Cost Statistics And Income Urban Household Year 2015 In Iran

Hanie Jalili

Table of Contents:

First chapter: Introduction of data and its preparation

Introduction

The plan to collect statistics on the expenditure and income of urban households in the country has been started since 1347 by the Statistics Center of Iran. Since 1353, in addition to the cost, the income of urban households has also been included in this census, and it has been implemented every year in the country except for the years 1355, 1357, and 1360. Every year with the necessary reviews and compliance with the standards advice Internationally, the changes needed in line with the evolution of the preparation and implementation of the plan are applied by the relevant experts.

Target

The aim of this project is to find the best predictive model to predict the income category of urban households using the data of the household income and expenditure plan in the year 1400, which was prepared by the Iranian Statistics Center.

The budget and planning organization of the country can use this model and according to the situation subsistence people and with the aim of increasing their welfare and income level, plan the next year's budget and consider facilities for people who are in low income percentiles. For example, if the pension has a great impact on the percentage of people's income, you can focus on it.

Questionnaire data and information

Questionnaire of expenditure and income plan of urban households includes the following sections:

- Social characteristics of family members
- The details of the place of residence and facilities and the main necessities of life
- costs of Food and non-food households
- Household incomes

Definition of variables

variable defining	Variable name
City code	**C.SH**

The gender of the head of the household	JENS.S
Age of the head of the household	SEN.S
Is the head of the household literate or not?	SAVAD.S
Educational qualification of the head of the household	M.T.S
The activity status of the head of the household	V.F.S
How to occupy a residential house	N.T.M
Number of rooms available	T.O
The basement level of the residence	S.Z
Major building materials for the residence	M.O.B
personal car	HERE
motorcycle	MO
Bike	DO
set voice	ZABT
Color TV	TV
All kinds of subsidies and tablets	PC
Cellular phone	TEL.H
vacuum cleaner	JAROO.B
washing machine	M. LEBAS
Sewing Machine	CHARKH.KH
Fan	PUNK
dishwasher	M. ENVELOPE
Oven	OJAGH.GAZ

variable defining	Variable name
Household food and drink expenses in the last month	H. KHORAKI. NOOSHIDANI
The cost of household communication in the last month	H.ERTEBATAT
Household health expenses in the last month	H.BEHDASHT
Household transportation expenses in the last month	H.HAMLONAGHL
The cost of miscellaneous goods or services of the household in the last month	H. KALA. MOT
Housing expenses – water, fuel, lighting, etc. in the last month	H. MASKAN
Costs of furniture and household appliances and their usual maintenance in the last month	H.MOBLEMAN
Household clothing expenses in the last month	H. POOSHAK
The cost of purchasing household durable goods in the last 12 months	KHARID.KALA.BADAVAM
Household investment cost in the last 12 months	H.SARMAYEGOZARI

variable defining	Variable name
Total continuous and non-continuous gross incomes of the last 12 months of employed members of the household before deductions	M.DARAMAD.NAKH
Wages and salaries for the past 12 months	HOOGHOOGH.MOSTAMAR
Discontinuous benefits of the last 12 months	GH. MOSTAMAR
Total net income from sales	M.DARAMAD.KH.F
Gross receipts from sales	DARYAFTI.NAKH.F
Income from renting business premises, gardens, land,real estate Home, business rights, movable and immovable property, etc. in the last 12 months	DARAMAD.M.KH.1
Proceeds Saving Fixed deposit, shares, insurance and the like in the last 12 months	DARAMAD.M.KH.2
RightsRetirement In the last 12 months	DARAMAD.M.KH.3
Scholarship in the last 12 months	DARAMAD.M.KH.4
Monthly net income	DARAMAD. KHALES

* **Appendix 1:** for data analysis because the time frameOne Month It is considered that we divide the variables of household durable goods purchase cost in the last 12 months and household investment cost in the last 12 months into 12.

* **Appendix 2:**According to the purpose and necessity of this project, we created a monthly net income variable using the following formula.

(continuous wages and salaries for the past 12 months (after deductions) + non-continuous benefits for the past 12 months (after deductions) + income from renting business premises, gardens, land,real estate Home, business rights, movable and immovable property and the like in the last 12 months + income from saving fixed deposits, shares, insurance and the like in the last 12 months + salaryRetirement in the last 12 months + scholarships in the last 12 months + total net income from sales) divided by 12

Then, using code, we converted it to a binary variable. (If the total monthly income was more than the 90th percentile, 1 and otherwise 0)

Defining the meanings of variables

- **Urban locations:**

In this plan, urban points are defined as all the points that had a municipality at the time of the general census of 2011.

- **Household (typical resident):**

A household consists of several people who live together in a fixed residence And they spend money with each other and usually eat together eat. A person who lives alone is also considered a household.

- **Households:**

The head of the household is one of the members of the household who is known as this in the household.

- **Literate:**

Any family member aged 6 years or older who can read simple texts in Farsi or any other language to writeWhether he has an official educational certificate or not, he is considered literate.

- **Studying:**

Every family member aged 6 years and older who is studying according to the official educational programs of the country is considered to be studying.

- **Student:**

People aged 10 years and more who are employed, unemployed, looking for work and haveIncome Those who were not without work, and were engaged in studies in the 7 days before the day of the census, are considered as students.

- **work:**

Work is any intellectual or physical activity for the purpose of earningIncome (cash and non-cash) and its purpose is to produce goods or provide services. Therefore, activities such as sports (except professionally), housekeeping, education, etc. are not considered work, because these activities are for the purpose of earning.Income It is not done.

- **My job status:**

1. Employer: is someone who employs at least one wage earner to perform his work activities.
2. Independent worker: He is a person who, within seven days before the visitagent A statistician, a salaried person should not be employed and himself also Do not take wages and salaries.

3. Get wages and salaries in the public sector: He is a person who works in the public sector and receives wages and salaries (cash and non-cash) for the work he does.
4. Get wages and salaries of the cooperative sector: He is a person who is active in the cooperative sector and receives wages and salaries (cash and non-cash) for the work he does.
5. Receive wages and salaries, private sector: is someone who receives wages and salaries (cash and non-cash) for individuals orInstitute Personalities of workMay Kind.
6. Unpaid family worker: An unpaid family worker is a person who works for a member of his family with whom he has a family relationship, and for this reason, he does not receive wages or salary.

- **How to occupy the residence of the family:**
  1. Ownership of the land and nobles: It is the way of occupying the place of residence where the family is the owner of the land and its building.
  2. Ownership of nobles: It is the way of occupying a place of residence where the household is only the owner of its building.
  3. Leasing: It is the way of occupying a residence that is rented.
  4. Mortgage: The method of occupying a place of residence is that the household has deposited money in the form of a good loan with the owner for its temporary possession.

5. Against service: It is the way of occupying a place of residence that is placed at the disposal of the household in exchange for doing the work of one or more members of the household.
6. Free: It is the way of occupying a place of residence that the household does not pay for its use and does not do anything.
7. Other: It is the way of occupying the place of residence that the household has in a way other than the above five modes.

- **room:**

In a normal residential unit, it means a room, an enclosed space, etcYou have a roof It has an area of at least four square meters and a height of two meters. Except for tents, capers, slums and the like as The room is used by the family, even if fourSquare meters. Even if it does not have an area, the room is considered.

- **Income:**

Income is all the funds and the value of the goods that are against the work done or the capital used or through other sources (rights retirement, transferable incomes and the like) have been assigned to the household at the desired statistical time.

1. Net income from public wages and salaries: the income of all members of the household in the public sector, includingReceipts Continuous and discontinuous Boolean and non-monetary They are after deduction of taxes and retirement.

2. Net income from wages and salaries of the cooperative: The income of all members of the household in the cooperative sector includes their monetary and non-monetary receipts after deduction of taxes and retirement.
3. Net income from private wages and salaries: the income of all members of the household who are wage earners and salary earners in the private sector, which includes wages and salaries and continuous and non-continuous monetary and non-monetary benefits after deduction of taxes and retirement.
4. Net income from agricultural self-employment: the income of people from the household who work in the agricultural sector as an employer or self-employed, after deducting employment expenses.Income It is considered as one of the free agricultural jobs.
5. Net income from freelance jobs non-agricultural: The income of people from the household who are employed as self-employed or self-employed in non-agricultural jobs, after deducting employment expenses, asIncome Net refers to non-agricultural jobs.
6. Miscellaneous income: all the funds and the value of the goods that the household received from a way other than the employment of its members, such as the rental of movable and immovable property.Transported، Rights Retirement and the like are considered miscellaneous incomes.

- **cost:**

The monetary value of goods or services provided by the household for the consumption of members or gifts to others is called cost. The goods or services provided can be provided to the household through purchase, home production, against service, from the place of purchase and free of charge (from organizations), which is calculated in monetary form and is included in the cost.

1. Gross cost: Gross cost is the value of the product without deducting second-hand sales of the same product.
2. Net cost: When the value of sold second-hand goods is subtracted from the value of similar goods purchased in the statistical period, the remainder will be the net cost of the household in the relevant item.

Actions taken to clean up data

- In the food and beverage cost variable, we had 28 missing data. We used the distribution of the values of this variable to find the appropriate method. The variable distribution of food and beverage costs is not symmetrical and has a skewed state. Therefore, the best method is the middle of the data; Therefore, we replaced the missing data with the median of the data.

Define categorical variables

Note: Changes have been made in the categories of M.T.S and M.O.B variables.

variable	NO	Yes
HERE	0	1
MO	0	1
DO	0	1
PC	0	1
TEL.H	0	1

M. LEBAS	0	1
M. ENVELOPE	0	1
ZABT	0	1
TV	0	1
OJAGH.GAZ	0	1
JAROO.B	0	1
CHARKH.KH	0	1
PUNK	0	1

m.t.s	Name
0	Literary
1	Ebtedai
2	Guide
3	Fashion balance
4	Diploma.Pishdaneshgahi
5	Foghe. Diploma
6	Licence
7	Foghe.lisans.Dr.herfei
8	Dr. specialist
9	Savad.amuzi.gheyre.asmi

variable	NO	Yes
SAVAD.S	2	1
jens.s	Name	
1	man	
2	woman	

n.t.m	Name
1	Melki. arse. ayan
2	Melki. ayan

3	Ejare
4	Rahn
5	Darbarabare.khedmat
6	Raygan
7	Sayer

v.f.s	Name
1	Shaghel
2	Use it
3	Daramad.Bedunekar
4	Mohasel
5	Khanedar
6	Sayer

m.o.b	Name
0	Tamam.chub/Khesht.gel
1	Ajor.ahan
2	Ajor.chub
3	Block.simani
4	ok.ajor
6	Khesht.chub
8	Sayer

Number of data

Type	Number of data
Training*	**1500**
Validation	**300**
Test	**121**

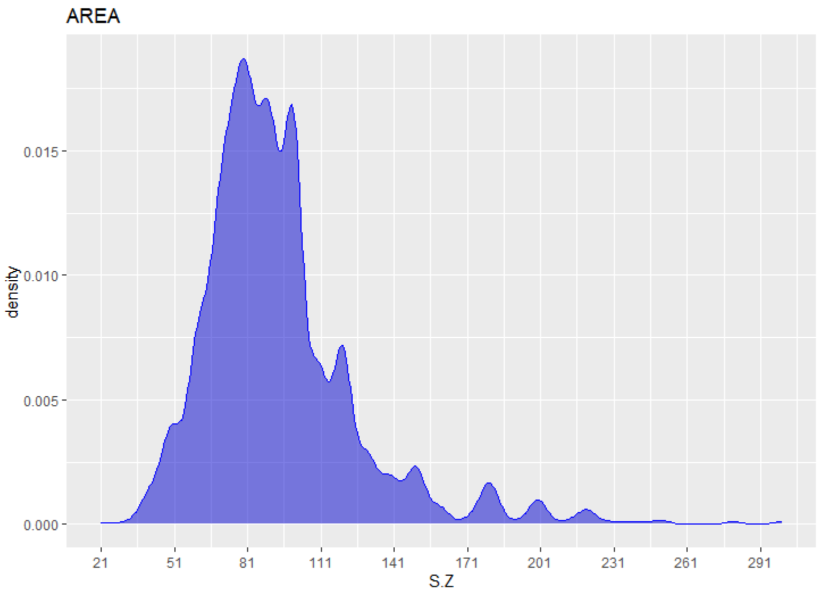
***Important note:** After classifying the response variable (net monthly income) into the mentioned categories, we find that category 1 is rare in this data. That is, the number of response variables = 1 in this data is very small compared to the total number of data. That is why it is necessary to use oversampling methods. For this, we use the combination of oversampling and undersampling methods, and from the 1500 data that we considered for training the model, we generate 10000 data with an almost equal number of categories zero and one, and from now on these data will be the training data of the model. we say. In fact, we train the model with these data and validate with the validation data on which we did not do any oversampling or undersampling.

The second chapter: data visualization

- univariate

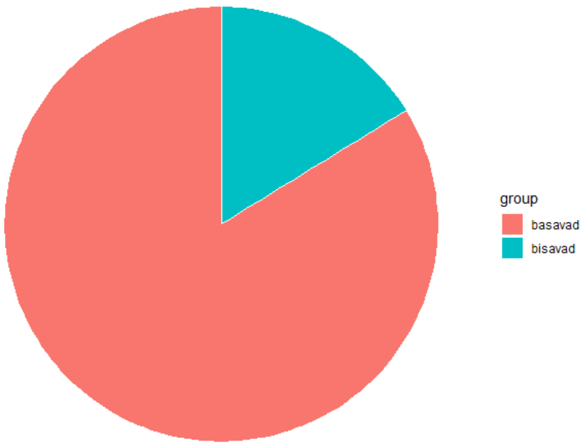
s.z variable diagram (land area)

Most people live in houses between 51 and 100 meters.



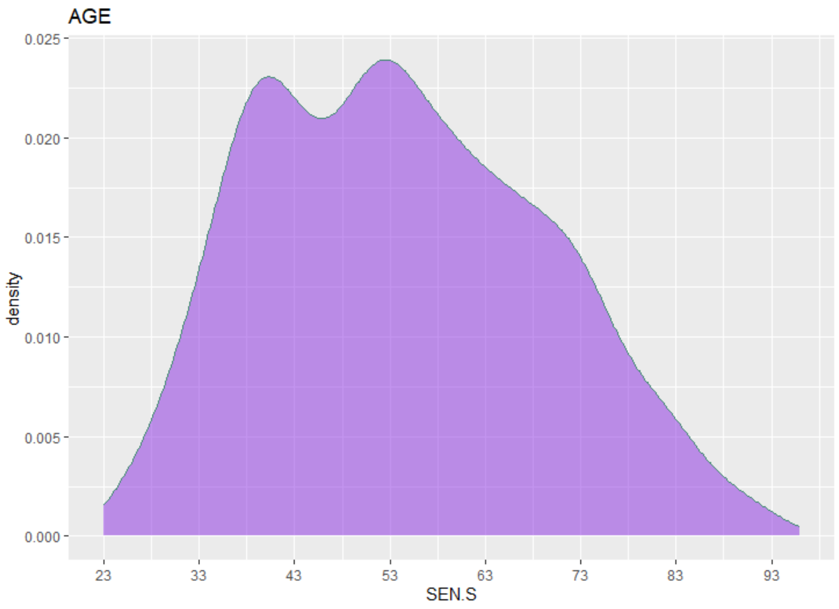
savad.s variable graph

Many people participating in this survey are literate.



