# Descriptive Data Analysis Notes

Before starting with descriptive data analysis, we contacted the client to solve the previous doubt about the field Senior. They told us that the Senior category should include 65 year old subscribers. Therefore, we applied the following conditional function to that column:

=SI([@Age]>65;"Yes";"No")

Once we had completed data cleaning and transformation, we copied the excel file into a new folder called Data Analysis and then we renamed that file to “amazon\_churn\_descriptive\_analysis.xlsx”. Descriptive data analysis was performed on that file.

For this purpose, we used the Data Analysis tool that is already integrated in Excel, which enabled us to obtain a quick summary of the most relevant descriptive statistics: mean, median, mode, standard deviation, variance, minimum, maximum, range, skewness coefficient and kurtosis coefficient.

## Univariant analysis: Insights on numeric columns

Insights are highlighted in yellow. Further analysis in blue.

* **Account Length (in months):** median (28 months) and mean (32.31 months) are similar, which suggests an equilibrated distribution of the accounts antiquities. On the other hand, a mode of 1 month indicates the majority of subscribers are new clients. However data dispersion is notorious: we have very new clients but also very old clients (with an account length up to 77 months (almost 6.5 years). This is confirmed by the big standard deviation and variance. Account Length data reveals the existence of two groups of customers: a significant group of new customers characterised by high rotation rates and an extended group of loyal customers with long-term retention rates. This variety of profiles suggest the need to apply different marketing strategies for each segment.
* **Customer Service Calls:** the average number of calls is less than 1, meaning the majority of subscribers did never called. This is confirmed by the mode (0) and the median (0) which suggest that more than half of the subscribers never called. On the other hand, data is really dispersed: there is a minority of customers that have called a lot of times, versus a majority of clients who are not calling at all. This is supported by the range from 0 to 5 and positive skewness. Clients who did multiple calls may require special attention and the reason for churn may be associated with these issues.
* **Average Monthly GB Download:** the average number of downloaded GB per month is approximately 6.7 GB. The 50% of the subscribers download 5GB per month or less. Moreover, the majority of the subscribers never downloaded films or series from the service, as the mode is equal to 0. Probably the streaming option is the favourite. On the other hand, data is really dispersed: there are customers that have downloaded a lot of GB per month versus customers with no downloads at all, which can be deduced from the range, standard deviation, variance and positive skewness. Data indicates the majority of subscribers prefer the streaming option while there is a significant group with a preference for downloading content. Segmented groups may need diversified marketing strategies and service plans.
* **Extra Data Charges:** the average extra data charges is almost 3.4 USD. Moreover, the majority of the subscribers have not been charged extra as the mode is equal to 0 USD. Same for the median. Most probably, this variable is positively correlated with the above (average monthly GB download): the more the GB downloaded, the more the extra data charge. Check this. On the other hand, data is really dispersed and grouped: there is a minority of customers that had been charged great quantities due to their particular need for big downloads (they may need special attention) versus a majority of customers who strictly control data usage and have no charges at all. This is confirmed by the range value (99) and extremely positive kurtosis.
* **Age:** the average age of the subscribers is equal to 47.53 years old and is mostly identical to the median (47), suggesting an equilibrated distribution, close to normality and, in general, mid-age subscribers. In addition, there is a majority of subscribers who are 29 years old (mode). However, data dispersion is notorious: the standard deviation is approximately 17 years old and the range goes from 19 and up to 85 years old. It is clear that Amazon Prime Video offers a variety of contents suitable for all ages and that the company plays in a very segmented market, needing to apply very diverse marketing strategies for assuring client retention.
* **Number of customers in Group**: the average number of customers in Group is less than 1, meaning the majority of customers are enrolled in an Individual Plan, which is confirmed by the mode and the median values (0). On the other hand, data dispersion is a fact: a significant group of customers who are enrolled in a Family Plan coexist with a majority of customers in Individual plans. This is confirmed by the standard deviation, variance, range (6), positive kurtosis and positive skewness. Both types of plans are suitable for Amazon Prime Video’s customers.
* **Monthly Charge (USD):** the average monthly charge is equal to 30.90 USD/month. The median is very similar: 31 USD/month, indicating an equilibrated distribution. On the other hand, the majority of subscribers pay a charge of 10 USD/month, which reflects that, in general, clients prefer more economic plans. However, the standard deviation and the very wide range (78) suggest that data dispersion is notorious, indicating a huge variety of plans, which can be strategic to reach more market segments.
* **Total charges (USD):** the average total charges is equal to 1084.78 USD. The median is much lower: 647 USD, meaning that the 50% of subscribers paid a total of 647 USD or less, which is relatively low. The majority of subscribers pay a charge of 10 USD/month. On the other hand, data dispersion is extreme: standard deviation (1129 USD) and range (5568 USD) have very significant values, meaning that total charges are really diversified, which may indicate a need to segment marketing strategies and customer services to cover necessities of all this groups.
* **Average Monthly Expenses (USD):** the average of the average monthly expenses per subscriber is equal to 30.98 USD. The median is very similar: 30.66 USD, meaning it is an equilibrated distribution. However, data dispersion is notorious: the standard deviation is approximately 16.84 USD. This idea is reinforced when we observe the range of monthly charges (80.67 USD). All in all, the variety of expenditures, with values from 4 USD to 80 USD distributed in an equilibrated manner, show a clearly segmented market where different marketing strategies need to be applied for customer retention.
* **Number of Complaints or Support:** A median (5) almost identical to the mean (5.01) and skewness coefficient close to 0 indicates that clients interaction with the customer service is more or less equilibrated, without significant segmentations. There is a majority of customers that have complained or requested support 6 times, which is a significantly high value. Moreover, data is dispersed, revealing customers who have registered from 0 up to 10 complaints or support. In general, data reveals there is a majority of customers with problems with the service. In addition, does ones who complain or request support often do it for multiple times, potentially indicating unresolved issues after the first interaction. This is something that requires special attention. A correlation analysis between churn and the number of complaints or support requests could help determine whether frequent complaints contribute to customer attrition.

