Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Data is a crucial component in the field of Machine Learning. It refers to the set of observations or measurements that can be used to train a machine-learning model. The quality and quantity of data available for training and testing play a significant role in determining the performance of a machine-learning model. Data can be in various forms such as numerical, categorical, or time-series data, and can come from various sources such as databases, spreadsheets, or APIs. Machine learning algorithms use data to learn patterns and relationships between input variables and target outputs, which can then be used for prediction or classification tasks.
Data is typically divided into two types:

1. Labeled data
2. Unlabeled data

Labeled data includes a label or target variable that the model is trying to predict, whereas unlabeled data does not include a label or target variable. The data used in machine learning is typically numerical or categorical. Numerical data includes values that can be ordered and measured, such as age or income. Categorical data includes values that represent categories, such as gender or type of fruit.

Data can be divided into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate the performance of the model. It is important to ensure that the data is split in a random and representative way. Data preprocessing is an important step in the machine learning pipeline. This step can include cleaning and normalizing the data, handling missing values, and feature selection or engineering.

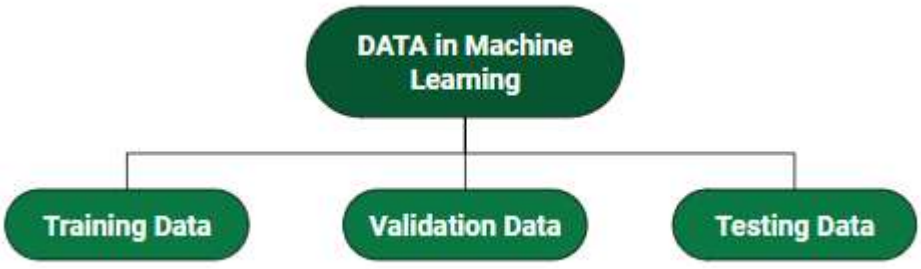
DATA: It can be any unprocessed fact, value, text, sound, or picture that is not being interpreted and analyzed. Data is the most important part of all Data Analytics, Machine Learning, and Artificial Intelligence. Without data, we can't train any model and all modern research and automation will go in vain. Big Enterprises are spending lots of money just to gather as much certain data as possible.
Example: Why did Facebook acquire WhatsApp by paying a huge price of $19 billion?
The answer is very simple and logical – it is to have access to the users' information that Facebook may not have but WhatsApp will have. This information about their users is of paramount importance to Facebook as it will facilitate the task of improvement in their services.

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

INFORMATION: Data that has been interpreted and manipulated and has now some meaningful inference for the users.
KNOWLEDGE: Combination of inferred information, experiences, learning, and insights. Results in awareness or concept building for an individual or organization.

How do we split data in Machine Learning?

- **Training Data:** The part of data we use to train our model. This is the data that your model actually sees(both input and output) and learns from.
- **Validation Data:** The part of data that is used to do a frequent evaluation of the model, fit on the training dataset along with improving involved hyperparameters (initially set parameters before the model begins learning). This data plays its part when the model is actually training.
- **Testing Data:** Once our model is completely trained, testing data provides an unbiased evaluation. When we feed in the inputs of Testing data, our model will predict some values(without seeing actual output). After prediction, we evaluate our model by comparing it with the actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.



Consider an example:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

There's a Shopping Mart Owner who conducted a survey for which he has a long list of questions and answers that he had asked from the customers, this list of questions and answers is **DATA**. Now every time when he wants to infer anything and can't just go through each and every question of thousands of customers to find something relevant as it would be time-consuming and not helpful. In order to reduce this overhead and time wastage and to make work easier, data is manipulated through software, calculations, graphs, etc. as per your own convenience, this inference from manipulated data is **Information**. So, Data is a must for Information.
Now **Knowledge** has its role in differentiating between two individuals having the same information. Knowledge is actually not technical content but is linked to the human thought process.

Different Forms of Data

- **Numeric Data** : If a feature represents a characteristic measured in numbers , it is called a numeric feature.
- **Categorical Data :** A categorical feature is an attribute that can take on one of the limited , and usually fixed number of possible values on the basis of some qualitative property . A categorical feature is also called a nominal feature.
- **Ordinal Data** : This denotes a nominal variable with categories falling in an ordered list . Examples include clothing sizes such as small, medium , and large , or a measurement of customer satisfaction on a scale from "not at all happy" to "very happy".

Properties of Data –

1. **Volume:** Scale of Data. With the growing world population and technology at exposure, huge data is being generated each and every millisecond.
2. **Variety:** Different forms of data – healthcare, images, videos, audio clippings.
3. **Velocity:** Rate of data streaming and generation.
4. **Value:** Meaningfulness of data in terms of information that researchers can infer from it.
5. **Veracity:** Certainty and correctness in data we are working on.

6. **Viability**: The ability of data to be used and integrated into different systems and processes.
7. **Security**: The measures taken to protect data from unauthorized access or manipulation.
8. **Accessibility**: The ease of obtaining and utilizing data for decision-making purposes.
9. **Integrity**: The accuracy and completeness of data over its entire lifecycle.
10. **Usability**: The ease of use and interpretability of data for end-users.

Some facts about Data:

- As compared to 2005, 300 times i.e. 40 Zettabytes (1ZB=10^21 bytes) of data will be generated by 2020.
- By 2011, the healthcare sector has a data of 161 Billion Gigabytes
- 400 Million tweets are sent by about 200 million active users per day
- Each month, more than 4 billion hours of video streaming is done by the users.
- 30 Billion different types of content are shared every month by the user.
- It is reported that about 27% of data is inaccurate and so 1 in 3 business idealists or leaders don't trust the information on which they are making decisions.

The above-mentioned facts are just a glimpse of the actually existing huge data statistics. When we talk in terms of real-world scenarios, the size of data currently presents and is getting generated each and every moment is beyond our mental horizons to imagine.

Example:

Imagine you're working for a car manufacturing company and you want to build a model that can predict the fuel efficiency of a car based on the weight and the engine size. In this case, the target variable (or label) is the fuel efficiency, and the features (or input variables) are the weight and engine size. You will collect data from different car models, with corresponding weight and engine size, and their fuel efficiency. This data is labeled and it's in the form of (weight,engine size,fuel efficiency) for each car. After having your data ready, you will then split it into two sets: training set and testing set, the training set will be used to train the model and the testing set will be

used to evaluate the performance of the model. Preprocessing could be needed for example, to fill missing values or handle outliers that might affect your model accuracy.

Advantages Or Disadvantages:

Advantages of using data in Machine Learning:

1. Improved accuracy: With large amounts of data, machine learning algorithms can learn more complex relationships between inputs and outputs, leading to improved accuracy in predictions and classifications.
2. Automation: Machine learning models can automate decision-making processes and can perform repetitive tasks more efficiently and accurately than humans.
3. Personalization: With the use of data, machine learning algorithms can personalize experiences for individual users, leading to increased user satisfaction.
4. Cost savings: Automation through machine learning can result in cost savings for businesses by reducing the need for manual labor and increasing efficiency.

Disadvantages of using data in Machine Learning:

1. Bias: Data used for training machine learning models can be biased, leading to biased predictions and classifications.
2. Privacy: Collection and storage of data for machine learning can raise privacy concerns and can lead to security risks if the data is not properly secured.
3. Quality of data: The quality of data used for training machine learning models is critical to the performance of the model. Poor quality data can lead to inaccurate predictions and classifications.
4. Lack of interpretability: Some machine learning models can be complex and difficult to interpret, making it challenging to understand how they are making decisions.

Use of Machine Learning :

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Machine learning is a powerful tool that can be used in a wide range of applications. Here are some of the most common uses of machine learning:

- **Predictive modeling:** Machine learning can be used to build predictive models that can predict future outcomes based on historical data. This can be used in many applications, such as stock market prediction, fraud detection, weather forecasting, and customer behavior prediction.
- **Image recognition:** Machine learning can be used to train models that can recognize objects, faces, and other patterns in images. This is used in many applications, such as self-driving cars, facial recognition systems, and medical image analysis.
- **Natural language processing:** Machine learning can be used to analyze and understand natural language, which is used in many applications, such as chatbots, voice assistants, and sentiment analysis.
- **Recommendation systems:** Machine learning can be used to build recommendation systems that can suggest products, services, or content to users based on their past behavior or preferences.
- **Data analysis:** Machine learning can be used to analyze large datasets and identify patterns and insights that would be difficult or impossible for humans to detect.
- **Robotics:** Machine learning can be used to train robots to perform tasks autonomously, such as navigating through a space or manipulating objects.

