CS6426 Data Visualization for Analytics Applications   Assignment 2

Attribute-based Visual exploration

Submitted By: Anasua Das   Student ID: 120220427 Year: March, 2022

----

1. INTRODUCTION:

Data visualization offers us to harness the huge amount of data that is created continuously all over the world by placing it in a visual context so that patterns, trends and correlations that might not otherwise be detected can be exposed. It helps us interpret available data; identify patterns, tendencies, and inconsistencies; make decisions, and analyze existing processes. Most of the times exploring and understanding the datasets become troublesome because of their high dimensionality where different Data visualization techniques come into rescue by creating easily understandable visual patterns from the complex data in an easily interpretable manner which would leave a strong pleasant impact on human brain. Finding hidden trends and patterns and meaningful insights out of complicated datasets is highly versatile a field of research and development which creates enormous opportunities for business intelligence in different industries, education and learning centers, in social network analysis, market research, healthcare developments, geospatial practices, genomic studies and where not.

We are already aware of different existing data visualization tools like Power BI, Tableau, Fusionchart, Infogram and their contribution to the field of data visualization, but in this article we are going to focus on different graphing libraries and packages offered by Python and how they are implemented on data visualization purpose.

Python being an open-source tool, it can be easily extended and is a great attraction for developers and data analysts. The ease of database connectivity, excellent scalability, robust libraries and user friendliness makes Python standout in competition.

A few versatile libraries offered by Python for data visualization purpose are: Bokeh, Geoplotlib, ggplot, Gleam, Leather, Matplotlib, missingno, Plotly , pygal, Seaborn.

We are going to discuss a few of them in details in the later part of this report.

For the data exploration purpose, following datasets have been used:

1. Election dataset from python plotly library.

2. Salaries dataset taken from github.

3. HDI.csv dataset taken from canvas (question)

4. Titanic dataset from github.

Libraries used:

1. Pandas, numpy, plotly, seaborn, matplotlib

Brief of tasks that have been performed in the process of building the report are given below:

2. Datasets are collected and explored.
3. Different visualizations such as: scatter plot, box plot, line graph, sunburst, Sankey Diagram, Tree map are performed on each of the datasets and the ones which best fits the respective data are saved for the report.
4. The patterns and specifications obtained in the visualization output are noted down.

2. DATA DESCRIPTION:

1. Election Dataset – plotly.express.data package offers the election() dataset where each record represents an electoral district in the 2013 Montreal mayoral election.[dataset1=px.data.election()]

Total no. of rows: 58

Total no. of columns: 8

[district', 'Coderre', 'Bergeron', 'Joly', 'total', 'winner', 'result','district_id']

2. **Salaries Dataset** – This dataset has been taken from github where each row describes rank, discipline, sex, phd , salary range and service tenure. [https://github.com/andvise/DataAnalyticsDatasets/blob/main/salaries.csv/]

Total no. of rows: 78

Total no. of columns: 6

['rank', 'discipline', 'phd', 'service', 'sex', 'salary']

3. '**hdi.csv'** dataset has been taken from the canvas uploaded version where each row holds human development index value from the year: 1990 to year: 2019 [From **Question**]

Total no. of rows: 189

Total no. of columns: 32

[**'Country_code', '1990', '1991', '1992', '1993', '1994', '1995', '1996', '1997', '1998', '1999', '2000', '2001', '2002', '2003', '2004', '2005', '2006', '2007', '2008', '2009', '2010', '2011', '2012', '2013', '2014', '2015', '2016', '2017', '2018', '2019', 'Country_name']**

4. **Titanic dataset** taken from github where each row describes the survival status of individual passengers on the Titanic.
[https://github.com/andvise/DataAnalyticsDatasets/blob/f4c1e07915ddfe98f0f5434ec3f0e7f3900f35ab/titanic.csv
]

Total no. of rows: 1000

Total no. of columns: 14

['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'Siblings/Spouses Aboard', 'Parents/Children Aboard', 'Fare']

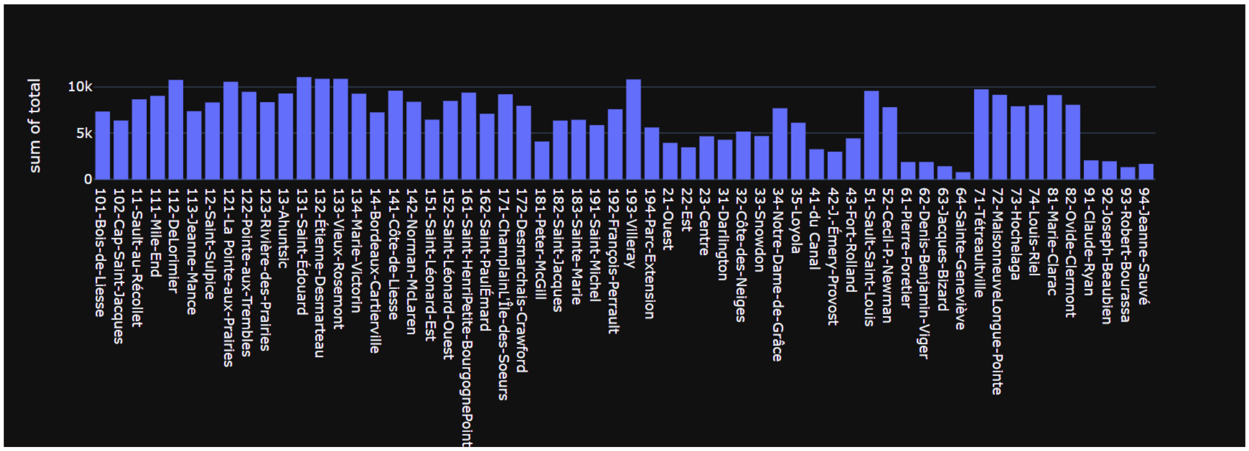
3. VISUAL OUTPUTS:

