

Psychology by Using AI

Prediction of a depression scale by using Machine learning models



by:Muhammad Ammad 2012350

Emad 2012339

Rafay Aziz 2012354

# Input Data:

The survey was open to anyone and people were motivated to take it to get personalized results. They participants were given an option to fill out a research to have their data used for the research, the dataset used for this project was based on those candidate’s submissions.

This data was collected 2017 - 2019.

The following items were included in the survey:

* Q1 I found myself getting upset by quite trivial things.
* Q2 I was aware of dryness of my mouth.
* Q3 I couldn't seem to experience any positive feeling at all.
* Q4 I experienced breathing difficulty (eg, excessively rapid breathing, breathlessness in the absence of physical exertion).
* Q5 I just couldn’t seem to get going.
* Q6 I tended to over-react to situations.
* Q7 I had a feeling of shakiness (eg, legs going to give way).
* Q8 I found it difficult to relax.
* Q9 I found myself in situations that made me so anxious I was most relieved when they ended.
* Q10 I felt that I had nothing to look forward to.
* Q11 I found myself getting upset rather easily.
* Q12 I felt that I was using a lot of nervous energy.
* Q13 I felt sad and depressed.
* Q14 I found myself getting impatient when I was delayed in any way (eg, elevators, traffic lights, being kept waiting).
* Q15 I had a feeling of faintness.
* Q16 I felt that I had lost interest in just about everything.
* Q17 I felt I wasn’t worth much as a person.
* Q18 I felt that I was rather touchy.
* Q19 I perspired noticeably (eg, hands sweaty) in the absence of high temperatures or physical exertion.
* Q20 I felt scared without any good reason.
* Q21 I felt that life wasn’t worthwhile.
* Q22 I found it hard to wind down.
* Q23 I had difficulty in swallowing.
* Q24 I couldn’t seem to get any enjoyment out of the things I did.
* Q25 I was aware of the action of my heart in the absence of physical exertion (eg, sense of heart rate increase, heart missing a beat).
* Q26 I felt down-hearted and blue.
* Q27 I found that I was very irritable.
* Q28 I felt I was close to panic.
* Q29 I found it hard to calm down after something upset me.
* Q30 I feared that I would be “thrown” by some trivial but unfamiliar task.
* Q31 I was unable to become enthusiastic about anything.
* Q32 I found it difficult to tolerate interruptions to what I was doing.
* Q33 I was in a state of nervous tension.
* Q34 I felt I was pretty worthless.
* Q35 I was intolerant of anything that kept me from getting on with what I was doing.
* Q36 I felt terrified.
* Q37 I could see nothing in the future to be hopeful about.
* Q38 I felt that life was meaningless.
* Q39 I found myself getting agitated.
* Q40 I was worried about situations in which I might panic and make a fool of myself.
* Q41 I experienced trembling (eg, in the hands).
* Q42 I found it difficult to work up the initiative to do things.

Each item was presented one at a time in a random order for each new participant along with a 4 point rating scale asking the user to indicate how often that had been true of them in the past week where

1 = Did not apply to me at all

2 = Applied to me to some degree, or some of the time

3 = Applied to me to a considerable degree, or a good part of the time

4 = Applied to me very much, or most of the time

This response is stored in variable A (e.g. Q1A).

Along with the question itself, it was recorded how long it took for that question to be filled and the question’s position in the survey. With their indication by Q1E and Q1I respectively.

These other durations were also recorded (measured on the server's side):

* introelapse The time spent on the introduction/landing page (in seconds)
* testelapse The time spent on all the DASS questions (should be equivalent to the time elapsed on all the indiviudal questions combined)
* surveyelapse The time spent answering the rest of the demographic and survey questions

TIPI1 Extraverted, enthusiastic.

TIPI2 Critical, quarrelsome.

TIPI3 Dependable, self-disciplined.

TIPI4 Anxious, easily upset.

TIPI5 Open to new experiences, complex.

TIPI6 Reserved, quiet.

TIPI7 Sympathetic, warm.

TIPI8 Disorganized, careless.

TIPI9 Calm, emotionally stable.

TIPI10 Conventional, uncreative.

The TIPI items were rated "I see myself as:" \_\_\_\_\_ such that

1 = Disagree strongly

2 = Disagree moderately

3 = Disagree a little

4 = Neither agree nor disagree

5 = Agree a little

6 = Agree moderately

7 = Agree strongly

The following items were presented as a check-list and subjects were instructed "In the grid below, check all the words whose definitions you are sure you know":

* VCL1 boat
* VCL2 incoherent
* VCL3 pallid
* VCL4 robot
* VCL5 audible
* VCL6 cuivocal
* VCL7 paucity
* VCL8 epistemology
* VCL9 florted
* VCL10 decide
* VCL11 pastiche
* VCL12 verdid
* VCL13 abysmal
* VCL14 lucid
* VCL15 betray
* VCL16 funny

A value of 1 is checked, 0 means unchecked. The words at VCL6, VCL9, and VCL12 are not real words and can be used as a validity check.

A bunch more questions were then asked:

Education "**How much education have you completed**?",

*1=Less than high school, 2=High school, 3=University degree, 4=Graduate degree*

Urban **"What type of area did you live when you were a child?",**

*1=Rural (country side), 2=Suburban, 3=Urban (town, city)*

Gender "**What is your gender**?",

*1=Male, 2=Female, 3=Other*

Engnat **"Is English your native language?",**

*1=Yes, 2=No*

Age **"How many years old are you?"**

Hand "**What hand do you use to write with?",**

*1=Right, 2=Left, 3=Both*

Religion "**What is your religion?",**

*1=Agnostic, 2=Atheist, 3=Buddhist, 4=Christian (Catholic), 5=Christian (Mormon), 6=Christian (Protestant), 7=Christian (Other), 8=Hindu, 9=Jewish, 10=Muslim, 11=Sikh, 12=Other*

Orientation **"What is your sexual orientation?", *1****=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual, 5=Other*

Race **"What is your race?",**

*10=Asian, 20=Arab, 30=Black, 40=Indigenous Australian, 50=Native American, 60=White, 70=Other*

Voted **"Have you voted in a national election in the past year?",**

*1=Yes, 2=No*

Married **"What is your marital status?",**

*1=Never married, 2=Currently married, 3=Previously married*

Familysize **"Including you, how many children did your mother have?"**

Major **"If you attended a university, what was your major**

*(e.g. "psychology", "English", "civil engineering")?"*

The following values were derived from technical information:

Country *ISO country code of where the user connected from*

Screensize *1=device with small screen (phone, etc),*

*2=device with big screen (laptop, desktop, etc)*

Uniquenetworklocation 1=only one survey from user's specific network in dataset,

2=multiple surveys submitted from the network of this user(2 does not necessarily imply duplicate records for an individual, as it could be different students at a single school or different members of the same household; and even if 1 there still could be duplicate records from a single individual e.g. if they took it once on their WIFI and once on their phone)

Source **how the user found the test,**

*1=from the front page of the site hosting the survey, 2=from google, 0=other or unknown*

