Project: Investigate a dataset of Medical Appointment

Dataset I used for Analysis is Medical Appointment dataset.

Name of dataset: noshowappointments-kagglev2-may-2016.csv

Questions I would pose for an analysis:

1. What is the shape of dataset?
2. What is the attributes behavior in the dataset?
3. How much difference are there for "having a Scholarship" and "not having a scholarship"?
4. Age wise distribution of Scholarship.
5. How is all the diseases looks like visually for the given patients? like what age group is suffering for which kind of disease more?
6. Which age-group patients doesn't show up?
7. Which gender have a more no-show ratio?
8. Gender vise distribution with Age distribution for No-show.

The steps I used to investigate the dataset.

Preparing the dataset

I got this data from given Udacity PDF.

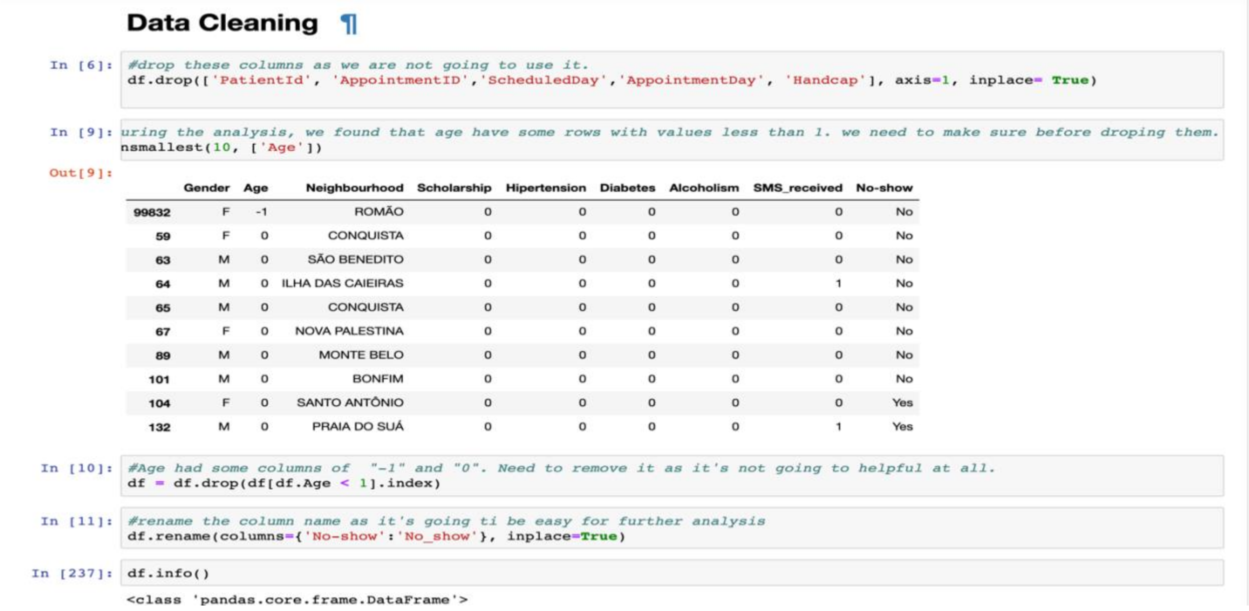
The description was on Kaggle website. So, I started to list all the data available on this page, understand their meaning. After having inventoried the data available on Kaggle.com and understanding the meaning of each data item, I started the data selection phase, that is, the data I want to keep for my Analysis.

Here are the data I want to keep:
Gender
Age
Neighborhood
Scholarship
Hipretension
Diabetes
Alcoholism
No_show
Firstly, I imported necessary packages and libraries which we need for analysis.

