Business Analytics Group Project Report

DS 227: Business Analytics (Section A)

Fall 2023

Instructor: Nelli Muradyan

Members: Hakob Janesian, Vahan Yeranosyan, Hasmik Sahakyan,

Aida Martirosyan, Hovhannes Hovhannisyan

Contents

Project Description

This project aims to do a survival analysis on the data of an internet service provider company and calculate the churn rate for various years. Our approach will include utilizing the ***Kaplan-Meier*** method for survival curve estimation. We will also conduct ***Logrank tests*** to assess survival differences between genders and extend this analysis to different age groups to identify distinct patterns.

By analyzing survival times, we aim to identify factors influencing customer churn, enabling the company to develop targeted strategies for improving customer loyalty and business sustainability. This insight will help us adjust our services and marketing efforts more effectively, ultimately boosting profitability and customer satisfaction.

Data Description

The dataset we were working with contains transactional data from an internet service provider company, covering the period from January 2016 up to the present. In total, there are 143869 records. Each record is made up of six columns:

1) **account_id - Id of the customer account**
2) **payment_amount - Payment amount**
3) **payment_method - Payment method**
4) **payment_date - Payment method**
5) **gender - gender of the customer (generated with Python)**
6) **age - Age of the customer (generated with Python)**
7) **age_group - Age group of the customer (generated with Python)**
8) **village - Village of the customer**

Exploratory analysis with visualization

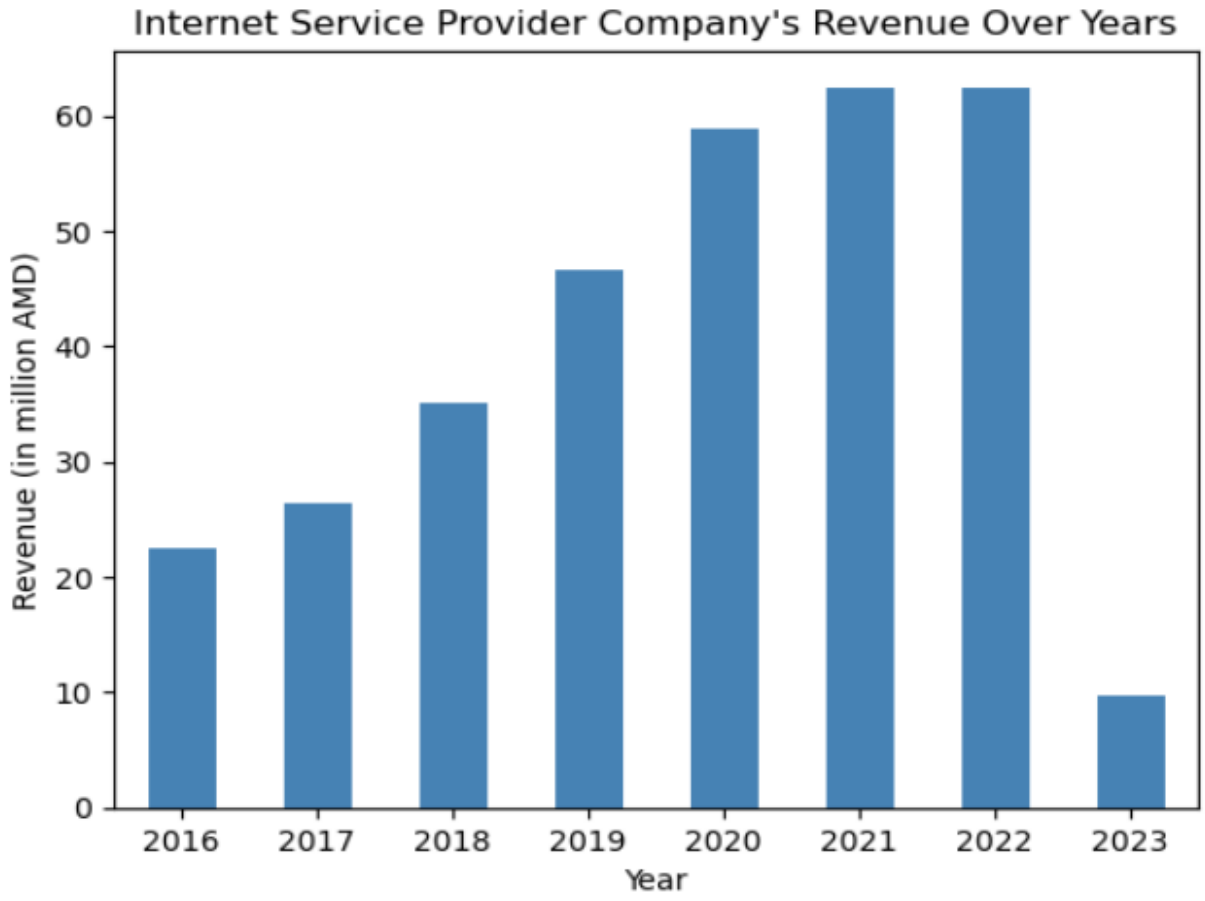

Figure 1 - "Barplot: Internet service provider company's revenue over years"

To get started, we have plotted the company's yearly revenue to identify anomaly years during the time period. Since the data was gathered in 2023 October, the last year's data is not available as of the time of the analysis. We can see dynamic patterns, when it goes up from 2016 up to 2021 and it goes down. We can identify growth and maturity stages from the plot and it is visually convincing that the company would likely experience stagnation or decline in the upcoming years under these circumstances.

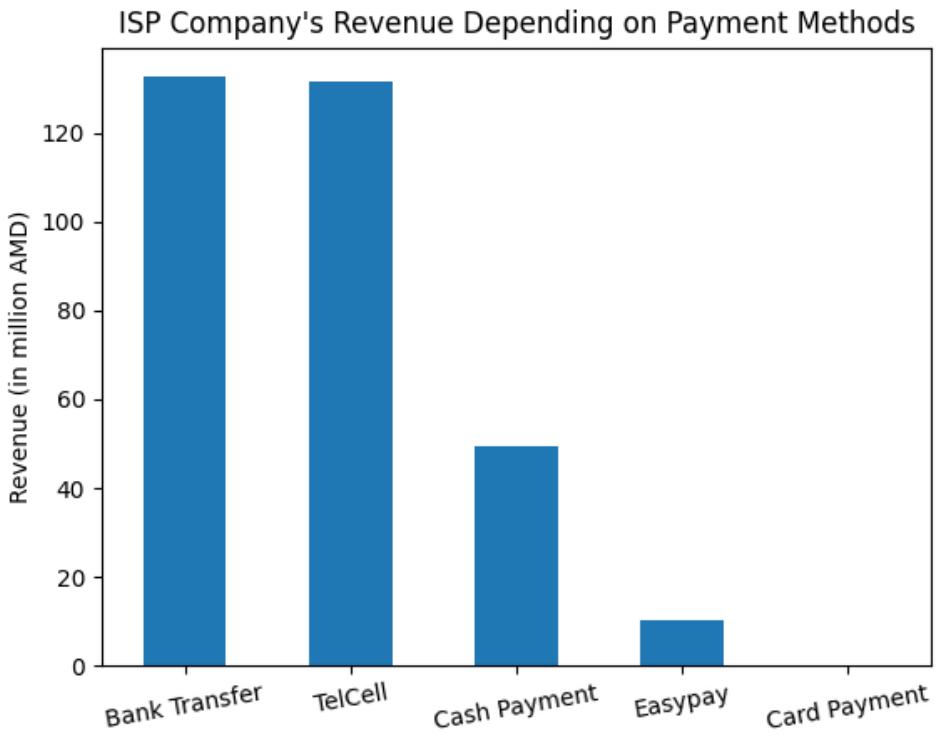

Figure 2 - "Barplot: Internet service provider company's revenue across payment methods"

The bar chart reveals that over the past eight years, the 'Bank Transfer' and 'TelCell' payment methods have been the top revenue generators, performing equally well. 'Cash payment' ranks third in generating revenue. Notably, 'Easypay' brings in considerably less revenue, and 'card payment' is the least effective, ranking fifth.

