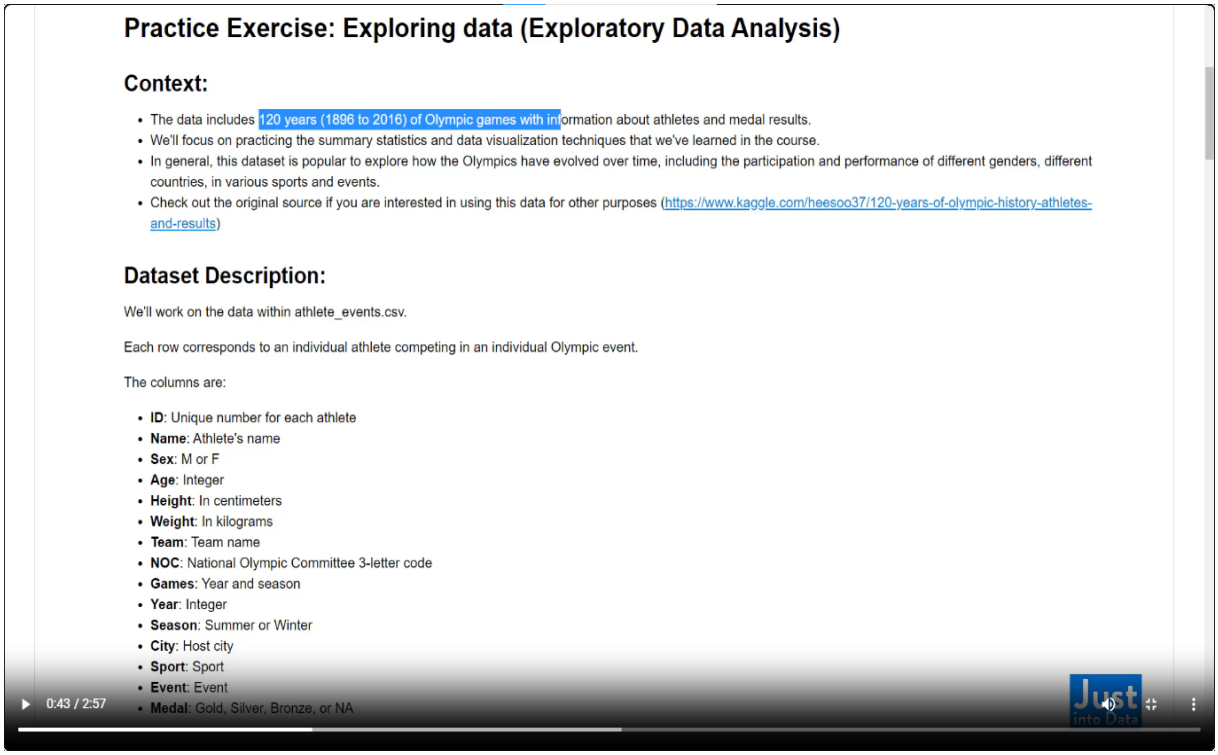
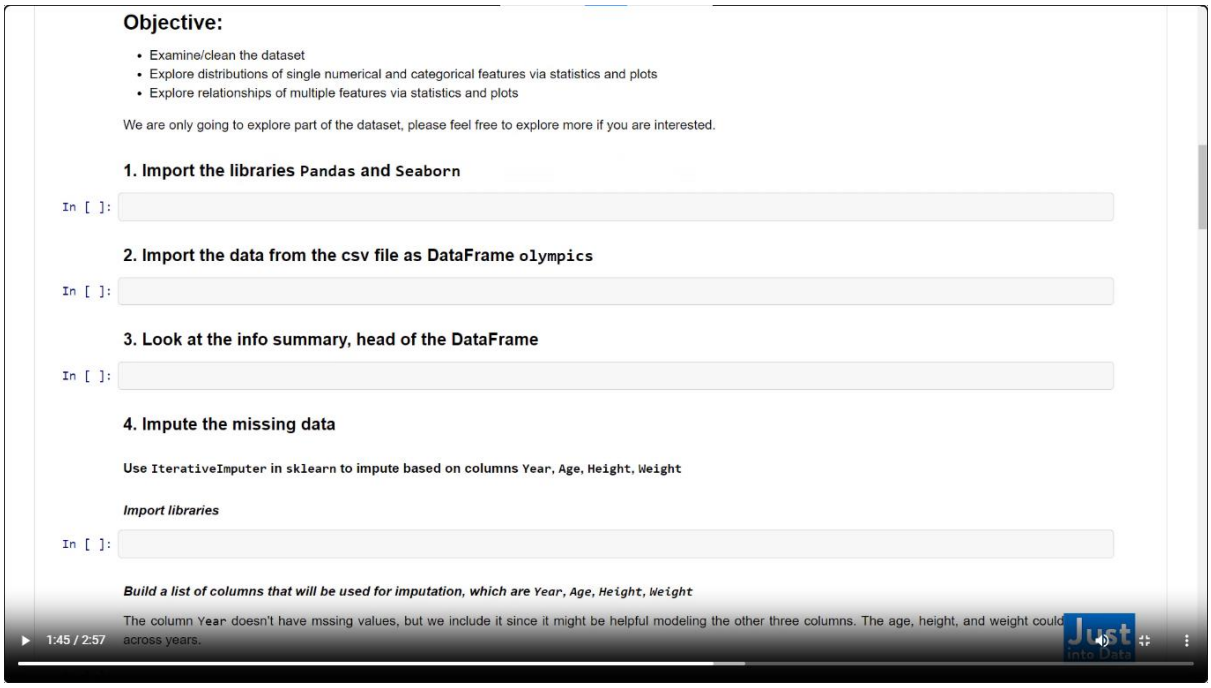
