CNG-514-Assignment-1

Muhammad Somaan 2528404

March 2024

1 Dataset Identification

The dataset is about a survey from 10753 participants with a total of 19 questions asked. The dataset mainly focuses on how the coronavirus has impacted their lives and contains information related to coronavirus activities in their lives, such as their concerns over it, their intentions against the preventative measures of coronavirus, their leading choice of support in case they want help, their symptoms, and their belief on whether they have the virus or not. The dataset also includes their personal details such as age, household income, gender, ethnicity, religion, etc., to identify any possible correlations.

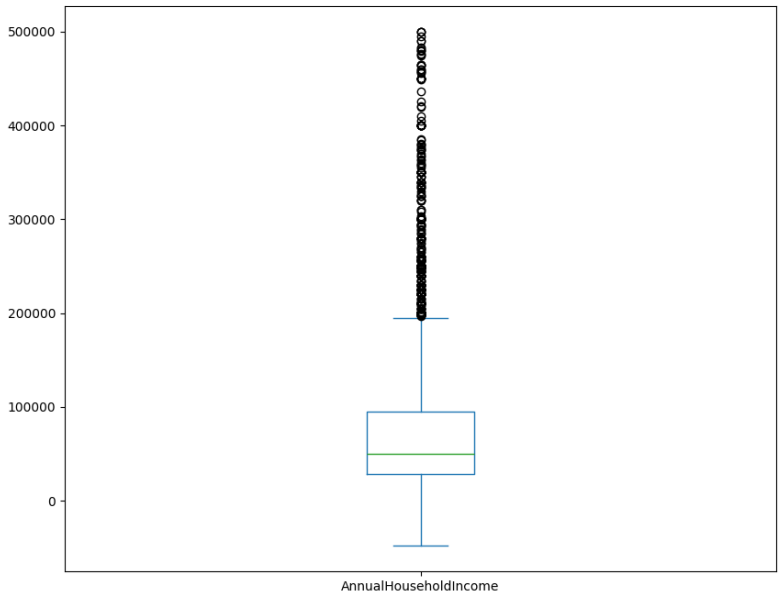
Table 1: Attributes Datatype

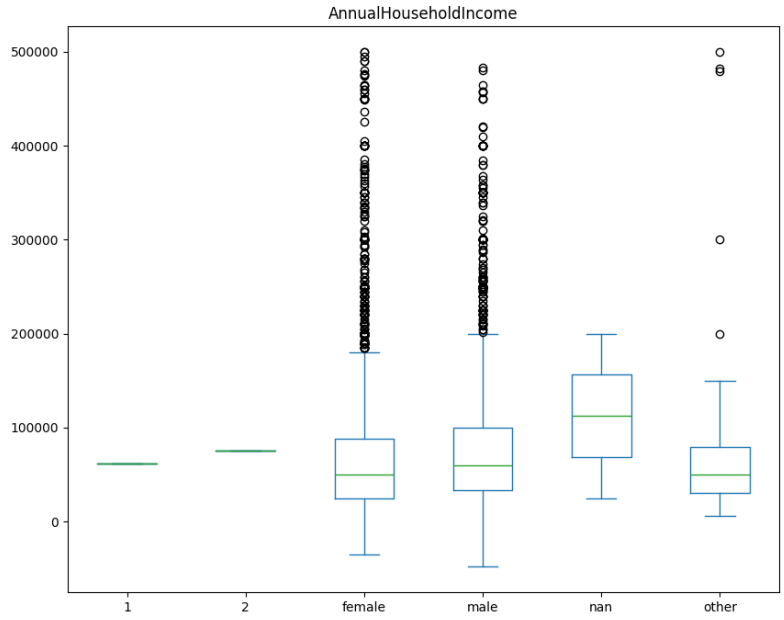
ParticipantId	Numeric	Interval	Integer
AnnualHouseholdIncome	Numeric	Ratio	Integer
BirthYear2020	Numeric	Interval	Integer
CoronavirusConcern2	Numeric	Ratio	Float
CoronavirusEmployment	Categorical	Nominal	List
CoronavirusIntent_Mask	Numeric	Ratio	Integer
CoronavirusIntent_SixFeet	Numeric	Ratio	Integer
CoronavirusIntent_StayHome	Numeric	Ratio	Integer
CoronavirusIntent_WashHands	Numeric	Ratio	Integer
CoronavirusLocalCommunity	Numeric	Ratio	Integer
CoronavirusSupportSystem	Categorical	Ordinal	List
CoronavirusSymptomSelect	Categorical	Ordinal	List
Education_Alt2	Categorical	Ordinal	List
Ethnicity	Categorical	Nominal	String
Gender	Categorical	Nominal	String
HasCoronavirusBelief	Numeric	Ratio	Float
PoliticalBeliefsNoOp	Numeric	Ratio	Float
Religion_Alt1	Categorical	Nominal	String
Religiosity_Alt2	Numeric	Ratio	Float
ZipCode	Categorical	Nominal	Integer

