Humana-Mays Healthcare Analytics 2023 Case Competition

Table of Contents

1. EXECUTIVE SUMMARY

1.1. Study proposal

The case involves analyzing case data provided by Humana consisting of their members in their first six months of Osimertinib (sold under the brand name **Tagrisso**) therapy, which is an effective drug against non-small cell lung cancer. The available data contains insurance claims during and before therapy and span across the years 2018-2022, with individual lookback 90 days previous to the first Osimertinib fill, through Osimertinib therapy.

Our study proposal intends to look at the patient, pharmacy and insurance data provided by Humana-Mays, and hopes to come up with solution(s) that would help predict members that are most likely to experience an Adverse Drug Event (ADE) and discontinue therapy.

In order to also ensure that our model was fair, we decided to leave null race and gender cells as unknowns, rather than populating them with their respective modes.

Some of the key questions we decided to look at were:

a. Are there any racial disparities in the cumulative costs? In other words, were different race groups incurring different costs in their drug therapy?
b. Is there any association between race and existence of an Adverse Drug Event (ADE) for a patient?
c. Is there any association between gender and existence of ADE for a patient?

Some of the key performance indicators (KPI) we used were derived features like the average number of medical visits, the number of insurance claims, the number of times a patient was diagnosed with ADE, and average days to process claims, along with several other featured engineered variables.

1.2. Modeling

In order to achieve the best performance of modeling, we carried out comprehensive studies in understanding the business issue we need to fix and all the features in the dataset.
First, we built a predictive model to identify members who are the most likely to withdraw from the treatment. We chose Gini Index, random forest and XGBoost to do feature selection based

on three models' intersection, developing a better understanding of the most important features included in our model.

Then we applied Decision Tree, Random Forest, Gradient Boosting Decision Tree, and XGBoost, along with parameter tuning to do preliminary prediction and compared their performances and corresponding AUC.

Finally, we got the best performance with an AUC of 0.97 with Random Forest. Further analysis and recommendations regarding improvement of identifying the potential patients that might withdraw from the treatment, is based on the features we derived.

1.3. Recommendations

Our recommendations are multi-fold, based on our observations and insights that we gathered from the data.

a. Look for treatment gaps between diagnoses in patients and corresponding treatments. Based on the data provided, there seem to be some diagnoses like fatigue that seem to have been left untreated. In addition, several patients seem to have discontinued after just 1 or 2 instances of recorded ADE, where effective management of side effects may have perhaps helped them to continue the therapy.

b. Reduce the number of days it takes to process insurance claims.

c. Keep Medicare costs reasonable and affordable.

d. Reassess patients coming through Outpatient and ER visits, to ensure quicker response times and turnaround times.

e. Consider adding additional variables to handle false-negatives and find any confounding variables to address the presence of Covid, and to see if any of the medical visits were related to that. Instead of accuracy, precision-recall would be a better metric to ensure there was a balance between the trade-offs of true positive and false positive rates, especially in cases where class imbalances are present.

f. If we had the time, we would have liked to explore all the other variables that had a high correlation with the seizure diagnosis. This is because the patients that did experience seizure do not seem to have completed the therapy, nor was there any evidence of treatment provided to these patients (13 in training, 5 discontinued).
   We would have also liked to keep track if a patient switched from one mode of visit to another, especially since outpatient and ER visits appear as features of importance.

2. CASE BACKGROUND

2.1. Context

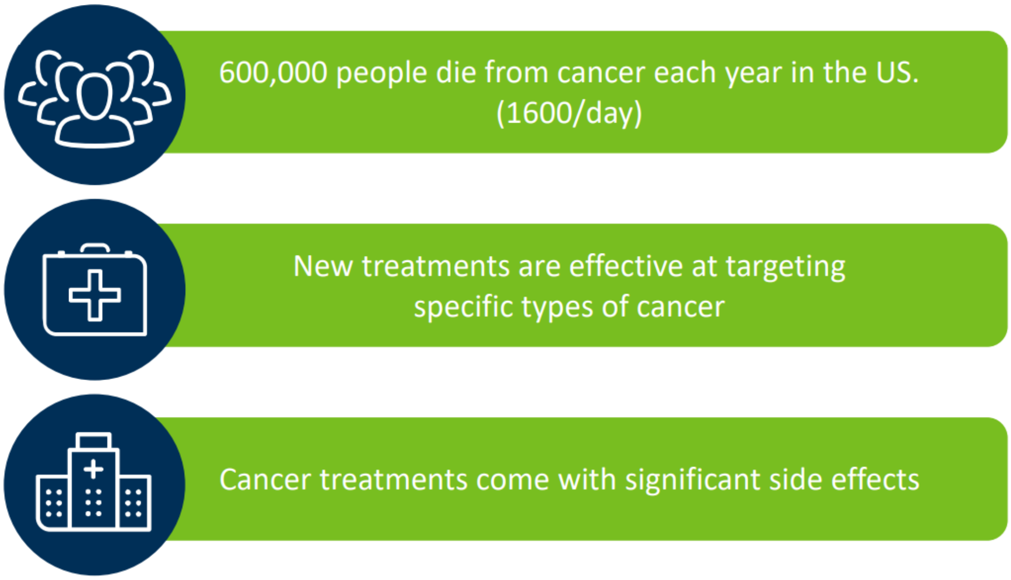
Humana Inc. (NYSE: HUM) is committed to helping millions of medical and specialty members achieve their best health. Their successful history in care delivery and health plan administration is helping to create a new kind of integrated care with the power to improve health and well-being and lower costs. Humana's efforts are leading to a better quality of life for people with Medicare, families, individuals, military service personnel, and communities at large.

This case involves tracking the effectiveness of the drug Osimertinib (TAGRISSO), which is a drug therapy used to treat non-small cell lung cancer. The population consists of Humana members that have been detected with the biomarker EGFR+R (Epidermal Growth Factor Receptor), indicating an early-stage non-small cell lung cancer (NSCLC), and are in their first six months of Osimertinib therapy.

People on Osimertinib therapy are twice as likely to survive vs. people who took no active medicine. Moreover, they are 80% less likely to have their cancer come back or die.

2.2. Problem Statement

A large number of people die from cancer every year, and while the effectiveness of cancer-fighting drugs is pretty high, they also come with serious side-effects that make it difficult for a patient to complete therapy and makes them quit mid-way through the treatment.



600,000 people die from cancer each year in the US. (1600/day)

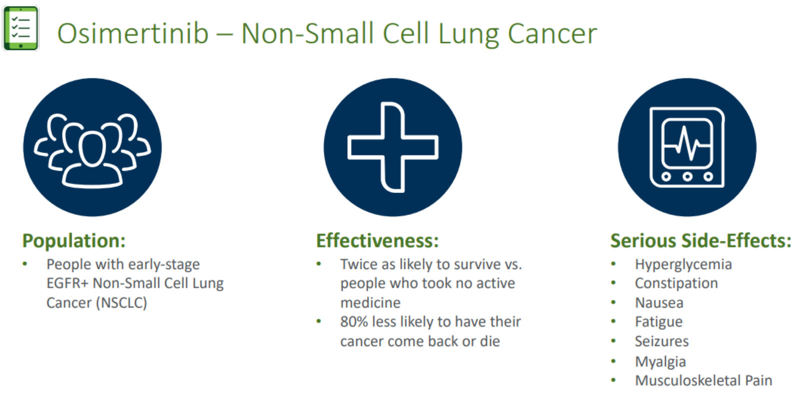
New treatments are effective at targeting specific types of cancer

Cancer treatments come with significant side effects

https://www.cdc.gov/cancer/dcpc/research/update-on-cancer-deaths/index.htm

Our problem statement:

To ensure the effectiveness of the Osimertinib drug, which is used to fight the non-small cell lung cancer, and increase the chances of survival, it is important that patients continue to adhere to the medication despite the serious side-effects that come with it (listed below). Approximately 24% of members taking Osimertinib have a side-effect and discontinue in the first 6 months of therapy. Our aim is to target at-risk members to improve adherence and survival.



Osimertinib – Non-Small Cell Lung Cancer

Population:

- People with early-stage EGFR+ Non-Small Cell Lung Cancer (NSCLC)

Effectiveness:

- Twice as likely to survive vs. people who took no active medicine
- 80% less likely to have their cancer come back or die

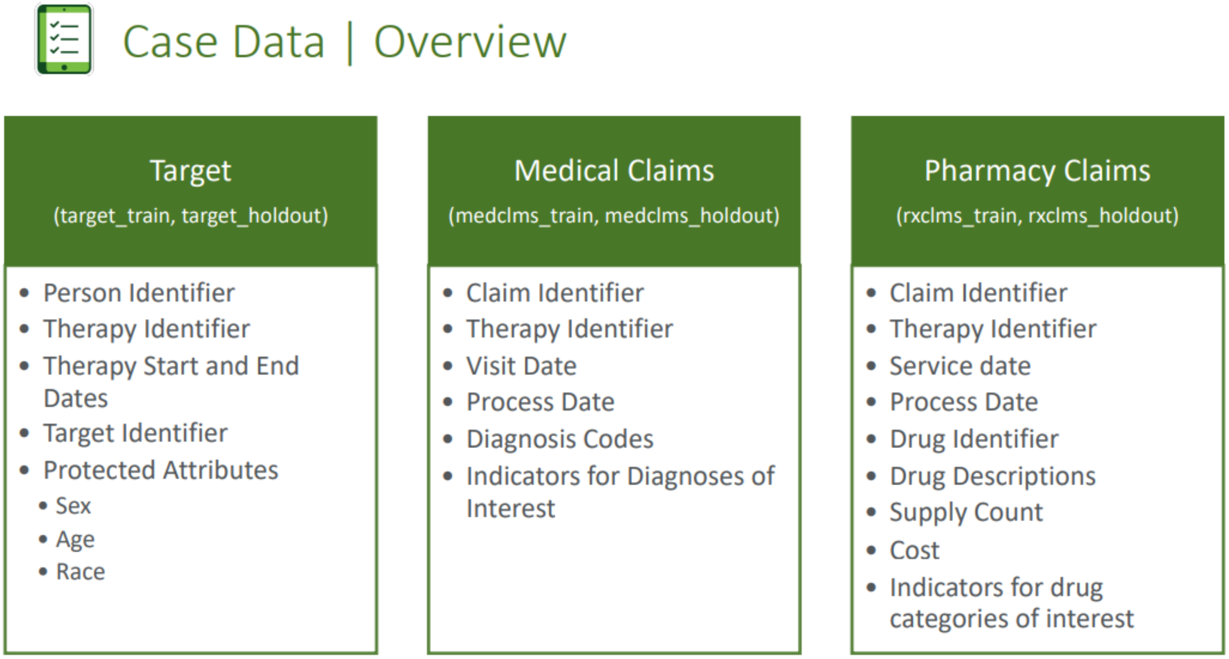
Serious Side-Effects:

- Hyperglycemia
- Constipation
- Nausea
- Fatigue
- Seizures
- Myalgia
- Musculoskeletal Pain

3. DATA ANALYSIS

3.1. Dataset Description

As part of the competition, we were given 3 files for training and 3 for the holdout (test) data, with the below overview:



Case Data | Overview

Target
(target_train, target_holdout)

- Person Identifier
- Therapy Identifier
- Therapy Start and End Dates
- Target Identifier
- Protected Attributes
  - Sex
  - Age
  - Race

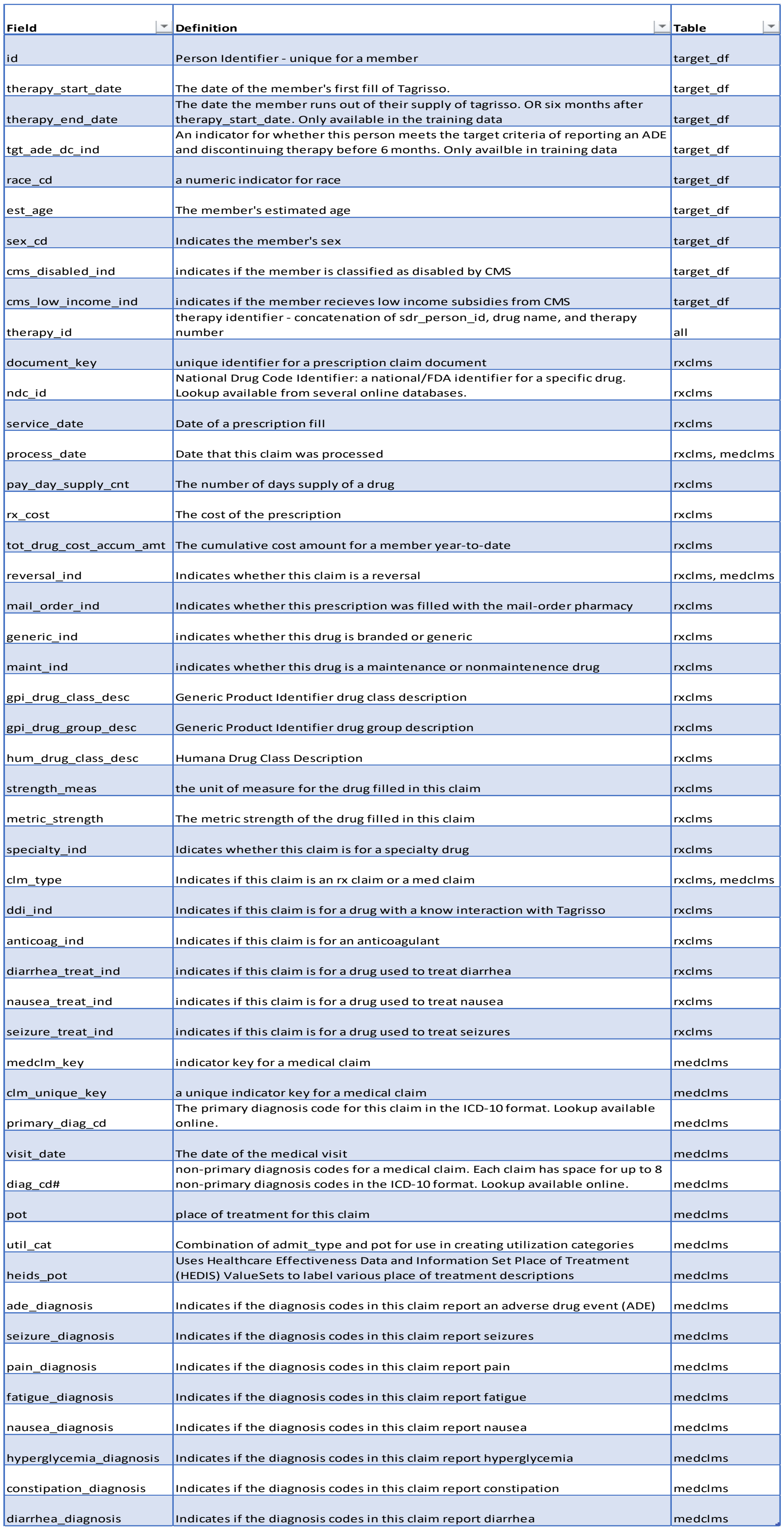
Medical Claims
(medclms_train, medclms_holdout)

- Claim Identifier
- Therapy Identifier
- Visit Date
- Process Date
- Diagnosis Codes
- Indicators for Diagnoses of Interest

Pharmacy Claims
(rxclms_train, rxclms_holdout)

- Claim Identifier
- Therapy Identifier
- Service date
- Process Date
- Drug Identifier
- Drug Descriptions
- Supply Count
- Cost
- Indicators for drug categories of interest

The data structure of all the three the original tables, along with the field names and descriptions are shown below:



Field	Definition	Table
id	Person Identifier - unique for a member	target_df
therapy_start_date	The date of the member's first fill of Tagrisso.	target_df
therapy_end_date	The date the member runs out of their supply of tagrisso. OR six months after therapy_start_date. Only available in the training data	target_df
tgt_ade_dc_ind	An indicator for whether this person meets the target criteria of reporting an ADE and discontinuing therapy before 6 months. Only availble in training data	target_df
race_cd	a numeric indicator for race	target_df
est_age	The member's estimated age	target_df
sex_cd	Indicates the member's sex	target_df
cms_disabled_ind	indicates if the member is classified as disabled by CMS	target_df
cms_low_income_ind	indicates if the member recieves low income subsidies from CMS	target_df
therapy_id	therapy identifier - concatenation of sdr_person_id, drug name, and therapy number	all
document_key	unique identifier for a prescription claim document	rxclms
ndc_id	National Drug Code Identifier: a national/FDA identifier for a specific drug. Lookup available from several online databases.	rxclms
service_date	Date of a prescription fill	rxclms
process_date	Date that this claim was processed	rxclms, medclms
pay_day_supply_cnt	The number of days supply of a drug	rxclms
rx_cost	The cost of the prescription	rxclms
tot_drug_cost_accum_amt	The cumulative cost amount for a member year-to-date	rxclms
reversal_ind	Indicates whether this claim is a reversal	rxclms, medclms
mail_order_ind	Indicates whether this prescription was filled with the mail-order pharmacy	rxclms
generic_ind	indicates whether this drug is branded or generic	rxclms
maint_ind	indicates whether this drug is a maintenance or nonmaintenence drug	rxclms
gpi_drug_class_desc	Generic Product Identifier drug class description	rxclms
gpi_drug_group_desc	Generic Product Identifier drug group description	rxclms
hum_drug_class_desc	Humana Drug Class Description	rxclms
strength_meas	the unit of measure for the drug filled in this claim	rxclms
metric_strength	The metric strength of the drug filled in this claim	rxclms
specialty_ind	Idicates whether this claim is for a specialty drug	rxclms
clm_type	Indicates if this claim is an rx claim or a med claim	rxclms, medclms
ddi_ind	Indicates if this claim is for a drug with a know interaction with Tagrisso	rxclms
anticoag_ind	Indicates if this claim is for an anticoagulant	rxclms
diarrhea_treat_ind	indicates if this claim is for a drug used to treat diarrhea	rxclms
nausea_treat_ind	indicates if this claim is for a drug used to treat nausea	rxclms
seizure_treat_ind	indicates if this claim is for a drug used to treat seizures	rxclms
medclm_key	indicator key for a medical claim	medclms
clm_unique_key	a unique indicator key for a medical claim	medclms
primary_diag_cd	The primary diagnosis code for this claim in the ICD-10 format. Lookup available online.	medclms
visit_date	The date of the medical visit	medclms
diag_cd#	non-primary diagnosis codes for a medical claim. Each claim has space for up to 8 non-primary diagnosis codes in the ICD-10 format. Lookup available online.	medclms
pot	place of treatment for this claim	medclms
util_cat	Combination of admit_type and pot for use in creating utilization categories	medclms
heids_pot	Uses Healthcare Effectiveness Data and Information Set Place of Treatment (HEDIS) ValueSets to label various place of treatment descriptions	medclms
ade_diagnosis	Indicates if the diagnosis codes in this claim report an adverse drug event (ADE)	medclms
seizure_diagnosis	Indicates if the diagnosis codes in this claim report seizures	medclms
pain_diagnosis	Indicates if the diagnosis codes in this claim report pain	medclms
fatigue_diagnosis	Indicates if the diagnosis codes in this claim report fatigue	medclms
nausea_diagnosis	Indicates if the diagnosis codes in this claim report nausea	medclms
hyperglycemia_diagnosis	Indicates if the diagnosis codes in this claim report hyperglycemia	medclms
constipation_diagnosis	Indicates if the diagnosis codes in this claim report constipation	medclms
diarrhea_diagnosis	Indicates if the diagnosis codes in this claim report diarrhea	medclms

Table 1 Original data structure

The training data set had 1,232 unique therapy IDs that were tied to a patient's record and was the primary key across all three files.

The holdout set had 420 unique therapy IDs.

The target outcome variable was tgt_ade_dc_ind, and present only in the training data set.

