Assignment 1

Part 1: Graphical Analysis

1. Scatter Diagram (10 points)

a. Choose a dataset (from UCI Machine Learning Repository).
b. Create a scatter diagram to visualize the relationship between two numerical variables.
c. Describe the relationship you observe. Is it linear, non-linear, or no correlation? Provide a brief interpretation of your findings.

Solution

The dataset used in this assignment is [Individual Household Electric Power Consumption] by *Georges Hebrail* and *Alice Berard*

The data was loaded from a .txt file into Jupyter notebook using pandas library.

```
# Getting the data from the dataset

import pandas as pd

df = pd.read_csv(
    "household_power_consumption.txt",
    sep=';',
    na_values='?',
    low_memory=False
)
```

```
[2]: df.shape

[2]: (2075259, 9)
```

The data was examined to see the presence of null values that could interfere with the analysis. There was the presence of missing values and they were handled by removing them from the dataset.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075259 entries, 0 to 2075258
Data columns (total 9 columns):
 #   Column                 Dtype
---  ------                 -----
 0   Date                   object
 1   Time                   object
 2   Global_active_power    float64
 3   Global_reactive_power  float64
 4   Voltage                float64
 5   Global_intensity       float64
 6   Sub_metering_1         float64
 7   Sub_metering_2         float64
 8   Sub_metering_3         float64
dtypes: float64(7), object(2)
memory usage: 142.5+ MB
```

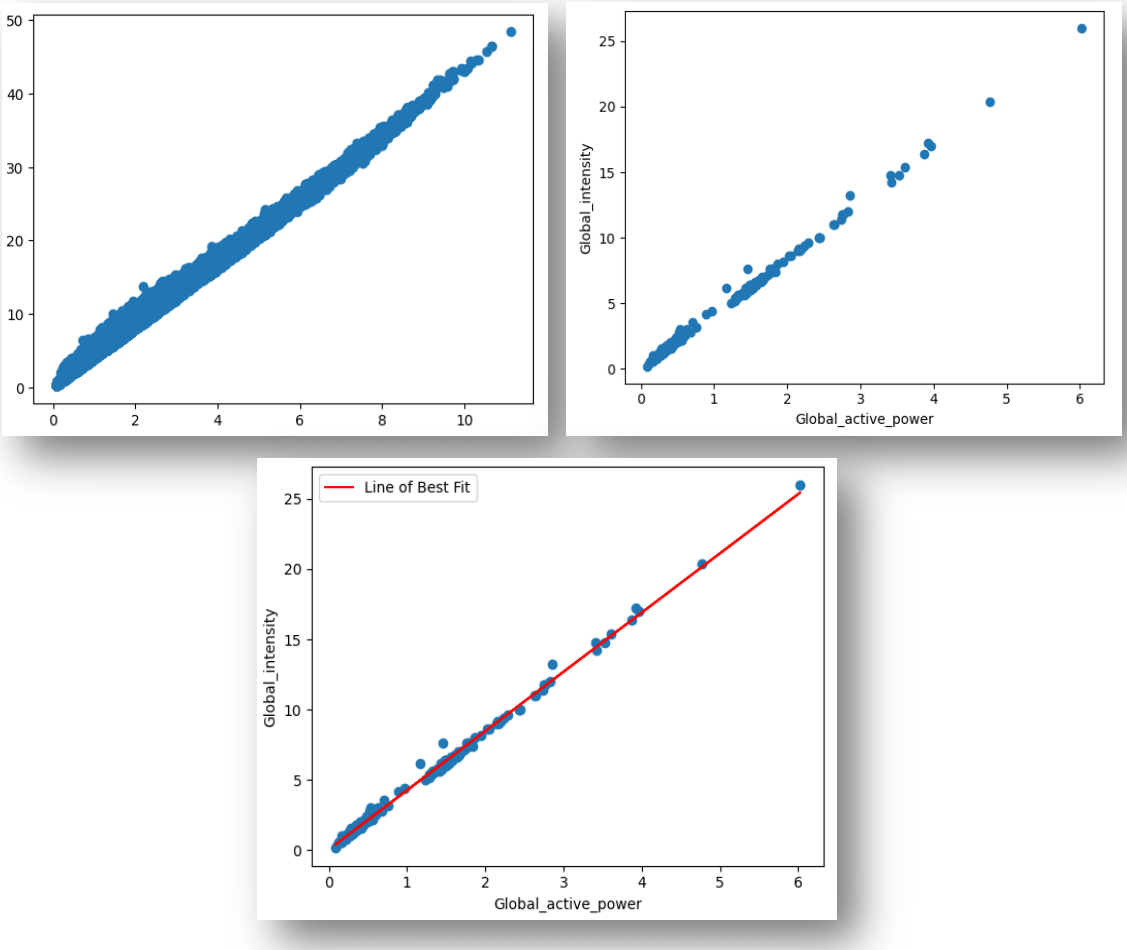
```
df.isnull().sum()

