American University of Armenia

College of Science and Engineering

Business Analytics for Data Science

**Final Project**

Nane Mambreyan

Fall 2023

**Abstract**

For this project, I collected historical data on Adidas sales and customers. For data cleaning, I combined two datasets in one Excel file, looked for NA, and duplicate values, fixed the format of columns and inconsistencies in all platforms (R, Python, Microsoft Excel) that I am going to use. I also generated some features randomly to be able to demonstrate the algorithms that are going to be used throughout the project.

Business Analytics is divided into 3 branches: **Descriptive, Predictive, and Prescriptive**. So, I define the goal of this project in the sense of 3 main components of BA:

**Descriptive** – Creating visualizations with R, Python, and Excel to describe historical data, more specifically some factors that were shaping Adidas sales and customer segments from 2020, to 2021. So, I am going to:

* Analyze Adidas business strategy(SWOT)
* Conduct Univariate Analysis, (Excel)
* Conduct Multivariate Analysis (R)
* Define some KPIs

**Predictive** – After creating visualizations, I will:

1. Conduct RFM analysis

2. Do K-means clustering based on RFM results

3. Do predictions with K-nn analysis based on RFM results

4. Do time series analysis

6. Sentiment analysis

**Prescriptive** - Based on the insights that will be gained in the above-mentioned two stages of the study, I will make some campaign suggestions to increase KPIs and address possible problems beforehand.

**Findings -** As a result of the study, I found out that there were many customers who were quite satisfied with Adidas and still were customers who were at risk. I investigated their characteristics and designed corresponding campaign strategy to attract them.

**Descriptive**

## **Adidas's Business Strategy:**

Adidas, a global sportswear and athletic footwear company, has adopted a differentiation strategy, which involves operating in a mass market with a unique position rather than aiming for the lowest cost. Adidas distinguishes itself through innovative product design and a strong brand image.

Why Adidas Chose the Differentiation Strategy:

*Brand Image***:** Adidas has established a reputation as a stylish and innovative brand. The differentiation strategy allows them to maintain this image and stand out in the competitive sportswear market.

*Collaborations:* Adidas collaborates with celebrities, athletes, and fashion designers to create limited-edition products. That approach supports the differentiation strategy by offering unique and exclusive items.

*Innovation*: Adidas invests in research and development to bring innovative products to the market. The differentiation strategy aligns with their commitment to offering technologically advanced and stylish products.

*Market Positioning*: By differentiating itself, Adidas targets consumers who value quality, style, and performance, allowing the company to command premium prices and build customer loyalty.

## **SWOT Analysis:**

*Strengths:* Strong brand image, Innovation in product design, Global presence and market share, Diverse product portfolio

*Weaknesses:* High prices compared to some competitors, Dependence on external suppliers for materials, Vulnerability to changing trends

*Opportunities:* Growing demand for activewear, Expansion in emerging markets, Increasing focus on sustainability in the industry

*Threats:* Intense competition from Nike and other sportswear brands, Economic downturn affecting consumer spending, Counterfeit products

## **Porter’s Five Forces Analysis:**

*Threat of New Entrants:*Moderate due to the strong brand image.

*Bargaining Power of Suppliers:* Moderate, Adidas has multiple suppliers and a global supply chain.

*Bargaining Power of Buyers:*Moderate/high, buyers have many options, but brand loyalty is a significant factor.

