Data science: a beginner's guide to the power of information

Aadhitya Sriram

12th Std student, The PSBB Millenium School, Gerugambakkam, Chennai, India

Abstract: Data science is a field of study that deals with abundant data resources using complex machine learning algorithms to derive meaningful information that is practically useful to the business world. In the past few years, data science has become one of the major talks of the century. For beginners, it may appear quite complicated to start learning data science, but with all the different ways by which it can be implemented, Data Science generates an overwhelming response from beginners. In this manuscript, we begin with a discussion on the basic definitions of data science and also look at why python is one of the best options for data science. This manuscript also attempts to implement data science using python's amazing collection of basic libraries and packages with several examples and discusses the fundamental pre-requisites and importance of data science in the 21st century. Finally, a comprehensive project with a detailed explanation of the same gives a summary of how data science helps acquire the power of information.

Keywords Data science, Python, NumPy, Seaborn, SciPy

I. INTRODUCTION

In the modern era, the one who holds control over information is considered powerful. A knowledgeable person is well informed and keeps himself updated on the current happenings in the world. The information that we have in hand not only gives us the power to use it but also enables us to flex and bend the future at our will. This can be done theoretically by predicting the probability of a future event and taking measures to prevent that event. What we need for such prediction is the 'Data source'.

II. EVOLUTION OF DATA SCIENCE

To organise data in our hands we have adopted various methods in the past. In the stone era, rocks, boulders and stone slabs were used to record and store data. But with the evolution of the human species, the discovery in ancient Egypt in 3000BC, 'Papyrus', changed the course of human history [1]. A paper-like material was used for maintaining records. Many experts claimed that this particular discovery hastened the course of human evolution. This public opinion is not directly attributed to papyrus, but it was the information stored in the material that helped develop the brain capacity of early humans.

This more durable and cost-efficient paper material, manufactured and brought to public use first in 100-200 AD, was a quick success wherever there was a need to store information officially, from local shops to imperial courts. Paper was transported to various continents across the globe through the famous silk routes and overseas trade. Information too followed the path of paper travel. This was a brief introduction about the birth of our beloved information.

The latter part of the 19th century and the 20th century witnessed one of the greatest revolutions in the field of data science. The invention of type-writers, audiotapes, photography, video camera devices and computers greatly enabled us to record and store all information. The significant milestone was the electronic storage as 'binary numbers', which paved way for the third industrial revolution. All of the information today is majorly stored as 0s and 1s in an allocated space called 'Memory'. Depending on the quantity of information to be stored, memory varies grossly. For example, a video is stored as information about pixels, which have 3 component colours called the primary colours [Appendix 1-A] and requires a lot of memory. Whereas a text file is composed of a list of encoding characters in a particular selected scheme such as Unicode or ASCII [Appendix 1-B]. Information is stored in various forms depending on the type of information that is to be stored.

In today's world, handling data and the creation of information is an art, needing a long learning curve to master that. In most instances, only raw data is available. In such settings, it is imperative to organise the Data available to make it ready-to-use information. In this manuscript, we shall discuss how Data Science can help us organize the data comprehensively and efficiently.

III. WHAT IS DATA SCIENCE?

Data science is a field that encompasses the conversion of unstructured data into structured data by using Mathematics, programming, and various algorithms [2]. It is the science of data that is now a major component of today's information-dependent world. It is mainly done by two key concepts, Data Analysis and Data visualisation. It is also a major component of Artificial Intelligence and Machine Learning and Deep Learning. A few of these are explained as we delve into future sections of the manuscript.

In general, Data science is a term used to describe a lot of things but the most interesting definition was given by, Drew Conway as a Venn Diagram [3]. A simplified version is given below. (Fig 1)

