3031_A6_v2_Template

November 11, 2021

1 Assignment 6

Instructions:

a. You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and not as a group activity.

b. Do not write your name on the assignment. (1 point)

c. Please include each question (or question number) followed by code and your answer (if applicable). Write your code in the 'Code' cells and your answer in the 'Markdown' cells of the Jupyter notebook. Ensure that the solution is written neatly enough to understand and grade. (1/2 point value of each question)

d. Export your Jupyter notebook as a PDF file. If you get an error, make sure you have downloaded the MikTex software (for windows) or MacTex (for mac). Note that after installing MikTex/MacTex, you will need to check for updates, install the updates if needed, and re-start your system. Submit the PDF file. (1 point)

This assignment is due at 11:59pm on Tuesday, November 23rd. Good luck! *(52 points overall – 50 points for code & answers, 2 points for anonymity and proper formatting)*

1.1 Part 1

Read the spotify dataset from the file *spotify_data.csv*.

What percentage of all the unique tracks are contributed by the top 3 artists of each genre, where the top artists are based on *artist_popularity*, and the unique tracks are based on unique values of *track_name*? *(8 points for code)*

A typical approach that will **not** work: If you group the data by genre, and filter the top 3 rows by *artist_popularity*, then you may not get 3 unique artists, as one artist can have multiple tracks.

Here is one way to answer this question:

(1) Group the data by genre, artist name and artist popularity. Find the number of unique tracks (by *track_name*) for each group.

(2) The dataset obtained in (1) is at artist-genre level, i.e., each row corresponds to a unique artist-genre combination. Group that dataset by genre, and filter the top 3 rows of each

group based on artist popularity.

(3) Sum up the number of unique tracks of the dataset obtained in (2) and divide it by the total number of unique tracks in the original dataset.

Note: (1) The functions *len()* and *unique()* will be useful.
(2) If you can propose a solution that is shorter than the one proposed above, on Monday - 15th Nov, in class, you will get 10% bonus points for this assignment.

1.2 Part 2

Read data from the file "Canadian_Fish_Biodiversity.csv" on Canvas. Each row records a unique fishing event from a 2013 sample of fish populations in Ontario, Canada. *(42 points overall)*

1.2.1 Question 1

To analyze the results of these fishing surveys, we need to understand the dynamics of projects, sites, and geographic locations. In large part the following questions deal with missing data. *(16 points total)*

a) Each site (identified by the column *SITEID*) represents a time and place at which fishing events occurred. Sites are grouped into broader projects (identified by the column *Project Name*). We want to understand the scope of these projects.

Using *.groupby*, find the top three projects by number of unique sites. *(2 points for code)*

Hint: The Pandas function *nunique()* may help

b) Find the top three and bottom three projects in terms of the proportion of unique sites of the total number of unique sites. *(3 points for code)*

c) (i) How many values are missing for the air temperature column? *(1 point for code)*

(ii) Impute the missing values of air temperature with the median air temperature of the corresponding water body (*Waterbody Name*) and month. *(2 points for code)*

(iii) How many missing values still remain for the air temperature column after the imputation in (ii)? *(1 point for answer)*

(iv) We will try to impute the remaining missing values for air temperature. Try impute the remaining missing values of air temperature with the median air temperature of the corresponding project (*Project Name*) and month. *(2 points for code)*

(v) How many missing values still remain for the air temperature column after the imputation in (iv)? *(1 point for answer)*

(vi) Find the correlation between air temperature and water temperature. *(1 point for code)*

(vii) As you found a high correlation between air temperature and water temperature in (vi), you can use water temperature to estimate the air temperature (using the trendline, like you did in assignment 4). Assuming you already did that, how many missing values will still remain for the air temperature column? *Note: Do not impute the missing values using the trendline, just assume you already did that. (1 point for code)*

(viii) Make a scatterplot of air temperature against water temperature. Higlight the points for which the air temperature was imputed in (ii) and (iv) with a different color. *(2 points for code and visualization)*

1.2.2 Question 2

This section begins to investigate the living conditions of fish at different locations and time periods. *(7 points total)*

a) Use a single *.groupby* statement to view the minimum, mean, standard deviation, and maximum air temperature and water temperature for each project during the month of August (use the *Month* column). *(2 points for code)*

b) Make lineplots showing maximum air temperature and water temperature by month and *Region*. To construct *Region*, use *pd.cut* to satisfy the following conditions:

- Rows with a latitude lower than 42.4 should have *Southern* in the *Region* column
- Rows with a latitude between 42.4 and 42.8 should have *Central* in the *Region* column
- Rows with a latitude higher than 42.8 should have *Northern* in the *Region* column

You can have the month on the horizontal axis, the temperature on the vertical axis, different colors for different regions, and different styles (solid line / dotted line) to indicate air/water temperature.

Does anything in the visualization surprise you? Why or why not? *(4 points for code and visualization, 1 point for answer)*

1.2.3 Question 3

Finally let's focus on the stars of this survey—the fish, of course. *(19 points total)*

a) Let's continue using our *Region* categorization. Find the top three fish species in each region by number captured. *(3 points for code)*

b) Are certain fish only found in some regions? Visualize how many species are in all three regions, how many are in two of three, and how many were only captured in one region. *(3 points for code and visualization)*

c) What percentage of all species are exclusively captured in the Southern region? How about the Northern Region? And the Central region? *(3 points for code)*

d) Turbidity quantifies the level of cloudiness in liquid. For fish in each of the three regions, is there a correlative relationship between turbidity and # of fish caught? *(2 points for code, 1 point for answer)*

e) Now let's turn to the length of fish captured, given by *Maximum (mm)* and *Minimum (mm)*. Find the overall maximum and minimum lengths of all fish in each region. Which region has the largest range in captured fish length? *(2 points for code, 1 point for answer)*

f) Find the inverse Simpson index of species counts for each waterbody type (*WaterbodyType*) within each region. Which combination of waterbody type and region has the greatest diversity of fish species? Which has the least?

The inverse Simpson index ($\frac{1}{\lambda}$) is a measure of ecological diversity, for which a larger index number indicates a greater diversity of species. The index is calculated as:

$\frac{1}{\lambda} = 1/(\sum_{i=1}^{R} p_i^2)$

where R is the number of unique species and p_i is the proportion of fish belonging to species i. *(3 points for code, 1 point for answer)*