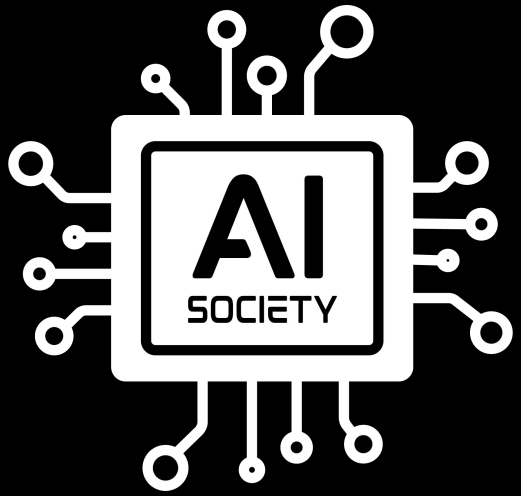


Machine Learning Lab

Week #2 - 18 Sept ‘24

The AI Society at Arizona State University

Objective

By the end of this workshop, you will understand

How to:

- Load and explore data.
- Handle missing values.
- Perform data transformations and feature extraction.
- Visualize data to uncover insights.

Quick Recap : Understanding Data

Data Pre-processing

It involves transforming raw data into an understandable format. It is a critical step before any machine learning model training, ensuring the data is clean and suitable for analysis.

1. Data Cleaning
2. Data Visualization
3. Feature Engineering
4. Handling Outliers, Normalizing data and Encoding

Data Cleaning

The process of detecting, correcting, or removing inaccuracies and inconsistencies in the dataset to improve data quality.

Handling Missing Values

The process of addressing incomplete data by removing or imputing missing entries to maintain data consistency.

```
# Fill missing values in 'host_org' with 'Unknown'
df['host_org'].fillna('Unknown', inplace=True)

# Fill missing values in 'event_perks' with 'None'
df['event_perks'].fillna('None', inplace=True)
```

Feature Engineering

Process of using domain knowledge to select, modify, or create new features from raw data, enhancing the machine learning model's performance.

Feature Engineering

Definition:

Feature engineering is the process of using domain knowledge to select, modify, or create new features from raw data, enhancing the machine learning model's performance.

Importance :

- Better Model Performance
- Reduce Overfitting
- Improved accuracy
- Faster Computation

Remove Irrelevant Data

Eliminating unnecessary or redundant features from the dataset that do not contribute to the model's performance, reducing the computational complexity.

```
# dropping irrelevant features.
df.drop(['web-scraper-order', 'web-scraper-start-url', 'event_link', 'event_link-href', 'date_extracted',
'time_extracted', 'date_time', 'description_normalized','location'], axis=1, inplace=True)
```

Handling Duplicates

Detecting and removing repeated data entries to avoid redundancy and ensure dataset accuracy.

```
1 df.groupby('Event_name')['datetime'].value_counts()[df.groupby('Event_name')['datetime'].value_counts() > 1]
```

		count
Event_name	datetime	
11th Annual Healthcare Panel Banquet	2024-11-14 18:00:00	3
11th Annual Poly Game Night	2024-08-23 18:00:00	4
2024 Annual SunMUN High School Conference	2024-11-15 08:00:00	4
A.T. Still University Campus Visit- Pre health programs	2024-10-18 12:00:00	2
ACF weekly event	2024-08-03 12:00:00	4
...	...	...
Welcome to West Valley!	2024-08-17 18:00:00	4
West Fest	2024-08-21 17:00:00	3
West Valley goes BIG (12)	2024-08-18 19:00:00	4
What is SoDA and Can I Drink It?	2024-08-27 19:30:00	8
What is TRIO Teacher Prep (and why should I attend this workshop)?	2024-08-07 16:00:00	2

```python
# Group by 'Event_name' and 'datetime', then aggregate categories and location_extracted
result = df.groupby(['Event_name', 'datetime', 'description']).agg({
    'host_org': lambda x: ', '.join(x[x != 'Unknown'].unique()),
    'event_perks': lambda x: ', '.join(x[x != 'None'].unique()),
    'categories': lambda x: ', '.join(x[x != "Not Specified"].unique()),
    'location_extracted': lambda x: ', '.join(x[x != 'Unknown'].unique())  # Exclude 'Unknown' locations
}).reset_index()
```

Data Aggregation

Data aggregation is the process of removing noise from the data by combining and organizing data from multiple sources into a single, unified body.