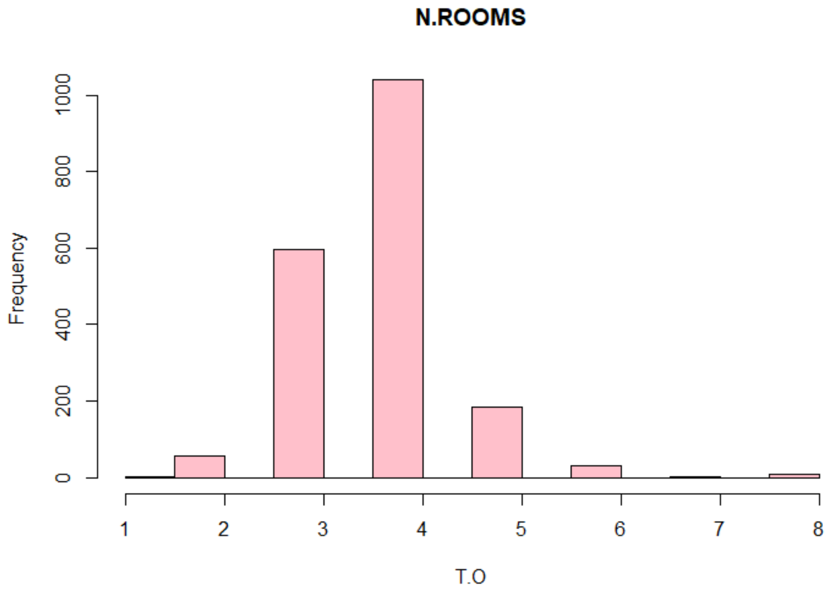
graph of sen.s variable (supervisor's age)

At the age of 3843 years old And we have the largest number of supervisors aged 49 to 58.



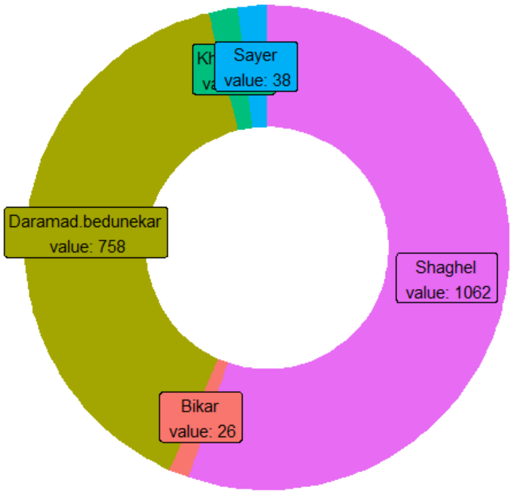
Variable graph t.o (number of rooms)

The highest number is related to houses with 4 or 3 rooms.



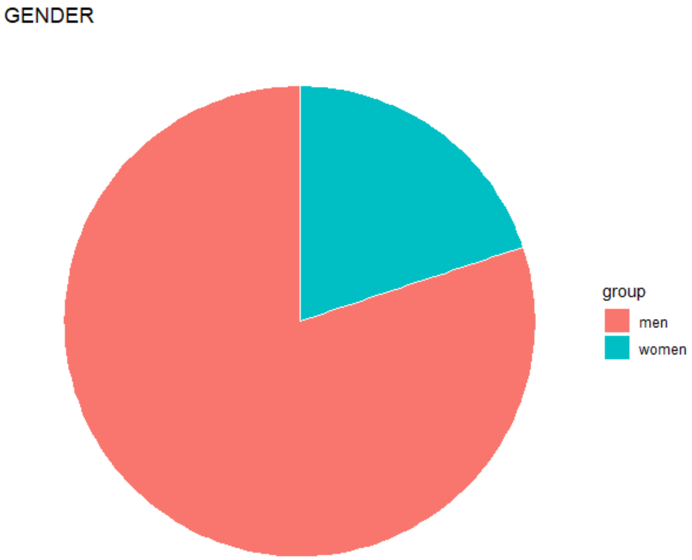
v.f.s variable chart (current status of supervisor)

Most supervisors are employed or have income without work.



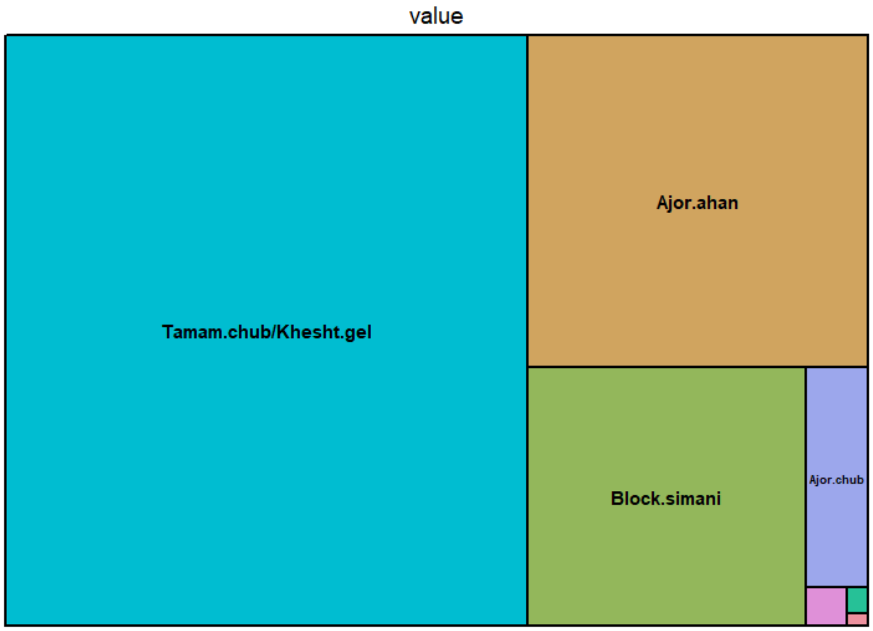
Jens.s variable graph (supervisor gender)

Most of the supervisors are men.



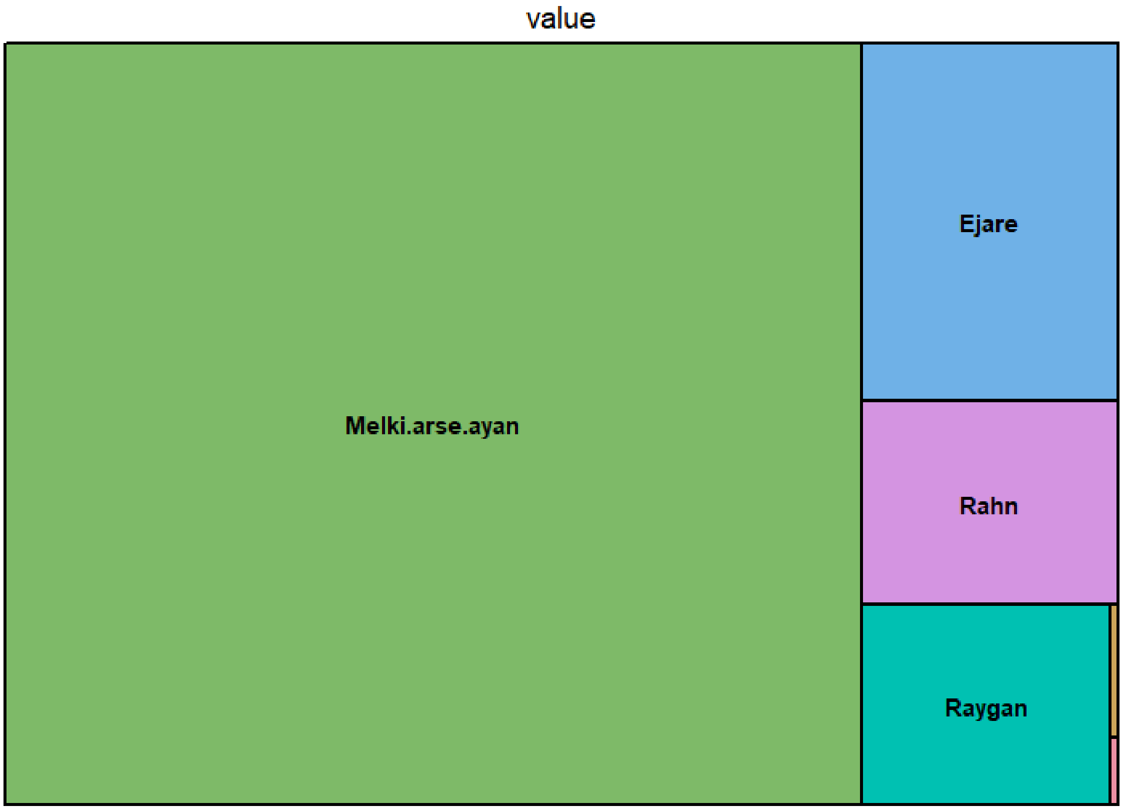
m.o.b variable diagram (major building materials)

Wood, clay and mud, brick and iron are the main building materials.



n.t.m variable diagram (how to occupy the house)

Most of the ways of occupying the house are related to rent and real estate.



variable diagram m.t.s (supervisor's educational qualification)

As we have seen before, many supervisors were literate and according to this graph, it can be seen that most of them have studied up to elementary, pre-university and middle school levels.

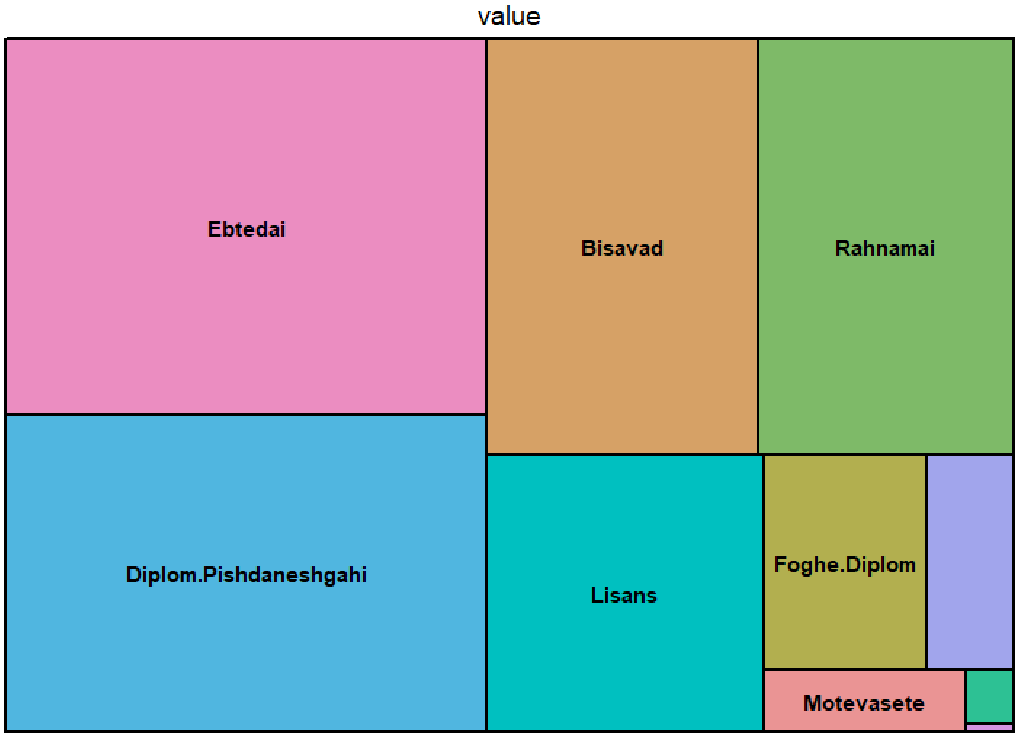
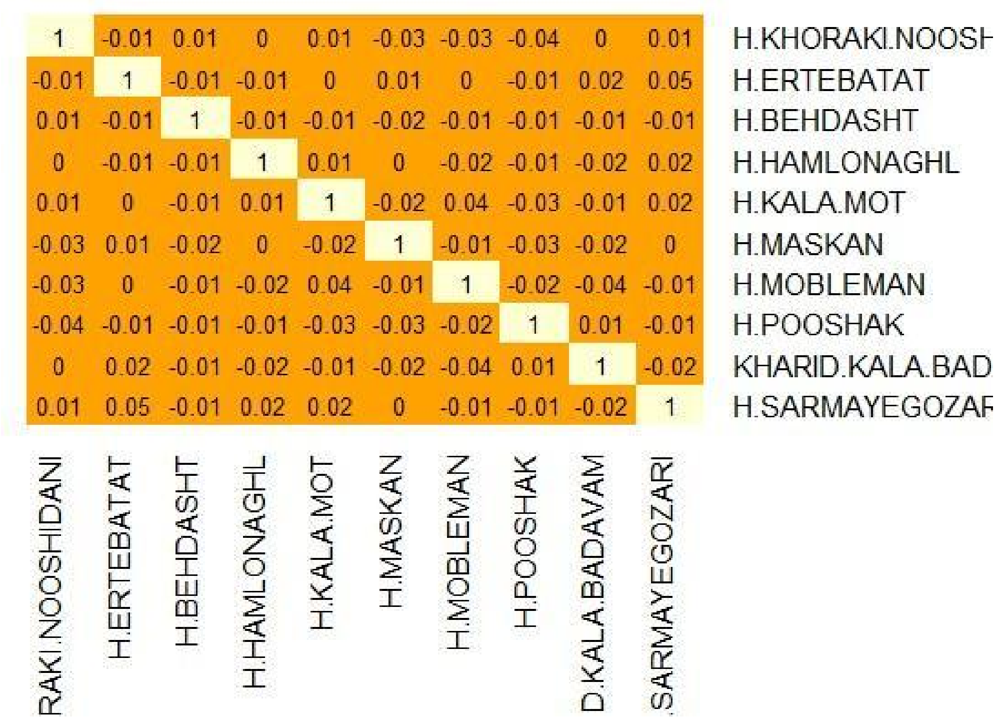
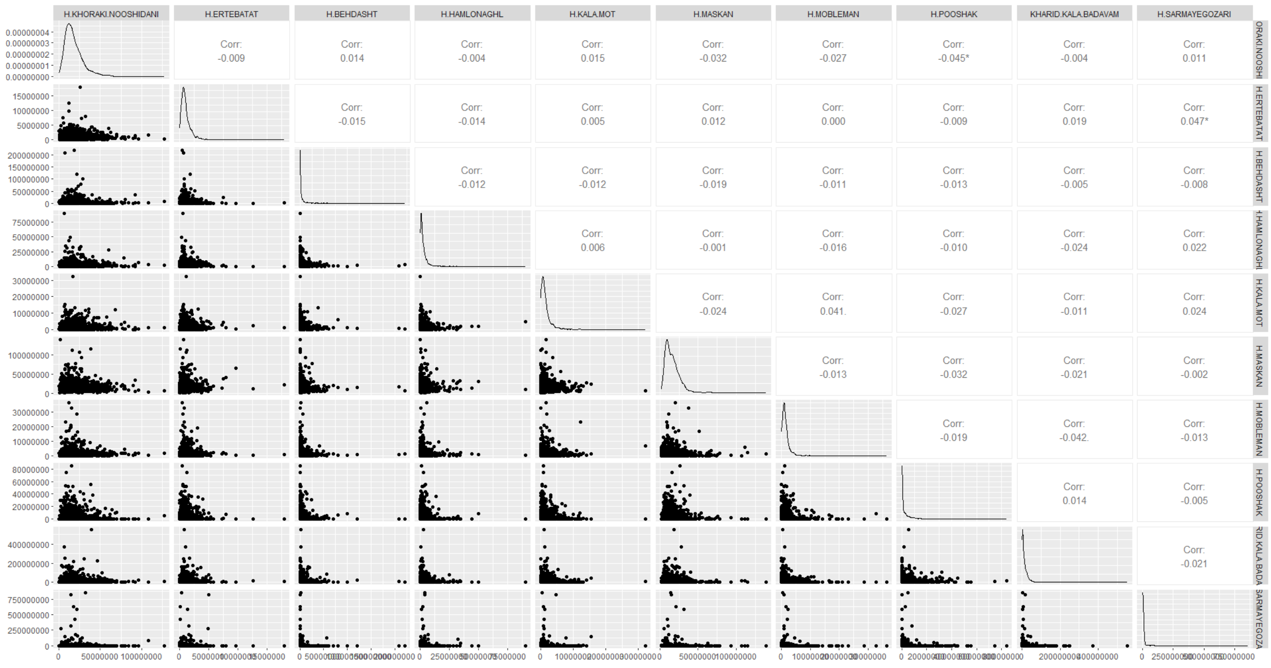


Diagram of cost variables (cost of food, clothing, furniture, etc.)

In this graph, you can see and compare the correlation between different costs. For example, the highest correlation is related to communication cost and investment cost. The cost of furniture and the cost of miscellaneous goods are also highly correlated. Also, the cost of clothing and the cost of food have a high correlation with each other, which is considered an obvious issue because food and clothing are among the basic needs that people pay for.