Data Cleaning

I have been thinking of several solutions to fix this dataset problem with missing values as follows:
Drop unnecessary columns
Rename columns
Delete the line with the missing values
Fill empty fields with specific values

You can see that below:



Data Cleaning ¶

```
In [6]: #drop these columns as we are not going to use it.
        df.drop(['PatientId', 'AppointmentID','ScheduledDay','AppointmentDay', 'Handcap'], axis=1, inplace= True)

In [9]: uring the analysis, we found that age have some rows with values less than 1. we need to make sure before droping them.
        nsmallest(10, ['Age'])
Out[9]:
```

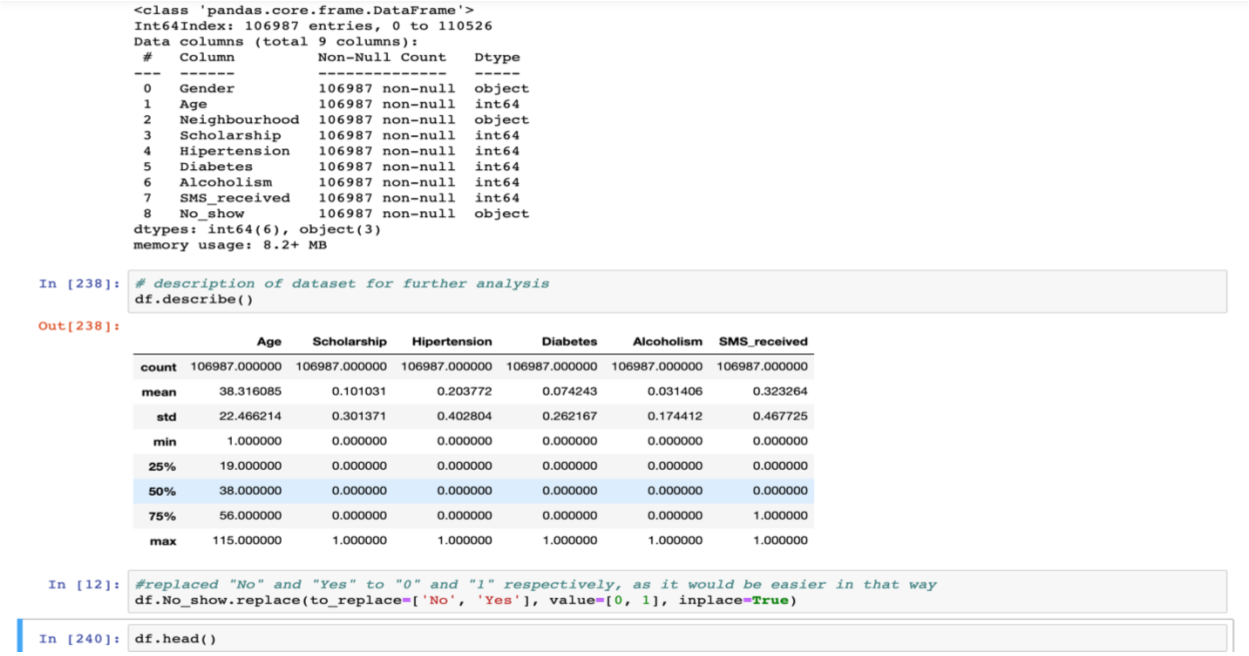
	Gender	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	SMS_received	No-show
99832	F	-1	ROMÃO	0	0	0	0	0	No
59	F	0	CONQUISTA	0	0	0	0	0	No
63	M	0	SÃO BENEDITO	0	0	0	0	0	No
64	M	0	ILHA DAS CAIEIRAS	0	0	0	0	1	No
65	M	0	CONQUISTA	0	0	0	0	0	No
67	F	0	NOVA PALESTINA	0	0	0	0	0	No
89	M	0	MONTE BELO	0	0	0	0	0	No
101	M	0	BONFIM	0	0	0	0	0	No
104	F	0	SANTO ANTÔNIO	0	0	0	0	0	Yes
132	M	0	PRAIA DO SUÁ	0	0	0	0	1	Yes

```
In [10]: #Age had some columns of  "-1" and "0". Need to remove it as it's not going to helpful at all.
         df = df.drop(df[df.Age < 1].index)

In [11]: #rename the column name as it's going ti be easy for further analysis
         df.rename(columns={'No-show':'No_show'}, inplace=True)

In [237]: df.info()

          <class 'pandas.core.frame.DataFrame'>
```



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 106987 entries, 0 to 110526
Data columns (total 9 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   Gender         106987 non-null  object
 1   Age            106987 non-null  int64
 2   Neighbourhood  106987 non-null  object
 3   Scholarship    106987 non-null  int64
 4   Hipertension   106987 non-null  int64
 5   Diabetes       106987 non-null  int64
 6   Alcoholism     106987 non-null  int64
 7   SMS_received   106987 non-null  int64
 8   No_show        106987 non-null  object
dtypes: int64(6), object(3)
memory usage: 8.2+ MB