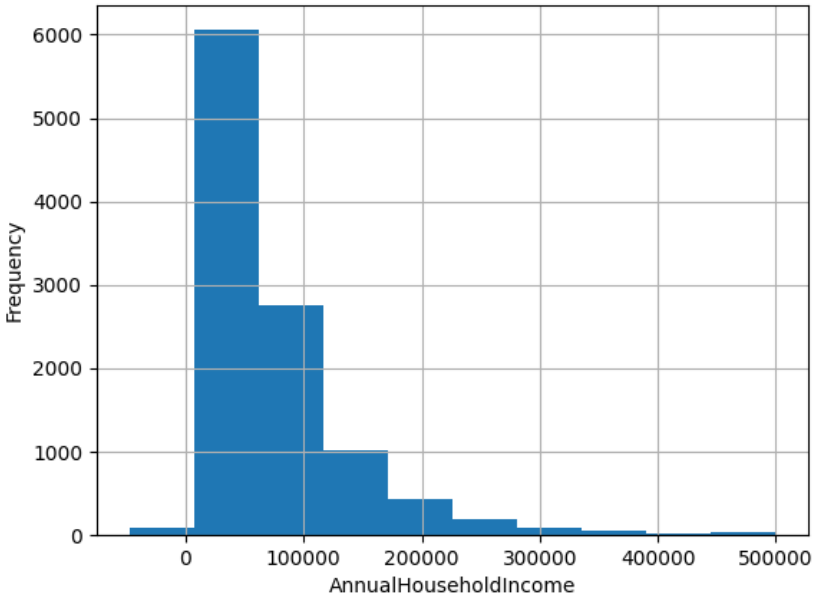
1.1 Data Mining Application

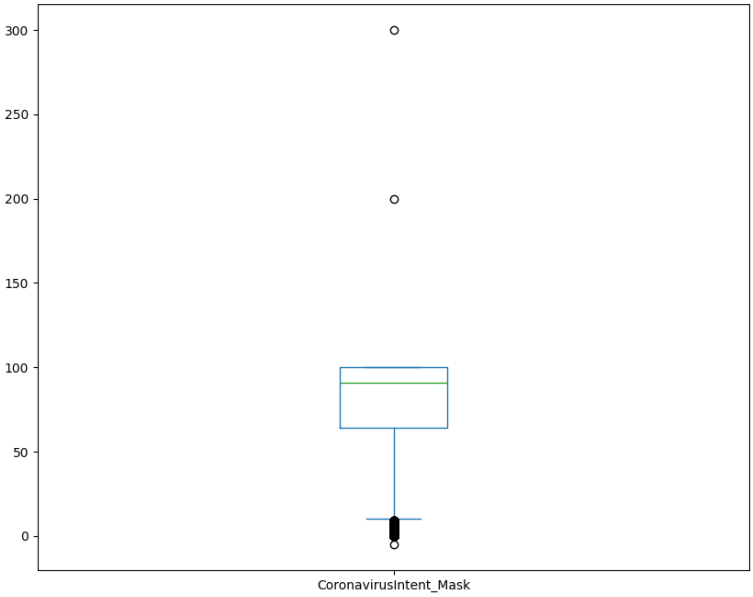
This dataset can be great for clustering applications. Employment status, household income, age, and support system attributes combined with others can be used to identify groups that are likely to need help in the pandemic against the coronavirus. This can be used to arrange help early on before the group's condition worsens.
Similarly, participants' intentions against preventative measures against coronavirus and their concerns over it could help us identify groups more likely to fall prey to the virus; this can be used to impose strict measures against the coronavirus, convey the seriousness of the situation and educate the importance of the preventative measures.
Another helpful measure could be the ethnical groups, the severity of symptoms they face, and their household income to identify the groups that would require more serious care so that those groups could be targeted early on to avoid their conditions getting serious.
Classification could be used to find correlations between the level of education and how seriously the participant takes the situation.

