UDEMY COURSE PREDICTION USING LINEAR REGRESSION

BY: M. RAFI SYAFRINALDI

About Udemy

Udemy is an online learning and teaching marketplace with over 213000 courses and 57 million students. Learn programming, marketing, data science and more.

Why This Project?

Problem Statement

- As a subscriber of this website, it makes me easier to understand the matter discussed in this project.
- I want to help people that want to start their own course by offering them insights about what an ideal course is.

Background

- Linear regression can be used to make predictions about the value of a dependent variable based on the values of one or more independent variables.
- Linear regression is easy to interpret
- The features of this datasets are not that complicated.
- Linear regression is widely used and well understood, so there is a wealth of resources available to help you understand and use the technique.

Problem Statement

Goal

Gaining buisness insights that helps to design an ideal course in Udemy

Research question

- Does the features have strong relationship with other features?

Assumption

The features can leverage a feature so it can be forecasted.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3681 entries, 0 to 1191
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   course_id             3676 non-null   float64
 1   course_title          3676 non-null   object
 2   url                   3676 non-null   object
 3   price                 3676 non-null   float64
 4   num_subscribers       3676 non-null   float64
 5   num_reviews           3676 non-null   float64
 6   num_lectures          3676 non-null   float64
 7   level                 3676 non-null   object
 8   Rating                3677 non-null   float64
 9   content_duration      3676 non-null   float64
 10  published_timestamp   3676 non-null   object
 11  subject               3677 non-null   object
 12  year                  3676 non-null   float64
 13  month                 3676 non-null   float64
 14  YearMonthDate         3676 non-null   object
dtypes: float64(9), object(6)
memory usage: 460.1+ KB
```

Attributes Explained

- course_title: The title of the Udemy course. (String)
- url: The URL of the Udemy course. (String)
- price: The price of the Udemy course. (Float)
- num_subscribers: The number of subscribers for the Udemy course. (Integer)
- num_reviews: The number of reviews for the Udemy course. (Integer)
- num_lectures: The number of lectures in the Udemy course. (Integer)
- level: The level of the Udemy course. (String)
- content_duration: The content duration of the Udemy course. (Float)
- published_timestamp: The timestamp of when the Udemy course was published. (Datetime)
- subject: The subject of the Udemy course. (String)

About The Dataset

