**Содержание**

**1.**

# Введение

В последнее время информация, растущая в колоссальных объёмах, рождает потребность в обработке больших объёмов данных. В этом направлении большое место отведено интеллектуальному анализу данных. Это направление включает в себя методы, отличные от классического анализа, основанные на моделировании, вероятностных, и решающие задачи обобщения, ассоциирования и отыскания закономерностей. В большой степени развитию этой дисциплины способствовало проникновение в сферу анализа данных идей, возникших в теории искусственного интеллекта.

В данной работе я хотел бы рассмотреть частную задачу подобного анализа, а именно задачу классификации

1

## Задача 1

Необходимо загрузить данные из указанного набора и произвести следующие действия.

1. Нормализовать данные, вычтя из каждого столбца среднее значение mean(x) и поделив на среднеквадратическое отклонение σ ~ sqrt(var(x)), где x – столбец данных.
2. Проверить, что в наборе данных нет линейной зависимости (построить зависимости между переменными, указанными в варианте, и проверить, что R2 в каждой из них не высокий). В случае, если R2 большой, один из таких столбцов можно исключить из рассмотрения.
3. Построить линейную модель зависимой переменной от указанных в варианте регрессоров по методу наименьших квадратов (команда lm пакета lmtest в языке R). Оценить, насколько хороша модель, согласно: 1) R2, 2) p-значениям каждого коэффициента.
4. Ввести в модель логарифмы регрессоров. Сравнить модели и выбрать наилучшую.
5. Ввести в модель всевозможные произведения из пар регрессоров, в том числе квадраты регрессоров. Найдите одну или несколько наилучших моделей по доле объяснённого разброса в данных R2.

Набор данных – Swiss.

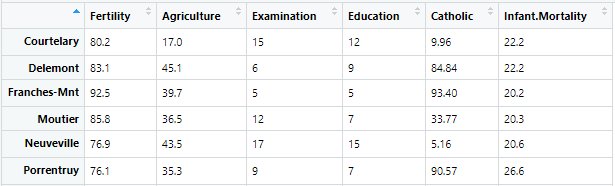
Объясняемая переменная – Fertility ().

Регрессоры (объясняющие переменные): Agriculture (%Мужчин в сельхозе), Catholic (%Католиков) и Infant.Mortality(%Смертность среди детей до 1 года)

2

## Решение Задача 1

В переменную data2 загружены данные из набора данных Swiss. На *рисунке 1* показана часть таблицы из набора данных.



*Рисунок 1. Набора данных в таблице data.*

Из данного набора данных в новую переменную data при помощи функции select выбираю столбцы из data2: Fertility, Agriculture, Catholic, Infant.Mortality. В дальнейшем буду работать с данной таблицей. Далее мне необходимо нормализовать регрессоры (, Agriculture, Catholic, Infant.Mortality), но перед этим логарифмирую их. Возьму логарифмы данных из столбцов, Agriculture, Catholic, Infant.Mortality. Полученные данные помещаю в созданные столбцы Log\_Agr, Log\_Cat и Log\_IM



***data["Log\_Agriculture»] =log(data$Agriculture)***

***data["Log\_Catholic»] =log(data$Catholic)***

***data["Log\_Infant.Mortality»] =log(data$Infant.Mortality)***

“Вычтя из каждого столбца среднее значение mean(x) и поделив на среднеквадратическое отклонение σ ~ sqrt(var(x)), где x – столбец данных. “

Нормализация данных(пример A*griculture)*

***Agr1<-as.character(data$Agriculture)***

***Agr2<-lapply(Agr1,as.integer)***

***Agr3<-as.numeric(unlist(Agr2))***

***data["Agr"]=(Agr3-mean(Agr3))/sqrt(var(Agr3))***

Также логарифмированы и нормализированы Cat и IM и нормализованы все регрессоры и их логарифмы

Создадим первую модель взяв в нее все самое лучшее, то есть Agr, IM и Cat. При помощи функции summary определяю, как регрессоры влияют на Fertility. Показатель R2 равен 0.3449. При это коэффициент p – () у Agr, (\*) у Cat и (\*\*) у IM