Dataset1: Election Dataset

For election dataset, two plots that the best explains the data are: **Histogram and bar plot**.

Histogram:

Histogram is plotted for total votes received for each of the district.



From this plot we can say, districts with the max vote received are: 131-Saint Edouard along with, 112-DeLorimier, 132-Etienne-Desmarteau while the district with the least vote received is: 64- Sainte-Genrevieve.

We can see that around four districts around 64- Sainte-Genrevieve consistently received less votes. Also from district code 91-94 vote received is quite low.

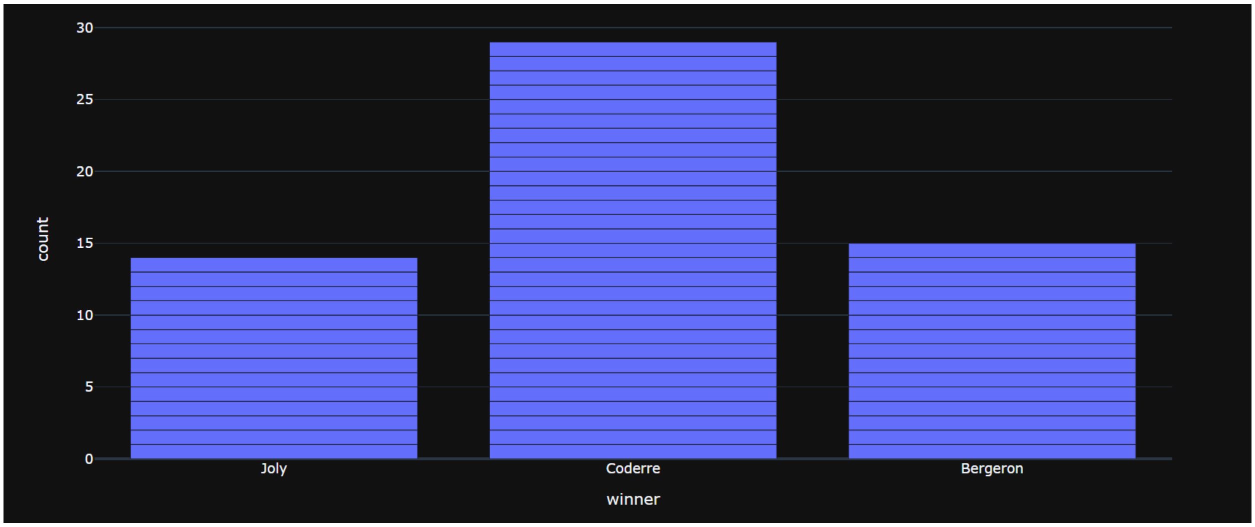
The frequency of the votes received is quite clear from this histogram plot. Based on the count of votes received we can try to understand about the political situation, population distribution . A lot of enhancements and decisions depend on these results.

Advantage of using Histogram:

- Histograms are simple and versatile.
- It is used for insightful look of frequency distribution.
- This is a major advantage for organizations to use histograms is it supports finding and dealing with process variation quickly.

Bar Plot:

Bar plot is used to see which candidate has received the majority of votes across all the districts.



From the bar plot, it can be said that Coderre has received the max of all votes receives across the districts whereas votes received for Joly is the least. There is a difference of around 1K votes by which Bergeron beat Joly. Coderre has won by a huge margin of the votes received.

