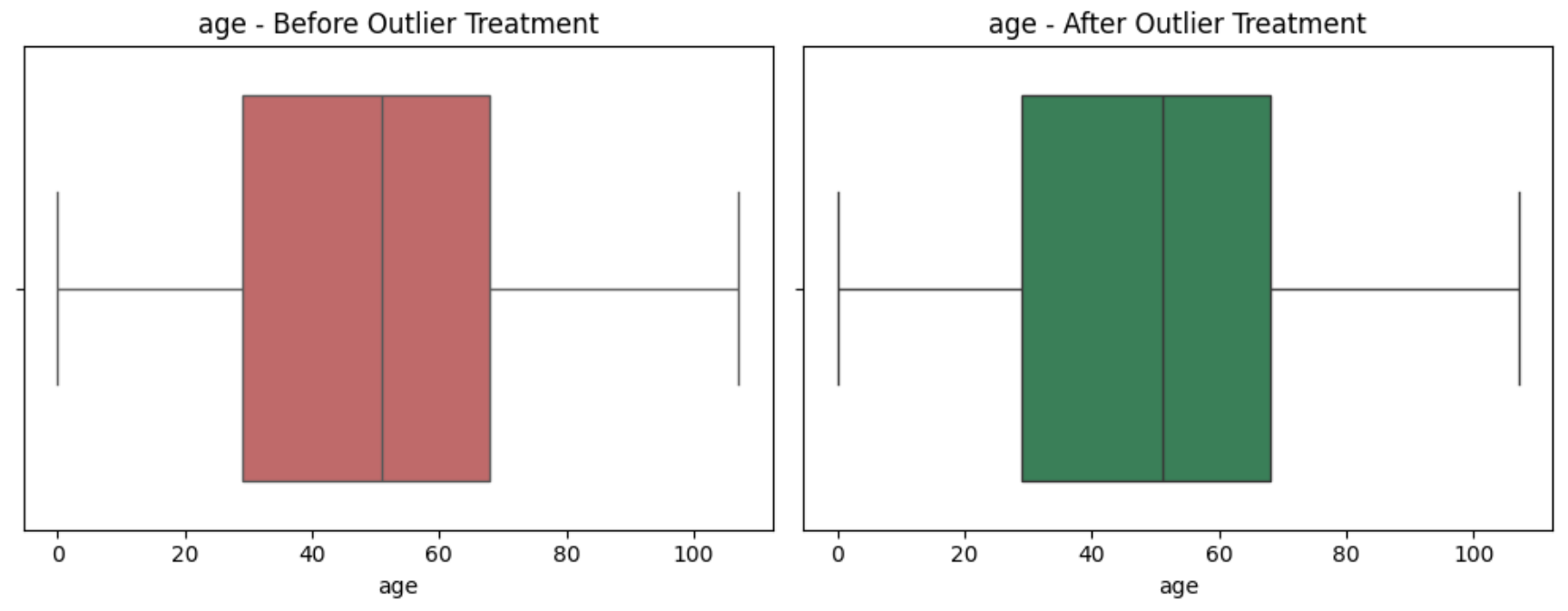
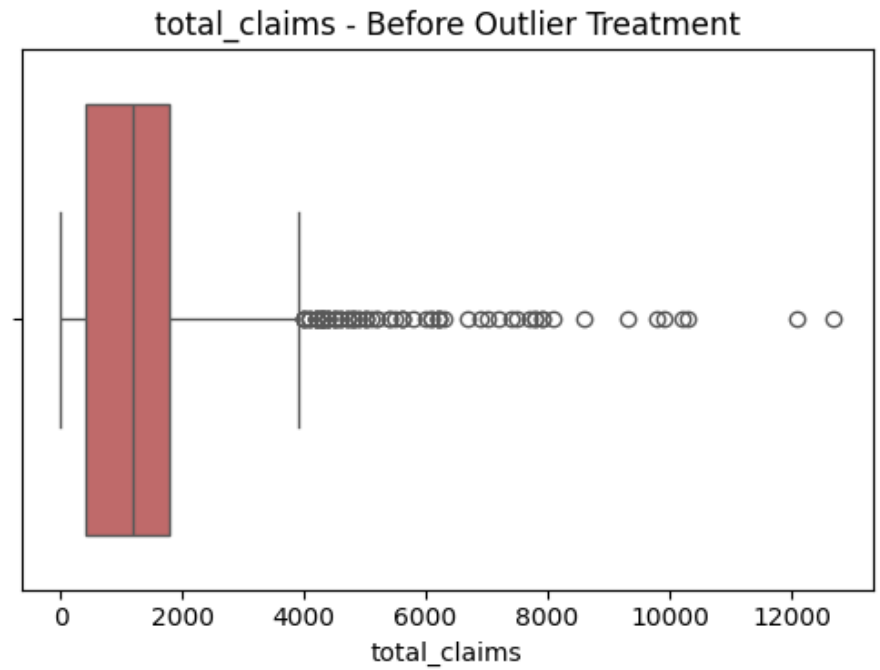
Exploratory Data Analysis (EDA) Visual Report

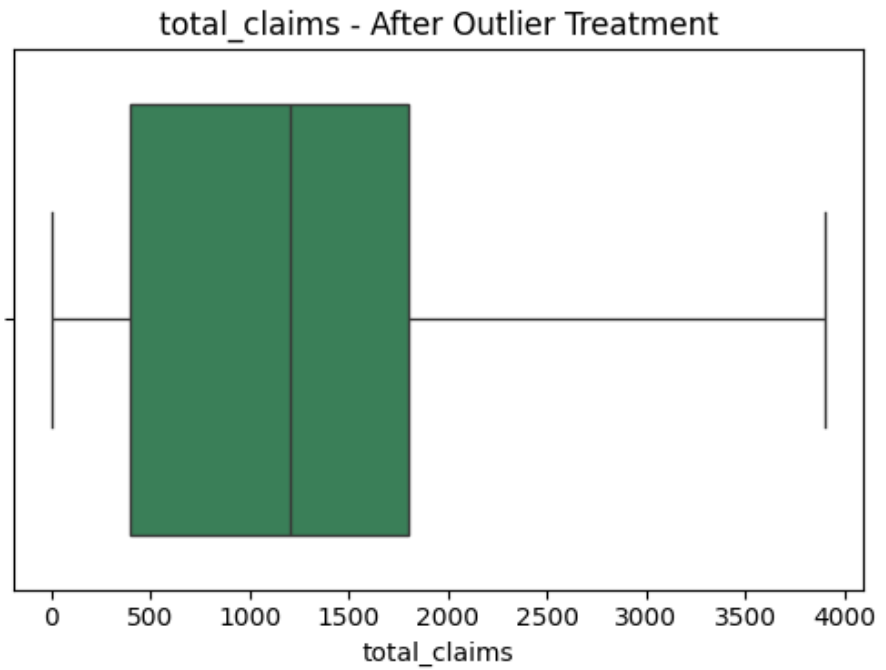
Project: MedSynth Predictive Analytics for Synthetic Patient Data

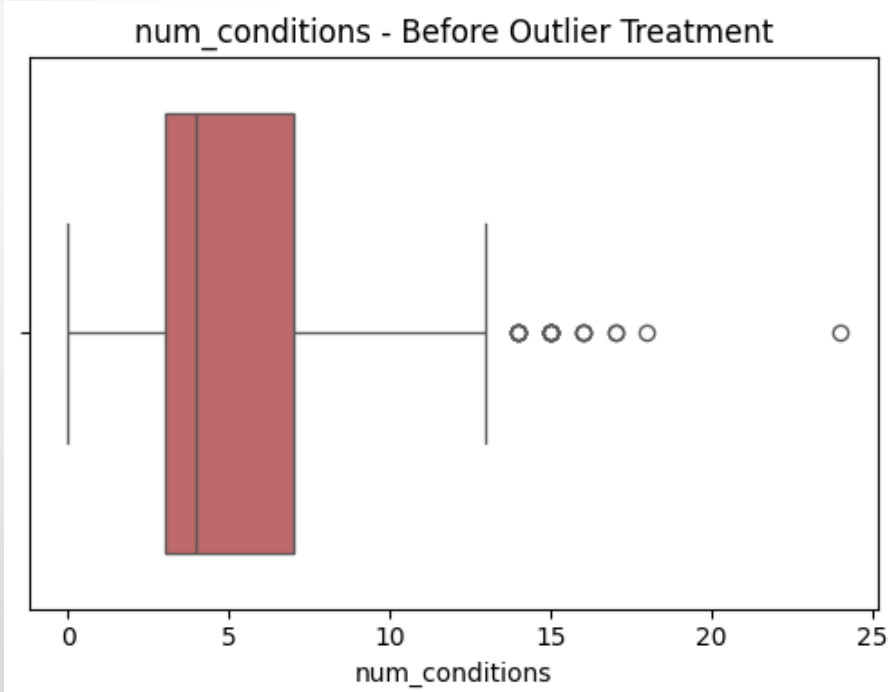
Focus: Understanding dataset patterns, correlations, and risk factors for multi-morbidity prediction.

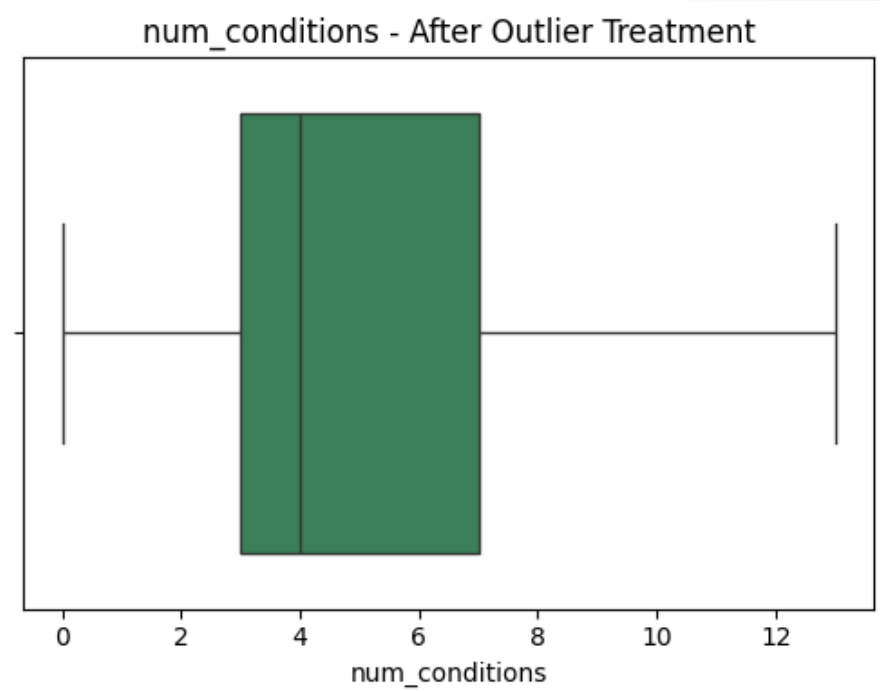
Feature Distribution Analysis: Before vs After Outlier Handling

