EDA Notebook (01_eda.ipynb)

Goal: Explore, clean, and understand data.

1. Introduction

This project explored the 'Telco Customer Churn' dataset to investigate the key factors that drive customer churn and to provide a guide for how to approach predictive modelling. The data used in this project were collected from [1], which was originally provided by IBM [2]. It contains 7043 observations and 21 variables collected and documented by the fictitious company 'Telco', an organisation that offers internet and home phone services to residents in California.

The objectives of this EDA were to:

- Understand what the data looks like.
- Identify patterns, relationships, or trends in the data.
- Highlight any errors, missing values, or outliers that might have been present.
- Discern what variables drive customer turnover (churn).
- Establish a forward path for predictive modelling.

It is important to analyse the relevant data and achieve the objectives above to predict the likelihood of customer defection and develop a strategised customer retention program.

2. Data Overview

The dataset contains 7043 rows and 21 columns. The columns were an amalgamation of categorical and numerical data types. Below is a brief description of the 21 variables contained within the dataset provided by [2].

2.1. Sorted by Demographics:

i. Gender: The customer's gender.
Data type: `Object`
Values: `['Female', 'Male']`

ii. Senior Citizen: Indicates if a customer is 65 or older.
Data type: `Integer`
Values: `['Yes', 'No']`

iii. Partner: Indicates if a customer has a partner.
Data type: `Object`
Values: `['Yes', 'No']`

iv. Dependents: Indicates if a customer lives with any dependents: children, parents, grandparents, etc.
Data type: `Object`
Values: `['Yes', 'No']`

2.2. Sorted by Status:

v. CustomerID: A unique ID that identifies each customer.
Data type: `Object`
Values: `Values are unique to each customer.`

vi. Churn Label: Indicates whether a customer has left the company. Directly related to Churn value and what we are trying to predict.
Data type: `Object`
Values: `['Yes', 'No']`

2.3. Sorted by Services:

vii. Tenure: Indicates the total number of months that a customer has been with the company by the end of the third quarter.
Data type: `Integer`
Values: `Values range from [0 - 72]`

viii. Phone Service: Indicates if a customer subscribes to a home phone service with the company.
Data type: `Object`
Values: `['Yes', 'No']`

ix. Multiple Lines: Indicates if a customer subscribes to multiple telephone lines with the company.
Data type: `Object`

Values: [‘Yes’, ‘No’, ‘No Phone Service’]

x. Internet Service: Indicates if a customer subscribes to an internet service with the company.

Data type: Object

Values: [‘DSL’, ‘Fibre Optic’, ‘Cable’]

xi. Online Security: Indicates if a customer subscribes to an additional online security service provided by the company.

Data type: Object

Values: [‘Yes’, ‘No’, ‘No Internet Service’]

xii. Online Backup: Indicates if a customer subscribes to an additional online backup service provided by the company.

Data type: Object

Values: [‘Yes’, ‘No’, ‘No Internet Service’]

xiii. Device Protection: Indicates if a customer subscribes to an additional device protection plan for their Internet equipment provided by the company.

Data type: Object

Values: [‘Yes’, ‘No’, ‘No Internet Service’]

xiv. Tech Support: Indicates if a customer subscribes to an additional technical support plan from the company.

Data type: Object

Values: [‘Yes’, ‘No’, ‘No Internet Service’]

xv. Streaming TV: Indicates if a customer uses their internet service to stream television programming from a third-party provider. The company does not charge an additional service fee for this.

Data type: Object

Values: [‘Yes’, ‘No’, ‘No Internet Service’]

xvi. Streaming Movies: Indicates if a customer uses their internet service to stream movies from a third-party provider. The company does not charge an additional service fee for this.

Data type: Object

Values: [‘Yes’, ‘No’, ‘No Internet Service’]

xvii. Contract: Indicates a customer’s current contract type.

Data type: Object

Values: [‘Month-to-Month’, ‘One year’, ‘Two year’]

xviii. Paperless Billing: Indicates if a customer has chosen paperless billing.

Data type: Object

Values: [‘Yes’, ‘No’]

xix. Payment Method: Indicates how a customer pays their bill.

Data type: Object

Values: [‘Electronic check’, ‘Mailed check’, ‘Bank transfer (automatic)’, ‘Credit card (automatic)’]

xx. Monthly Charge: Indicates a customer’s current total monthly charge for all their services from the company.

Data type: Float

Values: Values range from [18.25 – 118.75]

xxi. Total Charges: Indicates a customer’s total charges, calculated to the end of the third quarter.

Data type: Object

Value(s): Values range from [18.8 – 8684.8]

Upon initial analysis of these features, there were some key issues that stood out. The first is the data type for the ‘TotalCharges’ variable. [2] states that it is the total charges calculated at the end of the quarter for each customer. If tenure is the number of months each customer stayed with the company, and there is a total monthly charge billed to them for all the services they are subscribed to, one would expect ‘TotalCharges’ to be the product of these two. Instead, the column was stored as an object rather than a numeric type, which was corrected by converting it to a float.

However, correcting the data type revealed another issue: inconsistencies in the actual values. In theory, total charges should equal tenure multiplied by monthly charges, but closer inspection showed that some entries differed from this expected product by tens, and in some cases, even hundreds.

Another issue revealed during analysis was that the ‘TotalCharges’ variable contained 11 rows with undefined values represented as ‘NaN’. This suggested that the total amounts billed to these customers were either not recorded or not yet available. To investigate, the corresponding rows for ‘tenure’, ‘MonthlyCharges’, and ‘TotalCharges’ were examined. It was found that all customers with missing total charges had a tenure of ‘0’, meaning they were new customers who had only just subscribed to the service. Since ‘TotalCharges’ reflects the cumulative amount billed over time, it is logical that no value exists yet for these customers. Their ‘MonthlyCharges’, however, were still present, as these represent the expected recurring cost. To address the missing values, two options were considered: either dropping the rows with undefined ‘TotalCharges’ or imputing appropriate values to replace the ‘NaN’ entries.

Beyond issues with ‘TotalCharges’, there were also unusual values in other columns. For example, ‘MultipleLines’ included entries such as ‘Yes’, ‘No’, and ‘No phone service’, while he internet-related columns had values like ‘Yes’, ‘No’, and ‘No internet service’. To improve readability and reduce redundancy, it would be better to standardise the ‘No phone service’ and ‘No phone service’ entries to simply ‘No’. Similarly, the ‘PaymentMethod’ contained lengthy labels such as ‘Electronic check’, ‘Mailed check’, ‘Bank transfer (automatic)’, and ‘Credit card (automatic)’. These could be simplified by grouping them into categories: automatic (e.g., electronic check, bank transfer, credit card) and manual (e.g., mailed check). This simplification may help identify whether the payment method influences churn.

