SIX WEEKS SUMMER TRAINING REPORT

on

DATA SCIENCE WITH PYTHON

Submitted by

ASTUTI

Registration No. 11716658

Programme Name: Data Science

Under the Guidance of

Intershala

School of Computer Science & Engineering
Lovely Professional University, Phagwara
(June-July, 2019)

DECLARATION

I hereby declare that I have completed my six weeks summer training at Internshala from 1 June,2019 to 13 July,2019 under the guidance of Kunal Jain. I declare that I have worked with full dedication during these six weeks of training and my learning outcomes fulfil the requirements of training for the award of degree of Data Science with Python, Lovely Professional University, Phagwara.

Name of Student:

Registration no:

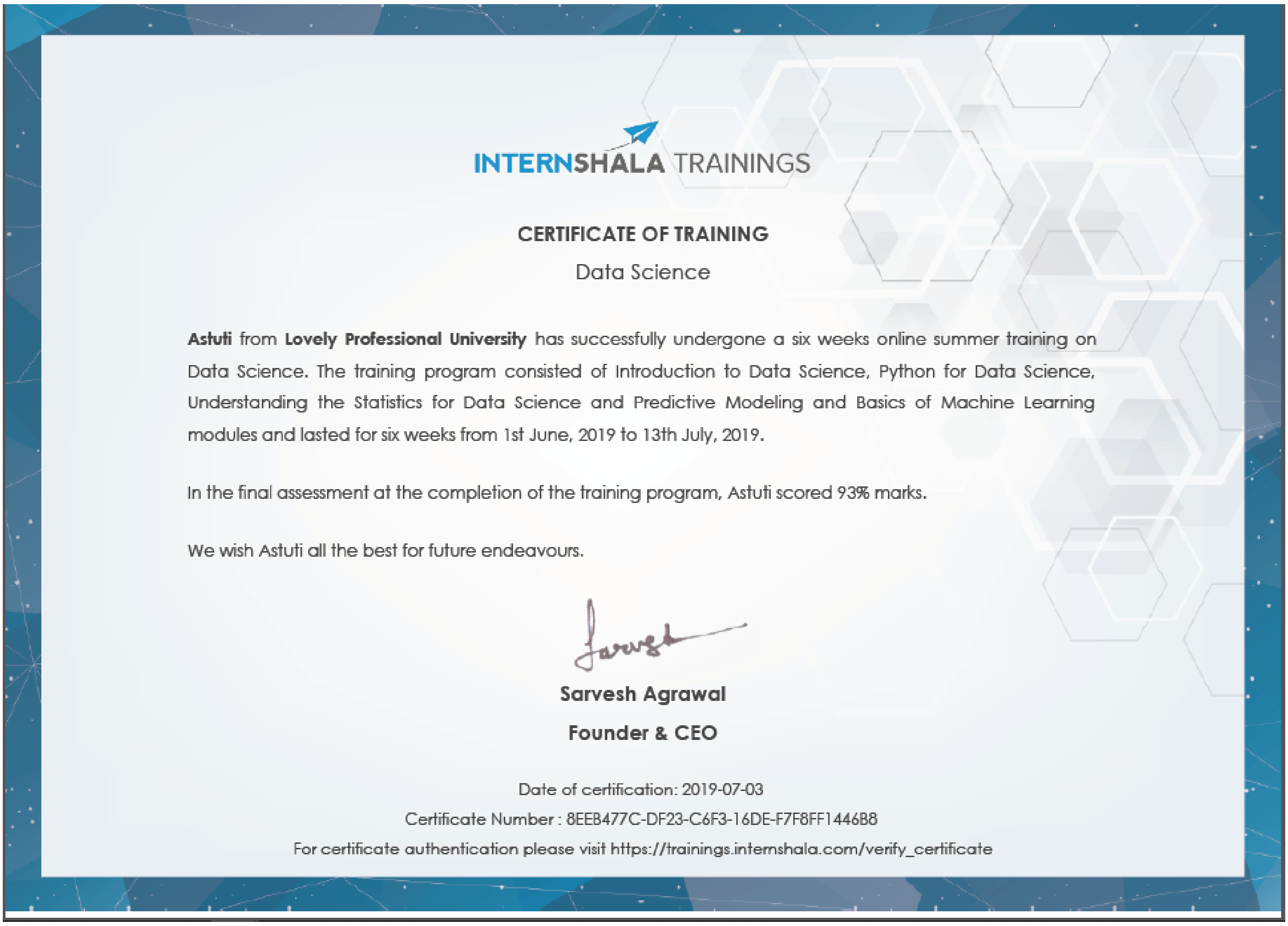
Date:

ACKNOWLEDGEMENT

It is with sense of gratitude; I acknowledge the efforts of entire hosts of well-wishers who have in some way or other contributed in their own special ways to the success and completion of the Summer Training.

Successfully completion of any type of technology requires helps from a number of people. I have also taken help from different people for the preparation of the report. Now, there is little effort to show my deep gratitude to those helpful people.

First, I express my sense of gratitude and indebtedness to our Training mentor- Kunal Jain. From the bottom of my heart, for his immense support and guidance throughout the training. Without his kind direction and proper guidance this study would have been a little success. In every phase of the project his Supervision and guidance shaped this training to be completed perfectly.



INTERNSHALA TRAININGS

CERTIFICATE OF TRAINING

Data Science

Astuti from **Lovely Professional University** has successfully undergone a six weeks online summer training on Data Science. The training program consisted of Introduction to Data Science, Python for Data Science, Understanding the Statistics for Data Science and Predictive Modeling and Basics of Machine Learning modules and lasted for six weeks from 1st June, 2019 to 13th July, 2019.

In the final assessment at the completion of the training program, Astuti scored 93% marks.

We wish Astuti all the best for future endeavours.

Sarvesh Agrawal

Founder & CEO

Date of certification: 2019-07-03

Certificate Number : 8EEB477C-DF23-C6F3-16DE-F7F8FF1446B8

For certificate authentication please visit https://trainings.internshala.com/verify_certificate

TABLE OF CONTENTS

Introduction to Data Science

Data Science

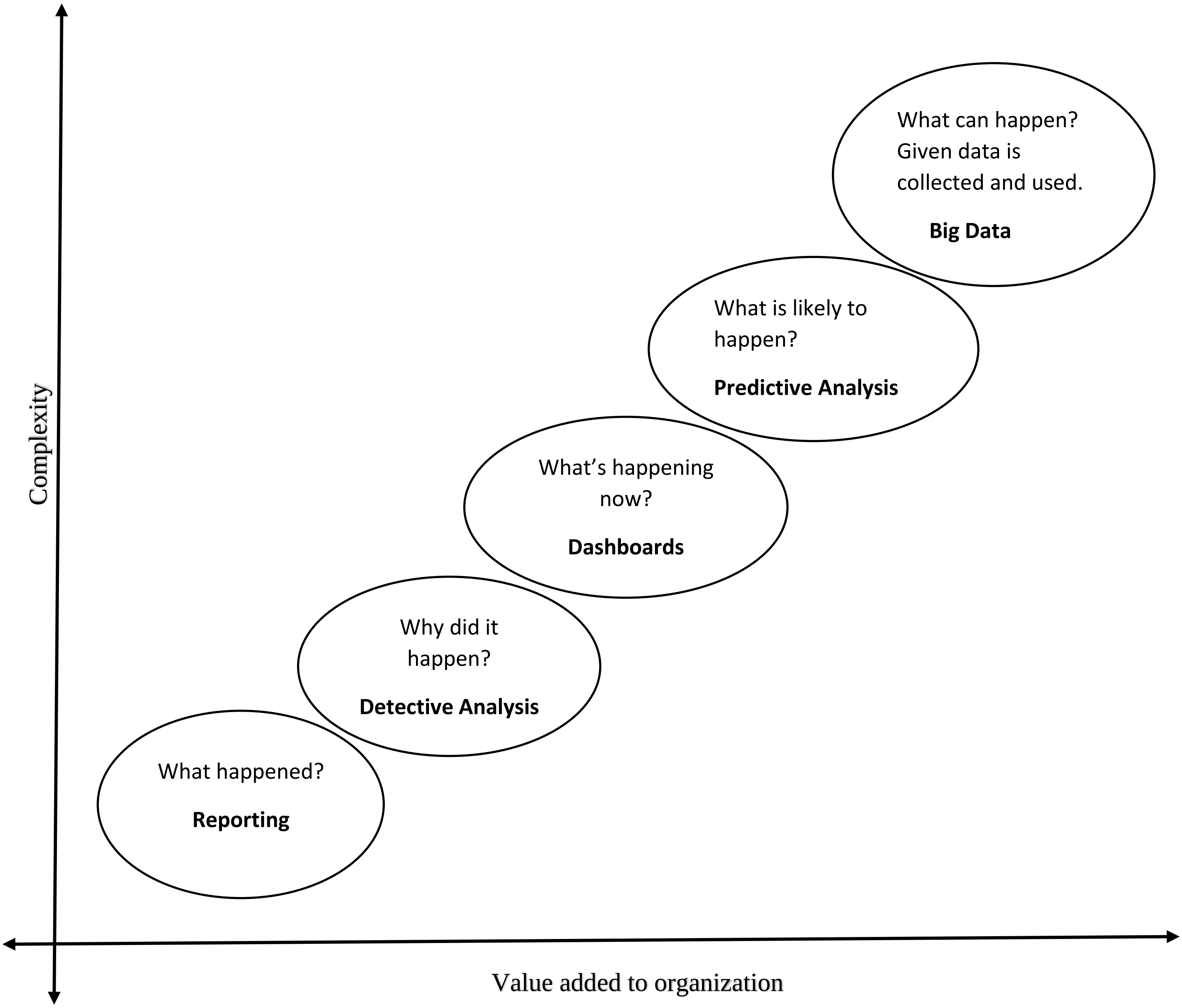
The field of bringing insights from data using scientific techniques is called **data science**.

Applications

Amazon Go – No checkout lines

Computer Vision - The advancement in recognizing an image by a computer involves processing large sets of image data from multiple objects of same category. For example, Face recognition.

Spectrum of Business Analysis



Reporting / Management Information System

To track what is happening in organization.

Detective Analysis

Asking questions based on data we are seeing, like. Why something happened?

Dashboard / Business Intelligence

Utopia of reporting. Every action about business is reflected in front of screen.

Predictive Modelling

Using past data to predict what is happening at granular level.

Big Data

Stage where complexity of handling data gets beyond the traditional system.

Can be caused because of volume, variety or velocity of data. Use specific tools to analyse such scale data.

Application of Data Science

- Recommendation System

  Example-In Amazon recommendations are different for different users according to their past search.

- Social Media

1. Recommendation Engine
2. Ad placement
3. Sentiment Analysis

- Deciding the right credit limit for credit card customers.
- Suggesting right products from e-commerce companies
    1. Recommendation System
    2. Past Data Searched
    3. Discount Price Optimization
- How google and other search engines know what are the more relevant results for our search query?
    1. Apply ML and Data Science
    2. Fraud Detection
    3. AD placement
    4. Personalized search results

Python Introduction

Python is an interpreted, high-level, general-purpose programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python's elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

Python for Data science:

Why Python???

1. Python is an open source language.
2. Syntax as simple as English.
3. Very large and Collaborative developer community.
4. Extensive Packages.

- UNDERSTANDING OPERATORS:

  Theory of operators: - Operators are symbolic representation of Mathematical tasks.

- VARIABLES AND DATATYPES:

  Variables are named bounded to objects. Data types in python are int (Integer), Float, Boolean and strings.

- CONDITIONAL STATEMENTS:

  If-else statements (Single condition)

  If- elif- else statements (Multiple Condition)

- LOOPING CONSTRUCTS:

  For loop

- FUNCTIONS:

  Functions are re-usable piece of code. Created for solving specific problem.

  Two types: Built-in functions and User- defined functions.

  Functions cannot be reused in python.

- DATA STRUCTURES:

  Two types of Data structures:

  LISTS: A list is an ordered data structure with elements separated by comma and enclosed within square brackets.

  DICTIONARY: A dictionary is an unordered data structure with elements separated by comma and stored as key: value pair, enclosed with curly braces {}.

Statistics

Descriptive Statistic

Mode

It is a number which occurs most frequently in the data series.

It is robust and is not generally affected much by addition of couple of new values.

Code

```
import pandas as pd
data=pd.read_csv( "Mode.csv")      //reads data from csv file
data.head()                         //print first five lines
mode_data=data['Subject'].mode() //to take mode of subject column
print(mode_data)
```

Mean

```
import pandas as pd
data=pd.read_csv( "mean.csv")      //reads data from csv file
data.head()                         //print first five lines
mean_data=data[Overallmarks].mean() //to take mode of subject column
print(mean_data)
```

Median

Absolute central value of data set.

```
import pandas as pd
data=pd.read_csv( "data.csv")      //reads data from csv file
data.head()                         //print first five lines
median_data=data[Overallmarks].median() //to take mode of subject column
print(median_data)
```

Types of variables

- Continous – Which takes continuous numeric values. Eg-marks
- Categorial-Which have discrete values. Eg- Gender
- Ordinal – Ordered categorial variables. Eg- Teacher feedback
- Nominal – Unorderd categorial variable. Eg- Gender

Outliers

Any value which will fall outside the range of the data is termed as a outlier. Eg- 9700 instead of 97.

Reasons of Outliers

- Typos-During collection. Eg-adding extra zero by mistake.
- Measurement Error-Outliers in data due to measurement operator being faulty.
- Intentional Error-Errors which are induced intentionally. Eg-claiming smaller amount of alcohol consumed then actual.
- Legit Outlier—These are values which are not actually errors but in data due to legitimate reasons. Eg - a CEO's salary might actually be high as compared to other employees.

Interquartile Range (IQR)

Is difference between third and first quartile from last. It is robust to outliers.

Histograms

Histograms depict the underlying frequency of a set of discrete or continuous data that are measured on an interval scale.

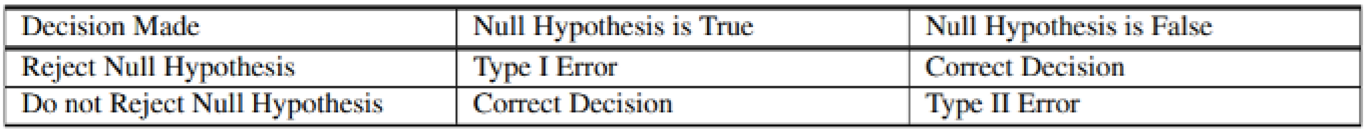
```
import pandas as pd
histogram=pd.read_csv(histogram.csv)
import matplotlib.pyplot as plt
%matplot inline
plt.hist(x= 'Overall Marks',data=histogram)
plt.show()
```

Inferential Statistics

Inferential statistics allows to make inferences about the population from the sample data.

Hypothesis Testing

Hypothesis testing is a kind of statistical inference that involves asking a question, collecting data, and then examining what the data tells us about how to proceed. The hypothesis to be tested is called the null hypothesis and given the symbol Ho. We test the null hypothesis against an alternative hypothesis, which is given the symbol Ha.



Decision Made	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	Type I Error	Correct Decision
Do not Reject Null Hypothesis	Correct Decision	Type II Error

T Tests

When we have just a sample not population statistics.

Use sample standard deviation to estimate population standard deviation.

T test is more prone to errors, because we just have samples.

