Assignment-1

Define Data Science. What are the missing values and errors in data?

Noah Raphael David Yuvaraja(C0846073)

Lambton College at Cestar College of Business, Health and Technology

AML 1114: Data Science and Machine Learning

Debashish Roy

8th March 2022

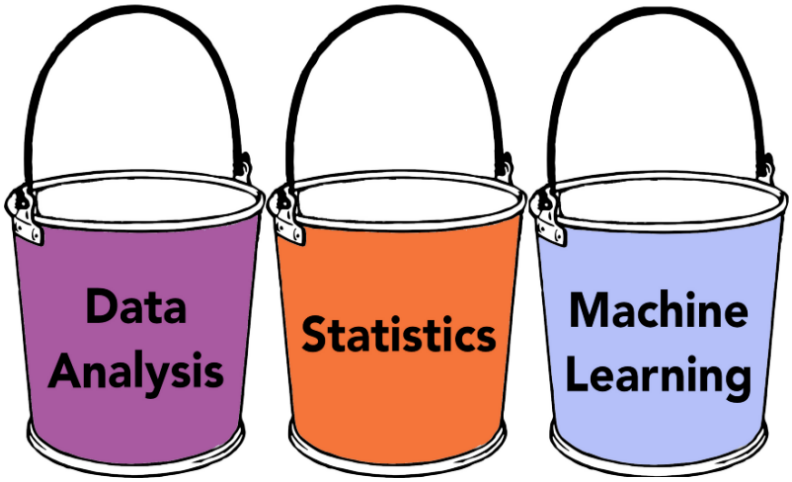
Define Data Science? Describe How Data Science Can Be Usable in Social Media and Banking and Financial Sector?

Source: (Hale, n.d.)

So, what is Data Science actually? Well, the question might sound simple but it doesn't really have a particular answer. There are various ways in which you can define Data Science and the simplest way in which I understood is within the terms "Data Science" itself. The two terms Data and Science simply emphasis Applying Scientific methods on Data! But what is the necessity behind that? Well, we will look in that in some time but before that here are the few other ways in which Data Science can be defined:

- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains - Wikipedia
- Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. - IBM
- Data science is the discipline of making data useful – Cassie Kozyrkov

If you closely observe the above definitions, even though they are written in a different way but the methods used and the end goals are almost same. Broadly, Data science can be divided into three areas i.e., Data analysis, Statistics and Machine Learning.



Source: Jeff Hale adapted from pixabay.com. Color-blind friendly palette from https://venngage.com/blog/color-blind-friendly-palette/

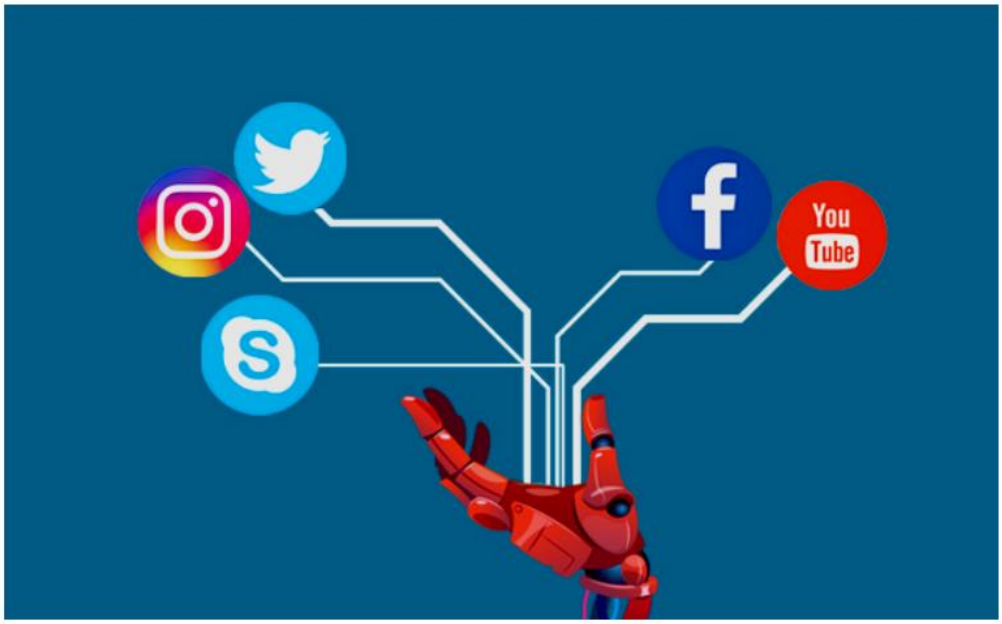
- Analysis is a technique for extracting and communicating information from existing data.
- Statistics is used to analyze and draw conclusions from data.
- Machine learning used for Predictive Modeling

Data Science is a critical component of a team that strives to improve business prospects through increased productivity, cost-cutting tactics, and a variety of other methods. The data would mean nothing unless the data collected from different sources is converted into actionable insights. Different domains in today's world rely heavily on data science to boost the value of their organisations. It's no surprise that the value of data has overtaken that of oil. Data science may aid businesses in a variety of ways, including assisting with quality decision-making in many sectors of the business, identifying the best possibilities to expand and improve, recognising trends in data, anomaly detection, recommendations, and so on.

Social Media

Source: (Active wizards, n.d.)

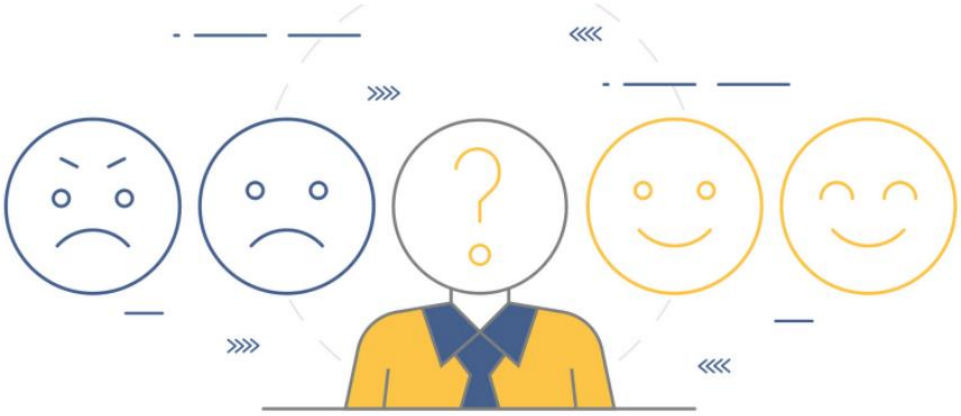
Data Science is being used by social media companies to improve user interaction and much more. For instance, let’s take a look at Instagram. Instagram, by Meta, is a popular social networking platform with a growing user base. Many users have been drawn to Insta- as millennials refer to it because of its visual appeal and aesthetics. The platform makes use of Meta's DeepText and DeepFace features, and it makes incredible use of data analytics. DeepText assists Instagram in detecting and filtering spam messages. When going through your Instagram feed and after you've viewed all of the most recent posts from accounts you follow, you may notice suggested posts. These recommendations are based on factors such as: What you've liked, saved, and remarked on, as well as who you've followed. Wonder embedding, a machine learning approach, aids Instagram in deciphering user interaction with pages and categorizes similar content and pages to provide recommendations.



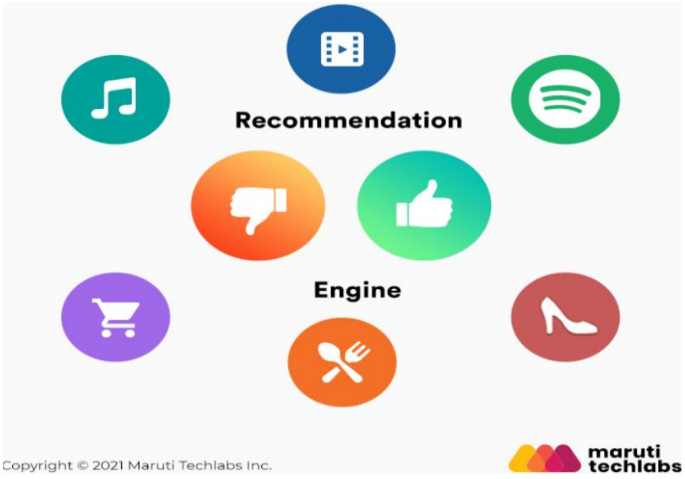
So social media in general relies on Data Science for various reasons but not limited to:

1) Customer Sentiment Analysis: All media and entertainment companies want to know how visitors react to their content, web pages, and online apps. This knowledge allows

you to tailor your presentation to the preferences of your audience. Customer sentiment analysis is commonly used for this purpose.



2) Real Time Analytics: Real-time analytics performs data processing and displays the results in incredibly short time frames. The output of real-time analytics algorithms is extraordinarily quick. As a result, critical decisions and content updates can be made right away.
3) Recommendation Engine: Recommendation engines allow entertainment and media companies to focus on the wishes and feelings of their customers. A provider pays special attention to the sensations associated with a user, in addition to the user's history inside a company.



Banking and Financial Sector

Source: (Active wizards, n.d.)

In the banking industry, data science is no longer a trend; it has become a must to stay competitive. Banks must recognise that big data technology may assist them in efficiently focusing their resources, making better decisions, and improving performance. Scotiabank and Google Cloud established a strategic agreement in 2021 to support the Bank's cloud-first strategy and accelerate its global data and analytics strategy. Google Cloud, as a Scotiabank trusted cloud partner for data and analytics, will assist Scotiabank clients in the Americas and throughout the world in creating a more personalised and predictive banking experience.

Few use cases in banking area are:

1) Fraud Detection: Machine learning is critical for detecting and preventing fraud in a variety of areas, including credit cards, accountancy, and insurance. In banking, proactive fraud detection is critical for ensuring the safety of both clients and personnel.

Image source: The Daily Star

2) Managing customer data: Digital banking is getting more and more common. Massive amounts of data must be collected, analysed, and stored by banks. Data specialists can unlock new revenue opportunities for banks by isolating and processing only the most relevant clients' information to improve business decision-making after being armed with information about customer behaviours, interactions, and preferences. With the help of accurate machine learning models, data specialists can unlock new revenue opportunities for banks by isolating and processing only this most relevant clients' information to improve business decision-making.
3) Personalized marketing: Making a personalised offer that meets the specific client's demands and preferences is the key to marketing success. We can build personalised marketing by using data analytics to provide the right product to the right person at the right time on the right device.

What are the missing values and errors in data? How can you handle missing values in data as a preprocessing step? Why is it important to handle? Explain with an example in detail.

Source: (Brownlee, n.d.)

Missing values in statistics means no value is stored for a variable in an observation and are generally encoded by -999, nan, null etc. It frequently arises when data is collected incorrectly, when there is a shortage of data or when data is entered incorrectly. Regardless of the reason, handling missing data is critical because statistical results based on a dataset with missing values may be biased. Furthermore, many machine learning algorithms do not work with input that has missing values.

There are three types of missing data

Source: (KESKES, n.d.)

1) Missing Completely and Random (MCAR) - This means that the missing values in any feature are unaffected by the values of other features. In the event of lacking data, this is the preferred scenario.

2) Missing at Random (MAR) - Missing values of one feature depends on another feature.

3) Missing Not at Random (MNAR) – Missing data are based on the missing column itself.

Experiment on Missing Values

Let’s take an example and see what is missing values and what are the different ways to handle them.

To understand this experimentally we have taken a Dataset of Melbourne Housing market from Kaggle.com (https://www.kaggle.com/anthonypino/melbourne-housing-market)

Data Description:

Suburb: Suburb
Address: Address
Rooms: Number of rooms
Price: Price in Australian dollars
Method: S - property sold; SP - property sold prior; PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed. N/A - price or highest bid not available.
Type: br - bedroom(s); h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential.

SellerG: Real Estate Agent
Date: Date sold
Distance: Distance from CBD in Kilometres
Region name: General Region (West, North West, North, North east …etc)
Property count: Number of properties that exist in the suburb.
Bedroom2: Scraped # of Bedrooms (from different source)
Bathroom: Number of Bathrooms
Car: Number of car spots
Land size: Land Size in Metres
Building Area: Building Size in Metres
Year Built: Year the house was built
Council Area: Governing council for the area
Latitude: Self explanatory
Longitude: Self explanatory

1) Importing Dataset and looking at the top five rows and columns. Also looking at the shape of Data frame to get overview on number of rows and columns in total.

```
import pandas as pd
df= pd.read_csv("melb_data.csv")
```

```
df.head()
```

	Suburb	Address	Rooms	Type	Price	Method	SellerG	Date	Distance	Postcode	...	Bathroom	Car	Landsize	BuildingArea	YearBuilt	Cou
0	Abbotsford	85 Turner St	2	h	1480000.0	S	Biggin	3/12/2016	2.5	3067.0	...	1.0	1.0	202.0	NaN	NaN	
1	Abbotsford	25 Bloomburg St	2	h	1035000.0	S	Biggin	4/02/2016	2.5	3067.0	...	1.0	0.0	156.0	79.0	1900.0	
2	Abbotsford	5 Charles St	3	h	1465000.0	SP	Biggin	4/03/2017	2.5	3067.0	...	2.0	0.0	134.0	150.0	1900.0	
3	Abbotsford	40 Federation La	3	h	850000.0	PI	Biggin	4/03/2017	2.5	3067.0	...	2.0	1.0	94.0	NaN	NaN	
4	Abbotsford	55a Park St	4	h	1600000.0	VB	Nelson	4/06/2016	2.5	3067.0	...	1.0	2.0	120.0	142.0	2014.0	

