statistics_project_v2

February 8, 2026

1 Statistics Project (M.Sc. Statistics for ML)

Submitters:

- Daniel Attali (3287808789)
- Sapir Bashan (214103368)

GitHub Repo Link:

- https://github.com/dattali18/statistics_msc_course

2 Project Task 1: Building a Database

Objective: To select a specific "population" of objects and build a statistical database for analysis.

Requirements:

1. **Population Selection:** - A distinct group of items (e.g., snacks, fruits, human performance data).
   - **Sample Size:** At least 36 objects. The total count must be a multiple of 4 (e.g., 100, 104).
   - **Sources:** Can be gathered manually or from existing databases (requires approval).
2. **Variables:** - The population must be measured by various parameters:
   - **Continuous:** Weight, height, length, etc.
   - **Categorical:** Color, type, etc.
   - **Nominal/Ordinal:** A ranking scale (e.g., Perfect, Good, Bad).
3. **Data Entry:** - All observations must be organized in a single Excel file.

The data we chose is the "bodyPerformance.csv" dataset from Kaggle, the data has 13393 observations of individuals with various physical performance metrics. and 12 variables. from this we chose a subset of 100 observations for our analysis. And we also chose the variables of 'age', 'gender', 'height', 'weight', 'body fat'. We also changed the 'gender' column values from 'M' and 'F' to '1' and '0' respectively to make it easier for analysis.

2.0.1 Phase 3: Primary Statistical Analysis

Goal: Perform initial statistical calculations on **one central continuous variable** (calculating for more than one yields a bonus).

Required Calculations:

- **Measures of Central Tendency:**

  - **Mean:** The average value.
  - **Median:** The middle value when sorted.
  - **Mode:** The most frequent value.
  - **Mid-range:** (Max + Min) / 2.
- **Measures of Dispersion:**
  - **Range:** Max value - Min value.
  - **Variance:** The average squared deviation from the mean.
  - **Standard Deviation (SD):** The square root of the variance.
  - **Mean Absolute Deviation (MAD):** The average of the absolute differences from the mean.
- **Other:**
  - **Error Percentage/Proportions:** As relevant to the data.

```
(100, 5)
```

	age	gender	height	weight	bf
0	21.0	0	167.4	72.18	40.0
1	42.0	1	162.3	67.30	18.0
2	36.0	1	178.5	90.50	14.7
3	23.0	1	180.9	77.10	25.4
4	53.0	1	177.3	88.48	35.6

2.0.2 Mean (Average)

Formula for Mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

```
age        36.93000
height    168.31600
weight     67.70460
bf         23.94307
Name: Mean, dtype: float64
```

2.0.3 Median

Formula for Median:

- If n is odd: Median = value at position $(n + 1)/2$
- If n is even: Median = average of values at positions $n/2$ and $(n/2) + 1$

```
age        34.50
height    169.05
weight     67.45
bf         24.70
Name: Median, dtype: float64
```

2.0.4 Mode

The mode is the value that appears most frequently in a data set.

```
age        21.0
height    167.2
weight     67.3
bf         25.1
Name: Mode, dtype: float64
```

2.0.5 Mid-range

The mid-range is calculated as:

$$Mid\text{-}range = \frac{Max + Min}{2}$$

```
age        42.50
height    167.15
weight     70.85
bf         24.50
Name: Mid-range, dtype: float64
```

2.0.6 Summery Table of Central Tendency

	Mean	Median	Mode	Mid-range
age	36.93000	34.50	21.0	42.50
height	168.31600	169.05	167.2	167.15
weight	67.70460	67.45	67.3	70.85
bf	23.94307	24.70	25.1	24.50

2.1 Range

The range is calculated as:

$$Range = Max - Min$$

```
age       43.0
height    32.1
weight    55.3
bf        34.6
Name: Range, dtype: float64
```

2.1.1 Variance

The formula for Variance is:

$$\mathbb{V}[X] = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

```
age       157.257677
height     63.911661
weight    143.930318
bf         53.909503
Name: Var, dtype: float64
```

2.1.2 Standard Deviation

The formula for Standard Deviation is:

$$\sigma = \sqrt{\mathbb{V}[X]}$$

```
age       12.540242
height     7.994477
weight    11.997096
bf         7.342309
Name: STD, dtype: float64
```

2.1.3 Mean Absolute Deviation (MAD)

The formula for MAD is:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

```
age       10.600200
height     6.634720
weight     9.743892
bf         5.925762
Name: MAD, dtype: float64
```

2.1.4 Summery Table of Dispersion

	Range	Var	STD	MAD
age	43.0	157.257677	12.540242	10.600200
height	32.1	63.911661	7.994477	6.634720
weight	55.3	143.930318	11.997096	9.743892
bf	34.6	53.909503	7.342309	5.925762

2.1.5 Phase 4: Frequency Distribution & Grouped Analysis

Goal: Organize the raw data into classes (groups) to analyze the frequency distribution.

Steps:

1. **Create a Frequency Table:**
   - Define **Class Width** and create classes (bins).
   - List **Apparent Limits** and **Real Limits (Boundaries)**.
   - Calculate **Frequencies** for each class.
   - Calculate **Cumulative Frequency**.
   - Calculate **Relative Frequency** (proportion of total) and **Cumulative Relative Frequency**.
2. **Calculate Parameters for Grouped Data:**
   - **Grouped Median:** Estimated using interpolation within the median class.
   - **Interquartile Range (IQR):** The difference between the 3rd Quartile (Q_3/P_{75}) and 1st Quartile (Q_1/P_{25}).
   - **Percentiles:** Specifically calculate the **10th Percentile (P_{10})** and the **90th Percentile (P_{90})**.

We will do the frequency distribution for the 'age' variable.

age	Freq	CF	RF	CRF
(20.957, 25.3]	21	21	0.21	0.21
(25.3, 29.6]	16	37	0.16	0.37
(29.6, 33.9]	12	49	0.12	0.49
(33.9, 38.2]	12	61	0.12	0.61
(38.2, 42.5]	6	67	0.06	0.67
(42.5, 46.8]	8	75	0.08	0.75
(46.8, 51.1]	11	86	0.11	0.86
(51.1, 55.4]	2	88	0.02	0.88
(55.4, 59.7]	4	92	0.04	0.92
(59.7, 64.0]	8	100	0.08	1.00

2.1.6 Parameters for Grouped Data

Median Median for grouped data is calculated using the formula:

$$Median = L + \left(\frac{\frac{n}{2} - CF}{f} \right) \times h$$

Where: - L = Lower boundary of the median class - n = Total number of observations - CF = Cumulative frequency of the class before the median class - f = Frequency of the median class - h = Class width

```
Grouped median: 13.47
```

Interquartile Range (IQR) For grouped data, Q1 and Q3 are calculated using the formulas:

$$Q1 = L1 + \left(\frac{(n/4) - CF1}{f1} \right) \times h$$

$$Q3 = L3 + \left(\frac{(3n/4) - CF3}{f3} \right) \times h$$

Where: - $L1$, $L3$ = Lower boundaries of the Q1 and Q3 classes - $CF1$, $CF3$ = Cumulative frequencies before the Q1 and Q3 classes - $f1$, $f3$ = Frequencies of the Q1 and Q3 classes - h = Class width - n = Total number of observations

```
IQR: 20.43
```

Percentiles (P10 and P90) Formula for Percentiles:

$$P_k = L + \left(\frac{(k \cdot n/100) - CF}{f} \right) \times h$$

Where: - P_k = k-th percentile - L = Lower boundary of the percentile class - n = Total number of observations - CF = Cumulative frequency before the percentile class - f = Frequency of the percentile class

```
P10: 23.03, P90: 57.55
```

