Islam Ahmed Afifi
Statistical report for Olympic winter games
16.06.2022

ii) Table of contents

iii) Motivation

1. definition of the problem

- Olympics has a rich history, spanning from 1896 till 2018, and has been a part of history[1]. So, it is an interesting topic to see how historical events have affected [1]. the specifics of the Olympics and how it has been changing to date[1]. Hence, this report attempts to build intuition around answering the following question :

- Are snowboard athletes aged to be younger than alpine skiing ages as we expect?.

- Is there a significant difference in the location of the variable age for the following two groups: Female athletes snowboarding and female athletes alpine skiing? Through the period from 2014 to 2016.

2- overview of the structure of the report.

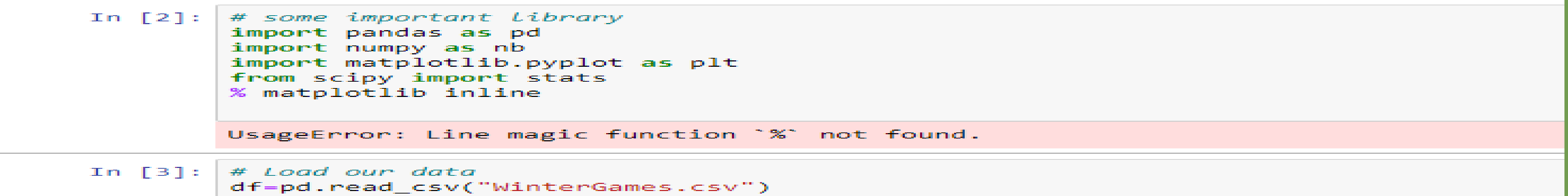
- 2.1 information about Olympic Data

- 2.2-tidy Data

- 2.3 Data cleaning

iv) Detailed description of the problem

- in this section, we want to explore and visualize the Olympic winter games through the period from 2014 to 2016 to get patterns from our data to answer some questions like, Are snowboard athletes aged to be younger than alpine skiing age as we expect? By answering some questions, we can extract some knowledge to help us know much more intuition about our world.
- **1- information about Olympic Data**
- In this data set, you find information about the Olympic winter games 2014, 2018, and 2022, as indicated by
- the variable Olympics. For each Olympic, we picked all starters of the (parallel) giant slalom in snowboarding and alpine skiing. The respective discipline is indicated by the variable discipline. For all athletes,
- the data set includes their sex (m for male, f for female) and their age. Age is defined as the difference between the birth year and the year the Olympics did take place.
- **2- Data Wrangling**
- **In this section of the report, we will load the data {Figure 1}.**

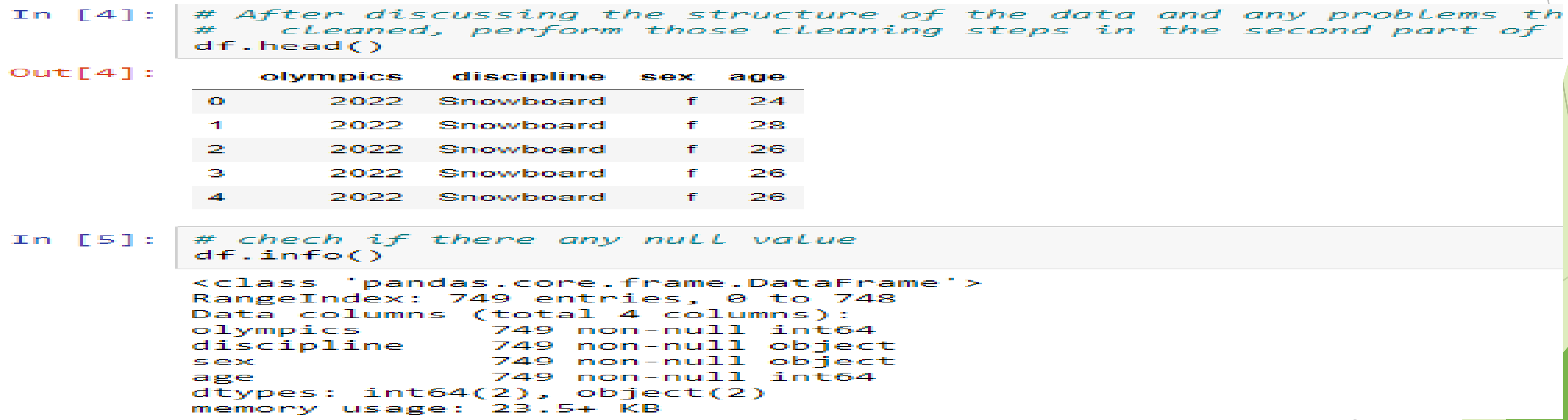


```
In [2]:  # some important library
         import pandas as pd
         import numpy as nb
         import matplotlib.pyplot as plt
         from scipy import stats
         % matplotlib inline

         UsageError: Line magic function `%` not found.

In [3]:  # Load our data
         df=pd.read_csv("WinterGames.csv")
```

