39_Project3_Part1

Tilak Poudel

2025-04-19

Part 1: Use "airquality" data of R and locate median and mode of "Temp" variable graphically. Validate the value of median and mode obtained from graph with median and mode functions in R. Which summary measure (average and dispersion) must be used for "Wind" and "Temp" variables? Why: Justify your decision with graphs and tests.

```
# Load the airquality dataset
data <- airquality
# Check the structure of the dataset
str(data)
```

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
# Check the first few rows of the dataset
head(data)
```

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

```
# Check the summary of the dataset
summary(data)
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7
```

```
##      Month            Day
##  Min.   :5.000   Min.   : 1.0
##  1st Qu.:6.000   1st Qu.: 8.0
##  Median :7.000   Median :16.0
##  Mean   :6.993   Mean   :15.8
##  3rd Qu.:8.000   3rd Qu.:23.0
##  Max.   :9.000   Max.   :31.0
##
```

```
# Check the names of the columns in the dataset
names(data)
```

```
## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
```

```
# Check the number of rows and columns in the dataset
dim(data)
```

```
## [1] 153   6
```

```
# Check the number of missing values in each column
colSums(is.na(data))
```

```
##   Ozone Solar.R    Wind    Temp   Month     Day
##      37       7       0       0       0       0
```

```
# Check the number of missing values in the "Temp" column
sum(is.na(data$Temp))
```

```
## [1] 0
```

```
# Check the number of missing values in the "Wind" column
sum(is.na(data$Wind))
```

```
## [1] 0
```

```
data$Temp
```

```
##   [1] 67 72 74 62 56 66 65 59 61 69 74 69 66 68 58 64 66 57 68 62 59 73 61 61 57
##  [26] 58 57 67 81 79 76 78 74 67 84 85 79 82 87 90 87 93 92 82 80 79 77 72 65 73
##  [51] 76 77 76 76 76 75 78 73 80 77 83 84 85 81 84 83 83 88 92 92 89 82 73 81 91
##  [76] 80 81 82 84 87 85 74 81 82 86 85 82 86 88 86 83 81 81 81 82 86 85 87 89 90
## [101] 90 92 86 86 82 80 79 77 79 76 78 78 77 72 75 79 81 86 88 97 94 96 94 91 92
## [126] 93 93 87 84 80 78 75 73 81 76 77 71 71 78 67 76 68 82 64 71 81 69 63 70 77
## [151] 75 76 68
```

```
data$Temp <- as.numeric(data$Temp)
class(data$Temp)
```

```
## [1] "numeric"
```

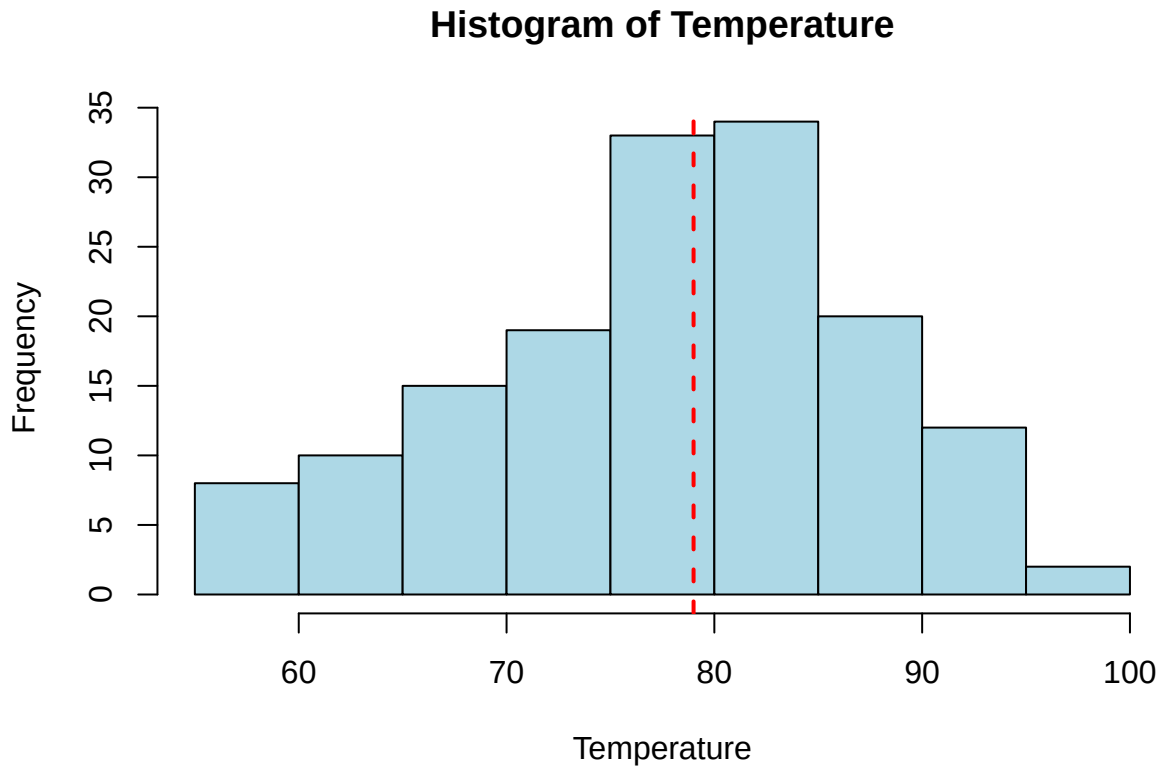
```r
# Check the summary of the "Temp" variable
summary(data$Temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   72.00   79.00   77.88   85.00   97.00
```