2.1.7 Summery Table of Grouped Data Parameters

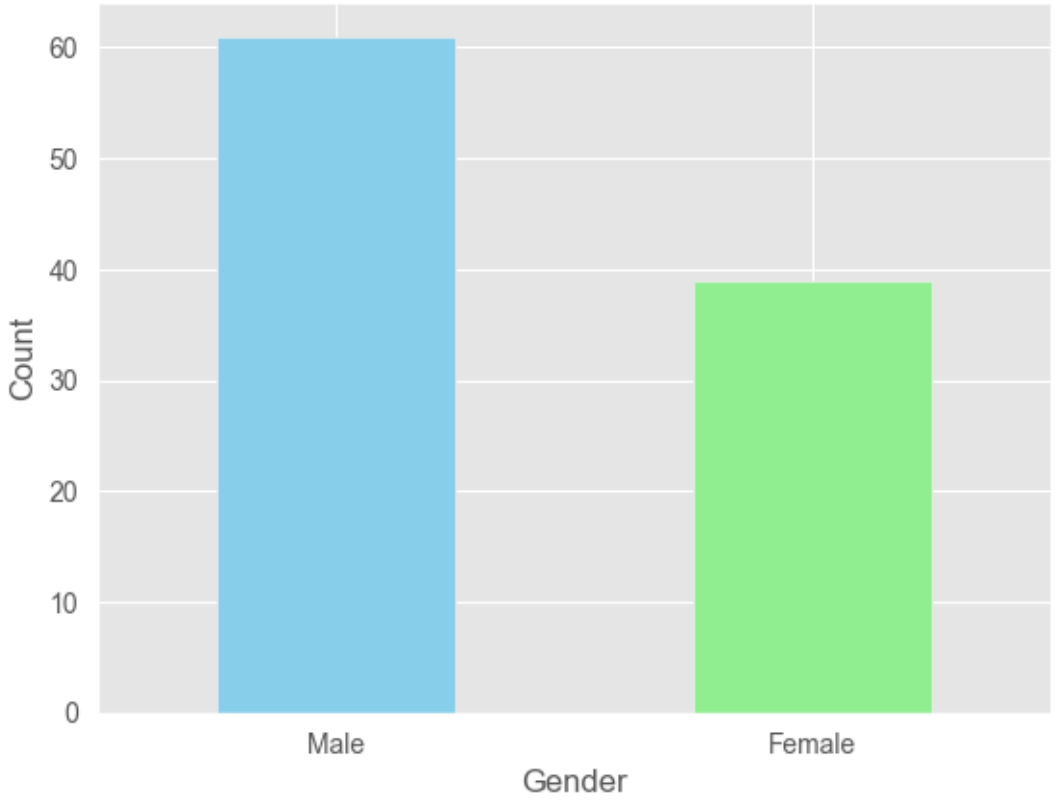
	Grouped Median	IQR	P10	P90
0	13.475	20.425	23.025095	57.55

2.2 Phase 5: Visualization

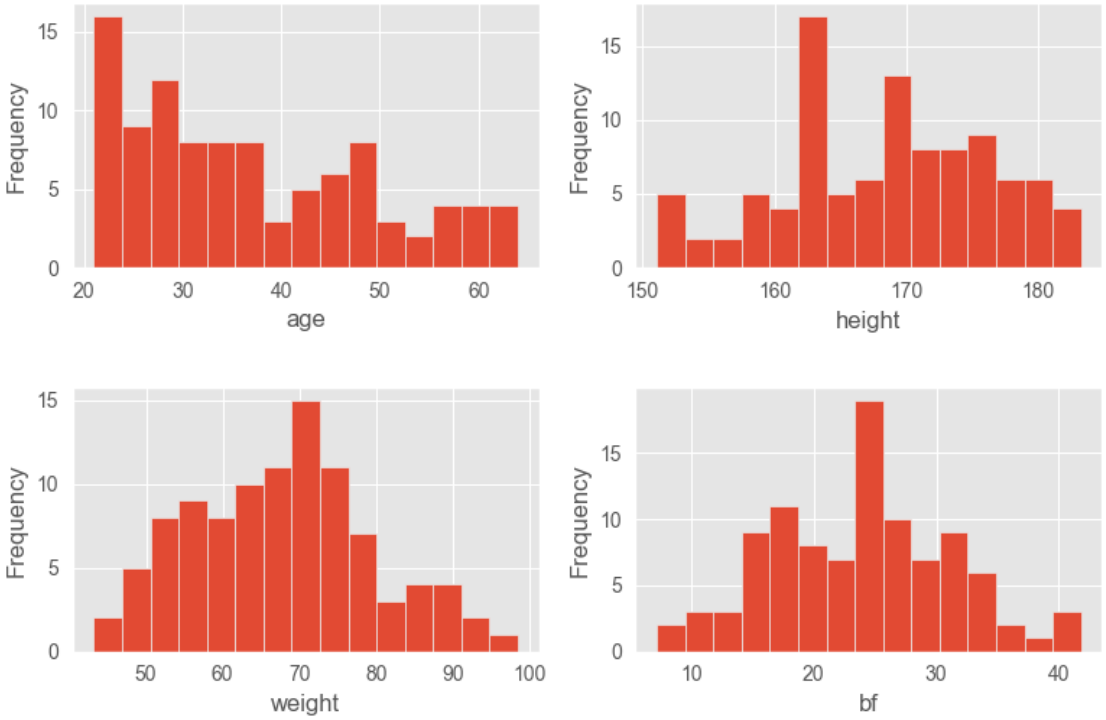
Goal: Present the data visually using specific chart types.

Required Charts: 1. **Bar / Column Chart:** Best for categorical data or simple comparisons. 2. **Histogram:** To show the distribution of the continuous variable (based on the frequency table from Phase 4). 3. **Pie Chart:** To show the relative proportions of categorical parts (e.g., the "Nominal Classification" groups). 4. **Frequency Polygon:** A line graph plotted using the midpoints of the histogram bars. 5. **Ogive (Cumulative Frequency Graph):** A line graph showing the cumulative frequency accumulation.

