

**PROJECT (P&S)**

**GROUP MEMBERS:**

**SYEDA KISAA FATIMA**

**TEHREEM ZUBAIR**

**JAWARIA KHALID**

**DEPARTMENT:**

**COMPUTER & INFORMATION SCIENCES**

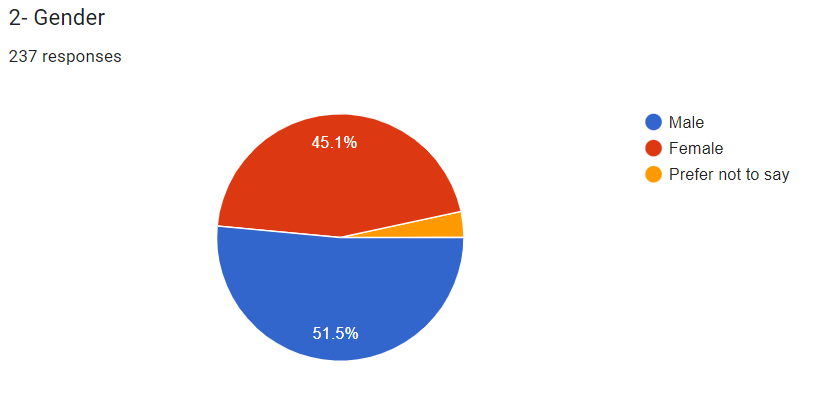
**TOPIC**

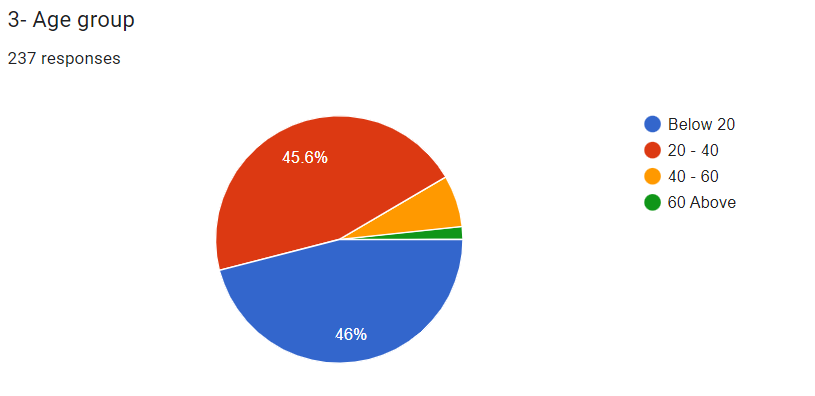
**SURVEY: RECYCLING AND ITS AWARENESS**

For the project, we were told to conduct a survey on any topic, for this purpose e decided to opt for a topic related to environmental issues, as it is a very crucial topic nowadays and we wanted to know about the views of the people on the threats being faced by our environment. So the topic we chose for the survey was "recycling and its importance."

**Summary:**

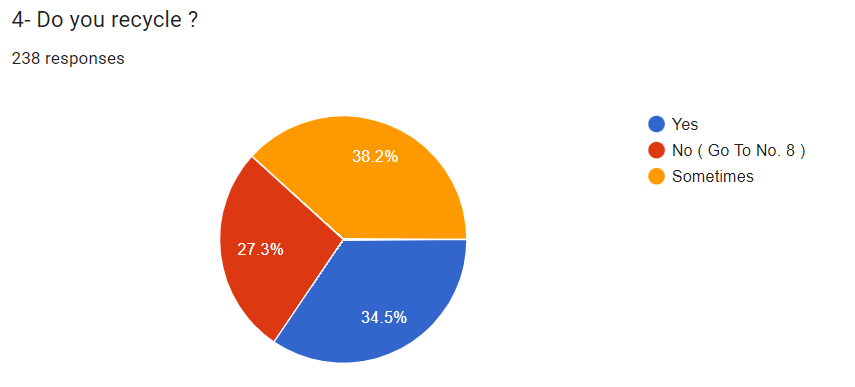
* The following pie charts are the overview of the total responses we got in terms of age grouping and gender segregation.

****



* Up next we did statistical analysis on a few of the survey questions to draw conclusions regarding:
  + - * + Mean, Median, Mode
* Frequency Distribution
* Standard deviation
* Variance
* Hypothesis testing

***Mean & Median***



* we computed MEAN & MEDIAN of the question where we asked whether that specific individual is into recycling or not.

Table below is used to calculate mean and median:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **Do you recycle?** | **Yes** | **Sometimes** | **No** | **Grand Total** |
| Female | 41 | 45 | 22 | 108 |
| Male | 40 | 41 | 41 | 122 |
| Prefer not to say | 1 | 5 | 2 | 8 |
| **Grand Total** | **82** | **91** | **65** | **238** |

Then using the formula we find mean and median for each separate option:

**MEAN = Σ x /n**

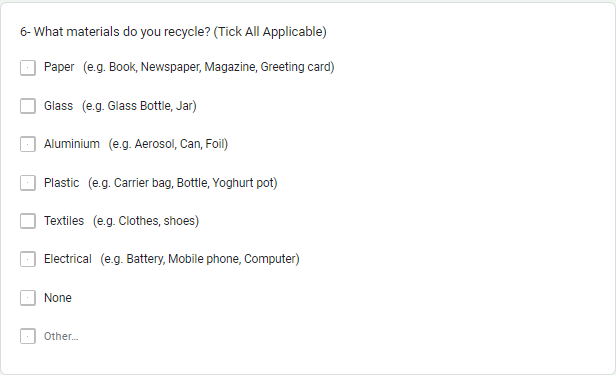
**MEDIAN = (n+1)/2**

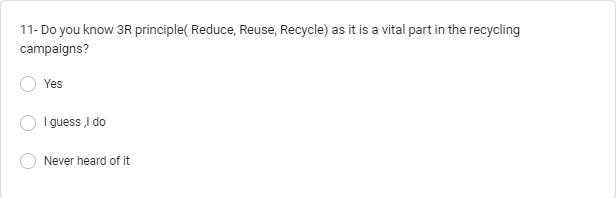
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Yes** | **Sometimes** | **No** |
| Mean | 27.333 | 29.583 | 23.645 |
| Median | 40 | 40.5 | 31.25 |

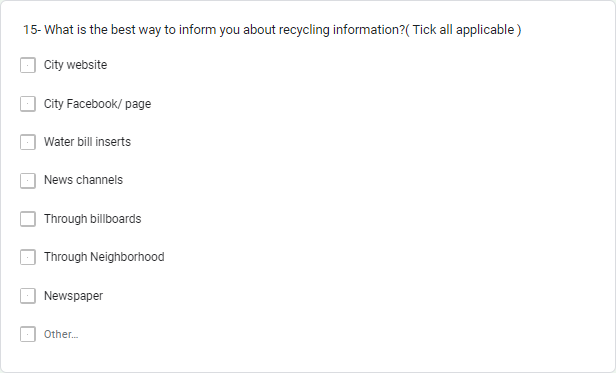
**Bar chart for mean & median:**

***SURVEY FORM***

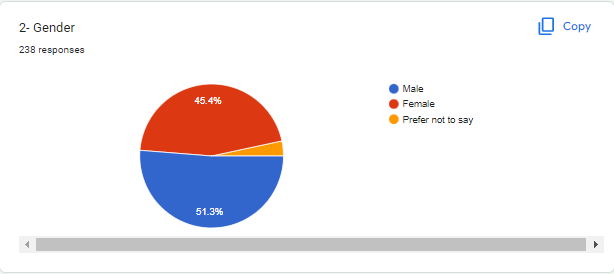
* The first and most important step was to conduct the survey, for this we made survey forms. While making the survey form we ensured that the questions are not too lengthy or boring, also their relevance to our topic was taken into view. Circulation of the form was done and luckily we had very good responses. In total, we succeeded in collecting more than 200 responses.
* Some of the snippets from the survey questions are given below:

