Homework Problem Set E Submission Form

# Overview

| Your Name | Nick Videtti |
| --- | --- |
| Your SU Email | nvidetti@syr.edu |

# Instructions

Put your name and SU email at the top. Answer these questions all from the lab. When asked to include screenshots, please follow the screenshot guidelines from the first homework.

Remember as you complete the homework that it is not only about getting it right/correct. We will discuss the answers in class so it’s important to articulate anything you would like to contribute to the discussion in your answer:

* If you feel the question is vague, include any assumptions you've made.
* If you feel the answer requires interpretation or justification, provide it.
* If you do not know the answer to the question, articulate what you tried and how you are stuck.
* Highlight any doubts or questions you would like me to review.

This how you receive credit for answering questions that might not be correct. In addition, you must complete the reflection portion of the homework assignment for full credit. Since most answers will be similar this is an important part of your individual submission.

Complete Part II of this document first, then go back and complete the Reflection in Part I.

# Part I: Reflection

Use this section to reflect on your learning. To achieve the highest grade on the assignment, you must be as descriptive and personal as possible with your reflection.

1. As you completed this assignment, identify what you learned.

**This assignment taught me how to troubleshoot Drill storage plugins, how to use the SPLIT and SPLIT\_PART functions, and how to reference columns with spaces in Spark SQL. It also gave me a refresher on some Python stuff during the last exercise, and that helped me get better with Spark and Pandas as well.**

1. What barriers or challenges did you encounter while completing this assignment?

**There were some issues with services that needed to be fixed, but my professor was luckily extremely quick and responsive and helped me fix that issue. I also had some troubles finding out how to configure the Drill storage plugin to extract headers and how to reference columns in Spark SQL after underscores in the column names were replaced with spaces. Lots of Googling was required for the last exercise, and other Googling was needed during some of the other exercises as well.**

1. How prepared were you to complete this assignment? What can you do to be better prepared?

**I probably took a little too much time between completing the asynchronous work and starting this assignment, but that could arguably be a good thing since I was required to relearn and revisit each topic while completing this assignment.**

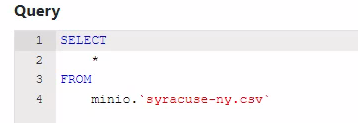
1. Rate your comfort level with this week’s material. Use the rubric provided.

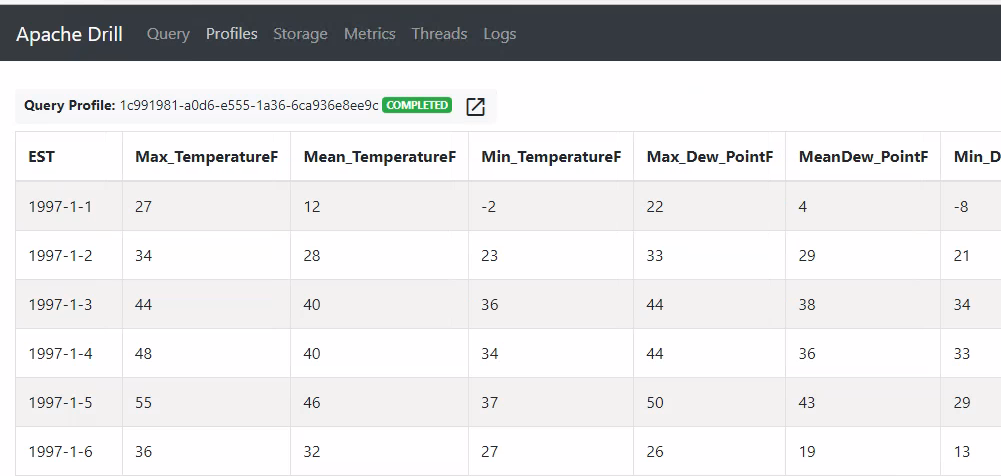
4 ==> I understand this material and can explain it to others.  
**3 ==> I understand this material.**  
2 ==> I somewhat understand the material but sometimes need guidance from others.  
1 ==> I understand very little of this material and need extra help.

