

## Technical Interview Transcript for DATA QUALITY ANALYST @ TURBOLAB TECHNOLOGIES

### Interviewers :

1) Joseph

2) Sumesh M , Python Developer

Date : 13-05-2022

Candidate Name :Swaroop N C

E-mail : [ncswaroop1997@gmail.com](mailto:ncswaroop1997@gmail.com)

Mobile : +91 99 1616 0739

LinkedIn : <https://www.linkedin.com/in/swaroop-n-c-a09499131/>

### Can you tell me about Yourself

Good morning to you both sir First of all thanks for this opportunity

Myself Swaroop , From Tumkur Near Bengaluru I have completed my B.E. in Electronics and Communication from CIT Tumkur

As I had social service in my mind , I kept 2 years for preparing UPSC Civil Services Exam. cleared Preliminary exam once and Written Mains once

As I couldn't make it through all stages in time I had , I decided to shape my career in IT . After exploring my interests , I decided to get into Data Science domain

Then I joined Data science training institute called Prime Intuit in Feb 2022 Here I have learnt Python , SQL , Statistics , And Data Analytics and course is going on yet will be done by end of this month

Apart from my academics , I have interest in Fitness , Astrology and Stock Market

### Can you explain about the project you have done On Data ?

Sir I will explain about Exploratory data analysis of stock prices using NIFTY 50 & Nifty Next 50 dataset from NSE India website

We read the dataset of NIFTY50 & NIFTY NEXT50 for 2021 year from NSE India Website and put into a Pandas Data-frame

Indexed on Date

Did initial analysis using .info and .describe

Observed on what days NIFTY50 was volatile and how it related to Volatility of Nifty next 50 based on High and Low

Computed mean , median, standard deviation for each month

Computed 30 day moving average

And plotted using Seaborn

Also observed one interesting insight like

When market starts after 2+ days of holidays , relatively more volatility expected ( Ex On Monday


**What Functions have you used in Pandas & illustrate one or 2 on a simple DataFrame example**

```
1 #consider this df as example
2 import pandas as pd
3
4 df = pd.DataFrame(
5     dict(
6         name=['John', 'Jacob', 'Tom', 'Tim', 'Ally','Tim'],
7         marks=[89, 23, 100, 56, 90,30],
8         subjects=["Math", "Physics", "Chemistry", "Biology", "English","Science"]
9     )
10 )
11 df
```

	name	marks	subjects
0	John	89	Math
1	Jacob	23	Physics
2	Tom	100	Chemistry
3	Tim	56	Biology
4	Ally	90	English
5	Tim	30	Science

functions which we can use are

```
1 df.head()
```

name marks subjects 

```
1 df.describe(include='all')
```

	name	marks	subjects
count	6	6.000000	6
unique	5	NaN	6
top	Tim	NaN	Math
freq	2	NaN	1
mean	NaN	64.666667	NaN
std	NaN	33.152174	NaN
min	NaN	23.000000	NaN
25%	NaN	36.500000	NaN
50%	NaN	72.500000	NaN
75%	NaN	89.750000	NaN
max	NaN	100.000000	NaN

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6 entries, 0 to 5
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  -
0    name         6 non-null     object
1   marks         6 non-null     int64
2  subjects      6 non-null     object
dtypes: int64(1), object(2)
memory usage: 272.0+ bytes
```

```
1 df.loc()
```

```
<pandas.core.indexing._LocIndexer at 0x7f8696f26fb0>
```

```
1 df.iloc()
```

```
<pandas.core.indexing._iLocIndexer at 0x7f8696f48530>
```

```
1 df.value_counts()
```

name	marks	subjects	
Ally	90	English	1
Jacob	23	Physics	1
John	89	Math	1
Tim	30	Science	1
	56	Biology	1

```
Tom    100    Chemistry    1
dtype: int64
```

```
1 df.drop_duplicates()
```

	name	marks	subjects
0	John	89	Math
1	Jacob	23	Physics
2	Tom	100	Chemistry
3	Tim	56	Biology
4	Ally	90	English
5	Tim	30	Science

```
1 df.groupby('name')['marks'].sum() # 2 repeated Tim marks has been summed
```

```
name
Ally    90
Jacob   23
John    89
Tim     86
Tom    100
Name: marks, dtype: int64
```

