Alexey Rybak
Monthly 22

Building an Online Education Platform: Lessons from Udemy

Contents

Executive Summary

- The purpose of this report is to assist development of a new online education company in Russia by analyzing the content offering of Udemy, a popular American digital learning platform. Analysis of over 3,500 courses offered at Udemy since 2011 is aimed at discovering correlations between the *number of subscribers* for each course (a direct measure of popularity and revenue), and content qualities such as number of lectures, price, difficulty level and subject; and prior customer engagement measured by the number of reviews.
- The number of reviews is the most important predictor of the course popularity, which means that the new company should invest in building social features. Customers should be encouraged to leave reviews and feedback – word-of-mouth will also help to boost the popularity of the platform as a whole.
- Web development courses are more popular, as are broadly accessible courses. On the other hand, there are already numerous companies and platforms offering professional software development courses, so there may be an opportunity to differentiate from the competition.
- Demand is relatively inelastic – price does not seem to highly correlate with the number of customers. This may mean that people are more willing to invest in education in order to advance their personal and professional knowledge (especially in times of COVID-19); however, Udemy is notorious for its aggressive pricing policy, which may explain the weakness in correlation, as these discounts are not reflected in the data.

Introduction

The global online education market is growing at a CAGR of 28.55%, expecting to reach $132.98 billion by 2023, and COVID-19 is only accelerating this trend. To seize this opportunity, ATO Events, a Moscow-based conference company, is launching an online education platform for airline professionals.

The purpose of this report is to help ATO Events develop a compelling content offering for the platform, by analyzing over 3,500 courses offered at Udemy (a popular American digital learning platform) since 2011. It aims to answer the following questions:

- What should the product managers focus on to attract more customers?
- How elastic is the demand?
- How should content be structured?

Dependent variable

- *Number of subscribers*. The number of people who have received access to the course. An objective measure of a course's popularity.

Independent variables (Quantitative)

- *Price* (in dollars). The amount paid by the customer for the course (not including discounts).
- *Number of reviews* submitted by customers for each course. Does not reflect the sentiment.
- *Number of lectures* included in each course.
- *Content duration* (in hours). Indicates how many lecture hours each course contains.
- *Time elapsed* (in days). This variables indicates the age of each course, as measured by the number of days between January 1, 2011, and the date a course was published.

Independent variables (Nominal)

- *Level of content*. A categorical variable describing the level of difficulty for each course.
- *Subject area*. A categorical variable describing the subject area of a course: 'Business & Finance', 'Graphic Design', 'Musical Instruments', or 'Web Development'.

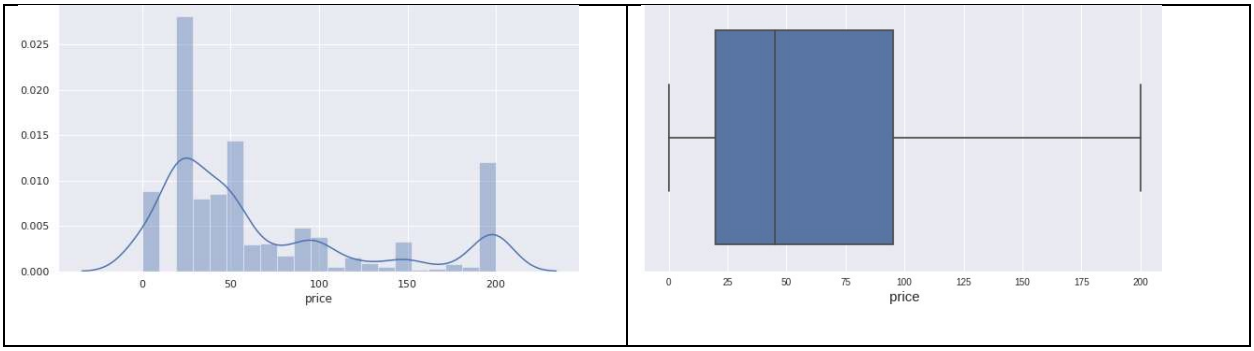
Analysis and Methods

Data Exploration

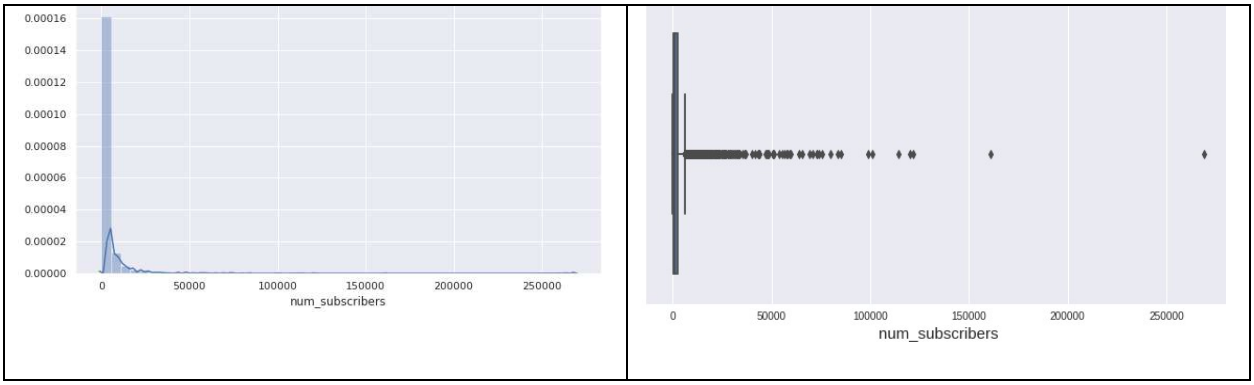
We start the exploration with price, number of reviews, and the number of subscribers:

	price	num_subscribers	num_reviews
count	3677.000000	3677.000000	3677.000000
mean	66.062007	3198.020125	156.301605
std	61.009324	9505.263339	935.575723
min	0.000000	0.000000	0.000000
25%	20.000000	111.000000	4.000000
50%	45.000000	912.000000	18.000000
75%	95.000000	2547.000000	67.000000
max	200.000000	268923.000000	27445.000000

