**Name: Chimge Tsendnyam**

**Project Title: Udemy Courses Statistical Tests & Data Analysis**

**Introduction:**

Udemy is an online learning and teaching platform aimed at professional adults and students. Udemy has over 100,000 courses and 24 million students. Sample dataset that I picked for my project is “Udemy Courses” from Kaggle which has over 3,600 courses listings from 4 different subjects. “Udemy Courses” dataset has course related information such as the cost, duration, number of lectures, difficulty, date the course was published, number of subscribers and number of reviews. I will be using R to do a statistical tests and analysis.

**Research Questions:**

Professional adults or students who take online classes might be interested in different types of courses available on Udemy, their price range, course components and they might ask what classes have higher number of subscribers, and reviews. The data analysis that look to derive from this dataset is to not only to help answer those questions, but ask more questions such as:

* Is there any correlation amongst Udemy Course variables?
* How much of an effect does cost or any of the courses related information have on the number of subscribers and number of reviews?
* Is there any variable(s) that have an impact on the price of the course?

My hypothesis is that courses that are free have higher number of subscribers and reviews compared to the courses that cost money. Another assumption is that the higher the number of the subscribers are, higher the number of reviews is. Answering all these questions by analyzing the data will help professional adults and student to make a decision on picking a course, if they are looking to do online classes on Udemy.

* + - **Summarize how you addressed this problem statement (the data used and the methodology employed).**

**Data and the methodology:**

After exploring and cleaning my dataset, my clean dataset and its summary looked like this:

A close up of text on a white background

Description automatically generated

**The methodologies I used:**

* Covariance and correlation analysis to look at the type of relationships, impacts as well as causalities between the Udemy Courses variables.
* For the variables that have strong positive linear relationships, coefficient determination is used measure the amount of variability in one variable that is shared by the other
* Partial correlation, controlling variables such as the year see if that changes the correlation outcome of the variables with strong relationship
* Linear Regression analysis to predict the variables such as the number of subscribers and the number of reviews
* Multiple Regression analysis where number of lectures and course duration predict the price.
* Logistic Regression to see which variable(s) had the greatest effect of z-statistic is significantly different from 0.
* KNN or the nearest algorithm to fit the model on price & number of subscribers
  + - **Summarize the interesting insights that your analysis provided.**

**Preliminary analysis findings show:**

* Top 10 courses with the highest number of subscribers are listed below. Out of top 10 courses with the highest number of subscribers, 9 are for web development and 1 is for musical instrument. Also, 6 are for free and 4 are for an average price of $151. Average number of lectures is 100 with an average duration of 12 hours. 7 out of top 10 courses had difficulty of all levels and remaining 3 only had difficulty of beginner level.

|  |  |
| --- | --- |
| **course\_title** | **num\_subscribers** |
| Learn HTML5 Programming From Scratch | **268923** |
| Coding for Entrepreneurs Basic | **161029** |
| The Web Developer Bootcamp | **121584** |
| Build Your First Website in 1 Week with HTML5 and CSS3 | **120291** |
| The Complete Web Developer Course 2.0 | **114512** |
| Free Beginner Electric Guitar Lessons | **101154** |
| Web Design for Web Developers: Build Beautiful Websites! | **98867** |
| Learn Javascript & JQuery From Scratch | **84897** |
| Practical PHP: Master the Basics and Code Dynamic Websites | **83737** |
| JavaScript: Understanding the Weird Parts | **79612** |

* Top 10 courses with the highest number of reviews are also listed below. 4 out of the top 10 courses with the highest number of reviews were also part of top 10 courses with the highest number of subscribers. It will be interesting to see the correlation analysis between *num\_subscribers* & *num\_reviews* later on. Similar to the previous findings, out of top 10 courses with the highest number of reviews, 9 are also for web development and 1 is for musical instrument. Also, 1 is for free and 9 are for an average price of $185. Average number of lectures is 196 with an average duration of 20 hours. All 10 courses had difficulty of all levels.