```
1 df.sort_values(by='name')
```

	name	marks	subjects
4	Ally	90	English
1	Jacob	23	Physics
0	John	89	Math
3	Tim	56	Biology
5	Tim	30	Science
2	Tom	100	Chemistry

```
1 df.dropna()
```

	name	marks	subjects
0	John	89	Math
1	Jack	92	Physics

**Download the dataset in given link and read the file into pandas and give some observations and anomalies in the Dataset ?**

```
1 df2=pd.read_csv('/content/qa_test.csv')
2 df2
```

	Product ID	Product Name	product_price	product_url	stock_count	Stock status
0	1	Laptop A	500	https://amazon.com/dp/ABCHDMSH	120	Stock
1	2	Laptop B	\$200	https://amazon.com/dp/DBAHSGSB	24	Stock
2	3	Laptop C	10	https://amazon.com/dp/ABCHDMSH	40	Stock

Observations & anomalies

1. In Product Price column , some values starts with \$ while others not
2. Laptop X is Out of stock , But Stock Status telling Its In stock

```
1 df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 6 columns):
#   Column          Non-Null Count  Dtype
---  -
0   Product ID      5 non-null     int64
1   Product Name    5 non-null     object
2   product_price   5 non-null     object
3   product_url     5 non-null     object
4   stock_count     5 non-null     int64
5   Stock status    5 non-null     object
dtypes: int64(2), object(4)
memory usage: 368.0+ bytes
```

3.Product price column is object type , ideally it should be int or float

```
1 df2.describe(include='all')
```

	Product ID	Product Name	product_price	product_url	stock_count
count	5.000000	5	5	5	5.000000
unique	NaN	5	5	3	NaN
top	NaN	Laptop A	500	https://amazon.com/dp/ABCHDMSH	NaN
freq	NaN	1	1	3	NaN
mean	3.000000	NaN	NaN	NaN	43.400000
std	1.581139	NaN	NaN	NaN	45.407048
min	1.000000	NaN	NaN	NaN	0.000000
25%	2.000000	NaN	NaN	NaN	24.000000
50%	3.000000	NaN	NaN	NaN	33.000000

4. max stock count is 120 Laptop A

### Take Your Written Test Answer Script & Explain your solution to Question #1 ?

Refer TurboLabs\_DataAnalyst(Quality)\_WritternTestQ&A\_SWAROOP.pdf for explanation

### What are Lists ?

Lists are one of 4 built-in data types in Python used to store collections of data, the other 3 are Tuple, Set, and Dictionary

Lists are used to store the elemtns / data items where each data item is separated by a comma (,)

A Python List can have data items of any data type,

One of the leading reasons why lists are being widely used is that Lists are mutable( editable )

Lists are created using square brackets:

Ex : l1 = [ 1, 'str1' , True , 3.4]

### Create a list contains numbers from 0 to 7 using List comprehension ?

```
1 l = [ x for x in range(0,8)]
2 l
```

```
[0, 1, 2, 3, 4, 5, 6, 7]
```

### Circular Rotate right the above list at nth element using function ?

```
1 def rotat_func(iterable,n_ele) :  
2     return (iterable[n_ele:]+iterable[:n_ele])
```

```
1 rotat_func(1,3)
```

```
[3, 4, 5, 6, 7, 0, 1, 2]
```

```
1 rotat_func(1,5)
```

```
[5, 6, 7, 0, 1, 2, 3, 4]
```

### Can you perform Same operation on a string ?

Yes sir , We can , example

```
1 rotat_func('turbolabs',4)
```

```
'olabsturb'
```

### What are limitations of Lists which led to NumPy ?

Lists supports heterogenous data types content (means strings , numbers , nested iterables can also be elements in list), So performing any mathematical operations involves type check , so consumes more time , And makes inefficient

Lists operations requires specific indexing and looping and it not supports Vectorisation, thus not convenient to perform such mathematical and statistical operations

These Limitaions are Overcame by Numpy which supports n dimensional array consumes less memory , supports vectorisation and very fast

### Which data types we left in this discussion ?

Sir , We left out Numeric Data types like int , float , complex

sequence data types like strings partially covered & Tuples we did not

And Sets and Dictionaries

### What are sets ?

Set is one of 4 built-in data types in Python used to store collections of data, the other 3 are List, Tuple, and Dictionary

Sets are used to store multiple items in a single variable. ,

A set is a collection which is unordered, unchangeable, and unindexed. And most importantly which doesn't allow duplicates