## Bivariant analysis: Insights on categorical columns

In order to extract valuable insights from categorical columns, we created dynamic contingency tables that enabled us to compare each variable with our key column, Churn Label.

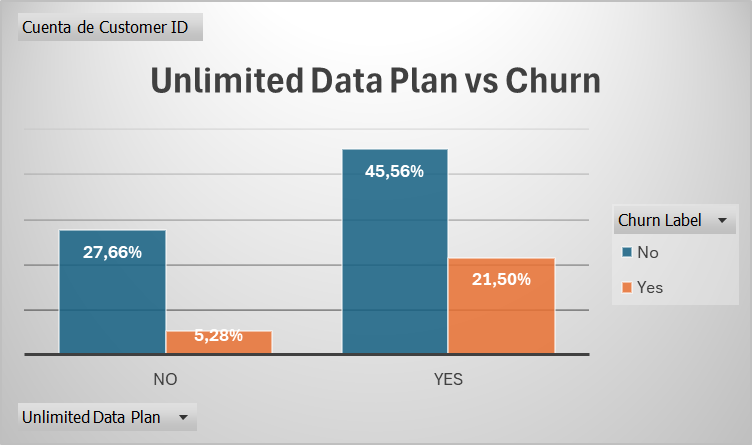
* **Unlimited Data Plan:** we observed that those subscribers enrolled in unlimited data plans duplicated the churn rate of those subscribers with limited data storage. Maybe subscribers are having issues related with that type of plans (maybe cumulative and non-solved), or they are dissatisfied cause it does not cover their particular necessities.
* **Gender:** it does not appear to be a determinant factor for churn rate.
* **Under30**: subscribers under 30 are less likely to cancel the subscription. This might be related with the content offered by the platform, which might be more suitable for this age group.
* **Senior**: these subscribers duplicate the churn rate with respect to the rest, which suggests there might be no suitable contents for this group or that the platform has a poor accessibility. This column could be very related with the above.
* **Group:** subscribers in a Family plan show very low churn rates, most probably because the charge per subscriber for this type of plans is also very low, which is beneficial for them. Another hypothesis is that, even if they don´t use the platform, subscribers in a Family plan tend to maintain the subscription to avoid annoying the rest of family members who still consume the content. Therefore, it is highly recommended to promote this kind of memberships.
* **Device protection & Online BackUp:** subscribers paying for this service are more likely to maintain the membership. This might be due to the feel of security when they acquire this particular service. On the other hand, it might be possible that customers paying for device protection and backup had already planned from the beginning staying enrolled for a long-term in their service plan.
* **Contract type:** subscribers retention is extremely related with the type of contract. The longer the contract, the greater the retention. Subscribers paying for month-to-month memberships may not continue for several reasons: they just wanted to take advantage of the free trial period, stationary membership (very often during Christmas, when people spend more time at home),...In order to reduce churn rates, it would be highly recommended to promote one-year or two-year memberships.
* **Payment method:** there are 16 null values. Subscribers paying by paper check show higher churn rates than the rest. However, this group represents just a small percentage of the total subscribers. High percentages are just a consequence of being less people. On the other hand, it is interesting how subscribers paying by debit card show higher churn rates than those paying by credit card. There are several hypothesis: Subscribers paying by debit card are more aware of the money they have remaining at their bank account, so they just cancel the service before they cannot pay. The second possibility might be a technical problem when paying by debit card. Finally, it would be interesting to analyse the income level of both groups, as it might be a contributing factor.
* **Churn category:** In this case, we haven’t compared the churn category with the churn column, because it had no sense. Instead, we have analysed which percentage represented each category within the total subscribers who had cancelled the membership. The majority of subscribers cancelled the membership due to competitors. There is a tiny minority of subscribers who did not specify the reason for cancelling.
* **Churn reason**: In this case, as before, we have analysed which percentage represented each churn reason within the total subscribers who had cancelled the membership. The majority of subscribers didn't specify the reason for the cancellation of subscription. Therefore, this particular analysis is not interesting.
* **Customer segment**: it does not appear to be a determinant factor for churn rate.
* **Preferred contact method**: it does not appear to be a determinant factor for churn rate.
* **Applied discount**: Surprisingly, the fact that we have applied or not a discount to our subscribers is not a determinant factor for churn rate. Discounts seem a poor strategy for customer retention.