|  |  |
| --- | --- |
| **course\_title** | **num\_reviews** |
| The Web Developer Bootcamp | **27445** |
| The Complete Web Developer Course 2.0 | **22412** |
| Angular 4 (formerly Angular 2) - The Complete Guide | **19649** |
| JavaScript: Understanding the Weird Parts | **16976** |
| Modern React with Redux | **15117** |
| Learn and Understand AngularJS | **11580** |
| Learn and Understand NodeJS | **11123** |
| Learn HTML5 Programming From Scratch | **8629** |
| Angular 2 with TypeScript for Beginners: The Pragmatic Guide | **8341** |
| Pianoforall - Incredible New Way To Learn Piano & Keyboard | **7676** |

**R Statistical Analysis:**

**Covariance:**

Covariance is used to measure how much two random variables are linearly related. As shown below, **number of subscribers** and **number of reviews** had the largest **positive covariance** of **5779871,** indicating our previous speculation of the higher the number of subscribers are, higher the number of the reviews is. On the other hand, **year courses were published,** and their **number of subscribers** had a strong **negative covariance** of **-2080,** indicating that more recently the courses were published, the lower their number of subscribers are. Similarly, the **year the courses were published** and the **number of reviews** those courses have based on **the negative covariance** of **-59**. It only makes sense, since the courses that were published long time ago had been on a market for a longer period compared to some of the recently published courses, giving it more time to gain higher number of subscribers and reviews.

> cov(u\_courses$num\_subscribers, u\_courses$num\_reviews)

[1] **5779871**

> cov(u\_courses$price, u\_courses$num\_subscribers)

[1] **29404.31**

> cov(u\_courses$price, u\_courses$num\_reviews)

[1] **6488.201**

> cov(u\_courses$duration, u\_courses$num\_subscribers)

[1**] 9303.382**

> cov(u\_courses$duration, u\_courses$num\_reviews)

[1] **1295.933**

> cov(u\_courses$duration, u\_courses$price)

[1] **108.229**

> cov(u\_courses$year, u\_courses$price)

[1] **9.613889**

> cov(u\_courses$year, u\_courses$num\_reviews)

[1] **-59.15352**

> cov(u\_courses$year, u\_courses$num\_subscribers)

[1] **-2080.235**

**Correlation Coefficients:**

As I do the correlation analysis on the Udemy Course variables, you will soon discover that the validity of the covariance analysis is being reaffirmed by the correlation analysis. As you can see on the below correlogram chart, the **number of lectures** and **course duration** have the highest correlation of **0.8,** meaning that when number of lectures moves in either direction, the length of the course follows that direction. Moreover, the **number of subscribers** and **number of reviews** which is **0.6,** meaningthat as the number of subscribers increase, the number of reviews also increases, vice versa. As we look at the correlations that are closer to 0, or almost no relationship, one of our research questions, whether the price has any impact on the number of subscribers and reviews is being answered. With correlation **of 0.1, price** has minimal impact on the **number of subscribers** and **reviews.** As we look at some of the negative relationships, **course year** and **number of subscribers** had negative correlation of **-0.2,** reaffirming our previous covariance analysis.

A screenshot of a cell phone

Description automatically generated

**Coefficient of Determination:**

Previously, I found 2 major positive linear relationships and thus, looked at the coefficient of determination (R^2) which is a useful measure of the substantive importance of an effect. As you can see below the **number of lectures** share **64.2%** of the variability in **course duration.** **Number of Subscribers** share **42%** of the variability in the **number of reviews.** Despite the one variable sharing the certain percentage in the other variable, we cannot infer a causal relationship from them.

> data\_frame <- data.frame(u\_courses$price, u\_courses$num\_subscribers, u\_courses$num\_reviews, u\_courses$num\_lectures, u\_courses$year, u\_courses$duration)

> data\_matrix <- data.matrix(data\_frame)

> cor(data\_matrix)^2\*100

A close up of text on a white background

Description automatically generated

**Partial Correlation:**

Partial correlation between number of subscribers and number of reviews is 0.65 which is exactly same as the correlation when the effect of year is not controlled for (0.6). The correlation is still statistically significant, because our p-value is 0.04, which is still below 0.05. In terms of variance, the value of R^2 is for the partial correlation is 0.4 (42%), which is exactly same as when the effects of year was not controlled for. Therefore, the inclusion of year did not change the amount of variation in number of reviews shared by number of subscribers.

pcor(c("num\_subscribers", "num\_reviews", "year"), var(courses\_matrix1))

