Here is homework assignment 2:  
  
Create a linear regression model using a data set of your choice.  
Examples of data sets are: iris, stock data we scraped, or another data set you "acquired."   
Return regression coefficient using the "score" function in scikit-learn's sklearn.linear\_model.LinearRegression.  
  
Due: Thursday, October 4 at 11:59pm

Data Scraping:

<http://nbviewer.ipython.org/urls/raw.github.com/tomoeyukishiro/GA_homework/master/Day_4_LinReg.ipynb>

Get the data

SELECT a.campaign\_id, count(a.campaign\_id) AS num\_actions, b.follows\_count,

SUM(CASE WHEN a.recruiter\_id IS NOT NULL THEN 1 ELSE 0 END) AS num\_recruited\_actions,

SUM(CASE WHEN c.context\_id = a.campaign\_id THEN 1 ELSE 0 END) AS num\_posts,

b.created\_at

FROM action\_credit\_silo.action\_credits AS a

JOIN descolada\_prod.campaigns AS b ON (a.campaign\_id = b.id)

LEFT JOIN descolada\_prod.post\_references AS c ON (b.id = c.context\_id)

WHERE b.id > 82

AND b.launched\_at IS NOT NULL

AND b.deleted\_at IS NULL

AND b.title NOT LIKE '%test%'

AND c.context\_type = 'Campaign'

GROUP BY campaign\_id

ORDER BY num\_actions

DESC

SELECT a.campaign\_id, count(a.campaign\_id) AS num\_actions, b.follows\_count,

SUM(CASE WHEN a.recruiter\_id IS NOT NULL THEN 1 ELSE 0 END) AS num\_recruited\_actions,

SUM(CASE WHEN c.context\_id = a.campaign\_id THEN 1 ELSE 0 END) AS num\_posts,

b.created\_at

FROM action\_credit\_silo.action\_credits AS a

JOIN descolada\_prod.campaigns AS b ON (a.campaign\_id = b.id)

LEFT JOIN descolada\_prod.post\_references AS c ON (b.id = c.context\_id)

WHERE b.id > 82

AND b.launched\_at IS NOT NULL

AND b.deleted\_at IS NULL

AND b.title NOT LIKE '%test%'

AND b.state NOT LIKE 'Inactive'

AND c.context\_type = 'Campaign'

GROUP BY campaign\_id

SELECT a.campaign\_id, count(a.campaign\_id) AS num\_actions, b.follows\_count,

SUM(CASE WHEN a.recruiter\_id IS NOT NULL THEN 1 ELSE 0 END) AS num\_recruited\_actions,

SUM(CASE WHEN c.context\_id = a.campaign\_id THEN 1 ELSE 0 END) AS num\_posts,

SUM(CASE WHEN d.author\_id = b.creator\_id THEN 1 ELSE 0 END) AS num\_auth\_posts,

b.created\_at

FROM action\_credit\_silo.action\_credits AS a

JOIN descolada\_prod.campaigns AS b ON (a.campaign\_id = b.id)

LEFT JOIN descolada\_prod.post\_references AS c ON (b.id = c.context\_id)

LEFT JOIN descolada\_prod.posts AS d ON (c.post\_id = d.id)

WHERE b.id > 82

AND b.launched\_at IS NOT NULL

AND b.deleted\_at IS NULL

AND b.title NOT LIKE '%test%'

AND b.state NOT LIKE 'Inactive'

AND c.context\_type = 'Campaign'

GROUP BY campaign\_id

SELECT count(campaign\_id) AS num\_pers\_campaigns, campaign\_id

FROM personal\_campaigns

GROUP BY campaign\_id;

SELECT a.campaign\_id, count(a.campaign\_id) AS num\_actions, b.follows\_count,

SUM(CASE WHEN a.recruiter\_id IS NOT NULL THEN 1 ELSE 0 END) AS num\_recruited\_actions,

SUM(CASE WHEN c.context\_id = a.campaign\_id THEN 1 ELSE 0 END) AS num\_posts,

SUM(CASE WHEN d.author\_id = b.creator\_id THEN 1 ELSE 0 END) AS num\_auth\_posts, e.num\_pers\_campaigns,

b.created\_at

FROM action\_credit\_silo.action\_credits AS a

JOIN descolada\_prod.campaigns AS b ON (a.campaign\_id = b.id)

LEFT JOIN descolada\_prod.post\_references AS c ON (b.id = c.context\_id)

LEFT JOIN descolada\_prod.posts AS d ON (c.post\_id = d.id)

LEFT JOIN analytics\_db.temp\_num\_pers\_campaigns\_Oct\_9\_sv AS e ON (b.id = e.campaign\_id)

WHERE b.id > 82

AND b.launched\_at IS NOT NULL

AND b.deleted\_at IS NULL

AND b.title NOT LIKE '%test%'

AND b.state NOT LIKE 'Inactive'

AND c.context\_type = 'Campaign'

GROUP BY campaign\_id

SELECT a.campaign\_id, count(a.campaign\_id) AS num\_actions, b.follows\_count,

SUM(CASE WHEN a.recruiter\_id IS NOT NULL THEN 1 ELSE 0 END) AS num\_recruited\_actions,

SUM(CASE WHEN c.context\_id = b.id THEN 1 ELSE 0 END) AS num\_posts,

SUM(CASE WHEN d.author\_id = b.creator\_id THEN 1 ELSE 0 END) AS num\_auth\_posts,

e.num\_pers\_campaigns, f.comment\_count, b.created\_at

FROM action\_credit\_silo.action\_credits AS a

JOIN descolada\_prod.campaigns AS b ON (a.campaign\_id = b.id)

LEFT JOIN descolada\_prod.post\_references AS c ON (b.id = c.context\_id)

LEFT JOIN descolada\_prod.posts AS d ON (c.post\_id = d.id)

LEFT JOIN analytics\_db.temp\_num\_pers\_campaigns\_Oct\_9\_sv AS e ON (b.id = e.campaign\_id)

LEFT JOIN descolada\_prod.comment\_threads AS f ON (c.post\_id = f.owner\_id)

WHERE b.id > 82

AND b.launched\_at IS NOT NULL

AND b.deleted\_at IS NULL

AND b.title NOT LIKE '%test%'

AND b.state NOT LIKE 'Inactive'

AND c.context\_type = 'Campaign'

AND f.owner\_type = 'Post'

GROUP BY campaign\_id

SELECT a.\*, SUM(CASE WHEN b.recruiter\_id IS NOT NULL THEN 1 ELSE 0 END) AS num\_recruiters

FROM analytics\_db.temp\_hw2\_Oct\_9\_sv AS a

LEFT JOIN action\_credit\_silo.action\_credits AS b ON (a.campaign\_id = b.campaign\_id)

GROUP BY campaign\_id;

Had some major collinearity problems – turned out that I mis-coded number of posts.

SELECT a.\*,

SUM(CASE WHEN b.context\_id = a.campaign\_id THEN 1 ELSE 0 END) AS true\_num\_posts

FROM analytics\_db.temp\_hw2\_Oct\_9\_sv AS a

LEFT JOIN descolada\_prod.post\_references AS b ON (a.campaign\_id = b.context\_id)

GROUP BY a.campaign\_id;

SELECT a.title, a.creator\_id, c.country\_code3 AS country, c.city\_name AS city, b.campaign\_id, b.num\_actions, b.num\_pers\_campaigns, b.num\_recruiters, b.num\_recruited\_actions, b.true\_num\_posts AS num\_posts, b.num\_auth\_posts, b.follows\_count, b.comment\_count, b.created\_at

FROM descolada\_prod.campaigns AS a

JOIN analytics\_db.temp\_hw2\_Oct\_9\_sv AS b ON (a.id = b.campaign\_id)

LEFT JOIN descolada\_prod.user\_locations AS c ON (a.creator\_id = c.user\_id)

SELECT a.\*, b.num\_auth\_posts

FROM temp\_hw2\_Oct\_9\_sv AS a

LEFT JOIN temp\_num\_auth\_posts\_sv AS b

ON (a.campaign\_id = b.campaign\_id)

----------------------------------------------

(Term)

saravera:~ saravera$ scp ./Desktop/hw2\_data\_pull.sql analytics-master-vip.prod.com:~/

(Tmux)

saravera:~ saravera$ ssh analytics-master-vip.prod.com

[sara@sfdb9010 ~]$ tmux

[sara@sfdb9010 ~]$ mysql -u analytics -A < hw2\_data\_pull.sql > hw2\_data.txt

(Term)

saravera:~ saravera$ scp analytics-master-vip.prod.com:~/hw2\_data.txt ./

--------------------------------------------------

Change txt files into csv files in R:

setwd("/Users/saravera/Desktop")

users <- read.table("hw2\_data.txt", header = TRUE, quote = "\"", dec = ".", sep="\t", fill=TRUE)

write.csv(users, file = " hw2\_data.csv", row.names = FALSE)

---------------------------------------------------

Do some data manipulations and run regression in R just to see what we get

Copy it from my local desktop to the aws node

scp Desktop/hw2\_data.csv [GA2@54.200.59.194:~](mailto:GA2@54.200.59.194:~)/GA\_Homework/hw2

Now log into aws node

ssh [GA2@54.200.59.194](mailto:GA2@54.200.59.194)

python

Run hw2.py script (line by line in Python to check that it’s working as I go, or I can write an entire script and run it using ‘python hw2.py’

To push to git

cd GA\_github/GA\_Homework/hw2/

ls

git status

git add hw2.py

git commit

git push

And then check github.com to make sure it’s there!