Stroke Data Analysis

Kai Wilson

2024-10-28

Stroke Data Analysis

In this project, I obtained this dataset from Kaggle.The data is based on people that have and haven't had a stroke. The dataset was used for training a model to make predictions if a person is likely to have a stoke.

The data is based on people that have and haven't had a stroke. The dataset was used for training a model to make predictions if a person is likely to have a stroke.

After I collected the data, cleaned it, and put it into a SQLite database. For the database schema, I made one table for person and another table for their medical history.

I queried the data through joining the two tables and completing my analysis. After I did three queries, I turned them into their own CSV files to quickly graph the data using ggplot. Taking a step further, I created two K-means clusters: one focusing on just the medical history and another with medical history based on gender.

All code can be found on my github: https://github.com/Kwilso3412/medical_data_analysis

Totals

Had a Stroke: 249

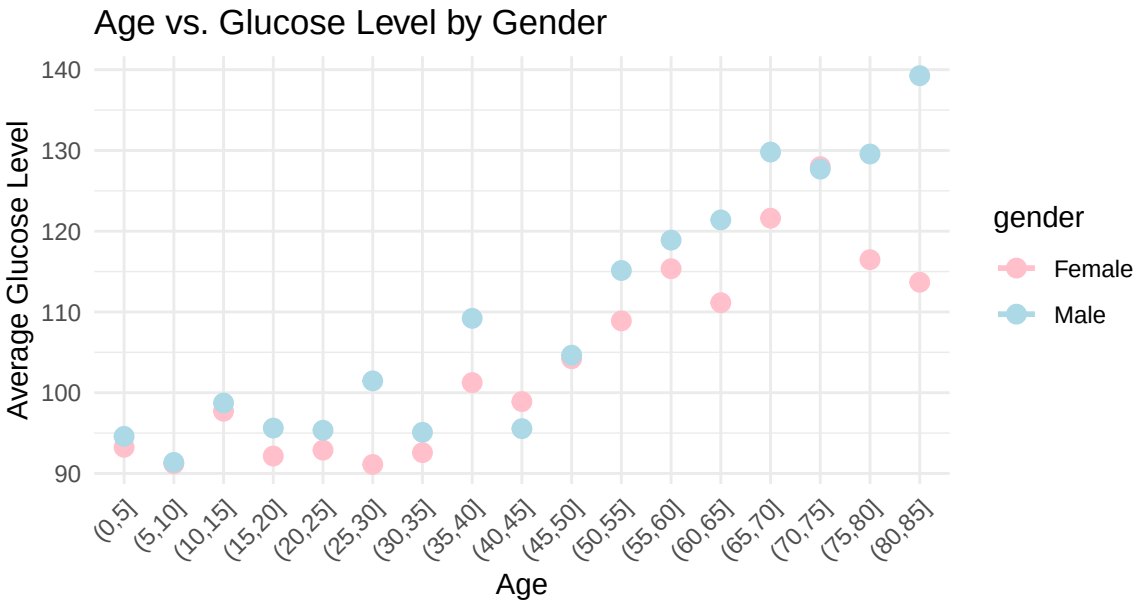
Havent had a Stroke: 4861

Men that have had a stroke: 108

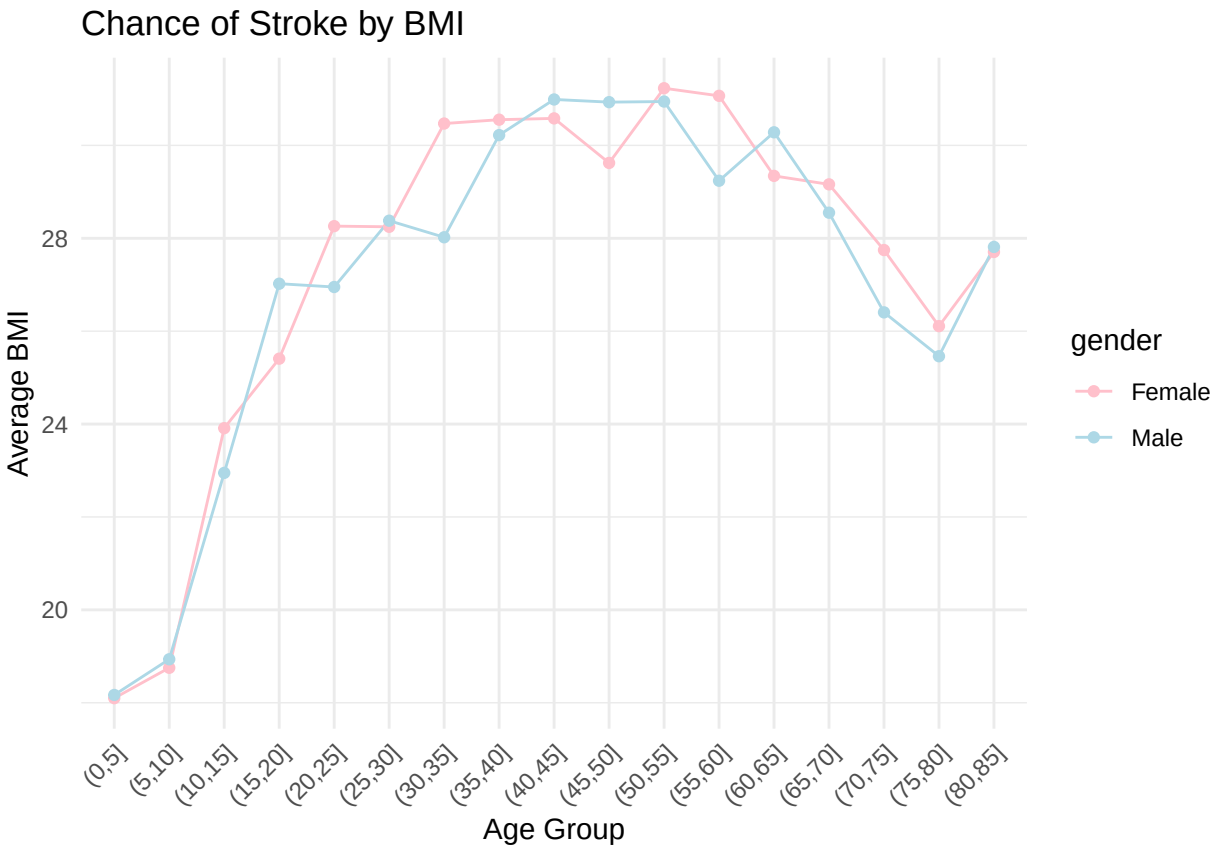
Men married had stroke: 100

Women that have had a stroke: 141

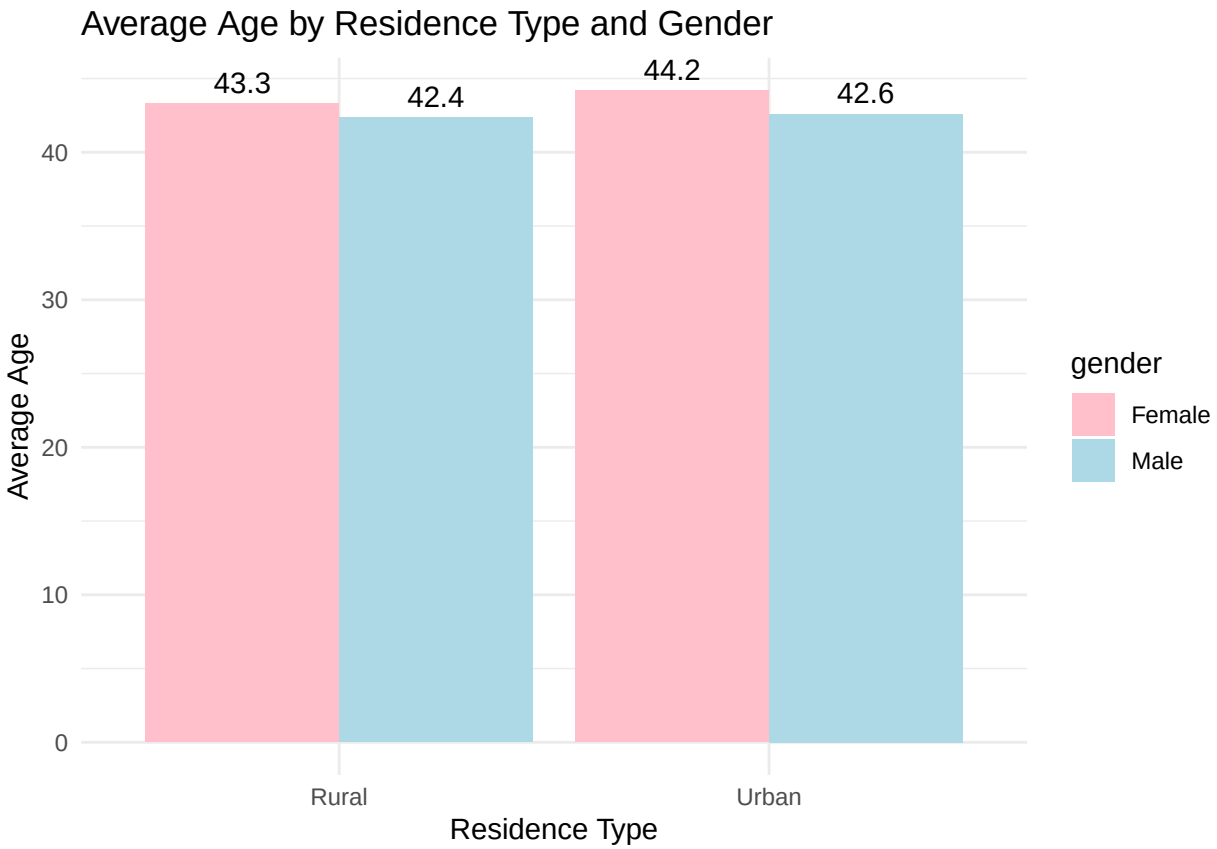
Women married had stroke: 120



The graph shows that blood glucose levels tend to increase with age for both men and women. While levels stay fairly stable until age 30, they begin rising after that, with a sharper increase after age 50. Men (blue line) generally show higher glucose levels than women (pink line), with this difference becoming more noticeable in older age groups, particularly after age 60.



This graph shows the relationship between age and BMI (body mass index) for men and women. BMI increases with age until around 45-50 years old, then gradually decreases in later years. Both genders follow similar patterns, with slight variations but no major differences between men and women.