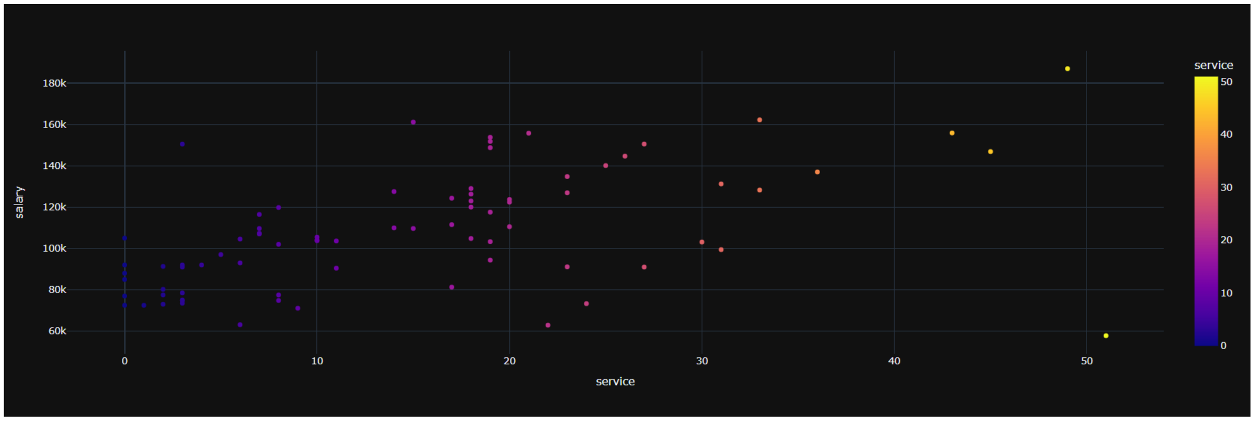
Advantage of using bar plot:

- Displays relative numbers or proportions of multiple categories.
- Summarize a large data set in visual form
- Estimate key values at a glance.
- Clarify trends better than do tables.
- It also shows changes over time.
-

Dataset2: Salaries dataset

For dataset2 , the following visualizations are chosen: **Scatter plot and Sunburst.**

Scatter Plot



In the x axis, **service period** has been taken and in the y-axis, **salary** range and a scatter plot is obtained.

We can see a pattern that overall as the service tenure increases, salary also increases although outliers are present.

As we see with service Tenure near to 50, the peak salary is above 180k

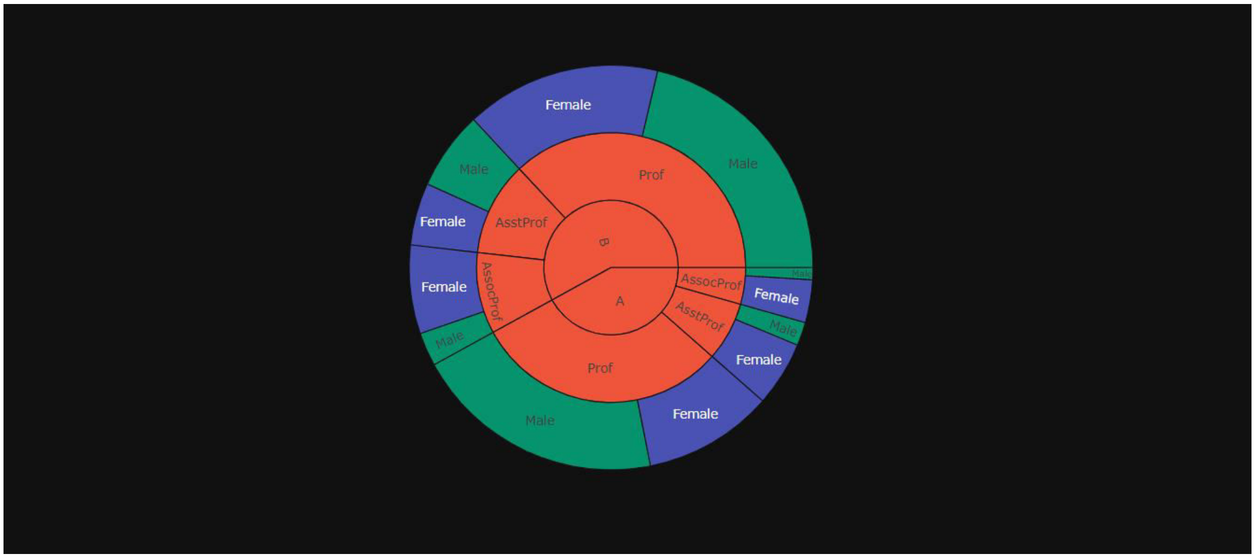
There are some distortions from the hypothesis that salary increases with higher value of service period which might be explained by some other factors playing at the back like rank or discipline.

This scatter plot shows a division in the plot with different colors represention different service tenure range.

Advantages of scatter plot:

- Shows relationship among the data.
- Shows all datapoints along with min and max outliers
- Retains exact data values and sample size.
- Shows both positive and negative type of graphical correlation.

Sunburst:



This plot shows association among discipline, rank and gender , each as subclass of their respective parent.

From the graph:

Level 1 parent or the root node is **discipline: A and B**

Child level of root node is **rank**: **Prof, AsstProf, AssocProf** ; this level is also the parent of third level.

The third level/ final level child is **Sex** : shows what part of male and female have become Professor, Associate Professor and Assistant professor

Hierarchy of the data explained above: **Discipline -> Rank -> Sex**

Discipline A: Professors are mostly male and comparatively smaller part is female.

While female portion in Associate and Assistant Professor rank is higher.

Discipline B: In the Professor rank, Male section is more. In the AsstProf rank male and female proprtion are almost equal while in the AssocProf rank Females are more than Males.

It has represented the higherarchy of the dataset pretty well.

