Usually you will expect these types of interview questions. Behavior questions are not listed because they are commonly seen in virtually all kinds of jobs. You are expected to meet up with both soft and hard-core problems.

**Soft Problems**

**Business Cases**

**Type 1: pure business cases (similar to consulting firm interviews)**

1. How to decide the location in Texas to set up a new cell tower (let’s say just one zipcode)
2. Our advertising department decided to invest $5 million in Texas for marketing campaign. How many additional customers can we acquire?
3. Before 2, how do you do due diligence to decide whether to invest in Texas or not?
4. Our website’s traffic is dropping down from 6,000/visit to 2,000/visit, what should you do?
5. Evaluating twitter’s new email campaign: 0.1% response rate, cost is $5 per email address. We can only send up to 10 emails to same user. if they response, twitter can get $500 returns. Is this a good campaign?

**Type 2: approaches to conduct AB testing**

1. Pricing problem: a pillow, original price is 9.99, cost is 4, how to determine sales price. Follow up include: 1. metrics, 2. why after doing linear models, our AB testing still failed (assuming all confounding variables have been controlled)
2. Search engine (make a change)---make the results more relevant to the users. What metrics and how to collect data?
3. AB testing, how to build metrics to test one of new functions on app (a big button called “XXX”)
4. Business Case (30 min), talking about A/B test design (Sephora’s original search engine vs our one)

**Term Explanation (in layman language)**

1. p-value
2. hypothesis testing (t-test)
3. K-means
4. after doing t-test, 95% of confidence of significant difference to a non-statistician
5. Explain R^2 in layman language
6. Explain Central Limit Theorem in layman language
7. explain random forest in layman language
8. explain hadoop in layman language
9. Do you know Bayes Theorem? Can you explain to me? Why it is important?

**Brain Teaser (refer to a quant’s book for more prep)**

1. 5 pirates distribute 100 coins
2. 3+2 hats
3. 10\*10\*10 cube, paint all surfaces, how many unpainted?

**Project Management or More Tricky Behavior Questions**

1. how do you collaborate with others?
2. What if your PM push you to finish a pj in a month?
3. how to design a recommendation pipeline and explain to the clients?
4. if your server failed and a big client’s 1day data is lost. How will you react?

**A more emphasizing should be given to “scenario modeling”, which is a combination of both hard core and soft problems.**

**Scenario Modeling**

1. Credit Fraud case. How do you build the model to determine the credit fraud occurrence? How do you conduct the experiment (A/B testing doesn’t work for their business since it costs too much) in our website?
2. How do you predict the U.S.’s housing price (in each location’s housing)?
3. how to build CLV model? How do you split train/test set?
4. how to distribute the cost on different marketing channel given their previous costs?
5. Demand forecasting for a seasonal product
6. We want to build a model to use credit attributes to predict insurance loss. Why do you think this makes sense? Or how should you interpret?

**Hard Problems**

**Statistical Inference & Probability**

1. 1000 hard drives. After running 6 months, 0 of them is fail. What can you infer? If your engineer colleague told you “Failure rate for hard drives is 50%”. How should you do?
2. Explain p-value=0.07 for H0: miu=10 vs H1: miu>10 both in non-statistical way and probability way
3. For 2’s setting. Originally we can get CI=[8,10]. What if your manager wants margin from 2->0.2. How should you do with the size?
4. Flip coins,14 out of 20 is head. How do you set up the hypothesis testing to test whether it is fair or not? How do you calculate the p-value?
5. A:8 pokers; B: 6 pokers

1 roll-6 sides

Roll Number: take number of chips from other one---i.e. 9,5

End of any round: 1 person has more chips

1. Prob that game goes beyond one round
2. General formula for prob that the game ends in the round n

6. what is the distribution of x1+x2, if they are iid poisson distributed?

7. Flip coins,14 out of 20 is head. How do you set up the hypothesis testing to test whether it is fair or not? How do you calculate the p-value?

**Experimental Design & Causal Inference**

1. What is causality inference? What’s the difference if you use propensity score to express your confounding variables (i.e. inverse Logit(T)=Beta\*Z and put T into the linear regressions) rather than simply put these confounding variables into your final regression models?
2. Evaluate pros & cons for two experimental designs. Goal is to do causal inference to see whether a new drug can decrease the blood pressure level

Design 1: 100+100 people , treatment+control group for 3 months

Design 2: 100 people, randomly split into 2 groups. treatment +control for 3 months, then come back and take a rest for 3 months, then come back and switch the group control+treatment for 3 months

**Linear Regression (most fundamental)**

1. Linear model 4 assumptions
2. For LM, total R^2 is OK, but each variable is not significant? (multicollinearity)? If that’s the case, how would you do?
3. For linear model, what if your y is not normal? How should you do?
4. For LM, we have already built a model with y=a+bx1+cx2. If we want to add another x3, under what condition will the coefficients of x1 and x2 won’t change?
5. Assumptions for linear regression. Does R^2=0.9 indicate a good model?

