Lab 3 - Document Databases

- Due Dec 10, 2024 by 11:59pm
- Points 100
- Submitting a file upload

Updates, clarifications, and corrections:

- Friday, December 6 - Modified question 7 to ask how Jennifers and Jennys -vote- for other people's reviews as cool, useful, or funny, rather than the number of times that their own reviews have been voted this way. The question now more accurately reflects the data associated with each record in the *yelp.user* collection. Specifically, according to the yelp dataset's data dictionary, *yelp.user.useful* field indicates the number of times that this user (e.g. Jenny or Jennifer) has voted other people's reviews as useful. Likewise for *yelp.user.funny* and *yelp.user.cool*.

    -
      ```
      // integer, number of useful votes sent by the user
         "useful": 21,
      ```

Overview:

For this lab, you will use MongoDB to retrieve, transform, and analyze data stored in various MongoDB *Document Databases*. The structure of this lab is very similar to **lab 1 (https://canvas.cmu.edu/courses/42925/assignments/768066)** , but using MongoDB instead of relational databases. Read the instructions for each section carefully, especially the directions regarding the format and expectations for deliverables.

- Part 1 - Use MongoDB to write aggregation pipelines to answer the questions posed. Interpret, synthesize, and summarize the results in a clear, crisp, concise English-language statement.
- Part 2 - Complete two deeper data analysis tasks and present your insights and conclusions from those analyses in a clear, concise, and compelling one-page written document.
- ~~Part 3 (separate submission) - Individual reflection exercise. Details will be posted soon. In the meantime, do not complete this step until *your team* has completed the assignment, not just your individual part of it~~. Cancelled.

This lab will require a significant amount of work to complete. Dedicate time up-front as a team to budget your time and coordinate your efforts. Waiting to start the lab until a few hours before it's due will likely end poorly.

Prerequisites

To do this lab, you need access to the MongoDB databases on our class's shared Atlas database server (**access instructions (https://canvas.cmu.edu/courses/42925/pages/mongodb-tool-setup-instructions-2)**). Although I encourage you to use Studio3T as your GUI for working with MongoDB, you may use a different client to construct and test your queries.

Datasets:

You will work with the datasets listed below. All are available on the class's shared MongoDB Atlas instance.

Data dictionaries and further details about the datasets can be found at the links below for most of the datasets. Unlike lab 1, which included a set of standardized data dictionaries, the documentation for these datasets is provided in their "native" format. Use the documentation provided by the dataset's creator/publisher, along with your own analysis and investigation into what the dataset contains. Part of this assignment is figuring out how to interpret the dataset documentation in the format provided (or where there is no data dictionary, by digging through the contents of the provided dataset).

- Global cities (no documentation provided - interpreting and understanding the dataset's contents is part of the assignment)
- **Sample_airbnb (https://www.mongodb.com/docs/atlas/sample-data/sample-airbnb/#std-label-sample-airbnb)** (MongoDB version)
- **NBA (http://thecodebarbarian.com/2014/02/14/crunching-30-years-of-nba-data-with-mongodb-aggregation.html)**
- **Yelp (https://www.yelp.com/dataset)** (MongoDB version)

Part 1 - MongoDB Aggregation Pipelines

For part 1, you need to write aggregation pipelines to run on the shared MongoDB Atlas server to answer the questions posed.

Answer format: You need to submit a single javascript text file (.js suffix - this is the format that Studio 3T saves your queries in - containing all answers to the part 1 questions. Submitting your answers in a text/javascript (.js) file rather than a formatted text file such as MS Word, Google Docs, or PDF is important because the formatting information invisibly added by MS Word, Google Docs, PDF/Acrobat, etc. will often break your queries when I copy them and try to run them for testing. Which makes them wrong. Which does not lead to a good grade for you. Besides, if you want to be a Data Ninja, you will not just have to learn to use a text editor; you should develop strong opinions about why Text Editor A is better than Text Editor B. So find one you like and learn to wield it as the tool and weapon it is (pro tip - if you're using a Mac, **BBEdit** **(https://www.barebones.com/products/bbedit/)** is the best text editor available. Disagree? Those are fightin' words :-) See? Strong opinions about this topic are expected of data ninjas :-)

Use the following template to answer each question:

```
-- --------------------------
-- Question #
-- --------------------------
```

MongoDB aggregation pipeline you used to compute the answer to this question

Your brief written English-language answer interpreting the results as requested by the question.

