

# Exploratory Data Analysis and Linear Regression on Body Performance Data

## (A Bio Statistics Project)

Submitted by:

Anjali Mehra

<https://github.com/mehraanjali/Exploratory-Data-Analysis-and-Linear-Regression-on-Body-Performance-Data-biostatistics>

### **Data Collection**

I collected data through Kaggle by exploring the platform's extensive repository of datasets and selecting one that aligned with my research interests. After identifying a suitable dataset, I downloaded it and conducted thorough preprocessing to ensure its quality and relevance. Kaggle's community discussions and resources proved invaluable in guiding me through challenges and providing insights on best practices.

## **About Data**

The dataset comprises a comprehensive collection of physical and health-related measurements from a diverse sample of individuals, with a total of 13,393 observations and 12 distinct variables. The dataset includes a wide range of demographic and physiological information, such as age, gender, height, weight, body fat percentage, diastolic and systolic blood pressure, grip force (a measure of hand strength), and exercise performance metrics like sit-ups counts and broad jump distance. Notably, the dataset covers individuals between the ages of 20 and 64, with gender information specifying "F" for females and "M" for males. Each individual is assigned a performance grade labeled as "A," "B," "C," or "D," reflecting their level of achievement. This dataset offers a comprehensive foundation for exploring relationships between various physical attributes, exercise capabilities, and performance outcomes, thereby providing valuable insights into the interplay of these factors and their potential implications for health and fitness.

- **Data Shape:** The dataset contains a total of 13,393 rows and 12 columns.
- **Columns:**
  1. **age:** The age of the individuals, ranging from 20 to 64.
  2. **gender:** The gender of the individuals, with possible values "F" (female) and "M" (male).

3. **height\_cm**: The height of the individuals in centimeters. If you want to convert to feet, divide by 30.48.
4. **weight\_kg**: The weight of the individuals in kilograms.
5. **body fat\_%**: The percentage of body fat.
6. **diastolic**: The diastolic blood pressure (minimum value).

**Diastolic Blood Pressure (diastolic)**: Diastolic blood pressure is the lower of the two numbers in a blood pressure reading. It represents the pressure in the arteries when the heart is resting between beats. A high diastolic blood pressure can indicate potential health concerns like hypertension (high blood pressure).

7. **systolic**: The systolic blood pressure (minimum value).

**Systolic Blood Pressure (systolic)**: Systolic blood pressure is the higher of the two numbers in a blood pressure reading. It represents the pressure in the arteries when the heart contracts or beats. High systolic blood pressure can also be a sign of hypertension.

8. **gripForce**: Grip force, which is a measure of hand strength.
9. **sit and bend forward\_cm**: Measurement of how far an individual can bend forward while sitting.
10. **sit-ups counts**: The number of sit-ups performed by each individual.
11. **broad jump\_cm**: The distance jumped in a broad jump, measured in centimeters.
12. **class**: A categorical variable indicating the grade of performance, with categories "A", "B", "C", and "D". "A" represents the best performance, while "D" represents lower performance. The class is stratified.

The dataset appears to contain various physical and health-related measurements, along with performance metrics such as sit-ups counts and broad jump distances. The "class" variable indicates the performance grade of each individual, with "A" being the highest grade.

## **Objective**

The objective of this analysis is to uncover and understand the relationships between key physical attributes (height\_cm, weight\_kg, body fat\_%) and exercise performance outcomes (sit-ups counts and broad jump\_cm). By conducting regression analysis and interpreting the coefficients, the goal is to identify the extent to which these predictor variables influence the corresponding exercise metrics. This analysis aims to provide actionable insights for exercise planning, fitness optimization, and performance enhancement. Furthermore, the exploration of these relationships contributes to a comprehensive understanding of how physiological factors impact specific exercise

capabilities, potentially guiding tailored training approaches and health management strategies.

## **Methodology**

### **1) Data Collection and Preparation:**

- Gather a dataset that includes variables of interest, such as height\_cm, weight\_kg, body fat\_%, sit-ups counts, and broad jump\_cm.

### **2) Descriptive Analysis:**

- Conduct initial exploratory data analysis to understand the distribution and summary statistics of the variables.

- Create histograms, box plots, and descriptive statistics to gain insights into the data.

### 3) **Regression Analysis:**

- Perform multiple linear regression analysis for each outcome variable (sit-ups counts and broad jump\_cm).
- Use predictor variables (height\_cm, weight\_kg, body fat\_%) as independent variables in separate regression models.
- Interpret the regression coefficients, p-values, and confidence intervals for each predictor.

### 4) **Visualizations:**

- Construct multiple visualizations to know more about the data.
- Create scatter plots with regression lines to visually depict the relationships between predictor variables and outcome variables.

### 5) **Interpretation of Results:**

- Interpret the regression coefficients to understand the impact of each predictor variable on the outcome variables.
- Evaluate the statistical significance of predictor variables based on p-values.
- Discuss the direction and strength of relationships (positive/negative) between variables.

### 6) **Conclusion:**

- Summarize the key findings and insights from the analysis.

- Reiterate the significance of understanding the relationships between physical attributes and exercise performance.

## Importing Libraries and loading the data

```
import math
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
[5] df = pd.read_csv('/content/bodyPerformance.csv')
df.head()
```

	age	gender	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend	forward_cm	sit-ups counts	broad jump_cm	class
0	27.0	M	172.3	75.24	21.3	80.0	130.0	54.9		18.4	60.0	217.0	C
1	25.0	M	165.0	55.80	15.7	77.0	126.0	36.4		16.3	53.0	229.0	A
2	31.0	M	179.6	78.00	20.1	92.0	152.0	44.8		12.0	49.0	181.0	C
3	32.0	M	174.5	71.10	18.4	76.0	147.0	41.4		15.2	53.0	219.0	B
4	28.0	M	173.8	67.70	17.1	70.0	127.0	43.5		27.1	45.0	217.0	B