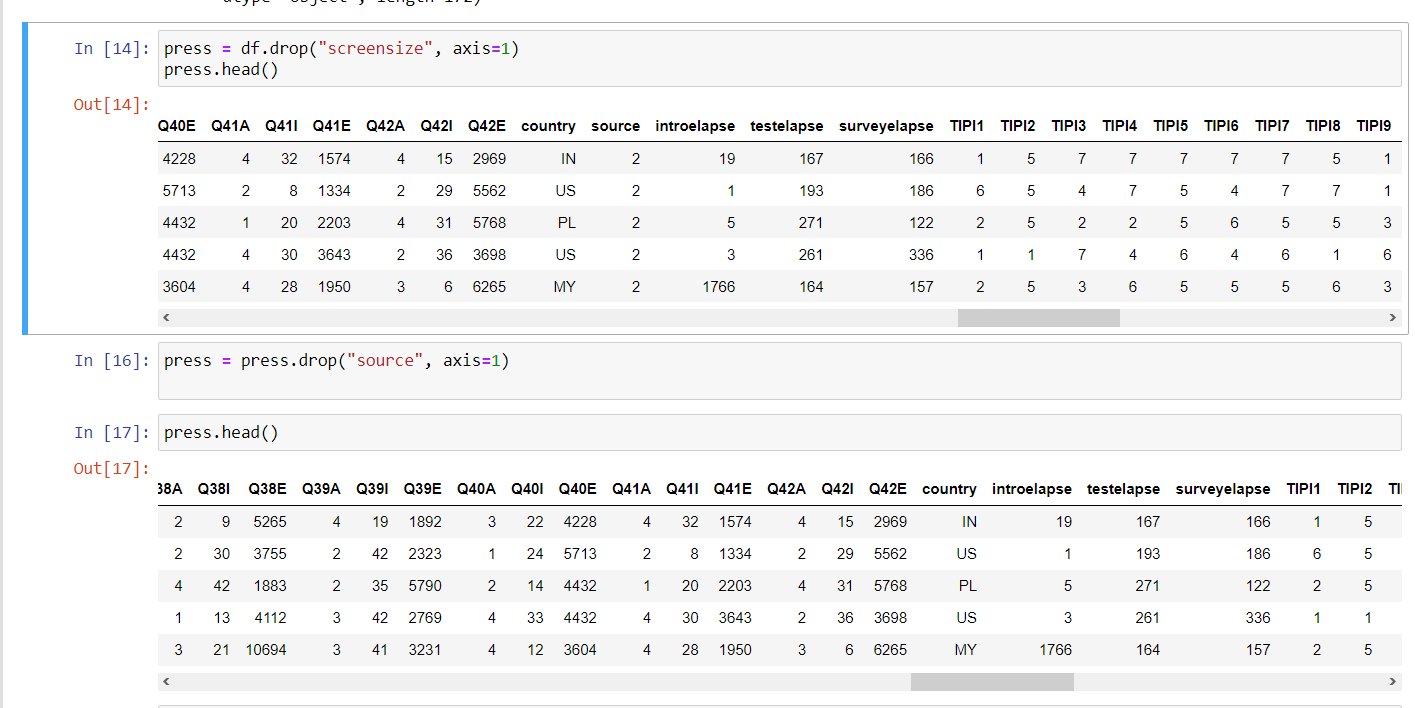
Now that the initial columns have been introduced, Lets begin the machine learning.

