Data Collection and Preprocessing Phase

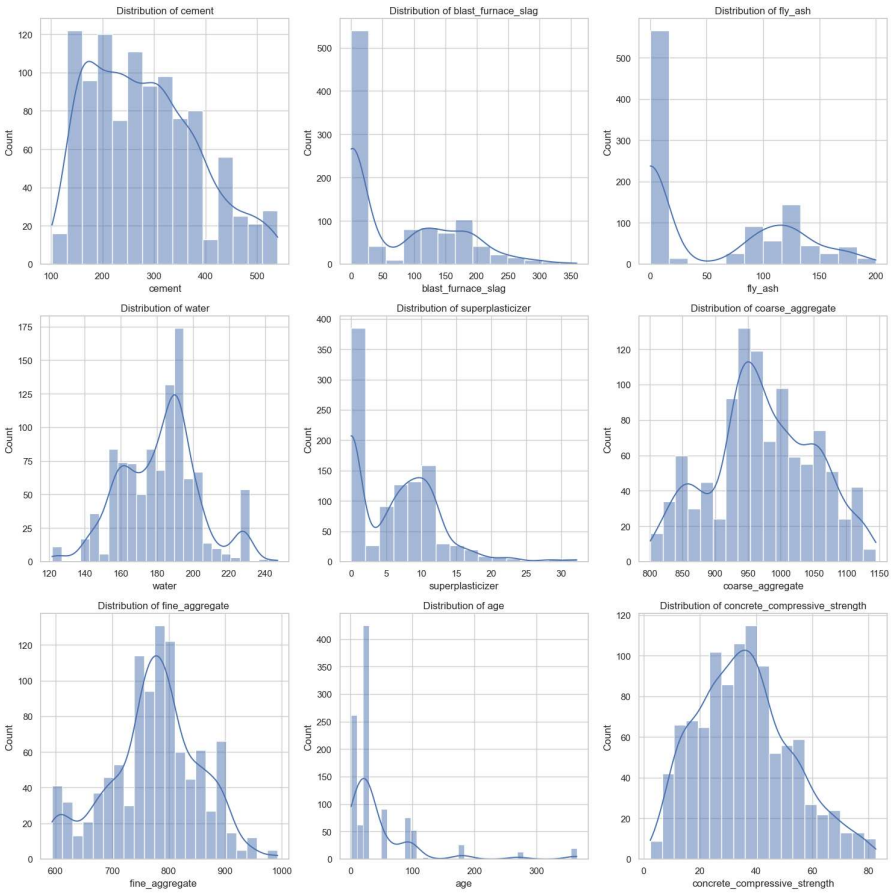
Date	14 July 2024
Team ID	SWTID1720171884
Project Title	Predicting Compressive Strength Of Concrete Using Machine Learning
Maximum Marks	6 Marks

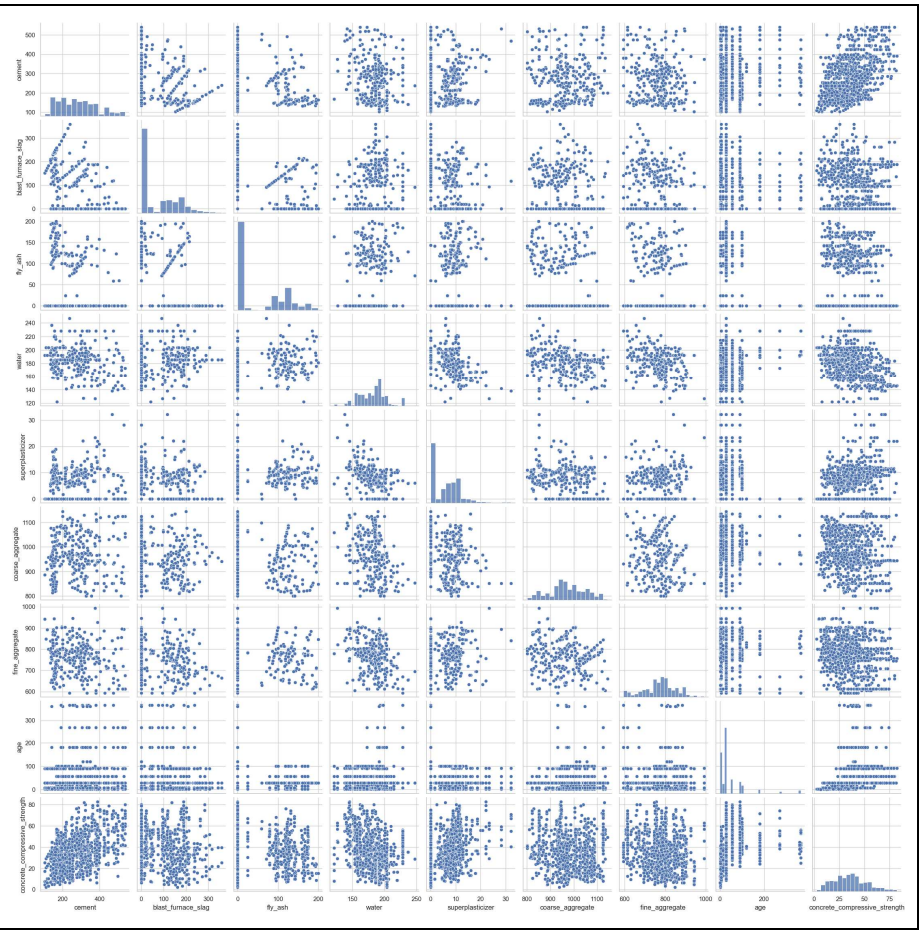
Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