## Descriptive Analysis

0s

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13393 entries, 0 to 13392
Data columns (total 12 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   age                                   13393 non-null  float64
1   gender                               13393 non-null  object
2   height_cm                           13393 non-null  float64
3   weight_kg                           13393 non-null  float64
4   body fat_%                           13393 non-null  float64
5   diastolic                           13393 non-null  float64
6   systolic                            13393 non-null  float64
7   gripForce                           13393 non-null  float64
8   sit and bend forward_cm             13393 non-null  float64
9   sit-ups counts                      13393 non-null  float64
10  broad jump_cm                       13393 non-null  float64
11  class                               13393 non-null  object
dtypes: float64(10), object(2)
memory usage: 1.2+ MB
```

df.describe()

	age	height_cm	weight_kg	body fat_%	diastolic	systolic	gripForce	sit and bend forward_cm	sit-ups counts	broad jump_cm
count	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000	13393.000000
mean	36.775106	168.559807	67.447316	23.240165	78.796842	130.234817	36.963877	15.209268	39.771224	190.129627
std	13.625639	8.426583	11.949666	7.256844	10.742033	14.713954	10.624864	8.456677	14.276698	39.868000
min	21.000000	125.000000	26.300000	3.000000	0.000000	0.000000	0.000000	-25.000000	0.000000	0.000000
25%	25.000000	162.400000	58.200000	18.000000	71.000000	120.000000	27.500000	10.900000	30.000000	162.000000
50%	32.000000	169.200000	67.400000	22.800000	79.000000	130.000000	37.900000	16.200000	41.000000	193.000000
75%	48.000000	174.800000	75.300000	28.000000	86.000000	141.000000	45.200000	20.700000	50.000000	221.000000
max	64.000000	193.800000	138.100000	78.400000	156.200000	201.000000	70.500000	213.000000	80.000000	303.000000

df.skew()

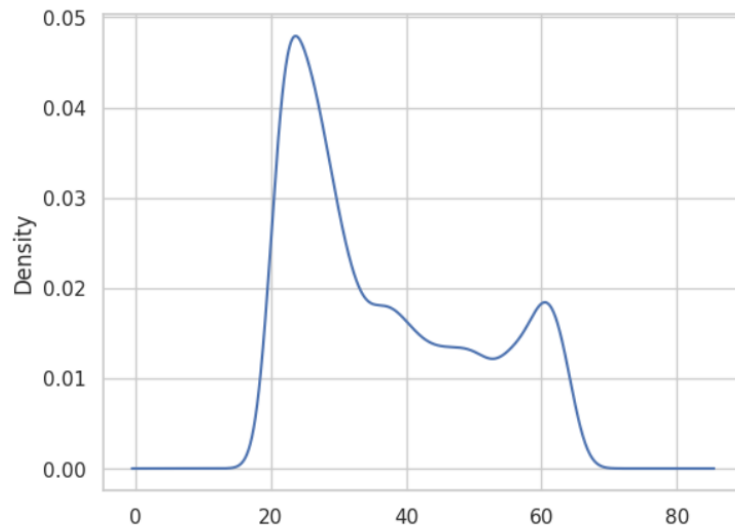
```
<ipython-input-58-9e0b1e29546f>:1: FutureWarning: The default value of numeric_only
df.skew()
age                                0.599896
height_cm                         -0.186882
weight_kg                         0.349805
body fat_%                        0.361132
diastolic                        -0.159637
systolic                         -0.048654
gripForce                        0.018456
sit and bend forward_cm          0.785492
sit-ups counts                   -0.467830
broad jump_cm                   -0.422623
dtypes: float64
```



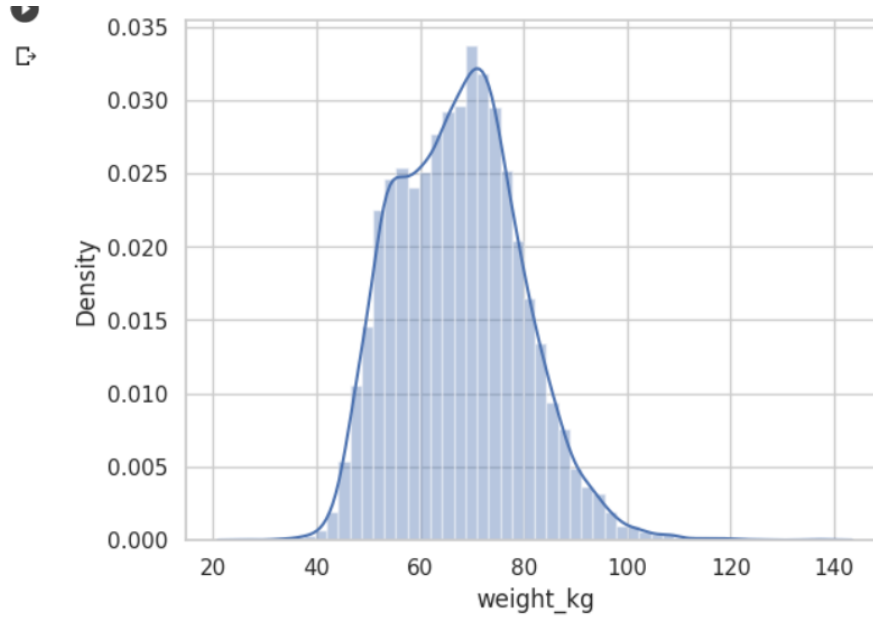
## **Visualizations**

```
df['age'].plot(kind = 'density')
```

<Axes: ylabel='Density'>

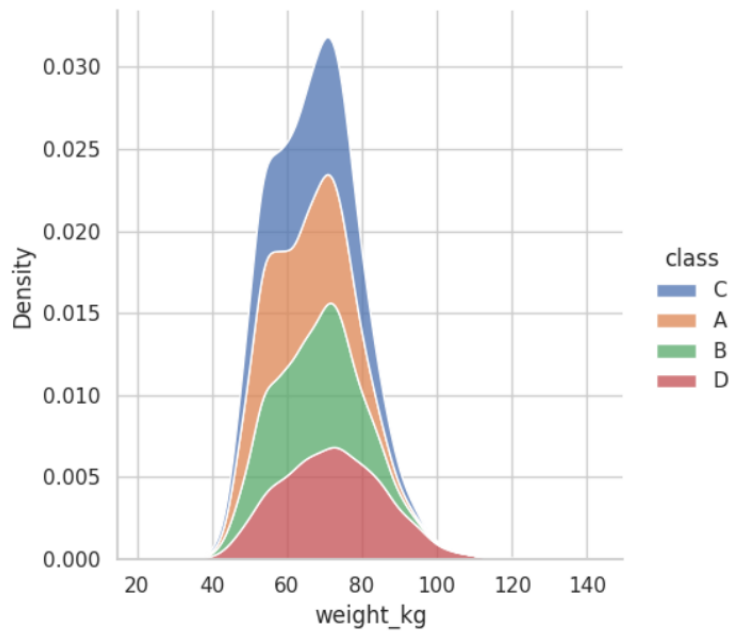


```
#Normal Distribution/Symmetric  
sns.distplot(df['weight_kg'], hist=True, kde=True)
```

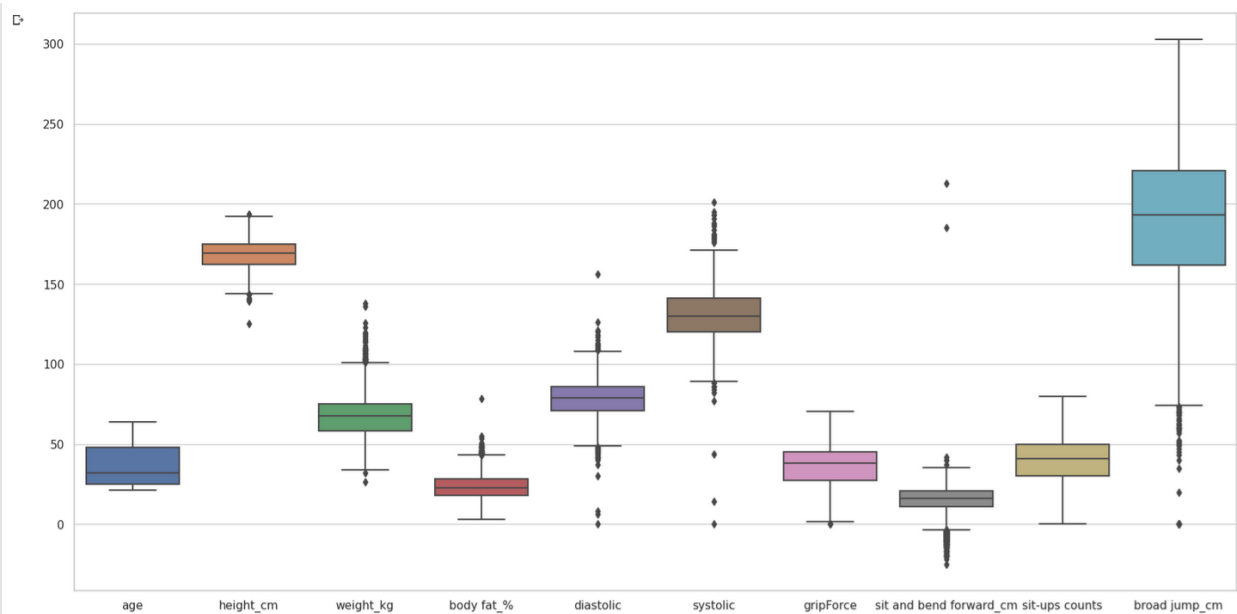


```
sns.displot(df, x="weight_kg", hue="class", kind="kde", multiple="stack")
```

<seaborn.axisgrid.FacetGrid at 0x7e5753ce5e10>

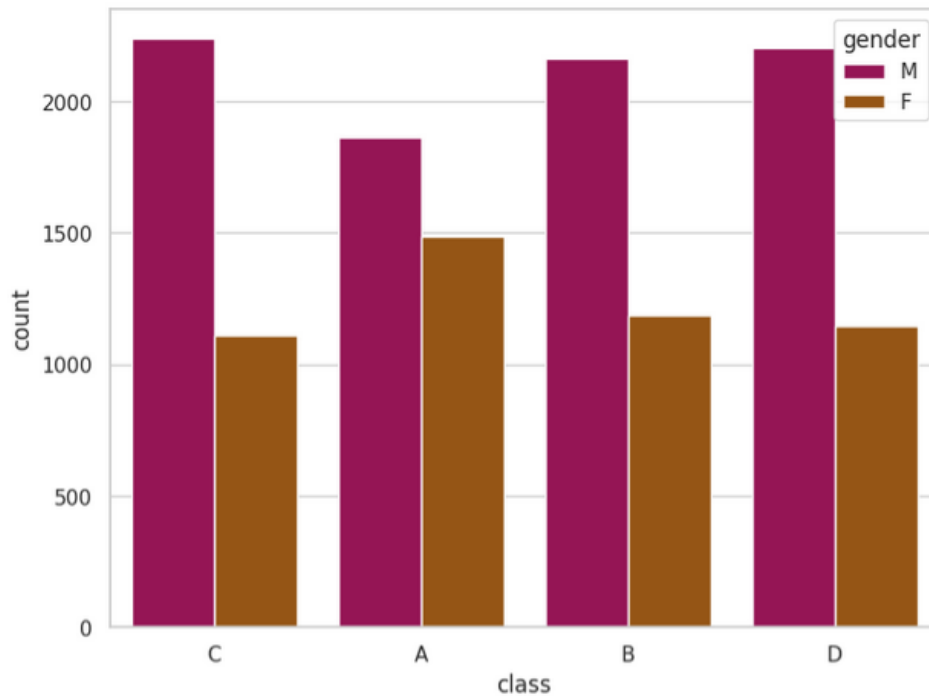