Finally, while no duplicate columns were found in the data, one irrelevant feature was identified: the ‘customerID’. As it is merely an identifier and not useful for modelling, it can be dropped.

3. Data Cleaning & Preprocessing for EDA Purposes

To address the missing entries in ‘TotalCharges’, the decision was made to drop the affected rows rather than impute values. With only 11 missing entries out of thousands, their removal was unlikely to have a significant impact on the analysis and modelling. Moreover, these rows all corresponded to customers with a tenure of zero. Since ‘TotalCharges’ is expected to equal tenure multiplied by monthly charges, the true value in such cases would logically be zero. This further justified removing the rows.

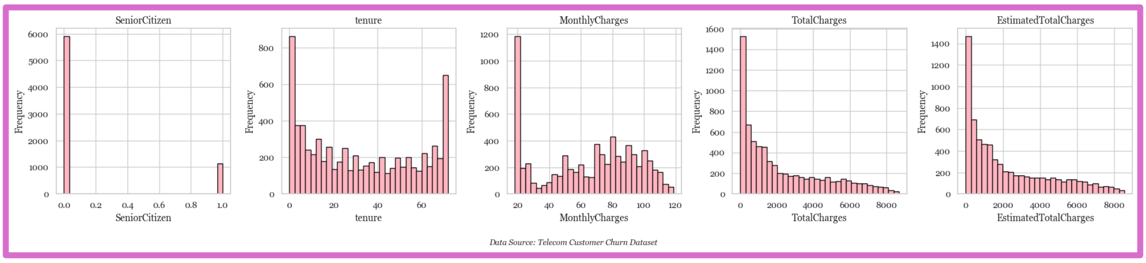
Still on the issue of ‘TotalCharges’, a new feature was engineered to further examine discrepancies found in the column. The feature is called ‘EstimatedTotalCharges’ and was formed as the true product of ‘tenure’ and ‘MonthlyCharges’. When juxtaposed with ‘TotalCharges’, it revealed the inconsistencies found within the column. Some values in ‘TotalCharges’ were slightly lower than the expected product, while others were slightly higher, highlighting that only 614 out of 7043 entries in the two variables have identical values. It was inferred that these discrepancies were most likely due to factors such as partial months of service, extra charges billed to the customers, fees, or credits that were not captured in the product. The ‘EstimatedTotalCharges’ was left in the data for further analysis. At this stage, it was speculated to highlight customers with unusual billing behaviours, which in turn might relate to churn.

Moving on to other issues, the atypical values for ‘MultipleLines’ and for the internet-related columns, ‘No phone service’ and ‘No internet service’, were standardised to ‘No’ for readability purposes and to reduce redundancy. Similarly, all lengthy labels belonging to ‘PaymentMethod’ were also changed. ‘Electronic check’, ‘Bank transfer (automatic)’, and ‘Credit card (automatic)’ were all categorised under ‘Automatic’, while ‘Mailed check’ was changed to ‘Manual’ to help the model capture what drives customer turnover more effectively.

Finally, further modifications of the dataset include dropping the ‘customerID’ column, as it won’t be useful when modelling.

4. Univariate Analysis

4.1. Numerical Variables



Data Source: Telecom Customer Churn Dataset

The diagram illustrates the distributions for the numerical variables in the ‘Telco Customer Churn’ Dataset.

The plots show the distributions of the values for the numerical variables: ‘SeniorCitizen’, ‘tenure’, ‘MonthlyCharges’, ‘TotalCharges’, and the new feature, ‘EstimatedTotalCharges’.The data shows that ‘SeniorCitizen’ is very imbalanced, indicating that the majority (more than half) of the company's customers are not senior citizens. A further investigation will need to be conducted to determine if older customers are at risk of high churn rates.

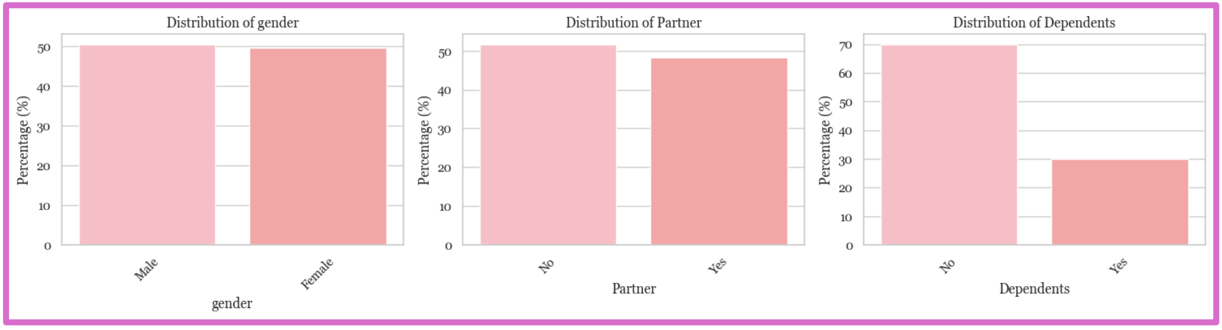
Moving on to the tenure distribution, the plot shows a U-shaped histogram that is slightly right-skewed (showing more customers have a shorter tenure). It would be valuable to examine whether shorter tenures are associated with higher churn rates and longer tenures with lower churn. The plot also shows high variability in customer loyalty duration. There’s a high frequency of customers at the start and end of the tenure range, suggesting that short-term and long-term customers are more common than mid-term customers. Feature engineering could involve transforming tenure into bins or derived features to help the model better capture churn patterns.

As for ‘MonthlyCharges’, the distribution shows a wide range of prices. The plot is skewed slightly to the left with one big spike at the low end, which most likely corresponds to customers on basic plans, while other prices are distributed broadly, suggesting that these customers may be on either minimal or premium plans. The next step could be to explore churn patterns by pricing tiers.

Initially, the ‘TotalCharges’ variable presented some issues. The plot did not align with the summary statistics, appearing cluttered and inconsistent. This was traced back to the column being stored in the wrong data type and containing null values. After correcting these issues, the distribution became clearer and revealed a strong right skew.

