1. Data gathering
2. Data preprocessing
3. Exploratory data analysis (EDA)
4. Feature engineering
5. Training machine learning models of the following kinds:
   * Regression
   * Classification
   * Clustering
6. Multivariate querying
7. Density estimation
8. Dimensionality reduction
9. Model / Algorithm selection
10. Testing and matching
11. Model monitoring
12. Model retraining

The above mentioned are the steps in machine learning.

**Data Preprocessing**: Before starting training the models, it is of utmost importance to prepare data appropriately. As part of data preprocessing, some of the following is done:

* + **Data cleaning**: Data cleaning requires one to identify attributes having not enough data or attributes which are not having variance. These data (rows and columns) need to be removed from the training data set.
  + **Missing data imputation**: Handling missing data using data imputation techniques such as replacing missing data with mean, median, or mode. Here is my post on this topic: [Replace missing values with mean, median or mode](https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/)

Answer 2 :

Quantitative data refers to any information that can be quantified — that is, numbers. If it can be counted or measured, and given a numerical value, it's quantitative in nature. Think of it as a measuring stick.

Quantitative variables can tell you "how many," "how much," or "how often."

Unlike quantitative data, qualitative data is descriptive, expressed in terms of language rather than numerical values.

Qualitative data analysis describes information and cannot be measured or counted. It refers to the words or labels used to describe certain characteristics or traits.

You would turn to qualitative data to answer the "why?" or "how?" questions. It is often used to investigate open-ended studies, allowing participants (or customers) to show their true feelings and actions without guidance.

Answer 3:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Airline** | **Date\_of\_Journey** | **Source** | **Destination** | **Route** | **Dep\_Time** | **Arrival\_Time** | **Duration** | **Total\_Stops** | **Additional\_Info** | **Price** |  |
| IndiGo | 24/03/2019 | Banglore | New Delhi | BLR → DEL | 22:20 | 01:10 22 Mar | 2h 50m | non-stop | No info | 3897 |  |
| Air India | 1/05/2019 | Kolkata | Banglore | CCU → IXR → BBI → BLR | 05:50 | 13:15 | 7h 25m | 2 stops | No info | 7662 |  |
| Jet Airways | 9/06/2019 | Delhi | Cochin | DEL → LKO → BOM → COK | 09:25 | 04:25 10 Jun | 19h | 2 stops | No info | 13882 |  |
| IndiGo | 12/05/2019 | Kolkata | Banglore | CCU → NAG → BLR | 18:05 | 23:30 | 5h 25m | 1 stop | No info | 6218 |  |
| IndiGo | 01/03/2019 | Banglore | New Delhi | BLR → NAG → DEL | 16:50 | 21:35 | 4h 45m | 1 stop | No info | 13302 |  |

Answer 4 :

* Inadequate Training Data. ...
* Poor quality of data. ...
* Non-representative training data. ...
* Overfitting and Underfitting. ...
* Monitoring and maintenance. ...
* Getting bad recommendations. ...
* Lack of skilled resources. ...
* Customer Segmentation.

The ramifications can be a poorly trained model with bad predictions.

Answer 5 :

We can explore categorical variables with

1.Mode

2. Groupby

3. Pie charts

Answer 6 :

It is important to handle the missing values appropriately.

* Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like K-nearest and Naive Bayes support data with missing values.
* You may end up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly.
* Missing data can lead to a lack of precision in the statistical analysis.

Answer 7:

There are 2 primary ways of handling missing values:

1. Deleting the Missing values
2. Imputing the Missing Values

**Deleting the entire row**

If a row has many missing values then you can choose to drop the entire row.

**Deleting the entire column**

If a certain column has many missing values then you can choose to drop the entire column.

Imputing:

**Replacing With Mode / Median or any arbitrary value**

Mode is the most frequently occurring value. It is used in the case of categorical features. We can also use median or any other arbitrary value. But mostly it depends on domain knowledge.

Answer 8 :

To ensure high-quality data, it's crucial to preprocess it. To make the process easier, data preprocessing is divided into four stages: **data cleaning, data integration, data reduction, and data transformation**.

Need of dimensionality reduction:

* Space required to store the data is reduced as the number of dimensions comes down
* Less dimensions lead to less computation/training time
* Some algorithms do not perform well when we have a large dimensions. So reducing these dimensions needs to happen for the algorithm to be useful
* It takes care of multicollinearity by removing redundant features. For example, you have two variables – ‘time spent on treadmill in minutes’ and ‘calories burnt’. These variables are highly correlated as the more time you spend running on a treadmill, the more calories you will burn. Hence, there is no point in storing both as just one of them does what you require

Techniques for Dimensionality reduction:

* Principal Component Analysis (PCA)
* Non-negative matrix factorization (NMF)
* Linear discriminant analysis (LDA)

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [statistics](https://en.wikipedia.org/wiki/Statistics), **feature selection**, also known as **variable selection**, **attribute selection** or **variable subset selection**, is the process of selecting a subset of relevant [features](https://en.wikipedia.org/wiki/Feature_(machine_learning)) (variables, predictors) for use in model construction.