# Homework 1

Your Name: Minguk Kim

Student ID: A20437179

**1. (6 points) Select appropriate data types from the two categories below, for each of the following attributes; note: you must pick up one answer from each category to describe the attributes**

Category1: Qualitative, Quantitative

Category2: Nominal, Ordinal, Symmetric Binary, Asymmetric Binary, Discrete, Continuous

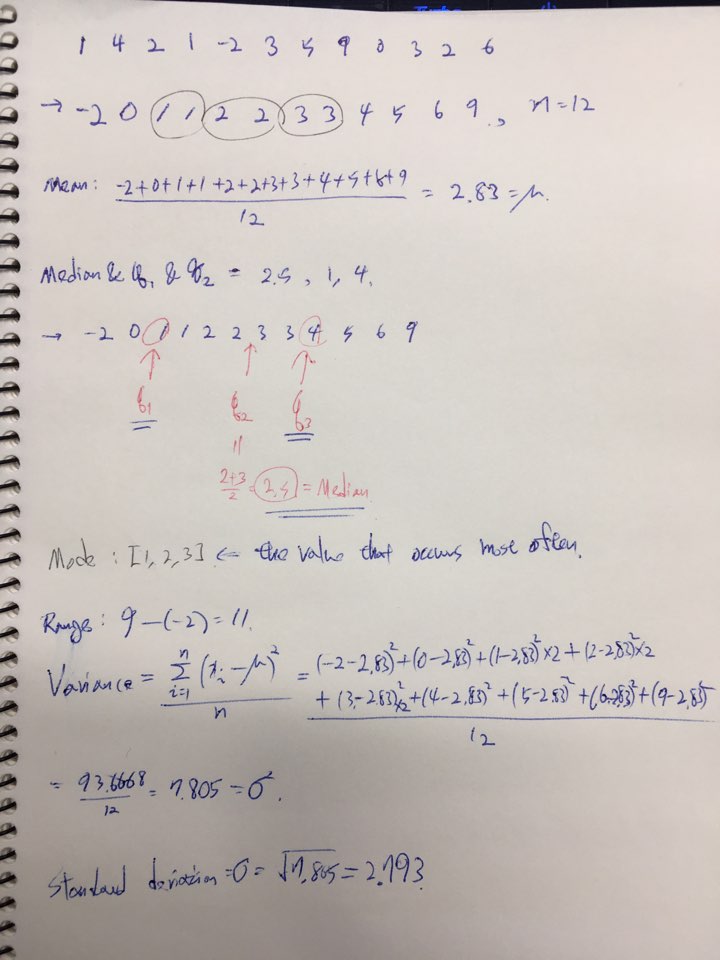
**Attributes:**

* Gender (M/F) : Qualitative - Symmetric Binary
* Degree of hungriness (light, medium, very hungry) : Qualitative – Nominal
* Department names : Qualitative - Nominal
* Grade (passed or failed) : Qualitative – Asymmetric Binary
* Nationality : Qualitative – Nominal
* The number of valid values : Quantitative – Discrete

**2. (10 points) Provide descriptive statistics on the following numbers: (manual calculations)**

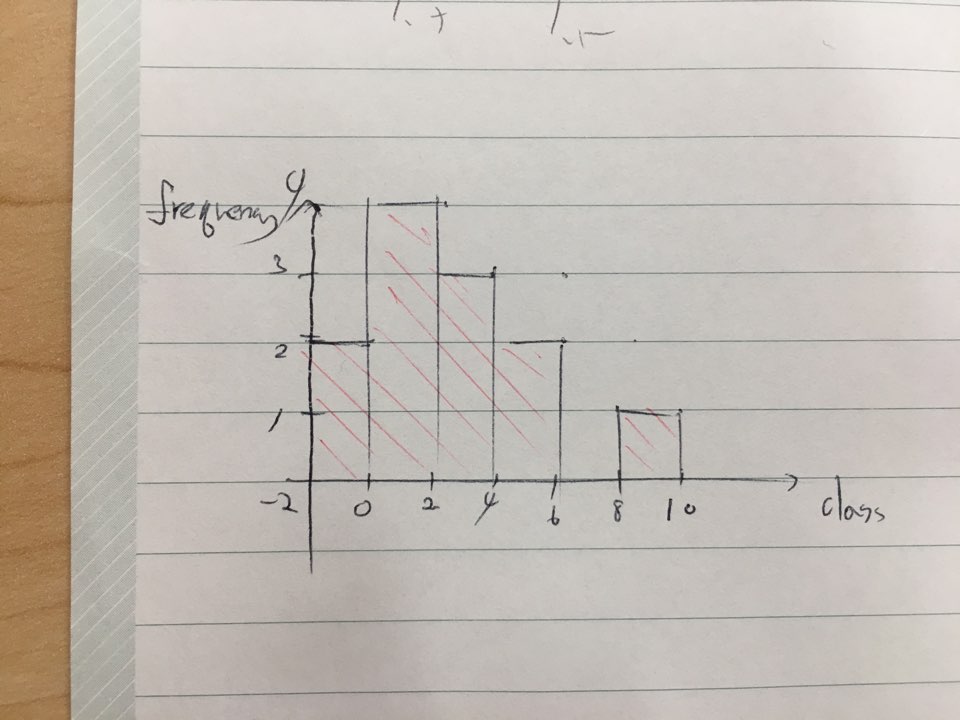
1, 4, 2, 1, -2, 3, 5, 9, 0, 3, 2, 6

* Mean : 2.83
* Median : 2.5
* Mode : [1,2,3]
* Range : 11
* Variance : 7.805
* Standard deviation : 2.793
* 1st Quartile : 1
* 3rd Quartile : 4

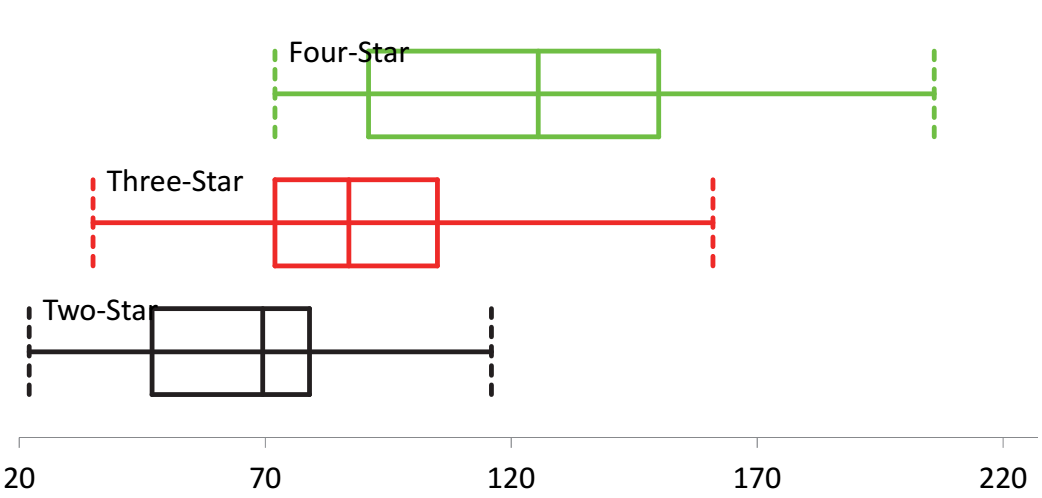


**Is this a symmetric/normal distribution? If not, is it positive or negative skewed?**

This is not a symmetric/normal distribution. It is positive skewed.