Figure 1

3- Data Cleaning

In this section of the report, we will check for cleanliness, and then trim and clean our dataset for analysis {Figure2}.



```
In [4]:  # After discussing the structure of the data and any problems th
         #    cleaned, perform those cleaning steps in the second part of
         df.head()

Out[4]:
```

	olympics	discipline	sex	age
0	2022	Snowboard	f	24
1	2022	Snowboard	f	28
2	2022	Snowboard	f	26
3	2022	Snowboard	f	26
4	2022	Snowboard	f	26

```
In [5]:  # chech if there any null value
         df.info()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 749 entries, 0 to 748
         Data columns (total 4 columns):
         olympics      749 non-null int64
         discipline    749 non-null object
         sex           749 non-null object
         age           749 non-null int64
         dtypes: int64(2), object(2)
         memory usage: 23.5+ KB
```

Figure 2

v) methods

- **1-Measures of Central Tendency**
- Central tendency is the value that describes the entire set of data as a single measurement. The three primary measures of central
- tendency are the mean, median, and mode.[2]
- The following example will be used to demonstrate these three measures.[2]
- Sample A (age in years) - 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60Sample B (age in years) - 52, 54, 54, 54, 55, 56, 57, 57, 58, 58,
- 60, 60 [2]
- Mean is the arithmetic average or the sum of values in a dataset divided by the total number of observations [2]. Using the above
- example, the mean of Sample A is 54 + 54 + 54 + 55 + 56 + 57 + 57 + 58 + 58 + 60 + 60 = 623, divided by 11 (total number of
- observations) = 56.6 years. [2]
- The mean should only be reported with interval and ratio data that are normally distributed (i.e., look like a “bell-shaped” curve) since
- this measure of central tendency is strongly affected by outliers and skewed distributions.[2]
- A formula for the Mean is given by : $\mu = \frac{\sum x_i}{n}$

- Median is the middle value in distribution when the data are ranked in order from highest to lowest (or vice versa)[2]. If there is an odd
- number of values, the median is the exact middle value; however, if there is an even number of values, the median is the average of
- the two middle values[2]. In the example above, the median for Sample A is 57 and for Sample, B is 56 + 57/2 = 56.5[2].
- Since the median is less affected by outliers and skewed distributions, it is the appropriate measure to report when data do not follow
- a “bell-shaped” curve. The median should also be reported with ordinal data[2].
- Mode is the most common value in a dataset[2]. In the above examples, the mode for Samples A and B is 54[2].
- Although the mode may be used for both qualitative and quantitative variables, it may not accurately represent the center of the
- Distribution [2]. Using the above example, the Sample A mode is 54, but the center of distribution is 57 years[2].
- Sometimes, there may not be a mode if all values are different or if there is a bimodal or multimodal sample (signifying peaks at two
- or more places in the data distribution)[2]. In such cases, one may report the mean or median as appropriate[2].
- As illustrated previously, the shape of the data distribution may influence the measures of central tendency[2]. When the distribution is
- symmetrical (i.e., “bell-shaped”), the mean, median, and mode are all in the middle {Figure 3}[2]. When the distribution is
- skewed toward the low end of values (positive skew), the mode remains the most common value, and the median remains the
- middle value, but the mean is pulled toward the right tail of the distribution {Figure 4}[2]. When the distribution is skewed toward
- the high end of values (negative skew), the mean is pulled toward the left tail of the distribution {Figure 5}[2].