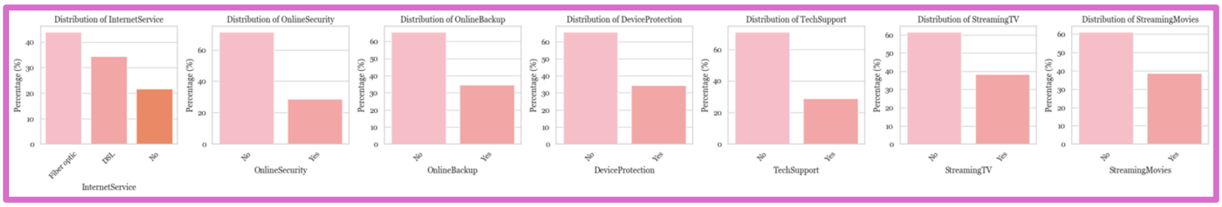
The cleaned plots for both ‘TotalCharges’ and the engineered feature ‘EstimatedTotalCharges’ are now nearly identical. Both highlight that while many customers make relatively small payments, there is a significant subset of long-term customers who accumulate very high charges. Since these two variables are closely related, it will be important to monitor for multicollinearity when building models.

4.2. Categorical Variables



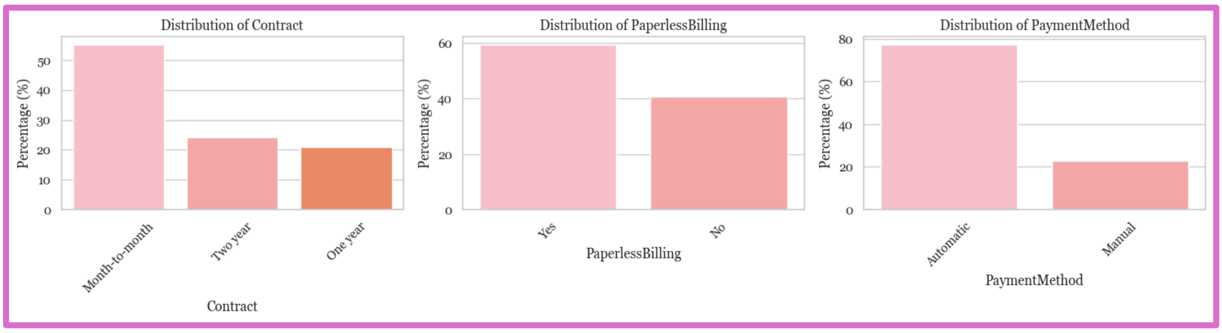
The diagram reveals three barplots that illustrate the distributions for ‘gender’, ‘Partner’, and ‘Dependents’ in the ‘Telco Customer Chur’ Dataset..

The section explores the distributions of categorical features such as gender, internet services, contract types, and more, as well as the target variable ‘Churn’. Gender appears relatively balanced across the dataset, indicating that it may not be a major factor influencing churn. In contrast, variables such as partnership and dependents status may provide more insight into customer stability and the effects it may have on turnover. For example, customers with partners or dependents may exhibit greater financial security and long-term commitment, making them less likely to churn compared to those without.



The diagram reveals seven barplots that illustrate the distributions for all internet-related services in the ‘Telco Customer Churn’ Dataset.

Beyond demographic factors, service-related features also play an important role in understanding churn. More customers are not subscribed to an internet service or any of the added services, such as online security. Because the added internet services like streaming TV and streaming movies do not come with an additional charge, customers might enjoy their subscriptions more if they were encouraged to add these services to their plan, thereby increasing commitment and thus lowering the risk of defection.

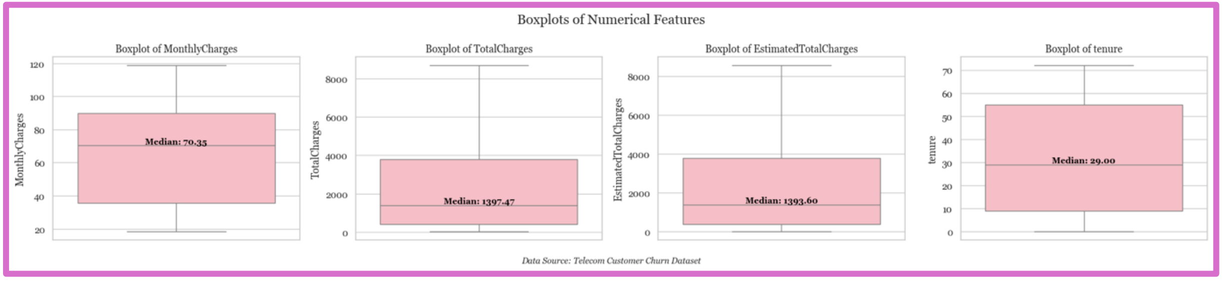


This diagram shows how the variables 'Contract', 'Paperlessbilling', and 'PaymentMethod' influence the target variable 'Churn'.

The majority of customers are on month-to-month contracts, which, while offering flexibility, also carry a higher churn risk than customers who have longer contract types. Billing preferences add another layer of insight. A larger share of customers still opt for paper billing, which may reflect demographic differences such as age or technology preferences, though its direct impact on churn requires further analysis. Payment methods also reveal meaningful patterns: many prefer automatic payments in lieu of manual ones. This behaviour could be a strong indicator that may contribute to higher retention rates.

As for the target variable, 'Churn', the dataset is imbalanced, with ~26% 'Yes' and ~74% 'No', which has implications for model training if not handled properly.

4.3. Insights



Data Source: Telecom Customer Churn Dataset

The above diagram corroborates the results of the histogram in section 4.1.

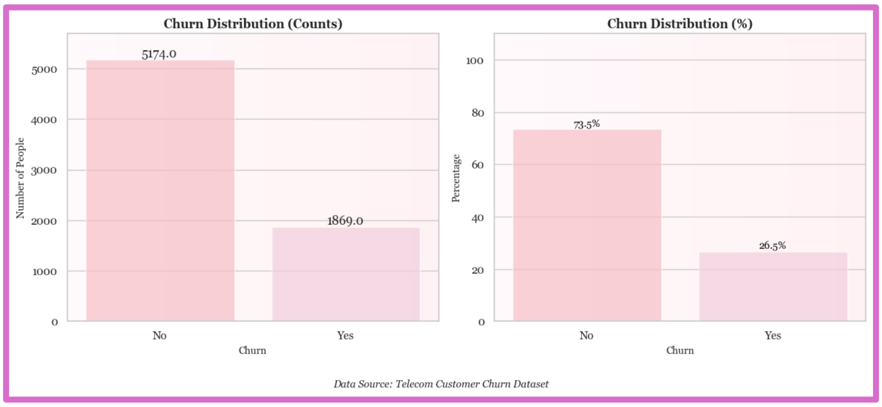
The boxplots for the numerical features reaffirm the results of the histograms. For instance, it shows that 'TotalCharges' and 'EstimatedTotalCharges' are very highly skewed. Although both distributions have wide spreads, their long upper whiskers suggest strong skewness, which could affect model sensitivity. A transformation, such as a log scale, may be needed to reduce their impact. They also likely contain outliers or customers with long tenure and high monthly charges, and could strongly distort the mean and affect classification.