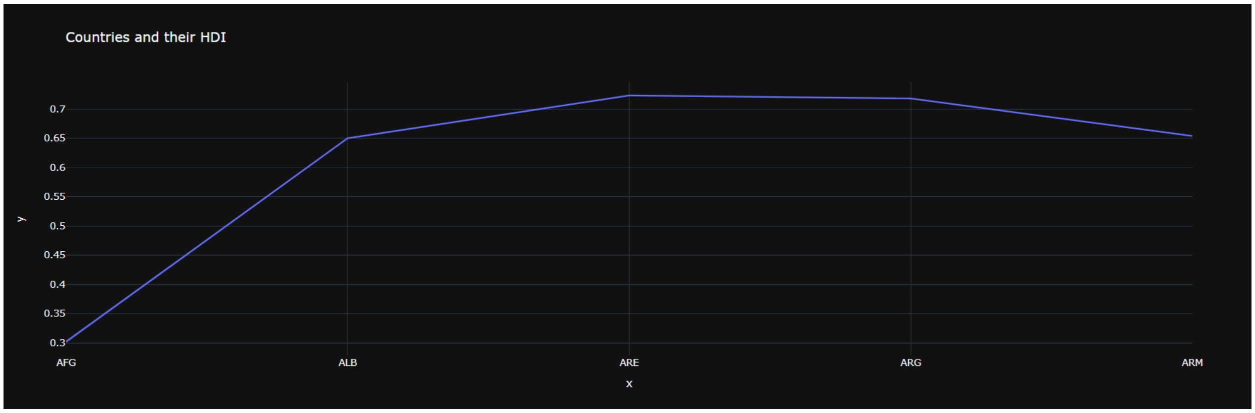
Advantages of Sunburst:

- It's represents hierarchy of data.
- As we expand outward from the center, we are going deeper in each tree
- The angle sweep corresponds to the value of each node
- Additional encodings possible like varying the width of the slice for additional dimensionality
- Color saturation can be used to define them as distinctly different.

Dataset 3:

For hdi.csv dataset we have taken: **Line plot and Scatter Plot.**

Line Plot:



The line graph shows HDI **Index of five countries in the year 1990**

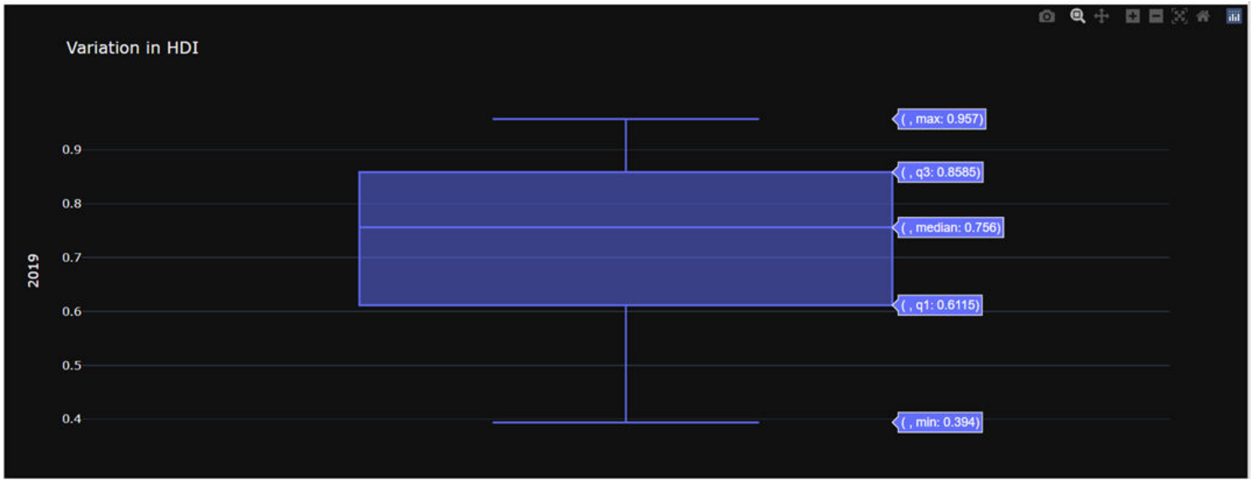
Country code is taken in x-axis and **HDI values in the year 1990 in y-axis**.

Country code out of these five countries with the highest HDI is ARE and the lowest HDI is AFG.

Advantages of Line Chart:

- It shows frequency of data along a number line.
- It is best to use a line plot when comparing fewer than 25 numbers. It is a quick, simple way to organize data.

Box Plot:



This box plot shows the variance in HDI value for all the countries in the year 2019.

Max HDI value is 0.957 while the minimum HDI value lies at 0.394 and the central value: median of all the HDIs is 0.756.