**3. (24 points) The box plot below describes and compares the average price of two-star, three-star and four-star hotel rooms in the city of Chicago. The values on the axis present the room price in US dollars**



Answer the following questions by your observations. Note that you do not need to do any calculations

1). Try to interpret the skewness of these three boxplots, and why? [9 points]

The criteria that can distinguish skewness of boxplot is where the median is skewed between q1 and q2.

So Four-Star is negative skew. Three-Star is positive skew. Two-Star is negative skew.

2). Try to compare the prices among these two-star, three-star and four-star hotels by following the instructions below [15 points]. Note: you do not need calculations, just infer the answers from the observations

2.1). compare the median values

Four-Star > Three-Star > Two-Star

2.2). compare the variance

Four-Star > Three-Star > Two-Star

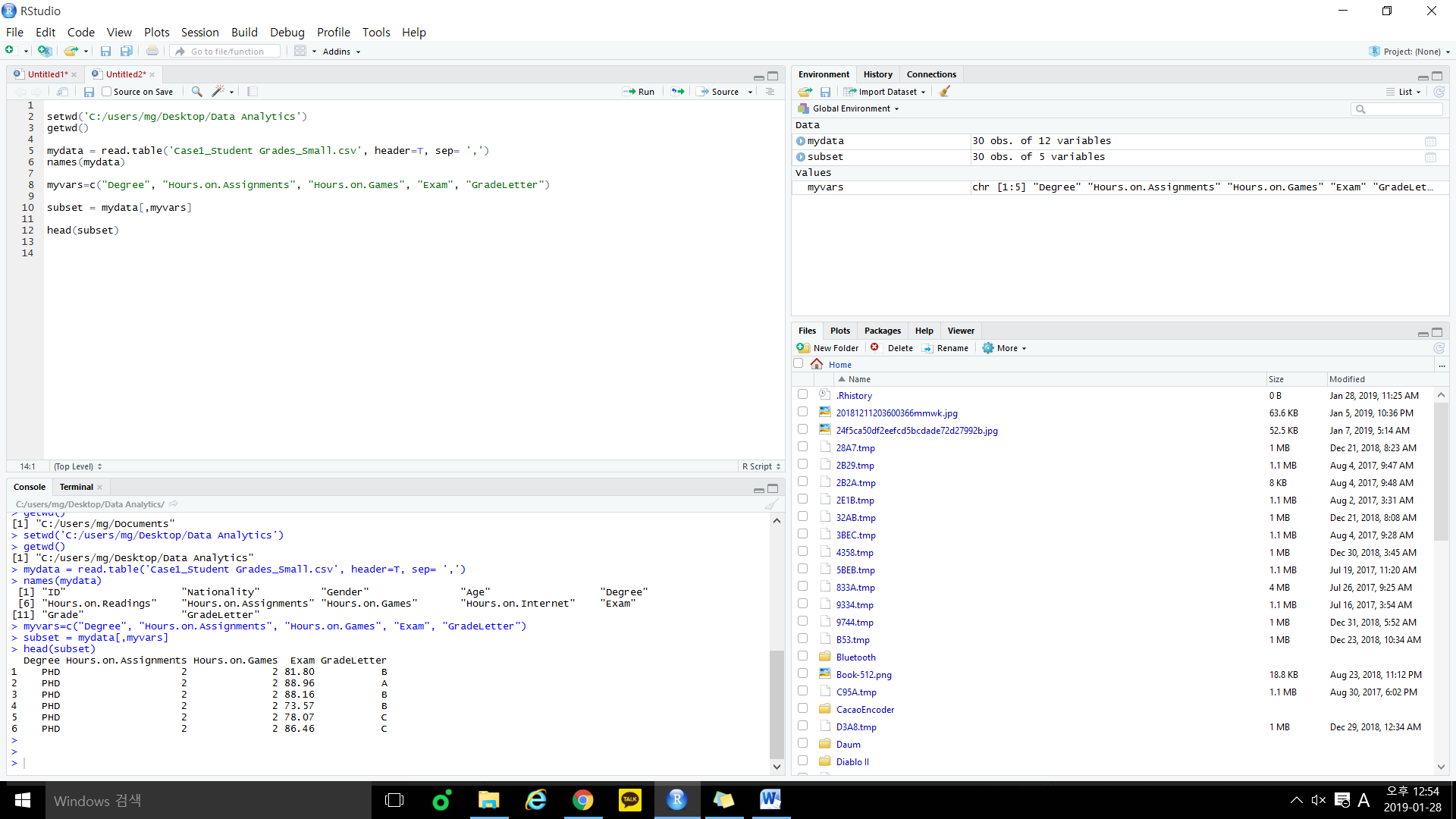
2.3). Do you think it is reliable to compare the hotel process based on the median value only? Why?

We can see that even few four-star hotels price is lower than three or two star hotel. Like this, although four-star hotel’s median value is higher than others, the higher the variance, the more prices are formed in various range. Therefore it is not right that compare the hotel process based on median value.

4. (60 points) Use R to produce the outputs by following the instructions below. Note that, for each question, you need to provide the codes and the snapshot of the outputs. Or just the snapshots with both codes and outputs – just like the examples given in the class/slides

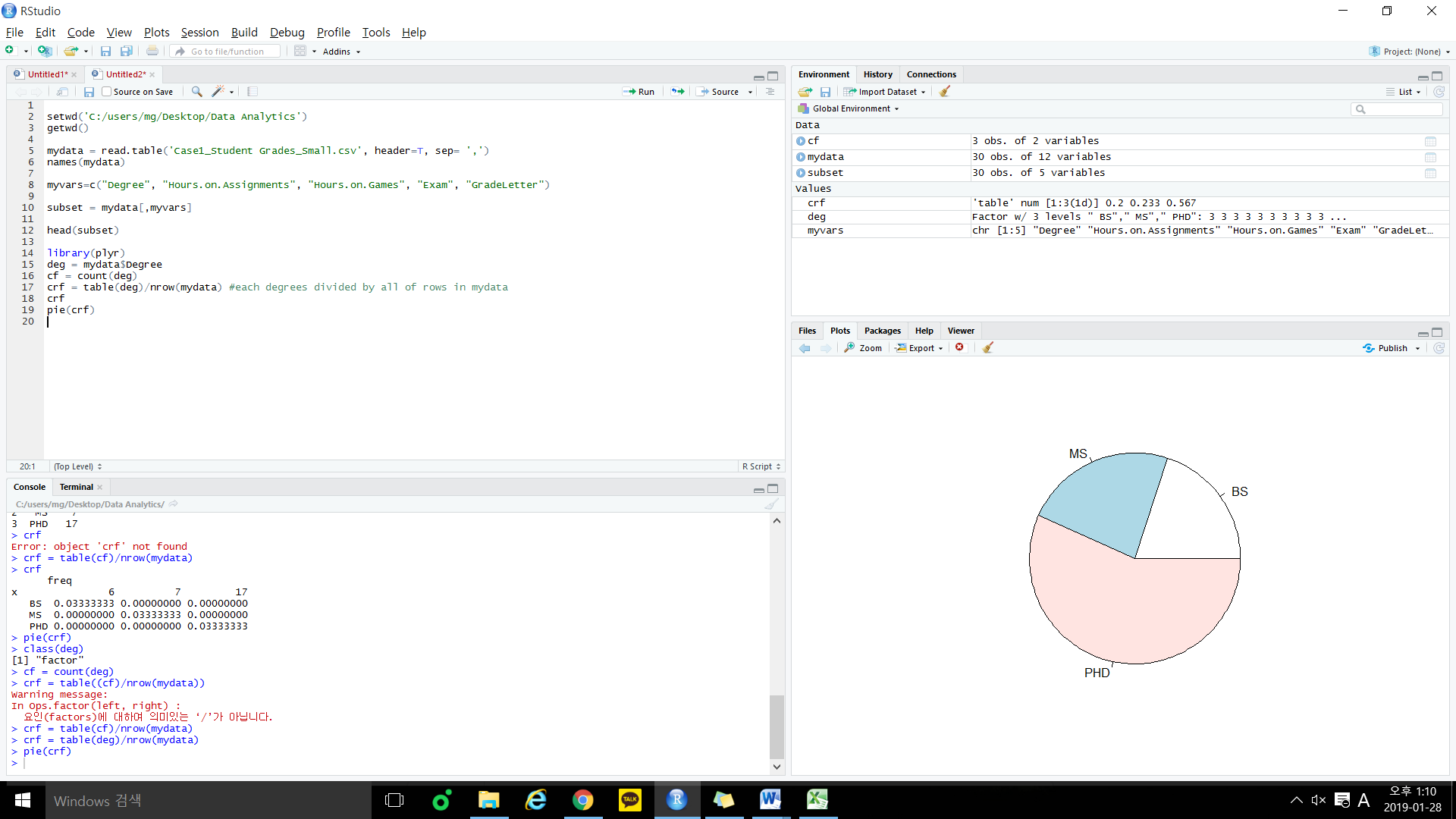
Let’s use the Case1\_Student Grades\_Small.csv data set which was uploaded to the blackboard system in the section of Case Studies

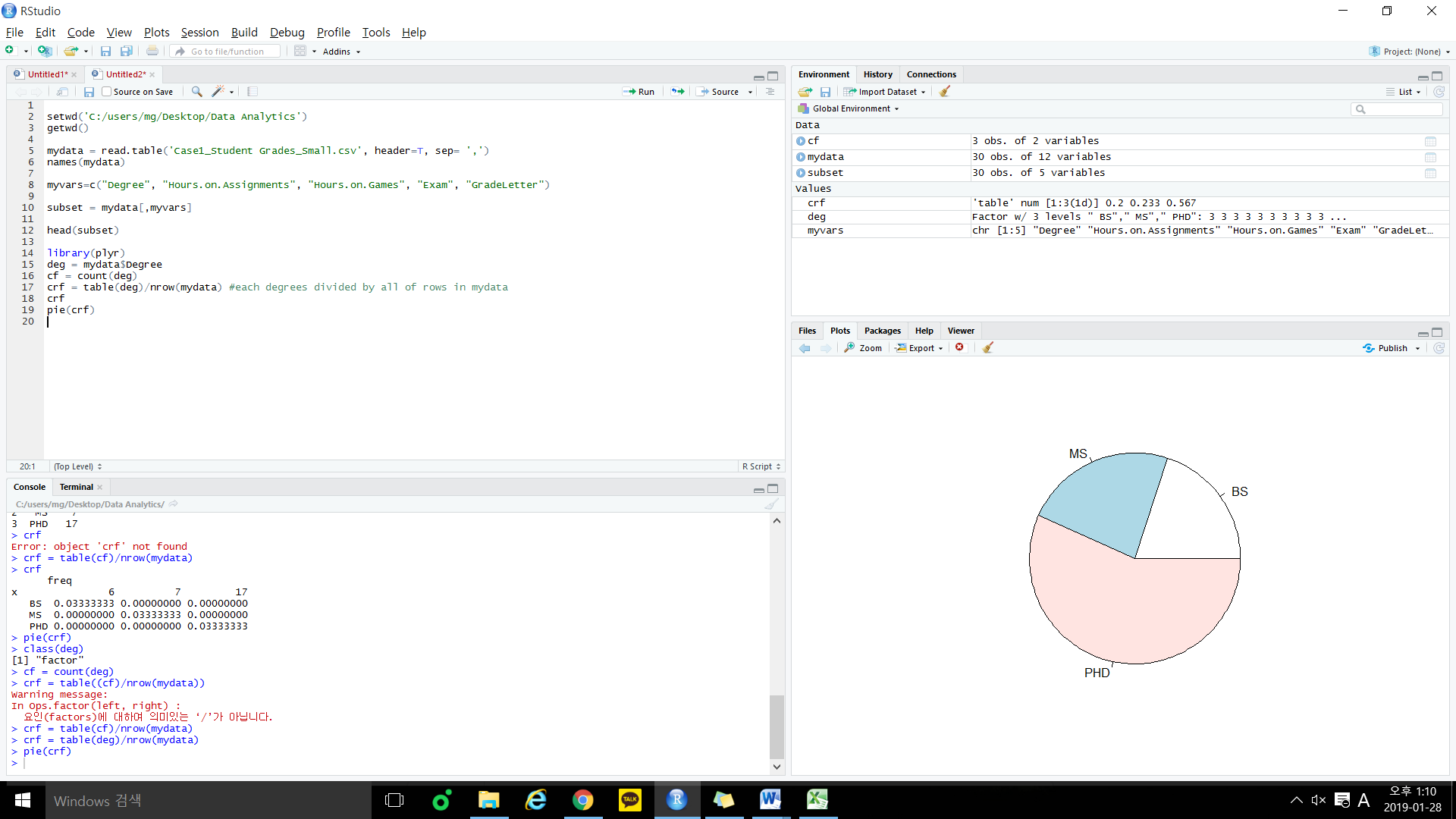
1). Load data into R, create a subset by using the columns "Degree", "Hours on Assignment", "Hours on Game", "Exam", and "GradeLetter". After that, output the first few of the lines (i.e., top lines) in this subset. [10]



Next, we analyze this subset.

