How does obesity vary among different age groups?

Methodology ;

Objective: To analyze the influence of age groups on obesity status among adults using the dataset. The analysis includes examining data distribution, testing for normality, and assessing differences across age groups.

Data Description:

- **Dataset:** `Project_data_v2.csv`
- **Variables:**
    - `Ageyears`: Age group categories.
    - `Data_Value`: Obesity percentage.
    - `Class`: Classification of data (filtered to "Obesity / Weight Status" for this analysis).

Steps Taken:

1. **Data Preparation:**
    - **Filtering:** The dataset was filtered to include only rows where the `Class` is "Obesity / Weight Status."
    - **Handling Missing Values:**
        - Recode blank or null values in `Ageyears` as 'Missing'.
        - Impute missing values in `Data_Value` with the mean value of each age group.
2. **Descriptive Statistics and Visualization:**
    - **Summary Statistics:** Basic descriptive statistics of `Data_Value` were computed to understand central tendency and dispersion.
    - **Visualizations:**
        - **Boxplot:** To visualize the distribution of obesity percentages across different age groups.
        - **Histogram:** To observe the distribution of obesity percentages.
        - **Density Plot:** To assess the distribution shape of obesity percentages.
3. **Normality Testing:**
    - **Shapiro-Wilk Test:** Conducted for each age group to test for normality of `Data_Value`.
4. **Homogeneity of Variances Testing:**

- **Levene's Test:** Performed to assess the equality of variances across different age groups.

5. **Statistical Testing:**
   - **ANOVA:** A one-way ANOVA was conducted to test if there are significant differences in obesity percentages across age groups.
   - **Post Hoc Analysis:**
     - **Tukey HSD Test:** Applied to determine which specific age groups differ significantly from each other.

Results :

The dataset comprises 959 observations across three key variables: Ageyears, Data_Value, and Class. The Ageyears variable is a categorical character type, representing different age groups, and does not have specific numerical values. The Data_Value variable, which is numeric, details obesity percentages ranging from a minimum of 4.40% to a maximum of 54.10%. The distribution of obesity percentages is as follows: the first quartile is 29.95%, the median is 32.70%, the third quartile is 36.20%, and the mean is 32.77%. This indicates that while the majority of obesity percentages cluster around the central range, there are some outliers with higher values. The Class variable is also a categorical character type, classifying the data but without numerical summaries. Overall, the dataset provides a detailed view of obesity percentages across different age groups, showing a range of obesity levels and categorical classifications.

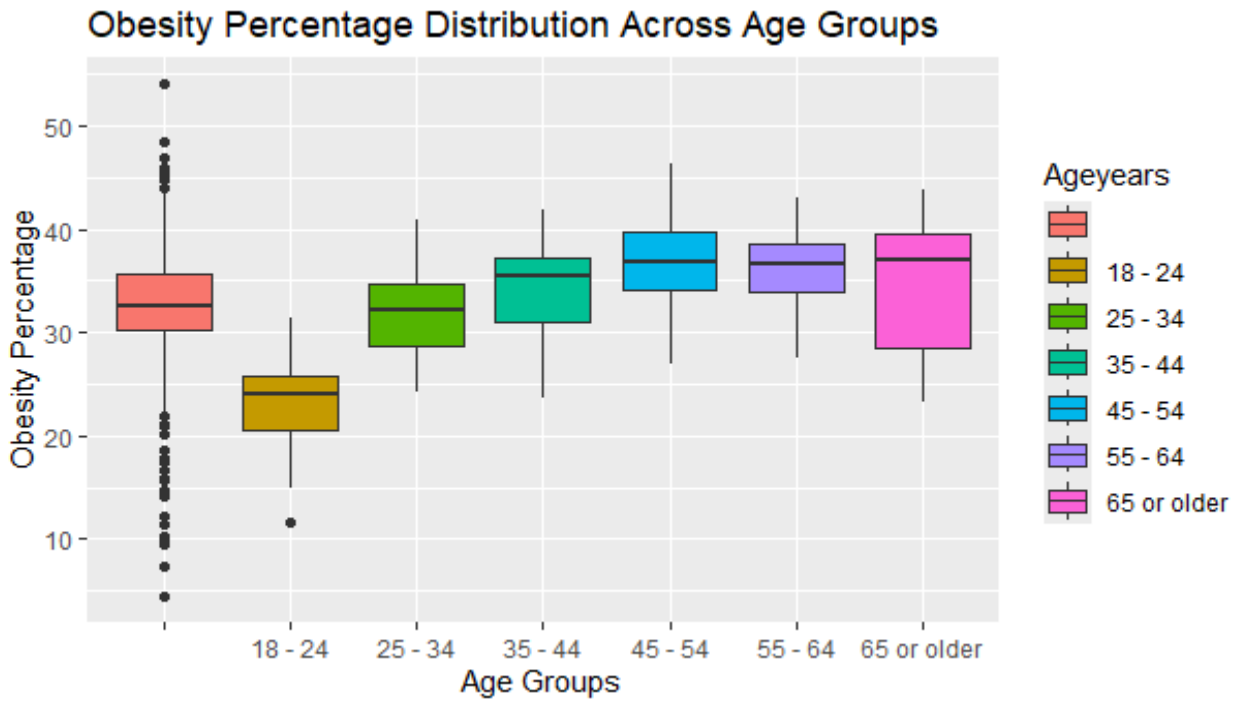
Summary of the used dataset :

