Data Visualization Assignment 2 Report

1st Subham Agarwala
IMT2022110
IIIT Bangalore
Subham.Agarwala@iiitb.ac.in

2nd Sarvesh Kumar
IMT2022521
IIIT Bangalore
SarveshKumar.A@iiitb.ac.in

3rdAyush Gupta
IMT2022546
IIIT Bangalore
Ayush.Gupta546@iiitb.ac.in

I. Introduction

This project encompasses two main categories of tasks: Information Visualization and Scientific Visualization.

A. Scientific Visualization

Scientific Visualization is further divided into the following sub-categories:

1) **Quiver Plots**
2) **Contour Maps**
3) **Colour Maps**

B. Information Visualization

Information Visualization includes the following sub-categories:

1) **Node-link Diagrams**
2) **Treemaps**
3) **Parallel Coordinate Plots**

C. Division of Work

Scientific Visualization

- **Quiver Plots** - Subham
- **Contour Maps** - Sarvesh
- **Colour Maps** - Ayush

Information Visualization

- **Node-Link Diagrams** - Subham
- **Treemaps** -Sarvesh
- **Parallel Coordinate Plots** - Ayush

II. Scientific Visualization

A. *Quiver Plots*

Quiver plot is basically a type of 2D plot which shows vector lines as arrows. Here we will use a quiver plot to plot wind data (direction and magnitude) over the US.

Dataset Description: The dataset consists of two NetCDF files, th_2001.nc and vs_2001.nc, each containing wind-related meteorological data for the year 2001.The data is organized spatially as well as temporally across latitudes and longitudes and across different days of a year respectively. The dataset provides us with sufficient information about daily wind conditions over the United States to perform analysis and make visualizations on wind speed and wind direction.

The dataset has variables that include latitude, longitude, day (unique days in 2001) and wind direction in degrees. We have only chosen selective dates from January- March for plotting and analysis, and chosen dates such that they include a range of wind speeds and directions.

The dates selected are:

- 10th of January, February and March
- 20th of January, February and March
- 30th of January and 28th of February and March.

Preprocessing and Data modelling: In the beginning of this project, two NetCDF datasets were loaded which represented wind parameters, i.e. ‘th_2001.nc‘ which represented wind direction denoted (‘wind_from_direction‘) while ‘vs_2001.nc‘ represented the wind zur speed – (‘wind_speed‘). These files were imported into python utilizing the ‘xarray‘ library which is a perfect fit for handling meteorological data arrays that are large and multi-dimensional with ease of indexing and manipulation along dimensions of time and latitudes and longitudes, when needed.

After they were imported, the data fields of interest such as the following were extracted,; latitude, longitude, dates, wind direction and Wind speed. We also restricted the range of analysis by considering only the dates within January to March that fell on the 10th, 20th, and 28th of the respective months so as to ease the present workload. This choice was made so as to avoid bulkiness of the time resolution and yet provide clarity for the animation sequence capturing most of the key moments within the timeline. So down sampled the latitude and longitude grids by taking every thirtieth data point, thereby decreasing density and improving clarity of the quiver plot visualization.

In order to convert the wind speeds and direction parameters to a vector form, it was necessary to convert the polar coordinates into U and V components signifying the eastward and northward wind vectors respectively. This aids plotting of the wind speed and wind direction data as vectors on the map.

Implementation: The implementation commenced with configuring a quiver plot that depicts wind flow all over the continental United States. For this task, libraries such as ‘matplotlib‘ and ‘cartopy‘ were utilized and appropriate projections (‘ccrs.PlateCarree‘) were employed and the area of interest was constrained to pertinent regions in order to improve focus on the matters at hand.

The quiver plot also required wind speed and direction to be presented in a vector form (U & V), which was done by first converting wind direction angles to trigonometric forms. To alleviate the problem of arrow congestion or prevent the arrows from overlapping, a sampling interval (‘SKIP‘) was employed

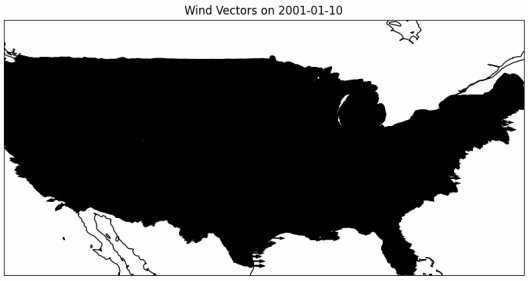

Fig. 1: Initial quiver plot without sampling of data.

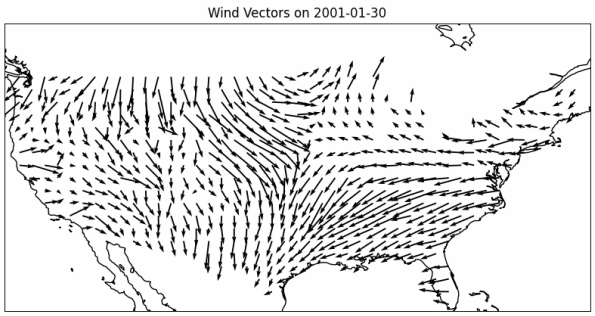

Fig. 2: Varying length quivers scaled to width of plot.

to bring clarity and comprehensibility in the visualizatior Fig.1 shows how the plot looks without sampling the dat points. The 'update_quiver' function was designed for th purpose of updating the figure for every date selected, whic contributed to the day animating smoothly as the U and V vectors were updated along with the quiver arrows on eac date.

***Visual exploration**:* We employed various techniques t plot the quivers and ensure clarity in the visualizations. Th techniques included downsampling data points, using vector lengths that vary with wind speed magnitude, varying lengt vectors with wind speed encoded using colour, constant length vectors with speed data encoded using colour etc.

The two broad categories could be considered as varying length and constant length quivers. Let us take a look at the varying length quivers first.

The scale_units parameter allows us to choose between scaling the quiver lengths (length encoded with wind speed magnitude) based on plot width as seen in Fig.2 and plot height as seen in Fig.3. we can clearly see that for the dimensions we are dealing with scaling on height of the plot yield more visually relaxing results, as there is lesser clutter as compared to the plot scaled on width. So for further exploration, we consider only height-scaled quiver plots.

Using only length to encode wind speed can cause difficulty in interpreting data in congested areas and also in areas where

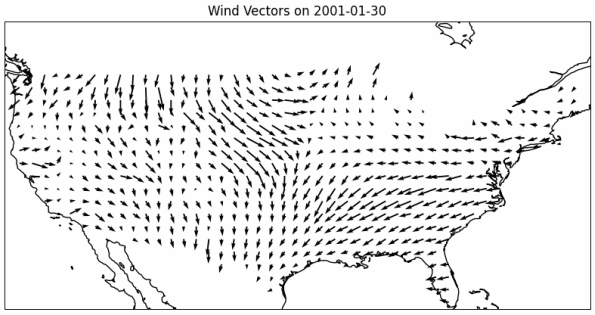

Fig. 3: Varying length quivers scaled to height of plot.

wind speed is low so the vector length is small. So we also bring in colour encoding to add additional interpretability. Colour encoding makes it easier to indentify low-magnitude regions, read data in dense regions and enhanced perception of gradients which enables identifying differences even when change in arrow length is subtle.

We have also chosen a global colour scale instead of a dynamic one, because it allows us to compare and read data across time without having the users to adjust to frequently changing colour scale for different days.

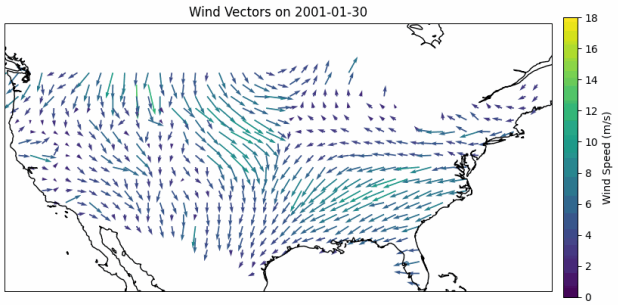

Fig. 4: Varying length quiver scaled to height of plot with colour encoding.

Fig.4 shows how using colour encoding we get a clearer understanding of data in the regions with higher magnitudes. But we can see that despite the visualization's ability to convey data, there still exist regions with overlapping and clustered quivers which degrade the the aesthetics of the visualization. We can solve this using constant length and then using colour encoding to provide wind speed data.

Fig.5 shows the use of constant-length vectors to visualize the data. We see that when we use constant length vectors it becomes necessary to use colour encoding to show speed data, otherwise we get information only about wind direction using constant length vectors. Also we can see that on using constant length vectors we have a scope of reducing the number of data

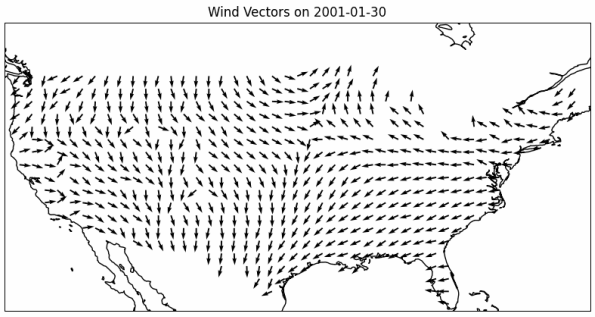

Fig. 5: Constant length quivers.

points we are skipping and plot more data points as the clutter is reduced.

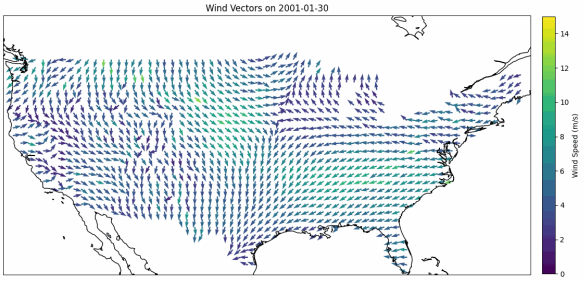

Fig. 6: Constant length quivers with colour denoting wind speed.

