**AZ Disconnections Database**

1. **Dataset extraction**

Dataset provided was in PDF format and only the useful tables were to be extracted from it. We used PDF tools to extract the raw data directly into an excel sheet. The extracted data was disjoint and needed to be properly joined on various join conditions according to the table format. We used Python script to join the tables.

1. **Data cleaning**

The only data cleaning required was to replace the incorrectly parsed values with correct numbers. For example, “9” used to get parsed as “g” at some places which needed replacing. From the zipcode data, some unnecessary fields like type, timezone, area code, etc.

1. **Dataset normalization**

After data extraction we were left with two sheets, one with disconnections data and other with zipcode data.

Using these tables, we first identified the entities and built an ERD from it. This gave use the perspective on how to normalize the dataset.

1. Created a table with different companies and their details like address and gave a unique id to each of the company.
2. The company field with company id in disconnections dataset.
3. Multiple sheets for each company were created after extraction and it had each month as separate field. These tables were pivoted to create just one month field and a year field.
4. These multiple sheets were appended to create a single sheet for all the companies.
5. The zipcode table was further normalized my creating a table for cities with unique ids and a table for counties with their unique ids.
6. All the normalization was done using Python script (pandas).