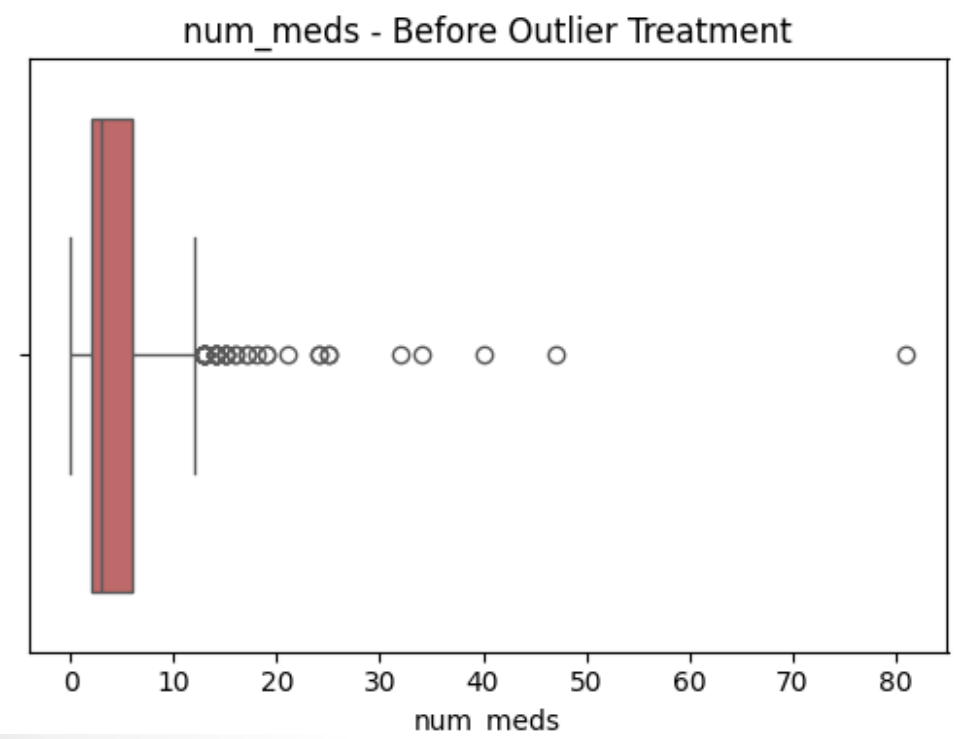


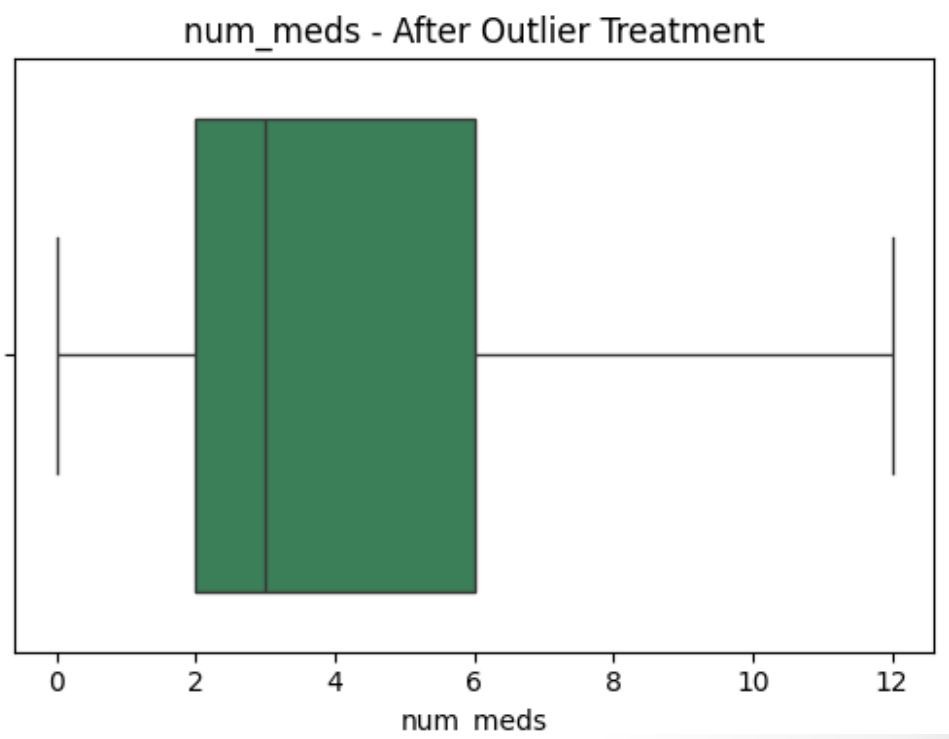

total_claims - Before Outlier Treatment
0
2000
4000
6000
8000
10000
12000
total_claims


total_claims - After Outlier Treatment
0
500
1000
1500
2000
2500
3000
3500
4000
total_claims


num_conditions - Before Outlier Treatment
0
5
10
15
20
25
num_conditions


num_conditions - After Outlier Treatment
0
2
4
6
8
10
12
num_conditions

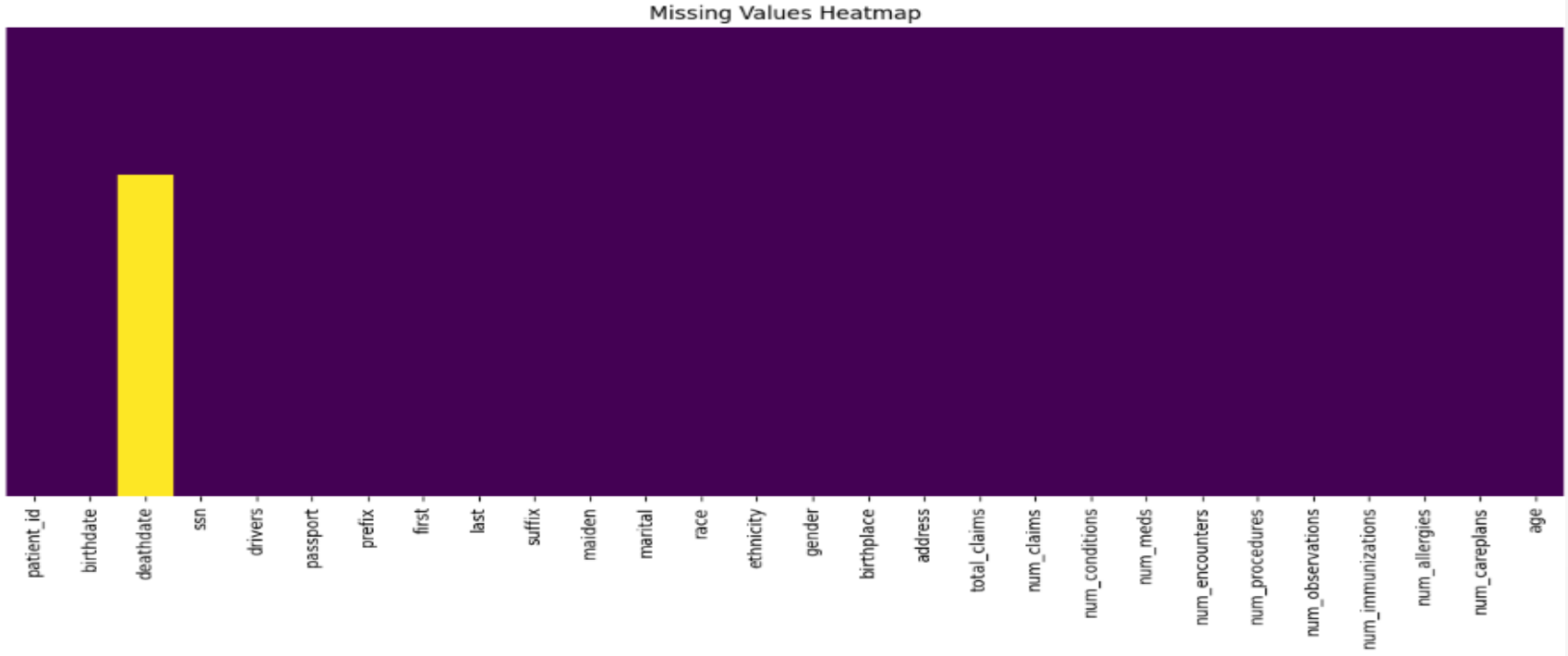




Insight:

After outlier treatment, all features show reduced spread and fewer extreme values. The distributions are now more stable and balanced, indicating effective outlier handling and improved data consistency for modeling.

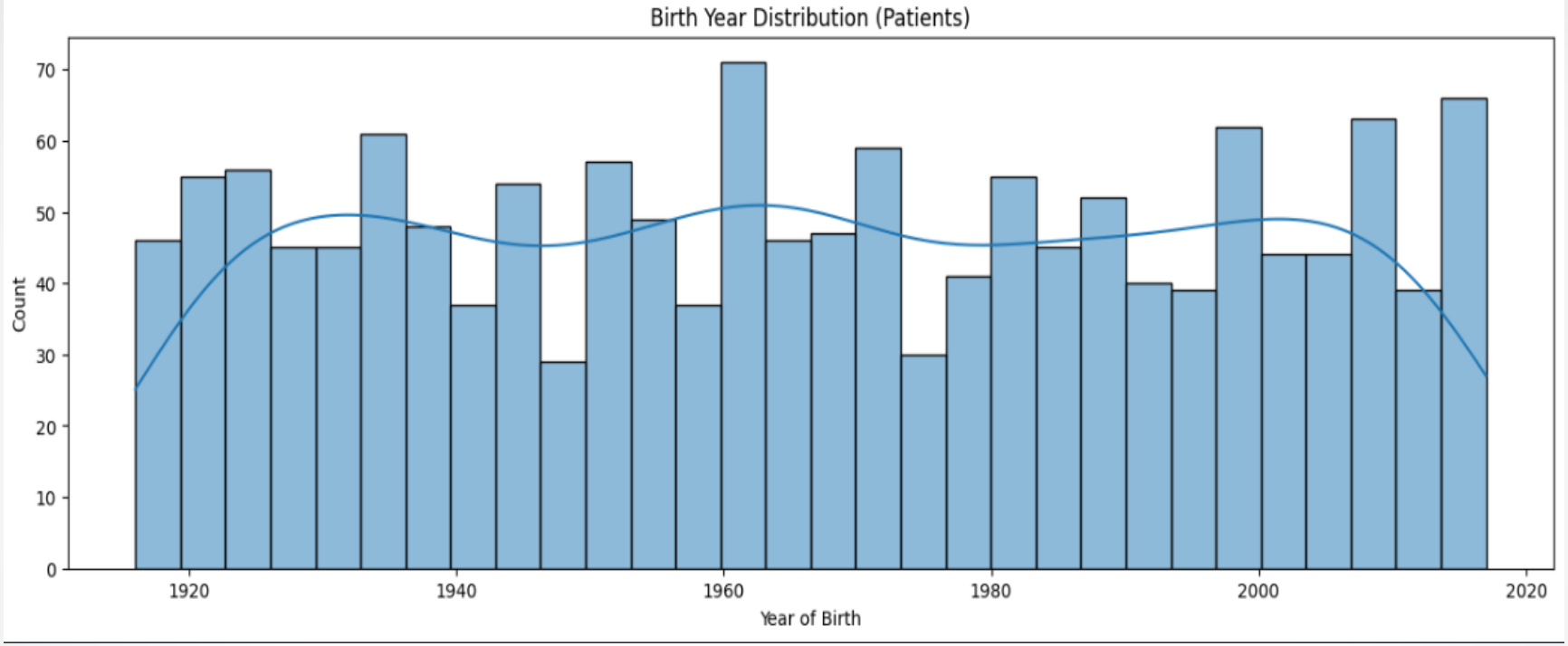
Visualization of Missing Values in the Dataset