Date                       0
Time                       0
Global_active_power    25979
Global_reactive_power  25979
Voltage                25979
Global_intensity       25979
Sub_metering_1         25979
Sub_metering_2         25979
Sub_metering_3         25979
dtype: int64
```

```
Percentage of data loss = 1.25 %
```

Scatter diagram was created between two numeric variables named Global_active_power and Global_intensity.

Figure below shows the scatter plot on population data, sample data of sample size 200 and deviation of points from the best fit line.



Interpretation:

From the above scatter plot of Global_active_power vs Global_intensity, we can conclude that the variables have strong positive linear correlation. Global_active_power while may not be the only factor that can predict Global_intensity values but it is a strong predictor.

2. Q-Q Plot (10 points)

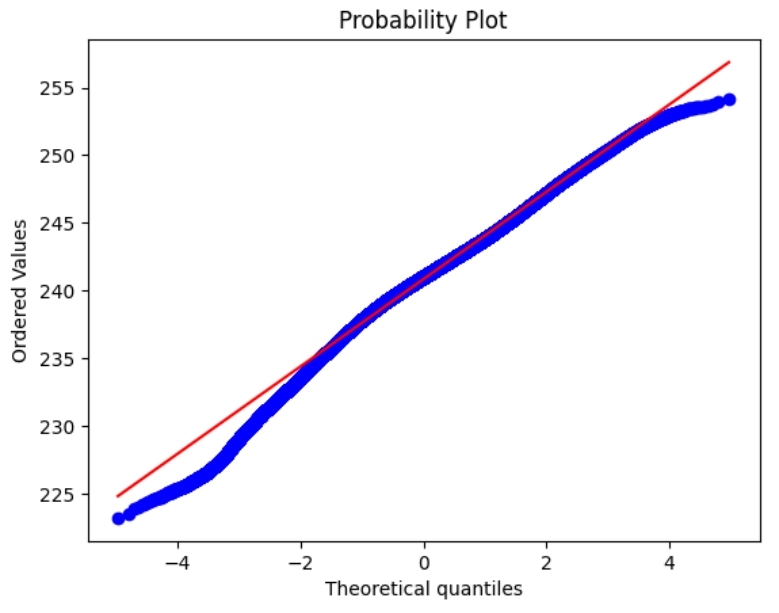
a. Using the same dataset, create a Q-Q plot to assess the normality of one of the numerical variables.

b. Explain what a Q-Q plot is and how to interpret it. Does the data appear to follow a normal distribution? Justify your answer.

Solution

Voltage was chosen as the test variable for the normality test using the Q-Q plot