'MonthlyCharges', however, shows that most customers are charged between \$35 and \$90 per month, with the median being \$70.35, indicating many customers are subscribed to mid-range pricing plans. In contrast with 'TotalCharges' and 'EstimatedTotalCharges', its distribution is fairly symmetric, with few or no visible outliers, suggesting that the variable needs minimal transformation before modelling.

As for 'tenure', its boxplot has a somewhat positively skewed distribution with no significant outliers present. The feature, however, may still benefit from binning or scaling.

All four features have different scales and can mislead distance-based models (e.g. KNN, SVM). They will need to be standardised so that models can assign equal attention to each feature.

In summary, the patterns revealed by the boxplots not only confirm earlier insights but also guide the preprocessing strategy for modelling. Addressing skewness, potential outliers, and scale differences will be key to building more robust and reliable, predictable models.



Data Source: Telecom Customer Churn Dataset

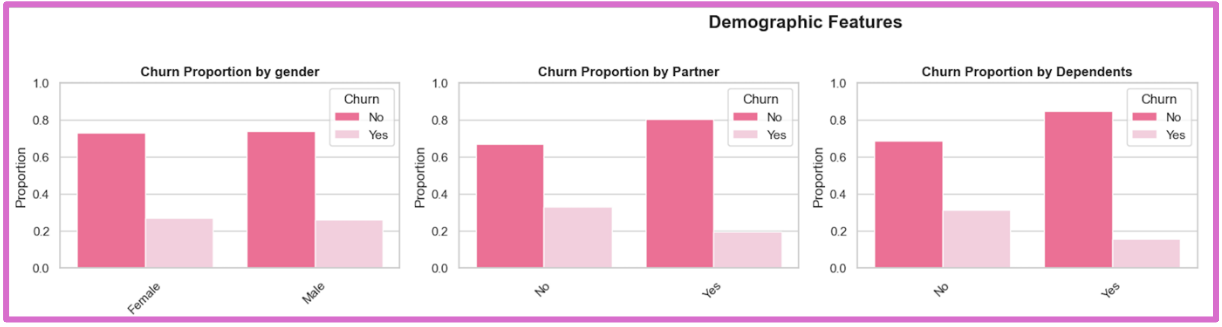
The diagram offers a visual representation of the class imbalance for the target variable 'Churn' in the 'Telco Customer Churn' Dataset.

Alongside these numerical considerations, the bar chart analysis of the target variable reveals a clear class imbalance: the majority of customers did not churn ('No'), while fewer customers churned ('Yes'). This imbalance poses a challenge, as models may default to predicting the majority class and still achieve high accuracy, but at the cost of misclassifying the minority churn cases, which is precisely the outcome that needs to be captured.

Together, these insights underline the importance of careful preprocessing: applying transformations and scaling for the features while also addressing class imbalance through techniques such as resampling, class weighting, or specified evaluation metrics. Both aspects are critical to ensuring the model learns meaningful churn patterns rather than being skewed by outliers, scaling differences, or class dominance.

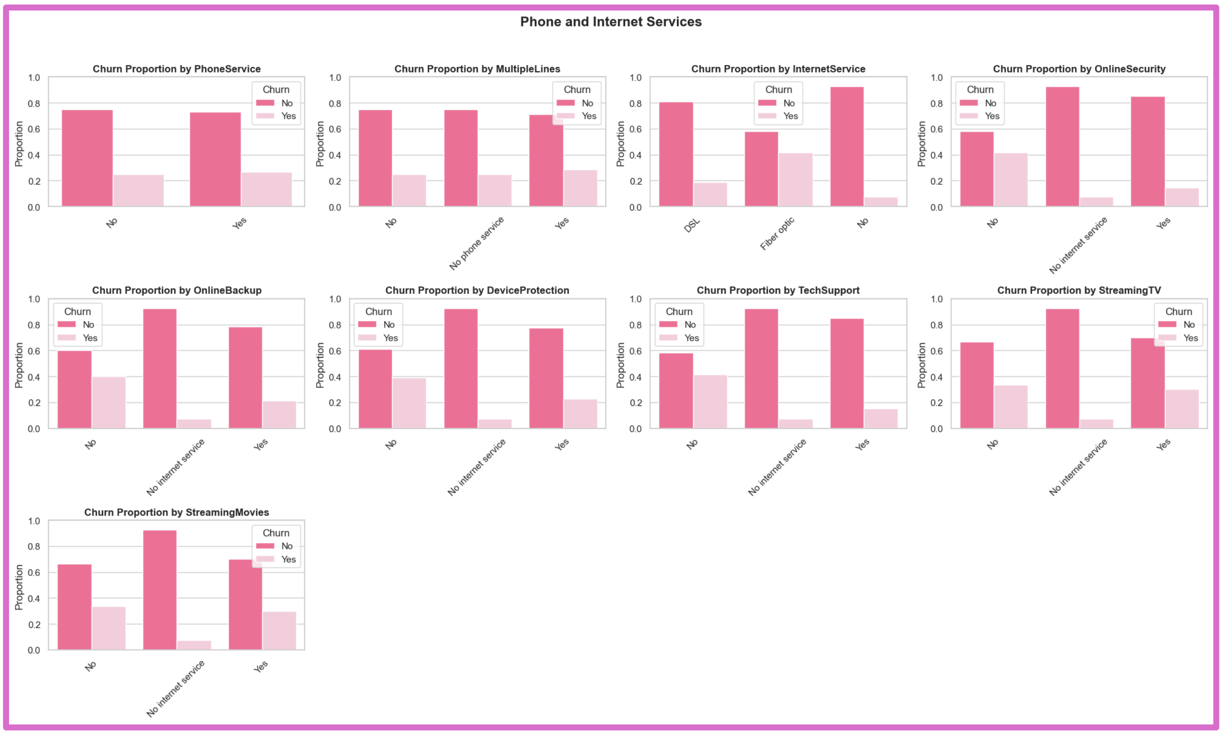
5. <u>Bivariate Analysis</u>

5.1. Churn Observations by Categorical Features



The figure above highlights the distribution of churn observations by demographic.