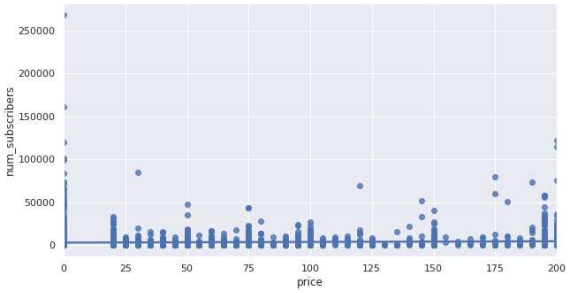
Price seems to follow approximately normal distribution (it can also be bimodally distributed; further analysis is needed). Mean price is $66, with standard deviation of $61.

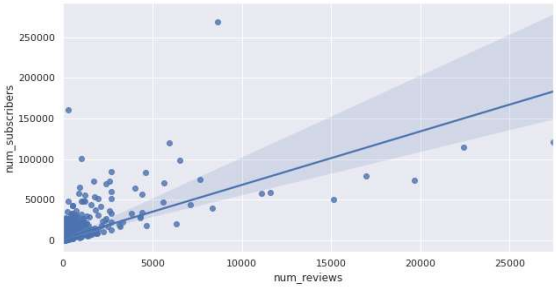


The *number of subscribers* and the *number of reviews* are both sharply skewed: most courses have relatively few of either (with mean of 3,200 and 156, respectively), but a few are extremely popular, creating large outliers (max of 268,923 subscribers and 27,445 reviews). All three variables show high standard deviation: number of subscribers and reviews are particularly affected due to many outliers.

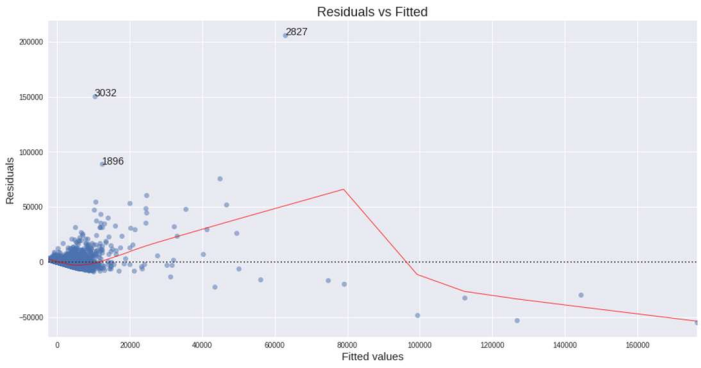


Surprisingly, there is little correlation between the number of subscribers and price (correlation coefficient of 0.05); however, correlation between the number of subscribers and the number of reviews is much stronger (correlation coefficient of 0.65).





We can expect the diagnostic plot of the multiple regression model to reflect the skewedness of the variables. Indeed, it does not seem to predict the higher values well.



To address this, we will apply logarithmic transformation to four variables: number of subscribers, number of reviews, content duration and the number of lectures. Before using the multiple regression, we will also convert difficulty levels and subject areas into indicator variables.

Multiple linear regression model

OLS Regression Results

Dep. Variable:	num_subscribers	R-squared:	0.667
Model:	OLS	Adj. R-squared:	0.666
Method:	Least Squares	F-statistic:	667.0
Date:	Sun, 14 Jun 2020	Prob (F-statistic):	0.00
Time:	21:28:47	Log-Likelihood:	-6380.8
No. Observations:	3677	AIC:	1.279e+04
Df Residuals:	3665	BIC:	1.286e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.5591	0.113	22.618	0.000	2.337	2.781
price	0.0012	0.000	2.800	0.005	0.000	0.002
num_reviews	0.9406	0.016	60.292	0.000	0.910	0.971
num_lectures	0.1915	0.050	3.852	0.000	0.094	0.289
content_duration	-0.4047	0.045	-9.067	0.000	-0.492	-0.317
time_elapsed	-0.0006	5.82e-05	-9.855	0.000	-0.001	-0.000
all_levels	0.8503	0.056	15.246	0.000	0.741	0.960
beginner_level	0.7910	0.059	13.412	0.000	0.675	0.907
expert_level	0.3888	0.147	2.654	0.008	0.102	0.676
intermediate_level	0.5291	0.071	7.411	0.000	0.389	0.669
business_finance	0.5084	0.043	11.959	0.000	0.425	0.592
graphic_design	0.4504	0.054	8.295	0.000	0.344	0.557
musical_instruments	0.3588	0.055	6.493	0.000	0.250	0.467
web development	1.2414	0.056	22.345	0.000	1.133	1.350

- A 10% increase in the number of reviews (controlling for other independent variables) is associated with the 9.4% increase in the number of subscribers with standard error of 1.6%.
- A 10% increase in the number of lectures is associated with the 1.9% increase in the number of subscribers with standard error of 5%).
- A 10-dollar price increase is associated with a 1% increase in the number of subscribers, with standard error of 0.0004. This is surprising: it is likely caused by bimodal distribution of price.
- A 10% increase in content duration is associated with a 40% decrease in the number of subscribers with standard error of 45%. This is surprising, and requires further analysis.
- All nominal variables show positive correlation with the number of subscribers, with ‘all levels’ and ‘web development’ having the strongest correlation.

