```
data <- read.csv("~/Desktop/final project VS code/data.csv", header = T)
```

Now, I would like to do some statistics on the data set. Specifically, can we predict rating based on the other factors like price, level, content duration, number of lectures, and subject? Also, how accurate is the prediction?

I am going to create the model scores using the lm function in R. *lm* is used to fit linear models. In order to make sure that the scores are accurate and statistically significant, I will perform a variable selection method. Specifically, the backwards elimination. I will begin by including each variable in the model. Then, I will slowly remove the **most** significant one. I will do this one by one and stop once no variables need to be dropped.

Since the web development courses have no ratings(it was all null values), I will use the model to predict their scores!

```
# create data set with all but WD because null model scores
data_business = subset(data, subject == 'Business Finance')
data_design = subset(data, subject == 'Graphic Design')
data_music = subset(data, subject == 'Musical Instruments')

data_all_noWD <- data_business
# add design
data_all_noWD <- rbind(data_all_noWD, data_design)

# add music
data_all_noWD <- rbind(data_all_noWD, data_music)
```

```
lmod <- lm(Rating ~ price + num_reviews + num_lectures + level +
             content_duration, data = data_all_noWD)

summary(lmod)$coefficients
```

```
##                              Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)              6.010344e-01 1.280858e-02 46.9243458 0.000000e+00
## price                    3.456898e-04 1.246429e-04  2.7734418 5.588427e-03
## num_reviews              1.908899e-05 2.693574e-05  0.7086864 4.785861e-01
## num_lectures            -1.178433e-03 2.488084e-04 -4.7363087 2.299942e-06
## levelBeginner Level     -2.673678e-02 1.523117e-02 -1.7553986 7.931536e-02
## levelExpert Level        1.269877e-02 5.283055e-02  0.2403679 8.100651e-01
## levelIntermediate Level -1.466619e-02 2.246219e-02 -0.6529277 5.138638e-01
## content_duration         6.748266e-03 1.945924e-03  3.4678986 5.335359e-04
```

```
# remove level

lmod2 <- lm(Rating ~ price + num_reviews + num_lectures +
             content_duration, data = data_all_noWD)

summary(lmod2)$coefficients
```

```
##                       Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)       5.900100e-01 1.079174e-02 54.6723591 0.000000e+00
## price             3.528867e-04 1.245557e-04  2.8331646 4.646393e-03
## num_reviews       1.902417e-05 2.693743e-05  0.7062357 4.801082e-01
## num_lectures     -1.181902e-03 2.488090e-04 -4.7502381 2.148242e-06
## content_duration  6.762352e-03 1.945564e-03  3.4757802 5.181642e-04
```

```
# remove num_reviews
final_model <- lm(Rating ~ price + num_lectures + content_duration,
            data = data_all_noWD)

summary(final_model)$coefficients
```

```
##                       Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)       0.5901000501 0.0107898941 54.690069 0.000000e+00
## price             0.0003564376 0.0001244415  2.864298 4.214642e-03
## num_lectures     -0.0011610837 0.0002470315 -4.700144 2.742670e-06
## content_duration  0.0067919358 0.0019449153  3.492150 4.875658e-04
```

So, final_model is the end linear model to predict the rating for each Udemy course. Each predicitor variable is significant since the p-value is smaller than 0.05.

Now, let's get the model scores (i.e. predictors) from the final_model.

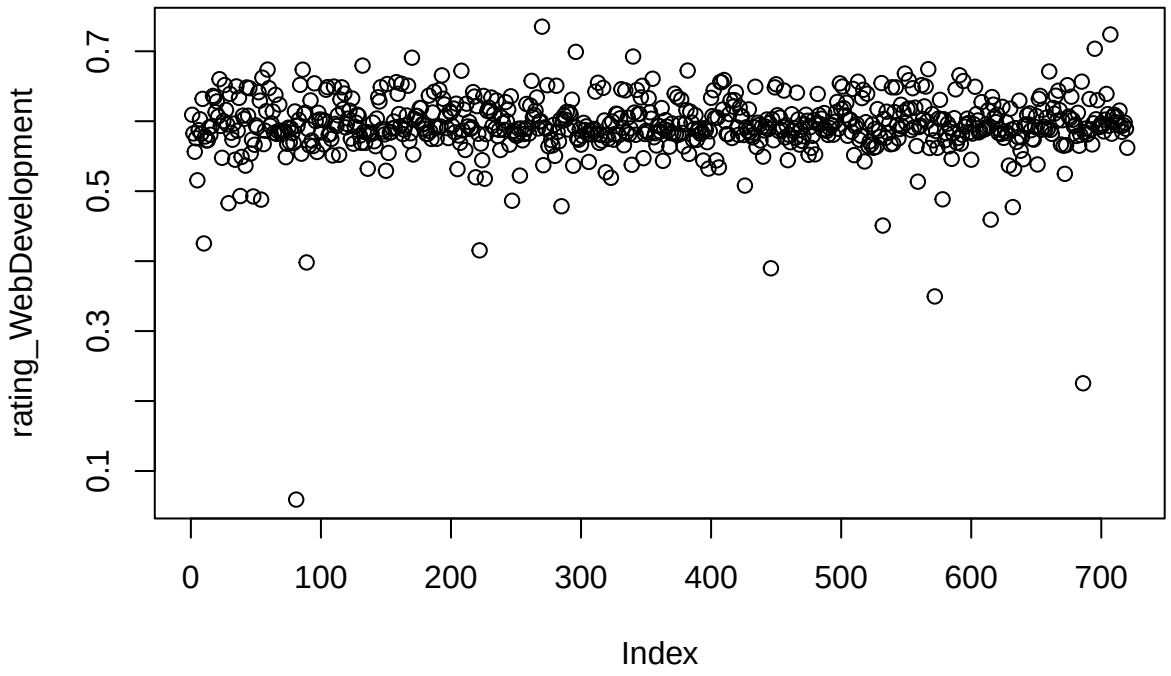
Following, I will export these model scores as a CSV file so that I can visualize them.

```
data_wd = subset(data, subject == 'Web Development')

rating_WebDevelopment <- predict.lm(final_model, data_wd)
summary(rating_WebDevelopment)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05914 0.58043 0.59225 0.59361 0.61254 0.73511
```

```
plot(rating_WebDevelopment)
```




```
ratings <- cbind(rating_WebDevelopment, data_business$Rating, data_music$Rating,
                 data_design$Rating)
```

```
## Warning in cbind(rating_WebDevelopment, data_business$Rating,
## data_music$Rating, : number of rows of result is not a multiple of vector length
## (arg 1)
```

```r
colnames(ratings) <- c('web development', 'business', 'music', 'design')
write.csv(ratings,"~/Desktop/final project VS code/NEW2.csv")
```