2.2.1 Bar Plot for gender

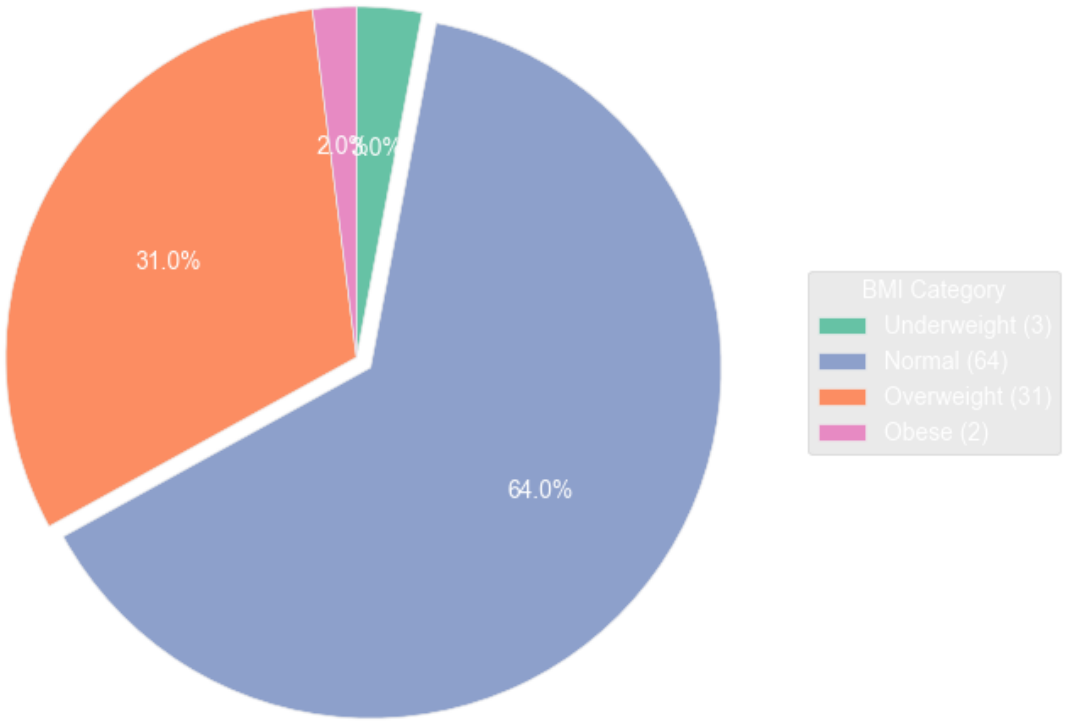


2.2.2 Histogram

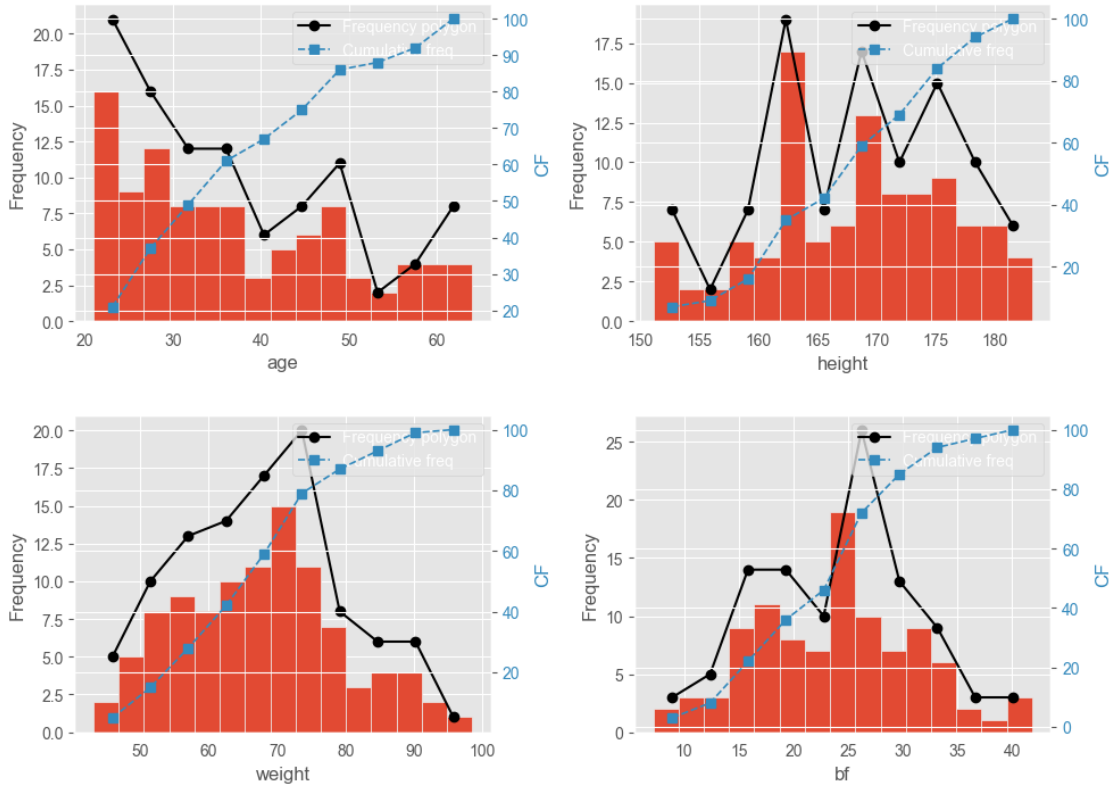


2.2.3 Pie chart

We will create a pie chart for height groups.


3.0%
64.0%
31.0%
2.0%
BMI Category
Underweight (3)
Normal (64)
Overweight (31)
Obese (2)

2.2.4 Frequency Polygon + Ogive



3 Project Task 2: Confidence Intervals (Interval Estimation)

Objective: To build Confidence Intervals (CI) for the central variable chosen in Task 1, under different conditions (known vs. unknown variance, large vs. small samples) and for a proportion.

General Requirements:

- **Target Variable:** Use the central variable from Task 1 (doing more than one is a bonus).
- **Confidence Levels:** For every part, you must calculate the interval for **90%, 95%, and 99%** confidence levels.
- **Interval Length:** In all cases, you must explicitly state the **length/size** of the confidence interval (Upper Bound - Lower Bound).
- **Submission:** Add to the existing Excel file and explain in the Word document.

3.1 Part 1: CI for Mean (Known Sigma)

Scenario: Construct a Confidence Interval for the population mean (μ) assuming the population Standard Deviation (σ) is **known**.

Methodology:

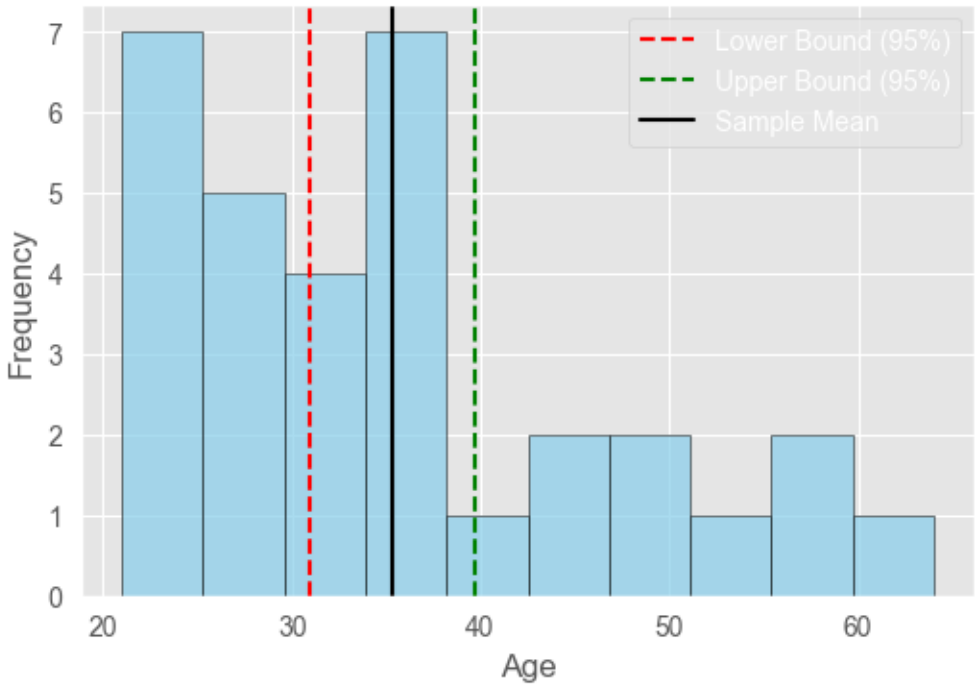
- **Population SD (σ):** Use the standard deviation calculated for the *entire* database in Task 1.
- **Sampling:** Randomly select **at least 32 observations** (but no more than 25% of the total population).
- **Distribution:** Use the **Normal Distribution (Z-table)** since σ is known.
- **Output:** Calculate intervals for 90%, 95%, 99% and state their lengths.

The formula for the confidence interval when sigma is known is:

$$CI = \bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

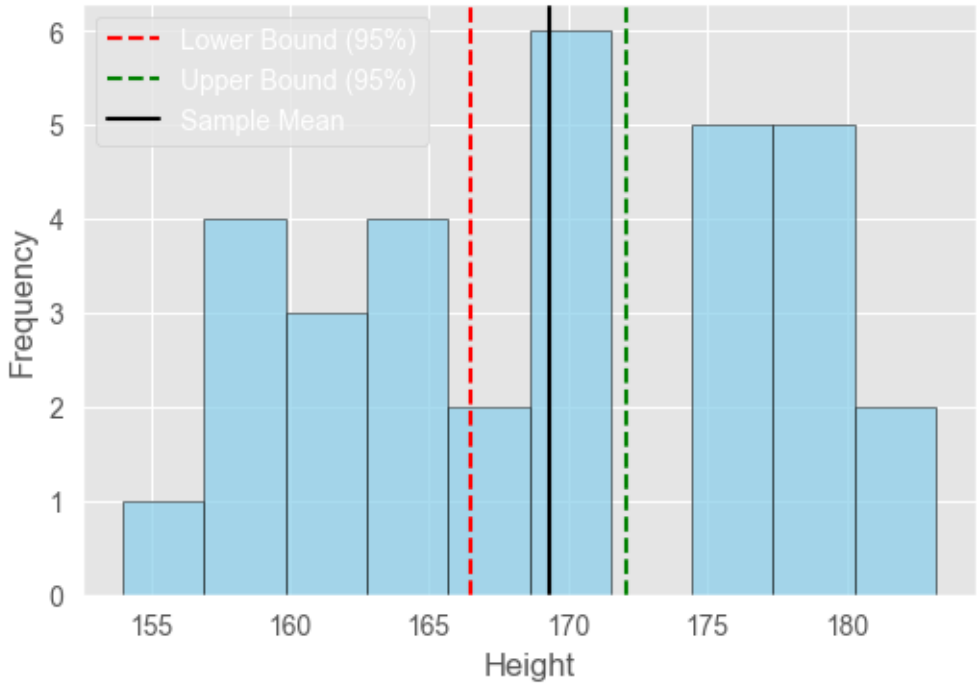
Where: - $\bar{x}$ = sample mean - $Z_{\alpha/2}$ = Z-score for the desired confidence level - σ = population standard deviation - n = sample size

Confidence Level	Lower Bound	Upper Bound	Length
0.90	31.697401	38.990099	7.292697
0.95	30.998858	39.688642	8.689785
0.99	29.633593	41.053907	11.420313



Confidence Level	Lower Bound	Upper Bound	Length
0.90	166.966057	171.615193	4.649137
0.95	166.520731	172.060519	5.539788
0.99	165.650367	172.930883	7.280516

We will try to visualize the confidence intervals for 'age' variable



3.2 Part 2: CI for Mean (Unknown Sigma - Large Sample)

Scenario: Construct a Confidence Interval for the population mean (μ) when the population Standard Deviation is **unknown**.

