Statistical Analysis of Obesity Data

MA-541

Nishant Singh

Nishant Upadhyay

Rajvir Singh

Jahaan Malik

Table of Contents

LIST OF VARIABLES

We've distilled our dataset down to five key variables: AGE, HEIGHT, GENDER, WEIGHT, and NObeyesdad. AGE, HEIGHT, GENDER, and WEIGHT act as the predictors, influencing the target variable NObeyesdad. This selection allows us to explore how factors such as age, height, gender, and weight contribute to predicting obesity levels.

AGE – The biological age of the individual.

HEIGHT – The Physical height of the individual.

GENDER – The sex/gender of the individual.

WEIGHT – The physical weight of the individual.

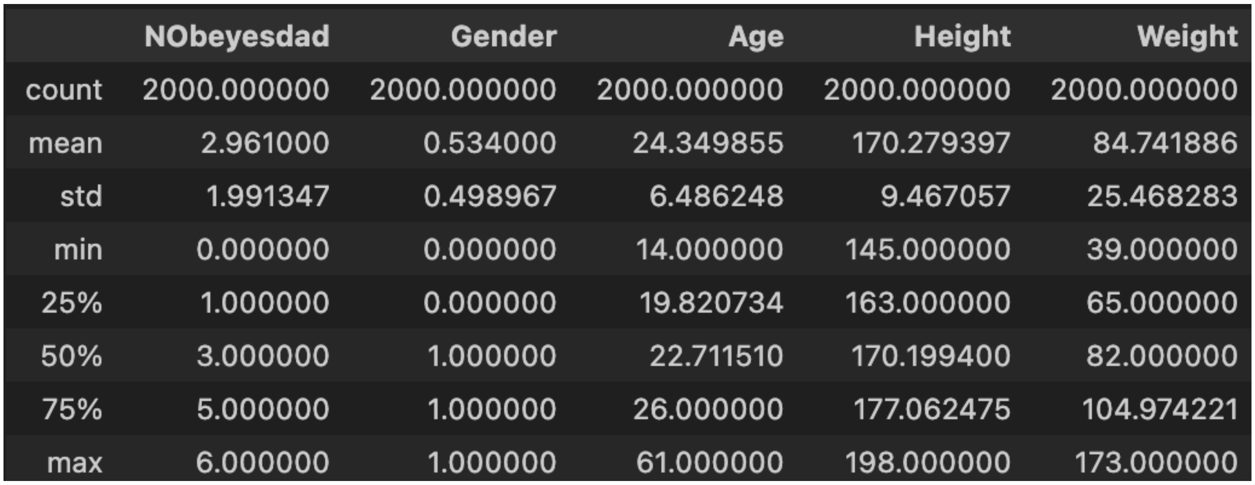
NObeyesdad – The type of Obesity the individual has.

INTRODUCTION

To observe how the theoretical basis of important statistical concepts influences the dynamics of various real-world applications, an analysis was conducted of a given obesity dataset encompassing n=2000 values for five variables of interest related to the physical attributes of individuals and key potential influencers thereof: Gender, Height, Weight, Age and NObeyesed. In this work, histograms and Q-Q plots were used to study the distribution of each variable, with Shapiro-Wilk and K-S testing as supplementary tools to reach a determination. These visualization methods were also used to confirm the principles of the Central Limit Theorem (CLT) and explore its consequences in this context. Confidence intervals were constructed to determine whether they were viable captures of the population mean of the Height variable. The analysis tested our knowledge of hypothesis testing to reach correct conclusions about the mean and standard deviation of Height, as well as when comparing the Height and Weight variables. Regression analysis and model assessment were also performed by fitting a line to the Height vs. Weight data and interpreting the results of the scatterplot, histogram, and Q-Q plot output of the residuals. In this work, we observe that the low correlation among the variables poses challenges to the predictive capabilities of linear regression models, and evidence of such is presented throughout. There is, however, not enough evidence to refute normality assumptions, or that the data contradicts the CLT in any way.

PRELIMINARY INVESTIGATION OF STATISTICAL MEASURES OF DATASET FEATURES

The Figure below shows a statistical summary of a dataset containing information about obesity levels, gender, age, height, and weight. The dataset has 2000 observations or rows. The columns show the count, mean, standard deviation, minimum, 25th percentile, 50th percentile (median), 75th percentile, and maximum values for each variable. This summary provides an overview of the central tendency and spread of various attributes related to obesity levels in the dataset.



	NObeyesdad	Gender	Age	Height	Weight
count	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	2.961000	0.534000	24.349855	170.279397	84.741886
std	1.991347	0.498967	6.486248	9.467057	25.468283
min	0.000000	0.000000	14.000000	145.000000	39.000000
25%	1.000000	0.000000	19.820734	163.000000	65.000000
50%	3.000000	1.000000	22.711510	170.199400	82.000000
75%	5.000000	1.000000	26.000000	177.062475	104.974221
max	6.000000	1.000000	61.000000	198.000000	173.000000

Some key observations from the data:

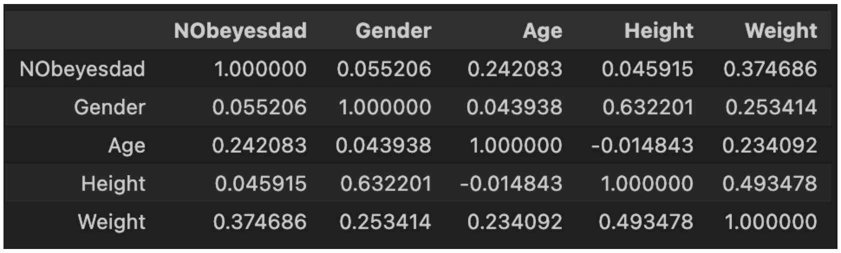
1. The mean obesity level is around 2.96, with a standard deviation of 1.99, indicating a widespread in the data.