# Preprocessing:-

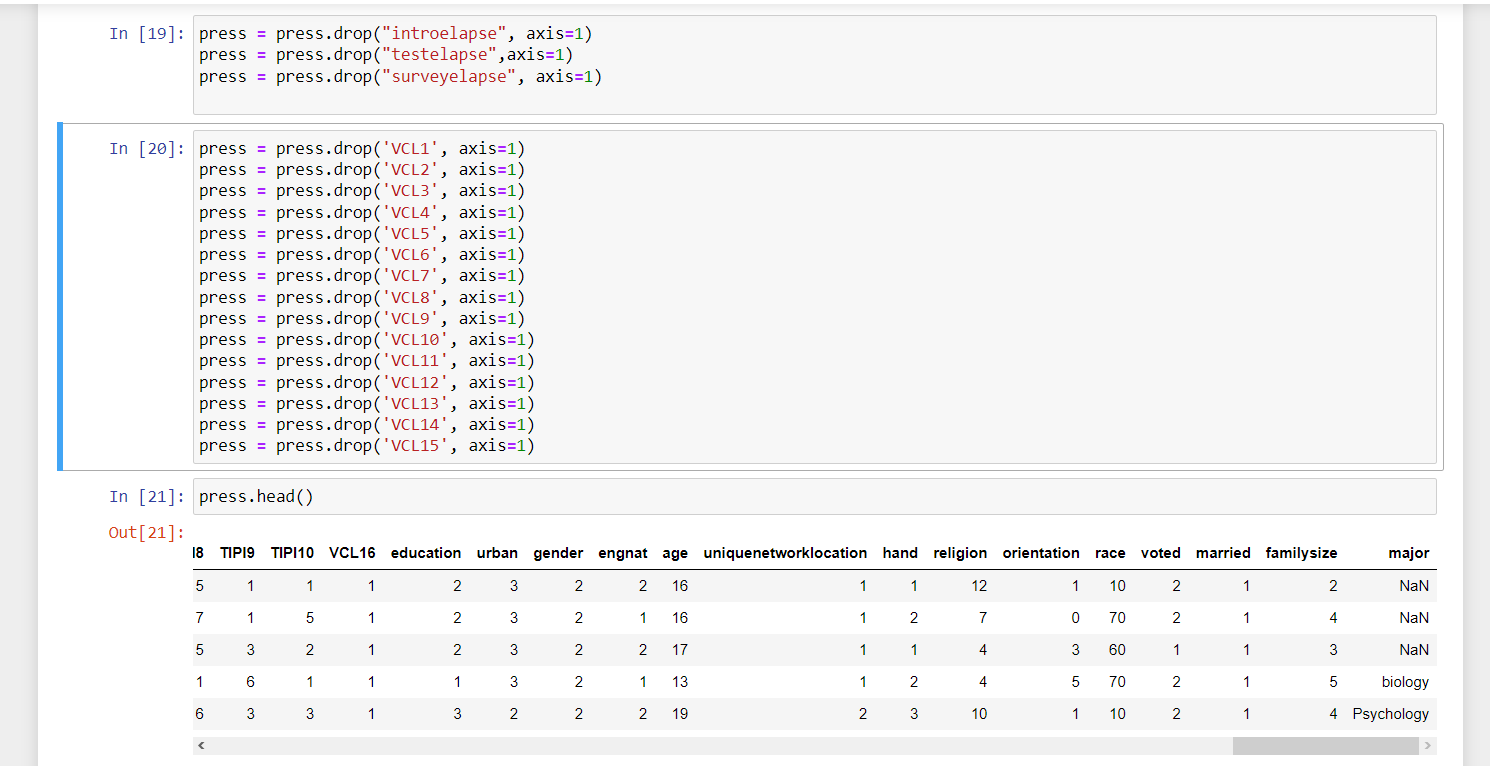


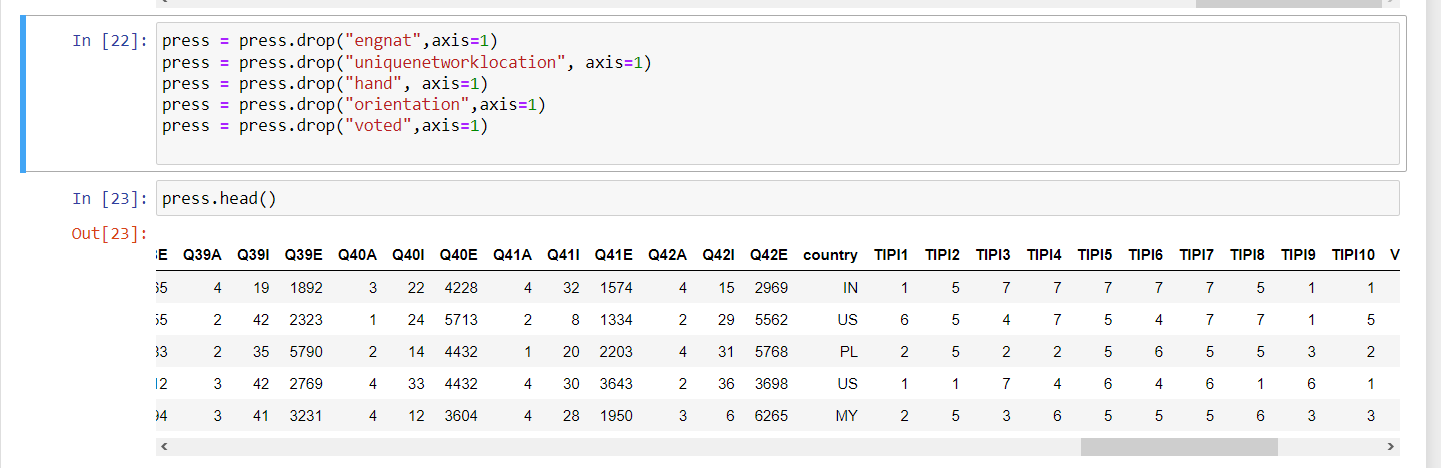
Reading the file as a delimiter.

We had to drop and clean some data in order to grasp a better understanding of end result using the point system.

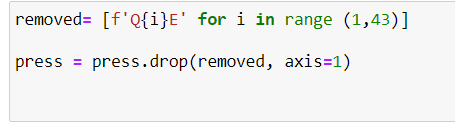


We started with screensize and source. Easy columns to drop as we weren’t categorically aware how it would affect depression.

We then dropped these columns cause they wouldn’t have provided data that was needed for psych analysis.



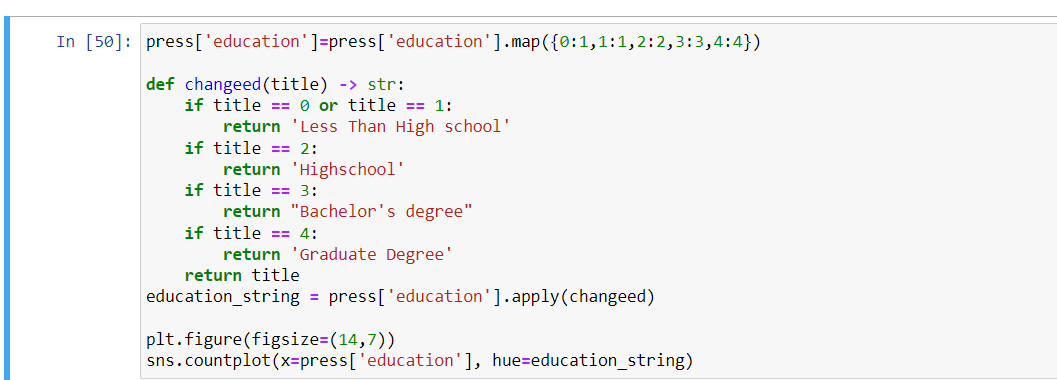
We then dropped the columns that would’ve made it more complicated for us and would require more data on this.

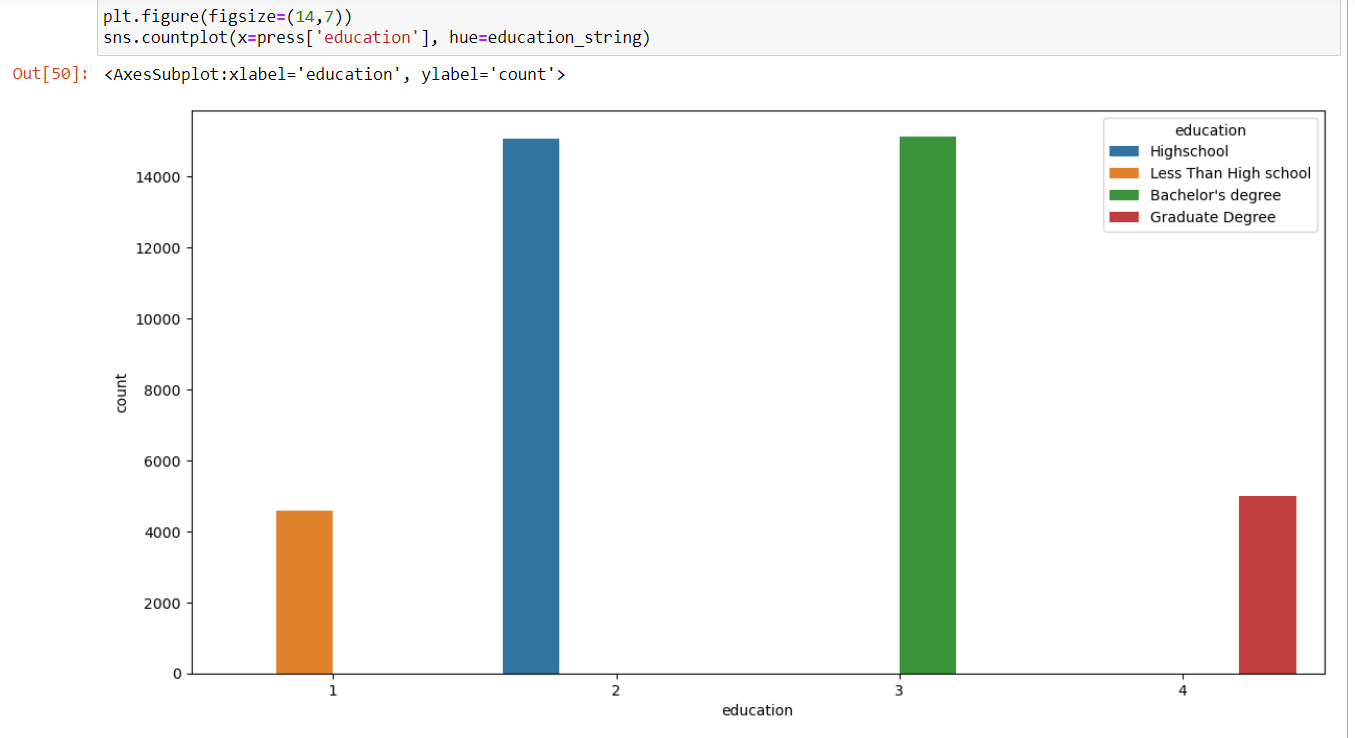


We then removed the columns that was included in every question regarding the time it took and the question’s position.

