PES University, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE17CS20
3

B.Tech, Sem III
Session : Aug-Dec, 2018

UE17CS203 – INTRODUCTION TO DATA SCIENCE

REPORT

EXPLORATORY ANALYSIS ON
National Health and Nutrition Examination Survey

DATA SET LINK :	**https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey/home**
TEAM MEMBERS	**NAME : Pranjal Srivastava** **SRN : PES1201700208** **EMAIL ID : pranjal270499@gmail.com**
	NAME: Anujeet Dubey **SRN: PES1201701294** **EMAIL ID : anujeet221@gmail.com**

ABSTRACT

Exploratory Data Analysis: In this assignment we have taken up a task of analyzing a pre-existing data set. The data set was not in the best way possible. There were missing values because the samples refused to give the data or the sample wasn't examined properly. There were fallacies in the data set which had to be dealt with and then performing the visualization tasks. Our Data set involved the data of US citizens sampled randomly by NAHNES. The data set is about health and nutrition of the US citizens. The conclusion drawn from the data set is that people below poverty line are not the only ones with bad health though it's a parameter. People who are rich and can afford a wellbeing are suffering problems like anxiety, food poisoning, and diarrhea.

Data Set

The data set that we collected is a NHANES data set from 2013-2014. It's a large data set collected from kaggle.com

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations. NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and has the responsibility for producing vital and health statistics for the Nation.

The survey examines a nationally representative sample of about 5,000 persons each year. These persons are located in counties across the country, 15 of which are visited each year.

The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.

Findings from this survey will be used to determine the prevalence of major diseases and risk factors for diseases. Information will be used to assess nutritional status and its association with health promotion and disease prevention. NHANES findings are also the basis for national standards for such measurements as height, weight, and blood pressure. Data from this survey will be used in epidemiological studies and health sciences research, which help develop sound public health policy, direct and design health programs and services, and expand the health knowledge for the Nation

Health interviews are conducted in respondents' homes. Health measurements are performed in specially-designed and equipped mobile centers, which travel to locations throughout the country. The study team consists of a physician, medical and health technicians, as well as dietary and health interviewers. Many of the study staff are bilingual (English/Spanish).

An advanced computer system using high-end servers, desktop PCs, and wide-area networking collect and process all of the NHANES data, nearly eliminating the need for paper forms and manual coding operations. This system allows interviewers to use notebook computers with electronic pens. The staff at the mobile center can automatically transmit data into data bases through such devices as digital scales and stadiometers. Touch-sensitive computer screens let respondents enter their own responses to certain sensitive questions in complete privacy. Survey information is available to NCHS staff within 24 hours of collection, which enhances the capability of collecting quality data and increases the speed with which results are released to the public.

In each location, local health and government officials are notified of the upcoming survey. Households in the study area receive a letter from the NCHS Director to introduce the survey. Local media may feature stories about the survey.

NHANES is designed to facilitate and encourage participation. Transportation is provided to and from the mobile center if necessary. Participants receive compensation and a report of medical findings is given to each participant. All information collected in the survey is kept strictly confidential. Privacy is protected by public laws.

Data Collection Procedures

NHANES uses a complex, multistage probability design to sample the civilian, noninstitutionalized population residing in the 50 states and D.C. Sample selection for NHANES followed these stages, in order:

1. Selection of primary sampling units (PSUs), which are counties or small groups of contiguous counties.
2. Selection of segments within PSUs that constitute a block or group of blocks containing a cluster of households.
3. Selection of specific households within segments.
4. Selection of individuals within a household.

The data set includes the following :

SEQUENCE - Respondent sequence number

Variable Name:
SEQUENCE
Target:
Both males and females 0 YEARS - 150 YEARS

Gender

Variable Name:
GENDER
Target:
Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative
1	Male	5003	5003
2	Female	5172	10175

.	Missing	0	10175

AGE - Age in years at screening

Variable Name:
AGE
English Text:
Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.
Target:
Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative
0 to 79	Range of Values	9823	9823
80	80 years of age and over	352	10175
.	Missing	0	10175

IN_MONTHS - Age in months at screening

Variable Name:
IN_MONTHS
Target:
Both males and females

Code or Value	Value Description	Count	Cumulative
0 to 24	Range of Values	673	673
.	Missing	9502	10175

RACE - Race/Hispanic origin

Variable Name:
RACE
Target:
Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative

1	Mexican American	1730	1730
2	Other Hispanic	960	2690
3	Non-Hispanic White	3674	6364
4	Non-Hispanic Black	2267	8631
5	Other Race - Including Multi-Racial	1544	10175
.	Missing	0	10175