Overall, the model is statistically significant: it passes the F test, with $p < 0.05$, explaining 66.7% of overall variance in the number of subscribers. Most variables are very highly statistically significant.

Logarithmic transformation has considerably improved the model (see Appendix 3 for comparison), and the results are reasonable. The correlation between price and the number of subscribers is surprising; more on that below.

Conclusion

The data analysis has yielded several important insights:

1. The number of reviews is the strongest predictor for the number of subscribers. This implies that ATO Events should encourage customers to leave reviews and invest in other features that help spread word of mouth.
2. Web development content is more popular, but competition in this segment is also higher. However, there may be a way for the company to capitalize on this popularity by offering technical courses for airline professionals.
3. Price analysis suggests that demand is relatively inelastic. However, the analysis does not capture the fact that Udemy frequently runs discounts and special offers. Further analysis is

Further refinements

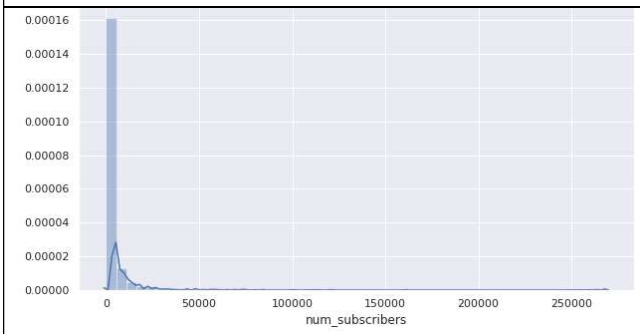
The model can be improved further to help obtain more accurate insights:

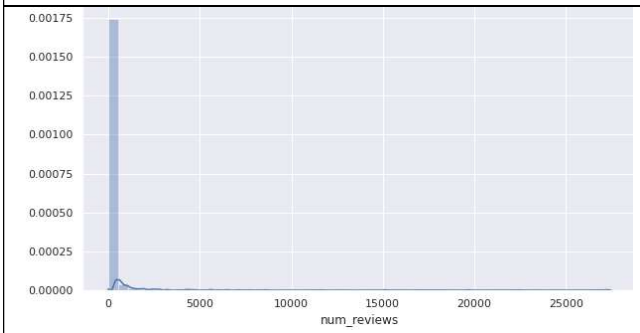
1. Applying the stepwise selection method to reduce the number of independent variables. This method has not yielded any materially different results (see Appendix 3).
2. Incorporate information on discounts to conduct more accurate price analysis. This will require web scraping, but should be relatively easy to do.
3. Price may be bimodally distributed, between longer, more expensive courses (with mean of $200), and shorter, cheaper courses (with mean of approximately $20-25). Splitting the dataset in two may yield more accurate analysis.

