THE DATA MONK PRESENTS

A COMPLETE DATA SCIENCE INTERVIEW WITH 100+ QUESTIONS

Data Science or Data Analyst interviews are all about
1. SQL
2. R/Python
3. Statistics
4. Visualization

This book contains questions from all these domain, we have kept the sequence as random because that's what it is in an actual interview.

Few of our best selling books are:-

1. The Monk who knew Linear Regression (Python): Understand, Learn and Crack Data Science Interview
2. 100 Python Questions to crack Data Science/Analyst Interview
3. Complete Linear Regression and ARIMA Forecasting project using R
4. 100 Hadoop Questions to crack data science interview: Hadoop Cheat Sheet
5. 100 Questions to Crack Data Science Interview
6. 100 Puzzles and Case Studies To Crack Data Science Interview
7. 100 Questions To Crack Big Data Interview
8. 100 Questions to Learn R in 6 Hours
9. Complete Analytical Project before Data Science interview
10. 112 Questions To Crack Business Analyst Interview Using SQL
11. 100 Questions To Crack Business Analyst Interview
12. A to Z of Machine Learning in 6 hours
13. In 2 Hours Create your first Azure ML in 23 Steps
14. How to Start A Career in Business Analysis
15. Web Analytics - The Way we do it
16. Write better SQL queries + SQL interview Questions
17. How To Start a Career in Data Science
18. Top Interview Questions And All About Adobe Analytics
19. Business Analyst and MBA Aspirant's Complete Guide to Case Study - Case Study Cheat sheet
20. What do they ask in top Data Science Interviews?

Do check out our website www.thedatamonk.com for Interview Questions,

SQL, Python, Case Studies and many more useful resources.

To start this book all you need is a basic understanding of SQL and Python. You can try your hands on these questions even if you are not that confident with the syntax.

If you are able to answer more than 70% of these questions then you are eligible to start applying for Data Science or Data Analyst profile :P

Back story – There are two characters in this book, **Kishmish Rajput (KR)** who is a candidate for the role of Data Scientist. The interview took place on 27th Oct'19 in Gurugram.
This was a stress interview which went on for close to 4 hours covering the surface of SQL, Python, Statistics, Linear Regression, and MS Excel.
Interviewer is the author of the book.
A total of 100+ questions were discussed.

Candidate – Kishmish Rajput
Location – Thoraipakkam, Chennai
Date – 27th Oct'19
Time – 10:00 AM

Interviewer : Hi myself Nitin and am a Data Scientist over here for the past 8 months. I look after the supply chain analytics for the firm, Can you walk me through your resume?
C: Myself Kishmish Rajput, born and brought up in Bangalore. I did my graduation in Computer Science stream from National Institute of Technology, Allahabad. I have a total of 2 years of experience in the Data Science domain. I am proficient in SQL, MS Excel, Python, Statistics and Predictive modeling.

I: Which is your current employer?
KR: Currently I am working for Sigma Mu

I: Sure! This round will mostly revolve around questions on SQL, Python, Statistics and Spreadsheet. You can pick the language of your choice from R or Python. You need not worry about the syntax as long as you are sound with the logic. Cool ?
KR: Yes !!

I:**So, Can you tell me the order of execution of SQL commands?(Q. 1)**
KR: Sorry, but are you asking about the order in which queries are written on the interface?

I: No, I am looking for the order in which different clauses are executed at the backend of SQL server.
KR: Okaay, It starts with FROM and JOIN, followed by WHERE, GROUP BY, HAVING,SELECT, DISTINCT,ORDER BY

Answer:
The SQL clauses are executed in the following order
FROM and JOIN
WHERE
GROUP BY
HAVING
SELECT
DISTINCT
ORDER BY

I: Sounds good to me. We have two columns(Revenue and Cost Price) in a table like below, get me the Profit column

Revenue	Cost Price	Profit
100	Null	100
200	20	180
300	50	250
Null	50	-50

KR: Direct adding the column values will not work, right?
I: Yeah, You can't add NULL to an integer

KR: I can either use a case when or a coalesce function
I: Can you write the query with either of it? **(Q. 2)**

KR:
Select (coalesce(Revenue,0)+coalesce(CostPrice,0)) as Profit
From Table

I: Now I want you to solve four questions with the same table(Table Name Employee) with columns as (EmployeeName, Department, and Salary)

Get a standard solution which can cater the duplicate values.
1. Get me the highest salary from the table
2. Second highest from the table
3. Highest salary from each department
4. Second highest salary from each department

