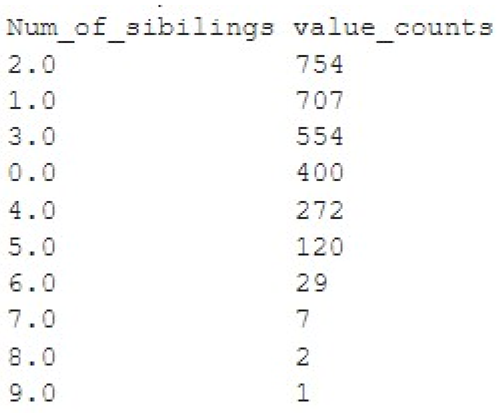
HW1 Report

Q1:

In the dataset we have 3000 rows and 29 columns.

Q2:



```
Num_of_sibilings value_counts
2.0              754
1.0              707
3.0              554
0.0              400
4.0              272
5.0              120
6.0              29
7.0              7
8.0              2
9.0              1
```

This feature refers in the real world to the number of siblings for an individual. The type of `Num_of_sibilings` is ordinal because the values of the feature obey an ordering of natural numbers. For example: someone with 5 has more siblings than someone with 6 siblings.

Q3:

Feature	Description	Type
patient_id	Id of the patient.	Ordinal
age	Age of the patient.	Ordinal
sex	Gender of the patient.	Categorical
weight	Weight of the patient.	Ordinal
blood_type	Blood type of the patient.	Categorical
address	Address of the patient.	Other
current_location	Coordinates of the current location of the patient.	Other
job	Job of the patient.	Other
num_of_siblings	Number of siblings of the patient.	Ordinal
happiness_score	A measure of the patient's happiness	Ordinal
household_income	The patient's household income.	Continuos
pcr_date	The date of the patient's PCR test.	Ordinal
symptoms	The patient's symptoms	Other
conversations_per_day	The number of conversations a patient has per day	Ordinal
sugar_levels	The sugar levels of the patient.	Ordinal
sport_activity	The number of sports activities the patient has been taking part in recently.	Ordinal
PCR_01	The results of the PCR_01 of the patient.	Continuos
PCR_02	The results of the PCR_02 of the patient.	Continuos
PCR_03	The results of the PCR_03 of the patient.	Continuos
PCR_04	The results of the PCR_04 of the patient.	Continuos
PCR_05	The results of the PCR_05 of the patient.	Ordinal
PCR_06	The results of the PCR_06 of the patient.	Continuos
PCR_07	The results of the PCR_07 of the patient.	Continuos
PCR_08	The results of the PCR_08 of the patient.	Continuos
PCR_09	The results of the PCR_09 of the patient.	Continuos
PCR_10	The results of the PCR_10 of the patient.	Ordinal

Q4:

Feature Name	Type	Explanation
patient_id	ordinal	ID is by definition an ordinal number, because it counts the people and each number represents different values (people)
age	ordinal	The data in this column counts (discreetly) the age of the patient
weight	ordinal	The data in this column counts (discreetly) the weight of the patient
address	other	Represents free-text data
current_location	other	Is represented by vectors in a two dimensional vector space, on which the concept of natural order is not defined between the vectors
job	other	Represents free-text data
num_of_siblings	ordinal	Counts how many siblings a specific patient has
happiness_score	ordinal	A score which is represented by natural numbers and therefore we can order it and the data type is ordinal
pcr_date	ordinal	Patients can be ordered and aggregated according to specific PCR dates and therefore is ordinal
symptoms	other	Is not categorical because the data is a combination (not selection) from the symptoms category
conversations_per_day	ordinal	Counts the patient's conversations per day by natural numbers
sugar_levels	ordinal	Counts the patient's sugar levels by natural numbers
sport_activity	ordinal	Counts the patient's sport activities by natural numbers
PCR_05	ordinal	Represent some test results by natural numbers, each of them has a meaning so the data type is ordinal
PCR_10	ordinal	Represent some test results by natural numbers, each of them has a meaning so the data type is ordinal

Q5:

It is important to use the same split for all the analysis because if we do not, we risk over-fitting our training models to our data,since every time we calculate parameters and hyper-parameters from the analysis done on the data, it may be on data that in a previous iteration was the test data.

Q6:

The length of the OHE vector is 8.

Q7:

Yes, it is possible to extract useful categorical features from the given `symptoms` feature. In a method similar to that of OHE, we add new features for each symptom that appears in the original `symptom` feature. We then place a 1 or 0 in the appropriate column based on whether the patient had that specific symptom. It is possible, therefore, for a specific patient to have a symptom vector containing more than one hot bit.

