ASSIGNMENT-1

Analysis and Visualization of a dataset

Made by-

Megha Agarwal

08101182021

ECEAI2, EA3

SUBMITTED TO-

MRS.NONITA SHARMA

Project Title: Analysis and visualization of H1N1 and seasonal vaccines

ABSTRACT

In 2020,the entire world got affected by a life-threatening disease called covid-19.India became one of the first countries in the world to roll out vaccines.

This project aims to analyze and visualize how likely an individual is to receive the H1N1 and seasonal flu vaccines. The dataset is collected from the 2009 H1N1 Flu Survey and consists of rows corresponding to a person who responded. Using python libraries, we are going to make different plots and reach to various conclusions. We are going to analyse various factors such as mass awareness, concerns, and behavioural changes.

KEYWORDS

INTRODUCTION

To execute our project on the dataset, we will use various python libraries such as:

Matplotib: It accompany you to create various interactive visualization using various kinds of plots such as histogram, bar plots, etc.

Pandas: It helps us in manipulation, modification and analysing our tabular data for graphical analysis. It is one of the most used libraries in python.

Seaborn: It is used for data visualization. It uses matplotlib.It is one of the most important libraries in python.

Sklearn: It is another python library used to selection, classify, etc of data.

Numpy: it used to manipulate and modify data in the form of an array.

```python
# Common libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Libraries for Feature Selection

from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.decomposition import PCA
```

Our dataset contains 26707 unique entries and 36 attributes corresponding to each entry. We have data of the following types:

Float64	23
Int64	1
object	12

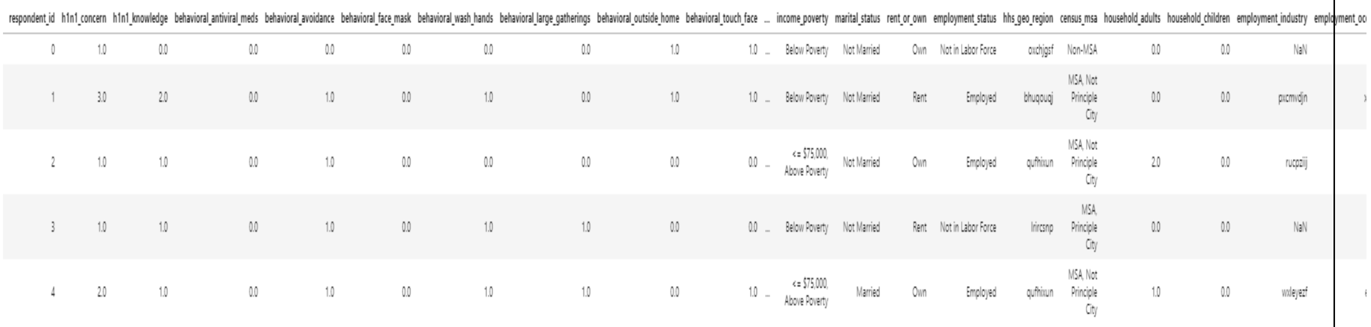
We need to encode the object into an int. There are no **duplicates** in our dataset. There are many **missing values** in our dataset which need to be imputed.

The objective of our project is to analyse and visualize how likely an individual is to receive the H1N1 and seasonal flu vaccines using the dataset collected from the 2009 H1N1 Flu Survey. We are going to analyse various attributes of our data such as mass awareness, concerns, and behavioural changes.

It became essentially imperative and absolutely important for us to adopt the various statistical techniques in order to effectively minimize human error in mathematical derivations and calculation and maximize the chances of a successful comparative analysis and volumetric evaluation.

