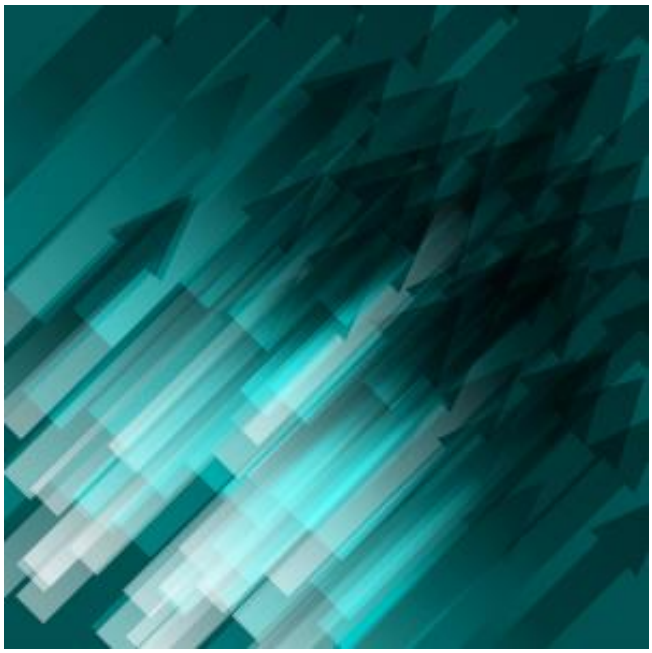


QUANTIC

an analytics company

WEEK 9 - DELIVERABLES

This project addresses a crucial organizational challenge, aiming to ensure alignment with business objectives. Guided by a well-structured project lifecycle with a defined deadline, this initiative incorporates a comprehensive data cleaning and preparation process. A rigorous data intake report is employed to safeguard the quality and relevance of the data, underpinning the project's success, and enhancing its capability to deliver actionable insights.

Ansel Vallejo | Data Scientist
LISUM25

QUANTIC

Table of Contents

About Us

Quantic is an analytics company that places a strong emphasis on healthcare. We are dedicated to the idea that data can be a catalyst for positive change in the healthcare industry. With a talented team of data scientists and analysts, our primary objective is to tackle complex healthcare challenges and enhance patient outcomes. Our distinctive approach combines state-of-the-art data analytics with deep healthcare sector knowledge to deliver actionable insights, fostering informed decisions and meaningful advancements in healthcare provision. At Quantic, we are committed to a future where healthcare is not only data-driven but also healthier and more efficient.

Our Team

Name	Email	Country	Institution	Specialization
Ansel Vallejo	msavg@hotmail.com	Japan	Flatiron School	Data Science

Overview

One of the persistent challenges faced by pharmaceutical companies lies in comprehending the duration of drug persistence as per physician prescriptions. To solve this problem, ABC Pharma Company recognized this issue and engaged Quantic to streamline and automate the identification process. By leveraging data analytics, the pharmaceutical company aimed to gain valuable insights into drug persistency patterns, ultimately enhancing their decision-making and ensuring better patient care. The collaboration between ABC Pharma and Quantic demonstrates a commitment to harnessing data-driven solutions to address critical industry challenges. Through this initiative, they strive to advance pharmaceutical practices and optimize patient outcomes.

Business Scope

The project scope entails the development of an automated system in collaboration with Quantic to analyze and identify drug persistency patterns within the pharmaceutical domain. This data-driven solution will enhance decision-making for ABC Pharma Company, ultimately leading to improved patient care and the advancement of pharmaceutical practices.

Data Intake Report

Name: *Healthcare_dataset*
Report date: *October 26th 2023*
Internship Batch: LISUM25
Version:1.0
Data intake by: Ansel Vallejo
Data intake reviewer: N/A
Data storage location: N/A

Tabular data details:

City

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	899 KB

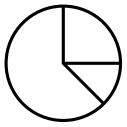
Proposed Approach:

- Check data for any missing values.
- Check for outliers.
- Check for skewed data.

