TOBB ETU – 2023 SPRING
YAP 101 – INTRODUCTION TO DATA SCIENCE
MIDTERM

1. (13 pts) Write the output of the codes given below. Assume that the following codes are run first.

```
import numpy as np
import pandas as pd
```

a. (2 pts)
```
np.arange(5) * 10 + np.ones(5)
```

b. (2 pts)
```
np.arange(2, 20, 3)
```

c. (2 pts)
```
my_array = np.arange(5)
my_array == 2
```

d. (2 pts)
```
my_data = pd.DataFrame({'Name':['a', 'b', 'c', ‘d’, ‘e’, ‘f’], 'Value':[0,10,20,30,40,50]})
my_data.iloc[1:4][[‘Value’]]
```

e. (2 pts)
```
my_data = pd.DataFrame({'Name':['a', 'b', 'c', ‘d’, ‘e’, ‘f’], 'Value':[0,10,20,30,40,50]})
my_data[‘Value2’] = my_data['Value'].apply(lambda x: x if x≤25 else 25)
my_data
```

f. (3 pts)
```
my_data1 = pd.DataFrame({'Name1':['a', 'b', 'c', ‘d’], 'Value1':[1,2,3,4]})
my_data2 = pd.DataFrame({'Name2':['a', 'b', ‘e’, ‘f’], 'Value2':[10,20,30,40]})
pd.merge(my_data1, my_data2, left_on=[‘Name1’], right_on=[‘Name2’], how="left")
```

2. (13 pts) Let ‘A’ be m-by-n numpy array which contains positive integers. Assume you are given the matrix A. Write python code to answer the questions below. You are **not** allowed to use any loops.

a. (2 pts) What is the maximum of the 5^{th} column of A?

b. (2 pts) What is the elements of the 4^{th} row of A which correspond to odd columns of A.

c. (3 pts) Let ‘B’ be the last 12 columns of A. What is the resulting numpy array when the mean of B is subtracted from each element of B.

d. (3 pts) Let ‘B’ be the rows of A between 10 and 20 (included). What is the number of 3’s in B?

e. (3 pts) What is the largest 15 values in the 6^{th} column of A?

3. (18 pts) The table ‘nba’ contains data for the 2016-2017 NBA Season. All numerical values in the table are integers. All other values are strings. The ‘nba’ table contains 8 columns. The first few rows are shown below. Note that ‘each’ prefix is an abbreviation for a team.

player	prefix	position	age	salary	games	minutes	points
Al Horford	BOS	C	30	2.65401e+07	68	2193	952
Amir Johnson	BOS	PF	29	1.2e+07	80	1608	520
Avery Bradley	BOS	SG	26	8.26966e+06	55	1835	894
Demetrius Jackson	BOS	PG	22	1.45e+06	5	17	10
Gerald Green	BOS	SF	31	1.4106e+06	47	538	262
Isaiah Thomas	BOS	PG	27	6.58713e+06	76	2569	2199
Jae Crowder	BOS	SF	26	6.28641e+06	72	2335	999
James Young	BOS	SG	21	1.8252e+06	29	220	68
Jaylen Brown	BOS	SF	20	4.743e+06	78	1341	515
Jonas Jerebko	BOS	PF	29	5e+06	78	1232	299

Write python code to answer the questions below. You are **not** allowed to use any loops.

a. (2 pts) What is the number of players who played more than 60 games?

b. (3 pts) What is the average salary after all players get a 20% raise?

c. (3 pts) What is the age of the oldest player(s) among the players whose position is center (C)?

d. (4 pts) What is the name(s) of the oldest player(s)?

e. (3 pts) What is the average salary of players for each team.

f. (3 pts) What is the number of teams that have players between 12 and 15?

4. (6 pts) For this question, use the table ‘nba’ given in question 3 and write python code for the following tasks.

Create a new table ‘nba_new’ which consists of players whose positions are ‘PG’ or ‘C’.

Add a column named ‘salary2’ to table ‘nba_new’ whose values are the randomly shuffled values of the column 'salary'.

Calculate the absolute difference of average values of ‘salary2’ for each ‘position’.

5. (8 pts) Write python code for the following tasks.

Create a numpy array ‘die’ whose values are 1, 2, 3, 4, 5, 6.

Assume the probability of each number is equal. Display 3 random samples from ‘die’.

