|  |  |  |  |
| --- | --- | --- | --- |
| **UNIT \_ 1 Data Science – A Definition** Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze,  visualize, interact with data to create data products. **Ben Fry’s Model: Visualizing Data Process**  1. Acquire  2. Parse (Analyze and put in proper format)  3. Filter  4. Mine (Discovering patterns or knowledge from datasets)  5. Represent  6. Refine  7. Interact **Jeff Hammerbacher’s Model**  1. Identify problem  2. Instrument data sources  3. Collect data  4. Prepare data (integrate, transform, clean, filter, aggregate)  5. Build model  6. Evaluate model  7. Communicate results  **7V’s of Big Data**  • Raw Data: Volume  • Change over time: Velocity  • Data types: Variety  • Data Quality: Veracity  • Information for Decision Making: Value  • Change in Data: Variability  • Presentation of Data: Visualization | **Applications of Data Scienece**   |  |  | | --- | --- | | **E Commerce**  Identifying Consumers  Recommending Products  Analyzing Reviews  **Manufacturing**  Predicting Potential problems  Monitoring Systems  Automating Manufacturing Units  Maintenance Scheduling  Anomaly Detection  **Banking**  Fraud Detection  Credit Risk  Customer Lifetime Value | **Healthcare**  Medical Analysis  frug Discovery  Bioinformatics  Virtual Assistants  **Transport**  Self-Driving Cars  Enhanced Driving Experience  Car Monitoring System  Enhancing the safety of passengers  **Finance**  Customer Segmentation  Strategic Decision Making  Algorithmic Trading  Risk Analytics |   **Data Science Life Cycle (Draw a Circle)**   * Data Collection * Data Preparation * Exploratory Data Analysis * Modelling * Model Evaluation * Model Deployment |
| |  |  |  | | --- | --- | --- | | **Area** | **BI Analyst** | **Data Scientist** | | Focus | Reports, KPIs, trends | Patterns, Correlations, models | | Process | Static, comparative | Exploratory, experimentation, visuals | | Data sources | Pre-planned, added slowly | On the fly, as needed | | Transform | Upfront, carefully planned | In-database, on-demand, enrichment | | Data quality | Single version of truth | “Good enough”,  probabilities | | Data model | Schema on load | Schema on query | | Analytics | Retrospective, Descriptive | Predictive, Prescriptive, Preventative |  |  |  | | --- | --- | | **Data Analyst Skills** | **Data Scientist Skills** | | Data Mining | Data Mining | | Data Warehousing | Data Warehousing | | Math, Statistics | Math, Statistics, Computer Science | | Tableau and Data Visualization | Tableau and Data Visualization/ Storytelling | | SQL | Python, R, Java, Scala, SQL, MATLAB, Pig | | Business Intelligence | Economics | | SAS | Big Data/Hadoop | | Advanced Excel Skills | Machine Learning | | **NumPy/Python**  NumPy is a Python library essential for working with arrays, providing a powerful array object called ND array. It is designed for high efficiency, with arrays stored in contiguous memory locations, allowing fast processing and manipulation compared to traditional Python lists. Besides array operations, NumPy includes functions for linear algebra, Fourier transforms, and matrix operations. Standing for Numerical Python, NumPy aims to offer an array object that is up to 50 times faster than standard Python lists, thanks to its optimized storage and extensive support functions that simplify array manipulations.  **Pandas**  Pandas, short for 'Python Data Analysis Library', is an open-source Python library widely used by data scientists for data exploration, manipulation, and visualization. It is renowned for its ease of use, speed, and powerful capabilities, making it comparable to Microsoft Excel in the Python ecosystem. Pandas integrates seamlessly with other visualization libraries and supports various data formats, including CSV, Excel, SQL databases, and even web pages. Its user-friendly interface and robust functionality make it an indispensable tool for efficient data analysis and processing. |

|  |  |
| --- | --- |
| **Data Preprocessing:** is a technique that is used to convert the raw data into a clean data set. Data is gathered from different sources it is collected in raw format which is not feasible for the analysis.  **Tasks of Data Preprocessing**  **Data Cleaning:** This is the first step which is implemented in Data Preprocessing. In this step, the primary focus is on handling missing data, noisy data, detection, and removal of outliers, minimizing duplication and computed biases within the data.  **Data Integration:** This process is used when data is gathered from various data sources and data are combined to form consistent data. This consistent data after performing data cleaning is used for analysis.  **Data Transformation:** This step is used to convert the raw data into a specified format according to the model.  **Normalization –** In this method, numerical data is converted into the specified range, i.e., between 0 and 1 so that scaling of data can be performed.  **Aggregation –** This method is used to combine the features into one. For example, combining two categories can be used to form a new group.  **Generalization –** In this case, lower-level attributes are converted to a higher standard (e.g. age 20, 40 – may be taken as Young, Old, etc.)  **Data Reduction:** After the transformation and scaling of data duplication, i.e., redundancy within the data is removed and efficiently organize the data. | ***How we can deal with the Missing data***  **Ignoring the missing record:** It is the simplest and efficient method for handling the missing data. But this method should not be performed at the time when the number of missing values are immense or when the pattern of data is related to the unrecognized primary root of the cause  **Filling the missing values manually:** This is one of the best-chosen methods. But there is one limitation that when there are large data set, and missing values are significant then, this approach is not efficient as it becomes a time-consuming task.  **Filling using computed values:** The missing values can also be occupied by computing mean, mode or median of the observed given values. Another method could be the predictive values that are computed by using any Machine Learning or Deep Learning algorithm.  ***How we can deal with the Noisy data***  **Data Binning:** In this approach sorting of data is performed concerning the values of the neighborhood. This method is also known as local smoothing.  **Clustering:** In the approach, the outliers may be detected by grouping the similar data in the same group, i.e., in the same cluster.  **Machine Learning:** A Machine Learning algorithm can be executed for smoothing of data. For example, Regression Algorithm can be used for smoothing of data using a specified linear function.  **Removing manually:** The noisy data can be deleted manually by the human being, but it is a time-consuming process, so mostly this method is not given priority. |
| **Data Wrangling**: is used in step during EDA and modeling to adjust data sets interactively while analyzing data and building a model. Data wrangling, also known as data munging, is the process of cleaning, restructuring, and transforming raw data from various sources into a more organized and usable format for analysis. This involves tasks like handling missing data, correcting inconsistencies, merging datasets, creating new variables, and ensuring data quality, ultimately preparing the data for further exploration and insights.  **Tasks of Data Wrangling • Discovering:** Firstly, data should be understood thoroughly and examine which approach will best suit. For example: if have a weather data when we analyze the data it is observed that data is from one area and so primary focus is on determining patterns.  **• Structuring:** As the data is gathered from different sources, the data will be present in various shapes and sizes. Therefore, there is a need for structuring the data in proper format.  **• Cleaning:** Cleaning or removing of data should be performed that can degrade the performance of analysis.  **• Enrichment:** Extract new features or data from the given  Validating: This approach is used for improving the quality of data and consistency rules so that transformations that are applied to the data could be verified.  **• Publishing:** After completing the steps of Data Wrangling, the steps can be documented so that similar steps can be performed for the same kind of data to save time. | **Understanding Data Attribute Types**  **Qualitative Attributes**   1. **Nominal Attributes:** Values are names or symbols without numeric value, representing categories or states with no inherent order or rank among them. 2. **Binary Attributes:** Binary data has two states (e.g., yes/no, true/false). Symmetric binary attributes have equally important values with no preference in coding, while asymmetric attributes have one value more important than the other, coded as 1. 3. **Ordinal Attributes:** Values have a meaningful sequence or ranking but the magnitude of differences between them is not known.   **Quantitative Attributes**   1. **Interval-scaled Attributes**: Concerned with order and differences between values, allowing measurement of standard deviation and central tendency. Values can be added and subtracted but not multiplied or divided. They have no true zero, so ratios cannot be calculated. Example: temperature in Celsius. 2. **Ratio-scaled Attributes**: Have all properties of interval-scaled attributes, with the addition of a true zero indicating absence of the quantity. Values can be added, subtracted, multiplied, and divided. Example: weight in kilograms. |
