BILAL EREN

Data Analysis Project on Age Difference in Different Disciplines of Olympic Winter Games

12.06.2022

TABLE OF CONTENTS

INTRODUCTION AND MOTIVATION

This report analyzes winter olympic games data and main focus is to understand if there is a significant age difference between snowboard and ski athletes. This problem is important because many winter sports equipment companies are interested in this problem. They want to know if there is a significant age difference between snowboard and ski athletes because they will shape their marketing strategies according to this age difference. The exact problem consists of different parts. Firstly, the age difference between snowboard and ski athletes is analyzed with descriptive statistics and visualizations such as box plot, histogram and KDE plot. Secondly, the age difference between female snowboard and ski athletes for each olympics (2014, 2018, 2022) are analyzed with descriptive statistics and visualizations such as box plot, histogram and KDE plot. Thirdly, the age difference between female snowboard and ski athletes for each olympics (2014, 2018, 2022) are analyzed with hypothesis testing to determine if there is a significant difference between the mean age of the groups. As a result, there is a significant age difference between the means of female snowboard and ski athletes in 2018 and 2022. On the other hand, there is no significant difference between the means of female snowboard and ski athletes' age in 2014.

DETAILED DESCRIPTION OF THE PROBLEM

Olympic winter games data consist of four variables. The first variable is dates of olympics which are 2014, 2018 and 2022, data type of it is integer. The second variable is discipline which consists of "Snowboard" and "Ski", data type of it is object. The third variable is genders of athletes with "f" letter which means female and "m" letter which means male, data type of it is object. The fourth and the most significant variable is age which consists of ages of athletes and data type of it is integer. Minimum age is 17 and maximum age is 45. There are 749 values (row) and there is no null value in the dataset. The research questions in the project are that is there a significant age difference between snowboarding and skiing athletes and is there a significant age difference between female snowboarding and female skiing athletes for all three olympics (2014, 2018 and 2022)? There are 3 main tasks: first task is descriptive statistical analysis to understand data which we are working on, second task is to illustrate the age difference between the research groups by box plots and histograms; third task is hypothesis testing. Hypothesis tests answer that there is or not a significant age difference between the means of research groups.

METHODS

The Jupyter Notebook environment in the Anaconda Navigator platform is used as a research environment. Also Python programming language is used to apply all methods in the project.

Methods Used in Descriptive Statistic Analysis

Arithmetic mean, standard deviation, variance, min value, max value, median.

Plots

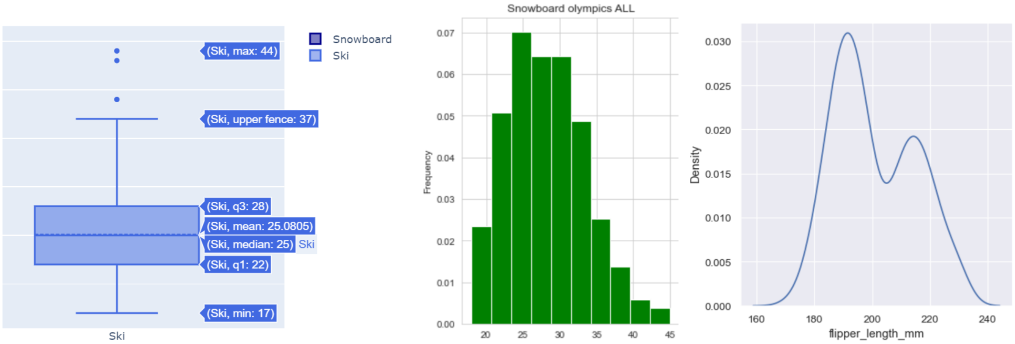

Figure 1: Box Plot, Histogram Plot and KDE Plot Example (Waskom, 2021)

First plot in figure 1 is a box plot which shows the concentration of the data. A box plot includes the min value, the first quartile, the median, the third quartile and the max value. These values are useful to compare two different data (OpenStaxCollege 2013). Second plot is a histogram plot which consists of adjacent boxes, the horizontal axis indicates the data represented and the vertical axis shows frequency (Rice University, 2022). "The histogram illustrates the shape of the data, the center, and the spread of the data" (Rice University, 2022). Third plot is "the Kernel Density Estimate plot which smooths the observations with a Gaussian kernel, producing a continuous density estimate" (Waskom, 2021).