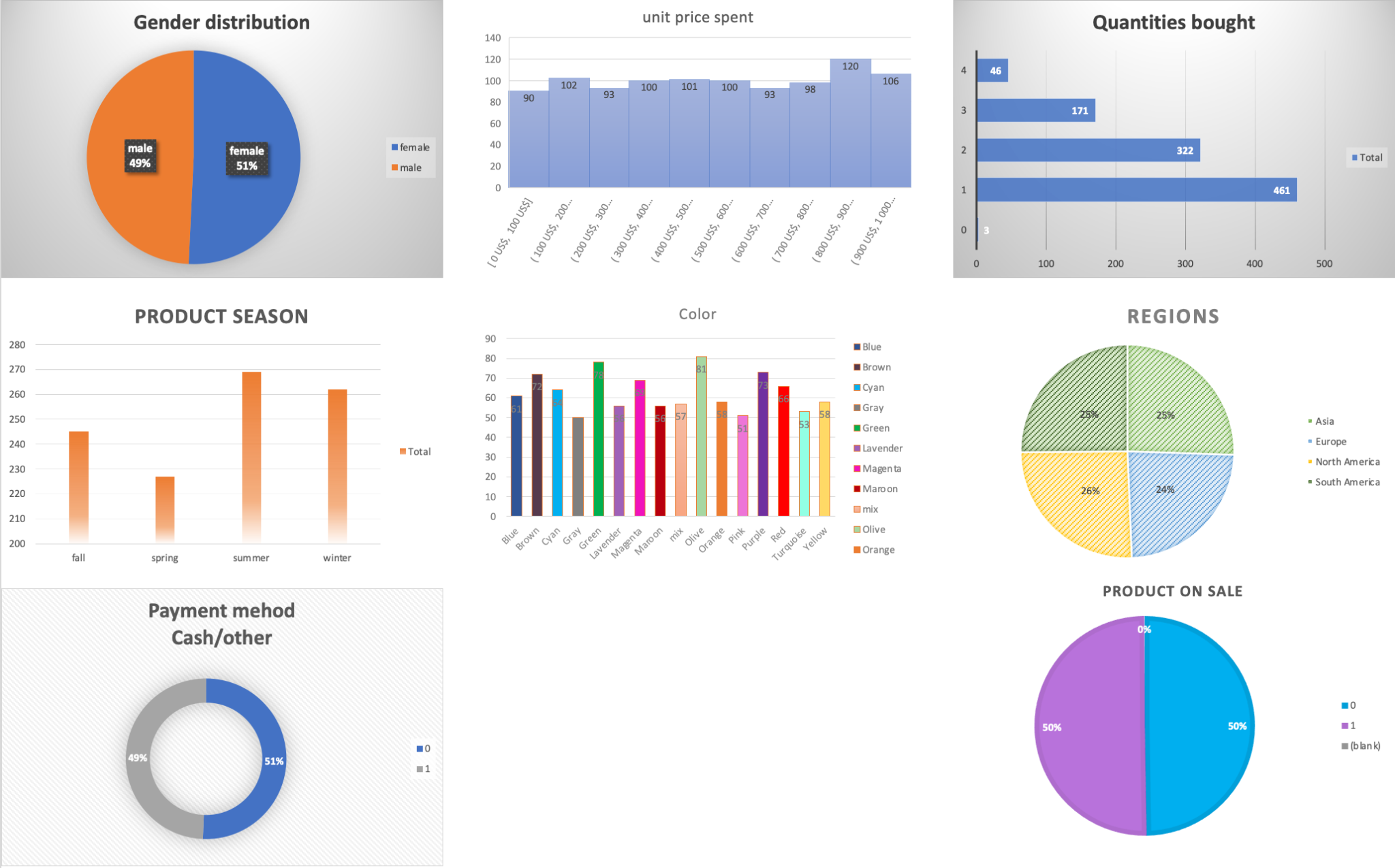
*Threat of Substitute Products:* Moderate; other sportswear brands exist, but Adidas's unique style and branding reduce the threat.

*Competitive Rivalry***:** High due to intense competition, especially with Nike, but Adidas's differentiation strategy helps maintain a competitive edge.

Adidas is positioned in the **“Market Penetration”** sector of Ansoff's Matrix, focusing on increasing market share and sales of its existing sportswear and athletic footwear products in an existing market.

## **Data Manipulation**

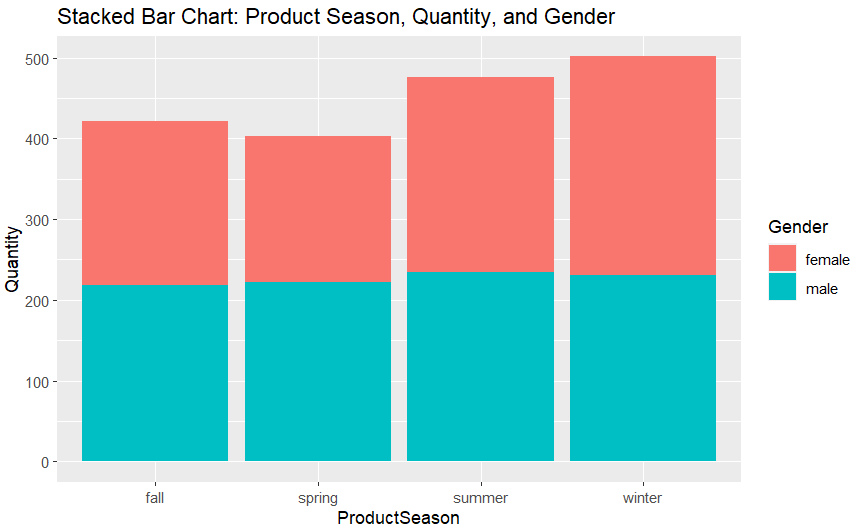
## I started generating the dataset using randomly generated values and original data. Many of the columns were generated using Excel formulas. Specifically, the 'randbetween' function facilitated the creation of columns with randomized numerical values in the given ranges. However, excel randomizes all of its random values when the page refreshes, so the randomly generated values are not constant. To keep the data consistent for analysis, I had to make these random numbers permanent by copying and pasting them again as constant values. I also generated data using Python code. For example, for the Reviews column, I have a Python script that takes the score of a product and generates a positive or negative review based on the score. Mixing these different data types helped to paint a fuller picture for the research.