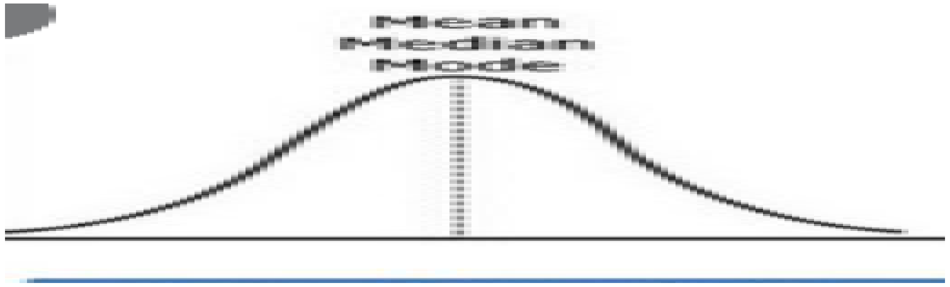


Figure 3

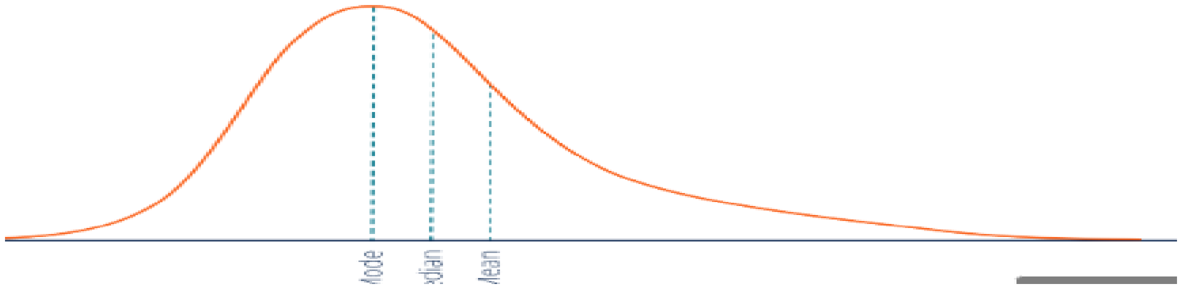


Figure 4

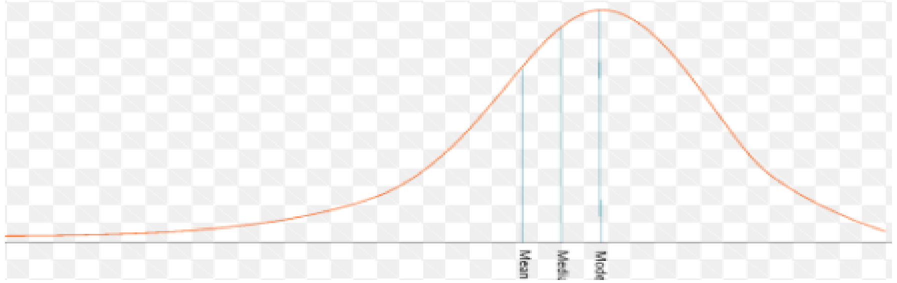


Figure 5

2-Measures of Dispersion/variation

- Although measures of central tendency provide important information when describing one's data, they fail to capture variability
- within a dataset.[3] Measures of dispersion/variation describe the degree to which a variable's values are similar or diverse [2]. This type
- of measure only applies to the ordinal, interval, and ratio data that can be ranked and includes the range, variance, and standard
- Deviation [2].
- The range is the difference between the lowest and the highest values in a dataset. For example, the range of Sample A above is 6
- (60–54 = 6), while the range of Sample B is 8 (60–52 = 8) [2].
- The variance and standard deviation are measures of spread that reveal how close each observed value is to the mean of the entire
- Dataset. In datasets with a small spread, all values are close to the mean, yielding smaller variance and standard deviation [2]. In contrast,
- Datasets with a greater spread of values away from the mean have larger variance and standard deviation. Therefore, if all values of a
- dataset is the same, the variance and standard deviation will be zero [2].
- In a normally distributed dataset, 68% of the values are within one standard deviation on either side of the mean, and 95% of values are
- within two standard deviations, and 99% of values are within three standard deviations[3].
- the variance formula : $\frac{\sum(xi - \bar{x})^2}{N}$

3-Measures of Position

- Determining the position of values in a dataset may be accomplished in three main ways [2].
- Percentiles divide the dataset into 100 equal sections, deciles divide it into ten equal parts, and quartiles divide an ordered dataset
- into four equal parts [2]. The differences between percentiles and quartiles are minor and often disappear with a large number of values
- in a dataset. One may clearly see how they are associated as follows:
- The lower quartile, Q1 (25th percentile), is the point between the lowest 25% and highest 75% of values[2]. The second quartile, Q2
- (50th percentile), is the median (middle of the dataset) [2]. The upper quartile, Q3 (75th percentile), is the point between the lowest 75%
- and highest 25% of values. If the quartile falls between two values, the average of those values represents the quartile value
- Box plots are often useful for interpreting descriptive data in the graphical form[3]. As seen in {Figure 6}, box plots are constructed using
- the 25th percentile (lower quartile), the median (50th percentile), the 75th percentile (upper quartile), the minimum data value, and
- the maximum data value[2]. Box plots also show outlier values{Figure 6}[2].