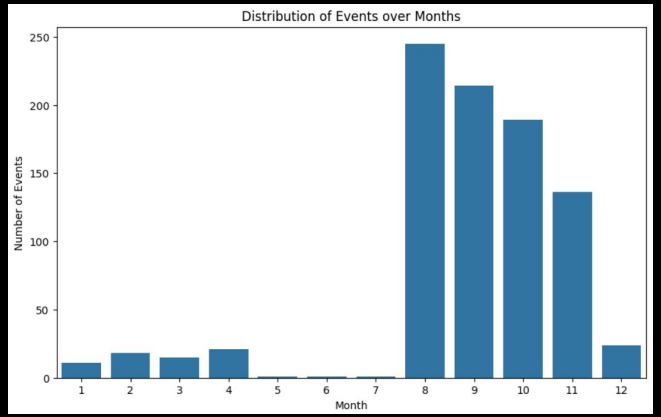
```
[ ]  # Group by 'Event_name' and 'datetime', then aggregate categories and location_extracted
     result = df.groupby(['Event_name', 'datetime', 'description']).agg({
         'host_org': lambda x: ', '.join(x[x != 'Unknown'].unique()),
         'event_perks': lambda x: ', '.join(x[x != 'None'].unique()),
         'categories': lambda x: ', '.join(x[x != "Not Specified"].unique()),
         'location_extracted': lambda x: ', '.join(x[x != 'Unknown'].unique())  # Exclude 'Unknown' locations
     }).reset_index()
```

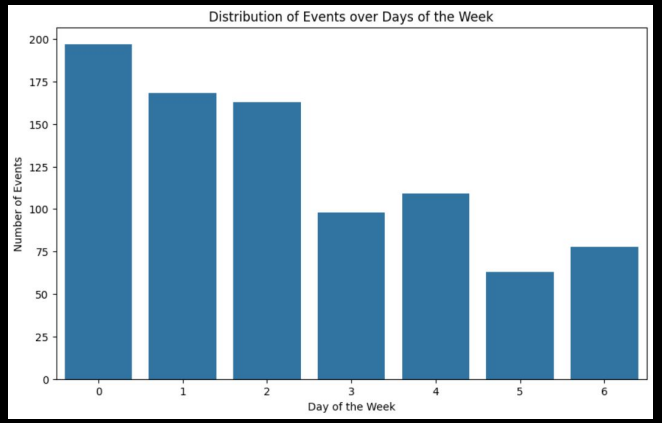
Exploratory Data Analysis

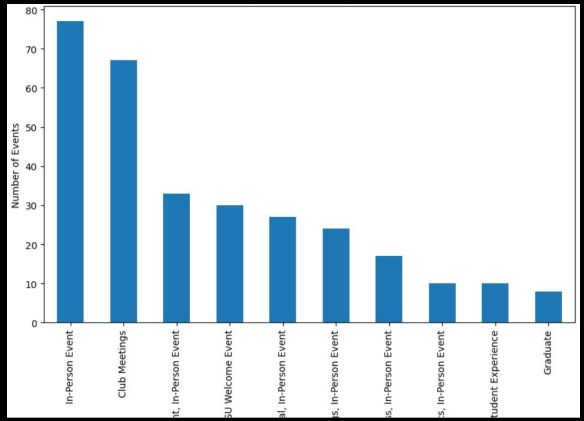
Process of examining and visualizing datasets to uncover patterns, detect anomalies, test hypotheses, and check assumptions using statistical summaries and graphical representations.