2. The mean age is around 24.35 years, with a minimum of 14 and a maximum of 61 years.

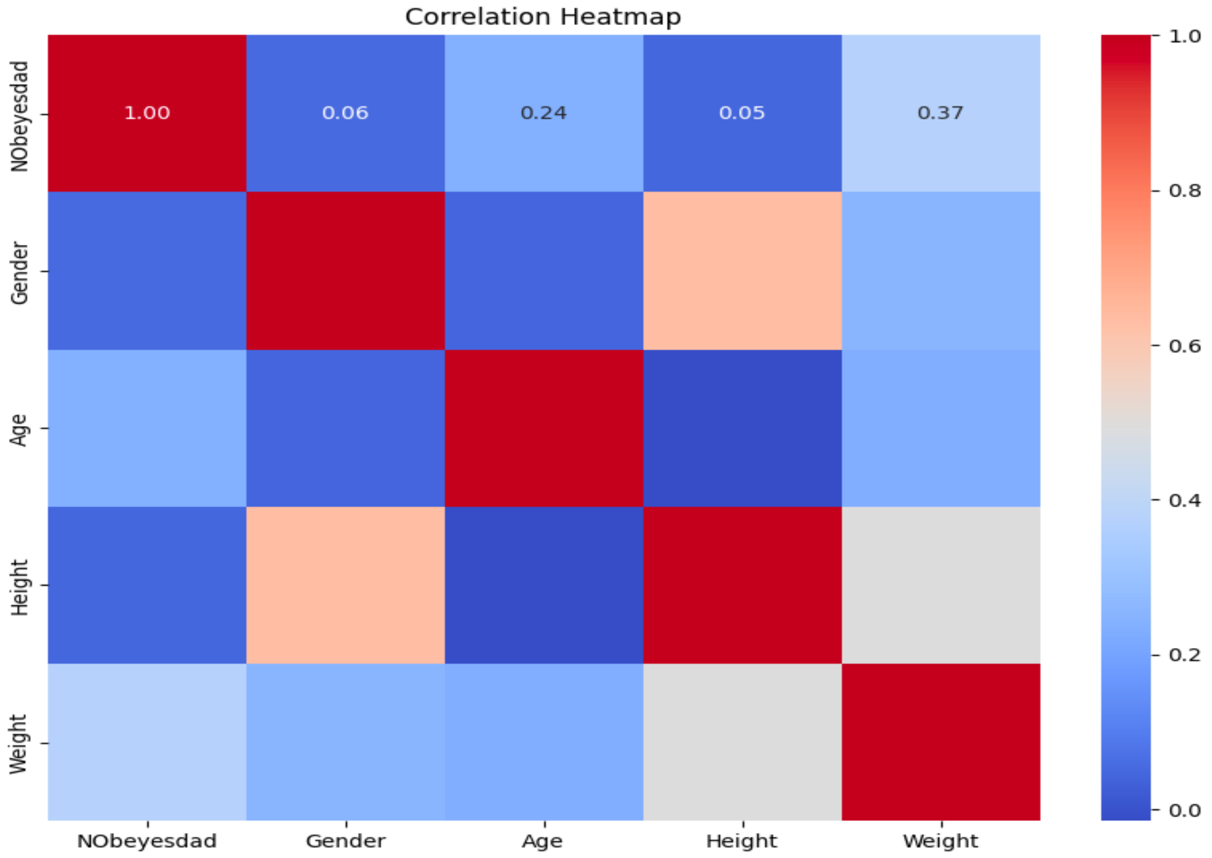
3. The mean height is around 170.28 cm (approximately 5'7"), with a minimum of 145 cm (4'9") and a maximum of 198 cm (6'6").

4. The mean weight is about 84.74 kg (186.6 lbs), with a minimum of 39 kg (86 lbs) and a maximum of 173 kg (381 lbs).

Then, the sample correlations among each pair of the random variables were obtained. All pairs except for those that examined the variable against itself produced a correlation of close to zero, indicating that a strong linear relationship in either the positive or negative direction is not present.



	NObeyesdad	Gender	Age	Height	Weight
NObeyesdad	1.000000	0.055206	0.242083	0.045915	0.374686
Gender	0.055206	1.000000	0.043938	0.632201	0.253414
Age	0.242083	0.043938	1.000000	-0.014843	0.234092
Height	0.045915	0.632201	-0.014843	1.000000	0.493478
Weight	0.374686	0.253414	0.234092	0.493478	1.000000



