Final Project

Zhan Gu

2023-12-12

Abstract

This project explores a synthetic healthcare dataset to build a multilevel multinomial model that predicts the results of a medical test from a wide selection of patient-related information. Due to the limitations of synthetic data, the model was not statistically significant, but the methods applied in this project could still be used in a similar context.

Introduction

The healthcare dataset used in this project is a synthetic dataset designed to mirror real-world healthcare scenarios. It comprises 10000 observations and key variables related to this project are listed below:

Name: name of the patient associated with the healthcare record.

Age: age of the patient at the time of admission, expressed in years.

Gender: gender of the patient, either "Male" or "Female."

Blood Type: patient's blood type, which can be one of the common blood types (e.g., "A+", "O-", etc.).

Medical Condition: the primary medical condition or diagnosis associated with the patient, such as "Diabetes," "Hypertension," "Asthma," and more.

Hospital: the healthcare facility or hospital where the patient was admitted.

Medication: the medication prescribed or administered to the patient during their admission. Examples include "Aspirin," "Ibuprofen," "Penicillin," "Paracetamol," and "Lipitor."

Test Results: the results of a medical test conducted during the patient's admission. Possible values include "Normal," "Abnormal," or "Inconclusive," indicating the outcome of the test.

Our primary aim is to build a predictive model that predicts test results based on these variables.

Method

First we conducted some data cleaning, particularly converting categorical data to factors and grouping ages. In the subsequent exploratory data analysis, we wanted to look at the distribution of various medical conditions across different demographic groups, including age, gender, and blood type. This is to examine the chance of medical condition confounding the association between test results and demographic groups.

We used stacked bar plots to observe the distribution of conditions among these demographics, which indicated a relatively uniform distribution of medical conditions across all demographic categories. Chi-square tests were conducted to assess the statistical significance of these observations, yielding high p-values over 0.5. This suggests that medical conditions are evenly distributed regardless of age, gender, or blood type in the dataset and medical condition will not confound the association between test results and demographic groups.

Then we build our predictive model and use 7000 observations for training and 3000 observationns for testing. Here we choose our model to be a multilevel multinomial model because the dependent variable "Test Results" is categorical with three levels - Normal, Abnormal, Inconclusive.

Fixed effects include:

Medication: to see how different medications relate to test results.

Medical Condition: to control for the type of medical condition being treated.

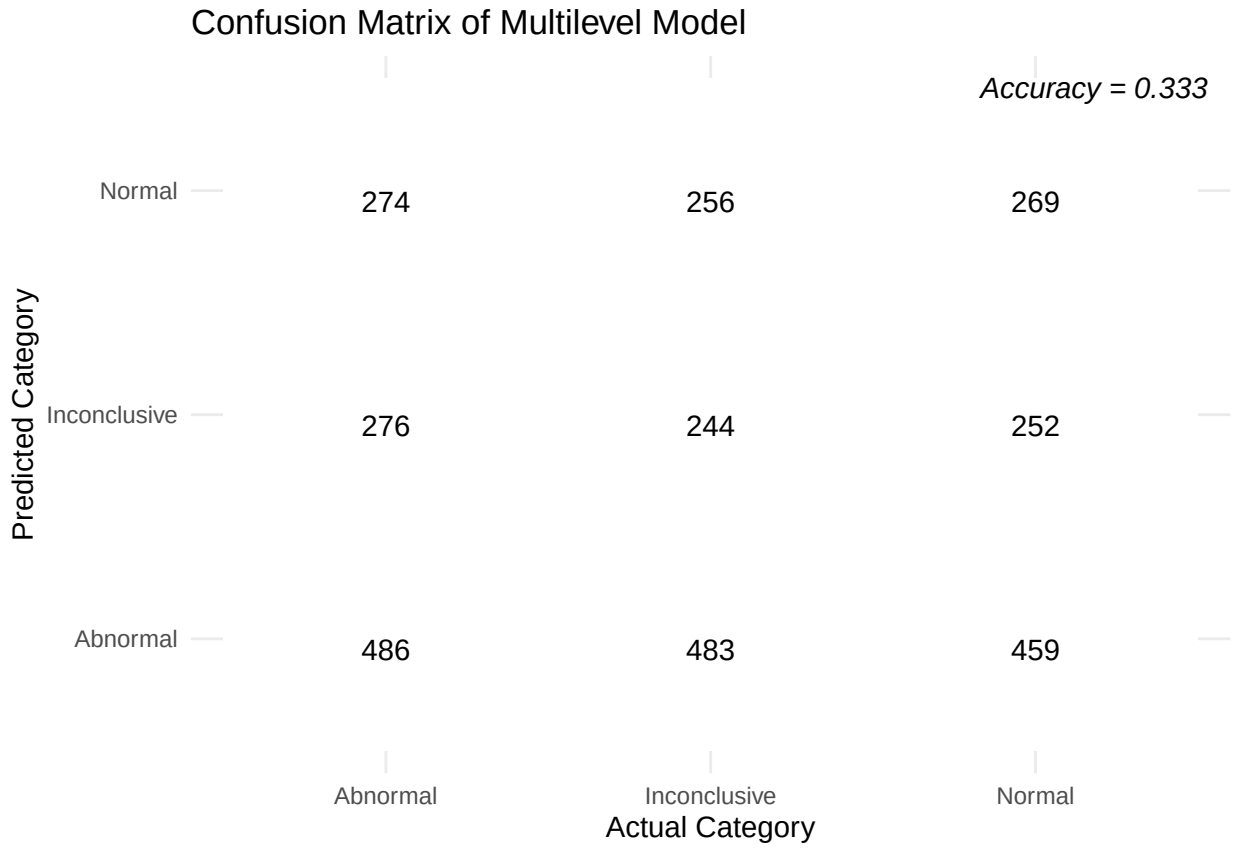
Age Group, Gender, Blood Type: to account for patient demographics.