Issues of using data in Machine Learning:

- **Data quality:** One of the biggest issues with using data in machine learning is ensuring that the data is accurate, complete, and representative of the problem domain. Low-quality data can result in inaccurate or biased models.
- **Data quantity:** In some cases, there may not be enough data available to train an accurate machine learning model. This is especially true for complex problems that require a large amount of data to accurately capture all the relevant patterns and relationships.
- **Bias and fairness:** Machine learning models can sometimes perpetuate bias and discrimination if the training data is biased or unrepresentative. This can

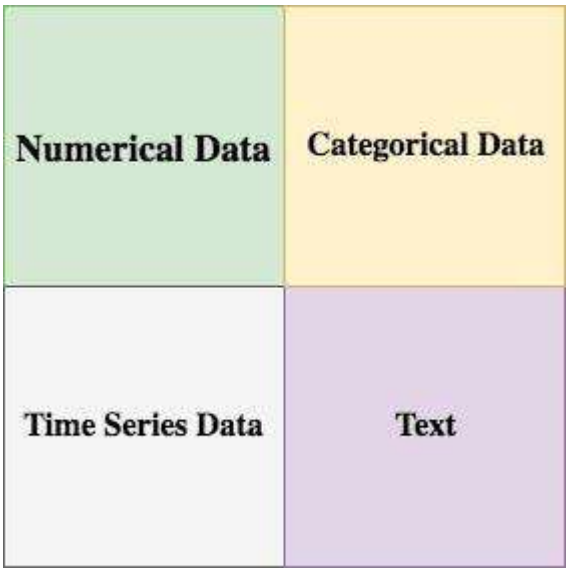
lead to unfair outcomes for certain groups of people, such as minorities or women.

- **Overfitting and underfitting:** Overfitting occurs when a model is too complex and fits the training data too closely, resulting in poor generalization to new data. Underfitting occurs when a model is too simple and does not capture all the relevant patterns in the data.
- **Privacy and security:** Machine learning models can sometimes be used to infer sensitive information about individuals or organizations, raising concerns about privacy and security.
- **Interpretability:** Some machine learning models, such as deep neural networks, can be difficult to interpret and understand, making it challenging to explain the reasoning behind their predictions and decisions.

- Almost anything can be turned into DATA. Building a deep understanding of the different data types is a crucial prerequisite for doing Exploratory Data Analysis (EDA) and Feature Engineering for Machine Learning models. You also need to convert data types of some variables in order to make appropriate choices for visual encodings in data visualization and storytelling.

- Most data can be categorized into 4 basic types from a Machine Learning perspective: numerical data, categorical data, time-series data, and text.

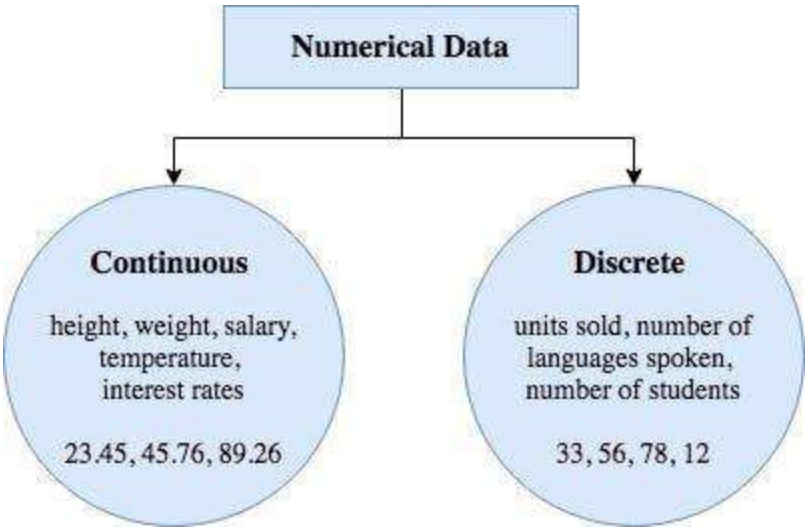
Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**



Numerical Data

Numerical data is any data where data points are exact numbers. Statisticians also might call numerical data, quantitative data. This data has meaning as a **measurement** such as house prices or as a count, such as a number of residential properties in Los Angeles or how many houses sold in the past year.

Numerical data can be characterized by continuous or discrete data. Continuous data can assume any value within a range whereas discrete data has distinct values.



For example, the number of students taking Python class would be a discrete data set. You can only have discrete whole number values like 10, 25, or 33. A class cannot have 12.75 students enrolled. A student either join a class or he doesn't. On the other hand, continuous data are numbers that can fall anywhere within a range. Like a student could have an average score of 88.25 which falls between 0 and 100.

The takeaway here is that numerical data is not ordered in time. They are just numbers that we have collected.

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Categorical Data

Categorical data represents characteristics, such as a hockey player's position, team, hometown. Categorical data can take numerical values. For example, maybe we would use 1 for the colour red and 2 for blue. But these numbers don't have a mathematical meaning. That is, we can't add them together or take the average.

In the context of super classification, categorical data would be the class label. This would also be something like if a person is a man or woman, or property is residential or commercial.

There is also something called ordinal data, which in some sense is a mix of numerical and categorical data. In ordinal data, the data still falls into categories, but those categories are ordered or ranked in some particular way. An example would be class difficulty, such as beginner, intermediate, and advanced. Those three types of classes would be a way that we could label the classes, and they have a natural order in increasing difficulty.

Another example is that we just take quantitative data, and splitting it into groups, so we have bins or categories of other types of data.

Population Bins
0 – 10 million, 10 – 100 million, 100 – 500 million, > 500 million

Ordinal Data

For plotting purposes, ordinal data is treated much in the same way as categorical data. But groups are usually ordered from lowest to highest so that we can preserve this ordering.

Time Series Data

Time series data is a sequence of numbers collected at regular intervals over some period of time. It is very important, especially in particular fields like finance. Time series data has a temporal value attached to it, so this would be something like a date or a timestamp that you can look for trends in time.

For example, we might measure the average number of home sales for many years. The difference of time series data and numerical data is that rather than having a bunch of numerical values that don't have any time ordering, time-series data does have some implied ordering. There is a first data point collected and the last data point collected.

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

CREA

Text

Text data is basically just words. A lot of the time the first thing that you do with text is you turn it into numbers using some interesting functions like the bag of words formulation.

These are four types of data from a Machine Learning perspective. Depending on exactly the type of data, this might have some repercussions for the type of algorithms that you can use for feature engineering and modeling, or the type of questions that you can ask of it.

12 Data Plot Types for Visualisation from Concept to Code

Introduction

When data is collected, there is a need to interpret and analyze it to provide insight into it. This insight can be about patterns, trends, or relationships between variables. Data interpretation is the process of reviewing data through well-defined methods. They help assign meaning to the data and arrive at a relevant conclusion. The analysis is the process of ordering, categorizing, and summarizing data to answer research questions. It should be done quickly and effectively. The results need to stand out and should be right in your face. Data Plot types for Visualization is an important aspect of this end. With growing data, this need is growing and hence data plots become very important in today's world. However, there are many types of plots used in data visualization. It is often tricky to choose which type is best for your business or data. Each of these

plots has its strengths and weaknesses that make it better than others in some situations.

This article provides a comprehensive list of data plots and their further subtypes. It discusses which one is right for the given problem.

Several packages can be used for this purpose. Popular packages widely used for this purpose are plotly and seaborn. This article will look at code that draws these plots in plotly and seaborn / matplotlib. The visual representation of these plots is given here for understanding. The code used in this article to generate plots and corresponding generated visual plots is posted on GitHub at: https://github.com/sameermahajan/MLWorkshop/tree/master/13.%20Visualization

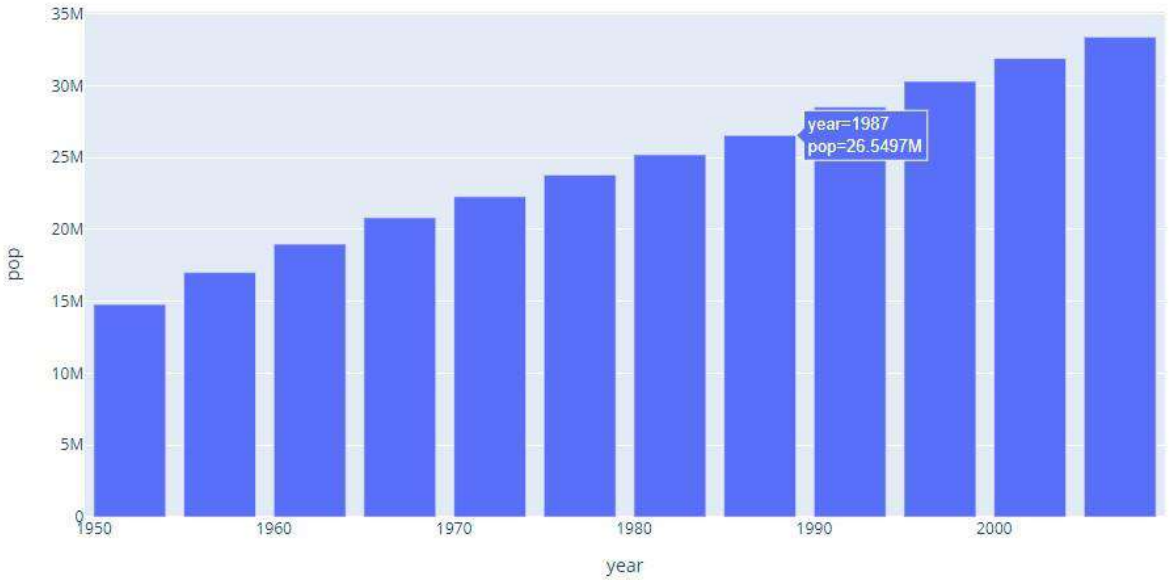
These data plot types for visualization are sometimes called graphs or charts depending on the context.

Bar Graph

A bar graph is a graph that presents categorical data with rectangle-shaped bars. The heights or lengths of these bars are proportional to the values that they represent. The bars can be vertical or horizontal. A vertical bar graph is sometimes called a column graph.

