Name: _____ USC ID: _____

Quiz 3: Apache Spark (10 points), **15 minutes**

TA handling this quiz: Abhishek Bhatt<abhishpb@usc.edu>

Consider again the two tables: Height(name, height) and Weight(name, weight), whose contents are stored in two text files: height.txt and weight.txt. Each line of the file contains a tuple separated by a comma. Consider an example data set where height.txt contains:

John,170
David,185
Jennifer,178
John,190
David,180

And weight.txt contains:

David,175
John,180
Mary,165

For each of the following SQL queries, write a Spark script in Python to implement the query.

1. [5 points]

```
SELECT H.name, height, weight
FROM Height H, Weight W
WHERE H.name = W.name and height > 175 and weight < 178
```

```
lines = spark.read.text("height.txt").rdd.map(lambda r: r[0])          [2 points]
height  = lines.map(lambda x: (x.split(',')[0],('h',int(x.split(',')[1]))))
height  = height.filter(lambda x: x[1][1]>175)
lines = spark.read.text("weight.txt").rdd.map(lambda r: r[0])          [2 points]
weight = lines.map(lambda x: (x.split(',')[0],('w',int(x.split(',')[1]))))
weight = weight.filter(lambda x: x[1][1]<178)
height.join(weight).collect()                                          [1 point]
```

2. [5 points]

```
SELECT name, avg(height)
FROM Height
GROUP BY name
```

```
lines = spark.read.text("height.txt").rdd.map(lambda r: r[0])   [1 point]

height  = lines.map(lambda x: (x.split(',')[0],int(x.split(',')[1]))) [1 point]

h2 = height.aggregateByKey((0,0), lambda U,v: (U[0] + v, U[1] + 1), lambda U1,U2: (U1[0] + U2[0], U1[1] +
        U2[1]))                                                  [2 point]

h2.map(lambda x:(x[0],float(x[1][0])/x[1][1])).collect()          [1 point]
```