W205 Exercise 1

This file is some of the observations, choices, and analysis that occurred during this exercise.

General information about project

- Used information from here for the measures information.
- Also this document for the HCAPS data.
- I extracted data from the 5 recommended files and converted them with transformations detailed later into the following tables with some limited filtering of unnecessary (for my analysis) columns and fields:

Filename	Intermediate Table	Table
Hospital General Information.csv	Hospitalgeneralinformation	Hospitals
hvbp_hcahps_05_28_2015.csv	hvbp_hcahps_05_28_2015	Survey
Measure Dates.csv	Measuredates	Procedures
Readmissions and Deaths - Hospital.csv Timely and Effective Care - Hospital.csv	readmissionsanddeathshospital timelyandeffectivecarehospital	ProcedureScores

I also created one extra table **HospitalScores** which was the result of calculations on a join of the Hospitals and ProcedureScores tables.

Decisions, observations, and cleaning

- I removed the following procedures from consideration:
    - EDV – this is simply a measure of the volume through the emergency department, this is not a procedure or measure of quality.
    - Any field that was a measure of time rather than percentage. There are a number of counts of the "Median time" in the Measure data. I (maybe personally) do not consider the length of time for some treatment to be related to the quality, however I could have ranked the hospitals or assigned ranges with scores for their speeds. Even doing this though, the median with no variance or mean has limited use.
    - OP_22 – These were individuals who left prior to being seen. Again not a measure of quality, there are a number of reasons they could have left which are unrelated to the hospital or the quality of care (started feeling better to name one)
- I observed that several procedures have "Lower percentages are better" type values. I updated these scores to be "higher is better" by subtracting them from 100(%). That allowed me to treat them all scores the same.
- I noted that the survey file hvbp_hcahps_05_28_2015.csv used providernumber rather than providerid used in all the other files. When I created the Survey table from it I converted the name of this field.

- I found 6 procedures which were in the “Timely and Effective Care - Hospital.csv” file but were not in the “Measure Dates.csv”. I added these in from the link above as they seemed to be regular procedures. I did not have any other way except using the internet for this since unlike many of the other fields the measure descriptions were not duplicated.
- The hospital names and addresses do not always match across the CSV files. This is likely as the data is filled in at different times by different people. I have worked on the assumption that the “Hospital General Information.csv” file contains the legitimate names and addresses for the hospitals and ignored these fields in all other files. I also assume that the ProviderID (which is a unique number for the hospital) is correct in all files.
- All data was type STRING so I converted (cast) fields to numeric values for the scores as I wanted to be able to run numerical calculations on these. I did this for the hospital and the survey scores.
- The HCAPS document indicated that a “Patient Experience of Care Domain” field was a sum of the Base and Consistency score. That seemed like a reasonable measure which was missing from our “hvbp_hcahps_05_28_2015.csv”. I chose to create the same in my Survey table by summing the base and consistency values in each row. I used this new field “PatientExperienceOfCareDomain” as the total measure from the survey table.
- In the HospitalScores table I calculated the following data for each hospital
    - Average: the mean of all scores
    - Maximum and minimum: the maximum and minimum score values for each hospital
    - Range: the difference between the maximum and minimum
    - Sum: the sum of all the scores
    - Number of Scores: a count of the number of scores the hospital had
- In the survey table the base score is supposed to be a sum of all the individual scores, taking either the dimensions field or the achievements (whichever is larger is used). During my verification of this I observed that the dimension field is always greater than or equal to the achievements field. I do not have a conclusion for this but it was odd.
- I do not know the mappings for the two digit state name abbreviations so I looked them up and created a “StateName” table to for this mapping. After that I used this to populate the full state name into the Hospitals table.
- A large number of the scores in the Readmissions and Deaths - Hospital.csv and Timely and Effective Care - Hospital.csv are “Not Available” (nearly half). I completely left these out as they do not provide any use in the evaluations.
- There are 10 conditions in the Timely and Effective Care

Implementation

I created two shell scripts for this project.

1. Under the loading_and_modelling directory I created a long script with a large number of command line options. The details are in the README.md of that directory, including how to use it.
   The script validates much of the environment, checks the availability of the files and directories, creates required directories for the files, copies over the files (efficiently, removing the header but not creating intermediate files), creates the SQL script from the headers and loads it into the Hive database.
2. A load.sh script for the transform and investigations directories. I created a convention that XYZ_hive.sql files should be loaded with hive and all others can be loaded with spark-sql. This script enforces it. Just run it in the directories with the sql files and it loads them in alphabetical order. (I also copied this into each folder because I was not sure how github would treat symbolic links.)

The transformation SQL files are documented in the README.md file in the transformations directory.

Calculations

There are further details in the .txt files in each of the investigations directories, here is a quick overview

1. Best hospital
   For this I was asked for the hospitals with consistent high scores. I chose to use a minimum of 10 scores and a maximum range of 15% between the top and the bottom score. The top 10 winners were:
   1) SURGERY SPECIALTY HOSPITALS OF AMERICA SE HOUSTON
   2) NOVANT HEALTH PARK HOSPITAL
   3) KANSAS SURGERY & RECOVERY CENTER
   4) GHS PATEWOOD MEMORIAL HOSPITAL
   5) QUAIL CREEK SURGICAL HOSPITAL
   6) NORTH CAROLINA SPECIALTY HOSPITAL
   7) WOMEN'S HOSPITAL THE
   8) ARIZONA SPINE AND JOINT HOSPITAL
   9) HOAG ORTHOPEDIC INSTITUTE
   10) THE ORTHOPEDIC SPECIALTY HOSPITAL
2. Best state
   For this I did not care about how many procedures per hospital. I calculated the total score for the state and divided by the number of scores for the state to get the mean score for each procedure in that state. This was independent of how many hospitals and how many procedures per-hospital in each state. The top 10 winners were:
   1) Utah
   2) Maine
   3) Virginia
   4) Delaware
   5) Colorado
   6) Connecticut

