INTERNSHIP REPORT

(March 2022-August 2022)

Internship report submitted in partial fulfillment of the requirement for the

degree of Bachelor of Technology

in

Computer Science and Engineering

By

Parthh Dikshit

181233

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology,

Waknaghat, 173234,

Himachal Pradesh, INDIA

DECLARATION

I hereby declare that this submission is my own work carried out at Cognizant Technology Solutions India Private Limited– A Consultancy Firm, Noida from March 2022 to August 2022 and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma from a university or other institute of higher learning, except where due acknowledgment has been made in the text.

Submitted by:
Parthh Dikshit
181233
Computer Science & Engineering and Information Technology Department
Jaypee University of Information Technology

ACKNOWLEDGEMENT

I would like to thank Cognizant Technology Solutions India Private Limited for allowing me to do an internship within the organization.

I also would like to thank all the people that worked along with me at Cognizant Technology Solutions India Private Limited with their patience and openness they created an enjoyable working environment.

I also would like to thank Ms Deeksha Sharma for mentoring me throughout my internship.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals.

I am highly indebted to Mr. Pankaj Kumar, Training & Placement Coordinator of our college for the facilities provided to accomplish this internship. I would also like to thank the Head of our Department Dr.Vivek Kumar Sehgal and the faculty for teaching us the skills required for this internship.

Finally, I must acknowledge the constant support and patients of my parents.

Parthh Dikshit
181233
Jaypee University of Information Technology

TABLE OF CONTENT

LIST OF FIGURES

Chapter 01:INTRODUCTION

1.1 Introduction

An internship is a professional teachable moment that gives students real-world experience in their field of study or professional goals. An internship enables students to learn new skills while exploring and enhancing their careers. This report is a description of my ongoing internship at Cognizant Technology Solutions India Private Limited- Noida. This internship report details the activities that helped me achieve a bunch of my stated objectives. I was assigned the profile of "Azure Data Engineer" for my internship. We were asked to learn about the languages in the first two months of the internship which was followed by one month of self paced learning of Azure cloud platform which is still in progress and in between there are sessions conducted by our technical trainer Ms Deeksha Sharma .After the completion of language learning, I was assigned a hands-on project which is called Data Analysis of Titanic Dataset.

1.2 Job Description

In a variety of circumstances, data engineers create systems that gather, process, and convert raw data into meaningful information for data scientists and business analysts to understand. Their main objective is to make data accessible to enterprises so that they may evaluate and enhance their effectiveness.

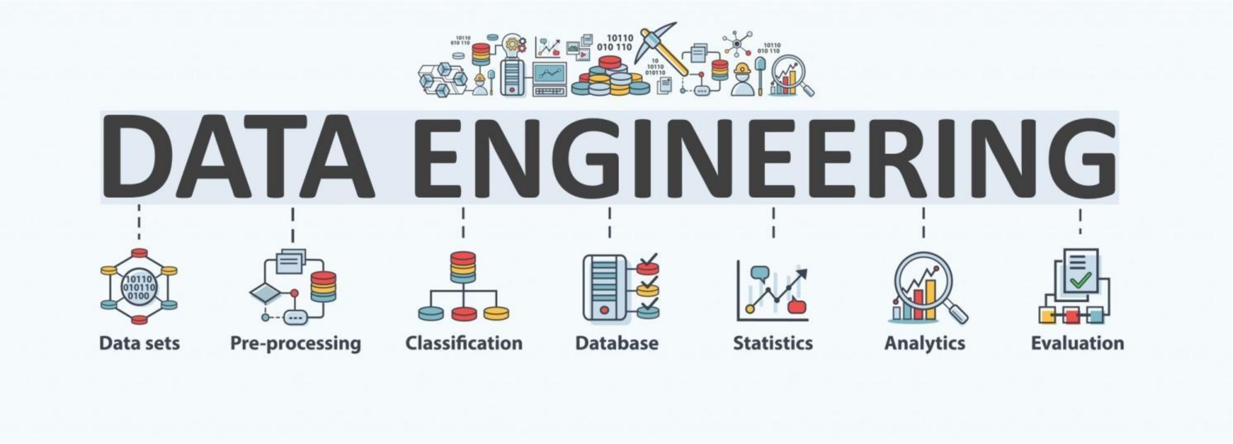


Fig 1. Objectives of Data Engineering

The Azure Data Engineer is responsible to utilize the services from azure cloud and perform data analysis using Azure services like Azure Data bricks and find meaningful insights from raw data. When working with data, you may encounter the following tasks:

- Gather data that is relevant to your industry.
- Make algorithms to turn data into useful information.
- Architectures for database pipelines must be created, tested, and maintained.
- Work with stakeholders to learn about the company's objectives.
- Create new validation rules and analysis software.
- Ascertain compliance with data compliance and management policies.

1.3 Objectives

The objective of the internship is to learn about Data Analysis and services in azure platform like Azure Data Factory , Azure Data Lake storage, Azure Databricks etc. How to upload and ustilise different languages like python and Sql to manipulate and reform raw data and learn to convert it into processed data. Perform various visualization techniques using library like plotly, matplotlib etc to recognize patterns and hidden meanings.

1.4 Internship Schedule

The internship plan was as follows:

➢ March 2022

This month, I revised Python basics and completed assignments that were allotted to our cohort by our trainer which included self-study of the language from Udemy course and hands-on practice questions which are of easy level. We are also allotted courses regarding data analysis on Udemy itself and are expected to learn various libraries like pandas , numpy etc by ourselves and clear out doubts (if any) in the technical connect.

In parallel, behavioral training sessions were also held 3 days a week to help us groom ourselves and to introduce us to the corporate culture and make us familiar with the expectations that the organization has for us.

- April 2022

After the completion of back Python training we were allotted a project regarding data analysis called Titanic data analysis which was given us to check our capabilities and grasp on the concepts which we learnt in previous month.

With that we were allotted a new trainer regarding Sql training. We were taught MySql and PostgreSQL. At the end of this month an interview was conducted by our sql trainer and bh trainer in the basis of which we were evaluated and marked.

- May 2022

Here we were provided with Azure cloud subscription and were allotted beginner level course on azure cloud namely Dp-900 where we were made familiar with the platform. Everyday trainer sessions are held were we are given detailed insights on various topics that we will be useful in our full time roles, like deployment of virtual machines , creation of storage account etc.

- June 2022

We are planning to work further on Azure platform ie DP 203 and learn databricks with pyspark.

Chapter 02: COMPANY DESCRIPTION

2.1 Cognizant

Cognizant is a multinational consultancy firm based in the United States. The company's headquarters are in Teaneck, New Jersey. Cognizant is a NASDAQ-100 company that trades under the symbol CTSH. It began as a Dun & Bradstreet in-house technology unit in 1994 and began its services in 1996.

In 1998, the company became public after a series of business reorganisations.

Cognizant experienced rapid growth in the 2000s and was named to the Fortune 500 , placed 185th.

Fig 2. About Cognizant

2.1.1 Our Services

- **Cloud Enablement** - Wherever you are on your cloud transformation journey, Cognizant is there to support you. We collaborate with you to build the best approach for your organisation so you can reap the full benefits and value of cloud computing.Our vision, industry experience, and client-centric approach deliver end-to-end cloud services and innovative solutions that result in short-term success, measurable business outcomes, and happy customers. Learn more about how Cognizant can help you accelerate your cloud journey.

- **Application modernization services**-Application modernization services cover the migration of legacy applications or platforms to new applications or platforms, as well as the integration of new functionality to give the business the most up-to-date features.
- **Data, Analytics and AI-** Artificial intelligence (AI) analytics is a subset of business intelligence that employs machine learning techniques to uncover new patterns and relationships in data. In practise, AI analytics is the process of automating much of the work that would normally be performed by a data analyst.

2.2 History of Cognizant

Cognizant was founded as Dun & Bradstreet Satyam Software (DBSS), with Srini Raju as the original CEO and MD, as Dun & Bradstreet's in-house technology arm focused on completing huge IT projects for Dun & Bradstreet enterprises. The company began targeting customers outside of Dun & Bradstreet in 1996.

Dun & Bradstreet split up some of its companies in 1996, including Erisco, IMS International, Nielsen Media Research, Pilot Software, Strategic Technologies, and DBSS,Cognizant Corporation, situated in Chennai, India, was formed. After a few months, in 1997, DBSS changed its name to Cognizant Software Solutions. In July 1997, Dun & Bradstreet spent $3.4 million for Satyam's 24 percent stake in DBSS. In March 1998, the company's headquarters was relocated to the The US, and Kumar Mahadeva was named CEO. As part of Cognizant Corporation, the company focuses on Y2K-related work and web development.

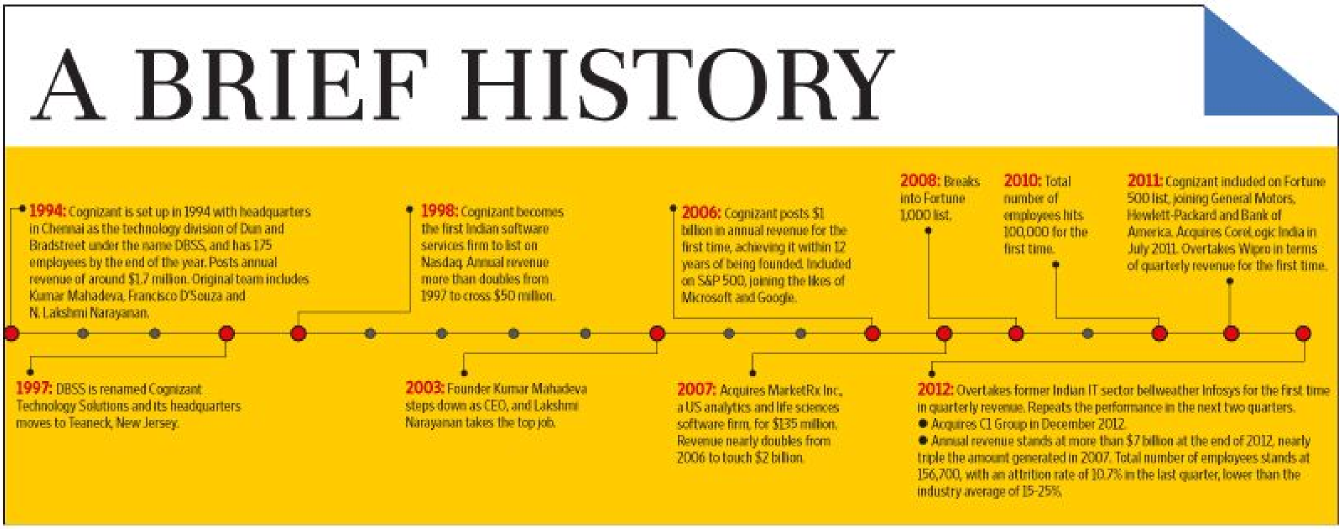
In 1998, the parent company, Cognizant Corporation, split into two companies: IMS Health and Nielsen Media Research. Following the restructure, Cognizant Technology Solutions became a public division of IMS Health. In June 1998, IMS Health began a public sale of Cognizant shares to partially divide the firm.

The firm can raise $34 million, which was less than the underwriters at IMS Health had hoped for. They set aside the funds for debt repayment and office upgrades.

Kumar Mahadeva chose to reduce the corporation's dependency upon Y2K contracts, with Y2K projects providing for only 26% of revenues in the first quarter of 1999, contrasted to 49% in early 1998. Since he believed the $16.6 billion corporate software market was saturated, Kumar Mahadeva avoided large-scale ERP deployment projects. He instead focused on application management, which accounted for 37% of Cognizant's sales in fiscal of 1999. Cognizant's revenues in 2002 were $229 million, with no debt and a total value of $100 million. During the dotcom bust, the company thrived by taking on routine maintenance that big IT services companies didn't desire.

IMS Health bought its entire 56 percent ownership in Cognizant in 2003, and the company implemented a poison pill policy to avoid aggressive takeovers. In 2003, Kumar Mahadeva stepped down as CEO and was succeeded by Lakshmi Narayanan. The company's products portfolio gradually expanded beyond IT to include outsourced (BPO) and consulting firm. Francisco D'Souza succeeded Lakshmi Narayanan in 2006. Cognizant achieved rapid expansion in the 2000s, as evidenced by its ten consecutive appearances on Fortune magazine's "100 Fastest-Growing Companies" list from 2003 to 2012.

Cognizant made its biggest acquisition in September 2014, when it paid $2.7 billion for healthcare IT services business TriZetto Corp. In pre-market trading, shares of Cognizant surged about 3%..



A BRIEF HISTORY

1994: Cognizant is set up in 1994 with headquarters in Chennai as the technology division of Dun and Bradstreet under the name DBSS, and has 175 employees by the end of the year. Posts annual revenue of around $1.7 million. Original team includes Kumar Mahadeva, Francisco D'Souza and N. Lakshmi Narayanan.

1997: DBSS is renamed Cognizant Technology Solutions and its headquarters moves to Teaneck, New Jersey.

1998: Cognizant becomes the first Indian software services firm to list on Nasdaq. Annual revenue more than doubles from 1997 to cross $50 million.

2003: Founder Kumar Mahadeva steps down as CEO, and Lakshmi Narayanan takes the top job.

2006: Cognizant posts $1 billion in annual revenue for the first time, achieving it within 12 years of being founded. Included on S&P 500, joining the likes of Microsoft and Google.

2007: Acquires MarketRx Inc., a US analytics and life sciences software firm, for $135 million. Revenue nearly doubles from 2006 to touch $2 billion.

2008: Breaks into Fortune 1,000 list.

2010: Total number of employees hits 100,000 for the first time.

2011: Cognizant included on Fortune 500 list, joining General Motors, Hewlett-Packard and Bank of America. Acquires CoreLogic India in July 2011. Overtakes Wipro in terms of quarterly revenue for the first time.

2012: Overtakes former Indian IT sector bellwether Infosys for the first time in quarterly revenue. Repeats the performance in the next two quarters.

- Acquires C1 Group in December 2012.
- Annual revenue stands at more than $7 billion at the end of 2012, nearly triple the amount generated in 2007. Total number of employees stands at 156,700, with an attrition rate of 10.7% in the last quarter, lower than the industry average of 15-25%.

Fig. 3 A Brief History

The company struck a multimillion-dollar arrangement with Escorts Group in India on June 24, 2015, to assist Escorts' enterprises with digital transformation and modernization across all business areas.

It teamed with Singapore-based market chain NTUC FairPrice on 30 June 2015 to digitally modernise NTUC's operations in order to improve personalised and consistent customer experience across numerous channels.

Cognizant sold its Oy Samlink acquisition to Kyndryl in January 2022.

Chapter 03: TOOLS AND TECHNOLOGIES USED

3.1 Python

Python is a high-level programming language that is dynamically semantic, interpreted, and object-oriented. Because of its elevated built-in data structures, dynamic typing, and dynamic binding, it's perfect for Faster Development and as a scripting for integrating existing components. Python's compact, simple syntax encourages readability, lowering programme maintenance costs. Python supports modules and packages, which help with programme versatility and code reuse. The Python interpreter as well as its extensive standard library are available in source and binary formats for all major systems.

- Pandas

  Pandas is an open-source library that makes working with relational or tagged data simple and intuitive. It includes a variety of data formats and methods for working with numerical data and time series. The NumPy Python library provides the foundation for this library. Pandas is fast and gives high performance and productivity to its users.

  1. Data management and analysis are simple and straightforward.
  2. Data can be loaded from a variety of file objects.
  3. In both floating - point numbers and non-floating point data, missing data (represented as NaN) is simply handled.
  4. Size mutability: DataFrame columns and higher-dimensional objects can also have their sizes changed.
  5. Connecting and combining data sets
  6. Time-series functionality allows for flexible data set shaping and pivoting.
  7. Procedures for splitting, applying, and combining data sets have been simplified.

- Numpy

  Numpy is a versatile array processing library. It comes with a high-performance multidimensional array object and facilities for manipulating it. For scientific computing, it is the most important Python package.

  In addition to its obvious scientific applications, It can be utilized as a multi-dimensional container of generic data.

- Plotly

  The Python Plotly Toolkit is an open-source library for quickly and easily visualising and interpreting data. Plotly offers a variety of plot types. So, why should you use Plotly over other visualisation tools or libraries? Here's the solution —

  Plotly offers a hover function that allows us to spot outliers or abnormalities in a huge number of data points.

  It is visually appealing and appealing to a wide spectrum of audiences.

  It enables us to personalise our charts in an unlimited variety of ways, making them more engaging and comprehensible to others.

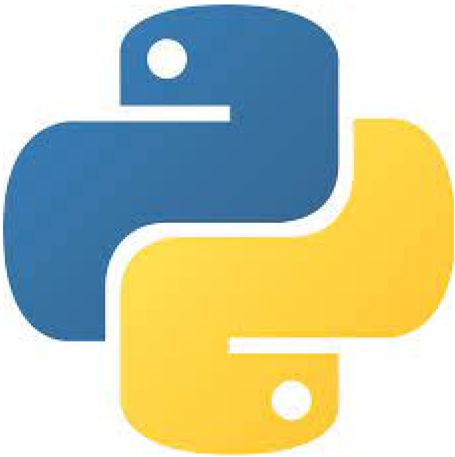


Fig. 4 Python Logo

3.2 MySQL

MySQL is an open-source relational database management system that is used extensively in online applications. Because data is stored and transmitted over the internet, databases and related tables are a key component of many websites and applications. Even the most popular social networking sites, such as Facebook, Twitter, and Google, rely on MySQL data, which is specifically created and optimised for this purpose. For these reasons, the MySQL server has become the standard for web applications.

MySQL is a database management system that allows you to query, sort, filter, group, alter, and join tables. Let's look at some of MySQL's benefits before studying the most regularly used queries.

Advantages of MySQL :

- Database with great performance.
- It's simple to set up, manage, and administrate.
- Database integrity is maintained and is easily accessible.
- Scalability, usability, and dependability are all advantages.
- Hardware that is inexpensive.

Fig.5 MySQL Logo

3.3 PostgreSql

PostgreSQL is an open-source database management system that is one of the most advanced general-purpose object-relational database management systems. Its source code is accessible under the PostgreSQL licence, which is a liberal open source licence. PostgreSQL can be used, modified, and distributed in any form by anyone with the necessary expertise.

Fig. 6 PostgreSQL Logo

3.4 Azure

Azure, like Google Cloud and Aws Web Service, is Microsoft's cloud - based platform (AWS.000). It is a platform that enables us all to utilize Microsoft's resources in general. Setting up a huge server, for example, will require a significant amount of money, time, and physical space. In such cases, Microsoft Azure can help. It will make our work easier by providing virtual machines, rapid data processing, analytical and monitoring tools, and so forth. Azure's pricing is also more straightforward and cost-effective. "Pay As You Go" is a popular expression that suggests you only pay for what you use.

- Services in cloud (IaaS. PaaS, SaaS)

Through an internet connection, a cloud service provider handles your infrastructure—the physical servers, networking, virtualization, and data storage—for you. The user gets access to the infrastructure via an API or dashboard and is effectively renting it. The user is in charge of the operating system, programmes, and middleware, while the provider is in charge of hardware, network, hard disks, storage systems, and server administration.

- Azure Region

Datacenters are placed within a lag perimeter in each Azure region. They're linked by a dedicated low-latency regional network. This architecture ensures that Microsoft azure services in any region provide the highest speed and security possible.

- High Availability

Azure availability zones are geographically distinct regions within each Azure region that are resilient to local failures. Software / hardware failures can occur, as well as natural disasters like earthquake, floods, and fires. The redundancy and logical separation of Azure services allow for failure tolerance. All availability zone-enabled regions must have at least three independent availability zones to provide robustness.

- Resource Group,

A resource group is a container for Azure solutions that houses connected resources. The resource group could contain all of the solution's resources or just the ones you would like to handle as a group.

- Subscription,

Free, pay-as-you-go, and membership offers are the three primary types of subscriptions accessible. Azure pass sponsorship is available.

- Management Group

Management groups are containers that allow you manage numerous subscriptions' access, policies, and compliance. Create those containers to create a hierarchical structure that can be utilised with Microsoft Azure Policy and Azure Role-Based Access Controls.

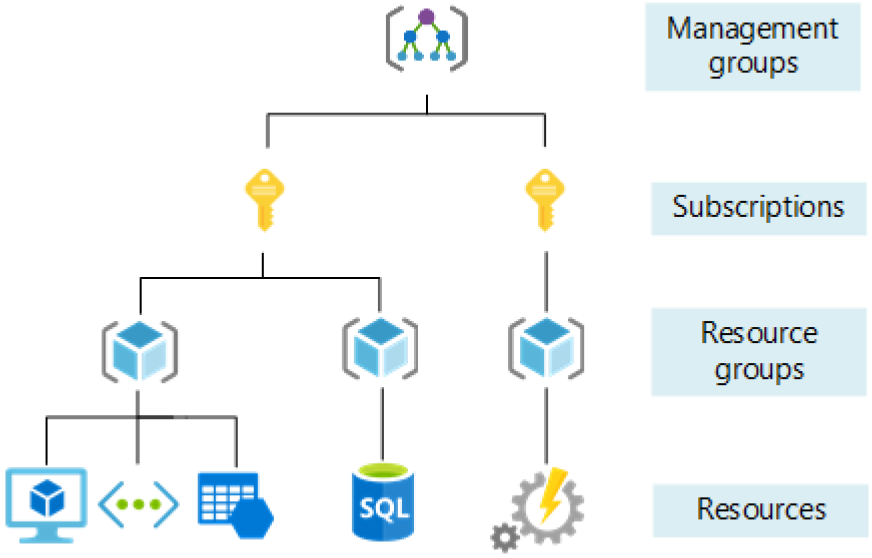


Fig. 7 Azure Hirarchy

- Azure Locks

  You could lock a subscription, resource group, or resource as an administrator to protect other users in your company from deleting or updating key resources by accident. Any permissions that user may have are overridden by the lock. You can choose between CanNotDelete and ReadOnly as the lock level.

- Azure Tags

  Azure tags are name-value pairings that are used in Azure Portal to group resources. You can tag individual resources or the resource group in which they are included.

- Different type of Azure SQL

Have 3 types-

➔ Manage instance- PAAS- not need to specify version etc but need to specif security only, only setup sql server (not os like in Sql VM), have to create virtual network , less administrator effort best for lift and shift ie pick up data and shift on azure

➔ Sql virtual machine - Sql is coming with virtual machine option , (IAAS), choose server and version for sql and windows, best when we have already existing setup of db abd server, easy to migrate to azure thru sql vm

➔ Databases- connect database with server so best for modern cloud applications , PAAS,has 2 types -

  ◆ Elastic pool - elastic space, automatically handle the workload,resource sharing wll happen.

  ◆ Single database

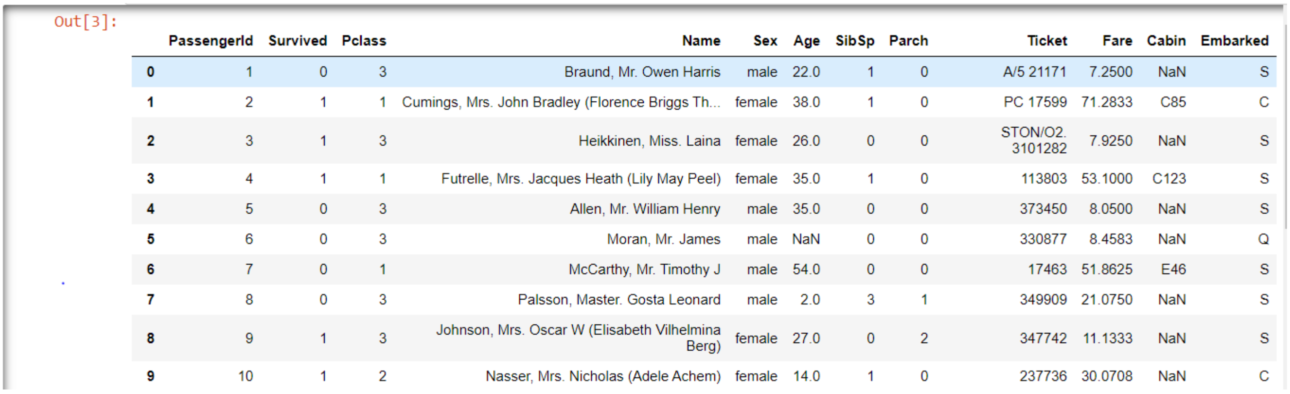
Fig 8 Azure Logo

Chapter 04: Python Hands-On PROJECT

4.1 Description

The hands on project that we were allotted with is called “Exploratory Data Analysis of Titanic Dataset”. The time given us to complete the data analysis was 1 week after which there was a small interview that was conducted by our trainer on the basis of which we were evaluated.

The dataset comprises of 11 columns namely – PassengerId , Survived, Pclass , Name, Sex, Age, SibSp,Parch, Ticket, Fare, Cabin, Embarked. Goal was to clean the data and find meaningful insights from it so that it can be used for a machine learning classification using logistic regression. Motive is to predict if a particular person boarded on titanic will survive or not.



Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

Table 1Titanic Dataset

The libraries used for visualization are matplotlib and seaborn.

4.2 Exploratory Data Analysis

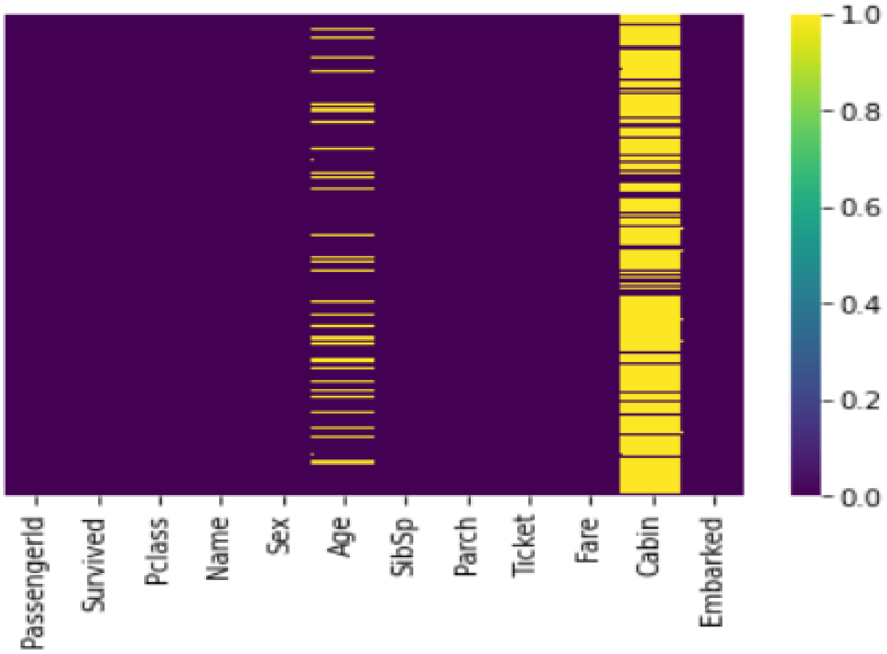
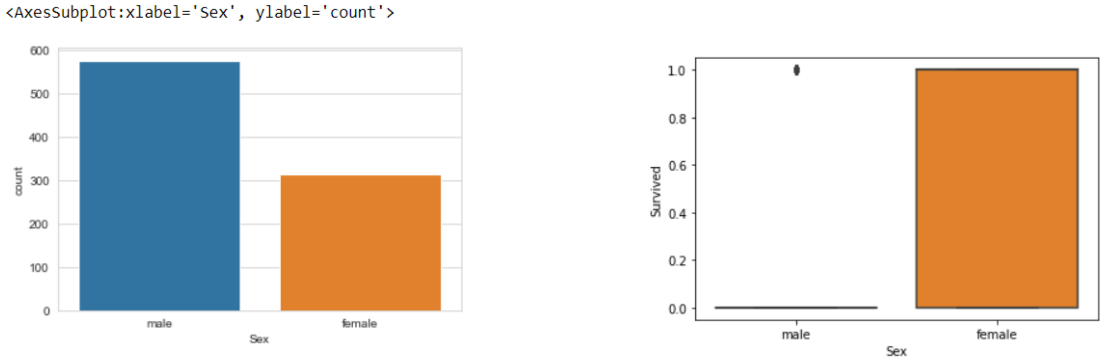


Fig. 9 Heatmap

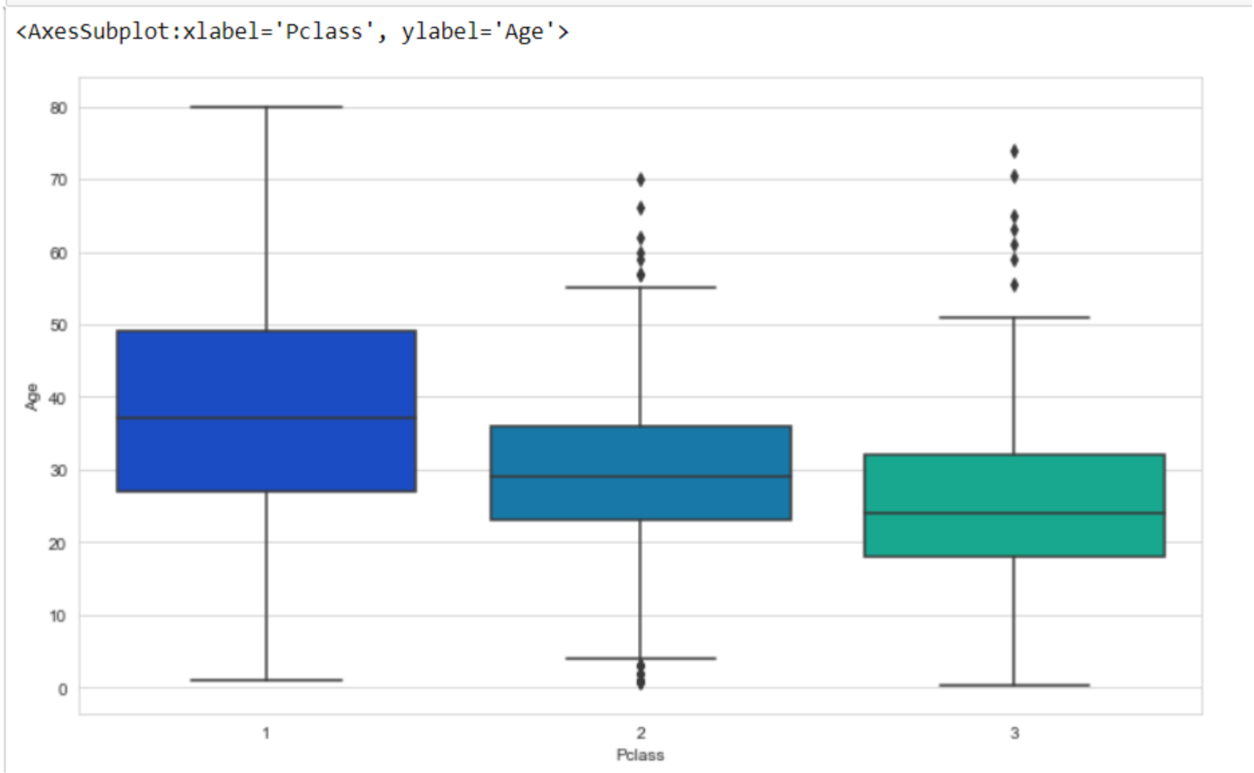
Approximately 20% of Age column is incomplete. The percentage of data lost is likely minimal enough to be replaced reasonably with imputation. Glancing at the Cabin column, it appears that we are lacking far too much information to perform anything useful with it at a basic level. We'll probably remove that or replace it with categorical values of 1 and 0.



```
<AxesSubplot:xlabel='Sex', ylabel='count'>
```

Fig 10 Bar charts

We can very well observe from this bar chart that the majority survivors of the titanic tragedy were women and no of men that survived were negligible. But the population of male passengers was considerably high than that of female passengers. These observation is important for predicting the survival rate of passenger, as if the passenger is female then the probability of survival increases considerably.

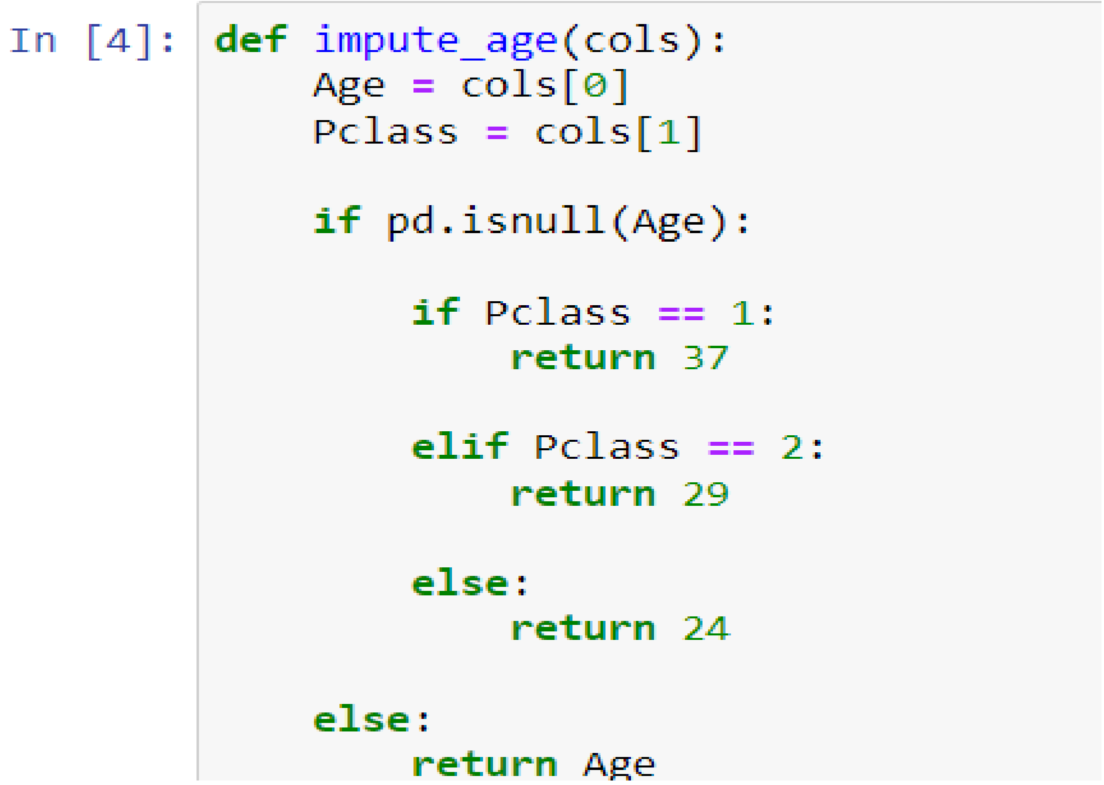


```
<AxesSubplot:xlabel='Pclass', ylabel='Age'>
```

Fig 11 Box Plot

Instead of discarding the missing age data rows, we wish to fill up the gaps. Filling in the average age of all passengers (imputation) is one way to achieve this. We can see that the wealthy passengers in the upper classes are older, which is understandable. We'll infer based on

Pclass for Age using these average age values.



```
In [4]: def impute_age(cols):
            Age = cols[0]
            Pclass = cols[1]

            if pd.isnull(Age):

                if Pclass == 1:
                    return 37

                elif Pclass == 2:
                    return 29

                else:
                    return 24

            else:
                return Age
```