Central Limit Theorem (CLT)

"If X_1 , X_2 , …X_n is a random sample from an infinite population with the mean μ, the variance σ^2 , and the moment-generating function $M_X(t)$, then the limiting distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

" (Miller & Miller, 2015, p. 236). "As n → ∞ is the standard normal distribution" (Miller & Miller, 2015, p. 236). Briefly, the CLT states that with a sufficiently large sample size, “the mean of sampling distribution approximately follows a normal distribution regardless of the distribution of population” (Miller & Miller, 2015, p. 237). In application, sufficiently large sample size is n ≥ 30 (Miller & Miller, 2015, p. 237).

Hypothesis Testing:

1. F-Test of Equality of Variances

F-test of equality of two variances requires two assumptions: first, the normality of the two samples’ population distribution second, independence of two populations. “We have two random samples from two independent normal populations and the unknown population variances are $\sigma_1^2$and σ_2^2; sample variances are $s_1^2$and s_2^2; sample sizes are n1 and *n*2” (OSCRiceUniversity, 2015). “*F* ratio for comparison of two sample variances: $F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}}$ *F* has the distribution $F \sim F(n1 - 1, n2 - 1)$, where $n1 - 1$ are the degrees of freedom for the numerator and $n2 - 1$ are the degrees of freedom for the denominator” (OSCRiceUniversity, 2015). When the H_0: $\sigma_1^2 = \sigma_2^2$, then the $F_c = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} = \frac{s_1^2}{s_2^2}$. There are different forms of hypothesis tests. **Two -tail test**: $H_0 : \sigma_1^2 = \sigma_2^2$ and H_1: $\sigma_1^2 \neq \sigma_2^2$. **One-tail test**: $H_0 : \sigma_1^2 \leq \sigma_2^2$ and H_1: $\sigma_1^2 > \sigma_2^2$ or $H_0 : \sigma_1^2 \geq \sigma_2^2$ and H_1: $\sigma_1^2 < \sigma_2^2$. Upper tail critical value: $F_{\alpha, df1, df2}$ and lower tail critical value: $1/ F_{\alpha, df2, df1}$. **α** is level of significance and df is degrees of freedom. (OSCRiceUniversity, 2015). “P value is the lowest level of significance at which the null hypothesis could have been rejected” (Miller & Miller, 2015, p. 362). If p value is less than or equal to **α**, reject null hypothesis. Also, other way is that if test statistic (F value) is between the critical values, we cannot reject H_0.

2. Student’s Two-sample T Test

Student’s two-sample t test is used to test differences between the means. “Two-sample t test can be used for independent random samples from two normal populations having the same unknown variance σ^2 , the likelihood ratio technique yields a test based on $t = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ ” (Miller & Miller, 2015, p. 367). H_0: $\mu_1 - \mu_2 = 0$, H_0: $\mu_1 - \mu_2 \leq 0$ or H_0: $\mu_1 - \mu_2 \geq 0$ and H_A: $\mu_1 - \mu_2 \neq 0$, $\mu_1 - \mu_2 > 0$ or $\mu_1 - \mu_2 < 0$. Critical regions of size alpha(α) are respectively, $|t| \geq t_{\frac{\alpha}{2}, n1 + n2 - 2}$, $t \geq t_{\alpha, n1 + n2 - 2}$, $t \leq -t_{\alpha, n1 + n2 - 2}$ according to hypotheses (Miller & Miller, 2015, p. 367). Evaluation of two sample t test is the same as the f-test. If p value is less than or equal to **α**, reject null hypothesis.

3. Welch’s test

Welch’s test is used when Student’s t test’s same variance assumption is failed. Calculation of the t-statistic used in the Welch test is similar to the Student test. “The difference between the sample means divided by some estimate of the standard error of that difference: $t = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$” (Navarro, 2020). “The standard error calculations are different; If the two populations have different standard deviations, then it's nonsense to try to calculate a pooled standard deviation estimate but you can still estimate the standard error of the difference between sample means: $\mathrm{SE}(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$ ” (Navarro, 2020). The other difference between Welch and Student is calculation of df. Calculation of df in Welch test: $\mathrm{df} = \frac{\left(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2\right)^2}{\left(\hat{\sigma}_1^2/N_1\right)^2/(N_1 - 1) + \left(\hat{\sigma}_2^2/N_2\right)^2/(N_2 - 1)}$ (Navarro, 2020). Hypotheses and evaluation of the test are the same as Student’s t test.