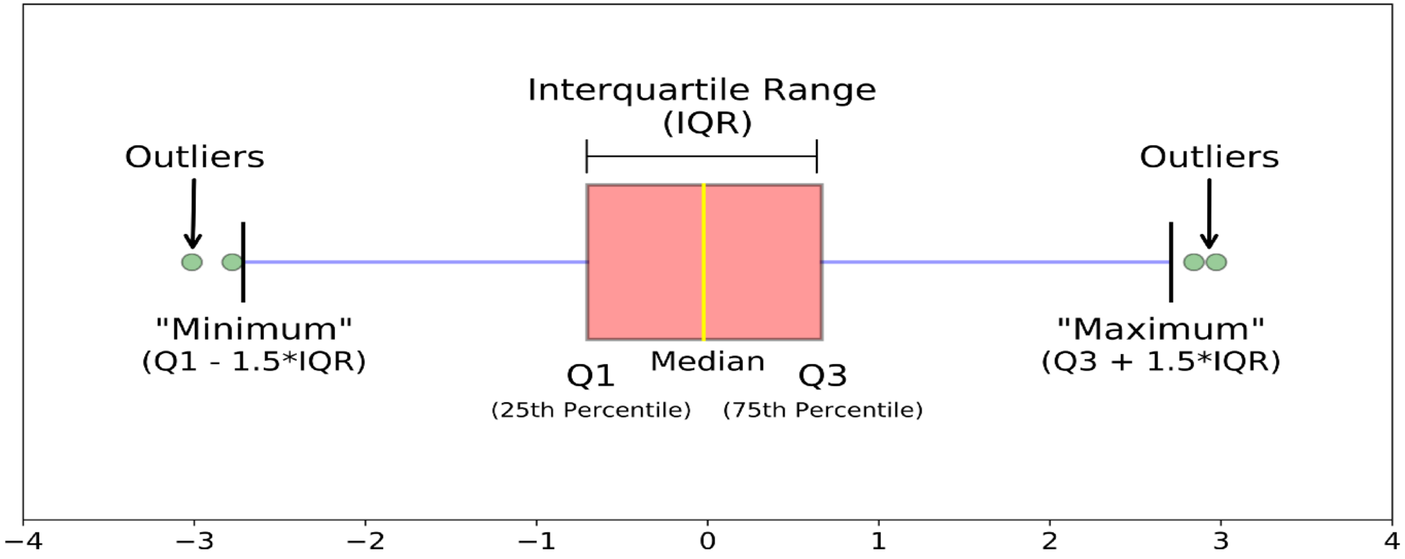


Figure 6

3- Histogram

The formula for the histogram revolves around the area of the bars [4]. It is simple. It is calculated by the summation of the product of the frequency density of each class interval and the corresponding class interval’s width[2]. The area of the histogram formula is mathematically represented as [4]:

Area of Histogram =

4- t-test

A t-test is a statistical $\sum frequence_i * classwidth_i$ means of two groups [5]. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another[5].

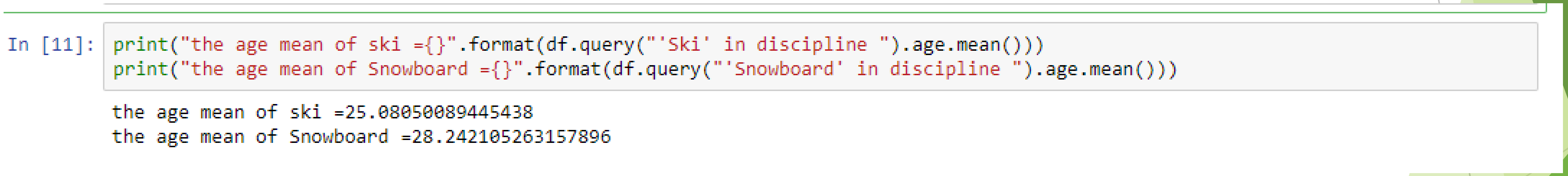
The t-test estimates the true difference between two group means using the ratio of the difference in group means over the pooled standard error of both groups[5]. You can calculate it manually using a formula, or use statistical analysis software[5].