Create a pandas dataframe ‘die_df’ from the numpy array ‘die’.

Assume the probability of each number is equal. Take 100 samples from ‘die_df’ and calculate their mean.

Repeat the last process 10.000 times and store the means in an array ‘means’.

6. (8 pts) There are two urns. Urn A contains 3 white, 2 black balls. Urn B contains 2 white, 4 black balls. One ball is drawn at random from urn A and placed in urn B. Then, one ball is drawn at random from urn B.

a. (4 pts) Find the probability that the transferred ball is also drawn from urn B.

b. (4 pts) Find the probability that the drawn ball is black.

7. (12 pts) There are 2 fair dice. The first die has 4 red and 2 blue faces. The second die has 2 red and 4 blue faces. A fair coin is flipped. If the result is heads, the first die is chosen. If the result is tails, the second die is chosen.

a. (5 pts) Chosen die is thrown twice. Find the probability that both results are red.

b. (7 pts) Chosen die is thrown once. Find the probability that the first die is chosen given that the result is red.

8. (12 pts)

a. (6 pts) What is confounding factor? How do we avoid cofounding? Explain the process.

b. (6 pts) What is the Bootstrap? Why and how do we use it?

9. (11 pts)

a. (3 pts) Suppose we have discovered an association between two variables in a dataset. Which of the following would be the best way to test whether it is causal? You do not need to justify your answers.

i. Run a randomized controlled experiment.
ii. Use hypothesis testing to check whether the association is statistically significant.
iii. Brainstorm some potential confounding factors and test whether any of them has an association with both variables.

b. (3 pts) A data scientist performs a statistical test. The null hypothesis is that a specified chance model is good and the alternative hypothesis is that the model is not good. The data scientist decides to use 3% as the cutoff for the P-value of the test. State whether below statements are true or false. You do not need to justify your answers.

i. If the model is good, the chance that the test will conclude that the data are consistent with the model is 0.03.
ii. If the model is good, the chance that the test will conclude that the data are consistent with the model is 0.97.
iii. It is not possible to approximate based on the information given.

c. (5 pts) Each person in a random sample of 1000 U.S. adults was asked if they agreed with the statement, "News organizations are growing in influence." Among the sampled men, 39% agreed. Among the sampled women, 43% agreed. Data scientists have used an A/B test to see whether or not the observed difference is due to chance. Their null hypothesis is: "In the U.S., the percent of men who agree is the same as the percent of women who agree. The difference in the sample is due to chance."

The data scientists are using a 1% cutoff for the p-value of the test. They run the test and the p-value comes out to be 0.5%, that is, 1 in 200. State whether below statements are true or false. You do not need to justify your answers.

i. The data scientists will conclude that the data are consistent with the null hypothesis.
ii. There is only a 1 in 200 chance that the null hypothesis is true.
iii. There is a 199 in 200 chance that the alternative hypothesis is true.
iv. The data scientists will reject the null hypothesis.
v. The assumptions made in the null hypothesis are used in the calculation of the p-value.

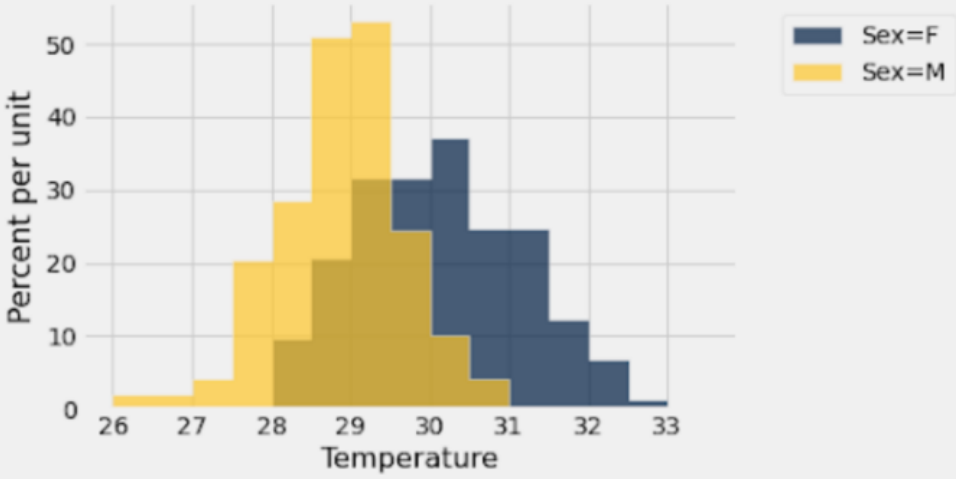
10. (8 pts) When hatching a baby turtle from an egg, we incubate the egg at some temperature. Ellen read that the temperature an egg is incubated at influences whether or not the turtle that hatches will be male or female.

