Introduction to Data Science

Midterm

CMSC320 – Fall 2020 – COVID Edition

Please **edit the header above** to contain your Name and Student ID, for example, "`John Dickerson, johnd`" and nothing else.

Please do not add extra pages to your answers, and stay within the allotted space – which is significantly larger than required for most questions. We'll be using GradeScope for grading, so writing outside of the expected location / adding pages of answers for a question will result in missed information. **Please do not shift questions around or add/remove pages from this document!**

Unless otherwise specified, a *Yes*/*No* answer without explanation will not get any points.

Show your reasoning. Write partial solutions. You will get a fair amount of the credit if we think that you know the concepts.

The number in square brackets indicate points (total of 50 points = 25% of the course grade).

Leave blank

THIS SHOULD BE ON PAGE 2:

Q1. [10pts]: Underline & bold **true** or **false** for each question below – or leave unedited. You will **gain 1 point** for each correct answer. You will **lose 0.5 points** for incorrect answers – that is, it is *worse* to answer a question incorrectly than it is to leave it unanswered.

A	The *mean* is a robust descriptive statistic.	True	False
B	The *range* is a robust descriptive statistic.	True	False
C	The *variance* is a robust descriptive statistic.	True	False
D	In general, it is best to maintain an index over each column in a database table.	True	False
E	Storing a graph via an adjacency list yields fast lookup time for relationships between two vertices in the graph.	True	False
F	Two variables, *X* and *Y*, have Pearson's correlation coefficient of +0.92. This means that an increase in *X* causes an increase in *Y.*	True	False
G	Complete case analysis as a method to deal with missing data will not induce bias into your analysis.	True	False
H	Multiple imputation methods for dealing with missing data will not induce bias into your analysis.	True	False
I	Branching in git is a heavyweight operation.	True	False
J	Data warehousing, where disparate data sources are centrally stored, is generally a good way to store and query quickly-changing data.	True	False

THIS SHOULD BE ON PAGE 3:
Q2. [6pts total]: Write *list comprehensions*, in Python, that describe the following ordered lists of elements.

Q2.i [1pt]: Even integers ranging from 0 to 100, inclusive.
Answer below:

Q2.ii [1pt]: The square of every odd number between 10 and 25, inclusive.
Answer below:

Q2.iii [2pts]: Given a list of characters, `c`, create a new list – via a single list comprehension – consisting of those characters in triplet *if they are alphanumeric*, and leaves them alone otherwise. For example, if:

```
 c = [‘*’, ‘c’, ‘m’, ‘s’, ‘c’, ‘3’, ‘2’, ‘0’, ‘!’]
```

then your code would return:

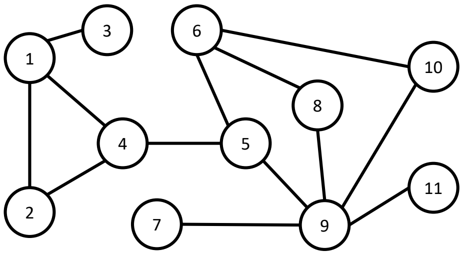
```
 [‘*’, ‘ccc’, ‘mmm’, ‘sss’, ‘ccc’, ‘333’, ‘222’, ‘000’, ‘!’]
```

Answer below:

Q2.iv [2pts]: In class, we discussed the relationship between `map` and list comprehensions. Write your solution to Q2.iii above – but using `map` / `filter` instead of a list comprehension.
Answer below:

THIS SHOULD BE ON PAGE 4:

Q3. [3 pts]: We'll use the graph below for a number of questions. For now, it is undirected and unweighted – that is, the edges are bidirectional and have unit weight.



