Data Exploration

Project Description

Important Considerations

Due to the dynamic and fluid nature of the company, particularly regarding the organizational structure and hierarchy, our project goals are unclear to us at this instant, and are subject to change based on company requirements with time. Consequently, our particular project statement is not known to us, and we have a choice of two projects, as given below:

Project 1: To detect the presence of background noise in cough samples.
Project 2: To create a model to detect the presence of COVID in cough samples as a baseline for comparison.

The project may change in January. For the purpose of this assignment, we considered the "master" dataset provided to us, that we hypothesize will be useful for both projects. As a result, our analysis is rather general, and is not particularly focused to a particular project.

Data Description

Data repository: https://github.com/iiscleap/Coswara-Data

Project Coswara by the Indian Institute of Science (IISc) Bangalore is an attempt to build a diagnostic tool for COVID-19 detection using audio recordings such as breathing, cough, and speech sounds of an individual. The data was collected through crowdsourcing.
Our project at Virufy will mainly utilize this data.

Assumptions

- Positive test results taken over a month ago are treated as recovered infections.
- This data is crowdsourced and consequentially may contain biases from this collection procedure such as sampling bias, convenience sampling, domain specificity, and so on.
- We hypothesize that NaNs in our comorbidity features are equivalent to False, but cannot confirm the same before conferring with our stakeholders. Consequently, we have not used this assumption in our EDA.

Column Descriptions and Feature Set

- id: User ID
- a: Age (number)
- covid_status: Health status (e.g. : positive_mild, healthy,etc.)
- record_date: Date when the user recorded and submitted the samples
- ep: Proficient in English (y/n)
- g: Gender (male/female/other)
- l_c: Country
- l_l: Locality
- l_s: State
- rU: Returning User (y/n)
- asthma: Asthma (True/False)
- cough: Cough (True/False)
- smoker: Smoker (True/False)
- test_status: Status of COVID Test (p->Positive, n->Negative, na-> Not taken Test)
- ht: Hypertension (True/False)
- cold: Cold (True/False)
- diabetes: Diabetes (True/False)
- diarrhoea: Diarrheoa (True/False)
- um: Using Mask (y/n)
- ihd: Ischemic Heart Disease (True/False)
- bd: Breathing Difficulties (True/False)
- st: Sore Throat (True/False)
- fever: Fever (True/False)
- ftg: Fatigue (True/False)
- mp: Muscle Pain (True/False)
- loss_of_smell: Loss of Smell & Taste (True/False)
- cld: Chronic Lung Disease (True/False)
- pneumonia: Pneumonia (True/False)
- ctScan: CT-Scan (y/n if the user has taken a test)
- testType: Type of test (RAT/RT-PCR)
- test_date: Date of COVID Test (if taken)
- vacc: Vaccination status (y->both doses, p->one dose(partially vaccinated), n->no doses)
- ctDate: Date of CT-Scan
- ctScore: CT-Score
- others_resp: Respiratory illnesses other than the listed ones (True/False)
- others_preexist: Pre-existing conditions other than the listed ones (True/False)
- bgnoise_detected: Indicator of presence of background noise in cough sample (number)

Feature Set

(SUBJECT TO CHANGE)
In this section, we are listing common features that we think are important for **both** projects.

- submission_date - date sample acquired
- pcr_test_date - date of test
  important note: positive tests results >1 month ago can be treated as recovered infections
- (inferred field) days_since_pcr_test - model developers can create this
- pcr_test_result - positive, negative, untested, pending
- cough_path - file path to cough
- bgnoise_detected - indicator of background noise in cough sample

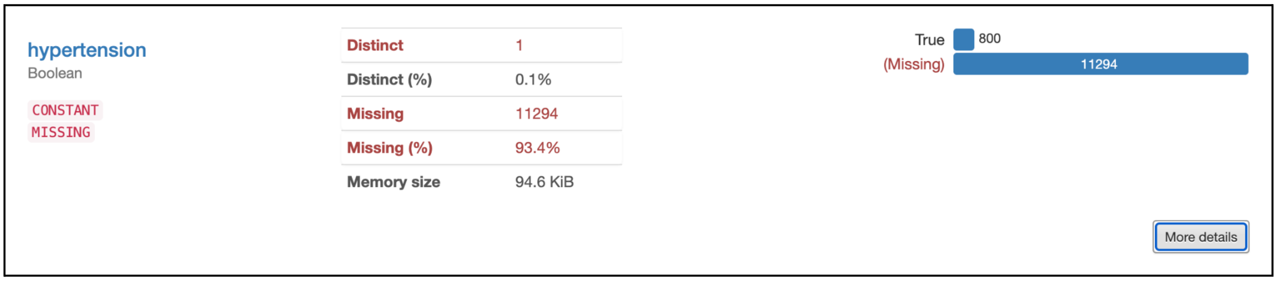
We think that these features will be included in our problem statement as our main object of analysis are th cough samples, with are desired feature of analysis being either the PCR test result or the extent to which background noise is present.