EVALUATION

Descriptive Analysis

Age of athletes data has 749 non-null values. The key statistics of athletes' age are mean: 25.883, standard deviation: 4.902 , min: 17 and max value: 45. There are 190 snowboard athletes and 559 ski athletes.

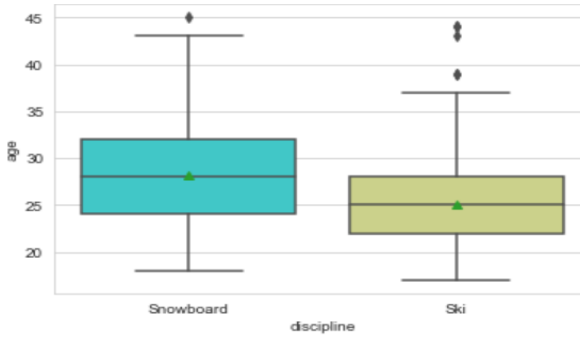

Figure 4: box plot, age difference between snowboarding and ski athletes

Boxplots in the figure 4 indicate the age difference between snowboard and ski athletes. Snowboard athletes' age has these key statistics: mean: 28.24, min: 18, median: 28, max: 45. On the other hand, ski athletes' age has a mean :25.08, min: 17, median: 25, max: 44. Snowboard athletes seem older than ski athletes.

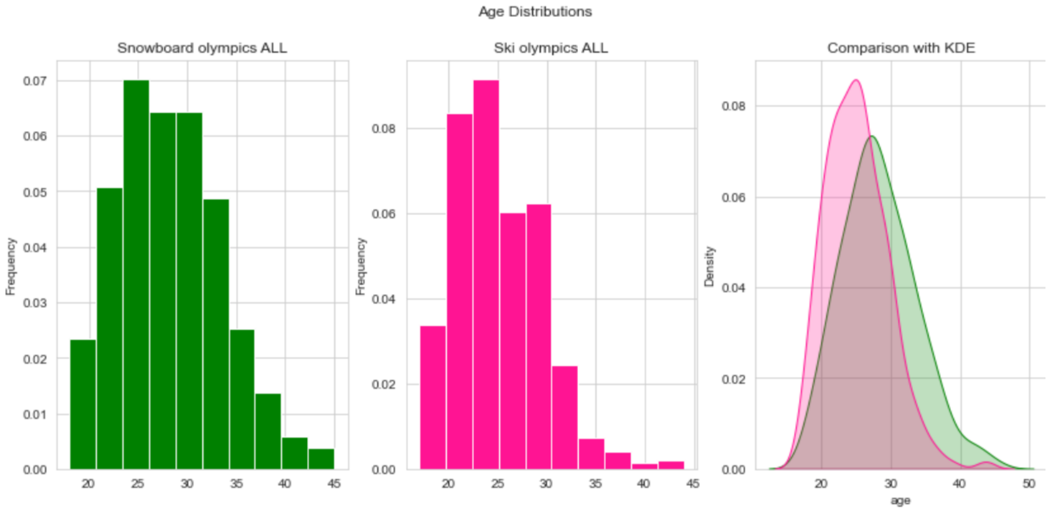

Figure 5: Histograms and KDE plots of snowboard vs ski athletes' age

Histogram and KDE plots in the figure 5 indicate the shape and spread of the data, it is seen that age of ski athletes is slightly concentrated on the left compared to age of snowboard athletes, which means ski athletes younger than snowboard athletes.

There were 121 female athletes in the 2014 Olympics. The key statistics of athletes' age are mean: 24.917, standard deviation: 4.36, min:18 and max: 41.