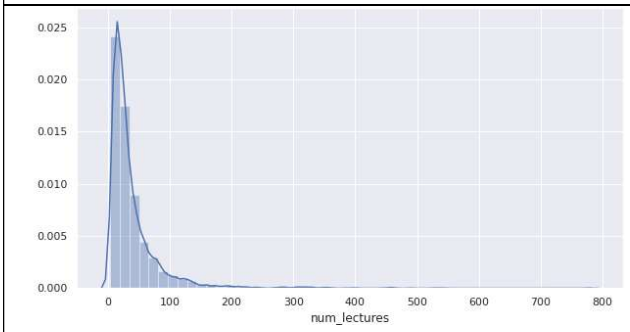
Appendix 1. Individual Plots

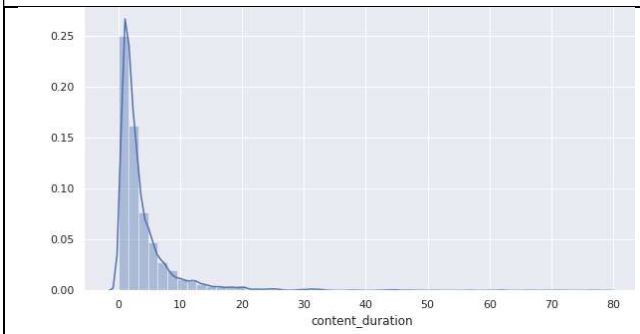
Histograms

Raw data

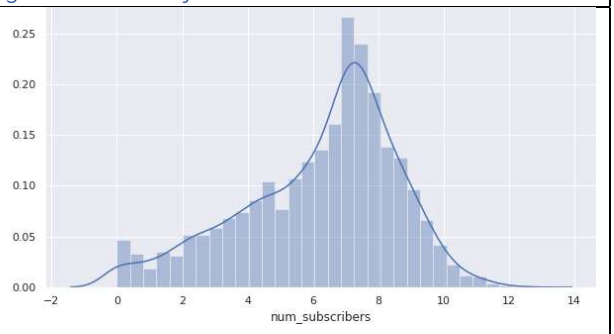


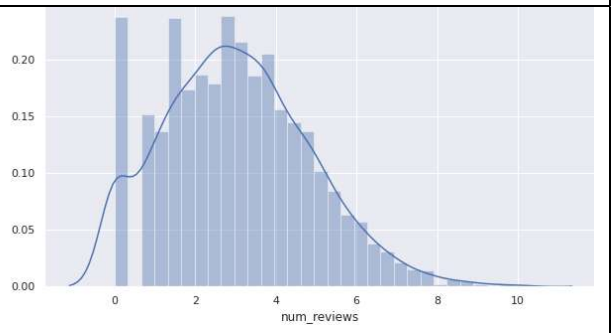


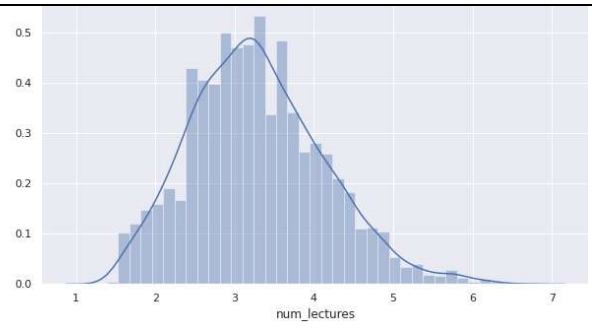


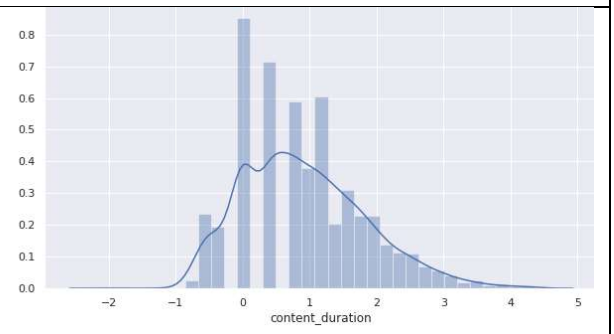


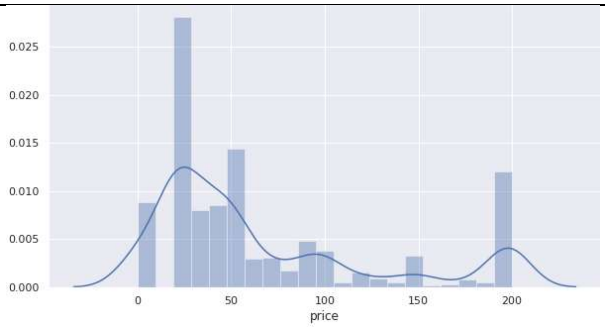
Logarithmic transformation

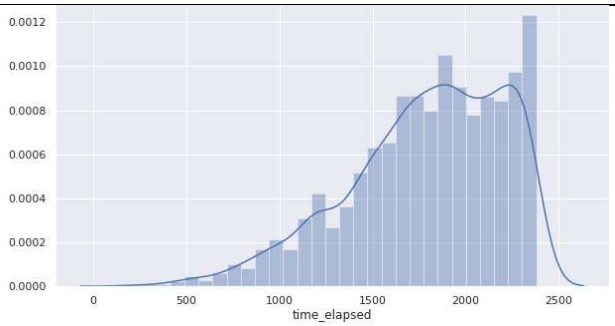






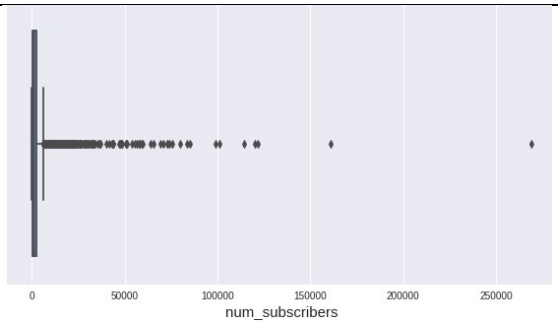


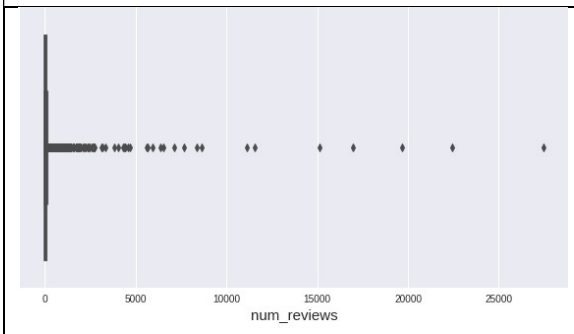




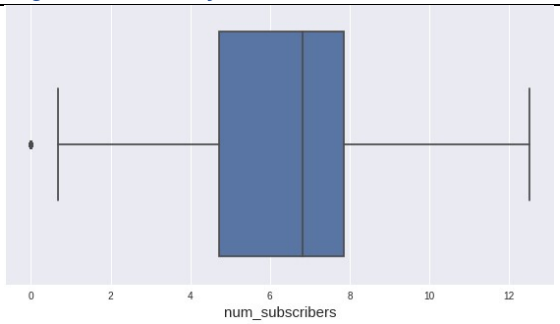
Boxplots

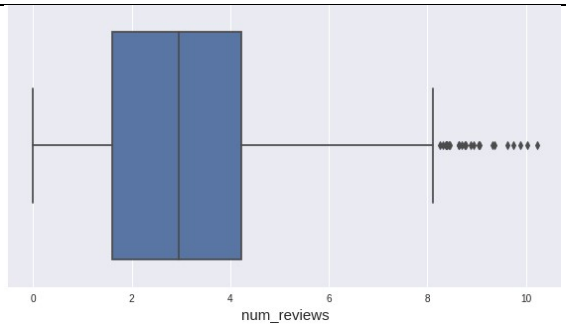
Raw data

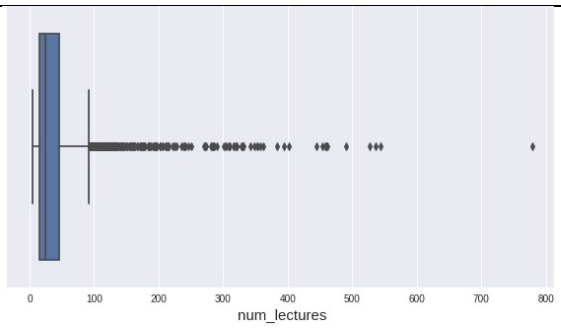


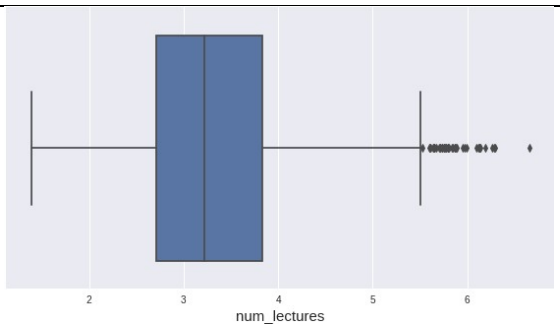


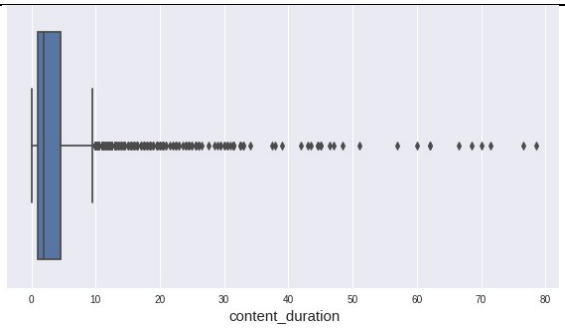
Logarithmic transformation

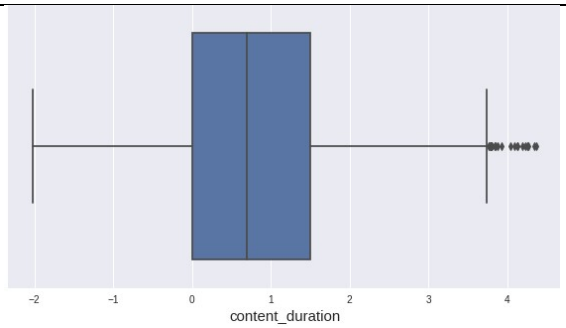


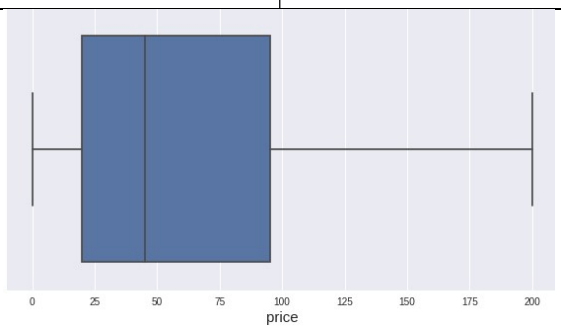


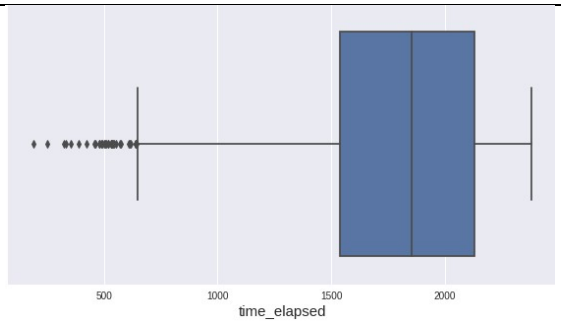

0
100
200
300
400
500
600
700
800
num_lectures


2
3
4
5
6
num_lectures


0
10
20
30
40
50
60
70
80
content_duration

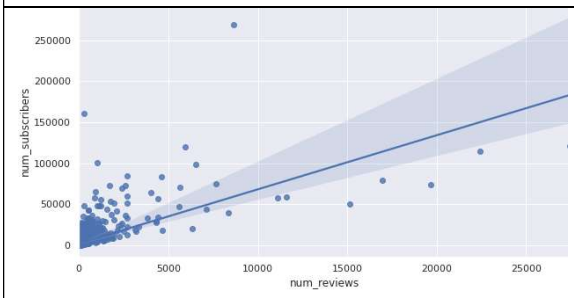

−2
−1
0
1
2
3
4
content_duration

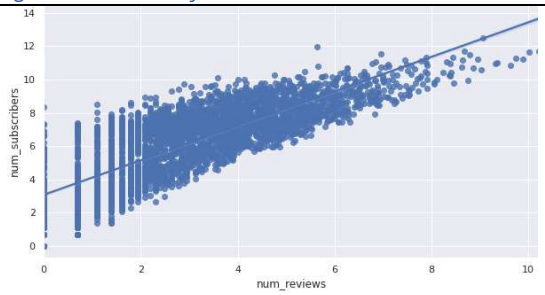

0
25
50
75
100
125
150
175
200
price

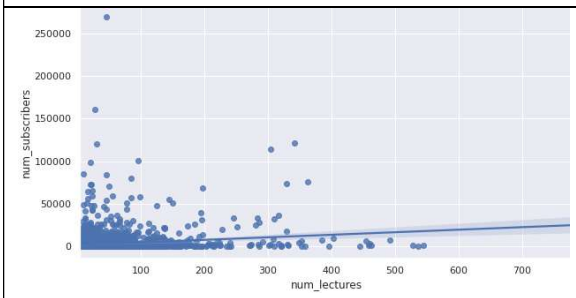