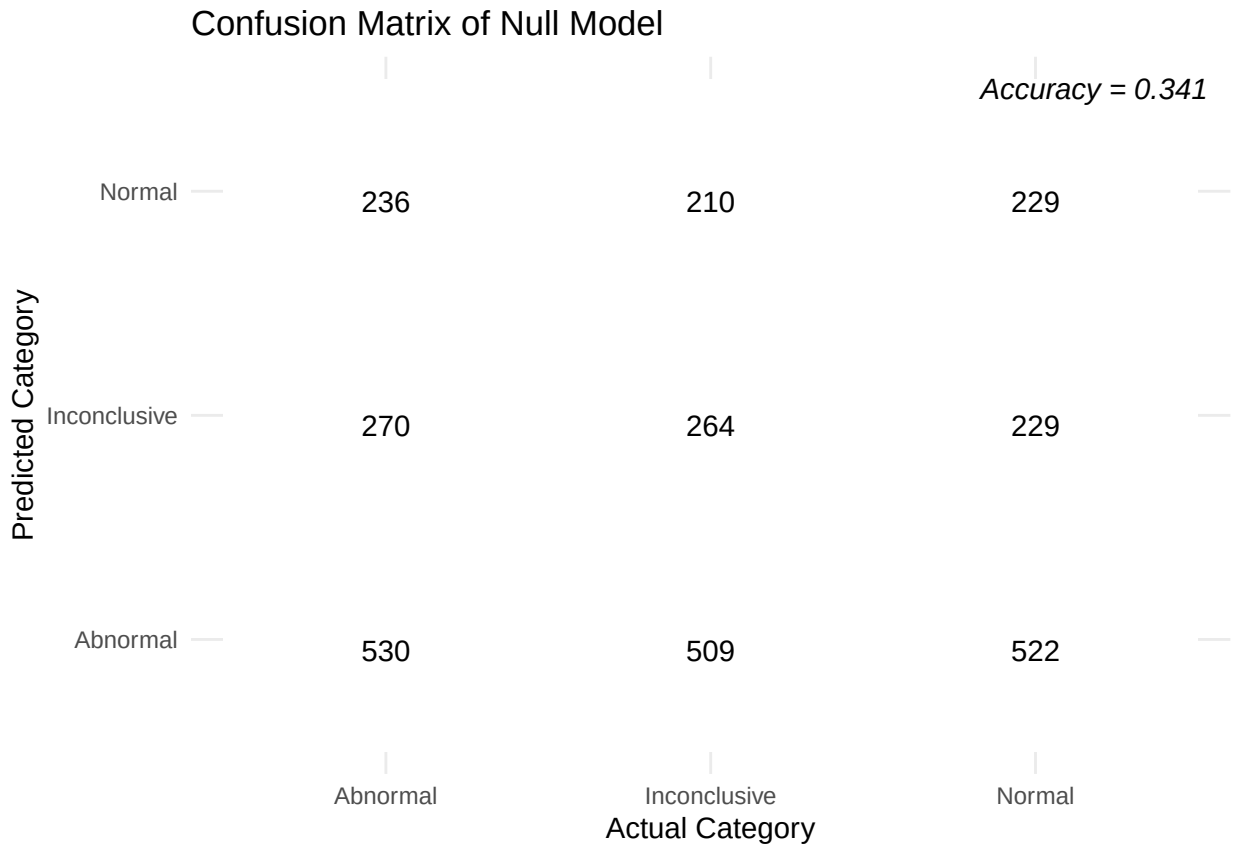
Random effects include:

Hospital: To account for variations between different hospitals, considering that healthcare practices might vary across institutions.

Finally we check the accuracy of the model and compare it with the null model. The null model includes intercept only and predict test results based solely on the overall distribution of the results.

Result


Confusion Matrix of Multilevel Model
Accuracy = 0.333
Normal
Inconclusive
Abnormal
274
256
269
276
244
252
486
483
459
Predicted Category
Abnormal
Inconclusive
Normal
Actual Category

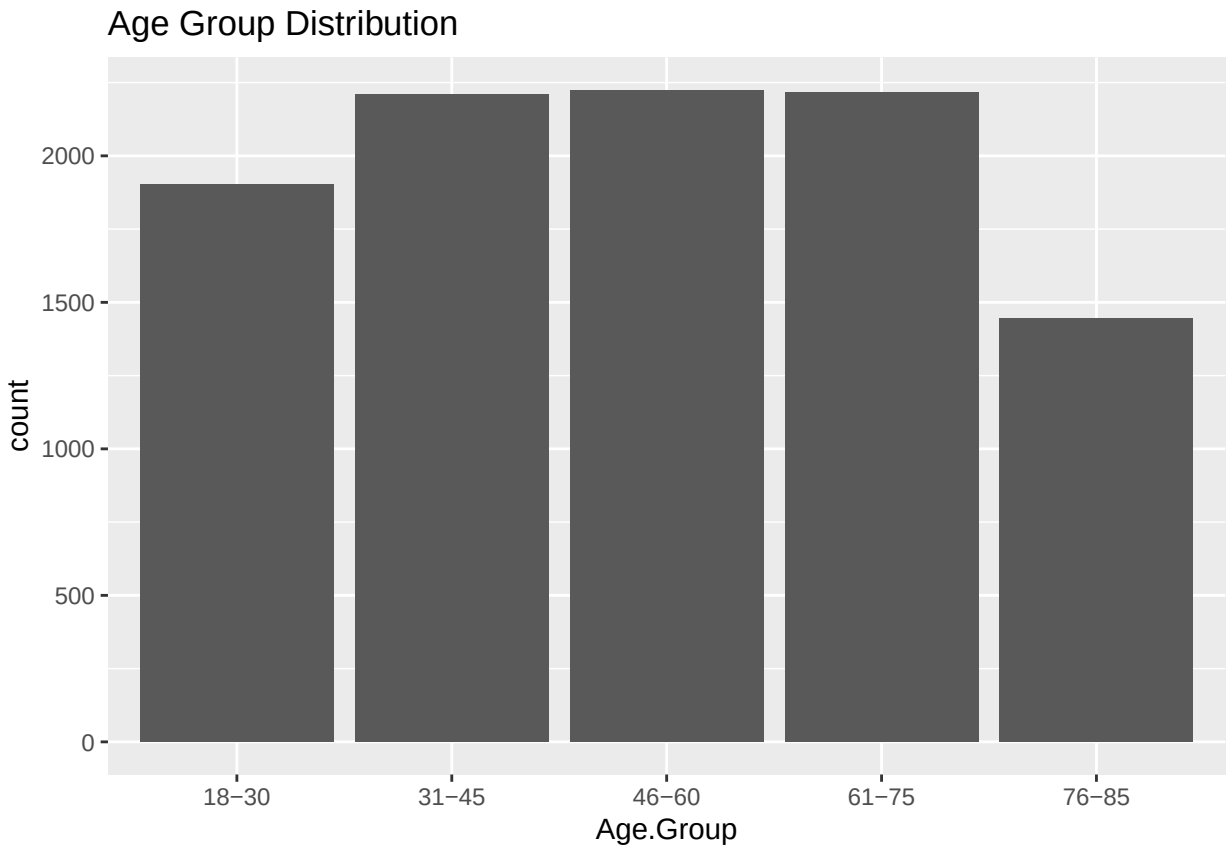

Confusion Matrix of Null Model
Accuracy = 0.341
Normal
Inconclusive
Abnormal
236
210
229
270
264
229
530
509
522
Predicted Category
Abnormal
Inconclusive
Normal
Actual Category

