Homework Problem Set E Submission Form

# Overview

|  |  |
| --- | --- |
| Your Name | Sana Khan |
| Your SU Email | Skhan53@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.

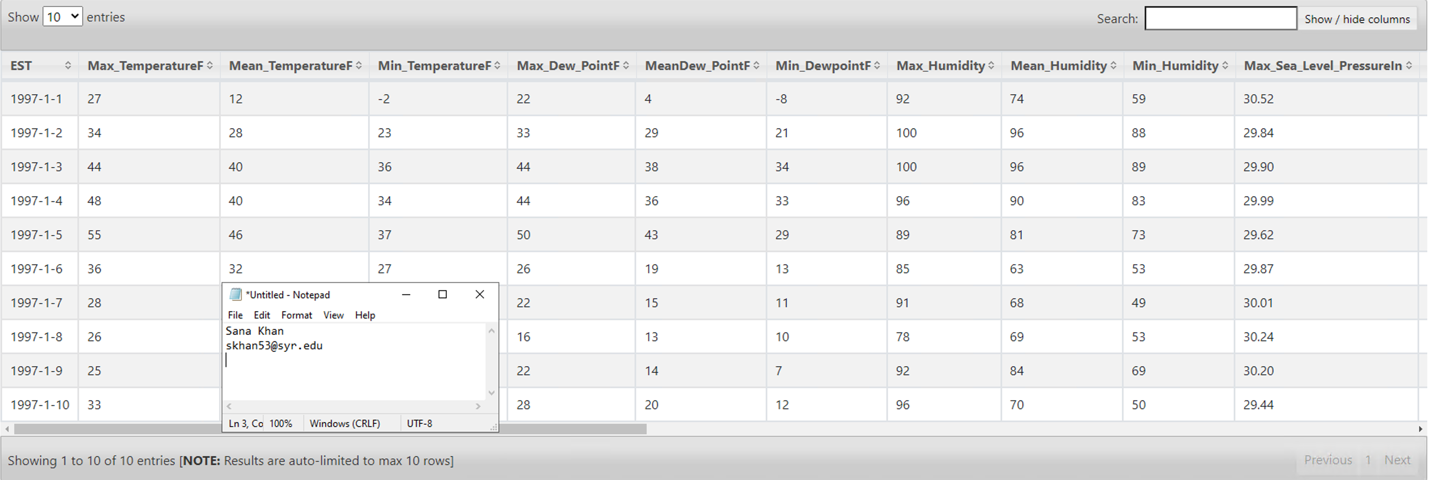
1. As you completed this assignment, identify what you learned.
2. What barriers or challenges did you encounter while completing this assignment? For number 7, I couldn’t get the data to load properly and the temperatures showed null values
3. How prepared were you to complete this assignment? What can you do to be better prepared?Undersgand spark a little bit better
4. Rate your comfort level with this week’s material. Use the rubric provided. 2

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.**

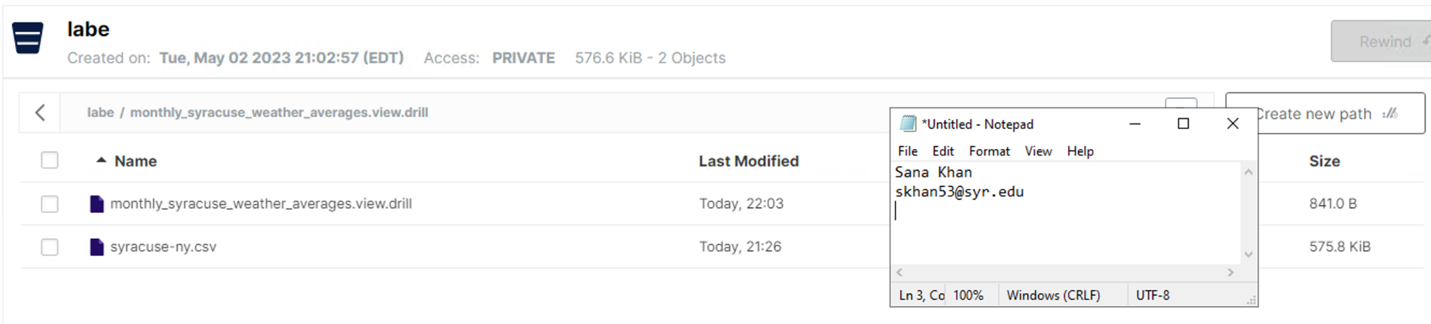
1. Configure a Drill storage plugin for the Minio **labe** bucket. Then write a Drill query for **syracuse-ny.csv** to demonstrate you can read the file with headers.   
   **SELECT \* FROM labe.`syracuse-ny.csv`**