In addition, we were provided a separate crosswalk file for race:

Table 2

race_cd	race_cd_desc
0	unknown
1	white
2	black
3	other
4	asian
5	hispanic
6	n amer native

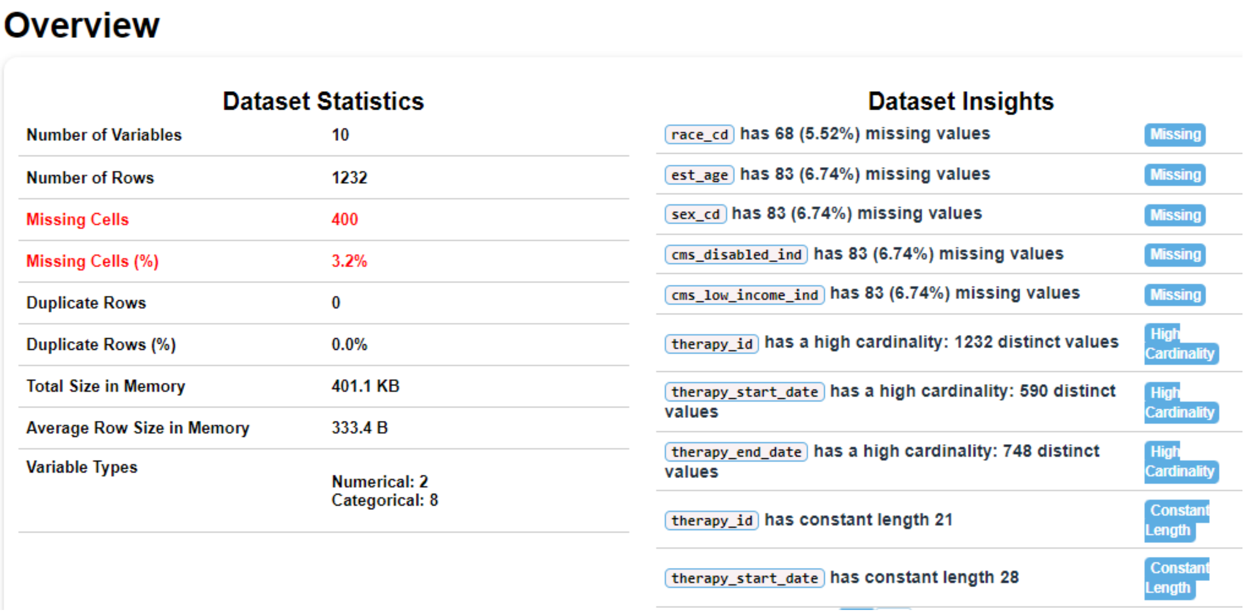
Table 3 Race crosswalk

3.2. Descriptive Statistics (exploratory data analysis)

Looking at each of the three files individually in training, and then for the holdout data, we gleaned some insights regarding their general distribution and any missing data.

Training

Target file:



Overview

Dataset Statistics

Number of Variables	10
Number of Rows	1232
Missing Cells	400
Missing Cells (%)	3.2%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	401.1 KB
Average Row Size in Memory	333.4 B
Variable Types	Numerical: 2 Categorical: 8

Dataset Insights

race_cd has 68 (5.52%) missing values	Missing
est_age has 83 (6.74%) missing values	Missing
sex_cd has 83 (6.74%) missing values	Missing
cms_disabled_ind has 83 (6.74%) missing values	Missing
cms_low_income_ind has 83 (6.74%) missing values	Missing
therapy_id has a high cardinality: 1232 distinct values	High Cardinality
therapy_start_date has a high cardinality: 590 distinct values	High Cardinality
therapy_end_date has a high cardinality: 748 distinct values	High Cardinality
therapy_id has constant length 21	Constant Length
therapy_start_date has constant length 28	Constant Length

Overview

Dataset Statistics

Number of Variables	10
Number of Rows	1232
Missing Cells	400
Missing Cells (%)	3.2%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	401.1 KB
Average Row Size in Memory	333.4 B
Variable Types	Numerical: 2 Categorical: 8

Dataset Insights

`tgt_ade_dc_ind` has constant length 1	Constant Length
`race_cd` has constant length 3	Constant Length
`sex_cd` has constant length 1	Constant Length
`cms_disabled_ind` has constant length 3	Constant Length
`cms_low_income_ind` has constant length 3	Constant Length
`therapy_id` has all distinct values	Unique

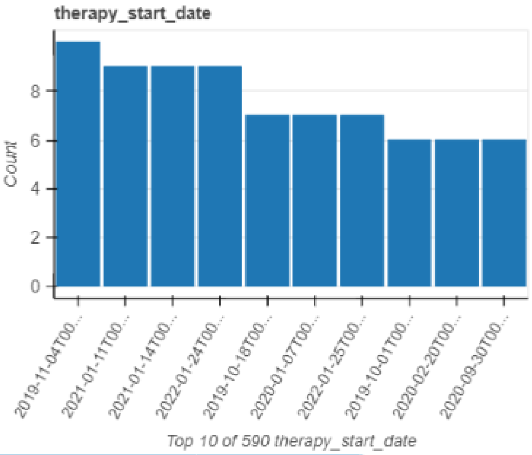
1 2

therapy_start_date

categorical

Hide Details

Approximate Distinct Count	590
Approximate Unique (%)	47.9%
Missing	0
Missing (%)	0.0%
Memory Size	111.9 KB



Stats | Pie Chart | Word Cloud | Word Frequency | Word Length

Length

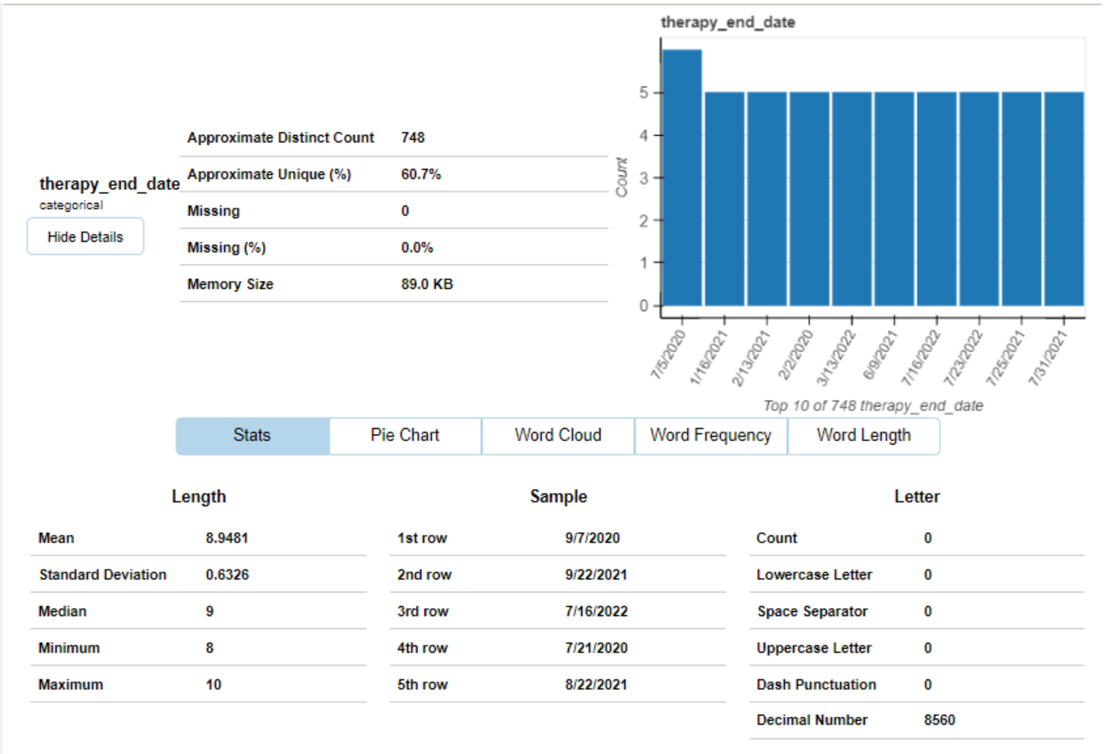
Mean	28
Standard Deviation	0
Median	28
Minimum	28
Maximum	28

Sample

1st row	2020-03-11T00:00:0...
2nd row	2021-08-23T00:00:0...
3rd row	2022-01-17T00:00:0...
4th row	2020-01-23T00:00:0...
5th row	2021-02-23T00:00:0...

Letter

Count	1232
Lowercase Letter	0
Space Separator	0
Uppercase Letter	1232
Dash Punctuation	2464
Decimal Number	25872



Approximate Distinct Count	748
Approximate Unique (%)	60.7%
Missing	0
Missing (%)	0.0%
Memory Size	89.0 KB

Length

Mean	8.9481
Standard Deviation	0.6326
Median	9
Minimum	8
Maximum	10

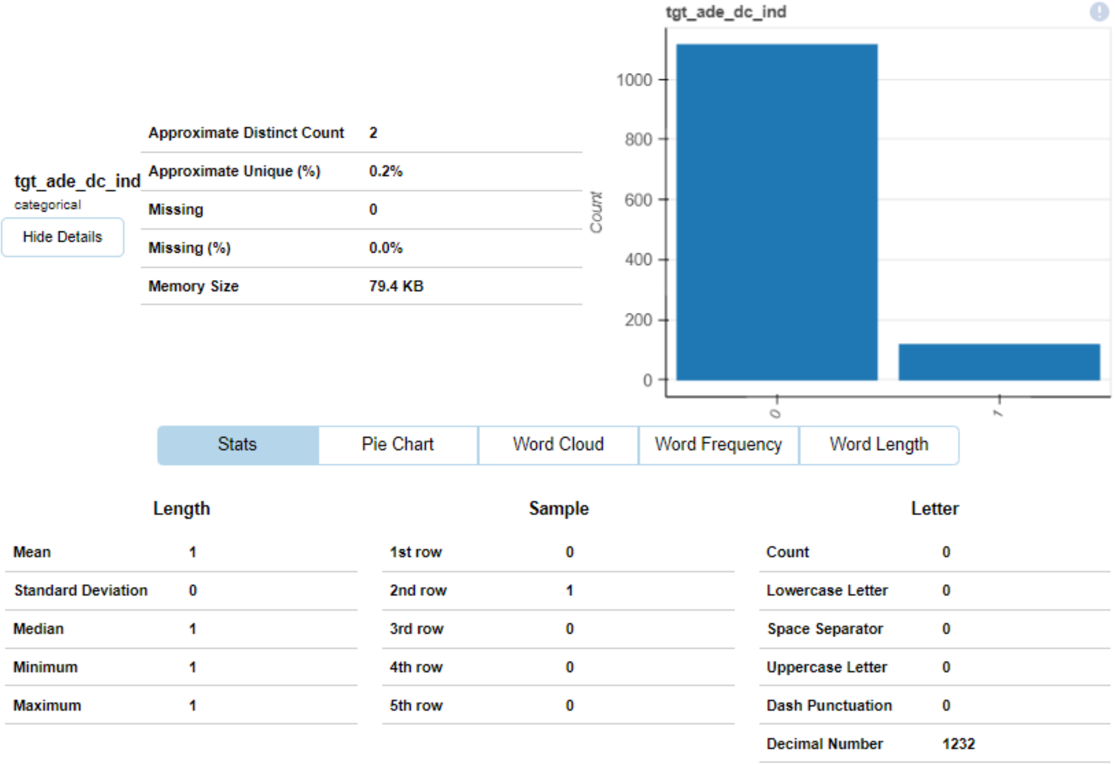
Sample

1st row	9/7/2020
2nd row	9/22/2021
3rd row	7/16/2022
4th row	7/21/2020
5th row	8/22/2021

Letter

Count	0
Lowercase Letter	0
Space Separator	0
Uppercase Letter	0
Dash Punctuation	0
Decimal Number	8560

The target outcome variable tgt_ade_dc_ind is imbalanced in the dataset with 2 distinct values: 0 (90.5%, 1115 occurrences) and 1 (9.5%, 117 occurrences). No missing values.



Approximate Distinct Count	2
Approximate Unique (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory Size	79.4 KB

Length

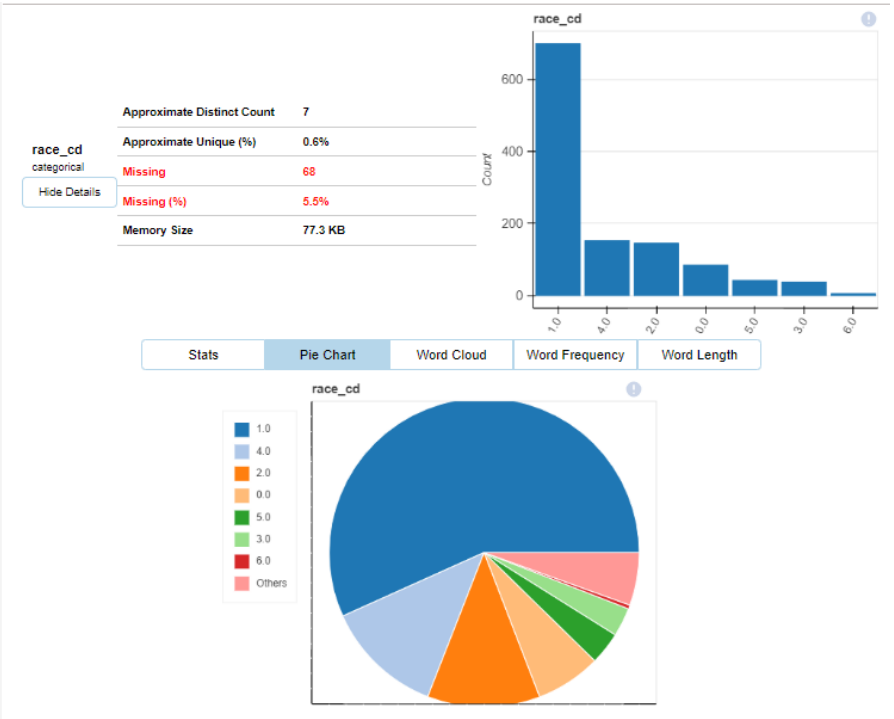
Mean	1
Standard Deviation	0
Median	1
Minimum	1
Maximum	1

Sample

1st row	0
2nd row	1
3rd row	0
4th row	0
5th row	0

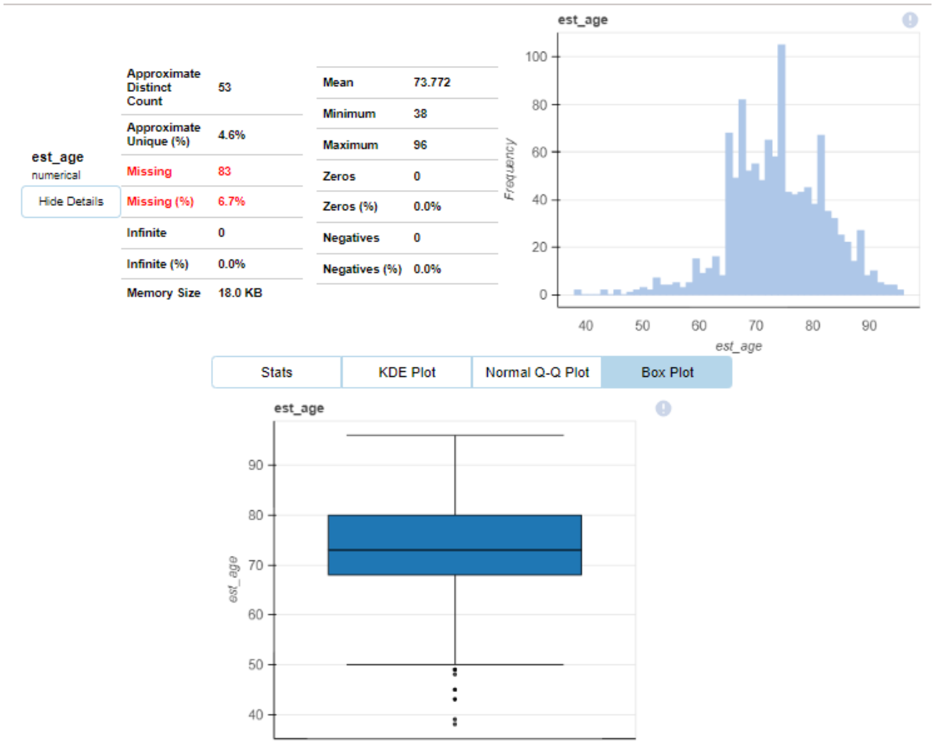
Letter

Count	0
Lowercase Letter	0
Space Separator	0
Uppercase Letter	0
Dash Punctuation	0
Decimal Number	1232



race_cd
categorical

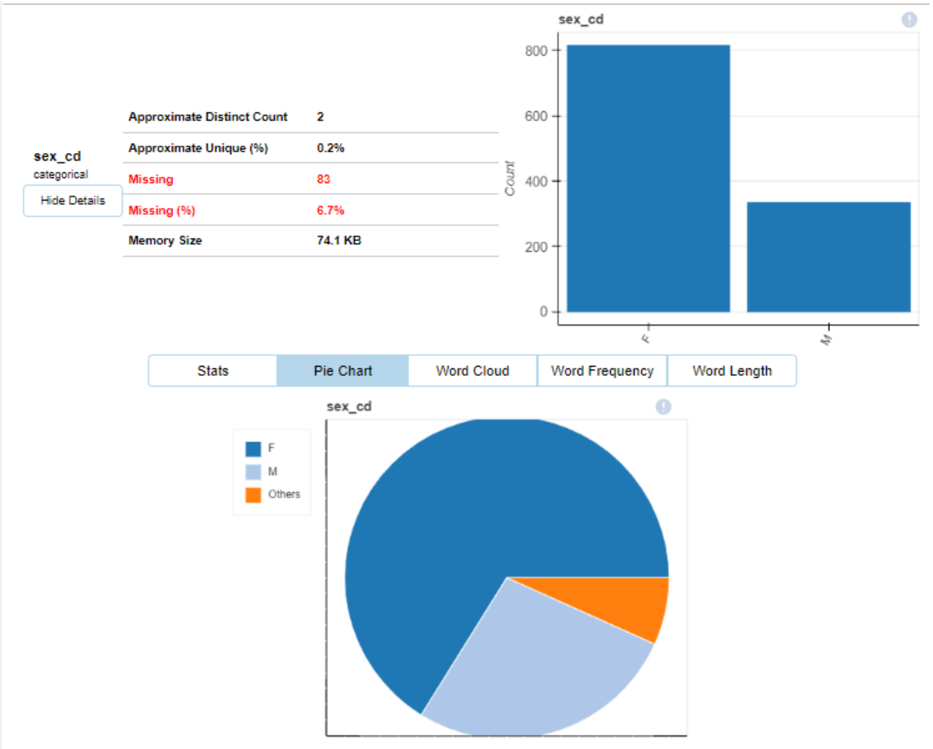
Approximate Distinct Count	7
Approximate Unique (%)	0.6%
Missing	68
Missing (%)	5.5%
Memory Size	77.3 KB



est_age
numerical

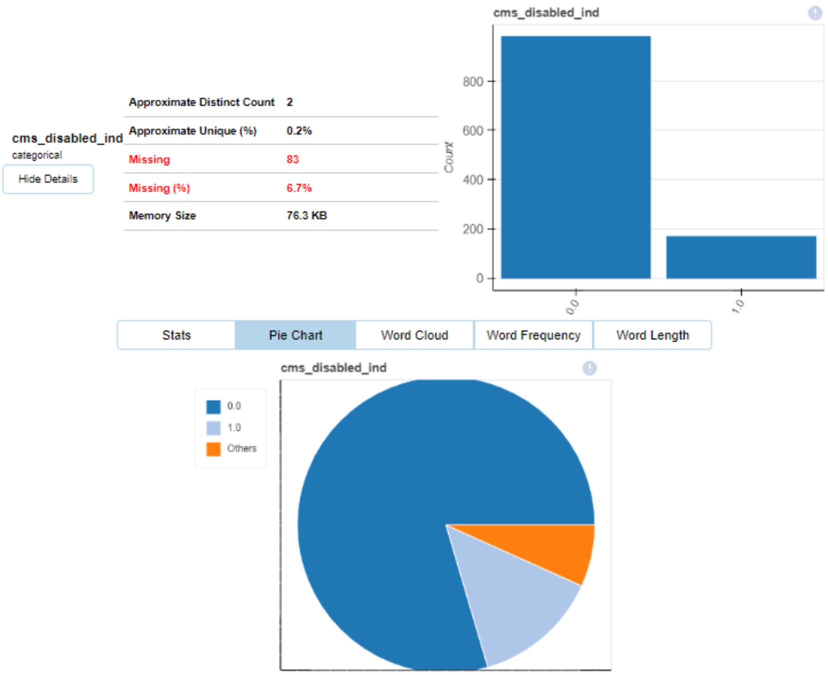
Approximate Distinct Count	53
Approximate Unique (%)	4.6%
Missing	83
Missing (%)	6.7%
Infinite	0
Infinite (%)	0.0%
Memory Size	18.0 KB

