Group Name:

Team Member Details:

- Corey Hamren, [hamhams86@gmail.com](mailto:hamhams86@gmail.com), USA, Self-teaching, Data Science

- Motamen MohammedAhmed, [Motamen.salih@hotmail.com](mailto:Motamen.salih@hotmail.com), UAE, Zayed University, Data Science

- Francis Kim, [fkim39@gmail.com](mailto:fkim39@gmail.com), [Fkim39@gmail.com](mailto:Fkim39@gmail.com), USA, University of Maryland, Data Science

Problem Description:

There are different diseases that affect people around the world. Pharmaceutical companies started to manufacture cures to defeat them. One of the challenges for these companies is to understand the persistency of drug as per the physician prescription and the related factors. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

Business Understanding:

Project Life-cycle and Deadline:

There are different diseases that affect people around the world. Pharmaceutical companies started to manufacture cures to defeat them. One of the challenges for these companies is to understand the persistency of drug as per the physician prescription and the related factors. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

Data Intake Report:

Name: Medicare Group Project

Report date: 09/29/2021

Internship Batch: LISUM03

Version: 1.0

Data intake by: Team Name

Data intake reviewer:

Data storage location: https://github.com/TheHamhams/dg\_group\_project

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 3424 |
| **Total number of files** | 1 |
| **Total number of features** | 69 |
| **Base format of the file** | .csv |
| **Size of the data** | 0.88 MB |

**Proposed Approach:**

* Mention approach of dedup validation (identification)
* Mention your assumptions (if you assume any other thing for data quality analysis)

In terms of dedup validation, we will used Excel’s duplicate identification feature to identify and remove duplicate addresses (if any). In addition, null values can be identified in either Excel or Python and removed. However, we found the column value: “Unknown” to be useful and not the same of a null value. The reason being is that column value “Unknown” is an important count for analyzing the distribution count of data we do have Vs. the unknown values. Dataset will also be cleaned through the lowering of column titles in Python.

GitHub Repository Link:

- <https://github.com/TheHamhams/dg_group_project>