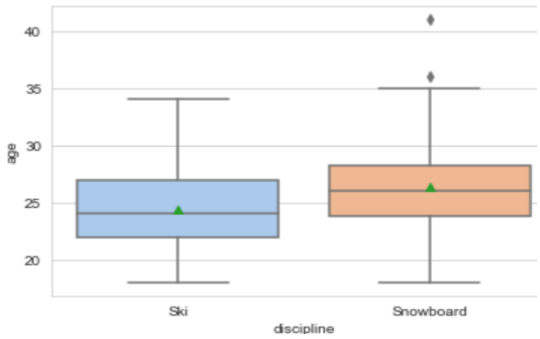

Figure 6: box plot, age difference between female snowboarding and ski athletes in 2014

There were 32 female snowboard athletes and 89 female ski athletes in 2014. Female snowboard athletes' age has these key statistics: mean: 26.38, min: 18, median: 26, max: 41 in 2014. On the other hand, female ski athletes' age has a mean :24.39, min: 18, median: 24, max: 34 in 2014. Female snowboard athletes seem older than female ski athletes in 2014.

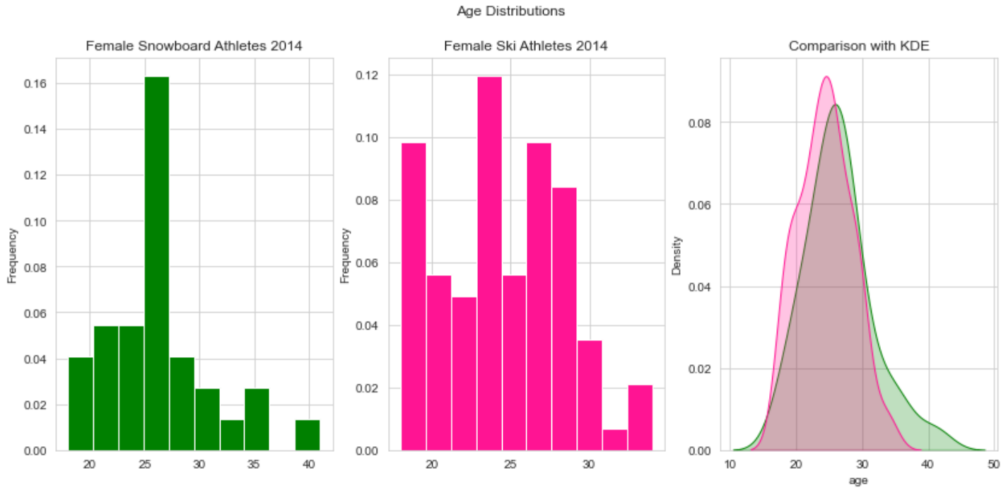

Figure 7: Histograms and KDE plots of female snowboard vs ski athletes' age 2014

According to figure 7, it is seen that the age of female ski athletes is slightly concentrated on the left compared to the age of female snowboard athletes, which means female ski athletes younger than female snowboard athletes in 2014.

There were 112 female athletes in the 2018 Olympics. The key statistics of athletes' age are mean: 25.29, standard deviation: 4.87, min:17 and max: 45.

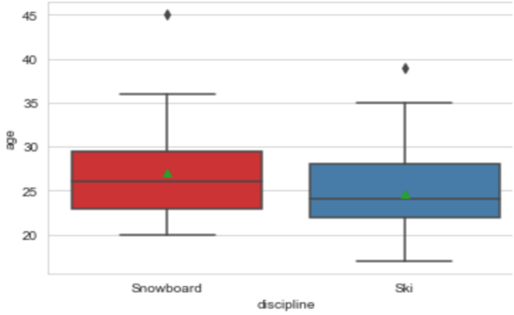

Figure 8: box plot, age difference between female snowboarding and ski athletes in 2018

There were 31 female snowboard athletes and 81 female ski athletes in 2018. Female snowboard athletes' age has these key statistics: mean: 27, min: 20, median: 26, max: 45 in 2018. On the other hand, female ski athletes' age has a mean :24.63, min: 17, median: 24, max: 39 in 2018. Female snowboard athletes seem older than female ski athletes in 2018.