Q8:

Original Feature	**Extracted Feature**	**Explanation**
address	U.S. state	The location of the patient in terms of what state they reside in may be useful in determining the target features. The city, street name, and postal code as features would have produced too many categories to be useful, in addition to producing duplicates.

Q9:

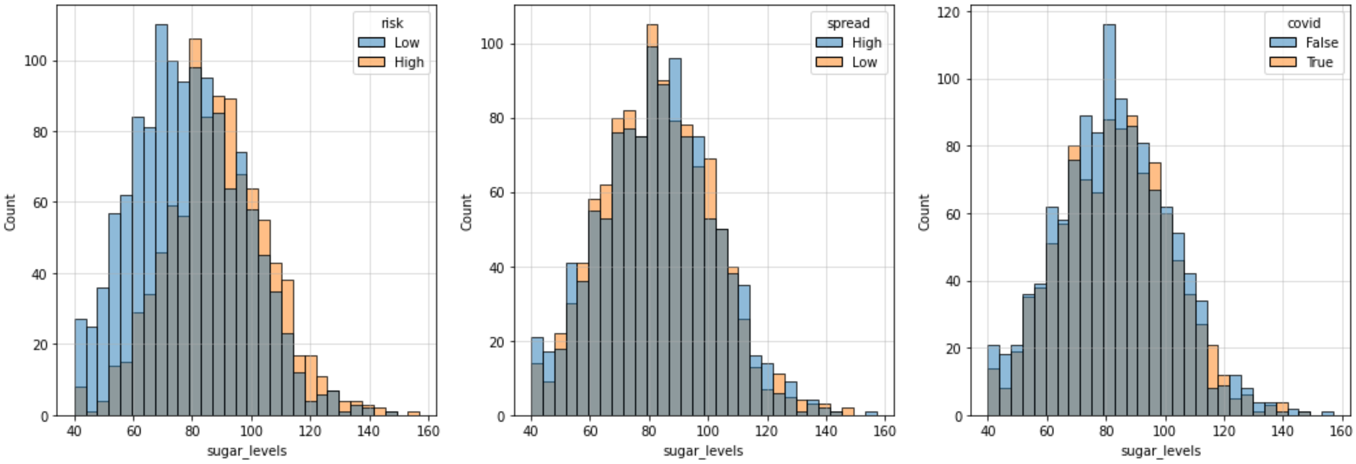


Figure 1: Histograms of patients' sugar levels according to infection of covid, risk of contracting covid, and potential for spreading covid

First of all, we can see in figure 1 that all the patients with sugar level below 80 have an approximately $\frac{2}{3}$ chance of being at a low risk of contracting covid, and ∼52% of them don't have covid. The peak is at a sugar level between 80-84: approximately 58% don't have covid. From sugar level 80 and on, the majority have a high risk but still test false for covid. We can also see a contextual outlier: a patient who has a sugar level of 155, high risk and level of spread but tests false for covid. Nevertheless, The spread plot is inconclusive. The largest group, containing approximately 200 patients, have sugar levels in the range of 80-84.

Q10:

For normal distributions, we employ a formula that relies on the STD of the distribution in order to set lower and upper bounds in-order to label outliers. Therefore, for histograms with a normal distribution, those with a tight (small) variance and therefore a smaller STD will set lower upper bounds and higher lower bounds, and therefore are more likely to contain outliers. For skewed distributions, we rely on the IQR rule to detect outliers, which in its formula utilizes IQR and 25th and 75th percentiles to set minimum and maximum bounds. Therefore, histogram with a skewed distribution and a more compact IQR will have higher minimum and lower maximum bounds, and are therefore more likely to contain outliers.

