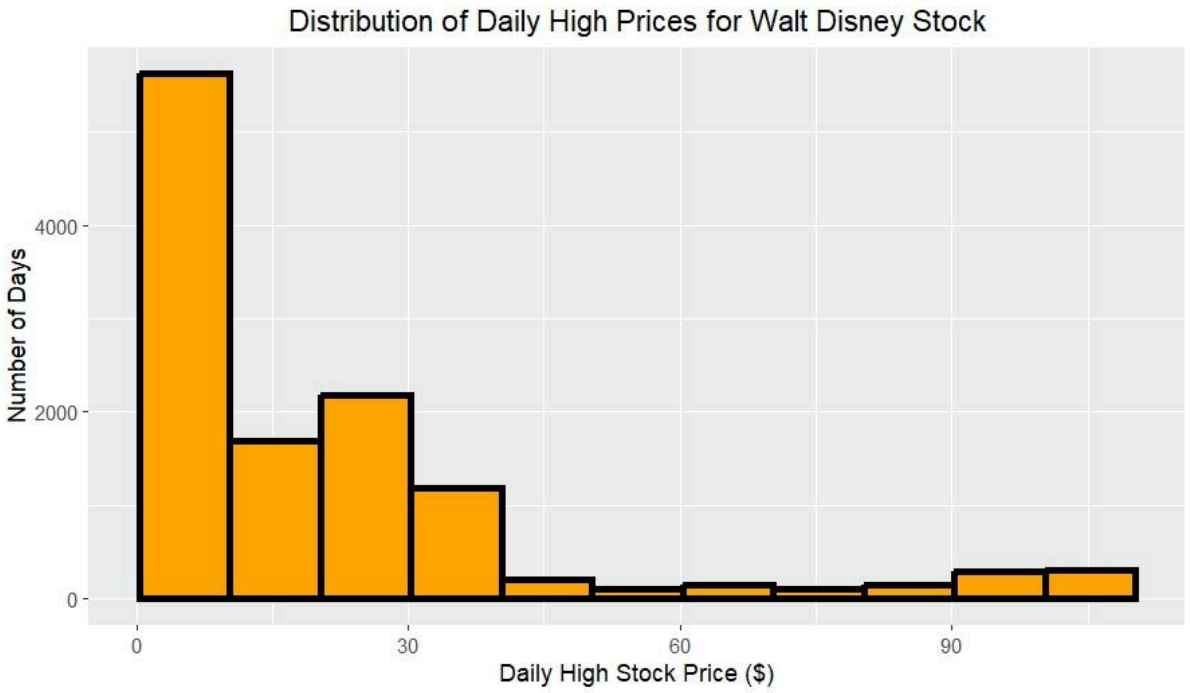
Data Analysis - The Central Limit Theorem

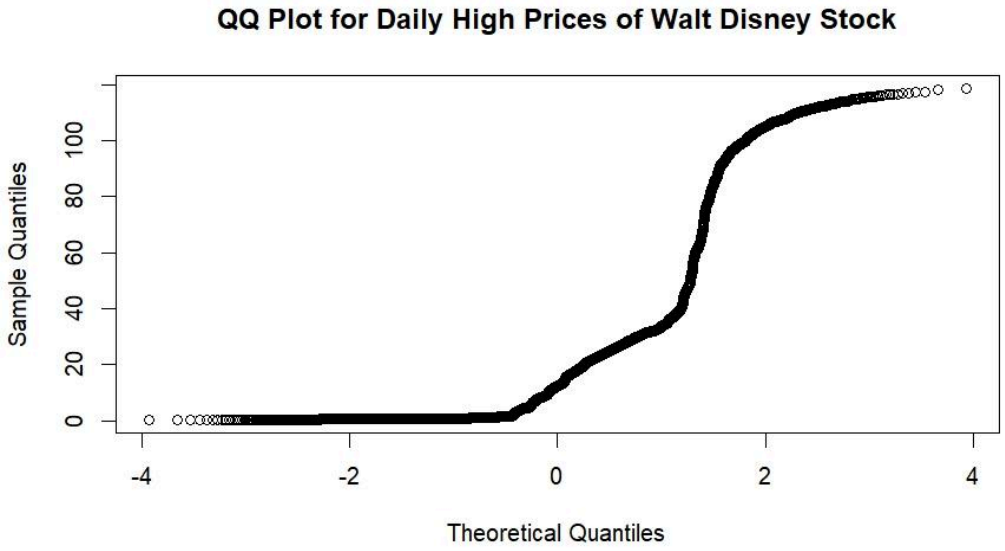
Student: Kevin Ivanov

Population Distribution

1) **Construct a population distribution in the form of a basic histogram for the High variable.**