Mean	73.772
Minimum	38
Maximum	96
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



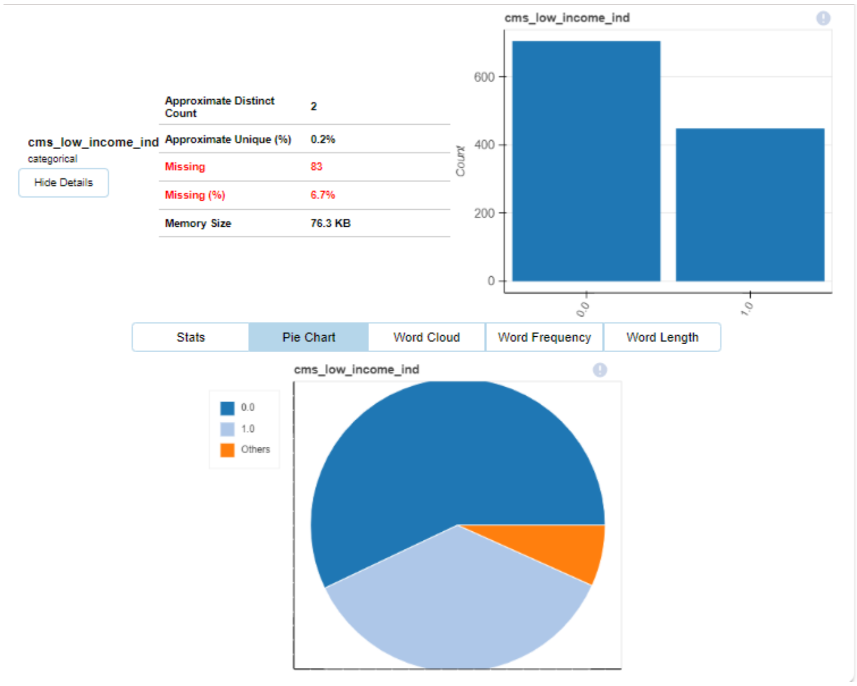
sex_cd
categorical

Approximate Distinct Count	2
Approximate Unique (%)	0.2%
Missing	83
Missing (%)	6.7%
Memory Size	74.1 KB



cms_disabled_ind
categorical

Approximate Distinct Count	2
Approximate Unique (%)	0.2%
Missing	83
Missing (%)	6.7%
Memory Size	76.3 KB

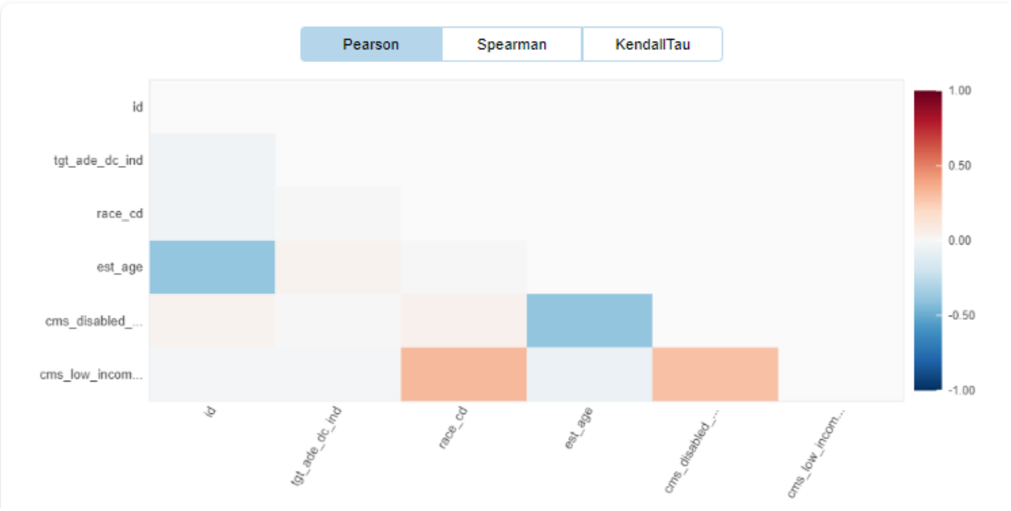


cms_low_income_ind
categorical
Hide Details

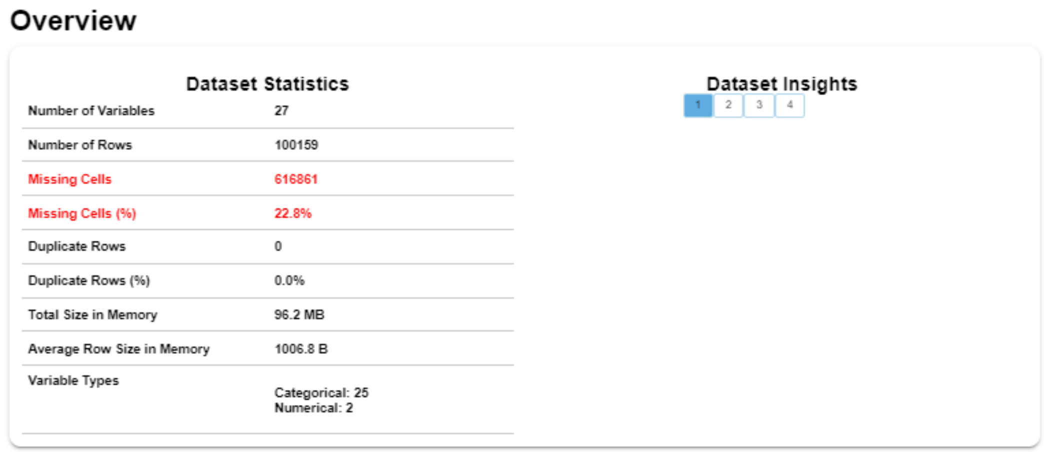
Approximate Distinct Count	2
Approximate Unique (%)	0.2%
Missing	83
Missing (%)	6.7%
Memory Size	76.3 KB

None of the variables in the target file are correlated very highly - neither positive nor negative.

Correlations



Insurance Claims:

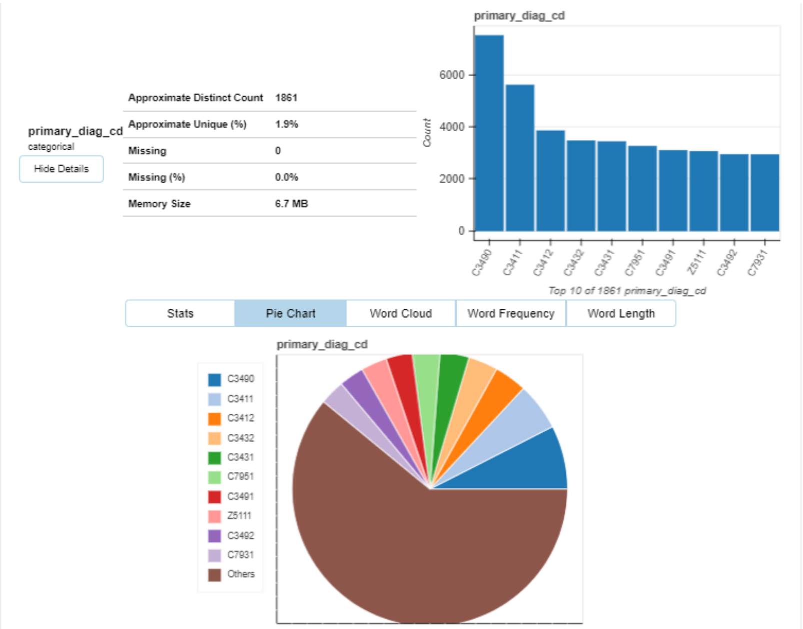


Overview

Dataset Statistics

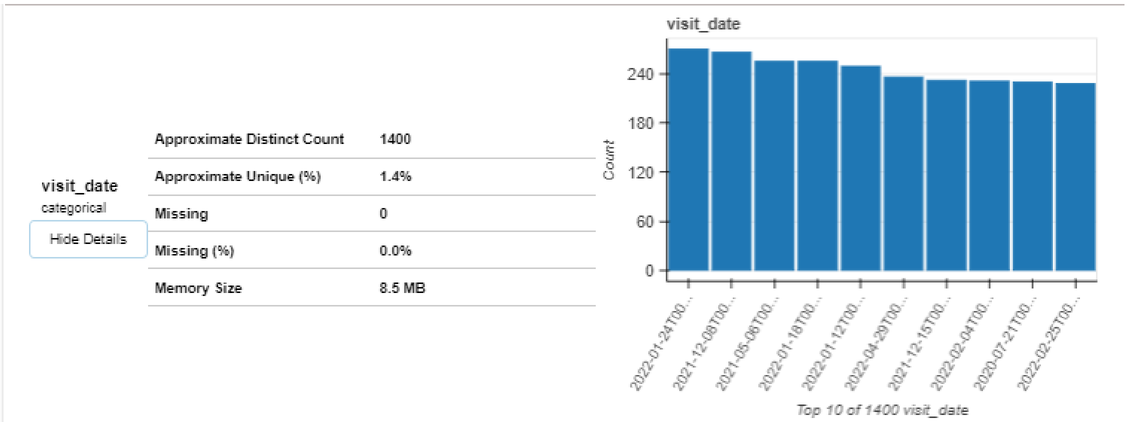
Number of Variables	27
Number of Rows	100159
Missing Cells	616861
Missing Cells (%)	22.8%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	96.2 MB
Average Row Size in Memory	1006.8 B
Variable Types	Categorical: 25 Numerical: 2

Dataset Insights



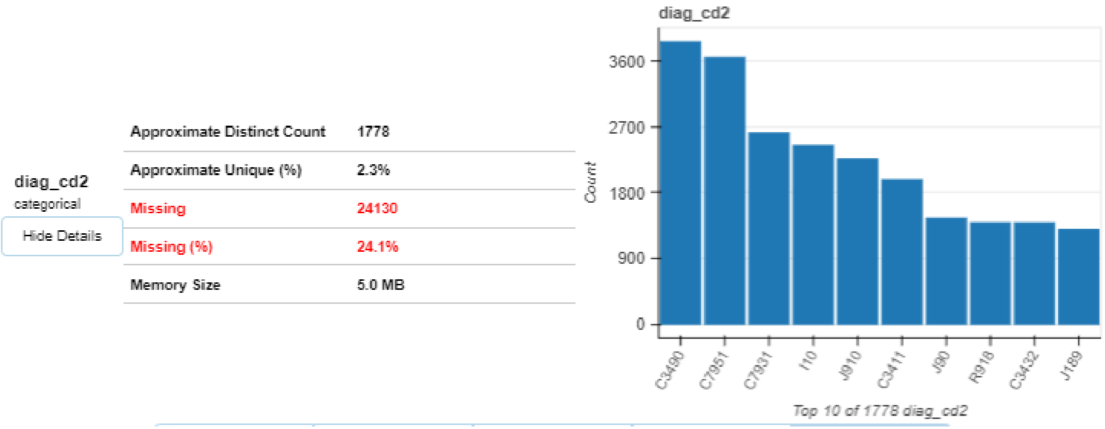
primary_diag_cd
categorical

Approximate Distinct Count	1861
Approximate Unique (%)	1.9%
Missing	0
Missing (%)	0.0%
Memory Size	6.7 MB



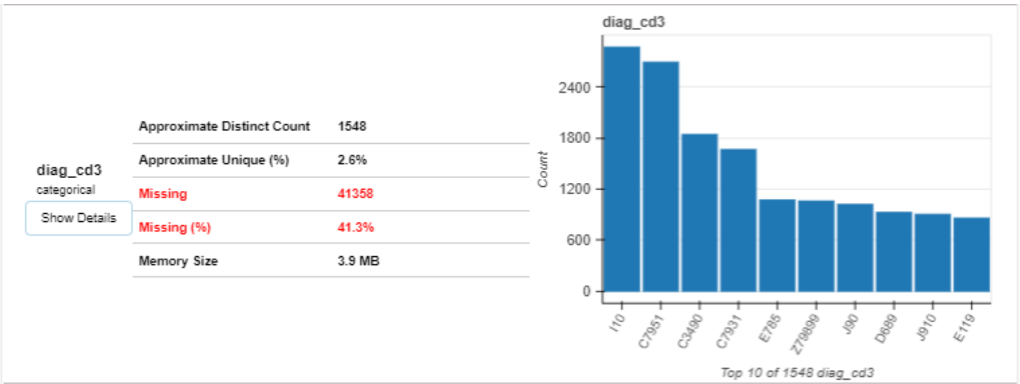
visit_date
categorical

Approximate Distinct Count	1400
Approximate Unique (%)	1.4%
Missing	0
Missing (%)	0.0%
Memory Size	8.5 MB



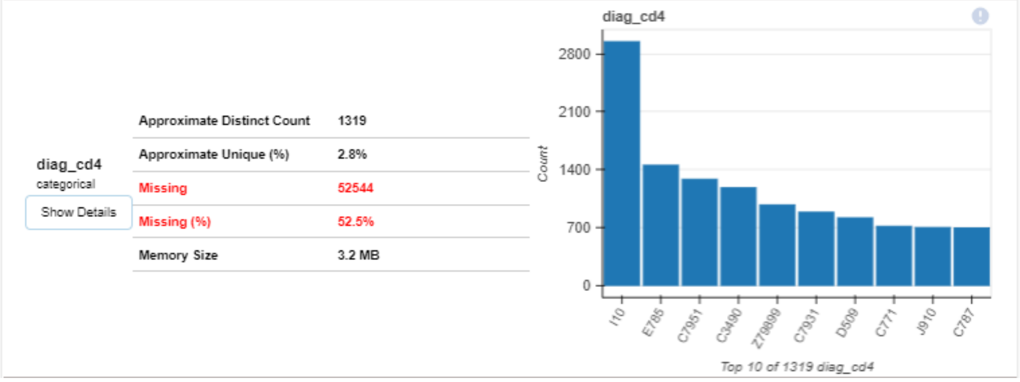
diag_cd2
categorical
Hide Details

Approximate Distinct Count	1778
Approximate Unique (%)	2.3%
Missing	24130
Missing (%)	24.1%
Memory Size	5.0 MB



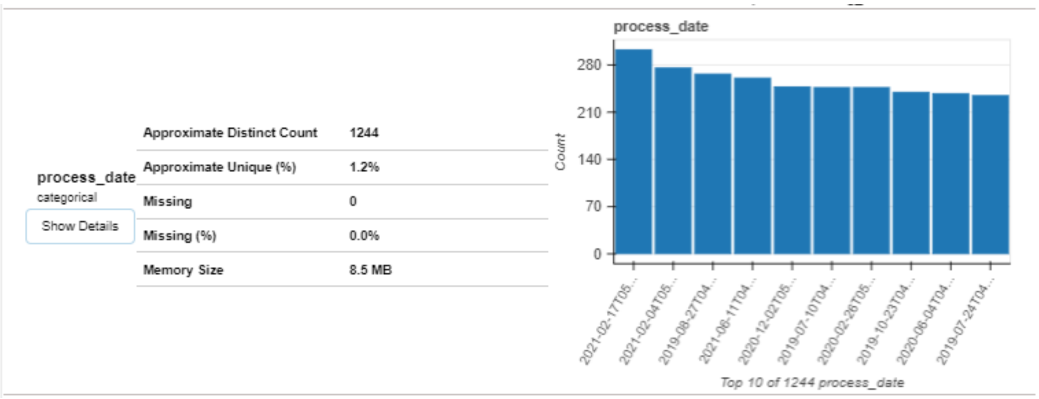
diag_cd3
categorical
Show Details

Approximate Distinct Count	1548
Approximate Unique (%)	2.6%
Missing	41358
Missing (%)	41.3%
Memory Size	3.9 MB



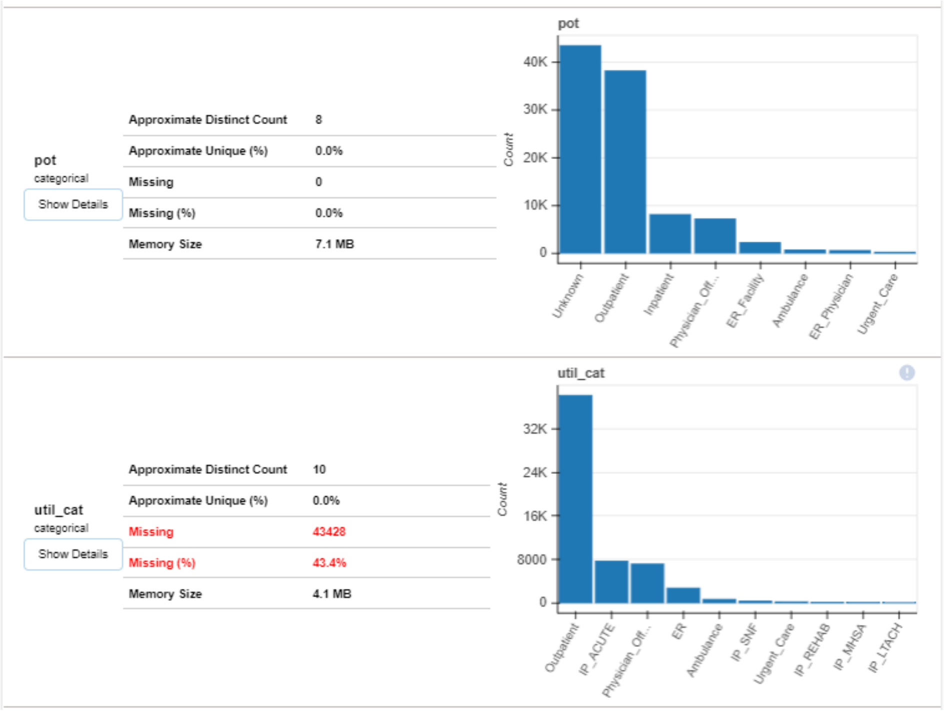
diag_cd4
categorical
Show Details

Approximate Distinct Count	1319
Approximate Unique (%)	2.8%
Missing	52544
Missing (%)	52.5%
Memory Size	3.2 MB



process_date
categorical
Show Details

Approximate Distinct Count	1244
Approximate Unique (%)	1.2%
Missing	0
Missing (%)	0.0%
Memory Size	8.5 MB

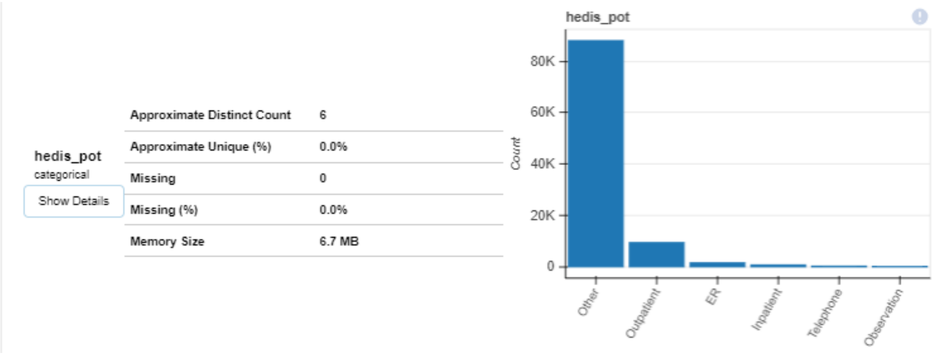


pot
categorical

Approximate Distinct Count	8
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	7.1 MB

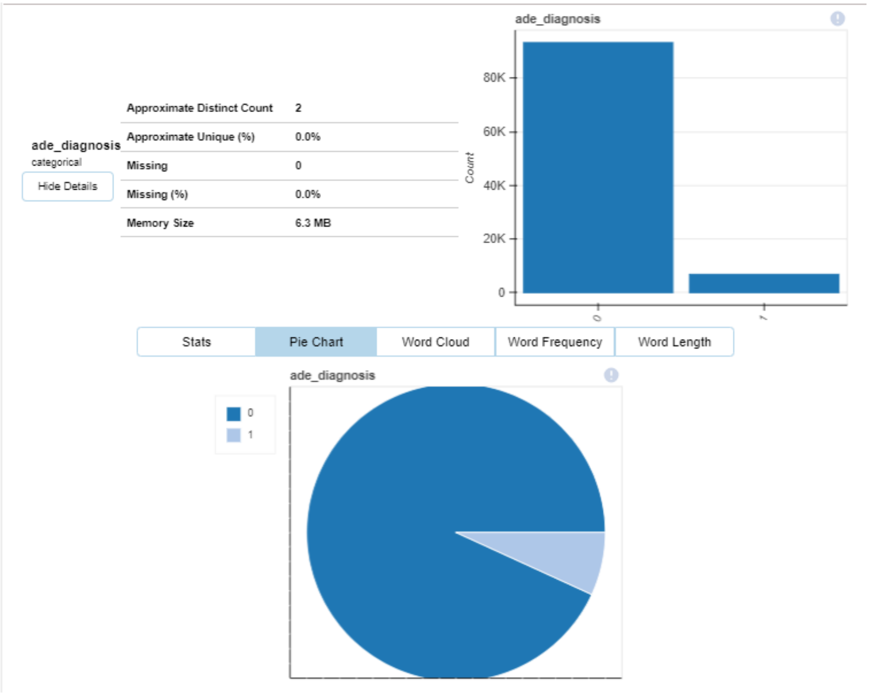
util_cat
categorical

Approximate Distinct Count	10
Approximate Unique (%)	0.0%
Missing	43428
Missing (%)	43.4%
Memory Size	4.1 MB



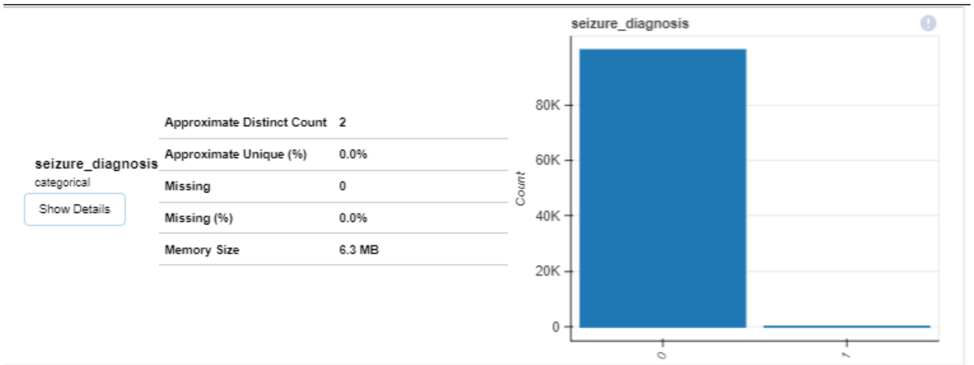
hedis_pot
categorical

Approximate Distinct Count	6
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	6.7 MB



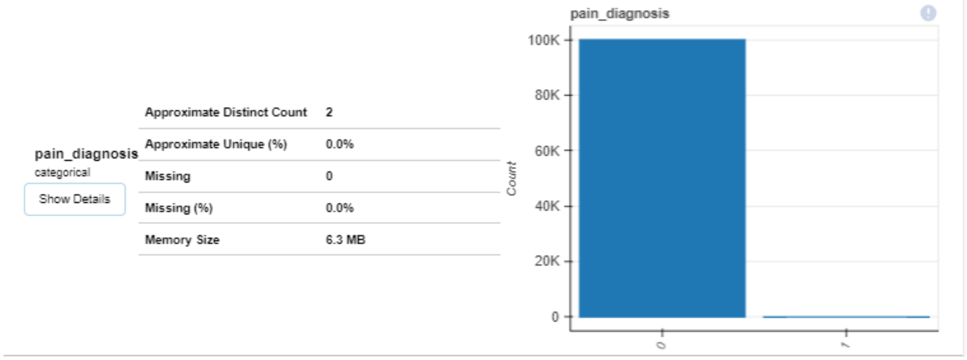
ade_diagnosis
categorical

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	6.3 MB



seizure_diagnosis
categorical

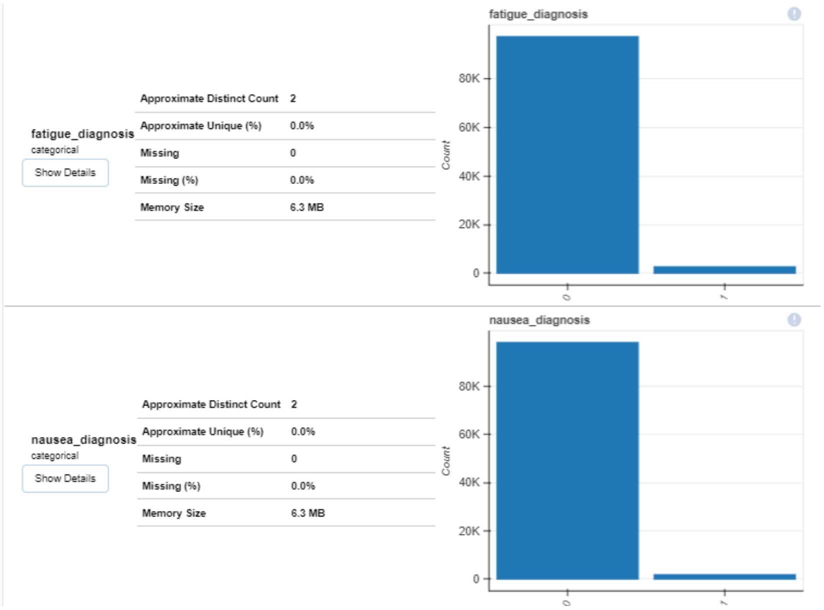
Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	6.3 MB



pain_diagnosis
categorical

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	6.3 MB

]]>

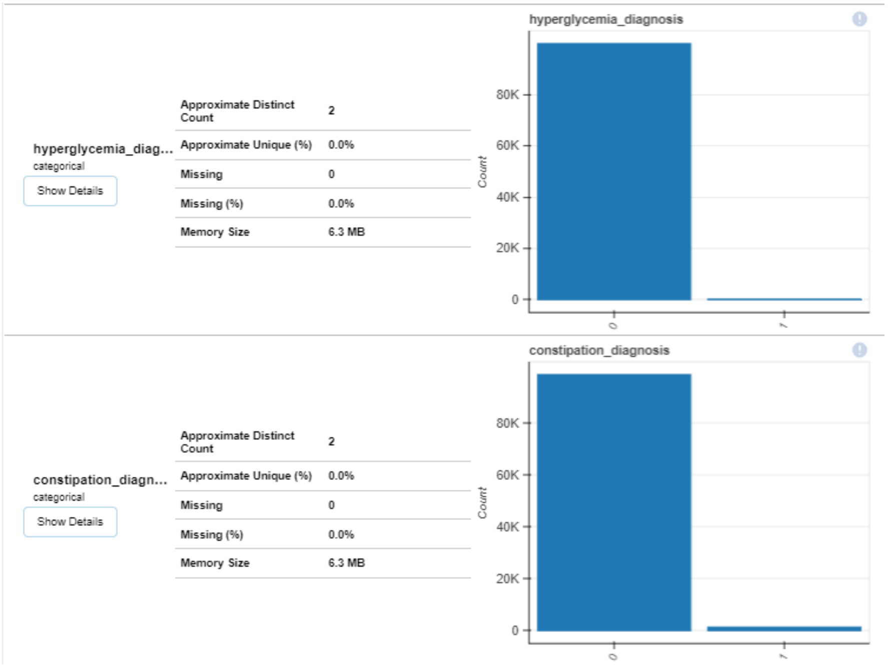


fatigue_diagnosis
categorical

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	6.3 MB

nausea_diagnosis
categorical

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	6.3 MB

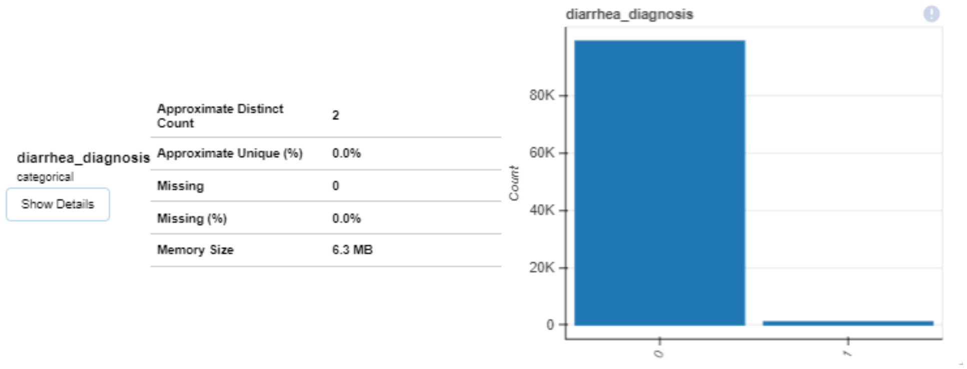


hyperglycemia_diag...
categorical

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	6.3 MB

constipation_diagn...
categorical

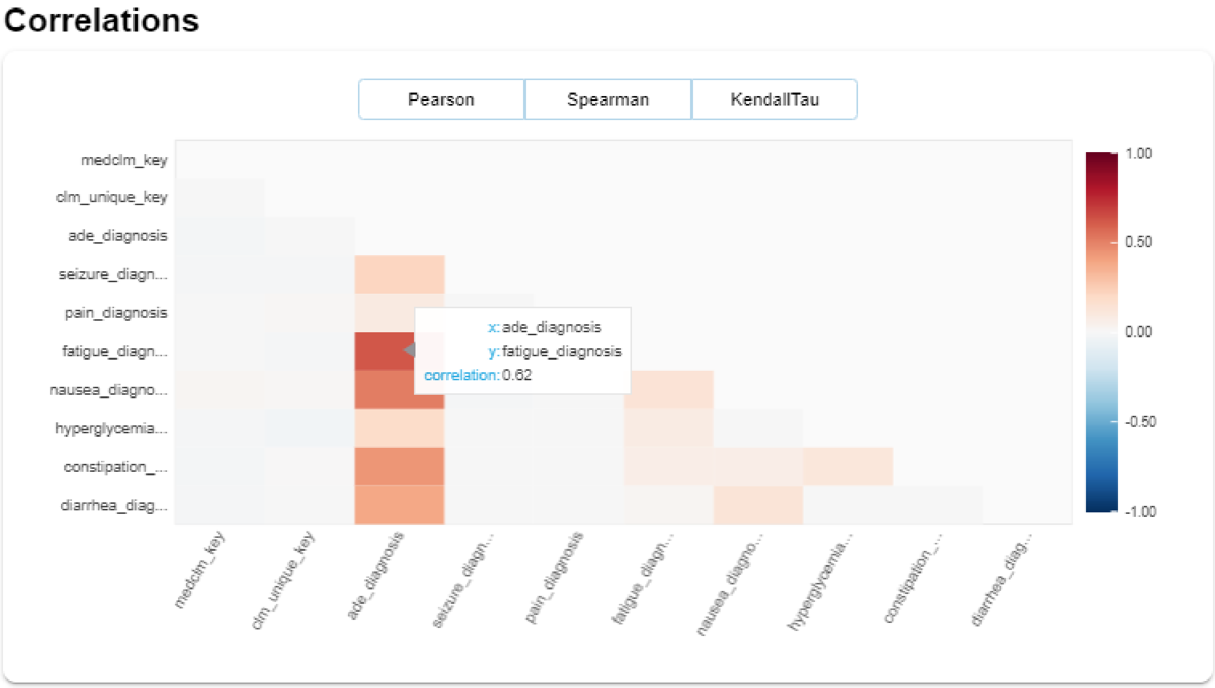
Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	6.3 MB



diarrhea_diagnosis
categorical

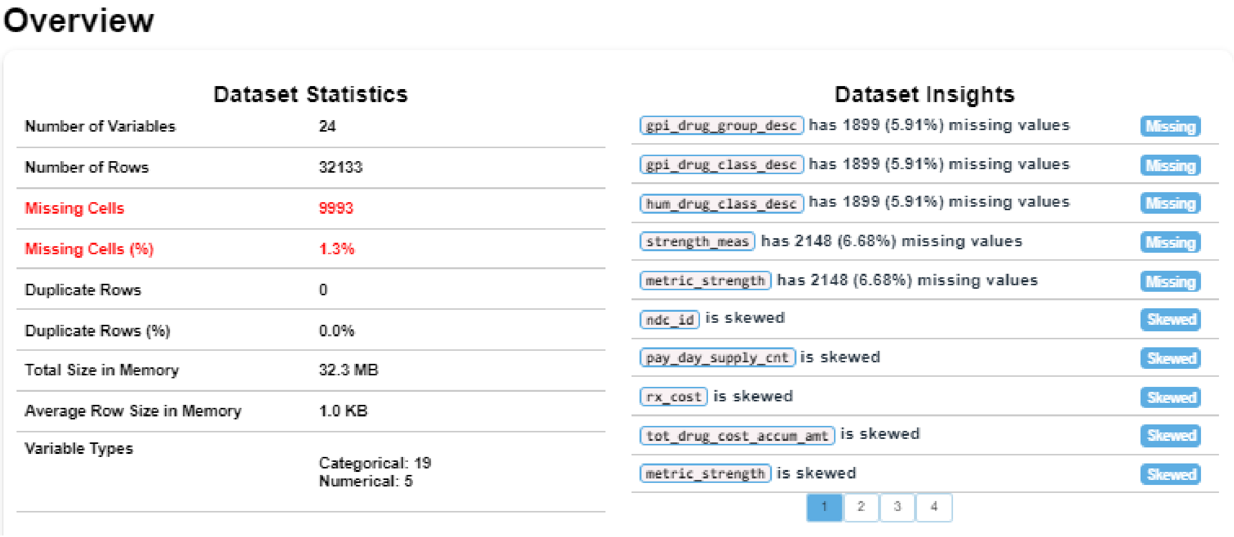
Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	6.3 MB

In the insurance file, we see a moderately positive correlation between fatigue_diagnosis and the ade_diagnosis, which may come in handy when we consider feature importances and recommendations.



Correlations

Pharmacy File:



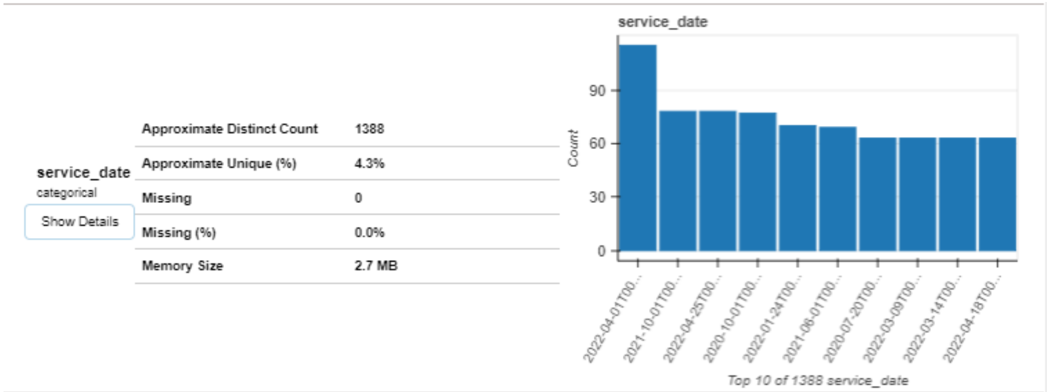
Overview

Dataset Statistics

Number of Variables	24
Number of Rows	32133
Missing Cells	9993
Missing Cells (%)	1.3%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	32.3 MB
Average Row Size in Memory	1.0 KB
Variable Types	Categorical: 19 Numerical: 5

Dataset Insights

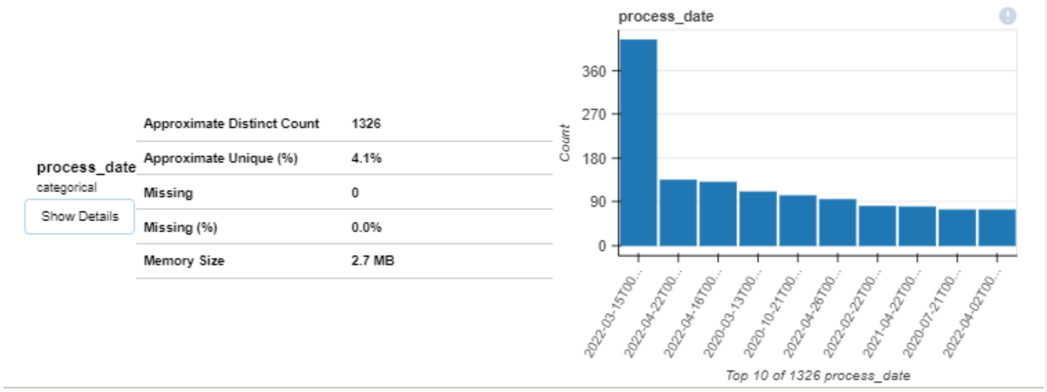
gpi_drug_group_desc has 1899 (5.91%) missing values	Missing
gpi_drug_class_desc has 1899 (5.91%) missing values	Missing
hum_drug_class_desc has 1899 (5.91%) missing values	Missing
strength_meas has 2148 (6.68%) missing values	Missing
metric_strength has 2148 (6.68%) missing values	Missing
ndc_id is skewed	Skewed
pay_day_supply_cnt is skewed	Skewed
rx_cost is skewed	Skewed
tot_drug_cost_accum_amt is skewed	Skewed
metric_strength is skewed	Skewed



service_date
categorical

Show Details

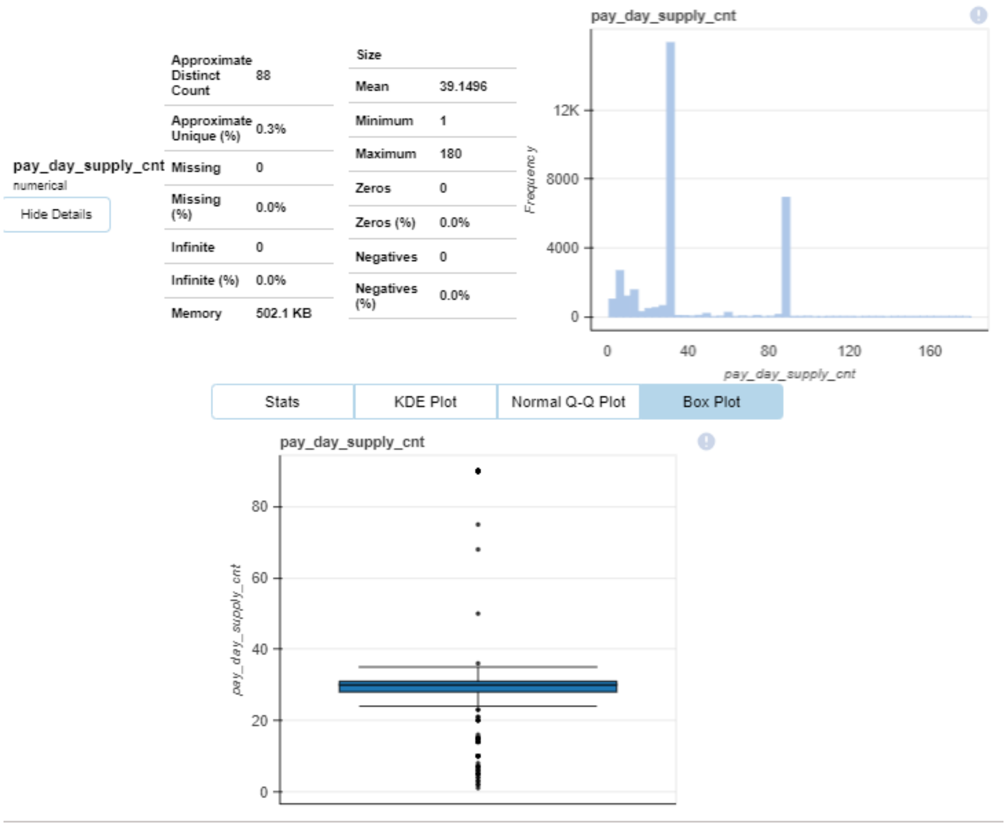
Approximate Distinct Count	1388
Approximate Unique (%)	4.3%
Missing	0
Missing (%)	0.0%
Memory Size	2.7 MB



process_date
categorical

Show Details

Approximate Distinct Count	1326
Approximate Unique (%)	4.1%
Missing	0
Missing (%)	0.0%
Memory Size	2.7 MB

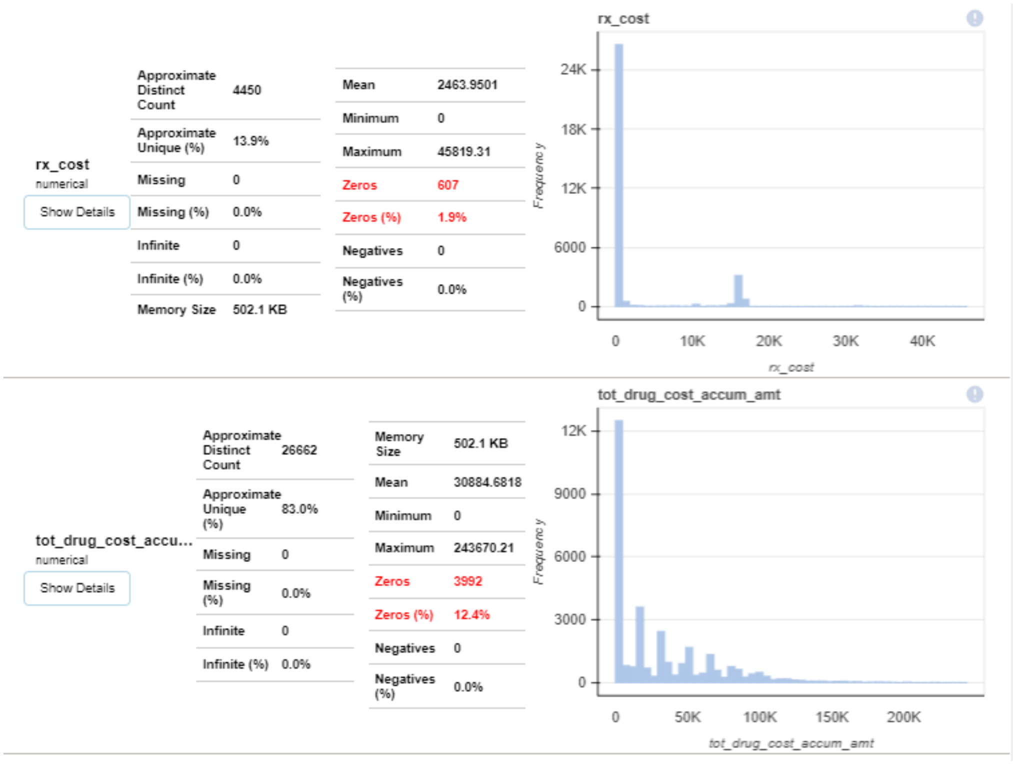


pay_day_supply_cnt
numerical

Hide Details

Approximate Distinct Count	88
Approximate Unique (%)	0.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory	502.1 KB

Size	
Mean	39.1496
Minimum	1
Maximum	180
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



rx_cost
numerical
Show Details

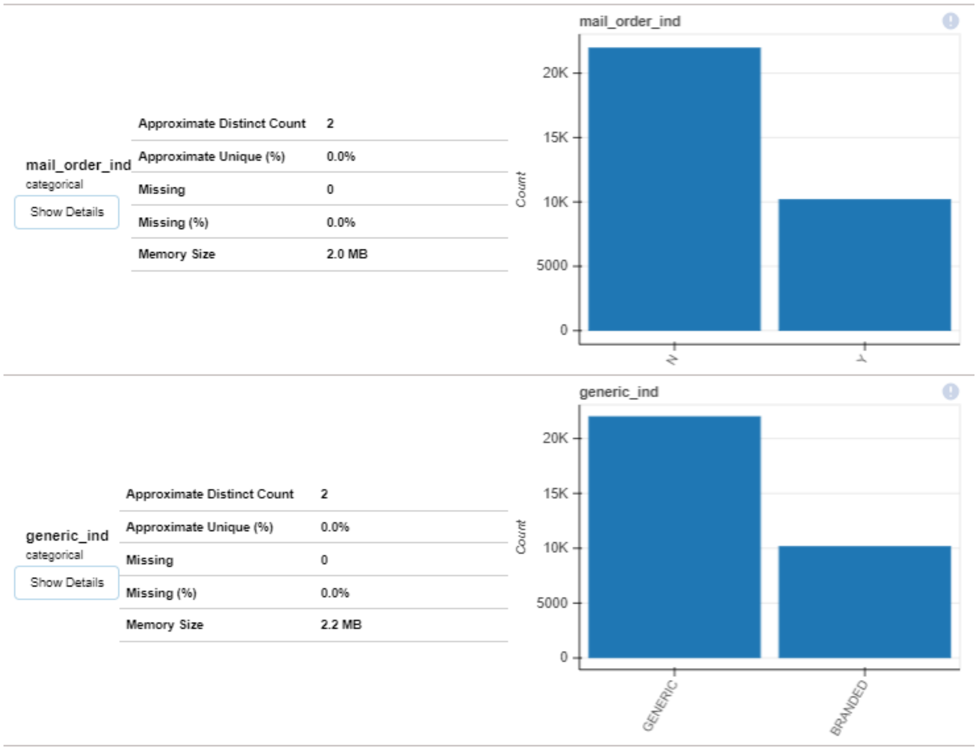
Approximate Distinct Count	4450
Approximate Unique (%)	13.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	502.1 KB

Mean	2463.9501
Minimum	0
Maximum	45819.31
Zeros	607
Zeros (%)	1.9%
Negatives	0
Negatives (%)	0.0%

tot_drug_cost_accu...
numerical
Show Details

Approximate Distinct Count	26662
Approximate Unique (%)	83.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Memory Size	502.1 KB
Mean	30884.6818
Minimum	0
Maximum	243670.21
Zeros	3992
Zeros (%)	12.4%
Negatives	0
Negatives (%)	0.0%

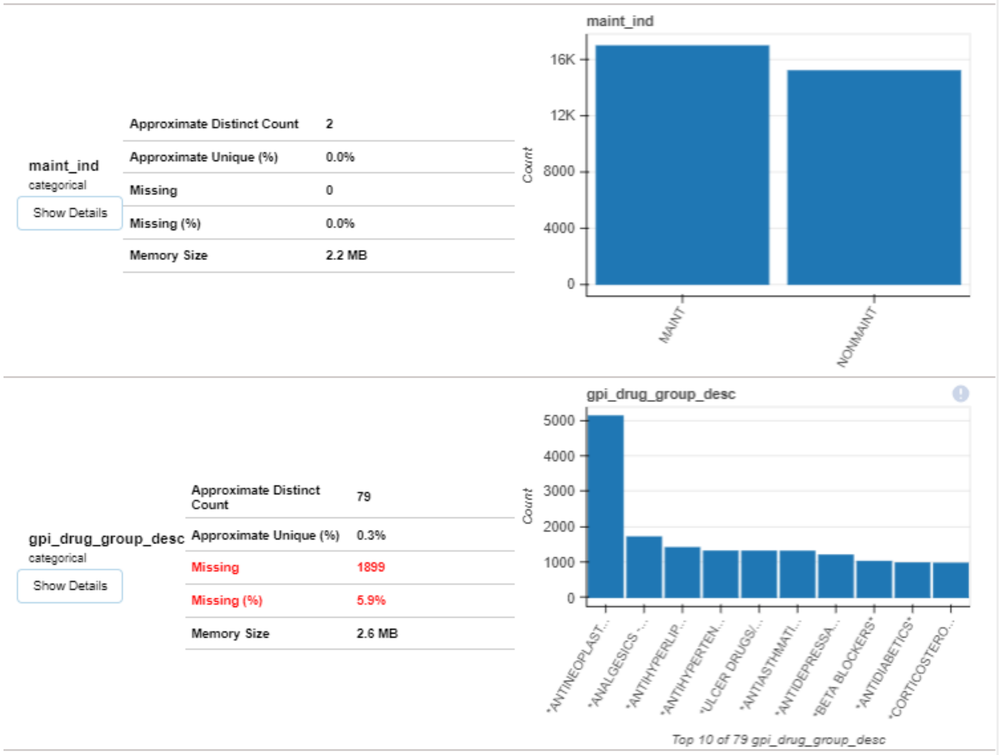


mail_order_ind
categorical
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	2.0 MB

generic_ind
categorical
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	2.2 MB

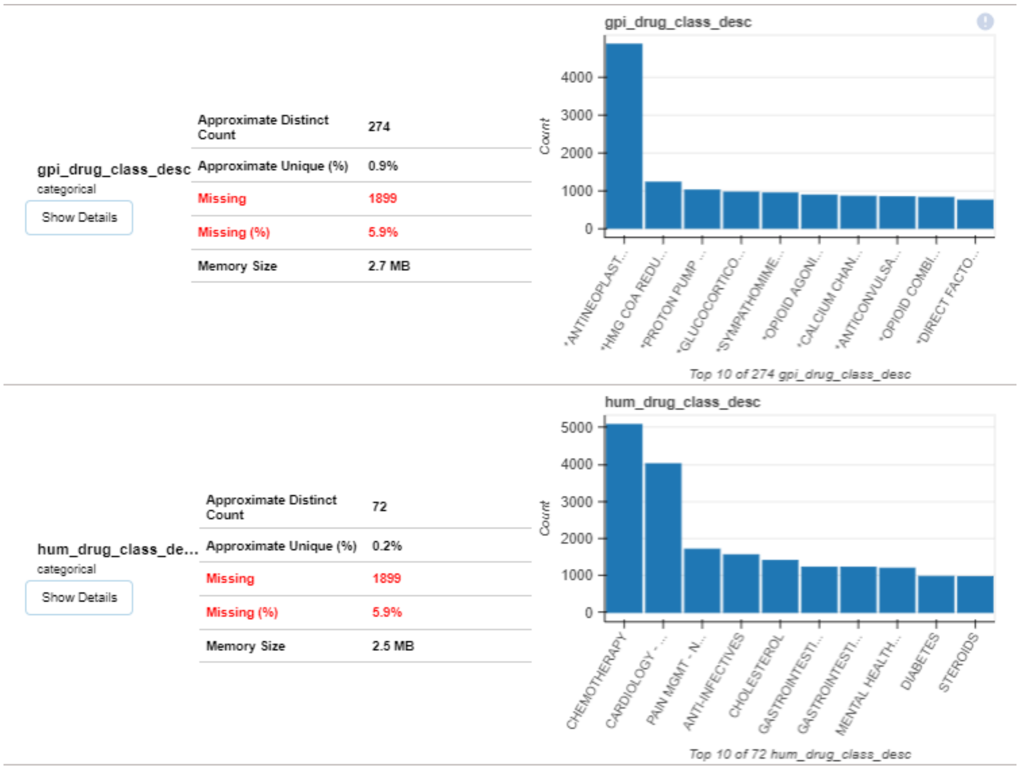


maint_ind
categorical
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	2.2 MB

gpi_drug_group_desc
categorical
Show Details

Approximate Distinct Count	79
Approximate Unique (%)	0.3%
Missing	1899
Missing (%)	5.9%
Memory Size	2.6 MB

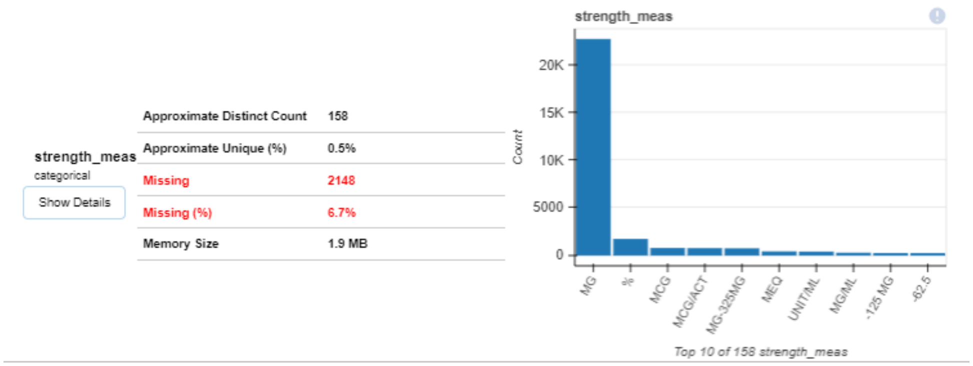


gpi_drug_class_desc
categorical
Show Details

Approximate Distinct Count	274
Approximate Unique (%)	0.9%
Missing	1899
Missing (%)	5.9%
Memory Size	2.7 MB

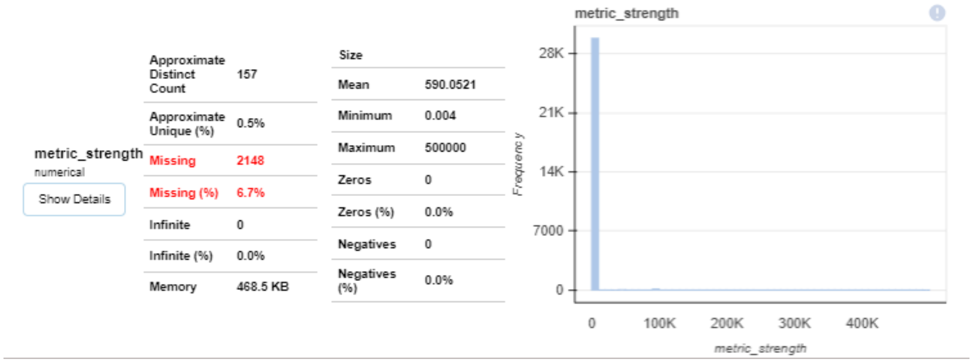
hum_drug_class_de...
categorical
Show Details

Approximate Distinct Count	72
Approximate Unique (%)	0.2%
Missing	1899
Missing (%)	5.9%
Memory Size	2.5 MB



strength_meas
categorical
Show Details

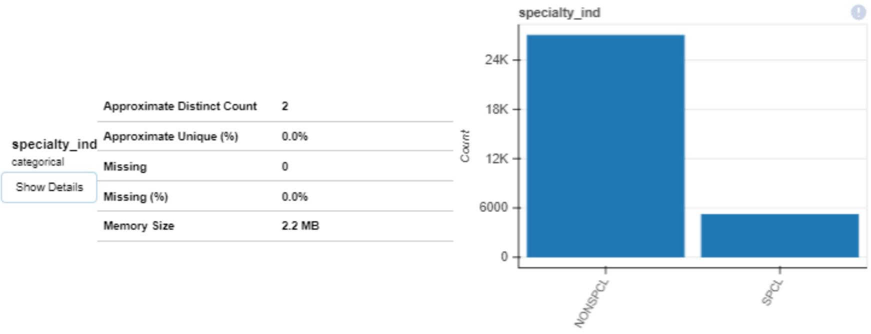
Approximate Distinct Count	158
Approximate Unique (%)	0.5%
Missing	2148
Missing (%)	6.7%
Memory Size	1.9 MB



metric_strength
numerical
Show Details

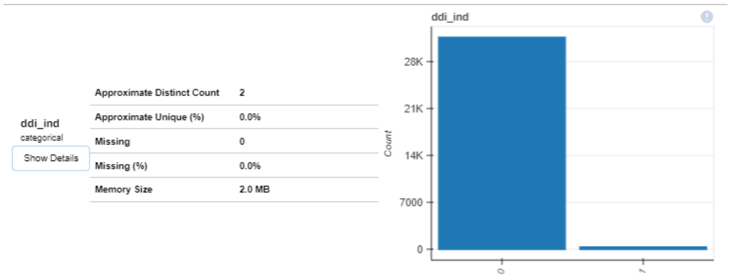
Approximate Distinct Count	157
Approximate Unique (%)	0.5%
Missing	2148
Missing (%)	6.7%
Infinite	0
Infinite (%)	0.0%
Memory	468.5 KB

Size	
Mean	590.0521
Minimum	0.004
Maximum	500000
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%



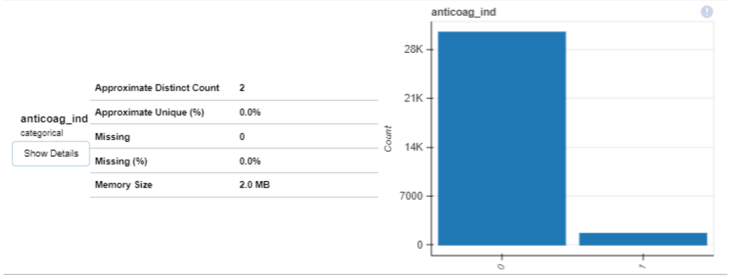
specialty_ind
categorical
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	2.2 MB



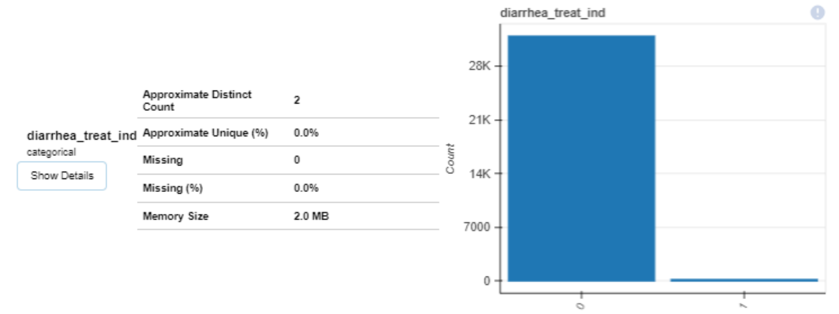
ddi_ind
categorical
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	2.0 MB



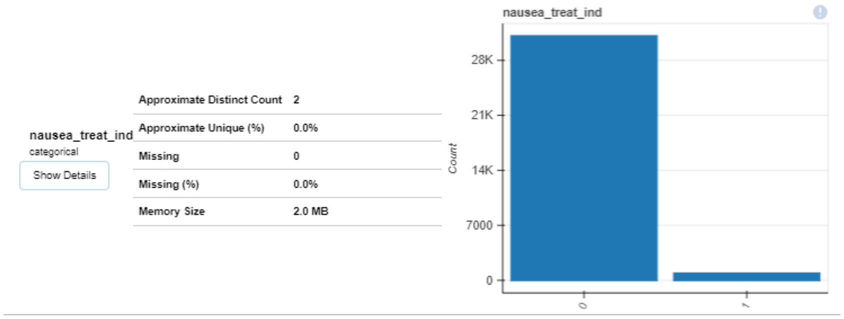
anticoag_ind
categorical
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	2.0 MB



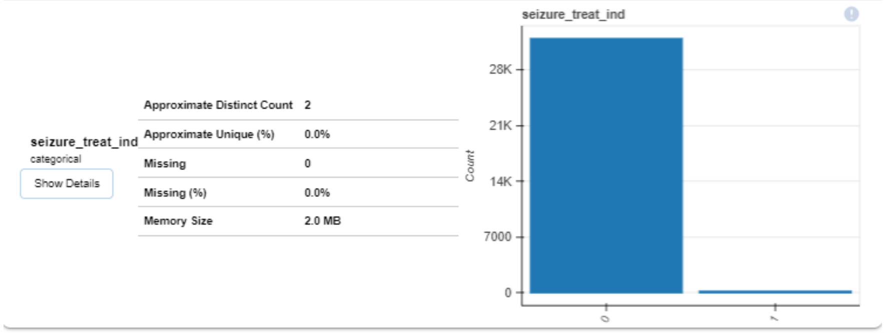
diarrhea_treat_ind
categorical
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	2.0 MB



nausea_treat_ind
categorical
Show Details

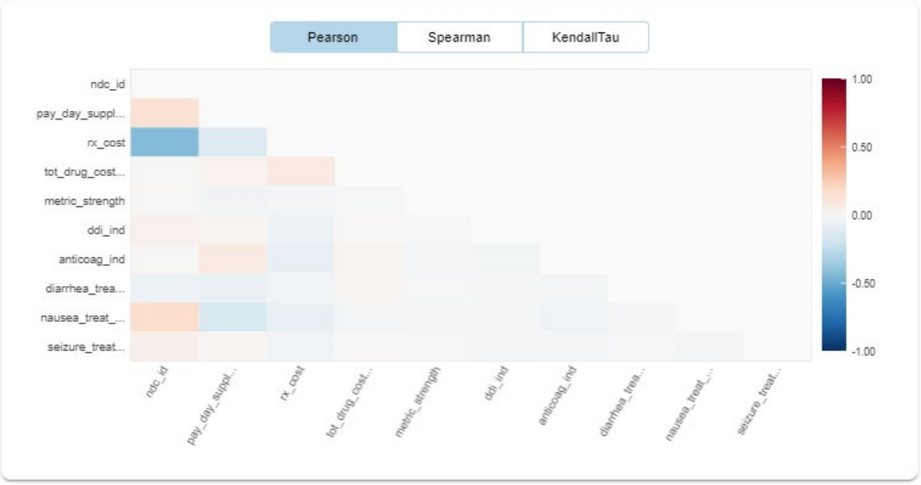
Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	2.0 MB



seizure_treat_ind
categorical
Show Details

Approximate Distinct Count	2
Approximate Unique (%)	0.0%
Missing	0
Missing (%)	0.0%
Memory Size	2.0 MB

Correlations



3.3. Data Cleaning and Imputation

3.3.1. Data Types Transformation

Datetime Conversion: - The columns `therapy_end_date` and `therapy_start_date` were converted to UTC datetime format.