So we reduced the SKIP to 20 from 30 qand plotted the data using colour encoding to denote wind speed as can be seen in Fig.6. We see that using constant length quivers with colour encoding is beneficial especially when animating or stitching together multiple instances of the data across time as the colour gradient helps us notice the difference and change in wind speed in a region over time and also the constant length quivers result in a neat and clear visualization.

B. Contour Plots

Dataset:

The dataset used is the gridMET dataset, which provides surface meteorological data. The time frame selected for this task is from January 1, 2001, to March 31, 2001. The specific dates chosen for generating plots are:

- 1st January 2001
- 11th January 2001
- 21st January 2001
- 31st January 2001
- 10th February 2001
- 20th February 2001
- 2nd March 2001
- 12th March 2001
- 22nd March 2001

Tools Used:

- **Matplotlib:** It is utilized for generating contour plots, where I used its *contour* and *contourf* functions. These functions rely on Marching Squares algorithm to compute contour locations. [2]
- **Cartopy:** A library providing cartographic tools for Python for plotting geospatial data.
- **Imageio:** For creating animations (GIFs) from a series of images.
- **netCDF4:** To read metadata from NetCDF files.
- **xarray:** A Python library for handling NetCDF data efficiently.

Pre-processing:

In this task, I utilized two NetCDF datasets: tmmx_2001.nc (maximum near-surface air temperature) and tmmn_2001.nc (minimum near-surface air temperature). The datasets were pre-processed using the **xarray library**.

Key steps in the pre-processing included:

- Data Extraction: Extracted essential features such as latitude, longitude, and air temperature for creating spatial visualizations.
- Time Range Filtering: Focused on the first quarter of 2001 by slicing the datasets to include dates from January 1, 2001, to March 31, 2001.

Implementation:

- Map Setup: Configured contour and contour fill plots using the *ccrs.PlateCarree()* projection. Geographic features like coastlines, borders, and states were added for clarity.
- Diurnal Temperature Range: Calculated the difference between maximum and minimum temperatures from the datasets to analyze daily temperature variations.
- Animations: Generated temporal visualizations by compiling the 10-day interval plots into GIFs using the **Imageio library**.

Visualisations:

Fig.7, Fig.8 and Fig.9 shows the temporal progression of maximum near-surface air temperatures across the U.S. for the 1st quarter of 2001. The visualizations offer detailed geographical representations of temperature variations, highlighting both spatial and temporal changes influenced by seasonal transitions.

Fig.7 reveals a stark latitudinal temperature gradient. Northern states have extremely low maximum temperatures-low enough to freeze in Minnesota and Wisconsin, whereas southern zones such as Arizona, Florida, and Texas see relatively warm conditions-reaching temperatures of up to 20°C. The mountainous areas seem cooler, being at higher elevations.

Visualisations show a clear and consistent warming pattern through the first half of January into March. The nation as a whole sees an increase in peak temperatures by around 10K. This fits the expected climatic change patterns as the Northern Hemisphere progresses from winter into spring.

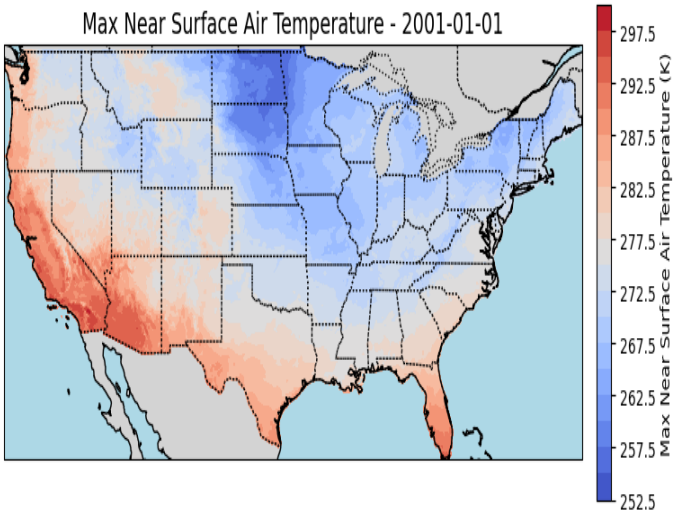

Fig. 7: Max Near-Surface Air Temperature on 1st Jan 2001

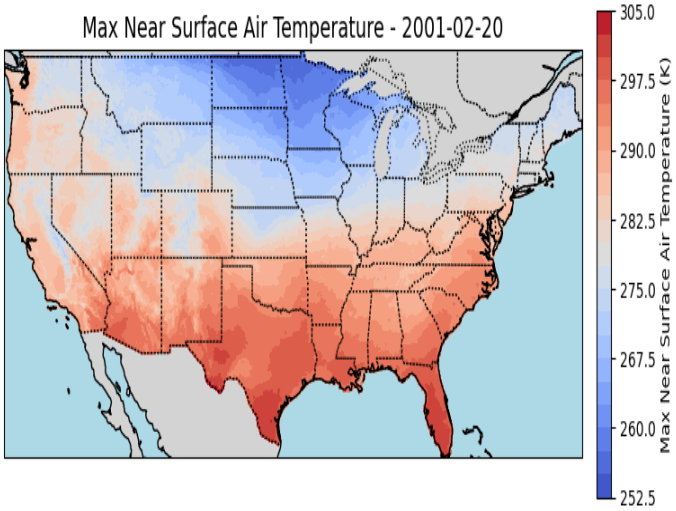

Fig. 8: Max Near-Surface Air Temperature on 20th Feb 2001

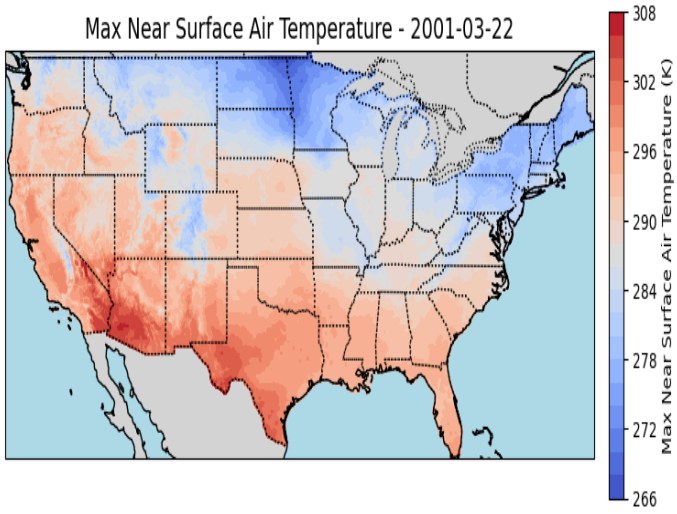

Fig. 9: Max Near-Surface Air Temperature on 22nd March 2001

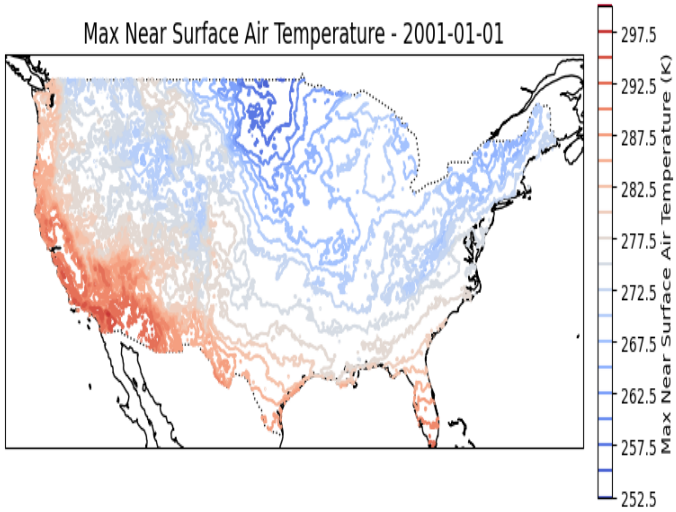

Fig. 10: Max Near-Surface Air Temperature on 1st Jan 2001

In Fig.7, the implementation of a contour fill algorithm provides a more intuitive and visually accessible representation of the data, thus making it easier to interpret the underlying temperature patterns compared to Fig.10, which displays only contour lines. Therefore, I have opted to use the contour fill approach for the remaining visualizations.

Building upon our earlier observations, we can delve deeper into the contour maps illustrating minimum near-surface air temperatures across the United States during the first quarter of 2001 in Fig.11, Fig.12, and Fig.13. These visualizations not only confirm the expected warming trend from winter to spring but also offer richer insights into the period's climatic dynamics.

Starting from a very cold situation, the northern states experience a significant warming trend by the time March arrives, making conditions more comfortable and even attracting tourists to places like Salt Lake City, where spring temperatures are ideal for outdoor activities. In the central plains, the moderate warming shift is beneficial for farming, as milder weather supports early-season crop growth. In the South, states such as Texas and Florida remain fairly warm all the time but they have also shown a gradual warming.

Geography, in addition to weather, also influences temperature distribution. Places with high altitudes like the Rocky mountains and the Appalachian mountains remain cold, while the coastal strips, which are influenced by the Pacific and Atlantic Oceans, have moderate temperatures due to their ability to absorb or dissipate heat thereby preventing extremes.

With the rise in temperature, what may happen is that the heating demands particularly in the colder region, may be reduced thus the energy costs may be saved.

