CS732/DS732: Data Visualization – Datathon 5

Kunika Valecha

November 05, 2020

Contents

Part I
Multivariate Data Visualization

1 Problem Statement

visualize using parallel coordinates and scatterplot matrices (with upto 7 variables at a time).

2 Data Description

This dataset was retrieved in the JSON-stat format using UNECE's API and transformed with jsonstat-conv (json-stat.com).

2.1 Metadata

- name: UNECE's Country Overview
- file type : json
- url: UNECE's Country Overview
- description: Changes in economical and social aspects of different countries.
- long-description: The dataset consists of 52 countries studied for their social and economic aspects in order to observe the overall growth of countries for 17 years.
- Each row of the data represents an observation for one country in one year and the columns hold the variables (data in this format is known as tidy data).

3 Data Reading

- import json file.
- store the json file content into a list.
- Convert the list into a dataframe.
- Convert the string values of each column into float values.

4 Multivariate Data Visualisation

Multivariate datasets with typically more than two attributes are hard to visualize since two variables restrict our data analysis. In Multivariate visualization, Parallel Coordinates and Scatterplots remain one of the widely used visual representations of them all.

4.1 Scatter plot matrix

A scatter plot matrix is a grid (or matrix) of scatter plots used to visualize bivariate relationships between combinations of variables. Each scatter plot in the matrix visualizes the relationship between a pair of variables, allowing many relationships to be explored in one chart.

A scatter plot matrix is composed of a grid of mini-plots and one larger preview plot that shows a selected mini-plot in more detail. Additionally, a histogram showing the distribution of each numeric variable can be added to the matrix by checking Show histograms in the Chart Properties pane.

4.1.1 Variables

A scatter plot matrix is made up of three or more Numeric fields. A scatter plot will be created for every pairwise combination of variables.

4.1.2 Statistics

A regression equation is calculated for every scatter plot in the matrix. The associated trend lines can be added to the scatter plots by checking Show linear trend in the Chart Properties pane. Alternatively, the mini-plots in the grid can be viewed as R^2 values with a color gradient corresponding to the strength of the R^2 value by checking Show as R2 in the Chart Properties pane.

4.1.3 Appearance

Titles and description Charts and axes are given default titles based on the variable names and chart type. These can be edited on the General tab in the Chart Properties pane. You can also provide a chart Description, which is a block of text that appears at the bottom of the chart window.

Visual formatting You can configure the look of your chart by formatting text and symbol elements, or by applying a chart theme. Format properties can be configured on the Format tab in the Chart Properties pane, or through the Chart Format context ribbon. Chart formatting options include the following:

- Size, color, and style of the font used for axis titles, axis labels, description text, legend title, legend text, and guide labels
- Color, width, and line type for grid and axis line

- Background color of the chart

Color Scatter plot points can be visualized using a single color, or with the colors specified in the layer's symbology. By default, scatter plots use layer colors and inherit their outline and fill colors from the source layer symbology.

