Top 30 Data Analyst Interview Questions & Answers

1) Mention what is the responsibility of a Data analyst?

Responsibility of a Data analyst include,

- Provide support to all data analysis and coordinate with customers and staffs
- Resolve business associated issues for clients and performing audit on data
- Analyze results and interpret data using statistical techniques and provide ongoing reports
- Prioritize business needs and work closely with management and information needs
- Identify new process or areas for improvement opportunities
- Analyze, identify and interpret trends or patterns in complex data sets
- Acquire data from primary or secondary data sources and maintain databases/data systems
- Filter and “clean” data, and review computer reports
- Determine performance indicators to locate and correct code problems
- Securing database by developing access system by determining user level of access

2) What is required to become a data analyst?

To become a data analyst,

- Robust knowledge on reporting packages (Business Objects), programming language (XML, Javascript, or ETL frameworks), databases (SQL, SQLite, etc.)
- Strong skills with the ability to analyze, organize, collect and disseminate big data with accuracy
- Technical knowledge in database design, data models, data mining and segmentation techniques
- Strong knowledge on statistical packages for analyzing large datasets (SAS, **Excel**, SPSS, etc.)

3) Mention what are the various steps in an analytics project?

Various steps in an analytics project include

- Problem definition
- Data exploration
- Data preparation
- Modelling
- Validation of data
- Implementation and tracking

4) Mention what is data cleansing?

Data cleaning also referred as data cleansing, deals with identifying and removing errors and inconsistencies from data in order to enhance the quality of data.

5) List out some of the best practices for data cleaning?

Some of the best practices for data cleaning includes,

- Sort data by different attributes
- For large datasets cleanse it stepwise and improve the data with each step until you achieve a good data quality
- For large datasets, break them into small data. Working with less data will increase your iteration speed
- To handle common cleansing task create a set of utility functions/tools/scripts. It might include, remapping values based on a CSV file or SQL database or, regex search-and-replace, blanking out all values that don't match a regex
- If you have an issue with data cleanliness, arrange them by estimated frequency and attack the most common problems
- Analyze the summary statistics for each column (standard deviation, mean, number of missing values,)
- Keep track of every date cleaning operation, so you can alter changes or remove operations if required

6) Explain what is logistic regression?

Logistic regression is a statistical method for examining a dataset in which there are one or more independent variables that defines an outcome.

7) List of some best tools that can be useful for data-analysis?

- Tableau
- RapidMiner
- OpenRefine
- KNIME
- Google Search Operators
- Solver
- NodeXL
- io
- Wolfram Alpha's
- Google Fusion tables

8) Mention what is the difference between data mining and data profiling?

The difference between data mining and data profiling is that

Data profiling: It targets on the instance analysis of individual attributes. It gives information on various attributes like value range, discrete value and their frequency, occurrence of null values, data type, length, etc.

Data mining: It focuses on cluster analysis, detection of unusual records, dependencies, sequence discovery, relation holding between several attributes, etc.

9) List out some common problems faced by data analyst?

Some of the common problems faced by data analyst are

- Common misspelling
- Duplicate entries
- Missing values
- Illegal values
- Varying value representations
- Identifying overlapping data

10) Mention the name of the framework developed by Apache for processing large data set for an application in a distributed computing environment?

Hadoop and MapReduce is the programming framework developed by Apache for processing large data set for an application in a distributed computing environment.

11) Mention what are the missing patterns that are generally observed?

The missing patterns that are generally observed are

- Missing completely at random
- Missing at random
- Missing that depends on the missing value itself
- Missing that depends on unobserved input variable

12) Explain what is KNN imputation method?

In KNN imputation, the missing attribute values are imputed by using the attributes value that are most similar to the attribute whose values are missing. By using a distance function, the similarity of two attributes is determined.

13) Mention what are the data validation methods used by data analyst?

Usually, methods used by data analyst for data validation are

- Data screening
- Data verification

14) Explain what should be done with suspected or missing data?

- Prepare a validation report that gives information of all suspected data. It should give information like validation criteria that it failed and the date and time of occurrence
- Experience personnel should examine the suspicious data to determine their acceptability
- Invalid data should be assigned and replaced with a validation code

- To work on missing data use the best analysis strategy like deletion method, single imputation methods, model based methods, etc.

15) Mention how to deal the multi-source problems?

To deal the multi-source problems,

- Restructuring of schemas to accomplish a schema integration
- Identify similar records and merge them into single record containing all relevant attributes without redundancy

16) Explain what is an Outlier?

The outlier is a commonly used terms by analysts referred for a value that appears far away and diverges from an overall pattern in a sample. There are two types of Outliers

- Univariate
- Multivariate

17) Explain what is Hierarchical Clustering Algorithm?