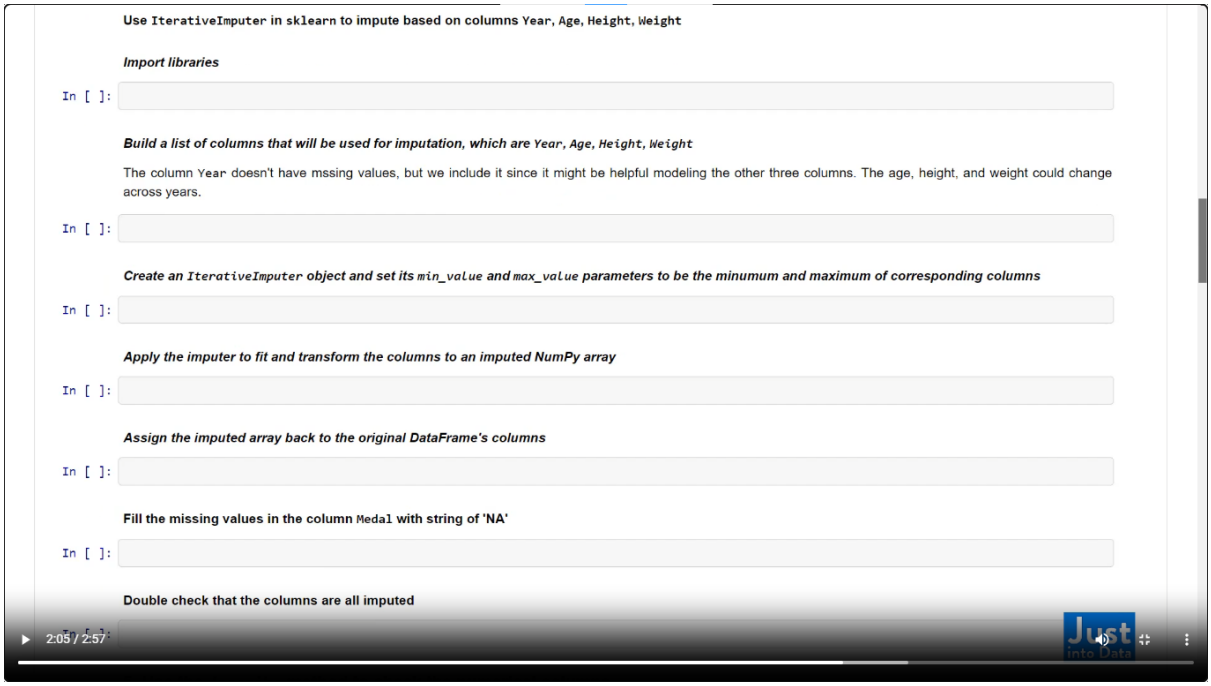
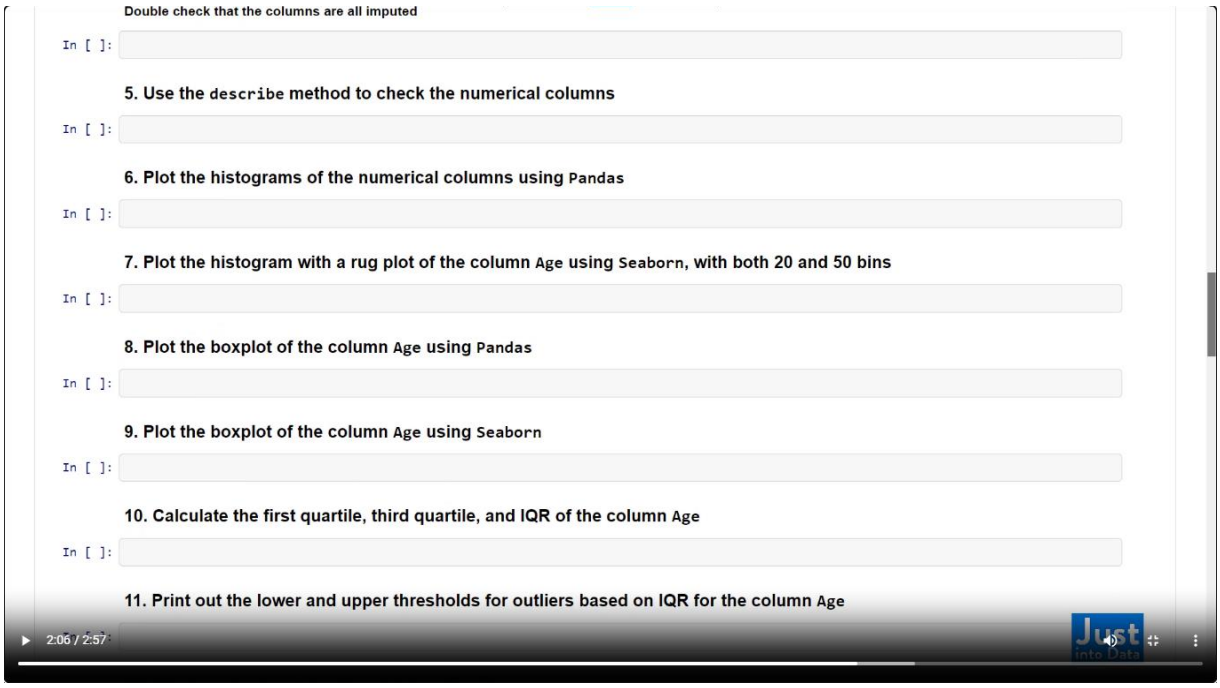