4.2 Scatter plot matrix Layout for UNECE's Country Overview

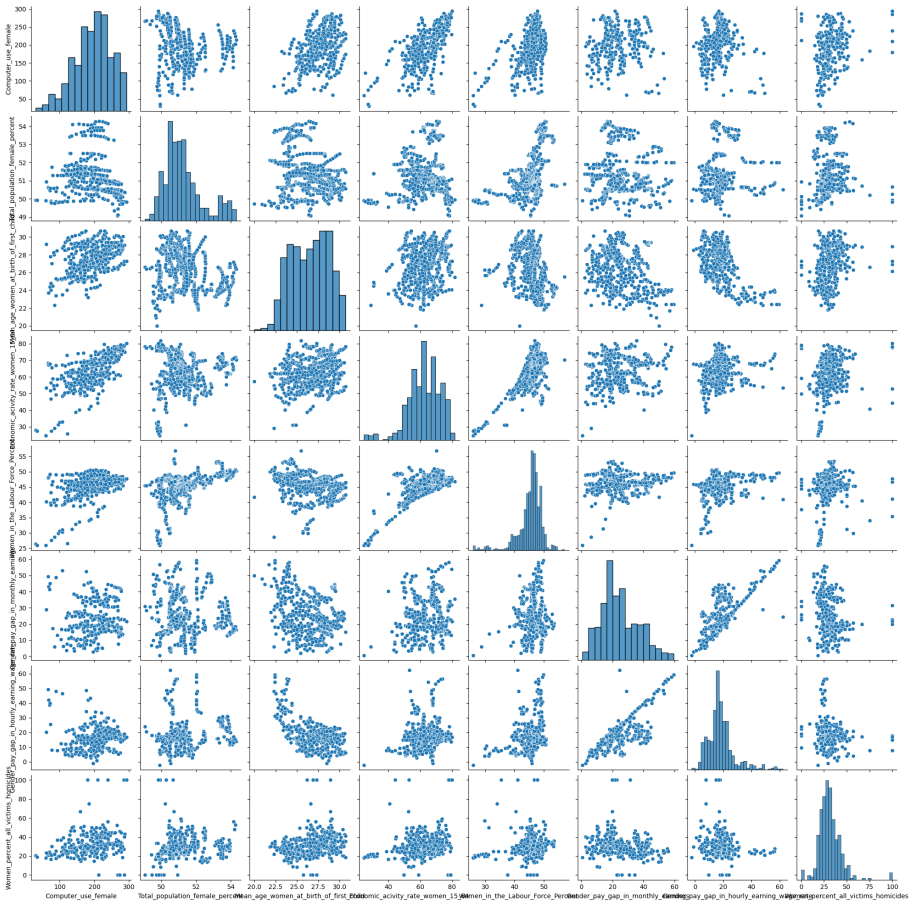


Figure 1: Figure showing scatter plot matrix for UNECE's Country Overview Data without hue.

The above pair plot is a result of pairing between following attributes of data,

- Use of Computer among all age groups of female

- Total population female percent
- Mean age of women at birth of first child
- Economic activity rate amongst women of age group 15 to 64
- Women in the Labour Force Percent
- Gender pay gap in monthly earnings
- Gender pay gap in hourly earning wage rate
- Women percent all victims homicides

The pairs plot builds on two basic figures, the histogram and the scatter plot. The histogram on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship (or lack thereof) between two variables.

The default pairs plot by itself often gives us valuable insights. We see that economic activity rate and female labour force are positively correlated showing that for female employment labour force forms a huge contribution.

While this plot alone can be useful in an analysis, we can find make it more valuable by coloring the figures based on a categorical variable such as country.