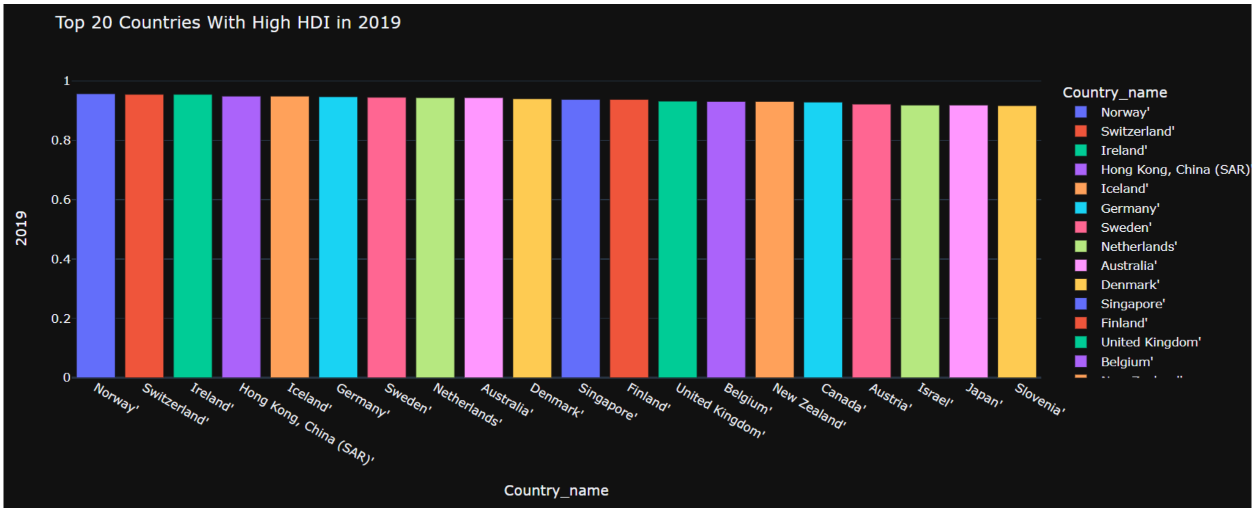
The data is not normally distributed and skewed.

Advantages of Box plot:

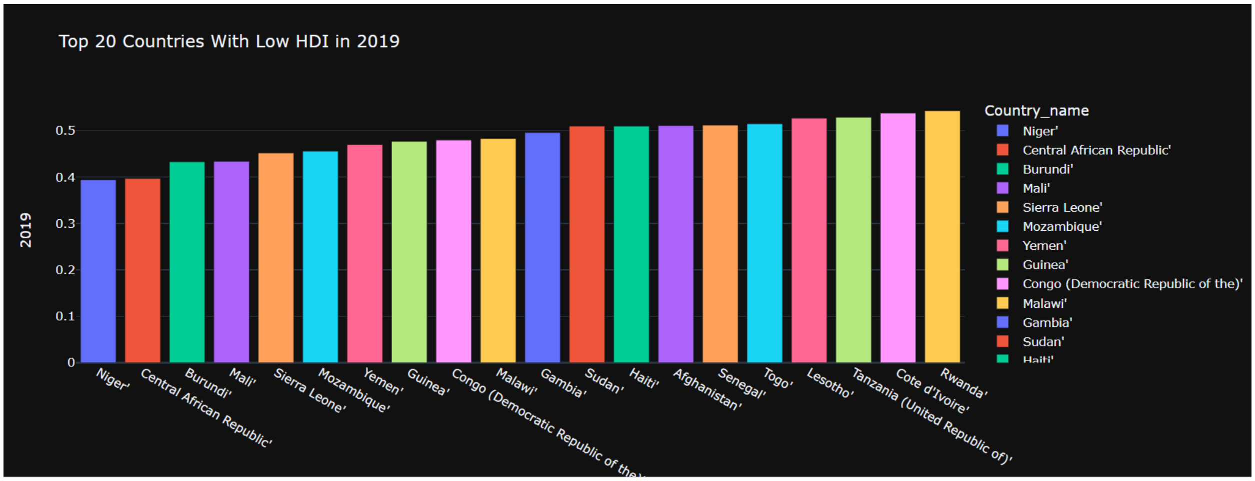
- Summarizes variation in large datasets visually.
- Shows outliers.
- Compares multiple distributions.
- Indicates symmetry and skewness to a degree.
- Simplicity is the greatest attraction.

Bar Plot:

This is another way of representing **Countries along with their HDIs in a bar plot** and each case is pretty easily distinguishable.



Here the top 20 countries with the highest HDI values in the year 2019 are plotted. To name: Norway, Switzerland, Ireland, Hong Kong,China, Iceland, Germany and so on.



Here the top 20 countries with the lowest HDI values in the year 2019 are plotted. Those are: Nigeria, Central African Republic, Burundi, Mali and so on.

Each country is coded with a unique color that is represented in the legend.

