«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Лабораторная работа №5 по теме: «Вычисление дескриптивных статистик»

Выполнил

Студент 1 курса

группы 09-115(3)

Зиновьев Е. А.

Преподаватель:

Шустова Е.П

Казань 2021

**Цель работы:** получить описательные статистики с помощью функций summary, describe, describeBy и проанализировать получившиеся результаты.

**План работы**

1. Выбрать данные из репозитория данных и привести их краткое описание
2. Получить описательные характеристики для всех признаков данных с помощью функции summary и проанализировать получившиеся результаты
3. Получить описательные характеристики для одного из признаков и только для некоторых признаков данных с помощью функции summary и проанализировать получившиеся результаты
4. Вывести не все статистики сразу, а запросить их по отдельности
5. Получить описательные статистики с помощью функций describe
6. Получить описательные статистики с помощью функций bescribeBy
7. Разместить полный код программы

**Ход работы**

1. Были выбраны следующие данные из репозитория:

<http://archive.ics.uci.edu/ml/datasets/Census+Income>

**Описание данных:**

Это данные переписи населения по гражданам разных стран, в которых с помощью имеющихся социально-экономически-демографических признаков надо определить, как они влияют на доход населения – дискретному бинарному параметру (больше или меньше 50 000 долларов в год). Полученные выводы помогли бы понять, как можно было бы улучшить состояние граждан своей страны.

Описание признаков (название: категории или тип параметра – непрерывный):

age: возраст (непрерывный).

workclass: частный, ЗАО, ООО, федеральное правительство, местное правительство, государственное правительство, без оплаты, никогда не работал.

fnlwgt: финальный вес (означает, какая количество человек с данными характеристиками), непрерывный.

образование: бакалавриат, колледж, 11 классов школы, законченная школа, профессиональное образование, незаконченное высшее, законченное высшее, 9 классов школы, 7-8 классов школы, 12 классов школы, магистратура, 1-4 классов школы, 10 классов школы, аспирантура, 5-6 классов школы, дошкольное образование.

По возрастанию степени образования: дошкольное образование. < 1-4 классов школы < 5-6 классов школы < 7-8 классов школы < 9 классов школы < 10 классов школы < 11 классов школы < 12 классов школы < законченная школа < профессиональное образование < незаконченное высшее < законченное высшее < колледж < бакалавриат < магистратура < аспирантура.

education-num: количество степеней образования (классы школы + степени высшего образования), непрерывное.

семейное положение: женат и супруг при этом - гражданский, разведен, не был в браке, разлучен, вдовец, лицо, состоящее в браке, которое не проживало в одном домохозяйстве., женат, но супруг находится на военной службе.

род занятий: Техподдержка, Ремесло-ремонт, Другое-обслуживание, Продажи, Управленческий-менеджер, Проф-специальность, Уборщики, Машиностроение, Адм.канцелярия, Сельское хозяйство-рыболовство, Транспорт, Обслуживание частных домов, Охранное предприятие, Вооруженные силы.

Отношения: жена, собственный ребенок, муж, не в семье, другой родственник, не состоящий в браке.

раса: белая, азиатско-тихоокеанская, американская-индийская-эскимосская, другая, черная.

пол: Женский, Мужской.

прирост капитала: в долларах, непрерывный.

убыток капитала: в долларах, непрерывный.

часов работы в неделю: непрерывно.

родная страна: США, Камбоджа, Англия, Пуэрто-Рико, Канада, Германия, окраина США (Гуам-ВИ-США и т. д.), Индия, Япония, Греция, Юг, Китай, Куба, Иран, Гондурас, Филиппины, Италия , Польша, Ямайка, Вьетнам, Мексика, Португалия, Ирландия, Франция, Доминиканская Республика, Лаос, Эквадор, Тайвань, Гаити, Колумбия, Венгрия, Гватемала, Никарагуа, Шотландия, Таиланд, Югославия, Сальвадор, Тринадад и Тобаго, Перу, Гонконг , Голландия-Нидерланды.

