DS-100 Midterm Exam

Spring 2017

Name: _____

Email address: _____

Student id: _____

Page:	2	3	5	6	7	8	9	10	11	12	13	14	16	17	Total
Points:	7	10	11	4	6	3	8	2	9	6	10	9	9	7	101

Instructions:

- Please fill in you name, email address, and student id at the top of both this exam booklet and your answer sheet.
- All answers must be written on the separate answer sheet.
- This exam must be completed in the **1.5 hour time** period ending at **2:00PM**.
- You may use a single page (two-sided) cheat sheet.
- Work quickly through each question there are a total of 101 Points on this exam.
- You must turn in both this exam booklet and your answer sheet.
- **Don't cheat!**

1 Data Science Overview and Problem Formulation

1. For each of the following circle **T** for true or **F** for false on the answer sheet.

(1) [1 Pt.] Data science is the study of answering questions with big data.

> **Solution:** **False.** In class we defined data science as the use of computational, inferential, and data centric thinking to answer questions and solve problems. In addition, in many settings there will be limited data.

(2) [1 Pt.] Data scientists do most of their work in Python and are unlikely to use other tools.

> **Solution:** **False.** Data scientists use many programming languages and tools. In class we discussed surveys that suggested that SQL and then R are the most commonly used languages.

(3) [1 Pt.] Data science combines computer science, statistics, and domain knowledge to answer questions and solve problems.

> **Solution:** **True.** Data science is interdisciplinary.

(4) [1 Pt.] Data science is often an iterative process that can generate new questions.

> **Solution:** **True.** We view data science as a cyclic process (we give a hint to this in a later question).

(5) [1 Pt.] Most data scientists spend the majority of their time developing new models.

> **Solution:** **False.** Sadly, data suggests that most data scientists spend the majority of their time collecting and cleaning data and doing exploratory data analysis.

(6) [1 Pt.] The use of historical data to make decisions about the future can reinforce historical biases.

> **Solution:** **True.** A key ethical challenge of data driven decision making is that we tend to reinforce trends in our data.

(7) [1 Pt.] By eliminating information about a persons race, gender, and ethnicity, we eliminate potentially unethical social biases in our data driven decisions.

Solution: **False.** While masking certain attributes such as race or gender may appear to eliminate biases, often these attributes can be correlated with other attributes that remain in our data (e.g., income, zip-codes, education) and can affect our data driven decisions.

(8) [1 Pt.] Randomization of subjects into treatment and control groups is the gold standard method for drawing causal conclusions about the treatment.

Solution: **True.** Randomly constructing the control and treatment groups helps eliminate confounding variables and is the ideal method for testing causality.

(9) [1 Pt.] The population in a study must be a well defined group of *people*.

Solution: **False.** While the population should be well defined it may not be a group of people. In many settings the population might be devices, specimens, documents, or images.

(10) [1 Pt.] In the 2016 US Presidential Election the population of interest was *all citizens of the United States*.

Solution: **False.** The population of interest was all people who vote in the election. This excludes people who are too young to vote or are unlikely to vote.

(11) [1 Pt.] In data science and statistics, validation is the process of getting someone to verify your calculations.

Solution: **False.** Validation is the process of verifying conclusions through data collection and statistical inference.

Multiple Choice: For each of the following questions ***circle all*** *of the appropriate answers* on the answer sheet.

2. [3 Pts.] To draw meaningful conclusions about data you must have (**circle all**):

   A. context or domain knowledge.

   B. an understanding of the data collection process.

   C. a representative sample of the population of interest.

   D. vast amounts of data.

   E. none of the above

3. [3 Pts.] A large sample covering $2/3$ of the population of a major city is (**circle all**):

   A. guaranteed to be representative of the population because we have enough data.

B. guaranteed to be representative of the population because fundamental probability theory.

C. *not* guaranteed to be representative.

D. none of the above

2 Exploratory Data Analysis and Data Wrangling

4. [3 Pts.] Which of the following are reliable ways to assess the granularity of a table. **Circle all that apply.**

   A. Build histograms on each column.

   B. Identify a primary key.

   C. Compare the number of rows in the table with the number of distinct values in subsets of the columns.

   D. Address outliers via trimming or winsorizing.

   E. All of the above.

   F. None of the above.

5. [5 Pts.] Please match each sentence prefix with the best suffix. **Each suffix may only be used once and select only one suffix (matching).**

   (1) In a sparse matrix, _____**E**_____

   (2) If the sole primary key for a relation is the set of all columns, _____**B**_____

   (3) In a nested data format like JSON or XML, _____**A**_____

   (4) In the standard relational representation of a matrix, _____**C**_____

   (5) In a relational table with NULL values, _____**D**_____

   (A) it is difficult to specify how to summarize the data.