The following plot shows the plot of normal line of the normal distribution against that of our test data's distribution.



Does the above data show normal distribution?

The data is approximately normally distributed in the center, but shows systematic deviations in the tails, indicating mild non-normality, likely due to skewness.

Justification:

Middle section aligns well with the red line:

- Around theoretical quantiles -1 to +3 Points closely follow the line
- Bulk of the data is roughly normal

What is Q-Q plot ?

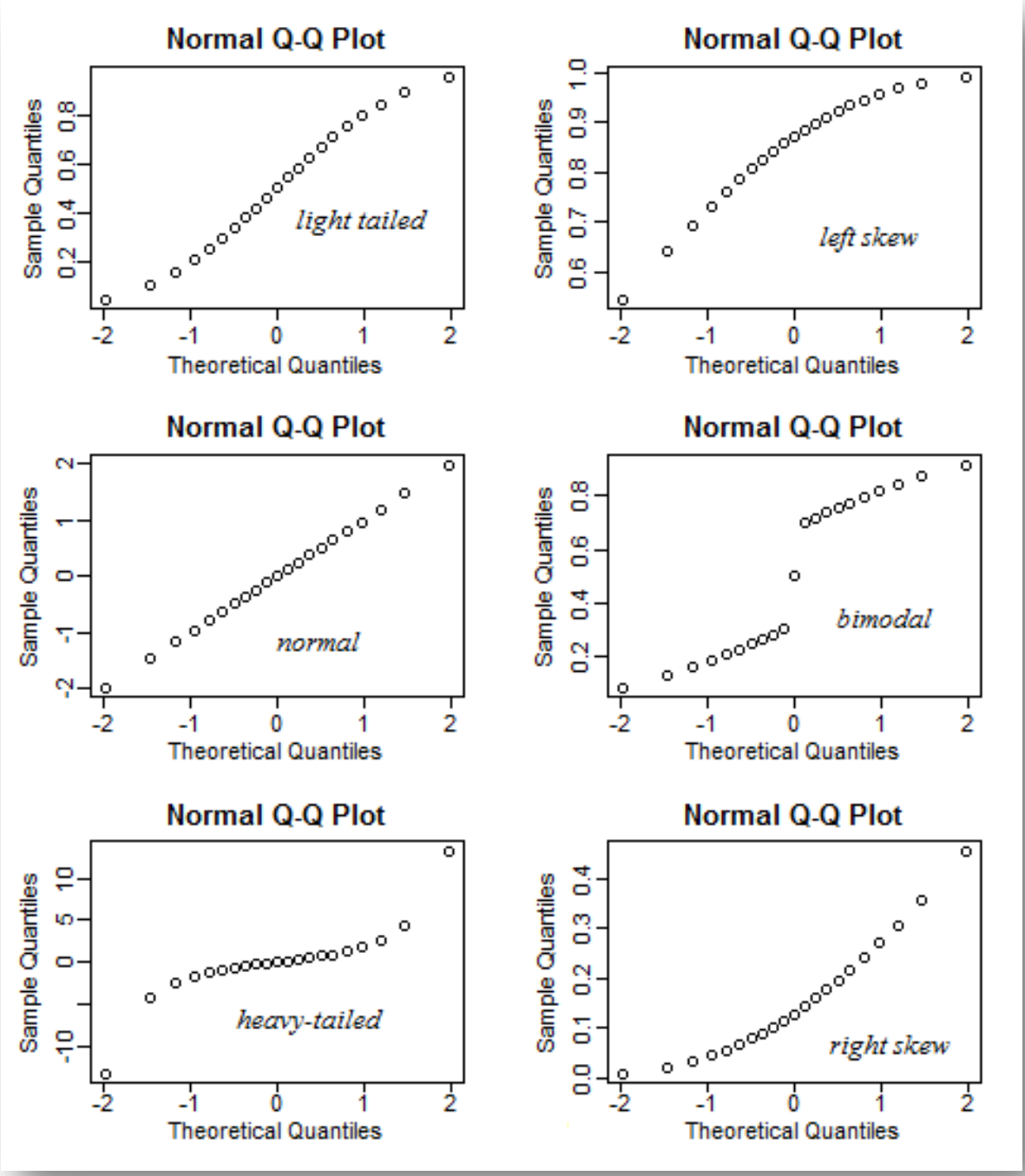
A Q-Q plot (Quantile-Quantile plot) is a plot of quantiles of test dataset against the quantiles of a theoretical distribution to visually check if the test data fits that distribution, with points forming a straight line indicating a good match, while deviations indicating skewness.

How to Interpret It

1. Perfect Fit
   - If the data perfectly matches the theoretical distribution, all points fall exactly on a straight, diagonal reference line (often $y = x$).
2. Normal Distribution (Common Use)
   - Linear Pattern: Points generally lying on a straight line suggest approximate normality.
3. Skewness
   - Right Skew (Positive Skew): An upward curve or S-shape at the end (upper tail).
   - Left Skew (Negative Skew): A downward curve or S-shape at the beginning (lower tail).