5 rows × 21 columns

```
df.shape #this line of code gives number of rows and columns
```

```
(13580, 21)
```

2) Second step is to identify missing values. This can be in done in multiple ways.

```
#identifying missing values
```

```
df.info() #from info we can see there are less non null values in Car,BuildingArea,YearBuilt and CouncilArea
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13580 entries, 0 to 13579
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Suburb         13580 non-null  object
 1   Address        13580 non-null  object
 2   Rooms          13580 non-null  int64
 3   Type           13580 non-null  object
 4   Price          13580 non-null  float64
 5   Method         13580 non-null  object
 6   SellerG        13580 non-null  object
 7   Date           13580 non-null  object
 8   Distance       13580 non-null  float64
 9   Postcode       13580 non-null  float64
 10  Bedroom2       13580 non-null  float64
 11  Bathroom       13580 non-null  float64
 12  Car            13518 non-null  float64
 13  Landsize       13580 non-null  float64
 14  BuildingArea   7130 non-null   float64
 15  YearBuilt      8205 non-null   float64
 16  CouncilArea    12211 non-null  object
 17  Lattitude      13580 non-null  float64
 18  Longtitude     13580 non-null  float64
 19  Regionname     13580 non-null  object
 20  Propertycount  13580 non-null  float64
dtypes: float64(12), int64(1), object(8)
```

```
df.isnull().sum() #this line of code gives number of missing values in each column
```

```
Suburb            0
Address           0
Rooms             0
Type              0
Price             0
Method            0
SellerG           0
Date              0
Distance          0
Postcode          0
Bedroom2          0
Bathroom          0
Car              62
Landsize          0
BuildingArea   6450
YearBuilt      5375
CouncilArea    1369
Lattitude         0
Longtitude        0
Regionname        0
Propertycount     0
dtype: int64
```

3) Why missing needs to be handled?

When a dataset has missing values, many machine learning methods fail. If missing values are not handled properly, you may end up with a biased machine learning model

that produces inaccurate results and can lead to a lack of precision in the statistical analysis.

```
#why missing values needs to be treated?
```

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
# load the dataset
df = pd.read_csv('melb_data.csv')

#for analysis purpose and easy demonstration we are removing categorical data
cat_cols = [col for col in df.columns if df[col].dtype=="O"]
df = df.drop(cat_cols, axis=1)

# split dataset into inputs and outputs
X = df.drop("Price",axis=1)
y = df.pop("Price")

#spliting data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
# define the model
reg = LinearRegression()
reg = reg.fit(X_train, y_train)
mean_squared_score = mean_squared_error(y_test, ytest_predict)
print(mean_squared_score)
```

```
ValueError: Input contains NaN, infinity or a value too large for dtype('float64').
```

Observation: As you can see from the output, Linear Regression failed to run due to too many NaN values in Dataset.

4) First method to treat missing values is Deletion.

   Rows or Columns with missing values are dropped using this method. If the number of missing values is modest, it is preferable to remove them. Although this is a simple method, it may result in a large reduction in the sample size. Furthermore, the data may not always be completely missing at random. So, estimation of parameters maybe biased.

```
#missing values treatment
```

```
#first method deletion
```

```
#removing rows with missing values
```

```
print(df.shape)
df.dropna(inplace=True)
print(df.shape)
```

```
(13580, 21)
(6196, 21)
```

```
#removing particular column that has many missing values.
```

```
df= pd.read_csv("melb_data.csv")
print(df.shape)
df.drop("YearBuilt",axis=1,inplace=True)
df.drop("CouncilArea",axis=1,inplace=True)
print(df.shape)
```

```
(13580, 21)
(13580, 19)
```

5) Second method to treat missing values is Imputation

Imputation is the process of using statistical tools to replace missing data. If the missing values is numerical, the mean of the variable can be used to impute the values and the missing values of a categorical feature could be replaced by the column's mode. The main disadvantage of this strategy is that the imputed variables' variance is reduced.

```
#second method Imputation
```

```
#imputing missing values with a value
```

```
df.fillna(df.mean(),inplace=True)
df.isnull().sum()
```

```
df.CouncilArea.fillna(df.CouncilArea.mode()[0],inplace=True)
```

```
df.CouncilArea.isnull().sum()
```

```
0
```

```
Suburb           0
Address          0
Rooms            0
Type             0
Price            0
Method           0
SellerG          0
Date             0
Distance         0
Postcode         0
Bedroom2         0
Bathroom         0
Car              0
Landsize         0
BuildingArea     0
YearBuilt        0
Lattitude        0
Longtitude       0
Regionname       0
Propertycount    0
dtype: int64
```

6) The SimpleImputer pre-processing class in the scikit-learn library can be used to replace missing values.