Z Score

The distance in terms of number of standard deviations, the observed value is away from mean, is standard score or z score.

$$z = \frac{\bar{x} - \mu}{\sigma}$$

+Z – value is above mean.

-Z – value is below mean.

The distribution once converted to z- score is always same as that of shape of original distribution.

Chi Squared Test

To test categorical variables.

Correlation

Determine the relationship between two variables.

It is denoted by r. The value ranges from -1 to +1. Hence, 0 means no relation.

Syntax

```
import pandas as pd
import numpy as np
data=pd.read_csv("data.csv")
data.corr()
```

Predictive Modelling

Making use of past data and attributes we predict future using this data.

Eg-

Past	Horror Movies
Future	Unwatched Horror Movies

Predicting stock price movement

1. Analysing past stock prices.
2. Analysing similar stocks.
3. Future stock price required.

Types

1. Supervised Learning

   Supervised learning is a type algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and response values.

   - Regression-which have continuous possible values. Eg-Marks
   - Classification-which have only two values. Eg-Cancer prediction is either 0 or 1.

2. Unsupervised Learning

   Unsupervised learning is the training of machine using information that is neither classified nor. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

   - **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.
   - **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Stages of Predictive Modelling

1. Problem definition
2. Hypothesis Generation
3. Data Extraction/Collection
4. Data Exploration and Transformation
5. Predictive Modelling
6. Model Development/Implementation

Problem Definition

Identify the right problem statement, ideally formulate the problem mathematically.

Hypothesis Generation

List down all possible variables, which might influence problem objective. These variables should be free from personal bias and preferences.

Quality of model is directly proportional to quality of hypothesis.

Data Extraction/Collection

Collect data from different sources and combine those for exploration and model building.

While looking at data we might come across new hypothesis.

Data Exploration and Transformation

Data extraction is a process that involves retrieval of data from various sources for further data processing or data storage.

Steps of Data Extraction

- Reading the data

  Eg- From csv file
- Variable identification
- Univariate Analysis
- Bivariate Analysis
- Missing value treatment
- Outlier treatment
- Variable Transformation

Variable Treatment

It is the process of identifying whether variable is

1. Independent or dependent variable
2. Continuous or categorical variable

Why do we perform variable identification?

1. Techniques like supervised learning require identification of dependent variable.
2. Different data processing techniques for categorical and continuous data.

Categorical variable- Stored as object.

Continuous variable-Stored as int or float.

Univariate Analysis

1. Explore one variable at a time.
2. Summarize the variable.
3. Make sense out of that summary to discover insights, anomalies, etc.

Bivariate Analysis

- When two variables are studied together for their empirical relationship.
- When you want to see whether the two variables are associated with each other.
- It helps in prediction and detecting anomalies.

Missing Value Treatment

Reasons of missing value

1. Non-response – Eg-when you collect data on people's income and many choose not to answer.
2. Error in data collection. Eg- Faculty data
3. Error in data reading.

Types

1. MCAR (Missing completely at random): Missing values have no relation to the variable in which missing value exist and other variables in dataset.
2. MAR (Missing at random): Missing values have no relation to the in which missing value exist and the variables other than the variables in which missing values exist.
3. MNAR (Missing not at random): Missing values have relation to the variable in which missing value exists

Identifying

Syntax: -

1. describe()
2. Isnull()

   Output will we in True or False

Different methods to deal with missing values

1. Imputation

   Continuous-Impute with help of mean, median or regression mode.

   Categorical-With mode, classification model.
2. Deletion

   Row wise or column wise deletion. But it leads to loss of data.

Outlier Treatment

Reasons of Outliers

1. Data entry Errors
2. Measurement Errors
3. Processing Errors
4. Change in underlying population

Types of Outlier

Univariate

Analysing only one variable for outlier.

Eg – In box plot of height and weight.

Weight will we analysed for outlier

Bivariate

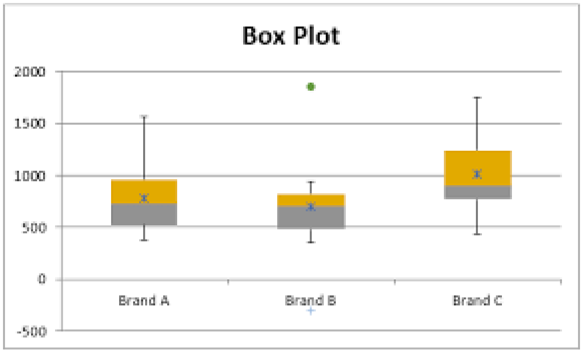
Analysing both variables for outlier.

Eg- In scatter plot graph of height and weight. Both will we analysed.

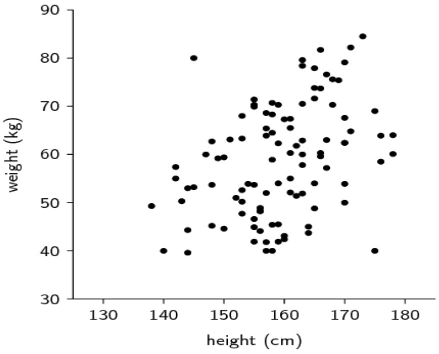
Identifying Outlier

Graphical Method

- Box Plot



- Scatter Plot



Formula Method

Using Box Plot

< Q1 - 1.5 * IQR or > Q3+1.5 * IQR

Where IQR= Q3 – Q1

Q3=Value of 3rd quartile

Q1=Value of 1st quartile

Treating Outlier

1. Deleting observations
2. Transforming and binning values
3. Imputing outliers like missing values
4. Treat them as separate

Variable Transformation

Is the process by which-

1. We replace a variable with some function of that variable. Eg – Replacing a variable x with its log.
2. We change the distribution or relationship of a variable with others.

Used to –

1. Change the scale of a variable
2. Transforming non linear relationships into linear relationship
3. Creating symmetric distribution from skewed distribution.

Common methods of Variable Transformation – Logarithm, Square root, Cube root, Binning, etc.

Model Building

It is a process to create a mathematical model for estimating / predicting the future based on past data.

Eg-

A retail wants to know the default behaviour of its credit card customers. They want to predict the probability of default for each customer in next three months.

- Probability of default would lie between 0 and 1.
- Assume every customer has a 10% default rate.

Probability of default for each customer in next 3 months=0.1

It moves the probability towards one of the extremes based on attributes of past information.

A customer with volatile income is more likely (closer to) to default.

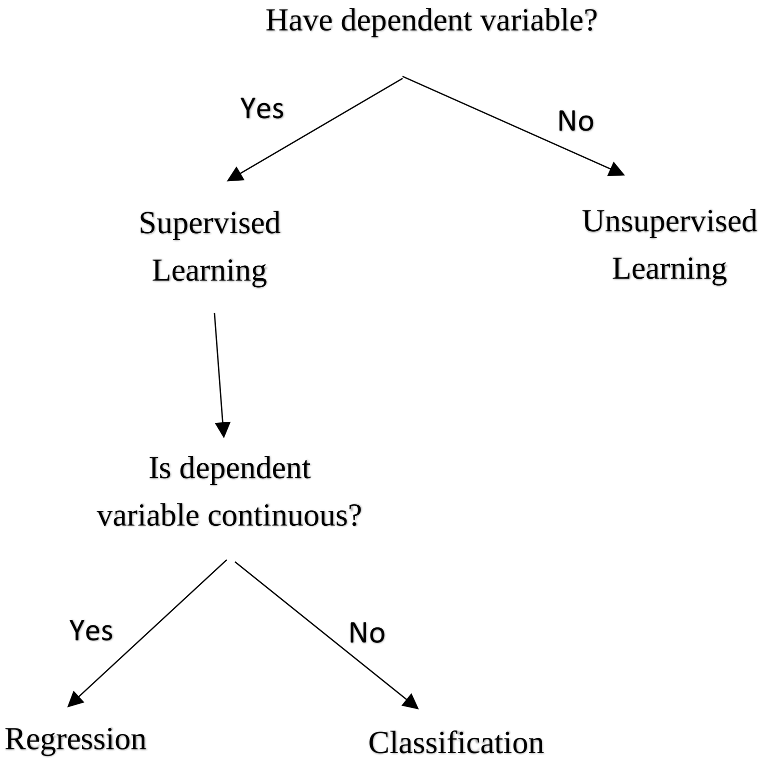
A customer with healthy credit history for last years has low chances of default (closer to 0).