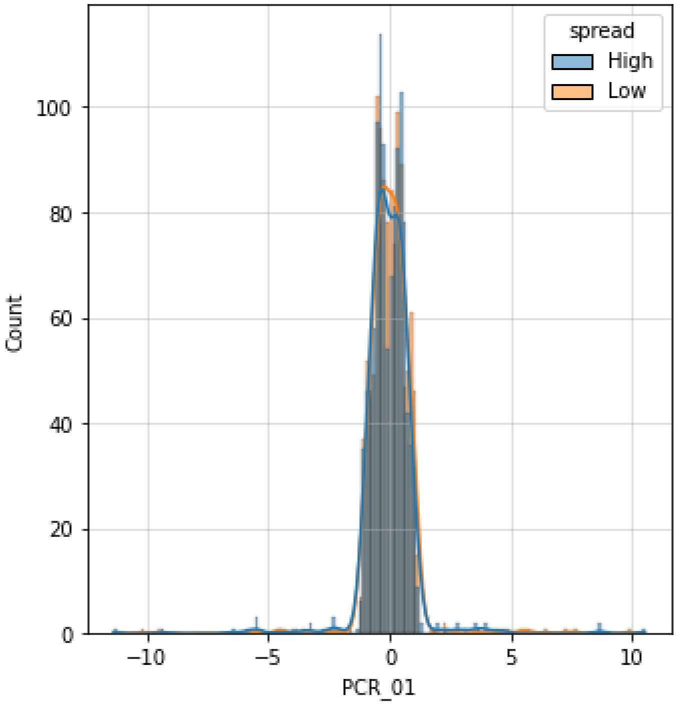


Figure 2: Histogram of PCR_01

The histogram in figure 7 of the `PCR_01` feature approximates a normal distribution with a small variance and therefore is more likely to have outliers, for example, the patient with a `PCR_01` and a high probability of spreading covid.

Q11:

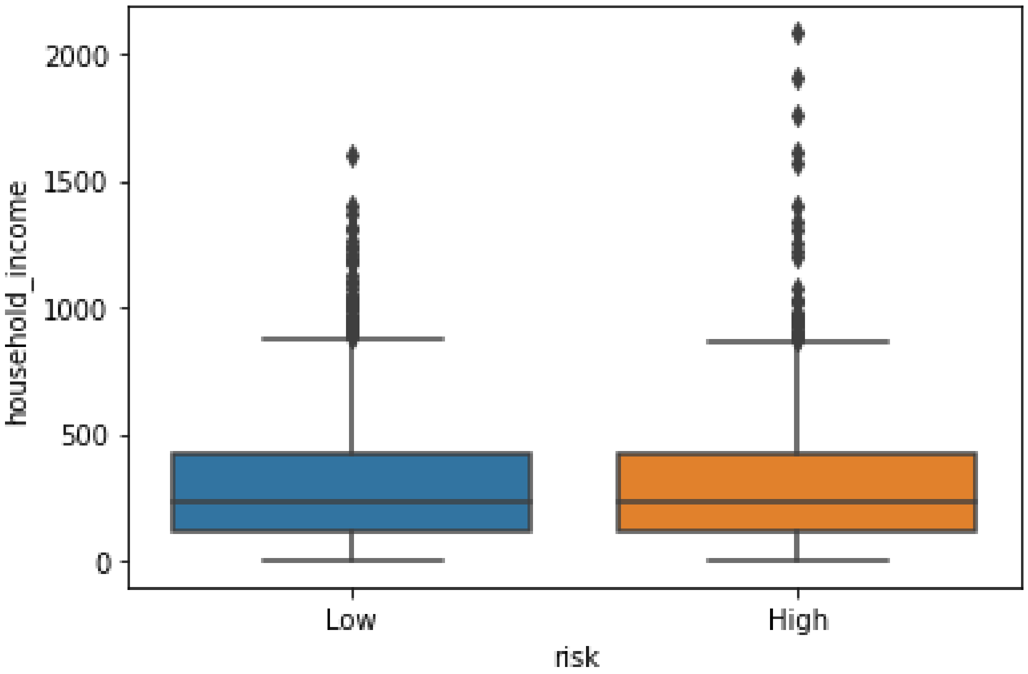


Figure 3: Box plot of household income grouped by risk

As can be seen in figure 3, there is no meaningful difference in the number of outliers between the risk levels. In addition, these outliers do not obviously appear to be erroneous data, and therefore any technique used to clean these data points should reflect in some sense the original values of the data points. Therefore, a reasonable technique to be used for these outliers would be to replace these values with blanks and thereafter treat them as missing data.

