Assignment 3 Instructions and Grading Rubric

The purpose of this assignment is to analyze tweets sentiment data. You will use the DB2 on Cloud service and the integrated RStudio environment in IBM Watson Studio. DB2 on Cloud uses wide columnar technology for advanced in-memory analytics.

**Assignment 3 is unavailable until Sentiment Analysis Tutorial is submitted.**

# Getting Started

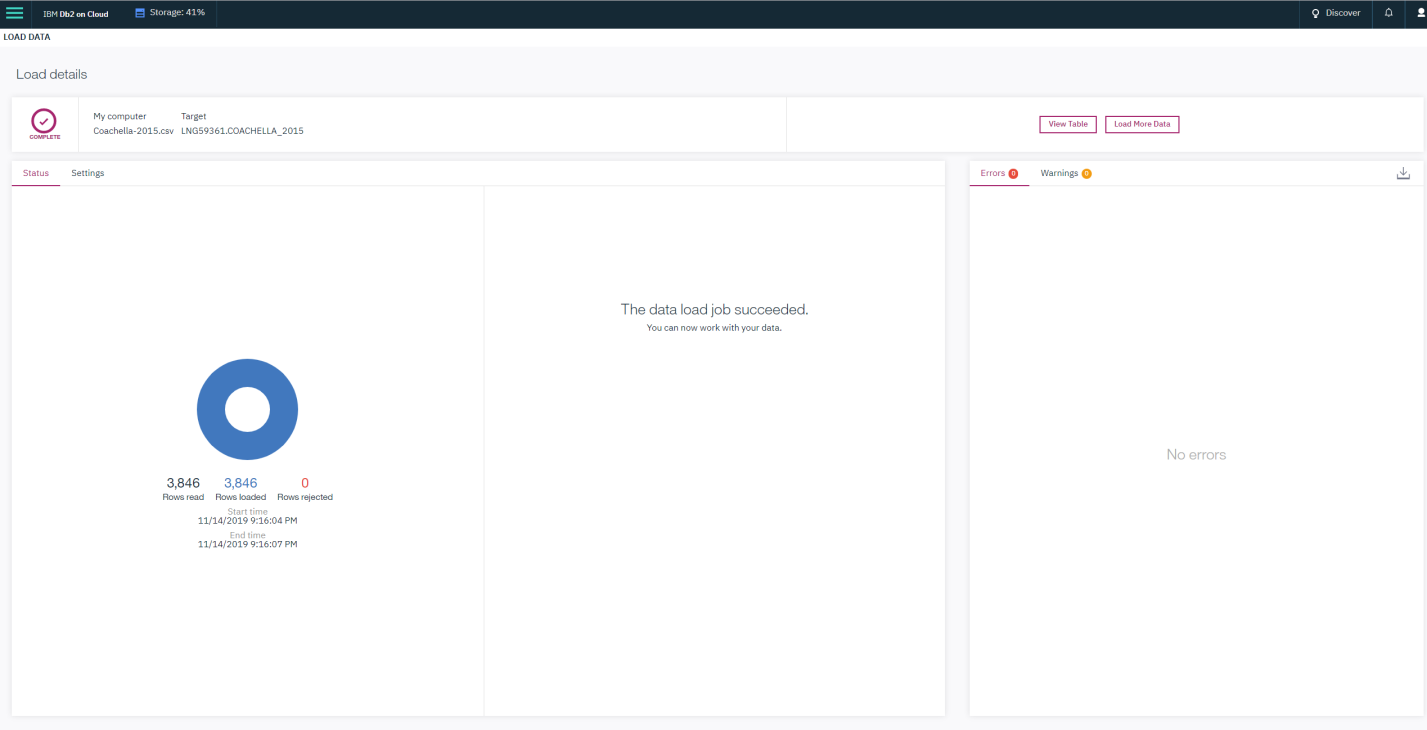
Before working on this assignment, you need to

* Your IBM Cloud account must be active
* Complete the readings in week 7-9 course content
* Earn 70% or higher on text mining and SQL review quiz in week 7-9 content.
* Submit the Sentimental Analysis Tutorial assignment
* **Choose one of the datasets from the list in Assignment 3 folder**. Download the csv file from assignment folder to your computer. **Do not open the file.**