**Univariate Analysis**

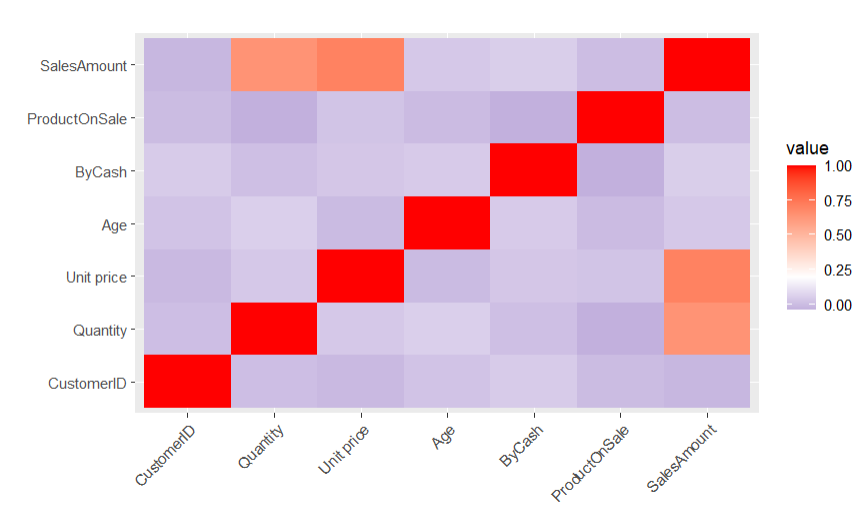
For the univariate analysis, I used Excel’s functionalities to create pivot tables and graphs. Pivot tables helped to explore individual variables ranging from customer details to specific product information. Then, I created graphs and charts based on the pivot tables, and these visualizations helped to understand the data better. The graphs, ranging from piecharts to column charts—were seamlessly integrated to visually elucidate the patterns, trends, and distributions uncovered during the univariate analysis.

The plots are quite simple, nevertheless, each one contains key information about the dataset. All of them show how an individual variable is distributed (e.g. the first piechart above shows what percentage of the customers is male and female).

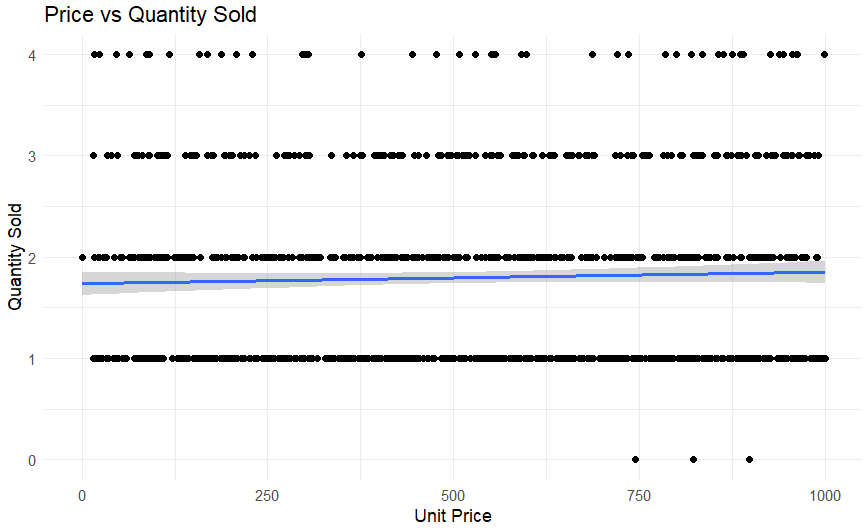
## **Multivariate Analysis**



The stacked bar chart above illustrates product quantities sold across different seasons, segmented by gender. It shows that male purchases (cyan) exceed female purchases (red) in fall and winter. while in spring and summer, the proportions are equal. This seasonal trend suggests varying consumer behavior and demand throughout the year, with potential implications for marketing and inventory planning.



From the correlation matrix heatmap, I can infer that increases in SalesAmount are closely tied to Quantity sold and Unit price rises. This suggests that selling more items at higher prices is a critical driver of revenue.