KR:
Highest Salary(Q. 3)

```
SELECT max(salary)
FROM emptable
```

Second Highest(Q. 4)

```
SELECT max(salary)
FROM emp_table
WHERE salary < (SELECT max(salary)
        FROM emptable);
```

Highest from each department(Q. 5)

We will be using dense rank to get it

```
WITH x AS (
   SELECT DeptID, EmpName, Salary,
        DENSE_RANK() OVER(PARTITION BY DeptID ORDER BY Salary
DESC) AS RowNum
      FROM EmpDetails
)
SELECT DeptID, EmpName, Salary
   FROM x
   WHERE RowNum = 1;
```

Second highest from each department(Ques 6)

```
WITH x AS (
   SELECT DeptID, EmpName, Salary,
        DENSE_RANK() OVER(PARTITION BY DeptID ORDER BY Salary
DESC) AS RowNum
      FROM EmpDetails
)
SELECT DeptID, EmpName, Salary
   FROM x
   WHERE RowNum = 2;
```

I: That sounds good to me, Tell me why did you use Dens rank? And **What is the difference between Rank, Dense Rank and Row Number(Q. 7)**
KR: Well Dense Rank will take care of the possibility that two employees can have the same salary which might be the highest in some case

Row number simply numbers each row. Rank gives discontinuous rank, Dense Rank gives continuous rank

Salary – 10000,10000,9000,9000,8000
Row Number – 1,2,3,4,5
Rank – 1,1,3,3,5
Dense Rank – 1,1,2,2,3

In this case Dense Rank was the best choice

I: Okaay, I have table with Employee Id, Employee Name and Manager Id where information of each employee is stored. Help me create a table with all the employee Names and Manager Names(Manager is also an employee da :P) Ques. 8

Employee_Name	Employee_Id	Manager_Id
A	1	2
B	2	3
C	3	2
D	4	3

I want you to create a table like the one given below

Employee_Name	Manager_Name
A	B
B	C
C	B
D	C

KR: We can use a self join to solve this
I: But which join will you use in the self join

KR: Sorry, I did not follow you
I: Basically, in a self join you join two table and you can do it using inner, left, right, etc. join.
I want to know if you will be using an inner join or any other join?

KR: It depends
I: On what ?
KR: Whether there is a possibility of having an employee without a manager ex. in case of CEO, he should not have a Manager, right?
I: Correct, So which one will you go with in this case?
KR: A left join should do

I: Great !! Get me the query**(Q. 9)**
KR:

```
SELECT e1.Employee_Id EmployeeId, e1.Employee_Name
EmployeeName,
    e1.Manager_id ManagerId, e2.Employee_Name AS ManagerName
FROM   Employee e1
    LEFT JOIN Employee e2
    ON e1.Manager_id = e2.Employee_id
```

<When I take an interview, I mix a lot of questions in between two questions. If a candidate is stuck in solving a question in SQL then I might ask a statistics/Python question to check his/her presence and to ease the pressure. These are mostly very basic questions which every Data Science enthusiast should know>

I: What percentage of value lies between Mean and one Standard deviation(both positive and negative)(Ques. 10)
KR: Around 68%

I: Between Mean and 2 Standard Deviation(both positive and negative) (Q. 11)
KR:95%

I: Mean and 3 Standard Deviation(both positive and negative) (Q. 12)
KR: 99%

I: Great, Now to get the third highest salary without using any window function? (Q. 13)
KR:

*Select * from Employee a Where 3 = (Select Count (distinct Salary) from Employee where a.salary<=b.salary*

I: Cool, I liked the way you used distinct :P
KR: yeah

I: What is the syntax of Vlookup? (Q. 14)
KR:

VLOOKUP(lookup_value,table_array,col_index_num,[range_lookup])

I: . What are the absolute measures of dispersion?(Q. 15)
KR: I don't know

I: You know the answer, okay, what is a dispersion in layman terms?
KR: The distribution of data points

I: Awesome, how do you measure this distribution?
KR: Variance?

I: Correct, What else?
KR: Standard Deviation?

I: Nicee, what more?
KR: Range, Inter quartile Range

I:Cool
Answer:
a. Range
b. Inter Quartile Range
c. Mean Deviation
d. Variance
e. Standard Deviation

I: What is the result of following query? (Q. 16)
KR:
select case when null=null then 'Amit' else 'Rahul' end from dual;

The null=null is always false. so the Answer of this query is Rahul.

