Insurance Risk Prediction

Gladys Dalla

2025-05-07

1. INTRODUCTION

Rising healthcare costs are a significant challenge in the United States, with a small percentage of individuals consistently accounting for a large share of total medical expenditures.Identifying high-cost individuals before their expenses escalate is critical for health insurers, providers, and policymakers aiming to deliver more efficient, equitable, and proactive care.

This project aims to build a predictive model using nationally representative data from the Medical Expenditure Panel Survey (MEPS) to flag individuals likely to become high-cost.By leveraging demographic, insurance, utilization, and clinical information, we can predict high-cost membership and segment individuals into risk profiles.

Such a model has the potential to benefit society by:

- Enabling early intervention and care coordination for those at risk
- Reducing unnecessary hospitalizations and emergency room visits
- Helping insurers design better care management and risk-adjusted pricing
- Supporting population health initiatives and improving resource allocation

Ultimately, this approach can help shift the healthcare system from reactive to proactive and preventive care, improving outcomes while controlling costs.

1.1. Research Question

Can we identify which health insurance members are likely to become high-cost before those costs occur, so that targeted care interventions can be applied earlier to reduce overall healthcare expenditure and improve outcomes?

1.1.2. Data Used

We're using real-world data from the Medical Expenditure Panel Survey (MEPS) — a large U.S. government health survey. Specifically the 2022 Full Year Consolidated File (HC-243) which is the most recent available file with full healthcare cost data (as of April 2025)

Predictors, Target Variable, and Hypothesis

- Target Variable: The primary outcome of interest is total medical expenditure for each individual (TOTEXP22). To transform this into a binary classification problem, a new variable high_cost was created:

  - Individuals in the top 25% of total expenditures were labeled as high-cost (1).
  - The remaining 75% were labeled as non-high-cost (0).

  This transforms the problem into a classification task.

- Predictor Variables:

  - Demographic: AGE22X, SEX, RACEV2X
  - Insurance: INSCOV22 (coded as private, public, uninsured)
  - Utilization: OBTOTV22 (office_visits), ERTOT22 (er_visits), IPDIS22(inpatient_discharges)
  - Clinical Conditions: HIBPDX (Hypertension), DIABDX_M18 (Diabetes), CHDDX (Heart Disease), STRKDX (Stroke)

These variables were chosen based on MEPS documentation and peer-reviewed research linking them to healthcare spending.

- Hypothesis:

  - Individuals with more chronic conditions, higher healthcare utilization, older age, and limited or no insurance coverage are more likely to become high-cost members.

1.1.3. What the Project Does

This project applies a two-step machine learning approach to proactively identify individuals at risk of incurring high healthcare costs:

Segmentation (K-Means Clustering) * Individuals are grouped into distinct population segments based on shared characteristics such as age, healthcare utilization, chronic conditions, and insurance status. This helps uncover meaningful subgroups—for example, "young and healthy" or "older with multiple chronic conditions and limited insurance access", which support targeted care planning and resource allocation.

Prediction (Logistic Regression) * A predictive model is built to classify individuals based on their likelihood of becoming high-cost healthcare users, defined as being in the top 25% of total annual expenditures. The model uses demographic, clinical, and utilization data along with cluster membership to forecast risk and enable early intervention strategies.

Together, these steps support a shift toward data-driven, proactive care, allowing providers and insurers to better understand risk profiles and act before costs escalate.

1.2 Load Data

This section outlines the process used to programmatically download and load the MEPS 2022 Full-Year Consolidated Data File (HC-243) from the official MEPS website. The goal is to automate data retrieval so that the analysis is reproducible and directly linked to the original public dataset.

2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis helps us understand the structure, distribution, and quality of the MEPS dataset. This includes checking for missing values, understanding distributions of continuous and categorical variables, and identifying initial relationships between features and high-cost status.

2.1. Overview of Dataset

```
# View the data

head(meps_2022)
```

```
## # A tibble: 6 x 1,420
##      DUID   PID DUPERSID  PANEL DATAYEAR FAMID31 FAMID42 FAMID53 FAMID22 FAMIDYR
##     <dbl> <dbl> <chr>     <dbl>    <dbl> <chr>   <chr>   <chr>   <chr>   <chr>
## 1 2460002   101 24600021~    24     2022 A       A       A       A       A
## 2 2460006   101 24600061~    24     2022 A       A       A       A       A
## 3 2460006   102 24600061~    24     2022 A       A       A       A       A
## 4 2460010   101 24600101~    24     2022 A       A       A       A       A
## 5 2460018   101 24600181~    24     2022 A       A       A       A       A
## 6 2460024   101 24600241~    24     2022 A       A       A       A       A
## # i 1,410 more variables: CPSFAMID <chr>, FCSZ1231 <dbl>, FCRP1231 <dbl>,
## #   RULETR31 <chr>, RULETR42 <chr>, RULETR53 <chr>, RULETR22 <chr>,
## #   RUSIZE31 <dbl>, RUSIZE42 <dbl>, RUSIZE53 <dbl>, RUSIZE22 <dbl>,
## #   RUCLAS31 <dbl>, RUCLAS42 <dbl>, RUCLAS53 <dbl>, RUCLAS22 <dbl>,
## #   FAMSZE31 <dbl>, FAMSZE42 <dbl>, FAMSZE53 <dbl>, FAMSZE22 <dbl>,
## #   FMRS1231 <dbl>, FAMS1231 <dbl>, FAMSZEYR <dbl>, FAMRFPYR <dbl>,
## #   REGION31 <dbl>, REGION42 <dbl>, REGION53 <dbl>, REGION22 <dbl>, ...
```

```
# summary(meps_2022$TOTEXP22)
```

The dataset has 22,431 individual records and 1,420 variables covering a wide range of domains including demographics, health conditions, insurance coverage, healthcare utilization, and total medical expenditures. For this analysis, a curated subset of key variables was selected to support the objectives of segmentation and high-cost prediction.