HISPANIC_GROUP - Race/Hispanic origin w/ NH Asian

Variable Name:
HISPANIC_GROUP
English Text:
Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category
Target:
Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative
1	Mexican American	1730	1730
2	Other Hispanic	960	2690
3	Non-Hispanic White	3674	6364
4	Non-Hispanic Black	2267	8631
6	Non-Hispanic Asian	1074	9705
7	Other Race - Including Multi-Racial	470	10175
.	Missing	0	10175

BIRTH_PLACE - Country of birth

Variable Name:
DMDBORN4
English Text:
In what country {were you/was SP} born?
Target:
Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative
1	Born in 50 US states or Washington, DC	8262	8262
2	Others	1908	10170
77	Refused	4	10174
99	Don't Know	1	10175
.	Missing	0	10175

EDUCATION_LEVEL_UNDER - Education level - Children/Youth 6-19

Variable Name:
EDUCATION_LEVEL_UNDER
English Text:
What is the highest grade or level of school {you have/SP has} completed or the highest degree {you have/s/he has} received?
Target:
Both males and females 6 YEARS - 19 YEARS

Code or Value	Value Description	Count	Cumulative
0	Never attended / kindergarten only	277	277
1	1st grade	237	514
2	2nd grade	231	745

3	3rd grade	215	960
4	4th grade	234	1194
5	5th grade	212	1406
6	6th grade	178	1584
7	7th grade	203	1787
8	8th grade	189	1976
9	9th grade	194	2170
10	10th grade	162	2332
11	11th grade	186	2518
12	12th grade, no diploma	45	2563
13	High school graduate	137	2700
14	GED or equivalent	9	2709
15	More than high school	81	2790
55	Less than 5th grade	3	2793
66	Less than 9th grade	9	2802
77	Refused	0	2802
99	Don't Know	1	2803
.	Missing	7372	10175

EDUCATION_LEVEL - Education level - Adults 20+

Variable Name:
EDUCATION_LEVEL
Target:
Both males and females 20 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative
1	Less than 9th grade	455	455
2	9-11th grade (Includes 12th grade with no diploma)	791	1246
3	High school graduate/GED or equivalent	1303	2549
4	Some college or AA degree	1770	4319
5	College graduate or above	1443	5762
7	Refused	2	5764
9	Don't Know	5	5769
.	Missing	4406	10175

MARITAL_STATUS - Marital status

Variable Name:
MARITAL_STATUS
Target:
Both males and females 20 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative
1	Married	2965	2965
2	Widowed	436	3401
3	Divorced	659	4060
4	Separated	177	4237

5	Never married	1112	5349
6	Living with partner	417	5766
77	Refused	2	5768
99	Don't Know	1	5769
.	Missing	4406	10175

FAMILY_SIZE - Total number of people in the Family

Variable Name:
FAMILY_SIZE
Target:
Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative
1	1	1297	1297
2	2	1610	2907
3	3	1737	4644
4	4	2027	6671
5	5	1723	8394
6	6	961	9355
7	7 or more people in the Family	820	10175
.	Missing	0	10175

TOTAL_FAMILY_INCOME - Annual family income

Variable Name:
TOTAL_FAMILY_INCOME
Target:
Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative
1	$ 0 to $ 4,999	375	375
2	$ 5,000 to $ 9,999	467	842
3	$10,000 to $14,999	691	1533
4	$15,000 to $19,999	706	2239
5	$20,000 to $24,999	897	3136
6	$25,000 to $34,999	1182	4318
7	$35,000 to $44,999	913	5231
8	$45,000 to $54,999	740	5971
9	$55,000 to $64,999	502	6473
10	$65,000 to $74,999	381	6854
12	$20,000 and Over	219	7073
13	Under $20,000	129	7202
14	$75,000 to $99,999	835	8037
15	$100,000 and Over	1701	9738
77	Refused	246	9984
99	Don't know	68	10052

.	Missing	123	10175

RATIO_TO_POVERTY_LINE - Ratio of family income to poverty

Variable Name:
RATIO_TO_POVERTY_LINE
Target:
Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
0 to 4.99	Range of Values	7974	7974	
5	Value greater than or equal to 5.00	1416	9390	
.	Missing	785	10175	

SEQUENCE_2 – Medication Sequences

Variable Name:
SEQUENCE ACCORDING TO MEDICAL HEALTHS
Target:
Both males and females 0 YEARS - 150 YEARS

