Experiment No. 1

Aim:

Data Wrangling, I

Perform the following operations using Python on any open source dataset (e.g., data.csv)

1) Import all the required Python Libraries.
2) Locate an open source data from the web (e.g., https://www.kaggle.com). Provide a clear description of the data and its source (i.e., URL of the web site).
3) Load the Dataset into pandas dataframe.
4) Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
5) Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.
6) Turn categorical variables into quantitative variables in Python.

Requirement:

- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook

Theory:

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

With the amount of data and data sources rapidly growing and expanding, it is getting increasingly essential for large amounts of available data to be organized for analysis. This process typically includes manually converting and mapping data from one raw form into another format to allow for more convenient consumption and organization of the data.

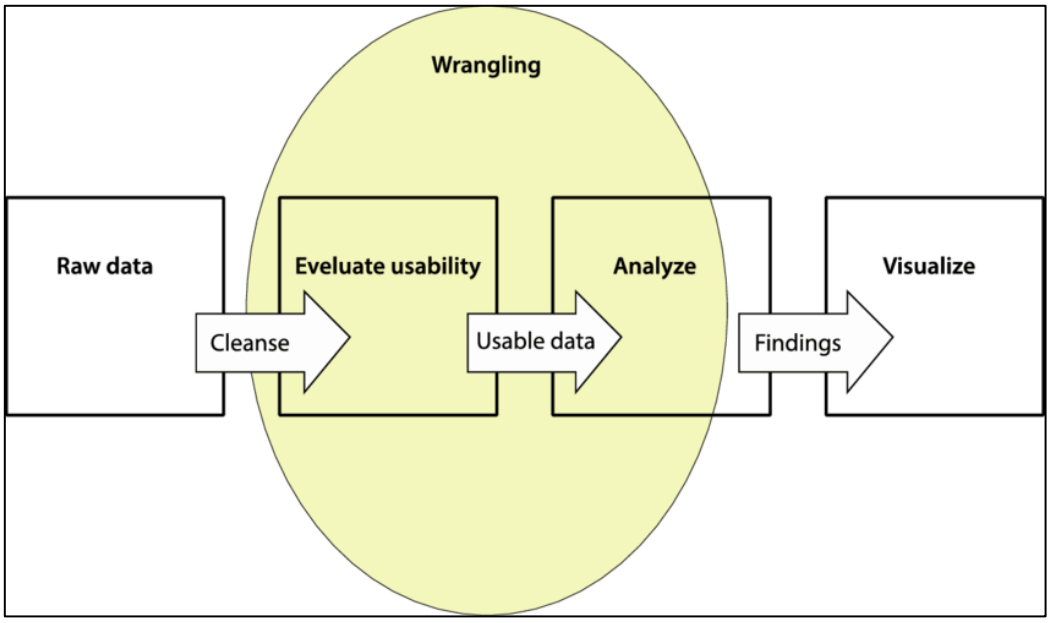
The Goals of Data Wrangling:

- Reveal a "deeper intelligence" by gathering data from multiple sources
- Provide accurate, actionable data in the hands of business analysts in a timely matter

- Reduce the time spent collecting and organizing unruly data before it can be utilized
- Enable data scientists and analysts to focus on the analysis of data, rather than the wrangling
- Drive better decision-making skills by senior leaders in an organization

Key steps to Data Wrangling:

- Data Acquisition: Identify and obtain access to the data within your sources.
- Joining Data: Combine the edited data for further use and analysis.
- Data Cleansing: Redesign the data into a usable and functional format and correct/remove any bad data.



Libraries Used:

Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring and manipulating data.

Numpy: NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Conclusion:

Hence, we have implemented data wrangling practical.

Experiment No. 2

Aim:

Data Wrangling II

Create an "Academic performance" dataset of students and perform the following operations using Python.

1) Scan all variables for missing values and inconsistencies. If there are missing values and/or inconsistencies, use any of the suitable techniques to deal with them.
2) Scan all numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them.
3) Apply data transformations on at least one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness and convert the distribution into a normal distribution.

Requirement:

- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook

Theory:

Data Wrangling:

Data wrangling can be defined as the process of cleaning, organizing, and transforming raw data into the desired format for analysts to use for prompt decision-making. Also known as data cleaning or data munging, data wrangling enables businesses to tackle more complex data in less time, produce more accurate results, and make better decisions. The exact methods vary from project to project depending upon your data and the goal you are trying to achieve. More and more organizations are increasingly relying on data wrangling tools to make data ready for downstream analytics.

Benefits of Data Wrangling:

- Data wrangling helps to improve data usability as it converts data into a compatible format for the end system.
- It helps to quickly build data flows within an intuitive user interface and easily schedule and automate the data-flow process.

- Integrates various types of information and their sources (like databases, web services, files, etc.)
- Help users to process very large volumes of data easily and easily share data-flow techniques.

Data Wrangling Vs. ETL:

ETL stands for Extract, Transform and Load. ETL is a middleware process that involves mining or extracting data from various sources, joining the data, transforming data as per business rules, and subsequently loading data to the target systems. ETL is generally used for loading processed data to flat files or relational database tables.

Though Data Wrangling and ETL look similar, there are key differences between data wrangling and ETL processes that set them apart.

- Users – Analysts, statisticians, business users, executives, and managers use data wrangling. In comparison, DW/ETL developers use ETL as an intermediate process linking source systems and reporting layers.
- Data Structure – Data wrangling involves varied and complex data sets, while ETL involves structured or semi-structured relational data sets.
- Use Case – Data wrangling is normally used for exploratory data analysis, but ETL is used for gathering, transforming, and loading data for reporting.

Data Wrangling Tools:

- Spreadsheets / Excel Power Query - It is the most basic manual data wrangling tool
- Tabula – It is a tool suited for all data types
- Google DataPrep – It is a data service that explores, cleans, and prepares data
- Data wrangler – It is a data cleaning and transforming tool

Conclusion:

Hence, we have implemented Data Wrangling Practical II.

Experiment No. 3

Aim:

1) Provide summery statistics for a dataset with numeric varibles grouped by one of the qualitative variables.