This graph shows the average age of people living in rural versus urban areas, broken down by gender. The differences are small, but urban women have the highest average age (44.2 years), while rural men have the lowest (42.4 years). Women in both locations are slightly older on average than men.

Clustering Based on Medical History

Cluster 1 (Red):

Primarily young individuals (average age ~29.4) Healthy BMI range (~25.6) Normal glucose levels (~98.2) Very low rates of hypertension (2.6%) and heart disease (2.3%) All marked as unknown smoking status Represents the "healthy young adult" group

Cluster 2 (Blue):

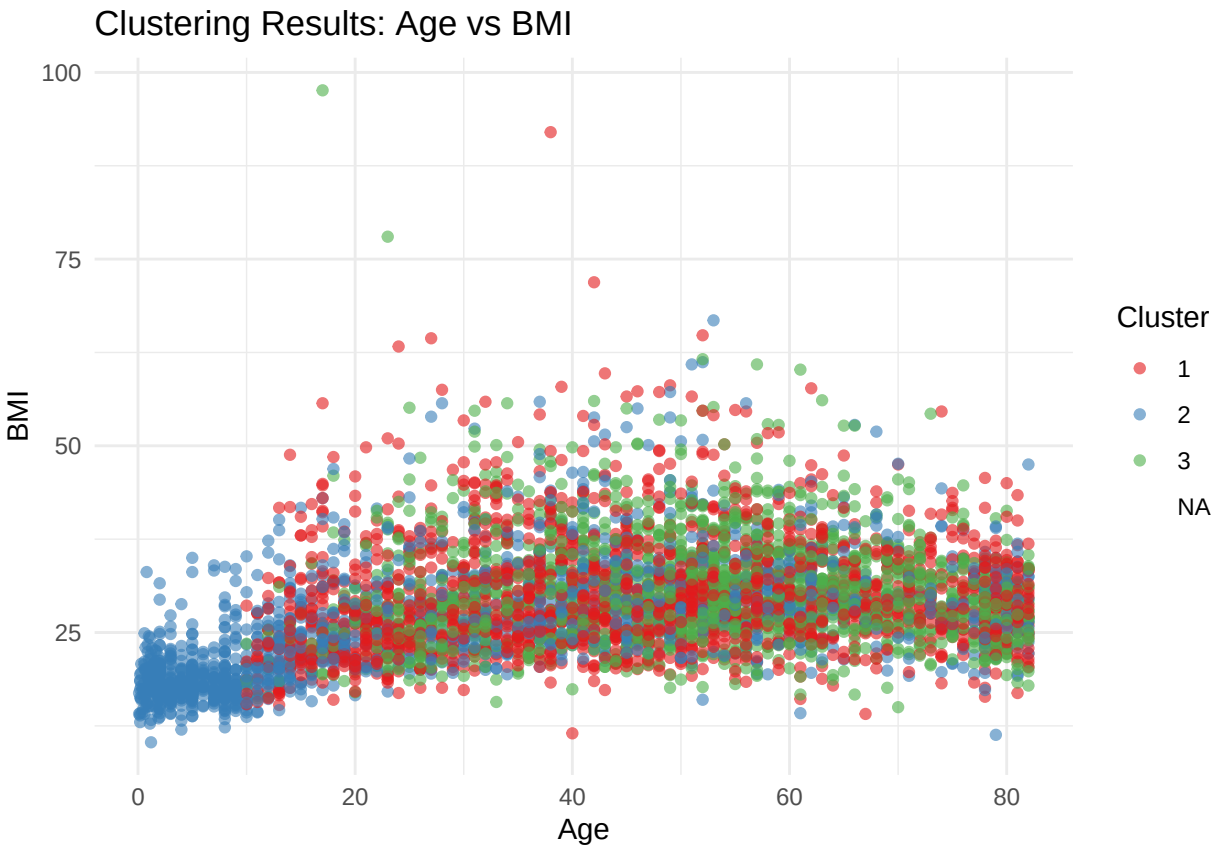
Oldest group (average age ~51.2) Highest BMI (30.7, classified as obese) Highest glucose levels (~109.8) Highest rates of hypertension (12.4%) and heart disease (8.1%) Mixed smoking status: 53% former smokers, 46.7% current smokers Represents the "high-risk" group

