Who am I ?

13 June 2023 07:01

Beginner → Intermediate #(Python)

AKASH PUSHKAR CHARAN

Quick witted | Tech-Savvy | Observer

- Machine Learning Enthusiast
- IIT Kanpur Alumnus
- Working as Lead Data Scientist with Accenture Strategy & Consulting

An instructor by passion and data scientist by profession who always tries to look at the problem in a different way. I started teaching when I was in first year of my graduation which amounts to **10+** years of experience.

"It's not who I am underneath but what I do that defines me"

started with Mathematics 2008.

↓

GATE|CAT

↓

Data Science

~15 yrs.

PROVIDE LINKEDIN PROFILE

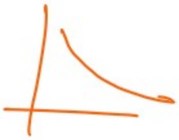
https://www.linkedin.com/in/akash-pushkar-04642925/

Setting the ground rules & expectations

13 June 2023 07:09

Ground rules:

1. Sheer focus, attention & interaction → doubts ↳ ask questions
   Mobile phone → DND



   - Two windows ↪ after break ↳ end of the session.

2. **Revision is the key to success.**
   ↳ Always revise all the notes (OneNote) & notebooks before you attend the next session

3. 4 times a week.
   In case, you happen to miss a class, pls cover the topics using LMS.

4. Real life application ⇒ Industrial examples.
   → (learning/experience from my past projects)

5. **Everyone will code.**

6. Assessment questions
   - → H/W
   - → QIC : Question in class
   - → Reading assignment ✓
   - → case-studies

7. Breaks # (10-15 mins)

Course Structure

13 June 2023 07:26

Pre-requisites: Python Basics
- → Conditional statements
- → Looping statements
- → Functions / Lambda functions

I

Module #1 NumPy.

Module #2 Pandas

} For data manipulation activities ⇒ EDA: Exploratory Data Analysis

- ↳ 2.0 has been released.

Module #3 Data visualization using Matplotlib
- → MATLAB style
- → Object oriented style
- ↳ seaborn
- ↳ plotly / Altair # Enterprise visualization

Storytelling with data.

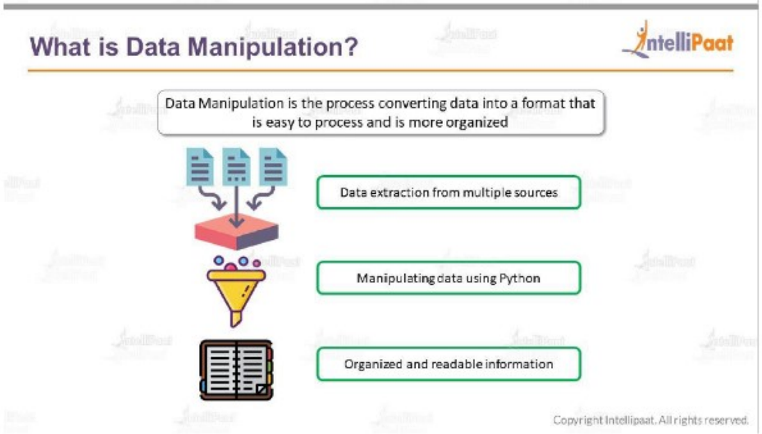
Module #4 Probability & statistics & a bit of maths : foundation for ML / DS.
- Descriptive
- Inferential

Module #5 Introduction to ML
- → Supervised
- → Unsupervised.

Module #1 - Data Manipulation - NumPy

13 June 2023 07:36

What is data manipulation?

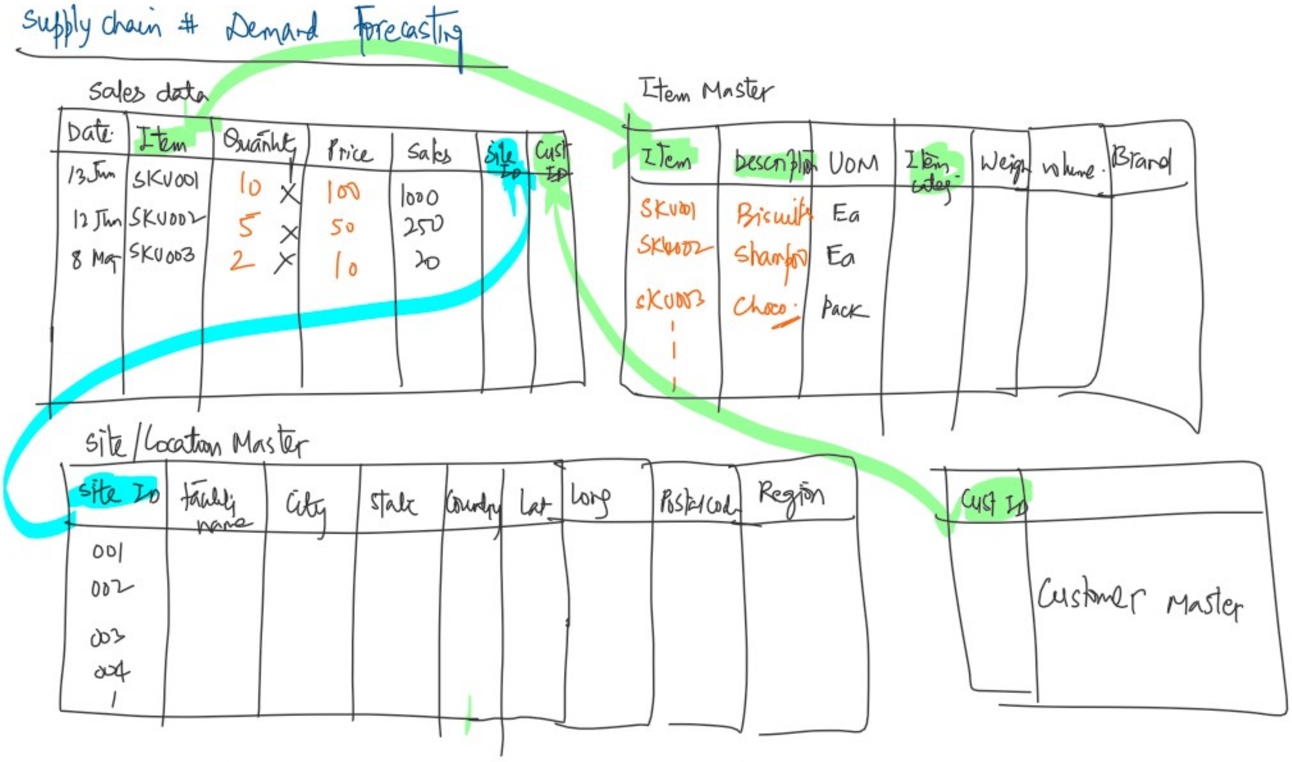


What is Data Manipulation?

Data Manipulation is the process converting data into a format that is easy to process and is more organized

- Data extraction from multiple sources
- Manipulating data using Python
- Organized and readable information

Supply chain # Demand Forecasting



Sales data

Date	Item	Quantity	Price	Sales	Site ID	Cust ID
13 Jun	SKU001	10 ×	100	1000		
12 Jun	SKU002	5 ×	50	250		
8 May	SKU003	2 ×	10	20		

Item Master

Item	Description	UOM	Item categ	Weigh	volume	Brand
SKU001	Biscuits	Ea				
SKU002	Shampoo	Ea				
SKU003	Choco.	Pack				

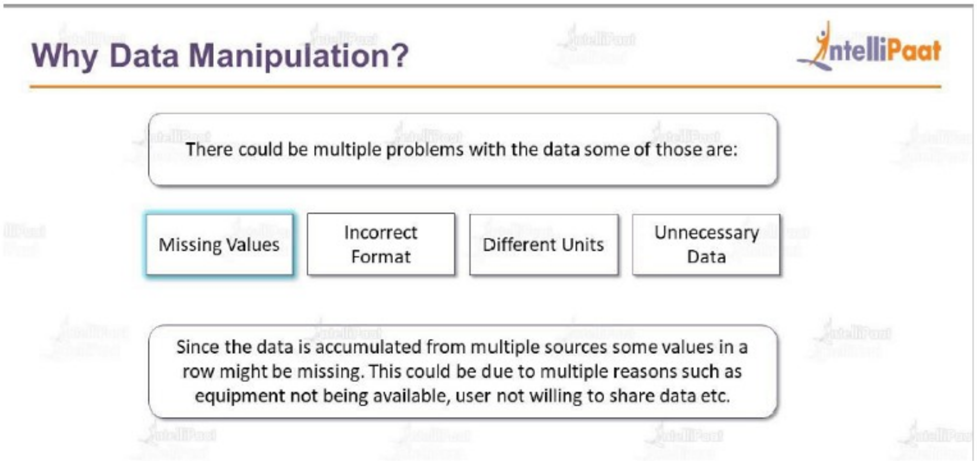
Site/Location Master

Site ID	Facility name	City	State	Country	Lat	Long	Postal Code	Region
001								
002								
003								
004								

Cust ID	
	Customer Master

Facts Table → Transaction Table (sales data)

Dimension Table → Master Files (Item | site | customer)



Why Data Manipulation?

There could be multiple problems with the data some of those are:

Missing Values | Incorrect Format | Different Units | Unnecessary Data

Since the data is accumulated from multiple sources some values in a row might be missing. This could be due to multiple reasons such as equipment not being available, user not willing to share data etc.

Missing Values Treatment

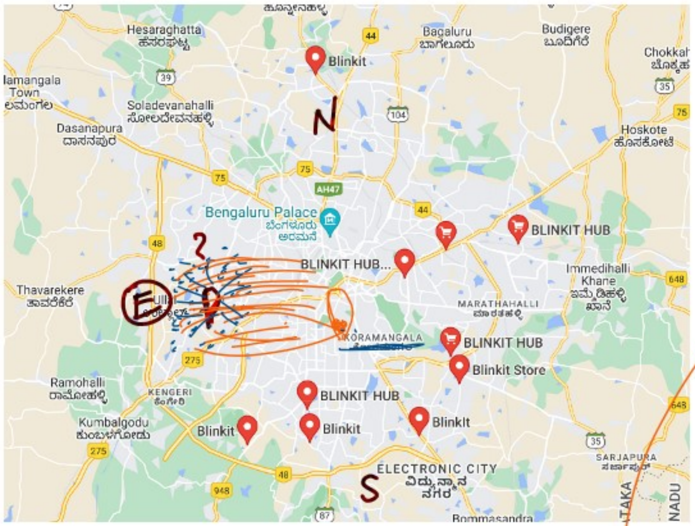
Location | site Master

Site ID	City	Address	State	Country	Region	Postal code	Lat.	Long.
loc001	BLR	-	KA	IN	South	560xxx	—	—
loc002	"	-	"	.	South	"	—	—
loc 003	"		"	'	East	~~560055~~		
'							—	—
|	'		'				—	—
			|	|				
loc 100	"		"		West	560023	—	—

How will you treat missing Lat & long

a) Drop the row(s)/column(s)

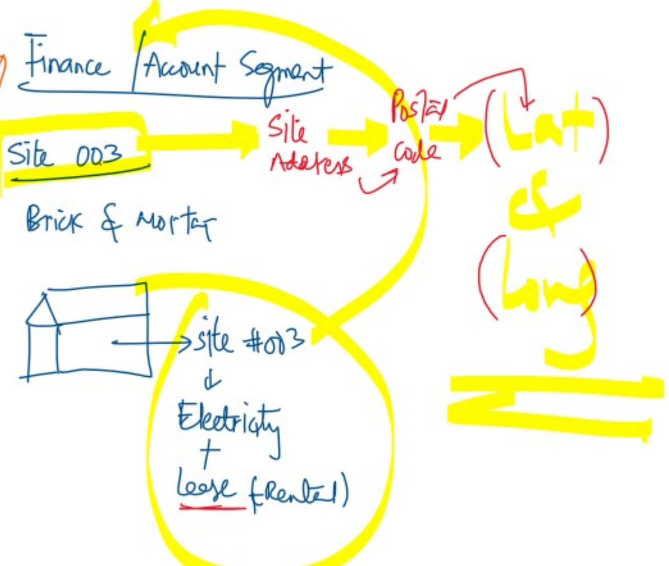
b) Impute with mean/median/mode.