- The used dataset taken from kaggle is from the 4 combined datasets which distinguished by the subjects, which are: Web Development, Musical Instruments, Graphic Design, Business Finance.
- Important features are: price, number of subscribers, number of reviews, Rating, content duration, year, number of lectures.

Datasets sourse:
https://www.kaggle.com/datasets/andrewmvd/udemy-courses

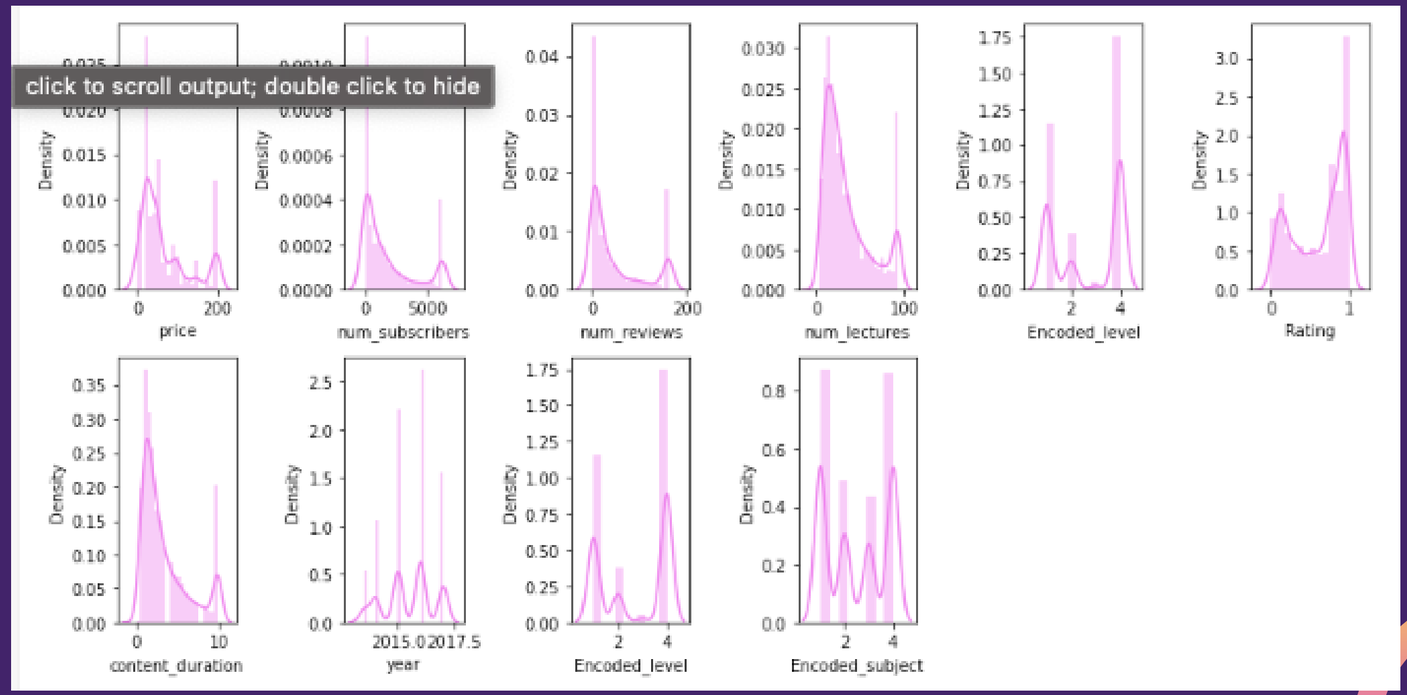
Data Preparation

- Data cleansing
  1. Combining the datasets
  2. Dropping Null and duplicated values
- Encoding
  1. Encoding the non numeric features
  2. Bining the features to be a classification
- Transformation and Standardization
  1. Transforming the features that have bad skewness or big outliers
  2. Standardize the features that don't have outliers but bad skewness.
- Correlation Analysis
  1. Splitting dataset into Train and test Dataset.
  2. Dropping the features that has big correlation

Data Visualization
Graph 1

INSIGHTS

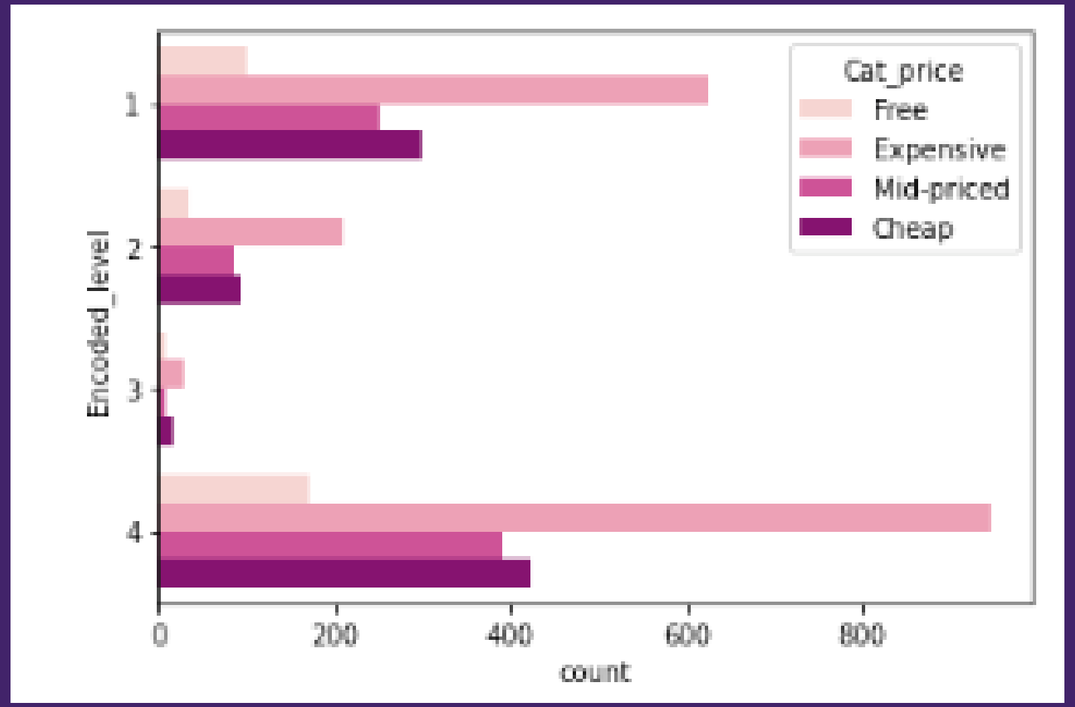
The distribution form of all attributes is not symmetrical the possible cause of this is the not handled outliers. Because the values that are outliers are many so it's safe to make the skewness handled with alternative ways to keep the originality of the data.



Data Visualization
Graph 2

INSIGHTS

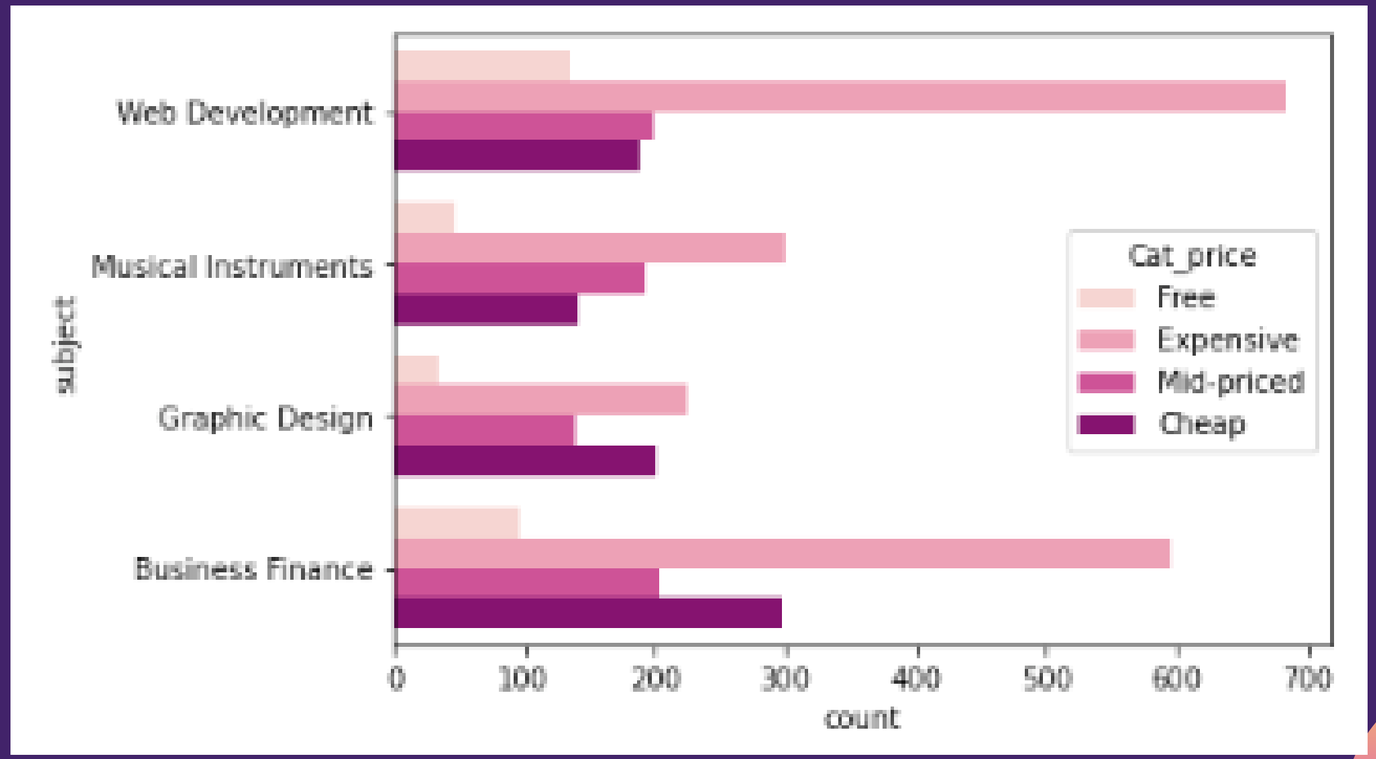
