</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 is the analysis of two variables (often denoted as X and Y), to determine empirical relationship between them.

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

Correlation is the measure of the strength of the relationship between two variables.

Causation is concerned with how change in one variable might affect the change in another variable.

**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:

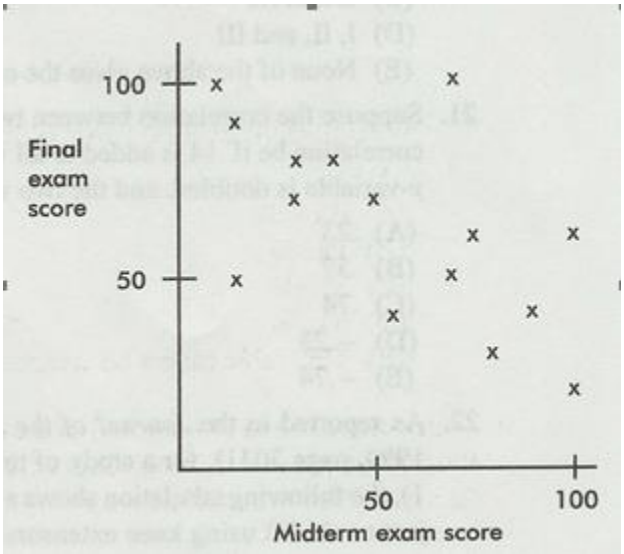
- 1
- 3
- 2
- 1 and 2 and 3
- 2 and 3
- 1 and 3
- 1 and 2

Statement 1 is True.

However, they exhibit a weak positive correlation. So, as the number of texts sent each day increases, the average credit debt increases slowly.

</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.

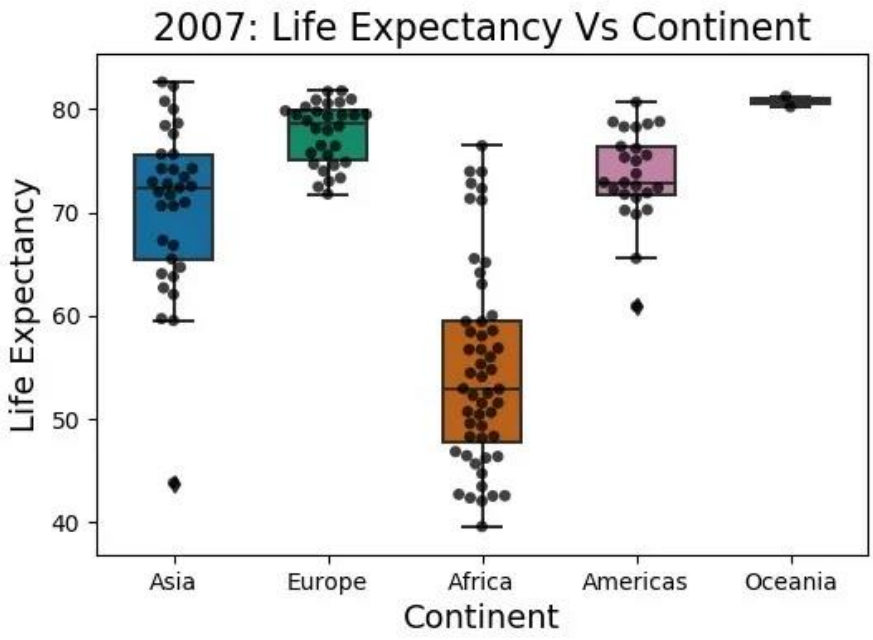
A. I and II
B. I and III
C. II and III
D. I, II, and III
E. None of the above gives the complete set of true responses.

</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.