[1] **0.6522195**

> pc <- pcor(c("num\_subscribers", "num\_reviews", "year"), var(courses\_matrix1))

> pc^2

**[1] 0.4253903**

> pcor.test(pc, 1, 11)

$tval

[1] 2.433617

$df

[1] 8

$pvalue

**[1] 0.04097039**

**Linear Regression:**

**Regression Analysis where number of subscribers predict number of reviews:**

**On average for courses with 10,000 subscribers, we expect to see 552 reviews.**

**Y = -48.2 + 0.06\*(10,000)= 552**

regresion1 <- lm(u\_courses$num\_reviews ~ u\_courses$num\_subscribers)

> summary(regresion1)

Call:

lm(formula = u\_courses$num\_reviews ~ u\_courses$num\_subscribers)

Residuals:

Min 1Q Median 3Q Max

-9974.1 -37.8 39.0 50.7 19715.3

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -**48.282114** 12.373346 -3.902 0.0000971 \*\*\*

u\_courses$num\_subscribers **0.063972** 0.001234 51.844 < 0.0000000000000002 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 711.1 on 3675 degrees of freedom

Multiple R-squared: 0.4224, Adjusted R-squared: 0.4223

F-statistic: 2688 on 1 and 3675 DF, p-value: < 0.00000000000000022

**Regression Analysis where course price predicts the number of subscribers:**

**On average for courses that are free, we expect 2676 people sign up for it.**

**Y = 2676 + 7.9\*(0)= 2676**

> regresion2 <- lm(u\_courses$num\_subscribers ~ u\_courses$price)

> summary(regresion2)

Call:

lm(formula = u\_courses$num\_subscribers ~ u\_courses$price)

Residuals:

Min 1Q Median 3Q Max

-4256 -2865 -2343 -607 266247

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) **2676.140** 230.796 11.595 <0.0000000000000002 \*\*\*

u\_courses$price **7.900** 2.567 3.078 0.0021 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9494 on 3675 degrees of freedom

Multiple R-squared: 0.002571, Adjusted R-squared: 0.0023

F-statistic: 9.473 on 1 and 3675 DF, p-value: 0.002101

**Multiple Regression:**

**R^2 for our multiple regression is 0.11, meaning that number of lectures and course duration account for the 11% of the variation in the course price.**

> multiple\_regression <- lm(price ~ num\_lectures + duration, data = u\_courses)

> summary(multiple\_regression)

Call:

lm(formula = price ~ num\_lectures + duration, data = u\_courses)

Residuals:

Min 1Q Median 3Q Max

-175.48 -36.84 -19.91 25.25 148.07

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 49.8376 1.2145 41.037 < 0.0000000000000002 \*\*\*

num\_lectures 0.3221 0.0315 10.226 < 0.0000000000000002 \*\*\*

duration 0.8056 0.2622 3.072 0.00214 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57.53 on 3674 degrees of freedom

Multiple R-squared**: 0.1112,** Adjusted R-squared: 0.1107

F-statistic: 229.9 on 2 and 3674 DF, p-value: < 0.00000000000000022

**Logistic Regression:**

**Based on the summary table, year that the course was published had the greatest impact on the price with z statistics of 66.93. Next, number of lectures had second biggest effect with z statistics of 60.71. Third, course duration with z statistics of 25.23.**

> log\_regression <- glm(price ~ num\_lectures + duration + year, data = u\_courses, family = "poisson")

> summary(log\_regression)

Call:

glm(formula = price ~ num\_lectures + duration + year, family = "poisson",

data = u\_courses)

Deviance Residuals:

Min 1Q Median 3Q Max

-17.939 -5.548 -2.714 2.914 17.908

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -237.60083218 3.61022076 -65.81 <0.0000000000000002 \*\*\*

num\_lectures 0.00271685 0.00004475 **60.71** <0.0000000000000002 \*\*\*

duration 0.00988115 0.00039157 **25.23** <0.0000000000000002 \*\*\*

year 0.11987940 0.00179110 **66.93** <0.0000000000000002 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 192513 on 3676 degrees of freedom

