Rivers of North America

Alberta “Albi” Kovatcheva

Task

- The built-in data set rivers has the length in miles of 141 major rivers in North America. You can build a data frame of this data set that is suitable for graphing as follows: rr = data.frame(rivers)
- Using the following command will provide the first 6 rows of data in the data frame, it should look something like this: head(rr)
- Create a histogram with suitable bin widths, a box plot, and a normal probability plot. Then answer the following questions:

Questions

1. Are there any outliers in this data set? Are they high or low outliers?

2. Do these data appear to come from a normal distribution?

Accessing the Data & Building the Data Frame

Code to Access Data & Build Data Frame:

```
# Access the built in data set 'rivers' which has the length in miles of 141

# major rivers in North America.

View(rivers)


# Build a data frame of this data set that is suitable for graphing:

rr <- data.frame(rivers)
```

Display the First 6 Rows of the Data Frame

Code to display the first 6 rows of the data frame:

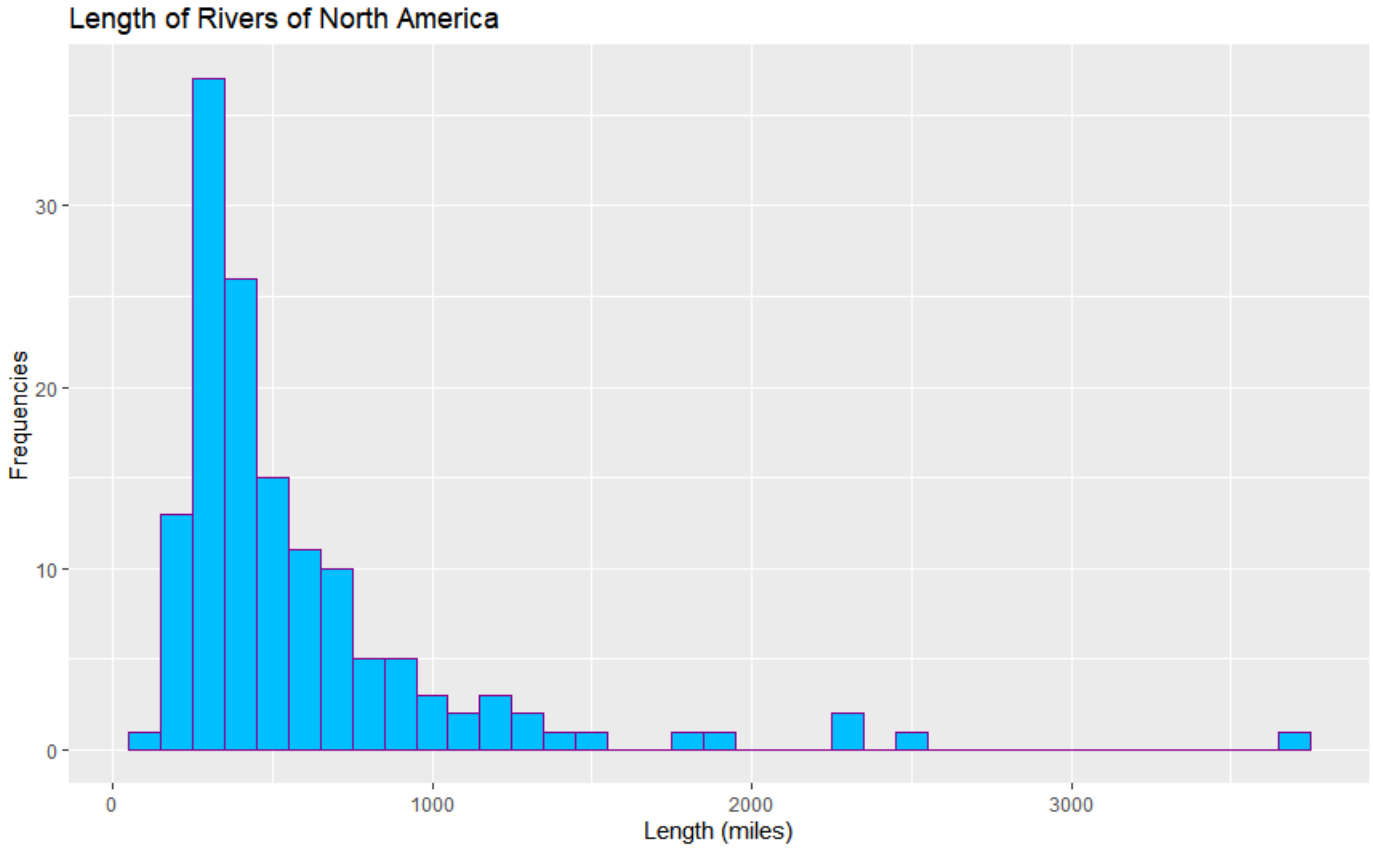
```
# Display the  first 6 rows of data in the data frame:

head(rr)
```