In [238]: # description of dataset for further analysis
          df.describe()
Out[238]:
```

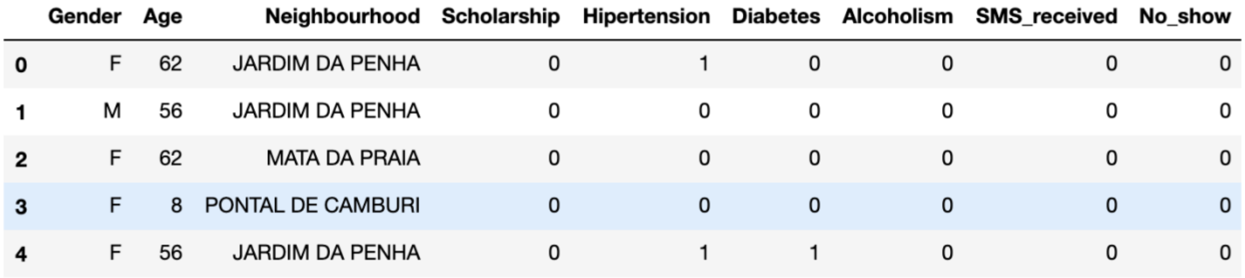
	Age	Scholarship	Hipertension	Diabetes	Alcoholism	SMS_received
count	106987.000000	106987.000000	106987.000000	106987.000000	106987.000000	106987.000000
mean	38.316085	0.101031	0.203772	0.074243	0.031406	0.323264
std	22.466214	0.301371	0.402804	0.262167	0.174412	0.467725
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	19.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	38.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	56.000000	0.000000	0.000000	0.000000	0.000000	1.000000
max	115.000000	1.000000	1.000000	1.000000	1.000000	1.000000

```
In [12]: #replaced "No" and "Yes" to "0" and "1" respectively, as it would be easier in that way
         df.No_show.replace(to_replace=['No', 'Yes'], value=[0, 1], inplace=True)