After this analysis we identified 5 categorical columns which showed significant differences in the churn rate between categories:

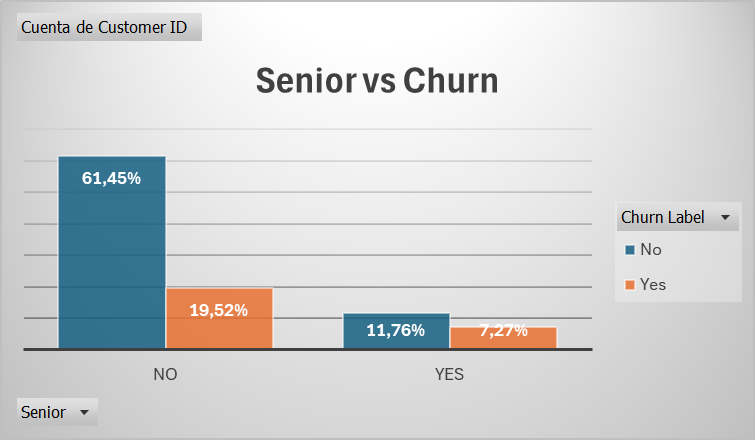
* Unlimited Data Plan
* Senior
* Group
* Contract Type
* Payment Method

In order to check whether the churn rate within each category of those categorical columns represented a significant amount with respect to the total customers, we transformed data in pivot tables into percentages of the total and we created the associated pivot plots:

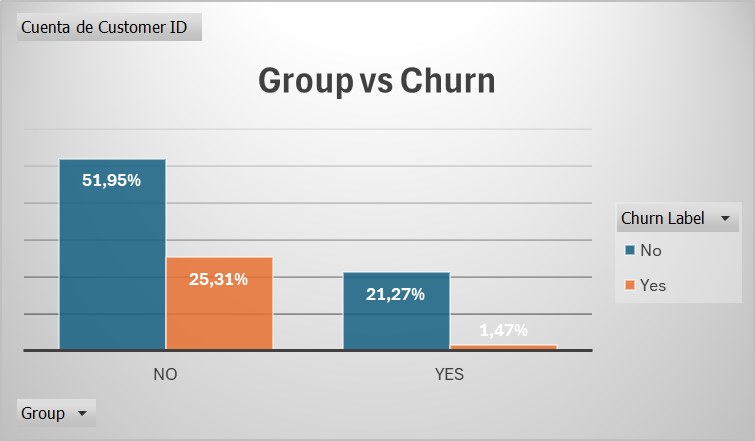
* **Unlimited Data Plan:** we observed thatsubscribers who were enrolled in unlimited data plans represented the 67% of the total customers (they are majority) and those who additionally cancelled the service represented the 22% of the total customers, which seems quite significant and worrying. These high percentages reinforce the necessity to take care of these type of customers.

****

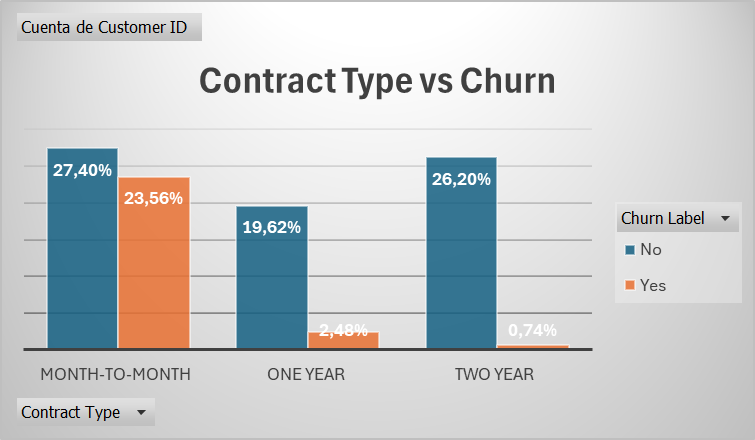
* **Senior:** despite the fact that senior customers duplicated churn rates versus non-seniors, they just represent the 18% of the total customers. Moreover, just seniors who left the service are just the 7% of the total. The company should evaluate whether is sensible to allocate resources for trying to retain this small amount of customers. Maybe is interesting or maybe not.

****

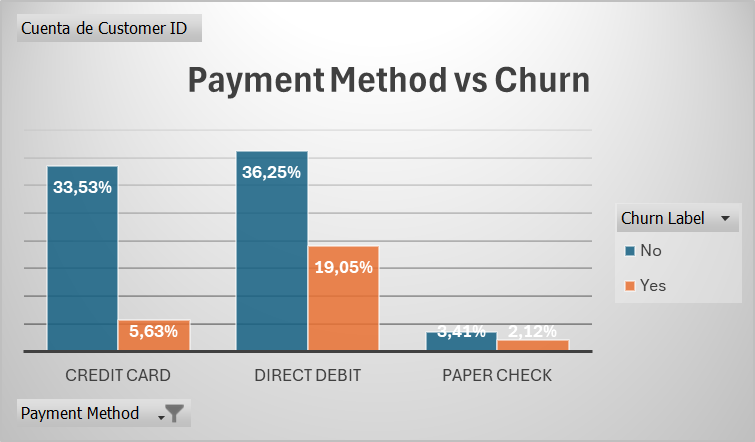
* **Group:** The majority of customers are enrolled in an Individual Plan (approx. 75%) and those which cancelled the subscription went up to the 25%, which is a high percentage. Therefore, it is important for the company to build a specific marketing strategy for this segment of customers.

****

* **Contract Type:** Half of the customers have a month-to-month contract and those which additionally cancelled the subscription went up to the 24%, which is a high percentage. Therefore, it is important for the company to build a specific marketing strategy for this segment of customers.