RAKI.NOOSHIDANI	H.ERTEBATAT	H.BEHDASHT	H.HAMLONAGHL	H.KALA.MOT	H.MASKAN	H.MOBLEMAN	H.POOSHAK	D.KALA.BADAVAM	SARMAYEGOZARI	
1	-0.01	0.01	0	0.01	-0.03	-0.03	-0.04	0	0.01	H.KHORAKI.NOOSH
-0.01	1	-0.01	-0.01	0	0.01	0	-0.01	0.02	0.05	H.ERTEBATAT
0.01	-0.01	1	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	H.BEHDASHT
0	-0.01	-0.01	1	0.01	0	-0.02	-0.01	-0.02	0.02	H.HAMLONAGHL
0.01	0	-0.01	0.01	1	-0.02	0.04	-0.03	-0.01	0.02	H.KALA.MOT
-0.03	0.01	-0.02	0	-0.02	1	-0.01	-0.03	-0.02	0	H.MASKAN
-0.03	0	-0.01	-0.02	0.04	-0.01	1	-0.02	-0.04	-0.01	H.MOBLEMAN
-0.04	-0.01	-0.01	-0.01	-0.03	-0.03	-0.02	1	0.01	-0.01	H.POOSHAK
0	0.02	-0.01	-0.02	-0.01	-0.02	-0.04	0.01	1	-0.02	KHARID.KALA.BAD
0.01	0.05	-0.01	0.02	0.02	0	-0.01	-0.01	-0.02	1	H.SARMAYEGOZAF



- Two variables

(In the cost graphs for various goods against net income, to understand and increase the readability of the graph, we drew the logarithm of costs against net income.)

A graph of the current position of the supervisor against the net income

It is obvious that the job status of supervisors has a high impact on their net income.

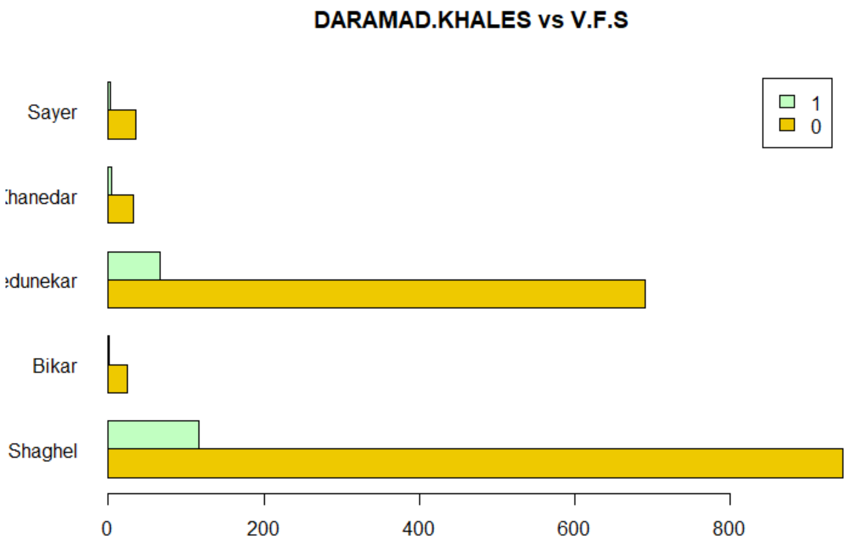
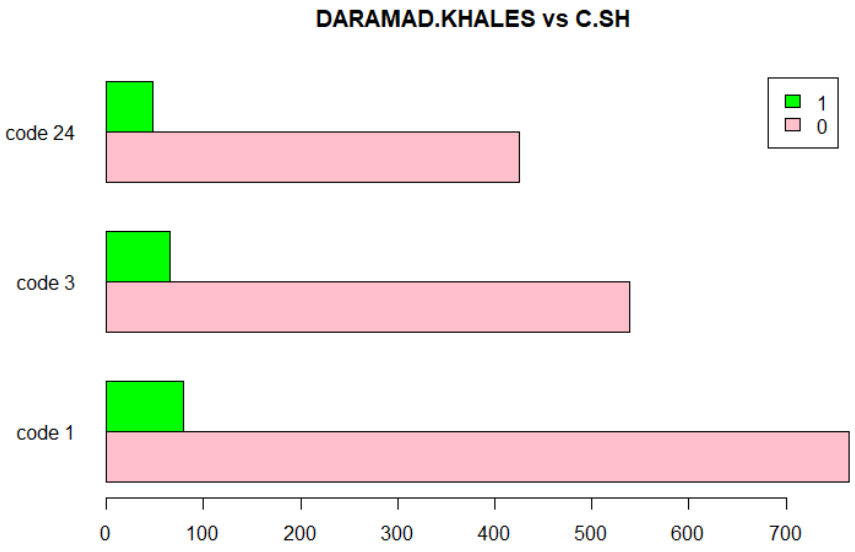


Chart of county code against net income

People's net income is highly related to where they live.



Graph of Gender vs. Net Income

The gender of supervisors is an influencing factor in their net income.

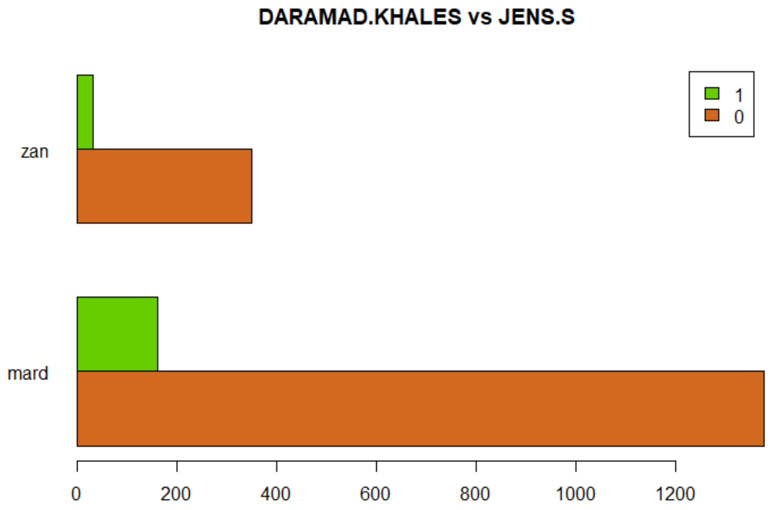


Chart of major building materials against net income

Cement, brick and wood, iron, clay and mud are building materials that have a great impact on people's net income.

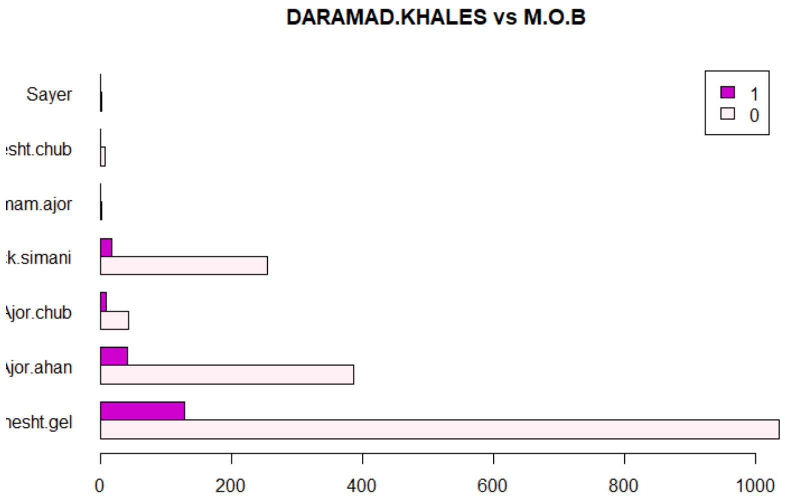


Chart of Superintendent's Education Degree vs. Net Income

Obviously, the supervisor's degree has a great impact on their net income.

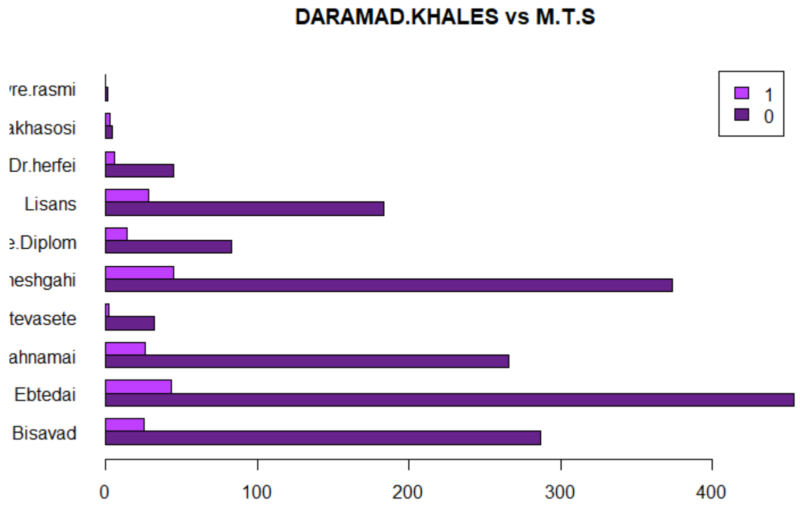
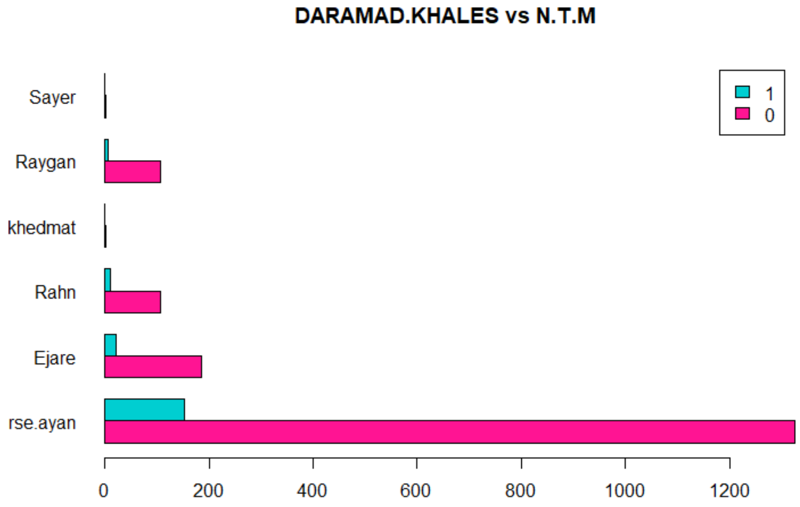
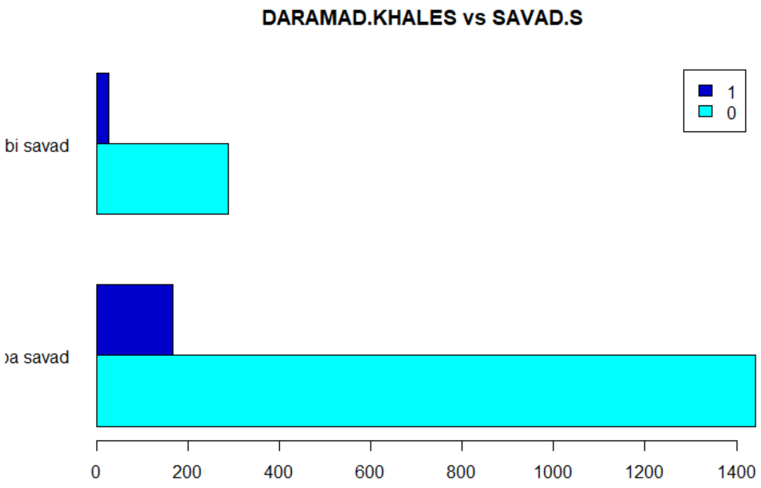


Chart of how to occupy the house against the net income

Among other factors that affect people's net income, we can mention the way of occupying the house.

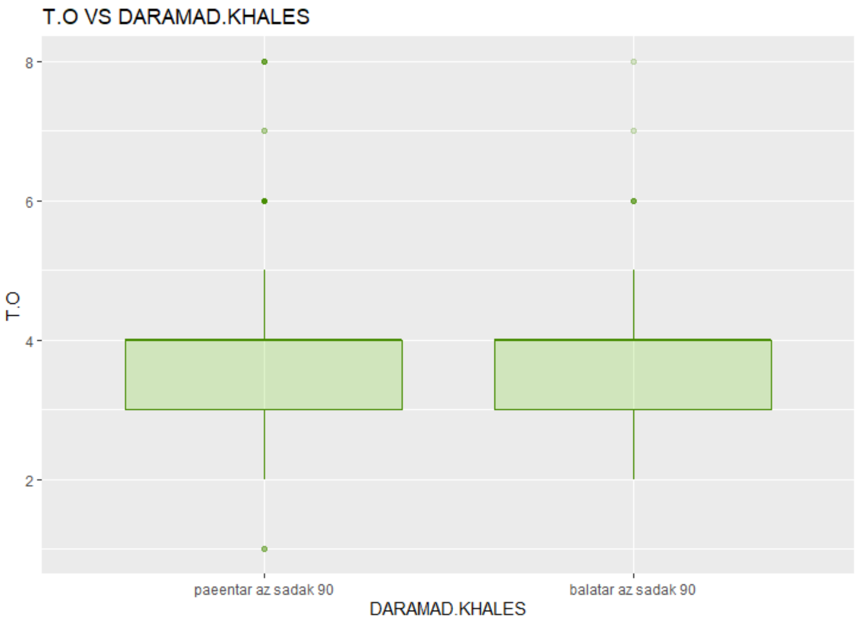


Superintendent's Literacy vs. Net Income Chart

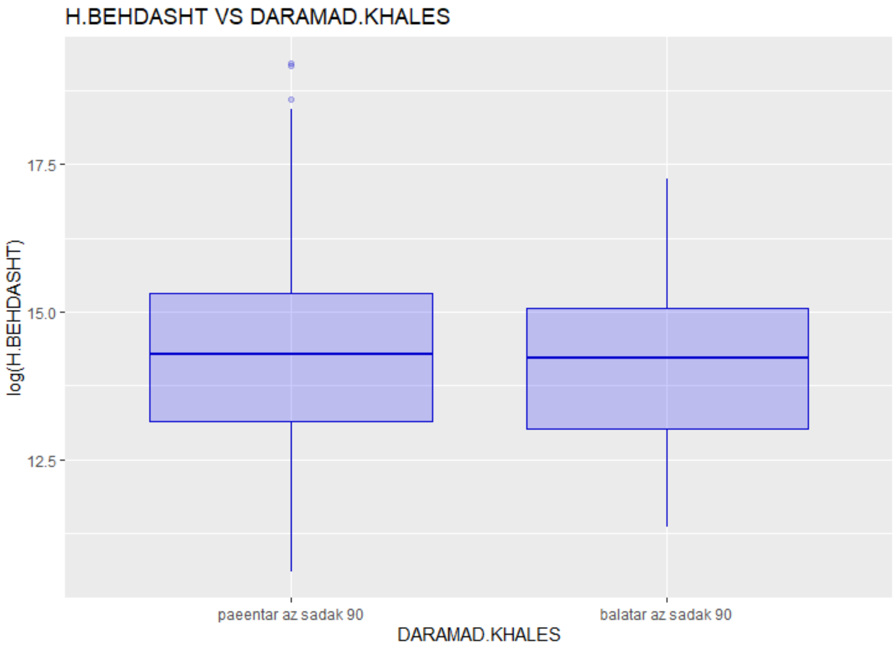


Graph of number of rooms against net income

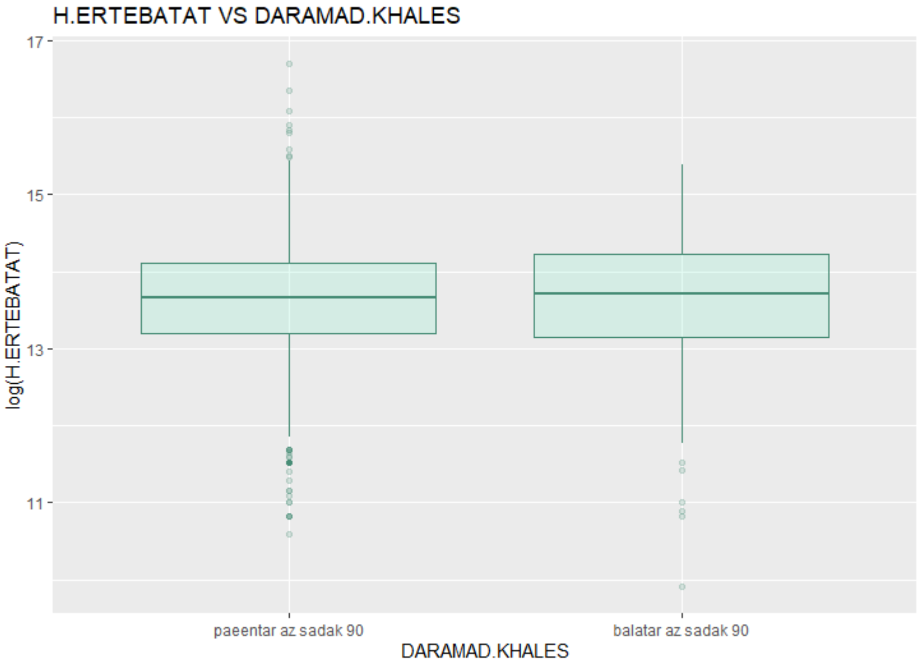
The median of the two graphs and their distribution are the same. That is, the number of rooms does not have much effect on people's net income.



