Week 1 Project Paper

C. Morgenstern, *Data Science student at Lewis University*

*Abstract*—This project paper was written for the assignment of week 1 in the course 70-530 Data Visualization at Lewis University in fall term 2, 2019. It is a report that summarizes results from the analysis of a chosen data set using the Tableau Desktop 2019.3 software. The data set used for explanatory data analysis is the Cleaveland Heart disease Data Set that was used by Machine Learning researchers attempting to distinguish the presence or absence of a heart disease in the patient based on certain attributes. This report summarizes an initial explanatory data analysis based on parameters like sex, age, indicators for chest pain types and raised blood sugar levels.

I. Description of Data Set

A

s a former biomedical scientist, my interests are especially in analyzing biological, biomedical or health care data sets. Thus, I was looking for an interesting life science data set on different platforms and came across a medical data set that contains information concerning heart disease diagnosis in patients on kaggle [1]. As it turned out, this data set has also been published on the UCI Machine Learning Repository [2] as well as on Harvard´s dataverse [3]. The latter source provided a cleaned and restructured version of the original data set without missing entries and clear features. The data set analyzed in this report was downloaded from the Harvard dataverse database [4]. Originally the database contained 76 attributes, however only 14 attributes have been referred to in the published experiments [5-7]. See Table 1 for the attributes outlined by Bartley [4], their description and associated values as well as data types. Further credit should be given to the Cleveland Clinic Foundation who curated the database with the following institutions and principal investigators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.  
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.  
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.  
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.

TABLE I

**Patient attributes in heart disease dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Row number** | **Attribute** | **Value and description** | **Data type** |
| 0 | Target | 0 = disease, 1 = no disease | Category |
| 1 | Age | in years | Continuous |
| 2 | Sex | 1=male, 0=female | Category |
| 3 | Indicator for typical angina | 1=true, 0=false | Category |
| 4 | Indicator for atypical angina | 1=true, 0=false | Category |
| 5 | Indicator for non-angina pain | 1=true, 0=false | Category |
| 6 | Resting blood pressure | in mmHg on admission to the hospital | Continuous |
| 7 | Serum cholesterol | in mg/ml | Continuous |
| 8 | Fasting blood sugar > 120 mg/dl | 1=true, 0=false | Category |
| 9 | Indicator for electrocardio value 1 | 1=true, 0=false | Category |
| 10 | Indicator for electrocardio value 2 | 1=true, 0=false | Category |
| 11 | Maximum heart rate achieved | in beats per minute | Continuous |
| 12 | Exercise induced angina | 1=true, 0=false | Category |
| 13 | Oldpeak = ST depression induced by exercise relative to rest | Electrocardiographic parameter | Continuous |
| 14 | Indicator for slope of the peak exercise up | 1=true, 0=false | Category |
| 15 | Indicator for slope of the peak exercise down | 1=true, 0=false | Category |
| 16 | Number of major vessels colored by fluoroscopy | 0,1,2,3 | Category |
| 17 | Thal (Thalassemia) reversible defect indicator | 1=true, 0=false | Category |
| 18 | Thal (Thalassemia) fixed defect indicator | 1=true, 0=false | Category |

II. Data Analysis Process

The data set was imported into Tableau Desktop 2019.3 using the Connect pane and the “To a file” – “Text file” selection. The first processing step required the splitting of the values into individual columns and renaming the headings according to Table 1. Individual graphics were designed on the sheets pages taken into account the learning of week 1 and then exported as .png for incorporation into this paper.

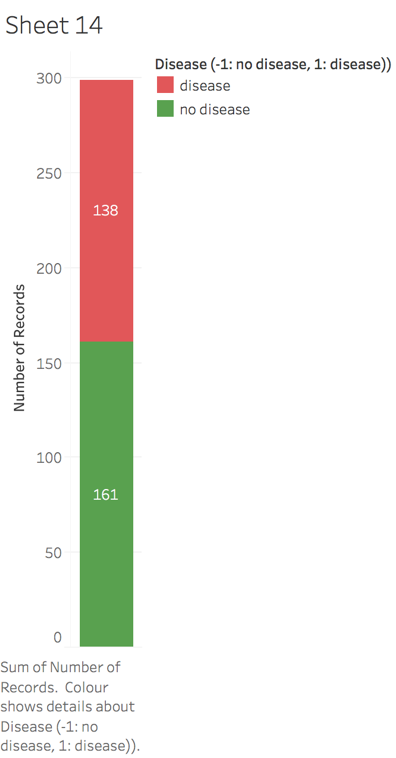
The basic question that should be answered by analyzing this data set are what patient attributes correlate with heart disease. Machine learning experts could use this labelled data set to construct a learning algorithm capable of predicting heart disease based on new patient data.