```r
# create breaks using min and max values
breaks <- seq(55, 100, by = 5)
breaks
```

```
##  [1]  55  60  65  70  75  80  85  90  95 100
```

```r
# Create a histogram of the "Temp" variable
hist(data$Temp,
     main = "Histogram of Temperature",
     xlab = "Temperature",
     ylab = "Frequency",
     col = "lightblue",
     border = "black",
     breaks = breaks
)
# Add a vertical line for the median
abline(v = median(data$Temp, na.rm = TRUE), col = "red", lwd = 2, lty = 2)
```



```r
# validate the value of median obtained from graph with median function in R
median(data$Temp, na.rm = TRUE)
```

```
## [1] 79
```

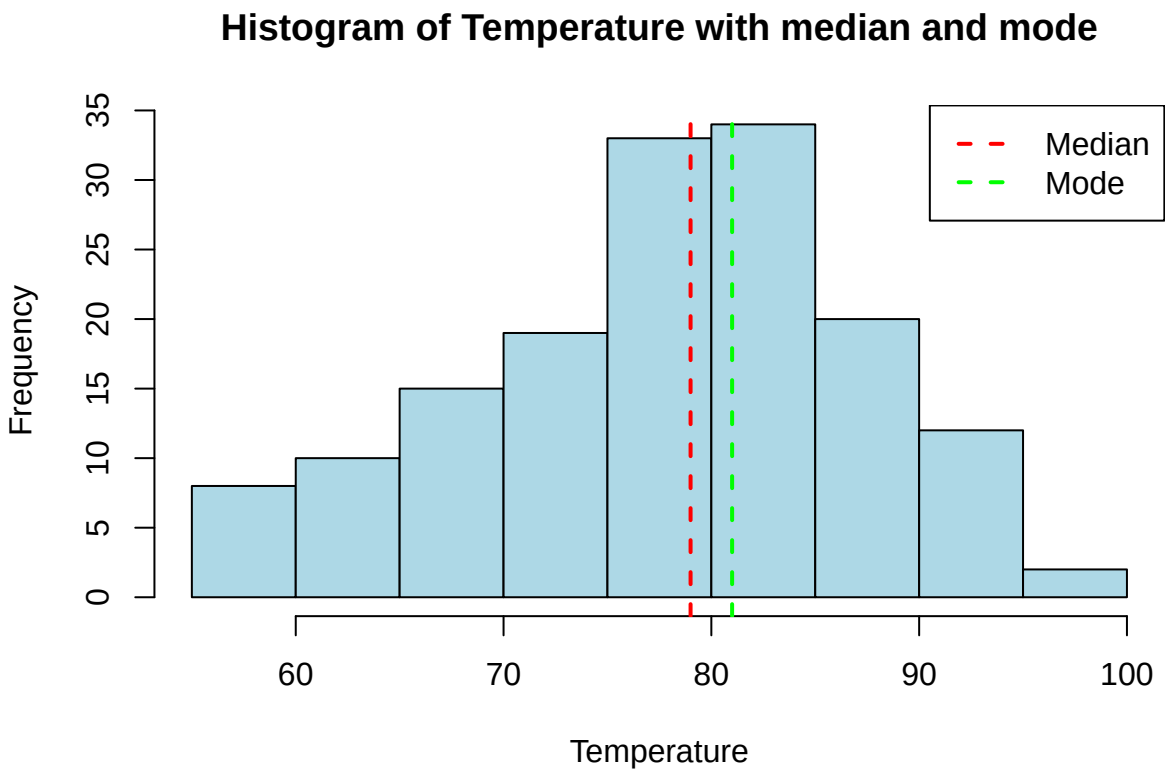
Here we can see that the median value of the "Temp" variable is 79, which is the same as the value obtained from the histogram.

Locate mode of the temp variable

