In this activity, you will use the Bitnami *containers* that you created in [Video 19.3](https://classroom.emeritus.org/courses/10605/pages/creating-spark-docker-images-and-containers-05-21), the data that you loaded in [Video 19.4](https://classroom.emeritus.org/courses/10605/pages/loading-flight-data-into-the-spark-docker-container-02-52), and PySpark to write SQL *queries*. The goal of this activity is for you to gain further insight into your data about the delays between flights across the United States.

Note that before you begin this activity, you will need to complete all of the steps demonstrated in the previous videos.

Before beginning this activity, review the submission instructions below to ensure that you collect the required screenshots as you progress through the activity.

**To complete this activity, follow these steps:**

1. Open Docker Desktop on your machine. Provide a screenshot to show that you have the Bitnami *containers* up and running.
2. Open the CLI for the bitnami\_spark\_1 *container*. Navigate to the root folder. Provide a screenshot to show that the departuredelays.csv file is in the root folder.
3. In the CLI window, type the correct command to open PySpark. Provide a screenshot to show that you successfully opened PySpark.
4. In the CLI window, type the correct command to import the correct *package* to start a PySpark session. Provide a screenshot to show that you successfully started a PySpark session.
5. In the CLI window, type the correct command to define a PySpark session named spark. Set the appName *argument* equal to Activity19.3. Provide a screenshot to show that you successfully defined the spark PySpark session.
6. In the CLI window, define a variable, activity19\_3\_data, to hold the path to the departuredelays.csv file. Provide a screenshot to show that you successfully defined the activity19\_3\_data variable.
7. In the CLI window, define a *dataframe*, df, that will contain all of the entries in the departuredelays.csv file. Provide a screenshot to show that you successfully defined the df *dataframe*.
8. In the CLI window, on the df *dataframe*, use the createOrReplaceTempView *method* to create a view of the *dataframe*. Name the view activity19\_3\_table. Provide a screenshot to show that you successfully created a view of the *dataframe* that is named activity19\_3\_table.
9. In the CLI window, type an SQL *query* to select the first 10 flights from John F. Kennedy International Airport (JKF) to Seattle-Tacoma International Airport (SEA) that had a delay of greater than 90 minutes. Provide a screenshot to show that you selected the correct entries from your data. Your data should display the first 10 flights from JFK to SEA that had a delay of greater than 90 minutes.
10. In the CLI window, type an SQL *query* to select the first five flights from San Francisco International Airport (SFO) to Miami International Airport (MIA) that had a delay of less than 60 minutes. Provide a screenshot to show that you selected the correct entries from your data. Your data should display the first five flights from SFO to MIA that had a delay of less than 60 minutes.

**Submission Instructions:**

Your submission for this activity should be a Word document that includes the following screenshots, each labeled for the step that the screenshot represents:

1. Provide a screenshot to show that you have the Bitnami *containers* up and running.
2. Provide a screenshot to show that the departuredelays.csv file is in the root folder.
3. Provide a screenshot to show that you successfully opened PySpark.
4. Provide a screenshot to show that you successfully started a PySpark session.
5. Provide a screenshot to show that you successfully defined the spark PySpark session.
6. Provide a screenshot to show that you successfully defined the activity19\_3\_data variable.
7. Provide a screenshot to show that you successfully defined the df *dataframe*.
8. Provide a screenshot to show that you successfully created a view of the *dataframe* that is named activity19\_3\_table.
9. Provide a screenshot to show that you selected the correct entries from your data. Your data should display the first 10 flights from JFK to SEA that had a delay of greater than 90 minutes.
10. Provide a screenshot to show that you selected the correct entries from your data. Your data should display the first five flights from SFO to MIA that had a delay of less than 60 minutes.