2) write a python program to display some basic statistical details like percentile, mean, standard devivation , etc.

Requirement:

- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook

Theory:

Step 1: Provide summery statistics such as mean, median, mode, standard deviation.

(1) Mean:

“Average” value is termed as mean of the dataset

means= sum of all data values / Total number of Data Values

(2) Median:

The middle values of sorted dataset is known as median.

(3) Mode:

mode refers to most frequently occuring values in the dataset.

e.g. Consider the weight (in kg) of 5children as 36,40,32,42,30.lets compute mean, median,mode

(1) Mean=(36+40+32+42+30) / 5

= 36 kg

(2) median: arrange the data in ascending order : 30,22,36,40,42

the middle value is 36. so median is 36 kg.

(3) mode: 36 kg occurs most number of times so mode=36 kg

Calculate mean using python:

```
df = pd.Dataframe(dict)
mean=df.[‘score’].mean()
print(mean)
```

Calculate median using python:

```
df = pd.Dataframe(dict)
mode=df.mode()
print(mode)
```

Calculating maximum and minimum python:

```
df=pd.Dataframe[[10,20,30,40],[7,14,21,28],[55,15,8,12],[15,14,1,8],[7,1,1,8],[5,4,9,2]]
coloumns=['Apple', 'Orange', 'Banana' , 'Peer']
index=['Basket1' , 'Basket2' , 'Basket3' , 'Basket4' , ' Basket5' , ' Basket6']

minimum=df.[['Apple' , 'Orange' , ' Banana', 'Peer']].min();
maximum=df.[['Apple' , 'Orange' , ' Banana', 'Peer']].max();

print(minimum);
print(maximum);
```

Standard Deviation:

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.

The standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. The standard deviation is calculated as the square root of variance by determining each data point's deviation relative to the mean.

Key Takeways:

- It is calculated as the square root of the variance.
- Standard deviation, in finance, is often used as a measure of a relative riskiness of an asset.
- A volatile stock has a high standard deviation, while the deviation of a stable blue-chip stock is usually rather low.
- As a downside, the standard deviation calculates all uncertainty as risk, even when it's in the investor's favor—such as above-average returns.

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

where:

x_i = Value of the i^{th} point in the data set

$\overline{x}$ = The mean value of the data set

n = The number of data points in the data

Central Tendency Measure	Pros	Cons
Mean	Sensitive as it takes all data values into account(reliable)	Biased output if outliers/extreme values exist in the data set
Median	Not affected by extreme values	-Less sensitive than Mean as it only focusses on giving out the middle data point irrespective of how far the other values are from the middle -Needs the data to be arranged in the ascending order before computing
Mode	Not affected by extreme values and can be used with non-numerical data	There may be more than one mode or no mode at all and it may not reflect data summary accurately

Libraries Used:

1. Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring and manipulating data.
2. Numpy: is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.
3. Seaborn: Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python.

Conclusion:

From this experiment we learnt how to calculate mean, median and mode.

Experiment No. 4

Aim: Create a Linear Regression model using python to predict home prise using Boston Housing dataset.

Requirement:

- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook

Theory:

Linear Regression in data science:

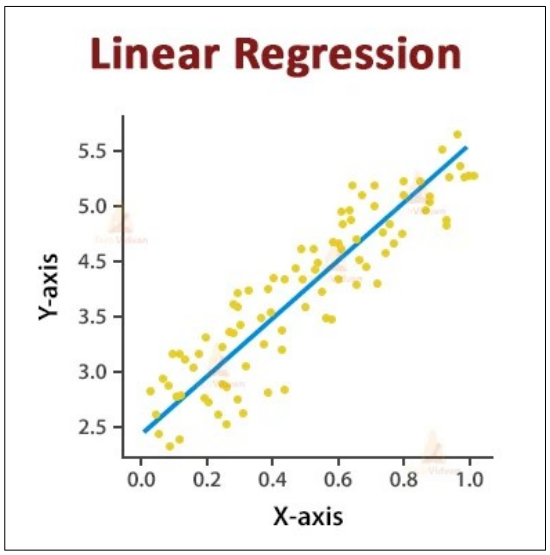


fig. Linear Regression

The term regression is used when you try to find the relationship between variables. In Machine Learning and in statistical modeling, that relationship is used to predict the outcome of events.

Simple Linear Regression:

Simple linear regression is an approach for predicting a response using a single feature. It is assumed that the two variables are linearly related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).

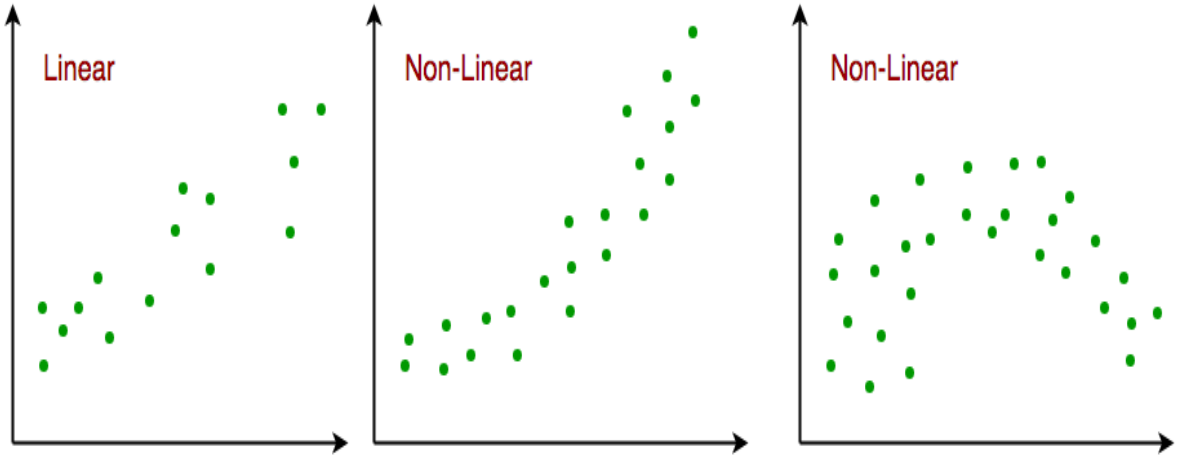
Multiple linear regression:

Multiple linear regression attempts to model the relationship between two or more features and a response by fitting a linear equation to the observed data. Clearly, it is nothing but an extension of simple linear regression.

