DA5020 – Practicum II

This assignment provides you with an opportunity to use the MongoDB document based data store. MongoDB is a commonly used non-relational database. In this practicum you will practice using MongoDB in the cloud[1]. Register for a free MongoDB Atlas account. Ensure that you select the free option (no credit card is needed). After registration is completed, create a free cluster (this will be done in question 1).

Alternatively, you can setup mongoDB on your computer. This will require software installations, command line servers, and keys to be installed. Have patience. However, it is recommended that you use MongoDB Atlas — which is free and does not require installing software on your computer.

This is a group practicum which means that you may choose to work in groups of up to three students. You may fully collaborate and submit the same work. However, you must put all students' names on all submitted work. If a group member is not adequately contributing, the remaining team members may "vote to eject" the student from the team by emailing me the reason. In such an event, the team member who was "fired" must still complete the project individually by the due date.

If you are working in groups, you can self-signup in Canvas using one of the Practicum 2 groups or notify me via email by Wednesday and I will create the group for you. Ensure that you include your name and the name(s) of your group member(s) in the email and cc them.

In this practicum, you will work with data from the Bureau of Transportation Statistics which reports the Airline/Carrier On-Time Performance from 1987 to present. The dataset contains airlines, which are also called operators, that voluntarily share its data and notify the organization about its flight hours, distance, arrival and departure delays at various airports across the US and its territories. For more information on the dataset and to learn more about the description of the fields, visit the following website.

You are given the task of performing a comprehensive analysis of this dataset. However, due to the volume of data, a subset has been extracted for you for all flights that originated or departed from the following states in 2019: Arizona: AZ, Nevada: NV and California: CA.

*You will need to Install the **mongolite** library in your R environment.*

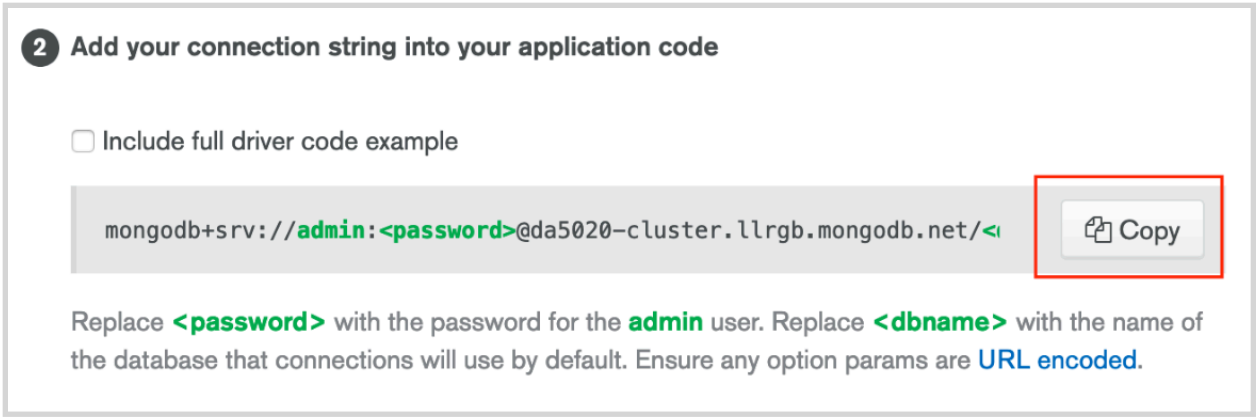
Question 1 — Configure the Database: MongoDB Atlas (10 points)

1. Create a free account with MongoDB Atlas
2. Watch the following tutorial entitled: Getting Started with MongoDB Atlas - Free Tier to create a free cluster and mongoDB will be installed automatically for you. When you watch the video, pay attention to the following terms: **cluster name**, **database user**, **network access** and **connection information**. You will need to perform the following when you create your cluster:
    - **Cluster name:** Ensure that your cluster is called **da5020-cluster**.

---

[1] You can perform this on your local machine. However, the cloud gives you the flexibility to scale your analysis (e.g use more or less computing resources for large-scale data sources). Additionally, you will only need to create one database for your group which you can all access at a central location.