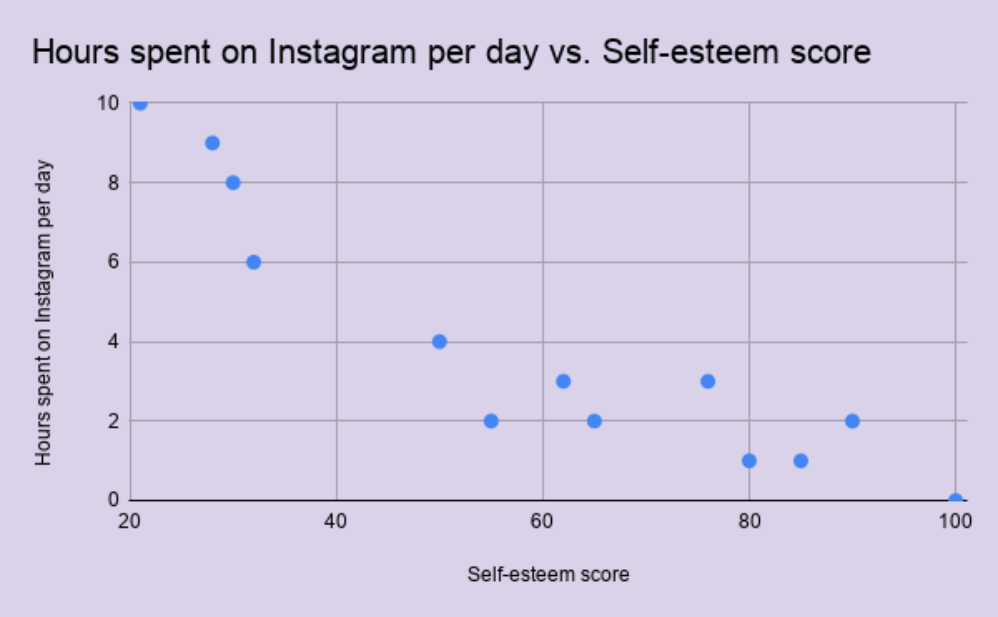

It is a bivariate box plot, which plots life expectancy for each continent category.

<u>Insights:</u>
Ocenia has the highest life expectancy mean at more than 80 years old.
Africa has the lowest life expectancy mean at around 50 years old.

</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.

Insights:

- There is a moderate negative correlation between hours spent on Instagram per day and self-esteem score.
- With lesser hours spent on Instagram per day, the higher the self-esteem score.