Cluster 3 (Green):

Middle-aged group (average age ~46.5) Elevated BMI (30.0, borderline obese) Elevated glucose levels (~107.1) Moderate hypertension (11.7%) and heart disease (4.4%) All never-smokers Represents the "moderate-risk" group

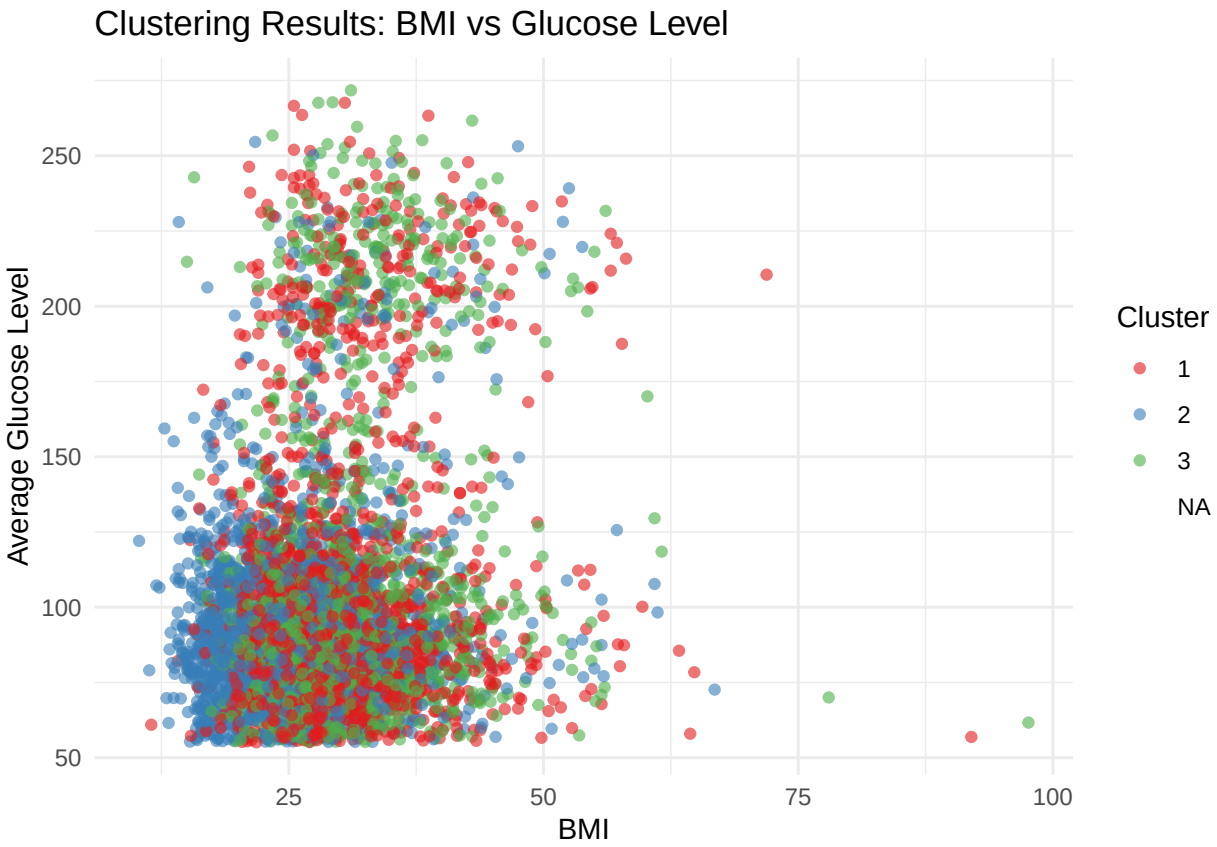
```
## [1] "Cluster Summary Statistics:"
```

```
##   cluster count      age avg_glucose_level      bmi hypertension heart_disease
## 1       1     1 46.46922         107.13621 29.98256   0.11663067    0.04373650
## 2       2     1 29.42522          98.17639 25.59905   0.02636917    0.02298851
## 3       3     1 51.24857         109.81642 30.70755   0.12428662    0.08116677
##   formerly_smoked never_smoked    smokes     Unknown
## 1       0.0000000            1 0.0000000 0.000000000
## 2       0.0000000            0 0.0000000 1.000000000
## 3       0.5301205            0 0.4673431 0.002536462
```



