</talentlabs>

Assignment 8 - Bivariate Analysis (21 points)

Instructions

1. Answer the below question in the boxes if needed.
2. For coding exercises, code in a single google colab notebook and zip all your code before submission.
3. Please submit the assignment through TalentLabs Learning System

Question 1 (1 point)
What do you understand by Bivariate Analysis?

Bivariate analysis involves the analysis of two variables for the purpose of determining the empirical relationship between them. It is used to test hypotheses of association or relationships between two variables and can be represented in a two-column data table.

Question 2 (2 points)
What are the differences between correlation and causation?

Correlation: A statistical measure that determines the strength of the relationship between two variables.

Causation: It implies that one event is the result of the occurrence of the other event; i.e., there is a cause-and-effect relationship between the two variables

Question 3 (1 point)
Which of the following correlation coefficients indicates the strongest relationship between variables?

- 0.2
- 0.01
- 0.8
- -0.1
- -0.9

- -0.9

</talentlabs>

Question 4 (1 point)
A national study on cell phone use found the following correlations

- The correlation between the number of texts sent each day and a person's average credit card debt is 0.35
- The correlation between the number of texts sent each day and the number of books read each month is −0.20

Which of the following statements are true?

1. As the number of texts sent each day increases, average credit card debt increases.
2. Sending more texts causes people to read less.
3. A person's average credit card debt is related more strongly to the number of texts sent each day than the number of books read each month is related to the number of texts sent each day.

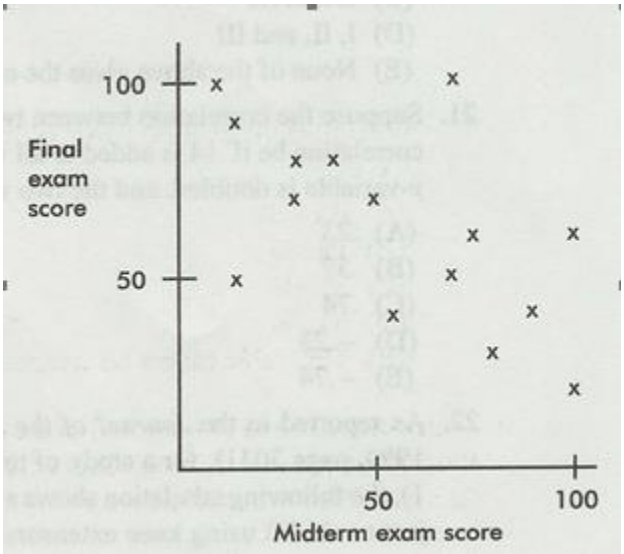
Possible Answers:

- 1
- 3
- 2
- 1 and 2 and 3
- 2 and 3
- 1 and 3
- 1 and 2

- 1 and 3

</talentlabs>

Question 5 (1 point)
Consider the following scatterplot of midterm and final exam scores for a class of 15 students. (2 point)



Which of the following are true statements?

I. The same number of students scored 100 on the midterm exam as scored 100 on the final exam.
II. Students who scored higher on the midterm exam tended to score higher on the final exam.
III. The scatterplot shows a moderate negative correlation between midterm and final exam scores.

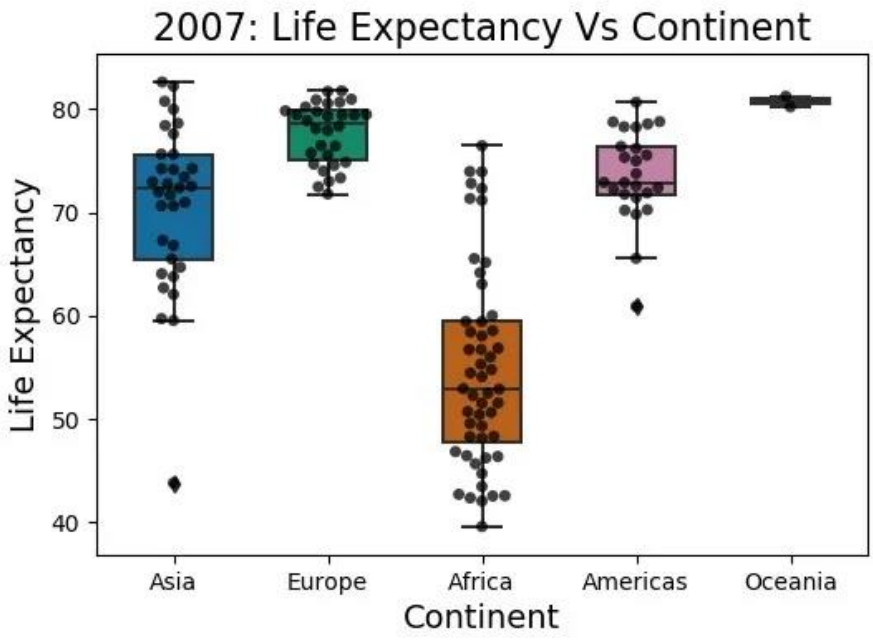
A. I and II [/]
B. I and III
C. II and III
D. I, II, and III
E. None of the above gives the complete set of true responses.

</talentlabs>

Question 6 (2.5 points)

The following data shows Life Expectancy in each continent in the year 2007.

What type of plot(s) is this (hint: there are two plots overlaid)? Give any two insights that you derive from the chart.



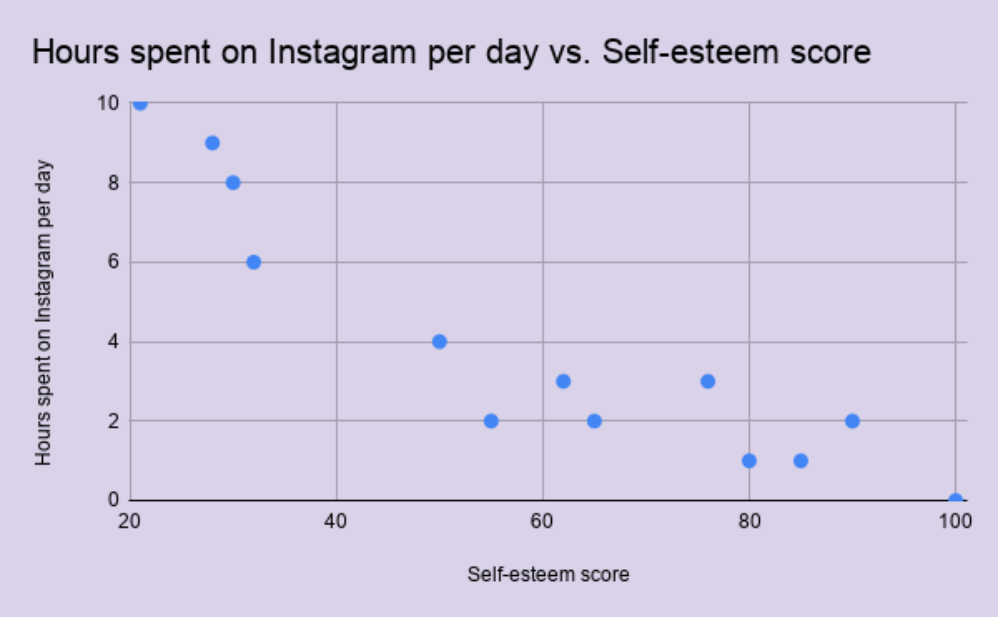
- Box plot overlaid with a strip plot

1. Africa has the lowest median of life expectancy.
2. The highest life expectancy here is from Asia at 85 years of age.

</talentlabs>

Question 7 (2.5 points)
The following data shows the number of hours spent on instagram per day vs self esteem score.

What type of a plot is this? Give any two insights that you derive from the chart.



- Scatter Plot

1. There seems to be a negative correlation between hours spent on Instagram and self-esteem scores, suggesting that increased time on Instagram is associated with lower self-esteem.
2. There is significant variability in self-esteem scores at lower usage levels, indicating that other factors may also influence self-esteem

</talentlabs>

Question 8 (10 points)
Note: Submit the code in a jupyter notebook or Google Colab with your assignment.

Load the titanic dataset using seaborn using and answer the questions below

```
import seaborn as sns
df = sns.load_dataset('titanic');
```

Study the dataset and the goal here: https://www.kaggle.com/competitions/titanic
You can use seaborn or matplotlib or both.