Assumptions:

Given below are the basic assumptions that a linear regression model makes regarding a dataset on which it is applied:

• Linear relationship: Relationship between response and feature variables should be linear. The linearity assumption can be tested using scatter plots. As shown below, 1st figure represents linearly related variables whereas variables in the 2nd and 3rd figures are most likely non-linear. So, 1st figure will give better predictions using linear regression.



• Little or no multi-collinearity: It is assumed that there is little or no multicollinearity in the data. Multicollinearity occurs when the features (or independent variables) are not independent of each other.

• Little or no auto-correlation: Another assumption is that there is little or no autocorrelation in the data. Autocorrelation occurs when the residual errors are not independent of each other. You can refer here for more insight into this topic.

•Homoscedasticity: Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. As shown below, figure 1 has homoscedasticity while figure 2 has heteroscedasticity.

Applications:

1. Trend lines: A trend line represents the variation in quantitative data with the passage of time (like GDP, oil prices, etc.). These trends usually follow a linear relationship. Hence, linear regression can be applied to predict future values. However, this method suffers from a lack of scientific validity in cases where other potential changes can affect the data.
2. Economics: Linear regression is the predominant empirical tool in economics. For example, it is used to predict consumer spending, fixed investment spending, inventory investment, purchases of a country's exports, spending on imports, the demand to hold liquid assets, labor demand, and labor supply.
3. Finance: The capital price asset model uses linear regression to analyze and quantify the systematic risks of an investment.
   Biology: Linear regression is used to model causal relationships between parameters in biological systems.

Dataset used:

- In this experminet we are going to use the boston housing dataset which contain information about various houses in boston through different parameters.
- There are total 506 samples ad 14 features (columns) in this dataset.
- Our objective is to predict the value of prices of the house using features with the help of linear regression.

Libraries Uesd:

1. Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring and manipulating data.
2. Sklearn: It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Conclusion:

In this experiment we have studied abbout linear regression and done house prise prediction using boston housing dataset.

Experiment No. 5

Aim: Implement logistic regression using python to perform classifiaction on social network_ads , cv dataset.

Requirement:

- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook

Theory:

Logistic Regression?

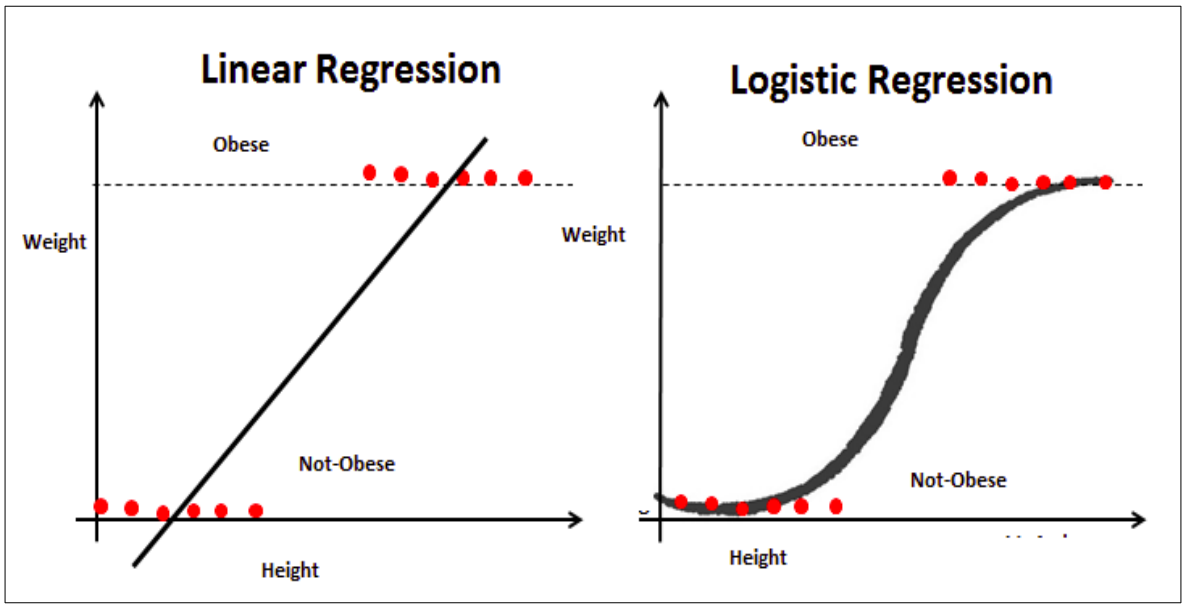


Fig. Linear Regression Vs Logistic Regression

- Logistic regression is a supervised learning algorithm used to predict a dependent categorical target variable. In essence, if you have a large set of data that you want to categorize, logistic regression may be able to help.
- For example, if you were given a dog and an orange and you wanted to find out whether each of these items was an animal or not, the desired result would be for the dog to end up classified as an animal, and for the orange to be categorized as not an animal.
- Animal is your target; it is dependent on your data in order to be able to classify the item correctly. In this example, there are only two possible answers (binary logistic regression), animal or not an animal. However, it is also possible to set up your logistic regression with more than two possible categories (multinomial logistic regression).
- To dive a little deeper into how your model might attempt to classify these two items directly, let's consider what else the model would need to know about the items in order to decide where they belong. Other similar aspects of these items

would need to be looked at when considering how to classify each item or data point. Aspects, or features, may include color, size, weight, shape, height, volume or amount of limbs.

- In this way, knowing that an orange's shape was a circle may help the algorithm to conclude that the orange was not an animal. Similarly, knowing that the orange had zero limbs would help as well.
- Logistic regression requires that the dependent variable, in this case whether the item was an animal or not, be categorical. The outcome is either animal or not an animal—there is no range in between.
- A problem that has a continuous outcome, such as predicting the grade of a student or the fuel tank range of a car, is not a good candidate to use logistic regression. Other options like linear regression may be more appropriate.