```r
# create frequency table
temp_freq <- table(data$Temp)

# Find the value with highest frequency
mode_temp <- as.numeric(names(temp_freq)[which.max(temp_freq)])
mode_temp
```

```
## [1] 81
```

```r
# Plot the histogram with mode value
hist(data$Temp,
     main = "Histogram of Temperature with median and mode",
     xlab = "Temperature",
     ylab = "Frequency",
     col = "lightblue",
     border = "black",
     breaks = breaks
)

# Add a vertical line for the mode
abline(v = mode_temp, col = "green", lwd = 2, lty = 2)
abline(v = median(data$Temp, na.rm = TRUE), col = "red", lwd = 2, lty = 2)

# Add a legend
legend("topright",
       legend = c("Median", "Mode"),
       col = c("red", "green"),
       lty = 2,
       lwd = 2
)
```



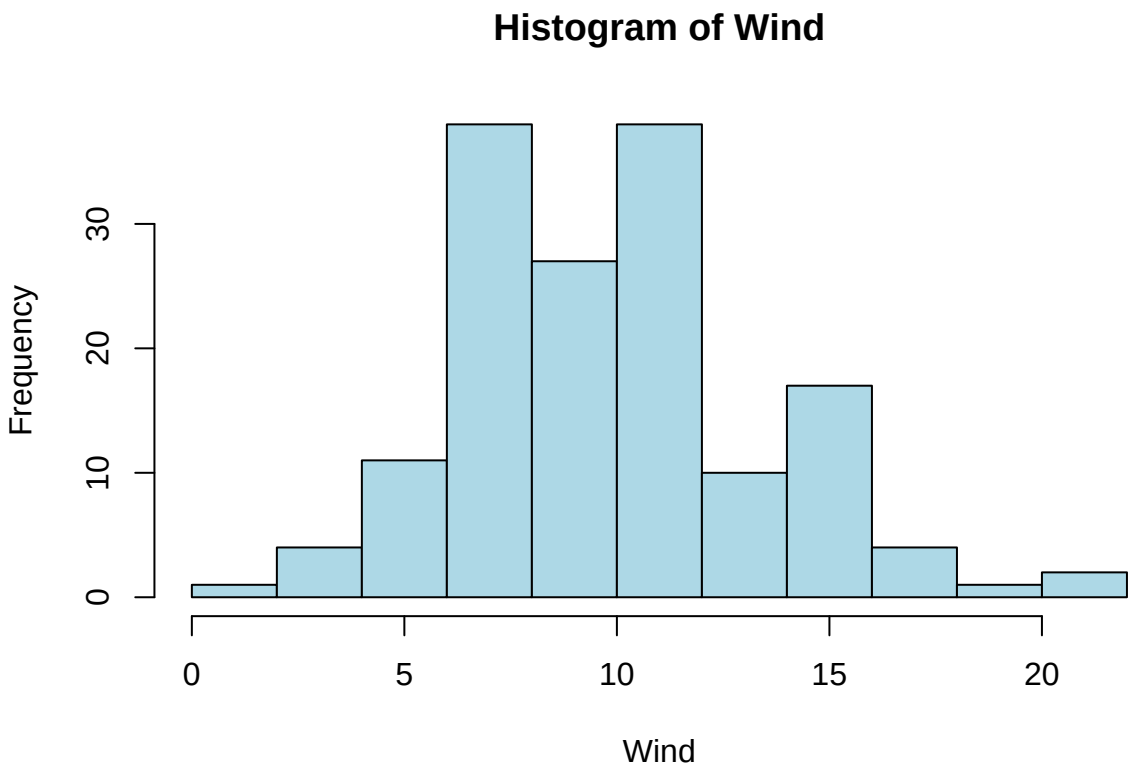

```
# validate the value of mode obtained from graph with mode function in R
mode_temp
```

```
## [1] 81
```

We can see that the mode value of the "Temp" variable is 81, which is the same as the value obtained from the histogram.

Apply appropriate tests to determine the average and dispersion measures for "Wind" and "Temp" variables

```
# Check the distribution of the "Wind" variable
# plot the scatter plot of "Wind" variable
hist(data$Wind,
     main = "Histogram of Wind",
     xlab = "Wind",
     ylab = "Frequency",
     col = "lightblue",
     border = "black"
)
```



From the graph, we can see that the "Wind" variable seems to be normally distributed. Lets apply shapiro-wilk test to check the normality of the data.