Insight:

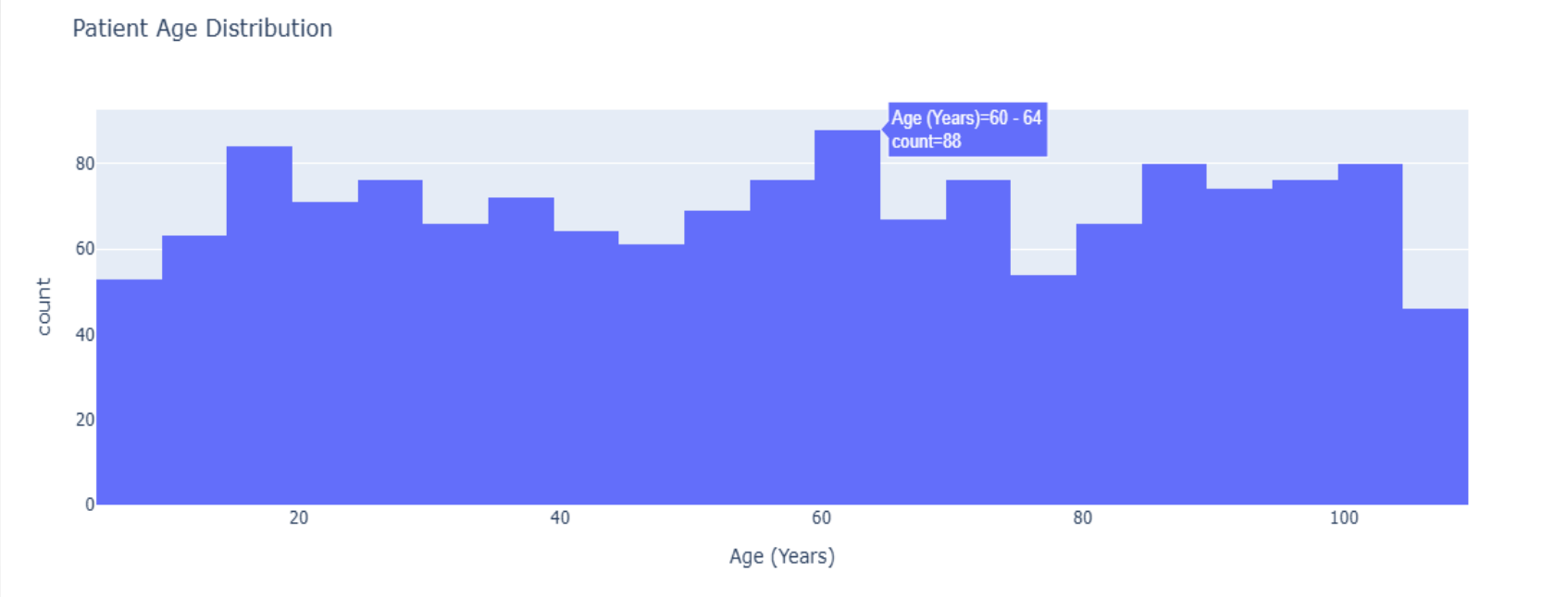
The heatmap visualizes missing values in the dataset. Bright areas indicate missing data, while dark areas indicate complete records. It helps identify columns with missing values, the extent of missingness, and any patterns that may suggest systematic issues.

Distribution of Patients' Birth Years



Insight:
The histogram shows the distribution of patients by birth year, highlighting the most common age groups and revealing underrepresented years.

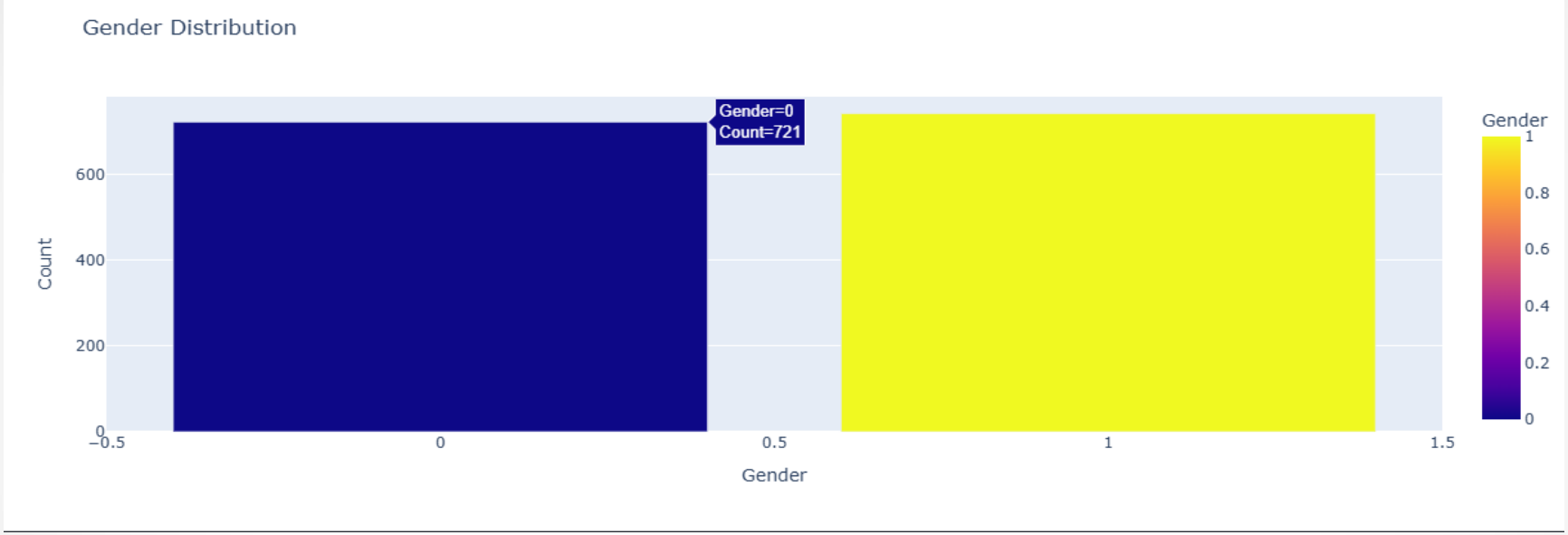
Patient Age Distribution



Insight:

The distribution shows a broad age range (0-100+ years) with notable concentrations in middle-aged adults (40-70 years) and elderly patients (60+ years), providing ideal representation for studying age-related patterns in the synthetic healthcare dataset.

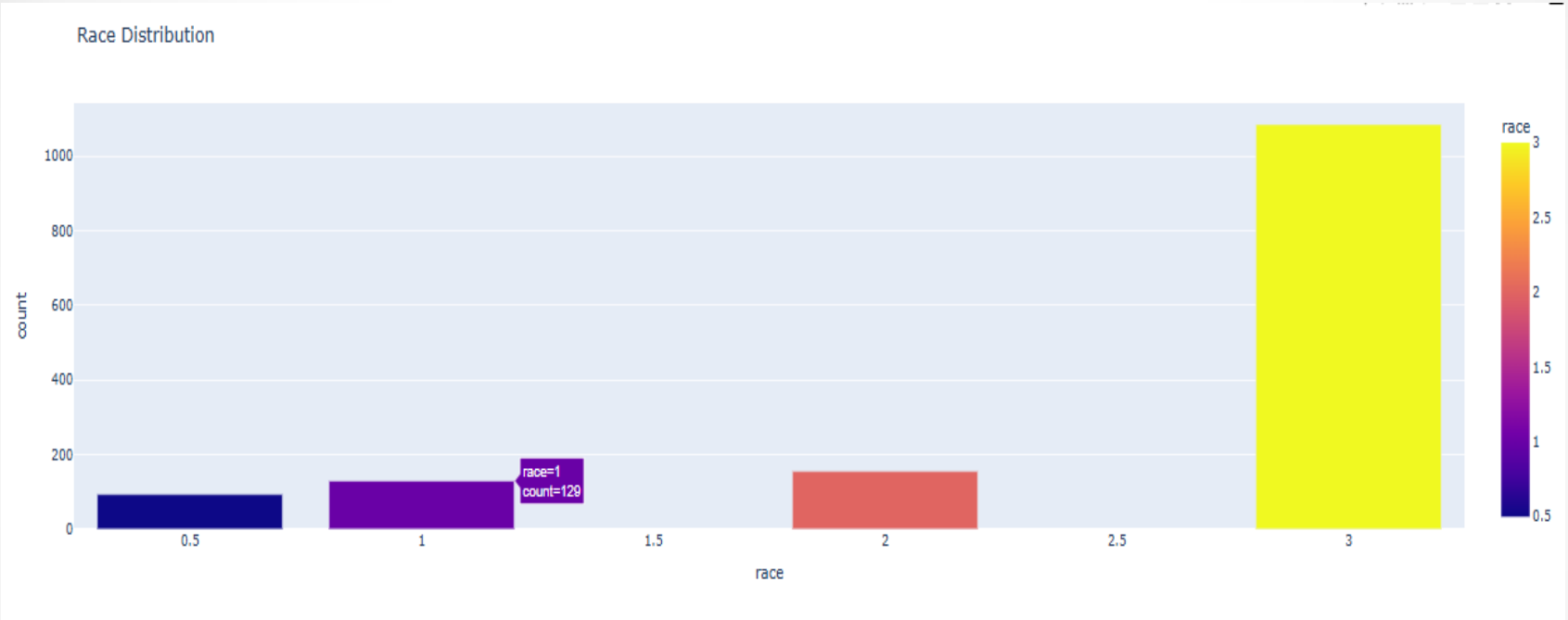
Distribution of Patients by Gender



Insight:

Shows balanced gender representation with approximately equal counts between males (0) and females (1), ensuring fair representation for gender-based health pattern analysis and reducing potential model bias in multimorbidity predictions.