The line suggests that most sales consist of one or two items per transaction, regardless of the item's unit price. This pattern could indicate that price does not strongly influence the quantity purchased, although it’s evident that purchases of more than 2 items are not as common.

## **Key Performance Indicators**

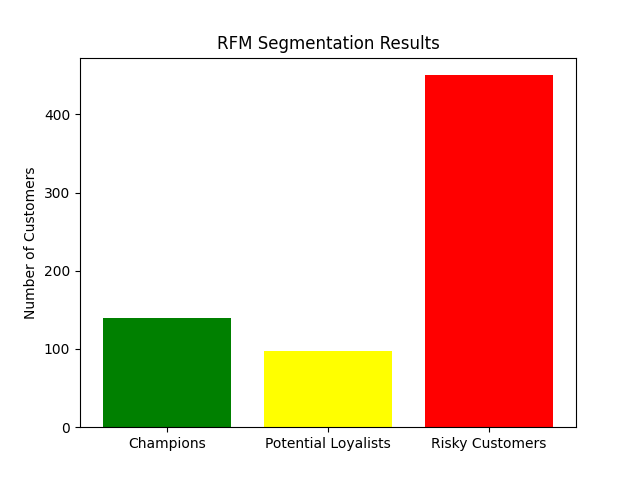
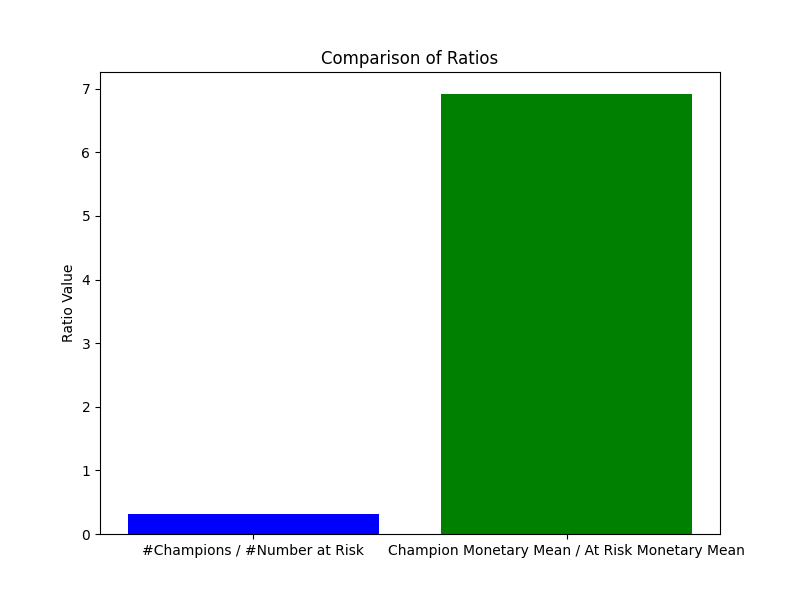
I take several quantitative measures as key performance indicators (KPI) to guide through the prescriptive and predictive stages of the study. Those are:

1. (Number of Champions / Number at Risk) VS (Champion Monetary Mean / At Risk Monetary Mean): Contrasting these two ratios will provide valuable insights into how the monetary aspects of champion customers differ from those of at-risk customers. This analysis can uncover patterns and assist in shaping strategies/campaigns.
2. Customer Satisfaction rate (CSAT), since it is a fundamental measure of how well a business meets customer expectations.

**Predictive Analysis and Clustering**

## **RFM**

I have used RFM analysis for customer segmentation, giving the highest importance (weight) to those customers, who had a higher index for Monetary value which demonstrates the spending power of the customer. Recency and Freqency of the purchase made by the customers were weighted equally. The segmentation is shown on the left barchart.



This result shows that about 150 of the considered customers were classified as Champions, about 100 as Potential Loyalists, and the vast majority, about 450 were classified as At Risk Customers. This is a useful finding since I can concentrate on this segment and try to change this picture with the help of further analysis of that segment and targeting campaigns later on.

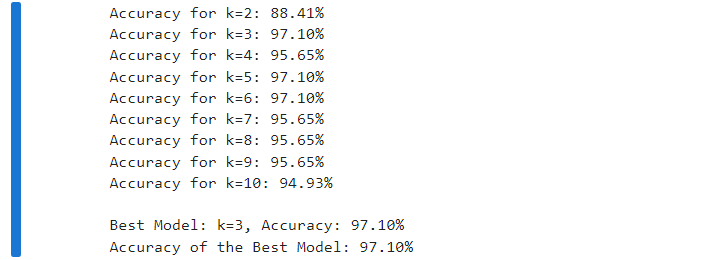
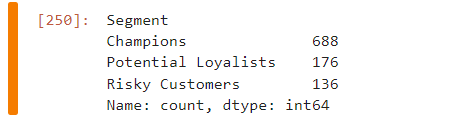
The relatively low proportion of champions (31.11%) compared to those at risk might indicate that a smaller subset of customers is driving a significant portion of the monetary value.