HEALTH – Medication Sequences

Variable Name:
HEALTH
Target:
Both males and females 0 YEARS - 150 YEARS

MEDICATION– PRESCRIBED MEDICINE OR HEALTHY

Variable Name:
MEDICATION
Target:
Both males and females 0 YEARS - 150 YEARS

DRUG_NAME – Drugs consumed

Variable Name:
DRUG_NAME
Target:
Both males and females 0 YEARS - 150 YEARS

Introduction

We started our research to select a data set which needed cleaning and gave good inferences. Adding on to that we wanted a data set that makes a difference to the society. The data set Targets the entire US population which makes it even more interesting to study.Our work was to first analyse the most important heading in our data set and go through the questionnaire that was asked to the sample. Then we headed towards the task of cleaning the data set using the python modules available. Next we did our part by making it more visuaally attractive by using data visualizing techniques of python. We asked some questions about the populationa dn the sample space like:
Mean age of the sample space?
What gender exists more in our sample?
What is the ratio of the races in our sample?
What is the mean education level of students currently pursuing studies?
What is the qualification of elderly men?
What is the average size of families?
What is the average range of Salaries of our data?
Average ratio to poverty line?
Analyzing the medical condition!
We were able to extract answers to such questions using our data set after cleaning it.

Processing (Data Cleaning)

- Age: There were some samples with ages that were beyond the expected values. So we cut down the ages of the data set to max 89. Beyond that we considered it as 89.

- In months: The age of our samples in months were not completely justified and there were many values missing. Our first task was to fill the missing values using the age column. Then take out the outliers

- Birth Place: There were many samples where the birth place were unknown thus was cleaned by assuming them to be non-citizens of USA.

- Marital Status: marital status for many of the respondents was unknown and hence was assumed to be unmarried.

- Education Level: Missing values of the data set were replaced by the mean.

- Family Size: The missing values were replaced by median

- Total Family Income: The missing values were replaced by median

- Ratio to poverty line: The missing values were replaced in accordance to the median value of the ratio to poverty line of the respondents with the same range of total family income.

- Health: The values were replaced by the number of prescribed medication.

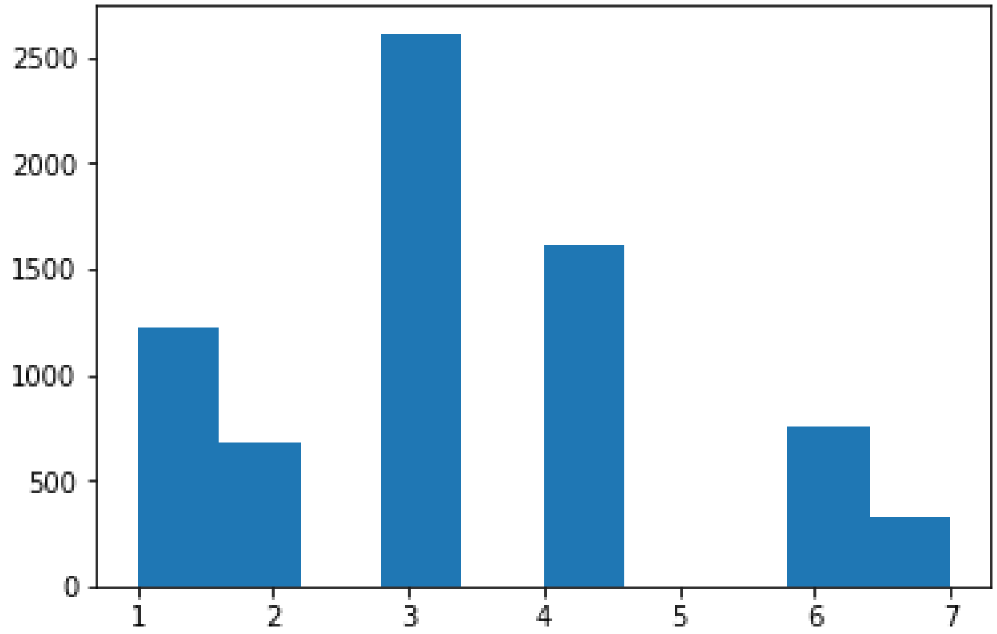
EXPLORATORY ANALYSIS:

INFERENCES:

