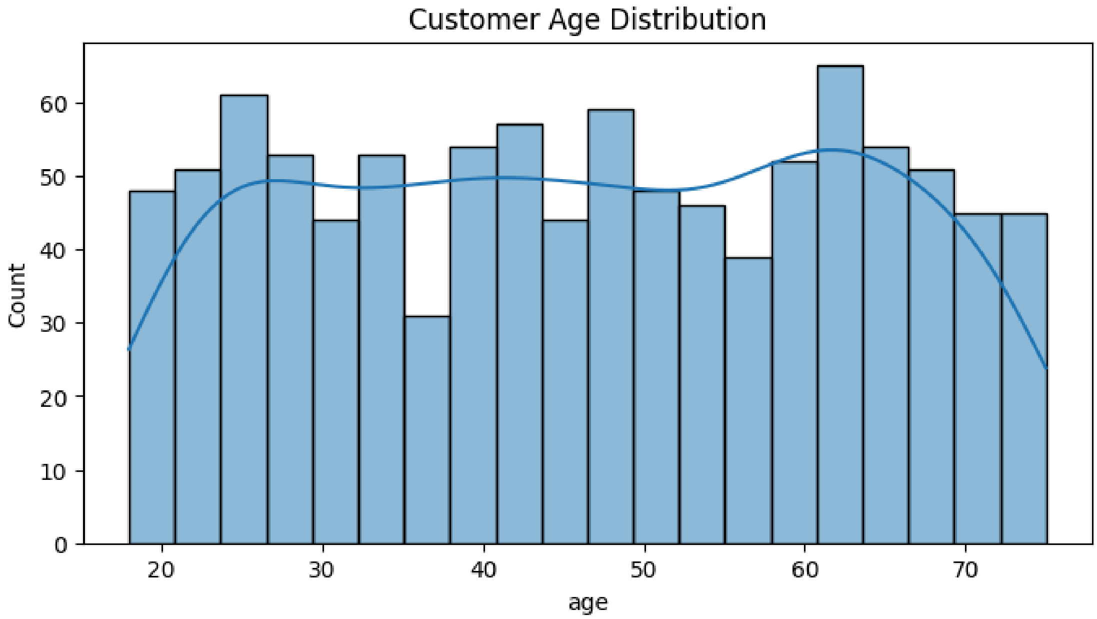
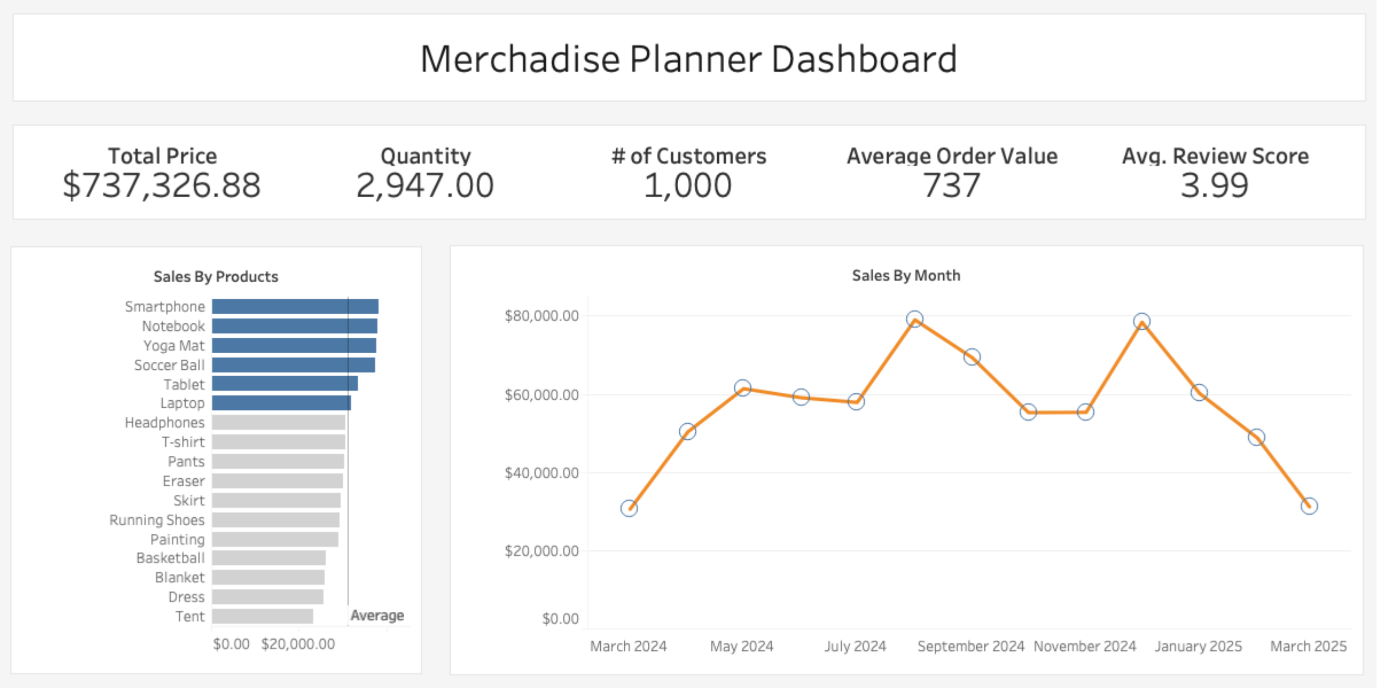