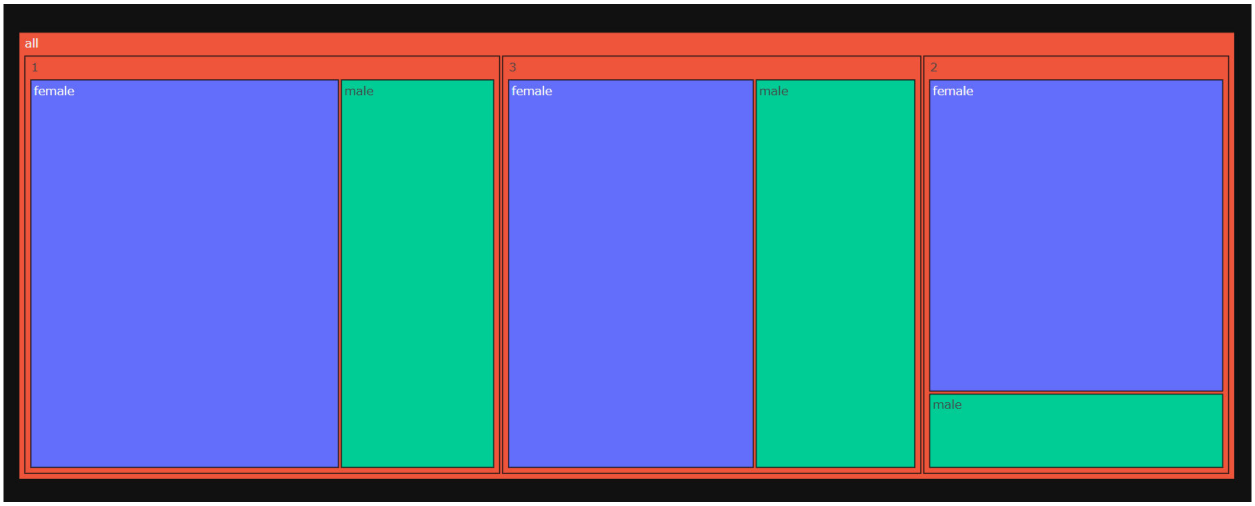
From the plot:

Nigeria with Country_code='NER' has the lowest HDI =0.394 in the year 2019 country with Country: Norway with Country_code=NOR has the highest HDI =0.957 in the year 2019

Dataset4:

For Titanic dataset two visualizations are chosen as: **Treemap and Sankey Diagram**

Treemap:



Here the entire constant space –'all' is first divided into 3 sections for Pclass1, Pclass2, Pclass3 coded as 1,2 and 3 respectively.

Each of these sections further breaks down in to two sections representing the Sex: Male or female.

Hence the plot clearly shows survival rate of male and female from the three classes.

Pclass1 and Plclass3 show the most of survival rate whereas boarders from Pclass3 mostly died.

Total survived: 341

Survived from Pclass1 is: 135

Male: 45; Female: 90

Survived from Pclass2 is: 87

Male: 17, Female: 70

Survived from Pclass3: 119

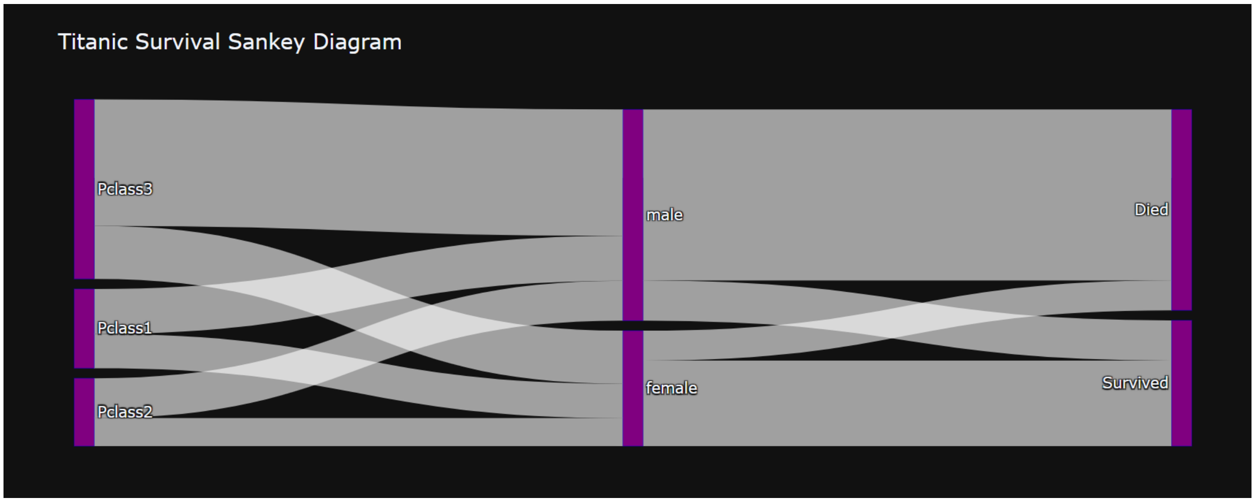
Male: 47, Female: 72

In three of the classes, survival rate of females are more than survival rate of males.

Advantage of Tree Map:

- Identify the relationship between two elements in a hierarchical data structure
- optimize the use of space
- accurately display multiple elements together
- show ratios of each part to the whole
- visualize attributes by size and color coding

Sankey Diagram:



Sankey diagrams are a type of flow diagram in which the width of the arrows is proportional to the flow rate. Sankey diagrams can also visualize the energy accounts, material flow accounts on a regional or national level, and cost breakdowns.