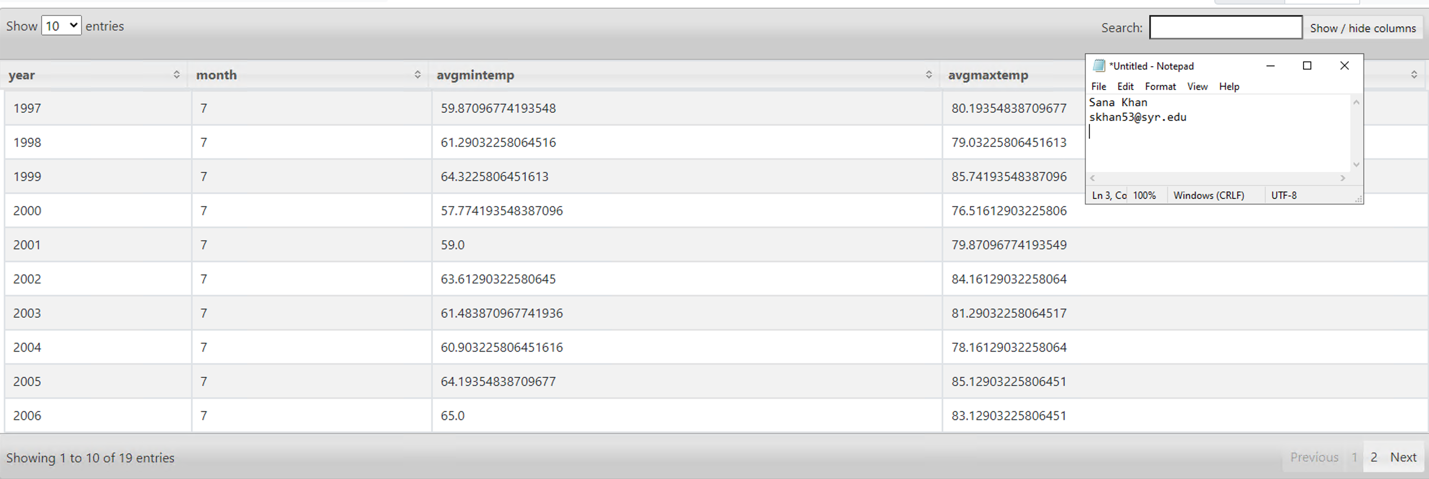
1. Write a Drill SQL query to get the overall average min and max temperatures by year and month. Use Drill’s SPLIT() function to separate Year, Month. You might need to use cast() to ensure the min and max temperatures are numeric types. You output should include four columns: Year, Month, the average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

**CREATE VIEW labe.monthly\_syracuse\_weather\_averages  
AS  
WITH  
Source  
AS  
(  
SELECT cast(split(EST, '-')[0] AS int) AS year,  
cast(split(EST, '-')[1] AS int) AS month,  
cast(Min\_TemperatureF AS int) as mintemp,  
cast(Max\_TemperatureF AS int) AS maxtemp  
FROM labe.`syracuse-ny.csv`  
)  
SELECT year, month, AVG(mintemp) AS avgmintemp, AVG(maxtemp) AS avgmaxtemp  
FROM Source  
GROUP BY year, month  
ORDER BY year, 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.

SELECT \* **FROM labe.monthly\_syracuse\_weather\_averages  
WHERE month = 7**  


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

**s3\_bucket = "labe"**

**file\_name = "syracuse-ny.csv"**

**print("hell")**

**source = f"s3a://{s3\_bucket}/{file\_name}"**

**weather = spark.read.option("header", True).option("inferSchema",True)\**

**.csv(source)**

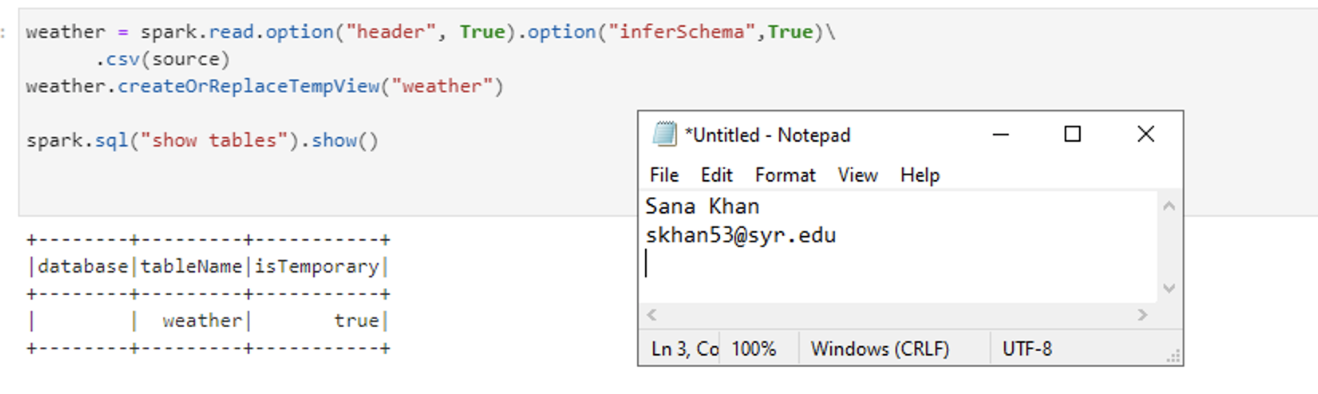
**weather.show()**

**weather = spark.read.option("header", True).option("inferSchema",True)\**

**.csv(source)**

**weather.createOrReplaceTempView("weather")**

**spark.sql("show tables").show()**

****

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.