The graph displays us that the most course in each level are expensive courses, and the free ones are the least one.



Data Visualization
Graph 3

INSIGHTS

The graph shows us that almost every subject the expensive ones are the most.

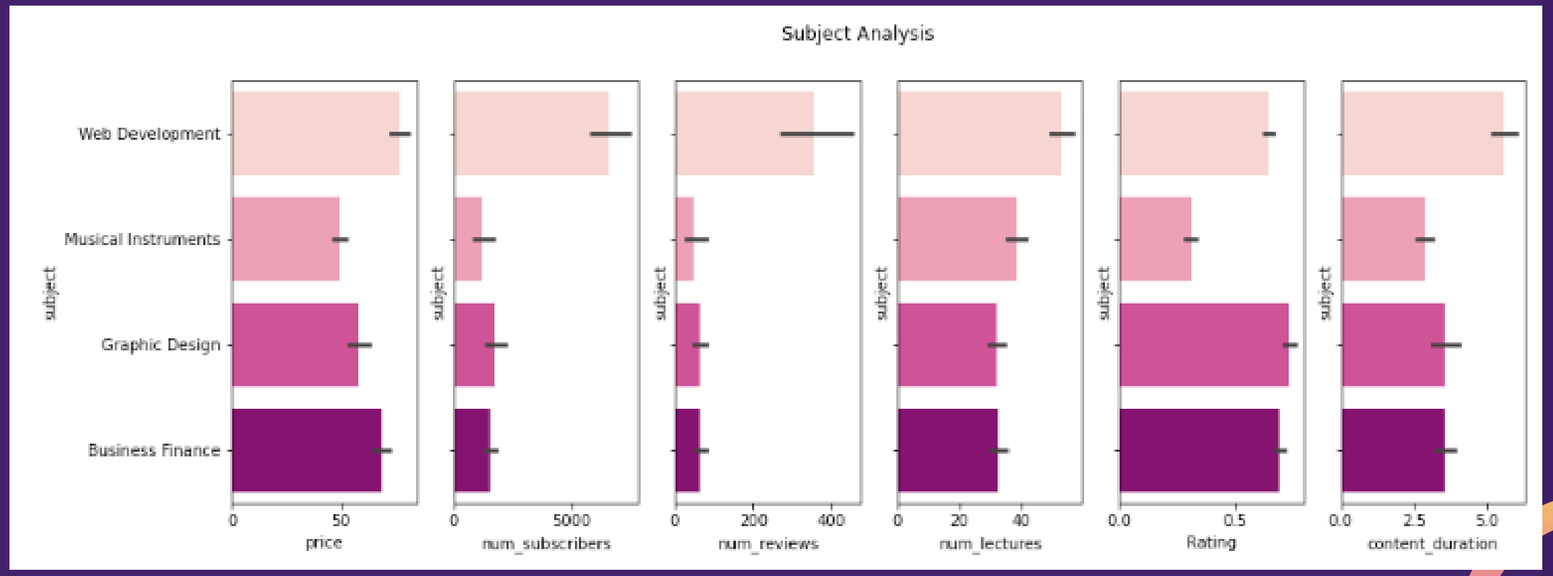


Data Visualization

Graph 4

INSIGHTS

- From this subplots we learn that Web Development dominates almost every category, and the opposite of it is the subject of Musical Instruments.
- The number of Rating is not determined from the most number of the course's num_subscribers.
- Even though Musical Instruments course placed the least in almost category but it offers the second most amount of num_lectures
- Web Development course can be considered as success because of how it placed in every category, so we can deduce that to be a great course it must increase almost each value in every category.

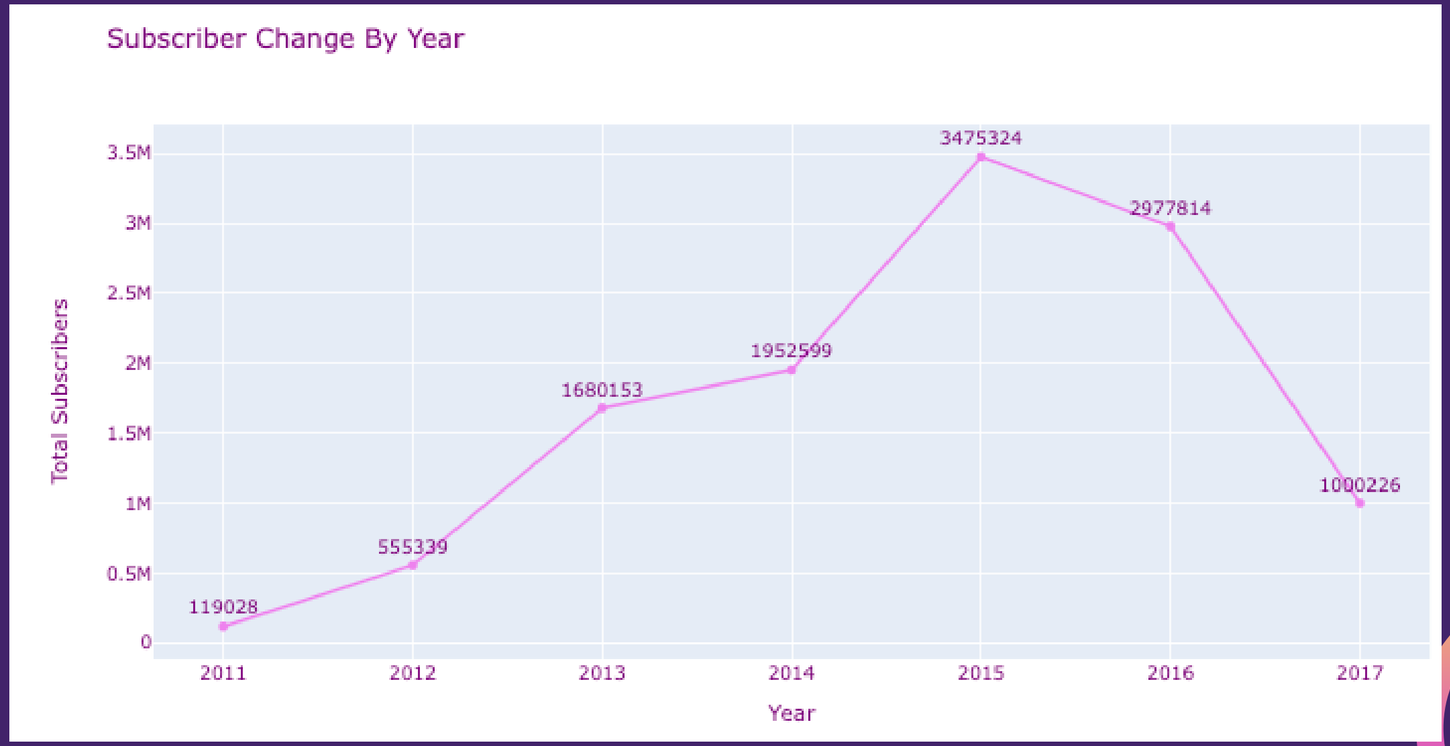


Data Visualization

Graph 5

INSIGHTS

The graph shows us that almost the peak clients on Udemy is in 2015.

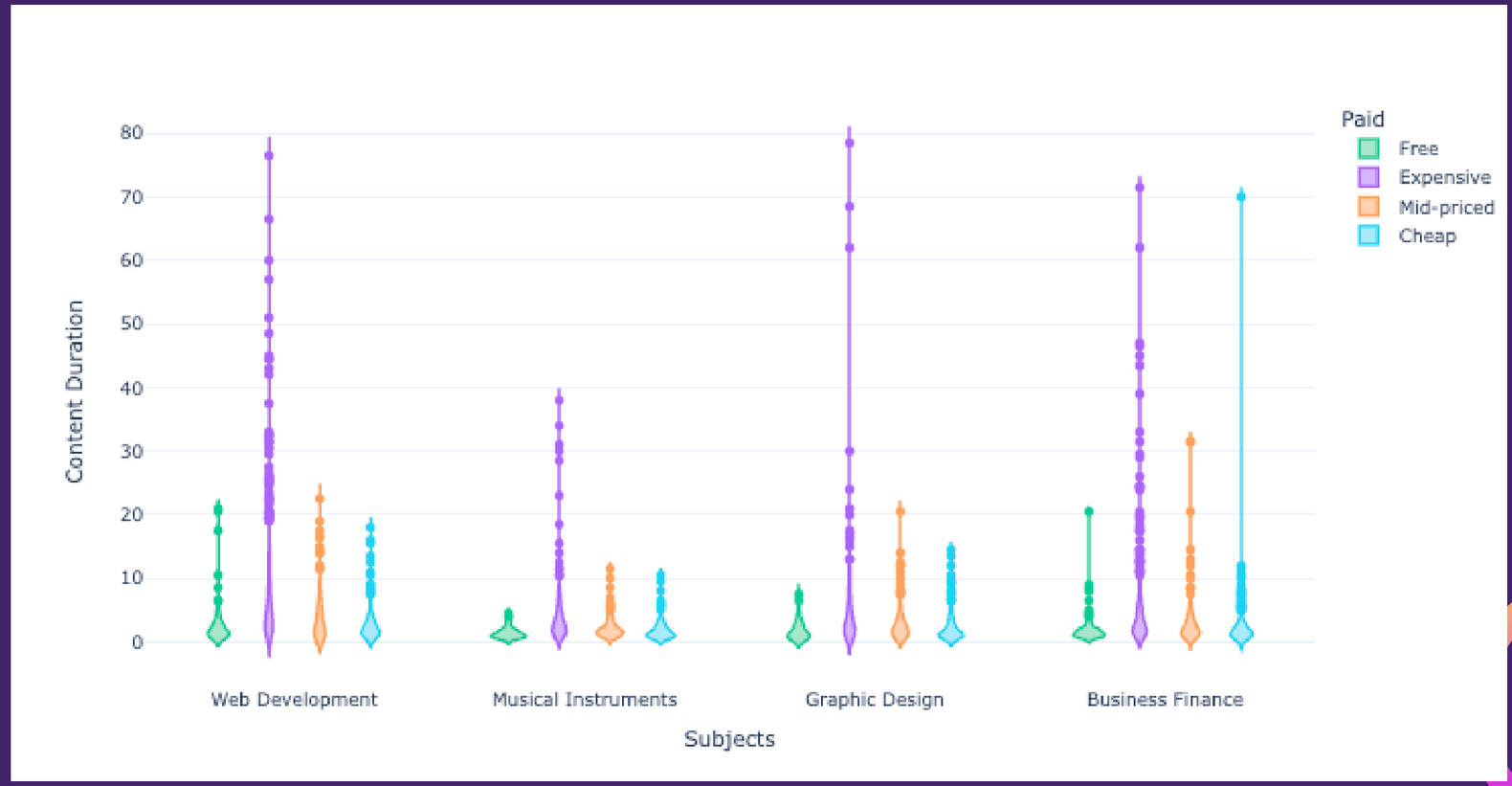


Data Visualization

Graph 6

INSIGHTS

- Free Courses have less content duration
- Courses from Musical instruments have less content duration
- Course with maximum content duration is from web development
- Average content duration is 2 for all subjects except web development(3.5)

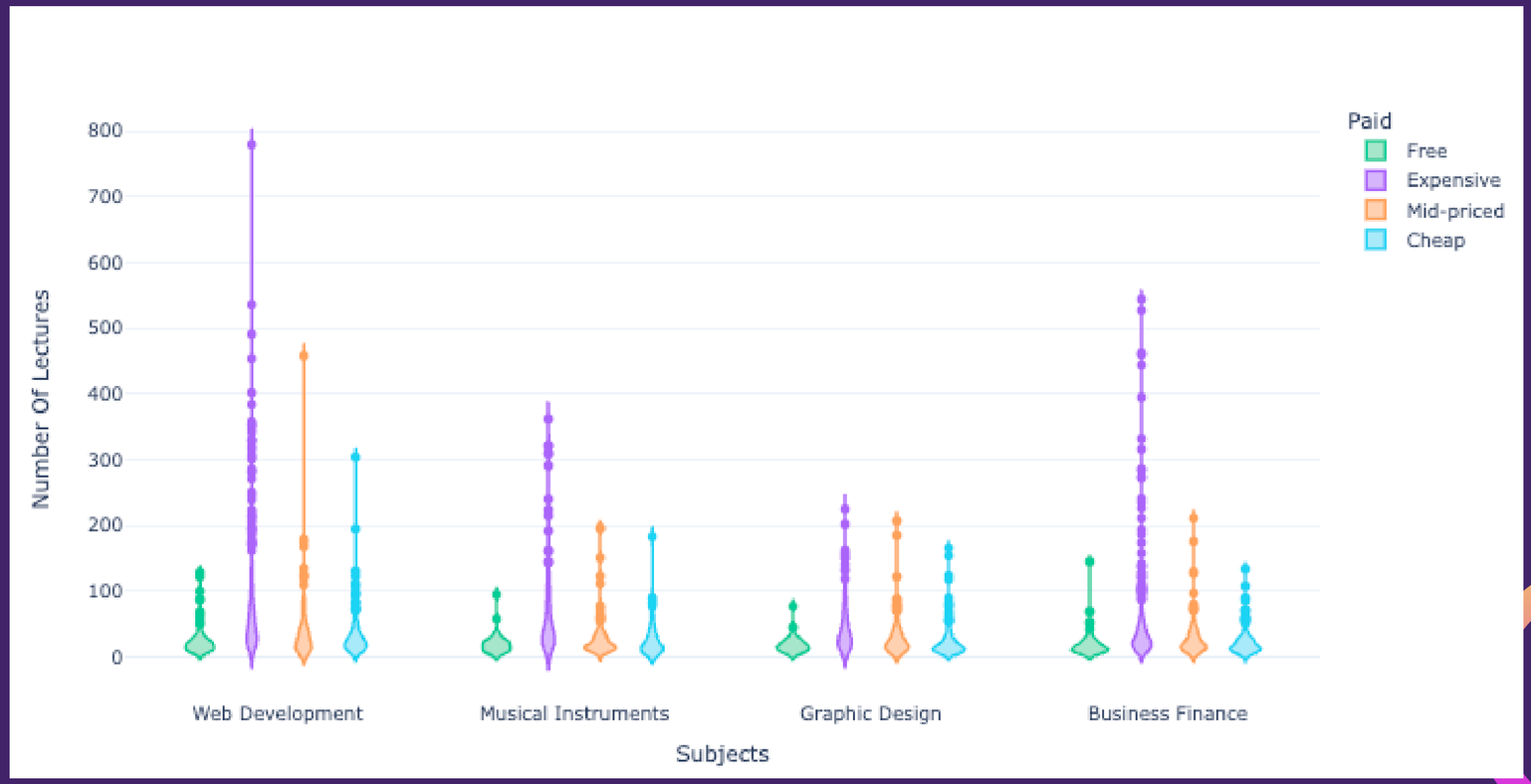


Data Visualization

Graph 7

INSIGHTS