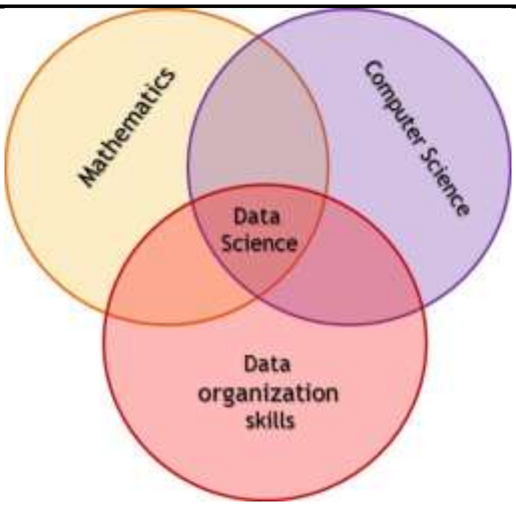


Fig 1 – A simplified Drew Conway's Data Science Venn diagram (created using canva)

The power of computers is mainly implemented in this paradigm using programming through specific languages. Some of the important languages that are primarily used for Data science are R, Python, Scala, Julia. Of these, there is heavy competition among R and python [Appendix 2-A]. Even though both of these are excellent choices, in this manuscript, let us discuss the usefulness of Python. The reason is that python can do much more than R. Another important reason is that Python has a web interface and can be easily interlinked into web applications whereas R cannot do so. Python has various support from its open-source community than R and hence has more modular package functions than R. Moreover, Python's syntax framing is a lot easier than that of R. For more information about Python in general, refer [Appendix 2-B]. That does not mean that python is superior to R. The R-Language excels in various other applications. Discussing the clinical utility of R language is beyond the scope of this paper.

After organising the data, one must find ways to visualise the structured data. There are many ways of representation that helps us to visualise all the data collected and organised. Some of the common ones include Line graphs, Bar graphs, Histograms, Pie charts, Scatter plots and Box plots (Fig 2).

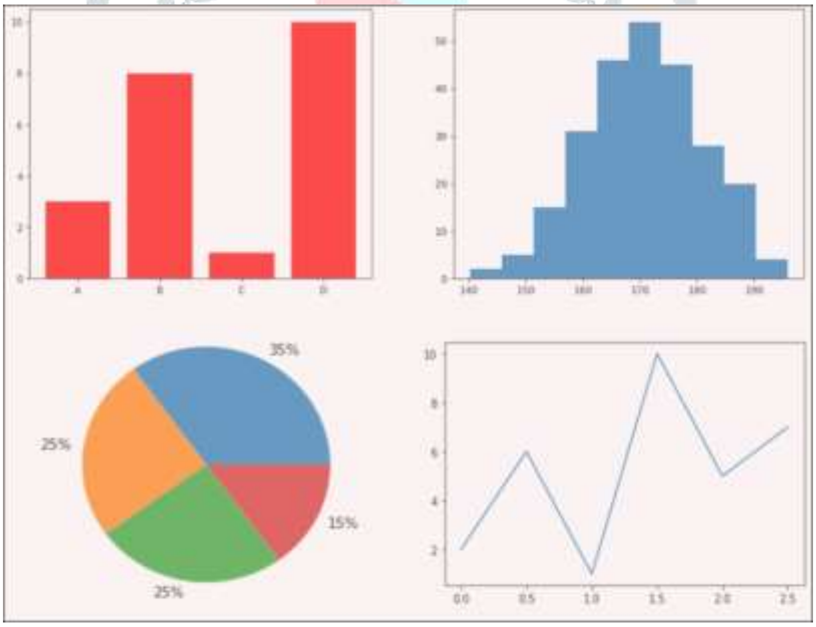


Fig 2 – Different ways to visualise data

As we can see, the visualised data is much more comprehensive than raw organised data. In many situations, it is important to visualise data so that it appeals to a larger viewership. When a programmer has mastery over visualising techniques, they can present their ideas in a much crisper and concise manner. This is the heart of data science, starting from analysing the data, to organising it in a structured format, finally to constructing and viewing it in visualised form. It is not as easy as it appears, but one of the most pressurizing and experience demanding skill. A Data Scientist must ensure that there is no loss of data while converting it into useful information. One small error can lead to drastic changes in the outcome. One can make it a bit easier by using some of the best Python packages available and are discussed in the subsequent sections.