Заработок: > 50 тысяч долларов в год, <= 50 тысяч долларов в год.

Для справки:

**Attribute Information :**

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

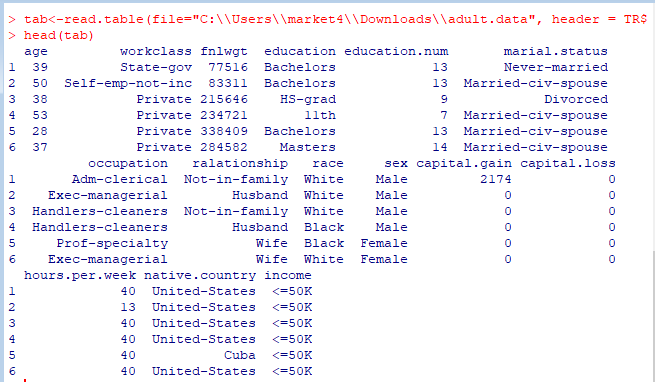
capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

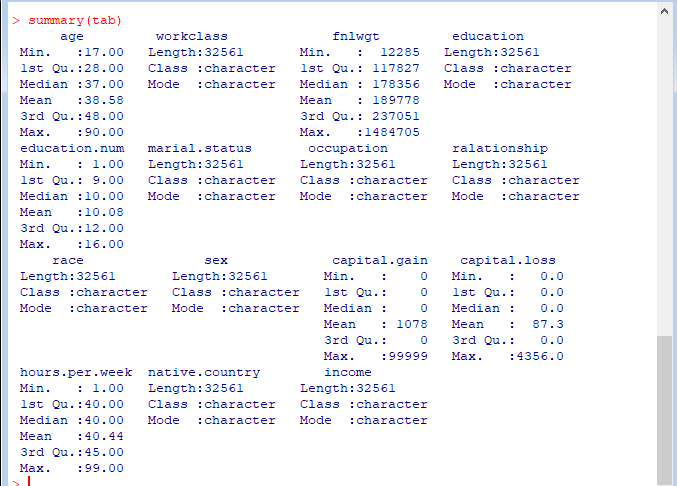
income: > 50 KB, <= 50 KB.

Загрузим датасет и посмотрим на него:

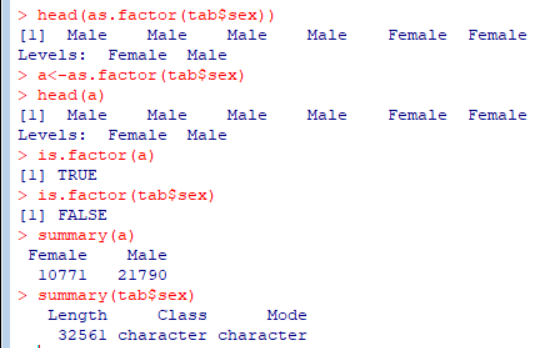


1. Для вывода описательных статистик в R есть специальная функция summary():

