Data Mining Assignment

On

Analysis of COVID Database

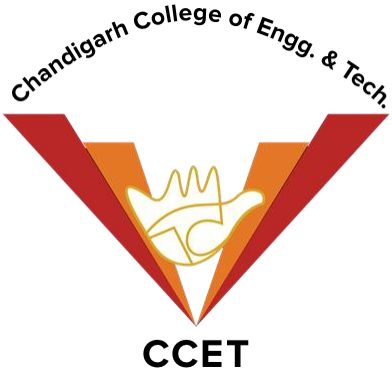
Project Report Submitted in Partial Fulfillment of
The Requirements for the Award of

Bachelor in Engineering

IN COMPUTER SCIENCE AND ENGINEERING
Submitted By

Amandeep	Ashish Upadhyay	Gaurav Kaushal
(CO17310)	(CO17315)	(CO17322)

Under the supervision of
Chandigarh College of Engineering and technology



Chandigarh College of Engineering and Technolgy
(Degree Wing)
Government Institute under Chandigarh (UT) Administration, Affiliated to Punjab University , Chandigarh
Sector-26, Chandigarh PIN 160019

ACKNOWLEDGEMENT

We have taken efforts in this project; however, it would have not been possible without the kind support and help of our mentor Mr. Ankit Gupta.

We are highly indebted to Chandigarh College Of Engineering and Technology (Degree Wing) for their guidance and constant supervision as well as for providing necessary information regarding the practical and constant supervision regarding the practical and also for their support in completing this report.

We would like to express our gratitude and thanks to institution (C.C.E.T.) persons for giving us such attention and time.

LIST OF FIGURES

CONTENT TABLE

CHAPTER 1: DATA COLLECTION

1.1 DATA SOURCES

The data that we are using is provided by John Hopkins University.

Source: https://github.com/CSSEGISandData/COVID-19

The data is in the form of comma-separated values files (.csv)