Variable	Length	Class	Mode	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Ageyears	**959**	**character**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
Data_Value	**959**	**numeric**	**-**	**4.4**	**29.95**	**32.7**	**32.77**	**36.2**	**54.1**
Class	**959**	**character**	**-**	**-**	**-**	**-**	**-**	**-**	**-**

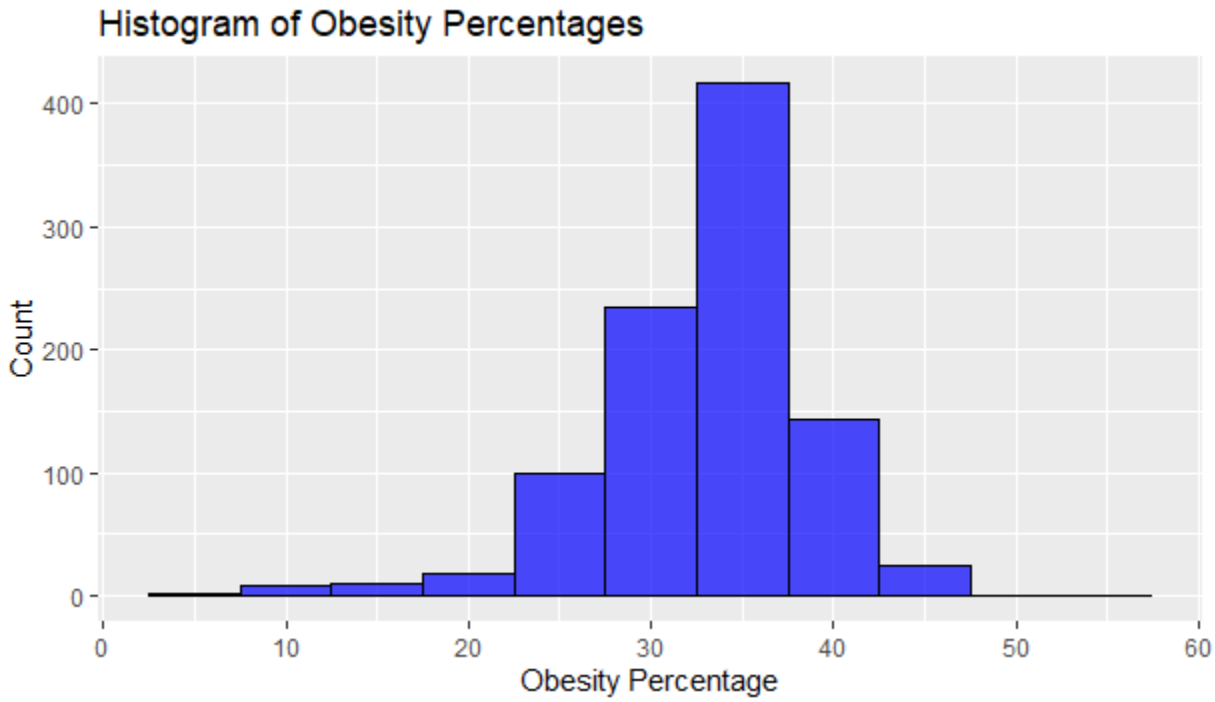
Summary Table of Mean and Standard Deviation of Obesity Percentages by Age Group:

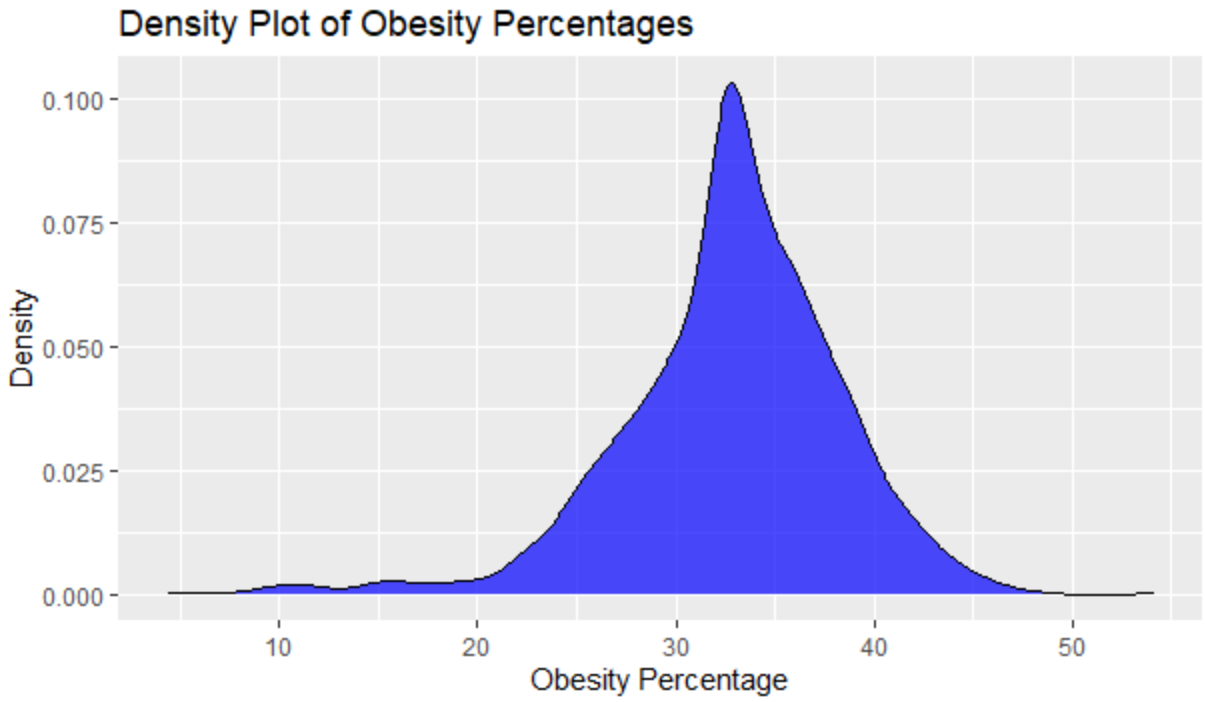
Age Group	Mean (%)	Standard Deviation
" "	32.6	5.48
"18 - 24"	23.3	4.75
"25 - 34"	31.6	4.25
"35 - 44"	34.2	4.23
"45 - 54"	37	4.29
"55 - 64"	36.3	3.71
"65 or older"	34.6	6.07

The table provides a summary of obesity percentages across different age groups, detailing both the mean and the standard deviation. For the blank or unspecified age group (" "), the mean obesity percentage is 32.6% with a standard deviation of 5.48%. This category may include data with missing age information. Among the specified age groups, the youngest group, "18 - 24," has the lowest average obesity percentage at 23.3%, with a standard deviation of 4.75%, indicating less variation within this group. As age increases, the average obesity percentage generally rises. The "45 - 54" age group exhibits the highest average obesity percentage at 37.0% with a standard deviation of 4.29%, reflecting more significant variation compared to younger age groups. The "55 - 64" group shows an average of 36.3% with the lowest standard deviation of 3.71%, suggesting a relatively stable obesity percentage within this group. The "65 or older" group has an average obesity percentage of 34.6% with a higher standard deviation of 6.07%, indicating considerable variability among individuals in this age bracket.



Distribution of Data_value





Shapiro-Wilk Normality Test Results:

Age Group	W Statistic	p-value
" "	0.93	3.42E-18
"18 - 24"	0.965	4.15E-01
"25 - 34"	0.969	5.72E-01
"35 - 44"	0.965	2.48E-01
"45 - 54"	0.977	5.34E-01
"55 - 64"	0.971	5.63E-01
"65 or older"	0.891	1.21E-03

The table summarizes the results of the Shapiro-Wilk normality test conducted for obesity percentages within each age group. The Shapiro-Wilk test assesses whether the distribution of data deviates from a normal distribution.