	rivers
1	735
2	320
3	325
4	392
5	524
6	450

Summary Statistics

	Length (miles)
Minimum	135.0
1st Quartile	310.0
Median	425.0
Mean	591.2
3rd Quartile	680.0
Maximum	3710.0

Histogram

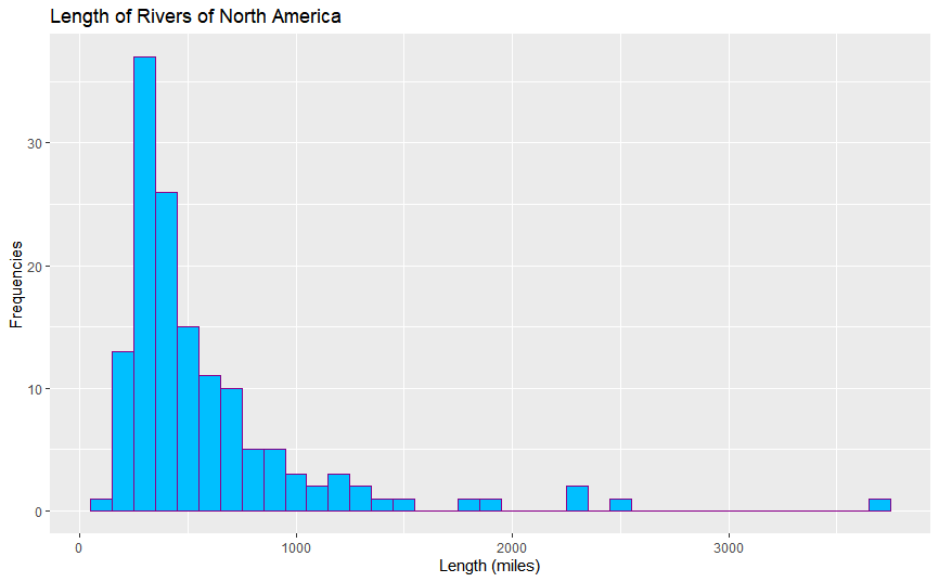


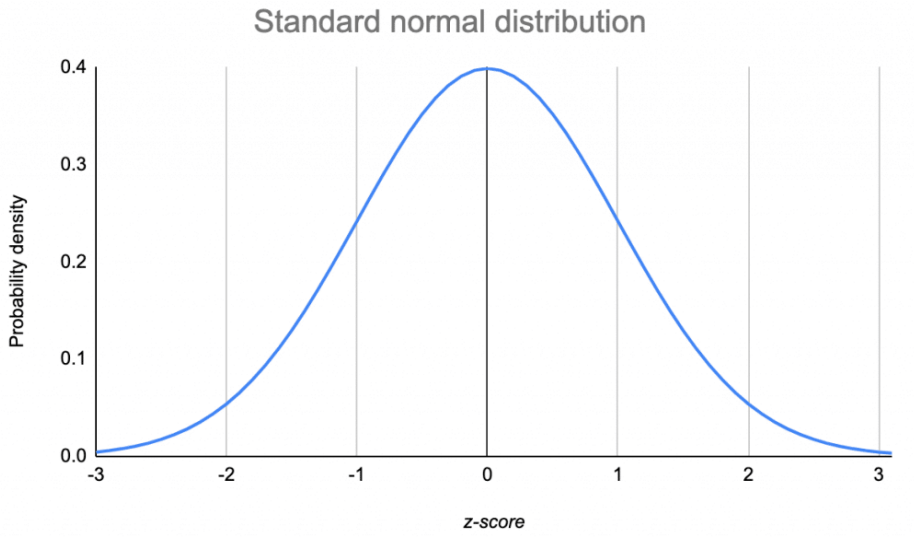
Code to create this histogram:

```
# Histogram

H <- ggplot(rr, aes(x = rivers))

H + geom_histogram(binwidth = 100, fill
= "deepskyblue", color = "darkmagenta")
+

  ggtitle("Length of Rivers of North
America") +

  xlab("Length (miles)") +
ylab("Frequencies")
```