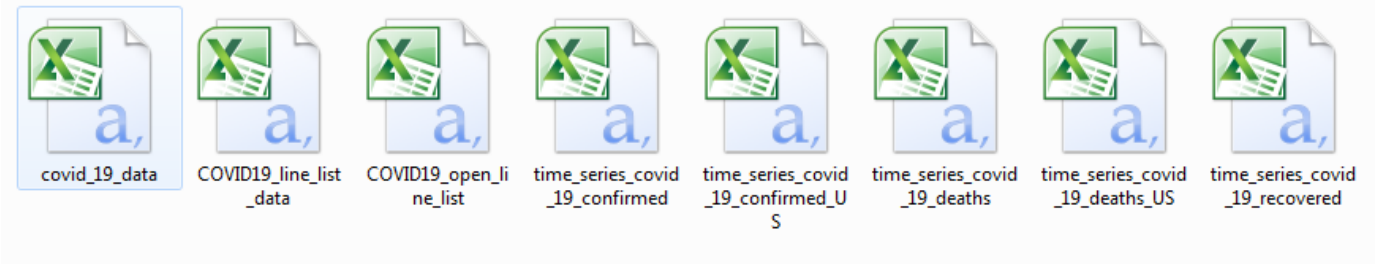


Fig. 1.1 List of the complete dataset

	A	B	C	D	E	F	G	H	I	J	K
7	84001001	US	USA	840	1001	Autauga	Alabama	US	32.53953	-86.6441	Autauga, Alabama, US
8	84001003	US	USA	840	1003	Baldwin	Alabama	US	30.72775	-87.7221	Baldwin, Alabama, US
9	84001005	US	USA	840	1005	Barbour	Alabama	US	31.86826	-85.3871	Barbour, Alabama, US
10	84001007	US	USA	840	1007	Bibb	Alabama	US	32.99642	-87.1251	Bibb, Alabama, US
11	84001009	US	USA	840	1009	Blount	Alabama	US	33.98211	-86.5679	Blount, Alabama, US
12	84001011	US	USA	840	1011	Bullock	Alabama	US	32.10031	-85.7127	Bullock, Alabama, US
13	84001013	US	USA	840	1013	Butler	Alabama	US	31.753	-86.6806	Butler, Alabama, US
14	84001015	US	USA	840	1015	Calhoun	Alabama	US	33.77484	-85.8263	Calhoun, Alabama, US
15	84001017	US	USA	840	1017	Chambers	Alabama	US	32.9136	-85.3907	Chambers, Alabama, US
16	84001019	US	USA	840	1019	Cherokee	Alabama	US	34.17806	-85.6064	Cherokee, Alabama, US
17	84001021	US	USA	840	1021	Chilton	Alabama	US	32.85044	-86.7173	Chilton, Alabama, US
18	84001023	US	USA	840	1023	Choctaw	Alabama	US	32.02227	-88.2656	Choctaw, Alabama, US
19	84001025	US	USA	840	1025	Clarke	Alabama	US	31.681	-87.8355	Clarke, Alabama, US
20	84001027	US	USA	840	1027	Clay	Alabama	US	33.26984	-85.8584	Clay, Alabama, US
21	84001029	US	USA	840	1029	Cleburne	Alabama	US	33.67679	-85.5201	Cleburne, Alabama, US
22	84001031	US	USA	840	1031	Coffee	Alabama	US	31.39933	-85.989	Coffee, Alabama, US
23	84001033	US	USA	840	1033	Colbert	Alabama	US	34.69847	-87.8017	Colbert, Alabama, US
24	84001035	US	USA	840	1035	Conecuh	Alabama	US	31.43402	-86.9932	Conecuh, Alabama, US
25	84001037	US	USA	840	1037	Coosa	Alabama	US	32.9369	-86.2485	Coosa, Alabama, US
26	84001039	US	USA	840	1039	Covington	Alabama	US	31.24779	-86.4505	Covington, Alabama, US
27	84001041	US	USA	840	1041	Crenshaw	Alabama	US	31.72942	-86.3159	Crenshaw, Alabama, US
28	84001043	US	USA	840	1043	Cullman	Alabama	US	34.1302	-86.8689	Cullman, Alabama, US
29	84001045	US	USA	840	1045	Dale	Alabama	US	31.43037	-85.611	Dale, Alabama, US
30	84001047	US	USA	840	1047	Dallas	Alabama	US	32.32688	-87.1087	Dallas, Alabama, US
31	84001049	US	USA	840	1049	DeKalb	Alabama	US	34.45947	-85.8078	DeKalb, Alabama, US
32	84001051	US	USA	840	1051	Elmore	Alabama	US	32.59785	-86.1442	Elmore, Alabama, US

Fig. 1.2 Example of a data file

CHAPTER 2: DATA CLEANING

2.1 DATA CLEANING:

The data we have collected may contain errors, missing values, noisy or inconsistent data. So, we get rid of such anomalies. This step was very time consuming, due to all the manual work required.

2.2 STEPS USED

2.2.1 RENAMING FILES AND REMOVING INCONSISTENCIES

In this step, we used Orange which is an open-source data visualization, machine learning and data mining toolkit to rename the files to represent the dates when the data was captured and to remove inconsistencies such as the use of abbreviations like "UK", "US" to represent countries, and disambiguation of names such as "China" and "China (Mainland)" to represent the same things.