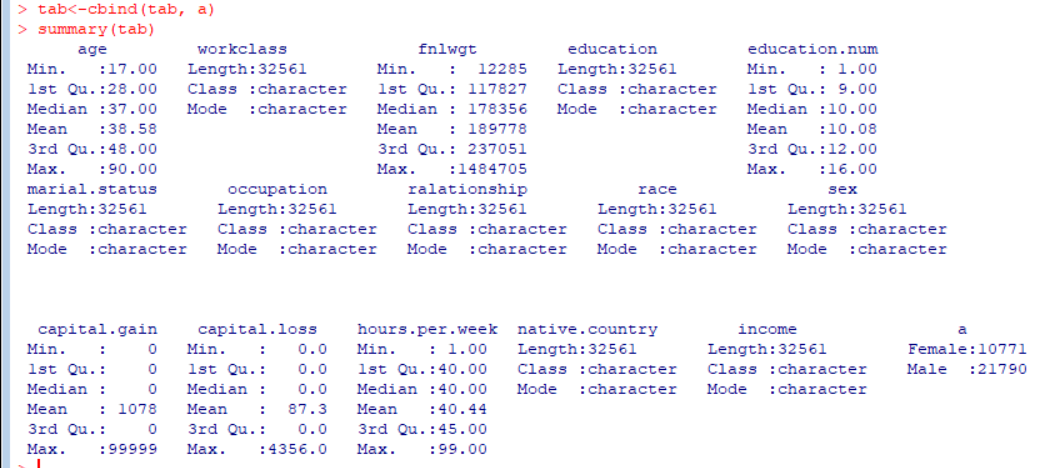
Как видно из нижележащего рисунка, была отражена полная таблица со всеми описательными статистиками. Для числовых параметров: минимум, 1-ый квартиль, медиана, среднее, 3-ий квартиль, максимум. Для категориальных только количество в разбивке уровням. Но возникла одна проблема, при которой для категориальных значения не считаются по уровням, а просто выводится тип character. Для ее решения нужно преобразовать все эти переменные в факторные и к ним уже применить эту функцию summary:



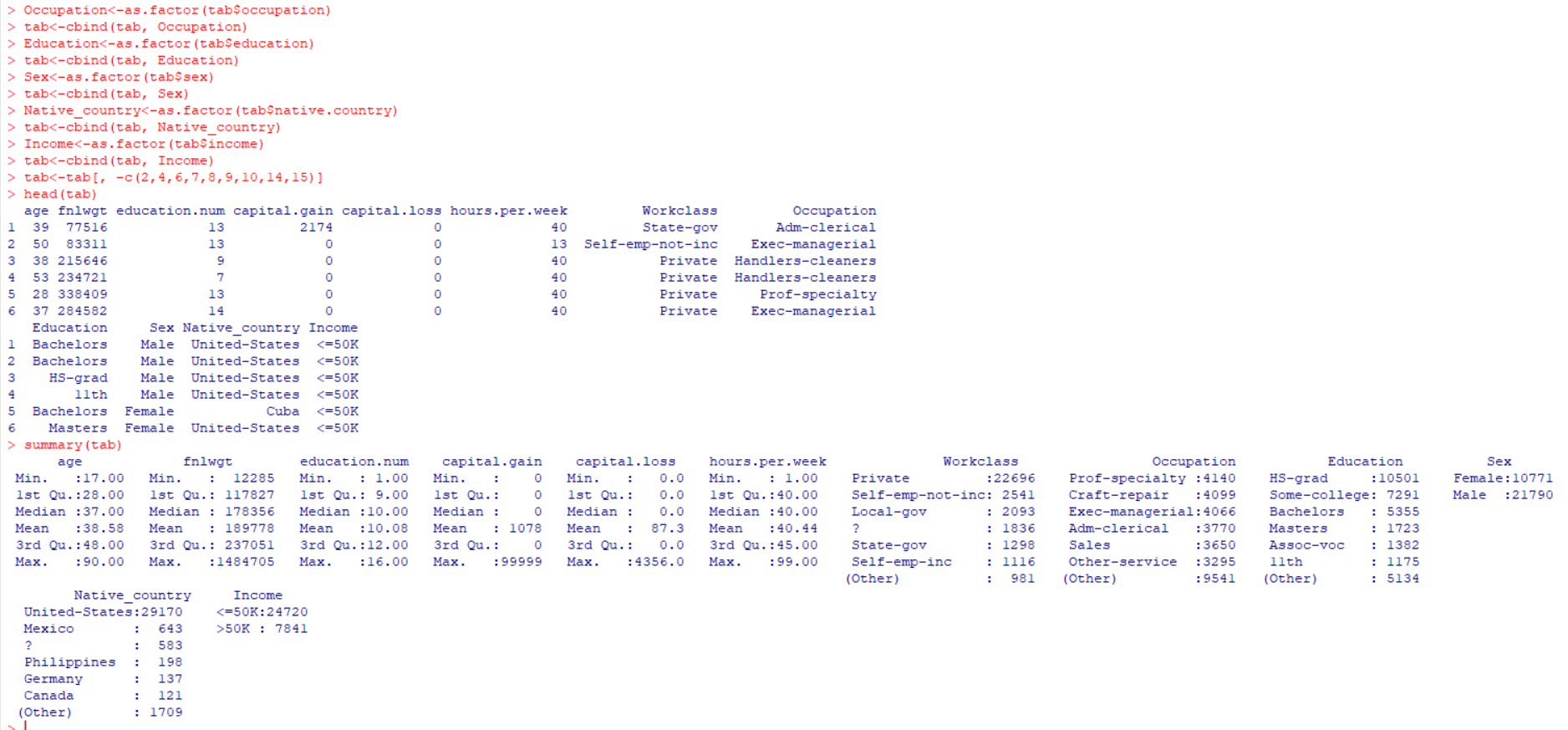
Но как увидим, значения в датафрейме inplace (на месте) не переносятся в факторные, поэтому приходится создавать новые переменные, то есть при той же операции summary(a) выводит разбивку по категориям, а summary(tab$sex) просто показывает тип параметра character:



Поэтому добавим новый столбец а и удалим старый, заменив на его старое название, вот так это выглядит, когда оба столбца в датафрейме: а и sex



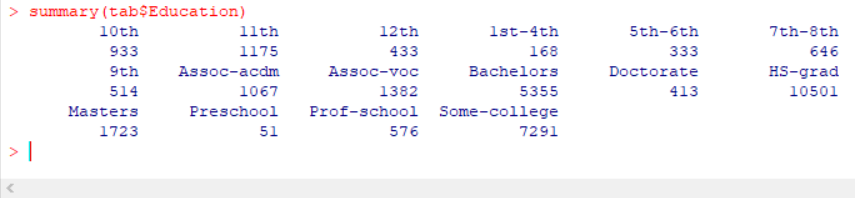
Так поступим с каждым столбцом:



Теперь видно, как поделились категориальные данные по уровням. Например, по уровню дохода в 3 раза тех, кто получает меньше 50000 долларов в год или в данной выборке мужчин вдвое больше, или почти основная масса населения данной выборки – представители США.

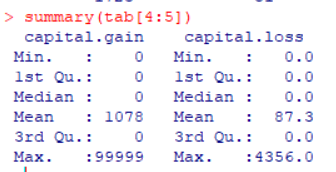
По числовым параметрам можно узнать гораздо больше. Так, минимальный возраст участника переписи составил – 17 лет, а максимальный - 90, медиана очень близка к среднеарифметическому и составила примерно средний возраст жизни человека: 37-38 лет, причем видно, что 90 лет находится довольно далеко от третьего квартиля, который составил 48 лет.

1. Функцию summary необязательно применять ко всему датафрейму, можно применить к одному признаку, например, к уровню образования:

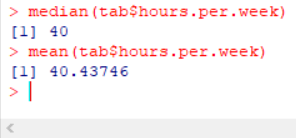


Как видим, что гораздо больше остальных представлено выпускников только школы (10501 человек), в 1.5 раза меньше выпускников колледжа (7291 человек) и вдвое меньше бакалавров (5355 человек) и самое меньшее, 51 человек вообще никогда не учились – ни одного класса в школе.

Для отбора нескольких столбцов надо в скобках перечислить их номера. Данное представление удобнее, когда нужно сравнить 2 параметра между собой. Так, например, на картинке ниже, сравниваются описательные характеристики дохода и убытка населения. Видно, что данных не очень много в исходной выборке, но по ним можно сказать, что доходы у одних сильно превысили потерь капитала у других.

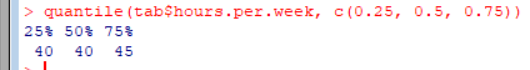


1. Можно не выводить все характеристики сразу, а запросить их по отдельности с помощью вызова конкретных описательных статистик. Допустим, хотим сравнить медиану со средним количеством отработанных населением часов:



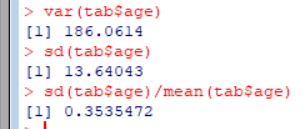
Как видим, значения практически совпадает, что говорит о том, что большинство работает на классической 8-часовой рабочей смене при 5 рабочих днях.

