Assignment 2

4 Oct 2021

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.

c. Please include each question followed by code and your answer. 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.

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.

This assignment is due at 11:59pm on Monday, October 11th.

I. Download the *Iris* dataset from canvas.

1) There are three species of flowers in the dataset. The sepal length-width combination of one of the species is quite distinct from that of the other two species. Make an appropriate visualization plot to find out the distinct species.

2) Make a plot that shows the distribution of petal length for each of the three species. It should show median, 25^{th} and 75^{th} percentile values and outliers. Based on the plot, answer the following questions:
    a. Which species does not have any outlying values of petal length?
    b. Which species always has a larger petal length than setosa?
    c. Which species seems to have the lowest variance of petal length?

3) Let us call Sepal length, Sepal width, Petal length, Petal width as *iris descriptors*. We wish to visualize the distribution of *iris descriptors* for each of the three species. Create a 2x2 grid of four subplots – one for each *iris descriptor*. In each subplot, make the density plots of the *iris descriptor* with respect to each of the three species (i.e., three density plots in each subplot).
    a. For which *iris descriptor* do the densities of two of the species have the highest overlap? Also identify those two species.

b. Do the mean values of the *iris descriptor* corresponding to each of the two species, found in (a), differ significantly? Justify your answer with a T-test. Assume a confidence level of 5%.
c. If you are given iris flowers of each of the three species, which species will be the easiest to identify? Justify your answer.

4) Make a pairwise plot of all the columns in the dataset, where data points of each species have a distinct color, and the diagonal contains density plots. Based on the plots:
   a. Which *iris descriptor* seems to distinguish the three species the most?
   b. Which *iris descriptor* seems to distinguish the three species the least?
   c. Which species has the least correlation among its sepal and petal *descriptors*? Justify your answer.

II. Download the *Movies* dataset from canvas and drop all rows with at least one missing value. Assume you are a movie producer.

1) You would like to make a movie that gets a high IMDB rating. Make boxplots to compare the distribution of IMDB ratings of movies based on their genres. Based on the plot:
   a. Mention a reason that you would choose to make a 'documentary'
   b. Mention a reason that you would choose to make an 'Action/Adventure' movie
   c. For which genres, at least 75% movies have an IMDB rating greater than 6?

2)
   a. Compute the ROI for each movie (as defined in assignment 1). Identify if there any outliers with respect to ROI. Remove the most extreme outlier.
   b. Draw a barplot showing the mean ROI (ROI is as defined in assignment 1) and its 95% confidence interval, for different genre types. Based on the plot, answer the questions from (c) to (e):
   c. You are a risk-taker. You would like to choose a genre that may lead to the highest possible ROI. Which genre will you choose? Justify your answer.
   d. You like to play it safe. You would like to have a high ROI, while simultaneously minimizing the risk of getting a low ROI? Which genre will you choose? Justify your answer.
   e. You would like to discard the genre that has the possibility of incurring a loss. Which genre will you discard? Justify your answer.
   f. Split the barplot created in part (b) by the 'Creative type' column to create two barplots side-by-side in a single plot – one showing mean ROI based on genre for non-fiction movies, and the other showing the same for fiction movies
   g. Which creative type / genre combination(s) may have negative mean ROI, based on the plot in (f)?

3)
   a. Compute profit for each movie as the difference between worldwide gross and production budget.

b. Identify any four variables that are positively correlated with profit. Draw the scatter plots along with the trendline showing the positive correlation for each of the four variables in a 2 x 2 grid.

4)

a. Scale the rotten tomatoes ratings appropriately to compare it with the IMDB ratings.
b. Draw an appropriate plot comparing the distributions of IMDB ratings and the scaled rotten tomatoes rating. Which rating seems to distinguish more among the movies?
c. The distribution for which rating is bi-modal? What is the approximate rating at the modes?

5)

a. Draw a histogram and kernel density estimate of the IMDB ratings in the same plot.
b. Do majority of the movies have a higher than mean IMDB rating? Justify your answer based on the plot.

III. Suppose you record your biking/running activities on Strava (exercising app). You download your data from the app, and put it in a tabular form as follows:

Day	Activity	Distance (miles)	Average speed (mph)	Max heart rate
1	Running	5	6	140
2	Biking	15	16	120
3	Running	3	7	155
4	Running	6	5	145

Choose the appropriate visualization from Line chart / Histogram / Bar chart / Scatter plot / boxplot to answer the following questions:

1) Is there an association between distance and speed?
2) How does the distance change over a year?
3) Are there more short runs than long runs?
4) How does the distribution of max heart rate vary between running and biking?
5) How does the average running distance compare with the average biking distance?