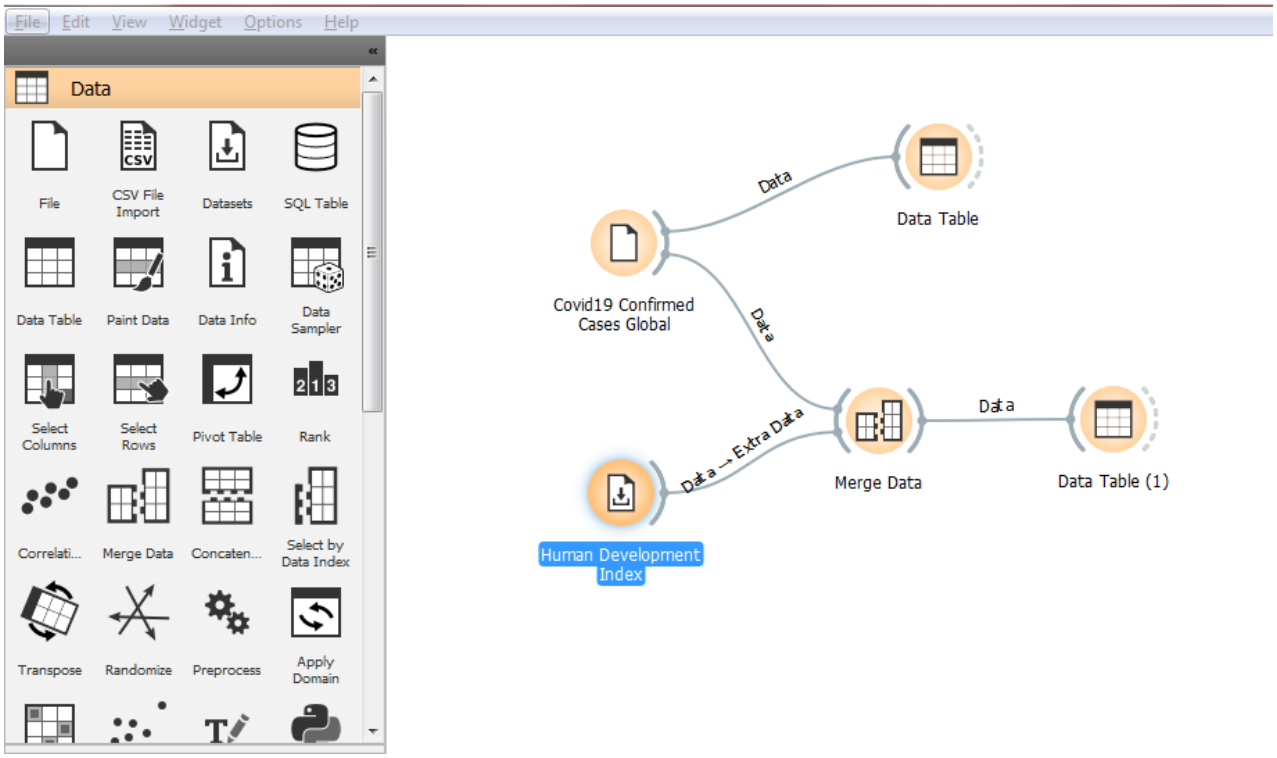
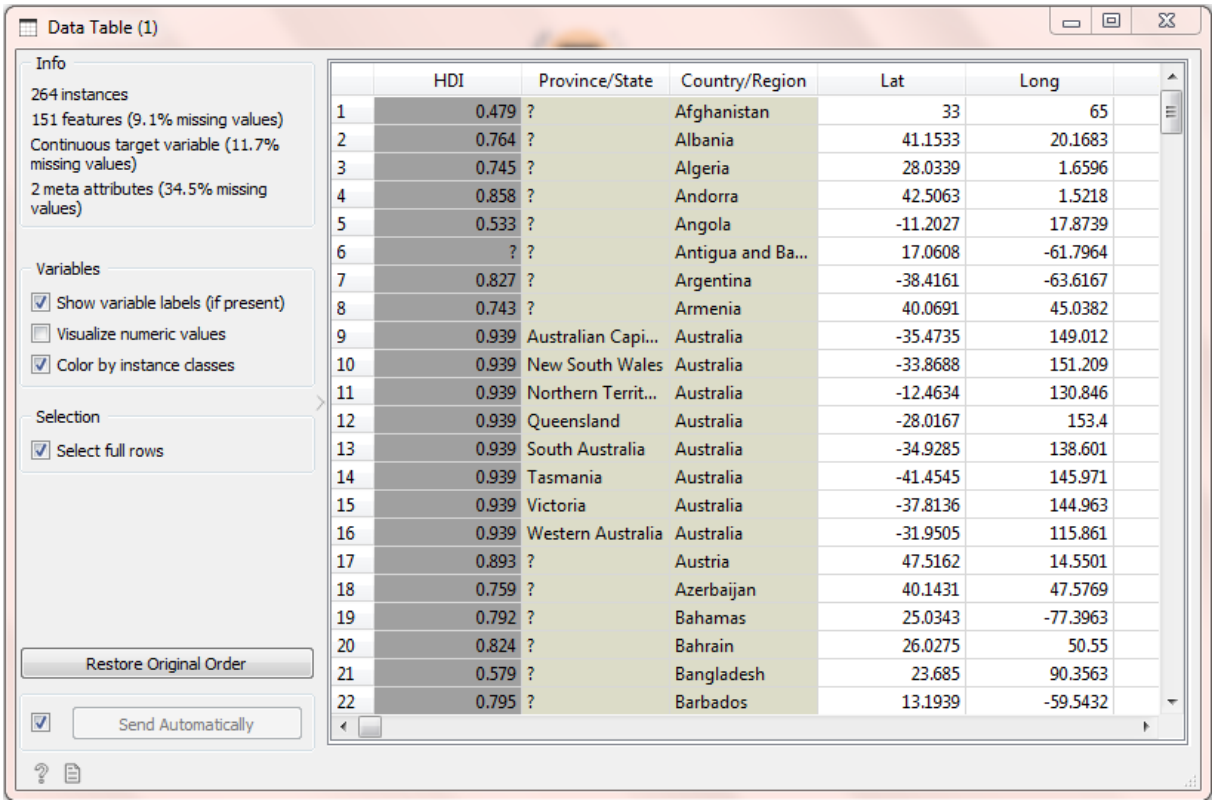


Fig. 2.1 Merging two datasets for example Covid19 Confirmed Cases Globally and Human Development Index