When analysing how much gender contributes to customer churn, it was observed that the churn rates for females and males are almost the same, with only a 0.76% difference. This suggests that gender may not have a strong influence on churn rates. In contrast, it was discovered that customers without partners have a higher churn rate of 32.96% compared to those with partners, who had churn rates of 19.66%. Comparably, customers without dependents also showed strong churn influence when compared to those who do not have, with a staggering 31.28% vs 15.45%. These values give the impression that customers with partners and/or dependents may be more committed, possibly due to shared decision-making or stability.

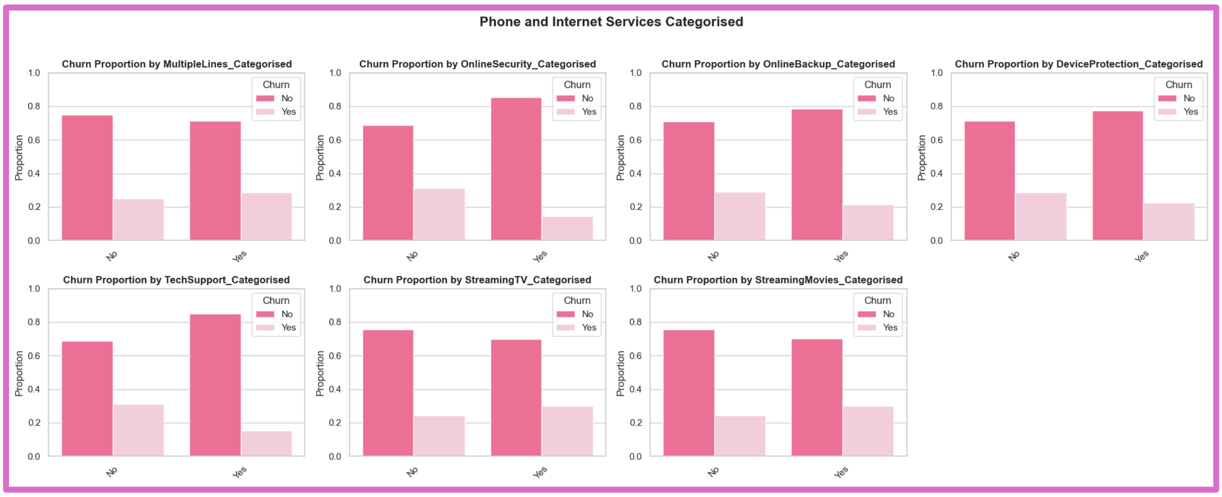


This displays the distribution of churn observations by phone and internet-related service before "No phone service" had been changed to "No".

Moving on to the service features before categorisation, the churn rates for customers with and without phone service are quite similar (24.93% vs 26.71%), indicating that whether or not a customer is subscribed to a phone service may not be a significant factor that drives churn. The same can be seen with the multiple lines feature. There is not much of a churn difference between customers who have multiple phone service lines and those who don't, suggesting that this feature might have some relationship with churn, but not a strong one.

For internet-related services, customers with fibre optic service show the highest churn rate (41.89%) compared to DSL (18.96%) and no internet service (7.40%), suggesting that internet service type strongly influences churn, with fibre users being the most at risk. Similarly, customers lacking online security face a churn rate of 41.77% versus just 14.61% for those with the service, highlighting security as a powerful retention factor. Online backup also appears influential, as churn is much higher among customers without it (39.93%) compared to those with it (21.53%). Device protection shows a similar trend, with churn at 39.13% for those without it and 22.50% for those with it, reinforcing the link between add-on services and customer loyalty.

Tech support emerges as one of the strongest factors, with customers lacking support churning at 41.64%, while only 15.17% of those with support leave. In contrast, entertainment services like StreamingTV and StreamingMovies show only a modest effect: churn is slightly higher among non-users (33.52% and 33.68%, respectively) compared to users (30.07% and 29.94%). This suggests that while engagement services contribute somewhat, security, backup, and support services play a far greater role in reducing churn.



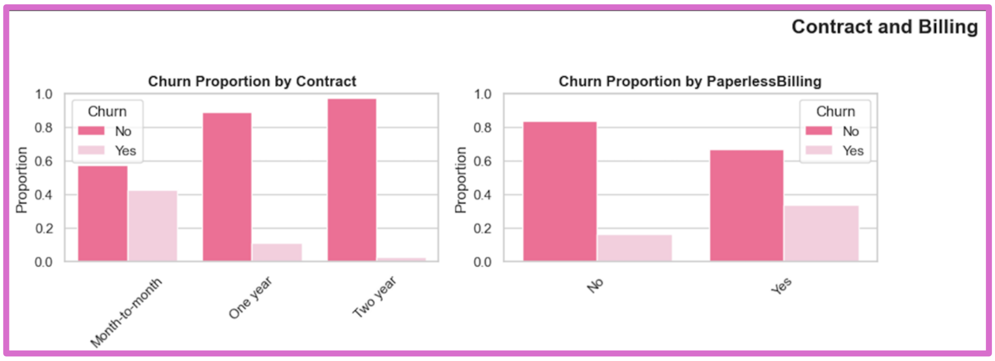
This displays the distribution of churn observations by phone and internet-related service after "No phone service" had been changed to "No".

When focusing on the re-categorised features (MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies), the patterns in churn become less distinct than before. Originally, many of these variables had three levels (Yes / No / No phone or internet service). Collapsing them into binary values (Yes/No) simplified the data, but it also blurred some important signals. This is because the "No phone/internet service" groups generally had much lower churn; once merged into the broader "No" category, they artificially boosted retention for that group, shrinking the gap against "Yes" and diluting the predictive power.

For example, 'MultipleLines' initially showed that "No phone service" customers churned the least. After collapsing, however, the distinction between "Yes" and "No" became weak. A similar effect occurred with 'OnlineSecurity' and 'TechSupport'. Before re-categorisation, both displayed a sharp contrast; customers with these services churned far less than those without. After merging, the separation remains but is less pronounced, reducing its actionability.

