# Pre-Analysis Notes

1. **Customer ID (string):** unique identifier for the subscriber. Check for duplicates.
2. **Churn Label (string):** Yes/No label that identifies whether the subscriber has cancelled the subscription or not.
3. **Churned (boolean):** 1 if the subscriber cancelled subscription, 0 if the subscriber continuous subscribed. Check if it coincides with the previous field Churn Label. If it’s the same, we can delete Churn Label.
4. **Account Length (in months) (integer):** duration of the current/last account of the subscriber.
5. **Customer Service Calls (integer):** number of calls to the customer service.
6. **Avg Monthly GB Download (integer):** average of monthly gigabytes downloaded. Amazon Prime enables download of films and episodes to the subscriber’s phone or tablet.
7. **Unlimited Data Plan (Boolean):** 1 stands for an unlimited data plan subscription and 0 for limited data plan subscription. Check this out.
8. **Extra data charges (integer):** I think this means extra payments for increasing data storage (in USD). Check this out.
9. **State (string):** US state abbreviation. It may be interesting to create a new column for the complete name of each state. Ask ChatGPT to associate state with abbreviation?
10. **Phone Number(string):** subscriber phone number. Not interesting for the analysis.
11. **Gender (string):** gender of the subscriber.
12. **Age (integer):** age of the subscriber. Maybe interesting to segment dataset on different groups of ages according to this field.
13. **Under30 (string):** Yes/No label that identifies whether the subscriber is under 30 years old or not.
14. **Senior (string):** Yes/No label that identifies whether the subscriber is 65 years old or more or not.
15. **Group (string):** Yes/No label that identifies whether the subscriber is included in a family subscription plan or not.
16. **Number of Customers in Group (integer):** number of subscriber included in the family subscription plan.
17. **Device Protection & Online Backup (string):** Yes/No label that identifies whether the subscription includes device protection and online backup or not.
18. **Contract Type (string):** type of contract according to the subscription period (month-to-month, one year, two year, etc.).
19. **Payment Method (string):** type of payment used by the subscriber for paying the last subscription period. Notice that there are several null values and inconsistencies between upper and lower-case values.
20. **Monthly Charge (integer):** monthly payment of each subscriber in USD. Notice inconsistencies regarding money suffix (USD included only for certain values). Notice null values.
21. **Total Charges (integer):** total money that has been charged for each subscriber since the beginning of the subscription period. Notice null values. Maybe null values could be calculated using other column values.
22. **Churn category (string):** category of the reason for subscription cancellation. Notice null values.
23. **Churn Reason (string):** reason for subscription cancellation. Notice null values.
24. **Contact Date (date):** last contact date. Notice the format: YYYY/MM/DD for some values and YYYY-MM-DD for other values.
25. **Last Transaction Date (date):** date of the last payment. Notice format DD/MM/YYYY which is different from previous field Contact Date.
26. **Customer Tenure (in months) (integer):** do not know exactly the meaning of this field. I suppose is the time since the first subscription. Need to investigate.
27. **Customer Segment (string):** some kind of customer level classification (high, medium or low). Do not know exactly what means. Need to investigate.
28. **Average Monthly Expenses (float):** average monthly expenses of each subscriber. So much decimal values!
29. **Number of Complaints or Support (integer**): number of times the subscriber has emitted a complaint or support request.
30. **Preferred Contact Method (string):** preferred contact method of each subscriber.
31. **Applied Discount (string):** True/False field that indicates whether a discount has been applied (VERDADERO) or not (FALSO).
32. **Favorite TV show (string):** favorite TV show of each subscriber.
33. **Internal Notes (string):** internal notes about each subscriber.

INTERESTING COLUMNS FOR THE ANALYSIS:

* **Churn label, Churn category, Churn Reason**: the key variables of the analysis.
* **Account length**: to analyse whether the cancellations are from old subscribers or new ones.
* **Customer service calls**: to analyse whether the cancellations are related with the frequency of customer service calls.
* **Number of complaints or Support**: to analyse whether the cancellations are related with the number of complaints or support.
* **State**: to analyse churn rates per US state. Maybe we find some geographical patterns.
* **Gender, Age, Under30, Senior**: interesting columns for performing dataset segmentation and analysing churn rate per groups.
* **Group**: to analyse the success or failure of the family plans.
* **Contract type**: to analyse which contract type favours customer retention. We believe it is the year contract.
* **Contact date, Last transaction date**: to analyse whether subscribers leave the subscription soon after the last contact with Amazon (it would give us hints about the nature of those last contacts: complaints, support requests, etc.) or soon before/after the payment.