Methodology:

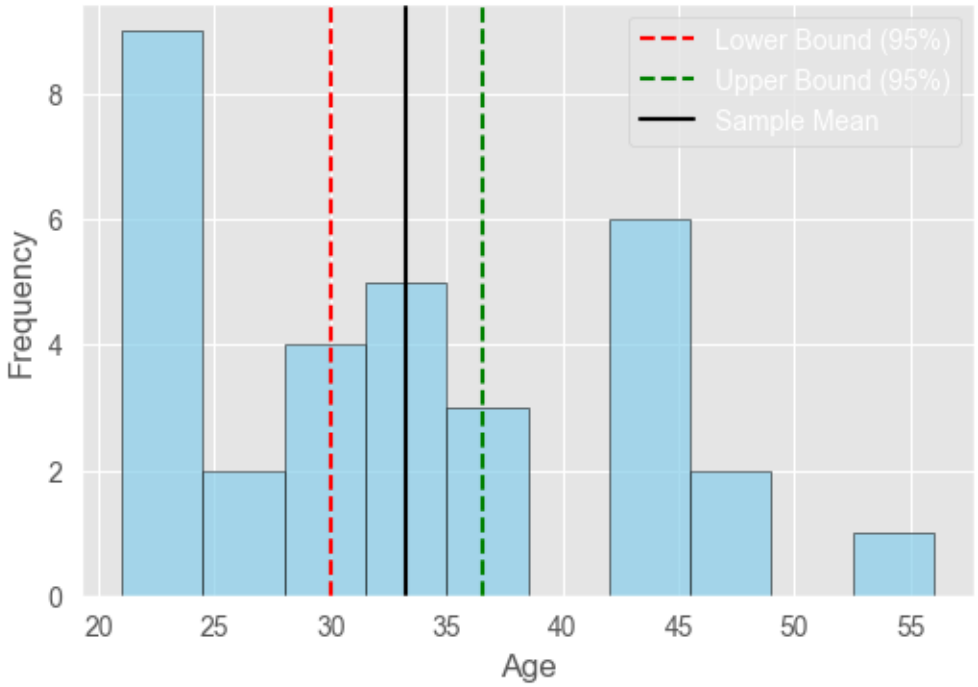
- **Sample SD (S):** Calculate the **unbiased sample standard deviation** (S) from the sample data itself.
- **Sampling:**
  - Randomly select **at least 32 observations** (max 25% of population).
  - **Note:** These observations should be mostly **different** from those selected in Part 1.
- **Output:** Calculate intervals for 90%, 95%, 99% and state their lengths.

The formula for the confidence interval when σ is unknown is:

$$CI = \bar{x} \pm Z_{\alpha/2} \times \frac{S}{\sqrt{n}}$$

Where: - $\bar{x}$ = sample mean - $Z_{\alpha/2}$ = Z-score for the desired confidence level - S = sample standard deviation - n = sample size

Confidence Level	Lower Bound	Upper Bound	Length
0.90	30.502394	35.997606	5.495212
0.95	29.976025	36.523975	6.547949
0.99	28.947268	37.552732	8.605464



3.3 Part 3: CI for Mean (Unknown Sigma - Small Sample)

Scenario: Construct a Confidence Interval for the mean assuming a **Normal Distribution** but with a **small sample size**.

Methodology:

- **Sampling:**
  - Randomly select **at most 16 observations**.
  - **Note:** These observations should be mostly **different** from previous sections.
- **Distribution:** Use the **Student's t-distribution** (t-table) because $n < 30$ and σ is unknown.

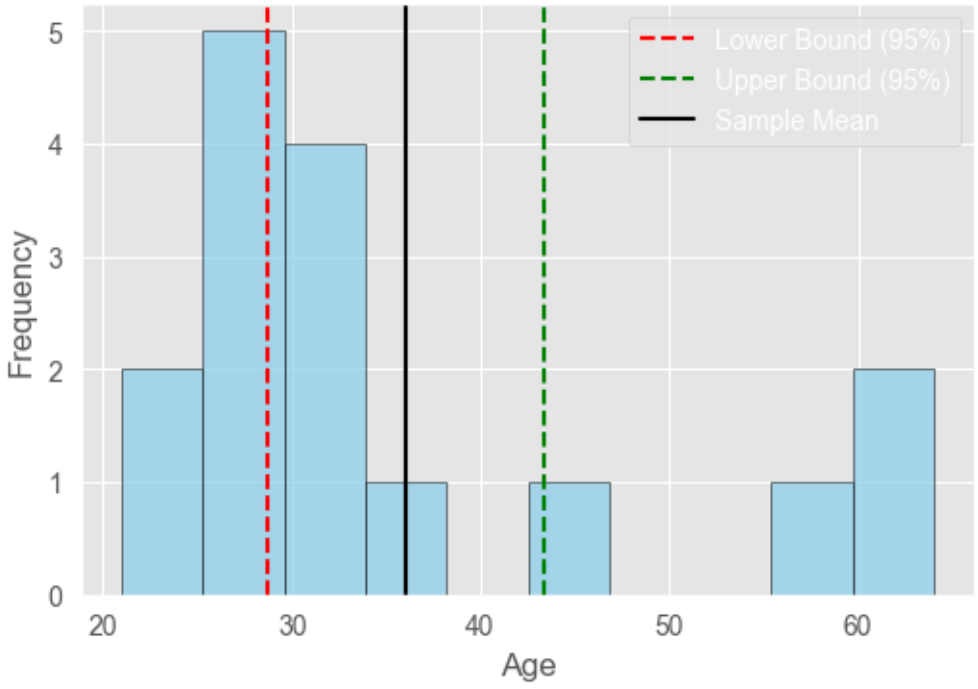
- **Output:** Calculate intervals for 90%, 95%, 99% and state their lengths.

The formula for the confidence interval using the t-distribution is:

$$CI = \bar{x} \pm t_{\alpha/2,df} \times \frac{S}{\sqrt{n}}$$

Where: - $\bar{x}$ = sample mean - $t_{\alpha/2,df}$ = t-score for the desired confidence level and degrees of freedom - S = sample standard deviation - n = sample size - df = degrees of freedom $(n-1)$

Confidence Level	Lower Bound	Upper Bound	Length
0.90	30.031079	42.093921	12.062841
0.95	28.729186	43.395814	14.666628
0.99	25.924249	46.200751	20.276502



3.4 Part 4: CI for Proportion

Scenario: Construct a Confidence Interval for a specific **Proportion (Ratio)** within your population.

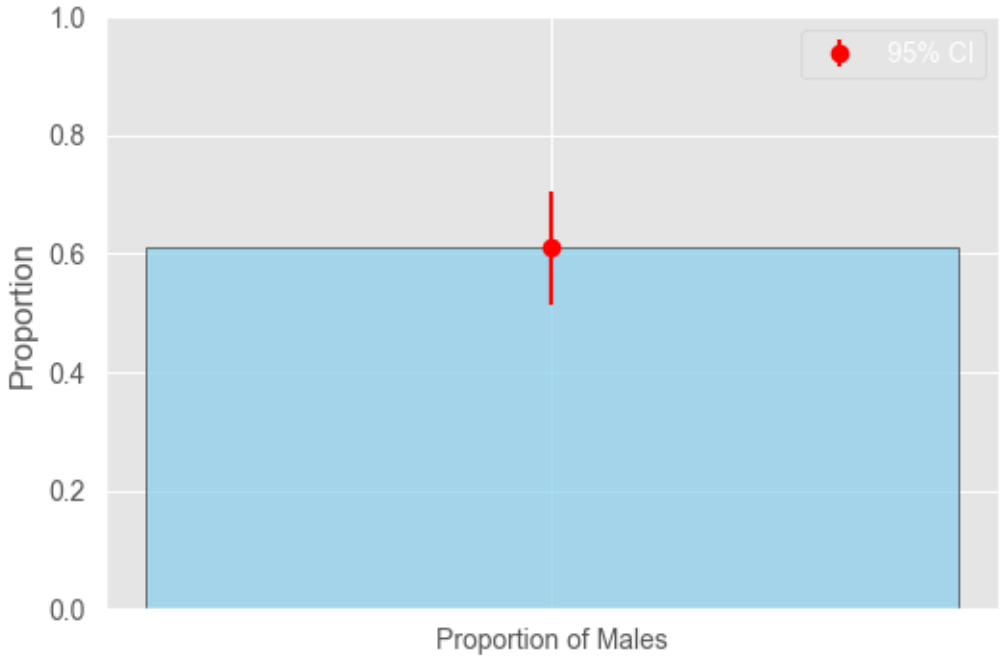
Methodology: * **Define Success:** Define a binary criteria for your variable (e.g., "% of athletes with Height > 180cm", "% of products classified as 'Grade A' "). * **Calculation:** Calculate the sample proportion ($\hat{p}$) and build the interval. * **Output:** Calculate intervals for 90%, 95%, 99% and state their lengths.

The formula for the confidence interval for a proportion is:

$$CI = \hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Where: - $\hat{p}$ = sample proportion - $Z_{\alpha/2}$ = Z-score for the desired confidence level - n = sample size

Confidence Level	Lower Bound	Upper Bound	Length
0.90	0.529772	0.690228	0.160455
0.95	0.514403	0.705597	0.191194
0.99	0.484364	0.735636	0.251272



4 Project Task 3: Hypothesis Testing, ANOVA, and χ^2 Tests

Objective: To perform various hypothesis tests on your database, including tests for differences in means, variance analysis (ANOVA), and Chi-Square tests for fit and independence.

General Requirements: * **Target Variable:** Continue using the central variable from previous tasks. * **Confidence Levels:** For every test, you must determine the conclusion (Reject/Fail to Reject H_0) at **90%, 95%, and 99%** confidence levels. * **Submission:** Add to the existing Excel file and explain in the Word document.

4.1 Part 1: Difference in Means (Small Sample / T-Test)

Scenario: Compare the means of **2 specific groups** (out of the 4+ groups in your data) using small samples.