***model1<-lm(Fertility~Agr+Cat+IM,data)***

Уберем из модели Agr так как у него наибольший коэффициент P

***model2<-lm(Fertility~Cat+IM,data)***

***model2***

***summary(model2) #R^2=0.302***

***#p-(\*\*)(\*\*)***

уменьшился => первая модель лучше

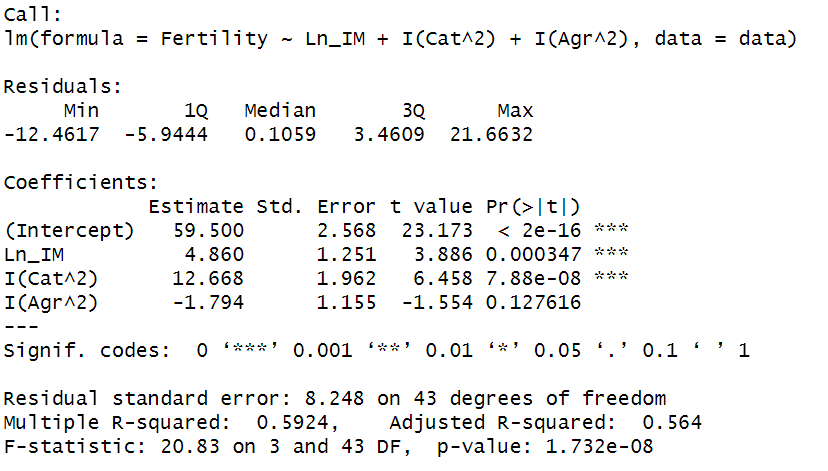
Далее в разделе «Код» (как у Петцольда только просто код)

Представлена часть моделей, которые я опробовал (какую-то часть я не записал так как сложно печатать много моделей, когда можно корректировать и «Играться» с одной пока не выйдет что-то интересное. Лучший результат, полученный мной за 20-25 моделей представлен в модели под номером 12.

***model12<-lm (Fertility~Ln\_IM+I(Cat^2)+I(Agr^2),data)#Best***

***model12***

***summary(model12) #R^2=0.564***



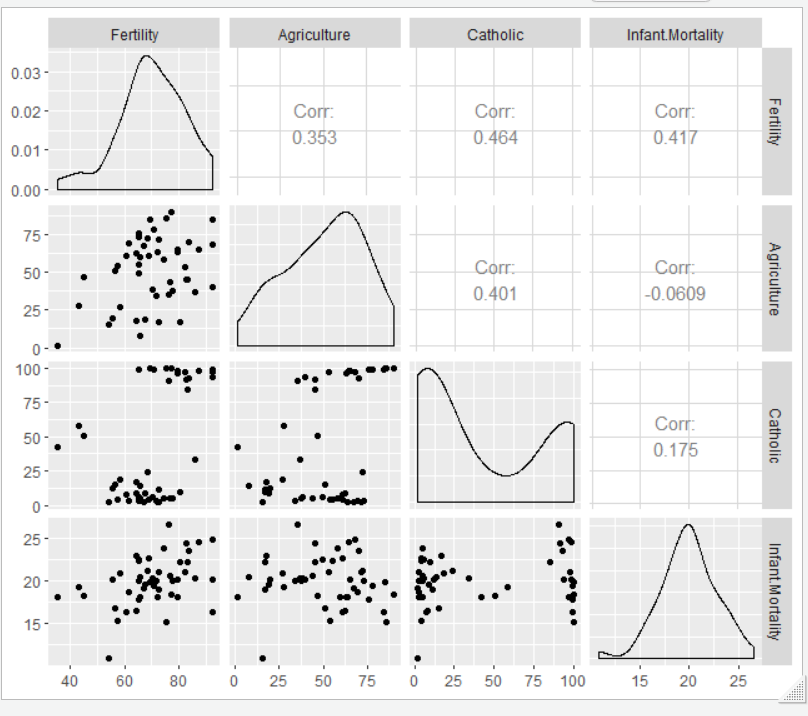


Рисунок 8. ggpairs

Так же при помощи функции “ggpairs(data)” можно получить графики зависимостей из базы данных, это тоже может помочь при анализе данных зачёт большей наглядности

Выполняя эту задачу по анализу данных, я пришел в следующим выводам:

1. Чем выше смертность среди детей до года, тем выше рождаемость (редко, но метко)
2. Плодятся либо только католики, либо только не католики (Сильное влияние)
3. Рождаемость не очень сильно зависима от агрокультуры (наиболее высокая рождаемость при уровне развития агрокультуры чуть выше среднего)

## Задача 2

В этой задаче необходимо проанализировать данные волны мониторинга экономического положения и здоровья населения РФ (данные обследования РМЭЗ НИУ ВШЭ).