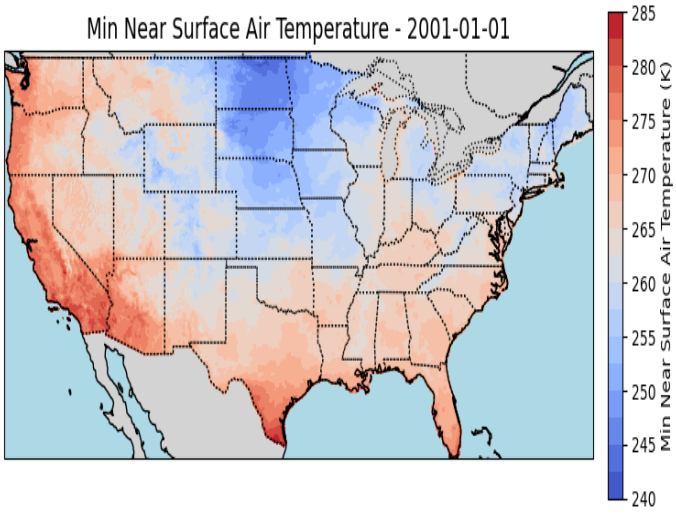

Fig. 11: Min Near-Surface Air Temperature on 1st Jan 2001

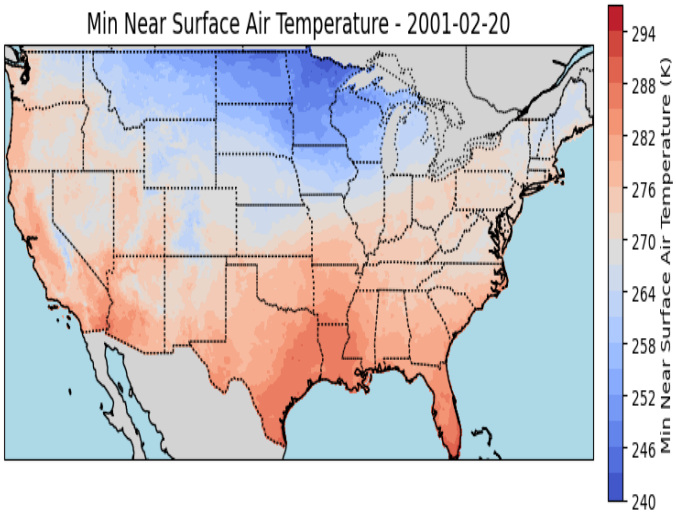

Fig. 12: Min Near-Surface Air Temperature on 20th Feb 2001

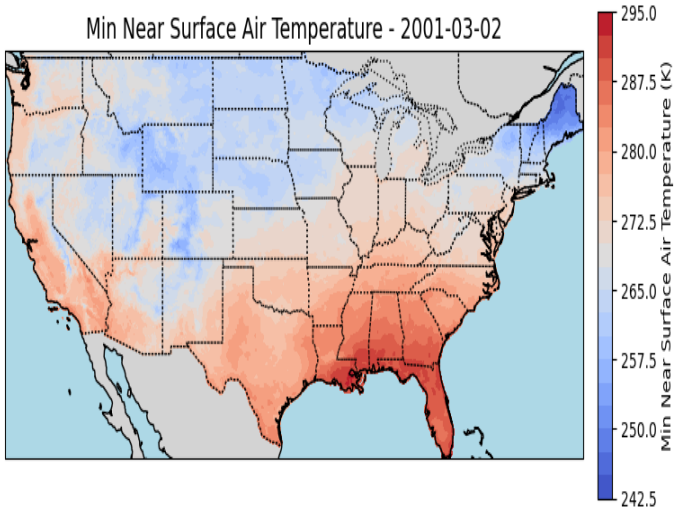

Fig. 13: Min Near-Surface Air Temperature on 2nd March 2001

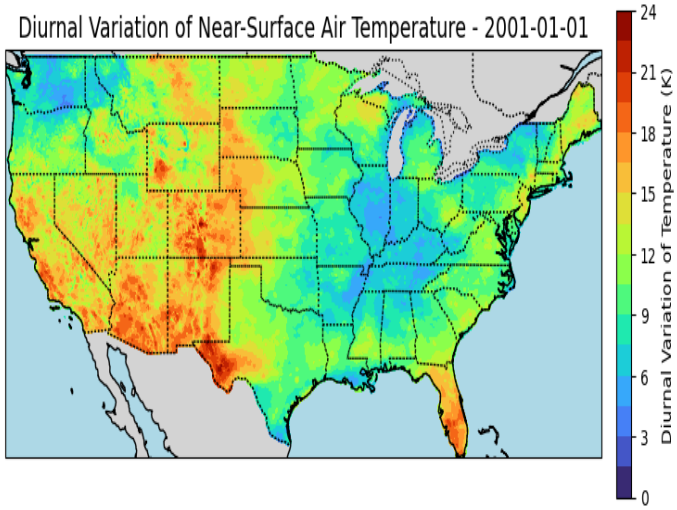

Fig. 14: Diurnal Variation of Near-Air Surface Air Temperature on 1st Jan 2001

The temperature range plots for the first quarter of 2001 shown in Fig.14, Fig.15 and Fig.16 reveal differences between maximum and minimum near-surface air temperatures across the United States, offering insights into various climate and environmental patterns.

In deserts of Arizona and New Mexico, the variation of diurnal temperature is reportedly wide. Due to low humidity, the days are hot and the nights very much cooler, these conditions tend to lead to the highest daily temperature changes.

Additionally, coastal areas benefit from land and sea breezes. During the day, cooler air from the ocean (sea breeze) moves inland as the land heats up faster than the water, cooling coastal areas. At night, as the land cools more quickly, warmer air from the ocean flows toward the shore (land breeze), which helps keep coastal temperatures mild even after sunset. This continuous exchange of air contributes to the stable, moderate temperatures characteristic of coastal regions.

These temperature fluctuations are of great importance for agriculture and energy usage. In agriculture, high daily temperature fluctuations can injure crops and change their growth cycles, which can influence planting decisions and the distribution of resources.

C. *Color Maps*

Color Maps are plotted for the gridMet dataset for the days between January of 2001 to March of 2001. The following variables yielded good visualizations through color maps:

- precipitation
- rmax : Maximum relative humidity

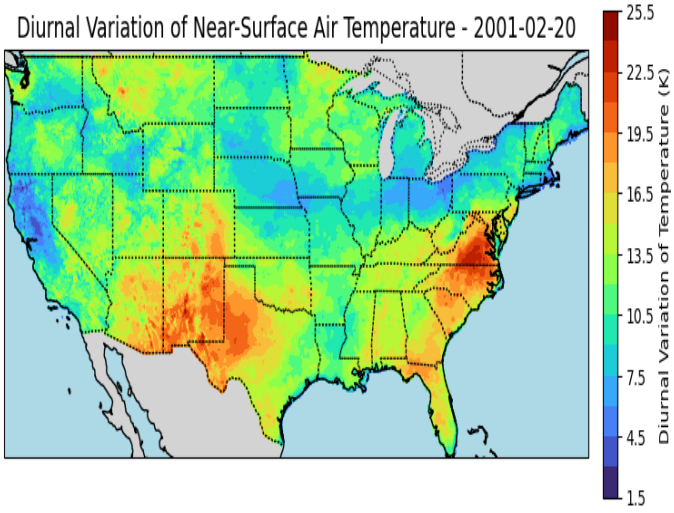

Fig. 15: Diurnal Variation of Near-Air Surface Air Temperature on 20th Feb 2001

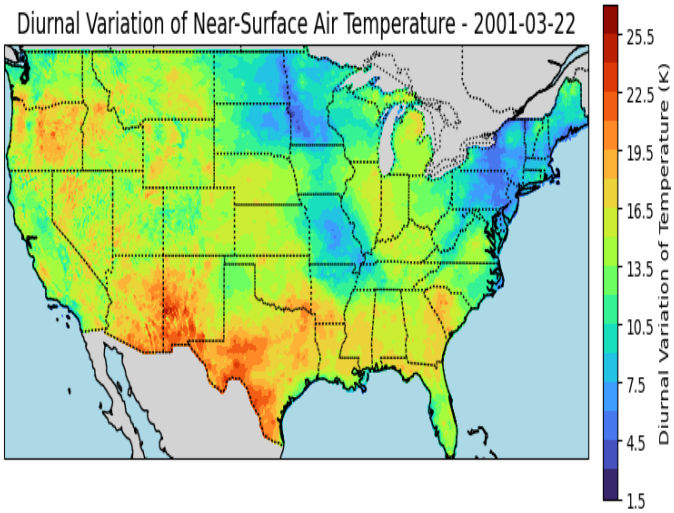

Fig. 16: Diurnal Variation of Near-Air Surface Air Temperature on 22nd March 2001

- rmin : Minimum relative humidity
- vpd: Mean Vapour Pressure Deficit

These variables were available as daily snapshots. Multiple colormaps were made from these variables, the following days are chosen for presentation:

- 1st Jan 2001
- 10th Jan 2001
- 19th Jan 2001
- 28th Jan 2001
- 6th Feb 2001
- 15th Feb 2001
- 24th Feb 2001
- 5th Mar 2001
- 14th Mar 2001
- 23rd Mar 2001

Multiple colormaps will be presented below. The flow of the variables over 3 months can be seen in gif form in the folders submitted.

Precipitation: : Two colormaps were made from the precipitation data. Precipitation data was unique in the sense that it was 0 for a majority of data points. There were also extremely high values such as 200. This meant that variances in the lower range of values could not be observed.

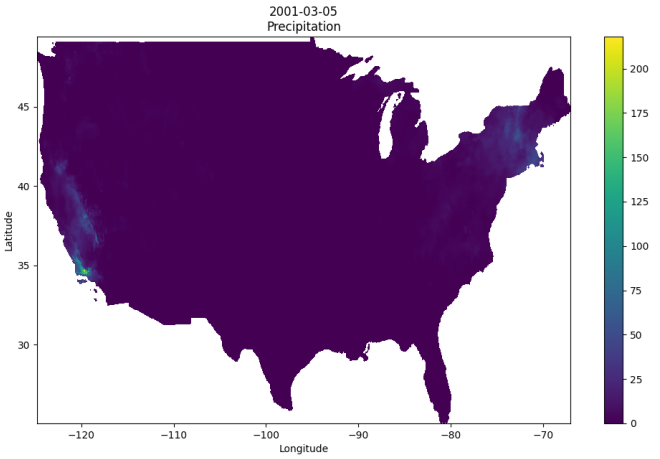

Fig. 17: Precipitation Graph

This can be seen in Fig. 17. Hence, it made sense to use a logarithmic colormap which shows both the high values and also the variance in lower values. Global minima and maxima are used in this graph for scaling the colormap to maintain consistency.