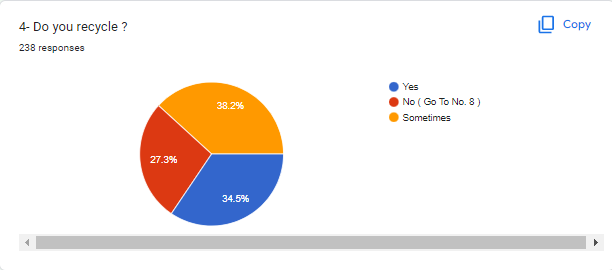


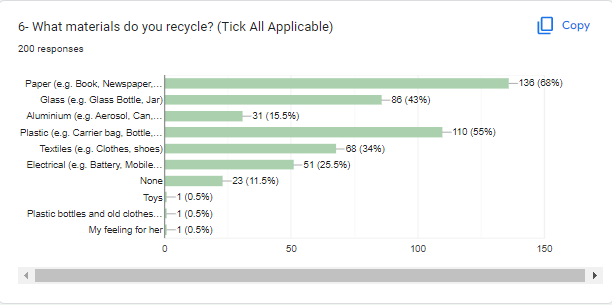


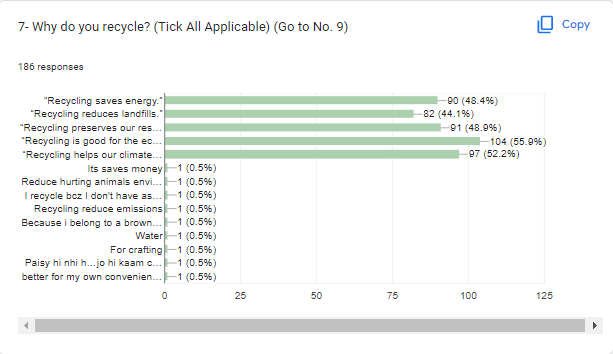


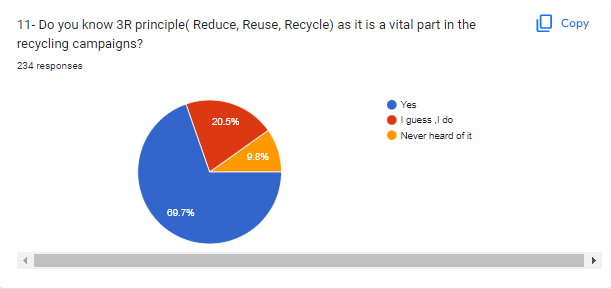
* Some of the survey results that we got from google form are given below:

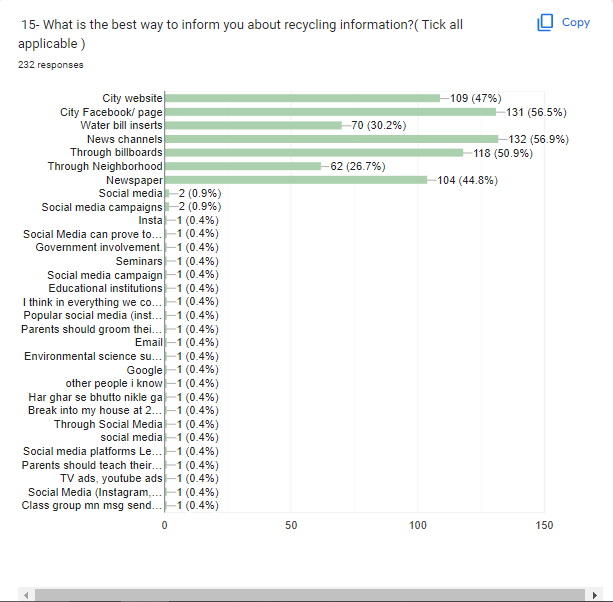












* the link of the form is given : <https://docs.google.com/forms/d/1yoXQaXyrLT0rP85wtnhz_WAksN1qUb7_Xa14bk_rX7I/edit?ts=63982056#responses>
* the csv file is also given
* and the html document of the statistical analysis is also attached

STATISTICAL ANALYSIS

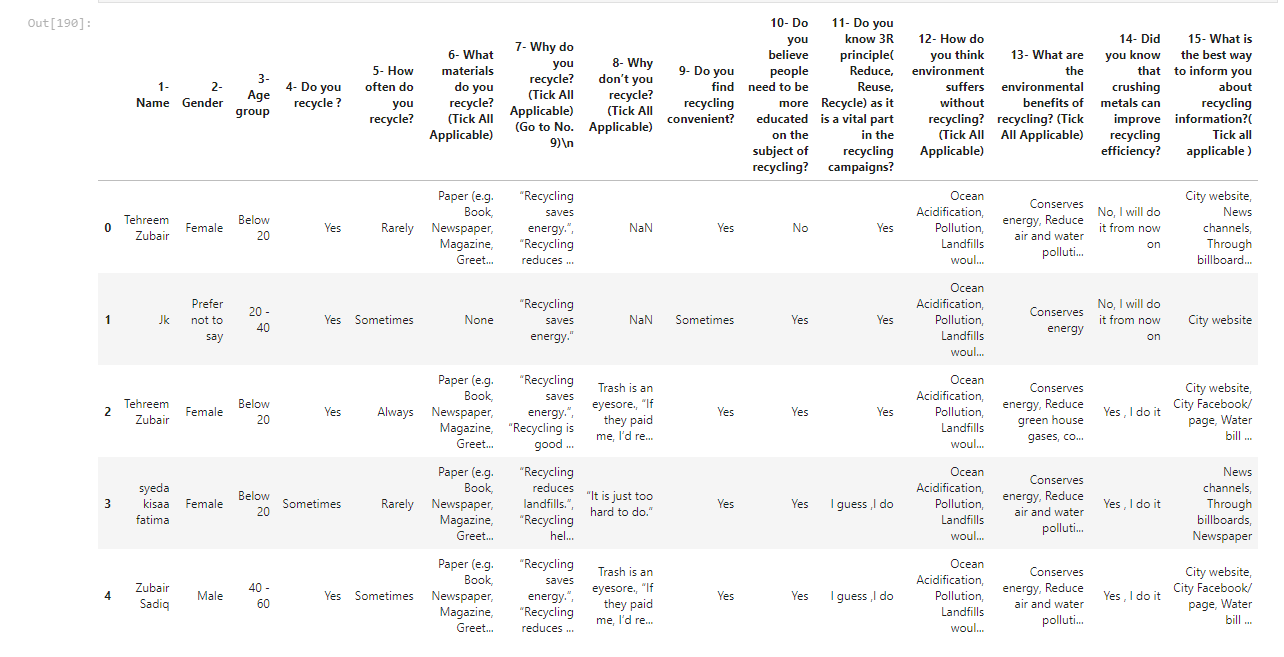
* After gathering all the responses the next step was to perform the statistical analysis of the data :
* for statistical analysis, we decided to opt for python as our programming language because it has proper data analytics and visualization tools which remained very helpful throughout our project.Python was implemented on Jupiter notebook using the kernel of anaconda.
* now the whole statistical analysis of data is given below:

The below figure shows some of the important libraries that we included

and the CSV file of the responses was included using pandas library



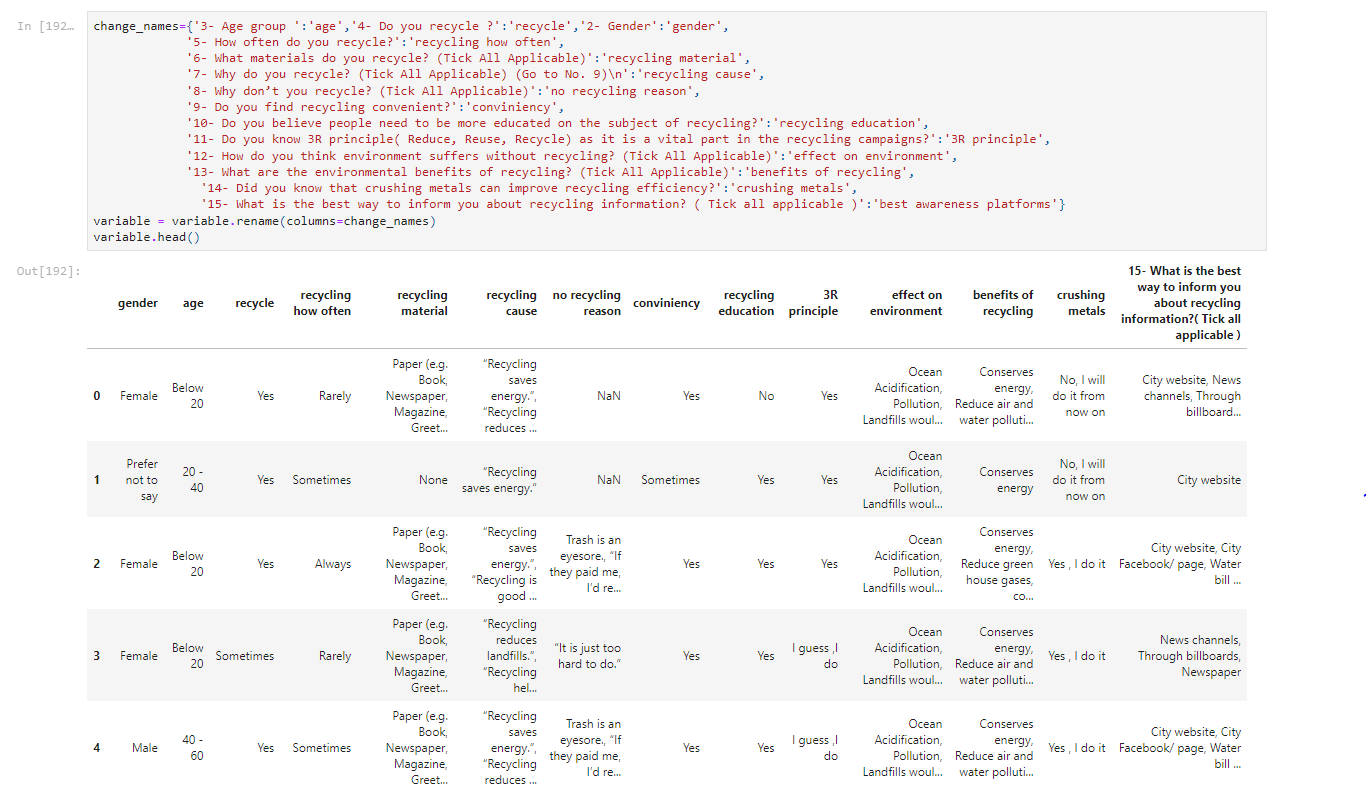
The general view of the csv file is given below:



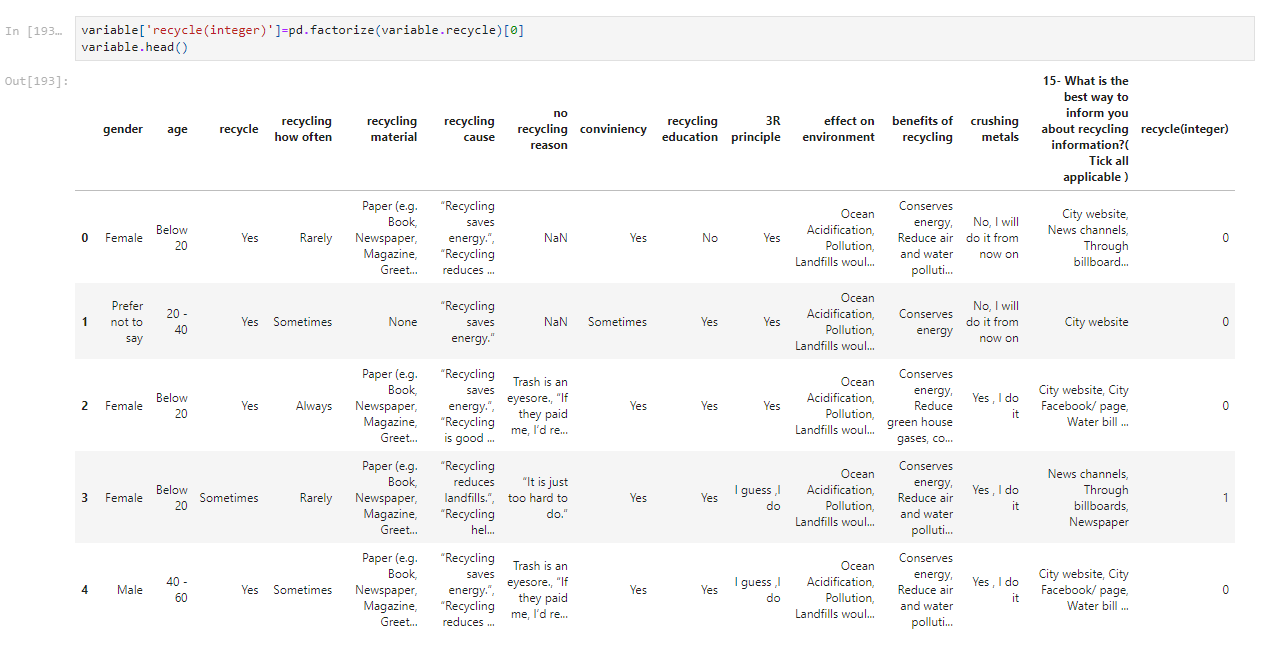
* as the ''name'' column was irrelevant so we dropped the name column