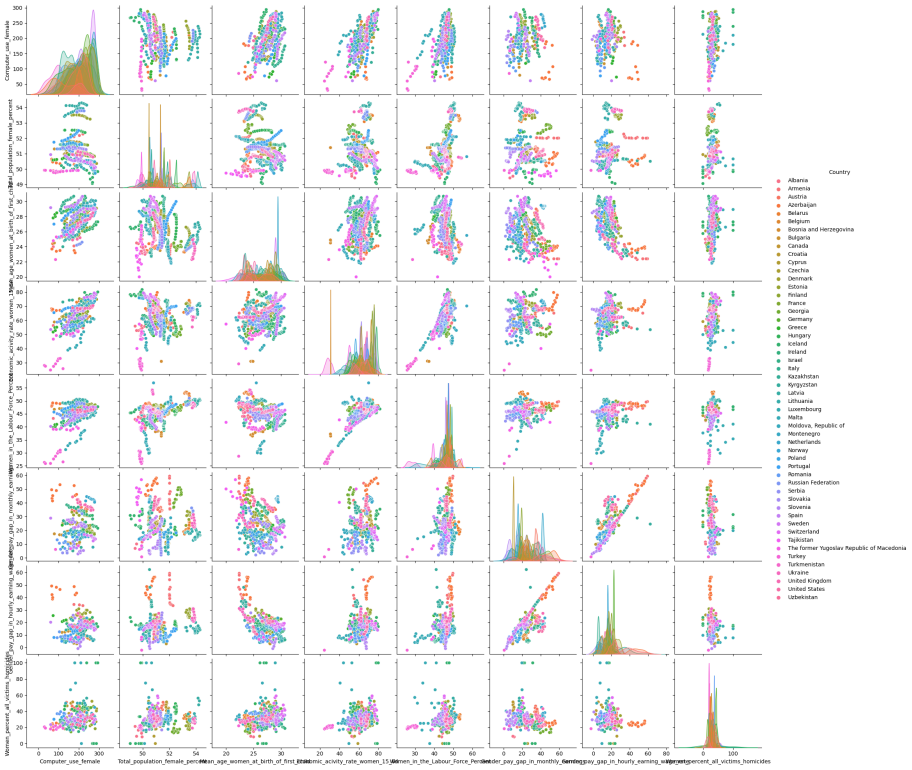
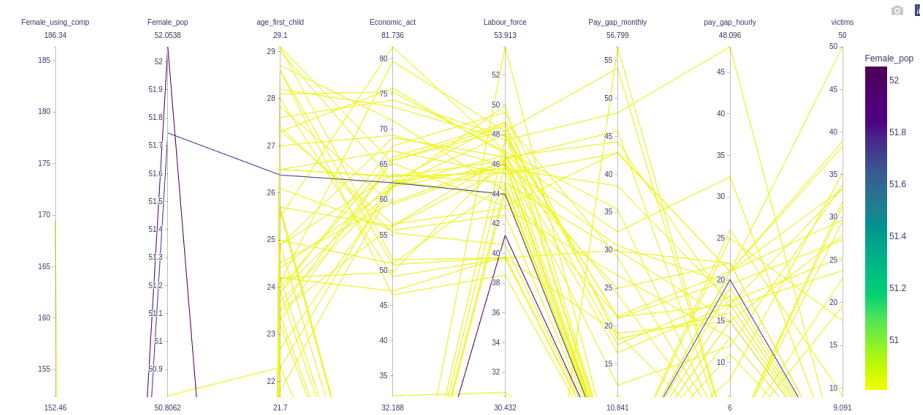


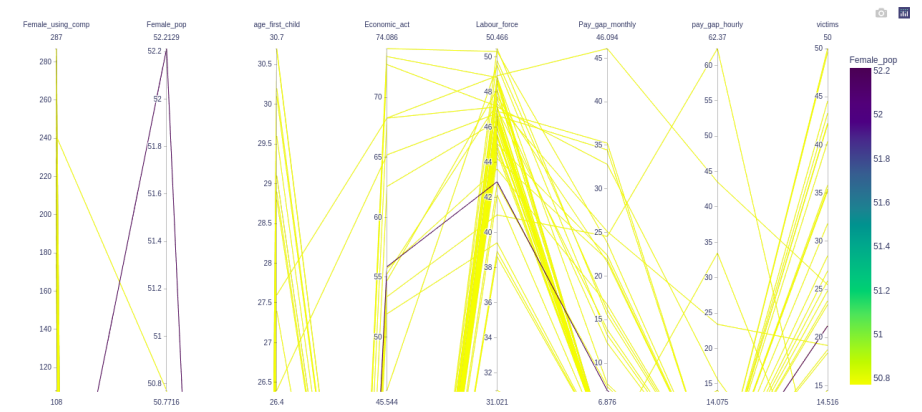
Figure 2: Figure showing scatter plot matrix for UNECE's Country Overview Data with hue.

4.3 Parallel Coordinates

Parallel Coordinates Plots is closely related to time series visualization, except that it is applied to data where axes points don't correspond to time. These are ideal for comparing many variables together and seeing the relationships between them. In plot, a set of parallel axes is drawn, each representing a variable. Each axis can have a different/uniform scale. Values are plotted as a series of lines that connected across all the axes. In other words, parallel coordinates plot is a collection of polylines where intersection points of axis lines and polylines are data values of variables for each data item in the dataset.



