INSTRUCTIONS

- You have 80 minutes to complete the exam.
- This exam has 6 pages and a total of 40 points.
- The exam is closed book, closed notes, closed computer, closed calculator, except one hand-written 8.5" × 11" crib sheet of your own creation and the official Data 100 reference sheet.
- Mark your answers on the exam itself. We will not grade answers written on scratch paper.
- Please put your name at the top of every page of the exam.

<table>
<thead>
<tr>
<th>Last name</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>First name</td>
<td></td>
</tr>
<tr>
<td>Student ID number</td>
<td></td>
</tr>
<tr>
<td>CalCentral email (<a href="mailto:_@berkeley.edu">_@berkeley.edu</a>)</td>
<td></td>
</tr>
<tr>
<td>Name of the person to your left</td>
<td></td>
</tr>
<tr>
<td>Name of the person to your right</td>
<td></td>
</tr>
</tbody>
</table>

*All the work on this exam is my own.* (please sign)
This page is intentionally left blank, but feel free to use it as scratch paper.
1. (12 points) Rush Hour

Fill in both the Python code and the SQL query to produce each result below, assuming that the following tables are stored both as Pandas DataFrames and SQLite tables. Only the first few rows are shown for each table. The trfc table contains one row per sensor recording of hourly average car speed in mph. The time column contains strings that encode the hour of day and whether the time occurred during rush hour. The dates table contains one row for all dates in 2019 with their days of the week.

<table>
<thead>
<tr>
<th>trfc</th>
<th>dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>dt</td>
</tr>
<tr>
<td>hr=01,rush=no</td>
<td>May 1</td>
</tr>
<tr>
<td>hr=13,rush=no</td>
<td>May 3</td>
</tr>
<tr>
<td>hr=08,rush=yes</td>
<td>May 29</td>
</tr>
<tr>
<td>hr=18,rush=yes</td>
<td>May 3</td>
</tr>
</tbody>
</table>

(a) (4 pt) Calculate the average speed during rush hour.

Python: `trfc.loc[trfc[__________].str.contains(___________), __________]._______________`

SQL: `SELECT AVG(_________) FROM ___________ WHERE _________ LIKE _______________________________;`

(b) (4 pt) Create a table t with one row per recording in trfc. Each row should contain the day of week, speed, hour of day as a two-character string, and whether the recording occurred during rush hour (either “yes” or “no”). Hint: The correct call to extract() takes in a single regex with two captured groups: one for hour of day and one for rush hour. Also, the provided SQL already computes the hr column.

Python: `m = trfc.merge(________________, ____________________________, _________________________)`

```
m[['hr', 'rush']] = m['time'].str.extract(r'______________________________________')
t = m[['day', 'spd', 'hr', 'rush']]
```

SQL: `CREATE TABLE t AS SELECT day, spd, SUBSTR(time, 4, 2) AS hr,
CASE WHEN _________ LIKE _____________ THEN _____________ ELSE _____________ END AS rush
FROM ____________________________________ ON __________________________________________;`

(c) (4 pt) Find the minimum speed in a cluster sample with two clusters: take a SRS of two unique values in day, then find the minimum speed across all recordings on those days of the week. Note that there are many speed recordings for every date in May. You may assume that the t table is correctly created.

Python: `days = np.random.choice(__________________________, size=2, replace=False)`

```
t.loc[_________________________________________________, _______________]._______________
```

SQL: `SELECT MIN(______________) FROM t WHERE ________________ IN (SELECT ______________ FROM _____________ GROUP BY _________________ ORDER BY ____________________________ LIMIT ____________)`
2. (5 points) Sampling

Suppose that there are ten people in a room and one of these people is named Sam. We will take random samples of these people and compute probabilities associated with these samples. Bubble in the circles corresponding to your answers.

(a) (1 pt) What is the probability that Sam is not in a simple random sample of 1 individual?

- $\frac{9}{10}$
- $\frac{1}{5}$
- $\frac{2}{5}$
- $\frac{1}{2}$
- $\frac{4}{5}$
- $\frac{9}{10}$
- None of these

(b) (2 pt) What is the probability that Sam is not in a simple random sample of 3 individuals?

- $\frac{1}{5}$
- $\frac{3}{10}$
- $\frac{7}{10}$
- $\frac{9}{10}$
- $\left(\frac{3}{10}\right)^3$
- $\left(\frac{9}{10}\right)^3$
- None of these

(c) (2 pt) Suppose we take a sample of 2 individuals by first drawing a simple random sample of size 5, then taking a simple random sample of size 2 from that sample. What is the probability that Sam is not in this sample?

- $\frac{1}{5}$
- $\frac{1}{2}$
- $\frac{3}{5}$
- $\frac{1}{2}$
- $\frac{4}{5}$
- $\left(\frac{1}{2} \cdot \frac{9}{10}\right)$
- None of these

3. (5 points) Go Bears?