I: Since this is the time of Diwali, I have a simple case study for you

A client has a Diwali-themed, e-commerce shop that sells five items. What are some potential problems you foresee with their revenue streams? (Q. 17)

KR: Give me a moment, according to me these are the potential problems of the shop
a. The immediate issue with the client's revenue stream is that it will take a severe hit once the holiday season is over.
b. How to generate revenue outside of the holiday season would be a key point to address with the client.
c. The other concern is with only offering five items.
d. The client is severely limiting their opportunity to generate revenue
e. A couple of bad reviews might create a lot of problem for them as they have very limited items

I: Okay, One more case study for you

Taj Group of Hotels is planning to start a new branch, What are the parameters it should consider to find the appropriate place?(Q. 18)

The following points were discussed with Kishmish:-

a. Find out the place where people have mostly searched for 5 or 7 star hotels
b. Find the place where the average annual income is high, may be Bangalore, Pune, Delhi, Hyderanad, etc.
c. Look for that place which is known for tourism as it will attract foreign customers
d. Look for that area which has good facilities around like popular restaurants, pubs, malls, etc.
e. Look for that city where there are all the necessary facilities like airport near the city, railway station, etc.
f. Look for that city where you can get good service from third party vendors for basic services like laundry, service employees, security service, etc.

I: Now we will slowly move towards Machine Learning thing, but will keep on solving SQL simple querie. Okay?
KR: Sure

I: What do you mean when I say "The model has high accuracy in Training dataset but low in testing dataset"(Q. 19)
KR: Well it means that the model is over fitting i.e. it is taking even the rare cases in consideration and fitting it with the output but is unable to work on new cases. It is mostly the case with Random forest where a number of trees are made on each of the different combination which results into very high accuracy on training dataset

I: What is the difference between NVL and NVL2? (Q. 20)
KR: I have not heard about NVL or NVL2 before

Ans.

In SQL, NVL() converts a null value to an actual value. Data types that can be used are date, character and number. Data type must match with each other i.e. expr1 and expr2 must of same data type.

NVL (expr1, expr2)
expr1 is the source value or expression that may contain a null.
expr2 is the target value for converting the null.

NVL2(expr1, expr2, expr3) : The NVL2 function examines the first expression. If the first expression is not null, then the NVL2 function returns the second expression. If the first expression is null, then the third expression is returned i.e. If expr1 is not null, NVL2 returns expr2. If expr1 is null, NVL2 returns expr3. The argument expr1 can have any data type.

I: What is the use of FETCH command?(Q. 21)
KR: I know this one

The FETCH argument is used to return a set of number of rows. FETCH can't be used itself, it is used in conjuction with OFFSET.

```
SELECT column_name(s)
FROM table_name
ORDER BY column_name
OFFSET rows_to_skip
FETCH NEXT number_of_rows ROWS ONLY;
```

I: OFFSET command?(Q. 22)
KR: The OFFSET argument is used to identify the starting point to return rows from a result set. Basically, it exclude the first set of records.
OFFSET can only be used with ORDER BY clause. It cannot be used on its own.
OFFSET value must be greater than or equal to zero. It cannot be negative, else return error.

```
SELECT column_name(s)
FROM table_name
WHERE condition
ORDER BY column_name
OFFSET rows_to_skip ROWS;
```

I: What is the difference between View and Store Procedure? (Q. 23)
KR: Answered

Providing a more detailed answer

View
Does not accept parameters
Can be used in FROM clause. Can be used as a building block in larger query
Contains only Select query
Cannot perform modification to any table

Store Procedure
Accept Parameters
Cannot be used in FROM clause. Hence, cannot be used a building block in larger query
Can contains several statements, IF-ELSE, Loop etc
Can perform modification to one or several tables

I: What is the difference between COUNT(*) and COUNT(ColName)? (Ques. 24)
KR: COUNT(*) count the number of rows in result set while COUNT(ColName) count the number of values
in column ignoring NULL values.

I: What is Partitioning? (Q. 25)
KR: SQL Server supports table and index partitioning. Partitioning is a way to divide a large table into smaller,
more manageable parts without having to create separate tables for each part. Data in a partitioned table is physically

stored in groups of rows called partitions and each partition can be accessed and maintained separately. Partitioning is
not visible to end users, a partitioned table behaves like one logical table when queried.

I: Why is it important ? (Q. 26)
KR: You can transfer or access subsets of data quickly and efficiently, while maintaining the integrity of a data collection
2. You can perform maintenance operations on one or more partitions more quickly. The operations are more efficient because they target only these data subsets, instead of the whole table.
It is mostly intended to aid in maintenance on larger tables and to offer fast ways to load and remove large amounts of data from a table. Partitioning can enhance query performance, but there is no guarantee

I:What is a Candidate Key? (Q. 27)
KR: A table can have multiple column (or Combination of Columns) which can uniquely identify a row. This column (or Combination of Columns) are referred as candidate keys. There can be multiple Candidate Keys in one table. Each Candidate Key can qualify as Primary Key.