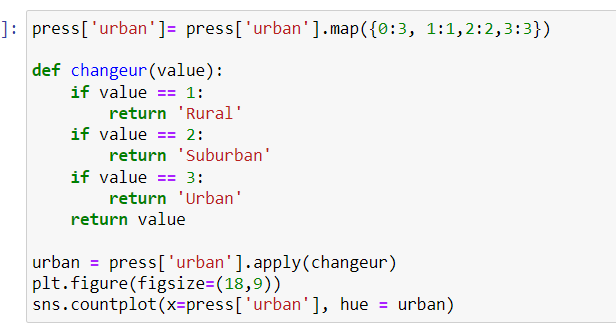


We noticed In the major column, That almost all of the columns weren’t filled so that means we didn’t have sufficient data to work with different majors.

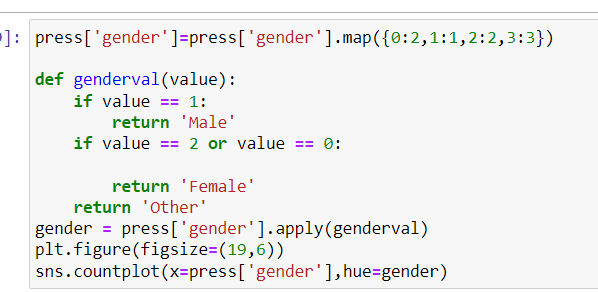


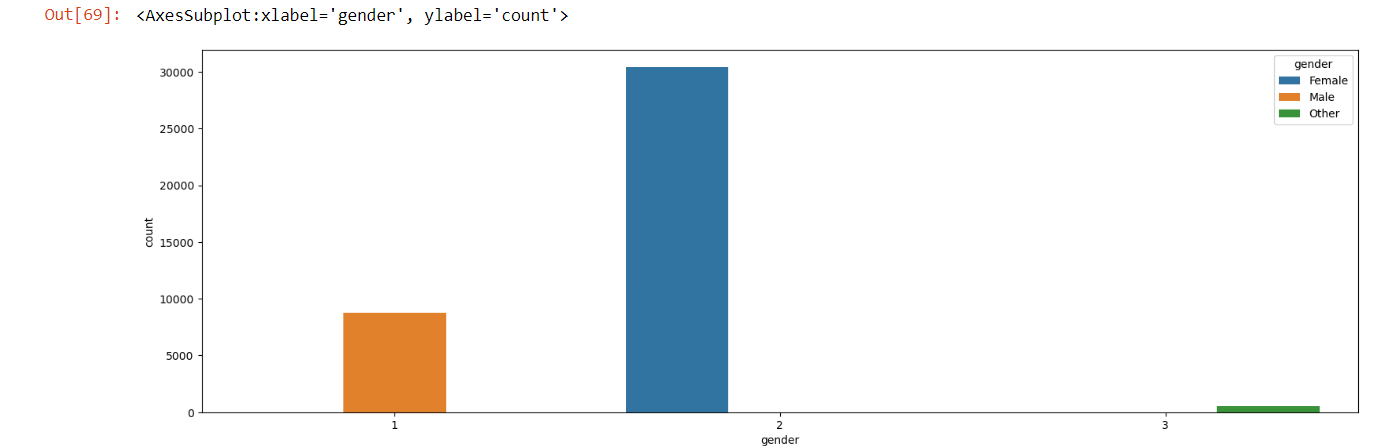


We took all the values of 0, and situated them in to “less than highschool”



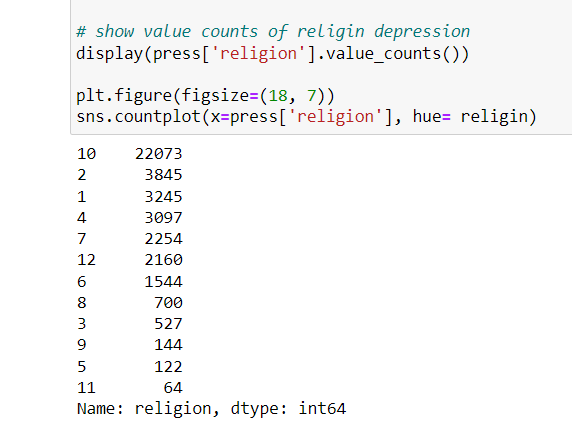
We changed all of the 0 values, to 3s as we assume that person might be urban due to industrial revolution.



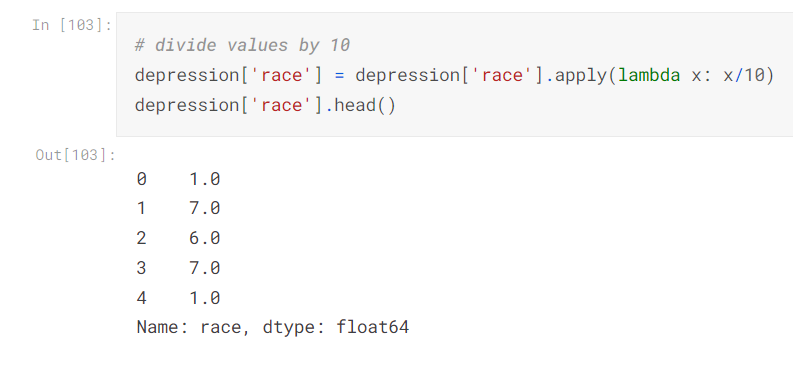


We took the values of 0, changed it into 2 as most of the candidates were females.