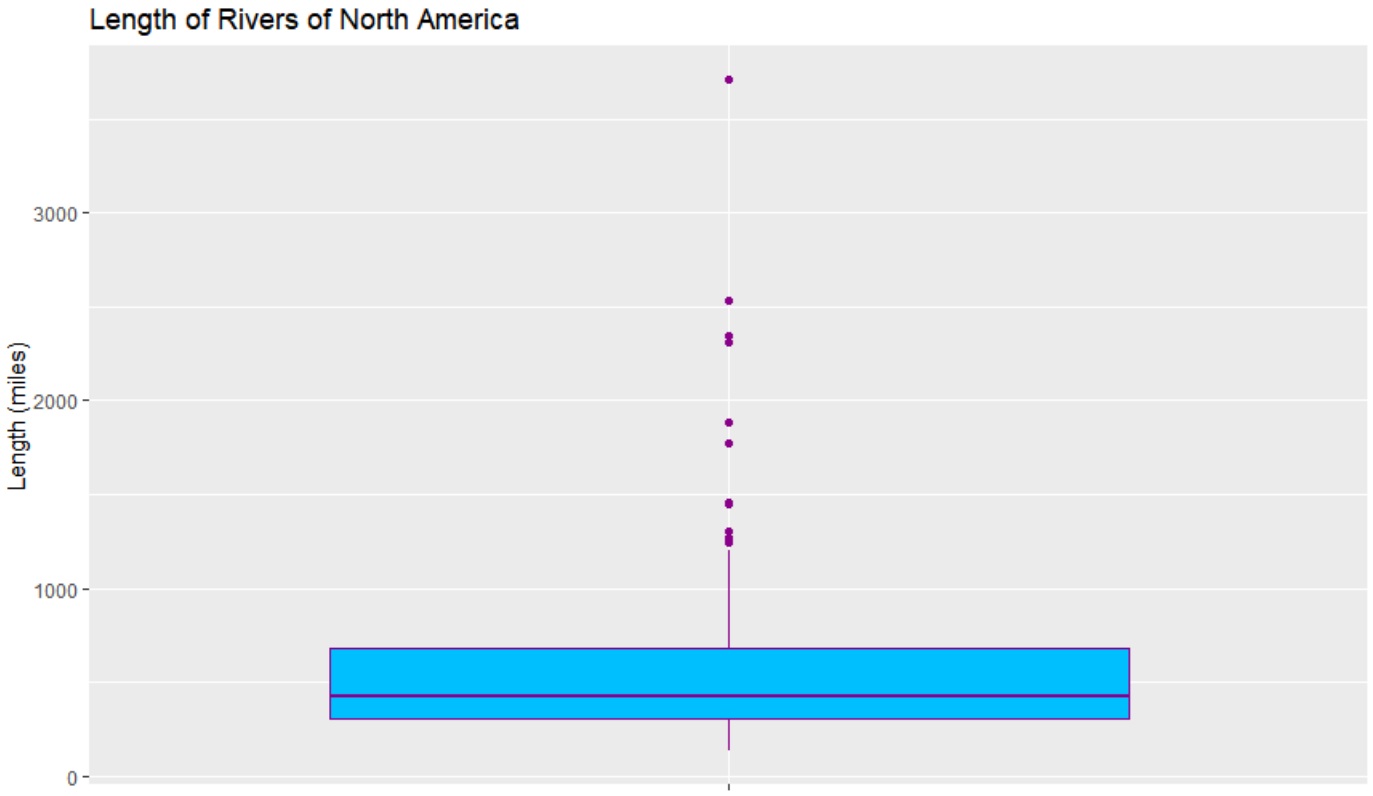
Analysis

- Because the histogram of the length of rivers in North America does not approximate a bell curve, the histogram shows that the data is severely skewed and not normally distributed.