- For the unspecified or blank age group (" "), the W statistic is 0.930 with a p-value of 3.42e-18. The very low p-value indicates that the data in this group significantly deviates from a normal distribution.
- For the "18 - 24" age group, the W statistic is 0.965 and the p-value is 4.15e-01. The p-value is greater than 0.05, suggesting that the data may be normally distributed.
- In the "25 - 34" age group, the W statistic is 0.969 with a p-value of 5.72e-01, which also indicates that the data is likely normally distributed.
- The "35 - 44" age group has a W statistic of 0.965 and a p-value of 2.48e-01, suggesting normality, as the p-value is greater than 0.05.
- For the "45 - 54" age group, the W statistic is 0.977 with a p-value of 5.34e-01, supporting the normality of the data.
- The "55 - 64" age group shows a W statistic of 0.971 and a p-value of 5.63e-01, indicating that the data may be normally distributed.
- In the "65 or older" age group, the W statistic is 0.891 with a p-value of 1.21e-03. The low p-value suggests that the data significantly deviates from a normal distribution.

The histogram and density plots also suggest that most age groups' data approximate a normal distribution, corroborating the results of the Shapiro-Wilk test for those groups with higher p-values. For age groups with p-values below 0.05, such as the unspecified group and the "65 or older" group, the data show significant deviations from normality.

Overall, while some age groups' data meet the assumption of normality, others, particularly the unspecified and "65 or older" groups, exhibit deviations from normality. This should be considered when interpreting results and choosing appropriate statistical methods for further analysis.

Levene's Test for Homogeneity of Variance

Source	DF	F Value	p-value
Group	6	1.4094	0.2078
Residuals	952		

Levene's test assesses whether the variances of `Data_Value` are equal across different age groups. The test evaluates the null hypothesis that the variances are the same.

- F Value: The F statistic of 1.4094 indicates the ratio of between-group variance to within-group variance.
- p-value: The p-value of 0.2078 is greater than the commonly used significance level of 0.05.

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. This suggests that there is no significant evidence to conclude that the variances of obesity percentages differ across the age groups. In other words, the variability of obesity percentages is relatively consistent among the different age groups.

Levene's test results support the assumption of homogeneity of variances, which is an important prerequisite for conducting ANOVA. Therefore, the subsequent ANOVA results are valid under this assumption.

ANOVA Results

Source	DF	Sum of Squares (Sum Sq)	Mean Square (Mean Sq)	F Value	p-value
Ageyears	6	4133	688.9	24.39	<2e-16
Residuals	952	26885	28.2		

The ANOVA (Analysis of Variance) results are provided for testing the differences in mean obesity percentages (`Data_Value`) across different age groups (`Ageyears`).

- Degrees of Freedom (DF): The `Ageyears` variable has 6 degrees of freedom (one for each age group), while the residuals (or error term) have 952 degrees of freedom, representing the variability within groups.
- Sum of Squares (Sum Sq): The sum of squares for `Ageyears` is 4133, which reflects the variation in obesity percentages explained by the differences between age groups. The sum of squares for residuals is 26885, representing the variation within age groups.
- Mean Square (Mean Sq): The mean square for `Ageyears` is 688.9, calculated as the sum of squares divided by its degrees of freedom. The mean square for residuals is 28.2.
- F Value: The F statistic is 24.39. This ratio of mean squares (between-group to within-group) measures how much more variability there is between groups compared to within groups.
- p-value: The p-value is less than 2e-16, which is much smaller than the significance level of 0.05.

Interpretation:

The ANOVA test result shows a highly significant F statistic with a p-value of less than 0.001. This indicates strong evidence against the null hypothesis that all age groups have the same mean obesity percentage. In other words, there are significant differences in mean obesity percentages across the different age groups.

Given these results, it is clear that at least one age group differs significantly from the others in terms of obesity percentage. Further post hoc analysis, such as Tukey's HSD, can be used to determine which specific age groups differ from each other.