- Expensive Courses have the most content or lectures
- Courses from Grapphic Design have the last contents
- Course with the most content is from web development

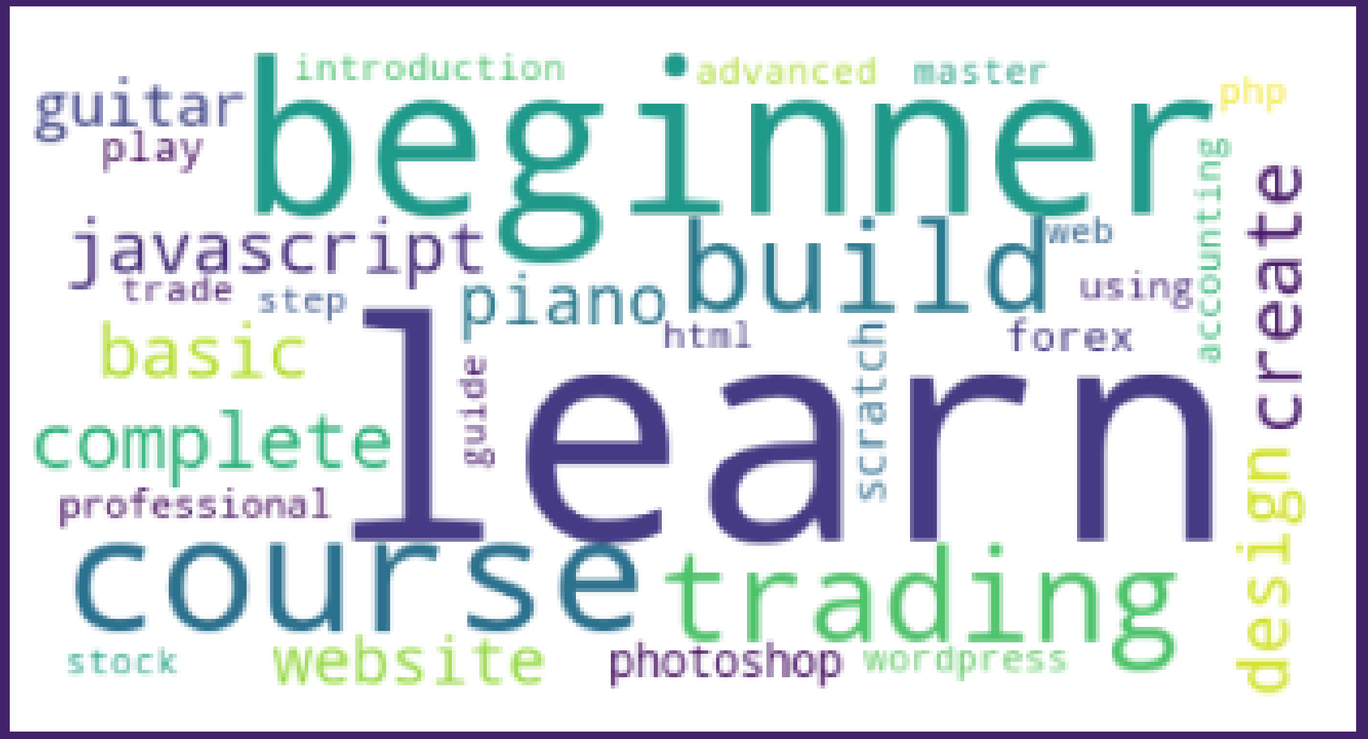


Data Visualization

Graph 8

INSIGHTS

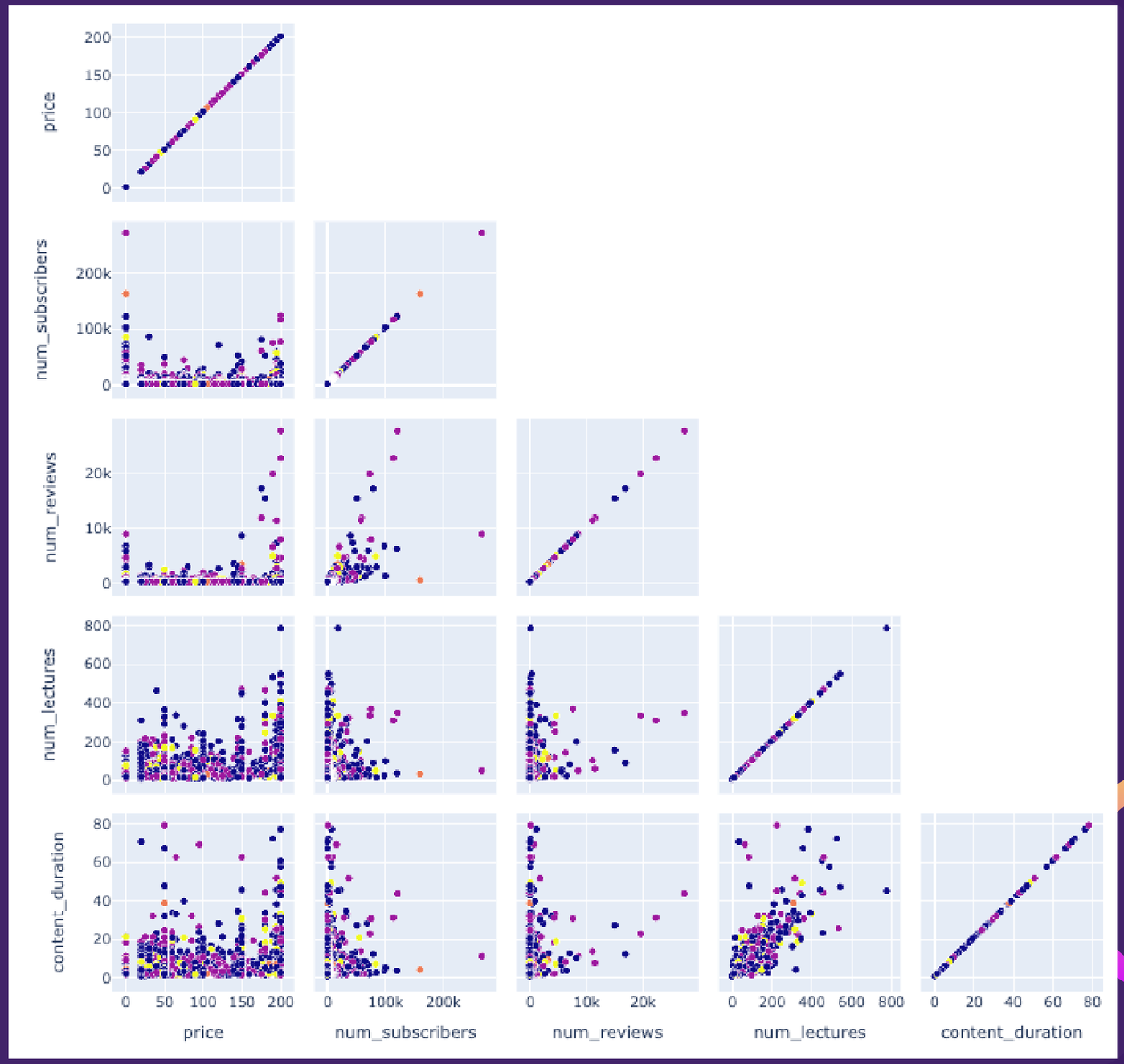
- These are the key words that are most common in `course_title`



Data Visualization
Graph 9

INSIGHTS

- There is positive but not so strong relationship between number of reviews and number of subscribers
- Also there is positive and almost strong (.80) relationship between number of lectures in the course and the duration of the course.



Data Visualization
Graph 10

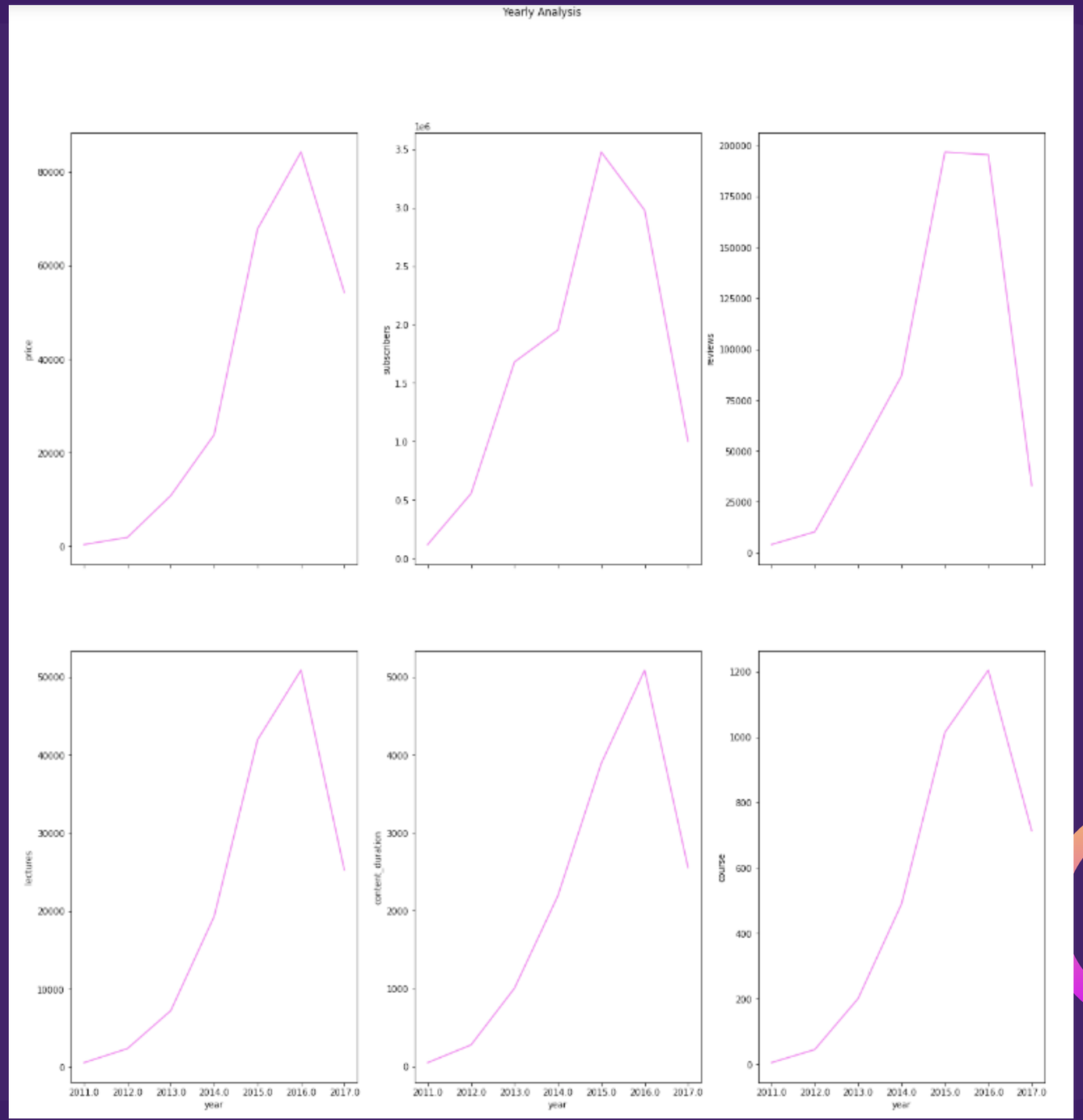
INSIGHTS

- Almost all of the top 10 paid courses are from Web Development area.

Data Visualization
Graph 11

INSIGHTS

- The Graph shows that almost every attribute records a decrement from 2016 to 2017
- Almost every attributes show the same fluctuation throug the years.
- The peak year of Udemy is at the year 2016 seen from



EDA
Conclusion

INSIGHTS

- Most Of The Courses Is Paid
- Most Of The Courses are All Levels , Need More From Expert Level
- Most Of Students care about Buisiness Finance & Web Development
- Course Price Does Not Affect Subscribers
- Price Does Not Affect Reviews
- Price Does Not Affect Course Duration
- Probaly course time affect course subscribers