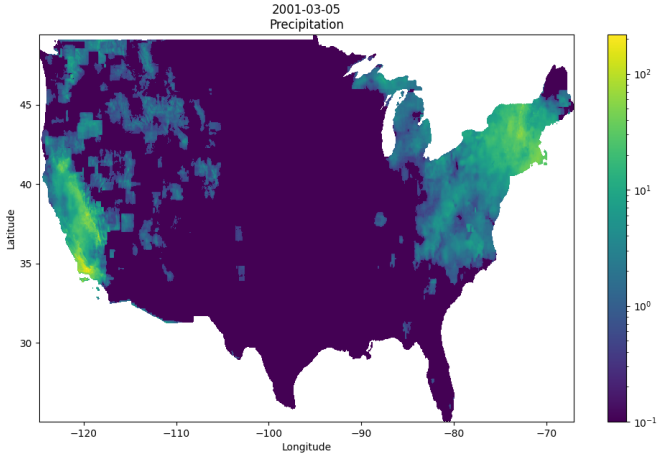

Fig. 18: Precipitation Graph with logarithmic scale

We see that Fig. 18 shows more details about the data. Together the two visualizations show some trends in the weather. We can see multiple waves of precipitation moving from the west to east for example between 19th and 28th January as seen in Fig. 19. This pattern can be observed across other dates and could be due to some weather phenomenon.

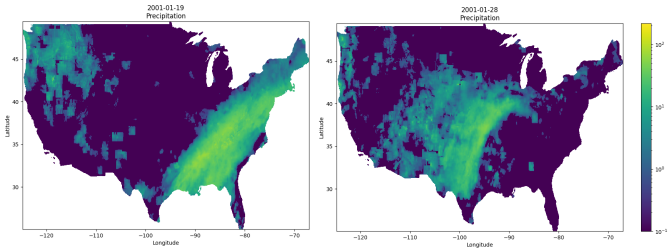

Fig. 19: Precipitation making its way across US from east to west

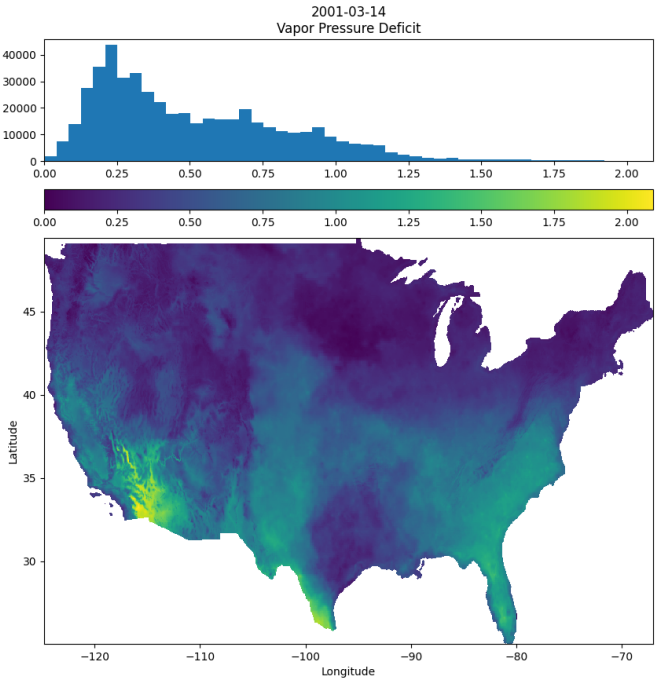

Fig. 20: VPD graph with histogram for distribution

Mean Vapour Pressure Deficit: : Vapor pressure deficit (vpd) refers to the difference in amount of water that can be held by the air and the amount of water that is actually held. Areas with high precipitation have a lower deficit as can be seen between Fig. 20 and Fig. 18 the vpd is measured 9 days after the precipitation and the lower values could be due to the rain.

Three colormaps were made from the data for this variable. Fig. 20 shows a colormap along with a histogram to show distribution of values. This distribution shows us that the values are not equally distributed. It might be helpful to highlight the percentile-wise extreme values to look at areas with extremely low and extremely high moisture. This is done by changing the colormap as seen in Fig. 21. There are two ways to find the extreme values:

- Percentile value globally
- Percentile value for that day

Both these approaches yield interesting results. Fig. 21 comes from the second approach. The same plot with the first

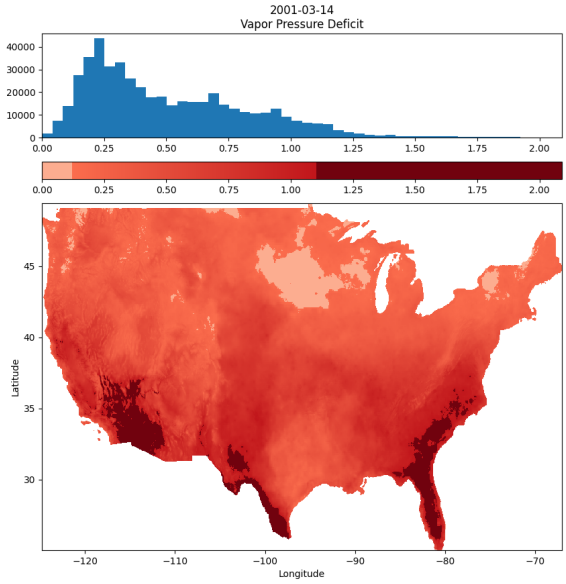

Fig. 21

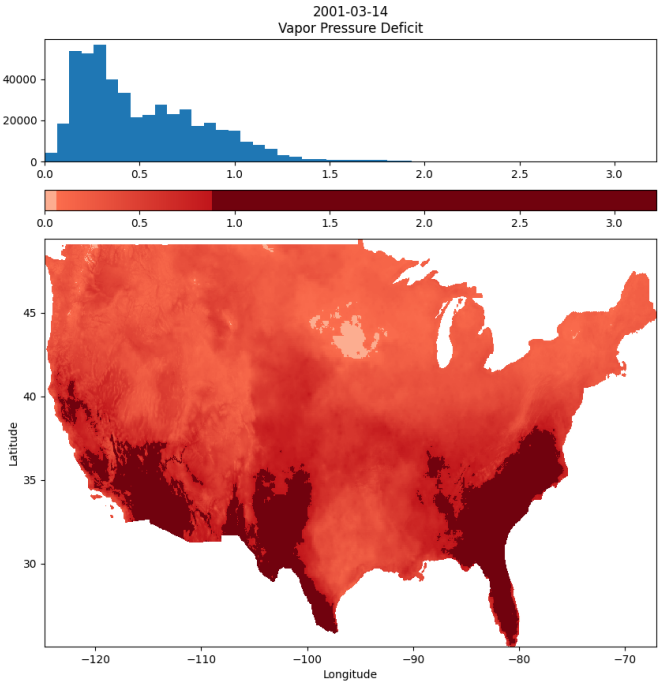

Fig. 22

approach is seen in Fig. 22. We see that the higher extreme is more present in Fig. x+5. This could be because it is a snapshot taken in mid March and there is less moisture when compared to the earlier months. Leading to a more than expected area to be coloured as an “extreme”. This shows the difference between local and global approaches. The local parameters are much more representative when comparing only for that

day. In case we want to check how a day compares to the rest of the sample data, a global approach would be preferred.

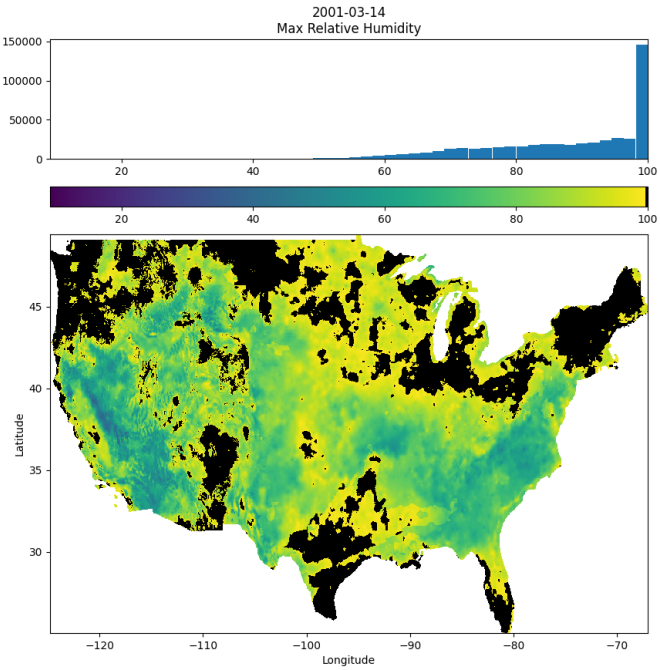

Fig. 23: Relative Humidit Max

Relative Humidity: : As the relative humidity seems to be similar to Mean Vapor Pressure deficit, we can expect to see some correlations in the maps. There are 2 variables given for Relative humidity - min and max for a specific day. Max value for a lot of areas throughout the 3 months is 100. There is also a huge difference in the number of locations with max as 100 and the others. Hence they've been colored black as seen in Fig. 23.

There are lesser black regions in this area as compared to early months as in Fig. 24 which supports our idea that the weather is drier in March.

When trying to correlate Fig. 23 to Fig. 21 we see that our initial expectation of vpd and relative humidity is incorrect. Multiple areas with 100 relative humidity are in the high extreme of the vpd which implies lesser moisture.

We now move to the minimum Relative humidity. We see that the distribution for the values of this is similar to a normal function. We can hence use a divergent colormap to differentiate between different ends of the distribution. This can be seen in Fig. 25.

The lack of blue highlights the general dryness in March. We also see a split in between the northern and southern parts of the country.