Unless otherwise directed, to answer each question, provide the query you've written to answer the question, followed by a brief English-language statement interpreting the analytic question posed at the end of the question, based on the results returned by your pipeline. You should not need more than a couple of sentences to get full credit for your answer to these questions.

Do not paste the output from running your query into your submitted javascript/query file! We will test your aggregation pipelines by running them in Studio3T. The output for the queries tends to get quite bulky. Adding it to your submitted file makes it difficult to grade and makes your answers harder to find and read. **You will lose two points (-2) for each question in the lab where you ignore this requirement.**

For some questions, you will be given a precise spec regarding the format of the return documents. You need to match that format in your result set, though **the order of the keys within each document does not matter because you cannot control the order in which MongoDB returns keys *inside* a document**. However, returning *precisely* the specified keys in *precisely* the specified document structure is an important part of the task. Returning extra fields (including _id), missing requested fields, or returning incorrectly named fields will all result in points being deducted.

Part 1 questions:

Global Cities dataset*: For the first three questions, write aggregation pipelines that use the countries_states_cities collection to answer the following questions about places stored in the global_cities dataset.*

This dataset was derived from a dataset published on kaggle.com at ***https://www.kaggle.com/datasets/darshangada/countries-states-cities-database?rvi=1*** ***(https://www.kaggle.com/datasets/darshangada/countries-states-cities-database?rvi=1)*** *. Use the data dictionary provided on the linked Kaggle.com page and some exploration of the dataset on your own to interpret the contents and answer the question in this section.*

Question 1: (4 points) Write a query that counts the number of countries in each subregion and returns a list of all subregions that contain at least ten countries. Order your result set from the subregion with the most countries to the subregion with the fewest (but still at least ten). Your result set should have the following structure ('*Caribbean'* sub-region shown as an example:

```
  {
    "region" : "Americas",
    "number of countries" : NumberInt(28),
    "subregion" : "Caribbean"
 }
```

Analytic question: Which subregions have 15 or 16 countries?

Question 2: (4 points) Write an aggregation pipeline to retrieve the country, state, city name, latitude, and longitude of all cities in the world named "Rochester". Your pipeline should return a set of documents with the following structure (the document for Rochester, Michigan is shown as an example):

```
  {
    "country" : "United States",
    "state" : "Michigan",
    "city" : "Rochester",
    "latitude" : "42.68059000",
    "longitude" : "-83.13382000"
 }
```

Analytic question: How many states in the United States contain a city named "Rochester"? Which of those cities (Rochester's in the United States) is located furthest north?

Question 3: (4 points) Write an aggregation pipeline that identifies all countries in the world that use the United States Dollar (USD) as their currency. For each such country, your pipeline should produce a result set that identifies the name of the country, the official currency of that country, and the total number of cities listed in the dataset for that country.

Order the list of countries in your result set from the country containing the most cities down to the country with the fewest.

Analytic question - The United States has the most cities listed in this result set. Which country has the second most cities listed? How many cities are listed for that country?

Yelp dataset*: For the next four questions, write queries that use the yelp database on the shared Atlas server to answer the following questions about businesses and users stored in that dataset.*

This dataset was derived from a dataset published by Yelp at https://www.yelp.com/dataset. Yelp provides a data dictionary describing the dataset's contents at ***https://www.yelp.com/dataset/documentation/main*** ***(https://www.yelp.com/dataset/documentation/main)*** *. An important part of this assignment is exploring the dataset, reading the data dictionary, and figuring out how to retrieve the correct information from it to do your calculations.*

Question 4: (4 points) Write an aggregation pipeline that identifies all cities in California with at least ten restaurants that offer outdoor seating. For each such city, your pipeline should list the number of restaurants there that offer outdoor seating. Your pipeline should return a document set where each such city is listed with the following structure (Santa Barbara shown as an example):

```
  {
      "Number of restaurants with outdoor seating" : NumberInt(510),
      "City" : "Santa Barbara"
  }
```

Order your result set from the California city with the most restaurants offering outdoor seating downwards.

