- State: The state where a customer comes from

- Account length: Number of days a customer has been using services

- Area code: The area where a customer comes from

- Phone number: The phone number of a customer

- International plan: The status of customer international plan

- Voicemail plan: The status of customer voicemail plan

- No. vmail msgs: Number of voicemail message sent by a customer

- Total day minutes: Total call minutes spent by a customer during day time

- Total day calls: Total number of calls made by a customer during day time

- Total day charge: Total amount charged to a customer during day time

- Total eve minutes: Total call minutes spent by a customer during evening time

- Total eve calls: Total number of calls made by a customer during evening time

- Total eve charge: Total amount charged to a customer during evening time

- Total night minutes: Total call minutes spent by a customer during night time

- Total night calls: Total number of calls made by a customer during night time

- Total night charge: Total amount charged to a customer during night time

- Total intl minutes: Total international call minutes spent by a customer

- Total intl calls: Total number of international calls made by a customer

- Total int charge: Total international call amount charged to a customer

- Customer service calls: Total number of customer service calls made by a customer

- Churn: Whether a customer is churned or not

Most phone numbers belong to Wyoming (WY) state.

5. 3010 out of 3333 (90%) people don't have international plan and 2411 out of 3333 (72%) don't have voice mail plan.

6. Majority of columns have outliers. We need to handle them so that certain models don't get affected by outliers.

7. We have highly imbalanced dataset: 85.5% customers didn't Churn while only 14.5% did. Thus:

- This is a **\*\* classification \*\*** problem.

- Use model evaluation criteria other than accuracy (since randomly stating no customer will churn will give us 85.5% accuracy). Metrics like precision, recall, F1-score, ROC-AUC curve will become important.

- If identifying churned customers is important, make sure we perform oversampling/undersampling so that our model learns the minority class as well. Not doing this will lead to our model getting biased towards majority class.

Handling Missing Data: **\*\*NOT\*\*** Required

- Handling Duplicate Data: **\*\*NOT\*\*** Required

- Correct Data Types

- Drop features if required