Types of Logistic Regression:

There are three main types of logistic regression:

1) binary
2) multinomial
3) ordinal.

They differ in execution and theory. Binary regression deals with two possible values, essentially: yes or no. Multinomial logistic regression deals with three or more values. And ordinal logistic regression deals with three or more classes in a predetermined order.

1. Binary logistic regression:

Binary logistic regression was mentioned earlier in the case of classifying an object as an animal or not an animal—it's an either/or solution. There are just two possible outcome answers. This concept is typically represented as a 0 or a 1 in coding.

Examples include:

• Whether or not to lend to a bank customer (outcomes are yes or no).
• Assessing cancer risk (outcomes are high or low).
• Will a team win tomorrow's game (outcomes are yes or no).

2. Multinomial logistic regression:

Multinomial logistic regression is a model where there are multiple classes that an item can be classified as. There is a set of three or more predefined classes set up prior to running the model.

Examples include:
• Classifying texts into what language they come from.
• Predicting whether a student will go to college, trade school or into the workforce.
• Does your cat prefer wet food, dry food or human food?

3. Ordinal logistic regression:

Ordinal logistic regression is also a model where there are multiple classes that an item can be classified as; however, in this case an ordering of classes is required. Classes do not need to be proportionate. The distance between each class can vary. Examples include:
• Ranking restaurants on a scale of 0 to 5 stars.
• Predicting the podium results of an Olympic event.
• Assessing a choice of candidates, specifically in places that institute ranked-choice voting.

4. Logistic regression assumptions:

• Remove highly correlated inputs.
• Consider removing outliers in your training set because logistic regression will not give significant weight to them during its calculations.
• Does not favor sparse (consisting of a lot of zero values) data.
• Logistic regression is a classification model, unlike linear regression.

Libraries Used:

1. Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring and manipulating data.

2. Sklearn: It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

3. Seaborn: Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python.

4. Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

Conclusion:

In this experiment we have studied about the logistic regression model. We ahve performed the classification on the social network. Ads dataset using various liatories of python.

Experiment No. 6

Aim: Implement simple Navie Bayes Classifications algorithm. Using python on iris.csv dataset.

Requirement:

- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook

Theory:

Naïve Bayes Classifier Algorithm:

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Why is it called Naïve Bayes?

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

- **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem:

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

- The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

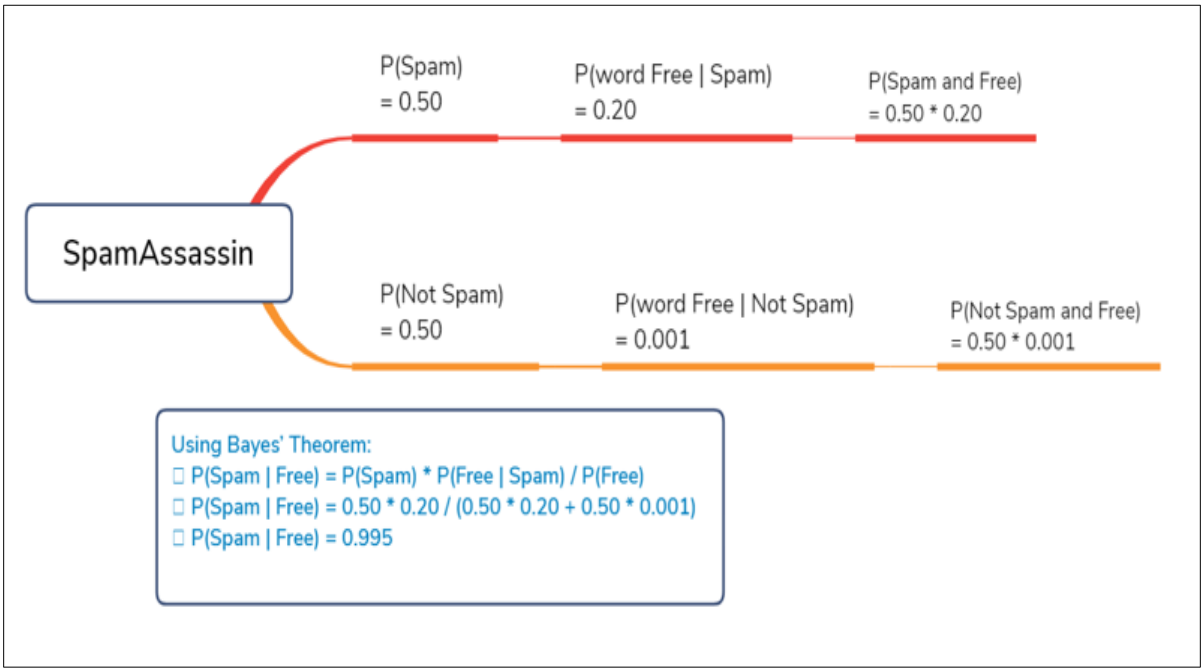
P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Example:



Advantages of Naive Bayes:

- This algorithm works very fast and can easily predict the class of a test dataset.

- You can use it to solve multi-class prediction problems as it's quite useful with them.
- Naive Bayes classifier performs better than other models with less training data if the assumption of independence of features holds.
- If you have categorical input variables, the Naive Bayes algorithm performs exceptionally well in comparison to numerical variables.

Disadvantages of Naive Bayes:

- If your test data set has a categorical variable of a category that wasn't present in the training data set, the Naive Bayes model will assign it zero probability and won't be able to make any predictions in this regard. This phenomenon is called 'Zero Frequency,' and you'll have to use a smoothing technique to solve this problem.
- This algorithm is also notorious as a lousy estimator. So, you shouldn't take the probability outputs of 'predict_proba' too seriously.
- It assumes that all the features are independent. While it might sound great in theory, in real life, you'll hardly find a set of independent features.

Applications of Naive Bayes Algorithm:

As you must've noticed, this algorithm offers plenty of advantages to its users. That's why it has a lot of applications in various sectors too. Here are some applications of Naive Bayes algorithm:

- As this algorithm is fast and efficient, you can use it to make real-time predictions.
- This algorithm is popular for multi-class predictions. You can find the probability of multiple target classes easily by using this algorithm.
- Email services (like Gmail) use this algorithm to figure out whether an email is a spam or not. This algorithm is excellent for spam filtering.

