Patient Segmentation for Healthcare Improvement

Intermediate Data Analytics

Alex Marsella

2024-02-12

```
##   patientID age      bmi blood_pressure chol_level glucose_
## 1         1  42 25.92133            148        293
## 2         2  24 20.46337            120        162
## 3         3  21 30.40683            158        108
## 4         4  89 33.78859            106        256
## 5         5  32 36.14481            156        196
## 6         6  29 20.48509            125        128
```

Section 1

Introduction

Consulting Project Background

Our team is tasked with aiding a healthcare provider in enhancing patient care and operational efficiency.

- **Goal**: Segment patients into groups based on health metrics.
- **Purpose**: Enable personalized care plans and efficient resource allocation.

Section 2

Objective

Project Objective

To perform patient segmentation using clustering analysis on health data.

- Identify distinct patient groups.
- Utilize metrics such as Age, BMI, Blood Pressure, etc.

Section 3

Dataset Overview

Patient Data

The dataset `patient_data` includes:

- `PatientID`: Unique identifier
- `Age`, `BMI`, `Blood Pressure`, `Cholesterol Level`, `Glucose Level`

Section 4

Data Preparation

Data Preparation

Steps for Data Preparation

- Remove any identifier or categorical variable(s) that we are not going to cluster by.
- Scale the data so all variables are measured in relative distance from their own means.

```
patient_data_scaled <- as.data.frame(scale(patient_data[,
    -1]))
```

Why Scale Data Before Clustering?

Understanding Scaling

- **Scaling** changes your numbers so that different measurements are on a similar scale.
  - Converts values to a z-score such that

$$x_{scaled} = \frac{x - \bar{x}}{s}$$

  - The new value is just how many standard deviations something is from its mean.
- Imagine you have two variables in a dataset of planets: `Distance from the sun` and `Distance from the moon`.
- Without scaling, distance from the sun will dominate the clustering algorithm because it's so much larger.

The Impact of Not Scaling

- Clustering algorithms like K-means use **distance measures** to form clusters.
- If one feature is much larger than another, it will dominate the distance calculation.
- **Result**: Clusters formed may be biased towards the larger-scale features.

Benefits of Scaling

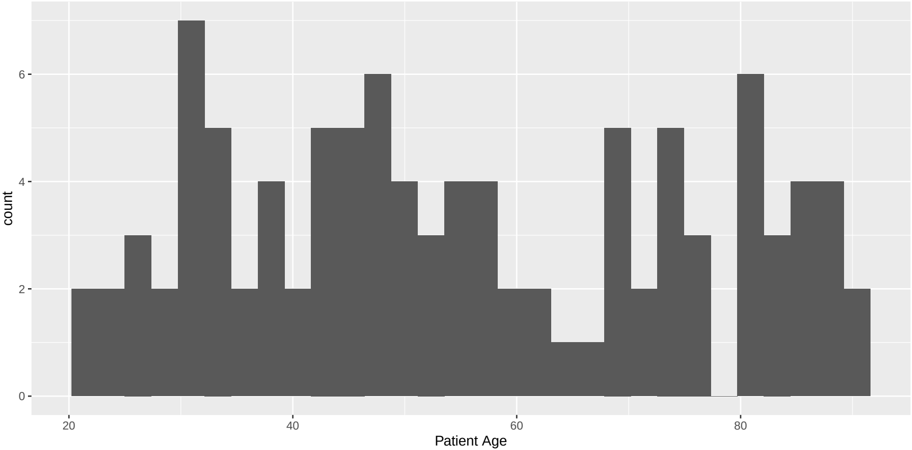
1. **Equal Footing**: Scaling puts all features on the same scale, ensuring no single feature dominates.
2. **Better Clusters**: With all features equally considered, clusters are formed based on true similarities.
3. **Faster Convergence**: Algorithms can find optimal clusters more efficiently on scaled data.

Section 5

Exploratory Data Analysis

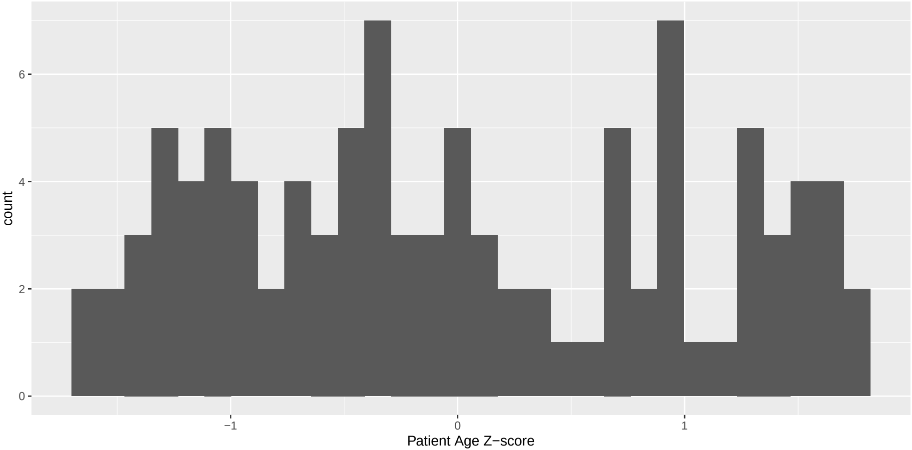
Examining the original data.

```r
ggplot(patient_data, mapping = aes(age)) +
  geom_histogram() + labs(x = "Patient Age")
```



Examining the scaled data.

```
ggplot(patient_data_scaled, mapping = aes(age)) +
    geom_histogram() + labs(x = "Patient Age Z-score")
```