I: Okay, tell me important conditions for joining two tables on a key? (Q. 28)
KR: Ideally there should have one common column(at least) between the tables like Roll No. in Student and Class_Topper Table (Assuming table names)

I: What Roll_no is in integer in Student table and Character in Class_Topper table, will it join?
KR: No, it won't, the data type of the columns should also be the same

I: Cool, What is NULLIF? (Q. 29)

KR: NULLIF (exp1,exp2): Return NULL if exp1=exp2 else return exp1.

I: IFNULL? (Q. 30)
KR: IFNULL(exp1,exp2): Return exp1 if it is not NULL else return exp2.

I: There is a function called ISNULL, what is it used for? (Q. 31)
KR: ISNULL(exp1,value): Return value if exp1 is NULL else return exp1.

I: Classic interview question before we move forward, What is the difference between WHERE and HAVING clause? (Q. 32)
KR:
Both are used to filter the dataset but WHERE is applied first and HAVING is applied at later stage of query execution.
WHERE can be used in any SELECT query, while HAVING clause is only used in SELECT queries, which contains aggregate function or group by clause.
Apart from SELECT queries, you can use WHERE clause with UPDATE and DELETE clause but HAVING clause can only be used with SELECT query

I: Nice !! You already know about RANK, ROW_NUMBER,DENSE_RANK function, now tell me what is NTILE() ? (Q. 33)
KR:
NTILE(): Divides the rows in roughly equal sized buckets.
Suppose you have 20 rows and you specify
NTILE(2). This will give you 2 buckets with 10 rows each. When using NTILE(3), you get 2 buckets with 7
rows and 1 bucket with 6 rows.

I: Okay, What is lag() function? (Q. 34)
KR: Provides access to a row at a given physical offset that comes before the current row. Use this
function in a SELECT statement to compare values in the current row with values in a previous row as specified by offset. Default offset is 1 if not specified. If Partition By clause is specified then it returns the offset Value in each partition after ordering the partition by Order By Clause.

I: And lead() function? (Q. 35)
KR: Provides access to a row at a given physical offset that comes after the current row. Use this
function in a SELECT statement to compare values in the current row with values in a subsequent row as specified by offset. Default offset is 1 if not specified. If Partition By clause is specified then it returns the offset Value in each partition after ordering the partition by Order By Clause

I: Chalo, let's have a rapid fire

I: SELECT 'NITIN'+1 **(Q. 36)**
KR: Throws error

I. SELECT 'NITIN'+'1' **(Q. 37)**
KR. NITIN1

I: SELECT(SELECT 'NITIN') **(Q. 38)**
KR: NITIN

I: SELECT '1'+1 **(Q. 39)**
KR: 2

I: SELECT 1+'1' **(Q. 40)**
KR: 2

I:SELECT NULL+98 **(Q. 41)**
KR: NULL

I: SELECT 0/9**(Q. 42)**
KR: 0

I:SELECT 0/0 **(Q. 43)**
KR: Throws error as the number is divided by 0

I:SELECT SUM('1') **(Q. 44)**
KR: 1

I: Get the name of all the restaurants that starts with any alphabet between a and k. (Q. 45)

KR: Regular expression?
I: Yeah !!

KR: Okay,

```
SELECT Restaurant
FROM food
WHERE Restaurant LIKE '[a-k]%'
```

<Python Questions ahead – These are basic Python questions which either you should know before going for a Data Science interview or learn it here itself>

I: We will together solve some basic Python questions, if you are more comfortable with R then you can switch the language
KR: I am good with Python

I:What is mutable and immutable data type? Q.46
KR: In a layman term, you can alter the content of a mutable data type after it is created whereas the content of an immutable data type cannot be changed after being created.
Objects like list, set and directory are mutable, on the other hand, int, float, bool, str, tuple and Unicode are immutable.