Или хотим посмотреть, какие значения попадают в каждый квартиль, чтобы исследовать, как меняется переменная:



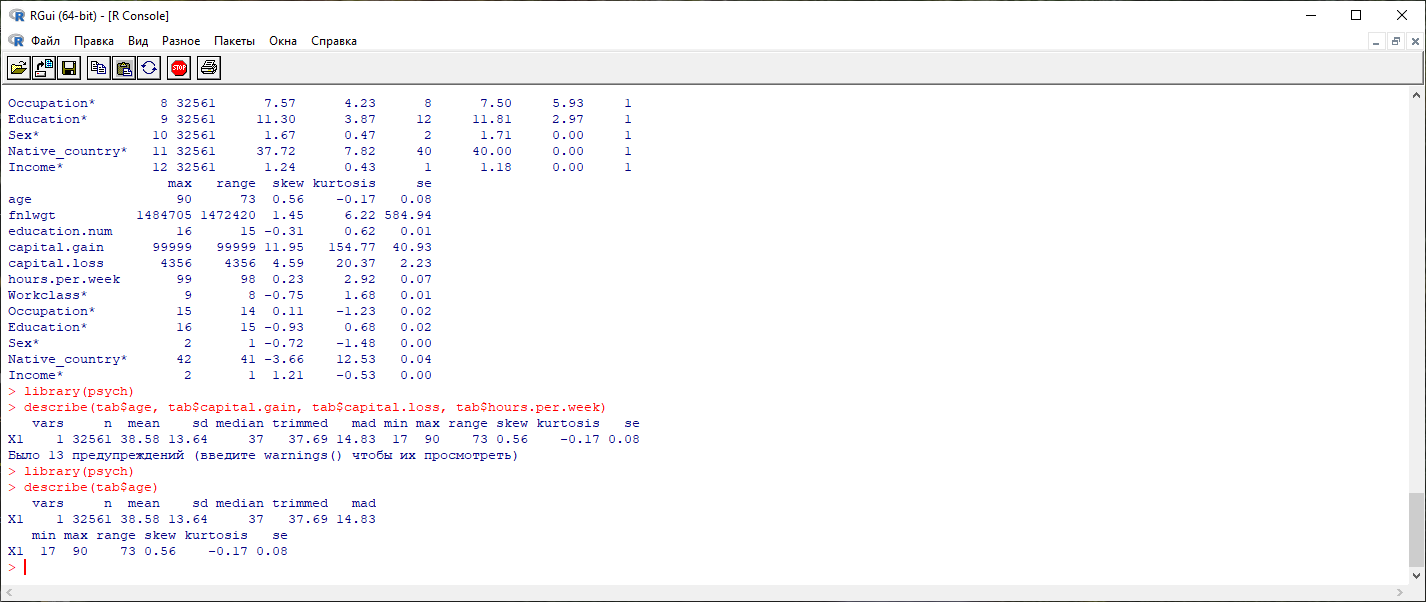
То есть 40-часовая рабочая неделя слабо изменяется в пределах разных квартилей.

Посмотрим на другие характеристики, которые покажут рассеяния данных:



Видно, что дисперсия составила – 186, а в нашей исходной единице измерения – 13.6 лет (стандартное отклонение), которое является корнем от дисперсии и коэффициент вариации – 0.35. То есть в среднем данные по возрасту меняются на 13.6 лет от среднего значения, а их вариативность составила 0.35, что говорит о довольно большой изменчивости данных. При таком значении коэффициента вариативности, например, в анализе торговых категорий такая бы категория попала бы в рискованную группу Z (по XYZ анализу продаж).