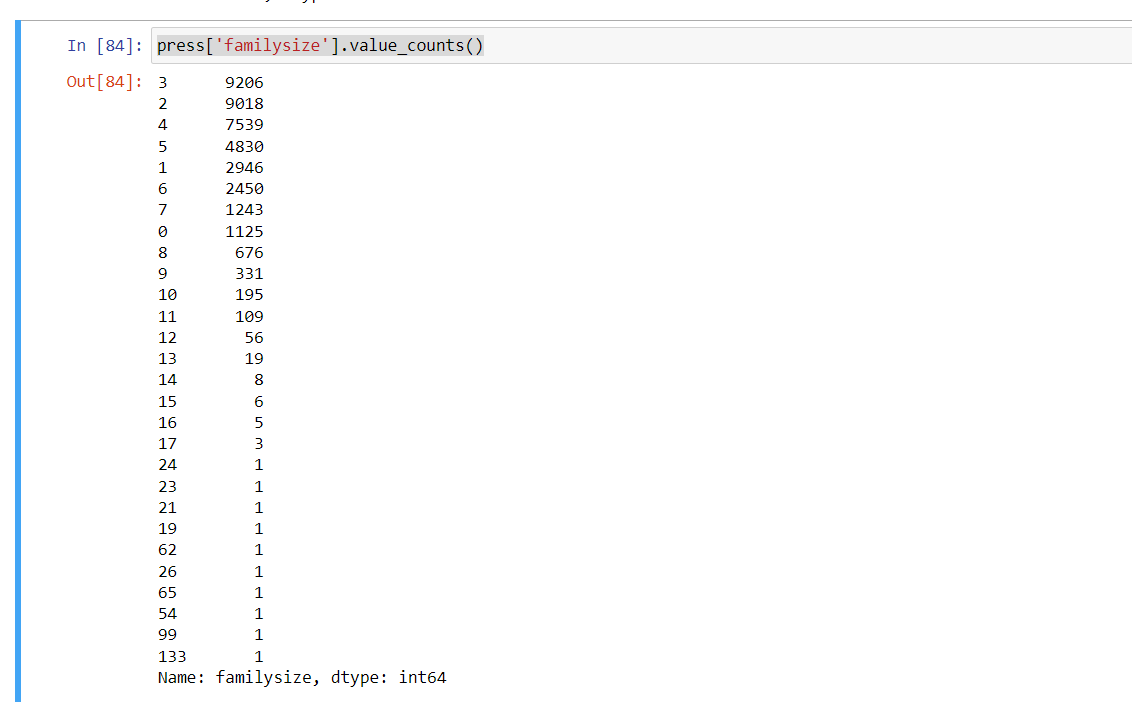
We took 0 values, and situated it onto the 12 value, which was the “Other” option.



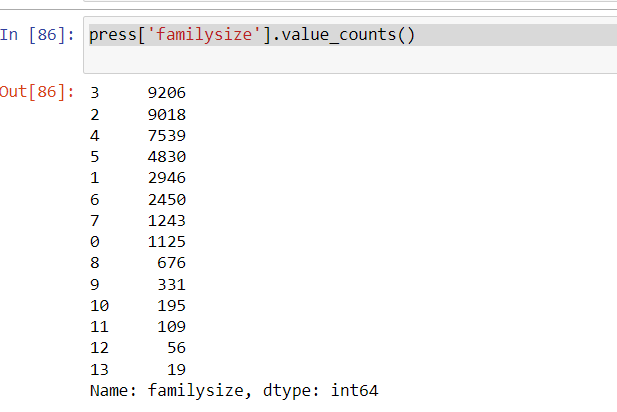
Since we needed the points to be added up from natural numbers, we simply divided the column of race by 10.



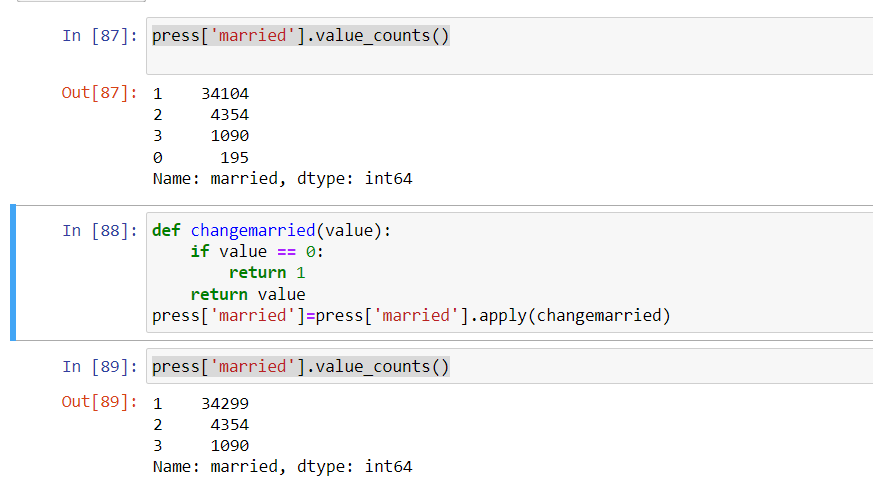
We changed the values of 0 in every TIPI question to 5 as it was the most common answer either way.



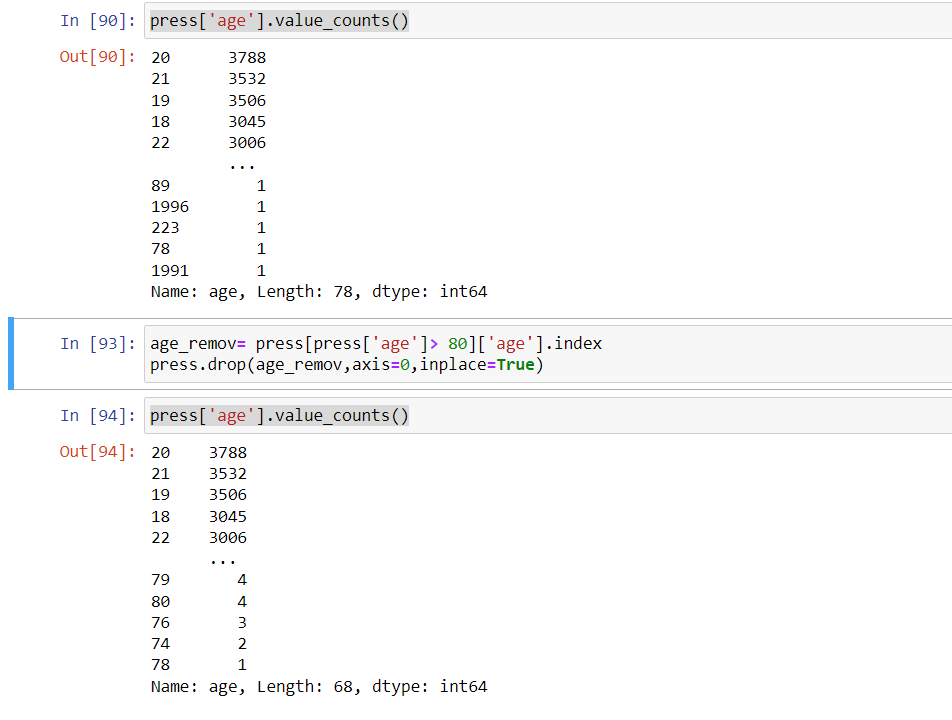
No way someone had a family of 133, or 90 even. We had to remove the outrageous answers.



A family of 13 still seemed a little too odd, but since it had double digits we went through with it.



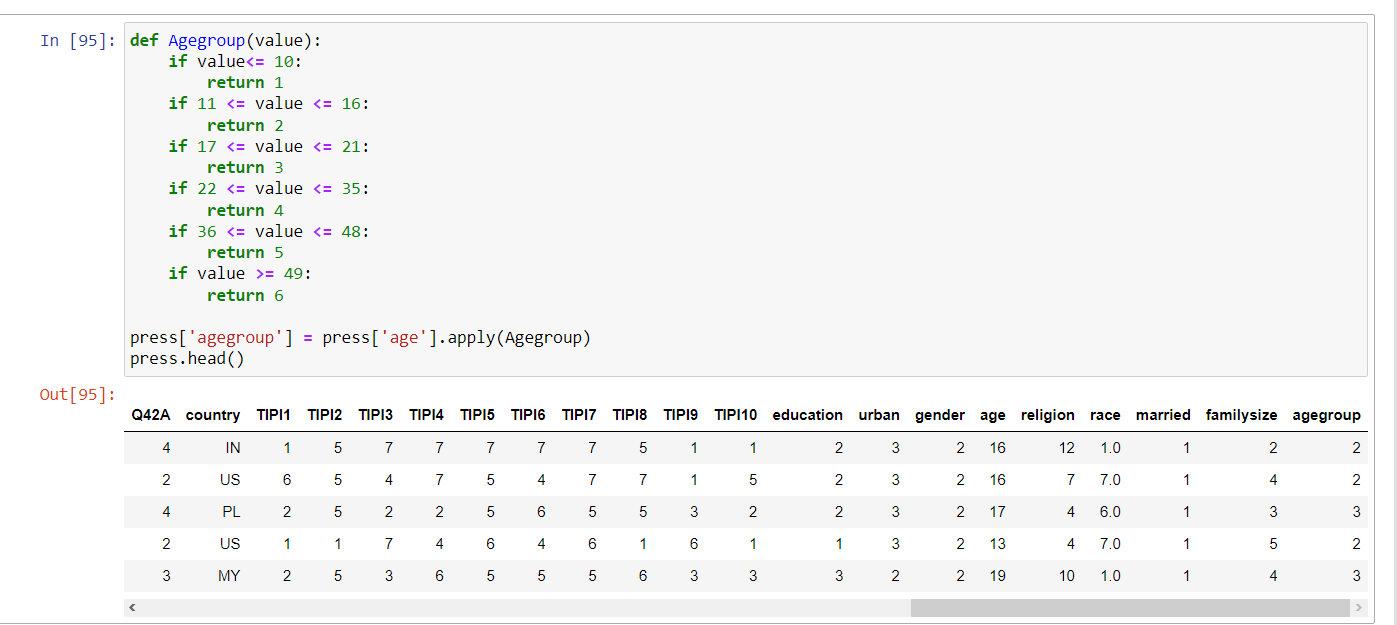
We changed values 0, to never married which was 1.



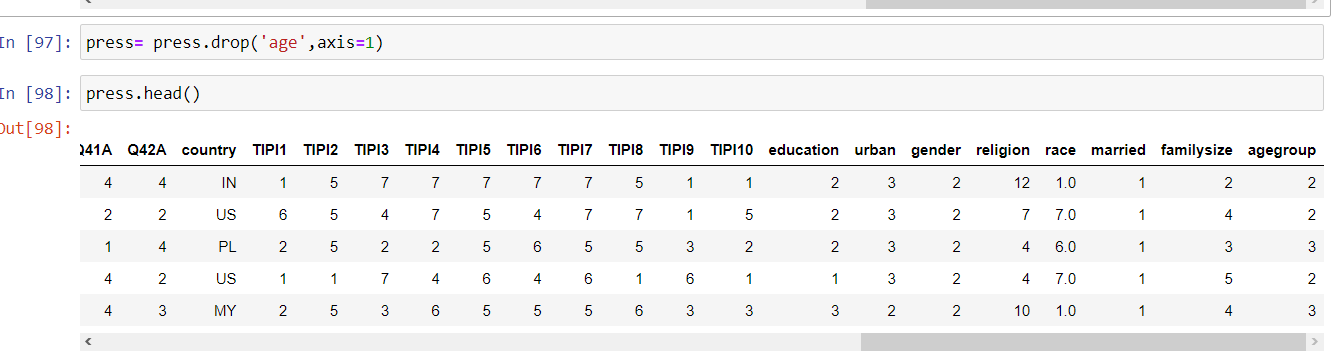
Another outrageous data(age 223, really?)

We removed any age before the number 80 for realistic analysis.

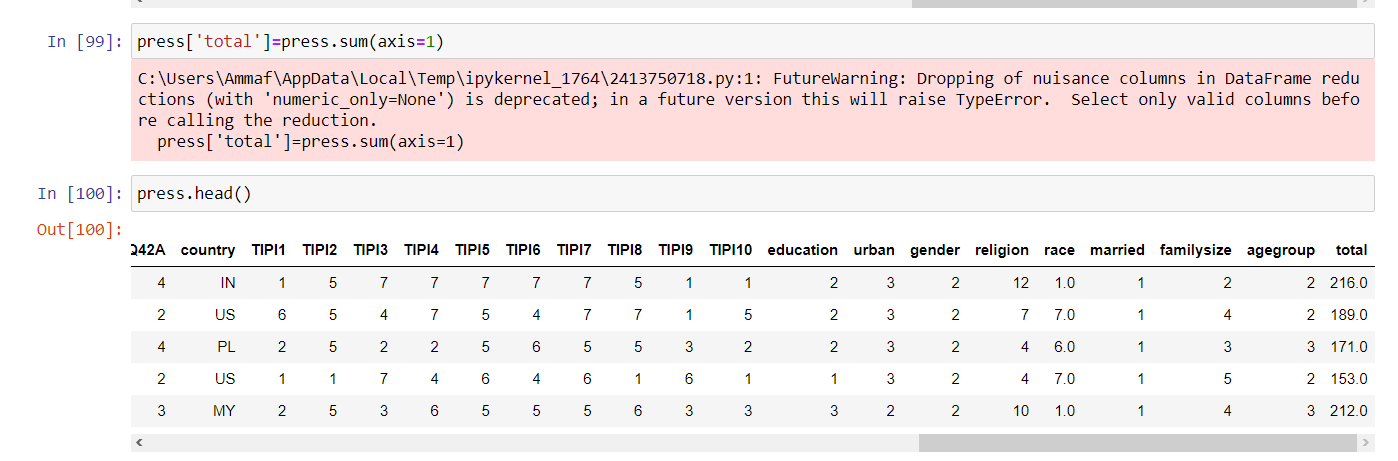
As the ages were shuffled from 1 to 80, we couldn’t have numbers that wide so to rectify that, we made age groups for a more controlled environment.



We situated the bigger numbers for a more experienced age and life.