Fig 15 : Customer Age Distribution

According to the *Figure 15* Customer Age Distribution chart above the histogram shows an age group ranging from 18 to 75. We can see a slight peak in the early 20s, late 40s, and early 60s. This chart can help us guide age-targeted marketing strategies.

c. Appendix C - Tableau Dashboard



Merchadise Planner Dashboard

Total Price: $737,326.88

Quantity: 2,947.00

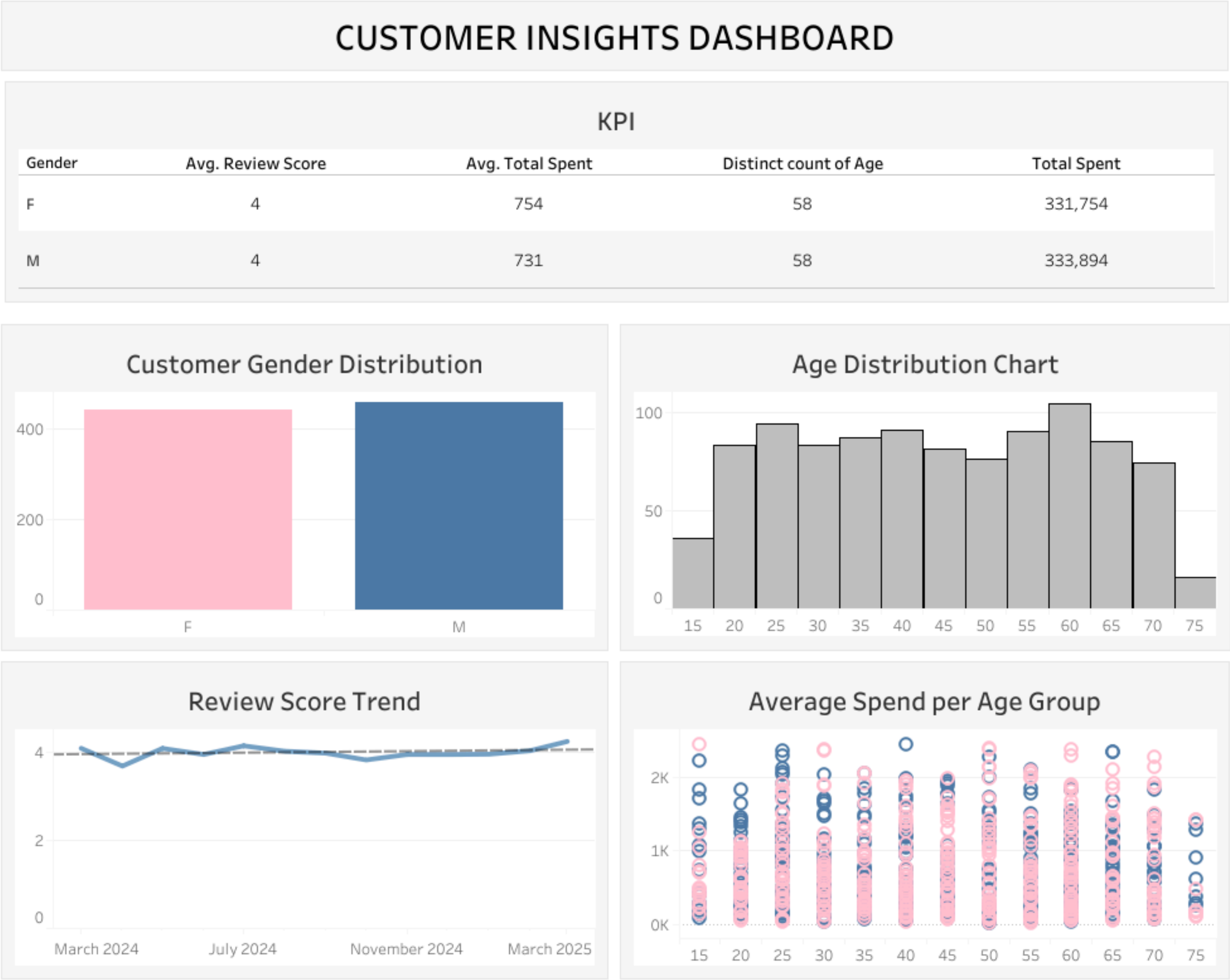
# of Customers: 1,000

Average Order Value: 737

Avg. Review Score: 3.99

Fig 16: Merchandise Planner Dashboard in Tableau

Link to Interactive Tableau Dashboard



CUSTOMER INSIGHTS DASHBOARD

KPI

Gender	Avg. Review Score	Avg. Total Spent	Distinct count of Age	Total Spent
F	4	754	58	331,754
M	4	731	58	333,894

Customer Gender Distribution

Age Distribution Chart

Review Score Trend

Average Spend per Age Group

Fig 17 : Customer Insight Dashboard in Tableau

Link to Interactive Table Dashboard