Health Cost vs. Net Income Chart



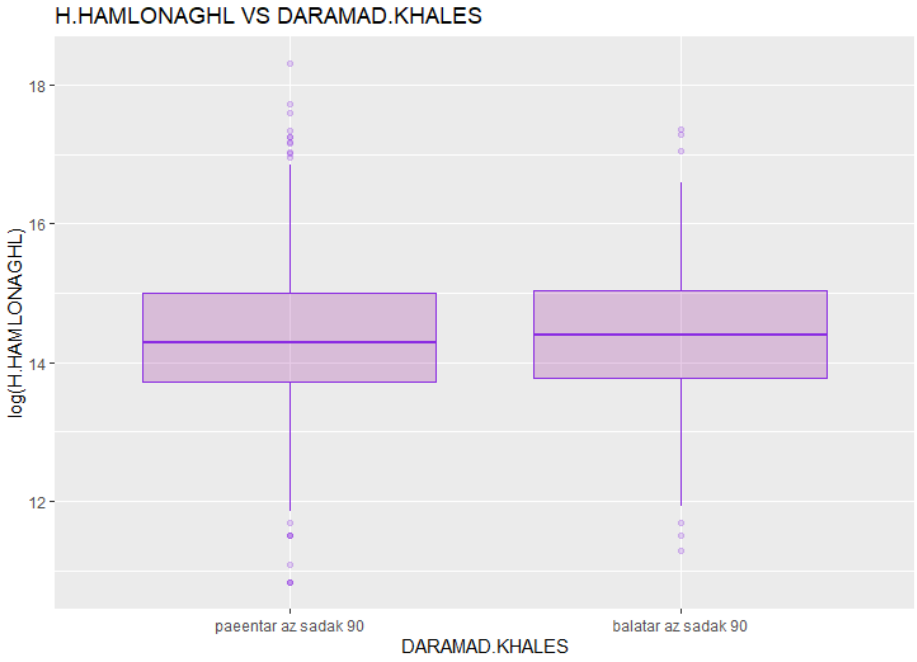
Communication cost versus net income graph



In the health cost graph, the distribution of two box plots and their median are different from each other. That is, this variable affects the net income. The cost of communication is the same.

Graph of shipping cost vs. net income

In addition to the effectiveness of this variable, it can be said that people who have an income above the 90th percentile spend more on transportation.



A graph of the cost of miscellaneous goods against net income

It is interesting to note that people with income less than the 90th percentile spend more on miscellaneous goods than other people.

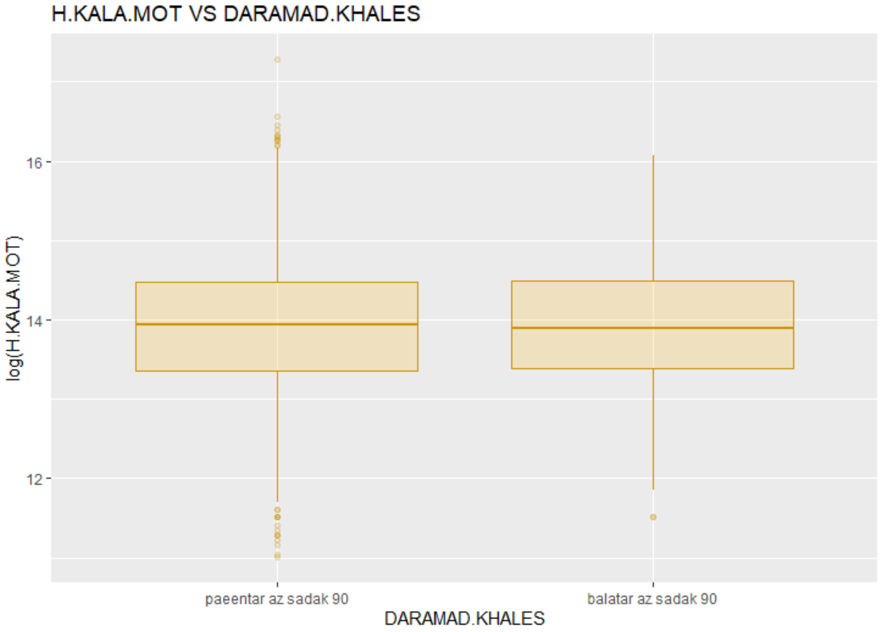
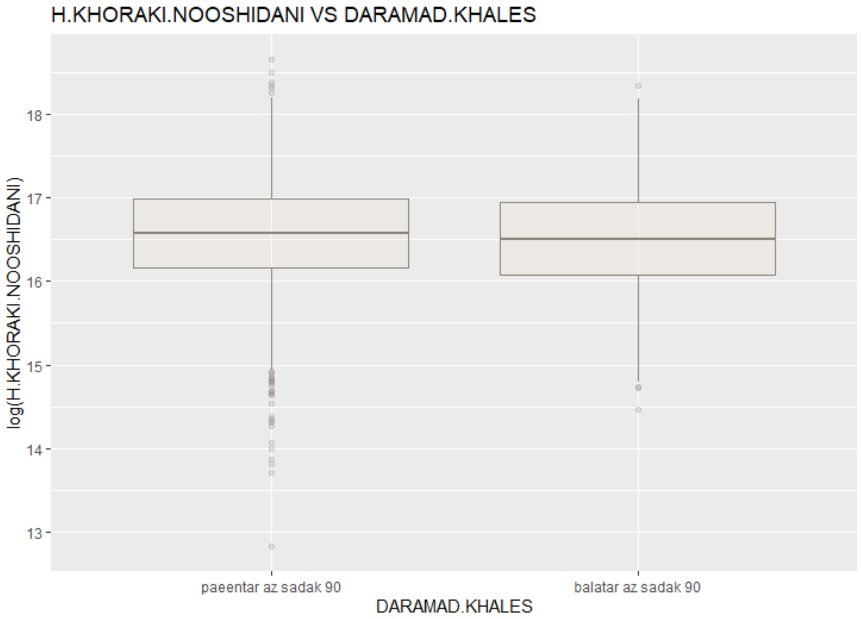


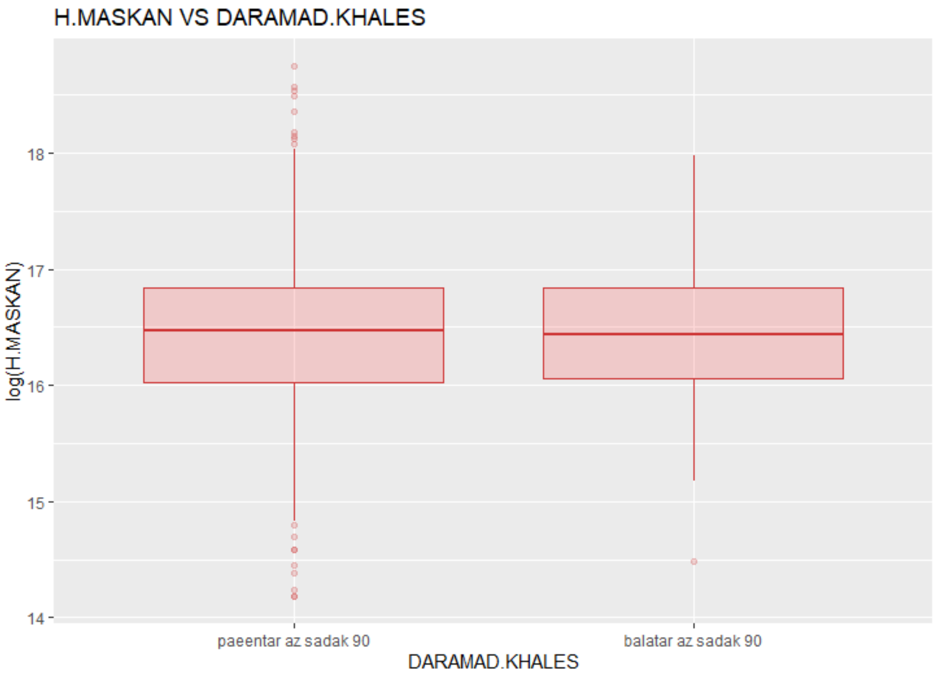
Chart of food and beverage expenses against net income

The median in both parts of the box plot is higher than the median in the rest of the plots. It means that people spend more on food than other consumer items and this is obvious.



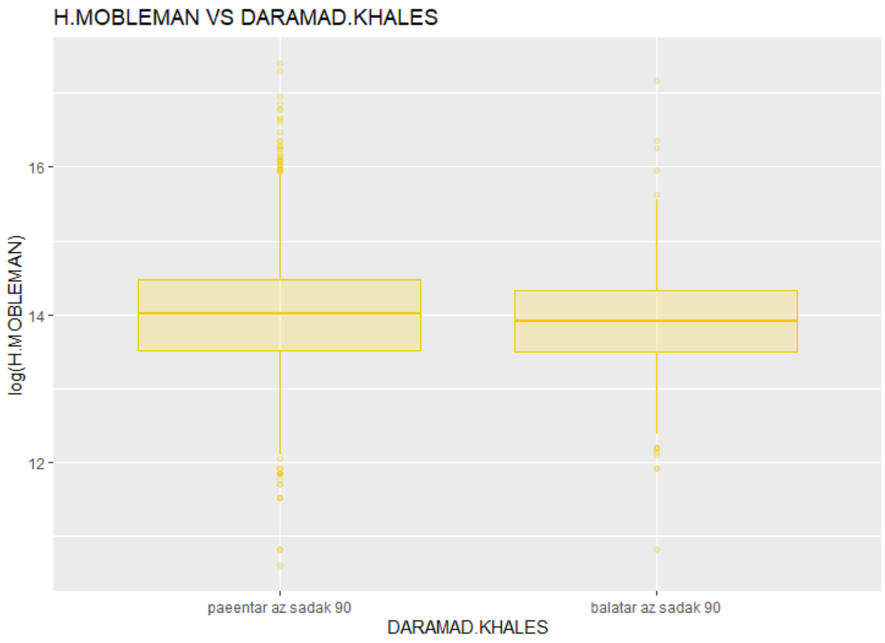
Housing cost vs net income chart

The cost of housing affects the net income of all people in different economic percentiles.



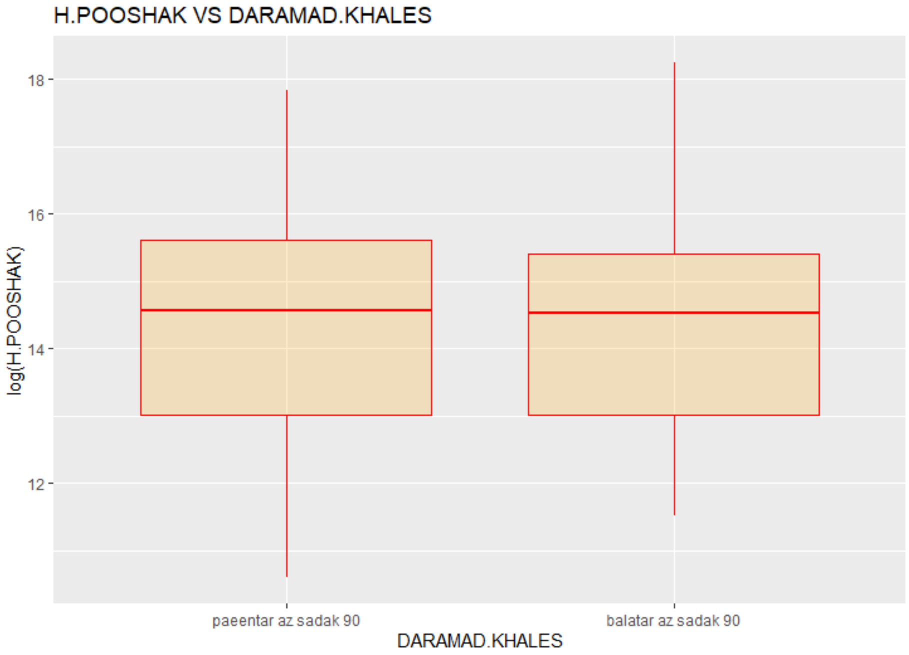
Furniture cost vs. net income chart

Spending on furniture by people with incomes below the 90th percentile is more dispersed. It can also be concluded that they spend more on furniture than people with income above the 90th percentile.



Clothing cost vs. net income chart

maximum People with income above the 90th percentile spend more to buy clothes than people with income below the 90th percentile, and their minimum spending for clothes is also higher. This means that these people pay more for clothes.



Investment cost versus net income graph

The middle of the two box plots have a drastic change with each other. This shows that people with incomes above the 90th percentile spend much more on investments than people with lower incomes. Outlier values are significant for people with incomes below the 90th percentile.

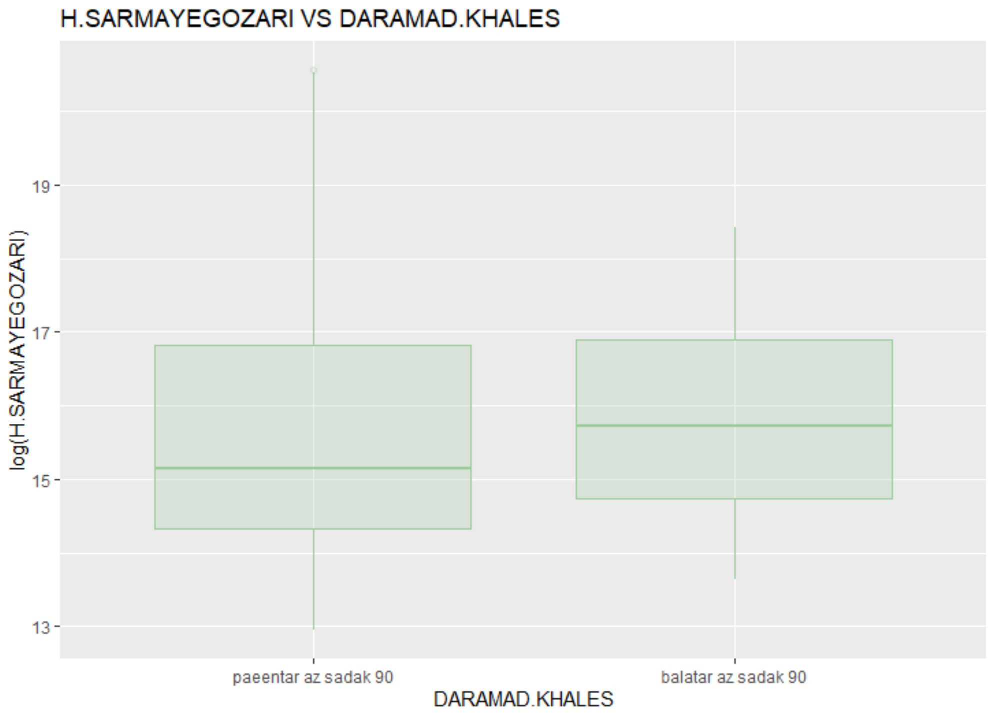
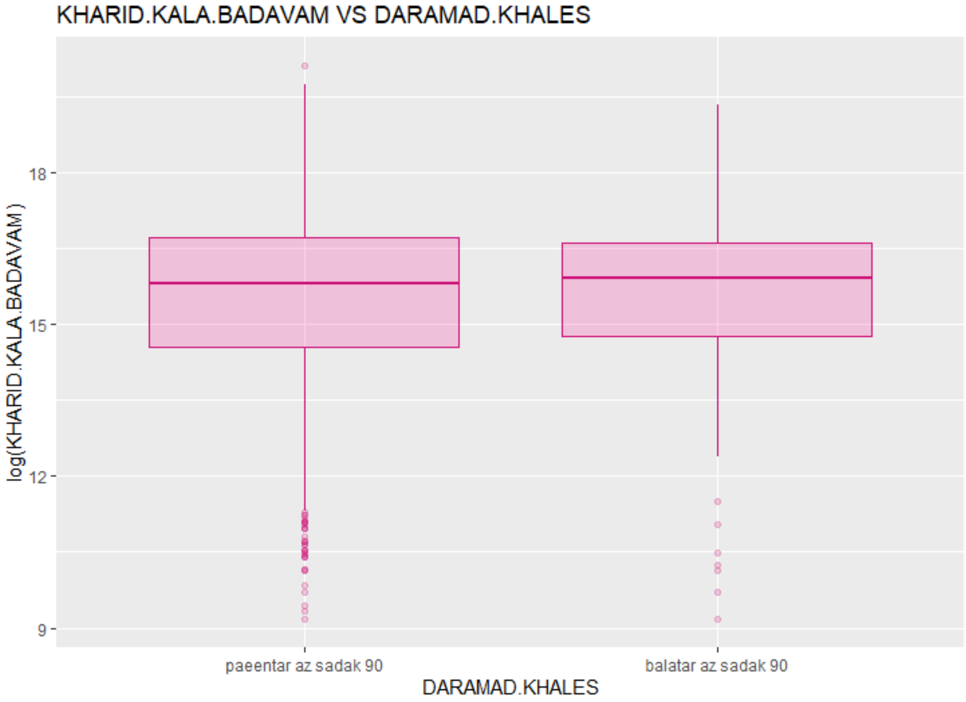


Chart of cost of purchasing durable goods against net income

It is clear that the cost of buying durable goods affects people's income.