So when we merge these two datasets there are some countries with a missing value in HDI column as shown in the figure below



	HDI	Province/State	Country/Region	Lat	Long
1	0.479	?	Afghanistan	33	65
2	0.764	?	Albania	41.1533	20.1683
3	0.745	?	Algeria	28.0339	1.6596
4	0.858	?	Andorra	42.5063	1.5218
5	0.533	?	Angola	-11.2027	17.8739
6	?	?	Antigua and Ba...	17.0608	-61.7964
7	0.827	?	Argentina	-38.4161	-63.6167
8	0.743	?	Armenia	40.0691	45.0382
9	0.939	Australian Capi...	Australia	-35.4735	149.012
10	0.939	New South Wales	Australia	-33.8688	151.209
11	0.939	Northern Territ...	Australia	-12.4634	130.846
12	0.939	Queensland	Australia	-28.0167	153.4
13	0.939	South Australia	Australia	-34.9285	138.601
14	0.939	Tasmania	Australia	-41.4545	145.971
15	0.939	Victoria	Australia	-37.8136	144.963
16	0.939	Western Australia	Australia	-31.9505	115.861
17	0.893	?	Austria	47.5162	14.5501
18	0.759	?	Azerbaijan	40.1431	47.5769
19	0.792	?	Bahamas	25.0343	-77.3963
20	0.824	?	Bahrain	26.0275	50.55
21	0.579	?	Bangladesh	23.685	90.3563
22	0.795	?	Barbados	13.1939	-59.5432

Fig. 2.2 Data Table consisting of missing values

To resolve this issue we use Edit Domain tool in Orange. We change the country name to their respective format to match them in both datasets. Example United States to US.

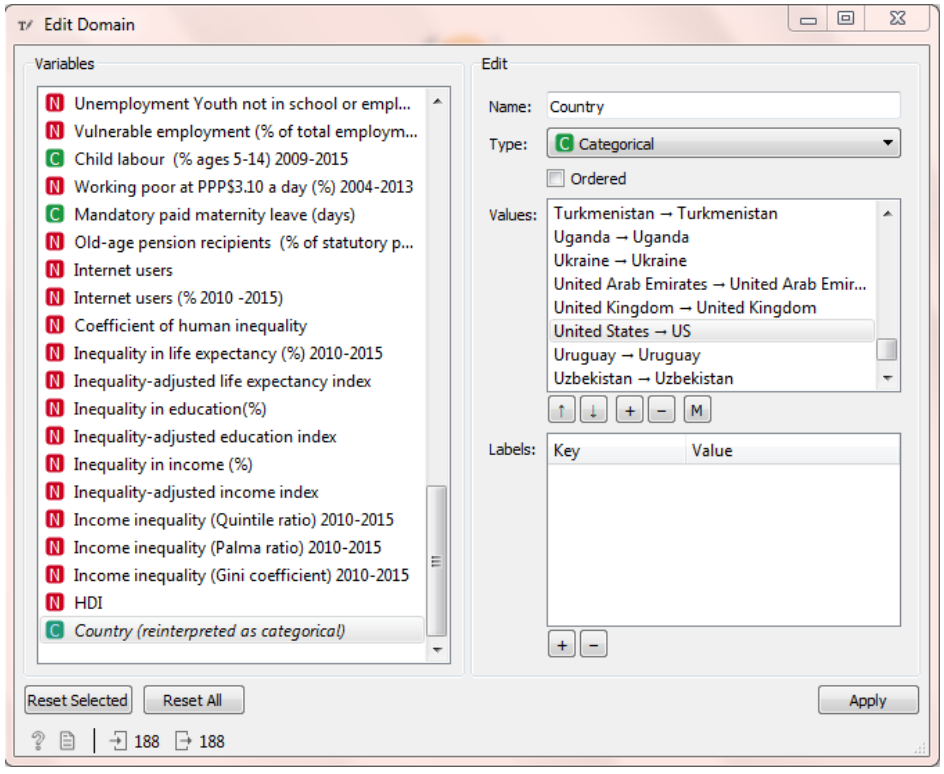


Fig. 2.3 Renaming country names to resolve missing values

Also there are some entries of the countries with unknown number of population. We resolve this using Select Rows widget. We will only select countries which have a population greater than zero.

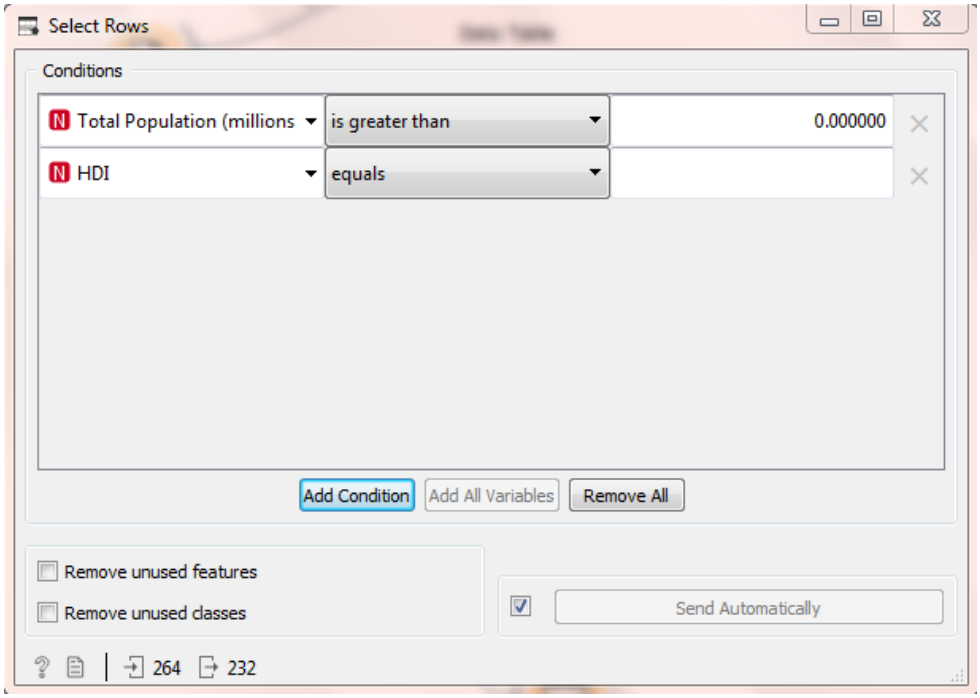


Fig. 2.4 Selecting Countries with non-zero population

Now we have resolved all the inconsistencies from out datasets and we will move to the next step which is data mining.

CHAPTER 3: DATA MINING

3.1 ORANGE

Orange is a component-based visual programming software package for data visualization, machine learning, data mining, and data analysis. Orange components are called widgets and they range from simple data visualization, subset selection, and preprocessing, to empirical evaluation of learning algorithms and predictive modeling. Visual programming is implemented through an interface in which workflows are created by linking predefined or user-designed widgets.