4. Kurtosis (Tail Behavior)
   - Heavy Tails (Leptokurtic): Points deviate upwards at both ends, indicating more extreme values than a normal distribution.
   - Light Tails (Platykurtic): Points deviate downwards at both ends, indicating fewer extreme values than a normal distribution.

Below is a diagram that shows what QQ-plots look like (for particular choices of distribution) on average:

Source: [stackexchange]



3. Box Plots (10 points)

a. Generate box plots for the numerical variable across different categories of a categorical variable in your dataset.

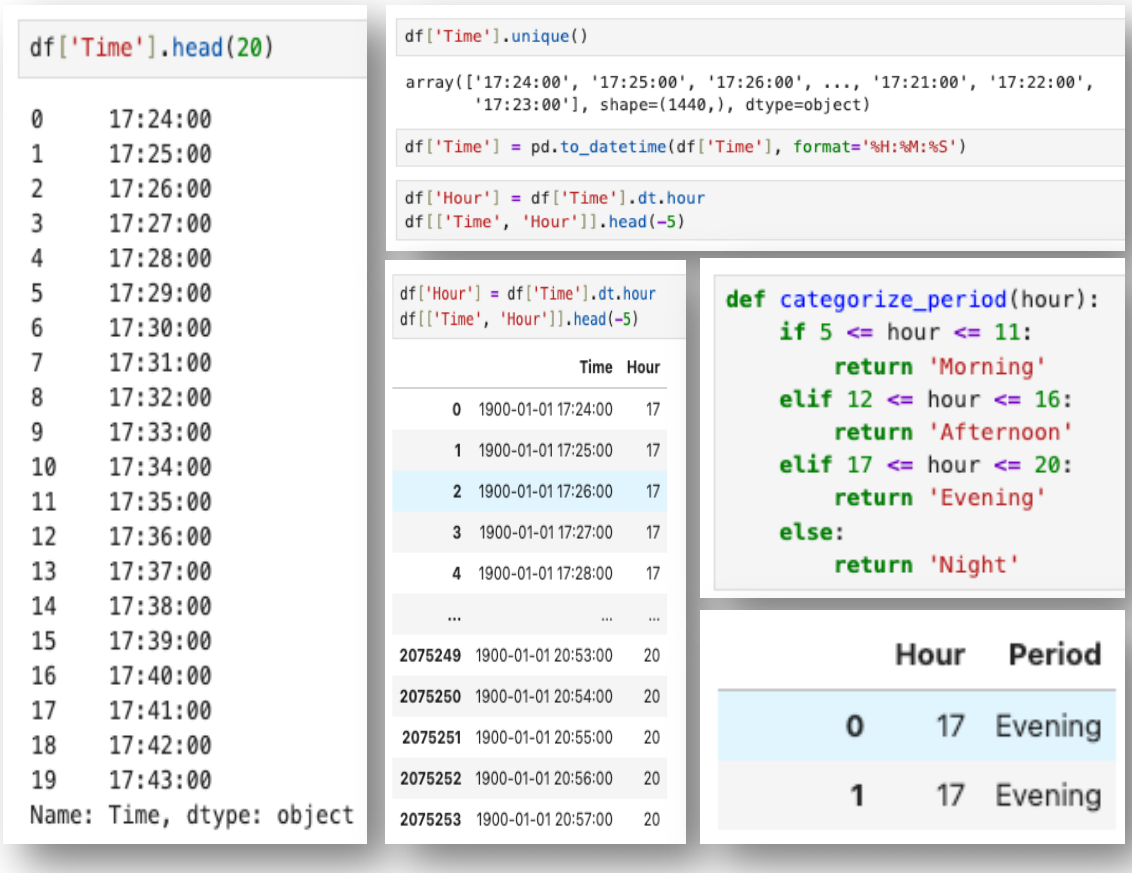
b. Analyze the box plots and discuss any outliers, the spread of the data.

Solution

The numeric variable for this test was chosen to be Global_active_power. It as a continuous variable. The categorical variable chosen was Time.

Since we are working with roughly 2 million Time data, where the unique timestamps are 1440, we need a way to categorize them into smaller quantity.

For scaling down this huge 1440 unique value, the values were first changed to just the hour value i.e. the timestamp 17:24:00 will now be just 17. This will give us 24 unique values. After that the time was changed to period like Morning, Afternoon, Evening, Night. Resulting in 4 categorical variables.