- Its assumption of feature independence, and its effectiveness in solving multi-class problems, makes it perfect for performing Sentiment Analysis. Sentiment Analysis refers to the identification of positive or negative sentiments of a target group (customers, audience, etc.)
- Collaborative Filtering and the Naive Bayes algorithm work together to build recommendation systems. These systems use data mining and machine learning to predict if the user would like a particular resource or not.

Types of Naive Bayes Classifier:
This algorithm has multiple kinds. Here are the main ones:

1. Bernoulli Naive Bayes:

Here, the predictors are boolean variables. So, the only values you have are 'True' and 'False' (you could also have 'Yes' or 'No'). We use it when the data is according to multivariate Bernoulli distribution.

2. Multinomial Naive Bayes:

People use this algorithm to solve document classification problems. For example, if you want to determine whether a document belongs to the 'Legal' category or 'Human Resources' category, you'd use this algorithm to sort it out. It uses the frequency of the present words as features.

3. Gaussian Naive Bayes

If the predictors aren't discrete but have a continuous value, we assume that they are a sample from a gaussian distribution.

Libraries Used:

1. Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring and manipulating data.
2. Sklearn: It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Conclusion:

In this experiment we have studied about Navie bayes algorithm and implemented it an the iris dataset for spam filteration

Experiment No. 7

Aim: Extract Sample document and apply following document preprocessing methods: Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization.

Requirement:

- Anaconda Installer
- Windows 10 OS
- Linux
- Jupyter Notebook

Theory:

Data Preprocessing in Python:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

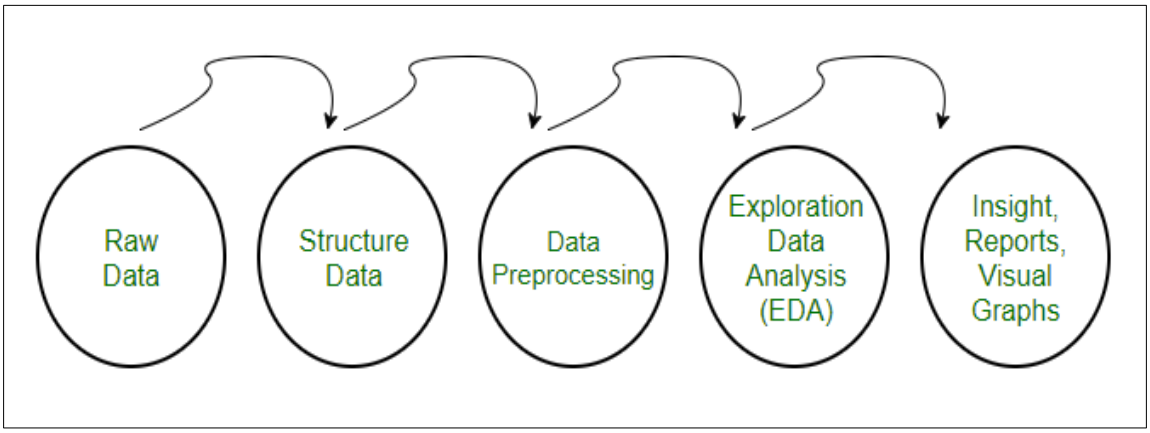
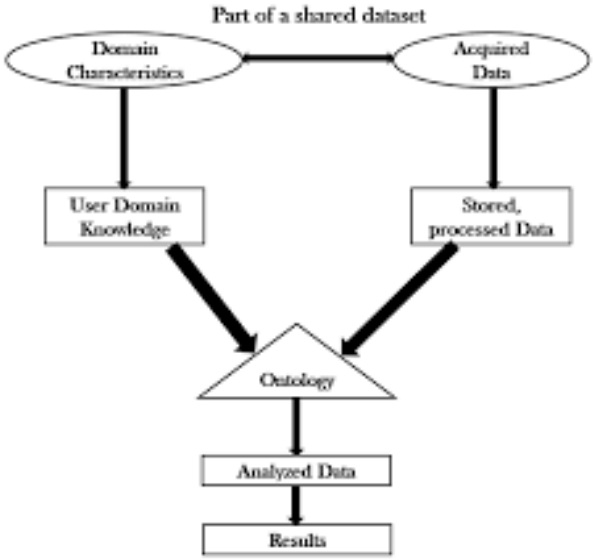


Fig. Data Preprocessing in Python

Need of Data Preprocessing:

- For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set.

- Another aspect is that the data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithm are executed in one data set, and best out of them is chosen.



NLP Techniques in data science:

1. Tokenize text using NLTK in python:

To run the below python program, (NLTK) natural language toolkit has to be installed in your system.The NLTK module is a massive tool kit, aimed at helping you with the entire Natural Language Processing (NLP) methodology. In order to install NLTK run the following commands in your terminal.

- sudo pip install nltk
- Then, enter the python shell in your terminal by simply typing python
- Type import nltk
- nltk.download('all')