2). Produce class relative frequency of the column "Degree", and draw a pie chart. What conclusions can you draw from the pie chart? [10]



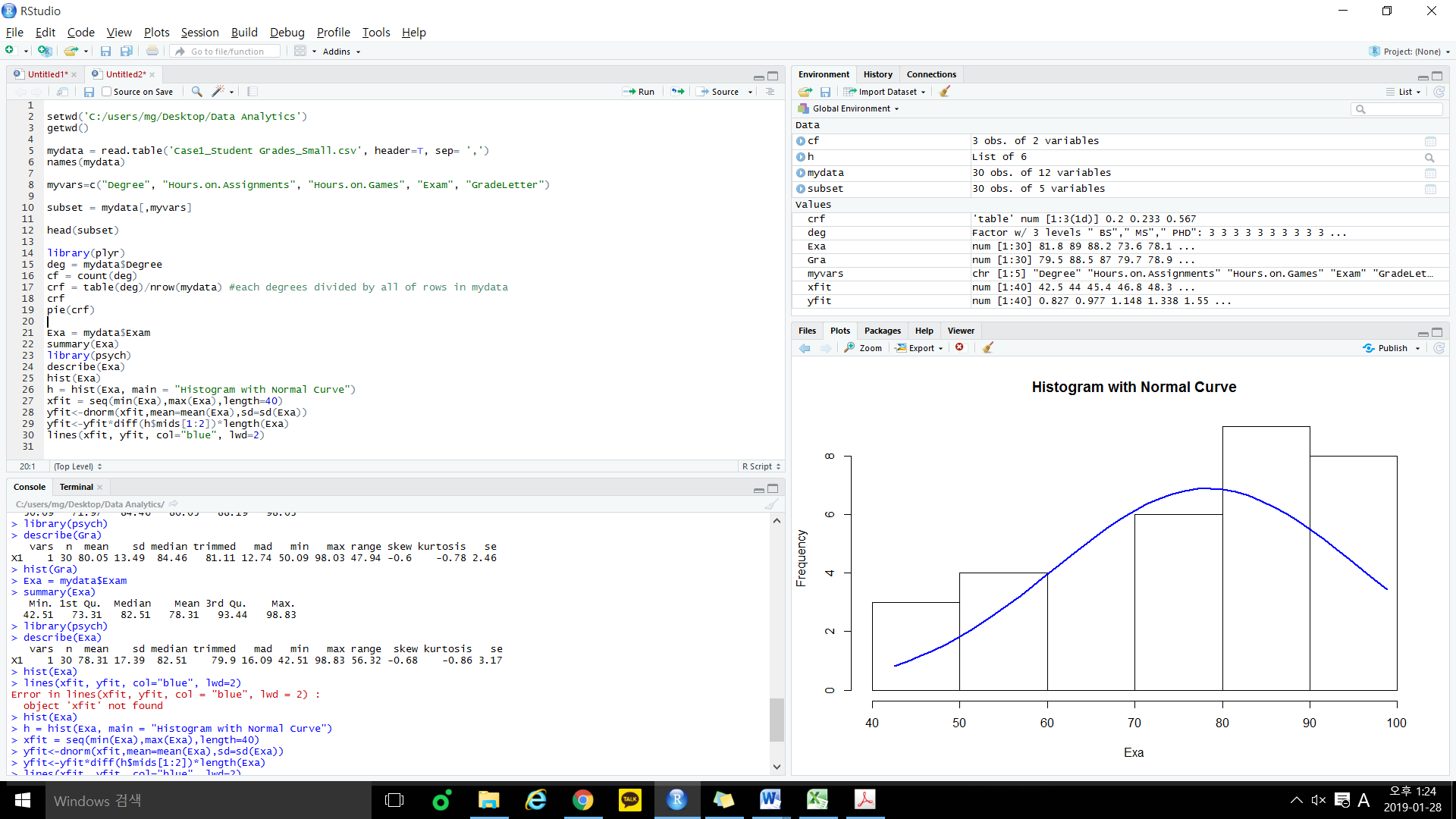


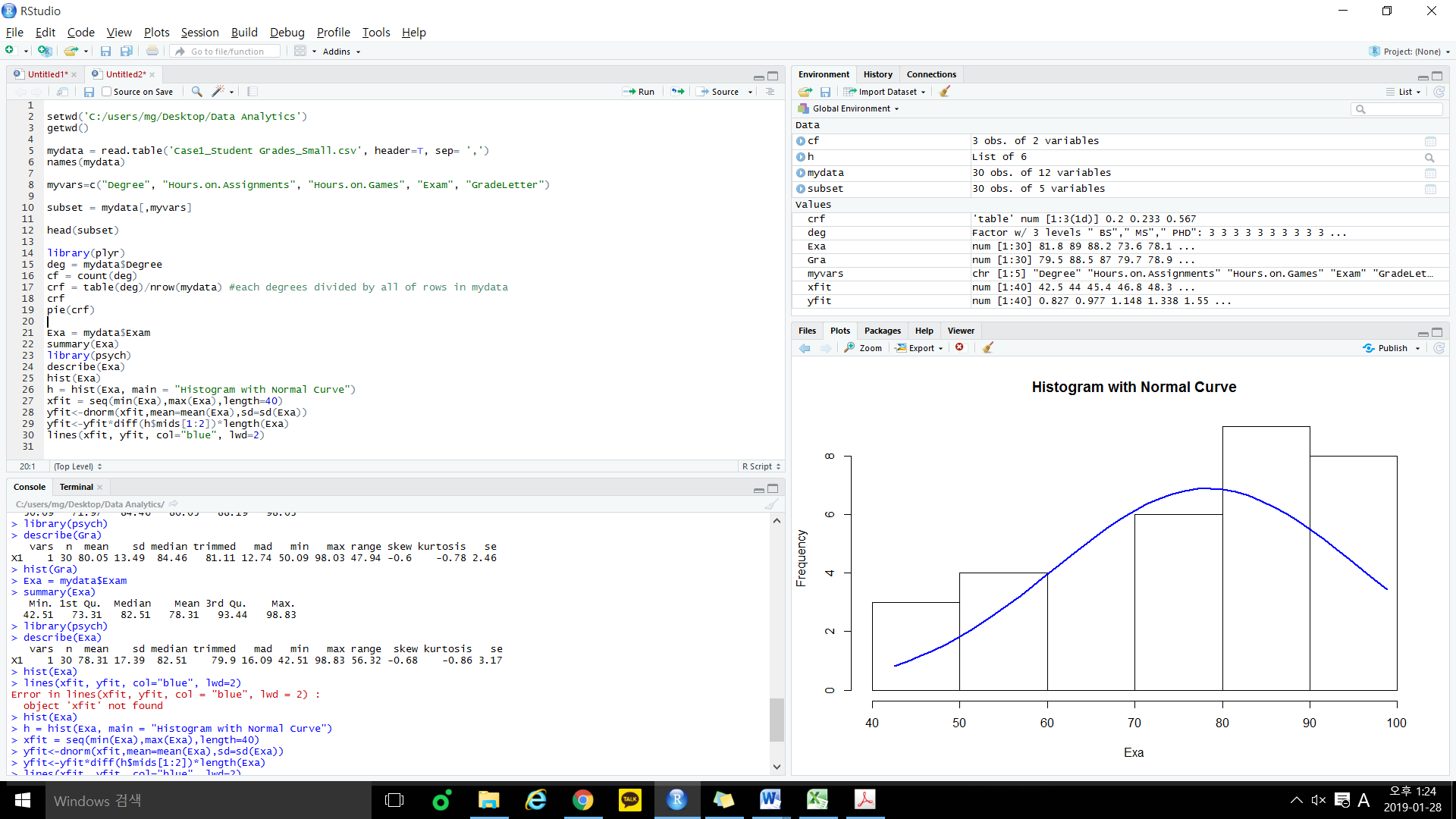
**What conclusions can you draw from the pie chart?**

Basically, This is pie chart about class relative frequency.

Here, we can see 3 classes(BS, MS, PHD). Each class point to their ratio from whole student numbers. Most of student are included in PHD, and the rest of the students have similar proportions.

3). Produce the histogram with a curve above the bars by using the values in the column "Exam". Interpret the plots and give your conclusions [20] Produce the descriptive statistics (q1, q2, q3, mean, median, variance, etc) of the variable "Exam". And answer the question "do you think 90 is a good grade in this class? Why?"



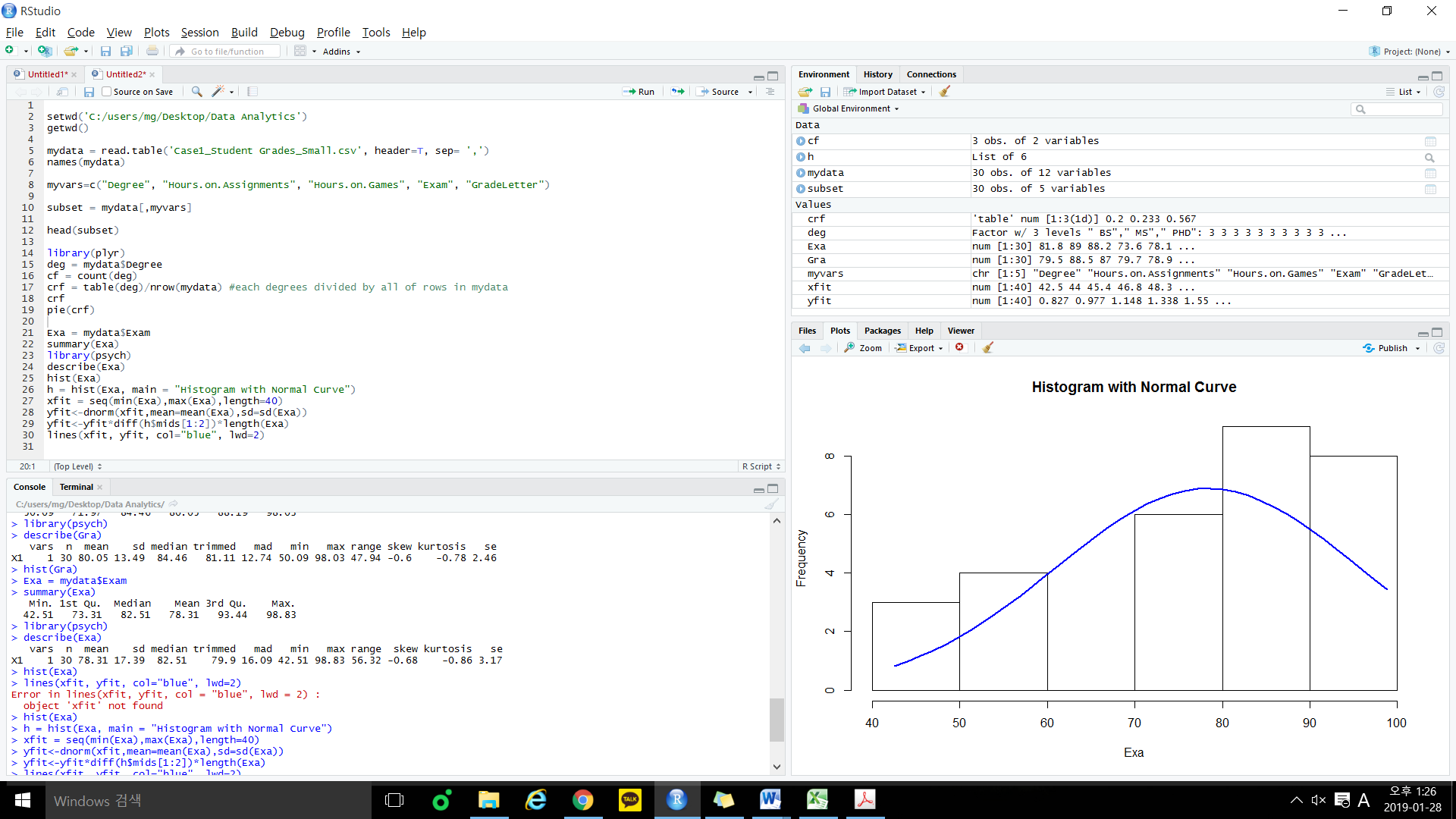


**Conclusions**

We can interpret this plot by using skewness and outlier.

Definition of outlier is that more(or less) than 3/2 times of upper(or lower) quartile so in this graph, there is no outliers. If you look at the overall graph, you can see that data is skewed to the left. So it points to negative skew and more than 50% of students are located in score 80-100. And only 7 students are located in score 40-60.