Methodology

1. **Data Collection:**
   The dataset is collected from 2009 H1N1 Flu Survey. There are 26707 unique ids and 36 attributes corresponding to them. Each row in the dataset is corresponding to one person. There are two target variables: H1N1 Flu vaccine and Seasonal Flu vaccine.



respondent_id	h1n1_concern	h1n1_knowledge	behavioral_antiviral_meds	behavioral_avoidance	behavioral_face_mask	behavioral_wash_hands	behavioral_large_gatherings	behavioral_outside_home	behavioral_touch_face	...	income_poverty	marital_status	rent_or_own	employment_status	hhs_geo_region	census_msa	household_adults	household_children	employment_industry	employment_oc
0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	...	Below Poverty	Not Married	Own	Not in Labor Force	oxchjgsf	Non-MSA	0.0	0.0	NaN	
1	3.0	2.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	...	Below Poverty	Not Married	Rent	Employed	bhuqouqj	MSA, Not Principle City	0.0	0.0	pxcmvdjn	x
2	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	<= $75,000, Above Poverty	Not Married	Own	Employed	qufhixun	MSA, Not Principle City	2.0	0.0	rucpziij	
3	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	...	Below Poverty	Not Married	Rent	Not in Labor Force	lrircsnp	MSA, Principle City	0.0	0.0	NaN	
4	2.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0	...	<= $75,000, Above Poverty	Married	Own	Employed	qufhixun	MSA, Not Principle City	1.0	0.0	wxleyezf	e

2. Data preprocessing:

It refers to manipulating the data and making it clean before analysing it. We look out for duplicate values and missing values. Then the missing values are either droped if not important for our analysis or imputed by **mean, median**, etc. Other categorical values are encoded as **Label** or **One hot encoding**.

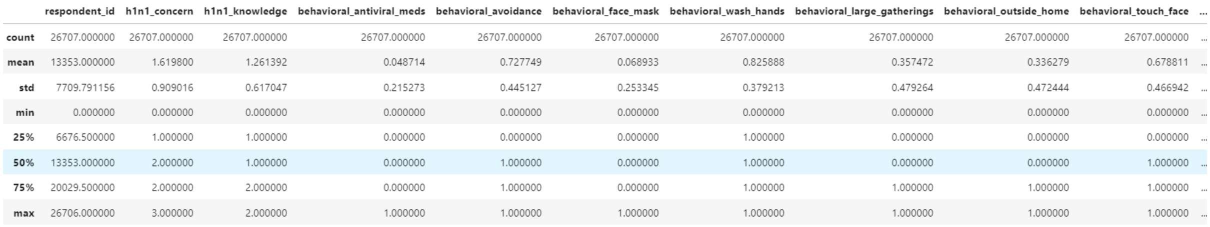
3. Statistical Analysis:

The implementation and execution has been performed with the help of various graphs that cater to the representation of outliers,maninimum and maximum values mainly to identify the fashion/pattern/trends in both univariate and multivariate datasets gor handling the categorical values.

Result and Discussion

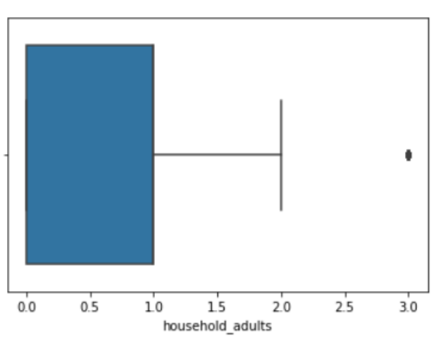
1.Statistical Analysis

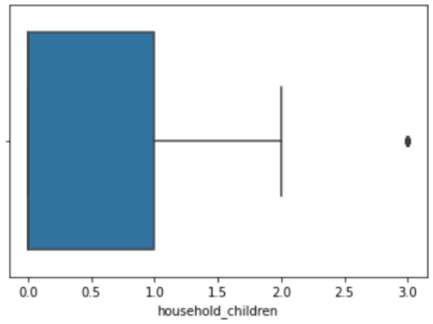
Describing our dataset(PREPROCESSING)