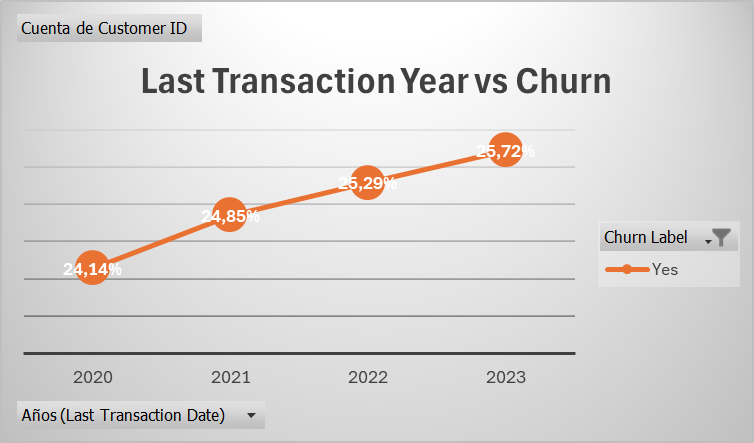
* **Payment Method:** the majority of the customers paid by direct debit (approx. 55%) and those who also cancelled the subscription represented almost the 20% of the total customers, which is a considerable amount. Therefore, it is recommendable to focus also on solving the problems or covering the necessities of this customer segment. Customers paying by credit card and cancelling subscription are just the 5% of the total. Finally, customers paying in paper checks are really insignificant.



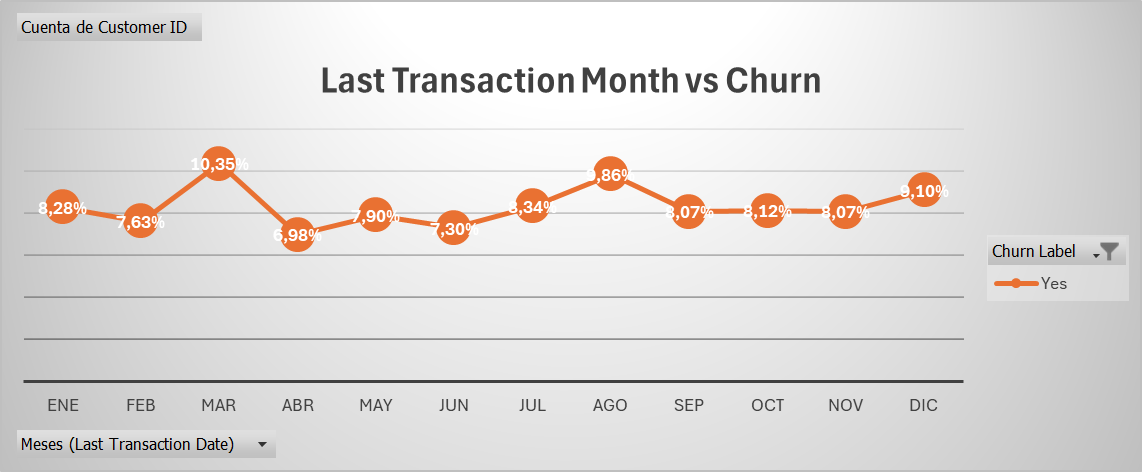
## Bivariant analysis: Insights on date columns

In order to extract valuable insights from date columns, we created dynamic contingency tables that enabled us to compare each date variable with our key column, Churn Label:

* **Contact date:** We observe there are no significant changes in the churn rate with respect to the contact year, quarter or month.
* **Last transaction date:** We observe there are no significant changes in the churn rate with respect to the last transaction year or quarter. However, if we observe the pivot plot for each year, we can see that the general churn rate is slightly increasing over time:



On the other hand, with respect to the last transaction month, in March and August, the percentage of subscribers cancelling subscription after last transaction is significantly higher. This pattern is also observed when representing percentages with respect to the total customers (more or less 10% of churn rate vs more or less 8% for the rest of the months).

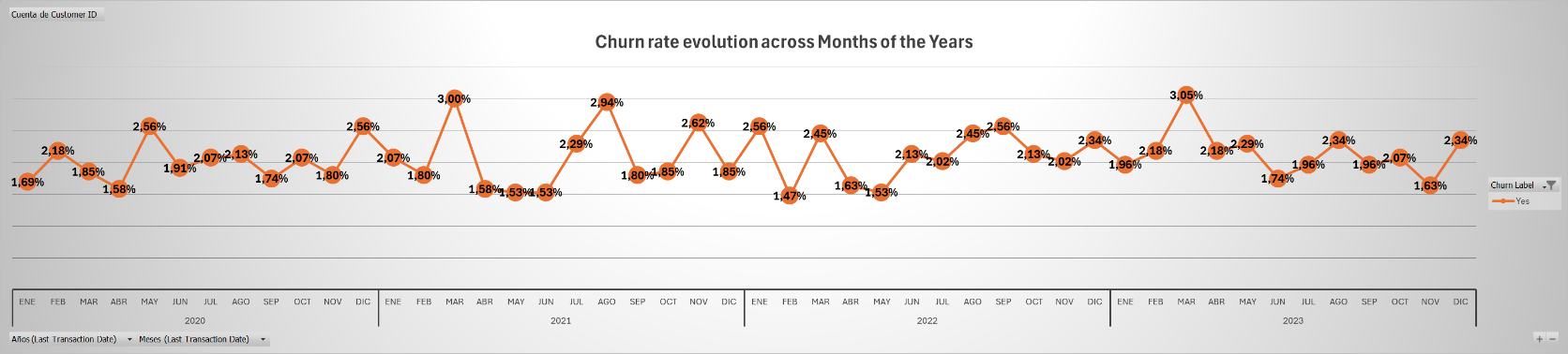


In March, it may happen that the weather starts getting better and people tend to spend less time at home. Therefore, they decide to cancel subscription. Some for August, when there is a majority of people who go on vacation and tend to prefer other ways of entertainment, very often in the outside. It might be interesting to consider this stationary behaviour when planning marketing campaigns.