# Assignment Requirements

The assignment deliverable must be an original work, must comply with graduate writing standards, and should include:

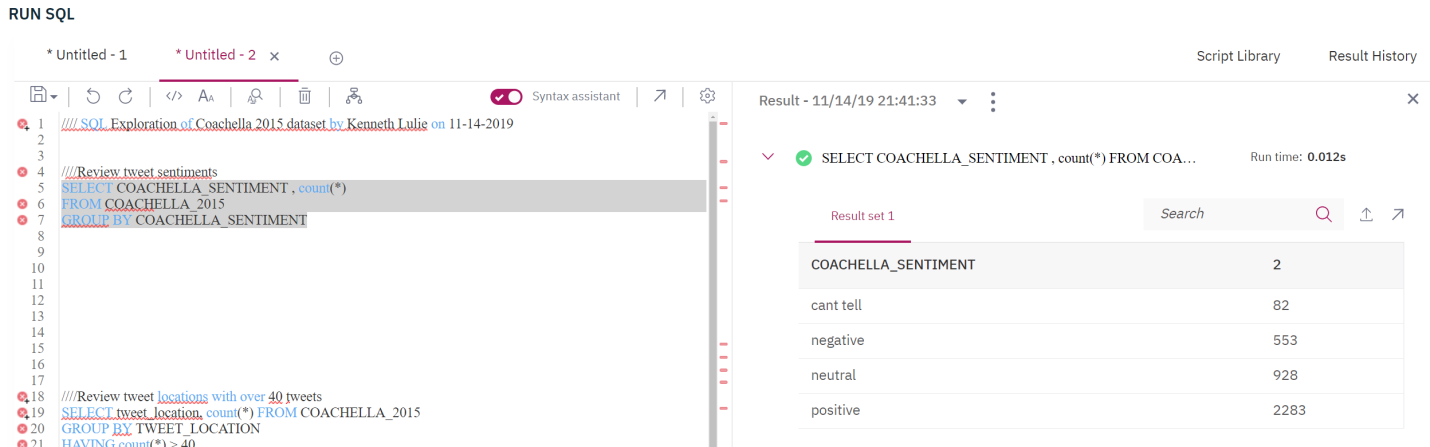
1. Introduction – Clearly explain the topic choice and the purpose of the study. Explain what data you are analyzing.
2. Read dataset descriptions first before choosing.
3. Use DB2 on Cloud web console to load the data. When the load completes, you will be redirected to the status page. Take a screenshot of that page and discuss the load status in the paper.



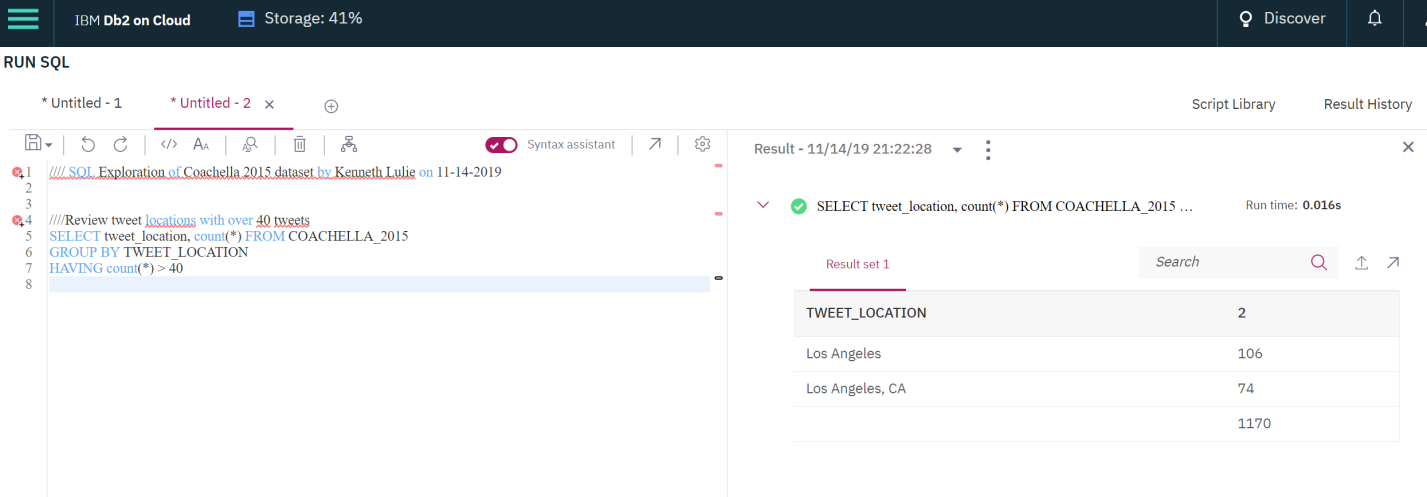
1. ez
2. Discuss the programmatic approaches you used to verify if the data was loaded as expected, including checking the row counts and table metadata.