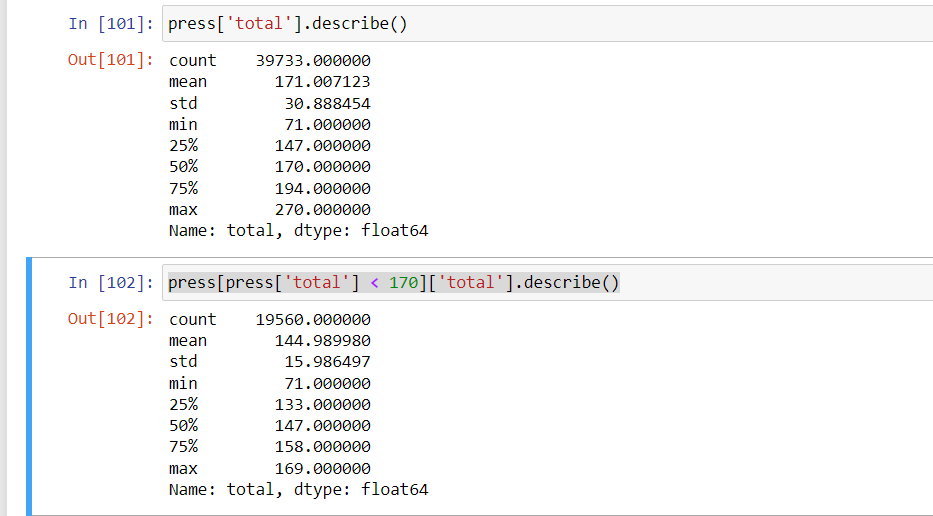


Therefore the age column had to be dropped.



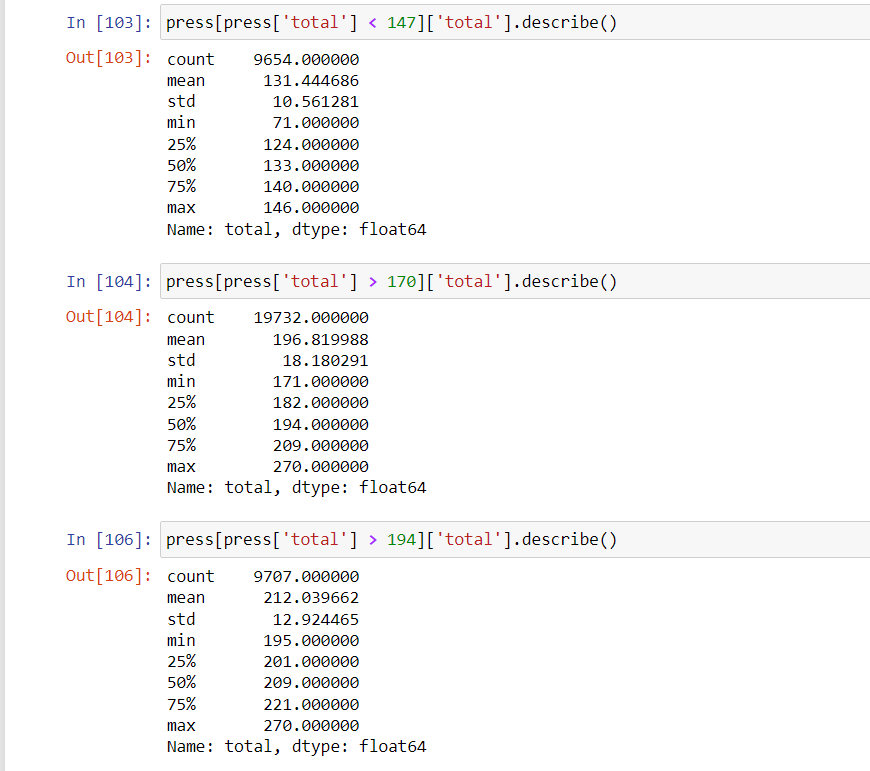
We made a new column that added all of the points into the total.

Except country, since there were too many countries and not enough data or time to situate them onto different continents. We removed it

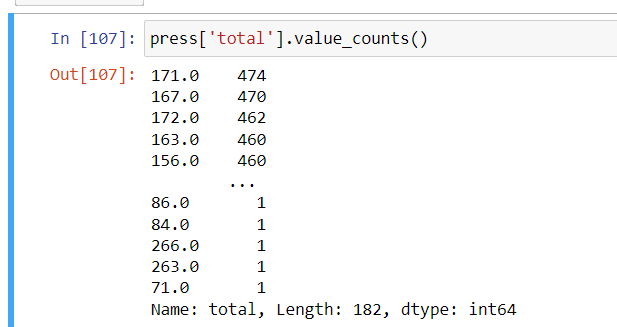


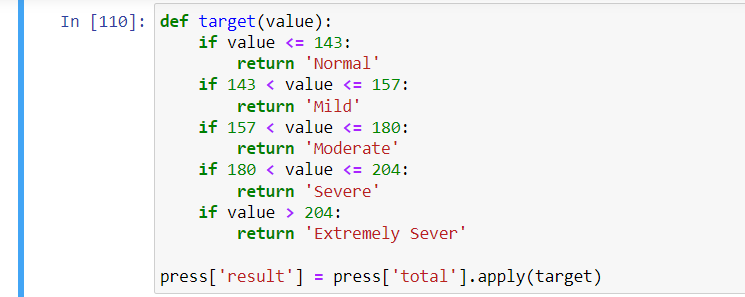
We wanted to find the mean value of total points so we could split the results out of 5 categories.

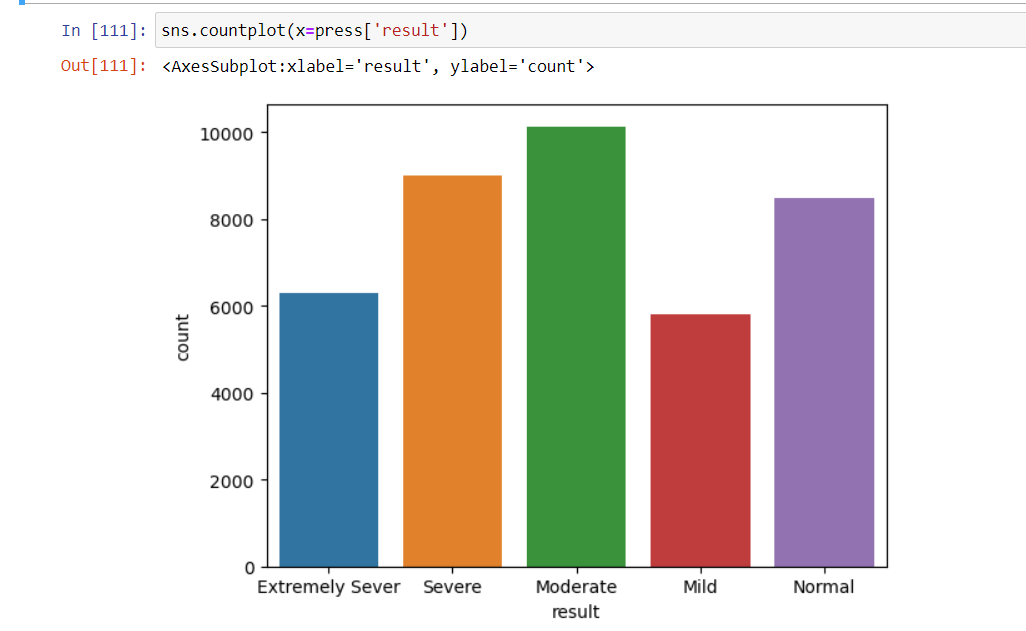
Normal, Mild, Moderate, Severe and Extremely Severe.



