This test has 6 problems on 12 pages. Be sure your test has them all. Do not spend too much time on one question — remember that this class lasts only 75 minutes and there are 75 points on the exam. That means you should spend no more than 1 minute per point.

Make sure to put your name and NetID in the space provided above and to put your NetID on the top of each page (worth 2 points).

Assume that the statements from datascience import * and import numpy as np have been executed. Don’t worry about getting function or method names exactly right. Assume that if a column contains numbers without a decimal point, then they are integers, if the numbers have a decimal point they are floats, and otherwise, it is a column of strings.

Assume that all libraries and packages we’ve discussed are imported in any code you write. You can write any helper functions you would like in solving the problems.

You should show your work on any questions involving arithmetic or analysis.

**Please write with a dark heavy pencil, or black pen and avoid writing too close to the edge of the paper.** Use the blank page at the end of the test if you cannot fit everything on a particular page. Make a note on the appropriate problem if you use an extra sheet.

All pages **including your two (2) sheets of notes** must be turned in.

Don’t panic. Just read all the questions carefully to begin with, and first try to answer those parts about which you feel most confident. Do not be alarmed if some of the answers are obvious.
PROBLEM 1: \( T/F \) (6 points)

In an August 2018 *Lancet* study, researchers studied 694 individual and population-level data sources detailing alcohol consumption among 28 million individuals and across populations. The figure from the paper below tracks the relationship between the risk of alcohol-related health problems and the number of alcoholic drinks consumed per day.

From the description of the study and the figure above, select either True or False for each of the following statements. *Briefly* justify your answer.

1.1. Drinking more alcohol causes more risk of health problems.

\( \text{True} \) \( \text{False} \)

1.2. This study shows that people who drink alcohol on average have a higher risk of alcohol-related health problems.

\( \text{True} \) \( \text{False} \)

1.3. This study shows that if you stop drinking alcohol, you will probably live longer.

\( \text{True} \) \( \text{False} \)
PROBLEM 2: \textit{(Expressions (12 points))}

The following code segment is executed.

\begin{verbatim}
a = make_array(0, 3, 2, 2)
b = make_array(2, 2, 2, 2)
c = make_array(’a’,’b’,’c’,’d’)
t = Table().with_columns(
    ’First’, c,
    ’Second’, a*b,
    ’Third’, c*b,
    ’Fourth’, make_array(c.item(a.item(0)), c.item(a.item(1)),
                           c.item(a.item(2)), c.item(a.item(3)))
)
\end{verbatim}

Fill in the table with the resulting values for \textit{t} below.

<table>
<thead>
<tr>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Fourth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PROBLEM 3: (Functions (8 points))

A square pyramid is a five-sided solid as shown.

The volume of a pyramid is given by this formula in terms of the length of a side of the base, $s$, and the height of the pyramid, $h$:

$$\frac{s^2 h}{3}$$

The surface area is given by this formula again in terms of the length of a side $s$ and the height $h$.

$$s^2 + 2s \times \sqrt{\frac{s^2}{4} + h^2}$$

You may use the `math.sqrt` function to calculate square roots, e.g., `math.sqrt(25)` evaluates to 5. Complete both functions below to return the volume and surface areas of a square pyramid.

```python
def pyrvolume(s, h):
    """
    return volume of pyramid with side s and height h
    """

def pyrsurface(s, h):
    """
    return surface area of pyramid with side s and height h
    """
```
PROBLEM 4: (Movie Tables (13 points))

Let imdb be a table with information about various movies.

<table>
<thead>
<tr>
<th>Votes</th>
<th>Rating</th>
<th>Title</th>
<th>Year</th>
<th>Decade</th>
</tr>
</thead>
<tbody>
<tr>
<td>88355</td>
<td>8.4</td>
<td>M</td>
<td>1931</td>
<td>1930</td>
</tr>
<tr>
<td>132823</td>
<td>8.3</td>
<td>Singin’ in the Rain</td>
<td>1952</td>
<td>1950</td>
</tr>
<tr>
<td>74178</td>
<td>8.3</td>
<td>All About Eve</td>
<td>1950</td>
<td>1950</td>
</tr>
<tr>
<td>635139</td>
<td>8.6</td>
<td>Léon</td>
<td>1994</td>
<td>1990</td>
</tr>
<tr>
<td>145514</td>
<td>8.2</td>
<td>The Elephant Man</td>
<td>1980</td>
<td>1980</td>
</tr>
<tr>
<td>425461</td>
<td>8.3</td>
<td>Full Metal Jacket</td>
<td>1987</td>
<td>1980</td>
</tr>
<tr>
<td>441174</td>
<td>8.1</td>
<td>Gone Girl</td>
<td>2014</td>
<td>2010</td>
</tr>
<tr>
<td>850601</td>
<td>8.3</td>
<td>Batman Begins</td>
<td>2005</td>
<td>2000</td>
</tr>
</tbody>
</table>

241 lines omitted...

Answer the following questions about the movie ratings below.

4.1. [4pts] Set lowest_rating to the Rating of the lowest-rated movie in the dataset.

4.2. [4pts] Set elite_movie to the number of movies whose rating is ≥ 9.0.