Methodology: * **Sampling:** * **Group A:** Randomly sample **6 to 9** observations. * **Group B:** Randomly sample **10 to 16** observations. * **Hypothesis (H_1):** Formulate an alternative hypothesis (e.g., $\mu_A \neq \mu_B$ or $\mu_A > \mu_B$) based on the sample data. * **Test:** Perform a Hypothesis Test for the difference of means (assumed Student's t-distribution due to small sample). * **Output:** Determine if H_0 is rejected at 90%, 95%, and 99%.

The formula for the t-test for difference of means is:

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

Where: - $\bar{X}_A$, $\bar{X}_B$ = sample means of groups A and B - S_A^2, S_B^2 = sample variances of groups A and B - n_A, n_B = sample sizes

We reject H_0 if the calculated p-value is less than the significance level (α).

Let's formulate the t-test for difference of means for 'weight' and 'height' between 'male' and 'female

$$H_0 : \mu_{male} - \mu_{female} = 0$$

The alternative hypothesis H_1 is that there is a difference in means:

$$H_1 : \mu_{male} - \mu_{female} \neq 0$$

	Group A (Female)	Group B (Male)
0	56.00	92.5
1	62.68	63.8
2	56.70	67.3
3	50.90	62.8
4	53.90	64.4

```
T-statistic: -4.2016, P-value: 0.0004
```

Alpha Level	Conclusion
0.10	Reject H0
0.05	Reject H0
0.01	Reject H0

Conclusion: Based on the p-values obtained from the t-test for the 'weight' variable, we can say for all significance levels (10%, 5%, 1%), we fail to reject H_0 meaning there is significant difference between the means of weight between the two groups, meaning the average weight for male and female are statistically different.

```
Group A Mean: 159.97777777777776
Group B Mean: 172.01875
Group A Variance: 39.016944444444476
Group B Variance: 40.952291666666696
Degrees of Freedom: 17.062915930758063
T-statistic: -4.5857, P-value: 0.0003

             Conclusion
 Alpha Level
 0.10         Reject H0
 0.05         Reject H0
 0.01         Reject H0
```

Conclusion: Based on the p-values calculated, for the 'height' variable, we can say for all significance levels (10%, 5%, 1%), we reject H_0 meaning there is a significant difference between the means of height between the two groups ("Female" and "Male").

4.2 Part 2: Difference in Means (Large Sample / Z-Test)

Scenario: Compare the means of the **other 2 groups** (those not used in Part 1) using large samples.

Methodology: * **Sampling:** * Randomly sample **at least 32 observations** from each group. * **Limit:** Do not sample more than 90% of the total group size. * **Hypothesis (H_1):** Formulate an alternative hypothesis, preferably different from the type used in Part 1 (e.g., if Part 1 was two-tailed, make this one one-tailed). * **Test:** Perform a Hypothesis Test for the difference of means (Z-test or Large Sample T-test). * **Output:** Determine if H_0 is rejected at 90%, 95%, and 99%.

The formula for the Z-test for difference of means is:

$$Z = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

Where: - $\bar{X}_A$, $\bar{X}_B$ = sample means of - σ_A^2, σ_B^2 = population variances of groups A and B - n_A, n_B = sample sizes

The p-value is calculated based on the Z-statistic and the nature of the alternative hypothesis.

We reject H_0 if the calculated p-value is less than the significance level (α).

Let's formulate the Z-test for difference of means:

We will try to compare the means of 'bf' (body fat percentage) between 'male' and 'female' populations.

The H_0 hypothesis is that there is no difference in means:

$$H_0 : \mu_{male} - \mu_{female} = 0$$

The alternative hypothesis H_1 is that the mean body fat percentage is higher for men

$$H_1 : \mu_{male} - \mu_{female} > 0$$

```
Z-statistic: 5.9267, P-value: 0.0000
```

Alpha Level	Conclusion
0.10	Reject H0
0.05	Reject H0
0.01	Reject H0

Conclusion: Based on the p-values obtained from the Z-test for the 'bf' variable, we can say for all significance levels (10%, 5%, 1%), we reject H_0 meaning there is a significant difference between the means of body fat percentage between the two groups, meaning the average body fat percentage for male and female are statistically different.

4.3 Part 3: Test for Variance Ratio (F-Test)

Scenario: Check if the variances of the two groups from **Part 1** (the small samples) are equal. This validates the assumptions made in the T-test.

Methodology: * **Data:** Use the exact same samples sampled in Part 1. * **Test:** Perform an F-test for the ratio of variances (S_1^2/S_2^2). * **Analysis:** * Was the assumption of equal/unequal variances in Part 1 correct? * Compare these sample variances to the **True Variances** (σ^2) calculated from the full population in Task 1. How much did they deviate? * **Output:** Determine if H_0 ($\sigma_1^2 = \sigma_2^2$) is rejected at 90%, 95%, and 99%.

The formula for the F-test for ratio of variances is:

$$F = \frac{S_1^2}{S_2^2}$$

Where: - S_1^2, S_2^2 = sample variances - n_1, n_2 = sample sizes - $df_1 = n_1 - 1$, $df_2 = n_2 - 1$ = degrees of freedom

Let's formulate the F-test for ratio of variances for 'weight' variable between male and female

The H_0 hypothesis is that the variances are equal:

$$H_0 : \sigma_{male}^2 = \sigma_{female}^2$$

The alternative hypothesis H_1 is that the variances are not equal:

$$H_1 : \sigma_{male}^2 \neq \sigma_{female}^2$$

```
F-statistic: 2.3753, P-value: 0.2186
```

Alpha Level	Conclusion
0.10	Fail to Reject H0

```
0.05          Fail to Reject H0
0.01          Fail to Reject H0
```

Conclusion: Based on the p-values obtained from the F-test for the 'weight' variable, we can say for all significance levels (10%, 5%, 1%), we fail to reject H_0 meaning there is no significant difference between the variances of weight between the two groups, meaning the variances for male and female are statistically equal. This validates the assumption made in Part 1 of equal variances.

4.4 Part 4: One-Way ANOVA

Scenario: Test if there is a statistically significant difference between the means of **all** your groups simultaneously.

Methodology: * **Sampling (Specific Percentages):** * **Group 1:** Sample ~15%. * **Group 2:** Sample ~25%. * **Group 3:** Sample ~35%. * **Other Groups:** Continue the pattern (round up to the nearest integer). * **Calculations:** Calculate Sum of Squares: * **SST** (Total) * **SSB** (Between Groups) * **SSW** (Within Groups) * **Output:** Perform the ANOVA F-test and conclude at 90%, 95%, and 99%.

The formulas for ANOVA calculations are:

$$SST = \sum_{i=1}^{N} (X_i - \bar{X})^2$$

$$SSB = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2$$

$$SSW = SST - SSB$$

The statistic for the ANOVA F-test is:

$$F = \frac{MSB}{MSW} = \frac{SSB/(k-1)}{SSW/(N-k)}$$

Where: - N = total number of observations - k = number of groups - $\bar{X}$ = overall mean - $\bar{X}_j$ = mean of group j - n_j = number of observations in group j - MSB = mean square between groups - MSW = mean square within groups

```
SST: 1936.4764, SSB: 887.0409, SSW: 1049.4355
F-statistic: 8.1708, P-value: 0.0004
```

```
             Conclusion
Alpha Level
0.10          Reject H0
0.05          Reject H0
0.01          Reject H0
```

Conclusion: Based on the p-values obtained from the One-Way ANOVA for the 4 different groups based on 'height' variable we reject the H_0 at all significance levels (10%, 5%, 1%) meaning there is a significant difference between the means of height across the different groups.

4.5 Part 5: χ^2 Goodness of Fit

Scenario: Test if your data follows a specific theoretical distribution (e.g., Uniform, Normal, or a specific ratio).

Methodology: * **Proposal:** Propose a distribution that fits your data logic (e.g., "The heights are distributed Normally" or "The product colors are distributed uniformly 25% each"). * **Sampling:** Sample ~**30%** of the total observations. * **Test:** Perform a Chi-Square Goodness of Fit test (χ^2). * **Output:** Determine if the data fits the proposed distribution at 90%, 95%, and 99%.

The formula for the χ^2 Goodness of Fit test is:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where: - O_i = observed frequency for category i - E_i = expected frequency for category i - k = number of categories - The degrees of freedom is $df = k - 1$.

Let's formulate the χ^2 Goodness of Fit test for our data:

We will test height and assume normal distribution. (the heights are distributed Normally, with $\mu = 171$ and $sigma = 7$ for males

We will sample ~30% of the total observations (for males only).