I:Can you help me understand mutable data type using an example (Q.47)
KR: The below example shows how the mutable list works:-

```
list_example = ['Amit','Sumit','Rahul']
print(list_example)
list_example[1] = 'Kamal'
print(list_example)
```

```
['Amit', 'Sumit', 'Rahul']
['Amit', 'Kamal', 'Rahul']
```

I: Nice, Can you take the same dataset, put it in a tuple and walk me through how an immutable data type rejects the idea of alteration? Q.48
KR:

```
tuple_example = ('Amit','Sumit','Rahul')
print(tuple_example)
tiple_example(1) = ('Kamal')
print(tuple_example)
```

```
File "<ipython-input-47-4a338933da64>", line 3
    tiple_example(1) = ('Kamal')
SyntaxError: can't assign to function call
```

I: How to define a dict? Q.49
KR:
Dictionaries are enclosed in curly braces {} and values can be assigned and accessed using square bracket. Below is an example:-

```
dict_example = {
'India':'New Delhi',
 'Pakistan' : 'Islamabad',
'Sri Lanka' : 'Colombo'
}
print(dict_example)
```

```
{'India': 'New Delhi', 'Pakistan': 'Islamabad', 'Sri Lanka': 'Colombo'}
```

I: What libraries are you comfortable with? Q.50
KR: It depends on the project I am involved in. By the way I have worked with
Pandas, SciPy, Matplotlib, Scikit and Numpy

I: Sure, Can you tell me the difference between *args and **kwargs? When are these used? Q.51
KR: Most of the times when we create a function, we need to specify how many arguments are going to be passed in the function.
*args is used when we don't know how many arguments are going to be passed to a function, or if we want to pass a stored list or tuple to the function.

We use **kwargs in function definitions to pass a keyworded variable-length argument list.
A keyword argument is where you provide a name to the variable as you pass it into the function.

I: Example of *Args? Q.52
First and foremost, the word args is used as a convention and you can very well use some other word in place of args. Following is the example where we use dynamic number of arguments and use them

```
def dynamo(*args):
        for x in args:
        print(x)

dynamo('The','Data','Monk')
```

Output
The
Data
Monk

I:kwargs? Q.53**

KR: def kwargs_example(**kwargs):
for key, value in kwargs.items():
print(key,value)
kwargs_example(first='the',mid='data',last='monk')

Output

first the
mid data
last monk

I: What is the function of describe() method? Q.54

KR: Describe method is implemented by dataset.describe() and it gives us the following results
Count
Mean
Standard Deviation
Minimum
25 percentile
50 percentile(Median)
75 percentile and maximum for each of the numerical columns in the data set

I: How to sort a dictionary by key? Q.55
KR:import operator

```
x = {1: 2, 3: 4, 4: 3, 2: 1, 0: 0}
sorted_x = sorted(x.items(), key=operator.itemgetter(0))
print(sorted_x)
```

[(0, 0), (1, 2), (2, 1), (3, 4), (4, 3)]

I: How to remove duplicates from a list? Q.56
We can use the set() data type to remove duplicates from the list

```
itemList = ['1', '2', '3', '3', '6', '4', '5', '6']
new_set = set(itemList)
print(new_set)
```

{'3', '2', '6', '5', '4', '1'}

I: How to remove duplicates from a list without using set data type? Q.57

KR:

```
items = ['1','2','3','4','4','5','5']
new_list = []
(new_list.append(item) for item in items if item not in new_list]
print(new_list)
```

I: What is the use of range() function? Q.58

KR:

The range() function is a built-in function used for iterating the sequence of number. Example below:-

```
for num in range (1, 5):
   print (num)
```

I: What is a lambda function ? Q.59

KR: A lambda function is a small function that can take any number of arguments but can have only one expression.

Example:-

```
x = lambda a:a+10
print(x(10))
```

I: Write a lambda function that multiplies two arguments Q.60

KR:

```
x = lambda a,b:a*b
print(x(23,45))
```

I: Write pattern or regular expression for all the names ending with t or tt. Q.61

KR:

name_pattern = '(t$|tt$)(,)'

I: What is extend() function? Q.62

KR: Extend() function is similar to append() and is used to extend the given list by appending elements from the iterable.

```
x=[1,2,3]
x.extend([4,5,6,7])
print(x)
```

[1,2,3,4,5,6,7]

I: What is append() function? Q.63

KR:

Append() function is used in a list to add/append objects at the end

```
x=[1,2,3]
x.append([4,5,6,7])
print(x)
```

[1,2,3,[4,5,6,7]]

I:You want to build a model in Python, a simple linear regression model. To create a model you need to have three types of data, training, test and cross-validation dataset. What are these datasets? Q.64

KR: Training dataset(60% of the original data) is the one on which you train your model, since we are talking about Linear Regression which is an example of supervised learning algorithm, so we will have both input and output values in the dataset.
Cross-Validation dataset(20% of the original data) is the one on which you tune your hyper parameter. This is the part of the dataset where you know the output but you use the model created on the training dataset to predict the value of the output variable. Looking at the accuracy which you get, you can tune the model. Basically validation dataset is used to validate the model in place.
Test dataset(20% of the original data) is the dataset which is supposed to follow the same probability distribution and your model needs to perform on it treating it as real data