IV. MATHEMATICS AND DATA SCIENCE

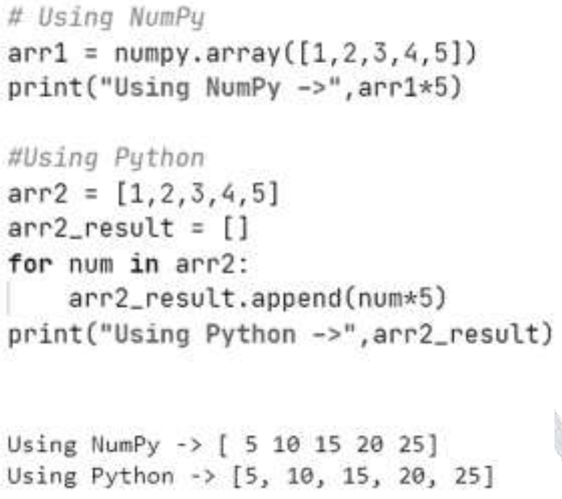
Mathematics is one of the basic pre-requisite for Data Science. Any field that is related to computer science is bound to have some form of mathematics. They are related so much that we can even refer to programming as 'fancy computer math'. We use many concepts of mathematics in Data Science such as Statistics, Probability, Vectors, Trigonometry, Geometry and Topology. In the following section, we introduce two packages in Python that allows us to manipulate data to a great extent using mathematics.

V. PYTHON PACKAGES FOR DATA SCIENCE

Python, being an open-source language can get a lot of support from the community. And many packages are being added every day by programmers across the globe. There are tons of packages for data science in python. But here we shall discuss the most effective and popular packages for Data Science such as NumPy, Pandas, Matplotlib, seaborn, SciPy etc.

5.1 ***NUMPY*:**

NumPy stands for Numerical Python. It is one of the most popular libraries that offers manipulation of data easily with less code in an easily interpretable array format [4]. It does that by creating a 'ndarray' object in which the numerical data is stored. The uniqueness of NumPy is that it allows operations on all of the data within them. Consider the following example where one must multiply the elements of a data holder (Fig 3).



```
# Using NumPy
arr1 = numpy.array([1,2,3,4,5])
print("Using NumPy ->",arr1*5)

#Using Python
arr2 = [1,2,3,4,5]
arr2_result = []
for num in arr2:
    arr2_result.append(num*5)
print("Using Python ->",arr2_result)
```

```
Using NumPy -> [ 5 10 15 20 25]
Using Python -> [5, 10, 15, 20, 25]
```

Fig 3 – Data manipulation using NumPy

As we can see, NumPy allows easier and faster manipulation than normal python manipulation with the same result in both cases. Imagine this result for all mathematical operations. This is why NumPy is one of the best tools for a Data Analyst.

5.2 ***PANDAS:***

Pandas is another important and indispensable package that offers other object types for storing data in form of pandas series and pandas dataframe. It allows easy access and retrieval of data from the dataframe which is essentially a tabular representation of data containing rows and columns. Pandas not only allow easy storage but also allow slicing a particular row or column or even both. It also allows us to modify them to our need. It has many in-built functions that support many Data Science processes like data cleaning that includes filling null values in a big data frame. It also has many functions related to statistics and probability. The possibilities become endless when we use these packages in correlation with each other.

5.3 ***MATPLOTLIB:***

Fig 2 had earlier illustrated visualisations of the processed data, using Matplotlib. Fig 4 illustrates the various visualization techniques in the form of bar graphs, line graphs, histograms, pie charts and other ways such as scatter plots, box plots, subplots, marginal plots, pairwise plots and violin plots (Fig 4).

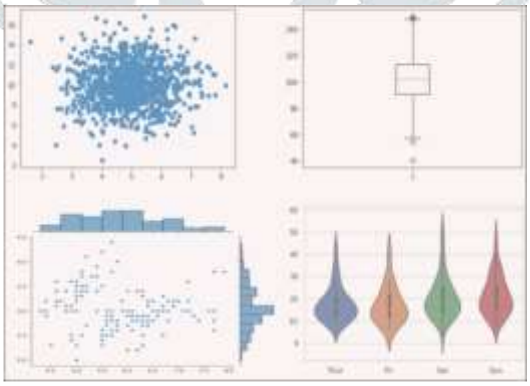


Fig 4 – Matplotlib Visualizations

Visualisations are much more comprehensive with Matplotlib and is a great way to represent one's data.