Row counts matched, 0 rejected, 0 warnings, reviewed column datatypes for any unexpected inferences of datatypes.

1. ez
2. Use the SQL Editor in the web console to analyze the data
   1. Write a query that returns the number of posts per each sentiment value. Include the query in the SQL script. Include a screenshot of the query output and output discussion in the paper.

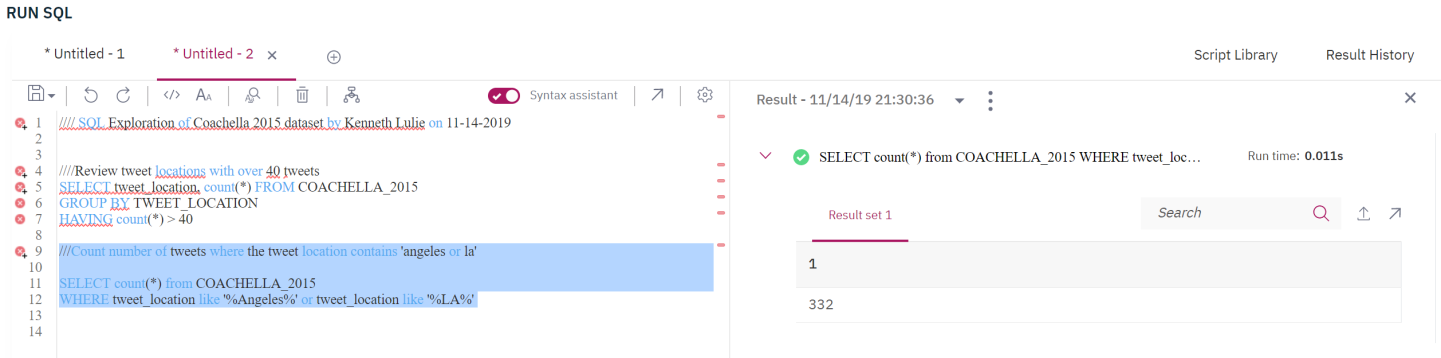


* 1. Write a query that returns the number of posts per user location with more than 40 tweets. Include the query in the SQL script. Include a screenshot of the query output/partial output and output discussion in the paper.

1. 
   1. Not a problem, can also use assignment 2 as a base
   2. Write four (4) additional meaningful queries. Use the subqueries, and aggregate functions. For each query, include the SQL statement in the SQL script. Include an output (or partial output if an output does not fit on a single page) screenshot, and **a paragraph** on query purpose and returned output.