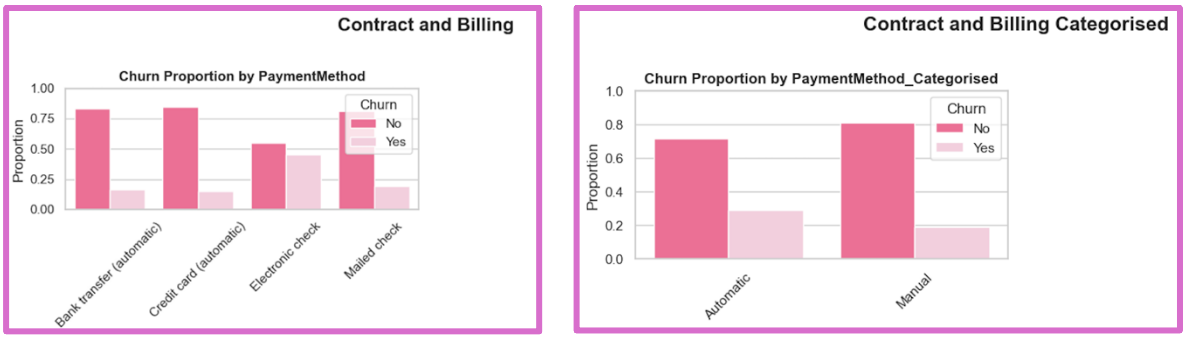
The same pattern applies to 'OnlineBackup' and 'DeviceProtection', although in their case the signal weakened more severely, losing much of the clarity observed before re-categorisation. 'StreamingTV' and 'StreamingMovies' were also affected: the very low-churn "No internet service" group inflated the "No" category, in some cases making differences between "Yes" and "No" almost disappear.

Overall, binary encoding made the plots cleaner but came at the cost of interpretability. For churn prediction or actionable insights, it may be better to retain the original three levels or engineer new flags that specifically indicate whether a customer subscribes to add-on services. This way, strong patterns like those seen for OnlineSecurity, TechSupport, and 'DeviceProtection' can be preserved without being washed out.



The figure above highlights the distribution of churn observations by the variables 'Contract' and 'PaperlessBilling'.

Moreover, month-to-month contracts show a much higher churn rate (42.71%) compared to one-year (11.27%) and two-year (2.83%) contracts. This conveys that the type of contract that a customer has strongly correlates with churn, with month-to-month customers being more likely to leave. Furthermore, the type of billing systems a customer is registered under seems to also be a strong factor that drives customer churn. Customers with paperless billing churn more at a rate of 33.57% than those without it at a rate of 16.33%.



A side-by-side comparison of the differences in the distribution of churn observations for the 'PaymentMethod' variable, before and after categorisation.

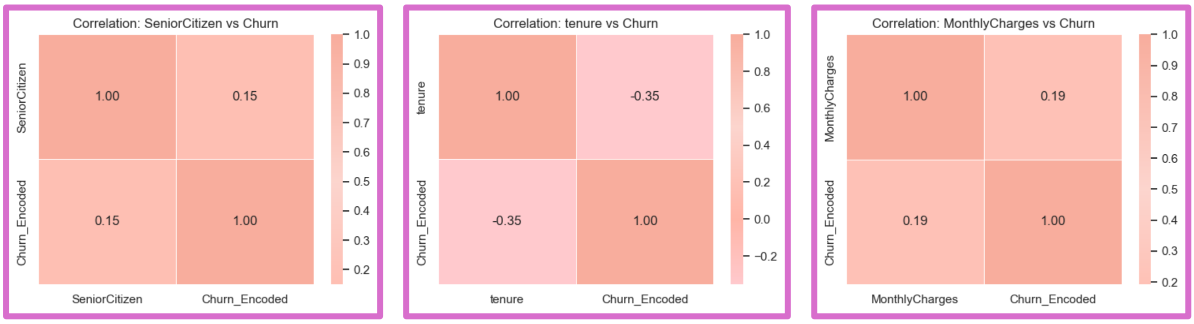
The above plots show the churn observations for 'PaymentMethod' before and after categorisation. Looking at the original PaymentMethod feature, churn patterns vary meaningfully across categories. Customers paying via electronic check exhibit noticeably higher churn compared to those using bank transfer, credit card, or mailed check, suggesting that electronic payment users may be more prone to leaving.

When these payment methods are re-categorised into broader groups, 'Automatic' (electronic check, bank transfer, credit card) versus 'Manual' (mailed check), the distinction becomes less pronounced. Churn differences between the two groups narrow, potentially masking the stronger signal carried by the original categories, particularly the elevated churn risk of electronic check users.

Because the two approaches tell slightly different stories, one highlighting a sharp risk factor, the other simplifying categories at the cost of nuance, Both versions will be includes during modelling. This will help determine whether preserving the granular detail of the original feature yields better predictive power or whether the simplified grouping improves generalisation.

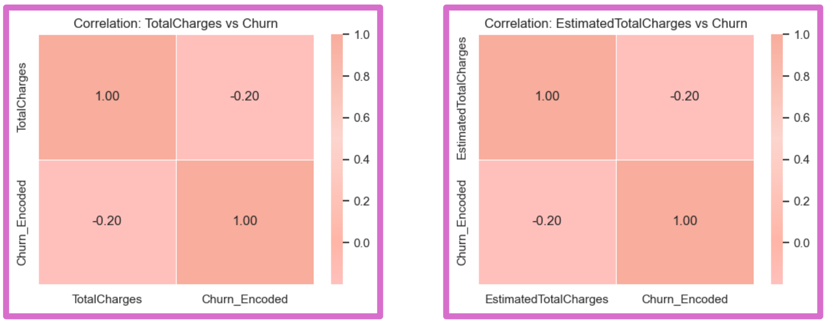
5.2. Churn Observations by Numerical Values

The following heatmaps illustrate how much of an influence the numerical values contribute to a higher churn rate, with a strong positive correlation being associated more with customer turnover and a negative correlation associated with less churn. Below are the maps that show how much the variables 'SeniorCitizen', 'tenure', and 'MonthlyCharges' correlate with 'Churn'.



The three heatmaps reflect the correlations for the variables 'SeniorCitizen', 'tenure', and 'MonthlyCharges' with the target variable 'Churn'.

The variable 'SeniorCitizen' shows a weak positive correlation with churn (0.15), suggesting that customers over the age of 65 are slightly more likely to churn. However, since the value is low, this relationship is not very strong. In contrast, 'tenure' has a moderate negative correlation with churn, meaning that the longer a customer stays with the company, the less likely they are to leave. This highlights tenure as an important factor in reducing churn risk. Meanwhile, 'MonthlyCharges' shows a weak positive correlation (0.19), indicating that higher monthly bills may contribute to churn, though the effect is relatively small.