```
#Boxplot can plot outliers in data  
plt.figure(figsize = (20, 10))  
sns.boxplot(data = df)
```



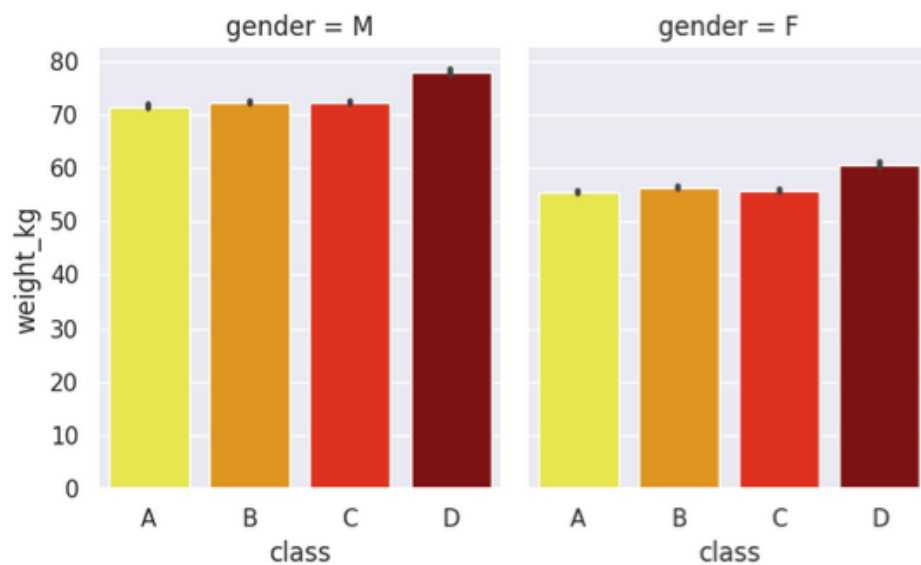
```
sns.set(rc = {'figure.figsize':(8,6)})
sns.set_style('whitegrid')
sns.countplot(x='class',hue='gender',data=df,palette='brg')
```

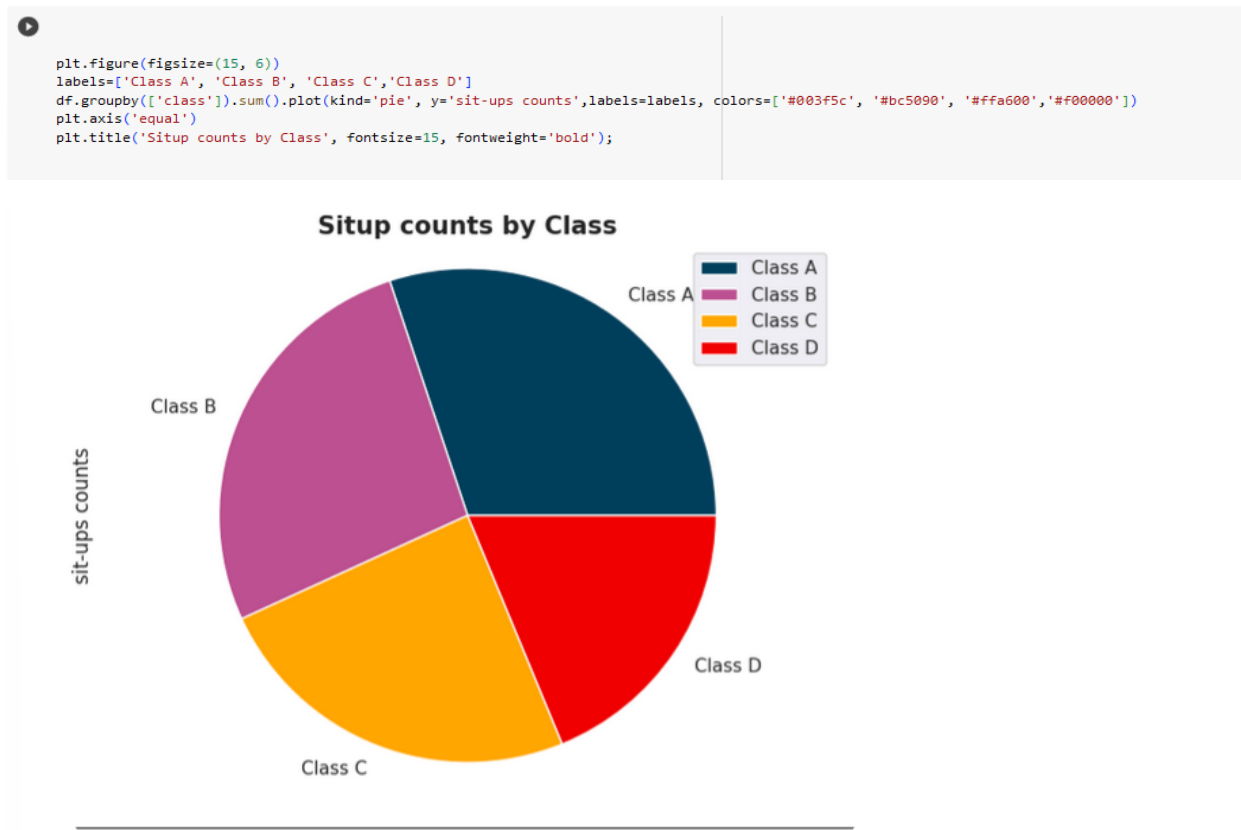
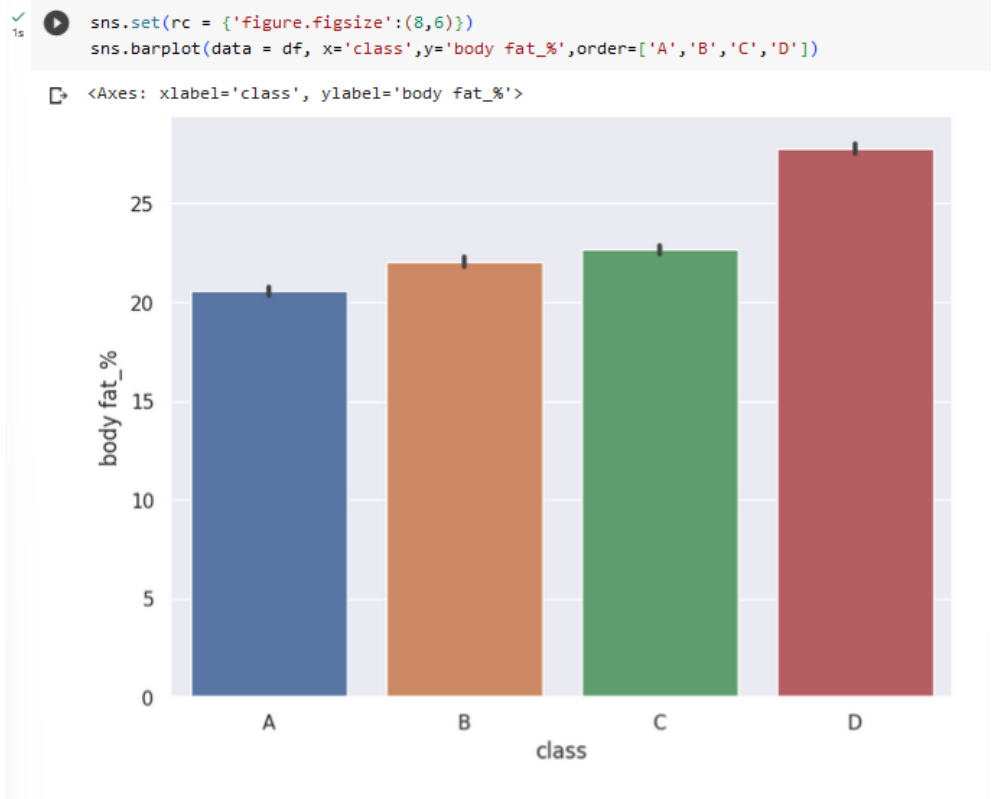
<Axes: xlabel='class', ylabel='count'>



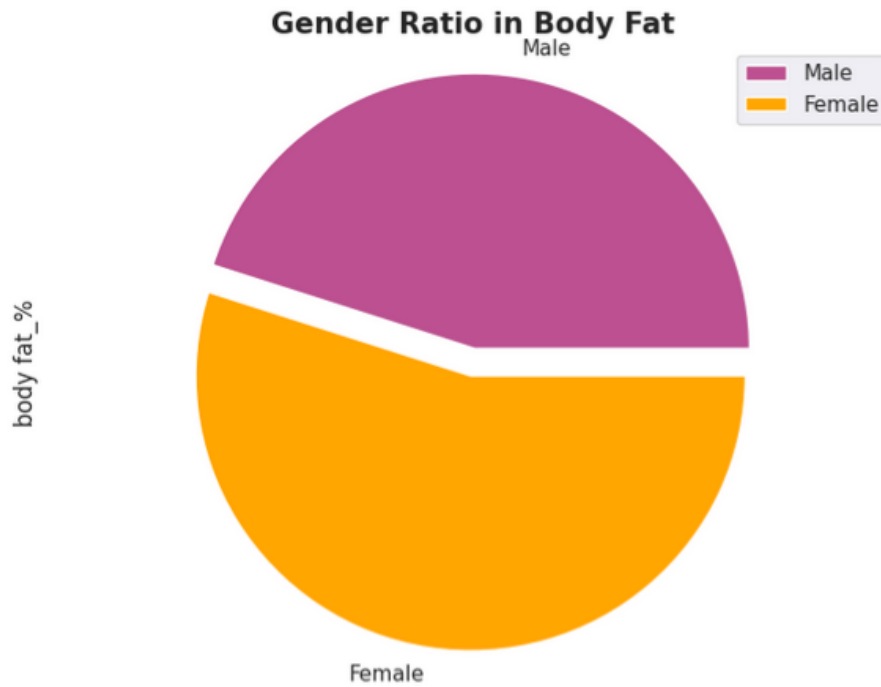
```
[66] sns.set(rc = {'figure.figsize':(8,6)})
g = sns.FacetGrid(df, col="gender", height=4, aspect=0.8 )
g.map(sns.barplot, "class", "weight_kg",order=['A','B','C','D'],palette='hot_r' )
```

<seaborn.axisgrid.FacetGrid at 0x7e57539aec50>