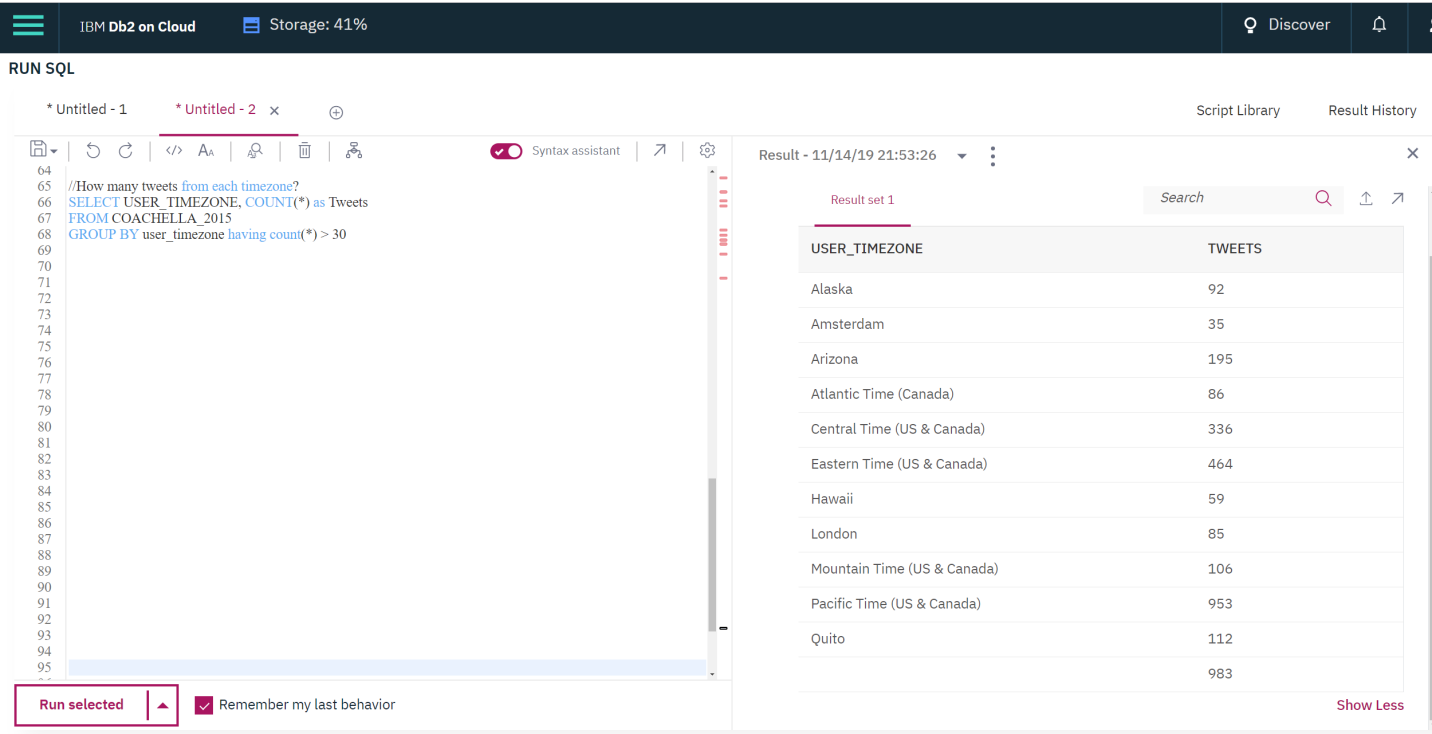
#1

In the first query we saw that Los Angeles showed up a few times under different names, this query shows that significantly more tweets are coming from LA than you might be led to believe based only on the original query.



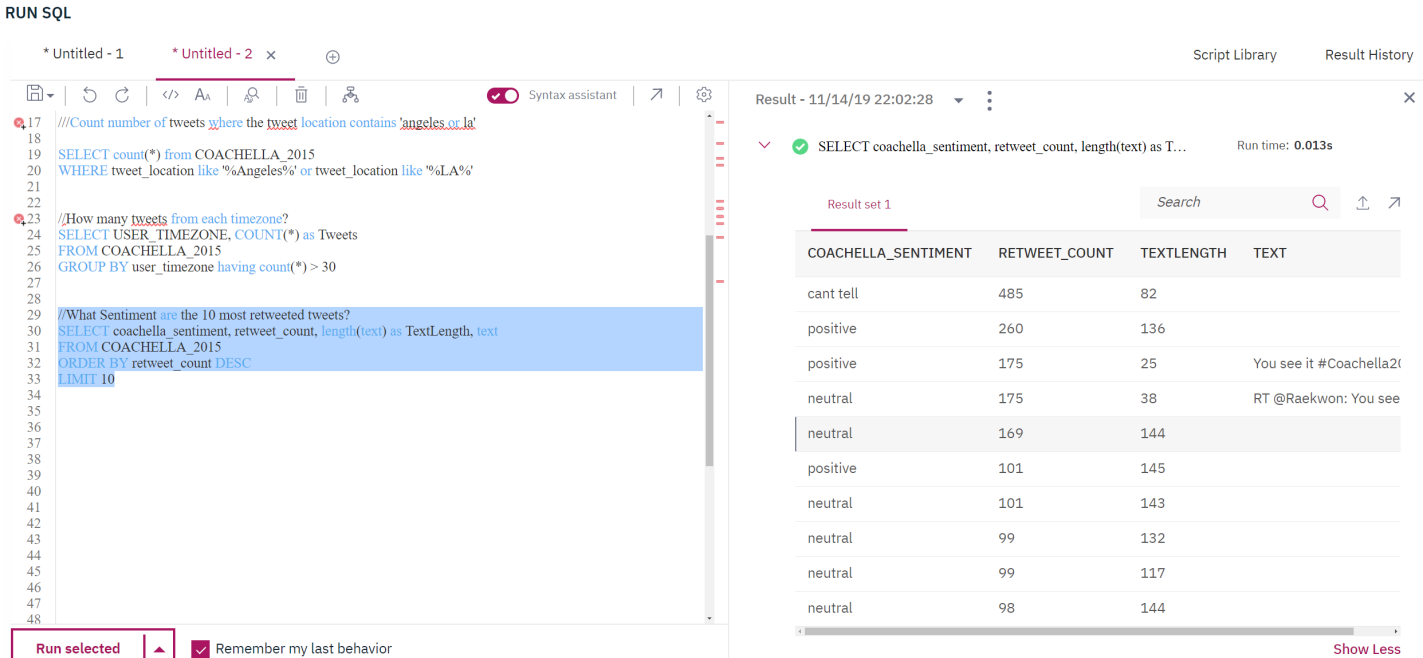
Here we look at what timezone tweets are coming from,

