Problem set 6 - CSCI E-88b Spring 2020

Due date: Friday, May 15 2020, 10pm EST.

Please write down the last 5 digits of your Harvard ID: 70196

Please write down the URL of your Renku project: https://renkulab.io/projects/stephen.t.knapp14/problem6 (https://renkulab.io/projects/stephen.t.knapp14/problem6)

Please indicate who you may have worked with on this problem and which section you worked with someone on. If you did not work with anyone then you can leave this blank: xxxxx

This problem set corresponds to 1/6 of your final grade. It is graded out of 10.

Question 6.0: 0pt

Question 6.1: 3pts

Question 6.2: 3pts

Question 6.3: 4pts

Question 6.0: Keep it private!

If your problem set is not private, your grade will be 0/10.

Question: Please confirm that your project is private. You can edit the markdown cell from `[ ]` to `[x]` to check a box.

Answer (6.0):

- [X] Yes, my project is private.
- [ ] No, my project is not private.

Objective

We are going to revisit Problem 5 using Spark/h2o instead of SQL, with the same datasets.

R or Python?

You are free to use R or Python.

Spark or h2o

You are free to use Spark or h2o.

Question 6.1

Use Spark or h2o via R or Python to calculate the mean drug cost (from `variable total_drug_cost`) per state (from variable `nppes_provider_state`) on the sample Medicare dataset `data/medicare/sample_medicare.csv`. You can remove missing values if any. Sort your answer by alphabetical order of `nppes_provider_state` and show the first 15 entries in your answer.

Answer (6.1):

In [12]:
```
library(h2o) #load h20 libarary
localH2O <- h2o.init(min_mem_size = "32g") #initialize a machine with h20 and
 32g of storage

medicare_path <- "/work/problem6/data/medicare/sample_medicare.csv" # full pat
h to the sample medicare data
medicare_to_h2o <- h2o.importFile(path = medicare_path,
                                  destination_frame = "medicare_from_r") #conv
erts the csv file to h2o type object
```

```
 Connection successful!

R is connected to the H2O cluster:
    H2O cluster uptime:         2 hours 19 minutes
    H2O cluster timezone:       Etc/UTC
    H2O data parsing timezone:  UTC
    H2O cluster version:        3.26.0.2
    H2O cluster version age:    9 months and 14 days !!!
    H2O cluster name:           H2O_started_from_R_rstudio_erm628
    H2O cluster total nodes:    1
    H2O cluster total memory:   30.16 GB
    H2O cluster total cores:    2
    H2O cluster allowed cores:  2
    H2O cluster healthy:        TRUE
    H2O Connection ip:          localhost
    H2O Connection port:        54321
    H2O Connection proxy:       NA
    H2O Internal Security:      FALSE
    H2O API Extensions:         Amazon S3, XGBoost, Algos, AutoML, Core V3, C
ore V4
    R Version:                  R version 3.6.1 (2019-07-05)

Warning message in h2o.clusterInfo():
“
Your H2O cluster version is too old (9 months and 14 days)!
Please download and install the latest version from http://h2o.ai/download/”

  |======================================================================| 10
0%
```

In [13]:
```
drug.cost.by.state <- h2o.group_by(data = medicare_to_h2o[c("nppes_provider_st
ate","total_drug_cost")], #point to the data and columns to use
                                   by = "nppes_provider_state", #column used t
o create groups, note this also autosorts the output alphabetically
                                   mean("total_drug_cost"), #take the mean tot
al_drug_cost by the group specified above
                                   gb.control=list(na.methods="rm")) #remove N
A values before computing the mean
```

In [14]:
```
head(drug.cost.by.state, 15) #display the first 15 rows
```

A data.frame: 15 × 2

nppes_provider_state <fct>	mean_total_drug_cost <dbl>
AE	142.565
AK	2084.695
AL	3308.722
AR	3886.480
AZ	3756.158
CA	3996.927
CO	4358.985
CT	4877.835
DC	3457.887
DE	3249.646
FL	3783.652
GA	3484.670
GU	1537.748
HI	3897.899
IA	4739.704

Question 6.2: Create a virtual machine (VM) on the Google Cloud Platform (GCP)

We want to create a VM on GCP, and then install Spark (or h2o) and, at your choice, the R or Python tools to use to use Spark (or h2o). (For consistency and easier troubleshooting, I recommend you use Ubuntu 18.04 LTS).

As mentioned in the notes of Module 6, a step-by-step guide can be found at: https://bit.ly/csci-e-88b-2020-gcp (https://bit.ly/csci-e-88b-2020-gcp).

To install packages, you can use (on an Ubuntu 18.04 system):

```
sudo apt update
sudo apt-get install -y unzip
# https://www.digitalocean.com/community/tutorials/how-to-install-java-with-apt-on-ubuntu-18-04
sudo apt install -y openjdk-8-jre-headless
# To install R packages
sudo apt-get -y install libssl-dev
sudo apt-get -y install libxml2-dev
sudo apt-get -y install libcurl4-openssl-dev
# add other packages as needed...
# sudo apt -y install r-base
```

For example, if you use R, you can install `sparklyr` with:

```
install.packages("sparklyr") # you should have time to grab a cup of tea...
sparklyr::spark_install()
```

You can also use Docker, for example with an image from https://renkulab.io/projects/cchoirat/problem6-image (https://renkulab.io/projects/cchoirat/problem6-image). You can find more information in the GitLab view of the project (https://renkulab.io/gitlab/cchoirat/problem6-image (https://renkulab.io/gitlab/cchoirat/problem6-image)): `Packages --> Container Registry` (https://renkulab.io/gitlab/cchoirat/problem6-image/container_registry (https://renkulab.io/gitlab/cchoirat/problem6-image/container_registry)). You could consider:

```
docker pull registry.renkulab.io/cchoirat/problem6-image:14a9f03
```

Question: Briefly explain the steps you followed and the choices you made (specs of the VM, Docker or not, R or Python, Spark or h2o).

