Analysis of annual incomes and spending capacity depending on gender

Anastasiia Vynychenko

22 April 2019

Table of Contents

### Intro

The main purpose of this analysis is to understand if there is some connections between “spending money capacity” and gender. For resolving this issue, data was taken from resource: <https://www.kaggle.com/> [[1]](#footnote-22)

Here you can see how the header of data looks like:

## CustomerID Gender Age annual\_income spending\_score  
## 1 1 Male 19 15 39  
## 2 2 Male 21 15 81  
## 3 3 Female 20 16 6  
## 4 4 Female 23 16 77  
## 5 5 Female 31 17 40  
## 6 6 Female 22 17 76

### Analyzing part

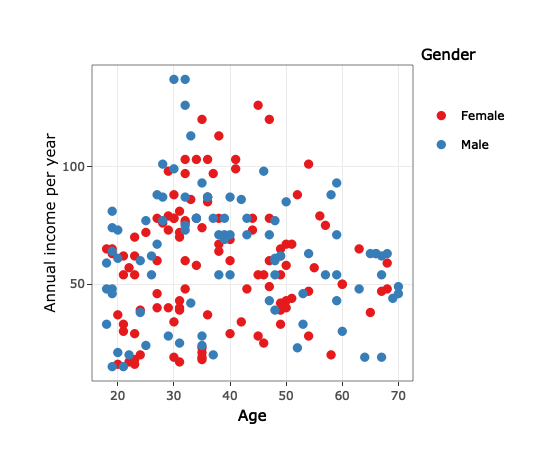
#### Correlation between gender and annual income

Let’s take a look to our data set. First of all, we want to understand quantity of Female & Male in our data and how age in each group is distributed.

Here is summary of data:

## # A tibble: 2 x 6  
## Gender quantity max\_age mean\_age min\_age mean\_income  
## <chr> <int> <int> <dbl> <int> <dbl>  
## 1 Female 112 68 38.1 18 59.2  
## 2 Male 88 70 39.8 18 62.2

From this tibble we can see that all indicators are almost equal. We also can see that mean age in male and female groups also equal, the same situation with income, etc. Only a small difference in quantity of participants, but it’s no so huge. That’s why we can carry on with our analysis. Despite the fact that we saw the summary table, it’s better to visualize our data.



From this graph we can see that in our data there is no correlation between income and age that people gained per year. Also there is no difference in both groupes: male and female.

Finally let’s try to prove this hypotesis with **ANOVA analysis**.

## Df Sum Sq Mean Sq F value Pr(>F)  
## Gender 1 437 436.8 0.632 0.428  
## Residuals 198 136840 691.1

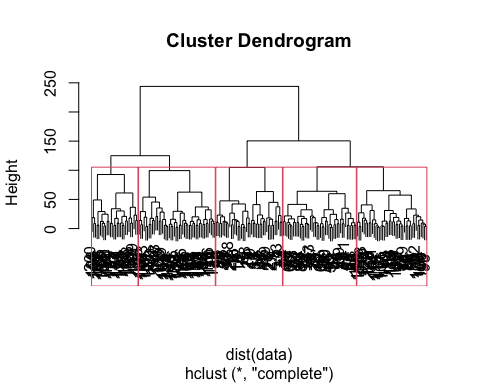
We received P-value = 0.428 that tell us that we can accept this hypotesis, telling us that people in two groups (Female and Male) don’t have differences in their incomes.

#### Clusterizing

The next step, which help us to understand our data better is to split customers to different groups, which will be different, in comparison with Gender characteristics.

First of all we should choose what amounts of clusters we want to pick out. Let’s do this with hierarchical clusterization.

Consider this graph:



Analyzing this graph, it will be optimally to choose 5 clusters. Let’s assign all customers to specific cluster and we will do this, using K-Means Model. Before clustering, I will normalize data, using Z-score method.

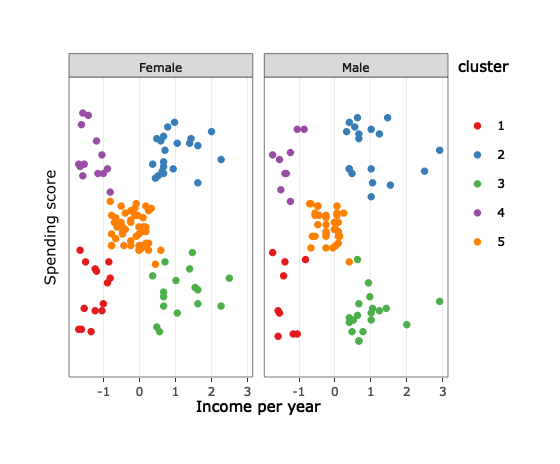
## CustomerID Gender Age annual\_income spending\_score cluster  
## 1 1 Male -1.4210029 -1.734646 -0.4337131 1  
## 2 2 Male -1.2778288 -1.734646 1.1927111 4  
## 3 3 Female -1.3494159 -1.696572 -1.7116178 1  
## 4 4 Female -1.1346547 -1.696572 1.0378135 4  
## 5 5 Female -0.5619583 -1.658498 -0.3949887 1  
## 6 6 Female -1.2062418 -1.658498 0.9990891 4

Here we can see the beggining of table with normalized numeric variables and assigned number of cluster.

#### Analyzing of spending money

Well, now we can look at graph and see what behaviour of “spending money capacity” for all 5 subgroups of customers.

## Warning: `group\_by\_()` is deprecated as of dplyr 0.7.0.  
## Please use `group\_by()` instead.  
## See vignette('programming') for more help  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.



Here we can see no difference in behaviour of spending money eather it female or male. We only can see difference depending on cluster, which we assigned to each customer.

### Conclusion

So, in conclusion we can claim:

1. There is no difference in “spending money capacity” between genders.
2. There is an existing difference in “spending money capacity”, depending on subgroups (clusters), which we can assign based on annual incomes and spending score.

1. [Direct link for downloading file](https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python) [↑](#footnote-ref-22)