$([go]|[bear])^+?!$

Shade in the box for all of the strings below that match the regular expression above. Only shade a box if the whole string matches the expression, not just a substring. Do not put a checkmark in the box; shade in the entire box.

- [ ] go!
- [ ] gear?
- [ ] garbs!
- [ ] bearsbears!
- [ ] gobears?!
4. (10 points) Ballers

Suppose you have a dataset of 30,000 basketball games played in the NBA over the last 20 years. After conducting EDA, you find that each row in this dataset corresponds to a single game with the six columns described below. You also find that all columns contain duplicate values.

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>date</td>
<td>Date of game</td>
<td>Numeric</td>
</tr>
<tr>
<td>team</td>
<td>Home team’s name</td>
<td>Nominal</td>
</tr>
<tr>
<td>opp</td>
<td>Visiting team’s name</td>
<td>Nominal</td>
</tr>
<tr>
<td>win</td>
<td>1 if the home team won, else 0</td>
<td>Ordinal</td>
</tr>
<tr>
<td>scr</td>
<td>Score of home team</td>
<td>Numeric, between 80 and 170</td>
</tr>
<tr>
<td>opp_scr</td>
<td>Score of visiting team</td>
<td>Numeric, between 80 and 170</td>
</tr>
</tbody>
</table>

For each of the following questions, shade in one or more boxes corresponding to your answer.

(a) (2 pt) Given the data types of this dataset, which of the following visualizations are not appropriate?

- [ ] A histogram of win.
- [ ] A histogram of scr.
- [x] A box plot with team on the x-axis and opp_scr on the y-axis.
- [x] A 2D KDE plot with scr on the x-axis and opp_scr on the y-axis.

(b) (2 pt) Which of the following plots will likely suffer from overplotting?

- [x] A scatter plot with dates on the x-axis and number of games played on that date on the y-axis.
- [ ] A scatter plot with scr on the x-axis and opp_scr on the y-axis.
- [ ] A scatter plot with scr on the x-axis and win on the y-axis.
- [ ] A dot plot with team on the x-axis and the average of scr for each team on the y-axis.

(c) (2 pt) Which of the following plots show all teams that improved in scoring? (These are the teams with higher scores at later dates.)

- [ ] A line plot with one line for every team with date on the x-axis and scr on the y-axis.
- [x] A line plot with one line for every team with date on the x-axis and opp_scr - scr on the y-axis.
- [ ] One separate line plot for each team with date on the x-axis and scr on the y-axis.
- [ ] A bar plot with one bar per team and average of the latest five games as bar lengths.

(d) (2 pt) Which of the following plots show that home teams scored more on average than visiting teams?

- [x] A 2D KDE plot with opp_scr - scr on the x-axis and scr on the y-axis.
(e) (2 pt) Suppose you find a linear relationship when you make a scatter plot with np.log(scr) on the x-axis and np.log(opp_scr) on the y-axis. When you fit a least-squares line on this plot, you find the slope of the line is 2 and the intercept is 5. Which of the following relationships hold?

- opp_scr = 2 · scr + 5
- log(opp_scr) = log(2 · scr + 5)
- opp_scr = e^5 · e^{2·scr}
- opp_scr = e^5 · scr^2

5. (8 points) Dim Matrices

You perform principal component analysis on a data matrix D using the following Python code from lecture.

```python
m = D.shape[0]
X = (D - np.mean(D, axis=0)) / np.sqrt(m)
u, s, vt = np.linalg.svd(X, full_matrices=False)
```

Here are the values of a few expressions executed after running the code above:

<table>
<thead>
<tr>
<th>Python</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>array([12, 6, 4, 2, 0])</td>
</tr>
<tr>
<td>u.shape</td>
<td>(40, 5)</td>
</tr>
<tr>
<td>vt[0]</td>
<td>array([0.8, 0, -0.6, 0, 0])</td>
</tr>
</tbody>
</table>

(a) (2 pt) What is the shape of D? Recall that a matrix with 10 rows and 3 columns has shape (10 × 3).

- (5 × 5)  
- (5 × 40)  
- (40 × 5)  
- (40 × 40)

(b) (2 pt) What is the rank of D?

- 0  
- 1  
- 2  
- 3  
- 4  
- 5  
- 6

(c) (2 pt) What percentage of D’s total variance is kept if PCA is used to reduce the number of dimensions to 3?

- 12%  
- 22%  
- 60%  
- 92%  
- 98%  
- Not enough information

(d) (2 pt) Suppose the last row in X is: array([10, 4, -5, 2, 1]). After projecting this point onto the first principal component, what is the location of this point on the principal component axis?

- 10  
- 11  
- 24  
- 100  
- 128  
- Not enough information