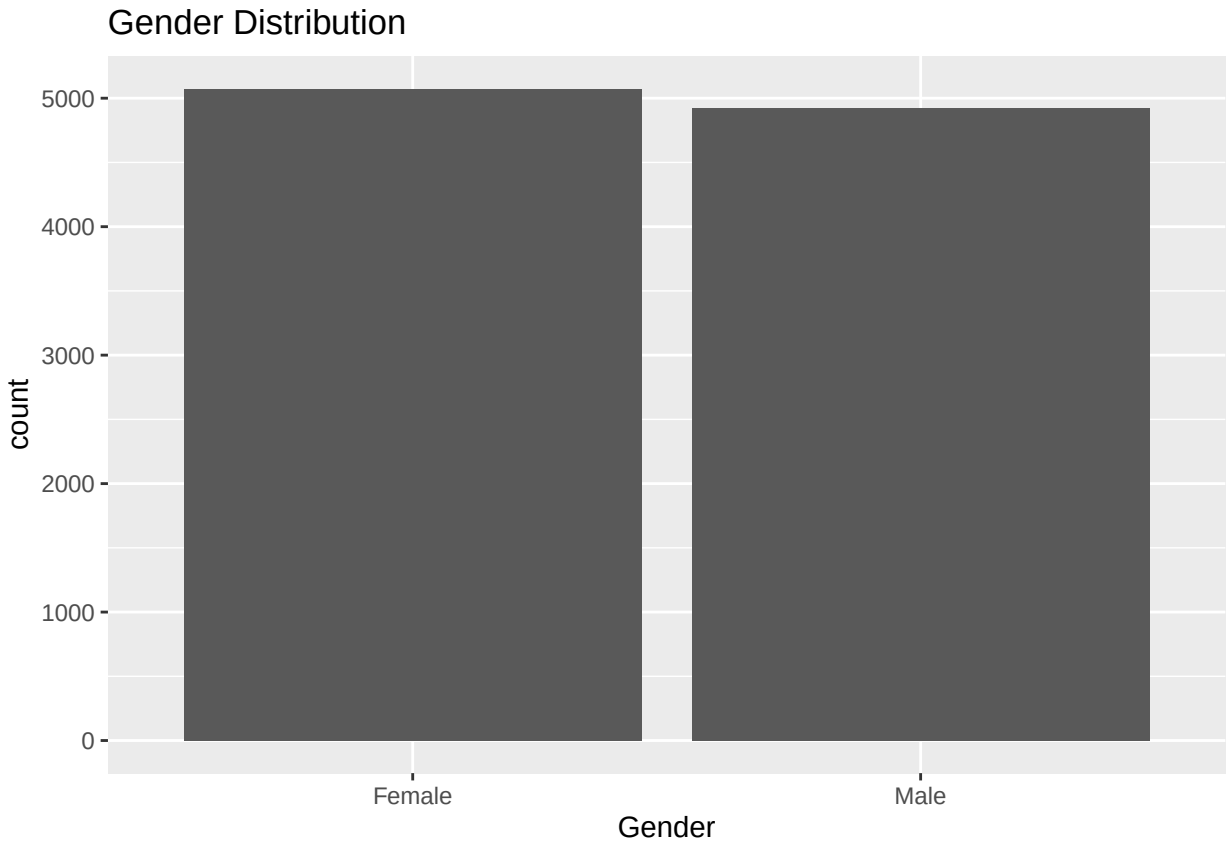
Discussion

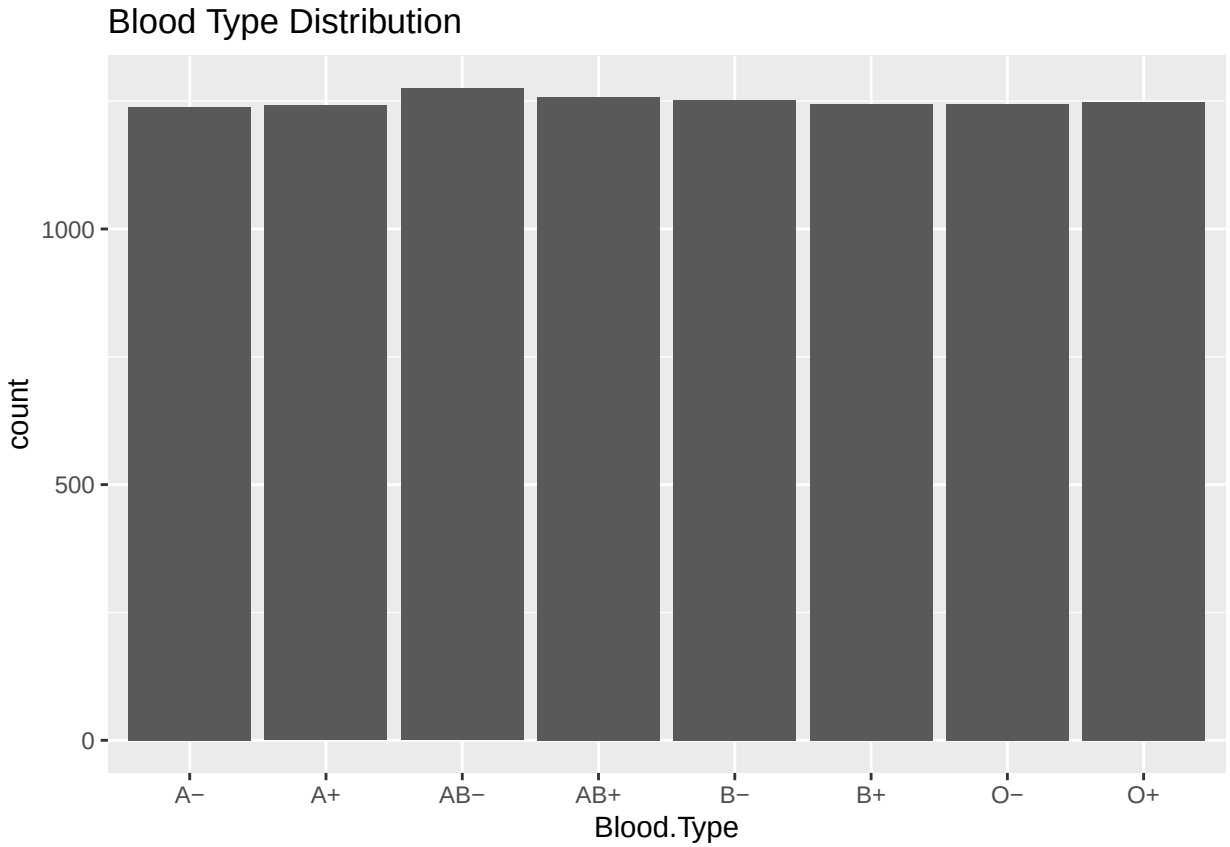
Most coefficients in our predictive model are not statistically significant, and the model accuracy is lower than the null model. This outcome is most likely due to the nature of the dataset. Synthetic data were generated by randomly assigning values to the variables and from our EDA, it seems that the values are uniformly distributed. As a result, there is no actual relationship between the variables, unlike real data where correlations exist. Here the outcome is only a dummy variable and thus the predictors in our model are simply irrelevant.