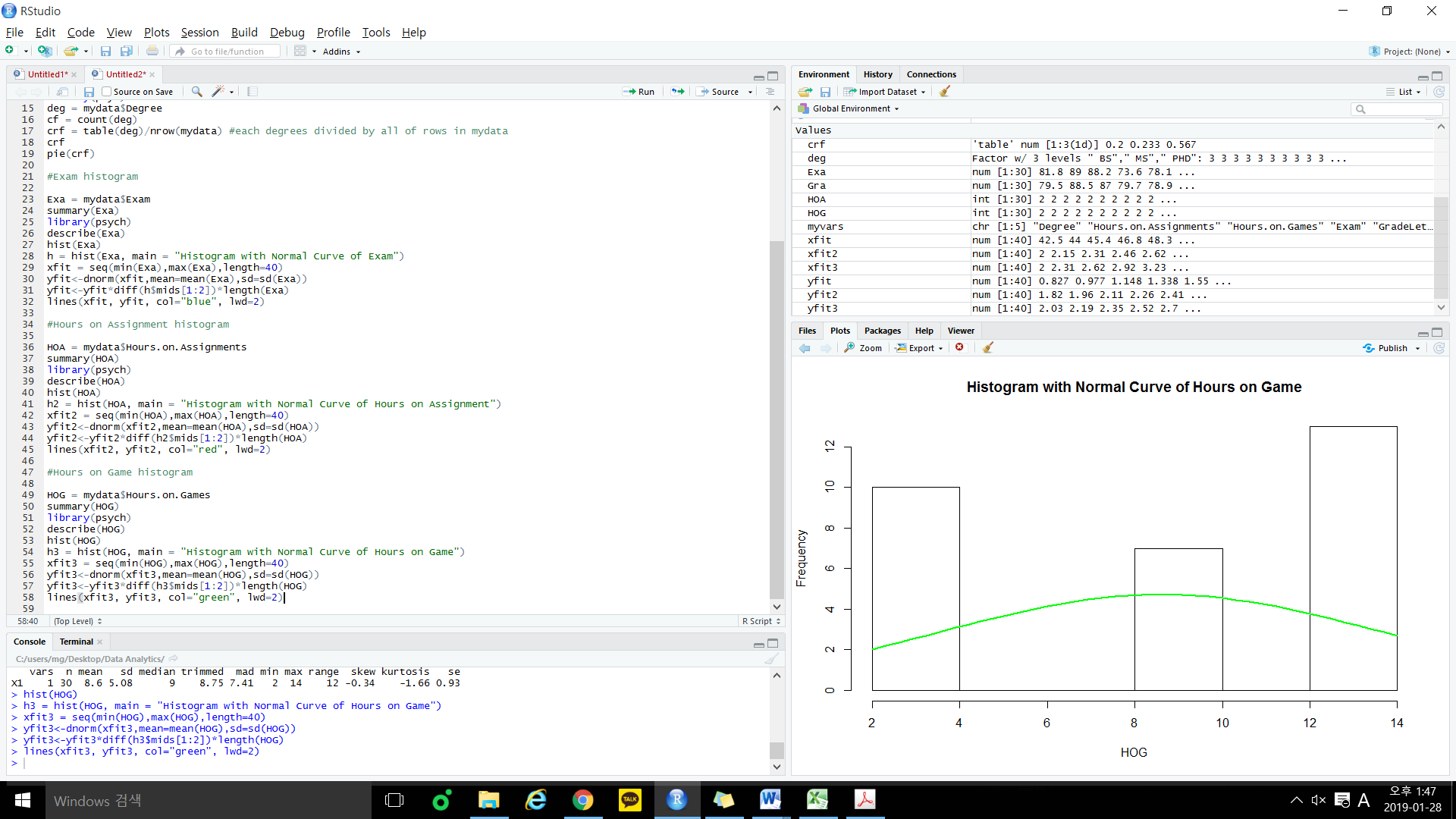
So we know that upper 50% of students received high score than average.