Steps in Model Building

1. Algorithm Selection
2. Training Model
3. Prediction / Scoring

Algorithm Selection

Example-



Eg- Predict the customer will buy product or not.

Algorithms

- Logistic Regression
- Decision Tree
- Random Forest

Training Model

It is a process to learn relationship / correlation between independent and dependent variables.

We use dependent variable of train data set to predict/estimate.

Dataset

- Train

  Past data (known dependent variable).

  Used to train model.

- Test

  Future data (unknown dependent variable)

  Used to score.

Prediction / Scoring

It is the process to estimate/predict dependent variable of train data set by applying model rules.

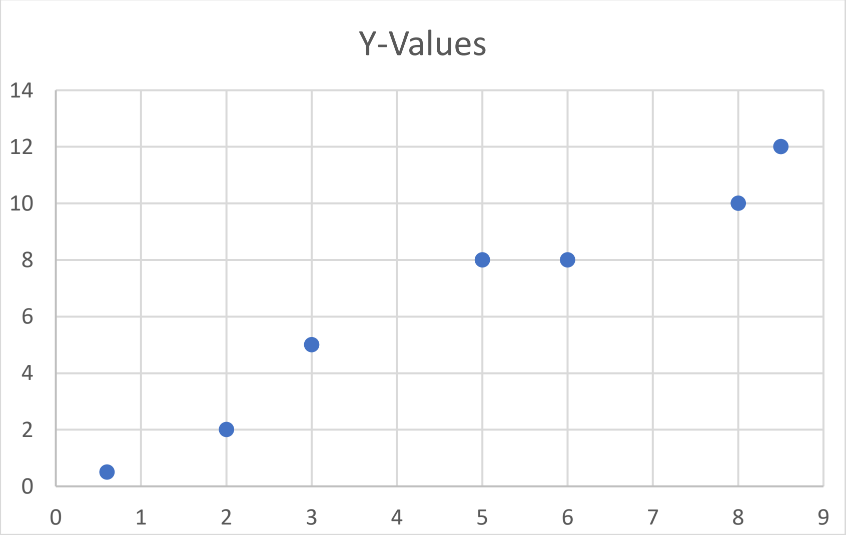
We apply training learning to test data set for prediction/estimation.

Algorithm of Machine Learning

Linear Regression

Linear regression is a statistical approach for modelling relationship between a dependent variable with a given set of independent variables.

It is assumed that the wo 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).



The equation of regression line is represented as:

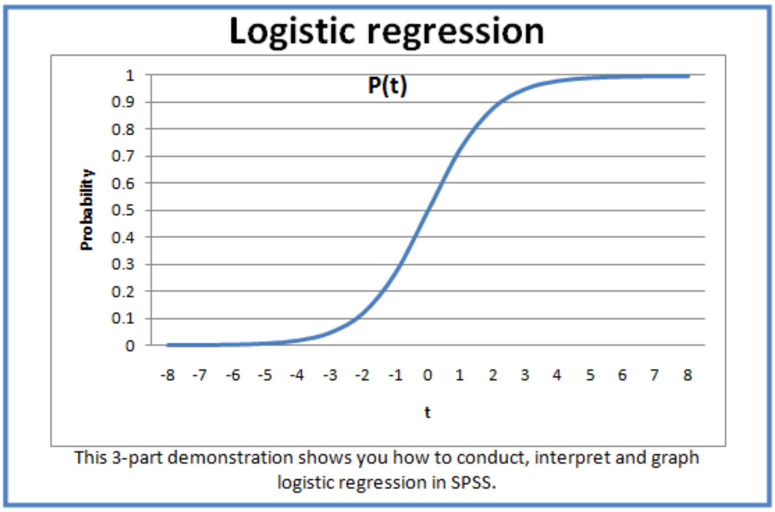
$$h(x_i) = \beta_0 + \beta_1 x_i$$

The squared error or cost function, J as:

$$J(\beta_0, \beta_1) = \frac{1}{2n} \sum_{i=1}^{n}$$

Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist.

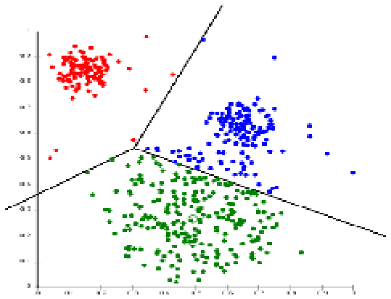


This 3-part demonstration shows you how to conduct, interpret and graph logistic regression in SPSS.

$$C = -y\ (\log(y) – (1-y)\ \log(1-y))$$

K-Means Clustering (Unsupervised learning)

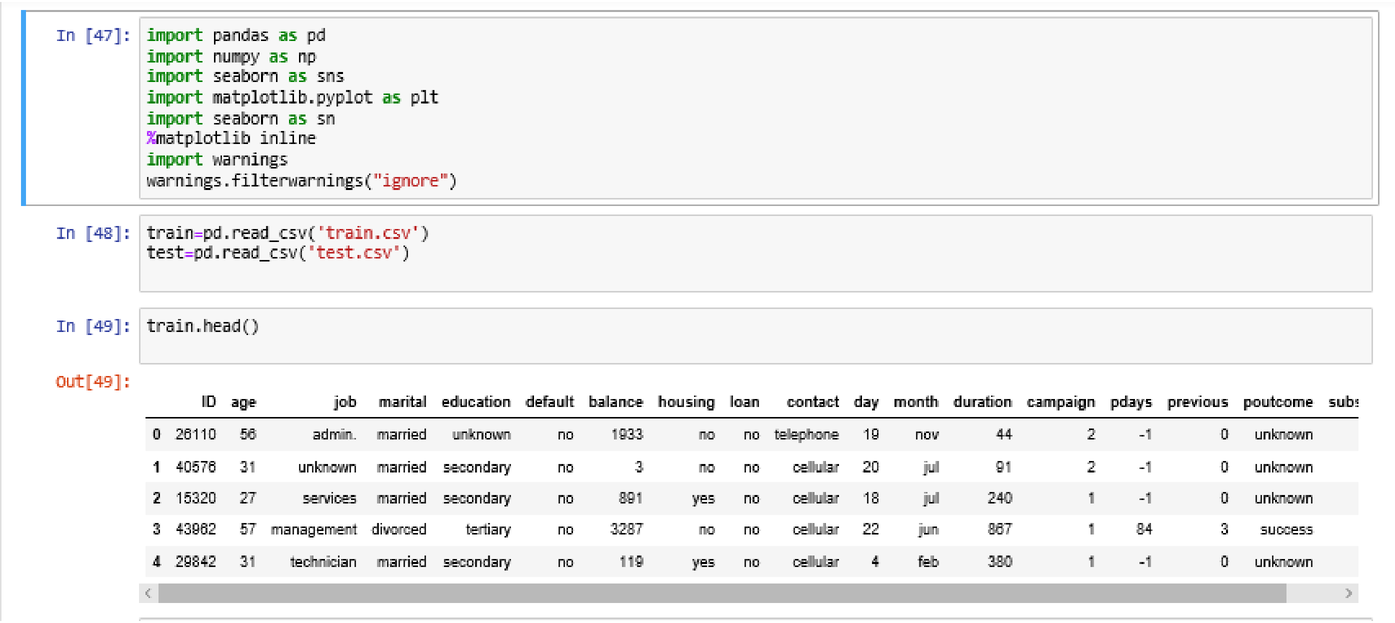
K-means clustering is a type of unsupervised learning, which is used when you have unlabelled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.