In [240]: df.head()
```

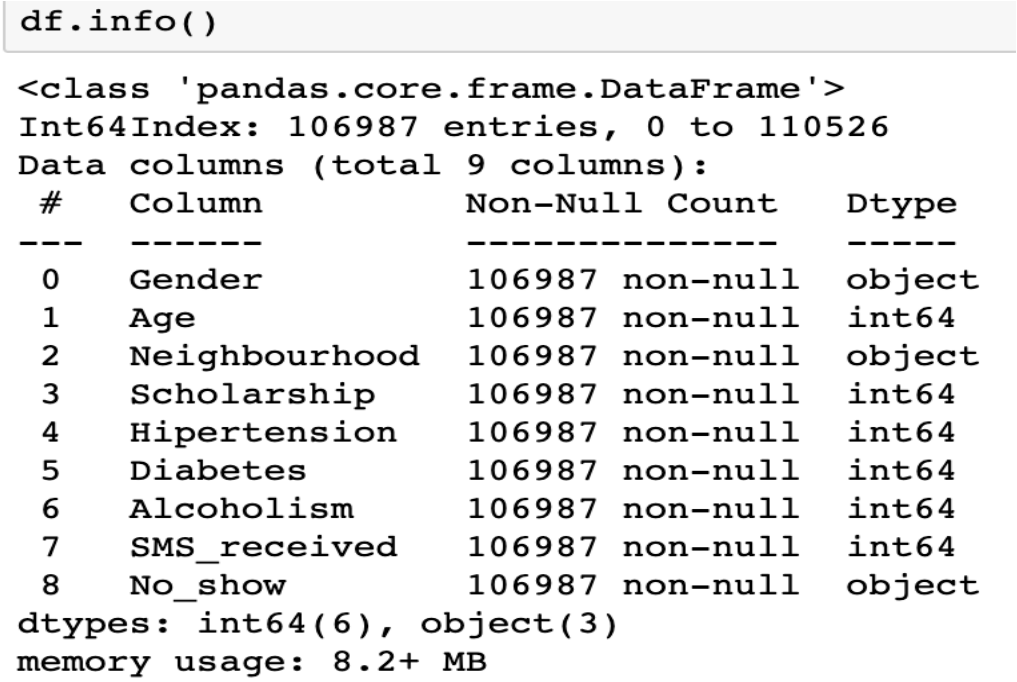
Data analysis

I thus recovered the dataset with the Python script. With the Pandas library, it is possible to have an overview of the dataset and by applying functions like info(), describe() and head(), I could check the contents of my dataset.
With the head() function applied to my dataset, I display a part of the dataset. I have displayed the first 5 data as below:



	Gender	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	SMS_received	No_show
0	F	62	JARDIM DA PENHA	0	1	0	0	0	0
1	M	56	JARDIM DA PENHA	0	0	0	0	0	0
2	F	62	MATA DA PRAIA	0	0	0	0	0	0
3	F	8	PONTAL DE CAMBURI	0	0	0	0	0	0
4	F	56	JARDIM DA PENHA	0	1	1	0	0	0

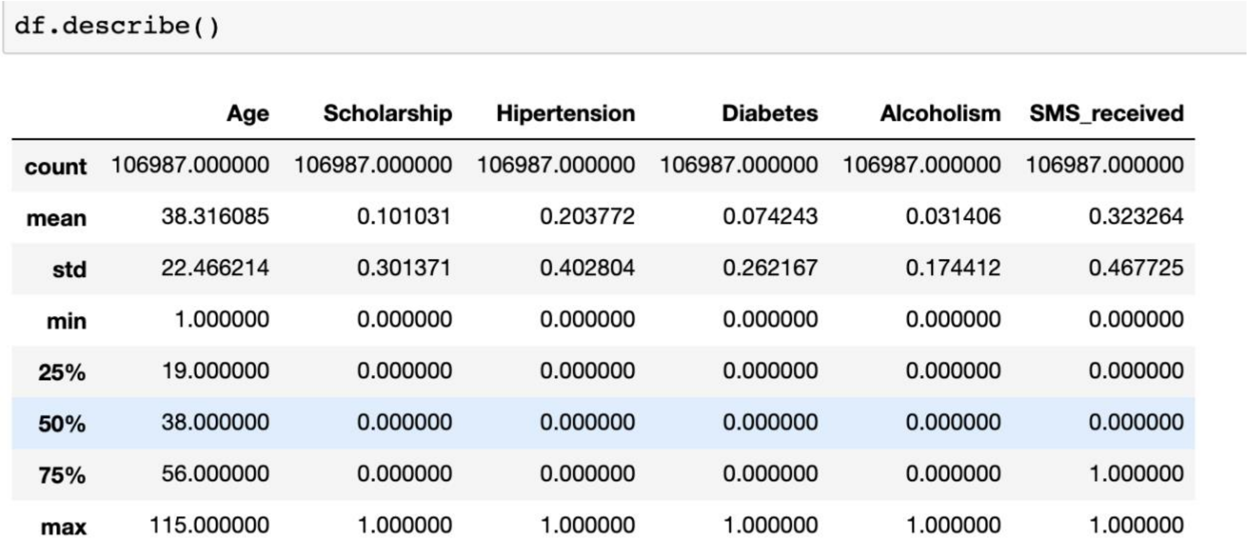
Then I apply the info() function on my dataset:



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 106987 entries, 0 to 110526
Data columns (total 9 columns):
 #   Column         Non-Null Count   Dtype 
---  ------         --------------   ----- 
 0   Gender         106987 non-null  object
 1   Age            106987 non-null  int64 
 2   Neighbourhood  106987 non-null  object
 3   Scholarship    106987 non-null  int64 
 4   Hipertension   106987 non-null  int64 
 5   Diabetes       106987 non-null  int64 
 6   Alcoholism     106987 non-null  int64 
 7   SMS_received   106987 non-null  int64 
 8   No_show        106987 non-null  object
dtypes: int64(6), object(3)
memory usage: 8.2+ MB
```

Then, I display the statistical summary of the dataset with describe().