One-hot encoding: The following variables were one-hot encoded to help capture the number of times a particular diagnosis appeared, as well as to prepare for modeling.

Gender (sex_cd), diagnosis ICD 10 codes (primary_diag_cd, diag_cd2 thru diag_cd9), hedis_pot, specialty_ind

3.3.2. Missing Value Imputation

Null Value Identification:

Identify Specific Null Patterns: - We identified rows in the target file where specific columns (`race_cd`, `est_age`, `sex_cd`, `cms_disabled_ind`, `cms_low_income_ind`) had null values simultaneously.

Fill NaN : - For rows where all the aforementioned columns were NaN simultaneously, We filled `race_cd`, `cms_disabled_ind`, and `cms_low_income_ind` with 0, so as not to assume any particular race, disability or low income status.
Once the three files were merged, any new feature engineered variables that were nulls were all filled with 0s.
Missing values in Training data
<u>**Target**</u>

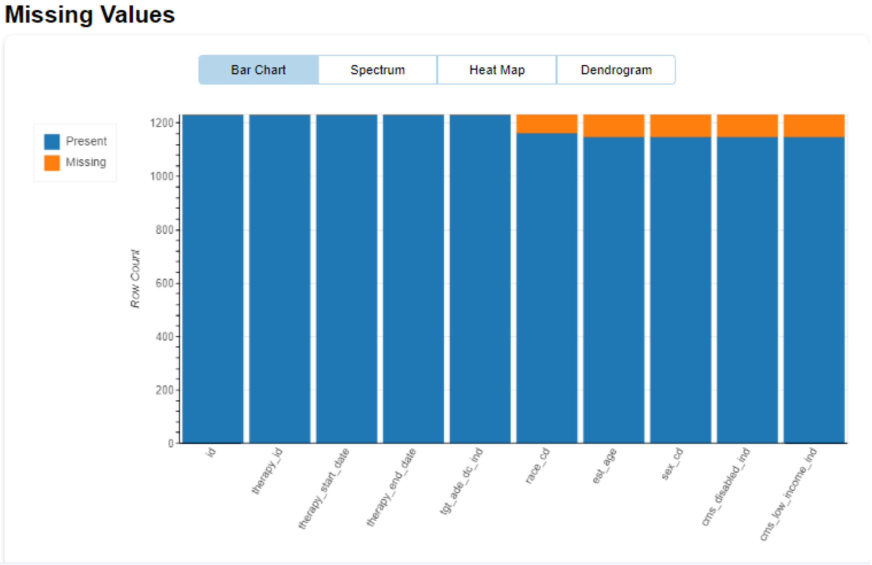


Table 4 Missing values in target file

Insurance Claims There are a number of missing values in the secondary diagnoses variables, as well as the util_cat (which was not used in the model). We resolved this by condensing the records per therapy id, and creating some feature engineered one-hot encoded variables that would indicate the number of diagnoses.

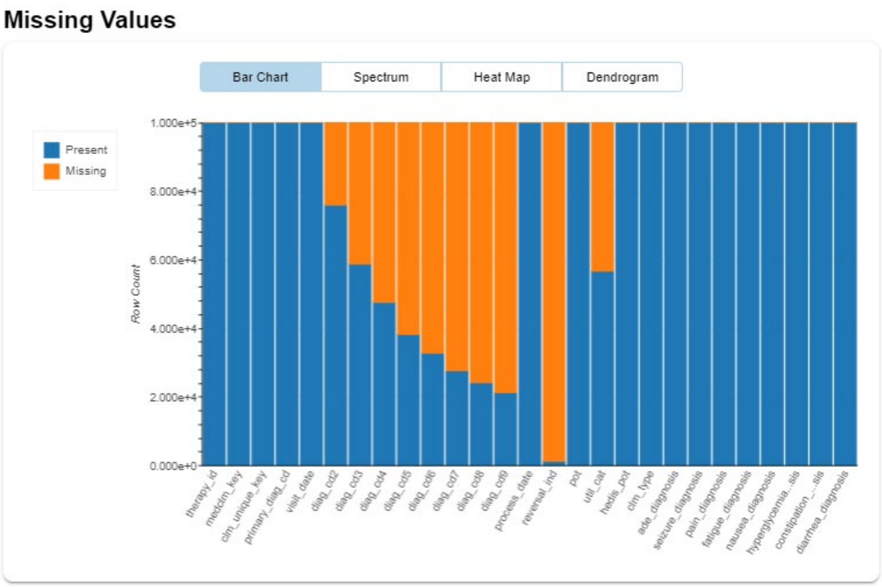


Table 5 Missing values in insurance file

Pharmacy Claims: The number of missing values in the pharmacy claims file is very low and were not considered in the modeling.

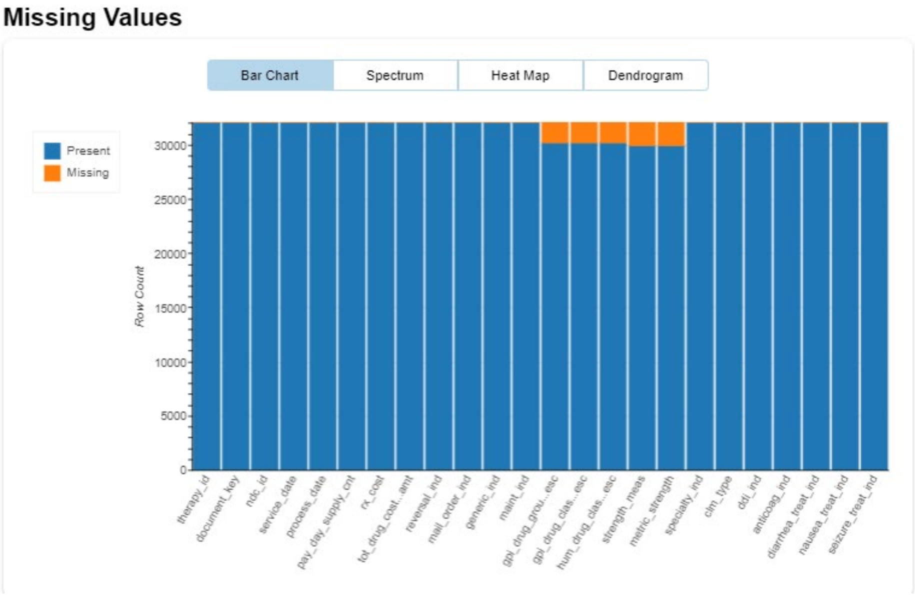


Table 6 Missing values in pharmacy file.

3.4. Feature Engineering

Feature engineering was performed on each of the three tables to get better insights and to be able to collapse multiple rows into one single row per patient therapy_id. The final tables, with newly feature engineered variables, along with their descriptions and data types are shown below.

Target file:

Column Name	Description	D Type
Id	Unique identifier for the data entry.	Int
therapy_id	Identifier for the therapy.	Object
tgt_ade_dc_ind	Target adverse drug event indicator.	Int
race_cd	Code representing the race of the individual.	Int
est_age	Estimated age of the individual.	Float
sex_cd	Code representing the sex of the individual.	Int
cms_disabled_ind	Indicator for disability status.	Int
cms_low_income_ind	Indicator for low-income status.	Int
therapy_start_mth	Starting month (cardinal number) of therapy	Int
therapy_start_dayOfWk	Starting day (1-7) of the week when therapy started.	Int
therapy_started_mth_beginning	Indicator if therapy started within first 15 days of the month	Int

Table 7 Target file structure

Medical Claims:

Column Name	Description	D Type
therapy_id	Patient therapy ID (primary key)	Object
Number_of_Ins_Claims	Number (count) of insurance claims	Int
Number_of_Primary_Diagnoses	Number of primary diagnoses	Int

primary_Certain infections and parasitic diseases	Number of primary diagnoses for Certain infections and parasitic diseases	Int
primary_Neoplasms	Number of primary diagnoses for Neoplasms	Int
primary_Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	Number of primary diagnoses for Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	Int
primary_Endocrine, nutritional and metabolic diseases	Number of primary diagnoses for Endocrine, nutritional and metabolic diseases	Int
primary_Mental, Behavioral and Neurodevelopmental disorders	Number of primary diagnoses for Mental, Behavioral and Neurodevelopmental disorders	Int
primary_Diseases of the nervous system	Number of primary diagnoses for Diseases of the nervous system	Int
primary_Diseases of the eye and adnexa	Number of primary diagnoses for Diseases of the eye and adnexa	Int
primary_Diseases of the ear and mastoid process	Number of primary diagnoses for Diseases of the ear and mastoid process	Int
primary_Diseases of the circulatory system	Number of primary diagnoses for Diseases of the circulatory system	Int
primary_Diseases of the respiratory system	Number of primary diagnoses for Diseases of the respiratory system	Int
primary_Diseases of the digestive system	Number of primary diagnoses for Diseases of the digestive system	Int
primary_Diseases of the skin and subcutaneous tissue	Number of primary diagnoses for Diseases of the skin and subcutaneous tissue	Int
primary_Diseases of the musculoskeletal system and connective tissue	Number of primary diagnoses for Diseases of the musculoskeletal system and connective tissue	Int
primary_Diseases of the genitourinary system	Number of primary diagnoses for Diseases of the genitourinary system	Int
primary_Pregnancy, childbirth, and puerperium	Number of primary diagnoses for Pregnancy, childbirth, and puerperium	Int
primary_Certain conditions originating in the perinatal period	Number of primary diagnoses for Certain conditions originating in the perinatal period	Int
primary_Congenital malformations, deformations and chromosomal abnormalities	Number of primary diagnoses for Congenital malformations, deformations and chromosomal abnormalities	Int
primary_Symptoms, signs, and abnormal clinical laboratory findings, not elsewhere classified	Number of primary diagnoses for Symptoms, signs, and abnormal clinical laboratory findings, not elsewhere classified	Int
primary_Injury, poisoning, and certain other consequences of external causes	Number of primary diagnoses for Injury, poisoning, and certain other consequences of external causes	Int
primary_External causes of morbidity	Number of primary diagnoses for External causes of morbidity	Int

primary_Factors influencing health status and contact with health services	Number of primary diagnoses for Factors influencing health status and contact with health services	Int
most_common_Certain infections and parasitic diseases	Indicator if this primary diagnosis was the most common	Int
most_common_Neoplasms	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	Indicator if this primary diagnosis was the most common	Int
most_common_Endocrine, nutritional and metabolic diseases	Indicator if this primary diagnosis was the most common	Int
most_common_Mental, Behavioral and Neurodevelopmental disorders	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the nervous system	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the eye and adnexa	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the ear and mastoid process	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the circulatory system	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the respiratory system	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the digestive system	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the skin and subcutaneous tissue	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the musculoskeletal system and connective tissue	Indicator if this primary diagnosis was the most common	Int
most_common_Diseases of the genitourinary system	Indicator if this primary diagnosis was the most common	Int
most_common_Pregnancy, childbirth, and puerperium	Indicator if this primary diagnosis was the most common	Int
most_common_Certain conditions originating in the perinatal period	Indicator if this primary diagnosis was the most common	Int
most_common_Congenital malformations, deformations and chromosomal abnormalities	Indicator if this primary diagnosis was the most common	Int
most_common_Symptoms, signs, and abnormal clinical laboratory findings, not elsewhere classified	Indicator if this primary diagnosis was the most common	Int

most_common_Injury, poisoning, and certain other consequences of external causes	Indicator if this primary diagnosis was the most common	Int
most_common_External causes of morbidity	Indicator if this primary diagnosis was the most common	Int
most_common_Factors influencing health status and contact with health services	Indicator if this primary diagnosis was the most common	Int
Avg_days_since_last_visit	Average number of days elapsed since last visit	Int
avg_medical_visits	Average number of medical visits (inc. all visit types)	Int
average_days_to_process_claim	Average number of days to process claim	Int
N_ER_visits	Number of visits to the ER	Int
N_Outpatient_visits	Number of outpatient visits	Int
N_Other_visits	Number of other types of visits	Int
N_Inpatient_visits	Number of inpatient visits	Int
N_Observation_visits	Number of visits by observation	Int
N_Telephone_visits	Number of visits by telephone	Int
N_ade_diagnosis	Number of times ADE was diagnosed per patient	Int
N_seizure_diagnosis	Number of times seizure was diagnosed	Int
N_pain_diagnosis	Number of times pain was diagnosed	Int
N_fatigue_diagnosis	Number of times fatigue was diagnosed	Int
N_nausea_diagnosis	Number of times nausea was diagnosed	Int
N_hyperglycemia_diagnosis	Number of times hyperglycemia was diagnosed	Int
N_constipation_diagnosis	Number of times constipation was diagnosed	Int
N_diarrhea_diagnosis	Number of times diarrhea was diagnosed	Int
Overall_Certain infections and parasitic diseases	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int

Overall_Neoplasms	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Endocrine, nutritional and metabolic diseases	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Mental, Behavioral and Neurodevelopmental disorders	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the nervous system	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the eye and adnexa	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the ear and mastoid process	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the circulatory system	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the respiratory system	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the digestive system	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the skin and subcutaneous tissue	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the musculoskeletal system and connective tissue	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Diseases of the genitourinary system	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Pregnancy, childbirth, and puerperium	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Certain conditions originating in the perinatal period	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Congenital malformations, deformations and chromosomal abnormalities	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Symptoms, signs, and abnormal clinical laboratory findings, not elsewhere classified	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Injury, poisoning, and certain other consequences of external causes	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_External causes of morbidity	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int
Overall_Factors influencing health status and contact with health services	Number of times this diagnosis was present in primary as well as all the secondary diagnoses	Int

Table 8 Insurance claims file structure

Pharmacy Claims:

Column Name	Description	D Type
therapy_id	Patient therapy ID (primary key)	Object
average_process_days_from_service	Average number of days it took to process the pharmacy claim from date of service	Int
Special_ind_N	Number of claims for specialty drug	Int
N_knownInteractionsDrug	Number of claims for known interactions with Tagrisso	Int
cum_cost	Cumulative cost for the duration of the therapy	Int
N_anticoag	Number of claims for an anticoagulant	Int
N_diarrhea_treat	Number of claims for diarrhea treatment	Int
N_nausea_treat	Number of claims for nausea treatment	Int
N_seizure_treat	Number of claims for seizure treatment	Int
comorbidities	Indicator of any comorbidities	Int

Table 9 Pharmacy claims file structure

4. METHODOLOGY

Once the feature engineering was done, we combined the three datasets – target_train, medclms (insurance) and rxclms (pharmacy). All the records in the insurance and pharmacy claims files were narrowed down to *only include records for each therapy_id that had visit date and service date* ***after*** *the patient's therapy started*. This was done to keep track of the period of activity after therapy started, like number of visits (outpatient, ER, etc.), number of primary diagnoses, and other factors.

Individual lookbacks and any diagnoses prior to the start of the therapy were treated as the existence of comorbidities in a patient's record.

The most recent year-to-date amount for a member once therapy started was picked as the cumulative cost incurred by a patient with the assumption that the therapy was still ongoing. Mapping was created for the broad diagnoses to the ICD 10 codes, so the categories were more layperson friendly.

ICD-10 Code	**Diagnosis**
A00-B99	Certain infections and parasitic diseases
C00-D49	Neoplasms
D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
E00-E89	Endocrine, nutritional, and metabolic diseases
F01-F99	Mental, Behavioral and Neurodevelopmental disorders
G00-G99	Diseases of the nervous system
H00-H59	Diseases of the eye and adnexa
H60-H95	Diseases of the ear and mastoid process
I00-I99	Diseases of the circulatory system
J00-J99	Diseases of the respiratory system
K00-K95	Diseases of the digestive system
L00-L99	Diseases of the skin and subcutaneous tissue
M00-M99	Diseases of the musculoskeletal system and connective tissue
N00-N99	Diseases of the genitourinary system
O00-O9A	Pregnancy, childbirth, and puerperium
P00-P96	Certain conditions originating in the perinatal period
Q00-Q99	Congenital malformations, deformations, and chromosomal abnormalities
R00-R99	Symptoms, signs, and abnormal clinical laboratory findings, not elsewhere classified
S00-T88	Injury, poisoning, and certain other consequences of external causes
V00-Y99	External causes of morbidity
Z00-Z99	Factors influencing health status and contact with health services

Table 10 ICD 10 to diagnoses crosswalk

Using the above mapping from the ICD 10 code to the diagnoses, we were able to feature engineer additional variables for primary diagnoses like Neoplasms, and other symptoms.

5. MODELING

5.1. Model Selection

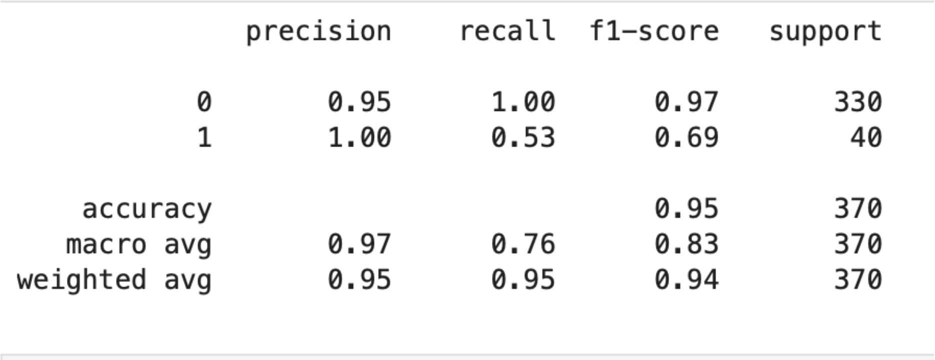
After we selected key factors with importance greater than zero, the model selection was the key to our accuracy of prediction. Our objective is to predict members most likely to withdraw from their treatment. First, we identified this as a classification prediction question.
Second, since we have a high dimensional dataset (after merging target dataset with medclaims and rxclaims dataset), we needed to choose a model with good flexibility, high predictive power and easy explainability. So, we mainly focused on tree-based algorithms to estimate the probability of members likely to withdraw from the treatment.
Therefore, we selected these four models: Decision Tree, Random Forest, Gradient Boosting Decision Tree, and XGBoost to do preliminary prediction and compared their performances after that.
Out of 1232 observations in the original training dataset, we split it into 70% as training data, 30% as testing data. Then, we performed cross- validation with 70% training data and tested all models' performances separately on 30% testing data. The evaluation metric used was the AUC-score, which measures the area under the Receiver Operating Curve (ROC) and generally reflects how well the model can distinguish the classes.
Finally, out of all the models we tested, Random Forest has the best performance of 0.9727 in terms of AUC score; Gradient Boosting Decision Tree had an AUC of 0.932; XGBoost returned a similar score of 0.932; Decision Tree returned a score of 0.903.

