SC1015 Mini Project

Udemy Courses

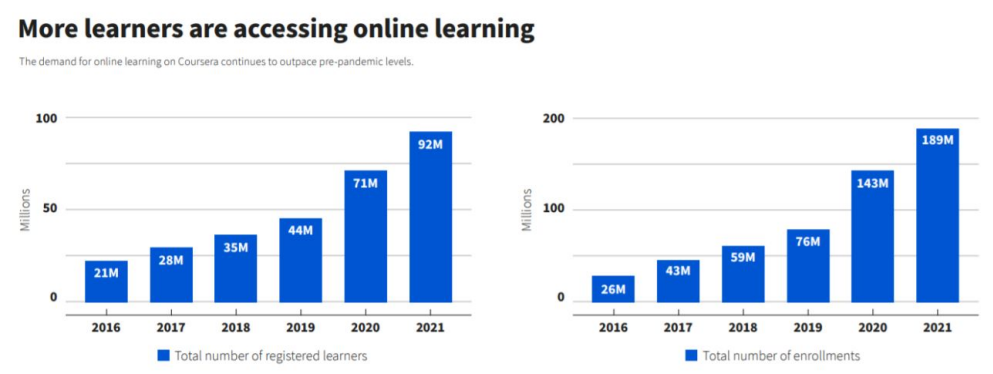
Amizzuddin MD Amin (U2222177J)
Sun Qing (U2220313A)
Ye Bowen (U2220893C)

PT1_TEAM 2

Problem Formulation

Goal

To predict if one was to create an online course, what would be its performance?



Problem Formulation

We want to :

- Identify which factors have a significant impact on the course's success

By achieving this goal, it could potentially :

- Help course creators design a more effective courses that meet the needs of their target audience
- Increase their chances of success on the Udemy platform.
  For example : making more revenue, have more influence by having more subscribers

01

Exploratory Analysis & Data Preparation

Sample Collection

title	is_paid	price	headline	num_subscribers	avg_rating	num_reviews	num_comments	num_lectures	content_length_min	published_time
Online Vegan Vegetarian Cooking School	True	24.99	Learn to cook delicious vegan recipes. Filmed ...	2231	3.75	134	42	37	1268	2010-08-05T22:06:13Z
The Lean Startup Talk at Stanford E-Corner	False	0.00	Debunking Myths of Entrepreneurship A startup ...	26474	4.50	709	112	9	88	2010-01-12T18:09:46Z
How To Become a Vegan, Vegetarian, or Flexitarian	True	19.99	Get the tools you need for a lifestyle change ...	1713	4.40	41	13	14	82	2010-10-13T18:07:17Z

Choosed Course_info file for further analysis

Size of Dataset - 76.15 MB

Preliminary Exploration

In [5]:

```python
# Create a copy of the Dataset
coursedata_clean = rawcoursedata.copy()

# Remove non-usable columns
coursedata_clean.drop(['course_url', 'instructor_url','id'],axis=1,inplace=True)

# Convert all Variable Names to UPPERCASE
coursedata_clean.columns = coursedata_clean.columns.str.upper()

# Print the Variable Information to check
coursedata_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209734 entries, 0 to 209733
Data columns (total 17 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   TITLE               209734 non-null  object
 1   IS_PAID             209734 non-null  bool
 2   PRICE               209734 non-null  float64
 3   HEADLINE            209707 non-null  object
 4   NUM_SUBSCRIBERS     209734 non-null  int64
 5   AVG_RATING          209734 non-null  float64
 6   NUM_REVIEWS         209734 non-null  int64
 7   NUM_COMMENTS        209734 non-null  int64
 8   NUM_LECTURES        209734 non-null  int64
 9   CONTENT_LENGTH_MIN  209734 non-null  int64
 10  PUBLISHED_TIME      209734 non-null  object
 11  LAST_UPDATE_DATE    209597 non-null  object
 12  CATEGORY            209734 non-null  object
 13  SUBCATEGORY         209734 non-null  object
 14  TOPIC               208776 non-null  object
 15  LANGUAGE            209734 non-null  object
 16  INSTRUCTOR_NAME     209729 non-null  object
dtypes: bool(1), float64(2), int64(5), object(9)
memory usage: 25.8+ MB
```

We firstly check the data type and removed unused columns

>course_url
>instructor_url
>id

Preliminary Exploration

```
coursedata_clean.isnull().sum()
```

```
TITLE                  0
IS_PAID                0
PRICE                  0
HEADLINE              27
NUM_SUBSCRIBERS        0
AVG_RATING             0
NUM_REVIEWS            0
NUM_COMMENTS           0
NUM_LECTURES           0
CONTENT_LENGTH_MIN     0
PUBLISHED_TIME         0
LAST_UPDATE_DATE     137
CATEGORY               0
SUBCATEGORY            0
TOPIC                958
LANGUAGE               0
INSTRUCTOR_NAME        5
dtype: int64
```

There are some NULL values we could further drop

Preliminary Exploration

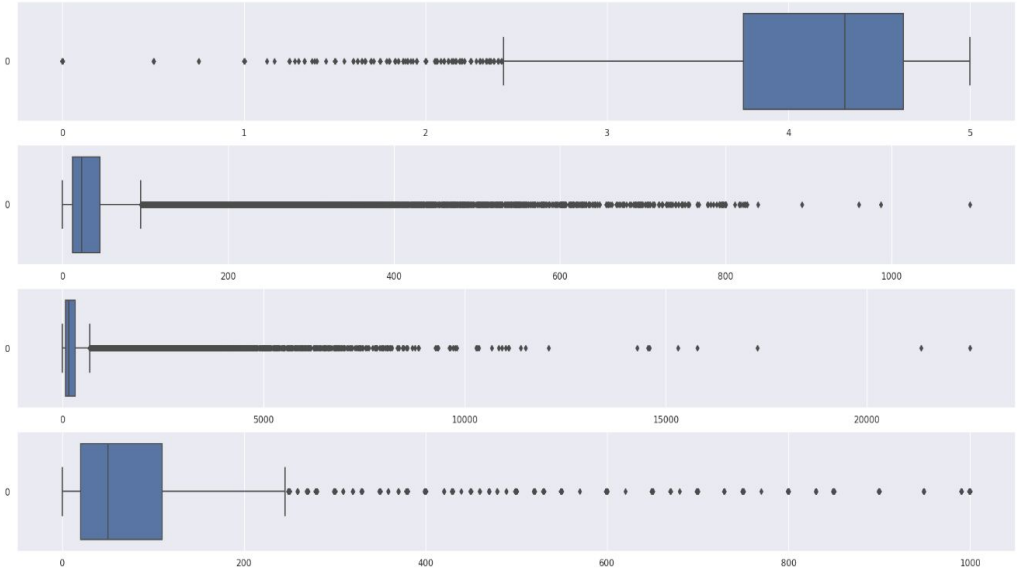
category
Lifestyle
Business
Lifestyle
Lifestyle
Design

One-hot encoding of course category

CATEGORY_Lifestyle	CATEGORY_Marketing	CATEGORY_Music	CATEGORY_Office Productivity	CATEGORY_Personal Development	CATEGORY_Photography & Video	CATEGORY_Teaching & Academics
1.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	1.0

Preliminary Exploration

```
<class 'str'>, AVG_RATING => 3.75, 4.63286525
[AVG_RATING] total outliers: 28166
<class 'str'>, NUM_LECTURES => 12.0, 45.0
[NUM_LECTURES] total outliers: 15034
<class 'str'>, CONTENT_LENGTH_MIN => 75.0, 315.0
[CONTENT_LENGTH_MIN] total outliers: 16648
<class 'str'>, PRICE => 19.99, 109.99
[PRICE] total outliers: 10888
```



Remove outliers for required numeric data

Preliminary Exploration

Before

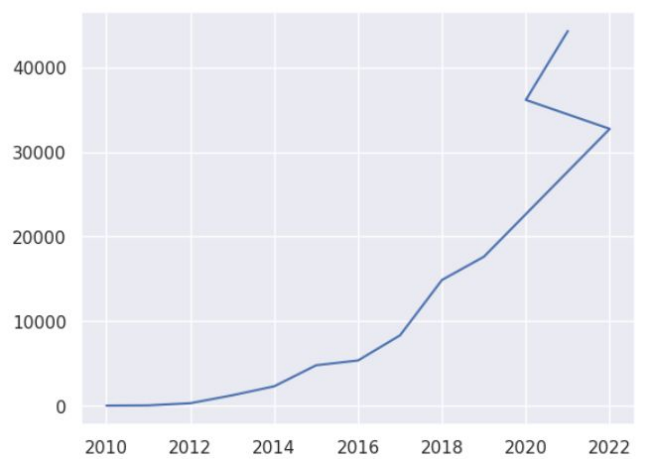
	AVG_RATING	NUM_LECTURES	CONTENT_LENGTH_MIN	PRICE
count	187996.000000	187996.000000	187996.000000	187996.000000
mean	3.700308	38.876987	285.666892	91.108524
std	1.584766	53.975131	473.472317	120.393588
min	0.000000	0.000000	0.000000	0.100000
25%	3.750000	12.000000	75.000000	19.990000
50%	4.310345	23.000000	151.000000	49.990000
75%	4.632865	45.000000	315.000000	109.990000
max	5.000000	1095.000000	22570.000000	999.990000

After

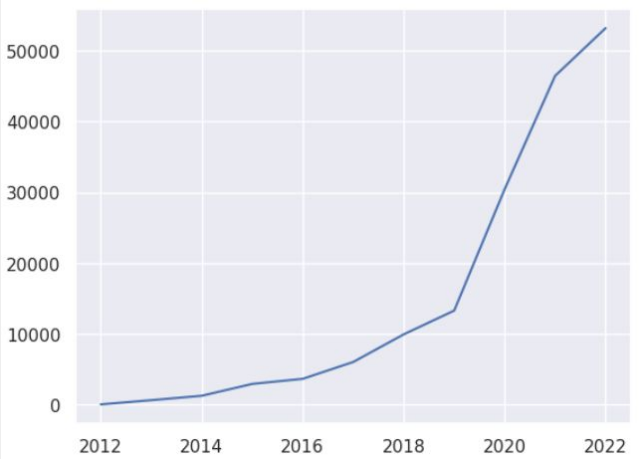
	AVG_RATING	NUM_LECTURES	CONTENT_LENGTH_MIN	PRICE
count	131501.000000	131501.000000	131501.000000	131501.000000
mean	4.318388	27.524270	179.871446	67.808275
std	0.492753	20.104563	143.397927	60.755752
min	2.428571	0.000000	0.000000	0.100000
25%	4.038462	12.000000	74.000000	19.990000
50%	4.400000	22.000000	136.000000	39.990000
75%	4.666666	38.000000	248.000000	99.990000
max	5.000000	94.000000	675.000000	244.990000

02

Data Analysis & Visualization



No. of courses created over years



No. of courses updated over years

Over the time…

Course Price Analysis

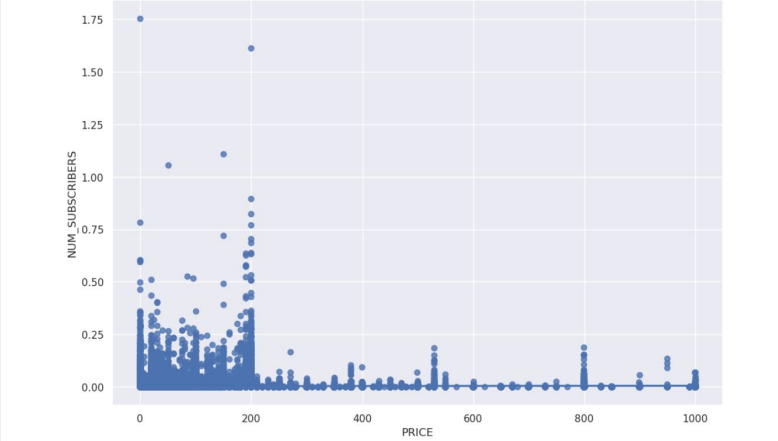
Since most of the courses are paid, we would like to observe the factors for instructors making revenue

Course Price Analysis

Identify the most popular paid courses using catplot

Regplot shows most people prefer courses below 200 which could be more affordable

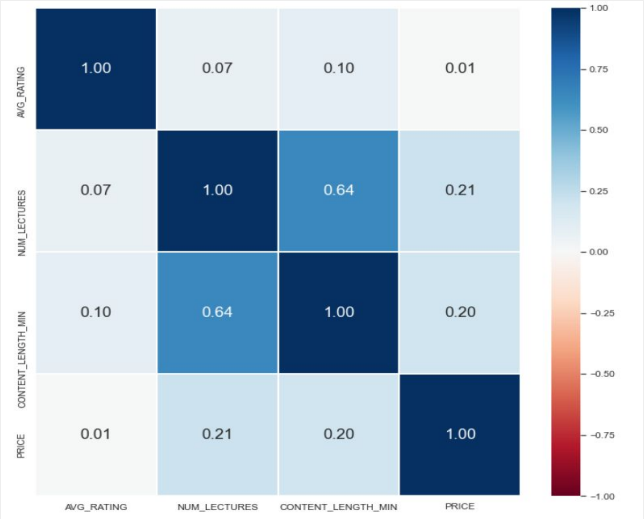
Paid Courses for each category



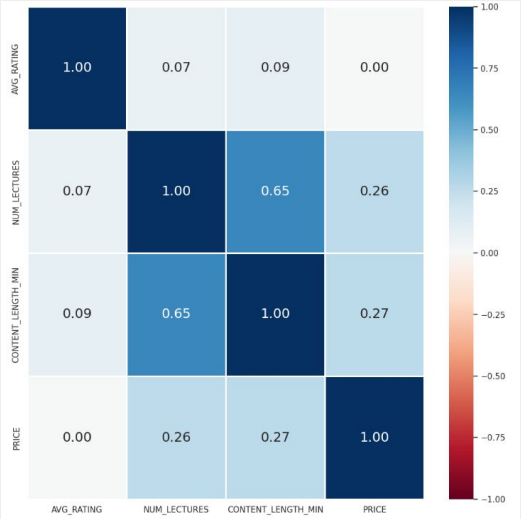
Subscriber distribution for range of price

Course Rating Analysis

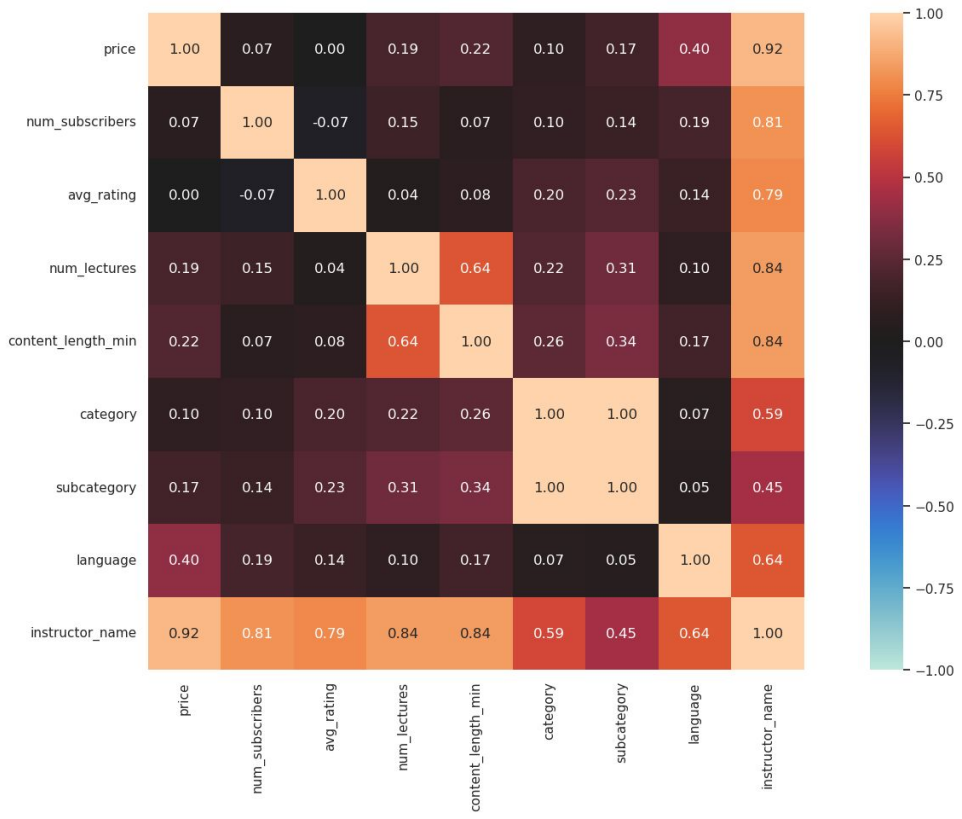
We expect the more number of lectures, longer content length will contribute to higher price.



Heatmap
Exclude 0 price



Nominal Association



Plot heat map to analyse the correlation for each variables

Instructor_name

Top 10 instructor by most gross sales ($):

INSTRUCTOR_NAME	EARN
Srinidhi Ranganathan	1,735,131,640
Learn Tech Plus	1,198,360,878
TJ Walker	1,171,615,935
Jose Portilla	818,904,784
YouAccel Training	801,222,163
Creative Online School	638,075,198
Robert (Bob) Steele	629,703,391
Kirill Eremenko	543,566,459
Joseph Delgadillo	543,485,234
365 Careers	535,805,190

Total gross sales of Udemy: 59.93 billion US dollar

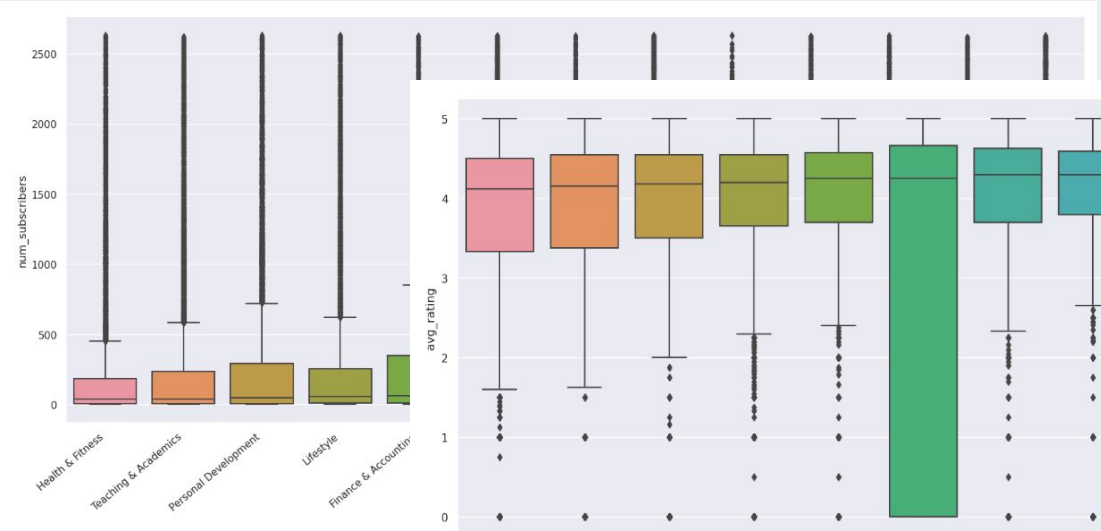
Instructor_name

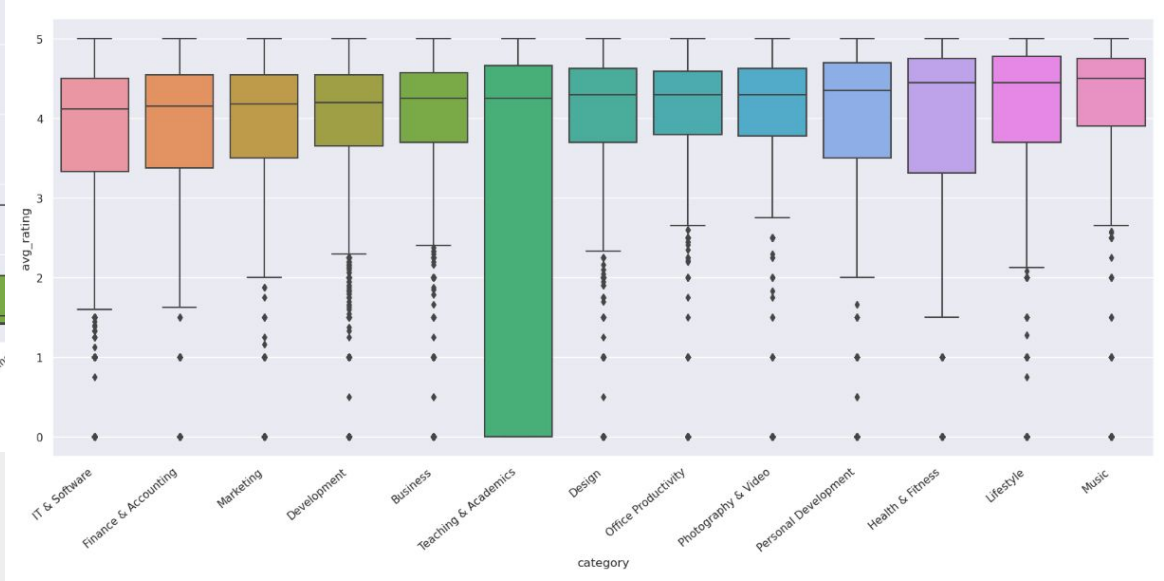
Top 5 productive instructor

	instructor_name	avg_rating	id	num_subscribers
49859	Packt Publishing	3.813512	1254	461689.0
9570	Bluelime Learning Solutions	3.998122	422	3072192.0
28218	Illumeo Learning	3.174521	410	78880.0
36875	Laurence Svekis	4.282094	327	3492822.0
28401	Infinite Skills	4.253677	323	2091770.0

Categories

If one was to create a course, which category to publish to establish the channel?