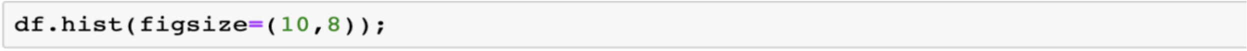


```
df.describe()
```

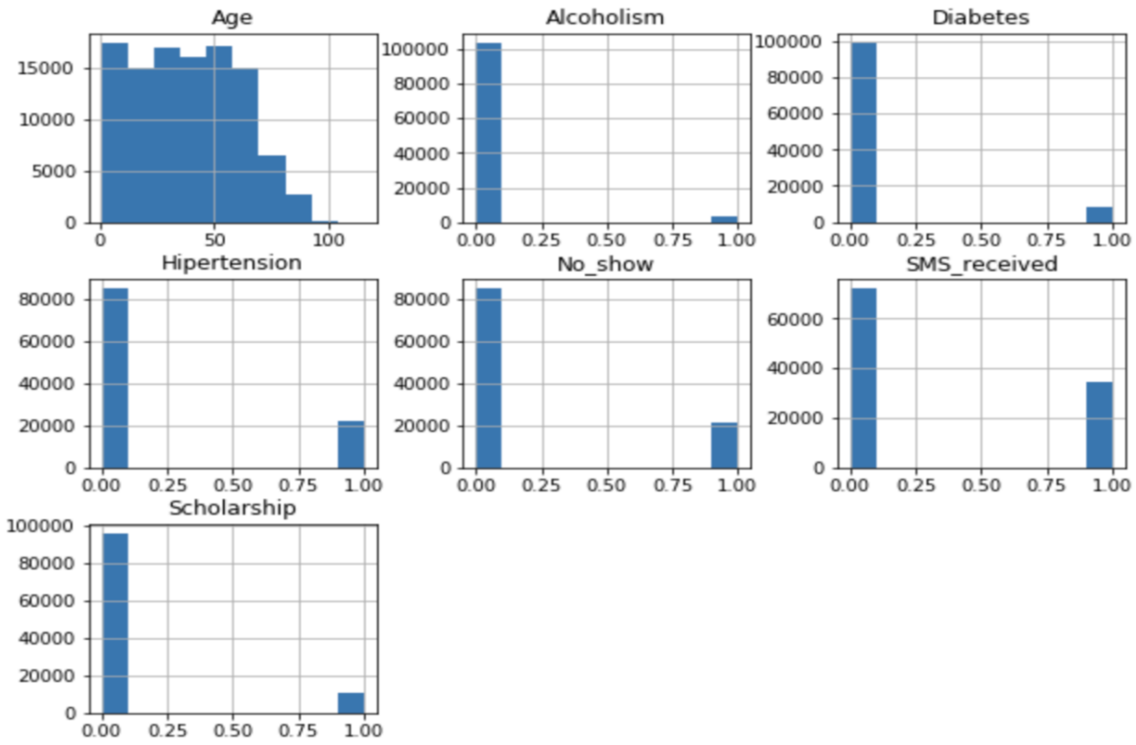
	Age	Scholarship	Hipertension	Diabetes	Alcoholism	SMS_received
count	106987.000000	106987.000000	106987.000000	106987.000000	106987.000000	106987.000000
mean	38.316085	0.101031	0.203772	0.074243	0.031406	0.323264
std	22.466214	0.301371	0.402804	0.262167	0.174412	0.467725
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	19.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	38.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	56.000000	0.000000	0.000000	0.000000	0.000000	1.000000
max	115.000000	1.000000	1.000000	1.000000	1.000000	1.000000

With this summary, I have access to a lot of information about my dataset, such as number of rows, average data, standard deviation, minimum, maximum, and all three quartiles.