```
plt.figure(figsize=(15, 6))
labels=['Male','Female']
df.groupby(['gender']).sum().plot(kind='pie', y='body fat_%',labels=labels, colors=['#bc5090', '#ffa600'],explode=(0.0, 0.1))
plt.axis('equal')
plt.title('Gender Ratio in Body Fat', fontsize=15, fontweight='bold');
```



# **Test Statistics**

## **Test Statistics 1:**

**Predictor: height\_cm**

Hypotheses 1:

- Null Hypothesis (H0): There is no relationship between height and the sit-ups counts.
- Alternative Hypothesis (H1): There is a relationship between height and the sit-ups counts.

**Predictor: weight\_kg**

Hypotheses 2:

- Null Hypothesis (H0): There is no relationship between weight and the sit-ups counts.
- Alternative Hypothesis (H1): There is a relationship between weight and the sit-ups counts.

**Predictor: body fat\_%**


Hypotheses 3:

- Null Hypothesis (H0): There is no relationship between body fat percentage and the sit-ups counts.

- Alternative Hypothesis (H1): There is a relationship between body fat percentage and the sit-ups counts.

## Code :

```
✓ [26] # Multiple Linear Regression
      x = df[['height_cm', 'weight_kg', 'body fat_%']]
      y = df[['sit-ups counts']]
```

```
✓  from sklearn.model_selection import train_test_split
  x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 100)
```

```
✓ [28] #Implenting Linear Model
0s    from sklearn.linear_model import LinearRegression
      mlr = LinearRegression()
      mlr.fit(x_train, y_train)
```

```
▼ LinearRegression
LinearRegression()
```

```
✓ [29] y_pred = mlr.predict(x_test).round()
```

```
✓ [30] import statsmodels.api as sm
0s    model = sm.OLS(y, x).fit()

      model.summary()
```





### OLS Regression Results

<b>Dep. Variable:</b>	sit-ups counts	<b>R-squared (uncentered):</b>	0.935			
<b>Model:</b>	OLS	<b>Adj. R-squared (uncentered):</b>	0.935			
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	6.392e+04			
<b>Date:</b>	Tue, 08 Aug 2023	<b>Prob (F-statistic):</b>	0.00			
<b>Time:</b>	06:06:34	<b>Log-Likelihood:</b>	-50868.			
<b>No. Observations:</b>	13393	<b>AIC:</b>	1.017e+05			
<b>Df Residuals:</b>	13390	<b>BIC:</b>	1.018e+05			
<b>Df Model:</b>	3					
<b>Covariance Type:</b> nonrobust						
	<b>coef</b>	<b>std err</b>	<b>t</b>	<b>P&gt; t </b>	<b>[0.025</b>	<b>0.975]</b>
<b>height_cm</b>	0.3098	0.004	73.336	0.000	0.302	0.318
<b>weight_kg</b>	0.1530	0.010	15.971	0.000	0.134	0.172
<b>body fat_%</b>	-0.9808	0.012	-83.196	0.000	-1.004	-0.958
<b>Omnibus:</b>	39.882	<b>Durbin-Watson:</b>	2.019			
<b>Prob(Omnibus):</b>	0.000	<b>Jarque-Bera (JB):</b>	42.539			
<b>Skew:</b>	-0.105	<b>Prob(JB):</b>	5.79e-10			
<b>Kurtosis:</b>	3.180	<b>Cond. No.</b>	23.5			

#### Notes:

[1]  $R^2$  is computed without centering (uncentered) since the model does not contain a constant.

[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In summary, for all three predictor variables (height\_cm, weight\_kg, and body fat\_%), the p-values are very close to zero, indicating strong evidence that these variables are statistically significant in predicting the outcome variable. The coefficients provide information about the magnitude and direction of the relationship between each predictor and the outcome variable, while controlling for other variables.

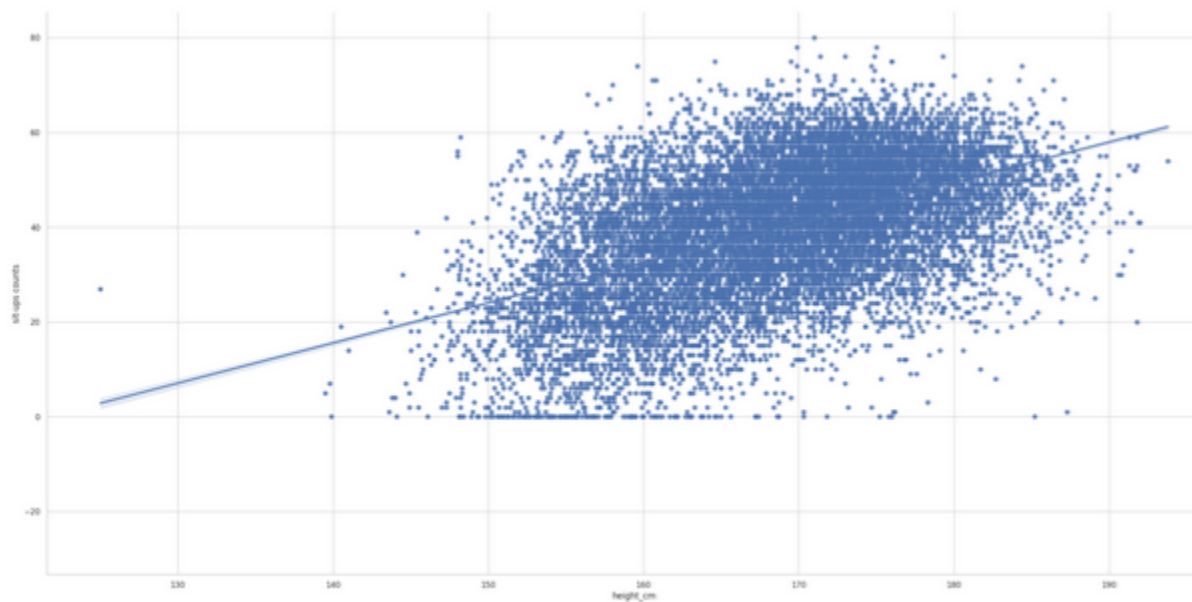
```