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Following is an illustration of a bar graph indicating the population in Canada by years.



This article was published as a part of the Data Science Blogathon

Introduction

When data is collected, there is a need to interpret and analyze it to provide insight into it. This insight can be about patterns, trends, or relationships between variables. Data interpretation is the process of reviewing data through well-defined methods. They help assign meaning to the data and arrive at a relevant conclusion. The analysis is the

process of ordering, categorizing, and summarizing data to answer research questions. It should be done quickly and effectively. The results need to stand out and should be right in your face. Data Plot types for Visualization is an important aspect of this end. With growing data, this need is growing and hence data plots become very important in today's world. However, there are many types of plots used in data visualization. It is often tricky to choose which type is best for your business or data. Each of these plots has its strengths and weaknesses that make it better than others in some situations.

This article provides a comprehensive list of data plots and their further subtypes. It discusses which one is right for the given problem.

Several packages can be used for this purpose. Popular packages widely used for this purpose are plotly and seaborn. This article will look at code that draws these plots in plotly and seaborn / matplotlib. The visual representation of these plots is given here for understanding. The code used in this article to generate plots and corresponding generated visual plots is posted on GitHub at: https://github.com/sameermahajan/MLWorkshop/tree/master/13.%20Visualization

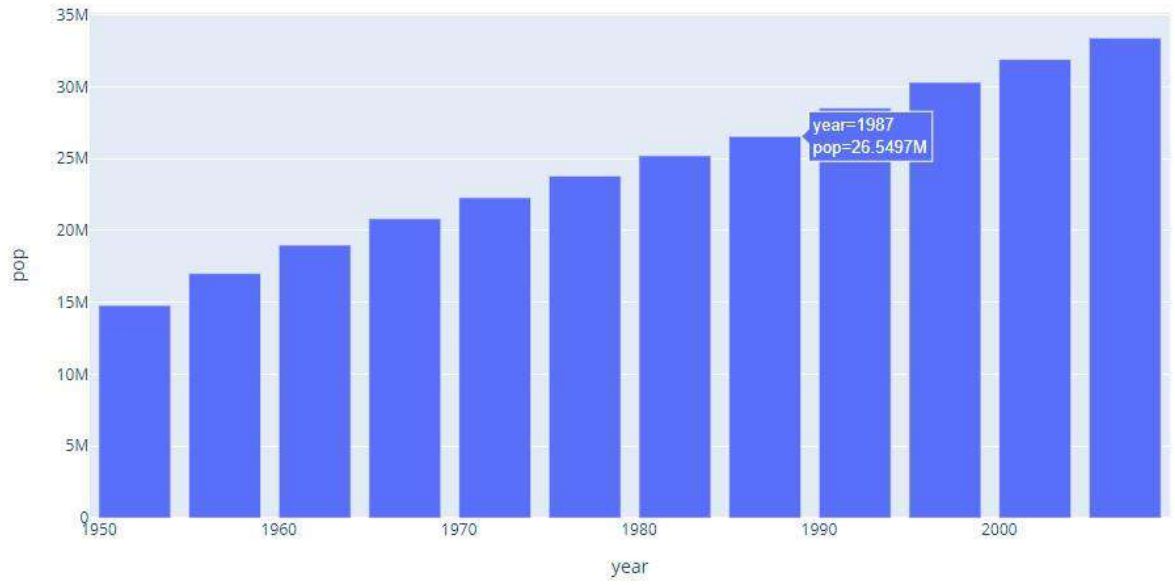
These data plot types for visualization are sometimes called graphs or charts depending on the context.

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Bar Graph

A bar graph is a graph that presents categorical data with rectangle-shaped bars. The heights or lengths of these bars are proportional to the values that they represent. The bars can be vertical or horizontal. A vertical bar graph is sometimes called a column graph.

Following is an illustration of a bar graph indicating the population in Canada by years.

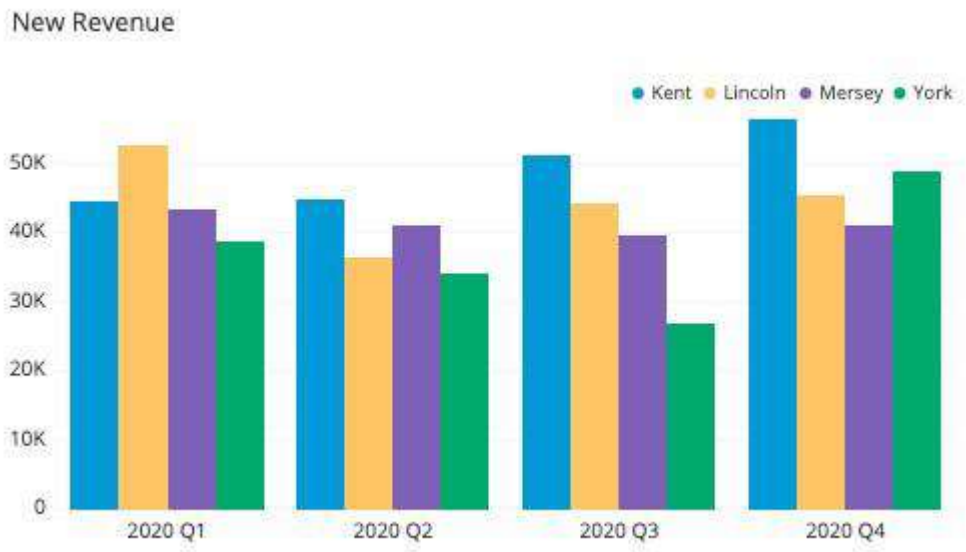


The following are types of bar graphs:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Grouped Bar Graph

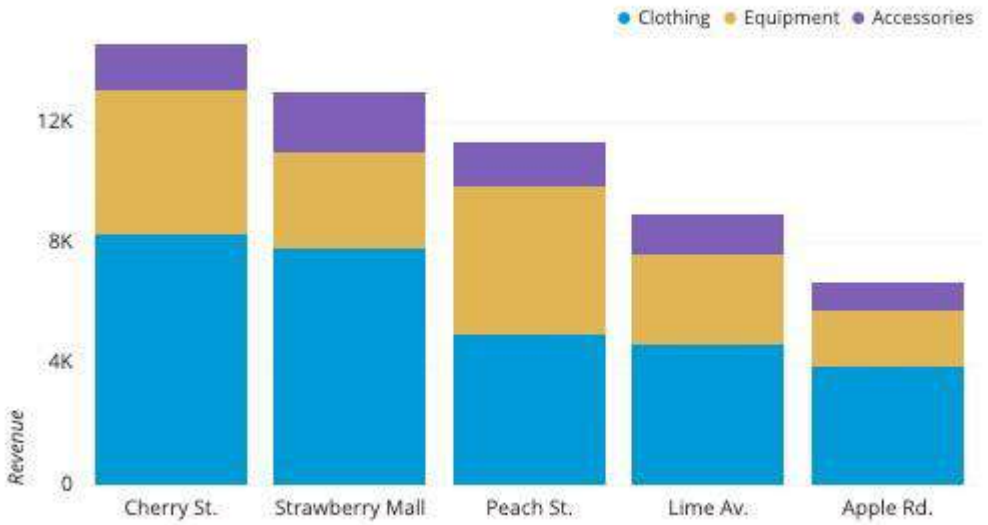
Grouped bar graphs are used when the datasets have subgroups that need to be visualized on the graph. The subgroups are differentiated by distinct colours. Here is an illustration of such a graph:



Stacked Bar Graph

The stacked bar graphs are used to show dataset subgroups. However, the bars are stacked on top of each other. Here is an illustration:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

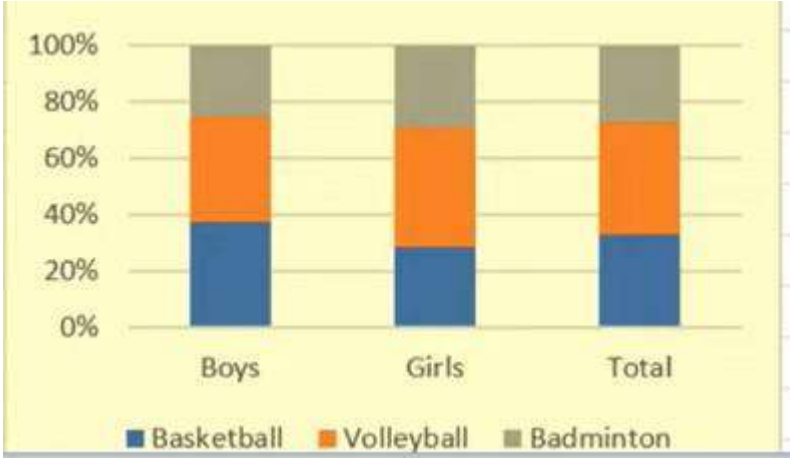


Here

Segmented Bar Graph

This is the type of stacked bar graph where each stacked bar shows the percentage of its discrete value from the total value. The total percentage is 100%. Here is an illustration:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

