Suppose a telecom company called ‘TeleX’ is the leading telecommunications provider in the country. The company earns most of its revenue by providing internet services. Based on the past and current customer information, the company has maintained a database of personal/demographic information, the services availed by a customer, and the expense information related to each customer. The missing values in the data sets are a result of data entry errors or other random errors.

As a data analyst, your aim is to automate the process of predicting whether a customer would churn (i.e. leave TeleX) and to find the factors affecting the churn. The collated data dictionary for the variables in the 3 data frames is given

Use the data quality checklist provided to identify the data quality issues in each file

**Q1**

**Look through the Customer dataset. Which of the following quality issues are present?**

1. **Lowercase and uppercase issues**
2. **CorrectYou missed this!**
3. **It contains a summary row**
4. **It contains NA values**
5. **It contains duplicated rows**
6. **It contains blank values, not interpreted as NA**
7. **It contains unnecessary header row(s)**

**Q2**

**How many duplicate rows are there in customer file?**

**Hint: Let's say there are 3 completely identical rows in a data set, i.e. the corresponding column values of these rows are exactly the same. Here, there are two duplicate rows or two rows which are an exact copy of the original row.**

**4**

**closeIncorrectdoneCorrect**

**2**

**Feedback :**

**There are 2 duplicated rows in the Customer dataset**