Box Plot



Code to create this box plot:

```
B <- ggplot(rr, aes(x = " ", y = rivers))

B + geom_boxplot(fill = "deepskyblue",
color = "darkmagenta") +

  ggtitle("Length of Rivers of North
America") +

  xlab(" ") +

  ylab("Length (miles)")
```

Analysis

Any data beyond the upper and lower outlier limits are considered outliers. Below is a mathematical assessment of the outliers of this data set.

1st Quartile = 310.0

3rd Quartile = 680.0

IQR = 680.0 - 310.0 = 370.0

IQR * 1.5 = 370.0 * 1.5 = 555

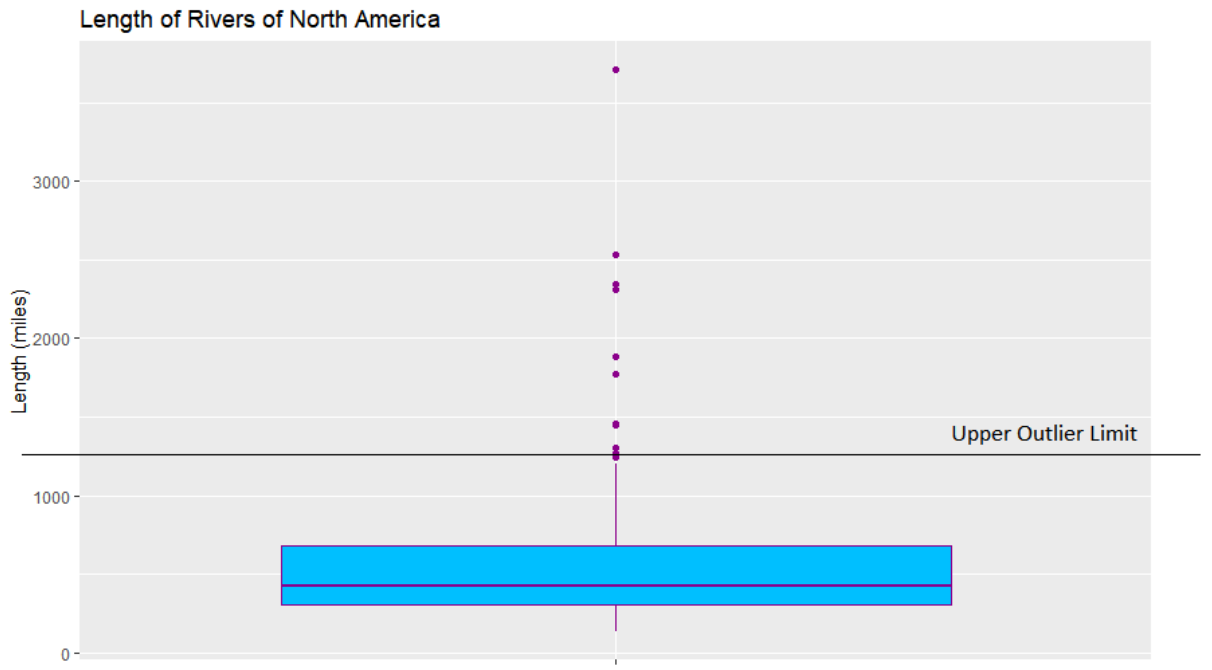
Lower Limit = 310.0 – 555 = -245

Upper Limit = 680.0 + 555 = 1,235

Because all the outliers are high, they inflate the mean significantly.