</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');
```

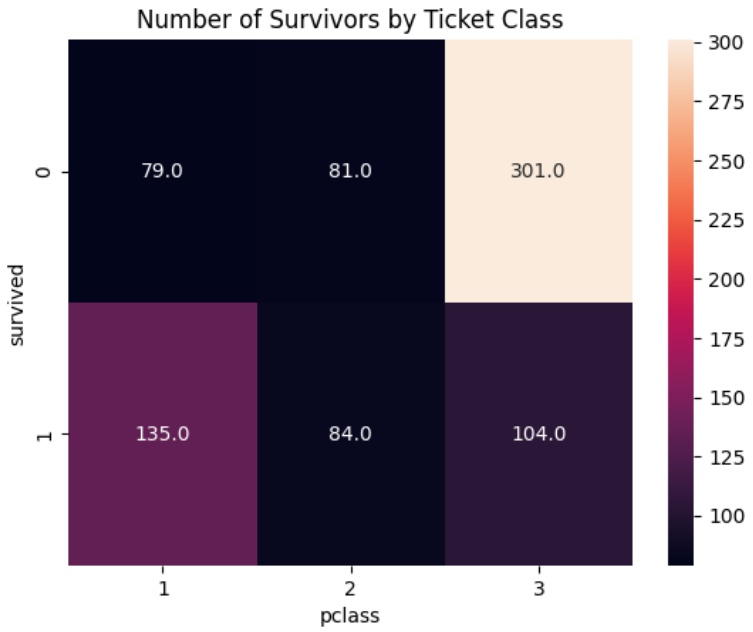
Study the dataset and the goal here: https://www.kaggle.com/competitions/titanic
You can use seaborn or matplotlib or both.

## 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 |

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

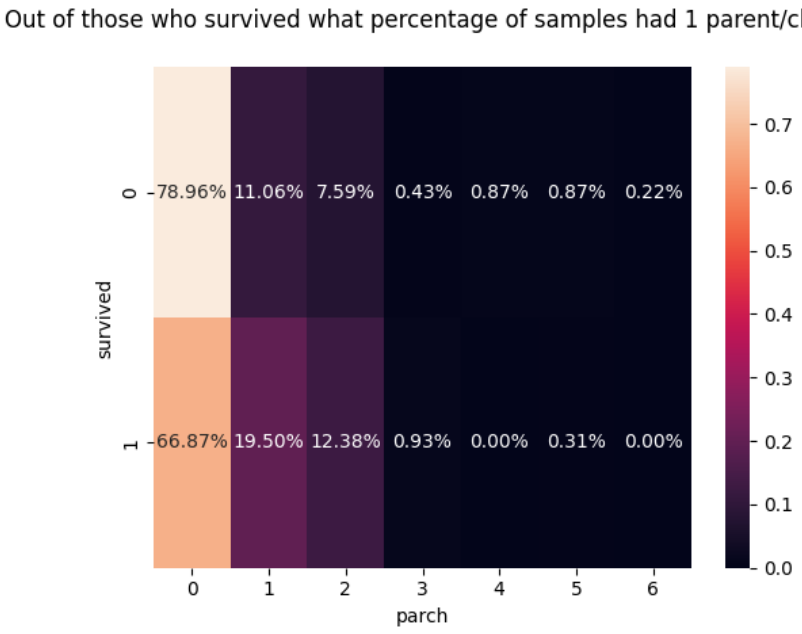
| Screenshot of the chart: |
|---|




Ticket class with most survivors: Ticket class 1

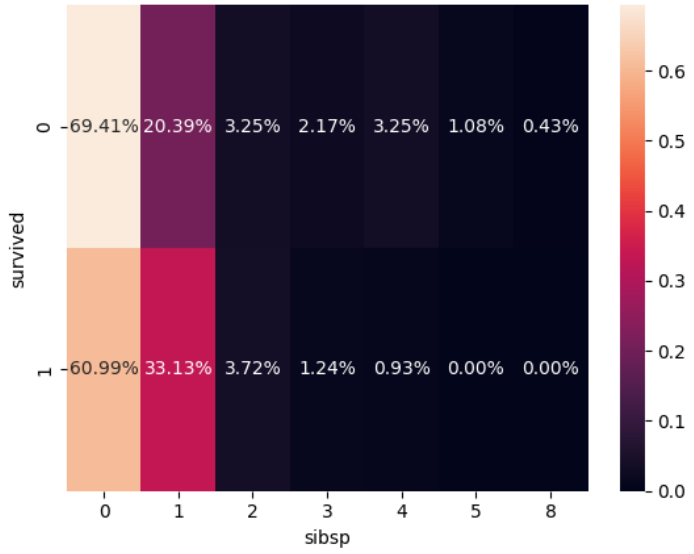
- Convert parch and sibsp variables to category. Out of those who **survived what percentage of samples had 1 parent/child**, and **what percentage of survivors had 1 sibling/spouse**? Round to percentage to 2 decimal places (2 points)

Screenshot of the charts:



</talentlabs>

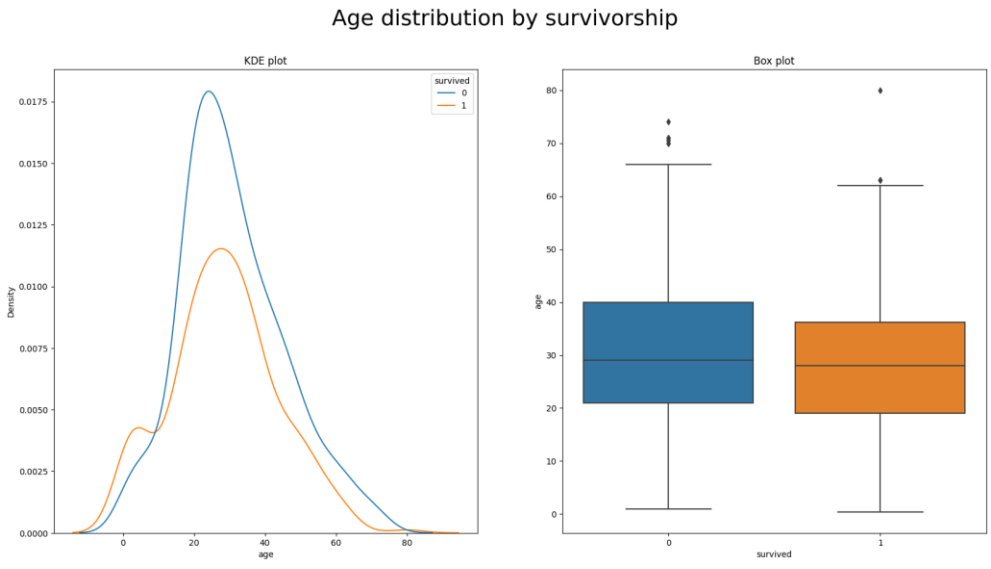
Out of those who survived what percentage of samples had 1 sibling/spouse




<u>Out of the survived:</u>
Percentage of samples that had 1 parent/child: 19.50%
Percentage of samples that had 1 sibling/spouse: 33.13%

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

Screenshot of the chart:




Interpretation:

To check whether Age is a determinant of survivorship, given that we were unable to establish statistically significant differences in the mean ages between survivors and non-survivors from the visual above, we conducted a t-test experiment.