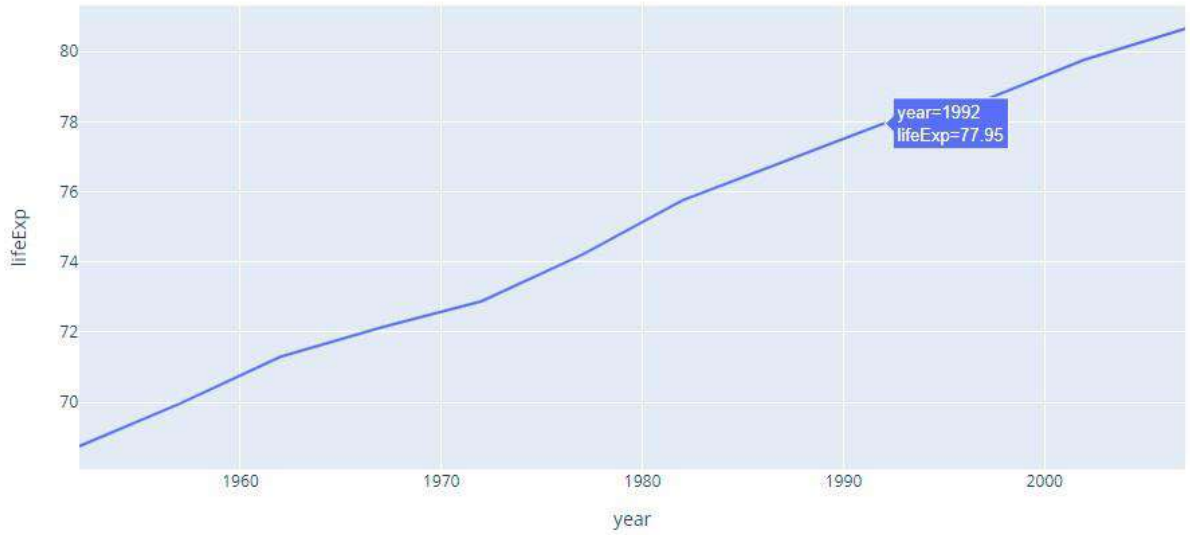


Line Graph

It displays a sequence of data points as markers. The points are ordered typically by their x-axis value. These points are joined with straight line segments. A line graph is used to visualize a trend in data over intervals of time.

The following is an illustration of Canadian life expectancy by years in Line Graph.

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**



Simple Line Graph

A simple line graph plots only one line on the graph. One of the axes defines the independent variable. The other axis contains a variable that depends on it.

Multiple Line Graph

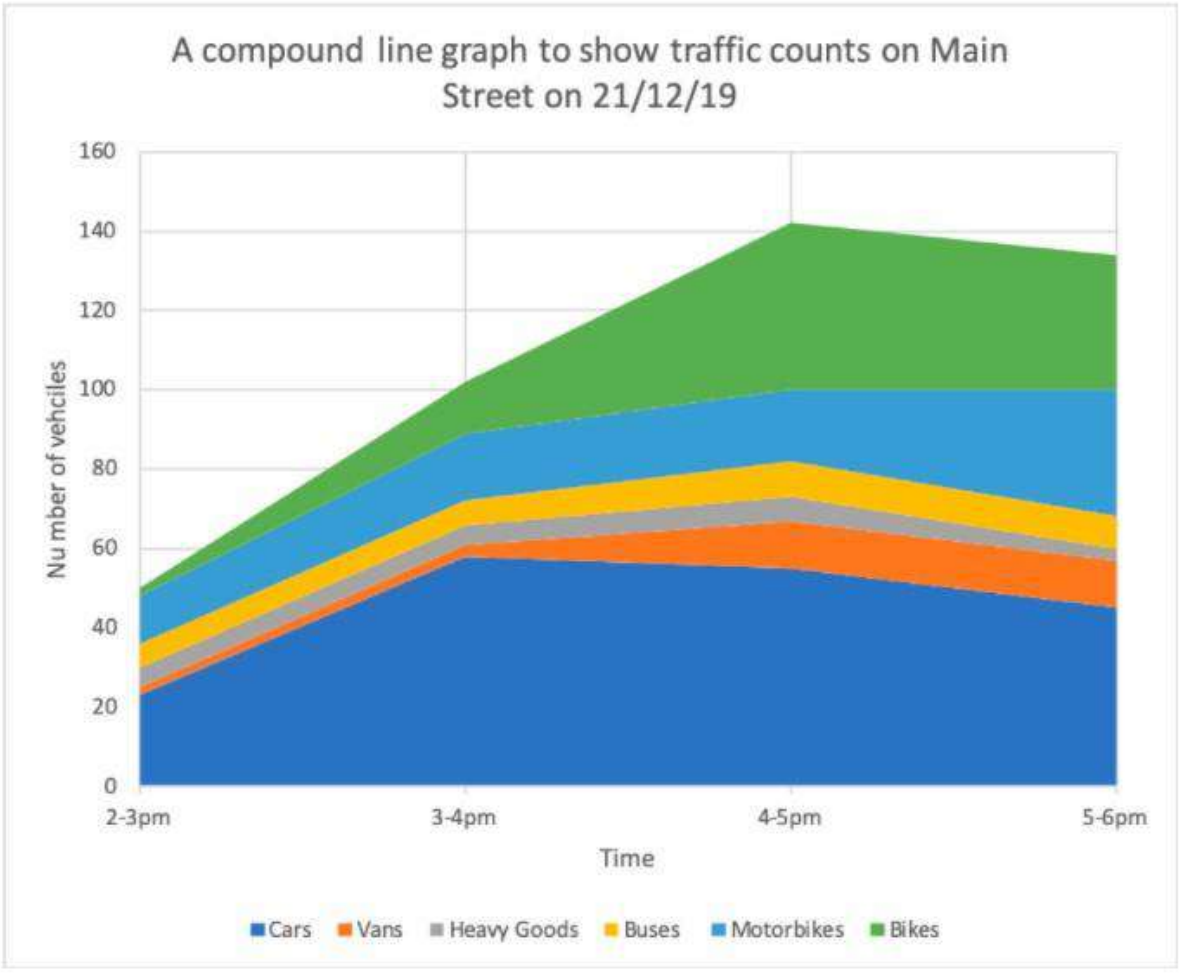
Multiple line graphs contain more than one line. They represent multiple variables in a dataset. This type of graph can be used to study more than one variable over the same period.

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Compound Line Graph

It is an extension of a simple line graph. It is used when dealing with different groups of data from a larger dataset. Its every line graph is shaded downwards to the x-axis. It has each group stacked upon one another.

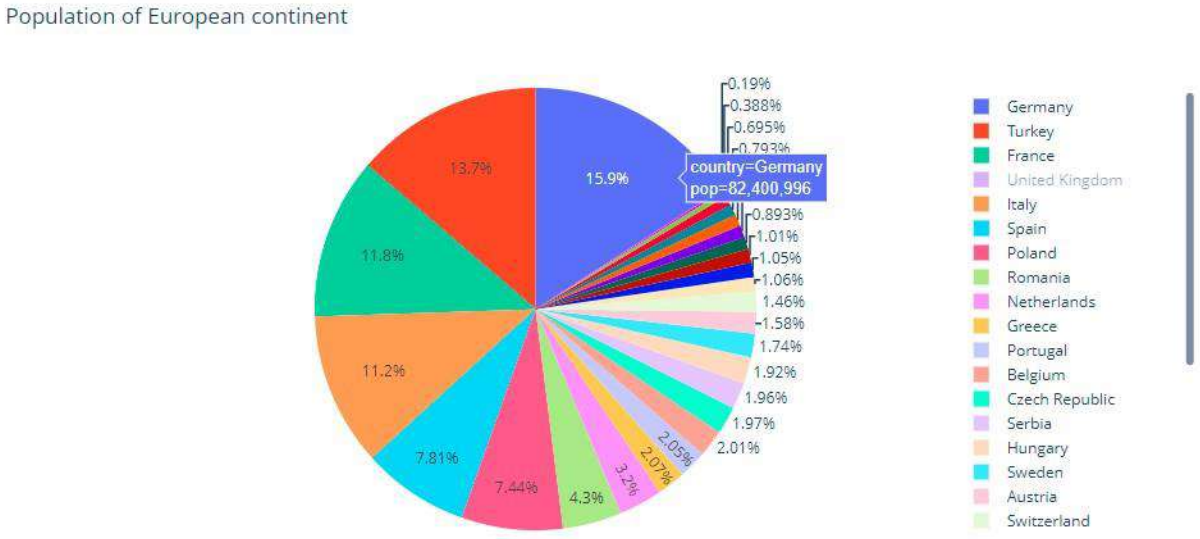
Here is an illustration:



Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Pie Chart

A pie chart is a circular statistical graphic. To illustrate numerical proportion, it is divided into slices. In a pie chart, for every slice, each of its arc lengths is proportional to the amount it represents. The central angles, and area are also proportional. It is named after a sliced pie.



Introduction

When data is collected, there is a need to interpret and analyze it to provide insight into it. This insight can be about patterns, trends, or relationships between variables. Data interpretation is the process of reviewing data through well-defined methods. They help assign meaning to the data and arrive at a relevant conclusion. The analysis is the process of ordering, categorizing, and summarizing data to answer research questions.

It should be done quickly and effectively. The results need to stand out and should be right in your face. Data Plot types for Visualization is an important aspect of this end. With growing data, this need is growing and hence data plots become very important in today's world. However, there are many types of plots used in data visualization. It is often tricky to choose which type is best for your business or data. Each of these plots has its strengths and weaknesses that make it better than others in some situations.

This article provides a comprehensive list of data plots and their further subtypes. It discusses which one is right for the given problem.