Analytic question: What common geographic trait do you notice about cities in California that have more than ten restaurants offering outdoor seating?

Question 5: (4 points) Write an aggregation pipeline that retrieves all restaurants in the city of Philadelphia, PA that meet the following criteria. (Words in "quotes" are the tags you need to match against):

- Classified as a "Restaurant" that serves "Dim Sum"
- Listed as "GoodForKids"
- Is "WheelchairAccessible"
- Has at least a 4.0 star rating and at least 250 reviews

For each restaurant that meets that criteria, return a document with the following structure:

```
    {
      "Restaurant name" : "Dim Sum Garden",
      "Full address" : "1020 Race St, Philadelphia, PA, 19107",
      "Star rating" : NumberInt(4)
    }
```

Your query must return the full address (street address, city, state, postal code) as a single attribute called "Full address" rather than simply returning each part of the address as an independent field.

Analytic question: Based on the results returned, which zip code in Philadelphia is most likely to contain Philadelphia's "Chinatown" neighborhood?

Question 6: (4 points) Write a query that calculates the number of businesses in the yelp dataset in each state that are categorized with both "Yoga" and "Meditation Centers" tags. Your query should return results for the five states with the largest number of businesses categorized with *both* tags, ordered from the state with the most such businesses down to the fifth most.

For each state returned, your result set should contain a document in the following format (Pennsylvania shown as an example):

```
        {
          "Number of Yoga and Meditation Centers" : NumberInt(15),
          "State" : "PA"
        }
```

Analytic question: Based on these results, do there appear to be more Yoga and Meditation Centers in Florida or California?

Question 7: (5 points) Write an aggregation pipeline to compare how users who list their name as "Jennifer" write and respond to reviews relative to users who list their names as "Jenny". To do so, your aggregation pipeline should return a result set that helps you determine the following about Jennifers and Jenny's:

- The total number of users in the dataset named "Jennifer" and "Jenny".
- The total number of reviews that reviewers with each of these names have written.
- On average, do users named Jennifer or Jenny write more reviews?
- Are Jennifers or Jenny's more likely to vote other people's reviews *useful*, *cool*, or *funny*?

Analytic question - provide a brief, clear, and concise 1-2 sentence English language statement that answers the questions above, using the data from your query results to support your answers.

> Note - This is the only query that should take more than one second to run with an efficient query. The solution I came up with takes about 10 seconds to complete. If you find that your answer is taking a lot longer than 30 seconds or so (ie, taking minutes), first make sure you are using the *yelp.users* collection to calculate your results. Second, make sure that you are filtering out all documents that are not for users named "Jennifer" or "Jenny" in an early $match stage (preferably the first stage of your pipeline) so that you don't need to do these calculations for all users, just for Jenny's and Jennifer's. Doing so will dramatically reduce the number of documents your aggregation pipeline needs to evaluate in subsequent stages, which should speed things up dramatically.

***Airbnb dataset:** Use the* ***[Sample_airbnb](https://www.mongodb.com/docs/atlas/sample-data/sample-airbnb/#std-label-sample-airbnb)** (https://www.mongodb.com/docs/atlas/sample-data/sample-airbnb/#std-label-sample-airbnb) dataset to answer the following questions.*

Question 8: (5 points) Write a query that calculates the highest, lowest, and average nightly price, along with the average cleaning fees for each of the following Airbnb markets (*address.market*): Barcelona, Porto, Sydney, New York, Montreal, and Istanbul.

Your result set should contain a set of documents with the following format (Istanbul is shown as an example). Order your results alphabetically by market name.

```
    {
        "Highest nightly price" : NumberDecimal("48842.00"),
        "Average nightly price" : NumberDecimal("367.94"),
        "Lowest nightly price" : NumberDecimal("26.00"),
        "Average cleaning fee" : NumberDecimal("82.53"),
        "Market" : "Istanbul"
    }
```

Analytic question - based purely on the results of this query, which of these markets seems to be "the most affordable"? Why?