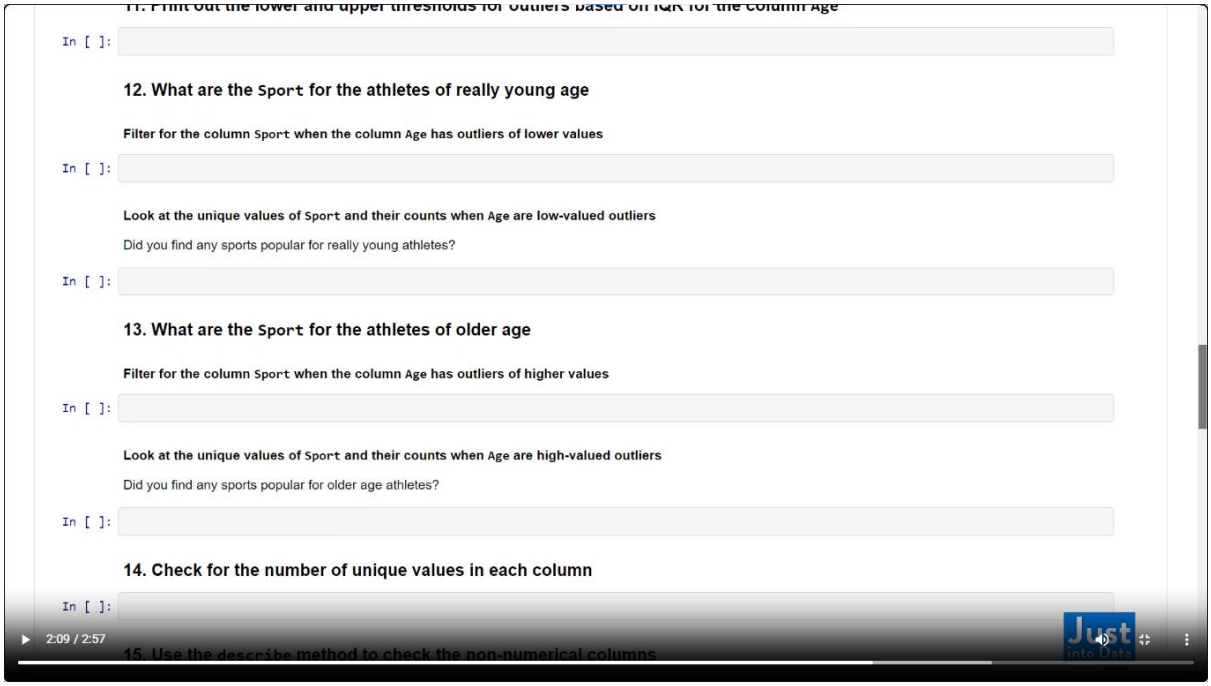
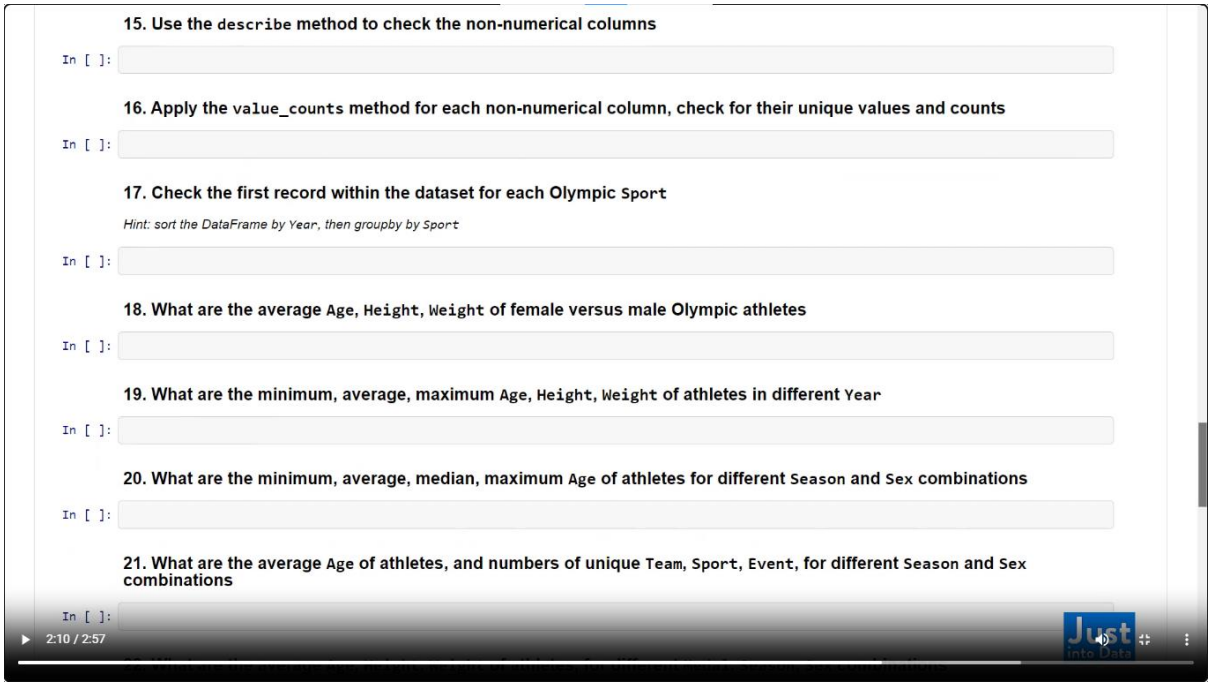