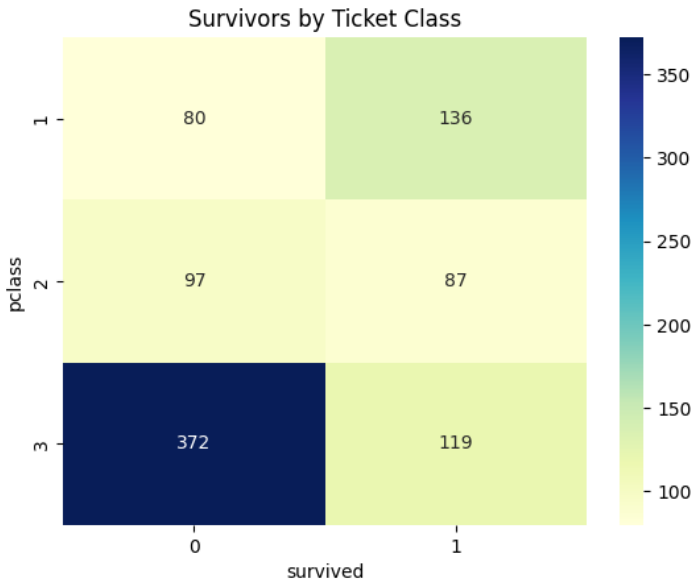
Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

</talentlabs>

- Using cross tabulations and heatmaps - find which ticket class had the most survivors. (2 points)

Screenshot of the chart:

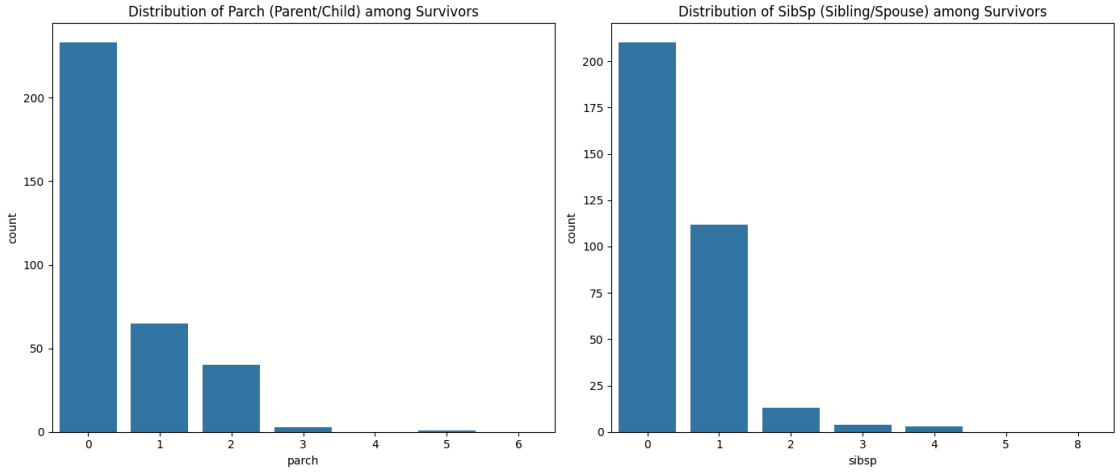


Ticket class with most survivors: 1

</talentlabs>

- Convert parch and sibsp variables to category. Out of those who **survived what percentage of samples had 1 parent/child**, and **what percentage of survivors had 1 sibling/spouse**? Round to percentage to 2 decimal places (2 points)

Screenshot of the charts:

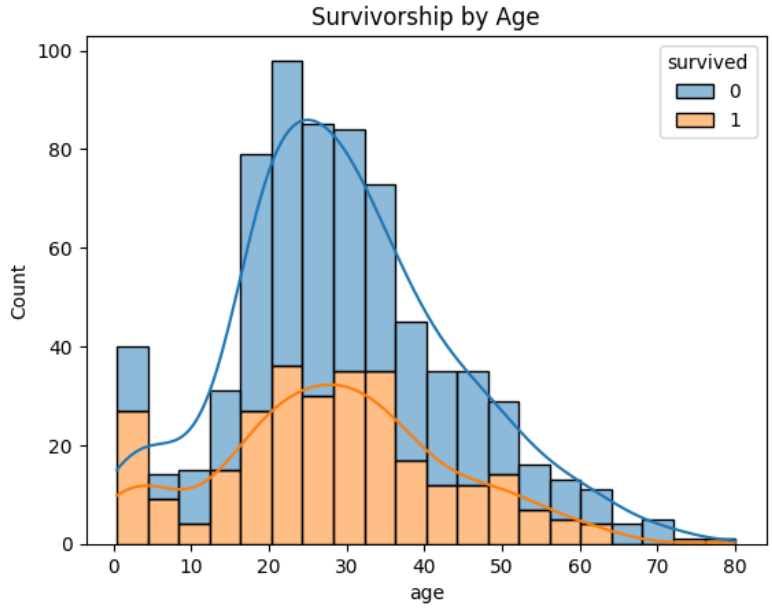


Out of the survived:
Percentage of samples that had 1 parent/child: 19.01%
Percentage of samples that had 1 sibling/spouse: 32.75%