**Time Series & Other Core Statistics Analysis**

1. What is the distributions or representations of the variables in your time series model? (Multivariate Gaussian)
2. In cross-section data, how do you compare different time series model?
3. What is pre-whitening? Why do you do that?
4. What is usual distance measurement? What if the variables (attributes) are highly correlated and can you still use Euclidian Distance? What other possible distance measures?
5. If we use the PCA to do linear transformation of your data points, are their Eclidian Distance still the same? (Yes, because of invariant property)
6. For Vector Space Model, what kind of distance measurement are you using for comparing different documents? Do you think this distance measurement (cosine distance) have the transient property ? i.e. If A &B is correlated, B&C is correlated, can we say A & C is correlated?

**Machine Learning**

1. Tell me the definition of overfitting
2. explain specificity and sensitivity, formula
3. How do you deal with imbalanced data (binary classification, few positive and many negative)? Explain your approach? (use logistic regression)
4. For logistic regression, can you directly use the probability into the test set?
5. For SVM, if you are dealing with multiclassification, how many classifiers do you need? (1v1 or 1v other)
6. What are the metrics you use to compare different learning methods? What is the difference between prediction error and AUC? When to use?
7. For ranking problem (i.e. credit score), what is the metric for evaluation? (Z-test)
8. Formula for LASSO? What’s the purpose for LASSO?
9. Bias-variance off
10. AUC
11. Decision Tree (CART)
12. How to choose split in CART
13. Conjugate Prior
14. Online Learning
15. you cannot see processed data before (2) stream SGD is online…(shuffle your data, read your each of your observation one by one)

15. Regret (I don’t know, cost function in online learning scenario)

16. How to validate your clustering results?

17. Explain the process or steps for complete statistical analysis for a model. Let's assume you have a data-set to do analysis, what steps do you think is important in order to a complete analysis of data-set.

18. How do you benchmark or compare your model against other models? How do you choose the best model from a set of different models?

19.If you use LDA to get topics, how do you evaluate the topic quality? If we want to use precision & recall to evaluate, how do you generate the label for “relevant”?

**Data\_Preprocessing**

1. How do you deal with missing data? What if an important predictor has around 70% missing rate? i.e. y=purchase (0 or 1), x=income(only first 6 has values, others are all missing), age(all), zip (all).
2. What is your general approach to deal with data cleaning or pre-processing before doing stats analysis?
3. A large dataset (>500 variables) with decent number of missing data (sparse). How do you determine whether the missing is at random or not at random?

**Data Manipulation (R,SAS, Python…)**

1. In SAS, we have a table, how should we delete the duplicate records (exactly the same)?
2. In SAS, assuming we have sorted the table, how do we conduct quadratic summation?

i.e. In a sheet, we have day, amount. How can we compute the summation of amount in each day (but from day 2, we want to know the summation for the amount within each day?)

|  |  |  |
| --- | --- | --- |
| Day | Amount | Sum |
| 1 | $ | D1 |
| 1 | $ | D1 |
| 2 | $ | D1+D2 |
| 2 | $ | D1+D2 |
| 3 | $ | D1+D2+D3 |
| 3 | $ | D1+D2+D3 |

3. (1) what does var x $10. mean?

(2) transfer string to numeric (input is not enough)

(3) what is the difference between these two commands:

x=a+b, x=sum(a,b) hint: considering missing value case

(4) we have a table like this:

index state

1 Cal

....

10 Cal

2 Flo

....

8 Flo

5 NY

....

10 NY

write sas code, return first observation in each state

4.

