1. **What does “Data Cleansing” mean? What are the best ways to practice this?**

Data cleaning means fixing bad data in your data set.

Bad data could be:

* Empty cells
* Data in wrong format
* Wrong data
* Duplicates

Missing Value handling drop():

Duplicates: dropduplicates()

Data in wrong format: pd.to\_datetime(df['timestamp'], format='%Y-%m-%d'); tostr()

Reference:  
  
<https://towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d>

1. **What is the difference between data profiling and data mining?**

**Data minin**g: refers to the process of identifying patterns in a pre-built database. It carries out analysis or knowledge discovery in the databases to evaluate the existing database and large datasets to turn raw data into useful information and find trends and patterns into it.

**Data Profiling** is to collect statistics or informative summaries about the data. Also called data archaeology, data profiling is used to derive information about the data itself and assess the quality of the data.

1. **Define Outlier with an example.**

an outlier is a data point that differs significantly from other observations.[1][2] An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.[3] An outlier can cause serious problems in statistical analyses.

We can use Histogram and scatter plot to identify the plot not follow the pattern

1. **What is “Collaborative Filtering”?**

**Collaborative filtering** is a technique that can filter out items that a user might like on the basis of reactions by similar users.

1. **What is “Time Series Analysis”?**

Time Series Analysis is the way of studying the characteristics of the response variable with respect to time, as the independent variable. To estimate the target variable in the name of predicting or forecasting, use the time variable as the point of reference. In this article we will discuss in detail TSA Objectives, Assumptions, Components (stationary, and Non- stationary). Along with the TSA algorithm and specific use cases in Python.

1. **Explain the core steps of a Data Analysis project?**

Step 1: frame the problem

Step 2: collect raw data

Step 3” collect the data for analysis

Step explore the data

Step 5 in depth analysis

Stemp 6 communicate result

1. **What are the characteristics of a good data model?**

The writer goes on to define the four criteria of a good data model: “ (1) Data in a good model can be easily consumed. (2) Large data changes in a good model are scalable. (3) A good model provides predictable performance.

1. **Explain and provide examples of univariate, bivariate, and multivariate analysis?**

**Univariate analysis**: We only have one variable in data set. We normally analysis the descriptive data such as mean medium distribution of the data

**Example**: we have a data set with a class of students’ height I would like know average /  Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation. the height of class.

**Bivariate analysis:** the data has two variables**.**

**Example**: we have a data set with a class of students’ height also the student age I would like know the correlation with student age and their height

**Multivariate analysis** The statistical study of data where multiple measurements are made on each experimental unit and where the relationships among multivariate measurements and their structure are important.

MANOVA

1. **What is a Linear Regression?**

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

1. **In terms of modelling data, what do we mean by Over-fitting and Under-fitting?**

**Overfitting** is an error that occurs in data modeling as a result of a particular function aligning too closely to a minimal set of data points.

 the model trains for too long on sample data or when the model is too complex, it can start to learn the “noise,” or irrelevant information, within the dataset. When the model memorizes the noise and fits too closely to the training set, the model becomes “overfitted,” and it is unable to generalize well to new data.

**Underfitting** is a scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen data