How many observations in each category?

```
# import data and make subsets
data <- read.csv("~/Desktop/final project VS code/data.csv", header = T)
data_business = subset(data, subject == 'Business Finance')
data_design = subset(data, subject == 'Graphic Design')
data_music = subset(data, subject == 'Musical Instruments')
data_wd = subset(data, subject == 'Web Development')

# Business Finance
nrow(data_business)
```

```
## [1] 1465
```

```
# Graphic Design
nrow(data_design)
```

```
## [1] 723
```

```
# Musical Instruments
nrow(data_music)
```

```
## [1] 770
```

```
# Web Development
nrow(data_wd)
```

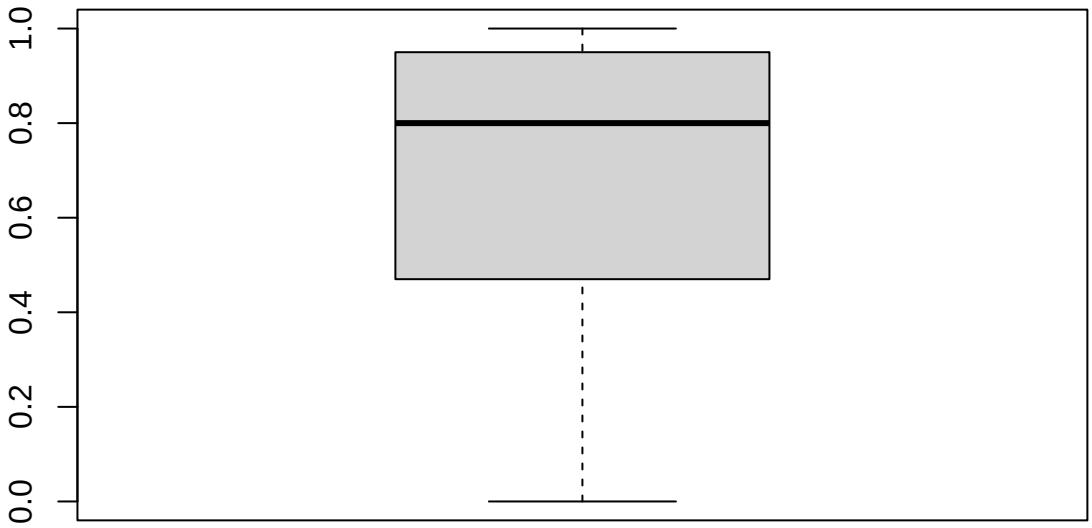
```
## [1] 720
```

What is the average rating per category?

```
# Business Finance
summary(data_business$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.4700  0.8000  0.6904  0.9500  1.0000     274
```

```
boxplot(data_business$Rating)
```

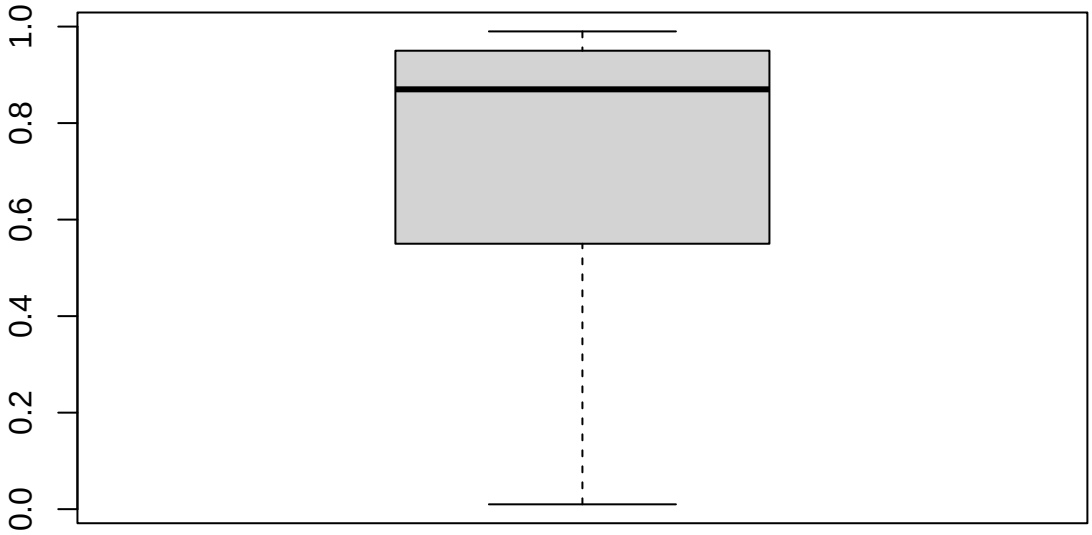



```
# Graphic Design
summary(data_design$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
```

```
##  0.0100  0.5500  0.8700  0.7304  0.9500  0.9900     121
```