The high ratio (6.92) for the Champion Monetary Mean / At Risk Monetary Mean suggests a substantial difference in spending patterns between champions and those at risk. Champion customers, on average, contribute significantly more to the monetary value compared to the at-risk group.

## **K-nn analysis**

In this part, I am modeling some situation where I have 1000 new customers. I assign them randomly generated Recency, Frequency, and Monetary values and try to understand what will be the frequency of customers in each segment. But before doing that, I am making sure that our methods are reliable y checking the accuracy of the algorithm.

The results demonstrated below mean that the method is reliable in 97% of cases, so I proceed to a randomized experiment and learn that out of 1000 customers, 68.8% are 97% likely to be customers at risk.



## **K-means**

Since 68.8% of 1000 customers are likely at risk, more research is required to target them. Therefore, a K-mean analysis was conducted to explore the risky customers. The customers were divided into groups by their age and salary. According to the silhouette method, when k was equal to 2, the score was 0.68. On the contrary, when k was 3, the score was 0.74.

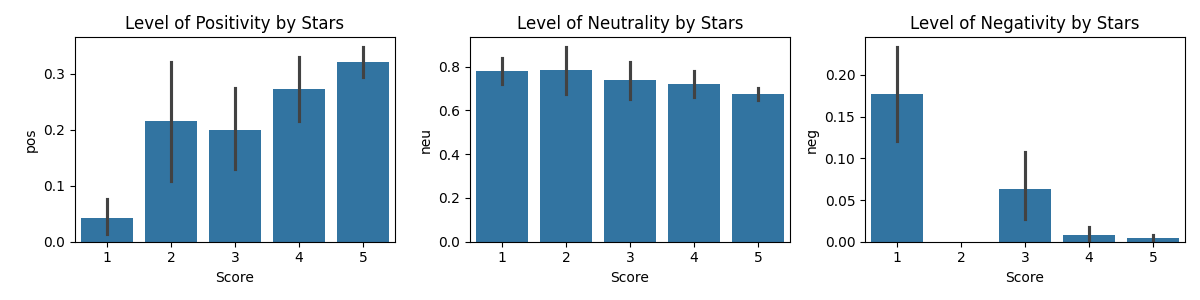
## The plots above represent the results of both approaches. Since the silhouette score for k=2 was higher, I will accept the result on the right picture.

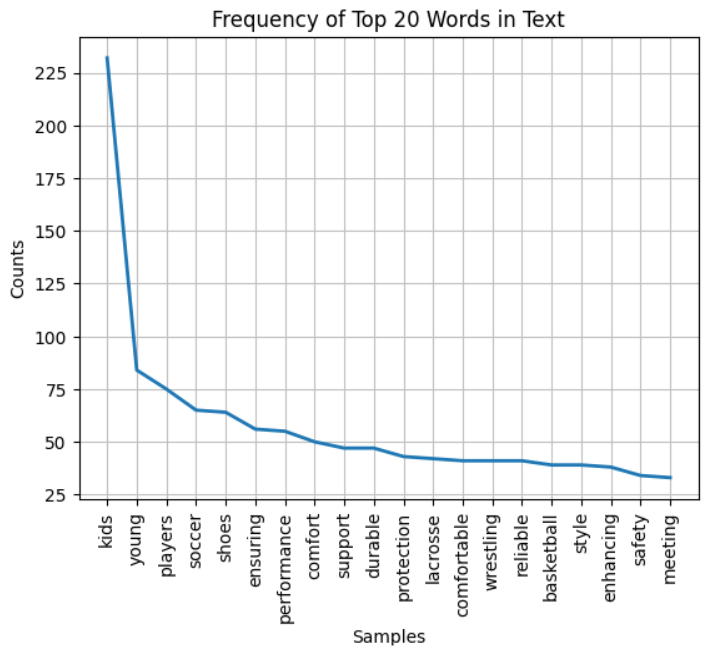
## Group 1 - customers are mostly people above 40 years old with salaries below 600$.