RFM analysis

Our group has initially done an RFM analysis. In the RFM analysis phase of our Survival Analysis project, we utilized customer transaction data to calculate Recency, Frequency, and Monetary values for each account. We first grouped the data by account, determining the first and last payment dates, total payment amount, and payment frequency. Using these metrics, we ranked and normalized each account's recency, frequency, and monetary scores. These normalized scores were then combined to form an overall RFM score. We calculated a weighted RFM score for each customer account, assigning weights of 25% to both Recency and Frequency

and 50% to Monetary value. The RFM score, rounded to the nearest integer, was then used to segment customers into four categories based on score ranges: scores below 25 were categorized as 'Leaving Customers,' those between 25 and 49 as 'Risky Customers,' scores from 50 to 74 were labeled 'Potential Loyalists', and scores 75 and above were classified as 'Champions.'

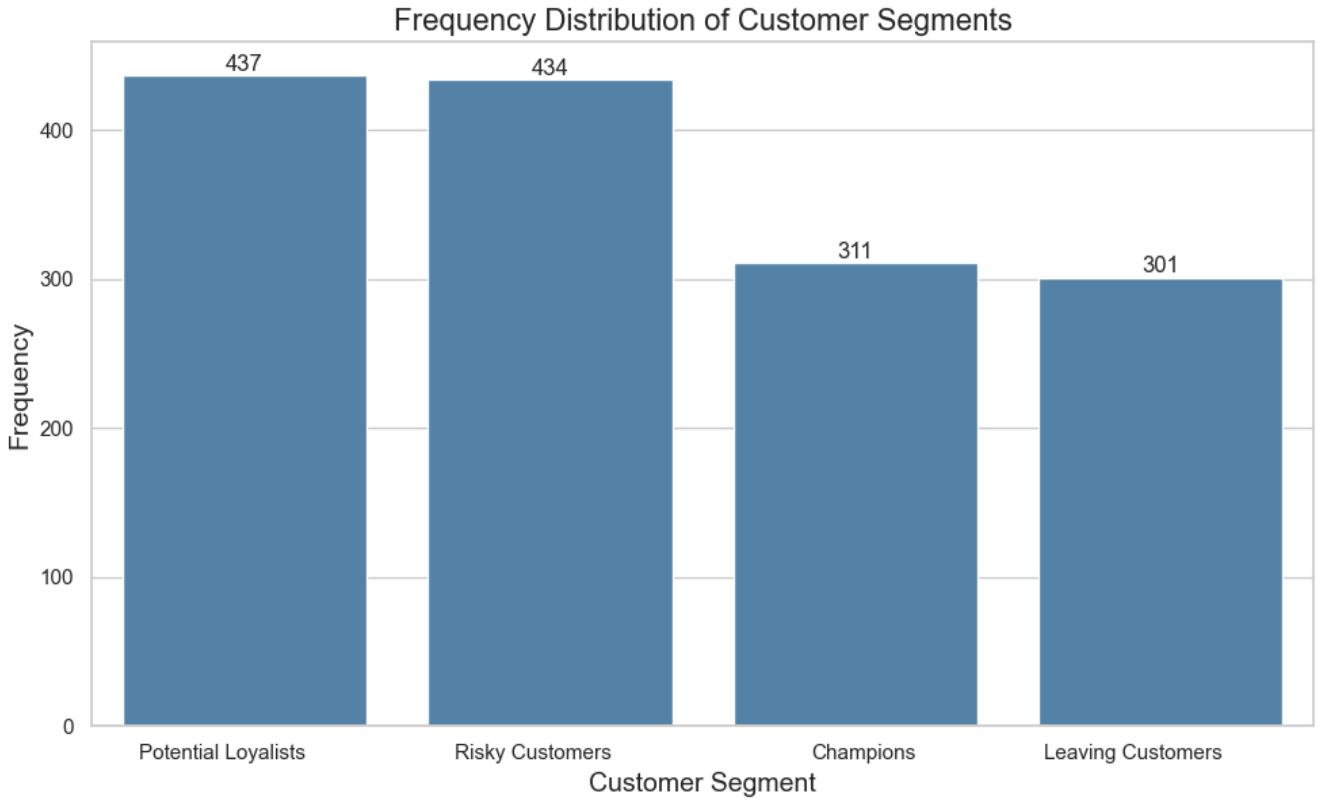

Figure 3 - "Barplot: Frequency Distribution of Customer Segments"

The data shows "Potential Loyalists" and "Risky Customers" have nearly identical frequencies, at 437 and 434 respectively. Similarly, "Champions" and "Leaving Customers" are close in count, with 311 and 301 frequencies. This suggests a balanced distribution between customer segments, indicating diverse customer engagement and retention challenges.

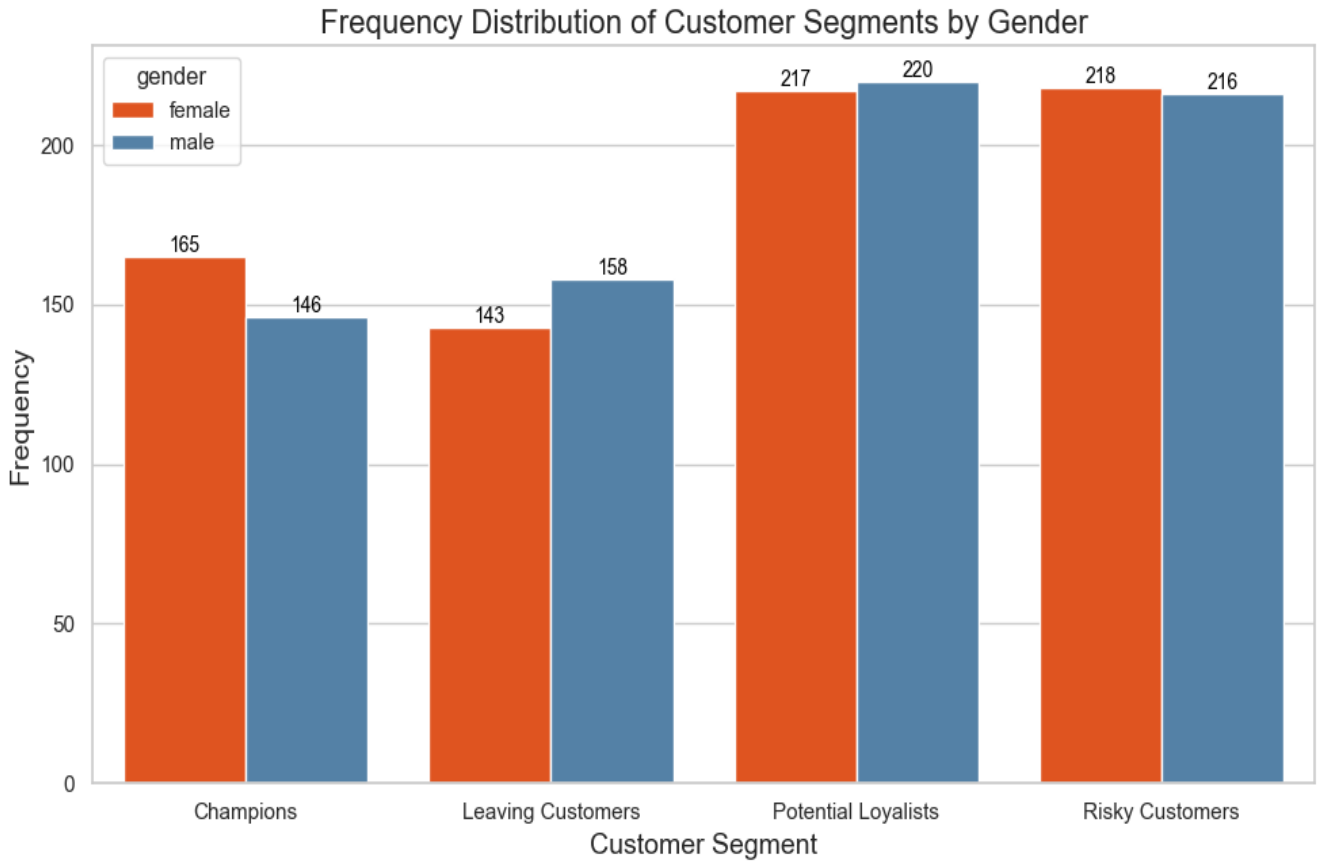

Figure 4 - "Barplot: Frequency Distribution of Customer Segments by Gender"

In the above plot, it is evident that the "Champions" segment comprises more females (165) compared to males (146). The "Leaving Customers" segment presents a balanced gender distribution, with females slightly outnumbering males (158 to 143). In the "Potential Loyalists" category, the gender frequency is almost equal, though males (220) marginally exceed females (217). Similarly, the "Risky Customers" segment shows a nearly identical count between genders, with a minimal male majority (218 males versus 216 females).