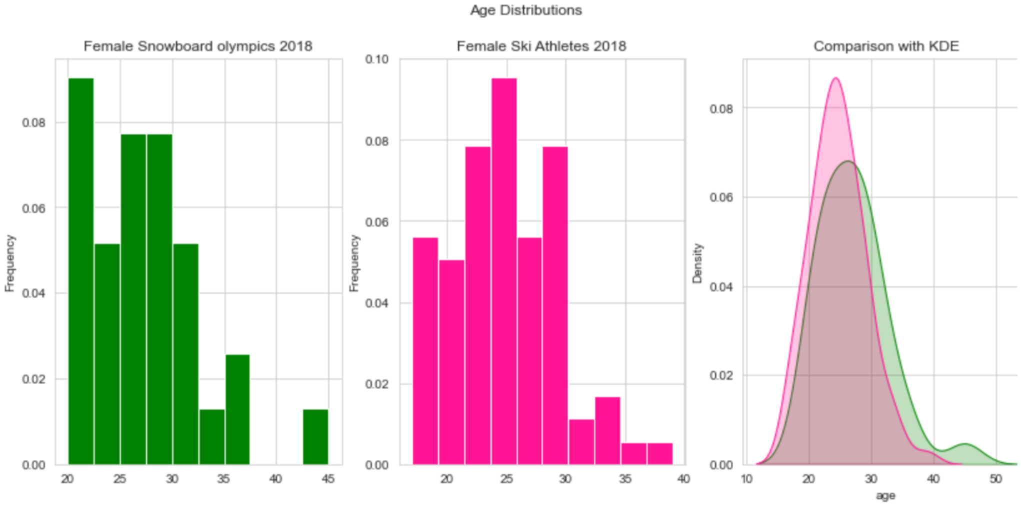

Figure 9: Histograms and KDE plots of female snowboard vs ski athletes' age 2018

According to figure 9, it is seen that the age of female ski athletes is slightly concentrated on the left compared to the age of female snowboard athletes, which means female ski athletes younger than female snowboard athletes in 2018.

There were 113 female athletes in the 2022 Olympics. The key statistics of athletes' age are mean: 25.49, standard deviation: 4.73, min:17 and max: 43.

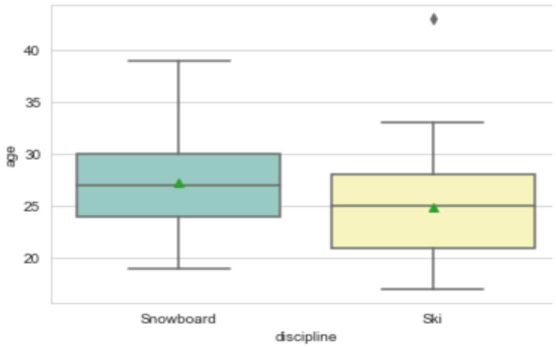

Figure 10: box plot, age difference between female snowboarding and ski athletes in 2022

There were 31 female snowboard athletes and 82 female ski athletes in 2022. Female snowboard athletes' age has these key statistics: mean: 27.26, min: 19, median: 27, max: 39. On the other hand, female ski athletes' age has a mean :24.82, min: 17, median: 25, max: 43. Female snowboard athletes seem older than female ski athletes in 2022.

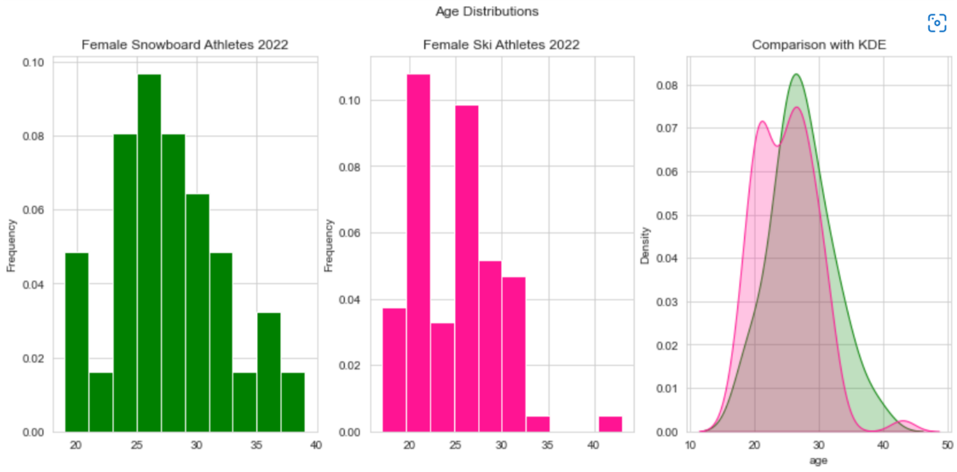

Figure 11: Histograms and KDE plots of female snowboard vs ski athletes' age 2022

According to figure 11, it is seen that the age of female ski athletes is slightly concentrated on the left compared to the age of female snowboard athletes, which means female ski athletes younger than female snowboard athletes in 2022.