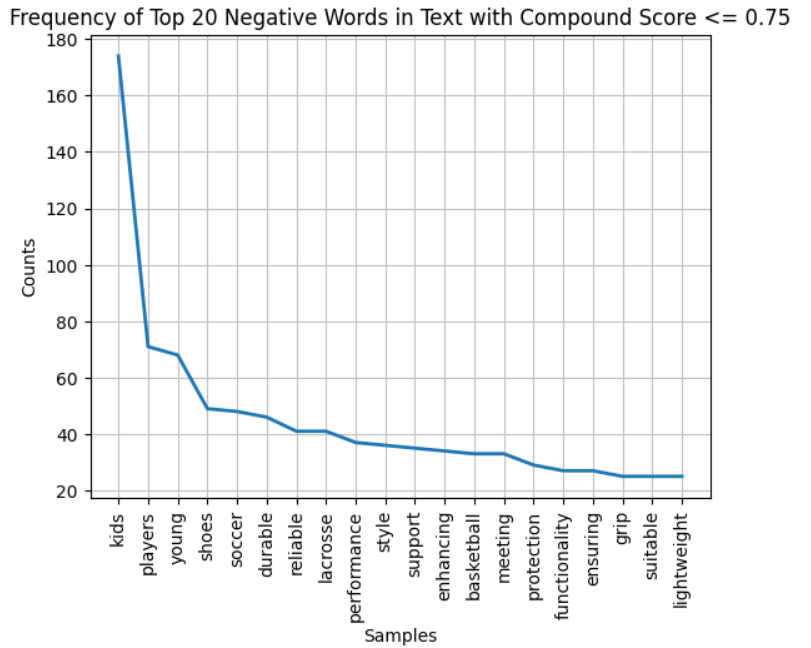
## Group 2 - customers are people 18-30 years old with a maximum income of 1200$.

## **Sentiment Analysis**

To be able to understand the patterns in reviews left by customers at risk and to identify their needs, I conduct sentiment analysis. The plot on the left shows that the context of the reviews was negative when people gave 1 star, somewhat neutral for 2,3,4 stars, and positive for 5 stars



The above dashboard shows that usually, ratings that had 2,4,5 stars had positive content.

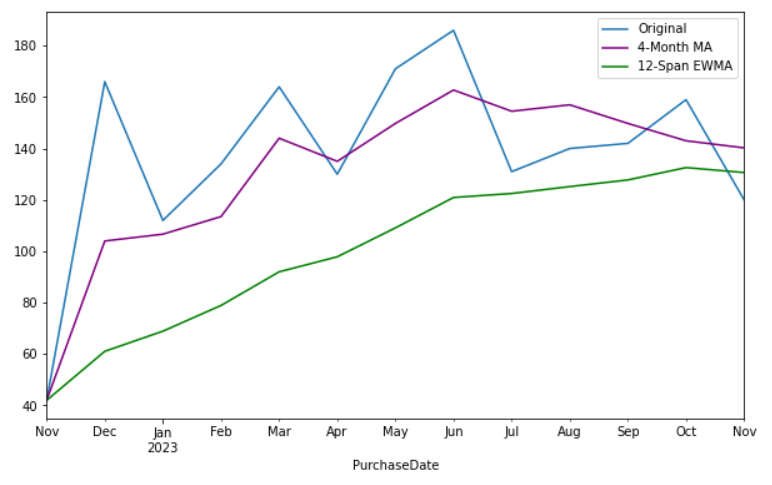
It also suggests that as the number of stars increases neutrality mainly decreases, meaning that people are more positive than neutral. For the last plot, I see that as expected whenever I had 1 star I usually had the highest negative context in the reviews. Still, the overall dashboard suggests that even though there are some low ratings, the negativity is not very obvious, this might be an indicator of different things, for example indicator that even unsatisfied customers mainly leave constructive feedback. To find out whether this assumption is reliable or not, I plot the words that have the highest occurrence (shown on the right). There are no negative words, but this can be because of having too many positive comments compared to negative ones, so I plotted the same graph for the reviews containing negative context(shown on the left) and got similar results:

no emphasized bad words. This means that the reason behind customers being at risk is not dissatisfaction.

Lastly, I investigated the second KPI mentioned above and found out that 77.61% of at risk customers were still satisfied. This is a good indicator of the fact that investment in engaging those customers won’t be a waste.