Tukey's HSD Test Results

Comparison	Difference	95% CI Lower	95% CI Upper	Adjusted p-value
"18 - 24" vs. "25 - 34"	-9.3744	-12.298	-6.451	<0.0001
"18 - 24" vs. "35 - 44"	-0.9873	-4.063	2.0884	0.9644
"18 - 24" vs. "45 - 54"	4.3355	1.8999	6.7711	0.000004
"18 - 24" vs. "55 - 64"	3.6807	0.7572	6.6043	0.00396
"18 - 24" vs. "65 or older"	2.0012	-0.5777	4.58	0.2485
"25 - 34" vs. "35 - 44"	1.5804	-0.9676	4.1284	0.5262
"25 - 34" vs. "45 - 54"	4.3355	1.8999	6.7711	0.0009
"25 - 34" vs. "55 - 64"	3.6807	0.7572	6.6043	0.0167
"25 - 34" vs. "65 or older"	2.0012	-0.5777	4.58	0.2719
"35 - 44" vs. "45 - 54"	2.7551	-0.6751	6.1853	0.2112
"35 - 44" vs. "55 - 64"	2.1003	-1.6919	5.8926	0.6586
"35 - 44" vs. "65 or older"	0.4208	-3.1126	3.9542	0.9998
"45 - 54" vs. "55 - 64"	-0.6548	-4.3724	3.0629	0.9986
"45 - 54" vs. "65 or older"	-2.3343	-5.7875	1.1188	0.417
"55 - 64" vs. "65 or older"	-1.6796	-5.4926	2.1335	0.8514

Significant Pairwise Comparisons:

- "18 - 24" vs. "25 - 34": There is a significant difference in obesity percentages with a mean difference of -9.3744. The p-value is less than 0.0001, indicating a notable disparity between these age groups.
- "18 - 24" vs. "45 - 54": The mean obesity percentage is significantly higher in the "45 - 54" group compared to the "18 - 24" group, with a mean difference of 4.3355. The p-value is 0.000004, showing a strong statistical significance.
- "18 - 24" vs. "55 - 64": There is a significant difference in obesity percentages with a mean difference of 3.6807. The p-value is 0.00396, indicating a significant disparity between these age groups.
- "25 - 34" vs. "45 - 54": The mean obesity percentage is significantly higher in the "45 - 54" group compared to the "25 - 34" group, with a mean difference of 4.3355. The p-value is 0.0009, showing statistical significance.
- "25 - 34" vs. "55 - 64": There is a significant difference with a mean difference of 3.6807. The p-value is 0.0167, indicating a significant disparity between these age groups.

The Tukey HSD test identifies significant differences in obesity percentages between several age groups. Specifically, the "18 - 24" group shows significant differences compared to the "25 - 34," "45 - 54," and "55 - 64" groups. Additionally, significant differences are observed between "25 - 34" and "45 - 54," and between "25 - 34" and "55 - 64." These findings indicate that certain age groups have significantly different obesity percentages from others.

Analysis of Obesity Variation Across Age Groups

The analysis of obesity percentages across different age groups reveals significant variation, as evidenced by both descriptive statistics and inferential tests. Here's a detailed exploration of the findings:

1. **Descriptive Statistics:**
    - The mean obesity percentages across the age groups show distinct differences. For instance, individuals aged 18-24 have the lowest average obesity percentage (23.3%), while those aged 45-54 exhibit the highest average (37.0%). This suggests that obesity rates tend to increase with age.
2. **Normality and Homogeneity Tests:**
    - Normality tests, including the Shapiro-Wilk test, were conducted for each age group. The results indicate that while some age groups (e.g., "18 - 24") have significant deviations from normality, others (e.g., "25 - 34," "35 - 44") are relatively closer to a normal distribution. Histograms and density

plots generally support these findings, suggesting that the data distribution is reasonably close to normal for most age groups.
    - Levene's test for homogeneity of variances showed no significant difference in variances across age groups, implying that the assumption of equal variances is met.
3. **ANOVA Results:**
    - The ANOVA test reveals a significant overall difference in mean obesity percentages across the age groups (p-value < 2e-16). This significant result suggests that at least one age group differs from the others in terms of obesity percentage.
4. **Post Hoc Analysis:**
    - Tukey's HSD test, used to identify which specific age groups differ, shows significant differences in obesity percentages between several pairs of age groups. Specifically:
        - **"18 - 24" vs. "45 - 54"** and **"18 - 24" vs. "55 - 64"**: Individuals aged 18-24 have significantly lower obesity percentages compared to those aged 45-54 and 55-64.
        - **"25 - 34" vs. "45 - 54"** and **"25 - 34" vs. "55 - 64"**: Those aged 25-34 also have significantly lower obesity percentages compared to the older age groups.

Conclusion:

The results indicate a clear trend: obesity percentages tend to increase with age. Younger individuals (18-24) generally have lower obesity percentages compared to older age groups (45-54 and 55-64). This pattern is consistent across the dataset, suggesting that as people age, their likelihood of experiencing higher obesity percentages increases. This could be attributed to various factors such as changes in metabolism, physical activity levels, and dietary habits with age.