DS-UA-112: Introduction to Data Science (Fall 2019)

Midterm Exam (October 23 4:55-6:10PM)

- You have 70 minutes to complete the exam.
- The exam is closed book, closed notes, closed computer, closed calculator, except one hand-written 8.5" × 11" reference sheet of your own creation and the provided DS-UA 112 study guide.
- The exam has 7 pages. Mark your answers on the exam itself. We will not grade answers written on scratch paper.

Name: _____

NYU NetID: _____

NYU Email: _____
(as it appears on Gradescope)

Question	Points	Score
1	5	
2	4	
3	8	
4	8	
5	12	
6	4	
Total:	41	

1. **Conceptual**

   (a) (1 point) ___**F**___ **True or False**: If a dataset is large, then it's less likely to be biased.

   (b) (1 point) ___**F**___ **True or False**: There are 1000 kibibytes in a mebibyte.

   (c) (1 point) ___**F**___ **True or False**: All the file formats `*.json, *.tsv, *.xml` are suited for nested data.

   (d) (1 point) ___**F**___ **True or False**: We cannot join two tables with different number of rows.

   (e) (1 point) ___**F**___ **True or False**: Suppose we have a table of data collected through simple random sampling. Assume that we have 135 rows with 17 rows containing missing values. If we drop each row that contains missing values, then we necessarily have a simple random sample of size 118.

2. **Variables**

   You are modeling the cafe-related preferences of your DS-UA 112 classmates. What type of variable would you use to encode each of the following data? Check all data types that apply to the question.

   (a) (1 point) Each student's response to "Over the last month, how much do you estimate you spent on coffee?"

   ☐ Nominal ☐ Ordinal ☐ Continuous ■ **Discrete**

   (b) (1 point) Each student's drink of choice from a list of ten options (drip coffee, espresso, cappuccino, matcha latte, ...)

   ■ **Qualitative** ☐ Quantitative

   (c) (1 point) The average, for each cafe, over all students, to the answer "How much do you like studying at [name of cafe]?"

   ☐ Nominal ☐ Ordinal ■ **Continuous** ☐ Discrete

   (d) (1 point) Each student's response to "What about your favorite cafe makes it your favorite?"

   ■ **Qualitative** ☐ Quantitative

3. **Sampling**

   After making the modeling decisions in Question 2 for your classmates' cafe preferences, you are ready to begin collecting data.

   (a) How can we describe the different approaches to data collection?

      i. (1 point) If you choose a single section at random, and then survey all of its members this is a _____**Cluster Sample**_____

      ii. (1 point) What if, instead, you give the survey to every student in every section? This is a _____**Census**_____

      iii. (1 point) If you assign a number to each student in the course, and draw numbers randomly to survey, this is a ___**Simple Random Sample**___

(b) Assume that the sections of the course have enrollment

Thursday	15	25
Friday	20	30

i. (1 point) If you pick the method from Question (a i), then how many students do you expect to have in your survey −− assuming everyone responds?

Solution: On average 22.5 students

ii. (2 points) You know your friend is in a section on Friday. Everyone in section responded to the survey. If the sample has 20 or more students, then how likely is it that your friend is in the sample?

Solution: $\frac{1}{3} \cdot \frac{2}{5} + \frac{1}{3} \cdot \frac{3}{5} = \frac{1}{3}$

(c) (2 points) Given the survey from Question (2), suggest some important characteristics you could use to stratify a sample. Moreover, suggest a sampling procedure that will generate such a stratified sample.

Solution: We use stratified sampling with an inhomogeneous group to avoid misrepresenting parts of the population. We need to determine a sampling design such that

- We can assign each person in the population to to exactly one statum
- We can choose at random from a stratum an amount proportional to the population
- We can determine strata relevant to the survey through the determination of appropriate proxies

For example, suppose that we want to capture the relationship between age, income, and obligations on coffee habits. We can take as proxies year at college (freshman, sophomore, junior, senior), meal plan (yes, no), and number of classes (3,4,5 excluding uncommon amounts). We would have 24 stata. Based on the number of students in each class, the number of outstanding meal plans, and enrollment numbers, we would have randomly sample the appropriate number from each stratum.