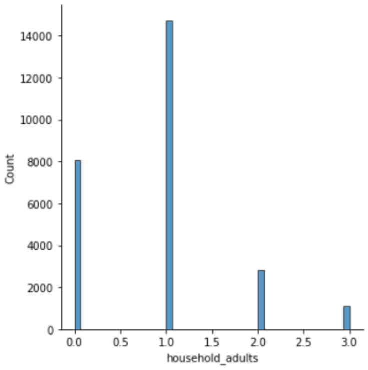
	respondent_id	h1n1_concern	h1n1_knowledge	behavioral_antiviral_meds	behavioral_avoidance	behavioral_face_mask	behavioral_wash_hands	behavioral_large_gatherings	behavioral_outside_home	behavioral_touch_face	...
count	26707.000000	26707.000000	26707.000000	26707.000000	26707.000000	26707.000000	26707.000000	26707.000000	26707.000000	26707.000000	...
mean	13353.000000	1.619800	1.261392	0.048714	0.727749	0.068933	0.825888	0.357472	0.336279	0.678811	...
std	7709.791156	0.909016	0.617047	0.215273	0.445127	0.253345	0.379213	0.479264	0.472444	0.466942	...
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	6676.500000	1.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	...
50%	13353.000000	2.000000	1.000000	0.000000	1.000000	0.000000	1.000000	0.000000	0.000000	1.000000	...
75%	20029.500000	2.000000	2.000000	0.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	...
max	26706.000000	3.000000	2.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...

Finding and vidualising the outliners:

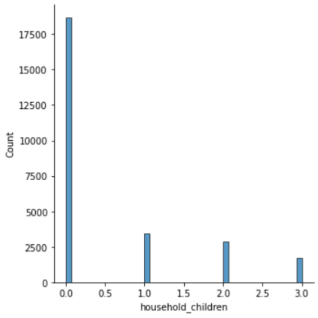




2.Visualisation Analysis

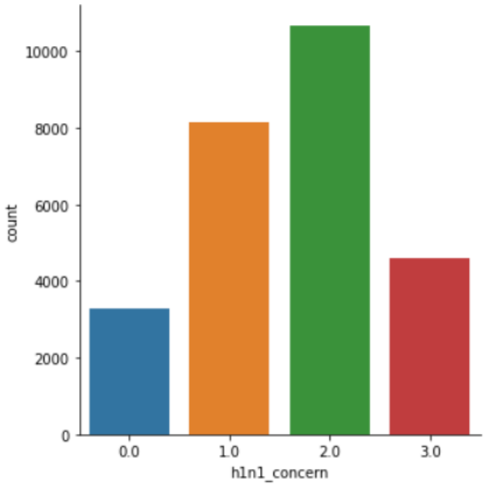


Number of other adults in household is mostly 1.

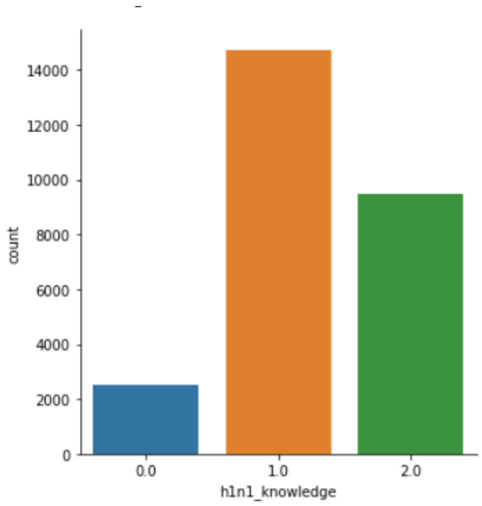


Number of children in household is mostly 0**.**

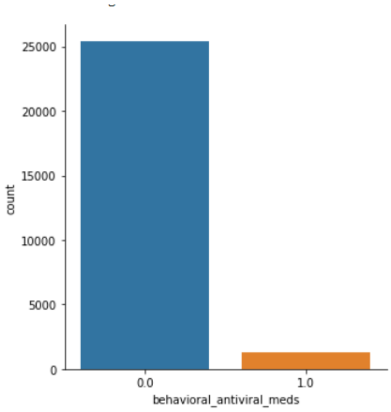
Visualising categorical data:



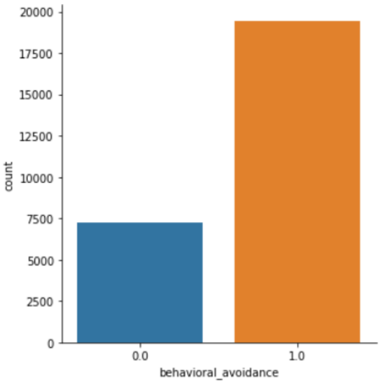
Most people are concerned about H1N1 vaccine.