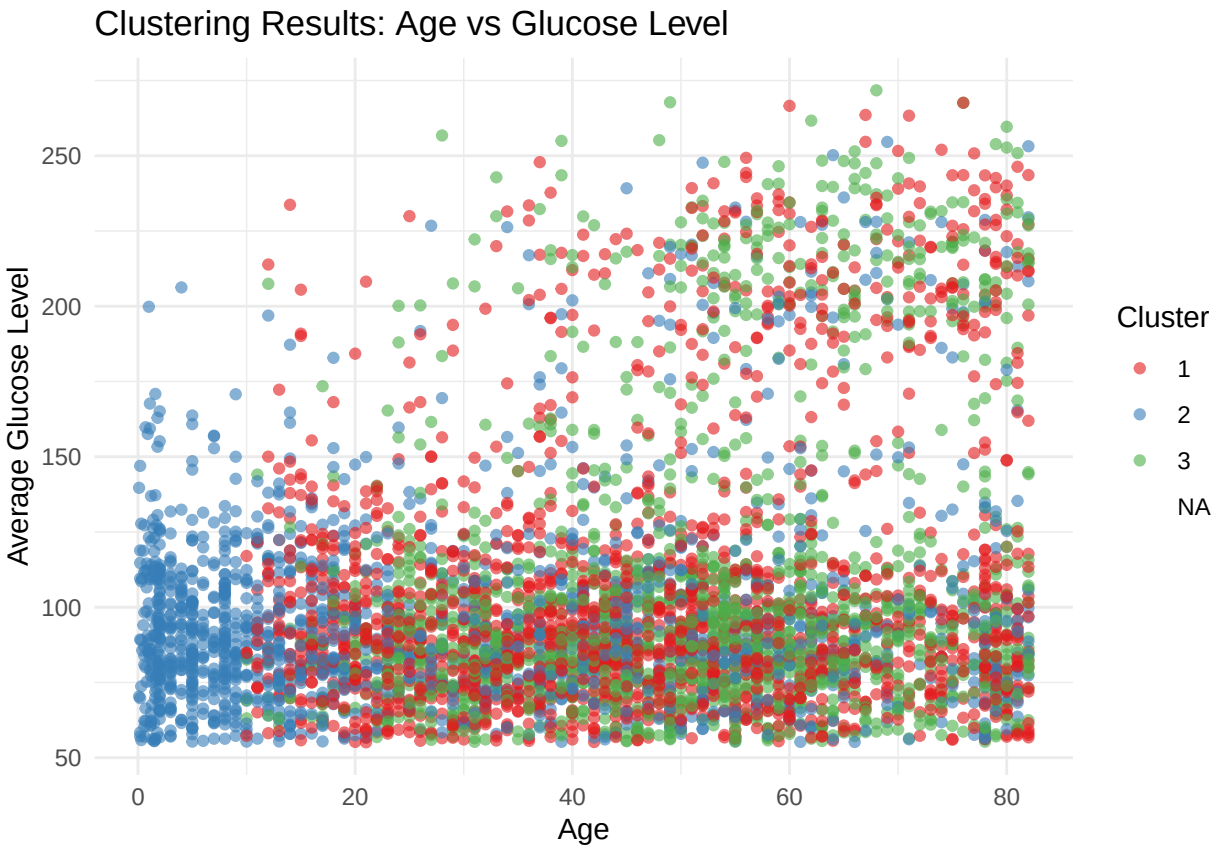
Age vs BMI:

Shows clear age-related BMI increases Younger cluster (1) maintains healthier BMI range More variation in BMI as age increases



BMI vs Glucose Level:

Shows positive correlation between BMI and glucose levels Higher BMI associated with more variable glucose levels Clear separation between healthy (Cluster 1) and at-risk groups (Clusters 2 and 3)



Age vs Glucose Level:

Strong positive correlation between age and glucose levels More scattered/variable glucose levels in older age groups Cluster 1 (young) shows tighter grouping of normal glucose levels

In Summary

Age is a strong determining factor in health outcomes Risk factors tend to cluster together (high BMI, glucose, hypertension, heart disease) Smoking status appears to be an important discriminating factor between clusters Clear progression from healthy young adults to higher-risk older adults Potential for early intervention in younger populations to prevent progression to higher-risk clusters

Clustering Based on Gender: Gender Distribution in Clusters

Cluster 1:

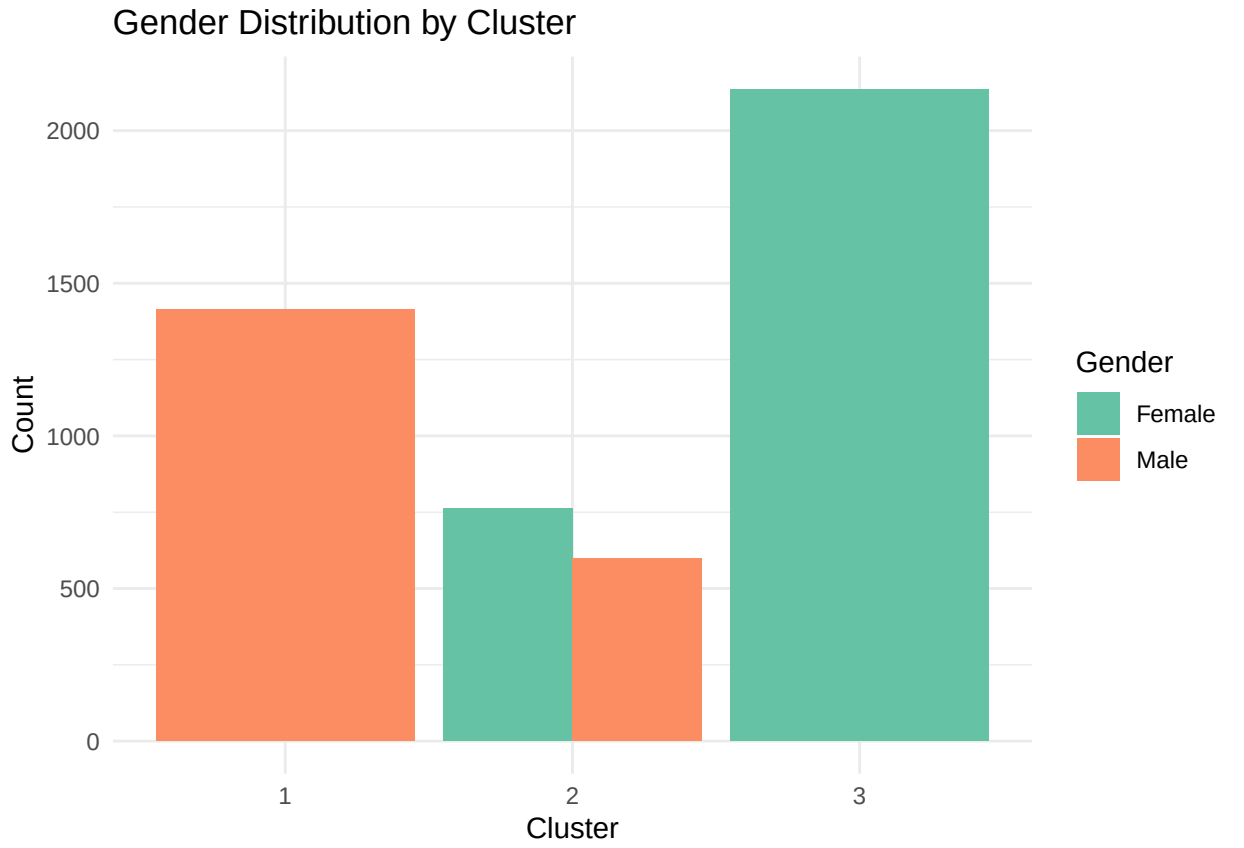
More balanced gender distribution Approximately 750 females and 600 males This was the "younger, healthier" cluster with lower BMI, lower glucose levels, and minimal health risks

Cluster 2:

Predominantly male Approximately 1400 males Very few or no females This represented the "higher risk" group with higher BMI, glucose levels, and cardiovascular risks

Cluster 3:

Predominantly female Approximately 2000 females Very few or no males This represented the female equivalent of Cluster 2, with similar age and health risk patterns



This visualization reveals that the clustering algorithm effectively separated the middle-aged/higher risk population by gender (Clusters 2 and 3), while keeping younger, healthier individuals of both genders together in Cluster 1. This gender segregation in the higher-risk clusters might be useful for developing targeted health interventions specific to each gender's risk patterns.

Clustering Based on Gender

Cluster 1:

Predominantly younger individuals (Female avg: 29.4 years, Male avg: 22.2 years) Lower BMI (Female: 25.3, Male: 23.9) - both in the normal/slightly overweight range Lower glucose levels (Female: 94.6, Male: 93.5) - both within normal range Very low rates of hypertension and heart disease (<1%) All marked as "Unknown" for smoking status (unknown_mean = 1) Represents the healthiest cluster overall

Cluster 2/3 (higher risk group): Females (Cluster 3):

