1-) Introduction to the problem.

In the health economics literature, moral hazard and supplier-induced demand are well-known issues that threaten the stability of health insurance schemes. According to moral hazard, insured consumers tend to be less cost-conscious and, as a result, overspend on health care. According to supplier-induced demand, health care providers can stimulate their patients' moral hazard behavior in order to increase their income or protecting themselves against malpractice lawsuits. Because of these two issues, health insurance companies need to monitor and evaluate the medical practice of doctors serving their clients. However, other factors such as specialty, locality and even time may also explain the variability of medical practice. The present study analyzes a sample of the database of a Colombian health insurance company. The sample contains the number of patients, visits and ambulatory surgeries performed for each doctor working with the company during the years between 2008 and 2012. The sample also informs the doctor's specialty and locality (more specifically, the Colombian department). There is also information regarding the cost of a visit and a surgery, but unfortunately that information is not disaggregated per doctor: the sample contains the mean cost of a visit and a surgery per specialty.

The main purpose of the present study is to understand the role of specialty, locality and perhaps time in explaining the variability of the medical practice in terms of number of patients, visits and ambulatory surgeries. If the role of specialty, locality and time is weak, then other factors such as moral hazard and supplier-induced demand may be relevant and worth further investigation.

2-) The data.

As explained, the data comes from a sample of the database of a Colombian health insurance company. Originally, the data was divided into seven csv files that can be found in the original_data directory (please visit https://github.com/tmivanus/Springboard/tree/master/Capstone_01):

- *datos_ambulatorios_2008*, *datos_ambulatorios_2009*, *datos_ambulatorios_2010*, *datos_ambulatorios_2011* and *datos_ambulatorios_2012* have the number of patients, visits and ambulatory surgeries performed per doctor in the respective year as well as the doctor's specialty.
- *datos_departamentos* informs the doctor's locality (Colombian department) for 2008-2012.
- *costos_promedios_ambulatorios* informs the mean cost of a visit and an ambulatory surgery per specialty for 2008-2012 in United States Dollar (USD).

The first task was to remove the names of the doctors from the first five csv files, substituting them for identification codes in order to preserve data confidentiality. The 00_erase_doc_name_datos_ambulatorios_ano code file performs that task and creates five new csv files to replace those with confidentiality issues: *datos_ambulatorios_2008_new*, *datos_ambulatorios_2009_new*, *datos_ambulatorios_2010_new*, *datos_ambulatorios_2011_new* and *datos_ambulatorios_2012_new*. For obvious reason, the old five csv files are not made available in the original_data directory.

After that, the 01_data_outpatient_years code file organize the newly created five csv files into python pandas dataframes, which are then saved in the data directory as the following csv files: *data_outpatient_2008*, *data_outpatient_2009*, *data_outpatient_2010*, *data_outpatient_2011* and *data_outpatient_2012*. The reader may wish to see step-by-step what tasks are being performed by opening the 00_optional_data_outpatient_year_details code file and choosing a year in the second cell.

The 02_mean_outpatient_costs_usd code file deals with the *costos_promedios_ambulatorios* csv file, organizing it into a python pandas dataframe and saving it in the data directory as the *mean_outpatient_costs_usd* csv file. That code file also translates the specialty of each doctor from Spanish to English using Google Translate.

Finally, the 03_data_departments code file organizes the *datos_departamentos* csv file into a python pandas dataframe, saving it in the data directory as the *data_departments* csv file.

All the recently created csv files in the data directory are put together to form a panel data by the 04_panel code file, which concatenates all the *data_outpatient* files from 2008 to 2012 and then merges it with *mean_outpatient_costs_usd* and *data_departments* files. The resulting panel data is saved in the data directory as the *data_panel* csv file. It contains 8294 rows and 12 columns of information that is going to be analyzed in the present study. It is worth noting that 10 rows were discarded because of errors involving the doctor's specialty name or specialty identification code.

Update: 05_first_analyses code file contain the first analyses made using the resulting panel data. After the first analyses, I was able to identify 8 specialties and 9 departments with enough data to be include in the next steps, as explained at the end of 05_first_analyses code file.