В случае, если не удаётся прочесть соответствующий файл, обратитесь к материалу, выложенному вместе с заданием.

Прочитайте данные, выберите столбцы, которые Вам кажутся необходимыми, чтобы описать социально-экономическое положение граждан Российской Федерации.

Минимальный набор параметров: зарплата, пол, семейное положение, наличие высшего образования, возраст, тип населенного пункта, длительность рабочей недели.

Из параметра, отвечающего семейному положению, сделать дамми-переменные (с помощью one-hot-encoding): 1) переменная wed1 имеет значение 1 в случае, если респондент женат, 0 – в противном случае; 2) wed2=1, если респондент разведён или вдовец; 3) wed3 = 1, если респондент никогда не состоял в браке; 4) если считаете необходимым, введите другие параметры. Следите за мультиколлинеарностью (убедитесь в её отсутствии, оценив вспомогательную регрессию любого параметра (например, зарплаты или одного из параметров wed) на эти переменные и использовав команду VIF для неё).

Из параметра пол сделайте переменную sex, имеющую значение 1 для мужчин и равную 0 для женщин.

Из параметра, отвечающего типу населённого пункта, создайте одну дамми-переменную city\_status со значением 1 для города или областного центра, 0 – в противоположном случае.

Введите один параметр higher\_educ, характеризующий наличие полного высшего образования.

Факторные переменные, «имеющие много значений», такие как: зарплата, длительность рабочей недели и возраст, - необходимо преобразовать в вещественные переменные и нормализовать их: вычесть среднее значение по этой переменной, разделить её значения на стандартное отклонение.

1. Постройте линейную регрессию зарплаты на все параметры, которые Вы выделили из данных мониторинга. Не забудьте оценить коэффициент вздутия дисперсии VIF.

2. Поэкспериментируйте с функциями вещественных параметров: используйте логарифм и степени (хотя бы от 0.1 до 2 с шагом 0.1).

3. Выделите наилучшие модели из построенных: по значимости параметров, включённых в зависимости, и по объяснённому с помощью построенных зависимостей разбросу adjusted R2 - R 2 adj.

4. Сделайте вывод о том, какие индивиды получают наибольшую зарплату.

5. Оцените регрессии для подмножества индивидов, указанных в варианте

## Задача 2

1 Обработка данных

В переменную data загружены данные волны № 19 мониторинга экономического положения и здоровья населения РФ.

Из данного набора данных при помощи функции select выбираю столбцы:

1. idind (индивидуальный номер),
2. oj13.2 (средняя зарплата за последние 12 месяцев),
3. oh5 (пол респондента),
4. o\_marst (семейное положение),
5. o\_educ (образование),
6. o\_age (количество полных лет),
7. status (тип населенного пункта)
8. oj6.2 (продолжительность рабочей недели).

Удаляю строки содержащие пустые значения (Чертов Unicode).

Далее веду работу с полученной «Датой»

If (был женат или замужем)

Wed2 == 1

If (Состоит в браке)

Wed1 == 1

If ((Никогда не был в браке)

Wed3 == 1

***data2["wed1"]=data2$o\_marst***

***data2["wed1"]=0***

***data2$wed1[which(data2$o\_marst=='2')] <- 1***

***data2$wed1[which(data2$o\_marst=='6')] <- 1***

***data2["wed2"]=data2$o\_marst***

***data2["wed2"]=0***

***data2$wed2[which(data2$o\_marst=='4')] <- 1***

***data2$wed2[which(data2$o\_marst=='5')] <- 1***

***#data2$wed2 = as.numeric(data2$wed2)***

***data2["wed3"]=data2$o\_marst***

***data2["wed3"]=0***

***data2$wed3[which(data2$o\_marst=='1')] <- 1***

***#data2$wed3 = as.numeric(data2$wed3)***

Переменная sex: пол человека 1 – мужчина, 0-женщина

***data2["sex"]=data2$oh5***

***data2$sex[which(data2$sex!='1')] <- 0***

***data2$sex[which(data2$sex=='1')] <- 1***

***data2$sex = as.numeric(data2$sex)***

Аналогично city\_status: 1-город, 0-нет

Далее Высшее образование да либо нет

Также возраст, логарифм от возраста, часов в неделю, и логарифм от часов в неделюю