From the Sankey diagram it can be told:

Total Died: 545:

Male: 464; Female: 81

Total Survived: 341

Male: 109; Female: 232

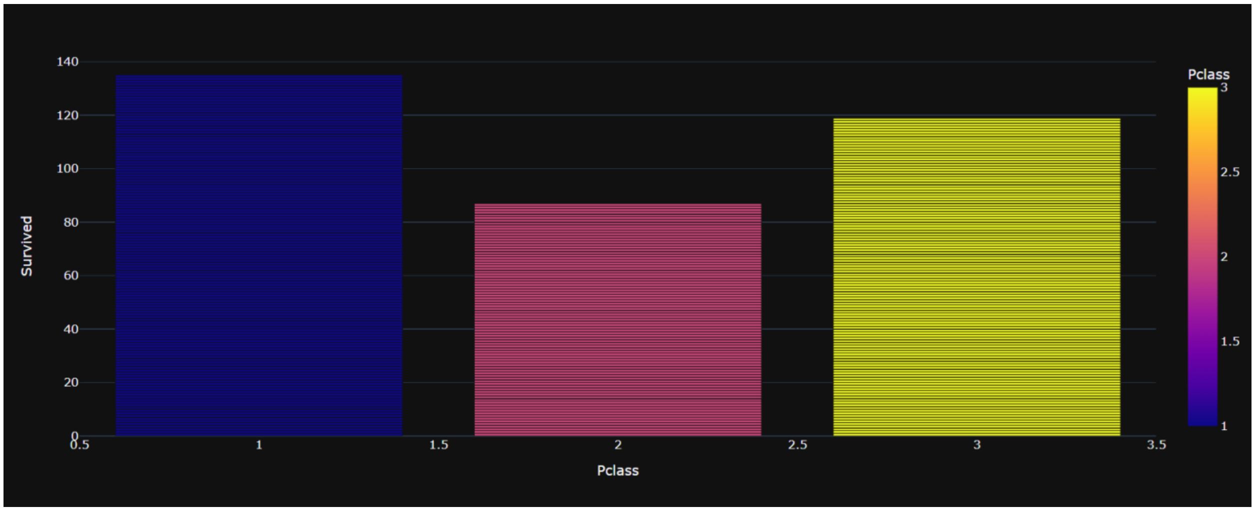
In Pclass1: Total Male: 122 Female: 93

In Pclass2: Total Male: 108 Female: 76

In Pclass3: Total Male: 343 Female: 144

Bar plot:

The bar plot shows survival rate based on Pclass:



From the plot:

Pclass1 has the maximum survival followed by Pclass3 and lastly Pclass2

Survived from Pclass1= 135

Survived from Pclass2= 97

Survived from Pclass3 = 119

4. Summary table and Discussion

Summary Of Observations				
Index	**Data Set**	**Visualizations**	**Data Observations**	**Utility of Visualizations**
1	Election Dataset	Histogram	Districts with highest votes received are: 131-Saint Edouard along with, 112-DeLorimier, 132-Etienne-Desmarteau with more than 10K votes. District with lowest votes received: 64- Sainte-Genrevieve with around 1K votes.	• Histograms are simple and versatile. • It is used for insightful look of frequency distribution. • This is a major advantage for organizations to use histograms is it supports finding and dealing with process variation quickly. • It has its use in transportation industry, superstores, urbanization process etc.

		Bar Plot	Highest vote receiver contestant is: Coderre . Joly receives the least number of votes. Bergeron beats Joly by 1K votes	• Bar chart representation helps in visualizing numerical variables / numerical and categorical variable in a dataset. Displays relative numbers or proportions of multiple categories by using rectangular bars with heights or lengths proportional to the values they represent. • summarize a large data set in visual form • estimate key values at a glance. • clarify trends better than do tables.
2	Salaries Dataset	Scatter Plot	We can see, that overall there's a tendency for salary to increase if the service duration increases. There are cases of deviations and outliers.	• Shows relationship among the data. • Shows all datapoints along with min and max outliers • Retains exact data values and sample size. • Shows both positive and negative type of graphical correlation.
		Sunburst Plot	Discipline A: Professors are mostly male and comparatively smaller part is female. While female portion in Associate and Assistant Professor rank is higher. Discipline B: In the Professor rank, Male section is more. In the AsstProf rank male and female proprtion are almost equal while in the AssocProf rank Females are more than Males.	• It’s represents hierarchy of data. • As we expand outward from the center, we are going deeper in each tree • The angle sweep corresponds to the value of each node. • Additional encodings possible like varying the width of the slice for additional dimensionality color saturation can be used to define them as distinctly different.
3	HDI Dataset	Line Plot	The line graph shows HDI Index of five countries in the year 1990 Country code out of these five countries with the highest HDI is ARE HDI is AFG	• It shows frequency of data along a number line. It is best to use a line plot when comparing fewer than 25 numbers. • It is a quick, a simple way to organize data.

