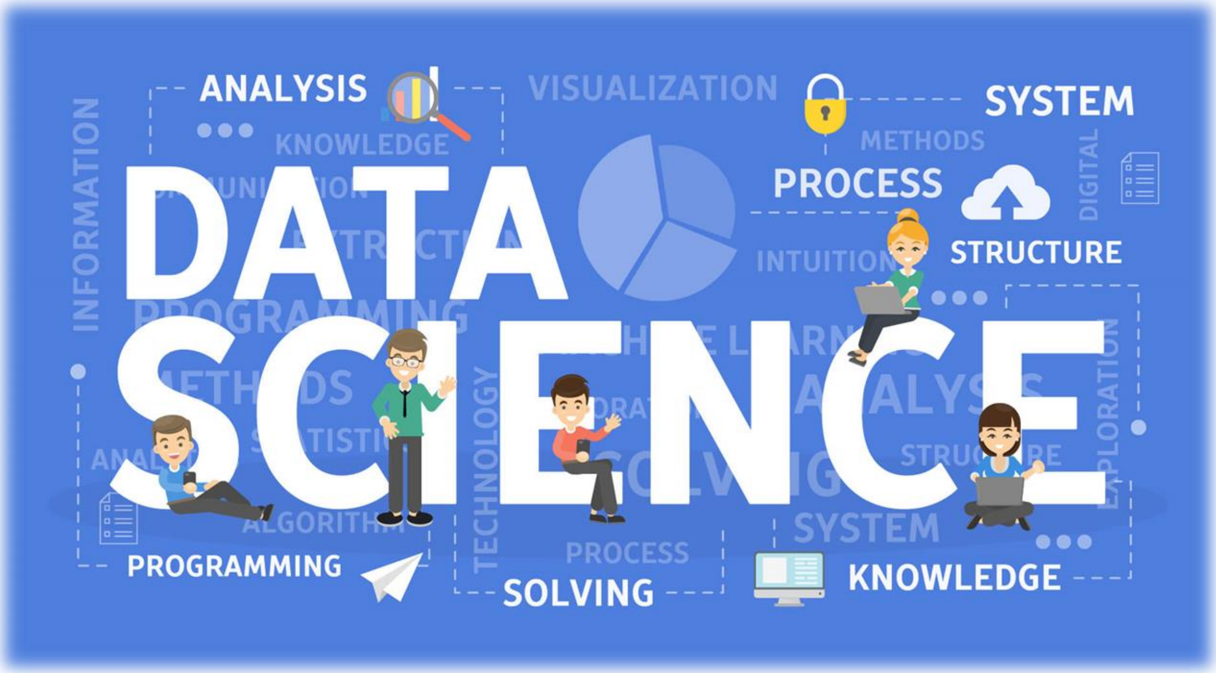
Data wrangling report



Faculty of Computers and Data Science - General

Data science methodology project

Introduction:

1- We are five students in the team, we all worked in this project from A to Z and shared our opinions while doing it

2- We tried to search about data that will satisfy our needs in any websites and we found it in Kaggle.

3- Our data is about the courses in Udemy, includes many features about the courses on the website like (courses name, rating, publishment date, URL of courses, discounts, price, date of creation, reviews, number of subscription)

4- This data had many issues like tidiness and quality issues and had some missing value, so we tried hardly to solve those problems and make it more better and cleaned, as we will mention in the next steps.

5- We worked using the methods that was said in the section videos in addition to asking google about some ways or techniques to complete cleaning the data set

6- We have two tidiness issues and eight quality issues.

Quality issues:

1- (publish time) type will be converted to datetime var
2- (creation time) type will be converted to datetime var
3- (is wishlisted) attribute will be removed.
4- Nan values in (discount price amount) attribute will be converted to 0
5- Nan values in (discount currency) attribute will be converted to INR
6- Nan values in (price amount) attribute will be converted to 0
7- Nan values in (price currency) attribute will be converted to INR
8- The row of course 13607 is paid but price is nan, so it will be removed from the dataset

Tidiness issues:

1- (price_detail__price_string) will be removed, repeated attribute
2- (discount_price__price_string) will be removed, repeated attribute

Code discussion:

1- Firstly, we read the file in CSV format and printed it to show the dataset clearly and define the issues and the problems in it.

2- We used info() and isnull() to define the null values and number of them and in which attributes are they located, and also the summation of those null values and not null elements

3- We needed to check if (avg_rating_recent) and (rating) have the same values, also needed to check if there are any paid courses that does not have a price listed

4- We removed the repeated attributes that causes tidiness issues like (discount_price__price_string) and(price_detail__price_string)

5- We removed (is_wishlisted) as its values are all false so it will be not effective, and also (id) and (URL) as they are irrelevant to our analysis

6- We removed (avg_rating_recent) as it has the same value of (rating) attribute

7- We changed the type of (published_time) and (created) to datetime from string.

8- We wanted to check for a free course that has a listed price, then

9- We filled the indexes that have NAN values to 0, and who have NAN currencies values to INR

10- We wanted to check if there are more null values, but we didn't find them anymore