2 Построение моделей

***model1 = lm(salary~wed1+wed2+wed3+sex+city\_status+higher\_educ+age+dur, data2)***

***model1***

***summary(model1) #R^2 = 0.137***

***vif(model1)***

***output:***

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.54373 0.04021 -13.521 < 2e-16 \*\*\*

wed1 0.01081 0.03674 0.294 0.7686

wed2 -0.01093 0.04710 -0.232 0.8164

wed3 -0.11441 0.04585 -2.495 0.0126 \*

sex 0.40666 0.02466 16.492 < 2e-16 \*\*\*

city\_status 0.30445 0.02608 11.674 < 2e-16 \*\*\*

higher\_educ 0.50632 0.02576 19.658 < 2e-16 \*\*\*

age -0.05955 0.01310 -4.545 5.59e-06 \*\*\*

dur 0.13043 0.01196 10.904 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.929 on 6415 degrees of freedom

Multiple R-squared: 0.1381, Adjusted R-squared: 0.137

F-statistic: 128.5 on 8 and 6415 DF, p-value: < 2.2e-16

wed1 wed2 wed3 sex city\_status higher\_educ age dur

2.434904 2.024549 1.935660 1.118280 1.022135 1.049440 1.277603 1.064891

***model2 = lm(salary~wed2+wed3+sex+city\_status+higher\_educ+age+dur, data2)***

***model2 #***

***summary(model2) #R^2 = 0.1371***

***vif(model2)***

Лучшая модель

***model35 = lm(salary~wed2+wed3+sex+city\_status+higher\_educ+I(age^2)+dur, data2)***

***model35***

***summary(model35) #R^2 = 0.148***

***vif(model35)***

***Output:***

Call:

lm(formula = salary ~ wed2 + wed3 + sex + city\_status + higher\_educ +

I(age^2) + dur, data = data2)

Residuals:

Min 1Q Median 3Q Max

-2.0180 -0.5156 -0.1828 0.2699 15.6246

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.459394 0.027609 -16.640 <2e-16 \*\*\*

wed2 -0.011660 0.034662 -0.336 0.737

wed3 0.007128 0.033949 0.210 0.834

sex 0.422936 0.024554 17.225 <2e-16 \*\*\*

city\_status 0.320224 0.025919 12.355 <2e-16 \*\*\*

higher\_educ 0.501966 0.025509 19.678 <2e-16 \*\*\*

I(age^2) -0.112735 0.011140 -10.120 <2e-16 \*\*\*

dur 0.124105 0.011897 10.432 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9231 on 6416 degrees of freedom

Multiple R-squared: 0.1489, Adjusted R-squared: 0.148

F-statistic: 160.3 on 7 and 6416 DF, p-value: < 2.2e-16

> vif(model35)

wed2 wed3 sex city\_status higher\_educ I(age^2) dur

1.110464 1.074906 1.123126 1.022625 1.042538 1.061074 1.066929

Вывод: Пол, Место жительства влияют сильно. Высшее образование сильнее всего.

## Задача 3

Необходимо провести анализ вашего датасета и сделать обработку данных по предложенному выше алгоритму. Код подготовить в виде питоновских файлов \*.py и сделать отчет в виде ноутбука.

* Сколько в датасете объектов и признаков? Дать описание каждому признаку, если оно есть.
* Сколько категориальных признаков, какие?
* Столбец с максимальным количеством уникальных значений категориального признака?
* Есть ли бинарные признаки?
* Есть ли пропуски?
* Сколько объектов с пропусками?
* Столбец с максимальным количеством пропусков?
* Есть ли на ваш взгляд выбросы, аномальные значения?
* Столбец с максимальным средним значением после нормировки признаков через стандартное отклонение?
* Столбец с целевым признаком?
* Сколько объектов попадает в тренировочную выборку при использовании train\_test\_split с
* параметрами test\_size = 0.3, random\_state = 42?
* Сколько признаков достаточно для объяснения 90% дисперсии после применения метода PCA?
* Какой признак вносит наибольший вклад в первую компоненту?

1.Загрузка и предварительный просмотр данных

Библиотеки:

***import numpy as np***

***import pandas as pd***

Загружаем данные из файла:

data = pd.read\_csv("winequality-red.csv") #Чтение данных

Рисунок.10

Первичный анализ :

Объем датасета



Рисунок 11.