</talentlabs>

- Does Age determine Survivorship? Plot and write your interpretation. (2 points)

Screenshot of the chart:

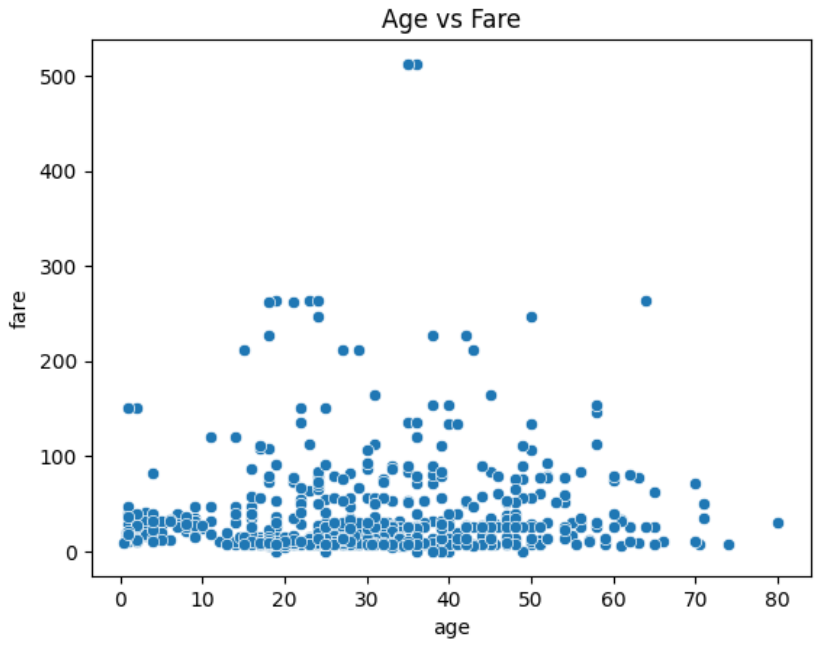


Interpretation:
Pearson correlation between Age and Survival: -0.08
Answer: No

</talentlabs>

- Is there a relation between Age and Fare? Find the Pearson correlation coefficient. Plot using a scatter plot and write your Interpretation. (2 points)

Screenshot of the charts:



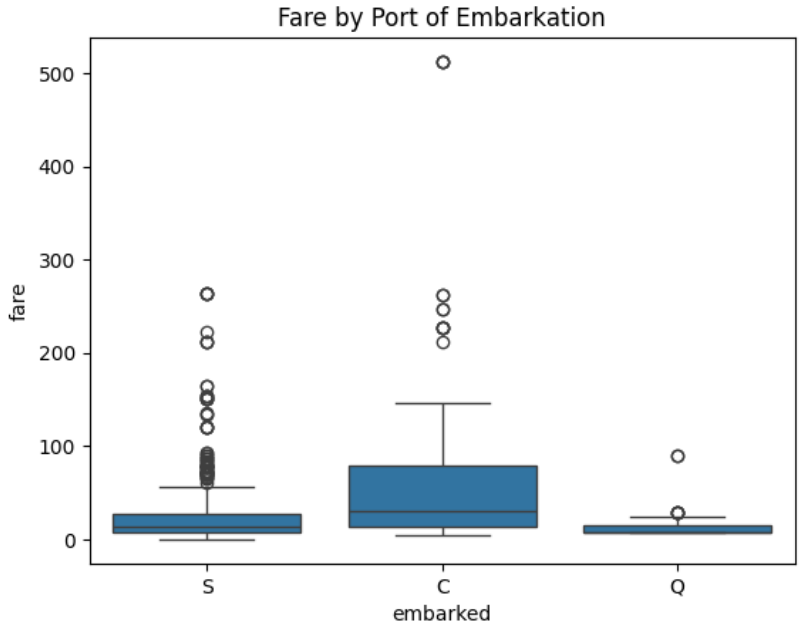
Interpretation between Age and Fare: Very weak positive correlation, so the answer is no.
Correlation Coefficient: 0.10