Q12:

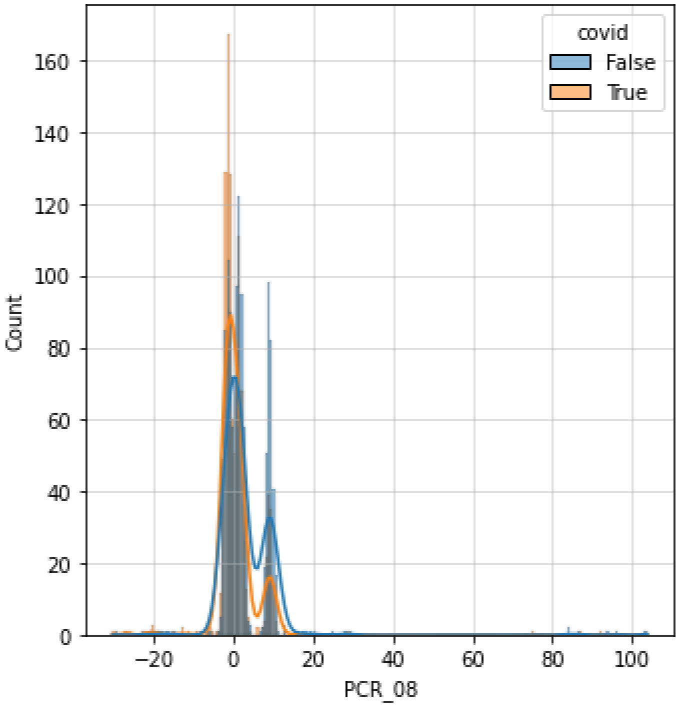


Figure 4: Histogram of PCR_08

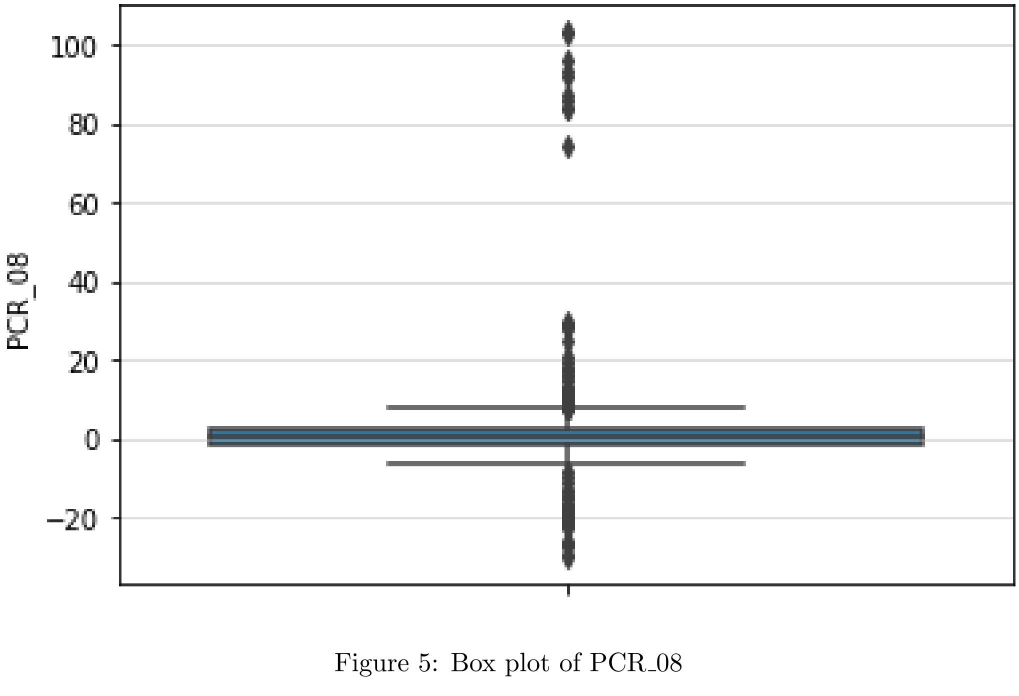


Figure 5: Box plot of PCR_08

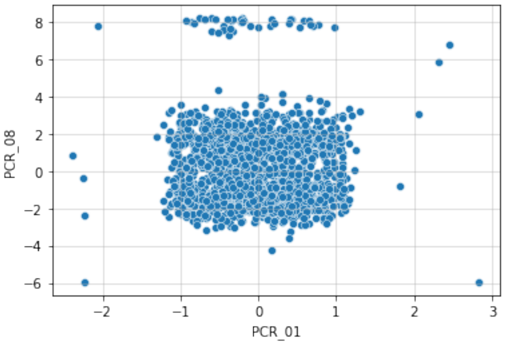


Figure 6: Scatter plot of PCR_08 and PCR_01 following global outlier removal

One of the features in which was found outlier values was PCR_08. Since, as can be seen in figure 4 that the feature does not follow a normal distribution, we decided to clean its global outliers that are clearly visible in figure 5 by reassigning to NaN all the values that fall on the outside of the IQR, and thus to treat them as missing values. We then noticed in the scatter plot in figure 6 two set of contextual outliers: Those that were above the value of 6 for PCR_08 ad those that were below -5. However, we noticed that the former group formed a relatively tight cluster, indicating a set of data points with possible significance, and therefore chose to not consider outliers. Whereas the latter group in comparison were more sparsely scattered and therefore we treated as outliers and applied to them the NaN labeling technique as was done to the global outliers. We used this particular method for both types of outliers because the reasoning for the existence of the outliers is not known, and therefore the data could not simply be removed.