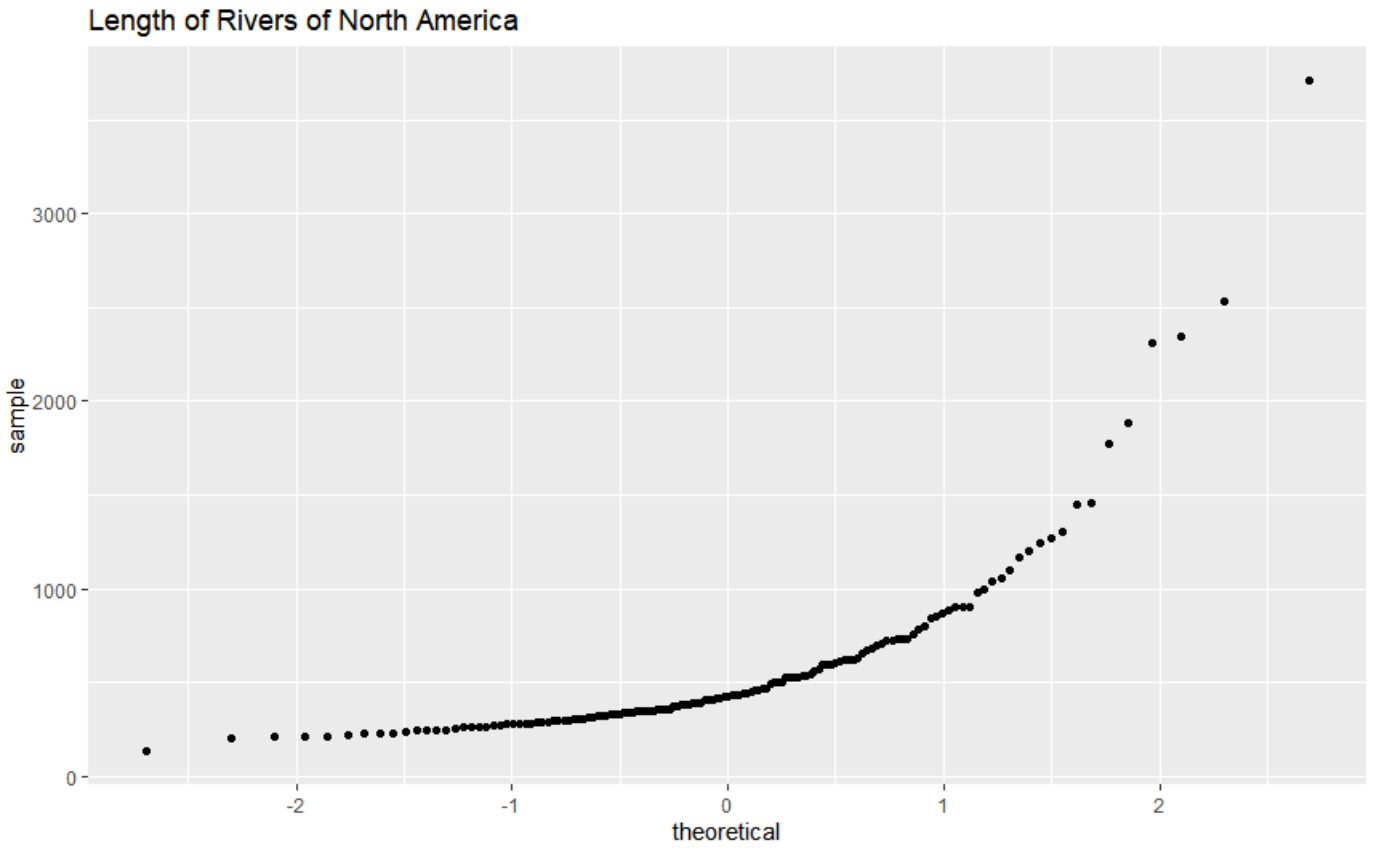
Mean with outliers: 591.2 miles

Mean without outliers: 477.4 miles



	Length (miles) With Outliers	Length (miles) Without Outliers
Minimum	135.0	135.0
1st Quartile	310.0	302.2
Median	425.0	408.5
Mean	591.2	477.4
3rd Quartile	680.0	603.8
Maximum	3710.0	1205.0