Problem statement: Is there a statistically significant relationship between age and survivorship? Specifically, does higher age correspond to a higher likelihood of survivorship, and vice versa?

Hypothesis:
H0: There is no statistically significant difference in the mean age between survivors and non-survivors, i.e., $\mu 1 = \mu 2$

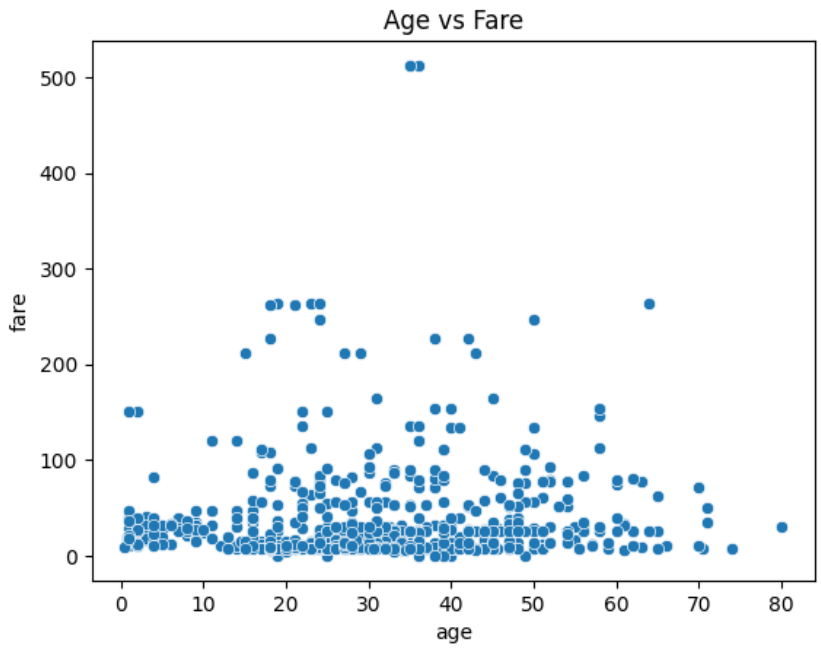
H1: There is statistically significant difference in the mean age between survivors and non-survivors, $\mu 1 \neq \mu 2$

Result:
p-value = 2.42% (less than 95% confidence interval)

We reject the null hypothesis, and it is evident that the age is the determinant of the survivorship.

- 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:



</talentlabs>

Interpretation between Age and Fare: There is weak positive correlation between age variable and fare variable.

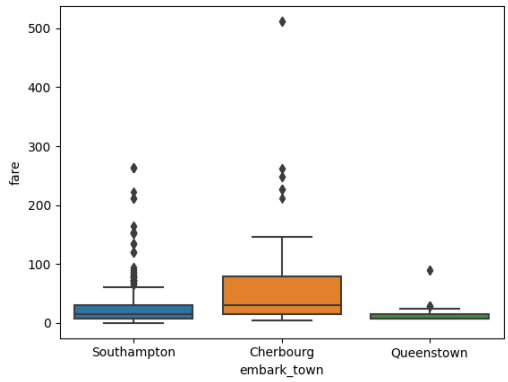
Correlation Coefficient: 0.092707

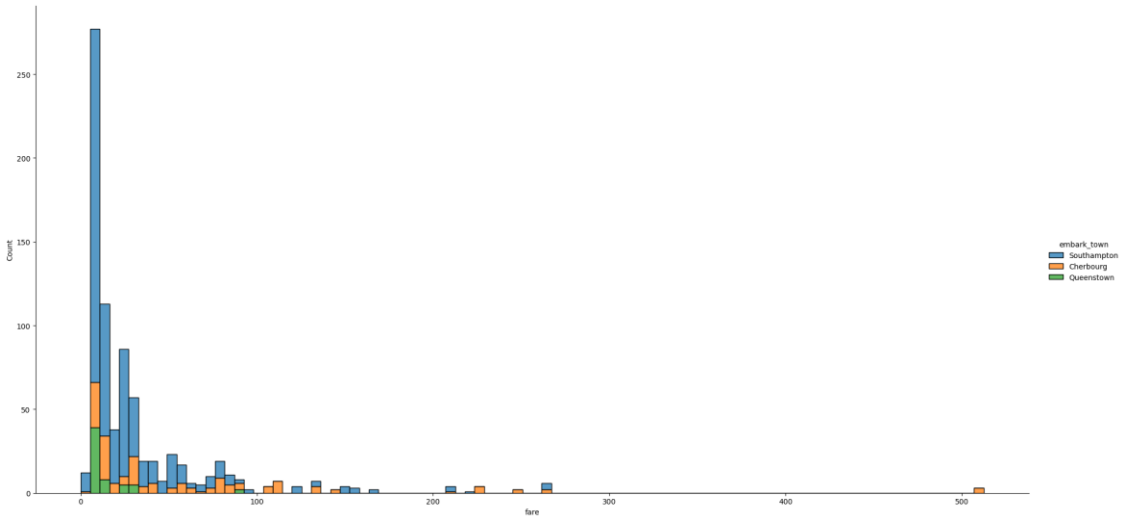
- Based on the port of embarkation, do you see any difference in median fares?

  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: Yes, there is obvious difference in median fares across 3 embark towns, where Cherbourg town has a highest median fares amongst the three.

People who paid 500 dollars and more are from? : **Cherbourg**