		Box plot	This box plot shows the variance in HDI value for all the countries in the year 2019. Max HDI value is 0.957 while the minimum HDI value lies at 0.394 and the central value: median of all the HDIs is 0.756. The data is not normally distributed and skewed.	• A box plot is a graph that gives you a good indication of how the values in the data are spread out. • It provides a visual summary of the data enabling researchers to quickly identify mean values, the dispersion of the data set, and signs of skewness.
		Bar Plot	All the countries along with their HDIs in a bar plot and each case is pretty distinguishable manner in a single frame. Top 5 countries with highest Index in 2019 are Norway, Switzerland, Ireland, Hong Kong, China, Iceland . Top 5 countries with the lowest HDI in 2019 are Nigeria, Central African Republic, Burundi and Mali.	Discussed for Election dataset.
4	Titanic Dataset	Tree map	Here the entire constant space – 'all' is first divided into 3 sections for Pclass1, Pclass2, Pclass3 coded as 1,2 and 3 respectively. Each of these sections further breaks down in to two sections representing the Sex: Male or female. Hence the plot clearly shows survival rate of male and female from the three classes. Pclass1 and Plclass3 show the most of survival rate whereas boarders from Pclass3 mostly died. Total survived: 341 Survived from Pclass1 is: 135 Male: 45; Female: 90 Survived from Pclass2 is: 87 Male: 17, Female: 70 Survived from Pclass3: 119 Male: 47, Female: 72	• Tree maps are best utilized in an interactive format, in which a user • Can drill deeper into the various categories and subcategories of interest and enlarge and reduce the size of the visual display (and the amount of detail revealed) as desired. • Tree maps can be several layers to dozens of layers deep and can allow drill-down to hundreds of sub-categories Properly represents hierarchy of data.
		Sankey Diagram	From the Sankey diagram it can be told: Survival of Female boarders is more than male boarders of Titanic. Pclass1 shows most of the survival portion followed by Pclass3 and then Pclass2	• Sankey diagram allows representation of complex processes with a focus on a single aspect or resource that we want to focus on. It is used to depict a flow

			Total Died: 545: Male: 464; Female: 81 Total Survived: 341 Male: 109; Female: 232 In Pclass1: Total Male: 122 Female: 93 In Pclass2: Total Male: 108 Female: 76 In Pclass3: Total Male: 343 Female: 144	from one set of values to another. The things being connected are called nodes and the connections are called links. • Sankey diagram can be used in improving organizational infrastructure, sales, business, super marts etc. • It is very useful in determining energy flows at power plants, incoming and outgoing flows in retail and logistics industry.
		Bar plot	Pclass1 has the maximum survival followed by Pclass3 and lastly Pclass2 Survived from Pclass1= 135 Survived from Pclass2= 97 Survived from Pclass3 = 119	Discussed for Election Dataset.

5. Conclusions

The Election dataset was examined carefully using Histogram and Bar chart. Using these two visualizations it can be inferred that which helped to find out meaningful frequency distribution about the count of votes received across the districts. It helped with a colorful and summarized visualization of who received maximum votes across the districts and won the election and also which districts saw a good example of active voting and which districts failed to achieve satisfactory election results.

For the salaries dataset, an existing correlation between salary scale and service period is found out using a scatter plot. On the other hand, drilling down to every layer of the dataset attributes to find hierarchy among the features using the sunburst plot was quite fascinating. The association of those interesting features in a definite hierarchical manner reveals interesting patterns and trends in the data. Overall, salary is highly influenced by factors : educational qualification(phd) and years of service.

From HDI dataset using Line chart, Box plot and Bar chart, it could actually be found out how a certain country performed in terms of human development for each year from 1990-2019. Norway looks pretty good in terms of human development whereas regions in Africa did not perform well, Nigeria at the lowest. Asian regions all over performed well with a slightly higher index than the average. The median lies at 0.756.

Finally Titanic dataset is explored using Sankey diagram and Tree map which revealed the statistics of survival from the Titanic tragedy based on the Class the boarders were travelling, their gender and age. Overall, Females survived more than male boarders while boarders from Pclass1 and Pclass3 survived more than Pclass2.

From the visualization results obtained above with the four datasets, we can reach to an understanding that line plots and bar plots are more suitable for time series data or observe patterns among numerical variables. Box plot is used to understand the spread of data and find out if skewed or normal. Scatter plots are used to find out correlation among various features within a dataset. Sankey Diagram is very effective in representing the relationship between different variables in the feature space and depicting a flow from one set of values to another to find hidden patterns.

6. REFERENCES

- Plotly Open Source Graphing Libraries
- https://venngage.com/blog/types-of-diagrams/
- https://github.com/andvise/DataAnalyticsDatasets
- https://www.analyticsvidhya.com/blog/2021/11/visualize-data-using-sankey-diagram/
- pandas.DataFrame — pandas 1.4.1 documentation (pydata.org)
- Matplotlib — Visualization with Python
- plotly.express.bar — 5.6.0 documentation
- Data Visualizing in Python. In this article, I will be focusing on… | by Rashmi Duleesha | LinkIT | Medium
- https://gilberttanner.com/blog/introduction-to-data-visualization-inpython
- https://www.ifu.com/e-sankey/sankey-diagram/