```
#handling missing value using skleanr module SimpleImputer
```

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer
df= pd.read_csv("melb_data.csv")
```

```
imputer = SimpleImputer(missing_values=np.NAN, strategy='mean')
```

```
df = imputer.fit_transform(df)
```

```
df.shape
```

```
(13580, 13)
```

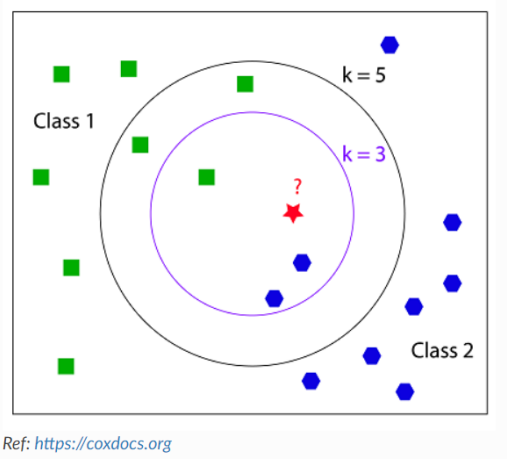
```
df = pd.DataFrame(df, columns=['Rooms', 'Price', 'Distance', 'Postcode', 'Bedroom2', 'Bathroom', 'Car',
       'Landsize', 'BuildingArea', 'YearBuilt', 'Lattitude', 'Longtitude',
       'Propertycount'])
```

```
df.isnull().sum()
```

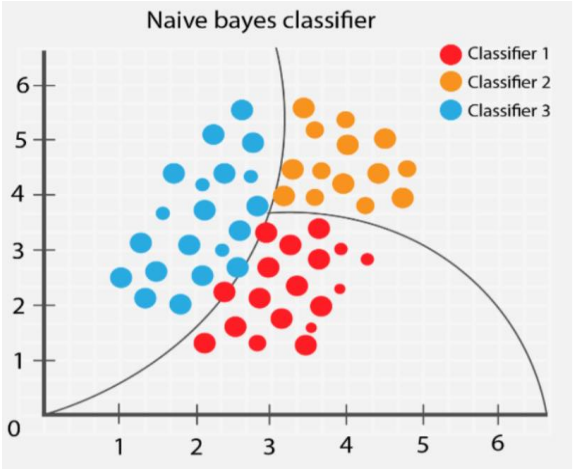
```
Rooms           0
Price           0
Distance        0
Postcode        0
Bedroom2        0
Bathroom        0
Car             0
Landsize        0
BuildingArea    0
YearBuilt       0
Lattitude       0
Longtitude      0
Propertycount   0
dtype: int64
```

7) Use Algorithms that are robust to missing values.

   Not all algorithms fail in the presence of missing values. When a value is missing, an algorithm like k-Nearest Neighbors can omit a column from a distance measure. Naïve Bayes can also handle missing values and give output.



Ref: https://coxdocs.org



References

Active wizards. (n.d.). Retrieved from https://activewizards.com/blog/top-9-data-science-use-cases-in-media-and-entertainment/

Active wizards. (n.d.). Retrieved from https://activewizards.com/blog/top-9-data-science-use-cases-in-banking/

Brownlee, J. (n.d.). *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/handle-missing-data-python/

EG, M. (n.d.). *analytics insight*. Retrieved from https://www.analyticsinsight.net/data-science-do-you-know-how-social-media-giants-leverage-it/

Hale, J. (n.d.). *Towards Data Science*. Retrieved from Towards Data Science: https://towardsdatascience.com/what-is-data-science-8c8fbaef1d37

KESKES, R. (n.d.). *Medium*. Retrieved from https://medium.com/@raoufkeskes/missing-data-its-types-and-statistical-methods-to-deal-with-it-5cf8b71a443f

Kinha, Y. (n.d.). *kdnuggets*. Retrieved from https://www.kdnuggets.com/2020/06/missing-values-dataset.html

Scotiabank. (n.d.). *News wire*. Retrieved from https://www.newswire.ca/news-releases/scotiabank-partners-with-google-cloud-to-create-more-personalized-and-predictive-banking-experiences-890582473.html

tecHindustan. (n.d.). *Medium*. Retrieved from https://medium.com/@techindustan/artificial-intelligence-in-social-media-edc04fc5f8d0