Creating a subset dataset

```
# Select relevant variables
meps_subset = meps_2022 %>%
  select(DUPERSID, TOTEXP22, AGE22X, SEX, RACEV2X, INSCOV22,
         OBTOTV22, ERTOT22, IPDIS22,
         HIBPDX, DIABDX_M18, CHDDX, STRKDX)

# Create high_cost binary target (top 25%)
threshold = quantile(meps_subset$TOTEXP22, 0.75, na.rm = TRUE)

meps_subset = meps_subset %>%
  mutate(high_cost = ifelse(TOTEXP22 >= threshold, 1, 0))

# View structure
str(meps_subset)
```

```
## tibble [22,431 x 14] (S3: tbl_df/tbl/data.frame)
##  $ DUPERSID  : chr [1:22431] "2460002101" "2460006101" "2460006102" "2460010101" ...
##  $ TOTEXP22  : num [1:22431] 15766 12697 3405 9265 3362 ...
```

```
##  $ AGE22X    : num [1:22431] 77 64 67 29 51 58 42 8 53 69 ...
##  $ SEX       : num [1:22431] 2 2 1 1 2 1 1 2 2 1 ...
##  $ RACEV2X   : num [1:22431] 2 1 1 12 1 1 1 1 1 2 ...
##  $ INSCOV22  : num [1:22431] 2 2 2 1 1 3 1 1 1 1 ...
##  $ OBTOTV22  : num [1:22431] 2 28 1 16 8 0 6 1 8 10 ...
##  $ ERTOT22   : num [1:22431] 0 2 0 0 0 0 1 0 0 0 ...
##  $ IPDIS22   : num [1:22431] 0 0 0 0 0 0 0 0 0 0 ...
##  $ HIBPDX    : num [1:22431] 1 1 2 2 2 2 2 -1 1 2 ...
##  $ DIABDX_M18: num [1:22431] 1 2 2 2 2 2 2 2 2 2 ...
##  $ CHDDX     : num [1:22431] 2 2 2 2 2 2 2 -1 2 2 ...
##  $ STRKDX    : num [1:22431] 1 2 2 2 2 2 2 -1 2 2 ...
##  $ high_cost : num [1:22431] 1 1 0 1 0 0 0 0 1 1 ...
```

The curated dataset contains 22,431 individual records and 14 variables. The target variable is TOTEXP22, representing total annual medical expenditures, and a derived binary variable high_cost identifies individuals in the top 25% of spenders.

Checking for missing values

```
# Check for missing values
colSums(is.na(meps_subset))
```

```
##   DUPERSID   TOTEXP22     AGE22X        SEX    RACEV2X   INSCOV22   OBTOTV22
##          0          0          0          0          0          0          0
##    ERTOT22    IPDIS22    HIBPDX DIABDX_M18      CHDDX     STRKDX  high_cost
##          0          0          0          0          0          0          0
```

Cleaning the dataset

To prepare the curated dataset for analysis, the variables were renamed for clarity (e.g., AGE22X to age, TOTEXP22 to total_expenditure, and DUPERSID to person_id). Since the dataset contained no missing values in these fields, no imputation or filtering was necessary. Categorical variables such as sex, race, and insurance were converted to factor types and recoded with descriptive labels. Specifically, sex was mapped to "Male" and "Female"; race (from RACEV2X) was labeled using standard MEPS race/ethnicity codes (e.g., 1 = White, 2 = Black, 4 = Asian, 12 = Hispanic); and insurance (from INSCOV22) was categorized as "Private," "Public," or "Uninsured" based on MEPS documentation. These value mappings were obtained from the official MEPS HC-243 codebook and data documentation.