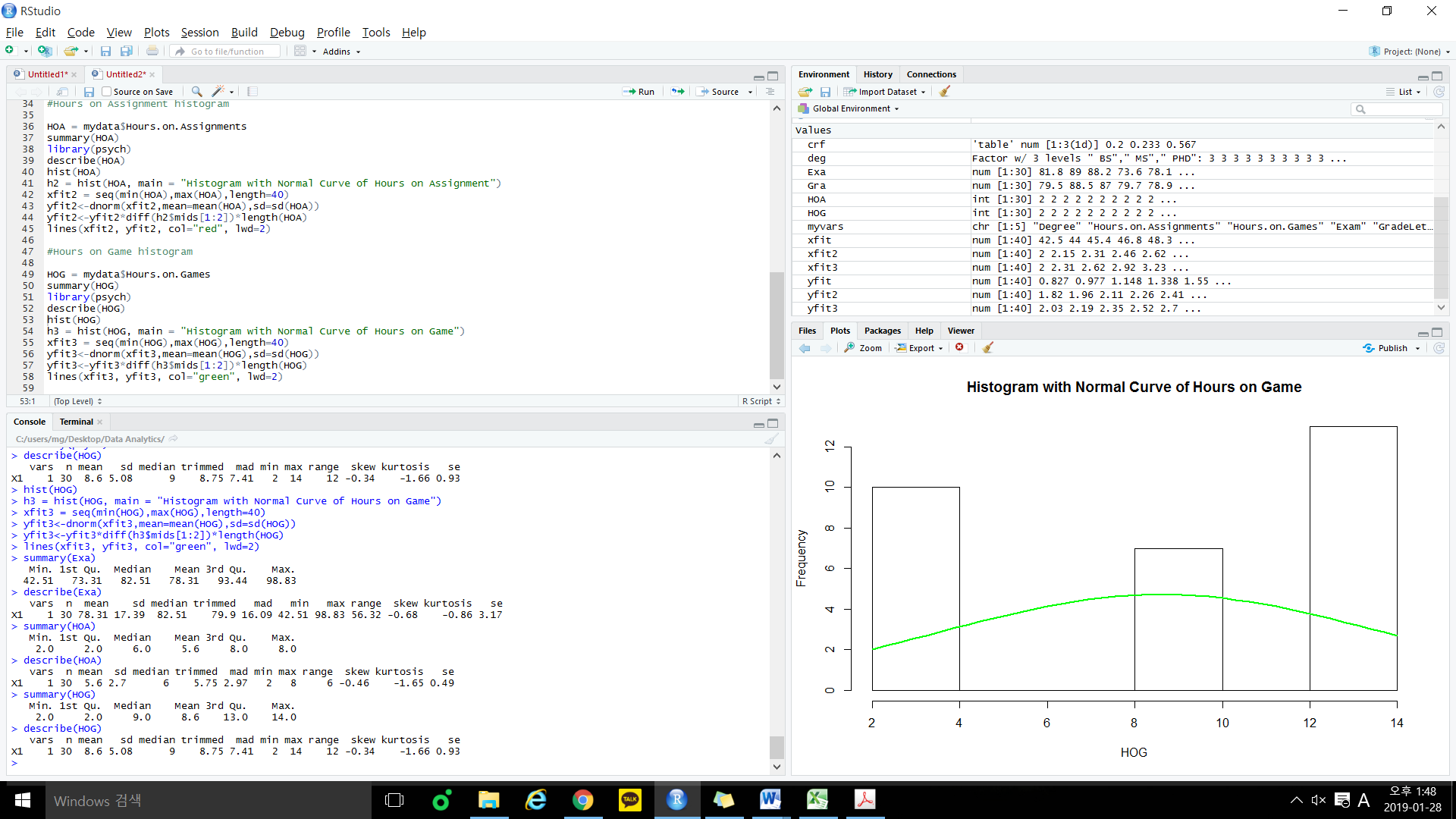
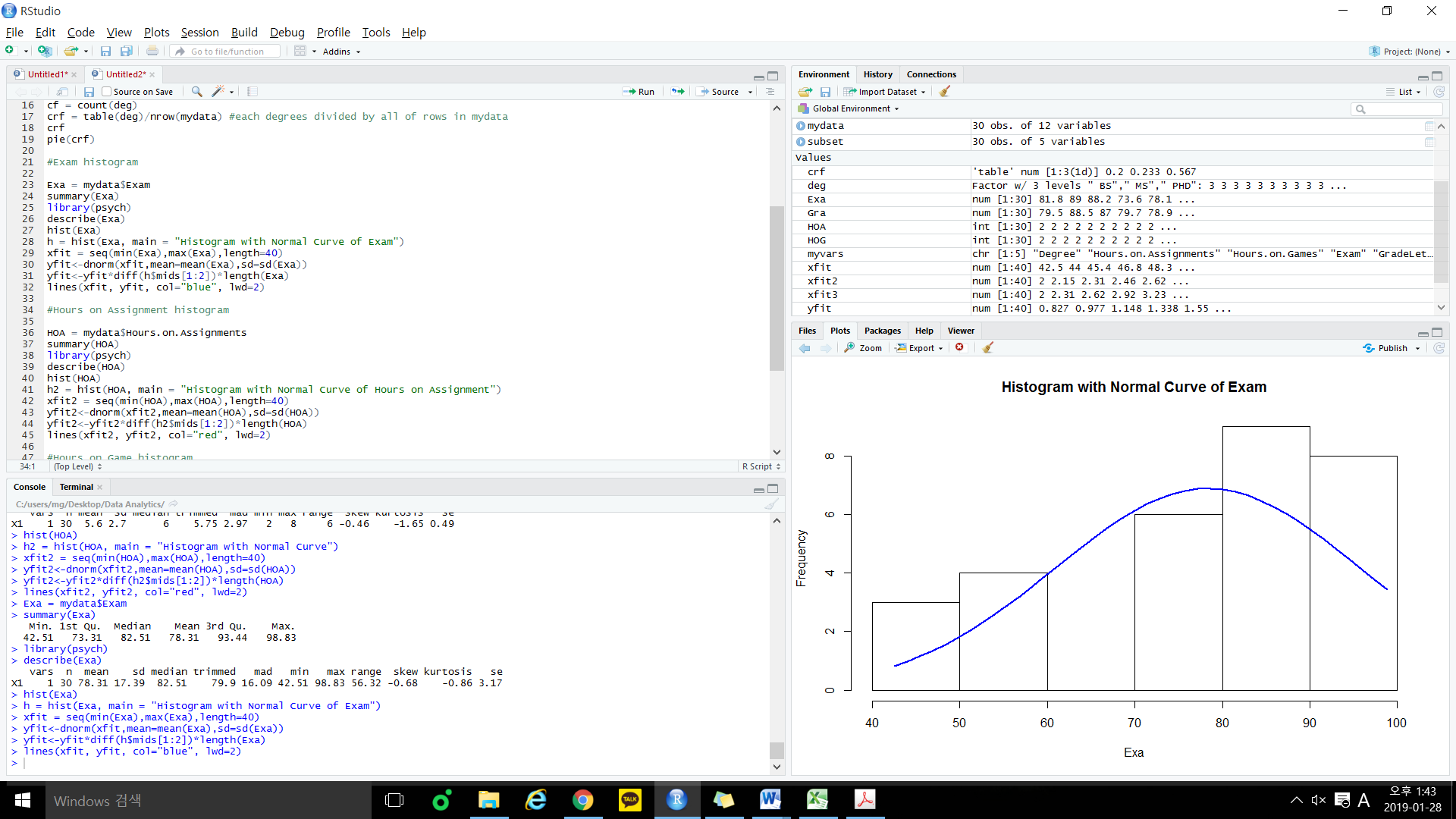
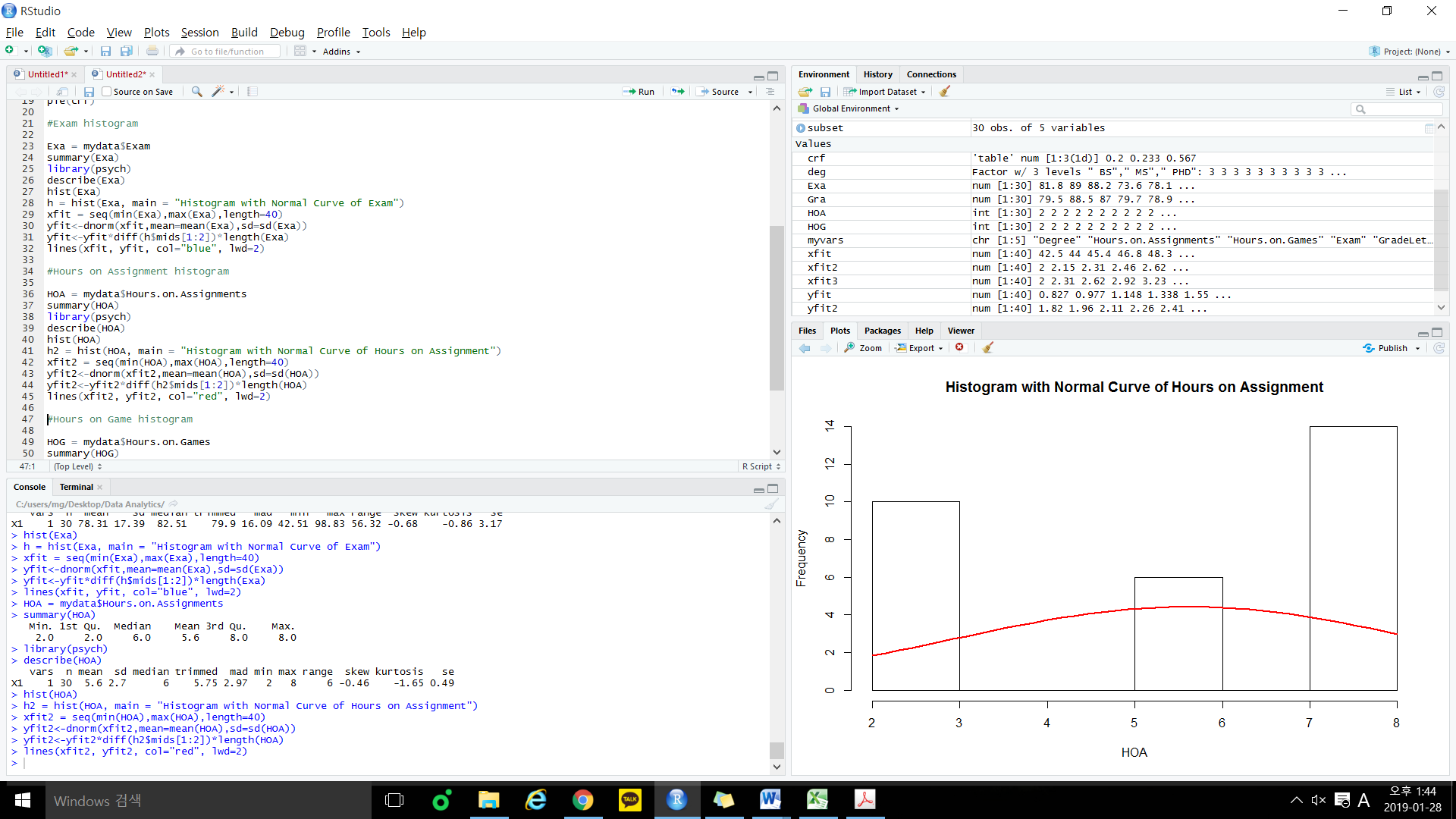
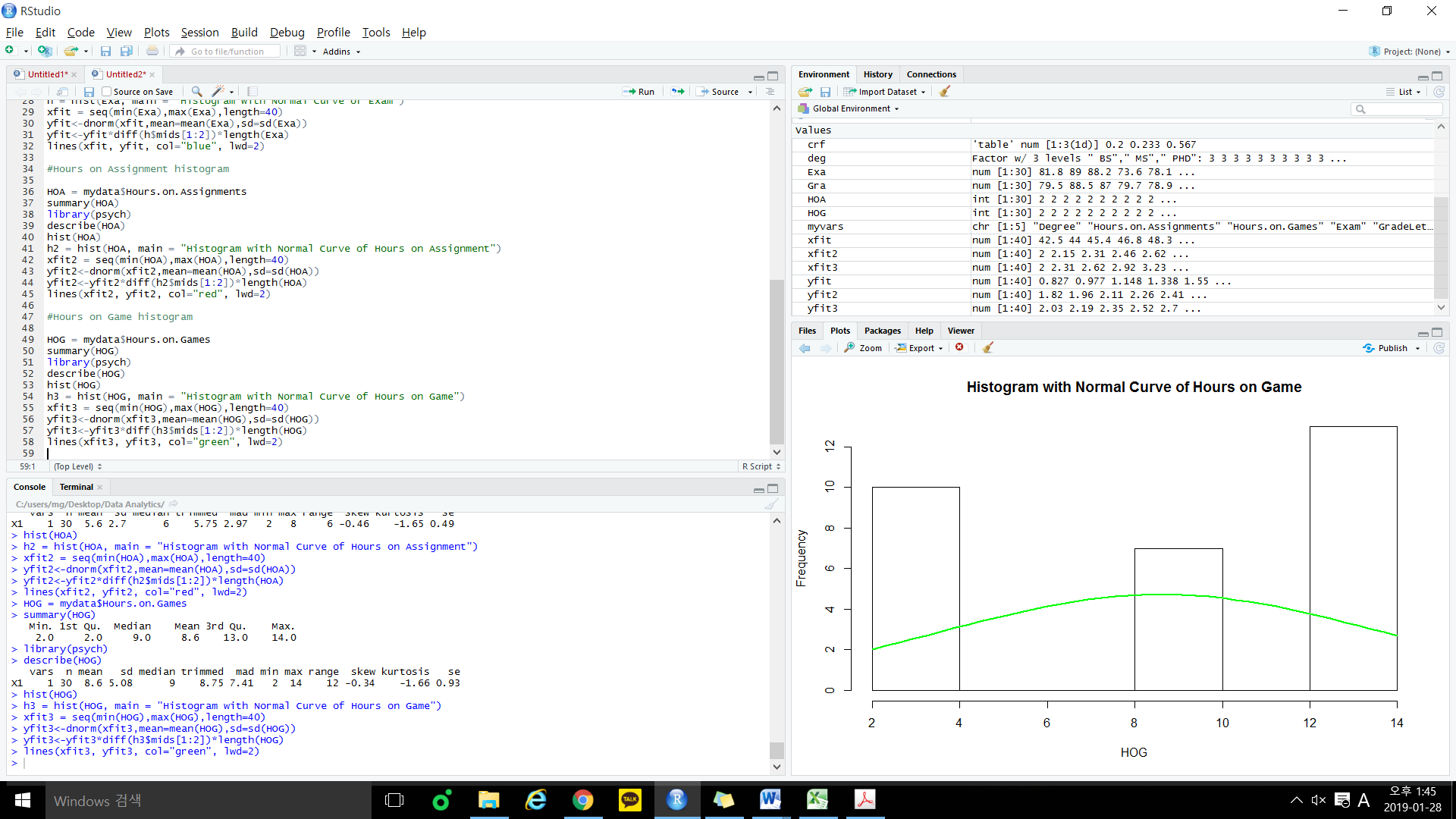


**"Do you think 90 is a good grade in this class? Why?"**

Even though the skewness is negative skew so more than half of students received grades upper than average grade, score 90 surpassed mean value and median value. Also it is located in the top 27%, and if someone scored 90, that person can get receive ‘A’. So i think 90 is quite good grade in the class.

4). Repeat the process in 3) but for the variables "Hours on Assignment" and "Hours on Game". Interpret these plots, and compare them with the plot in 3), and give your findings or patterns related to the observations. [20]





**findings or patterns related to the observations**

Hours on Game and Hour on Assignment show the similar shape, have a high value on both sides and all these three graphs show negative skew. as well, they all have a minus kurtosis.