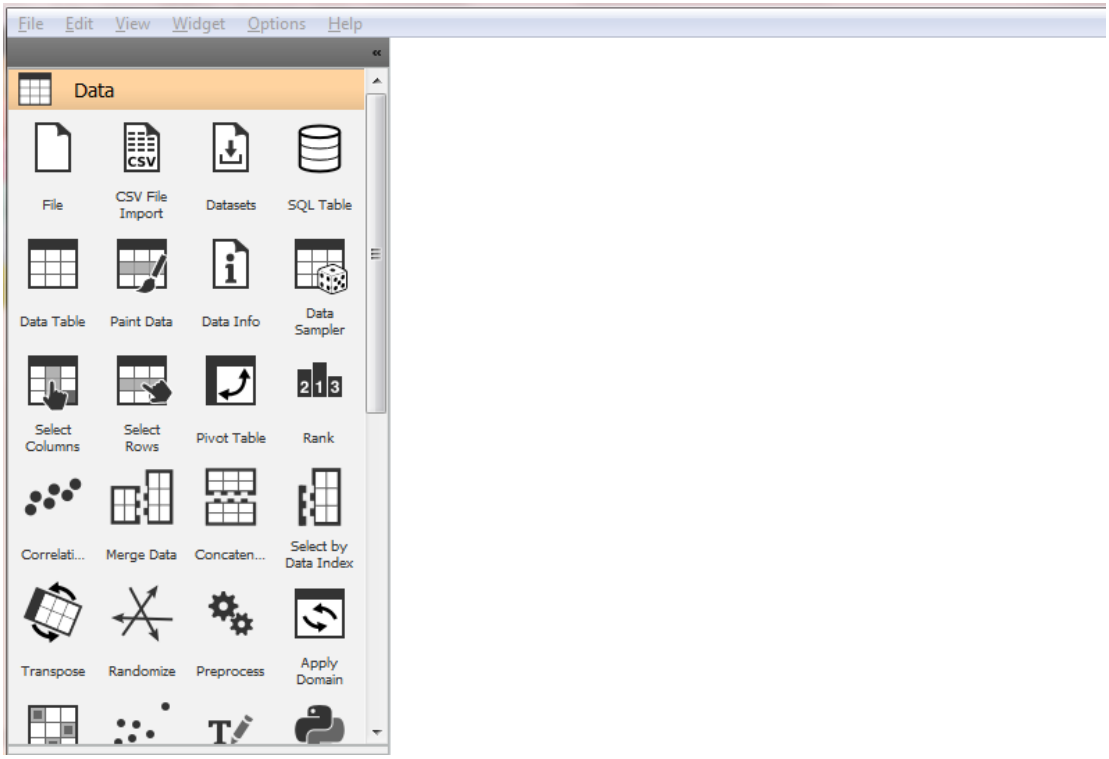


Fig. 3.1 Basic Orange Interface

To visualize the relation between different attributes of the datasets we will use a widget named Feature Constructor which will compute the cases to population ratio.

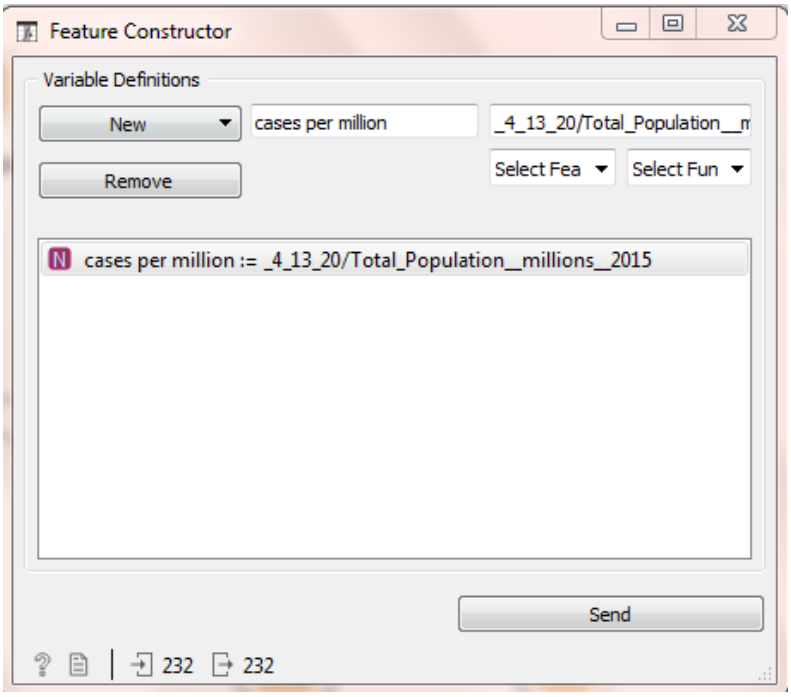


Fig. 3.2 Feature Constructor

Now we will connect Scatter Plot widget to this Feature Constructor widget to visualize the data.