Plot of land area versus net income

The area of land for people with income above 90th percentile has more dispersion.

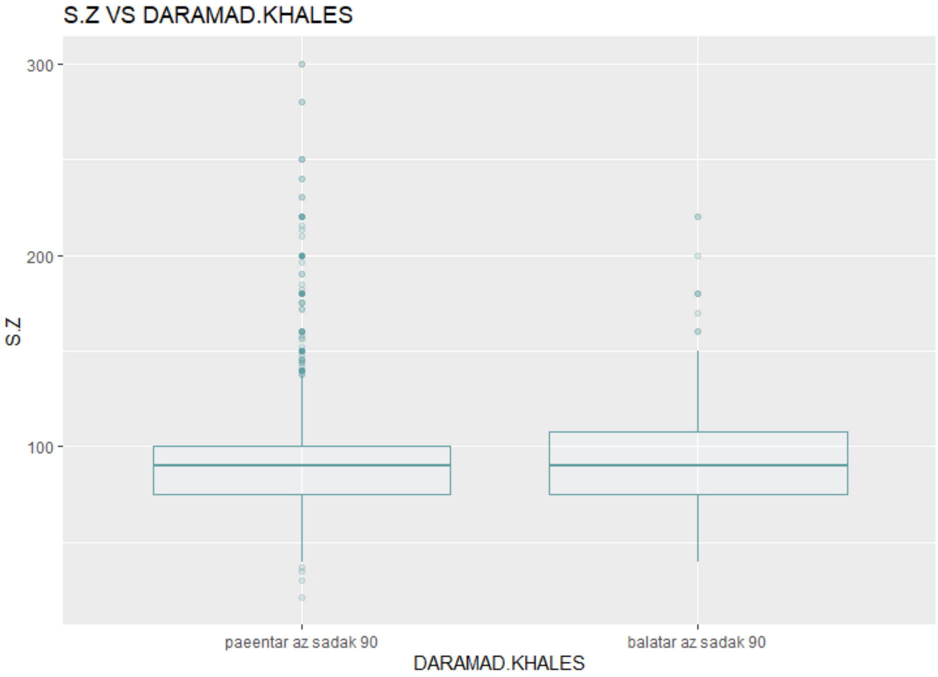
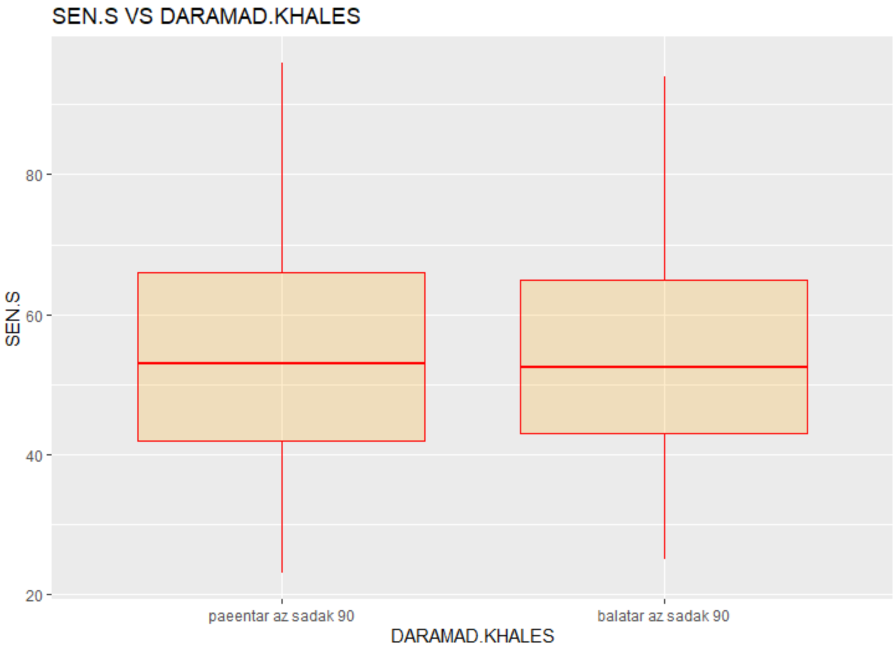


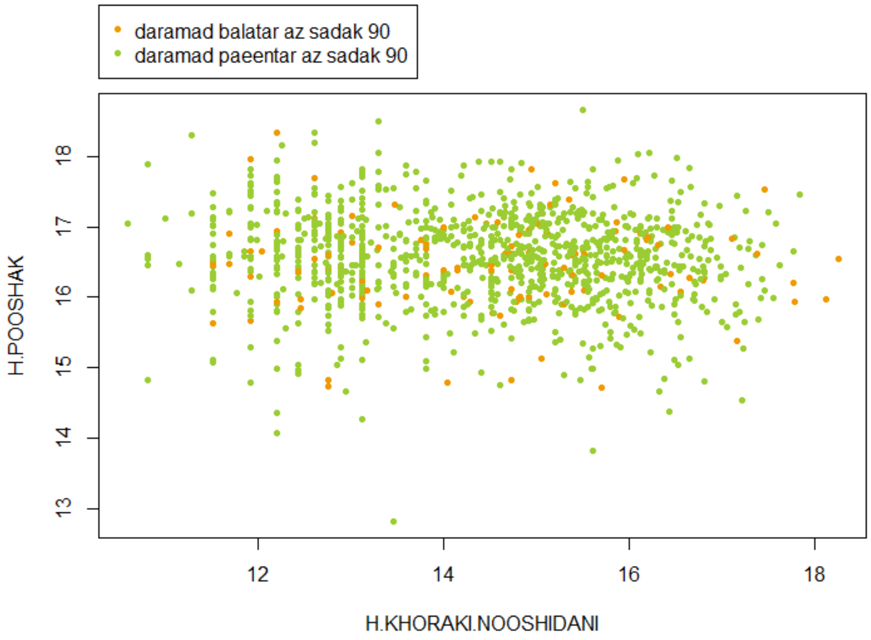
Chart of Supervisor's Age vs. Net Income



- Multivariate

Chart of clothing expenses and food expenses against net income

The high correlation between the cost of clothing and food and the large number of people with income below the 90th percentile is the information obtained from this graph.



Graph of communication cost and investment cost against net income

As it is clear from this graph, these two variables do not have a high correlation with each other.

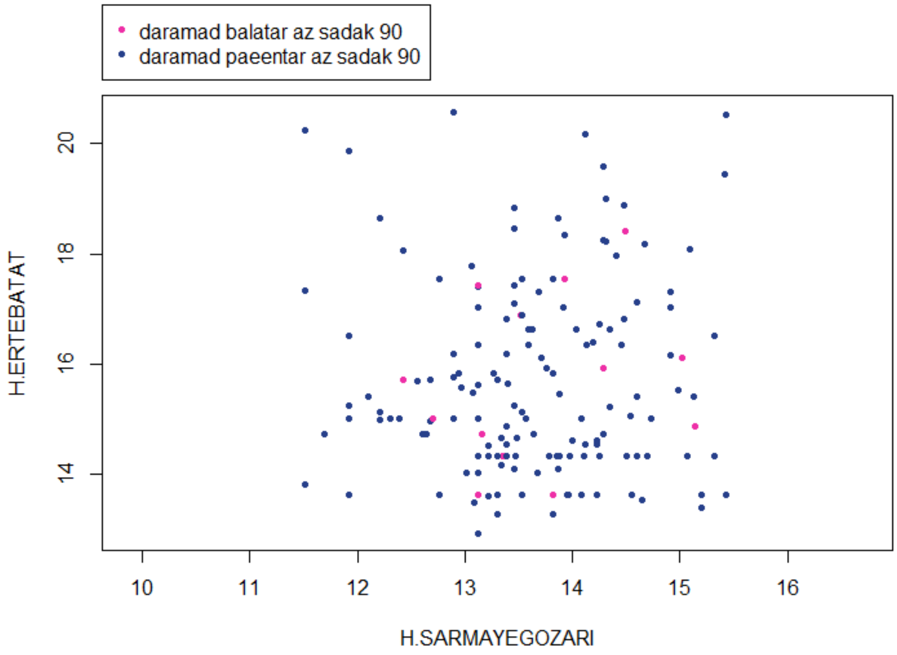


Chart of the cost of miscellaneous goods and the cost of furniture against net income

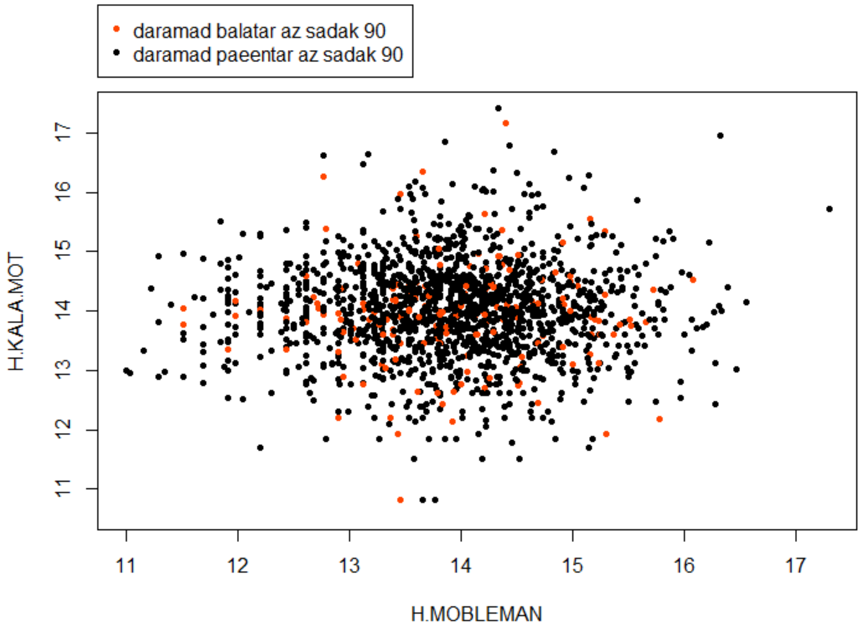
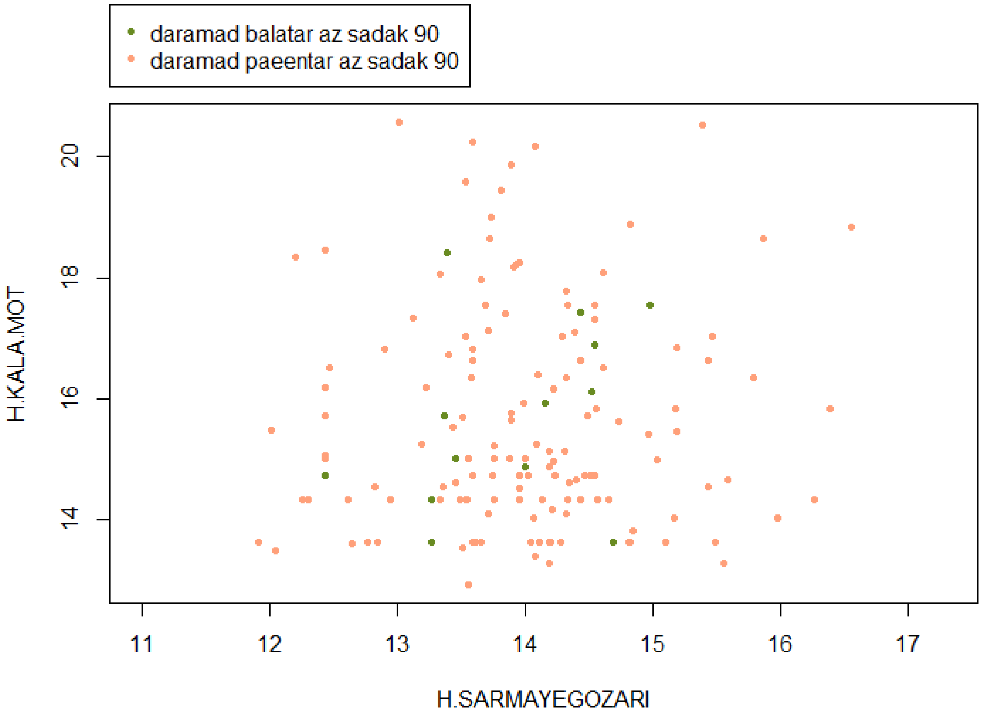
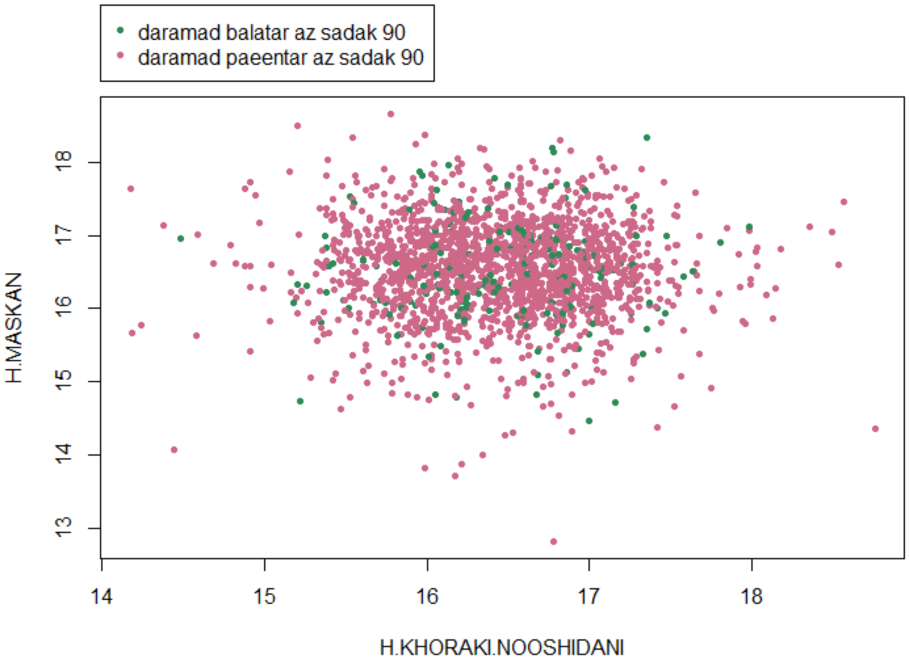


