<u>Statistics Assignment 2</u>

Prepared by Sohana Tasneem (Batch 2412)

Python Code File : Statistics A2.ipynb

Task 1: Find the following information of all the features: (40 marks)
Mean, Median, Mode, Range, Variance, Standard deviation, Interquartile range, skewness and kurtosis.
Bonus point: If you use a for loop to obtain all the values.

Code :

```
# Clean column names
df.columns = [col.strip() for col in df.columns]

# Initialize list to collect rows
stats_list = []

# Loop through columns and compute stats
for col in df.columns:
    data = df[col]
    stats_list.append([
        col,
        round(data.mean(), 3),
        round(data.median(), 3),
        round(mode(data, keepdims=False).mode, 3) if not data.empty else
np.nan,
        round(data.max() - data.min(), 3),
        round(data.var(), 3),
        round(data.std(), 3),
        round(data.quantile(0.75) - data.quantile(0.25), 3),
        round(skew(data), 3),
        round(kurtosis(data), 3)
    ])

# Create a DataFrame from the list
stats_df = pd.DataFrame(stats_list, columns=[
    'Feature',
    'Mean',
```

```python
    'Median',
    'Mode',
    'Range',
    'Variance',
    'Standard Deviation',
    'IQR',
    'Skewness',
    'Kurtosis'
])

# Display the final formatted table
print("Descriptive Statistics Table:\n")
print(stats_df.to_string(index=False))
```

Output :

Descriptive Statistics Table:

Feature	Mean	Median	Mode	Range	Variance	Standard Deviation	IQR	Skewness	Kurtosis
Cement (component 1)(kg in a m^3 mixture)	281.168	272.900	362.6	438.00	10921.580	104.506	157.625	0.509	-0.524
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	73.896	22.000	0.0	359.40	7444.125	86.279	142.950	0.800	-0.512
Fly Ash (component 3)(kg in a m^3 mixture)	54.188	0.000	0.0	200.10	4095.617	63.997	118.300	0.537	-1.328
Water (component 4)(kg in a m^3 mixture)	181.567	185.000	192.0	125.20	456.003	21.354	27.100	0.075	0.116
Superplasticizer (component 5)(kg in a m^3 mixture)	6.205	6.400	0.0	32.20	35.687	5.974	10.200	0.906	1.399
Coarse Aggregate (component 6)(kg in a m^3 mixture)	972.919	968.000	932.0	344.00	6045.677	77.754	97.400	-0.040	-0.602
Fine Aggregate (component 7)(kg in a m^3 mixture)	773.580	779.500	594.0	398.60	6428.188	80.176	93.050	-0.253	-0.108
Age (day)	45.662	28.000	28.0	364.00	3990.438	63.170	49.000	3.264	12.104
Concrete compressive strength(MPa, megapascals)	35.818	34.445	33.4	80.27	279.082	16.706	22.425	0.416	-0.318

Task 2: Explain all the features whether each feature shows what kind of characteristics based on the measures you got from question 01. This must include the characteristics such as their central tendency, spread, and shape of the data. (30 marks)

Below is the interpretation of each feature's characteristics, including **central tendency**, **spread**, and **shape**, based on the statistics computed in Task 1.

1. Cement (kg/m^3)

- Values are fairly evenly distributed, with no extreme skew.
- Most mixes use a moderate amount of cement.
- Indicates balanced and consistent usage across the dataset.

2. Blast Furnace Slag (kg/m^3)

- Most values are low or zero, with a few mixtures containing high amounts.
- This creates a right-skewed distribution.
- Suggested slag is optional and used selectively.

3. Fly Ash (kg/m^3)

- Frequently not used in mixes (many values are zero).
- When present, it varies widely in amount.
- Distribution is heavily right-skewed due to infrequent but high values.

4. Water (kg/m^3)

- Values are centered and consistent across samples.
- The distribution is nearly symmetrical.
- Indicates standard water usage in most concrete mixes.

5. Superplasticizer (kg/m^3)

- Many samples do not contain any (lots of zeros).
- Used in varying quantities when present.
- Right-skewed distribution shows it's not essential in all mixes, but impactful when used.

