1. Start the PySpark console with CSV support

Load the Yelp dataset used in the lessons, using:

```
yelp_df = sqlCtx.load(source='com.databricks.spark.csv',

header = 'true',

inferSchema = 'true',

path = 'file:///usr/lib/hue/apps/search/examples/collections/solr_configs_yelp_demo/index_data
.csv')
```

Answer the following questions by running analytics functions on this DataFrame.

1. What is the mean of the "cool" column across all of the dataset?

- 0.991
- 0.993
- 0.998
- 0.996

2. Using again the Yelp dataset, take into consideration only the records with a "review count" of 10 or more. What is the average of the "cool" column for venues with 4 "stars"? (Hint: use grouping)

Choose the closest answer

- 1.095
- 1.078
- 1.067
- 0.985

3. Using again the Yelp dataset, take into consideration only the records with a "review count" of 10 or more and only records for which the venue is still open (see the "open" column).

What is the average of the "cool" column for venues with 5 "stars"?

Choose the closest answer

- 2.22
- 2.34
- 2.32
- 2.25

4. Using again the Yelp dataset, take into consideration only the records with a "review count" of 10 or more and only records for which the venue is still open (see the "open" column).

Count the records for each "state", which state has the 3rd highest number of reviews?

- CA
- CO
- LA
- GA

5. Using again the Yelp dataset, but taking into consideration the complete dataset, what is the maximum number of rows per venue (identified by "business_id")?

- 9
- 7
- 8
- 6

What to submit to your TA?

In addition to the correct option for the above five questions (which you need to choose from the given options), you need to submit the queries or code which you used to reach to the answer.