7) New Hampshire
8) North Carolina
9) South Carolina
10) Massachusetts

3. Greatest variability in procedures
   Here I checked these independently of the hospitals. For each procedure, I checked the difference between the maximum and minimum recorded value and then ordered the results by this calculation. Given that more than 10 (15) procedures had a range of 100% (some hospitals scored 0 and some 100). The following is a list of all 15:
   1) Stroke Education
   2) Venous Thromboembolism Prophylaxis
   3) Intensive Care Unit Venous Thromboembolism Prophylaxis
   4) Influenza Immunization
   5) Evaluation of LVS Function
   6) Head CT Scan Results for Acute Ischemic Stroke or Hemorrhagic Stroke Patients who Received Head CT or MRI Scan Interpretation Within 45 Minutes of ED Arrival
   7) Elective Delivery
   8) Venous Thromboembolism Patients Receiving Unfractionated Heparin with Dosages/Platelet Count Monitoring by Protocol or Nomogram
   9) Antithrombotic Therapy By End of Hospital Day 2
   10) Surgery Patients Who Received Appropriate Venous Thromboembolism Prophylaxis Within 24 Hours Prior to Surgery to 24 Hours After Surgery
   11) Prophylactic Antibiotic Received Within 1 Hour Prior to Surgical Incision
   12) Thrombolytic Therapy
   13) Discharge Instructions
   14) Venous Thromboembolism (VTE) Prophylaxis
   15) Venous Thromboembolism Warfarin Therapy Discharge Instructions
4. The correlation between the hospital reported scores and the survey scores, and the procedure range in the hospitals and the survey scores. Here I use the "Patient Experience of Care Domain" (discussed in the decisions and cleaning section) as the survey score and then got the correlation for it against the hospital mean score and the hospital range score. In both cases I got no significant correlation (results of 0.06 and -0.12). Looking over the data afterwards I think this could have been expected. The hospital scoring is about procedures that the hospitals perform for patients, whereas the survey is only 9 questions (split into 3 parts each) and 5 of them are about how well the staff communicated with the patient.

Possible future analysis

These are a list of things I considered or would like to have had the time to do:

- Analyze the hospitals on a type of care basis

- I did not put in any weighting on the procedures or categorize them in any way. If I had more domain experience, then this might be a useful extension. I think particularly the readmissions and deaths should probably have a higher weight since they are likely more indicative of care quality than procedures such % of patients who received asprin within 30 minutes
- Analyze the care by the type of illness. This could show if a hospital is a leader at dealing with one type of illness. Also given that I have removed hospitals that did not have enough procedures, this would allow specialist hospitals to be considered.
- The footnote in the "Timely and effective care file" needs more investigation. There are a number of scores for footnotes that would seem to indicate that any data is invalid. I was investigating these because the range of values for many procedures is 0->100%, which is a very large range. It may be that some of the 0 values are misinterpretations of how to fill in "Not Available".

Initial observations of files

I created the below table to initially analyze the data. This shows the sets of fields that were present in the csv files and with some preliminary analysis. I was able to figure out what appeared to be the primary keys in the files. The field entries in red indicate fields that I took to be primary keys and ones in blue are fields that only showed up in one file.

The column names in the table are the following:

Hospital General Information.csv -> hospital

hvbp_hcahps_05_28_2015.csv -> survey

Measure Dates.csv -> measures

Readmissions and Deaths - Hospital.csv -> readmissions

Timely and Effective Care - Hospital.csv -> effective care

Field	hospital	effective care	readmissions	measures	survey
ZIP Code	present	Present	present		present
State	present	Present	present		present
Hospital Name	present	Present	present		present
County Name	present	Present	present		present
City	present	Present	present		present
Address	present	Present	present		present
Provider ID	**present**	**Present**	**present**		present as Provider number

Phone Number	present	Present	present		
Measure Start Date		Present	present	present	
Measure Name		**Present**	**present**	present	
Measure ID		**Present**	**present**	**present**	
Measure End Date		Present	present	present	
Score		Present	present		
Footnote		Present	present		
Sample		Present			
Responsiveness of Hospital Staff Improvement Points					present
Responsiveness of Hospital Staff Dimension Score					present
Responsiveness of Hospital Staff Achievement Points					present
Provider Number					**present**
Pain Management Improvement Points					present
Pain Management Dimension Score					present
Pain Management Achievement Points					present
Overall Rating of Hospital Improvement Points					present
Overall Rating of Hospital Dimension Score					present
Overall Rating of Hospital Achievement Points					present
Measure Start Quarter				present	
Measure End Quarter				present	
Lower Estimate			present		
Hospital Type	present				
Hospital Ownership	present				
Higher Estimate			present		
HCAHPS Consistency Score					present

HCAHPS Base Score					present
Emergency Services	present				
Discharge Information Improvement Points					present
Discharge Information Dimension Score					present
Discharge Information Achievement Points					present
Denominator			**present**		
Condition		Present			
Compared to National			present		
Communication with Nurses Improvement Points					present
Communication with Nurses Dimension Score					present
Communication with Nurses Achievement Points					present
Communication with Doctors Improvement Points					present
Communication with Doctors Dimension Score					present
Communication with Doctors Achievement Points					present
Communication about Medicines Improvement Points					present
Communication about Medicines Dimension Score					present
Communication about Medicines Achievement Points					present
Cleanliness and Quietness of Hospital Environment Improvement Points					present

Cleanliness and Quietness of Hospital Environment Dimension Score					present
Cleanliness and Quietness of Hospital Environment Achievement Points					present