Chart of cost of miscellaneous goods and investment cost against net income

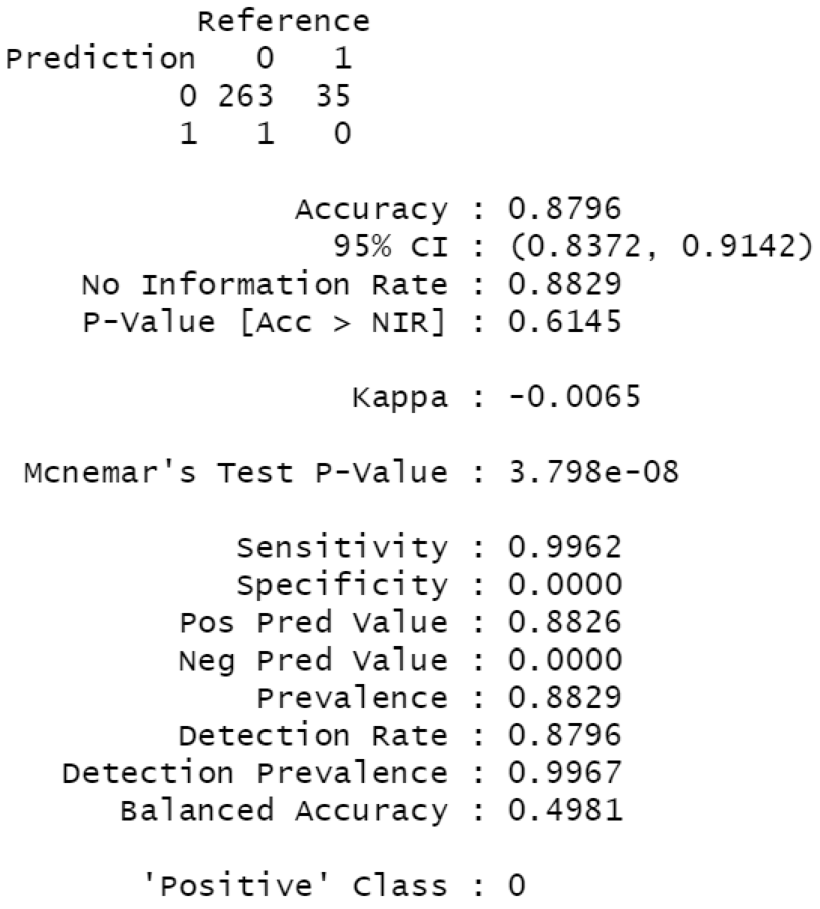


Graph of housing cost and food cost against net income



The third chapter: Implementation of the logistic regression model

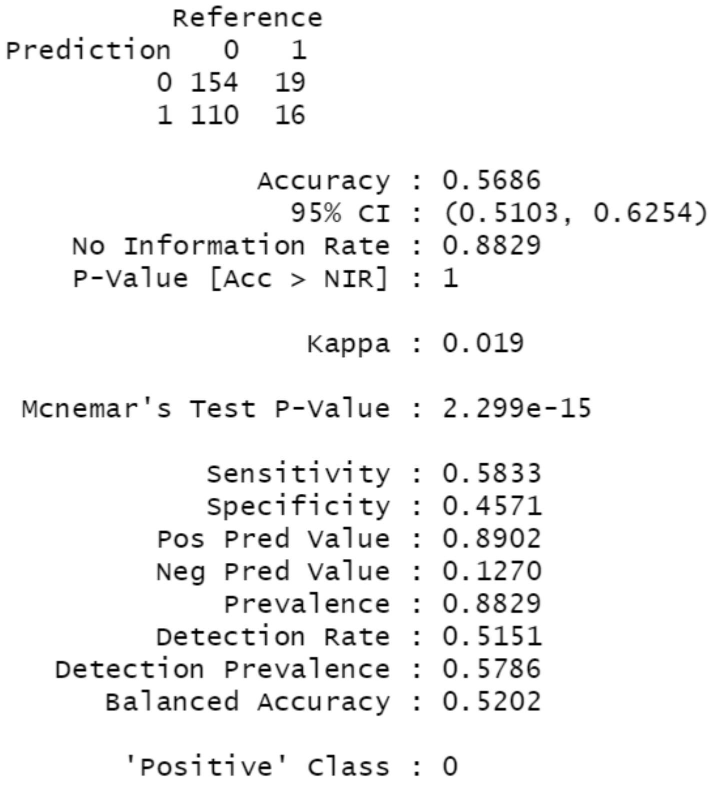
Firstonce We apply the logistic regression model to the training data on which we did not perform oversampling:



```
          Reference
Prediction   0   1
         0 263  35
         1   1   0
                                          
               Accuracy : 0.8796          
                 95% CI : (0.8372, 0.9142)
    No Information Rate : 0.8829          
    P-Value [Acc > NIR] : 0.6145          
                                          
                  Kappa : -0.0065         
                                          
 Mcnemar's Test P-Value : 3.798e-08       
                                          
            Sensitivity : 0.9962          
            Specificity : 0.0000          
         Pos Pred Value : 0.8826          
         Neg Pred Value : 0.0000          
             Prevalence : 0.8829          
         Detection Rate : 0.8796          
   Detection Prevalence : 0.9967          
      Balanced Accuracy : 0.4981          
                                          
       'Positive' Class : 0               
```

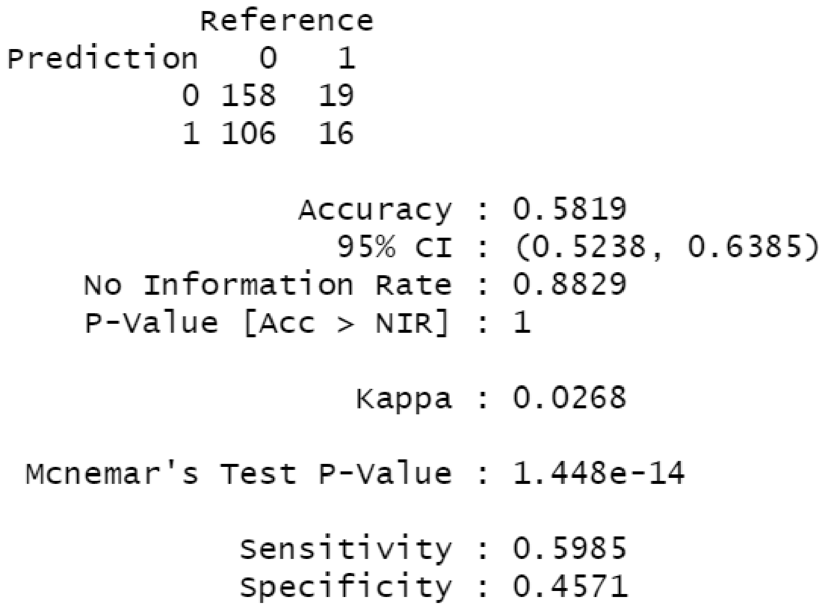
The accuracy of the model is high, but the specification factor is very low, which is not our desired case. (That is, the values of net income, which in reality are equal to 1, are mistakenly predicted to be equal to 0 in this model, which is wrong.)

Now we apply the model to the training data on which we performed oversampling. As it is known, the accuracy is reduced, but the values of the characterization and sensitivity factors are acceptable values. (that is, the values of net income that are equal to 1 in reality (income above the 90th percentile), are also predicted to be approximately equal to 1 in this model, as well as the values of net income that are actually equal to 0 (income below the 90th percentile), in this The model was also predicted to be almost equal to 0.)



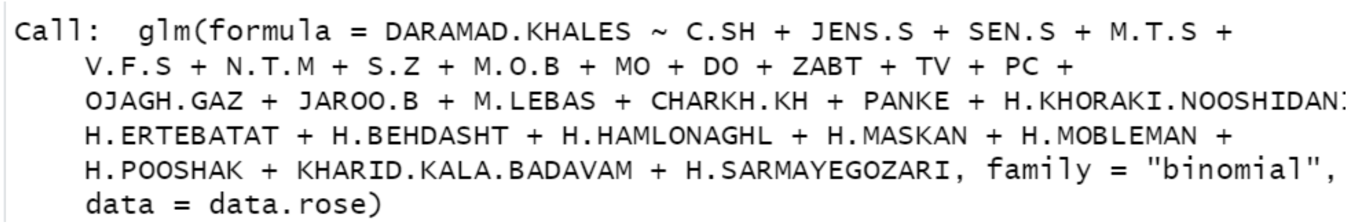
```
          Reference
Prediction   0   1
         0 154  19
         1 110  16
                                          
               Accuracy : 0.5686          
                 95% CI : (0.5103, 0.6254)
    No Information Rate : 0.8829          
    P-Value [Acc > NIR] : 1               
                                          
                  Kappa : 0.019           
                                          
 Mcnemar's Test P-Value : 2.299e-15       
                                          
            Sensitivity : 0.5833          
            Specificity : 0.4571          
         Pos Pred Value : 0.8902          
         Neg Pred Value : 0.1270          
             Prevalence : 0.8829          
         Detection Rate : 0.5151          
   Detection Prevalence : 0.5786          
      Balanced Accuracy : 0.5202          
                                          
       'Positive' Class : 0               
```

To achieve higher accuracy, we use different methods of forward, backward, or both, and we observe that the forward and both methods do not have higher accuracy, and only the backward method has higher accuracy.

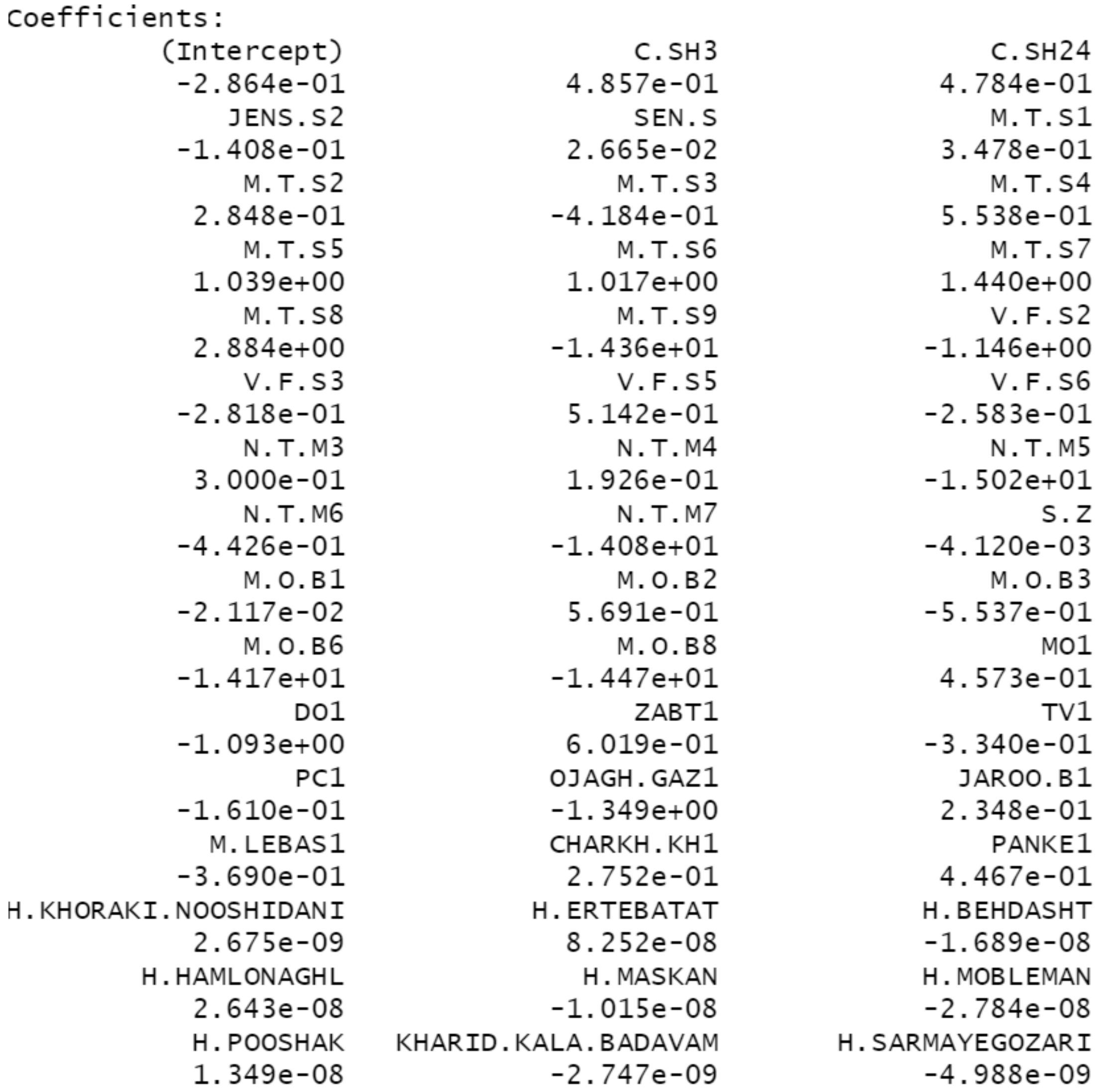


```
          Reference
Prediction   0   1
         0 158  19
         1 106  16

               Accuracy : 0.5819
                 95% CI : (0.5238, 0.6385)
    No Information Rate : 0.8829
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0268

 Mcnemar's Test P-Value : 1.448e-14

            Sensitivity : 0.5985
            Specificity : 0.4571
```

The variables that influence this model are:



```
Call:  glm(formula = DARAMAD.KHALES ~ C.SH + JENS.S + SEN.S + M.T.S +
    V.F.S + N.T.M + S.Z + M.O.B + MO + DO + ZABT + TV + PC +
    OJAGH.GAZ + JAROO.B + M.LEBAS + CHARKH.KH + PANKE + H.KHORAKI.NOOSHIDAN
    H.ERTEBATAT + H.BEHDASHT + H.HAMLONAGHL + H.MASKAN + H.MOBLEMAN +
    H.POOSHAK + KHARID.KALA.BADAVAM + H.SARMAYEGOZARI, family = "binomial",
    data = data.rose)
```

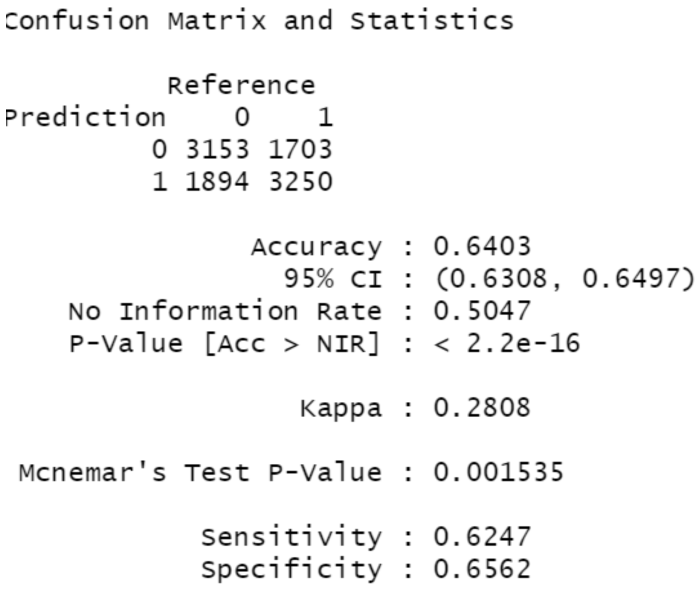


```
Coefficients:
          (Intercept)                 C.SH3                 C.SH24
           -2.864e-01             4.857e-01              4.784e-01
              JENS.S2                 SEN.S                 M.T.S1
           -1.408e-01             2.665e-02              3.478e-01
               M.T.S2                M.T.S3                 M.T.S4
            2.848e-01            -4.184e-01              5.538e-01
               M.T.S5                M.T.S6                 M.T.S7
            1.039e+00             1.017e+00              1.440e+00
               M.T.S8                M.T.S9                 V.F.S2
            2.884e+00            -1.436e+01             -1.146e+00
               V.F.S3                V.F.S5                 V.F.S6
           -2.818e-01             5.142e-01             -2.583e-01
               N.T.M3                N.T.M4                 N.T.M5
            3.000e-01             1.926e-01             -1.502e+01
               N.T.M6                N.T.M7                    S.Z
           -4.426e-01            -1.408e+01             -4.120e-03
               M.O.B1                M.O.B2                 M.O.B3
           -2.117e-02             5.691e-01             -5.537e-01
               M.O.B6                M.O.B8                    MO1
           -1.417e+01            -1.447e+01              4.573e-01
                  DO1                 ZABT1                    TV1
           -1.093e+00             6.019e-01             -3.340e-01
                  PC1            OJAGH.GAZ1               JAROO.B1
           -1.610e-01            -1.349e+00              2.348e-01
             M.LEBAS1            CHARKH.KH1                 PANKE1
           -3.690e-01             2.752e-01              4.467e-01
H.KHORAKI.NOOSHIDANI            H.ERTEBATAT             H.BEHDASHT
            2.675e-09             8.252e-08             -1.689e-08
          H.HAMLONAGHL              H.MASKAN             H.MOBLEMAN
            2.643e-08            -1.015e-08             -2.784e-08
             H.POOSHAK   KHARID.KALA.BADAVAM       H.SARMAYEGOZARI
            1.349e-08            -2.747e-09             -4.988e-09
```