import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(12, 8))
# Create a pair plot for selected variables
sns.pairplot(
    df,
    x_vars=["height_cm", "weight_kg", "body fat_%"],
    y_vars=["sit-ups counts"],
    kind="reg", # Use regression line for scatter plots
    height=12, # Height of each subplot
    aspect=2,  # Height of each subplot
)

plt.show()

```

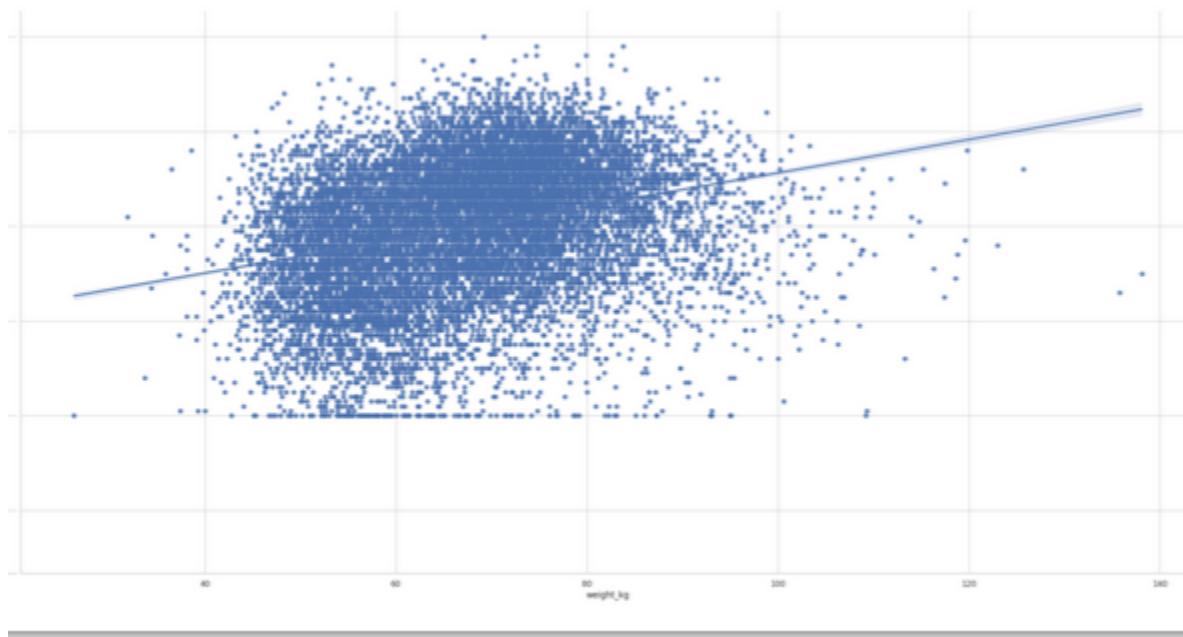


### Interpretation 1:

- The p-value for height\_cm is very close to zero ( $p < 0.05$ ), which is typically the significance level used for hypothesis testing.
- Since the p-value is less than the significance level, you would reject the null hypothesis.

From the p-value and the above graph, we can conclude that there is a positive correlation between the height and the sit-ups counts. So, as the body height increases the sit-up counts increases.

- Interpretation of Coefficient: For a one-unit increase in height (in centimeters), the expected outcome variable (e.g., sit-ups count) increases by approximately 0.3098, while holding other variables constant.

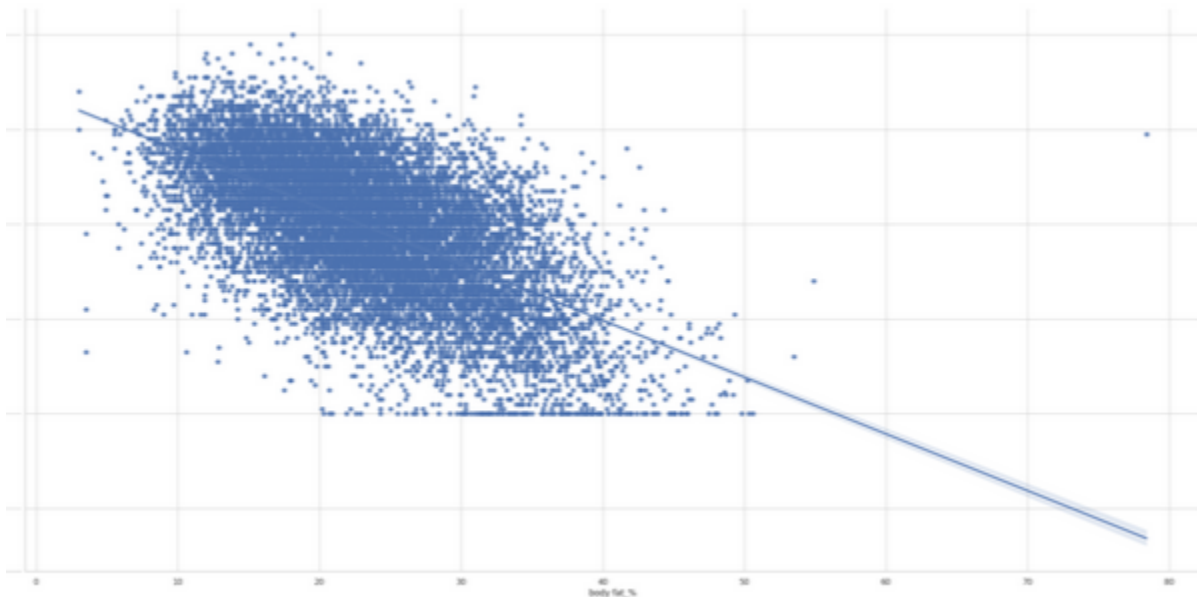


### Interpretation 2:

- The p-value for weight\_kg is very close to zero ( $p < 0.05$ ), indicating that the relationship between weight and the outcome variable is statistically significant.
- Given the low p-value, you would reject the null hypothesis.

From the p-value and the above graph, we can conclude that there is a positive correlation between the weight and the sit-ups counts. So, as the weight increases the sit-up counts increases.

- Interpretation of Coefficient: For a one-unit increase in weight (in kilograms), the expected outcome variable increases by approximately 0.1530, while holding other variables constant.



### Interpretation 3:

- The p-value for body fat\_% is very close to zero ( $p < 0.05$ ), suggesting that the relationship between body fat percentage and the outcome variable is statistically significant.
- Since the p-value is below the significance level, you would reject the null hypothesis.

From the p-value and the above graph, we can conclude that there is a negative correlation between the body fat and the sit-ups counts. So, as the body fat increases the sit-up counts decreases.

- Interpretation of Coefficient: For a one-unit increase in body fat percentage, the expected outcome variable decreases by approximately 0.9808, while holding other variables constant.

### **Test Statistics 2:**

#### **Predictor: height\_cm**

Hypotheses 1:

- Null Hypothesis (H0): There is no relationship between height and the broad jump\_cm.
- Alternative Hypothesis (H1): There is a relationship between height and the broad jump\_cm

#### **Predictor: weight\_kg**

Hypotheses 2:

- Null Hypothesis (H0): There is no relationship between weight and the broad jump\_cm.
- Alternative Hypothesis (H1): There is a relationship between weight and the broad jump\_cm

#### **Predictor: body fat\_%**

### Hypotheses 3:

- Null Hypothesis (H0): There is no relationship between body fat percentage and broad jump\_cm
- Alternative Hypothesis (H1): There is a relationship between body fat percentage and broad jump\_cm

