**Oregon Utility Disconnections   
Data Normalization**

**Step 1:** We have a database with zip code details for the entire United States. Our specific requirement is to extract information solely related to the state of Oregon (OR). Therefore, our initial step involves filtering and isolating the data specific to Oregon. We use filter option on state to filter out the Oregon state Data.

A screenshot of a computer

Description automatically generated

**Step 2:** Next, we'll review all the details in the zip code database and get rid of any columns that we don't need or that aren't relevant.

A screenshot of a data

Description automatically generated

We have removed the following columns, which doesn’t serve much purpose as we are solely focusing on Oregon state, we don’t need country and state.

* acceptable\_cities
* unacceptable\_cities
* state
* timezone
* area\_codes
* world\_region
* country

After removing the irrelevant columns, we have the data which needs to be normalized.   
  
**Step 3:** From the raw data, will create a new data sheet called “County”. We have 2 columns in “County” table.

1. County\_ID
2. County\_Name

* Copy all the data into “county” column from the zipcode Database to County\_Name in the new datasheet “county”.
* To eliminate redundancy, we need to remove the duplicates present in the County\_Name.
* We can click on Data tab -> Remove Duplicates.
* Fill the County\_ID column, enter the value 1 in the first row and drag down to auto increment the value.
* County\_ID will be our primary key here.

A table of names and numbers

Description automatically generated with medium confidence

**Step 4:** Now create another new sheet called “City”. We have 3 columns in City table.

1. City\_ID
2. City\_Name
3. County\_ID\_FK

* First, we need to copy the City\_Name, County\_Name data from zipcode database.
* In order to get County\_ID\_FK, we will need County\_Name field as well in our datasheet.
* To remove redundancy, we will delete the duplicates.
* We can click on Data tab -> Remove Duplicates.
* Fill the City\_ID column, enter the value 1 in the first row and drag down to auto increment the value.
* To get the County\_ID data, we use vlookup. To do this, we need to paste the County\_ID and County\_Name in any column beside the data.
* Click on the first row of County\_ID\_FK, and paste the following vlookup formula

**Function f(x) = VLOOKUP($C2,$I$2:$J$38,2,FALSE)**  
  
A screenshot of a table

Description automatically generated  
  
  
A screenshot of a table

Description automatically generated  
  
**Step 5:** Now create another new sheet called “Address”. We have 7 columns in Address table.

1. Zipcode
2. Latitude
3. Longitude
4. Population
5. City\_ID\_FK
6. Type\_ID\_FK
7. Decommissioned

* Copy the data Zipcode, Latitude, Longitude, Population and Decomissioned data from zipcode database.
* To get City\_ID\_FK and Type\_ID\_FK, we need to initially copy “City\_Name” and “type” from zipcode database.
* To remove redundancy from type, lets remove the duplicates.
* After removing the duplicates we will get only 3 rows as follows

A close-up of a type of item

Description automatically generated

To get the City\_ID\_FK data, we use vlookup. To do this, we need to paste the City\_ID and City\_Name in any column beside the data as follows:

A screenshot of a table

Description automatically generated  
  
A screenshot of a computer

Description automatically generated

**Function f(x) = VLOOKUP($H2,$K$2:$L$382, 2, FALSE)**

Similarly, to get the Type\_ID\_FK, we need to paste the Type\_ID and Type\_Name in any column beside the data as follows:

A screenshot of a computer

Description automatically generated  
  
A screenshot of a computer

Description automatically generated  
  
**Function f(x) = VLOOKUP($I2,$N$2:$O$4, 2, FALSE)**

Once, we get the City\_ID\_FK and Type\_ID\_FK, we can save the data in a new datasheet with the values and save this datasheet as “Address”.  
 **Step 6:** Create a new data sheet and name it as “Month”. It contains two columns.

1. Month\_ID
2. Month\_Name

* To get the values for Month\_ID and Month\_Name. Enter 1 in Month\_ID and drag it down to auto fill the series.
* Similarly, enter January in the first row of Month\_Name and drag it down to auto fill the series. Save the datasheet with the values as follows:  
    
   **A screenshot of a calendar

  Description automatically generated**

**Step 7:** Create a new data sheet and name it as “Company”. It contains two columns.

1. Company\_ID
2. Company\_Name

* To get the values for Company\_ID and Company\_Name.
* We have 6 companies in total, we enter the 6 company names manually under Company\_name column.
* Similarly, we enter the respective Company\_ID’s for respective Company\_name as below.
* Save the datasheet with the values as follows:

**A screenshot of a computer

Description automatically generated**

**Step 8:** Data extraction from PDF files  
  
We have created a new datasheet called “Disconnections” to get the data from pdf files.  
Disconnections table contains

1. Disconnection\_ID
2. Company\_ID
3. Zipcode
4. Month\_ID
5. Year
6. Number\_of\_disconnections

In order to get the above data. We have downloaded all the pdf files related to 6 companies.   
  
**A screenshot of a computer

Description automatically generated**

* Each company folder contains their respective pdf files.
* For companies like Avista, Cascade, Idaho, Pacific and Portland General Electric Company, we have used tools like Adobe Acrobat, ilovepdf, smallpdf to convert the pdf file into excel.
* Once the file is converted into excel, we have saved them.   
    
  A screenshot of a computer

  Description automatically generated
* Using these tools, we were able to get the data in pdf to excel.
* We normalized the data by deleting the irrelevant fields and just copied the data which is required.
* If there are any empty cells, we replaced them with 0.  
    
  **Sample snippet of disconnections table:**   
    
  A screenshot of a table

  Description automatically generated

**Failed Attempts:**We faced an issue while extracting the data from NW Naturals using the tools.   
  
**Sample snippet of NW Naturals data:**  
  
**A table with numbers and text

Description automatically generated**

* The above pdf format is not properly converting to excel. When we tried to convert it, we were losing and missing parts of the data.
* To overcome this issue, we have edited the pdf and removed the table, which is an image and got only the data in text.
* Then, we used Adobe Acrobat to convert the data into Excel.
* Using this approach, we were able to extract the data without any data loss.