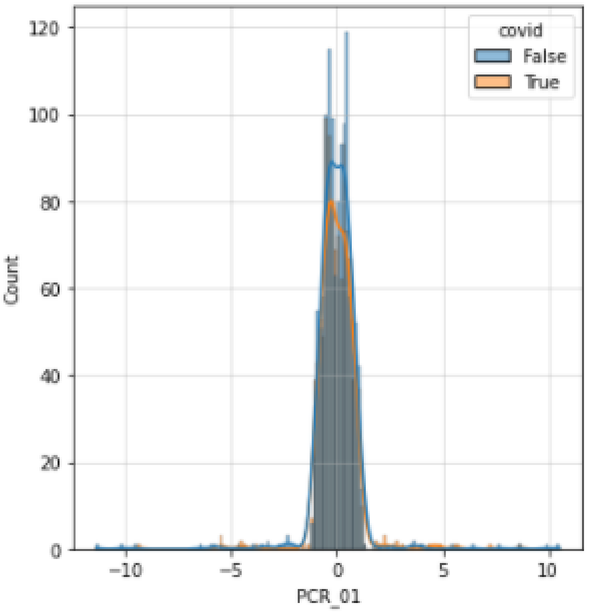


Figure 7: Histogram of PCR_01

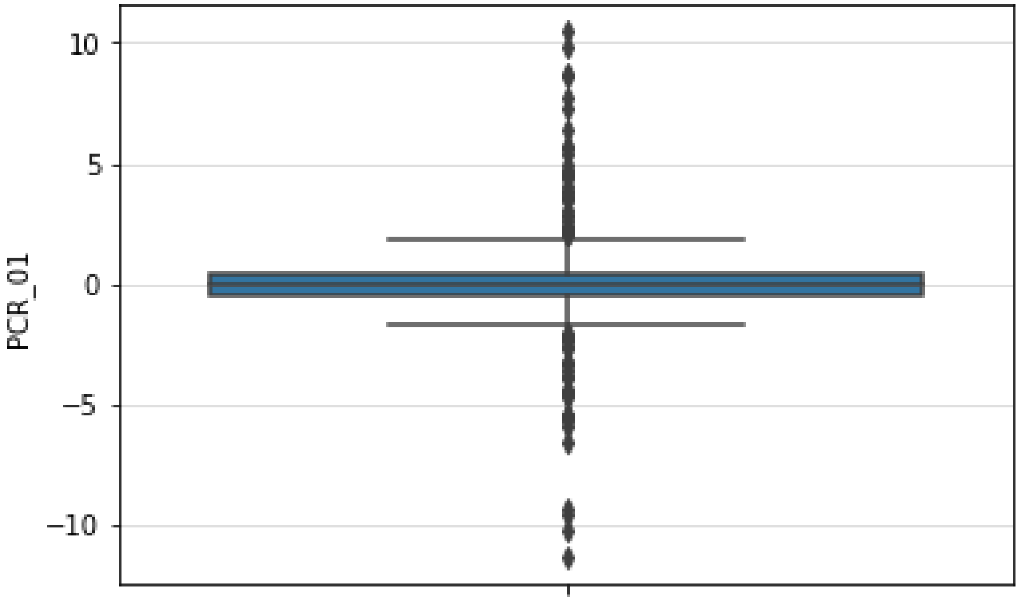


Figure 8: Box plot of PCR_01

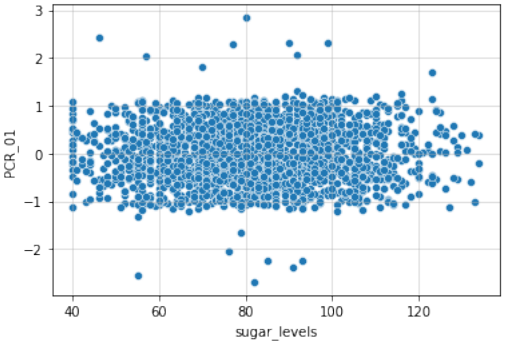


Figure 9: Scatter plot of PCR_01 and sugar_levels following global outlier removal