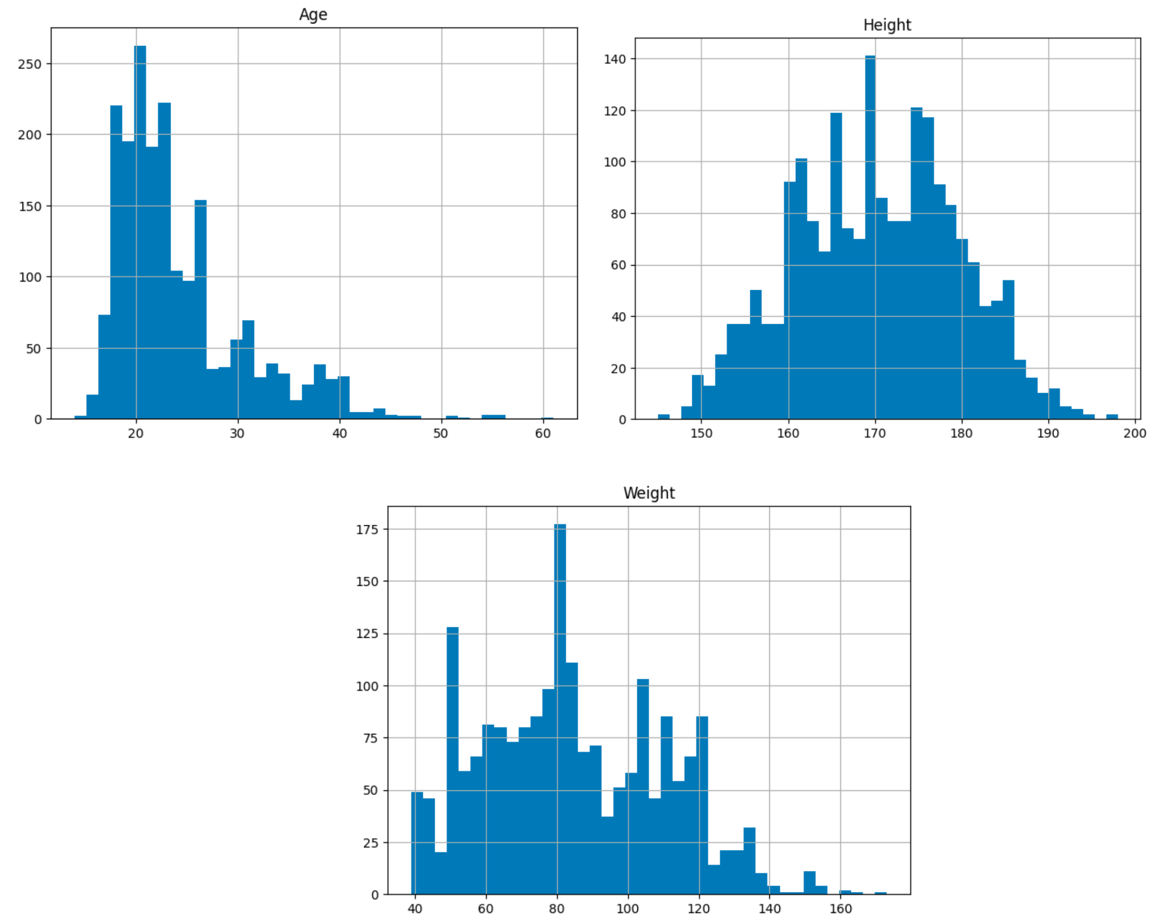
INITIAL VISUALIZATIONS AND ASSUMPTIONS REGARDING VARIABLE BEHAVIOR

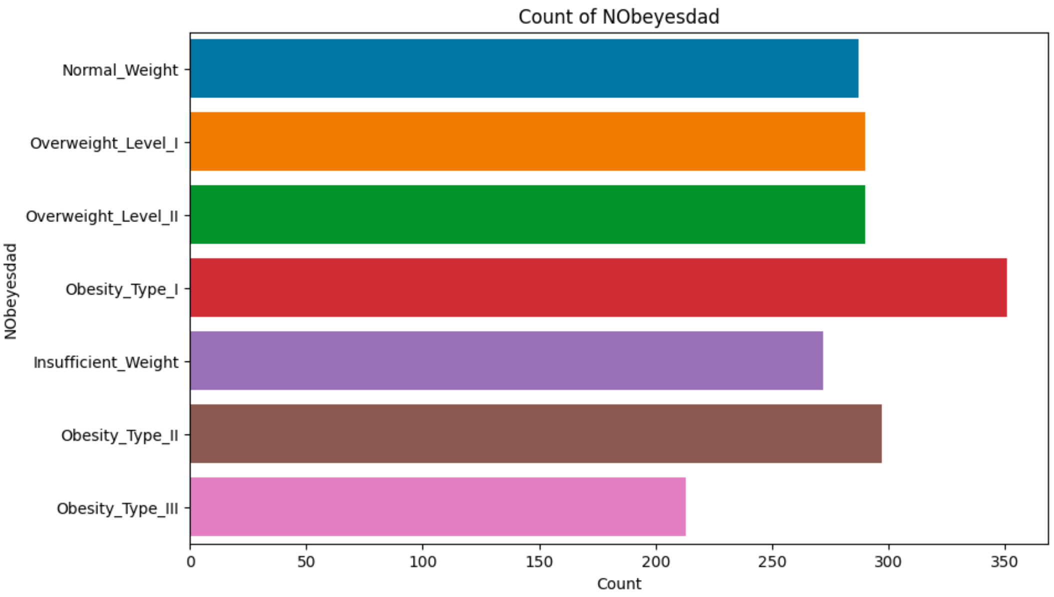
Various visualization tools including histograms, and barplots were formed for the purpose of extracting general information about individual feature distribution and behavior. The age histogram shows a roughly bell-shaped distribution, but with some positive skewness. This suggests that the age distribution is slightly skewed towards higher ages, with a longer tail on the right side of the distribution.

The height histogram appears to be approximately symmetric and bell-shaped, indicating that the heights in the dataset may be normally distributed or close to a normal distribution. The weight histogram exhibits a slight positive skewness, with a longer tail towards higher weight values. This indicates that the distribution of weights is skewed towards the right, with more individuals having higher weights than a perfectly normal distribution would suggest.