Preprocessing

For columns indicating comorbidities, we have a large number of NaN values - over 95%.

In this regard, this data cannot be used for analysis unless further conferred upon with the stakeholders - do NaNs translate to False, or is this data actually missing from the dataset?

The following picture demonstrates an example of one such comorbidity feature -



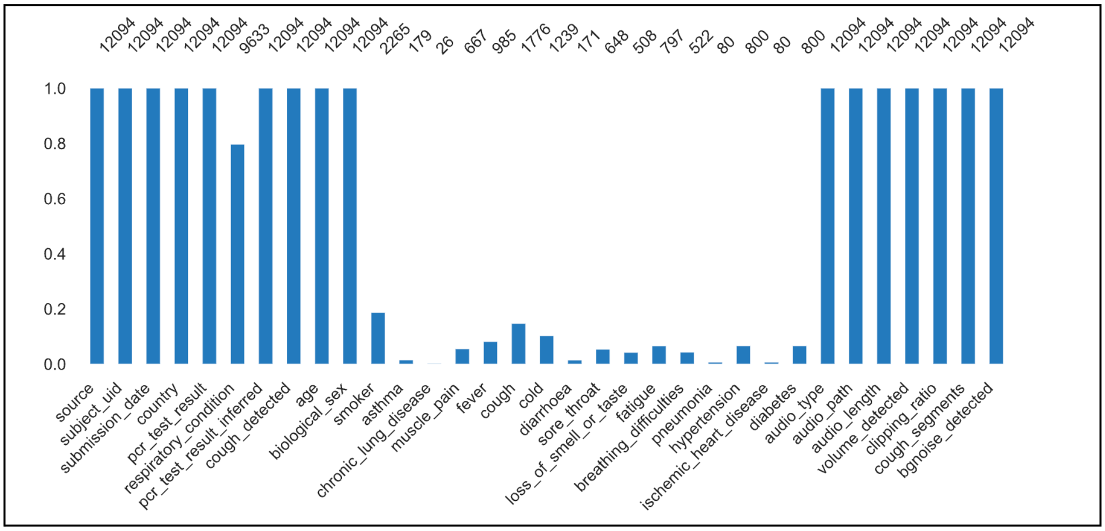
hypertension
Boolean

CONSTANT
MISSING

Distinct	1
Distinct (%)	0.1%
Missing	11294
Missing (%)	93.4%
Memory size	94.6 KiB

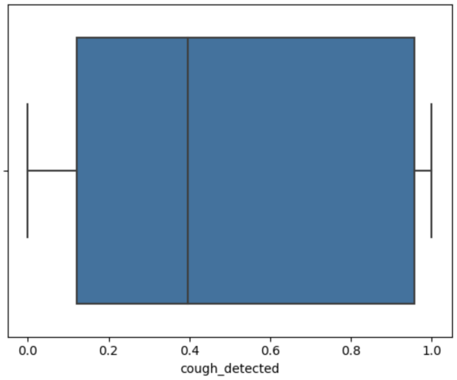
Missing Values

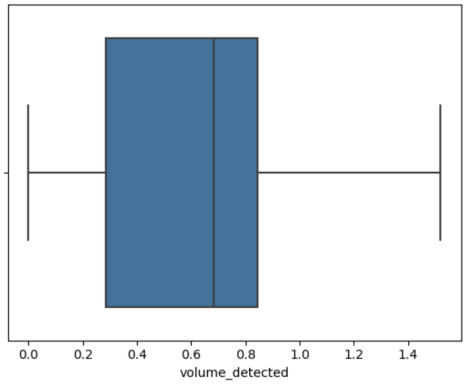
The graph below shows the percentage of missing values for each column attribute from our dataset.