Practice Exercise: Exploring data (Exploratory Data Analysis)
Context:
The data includes 120 years (1896 to 2016) of Olympic games with information about athletes and medal results.
We'll focus on practicing the summary statistics and data visualization techniques that we've learned in the course.
In general, this dataset is popular to explore how the Olympics have evolved over time, including the participation and performance of different genders, different countries, in various sports and events.
Check out the original source if you are interested in using this data for other purposes (https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results)
Dataset Description:
We'll work on the data within athlete_events.csv.
Each row corresponds to an individual athlete competing in an individual Olympic event.
The columns are:
ID: Unique number for each athlete
Name: Athlete's name
Sex: M or F
Age: Integer
Height: In centimeters
Weight: In kilograms
Team: Team name
NOC: National Olympic Committee 3-letter code
Games: Year and season
Year: Integer
Season: Summer or Winter
City: Host city
Sport: Sport
Event: Event
Medal: Gold, Silver, Bronze, or NA
0:43 / 2:57


Objective:
Examine/clean the dataset
Explore distributions of single numerical and categorical features via statistics and plots
Explore relationships of multiple features via statistics and plots
We are only going to explore part of the dataset, please feel free to explore more if you are interested.
1. Import the libraries Pandas and Seaborn
In [ ]:
2. Import the data from the csv file as DataFrame olympics
In [ ]:
3. Look at the info summary, head of the DataFrame
In [ ]:
4. Impute the missing data
Use IterativeImputer in sklearn to impute based on columns Year, Age, Height, Weight
Import libraries
In [ ]:
Build a list of columns that will be used for imputation, which are Year, Age, Height, Weight
The column Year doesn't have mssing values, but we include it since it might be helpful modeling the other three columns. The age, height, and weight could
across years.
1:45 / 2:57