Residual deviance: 172049 on 3673 degrees of freedom

AIC: 191574

Number of Fisher Scoring iterations: 5

**Binary data and KNN Accuracy:**

**I picked is\_paid, price and number of subscribers for my binary data got a result of 91.7% of accuracy, meaning that 91.7% of the time this model predicts the correct result. Also, for k-values between 3 and 25, I also received 93.8-91.9% accuracies, which is not a significant change from result of k=50.**

> head(udemy\_courses\_binarydata)

# A tibble: 6 x 3

is\_paid\_numeric price num\_subscribers

<dbl> <dbl> <dbl>

1 1 65 1540

2 1 200 827

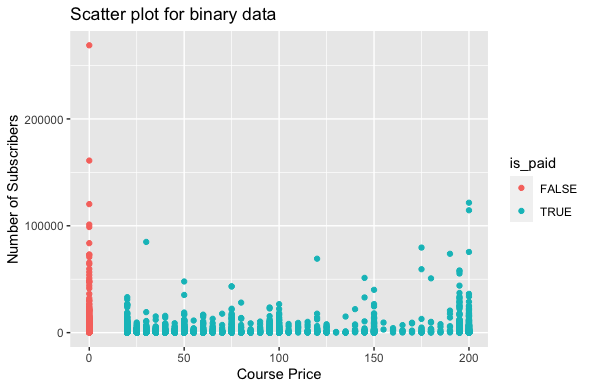
3 1 200 1380

4 1 50 1916

5 1 115 7489

6 1 30 6219

> ggplot(udemy\_courses\_binarydata, aes(x = price, y = num\_subscribers, color = is\_paid)) + geom\_point() + ggtitle("Scatter plot for binary data") + labs(x = "Course Price", y = "Number of Subscribers")

****

> normalize <- function(price) {

+ return ((price - min(price)) / (max(price) - min(price))) }

> binary\_data.n <- as.data.frame(lapply(udemy\_courses\_binarydata[,2:3], normalize))

> binary.subset.n <- as.data.frame(lapply(udemy\_courses\_binarydata[,2:3], normalize))

> set.seed(123)

> dat.d <- sample(1:nrow(binary.subset.n),size=nrow(binary.subset.n)\*0.7,replace = FALSE)

> train.binary.data <- udemy\_courses\_binarydata[dat.d,]

> test.binary.data <- udemy\_courses\_binarydata[-dat.d,]

> train.binary.is\_paid\_numeric <- udemy\_courses\_binarydata[dat.d,1]

> test.binary.is\_paid\_numeric <- udemy\_courses\_binarydata[-dat.d,1]

> NROW(train.binary.is\_paid\_numeric)

[1] 2573

> knn.50 <- knn(train=train.binary.data, test=test.binary.data, cl=train.binary.is\_paid\_numeric$is\_paid\_numeric, k=50)

> ACC.50 <- 100 \* sum(test.binary.is\_paid\_numeric$is\_paid\_numeric == knn.50)/NROW(test.binary.is\_paid\_numeric$is\_paid\_numeric)

> ACC.50

**[1] 91.75725**

> i = 1

> k.optm = 1

> for (i in 3:25) {

+ knn.mod <- knn(train = train.binary.data, test = test.binary.data, cl = train.binary.is\_paid\_numeric$is\_paid\_numeric, k = i)

+ k.optm[i] <- 100 \* sum(test.binary.is\_paid\_numeric$is\_paid\_numeric == knn.mod)/NROW(test.binary.is\_paid\_numeric$is\_paid\_numeric)

+ k = i

+ cat(k, '=', k.optm[i],'')

+ }

**3 = 93.84058 4 = 93.20652 5 = 92.30072 6 = 92.21014 7 = 92.21014 8 = 91.66667 9 = 91.84783 10 = 92.3913 11 = 92.21014 12 = 91.75725 13 = 92.02899 14 = 92.11957 15 = 92.11957 16 = 92.11957 17 = 92.02899 18 = 91.93841 19 = 91.75725 20 = 91.84783 21 = 91.84783 22 = 91.84783 23 = 91.84783 24 = 91.84783 25 = 91.93841**

* + - **Summarize the implications to the consumer (target audience) of your analysis.**

In summary, most of my hypothesis are turned out to be true. Some have higher degree of affirmation than the others. For example, one of my previous assumptions was that the courses that are free have higher number of subscribers. As you can see on below scatter plot, there is more free courses that have higher number of subscribers. Top 2 courses that have the highest number of subscribers are free, confirming the hypothesis is true. Top 5 courses that have the highest number of subscribers are Web Development courses. However, to answer the question on whether, price has an impact on people’s decision, we have previously gotten 0.1 correlation coefficient, meaning that, if price is any impact, it is very minimal.

