Questions from Sriram Natarajan

1.  Give an example of a SQL query  that resembles a query that you have actually used in some project. Pick an example that is challenging and in 1 paragraph explain what this does.

Please refer to the attached file.

2a. Consider the string "XXXXYYYZZX" with 10 characters. Write a function (R or Python or both) that takes a string like this and returns the count of each character as the output [("X", 5), ("Y",3"), ("Z", 2)].  The  function should work properly on empty string "" and the output should be [].  
2b. Same input as 2a "XXXXYYYZZX" but the count is the number of consecutive characters. Output is [("X",4), ("Y",3), ("Z",2), ("X",1)].

Please refer to the attached file.

3. You have a 5 faced dice with numbers on the 5 faces.  An experiment was done with 100 consecutive tosses (each toss results in a number 1-5). The sum across 100 tosses = 400. Is the dice fair?  Use probability theory OR computer simulation (or both) to arrive at your conclusion from a hypothesis testing perspective.



Answer:

Use probability to make conclusion the die is not fair

* The total number of rolls = 100
* The expected number of times each side should have come up for a fair die  = 100/5 = 20
* If the die is fair, the sum across 100 tosses = 20 \* ( 1 + 2 +3 + 4 + 5) = 20 \* 15 = 300
* The die is not fair, if the actual sum across 100 tosses <> 300

Use Pearson's χ2 test whether a die is fair

# I adapted the idea from the article “[How can I test whether a die is fair?](https://rpg.stackexchange.com/questions/70802/how-can-i-test-whether-a-die-is-fair)”.

* Hypothesis

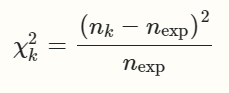
H0: The die is fair

H1: The die is not fair

* The total number of rolls = 100
* How many times each side came up

The number of times side 1 came up n1, the number of times side 2 came up n2, and so on up to n5 for a die with 5 faces.

* Calculate the total number of rolls, i.e.  N=n1+n2+n3+n4+n5.
* Calculate nexp the expected number of times each side should have come up for a fair die Nexp = 100/5 = 20
* Calculate χ2 for each face:



* Add up all the results from the previous step to obtain the test statistic χ2
* The degrees of freedom = (number of faces -1) = 4
* Determine the alpha level to get the critical value of the critical region
* If χ2 >= the critical value, reject H0. Otherwise, accept H0.

 4. Comment on the following statements about logistic regression.

(i) Logistic regression is a linear model.

Logistic regression is not a linear model.  Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters.

(ii) Why do we need a logistic function? What is the purpose?

Linear Regression can’t handle non-continue feature variable.

* Linear regression is used for predicting the continuous dependent variable using a given set of independent features. Linear regression is used to solve regression problems. Linear Regression is all about fitting a straight line in the data.
* Logistic Regression is used to predict the categorical dependent variable. Logistic regression is used to solve classification problems. Logistic Regression is about fitting a curve to the data.

(iii) Are we always guaranteed to find the best (most optimal) solution in the case of logistic regression. Why?

Logic regression would not guarantee to find the best (most optimal) solution.

A frequent problem in estimating logistic regression models is a failure of the likelihood maximization algorithm to converge. In most cases, this failure is a consequence of data patterns known as complete or quasi-complete separation. For these patterns, the maximum likelihood estimates simply do not exist

 5.  Across all your projects pick one example of something you did that was innovative (in your opinion) and describe it (without breaking confidentiality) in 1 or 2 paragraphs.

When MBNA was acquired by Bank of America (BofA) , I was assigned to convert their model codes written in SAS to fit BofA environment. After converting one code which was trying to get/manipulate data to fit models, I did the test. The code ran slowly and the performance was not acceptable.

After investigations, research, and testing, I came out a two phases solution to modularize the code. Phase 1 was ETL part for data extraction, manipulating data and loading the cleaned data to data warehouse. The code of phase I was written in Unix. Unix code was much better for the ETL than SAS code. The phase II was model build which would get data from phase I and build models.

The performance was much better, and this approach was easy to maintain or handle job failures. BofA adapted this approach and gave a nick name PnP (Plug and Play).

6. Describe, in 1 page or less, how Autosys JIL files work.

* Autosys **JIL (Job Information Language)** is used to define a job in Autosys.
* There are three types of jobs: command, file watcher, and box.

1. Command jobs execute commands.
2. Box job

* Box jobs are containers that hold other jobs (including other boxes).
* A box job can be used to organize and control process flow. The box itself performs no actions, although it can trigger other jobs to run.
* Boxes can be put inside of other boxes.
* By default, a box will return a status of FAILURE only when all jobs in the box have run and the status of one or more of the jobs is “failure.”

1. File watcher jobs watch for the arrival of a specified file to starts a process that will check the presence of a particular file in the operating system.

* The status of Autosys job is inactive, active, starting, running, success, failure, terminated, restart, on-hold, on-ice.
* Runs the Job Information Language (JIL) processor to add, update, and delete AutoSys jobs, machines, monitors, and reports.
* The jobs that have the attribute run\_calendar will run on the day mentioned in the calendar in a specified time zone.
* If job failed, it could rerun a couple of times and send an email regarding the failure.
* Example of Jil with type CMD

insert\_job: SAP\_UAT\_MU03\_C job\_type: cmd

command: /local/SAP/processCheckUAT.sh

machine: MU03-UAT

owner: admin@MU03-UAT

permission: gx,wx,mx,me

days\_of\_week: all

start\_times: “15:00, 14:00”

description: “Job used for Run testing of process”

alarm\_if\_fail: 1

max\_exit\_success: 1