Наша таблица содержит 1599 столбцов и 9 строк

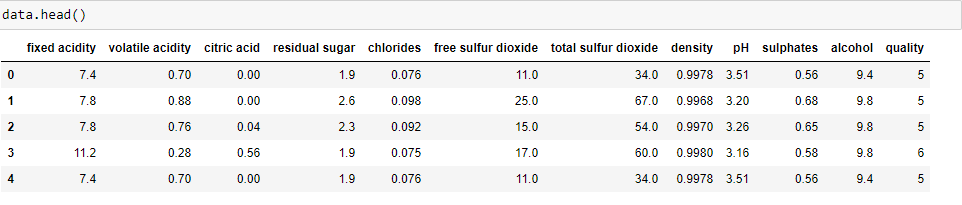
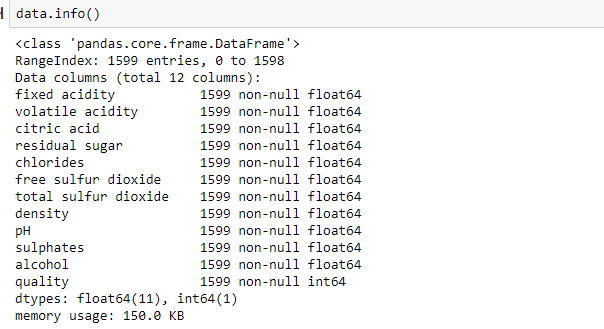
Смотрим на верхушку

Рисунок 12.

Выводим информацию о наших данных:

Рисунок 13.

Делая вывод из проделанных операций, можно сказать, что датасет содержит 1599 объект и у каждого объекта 12 признаков. Описание признаков:

1) fixed acidity - дробный признак, фиксированная кислотность вина

2) volatile acidity - дробный признак,летучая кислотность вина

3) citric acid - дробный признак, лимонная кислота

4) residual sugar - дробный признак,остаточный сахар

5) chlorides, - дробный признак,хлориды

6) free sulfur dioxide, - дробный признак,св диоксид серы

7) total sulfur dioxide - дробный признак,весь диоксид серы

8) density - дробный признак,плотность

9) pH- дробный признак,кислотность и щелочность

10) sulphates - дробный признак, сульфаты

11) alcohol - дробный признак, алкоголь

12) quality - целочисленный признак , качество вина

Смотрим на наши данные и видим, что категориальным признаком является *quality.*

Столбцом с максимальным количеством уникальных значений категориального признака так же является столбец *quality*

(он принимает значения от 0 до 10)

По нашей таблице мы так же можем заметить то что у нас нет бинарных признаков.

Далее проверяем нашу таблицу на пропуски:

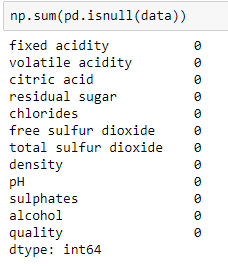


Рисунок 13.

Здесь видно, что количество пропусков в каждом признаке равно 0

Далее смотрим на наши данные снова на наличие аномалий функцией data.describe():

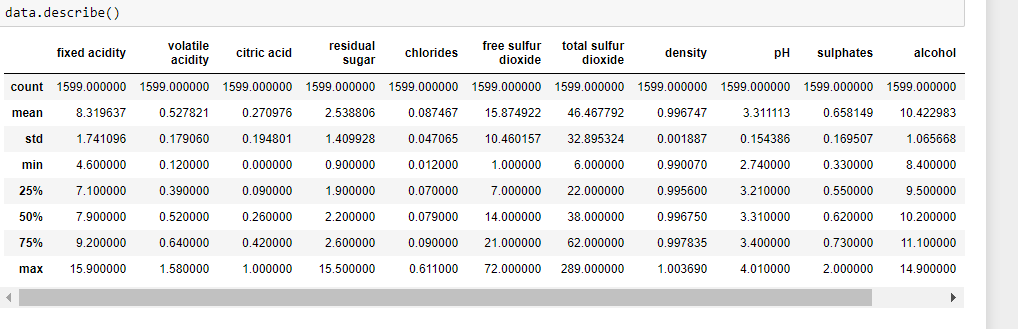


Рис.14

Просматривая данные таблицы, можно сделать вывод о том, что возможно есть аномалия в total sulfur dioxide очень высокий в max.