Finally, from the evolution of the churn rates throughout the months of the years (2020-2023) we observed that the 3 top months with higher churn rates were:

1. March-2023
2. March-2021
3. August-2021

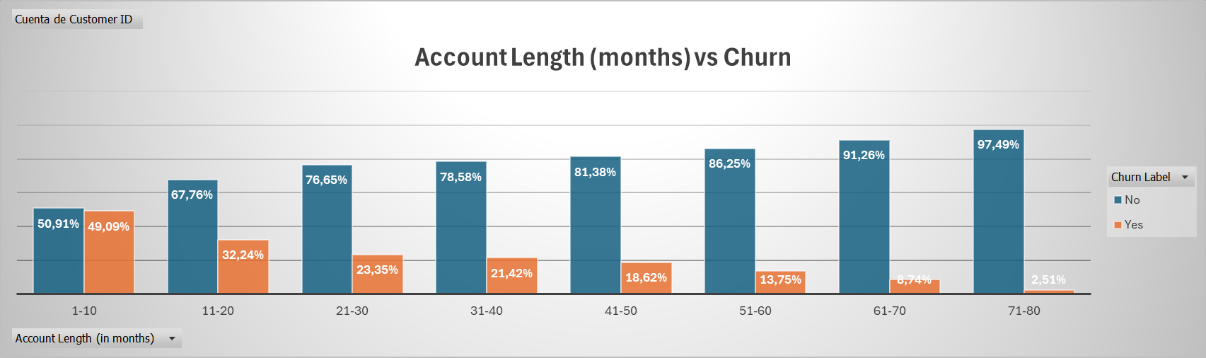
It would be interesting if we could associate some of the subscription cancellations during those specific dates to a particular event affecting the company and propose more accurate insights on this issue:



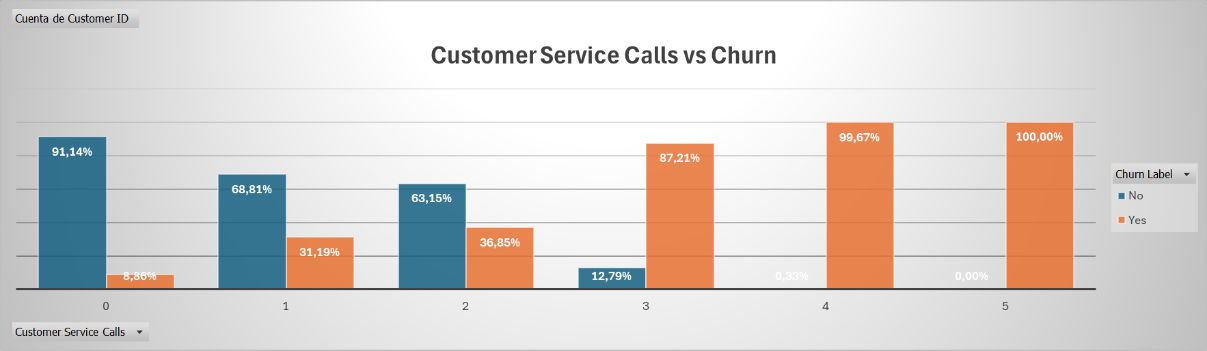
## Bivariant analysis: Insights on numeric columns

Once we had performed univariant analysis on numeric columns, we went further and compared those variables with the target column Churn Label. In this way, we wanted to extract valuable insights on how numerical variables related with the churn rate. For this purpose, we created pivot tables and pivot plots.

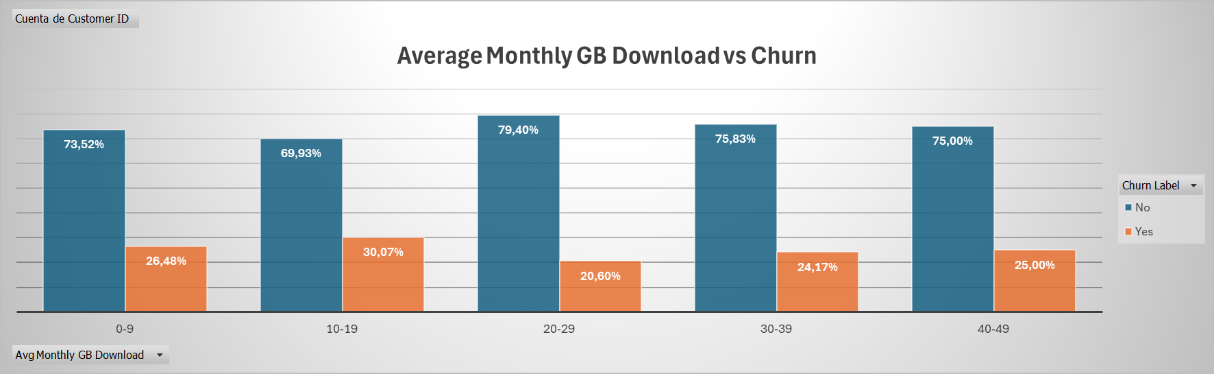
* **Account length (months):** we observed the older the subscriber’s account, the lower the churn rate (lower possibility to cancel subscription). Moreover, subscribers with account lengths between 1 – 5 months, have more possibilities to cancel subscription than to continue in the service. We would highly recommend to promote some kind of benefit plans for loyal customers to encourage keeping the membership.



