COMM 335 Data Viz Memo

Section 201, Group 13

95970, 37531, 74542, 64813, 95972

Memo: What set of factors leads to high serum cholesterol levels within male patients in Cleveland?

The data used are datasets for patients undergoing angiography tests within the Cleveland Clinic in Cleveland, Ohio, the Hungarian Institute of Cardiology in Budapest, Hungary, and the Veterans Administration Medical Center in Long Beach, California. (Detrano, Janosi, Steinbrunn, Pfisterer, Schmid, Sandhu, Guppy, Lee, Froelicher, 1989) These datasets were retrieved from the Heart Disease dataset from the UCI machine learning repository and converted from .data files to Excel files, then CSV files.

Within R, each CSV file is read separately, then merged into a single dataframe, with a “location” column added for each clinic’s dataset. Specific columns are missing the majority of their data and thus have been removed under the assumption that they are irrelevant. Any patients with “?” for any variables, trestbps = 0 or chol = 0, are assumed to be invalid and have been removed. Any values of num ≥ 1 provide the same result and thus have been converted to 1. The updated data frame is converted to an Excel spreadsheet with their respective definitions to be uploaded into Power BI.

The following variables were provided in both the Excel data frame and the dashboard:

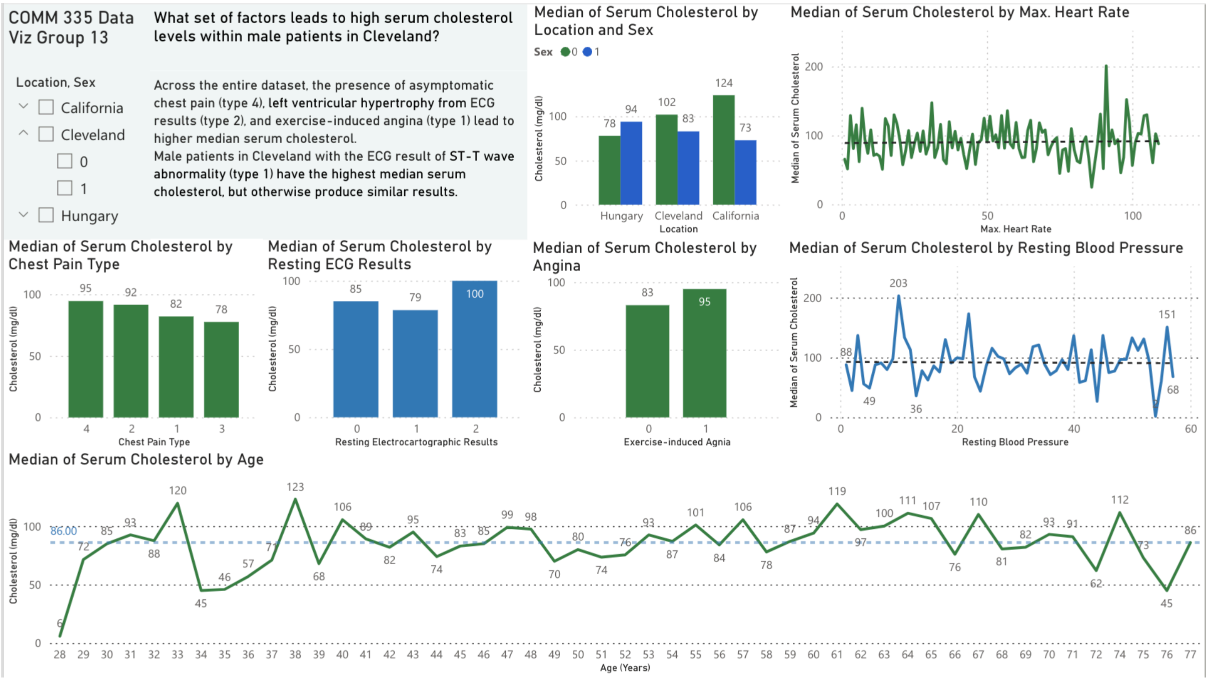
Variable	Definition	Unit	Categories
age	Age	Years	N/A
sex	Sex	N/A	0: Female; 1: Male
cp	Chest pain type	N/A	1: Typical angina; 2: Atypical angina; 3: Non-anginal pain; 4: Asymptomatic
trestbps	Resting blood pressure on admission to hospital	mmHg	N/A
chol	Serum cholesterol	mg/dl	N/A
restecg	Resting electrocardiographic results	N/A	0: Normal; 1: Having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); 2: Showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	Maximum heart rate achieved	BPM	N/A
exang	Exercise-induced angina	N/A	0: No; 1: Yes
num	Diagnosis of heart disease	N/A	0: < 50% diameter narrowing; 1+: > 50% diameter narrowing

To answer our question, we selected the median serum cholesterol level as our test statistic. The proportions for each variable category may differ, and prior exploratory data analysis indicates that serum cholesterol levels contain many outliers that are not representative of the data. As the diagnosis of heart disease was a result produced by a predictive model based on the Cleveland dataset, it has been ignored.

Across all patients, the presence of asymptomatic chest pain left ventricular hypertrophy from ECG results, as well as the presence of exercise-induced angina, leads to higher median serum cholesterol levels. Both resting blood pressure and maximum heart rate produce a trend line with median serum cholesterol level with positive gradients close to 0. Male patients in Cleveland with the ECG result of ST-T wave abnormality have the highest median serum cholesterol, as well as a stronger correlation between resting blood pressure and serum cholesterol, but otherwise produce similar results.

In conclusion, the set of factors that relate to high serum cholesterol levels within male patients in Cleveland are ST-T wave abnormality, exercise-induced angina, high resting blood pressure and high maximum heart rate.

Dashboard



COMM 335 Data Viz Group 13

What set of factors leads to high serum cholesterol levels within male patients in Cleveland?

Location, Sex

- ☐ California
- ☐ Cleveland
  - ☐ 0
  - ☐ 1
- ☐ Hungary

Across the entire dataset, the presence of asymptomatic chest pain (type 4), left ventricular hypertrophy from ECG results (type 2), and exercise-induced angina (type 1) lead to higher median serum cholesterol.
Male patients in Cleveland with the ECG result of ST-T wave abnormality (type 1) have the highest median serum cholesterol, but otherwise produce similar results.

Sources

- Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X
- Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American journal of cardiology, 64(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9 (Detrano, Janosi, Steinbrunn, Pfisterer, Schmid, Sandhu, Guppy, Lee, Froelicher, 1989)