T-test formula

The formula for the two-sample t-test (a.k.a. the Student’s t-test) is shown below[5]. $t=\frac{\mu_1-\mu_2}{\sqrt[2]{(s^2(\frac{1}{n_1}+\frac{1}{n_2})}}$

vi) Evaluation (Exploratory and visualization Data)

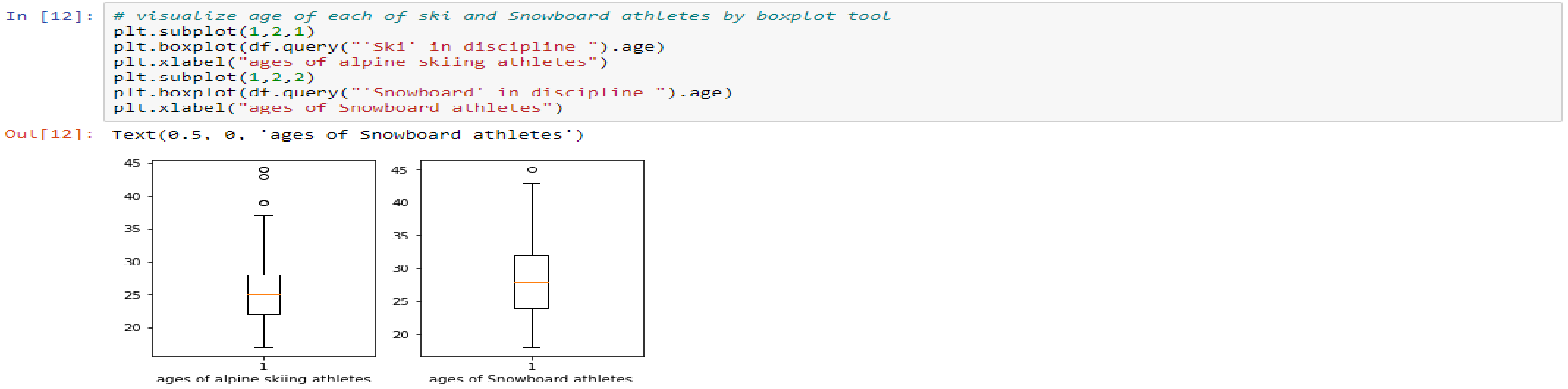
- Now that we've trimmed and cleaned our data, we're ready to move on to exploration. **Compute statistics** and **create visualizations** with the goal of addressing the research questions that we posed in the Introduction section. we should compute the relevant statistics throughout the analysis when an inference is made about the data.
- **1- Research Question 1 (Are snowboard athletes aged to be younger than alpine skiing ages as we expect!)**
- After making exploration and visualization of our Data we noticed that: the mean of alpine skiing ages was less than Snowboard athlete ages {Figure 7}, and visualize each of them by boxplot, which shows that the interval of the ages of alpine skiing is less than Snowboard ages too {Figure 8}, and visualizes each of them by histogram which shows us that each of them is a positively skewed distribution {Figure 9, Figure 10}, which means that each of them Tends to Younger age groups but the distribution of alpine skiing ages are more skewed than Snowboard ages so alpine skiing ages have ages younger than Snowboard ages.



```
In [11]: print("the age mean of ski ={}".format(df.query("'Ski' in discipline ").age.mean()))
         print("the age mean of Snowboard ={}".format(df.query("'Snowboard' in discipline ").age.mean()))
```