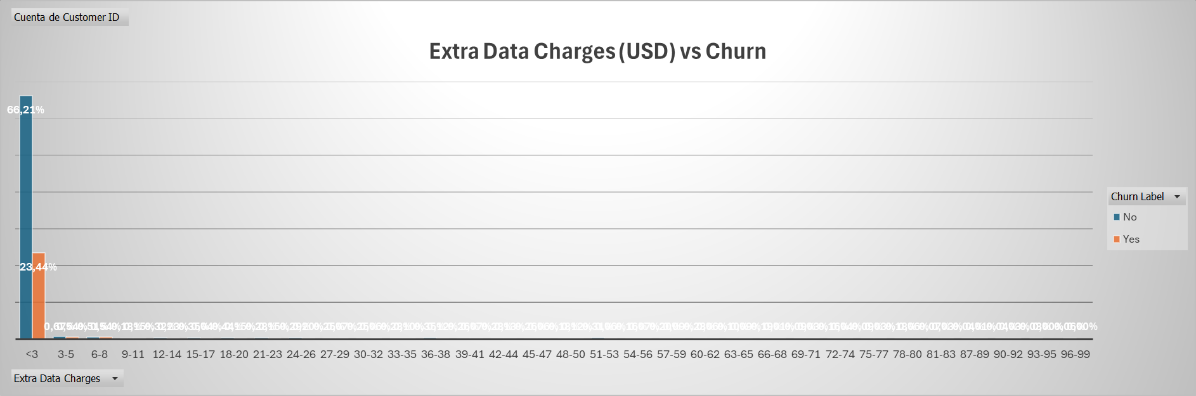
* **Customer Service Calls:** the greater the number of calls to customer service, the higher the probability to cancel the membership. Particularly significant is the pattern change when customers call more than 2 times: the probability of cancelling the subscription increases by a 50% and keeping the membership turns to be the less probable scenario. Therefore, once the customer has called the service, it is critical that the issue gets resolved in a maximum of 2 calls, moment from which the customer very probably gets annoyed and tends to cancel the subscription.

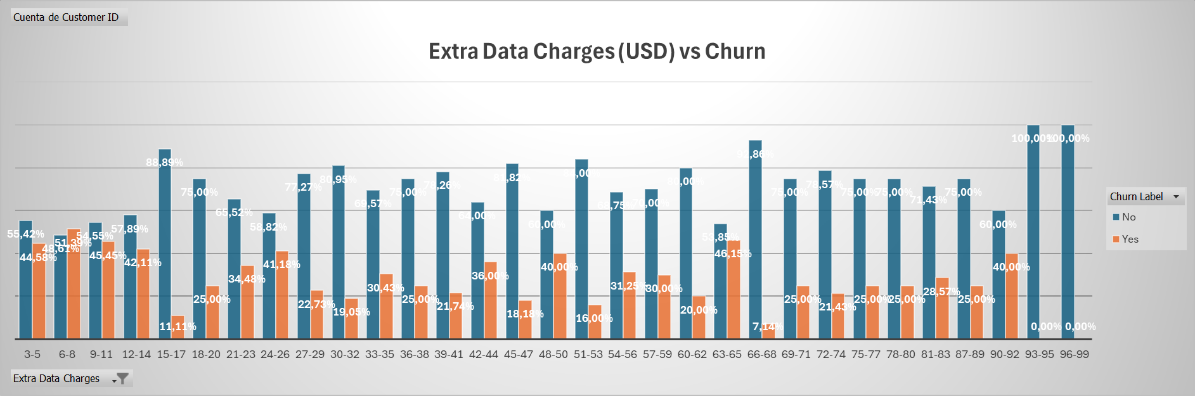


* **Average Monthly GB Download:** there is no significant relationship between the gigabytes downloaded per month and the probability of cancelling the membership. In other words, customer making no use of this feature have the same probability of cancelling the membership than customers downloading great amounts of gigabytes.

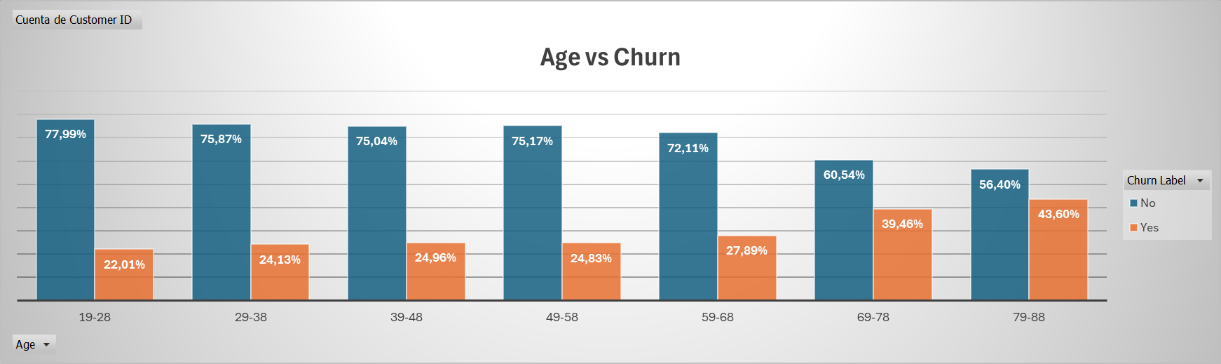


* **Extra Data Charges:** the vast majority of subscribers (90% approx.) do not have extra data charges. Therefore, in this case, the only interesting analysis would be to focus on the remaining 10%. After filtering, we observe there is significant relationship between the variables, however, there is not enough data per segment to establish strong conclusions on this particular case.