# Part II: Questions

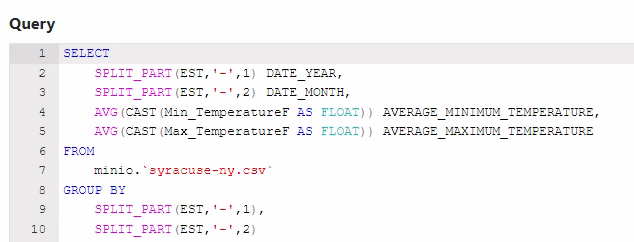
**For each question, include a copy of the code required to complete the question along with a screenshot of the code and a screenshot of the output.**

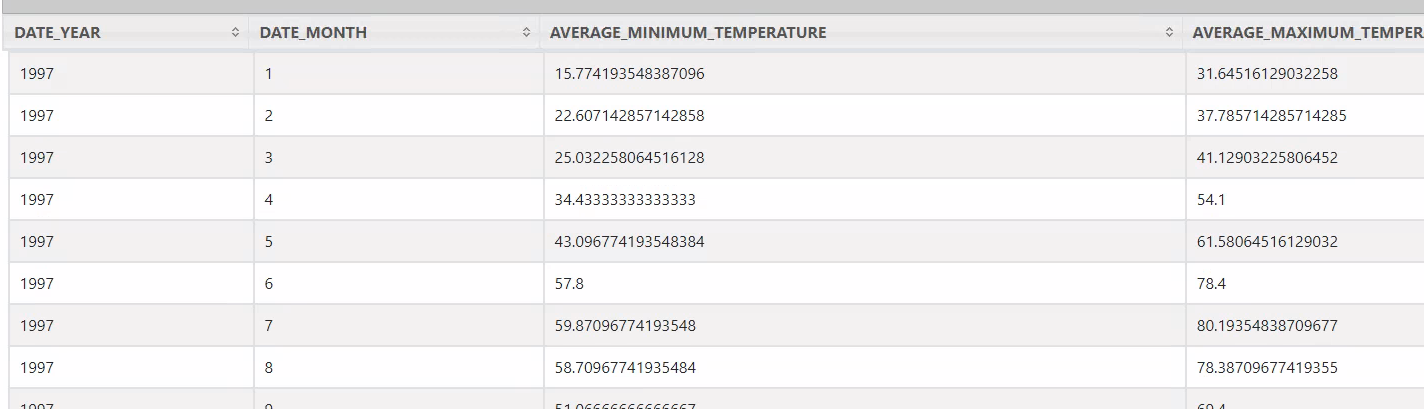
1. Configure a Drill storage plugin for the Minio **labe** bucket. Then write a Drill query for **syracuse-ny.csv** to demonstrate you can read the file with headers.



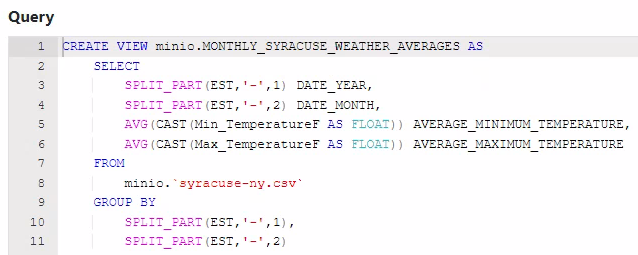


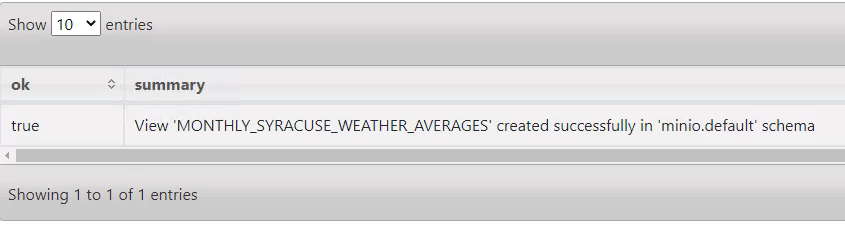
1. Write a Drill SQL query to get the overall average min and max temperatures by year and month. Use Drill’s SPLIT() function to separate Year, Month. You might need to use cast() to ensure the min and max temperatures are numeric types. You output should include four columns: Year, Month, the average minimum temperature for that month, and the average maximum temperature for that month.