4.3. [5pts] Set med_movie to the Title of the movie whose rating is closest to the average rating of all movies in the database.
PROBLEM 5:  (Uber Tables (21 points))

In Homework 3, you were given data of 200,000 Uber rides in the Manila, Philippines, and Boston, Massachusetts metropolitan areas. The sourceid and dstid columns contain integers corresponding to start and end locations of each ride. The hod column contains integers corresponding to the hour of the day the ride took place. The ride time column contains the length of the ride as a float, in minutes.

```python
boston = Table.read_table("boston.csv")
bins = np.arange(0, 120, 5)
boston.hist("ride time",bins=bins,unit="Minutes")
```

5.1.  [4pts] Based on the histogram below, what percentage of rides took less than 15 minutes in Boston? Show your work.
5.2. [3pts] The histogram below shows the ride times for Manila.

```python
manila = Table.read_table("manila.csv")
manila.hist("ride time", bins=bins, unit="Minutes")
```

Based on the two histograms, which city has the higher median duration of Uber rides? The median is the middle number in a sorted range. Justify your answer.

5.3. [4pts] Write code below to set the value of `short_rides` to the proportion of rides that have a duration of less than 15 minutes in Boston. You need to actually create a table with all of the short rides - not just just repeat the arithmetic in the histogram question above.
5.4. [5pts] Write code to set maxcars to the number of rides that begin at the most popular location (sourceid) and time (hod) to start rides (e.g, 70 rides start at 11pm from sourceid 435) in Boston. You will need to determine what start location and hour of day is the most popular location and time to begin rides and then count how many rides originate there.

5.5. [5pts] You are given a working version of a function dist that tells you how far apart two locations are.

```python
def dist(start, end):
    """
    Returns the distance in miles between sourceid start and the destid end
    """
    [ implementation omitted ]
```

You want to compute how long it takes to travel on average in both cities. The speed of a trip is the ratio of distance divided by the duration of the trip (ride time) Set the variable bs to the average speed of trips in Boston and ms to the average speed of trips in Manila.
PROBLEM 6: (NBA (13 Points))

Assume we have salary information for the 2015-2016 NBA Season. The nba table contains 4 columns: the name of the player, the player’s position, team, and salary in millions of dollars. The first eight rows are shown below.

\[
\text{nba = Table.read_table('nba_salaries_16.csv').relabeled(3, 'SALARY')}
\]

<table>
<thead>
<tr>
<th>PLAYER</th>
<th>POSITION</th>
<th>TEAM</th>
<th>SALARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paul Millsap</td>
<td>PF</td>
<td>Atlanta Hawks</td>
<td>18.6717</td>
</tr>
<tr>
<td>Al Horford</td>
<td>C</td>
<td>Atlanta Hawks</td>
<td>12</td>
</tr>
<tr>
<td>Tiago Splitter</td>
<td>C</td>
<td>Atlanta Hawks</td>
<td>9.75625</td>
</tr>
<tr>
<td>Jeff Teague</td>
<td>PG</td>
<td>Atlanta Hawks</td>
<td>8</td>
</tr>
<tr>
<td>Kyle Korver</td>
<td>SG</td>
<td>Atlanta Hawks</td>
<td>5.74648</td>
</tr>
<tr>
<td>Thabo Sefolosha</td>
<td>SF</td>
<td>Atlanta Hawks</td>
<td>4</td>
</tr>
<tr>
<td>Mike Scott</td>
<td>PF</td>
<td>Atlanta Hawks</td>
<td>3.33333</td>
</tr>
<tr>
<td>Kent Bazemore</td>
<td>SF</td>
<td>Atlanta Hawks</td>
<td>2</td>
</tr>
</tbody>
</table>

... (409 rows omitted)

6.1. [4pts] Create a salary function that will return the salary for a player in dollars. For example, 
\[
salary('Al Horford') \]
should return 12000000.0 or 1.2e7. You can assume nba is defined above.
6.2. [4pts] Set the variable `superteams` equal to the number of teams with at least three (3) players who made more than 10 million dollars in salary in 2015-2016.

6.3. [5pts] You would like to determine the starters for each team. Each team has one starter per position. Assume the starter for a team at a position is the player with the highest salary on that team in that position. Set `startsal` to a table with the salaries for every starter on every team, as in the picture below.

<table>
<thead>
<tr>
<th>TEAM</th>
<th>C</th>
<th>PF</th>
<th>PG</th>
<th>SF</th>
<th>SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlanta Hawks</td>
<td>12</td>
<td>18.6717</td>
<td>8</td>
<td>4</td>
<td>5.74648</td>
</tr>
<tr>
<td>Boston Celtics</td>
<td>2.61698</td>
<td>5</td>
<td>7.73034</td>
<td>6.79612</td>
<td>3.42551</td>
</tr>
<tr>
<td>Brooklyn Nets</td>
<td>1.3629</td>
<td>11.236</td>
<td>6.3</td>
<td>24.8949</td>
<td>3.42551</td>
</tr>
<tr>
<td>Charlotte Hornets</td>
<td>13.5</td>
<td>7</td>
<td>12</td>
<td>6.3314</td>
<td>13.1253</td>
</tr>
<tr>
<td>Chicago Bulls</td>
<td>13.4</td>
<td>5.54373</td>
<td>20.0931</td>
<td>2.38044</td>
<td>16.4075</td>
</tr>
<tr>
<td>Cleveland Cavaliers</td>
<td>14.2609</td>
<td>19.689</td>
<td>16.4075</td>
<td>22.9705</td>
<td>8.98876</td>
</tr>
<tr>
<td>Dallas Mavericks</td>
<td>5.2</td>
<td>15.4937</td>
<td>5.37897</td>
<td>15.3615</td>
<td>1.449</td>
</tr>
<tr>
<td>Denver Nuggets</td>
<td>5.6135</td>
<td>11.236</td>
<td>4.345</td>
<td>14</td>
<td>1.58448</td>
</tr>
</tbody>
</table>

... (22 rows omitted)
This page is blank. Use as extra space if needed.
End of test