- More Subscribers leads to more reviews

	index	course_title	year	subject	num_subscribers	price	Rating	content_duration	level
0	2	The Web Developer Bootcamp	2015.0	Web Development	121584.0	200.0	0.89	43.0	Beginner Level
1	4	The Complete Web Developer Course 2.0	2016.0	Web Development	114512.0	200.0	0.55	30.5	Beginner Level
2	1	Pianoforall - Incredible New Way To Learn Pian...	2014.0	Musical Instruments	75499.0	200.0	0.96	30.0	Beginner Level
3	3	Photoshop for Entrepreneurs - Design 11 Practi...	2016.0	Graphic Design	36288.0	200.0	0.96	5.0	All Levels
4	32	Ultimate Web Designer & Developer Course: Buil...	2015.0	Web Development	33788.0	200.0	0.31	32.5	Beginner Level
5	41	PHP for Beginners -Become a PHP Master - Proje...	2015.0	Web Development	28880.0	200.0	0.40	30.5	Beginner Level
6	53	The Ultimate Web Developer How To Guide	2015.0	Web Development	24861.0	200.0	0.85	22.5	Beginner Level
7	7	How To Make Graphics For A Website	2014.0	Graphic Design	24857.0	200.0	0.89	1.5	All Levels
8	59	1 Hour JavaScript	2013.0	Web Development	22999.0	200.0	0.24	1.0	All Levels
9	64	Become A Web Developer And Seller - Build Webs...	2013.0	Web Development	21730.0	200.0	0.36	2.5	All Levels
10	5	The Professional Guitar Masterclass	2015.0	Musical Instruments	21701.0	200.0	0.30	9.5	All Levels
11	7	Black Algo Trading: Build Your Trading Robot	2014.0	Business Finance	20195.0	200.0	0.21	16.0	All Levels