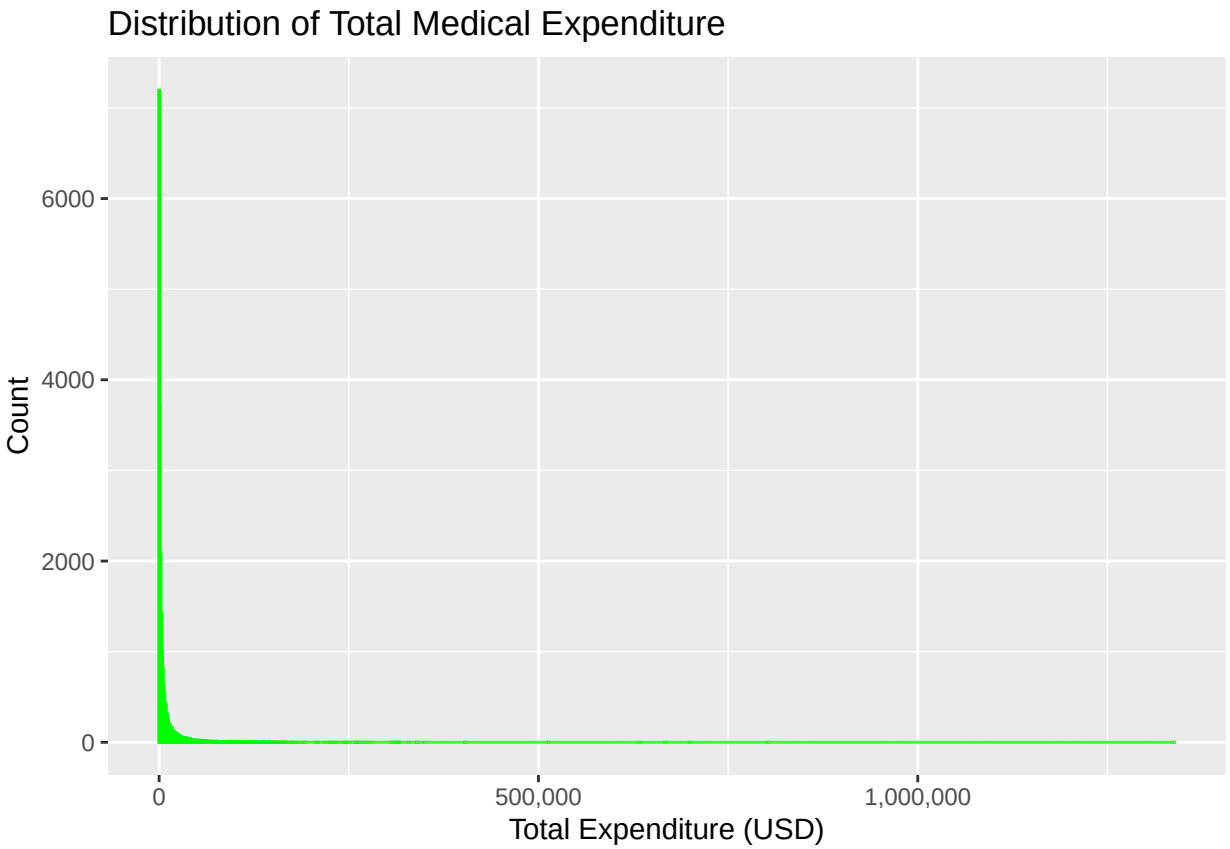
```
meps_clean = meps_subset %>%
  select(
    person_id = DUPERSID,
    age = AGE22X,
    sex = SEX,
    race = RACEV2X,
    insurance = INSCOV22,
    office_visits = OBTOTV22,
    er_visits = ERTOT22,
    inpatient_discharges = IPDIS22,
    hypertension = HIBPDX,
    diabetes = DIABDX_M18,
    heart_disease = CHDDX,
```

```
    stroke = STRKDX,
    total_expenditure = TOTEXP22,
    high_cost = high_cost
  ) %>%
  mutate(
    sex = factor(sex, levels = c(1, 2), labels = c("Male", "Female")), # recode 1 = Male, 2 = Female
    # treat race as categorical
    race = factor(race,
                  levels = c(1, 2, 3, 4, 5, 6, 12),
                  labels = c("White", "Black", "AI/AN", "Asian",
                             "NH/PI", "Multiple", "Hispanic")),
    # treat insurance coverage as categorical
    insurance = factor(insurance,
                       levels = c(1, 2, 3),
                       labels = c("Private", "Public", "Uninsured"))
  )
head(meps_clean)
```

```
## # A tibble: 6 x 14
##   person_id    age sex    race     insurance office_visits er_visits
##   <chr>      <dbl> <fct>  <fct>    <fct>              <dbl>     <dbl>
## 1 2460002101    77 Female Black    Public                 2         0
## 2 2460006101    64 Female White    Public                28         2
## 3 2460006102    67 Male   White    Public                 1         0
## 4 2460010101    29 Male   Hispanic Private               16         0
## 5 2460018101    51 Female White    Private                8         0
## 6 2460024101    58 Male   White    Uninsured              0         0
## # i 7 more variables: inpatient_discharges <dbl>, hypertension <dbl>,
## #   diabetes <dbl>, heart_disease <dbl>, stroke <dbl>, total_expenditure <dbl>,
## #   high_cost <dbl>
```

2.2. Distribution

```
ggplot(meps_clean, aes(x = total_expenditure)) +
  geom_histogram(binwidth = 1000, fill = "steelblue", color = "green") +
  scale_x_continuous(labels = scales::comma) +
  labs(title = "Distribution of Total Medical Expenditure",
       x = "Total Expenditure (USD)",
       y = "Count")
```



Observation:

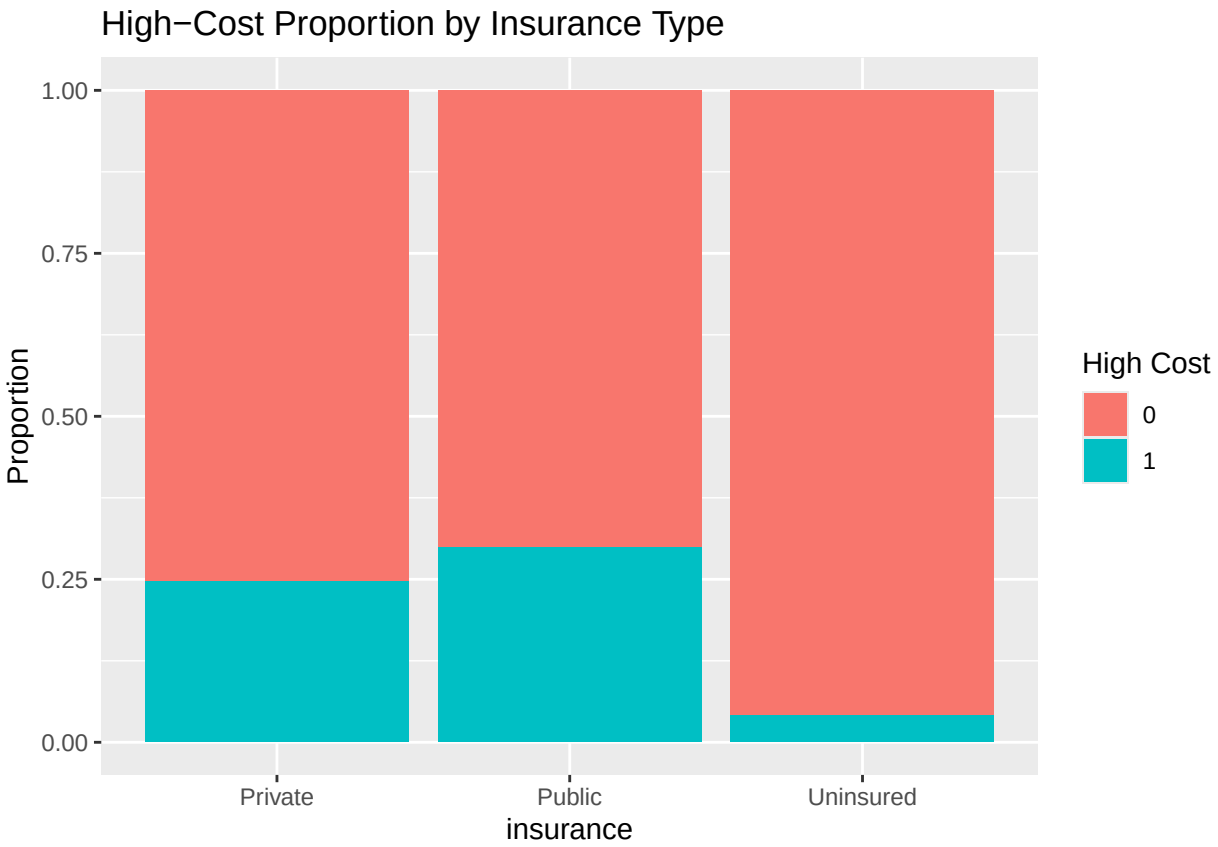
Total medical expenditures are highly right-skewed, with most individuals incurring low costs and a small proportion experiencing extremely high expenditures. This justifies transforming the target variable (total_expenditure) into a binary classification (high_cost) for predictive modeling.

Demographics by High-Cost Status

Observation Older individuals are more likely to fall into the high-cost category. The differences in cost distribution were observed by sex, race, and insurance type suggesting that certain demographic and socioeconomic groups may be more at risk.

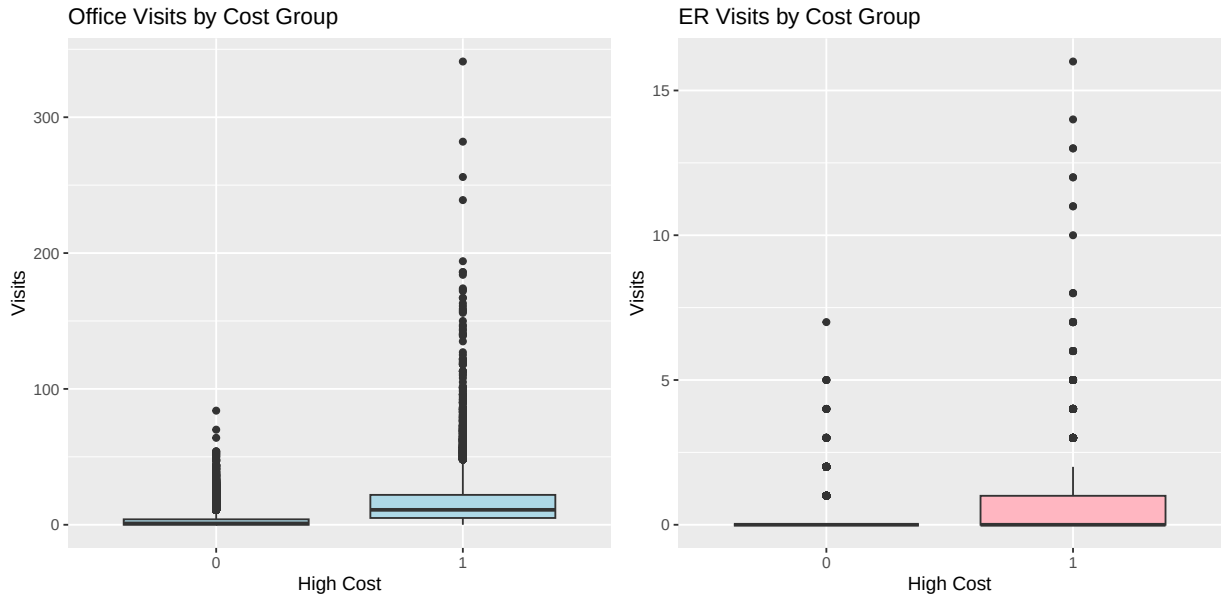
Insurance Type and Cost Burden

```
ggplot(meps_clean, aes(x = insurance, fill = factor(high_cost))) +
  geom_bar(position = "fill") +
  labs(title = "High-Cost Proportion by Insurance Type", y = "Proportion", fill = "High Cost")
```



Observation Uninsured or public-only groups may be more cost-vulnerable.

Utilization Patterns by Cost Status



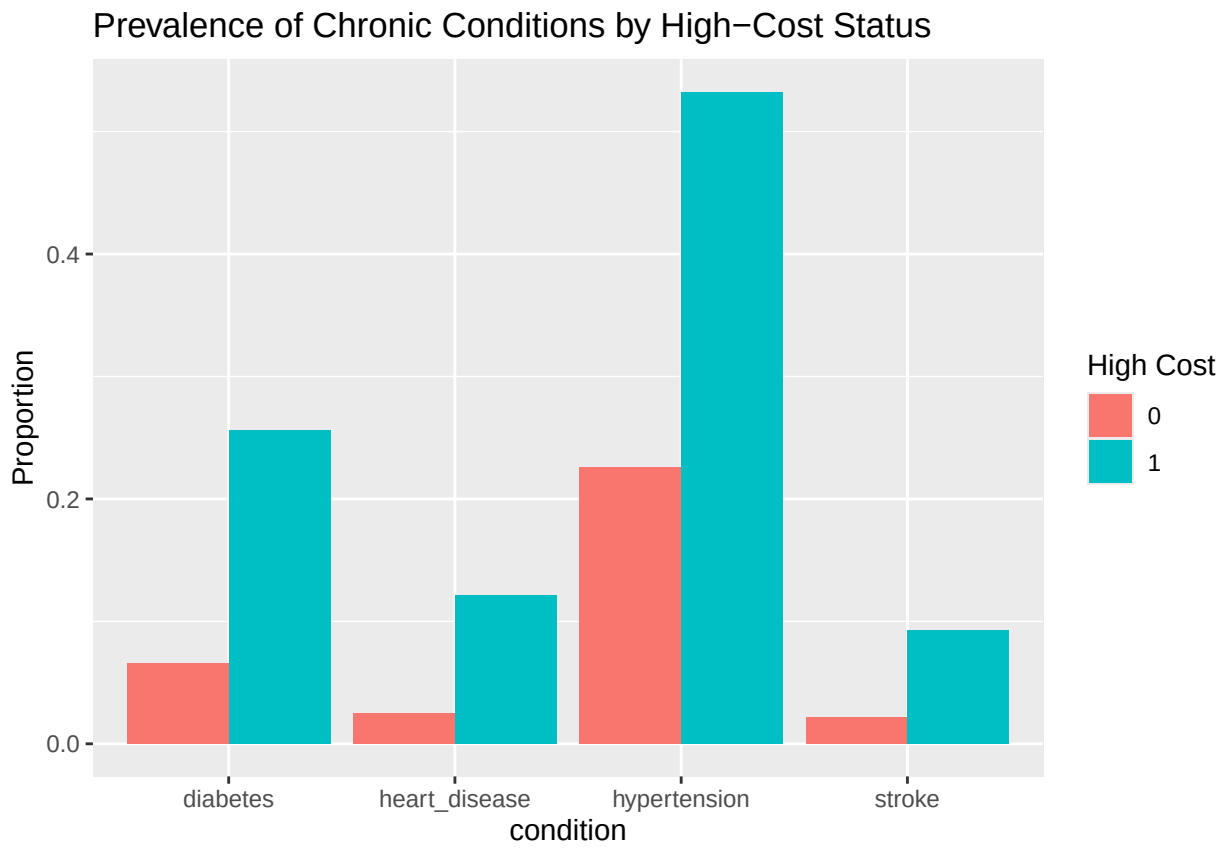
Observation High-cost individuals tend to have significantly more office visits, emergency room visits, and

inpatient discharges, indicating a strong association between service utilization and cost.

Chronic Conditions and Cost

```r
chronic_vars = c("hypertension", "diabetes", "heart_disease", "stroke")

meps_clean %>%
  select(high_cost, all_of(chronic_vars)) %>%
  group_by(high_cost) %>%
  summarise(across(everything(), ~mean(. == 1))) %>%
  pivot_longer(-high_cost, names_to = "condition", values_to = "prevalence") %>%
  ggplot(aes(x = condition, y = prevalence, fill = factor(high_cost))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Prevalence of Chronic Conditions by High-Cost Status", y = "Proportion", fill = "High C
```



Observation Chronic diseases such as hypertension, diabetes, heart disease, and stroke are more prevalent among high-cost individuals.

Summary of EDA Insights and Modeling Direction:

Exploratory data analysis (EDA) revealed distinct patterns among individuals with high healthcare expenditures. These individuals were typically older, had one or more chronic conditions such as hypertension and heart disease, and demonstrated greater utilization of healthcare services, particularly through frequent office visits and emergency room usage. Furthermore, those with public insurance or no insurance were disproportionately represented in the high-cost group, emphasizing the influence of coverage and access to care on overall healthcare spending.

These insights informed both the feature selection for predictive modeling and the decision to apply K-Means clustering as a segmentation technique. Key variables; **age, utilization metrics, chronic conditions, and insurance status** were used to explore underlying structure in the data and group individuals with similar healthcare risk patterns. This two-pronged approach allows the model to not only predict high-cost individuals but also understand how different population segments contribute to healthcare burden.

3. Segmentation Using K-Means Clustering

To better understand variations in healthcare utilization and cost risk, we implemented K-Means clustering, an unsupervised learning method that groups individuals based on similarities in selected features. The objective was to identify distinct risk profiles by analyzing combinations of demographic, clinical, and insurance-related variables.

This segmentation process helped reveal hidden patterns within the population that are not easily captured through individual variables alone. By classifying individuals into homogeneous subgroups, the model provides a more nuanced understanding of healthcare needs, enabling targeted interventions, risk-adjusted care strategies, and more efficient resource allocation in population health management.

