Analysis of annual incomes and spending capacity depending on gender

Anastasiia Vynychenko

21 February 2019

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]

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

```
##   CustomerID Gender Age annual_income spending_score
## 1          1   Male  19         15000             39
## 2          2   Male  21         15000             81
## 3          3 Female  20         16000              6
## 4          4 Female  23         16000             77
## 5          5 Female  31         17000             40
## 6          6 Female  22         17000             76
```

Analyzing part

Correlation between gender and annual income

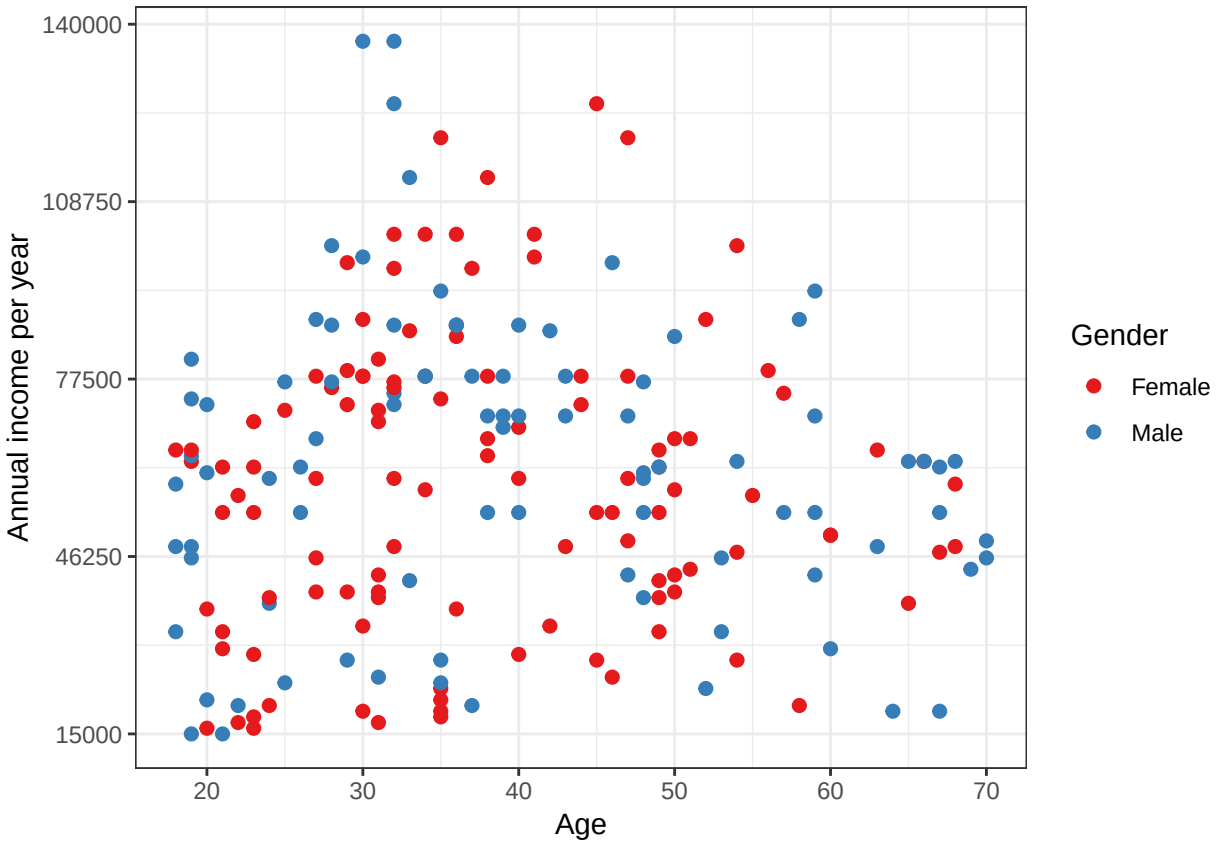
Let's take a look to our data set. First of all, we want to understand quantity of Female & Male in our data and how age in each subgroup is distributed.