I: How to split a dataset into train and test in python? Q.65
KR: Splitting a dataset into train and test is one of the initial stage of most of the machine learning models. Following is how you can split dataset in python:-

```
from sklearn.model_selection import train_test_split

X_train,X_test,Y_train,Y_test =
train_test_split(dataframe_name,target_variable,test_size=0.3,
random_state=42)
```

dataframe_name = the complete dataset as a panda dataframe
target_variable = the name of the target variable
test_size = 0.3 denotes 70-30 split of the dataset in train and test
random_state = 42, Look for the explanation in the next question

I: What all visualizations are helpful in EDA? Q.66

KR: Any visualization can help in EDA as long as it is able to provide you insights. But the three most important graphs used by Data Scientists are:-

Histogram
Box Plot
scatter plot
Line chart
Stack bar chart
Heat Map

I: How useful is box plot graph? Q.67

KR: Box plot graph is quite useful as it gives us the following information about the data set:-
Outliers
Min/Max
25,50,75 percentile

Box plot of population distribution grouped by continent

df.boxplot(column='population',by='continent')

I: What will be the output of the below code
word = 'aeioubcdfg'
print word [:3] + word [3:] Q.68

KR: The output for the above code will be: 'aeioubcdfg'.
In string slicing when the indices of both the slices collide and a "+" operator is applied on the string it concatenates them.

I:Use while loop to print from 1 to 50, but skip at 5,10 and 15. (Q.69)

KR:

```
x=0
while(x<51)
    x=x+1
    if(x==5)|(x==10)|(x==15):
        continue
        print(x)
```

I:How to join tables in python? (Q.70)

KR: We use the merge function from pandas library. Following is the syntax for the same

Merge the two table
Table 1 State_population
State_name, population

Table 2 State_Capital
Name_State. Capital

m = pd.merge(left=State_population, right = State_Capital, on=None, left_on='State_name',right_on='Name_State')

We use on when the column names of the key column on which you want to merge the data are different. Left_on and right_on are used to specify the column names in the two tables

I: Explain split(), sub(), subn() methods of "re" module in Python. (Q.71)

KR:

To modify the strings, Python's "re" module is providing 3 methods. They are:
split() – uses a regex pattern to "split" a given string into a list.
sub() – finds all substrings where the regex pattern matches and then replace them with a different string
subn() – it is similar to sub() and also returns the new string along with the no. of replacements.

I: What will be the output of the print(str*3) if str="TheDataMonk"? Q.72

KR: It will print TheDataMonk three times

```
str='TheDataMonk'
print(str*3)
```

```
TheDataMonkTheDataMonkTheDataMonk
```

I:As a Data Scientist, you will come across multiple instance where you need to remove the trailing and leading white spaces. How to strip the string? Q.73

KR:

```
str = "     TheDataMonk    "
print(str)
print(str.strip())
```

```
    TheDataMonk
TheDataMonk
```

I: What is the output of [1, 2, 3] + [4, 5, 6]? Q.74

KR:

[1, 2, 3, 4, 5, 6]

I: How will you remove last object from a list? Q.75

KR:

list.pop(obj=list[-1]) − Removes and returns last object or obj from list.

I: Write a program in Python to reverse a string without using inbuilt function reverse string? (Q.76)

KR:

```
def string_reverse(str1):
rev_str = ''
index = len(str1) #defining index as length of string.
while(index>0):
rev_str = rev_str + str1[index-1]
index = index-1
return(rev_str)
print(string_reverse('1tniop'))
```

I: What is the output of 3*32? (Q.77)**

KR:

27

The order of precedence is ** then *. Thus 3**2 = 9 and then 9*3 = 27.

I: Though you are comfortable with string, so we will quickly check your fluency to work with string (Q.77)

KR:

```
str="TheDataMonk"
print (str)
print (str*2)
print (str[2:5])
print (str[3:])
print (str + ".com")
print ("www."+str+".com")
```

```
TheDataMonk
TheDataMonkTheDataMonk
eDa
DataMonk
TheDataMonk.com
www.TheDataMonk.com
```

I: How to convert a tuple into a list? (Q.78)

KR:

```
tup = ('the','Data','Monk')
list_example = list(tup)
print(list_example)
```

['the','Data','Monk']

I: What is a global and local variable? (Q.79)

KR: The understanding of global and local variable is very important to use the capability of a programming language.