```
# Shapiro-Wilk test for normality
shapiro_test_wind <- shapiro.test(data$Wind)
shapiro_test_wind
```

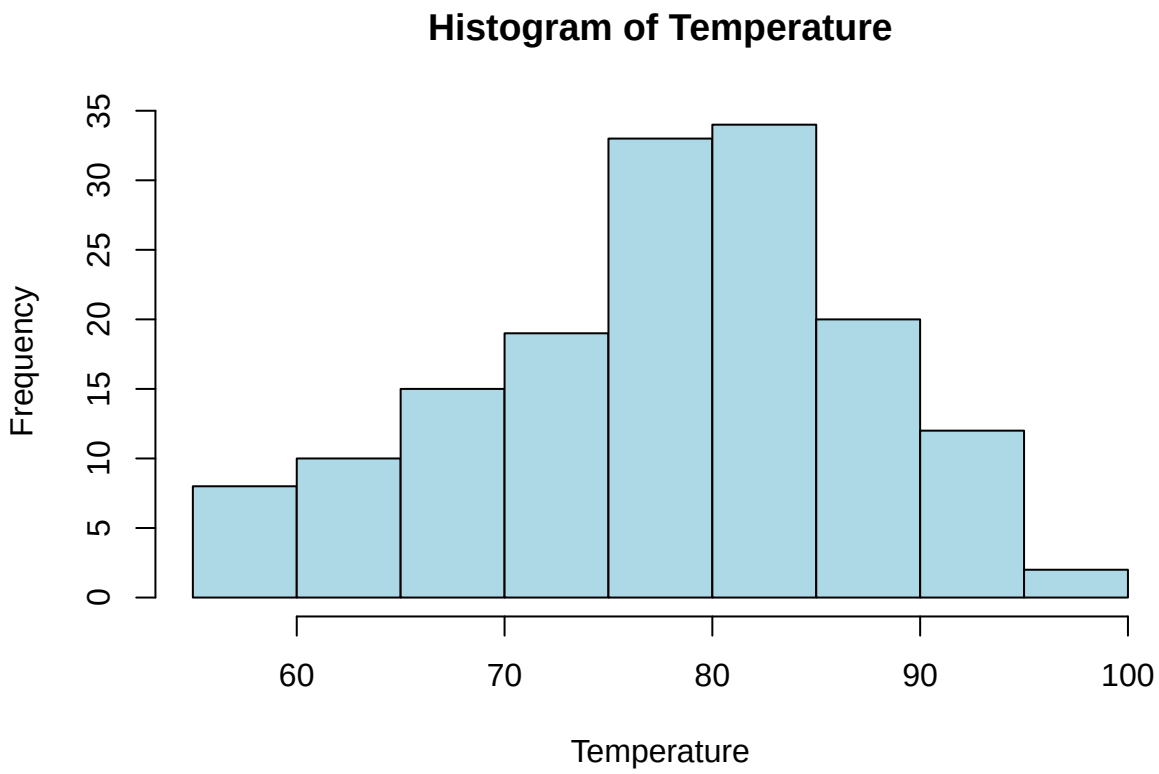
```
##
##  Shapiro-Wilk normality test
##
## data:  data$Wind
## W = 0.98575, p-value = 0.1178
```

The p-value is greater than 0.05, which indicates that we fail to reject the null hypothesis and conclude that the "Wind" variable is normally distributed. So we can use mean and standard deviation as the average and dispersion measures for the "Wind" variable.

Now lets check the distribution of the "Temp" variable

```
# Check the distribution of the "Temp" variable
hist(data$Temp,
     main = "Histogram of Temperature",
     xlab = "Temperature",
     ylab = "Frequency",
```

```
    col = "lightblue",
    border = "black"
)
```



From the graph, we can see that the "Temp" variable seems to be normally distributed. Lets apply shapiro-wilk test to check the normality of the data.

```
# Shapiro-Wilk test for normality
shapiro_test_temp <- shapiro.test(data$Temp)
shapiro_test_temp
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Temp
## W = 0.97617, p-value = 0.009319
```

The p-value is less than 0.05, which indicates that we reject the null hypothesis and conclude that the "Temp" variable is not normally distributed. So we can use median and inter-quartile range as the average and dispersion measures for the "Temp" variable.