Question 9: (5 points) Find the three hosts in this dataset with the most listings in the Porto market. For each of those hosts, list their name, host_id, number of listings, the average nightly price of their listings, and the host's host_location (which may not be the same as the location of the listings).

The documents in your result set should have the following structure and format (example result shown). Order your results from the host with the most listings in Porto downward.

```
      {
        "Host name" : "Feels Like Home",
        "Host location" : "Lisbon, Lisboa, Portugal",
        "Average nightly price" : NumberDecimal("61.14"),
        "Number of listings in Porto" : NumberInt(7),
        "Host ID" : "3953109"
      }
```

Analytic question - based on the results of this query, does the Airbnb rental market in Porto appear to be dominated by a small number of large rental companies, or does it appear to mostly consist of landlords with a small number of rentals listed?

Question 10: (5 points) Retrieve data about all reviewers who have written five or more reviews for listings in the Montreal market. For each of these reviewers, list the number of reviews they have written, along with their reviewer_id and name.

Your result set should have the following structure (shown here for "Dan"). Order your results from the reviewer who has written the most reviews for listings in Montreal downward.

```
      {
        "Reviewer name" : "Dan",
        "Number of reviews" : NumberInt(6),
        "Reviewer ID" : "34005800"
      }
```

Analytic question - how many reviewers meet these criteria?

Remember that multiple reviewers may have the same name, but each person writing reviews has a unique reviewer_id that should be consistent from review to review. Calculate accordingly.

> Hint #1 - you will probably need to have an $unwind step in your pipeline somewhere to calculate this properly.
>
> Hint #2 - It's easiest to do the calculation if you restructure and simplify the documents flowing through the pipeline after you've done your $unwind but before you $group and aggregate

NBA dataset: *Use the* [***NBA***](http://thecodebarbarian.com/2014/02/14/crunching-30-years-of-nba-data-with-mongodb-aggregation.html) ***(http://thecodebarbarian.com/2014/02/14/crunching-30-years-of-nba-data-with-mongodb-aggregation.html)*** *dataset to answer the following questions.*

Question 11: (5 points) Write an aggregation pipeline to determine whether Kobe Bryant or Tim Duncan played in more NBA games during the dates covered by this dataset.

Analytic question - state the answer to this question in one sentence.

Question 12: (5 points) Write an aggregation pipeline that lists all games played during the 2003-2004 NBA season (July 1, 2003 - June 30, 2004) in which the losing team scored less than 70 points. For each such game, your query should return the date of the game, the two teams that played, and the final score.

Analytic question - how frequently does your query suggest that teams score less than 70 points in an NBA basketball game? Find the total number of total games played in a typical NBA season to use as a baseline to answer this question.

Question 13: (5 points) There is a perception amongst NBA fans that the team playing on their home court has an advantage. Evaluate whether such a home-court advantage exists, and if so, how strong it is in the NBA. To do so, write a query to calculate the percentage of all games won by the home team between July 1, 1995, and June 30, 2005. Does the result indicate a strong advantage, a weak advantage, no difference, a weak disadvantage, or a strong disadvantage to playing at home in the NBA?

Analytic question - Your answer to this question should be a short statement that answers the question posed in no more than a few sentences. This statement should clearly state your conclusion and support that answer through rigorous data analysis. Following your English-language statement, provide the query(ies) you used to perform your calculations.

Question 14: (5 points) Looking only at games played in the 2000's (January 1, 2000 through December 31, 2010, inclusive), what is the likelihood that a team with the higher number of steals in a given game won that game, across all of the games in the dataset?

Note: The number of steals for each team in a given game is stored as an integer value in the field box > team > stl.

Analytic question - Your answer to this question should be a short statement that answers the question posed in no more than a few sentences. This statement should clearly state your conclusion and support that answer through rigorous data analysis. Following your English-language statement, provide the query(ies) you used to perform your calculations.

Part 1 deliverables and submission instructions:

Each group should submit one Javascript text file (.js) named *Lab 3 Part 1 Queries - group XX.js* containing the answers to all part 1 questions. Replace the *XX* with your group number.