Fig 12 Code for imputation

```
Out[7]: <AxesSubplot:>
```

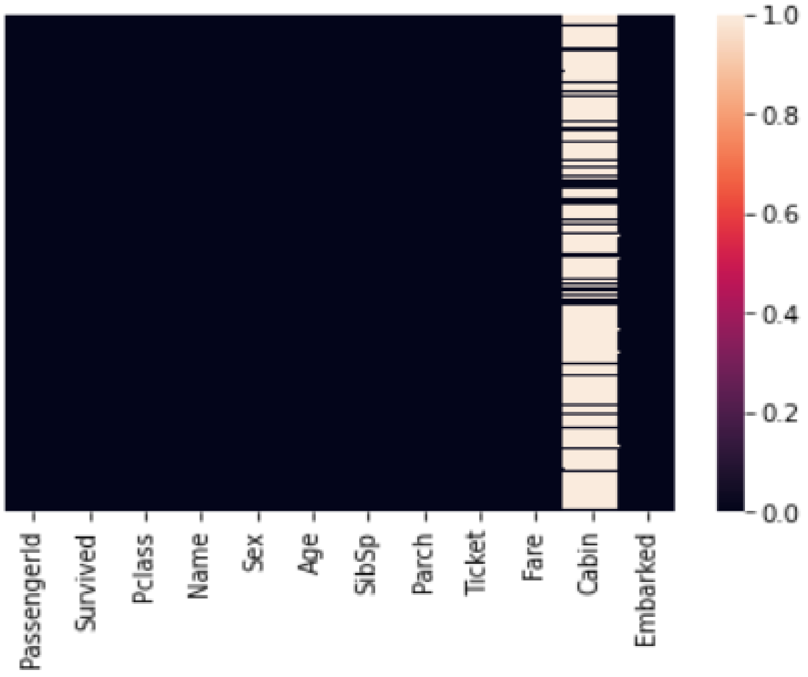
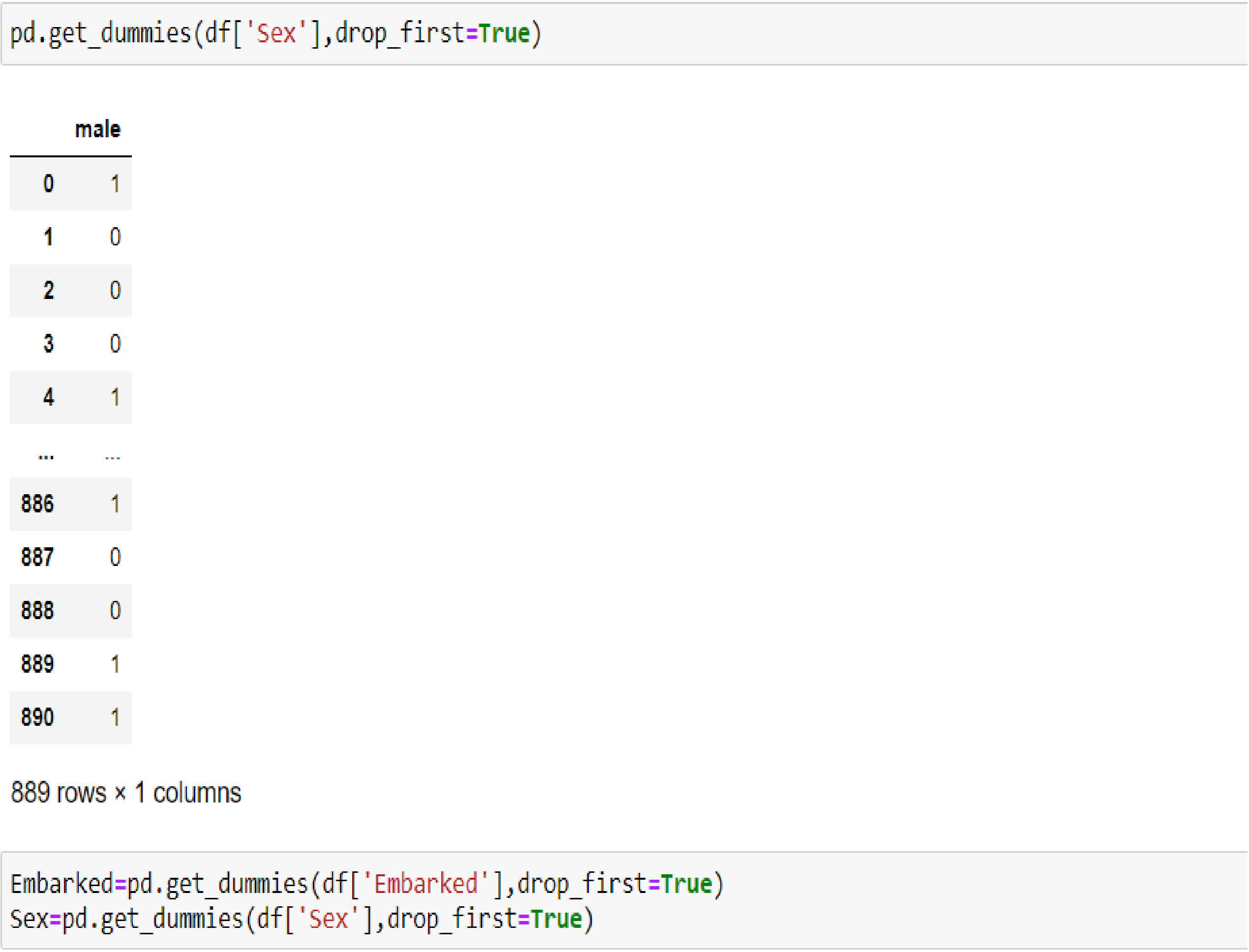


Fig 13 Heatmap

We can observe that by using impute function , the problem of missing values in titanic dataset got solved but the cabin column is still largely empty. And cabin have not correlation with the survival rate so we can drop it.

The sex column in the dataset contains 2 values - male and female and is of object type, so have to be converted to numerical by using the get_dummies method of pandas library so that it can be used in the Machine Learning model as it is a key feature in deciding whether a person will survive or not. Replacing the sex column with male column which contains integer categorical value of 0 and 1 where 0 means female and 1 implies male.Later we drop the features which are not necessary for our model building-'Sex', 'Embarked', 'Name',' Ticket'.



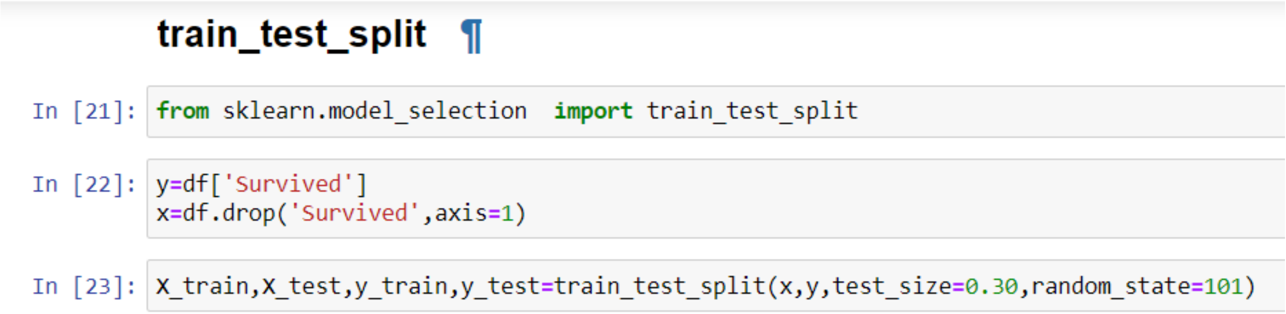
```
pd.get_dummies(df['Sex'],drop_first=True)
```

```
     male
  0     1
  1     0
  2     0
  3     0
  4     1
...   ...
886     1
887     0
888     0
889     1
890     1

889 rows × 1 columns
```

```
Embarked=pd.get_dummies(df['Embarked'],drop_first=True)
Sex=pd.get_dummies(df['Sex'],drop_first=True)
```

Fig 14 Code for conversion of categorical values

4.3 Model Building

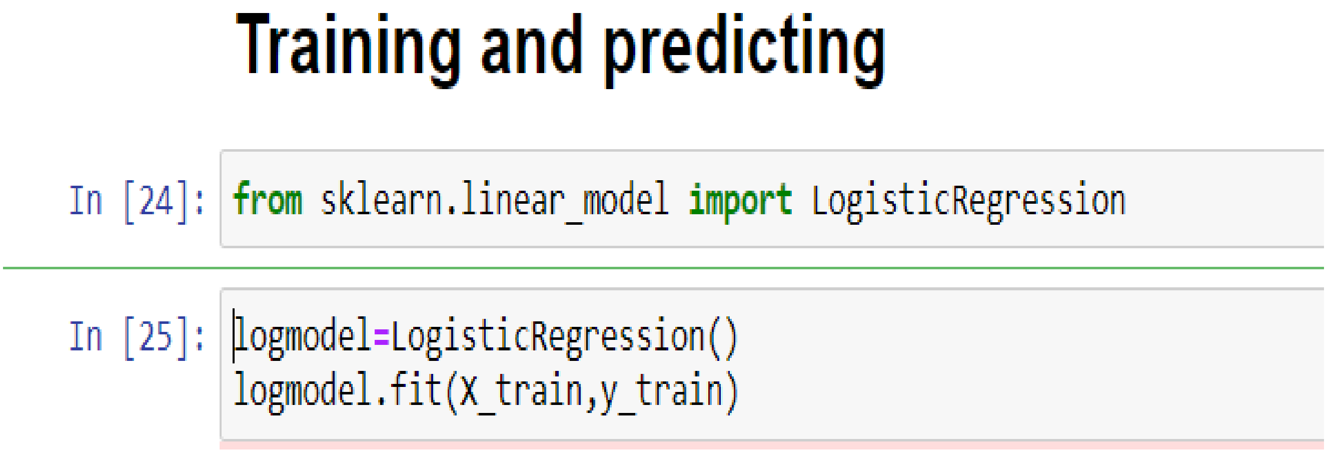


train_test_split ¶

```
In [21]: from sklearn.model_selection  import train_test_split

In [22]: y=df['Survived']
         x=df.drop('Survived',axis=1)

In [23]: X_train,X_test,y_train,y_test=train_test_split(x,y,test_size=0.30,random_state=101)
```

Fig 15 Train Test Split

4.4 Results



Training and predicting

```
In [24]: from sklearn.linear_model import LogisticRegression

In [25]: logmodel=LogisticRegression()
         logmodel.fit(X_train,y_train)
```

Fig 16 Machine Learning Model

When a decision criterion is introduced, logistic regression transforms into a classification procedure. Setting the threshold value is a crucial part of logistic regression, and it is determined by the classification problem.

The precision and recall levels have a significant influence on the threshold value determination. Both precision and recall should ideally be 1, but this is rarely the case.

Logistic regression is classified as follows based on the number of categories:

- Binomial
- Multinomial
- Ordinal

Here we used Binomial Regression because we have two classes that we need to classify on.

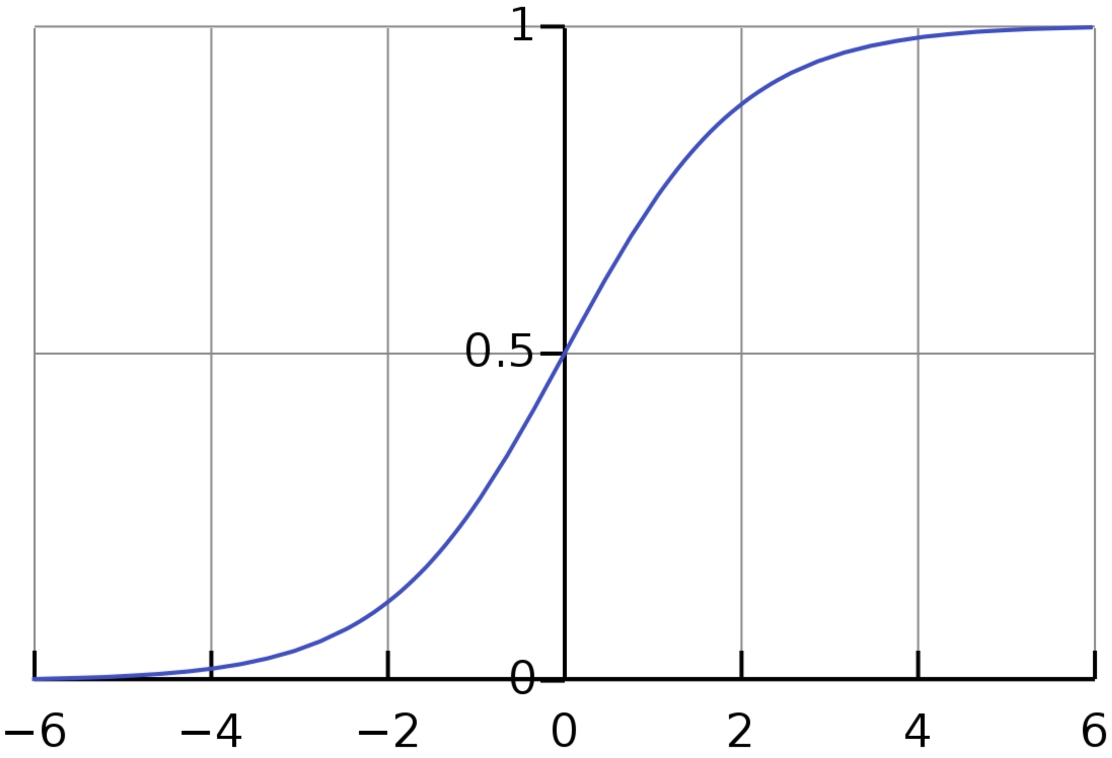
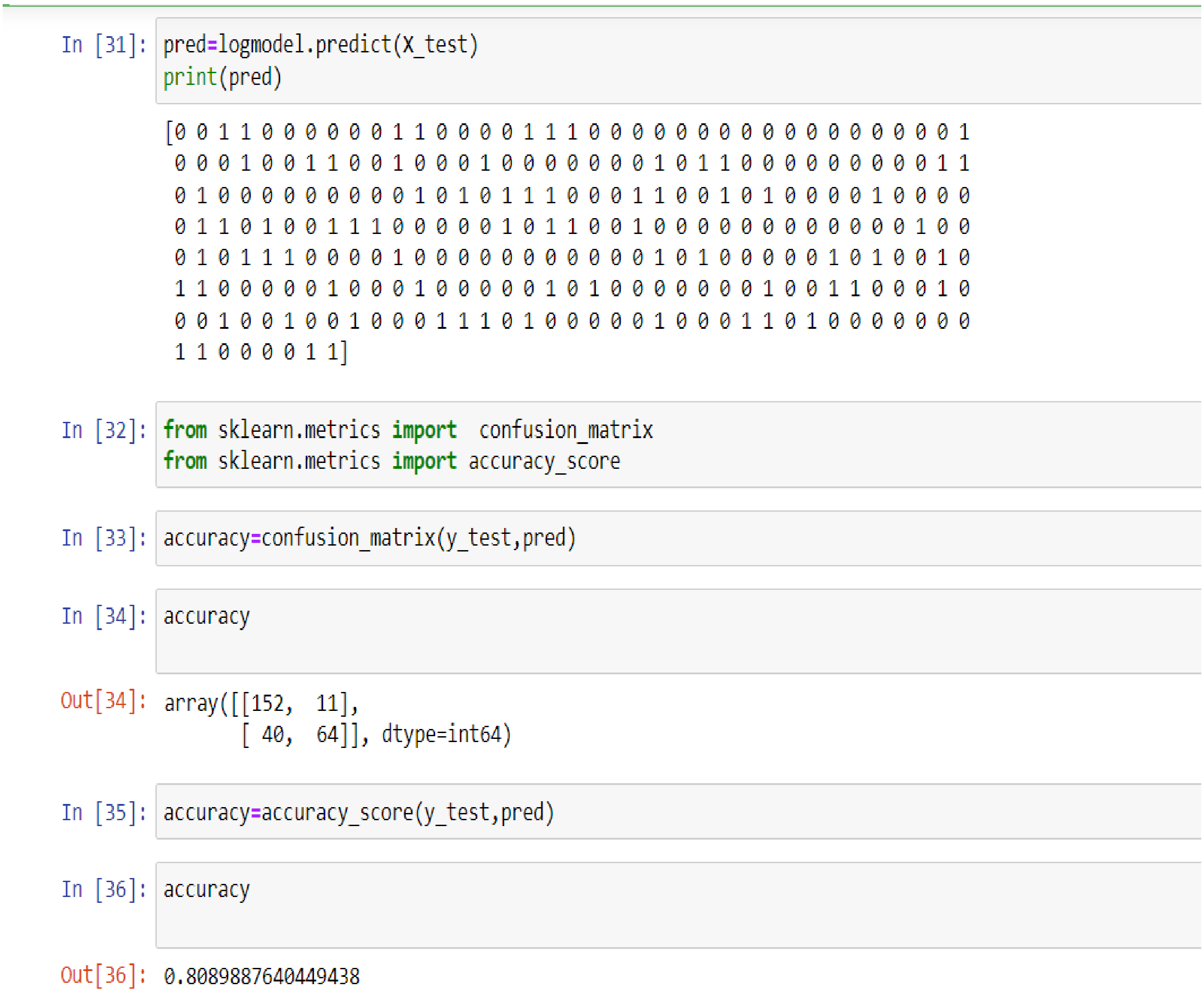


