Demystifying
AI, ML, Data Science

Date : 7th June 2020 | Speaker : Ayon Roy |
Event : Webinar by IIT ISM Dhanbad

Visit - AYONROY.ML

Hello **Buddy!**

I am **Ayon Roy**

B.Tech CSE (2017-2021)

Data Science Intern @ **Lulu International Exchange**, Abu Dhabi

(**World's Leading Financial Services Company**)

Brought **Kaggle Days Meetup** Community in India for the 1st time

If you haven't heard about me yet, you might have been living under the rocks. Wake up !!

Agenda (7-6-2020)

- Brief discussion on Career options
- How to start Machine Learning ?
- A brief Intro to Data Pre-Processing, Exploratory Data Analysis, Data Visualization
- How to do projects in ML ?
- Why should you start Kaggle ?
- How to crack Internships ?

All the Resource links for the discussion will be shared with you at the end of this webinar by me.

What's the Current Scenario ?

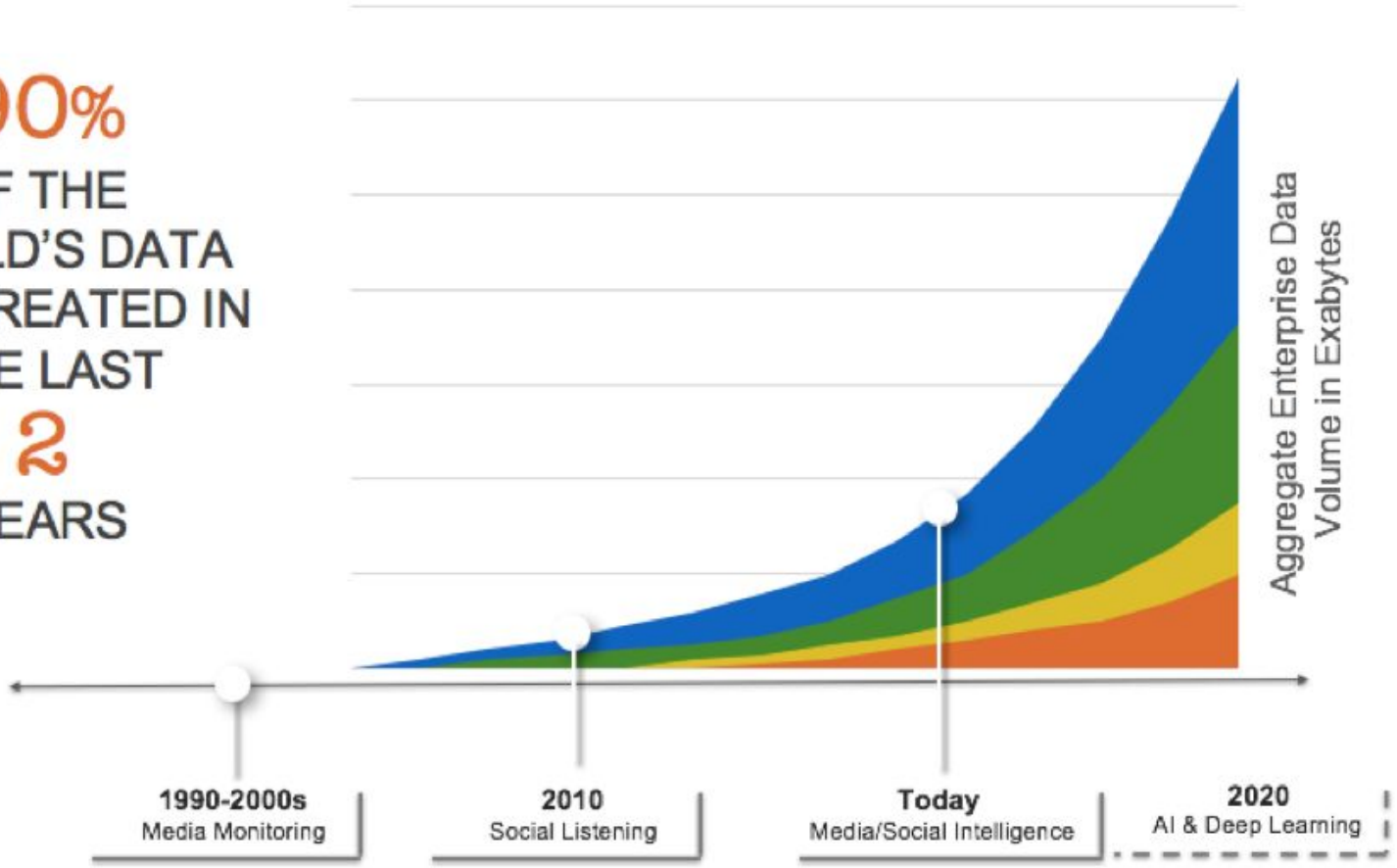
Data is increasing exponentially

90%

OF THE WORLD'S DATA WAS CREATED IN THE LAST

2

YEARS



Focus on Optimal Applications, User Experience is also increasing exponentially

INCREASED FOCUS ON USER EXPERIENCE

DIGITAL TRENDS

Companies know they need to improve user experience. It's an obvious pain point for over 97% of the top 1,500 websites. Improving user experience, though, is an involved task not easily undertaken.

The first step in improving user experience is to understand the user. Developing consumer personas is a process in and of itself that requires key insights and good resources. With well thought-out personas, though, discovering the pitfalls in a user's journey becomes much easier and the process is much smoother.

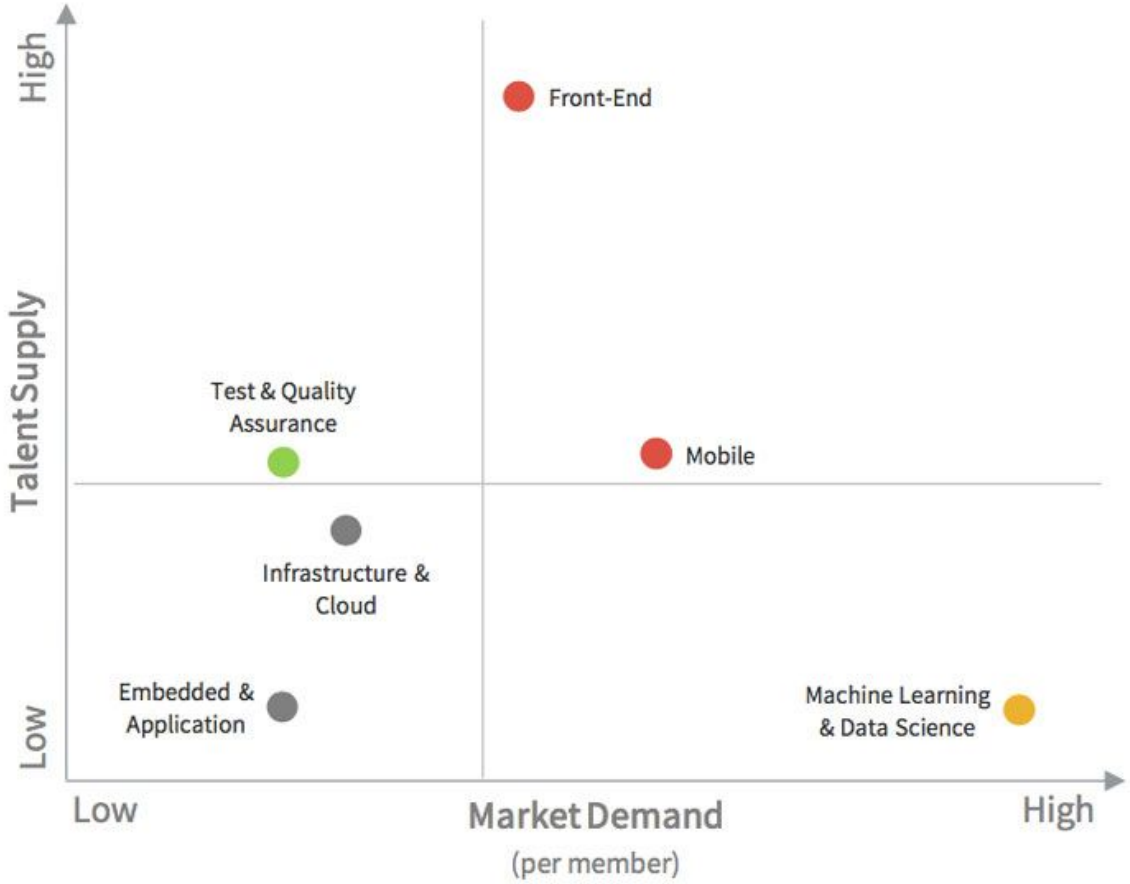
COMPANIES FOCUSING ON USER EXPERIENCE

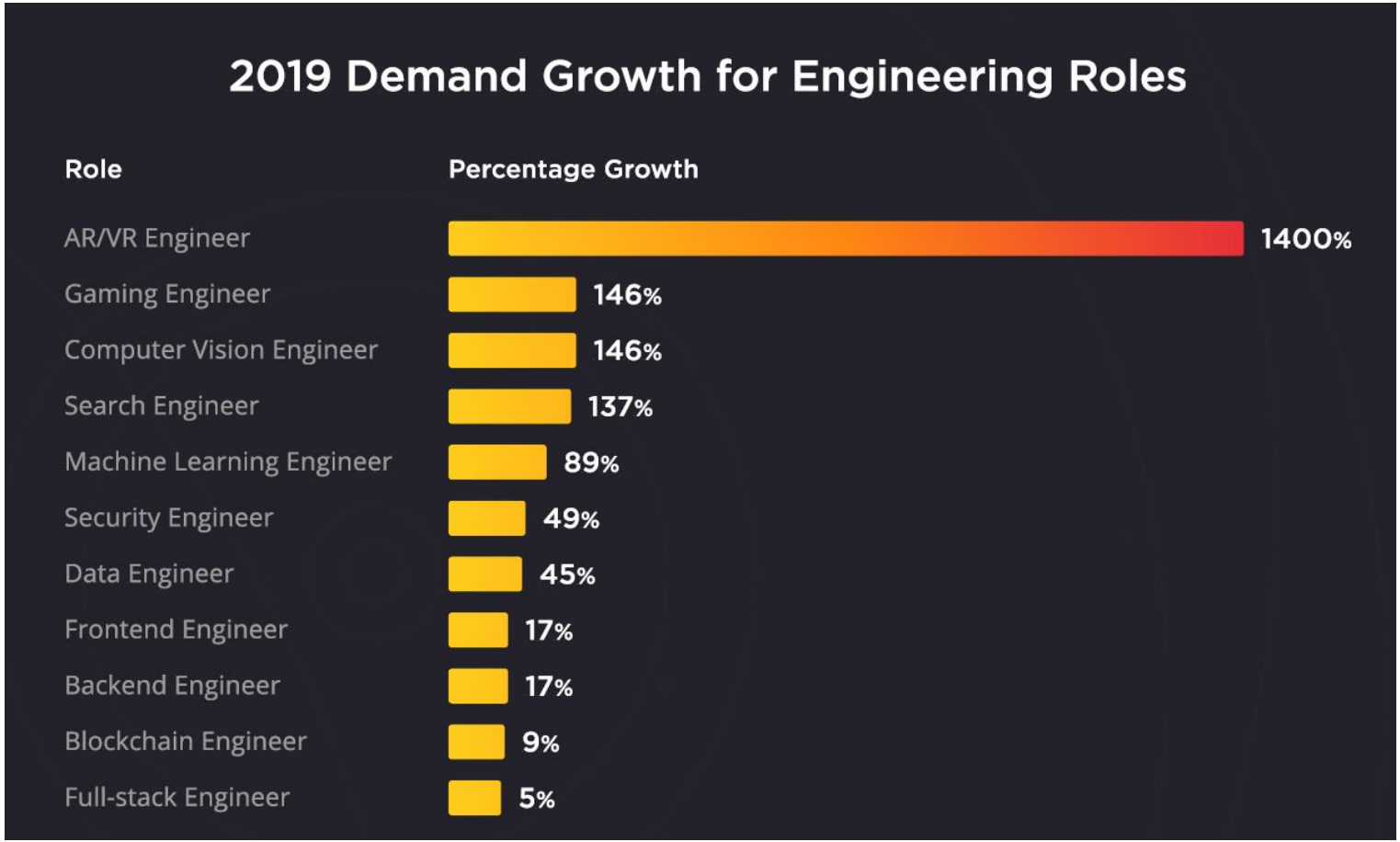
of companies that don't conduct UX testing plan to do so in next 12 months

of companies plan to boost focus on customer experience metrics

of sites fail at User Experience

Supply & Demand by Specialty





What to do now ?

In the end, it all just boils down to your personal preference and interest

If you **like creating things and building algorithms** that have a set outcome where you know what to expect, then **competitive programming is right for you**.

But if you **like the unpredictable, are in love with statistics and trends**, and have innate business acumen, then **Machine Learning is right for you**.

How to start Machine Learning

Visit - AYONROY.ML

Start with

Maths for Machine Learning

Visit - AYONROY.ML

But why should I do Maths first for Machine Learning ?

- Week 1 : Linear Algebra [B] https://www.khanacademy.org/math/linear-algebra
- Week 2 : Calculus [B] https://www.youtube.com/playlist?list=PLZHQObOWTQDMsr9K-rj53DwVRMYO3t5Yr or https://www.mathsisfun.com/calculus/ ; want theoretical notes , find it at https://the-learning-machine.com/article/machine-learning/calculus .
- Week 3 : Probability [B] https://www.edx.org/course/introduction-probability-science-mitx-6-041x-2
- Week 4 : Statistics [B] http://alex.smola.org/teaching/cmu2013-10-701/stats.html
- Algorithms (Only if you want to learn proper software development) [Highly optional]
  This is an overview of what the students study as the subject Data Structures & Algorithm . So if you are fluent with this part , you can skip this !! https://www.edx.org/course/algorithm-design-analysis-pennx-sd3x

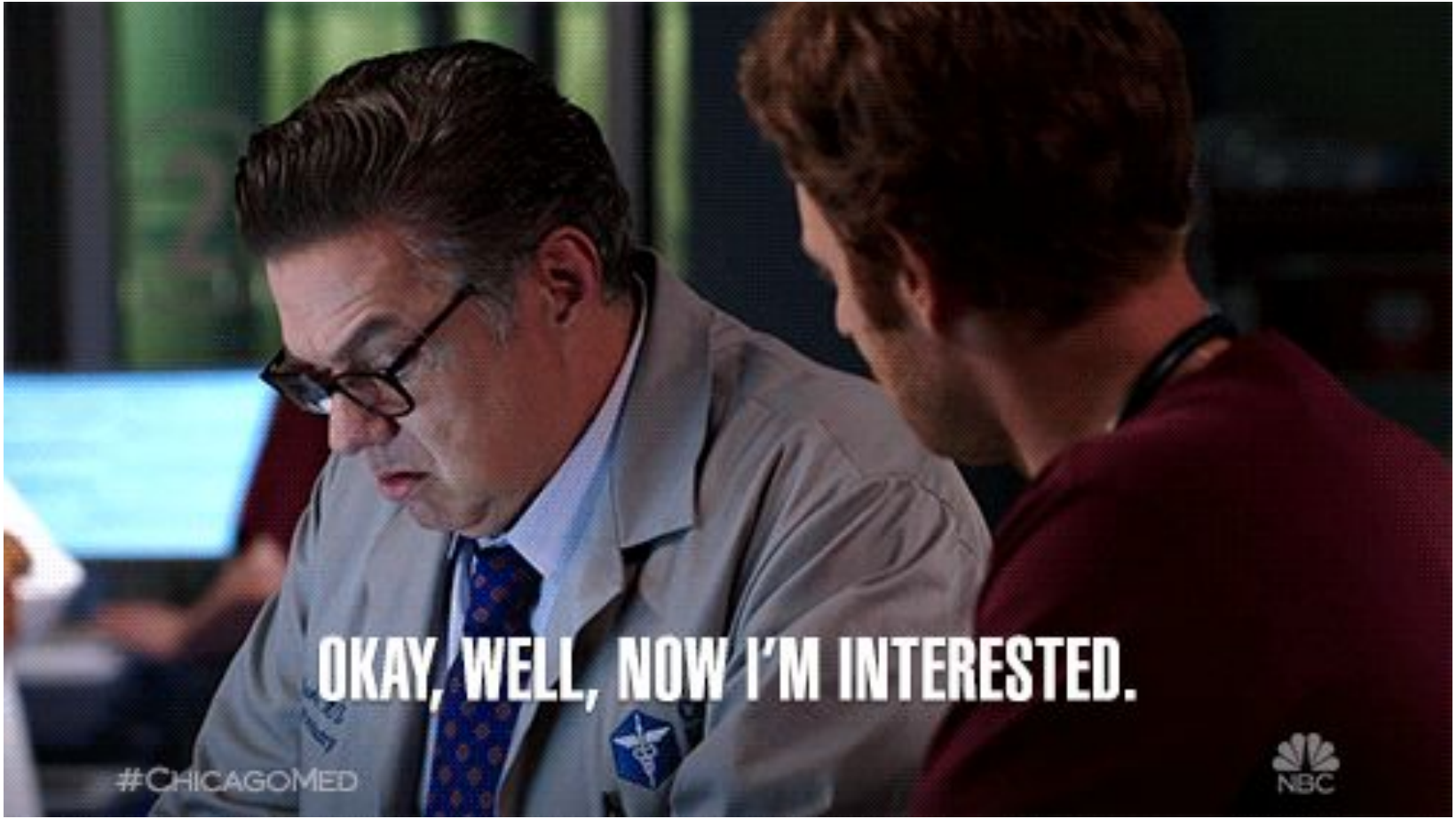
Start with Python

&

try to implement those

Mathematical Concepts

Start exploring Libraries & then start Machine Learning Courses

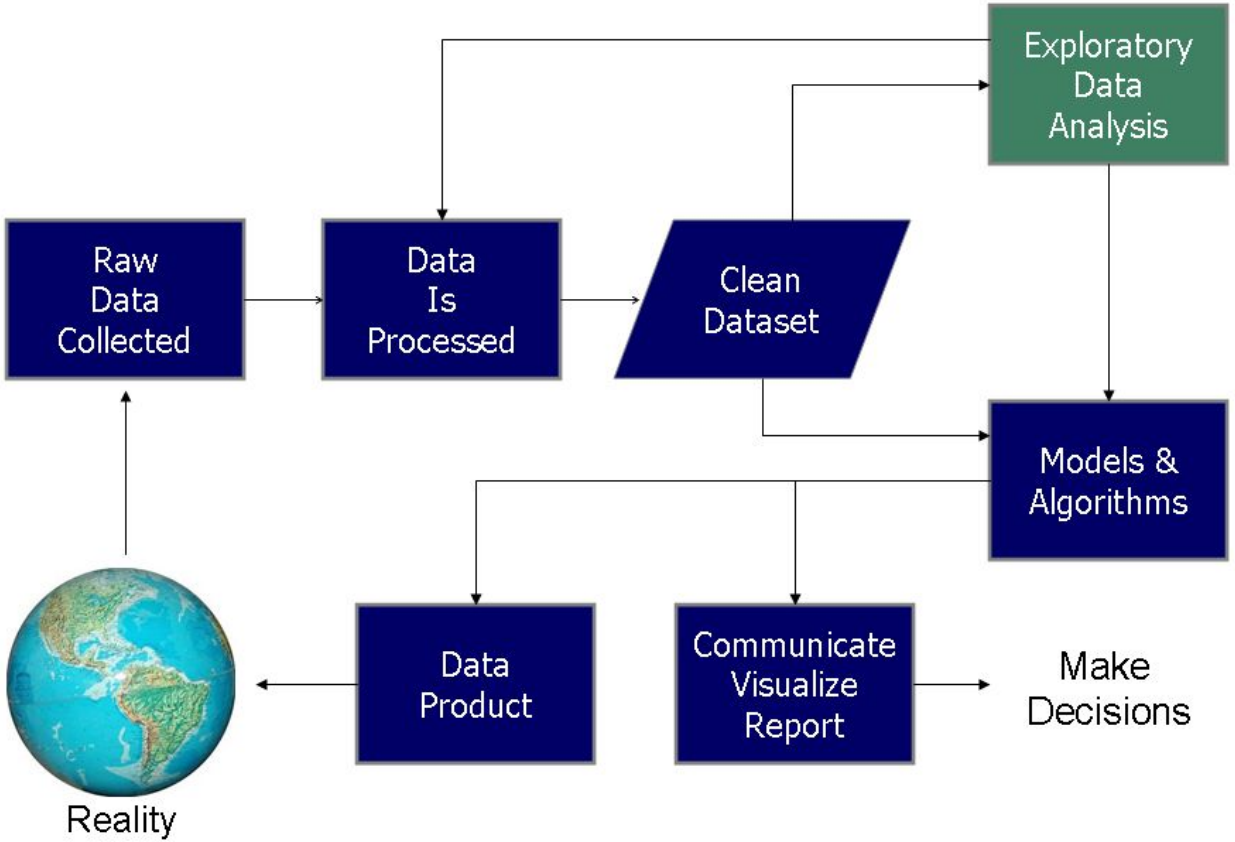


- Introduction to python for data science [B] https://www.datacamp.com/courses/intro-to-python-for-data-science
- Want to dive deeper into Data Visualization & Pre-Processing ? Look into Data Visualization & Pre-Processing section in miscellaneous resources . [Highly optional]
- Want to explore the field of Deep Learning ? See the Deep Learning Section in miscellaneous resources . [Highly optional]
- Want to explore the field of Natural Language Processing [NLP } ? See the Natural language Processing Section in miscellaneous resources . [Highly optional]
- See how ML codes are written and made to work at - > https://github.com/maykulkarni/Machine-Learning-Notebooks or https://github.com/GokuMohandas/practicalAI/blob/master/README.md . [Highly optional]
- Find useful resources here at https://github.com/ujjwalkarn/Machine-Learning-Tutorials/blob/master/README.md . [Highly optional]

Don’t rush behind completing Courses & add them to Resume

Understand the concepts well before starting Projects

Data Science Process



What is

Data Pre-Processing ?

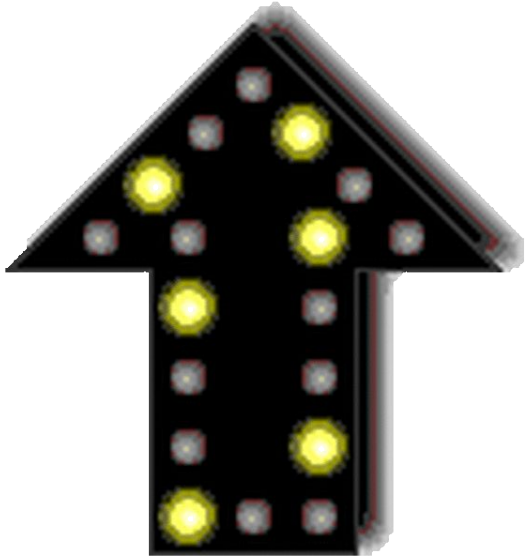
It is a technique that transforms raw data into an understandable format.

Why do we need it ?

Raw data (Real world data) is always messy and that data cannot be sent through a model. That would cause certain errors.

So we need to preprocess data before sending through further analysis.

Steps to be followed



Get the data & Import the Libraries

```
# main libraries
import pandas as pd
import numpy as np
import time


# visual libraries
from matplotlib import pyplot as plt
import seaborn as sns
from mpl_toolkits.mplot3d import Axes3D
plt.style.use('ggplot')


# sklearn libraries
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import normalize
from sklearn.metrics import
confusion_matrix,accuracy_score,precision_score,recall_score,f1_score
,matthews_corrcoef,classification_report,roc_curve
from sklearn.externals import joblib
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
```

Read the data

```
# Read the data in the CSV file using pandas
df = pd.read_csv('../input/creditcard.csv')
df.head()
```

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.12
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.16
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.32
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.64
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.20

Fig 1 : Dataset

Checking the Missing Values

```
# Looking at the ST_NUM column
print df['ST_NUM']
print df['ST_NUM'].isnull()


Out:
0    104.0
1    197.0
2      NaN
3    201.0
4    203.0
5    207.0
6      NaN
7    213.0
8    215.0

Out:
0    False
1    False
2     True
3    False
4    False
5    False
6     True
7    False
8    False
```

Replacing the Missing Values

A very common way to replace missing values is using a median.

```
# Replace using median
median = df['NUM_BEDROOMS'].median()
df['NUM_BEDROOMS'].fillna(median, inplace=True)
```

Standardizing the data

```
# Standardizing the features
df['Vamount'] =
StandardScaler().fit_transform(df['Amount'].values.reshape(-1,1))
df['Vtime'] =
StandardScaler().fit_transform(df['Time'].values.reshape(-1,1))

df = df.drop(['Time','Amount'], axis = 1)
df.head()
```

V22	V23	V24	V25	V26	V27	V28	Class	Vamount	Vtime
0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	0	0.244964	-1.996583
-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	0	-0.342475	-1.996583
0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	0	1.160686	-1.996562
0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	0	0.140534	-1.996562
0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	0	-0.073403	-1.996541

Fig 7 : Standardized dataset

Data Splitting

```
# splitting the faeture array and label array keeping 80% for the
trainnig sets
X_train,X_test,y_train,y_test =
train_test_split(feature_array,label_array,test_size=0.20)

# normalize: Scale input vectors individually to unit norm (vector
length).
X_train = normalize(X_train)
X_test=normalize(X_test)
```

Exploratory Data Analysis

Visit - AYONROY.ML

What is Exploratory Data Analysis ?