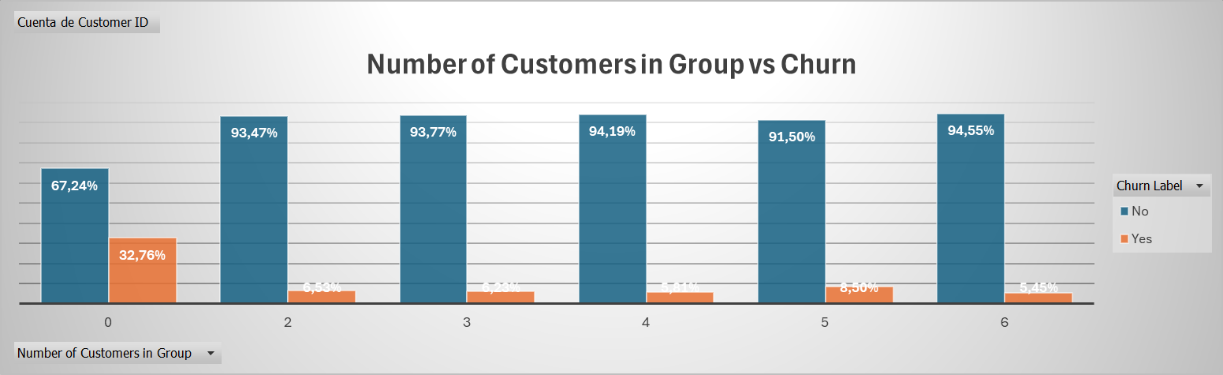




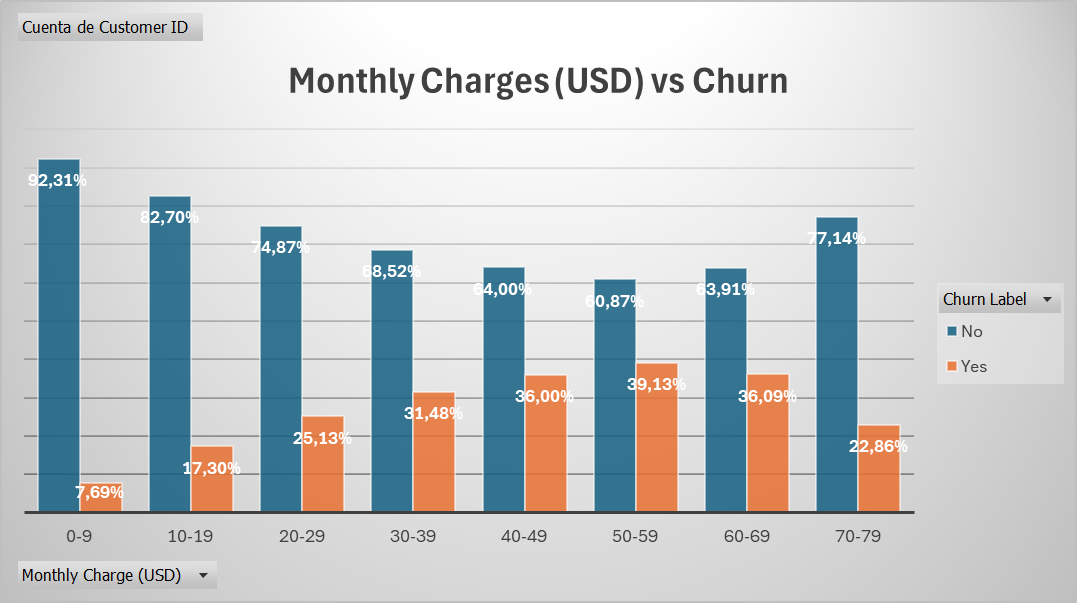
* **Age**: churn rates keep constant (approx. 24%) until subscribers reach the senior group, when the probability of cancelling subscription starts increasing significantly. This could be related with two issues: the majority of audiovisual content that is offered does not suit elder people or the platform lacks accessibility for this age group. It would be interesting to further investigate the reason why elder people cancel the subscription and, maybe, launch a specific marketing campaign for seniors applying strategies to reduce their churn rates.



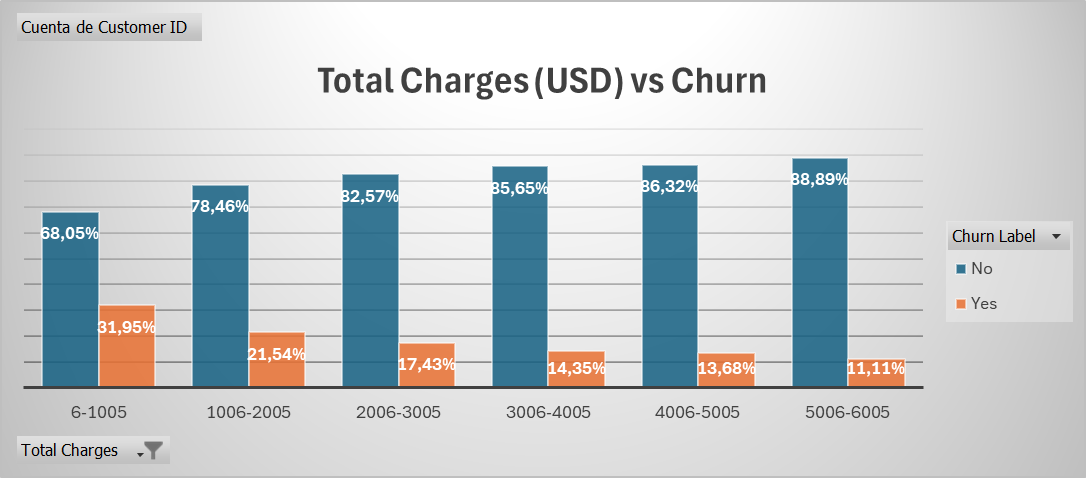
* **Number of customers in Group:** almost the 80% of subscribers pay for an Individual Plan. However, that is the plan with higher churn rates (32.76%). The probability of cancelling subscription decreases abruptly (down to approx. 6%) if the subscriber pays for a Family Plan, no matter how many customers integrate the family plan. The main reason for these patterns might be related with the compromise that subscribers feel when paying for a Family plan, preferring to maintain subscription if one of the group wants to keep enjoying the service. It is highly recommended to foster Family Plans rather than individual plans to reduce general churn rates.