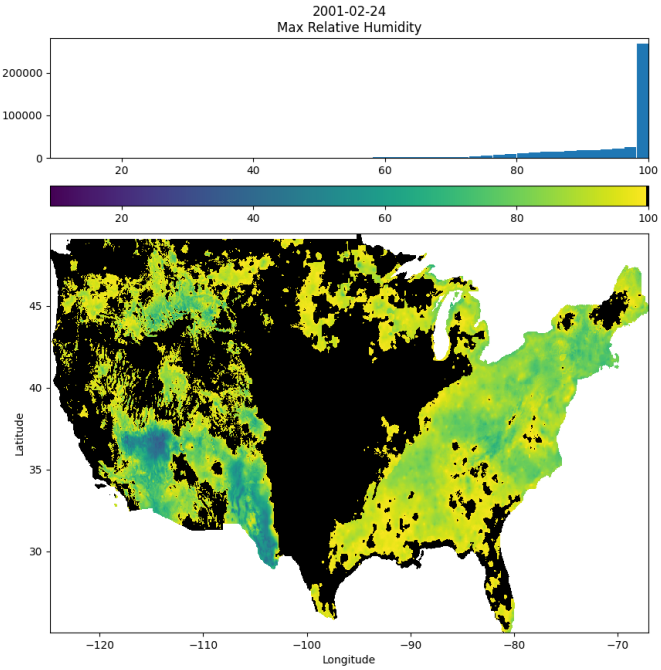

Fig. 24: Relive Humidity Max

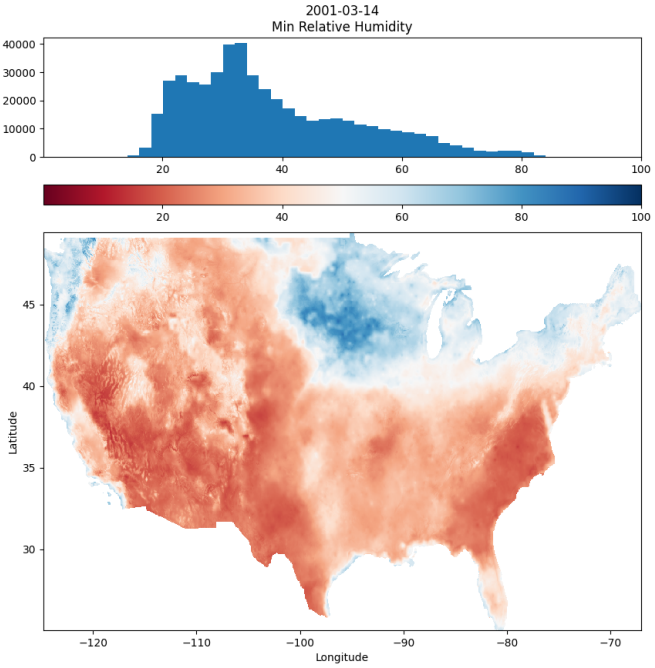

Fig. 25: Minimum Relative Humidity

III. Information Visualization

A. *Node-link Diagram*

Dataset Description: This dataset [4] encompasses the 1880-1881 School class friendships in a German school, and helps to reconstruct, to some extent, the friendships and social architecture among 53 boys in a German primary school. The data was collected by their teacher, Johannes Delitsch [1]. The sociometric research collects information on friendship perceptions, with directional ties illustrating mutual and non-reciprocal friendships. Some students took specific roles in the class: "repeaters" (those held back a grade who formed

a cohesive clique), and a ”sweets giver” (Lasch, who gained friends by distributing sweets).Such ideologies also exist in the data when for instance, there is mention of boys who were said to have some physical, psychological or even economic difficulties which led to their exclusion from the society. These illustrations help to understand how various characteristics influenced the formation of bonds between children.

Preprocessing: Since the dataset consists of only 53 data points, we did not have to perform any preprocessing on the dataset. Only while plotting the data using nodes and edges, the sizes and colour of the nodes and edges were changed to reveal useful information and enhance analysis.

We have used Gephi for all visualizations made in this section which provides a variety of graph layout algorithms and also parameters that can be tweaked to customize the node-link diagrams as per need. Gephi also provides tools that help us perform statistical analysis on the data. Some of the statistical outcomes utilized by us are listed as follows:

- Network diameter: It measures the greatest ditance needed to travel from one node to another, in terms of steps or edges. The value for this dataset is 8 as calculated by Gephi which tells us that the farthest a student is connected to another by friendship links is 8. showing moderate connectivity in the class.
- Average weighted degree: This value is a measure of how strong the connections are in a network. Lower values indicate that weak or lesser interactions while higher values indicate strong and frequent interactions. The value for this dataset is 3.377, which tells us that on an average a node is connected to 3-4 other nodes where the connections indicate friendships.
- Modularity: It measures the strength of community structures. A higher value indicates well formed clusters. Due to the small size of dataset and not many features we do not see well-defined clusters in the visualizations for this dataset. The modularity score was low at 0.2 in a scale of 0 to 1.

Plotting and analysis: Initially, on loading the data onto Gephi we get a cluttered visualization of the nodes and edges as shown in Fig. 26. We used different graph layout algorithms to organize the nodes and edges in a way that is more comprehensible and aesthetically pleasing. The different layout algorithms used and that will be discussed later are Fruchterman Reinggold, Yifan Hu. and Force Atlas. Before we move to the details of the algorithms, listed below are a few channels used in the node-link diagrams to encode and convey useful information:

- **Size**: Size has been used to encode information pertaining to the number of friends each student (here represented by nodes) has. The in-degree of the nodes have been used to do the same. Larger size of the node indicates a larger in-degree which in turn indicates a larger number of friends.
- **Colour**: Colour has been used to highlight certain characteristics (repeater, handicap, sweet-giver) about a student creating categories in the node-link diagram which makes it easier to perform characteristic-wise analysis on the data and better understand the reasoning behind the formation of friendships between students. Since no student in the dataset has more than one feature listed above, simple choice of colours helps us create a distinct categorization.

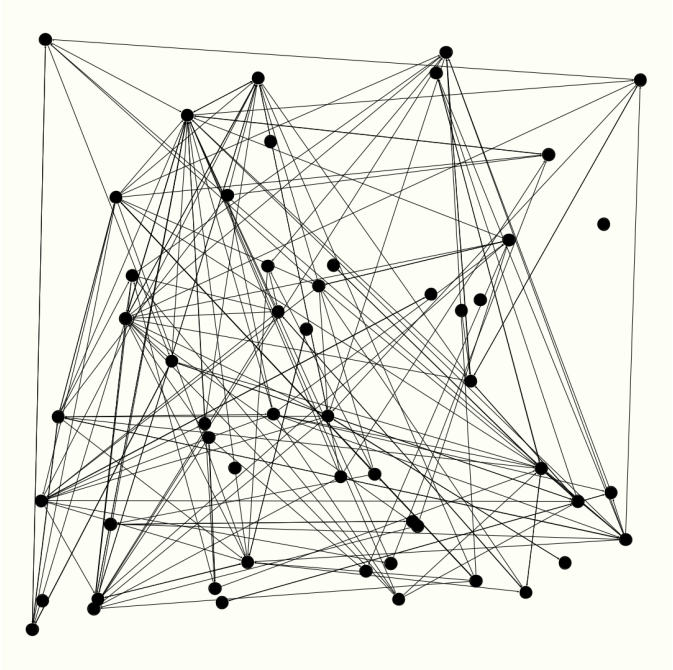

Fig. 26: Initial look at the network of nodes and edges.

Along with these labels have been added to the nodes with a characteristic feature.

Now we will discuss each layout algorithm and their features in detail.

- **Fruchterman Reingold**: The Fruchterman-Reingold algorithm works by simulating a physical system in which the nodes of a graph are treated as objects with electrical charges and the edges of the graph are treated as springs. The nodes are initially placed at random positions in the 2D or 3D space, and then the algorithm iteratively adjusts the position of the nodes based on the repulsion between the nodes and the attraction between the connected nodes. The nodes that are connected by an edge are pulled closer together, while the nodes that are not connected are pushed apart. It provides a balanced visualization with nodes evenly spread out as can be seen in Fig. 27 and Fig. 28.

  Parameters:

  – Area: This controls the layout space we want to allocate to the graph. Increasing its value helps prevent node overlap.
  – Gravity: This parameter helps control the dispersal of unconnected nodes by providing a pull toward the center of the layout.

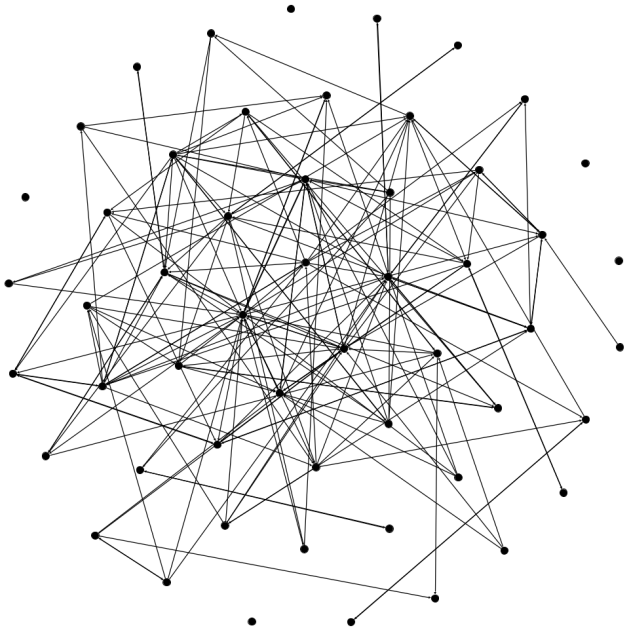

Fig. 27: Node-link diagram using Fruchterman Reingold layout algorithm.

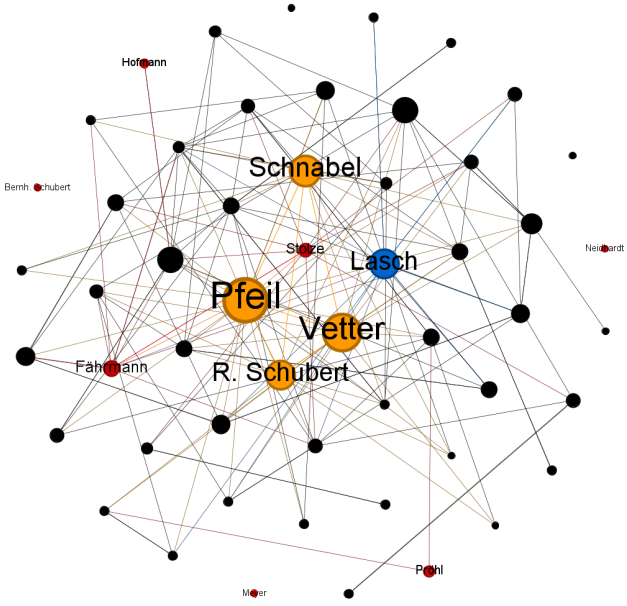

Fig. 28: Node-link diagram using Fruchterman Reingold layout algorithm along with size and colour channels.