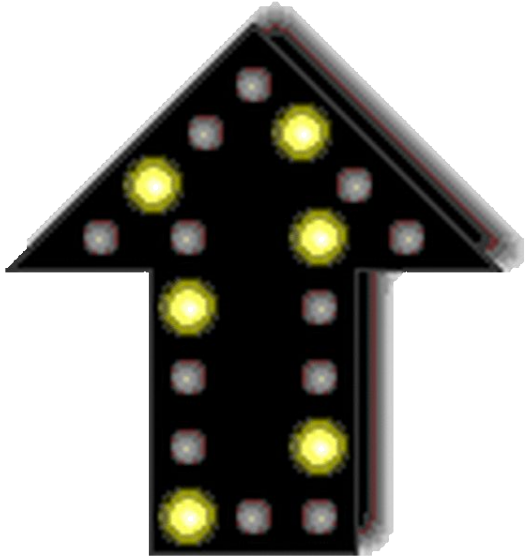
A critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Why do we need it ?

1. Detection of mistakes & missing data
2. Checking of assumptions
3. Preliminary selection of appropriate models
4. Determining relationships among the explanatory variables

With EDA, we can make sense of the data we have and then figure out what questions we want to ask and how to frame them

Major Steps to be followed



Import the Libraries

```
# Importing required libraries.
import pandas as pd
import numpy as np
import seaborn as sns #visualisation
import matplotlib.pyplot as plt #visualisation
%matplotlib inline
sns.set(color_codes=True)
```

Check the type of Data

```
# Checking the data type
df.dtypes
```

```
Make                object
Model               object
Year                 int64
Engine Fuel Type    object
Engine HP          float64
Engine Cylinders   float64
Transmission Type   object
Driven_Wheels       object
Number of Doors    float64
Market Category     object
Vehicle Size        object
Vehicle Style       object
highway MPG          int64
city mpg             int64
Popularity           int64
MSRP                 int64
dtype: object
```

Dropping Irrelevant Columns

```
# Dropping irrelevant columns
df = df.drop(['Engine Fuel Type', 'Market Category', 'Vehicle Style',
'Popularity', 'Number of Doors', 'Vehicle Size'], axis=1)
df.head(5)
```

	Make	Model	Year	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	highway MPG	city mpg	MSRP
0	BMW	1 Series M	2011	335.0	6.0	MANUAL	rear wheel drive	26	19	46135
1	BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	19	40650
2	BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	20	36350
3	BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	18	29450
4	BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	18	34500

Dropping irrelevant columns.

Renaming the Columns

```
# Renaming the column names
df = df.rename(columns={“Engine HP”: “HP”, “Engine Cylinders”:
“Cylinders”, “Transmission Type”: “Transmission”, “Driven_Wheels”:
“Drive Mode”,”highway MPG”: “MPG-H”, “city mpg”: “MPG-C”, “MSRP”:
“Price” })
df.head(5)
```

	Make	Model	Year	HP	Cylinders	Transmission	Drive Mode	MPG-H	MPG-C	Price
0	BMW	1 Series M	2011	335.0	6.0	MANUAL	rear wheel drive	26	19	46135
1	BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	19	40650
2	BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	20	36350
3	BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	18	29450
4	BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	18	34500

Renaming the column name.

Removing the Duplicates

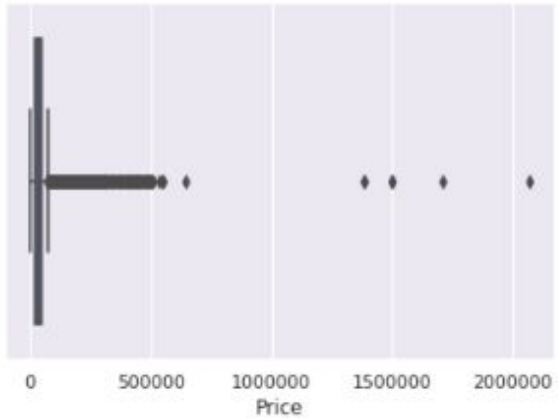
```
# Dropping the duplicates
df = df.drop_duplicates()
df.head(5)
```

	Make	Model	Year	HP	Cylinders	Transmission	Drive Mode	MPG-H	MPG-C	Price
0	BMW	1 Series M	2011	335.0	6.0	MANUAL	rear wheel drive	26	19	46135
1	BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	19	40650
2	BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	20	36350
3	BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	18	29450
4	BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	18	34500

Detecting the Outliers

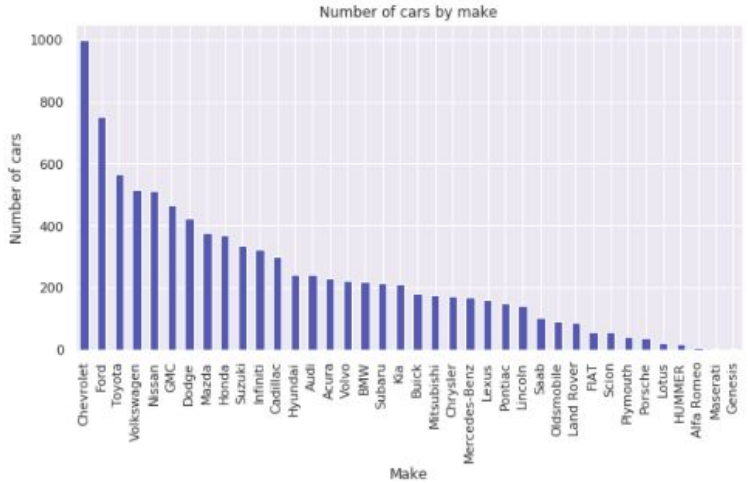
```
sns.boxplot(x=df['Price'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f69f68edc18>
```