12	10	Canva Graphics Design for Entrepreneurs - Desi...	2016.0	Graphic Design	18303.0	200.0	0.81	3.5	All Levels
13	80	Back to School Web Development and Programming...	2013.0	Web Development	18170.0	200.0	0.40	44.5	All Levels
14	94	Adobe Flash for Beginners - Build Flash Websit...	2014.0	Web Development	17071.0	200.0	0.76	1.5	Intermediate Level

	year	num_subscribers	price	Rating	content_duration
count	15.000000	15.000000	15.0	15.000000	15.000000
mean	2014.533333	40029.200000	200.0	0.592667	18.266667
std	1.060099	34770.309118	0.0	0.287190	15.906947
min	2013.000000	17071.000000	200.0	0.210000	1.000000
25%	2014.000000	20948.000000	200.0	0.335000	3.000000
50%	2015.000000	24857.000000	200.0	0.550000	16.000000
75%	2015.000000	35038.000000	200.0	0.870000	30.500000
max	2016.000000	121584.000000	200.0	0.960000	44.500000

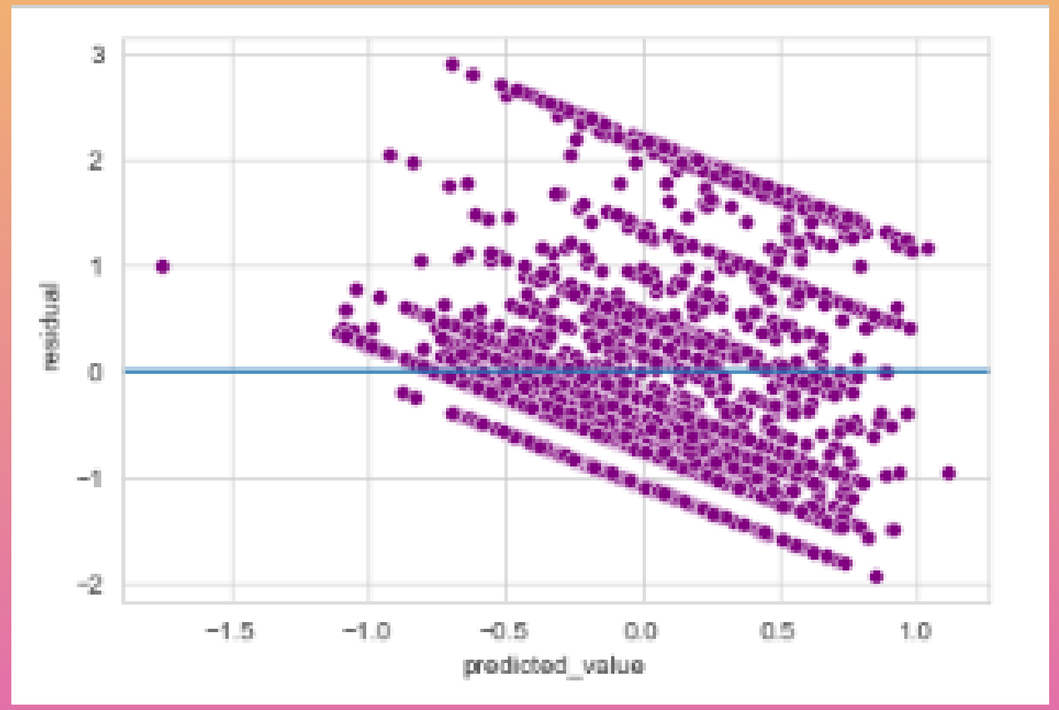
Modelling

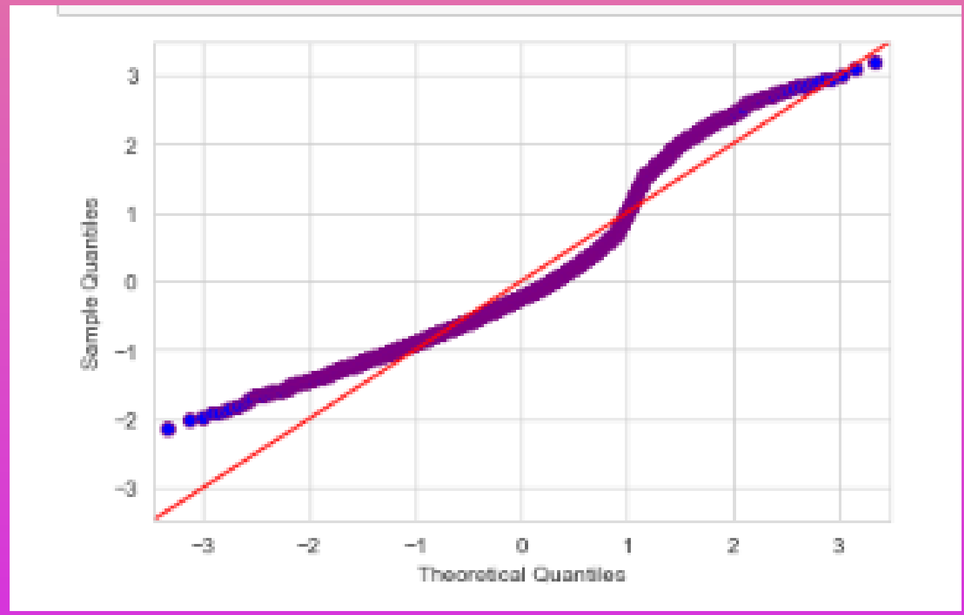
- What's the Target?
  In this matter, the target used is the price.
- Model fitting
- Dropping high correlation feature
- Training Error
- Testing Error
- Finding the coefficient

Dropping high correlation feature

Dropping content duration feature.

Training, and Residuals





Training Error

- MAE = On average, our prediction deviates the true price by 0.69
- MAPE = With 1.14 is considered Good to forecast, and this 0.69 is equivalent to 1.14 deviation relative to the standardized price