	Bins	Observed Frequency
0	166.5	2
1	169.5	4
2	172.5	5
3	175.5	4
4	178.5	3

	Bins	Observed Frequency	Expected Frequency	(O-E)^2 / E
0	166.5	2	2.674275	0.170007
1	169.5	4	3.408680	0.102579
2	172.5	5	3.418076	0.732132
3	175.5	4	2.696453	0.630174
4	178.5	3	1.673428	1.051610

```
Chi-Square Statistic: 2.6865, P-value: 0.6116
```

Alpha Level	Conclusion
0.10	Fail to Reject H0
0.05	Fail to Reject H0
0.01	Fail to Reject H0

Conclusion: Based on the p-values obtained from the Chi-Square Goodness of Fit test for the 'height' variable among males is distributed normally, we fail to reject the H_0 at all significance levels (10%, 5%, 1%) meaning there is no significant difference between the observed and expected frequencies, indicating that the heights of males follow a normal distribution.

4.6 Part 6: χ^2 Test for Independence

Scenario: Check if two categorical variables in your database are dependent (related) or independent.

Methodology: * **Selection:** Choose 2 variables (e.g., "Gender" and "Performance Class", or "Brand" and "Quality Rating"). * **Hypothesis:** H_0: The variables are independent. * **Test:** Perform a χ^2 Test for Independence. * **Output:** Determine if there is a relationship between the variables at 90%, 95%, and 99%.

The formula for the χ^2 Test for Independence is:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where: - O_{ij} = observed frequency in cell (i,j) - $E_{ij} = \frac{n_i \cdot n_j}{n}$ - n_i = total observations in row i - n_j = total observations in column j - n = total observations - r = number of rows - c = number of columns

Let's formulate the χ^2 Test for Independence for our data:

We will try to check if there is a relationship between 'height' category and 'weight' category.

weight_cat	Light	Medium	Heavy
height_cat			
Short	15	3	0
Medium	17	30	2
Tall	0	21	12

```
Chi-Square Statistic: 48.9326, P-value: 0.0000
```

Alpha Level	Conclusion
0.10	Reject H0
0.05	Reject H0
0.01	Reject H0

Conclusion: Based on the p-values obtained from the Chi-Square Test for Independence between 'height' category and 'weight' category, we reject the H_0 at all significance levels (10%, 5%, 1%) meaning there is a significant relationship between height and weight categories, indicating that these two categorical variables are dependent.

5 Project Task 4: Linear Regression

Objective: To perform simple linear regression analysis on different pairs of variables from your dataset. You will calculate regression lines, visualize the data, predict values, and test for statistical significance.

General Requirements: * **Method:** Simple Linear Regression (1 independent variable predicting 1 dependent variable). * **Significance Levels:** All hypothesis tests must be evaluated at **10% ($\alpha = 0.10$), 5% ($\alpha = 0.05$), and 1% ($\alpha = 0.01$).** * **Calculations:** You must show/output

the intermediate calculations (slope, intercept, means) not just the final result. * **Structure:** * **Model 1:** First pair of variables ($X_1 \to Y_1$). * **Model 2:** Second pair of completely different variables ($X_2 \to Y_2$). * **Model 3:** New independent variable affecting a previous dependent variable ($X_3 \to Y_1$ or Y_2).

5.1 Part 1: Regression Model A (Variables $X_1 \to Y_1$)

1. Variable Selection * Identify two variables where a relationship exists. * Define the **Independent Variable (X)** (Explanatory/Influencer). * Define the **Dependent Variable (Y)** (Explained/Affected).

2. Calculations * Calculate the regression coefficients: **Slope (b_1)** and **Intercept (b_0)**. * Print intermediate calculations (e.g., Means $\bar{x}, \bar{y}$, Sum of Squares SS_{xy}, SS_{xx}). * Formulate the Regression Equation: $\hat{y} = b_0 + b_1 x$

3. Visualization * Plot the **Regression Line**. * Scatter plot **at least 5 actual data points** (observations) from your dataset on the same graph to visualize the fit.

4. Prediction * Choose **3 specific values** for X. * Calculate the predicted $\hat{y}$ for these values using your equation.

5. Significance Testing * **Regression Significance:** Perform a hypothesis test on the model (F-test or T-test for slope). * *Check at:* 10%, 5%, and 1% significance levels. * **Pearson Correlation:** Calculate Pearson's Correlation Coefficient (r) and test its significance. * *Check at:* 10%, 5%, and 1% significance levels.

```
['age',
 'gender',
 'height',
 'weight',
 'bf',
 'gender_str',
 'BMI',
 'BMI_cat',
 'height_cat',
 'weight_cat']
```

From working with our dataset we started with 'age', 'gender', 'height', 'weight', 'bf' but we added more variables like 'BMI', and categories for height, weight, gender, and BMI.

So we will try to find relationships between the following variables:

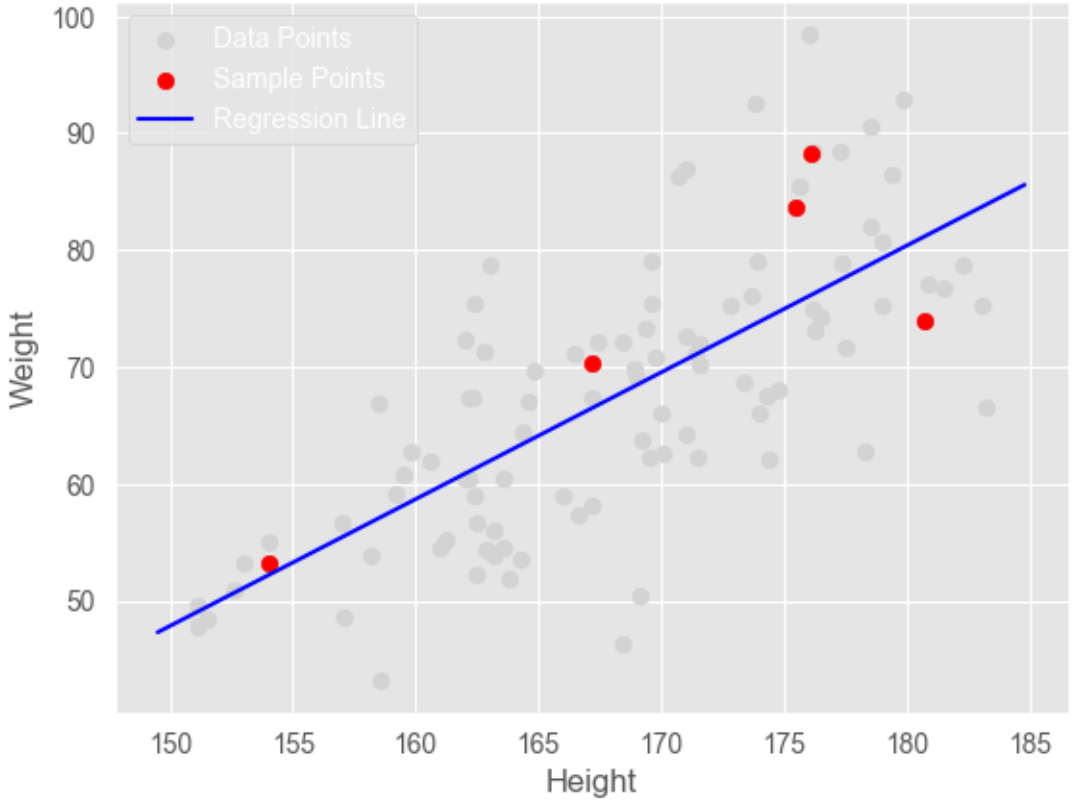
1. Part 1: X1 = height, Y1 = weight
2. Part 2: X2 = weight, Y2 = bf
3. Part 3: X3 = BMI, Y1 = weight

	X1 (height)	Y1 (weight)
0	167.4	72.18
1	162.3	67.30
2	178.5	90.50
3	180.9	77.10

```
 4        177.3        88.48

n: 100
Mean X1 (height): 168.3160
Mean Y1 (weight): 67.7046
SS_xy: 6862.6186
SS_xx: 6327.2544

Slope (b1): 1.0846
Intercept (b0): -114.8530
Regression Equation: ŷ = -114.8530 + 1.0846x
```



Let's now apply the statistical tests for significance.

Let's define the hypotheses for regression significance:

$$H_0 : b_1 = 0$$

$$H_1 : b_1 \neq 0$$

```
F-statistic: 107.1791, P-value (Regression Significance): 0.0000
```

Alpha Level	Conclusion
0.10	Reject H0
0.05	Reject H0
0.01	Reject H0

Conclusion: Based on the p-values obtained from the regression significance test for the relationship between height and weight, we reject the H_0 at all significance levels (10%, 5%, 1%) meaning there is a significant linear relationship between height and weight.

Let's now calculate Pearson's Correlation Coefficient (r) and test its significance. The formula for Pearson's Correlation Coefficient is:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

Where: - $SS_{yy} = \sum(y_i - \bar{y})^2$ - r ranges from -1 to 1, indicating the strength and direction of the linear relationship.

```
Pearson's Correlation Coefficient (r): 0.7228
```

From a $\rho = 0.72$ we can see the there is a strong positive relationship between height and weight.