Another feature in which was found outlier values was PCR_01. Since, as can be seen in figure 7 that the feature does follow a normal distribution, we decided to clean its global outliers that are clearly visible in figure 8 by reassigning to NaN all the outliers found via use of the z-score technique, and thus to treat them as missing values. We then noticed in the scatter plot in figure 9 two sets of contextual outliers: Those that were above the value of 1.5 for PCR_01 ad those that were below -1.5. However, in this case we noticed that both sets were sparsely scattered and therefore we treated as outliers and applied to them the NaN labeling technique as was done to the global outliers. We used this particular method for both types of outliers because the reasoning for the existence of the outliers is not known, and therefore the data could not simply be removed.

Q14:

One advantage for using median imputation is in its ease of implementation. One disadvantage is that it distorts the original variable distribution and variance.

Q15:

By using "missing category" technique on a variable, we gain an additional label for that categorical variable "missing", which is applied to all the data points in that variable that where missing a value.

Q16:

We chose the median imputation technique because of the following reasons:

- The missing values are numerical, meaning we will impute them with a numerical value and we want the data's distribution to be kept as much as possible without using trimming so in that case, median imputation does the minimal changes to the distribution (by minimal change to the mean, median,...).
- Our data is missing at random because there's no informative logic which explains the missing values and the samples with the missing values (of num_of_siblings) looks like the majority of the observations of the other samples.
- The missing data is 6.708333333333333% (while $5\frac{1}{3}\%$ is really missing and the other samples are outliers we decided to fill with NaN) of the training data, it's in range of the 5% boundary and not too far from it.

Q17:

We would like to use `RandomSampleImputer` to get results (within the imputation) that will keep the data's distribution as it was before the imputation.

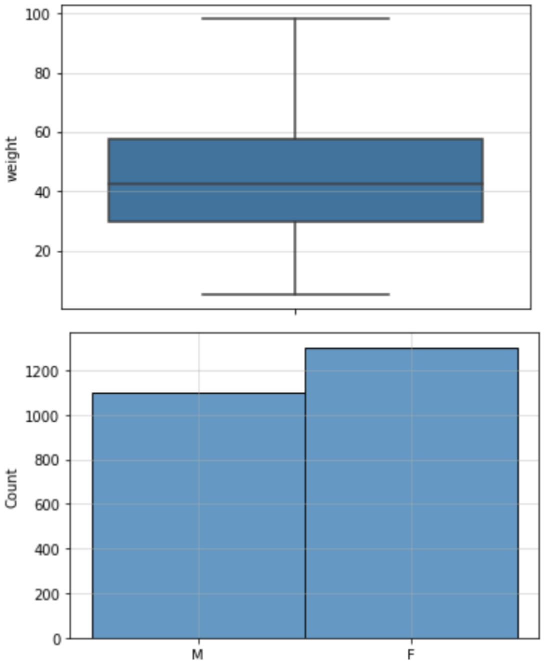


Figure 10: Histograms for weight and sex before imputation

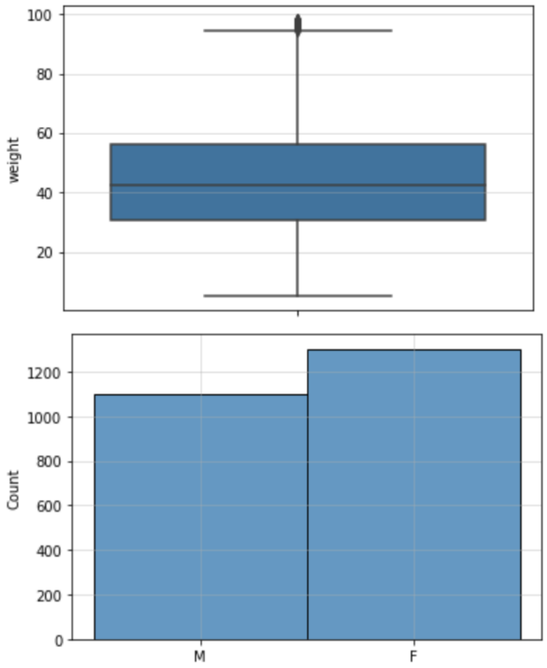


Figure 11: Histograms for weight and sex after imputation

Weight: We can see that there's more outliers in the plot which represent the weight column after the imputation because we filled the NaN values with the column's median, which caused the IQR limits to get shorter because there is more mass on the center of the values so the data behaves more strictly, resulting in more global outliers.

