# What the EDA shows

* No duplicate IDs; no full-row dups (including ID). After removing customerID, there are 42 full-row duplicates across features, expected in real life when multiple customers share identical attributes. We’ll de-duplicate after dropping ID since duplicated data won’t offer additional information and might affect findings.
* Hidden missingness is confined to TotalCharges: 11 blanks/whitespace. These rows all have tenure == 0, and no rows have MonthlyCharges == 0. We’ll drop these 11 rows as “not yet billable.” This is not indicative of data errors since tenure 0 aligns with no monthly charges. However, those customers are not relevant to our research.
* Churn prevalence ≈ 26.5%, so the problem is imbalanced but not extreme. We can tackle this problem in data sampling and choosing which metrics to tackle (beyond just accuracy)
* Contract & billing patterns line up with prior knowledge: Month-to-month and Electronic check display the highest churn rates; Two-year and automatic payments the lowest. This is expected correlation.
* Contract type similar to the logic above where longer contracts correlate to low churn rates.
* Fiber optic churn is higher than DSL, and lack of OnlineSecurity/TechSupport correlates with higher churn. These will be strong features.
* Features with little to no effect: MultipleLines, PhoneService, Gender. (Also redundancy between MultipleLines and PhoneService)

# Data types to enforce

Numeric:

* tenure: int32
* MonthlyCharges: float32
* (Temporary) TotalCharges: float32 to compute avg\_monthly\_spend, then drop raw TotalCharges

Binary:

* Binary categorical (0/1): Churn (1 = churn “Yes”), Partner, Dependents, PaperlessBilling, SeniorCitizen.

Multi-category:

* InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaymentMethod
  + Keep tri-levels for the add-ons (Yes / No / No internet service).

# Feature engineering

Columns to add:

* avg\_monthly\_spend = TotalCharges / tenure (dropping tenure==0 rows makes this safe)
* streaming\_count = 1\*(StreamingTV=="Yes") + 1\*(StreamingMovies=="Yes") → {0,1,2}
* security\_support\_count = Yes-count over {OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport}

Columns to drop:

* customerID
* Gender
* PhoneService
* MultipleLines
* TotalCharges(raw)