1. Для отображения описательных статистик можно использовать другую функцию describe(), для этого установим библиотеку psych, которая используется для психометрических исследований и посмотрим, какие у неё есть возможности, связанные с описанием данных.



Расшифровка выходных параметров функции describe:

* vars: порядковый номер признака
* n: число наблюдений;
* mean: среднее арифметическое, выборочное среднее;
* sd: стандартное отклонение;
* median: медиана;
* trimmed: усечённое среднее, среднее по цензурированной выборке;
* mad: медианное значение абсолютного отклонения от медианы;
* min, max: минимальное и максимальное значение;
* range: размах;
* skew: коэффициент асимметрии или скошенности ;
* kurtosis: коэффициент эксцесса;
* se: стандартная ошибка среднего;

От ранее рассмотренной функции summary эта функция отличается наличием:

sd ( стандартное отклонение, но уже была рассмотрена);

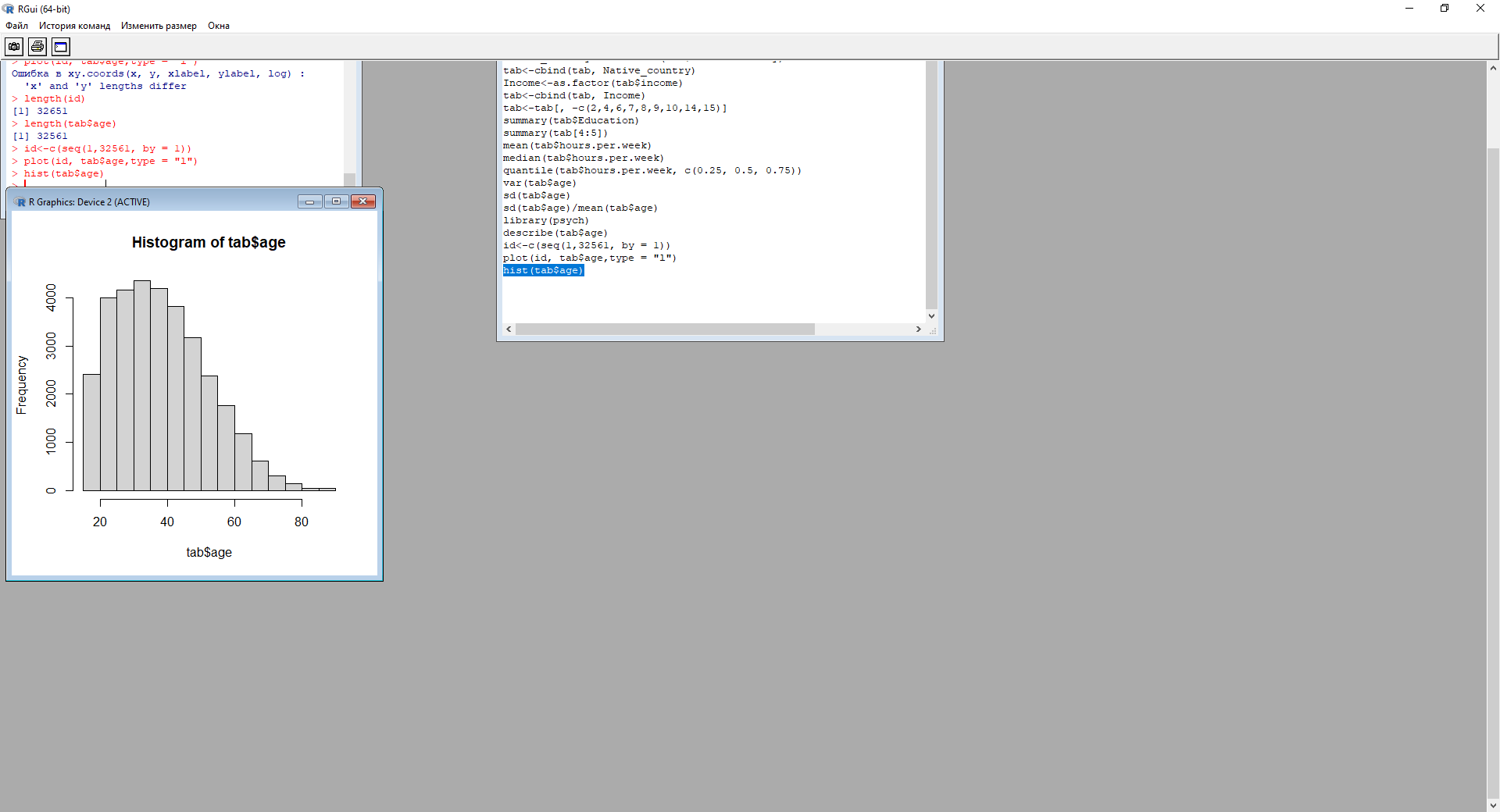
trimmed (усеченное среднее) – величина, показывающая изменение среднего при удалении х% значений выборки спереди и сзади. Уменьшив с обоих концов нашу выборку, видно, что среднее немного уменьшилось, что может говорить о наличии завышенных значений, таких как 90 лет (max);