   (B) we know the granularity is as fine as possible.

   (C) the primary key consists of two columns.

   (D) there is an explicit representation of what is unknown.

   (E) the missing cells are assumed to have value 0.

   (F) each column sums to 1

   (G) the missing cells are assumed to have value `NaN`.

6. [3 Pts.] Recall the Ta Feng dataset from Homework 3. In one of the charts you built, you studied Repeat Business, by summarizing the number of shopping trips per customer in a histogram. The x axis was binned by "Number of Transactions in 4 months," the y axis showed the number of customers per bin.

   (1) The distribution in this histogram was ____**right**____ skewed.

   (2) The mode of the distribution had value _____**1**_____.

   (3) Because the data was skewed, we tried a _____**log**_____ transform to stretch the x axis.

data.csv

```
Reported crime in Alabama,
,
2004,4029.3
2005,3900
2006,3937
2007,3974.9
2008,4081.9
,
Reported crime in Alaska,
,
2004,3370.9
2005,3615
2006,3582
2007,3373.9
2008,2928.3
2009,3639.8
<End of File>
```

7. [4 Pts.] Consider the following CSV file on crime rates provided by the US Department of Justice. Assume you have been told that the *row delimiter is "newline"*, and the *column delimiter is ","*. Which of the following statements are true? **Circle all that apply**:

   A. In its current form, this data cannot be loaded into a relational database.

   B. If we load this data into a relational database without transformation, we will lose important information.

   Solution: We would lose order information.

   C. We could transform the structure of this data into a matrix of real numbers, with a row for each state and column for each year.

   D. 2009 is an outlier.

   E. This data appears to have limited scope.

   Solution: We are missing states.

   F. All of the above.

   G. None of the above.

3 Visualization and Communication

8. [3 Pts.] Suppose we want to make a scatter plot for the houses sold in the SF Bay Area in 2016. The x-axis is the size of the house in square-meters and the y-axis is the cost of the house per square-meter. Over 20,000 homes were sold last year. What techniques would you employ to avoid problems related to over plotting. **Circle all that apply.**

   A. jitter the values for cost

   B. plot a smoothed curve of average cost for the size of the house

   C. make the plotting symbols semi-transparent

   D. use color to distinguish the city where the house was sold

   E. none of the above

9. [3 Pts.] Which visualization would be appropriate for examining the relationship between the birth weight of a baby and the number of siblings of the baby? (Assume there are a few hundred observations) **Circle all that apply.**

   A. side-by-side box plots of weight by number of siblings

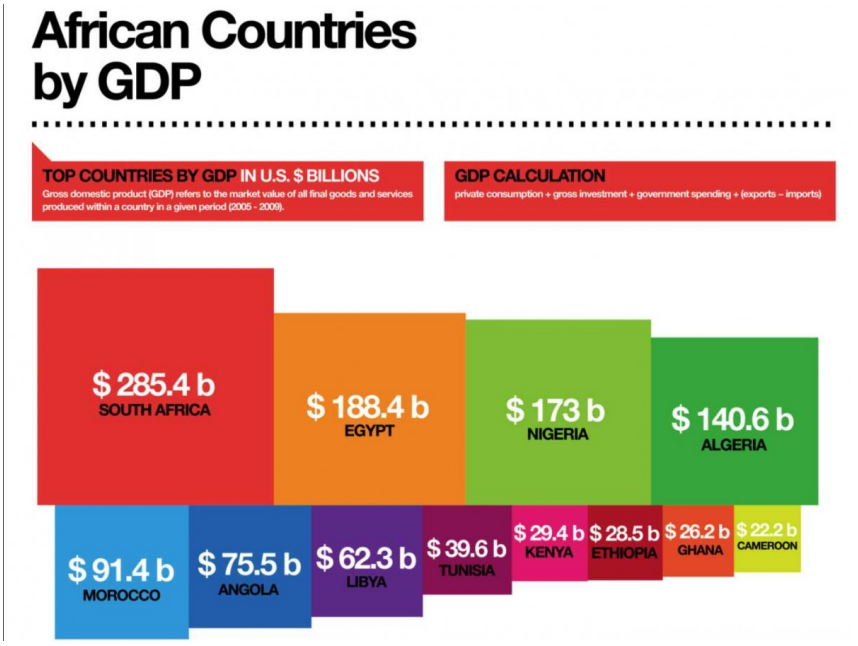
   B. scatter plot of weight by number of siblings

   C. bar plot of weight by number of siblings

   D. overlaid density curves of weight, one for each number of siblings

   E. mosaic plot of weight by number of siblings

   F. none of the above



10. [3 Pts.] Consider the above figure. Which of the following suggestions would better facilitate comparisons of the GDP for African countries. **Circle all that apply.**

    A. arrange the countries in alphabetical order to make it easier to find a country's GDP

    B. choose a sequential color palette to match size of the GDP

    C. make a box plot of GDP to show the skew and spread in GDP

    D. make a dot chart of GDP

    E. none of the above

4 Prediction and Inference

11. For each of the following circle **T** for true or **F** for false on the answer sheet.

(1) [1 Pt.] Clustering is a form of unsupervised learning.