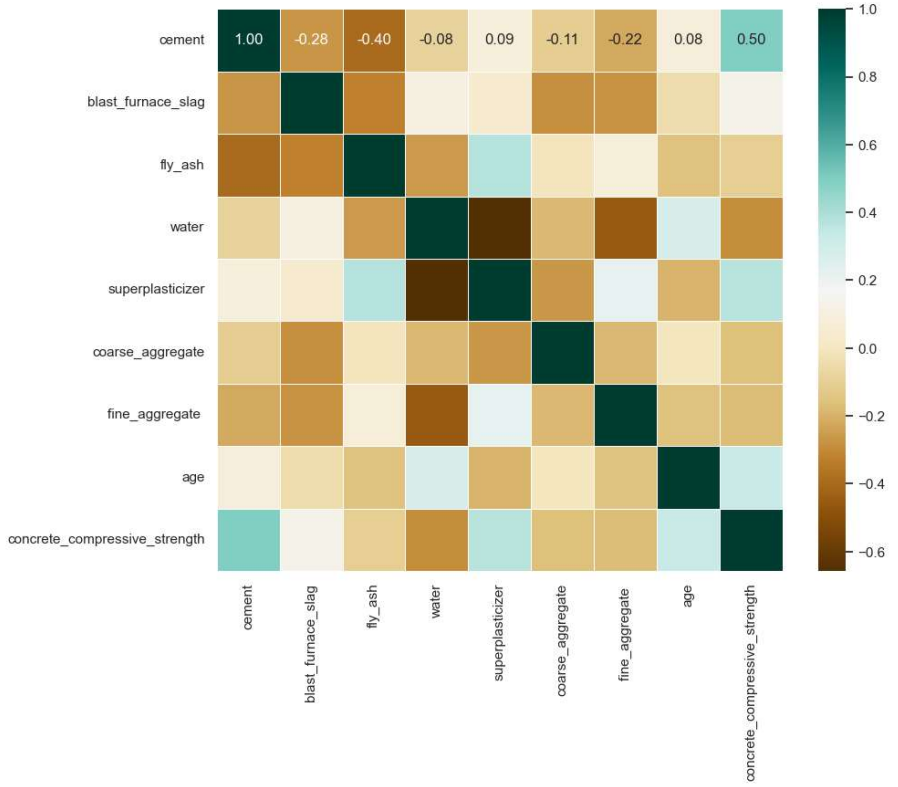
Section	Description
Data Overview	Dimension: 1030 x 9 Descriptive statistics: [9]:

	cement	blast_furnace_slag	fly_ash	water	superplasticizer	coarse_aggregate	fine_aggregate	age	concrete_compressive_strength
count	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000	1030.000000
mean	281.167864	73.895825	54.188350	181.567282	6.204660	972.918932	773.580485	45.662136	35.817961
std	104.506364	86.279342	63.997004	21.354219	5.973841	77.753954	80.175980	63.169912	16.705742
min	102.000000	0.000000	0.000000	121.800000	0.000000	801.000000	594.000000	1.000000	2.330000
25%	192.375000	0.000000	0.000000	164.900000	0.000000	932.000000	730.950000	7.000000	23.710000
50%	272.900000	22.000000	0.000000	185.000000	6.400000	968.000000	779.500000	28.000000	34.445000
75%	350.000000	142.950000	118.300000	192.000000	10.200000	1029.400000	824.000000	56.000000	46.135000
max	540.000000	359.400000	200.100000	247.000000	32.200000	1145.000000	992.600000	365.000000	82.600000

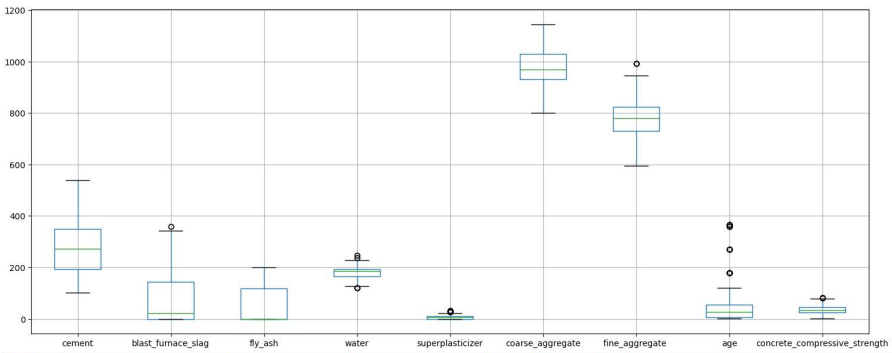
Univariate Analysis	

Bivariate Analysis	

Multivariate Analysis



Outliers and Anomalies




```python
# Define a function to handle outliers using the IQR method
def handle_outliers(df):
    for column in df.columns:
        if df[column].dtype != 'object':
            Q1 = df[column].quantile(0.25)
            Q3 = df[column].quantile(0.75)
            IQR = Q3 - Q1
            lower_bound = Q1 - 1.5 * IQR
            upper_bound = Q3 + 1.5 * IQR
            df[column] = np.where(df[column] < lower_bound, lower_bound, df[column])
            df[column] = np.where(df[column] > upper_bound, upper_bound, df[column])
    return df
```

Data Preprocessing Code Screenshots

Loading Data	`data=pd.read_csv('concrete_data.csv')`

Handling Missing Data	`data.isnull().sum()`
	`[13]:`
	cement 0
	blast_furnace_slag 0
	fly_ash 0
	water 0
	superplasticizer 0
	coarse_aggregate 0
	fine_aggregate 0
	age 0
	concrete_compressive_strength 0
	dtype: int64
	No null values in the data

Data Transformation	No need of transformation as there are no categorical data in the dataset
Feature Engineering	All 8 features are considered as they are the components of concrete mixture.
Save Processed Data	`: data.to_csv('preprocessed_data.csv')`