#2



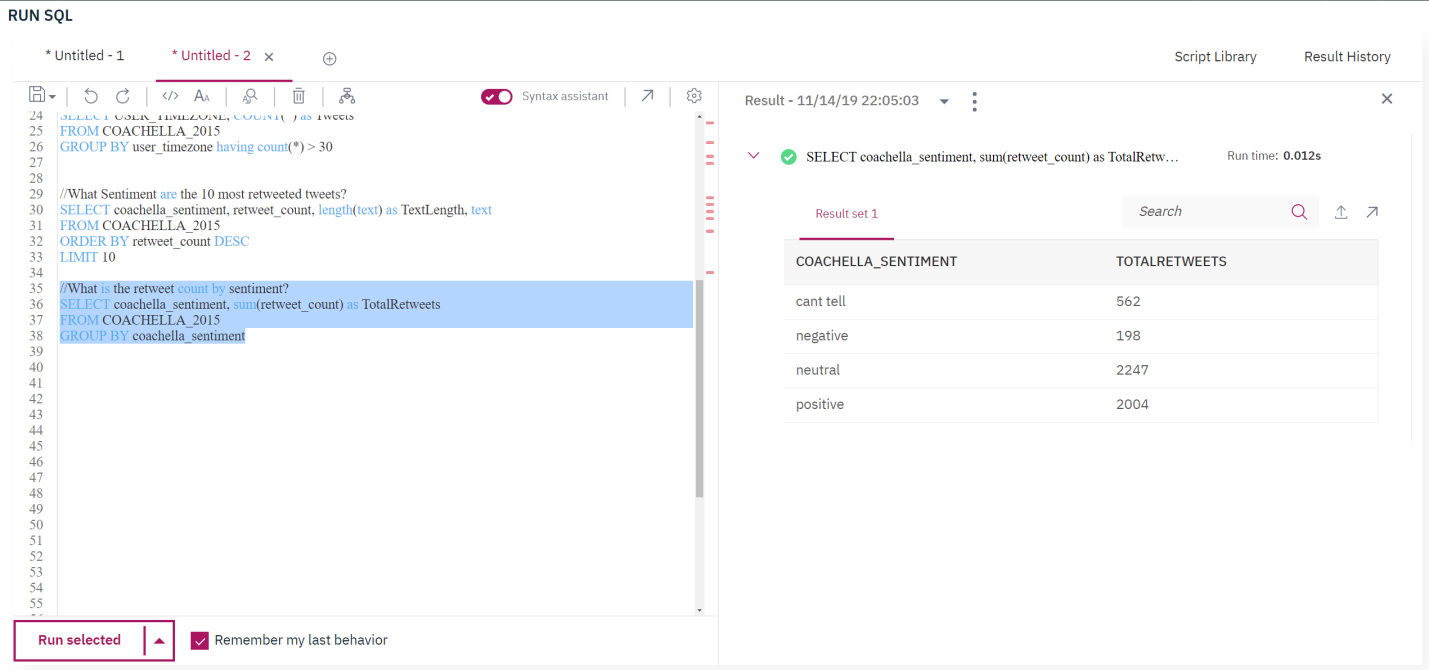
Note – Text being blank appears to be issue with IBM cloud, as the text length shows that there are characters in there.