**SQL(extremely important)**

1. Tell me the difference among left join vs Inner join vs Full outer join
2. How to select the elements in Table A but not in Table B?
3. One table with two columns: employee name + location. find all the employee names whose location is same with Joe’s (Self join or subquries)
4. For two tables. Salesmen table: salesman name +salesman id. Order table: salesman id (many replicants)+order id+order amount. Find all salesmen who have ordered more than once. (having, join)
5. One table with three columns: product id, date and price. find the average prices for the different products?
6. Table A Table B

num 1 num2

1 1

2 2

3 2

3

(1) what's the result for this query: select \* from A join B on (A.num1=B.num2)

(2) What's the result for this query: select\* from A join B;

(3) what if I change to another table and run the same query with (2):

A B

1 1

2 2

3 2

4 3

7. a table looks like this:

yyyymm pop

200001 40,000

....

201012 342342

return a table with yyyyymm, pop, but each row is the month with largest population

8. 1 table, customer\_id,order\_id,order\_total,order\_date (each transaction is unique)

1. select sum order in each date
2. select highest order\_sum for each date
3. which date have total order greater than 1 million?
4. find most popular item?

9. rank or row\_number function? difference between having and qualify?

10. TableA

UID Date Spend country

123 1/21/2015 20.1 US

123 1/22/2015 50.7 US

123 1/23/2015 30.1 US

234 1/2/2015 0.15 FR

… … … …

234 1/8/2015 0.14 FR

… … … …

TableB

UID Signup\_date

123 1/1/2015

234 12/2/2014

1. UID, Max(Spend) on a given value
2. first signup date per country.
3. Maximum spend date, per user

11. I. Members  
 member\_id | join\_date  
123 | 2009-01-04  
234 | 2011-12-01  
…  
  
II. Pages  
page\_id | page\_group | is\_active  
0 | A | Y  
1 | A | N  
2 | B | Y  
…  
  
III. Page\_views  
member\_id | page\_id | session\_id | date  
123 | 0 | 52347 | 2013-02-04  
234 | 2 | 7223 | 2011-01-04  
234 | 1 | 7223 | 2011-01-04  
234 | 1 | 8435 | 2011-01-05  
…

1. return the average page views for each page group
2. simple join table II and III based on ID?
3. Top 5 most reviewed page\_views for each page group?

12. user\_walkin table

• user\_id:int

• day (date format, ‘2015-01-01’)

• qty:int

1. Write me a query that computes the total number of walkins for each user in Nov 2014

2 columns: user\_id, Nov 2014 walkins

(2) Return a list of users whose Dec 2014 walkins were zero or less than half their Nov 2014 walkins: 1 column: user\_id

13. SQL question:

Table 1 search: pageid, query, time\_search;

Table 2 click: pageid, time\_click;

We want to get a third table: query, click\_through\_rate(no. of click/no. of search)?

**Hadoop**

1. Suppose we have 1 data center containing informations per user. And we have many clusters that stores all the user information. How to compute the grand mean of number of queries per user? (Mapreduce concepts, use user id as key during Map processes)
2. Write a map-reduce function, mapReduce(), that counts repetition of each character in alphabetical order for a line of text.

**Pure Coding**

1. Given N points in a 2D plane, return the equation of a line which passes maximum number of points
2. 2sum
3. merge sort
4. merge 2 sorted array
5. swap elements in a list
6. linux shell scripting, head, tail
7. if you have a file which like this:

0 aasdfsfsa

0 asdfsafadsfa…

how do you transfer all the 0 to 1 in shell?

1. Explain hash table
2. Binary Search Tree find the node
3. 2 lists. 1st list: Amazon Prime User names 2nd: Users who registered their devices. Find their common list.
4. If these lists are stored in local database. 你如何调用你写的function?
5. use hashtable to achieve 10.
6. 如果你两个list都很大memory 都读不进去怎么办
7. Create a function that takes as input the nodes and edges of a graph and returns the Degree Centrality and Closeness Centrality, or Betweeness Centrality of the graph. You can assume type of graph.
8. Suppose we have multiple users writing different queries, stored in a file called log.txt:

(user1, not shown in thelog file) “hadoop”,”java”,”python”,”splunk”

(user2) “scala”,”C++”,”C#”,”go”

(user3) “hive”,”java”, “python”

…..

1. do word count for each unique word
2. compute the probability of each word
3. find all possible doublets in each line( i.e. hadoop, java is a doublets, but hadoop,scala is not), all doublets are unique (a,b is ok, but not b,a or a,a, b,b). Compute the frequecy of the doublets
4. compute conditional probability of (hadoop, java) given hadoop

16. Unix shell scripting

1. read file
2. store last 5 lines for all files (log1.txt, log2.txt….log100.txt)

17. Give a binary tree. You do traversal from top->bottom, left->right? Describe the algorithm and write down the code.

18. In a streaming context, compute the variance