- **MEAN AGE**: 34.3 years
- **GENDER**: 49.1% MALE ,50.9% FEMALE
- **RACE**: 17% MEXICAN AMERICAN ,36% NON-HISPANIC WHITE,22.2% NON-HISPANIC BLACK
- **FAMILY SIZE**: 3.74
- **TOTALFAMILY INCOME(AVERAGE)**:62,992 USD
- **MARITAL STATUS:** 30% MARRIED, 7.3% DIVORCED,53.8% UNMARRIED
- **EDUCATION**: 16% UNDERGRADUATE,8% NOT EVEN 12STD
- **BELOW POVERTY LINE**: 19.2%
- **HEALTH:**71.8% HEALTHY

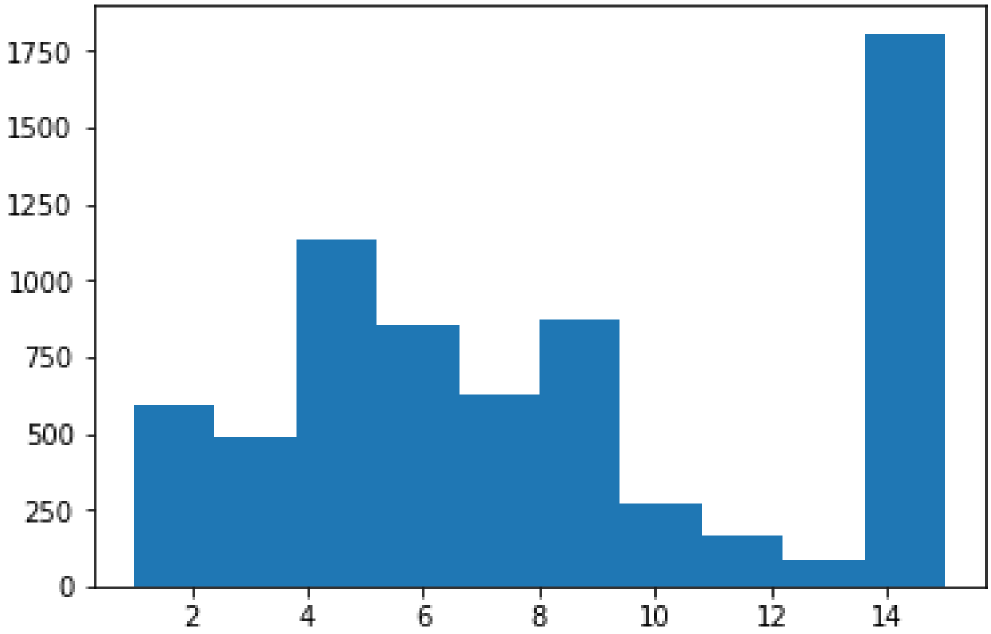
Here are different inferential Graphs:

RACE :



Code or Value	Value Description	Count	Cumulative
1	Mexican American	1730	1730
2	Other Hispanic	960	2690
3	Non-Hispanic White	3674	6364
4	Non-Hispanic Black	2267	8631
6	Non-Hispanic Asian	1074	9705
7	Other Race - Including Multi-Racial	470	10175

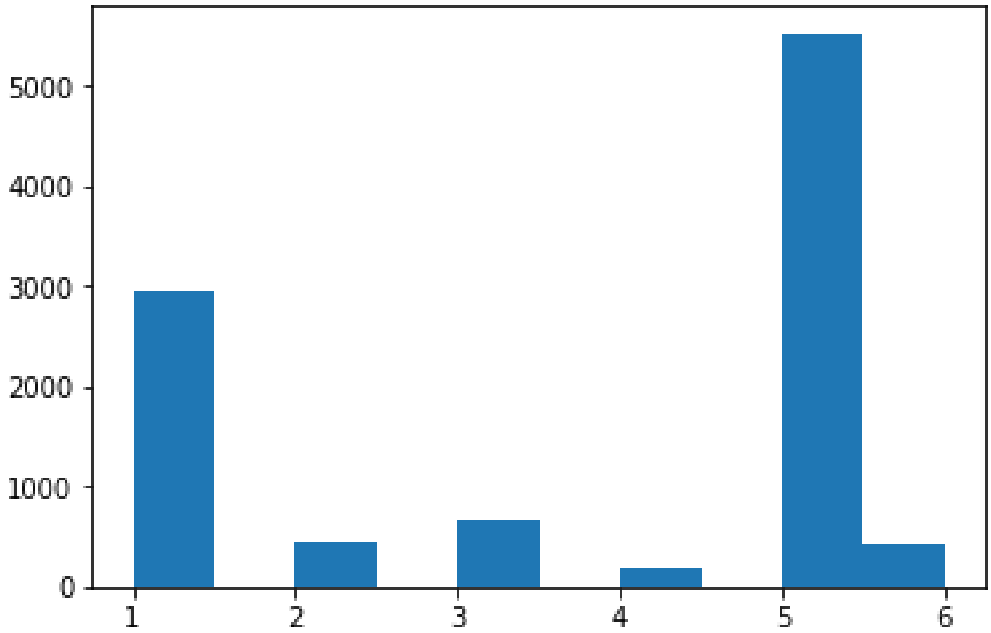
TOTAL FAMILY INCOME:



1	$ 0 to $ 4,999	375	375
2	$ 5,000 to $ 9,999	467	842
3	$10,000 to $14,999	691	1533
4	$15,000 to $19,999	706	2239
5	$20,000 to $24,999	897	3136
6	$25,000 to $34,999	1182	4318
7	$35,000 to $44,999	913	5231
8	$45,000 to $54,999	740	5971
9	$55,000 to $64,999	502	6473
10	$65,000 to $74,999	381	6854
12	$20,000 and Over	219	7073
13	Under $20,000	129	7202

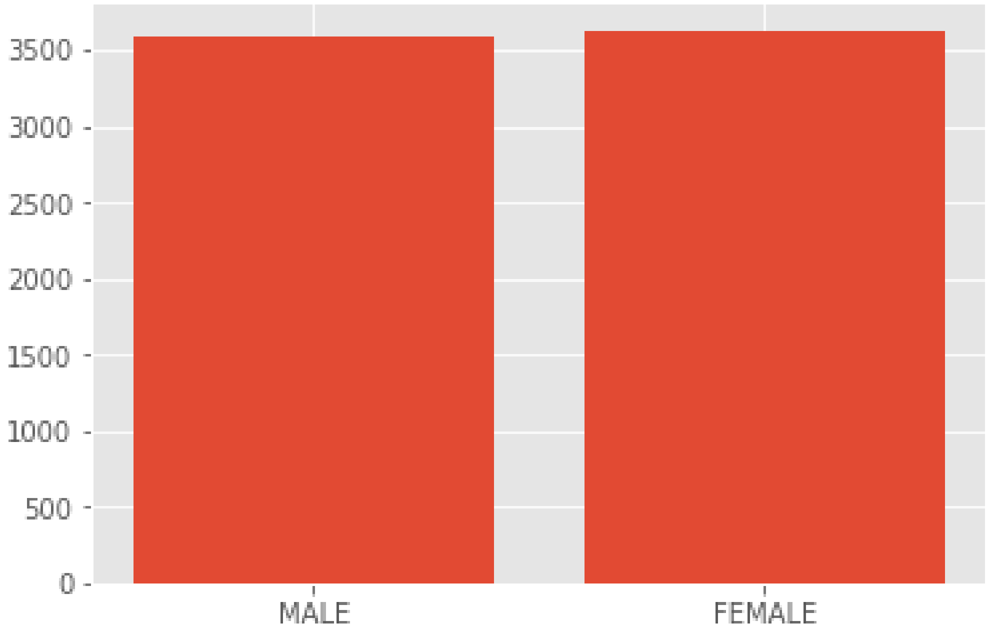
14	$75,000 to $99,999	835	8037
15	$100,000 and Over	1701	9738

MARITAL STATUS:



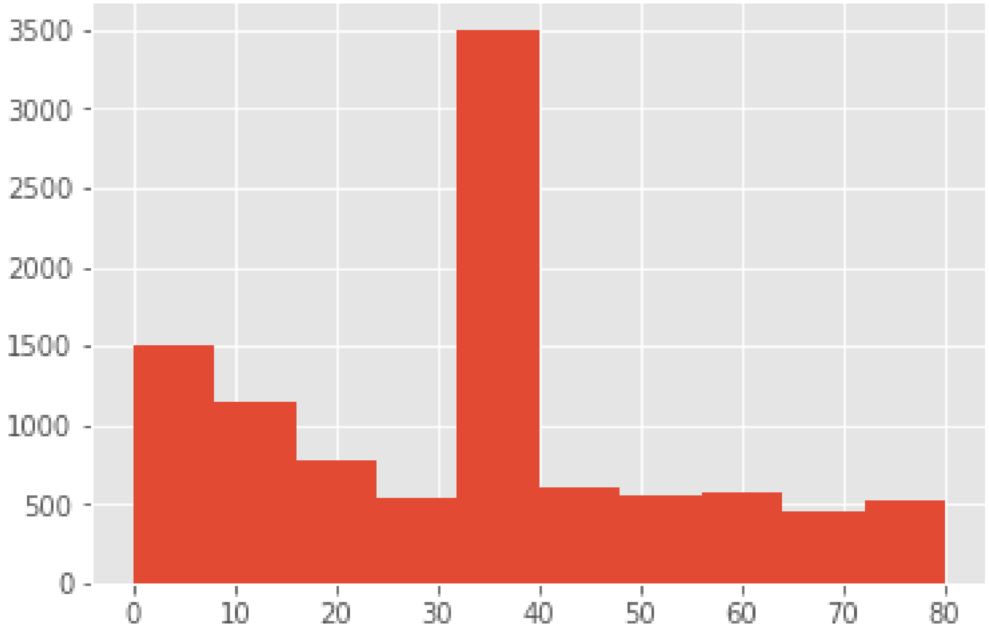
Code or Value	Value Description	Count	Cumulative
1	Married	2965	2965
2	Widowed	436	3401
3	Divorced	659	4060
4	Separated	177	4237
5	Never married	5112	9349
6	Living with partner	417	9766