#3



What is the retweet count by sentiment?

#4



1. Use the integrated RStudio environment in Watson Studio to connect to DB2 on Cloud and to load data into R data frame. Run the commands to validate the load and discuss the data load validation approaches you used. Include all commands and in the correct order in the R script.

Essentially I loaded the data, compared counts of rows and columns to the csv file on my desktop, and the database which was uploaded, and also manually viewed the dataframe in rstudio. All counts were as expected and all the data looked as expected.

I also confirmed that the tables existed and were the right tables.

1. Analyze the data in R. Present and interpret the findings. Include all code in the correct order in the R script. The analysis method should include but not limited to
   1. Frequency counts and plots that support the purpose of the study

*Purpose is to provide information to the social media team regarding the reception of the 2015 Coachella line up.*

Important to come up with a useful analysis reason

Looks like these are the tweets of the two days after the Coachella line up announcement on January 6th, 2015 and January 7th 2015.

Social Media team of Coachella wanting to understand the reaction to the line up announcement.

Can do an analysis of the most influential tweeters, perhaps partner with them to increase hype?

Can try to identify the most negative and influential (as counted by retweets) tweets, why are people being negative, is it actually negative, is it something we can address?

Are any particular bands getting a lot of hype?

* 1. Run **2-3 methods** from the following list. For each method, discuss the rationale, pre-processing, input parameters you used, and interpret the results.

Probably trend analysis, word cloud/term frequency analysis, maybe dendogram?

Trend analysis might not be super useful since only two days? Maybe it is anyway.

* + 1. Apriori rules method to analyze the tweets metadata
    2. Trend analysis – the number of posts of each sentiment over time
    3. Dendogram
    4. Word cloud, and term frequency analysis
    5. K-means clustering
    6. Network of terms analysis
  1. Be creative and run analysis beyond examples in the walk through. **Explore the method input parameters not discussed in the walk through**. However, **please check with the instructor and TA before installing packages** besides SnowballC

Wild card, provide additional value. (I think this is addressed with my ideas higher up)

1. Discuss the study accomplishments and 4-5 limitations. Propose further research.

# Assignment Deliverables

Submit three **(3) files** into Assignment 3 folder

1. The paper in a Word or PDF format
2. A text file with the SQL code. The code needs to have comments and needs to include all queries in the same order as discussed in the paper.
3. A text file with the R code. The code needs to have comments and needs to include all statements in the same order as in the paper.

**Only one assignment submission is allowed. Make sure to attach all 3 files before hitting the submit button.**

**Do not update your IBM cloud account and do not enter the credit card number.**

**Do not submit any tickets unless instructed by the faculty to do so.**

**The assignment due date cannot be changed. Late submissions will incur a5% points earned penalty per each day late.**

# Assignment 3 Grading Rubric

|  |  |  |
| --- | --- | --- |
| Criteria | Weight  (%) | Score(0-100) |
| Introduction and Purpose – provide a clear and purposeful background to the study. | 10 |  |
| Method - describe your experimental design, including data load verification, SQL queries, R methods selection, data pre-processing steps. The experimental design must address the purpose of the study and must include the methods outlined in the assignment requirements.  The assignment submission must include a text file with SQL code and a text file with an R code. **If one or both files are missing, the assignment cannot be graded.** | 35 |  |
| Analyze data in R and interpret your findings. Demonstrate the concepts application through clear explanation and elaboration. Go beyond examples in the walk through. | 25 |  |
| Discuss the accomplishments and limitations of the study, including the limitations of selected methods. Propose further research and state the conclusions. Use references effectively. | 25 |  |
| Presentation & Articulation, APA format, paper flow and mechanics (grammar, spelling, and punctuation); all figures and tables should be in chronological order and should support the findings. The references in your reference list should be cited in the paper. | 5 |  |
| Total | 100 |  |