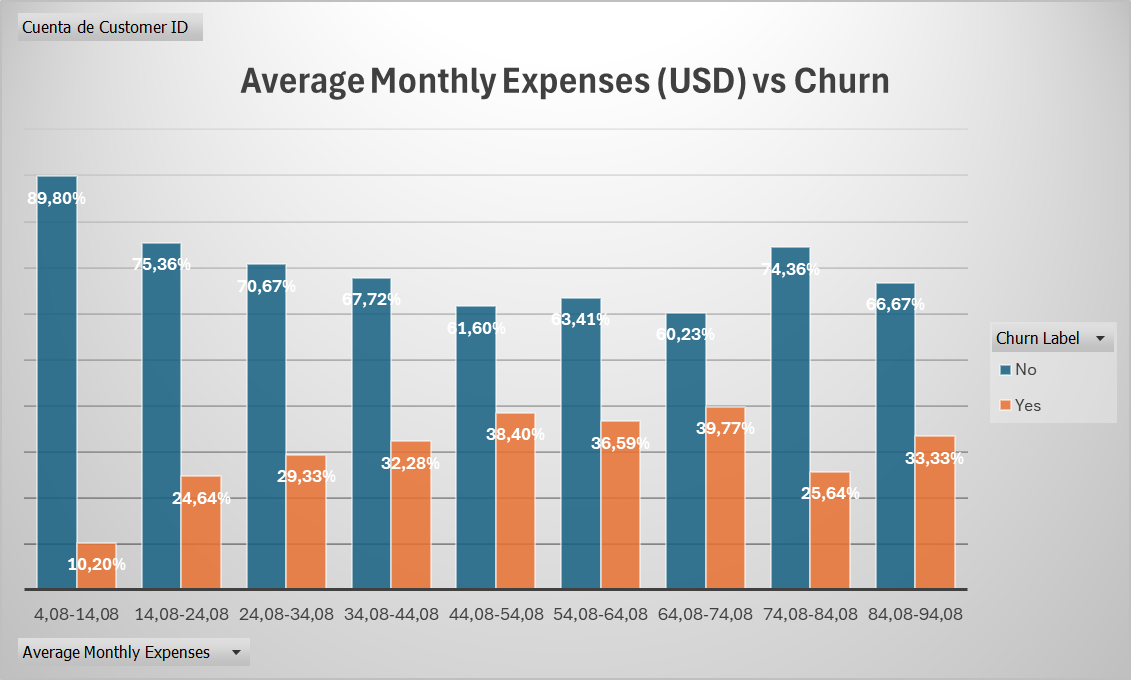
* **Monthly Charges (USD):** the greater the monthly charges, the higher the probability of subscription cancellation (this rule is not followed for the last two segments, however, those segments represent a tiny proportion of the total subscribers). Customers paying greater amounts of money may keep more track of their expenses and evaluate more deeply the quality-price ratio for what they have paid, finally switching for alternatives. It is recommendable to investigate whether there are significant benefits to those customer paying such quantities for the service and evaluating if they are really worth it.



* **Total charges (USD):**  we observed the greater the total charges (USD), the lower the probability of cancelling membership. However, this is probably related with the time from which the customers have been paying for the membership: the greater the account length, the greater the accumulated charges, and the lower the churn rates. As said before, we would highly recommend to promote some kind of benefit plans for loyal customers to encourage keeping the membership.



* Average Monthly Expenses (USD): this column behaves very similarly to the Monthly Expenses variable and the same insights are applicable.



* **Number of Complaints or Support**: Unexpectedly, customers with no registered complaints or support requests show similar churn rates to those with 1 to 10 requests. In contrast, the number of customer service calls is positively correlated with churn.

This discrepancy may be due to the nature of the issues. Customers who call customer service are likely dealing with critical problems that require immediate resolution or have reached a higher level of dissatisfaction. When these issues remain unresolved after multiple attempts, customers are more likely to cancel their memberships.

On the other hand, customers submitting complaints or support requests may be facing less urgent issues or have not yet reached the frustration threshold that prompts a call. Additionally, the complaint/support platform may provide a more effective and satisfying resolution process compared to customer service calls.

We recommend evaluating the effectiveness of the customer service process, particularly in resolving issues within the first two calls, as unresolved problems are a key driver of churn. It may also be beneficial to analyze the strengths of the complaint/support platform to identify practices that could be applied to customer service calls.