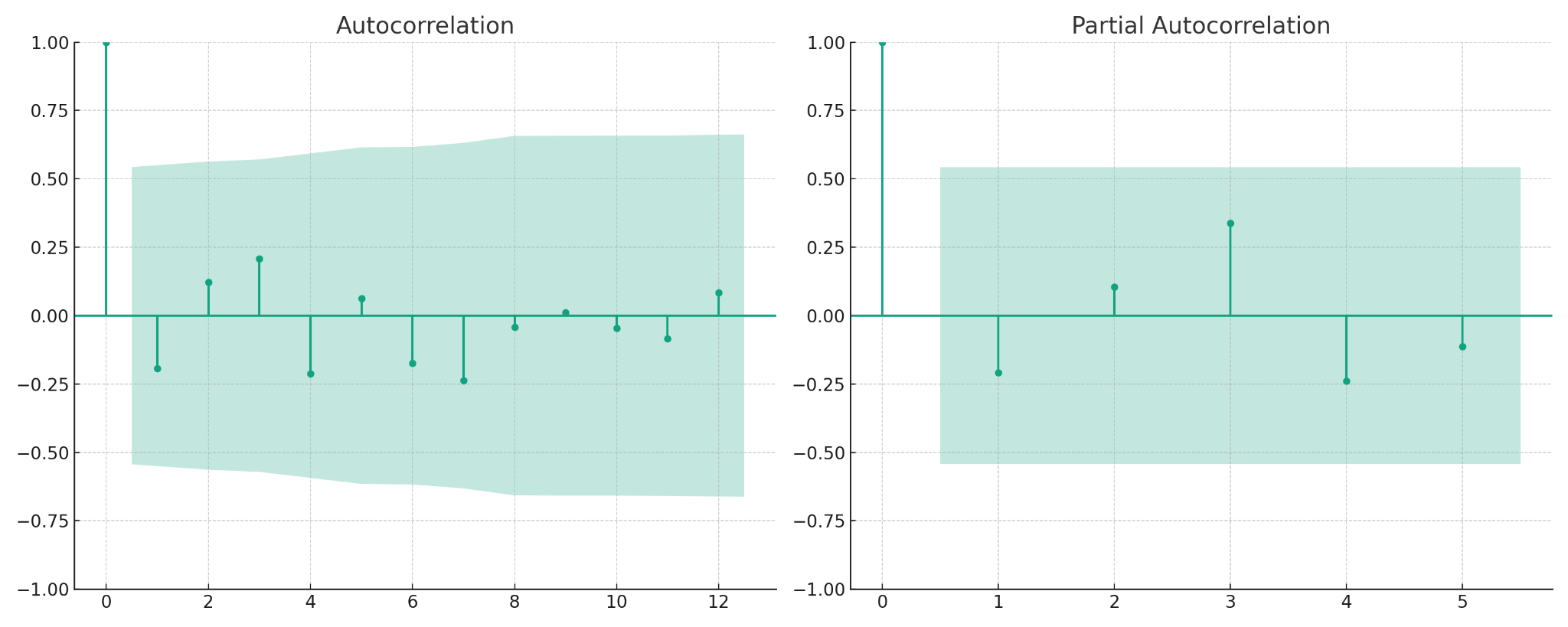
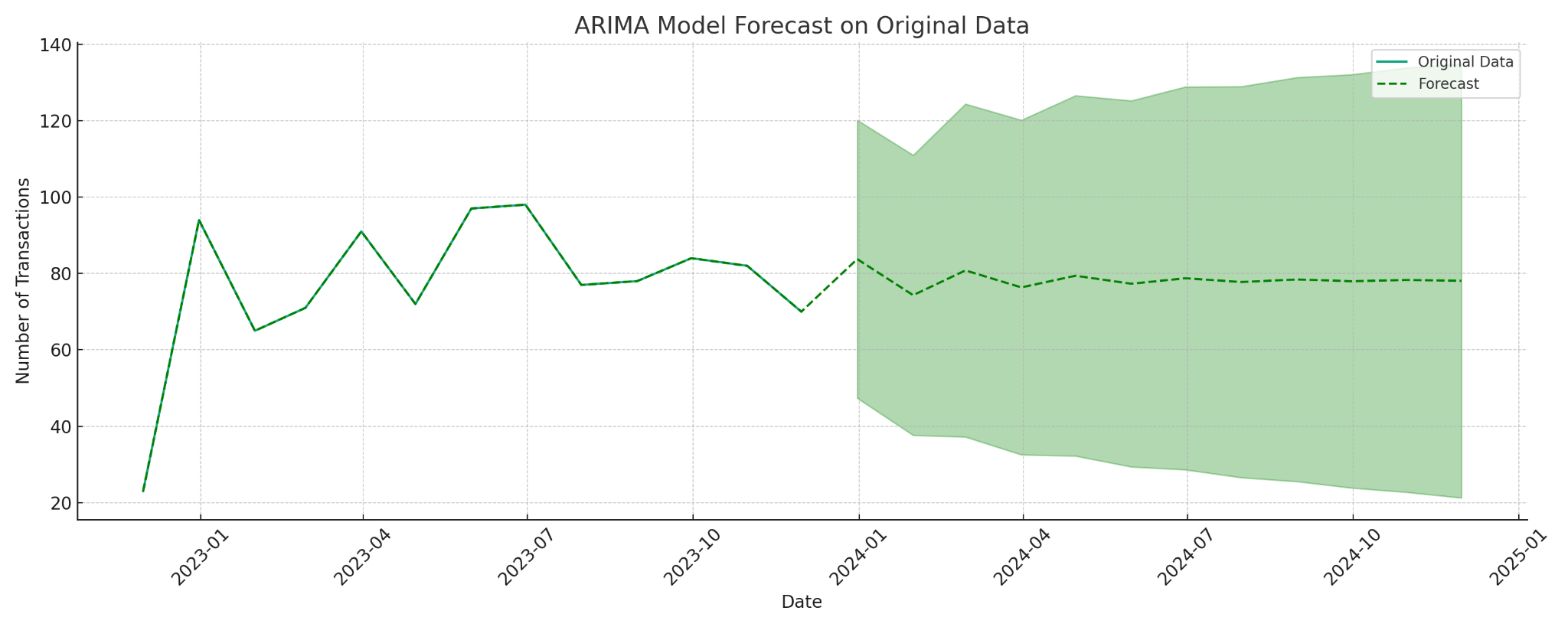
## **Time Series**