500
1000
1500
2000
time_elapsed

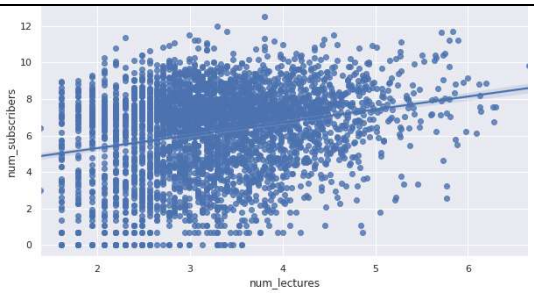
Scatterplots

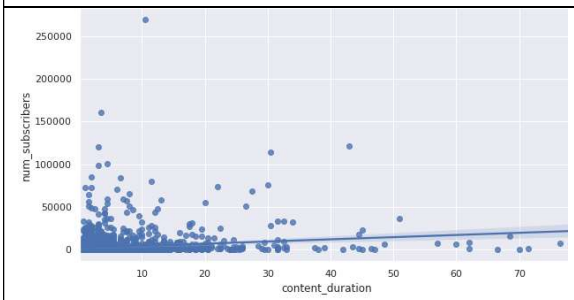
Logarithmic transformation

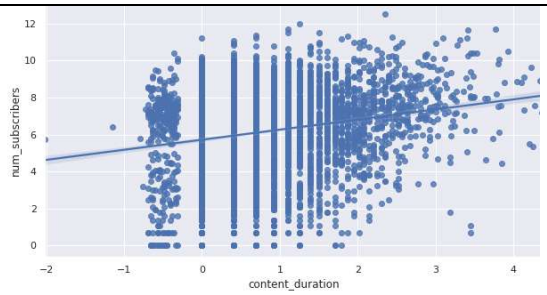


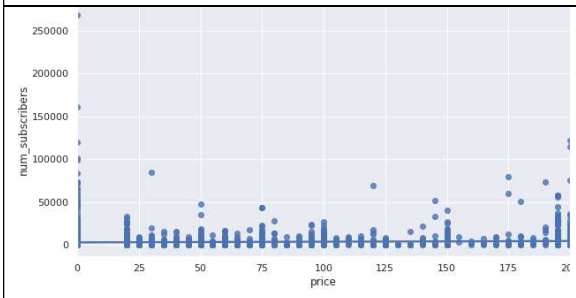


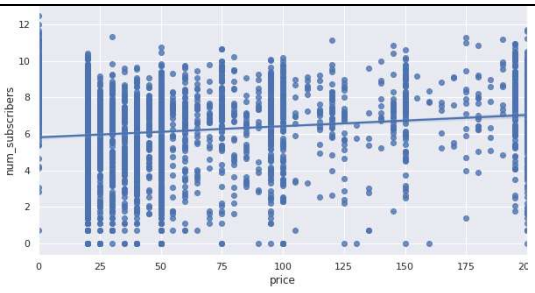


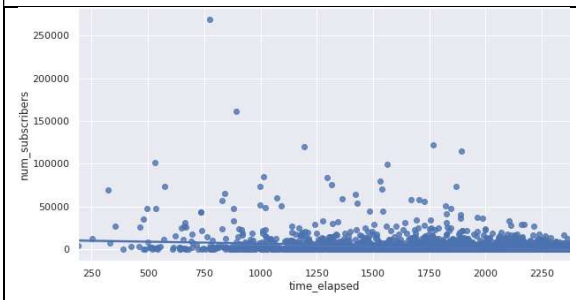


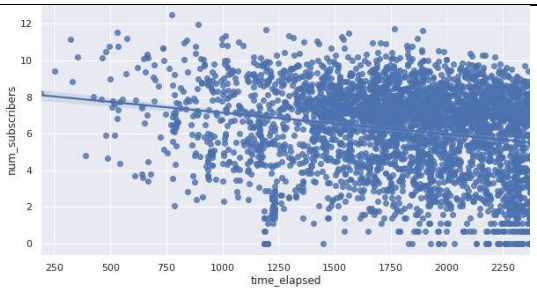






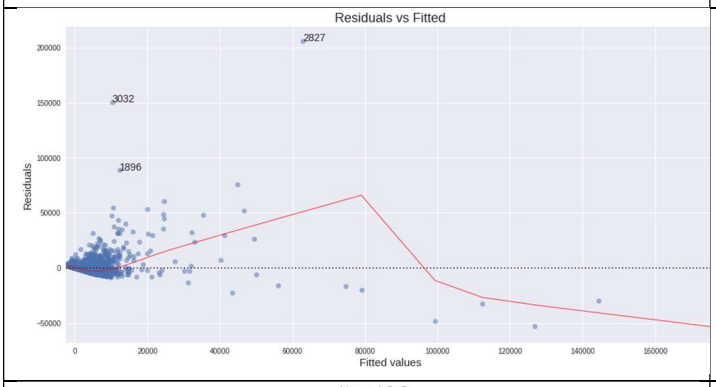


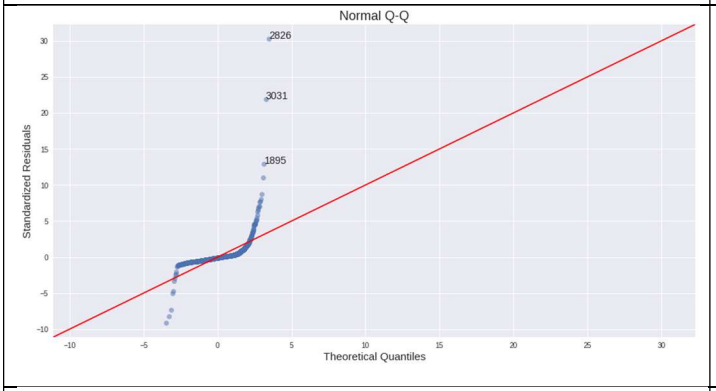


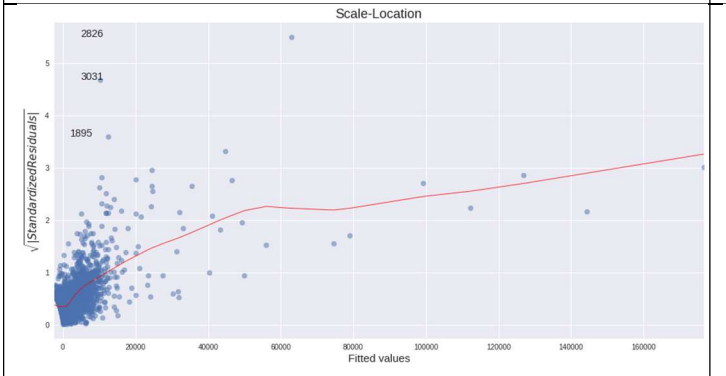


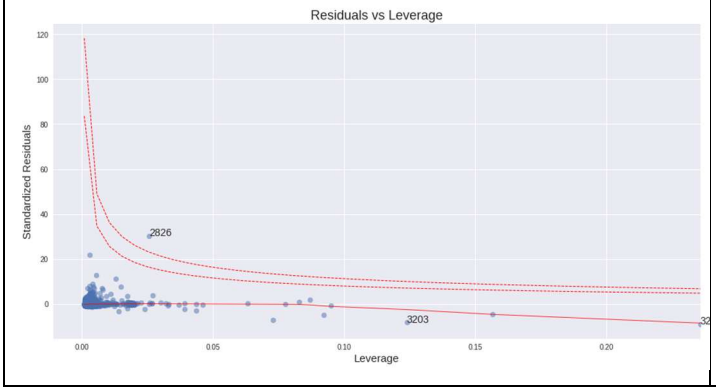
Appendix 2. Diagnostic Plots

Raw data

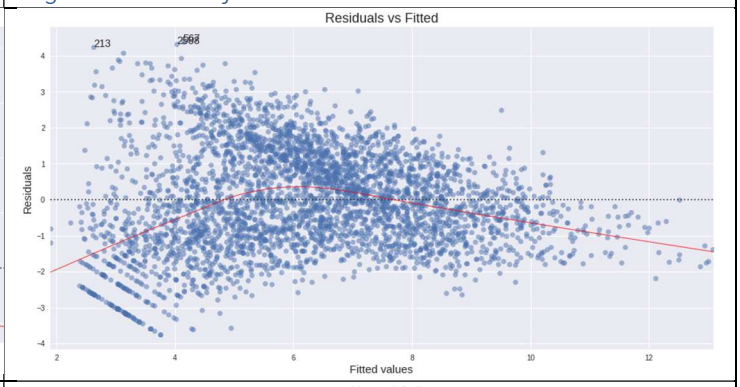


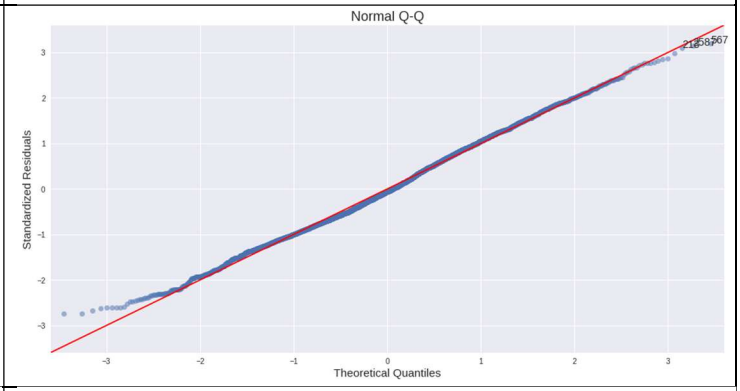


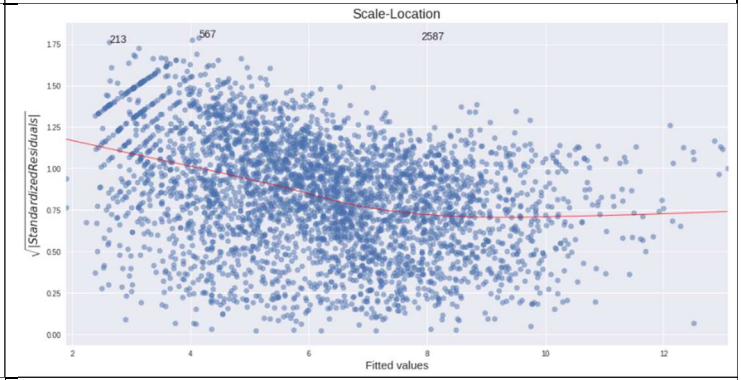


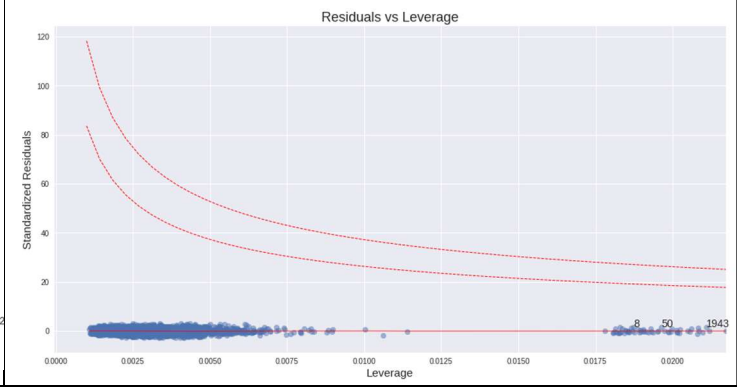


Logarithmic transformation









Appendix 3. Multiple Regression Results

Raw data

OLS Regression Results

Dep. Variable:	num_subscribers	R-squared:	0.477
Model:	OLS	Adj. R-squared:	0.475
Method:	Least Squares	F-statistic:	303.9
Date:	Sun, 14 Jun 2020	Prob (F-statistic):	0.00
Time:	18:04:19	Log-Likelihood:	-37705.
No. Observations:	3677	AIC:	7.543e+04
Df Residuals:	3665	BIC:	7.551e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5669.9521	379.159	14.954	0.000	4926.568	6413.336