Several packages can be used for this purpose. Popular packages widely used for this purpose are plotly and seaborn. This article will look at code that draws these plots in plotly and seaborn / matplotlib. The visual representation of these plots is given here for understanding. The code used in this article to generate plots and corresponding generated visual plots is posted on GitHub at: https://github.com/sameermahajan/MLWorkshop/tree/master/13.%20Visualization

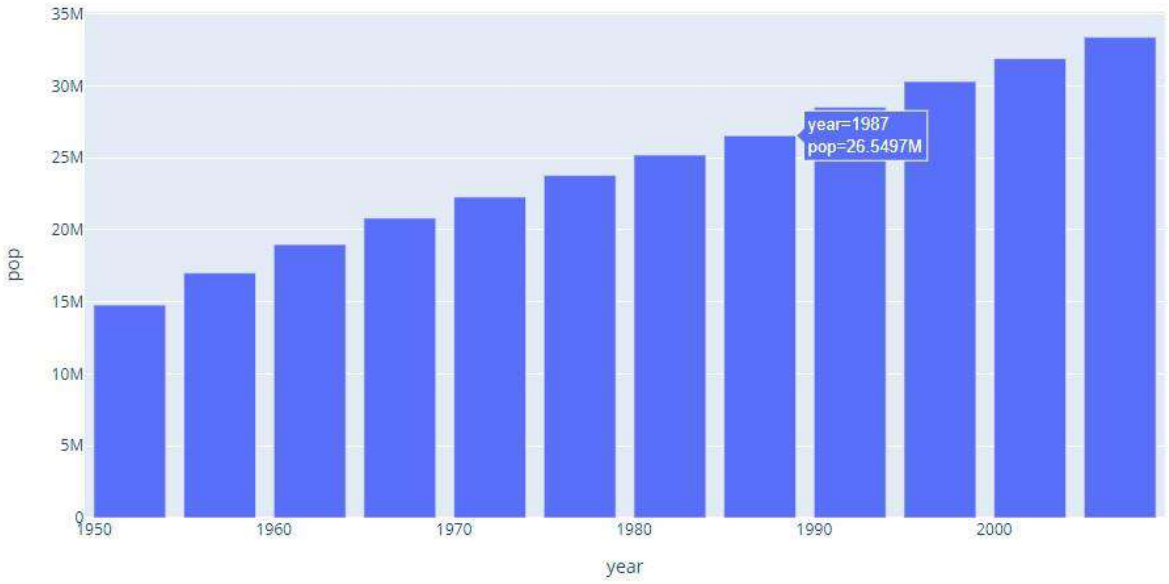
These data plot types for visualization are sometimes called graphs or charts depending on the context.

Bar Graph

A bar graph is a graph that presents categorical data with rectangle-shaped bars. The heights or lengths of these bars are proportional to the values that they represent. The

bars can be vertical or horizontal. A vertical bar graph is sometimes called a column graph.

Following is an illustration of a bar graph indicating the population in Canada by years.



Following is the code indicating how to do it in plotly.

```
import plotly.express as px
data_canada = px.data.gapminder().query("country == 'Canada'")
fig = px.bar(data_canada, x='year', y='pop')
fig.show()
```

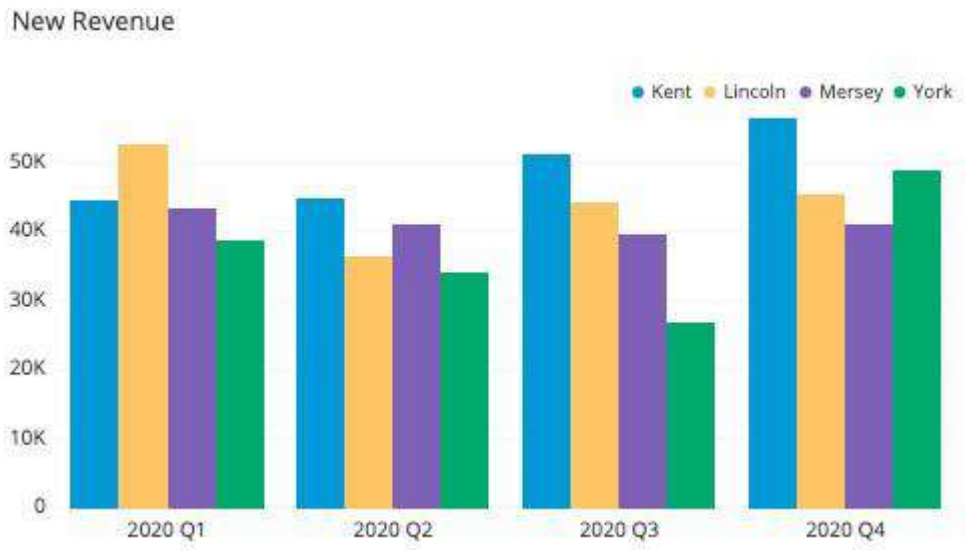
Following is the representational code of doing it in seaborn.

The following are types of bar graphs:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Grouped Bar Graph

Grouped bar graphs are used when the datasets have subgroups that need to be visualized on the graph. The subgroups are differentiated by distinct colours. Here is an illustration of such a graph:



Here is a code snippet on how to do it in plotly:

```
import plotly.express as px
df = px.data.tips()
fig = px.bar(df, x="sex", y="total_bill", color="time")
fig.show()
```

Here is a code snippet on how to do it in seaborn:

```
import seaborn as sb
df = sb.load_dataset('tips')
```

```
df = df.groupby(['size', 'sex']).agg(mean_total_bill=("total_bill",
'mean'))
df = df.reset_index()
sb.barplot(x="size", y="mean_total_bill", hue="sex", data=df)
```

Stacked Bar Graph

The stacked bar graphs are used to show dataset subgroups. However, the bars are stacked on top of each other. Here is an illustration:

Here is a code snippet on how to do it in plotly:

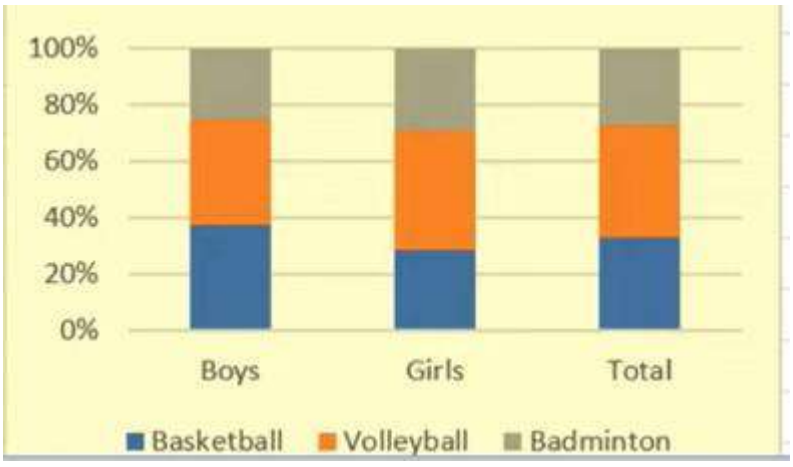
```
import plotly.express as px
df = px.data.tips()
fig = px.bar(df, x="sex", y="total_bill", color='time')
fig.show()
```

Seaborn code snippet:

```
import pandas
import matplotlib.pylab as plt
import seaborn as sns
plt.rcParams["figure.figsize"] = [7.00, 3.50]
plt.rcParams["figure.autolayout"] = True
df = pandas.DataFrame(dict(
   number=[2, 5, 1, 6, 3],
   count=[56, 21, 34, 36, 12],
   select=[29, 13, 17, 21, 8]
))
bar_plot1 = sns.barplot(x='number', y='count', data=df, label="count",
color="red")
bar_plot2 = sns.barplot(x='number', y='select', data=df,
label="select", color="green")
plt.legend(ncol=2, loc="upper right", frameon=True)
plt.show()
```

Segmented Bar Graph

This is the type of stacked bar graph where each stacked bar shows the percentage of its discrete value from the total value. The total percentage is 100%. Here is an illustration:



Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Line Graph

It displays a sequence of data points as markers. The points are ordered typically by their x-axis value. These points are joined with straight line segments. A line graph is used to visualize a trend in data over intervals of time.

The following is an illustration of Canadian life expectancy by years in Line Graph.

Here is how to do it in plotly:

```
import plotly.express as px
df = px.data.gapminder().query("country=='Canada'")
fig = px.line(df, x="year", y="lifeExp", title='Life expectancy in
Canada')
fig.show()
```

Here is how to do it in seabron:

```
import seaborn as sns
sns.lineplot(data=df, x="year", y="lifeExp")
```

Here are types of line graphs:

Simple Line Graph

A simple line graph plots only one line on the graph. One of the axes defines the independent variable. The other axis contains a variable that depends on it.

