**Preliminary Analysis**

# Overview of our dataset

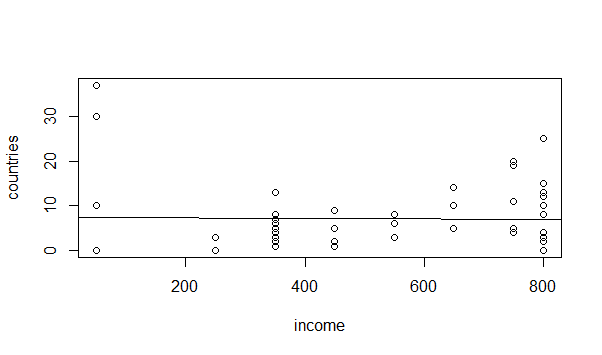
Our dataset contains information about our class. The following list is the data that was collected:

* Gender
* International or Korean student
* How many countries visited
* How many times watched a movie during a month
* Income per month (in USD)
* How they get their income (family, work, both)
* Categories they spend their income (in percentage)

# Observations

## Observation 1: Income vs countries visited

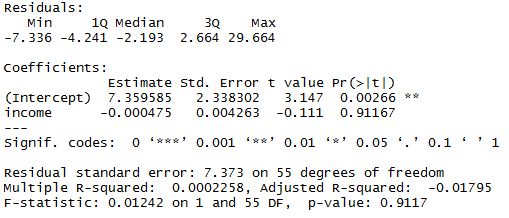
We wanted to find out if there is a relationship between the income of a student and the countries they already visited. All students were considered in this plot.



The plot shows us no clear trend. Because of this we can assume that a Student is not effected by higher income to go to a foreign country.

### Fact of observation 1

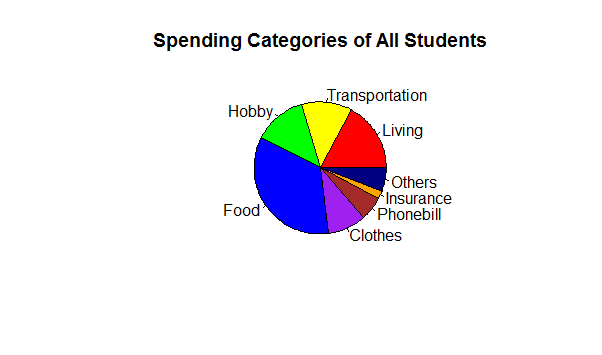
On the following picture you can see the fact of this observation:



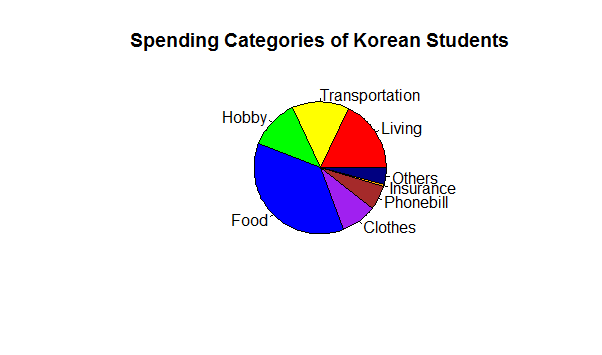
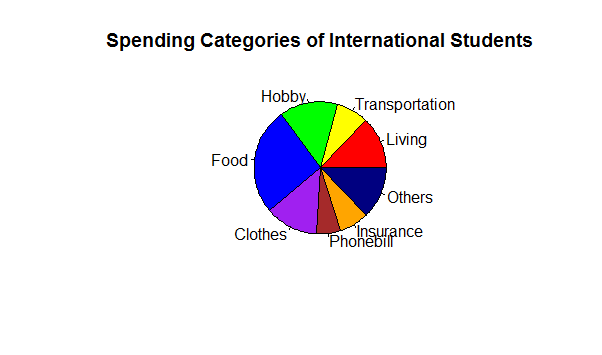
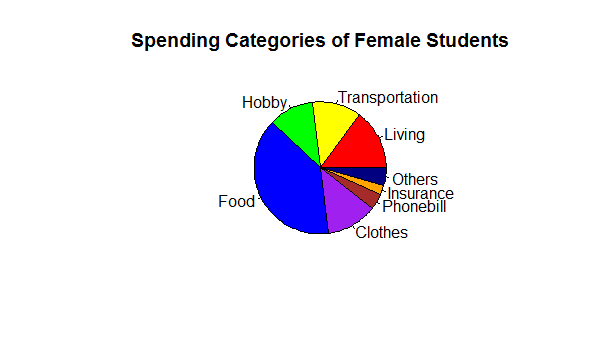
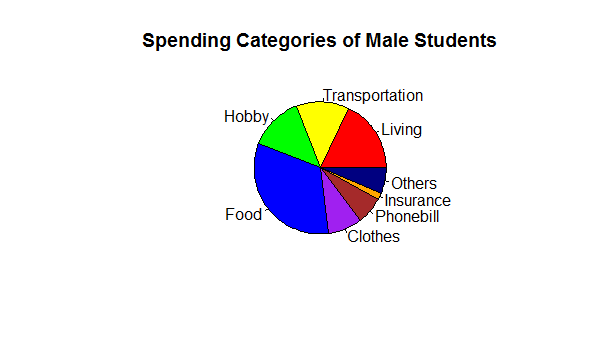
We can read that the income and countries have no linear relationship and the coefficient of the intercept value is highly significant. Therefore, the number of visited countries do not depend on the amount of income of a student. The R-squared is about 0.0002 which mean that only 0.022 % of the visited countries was effected by the amount of income.