4. **Probability**

(a) You are getting ready to deal with trick-or-treaters in a few weeks. You purchased the following candy, and tossed it all on a mysterious-looking jar:

Type	Amount	Flavor
Haribo Gummy	20	Chewy
Twix	15	Chocolatey
Mars	15	Chocolatey
Cheez-its	20	Cheesy
Goldfish	30	Cheesy

What is the probability that? Do not simplify the expressions.

i. (1 point) The first kid who comes receives a chocolatey snack?

Solution: 0.3

ii. (1 point) The first two kids both receive cheesy snacks?

Solution: $\frac{50}{100}\frac{49}{99}$

iii. (1 point) The first two kids receive snacks from different groups (neither both chocolatey, both cheesy, nor both chewy)?

Solution: $1 - \frac{20}{100}\frac{19}{99} - \frac{30}{100}\frac{29}{99} - \frac{50}{100}\frac{49}{99}$

iv. (1 point) The first two kids receive identical snacks?

Solution: $\frac{20}{100}\frac{19}{99} + \frac{15}{100}\frac{14}{99} + \frac{15}{100}\frac{14}{99} + \frac{20}{100}\frac{19}{99} + \frac{30}{100}\frac{29}{99}$

(b) (1 point) Let A be indicate which candy the first kid receives, and B be indicate which candy the second kid receives. Note that A and B are random.
__F__ **True or False** Are these independent events?

(c) (3 points) Your little sister also went trick-or-treating. One house of neighbors, who live fairly far, often give out full-sized bars, rather than mini-sized. Because they live far, you estimate there's only a 15% probability that your sister went to their house. If she did, you estimate the probability she received a full-sized bar to be 80%. If she didn't go, you estimate the probability she receives a full-sized bar to be only 5%. Your sister returns with a full-sized bar! What is the probability she went all the way to their house?

Solution:

$$\frac{0.8 \cdot 0.15}{0.8 \cdot 0.15 + 0.85 \cdot 0.05}$$

5. **Coding**

(a) (1 point) __T__ **True or False:** Does the following snippet of code throw an error?

```
import numpy as np
x = np.arange(7)
x[:3] += x[3:]
print(x ** 2)
```

(b) i. (1 point) Which of the following string match the regular expression `[bcr]1,2a+ts?`
Select all that do:
☐ bat ☐ Bat ■ **rats** ☐ bts ☐ cccaaaaats ☐ braaaaat

ii. (1 point) Using the `re` library, write a snippet of code that receives the string `s` and replaces

- one-digit numbers with `X`,
- two-digit numbers with `YY`, and
- three-digit numbers with `ZZZ`.

For example, if `s = 'I bought 300 grams of ground beef, 2 onions, and 12 cans of beer'`, at the end of the snippet it should be `s = 'I bought ZZZ grams of ground beef, X onions, and YY cans of beer'`

Solution: Not assuming surrounded by whitespace

```
s = re.sub(r'(\D{1})\d{1}(\D{1})', r'\1X\2', s)
s = re.sub(r'(\D{1})\d{2}(\D{1})', r'\1YY\2', s)
s = re.sub(r'(\D{1})\d{3}(\D{1})', r'\1ZZZ\2', s)
```

Assuming surrounded by whitespace

```
s = re.sub(r'[ ]{1}\d{1}[ ]{1}', r' X ', s)
s = re.sub(r'[ ]{1}\d{2}[ ]{1}', r' YY ', s)
s = re.sub(r'[ ]{1}\d{3}[ ]{1}', r' ZZZ ', s)
```

(c) Consider the file `classes.json` below:

```
[{
"Student": "sc2367",
"Details":
{
"Name": "Smith Carter",
"Year": 2,
"Classes":[112,113,114]
}},
{"Student": "zx1212",
"Details":
{
"Name": "Zexi Xu",
"Year": 2,
"Classes":[112,113,114]
}},
{"Student": "tr5564",
"Details":
{
"Name": "Tracy Rhodes",
"Year": 3,
"Classes":[121,125]
}}]
```

What would be the output of the following block of code:

```
import json
with open("classes.json", "r") as f:
    x = json.load(f)
print('A:{}'.format(len(x["Details"]["Classes"])))
print('B:{}'.format(len(x[2]["Details"]["Name"])))
```

i. (1 point) `A:` **(112,113,114)**

ii. (1 point) `B:` **Tracy Rhodes**

(d) (4 points) Consider the following `pandas DataFrame`.

id	age	color	fur	name
123	4	brown	shaggy	odie
456	3	grey	short	gabe
821	6	golden	curly	samosa
198	4	grey	shaggy	gabe
3	2	black	curly	bob barker
42	5	brown	shaggy	odie

i. (1 point) Which of the following expressions returns a `Series` containing only the names of all the grey dogs in the dogs `DataFrame`? Select all that apply.