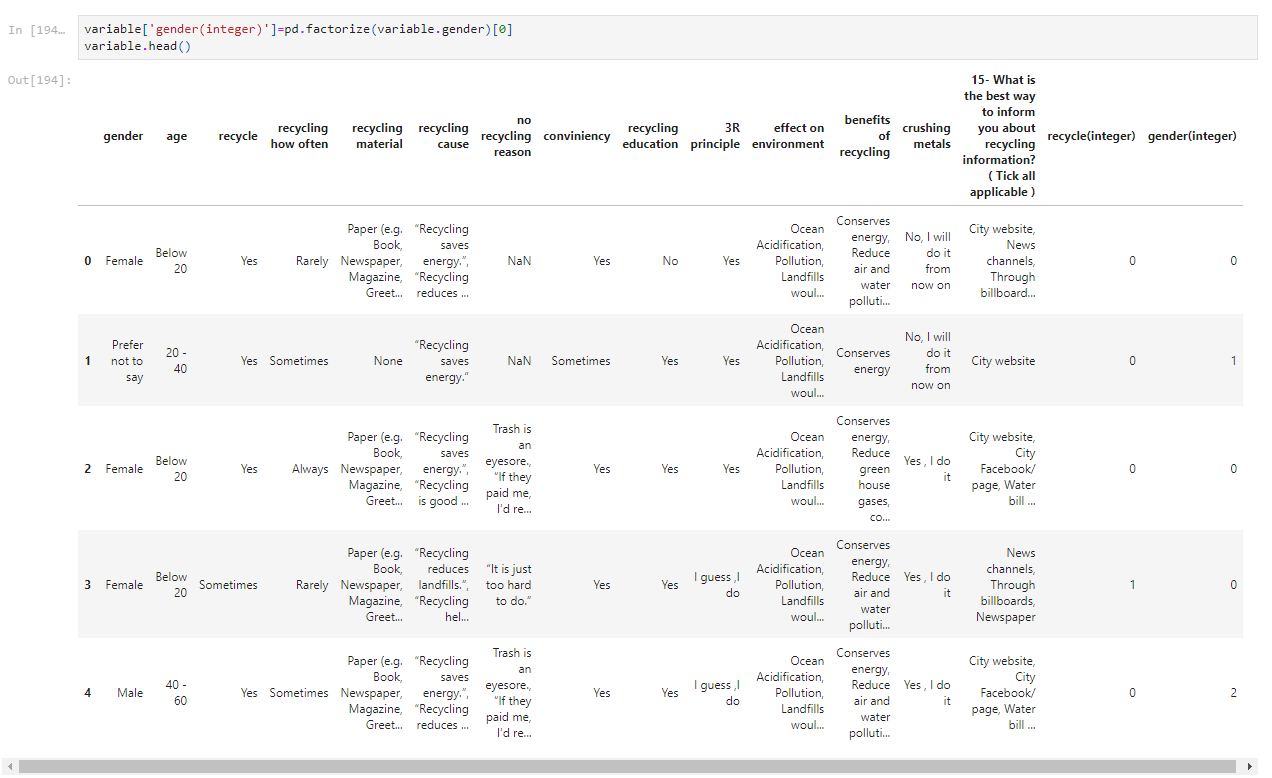


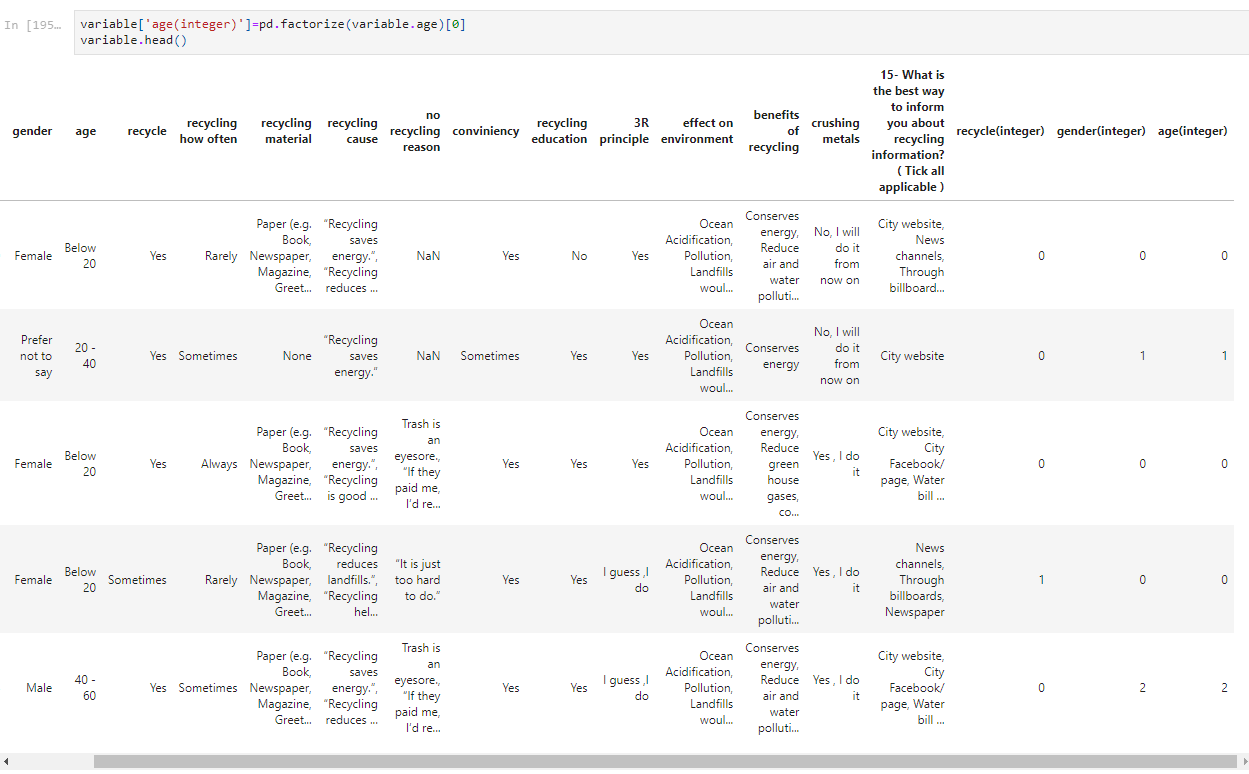
* the names of the columns were changed to shorter names for easy handling

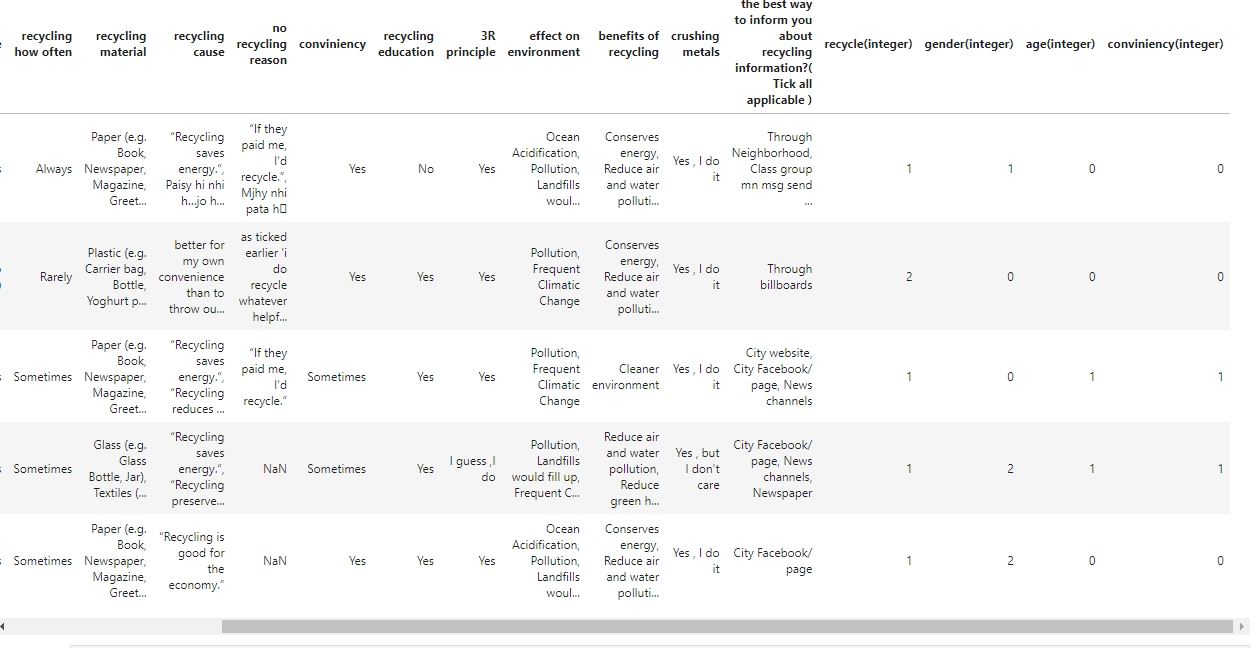


* for statistical analysis we needed data in numeric form but we only had had data in the form of strings so we did some coding and generated code for the columns like in column "recycle" ,when we coded the coded column recycle(integer) was created that contained a code for the column in numeric form .here in the given figure 1 means Sometimes , 2 means no,0 means yes.