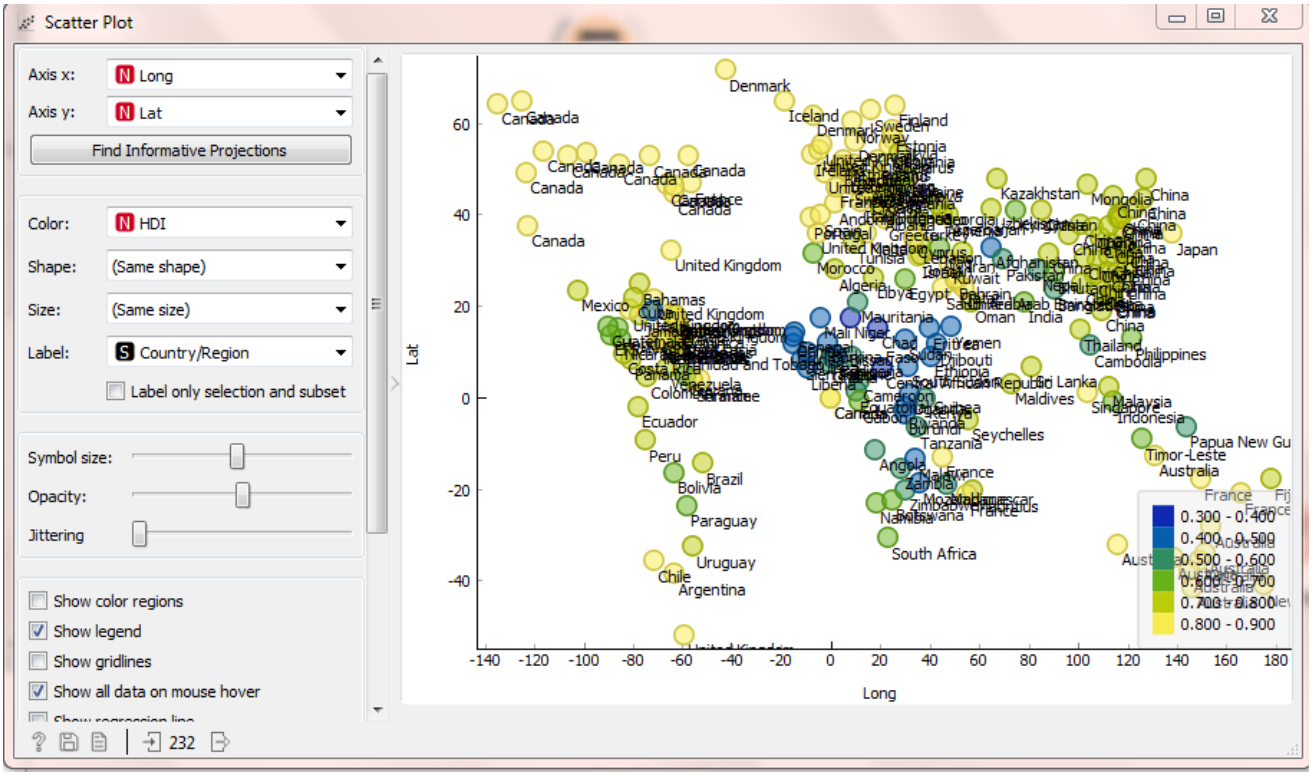


Fig. 3.3 Visualizing the data using Scatter Plot

CHAPTER 4: PATTERNS FOUND

4.1 CONDITION OF HEALTHCARE SYSTEM

The countries which have a low number of Covid19 cases with high number doctors will have a good healthcare system. Means that they can handle this crisis pretty well.