> **Solution:** **True.**

(2) [1 Pt.] If the p-value is less than our chosen significance level, then we may reject the null hypothesis.

> **Solution:** **True.**

(3) [1 Pt.] It never makes sense to make predictions about the past.

> **Solution:** **False.** We may want to make predictions about past events that we did not observe.

12. Use the following hypothesis:

> *Berkeley students who have taken Data8 are more likely to be hired as data scientists than those who have not taken Data8.*

to answer each of the following questions. For each of the following questions **circle all of the appropriate answers:**

(1) [1 Pt.] Which of the following is the population:

A. All students in the US

B. Berkeley students

C. Students who have taken Data8

D. Berkeley students with job offers.

E. none of the above

(2) [1 Pt.] A dataset was constructed by inviting Data8 students to complete a voluntary survey. Such a dataset would most likely be described as a:

A. Sample

B. Census

(3) [3 Pts.] Which of the following are reasons the voluntary survey of Data8 students would be insufficient to make a conclusion about the hypothesis?

A. The sample size is guaranteed to be too small.

B. The survey may not be representative of Data8 students overall.

C. The survey would tell us nothing about non-Berkeley students.

D. The survey would tell us nothing about students who have not taken Data8.

E. The survey would tell us nothing about students who were not hired as data scientists.

F. None of the above.

(4) [2 Pts.] A second analysis was conducted by asking Berkeley graduates employed as data scientists. Together with the survey of Data8 students, would this be sufficient to make a conclusion about the hypothesis?

A. Yes

B. No

Solution: This problem is slightly tricky. The survey of Data 8 students would not give us any data about students that did not take Data 8. While the survey of data scientists would not provide information about students who did not become data scientists. In particular neither of these samples would contain the Berkeley students who did not take Data8 and did not get a job as a data scientist.

5 Pandas, Relational Algebra, and SQL

Consider the following simplified schema based on the FEC data for candidates and committees.

```
-- donations records money given by donors to committees.
donations(did, name, address, don_state, don_date, comm_id, amount)

-- comm records information about committees,
-- which are the entities that accept donations.
-- Some (but not all) committees are associated with a specific
-- candidate (via the cand_id field).
comm(comm_id, comm_nm, comm_state, comm_party, cand_id)

-- cand records information about each candidate for office
cand(cand_id, name, party, elec_year, state, office)
```

13. Write a relational algebra expression for the following statements.

(1) [1 Pt.] All attributes of candidates affiliated with the `'R'` party.

Solution:

$$\sigma_{\texttt{party = 'R'}}(\texttt{cand})$$

(2) [2 Pts.] Distinct offices for which candidates are running.

Solution:

$$\pi_{\texttt{office}}(\texttt{cand})$$

(3) [2 Pts.] Names of committees for party `'D'`.

Solution:

$$\pi_{\texttt{comm\_nm}}(\sigma_{\texttt{party = 'D'}}(\texttt{comm}))$$

(4) [4 Pts.] Names of committees and candidates where the committee is affiliated with the candidate, but the committee is not located in the candidate's state.

Solution:

$$\pi_{\texttt{name, comm\_nm}}(\sigma_{\texttt{state != comm\_state}}(\texttt{cand} \bowtie \texttt{comm}))$$

14. Complete the following SQL queries *on the answer sheet*.

(1) [3 Pts.] All attributes of candidates with the `'R'` party:

```
SELECT  *
  FROM  cand
 WHERE  party = 'R'
```

(2) [3 Pts.] Distinct offices for which candidates from the `'D'` party are running:

```
SELECT  DISTINCT office
  FROM  cand
 WHERE  party = 'D'
```

The following is a copy of the schema from previous page for quick reference.

donations(did, name, address, don_state, don_date, comm_id, amount)

comm(comm_id, comm_nm, comm_state, comm_party, cand_id)

cand(cand_id, name, party, elec_year, state, office)

(3) [5 Pts.] Names of committees headquartered in CA that are *not* affiliated with a candidate for office in CA:

```
SELECT comm_nm
  FROM comm ___LEFT JOIN___ cand
    ON comm.cand_id = cand.cand_id
 WHERE ____comm_state = 'CA'____
   AND (____comm.cand_id IS NULL____
        OR ____cand.state <> 'CA'____)
```