```
df['Time'].head(20)

0     17:24:00
1     17:25:00
2     17:26:00
3     17:27:00
4     17:28:00
5     17:29:00
6     17:30:00
7     17:31:00
8     17:32:00
9     17:33:00
10    17:34:00
11    17:35:00
12    17:36:00
13    17:37:00
14    17:38:00
15    17:39:00
16    17:40:00
17    17:41:00
18    17:42:00
19    17:43:00
Name: Time, dtype: object
```

```
df['Time'].unique()

array(['17:24:00', '17:25:00', '17:26:00', ..., '17:21:00', '17:22:00',
       '17:23:00'], shape=(1440,), dtype=object)

df['Time'] = pd.to_datetime(df['Time'], format='%H:%M:%S')

df['Hour'] = df['Time'].dt.hour
df[['Time', 'Hour']].head(-5)
```

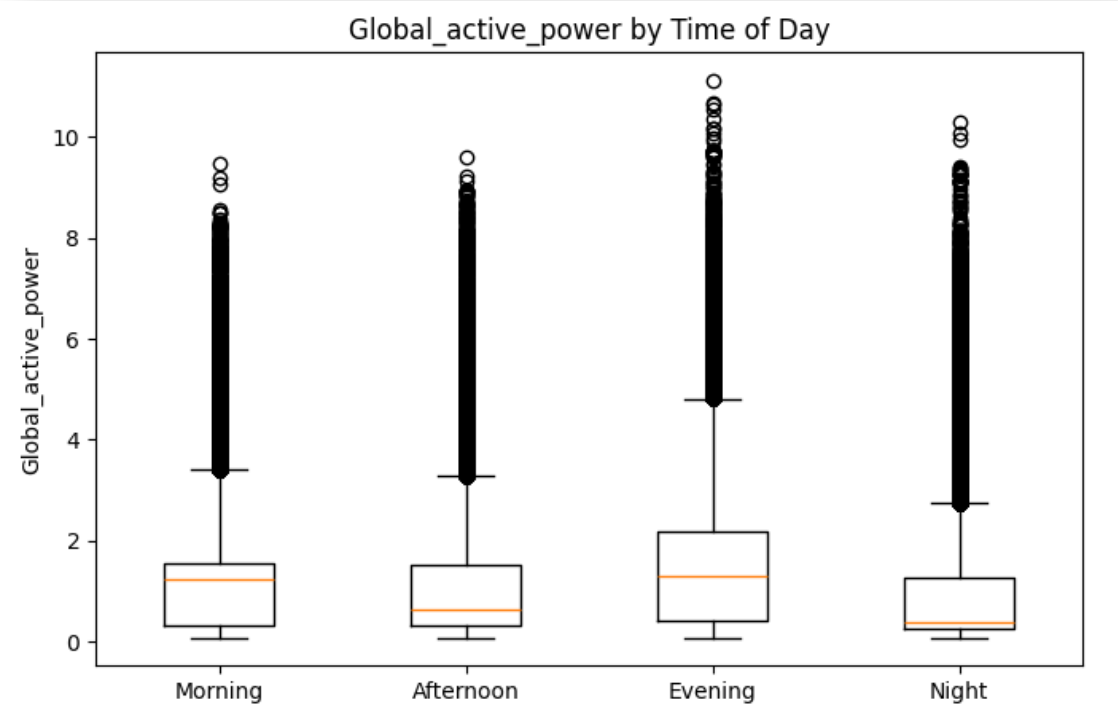
```
df['Hour'] = df['Time'].dt.hour
df[['Time', 'Hour']].head(-5)
```

	Time	Hour
0	1900-01-01 17:24:00	17
1	1900-01-01 17:25:00	17
2	1900-01-01 17:26:00	17
3	1900-01-01 17:27:00	17
4	1900-01-01 17:28:00	17
...	...	...
2075249	1900-01-01 20:53:00	20
2075250	1900-01-01 20:54:00	20
2075251	1900-01-01 20:55:00	20
2075252	1900-01-01 20:56:00	20
2075253	1900-01-01 20:57:00	20

```
def categorize_period(hour):
    if 5 <= hour <= 11:
        return 'Morning'
    elif 12 <= hour <= 16:
        return 'Afternoon'
    elif 17 <= hour <= 20:
        return 'Evening'
    else:
        return 'Night'
```

	Hour	Period
0	17	Evening
1	17	Evening

Finally, after creating the categorical variable Period from the Time column, box plot was plotted against the numeric variable Global_active_power



Analysis of the plot:

1. The spread of the data
   - The spread of the data is indicated by the Interquartile Range and the length of the whiskers.
   - Evening Peak: The Evening category shows the largest spread. Both the box and the upper whisker are significantly taller than the other time periods. indicating that the power consumption is generally higher during these hours.
   - Night Lows: The Night category has the smallest spread and the lowest median. The box is compressed toward the bottom, indicating that for most of the night, power usage remains low.
   - Morning vs. Afternoon: These two are relatively similar in their range, though the Morning has a slightly higher median and a more balanced distribution within the box compared to the Afternoon.
   -

2. Presence of Outliers
    - Positive Skew: All distributions are heavily right-skewed. While the "normal" consumption stays mostly below 3 units, there are frequent instances where power usage spikes up to 8, 10, or even 11 units.
    - Extreme Events: The Evening session contains the most extreme outliers, reaching the highest point on the graph (above 11). Representing times when multiple high energy appliances are running simultaneously.
    - The Night Exception: Even though Night has the lowest average usage, it still shows a massive column of outliers. This suggests that while most people are asleep, certain households or specific nights involve high energy activity.