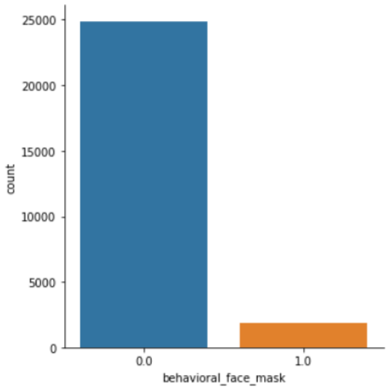
Almost 90% of people have little or lots of knowledge about h1n1 vaccine.



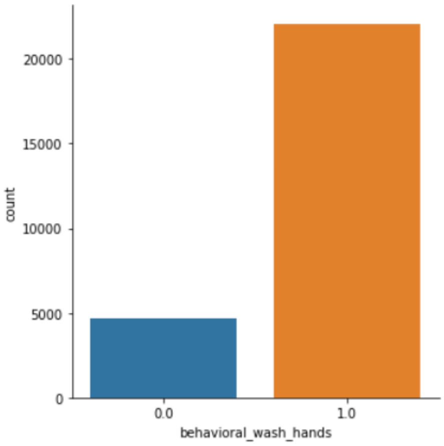
Almost non of the people took antiviral medications.



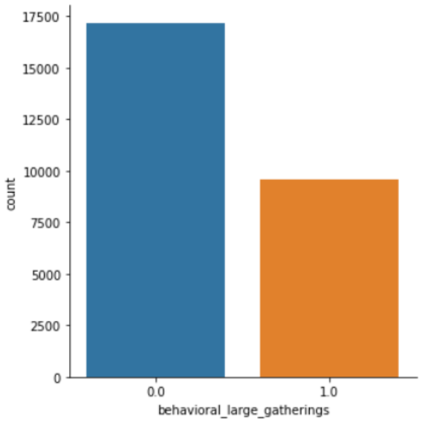
70% of people have avoided contact with people with symptoms.



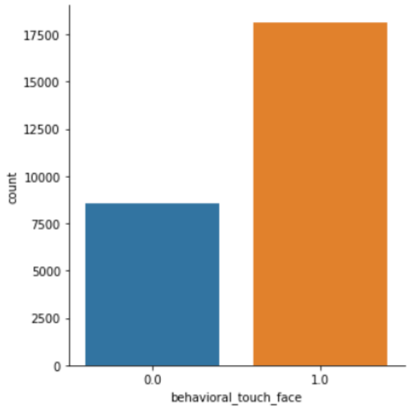
Very few people have bought a face mask.



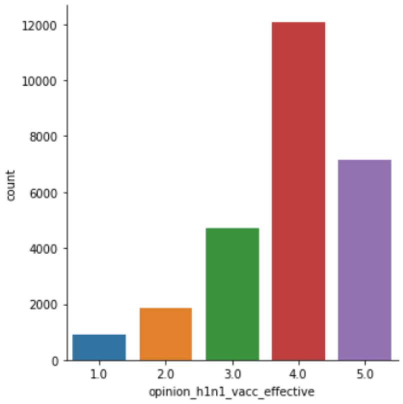
80% of people have frequently washed their hands.



Few people have reduced time at large gatherings.



70% of people have avoided touching eyes, nose, or mouth.



Most of the people think h1n1 is somewhat effective.¶

CONCLUSION

Based on the results derived by the above statistical techniques, we have effectively implemented and executed the visualization of the dataset through the various python libraries catering to the enormous number of features available for the classification of attributes.