Hierarchical clustering algorithm combines and divides existing groups, creating a hierarchical structure that showcase the order in which groups are divided or merged.

18) Explain what is K-mean Algorithm?

K mean is a famous partitioning method. Objects are classified as belonging to one of K groups, k chosen a priori.

In K-mean algorithm,

- The clusters are spherical: the data points in a cluster are centered around that cluster
- The variance/spread of the clusters is similar: Each data point belongs to the closest cluster

19) Mention what are the key skills required for Data Analyst?

A data scientist must have the following skills

- **Database knowledge**

- Database management
- Data blending
- Querying
- Data manipulation

- **Predictive Analytics**

- Basic descriptive statistics
- Predictive modeling
- Advanced analytics

- **Big Data Knowledge**

- Big data analytics
- Unstructured data analysis
- Machine learning

- **Presentation skill**

- Data visualization
- Insight presentation
- Report design

20) Explain what is collaborative filtering?

Collaborative filtering is a simple algorithm to create a recommendation system based on user behavioral data. The most important components of collaborative filtering are **users- items- interest**.

A good example of collaborative filtering is when you see a statement like “recommended for you” on online shopping sites that’s pops out based on your browsing history.

21) Explain what are the tools used in Big Data?

Tools used in Big Data includes

- Hadoop
- Hive
- Pig
- Flume
- Mahout
- Sqoop

22) Explain what is KPI, design of experiments and 80/20 rule?

KPI: It stands for Key Performance Indicator, it is a metric that consists of any combination of spreadsheets, reports or charts about business process

Design of experiments: It is the initial process used to split your data, sample and set up of a data for statistical analysis

80/20 rules: It means that 80 percent of your income comes from 20 percent of your clients

23) Explain what is Map Reduce?

Map-reduce is a framework to process large data sets, splitting them into subsets, processing each subset on a different server and then blending results obtained on each.

24) Explain what is Clustering? **What are the properties for clustering algorithms?**

Clustering is a classification method that is applied to data. Clustering algorithm divides a data set into natural groups or clusters.

Properties for clustering algorithm are

- Hierarchical or flat
- Iterative
- Hard and soft
- Disjunctive

25) What are some of the statistical methods that are useful for data-analyst?

Statistical methods that are useful for data scientist are

- Bayesian method
- Markov process
- Spatial and cluster processes
- Rank statistics, percentile, outliers detection
- Imputation techniques, etc.
- Simplex algorithm
- Mathematical optimization

26) What is time series analysis?

Time series analysis can be done in two domains, frequency domain and the time domain. In Time series analysis the output of a particular process can be forecast by analyzing the previous data by the help of various methods like exponential smoothening, log-linear regression method, etc.

27) Explain what is correlogram analysis?

A correlogram analysis is the common form of spatial analysis in geography. It consists of a series of estimated autocorrelation coefficients calculated for a different spatial relationship. It can be used to construct a correlogram for distance-based data, when the raw data is expressed as distance rather than values at individual points.

28) What is a hash table?

In computing, a hash table is a map of keys to values. It is a data structure used to implement an associative array. It uses a hash function to compute an index into an array of slots, from which desired value can be fetched.

29) What are hash table collisions? How is it avoided?

A hash table collision happens when two different keys hash to the same value. Two data cannot be stored in the same slot in array.

To avoid hash table collision there are many techniques, here we list out two

- **Separate Chaining**:

It uses the data structure to store multiple items that hash to the same slot.

- **Open addressing**:

It searches for other slots using a second function and store item in first empty slot that is found

29) Explain what is imputation? List out different types of imputation techniques?

During imputation we replace missing data with substituted values. The types of imputation techniques involve are

- **Single Imputation**

- Hot-deck imputation: A missing value is imputed from a randomly selected similar record by the help of punch card
- Cold deck imputation: It works same as hot deck imputation, but it is more advanced and selects donors from another datasets
- Mean imputation: It involves replacing missing value with the mean of that variable for all other cases
- Regression imputation: It involves replacing missing value with the predicted values of a variable based on other variables
- Stochastic regression: It is same as regression imputation, but it adds the average regression variance to regression imputation

- **Multiple Imputation**

- Unlike single imputation, multiple imputation estimates the values multiple times

30) Which imputation method is more favorable?

Although single imputation is widely used, it does not reflect the uncertainty created by missing data at random. So, multiple imputation is more favorable then single imputation in case of data missing at random.

31) Explain what is n-gram?

N-gram:

An n-gram is a contiguous sequence of n items from a given sequence of text or speech. It is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n-1).

32) Explain what is the criteria for a good data model?

Criteria for a good data model includes

- It can be easily consumed
- Large data changes in a good model should be scalable
- It should provide predictable performance
- A good model can adapt to changes in requirements