```
RMSE for training data is 0.9064133222305405
MAE for training data is 0.6991078726347728
MAPE for training data is 1.1430941009191986
```

Testing Error

- MAE = On average, our prediction deviates the true price by 0.7
- MAPE = With 1.22 is considered Good to forecast, and this 0.7 is equivalent to 1.22 deviation relative to the standardized price

```
RMSE for testing data is 0.9199733549765127
MAE for testing data is 0.7087642300071461
MAPE for testing data is 1.223617267012635
```

Model Conclusion

- A 1 unit increase in log_subscribers is associated with an increase of 0.008199 total price (standardized form).
- A 1 unit increase in log_reviews is associated with an increase of 0.087175 total price (standardized form).
- A 1 unit increase in log_lectures is associated with an increase of 0.441011 total price (standardized form).
- A 1 unit increase in log_rating is associated with an increase of 0.119606 total price (standardized form).
- A 1 unit increase in year is associated with an increase of 0.151768 total price (standardized form).
- A 1 unit increase in Encoded_subject is associated with an increase of 0.065093 total price (standardized form).

	feature	coefficient
0	log_subscribers	0.008199
1	log_reviews	0.087175
2	log_lectures	0.441011
3	log_rating	0.119606
4	year	0.151768
5	Encoded_subject	0.065093

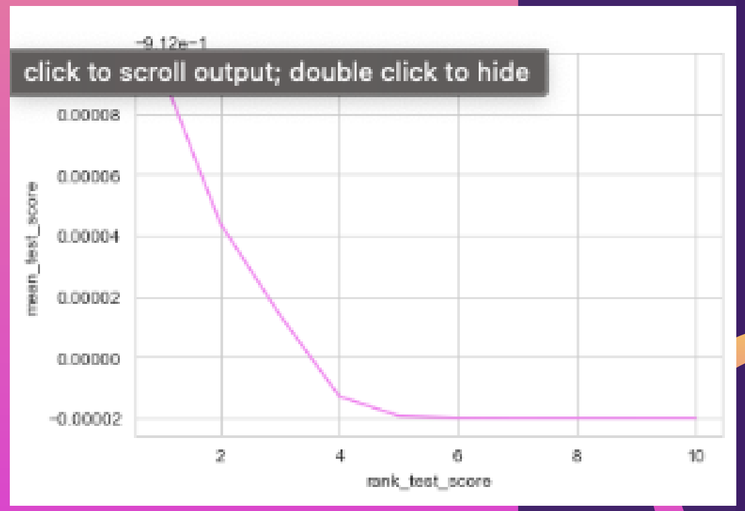
Hyperparameter Tuning

- A 1 unit increase in log_subscribers is associated with an increase of 0.021685 total price (standardized form).
- A 1 unit increase in log_reviews is associated with an increase of 0.068339 total price (standardized form).
- A 1 unit increase in log_lectures is associated with an increase of 0.287351 total price (standardized form).
- A 1 unit increase in log_rating is associated with an increase of 0.079715 total price (standardized form).
- A 1 unit increase in log_content is associated with an increase of 0.247819 total price (standardized form).
- A 1 unit increase in year is associated with an increase of 0.164968 total price (standardized form).
- A 1 unit increase in Encoded_subject is associated with an increase of 0.060542 total price (standardized form).

	params	mean_test_score	rank_test_score
0	{'alpha': 1e-06}	-0.912020	10
1	{'alpha': 1e-05}	-0.912020	9
2	{'alpha': 0.0001}	-0.912020	8
3	{'alpha': 0.001}	-0.912020	7
4	{'alpha': 0.01}	-0.912020	6
5	{'alpha': 0.1}	-0.912019	5
6	{'alpha': 1}	-0.912013	4
7	{'alpha': 5}	-0.911986	3
8	{'alpha': 10}	-0.911956	2
9	{'alpha': 20}	-0.911906	1

click to scroll output; double click to

	feature	coefficient
0	intercept	-334.274109
1	log_subscribers	0.021685
2	log_reviews	0.068339
3	log_lectures	0.287351
4	log_rating	0.079715
5	log_content	0.247819
6	year	0.164968
7	Encoded_subject	0.060542



Model Simulation

```
model = ridge_reg.fit(y_df_train.reshape(-1, 1), X_df_train)
```

```
import joblib

filename = "model.sav"
joblib.dump(model, filename)
```

```
['model.sav']
```

```
loaded_model = joblib.load(filename)
```

```
loaded_model.predict([[20]])
```

```
array([[1.27304613e+01, 9.30989840e+00, 8.56688418e+00, 5.40397551e-01,
        2.01847973e+03, 1.53922303e+00]])
```

Business Recommendations:

So, in order to have an ideal course on Udemy It would be:

- ***Under Web Development***
- ***priced between 175-200USD***
- ***tagged as all levels***
- ***Content Duration between 15-18 hours long***
- ***Put popular keywords in the title***

THANK YOU