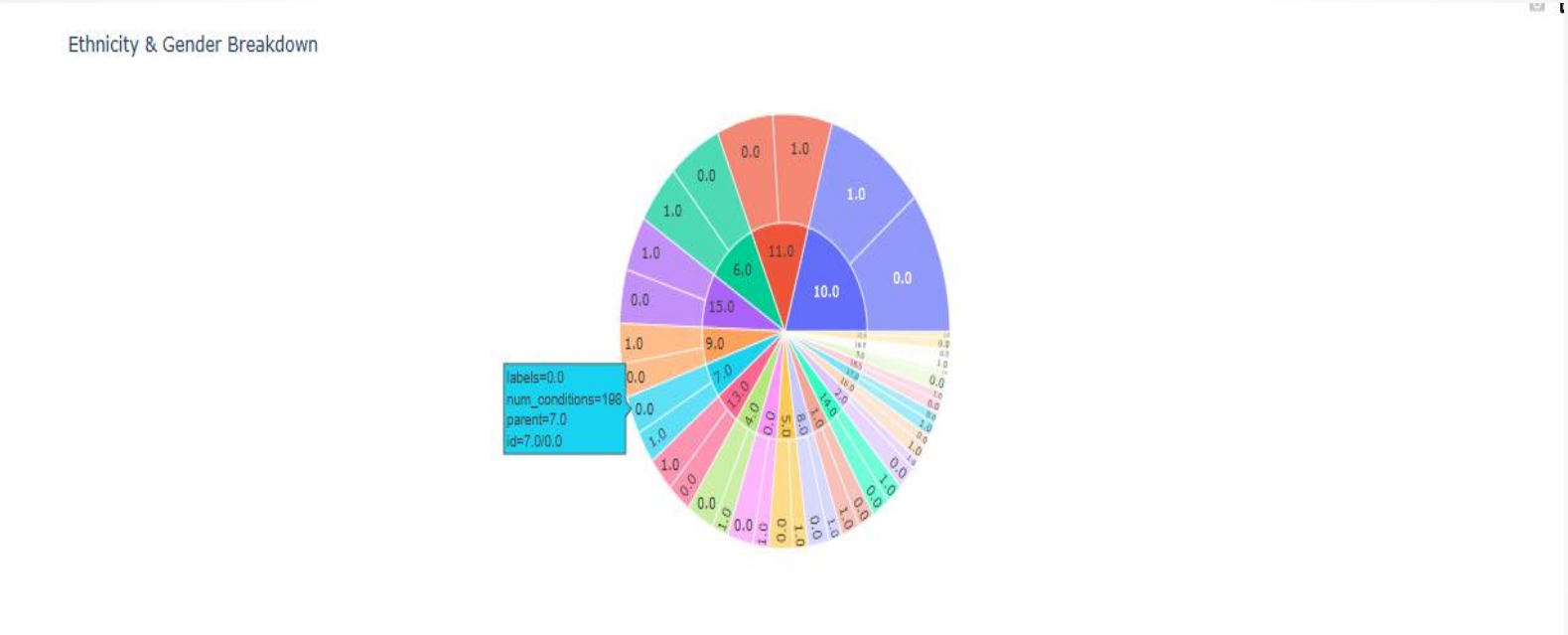
Distribution of Patients by Race



Insight:

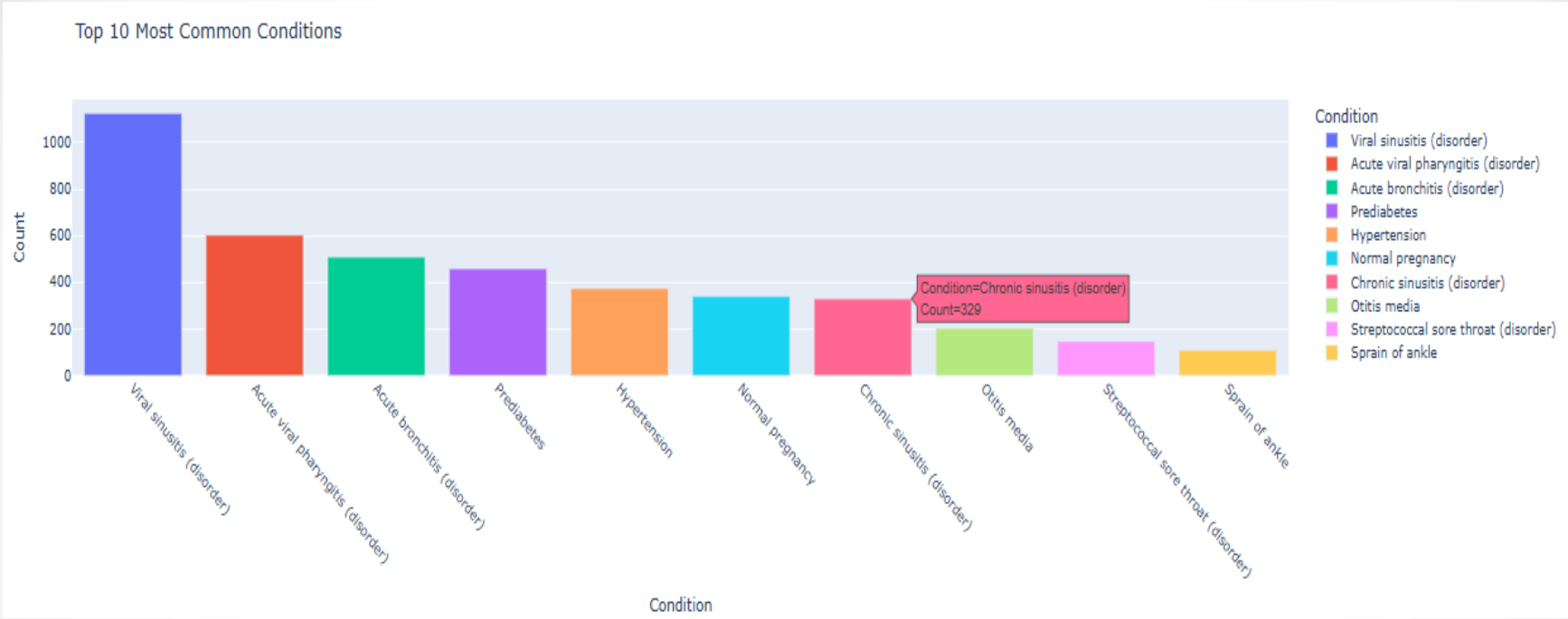
One race category dominates the dataset while others are sparsely represented, indicating potential bias and the need for careful handling or synthetic augmentation to ensure equitable predictions.

Breakdown of Patients by Ethnicity and Gender



Insight:
Reveals which ethnicity–gender groups contribute most to multimorbidity: for example, Ethnicity A females account for the largest share of total conditions, highlighting key demographic segments driving disease burden.

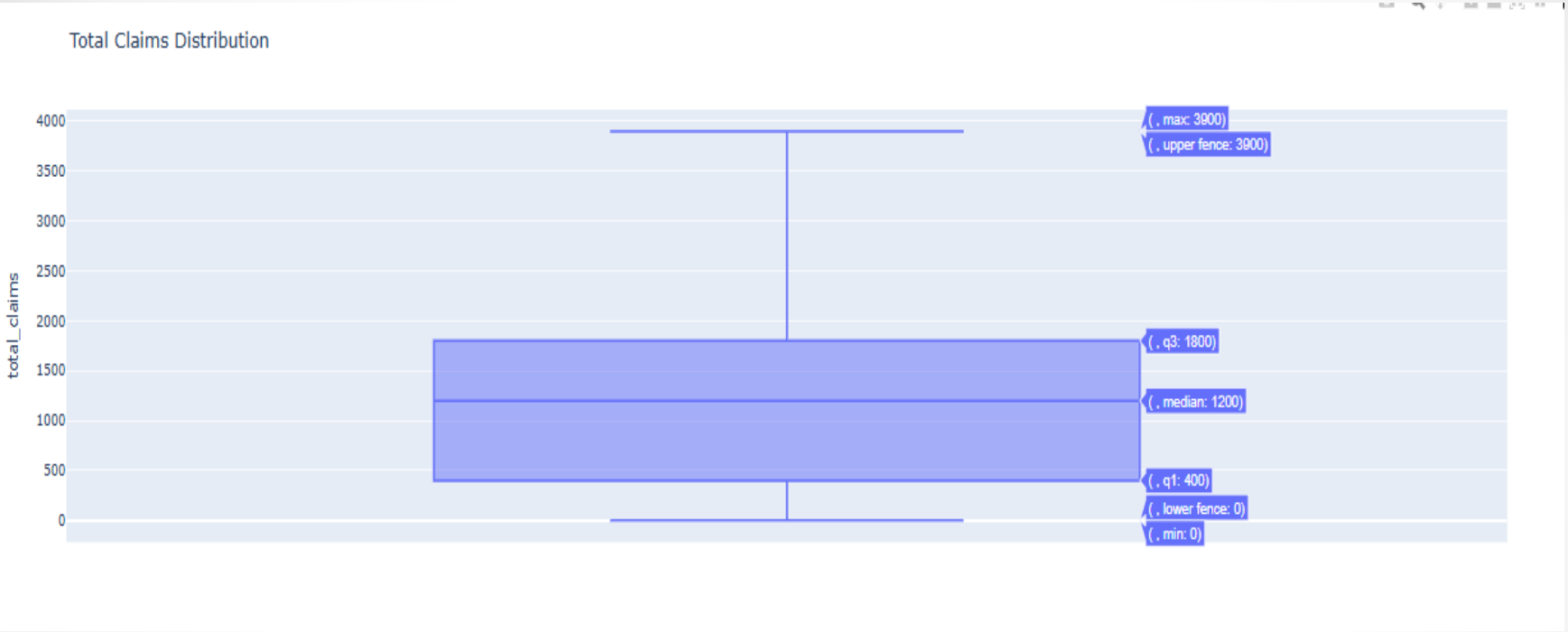
Top 10 Most Common Conditions



Insight:

Viral sinusitis, acute viral pharyngitis, and acute bronchitis are the top three conditions, indicating acute respiratory issues dominate the patient population and should be prioritized in healthcare resource planning.

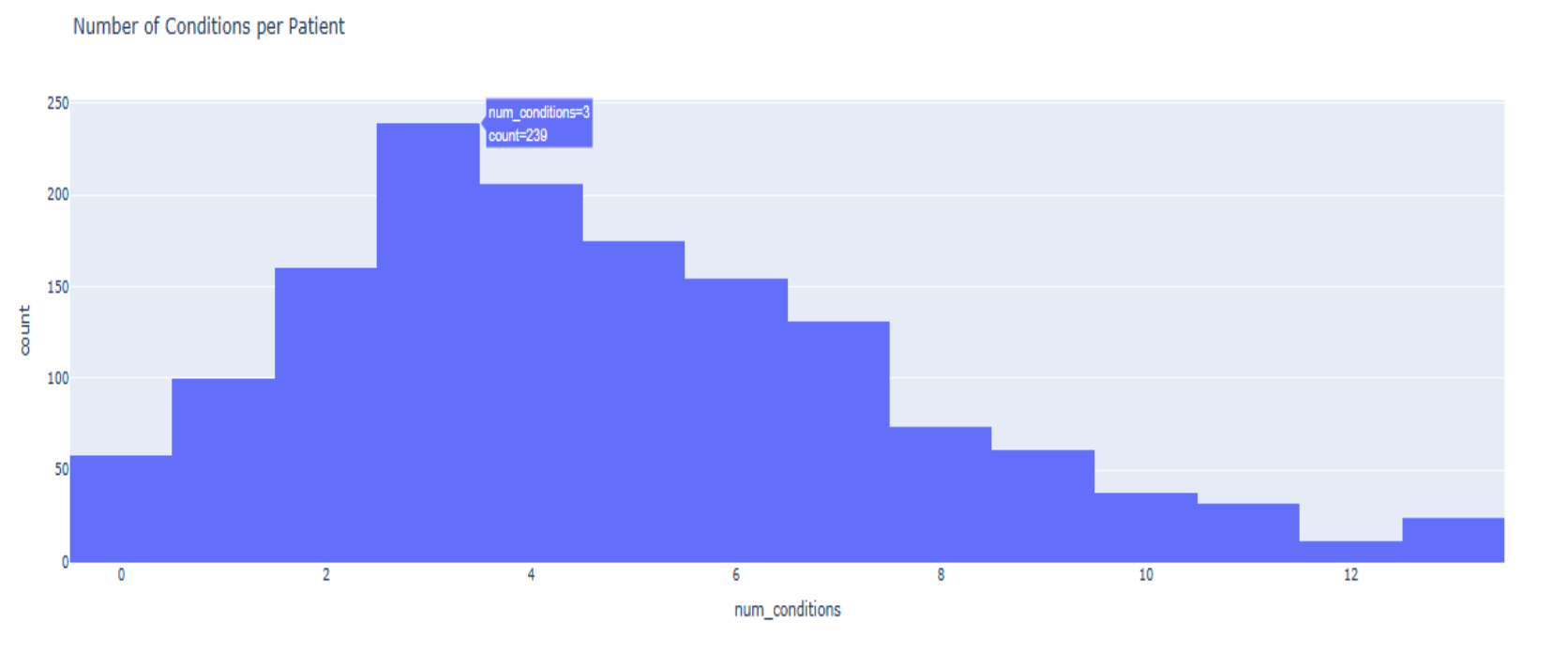
Claims cost distribution



Insight:

Median total claims is 1,200 with an IQR from 400 to 1,800, and capped at 3,900; this highlights moderate variability in patient costs and identifies extreme high-cost cases for targeted analysis.

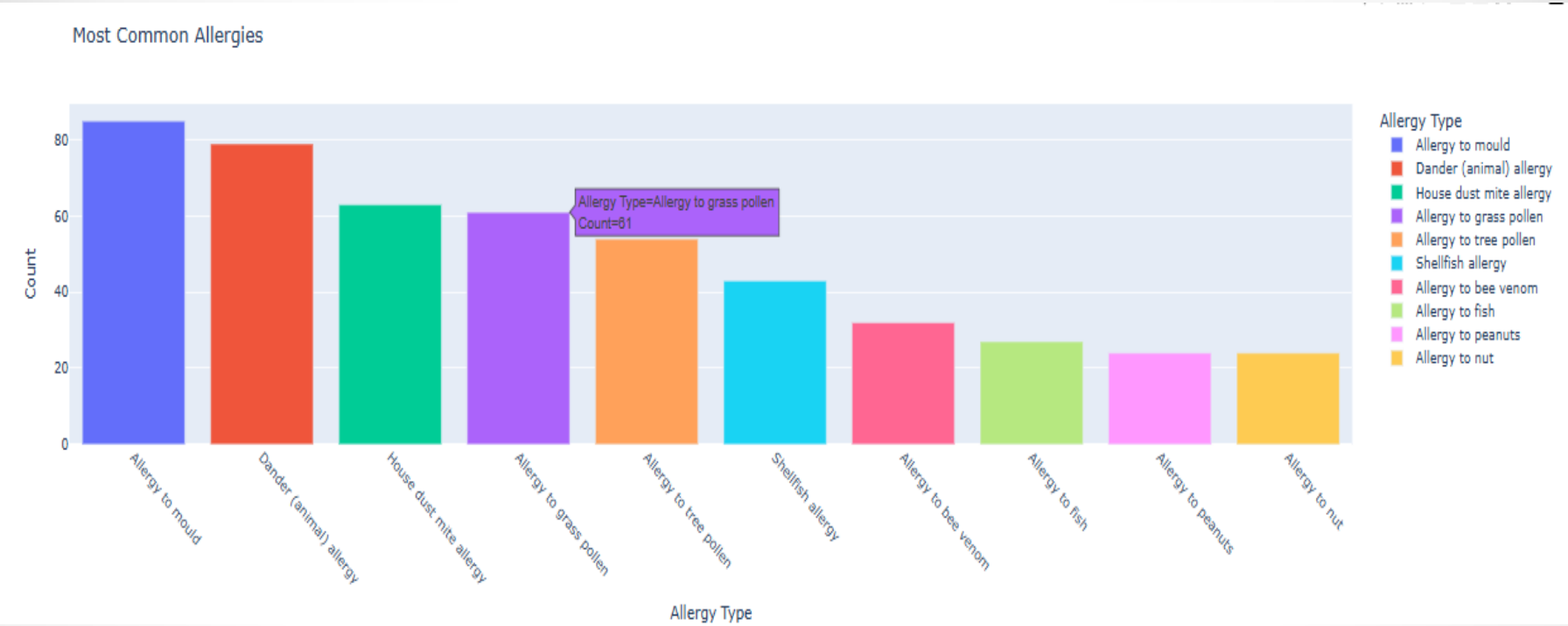
Conditions per patient



Insight:

Most patients have 2–5 conditions, peaking at three conditions, while a smaller subset exhibits high multimorbidity (8+ conditions), highlighting the extreme complexity cases that may require specialized care.

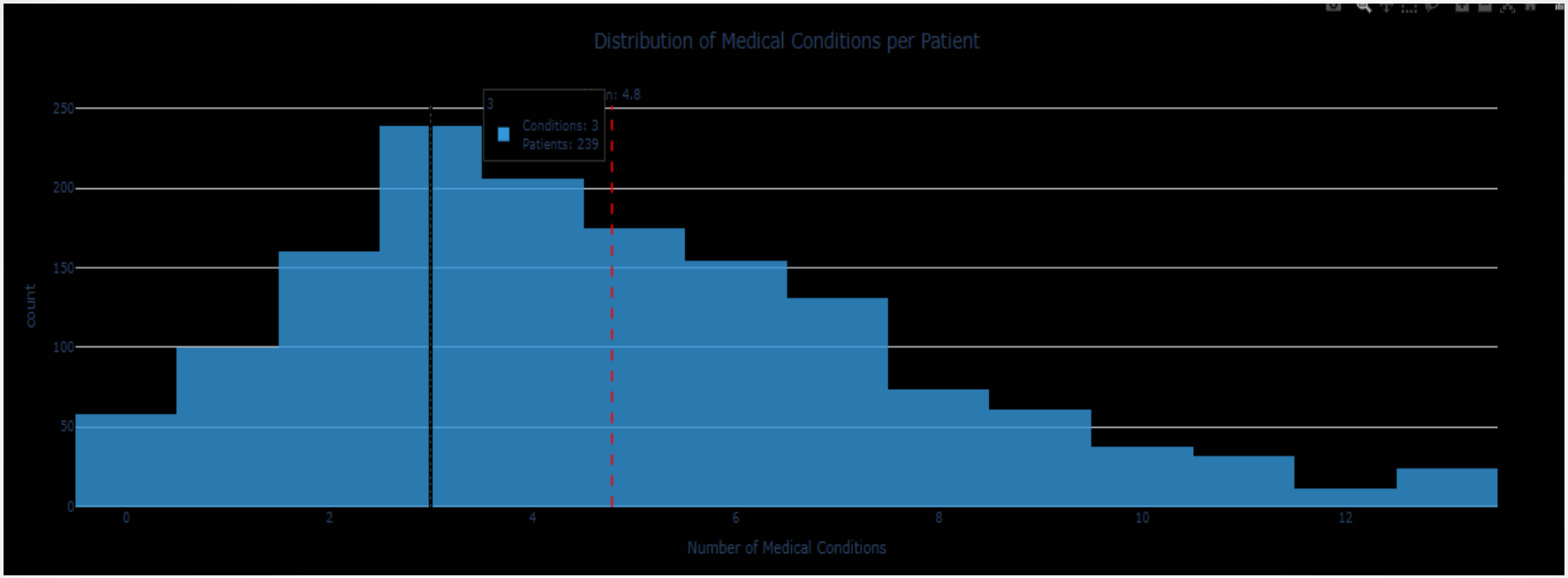
Most Common Allergies



Insight:

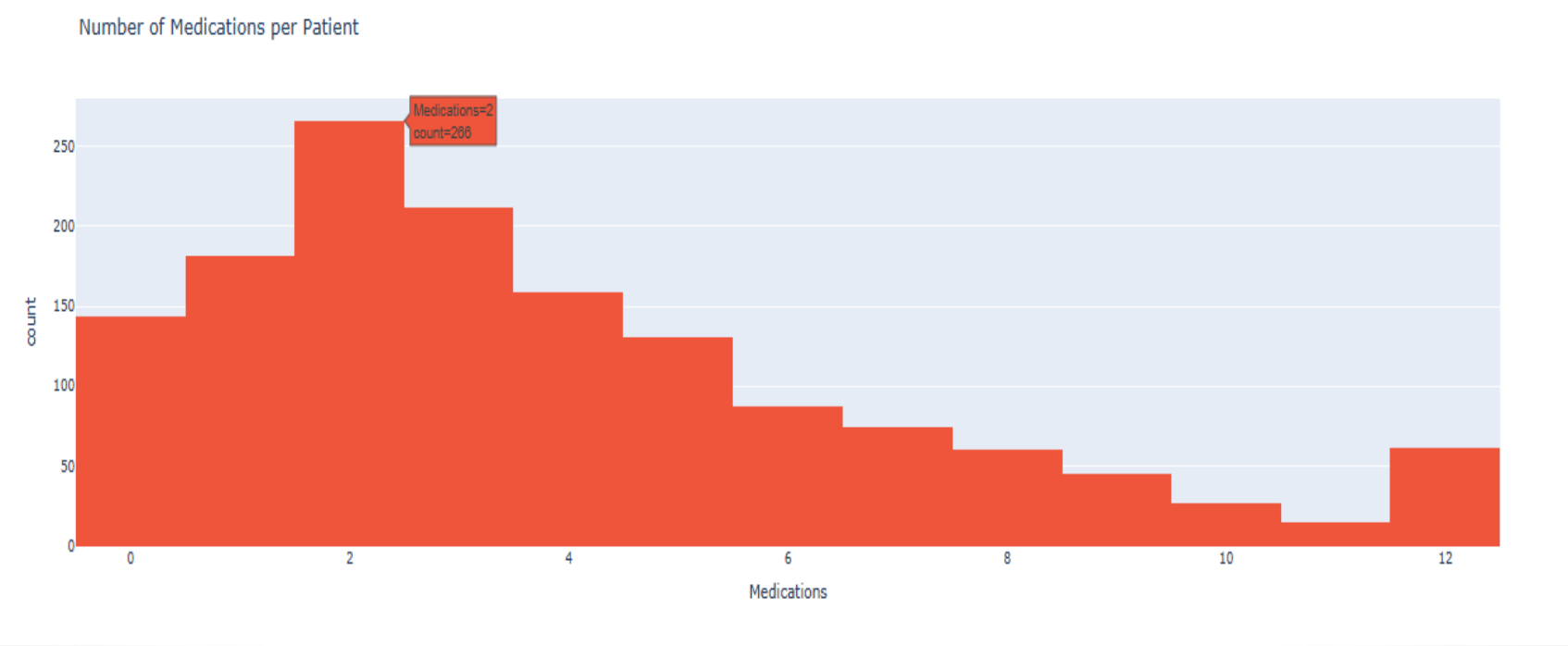
Allergy to mould, animal dander, and house dust mite are the top three allergies, highlighting common environmental triggers in the synthetic population and guiding focus for allergy management strategies.