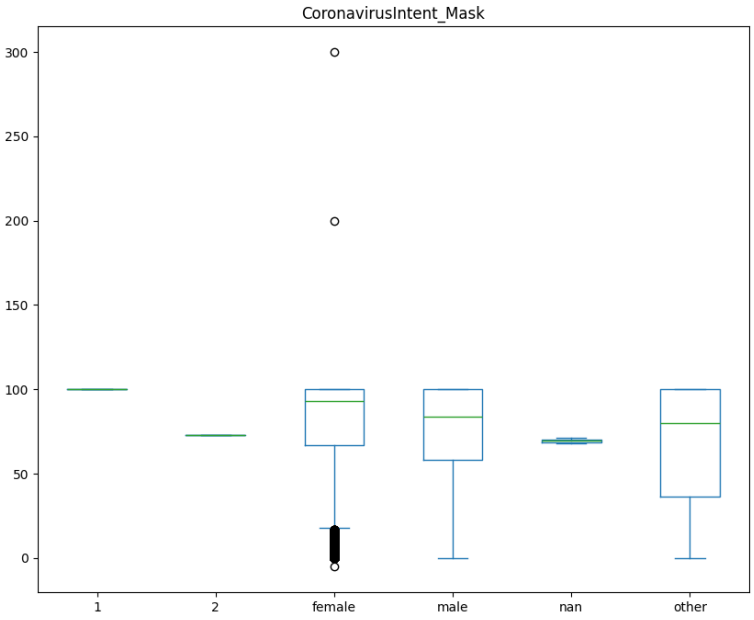
2 Attribute Characterization

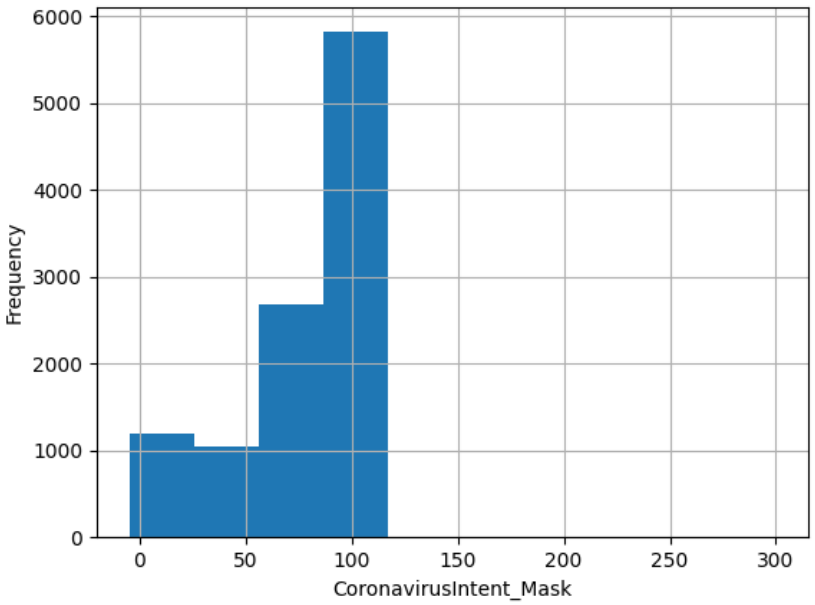


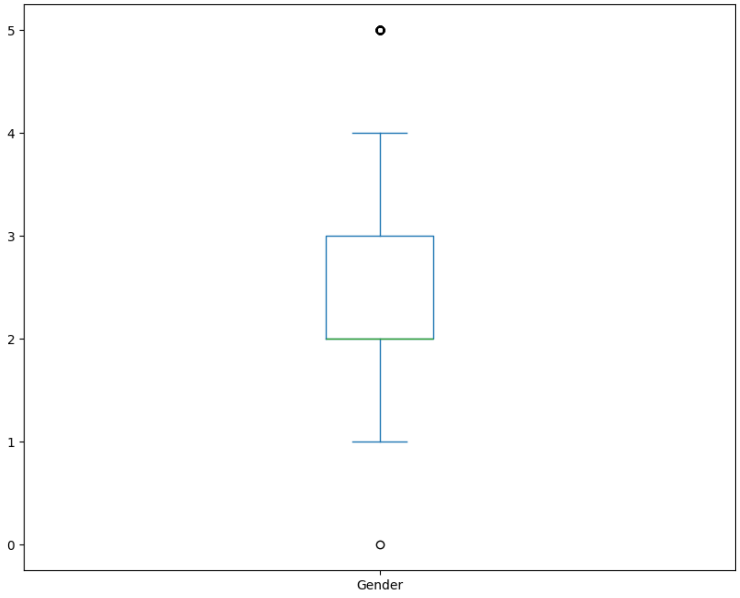

AnnualHouseholdIncome
500000
400000