Multiple Line Graph

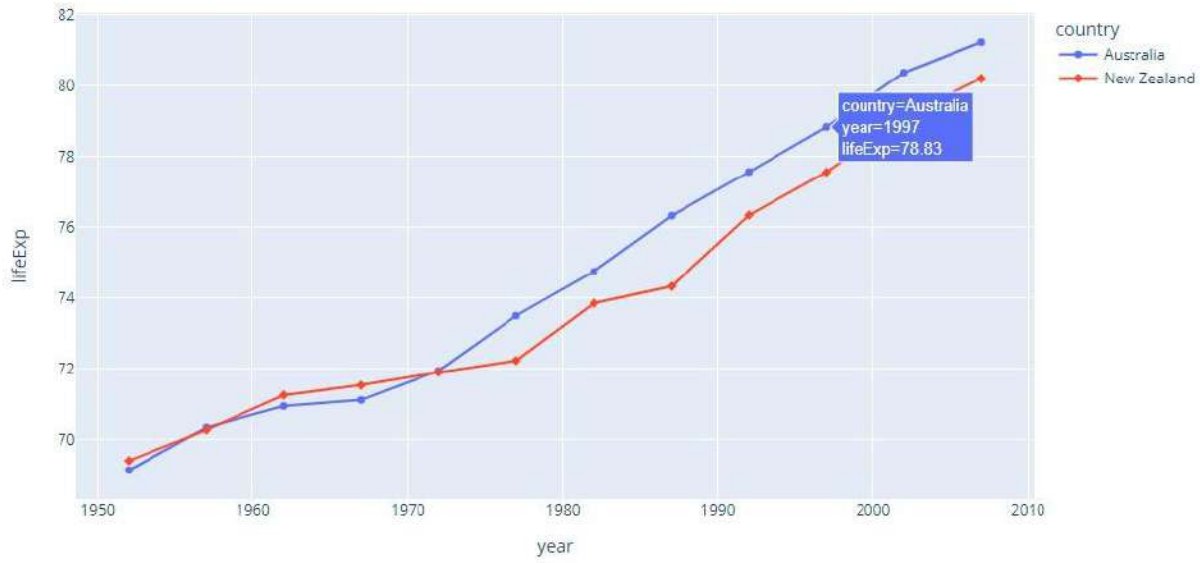
Multiple line graphs contain more than one line. They represent multiple variables in a dataset. This type of graph can be used to study more than one variable over the same period.

It can be drawn in plotly as:

```
import plotly.express as px
df = px.data.gapminder().query("continent == 'Oceania'")
fig = px.line(df, x='year', y='lifeExp', color='country',
symbol="country")
fig.show()
```

Here is the illustration:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

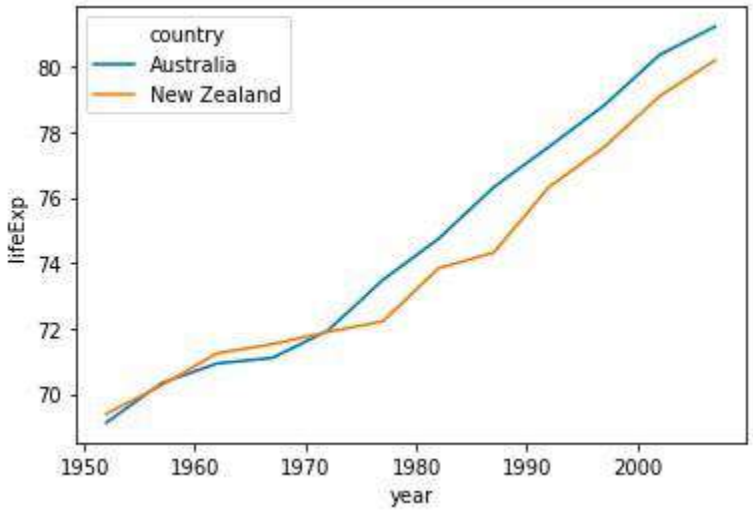


In seaborn as:

```
import seaborn as sns
sns.lineplot(data=df, x='year', y='lifeExp', hue='country')
```

Here is the illustration:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

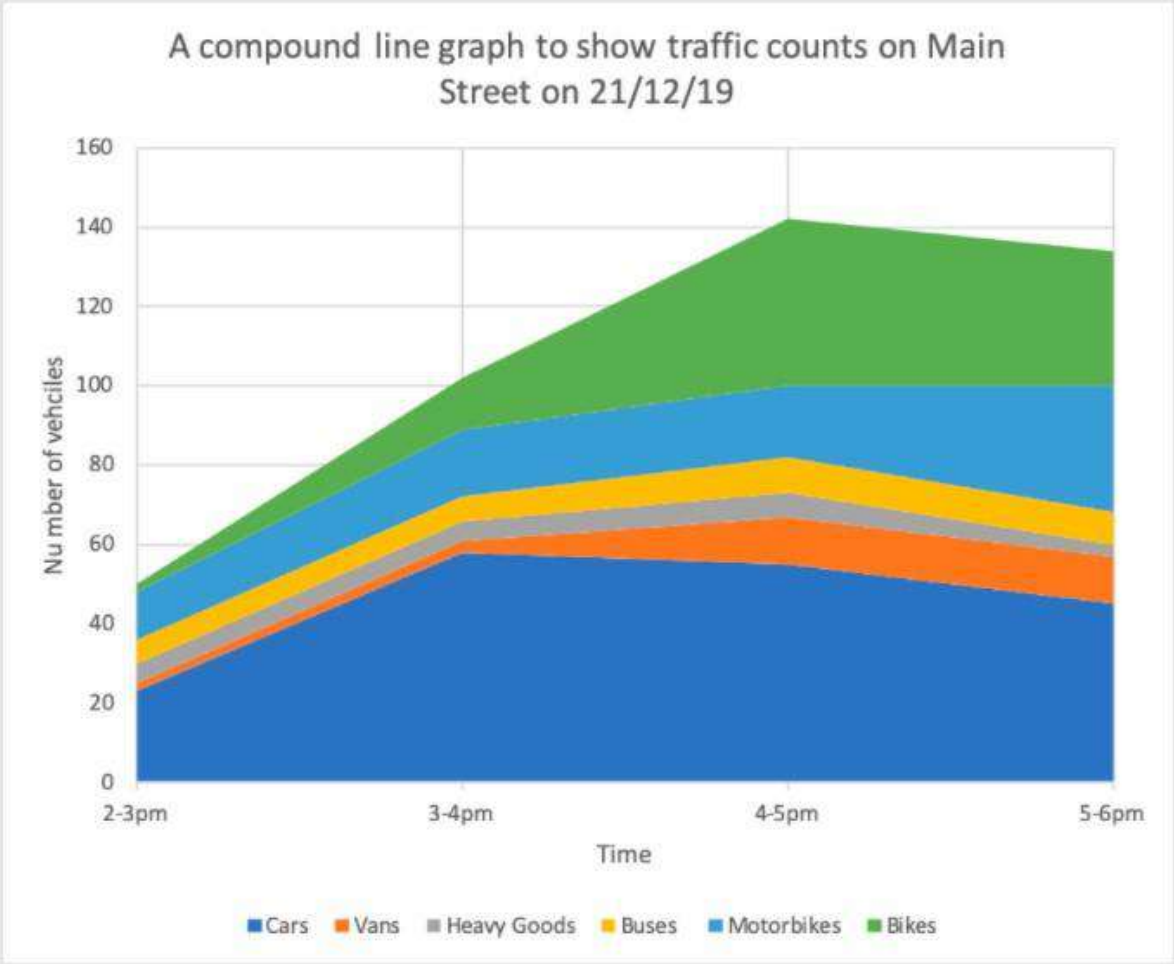


Compound Line Graph

It is an extension of a simple line graph. It is used when dealing with different groups of data from a larger dataset. Its every line graph is shaded downwards to the x-axis. It has each group stacked upon one another.

Here is an illustration:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**



Pie Chart

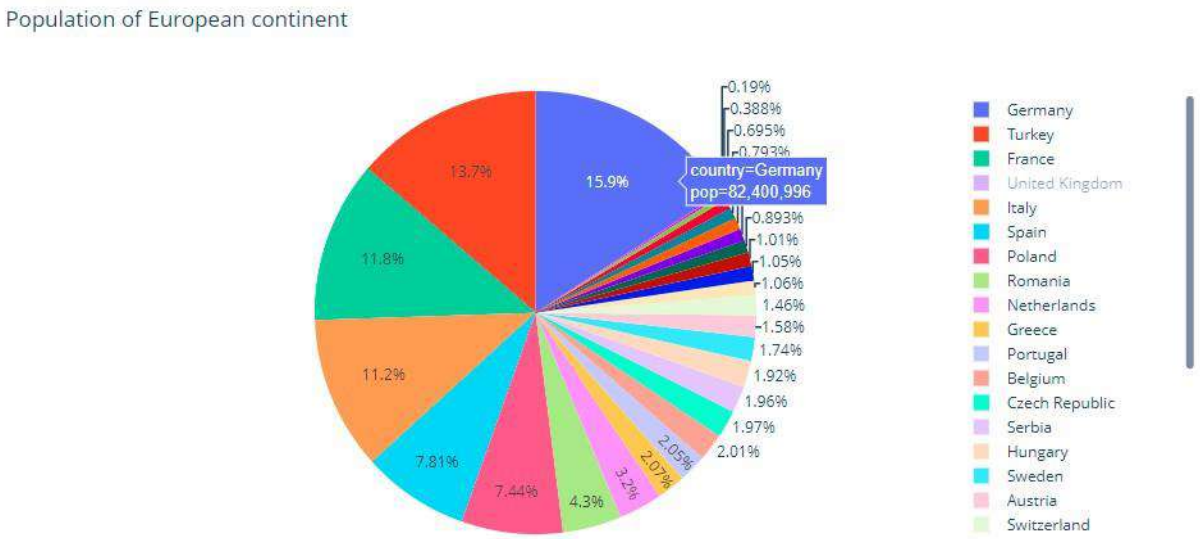
A pie chart is a circular statistical graphic. To illustrate numerical proportion, it is divided into slices. In a pie chart, for every slice, each of its arc lengths is proportional

to the amount it represents. The central angles, and area are also proportional. It is named after a sliced pie.