Plotting different features

```
# Plotting a Histogram
df.Make.value_counts().nlargest(40).plot(kind='bar', figsize=(10,5))
plt.title("Number of cars by make")
plt.ylabel('Number of cars')
plt.xlabel('Make');
```



Histogram

Correlation Matrix etc.

Visit - AYONROY.ML

What's Data Visualization ?

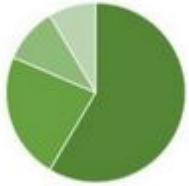
Data visualization is the graphical representation of information and data.

By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Difft. Types of Data Visualization methods

Charts ->

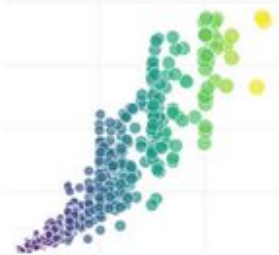
Line



Pie

Bar

Plots ->



Bubble

Scatter

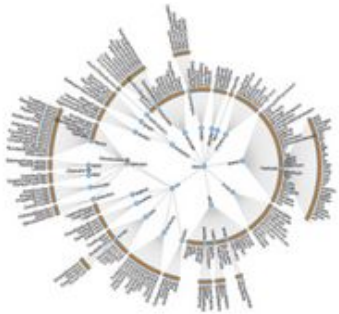
Difft. Types of Data Visualization methods

Maps ->

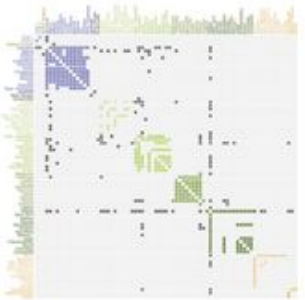
Heat

Dot distribution

Trees & Matrix ->



Tree



Matrix

But

What projects to start with

?

- Beginners Section [B] : Brush your basic concepts and revise them to start doing projects

Titanic Dataset

Iris Dataset

Stock Price Prediction

Stores Sales Forecasting

Housing Price Prediction

Guide for Beginner Projects:

First of all see Below 2 videos to get an idea on how to make projects of Data Science and Machine Learning And then Move to Kaggle for Making your own project.Its is Good if you Make Minimum 2-3 Projects on your own.

- Titanic Survivor : https://www.youtube.com/watch?v=fS70iptz-XU&t=
- Credit Card Fraud Detection : https://www.youtube.com/watch?v=gCWBFyFTxVU

Intermediate & Advanced Section

- Learn libraries like Opencv , Tensorflow , SkLearn

1) Natural Language Processing : MNIST Handwritten Digit Classification , Twitter Sentiment Analysis

2) Email Spam Classifier

3) Fraud Detection System

4) Computer Vision : Face Recognition , Face Detection

Ayon Roy

Speaker 🏆 Let's talk ML, AI, Data Science, DL, Python 🏆 Catch me @ Hackat...

1w •

3 Major types of projects you should do if you are just diving into **#datascience**, **#machinelearning**, **#artificialintelligence**. Here are a few pointers :

For Exploratory Data Analysis (EDA) Projects -
Practice on the dataset at
- **https://lnkd.in/gztCfy3**
- **https://lnkd.in/gFasqNi**
- **https://lnkd.in/grvF-jc**
- **https://lnkd.in/gPsxf5y**
- **https://lnkd.in/gDKuhEf**
- **https://lnkd.in/g_SRS7F**

For Prediction Modelling Projects -
Practice on the dataset at
- **https://lnkd.in/gQh6SRZ**
- **https://lnkd.in/g5JfbeA**
- **https://lnkd.in/gPG6Wgf**
- **https://lnkd.in/gYBE6DY**

For Data Visualization Projects -
Practice on the dataset at
- **https://lnkd.in/gWZJ3TZ**
- **https://lnkd.in/gih7YDd**
- **https://lnkd.in/gcv2xar**

" I am a beginner in AI,ML,Data Science & trying to do projects; but not succeeding as I get stuck more often "

Here is my way ahead if you are facing the same.

" Start Simple Projects & Be Motivated "

We all usually want to do the best projects & showcase them in our resume & hence sometimes end picking up a complex project at the first go. But do understand that while it's very normal to pick complex projects as a beginner, because we can't analyze the scale of project at first go. Picking up a complex project at a first sight may demotivate you as they have a lot of details, requires lot of studies to progress, thus a beginner ends up leaving the project midway & be traumatized. So start your journey with Simpler & Smaller projects as they require comparatively less details & can be achieved over a short period of time, thus helping us to stay motivated & keep doing projects. And as the learning in AI, machinelearning, datascience never stops, so as we get motivated with completion of small projects; we learn & practice more while increasing the complexity of our upcoming projects. Still waiting to start ? Start today !! All the best !!