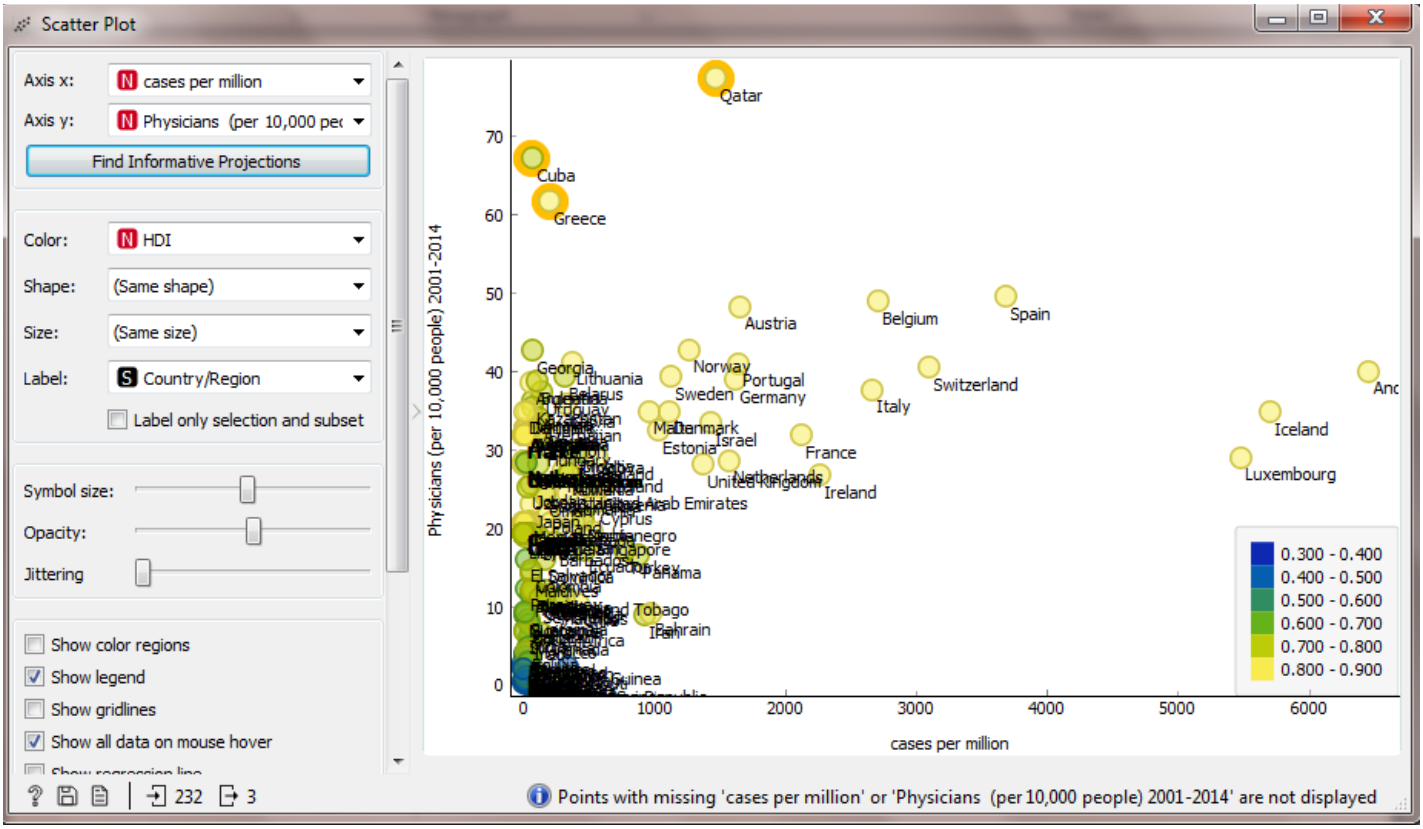


Fig. 4.1 Plot for Condition of Healthcare System

In the above graph we can see that the countries like Cuba, Greece and Qatar have a high number of doctors as compared to the number of infected patients which is very low. So their healthcare system is in a good shape.

4.2 DEPENDENCE ON OLD PEOPLE

It is shown that the old people are more effected by this Covid19. So in our graph we can see that Japan has the most number of old people per million which can lead to a dangerous outcome if not prevented properly.

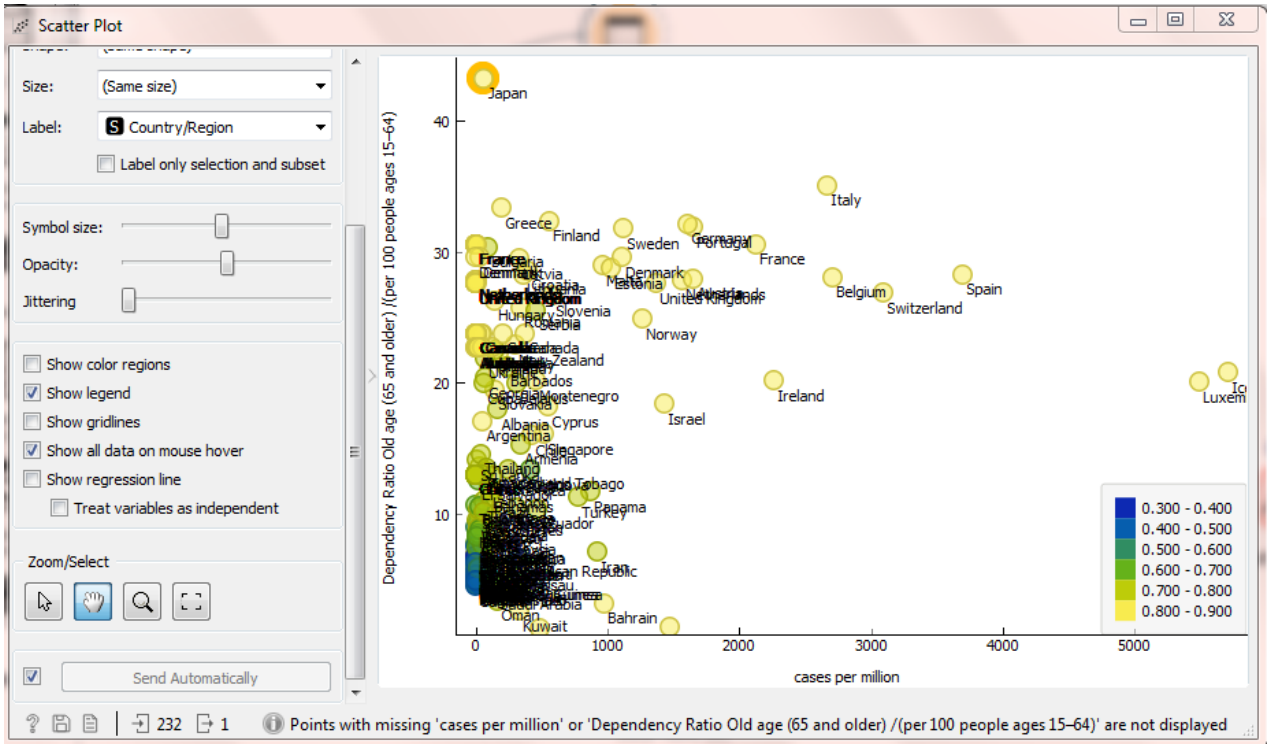


Fig. 4.2 Plot for Dependency on Old People