Section 6

Clustering Analysis

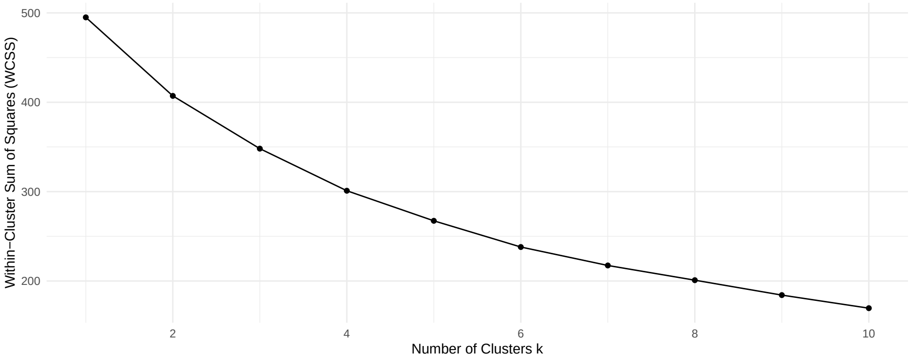
K-means Clustering

Finding Optimal Number of Clusters

```
wcss <- numeric(10)
for (k in 1:10) {
    # k will take on the value 1,
    # then 2, and so on, repeating
    # to 10
    kmeans_result <- kmeans(patient_data_scaled,
        centers = k, nstart = 20)
    wcss[k] <- kmeans_result$tot.withinss
}
```

The elbow plot

```
ggplot(, aes(x = 1:10, y = wcss)) + geom_line() +
    geom_point() + xlab("Number of Clusters k") +
    ylab("Within-Cluster Sum of Squares (WCSS)") +
    theme_minimal() + scale_x_continuous(breaks = pretty_breal
```



Clustering

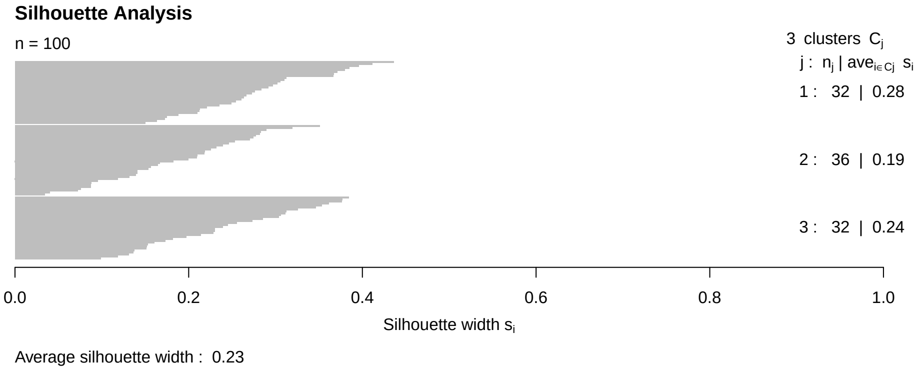
```
kmeans_result <- kmeans(patient_data_scaled,
    3, nstart = 20)
# Examine cluster centroids
kmeans_result$centers
```

```
##          age        bmi blood_pressure chol_level glucose_l
## 1 -0.7220577  0.8661437      0.3886815  0.3197159    -0.110
## 2  1.0206135  0.1450682     -0.4583293 -0.1462330    -0.314
## 3 -0.4261324 -1.0293454      0.1269390 -0.1552037     0.464
```

```
# Assign cluster membership back to
# the original data
patient_data_scaled$cluster <- as.factor(kmeans_result$cluster
```

Silhouette Analysis

```
sil <- silhouette(kmeans_result$cluster,
    dist(patient_data_scaled))
plot(sil, main = "Silhouette Analysis")  # base R plotting is
```



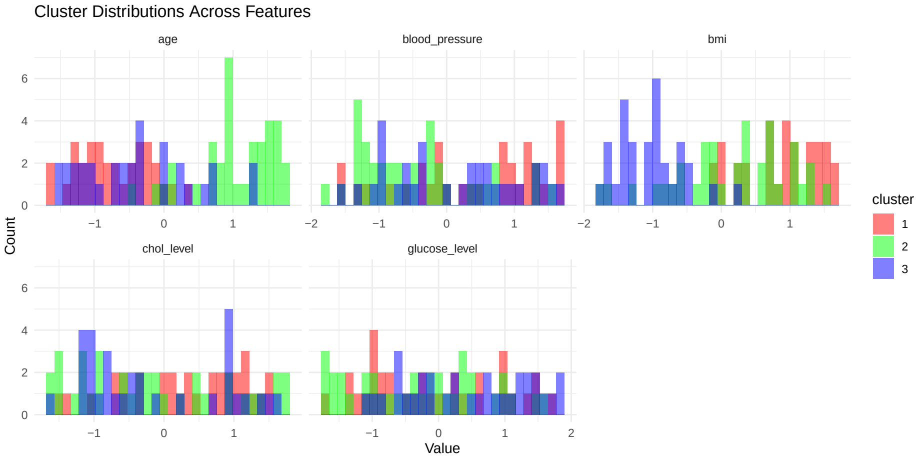
Section 7

Advanced Visualization Methods

Making our data long using gather() ()

```
patient_data_long <- patient_data_scaled %>%
    gather(key = "Feature", value = "Value",
        -cluster)
```

Visualizing it



Section 8

Takeaway

Analysis and Strategic Recommendations

- Interpretation of clusters.
- Suggest healthcare interventions.

Subjectivity of it all.

- Clustering analysis is practically tea leaf reading, in my opinion.
- But many firms use it. You may very well be asked to do it as an analyst.
- Lots of subjectivity, up to you to visualize and come up with your own thoughts.