Model	AUC score
Random Forest	0.9727
Gradient Boosting Decision Tree	0.932
XGBoost	0.932
Decision Tree	0.903

Table 11 Models to predict members most likely to experience an ADE and discontinue therapy

5.2. Final Model Construction

Based on the ROC metric for the Random Forest model below, the AUC score of the Random Forest Classifier is 0.9727 and outperforms the rest, so we decided to use Random Forest to predict on our holdout dataset. Besides excellent prediction performance and fast processing speed, the Random Forest can deal with the imbalanced problem existing in our dataset, where out of 1232 observations, only 9.5% of members (117) have withdrawn or discontinued from the Osimertinib therapy, and most members are continuing the treatment. As Random Forest shakes the data each time, it handles the forementioned imbalance problem very well. To better analyze the performance of our model, we also calculated the confusion matrix and created a classification report with precision, recall and F1-score for both classes: 0, 1. From the ROC curve below, we can see that the true positive rate for both classes {0,1} is around 90% which shows the excellent performance of our model.



```
              precision    recall  f1-score   support

           0       0.95      1.00      0.97       330
           1       1.00      0.53      0.69        40

    accuracy                           0.95       370
   macro avg       0.97      0.76      0.83       370
weighted avg       0.95      0.95      0.94       370
```

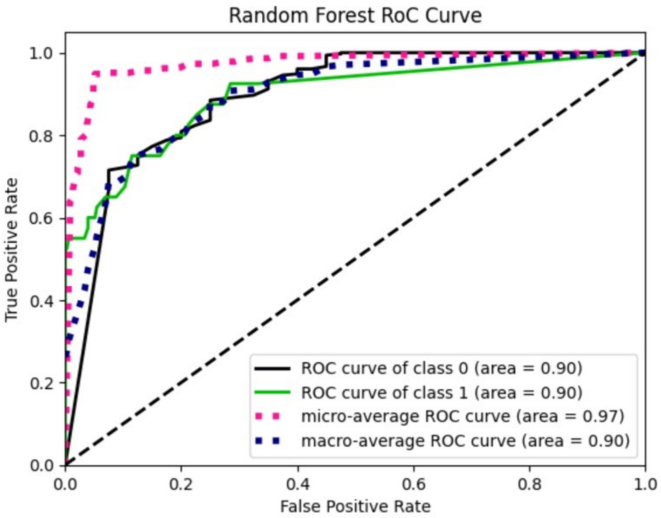


Table 12 Accuracy, Precision-Recall, AUC & ROC for Random Forest model

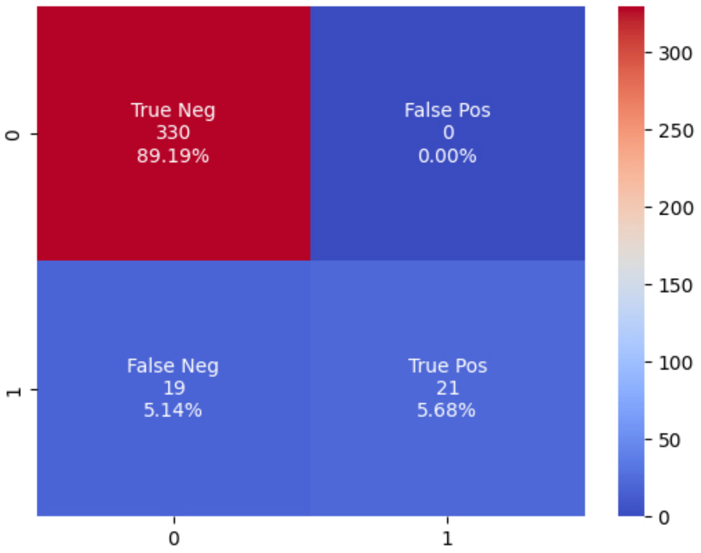


Table 13 Confusion Matrix for Random Forest model

To improve the model's performance. We did parameter tuning and found most of the parameters did well with their default values except:

a. '**n_estimators'**: **73**. n_estimator is the hyperparameter that defines the number of trees to be used in the model. The tree can also be understood as the sub-divisions.

6. KEY PERFORMANCE INDICATOR ANALYSIS

6.1. Feature Importance

To better understand the model and important features, and drive insights from the model. We looked up the top 30 important features in XGBoost gain importance and top 20 important SHAP values.

Gini Importance:

We used the built-in Random Forest feature importance function (Gini importance) to get the most important features after tuning and training the model. The calculated average numerical value of "information gain" or decrease in "gini impurity" to take each feature's contribution to each tree in the model is the most common method to evaluate the importance of the features in the model. The top 30 important features of gain importance are shown in the following figure:

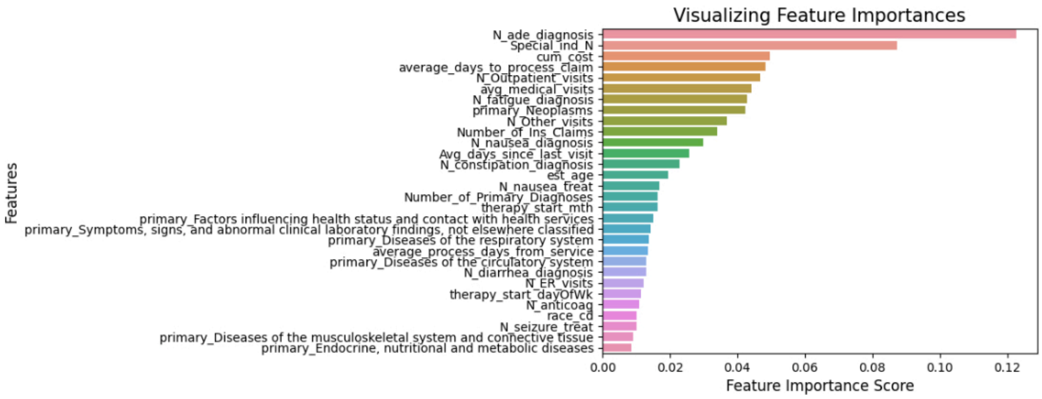


Table 14 Feature Importance list to predict target outcome

SHAP Value:

In our feature importance analysis, SHAP is also a well-known method in post-model analysis to compare and analyze the final features, since it generates numeric values which can be used to calculate the important role of the features to the model. The top 20 features of Shap value are shown in the following figure:

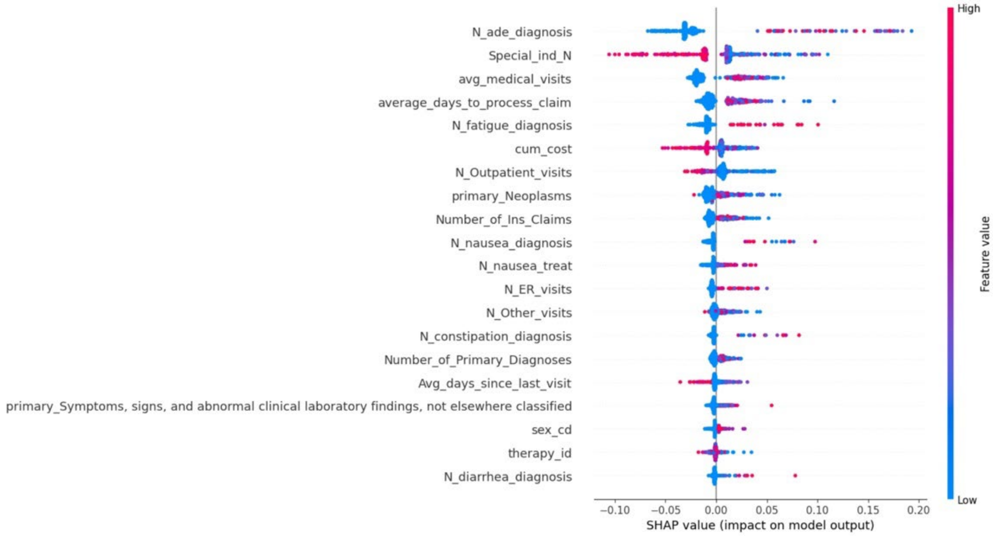


Table 15 SHAP values to show positive/negative impact on outcome variable

As can be seen in the gain importance and SHAP value figures, some features stand out and have high importance in both figures. And comparing the aspects of all variables, the features can be categorized as the following:

1. Adverse Drug Event Factor: 'N_ADE_DIAGNOSIS' is the variable ranked first in both figures, which is the most important feature in our Random Forest model. We can see that the member having high number of adverse drug events in conjunction with other factors, is highly related to withdrawal of the treatment.

2. Specialty drug Factor: ‘SPECIAL_IND_N’ is the second important feature in terms of both importance and SHAP value. The variable indicates the number of claims filed for the specialty drug. We can see that members filed higher number of specialty drug claims, are less likely to withdraw from the treatment.

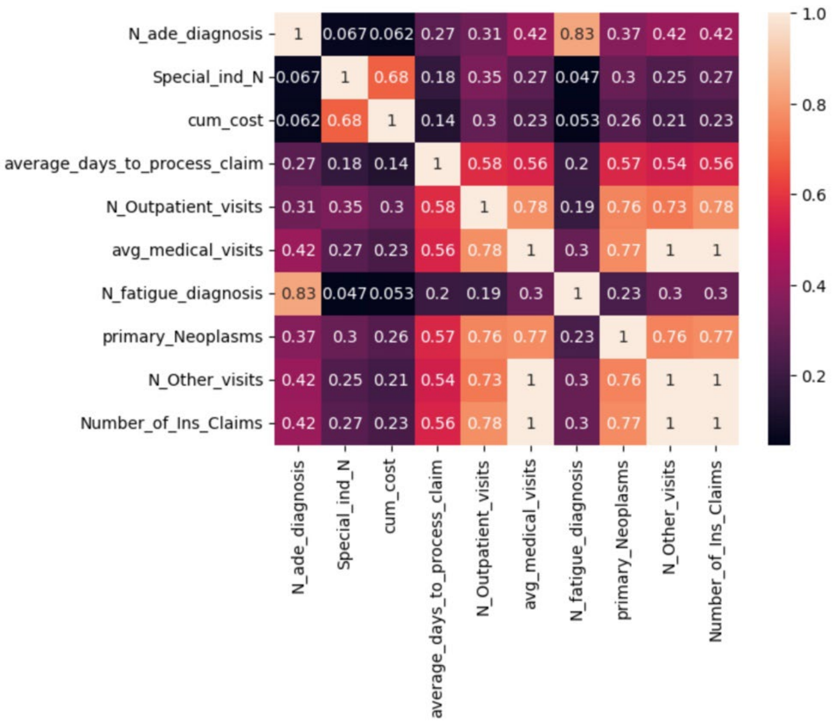
3. Health (Other Diagnosis Factor): ‘N_FATIGUE_DIAGNOSIS’ is the variable which is in the Top-6 in both figures which indicates the number of times a patient is diagnosed with fatigue. There are other diagnosis variables like ‘N_NAUSEA_DIAGNOSIS’ and ‘N_DIARRHEA_DIAGNOSIS’ that are present in the Top-30 important variables. This shows that patients who experienced these effects, are potential candidates that could withdraw from the treatment.

4. The Financial Factor: ‘CUM_COST’ is the variable which indicates the total cumulative drug cost and is in the Top-5 in both figures. We can see that the higher the cumulative cost, members are less likely to withdraw from the treatment. It could be because higher drug costs are covered by insurance once deductibles are accounted for (although no deductibles were mentioned, our assumption is every patient has to pay out of pocket for the deductible). It could also indicate that since the patients are continuing their therapy, the costs are going to increase, which makes sense.

5. Demographic factor: ‘SEX_CD’ is the variable which indicates member’s sex. We can see that Male members (sex_cd=1) are more likely to withdraw from the treatment as compared to Female members.

6.2. Relationship between Factors

To further analyze the important features and the relationship between these factors, we generated a heatmap to see the correlation relationship of each factor. As can be shown in the following heatmap, Fatigue_diagnosis is highly positively related to the ADE_diagnosis which is the most influential factor. Also, average_days_to_process_claim and avg_medical_visits are positively related to ADE_diagnosis.

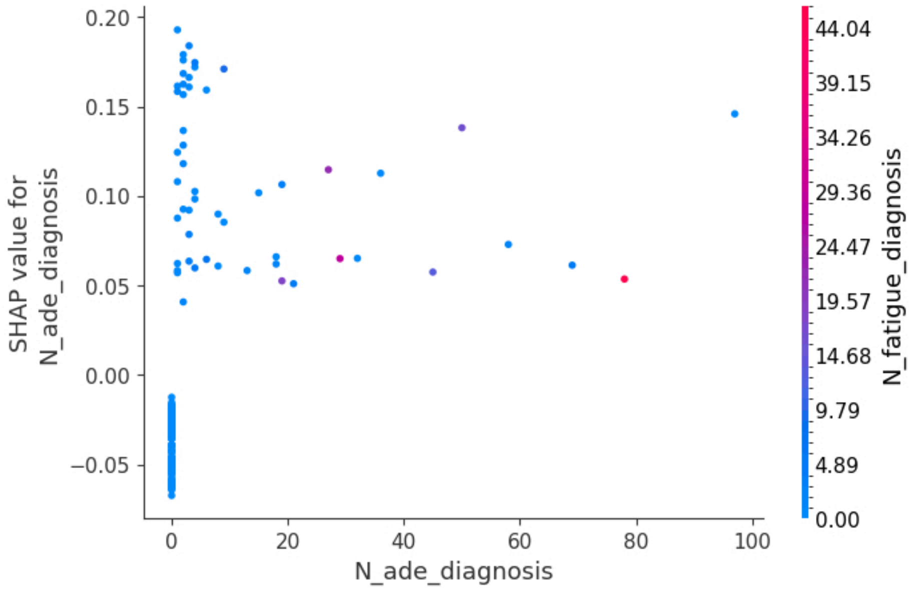


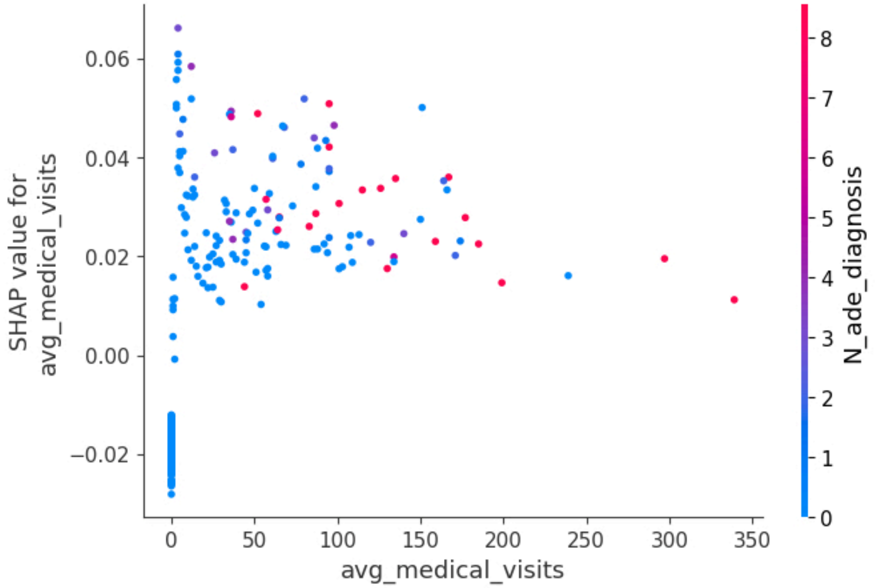
	N_ade_diagnosis	Special_ind_N	cum_cost	average_days_to_process_claim	N_Outpatient_visits	avg_medical_visits	N_fatigue_diagnosis	primary_Neoplasms	N_Other_visits	Number_of_Ins_Claims
N_ade_diagnosis	1	0.067	0.062	0.27	0.31	0.42	0.83	0.37	0.42	0.42
Special_ind_N	0.067	1	0.68	0.18	0.35	0.27	0.047	0.3	0.25	0.27
cum_cost	0.062	0.68	1	0.14	0.3	0.23	0.053	0.26	0.21	0.23
average_days_to_process_claim	0.27	0.18	0.14	1	0.58	0.56	0.2	0.57	0.54	0.56
N_Outpatient_visits	0.31	0.35	0.3	0.58	1	0.78	0.19	0.76	0.73	0.78
avg_medical_visits	0.42	0.27	0.23	0.56	0.78	1	0.3	0.77	1	1
N_fatigue_diagnosis	0.83	0.047	0.053	0.2	0.19	0.3	1	0.23	0.3	0.3
primary_Neoplasms	0.37	0.3	0.26	0.57	0.76	0.77	0.23	1	0.76	0.77
N_Other_visits	0.42	0.25	0.21	0.54	0.73	1	0.3	0.76	1	1
Number_of_Ins_Claims	0.42	0.27	0.23	0.56	0.78	1	0.3	0.77	1	1

Table 16 Correlation matrix of top features

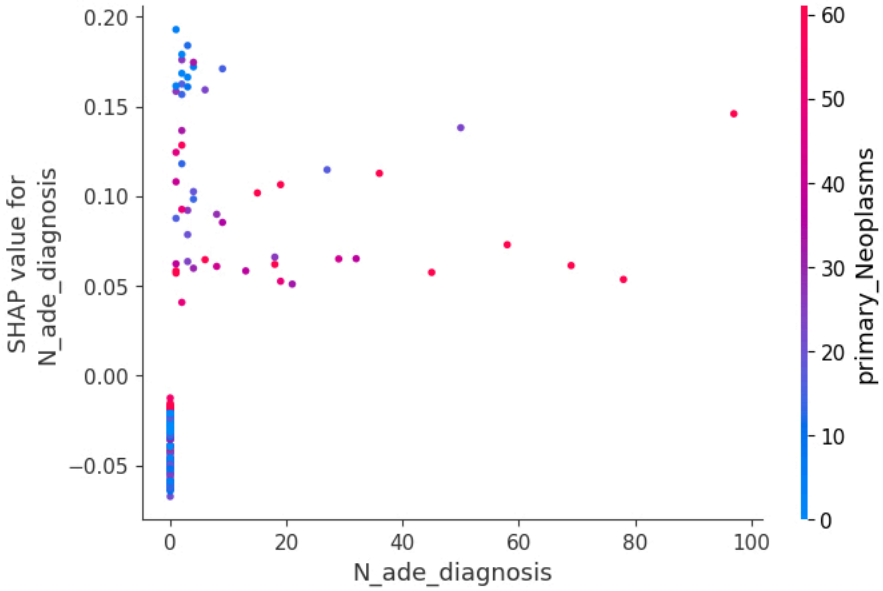
We also used SHAP dependence plots to study the individual effects and interaction effects of key variables:

- N_ade_diagnosis: The following two dependency plots show the relationship between N_ade_diagnosis and N_fatigue_diagnosis (health_factor), avg_medical_visits. We can see that the N_ade_diagnosis correlates with increase of N_fatigue_diagnosis and avg_medical_visists correlate with increase of N_ade_diagnosis. This is aligned with the observation we figured out and information shown in figure 3 and it is reasonable to explain the positive relationship among N_Fatigue_diganosis, N_ade_dignosis, avg_medical_visists and member likely to withdraw from the treatment.





- primary_Neoplasms: From the following dependency plot between primary_Neoplasms and N_ade_diagnosis, we can see that N_ade_diagnosis is positively correlates with primary_Neoplasms. The higher the primary_Neoplasms, the higher number of ADE_diagnosis which makes member more likely to withdraw from the treatment.