To grade your lab, I will copy and paste your queries from this file to test them on the shared MongoDB Atlas instance. Submissions of query statements and results submitted in plain text format are much easier to grade in this manner, as cutting and pasting from an MS Word or PDF document often introduces odd (and sometimes invisible) characters that can mess up the query, causing it to either fail outright or to return the wrong values, neither of which are good outcomes for grading your assignment. So please submit your queries and results in a plain text file with a .txt or .js suffix. Simply saving the file you edit in Studio3T to run your queries as *Lab 3 Part 1 Queries.js* is sufficient to create the correct file format.

Grading Criteria: We will use the grading criteria in the attached rubric to assess your answer to each question in part 1.

Part 2 - Brief analytic reports

For the second part of this assignment, you will revisit the one-page analytic brief exercise we did in the first lab. You will be given a similar template and a pair of broad analytic tasks. You need to refine and frame the analytic questions, analyze the data to answer them, and present your insights and conclusions from that analysis in clear, concise, single-page briefs.

This MS Word template file contains the questions you need to answer and templates for your briefs.

> **MDM M2-24 Lab 3 Part 2 - Analysis Questions.docx (https://canvas.cmu.edu/courses/42925/files/12011614?wrap=1) (https://canvas.cmu.edu/courses/42925/files/12011614/download?download_frd=1)**

As described in the template, you will need to retrieve and analyze data in the assigned datasets to generate a compelling answer to the question posed. These questions are much more open-ended than the questions in Part 1. This assignment is less about writing individual SQL queries and more about figuring out how to structure your analysis, where the data that you need to use for that analysis lives in the datasets, how to complete the analyses in a statistically appropriate way, and how to present those results in a concise, clear, and compelling way.

The first page of the template file provides details on what is expected. Read the instructions and the grading rubric attached to this Canvas assignment carefully to ensure you approach the assignment appropriately.

Each analysis brief is worth 16 points (32 points total for part 2)

Part 2 deliverables and submission instructions:

Upload a single PDF file containing your analyses. The one-page limit for presenting your analysis results and conclusions is strict. When grading, we will stop reading after the first page of your response for each question and grade your submission based only on what you have presented in a single page (though we may look at the appendix to see how you derived your answers).

Grading criteria: We will grade this part of the lab using the criteria in the Canvas assignment rubric.

Part 3 - Individual reflective exercise

~~As with the second lab, reflective essays should be written and submitted individually through a separate Canvas Assignment.~~

Part 3 of this lab has been canceled. There is no need to hand anything in for it.

Score calculations: Parts 1 and 2 of this lab are worth 95 points. To calculate each student's score for the lab, your individual score for your reflective essay (up to 5 points) will be added to your team score (of up to 95 points) for a total of 100 possible points.

MDM Lab 3 Rubric (M2-24)

Criteria	Ratings				Pts
Part 1, Q1	**4 to >3.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is clear, concise, gramatically correct, and easily understood.	**3 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is clearly written.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, poorly written, or incorrect analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	4 pts
Part 1, Q2	**4 to >3.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is clear, concise, gramatically correct, and easily understood.	**3 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is clearly written.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, poorly written, or incorrect analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	4 pts
Part 1, Q3	**4 to >3.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is clear, concise, gramatically correct, and easily understood.	**3 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is clearly written.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, poorly written, or incorrect analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	4 pts
Part 1, Q4	**4 to >3.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is clear, concise, gramatically correct, and easily understood.	**3 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is clearly written.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, poorly written, or incorrect analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	4 pts
Part 1, Q5	**4 to >3.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is clear, concise,	**3 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is clearly written.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, poorly written, or incorrect analytic statement can	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	4 pts