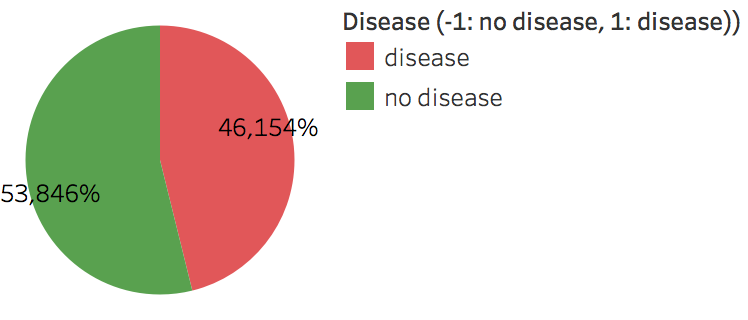
III. Results

*A. Analysis of Target Data*

The heart data set comprises a total of 299 patient entries with labels assigned to each patient showing whether the patient suffers from a heart disease (labelled with 1) or not (labelled with -1). Analyzing the data with regards to this feature shows that there are 161 patients without a heart disease and 138 patients with a heart condition in the data set (Fig. 1). A bar chart was used to visualize absolute numbers whereas the pie chart demonstrates percentage of total (Fig. 1). To conclude these insights, we are faced with an almost balanced data set, with the group of healthy patients exceeding those of diseased patients by 2.6%.

a)



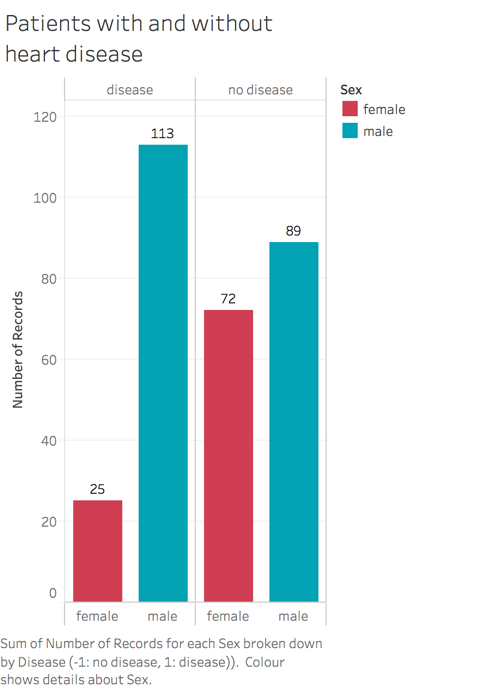


b)

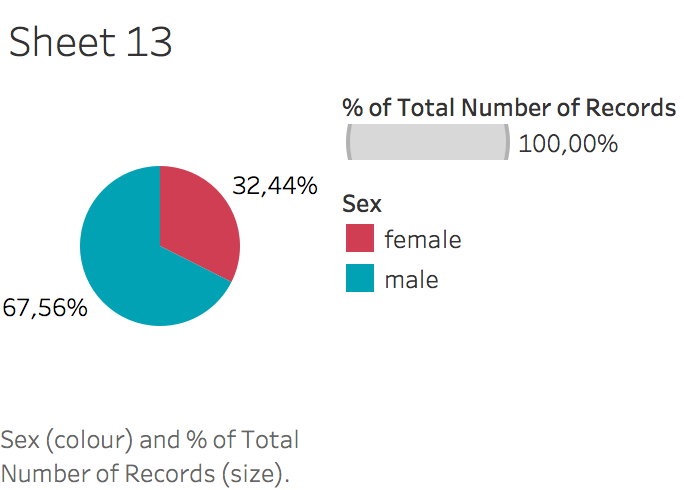
Fig. 1. Distribution of patients with and without heart disease in the data set. a) The bar chart shows the two target groups – no disease (green) and disease (red) in absolute numbers, b) the pie chart demonstrates the relative figures for the two groups.

