**Homework 10: Cleaning and Wrangling Data with OpenRefine**

**Part 1: Cleaning and Wrangling Data**

**Objective: To discuss the approach and techniques for data wrangling**

**Data Cleaning and Wrangling:** The process of preparing/transforming the raw data into more meaningful/accurate data so that we can efficiently draw analytics out of it.

**Why we need Data Wrangling:**

* Most often the raw/source data is not clean and may contain missing, null and erroneous data.
* This data can not be used as is to make any meaningful decisions.
* If the data is used as is, then it might not lead to accurate insights/predictions.

**Activities involved in Data Wrangling:**

* **Missing values removal:** Missing values for the key features (columns) can impact the analysis hence, it is important to analyse why the values are missing and remove them if it does not make sense.
* **Reformatting Features:** Sometimesthe features of interest may not have an appropriate feature. It is important that we reformat it to be able to use it efficiently.
* **Removing unwanted columns:** A huge dataset may contain hundreds of columns. We need to analyse the columns and get rid of the unwanted ones as they may not be helpful to solve the given problem but unnecessarily increase the data and processing resources.
* **Enriching Data:** Sometimes we may not have a meaningful data in the given columns but we can combine a few columns to get a meaningful insight. This is referred as enriching data. In some cases, we might also need to join external data sources to enrich the data.
  + **Example:** Joiningthe locations dataset to identify the area based on ZipCodes.
* **Assuring data consistency:**  Once we have an enriched data, we need to confirm the consistency of the data. This step may need some level of domain knowledge but it is essential to be assured of the data consistency to move forward.
* **Availability of the data:** Once each of the above activities are performed and we are confident about the data, it is time to make the data available for analysis. This essentially means that we need to publish the data so that others like data scientists and analysts can access it for further analysis and predictions. This is considered as the output of the Data Wrangling process.

**Conclusion:** Data Wrangling is an important step in data preparation and it helps to restructure and reduce the data to the required format which can be further used to draw meaningful outcomes/analytics.

**Task 2: Exploring, Cleaning and Wrangling datasets with OpenRefine**

**Objective:** To use OpenRefine to perform data wrangling on a dataset.

**About OpenRefine:** it is an opensource web-based tool (works on a browser) which can be used to perform various steps and activities involved in data wrangling to get a clean dataset out of the raw dataset. It runs on a browser but the data does not get loaded to any other external system hence, the data security is not compromised.

**Step 1: Installation**

* Go to <https://openrefine.org/> and click on the Download button.
* A zipped folder “openrefine-win-with-java-3.7.9 “ gets downloaded to your machine.
* Now, unzip the folder and double click on the openrefine.exe file.
* It opens a new command window as shown below.

A screen shot of a computer screen

Description automatically generated

* Also a new tab is open in your browser as shown below.

A screenshot of a computer

Description automatically generated

**Step 2: Upload Data and Create Project**

* Download the “Consumer\_Complaints.csv” file from Canvas
* Go to the Openrefine web browser.

A screenshot of a computer

Description automatically generated

* Click on the “Choose file” button and select the file and click on Next.

A screenshot of a computer

Description automatically generated

* It detects the file format, null and blanks etc.
* If you wish to have specific name for your project, update the project name and click on Create Project.
* The project is now created, and the below screen gets displayed.

A screenshot of a computer

Description automatically generated

* Now, it shows the number of records at the top left corner, we can select the number of records to be displayed on the page and we can use the “next” button to go to the next page of the records.

**Step 3: Check states with Text Facets**

* From the State column, expand the dropdown and select Facet🡪Text Facet. The below screen gets loaded.

A screenshot of a computer

Description automatically generated

* In the above screenshot, we can see the unique values for the State column and the number of records having that unique value.
* This view can help us to understand that from which states we have how many customer complaints and identify the highest and the lowest.

**Step 4: Wrangling/Munging – Transforming data**

1. **Text Facet for Zip code**

* When we go to Zip code🡪dropdown 🡪 Facet 🡪 Text Facet, the below screen gets displayed.

**A screenshot of a computer

Description automatically generated**

* Click on “Set choice count limit”

A screenshot of a computer

Description automatically generated

* In the above screen, we can see that there are some values with 1 digit, 2 digit, 3 digit etc. This is not appropriate for Zip code and it means that the zip codes are not consistent and we need to make it consistent.

1. **Numeric Facet for Zip Code**

* Select Zip code 🡪dropdown 🡪 facet🡪 numeric facet
* It displays the below screen with a message that “No numeric values present”.

**A screenshot of a computer

Description automatically generated**

* This means that the Zip Code field has the data type as text and not numeric which can be converted to numeric for better analysis.