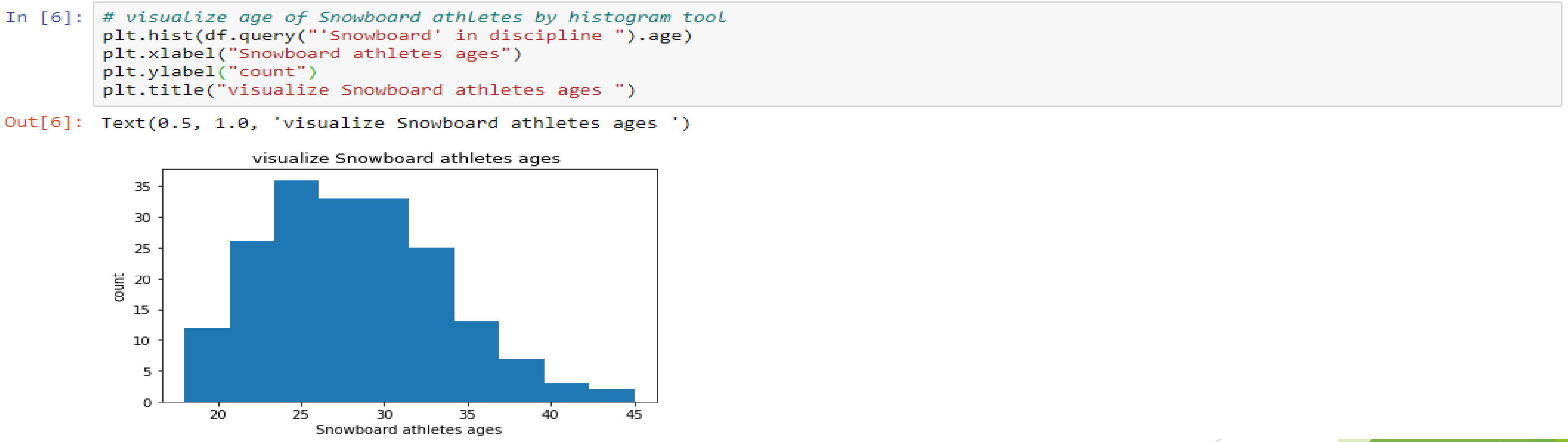
```
the age mean of ski =25.08050089445438
the age mean of Snowboard =28.242105263157896
```

Figure 7



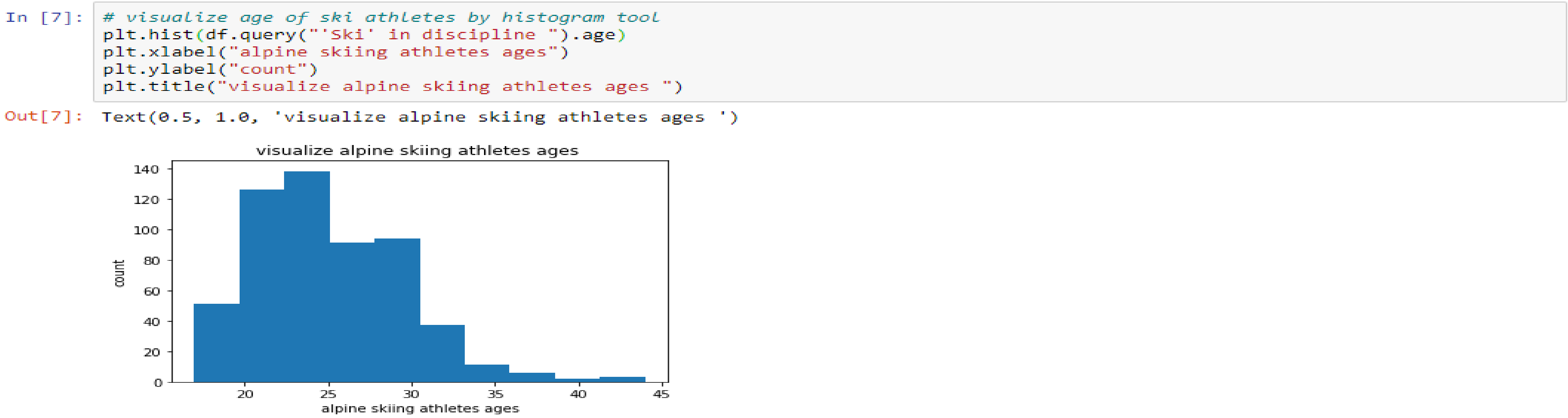
```
In [12]: # visualize age of each of ski and Snowboard athletes by boxplot tool
         plt.subplot(1,2,1)
         plt.boxplot(df.query("'Ski' in discipline ").age)
         plt.xlabel("ages of alpine skiing athletes")
         plt.subplot(1,2,2)
         plt.boxplot(df.query("'Snowboard' in discipline ").age)
         plt.xlabel("ages of Snowboard athletes")

Out[12]: Text(0.5, 0, 'ages of Snowboard athletes')
```

Figure 8



```
In [6]: # visualize age of Snowboard athletes by histogram tool
        plt.hist(df.query("'Snowboard' in discipline ").age)
        plt.xlabel("Snowboard athletes ages")
        plt.ylabel("count")
        plt.title("visualize Snowboard athletes ages ")

Out[6]: Text(0.5, 1.0, 'visualize Snowboard athletes ages ')
```