price	-1.0930	2.036	-0.537	0.591	-5.085	2.899
num_reviews	6.3121	0.126	49.938	0.000	6.064	6.560
num_lectures	-4.8509	3.888	-1.248	0.212	-12.474	2.772
content_duration	8.0285	31.889	0.252	0.801	-54.493	70.550
time_elapsed	-3.7675	0.281	-13.409	0.000	-4.318	-3.217
all_levels	1746.1367	252.846	6.906	0.000	1250.404	2241.869
beginner_level	2327.0677	270.390	8.606	0.000	1796.938	2857.197
expert_level	532.4428	728.836	0.731	0.465	-896.522	1961.408
intermediate_level	1064.3050	344.320	3.091	0.002	389.227	1739.383
business_finance	516.5875	201.137	2.568	0.010	122.235	910.940
graphic_design	911.7070	259.376	3.515	0.000	403.171	1420.243
musical_instruments	210.2958	242.056	0.869	0.385	-264.282	684.874
web_development	4031.3618	215.326	18.722	0.000	3609.192	4453.531

Omnibus:	6560.342	Durbin-Watson:	1.843
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14318649.472
Skew:	12.417	Prob(JB):	0.00
Kurtosis:	307.700	Cond. No.	3.17e+19

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.25e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Logarithmic transformation

OLS Regression Results

Dep. Variable:	num_subscribers	R-squared:	0.667
Model:	OLS	Adj. R-squared:	0.666