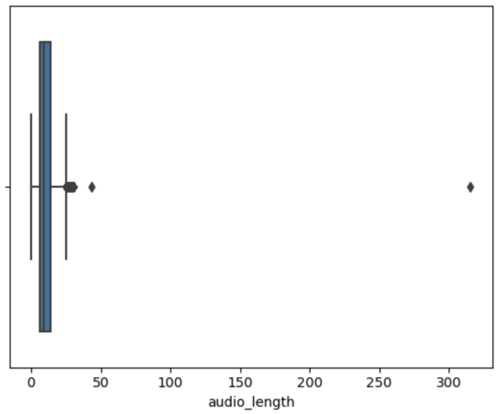


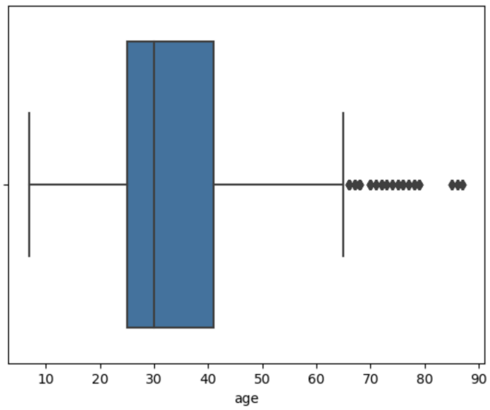
Outliers

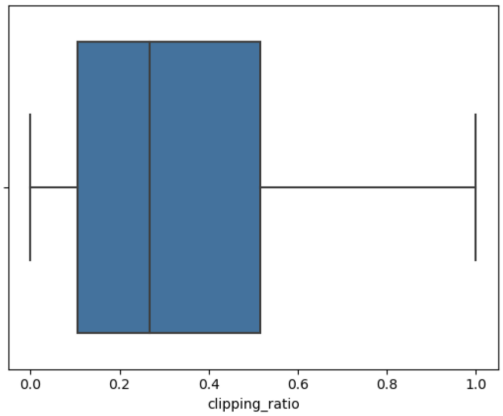
Outliers for the features of interest to us have been visualized as follows: -

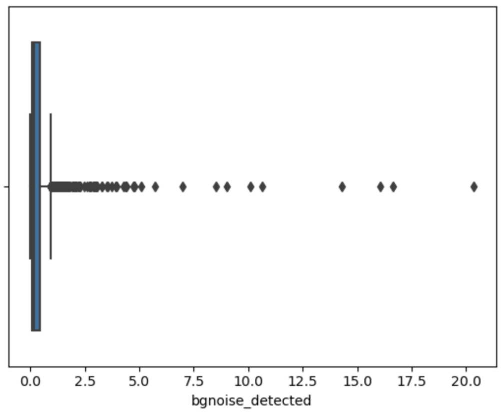




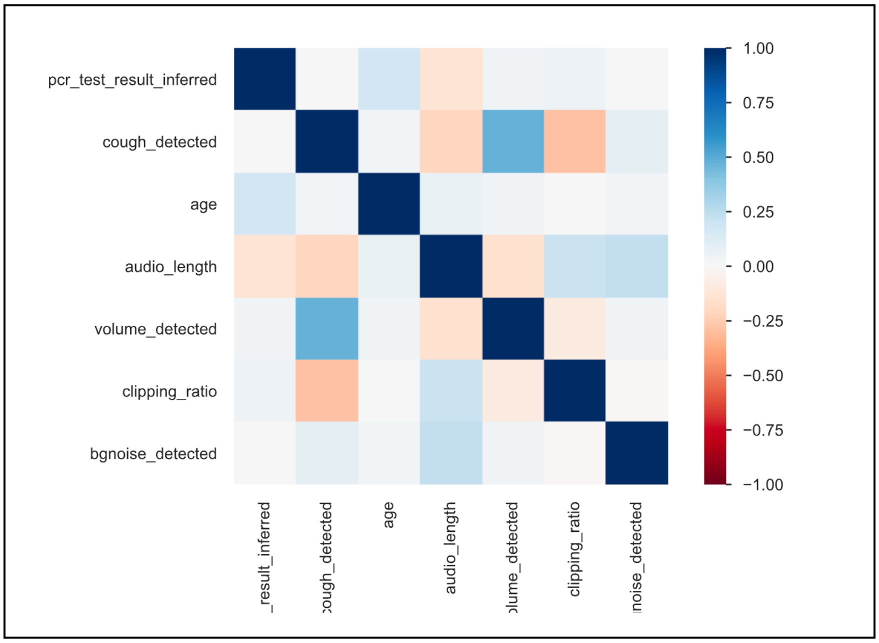








Correlated Features



Spearman's ρ

The Spearman's rank correlation coefficient (ρ) is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's r. It's value lies between -1 and +1, -1 indicating total negative monotonic correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation.

To calculate ρ for two variables X and Y, one divides the covariance of the rank variables of X and Y by the product of their standard deviations.

Data Flaws

We have several datasets provided to us for exploration. There is one “master” dataset (the “Coswara” dataset), which includes demographic and COVID-specific symptomatic data about the cough samples collected for COVID detection. This is the dataset we will consider for exploration.

- While investigating these datasets, we noticed that some columns were split and combined differently across different datasets - there needed to be more consistency in this manner. There were also some naming inconsistencies - what was called “gender” in one dataset was called “biological_sex” in another.
- There are several columns that have very high percentages of missing values.
- No schema/dataset description provided.
- For one of the fields (bgnoise_detected), the values are real numbers with an unspecified range or magnitude. As a result, the meaning of these values aren’t entirely clear.