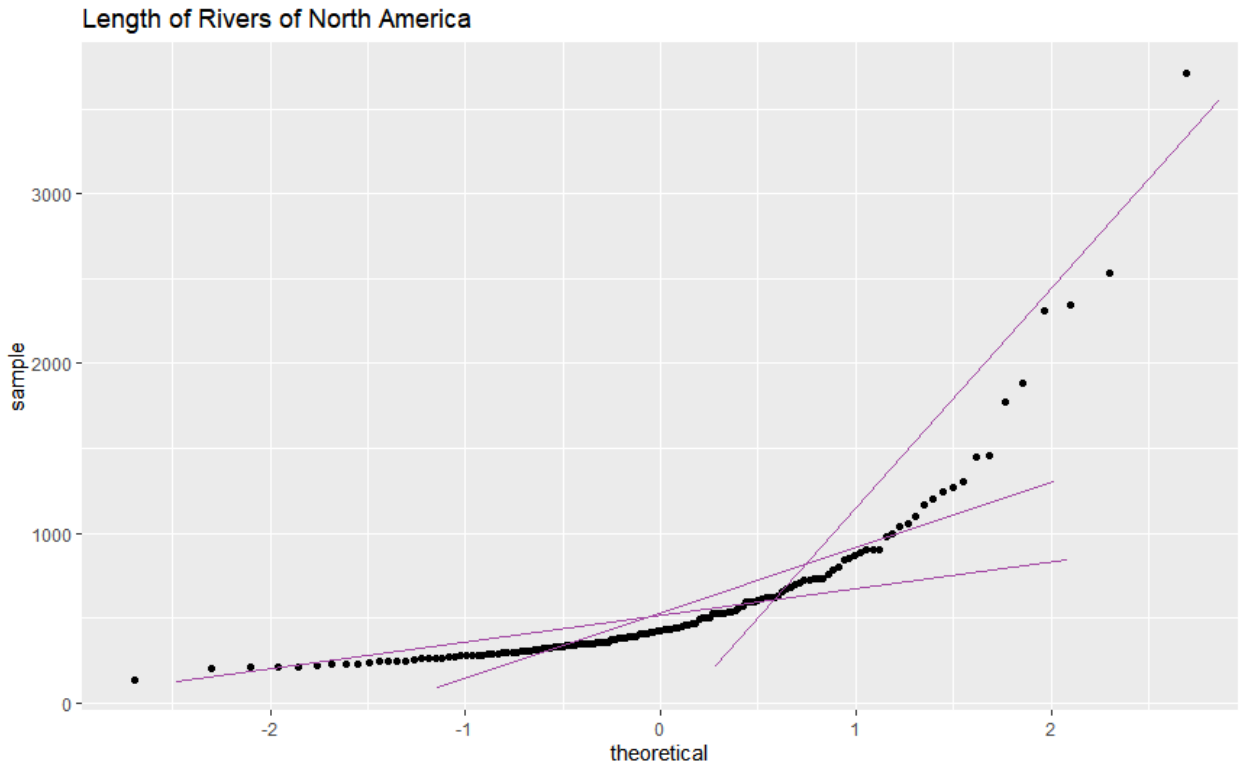
Normal Probability (QQ) Plot



Code to create this QQ plot:

```
# QQ Plot

QQ <- ggplot(rr, aes(sample = rivers))

QQ + geom_qq() +
ggtitle("Length of Rivers of North America")
```

Analysis



- QQ Plot explained:
  - Theoretical: The data set, as if it were part of a standard normal distribution.
  - Sample: The actual data.
- If the data in the QQ plot were approximately linearly distributed, it could be considered normally distributed. However, this is not the case.

Answers

1. Are there any outliers in this data set? Are they high or low outliers?

Based on the Box Plot, there are outliers in this data set, and they are high outliers.

2. Do these data appear to come from a normal distribution?

Based on the histogram, the data does not appear to be normally distributed.

Upon further examination of the QQ Plot, the data fails the “fat pencil test.” Since the data in the QQ plot does not approximate a straight line, the data is not normally distributed.

Conclusion

The length of rivers in North America significantly varies and is not normally distributed.