5.4 ***SEABORN:***

This package is another important visualisation tool that allows us to create much fancier and creative visualisations. It serves as an interface to create high-level statistical graphs. Some of the important ones include distplots, countplots, histplots, jointplots, density plots, heat maps etc (Fig 5). As mentioned earlier, reading these graphical representations requires knowledge of advanced mathematics. The uniqueness of Seaborn makes it an irreplaceable part of advanced data science problems often in fields related to other sciences.

5.5 ***SCI-PY:***

Data analysts often come across complicated and high level mathematical, scientific, engineering and technical problems. This is one scenario where SciPy shines and scores over all others. The word SciPy means Scientific Python. You can imagine SciPy to be a scientific calculator that we use for complex problems. SciPy provides many mathematical operations from basic Linear algebra, trigonometric functions, complex calculus functions, optimisation functions, Interpolation functions, Fourier transformations, etc.

It also has functions relating to Statistics, probability and Matrix-Determinant manipulations and many other advanced mathematical concepts [Appendix 3]. It is worthy to note that SciPy is not only used for Data Science but also advanced machine learning and deep learning concepts. This particular package (with many other sub-packages) forms the basis of many python projects.

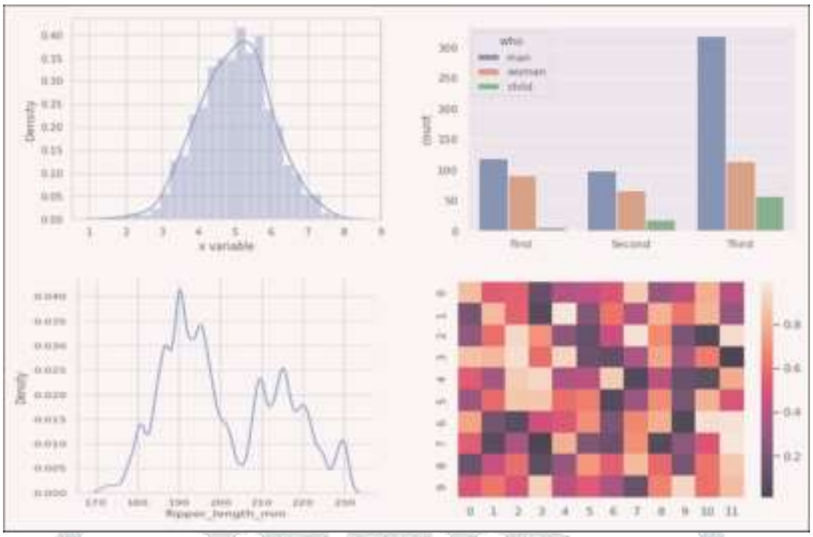


Fig 5 – Seaborn Visualizations

It is worthy to mention that many other packages such as TensorFlow, Keras, Scikit-learn are used in advanced Artificial intelligence, machine learning, and deep learning projects and could not be discussed here as it is out of the scope of this manuscript. To know more about the packages mentioned in the previous section refer [Appendix 3].

VI. IMPORTANCE OF DATA SCIENCE

As mentioned earlier, Information is just data that is organised and even visualised. In the modern world, data is anything and everything. From the records of our groceries to ledger statements in blockchains, everything is data, most of which are digitally programmable. And understanding the structure of the organisation is one of the most important factors that lead to better and efficient working, classification and retrieval of information stored in databases. It is one of the most demanding jobs that have a direct influence on the global economy.

All companies and institutions need Data Analysts working for them. The purpose is to classify the data coming in and organise the data going out so that the customers have a smooth experience and the companies and institutions acquire profit. Just imagine if you're accidentally charged twice the retail price just because there was a problem in retrieving the information of a product. The occurrence of such a scenario is very rare because most of these are automated functions. But for understanding, let us consider that it did occur.

Data Science does not apply to specific fields of work alone. We use data science almost everywhere to predict the future and analyse the past. It is done in small shops to get a record of their profit or loss. It helps us to get to know under what circumstances there was a lot of profit and gives us a clear picture of what is going on and how to optimise our methods and refine our plans by dividing a bigger part into many components and sub-components.

The Stock market is one such place where data science comes in handy. It allows us to predict or gamble on what might be the state of a particular stock in a month or year up to a good level of certainty. It is done by accessing all the previous year's data of a particular stock and representing them visually using the above-mentioned modules. It is well-advised to not entirely depend on data science to gamble in the stock market as it is very unpredictable.

Data science also prevails in the field of Physics, especially in Atomic physics and Astronomy. In Atomic physics, physics often needs a clear picture of what they cannot see or feel since they deal with subatomic particles of the order 10-15 metres. Here, analysis of larger sets of data and visualisation is very useful. For example, in 2012 physicists found experimentally, the presence of a subatomic particle known as the 'Higgs Boson' which was possible due to the advancements in computer technology implemented with data science [5]. On the other hand, in Astronomy, data science is used to predict the presence of exoplanets, pulsar stars, the brightness of distant stars, age and various other mysteries of our universe.

Setting aside all the examples given above, the most important application of data science is in correlation with another important concept called Machine learning. The meaning of this term is just how it sounds. Teaching machines. This is done for one important reason – Automation. There is one basic way to do this which is to first import and process a dataset and then implement a model to create what data scientists call train and test datasets. By doing this, we are training a program to do something particular. Machine learning is an advanced concept and hence is out of the scope of this manuscript. For further information refer [Appendix 4]. Data science plays a major role here by acquisition (creating access) and processing of data for the program to understand the data without any complications.

VII. DATA CLEANING

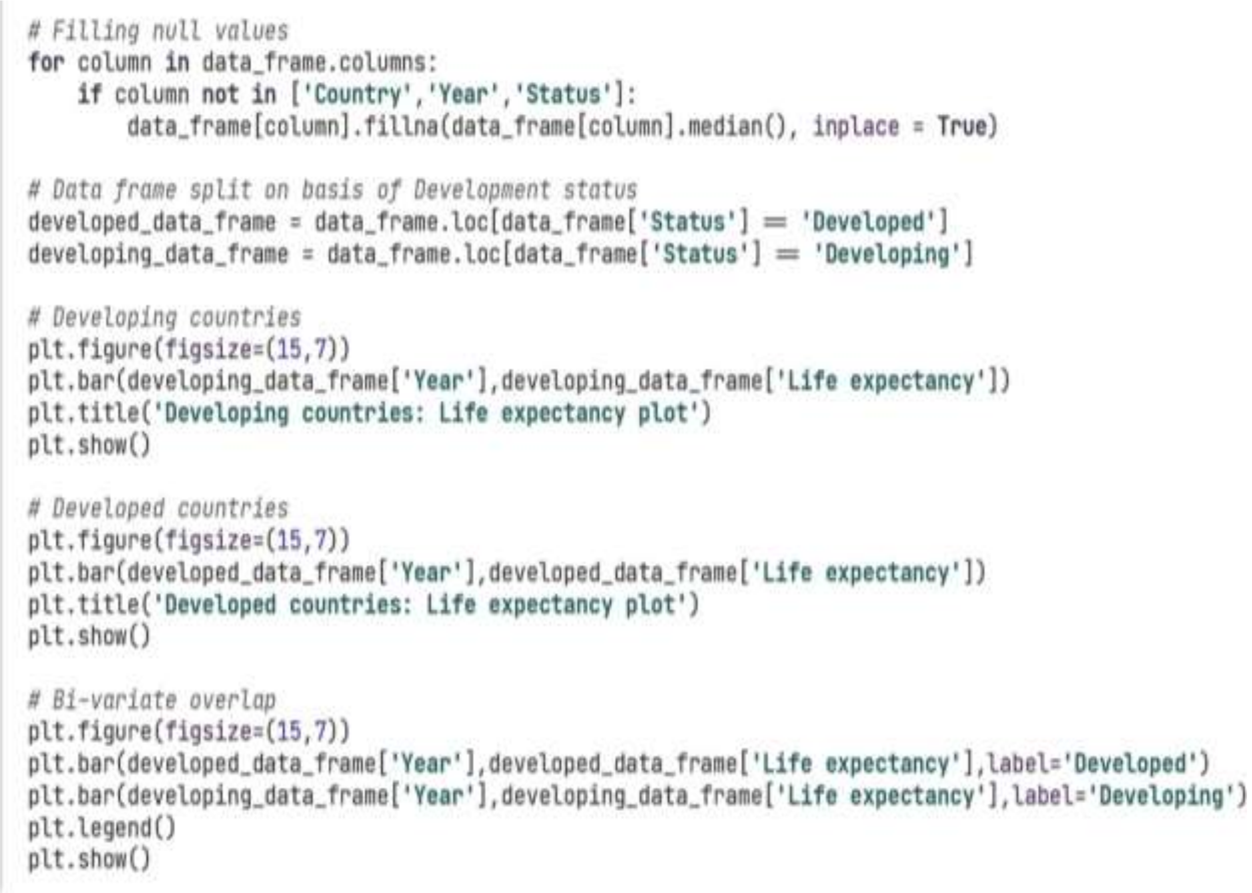
All the data we acquire is collected by actual humans through surveys and assessment of government data on us. While most of today's data may even be collected through online portals such as survey-monkey or google-forms, they are also stored in databases and there are bound to be errors in data handling. A bad dataset contains a lot of data that is either wrong or null data. And it is for this reason that we have to do something that we call 'data cleaning'. The Process might range from simple to complex depending on the number and type of error. But for the sake of simplicity that is followed throughout this manuscript, we will look at the simple end of the spectrum.

In the process, which is demonstrated as a part of the project in the next section, we first check for null values in columns. If there are any and if that particular row or column is necessary for our analysis, we ‘clean’ the column by manipulating it by filling in the null values as the mean or median or mode depending on the type of the column or else, we delete or technically ‘drop’ that row or column.

VIII. PROJECT DISCLAIMER

Many of the functions used in the following project are not from core python. They are from the packages that we import in the beginning. The project is to graphically represent the life expectancy of developing and developed countries in graphs by manipulating data in a dataset. Also, the code of the project focuses on readability and isn’t necessarily the best way to do it. This is done to encourage beginners. All of the code is written in DataLore [Appendix 5] For more information on modules, refer Appendix 3 as mentioned earlier.

IX. PROJECT CODE



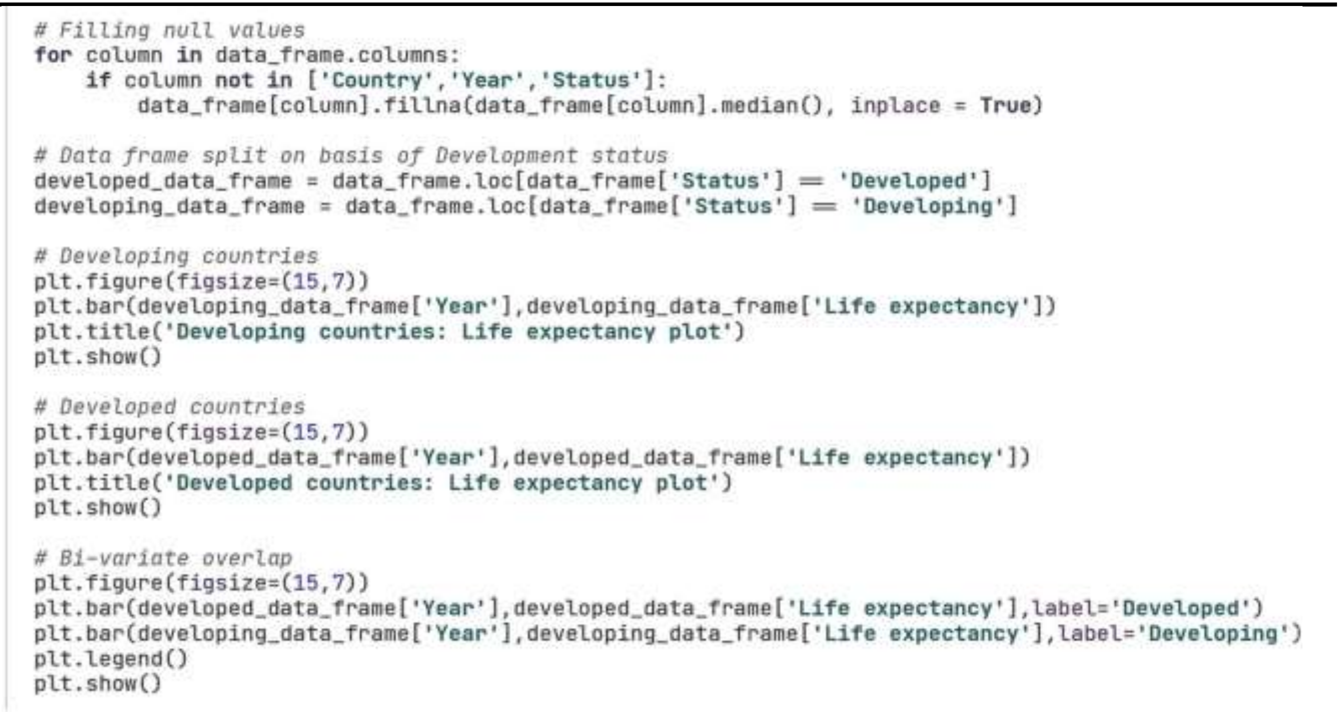
```
# Filling null values
for column in data_frame.columns:
    if column not in ['Country','Year','Status']:
        data_frame[column].fillna(data_frame[column].median(), inplace = True)

# Data frame split on basis of Development status
developed_data_frame = data_frame.loc[data_frame['Status'] == 'Developed']
developing_data_frame = data_frame.loc[data_frame['Status'] == 'Developing']

# Developing countries
plt.figure(figsize=(15,7))
plt.bar(developing_data_frame['Year'],developing_data_frame['Life expectancy'])
plt.title('Developing countries: Life expectancy plot')
plt.show()

# Developed countries
plt.figure(figsize=(15,7))
plt.bar(developed_data_frame['Year'],developed_data_frame['Life expectancy'])
plt.title('Developed countries: Life expectancy plot')
plt.show()

# Bi-variate overlap
plt.figure(figsize=(15,7))
plt.bar(developed_data_frame['Year'],developed_data_frame['Life expectancy'],label='Developed')
plt.bar(developing_data_frame['Year'],developing_data_frame['Life expectancy'],label='Developing')
plt.legend()
plt.show()
```



```
# Filling null values
for column in data_frame.columns:
    if column not in ['Country','Year','Status']:
        data_frame[column].fillna(data_frame[column].median(), inplace = True)

# Data frame split on basis of Development status
developed_data_frame = data_frame.loc[data_frame['Status'] == 'Developed']
developing_data_frame = data_frame.loc[data_frame['Status'] == 'Developing']

# Developing countries
plt.figure(figsize=(15,7))
plt.bar(developing_data_frame['Year'],developing_data_frame['Life expectancy'])
plt.title('Developing countries: Life expectancy plot')
plt.show()

# Developed countries
plt.figure(figsize=(15,7))
plt.bar(developed_data_frame['Year'],developed_data_frame['Life expectancy'])
plt.title('Developed countries: Life expectancy plot')
plt.show()

# Bi-variate overlap
plt.figure(figsize=(15,7))
plt.bar(developed_data_frame['Year'],developed_data_frame['Life expectancy'],label='Developed')
plt.bar(developing_data_frame['Year'],developing_data_frame['Life expectancy'],label='Developing')
plt.legend()
plt.show()
```