300000
200000
100000
0
1
2
female
male
nan
other

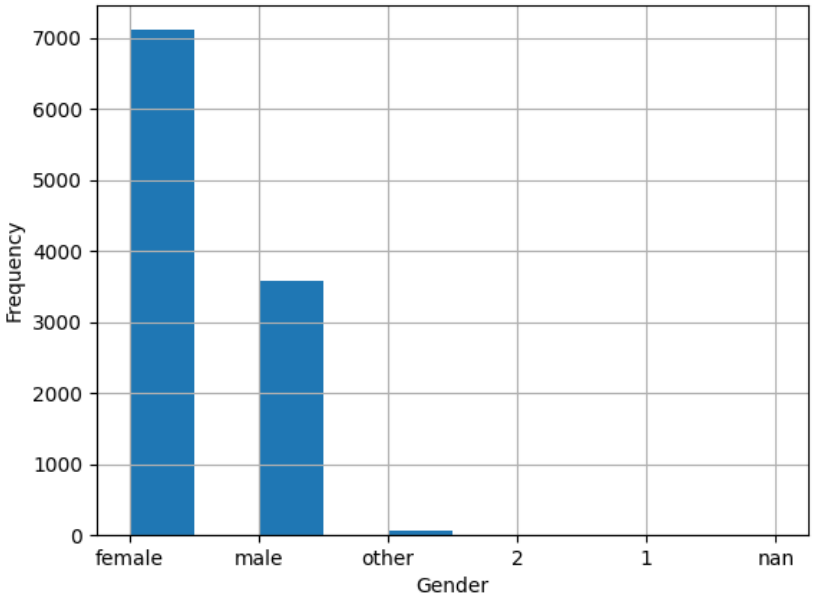

6000
5000
4000
3000
2000
1000
0
Frequency
0
100000
200000
300000
400000
500000
AnnualHouseholdIncome