Data Visualization

The graphical representation of data to identify patterns, trends, and insights using charts, graphs, and plots.

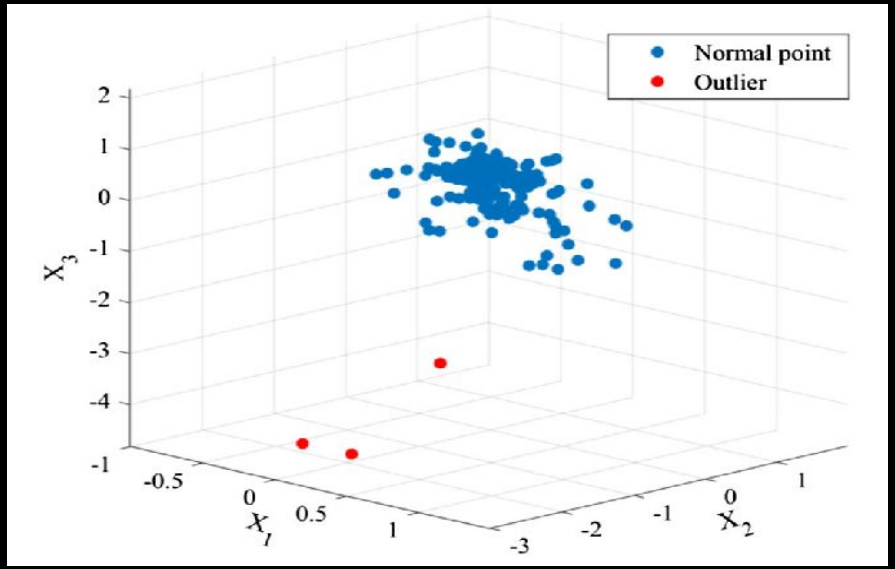


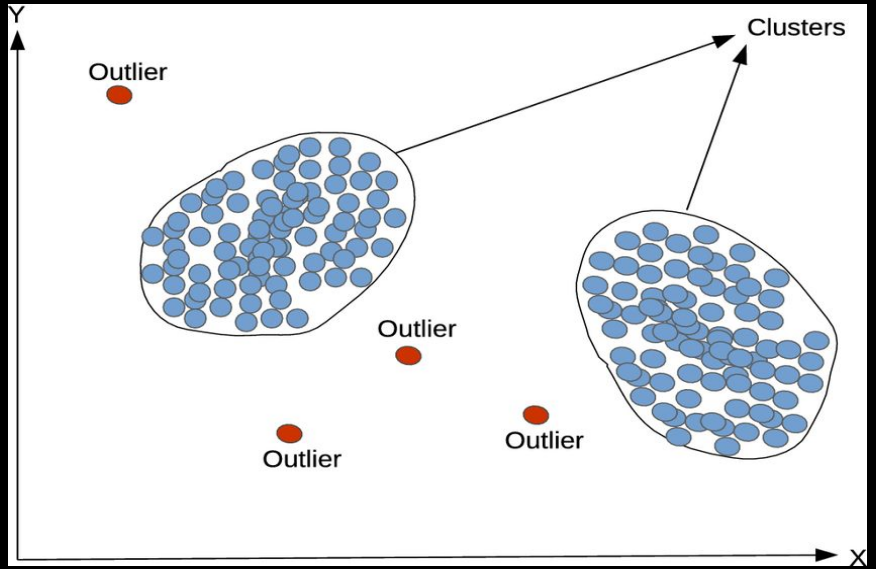




Advantages of Data Visualization

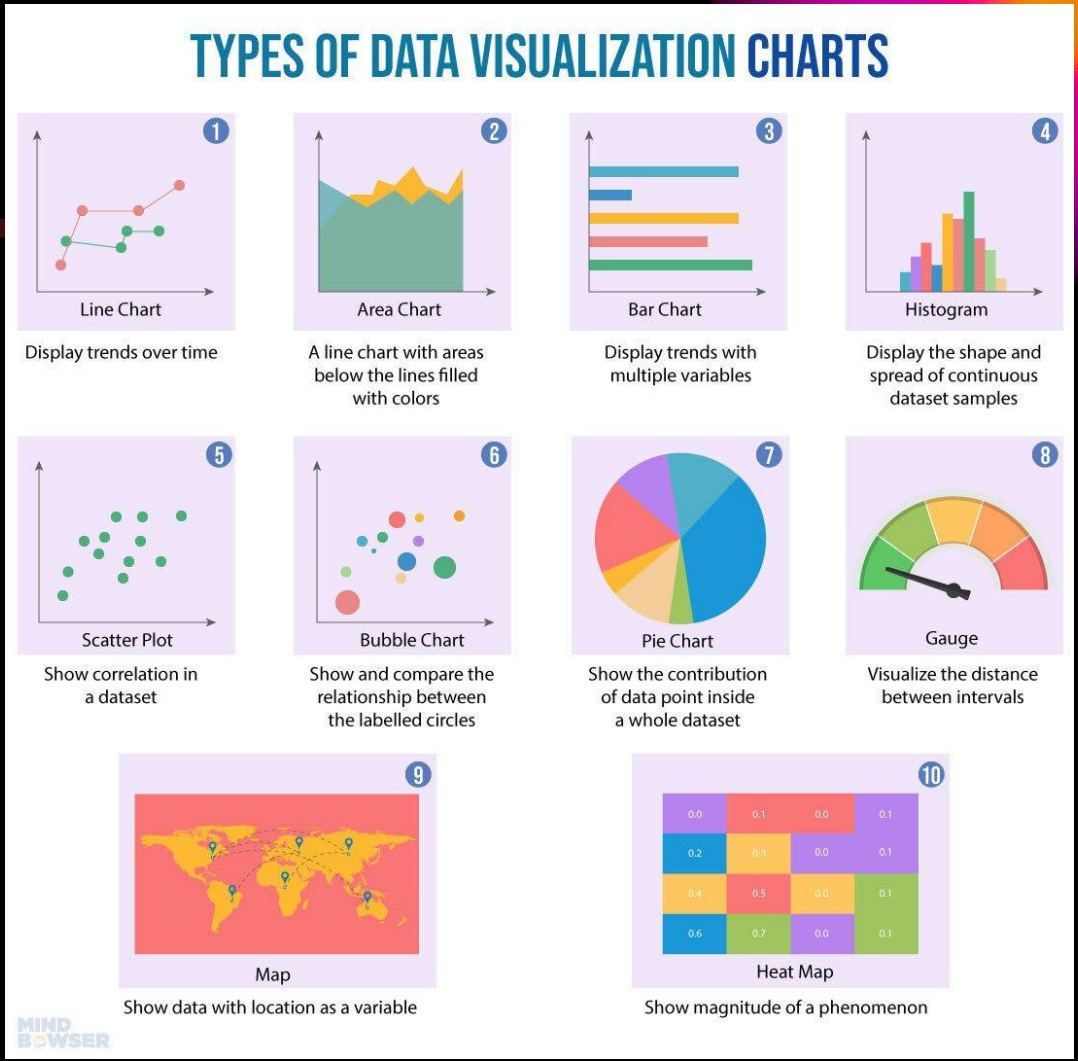
Identify Outliers





Advantages of Data Visualization

Identify Patterns and Trends



Additional Resources

Python: https://www.w3schools.com/python/

Python Video: https://www.youtube.com/watch?v=_uQrJ0TkZlc

Libraries: https://www.geeksforgeeks.org/libraries-in-python/

Pandas: https://www.w3schools.com/python/pandas/default.asp

Join our Discord for more updates and resources!

https://www.bit.ly/AIS-Links

Test your knowledge!

Group Picture!