Answer (6.2):

1) I created a Google GPC project and started "instance-2" VM compute engine. I used the following settings: 8CPU, 30GB RAM, Ubuntu 18.04 LTS, and 50g on the boot disk. 8 CPUs will provide an affordable amount of adequate computing speed, 30GB RAM is sufficient for the h2o to initialize, the Ubuntu 18.04 LTS is well documented and user friendly and 50g on the boot disk to allow the machine to boot properly.

2) I then SSH'd into the VM and ran the folling commands to set-up the environment, install the necessary software and create a user account:

```
$ sudo apt update

$ sudo apt install apt-transport-https ca-certificates curl software-properties-com
mon

$ curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add -

$ sudo add-apt-repository "deb [arch=amd64] https://download.docker.com/linux/ubunt
u bionic stable"

$ sudo apt update

$ apt-cache policy docker-ce

$ sudo apt install docker-ce

$ sudo systemctl status docker

$ sudo apt-get update -y

$ sudo apt-get install -y unzip

$ sudo adduser steve

$ sudo usermod -aG docker steve

$ su - steve

$ id -nG
```

3) Next I forked the problem6-image to my Renku account and pulled the problem 6 docker image from Renku to my VM with the following code:

```
$ docker pull registry.renkulab.io/stephen.t.knapp14/problem6-image:14a9f03
```

```
I used the docker image because I am not working with any proprietary data (docker
 has security limitations).  It is also an efficient, 1 step solution that provides
the exact capabilities I need to complete the project.  It does not make sense to r
einvent the wheel if I don't have to.
```

4) Run the docker image:

```
$ docker run -it registry.renkulab.io/stephen.t.knapp14/problem6-image:14a9f03 /bi
n/bash
```

5) Confirm R is working in the docker image:

```
$ R
```

```
R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

```
I made the choice to use R because my R skills are much stronger than my Python ski
lls.
```

6) Load the h2o library inside of R:

```
> library(h2o)
```

Question 6.3

Use Spark or h2o (via R or Python) to answer Question 6.1 again, but on the whole Medicare dataset (that you can get from http://download.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/PartD_Prescriber_PUF_NPI_DRUG_15.zip (http://download.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/PartD_Prescriber_PUF_NPI_DRUG_15.zip). You can use for example `wget` to retrieve the zip file and `unzip` to unzip the archive.

Answer (6.3):

1) Download data to the docker image inside the VM and unzip it:

```
$ wget http://download.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/PartD_Prescriber_PUF_NPI_DRUG_15.zip

$ unzip PartD_Prescriber_PUF_NPI_DRUG_15.zip
```

Confirm the files are there: $ ls

```
   PartD_Prescriber_PUF_NPI_DRUG_15.zip            PartD_Prescriber_PUF_NPI_Drug_15.txt
   PartD_Prescriber_PUF_NPI_Drug_15-SAS-Infile.sas
```

2) Mount the artifacts folder and run the docker:

```
$ docker run -it -v ${PWD}/artifacts:/home/rstudio/artifacts registry.renkulab.io/stephen.t.knapp14/problem6-image:14a9f03 /bin/bash
```

2) Go back into R:

```
$ R
```

3) Load the h2o library and initialize a machine with h20 and 32g of storage

```
> library(h2o)

> localH2O <- h2o.init(min_mem_size = "30g")

I selected 30g here because that is what the VM is capable of when it was set-up.
```

4) Define path to data and create an h2o object from it

```
> medicare_path <- "/home/rstudio/artifacts/PartD_Prescriber_PUF_NPI_Drug_15.txt"

> medicare_to_h2o <- h2o.importFile(path = medicare_path, destination_frame = "medicare_from_r")
```

5) Aggregate the data by group, sort it by state and determine the means by group. Also remove NA values prior to mean calculation.

```
> drug.cost.by.state <- h2o.group_by(data = medicare_to_h2o[c("nppes_provider_stat
e","total_drug_cost")],
                                   by = "nppes_provider_state",
                                   mean("total_drug_cost"),
                                   gb.control=list(na.methods="rm"))
```

6) Print the first 15 entries:

```
> head(drug.cost.by.state, 15)

   nppes_provider_state mean_total_drug_cost
```

1 AA 498.4405

2 AE 2785.6377

3 AK 2876.4162

4 AL 4721.6955

5 AP 679.9736

6 AR 3629.1984

7 AS 133.1600

8 AZ 4073.9198

9 CA 4636.7208

10 CO 3666.7932

11 CT 4833.8620

12 DC 5480.7810

13 DE 4971.0013

14 FL 4770.0165

15 GA 4617.4674

Self-assessment (not graded)

Questions

1. What do you think I was hoping for you to learn through this homework?
2. Did you find anything particularly challenging?

Your answer

1. Basic application of big data tools
2. Creating the VM, docker image, running the docker image, and mounting the drive between the VM and docker image.

Submitting your work To submit the problem set, export the notebook to HTML and upload the file to Canvas. [`File` -> `Export Notebook As...` ->`Export Notebook to HTML`]