(a) 2000



(b) 2015

Figure 3: Figure a, b showing all the correlations amongst given variables in year 2000 and 2015

Part II
Hierarchical Data Visualization

1 Problem Statement

To build hierarchical relationships based on time and/or space, and use the treemap and sunburst visualizations,

2 Data Description

The data source is COVID-19 dataset.

- **COVID-19 Metadata**: The tabular datasets published by the World Health Organization to create new networks (similarity networks, correlation networks, etc.) and visualize network communities.
- The data is in tabular format and is contained inside multiple csv files in multiple formats.
- One of the formats is the time-series format in which the number of cases(deaths, recovered, and confirmed) are being recorded day-wise and stored in accordance with there positions in world coordinates.
- The time-series data for United states particularly is also available.
- The data has been recorded from 22nd January, 2020 to 23rd September,2020.
- Three categories of cases are recorded in three separate files viz.
  - time_series_covid_19_deaths.csv,
  - time_series_covid_19_recovered.csv, and
  - time_series_covid_19_confirmed.csv.

3 Data Reading and Cleaning

- Since the dates of data are represented by columns and rows represent the States and their respective countries worldwide, so the data has been transposed in order to labels of nodes in nodelink diagram as name of countries.
- Due to presence of a large amount of States the node-link diagram might look cluttered. Also, for some of the observations the information of the particular state is absent, leaving a NaN value in the data frame, this is why the countries are considered to be a better choice in order to represent the labels on the nodes.

- As the the rows of the dataset are unique from each other due to State identity, but as we consider the countries for the basis of our labelling then there might occur repetitions as several state can have a common country. In order to avoid this repetitions, we have taken the sum of all the values belonging to a particular country on a specific date.

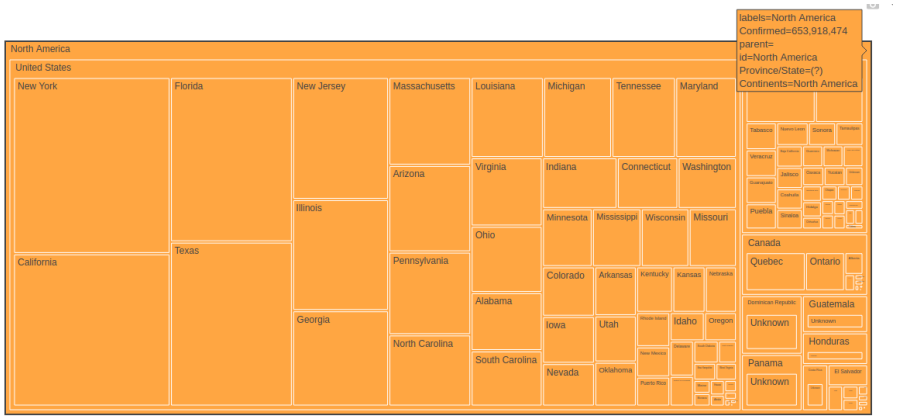
4 Hierarchical Data Visualization

Hierarchical Visualization methods that show how data or objects are ranked and ordered together in an organization or system. Here we perform Treemap and Sunburst Diagrams on the countries of world data following the hierarchy of countries based on region. We visualize a few attributes like recovery, death, confirmation rate of COVID 19 cases across the world.

