Assignment 2: Data Preprocessing for Case Competition

# Instructions

Learn, firsthand, how top teams tackle analytics problems and extract strategies you can reuse in your case competition. Choose one prize-winning report from the list provided (but make sure no one else in your group uses the same report): Focus on the approach, not the application domain, so you can generalize what you learn to your own case.

* Analyze the Data preparation sections of this report (look for sections titled data preprocessing, data cleaning, data transformation, or feature engineering)
  + What specific steps did they take to create, remove, or transform variables?
  + Why did they use each step? (Provide the rationale.)
  + Which of these steps will you try on your dataset, and why?

# Rellikson Kisyula

HUMANA-MAYS HEALTHCARE ANALYTICS CASE COMPETITION 2020: TRANSPORTATION ISSUES PREDICTION ANALYSIS

MAYS BUSINESS SCHOOL  
FINALIST 2020

## A. Specific Steps They Took to Create, Remove, or Transform Variables

### 1. Exploratory Data Analysis

* Dataset structure: The data had 69,572 Medicare members and 826 fields, including consumer data, medical claims, pharmacy claims, lab claims, demographics, credit data, condition-related features, and CMS features.
* Age analysis: The average age was 70.81 years, with most members between 66 and 77 years. Younger members reported more transportation issues than older members.
* Disability patterns: 23.18% of disabled members had transportation issues compared to 12.19% of non-disabled members.
* Health score correlation: Higher health risk scores (CCI, DSCI, FCI, HCC) were linked with more transportation issues.
* Other patterns: Smokers, renters, single-parent households, and high prescription drug users showed higher transportation issues than others.

### 2. Feature Engineering Techniques

* Group Binning and Ranking: The team grouped categorical variables such as education, household type, homeownership, and language into clearer categories. Numerical values were converted into percentile ranks to show relative standing.
* Weighted Metrics Creation: Built composite scores like StressIndex and MobilityIndex, normalized to a 0–100 scale for clarity.
* K-Means Clustering: Standardized variables and created 30 new cluster-based features to group members by credit, health, stress, or age patterns.
* Isolation Forest for Anomaly Detection: Identified 5,600 anomalies with higher transportation issues and created anomaly scores and binary flags.

### 3. Feature Selection Process

* Forward selection with Logistic Regression to add useful features.
* Random Forest importance scores used to select 250 features.
* Refined final 74 features with XGBoost for optimal performance.

## B. Rationale for Each Step

Exploratory Data Analysis (EDA) Rationale: Helped make sense of 800+ features and revealed domain patterns such as age, disability, and prescription use.

Feature Engineering Rationale: Grouping, ranking, and composite scores simplified raw data; clustering revealed hidden groups; anomaly detection flagged unusual records.

Feature Selection Rationale: Reduced dimensionality from 8,000+ to 74 features, balancing performance and efficiency. Used ROC AUC to guide decisions.

## C. Steps to Apply to Our Dataset and Why

* Systematic Exploratory Data Analysis We will begin by reviewing the dataset shape, target distribution, and basic descriptive statistics. Like their disability and age analysis, we will examine default rates by age, income groups, and gender to reveal useful patterns.
* Group Binning and Percentile Ranking We will bin categories like education into Basic, Intermediate, and Advanced groups. We will also convert key numeric variables such as income, credit, and annuity into percentile ranks. Finally, we will combine these into a Financial Profile Score that summarizes overall financial standing.
* K-Means Clustering for Customer Segmentation We will apply K-Means clustering on standardized features such as stress, stability, income, age, and credit. This will divide customers into 3–4 clusters, helping us identify risk segments with different default rates.
* Three-Step Feature Selection Process We will select features in three stages. First, apply statistical tests to filter. Second, use Random Forest importance to keep the strongest predictors. Third, refine with XGBoost to choose the final 50–75 features. This ensures a strong but efficient model.

# Ernestina Hooper

Assignment 2  
HUMANA-MAYS  
HEALTHCARE ANALYTICS CASE COMPETITION 2022: PREDICTION OF HOUSING INSECURITY  
MAYS BUSINESS SCHOOL  
TEAM-D-2022

## Data Preparation

What specific steps did they take to create, remove, or transform variables?

* Column Removal: Removed 119 columns that never changed and 12 columns with >80% missing values.
* Missing Data: Filled missing categories with 'Unknown' or most common value; imputed zeros/ones contextually; used KNN for county variables; median for normal/skewed distributions.
* Added New Data: Brought in external county-level data (health, poverty, food access).
* Created New Features: Composite scores and interaction terms (e.g., homeownership × low income).
* Feature Selection: Boruta algorithm reduced 810 to 134 variables.
* Outlier Removal: Isolation Forest removed 0.3% of unusual rows.

Why did they use each step? (Provide the rationale).

* Remove columns: Non-informative features clutter data and slow models.
* Missing Data: Preserved categories and applied domain logic (0 means 'none'). KNN leveraged similarities between counties.
* External Data: Poverty, housing, and food insecurity are strong predictors.
* New Features: Highlighted important relationships not directly visible.
* Feature Selection: Reduced noise and improved accuracy.
* Outlier Removal: Prevented unusual rows from distorting models.

Which of these steps will you try on your dataset, and why?

* Column Removal: Drop features with no variation or >80% missing.
* Missing Data Imputation: Apply 'Unknown' for categorical, median or 0 for numeric based on context.
* Feature Enrichment: Add external datasets (census, demographic, environmental).
* Engineering: Interaction terms from domain knowledge (e.g., age × income).
* Feature Selection: Use Boruta or Random Forest to reduce dimensions.
* Outlier Removal: Apply Isolation Forest or z-score filtering.

# Andrew T Mavizha

ASSIGNMENT 2  
HUMANA-MAYS  
APPLYING ADVANCED ANALYTICS TO IMPROVE TAGRISSO ADHERENCE AMONG MEMBERS WITH NON-SMALL CELL LUNG CANCER  
HUMANA-MAYS 2023 CASE COMPETITION  
FINALIST E -2023

What specific steps did they take to create, remove, or transform variables?  
Answer

* Handling Null Values: 'Unknown' for race, mode for categorical, median for continuous variables.
* Categorical Transformation: One-hot encoding for demographic variables.
* Derived Variables: Aggregated claims-level data into 76 member-level features.
* Improving Generalizability: SMOTE for imbalance, cross-validation and regularization, standardized continuous variables.

Why did they use each step? (Provide the rationale.)  
Answer

* Imputation reduced data loss while preserving incomplete records.
* One-hot encoding prepared categorical data for ML models.
* Derived features summarized claims into meaningful predictors.
* SMOTE balanced classes; cross-validation improved generalizability; standardization stabilized models.

c) Which of these steps will you try on your dataset, and why?  
Answer

* Imputation: Mode for categorical, median for continuous.
* One-hot encoding for demographic variables.
* Feature engineering via aggregation (e.g., totals, averages).
* Balancing with SMOTE for class imbalance.
* Cross-validation and regularization for better generalization.
* Standardization for continuous features in sensitive models.