Here is how to do it in plotly:

```
import plotly.express as px
df = px.data.gapminder().query("year == 2007").query("continent ==
'Europe'")
df.loc[df['pop'] < 2.e6, 'country'] = 'Other countries' # Represent
only large countries
fig = px.pie(df, values='pop', names='country', title='Population of
European continent')
fig.show()
```

And here is how it looks:



Seaborn doesn't have a default function to create pie charts, but the following syntax in matplotlib can be used to create a pie chart and add a seaborn color palette:

```
import matplotlib.pyplot as plt
import seaborn as sns

data = [15, 25, 25, 30, 5]
labels = ['Group 1', 'Group 2', 'Group 3', 'Group 4', 'Group 5']

colors = sns.color_palette('pastel')[0:5]

plt.pie(data, labels = labels, colors = colors, autopct='%.0f%%')
plt.show()
```

This is how it looks:

These are types of pie charts:

Simple Pie Chart

This is the basic type of pie chart. It is often called just a pie chart.

Exploded Pie Chart

One or more sectors of the chart are separated (termed as exploded) from the chart in an exploded pie chart. It is used to emphasize a particular element in the data set.

Donut Chart

In this pie chart, there is a hole in the centre. The hole makes it look like a donut from which it derives its name.

Donut Chart

In this pie chart, there is a hole in the centre. The hole makes it look like a donut from which it derives its name.

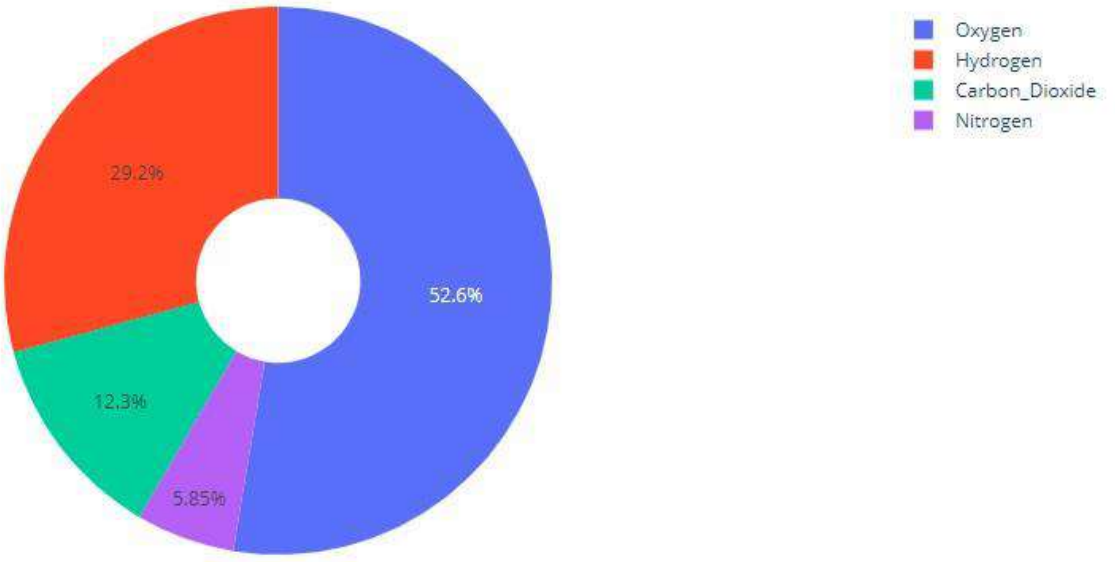
The way to do it in plotly is:

```
import plotly.graph_objects as go

labels = ['Oxygen','Hydrogen','Carbon_Dioxide','Nitrogen']
values = [4500, 2500, 1053, 500]

# Use `hole` to create a donut-like pie chart
fig = go.Figure(data=[go.Pie(labels=labels, values=values, hole=.3)])
fig.show()
```

And this is how it looks:



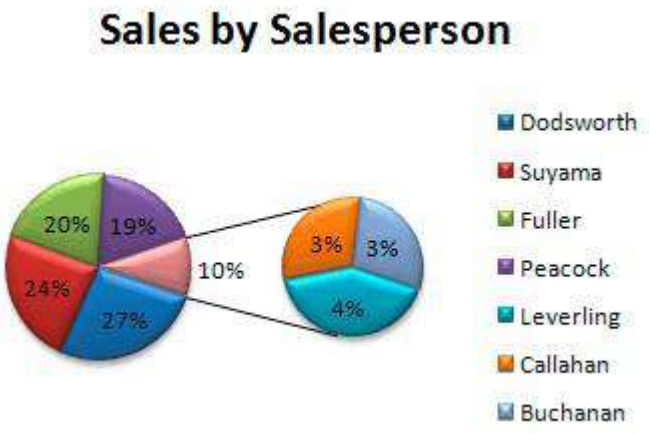
This is how it is done in seaborn:

```
import numpy as np
import matplotlib.pyplot as plt
data = np.random.randint(20, 100, 6)
plt.pie(data)
circle = plt.Circle( (0,0), 0.7, color='white')
p=plt.gcf()
p.gca().add_artist(circle)
plt.show()
```

Pie of Pie

A pie of pie is a chart that generates an entirely new pie chart detailing a small sector of the existing pie chart. It can be used to reduce the clutter and emphasize a particular group of elements.

Here is an illustration:

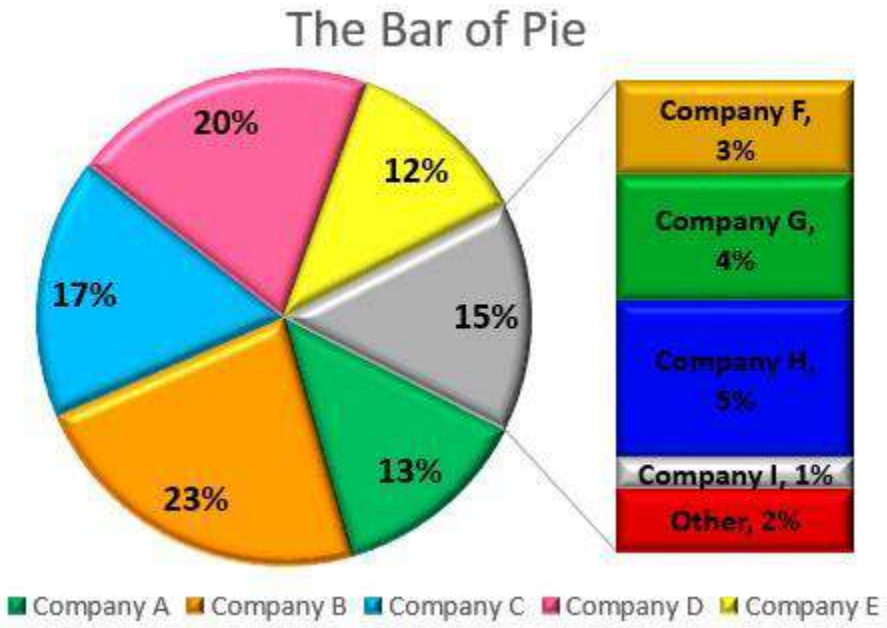


Bar of Pie

This is similar to the pie of pie, except that a bar chart is what is generated.

Here is an illustration:

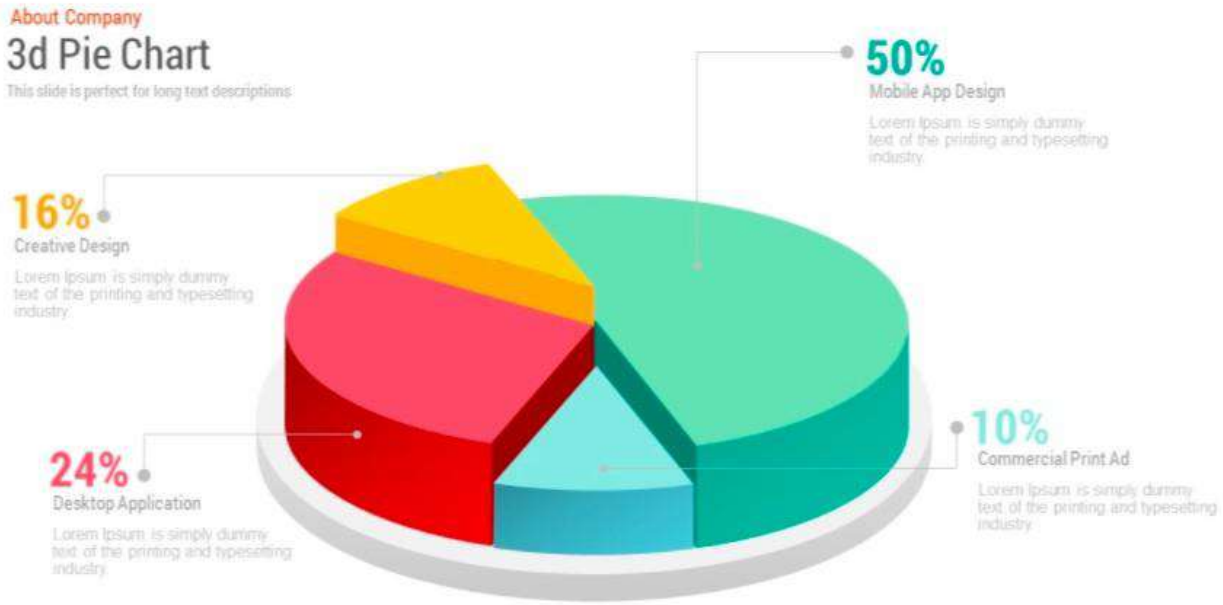
Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**