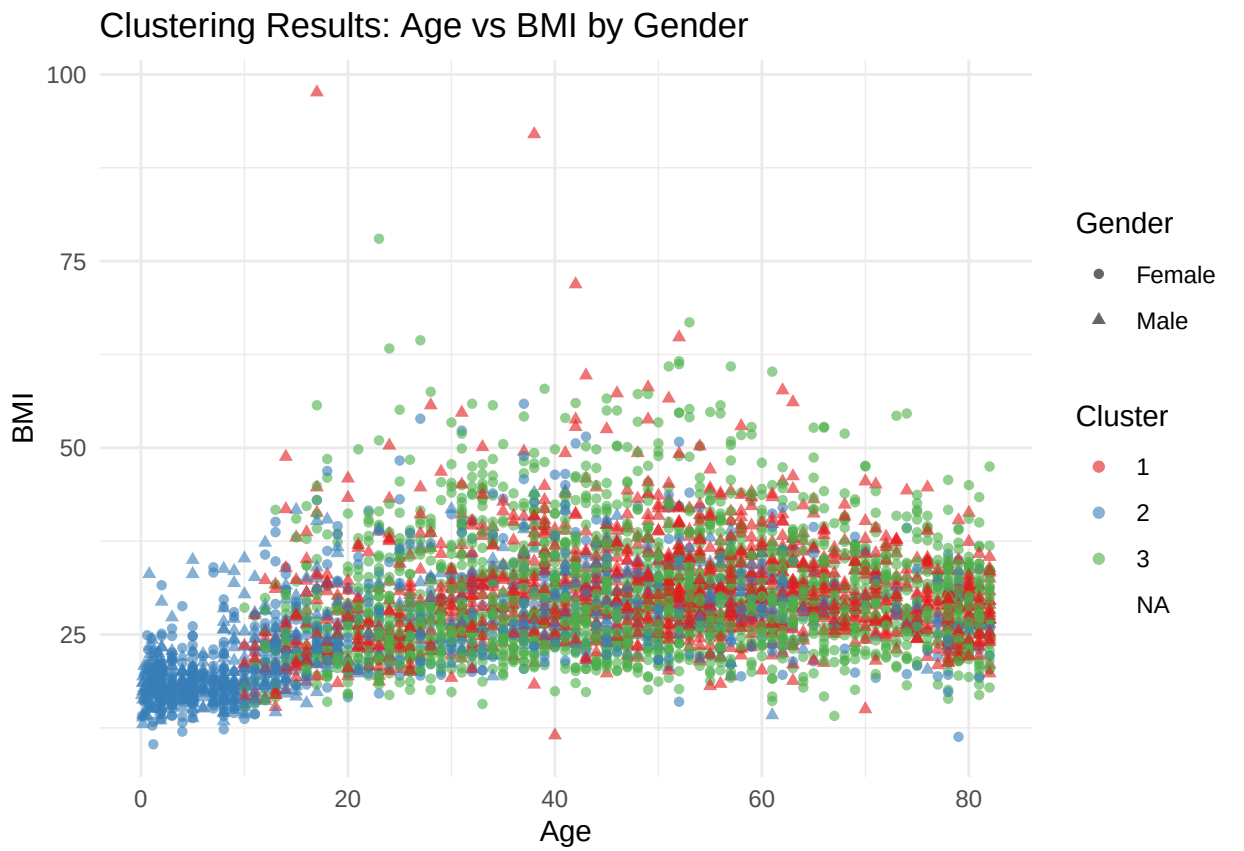
Middle-aged (avg: 48.4 years) Higher BMI (30.4) - in the obese range Elevated glucose (106.5) Higher rates of hypertension (11.6%) and heart disease (4.6%) Mixed smoking status: 21.4% formerly smoked, 56.3% never smoked, 19.9% current smokers

Males (Cluster 2):