The bar chart provides insights into the distribution of obesity levels within the dataset. It reveals that a considerable portion of the individuals fall within the "Normal Weight" category, forming the largest group. However, a significant number of individuals are also classified as having "Obesity_Type_I", indicating a substantial presence of obesity in the dataset. Additionally, the chart highlights the existence of individuals who are overweight but not yet obese, as seen in the "Overweight_Level_I" and "Overweight_Level_II" categories. While the counts for "Insufficient_Weight" and "Obesity_Type_III" are relatively lower, they represent the extremes of the weight spectrum, encompassing underweight and severe obesity cases. Overall, the bar chart offers a visual representation of the varying weight and obesity levels present within the dataset, allowing for a comprehensive understanding of the distribution across different categories.

Additionally, by observing the scales and ranges of the histograms, we can make inferences about the typical values and spread of each variable within the dataset. For example, the age histogram indicates that most individuals in the dataset are between 15 and 35 years old, while the weight histogram shows a wide range of weights, from around 40 kg to over 170 kg.

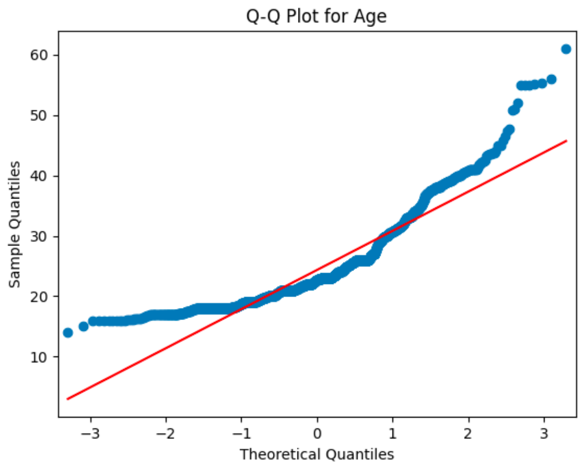

Age
0
50
100
150
200
250
20
30
40
50
60
Height
0
20
40
60
80
100
120
140
150
160
170
180
190
200
Weight
0
25
50
75
100
125
150
175
40
60
80
100
120
140
160

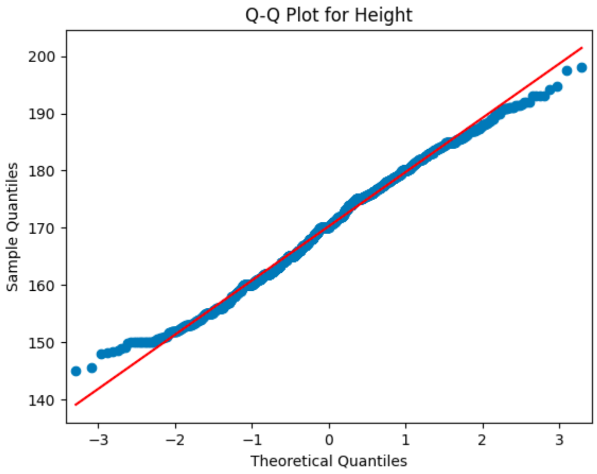

Count of NObeyesdad
NObeyesdad
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III
0
50
100
150
200
250
300
350
Count

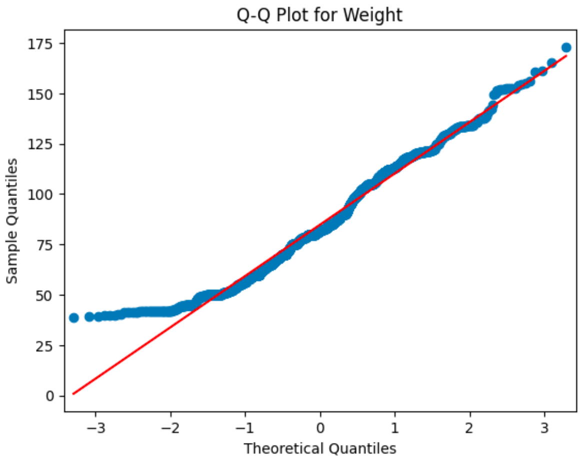
DETERMINING THE DISTRIBUTION OF EACH RANDOM VARIABLE

Histograms and Q-Q plots were used to assess the normality of the variables. The histograms showed approximately bell-shaped distributions for age, height, and weight, while the NObeyesdad variable exhibited some deviations from normality.

The Shapiro-Wilk test were performed to formally test the normality assumption. These tests generally rejected the null hypothesis of normality for the variables, likely due to the large sample size. However, visual inspection of the Q-Q plots suggested that the variables were approximately normal, with some deviations in the tails except







```
Age Shapiro-wilk Test:  Statistic = 0.8649037480354309, p-value = 1.8860056413169212e-
38

Height Shapiro-wilk Test:  Statistic = 0.9925702810287476, p-value =
1.5525527530257932e-08
```

```
Weight Shapiro–wilk Test:  Statistic = 0.9756847023963928, p–value =
6.083075057776696e–18
```