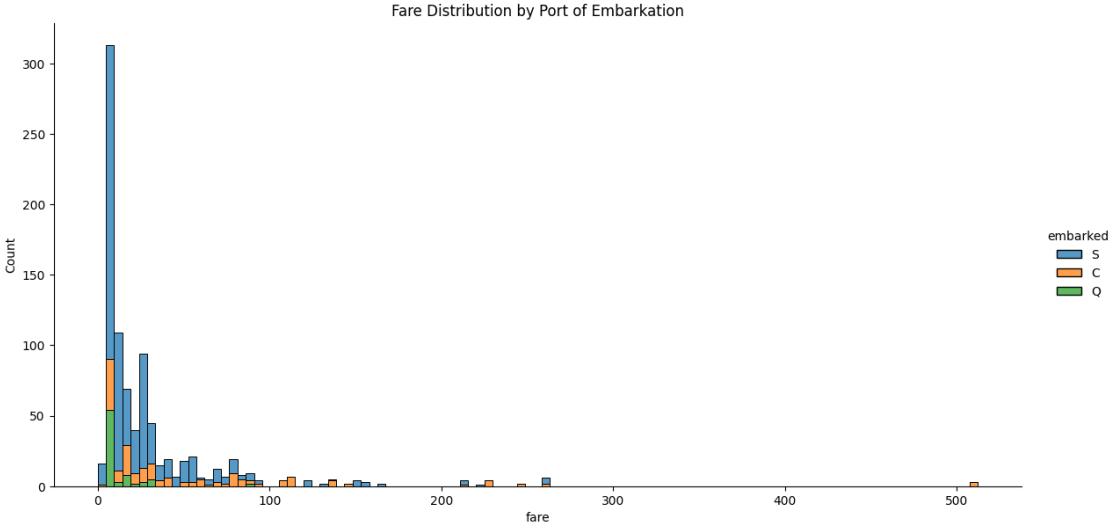
- Based on the port of embarkation,do you see any difference in median fares?

  Plot a box plot and a distribution plot (hint: use port as color here for distribution plot and the `sns.displot` function) showing the different distributions of fare for each port of embarkation?

  In the distribution plot where are the people who paid more than 500 dollars in fare from? For `sns.displot` use `multiple='stack',height=10` and `aspect=2`. (2 points)

Screenshot of the chart:





</talentlabs>

Insights:

1. Box Plot for Fare by Port of Embarkation:

- **Median Fare:** The median fare is highest for passengers who embarked from Cherbourg (C), followed by Southampton (S), and is lowest for those who embarked from Queenstown (Q).
- **Fare Distribution:**
  - Passengers who embarked from Cherbourg (C) show a wider range of fare prices, indicating that both high and low fares were common among these passengers.
  - Passengers who embarked from Southampton (S) generally paid lower fares, but there are outliers who paid higher fares.
  - Passengers from Queenstown (Q) mostly paid lower fares, with fewer outliers indicating that higher fare purchases were rare among them.
- **Outliers:** The box plot highlights several outliers, particularly in the Cherbourg (C) group, where some passengers paid significantly higher fares.

2. Distribution Plot for Fare by Port of Embarkation:

- **Concentration of Fares:**
  - The majority of passengers paid fares below 100, with the highest concentration of passengers embarking from Southampton (S).
  - Cherbourg (C) also shows a noticeable number of passengers in the higher fare categories (100-200 and even above), which aligns with the outliers observed in the box plot.
  - Queenstown (Q) passengers are primarily concentrated in the lower fare categories.
- **High Fare Observations:**
  - The distribution plot indicates that passengers who paid over $500 in fares almost exclusively embarked from Cherbourg (C). This suggests that Cherbourg was likely the port of departure for wealthier passengers, or that it had more first-class passengers.

General Insights:

- **Economic Class Differences:** The data suggests that Cherbourg was a port of departure for more affluent passengers, possibly due to it having more first-class accommodations or because it was a departure point for wealthier individuals. Southampton had a mix, but with many passengers paying lower fares, likely representing third-class passengers. Queenstown appears to have been a departure point mainly for lower fare-paying passengers.

People who paid 500 dollars and more are from? : Cherbourg (‘C’)