Here is summary of data:

```
## # A tibble: 2 x 6
##   Gender quantity max_age mean_age min_age mean_income
##   <fct>     <int>   <dbl>    <dbl>   <dbl>       <dbl>
## 1 Female      112      68     38.1      18      59250
## 2 Male         88      70     39.8      18      62227.
```

From this tibble we can see that all indicators 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.

[1] Direct link for downloading file



From this graph we can see that in our data we almost don't have difference between gender and income that people gain per year.

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

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Gender        1 4.368e+08 436825455   0.632  0.428
## Residuals   198 1.368e+11 691113407
```

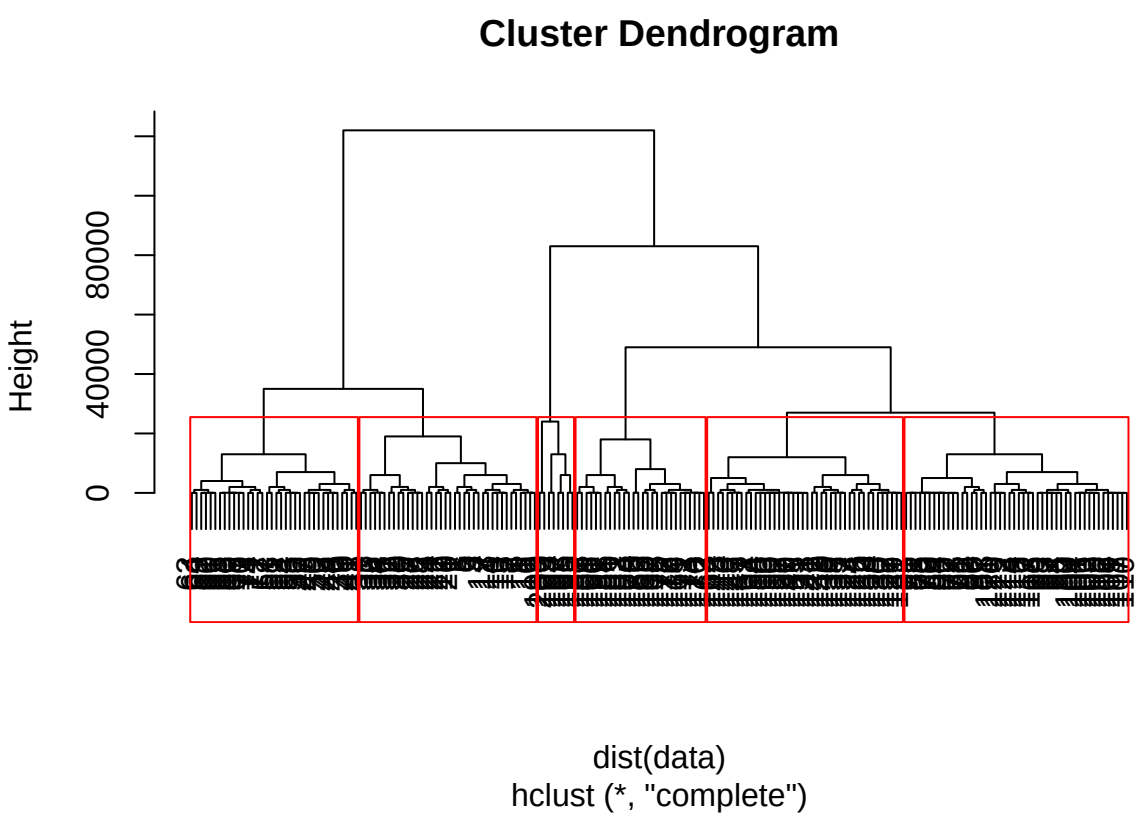
We received P-value = 0.428 that tell us that we can accept this hypotesis, telling us that people in two groups (Female and Male) don't have differences in their incomes.

Clusterizing

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

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

Consider this graph:



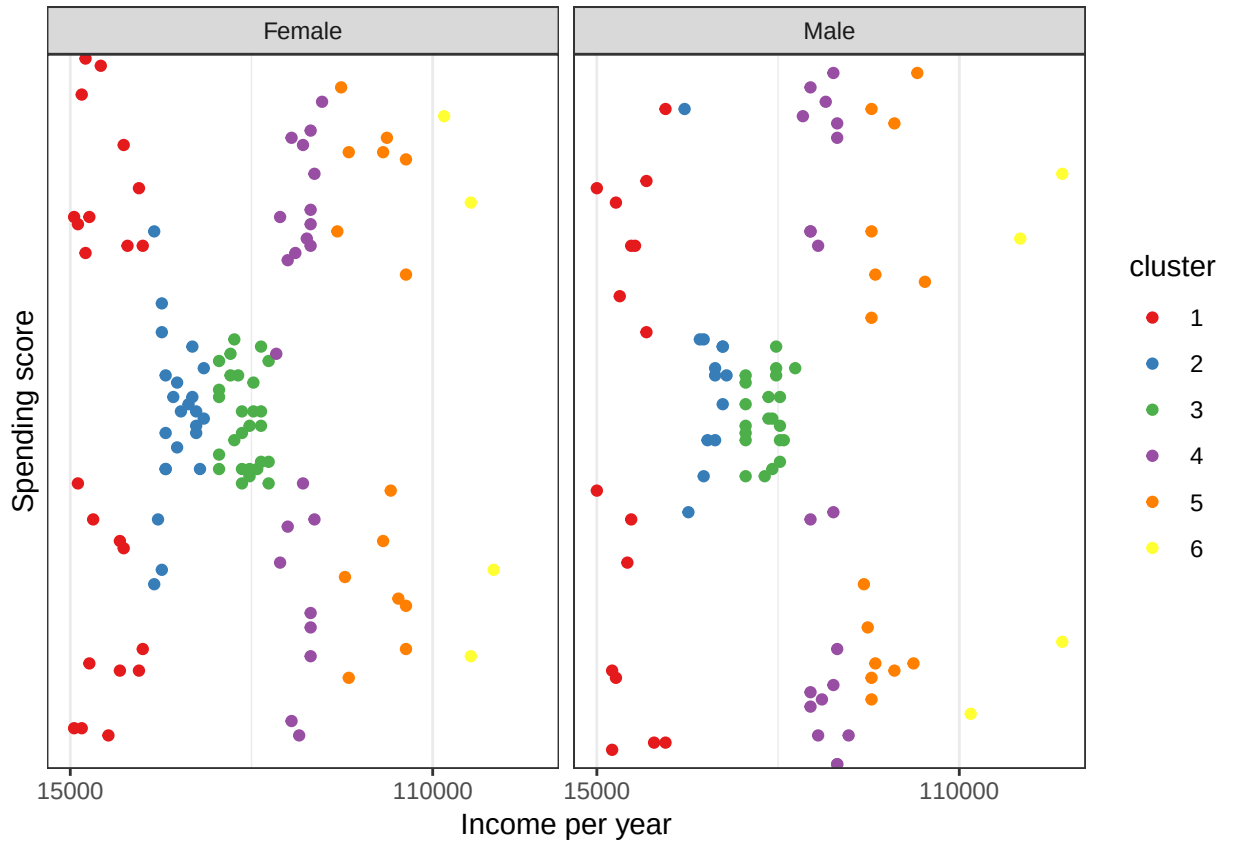
Analyzing this graph, it will be optimally to choose 6 clusters. Let's assign all customers to specific cluster and compare these groups again with ANOVA analysis.

```
##    CustomerID           Age annual_income
## 4.951777e-135  2.210468e-02 3.134022e-136
```

Here we can see p-value for variables that less then 0.05. It means that differences in all these groups (dividing by clusters) significant and we can't accept hypotesis that all groups are equal.

Analyzing of spending money

Well, now we can look at graph and see what behaviour of "spending money capacity" for all 6 subgroups of customers.



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 ages of customers.
    - customers from 1, 4, 5, 6 groups prone to spend or too much (high score), or too small (low score) ammount of money. But it's interesting, that the 1st group has the smalest income, while 4-6 groups have the highest incomes.
    - customers from 2 and 3 groups prone to spend money in the average level of "spending score" and this type of clients also have average incomes per year.
3. It requires more data to analyze why customers from 1, 4, 5, 6 groups behaves differently. Perhaps we will have influence of other factors, which we don't have in this dataset.