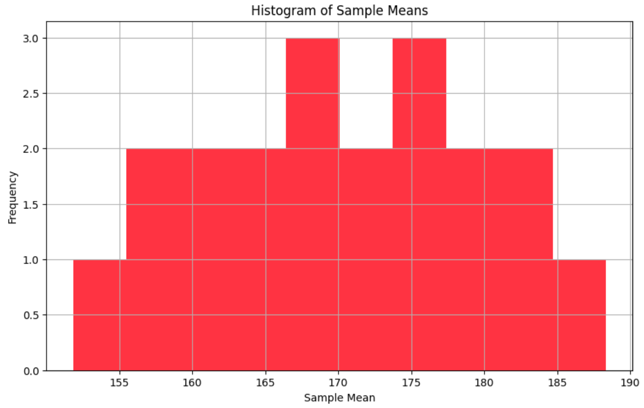
The Q-Q (Quantile-Quantile) plots provide insightful comparisons for the distributions of Age, Height, and Weight. In terms of Age, distinct deviations from a perfect normal distribution are evident, with the upper quantiles indicating positive skewness and the lower quantiles suggesting negative skewness, alongside a leptokurtic shape. Conversely, Height demonstrates a closer adherence to a normal distribution, with minimal deviations observed, indicating a distribution that is largely in line with normality. However, Weight stands out with significant deviations from normality, particularly in the upper quantiles, suggesting positive skewness and a more peaked distribution.

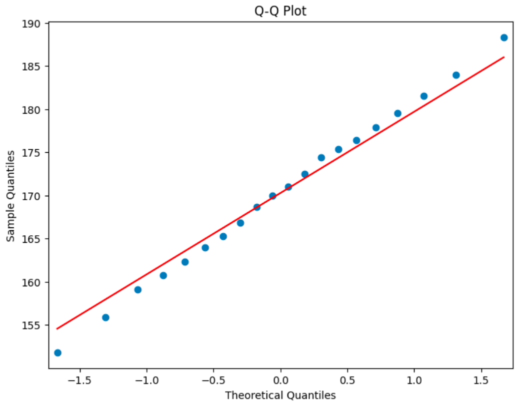
In the Shapiro-Wilk Test, the null hypothesis of normality is rejected for all variables, and in. It is easy for the test to reject null hypotheses of normality when the sample is sufficiently large, as it is in this case for n=2000. Therefore, we do not rely solely on them to arrive at any conclusion but evaluate them in conjunction with other tools discussed.

CENTRAL LIMIT THEOREM EXPLORATION

The Central Limit Theorem (CLT) states that for a sufficiently large sample size, the distribution of a sample variable approximates a normal distribution, and the sample mean and variance will be approximately equal to the mean of the whole population. Since n ≥ 30 is generally considered to be sufficiently large, we study the Height column and consider one case in which the 2000-value column is divided into 20 sequential groups of n=100 samples each, and another in which it is divided into 20 simple random groups of n=100 samples each. Histograms and Q-Q plots were generated in each case, as well as a comparison between the population mean (μ_x), sample mean $(\mu_{\bar{x}})$, sample standard deviation $(\sigma_{\bar{x}})$, and population standard deviation $(\sigma_x/\sqrt{n})$.

Let us first investigate the sequential case.

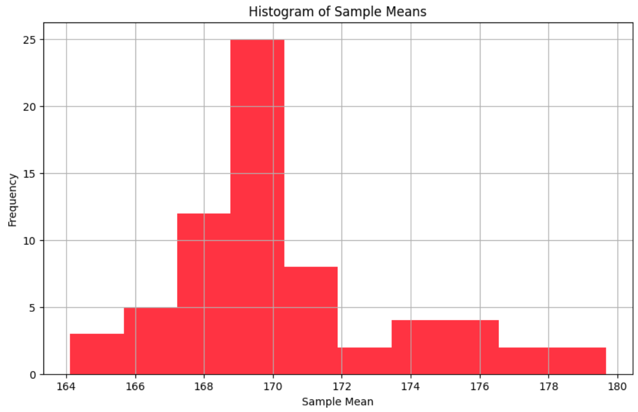


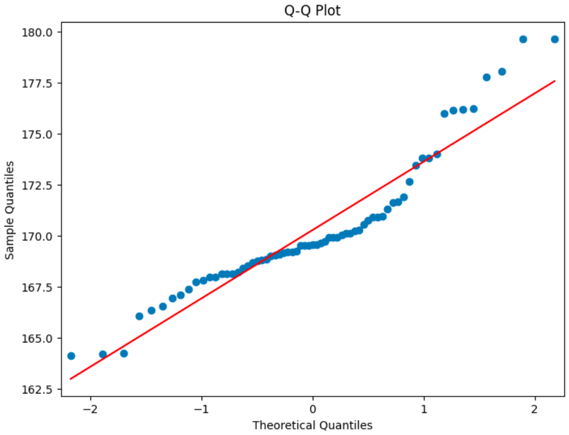


```
Population mean (μx):170.27939705
Mean of sample means (μ x−bar):170.27939705
Population standard deviation (σx):9.46705673863279
Standard deviation of sample mean (σ x−bar):9.421936230962842
Population standard deviation divided by sqrt(n) (σx/sqrt(n)):1.4968731016062493
```

From the, we determine that although the histogram and Q-Q plot generated for the sequentially obtained groups are indicative of normality, the fact that $\sigma_{\bar{x}}$ is much larger than $\sigma_x/\sqrt{n}$ is not consistent with the CLT.

Now for the simple random group case.