– Speed: Determines how fast the nodes move to attain equilibrium. Higher speed can make the layout reach equilibrium faster but can also cause instability.

All of these parameters were adjusted to provide a favorable and aesthetic layout.

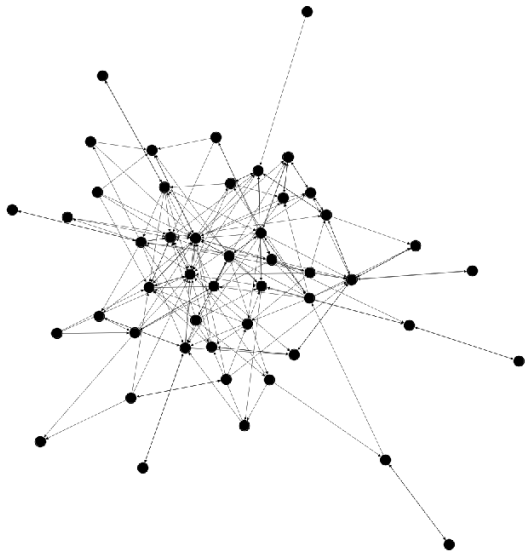

Fig. 29: Node-link diagram using Yifan Hu layout algorithm.

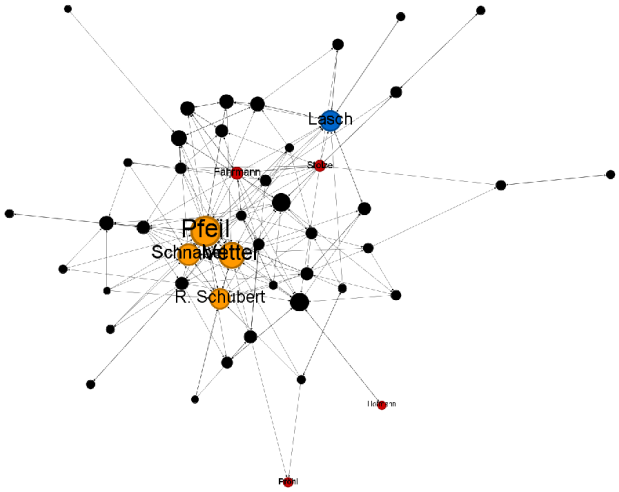

Fig. 30: Node-link diagram using Yifan Hu layout algorithm along with size and colour channels.

- **Yifan Hu**: The Yifan Hu layout algorithm belongs to the category of force-directed algorithms, which includes the Force Atlas and Fruchterman Reingold algorithms. It is generally used for large and dense networks due to which we don't see very impressive results for the dataset of our concern (Fig. 29 and Fig. 30). It has a multi-level approach wherein the network is first simplified and then expanded.

  Parameters:

  – Optimal Distance: This parameter specifies how closely or loosely the connected nodes should be placed.

- Relative Strength: This is used to control the relative force between the nodes; a higher value would result in an increase of repulsive force between the nodes.
- Adaptive cooling: enabling this dynamically adjusts the speed of the layout stabilization process.

To solve the clutter in the central region, we tried increasing the optimal distance, but that resulted in excessive scattering of the nodes that were already far from the center. Also the unconnected have been filtered out because they were too far away from the larger connected component.

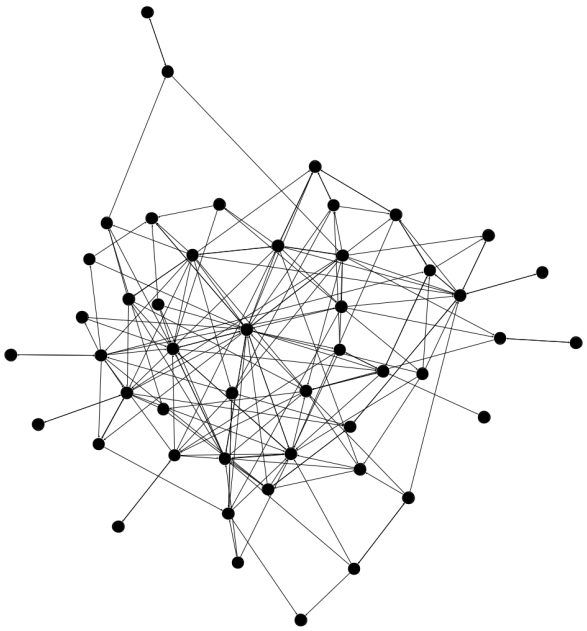

Fig. 31: Node-link diagram using Force Atlas layout algorithm.

- **Force atlas:** This is another force-directed algorithm that focuses cluster and community detection. Nodes are repelling entities and edges attract the nodes, as seen in Fruchterman Reingold, but unlike the latter, it includes specific parameters to handle node overlap, emphasize clusters etc. It was specially designed to highlight communities in the network. Our node-link diagram does not contain structured communities due to minimal data points as can be seen in Fig. 31 and Fig. 32. This layout algorithm is optimum for moderately sized networks and in cases where we want to study community clusters and relationships.
  Parameters:
  - Gravity and repulsive strength same as Fruchterman Riengold and Yifan Hu.
  - LinLog Mode: Adjusts the force model to promote better community separation, particularly for networks where clustering is important.

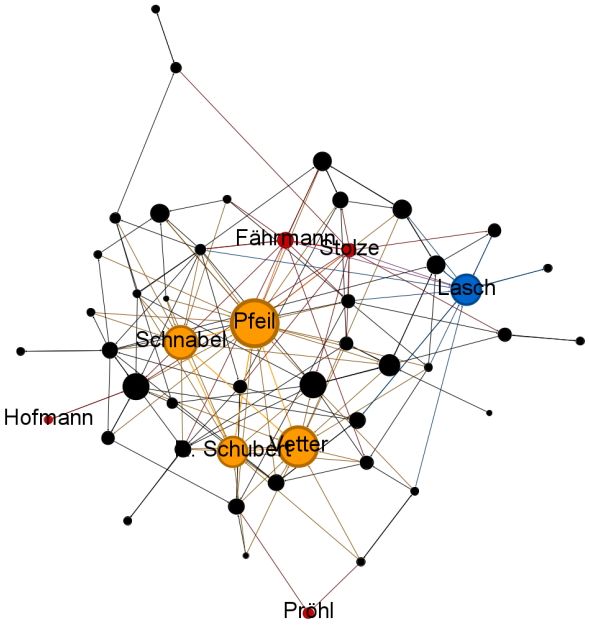

Fig. 32: Node-link diagram using Force Atlas layout algorithm along with size and colour channels.

  - Prevent Overlap: Ensures nodes don't overlap, which can improve readability, particularly in dense sections.

Inferences [1]: :

- Repeaters: Repeaters (e.g., Pfeil, Vetter) exhibit a high degree of connectedness and form a dense cluster in the node link diagram, thereby indicating their close-knit clique. A few incoming ties surrounding them are also quite visible, in all likelihood due to their games and other group activities, which in turn earned them social prominence.
- Sweets Giver (Lasch): The node representing Lasch is very busy as many directed ties lead towards him implying that he was quite good at making connections and this he achieved by distributing sweets to his classmates and gaining popularity.
- Handicapped Students: Illustrations of handicapped students (Meyer, Neidhardt) for example, are situated away from the central perspective, receive less incoming ties and do not have many connections to other nodes. This position indicates their marginality and failure to belong to the core or even intermediate groups which is perhaps due to rejection or lack of chances for interaction.
- General Trends: The overall diagram exhibits a number of clusters and has a pretty connected but sparse network. The central nodes depict the location of active social participants such as the repeaters while the outer or peripheral nodes depict participants with few connections or one-sided connections. This indicates that social interactions were more dependent on a certain characteristic

such as popularity or the ability to give rather than a group solidarity.

B. *Treemaps*

Dataset:

The same dataset from Assignment 1 *'USA Big City Crime Data'* [3] was utilized to perform this analysis.

Preprocessing:

The initial data pre-processing for this analysis was conducted in Assignment 1 using Tableau. Hence, the cleaned CSV file titled *Cleaned_LA_Dataset.csv* will be the basis for this task. Then, the Python script titled *Treemaps.py* was used for additional modifications of the treemaps in this report.

Interaction:

To make the data exploration and user interaction more engaging, the treemaps included interactive features.

- Hovering over a node reveals additional data specific to that node, enhancing the information available at a glance.
- The entire size of a node becomes easily readable by full screening the node through a mouse click.
- In the case of nested treemaps, the user is directed to a more specific level of an item by clicking on the node in it. It shows a focused treemap view of that segment only.

All these interactive features make the treemaps intuitive and easy to use, which results in smooth and fast exploration of the data and insight discovery.

Tools Used:

- **Plotly Python**: Has been applied for creating highly interactive and customizable treemaps within Python, thus it has been easy to integrate into data processing workflows.
- **Plotly.js**: It is used for the addition of client-side interactivity to the treemaps such that the users achieve smoother and more responsive experiences on the web visualizations.
- **HTML**: Used to structure the layout and present the treemaps in a web-friendly format, thus providing the base on which interactive visualizations are built.
- **Pandas**: The dataset was cleaned and preprocessed by Pandas.
- **Tableau**: Used for the initial pre-processing of data in Assignment 1.

Visualizations:

Fig.33, is a treemap depicting the Gender-wise Distribution of Crimes by Time of Day and Crime Seriousness. In this treemap, each gender (Male, Female, Unknown) forms a primary category, with time intervals as subcategories within each gender. Each time slot (e.g., 22:00) represents a 30-minute window (from 21:30 to 22:30), as modified in Python. The seriousness of crimes is further classified within each time slot, with "1" representing serious crimes and "2" representing non-serious crimes. The color corresponds to the victim's gender, and the size of each rectangle reflects the number of crimes committed. The spatial partitioning technique, set to 'total,' is a predefined Plotly setting that is used to plot the treemap.