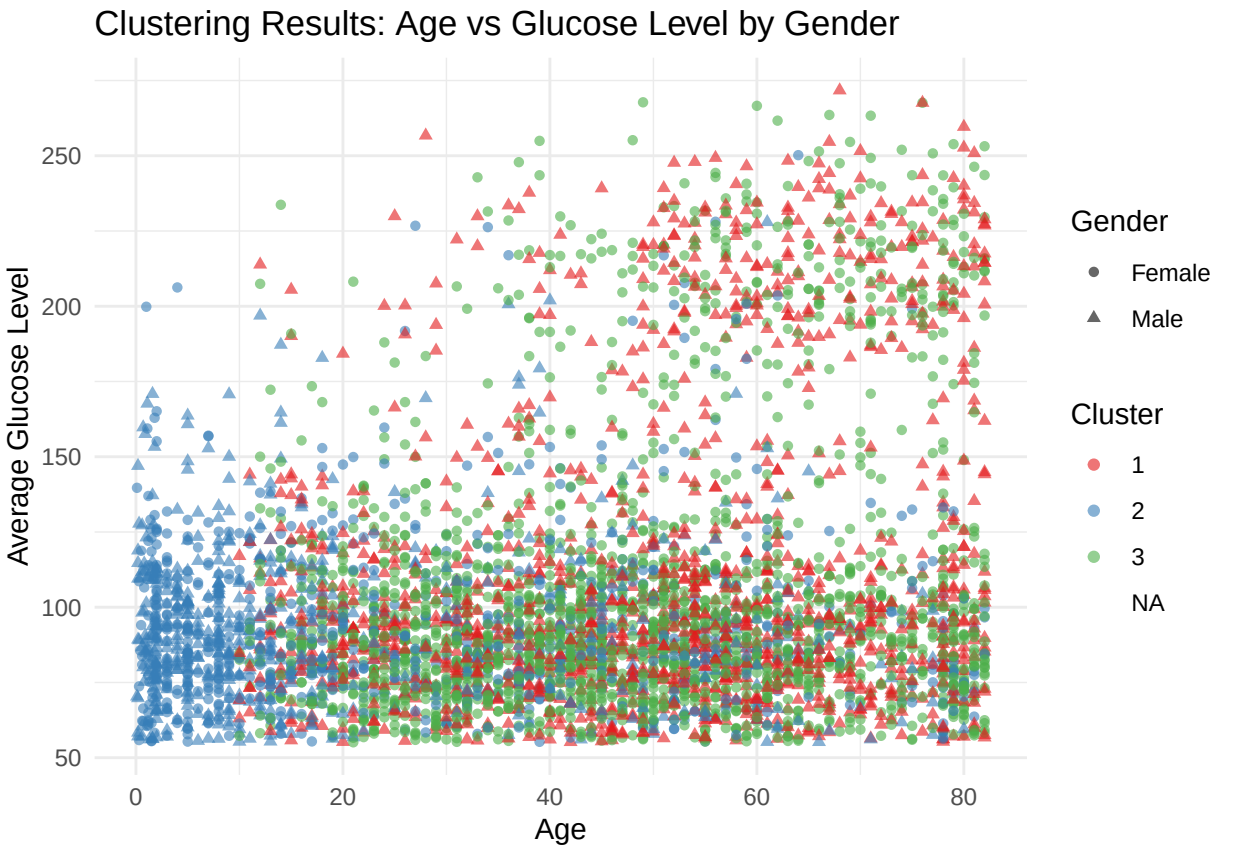
Slightly older (avg: 50.5 years) Similar BMI (30.7) - also in obese range Highest glucose levels (114.3) Highest rates of hypertension (14.1%) and heart disease (10%) Different smoking pattern: 26.8% formerly smoked, 45.9% never smoked, 22.1% current smokers

```
## [1] "Cluster Summary Statistics by Gender:"
```

```
##   cluster gender count age_mean glucose_mean bmi_mean hypertension_mean
## 1       2 Female     1 29.39448     94.55197 25.32431       0.003942181
## 2       3 Female     1 48.44101    106.45726 30.39874       0.116104869
## 3       1   Male     1 50.46709    114.32844 30.66617       0.140835103
## 4       2   Male     1 22.15836     93.48965 23.87910       0.001672241
##   heart_disease_mean formerly_smoked_mean never_smoked_mean smokes_mean
## 1        0.002628121            0.0000000         0.0000000   0.0000000
## 2        0.045880150            0.2144195         0.5632022   0.1989700
## 3        0.100495400            0.2675159         0.4593064   0.2208068
## 4        0.001672241            0.0000000         0.0000000   0.0000000
##   unknown_mean
## 1   1.00000000
## 2   0.02340824
## 3   0.05237084
## 4   1.00000000
```



This scatter plot shows two main clusters in the Age vs BMI relationship: Cluster 1 (red) shows younger people (mostly under 40) with lower BMI ranges (mostly between 20-30), suggesting a healthier weight group. Cluster 2 (green) shows a wider age range with higher BMI values (mostly between 25-40), suggesting an overweight/obese group across different ages. The pattern suggests that while younger people tend to have lower BMI, there's more variation in BMI as age increases, with some people maintaining healthy weights while others show higher BMI values.



This scatter plot shows Age vs Glucose Levels clustered into two main groups: Cluster 1 (red) shows people with normal glucose levels (mostly between 60-120), appearing more commonly in younger ages. Cluster 2 (green) shows elevated glucose levels (above 120), becoming more common with increasing age, possibly indicating pre-diabetic or diabetic ranges. The pattern suggests that glucose levels tend to increase with age, and older individuals are more likely to have higher glucose readings. The separation between clusters could help identify potential diabetes risk groups across age ranges.

In Summary

The data shows clear age-related clustering and notable gender differences in health risks, with males showing higher glucose levels and cardiovascular risks. There are similar BMI patterns across genders within the same age groups. Risk factors tend to cluster together, as higher age correlates with higher BMI and glucose levels. Additionally, smoking status varies significantly by both age group and gender.