Let's use the table for ρ values:

The $df = n - 2 = 100 - 2 = 98$

The α values are $0.10, 0.05, 0.01$

So from the table $0.164, 0.195, 0.254$

So a value of $|0.72|$ is greater than all the critical values so we can reject H_0 at all significance levels.

5.2 Part 2: Regression Model B (Variables $X_2 \rightarrow Y_2$)

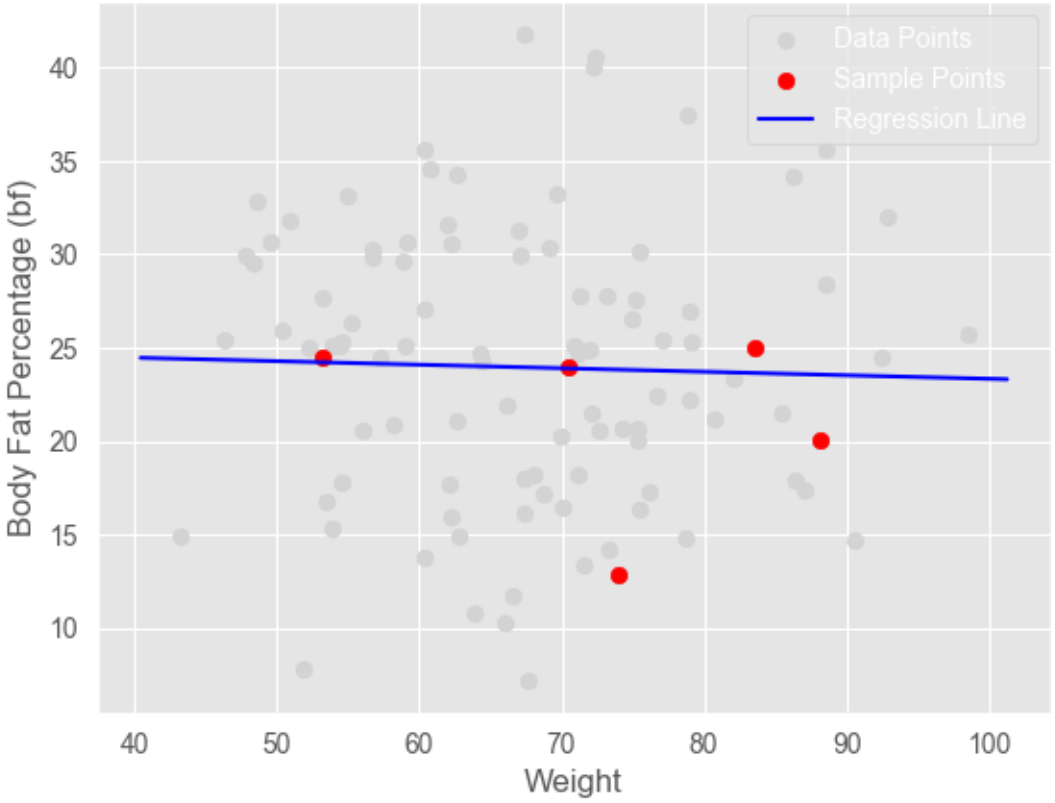
Requirement: Repeat the entire process from Part 1 but with **two new variables** (different columns from the dataset).

Steps: 1. **Define Variables:** Select new Independent (X_2) and Dependent (Y_2) variables. 2. **Equation:** Calculate coefficients and state the equation $\hat{y} = b_0 + b_1 x$. 3. **Plot:** Graph the line and overlay at least 5 sample data points. 4. **Predict:** Choose 3 new X values and predict $\hat{y}$. 5. **Test:** * Test Regression Significance (at 10%, 5%, 1%). * Test Pearson Correlation Significance (at 10%, 5%, 1%).

	X2 (weight)	Y2 (bf)
0	72.18	40.0
1	67.30	18.0
2	90.50	14.7
3	77.10	25.4
4	88.48	35.6

```
n: 100
Mean X2 (weight): 67.7046
Mean Y2 (bf): 23.9431
SS_xy: -269.1473
SS_xx: 14249.1015

Slope (b1): -0.0189
Intercept (b0): 25.2219
Regression Equation: ŷ = 25.2219 + -0.0189x
```




```
F-statistic: 0.0934, P-value (Regression Significance): 0.7605

                    Conclusion
Alpha Level
0.10         Fail to Reject H0
0.05         Fail to Reject H0
0.01         Fail to Reject H0
```

Conclusion: Based on the p-values obtained from the regression significance test for the relationship between weight and body fat percentage, we fail to reject the H_0 at all significance levels (10%, 5%, 1%) meaning there is no significant linear relationship between weight and body fat percentage.

Let's calculate the Pearson's Correlation Coefficient (ρ) and test its significance.

```
Pearson's Correlation Coefficient (r): -0.0309
```

For a $\rho = -0.0309$ we can see the there is a very weak negative relationship between weight and body fat percentage.

The values for the degrees of freedom is $df = n - 2 = 100 - 2 = 98$ are the same as before.

So for all significance levels (10%, 5%, 1%) we fail to reject H_0 meaning there is no significant correlation between weight and body fat percentage.

5.3 Part 3: Regression Model C (Variables $X_3 \rightarrow Y_1$ or Y_2)

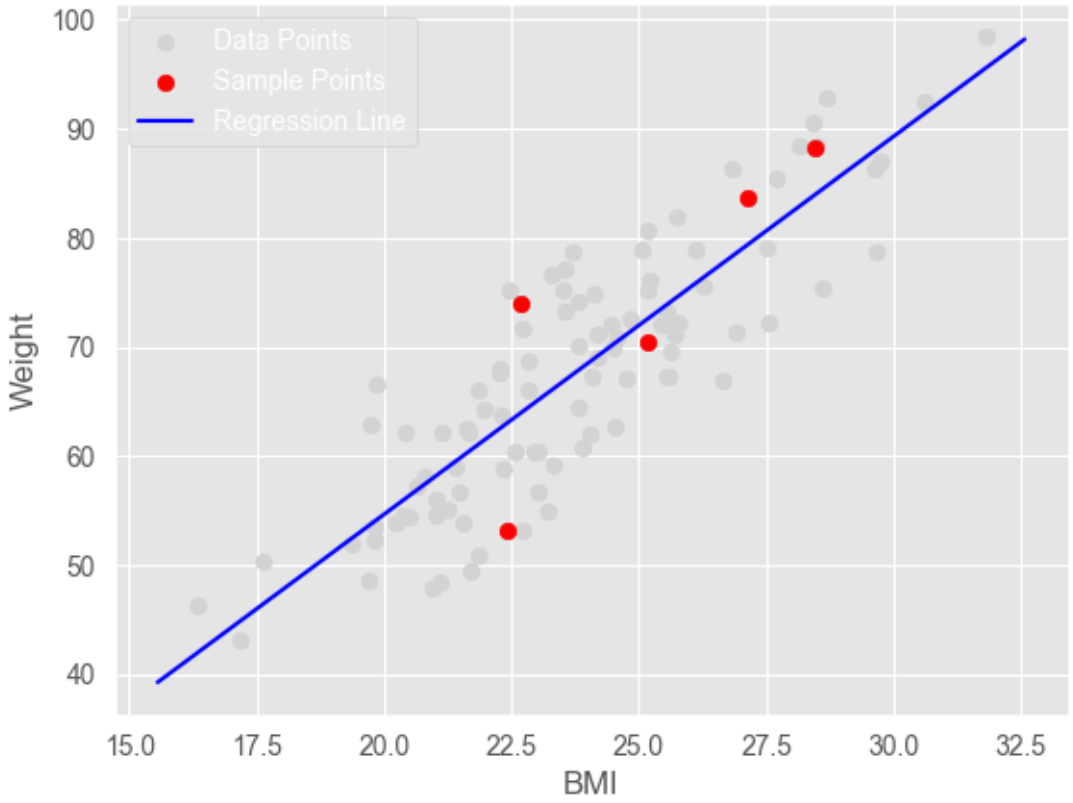
Requirement: Find a **new independent variable** (X_3) that affects **one of the dependent variables** you already used in Part 1 or Part 2.

Steps: 1. **Define Variables:** Select a new Independent variable (X_3) and reuse a previous Dependent variable (Y_{old}). 2. **Equation:** Calculate coefficients and state the equation $\hat{y} = b_0 + b_1 x$. 3. **Plot:** Graph the line and overlay at least 5 sample data points. 4. **Predict:** Choose 3 new X values and predict $\hat{y}$. 5. **Test:** * Test Regression Significance (at 10%, 5%, 1%). * Test Pearson Correlation Significance (at 10%, 5%, 1%).

```
    X3 (BMI)  Y1 (weight)
0  25.757634        72.18
1  25.549242        67.30
2  28.403518        90.50
3  23.560106        77.10
4  28.146710        88.48

n: 100
Mean X3 (BMI): 23.7661
Mean Y1 (weight): 67.7046
SS_xy: 3082.4739
SS_xx: 890.0484

Slope (b1): 3.4633
Intercept (b0): -14.6036
Regression Equation: ŷ = -14.6036 + 3.4633x
```




```
F-statistic: 292.7494, P-value (Regression Significance): 0.0000
```

Alpha Level	Conclusion
0.10	Reject H0
0.05	Reject H0
0.01	Reject H0

Conclusion: Based on the p-values obtained from the regression significance test for the relationship between BMI and weight, we reject the H_0 at all significance levels (10%, 5%, 1%) meaning there is a significant linear relationship between BMI and weight.