3D Pie Chart

This is a pie chart that is represented in a 3-dimensional space. Here is an illustration:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

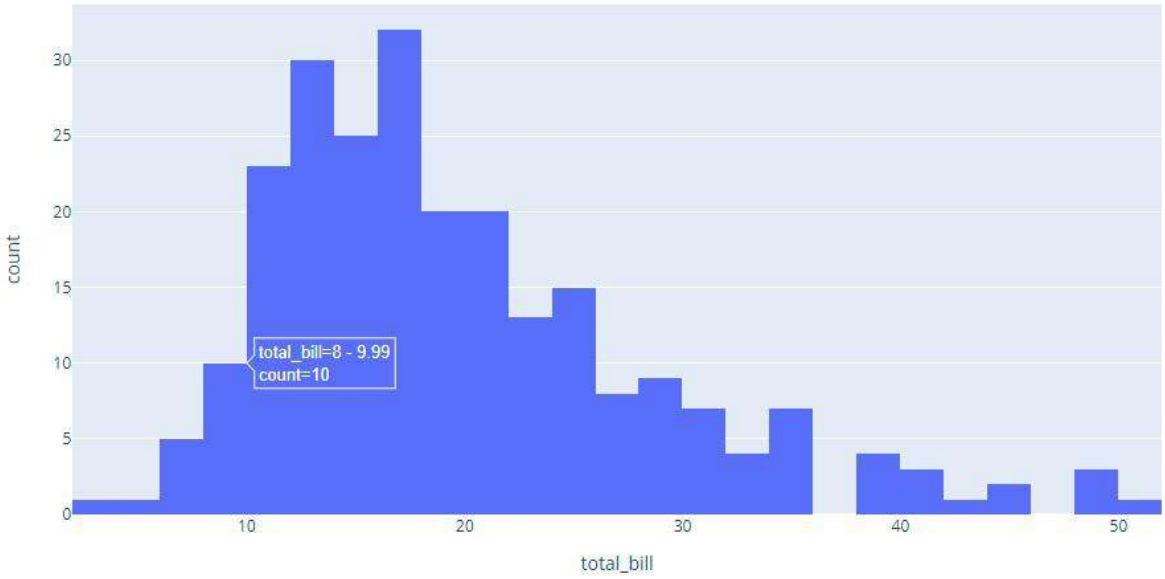


Histogram

A histogram is an approximate representation of the distribution of numerical data. The data is divided into non-overlapping intervals called bins and buckets. A rectangle is erected over a bin whose height is proportional to the number of data points in the bin. Histograms give a feel of the density of the distribution of the underlying data.

Here is a visual:

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**



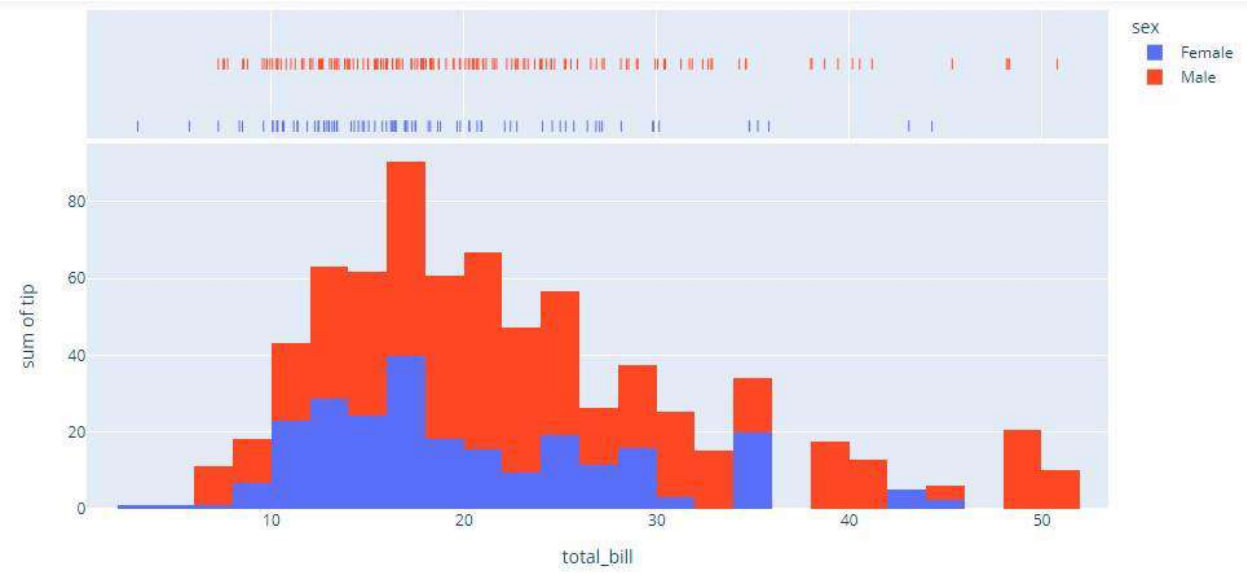
Normal Distribution

This chart is usually bell-shaped.

Bimodal Distribution

In this histogram, there are two groups of histogram charts that are of normal distribution. It is a result of combining two variables in a dataset.

Visualization:



Skewed Distribution

This is an asymmetric graph with an off-centre peak. The peak tends towards the beginning or end of the graph. A histogram can be said to be right or left-skewed depending on the direction where the peak tends towards.

Random Distribution

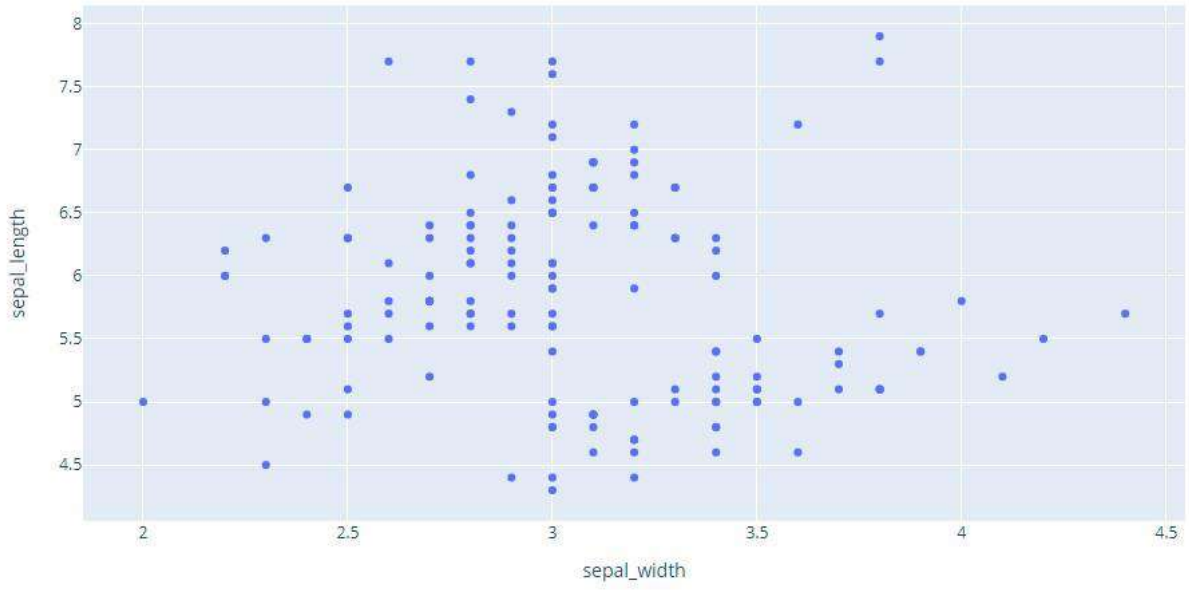
This histogram does not have a regular pattern. It produces multiple peaks. It can be called a multimodal distribution.

Subject: Honors: Mathematics for AI-ML (HAIMLC501) **SEM : V**

Scatter Plot

It is a type of plot using Cartesian coordinates to display values for two variables for a set of data. It is displayed as a collection of points. Their position on the horizontal axis determines the value of one variable. The position on the vertical axis determines the value of the other variable. A scatter plot can be used when one variable can be controlled and the other variable depends on it. It can also be used when both continuous variables are independent.

Visual:



According to the correlation of the data points, scatter plots are grouped into different types. These correlation types are listed below

Positive Correlation

In these types of plots, an increase in the independent variable indicates an increase in the variable that depends on it. A scatter plot can have a high or low positive correlation.

Negative Correlation

In these types of plots, an increase in the independent variable indicates a decrease in the variable that depends on it. A scatter plot can have a high or low negative correlation.

No Correlation

Two groups of data visualized on a scatter plot are said to not correlate if there is no clear correlation between them.