Problem Description

Provided with following files: train.csv and test.csv.

Use train.csv dataset to train the model. This file contains all the client and call details as well as the target variable “subscribed”. Then use the trained model to predict whether a new set of clients will subscribe the term deposit.



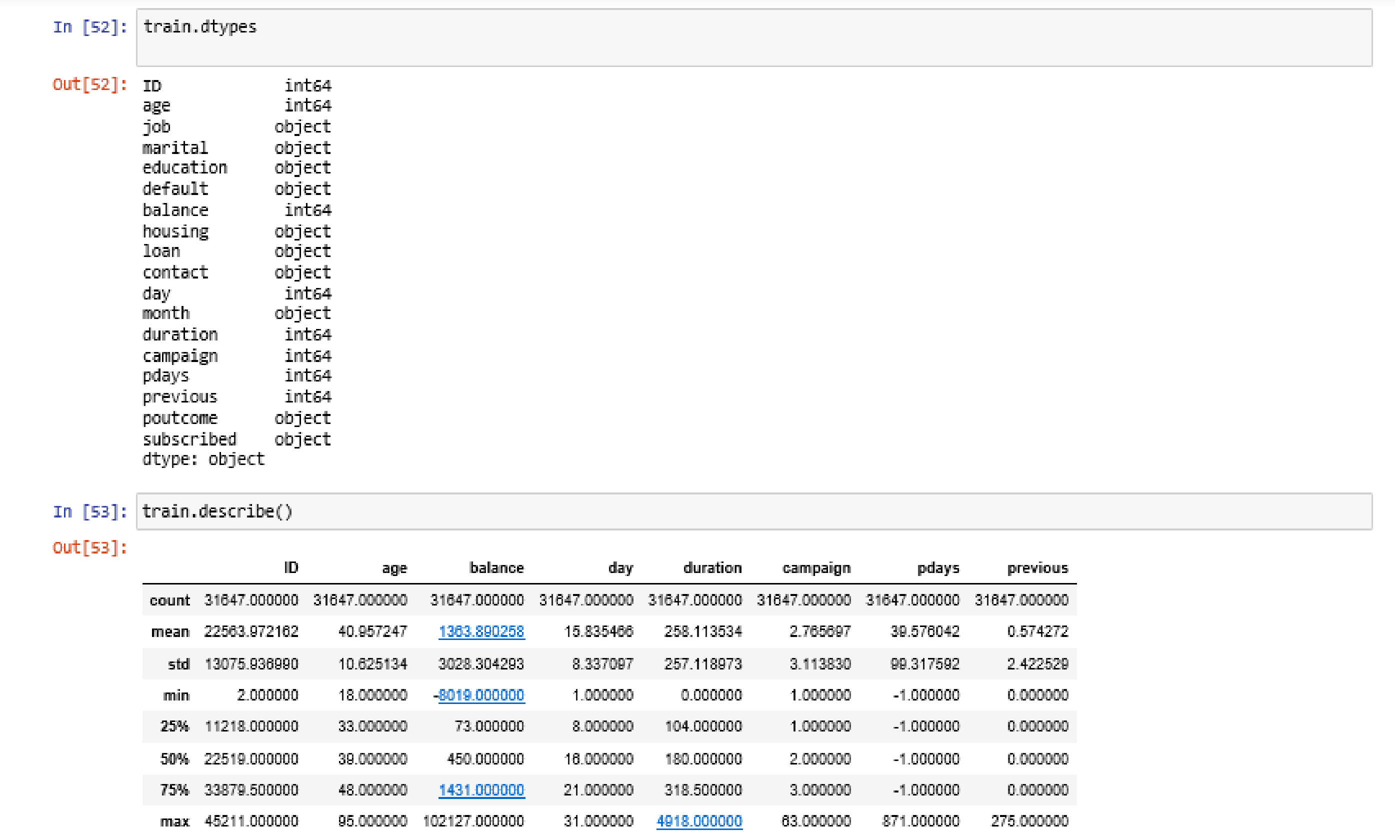
In [47]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sn
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
```

In [48]:
```
train=pd.read_csv('train.csv')
test=pd.read_csv('test.csv')
```

In [49]:
```
train.head()
```

Out[49]:

	ID	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	subs
0	26110	56	admin.	married	unknown	no	1933	no	no	telephone	19	nov	44	2	-1	0	unknown	
1	40576	31	unknown	married	secondary	no	3	no	no	cellular	20	jul	91	2	-1	0	unknown	
2	15320	27	services	married	secondary	no	891	yes	no	cellular	18	jul	240	1	-1	0	unknown	
3	43962	57	management	divorced	tertiary	no	3287	no	no	cellular	22	jun	867	1	84	3	success	
4	29842	31	technician	married	secondary	no	119	yes	no	cellular	4	feb	380	1	-1	0	unknown	



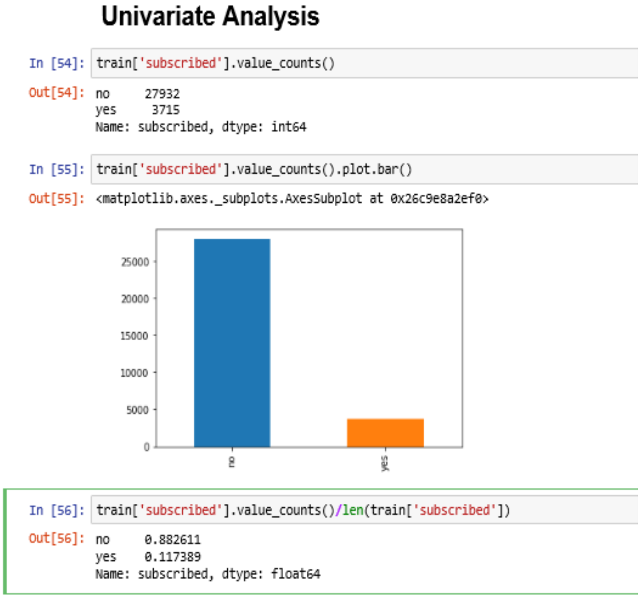
In [52]:
```
train.dtypes
```

Out[52]:
```
ID              int64
age             int64
job            object
marital        object
education      object
default        object
balance         int64
housing        object
loan           object
contact        object
day             int64
month          object
duration        int64
campaign        int64
pdays           int64
previous        int64
poutcome       object
subscribed     object
dtype: object
```

In [53]:
```
train.describe()
```

Out[53]:

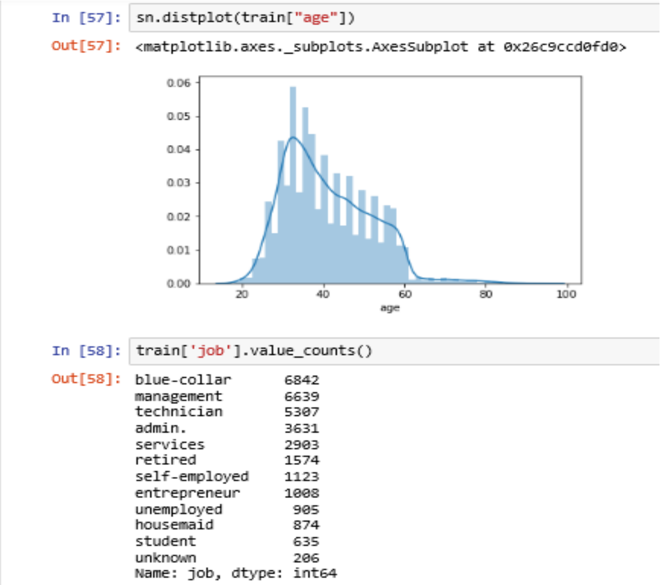
	ID	age	balance	day	duration	campaign	pdays	previous
count	31647.000000	31647.000000	31647.000000	31647.000000	31647.000000	31647.000000	31647.000000	31647.000000
mean	22563.972162	40.957247	1363.890258	15.835466	258.113534	2.765697	39.576042	0.574272
std	13075.936990	10.625134	3028.304293	8.337097	257.118973	3.113830	99.317592	2.422529
min	2.000000	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	11218.000000	33.000000	73.000000	8.000000	104.000000	1.000000	-1.000000	0.000000
50%	22519.000000	39.000000	450.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	33879.500000	48.000000	1431.000000	21.000000	318.500000	3.000000	-1.000000	0.000000
max	45211.000000	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000



Univariate Analysis

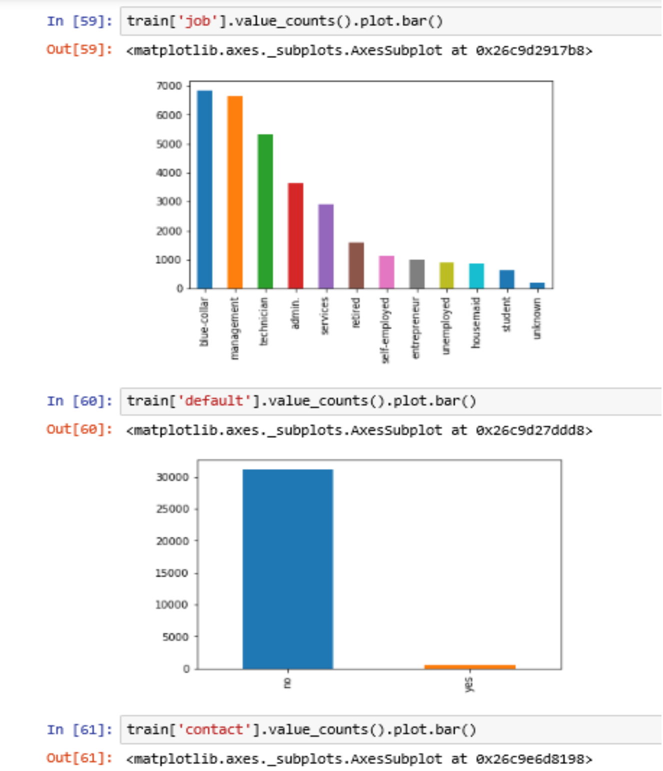
```
In [54]: train['subscribed'].value_counts()