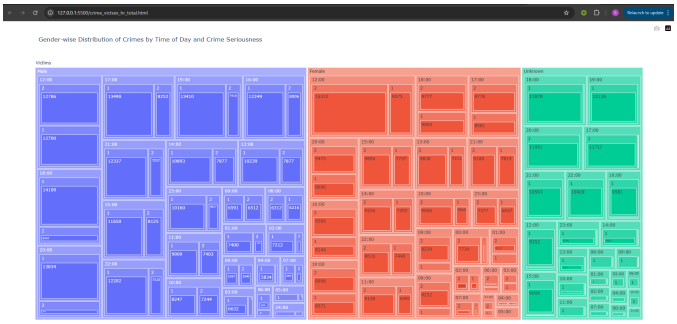

Fig. 33: Gender-wise Crime Distribution by Time of Day and Seriousness using the Total Technique (Plotly)

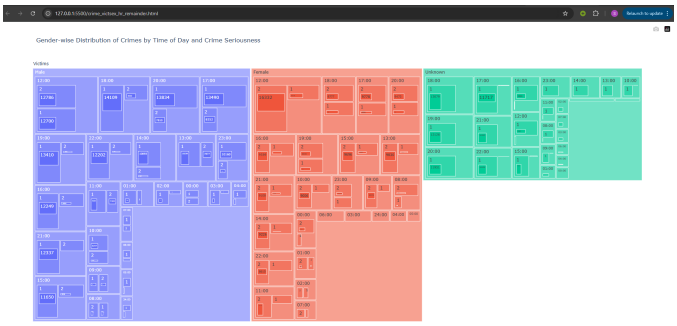

Fig. 34: Gender-wise Crime Distribution by Time of Day and Seriousness using the Remainder Technique (Plotly)

Fig.34, uses the remainder partitioning technique in Plotly to create the treemap. Unlike the total technique, where the value of each parent node is treated as the sum of all its descendant leaf nodes, in the remainder technique the value of each parent node represents only the extra part of its value that isn't captured by the sum of its child nodes.

Remainder technique is used when you want to highlight differences at each level of the hierarchy rather than just cumulative values.

Inferences:

The data reveals a notable spike in crime reports at 12:00, which may indicate potential data inaccuracies or a default time entry. Higher incident rates occur during late hours, while fewer crimes are reported in the earlier parts of the day.

This visualization, Fig.35, is a treemap illustrating the Gender-wise Distribution of Crimes by Premises Type and Crime Seriousness. Each gender (Male, Female, Unknown) serves as a primary category, with types of premises (e.g., Street, Single Family Dwelling, Parking Lot) acting as subcategories. The seriousness of crimes is further distinguished within each premises type, where "1" indicates serious crimes and "2" represents non-serious crimes. The color coding reflects the gender of the victims, while the size of each rectangle corresponds to

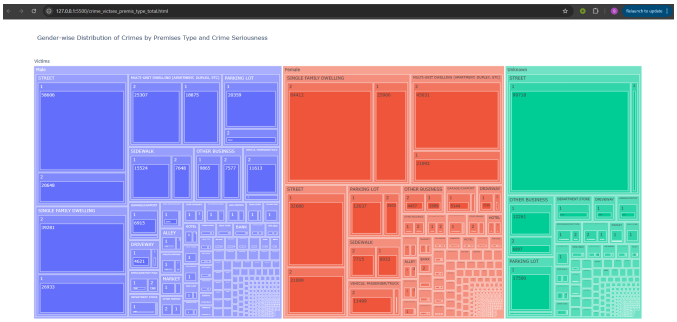

Fig. 35: Gender-wise Distribution of Crimes by Premises Type and Crime Seriousness using the Total Technique (Plotly)

the number of crimes committed. The 'total' partitioning technique is used to display the hierarchical relationships across the attributes.

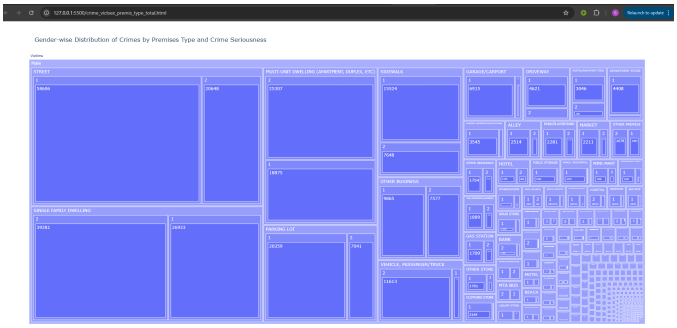

Fig. 36: Gender-wise Distribution of Crimes by Premises Type and Crime Seriousness using the Total Technique -Male Node (Plotly)

Fig.36 is what you will see once you click on the 'Male' node shown in Fig.35. You will notice that the rectangles automatically resize to reflect the selected level, ensuring that each segment remains clearly visible and proportional.

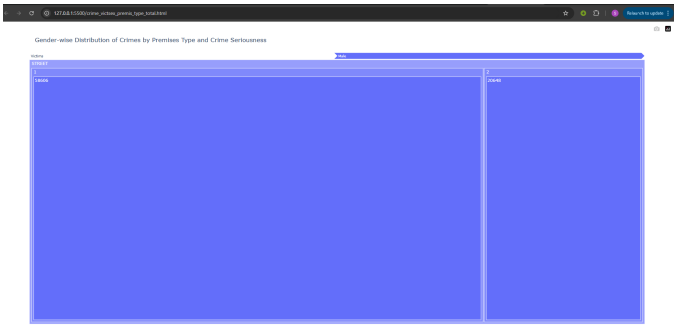

Fig. 37: Gender-wise Distribution of Crimes by Premises Type and Crime Seriousness using the Total Technique - Street Node (Plotly)

You will see Fig.37 after clicking the 'Street' node in Fig.36. In this manner, one can keep going through deeper levels of hierarchy and get a focused view of each segment.

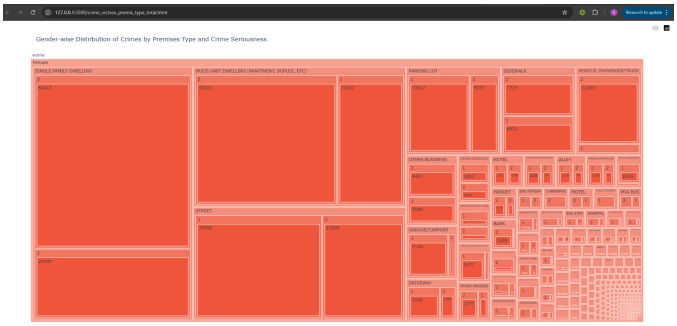

Fig. 38: Gender-wise Distribution of Crimes by Premises Type and Crime Seriousness using the Total Technique - Female Node (Plotly)

Similarly, by selecting the 'Female' node within the interactive treemap in Fig.35, users can zoom into this specific category and view Fig.38 in greater detail. The interactive features of Plotly enable a closer examination of specific segments within the treemap, allowing for a detailed analysis of particular crime characteristics based on user selection. This interactivity enhances the effectiveness of the visualization by enabling users to explore and focus on specific categories and subcategories as required.

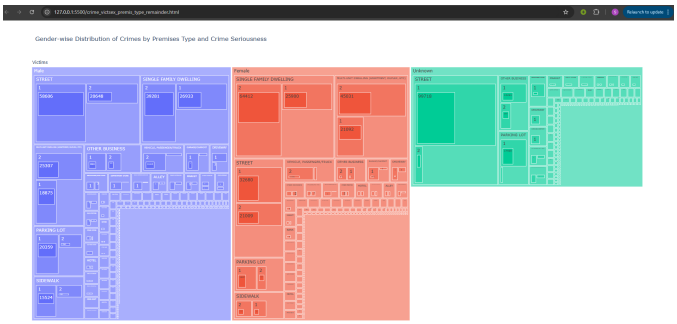

Fig. 39: Gender-wise Distribution of Crimes by Premises Type and Crime Seriousness using the Remainder Technique (Plotly)

In Fig.39, the same parameters as those in Fig.35 are used. However, instead of employing the 'total' spatial partitioning technique, the 'remainder' technique is utilized to construct the treemap.

Inferences: The data reveals that certain premises, such as streets, single-family dwellings, and multi-unit dwellings, have higher reported crime rates for both male and female victims. Additionally, the sparse reporting for the "unknown" gender category suggests that authorities are successfully identifying the victim's gender in most

cases. Within this category, a significant portion of incidents occur on streets, often involving serious crimes (indicated by ”1”). This aligns with the possibility that, in cases of serious crimes, victims may be less identifiable, contributing to the ”unknown” gender count.

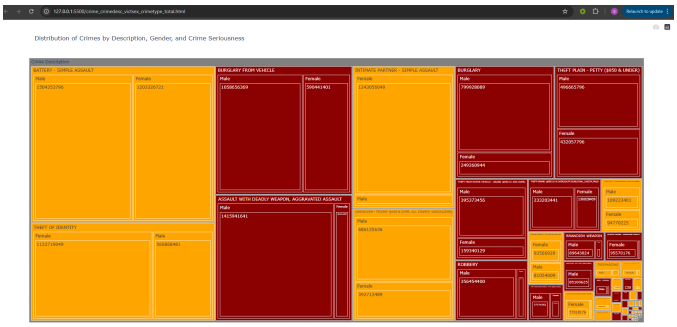

Fig. 40: Distribution of Crimes by Description, Gender, and Crime Seriousness using the Total Technique (Plotly)

This visualization, Fig.40 is a treemap showing the Distribution of Crimes by Description, Gender, and Crime Seriousness. Each type of crime, such as ”Battery - Simple Assault” or ”Burglary from Vehicle,” forms a primary category, with subcategories for the gender of the victim (Male, Female). The colors in the treemap indicate the seriousness of each crime: red represents serious crimes, while yellow denotes non-serious crimes. The size of each rectangle reflects the count of crimes within each category. The ‘Total’ spatial partitioning technique was used for the treemap.