Далее копируем нашу таблицу для дальнейшего использования без потери данных и кореллируем наши данные:

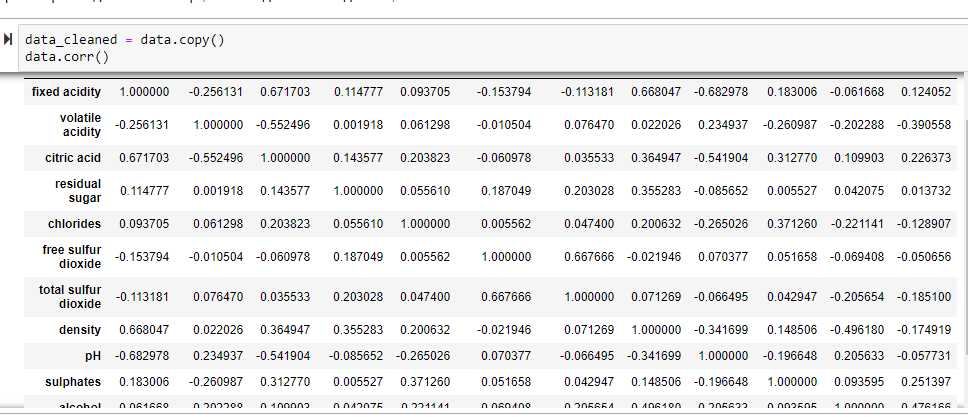




Рис.15

Оценив наши данные мы увидим , что можно убрать слишком маленькие значения quality .

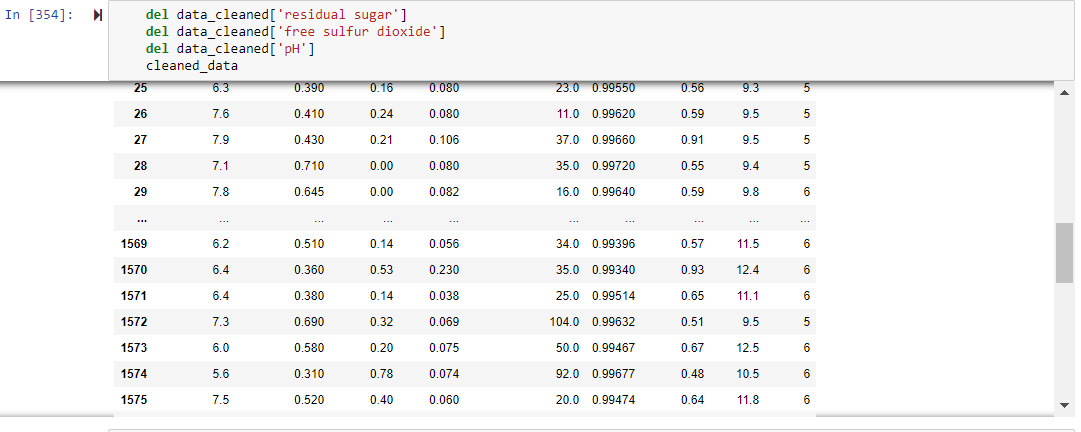


Рис.15

2.Стандартизация переменных

Теперь нормализируем признаки:

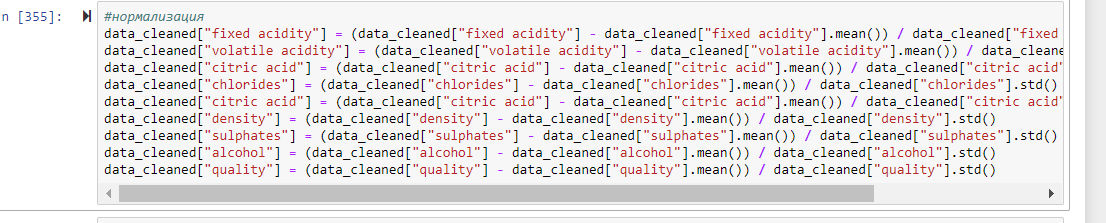


Рис.16

Теперь нормализируем признаки:

Далее загрузив наши данные видим ,что максимальное среднее значение нормированных признаков в total sulfur dioxide:

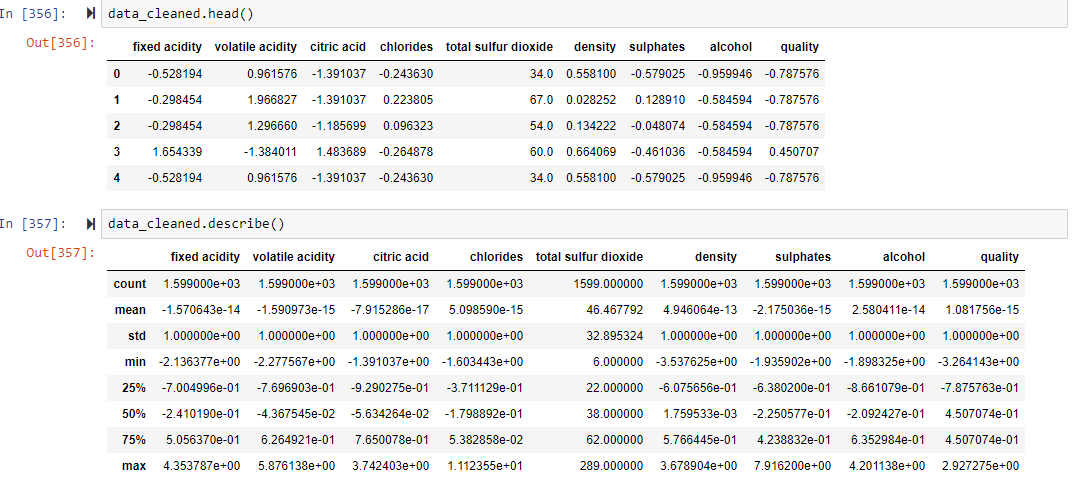


Рис.17

Целевой переменной делаем quality и смотрим сколько объектов попадает в тренировочную выборку:

Рис.18

В тренировочную выборку попадает 1119 объектов.

3.4.Метод главных компонент.