2) **Construct a QQ Norm Plot for the High variable.**



3) **Construct a nicely formatted table showing the mean and standard deviation of the variable. Refer to prior assignments for a nicely formatted table. These values are population parameters.**

Statistic	Value
Mean	20.81493
Standard Deviation	26.66029

4) **Is the variable normally distributed? If is not normally distributed, is the distribution skewed right? skewed left? Something else?**

R: The daily prices of Walt Disney Stock aren't normally distributed; they are right-skewed.

Sampling Distribution

5) Randomly select a sample of size 100 from the population of High data.

```
> sample(x=DisneyData$High, size = 100, replace = TRUE)
  [1]   0.79400   0.72110 103.24000   0.81240   0.79400   0.75760  74.99900  12.34000
  [9]  16.84900   0.83070  12.10400  36.06600  22.83700  33.81600  25.04600  38.84800
 [17]  32.88500  12.45700   1.03130  17.79900   0.84870  28.68700  27.01700   8.32420
 [25]   1.20470   1.34160  25.38400   0.87630  20.45500  27.07300   7.26540  33.03100
 [33]  16.24700 113.13000  31.38800  23.12800   0.78500 102.82000  27.74800   1.36000
 [41]  34.15500  17.97900  39.81400  10.25900   0.49290   0.56600   0.47465   4.30820
 [49]  13.79100  16.65800   1.06790   3.18550  15.35200   2.09920   8.74410  48.28400
 [57]   0.76660   0.61160   0.99490  21.23000  42.50600   8.41540   1.48770  23.84100
 [65]   0.83070  17.12500   8.14160  31.46300   6.06050  13.13300   1.30520  23.44900
 [73]  31.01400   0.64810  17.45200  23.64000   0.82140  22.72700 101.14000   0.78500
 [81]  23.75000  22.67200   0.76660  13.28000  43.27600 102.82000   8.44270   0.85800
 [89]   7.62130   0.73020  22.31600  38.95200   1.11360   4.46320  16.42900   0.66610
 [97]   1.55170   0.60250  11.02600   8.98130
```

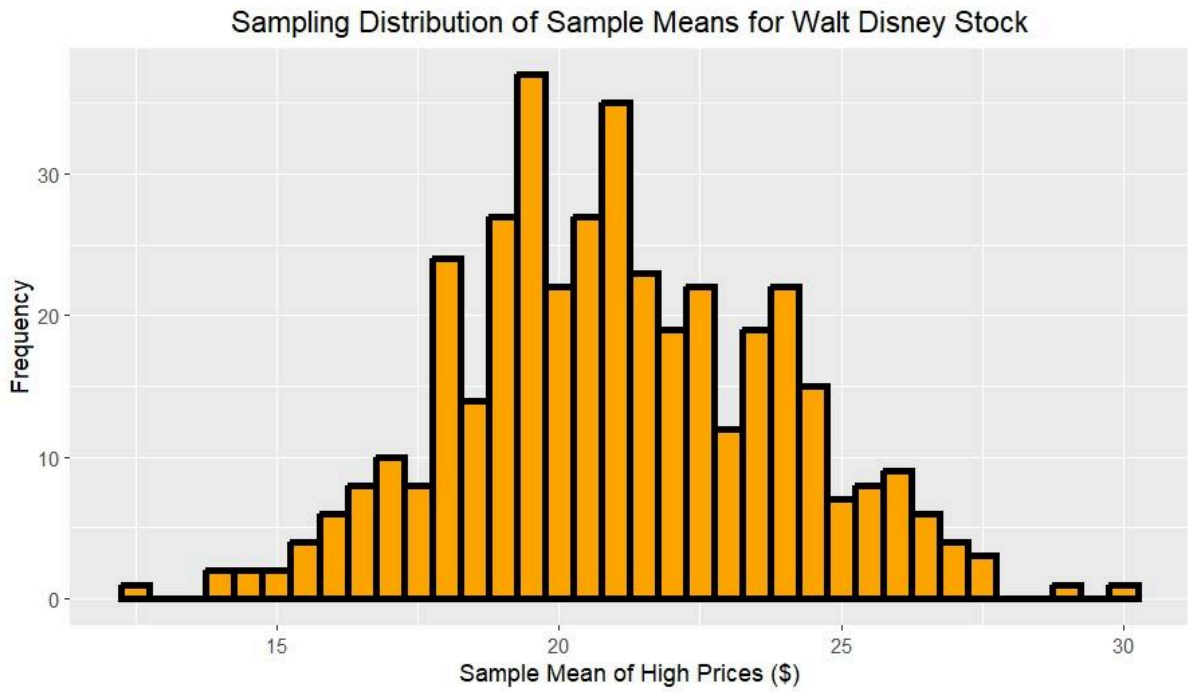
6) Find the mean of a random sample of size 100 from the population

```
> mean(sample(x=DisneyData$High, size = 100, replace = TRUE))
[1] 19.37262
```