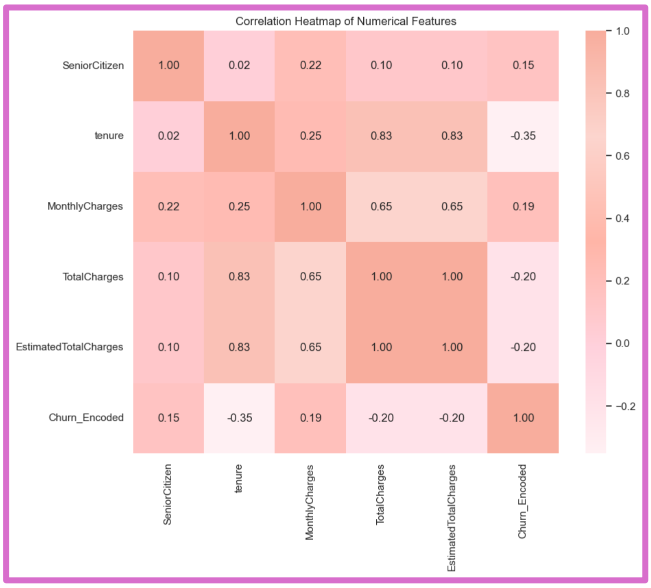
A comparative view of the two heatmaps reflects the correlations for the variables 'TotalCharges' and 'EstimatedTotalCharges' with the target variable 'Churn'.

The plots for 'TotalCharges' and 'EstimatedTotalCharges' are identical, as expected by now, and show a negative correlation with churn. This suggests that customers who have accumulated higher total charges over time are less likely to leave.

In summary, senior citizens are slightly more likely to churn compared to younger customers, while those with longer tenure or higher total charges are less likely to leave. On the other hand, higher monthly charges appear to increase churn risk, though the effect is relatively small.

6. Correlation Analysis

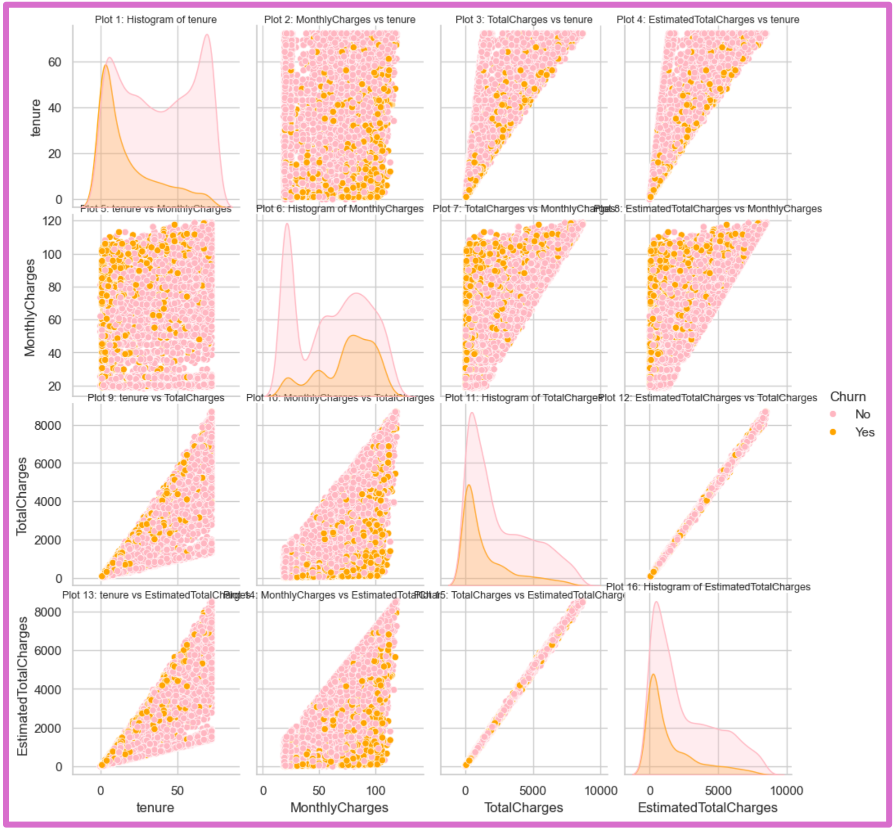
6.1. Correlation of Features



This heatmap highlights the correlation between the numerical features.

The correlation heatmaps reveal that 'tenure', 'TotalCharges', and 'EstimatedTotalCharges' are highly correlated (≈0.83 to 1.00), highlighting that longer-tenure customers naturally accumulate higher charges. This signals an overlap, and that one of these features might be redundant in modelling. In contrast, monthly charges, moderately correlated with 'TotalCharges', offer a more distinct signal as they reflect current billing patterns rather than historical accumulation. Weak correlations are seen between 'SeniorCitizen' and most variables, suggesting that age status is relatively independent. Overall, tenure and spending-related features overlap substantially, while monthly charges provide additional nuance. The most actionable signal here is that short-tenure, high-monthly-charge customers represent the group most vulnerable to churn.

6.2. Further Analysis of Correlations



The diagram gives a visual representation of the correlations between the numerical features via scatterplots and histograms.

The histogram of 'tenure' shows that churned customers are concentrated at low tenure values, which aligns with the heatmap, indicating that early customers are far more likely to leave. This suggests that feature engineering could help capture early churn risk more explicitly. The relationships between 'MonthlyCharges' and tenure reveal that customers with shorter tenure often face higher monthly charges, and these high upfront costs appear to increase churn risk. Similarly, plots involving 'TotalCharges' and 'EstimatedTotalCharges' show that churners tend to have low accumulated charges, simply because they leave early. While 'MonthlyCharges' highlights the immediate financial burden, 'TotalCharges' better reflects the overall customer lifespan. Together, these features suggest that combining monthly and total charges may provide stronger insights into early churn patterns. Since 'TotalCharges' is perfectly correlated with 'EstimatedTotalCharges'**,** engineering a refined feature could help capture these dynamics more effectively.