Out[54]: no     27932
         yes     3715
         Name: subscribed, dtype: int64

In [55]: train['subscribed'].value_counts().plot.bar()

Out[55]: <matplotlib.axes._subplots.AxesSubplot at 0x26c9e8a2ef0>
```

```
In [56]: train['subscribed'].value_counts()/len(train['subscribed'])

Out[56]: no     0.882611
         yes    0.117389
         Name: subscribed, dtype: float64
```



```
In [57]: sn.distplot(train["age"])

Out[57]: <matplotlib.axes._subplots.AxesSubplot at 0x26c9ccd0fd0>
```

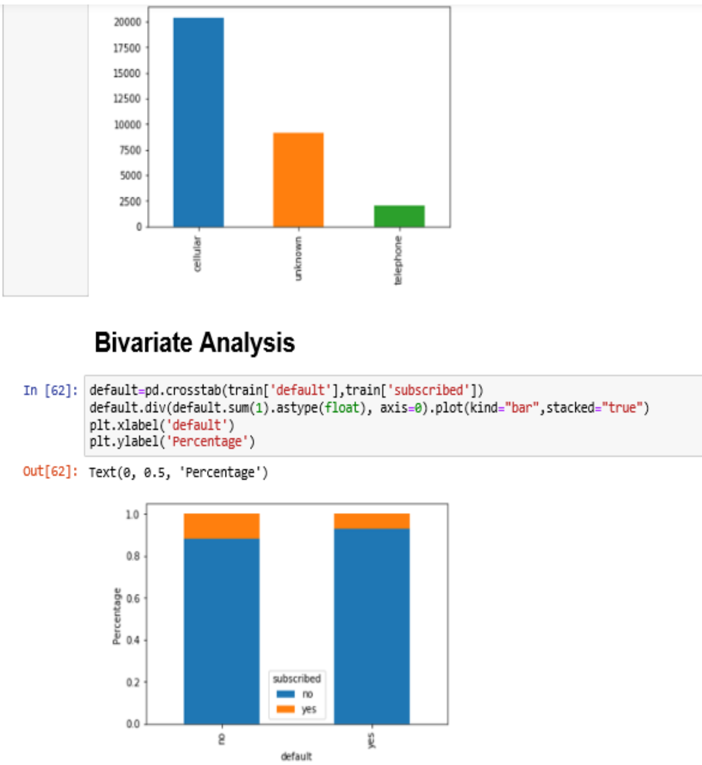
```
In [58]: train['job'].value_counts()

Out[58]: blue-collar      6842
         management       6639
         technician       5307
         admin.           3631
         services         2903
         retired          1574
         self-employed    1123
         entrepreneur     1008
         unemployed        905
         housemaid         874
         student           635
         unknown           206
         Name: job, dtype: int64
```



```
In [59]: train['job'].value_counts().plot.bar()

Out[59]: <matplotlib.axes._subplots.AxesSubplot at 0x26c9d2917b8>
```

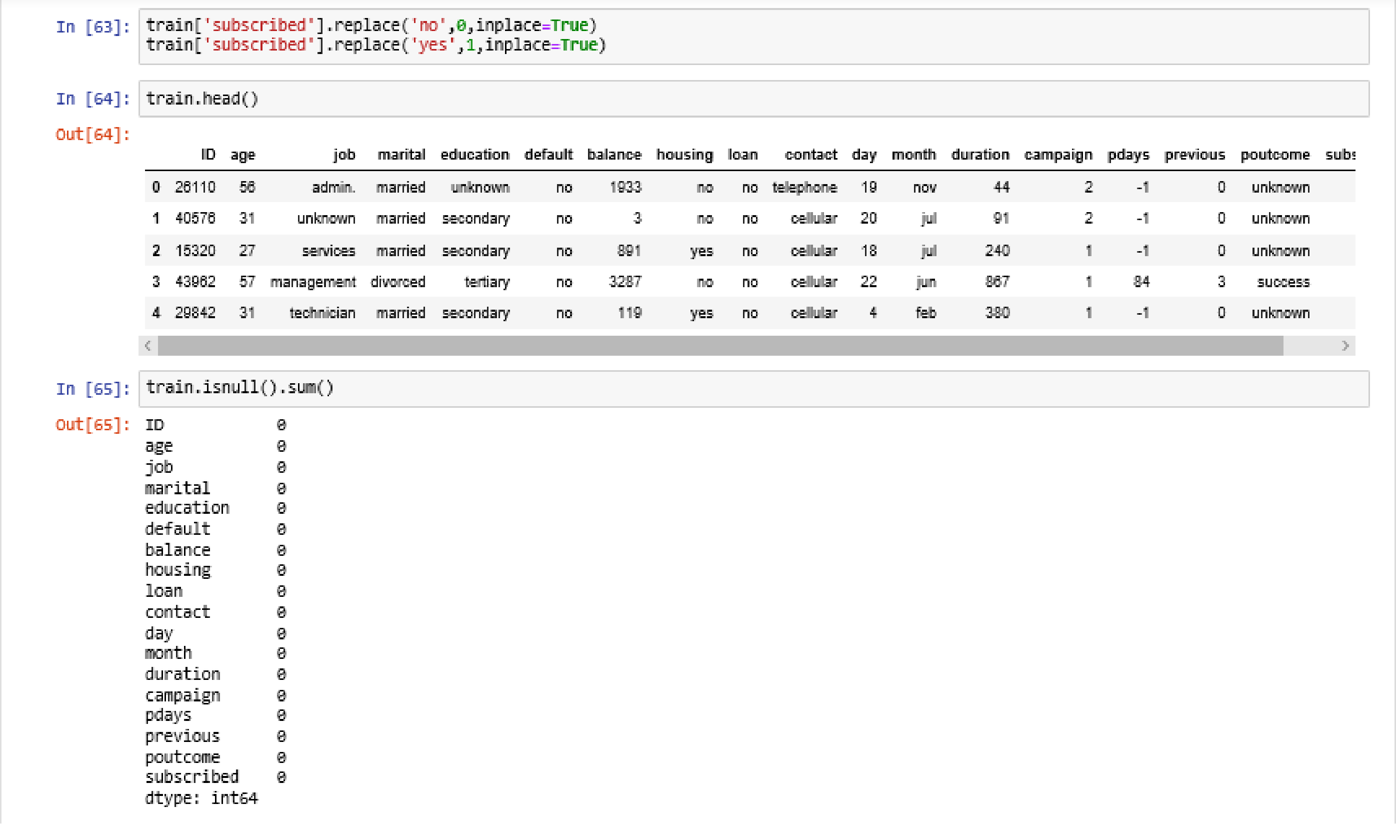
```
In [60]: train['default'].value_counts().plot.bar()