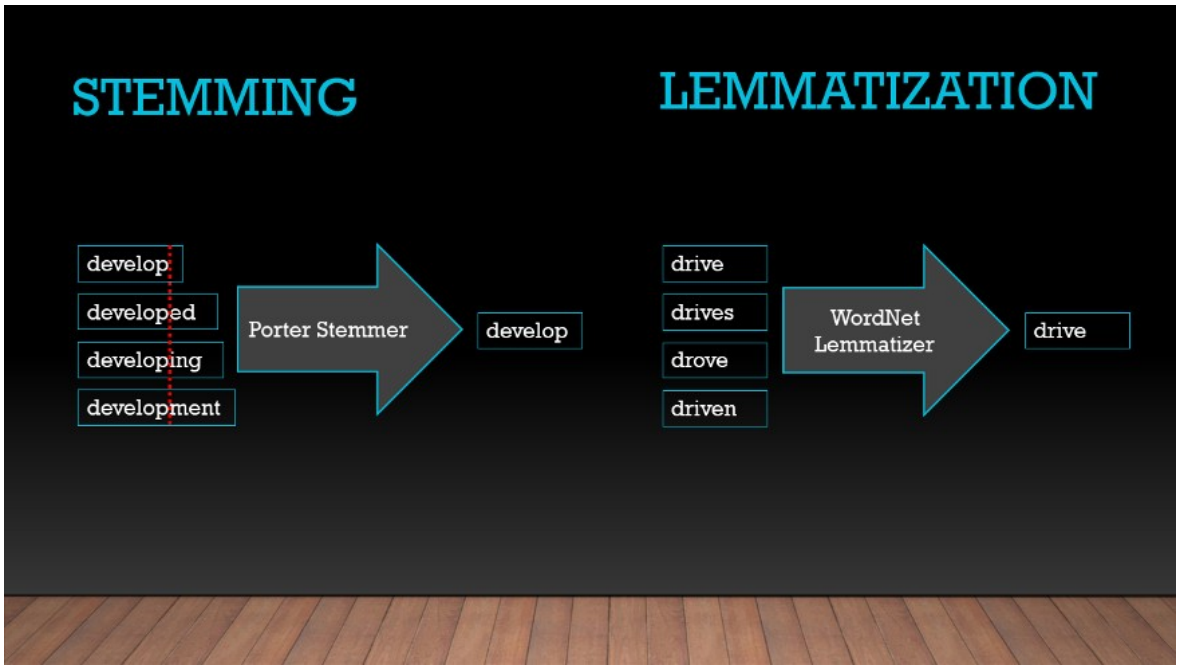
2. Removing stop words with NLTK in Python:

The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

What are Stop words?

Stop Words: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. We would not want these words to take up space in our database, or taking up valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK(Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages. You can find them in the nltk_data directory.

3. Lemmatization with NLTK:



Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So it links words with similar meanings to one word.Text preprocessing includes both Stemming as well as Lemmatization. Many times people find these two terms confusing. Some treat

these two as the same. Actually, lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

Applications of lemmatization are:

- Used in comprehensive retrieval systems like search engines.
- Used in compact indexing

Advantages and Disadvantages of Lemmatization:

As you could probably tell by now, the obvious advantage of lemmatization is that it is more accurate. So if you're dealing with an NLP application such as a chat bot or a virtual assistant where understanding the meaning of the dialogue is crucial, lemmatization would be useful. But this accuracy comes at a cost.

Because lemmatization involves deriving the meaning of a word from something like a dictionary, it's very time consuming. So most lemmatization algorithms are slower compared to their stemming counterparts. There is also a computation overhead for lemmatization, however, in an ML problem, computational resources are rarely a cause of concern.

4. Stemming words with NLTK:

Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers.

A stemming algorithm reduces the words "chocolates", "chocolatey", "choco" to the root word, "chocolate" and "retrieval", "retrieved", "retrieves" reduce to the stem "retrieve".

Some more example of stemming for root word "like" include:

-> "likes"

-> "liked"

-> "likely"

-> "liking"

Errors in Stemming:

There are mainly two errors in stemming – *Overstemming* and *Understemming*. Overstemming occurs when two words are stemmed to same root that are of different stems. Under-stemming occurs when two words are stemmed to same root that are not of different stems.

Libraries Used:

1. NLTK: NLTK is a standard python library that provides a set of diverse algorithms for NLP. It is one of the most used libraries for NLP and Computational Linguistics.
2. Sklearn: It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.
3. Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring and manipulating data.
4. String: Python String module contains some constants, utility function, and classes for string manipulation.

Conclusion:

In this experiment we have studied about data preprocessingusing natural language processing technique.

Experiment No. 8

Aim: Use the inbuilt dataset “titanic” the dataset contains 891 rows and contains informationabout the passengers who boarded the unfortunate titanic ship use the seaborn liabrary to see if we can find any patterns in the data.

Requirement:

- Anaconda Installer
- Windows 10 OS
- Linux
- Jupyter Notebook

Theory:

What is Data Visualization?

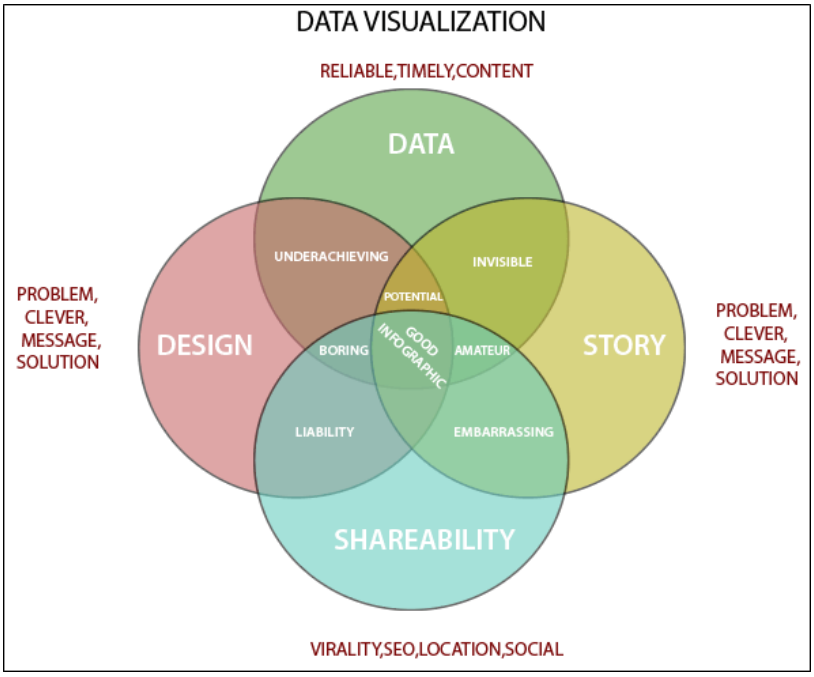


Fig. Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

The benefits of data visualization:

When considering business strategies and goals, data visualization benefits decision makers in several ways to improve data insights. Let's explore seven major benefits in detail:

- Better analysis
- Quick action
- Identifying patterns
- Finding errors
- Understanding the story
- Exploring business insights
- Grasping the Latest Trends

Better analysis:

Data visualization helps business stakeholders analyze reports regarding sales, marketing strategies, and product interest. Based on the analysis, they can focus on the areas that require attention to increase profits, which in turn makes the business more productive.