Check more fit:

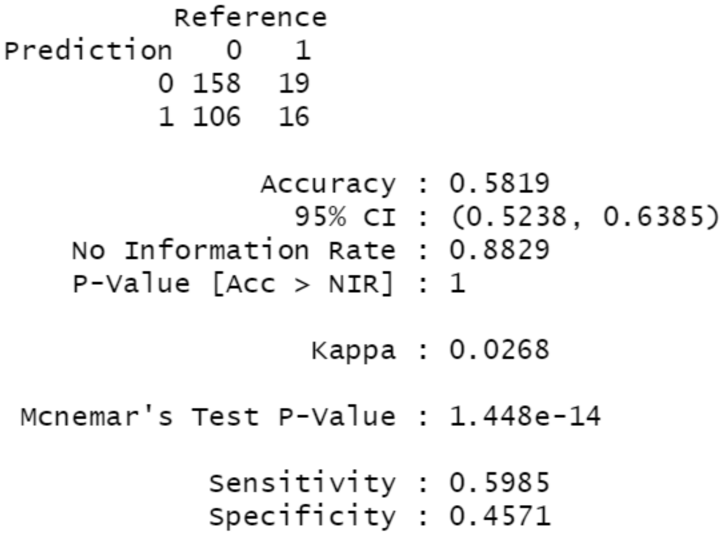
By applying the model to the training data, it can be seen that no overfitting has been done in this model.

Accuracy of the model on the training data:



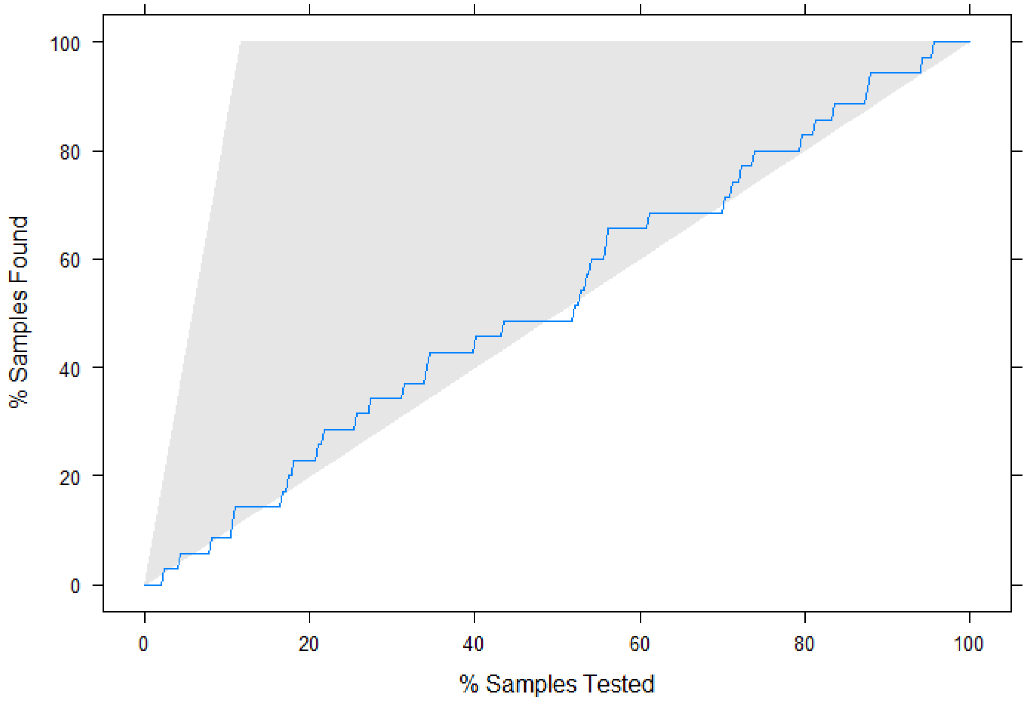
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3153 1703
         1 1894 3250

               Accuracy : 0.6403
                 95% CI : (0.6308, 0.6497)
    No Information Rate : 0.5047
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2808

 Mcnemar's Test P-Value : 0.001535

            Sensitivity : 0.6247
            Specificity : 0.6562
```

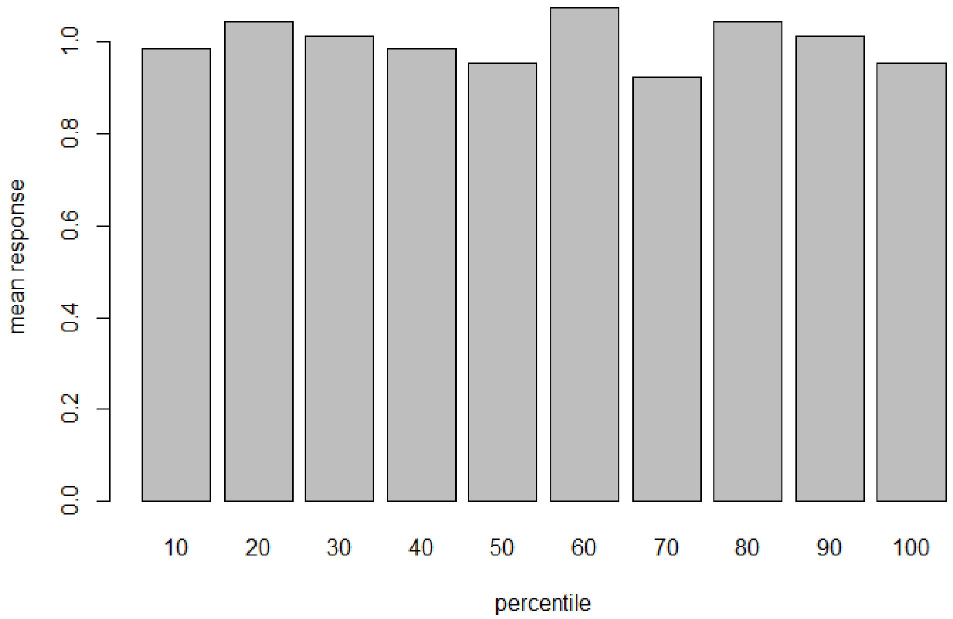
Accuracy of the model on validation data:

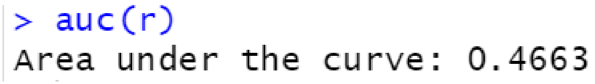


```
          Reference
Prediction   0   1
         0 158  19
         1 106  16

               Accuracy : 0.5819
                 95% CI : (0.5238, 0.6385)
    No Information Rate : 0.8829
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0268

 Mcnemar's Test P-Value : 1.448e-14

            Sensitivity : 0.5985
            Specificity : 0.4571
```

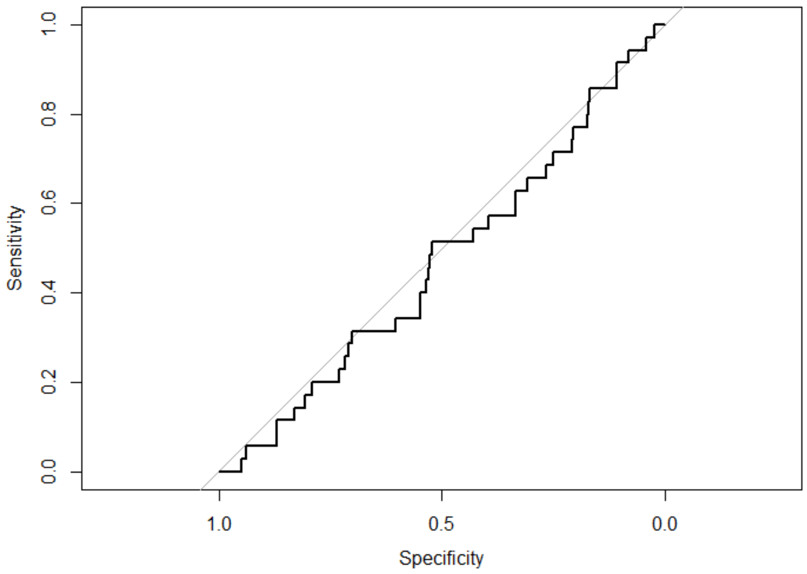
The lifting chart, the percentile lifting chart and the roc chart for the logistic regression model with regression removal show the acceptability of the model.


% Samples Found
100
80
60
40
20
0
0
20
40
60
80
100
% Samples Tested


mean response
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
70
80
90
100
percentile

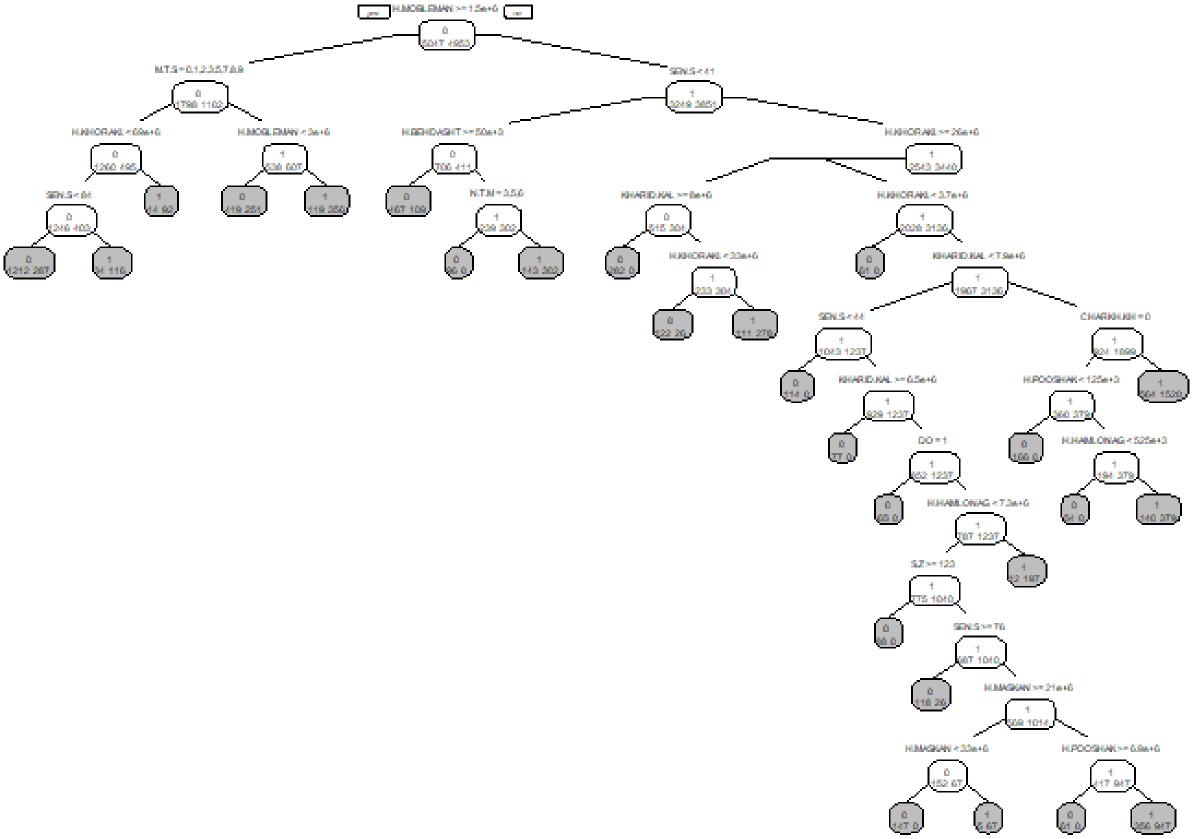


```
> auc(r)
Area under the curve: 0.4663
```



Chapter 4: Implementation of the decision tree model

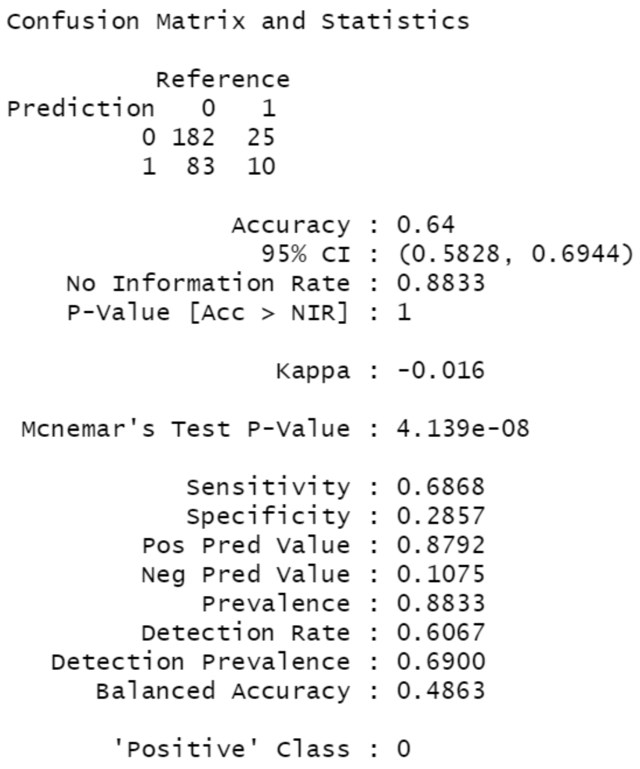
To select the decision tree model, we tried random forest, boosted tree and single tree models; Both the random forest and the enhanced tree had high accuracy, but their sensitivity and specificity indices did not have acceptable values. So we ran the single tree model as the final decision tree model on the data:



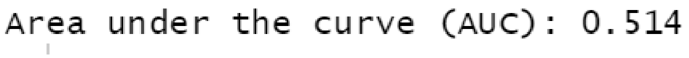
One of the above single tree decision rules is as follows:

H.MOBLEMAN >= 1492500 and **M.T.S** = 0,1,2,3,5,7,8,9 and **H. KHORAKI. NOOSHIDANI**< 6.8856e+07 **daramad. khales** = 1 (above the 90th percentile)

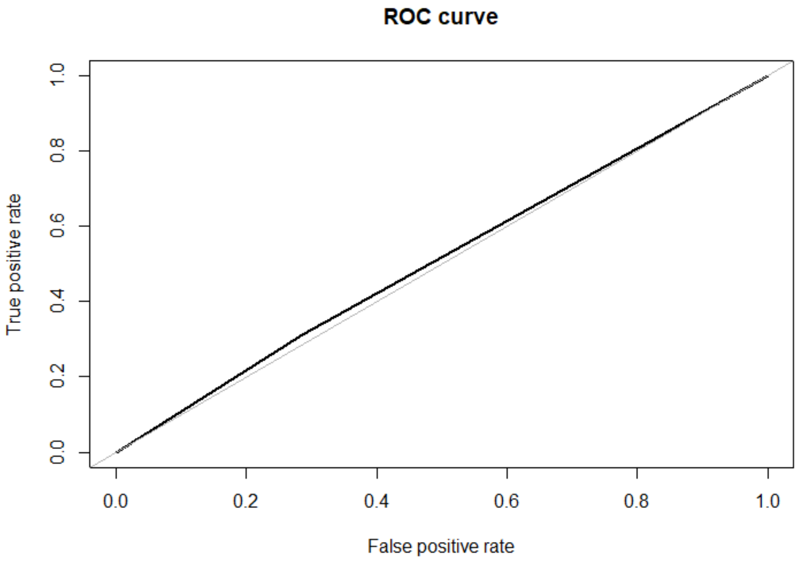
The accuracy of the single tree model and the roc curve are as follows:



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 182  25
         1  83  10
                                          
               Accuracy : 0.64            
                 95% CI : (0.5828, 0.6944)
    No Information Rate : 0.8833          
    P-Value [Acc > NIR] : 1               
                                          
                  Kappa : -0.016          
                                          
 Mcnemar's Test P-Value : 4.139e-08       
                                          
            Sensitivity : 0.6868          
            Specificity : 0.2857          
         Pos Pred Value : 0.8792          
         Neg Pred Value : 0.1075          
             Prevalence : 0.8833          
         Detection Rate : 0.6067          
   Detection Prevalence : 0.6900          
      Balanced Accuracy : 0.4863          
                                          
       'Positive' Class : 0               
```