Criteria	Ratings				Pts
	gramatically correct, and easily understood.		also bring the score for this question down to this level.		
Part 1, Q6	**4 to >3.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is clear, concise, gramatically correct, and easily understood.	**3 to >2.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is clearly written.	**2 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, poorly written, or incorrect analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	4 pts
Part 1, Q7	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is correct, clear, concise, gramatically correct, and easily understood.	**4 to >3.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is correct and clearly written.	**3 to >0.0 pts** **Materially incorrect** The query has significant problems even though it may be "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, incorrect, or poorly written analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Q8	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is correct, clear, concise, gramatically correct, and easily understood.	**4 to >3.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is correct and clearly written.	**3 to >0.0 pts** **Materially incorrect** The query has significant problems even though it may be "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, incorrect, or poorly written analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Q9	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is correct, clear, concise, gramatically correct, and easily understood.	**4 to >3.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is correct and clearly written.	**3 to >0.0 pts** **Materially incorrect** The query has significant problems even though it may be "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, incorrect, or poorly written analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Q10	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is correct, clear,	**4 to >3.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is correct and clearly written.	**3 to >0.0 pts** **Materially incorrect** The query has significant problems even though it may be "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, incorrect, or poorly written analytic statement can	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts

Criteria	Ratings				Pts
	concise, gramatically correct, and easily understood.		also bring the score for this question down to this level.		
Part 1, Q11	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is clear, concise, gramatically correct, and easily understood.	**4 to >3.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is clearly written.	**3 to >0.0 pts** **Materially incorrect** The query has significant problems even though it is "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, poorly written, or incorrect analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Q12	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is correct, clear, concise, gramatically correct, and easily understood.	**4 to >3.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is correct and clearly written.	**3 to >0.0 pts** **Materially incorrect** The query has significant problems even though it may be "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, incorrect, or poorly written analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Q13	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is correct, clear, concise, gramatically correct, and easily understood.	**4 to >3.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is correct and clearly written.	**3 to >0.0 pts** **Materially incorrect** The query has significant problems even though it may be "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, incorrect, or poorly written analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 1, Q14	**5 to >4.0 pts** **Exactly correct** The query runs perfectly, returns exactly what was requested, and would return exactly what was requested even if additional data was inserted, updated, or removed from the dataset. The analytic statement is correct, clear, concise, gramatically correct, and easily understood.	**4 to >3.0 pts** **Almost correct** The query runs almost correctly. Small errors lead to a slightly incorrect result set, but the values calculated and retrieved are materially correct. The analytic statement is correct and clearly written.	**3 to >0.0 pts** **Materially incorrect** The query has significant problems even though it may be "directionally correct". The result set returned has errors that make the values calculated or retrieved materially incorrect. A missing, incorrect, or poorly written analytic statement can also bring the score for this question down to this level.	**0 pts** **Not attempted or substantially incorrect** The query is substantially incorrect, or was not attempted at all.	5 pts
Part 2 - Airbnb analysis - problem framing and metric specification	**6 to >4.0 pts** **Sophisticated and nuanced** Findings are presented in a clear, concise, and compelling manner. The single-sentence answer is direct, insightful, and fully supported by the presented evidence. Visualizations are effective, well-chosen, and easy to interpret, enhancing the clarity	**4 to >2.0 pts** **Meets baseline expectations** Findings are presented adequately, with the main points understandable. The single-sentence answer is present and generally supported by the evidence. Visualizations are included and generally interpretable.	**2 to >0 pts** **Not yet up to expectations** Findings are presented poorly, making it difficult to understand the conclusions. The focused statement answering the question is missing, weak, or unsupported by the evidence. Visualizations are missing, unclear, or ineffective.		6 pts