Problem Description

The problem at hand pertains to the pharmaceutical industry's struggle to gain insights into drug persistency duration as influenced by physician prescriptions. It is crucial to comprehend the duration of patient adherence to prescribed drug regimens, as this knowledge is vital for enhancing treatment effectiveness, patient well-being, and overall pharmaceutical strategies. The key issue in this context lies in the absence of an efficient and automated system for identifying persistency patterns, hampered by data anomalies such as missing values (NA), outliers, and skewed distributions. Addressing these data quality concerns is paramount to improving the understanding of patient drug persistency and ultimately enhancing pharmaceutical practices.

Data Cleaning and Preparation



Scope:
The scope of data understanding encompasses a thorough examination and resolution of critical data issues, including missing values, outliers, skewed distributions, and the handling of categorical data, achieved through a comprehensive data cleaning and preparation process. By identifying, addressing, and mitigating these data challenges, the analysis aims to ensure the dataset's integrity and improve the accuracy of insights into drug persistency patterns influenced by physician prescriptions.

Observation:
The dataset presented contains *3,424* datapoints and *69* variables.
For better understanding, the data is grouped in buckets:

- Demographics
- Provider Attributes
- Clinical Factors
- Disease and Treatment Factors

Target:
Persistency Flag

Data Cleaning and Preparation

A file named **clean_data.py** contains the script responsible for cleaning the *healthcare_data.xlsm* file and saving the newly cleaned file as *healthcare_dataset_cleaned.csv*.

Demographics

Race

Categorized	Uncategorized	Percentage
3327	97	**2.83**
In this feature, there are 97 Uncategorized values with the alias “Unknown”, “Other” or “Unknown/Others”, and 3327 Categorized values, resulting in a **2.83% of missing values**. **Actions Taken:** *Mode Imputer* A mode imputer is a data preprocessing technique that replaces missing values in a dataset with the most frequently occurring value (the mode) in the respective feature or column, ensuring that the imputed data reflects the prevailing category or value in the dataset.		

Ethnicity

Categorized	Uncategorized	Percentage
3333	91	**2.66**
In this feature, there are 91 Uncategorized values with the alias “Unknown”, “Other” or “Unknown/Others”, and 3333 Categorized values, resulting in a **2.66% of missing values**. **Actions Taken:** *Mode Imputer* A mode imputer is a data preprocessing technique that replaces missing values in a dataset with the most frequently occurring value (the mode) in the respective feature or column, ensuring that the imputed data reflects the prevailing category or value in the dataset.		

Region

Categorized	Uncategorized	Percentage
3364	60	**1.75**

In this feature, there are 60 Uncategorized values with the alias “Unknown”, “Other” or “Unknown/Others”, and 3364 Categorized values, resulting in a **1.75% of missing values**.

Actions Taken:
Mode Imputer

A mode imputer is a data preprocessing technique that replaces missing values in a dataset with the most frequently occurring value (the mode) in the respective feature or column, ensuring that the imputed data reflects the prevailing category or value in the dataset.

Provider Attributes

Ntm_Specialty

Categorized	Uncategorized	Percentage
3114	310	**9.05**

In this feature, there are 310 Uncategorized values with the alias “Unknown”, “Other” or “Unknown/Others”, and 3114 Categorized values, resulting in a **9.05% of missing values**.

Actions Taken:
Mode Imputer

A mode imputer is a data preprocessing technique that replaces missing values in a dataset with the most frequently occurring value (the mode) in the respective feature or column, ensuring that the imputed data reflects the prevailing category or value in the dataset.

Clinical Factors

Features	Categorized	Format
Gluco_Record_Prior_Ntm Gluco_Record_During_Rx Dexa_Freq_During_Rx Dexa_During_Rx Frag_Frac_Prior_Ntm Frag_Frac_During_Rx Risk_Segment_Prior_Ntm Tscore_Bucket_Prior_Ntm Change_Risk_Segment Adherent_Flag	Yes	**Binary**

The data in the features listed above are Categorized in Binary data format, containing values of "N" for No, and "Y" for Yes, or other categories than is considered **Binary**.