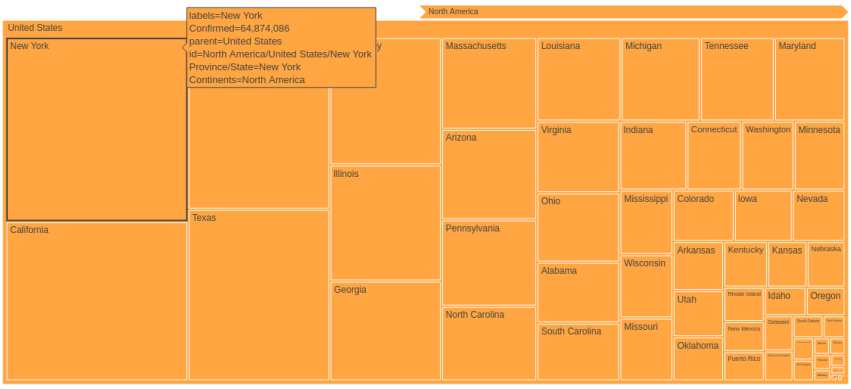
4.1 Tree-Map

Treemaps are a more compact way of visualizing the hierarchical structure of a tree diagram. It helps us to compare the proportions of data between categories via their area size. It displays quantities under each category via rectangular area size, which is the proportion of quantity with other quantities in the same parent category and subcategory rectangles nested in it. The tiling algorithm is used to order and divide the rectangles. But to the darker side, treemap doesn't show the hierarchal levels as clearly as the basic Tree Diagram or Sunburst Diagram.

(a) World



(b) North America

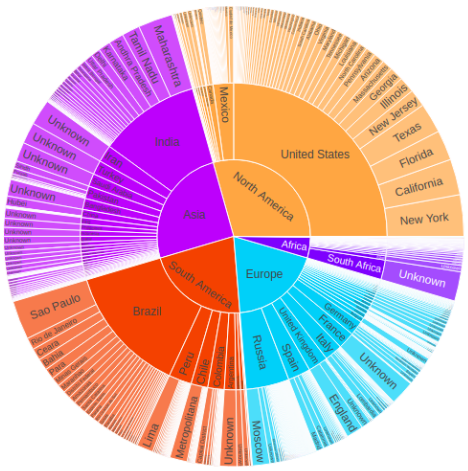


(c) USA

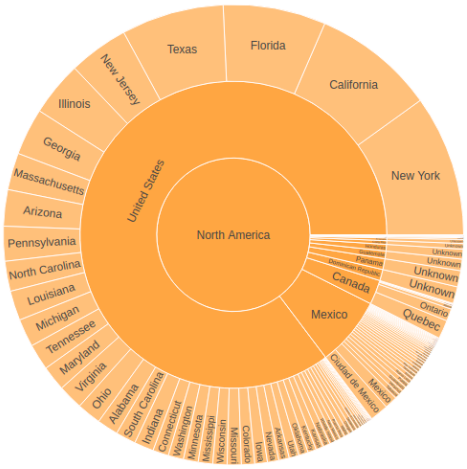
Figure 4: Figure a, b, c showing the hierarchy of the regions and cases of confirmation there.

4.2 Sunburst Diagram

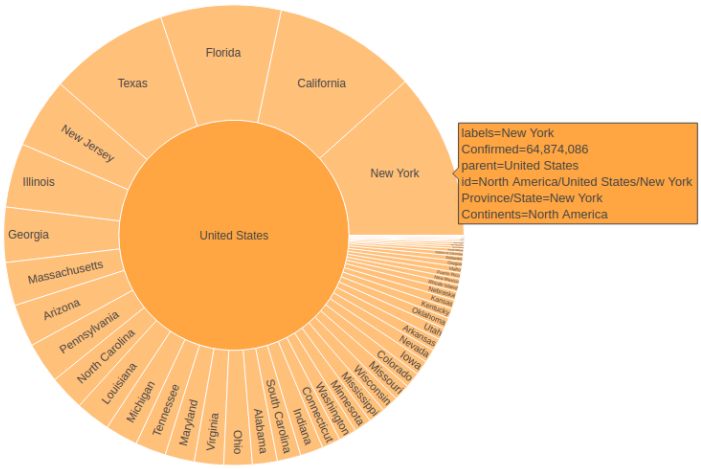
The Sunburst Diagram or multi-level pie chart is an alternative, area filling visualization that uses a radial layout. The hierarchy regions depicted by concentric circles. The center of the circle is a root node, and a segment of the inner circle bears a hierarchical relationship to those segments of the outer circle which lie within the angular sweep of the parent segment. The color of arc corresponds to some attribute of the data.



(a) World



(b) North America



(c) USA

Figure 5: Figure a, b, c showing the hierarchy of the regions and cases of confirmation there.

5 Technologies Used

- seaborn library python
- plotly library python
- matplotlib python