Criteria	Ratings			Pts
	and impact of the findings. Excellent use of whitespace and formatting.			
Part 2 - Airbnb analysis - data retrieval, calculation, and analysis	**6 to >4.0 pts** **Sophisticated and nuanced** The analysis is thorough and accurate, employing appropriate techniques and calculations. Considers potential confounding factors or limitations of available data. Calculations are accurate and provided code runs properly, generating results presented. Demonstrates critical thinking and insightful interpretation of the data. Addresses potential limitations of the data available and metrics chosen.	**4 to >2.0 pts** **Meets baseline expectations** The analysis is generally accurate and addresses the question, but may have minor errors or omissions. Calculations are broadly correct, but may not demonstrate a subtle and nuanced understanding of the dataset or limitations of the data it contains. Provided code runs correctly and generates results presented. Analysis presents a straightforward analysis of the data and interpretation of metrics.	**2 to >0 pts** **Not yet up to expectations** The analysis is flawed, inaccurate, or incomplete. Calculations and/or code are incorrect or do not produce the results presented. Demonstrates a significant misunderstanding of the data or analytical techniques. Fails to address limitations in data or metrics.	6 pts
Part 2 - Airbnb analysis - presentation, writing, and visualization(s)	**6 to >4.0 pts** **Sophisticated and nuanced** Findings are presented in a clear, concise, and compelling manner. The concise statement answering the question is direct, insightful, and fully supported by the presented evidence. Visualizations are effective, well-chosen, and easy to interpret, enhancing the clarity and impact of the findings. Excellent use of whitespace and formatting.	**4 to >2.0 pts** **Meets baseline expectations** Findings are presented adequately, with the main points understandable. The single-sentence answer is present and generally supported by the evidence. Visualizations are included and generally interpretable.	**2 to >0 pts** **Not yet up to expectations** Findings are presented poorly, making it difficult to understand the conclusions. The concise answer statement is missing, weak, or unsupported by the evidence. Visualizations are missing, unclear, or ineffective.	6 pts
Part 2 - NBA analysis - problem framing and metric specification	**6 to >4.0 pts** **Sophisticated and nuanced** Findings are presented in a clear, concise, and compelling manner. The single-sentence answer is direct, insightful, and fully supported by the presented evidence. Visualizations are effective, well-chosen, and easy to interpret, enhancing the clarity and impact of the findings. Excellent use of whitespace and formatting.	**4 to >2.0 pts** **Meets baseline expectations** Findings are presented adequately, with the main points understandable. The single-sentence answer is present and generally supported by the evidence. Visualizations are included and generally interpretable.	**2 to >0 pts** **Not yet up to expectations** Findings are presented poorly, making it difficult to understand the conclusions. The focused statement answering the question is missing, weak, or unsupported by the evidence. Visualizations are missing, unclear, or ineffective.	6 pts
Part 2 - NBA analysis - data retrieval, calculation, and analysis	**6 to >4.0 pts** **Sophisticated and nuanced** The analysis is thorough and accurate, employing appropriate techniques and calculations. Considers potential confounding factors or limitations of available data. Calculations are accurate and provided code runs properly, generating results presented. Demonstrates critical thinking and insightful interpretation of the data. Addresses potential limitations of the data available and metrics chosen.	**4 to >2.0 pts** **Meets baseline expectations** The analysis is generally accurate and addresses the question, but may have minor errors or omissions. Calculations are broadly correct, but may not demonstrate a subtle and nuanced understanding of the dataset or limitations of the data it contains. Provided code runs correctly and generates results presented. Analysis presents a straightforward analysis of the data and interpretation of metrics.	**2 to >0 pts** **Not yet up to expectations** The analysis is flawed, inaccurate, or incomplete. Calculations and/or code are incorrect or do not produce the results presented. Demonstrates a significant misunderstanding of the data or analytical techniques. Fails to address limitations in data or metrics.	6 pts
Part 2 - NBA analysis - presentation, writing, and visualization(s)	**6 to >4.0 pts** **Sophisticated and nuanced** Findings are presented in a clear, concise, and compelling manner. The concise statement answering the question is direct, insightful, and fully supported by the presented evidence. Visualizations are effective, well-chosen, and easy to interpret, enhancing the clarity and impact	**4 to >2.0 pts** **Meets baseline expectations** Findings are presented adequately, with the main points understandable. The single-sentence answer is present and generally supported by the evidence. Visualizations are included and generally interpretable.	**2 to >0 pts** **Not yet up to expectations** Findings are presented poorly, making it difficult to understand the conclusions. The concise answer statement is missing, weak, or unsupported by the evidence. Visualizations are missing, unclear, or ineffective.	6 pts

Criteria	Ratings			Pts
	of the findings. Excellent use of whitespace and formatting.			
				Total Points: 100