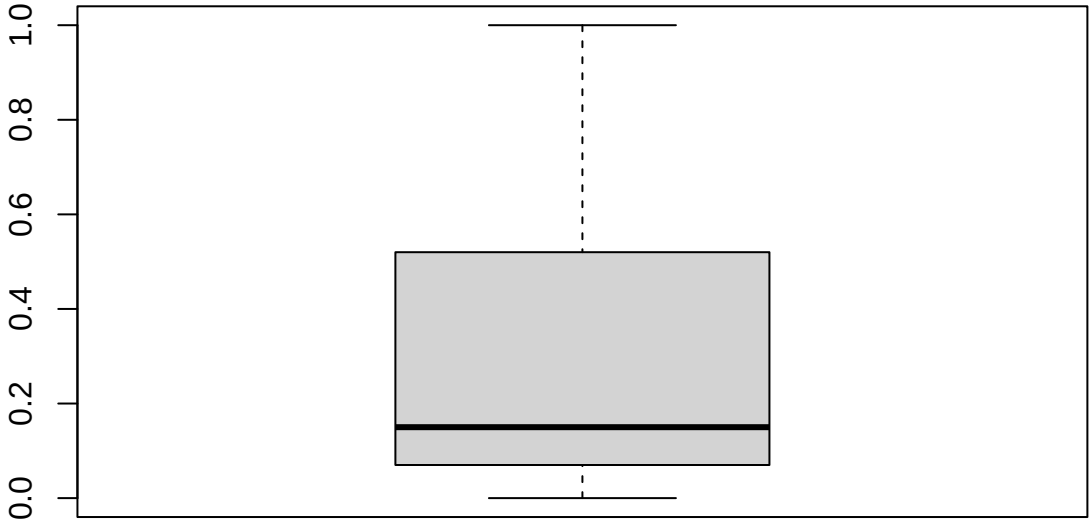
```
boxplot(data_design$Rating)
```




```
# Musical Instruments
summary(data_music$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0700  0.1500  0.3089  0.5200  1.0000      90
```

```
boxplot(data_music$Rating)
```




```
# Web Development
summary(data_wd$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      NA      NA      NA     NaN      NA      NA     720
```

The category with the highest mean is Graphic Design. The category with the lowest average is Musical Instruments. For the courses on Web Development, the data set has all Null Values. Also, none of the box plots show any outliers.

```
# Average rating across all courses
summary(data$Rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.2100  0.7600  0.5953  0.9400  1.0000    2412
```

The average rating across all categories is 59.53%.

How many courses cost money? And what are there prices?

```
# how many are not free
nrow(subset(data, price > 0))
```

```
## [1] 4319
```

```
# how many are free
nrow(subset(data, price == 0))
```

```
## [1] 562
```

```
# what are the prices for those that cost money
hi <- subset(data, price > 0)$price
summary(hi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   25.00   50.00   77.13  100.00  200.00
```

They have an average price of 77. The maximum price is 200 and the minimum is 20.

Let's look at the popularity of the courses in the data set.

```
summary(data$num_subscribers)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     264    1730    4564    4407  268923       5
```

So, there are some with 0 subscribers and some that are very popular with 268,923 subscribers. This shows the vastness of the large data set.

```
# how many have 0 subscribers?
zero_sub <- subset(data, num_subscribers == 0)
nrow(zero_sub) #65
```

```
## [1] 65
```

```
library(lubridate)
```

```
## Loading required package: timechange
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
# When were these courses published?
unique(year(unique(zero_sub$published_timestamp)))
```

```
## [1] 2017 2016 2015 2014
```

```
# What subject are these courses?
unique(zero_sub$subject) # all but web development
```

```
## [1] "Business Finance"    "Graphic Design"      "Musical Instruments"
```