DATA ANALYSIS INTERVIEW QUESTIONS

1. Python or R – Which one would you prefer for text analytics?

We will prefer Python because of the following reasons:

- ***Python*** would be the best option because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.
- ***R*** is more suitable for machine learning than just text analysis.
- Python performs faster for all types of text analytics.

2. How does data cleaning plays a vital role in the analysis?

Data cleaning can help in analysis because:

- Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.
- Data Cleaning helps to increase the accuracy of the model in machine learning.
- It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.
- It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

3. Differentiate between univariate, bivariate and multivariate analysis.

Univariate *analyses* are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can the analysis can be referred to as univariate analysis.

The ***bivariate*** *analysis* attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.

Multivariate analysis deals with the study of more than two variables to understand the effect of variables on the responses.

4. Explain Star Schema.

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

5. What is Cluster Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

6. What is Systematic Sampling?

Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, itis progressed from the top again. The best example of systematic sampling is equal probability method.

7. What are Eigenvectors and Eigenvalues?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching.

Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

8. Can you cite some examples where a false positive is important than a false negative?

Let us first understand what false positives and false negatives are.

- **False Positives** are the cases where you wrongly classified a non-event as an event a.k.a Type I error.
- **False Negatives** are the cases where you wrongly classify events as non-events, a.k.a Type II error.

Example 1: In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn't have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.

Example 2: Let's say an e-commerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not actually purchased anything but are marked as having made $10,000 worth of purchase.

9. Can you cite some examples where a false negative important than a false positive?

***Example 1*:** Assume there is an airport 'A' which has received high-security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to a shortage of staff, they decide to scan passengers being predicted as risk positives by their predictive model. What will happen if a true threat customer is being flagged as non-threat by airport model?

***Example 2*:** What if Jury or judge decides to make a criminal go free?
***Example 3*:** What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

10. Can you cite some examples where both false positive and false negatives are equally important?

In the **Banking** industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

11. Can you explain the difference between a Validation Set and a Test Set?

A ***Validation set*** can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting of the model being built.

On the other hand, a ***Test Set*** is used for testing or evaluating the performance of a trained machine learning model.

In simple terms, the differences can be summarized as; training set is to fit the parameters i.e. weights and test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.

12. Explain cross-validation.

Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will **generalize** to an **independent dataset**. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.

The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.

13. What are the various steps involved in an analytics project?

The following are the various ***steps involved in an analytics project***:

1. Understand the Business problem
2. Explore the data and become familiar with it.
3. Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.
4. After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step until the best possible outcome is achieved.
5. Validate the model using a new data set.
6. Start implementing the model and track the result to analyze the performance of the model over the period of time.

14. During analysis, how do you treat missing values?

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights.

If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.

If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.

If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

15.What are the feature selection methods used to select the right variables?

There are two main methods for feature selection:

Filter Methods

This involves:

- Linear discrimination analysis
- ANOVA
- Chi-Square

The best analogy for selecting features is "bad data in, bad answer out." When we're limiting or selecting the features, it's all about cleaning up the data coming in.

Wrapper Methods

This involves:

- Forward Selection: We test one feature at a time and keep adding them until we get a good fit
- Backward Selection: We test all the features and start removing them to see what works better
- Recursive Feature Elimination: Recursively looks through all the different features and how they pair together

Wrapper methods are very labor-intensive, and high-end computers are needed if a lot of data analysis is performed with the wrapper method.

16.What are dimensionality reduction and its benefits?

Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.

This reduction helps in compressing data and reducing storage space. It also reduces computation time as fewer dimensions lead to less computing. It removes redundant features; for example, there's no point in storing a value in two different units (meters and inches).

17. What is root cause analysis?

Root cause analysis was initially developed to analyze industrial accidents but is now widely used in other areas. It is a problem-solving technique used for isolating the root causes of faults or problems. A factor is called a root cause if its deduction from the problem-fault-sequence averts the final undesirable event from recurring

18. What are the important skills to have in Python with regard to data analysis?

The following are some of the important skills to possess which will come handy when performing data analysis using Python.

- Good understanding of the built-in data types especially lists, dictionaries, tuples, and sets.
- Mastery of N-dimensional NumPy Arrays.
- Mastery of Pandas dataframes.
- Ability to perform element-wise vector and matrix operations on NumPy arrays.
- Knowing that you should use the Anaconda distribution and the conda package manager.
- Familiarity with Scikit-learn. ****Scikit-Learn Cheat Sheet****
- Ability to write efficient list comprehensions instead of traditional for loops.
- Ability to write small, clean functions (important for any developer), preferably pure functions that don't alter objects.
- Knowing how to profile the performance of a Python script and how to optimize bottlenecks

.

19.What is Z-Score ?

Measuring deviation of mean in terms of standard deviation (Z)

gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is
Z= x-Mu/sigma

20. What is Inter Quartile Range (IQR) ?

In descriptive statistics, the interquartile range (IQR), also called the midspread, middle 50%, or H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles,[1][2] IQR = $Q_3 - Q_1$.

21. What are Outliers and Extreme Values ?

An outlier is a data point that differs significantly from other observations.
These characteristic ***values*** are the smallest (minimum ***value***) or largest (maximum ***value***), and are known as ***extreme values***.

22. Rejection Region and acceptance region ?

Acceptance Region/Rejection Region. ... The range containing values that are consistent with the null hypothesis is the "acceptance region"; the other range, in which the null hypothesis is rejected, is the rejection region (or critical region).