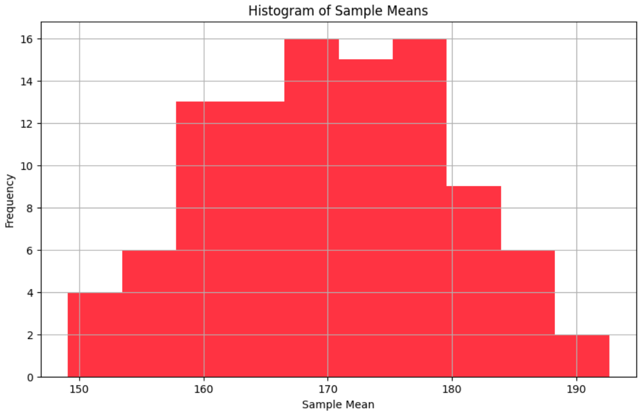
```
Population mean (μx):170.27939705
Mean of sample means (μ x-bar):170.2826378358209
Population standard deviation (σx):9.46705673863279
Standard deviation of sample mean (σ x-bar):3.3497480728613693
Population standard deviation divided by sqrt(n) (σx/sqrt(n)):1.4968731016062493
```

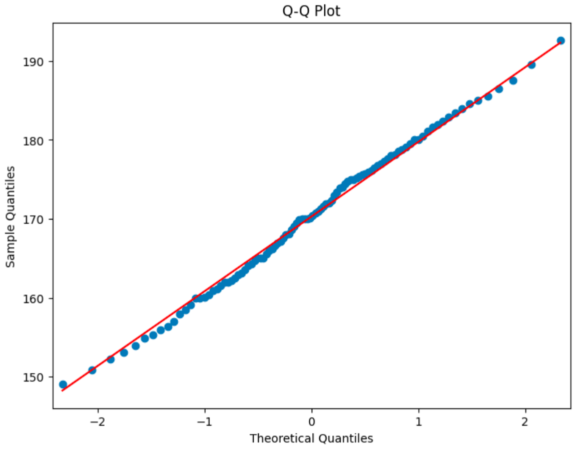
Figure shows that, although the histogram shows some left-skewness for this group, the Q-Q plot is indicative of normality. In addition, the sample and populations are approximately equal, now demonstrating consistency with the CLT.

While minor deviations are observable in extreme quantiles, the overall alignment suggests a robust approximation to normality. This finding corroborates the Central Limit Theorem, asserting that as sample size increases, the distribution of sample means approaches normality, irrespective of the underlying population distribution.

For this portion of the analysis, the same procedure was conducted with groups consisting of n=20 values. Although n = 20 is less than n = 30, meaning that the samples are not considered sufficiently large enough, the output upon dividing sequentially and into simple random samples was the same. $\sigma\bar{x}$ was larger than $\sigma x/\sqrt{n}$ when dividing sequentially and they were approximately equal when dividing into simple random samples.

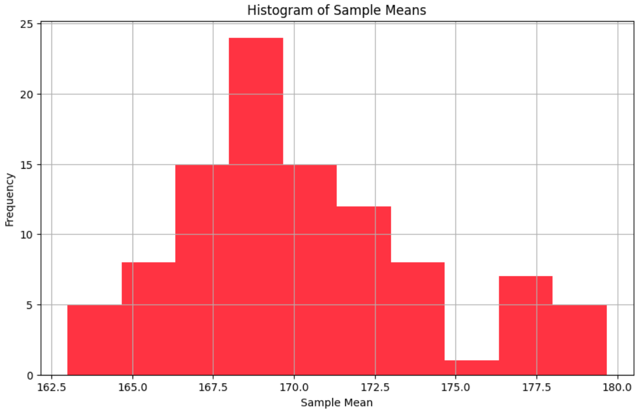
Sequentially:

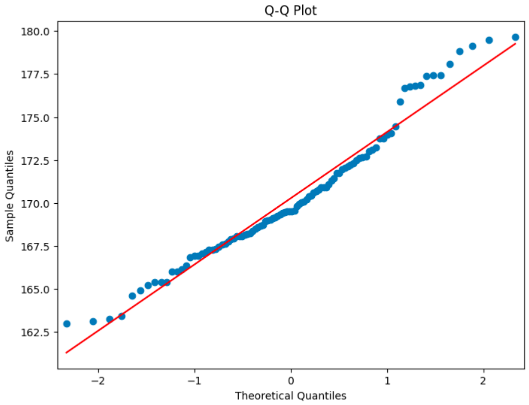




```
Population mean (μx):170.27939705
Mean of sample means (μ x-bar):170.27939705000006
Population standard deviation (σx):9.46705673863279
Standard deviation of sample mean (σ x-bar):9.459878127400167
Population standard deviation divided by sqrt(n) (σx/sqrt(n)):1.4968731016062493
```

S.R





```
Population mean (μx):170.27939705
Mean of sample means (μ x-bar):170.27939705
Population standard deviation (σx):9.46705673863279
Standard deviation of sample mean (σ x-bar):3.85187923392207
Population standard deviation divided by sqrt(n) (σx/sqrt(n)):1.4968731016062493
```

CONSTRUCTING A CONFIDENCE INTERVAL FOR HEIGHT FROM RANDOM SAMPLES

After selecting one of the random samples from the n=20 and n=100 cases, confidence intervals were constructed by using the critical t-value in the case of n=20 and using the Z-score approach for the n=100 case.

```
Confidence Interval for n=20: (166.13035273783547,174.42844136216453)
Confidence Interval for n=100: (168.42388802526824,172.13490607473176)
```

This shows that both intervals are accurate in capturing the population mean of 170.27939705. However, since the n=100 interval is narrower, it is also the more accurate of the two.

Hypothesis Testing Using the Same Random Sample of n=20 and n=100 values

Using the random sample extracted from the n=100 case as was for the confidence intervals calculated in the previous section, we perform the following two-tailed Z-test for a single sample and large sample size.