Fig 17 Sigmoid Function



```
In [31]: pred=logmodel.predict(X_test)
         print(pred)

         [0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
          0 0 0 1 0 0 1 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 1 1
          0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 0 0 1 1 0 0 1 0 1 0 0 0 0 1 0 0 0 0
          0 1 1 0 1 0 0 1 1 1 0 0 0 0 0 1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
          0 1 0 1 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 0 1 0
          1 1 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 1 0
          0 0 1 0 0 1 0 0 1 0 0 0 1 1 1 0 1 0 0 0 0 0 1 0 0 0 1 1 0 1 0 0 0 0 0 0 0
          1 1 0 0 0 0 1 1]

In [32]: from sklearn.metrics import  confusion_matrix
         from sklearn.metrics import accuracy_score

In [33]: accuracy=confusion_matrix(y_test,pred)

In [34]: accuracy

Out[34]: array([[152,  11],
                [ 40,  64]], dtype=int64)

In [35]: accuracy=accuracy_score(y_test,pred)

In [36]: accuracy

Out[36]: 0.8089887640449438
```

Fig 18 Results of Machine Learning model

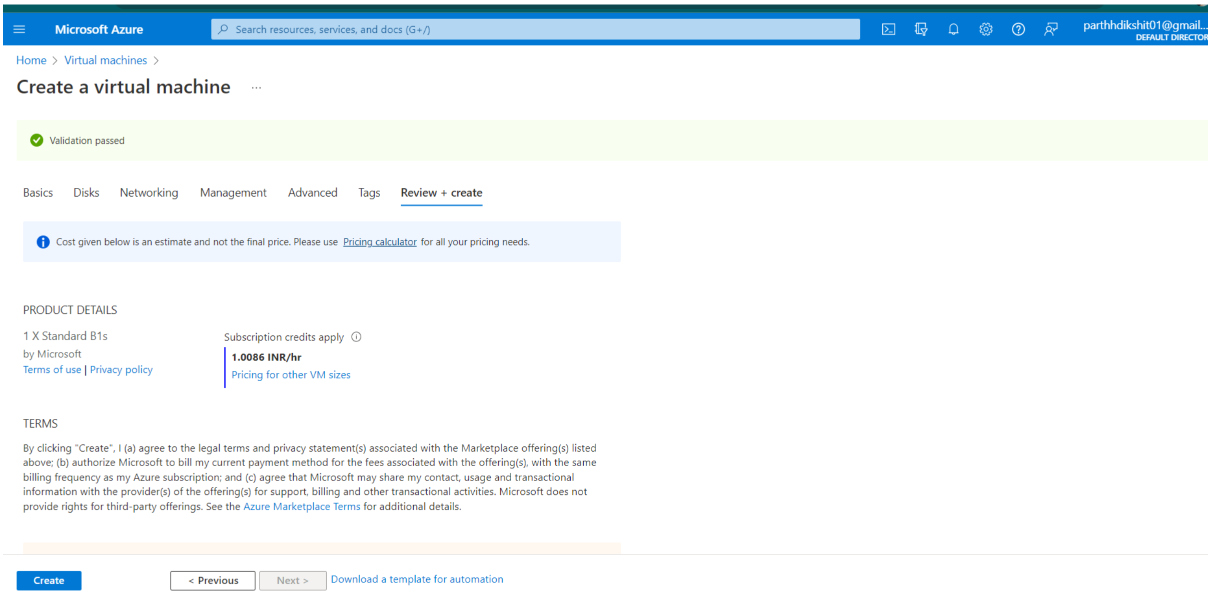
We can observe that the accuracy score comes out to be 80%. Hence we can conclude that out model is neither overfit nor underfit and is working perfectly fine.

4.5: Azure Training

4.5.1 Description

There are various components in azure that we were introduced and the python training and the sql training were the pre-requisite of using azure cloud. We were first introduced to the azure hierarchy which included talks about subscription, resource group and resource. Later we were given practical knowledge by giving us a live description of azure platform and its various functions. Assignments were alloted to us in which we were asked to create virtual machines storage accounts and resource groups.

4.5.2 Screenshots



Create a virtual machine

Validation passed

Cost given below is an estimate and not the final price. Please use Pricing calculator for all your pricing needs.

PRODUCT DETAILS

1 X Standard B1s
by Microsoft
Terms of use | Privacy policy

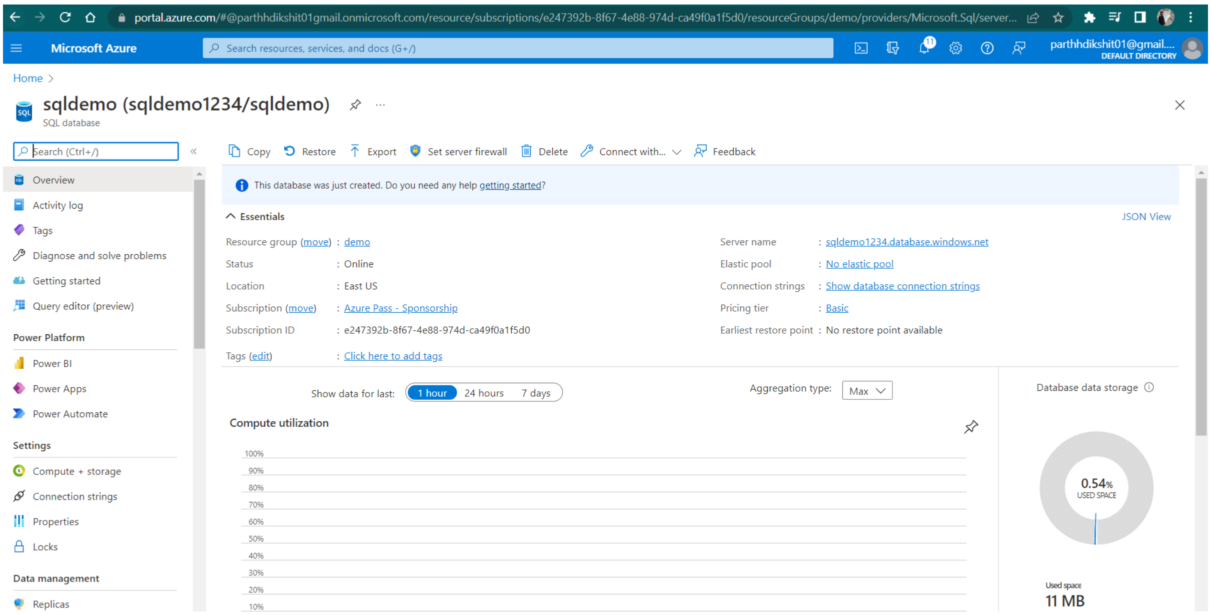
Subscription credits apply
1.0086 INR/hr
Pricing for other VM sizes

TERMS

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. See the Azure Marketplace Terms for additional details.

Fig 19 Creation of Virtual Machine

A virtual machine in Azure is an on-demand, scalable computer resource. Virtual machines are frequently used to host applications when a customer requires greater control over the computing environment than conventional compute resources can supply. When you utilise a vm to host your application, you receive the flexibility of virtualisation without having to buy or maintain any underlying real hardware. You will, however, be in charge of administering the normal chores associated with other servers, such as configuration and monitoring. patching, and application deployment.



portal.azure.com/#@parthhdikshit01gmail.onmicrosoft.com/resource/subscriptions/e247392b-8f67-4e88-974d-ca49f0a1f5d0/resourceGroups/demo/providers/Microsoft.Sql/server...

sqldemo (sqldemo1234/sqldemo)

SQL database

This database was just created. Do you need any help getting started?

Essentials

Resource group (move) : demo
Status : Online
Location : East US
Subscription (move) : Azure Pass - Sponsorship
Subscription ID : e247392b-8f67-4e88-974d-ca49f0a1f5d0
Tags (edit) : Click here to add tags

Server name : sqldemo1234.database.windows.net
Elastic pool : No elastic pool
Connection strings : Show database connection strings
Pricing tier : Basic
Earliest restore point : No restore point available

Show data for last: 1 hour 24 hours 7 days

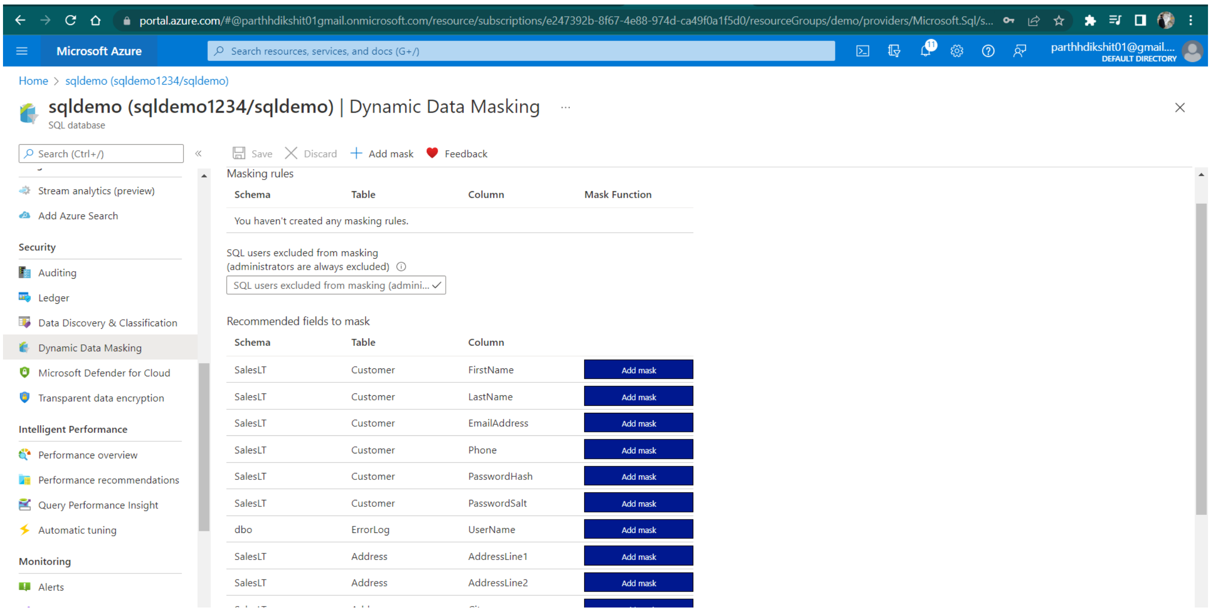
Aggregation type: Max

Fig 20 Creation of SQL Database

Azure SQL Database is a managed service database service, that means Microsoft manages and maintains SQL Server for you. SQL Database also contains new features like built-in high availability to help you maintain business continuity.