6. Coarse Aggregate (kg/m³)

- Values are tightly clustered with little skew.
- Suggests uniform use of coarse aggregate across all mixtures.
- Reflects standard design practice in concrete formulation.

7. Fine Aggregate (kg/m³)

- Distribution is centered with a slight left skew.
- Indicates consistency in usage, similar to coarse aggregate.
- Most mixes contain moderate amounts.

8. Age (days)

- Most tests are conducted early (e.g., at 7 or 28 days).
- A few are tested at much later ages (up to 365 days), causing a long right tail.
- Strong right-skewed distribution reflects industry testing practices.

9. Concrete Compressive Strength (MPa)

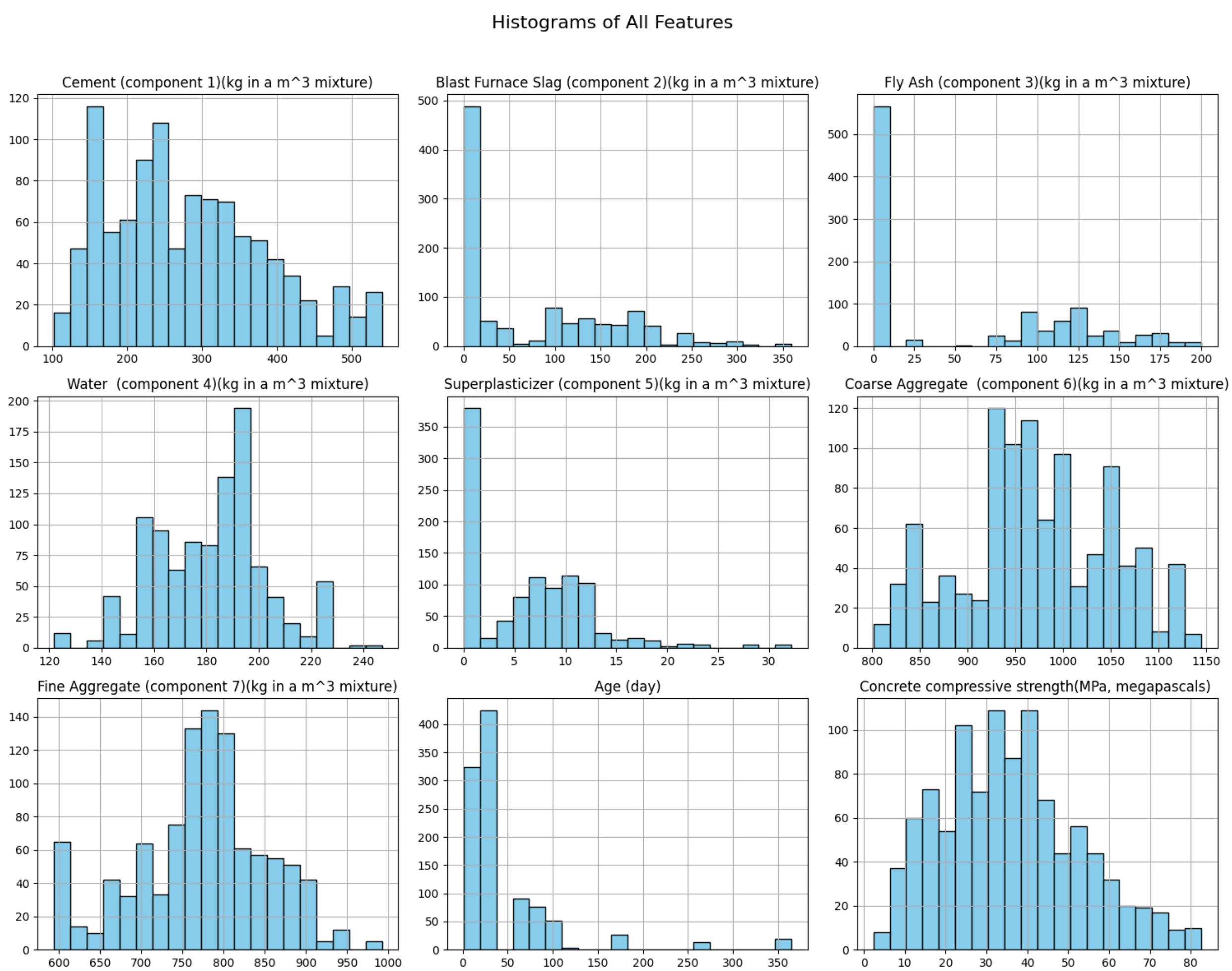
- Distribution is slightly right-skewed.
- Most samples fall in a moderate strength range, with a few high-strength outliers.
- Indicates that while variation exists, extreme values are rare.

Task 3: Draw figure of the distributions of all the features such as histogram, scatter plot, boxplot and explain the diagrams. (30 marks)

Code:

```
# Plot histograms
df.hist(bins=20, figsize=(15, 12), edgecolor='black', color='skyblue')
plt.suptitle("Histograms of All Features", fontsize=16)
plt.tight_layout(rect=[0, 0, 1, 0.96])
plt.show()
```

Output:



Explanation:

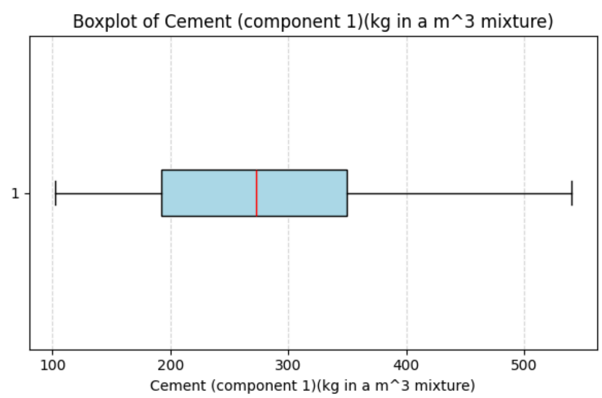
- The histograms provide an overview of the distribution of values for each feature in the dataset.

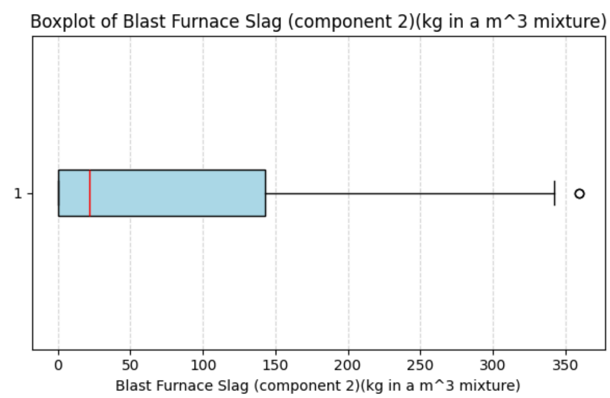
- Some features like **Water** exhibit a nearly symmetric distribution, suggesting consistent usage across concrete mixes.
- Others, such as **Fly Ash**, **Superplasticizer**, and especially **Age**, show **strong right skewness**, with most values clustered at the lower end and a few stretching into much higher ranges.
- **Cement** is slightly right-skewed, indicating a tendency for mixes to use moderate to high amounts.
- These histograms are useful for identifying distribution patterns, skewness, common value ranges or gaps in the data.

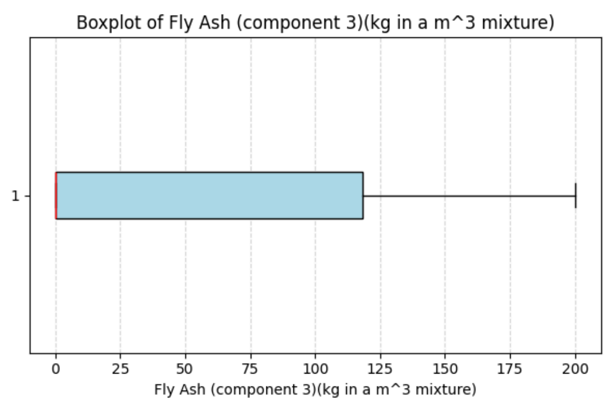
Code:

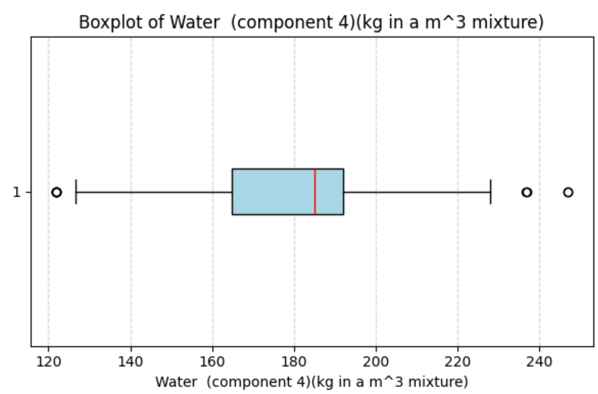
```
# Create a boxplot for each feature with labels
for col in df.columns:
    plt.figure(figsize=(6, 4))
    plt.boxplot(df[col], vert=False, patch_artist=True,
                boxprops=dict(facecolor='lightblue', color='black'),
                medianprops=dict(color='red'))
    plt.title(f"Boxplot of {col}")
    plt.xlabel(col)
    plt.grid(axis='x', linestyle='--', alpha=0.5)
    plt.tight_layout()
    plt.show()
```

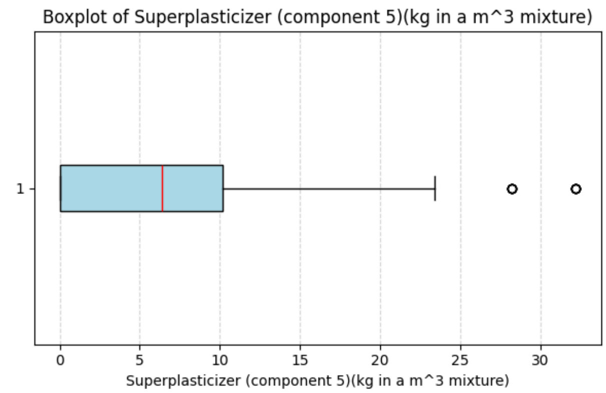
Output:

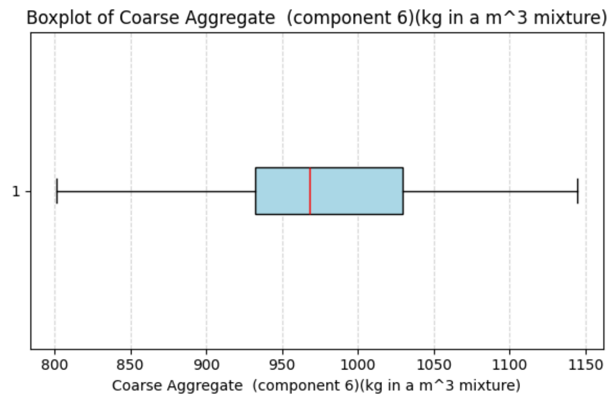


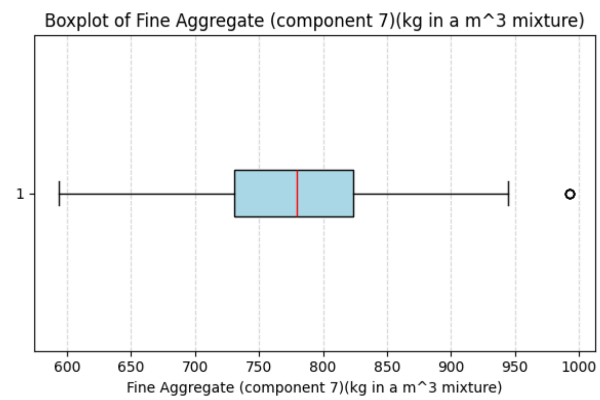


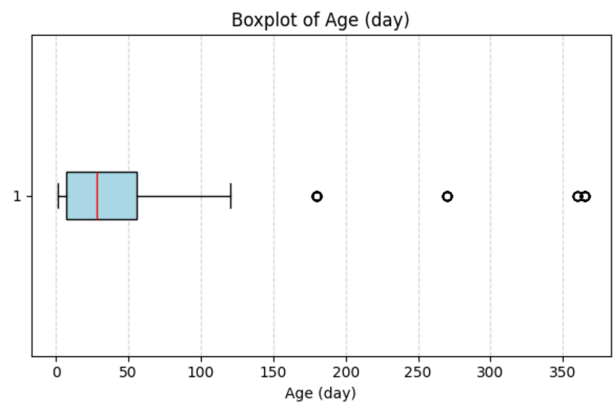


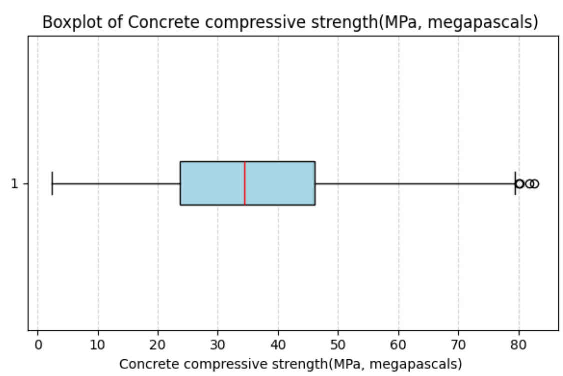












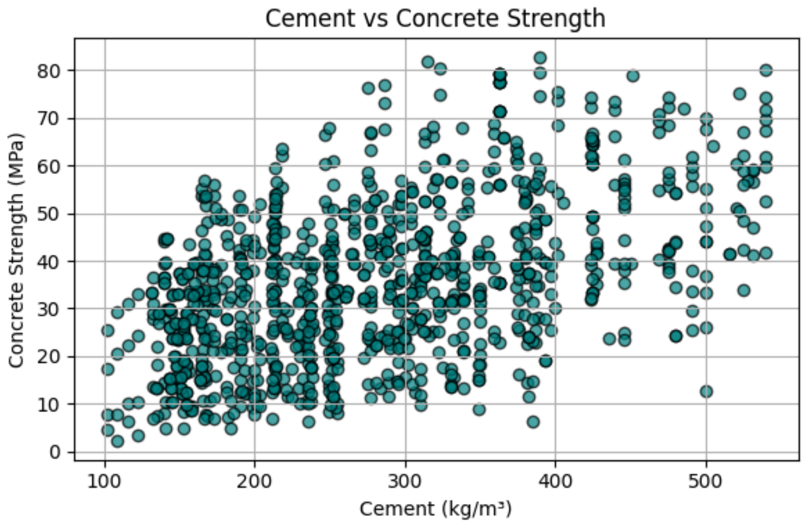
Explanation:

- The boxplots visualize the spread and central tendency of each feature, along with potential **outliers**.
- Features such as **Superplasticizer**, **Fly Ash**, and **Age** show clear outliers, while features like **Coarse Aggregate** and **Water** appear more stable and tightly distributed.
- This helps identify variability and potential anomalies in the dataset.

Code:

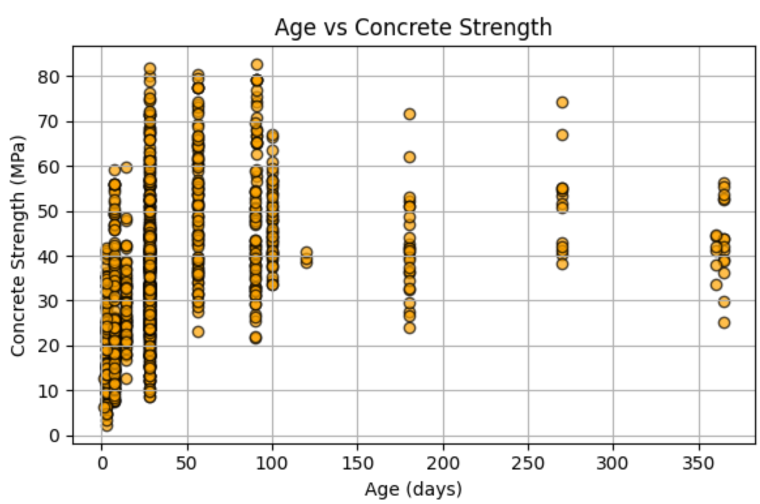
```python
# Scatter Plot 1: Cement vs Strength
plt.figure(figsize=(6, 4))
plt.scatter(df['Cement (component 1)(kg in a m^3 mixture)'], df['Concrete compressive strength(MPa, megapascals)'],
            alpha=0.7, edgecolor='black', color='teal')
plt.xlabel("Cement (kg/m³)")
plt.ylabel("Concrete Strength (MPa)")
plt.title("Cement vs Concrete Strength")
plt.grid(True)
plt.tight_layout()
plt.show()

# Scatter Plot 2: Age vs Strength
plt.figure(figsize=(6, 4))
plt.scatter(df['Age (day)'], df['Concrete compressive strength(MPa, megapascals)'],
            alpha=0.7, edgecolor='black', color='orange')
plt.xlabel("Age (days)")
plt.ylabel("Concrete Strength (MPa)")
plt.title("Age vs Concrete Strength")
plt.grid(True)
plt.tight_layout()
plt.show()

# Scatter Plot 3: Water vs Strength
plt.figure(figsize=(6, 4))
plt.scatter(df['Water  (component 4)(kg in a m^3 mixture)'], df['Concrete compressive strength(MPa, megapascals)'],
            alpha=0.7, edgecolor='black', color='purple')
plt.xlabel("Water (kg/m³)")
plt.ylabel("Concrete Strength (MPa)")
plt.title("Water vs Concrete Strength")
plt.grid(True)
plt.tight_layout()
plt.show()
```

Output:



Cement vs Concrete Strength

The scatter plot shows a generally **positive correlation** — as the amount of cement increases, concrete strength tends to increase as well.
This makes sense as **cement** is the **primary binding component** that contributes to strength.

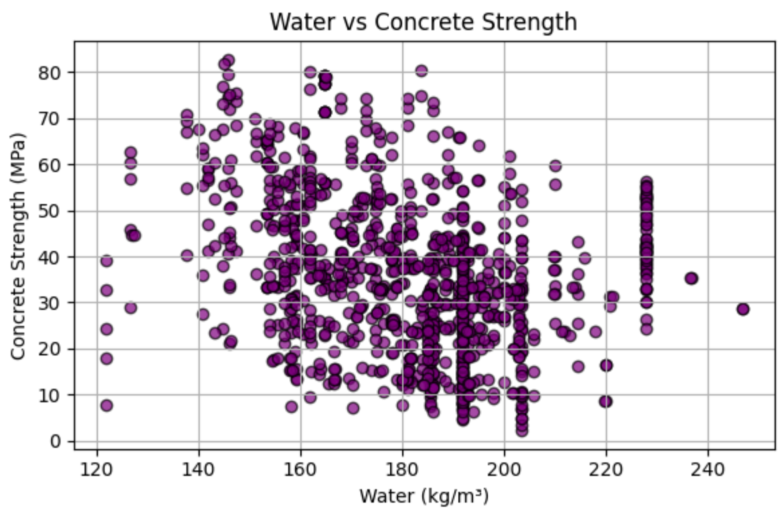


Age vs Concrete Compressive Strength

The scatter plot reveals a **clear upward trend**, indicating a **positive correlation** between the two variables.
This means that as the **age of the concrete increases**, its **compressive strength generally improves.**

This pattern confirms a fundamental principle in concrete science: **the longer the concrete is allowed to cure, the stronger it becomes**, as hydration reactions continue over time.

The plot also highlights that while most samples are tested early (e.g., 7 or 28 days), a smaller number of samples extend into high-age values. This imbalance explains the **right-skewed distribution** of the Age variable observed in earlier analysis.



Water vs Concrete Compressive Strength

This shows a **noticeable negative relationship** between the two variables.As the **amount of water** in the concrete mix **increases,** the **strength of the concrete generally decreases.**

On the plot, we can see that many high-strength samples are associated with **lower water content**, while mixes with **higher water content** tend to produce **weaker concrete.**

The relationship is not perfectly linear — due to the presence of other influencing variables like cement content or additives — but the overall **downward slope** is clear.