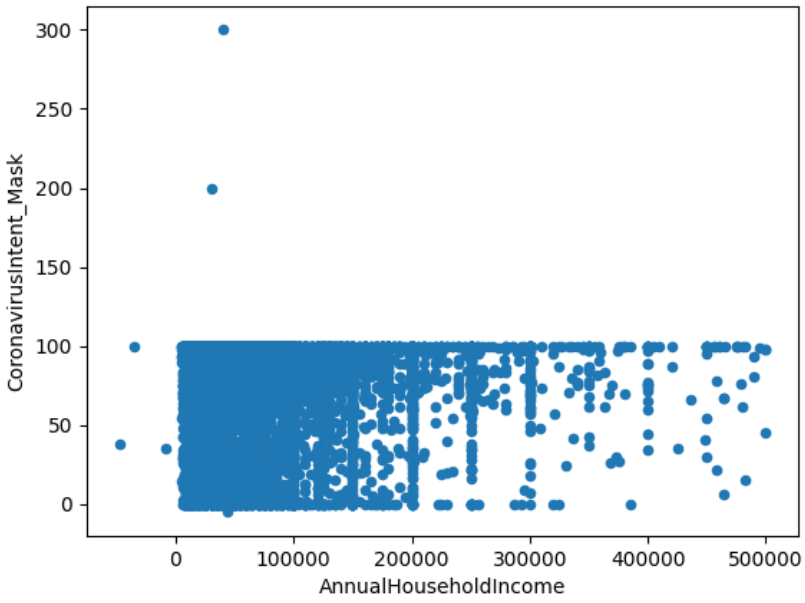

300
250
200
150
100
50
0
CoronavirusIntent_Mask

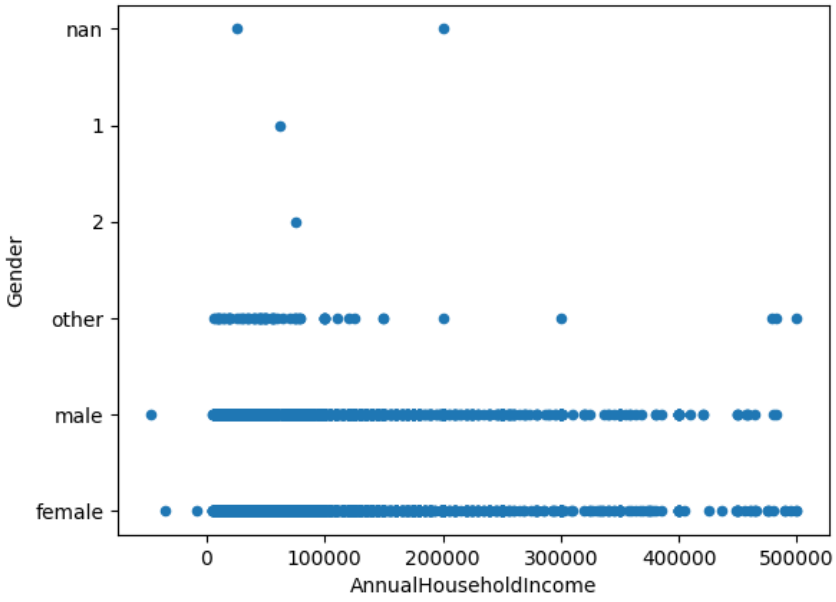

CoronavirusIntent_Mask
300
250
200
150
100
50
0
1
2
female
male
nan
other


6000
5000
4000
3000
2000
1000
0
Frequency
0
50
100
150
200
250
300
CoronavirusIntent_Mask


5
4
3
2
1
0
Gender


7000
6000
5000
4000
3000
2000
1000
0
Frequency
female
male
other
2
1
nan
Gender


CoronavirusIntent_Mask
300
250
200
150
100
50
0
0
100000
200000
300000
400000
500000
AnnualHouseholdIncome


Gender
nan
1
2
other
male
female
0
100000
200000
300000
400000
500000
AnnualHouseholdIncome