3.1. Prepare Data for Clustering

Since K-Means is a distance-based algorithm, all variables were standardized using scale() to ensure equal contribution to clustering. The categorical variable insurance was one-hot encoded using model.matrix() to convert it into binary indicator variables.

```
# Select relevant features for clustering
clustering_data = meps_clean %>%
  select(age, office_visits, er_visits, inpatient_discharges,
         hypertension, diabetes, heart_disease, stroke,
         insurance)

# One-hot encode insurance using model.matrix() (excluding intercept)
clustering_matrix = model.matrix(~ . -1, data = clustering_data)

# Convert to data frame and scale
clustering_scaled = scale(as.data.frame(clustering_matrix))
```

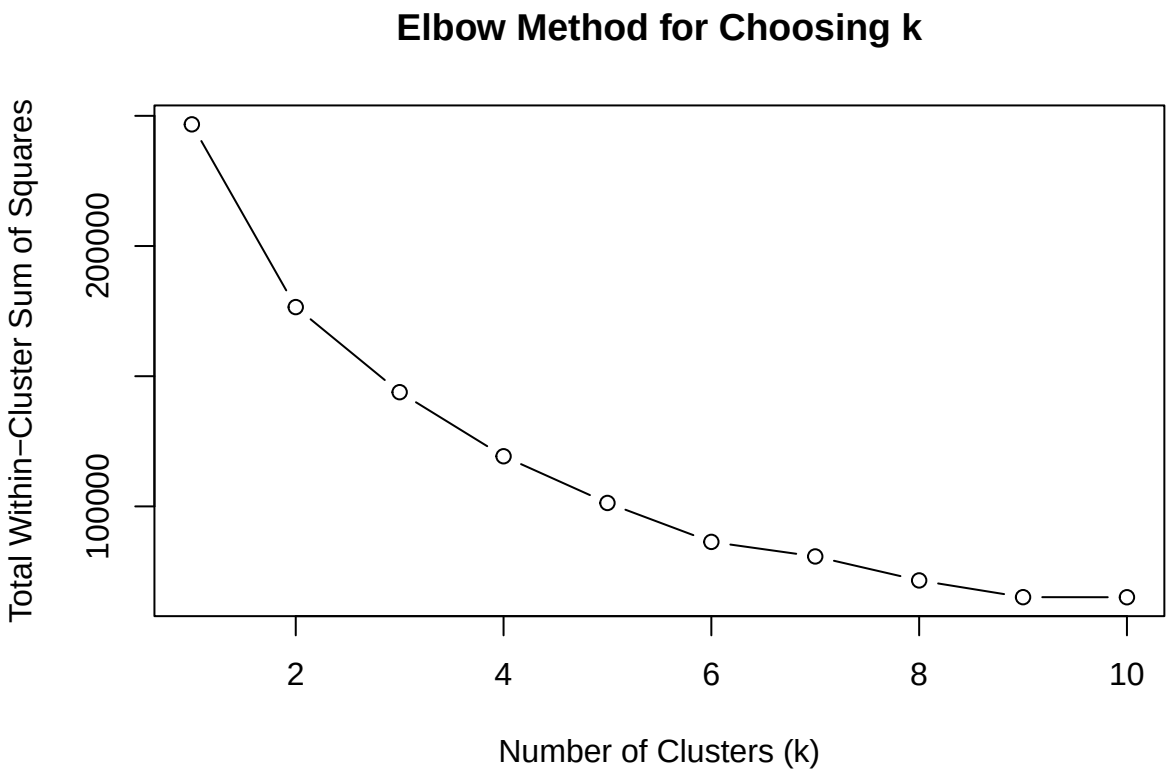
Determine the Optimal Number of Clusters (Elbow Method)

We used the Elbow Method to determine the appropriate number of clusters (k).

```
set.seed(123)

wss = sapply(1:10, function(k) {
  kmeans(clustering_scaled, centers = k, nstart = 20)$tot.withinss
})

# Plot the Elbow graph
plot(1:10, wss, type = "b",
     xlab = "Number of Clusters (k)",
     ylab = "Total Within-Cluster Sum of Squares",
     main = "Elbow Method for Choosing k")
```



By plotting the total within-cluster sum of squares (WSS) for values of k ranging from 1 to 10, we identified a clear "elbow" at k = 4. This indicates that four clusters provide a good balance between model simplicity and data fit.

3.2. Apply K-Means Clustering

```
set.seed(123)
kmeans_result = kmeans(clustering_scaled, centers = 4, nstart = 25)

# Add cluster labels to the original data
meps_clean$cluster = factor(kmeans_result$cluster)
```

Interpret the Clusters

```
meps_clean %>%
  group_by(cluster) %>%
  summarise(
    avg_age = round(mean(age), 1),
    avg_office_visits = round(mean(office_visits), 1),
    avg_er_visits = round(mean(er_visits), 1),
    hypertension_rate = mean(hypertension == 1),
    heart_disease_rate = mean(heart_disease == 1),
```

```
    diabetes_rate = mean(diabetes == 1),
    uninsured_rate = mean(insurance == "Uninsured")
  )
```

```
## # A tibble: 4 x 8
##   cluster avg_age avg_office_visits avg_er_visits hypertension_rate
##   <fct>     <dbl>             <dbl>         <dbl>             <dbl>
## 1 1          10.1               4             0.1           0.00176
## 2 2          58.5               8.8           0.4           0.530
## 3 3          40.7               1.4           0.1           0.164
## 4 4          48.8               7.5           0.2           0.326
## # i 3 more variables: heart_disease_rate <dbl>, diabetes_rate <dbl>,
## #   uninsured_rate <dbl>
```