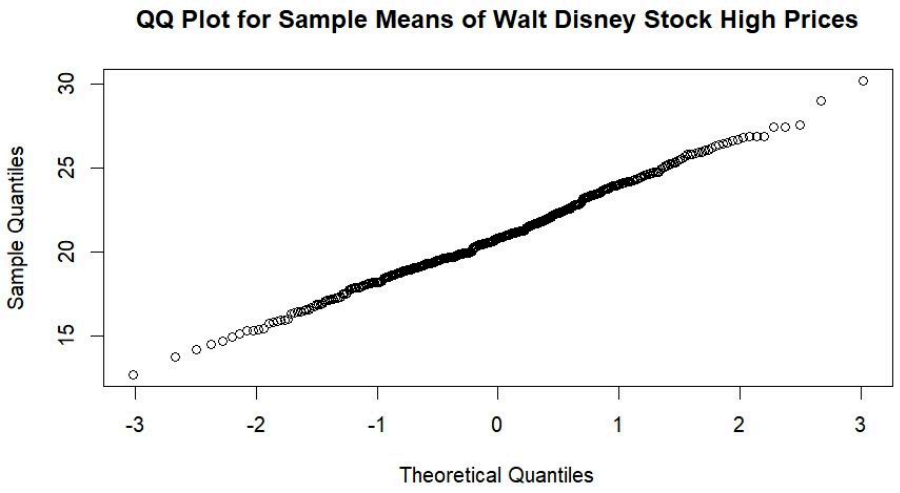
7) Repeat this process so that you have computed 400 means from 400 randomly selected samples of size 100.

DFrame	400 obs. of 1 variable

8) Construct a basic histogram of the sample means computed and stored in DFrame.



9) Construct a QQ Norm Plot for the sample means. Use the DFrame - not LabData in the construction of the QQ Plot.



10) Construct a nicely formatted table showing the mean and standard deviation of the sample means. Refer to prior assignments for a nicely formatted table.

Statistic	Value
Mean	20.99447
Standard Deviation	2.850145

11) Is the sampling distribution of the sample means normally distributed? If it is not normally distributed, is the distribution skewed right? skewed left? Something else?

R: The sampling distribution of the sample means for Walt Disney stock high prices is normally distributed. This means we can use standard statistical methods that habdle the data is normal.

The Central Limit Theorem

12)a) Complete the table

	Shape	**Mean**	**Std Dev**
Population Distribution	Right-skewed	20.81493	26.66029
Sampling Distribution	Approximately Normal	20.99447	2.850145

b) How do the two means compare to each other? Are they approximately equal? Is one (significantly) larger than the other? If one is larger than the other, which one is larger and by how much?