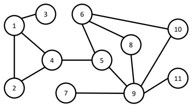
Q3.i [1 pt]: How would you store this graph *assuming it will not change*? Why?
Answer below:

Q3.ii [1 pt]: What if the graph were qualitatively similar but 100,000,000 times larger. Would that change how you store the graph *still assuming it will not change*?
Answer below:

Q3.ii [1 pt]: What if the graph will change over time (that is, vertices and edges may be added or deleted). Describe how this would impact your decision on storage method, if at all.
Answer below:

THIS SHOULD BE ON PAGE 5:

Q4. [8 pts]: Continue using the graph above (Q3), reproduced in much smaller form below:



Q4.i [1 pt]: What is the degree centrality of vertex 3, vertex 1, and vertex 9?
Answer below:

- V3:
- V1:
- V9:

Q4.ii [2 pt]: What is the closeness centrality of vertex 3, vertex 1, and vertex 9?
Answer below:

- V3:
- V1:
- V9:

Q4.iii [2 pt]: What is the betweenness centrality of vertex 3, vertex 1, and vertex 9?
Answer below:

- V3:
- V1:
- V9:

THIS SHOULD BE ON PAGE 6:

Q4.iv [3 pt]: We discussed in class how “centrality” is a qualitative notion, and could be quantitatively defined in many ways (e.g., proportional to degree, closeness, betweenness, as above). Define your own (non-trivial) notion of centrality here. Compute the centrality of vertices 1, 3, and 9, as above; how does your notion differ from those above?

THIS SHOULD BE ON PAGE 7:

Q5. [3 pts]: Create a table storing the following collection of documents – in this case, texts from posts on John's Instagram feed (Oct 28) – using a bag of words (BoW) representation. *Ignore punctuation and capitalization, and any word with three or fewer characters.*

1. Won't let a little thing like a broken ankle stop me from voting. #bootandall #vote #ivoted
2. ICYMI: As the season changes and the temperature drops in Front Royal, Virginia, cheetah
3. Have you tried yet?
4. Not really into espresso? Just get a Matcha Latte!
5. It never hurts to keep looking for sunshine.

Answer below:

THIS SHOULD BE ON PAGE 8:
Q6. [3 pts]: Recall term frequency-inverse document frequency (tf-idf). Your table above for Question 5 lists the term frequencies for each term in each document. Translate that table to its tf-idf representation. You may use the idf function from class, or define your own, so long as it makes sense. *You may leave answers "unreduced", or round to the nearest tenth.*
Answer below:

THIS SHOULD BE ON PAGE 9:

Q7. [3 pts]: You are tasked with building a RESTful API for the sell side (i.e., the stores selling things to consumers) of Stripe; stores want to be able to list new items, update pricing on items, report out of stock items, et cetera. Give reasonable functionality for the following HTTP methods:

Answer below:

GET:

PUT:

POST:

DELETE:

THIS SHOULD BE ON PAGE 10:
Q8. [6 pts]: This problem focuses on missing data. You are a 2020 Census worker. Wearing full protective gear, you go door to door asking questions. Assume everyone opens the door for you, but not everyone answers every question you ask. Mark the following as MCAR, MAR, or MNAR – or describe why you think none, or one, or more than one answer is relevant.

Q8.i [2 pt]: A smoke alarm goes off in the house and the survey respondent closes the door in your face, with no further answers given.
Answer below:

Q8.ii [2 pt]: You ask a question about paying off the mortgage; specifically, if this person has trouble paying down the debt on their house. You know that many people have an aversion to responding to questions pertaining to their net worth.
Answer below:

Q8.iii [2 pt]: You know that, on this particular day, residents of this building have a test fire alarm that activates. You knock, the alarm activates, and you cannot hear the response.
Answer below:

THIS SHOULD BE ON PAGE 11:

Q9. [3 pts]: We discussed many ways of obtaining data in class. List two of them and give an example of data we have used or discussed in labs or lecture.

Answer below:

THIS SHOULD BE ON PAGE 12:

Q10. [5 pts]: Describe your course project, i.e., your final tutorial. If you are working with a student, name that student (or students).

Answer below: