2017(10-12月) **分析|数据科学类** **硕士** **全职**@**Linkedin** - 网上海投 - **技术电面**|**Fail**在职跳槽

|  |
| --- |
| 不说废话，直接上题：  -- INPUT DATA -- table t1.鏈枃鍘熷垱鑷�1point3acres璁哄潧 -- date  qty\_prod\_a  qty\_prod\_b  qty\_prod\_c -- 1/1/2013 100    200          300 -- 1/2/2013 101      0      301 -- 1/3/2013 102    202      302 鏉ユ簮涓€浜�.涓夊垎鍦拌鍧�.  -------------------------------------------------------------------------------------------------------- -- QUESTION --  # you have a table t1 with the quantity of product A, B, and C sold per day, as shown above --  # there are only 3 possible products in the table: A, B, and C --  # write SQL code to reformat the data as shown below --  # the resulting data should be in 3 columns: {date, product name, quantity sold}. |

|  |
| --- |
| data qty;  input date qty\_prod\_a qty\_prod\_b qty\_prod\_c;  datalines;  1 100 200 300  2 200 0 300  3 300 400 500  ;  run;  proc print data=qty;  run;  proc sql;  create table pre1 as  select sum(qty\_prod\_a) as sum\_a, sum(qty\_prod\_b) as sum\_b, sum(qty\_prod\_c) as sum\_c  from qty  group by date  order by date  ;  quit;  proc sql;  create table final as  select t\_a.\* from (select date, "a" as product\_name, sum\_a as quantity\_sold from pre1) as t\_a  union  t\_b.\* from (select date, "b" as product\_name, sum\_b as quantity\_sold from pre1) as t\_b  union  t\_c.\* from (select date, "c" as product\_name, sum\_c as quantity\_sold from pre1) as t\_c  ;  quit; |

[**[面试经验]**](http://www.1point3acres.com/bbs/forum.php?mod=forumdisplay&fid=259&filter=sortid&sortid=311)**twitter 电面1**

2017(10-12月) 分析|数据科学类 博士 全职@Twitter - 网上海投 - 技术电面 |Pass在职跳槽

lz面的广告组ds，第一面是个国人妹子，面试完不到15分钟hr发邮件说通过。题目如下：1. 概率：

For a twitter user, their daily log in prob is: ⅓, if they log in yesterday and log in today, the probability they spend more than 5 mins on their page is ⅘. If a user did not log in yesterday and log in today, the probability they spend more than 5 mins on their page is ½

Question: what is the probability of a user log in for both today and yesterday if we see a user spend more than 10 mins on their page today

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | 1, log in both yesterday and today | 2, log in yesterday but not today | 3 log in only today | Not log in for two days |
| p | (1/3)\*\*2=1/9 | 1/3\*2/3=2/9 | 2/9 | (2/3)\*\*2=4/9 |

Y=1 to indicate log time>5 minutes

Y=0 to indicate <5 minutes

The conditional probability of P(X=1|y=1)=(P(x=1,y=1)/p(y=1))=4/9

2. sql

Table 1: Campaigns

Account\_id (AID) | Campaign\_id (CID)

1     123

1    234

2    235

Table 2:  Spend

Campaign\_id (CID) | Date | Spend\_amount | Currency

123  2017-08-01 200 USD

123 2017-08-02 150 USD

234 2017-09-01 500 USD

235 2017-07-01 100 CAD

Table 3: Exchange\_rate

Currency | Rate(to USD)

CAD  0.79

USD 1.00

. 涓€浜�-涓夊垎-鍦帮紝鐙鍙戝竷

q1: CID, total spend in USD

q2: AID, number of days from first spend date to highest spend date

**Measurement DS跪经**

2017(7-9月) 分析|数据科学类 硕士 全职@Pinterest - 内推 - 技术电面 |Fail在职跳槽

回馈地里，攒点人品。之前在地里找P家DS的面经没找到。我面的这个是在marketing science组的DS。

其他顺利回答的记不得了，我就在两个问题上磕了一下记得很清楚：

1. Random forest里面可以有colinearity吗？

When building random forest trees, for each tree, it selects a subset of features to build the tree. By doing this, it avoids that the strong feature dominates each tree. So that we will reduce the variance by averaging all the trees.

2）ANOVA, regression, chi-square的区别。。。跪在chi-square上，紧张了也好久没用了，几秒钟之内居然想不起来chi-square的原理，这个问的不偏全怪自己。

这个组的HM感觉不是很有经验，感觉她手里有一个各种stat概念性问题的list（可以马上google出来的那种），选着问，问题和问题之间完全没有联系。这种问题new grad来回答比较有优势吧。

ANOVA and regression are parametric tests.

ANOVA: categorical independent variable, continuous dependent variable. Used to compare whether all the groups means are the same.

Regression: continuous dependent and independent variables. Use to study the relationship between the dependent variable and independent variables.

Chi-square goodness of fit test: non-parametric test. Use the **chi**-**square goodness of fit test**to determine whether observed sample frequencies differ significantly from expected frequencies specified in the null hypothesis.

[**[面试经验]**](http://www.1point3acres.com/bbs/forum.php?mod=forumdisplay&fid=259&filter=sortid&sortid=311)**twitter 2面**

2017(10-12月) 分析|数据科学类 博士 全职@Twitter - 网上海投 - 技术电面 |Fail在职跳槽

推特2面，这次考的是case，国人大哥面的，挂了。。。

. 鐣欏鐢宠璁哄潧-涓€浜╀笁鍒嗗湴

考了两个case，不多废话了，上题目

1. 推特要给business们发广告，想从第三方那里买各个business的email联系方式，买一个花5块。他们买完后会给这些business发10次email，每一次的conversion rate是0.1%，每个转化带给推特500块的收益。问要不要买

2. 有个ds做了一个模型，他用clustering的方法把100k个business给group成5个segment，然后他辞职走人了，模型啥的都不available了，你现在接替他的工作，然后marketing team想给新的100k的business发email campaign，问你该怎么办，如果想用原来的模型，你该怎么办，前提是原来模型的code都没有了，只有最后的结果，也就是business对应的哪个segment。根据你怎么回答的会有一些follow up问题，比如k mean如何选的centroid，如何重建centroid这些

整个过程感觉挺instense的，国人大哥一上来就说我们会run out of time，are you ok with it? 我答的磕磕巴巴，从他的反应来看应该也是个feasible solution，本来以为能混过去，还是给挂了。个人感觉推特的bar还是挺高的。考察的特别全面，比flag面试要难

1) 不买，the actual conversion rate is <10\*0.1%, so not break even. the actual conversion rate is: 1\*0.1%+99.9%\*0.1%+.......(1-0.1%-99.9%\*0.1%)\*0.1%+......

2) using the clustering info as training data, then build classification model.

The population for each sending the first,…,10th email is n, 0.999n, 0.999^2\*n, 0.999^9\*n

So the profit is 500n\*0.001(1+0.999+0.999\*\*2+…0.999^9)~4.978n<the cost of 5n, so not investment.

[**[面试经验]**](http://www.1point3acres.com/bbs/forum.php?mod=forumdisplay&fid=259&filter=sortid&sortid=311)**Houzz DS + Google QA 挂经**

2018(7-9月) 分析|数据科学类 博士 全职@Google - 内推 - 技术电面 |Failfresh grad应届毕业生

长期从地里收获信息，今天来回馈啦。上周面了Houzz DS和狗家QA，来贡献挂经。

Houzz

Q1: probability of a full house hand? (no jokers, 5 out of 52 cards. The inteviewer explained what is a full house in poker games)

Poker probability

<https://en.wikipedia.org/wiki/Poker_probability>

Q2: SQL. Two tables,

Emails: email\_id int, email\_type int, title string, text string, pulish\_date date

Email\_event\_log: email\_id int, user\_id int, event\_type string (“sent”, “open”, ”click”, etc.), time datetime

Compute open rate for each email\_type. Open rate = open/send in each type. Mentioned that one can open the same email multiple times.

Google

Q1: tell me about the most interesting data project you’ve done. 鐗涗汉浜戦泦,涓€浜╀笁鍒嗗湴

Q2: linear regression: income ~ gender + age + highest degree + years in that degree. Interested in the relationship between income and gender. Do we need to consider multicollinearity? Then what is the influence of multicollinearity? Suppose there is multicollinearity, we only care about the prediction of income. Is the prediction good? And why? How to deal with multicollinearity? Compare different variable selection methods.

Q3: We have web pages with/without ads. In the same browser, there is a survey with 5 options for satisfaction degree towards the page. How do you know which one is better? When I said use socres 1 to 5, the interviewer asked why. If you have the correct metric now, how to perform the test? (details)

Q4: easy data manipulation coding. I used R.