Use IterativeImputer in sklearn to impute based on columns Year, Age, Height, Weight
Import libraries
In [ ]:
Build a list of columns that will be used for imputation, which are Year, Age, Height, Weight
The column Year doesn't have mssing values, but we include it since it might be helpful modeling the other three columns. The age, height, and weight could change across years.
In [ ]:
Create an IterativeImputer object and set its min_value and max_value parameters to be the minumum and maximum of corresponding columns
In [ ]:
Apply the imputer to fit and transform the columns to an imputed NumPy array
In [ ]:
Assign the imputed array back to the original DataFrame's columns
In [ ]:
Fill the missing values in the column Medal with string of 'NA'
In [ ]:
Double check that the columns are all imputed
2:05 / 2:57


Double check that the columns are all imputed
In [ ]:
5. Use the describe method to check the numerical columns
In [ ]:
6. Plot the histograms of the numerical columns using Pandas
In [ ]:
7. Plot the histogram with a rug plot of the column Age using Seaborn, with both 20 and 50 bins
In [ ]:
8. Plot the boxplot of the column Age using Pandas
In [ ]:
9. Plot the boxplot of the column Age using Seaborn
In [ ]:
10. Calculate the first quartile, third quartile, and IQR of the column Age
In [ ]:
11. Print out the lower and upper thresholds for outliers based on IQR for the column Age
2:06 / 2:57


11. Print out the lower and upper thresholds for outliers based on IQR for the column Age
In [ ]:
12. What are the Sport for the athletes of really young age
Filter for the column Sport when the column Age has outliers of lower values
In [ ]:
Look at the unique values of Sport and their counts when Age are low-valued outliers
Did you find any sports popular for really young athletes?
In [ ]:
13. What are the Sport for the athletes of older age
Filter for the column Sport when the column Age has outliers of higher values
In [ ]:
Look at the unique values of Sport and their counts when Age are high-valued outliers
Did you find any sports popular for older age athletes?
In [ ]:
14. Check for the number of unique values in each column
In [ ]:
2:09 / 2:57
15. Use the describe method to check the non-numerical columns


15. Use the describe method to check the non-numerical columns
In [ ]:
16. Apply the value_counts method for each non-numerical column, check for their unique values and counts
In [ ]:
17. Check the first record within the dataset for each Olympic Sport
Hint: sort the DataFrame by Year, then groupby by Sport
In [ ]:
18. What are the average Age, Height, Weight of female versus male Olympic athletes
In [ ]:
19. What are the minimum, average, maximum Age, Height, Weight of athletes in different Year
In [ ]:
20. What are the minimum, average, median, maximum Age of athletes for different Season and Sex combinations
In [ ]:
21. What are the average Age of athletes, and numbers of unique Team, Sport, Event, for different Season and Sex combinations
In [ ]:
2:10 / 2:57

In [ ]:

22. What are the average `Age`, `Height`, `Weight` of athletes, for different `Medal`, `Season`, `Sex` combinations

In [ ]:

23. Plot the scatterplot of `Height` and `Weight`

In [ ]:

24. Plot the scatterplot of `Height` and `Weight`, using different colors and styles of dots for different `Sex`

In [ ]:

25. Plot the pairwise relationships of `Age`, `Height`, `Weight`

In [ ]:

26. Plot the pairwise relationships of `Age`, `Height`, `Weight`, with different colors for `Sex`

In [ ]:

27. Print out the correlation matrix of `Age`, `Height`, `Weight`

In [ ]:

28. Use heatmap to demonstrate the correlation matrix of `Age`, `Height`, `Weight`, use a colormap (`cmap`) of 'crest'

In [ ]:

29. Plot the histograms of `Age`, with different colors for different `Sex`

2:12 / 2:57

In [ ]:

29. Plot the histograms of `Age`, with different colors for different `Sex`

In [ ]:

30. Plot the histograms of `Age`, on separate plots for different `Sex`

In [ ]:

31. Look at the changes of average `Age` across `Year` by line charts, with separate lines for different `Season` using different colors

In [ ]:

32. Look at the distributions of `Age` for different `Sex` using boxplots

In [ ]:

33. Look at the distributions of `Age` for different `Sex` using violin plots

In [ ]:

34. Look at the distributions of `Age` for different `Sex` using boxplots, with different colors of plots for different `Season`

In [ ]:

35. Use count plots to look at the changes of number of athlete-events across `Year`, for different `Sex` by colors, and different `Season` on separate plots

In [ ]:

2:14 / 2:57