H0: μ = 170

Ha: μ ≠ 170

```
Z Statistic: -0.4191938645176657
p value: 0.6750744557398731
Fail to reject the null hypothesis
```

Now, the following hypothesis tests were performed on the random sample extracted from the n = 20 case.

H0: μ = 170

Ha: μ ≠ 170

```
Z Statistic: -0.4191938645176657
p value: 0.6750744557398731
Fail to reject the null hypothesis
```

Chi-square investigates variance, and in this case, is used to make inferences about the standard deviation. Using a two-tailed χ^2 test:

H0: σ = 15

Ha: σ ≠ 15

```
Chisqr Statistic: 15.280740740740738
p value: 0.6750744557398731
Reject the null hypothesis
```

Using a one-tailed χ^2 test:

H0: σ = 15

Ha: σ ≠ 15

```
Chisqr Statistic: 15.280740740740738
p value: 0.6750744557398731
Reject the null hypothesis (one-tailed)
```

COMPARING DIFFERENT DATASETS: WEIGHT AND HEIGHT

We now consider the entirety of the gold and oil columns and perform the appropriate two-tailed hypothesis test with α = 0.05.

Using the Z-statistic for large sample size:

1. H0: The means height of male and female are equal.

   Ha: The means height of male and female are not equal.

   ```
   Z-score: 34.98567772147795
   P-value: 0.0
   ```

Since the p-value is less than any reasonable significance level, we reject the null hypothesis.

2. H0: The mean difference between male and female height is zero.

   Ha: The mean difference between male and female height is not zero.

   ```
   Z-score: 3.1667223235198976
   P-value: 0.0
   ```

Since the p-value is less than any reasonable significance level, we reject the null hypothesis.

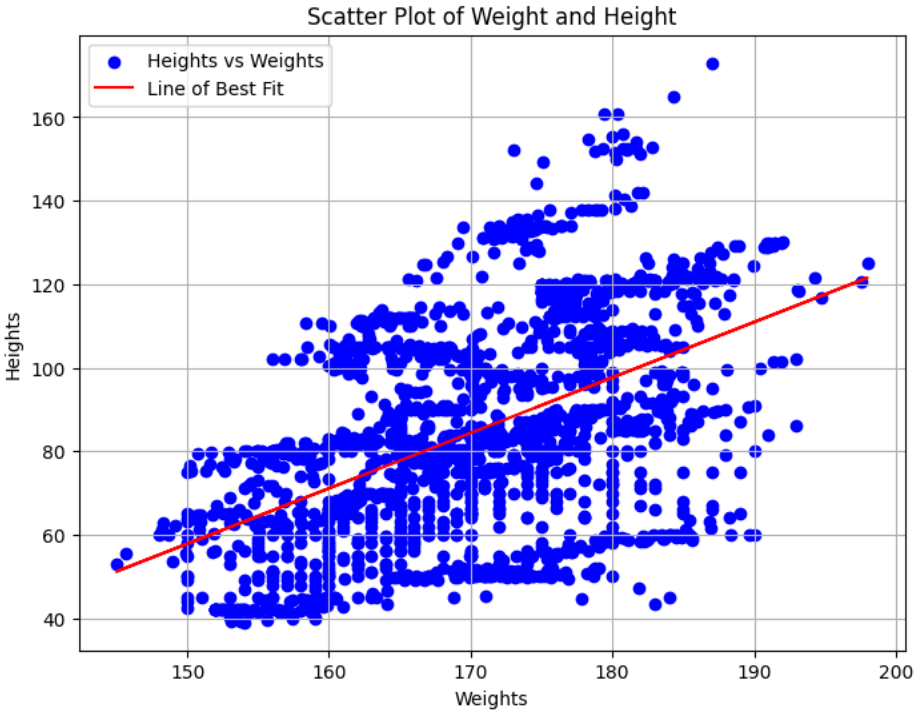
Using the F-statistic for two different variances:

3. H0: The standard deviation of male and female height are equal.
   Ha: The standard deviation of male and female height are not equal.

   ```
   F-statistic: 0.9215277760190127
   Critical value: 1.1371773182072882
   ```

Fail to reject the null hypothesis. The standard deviations of male and female heights are equal.

FITTING A LINE TO HEIGHT AND WEIGHT DATA



```
Slope: 1.3275557912740032
Intercept: -141.3135141688734
Correlation Coefficient: 0.4934783372560992
```

A scatter plot of height versus weight showed a moderate positive correlation, which was confirmed by calculating the correlation coefficient. Linear regression analysis revealed a statistically positive relationship between height and weight.

We now conduct a two-tailed t-test regarding the slope of the linear regression model β_1 with $\alpha = 0.01$.

H0: $\beta_1 = 0$
Ha: $\beta_1 \neq 0$

```
Standard Error of the Slope: 0.05234623163859891
T-statistic: 25.36105751488513
P-value: 0.0
```

At the 1% significance level, we reject H0 since p-value < α = 0.01. Therefore, the linear relationship between Height and Weight is significant, suggesting that changes in the Height can predict changes in the Weight returns.

```
Coefficient of Determination (R^2): 0.24352086934104425
```

The (R^2) value of around 0.244 indicates a moderate level of fit of the regression model to the data. It also suggests that approximately 24.4% of the variance in the dependent variable is explained by the independent variable in the regression model. This means that the model captures some, but not all, of the variation in the dependent variable.