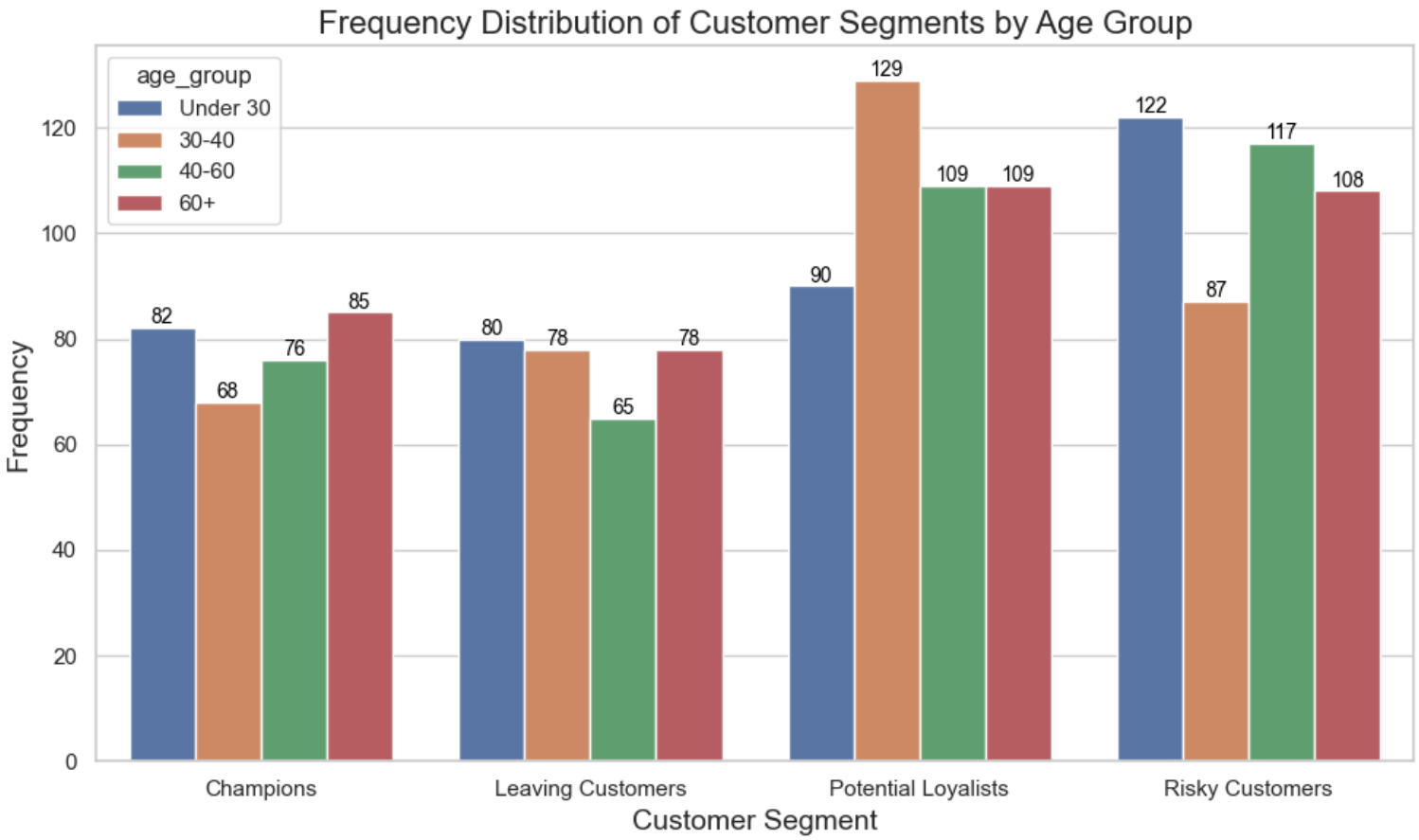

Figure 4 - "Barplot: Frequency Distribution of Customer Segments by age groups"

'Champions' are most frequent above 60 years, suggesting loyalty grows with age. 'Potential Loyalists' are largest in the 30-40 age group, indicating a strong middle-aged customer base. 'Risky Customers' are mainly under 30, hinting at uncertainty with younger clients. 'Leaving Customers' are spread evenly across age groups, with slightly fewer in the 40-60 range.

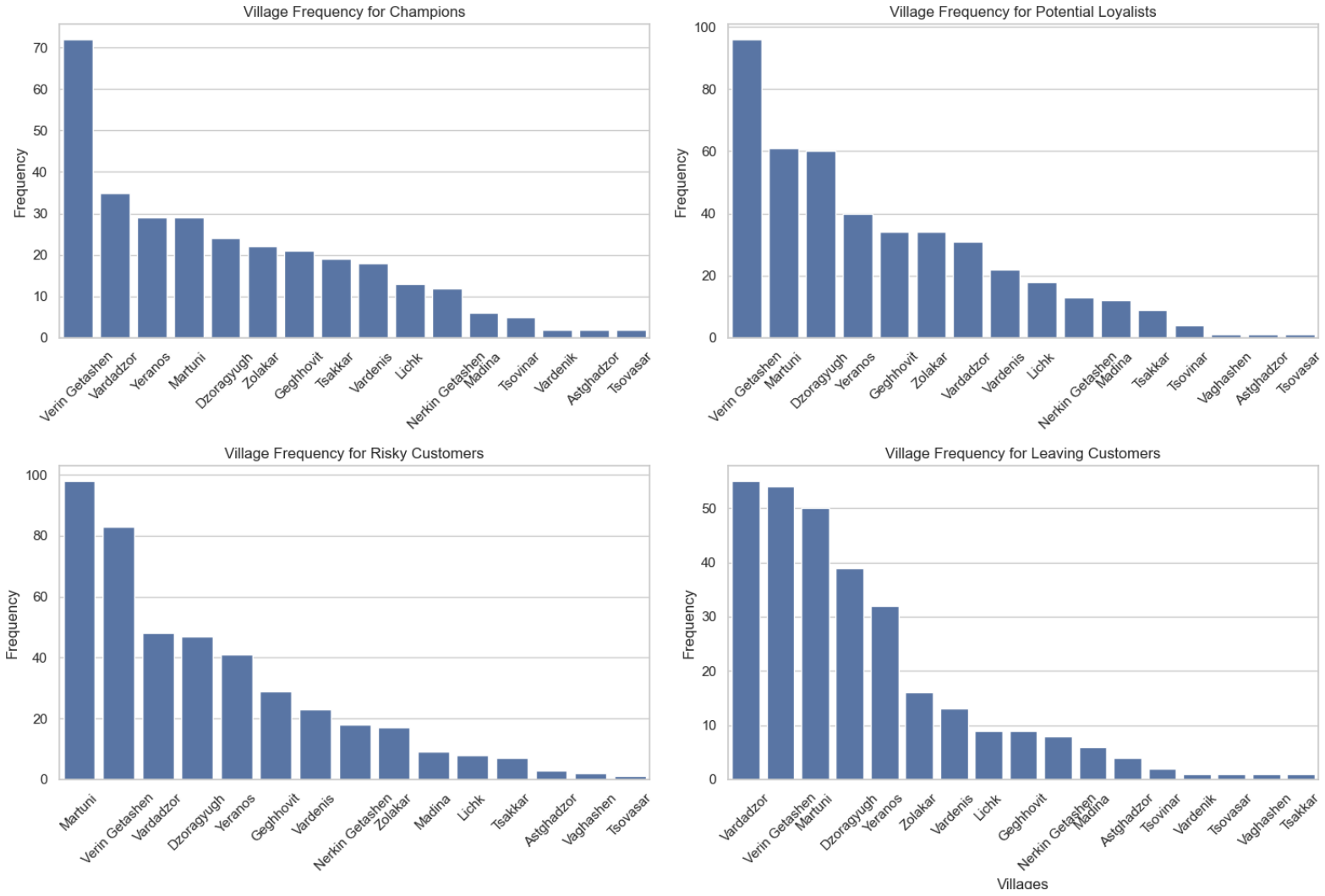

Figure 5 - "Barplots: Frequencies of Distributions of Customer Segments by Villages"

Martuni has the highest number of 'Risky Customers', while Verin Getashen leads with 'Potential Loyalists' and is also prominent across other segments. 'Leaving Customers' are most numerous in Vardadzor, and Verin Getashen again ranks high in this segment. Smaller villages like Tsovasar and Vaghashen have very few customers across all segments, indicating either smaller populations or less engagement. Among the villages, Verin Getashen holds a substantial number of 'Champions', though fewer than its 'Potential Loyalists' and 'Risky Customers', while Martuni, despite its high-risk customer count, has a comparatively moderate presence of 'Champions'.