R: There is a 0.18 discrepancy between the sampling and population means, This difference of 0.18 is small and not considered significant, It shows that the method used to sample the data is effective, the small difference supports the reliability of using sample means to estimate the population mean

c) How do the two standard deviations compare? Are they approximately equal? Is one (significantly) larger than the other? If one is larger than the other, which one is larger and by how much?

R: The smaller std in the sampling distribution confirms the expected behavior described by statistical theory, which predicts a decrease in variability with averaged data The population standard deviation is approximately 9.35 times larger than that of the sampling distribution.

13) Compute: . Here n is the sample size of 100. How does this value compare to the standard deviation of the sample means ?

σ = 26.66 n = 100

Calculated Standard Error: 2.666

Sample Means: 2.85

This tiny discrepancy suggests that the standard error, or theoretical prediction, closely resembles the actual data.

14) The Central Limit Theorem predicts the shape of the sampling distribution. What shape does the Central Limit Theorem predict for the sampling distribution? What is the shape of the sampling distribution? Did the work done in this assignment confirm the conclusion of the Central Limit Theorem? Briefly explain your response.

R: Yes, the work done in this assignment supports the CLT. The CLT states that with a large enough sample size, the sampling distribution of the sample means will be approximately normal, regardless of the population's distribution. This was confirmed by the bell-shaped histogram and the QQ plot, both indicating normality, thus validating the CLT's prediction.

15) The Central Limit Theorem predicts the mean of the sampling distribution. What mean does the Central Limit Theorem predict for the sampling distribution? What is the mean of the sampling distribution? Did the work done in this assignment confirm the conclusion of the Central Limit Theorem? Briefly explain your response.

R: This assignment supports the CLT, showing that a large sample size makes the sampling mean closely match the population mean, despite a minor discrepancy.

16)The Central Limit Theorem predicts the standard deviation of the sampling distribution. What standard deviation does the Central Limit Theorem predict for the sampling distribution? What is the standard deviation of the sampling distribution? Did the work done in this assignment confirm the conclusion of the Central Limit Theorem? Briefly explain your response.

R: The CLT predicts the sampling distribution's standard deviation to be approximately 2.666. In this assignment, the observed standard deviation is 2.85. This slight difference supports the CLT's accuracy, confirming its reliability in practical scenarios

Code

```
Histogram.1 <- ggplot(DisneyData, aes(x = High)) +

geom_histogram(col = "black", fill = "grey", breaks = seq(min(DisneyData$High),
max(DisneyData$High), by = 10), lwd = 1.5) +

labs(x = "Daily High Stock Price ($)", y = "Number of Days", title = "Distribution of Daily High
Prices for Walt Disney Stock") +

theme(plot.title = element_text(hjust = 0.5))

qqnorm(DisneyData$High, main = "QQ Plot for Daily High Prices of Walt Disney Stock")

> mean(DisneyData$High , na.rm = T)

[1] 20.81493

> median((DisneyData$High , na.rm = T))

Error: unexpected ',' in "median((DisneyData$High ,"

> median(DisneyData$High ,na.rm = T)

[1] 12.363
```

```
> sd(DisneyData$High ,na.rm = T)

[1] 26.66029

> quantile(DisneyData$High ,na.rm = T)

     0%      25%      50%      75%

0.28298  1.03130 12.36300 28.29600

    100%

118.63000

sample(x=DisneyData$High, size = 100, replace = TRUE)

mean(sample(x=DisneyData$High, size = 100, replace = TRUE))

SampleMeans <- replicate(400,mean(sample(x=DisneyData$High, size=100, replace=TRUE)) )

DFrame <-data.frame(SampleMeans)

Histogram2 <- ggplot(DFrame, aes(x = SampleMeans))

Histogram2  +  geom_histogram(col = "black", fill = "orange", binwidth = 0.5, lwd = 1.5) +  #
Adjust binwidth to better suit the distribution of sample means

labs(x = "Sample Mean of High Prices ($)",

y = "Frequency",

title = "Sampling Distribution of Sample Means for Walt Disney Stock") +

theme(plot.title = element_text(hjust = 0.5))

qqnorm(DFrame$SampleMeans, main = "QQ Plot for Sample Means of Walt Disney Stock High
Prices")
```