**Data: Oil of the Digital Age**

Suryam Gupta

I19MA038

SVNIT

[i19ma038@ahmd.svnit.ac.in](mailto:i19ma038@ahmd.svnit.ac.in)

**Introduction**

We are living in the age of Information and Computerization. With the availability of Internet and its access to people all over the world, huge amount of data is produced and stored every day. To give you an idea, in 1 Internet Minute, 400,000 hours of video is streamed on Netflix, 500 hours of video is uploaded by users on YouTube and 42 million messages are shared on WhatsApp ­­– this data keeps on increasing exponentially.

But hold on, what exactly is data?

Well, it could be something as simple as when you go to buy a gadget in an electronic store, and the salesperson stores your unique personal information such as Customer\_Id, Name, Phone Number, Address, Product, Quantity, etc. into an Excel Sheet named ‘April\_Sales\_2021.xls’. Your name, along with your various attributes and preferences, is a record or a data. Now you can only imagine with the amount of people and their interactions on so many different platforms available these days, that handling and processing this data becomes a topic of utmost importance. That’s where Data Science comes in.

**Data Science**

Data Science is said to be an umbrella term which has various fields under it, namely **Big Data Analytics, Machine Learning, Neural Networks (Deep Learning), Artificial Intelligence (AI), Statistics and Probability, Modelling**, amongst others.

This aforementioned enormous data generated is extremely valuable and if one can observe trends and patterns in it and subsequently make data-driven decisions accordingly, businesses can greatly increase their efficiency and productivity. This is the purpose of Data Science. It can be defined as **a field that uses the Knowledge of Mathematics, Business and Domain Knowledge, Programming Skills, Statistics and Models to extract meaningful insights from raw data**. These insights can then be used to see patterns which can be modelled for making predictions and hence help in making better business decisions in the future. Note that ‘meaningful insights’ doesn’t necessarily mean that you will directly have the answer of what needs to be done, but rather it can give you an idea on how a problem should be approached. This accounts for the primary difference between Data Science and Data Analytics.

**Data Science deals with asking the right questions to understand the problem, and Data Analytics deals with answering these questions.**

**Machine Learning (ML)**

Machine Learning is a subset/ an application of Artificial Intelligence (AI) that offers the ability to the system to learn and improve from experience without being explicitly programmed to that level. Machine Learning uses data to train and find accurate results. It focuses on the development of a computer program/ algorithm that accesses the data and uses it to learn from themselves. Basically, a ‘Machine Learning Algorithm’ is chosen, and the process starts with inputting ‘training set’ into it to build a ‘Machine Learning Model’. This model is then tested by inputting the new ‘validation set’ to check if it is giving predicted results. If not, the model may be re-trained multiple number of times or the algorithm may be tweaked to give better results. This makes the Machine Learning Algorithm continuously learn on its own and produce the most optimal answers that will gradually increase in accuracy over time.

A simple example

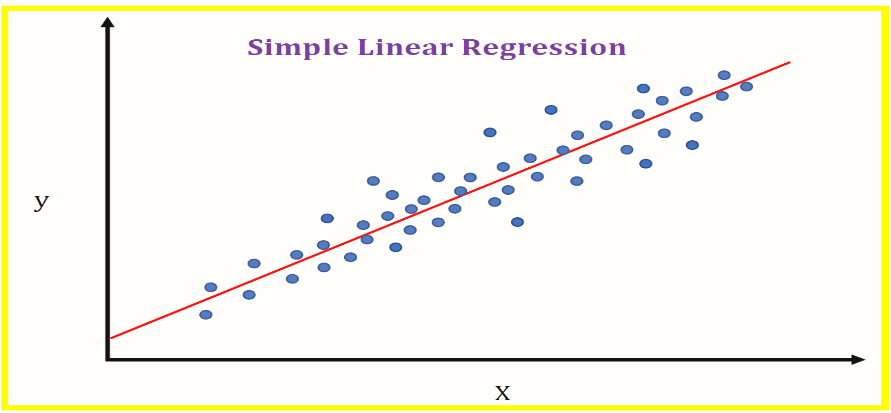
Suppose you have the following data points –

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| y | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 |

Random collection of points in the xy plane

Looking at this data, you begin by choosing one of the most common algorithms in Machine Learning, known as Linear Regression. This whole data is then split into ‘training set’ and ‘validation set’. So, say you input values of both the co-ordinates from x=1 to x=7 into your model as training data, it recognizes a pattern and concludes that all these data points fall on the line y = 5x. Now you input your validation data, i.e. both co-ordinates of x = 8,9,10 into this model to check if the model can correctly predict the accurate answers. It predicts 40, 45 and 50 respectively as answers, which matches with the values in the original data. You conclude that you have built an accurate Machine Learning Model which can now be used for any number of values of x to predict correct values of y in the future.

(Note that the above example is very basic, just to give you an idea of how things work on surface level. It is very important to first look and understand the raw data and then choose an algorithm which might go best with it. The data is split into one more section known as ‘Testing Set’ and there are various rules regarding the proportion of this splitting, as it all depends on the original data set. Another point to note is that the data used above is absolutely perfect, hence it builds a perfect model which gives answers with 100% accuracy. But in practicality, the original data is very random and you almost never get a perfect model. Hence even after building one, you need to keep working on it so that it becomes more and more accurate.)



Classification of Machine Learning

There are broadly two types of machine learning algorithms: Supervised and Unsupervised.

1. **Supervised:** Supervised learning algorithms are used when the data is labelled, i.e. it has input and output parameters in a completely machine-readable pattern. There are two types:
2. **Regression:** When you need to predict continuous values (variables may or may not be linearly dependent), algorithms used are Linear and Multiple Regression, Decision Trees and Random Forest.
3. **Classification:** When you need to predict categorical values, some of the classification algorithms used are K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM) and Naïve-Bayes.
4. **Unsupervised:** Unsupervised learning algorithms are used when the data is unlabelled, i.e. it only has one or none of the parameters in a machine-readable form (no target value) and hence there is no labelled data to learn from. There are two types:
5. **Clustering:** This is the method of dividing the objects which are similar between them and dissimilar to others. K-Means and DBSCAN clustering algorithms are two such commonly used models.
6. **Association-rule analysis:** This is used to discover interesting relations between variables. Apriori and Hidden Markov Model algorithm can be used.

Applications of Machine Learning

1. **Recommendation Systems:** Netflix recommending shows you might like and Amazon recommending similar products are great examples of how their models were trained by your previous searches and preferences, resulting in giving you a better personalized experience.
2. **Speech Recognition:** Google’s ‘Search by voice’ option, Google Assistant, Siri, Cortana, Alexa, all use ML Algorithms for Speech Recognition and converting voice instructions into text.
3. **Predictive Algorithms:** Dating sites match people by predicting which individuals might be compatible for each other based on their likes and interests.
4. Others include Image Recognition, Traffic Prediction, Self-Driving Cars, Email Spam, Malware Filtering and Medical Diagnosis.

**Data Analytics**

Data Analysis is defined as a process of cleaning, transforming, and modelling data to discover useful information for business decision-making. It doesn’t matter whether you are working in Big Tech companies like Facebook, Amazon, Apple, Netflix, Google (FAANG) or Microsoft, or you have a small business, say, an electronic store of your own, everyone needs to make use of Data Analysis to make better decisions for more productivity. Analysing and correcting the mistakes made in the past, eventually results in the growth of the business.

A real-life example

Suppose you made some sales for the month of April from your electronic store and now you want to come up with a strategy which would ensure more sales in the following months. You give your data, i.e. the Excel Sheet named ‘April\_Sales\_2021.xls’ to a data scientist. They build a model, run your data through it to see for some potential patterns, and conclude that finding the 2 products which are sold together the most, could be helpful. To find this pair of products, the data analyst goes through the steps of various processes and comes to the conclusion that the laptop ‘Lenovo Legion 5i’ and the mouse ‘Lenovo Legion M300 RGB Gaming Mouse’ are sold together the most. In this case, an effective marketing strategy would be to teach the salespeople in your store that whenever a customer is interested in buying the aforementioned laptop, the aforementioned mouse should be recommended to them as they are very likely to buy it, according to our data analysis. This strategy will lead to an increase in sales and an eventual growth in business.

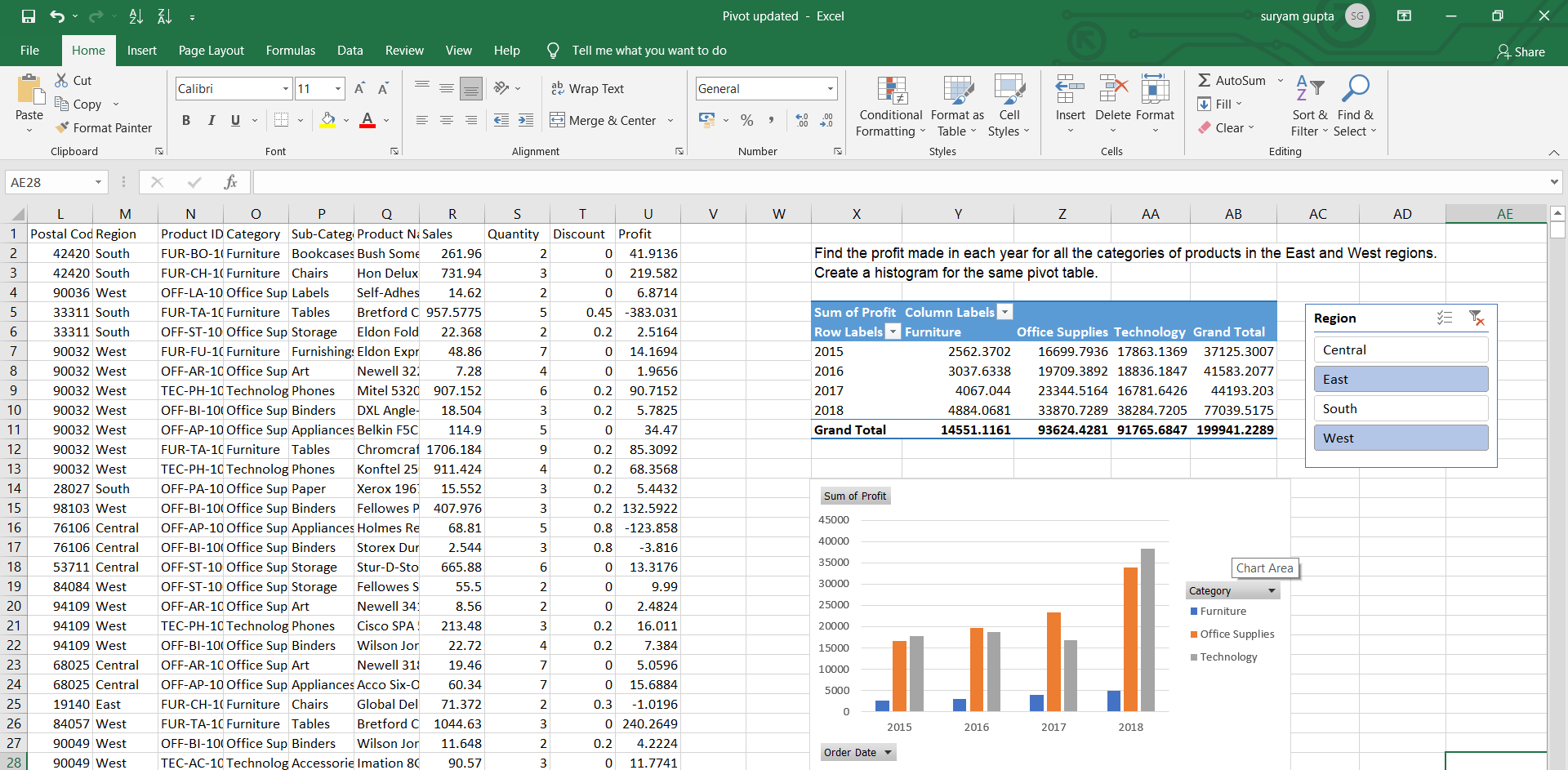
Just imagine how many more such complex and interesting business problems could be questioned and answered through Data Science and Data Analysis!

Data Analysis Process

1. **Define Goals:** First of all, you need to have a clear idea of why you are analysing, i.e. the aim or purpose of this analysis and what type of data analysis (text, statistical, diagnostic, predictive, etc.) you are going to implement. This helps with the type of data you will need to collect and analyse.
2. **Data Collection:** You now bring all the required data into one place for organizing, cleaning and analysis. Excel is a great platform for storing your data.
3. **Data Cleaning:** One of the most important processes that has to be compulsorily done before beginning your analysis. This includes removing some of the data which is irrelevant to your aim of analysis, deleting duplicate records and extra white spaces which might give calculation errors, checking for spelling mistakes, and making sure that the overall data is clean and free of errors and outliers.
4. **Data Analysis:** This is the part where you use various Data Analysis Tools and software to manipulate the data, so as to understand, interpret and derive insights and conclusions by finding the exact information you needed to answer.
5. **Data Visualization:** Sometimes it can get difficult to look for trends and patterns among so many values in a huge dataset, and hence, visualizing the data by converting it into charts and graphs makes it easier for our brain to spot them. Also, as a data analyst, you understand all the numbers and processing, but your superiors, who ultimately have to take the decision on whether to implement your analysis or not, might not have that level of understanding. Hence it becomes very important for you to show your analysis through convincing charts and graphs to make it easier for them to understand and process.

Data Analysis Tools

1. **Microsoft Excel:** Excel is one of the most common tools and a ‘must know’ for manipulating spreadsheets and for doing simple analysis. Arithmetic Manipulation, Functions and Formulas, Absolute and Relative references, Filtering and Sorting, VLOOKUP and HLOOKUP functions and Pivot Tables are some of its most common features.

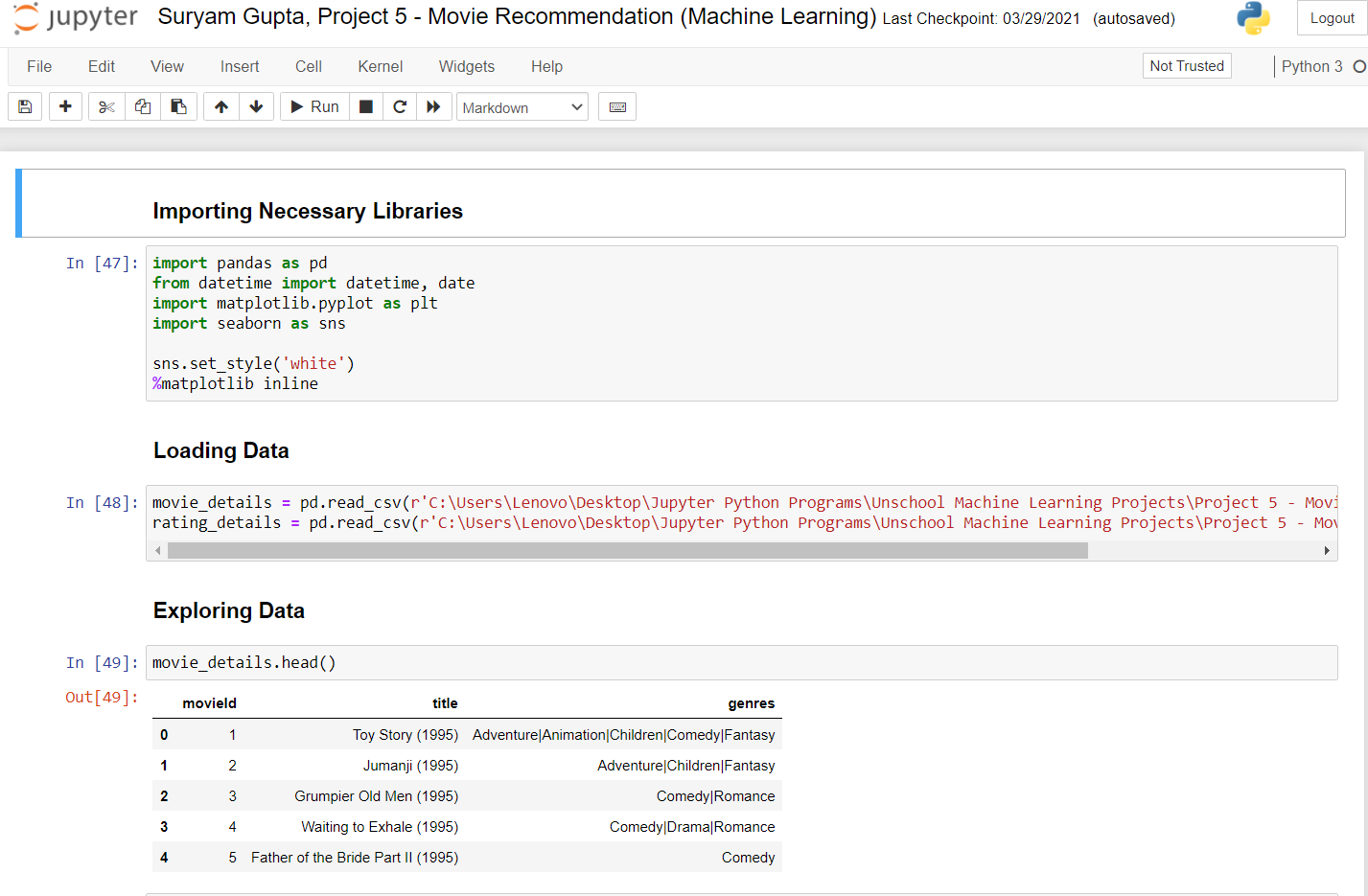
****

1. **Python:** Along with basic programming skills, one also needs to have knowledge of the wide varieties of libraries and packages that Python provides:

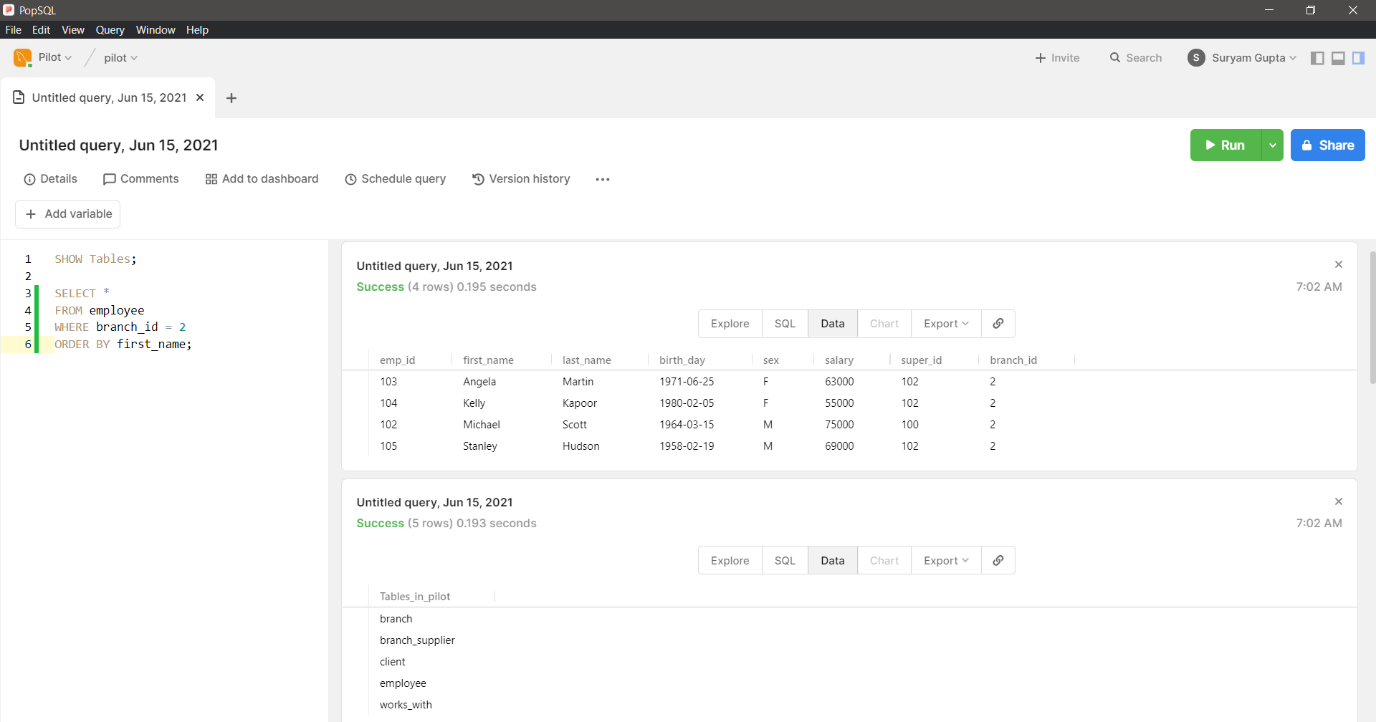
* Pandas for data manipulation and analysis
* NumPy and SciPy for mathematical and scientific computations
* Matplotlib, Seaborn and Plotly for data visualization
* SciKit-Learn for building ML models
* TensorFlow and Keras for Deep Learning models

Note that instead of Python IDEs like PyCharm or text editors like Visual Studio Code, Jupyter Notebook is preferred for Data Science purposes.

Other popular programming languages include R and SAS.



1. **SQL:** Structured Query Language is one of the most requested skills in Data Science. It is a programming language used to query and manage data in relational databases. SQL can interact with various Relational Database Management Systems (RDBMS) like MySQL, SQLite, PostgreSQL, Oracle, etc.

****

1. **Tableau:** Tableau is a powerful Data Visualization and Analytics software. It can easily connect to a data source and create data visualizations, maps and interactive dashboards which update in real-time.

Other popular and important tools include Microsoft Power BI, Apache Spark, Apache Hadoop, RapidMiner, KNIME and Qlik.

**Conclusion**

Data Science is referred to as “The Sexiest Job of the 21st century” by Harvard Business Review. With the amount of data increasing day by day, it has definitely emerged as a growing field with a bright future and people have come to realise its importance. Considering its applications in multiple sectors including but not limited to, Security, Transportation, Delivery, Healthcare, Travelling, Banking, Education, Energy, Agricultural and many others, Data Science is truly the future.

**Bibliography**

1. Claire Jenik, A Minute on the Internet in 2020, statista, <https://www.statista.com/chart/17518/data-created-in-an-internet-minute/>, accessed on 5th July 2021
2. Julian Birkinshaw, Beyond the Information Age, wired, <https://www.wired.com/insights/2014/06/beyond-information-age/>, accessed on 5th July 2021
3. Alex Castrounis, Data Science and Big Data Explained, kdnuggets, <https://www.kdnuggets.com/2016/11/big-data-data-science-explained.html>, accessed on 5th July 2021
4. Rohit Sharma, Data Mining Vs. Data Analytics, upgrad, <https://www.upgrad.com/blog/data-mining-vs-data-analytics/>, accessed on 5th July 2021
5. Bob Laurent, Data Science: Still the sexiest job of 21st Century, dominodatalab, <https://www.dominodatalab.com/blog/data-scientist-still-the-sexiest-job-of-the-21st-century/>, accessed on 6th July 2021
6. <https://www.proschoolonline.com/blog/top-10-data-analytics-tools>, accessed on 6th July 2021
7. <https://www.sisense.com/blog/data-science-vs-data-analytics/>, accessed on 6th July 2021
8. <https://www.javatpoint.com/applications-of-machine-learning>, accessed on 6th July 2021