The time series plot, enhanced with Moving Average and Exponentially Weighted Moving Average (EWMA) lines, offers a detailed view of the transaction trends over time. The Moving Average provides a clearer understanding of the overall trend from the outset. The EWMA, with its greater emphasis on more recent observations, adapts more quickly to changes in the trend. This is particularly useful for identifying turning points in the data, such as sudden increases or decreases in monthly transactions.



While the Moving Average gives a general indication of the trend over a fixed period, the EWMA provides insights into how recent events may be influencing the trend. This plot is particularly valuable in understanding the underlying patterns and dynamics of the transactions, enabling better forecasting and strategic decision-making. The overlay of these averages on the original time series data provides insights into the transaction activity, revealing both short-term and long-term patterns that are crucial for informed business analysis.

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plot provide insights into the autocorrelation structure of the time series data. The ACF plot shows a non-stationary time series, which is typical for real-world data like monthly transactions. The significant spikes at initial lags indicate a strong autocorrelation, implying that past values substantially influence future discounts. However, The PACF plot shows a sharp cut-off after the first lag, which is indicative of an autoregressive component in the time series data. These plots are instrumental in ensuring that the selected model effectively captures the underlying patterns in the data, which is essential for accurate forecasting.

The ARIMA model's forecast, based on the transaction data from early 2023 to mid-2024, indicates that the volume of transactions the company handles is expected to hold relatively steady in the near term. This forecast line does not show significant upward or downward trends but rather fluctuates around a consistent level, similar the pattern observed in the historical data. The confidence interval, represented by the shaded area around the forecast, does not widen dramatically as time progresses, which would indicate increasing uncertainty. Instead, it maintains a relatively uniform breadth, reinforcing the idea of a stable forecast. If the model predicted a volatile future, this interval would typically expand significantly, reflecting greater uncertainty in the further-out predictions. Therefore, the plot visually communicates a message of near-term consistency in transaction activity, as interpreted by the ARIMA model shown above.

**Prescriptive**

* The analysis suggests that focusing efforts on retaining and engaging with at-risk customers can have a substantial impact on overall monetary performance.
* Strategies and campaigns could be tailored to the specific needs and preferences of champions to further enhance their satisfaction and loyalty.
* Identifying and addressing the factors that contribute to the at-risk group could be crucial for preventing potential churn and maximizing their lifetime value.
* Based on sentiment analysis, doing campaigns for this segment will not be a waste of time and money, since the targeted group was mostly giving positive feedback.



According to the plot of two different segments of customers, the number of female and male customers is not significantly different within each group. In addition, the customers in both groups are distributed in the regions evenly.



### The plot shows the distribution of the salaries for female and male risky and champion customers. I can see that Champion customers have higher median and mean salaries than Risky customers.

### 

## **Campaign Strategy**

I have designed several business strategies to attract at-risk customers.

Segment-Focused Approach: Utilize the findings obtained through the RFM analysis to craft campaigns specifically targeting the 'At Risk' segment, focusing on their distinct spending behaviors and preferences.

Targeted Loyalty Programs: Implement targeted loyalty programs offering incentives, early access to sales, and personalized rewards for repeat purchases, aiming to cultivate customer loyalty among the 'At Risk' segment.

Feedback Utilization: Use constructive feedback from reviews to tailor offerings and services, showing customers that their opinions are valued and implemented.