Out[60]: <matplotlib.axes._subplots.AxesSubplot at 0x26c9d27ddd8>
```

```
In [61]: train['contact'].value_counts().plot.bar()

Out[61]: <matplotlib.axes._subplots.AxesSubplot at 0x26c9e6d8198>
```



Bivariate Analysis

```
In [62]: default=pd.crosstab(train['default'],train['subscribed'])
         default.div(default.sum(1).astype(float), axis=0).plot(kind="bar",stacked="true")
         plt.xlabel('default')
         plt.ylabel('Percentage')

Out[62]: Text(0, 0.5, 'Percentage')
```



In [63]:
```
train['subscribed'].replace('no',0,inplace=True)
train['subscribed'].replace('yes',1,inplace=True)
```

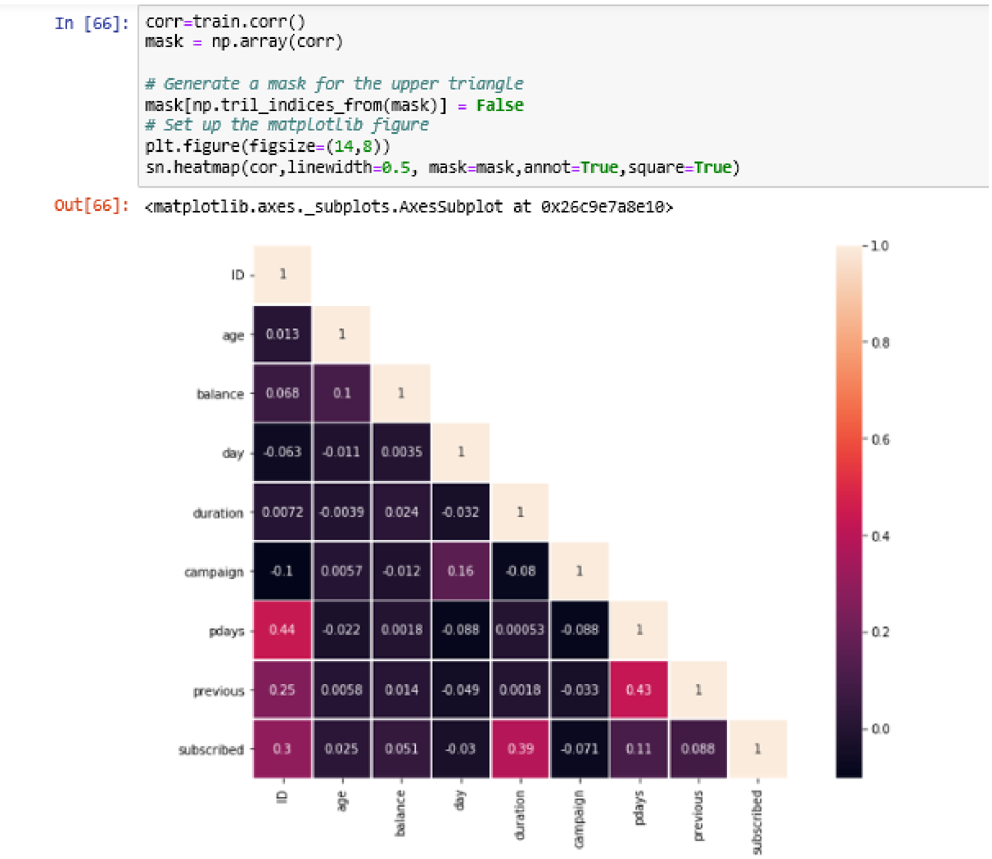
In [64]:
```
train.head()
```

Out[64]:

	ID	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	subs
0	26110	56	admin.	married	unknown	no	1933	no	no	telephone	19	nov	44	2	-1	0	unknown	
1	40576	31	unknown	married	secondary	no	3	no	no	cellular	20	jul	91	2	-1	0	unknown	
2	15320	27	services	married	secondary	no	891	yes	no	cellular	18	jul	240	1	-1	0	unknown	
3	43962	57	management	divorced	tertiary	no	3287	no	no	cellular	22	jun	867	1	84	3	success	
4	29842	31	technician	married	secondary	no	119	yes	no	cellular	4	feb	380	1	-1	0	unknown	

In [65]:
```
train.isnull().sum()
```

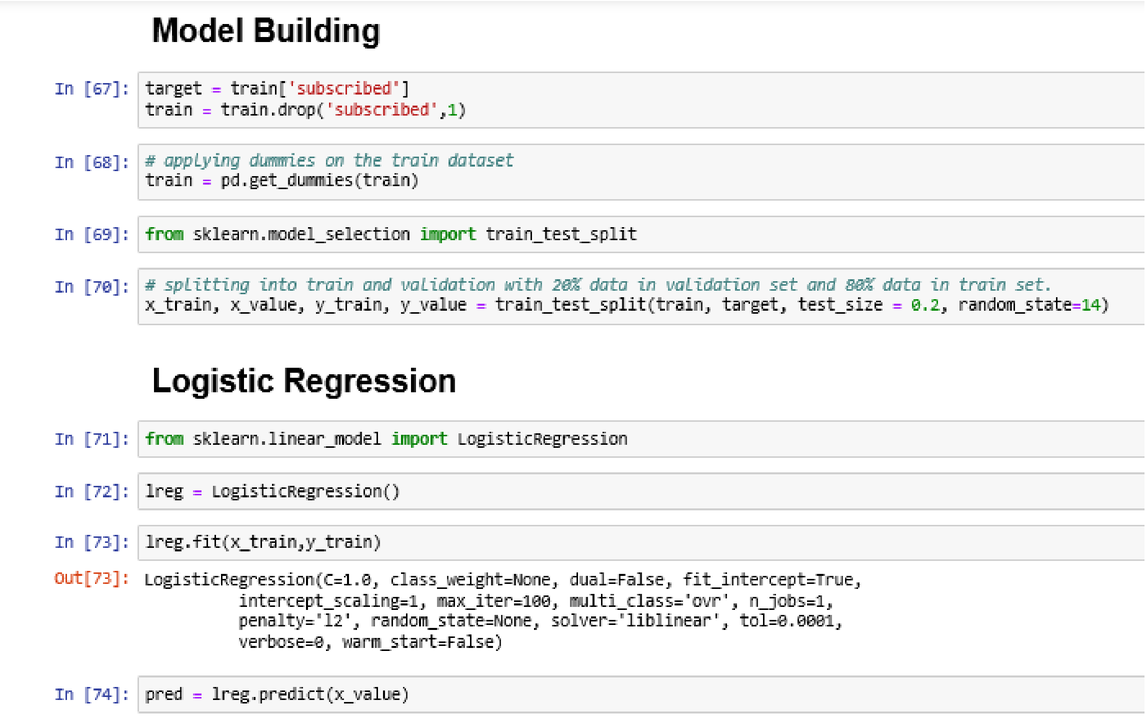
Out[65]:
```
ID            0
age           0
job           0
marital       0
education     0
default       0
balance       0
housing       0
loan          0
contact       0
day           0
month         0
duration      0
campaign      0
pdays         0
previous      0
poutcome      0
subscribed    0
dtype: int64
```



In [66]:
```
corr=train.corr()
mask = np.array(corr)

