UDEMY COURSES

|  |  |  |
| --- | --- | --- |
| Name | SRN | Section |
| Aakansha Singh | PES1UG19CS004 | A |
| Aanchal Narendran | PES1UG19CS006 | A |
| Aishwarya B.S | PES1UG19CS036 | A |
| Anish Surendra | PES1UG19CS066 | A |

Table of Contents

[**Abstract**](#_x9cctclhtwnn) **1**

[**Introduction**](#_v7yu9sy9pbkv) **1**

[**Dataset Description**](#_9vbvyvnesmuz) **1**

[**Preprocessing and Data cleaning**](#_9cjoht1ezb6g) **2**

[**Exploratory Data Analysis**](#_2hk1kgm9h11b) **3**

[**Hypothesis Testing**](#_nzxuba63lk7h) **7**

[**Results and Conclusion**](#_769np8nl8u24) **8**

# 

# Abstract

The dataset being studied is the Udemy Courses dataset containing attributes which describes each course thoroughly. The dataset is cleaned by dropping the rows that corresponded with the same title and the missing values are handled. The prediction is carried out on the number of subscribers on the basis of numerous factors like the duration, price, level etc. Data visualization is done to illustrate the predictions. Hypothesis testing is enforced on the dataset where the null hypothesis is “**The average number of subscribers for Web development courses is lesser than or equal to 5500*”.*** The hypothesis test indicated that H0 cannot be rejected, so the conclusion is drawn that both H0 and H1 are plausible.

# Introduction

2020 can be rightly called “the Year of the MOOC” (Massive Open Online Courses). MOOC providers have seen a drastic growth due to the increased interest in online education.**Udemy, Inc.**  is one such American [massive open online course](https://en.wikipedia.org/wiki/Massive_open_online_course) (MOOC) provider aimed at professional adults and students. As of Jan 2020, the platform has more than 35 million students and 57,000 instructors teaching courses in over 65 languages. There have been over 400 million course enrollments. Students and instructors come from 180+ countries .

Learners expect [benefits](https://er.educause.edu/articles/2019/6/mooc-based-alternative-credentials-whats-the-value-for-the-learner), like boosting their academic or job performance, helping them start a new business and improving their application for a new job. Hence, it becomes crucial to **identify the RIGHT course** for you among the sea of courses available that caters to your skill level and the content being offered. Platforms like Udemy also need to identify the type of courses that perform the best in terms of popularity.

We have tried to harness the power of data to derive useful insights which can assist a learner find the perfect course and also help MOOC platforms in knowing what kind of courses appeal to the target audience.

**Problem Statement:** Prediction of the number of subscribers of a course based on various factors such as price, content duration and subject.

# Dataset Description

This dataset was compiled by Kaggle volunteer Larxel. It consists of information on various courses available on Udemy. It contains information about the prices, duration,level and other qualities of a course.

Each course has multiple different attributes.

* course\_id: It is a *unique integer* value assigned by udemy to the certain course
* course\_title: It is a *string value* denoting the *name* of the course
* url: It is a string value denoting the specific url of the course
* is\_paid: It is a *boolean value* denoting whether a certain course is paid or free
* price: It is an *integer value* denoting the *price* of the course
* num\_subs: It is an *integer value* denoting the number of subscribers of a course
* num\_reviews: It is an *integer value* denoting the number of reviews a course has (if any)
* num\_lectures: It is an *integer value* denoting the number of lectures available for a certain course
* level: The recommended difficulty level of each of the courses
* content\_duration: It is a decimal value denoting length of each lecture
* published\_timestamp: This is a datetime string that holds the value of the date and time when the course was published
* subject: As mentioned before, the courses can fall under the 5 main subjects

# Preprocessing and Data cleaning

There are certain attributes we do not require to meet our problem statement. We concluded after thorough analysis that course\_id, url and published date are not relevant to this problem statement so we have decided to eliminate these columns.

We also had to handle missing values. After testing for null values in the data set, we pin-point columns with null values. We see that there are 2 columns with missing values i.e price and level. We rectify the price values by imputing with the mean of all prices since it is a numerical value. We rectify the level values by filling it with the mode of level since it is a categorical value.

We handled duplicate values by dropping those rows with the same course title. We tested for typos or incorrect capitalization. Since there were none in the dataset we saved it as the processed csv.