```
Ex a=set([1,2,2,3])
```

```
print(a)
```

Output

```
(1,2,3)
```

### **What are Dictionaries ?**

Dictionary is one of 4 built-in data types in Python used to store collections of data, the other 3 are List, Tuple, and sets

Dictionaries also known as an associative array.

A dictionary consists of a collection of key-value pairs.

Each key-value pair maps the key to its associated value

### **Other than pandas numpy , what packages and libraries you know and worked on ?**

Other than Pandas and Numpy , there are packages like Seaborn , matplotlib , SciPy

Recently as part of our Data Visualisation sessions , I'm currently working on Seaborn

### **Then , can you plot any graph using seaborn for a dataset ?**

Yes Sir , I will , Consider the dataset

```
1 import seaborn as sns
```

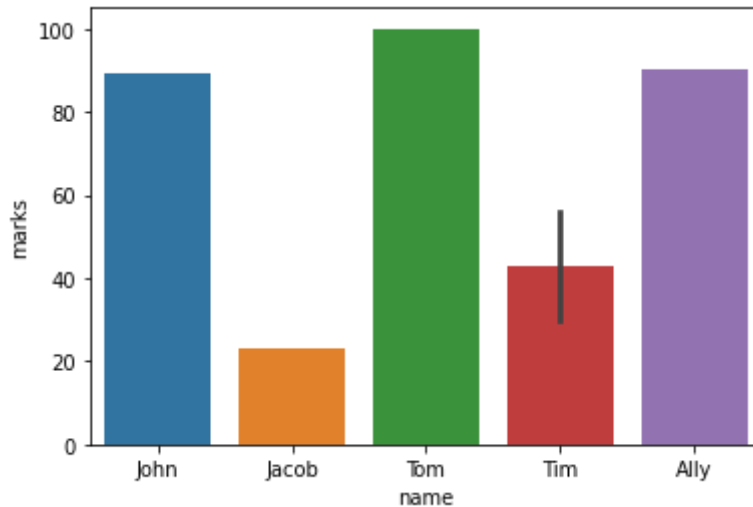
```
1 df
```



name marks subjects 

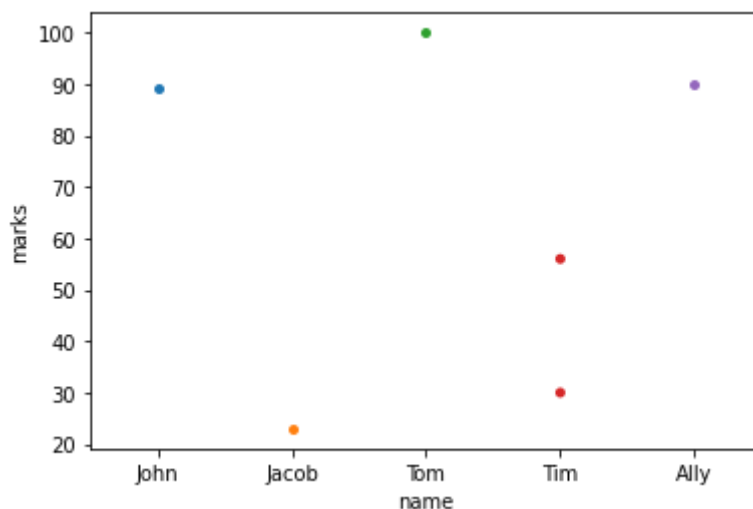
```
1 sns.barplot(x='name',y='marks',data=df)
```

<matplotlib.axes.\_subplots.AxesSubplot at 0x7f868529add0>



```
1 sns.swarmplot(x='name',y='marks',data=df)
```

<matplotlib.axes.\_subplots.AxesSubplot at 0x7f8685265850>



## Do You Know anything about Scraping ?

Web scraping is the process of collecting structured web data in an automated fashion.

It's also called web data extraction.

Some of the main use cases of web scraping include price monitoring, price intelligence, news monitoring, lead generation, and market research among many others

And Turbolabs works on this most

I also did a small project on Web Scraping of Populations Table from Wikipedia using Beautiful Soup and HTML attributes like table tag

**Okay Swaroop Your interview is over , Our HR will get back to you**

Okay sir

Thank You Joseph and Sumesh Sir , it was pleasure interacting with you sir

Will be waiting for response from HR