Distribution of Medical Conditions per Patient



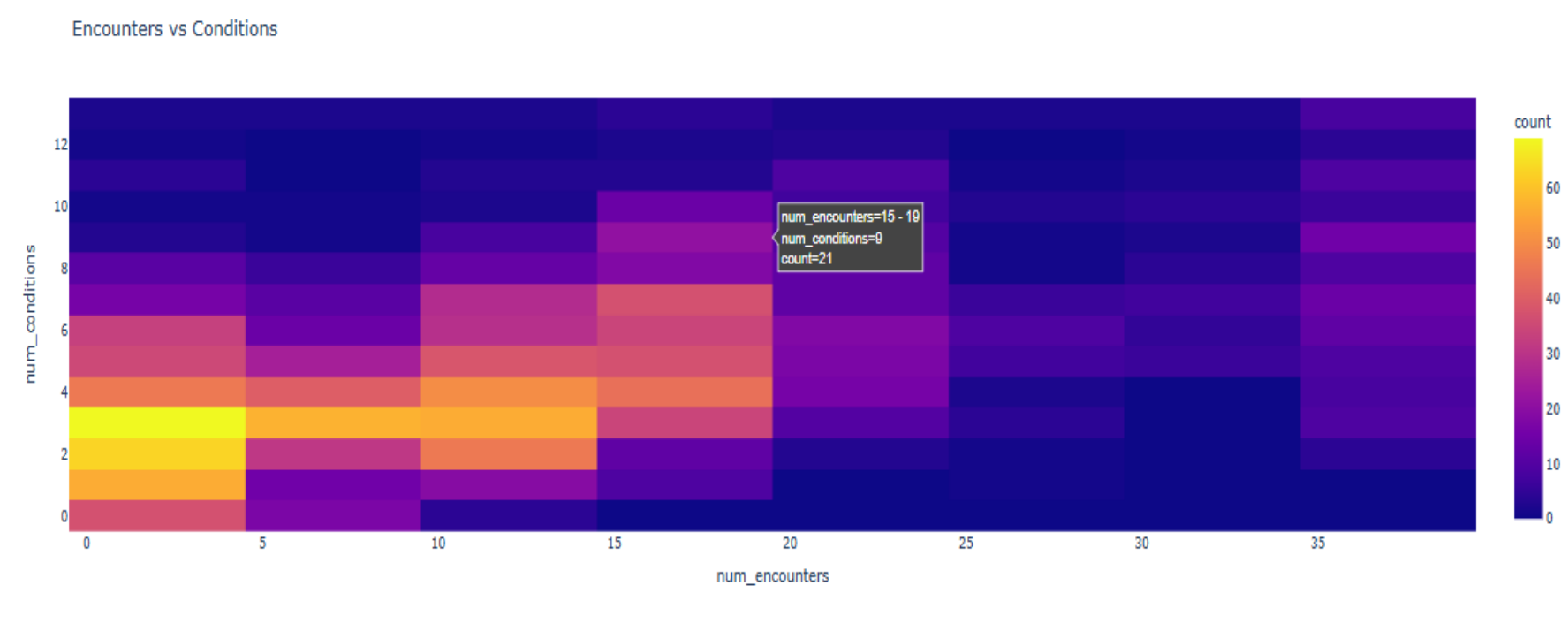
Insight:
Average multimorbidity is 4.8 conditions, with most patients clustered around 3–6 conditions, revealing a moderate disease burden and identifying those above average for targeted intervention.

Number of Medications per Patient



Insight:
Most patients take 1–3 medications, peaking at 2, while a smaller group on 8+ drugs indicates high polypharmacy risk and need for medication review.

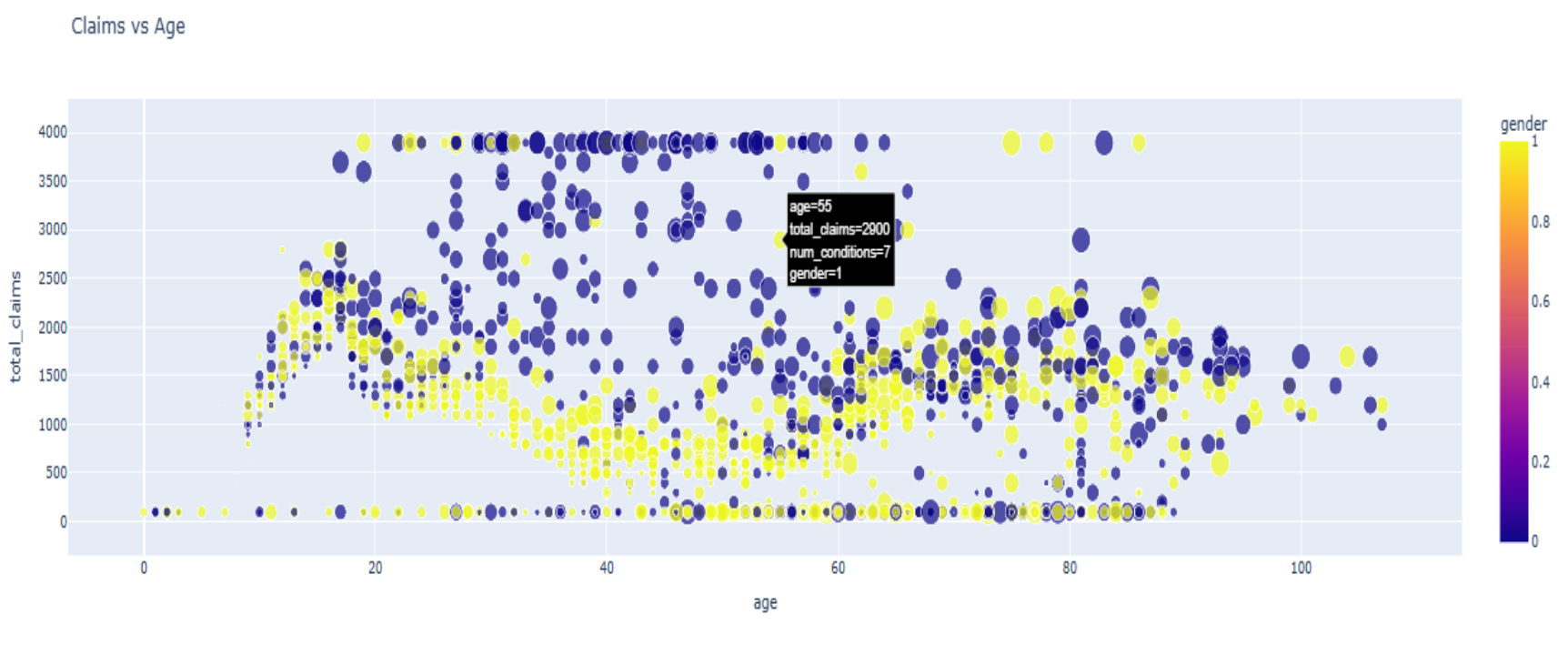
Encounters vs Conditions



Insight:

Patients with 2–5 conditions most commonly have 5–15 encounters, indicating moderate multimorbidity drives average utilization; extremely high encounter counts (>25) align with both low and high condition counts, suggesting frequent visits are not solely driven by multimorbidity but possibly by acute events or complex care needs.

Claims vs Age

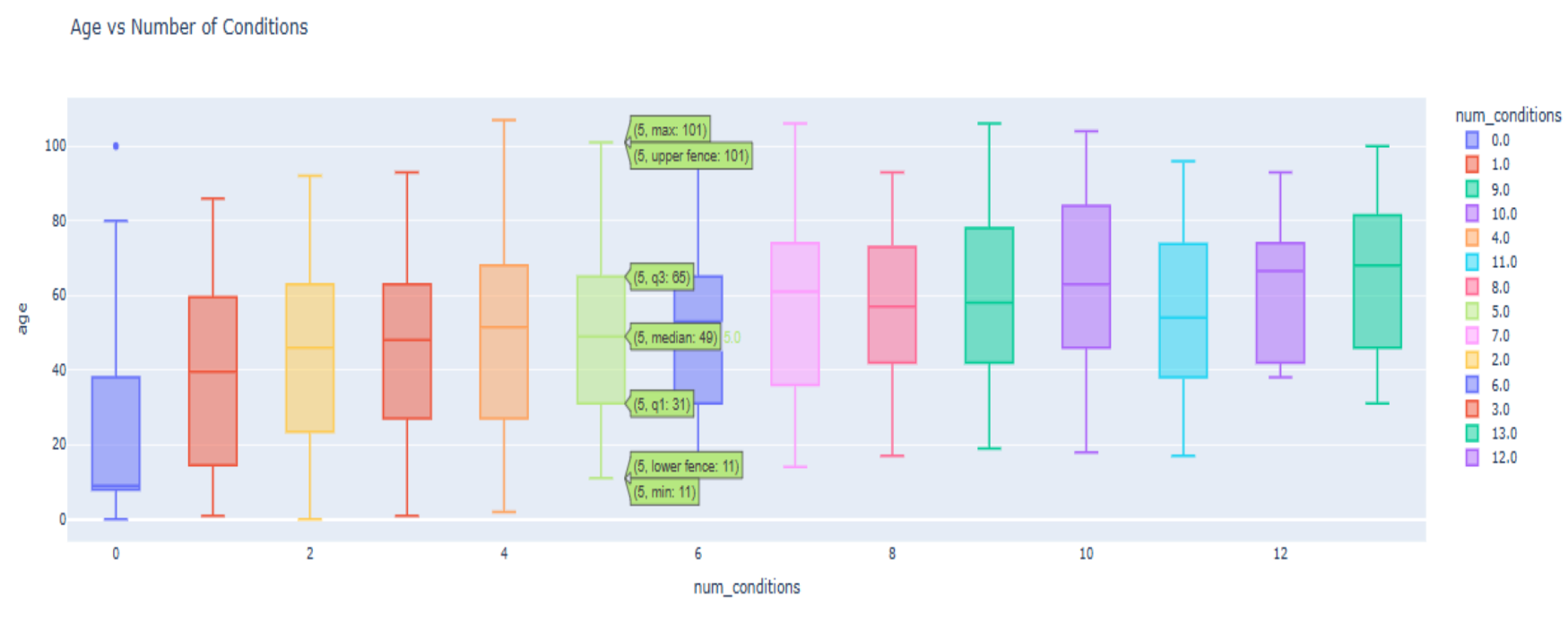


Insight:

Older patients (60+) tend to have higher claim totals, while younger patients cluster at lower costs.

No major gender-specific cost differences appear, confirming age as the primary driver of healthcare spending in this cohort.