In layman terms the global variable can be used anywhere in the complete program, whereas a local variable is used only in the vicinity of its declaration. Variables that are defined inside a function has a local scope. This means that local variables can be accessed only inside the function in which they are declared, whereas global variables can be accessed throughout the program body by all functions. When you call a function, the variables declared inside it are brought into scope.

I: Explain global and local variable using an example. (Q.80)

KR:

```
summ = 0#Global Variable
def sum(a,b):
    summ=a+b#Local variable
    print("Local Variable",summ)
    return sum
sum(40,50)
print("global variable",summ)
```

Local Variable 90
global variable 0

I:What is the difference between (a-z) and [A-Z]? Q.81
KR: When you specify (a-z), it will only match the string "a-z". But when you specify [A-Z] then it covers all the alphabet between upper case A and Z

I: How is Memory managed in Python? Q.82
KR: In Python, memory is managed in a private heap space. This means that all the objects and data structures will be located in a private heap. However, the programmer won't be allowed to access this heap. Instead, the Python interpreter will handle it. At the same time, the core API will enable access to some Python tools for the programmer to start coding.
The memory manager will allocate the heap space for the Python objects while the inbuilt garbage collector will recycle all the memory that's not being used to boost available heap space.

I: Your Python knowledge is decent, let's check some questions on Excel, I hope you are comfortable with one of MS Excel or Google Sheet
KR: I am fine with Excel

I: Will keep this section quick, okay**? Tell me the formula of HLOOKUP. (Q.83)**
KR:
=HLOOKUP(value to look up, table area, row number)

I: Can you help me extract email id from the text given below? (Q.84)
"Hi myself Nitin and you can reach out to me at nitinkamal132@gmail.com , and am awesomeeeee"
Make sure that the approach is generalized and am not looking for a syntax correct solution, so you can speak your thoughts
KR: Okaay !! I will start with indexing the @, once I get the index of @ I will look for the first space in the right hand side using the RIGHT function, then will extract the index of the first blank space in the left hand side. This will get me the two ends of the email

I: Great !!

Answer
TRIM(RIGHT(SUBSTITUTE(LEFT(A1,FIND (" ",A1&" ",FIND("@",A1))-1)," ", REPT(" ",LEN(A1))),LEN(A1)))

KR: Thanks

I: Can we have another column in a table other than a primary key column which will act as a primary key?(**Q.85)**

KR: Yes we can have another column in the table other than primary key which acts like primary key.But in database design this is not the recommended. One table can have only one primary key. But if you create other column with unique and not null constraint it acts like primary key.

I: You were good with SQL, decent with statistics and Python. Choose any algorithm of your choice.
Warning, be very particular about what you select here :P
KR: I would like to go for Natural Language Processing.

I: Sure ??
KR: No, no, I would go for Linear Regression.

I: Sure, this time?
KR: Yessss

I: Okay, What is a Linear Regression? Q.86
KR: In a layman term, Regression is a way to extrapolate your dataset to predict values on the basis of independent variables.

I: Any example of the same? Have you ever worked on a problem involving Regression?Q.87
KR: If you know the attendance, class test marks, last year records of a student, then you can predict his/her performance in the current semester. If you know the weight, BMI, sugar level, etc. of a patient, then you can predict if he/she will suffer from Diabetes in the future.

I: What are the assumptions of Linear Regression? Q.88
KR: I know this answer, I know this for sure

I: Okayy, then go ahead.
KR: Assumptions of Linear Regression :-

1. Linear relationship between independent and dependent variable
2. Less or no Multicollinearity
3. No auto-correlation – Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of y(x+1) is not independent from the value of y(x).
4. Homoscedasticity

I: What are the residuals? Q.89
KR: In regression analysis, the difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). Each data point has one residual. Both the sum and the mean of the residuals are equal to zero.

I: What should be the value of a good variable which we should include in our model? Q.90
KR:A good variable for the model will have the following attributes:-
i. High Coefficient – t value

ii. Very low Pr(>|t|) value

I: How do you evaluate a Linear Regression model? Q.91

KR:

There multiple ways in which you can evaluate an LR model

1. R Square
2. Adjusted R-Square
3. F-Test
4. RMSE

I: What is R-Squared? Q.92

KR:

R Squared = (SST-SSE)/SST

It indicates the goodness of fit of the model.

R-squared has the useful property that its scale is intuitive: it ranges from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model, and one indicating perfect prediction. Improvement in the regression model results in proportional increases in R-squared.

I: More the predictor, better the R-Squared error. Is this statement true? If, Yes then how to counter this? Q.93

KR:

This is true, that's why we do not use R-Squared error as a success metric for models with a lot of predictor variables. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.

Hence, if you are building Linear regression on multiple variable, it is always suggested that you use Adjusted R-squared to judge goodness of model.