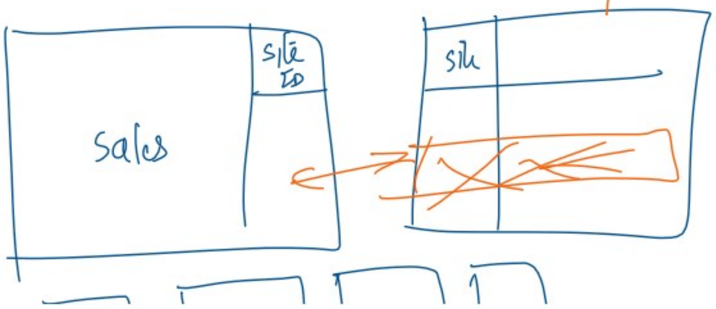


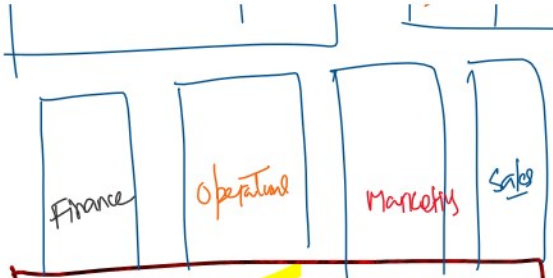
KPI

cost/order → [₹ 10 to ₹ 20]

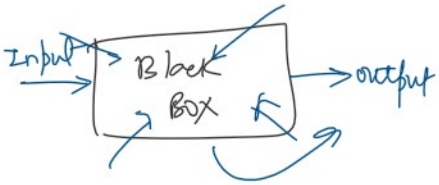
100%

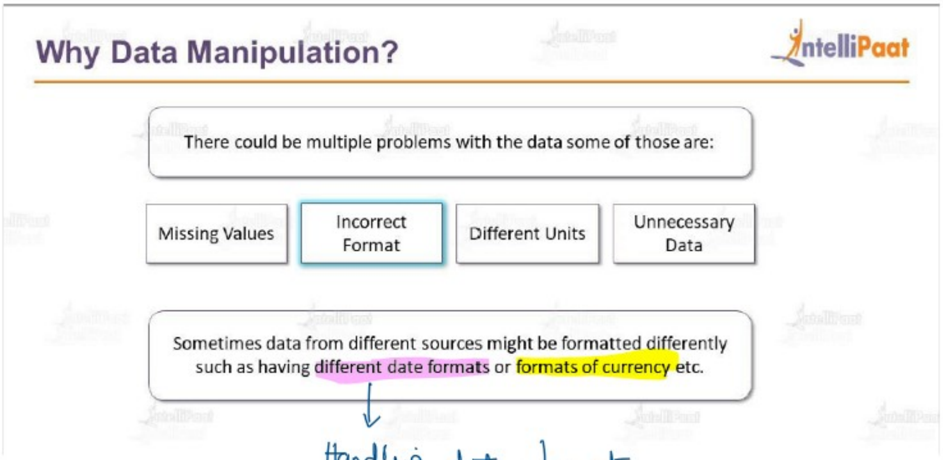






Data Science is an enabler !



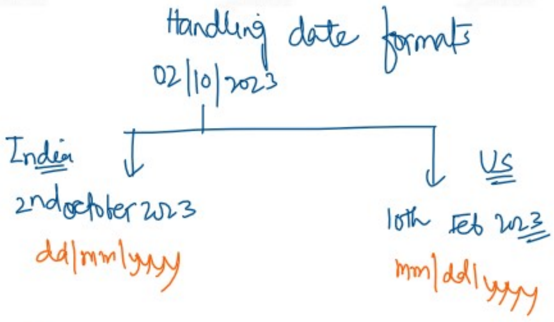


Why Data Manipulation?

There could be multiple problems with the data some of those are:

Missing Values | Incorrect Format | Different Units | Unnecessary Data

Sometimes data from different sources might be formatted differently such as having different date formats or formats of currency etc.