Sex: As shown (and counted) the differences is within 15 samples because we imputed this column's missing data using RandomSampleImputer because we wanted to keep its proportion (distribution) for more reliable analysis.

Q18:

We can see that if we clean outliers after data impute, there could be some issue:

- Outliers could include extremely high/low ordinal/continuous values, which affects our distribution parameters (mean, median,...) so if we'll choose to impute data based on median/mean technique, the chosen parameter's calculations will include the outlier's extreme values, which we want to avoid since we removing outliers anyway.
- Bivariate outliers can be a group which has a majority of some groups over others, meaning that if we'll clean after the imputation, we could mistakenly impute the wrong majority over missing values. For example: 7 samples which 3 of them describe females and 4 of them describe males but 2 of the samples are bivariate outliers so after cleaning them we left with a different majority group.

Q19:

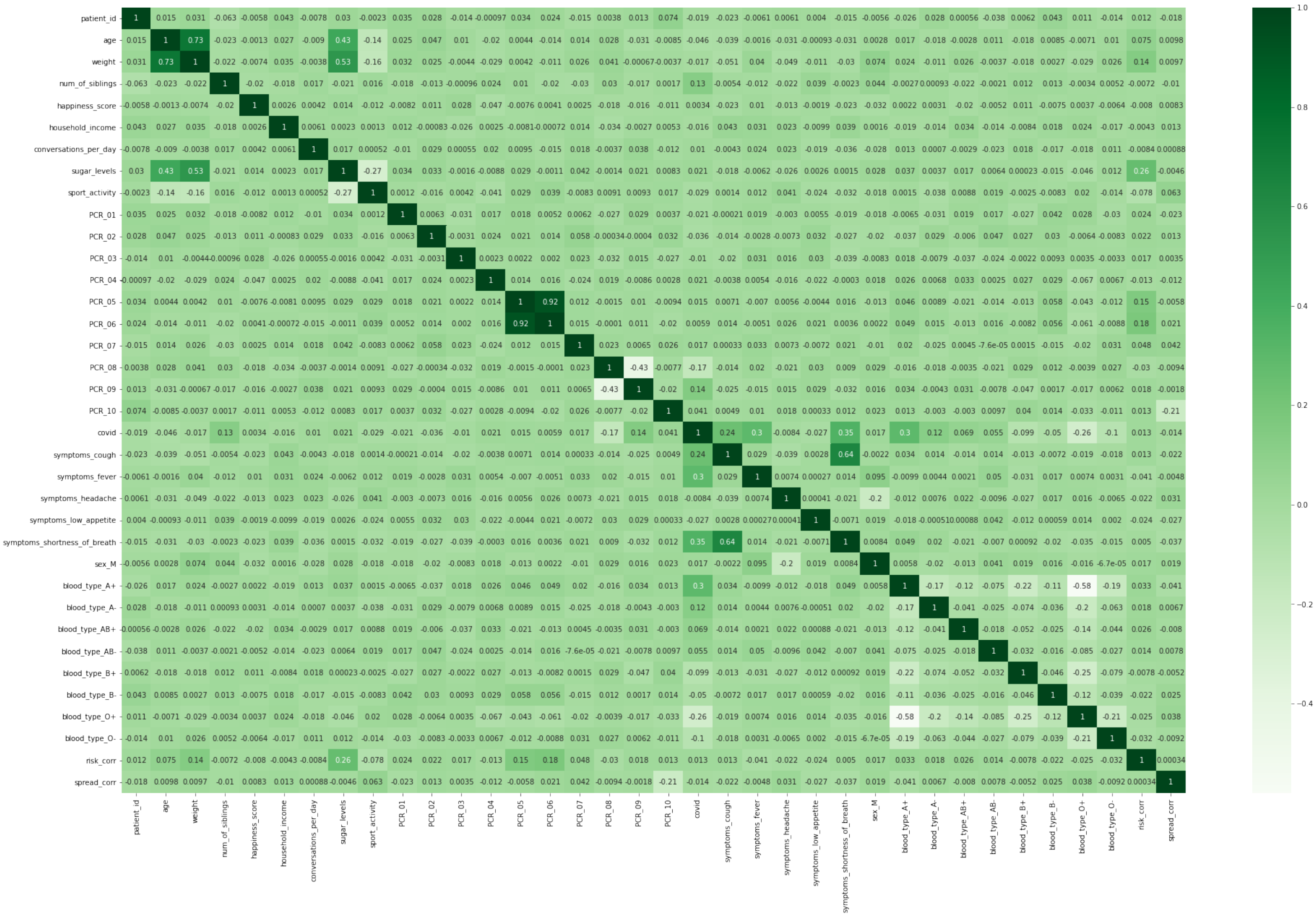


Figure 12: Correlation heat map of dataset