☐ `dogs[(dogs["color"] == "grey") | (dogs["fur"] == "shaggy")]["name"]`

☐ `dogs[(dogs["color"] == "grey") & (dogs["fur"] == "shaggy")]["name"]`

☐ `dogs[(dogs["color"] == "grey") & (dogs["fur"] == "shaggy")]`

☐ `dogs[(dogs["name"].isin(['grey'])) & (dogs["fur"] == "shaggy")]`

■ **None of the above.**

ii. (1 point) Select all of the following expressions that generate a `DataFrame` containing only rows of brown dogs.

■ `dogs.set_index("color").loc["brown", :]`

☐ `dogs.if(dogs["color"] == "brown")`

■ `dogs[dogs["color"] == "brown"]`

☐ `dogs["color"] == "brown"`

☐ None of the above.

iii. (1 point) What would be the output of the following block of code

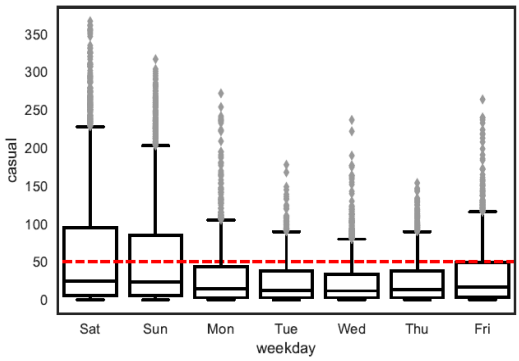
```
result = ( dogs[['age','color','fur']]
             .groupby(by=['color'])
             .agg({'age': sum, 'fur': len}) )
print(len(result))
```


6. **Visualization**

(a) (1 point) Suppose you wish to compare the number of children per household in the US and monthly earnings of households. Which style of plot would be the best?

☐ Scatter Plot
☐ Overlaid Line Plots
■ **Side-by-Side Box Plots**
☐ Stacked Bar Chart

(b) Consider the following visualization of the number of casual riders per hour by day of the week, which has been constructed from the bike sharing data used in Homework 3.



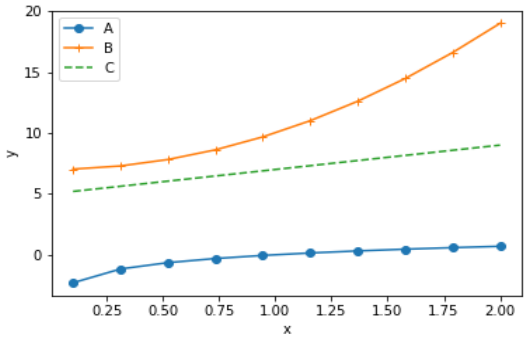
i. (1 point) Which day of the week had the least maximum number of casual riders?

- ☐ Sunday
- ■ **Thursday**
- ☐ Monday
- ☐ Tuesday
- ☐ None of the above.

ii. (1 point) Which of the following describe conclusions that we can draw about the distribution of rider counts on Fridays using the above plot? Select all that apply.

- ■ **Has outliers**
- ☐ Skewed left
- ■ **Skewed right**
- ☐ Symmetric
- ☐ None of the above

(c) (1 point) Consider the lineplot given below.



Which of the lines map to which function? Select the correct mapping below

- ☐ A - Quadratic, B - Linear, C - Logarithmic
- ☐ A - Linear, B - Logarithmic, C - Quadratic
- ☐ A - Quadratic, B - Logarithmic, C - Linear
- ■ **A - Logarithmic, B - Quadratic, C -Linear**
- ☐ None of the above matches are correct.

END OF EXAM – PRESENT YOUR NYU ID AT SUBMISSION