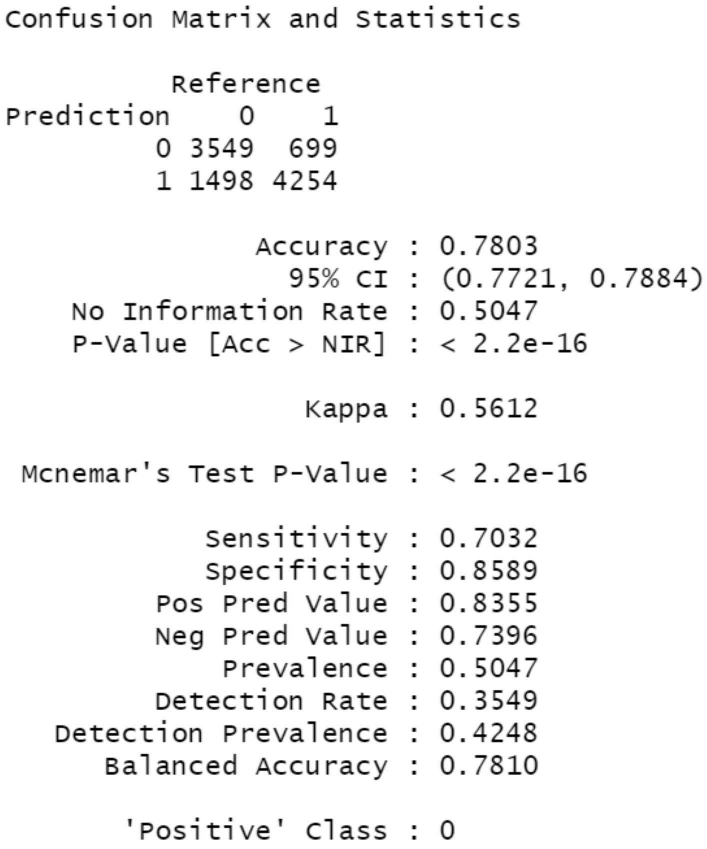
```
Area under the curve (AUC): 0.514
```



Check more fit:

By applying the model to the training data, it can be seen that no overfitting has been done in this model.

Accuracy of the model on the training data:



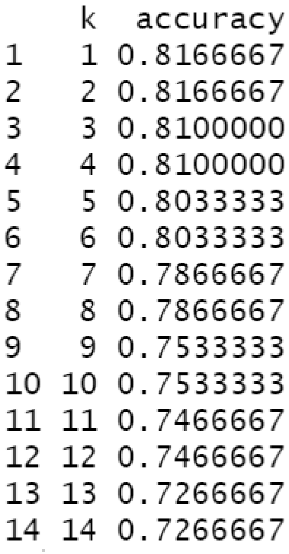
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3549  699
         1 1498 4254

               Accuracy : 0.7803
                 95% CI : (0.7721, 0.7884)
    No Information Rate : 0.5047
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5612

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7032
            Specificity : 0.8589
         Pos Pred Value : 0.8355
         Neg Pred Value : 0.7396
             Prevalence : 0.5047
         Detection Rate : 0.3549
   Detection Prevalence : 0.4248
      Balanced Accuracy : 0.7810

       'Positive' Class : 0
```

Chapter 5: Implementation of the KNN model

Before implementing the KNN model, because this model is based on distance calculation, the data must be pre-processed. That is, they must be numerical and also the normalization process must be done on them so that the output obtained from the model can be relied upon.

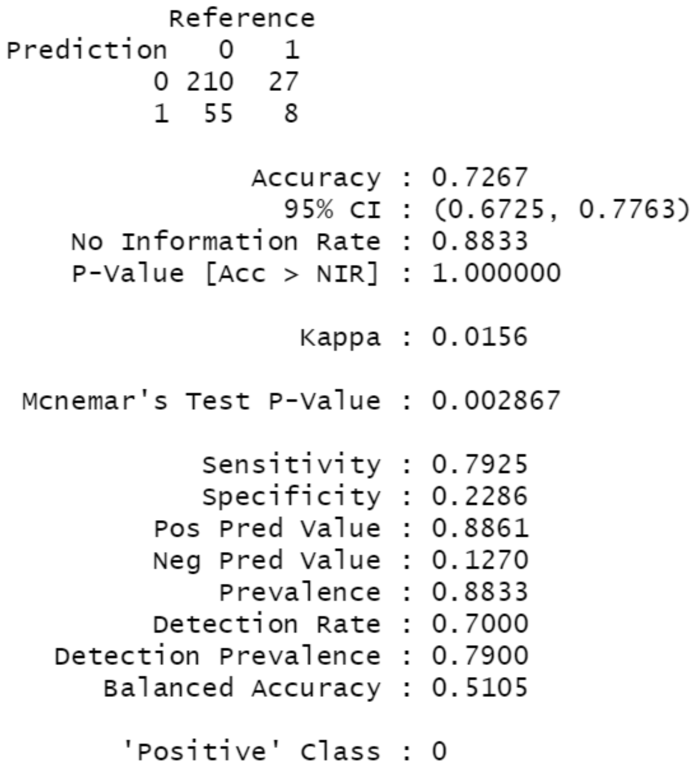
For this purpose, we first converted the classified variables into binary variables (dummy variable) and then normalized all the variables and finally applied the model to the data.

(Note: we applied the knn model only to the community of variables that were significant in the single tree model and logistic regression.)



```
    k  accuracy
1   1 0.8166667
2   2 0.8166667
3   3 0.8100000
4   4 0.8100000
5   5 0.8033333
6   6 0.8033333
7   7 0.7866667
8   8 0.7866667
9   9 0.7533333
10 10 0.7533333
11 11 0.7466667
12 12 0.7466667
13 13 0.7266667
14 14 0.7266667
```

According to the table above and considering the valuesIndicator sensitivity and specificity, it seems that k=14 is suitable for all three factors of accuracy, sensitivity and specificity.

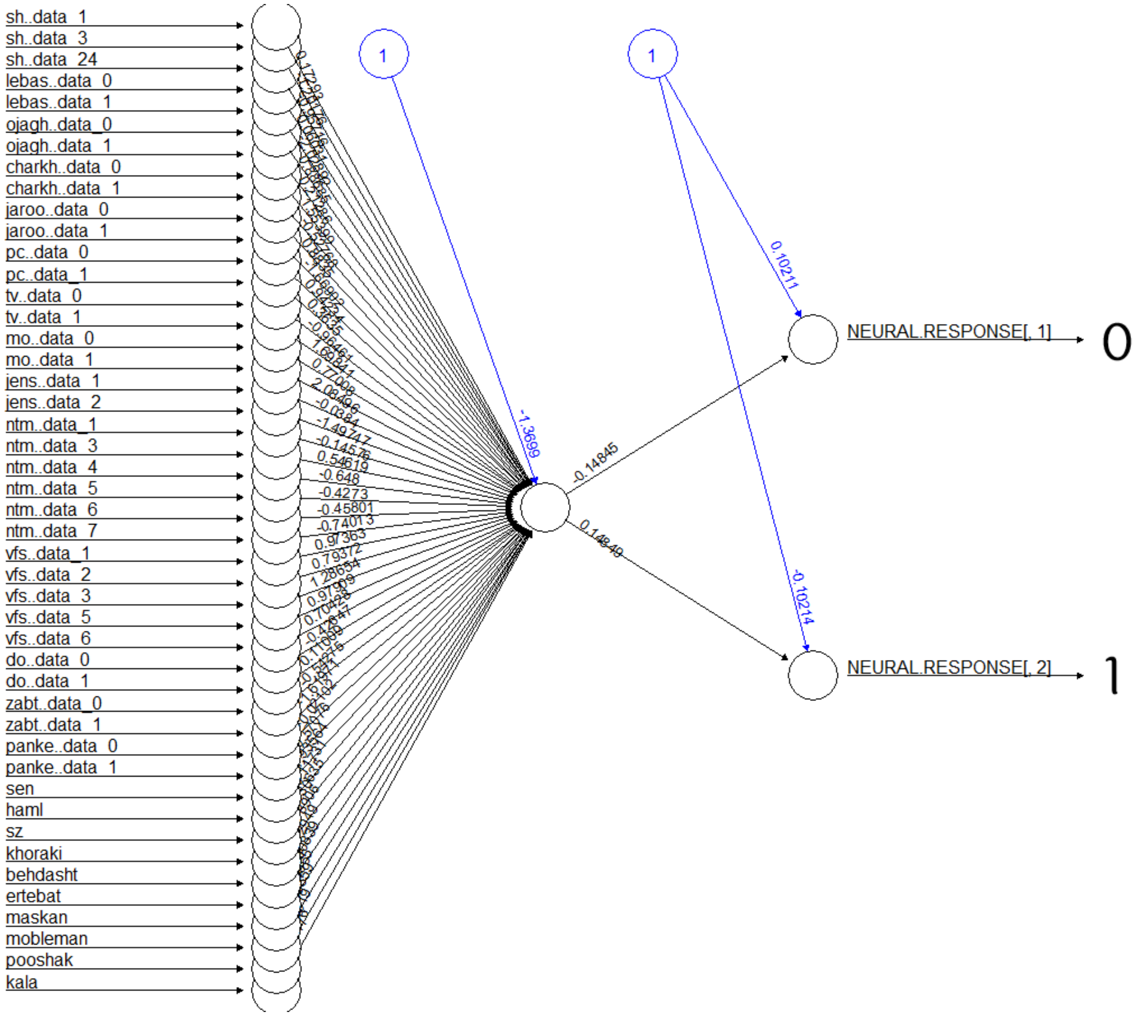


```
            Reference
Prediction    0   1
         0 210  27
         1  55   8

               Accuracy : 0.7267
                 95% CI : (0.6725, 0.7763)
    No Information Rate : 0.8833
    P-Value [Acc > NIR] : 1.000000

                  Kappa : 0.0156

 Mcnemar's Test P-Value : 0.002867

            Sensitivity : 0.7925
            Specificity : 0.2286
         Pos Pred Value : 0.8861
         Neg Pred Value : 0.1270
             Prevalence : 0.8833
         Detection Rate : 0.7000
   Detection Prevalence : 0.7900
      Balanced Accuracy : 0.5105

       'Positive' Class : 0
```

Chapter 6: Implementation of the neural network model

The output of the neural network model with a hidden layer is as follows:

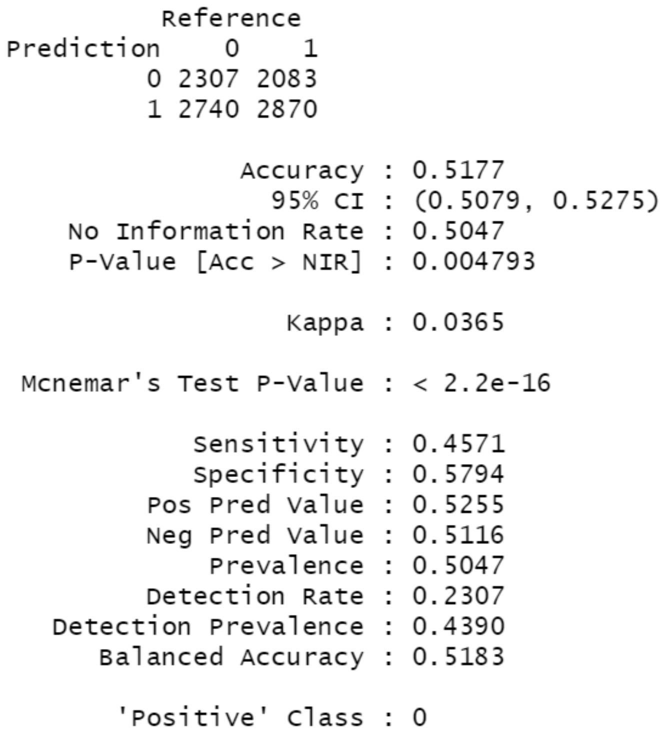
(Note: We applied the neural network model only to the community of variables that were significant in the single tree model and logistic regression.)



Check more fit:

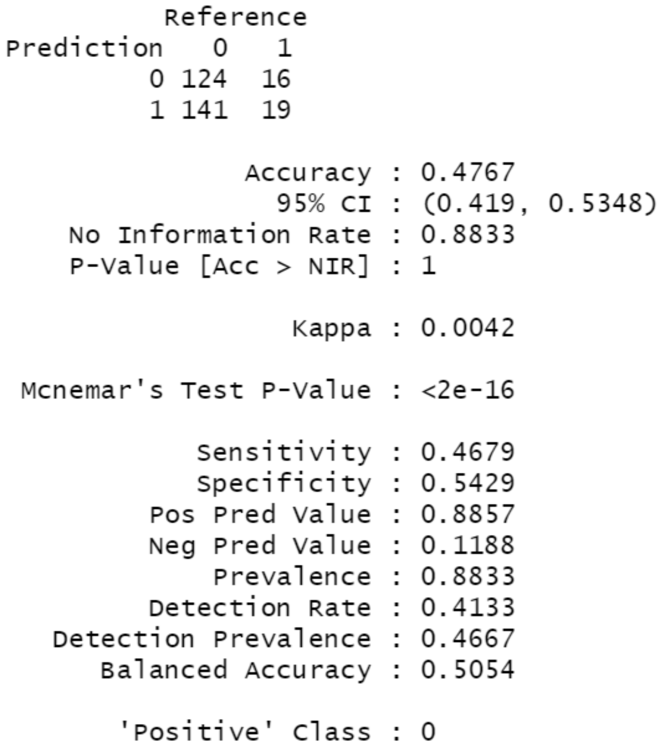
By applying the model to the training data, it can be seen that no overfitting has been done in this model.

Accuracy of the model on the training data:



```
          Reference
Prediction    0    1
         0 2307 2083
         1 2740 2870

               Accuracy : 0.5177
                 95% CI : (0.5079, 0.5275)
    No Information Rate : 0.5047
    P-Value [Acc > NIR] : 0.004793

                  Kappa : 0.0365

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4571
            Specificity : 0.5794
         Pos Pred Value : 0.5255
         Neg Pred Value : 0.5116
             Prevalence : 0.5047
         Detection Rate : 0.2307
   Detection Prevalence : 0.4390
      Balanced Accuracy : 0.5183

       'Positive' Class : 0
```

Accuracy of the model on validation data:



```
          Reference
Prediction   0   1
         0 124  16
         1 141  19

               Accuracy : 0.4767
                 95% CI : (0.419, 0.5348)
    No Information Rate : 0.8833
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0042

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.4679
            Specificity : 0.5429
         Pos Pred Value : 0.8857
         Neg Pred Value : 0.1188
             Prevalence : 0.8833
         Detection Rate : 0.4133
   Detection Prevalence : 0.4667
      Balanced Accuracy : 0.5054

       'Positive' Class : 0
```

The seventh chapter: summary and conclusion about the final model

Considering that our problem is classification, the factors of accuracy, sensitivity and specification are effective in choosing the final model.

precision	**Model**
0.58	**Logistic regression with regression elimination**
0.64	**Single decision tree**
0.72	**Knn**
0.47	**neural network**
allergy	**Model**
0.59	**Logistic regression with regression elimination**
0.68	**Single decision tree**
0.79	**Knn**
0.46	**neural network**
specification	Model
0.45	**Logistic regression with regression elimination**
0.28	**Single decision tree**
0.22	**Knn**
0.54	**neural network**

According to these indicators,Can concluded that the logistic regression model is the best model for this problem.