Survival analysis (Kaplan-Meier and Logrank tests)

During the Kaplan-Meier analysis, we focused on customer retention by analyzing their purchase history between 2020 and 2022. We prepared and grouped the data by individual account IDs, taking into account various factors like purchase dates, amounts, and key demographic details. This grouping helped us understand payment behaviors and demographic patterns. We then calculated survival times for each account, measured as the duration between their first and last purchase. Using the Kaplan-Meier Fitter, we plotted the survival probability over time, providing a clear view of customer retention trends.

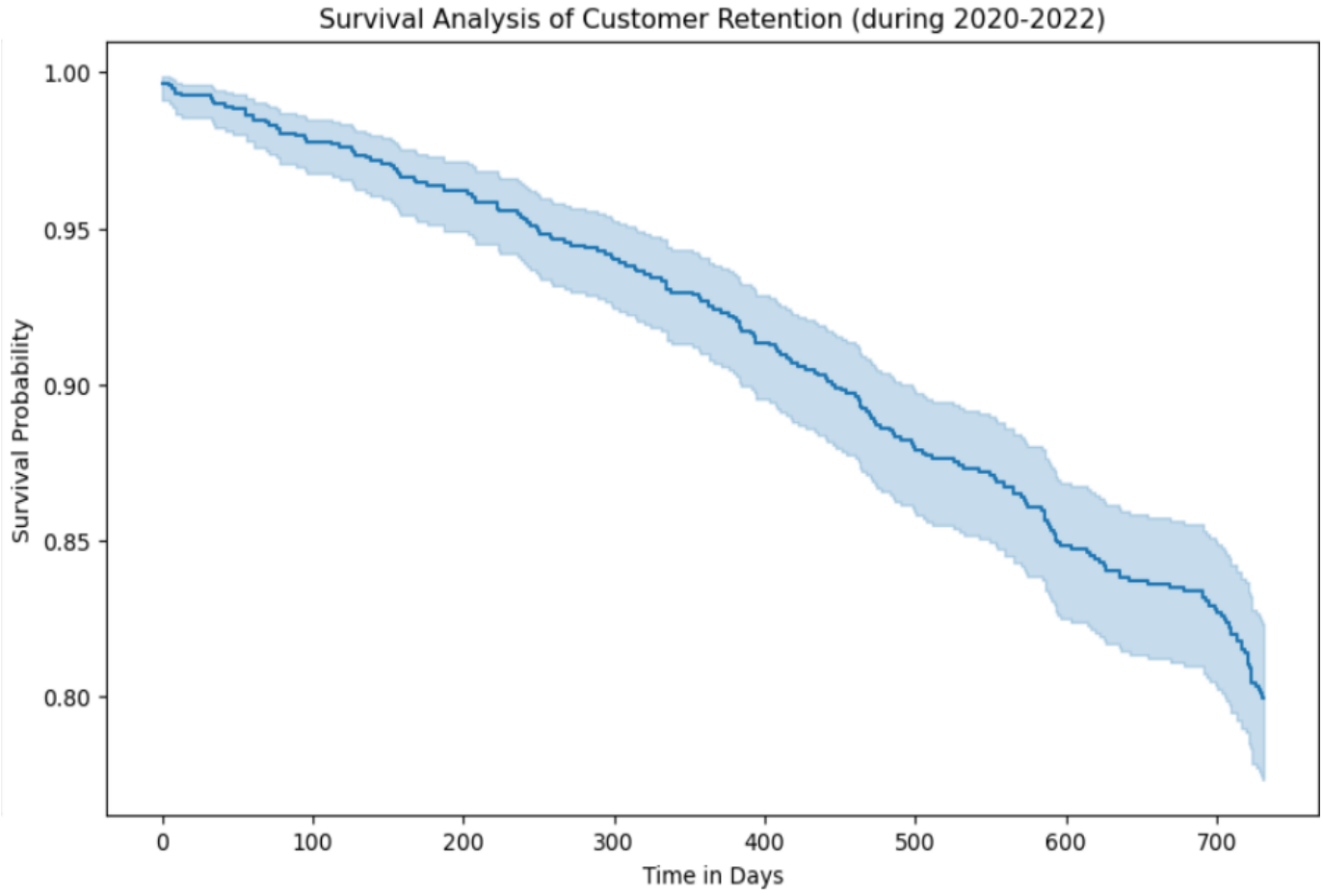

Figure 6 - "Survival curve: Survival analysis of Customer Retention (during 2020 - 2022)"

This plot above revealed that around 20% of customers churned during the period 2020 to 2022. The downward trend in the survival chart suggests that customer involvement decreased

over time. These findings are crucial for developing future strategies to improve customer loyalty.

In the Log-Rank test, we compared survival curves across different demographics, focusing initially on gender. By plotting Kaplan-Meier survival functions for male and female customers, we visually examined differences in their retention over time.

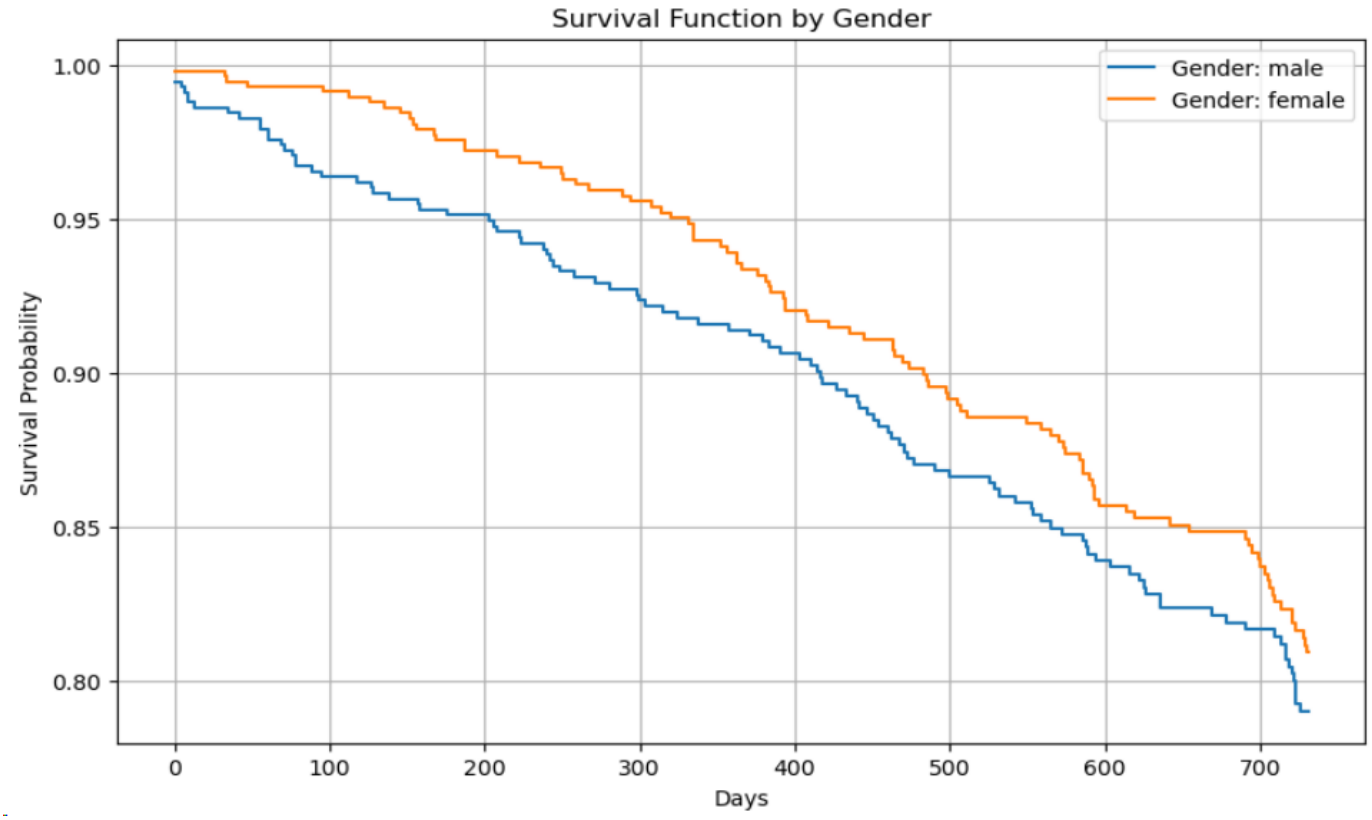

Figure 7 - "Survival curves: Survival curves by gender (during 2020 - 2022)"