Azure SQL Database contains some pretty important functions. The following are a handful of them:

- Azure SQL Database has more freedom because Azure will manage SQL Server behind the scenes while the user concentrates on the database.
- Azure can provide an elastic pool where you can share a bunch of databases, pay for them as a group, and administer them as a whole. In hoster circumstances, this is really useful.
- Azure will also provide you with choices such as HyperScale, which will allow you to create databases of any size you require for maximum performance with really large-scale applications.
- Things like serverless computing, where Azure can even suspend an idle SQL Server and you just pay for the resources you utilize.


portal.azure.com/#@parthhdikshit01gmail.onmicrosoft.com/resource/subscriptions/e247392b-8f67-4e88-974d-ca49f0a1f5d0/resourceGroups/demo/providers/Microsoft.Sql/s...
Microsoft Azure
Search resources, services, and docs (G+/)
parthhdikshit01@gmail....
DEFAULT DIRECTORY
Home > sqldemo (sqldemo1234/sqldemo)
sqldemo (sqldemo1234/sqldemo) | Dynamic Data Masking
SQL database
Search (Ctrl+/)
Stream analytics (preview)
Add Azure Search
Security
Auditing
Ledger
Data Discovery & Classification
Dynamic Data Masking
Microsoft Defender for Cloud
Transparent data encryption
Intelligent Performance
Performance overview
Performance recommendations
Query Performance Insight
Automatic tuning
Monitoring
Alerts
Save
Discard
Add mask
Feedback

Masking rules

Schema	Table	Column	Mask Function
You haven't created any masking rules.			

SQL users excluded from masking (administrators are always excluded)

SQL users excluded from masking (admini...

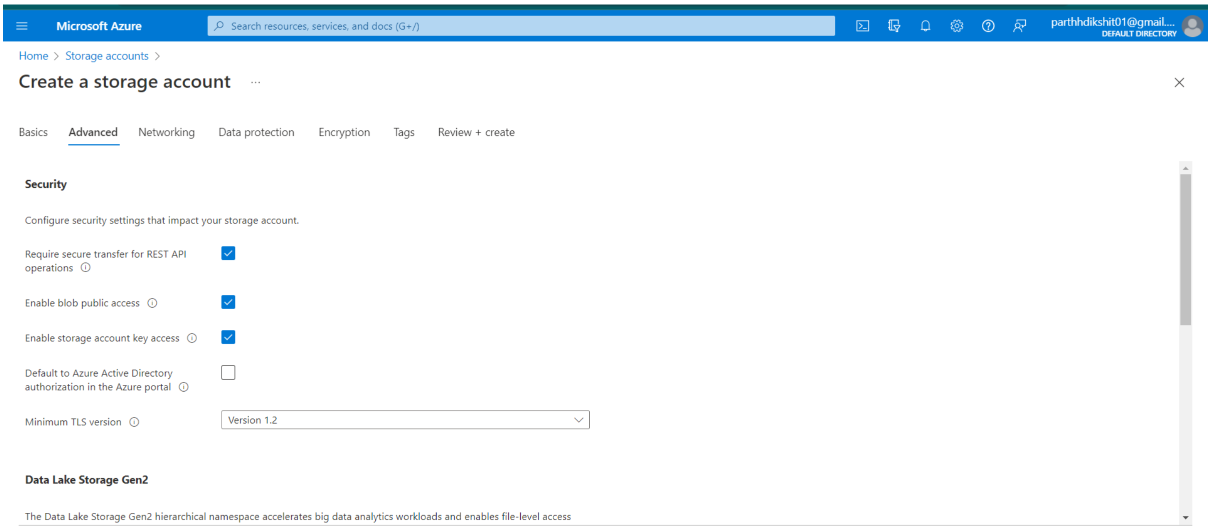
Recommended fields to mask

Schema	Table	Column	
SalesLT	Customer	FirstName	Add mask
SalesLT	Customer	LastName	Add mask
SalesLT	Customer	EmailAddress	Add mask
SalesLT	Customer	Phone	Add mask
SalesLT	Customer	PasswordHash	Add mask
SalesLT	Customer	PasswordSalt	Add mask
dbo	ErrorLog	UserName	Add mask
SalesLT	Address	AddressLine1	Add mask
SalesLT	Address	AddressLine2	Add mask

Fig 21 Masking

Data masking is just a method for obtaining a fictitious but realistic representation of your user's data. When genuine data is not required, such as in user training, sales demos, or software testing, the purpose is to secure sensitive data while offering a functioning replacement.

Data masking processes alter the data's values while maintaining the same format. The idea is to develop a version that is impossible to interpret or reverse engineer.



Create a storage account

Security

Configure security settings that impact your storage account.

Require secure transfer for REST API operations ☑

Enable blob public access ☑

Enable storage account key access ☑

Default to Azure Active Directory authorization in the Azure portal □

Minimum TLS version: Version 1.2

Data Lake Storage Gen2

The Data Lake Storage Gen2 hierarchical namespace accelerates big data analytics workloads and enables file-level access

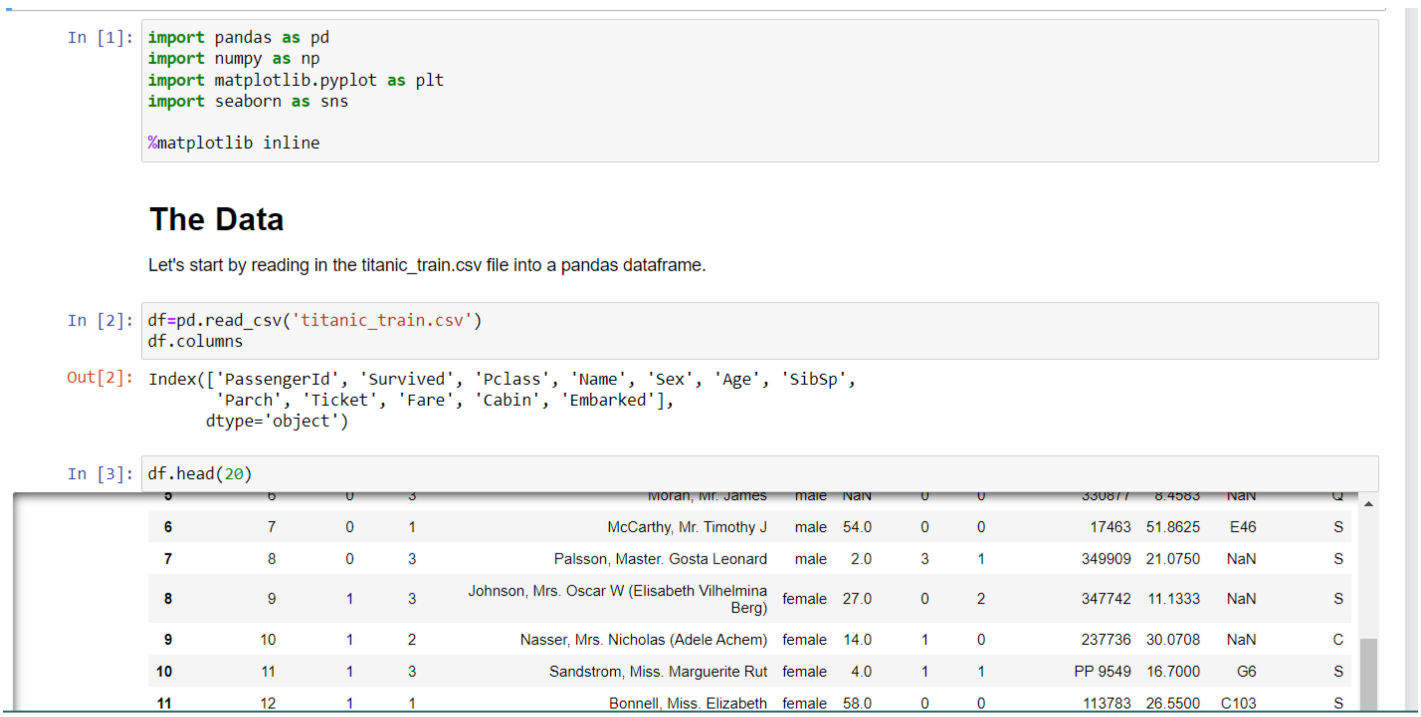
Fig 22 Storage Account

A storage account is a container that joins several Azure Storage services together. Only Azure Storage data services are allowed in a storage account. The user can handle data services as a group when they are connected into a storage account. The options you select when creating your account, or changes you make afterward, apply everywhere. All data stored in the storage account is erased when it is deactivated.

Several types of data accounts are available in Azure Storage. Each type has its own set of features and pricing structure. To determine the optimum storage account for the applications, consider these differences before creating one. The following are the different types of storage accounts:

- General-purpose v2 accounts
- General-purpose v1 accounts
- Block Blob Storage accounts
- File Storage accounts
- Blob Storage accounts

4.6: Code Screenshots



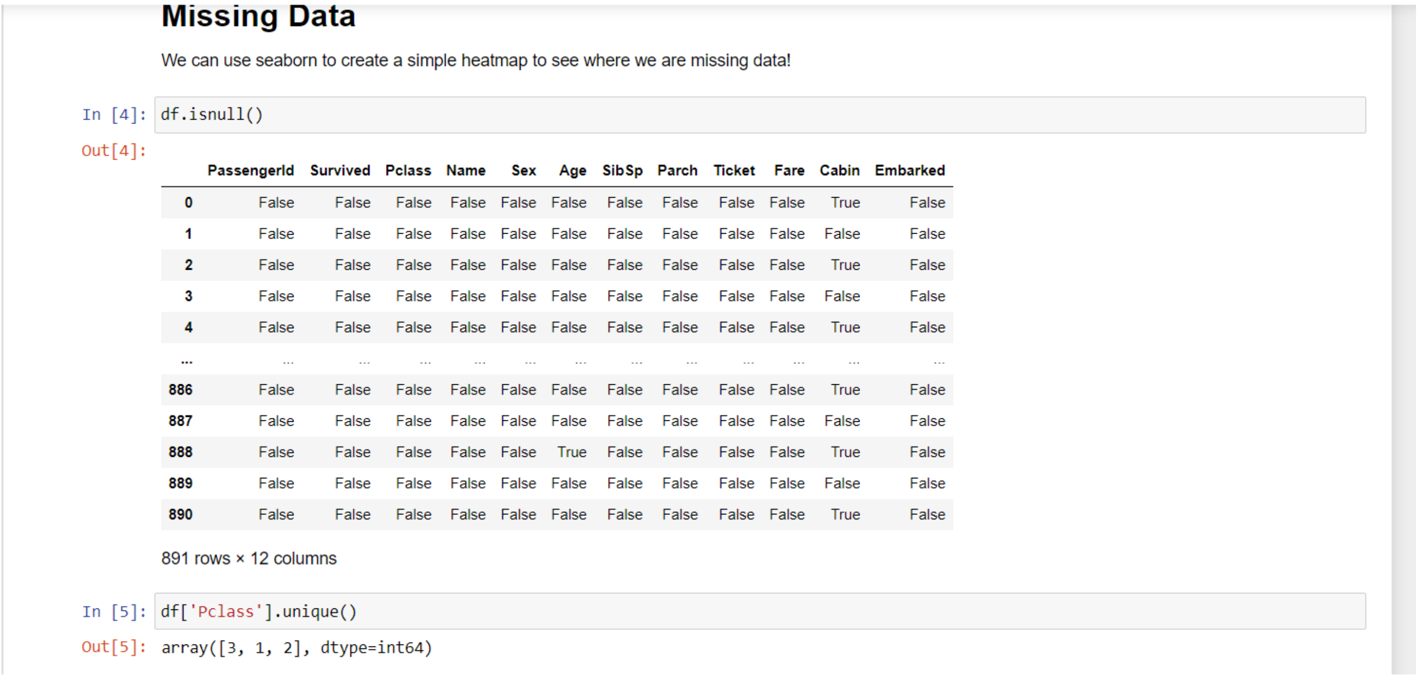
```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns

        %matplotlib inline
```

The Data

Let's start by reading in the titanic_train.csv file into a pandas dataframe.

```
In [2]: df=pd.read_csv('titanic_train.csv')
        df.columns

Out[2]: Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
               'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
              dtype='object')

In [3]: df.head(20)
```

5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
10	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S