Let's calculate the Pearson's Correlation Coefficient (ρ) and test its significance.

```
Pearson's Correlation Coefficient (r): 0.8656
```

The value for $\rho = 0.86$ indicates a very strong positive relationship between BMI and weight.

Using the table provided earlier for hypothesis testing of Pearson's Correlation Coefficient, with $df = n - 2 = 100 - 2 = 98$, we can see that for all significance levels (10%, 5%, 1%) we can reject H_0 meaning there is a significant correlation between BMI and weight.

6 Project Task 5: Non-Parametric Tests & Paired Samples

Objective: To perform advanced hypothesis testing including Paired Sample tests, Sign Tests, and Rank tests (Wilcoxon/Mann-Whitney). These tests often focus on the Median rather than the Mean and do not necessarily assume a Normal distribution.

General Requirements: * **Variable:** Use one or two central variables from your database. * **Confidence Levels:** For every test, determine the conclusion (Reject/Fail to Reject H_0) at **90%, 95%, and 99%**. * **Methodology:** * **Small Samples:** Use exact probability tables or specific small-sample formulas. * **Large Samples:** Use the **Normal Approximation** with **Continuity Correction**. * **Choice:** Questions 1-3 are mandatory. You must choose **one** additional question from 4, 5, or 6 (doing more is a bonus).

6.1 Part 1: Paired Samples Test (Parametric)

Scenario: Test the difference between two paired samples (e.g., "Before vs. After" or two different measurements on the *same* object).

Methodology: 1. **Small Sample Case:** * Randomly select **26 observations** (pairs). * Calculate the differences ($d = x_1 - x_2$). * Perform a T-test on the mean of differences (μ_d). 2. **Large Sample Case:** * Randomly select **72 observations**. * Use the **Normal Approximation** (Z-test) and apply **Continuity Correction**. 3. **Hypothesis (H_1):** Formulate a logical alternative hypothesis based on your data. 4. **Output:** Determine rejection at 90%, 95%, and 99%.

```
Mean Difference (d_bar): 0.7304
Std Dev of Diff (s_d):   1.5650
Standard Error (SE):     0.3069
T-Statistic:             2.3798
P-Value:                 0.0253

Conclusions (Small Sample):
                    Conclusion
Alpha Level
0.10                 Reject H0
0.05                 Reject H0
0.01         Fail to Reject H0

Mean Difference (d_bar): 0.6905
Z-Statistic:             3.4331
P-Value:                 0.0006

Conclusions (Large Sample):
            Conclusion
Alpha Level
0.10         Reject H0
0.05         Reject H0
0.01         Reject H0
```

6.2 Part 2: Sign Test (One Sample - Median)

Scenario: A non-parametric test to check if the population **Median** differs from a specific value ($H_0 : Median = M_0$).

Methodology: 1. **Small Sample Case:** * Randomly select **18 observations**. * Count the number of "Positive" (+) and "Negative" (-) signs relative to the hypothesized median. * Use the Binomial distribution (exact calculation). 2. **Large Sample Case:** * Randomly select **42 observations**. * Use the **Normal Approximation** for the Binomial distribution with **Continuity Correction**. 3. **Hypothesis:** Formulate an alternative hypothesis. 4. **Output:** Determine rejection at 90%, 95%, and 99%.

```
Variable: height
Hypothesized Median (M0): 169.05
H0: Median = M0
H1: Median != M0 (Two-tailed)

--- Small Sample (n=18) Results ---
Positive Signs (+): 9
Negative Signs (-): 9
Effective n:        18
Test Statistic (S): 9
P-Value:            1.1855

Conclusions:
Alpha 0.1: Fail to Reject H0
Alpha 0.05: Fail to Reject H0
Alpha 0.01: Fail to Reject H0

--- Large Sample (n=42) Results ---
Positive Signs (+): 17
Negative Signs (-): 25
Effective n:        42
Expected Mean (mu): 21.00
Std Dev (sigma):    3.2404
Z-Statistic:        1.0801
P-Value:            0.2801

Conclusions:
Alpha 0.1: Fail to Reject H0
Alpha 0.05: Fail to Reject H0
Alpha 0.01: Fail to Reject H0
```

6.3 Part 3: Sign Test (Two Paired Samples)

Scenario: A non-parametric test to compare two paired groups to see if one tends to be larger than the other (testing the Median of differences).

Methodology: 1. **Small Sample Case:** * Randomly select **21 observations** (pairs). * Determine the sign of the difference for each pair (+ if $A > B$, − if $A < B$). 2. **Large Sample Case:** * Randomly select **36 observations**. * Use the **Normal Approximation** with **Continuity Cor-**

rection. 3. **Hypothesis:** Formulate an alternative hypothesis. 4. **Output:** Determine rejection at 90%, 95%, and 99%.

```
--- Small Sample (n=21) Results ---
Positive Differences (+): 15
Negative Differences (-): 6
Effective n:              21
Test Statistic (S):       6
P-Value:                  0.0784

Conclusions:
Alpha 0.1: Reject H0
Alpha 0.05: Fail to Reject H0
Alpha 0.01: Fail to Reject H0
```

```
--- Large Sample (n=36) Results ---
Positive Differences (+): 26
Negative Differences (-): 10
Effective n:              36
Expected Mean (mu):       18.00
Std Dev (sigma):          3.0000
Z-Statistic:              2.5000
P-Value:                  0.0124

Conclusions:
Alpha 0.1: Reject H0
Alpha 0.05: Reject H0
Alpha 0.01: Fail to Reject H0
```

6.4 Option A (Q4): Wilcoxon Signed-Rank Test (One Sample)

- **Goal:** Test the Median of a single population (more powerful than the Sign Test).
- **Small Sample:** 17 observations.
- **Large Sample:** 49 observations (use Normal Approx + Continuity Correction).

```
Task 5 Q4: One-Sample Wilcoxon Test
Testing if Median(height) != 170.0

--- CASE 1: Small Sample (n=17) ---
Effective n:     16
Sum Ranks (+):   63.5
Sum Ranks (-):   72.5
Min Rank Sum (T):63.5
P-Value:         0.8160

--- CASE 2: Large Sample (n=49) ---
Effective n:     48
Sum Ranks (+):   423.0
Sum Ranks (-):   753.0
```

```
Expected Mean:   588.00
Std Dev (sigma): 97.4987
Z-Statistic:     1.6872
P-Value:         0.0916
```

6.5 Option B (Q5): Wilcoxon Signed-Rank Test (Two Paired Samples)

- **Goal:** Compare two paired groups taking into account the *magnitude* of differences, not just the sign.
- **Small Sample:** 10 observations.
- **Large Sample:** 52 observations (use Normal Approx + Continuity Correction).

```
========================================
Task 5 Q5: Paired Wilcoxon Test
Comparing 'weight' vs 'weight_after'

--- CASE 1: Small Sample (n=10) ---
Sum Ranks (+):   44.0
Sum Ranks (-):   11.0
Min Rank Sum (T):11.0
P-Value:         0.1055

--- CASE 2: Large Sample (n=52) ---
Effective n:     52
Sum Ranks (+):   1032.0
Sum Ranks (-):   346.0
Z-Statistic:     3.1191
P-Value:         0.0018
Conclusions:
  Alpha 0.1: Reject H0
  Alpha 0.05: Reject H0
  Alpha 0.01: Reject H0
```

6.6 Option C (Q6): Mann-Whitney U Test (Two Independent Samples)

- **Goal:** Compare distributions of two *independent* groups (Non-parametric equivalent to the independent T-test).
- **Small Sample:** 8 observations.
- **Large Sample:** 72 observations (use Normal Approx + Continuity Correction).

```
Group 0 count: 3
Group 1 count: 5
--- Small Sample (n=8) Results ---
U-Statistic: 0.0
P-Value:     0.0357

Conclusions:
Alpha 0.1: Reject H0
```

```
Alpha 0.05: Reject H0
Alpha 0.01: Fail to Reject H0

--- Large Sample (n=72) Results ---
n1: 31, n2: 41
Rank Sum R1:    585.5
U-Statistic:    89.5
Expected Mean:  635.50
Std Dev (sigma):87.9313
Z-Statistic:    6.2037
P-Value:        0.0000

Conclusions:
Alpha 0.1: Reject H0
Alpha 0.05: Reject H0
Alpha 0.01: Reject H0
```