```
✓ [48] x = df[['height_cm', 'weight_kg', 'body fat%']]  
0s      y = df[['broad jump_cm']]
```

```
✓ [50] mlr = LinearRegression()  
0s      mlr.fit(x_train, y_train)
```

```
▼ LinearRegression  
LinearRegression()
```

```
✓ [51] y_pred = mlr.predict(x_test).round()  
0s
```

```
import statsmodels.api as sm
model = sm.OLS(y, x).fit()

model.summary()
```

OLS Regression Results

Dep. Variable: broad jump\_cm      R-squared (uncentered): 0.985  
Model: OLS      Adj. R-squared (uncentered): 0.985  
Method: Least Squares      F-statistic: 2.853e+05  
Date: Tue, 08 Aug 2023      Prob (F-statistic): 0.00  
Time: 07:09:33      Log-Likelihood: -61629.  
No. Observations: 13393      AIC: 1.233e+05  
Df Residuals: 13390      BIC: 1.233e+05  
Df Model: 3  
Covariance Type: nonrobust

	coef	std err	t	P> t	[0.025	0.975]
height_cm	1.1678	0.009	123.771	0.000	1.149	1.186
weight_kg	0.8809	0.021	41.175	0.000	0.839	0.923
body fat_%	-2.8485	0.026	-108.186	0.000	-2.900	-2.797

Omnibus: 1802.079      Durbin-Watson: 2.012  
Prob(Omnibus): 0.000      Jarque-Bera (JB): 6430.891  
Skew: -0.661      Prob(JB): 0.00  
Kurtosis: 6.127      Cond. No. 23.5

Notes:  
[1] R<sup>2</sup> is computed without centering (uncentered) since the model does not contain a constant.  
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In summary, all three predictor variables (height\_cm, weight\_kg, and body fat\_%) have p-values very close to zero, indicating strong evidence that these variables are statistically significant in predicting the outcome variable. The coefficients provide information about the magnitude and direction of the relationship between each predictor and the outcome variable, while controlling for other variables.

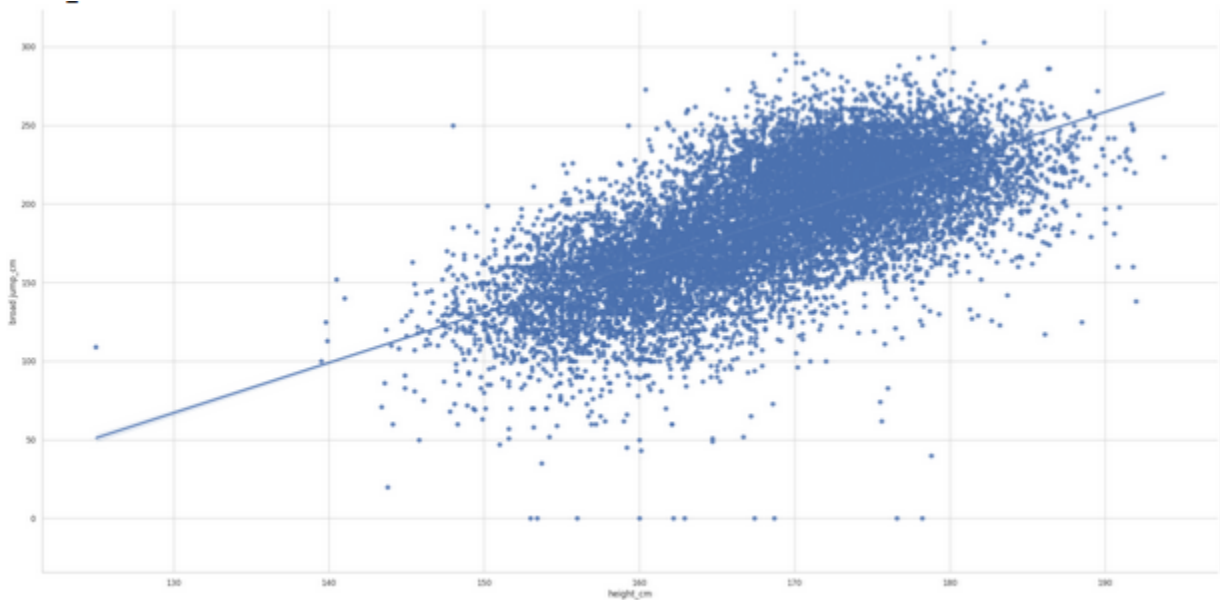
## Code :

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(12, 8))
# Create a pair plot for selected variables
sns.pairplot(
    df,
    x_vars=["height_cm", "weight_kg", "body fat%"],
    y_vars=["broad jump_cm"],
    kind="reg", # Use regression line for scatter plots
    height=12, # Height of each subplot
    aspect=2, # Height of each subplot
)

plt.show()
```

## Interpretation 1:



- The p-value for height\_cm is very close to zero ( $p < 0.05$ ), which is typically the significance level used for hypothesis testing.

