</talentlabs>

Assignment 7 - Univariate Analysis (23.5 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 (3.5 points)
Interpret and write your insights for the following plots, you are given data description for your reference and the goal.

Our data for this question, adult overweight/obesity rates by state, come from the Kaiser Family Foundation. The data file, a comma-separated values file called adult_data.csv is provided to you.

Data contains the name of the state; then all_adults, male, and female, which contain the overweight/obesity rates for that state overall and by gender. Snippet below.

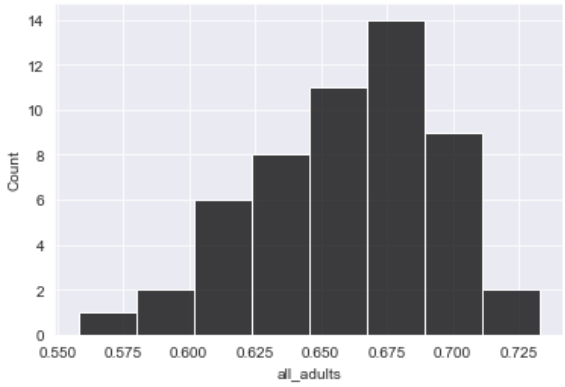
	state	all_adults	male	female
0	Alabama	0.697	0.721	0.674
1	Alaska	0.642	0.703	0.568
2	Arizona	0.647	0.710	0.583
3	Arkansas	0.705	0.721	0.688
4	California	0.622	0.681	0.560

Now given this data, answer the following:

</talentlabs>

```
sns.histplot(data = df, x = "all_adults", color = "black")
```

```
<AxesSubplot:xlabel='all_adults', ylabel='Count'>
```



1. What kind of variable is "all_adults"? (Numerical: discrete or continuous, Categorical: ordinal or nominal) (1 point)

Numerical: continuous

2. What kind of plot/chart is this? (0.5 point)

Histogram chart

3. What does the plot show? (1 point)

It shows the frequency distribution of adult overweight/obesity rates by state, come from the Kaiser Family Foundation.

4. How many states are having obesity rates between 0.625 and 0.675 throughout the US according to the dataset roughly(see the plot)? (1 point)

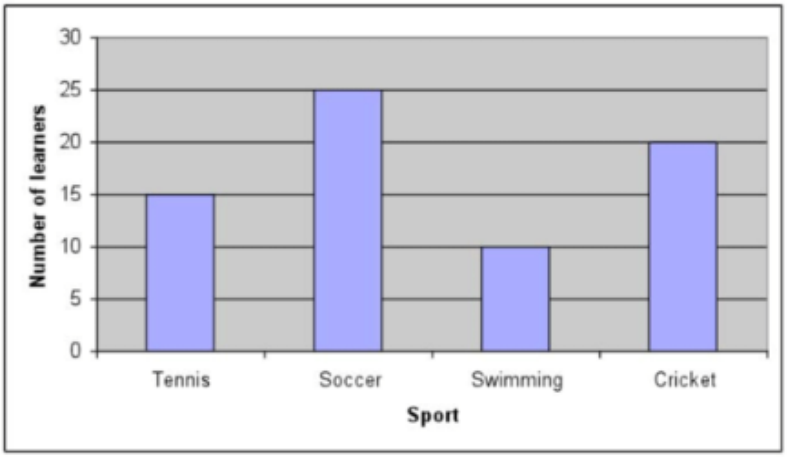
8 states

Question 2 (2 point)

Read the description below and answer the following questions:

</talentlabs>

Mr Khosa asked each learner in his class to name their favorite sport and he plotted it out below.



1. What kind of a plot/chart is this? What is the data type of variable being analyzed here (categorical or numerical)? (0.5 point)

Bar plot

There are 2 variables analyzed as below:

Variable	Data type
Sport	String - categorical
Number of learners	Numerical - continuous

2. How many learners are there in Mr Khosa's class? (0.5 point)

Assuming each learner has only 1 class, there are 70 learners in total.

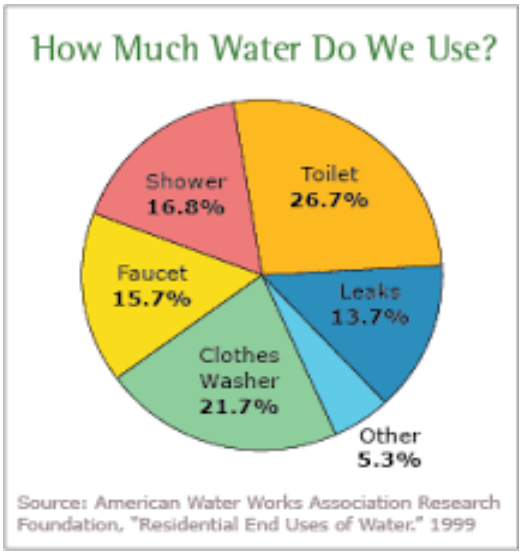
3. Which is the most popular sport and the least popular sport here? How many learners are in each?(1 point)