Method:	Least Squares	F-statistic:	667.0
Date:	Sun, 14 Jun 2020	Prob (F-statistic):	0.00
Time:	17:26:41	Log-Likelihood:	-6380.8
No. Observations:	3677	AIC:	1.279e+04
Df Residuals:	3665	BIC:	1.286e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.5591	0.113	22.618	0.000	2.337	2.781
price	0.0012	0.000	2.800	0.005	0.000	0.002
num_reviews	0.9406	0.016	60.292	0.000	0.910	0.971
num_lectures	0.1915	0.050	3.852	0.000	0.094	0.289
content_duration	-0.4047	0.045	-9.067	0.000	-0.492	-0.317
time_elapsed	-0.0006	5.82e-05	-9.855	0.000	-0.001	-0.000
all_levels	0.8503	0.056	15.246	0.000	0.741	0.960
beginner_level	0.7910	0.059	13.412	0.000	0.675	0.907
expert_level	0.3888	0.147	2.654	0.008	0.102	0.676
intermediate_level	0.5291	0.071	7.411	0.000	0.389	0.669
business_finance	0.5084	0.043	11.959	0.000	0.425	0.592
graphic_design	0.4504	0.054	8.295	0.000	0.344	0.557
musical_instruments	0.3588	0.055	6.493	0.000	0.250	0.467
web_development	1.2414	0.056	22.345	0.000	1.133	1.350

Omnibus:	30.463	Durbin-Watson:	1.319
Prob(Omnibus):	0.000	Jarque-Bera (JB):	25.254
Skew:	0.135	Prob(JB):	3.28e-06
Kurtosis:	2.696	Cond. No.	2.54e+19

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.92e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Stepwise regression