Quick action:

As mentioned previously, the human brain grasps visuals more easily than table reports. Data visualizations allow decision makers to be notified quickly of new data insights and take necessary actions for business growth.

Identifying patterns:

Large amounts of complicated data can provide many opportunities for insights when we visualize them. Visualization allows business users to recognize relationships between the data, providing greater meaning to it. Exploring these patterns helps users focus on specific areas that require attention in the data, so that they can identify the significance of those areas to drive their business forward.

Finding errors:

Visualizing your data helps quickly identify any errors in the data. If the data tends to suggest the wrong actions, visualizations help identify erroneous data sooner so that it can be removed from analysis.

Understanding the story:

Storytelling is the purpose of your dashboard. By designing your visuals in a meaningful way, you help the target audience grasp the story in a single glance. Always be sure to convey the story in the simplest way, without excessive complicated visuals.

Exploring business insights:

In the current competitive business environment, finding data correlations using visual representations is key to identifying business insights. Exploring these insights is important for business users or executives to set the right path to achieving the business' goals.

Grasping the latest trends:

Using data visualization, you can discover the latest trends in your business to provide quality products and identify problems before they arise. Staying on top of trends, you can put more effort into increasing profits for your business.

How data visualization works:

Data visualization involves handling tons of data that will be converted into meaningful visuals using widgets. To achieve this, we require the best software tools to operate various types of data sources such as files, web API data, database-maintained sources, and others. Organizations should choose the best data visualization tool to meet all their requirements.

At a minimum, the tool should support interactive visual creation, flexible connectivity to data sources, combining data sources, automatic refresh of data, sharing visuals with others, secured access to data sources, and exporting widgets. These features allow you to make the best visuals of your data and also save your business time.

Trusted, real-time data visualization:

You need a way to quickly pivot your company's efforts in response to world and customer-evolving expectations. You also need a way to make these quick business decisions using big data. But big data has been increasing in volume-becoming even bigger data. As a result, the massive amount of data is slow to sort through, comprehend and especially explain. And, if you can pull outcomes from disparate sources, it isn't easy to interpret their numerical outputs.

Libraries Used:

1. Seaborn: Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python.

<u>Conclusion</u>:

In this experiment we have studied about what is data visualization. How data visualization helps us in different ways. We have also used the inbuilt data set i.r., Titanic from the seaborn liabrary to perform data visualization on it.

Experiment No. 9

Aim: Use the inbuilt dataset ‘Titanic’ as used in previous experiment . Plot a box plot for distribution of age with respect to each gender along with the information anout wheather they survived or not.(Cloumn name ‘sex’ and ‘age’)

Requirement:

- Anaconda Installer
- Windows 10 OS
- Linux
- Jupyter Notebook

Theory:

What is Data Visualization:

Data visualization is defined as a graphical representation that contains the information and the data. By using visual elements like charts, graphs, and maps, data visualization techniques provide an accessible way to see and understand trends, outliers, and patterns in data.

In modern days we have a lot of data in our hands i.e, in the world of Big Data, data visualization tools, and technologies are crucial to analyze massive amounts of information and make data-driven decisions.

It is used in many areas such as:

• To model complex events.
• Visualize phenomenons that cannot be observed directly, such as weather patterns, medical conditions, or mathematical relationships.

Data Visualization Technique:

1. Box and Whisker Plot

• This plot can be used to obtain more statistical details about the data.
• The straight lines at the maximum and minimum are also called whiskers.
• Points that lie outside the whiskers will be considered as an outlier.
• The box plot also gives us a description of the 25th, 50th,75th quartiles.
• With the help of a box plot, we can also determine the Interquartile range(IQR) where maximum details of the data will be present. Therefore, it can also give us a clear idea about the outliers in the dataset.

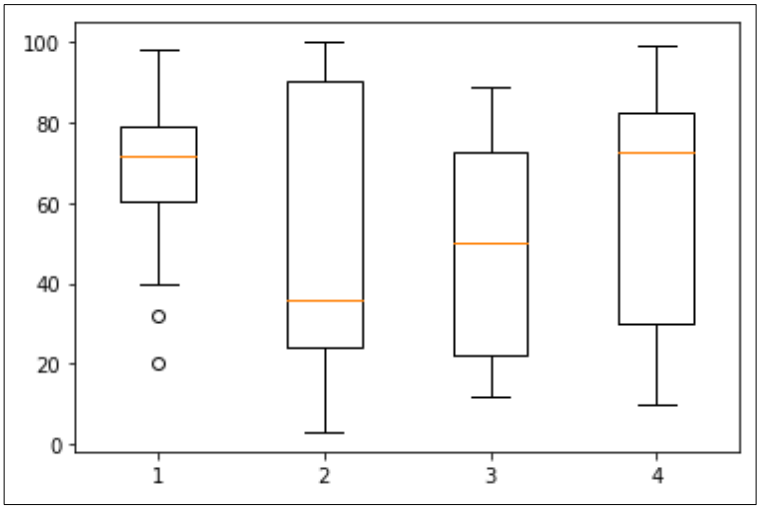


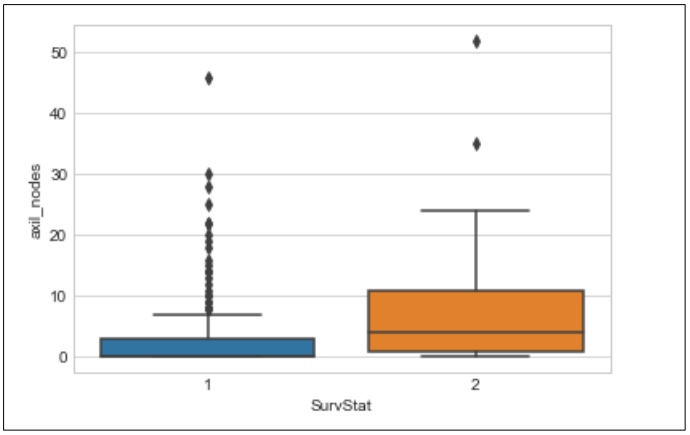
Fig. General Diagram for a Box-plot

Implementation:

- Boxplot is available in the Seaborn library.
- Here x is considered as the dependent variable and y is considered as the independent variable. These box plots come under univariate analysis, which means that we are exploring data only with one variable.
- Here we are trying to check the impact of a feature named “axil_nodes” on the class named “Survival status” and not between any two independent features.