6.3. Segmented Analysis of Numerical Features

'tenure', 'MonthlyCharges', and 'TotalCharges' were segmented into the following groups, respectively: ([0-12], [13-24], [25-48], [49-72]), ([0-35], [36-70], [71-105], [106-140]), and ([0-500], [501-1500], [1501-3500], [3501-6000]). Customers with tenures between 0 to 12 months exhibited the highest churn rate, indicating that newer customers are more likely to leave. This could mean customers aren't fully satisfied or their expectations aren't met early on. It might also point to onboarding or early service issues, like if customers have a bad initial experience, they churn quickly. Alternatively, some customers might just be trying the service but don't find it valuable enough to stay. The company could focus on improving the first-year experience to reduce early churn, such as better onboarding, clearer communication, or targeted offers.

Customers with monthly charges in the $71 to &105 range showed the highest churn rates compared to other price groups. This suggests that customers paying in this mid-to-high price range might feel the service is not worth the cost, leading to them churning more frequently. It could also be that these customers expect better service or more features for the higher monthly charges, and if their expectations aren't met, they may decide to leave.

Another factor could be that customers in this price segment could be more tempted by competitors' similar or better services at a lower price, causing higher churn. Additionally, there is also the factor of billing surprise. If these charges are higher than they expected or variable, it might create dissatisfaction. With all these considerations, the company should investigate the needs and experiences of these customers and consider tailored retention strategies such as enhanced features, improved service quality, or targeted promotions to increase perceived value and loyalty.

For ‘TotalCharges’ it the analysis showed that lower total charges (0-500) tend to have a higher churn rate, and the churn rate decreases as total charges increase, suggesting that customers who have spent more overall are less likely to churn and that customers who have spent less total money with the company tend to leave more. This could be for a myriad of reasons, including them being less invested or engaged with the service. It could also be that higher total charges usually indicate longer tenure or more usage, so these customers might be more loyal or satisfied. Low total charges might also correspond to short tenure customers, who churn more frequently (which matches the first point. To curb this, it may be helpful to encourage customers to engage more or increase their usage to help improve retention.

7. Key Insights/Findings

7.1. Overall churn picture

- The target is imbalanced (≈26% churn vs ≈74% non-churn), so accuracy alone would be misleading without addressing class imbalance.

7.2. Numerical drivers

- ‘tenure’ is the single strongest protective factor: churn falls as tenure rises (moderate negative correlation). New customers (0–12 months) have the highest churn, confirming early-life risk.
- ‘MonthlyCharges’ shows a weak positive correlation with churn: risk increases with higher bills; the $71–$105 band is the most churn-prone among pricing tiers.
- ‘TotalCharges/EstimatedTotalCharges’ are strongly right-skewed and negatively associated with churn: customers who have accumulated more spend (typically longer tenure) churn less; these variables are highly collinear with tenure and with each other.
- ‘SeniorCitizen’ has only a weak positive correlation (~0.15) with churn: age status adds a limited incremental signal.

7.3. Categorical drivers

- ‘Contract’ type and ‘PaymentMethod’ are the strongest categorical signals: month-to-month customers churn far more (≈42.71%) than one-year (≈11.27%) and two-year (≈2.83%) contracts. As well as electronic check users (≈45.29%)
- ‘InternetService’ matters: fibre-optic users churn most (≈41.89%) vs DSL (≈18.96%) and no-internet (≈7.40%).
- Value-add support/security services are protective: not having ‘OnlineSecurity’, ‘OnlineBackup’, ‘DeviceProtection’, or ‘TechSupport’ roughly doubles churn versus having them (e.g., ‘TechSupport’: ~41.64% without vs ~15.17% with).
- Re-categorising phone and internet service features simplifies them, but sacrifices their interpretability, making strong signals less evident.
- Billing & payments show meaningful patterns: paperless billing users churn more (~33.57% vs ~16.33%); electronic check users churn higher than other payment methods. Grouping methods into “Automatic” vs “Manual” simplifies features but dilutes the strong electronic-check effect.
- Household context helps: customers without partners (~32.96%) or dependents (~31.28%) churn substantially more than those with partners (~19.66%) or dependents (~15.45%).
- Low signal features: ‘PhoneService’ and ‘MultipleLines’ show little churn separation; Gender differences are negligible (~0.76% gap).

7.4. Key relationships & vulnerable segments

- Short-tenure + high MonthlyCharges = highest risk. Early-lifecycle customers facing higher monthly bills are the most vulnerable to churn.
- A broader high-risk cluster combines month-to-month contracts, fibre-optic service, no security/backup/tech-support add-ons, paperless billing, and electronic check payments, especially within the first year.

7.5. Implications for modelling & action

- Modelling: address class imbalance; standardise features of differing scales; manage multicollinearity among ‘tenure’/’TotalCharges’/’EstimatedTotalCharges’ (retain one or regularise); keep both granular ‘PaymentMethod’ and the simplified grouping to test which yields better generalisation; consider tenure binning.
- Service-related features will be left as they are to avoid signal dilution and allow models to learn and capture which signals drive customer turnover.
- Business levers: focus retention on the first 12 months; price-packaging reviews for mid-to-high ‘MonthlyCharges’; promote/ bundle security/backup/tech-support; incentivise longer contracts; revisit paperless + electronic-check customer experience.

In short, longer-tenure customers are less likely to churn, while short-tenure, high-monthly-charge customers, often on flexible contracts without protective add-ons, form the core at-risk group.

8. Conclusion & Next Steps

This exploratory data analysis has highlighted clear churn drivers: short tenure, high monthly charges, month-to-month contracts, lack of support/security add-ons, and certain billing/payment practices. In contrast, longer-tenure customers with bundled services and stable contracts are far less likely to churn. These findings provide a strong foundation for the predictive modelling phase.

Next steps will focus on:

- Feature engineering: creating tenure bins, aggregating and simplifying payment/billing categories, and encoding service add-ons to capture churn risk more effectively.
- Data preparation: addressing class imbalance, handling multicollinearity among tenure-related variables, and scaling numerical features.
- Modelling: developing and comparing predictive models (e.g., logistic regression, tree-based methods) to estimate churn probability and prioritise interpretability.
- Domain exploration**:** engaging with business stakeholders to contextualise findings (e.g., reasons for dissatisfaction with fibre-optic plans or electronic check payments).

By combining robust feature engineering with churn-focused modelling and domain feedback, the analysis will evolve into actionable strategies to reduce attrition and improve customer retention.

9. References and Sources

[1] Y. Yean, "Telco Customer Churn - IBM Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset. [Accessed: 20-Aug-2025].

[2] S. Macko, "Telco customer churn," IBM Community Blog, Jul. 11, 2019. [Online]. Available: https://community.ibm.com/community/user/blogs/steven-macko/2019/07/11/telco-customer-churn-1113. [Accessed: 20-Aug-2025].