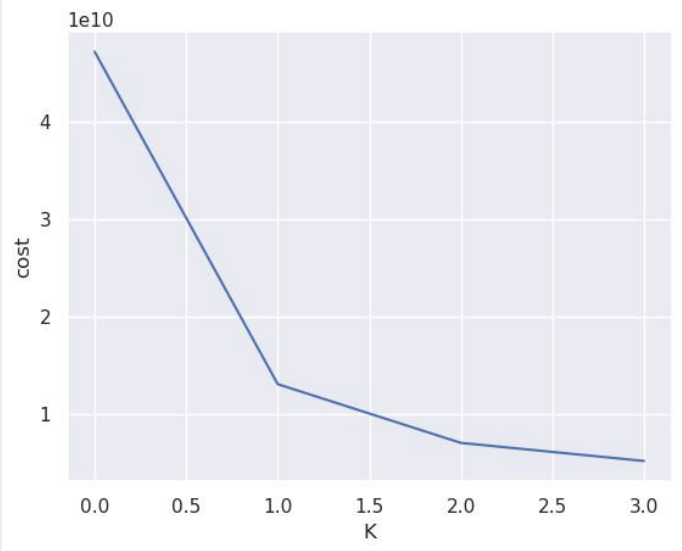
03

Machine Learning

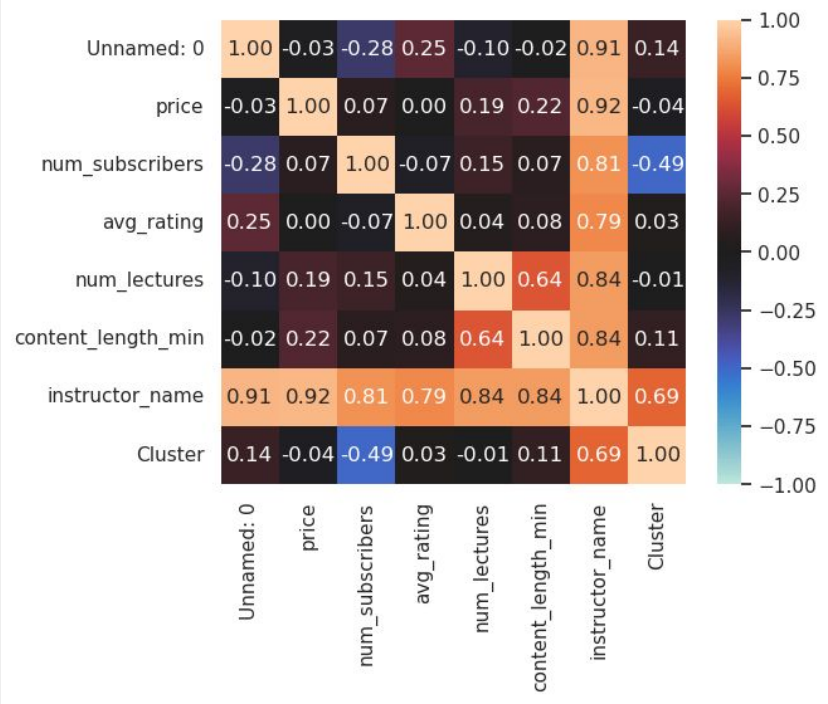
K-prototypes

Unsupervised machine learning techniques is being deployed to cluster instructor_name based on the numerical features.

```
Best run was number 5
[['84.25810819672034' '2643.7432786885247' '4.288409089868769'
  '33.32819672131148' '205.70262295081966' 'Laurence Svekis']
 ['83.0687173076893' '1503.5188461538462' '4.286812085038472'
  '32.68865384615385' '199.2528846153846' 'Pradeep Aggarwal']
 ['83.18129568105738' '2036.2311343141907' '4.294782132653062'
  '32.231371618414805' '193.31347887992408' 'Bluelime Learning Solutions']
 ['70.21661170172773' '305.469805453099' '4.3041623249872325'
  '25.122079988428435' '136.19953713748464' 'Packt Publishing']
 ['86.77670555897352' '70.27884669321816' '4.407841317815702'
  '31.970382004824142' '238.70208111291862' 'Packt Publishing']
 ['101.7579110083751' '562.2403929500144' '4.373869762843109'
  '51.60849465472407' '431.8803813926611' 'Packt Publishing']
 ['56.18114416859212' '52.420106915950626' '4.359117913314102'
  '16.077556046908672' '79.92543871924767' 'Illumeo Learning']
 ['76.55428663904027' '1056.15108751942' '4.302352738684602'
  '29.81214396685655' '179.12687726566546' 'Pradeep Aggarwal']
 ['69.4374739124496' '645.2549125897818' '4.292121978181319'
  '24.656321994850252' '125.57948231467678' 'Packt Publishing']
 ['94.94873036147257' '113.96500191595351' '4.410181661476563'
  '47.6887214203602' '480.0114957210372' 'Packt Publishing']]
```