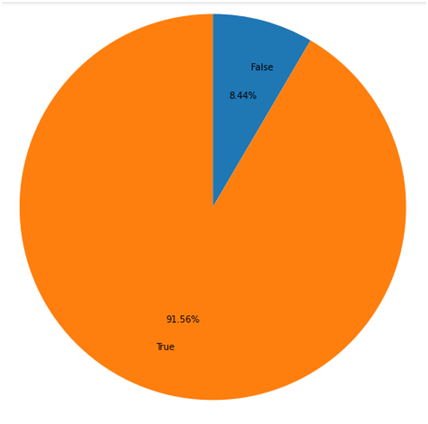
# Exploratory Data Analysis

# **1.Histogram of Price**

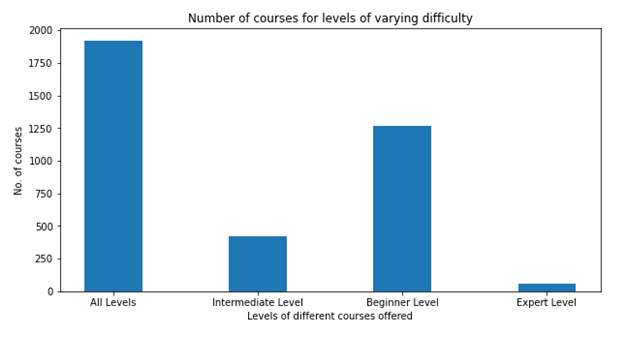
***Insights:*** This plot shows that most of the courses in Udemy lie between the price range of 25 to 50 dollars The least number between 150 to 175 dollars.

This data thus suggests that Udemy is an affordable platform for up-skilling.

**2.Paid or Unpaid**

***Insight:*** We can see that about 90.9% of the courses under the 4 subjects we deal with in this dataset are paid.This could potentially be due to the fact that courses are taught by professionals or professors of popular universities.This suggests that courses available are of good quality.

**3.Number of Courses for varying level of difficulty**

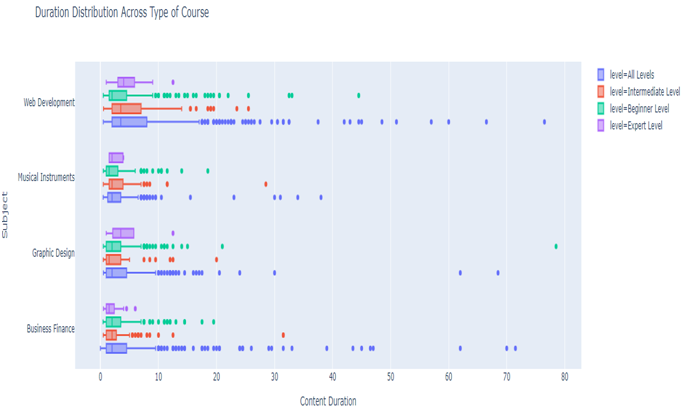
****

***Insight:***

* The number of courses applicable to “all levels” is the most with more than 1750 courses .
* The number of courses in the expert level is the least with less than 250.

The above plot makes it clear that the content available on Udemy caters to people belonging to all skill levels.

**4.** **Boxplot showing duration distribution across type of course**



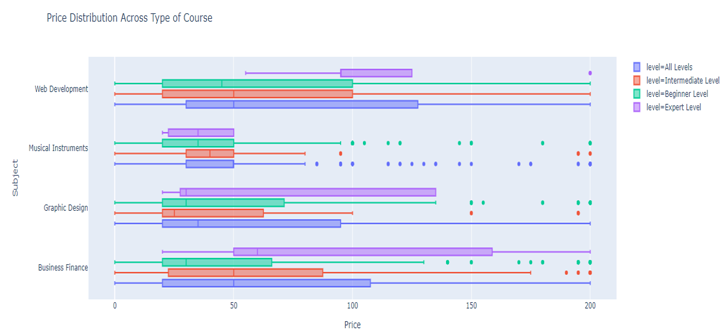
***Insights:***

* Here we can observe that most of the box plots obtained are right skewed which implies that mode<median<mean.
* Presence of outliers is observed in all levels of every subject.

The trends observed in the data suggest that Web Development has longer lectures than any other subject. Since more time on the platform is dedicated to Web Development , one may intuitively conclude that course material covered will be more along with detailed explanations which makes Udemy an ideal platform for learning this Web Development.