Visually we observe that there may be some difference between the survival curves of male and female customers. Noticeably the orange curve (females) is in a higher position during the whole period from 2020 to 2022, which means that female customers tend to have a higher retention rate over time compared to males. However, towards the end of the period (around t = 700), this gap diminishes. This observation aligns with the Log-rank test's p-value for gender, which is 0.3018. Since this is above the 0.05 threshold, it indicates no statistically significant difference in retention between genders over the entire period analyzed. It is important to keep in mind that the data we analyzed was generated. Because of this, the results showing no significant

difference in gender retention rates are not surprising. Generated data like this typically lacks variations usually present in real-world data.

We next applied the Log-Rank test to examine differences in survival curves across various age groups.

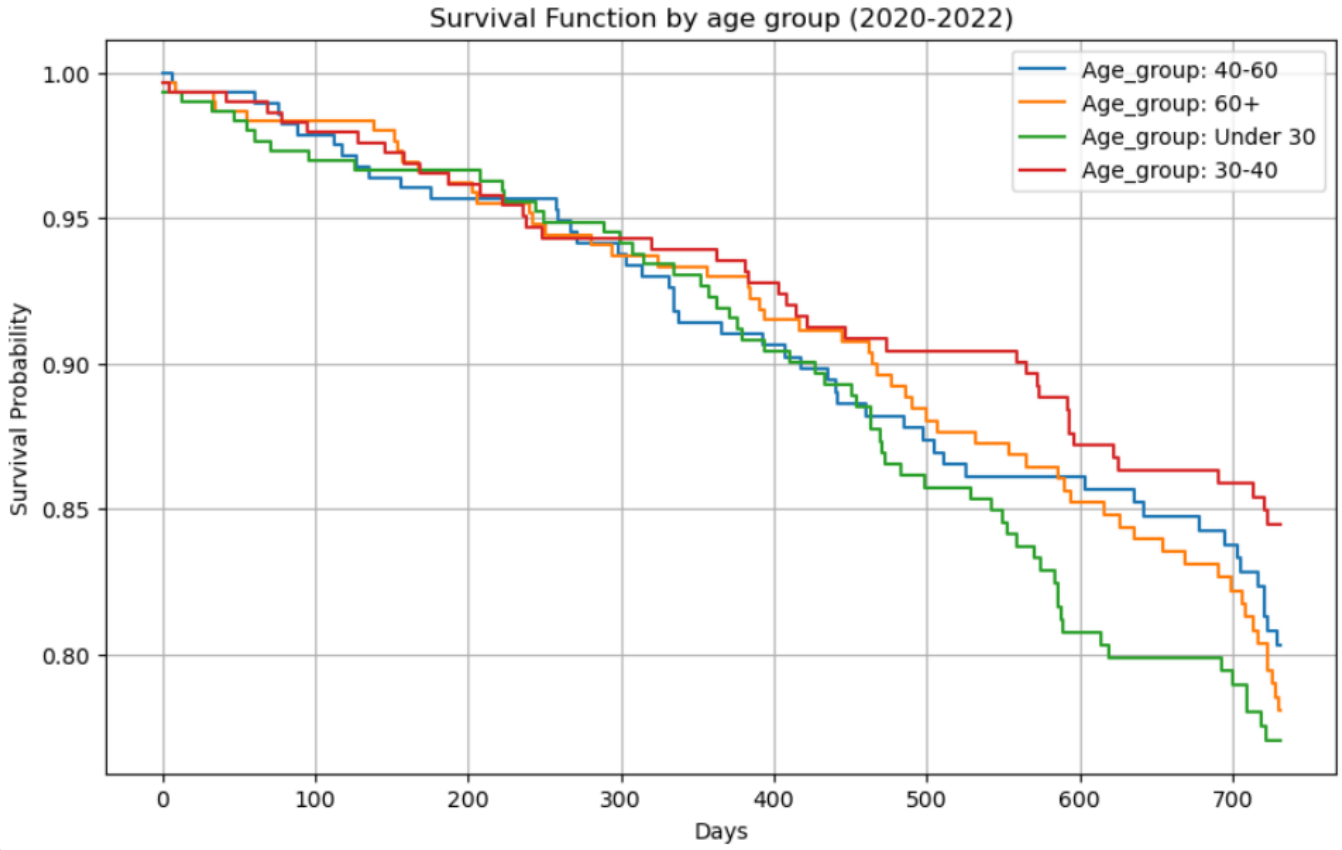

Figure 8 - "Survival curve: Survival curves by four age groups (during 2020 - 2022)"

Upon examining the plot above, both visual inspection and Log-Rank test p-values indicate no significant difference in customer retention across the age categories. The survival curves overlap substantially throughout the observed period, and the p-values from the Log-Rank test confirm this lack of statistical significance, reinforcing the conclusion that the age group does not play a noticeable role in customer retention.

Moving on to regional analysis, we first reviewed a histogram to determine the most frequent villages from 2020 to 2022.