*B. Assessing Population specific differences*

Sex and age are population specific parameters that are relevant in addressing heart disease. Out of the 299 patients in the database, 97 (or 32,44%) are women and 202 (or 67,56%) are male (see Fig. 2a). In order to find out, if men and women are equally affected by heart disease based on the given dataset, the target distribution in the two groups was assessed. Out of 138 patients diagnosed with heart disease there are 113 male patients and 25 females (see Fig. 2b). Given this data, it suggests that males are four times more likely to suffer from a heart disease than females. 161 patients from the data set were diagnosed without heart condition. The balance between male and female patients in this group is almost even (see Fig. 2b).



b)



a)

Fig. 2. Distribution of female and male patients in the data set as well as females and males meeting the target. a) The pie chart on the left shows the percentage of male (blue) and female (red) patients in the data set. b) The bar chart on the right highlights the number of females (red) and males (blue) with a diagnosed heart disease (2 left most columns) and without a heart disease (2 right most columns).

Age is an important factor when facing disease, with older age being a risk factor of acquiring a heart disease. For the analysis of the target heart disease in different age groups, the age dimension was grouped in bins of 10 years. The top histogram in Figure 3 shows the distribution of patients in the respective bins. The lower histogram has the percentages of healthy (green) patients and patients with heart disease (red) demonstrating an increase in heart disease with age, with most affected people in the group between 51 and 60 years. However, this result can be misleading since this age group has also the most patients in the dataset (see Fig. 3).