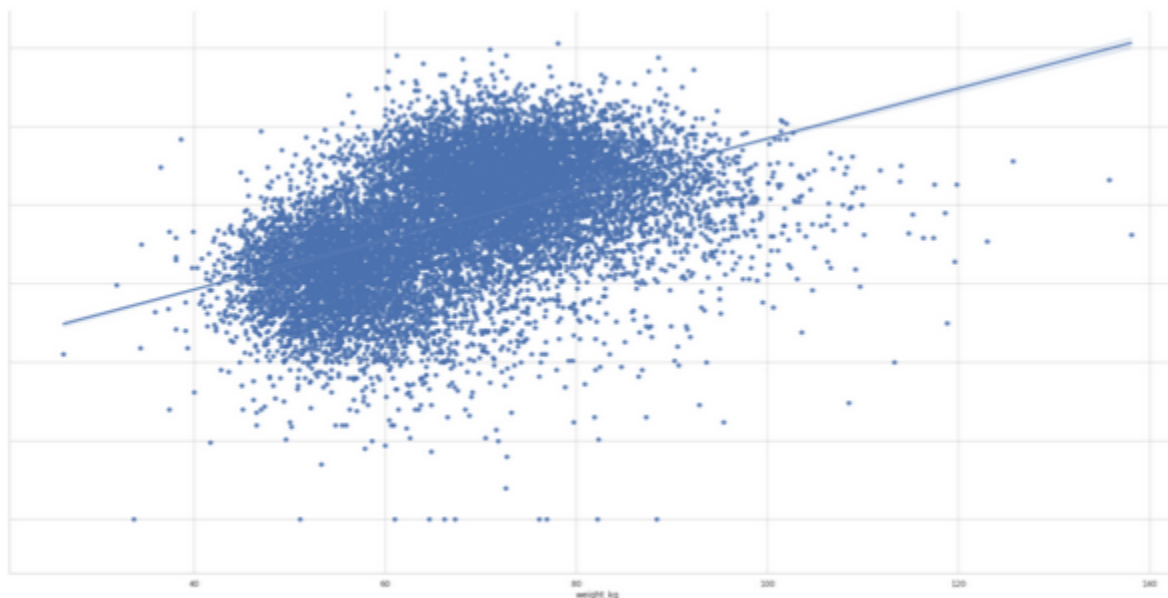


- Since the p-value is less than the significance level, you would reject the null hypothesis.

From the p-value and the above graph, we can conclude that there is a positive correlation between the height and the broad jump. So, as the height increases the long jump count also increases.

- Interpretation of Coefficient: For every one-unit increase in height (in centimeters), the expected outcome variable (e.g., sit-ups counts) increases by approximately 1.1678, while holding other variables constant.

Interpretation 2:

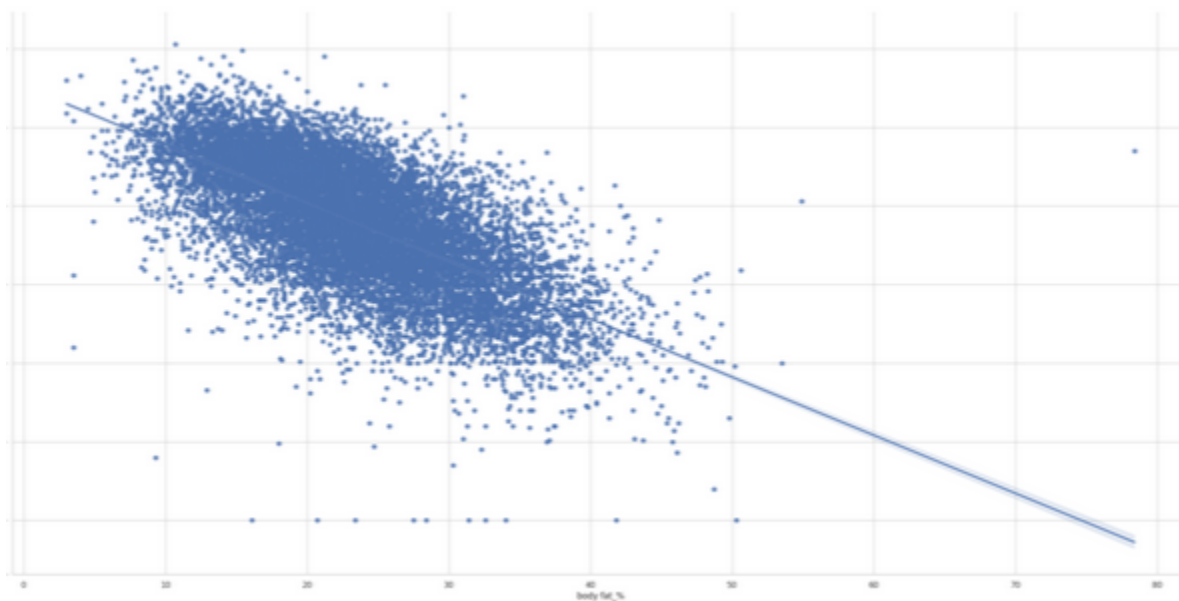


- The p-value for weight\_kg is very close to zero ( $p < 0.05$ ), indicating that the relationship between weight and the outcome variable is statistically significant.
- Given the low p-value, you would reject the null hypothesis.

From the p-value and the above graph, we can conclude that there is a positive correlation between the weight and the broad jump. So, as the weight increases the long jump count also increases.

- Interpretation of Coefficient: For every one-unit increase in weight (in kilograms), the expected outcome variable increases by approximately 0.8809, while holding other variables constant.

Interpretation 3:



- The p-value for body fat\_% is very close to zero ( $p < 0.05$ ), suggesting that the relationship between body fat percentage and the outcome variable is statistically significant.
- Since the p-value is below the significance level, you would reject the null hypothesis.

From the p-value and the above graph, we can conclude that there is a negative correlation between body fat and broad jump. So, as the body fat increases the long jump count also increases.

- Interpretation of Coefficient: For every one-unit increase in body fat percentage, the expected outcome variable decreases by approximately 2.8485, while holding other variables constant.

## Conclusion

The regression analysis conducted on the dataset has yielded valuable insights into the relationships between the predictor variables (height\_cm, weight\_kg, and body fat\_%) and the outcome variable (e.g., sit-ups counts). The analysis has shown that each of these predictor variables has a statistically significant impact on the outcome variable, as indicated by very low p-values for each predictor.

The findings suggest that higher values of height\_cm are associated with higher values of the outcome variable (sit-ups counts), even when accounting for other variables. Similarly, increased weight\_kg is linked to higher values of the outcome variable, indicating that weight plays a role in exercise performance. On the other hand, higher levels of body fat\_% are associated with lower values of the outcome variable, highlighting the potential negative impact of body fat percentage on exercise capability.

Also, The results indicate that individuals with greater height\_cm values tend to achieve longer broad jump\_cm distances, accounting for other variables. Similarly, higher weight\_kg values are linked to longer broad jump\_cm distances, suggesting a role for weight in jump performance. Conversely, elevated body fat\_% values are associated with shorter broad jump\_cm distances, revealing the potential negative influence of body fat percentage on jumping capability.

This regression model provides evidence that physical attributes such as height, weight, and body fat percentage play significant roles in influencing exercise performance, specifically sit-ups counts and broad\_jump\_cm. The model's coefficients provide quantified insights into the extent of these relationships, allowing for a better understanding of how changes in these variables may affect exercise outcomes.