mad - медианное значение абсолютного отклонения от медианы. Используется для оценки среднеквадратического отклонения, но с помощью медианы, а не среднего. По сравнению со среднеквадратическим отклонением, mad немного выше (14.83 и 13.64), но и это понятно, так как медиана меньше среднего (38.58 и 37), а выборка, как было видно по параметру trimmed, немного скошена вправо;

range - размах, то есть насколько широка выборка, просто максимальное значение – минимальное (90-17=73);

skew - коэффициент асимметрии, показывающий скошенность выбоки по горизонтали. Изменяется от -3 до 3, при котором 0 – симметричная выборка, значения меньше 0 – скошено влево, больше 0 – вправо. По нашей таблице оно равно 0.56, что доказывает предыдущие выводы при рассмотрении параметров mad и trimmed о том, что выборка немного скошена направо.

Kurtosis - коэффициент эксцесса, показывающий скошенность выборки по вертикали. Также изменяется от -3 до 3, при этом 0 – симметричная выборка, значения меньше 0 – более пологая кривая, больше 0 – кривая с выраженным пиком. По нашей таблице оно равно -0.17, что говорит о слегка пологой кривой. Посмотрим, как это выглядит:

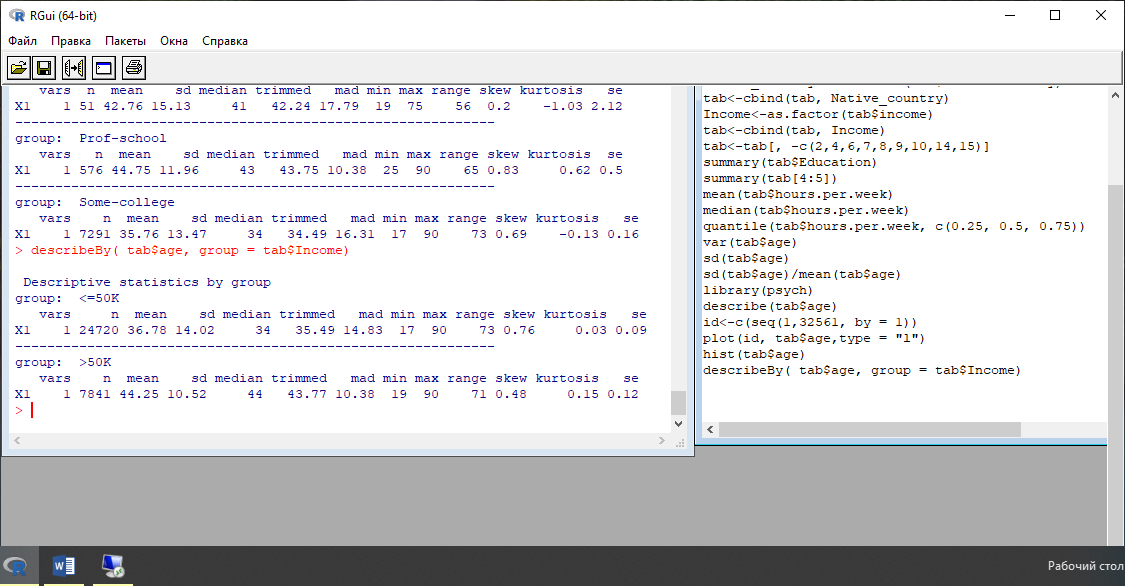


По гистограмме виден хвост справа и немного более пологое размещение кривой относительно нормального распределения.

