CMSC178DA

*Activity 2*

# EDA by Calculating 5-Number Summary and Constructing Box-and-Whiskers Plot

Submitted by:

Peladas, Daenielle Rai

Section B (TF 10:30 AM - 12:00 PM)

# The Case

A data analyst was given the task by the head of office to look into the stability of the network connections in two (2) locations of their office. She gathered the following samples of data for round-trip transit times (also known colloquially as “pings”) from the two locations:

**Location I:** 1.6, 4.0, 3.6, 4.8, 4.2, 3.4, 5.4, 3.5, 2.8, 2.1, 3.1, 3.3, 2.2, 4.4, 4.8 seconds
**Location II:** 3.0, 6.7, 3.9, 7.9, 4.7, 7.1, 0.5, 1.7, 6.6, 1.1, 2.0, 0.4 seconds

*The data analyst poses this question to herself: which of the two (2) locations in their office has an unstable network connection?*

# The 5-Number Summary

The pandas library was imported for data manipulation and analysis.

Python
```python
import pandas as pd
```

Two dictionaries were created to store ping rate data for Location 1 and Location 2. Each dictionary was then converted into a separate DataFrame using the DataFrame constructor.

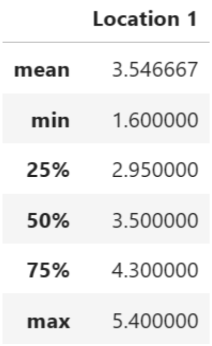
Python
```python
ping_location_1 = {"Location 1": [1.6, 4.0, 3.6, 4.8, 4.2, 3.4, 5.4, 3.5,
2.8, 2.1, 3.1, 3.3, 2.2, 4.4, 4.8]}
ping_location_2 = {"Location 2": [3.0, 6.7, 3.9, 7.9, 4.7, 7.1, 0.5, 1.7,
6.6, 1.1, 2.0, 0.4]}

df_loc1 = pd.DataFrame(ping_location_1)
df_loc2 = pd.DataFrame(ping_location_2)
```

The `describe` function was applied to each DataFrame to generate descriptive statistics. The `loc` attribute was then used to isolate specific metrics for the 5-number summary (mean, minimum, 25th percentile, 50th percentile, 75th percentile, and maximum), providing a concise overview of the ping rate distributions for each location.

Python

```python
df_loc1.describe().loc[['mean', 'min', '25%', '50%', '75%', 'max']]
```

*Table 1. The 5-Number Summary of Location 1.*

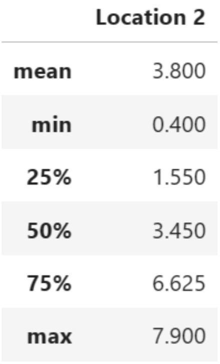


| | Location 1 |
|---|---|
| **mean** | 3.546667 |
| **min** | 1.600000 |
| **25%** | 2.950000 |
| **50%** | 3.500000 |
| **75%** | 4.300000 |
| **max** | 5.400000 |

Python

```python
df_loc2.describe().loc[['mean', 'min', '25%', '50%', '75%', 'max']]
```

*Table 2. The 5-Number Summary of Location 2*

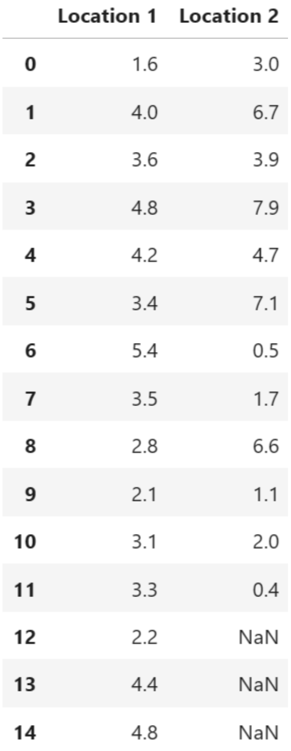


| | Location 2 |
|---|---|
| **mean** | 3.800 |
| **min** | 0.400 |
| **25%** | 1.550 |
| **50%** | 3.450 |
| **75%** | 6.625 |
| **max** | 7.900 |

The concat function was utilized to merge the two DataFrames (df_loc1 and df_loc2) into a single DataFrame (df_combined). The axis parameter was specified as 1 to ensure vertical concatenation, aligning the data for comparison and subsequent operations.

```Python
df_combined = pd.concat([df_loc1, df_loc2], axis=1)
df_combined
```

*Table 3. The concatenated dataframes of the ping rates in Location 1 and Location 2.*



| | Location 1 | Location 2 |
|---|---|---|
| **0** | 1.6 | 3.0 |
| **1** | 4.0 | 6.7 |
| **2** | 3.6 | 3.9 |
| **3** | 4.8 | 7.9 |
| **4** | 4.2 | 4.7 |
| **5** | 3.4 | 7.1 |
| **6** | 5.4 | 0.5 |
| **7** | 3.5 | 1.7 |
| **8** | 2.8 | 6.6 |
| **9** | 2.1 | 1.1 |
| **10** | 3.1 | 2.0 |
| **11** | 3.3 | 0.4 |
| **12** | 2.2 | NaN |
| **13** | 4.4 | NaN |
| **14** | 4.8 | NaN |

# The Box-and-Whiskers Plot

The `seaborn` library is imported for creating boxplots, a visualization technique suitable for comparing distributions across multiple categories. The `matplotlib.pyplot` library (imported as `plt`) is imported for further customization of the plot.