**query = '''**

**WITH**

**Source**

**AS**

**(**

**SELECT cast(split(EST, '-')[0] AS int) AS year,**

**cast(split(EST, '-')[1] AS int) AS month,**

**cast('Min TemperatureF' AS int) as mintemp,**

**cast('Max TemperatureF' AS int) AS maxtemp**

**FROM weather**

**)**

**SELECT year, month, AVG(mintemp) AS avgmintemp, AVG(maxtemp) AS avgmaxtemp**

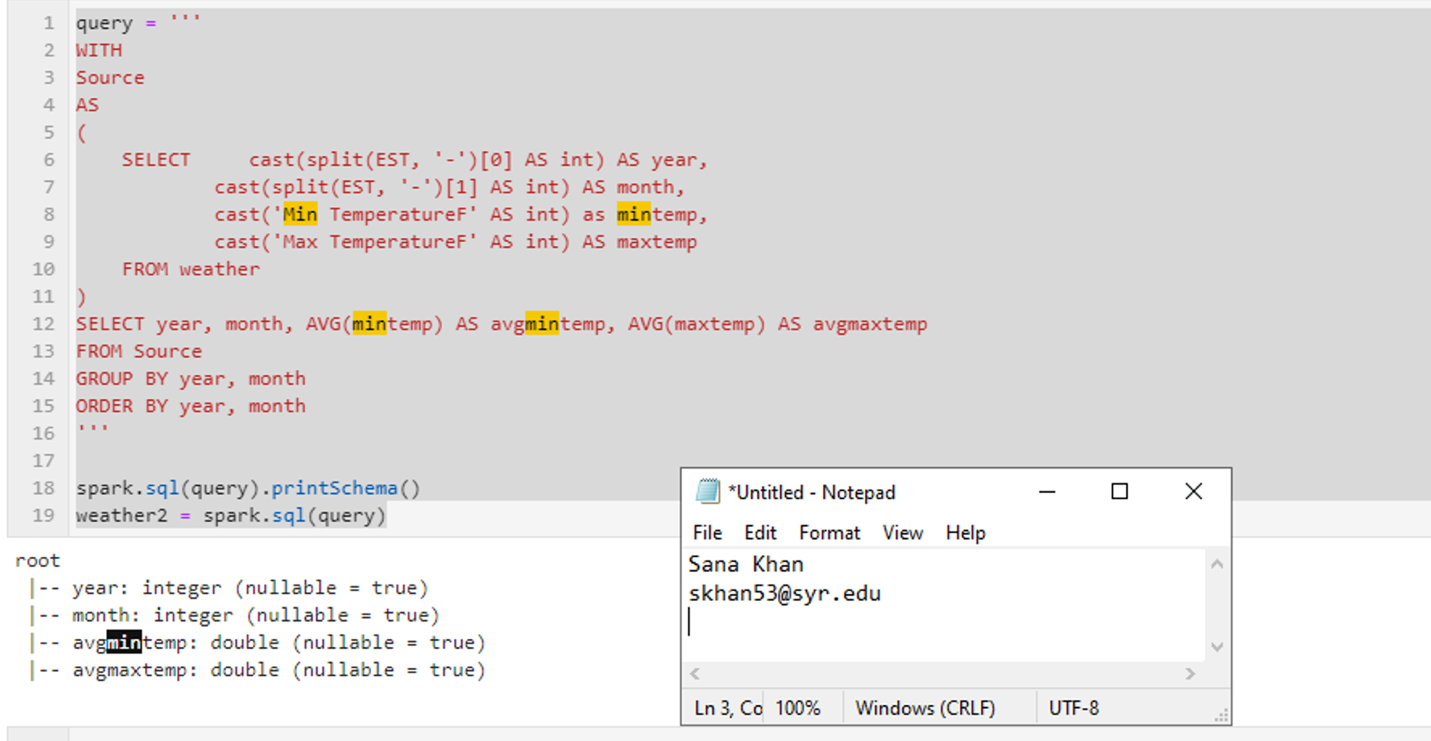
**FROM Source**

**GROUP BY year, month**

**ORDER BY year, month**

**'''**

**spark.sql(query).printSchema()**

**weather2 = spark.sql(query)  
**

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.

**weather2.createOrReplaceTempView("monthly\_syracuse\_weather\_averages")**

**query2 = '''**

**select \* from monthly\_syracuse\_weather\_averages**

**'''**

**spark.sql(query2).show()**

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