Ellen loves turtles and is wondering whether this is really right, or whether differences might just be due to chance. She collects data on 100 randomly drawn turtles. She records the incubation temperature (in Celsius) and the sex of the turtle that hatches in the table turtles:

Temperature	Sex
30.8	M
31.5	F
32.4	F

(. . . 97 more rows)

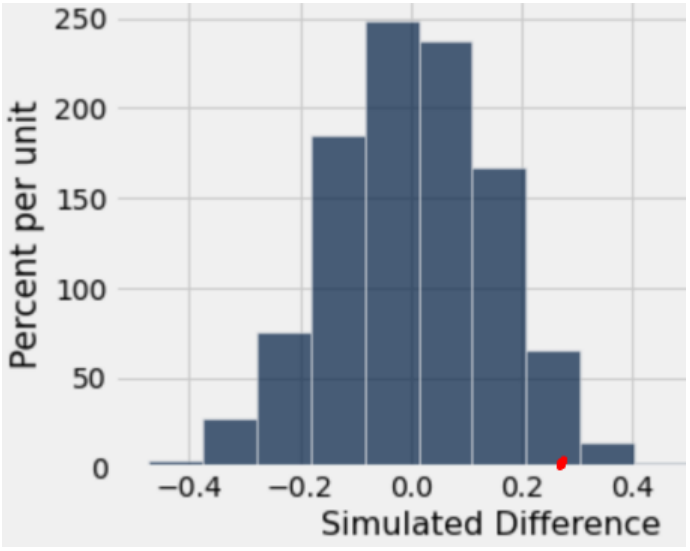
a. (3 pts) Ellen decides to visualize her data before doing any inference. She creates the following histograms, using the same bins for female and male turtles. All bars of the histograms are clearly visible.



State whether below statement is true or false. Why?

"Males and female turtles have different distributions of incubation temperatures."

b. (5 pts) Ellen performs an A/B test to see whether females in the population in general have higher incubation temperatures than the males, or if the observed difference in distributions is due to chance. Ellen's test statistic is the difference between average incubation temperatures, defined as "female average minus male average". She simulates the statistic 1000 times under the null hypothesis. The histogram below shows the 1000 simulated differences. The red dot shows the observed difference.



State whether below statements are true or false. You do not need to justify your answers.

i. Based on the test, a reasonable conclusion is that the average incubation temperature of females in the population is higher than the average for males in the population.
ii. Based on the test, Ellen cannot reasonably decide between her two hypotheses.
iii. Based on the test, a reasonable conclusion is that the difference observed in the sample is due to chance if p-value is 0.01.
iv. Based on the test, a reasonable conclusion is that the difference observed in the sample is due to chance if p-value is 0.04 and cut-off is 0.05.
v. Based on the test, a reasonable conclusion is that the average incubation temperature of females in the population is higher than the average for males in the population if p-value is 0.04 and cut-off is 0.05.

11. (7 pts) Shmuel has collected some data about cells that have been infected with the ILLNESS virus and cells that have not been infected. For each cell, he has collected data about the number of nuclei in the cell. In some cases, ILLNESS may cause some cells to form multiple nuclei. Shmuel would like to determine if the number of nuclei in a cell is higher in cells that have been infected with ILLNESS compared to those that have not been infected.

a. (2 pts) What should Shmuel's null hypothesis be?
b. (2 pts) What should Shmuel's alternative hypothesis be?
c. (3 pts) What is a reasonable test statistic that Shmuel can use?

12. (4 pts) Suppose we want to find the 95% confidence interval for the mean of a population. We take 500 samples and obtain a dataset. We use bootstrapping and obtain the mean of the resampled data 10.000 times. Resulting means are stored in a numpy array 'means'. Write python code to find the lower and upper bound of the 95% confidence interval.