Additionally, the results upon computing the 99% confidence interval of the mean Mean Predicted Weight, and the 99% prediction interval of the Individual Predicted Weights are given below

```
99% Confidence Interval for Mean Predicted Weight: (85.46939929661534,
84.01437174238472)
99% Prediction Interval for Individual Predicted Weights: (117.27729072344903,
117.27729045976654)
```

FITTING A MULTIPLE LINEAR REGRESSION MODEL WITH HEIGHT, NOBESYESDAD, AGE, GENDER AS PREDICTOR VARIABLES, WEIGHT AS RESPONSE VARIABLE

```
Coefficients:
const        -197.403586
Gender         -6.458186
Height          1.510792
Age             0.674997
NObeyesdad      4.019352
R^2: 0.4030643554472395
Adjusted R^2: 0.40186749199951477
```

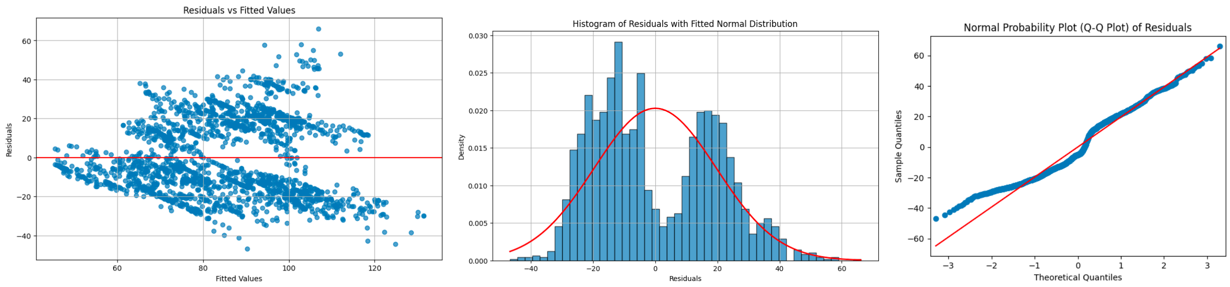
The coefficient of determination is approximately 0.403. This means that approximately 40.3% of the variance in the target variable (weight) is explained by the independent variables (Gender, Height, Age, NObeyesdad) in the model. The adjusted R-squared value is approximately 0.402. This adjusts the R-squared value for the number of predictors and provides a more accurate measure of the model's goodness-of-fit.

The model explains a moderate amount of the variance in weight, with an R-squared value of 0.403.

The weight variation can be explained by factors like

- Gender is negatively associated with weight, indicating that, on average, males tend to weigh less than females.
- Height is positively associated with weight, suggesting that taller individuals tend to weigh more.
- Age is positively associated with weight, indicating that weight tends to increase with age.
- NObeyesdad (likely a categorical variable related to obesity status) is positively associated with weight, suggesting that individuals classified with higher levels of obesity tend to weigh more.

CHECKING THE RESIDUALS OF THE MODEL FITTING



```
Mean residual – −1.8189894035458565e−14
```

The model explains a moderate amount of the variance in weight, with an R-squared value of 0.403.

The four assumptions made for the error terms of the multiple regression model are:

- Mean zero
- Constant Variance
- Normality
- Independence

A mean residual of approximately -1.82e-14 (or close to zero) indicates that, on average, the residuals (the differences between the observed values and the predicted values) are very close to zero. This suggests that the first assumption of the multiple regression model, which states that the mean of the residuals should be zero, is effectively satisfied.

The red curve represents the fitted normal distribution, which suggests that the residuals are approximately normally distributed. The histogram is roughly bell-shaped and symmetric, further indicating that the residuals follow a normal distribution.

The scatter of points appears to be randomly distributed around the horizontal line with no obvious patterns or trends. The spread of the residuals seems to be relatively constant across the range of fitted values, indicating that the assumption of homoscedasticity (constant variance of residuals) is likely met.

Also in the QQ-Plot, we can see that the data points closely follow the diagonal red line, suggesting that the residuals are normally distributed. The points do not deviate substantially from the line, providing further evidence that the normality assumption for the residuals is valid.

One way of improving our model is by using techniques like Regularization which can help improve the generalization of the model by penalizing large coefficients and reducing overfitting. By Scaling our features, we ensure that each feature contributes proportionately to the model's learning process. This helps prevent biases in the model's parameter estimates and ensures that no single feature dominates the others solely due to its scale.In our case, the dataset only had 2000 values to work on. Collecting more data can improve the accuracy of the model.

CONCLUSION

In this statistical analysis of obesity data, we explored various aspects of the dataset consisting of age, height, gender, weight, and obesity levels. Through visualizations and statistical tests, we investigated the distributions of the variables, confirming approximate normality for most features except for some deviations in the tails.

The Central Limit Theorem principles were evaluated, and the findings aligned with the theorem's assertions when samples were drawn randomly. Confidence intervals and hypothesis testing techniques were employed to make inferences about the population parameters, such as the mean height and standard deviation.

Regression analysis revealed a moderate positive correlation between height and weight, with changes in height being able to predict changes in weight to some extent. However, when considering multiple predictors (gender, height, age, and obesity level) in a multiple linear regression model, only a moderate amount of variance in weight could be explained (R-squared of 0.403).

The residual analysis confirmed the validity of the assumptions underlying the multiple regression model, such as normality, constant variance, and independence of residuals. However, the moderate R-squared value suggests that there may be other factors influencing weight that are not captured by the current set of predictors.

While the analysis provided insights into the relationships between the variables, the low correlations among the predictors pose challenges in developing highly accurate predictive models for obesity levels or weight using linear regression techniques. Future work could explore more advanced modeling approaches, feature engineering, or the incorporation of additional relevant variables to enhance the predictive capabilities of the models.