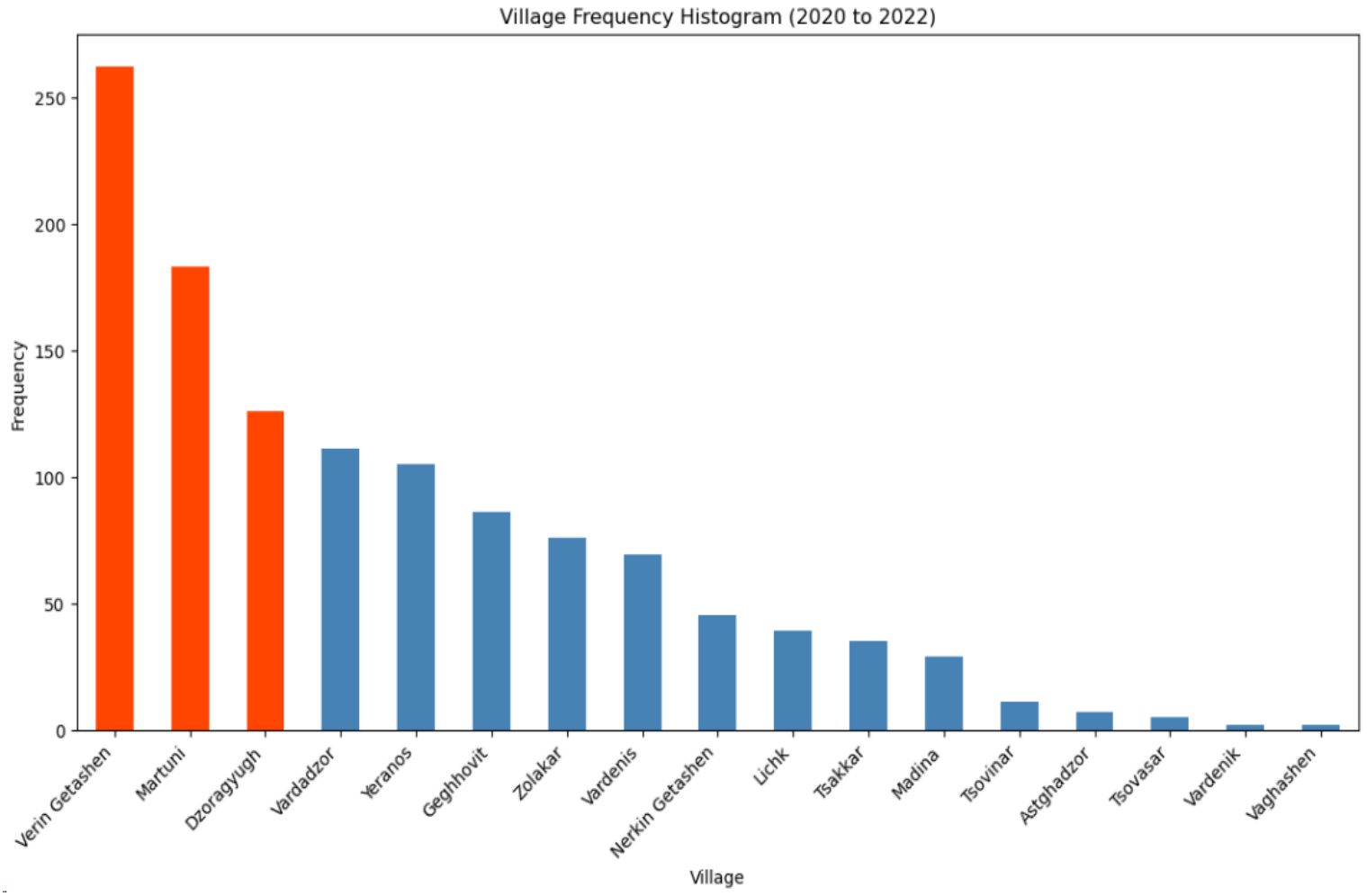

Figure 9 - "Barplot: Identifying top three active villages during 2020 - 2022"

Hence, we identified Verin Getashen, Martuni, and Dzoragyugh as the top three. Focusing on these, we will perform a Log-Rank test to compare customer retention rates across these villages. Our aim is to understand how regional differences impact customer behavior.

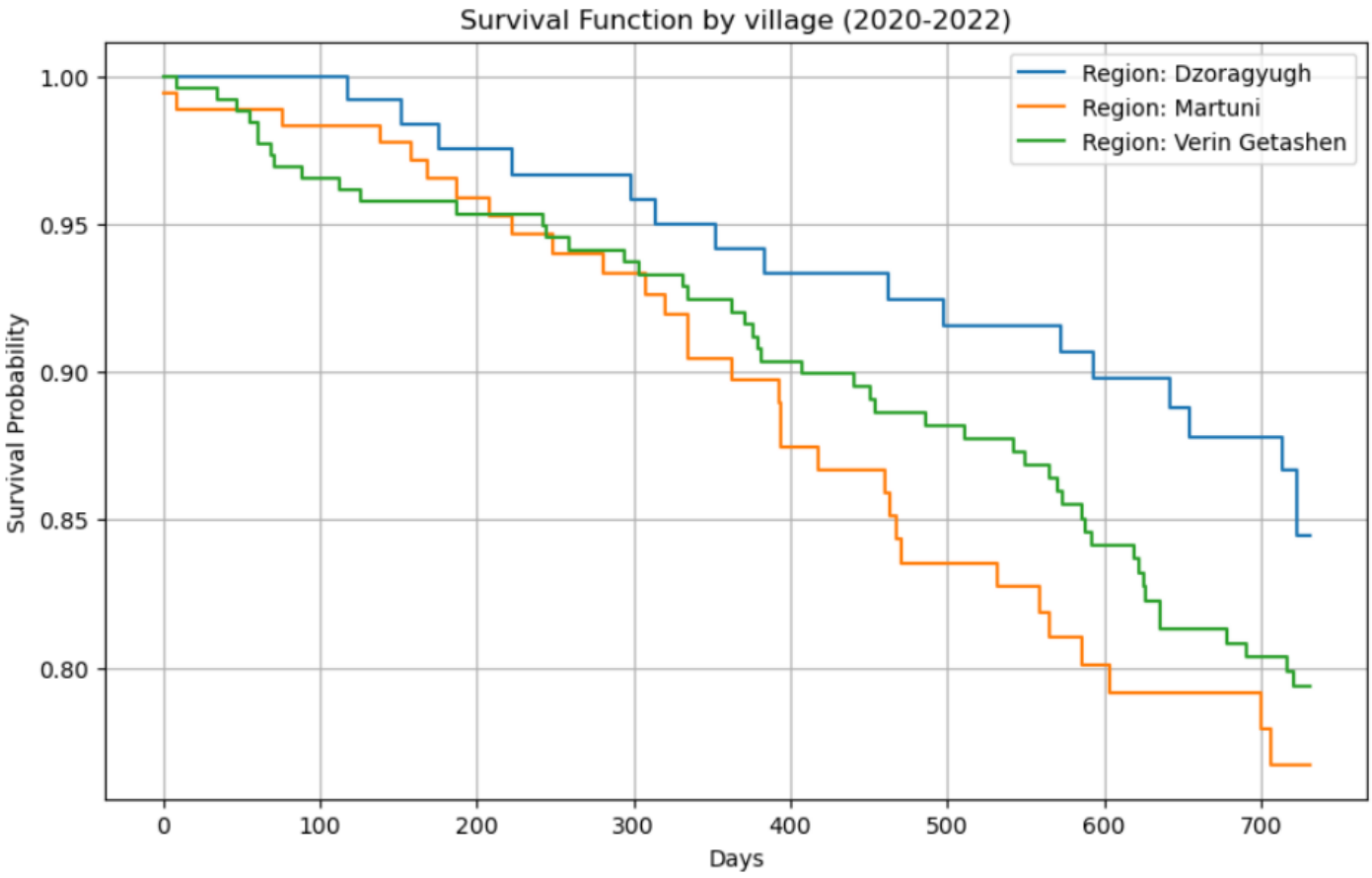

Figure 10 - "Survival curve: Survival curves by top three villages (during 2020 - 2022)"

From the plot above we can visually claim that the blue curve (Dzoragyugh) is in a higher position during the whole period from 2020 to 2022, which means that customers from Dzoragyugh tend to have a higher retention rate over time compared to customers from Martuni and Verin Getashen. After the 300th day, the Verin Getashen's curve is in a consistently higher position than Martuni's curve, thus Martuni is in an inferior position compared with the two regions. In the end, Dzoragyugh's curve stops on survival probability nearly equal to 0.84, but Verin Getashen's and Martuni's curves nearly stop at 0.79 and 0.76 respectively. However, the Log-Rank test p-values, being 0.074 for Dzoragyugh versus Martuni, 0.198 for Dzoragyugh versus Verin Getashen, and 0.501 for Martuni versus Verin Getashen, indicate that these differences in retention rates are not statistically significant, as all values are above the 0.05

threshold. This implies that while there are visual differences in the survival curves, they do not indicate significant statistical disparities in retention across these regions.

Next, we extended our analysis to include Log-Rank testing on customer segments defined by RFM analysis, which are - 'Leaving Customers,' 'Risky Customers,' 'Potential Loyalists,' and 'Champions.' We plotted Kaplan-Meier survival functions for each RFM-based category to visually assess differences in customer retention.