Missing Data

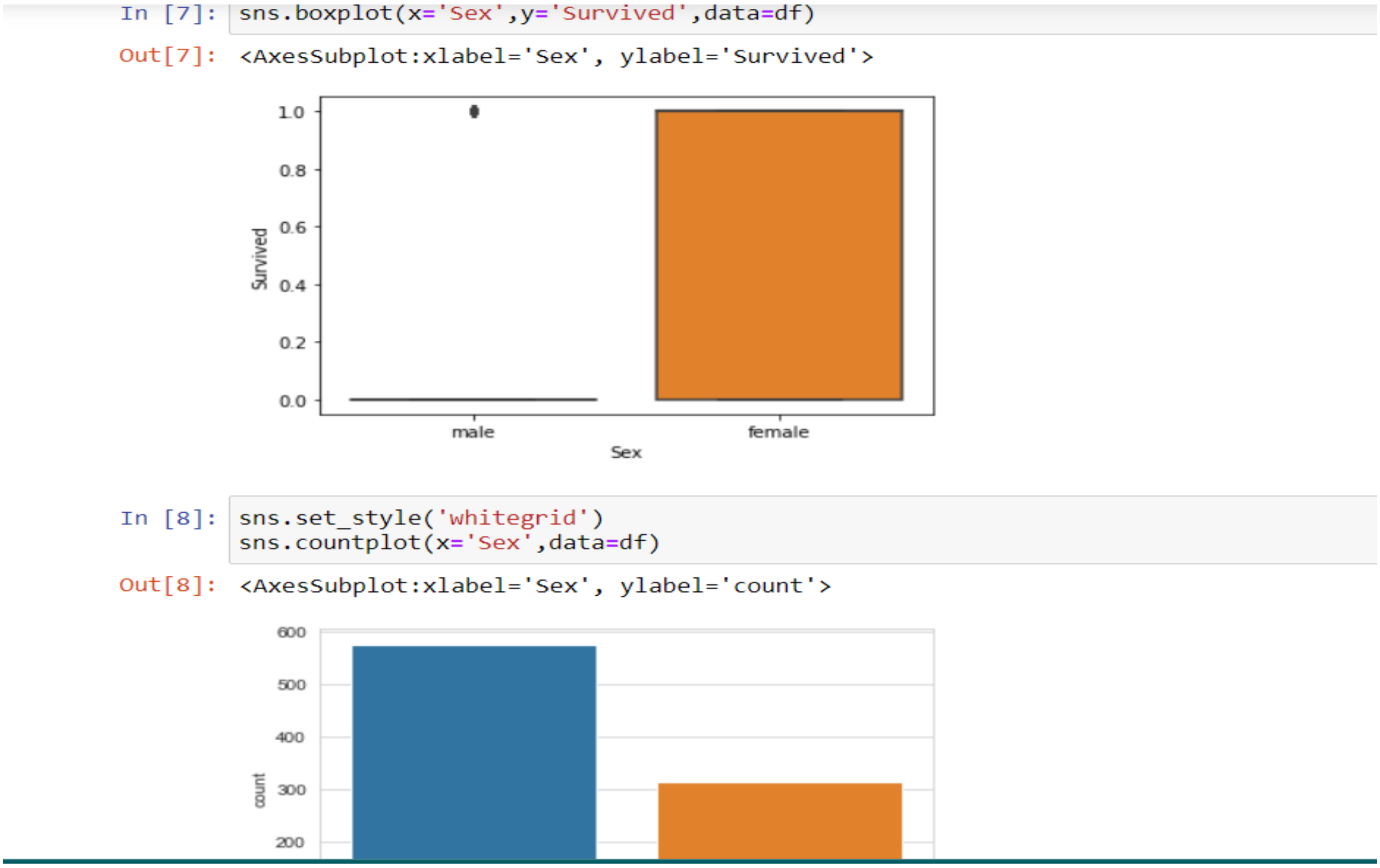
We can use seaborn to create a simple heatmap to see where we are missing data!

```
In [4]: df.isnull()

Out[4]:
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	True	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False	False	True	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	True	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	True	False

891 rows × 12 columns

```
In [5]: df['Pclass'].unique()

Out[5]: array([3, 1, 2], dtype=int64)
```

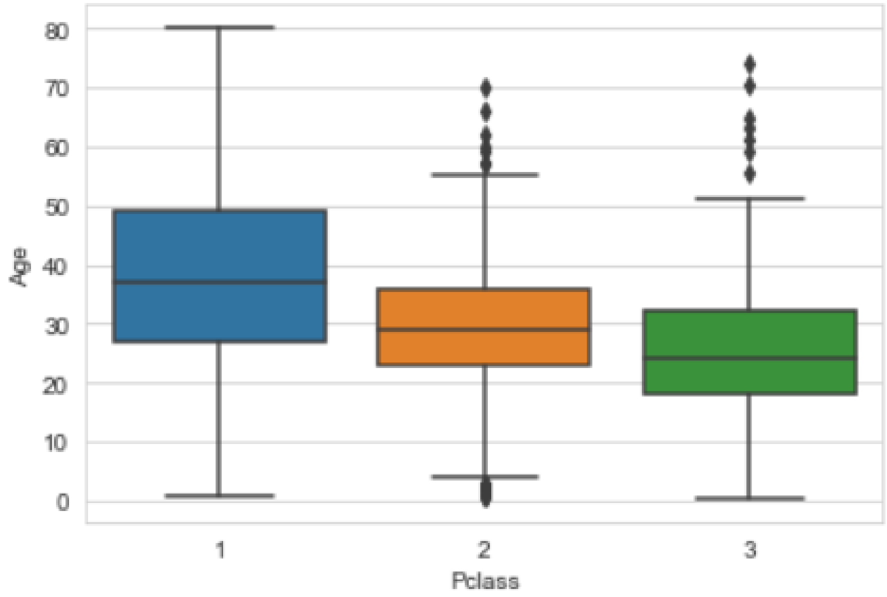


```
In [7]: sns.boxplot(x='Sex',y='Survived',data=df)

Out[7]: <AxesSubplot:xlabel='Sex', ylabel='Survived'>
```

```
In [8]: sns.set_style('whitegrid')
        sns.countplot(x='Sex',data=df)

Out[8]: <AxesSubplot:xlabel='Sex', ylabel='count'>
```

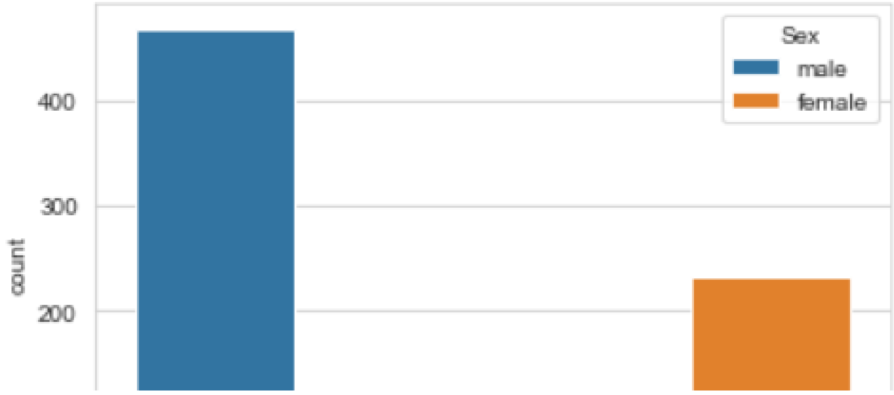
```
In [9]: sns.boxplot(x='Pclass',y='Age',data=df)
```

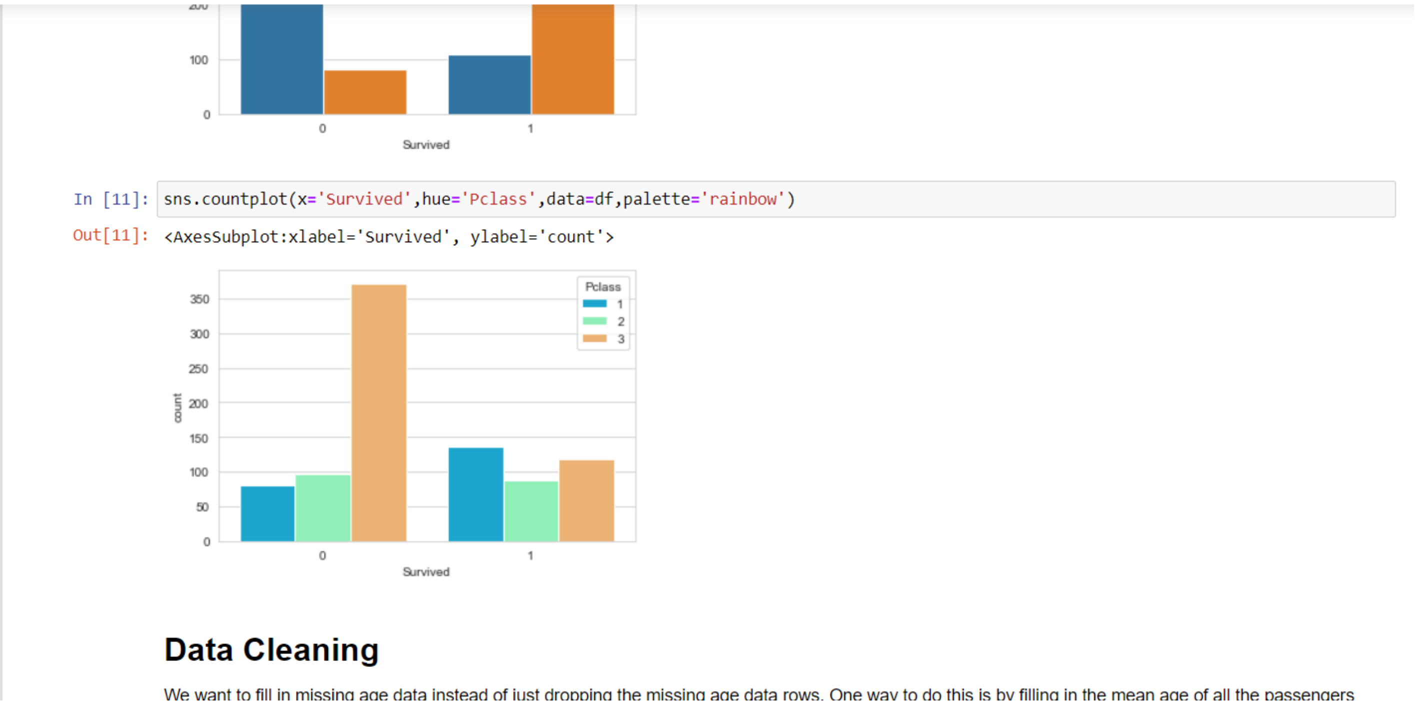
```
Out[9]: <AxesSubplot:xlabel='Pclass', ylabel='Age'>
```



```
In [10]: sns.countplot(x='Survived',hue='Sex',data=df)
```

```
Out[10]: <AxesSubplot:xlabel='Survived', ylabel='count'>
```

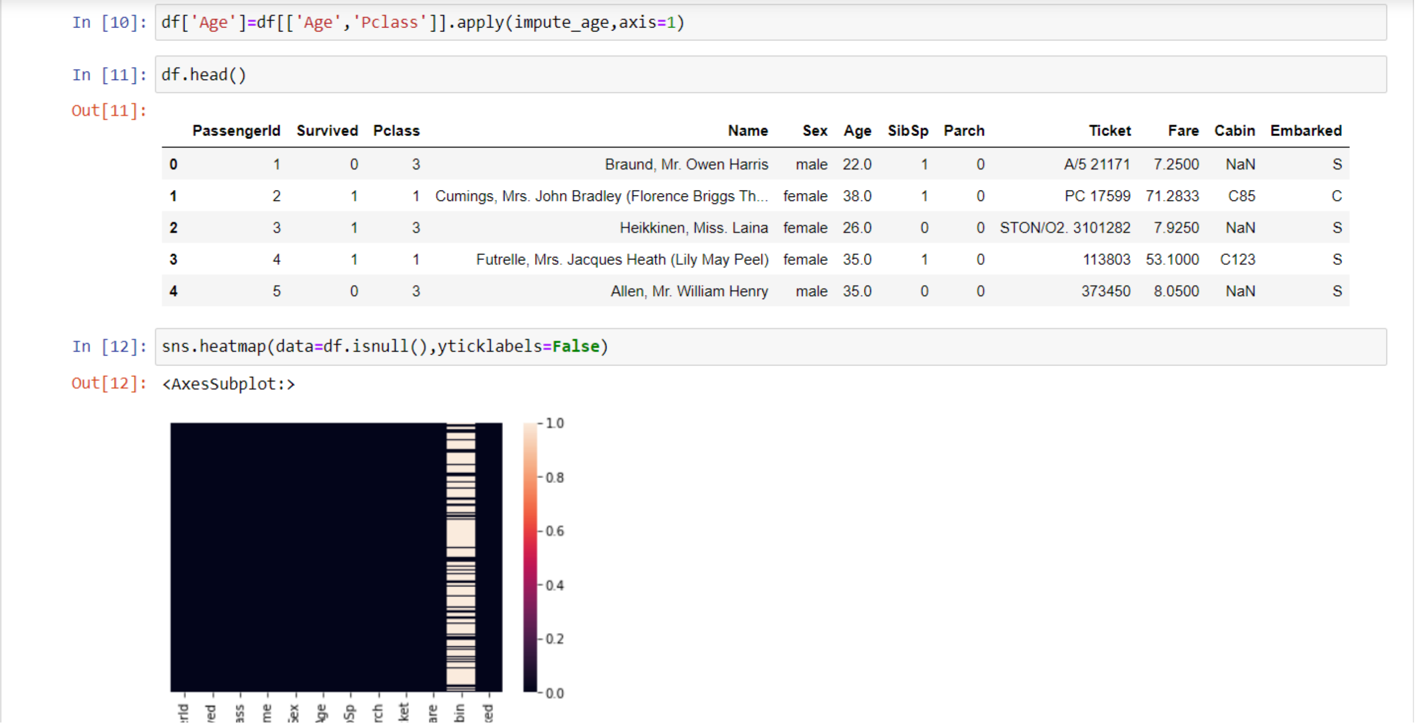