Se - стандартная ошибка среднего (теоретическое стандартное отклонение всех средних выборки размера, извлекаемое из совокупности). По значению из нашей таблицы 0.08 можно сказать, что среднее генеральной совокупности практически совпадает со средним данной выборки.

1. Улучшение функции summary - функция describeBy(), которая позволяет выводить описательные статистики по группам. Нет необходимости отфильтровывать нужные строки и сохранять их в отдельные датасеты, можно просто указать группирующую переменную. Например, так можно просмотреть, как меняется описательные характеристики возраста для 2 уровней дохода: больше и меньше 50 000 долларов:

.



Как видим, при переходе границы дохода в 50000 долларов описательные характеристики среднего уровня увеличиваются на 8-10 лет (mean: 36 и 44, median: 34 и 44 года , trimmed: 36 и 44 года), а меры рассеяния в 1.5 раза уменьшаются (sd: 14 и 10.5, mad: 15 и 10), то есть возраст большего количества людей у более богатого населения лежит ближе к среднему значению. Об этом же можно убедиться, посмотрев на параметр skew (коэффициент асимметрии), он у менее богатого населения выше (0.76 и 0.48), то есть гистограмма возраста людей с меньшим доходом более сильно скошена направо.

Таким образом, были рассмотрены 2 функции: summary и describe (describeBy). С помощью summary можно получить следующие описательные характеристики: минимум, 1-ый квартиль, медиана, среднее, 3-ий квартиль, максимум, а для категориальных только количество в разбивке уровням. Для функции describe можно получить также вышеназванные характеристики (кроме квартилей) вместе с добавлением стандартного отклонения, усеченного среднего, коэффициента асимметрии, эксцесса, размаха и медианного значения абсолютного отклонения.

Такие функции можно использовать, как ко всем признакам сразу, так и по некоторым или даже одному, при необходимости также можно использовать только некоторые описательные характеристики или использовать с группировкой по категориальной переменной с помощью функции describeBy.

Общий вывод, который можно было сделать по этим данным с помощью вышесказанных функций, - это то, что более богатые люди – имеют более высокую степень образования, старше, причем возраст слабее колеблется возле среднего относительно более бедных, с преобладанием мужчин и граждан развитых стран, таких как США.

Код программы:

tab<-read.table(file="C:\\Users\\market4\\Downloads\\adult.data", header = TRUE, sep=",")

head(tab)

summary(tab)

Workclass<-as.factor(tab$workclass)

tab<-cbind(tab, Workclass)

Occupation<-as.factor(tab$occupation)

tab<-cbind(tab, Occupation)

Education<-as.factor(tab$education)

tab<-cbind(tab, Education)

Sex<-as.factor(tab$sex)

tab<-cbind(tab, Sex)

Native\_country<-as.factor(tab$native.country)

tab<-cbind(tab, Native\_country)

Income<-as.factor(tab$income)

tab<-cbind(tab, Income)

tab<-tab[, -c(2,4,6,7,8,9,10,14,15)]

summary(tab$Education)

summary(tab[4:5])

mean(tab$hours.per.week)

median(tab$hours.per.week)

quantile(tab$hours.per.week, c(0.25, 0.5, 0.75))

var(tab$age)

sd(tab$age)

sd(tab$age)/mean(tab$age)

library(psych)

describe(tab$age)

id<-c(seq(1,32561, by = 1))

plot(id, tab$age,type = "l")

hist(tab$age)

describeBy( tab$age, group = tab$Income)