**Data Science Interview Questions.**

1.Differnce between *Datamining* and *Data Profiling*?

A. *Datamining:*

Data mining is the process of finding relevant information which has not been found before.

It is the way in which raw data is turned is turned into valuable information.

b. *Data Profiling:*

Data Profiling is usually done to assess a dataset for its uniqueness consistency and logic.

It cannot identity incorrect or inaccurate data values.

*Summary:*

Data mining is the process of finding relevant information or data. It’s mostly raw data.

Data profiling is whatever the data we found, we use the data to assess its uniqueness, consistency and logic.

2. Define the term *data wrangling* in data analytics?

A. *Data wrangling* is the process of cleaning, Structuring and enriching the raw data into a desired useable format for better decision making.

Summary:

Data wrangling is the like tidying up a messy room. You are organizing cleaning and rearranging the ate so that it’s easier to work and analyse. It involves takes like cleaning errors, filling in missing information and reshaping the data to fit the needs of your analysis.

Think of it as preparing the ingredients before cooking a meal- you want everything ready and in the right form you start.

A diagram of a diagram

Description automatically generated with medium confidence

3. What are the common problems that data analyst encounter during analysis?

* Handling duplicates and missing values
* Collecting the meaningful right data and at the right time
* Making data secure and dealing with compliance issues
* Handling data purging and storage problems

4. What are the various steps involved in any analytics project?

1. Understanding the problem
2. Data collection
3. Data cleaning’\
4. Data exploration and analysis (EDA)
5. Interpret the results.

5. which technical tools does a data analyst use for analysis and presentation purposes?

* Sql Server
* My SQL
* Excel
* IBM SPSS
* Tableau
* Python
* PowerPoint Presentation

6. What are the best practices for data cleaning?

1. Make a data cleaning plan by understanding where the common errors take place and keep communications open.
2. Identify and remove duplicates before working with the data. This will lead to an effective data analysis process.
3. Focus on the accuracy of the data. Maintain the value types of the data, provide mandatory constrains and set cross field validation.
4. Standardise the data at the point of entry so that it is less chaotic, and you will be able to ensure that all information is standardised leading to fewer errors on entry.

7. How can you handle missing values in dataset?

1. Listwise deletion:

In this method, an Entire record is excluded from analysis if any single value is missing.

1. Average imputation:

Use the average values of the responses from the other participants to fill the missing values.

1. Regression substitution:

You can use the multiple regression analysis to estimate a missing value

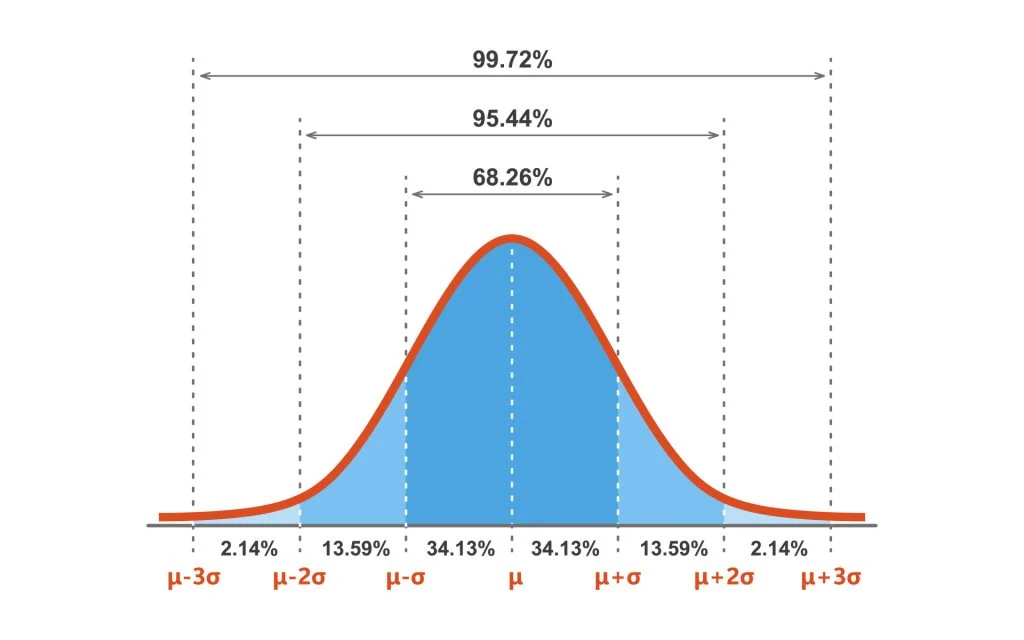
1. Multiple imputation:

Multiple imputation values based on the correlation for the missing data and then averages the simulated data sets by incorporating random errors in your prediction.

8. What do you understand by the term Normal distribution?

A. normal distribution is a type of continues probability distribution that is symmetric about the mean and in a graph.

A normal distribution will appear as a bell curve:



* The Mean, Median and Mode are Equal
* All of them are located at the center of distribution
* 68% of the data lies within 1st standard deviation of the mean
* 95% of the data falls within 2nd standard deviation of the mean
* 99.7% of the data lies within 3 standard deviations of the mean

9. What is the time series analysis?

Time series analysis is a statistical Method that deals with ordered sequence of values of a variable at Equally spaced time intervals.

10. How is data joining different from data blending in tableau?

|  |  |
| --- | --- |
| Data Joining | Data blending |
| 1. Data joining is done when the data comes from the same sources.  Eg: Combining two tables from same databases or 2 or more different worksheets from same excel file.  2. All the combined tables or sheets contain common sets of dimensions and measures | 1. Data blending is used when the data is from 2 or more sources.  Eg: combining the oracle table with SQL server or two sheets from excel or combining excel sheets and oracle table.  2. In data blending source contains its own set of dimensions and measures |

11. How is overfitting different from underfitting?

|  |  |
| --- | --- |
| Overfitting | Underfitting |
| 1. Model trains the data too well using training set. | 1. The model neither trains the data well nor can generalize to new data |
| 2. Overfitting performs significantly over the test set | 2. Underfitting performs poorly both on train and the test set. |
| 3. Happens when the model learns the noise and random fluctuations in the training dataset in details. | 3. Happens when there is less data to build accurate models and also when we try to build a linear model with a nonlinear data. |



12. In MS Excel, a numeric value can be treated as a text value it precedes with \_\_\_\_\_\_\_\_\_?

A. Apostrophe (‘)

13. What is the difference between Count, Count A, Count Blank and Count IF in Excel?

A. *Count function* returns the count of numeric cells in a range.

*Count A function* Returns the count of non-blank cells in a range

*Count Blank function* returns the count blank cells in a range.

*Count IF function* returns the count of values by checking a given condition.