**Iranian Churn Data Analysis Proposal**

Tim Pulfer, Jacob Rosen, Dhun Sheth

**Introduction**

The project will aim to analyze churn data from an Iranian telecommunications company. The goal is to understand the relationship between customer interactions/profiles to the rate at which customers churn. By modeling this relationship, the company can then predict the expected reduction in the number of customers that churn based on some reduction in call failures or complaints.

**Statistical Description**

The dataset we are using is publicly available from the UC Irvine Machine Learning repository. Data was randomly collected from an Iranian telecom company database over a period of 12 months in 2020. The database has a total of **3150 rows**, each representing a customer. The following are the **13 columns** in the database:

* Call Failures (numerical): the number of call failures over 9 month period
* Complains (binary): 0: No complaint, 1: complaint
* Subscription Length (numerical): total months of subscription
* Charge Amount (ordinal): 0: lowest amount, 9: highest amount
* Seconds of Use (numerical): total seconds of calls over 9 month period
* Frequency of use (numerical): total number of calls over 9 month period
* Frequency of SMS (numerical): total number of text messages over 9 month period
* Distinct Called Numbers (numerical): total number of distinct phone calls
* Age Group (ordinal): 1: younger age, 5: older age (10-19 is 1, 20-29 is 2, etc.)
* Tariff Plan (binary): 1: Pay as you go, 2: contractual
* Status (binary): 1: active, 2: non-active
* Customer Value (numerical): The calculated value of customer 9 month period
* Churn (binary): 1: churn, 0: non-churn - the state of the customers at the end of 12 months

All of the attributes except for churn is aggregated data of the first 9 months. The churn labels are the state of the customers at the end of 12 months. The three months is the designated planning gap.