IV. American Gourds

17 points total

For this part of the assignment, please load and use the data set "GourdData.csv" from Files on Canvas. This data was collected from the US Department of Agriculture's Quick Stats tool (https://quickstats.nass.usda.gov/). Each row includes results from a survey on a given state's gourd production in a given year.

Rows with "NaN" in certain columns indicate that we don't have complete information on a state's crop in a certain year but be sure to keep all rows when you're working with this data set.

Be sure to format all your visualizations with appropriate titles and axis labels.

Question 1

(4 points for code incl. visualization, 1 point for answer)

Find the top five pumpkin-producing states in the year 2020. Say we're interested in those five states' pumpkin production from 2016-2020.

Assuming that the amount of pumpkin per acre harvested is constant, produce an appropriate seaborn visualization to meet this goal. Briefly provide at least two observations about your visualization and the US pumpkin market.

Question 2

(3 points for code incl. visualization, 1 point for answer)

Suppose we're considering Pacific states California, Oregon, and Washington. For each state, we want to know the mean number of acres of pumpkins planted across 2000-2020. We're also interested in looking at each state's median number of acres of pumpkins planted in the same time horizon.

Produce one appropriate visualization to meet both goals. Briefly provide at least one observation about this visualization and the US pumpkin market.

Question 3

(3 points for code incl. visualization, 1 point for answer)

Consider the acres of pumpkins planted and acres of pumpkins harvested in Illinois from 2000-2020. Are there any trends in the number of acres planted and harvested over time? Produce a visualization to answer this question, along with a short written answer.

Hint: If you're having trouble with label sizes, before producing a visualization play around with running: sns.set(rc = {'figure.figsize': (30, 30), 'axes.labelsize': 15})

Question 4

(3 points for code incl. visualization, 1 point for answer)

Suppose we're curious about the relationship between acres of squash planted, acres of squash harvested, and price received in $/cwd (the price per 100lbs of squash) across all states and years.

Produce a single visualization to demonstrate the relationship between the three variables. Briefly describe what you learn from your visualization. Does there seem to be a relationship between price and acres of pumpkins harvested?

Hint: if you're having trouble with title placement, look up and experiment with the suptitle() function.

V. Social Indicator Data

1. Download the data set "social_indicator.csv" from the modules section in Canvas.
   a. Read it with python. Set the first column as the index when reading csv. (1 point for code)
   b. How many observations are there in the dataset? (1 point for answer)
   c. Sort the row index into ascending order. (1 point for code)
   d. Sort the column index lexicographically. (1 point for code)
   e. Make a column 'West' with True and False values indicating whether the region is west. Then delete the original region column. (1 point for code)
   f. We are not interested in studying fertility rate ('totalfertitlyrate') and contraception ('contraception'). Drop these two columns. (1 point for code)
   g. Remove the rows with missing values in the dataset. How many observations are there in the dataset now? Use the new data frame to answer the following questions. (1 point for answer)

2. Find the summary statistics of the data. (1 point for code)
   Based on the statistics, answer the following questions. Justify your answer briefly: (1 point for each answer; 1 point for justification)
   a. What is the maximum infant mortality among the set of 25% countries with the least infant mortality rate?
   b. Do the majority of the countries have GDP per capita lower than the average.
   c. Is the following claim true? *Across the countries, there are about 36.4 deaths per 1,000 individuals per year.* The infant mortality in this dataset is expressed in units of deaths per 1,000 infants.

3. We are interested in studying life expectancy and want to find the countries with high life expectancy for both males and females. We define having a high life expectancy to be being in the top 25% for its corresponding gender. (the top 25th percentile included)
   a. Make a subset of the data frame that only contains the countries with high life expectancy for both males and females. (1 point for code)
   b. Sort the data frame from (a) by GDP per capita. List the top three countries. (1 point for code; 1 point for answer)
   c. Make another subset of the data frame that only contains the countries with high life expectancy for males but not females. List the names of the countries. (1 point for code; 1 point for answer)

4. Use function application method to find i) interquartile range ii) Max-Min for all the numeric columns. (2 points for code)

Answer the following questions.

For which gender:

a. the range of economic activity is larger? (1 point for answer)
b. the interquartile range of illiteracy is smaller? (1 point for answer)

5.

a. Among economic activity, education, and illiteracy, which variable has the largest correlation between male and female? (1 point for code; 1 point for answer)
b. Find the correlation between GDP per capita and infant mortality rate. Are they positively or negatively correlated? (1 point for code; 1 point for answer)
c. Find the pair of most correlated variables in the data. (1 point for code; 1 point for answer)