Actions Taken:

Binary Encoding

Binary encoding is a data transformation technique where categorical or text data is converted into a binary format, typically using 0s and 1s, to represent different categories or options. Each category is assigned a unique binary pattern, making it suitable for machine learning algorithms and data analysis.

Disease and Treatment Factor

Features	Categorized	Format
Idn_Indicator *Injectable_Experience_During_Rx* *Comorb_Encounter_For_Screening_For_Malignant_Neoplasms* *Comorb_Encounter_For_Immunization* *Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx* *Comorb_Vitamin_D_Deficiency* *Comorb_Other_Joint_Disorder_Not_Elsewhere_Classified* *Comorb_Encntr_For_Oth_Sp_Exam_W_O_Complaint_Suspected_Or_Reprtd_Dx* *Comorb_Long_Term_Current_Drug_Therapy* *Comorb_Dorsalgia* *Comorb_Personal_History_Of_Other_Diseases_And_Conditions* *Comorb_Other_Disorders_Of_Bone_Density_And_Structure* *Comorb_Disorders_of_lipoprotein_metabolism_and_other_lipidemias* *Comorb_Osteoporosis_without_current_pathological_fracture* *Comorb_Personal_history_of_malignant_neoplasm* *Comorb_Gastro_esophageal_reflux_disease* *Concom_Cholesterol_And_Triglyceride_Regulating_Preparations* *Concom_Narcotics* *Concom_Systemic_Corticosteroids_Plain* *Concom_Anti_Depressants_And_Mood_Stabilisers* *Concom_Fluoroquinolones* *Concom_Cephalosporins* *Concom_Macrolides_And_Similar_Types* *Concom_Broad_Spectrum_Penicillins* *Concom_Anaesthetics_General* *Concom_Viral_Vaccines* *Risk_Type_1_Insulin_Dependent_Diabetes* *Risk_Osteogenesis_Imperfecta* *Risk_Rheumatoid_Arthritis* *Risk_Untreated_Chronic_Hyperthyroidism* *Risk_Untreated_Chronic_Hypogonadism* *Risk_Untreated_Early_Menopause* *Risk_Patient_Parent_Fractured_Their_Hip* *Risk_Smoking_Tobacco* *Risk_Chronic_Malnutrition_Or_Malabsorption* *Risk_Chronic_Liver_Disease* *Risk_Family_History_Of_Osteoporosis* *Risk_Low_Calcium_Intake* *Risk_Vitamin_D_Insufficiency* *Risk_Poor_Health_Frailty* *Risk_Excessive_Thinness* *Risk_Hysterectomy_Oophorectomy* *Risk_Estrogen_Deficiency* *Risk_Immobilization* *Risk_Recurring_Falls*	Yes	**Binary**

The data in the features listed above are Categorized in Binary data format, containing values of “N” for No, and “Y” for Yes, or other categories than is considered **Binary**.

Actions Taken:

Binary Encoding

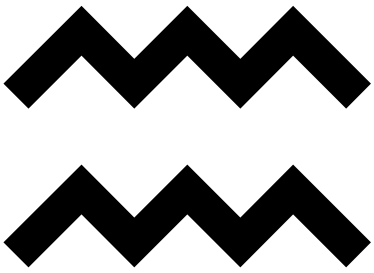
Binary encoding is a data transformation technique where categorical or text data is converted into a binary format, typically using 0s and 1s, to represent different categories or options. Each category is assigned a unique binary pattern, making it suitable for machine learning algorithms and data analysis.

References

Github

https://github.com/anvadev/Healthcare-Drug_Persistency/tree/main/Week%209%20-%20Data%20Cleaning%20and%20Preparation

End of Documentation



QUANTIC