Most popular - Soccer
Least popular - Swimming

</talentlabs>

Question 3 (2 points)

Distribution of water usage from different water sources. Study and answer questions below(2 point):



1. What kind of plot/chart is this? What is the data type of variable being analyzed here (categorical or numerical)? (0.5 point)

Categorical

2. Which source of water utilizes the most water vs least water? What is the percentage of water used in each? (1 point)

The most - Clothes Washer (21.7%) The least – Leaks (13.7%), if exclude Others (5.3%)

3. Why are pie charts not used that often? What are the alternatives(0.5 point)

It is messy to compare too much categorical variables (e.g., >10) inside pie chart. Bar chart could be the alternative when the data has too much categorical variables to compare with.

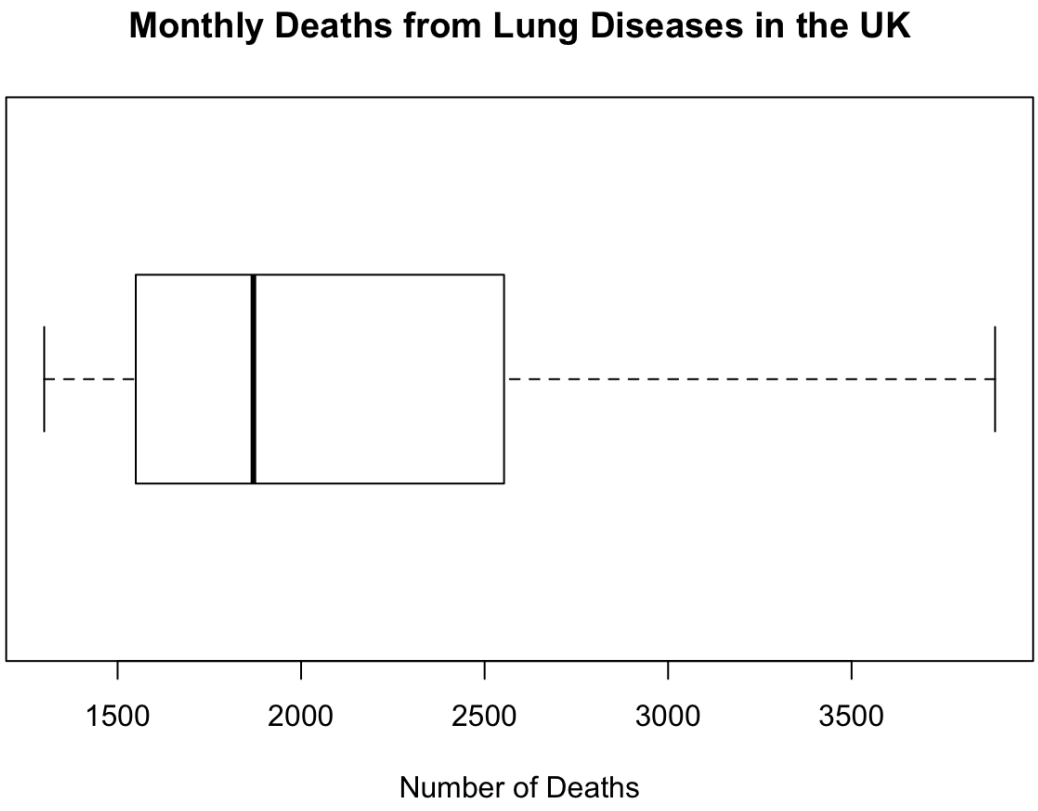
</talentlabs>

Question 4 (2 points)

1. Which of these is one reason to run a univariate analysis?
    a. To understand relationships between variables
    b. To understand and compare groups.
    c. To understand a variable and glean insights
    d. To predict one variable based on another

2. Univariate data cannot answer research questions about relationships between variables, but rather, it is used to describe:
    a. All related variables to the outcome
    b. The characteristics that vary randomly
    c. One characteristic or attribute that varies from observation to observation
    d. Relationships among variables

Question 5 (4 points)
With the plot given below, answer the following questions:



- What type of plot do we see? (0.5 point)

Box plot

- What are the five summary statistics values here roughly? (2.5 point)

Median
Q1
Q3
Lower boundary (Q1 – 1.5 * IQR)
Upper boundary (Q3 + 1.5 * IQR)

- Give the Interquartile Range and does the plot above have outliers?(1 point)

Interquartile Range = Q3 – Q1
= 2,600 – 1,600
= 1,000

The plot above doesn't have outliers, as no data points beyond upper and lower boundary of box plot.

</talentlabs>

Question 6 (10 points)
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 plotly or all of them.