Figure 9

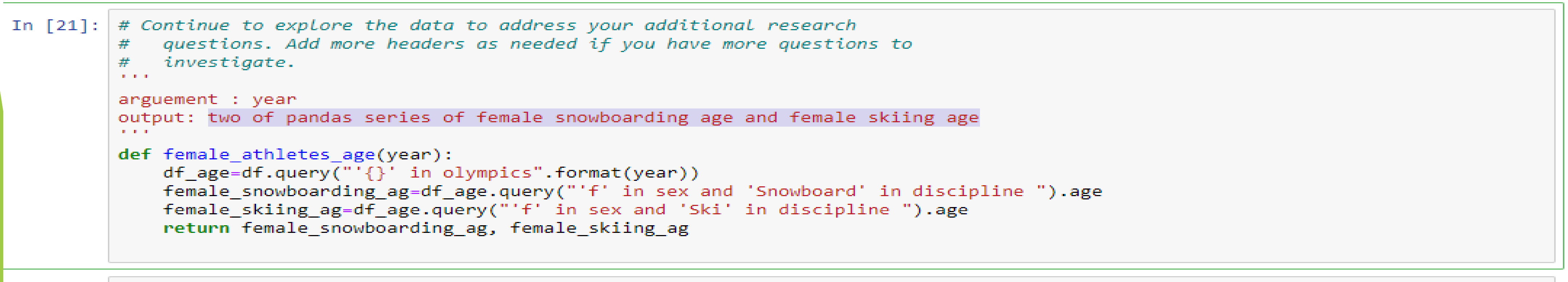


```
In [7]: # visualize age of ski athletes by histogram tool
        plt.hist(df.query("'Ski' in discipline ").age)
        plt.xlabel("alpine skiing athletes ages")
        plt.ylabel("count")
        plt.title("visualize alpine skiing athletes ages ")

Out[7]: Text(0.5, 1.0, 'visualize alpine skiing athletes ages ')
```

visualize alpine skiing athletes ages
count
0
20
40
60
80
100
120
140
20
25
30
35
40
45
alpine skiing athletes ages

Figure 10

Techincal Hint: This function takes one argument (year) and returns two of pandas series of female snowboarding ages and female skiing ages{figure 11}.



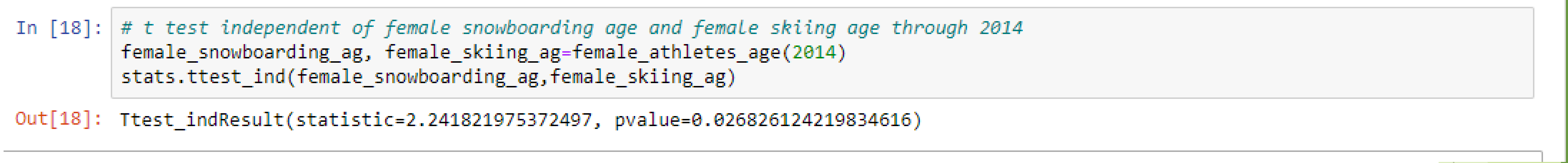
```
In [21]: # Continue to explore the data to address your additional research
         #   questions. Add more headers as needed if you have more questions to
         #   investigate.
         '''
         arguement : year
         output: two of pandas series of female snowboarding age and female skiing age
         '''
         def female_athletes_age(year):
             df_age=df.query("'{}' in olympics".format(year))
             female_snowboarding_ag=df_age.query("'f' in sex and 'Snowboard' in discipline ").age
             female_skiing_ag=df_age.query("'f' in sex and 'Ski' in discipline ").age
             return female_snowboarding_ag, female_skiing_ag
```

Figure 11

2- Research Question 2 (Is there a significant difference in the location of the variable age for the following two groups: female athletes snowboarding and female athletes alpine skiing? Through the period from 2014 to 2016!)

2.1-Through 2014

- H0:female snowboarding ages = female skiing ages.
- H1:female snowboarding ages !=female skiing ages.
- Notice from {Figure 12} that p-value <0.05 so we reject H0.

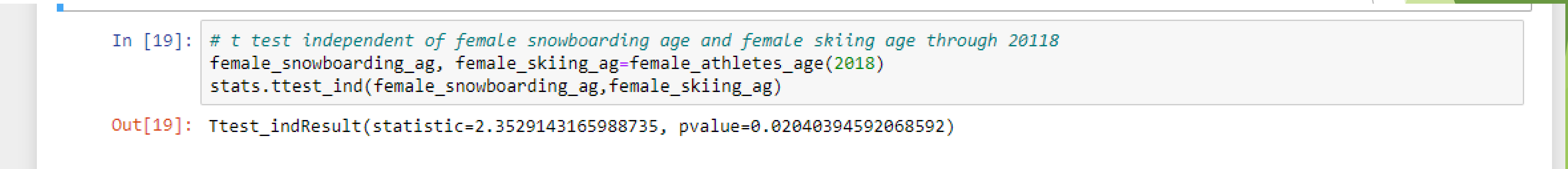


```
In [18]: # t test independent of female snowboarding age and female skiing age through 2014
         female_snowboarding_ag, female_skiing_ag=female_athletes_age(2014)
         stats.ttest_ind(female_snowboarding_ag,female_skiing_ag)

Out[18]: Ttest_indResult(statistic=2.241821975372497, pvalue=0.026826124219834616)
```