It is also seen that expert level courses have very less number of data points near higher quantities. This suggests that lectures for most expert level courses are short and do not contain detailed explanation of course material thereby increasing the difficulty level of the course.

**5.** **Boxplot showing Price distribution across each subject**

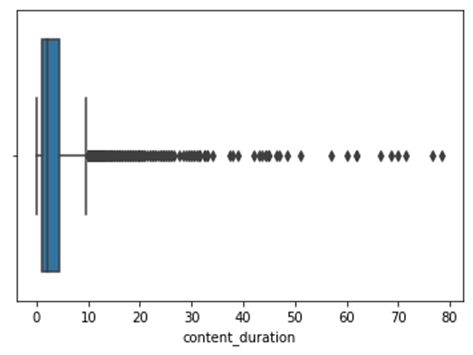
****

***Insights:***

* The box plots obtained are right skewed which implies that the mode < median < mean.
* Presence of outliers in prices are observed in all the subjects except Web Development.

We can observe that most of the courses belonging to the subject Musical Instruments are not very expensive. The expert-level courses are usually more expensive as compared to the other levels.

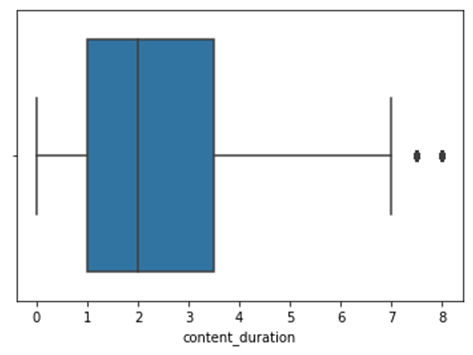
**6. Analyzing Course Duration Distribution:**

****

***Insights:***

* Since we observe the presence of many outliers we remove the outliers based on the z-score value being less than 3.

**Box-Plot on removal of outliers**

****

Post outlier removal with a threshold of three standard deviations (i.e 97% of the data), we notice the boxplot has a much more symmetric appearance.

# Hypothesis Testing

In this dataset, the number of subscribers play a great importance as most online education platforms are heavily reliant on their subscribers. Hence, it was only appropriate that our hypothesis was based on the number of subscribers.

Our Research Hypothesis is

***Ha: The average number of subscribers for Web Development courses is greater than 5500.***

The contradictory Null Hypothesis would be

***H0: The average number of subscribers for Web development courses is lesser than or equal to 5500.***

Here the population parameter in consideration is the population mean.

We take a random sample from the population. The population in question here consists of all the Web Development courses specified in the dataset. The sample consists of 141 randomly chosen courses from this population. On calculation we get the below statistics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistic | Sample mean | Sample Standard deviation | Size | Alpha |
| Value | 5708 | 8748 | 141 | 0.05 |

We start with the P-Value test. Let us assume that the Null Hypothesis is true. Then we find the Null distribution of the data. Under the Null distribution, we obtain a mean that is equivalent to the border value of the hypothesis and a standard deviation derived from that of the sample. Hence we obtain the below statistics.

|  |  |  |
| --- | --- | --- |
| Statistic | Null distribution mean | Null distribution standard deviation |
| Value | 5500 | 736 |

We now perform a Z-test since the sample size is greater than 30 in order to obtain a Z-score value. We then use this Z-Score value to obtain the p-value which is the probability value for the sample.

|  |  |  |
| --- | --- | --- |
| Statistic | Z-score | P-value |
| Value | 0.283 | 0.389 |

Here, we see that the P-value is greater than the alpha (0.05) value. Hence, we conclude that both H0 and Ha are plausible. In this case, a sample has a 38.9% probability of disagreeing with H0.

# Results and Conclusion

This project was truly successful in bringing and pinpointing insights into various aspects that go into course selection and segregation. It also goes to prove that one requires far too many features than available to predict if a certain course would be able to garner the attention of users and push them to subscribe to the course. That said the correlational properties of various columns as well as the hypothesis test, proves that subscribers can be correlated to the subject at hand. The graphs demonstrated in exploratory data analysis show that the payment aspect of a course does not hinder a humans passion to learn the subject. This is due to the fact that courses priced at 175-200 also manage to have the same number of subscribers as those between 0-25. It also shows that courses with content that can be classified as All levels tend to have more subscribers as opposed to courses at expert level. In conclusion, we believe the project was very insightful in demonstrating how generic and common assumptions could be wrong such as the correlation between price and subscribers which was proven to be false.