What is Kaggle & why to start it ?

Kaggle is an ecosystem for doing and sharing Data Science stuffs.

Kernels is a cloud computational environment that enables reproducible and collaborative analysis.

Datasets is what you need. Whether it is CSVs/tabular data, or you are looking for an image data, or you are interested in NLP and speech and looking for those datasets, everything is here.

Discussions is a great place to ask questions, answer questions and interact with the community.

Competitions involve a very friendly community that is there to work on a problem statement by using start of the art concepts.

Why you should start Kaggling?

1. **Steep learning curve:** Participating on Kaggle provides much more wider range of learning than you will find anywhere else online. People come up with great ideas and share the ideas in a public kernel.

2. **Variety in problem sets:** You will easily find data related to almost every field in data science and machine learning on Kaggle. One single place to get your hands dirty in every area.

3. **An awesome community:** The true power of Kaggle comes from its community. The people there are really helpful and they try to help you in every aspect without judging you for a second.

Where to get guided for starting with Kaggle?

This tab contains **free, practical, hands-on courses that cover the minimum prerequisites needed to quickly get started in the field**. The best thing about them?—everything is done using Kaggle's kernels (described above). This means that you can interact and learn

How to
Crack Internships ?

But

No CONTENT found !!

As it will be a DISCUSSION

" I am doing Online courses & learning techniques in ML,AI, Data Science ; but still I can't bag an internship 😔 "

Here is something that may boost your strategy to bagging an internship 🚀

.... Usual Process - -

You do online courses to further deepen your knowledge

- You earn a certificate of knowledge

- You go to university - You learn computer science

- You learn data science tools and techniques

--> This is not working so great!

Now, let's follow this process

- - You research job postings (Like datascience / machinelearning Internships)

- You become your own intern by doing a project that uses skills and technologies from those job postings

- You create your own certificate of knowledge through documenting/showcasing your work and building a professional profile. [View mine at https://ayonroy.ml/]

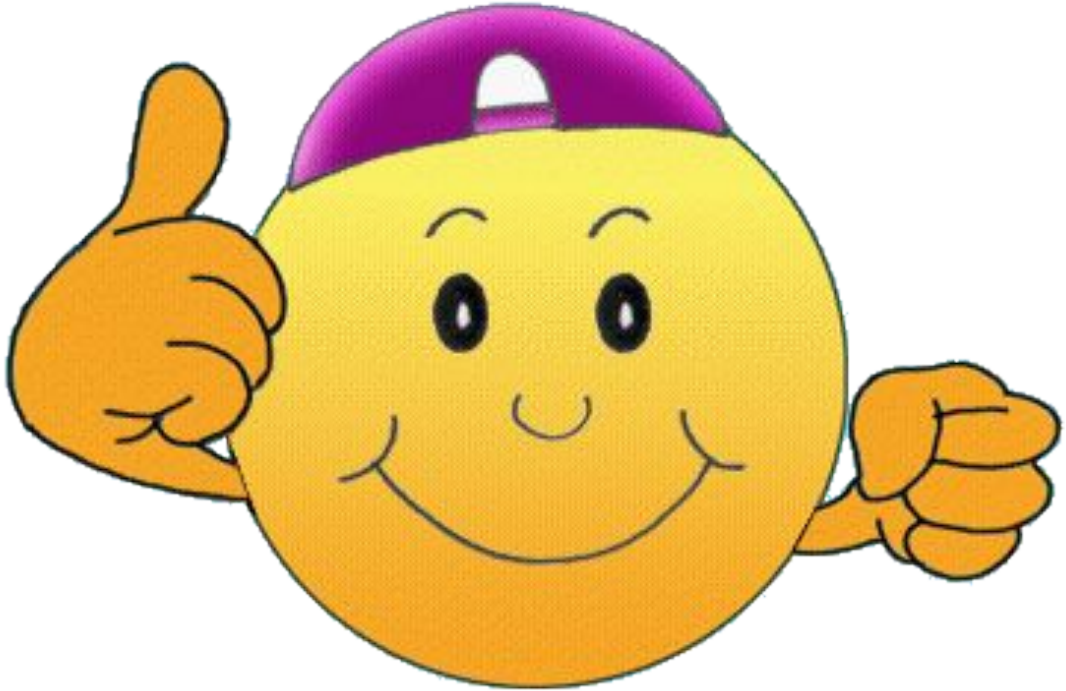
- You apply to those researched jobs. But make sure you are showing off your skills beforehand. And I think you will soon bag an internship in your domain of interest.

Get the resources at

1. https://github.com/ayonroy2000/100DaysOfMLCode
2. http://bit.do/aroy
3. https://www.linkedin.com/in/ayonroy2000/

Visit - AYONROY.ML

GO FOR IT !



GOOD LUCK !

Let me answer your Questions now.

Finally, it’s your time to speak.

Danke Scheon

Questions ? Any Feedbacks ? Did you like the talk?
Tell me about it.

If you think I can help you,
connect with me via

Email : ayonroy2000@gmail.com

LinkedIn / Github / Telegram Username : ayonroy2000

Website : https://AYONROY.ML/