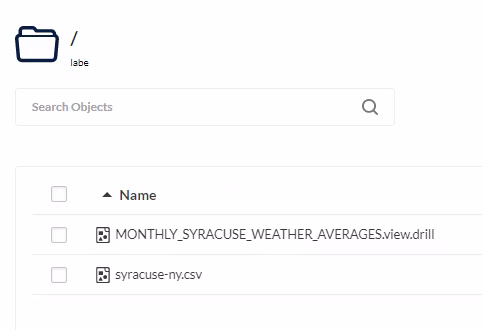


1. Create a view called **monthly\_syracuse\_weather\_averages** from the query you wrote in Question 2 and store it back on the **labe** bucket. (If you cannot get Question 2 working, use a similar query.) Provide your Drill SQL code and a screenshot showing the view file is on the Minio bucket.  
   NOTE: If you get an error about an immutable object, you need to change your storage config so you can write to the storage location.

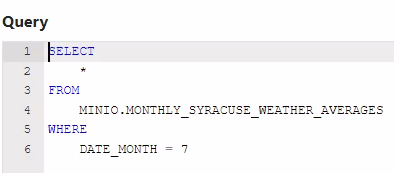


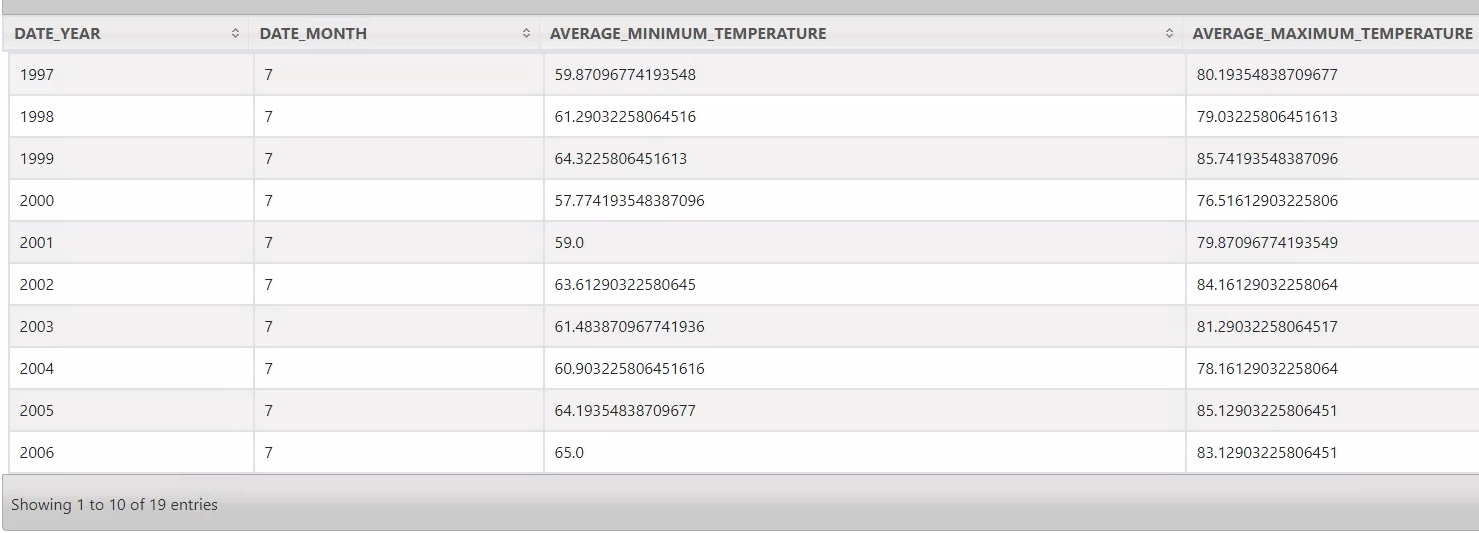




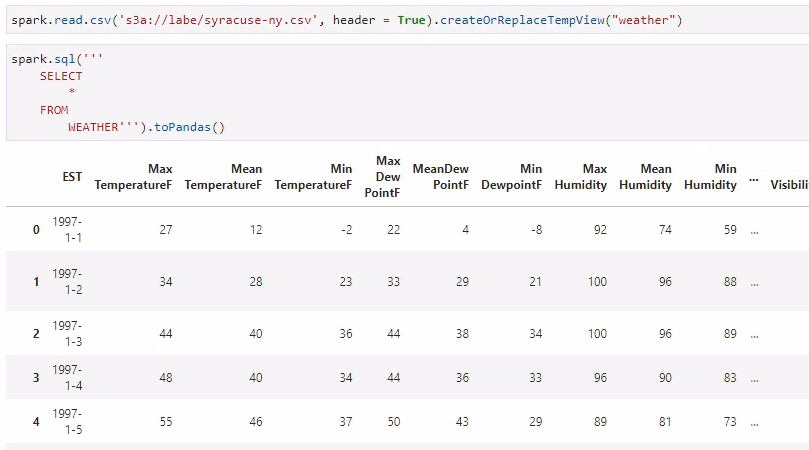


1. Use the view you created in Question 3 to show the weather data for only the month of July.

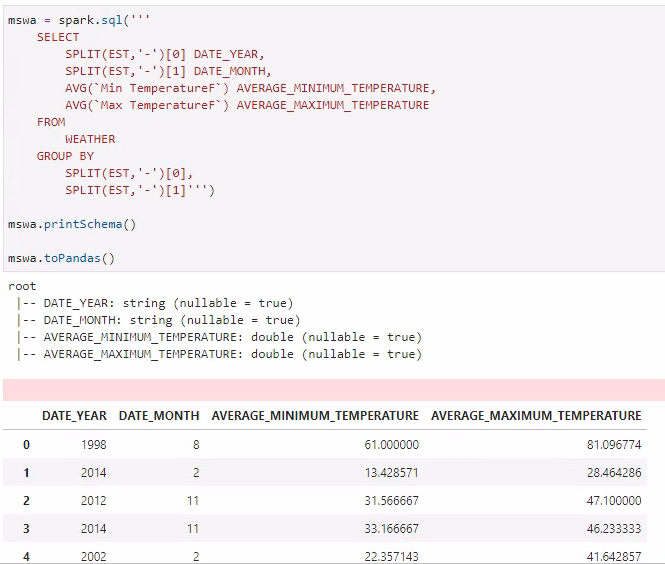




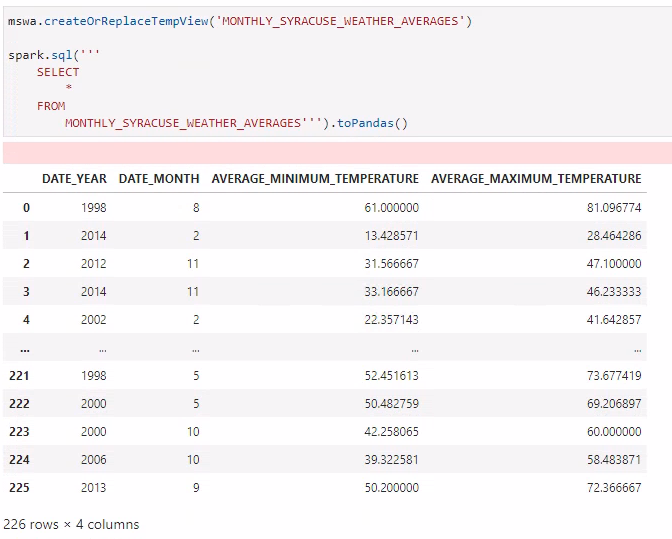
1. Configure Spark to read from Minio **labe** bucket, then load **syracuse-ny.csv** into a DataFrame and register it as the table **weather**.

****

1. Rewrite Question 2 using pure Spark SQL and the **weather** temp view. NOTE: There will be some subtle differences with how you must write the code, so be sure to **printSchema()** so you can see what the columns are.



1. Save the output from the DataFrame in Question 6 to the temp view **monthly\_syracuse\_weather\_averages**. Prove the view is there by querying it.



1. CHALLENGE YOURSELF! At the bottom of the **work/content/E-Drill-Spark.ipynb** file there is a section called Big Data to Small Data. Try to write a complete program that:
   1. Inputs a month 1–12 at run-time
   2. Displays a scatter plot of min/max average monthly temperatures, where year is on the X-axis