X. PROJECT OUTPUT

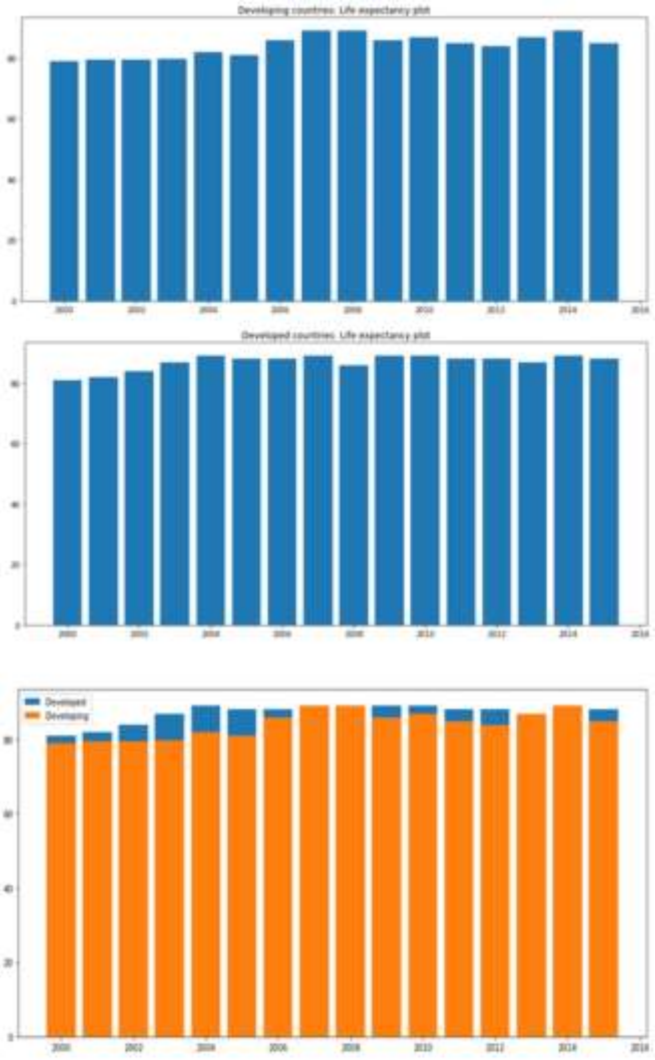


Fig 6 – Output of the project code given above

The above graphs just represent the average of many developing countries vs many developed countries in a single plot. This is done for the sake of simplicity. But it can be visualised in at least 1000 other ways. This is explained in the official code documentation of this manuscript. Refer appendix.

XI. CONCLUSIONS

Data Science is one field that takes your analysing skills and observations to the next level. This is one field that is going to be there around for a long time till automation takes place. Data science has established its utility for individuals who work on solving puzzles, creating visuals and decoding patterns. It is my personal belief that every one of us is a data scientist in some form or the other. Let us bring out the data scientist inside us and aim to acquire/conquer the power of information.

XII. APPENDIX

1-A [PRIMARY COLORS]
https://en.wikipedia.org/wiki/RGB_color_model
https://web.stanford.edu/class/cs101/image-1-introduction.html
1-B [ENCODING SCHEMES]
https://en.wikipedia.org/wiki/Character_encoding
https://medium.com/jspoint/introduction-to-character-encoding-3b9735f265a6
2-A [WHAT R DOES BETTER!]
https://www.kdnuggets.com/2017/09/python-vs-r-data-science-machine-learning.html
https://flatironschool.com/blog/r-vs-python
2-B [ADDITIONAL PYTHON]
https://www.ijraset.com/fileserve.php?FID=34185
https://www.python.org/doc/
3 [PACKAGE DOCUMENTATION]
Matplotlib - https://matplotlib.org/
Pandas - https://pandas.pydata.org/
NumPy -https://numpy.org/
SciPy -https://www.scipy.org/
Seaborn - https://seaborn.pydata.org/
4 [MACHINE LEARNING]
https://www.sas.com/en_us/insights/analytics/machine-learning.html
https://www.expert.ai/blog/machine-learning-definition/
5 [DATALORE + Alternates for IPython motebooks]
https://datalore.jetbrains.com/
https://jupyter.org/
https://colab.research.google.com/
[CODE DOCUMENTATION]
Contact me at aadhityasriram08@gmail.com to get access to all the code of this manuscript and for further clarifications.

XIII. REFERENCES

1. Aly Saber (2010) Ancient Egyptian Surgical Heritage, Journal of Investigative Surgery, 23:6, 327-334.
2. Song IY and Zhu Y. Big data and data science: What should we teach? Expert Systems, August 2016, Vol. 33, No. 4. 364-373.
3. Shcherbakov M, Shcherbakova N, Brebels A, Janovsky T, Kamaev V. JCKBSE 2014, CCIS 466, pp. 708–716, 2014.
4. Raschka S, Patterson J, Nolet C. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. Information 2020, 11, 193. 1-44.
5. Cern. Measurements of properties of the Higgs boson decaying to a Wboson pair in pp collisions at $\sqrt{s}$=13TeV. Physics Letters B791(2019)96 – 129.