(4) [5 Pts.] For each state, report the state name and the total amount of donations from that state that went to committees affiliated with a candidate:

```
SELECT __D.don_state__, SUM(amount)
  FROM ____donor D, comm CM____
 WHERE ____D.comm_id = CM.comm_id____
   AND ____CM.cand_id IS NOT NULL____
 GROUP BY ____D.don_state____
```

15. Consider the following SQL expressions.

(A) `SELECT DISTINCT count(*) FROM cand`

(B) `SELECT count(*) FROM cand WHERE state = 'CA'`

(C) `SELECT count(*) FROM cand WHERE state = 'CA' GROUP BY party`

(D) `SELECT COUNT(DISTINCT state) FROM cand`

(E) `SELECT count(DISTINCT name) FROM cand WHERE state = 'CA' GROUP BY party`

(F) `SELECT count(*) FROM cand GROUP BY party HAVING state = 'CA'`

Assume that `cand` is also a dataframe in Pandas. For each of the following Pandas expression identify which of the above SQL expressions always produces the same result assuming that the output order does not matter.

(1) [1 Pt.] `cand['state'].nunique()`

> **Solution:** **(D)** The `DISTINCT`, `ORDER BY` and `LIMIT` clauses are irrelevant here because there is only one group, so only one row to return.

(2) [2 Pts.] `len(cand)`

> **Solution:** **(A)** The `DISTINCT`, `ORDER BY` and `LIMIT` clauses are irrelevant here because there is only one group, so only one row to return.

(3) [3 Pts.]
```
cand[cand['state'] == 'CA'][['party', 'name']]
    .groupby('party')['name']
    .nunique()
```

> **Solution:** **(E)** This expression considers only the rows which have `state = 'CA'` and then groups those rows by `party` and computes the number of unique names in each group.

(4) [3 Pts.]
```
cand.groupby('state')
    .get_group('CA')['name']
    .count()
```

Solution: **(B)** This expression is slightly tricky. Notice that the Pandas operation requests the group corresponding to `party = 'R'` and counts the number of records in that group. This is equivalent to computing the number of records where the party is `'R'`.

6 Probability and Maximum Likelihood

16. [2 Pts.] Which of these are true statements about any events A and B. **Select all that apply.**

A. $\mathbb{P}(A \textbf{ and } B) = \mathbb{P}(A|B) \times \mathbb{P}(B)$

Solution: This is true by the definition of conditional probability.

B. $\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B)$

C. $\mathbb{P}(A) + \mathbb{P}(\textbf{not } A) = 1$

Solution: This is the rule of complements.

D. None of the above.

17. [3 Pts.] Suppose a robotic arm in a car factory has a daily failure probability of p. Which of the following probability mass functions would model the probability that the part fails for the first time on day x. (Note that C is a normalizing constant).

A. $P(x) = Cp^x$

B. $P(x) = C(1-p)p^x$

C. $P(x) = C(1-p)^{(x-1)}p$

D. $P(x) = Cpe^{-px}$

18. [4 Pts.] A town has 200 families, where 20% have 0 children, 30% have 1 child, and 50% have 2 children. The names of all the children are written on tickets and placed in a glass bowl. The tickets are well mixed. One ticket is drawn. What is the chance the child is from a 2-child family? Assume the children's names are unique.

A. $1/3$

B. $1/2$

C. $5/8$

D. $10/13$

E. none of the above

Solution: We can compute the solution by looking at the fraction of tickets in the barrel that come from 2 children families. It is important to note the following two conditions

- There will be no tickets corresponding to families with no children

- There will be two tickets for each family with two children

$$\frac{200 \cdot \frac{5}{10} \cdot 2}{200(\frac{5}{10} \cdot 2 + \frac{3}{10})} = \frac{200}{260} = \frac{10}{13}$$

OR

$$\frac{\frac{5}{5+3} \cdot 2}{\frac{5}{5+3} \cdot 2 + \frac{3}{5+3}} = \frac{5 \cdot 2}{5 \cdot 2 + 3} = \frac{10}{13}$$

19. [3 Pts.] Suppose a random variable has the following probability mass function:

$$P(X = k) = (k-1)p^2(1-p)^{k-2} \text{ for } k = 2, 3, 4, \ldots$$

We observe 3 independent values from the pmf to be 3, 7, 2. What is the likelihood for p?

A. $12p^2(1-p)^6$

B. $12p^6(1-p)^6$

C. $54p^6(1-p)^{12}$

D. $12p^8(1-p)^5$

E. none of the above

20. [4 Pts.] Assume you have a single data observation k, what is the maximum likelihood estimate for θ given the following likelihood function:

$$\mathcal{L}(\theta) = \frac{\theta^k}{k!}e^{-\theta}$$

A. $1/k$

B. k

C. k/n

D. $k!$

E. none of the above

End of Exam