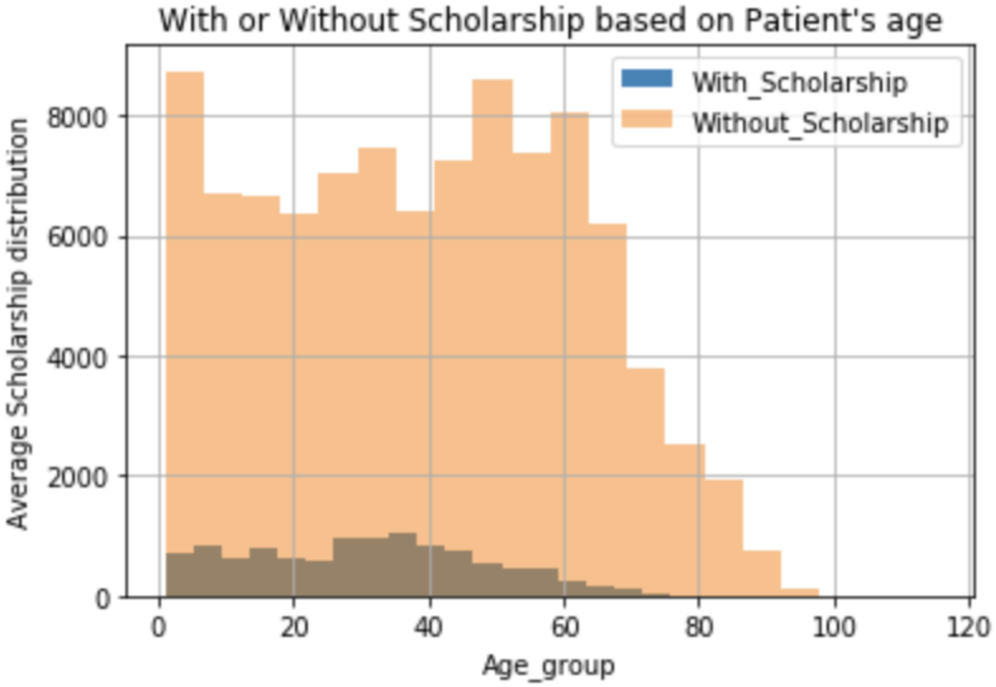
Then, I displayed histograms for all the attributes in dataset to roughly calculate the behavior of data.



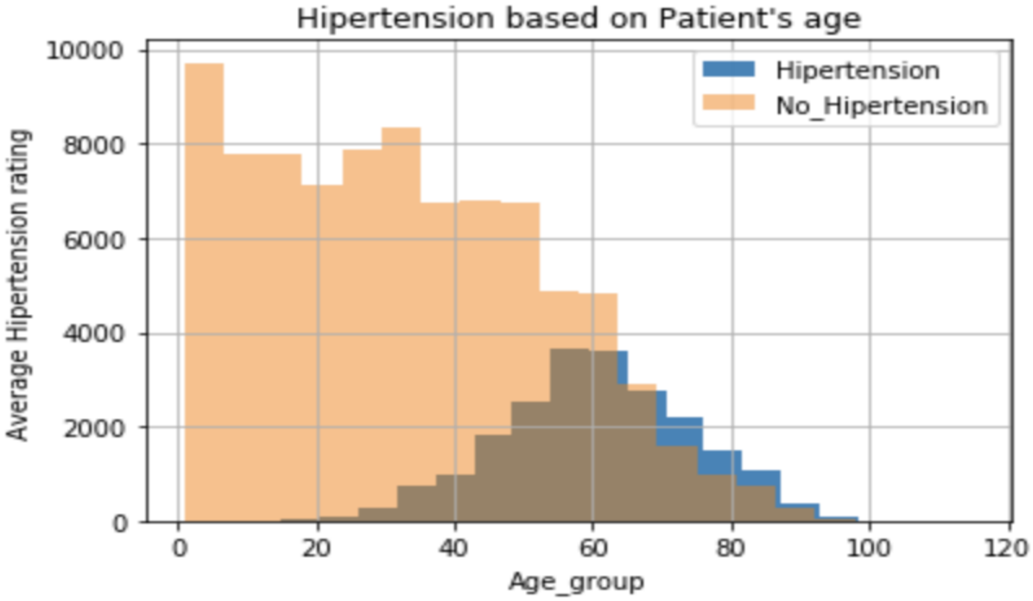
```
df.hist(figsize=(10,8));
```



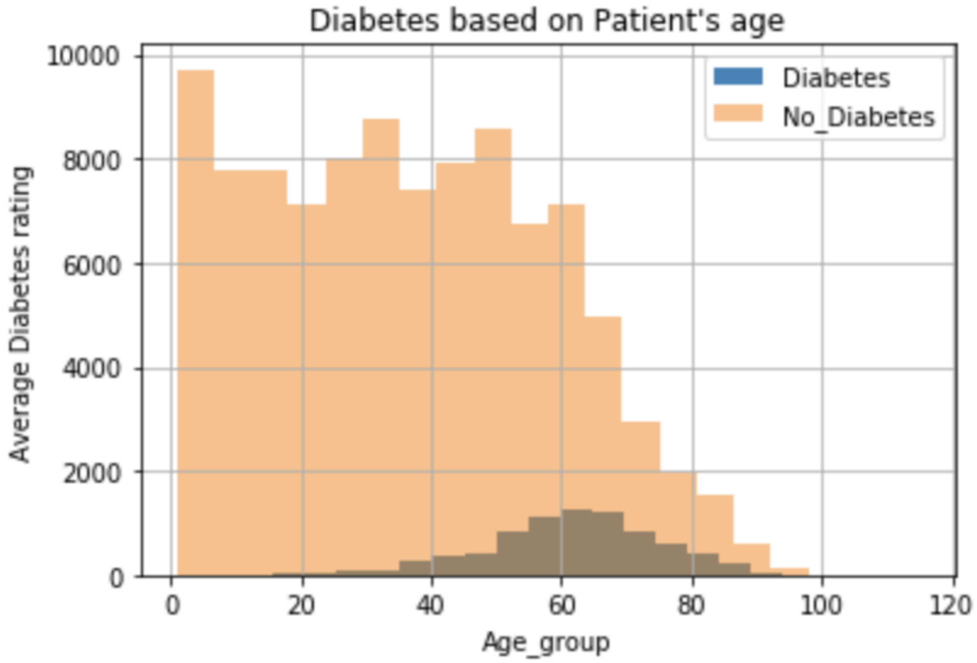
As histogram shows us there is a big difference between having a scholarship and not having a scholarship. Let's see the age distribution for scholarship.



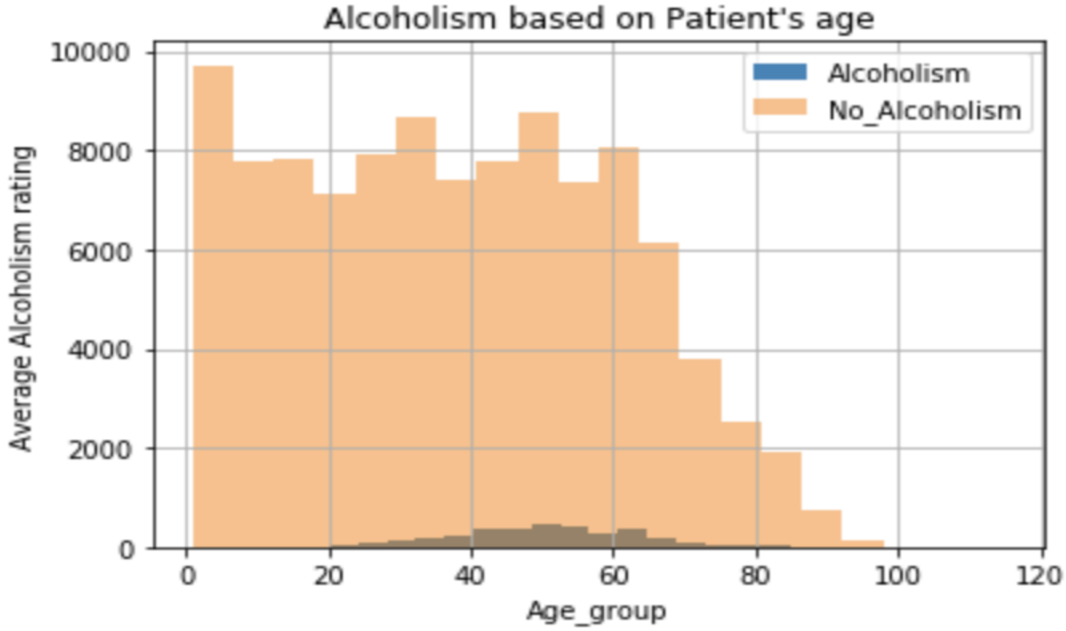
Now, it's important to see that what kind of diseases are there in patients? in which age group? This is for Hipertension:



Let’s see for Diabetes:



Now see for Alcoholism:

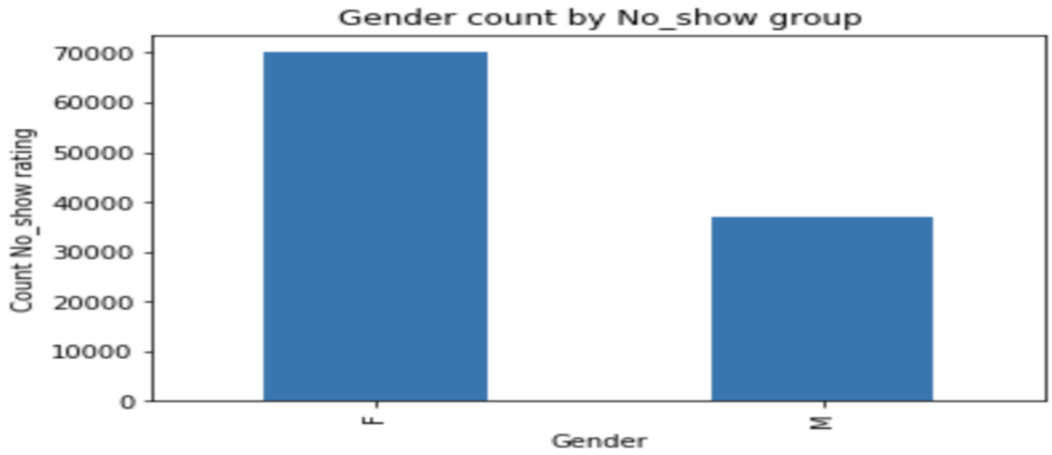


Which age-group patients doesn't show up?

We see that there is a high concentration of results for below 38 years patients, which means that in most cases, young people are more likely to not show up on appointment days.

In this graph, we can conclude that the patients more likely to show up when their age is more than 38 years.

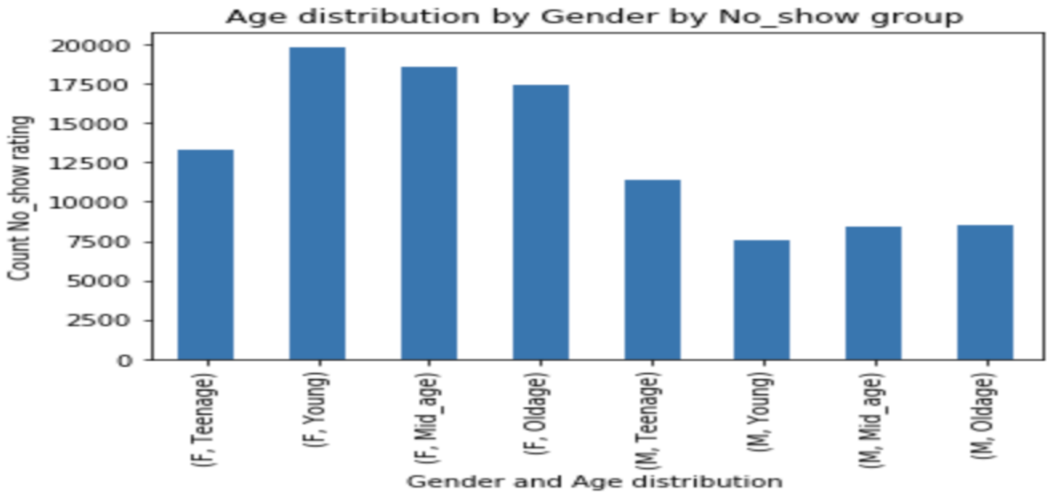
Which gender have a more no show ratio?



We see that there is a high concentration of results for Female patients, which means that in most cases, Female patients are more likely to not show up on appointment days.

In this graph, we can conclude that the Male patients more likely to show up on their appointment days.

Gender vise distribution with Age_distribution for No_show.



We see that there is a high concentration of results for female young age group for not to show up on the appointment day. We can see all the distribution from the chart based on the age group as well as their gender.

In this graph, we can conclude that the Male patients more likely to show up on their appointment days. To be precise, Male young patients are more punctual then other male patients.

Conclusion

Results: Our data suggest that

1. There is not big difference between the distribution of Age between patients who showed up for the appointment versus the patients that did not show up for the appointment.
2. Being enrolled in the scholarship program does not seem to make people more likely to show up to the appointment.
3. People that have a disease are more likely to show up for the appointment than people who do not have a disease.

Limitations: There are couple of limitations with our data:

1. Most of our variables are categorial, which does not allow for a higher level of statistical method that can be used to provide correlations etc.
2. Cannot show strong correlations between factors since most of our data is categorial.
3. The statistics used here are descriptive statistics, not inferential, meaning that we did not create any hypotheses or controlled experiments or inferences with our data.

References:

Google.com
Stackoverflow.com