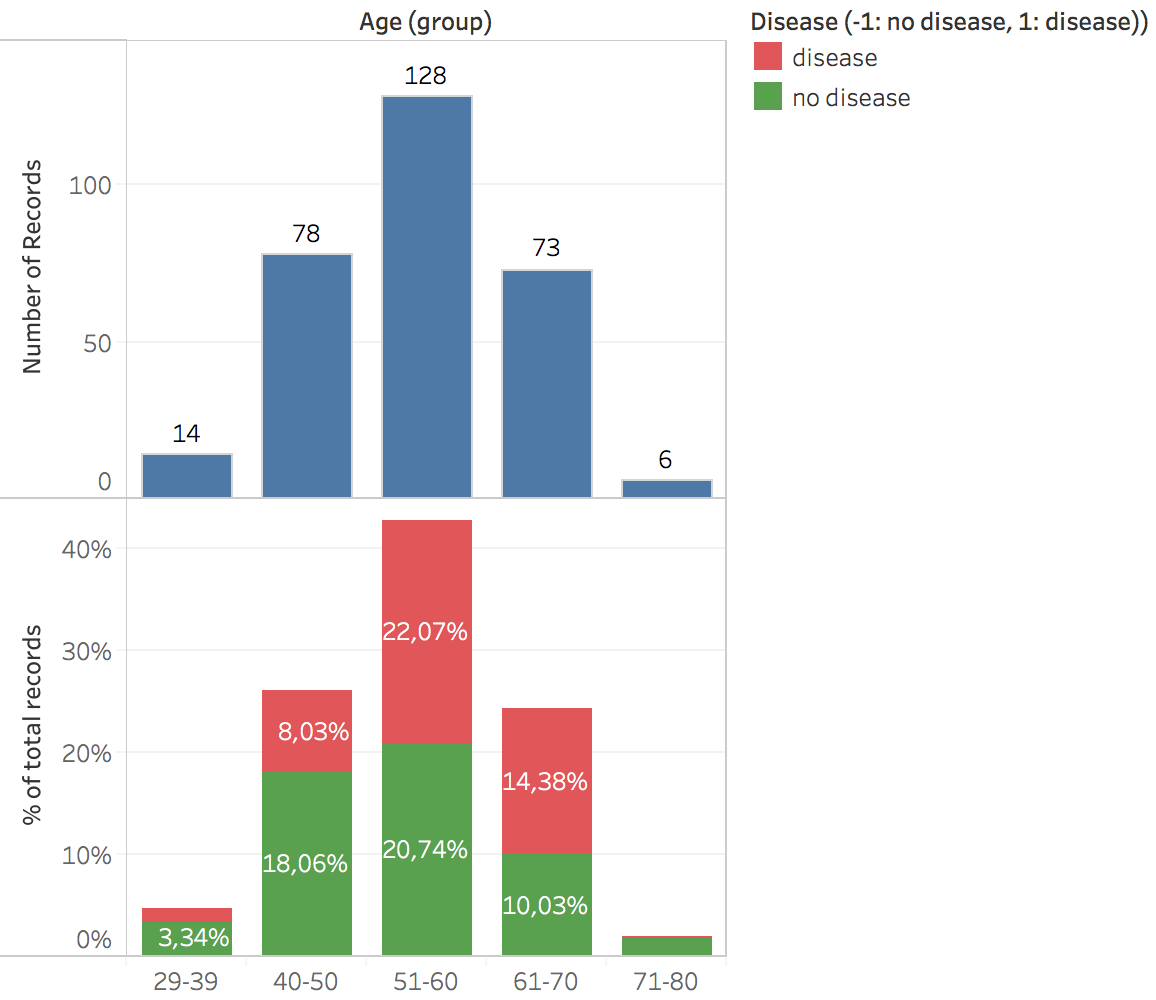
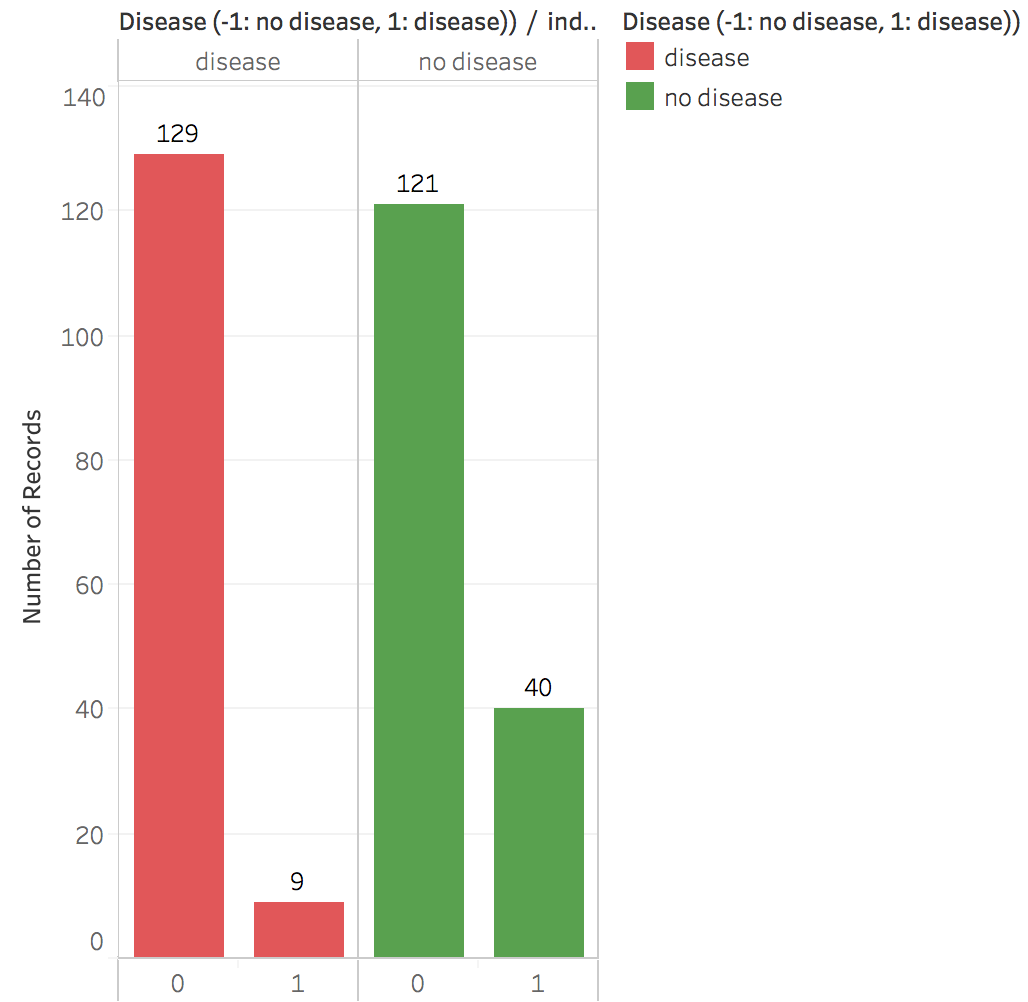


