W205, Information Storage and Retrieval

Week #: 5
Exercise #: 1
Word Count: 1033
Name: Gregory Ceccarelli
Date: 10/18/2015

Outline of General Assumptions / Steps Taken

As instructed, data was downloaded from the Centers for Medicare and Medicaid Services (CMS) Hospital Compare and stored on a remote AWS instance in [/usr/local]. For step 1, the data was renamed, had it's header stripped and moved to the loading_and_modeling directory of the remote machine. The provided data dictionary was thoroughly reviewed and it's interpretation formed the basis for a folder structure in HDFS that was subsequently created corresponding to the ultimate staging environment envisioned.The staging env. was based on a selection of files deemed necessary to answer the four questions. Their renamings are as follows:

- Hospital General Information.csv > hospitals.csv
- Timely and Effective Care Hospital.csv > effective_care.csv
- Readmissions and Deaths Hospital.csv > readmissions.csv
- Measure Dates.csv > measuredates.csv
- hvbp_hcahps_05_28_2015.csv > hosp_patientexperience.csv
- hvbp_tps_05_28_2015.csv > hosp_totalperformance.csv
- READMISSION REDUCTION.csv > readmissionreduction.csv
- HCAHPS Hospital.csv > hosp_surveys.csv

After the data was loaded into HDFS, a helper program was written in python using pandas in order to load the data into postgressql to automatically generate a table schema. This schema was used for two reasons 1) to help generate the ERD diagram for the staging environment which was created using Vertebelo.com and 2) to create a schema that could be easily translated into Hive SQL. Once generated, the staging ERD was created and the corresponding SQL necessarily to create EXTERNAL tables in Hive was translated

Step 2 began with data exploration. Various relationships between the chosen files were explored - these relationships were modeled into the ultimate transformed ERD that was to be used for the basis of the required analysis. The tables ultimately created were aggregations and joins resulting in a hospital schema conducive to creating a procedure score to understand high-quality care and to figure out how that extended at various levels of aggregation.

The final step of the exercise was to actually conduct the analysis. A few top level assumptions were made that warrant being called out here:

- A sample of 9 procedures (1 from each condition category) were selected to answer question 1
    - It was determined that this should be representative of the overall corpus of procedures
- The measurement period for those procedures was determined to be 2013Q4 - 2014Q3

Following from these pre-conditions, an analysis was conducted iteratively in iPython notebook connected to a kernel running on the same remote machine using both Spark Dataframes and Spark SQL -- allowing for flexibility in computation and expression. The output notebook and resultant .py files can be found in the root of the [exercise_1/investigations] as well as in each of the subdirectories: best_hospitals, best_states, hospital_variability and hospitals_and_patients. This narrative is also saved there as a PDF.

Answers to Questions

1. What hospitals are models of high-quality care—that is, which hospitals have the most consistently high scores for a variety of procedures?

   To compute those hospitals that are models of high quality care, a few operations were undertaken in addition to abiding by the assumptions provided above.
   - The effective score range was matched against each measure (procedure) of the 9 selected
   - By provider, procedure scores with a sample size less than 30 were eliminated from the analysis
   - Of the remaining providers, aggregate scores (perc_score, potential score maximum, count of procedures contributing to the score and score standard deviation) were created such that each provider had one entry in the result set
   - Data corresponding to excess readmissions was incorporated into the analysis and those providers who had a ratio of readmissions (predicted vs. expected) greater than 1 were removed from the result set
   - Once this was computed, those providers with score standard deviations greater than 2 were also eliminated, in an effort to reward those providers who had consistent marks in the scores randomly sampled.
   - Finally the results ordered by score and were written to the output file.
   - **The top 10 providers can be reviewed in the file best_hospitals.txt**

2. What states are models of high-quality care?

   This computation piggy backed on the assumptions outlined above. Much of the same logic was employed, however; the level of aggregation from the previously computed “best_hospitals” result set was changed to reflect state level.

- Percent Scores, Standard Deviations and other provider level data was recomputed from the score data available at the provider line level
- Ultimately this approach is justified by the fact that it is consistent with my results for hospitals as models of quality
- **The top 10 states can be reviewed in the file best_states.txt**

3. Which procedures have the greatest variability between hospitals?

   Unlike the above two questions where a ranking methodology was employed based on sampling coupled with reasonable constraints applied to the data, this question was approached slightly differently.
   - Instead of limiting the data based on a sampled selection of procedures, **all procedures by condition (category)** were included in the case they were scored on the same scale (max score = 100) and measured in the time period between 2013Q4 - 2014Q3.
   - Using similar aggregation logic, the following statistics were computed by procedure: variance, standard deviation and count of providers.
   - The data was ranked and the top 10 procedures were written to an output file sorted in descending order based on total variance
   - **The top 10 procedures with the most variability can be reviewed in the file score_variability.txt**

4. Are average scores for hospital quality or procedural variability correlated with patient survey responses?

   The short answer is not really, there is only a slightly positive (3%) correlation between hospital quality (as calculated according to the assumptions outlined in Question 1) and scaled average responses (taken from the transformed table t_hospitalspatientexperiencescaledscore)
   - The data precomputed that served as the basis for the answer to Question 1 was matched at the provider level to the scaled patient experience score
   - The output was saved to a data frame
   - A correlation matrix was built using pandas on the columns: perc_score (corresponding to hospital quality) and scaledscoreaverage (corresponding to patient survey responses).
   - **The output as provided below was recorded and can be reviewed in the file quality correlation.txt**
	perc_score	scaledscoreaverage
perc_score	1.000000	0.029226
scaledscoreaverage	0.029226	1.000000