Inferences:

The data shows that certain types of crimes, such as ”Battery - Simple Assault,” ”Theft of Identity,” and ”Intimate Partner - Simple Assault,” predominantly affect female victims, while crimes like ”Battery - Simple Assault” , ”Burglary from Vehicle” and “Assault With Deadly Weapon, Aggravated Assault” are more prevalent, particularly for male victims. The color scheme highlights that serious crimes (in red) are frequently reported. This treemap effectively highlights the types of crimes different genders experience, with specific focus areas for both male and female victims, while also reflecting crime severity across categories.

Fig.41 is a treemap showing the Distribution of Crimes by Victim Descent, Premises, Weapon Description, and Crime Seriousness. It categorizes crimes by victim descent (e.g., Hispanic/Latin/Mexican, Black, White), then by premises (e.g., Single Family Dwelling, Street), weapon type (e.g., Strong-Arm, Hand Gun), and crime seriousness (1 for serious, 2 for non-serious). Colors indicate victim descent, and rectangle sizes represent crime counts. This Plotly treemap uses ’total’ spatial partitioning to display hierarchical relationships across the attributes.

Inferences:

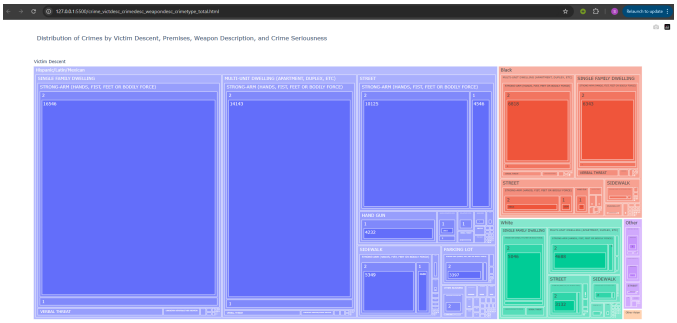

Fig. 41: Distribution of Crimes by Victim Descent, Premises, Weapon Description, and Crime Seriousnes using the Total Technique (Plotly)

The data highlights a notable concentration of crimes involving Hispanic/Latin/Mexican, White, and Black victims within specific premises, including single-family dwellings, multi-unit dwellings, and streets. In most of these instances, the weapon used is either unknown or there is no weapon involved. However, due to the high volume of such cases, they were excluded from the visualization to avoid sparsity. Among the remaining cases, the most prevalent weapon type is strong bodily force, indicating the use of hands, fists, or other physical means. There are a few instances where a handgun is used, which is particularly concerning as it poses a greater potential for serious harm.

Across all categories, the number of non-serious crimes (classified as Part 2 and labeled ”2” in the treemap) is substantially higher than the number of serious crimes (classified as Part 1 and labeled ”1”). This trend is a positive indicator, as it suggests that the majority of reported incidents are not of a highly serious nature.

C. Parallel Category Plots

In this assignment, we further go into the dataset analyzed in A1. In A1, using only conventional plots, we were only able to correlate between 2-3 variables at once. The dataset includes various details about the victims of crimes done in Los Angeles.

Finding a correlation between the various characteristics of victims could be useful for helping increase safety. Multiple variables like this can be correlated using a Parallel Category Plot(PCP). If it can be made interactive, it can lead to a lot of learning from the data. This is done using Plotly.

All the demographic related data is either categorical(Area Name, Part of crime, Victim Sex, Victim Descent) or can be made categorical by binning(Age). Once all this preprocessing is done, it can be converted to a html based PCP. Fig. 42 shows the plot.

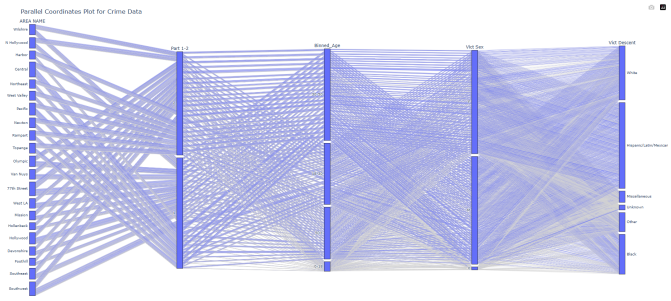

Fig. 42: Overall image of PCP

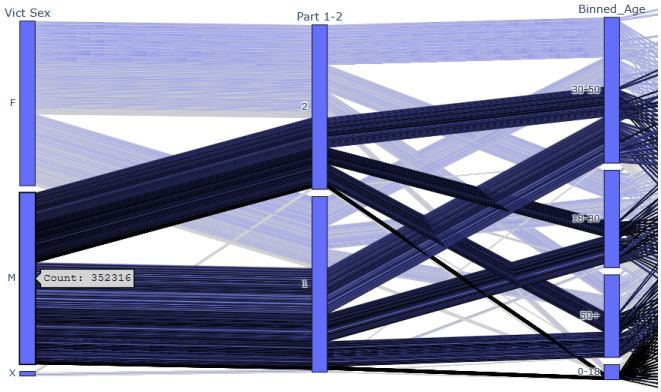

Fig. 43: Image highlighting all lines with Male victims, comparing with Part

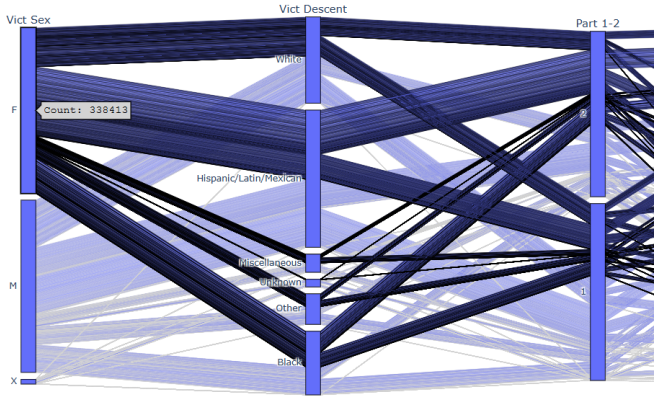

Fig. 44: Image highlighting all lines with Female victims, comparing with Descent

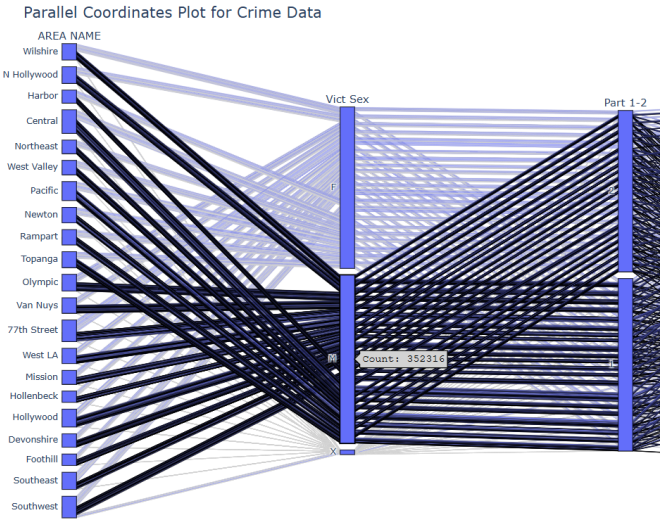

Fig. 45: Image highlighting all lines with Male victims, comparing with Area

Let us look at some correlations between various demographics.

In Fig. x+1 we can see the correlation between being Male and the Part of Crime and Age. Variables after that such as victim descent have chaotic connections and reordering needs to be done to make correlations.

Fig. x+1 shows more crimes against men are of Part 1 which are more aggressive crimes. We also see that most men that have Part 2 crimes against them, are between the age 30-50. This is also the case for Part 1 crimes. Hence we can conclude most Male victims are of age 30-50.

We can now rearrange the axes and compare Sex and Descent in Fig. 44. We see that the most frequent descent for woman victims is Hispanic/Latin/Mexican. We also see a proportionately large number of black descent.

Let us now look at different areas and how they're victims are distributed by Sex. in Fig. 45, we see that there are regions with a majority of men victims while some with a minority. Central for example seems to have more men. Harbor has more women as seen in the plot.

We can now look at how age ranges correspond to Part of crime. in Fig. 46 we can see most Part 1 victims are aged 30-50. This shows violent crime is most prevalent against those ages.

We see that multiple parameters can be related using PCP's and hence are great tools for data exploration.

REFERENCES

[1] Richard Heidler, Markus Gamper, Andreas Herz, and Florian Eßer. Relationship patterns in the 19th century: The friendship network in a german boys' school class from 1880 to 1881 revisited. *Social Networks*, 37:1–13, 05 2014.

[2] https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.contour.html.

[3] https://www.kaggle.com/datasets/middlehigh/los-angeles-crime-data-from-2000?resource=download&select=Crime_Data_from_2020_to_Present.csv.

[4] https://zenodo.org/records/4612153.Yk8HW25ueLp.

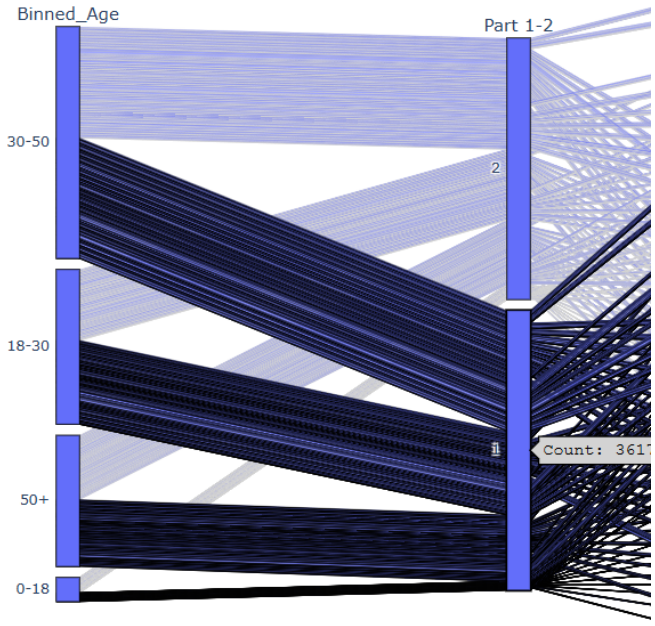

Fig. 46: Image highlighting all lines with Male victims, comparing with Area