# Online Session :

## Dataset Description:

### Dataset Information :

This dataset comprises of sales transactions captured at a retail store. It’s a classic dataset to explore and expand your feature engineering skills and day to day understanding from multiple shopping experiences. This is a regression problem. The dataset has 550,069 rows and 12 columns.

Problem: Predict purchase amount

### Gender:

This column represents the gender of the individuals participating in the Black Friday sales. It typically has values like 'Male' and 'Female.'

### Age:

The 'Age' column indicates the age group of the customers. It may be categorized into different age ranges such as '18-25,' '26-35,' '36-45,' '46-50,' '51-55,' '55+,' etc.

### Occupation:

This column represents the occupation of the customers involved in the Black Friday sales. Occupations are usually coded with numerical values or specific labels.

### City_Category:

The 'City_Category' column denotes the category of the city where the customer resides or is making the purchase. Cities may be categorized as 'A,' 'B,' or 'C,' indicating different levels of development or size.

### Stay_In_Current_City_Years:

This column represents the number of years the customer has been residing in their current city. It may have values like '0,' '1,' '2,' '3,' '4+,' indicating the duration of stay.

### Marital_Status:

The 'Marital_Status' column indicates whether the customer is married or not. It typically has binary values like '0' for unmarried and '1' for married.

Purchase:

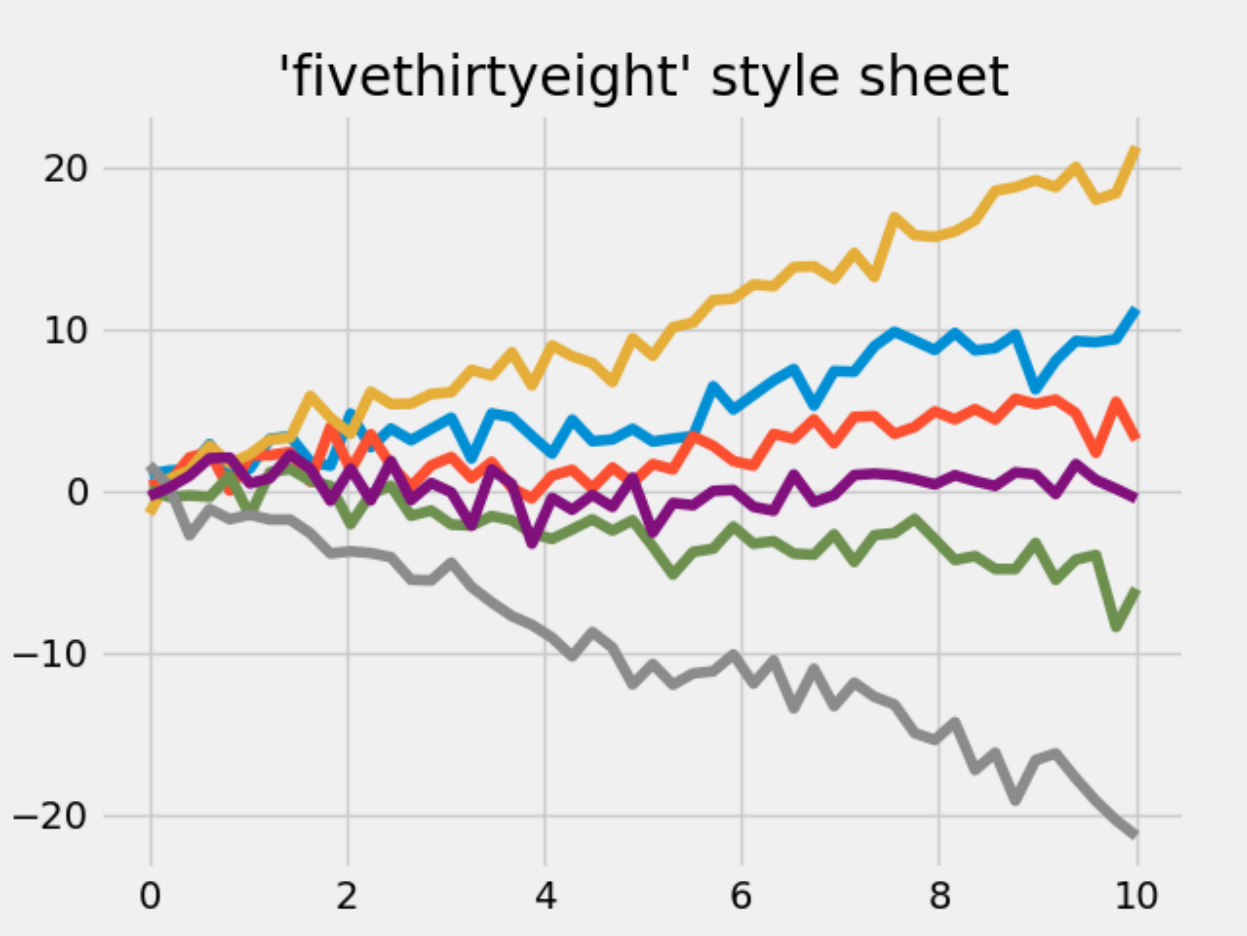
The 'Purchase' column contains the amount spent by the customer during the Black Friday sales. It is a numerical value representing the purchase amount in the local currency.

first step :

```
# find unique values using lambda function :
df.apply(lambda x: len(x.unique()))
```

# FiveThirtyEight style sheet

This shows an example of the "fivethirtyeight" styling, which tries to replicate the styles from FiveThirtyEight.com.




Here are some key points about central tendency measures:

- Outliers will affect only the mean value of the data and less affect the median or mode value of data.
- The median value is useful for skewed data such as income data.
- The mean value is useful for symmetric data because, in the case of skewed data, it gets influence by outliers.
- If the mean=median=mode then we can say that the data is symmetrically distributed about the mean. So the greater the difference among these measures the more asymmetrical the data.

Data Preprocessing

data integration?

Data preprocessing is a important step in the machine learning where raw data is transformed, cleaned, and organized to enhance its quality and prepare it for training machine learning models. The goal of data preprocessing is to make the data more suit for training the machine learning model by removing noise, handling missing values, and addressing other issues that might affect the performance of machine learning algorithms.

## Missing Values Imputation:

**Average_Age = 26.0**

| ID | City | Age | Married ? |
|---|---|---|---|
| 1 | Lisbon | 25 | 0 |
| 2 | Berlin | 25 | 1 |
| 3 | Lisbon | 30 | 1 |
| 4 | Lisbon | 30 | 1 |
| 5 | Berlin | 18 | 0 |
| 6 | Lisbon | NaN | 0 |
| 7 | Berlin | 30 | 1 |
| 8 | Berlin | NaN | 0 |
| 9 | Berlin | 25 | 1 |
| 10 | Madrid | 25 | 1 |

| ID | City | Age | Married ? |
|---|---|---|---|
| 1 | Lisbon | 25 | 0 |
| 2 | Berlin | 25 | 1 |
| 3 | Lisbon | 30 | 1 |
| 4 | Lisbon | 30 | 1 |
| 5 | Berlin | 18 | 0 |
| 6 | Lisbon | 26 | 0 |
| 7 | Berlin | 30 | 1 |
| 8 | Berlin | 26 | 0 |
| 9 | Berlin | 25 | 1 |
| 10 | Madrid | 25 | 1 |

Why we impute missing values:

- **Improving Model Performance.**

- **Compatibility with Algorithms (Many machine learning algorithms and statistical models require complete datasets to function properly).**