We decided to bump up the numbers by 10 in each category because most of the values were summing up to under 170 so the data would get unbalanced.

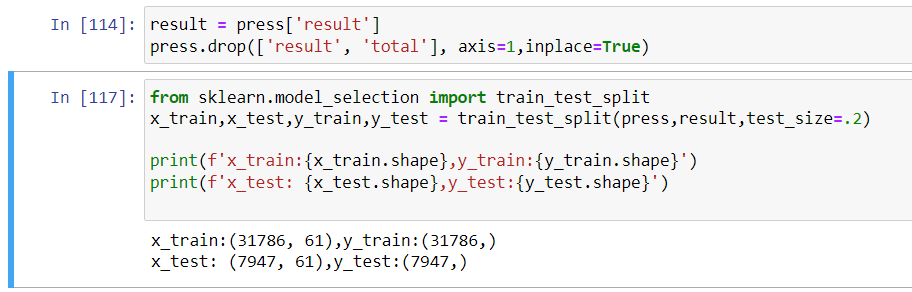






We made a new column named result.

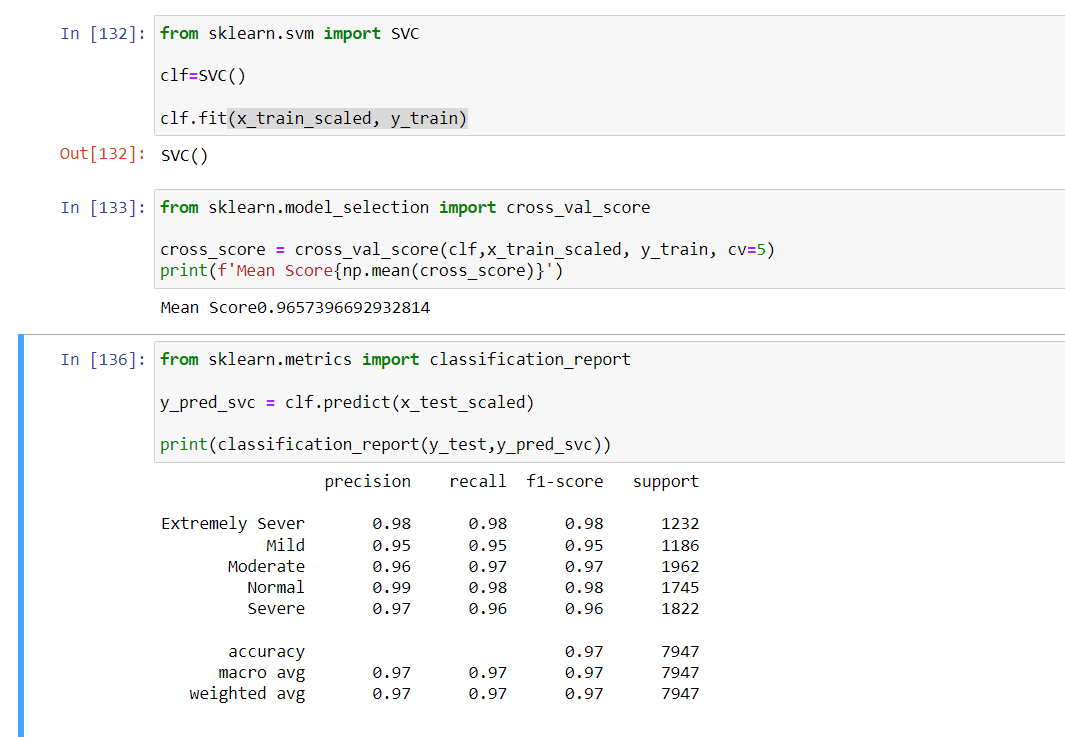
# Splitting



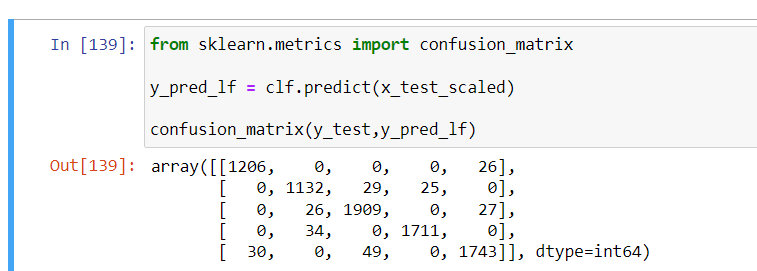
We split up the data into X and Y, while also removing the columns total and result.

# Model Selection:-

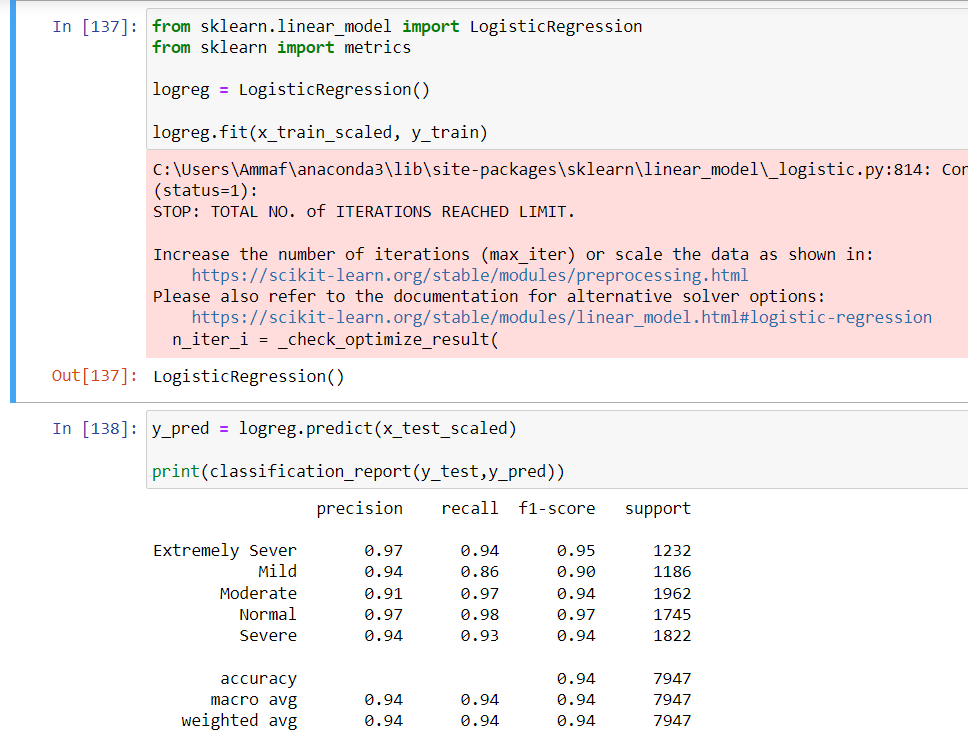
## SVM Model Build:



With its confusion matrix.



### Logistic Regression:-



With its confusion matrix