# Generate a mask for the upper triangle
mask[np.tril_indices_from(mask)] = False
# Set up the matplotlib figure
plt.figure(figsize=(14,8))
sn.heatmap(cor,linewidth=0.5, mask=mask,annot=True,square=True)
```

Out[66]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x26c9e7a8e10>
```

ID 1
age 0.013 1
balance 0.068 0.1 1
day -0.063 -0.011 0.0035 1
duration 0.0072 -0.0039 0.024 -0.032 1
campaign -0.1 0.0057 -0.012 0.16 -0.08 1
pdays 0.44 -0.022 0.0018 -0.088 0.00053 -0.088 1
previous 0.25 0.0058 0.014 -0.049 0.0018 -0.033 0.43 1
subscribed 0.3 0.025 0.051 -0.03 0.39 -0.071 0.11 0.088 1
ID
age
balance
day
duration
campaign
pdays
previous
subscribed
1.0
0.8
0.6
0.4
0.2
0.0

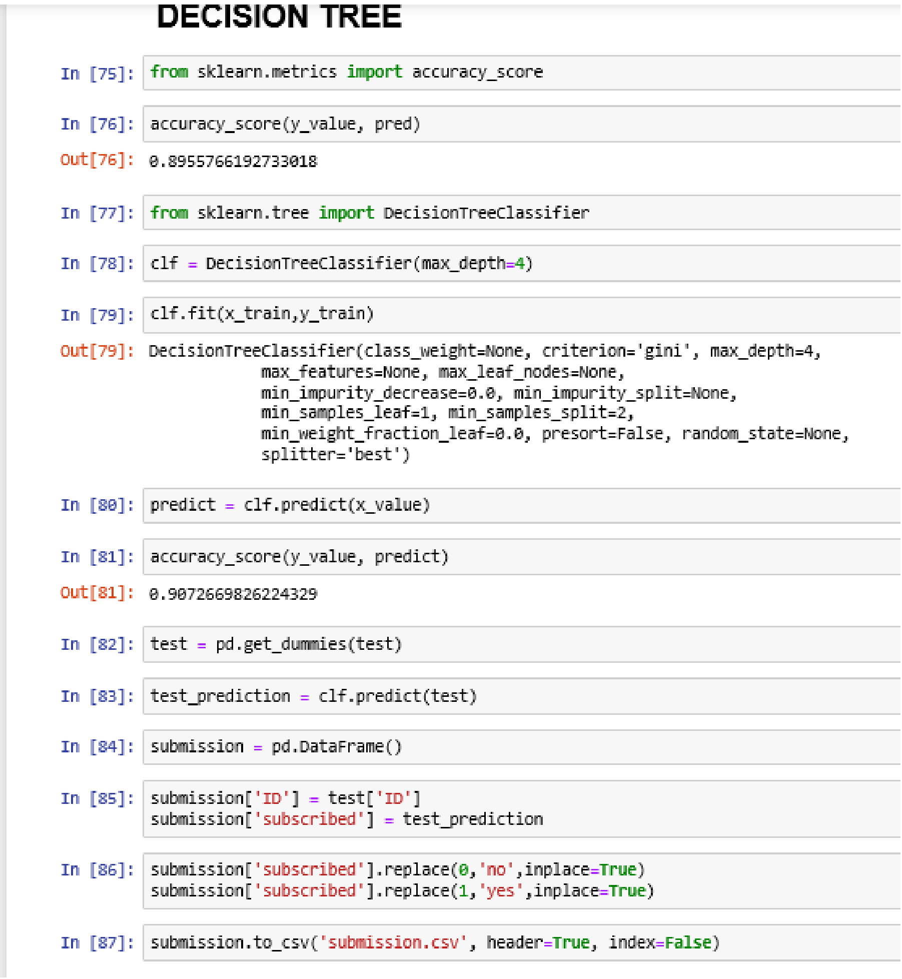


Model Building

```
In [67]: target = train['subscribed']
         train = train.drop('subscribed',1)

In [68]: # applying dummies on the train dataset
         train = pd.get_dummies(train)

In [69]: from sklearn.model_selection import train_test_split

In [70]: # splitting into train and validation with 20% data in validation set and 80% data in train set.
         x_train, x_value, y_train, y_value = train_test_split(train, target, test_size = 0.2, random_state=14)
```

Logistic Regression

```
In [71]: from sklearn.linear_model import LogisticRegression

In [72]: lreg = LogisticRegression()

In [73]: lreg.fit(x_train,y_train)

Out[73]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
                   penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
                   verbose=0, warm_start=False)

In [74]: pred = lreg.predict(x_value)
```



DECISION TREE

```
In [75]: from sklearn.metrics import accuracy_score

In [76]: accuracy_score(y_value, pred)

Out[76]: 0.8955766192733018

In [77]: from sklearn.tree import DecisionTreeClassifier

In [78]: clf = DecisionTreeClassifier(max_depth=4)

In [79]: clf.fit(x_train,y_train)

Out[79]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=4,
                     max_features=None, max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                     splitter='best')

In [80]: predict = clf.predict(x_value)

In [81]: accuracy_score(y_value, predict)

Out[81]: 0.9072669826224329

In [82]: test = pd.get_dummies(test)

In [83]: test_prediction = clf.predict(test)

In [84]: submission = pd.DataFrame()

In [85]: submission['ID'] = test['ID']
         submission['subscribed'] = test_prediction

In [86]: submission['subscribed'].replace(0,'no',inplace=True)
         submission['subscribed'].replace(1,'yes',inplace=True)

In [87]: submission.to_csv('submission.csv', header=True, index=False)
```

Reason for choosing data science

Data Science has become a revolutionary technology that everyone seems to talk about. Hailed as the 'sexiest job of the 21st century'. Data Science is a buzzword with very few people knowing about the technology in its true sense.

While many people wish to become Data Scientists, it is essential to weigh the pros and cons of data science and give out a real picture. In this article, we will discuss these points in detail and provide you with the necessary insights about Data Science.

Advantages: -

1. It's in Demand
2. Abundance of Positions
3. A Highly Paid Career
4. Data Science is Versatile

Disadvantages: -

1. Mastering Data Science is near to impossible
2. A large Amount of Domain Knowledge Required
3. Arbitrary Data May Yield Unexpected Results
4. The problem of Data Privacy

Learning Outcome

After completing the training, I am able to:

- Develop relevant programming abilities.
- Demonstrate proficiency with statistical analysis of data.
- Develop the skill to build and assess data-based model.
- Execute statistical analysis with professional statistical software.
- Demonstrate skill in data management.
- Apply data science concepts and methods to solve problem in real-world contexts and will communicate these solutions effectively.

Gantt Chart

Table	Week 1	Week 2	Week 3	Week 4
Introduction to Data science	■			
Python for data science	■			
Operations of python		■		
Functions in python		■	■	
Containers of python			■	
Statistics of Python				■

BIBLOGRAPHY

- Google
- Internshala
- Wikipedia
- Geeks for Geeks