GENDER:



Code or Value	Value Description	Count	Cumulative
1	Male	5003	5003
2	Female	5172	10175

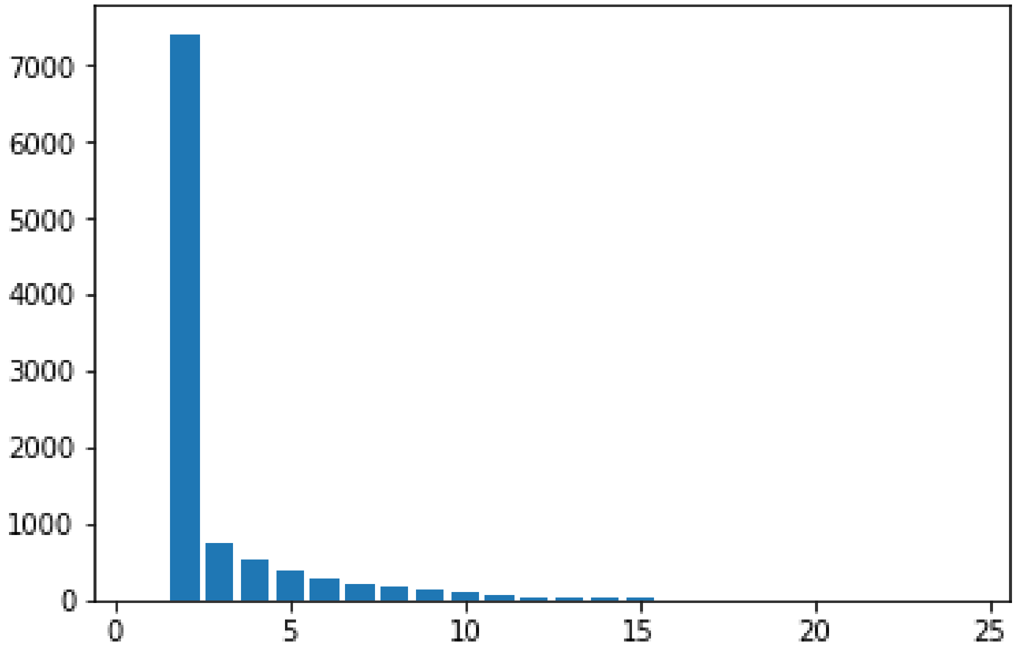
AGE:



Code or Value	Value Description	Count	Cumulative

0 to 79	Range of Values	9823	9823
80	80 years of age and over	352	10175

HEALTH:



1:HEALTHY or takes 1 medicines
Rest is the number of people taking the number of prescribed medicines

CONCLUSION

- We concluded a lot many details about the target population using a good representation sample. This study was very helpful to know the various aspects of Data Science in a way that we faced a lot of problems solving small issues. This made us realize the importance of a good unbiased sample as well as the value of various small parameters in a data set.
- Majorly people in the given data set are taking medication due to some or the other reason.
- But many of them take 1 or 2 prescribed medicines which we considered as healthy. For the rest of the respondents the reason of being unhealthy vary with different sectors, regions and lifestyle.
- We concluded that the missing data values when replaced with some amount using logics gives us better representation of our target population

- With this small exercise we managed to develop data exploration mindset. We enhanced our ability to read, clean and visualize the data sets. It helped us learn about the importance of sampling methods used and the importance of good representation of every section of our population.