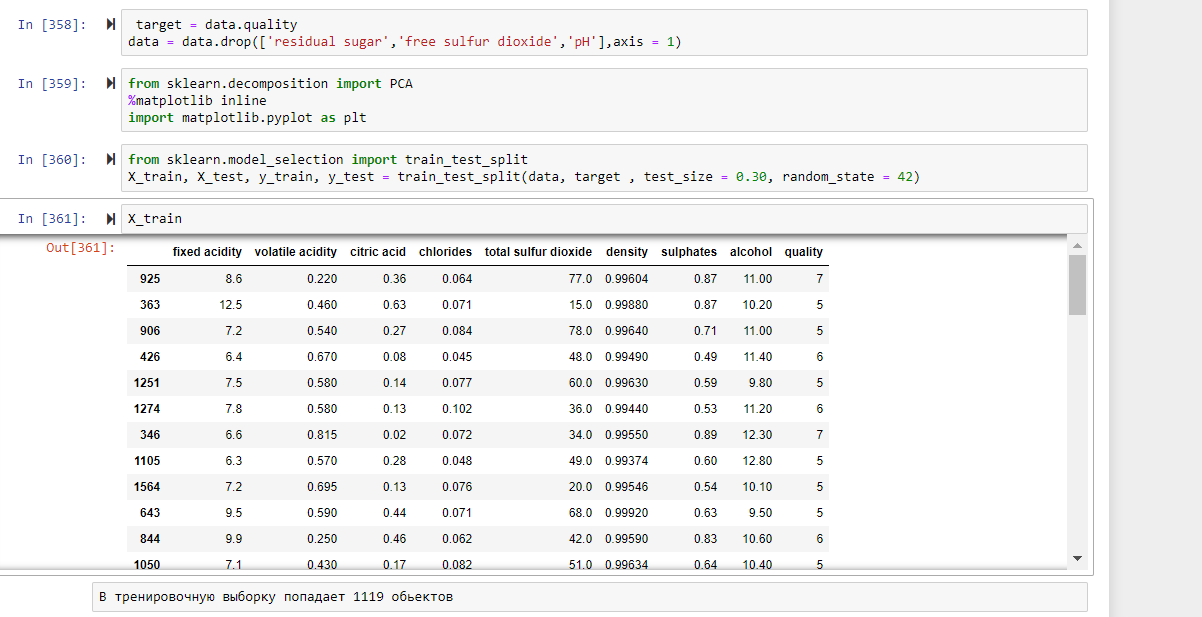


Рис.19

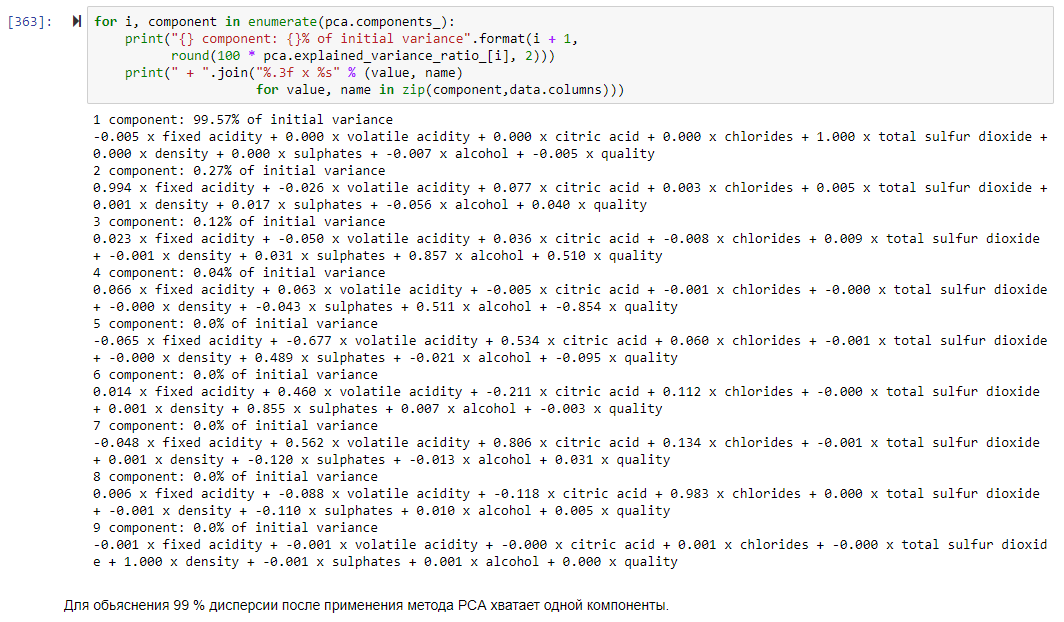


Рис.20

Для обьяснения 99%  дисперсии после применения метода PCA хватает одной компоненты.

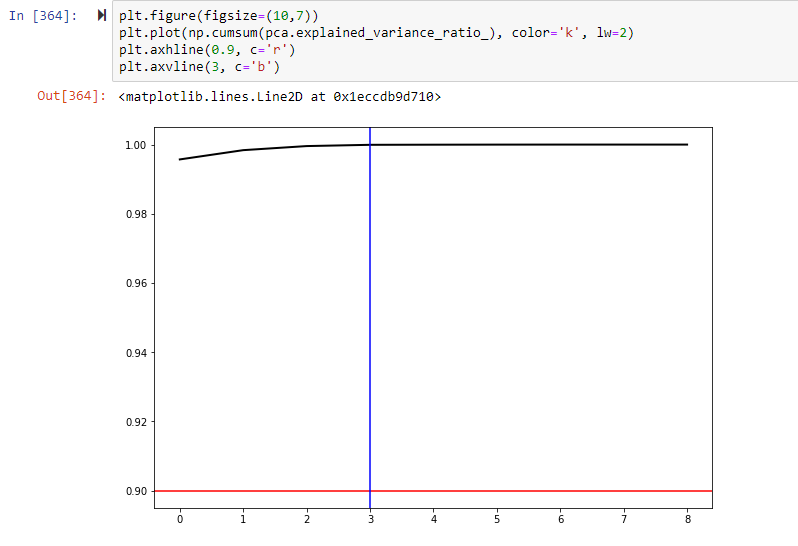


Рис.21

Вывод:

В анализе данных, как и в любом другом анализе, порой бывает не лишним создать упрощенную модель, максимально точно описывающую реальное положение дел. Часто бывает так, что признаки довольно сильно зависят друг от друга и их одновременное наличие избыточно и для упрощения нашего вывода существует метод главных компонент PCA.

**Метод главных компонент (англ. principal component analysis, PCA)** — один из основных способов уменьшить размерность данных, потеряв наименьшее количество информации.В данном случае мы обработали наши данные и получили что одной переменной можем описать 99% наших данных.