Figure 12

2.2-Through 2018

- H0:female snowboarding ages = female skiing ages.
- H1:female snowboarding ages !=female skiing ages.
- Notice from {Figure 13} that p-value <0.05 so we reject H0.

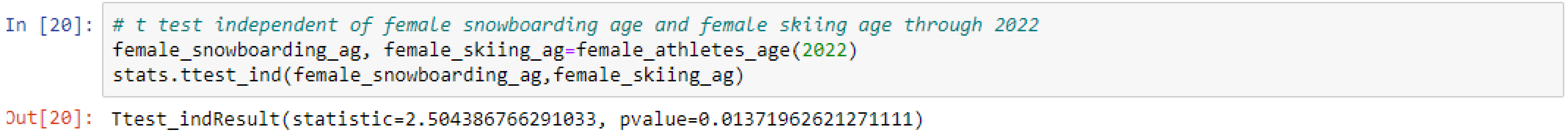


```
In [19]: # t test independent of female snowboarding age and female skiing age through 20118
         female_snowboarding_ag, female_skiing_ag=female_athletes_age(2018)
         stats.ttest_ind(female_snowboarding_ag,female_skiing_ag)

Out[19]: Ttest_indResult(statistic=2.3529143165988735, pvalue=0.02040394592068592)
```

Figure 13

2.3-Through 2022

- H0:female snowboarding ages = female skiing ages.
- H1:female snowboarding ages !=female skiing ages.
- Notice from {Figure 14} that p-value <0.05 so we reject H0.



```
In [20]: # t test independent of female snowboarding age and female skiing age through 2022
         female_snowboarding_ag, female_skiing_ag=female_athletes_age(2022)
         stats.ttest_ind(female_snowboarding_ag,female_skiing_ag)

Out[20]: Ttest_indResult(statistic=2.504386766291033, pvalue=0.01371962621271111)
```

Figure 14

vii) Summary (Conclusions)

- Finally, we summarize our findings and the results that have been performed in relation to the question(s) provided at the beginning of the analysis. We Summarize the results accurately and point out where additional research can be done or where additional information could be useful.
- **1-Results are our Data suggest that :**
- 1.1- alpine skiing ages are younger than a snowboard age, Contrary to what we expected.
- 1.2- there is a significant 95% that be difference in the location of the variable age for the following two groups: female athletes snowboarding and female athletes alpine skiing through the period from 2014 to 2016.
- **2-limitation: there is a couple of limitations with our data :**
- limitation of features don’t let us answer some questions like :
- 2.1- Are there some sports that tend to fall into certain categories (Male or Female)? we should have a lot of types of Olympic games to use the chi test to answer this question.
- 2.2- What effect does the host country have on the medals won at the Olympics?[1].
- 2.3- Is the performance of countries in Olympic games affected by the economic factors of the country?[1].

viii) Bibliography

- 1- towards data science, Visual Analysis of Olympics Data, Jul 21, 2020, Available from:https://towardsdatascience.com/visual-analysis-of-olympics-data-16273f7c6cf2.
- 2- Vikas Yellapu,_Descriptive statistics, April 2018, International Journal of Academic Medicine, researchgate.
- 3- Sonnad SS. Describing data: Statistical and graphical methods. Radiology 2002;225:622-8.
- 4- wallstreetmojo, Histogram formula, (Article by Dheeraj Vaidya, CFA, FRM), Available from Histogram Formula | Calculate Area using Histogram Equation (Examples) (wallstreetmojo.com).
- 5- scribbr, An Introduction to T-Tests | Definitions, Formula, and Examples,(Published on January 31, 2020, by Rebecca Bevans. Revised on May 23, 2022.),
- Available from: https://www.wallstreetmojo.com/histogram-formula/.
- 6- pandas library, Available from:https://pandas.pydata.org/docs/reference/index.html.
- 7- matplotlib pyplot, Available from: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html.