7. SEGMENTATION

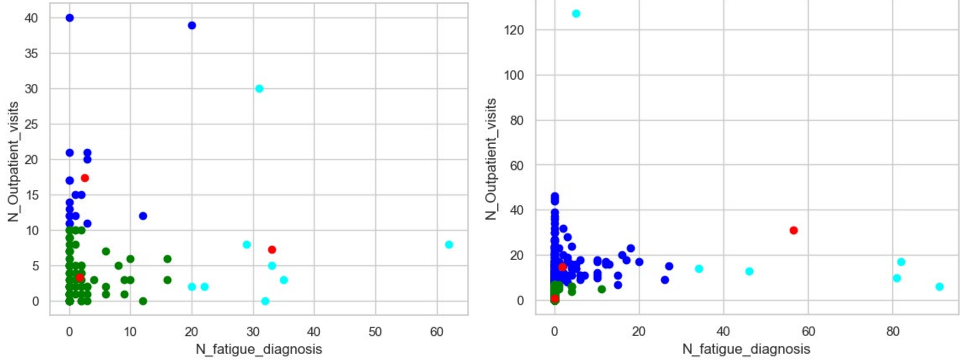
In analyzing the data, we wanted to look for any clusters or patterns that might exist, especially with race, gender, and disability.

We looked at clustering for different combinations, like cumulative costs by age and race (White and Black), as well as the number of fatigue diagnosis and outpatient visits for both the continued & discontinued populations.

Segment 1 Features:

Number of outpatient visits, and number of fatigue diagnoses.

The number of clusters for the discontinued population was 4, and for the completed population was 3.



Discontinued therapy population   Completed therapy population

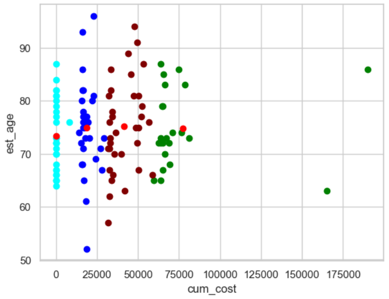
Although the clusters are not very clearcut in the above plots, they show that for the discontinued population, the number of outpatient visits were a lot higher when the patient had 0 to 5 diagnoses of fatigue, but dipped considerably when the number of fatigue diagnoses for a patient was more than 5 times. It might be interesting to see if these patients chose an

alternative visit type before they discontinued therapy. In other words, tracking the switch between visit types might yield some insights, and potential actionable items on helping patients deal with fatigue or other symptoms.
For the patients completing therapy, the increase in the number of fatigue diagnoses did see a decrease in the number of outpatient visits.

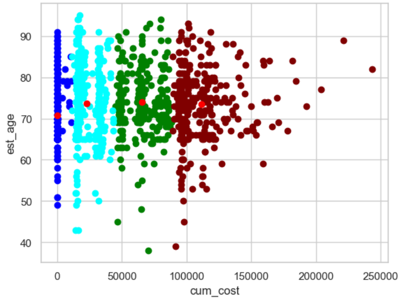
Segment 2 Features:

Cumulative cost, estimated age

Overall cumulative cost stayed under 100K across all ages in patients that discontinued therapy, whereas for those who completed, cumulative costs ran much higher.



Discontinued therapy population



Completed therapy population

Although we did only a couple of feature combinations for segmentation analysis, with more time, there is potential for more interesting insights.

8. STATISTICAL TESTS TO DETERMINE ASSOCIATIONS

a. Are there any racial disparities in the cumulative costs?

Hypothesis to check if there are significant differences in cumulative costs between racial groups.

Null Hypothesis (H0): There is no significant difference in cumulative costs between racial groups.
Alternative Hypothesis (H1): There is a significant difference in cumulative costs between racial groups.
Significance Level (Alpha) (Probability of making a Type I error): 0.05
p-value: 0.71

We ran a t-test to test for differences in cumulative costs between two racial groups, White and Black, and failed to reject the null hypothesis. We extended this to other groups as well and found there is no evidence of a significant difference between races for cumulative costs.

b. Is there any association between race and existence of an Adverse Drug Event (ADE) for a patient?

Hypotheses to check if there is any significant association between race and the existence of an ADE diagnosis.

Null Hypothesis (H0): There is no significant association between race and the existence of an ADE diagnosis.

Alternative Hypothesis (H1): There is a significant association between race and the existence of an ADE diagnosis.

Significance Level (Alpha): 0.05
p-value: 0.00

We ran a Chi-square test to assess if there was an association between race and existence of ADE for a patient and rejected the null hypothesis: There is evidence of a significant association between race and existence of an ADE diagnosis.

c. Is there any association between gender and existence of ADE for a patient?

Hypotheses to check if there is any significant association between gender and existence of ADE for a patient.

Null Hypothesis (H0): There is no significant association between gender and existence of ADE for a patient.
Alternative Hypothesis (H1): There is a significant association between gender and existence of ADE for a patient.
Significance Level (Alpha): 0.05
p-value: 0.0003

There is evidence of a significant association between gender and existence of ADE for a patient.

d. Since the average number of medical visits came up as one of the important features, along with the number of primary diagnoses, we ran a correlation test to see if the two features were positively correlated in two separate tests for the population completing therapy and those that did not.
   We found that in both these groups, the average number of medical visits was positively correlated with the number of primary diagnoses that a patient had, with a Pearson correlation value of 0.88 (for patients that completed therapy) and 0.87 (for patients that did not complete therapy).
e. The number of outpatient visits played a significant role in the continuation or withdrawal form therapy. Given that 14% in training (169) and 13% in test patients were disabled, another relationship we explored was between the number of outpatient visits and a patient's disability status.
   For patients that completed therapy, there was a significant difference in the mean outpatient visits among differently abled groups (p-value was 0.0003).
   For patients discontinuing therapy, there was no significant difference in the mean outpatient visits among differently abled groups (p-value was 0.554).

9. RECOMMENDATIONS

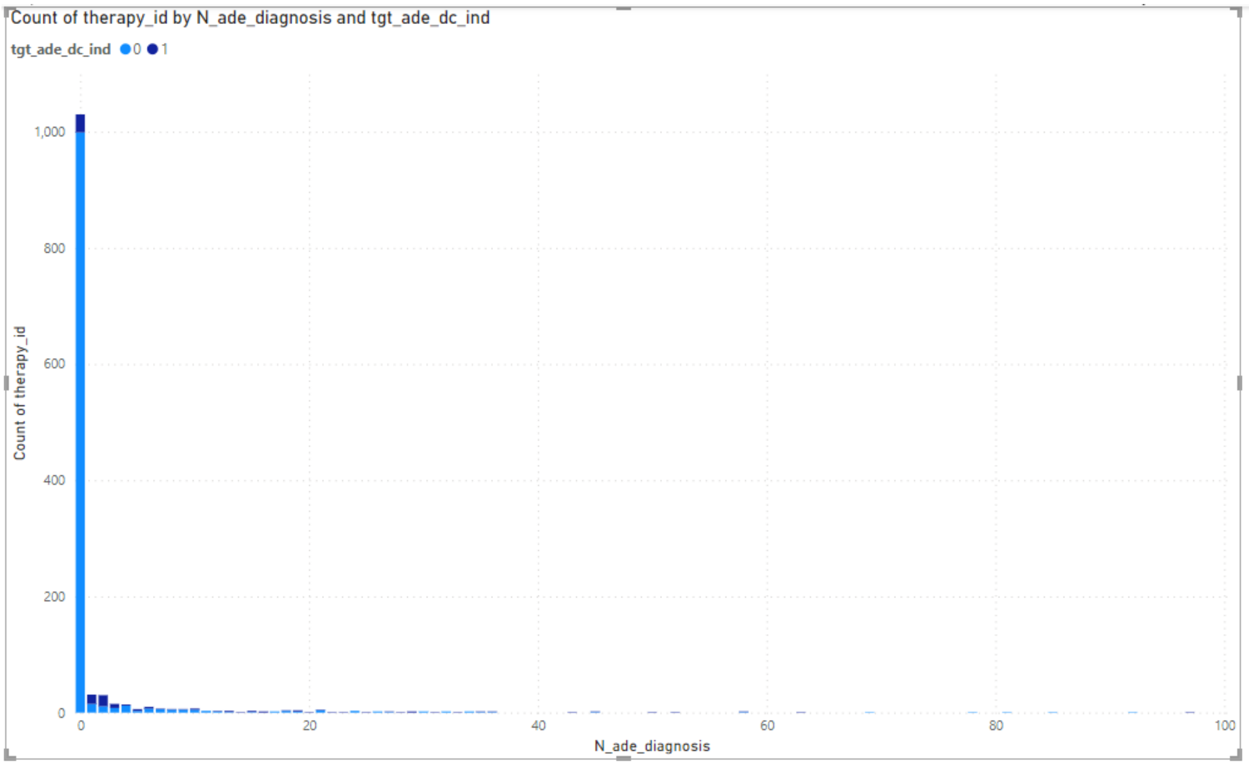
a. Address treatment gaps and side effects.

Are there any treatment gaps that could be reduced?
Between the diagnoses that a patient has been identified to have and the treatment for them, there seems to be some disparity. For example, 64 patients out of 1232 in the training dataset were identified as having been diagnosed with diarrhea, and out of these, only 16 patients were treated for the condition, and 11 of the 16 continued the therapy. Breaking down the treatments for other diagnoses, we have the table below:

Table 17

Diagnosis	# of patients with the diagnosis	# of patients treated	# of treated patients that completed therapy	# of untreated patients that completed therapy
Diarrhea	64	16	11 (69%)	66 % (32 of 48)
Nausea	52	38	15 (39.5%)	71.4% (10 of 14)
Seizure	13	11	8 (73%)	0

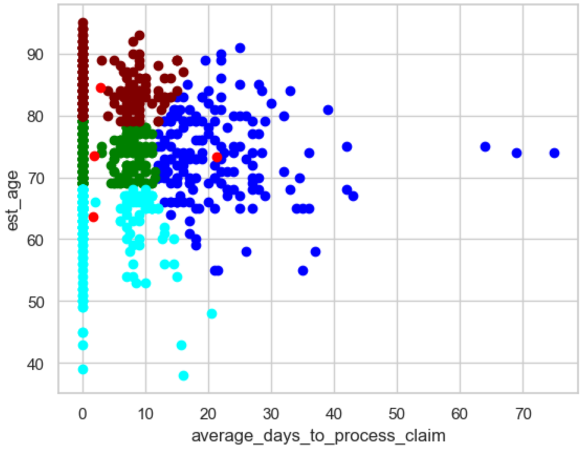
202 patients out of 1232 in the training data set were diagnosed with an Adverse Drug Event (ADE), out of which only 57% completed the therapy. and out of those, 36 had known interactions with the drug, and 86 completed the therapy. Despite this, treatment options did not seem to be either not offered or recorded.



N_ade_diagnosis by target outcome

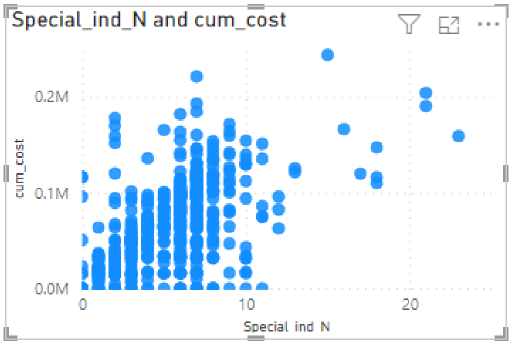
The number of patients dropping out after just 1 or 2 adverse drug events (ADE) was higher than those continuing the therapy. This further points to effective management of the side effects.

b. **Reduce the number of days it takes to process insurance claims.**
We wanted to check if there are any significant clusters or age groups among the patients that did not complete the Osimertinib therapy. In identifying clusters for the average number of days to process insurance claims and the estimated age, we found that insurance claims took longer to process for patients in a higher age range (above 70+) among those that did not complete the therapy.



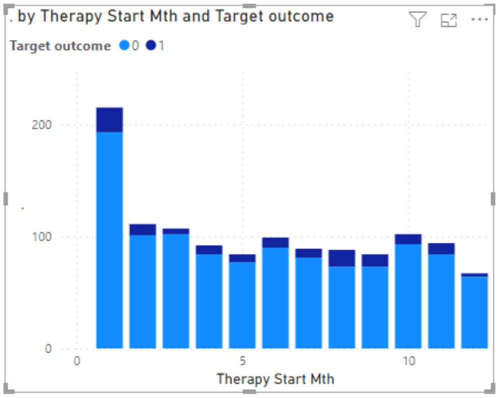
Average days to process claim by estimated age

c. **Make healthcare affordable and cost-effective.** The number of specialty drugs for a patient can be prohibitively expensive, even if all or a portion of it is covered by insurance. There is no mention of what deductibles are expected from a patient, but assuming they have to pay deductibles just like the rest of us, that could perhaps be a constraint staying in the therapy. The plot below shows a direct correlation between the number of specialty drugs a patient is on, and the cumulative cost incurred.



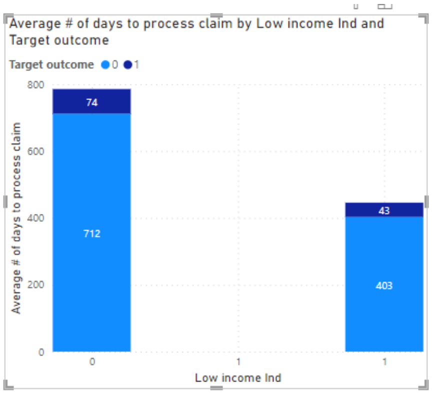
Specialty drug indicator by cumulative cost

The majority of therapy started in January, so it makes sense that the number of patients that discontinued seems pretty high in January. However, the therapy-start months of August and September (months 8 & 9 respectively) also show a higher proportion of patients that discontinued therapy. Understanding that the therapy is for 6 months, and without knowing the geographical location of the patient, it is possible that the discontinuation was because of weather (severe winter) or the Holiday season making it harder to follow-up on visits, or there were other factors related to coverage that might have run out by the end of the year.



Therapy start month by target outcome

In addition, 11% of low-income patients did not complete the therapy. It is not clear if cost was a factor.

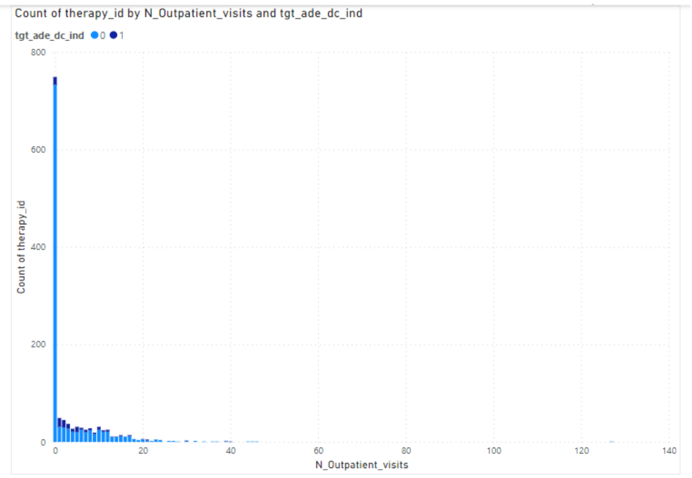


Average number of days to process claims by low-income and target outcome

d. **A closer look at the Outpatient and ER visits**: Looking at the column chart, we see that for the initial number of outpatient visits (1 through 5), the number of patients that dropped off therapy seems pretty high. Could there be a significant wait time for outpatient visits? Or was there a lot of paperwork to be filled out. It was also possible that because of the prevalence of Covid, everything was taking longer to process.

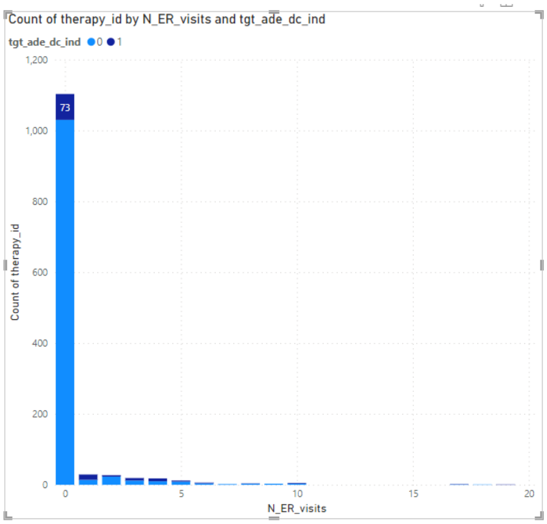
Number of Outpatient visits (n)	Total # of patients	# of patients that discontinued after n visits
1	49	18 (37%)
2	45	15 (33%)
3	37	10 (27%)
4	27	6 (22%)
5	31	11 (35%)

Comparison of number of discontinued therapies after n outpatient visits



Number of outpatient visits by target outcome

Similarly, more than 50% of people (15 of 29) discontinued therapy after just one ER visit. While there were dropouts from other types of visits, the number of people in those visit types were fewer than those found in Outpatient and ER.



Number of ER visits by target outcome

e. Since the case data spans from 2018-2022, out of which two of the years were impacted by the pandemic due to Covid-19, we may have confounding variable(s) related to the presence of the pandemic that may have reduced the number of in-person visits or might have increased hospitalizations and ER visits due to covid symptoms. To ensure that the features coming out as top factors for (dis)continuing the therapy are correctly represented, we would suggest increasing the number of years for the analysis. If data is absent, then we suggest ongoing analysis to ensure there are no hidden factors that might have surfaced because of Covid.

10. CONCLUSION

In this comprehensive study, we embarked on an in-depth analysis of data provided by Humana, focusing on their members who underwent Osimertinib (Tagrisso) therapy, a potent treatment for non-small cell lung cancer. The goal of our study was to develop predictive models that could help identify patients at risk of experiencing Adverse Drug Events (ADEs) and potentially discontinuing their therapy. Our commitment to fairness was upheld by treating unknown race and gender values as such, without making any assumptions.

Throughout our analysis, we addressed critical questions, such as racial disparities in cumulative costs, associations between race and ADEs, as well as gender and ADEs. We harnessed a broad range of performance indicators, including the number of medical visits, insurance claims, ADE diagnoses, and days to process claims, among other engineered variables.

Our modeling efforts, which leveraged techniques such as Gini Index, Random Forest, and XGBoost, culminated in the selection of the Random Forest model as the most effective, boasting an impressive AUC of 0.9727.

** Recommendations: **

Based on our extensive analysis, we have formulated several key recommendations for enhancing the efficacy and success of Osimertinib therapy and patient adherence:

1. **Addressing Treatment Gaps and Side Effects: **
   - Identify and address discrepancies between diagnosed conditions and treatments offered. For instance, there is evidence of undertreated conditions like diarrhea and nausea.
   - Focus on patients who discontinued therapy after only a few ADEs, emphasizing effective side effect management.

2. **Reducing Insurance Claim Processing Time: **
   - Implement strategies to expedite the processing of insurance claims, which could enhance patient experience and potentially improve adherence.

3. **Affordable and Cost-effective Healthcare: **
   - Explore ways to make specialty drugs more affordable for patients, possibly through deductibles and insurance coverage.
   - Assess the impact of healthcare costs, including deductibles, on patient adherence.

4. **Outpatient and ER Visit Optimization: **
   - Investigate the reasons behind the high dropout rates after just a few outpatient or ER visits.
   - Explore whether factors such as waiting times, paperwork, or the COVID-19 pandemic influenced the decision to discontinue therapy.

5. **Long-term Data Analysis: **
   - Extend the data analysis to more years to ensure that potential confounding variables, such as the COVID-19 pandemic, are thoroughly understood and accounted for.

6. **Exploration of Other High-Correlation Variables: **
   - Continue investigating variables that exhibit high correlations with seizure diagnoses to better understand their role in therapy discontinuation.

7. **Patient Demographics and Specialty Drug Costs: **
   - Assess the impact of age and disability status on insurance claim processing time.
   - Explore the effect of gender on patient adherence.

Our findings emphasize the importance of addressing specific issues like treatment gaps, claim processing time, healthcare affordability, and the impact of outpatient and ER visits on patient adherence. By implementing these recommendations, Humana can further enhance the effectiveness of Osimertinib therapy and contribute to better patient outcomes.

11. REFERENCES

1. https://mays.tamu.edu/wp-content/uploads/2023/09/Humana_Mays_Case_Competition_Informational_2023.pdf
2. ICD 10 to diagnosis mapping https://www.medicalbillingandcoding.org/icd-10-cm/