| **Major Tasks in Data Preprocessing**  **• Data Cleaning:** Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies  **• Data Integration:** Integration of multiple databases, or files  **• Data Transformation:** Normalization and aggregation  **• Data Reduction:** Obtains reduced representation in volume but produces the same or similar analytical results  **• Data Discretization** (for numerical data)  **Data Cleaning:** also known as **data cleansing**or **data preprocessing**, is a crucial step in the data science pipeline that involves identifying and correcting or removing errors, inconsistencies, and inaccuracies in the data to improve its quality and usability. Data cleaning is essential because raw data is often noisy, incomplete, and inconsistent, which can negatively impact the accuracy and reliability of the insights derived from it.  **Python Collections to Store the Data**  **List:** ordered and changeable. Allows duplicate members.  e.g. [1, 2, 3, 4, 5]  **Tuple:** ordered and unchangeable Allows duplicate members. e.g. (1, 2, 3, 4, 5)  **Set:** unchangeable and unindexed. No duplicate members.  e.g. {1, 2, 3, 4, 5}  **Dictionary:** ordered and changeable. No duplicate members. e.g. {1: “a”, 2: “b”, 3: “c”, 4: “d”, 5: “e”} | **Data integration** is the process of combining data from different sources to provide a unified view, making it accessible and useful for analysis and decision-making. This involves merging datasets from various databases, applications, and systems, ensuring consistency and coherence across the integrated data. Data integration techniques include ETL , data warehousing, and data virtualization, among others. Effective data integration addresses issues such as data redundancy, discrepancies, and fragmentation, enabling organizations to harness comprehensive and accurate insights from their data.  **Redundancy Analysis**  Redundancy analysis involves identifying and addressing redundant data that often arise when integrating multiple databases. Redundant data can occur when the same attribute or object has different names across databases, a situation known as object identification. Additionally, redundancy can appear when one attribute in a dataset is a derived attribute in another table, such as when annual revenue is calculated from monthly revenues.  **Correlation Analysis**  Correlation analysis is a statistical method used to measure and describe the association between random variables, helping to predict one quantity based on another. While correlation can suggest the presence of a causal relationship, it does not confirm causality. It serves as a foundational tool in many modeling techniques, providing a basic measure of association strength. Various methods exist to calculate the correlation coefficient, each capturing different types of relationships between variables. |
| **Data Transformation**  Data transformation is the process of converting data from one format or structure into another, facilitating its compatibility and usability across different systems and applications. This process often involves data cleaning, normalization, aggregation, and enrichment to ensure the data is accurate, consistent, and in a suitable format for analysis or integration. Transformation can include converting data types, standardizing values, and creating derived attributes. Effective data transformation enhances data quality and interoperability, enabling more efficient data analysis, improving insights, and supporting better decision-making. It is a critical step in data processing workflows, such as ETL (Extract, Transform, Load), which prepares data for storage in data warehouses or for use in advanced analytics and machine learning models.  **1) Normalization by Decimal Scaling**  Decimal scaling normalization is a data transformation method where the decimal point of attribute values is shifted based on the maximum value in the dataset. This shift ensures that all values are within a certain range, making comparisons and analyses more manageable. For example, if ages range from 25 to 70, dividing all ages by 10 yields values like 7.0 and 3.6, simplifying comparisons and analysis. This would result in age values like 7.0, 3.6, and 2.5, making them easier to work with and interpret, especially when combined with other normalized attributes or datasets. | **2) Data Transformation by min-max Normalization**  Min-max normalization is a technique in data transformation that scales numeric data to a specific range, typically between 0 and 1. For instance, if we have a dataset of exam scores ranging from 0 to 100, min-max normalization would convert these scores into values between 0 and 1 based on their relation to the minimum and maximum scores in the dataset. For example, if a student's score was 80 out of 100, after min-max normalization, it might become 0.8. This method is useful for standardizing data across different scales, ensuring that smaller and larger values contribute equally to analyses like machine learning models without losing the relationships between the original data points.  **3) Data Transformation by z-score Normalization**  Z-score normalization, also known as zero-mean normalization, is a technique in data transformation that standardizes numerical data based on the mean and standard deviation of the dataset. This method is particularly useful when dealing with data where the actual minimum and maximum values are unknown, or when outliers may significantly affect min-max normalization. For example, if we have a dataset of exam scores with a mean of 70 and a standard deviation of 10, a score of 80 would have a z-score of 1 (indicating one standard deviation above the mean), while a score of 60 would have a z-score of -1.d |