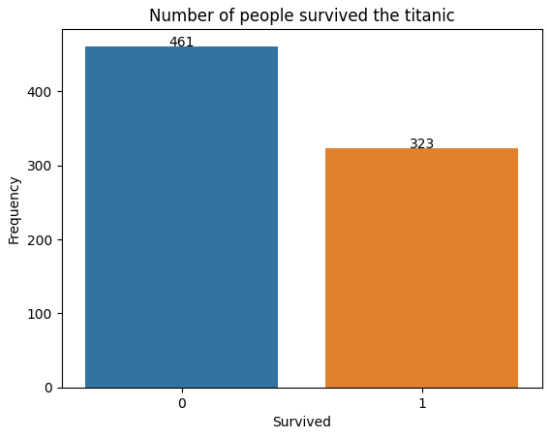
Data Dictionary

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

- How many people survived the titanic? Plot a graph. (2 points)

Screenshot of the chart:

</talentlabs>

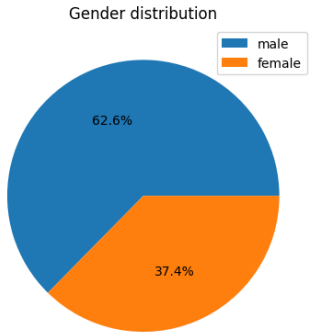


Number of people who survived: 323

</talentlabs>

- What was the ratio of Males to Females on the titanic? Plot a graph (2 points)

Screenshot of the chart:

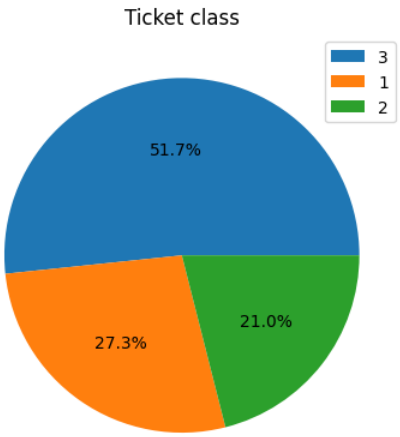


<u>Male - Female Ratio:</u>

Male: 62.6%
Female: 37.4%

- Make a pie chart of the number of people from different ticket classes. Make sure you **deduplicate the data as a whole and remove any null values in the class variable and generate the chart.** Use the 'class' variable. (1 point)

Screenshot of the chart:

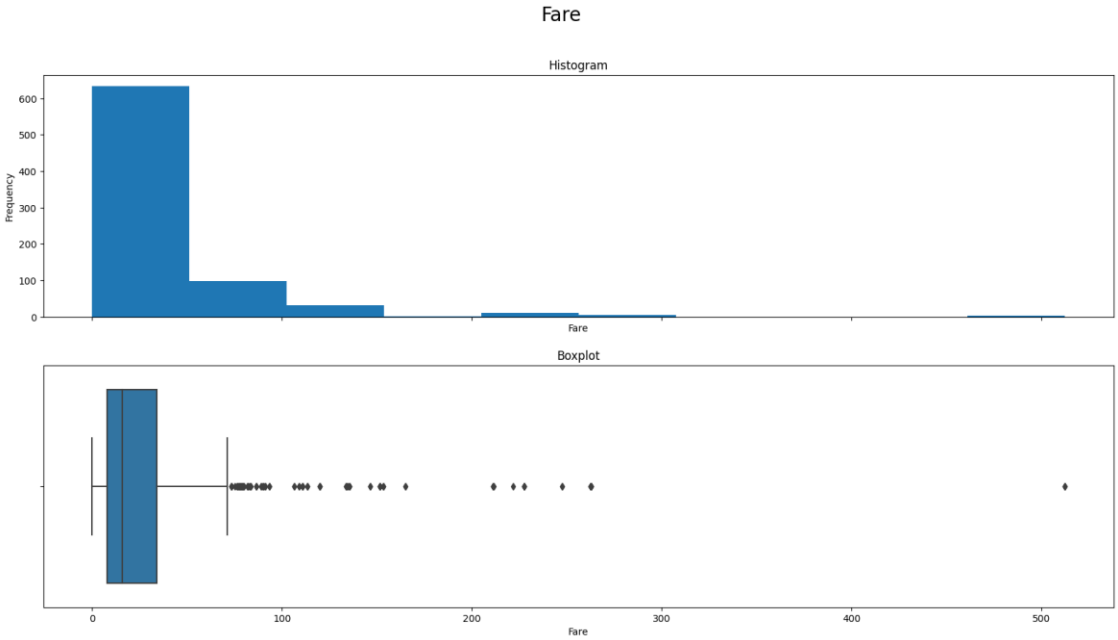


- What is the distribution of the Fare paid by people? Plot a histogram and tell us what you see? How many people paid between 0 and 100 dollars, while how many paid more than 300 dollars? Also make a boxplot? Make sure they align in terms of the x axis. Give us the interquartile range by using quantile function in pandas. (5 points)

Hint: Use plt.subplots with sharex=True and figsize=(20,10)

</talentlabs>

Screenshot of the charts:



Insight: The distribution of fare is right skew, and a lot of outliers beyond its upper boundary.

People who paid Fare between 0 and 100: 721

People who paid Fare above 300: 3

IQR: 26.05935