OLS Regression Results

Dep. Variable:	num_subscribers	R-squared:	0.666
Model:	OLS	Adj. R-squared:	0.665
Method:	Least Squares	F-statistic:	1044.
Date:	Sun, 14 Jun 2020	Prob (F-statistic):	0.00
Time:	21:25:35	Log-Likelihood:	-6386.9
No. Observations:	3677	AIC:	1.279e+04
Df Residuals:	3669	BIC:	1.284e+04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.8614	0.162	23.907	0.000	3.545	4.178
num_reviews	0.9452	0.015	61.166	0.000	0.915	0.975
web_development	0.7823	0.056	13.896	0.000	0.672	0.893
time_elapsed	-0.0006	5.78e-05	-9.914	0.000	-0.001	-0.000
content_duration	-0.3873	0.044	-8.788	0.000	-0.474	-0.301
intermediate_level	-0.2971	0.072	-4.153	0.000	-0.437	-0.157
num_lectures	0.1723	0.049	3.546	0.000	0.077	0.268
price	0.0013	0.000	3.047	0.002	0.000	0.002

Omnibus:	33.888	Durbin-Watson:	1.324
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28.759
Skew:	0.155	Prob(JB):	5.69e-07
Kurtosis:	2.696	Cond. No.	1.35e+04

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.35e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Attachments

Attachment 1. Raw Data Set (Excel)

Attachment 2. Jupyter Notebook (Python) used for modeling