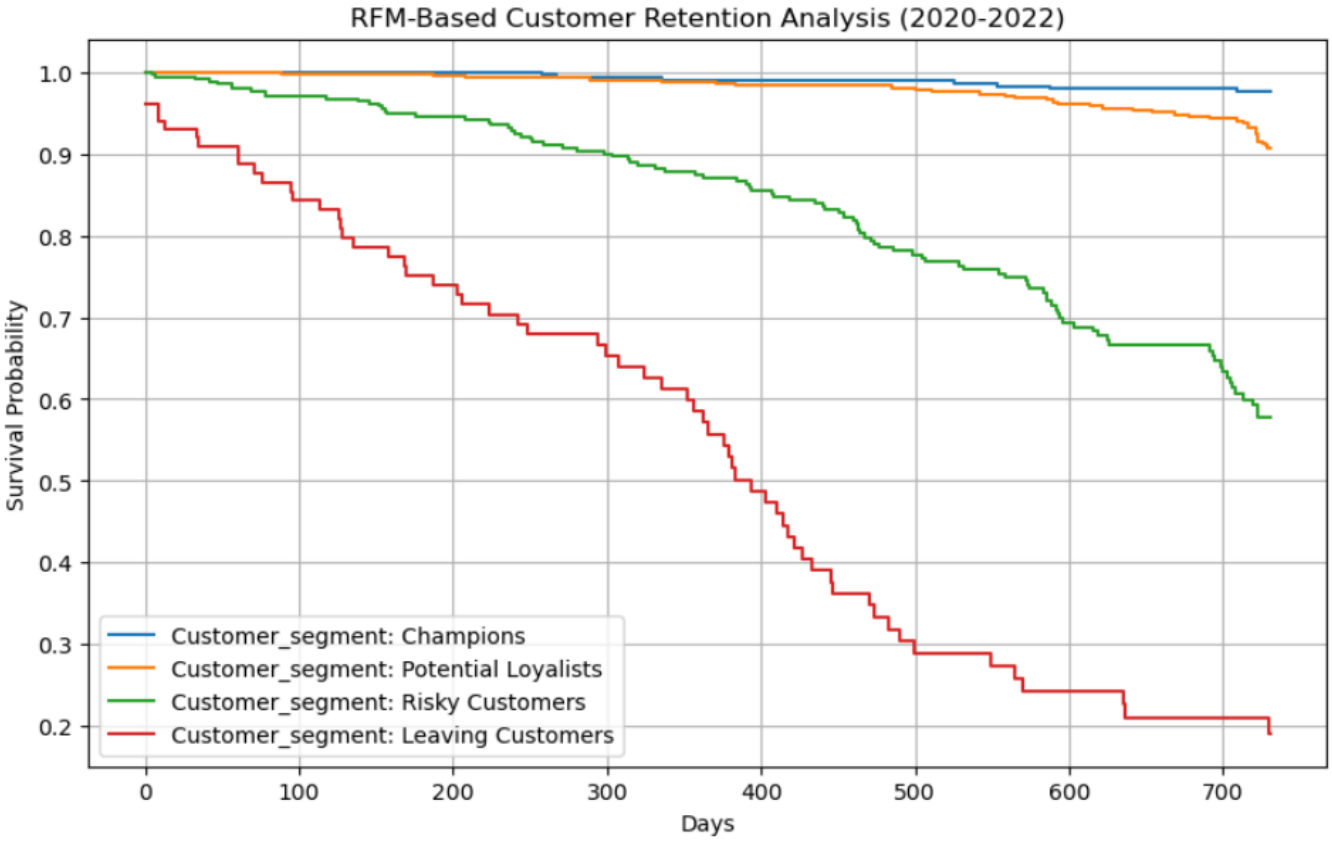

Figure 11 - " Survival curves by customer segments by RFM (during 2020 - 2022)"

Our analysis revealed distinct variations in churn rates among different customer segments from 2020 to 2022. The 'Champions' segment displayed the lowest churn rate, with only 3% leaving, followed by 'Potential Loyalists' at 10%, 'Risky Customers' at 42%, and 'Leaving Customers' with a high churn rate of nearly 80%. This pattern aligns with the expected behavior of these segments. The Log-Rank test results further substantiate these observations. The test showed highly significant differences between each pair of customer segments. The p-value for 'Champions' vs. 'Potential Loyalists' was 0.0007, indicating a notable difference in

retention between these two segments. More strikingly, the comparison between ‘Champions’ and ‘Leaving Customers’ resulted in a p-value of approximately 1.63e-76, confirming a substantial disparity in retention rates. Similarly, ‘Potential Loyalists’ compared to ‘Leaving Customers’ yielded a p-value of around 4.50e-84. Even between ‘Risky Customers’ and ‘Leaving Customers,’ the difference was significant, with a p-value of 1.20e-18. These p-values, being well below the 0.05 threshold, clearly indicate significant differences in customer retention across these RFM segments.

Churn rate

Next our group decided to calculate the churn rate based on the dataset. Churn rate is a crucial business metric that measures the percentage of customers who stop using a product or service within a specified time period. The churn rate is calculated using the following formula:

$$\text{Churn Rate} = \frac{\text{Number of Clients at the Beginning} - \text{Number of Clients at the End}}{\text{Number of Clients at the Beginning}}$$

We identified users present in January and December of each year and calculated the yearly churn rate. Users who were active in January but were not present in December are considered churned for that year. The churn rates were then visualized to compare yearly changes. Let us go through the below visualizations and see what findings we have.

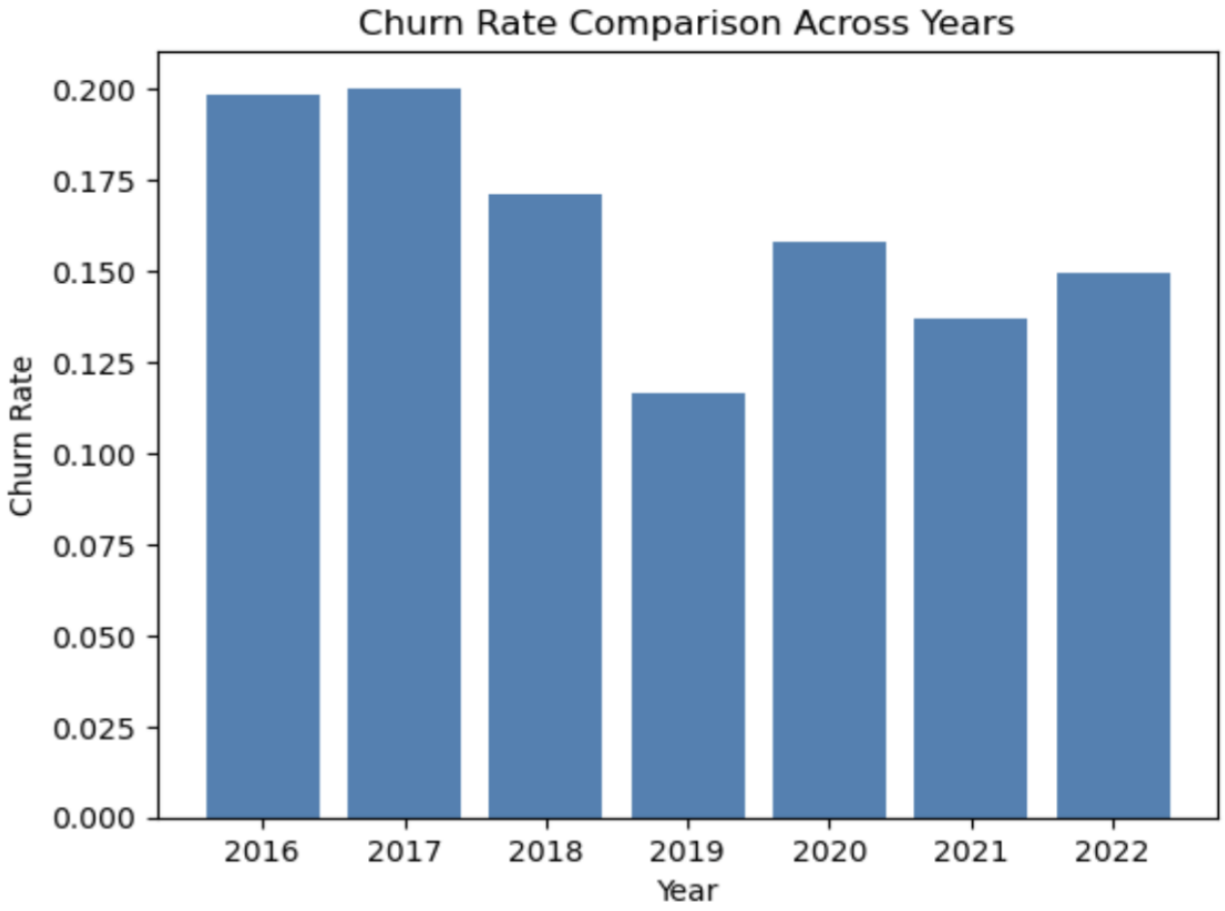

Figure 12 - " Barplot: Churn rates from 2016 to 2022"

The above barplot illustrates the yearly churn rates from 2016 to 2022. During the initial years, 2016 and 2017, the business experienced relatively higher churn rates, with around 19.85% and 20.03%. However, in the upcoming years, the churn rates decline. The lowest churn rate of 11.66% was observed in 2019, indicating a notable improvement in customer retention.