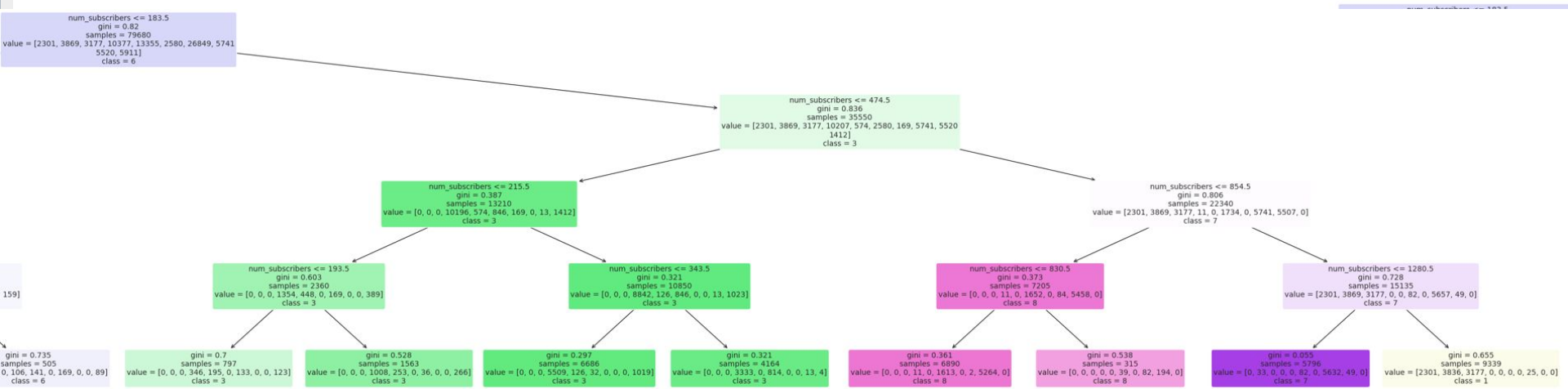


Nominal Association after Kprototypes



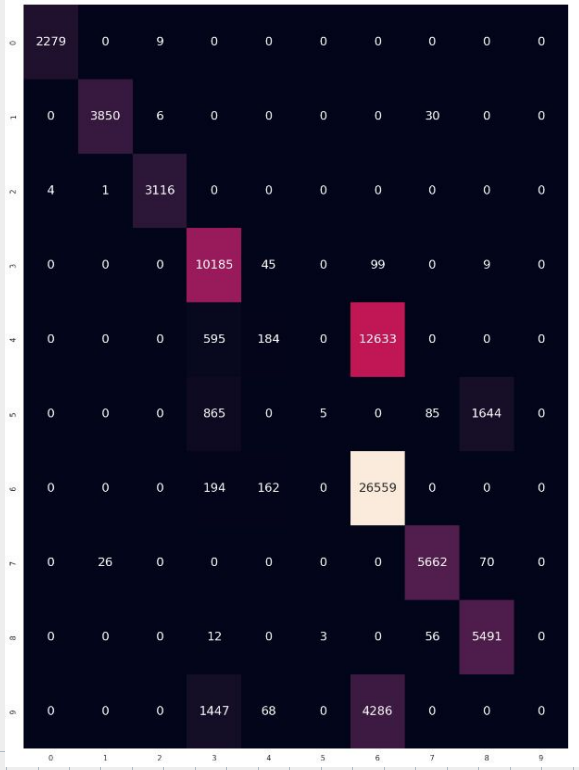
Decision Tree

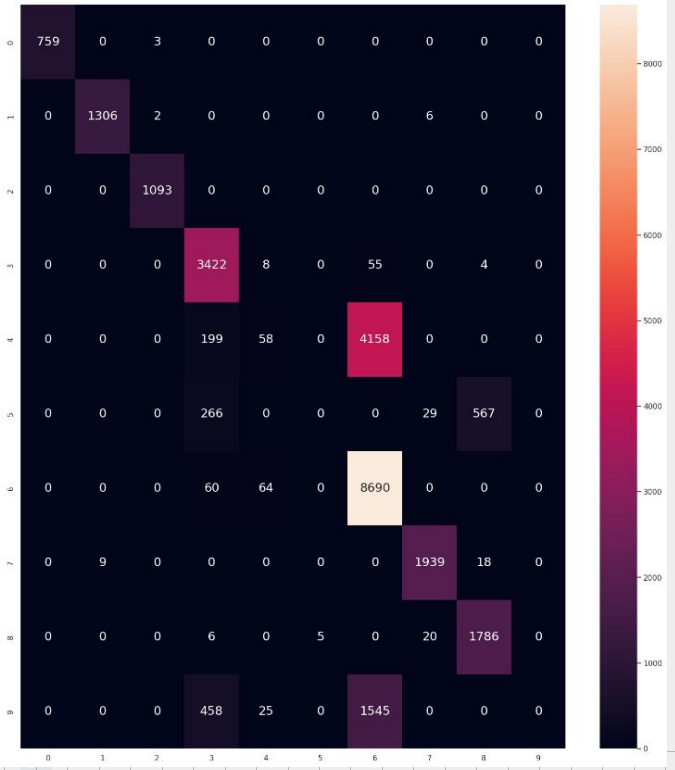
Investigate how accurate the clusters generated by Kprototypes by using decision tree (based on depth of 4)



Confusion matrix

Depth of 10 provide the following confusion matrix





04

Conclusion

What we learned

One-hot encoding

Learnt how to build decision tree with categorical data

New Library

> StandardScaler
> SimpleImputer

K-Means & K-Prototypes

Clustering with Mixed Data Types

Clearance of Bulk Data

Various data type to exclude, convert and remove outliers. Huge data lead to difficulty to choose the correct ones.

Thank You

References

Dataset: https://www.kaggle.com/datasets/hossaingh/udemy-courses

Price Prediction Case Study:
https://towardsdatascience.com/mercari-price-suggestion-97ff15840dbd

Random Forest:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html https://www.datacamp.com/tutorial/random-forests-classifier-python

K-Means & K-Prototypes: https://antonsruberts.github.io/kproto-audience/
SC1015 Course Content under Xtra Module