| **Data Reduction** : refers to the process of reducing the volume or complexity of a dataset while retaining its meaningful information. This can involve techniques such as dimensionality reduction, where the number of variables or features is decreased, or sampling methods, where a subset of data is selected to represent the whole. Data reduction aims to improve efficiency in data storage, processing, and analysis, making it easier to work with large datasets and extract relevant insights.  **Data Reduction Strategies**  **1) Data Cube Aggregation**  Data cube aggregation involves summarizing data at multiple levels of granularity within a data cube, effectively reducing the size of the dataset to be managed. By referencing appropriate levels of aggregation, it ensures that queries regarding summarized information are efficiently answered using the data cube. This technique leverages the smallest representation capable of solving the given task, optimizing storage and processing resources. Data cube aggregation is especially useful in scenarios where quick access to aggregated data is crucial, such as in OLAP (Online Analytical Processing) systems.  • Attribute subset selection/Feature subset selection/feature creation: Irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed. | **2) Dimensionality Reduction**  Dimensionality reduction is the process of decreasing the number of random variables or attributes considered in a dataset, often through data encoding or transformations. Techniques like wavelet transforms and principal components analysis (PCA) project the original data onto a smaller space, yielding a compressed representation. This process includes attribute subset selection or feature creation, where irrelevant, weakly relevant, or redundant attributes are identified and removed. By simplifying the data structure, dimensionality reduction enhances computational efficiency and improves the performance of machine learning models.  **3) Numerosity Reduction**  Numerosity reduction techniques aim to decrease the original data volume by representing it in a more compact form. Both parametric and non-parametric methods are employed in this process. Parametric methods assume the data fits a particular model, estimate the model parameters, and store only these parameters, discarding the original data except for potential outliers. Non-parametric methods, which do not assume any specific model, include techniques such as histograms, clustering, and sampling. These methods effectively compress the data while preserving essential information, making it more manageable for analysis and storage |
| **4) Data Compression**  **String compression** • There are extensive theories and well-tuned algorithms • Typically lossless • But only limited manipulation is possible  **Audio/video, image compression** • Typically lossy compression, with progressive refinement • Sometimes small fragments of signal can be reconstructed without reconstructing the whole  **Clustering** is a data reduction technique that involves grouping a dataset into clusters based on similarity and storing only the representative information for each cluster, such as centroids or cluster diameters. The quality of clusters is assessed by metrics like diameter (max distance within a cluster) or centroid distance (average distance from objects to cluster centers), making clustering effective when data is not widely dispersed.  **Sampling** is another data reduction method that simplifies large datasets by selecting representative subsets. Simple random sampling chooses items with equal probability, but it can perform poorly with skewed data distributions. Sampling without replacement removes selected items from consideration, while sampling with replacement keeps them in the pool. Stratified sampling partitions the dataset and draws samples proportionally from each partition, making it useful for skewed datasets where balanced representation is crucial. | **5) Data Discretization**  **Discretization and Concept Hierarchies**   1. **Discretization**   Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.  **Three types of attributes:**  • Nominal — values from an unordered set  • Ordinal — values from an ordered set  • Continuous — real numbers  **Discretization/Quantization:**  divide the range of a continuous attribute into intervals Some classification algorithms only accept categorical attributes. Reduce data size by discretization Prepare for further analysis   1. **Concept Hierarchies**   Reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).  • Hierarchical and recursive decomposition using:  • Binning (data smoothing)  • Histogram analysis (numerosity reduction)  • Clustering analysis (numerosity reduction)  • Entropy-based discretization  • Segmentation by natural partitioning |