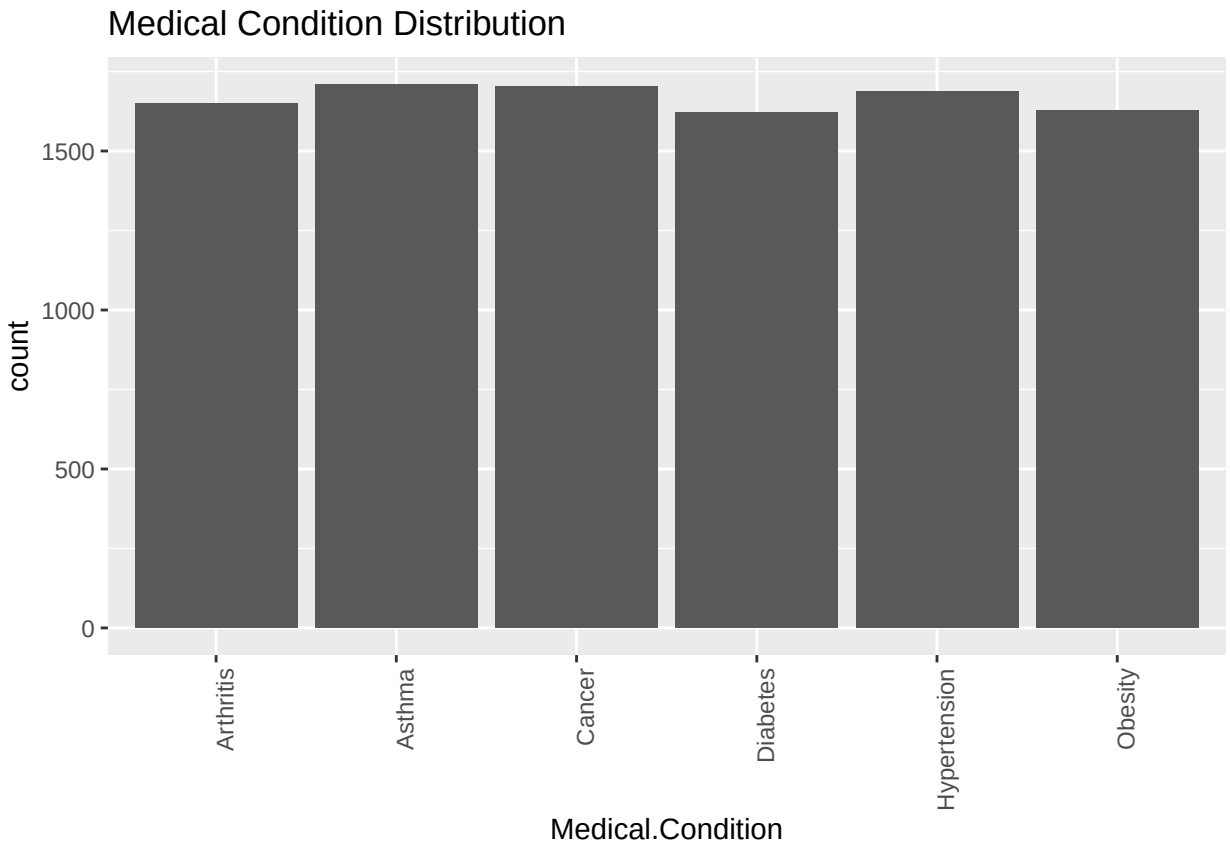
Due to the limitations of time and access to real data, this project can only explore multilevel models in healthcare data analysis to this extent. Despite the lack of statistically significant outcomes, the applied methods could still be of use in contexts where more comprehensive, real-world data are available.

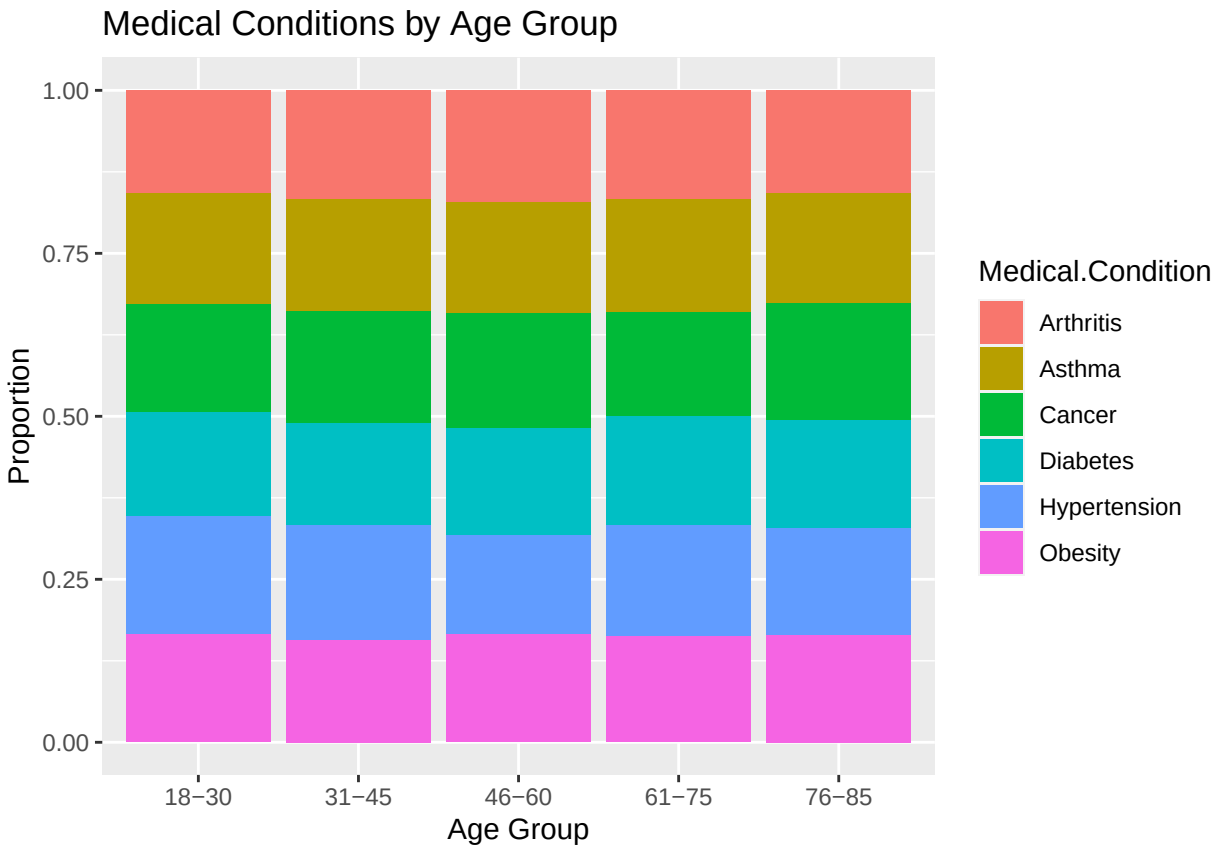
Appendix

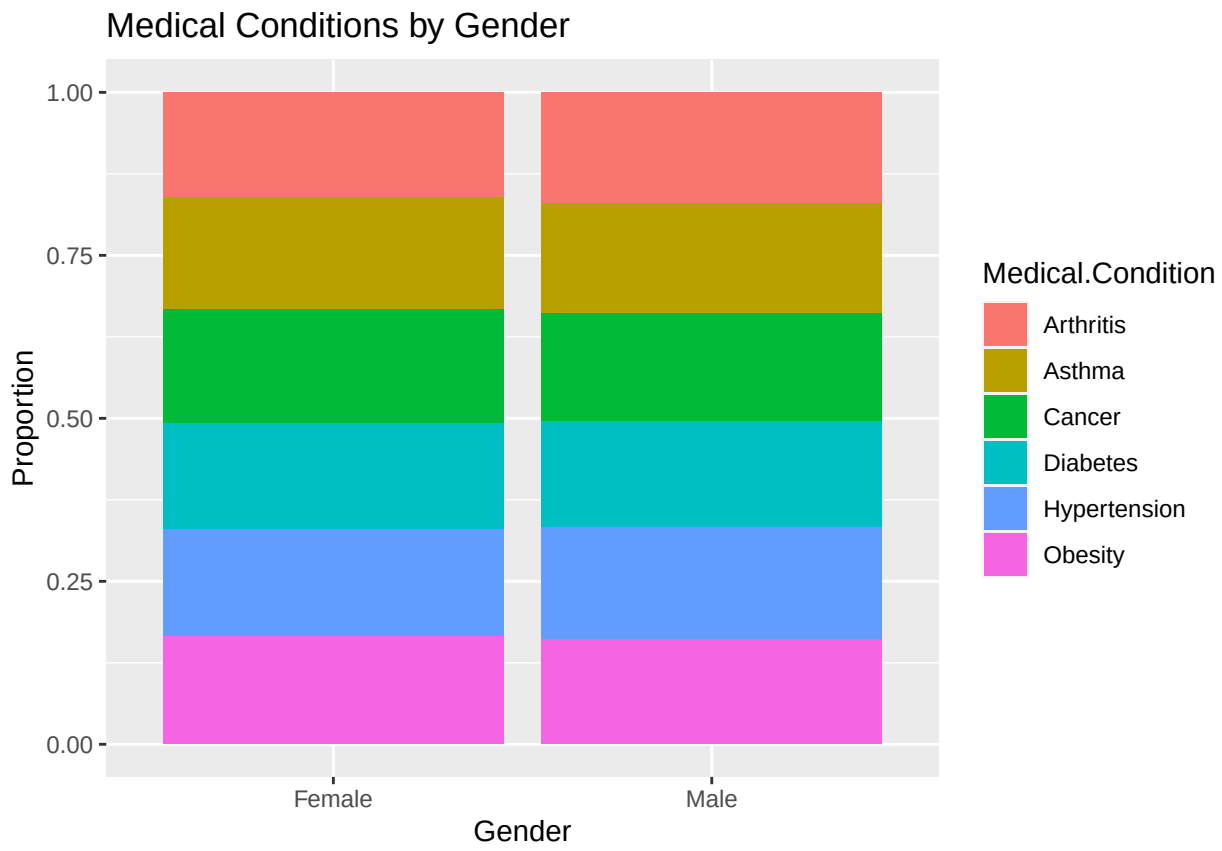

Age Group Distribution
2000
1500
1000
500
0
count
18−30
31−45
46−60
61−75
76−85
Age.Group