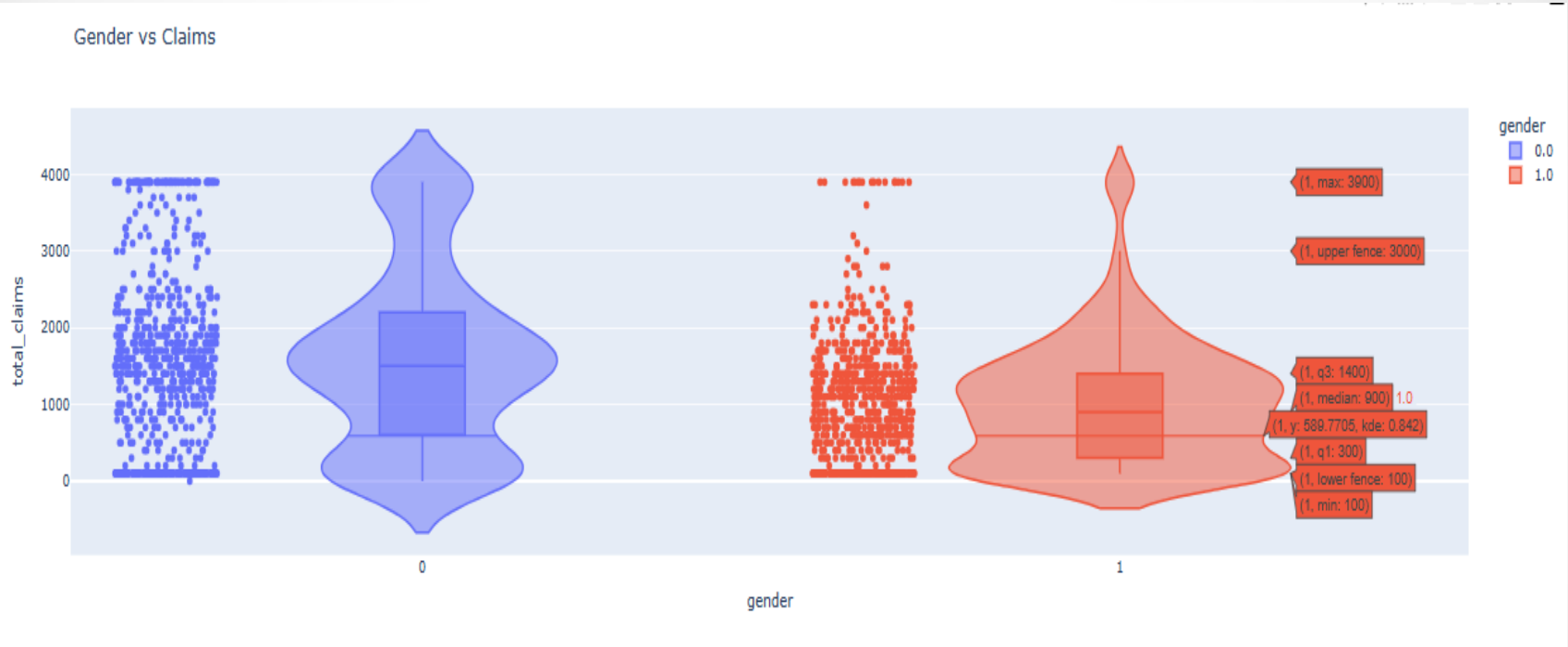
Age vs Conditions



Insight:

As condition count increases, median age rises and spread widens—patients with more conditions are generally older and more variable in age, confirming age as a key driver of multimorbidity.

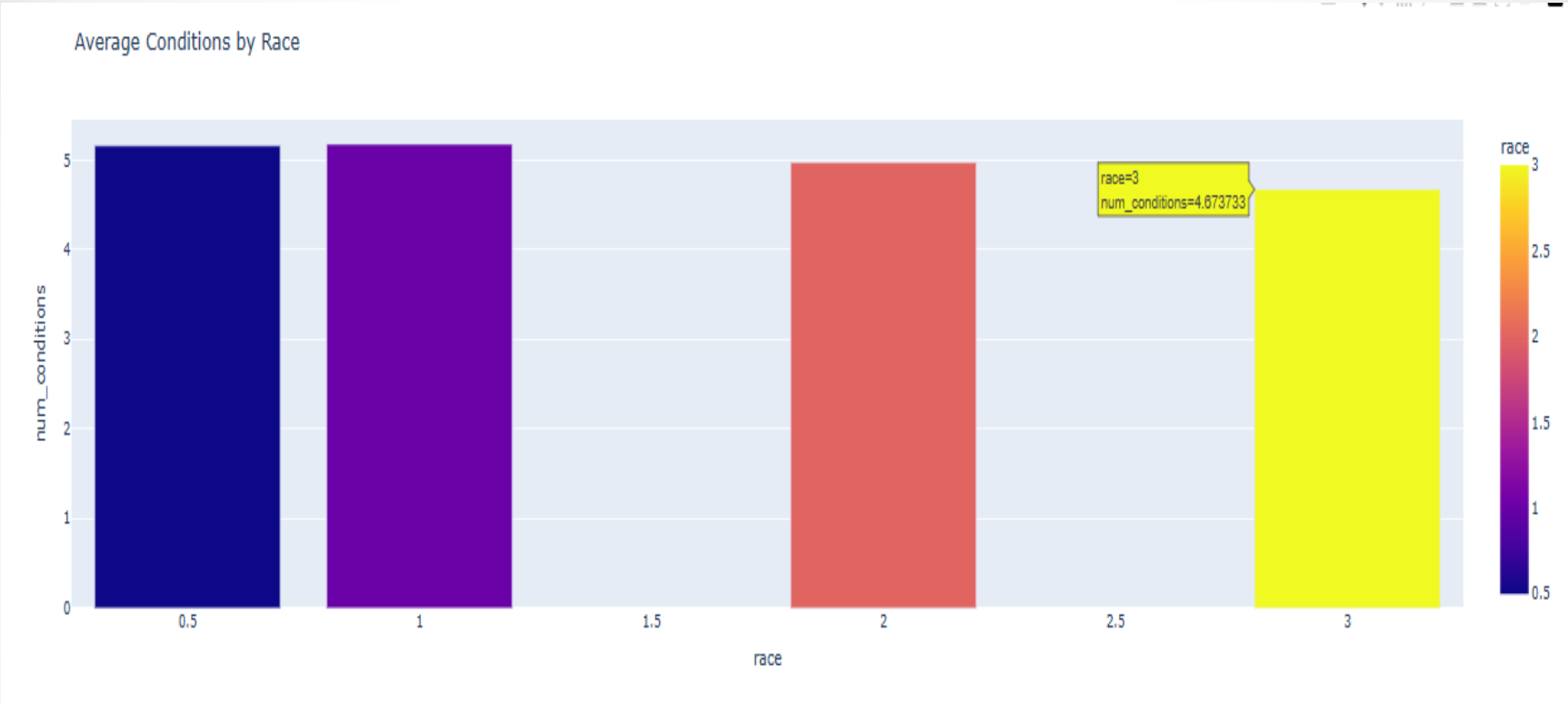
Gender vs Claims



Insight:

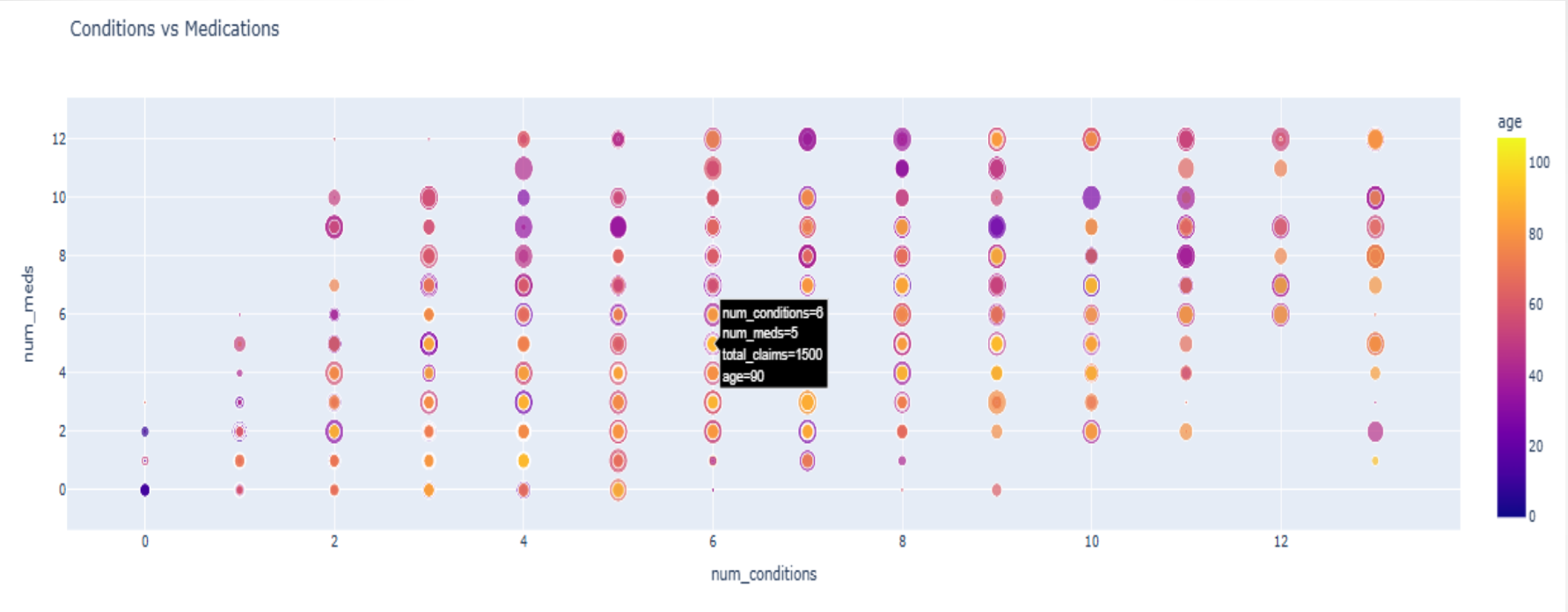
Both genders show similar median claim totals, but males (0) exhibit a wider spread and more extreme high-cost cases, suggesting slightly higher variability in healthcare spending among male patients.

Avg conditions by race



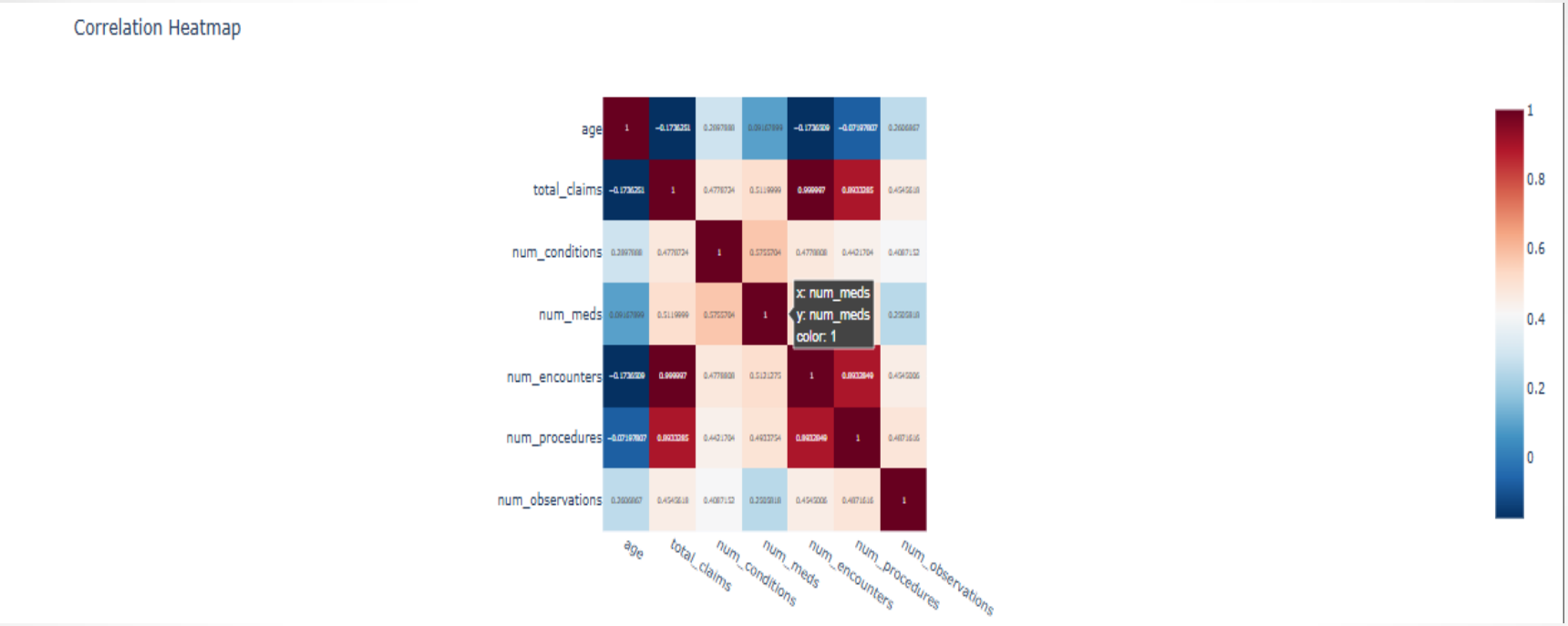
Insight:
Two race categories exhibit the highest mean conditions (5.2), while another shows the lowest (4.7), indicating meaningful variation in multimorbidity by race that may warrant targeted interventions.