1. **Transforming Zip code from text to numeric**

**Data Wrangling:** Data wrangling is the process of making the raw data ready to be used for analysis by performing a few steps which may include:

* Removing missing values
* Identifying and discarding outliers(if required)
* Remove duplicate data
* Reformat the columns(if required)
* Drop the unwanted columns.
* Create derived columns(if required)
* Make the cleaned and processed dataset available for the data analyst/Scientist)

**Converting text field to numeric field**

* Select Zip code 🡪Edit Cell🡪Common Transforms🡪To number
* The below screen gets loaded.

**A screenshot of a computer

Description automatically generated**

* The bottom left corner shows that the 380136 zip codes are converted to numeric and there are 4362 blank values.

**Step 5: Handling missing zip codes**

1. **Fill the missing value :** Use the above value to fill the blank cell.

* Select Zip code🡪edit cells🡪Fill Down

**A screenshot of a computer

Description automatically generated**

* The result shows that there are no missing values now.

1. **Rescind the Wrangling Data:** We can undo the action to revert it back to the previous state.

* Click on the Undo/Redo tab and click on the step 1.

**A screenshot of a computer

Description automatically generated**

* Now the blank values are restored for Zip code.

1. **Handling missing zip codes, create a new column**

* Select Zip code🡪 edit column🡪 add column based on this column.

**A screenshot of a computer

Description automatically generated**

* Fill the below details and click OK.
  + New column name: Zipcode5
  + if(value.length() > 4, value, “99999”)

**A screenshot of a computer

Description automatically generated**

* The below screenshot shows that the blanks are replaced with ‘99999’ in the newly created Zipcode5 field.

**A screenshot of a computer

Description automatically generated**

* Go to the Undo/Redo tab to revert the last action which would delete the newly created column.

**A screenshot of a computer

Description automatically generated**

* We can clearly see in the above screen that the Zipcode5 filed is not present anymore.
* Again, recreate the Zipcode5 filed by repeating the previous steps.
* The below screenshot displays the recreated Zipcode5 column.

**A screenshot of a computer

Description automatically generated**

**Step 6: Getting Started:**

1. **Start a new project**

* Repeat “Step 2: Upload Data and Create Project“ and select eq2015.csv and Project name as “Earthquake 2015”.

A screenshot of a computer

Description automatically generated

1. **Review the column “nst”**

* We can see that there are 8061 blank cells for nst and the total number of rows in the dataset are 8708.
* So this column has a majority of the blanks. We can think about ignoring this field.

**Step 7. Wrangling the Place column**

1. **Overview:** The place column has coma separated values and we can get the state and country as values and store them in a separate column.
2. **Data Transformation:**

* Select Place 🡪 Edit Column 🡪 Add Column based on this column
* Enter the values as:
  + New column: Location
  + Expression: value.split(‘,’)[1]
* The new filed Location is created and displays the last part of the coma separated value.

A screenshot of a computer

Description automatically generated

* We identify that some rows have blank data as there is no coma for the parent field.
* We need to rewrite the expression to deal with these kinds of cells.
* We rewrite the expression as per the below screenshot.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* The above screenshot shows the filed created with the new expression.

**Step 8: Cleaning the Location column**

1. **Exploring the Location Column:**

* The data is not consistent in the Location Column and the same value is spelled differently in different cells.
* This would create discrepancy in the analytics if not corrected.
* Select Location🡪Facet🡪Text Facet
* Note: I have my Location column named as Location1

**A screenshot of a computer

Description automatically generated**

**Step 9: Wrangling data by Clustering**

1. **Description:** With this functionality we can group the similar string or patterns and assign them a common value. This will solve the inconsistency problem to a greater extent. While using the clustering functionality we decide to make the feature conservative, so it doesn't group the unwanted entries in one cluster.
2. **Clustering the Location data**

* There are different options to get the clustering done.
* The below screenshot shows some of them.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated** **A screenshot of a computer

Description automatically generated**

**Step 10: Editing Cells**

1. **Overview:** Now we have an idea how we can use clustering functionality to correct misspellings and get a clean data.
2. **Editing Cells Using Key Collision Method & Metaphone3 Keying Function**

* Select Key Collision and Metaphone3 , select the clustering and click on “Merger Selected & Recentre” button.

A screenshot of a computer

Description automatically generated

* Now, the data is clustered. In the below screenshot, we can see that the two ticked clusters are not showing anymore.

A screenshot of a computer

Description automatically generated

* Now we can verify if the clustering is complete.
* We can see that there are still some typos for Alaska and California

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Step 11: Manually Editing the cell**

1. **Description:** In some scenarios, we might need to edit the cells manually.

* Though it is not a best practice to do manual updates, it can sometimes be the only option left depending on the data.

1. **Edit Cell**

* Select the column 🡪 Text Filter
* In the left side pane, enter “Alaka”
* Now it shows the row with that data.
* Verify the details from the aren’t column before editing it.
* Once verified, select the cell and click on edit, enter “Alaska” and click “Apply” as shown in the below screenshot.

**A screenshot of a computer

Description automatically generated**

* Once you click apply, the value gets updated. Now we can not see any value with “Alaka”.

A screenshot of a computer

Description automatically generated