A screenshot of a cell phone

Description automatically generated

Similar to the previous analogy, another assumption was that the courses that are free have higher number of reviews. As you can see below on the scatter plot, the courses that have the highest number of reviews are actually courses that are in the price range of $150-$200, confirming that the hypothesis is not true. Top 7 courses that have the highest number of reviews are Web Development courses.

Also, we have seen a linear, positive and somewhat stronger relationship between number of subscribers and number of reviews. The courses that have higher number of subscribers tend to have higher number of reviews, confirming our previous hypothesis as well.

**A picture containing photo, different, table, large

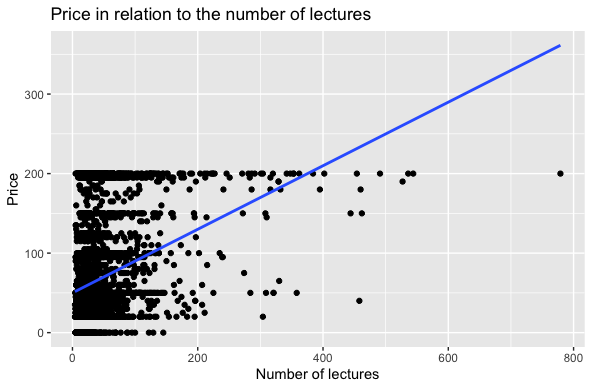
Description automatically generated**

The strongest, positive and linear relationship we seen is the number of lectures and course duration, which is something that I had not thought of before starting my data analysis. It totally makes sense for courses to have lengthier duration for the ones that have larger number of lectures.

**A picture containing photo, people, different, water

Description automatically generated**

Another big question that I had in the beginning was that is there any variable(s) that have an impact on the price. I have done multi-regression, logistic regression analysis to find out number of lectures and course duration have somewhat of an impact on the price. As you can see below, below 2 graphs show somewhat positive and linear relationships. You can see that as the number of lectures get higher, especially if courses have 200 or more lectures in a course, it is almost guaranteed that there is a cost associated with it. Similar to that, courses that have 25 or more hours in course duration, almost all of it have price associated with them.

****

**A screenshot of a cell phone

Description automatically generated**

In conclusion, in other to help with professional adults or students who take online classes might be interested in different types of courses available on Udemy, their price range, course components, I put together below graphs:

Below graph visualizes the type of courses by count using bar plot which helps answer the type of courses are making up our over 3600 courses in our dataset. As you can see, almost 2/3 of the course counts are composed of Business Finance & Web Development courses.

A screenshot of a cell phone

Description automatically generated

Below graph visualizes the course price by count using histogram which helps understand the price range of Udemy courses. As you can see their courses cost between $0-200 and majority of the courses or about 800 out of 3600 courses cost $20. Only about 300 out of our 3600 courses are free of charge.

A picture containing clock, drawing

Description automatically generated

Similar to the previous graph, below graph visualizes the course price by count using histogram. It also makes the distinction of course subject by color. As you can see our large chunk of courses that are free are Web Development courses, and the largest chunk of our courses that cost $20 is Business Finance courses.

A picture containing screenshot

Description automatically generated

* + - **Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.**

Even though the Udemy courses dataset was pretty straightforward in terms of variables that are involved and the number of samples in the data, I felt like it lacked more insights. For instance, instead of number of reviews each course has, it would have been great to have the actual score reviews. Also, one variable that I did not touch on was course difficulty levels. I’m not sure how I would have used simple definitions such as “beginner level”, “intermediate level”, and “all levels” to derive greater insight in relation to my other variables such as price and number of reviews. Not sure, if this would have given any insights, but It might have been interesting to see if the month of the year that course was published have anything to the with number of subscribers. For instance, there could be a case of drop in large number of subscribers for the summer months or courses that are published in the summer just because people tend to have summer vacations, travels, etc during summer.