Hypothesis Tests

Unless otherwise stated first group is snowboard, second group is ski and α=0.05 for all hypothesis tests. This hypothesis test on difference of means of female snowboard and ski athletes' age in 2014 olympics. We have 2 independent random samples which sizes, means and variances are respectively; n1=32, n2=89, μ_1=26.38, μ_2=24.39 and σ_1^2 =25.4 and σ_2^2 = 15.92. According to the CLT, the means of our two samples follow approximately normal distribution because sample sizes are sufficiently large ($n \geq 30$). First, we will apply the F-test to check equality of variances. $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$. After applying F-test in python, F value: 1.59 and p=0.047 are found. P <α (0.05) so, H_0 is rejected, variances are not equal. We should apply Welch's test. H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$. Test statistics: 2.01 and p=0.0504 are found. P >α so, we cannot reject H_0. There is no significant difference between the means of female snowboard and ski athletes' ages in 2014.

This hypothesis test on the difference of means of female snowboard and ski athletes' age in 2018 Olympics. We have 2 independent random samples which sizes , means and variances are respectively; n1=31, n2=81, μ_1=27, μ_2=24.63 and σ_1^2 =30.53 and σ_2^2 = 19.84. According to the CLT, the means of our two samples follow approximately normal distribution ($n \geq 30$). First, we will apply the F-test to check equality of variances. $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$. F value: 1.54 and p=0.066 are found. P >α so, we cannot reject H_0 which means we assume that variances are equal. We will apply Student two samples t test. H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$. T value: 2.35 and p=0.02 are found. P < α so H_0 is rejected, there is a significant difference between the means of female snowboard and ski athletes' ages in 2018. Also , Test statistic is positive so, the difference in favor of female snowboard athletes, we can say that female snowboard athletes are older than female ski athletes.

This hypothesis test on the difference of means of female snowboard and ski athletes' age in 2022 Olympics. We have 2 independent random samples which sizes , means and variances are respectively; n1=31, n2=82, μ_1=27.26, μ_2=24.82 and σ_1^2 =22.6 and σ_2^2 = 20.92. According to the CLT, the means of our two samples follow approximately normal distribution ($n \geq 30$). First, we will apply the F-test to check equality of variances. $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$. F value: 1.08 and p=0.38 are found. P >α so, we cannot reject H_0 which means we assume that variances are equal. We will apply Student two samples t test. H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$. T value: 2.5 and p=0.014 are found. P < α so H_0 is rejected, there is a significant difference between the means of female snowboard and ski athletes' ages in 2022. Also , Test statistic is positive so, the difference in favor of female snowboard athletes, we can say that female snowboard athletes are older than female ski athletes.

REFERENCES

1. Miller, I., & Miller, M. (2015). *John E. Freund's mathematical statistics with applications* (8th ed.). Pearson.
2. Navarro, D. (2020, August 11). *13.4: The independent samples T-TesT (Welch test)*. Statistics LibreTexts. https://shorturl.ae/1psRg
3. OpenStaxCollege. (2013, July 18). *Box plots – Introductory statistics*. BCcampus Open Publishing – Open Textbooks Adapted and Created by BC Faculty. https://shorturl.ae/8wMLK

4. OSCRiceUniversity. (2015, March 31). *Test of two variances – Introductory business statistics*. BCcampus Open Publishing – Open Textbooks Adapted and Created by BC Faculty. https://opentextbc.ca/introbusinessstatopenstax/chapter/test-of-two-variances/

5. Rice University. (2022, January 27). *2.2 histograms, frequency polygons, and time series graphs - Introductory statistics | OpenStax*. OpenStax. https://openstax.org/books/introductory-statistics/pages/2-2-histograms-frequency-polygons-and-time-series-graphs

6. Waskom, M. (2021). *Seaborn.kdeplot — Seaborn 0.11.2 documentation*. seaborn: statistical data visualization — seaborn 0.11.2 documentation. https://seaborn.pydata.org/tutorial/distributions.html#tutorial-kde