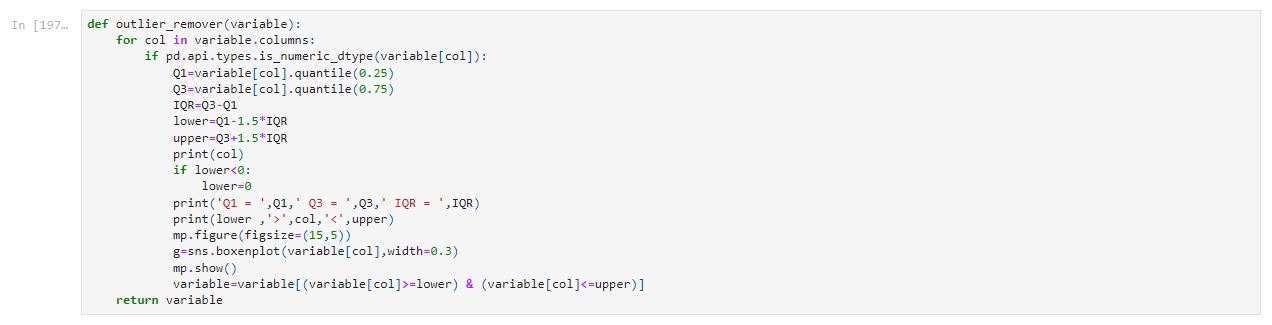


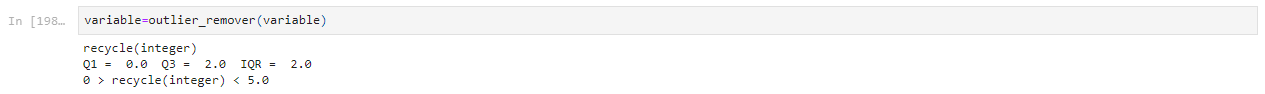


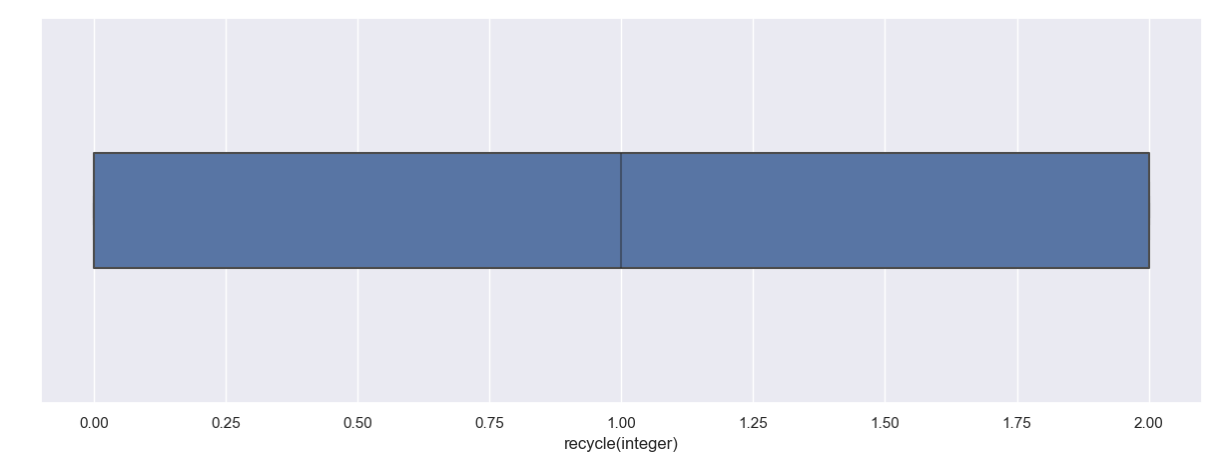




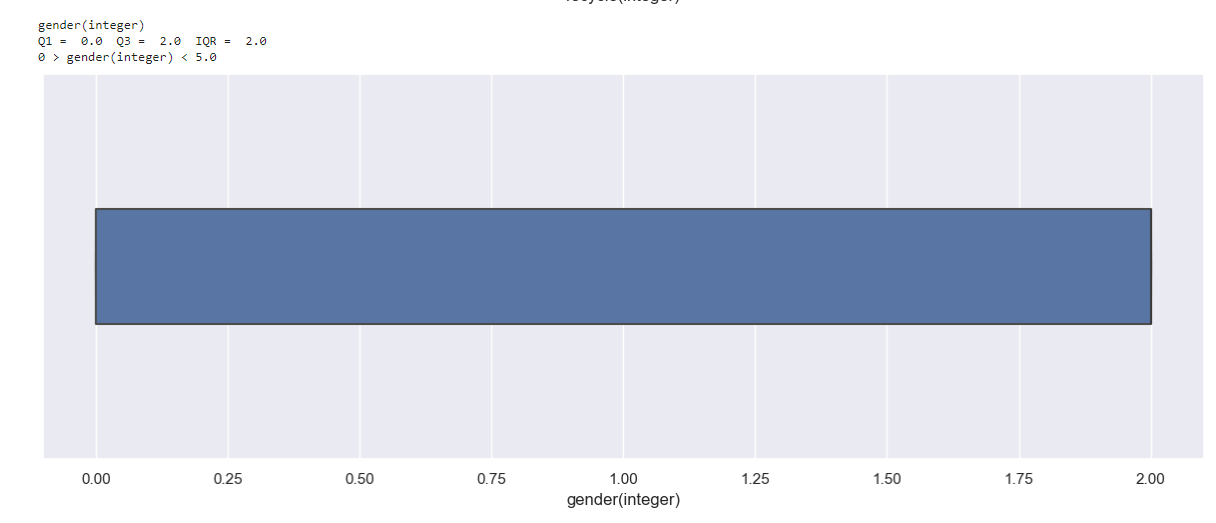
* next was outliers removal





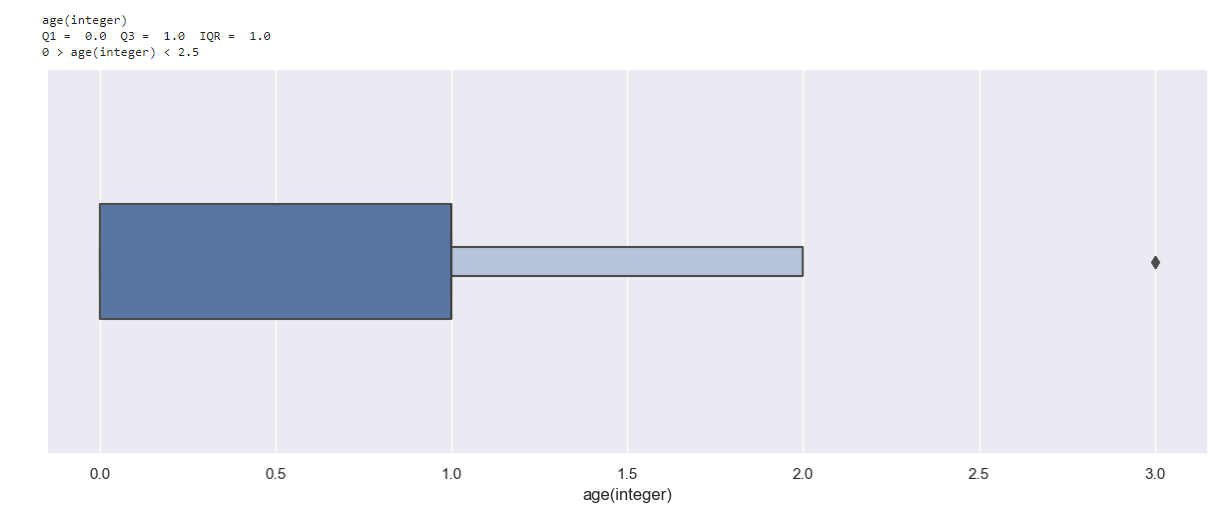


age(integer) key: 0->yes, 1->sometimes, 2->no



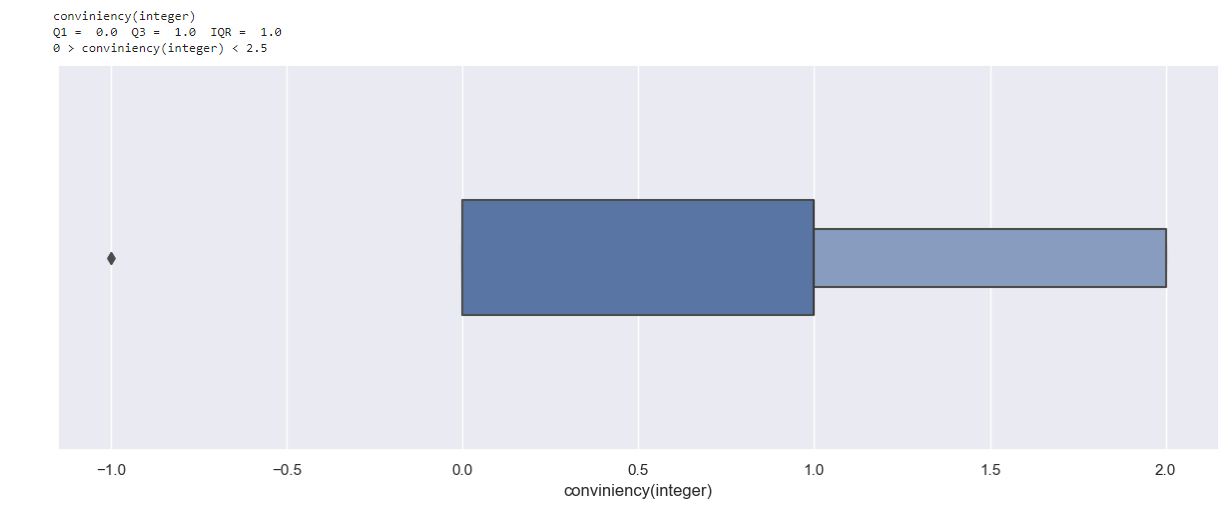
gender(integer) key:

* 0->female
* 1->prefer not to say
* 2->male



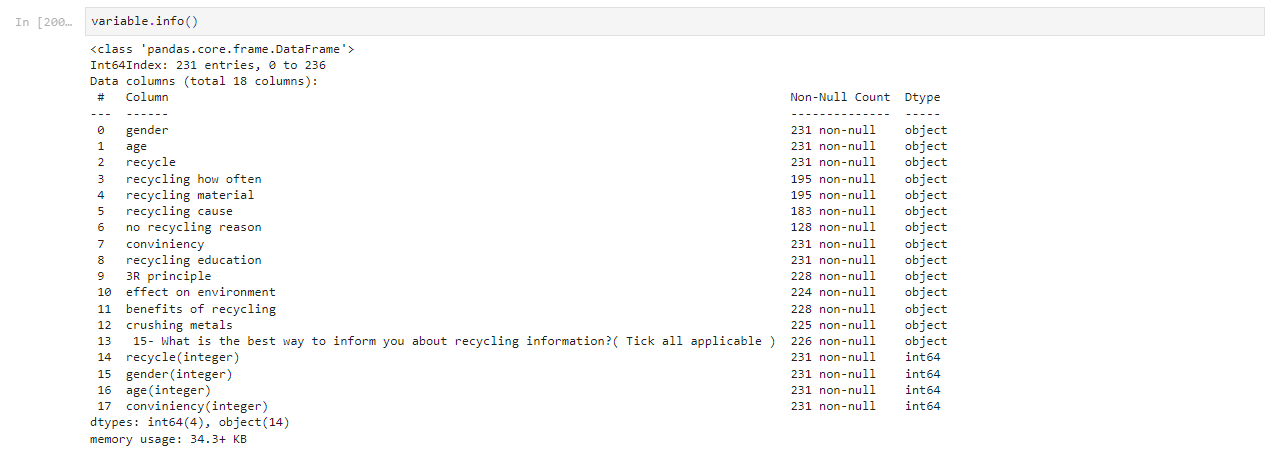
age(integer) key:

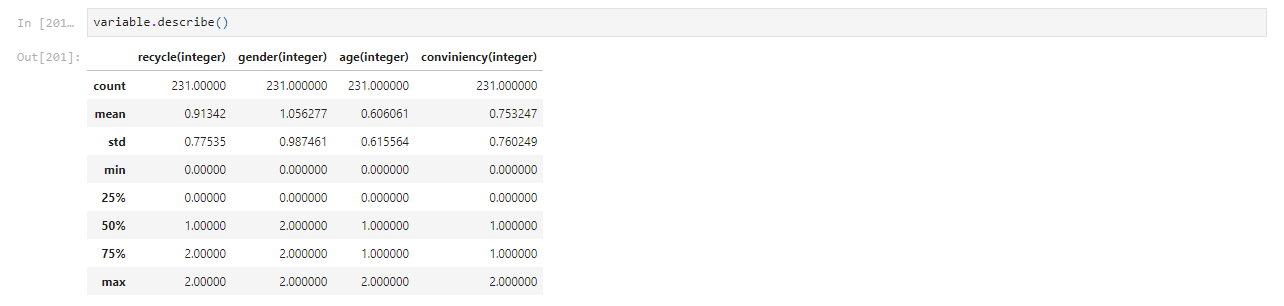
* 0->below 20
* 1->20-40
* 2->40-60
* 3->above 60



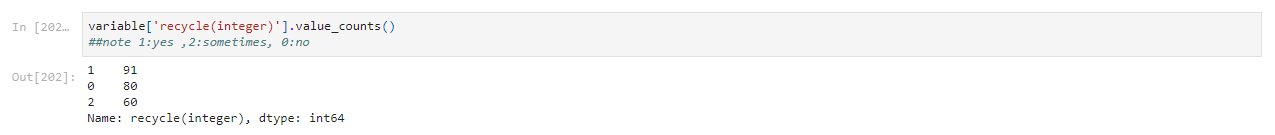
conviniency(integer) key :

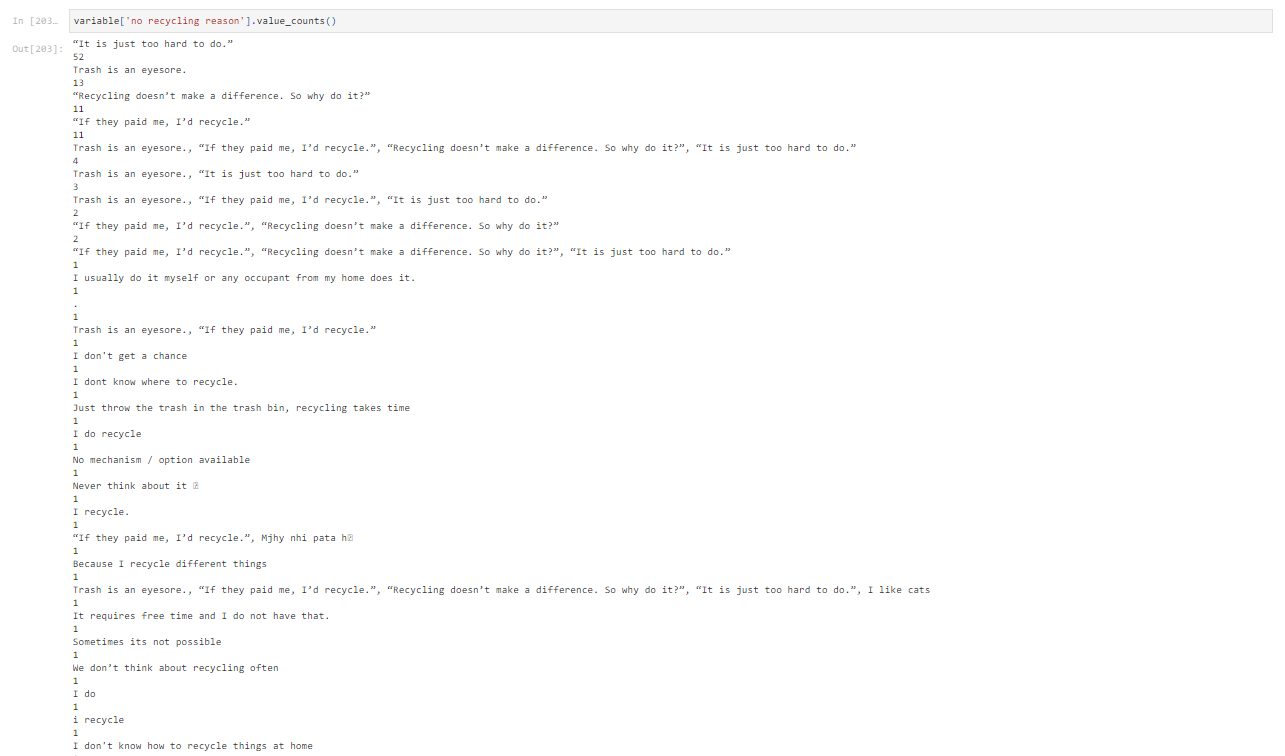
* -1->rarely
* 0->yes
* 1->sometimes
* 2->never

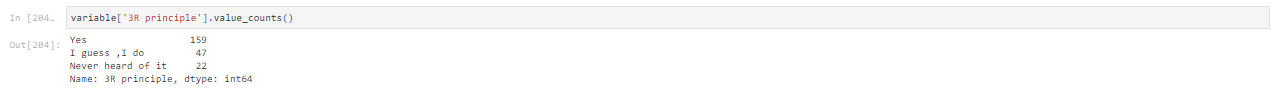




* counts of the data in each column:

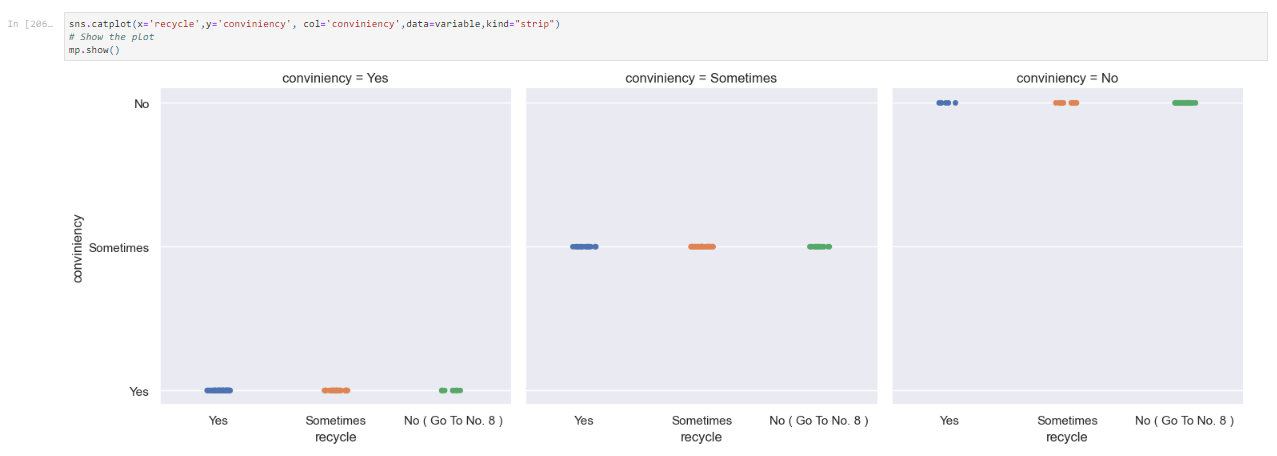




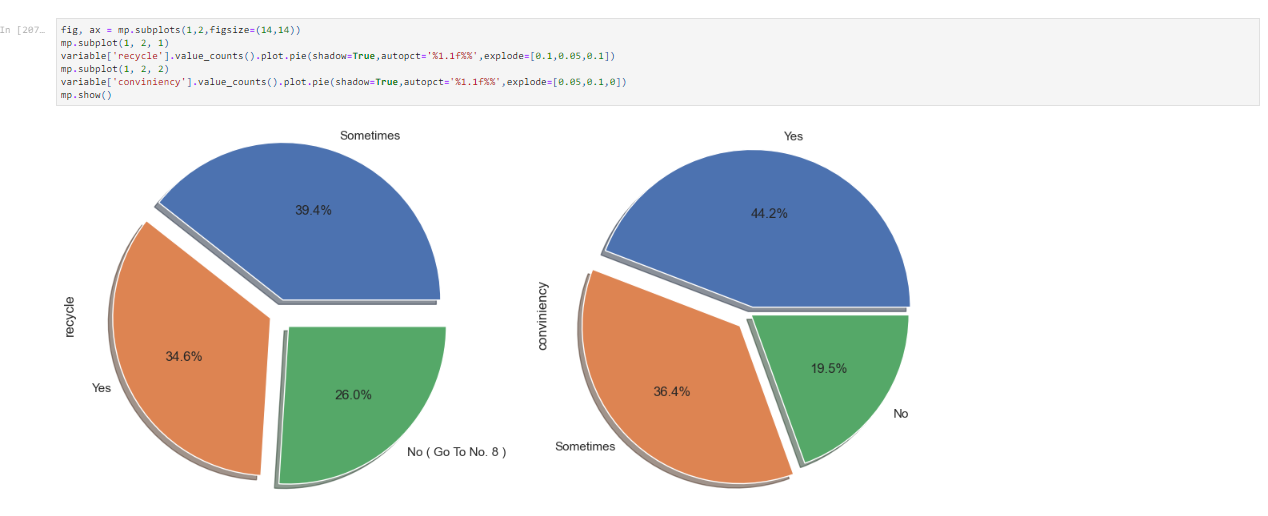


* graphs for statistical analysis
* strip plot



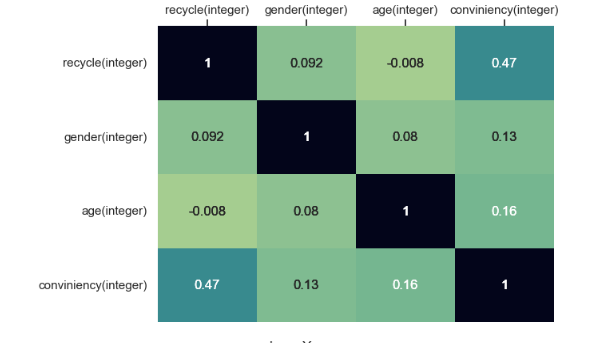


* pie charts

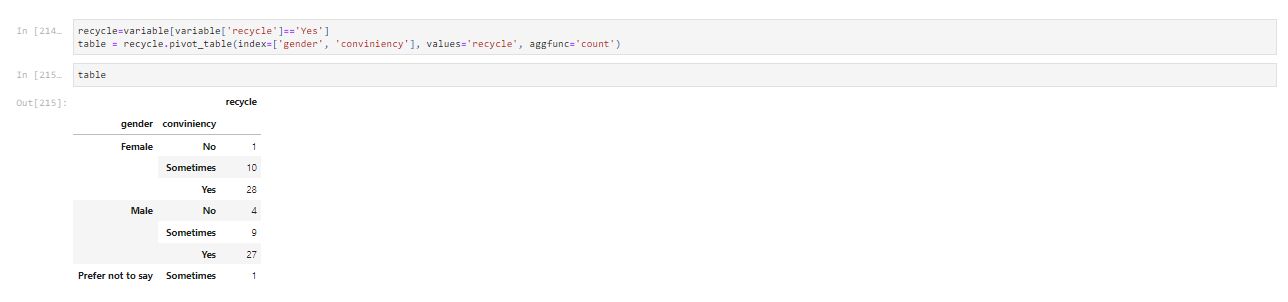


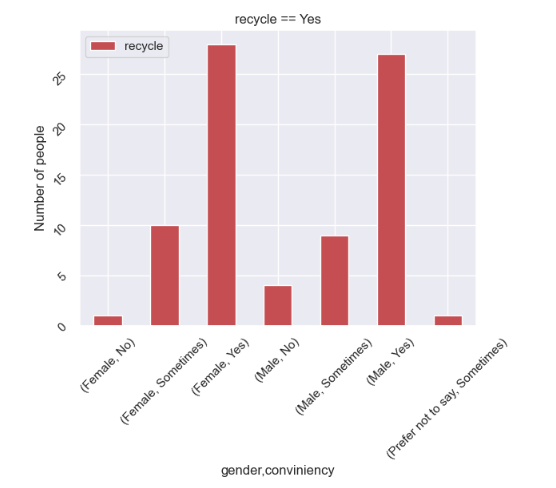
correlation diagrams:

one thing to be noted here is that 1 means strong relation 0 means no relation



* tables and bar charts





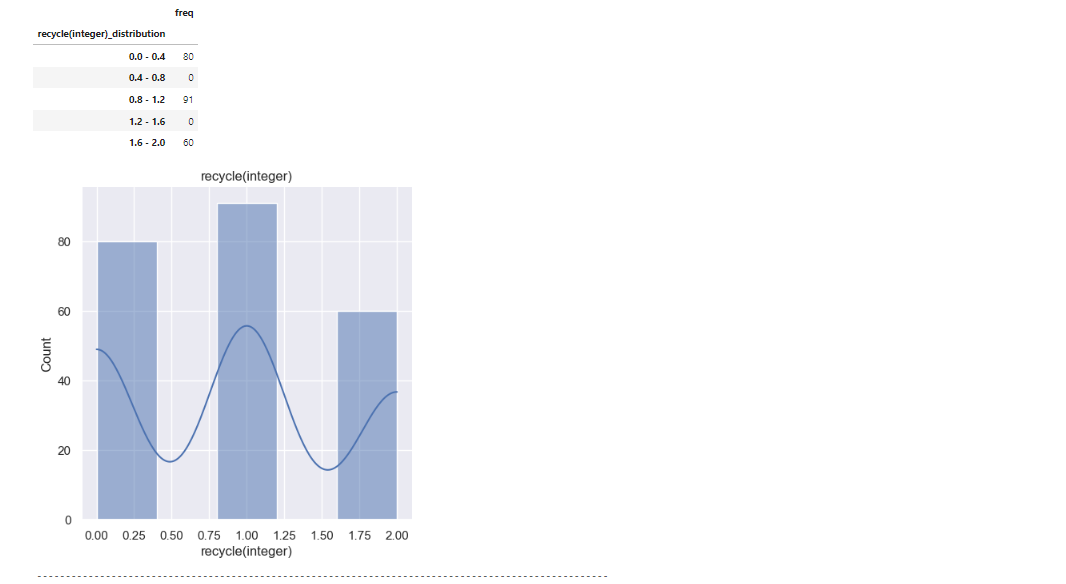


frequency distribution:

after making all the plots for the data the next step was frequency distribution of the data in numeric form

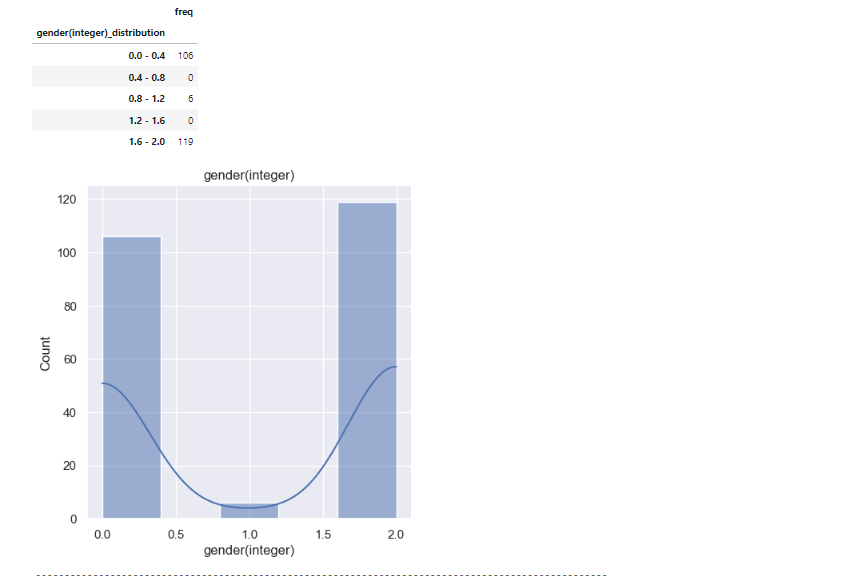
for this purpose we used data that was converted in numeric form

