Week 4 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

I. Description of Data Set

A

s

Found data on gapminder <https://www.gapminder.org/data/>

Under the indicator Health.

The world map was taken from the following site:

<https://simplemaps.com/resources/svg-world>

svg graphic is not optimal. Find png graphic: <https://commons.wikimedia.org/wiki/File:BlankMap-World.png> Download full resolution .png. This work has been released into the **[public domain](https://en.wikipedia.org/wiki/en:public_domain" \o "w:en:public domain)** by its author, **[Vardion](https://commons.wikimedia.org/wiki/User:Vardion" \o "User:Vardion)**.

Data on different cancers in female was obtained through the following site:

<https://www.gapminder.org/data/>

Latitude and Longitude data taken from: <https://www.mapsofworld.com/world-maps/world-map-with-latitude-and-longitude.html>

Since I couldn´t map the countries with the longitude and latitude coordinates onto this map, I was looking for alternatives and found a library for Processing capable of creating maps:

<https://www.gicentre.net/geomap/using>

geoMap allows you to draw vector maps in Processing. Unlike images, vectors define the shape of each object in the map and so preserve their detail when magnified. A shapefile actually consists of several files all with the same name, but with the extensions **.shp**(storing the geometry of the map), **.dbf** (storing the attributes) and optionally **.shx** (an index file). To use a shapefile with geoMap you need to place at least the .dbf and .shp files somewhere where your sketch can read them (e.g. by dragging them into your sketch).

The library was downloaded and installed in Processing.

I was making use oft he world shape file that comes with the geoMap library.

One of the advantages of vector-based mapping via shapefiles is that each item in the map is individually identifiable. We can exploit this to make our map interactive, highlighting the country under the current mouse position.

Each item (a country in this example) in a geoMap object will always have its own unique numeric ID. There are a number of methods that can use this ID. The key line in the example above is the one that calls **getID(mouseX, mouseY)**. Here geoMap will return the numeric ID of the map item at the given screen coordinates (mouse position in this example). We can then use this ID to draw, in a different colour, just the item selected. If there is no item at the given (x,y) coordinates, the method will return a -1, so we test for that to ensure we only draw when a map item has been found.

Start new with data set for US: <https://github.com/fivethirtyeight/data/blob/master/hate-crimes/hate_crimes.csv>

| **median\_household\_income** |
| --- |

**share\_non\_white**

| **share\_voters\_voted\_trump** |
| --- |

Houshold doesn´t work because too large values replace with:

| **share\_white\_poverty** |
| --- |

II. Data Cleaning

Select only data from year 2016 of all 3 data sets. Delete columns from the data sets that leave only values for 2016.

See if countries are the same.

Life expectancy female: 185 countries

Breast cancer female: 188 countries

Cervical cancer female: 188 countries

Delete Andorra, Dominica and Marshall Islands from breast cancer and cervical cancer data set. Save as .csv files.

Clean locations data to reflect only countries that are in data set.

Use the following image size to capture the entire world: 1500 × 740

III. Results

*A. Importing the world map*

PImage mapImage;

void setup( ) {

size(1500,740);

mapImage = loadImage("world.png");

}

void draw( ) {

background(255);

image(mapImage, 0, 0);

This works.

Fig. 1.

*B. Assessing Population specific differences*

Sex

b)

*D. Impact of symptoms on 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

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.