**Data Engineering case study**

**Instructions :**

* The purpose of this exercise is to consistently and carefully assess the relevant job skills for each applicant.
* This exercise will be completed on your own time.
* You will have 72 hours to complete this exercise. Be sure to keep track of your own time and plan accordingly. You should be sending back your output within 72 hours of receiving it, failing which your candidature will be removed from consideration.
* Be prepared to discuss the Job Simulation and your findings during your scheduled interview time.

**Task:**

* + Download the data onto your desktop from <https://github.com/aghosh704/tickerdata/blob/main/tickersData.csv.zip?raw=true>
  + Setup Spark on your personal laptop or computer. Note: Minimum memory requirement is at least 8 GB. You may use Scala Spark, but PySpark is preferable. You may not use other environments than Spark.)
* Answer the following questions in a Word Document, using the data that falls between 11/24/2017 (inclusive) and 12/23/2017 (inclusive)

**Questions:**

1. How many types of ETF products are available?
2. What are the top three heavily traded stocks?
3. Which stock had the highest number of consecutive up days (in terms of end of the day prices)? What were the assumptions/adjustments that led to your solution to 6.c.?
4. How many stocks had at least 5 consecutive up days and at least 5 consecutive down days (in terms of end of the day prices)?
5. What are 3 or 4 insights that you have identified in this data set?
6. What are 3 or 4 data quality issues that you have identified?

* Please share back the word doc with answers to the above questions and any table / visuatlization to show the results.
* Please share your code (github link or Jupyter notebook). If Jupyter NB, please share both .html and .ipynb versions
* Please email the above two to the WF person you are in contact with.

**Context:** This assignment tests basic ETL data pipeline skills and exploratory data analysis.

**Answers:**

|  |  |  |
| --- | --- | --- |
| S.NO | Question | Answer |
| 1 | How many types of ETF products are available? | 1123 |
| 2 | What are the Top three heavily traded stocks? | DE000A0KRJ36| OD7L| ETC  SE0006027546| SVAB|Common stock  SE0006027546| SVAB|Common stock |
| 3 | Which stock had highest number of Consecutive  up days? | ETF |
| 4 | Which stock had Five Consecutive up days and  Down days? | ETF |
| 5 | What are 3 or 4 Insights about data sets? | Provided report of insghts about the data sets in HTML format |
| 6 | What are 3 0r 4 data quality issues? | Missing values filled with Nan |