## Observation 2: Spending categories of all students

In this observation we compared the categories. On the following pie you can see the overview of all students and the mean distribution of their income spending in the different categories.



In Addition, we separated the students based on gender and nationality:

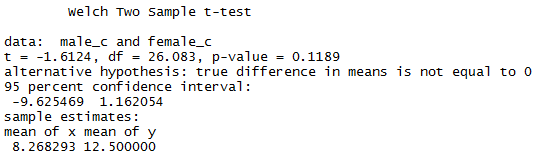


### Hypothesis test of observation 2

In Order to find out a relationship between male and female students, we set up a hypothesis test of percentage of spending for clothes between males and females:

* H0: The percentages are the same in males and females
* H1: The percentages are not the same

The t-test show us following facts:

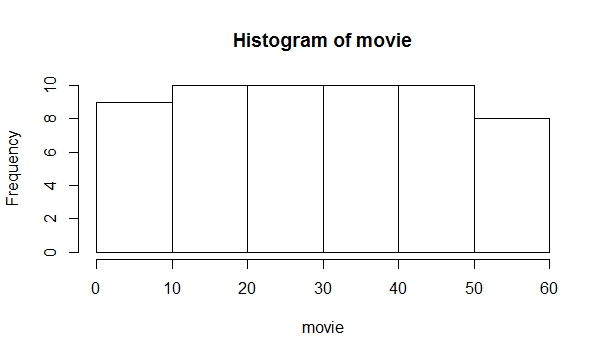


The p-value for this test is 0.1189, so we can't reject null hypothesis. Therefore, the percentage of spending for clothes between males and females are the same. We can not assume a high difference between male and female in this category.

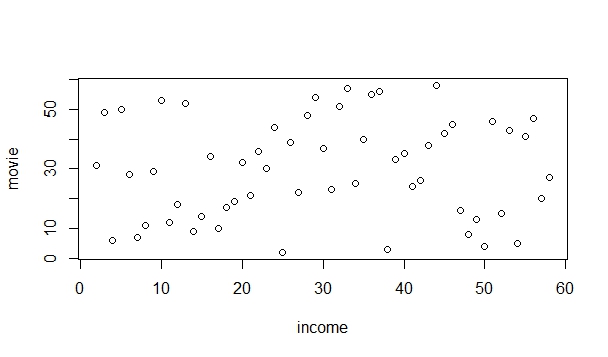
# Observation 3: income vs movie

In the third observation we compared the movie data with the income data to spot any noticeable relationship. The following two histogram shows the distribution of this two subjects.





When we look at the two histogram, we can assume that there is no correlation between income and movies. As a consequence of this, the next plot graph shows, that there's no particular correlation between income and frequency of watching movie.



# List of the next tasks

The next step is to analyse the other categories to find any anomalies that could be interesting. In particular we want to compare the gender and nationality with the categories who are left. In this case we want to be able to answer following questions:

* Are there any salient relationships between male and female students?
* Are there any salient relationships between korean and foreign students?
* Have the income of a student any influence of the activity of a student (e.g. going to cinema)?
* Does any similarities or differences exist between Korean and foreign male students?
* Does any similarities or differences exist between Korean and foreign female students?

# Project Members

**Gayun Lee Javier Rodriguez**



**Taewoo Kim Philipp Merrouche**

** **

**Sung Mo Koo**