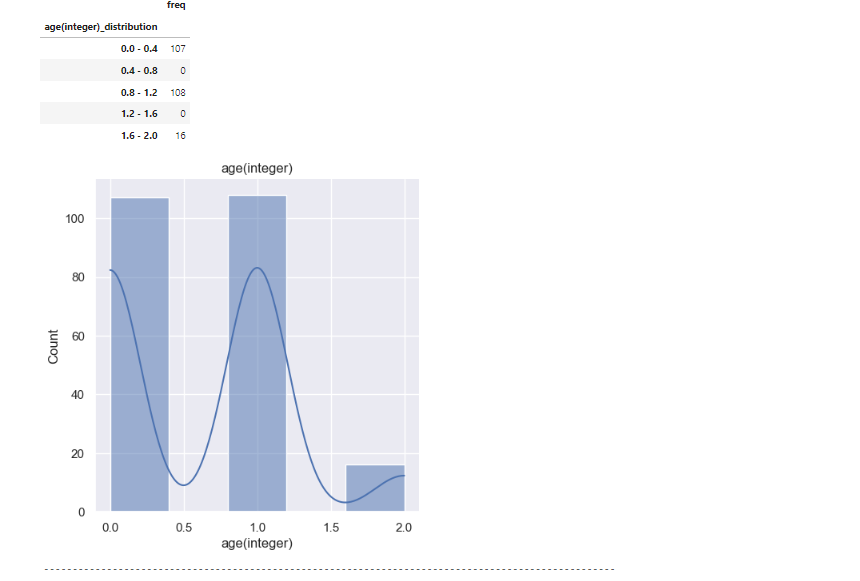


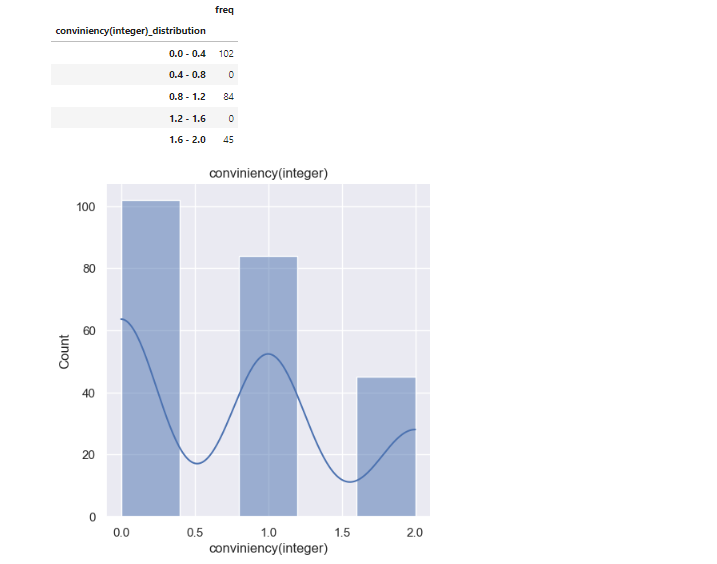


in the above diagram on x - axis 0 indicate yes, 1 indicate sometimes, 2 indicates no,

similarly all the frequency distribution plots given below are according to the coding keys given above in the outlier handling part.







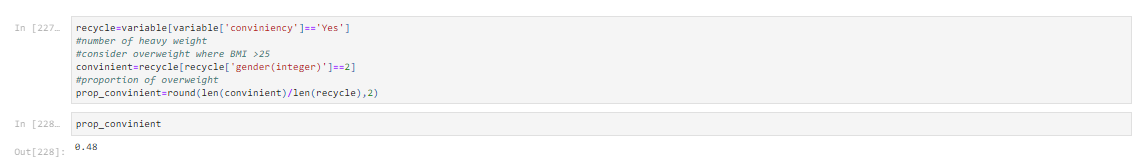
hypothesis testing:  
1. recycling is more convinient for males as compared to other genders

lets take our null hypothesis as

h0: p=theproprtion of people who find recycling convenient are males

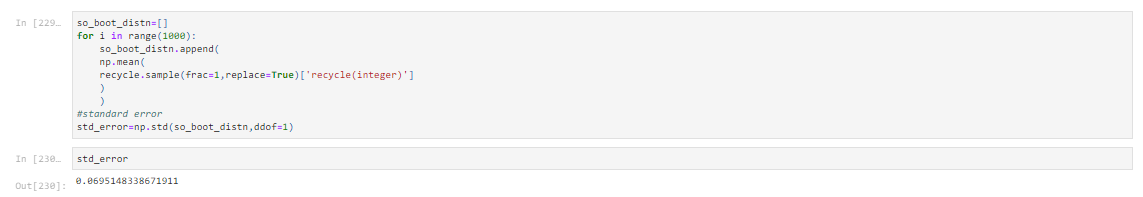
h1: p=the proportion of people who find recycling convinient are not males

the hypothesis testing of the propotion that we will use in our z and t distribution are given below:



standard error is calculated using bootstrap distribution

let us suppose the value of alpha = 0.05 i.e at 95% confidence interval





conclusion:  
p=4.22e-6<0.05 so we can say that we have sufficient evidence to reject the null hypothesis

that the people who find recycling convenient are males.

hypothesis 2:  
in the second hypothesis we used t distribution instead of t distriution

again the confidence interval is at 95% i.e alpha:0.05

for now our null hypothesis is :

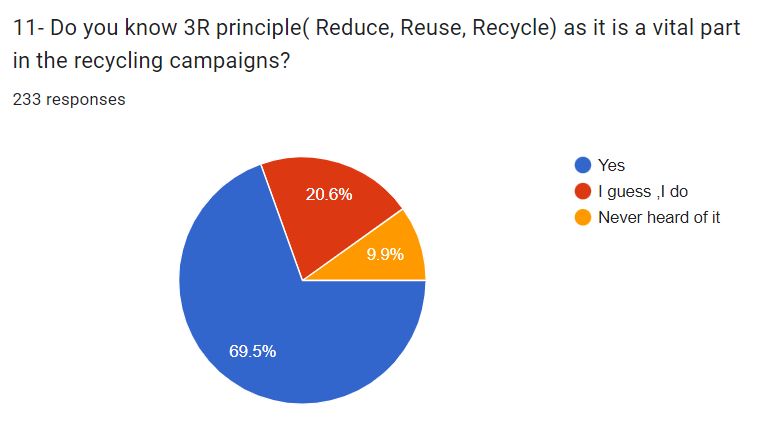
h0: those who dont recycle are not males

h1:those who recyclea re males



conclusion:  
as 8.77>0.05 so we fail to reject the null hypothesis that males find recycling inconvenient for themselves.

***Applying frequency distribution***

****

* First of all, we computed the separate frequencies of all age groups and listed them in x1 representing YES, x2 representing NEVER HEARD OF IT, and x3 representing I GUESS, I DO.
* Then calculated the x1.f(x), x2.f(x), x3.f(x) and their respective sum in bottom-most row.

The table below is frequency distribution table comprising of x1, x2, x3 & f(x).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age Limit** | **X1(YES)** | | | **X2(NEVER HEARD OF IT)** | **X3(I GUESS, I DO)** | **f(x)** | **X1.f(x)** | **X2.f(x)** | **X3.f(x)** |
| Below 20 | | 83 | 8 | | 16 | 107 | 8881 | 856 | 1712 |
| 60 Above | | 2 | 1 | | 1 | 4 | 8 | 4 | 4 |
| 40 - 60 | | 9 | 2 | | 5 | 16 | 144 | 32 | 80 |
| 20 - 40 | | 68 | 12 | | 26 | 106 | 7208 | 1272 | 2756 |
| **Grand Total** | | **162** | **23** | | **48** | **233** | **16241** | **2164** | **4552** |

**Bar chart for frequency distribution:**

**Line chart for frequency distribution:**