Conditions vs Medications



Insight:
Older patients (larger, darker bubbles) tend to have more conditions and medications, confirming that polypharmacy correlates strongly with both age and disease burden.

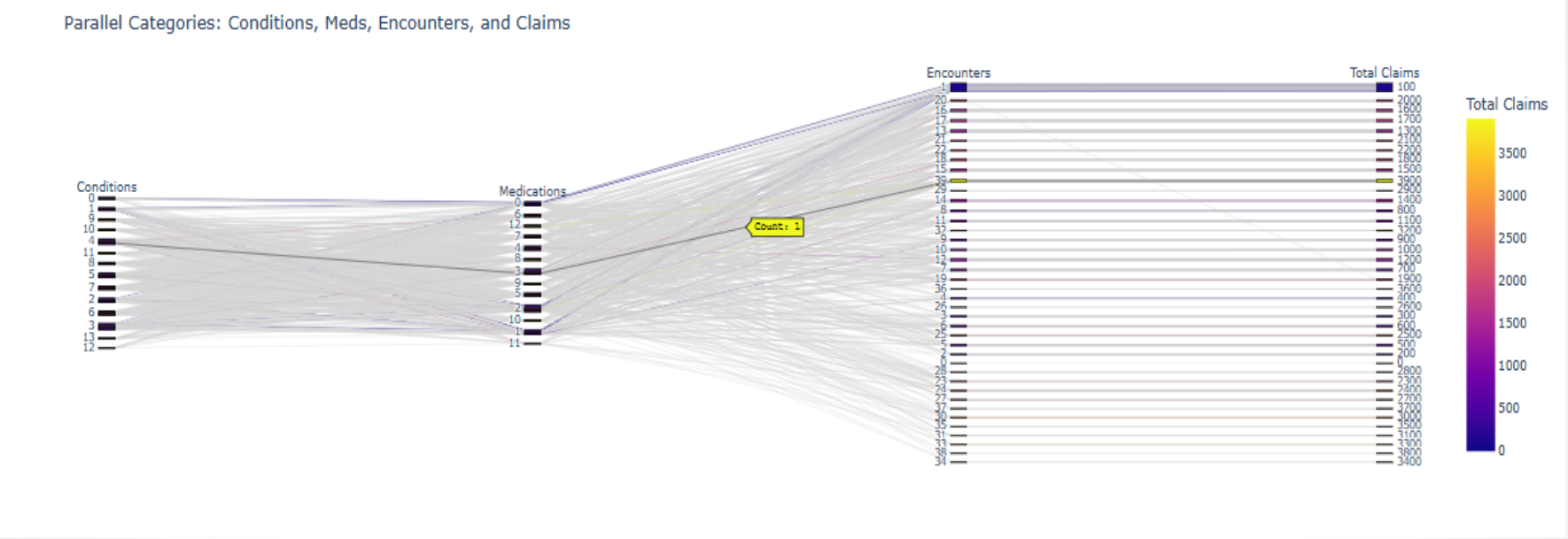
Correlation Heatmap



Insight:

Total encounters and procedures are almost perfectly correlated (0.99), indicating potential redundancy, while age negatively correlates with encounters and claims, suggesting older patients have fewer but costlier encounters.

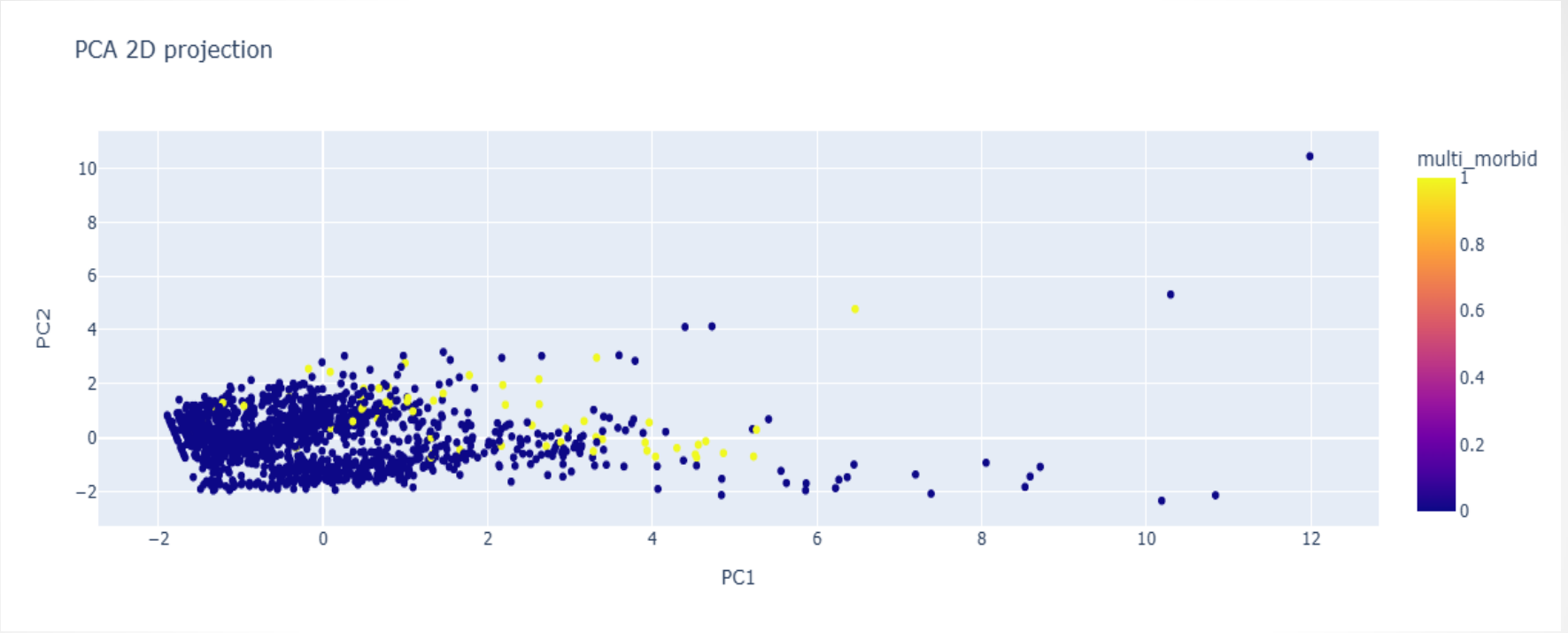
Parallel Categories: Conditions, Meds, Encounters, and Claims



Insight:

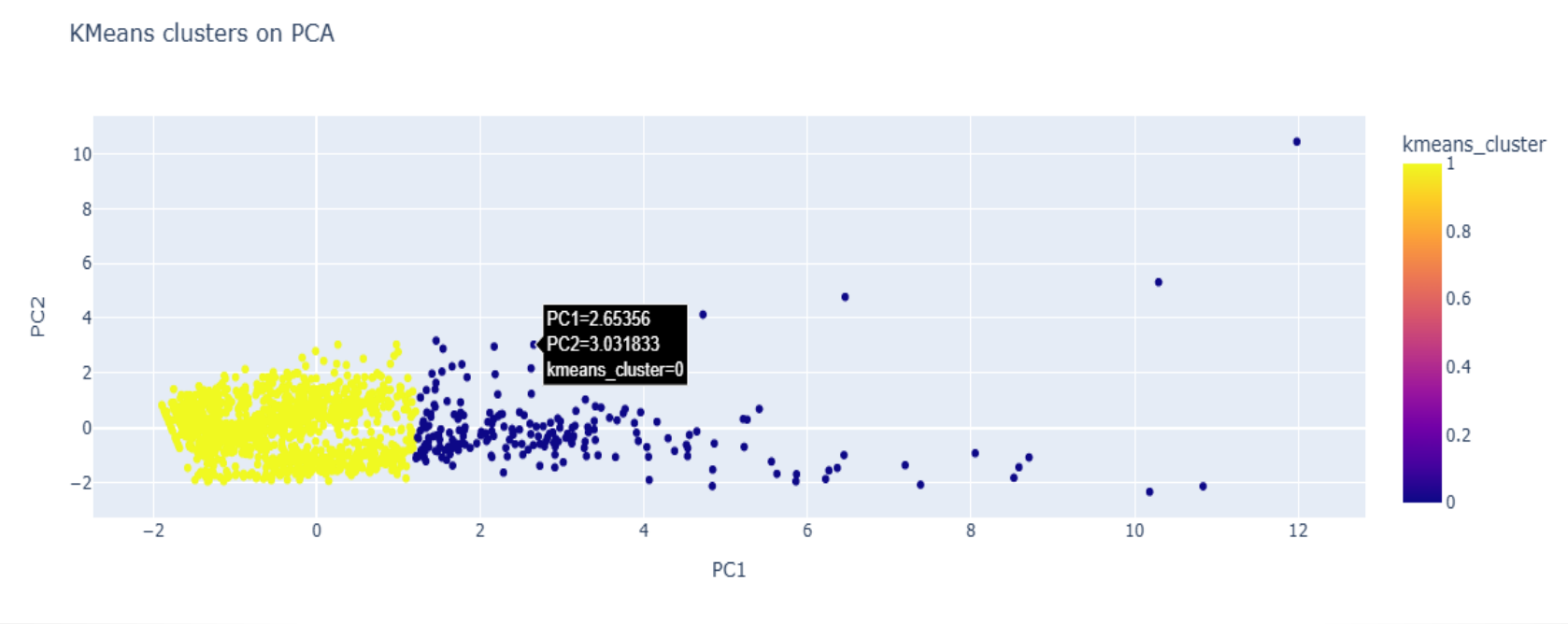
Patterns show that patients with high claims often have many encounters, conditions, and medications simultaneously, highlighting the subgroup most likely to drive healthcare costs and multimorbidity burden.

2D PCA Projection of Patients



Insight:
The scatter plot visualizes patients on the first two PCA components, highlighting patterns that separate multi-morbid from non-multi-morbid patients, with some overlap between groups.

KMeans Clustering Visualization on PCA Projection



Insight:

The scatter plot visualizes KMeans clusters on the first two PCA components, showing distinct patient subgroups with similar feature profiles.