I: What is RMSE?Q.94

KR:

Root Mean Square Error takes the difference between the predicted and actual value and square it before dividing the value with the total number of terms

I: What is a F-Test?Q.95

The F-test evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one is not. An equivalent null hypothesis is that R-squared equals zero.

I: Good good, Let's combine the interview with a few case studies, cool ?
KR: Sure

I: Profit of a company selling mobile back cover is declining. List out all the possible reasons. Q.96

KR:

Following is the way in which discussion proceeded with the interviewer:-
1. The demand itself has declined i.e. customers are not using cover that much. Asked to think more by the interviewer
2. Maybe the competitor is also facing loss which again means that the demand is low. Competitors are making a decent profit
3. Bad Marketing – The company is not putting stalls or shops in a crowded place. The interviewer told that the company was making a decent profit 6 months back
4. Maybe the footfall of the mall or place decreased. Could be(first positive response)
5. Maybe a popular mobile phone shop has shifted somewhere else. Could be(again a so-so response)
6. Maybe the other companies have reduced the price of their product which is why customers are drifting to these companies. The interviewer seemed pleased
7. New technology in the cover market to make covers more durable and the company we are talking about is using the same old technology. Seemed good enough point

8. Since we are talking about back covers, there could be new or trending designs which are not produced by the company
9. The company has not registered on different e-commerce websites and the website they are present on is not doing good business. He looked satisfied with the point

This was one of the few interviews where the candidate performed really well and covered a lot of point

I: How would you design the people you may know feature on LinkedIn or Facebook? Q.97
KR:
1. Find strong unconnected people in weighted connection graph
2. Define similarity as how strong the two people are connected
3. Given a certain feature, we can calculate the similarity based on
   -friend connections (neighbors)
   -Check-in's people being at the same location all the time.
   -same college, workplace
4. Have randomly dropped graphs test the performance of the algorithm
5. News Feed Optimization
6. Affinity score: how close the content creator and the users are
7. Weight: weight for the edge type (comment, like, tag, etc.).
8. Emphasis on features the company wants to promote
9. Time decay: the older the less important

I: How would you suggest to a franchise where to open a new store? Q.98
KR:
-Build a master dataset with local demographic information available for each location.
-Local income levels
-Proximity to traffic, weather, population density, proximity to other businesses
-A reference dataset on local, regional, and national macroeconomic conditions (e.g. unemployment, inflation, prime interest rate, etc.)
-Data on the local franchise owner-operators, to the degree the manager identify a set of KPIs acceptable to the management that had requested the analysis concerning the most desirable factors surrounding a franchise
-Quarterly operating profit, ROI, EVA, pay-down rate, etc.
-Run econometric models to understand the relative significance of each variable
-Run machine learning algorithms to predict the performance of each location candidate

I: How do you think TVF makes a profit? Did moving to it's own website advantageous to TVF?

KR:

Following are the points on which we discussed:-

1. TVF has some 10Million subscriber on Youtube, and it release it's video on Youtube after a week of it's original release on the TVF website. These videos give it a good amount of money to keep the show running
2. The main reason for TVF to move to it's own website was to create an ecosystem comparable to Netflix so that people buy subscription to watch the show.
3. Netflix charges some $9 for subscription, TVF could be planning to launch it's series exclusively to any of these and can get some part of the subscription. Even a dollar per person can get them close to 10Million dollars
4. The estimated revenue of a Youtube channel with 10 Million subscriber is ~500,000 dollars per year.
5. Apart from these, a major chunk of the production cost is taken care by the sponsor of the show. For example Tiago in Trippling, Kingfisher in Pitchers, etc. So the production cost is next to zero for the episodes
6. TVF is also going for it's own website and raising funding to acquire customers and drive them to their website

It's hard to get a $10 subscription, but even a basic subscription or tie-up with some other production can get them a handful of money

I: Kishmish, Do you have any question for me? Try to avoid that regular question “What is the role about” :P
KR: I was about to ask the same question :P
Okay, so is there an opportunity to work with development team in my spare time?

I: Believe me, You would run from the office the moment your work is done :P
Short answer – Yess there is a lot

KR: Any feedback for me?
I: Work on your SQL, Python, Statistics skills. HR will get back to you. Thank you

Overall feedback
Kishmish was **confident**, **had good command on SQL and Python**, **clear on thought in the case studies**.
Overall a strong candidate for the position.

Final result – Hired □

This is the width which you should cover before thinking about switching into Data Science. Once the width is done, deep dive in the domain of your interest.

Keep Learning □

The Data Monk