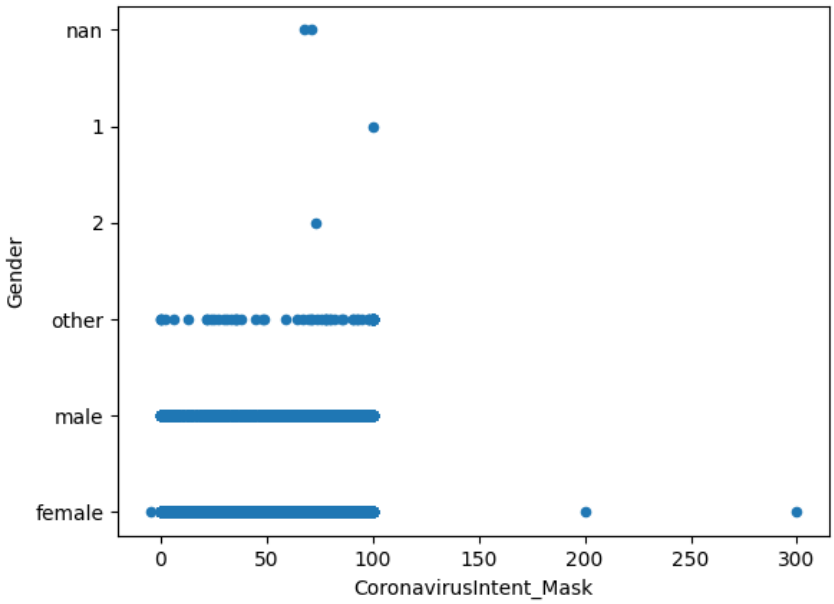


Table 2: Attributes Data

Attribute Name	AnnualHouseholdIncome	CoronavirusIntent_Mask	Gender
Mean	72230	76	-
Median	50000	91	-
Mode	50000	100	female
Range	547857	305	-
Standard Deviation	65599	31	-
Variance	4303244049	938	-
Q1	28000	64	-
Q3	95000	100	-

3 Data Preprocessing

3.1 Missing Values

AnnualHouseholdIncome, CoronavirusIntent_Mask and Gender, all three of them have two missing values each. Since the amount of missing values is very minute compared to the dataset size, the easiest approach to deal with this will be to drop those rows. A better approach would be to use the Bayesian formula to generate the values, but it will be too much hassle for too little.

3.2 Noisy Data

From the plots above, we can see that each attribute has some noisy data.

3.2.1 AnnualHouseholdIncome

The AnnualHouseholdIncome has some negative values; since they are still integers we can assume it to be a sign mistake and convert them into positive values.

3.2.2 CoronavirusIntent_Mask

The CoronavirusIntent_Mask attribute has values greater than 100, we can use clustering to identify and remove the outliers.

3.2.3 Gender

The Gender attribute contains some numerical data, they are classified as an outlier and hence clustering can be used to remove them.

3.3 Discretization

3.3.1 Age

Age can be discretized by the concept of hierarchy formation, such that we can divide the data between age groups that can be categorized as young, middle-aged, or senior.

3.3.2 CoronavirusIntent_Mask

CoronavirusIntent_Mask can be discretized by using ChiMerge since a high enough intention would most likely lead to the same results.

3.3.3 Gender

The gender attribute can also be discretized by using ChiMerge, since similar gender may produce similar results.

3.4 Normalization

```
# Normalization
scaler = preprocessing.MinMaxScaler()
data_normalized_min_max = data.copy()
data_normalized_min_max["AnnualHouseholdIncome"] = scaler.fit_transform(data_normalized_min_max[["AnnualHouseholdIncome"]])
data_normalized_min_max["CoronavirusIntent_Mask"] = scaler.fit_transform(data_normalized_min_max[["CoronavirusIntent_Mask"]])
print(data_normalized_min_max)

scaler = preprocessing.StandardScaler()
data_normalized_z_score = data.copy()
data_normalized_z_score["AnnualHouseholdIncome"] = scaler.fit_transform(data_normalized_z_score[["AnnualHouseholdIncome"]])
data_normalized_z_score["CoronavirusIntent_Mask"] = scaler.fit_transform(data_normalized_z_score[["CoronavirusIntent_Mask"]])
print(data_normalized_z_score)
```

3.5 Correaltion

```
#Correlation
correlation = income.corr(mask_intent)
print("Correlation between Income and Mask Intent: ", correlation)
correlation = income.corr(gender_labeled)
print("Correlation between Income and Gender: ", correlation)
correlation = mask_intent.corr(gender_labeled)
print("Correlation between Mask Intent and Gender: ", correlation)
```

```
Correlation between Income and Mask Intent:  0.04091001258486929
Correlation between Income and Gender:  0.08921718776512559
Correlation between Mask Intent and Gender:  -0.07907802816030307
```