The code snippet is as follows:

```
sns.boxplot(x='SurvStat',y='axil_nodes',data=hb)
```



Libraries Used:

Seaborn: Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python.

Conclusion: In this experiment we have studied about the data visualization technique and implemented data visualization on the built in data set in seaborn laibrary i.e., titanic dataset

Experiment No. 10

Aim:

Download the iris flower dataset or any other dataset intoa dataframe. Scan the dataset and given the interference as

1) List down the feature and types
2) Create histogram for each feature in the dataset
3) Create a boxplot for each seature in the dataset
4) Compare distributions and identify outliers.

Requirement:

- Anaconda Installer
- Windows 10 OS
- Linux
- Jupyter Notebook

Theory:

What is Data Visualization?

Data visualization simply refers to techniques that are used to communicate with insights from data through a visual representation. To simplify the thing, you can say that data visualization visually puts data. So we can easily understand it. One has to master different data visualization tools to become effective.

The main goal of data visualization is to put large datasets into visual graphics. And it is one of the important steps when it comes to data science. Also, it is a simple way to track different data points. No matter if you wish to track website metrics, sales team performance, marketing campaign, product adoption rate, or any other thing.

Data visualization is what will help you out. To learn more about data visualization and other visual representation in data science, check out our data science certifications from recognized universities.

Categorical Data:

Categorical data can be:

- nominal, qualitative
- ordinal

For visualization, the main difference is that ordinal data suggests a particular display order.Purely categorical data can come in a range of formats. The most common are:

• raw data: individual observations;
• aggregated data: counts for each unique combination of levels
• cross-tabulated data

Working With Categorical Variables:

Categorical variables are usually represented as:

- character vectors
- factors.

Some advantages of factors:

- more control over ordering of levels
- levels are preserved when forming subsets

Most plotting and modeling functions will convert character vectors to factors with levels ordered alphabetically. Some standard R functions for working with factors include:

- factor creates a factor from another type of variable
- levels returns the levels of a factor
- reorder changes level order to match another variable
- relevel moves a particular level to the first position as a base line
- droplevels removes levels not in the variable.

The tidyverse package forcats adds some more tools, including,

- fct_inorder creates a factor with levels ordered by first appearance
- fct_infreq orders levels by decreasing frequency
- fct_rev reverses the levels
- fct_recode changes factor levels
- fct_relevel moves one or more levels
- fct_c merges two or more factors

Seaborn:

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. To see the code or report a bug, please visit the GitHub repository. General support questions are most at home on stackoverflow or discourse, which have dedicated channels for seaborn.

Why data visualization is important for any career:

It's hard to think of a professional industry that doesn't benefit from making data more understandable. Every STEM field benefits from understanding data—and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on. While we'll always wax poetically about data visualization (you're on the Tableau website, after all) there are practical, real-life applications that are undeniable. And, since visualization is so prolific, it's also one of the most useful professional skills to develop.

The better you can convey your points visually, whether in a dashboard or a slide deck, the better you can leverage that information. The concept of the citizen data scientist is on the rise. Skill sets are changing to accommodate a data-driven world. It is increasingly valuable for professionals to be able to use data to make decisions and use visuals to tell stories of when data informs the who, what, when, where, and how. While traditional education typically draws a distinct line between creative storytelling and technical analysis, the modern professional world also values those who can cross between the two: data visualization sits right in the middle of analysis and visual storytelling.

Libraries Used:

1. Seaborn: Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python.
2. Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring and manipulating data.
3. Numpy: is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.
4. Sklearn: It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Conclusion:

In this experiment we have studied about the data visualization and performed data visualization on the iris data set in different ways.

Experiment No. 11 (Mini Project)

Aim:

Use the following dataset and classify tweets into positive and negative tweets. https://www.kaggle.com/ruchi798/data-science-tweet

---

Requirement:

- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook

Theory:

What is Sentiment Analysis?

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as **opinion mining**, deriving the opinion or attitude of a speaker.

Why Sentiment Analysis?

- Business: In marketing field companies use it to develop their strategies, to understand customers' feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don't buy some products.
- Politics: In political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!
- Public Actions: Sentiment analysis also is used to monitor and analyse social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

Libraries Used:

1. Pandas: Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays.

2. String: The string module contains a number of functions to process standard Python strings.

3. Sklearn: It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Conslusion:

Hence, we successfully implemented sentiment analysis using python.

Experiment No. 12 (Mini Project)

Aim:

Use the following covid_vaccine_statewise.csv dataset and perform following analytics on given dataset.
https://www.kaggle.com/sudalairajkumar/covid19-in-india?select=covid_vaccine_statewise.csv

a. Describe the dataset
b. Number of persons statewise vaccinated for first dose in India
c. Number of persons statewise vaccinated for second dose in India
d. Number of Males vaccinated
e. Number of Females vaccinated

Requirement:

- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook

Theory:

About this dataset:

Coronaviruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19 - World Health Organization The number of new cases are increasing day by day around the world. This dataset has information from the states and union territories of India at daily level.

State level data comes from Ministry of Health & Family Welfare Testing data and vaccination data comes from covid19india. Huge thanks to them for their efforts! Update on April 20, 2021: Thanks to the Team at ISIBang, I was able to get the historical data for the periods that I missed to collect and updated the csv file.

Libraries Used:

1. Pandas: Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays.

2. NumPy: NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structure to python that guarantee efficient calculations with arrays and matrices and it supplies on enormous library of high-level mathematical functions that operate on these arrays and matrices.

Conclusion:

Hence, we successfully implemented mini project using covid_vaccine_statewise.csv dataset.