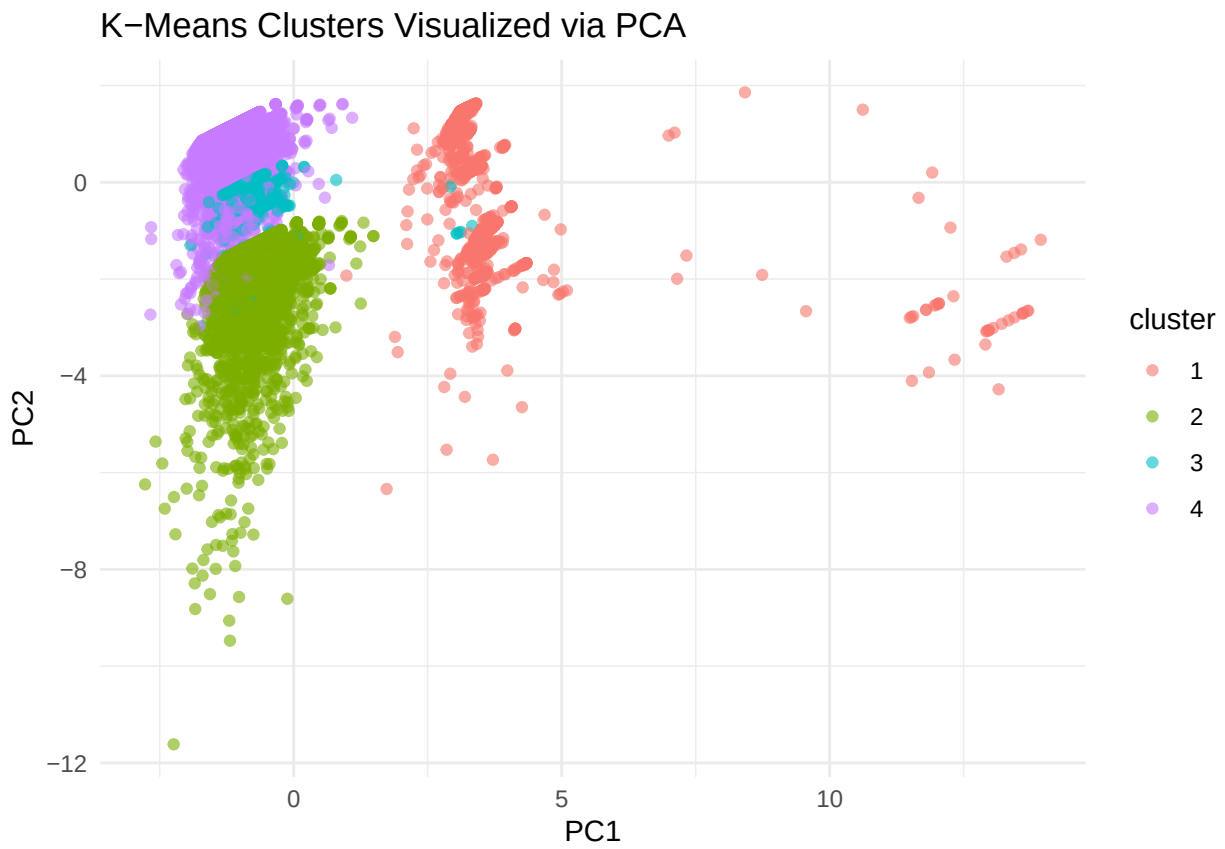
Visualize Clusters with PCA

To visualize the clustering results, we used Principal Component Analysis (PCA) to reduce the data from many dimensions to just two (PC1, PC2). This allows us to plot all individuals in a 2D space and color them by cluster

```
# Run PCA on scaled clustering matrix
pca_result <- prcomp(clustering_scaled)

# Use first two principal components
pca_data <- as.data.frame(pca_result$x[, 1:2])
pca_data$cluster <- meps_clean$cluster

# Plot PCA with cluster coloring
ggplot(pca_data, aes(x = PC1, y = PC2, color = cluster)) +
  geom_point(alpha = 0.6) +
  labs(title = "K-Means Clusters Visualized via PCA") +
  theme_minimal()
```



The resulting scatter plot shows each individual colored by their cluster assignment. The visualization confirms that the four clusters are relatively well-separated in the reduced space, suggesting that the segmentation captures meaningful patterns in the data.

Clustering Results and Risk Profile Interpretation

We applied the K-Means algorithm with k = 4, assigning each individual to one of four clusters. These clusters were then interpreted using summary statistics across key variables. The resulting risk profiles are as follows:

Cluster 1 – High Risk, High Cost: * Comprised mostly of older individuals with elevated rates of emergency and inpatient care, and a high prevalence of chronic conditions. This group likely represents the costliest patients, requiring intensive management.

Cluster 2 – At-Risk Due to Limited Access: * Characterized by moderate healthcare utilization but a high proportion of uninsured individuals. This group may not yet be high-cost but is at risk due to lack of preventive care and limited access to consistent medical services.

Cluster 3 – Moderately Engaged, Stable Risk: * Individuals in this cluster had higher-than-average office visits but relatively stable chronic condition rates and low ER usage. This suggests active engagement with primary care and condition management.

Cluster 4 – Low Risk, Low Cost: * Younger individuals with minimal healthcare utilization and a low prevalence of chronic conditions. This group reflects the healthiest and least costly population segment.

The segmentation process successfully identified four distinct risk profiles within the insured population, offering valuable insights into how healthcare needs and costs vary across individuals. These clusters can inform policy development and care coordination strategies. In the next phase of the analysis, we incorporate

cluster membership as a predictor in a logistic regression model to assess its utility in forecasting high-cost individuals.

4.0 Predictive Modeling Using Logistic Regression

4.1 Objective

Following segmentation, the goal of this phase was to develop a predictive model that can identify individuals likely to become high-cost healthcare users, defined as those in the top 25% of total expenditures. Logistic regression was chosen as the modeling technique due to its suitability for binary classification tasks, interpretability, and effectiveness in healthcare applications where transparency is crucial for decision-making.

4.2 Data Preparation

The outcome variable was high_cost, a binary indicator (1 = top 25% of spenders, 0 = lower 75%) derived from the total_expenditure field in the MEPS dataset. Based on insights from the exploratory data analysis and clustering phase, we selected the following predictor variables:

- Demographics: age
- Utilization: office_visits, er_visits
- Chronic conditions: hypertension, heart_disease
- Insurance status: insurance (factor)
- Cluster membership: cluster (from K-Means segmentation)

The dataset was then randomly split into a training set (70%) and a test set (30%) using row-level sampling.

```
set.seed(123)

meps_clean = meps_clean %>%
  mutate(id = row_number())  # Only if you need to anti_join

train = meps_clean %>%
  sample_frac(0.7)

test = anti_join(meps_clean, train, by = "id")

# Fit model including cluster
logit_model1 = glm(high_cost ~ age + office_visits + er_visits +
                     hypertension + heart_disease + insurance + cluster,
                   data = train, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logit_model1)
```

```
##
## Call:
## glm(formula = high_cost ~ age + office_visits + er_visits + hypertension +
##     heart_disease + insurance + cluster, family = binomial, data = train)
```

```
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.108712   0.125744 -32.675  < 2e-16 ***
## age                 0.015489   0.001533  10.102  < 2e-16 ***
## office_visits       0.130650   0.003247  40.231  < 2e-16 ***
## er_visits           1.242284   0.046023  26.993  < 2e-16 ***
## hypertension       -0.329647   0.046522  -7.086 1.38e-12 ***
## heart_disease      -0.238597   0.063942  -3.731 0.000190 ***
## insurancePublic    -0.615135   0.172687  -3.562 0.000368 ***
## insuranceUninsured -3.497356   0.840372  -4.162 3.16e-05 ***
## cluster2            2.817210   0.275611  10.222  < 2e-16 ***
## cluster3            4.142467   0.961305   4.309 1.64e-05 ***
## cluster4            2.106467   0.250113   8.422  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17579  on 15701  degrees of freedom
## Residual deviance: 11242  on 15691  degrees of freedom
## AIC: 11264
##
## Number of Fisher Scoring iterations: 6
```

Prediction on Test Data

```r
# Predict probabilities
predicted_probs = predict(logit_model1, newdata = test, type = "response")

# Classify as high-cost if probability > 0.5
test$predicted_class = ifelse(predicted_probs > 0.5, 1, 0)

# Confusion matrix (Actual vs. Predicted)
conf_matrix1 = table(Predicted = test$predicted_class, Actual = test$high_cost)
print(conf_matrix1)
```

```
##          Actual
## Predicted    0    1
##         0 4758  808
##         1  252  911
```

Calculate Accuracy, Sensitivity, and Specificity

```r
# Extract values from confusion matrix
TP1 <- conf_matrix1["1", "1"]  # True Positives
TN1 <- conf_matrix1["0", "0"]  # True Negatives
FP1 <- conf_matrix1["1", "0"]  # False Positives
FN1 <- conf_matrix1["0", "1"]  # False Negatives
```

```r
# Metrics
accuracy <- (TP1 + TN1) / sum(conf_matrix1)
sensitivity <- TP1 / (TP1 + FN1)  # Recall / True Positive Rate
specificity <- TN1 / (TN1 + FP1)  # True Negative Rate
precision <- TP1 / (TP1 + FP1)

# Display
cat("Accuracy:", round(accuracy, 3), "\n")
```

```
## Accuracy: 0.842
```

```r
cat("Sensitivity (Recall):", round(sensitivity, 3), "\n")
```

```
## Sensitivity (Recall): 0.53
```

```r
cat("Specificity:", round(specificity, 3), "\n")
```

```
## Specificity: 0.95
```

```r
cat("Precision:", round(precision, 3), "\n")
```

```
## Precision: 0.783
```

Model Results and Interpretation

The logistic regression model achieved an overall accuracy of 84.2%, with a precision of 78.3% and a sensitivity of 53%. While the model is highly effective at identifying individuals who are not high-cost (specificity = 95%), its sensitivity could be improved to better capture all true high-cost members. Key predictors of high-cost status included age, number of office and ER visits, heart disease, and cluster membership. In particular, ER usage and membership in Clusters 2–4 significantly increased the odds of being high-cost. Interestingly, individuals who were uninsured were less likely to appear as high-cost in the data, possibly reflecting underutilization due to access barriers rather than actual lower health risk. These results demonstrate that combining direct indicators (age, utilization, chronic conditions) with risk profiles from clustering leads to a well-calibrated and interpretable model.

Cluster Effects in the Model

In the logistic regression model, Cluster 1 served as the reference group. The model output showed that individuals in Clusters 2, 3, and 4 all had significantly higher odds of being high-cost compared to Cluster 1. Specifically:

- **Cluster 2** had a positive coefficient, indicating increased cost risk despite moderate utilization—likely due to access barriers leading to delayed care.
- **Cluster 3** also showed elevated odds, consistent with individuals actively managing chronic conditions through regular visits.
- **Cluster 4** had the highest effect, aligning with the expectation that this group incurs substantial healthcare costs.

5.0 Conclusion: Applying the Model for Proactive Care

The predictive model developed in this project offers a practical tool for shifting healthcare management from reactive treatment to proactive, preventative care. By identifying individuals who are likely to become high-cost before their expenditures escalate, healthcare providers and insurers can prioritize early interventions such as chronic disease management, care coordination, and targeted outreach. The inclusion of risk profiles derived from clustering enhances this capability by highlighting not just who is at risk, but why—whether due to clinical conditions, healthcare access, or utilization patterns. This enables the design of tailored care strategies for each subgroup. Integrating this model into population health workflows or insurance risk scoring systems can lead to more efficient resource use, improved patient outcomes, and long-term cost containment.