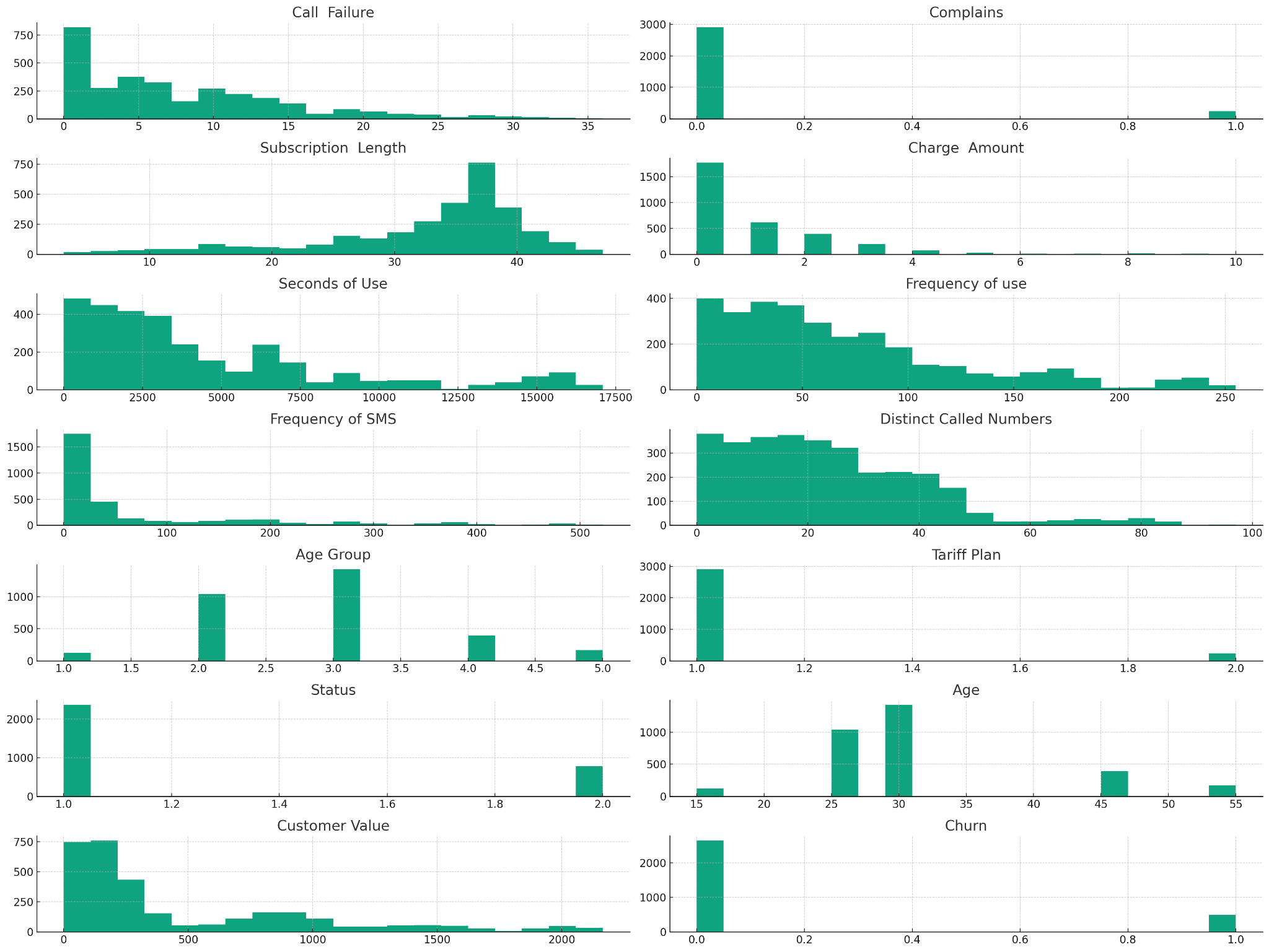
**Potential Supplementary Analysis**

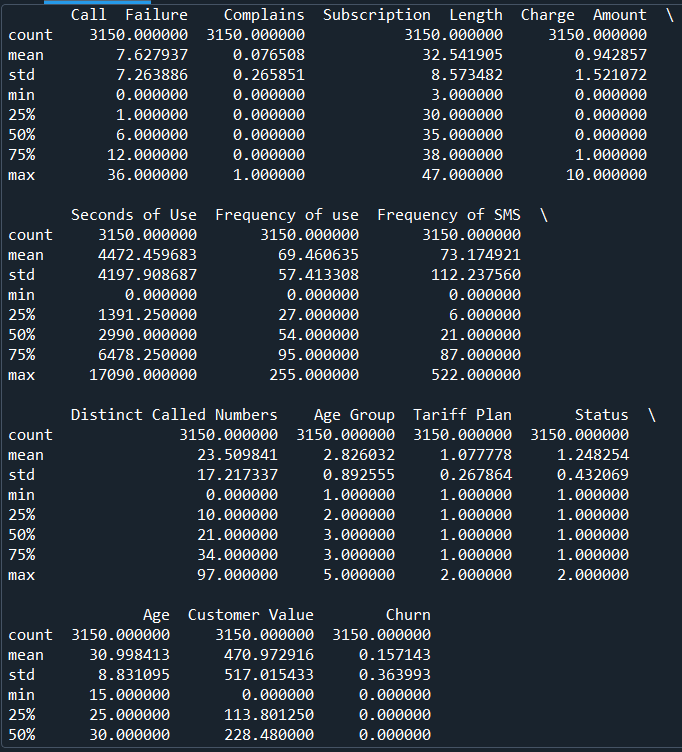
Additional questions to be answered include:

1. Is there any collinearity between predictors?
2. What is the expected reduction in churn if a proposed business solution can reduce call failures by 50%?
3. Which predictors dominate churn for high-value customers?

**Dataset Reference:**

[1] Iranian Churn Dataset. (2020). UCI Machine Learning Repository. <https://doi.org/10.24432/C5JW3Z>.





**Appendix:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Call Failure | Complains | Subscription Length | Charge Amount | Seconds of Use | Frequency of use | Frequency of SMS | Distinct Called Numbers | Age Group | Tariff Plan | Status | Age | Customer Value | Churn |
| 8 | 0 | 38 | 0 | 4370 | 71 | 5 | 17 | 3 | 1 | 1 | 30 | 197.64 | 0 |
| 0 | 0 | 39 | 0 | 318 | 5 | 7 | 4 | 2 | 1 | 2 | 25 | 46.035 | 0 |
| 10 | 0 | 37 | 0 | 2453 | 60 | 359 | 24 | 3 | 1 | 1 | 30 | 1536.52 | 0 |
| 10 | 0 | 38 | 0 | 4198 | 66 | 1 | 35 | 1 | 1 | 1 | 15 | 240.02 | 0 |
| 3 | 0 | 38 | 0 | 2393 | 58 | 2 | 33 | 1 | 1 | 1 | 15 | 145.805 | 0 |
| 11 | 0 | 38 | 1 | 3775 | 82 | 32 | 28 | 3 | 1 | 1 | 30 | 282.28 | 0 |
| 4 | 0 | 38 | 0 | 2360 | 39 | 285 | 18 | 3 | 1 | 1 | 30 | 1235.96 | 0 |
| 13 | 0 | 37 | 2 | 9115 | 121 | 144 | 43 | 3 | 1 | 1 | 30 | 945.44 | 0 |
| 7 | 0 | 38 | 0 | 13773 | 169 | 0 | 44 | 3 | 1 | 1 | 30 | 557.68 | 0 |
| 7 | 0 | 38 | 1 | 4515 | 83 | 2 | 25 | 3 | 1 | 1 | 30 | 191.92 | 0 |
| 6 | 0 | 38 | 0 | 5918 | 95 | 7 | 12 | 3 | 1 | 1 | 30 | 268.52 | 0 |
| 9 | 0 | 38 | 0 | 2238 | 54 | 8 | 17 | 3 | 1 | 2 | 30 | 123.68 | 0 |
| 25 | 0 | 38 | 3 | 15140 | 225 | 54 | 32 | 3 | 1 | 1 | 30 | 830.6 | 0 |
| 4 | 0 | 38 | 1 | 3095 | 27 | 483 | 8 | 3 | 1 | 1 | 30 | 2056.88 | 0 |
| 9 | 0 | 37 | 0 | 15485 | 182 | 150 | 30 | 2 | 1 | 1 | 25 | 1380.015 | 0 |