```
In [11]: sns.countplot(x='Survived',hue='Pclass',data=df,palette='rainbow')
Out[11]: <AxesSubplot:xlabel='Survived', ylabel='count'>
```

Data Cleaning

We want to fill in missing age data instead of just dropping the missing age data rows. One way to do this is by filling in the mean age of all the passengers



```
In [10]: df['Age']=df[['Age','Pclass']].apply(impute_age,axis=1)
```

```
In [11]: df.head()
Out[11]:
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

```
In [12]: sns.heatmap(data=df.isnull(),yticklabels=False)
Out[12]: <AxesSubplot:>
```

Chapter 06: CONCLUSION

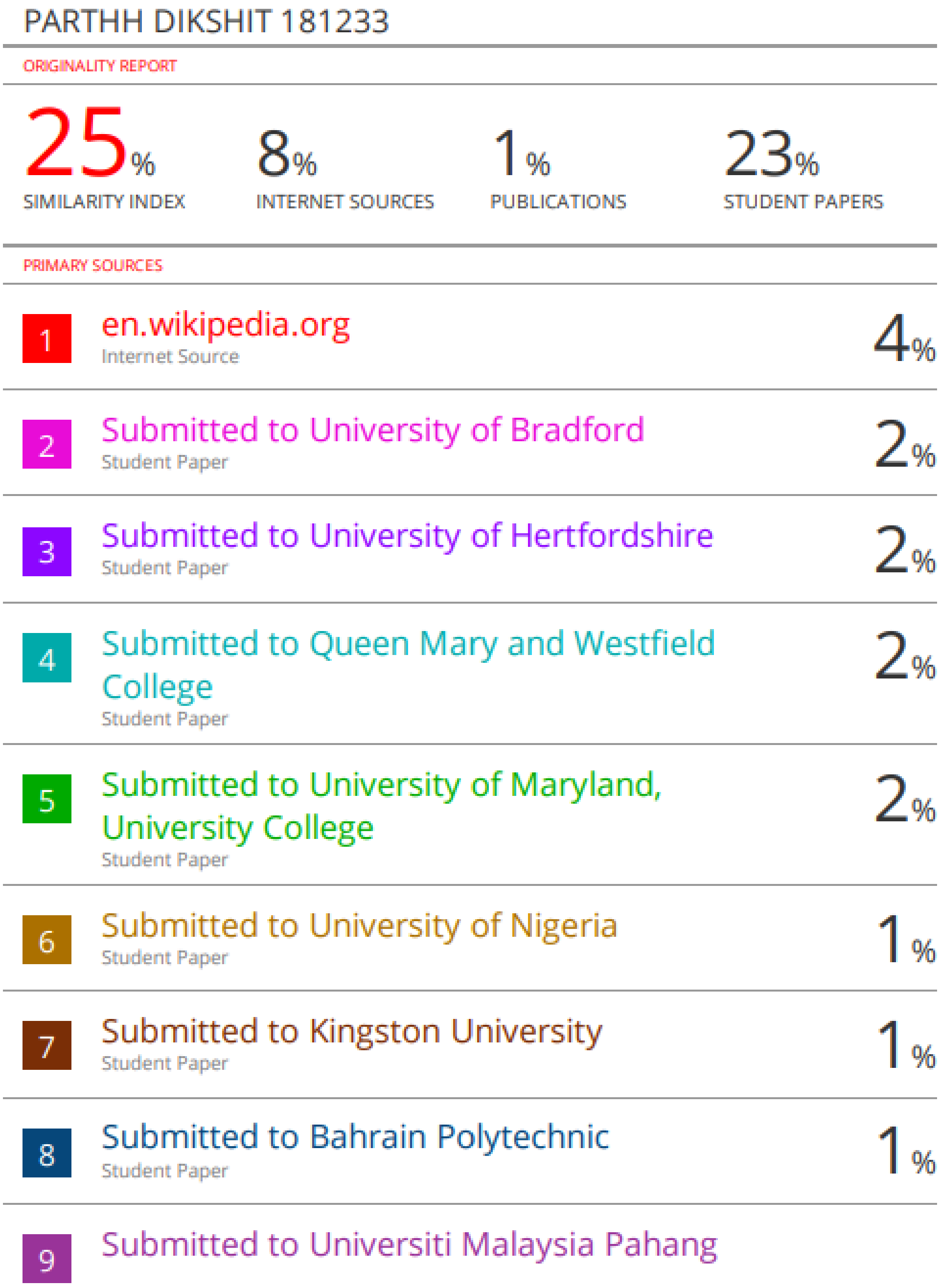
6.1 Conclusion

In conclusion, this internship has become a wonderful and rewarding experience. I can say that my stay with Cognizant Technology Solutions India Private Limited was quite beneficial to me. Needless to say, the technical aspects of my work aren't ideal and might be improved with more time. I believe the time I spent studying and comprehending Azure Cloud was well spent as one who has no prior familiarity with it. Two of the most significant lessons I've learned are time management and self-motivation.

REFERENCES

1. https://www.cognizant.com/
2. https://www.python.org/
3. https://www.w3schools.com/python/
4. https://www.w3schools.com/mySQl/default.asp
5. https://www.postgresqltutorial.com/
6. https://azure.microsoft.com/en-in/
7. https://www.tutorialspoint.com/microsoft_azure/index.htm
8. https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
9. https://towardsdatascience.com/step-by-step-guide-building-a-prediction-model-in-python-ac441e8b9e8b
10. https://azure.microsoft.com/en-in/services/

PLAGIARISM REPORT



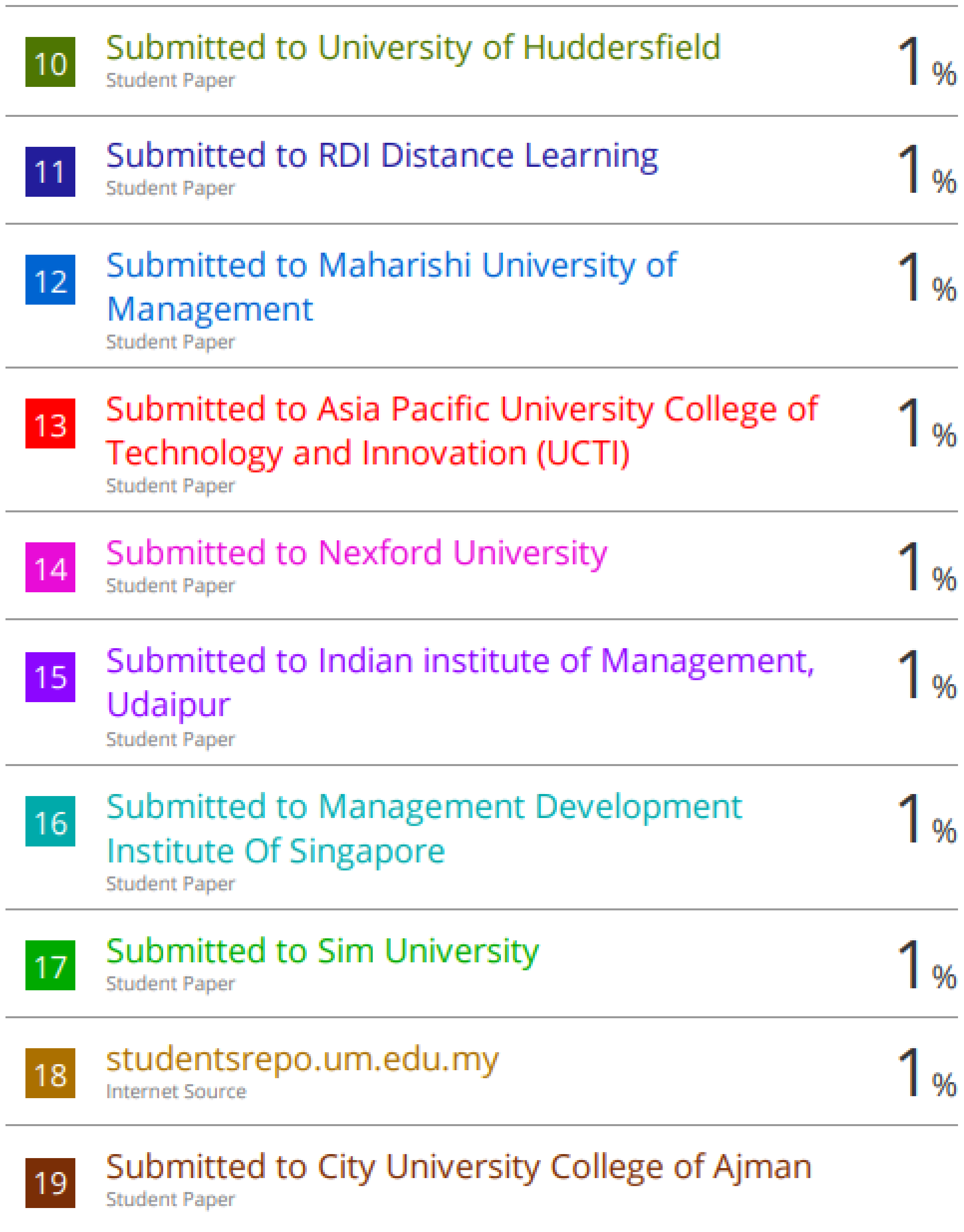
PARTHH DIKSHIT 181233

ORIGINALITY REPORT

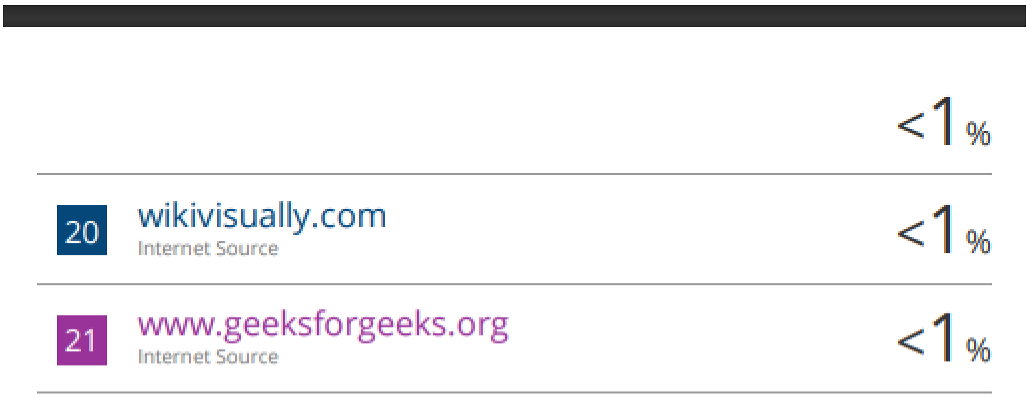
25%	8%	1%	23%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	en.wikipedia.org Internet Source	4%
2	Submitted to University of Bradford Student Paper	2%
3	Submitted to University of Hertfordshire Student Paper	2%
4	Submitted to Queen Mary and Westfield College Student Paper	2%
5	Submitted to University of Maryland, University College Student Paper	2%
6	Submitted to University of Nigeria Student Paper	1%
7	Submitted to Kingston University Student Paper	1%
8	Submitted to Bahrain Polytechnic Student Paper	1%
9	Submitted to Universiti Malaysia Pahang	



10	Submitted to University of Huddersfield Student Paper	1 %
11	Submitted to RDI Distance Learning Student Paper	1 %
12	Submitted to Maharishi University of Management Student Paper	1 %
13	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1 %
14	Submitted to Nexford University Student Paper	1 %
15	Submitted to Indian institute of Management, Udaipur Student Paper	1 %
16	Submitted to Management Development Institute Of Singapore Student Paper	1 %
17	Submitted to Sim University Student Paper	1 %
18	studentsrepo.um.edu.my Internet Source	1 %
19	Submitted to City University College of Ajman Student Paper	



		<1 %
20	wikivisually.com Internet Source	<1 %
21	www.geeksforgeeks.org Internet Source	<1 %