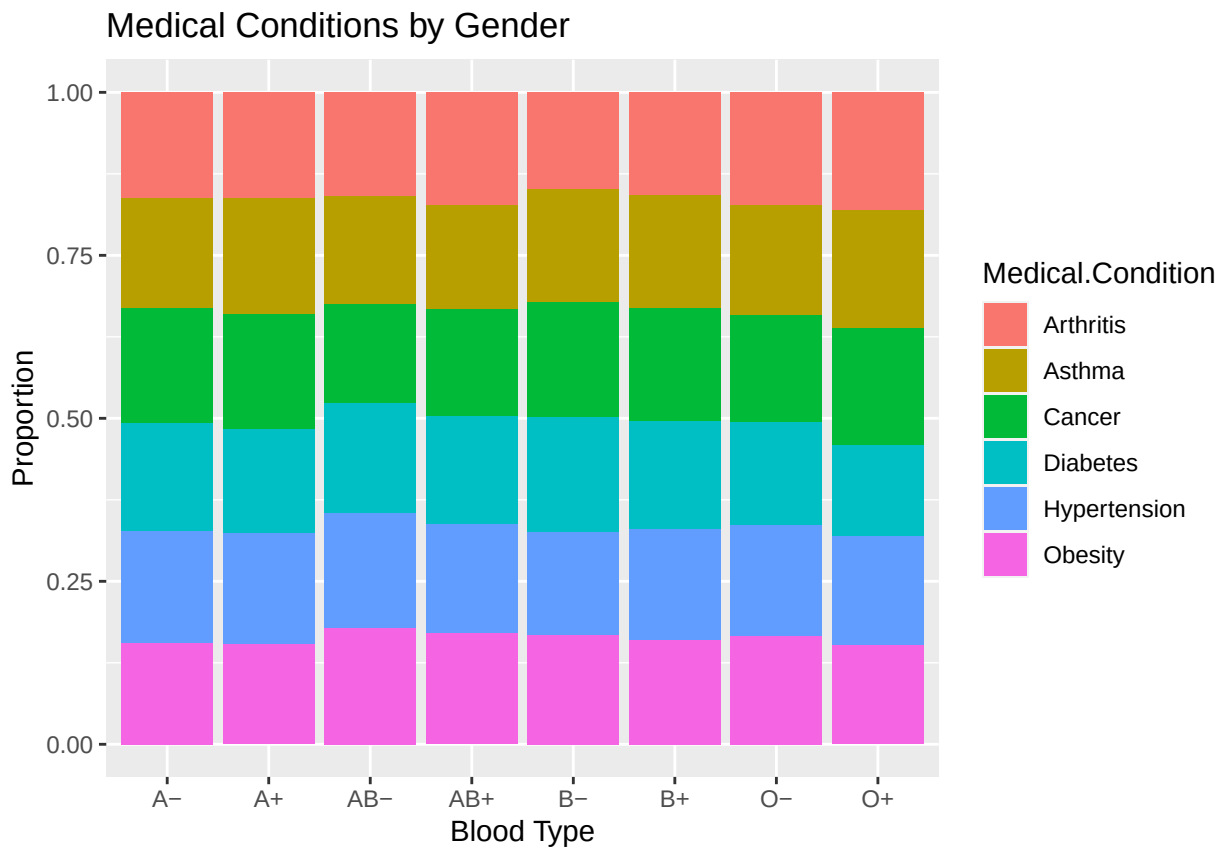

Gender Distribution
5000
4000
3000
2000
1000
0
count
Female
Male
Gender


Blood Type Distribution
count
1000
500
0
A−
A+
AB−
AB+
B−
B+
O−
O+
Blood.Type


Medical Condition Distribution
count
1500
1000
500
0
Arthritis
Asthma
Cancer
Diabetes
Hypertension
Obesity
Medical.Condition


Medical Conditions by Age Group
1.00
0.75
0.50
0.25
0.00
Proportion
18−30
31−45
46−60
61−75
76−85
Age Group
Medical.Condition
Arthritis
Asthma
Cancer
Diabetes
Hypertension
Obesity


Medical Conditions by Gender
1.00
0.75
0.50
0.25
0.00
Proportion
Female
Male
Gender
Medical.Condition
Arthritis
Asthma
Cancer
Diabetes
Hypertension
Obesity


Medical Conditions by Gender
Proportion
1.00
0.75
0.50
0.25
0.00
A−
A+
AB−
AB+
B−
B+
O−
O+
Blood Type
Medical.Condition
Arthritis
Asthma
Cancer
Diabetes
Hypertension
Obesity