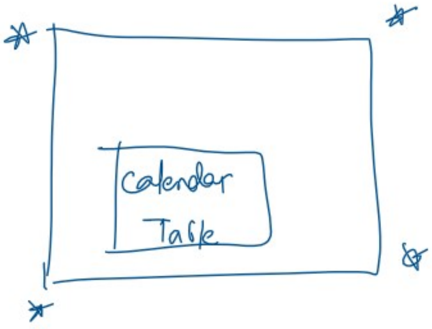
Income Tax

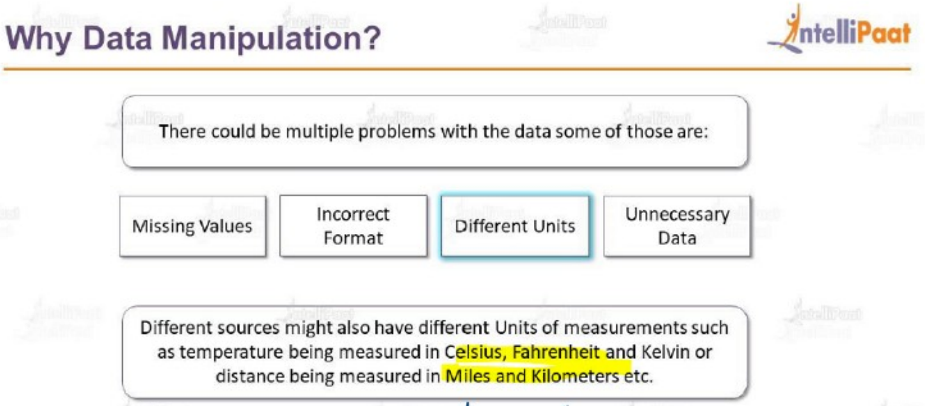
fiscal year #

(April — Mar)

04/2023 — 03/2024

01/2023 — 12/2024





Why Data Manipulation?

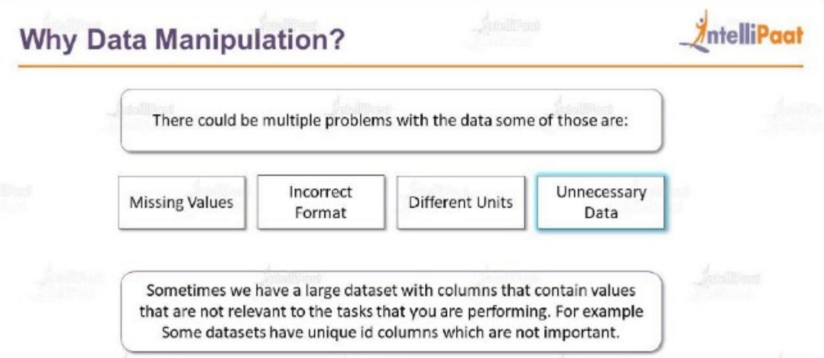
IntelliPaat

There could be multiple problems with the data some of those are:

Missing Values | Incorrect Format | Different Units | Unnecessary Data

Different sources might also have different Units of measurements such as temperature being measured in Celsius, Fahrenheit and Kelvin or distance being measured in Miles and Kilometers etc.

Kg and litres



Why Data Manipulation?

IntelliPaat

There could be multiple problems with the data some of those are:

Missing Values | Incorrect Format | Different Units | Unnecessary Data

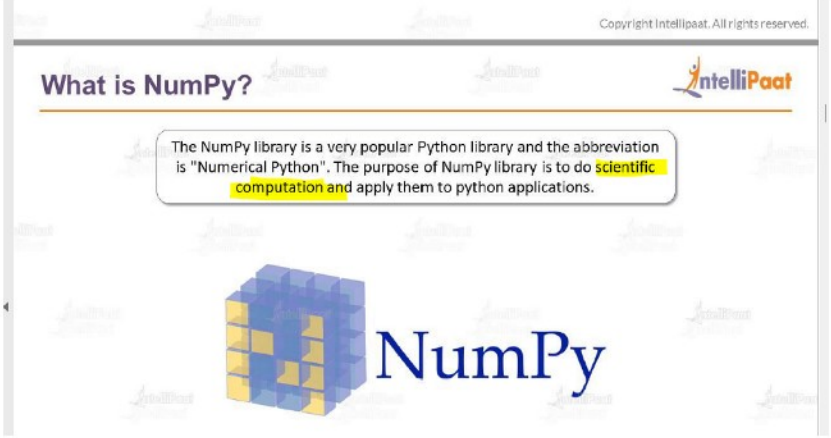
Sometimes we have a large dataset with columns that contain values that are not relevant to the tasks that you are performing. For example Some datasets have unique id columns which are not important.

Feature engineering

Variables | columns & importance

⇓

Rest evrything we drop .



What is NumPy?

IntelliPaat

The NumPy library is a very popular Python library and the abbreviation is "Numerical Python". The purpose of NumPy library is to do scientific computation and apply them to python applications.

How will NumPy help in Data Science?

1. First of all, It is a open source Python library
2. It is fast because it is written in C and Python
3. In Python, there is no in-built array capabilities
4. You can use List as an alternative for arrays, but NumPy is better. But how is it better? We will discuss that now.