- **Database User:** Create a database user[2] and password so that you can access the cluster. This is different from the credentials for your mongoDB Atlas account.
  - If you work in a group, create a database user for each group member. You only need to create one cluster for your group; however, each group member is expected to use their respective username and password to access the database.
  - Create a database user for the TA with the following information username: **da5020_ta** and password: **da5020NEU**; grant them **read and write access to any database**.
  - **Take a screenshot of the database users that were created and submit it with the practicum.**
- **Network Access:** In the Mongo DB Atlas Dashboard, click Network access => IP Address and select "Allow Access From Anywhere". This setting is not very secure but is only required for grading purposes to allow the TAs access to your database (and you won't need to whitelist their IP address) .
- **Connection String:** Get your connection information so that you can connect to your database in R studio. In the Mongo DB Atlas Dashboard, click clusters => connect => connect to your application. You will be presented with information on how to connect to your cluster. Copy the connection string (view the image below). This will be your url to connect to the database in R. Open R studio and create a variable called mongo_url and paste the connection string. Ensure that you replace **<password>** with your actual password and **<dbname>** with the name of the database that you will create in question 2.



Note: if you decided to install mongoDB on your computer, skip steps 1 and 2 above and follow the instructions at this link to install the necessary software.

3. Create the database and load the data.

Open R studio, connect to your mongoDB instance and create a database called **airline_performance** and a collection called **flights_2019**. You will need to use your connection string from Question 1.2 above. After which, insert the attached CSV data *"2019_ONTIME_REPORTING_FSW.csv"* using the insert

[2] Ensure that you will remember the username and password for the database user, because you will use it in your R environment. Note: the credentials for the database user is different from your MongoDB Atlas login credentials.

function from the mongolite package. It will take a few minutes to insert all the data because it contains approximately 2 million observations.

Using R Studio, create queries using mongoDB (via the mongolite library) to answer the remaining questions. Ensure that you write queries to answer each question below and do not load the entire dataset in your R environment.

Question 2 — (30 points)

Let's explore patterns in the region. For each of the original 3 states (i.e. AZ, NV, CA), analyze the most popular outbound/destination airports. For example, if a flight originated in CA (at any of its airports), where do they often go? Comment on your findings and visualize the top results.

Question 3 — (30 points)

Let's explore the carriers. Calculate the total flights for each airline/operator.

a. Ensure that you indicate the full name of each carrier, in lieu of the carrier code. *This will require web scraping*. Here is a helpful resource with the list of airline codes and the respective names. You can also use an alternative webpage of your choice.

b. Visualize the **top 10 results** and show the carrier name and the frequency. Explain the results.

Question 4 — (15 points)

Select the top 5 airlines, from the previous question, and calculate the ***total flight hours*** for each month (grouped by airline). Explain and visualize the results. Hint: the total flight hours is not equivalent to the frequency of flights and ensure that you display the total hours and not the total minutes.

Question 5 — (15 points)

Select any (1) aircraft, and explore the data to determine where it often travels. Calculate its average arrival and departure delays at the airports. After which analyze all the results to identify any patterns that are evident and also indicate which airline operates that aircraft. Explain your findings and visualize the results. Note: the TAIL_NUM can help you to identify each unique aircraft.

Question 6 — (+10 optional/bonus points)

Build one additional query to test a hypothesis or answer a question that you have about the dataset. Your query should retrieve data from MongoDB and evaluate the pattern/trend. Prepare supporting visualizations for your analysis. If necessary, you can integrate any additional data that provide more details or support your analysis/findings.

Note: all charts that are displayed should have the following:

- An informative title (and subtitle if applicable)
- Labels on the x-axis and y-axis that indicate the units of measurement.

- A caption that indicates the purpose of the chart.

Submission Details

- Your submission must contain three files: the .Rmd file, a knitted PDF or HTML and a screenshot of the database users. Name your .Rmd file, DA5020.P2.FirstName.LastName.Rmd and your PDF/HTML DA5020.P2.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name.
- The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it). Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.
- Not submitting a knitted PDF or HTML will result in reduction of 30 points.
- Not submitting the .Rmd file (or both) will result in a score of 0.
- **Include the names of all group members in your RMD file.**

Useful Resources

- MongoDB Atlas
- Getting Started with MongoDB Atlas - Free Tier
- Install mongo DB on your computer
- Mongo DB Atlas and R
- Query the days of the week in MongoDB
- List of airline codes
- Mapping the US
- Advanced Mapping
- Bureau of Transportation Statistics
- Data dictionary

** Note: Do not merely print the output for each question but always explain the results from your code. Points will be deducted for any missing explanation and any visualizations that are not clearly labeled.