Fig. 3 Analysis of age-related heart disease in data set. The histogram on the top show the distribution of patients into bins of 10 years starting from 29 and ending at 80. The lower histogram shows the percentages of patients with an without disease in the respective age groups.

*D. Impact of symptoms on heart disease*

Angina, a type of chest pain caused by reduced blood flow in the heart, can be an indicator of a heart attack. Thus, analyzing the data set in respect of these indicators can provide insights if people with chest pain are more likely to suffer from a heart disease. Using typical and atypical angina as indicators, we can see that this type of chest pain is not a reliable feature to characterize people with a heart disease as most people with a heart indication have no prior chest pain (see Fig. 4a for typical angina and Fig. 4b for atypical angina).



a) Indicator of typical Angina

b) Indicator of atypical Angina

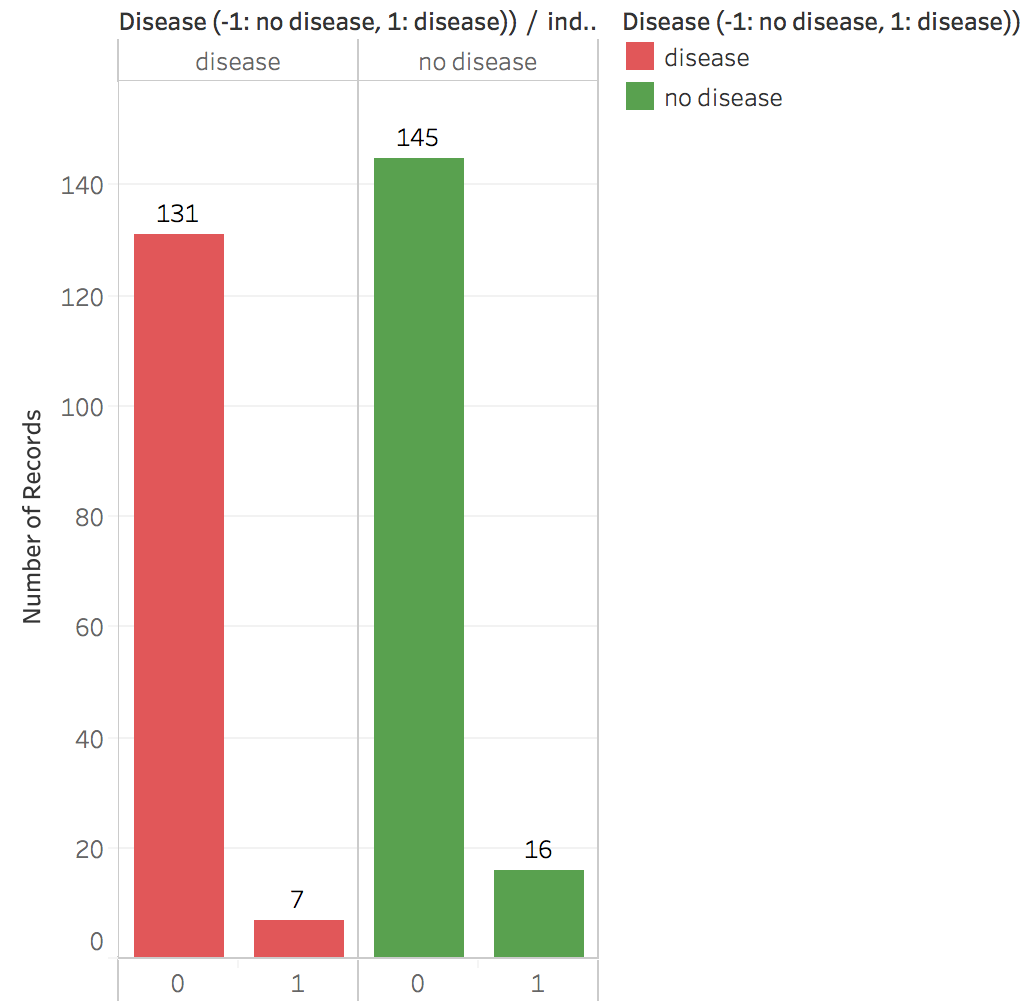


Fig. 4. Incidence of typical (left bar chart) and atypical (right bar chart) angina in patients with and without heart disease.

An elevated level of blood sugar might be an indicator for a heart disease. Thus, a fasting blood sugar concentration of greater than 120 mg/dl is a positive indicator (1) whereas blood sugar levels below (0) are considered healthy. The distribution of patients with elevated blood sugar levels does not correspond with disease occurrence. Patients without a heart disease are equally likely to have an elevated blood sugar concentration (see Fig. 5).

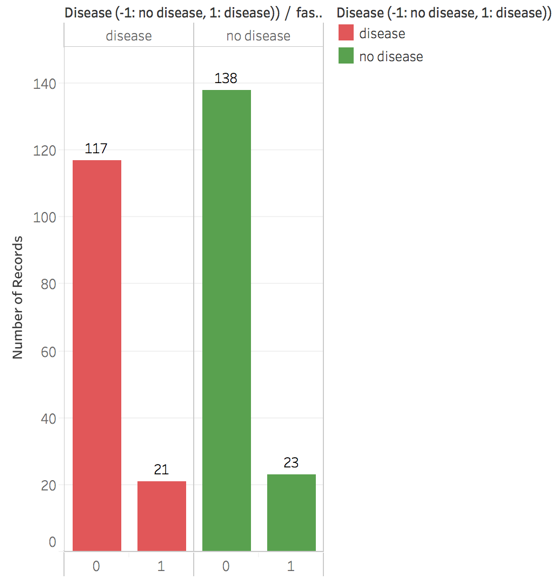


Fig. 5. Indication of high (1) and normal (0) blood sugar levels in patients with (red) and without (green) heart disease.

IV. Conclusions

In this paper, I take an attempt to an exploratory data analysis on the Cleaveland Heart Disease Data Set [4] based on a few selected features. Due to time restrictions, this analysis can only provide a start for further characterization of the information contained. The aim of this project paper was to design data visualizations using the Tableau Desktop software in order to gain insights into the data. The handling of the Tableau software is reasonably user-friendly with many elements that are self-explanatory and learning works by trial-and-error. Since this analysis was done for the purpose of drawing insights from the data set for myself, I used chart types and colors, that support my understanding. E.g. I was using green and red colors in the graphs for healthy and disease status because for me red always indicates danger whereas green suggests everything is ok (e.g. see Fig 1). Hower, one might have to think about this color-coding scheme as there are people with a diagnosed red-green blindness who would not be able to read my graphs.

From the data analysis itself, I cannot draw any meaningful conclusions because, I didn´t explore all of the features and their correlation with heart disease. At this point, I can only say that there is an almost balance between the two target groups (disease vs no disease), that there are more males than females in the data set as well as that chest pain and elevated blood sugar levels didn´t seem to impact this patient group.

References

*Weblinks*

[1] [www.kaggle.com](http://www.kaggle.com)

[2] <https://archive.ics.uci.edu/ml/index.php>

[3] <https://dataverse.harvard.edu>

[4] Bartley, Christopher, 2016, "Replication Data for: Cleveland Heart Disease”, <https://doi.org/10.7910/DVN/QWXVNT>, Harvard Dataverse, V1, UNF:6:uUXnE2XOKvaGcPfH8fzDpw==

*Journals*

[5] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304–310.

[6] David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."

[7] Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11–61.