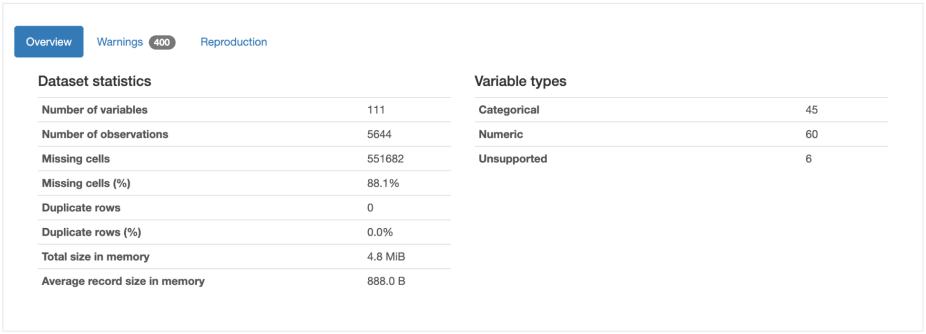
Team 2 Updates

Update 1 (9/15)

- Only very simple EDA done at this point
- Mentioned that we noticed grouping in lab results
- Pandas Profiling package (file too large to put on Github)
    - Demonstrated how the package is a starting point for EDA on datasets with many variables
    - Example information, data overview:



Overview | Warnings 400 | Reproduction

Dataset statistics

Number of variables	111
Number of observations	5644
Missing cells	551682
Missing cells (%)	88.1%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	4.8 MiB
Average record size in memory	888.0 B

Variable types

Categorical	45
Numeric	60
Unsupported	6

Update 2 (9/22)

- Presented data dictionary
    - Example:

Data dictionary

1. **'Patient ID'**-unique ID identifying an anonymized patient
2. **'Patient age quantile'**- bins representing the patients age,as interpreted on kaggle by others
    1 : 0-5
    2 : 6-10
    3 : 11-15
    4 : 16-20
    5 : 21-25
    6 : 26-30
    7 : 31-35
    8 : 36-40
    9 : 41-45
    10 : 46-50
    11 : 51- 55
    12 : 56-60
    13 : 61-65
    14 : 66-70
    15 : 71-75
    16 : 76-80
    17 : 81-85
    18 : 86-90
    19 : 91-95
3. **'SARS-Cov-2 exam result'**-Positive or negative PCR result, 558 positive results, categorical variable
4. **'Patient admitted to regular ward (1=yes, 0=no)'** self-explanatory
5. **'Patient admitted to semi-intensive unit (1=yes, 0=no)'**-semi-intensive or intermediate care unit, is usually the place to move improving ICU patients or deteriorating regular ward patients. The semi-intensive unit can be for a patient who is bad but not deteriorating rapidly.
6. **'Patient admitted to intensive care unit (1=yes, 0=no)'**- self explanatory, for the most critical, on the edge of death patients, ICU beds typically cost between $25,000 and $30,000. The cost of an ICU bed per night is $1,107, according to a recent study of two Washington hospitals.
7. **'Hematocrit'** -volume percentage of red blood cells in blood,

- Presented our lab test groupings based on 1) tests that appear grouped together in our data dictionary and 2) numbers of missing values
    - Data dictionary at that point had highlights grouping different lab groups together
- Data quality biggest issue at this point, which is why a majority of project spent trying to create a usable dataset for modeling

Update 3 (9/27)

- Did not present visuals but talked through our final dataset: COVID-19 positive patients with several new features indicating whether a set of lab tests were performed (based on previous week's presentation)
- Also decided to collapse our prediction problem from 4 classes (discharged, general, semi-intensive, ICU) to just admitted vs discharged due to lack of data & data quality)
- Mentioned combined over- & under-sampling
- Mentioned models to run being KNN & Random Forests