Python

```
import seaborn as sns
import matplotlib.pyplot as plt
```

The `sns.boxplot` function is used to generate the boxplot. The `df_combined` dataframe is passed as input, containing the ping rate data for both locations. The `width` of the boxes is set to 0.2 for better visual clarity, allowing for easier side-by-side comparisons. The color `palette` is specified as 'magma' to provide visually appealing color gradients within the boxes.

Python

```
sns.boxplot(data=df_combined, width = 0.2, palette = 'magma')
```

The customization of the plot starts with a descriptive title using the `plt.title()`, "Comparison of Network Round-Trip Times (RTTs) Between Two Locations" added to the top of the plot, clarifying the purpose of the visualization. The label "Ping Rate" is set for the y-axis, indicating the units of the plotted values (ping time in seconds) through `plt.ylabel()`. The figure size, `plt.figure()`, is adjusted to a width of 15 and a height of 25 units, enhancing readability and allowing for a larger, clearer representation. The final boxplot is displayed using `plt.show()`.

Python

```python
plt.title("Comparison of Network Round-Trip Times (RTTs) Between Two
Locations")
plt.ylabel("Ping Rate")
plt.figure(figsize=(15,25))
plt.show()
```

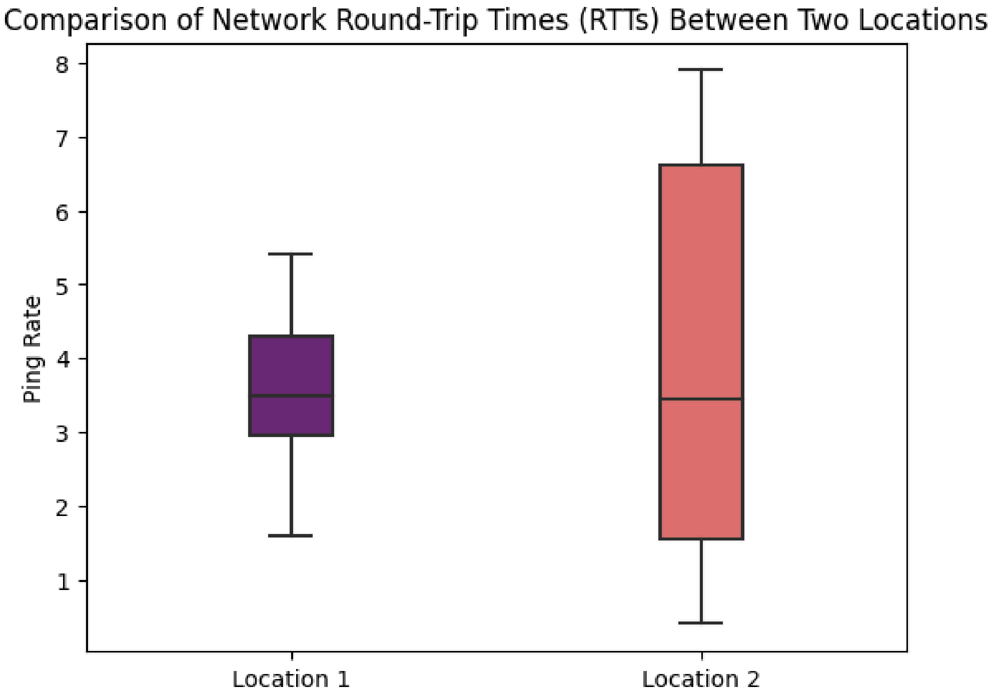


*Figure 1. The parallel boxplots of the ping measurements from Location 1 and Location 2.*

# Interpretation

This report analyzes the ping rates of Location 1 and Location 2, aiming to compare their variability and stability. It's important to note that Location 2 has a smaller sample size compared to Location 1, potentially impacting the generalizability of the comparisons. Ideally, the dataframes should have had equal data points for a more balanced comparison.

The analysis is based on the five-number summary, which includes the minimum, first quartile, median, third quartile, and maximum values. Location 2 has a higher average ping rate (3.80) compared to Location 1 (3.55), its data is spread wider across the range as seen by the larger area it is occupying in the boxplot, as is also indicated by the larger difference between the third quartile (Q3: 6.63) and the first quartile (Q1: 1.55) compared to Location 1 (Q3: 4.30, Q1: 2.95). This suggests that Location 2 experiences more significant ping rate fluctuations than Location 1. Additionally, both locations have higher maximum ping rates in the upper quartile compared to the lower quartile, indicating a positive skew towards higher ping rates.

While boasting a slightly lower average ping rate, Location 1 offers a more stable network due to its tighter data distribution. This translates to fewer significant ping fluctuations. Conversely, Location 2, despite having a higher average ping rate, exhibits a wider data spread, suggesting more pronounced ping fluctuations and a less stable network connection. However, it's crucial to acknowledge the potential impact of the unequal sample sizes when interpreting these results. Future comparisons should utilize dataframes with equal data points for a more robust and definitive assessment.