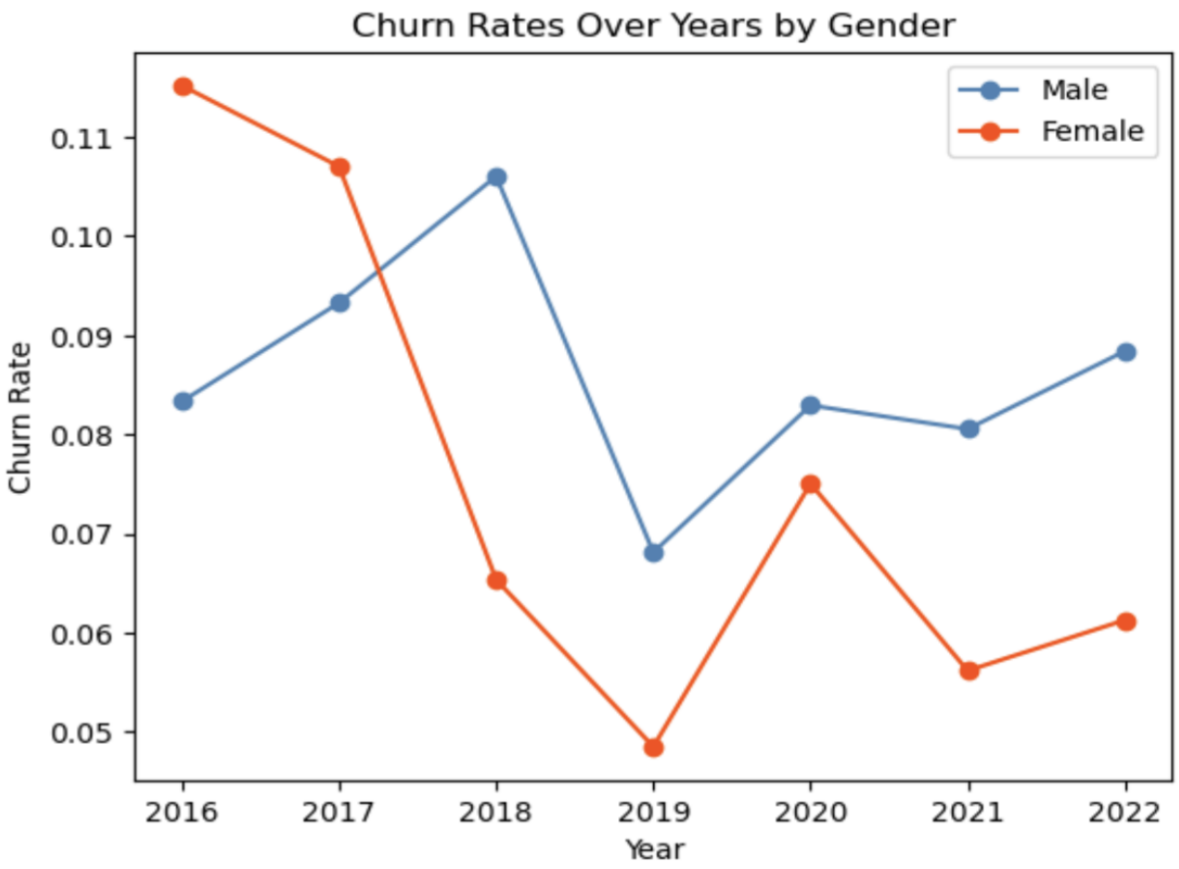

Figure 12 - "Lineplot: Churn rates over years by gender"

The chart shown above indicates a higher churn rate among female users compared to male users in 2016 and 2017. This suggests that during these years, a larger proportion of our service's users who churned were females. However, the trend shifted after 2017. From that point onwards, the churn rate for males surpassed that for females, indicating that a greater number of males were ceasing to use our service in comparison to females.

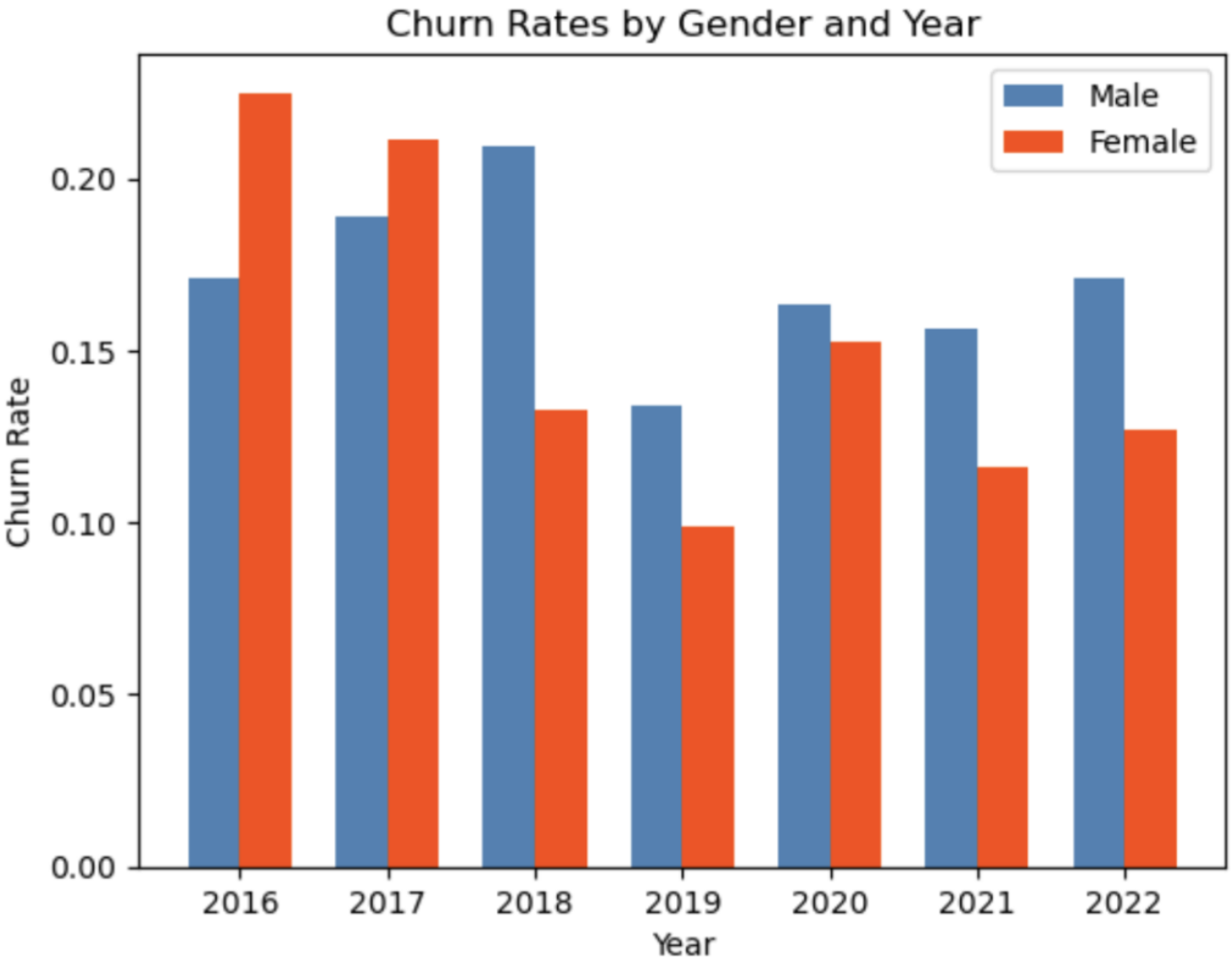

Figure 12 - "Barplot: Churn rates by gender over years "

After calculating general churn rates we decided to also calculate churn rates for male and female users separately over multiple years. The above plot represents the percentage of male and female users who stopped using the service, let us analyze data for the year 2022. The male churn rate is 17.12%, while the female churn rate is 12.67%.

Finally, we also calculated the overall churn rate between 2016 and 2023. The churn rate from January 2016 to January 2023 is 59.31%. This percentage represents the proportion of customers who stopped using the service during the whole time period of our analysis.

Conclusion

The RFM analysis revealed a balanced distribution among customer segments, with "Potential Loyalists" and "Risky Customers" showing nearly identical frequencies, suggesting a need for targeted strategies to convert risky customers to loyal ones. Kaplan-Meier and Logrank tests indicated no significant differences in retention across gender or age groups, but highlighted substantial variation in churn rates among RFM segments, particularly between "Champions" and "Leaving Customers." Churn rate analysis revealed a declining trend over the years, yet a higher overall churn rate for males in recent years. To improve business performance and reduce churn, the company should focus on nurturing 'Potential Loyalists' and devising personalized retention strategies for high-risk segments. Additionally, gender-specific marketing tactics could be beneficial in addressing the differing churn trends among male and female customers.