Purely from looking at the row for PCR_10 in figure 12 it is clear that there are no redundant features, as the highest absolute value of correlation is -0.21 with the target feature spread, which is according to our estimation a value that does suggest a linear relationship between values.

Q20:

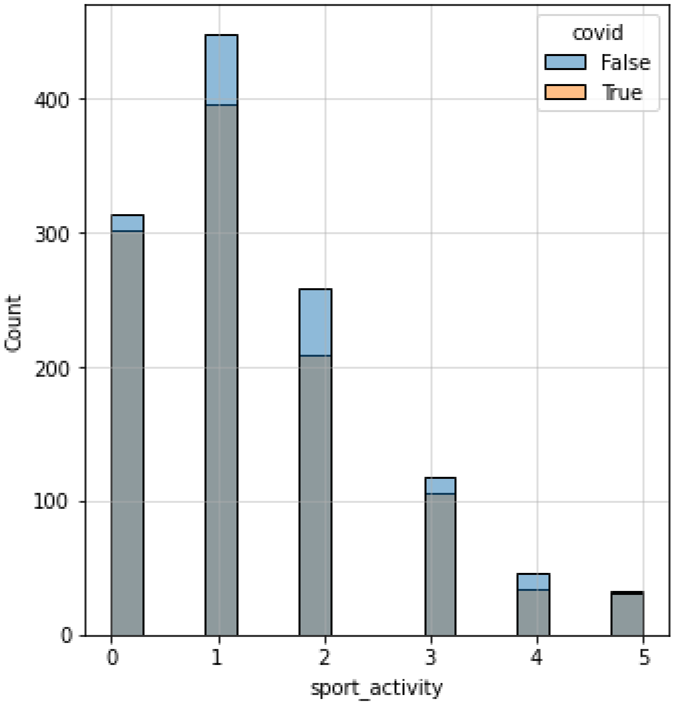


Figure 13: Histogram plot from univariate exploration stage of sport_activity according to covid

Solely based off of the histogram in figure 13, it would be extremely difficult to impossible to classify covid according to the sport_activity feature, since for every level of sport activity, there is a near perfect overlap between the number of patients with and without covid.

Q21:

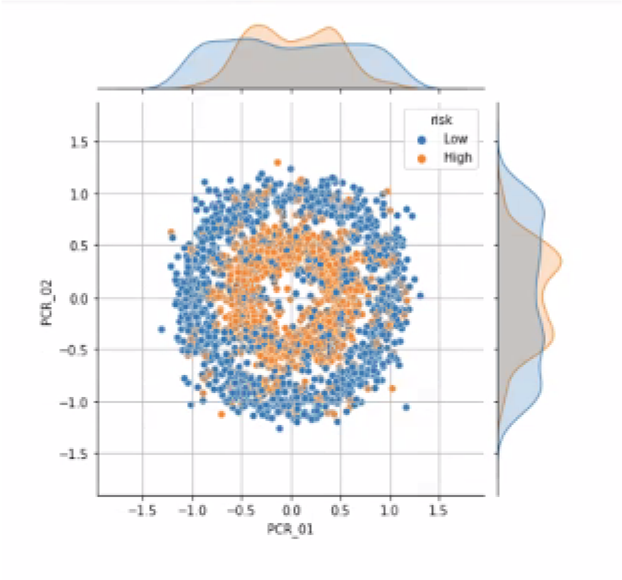


Figure 14: Jointplot of PCR_01 and PCR_02 according to risk

We can see from the jointplot in figure 14 that in the histograms for both PCR_01 and PCR_02 the high and low risk labels completely overlap with each other and therefore it would be early impossible to use either of these features on their won to classify risk. On-the-other-hand, as can be seen from the scatter portion of the jointplot in figure 14, the plot is mostly separable into radiuses, which will be useful in the next stage of the machine learning pipeline.

Q22:

Interesting Findings from Univariate Analysis:

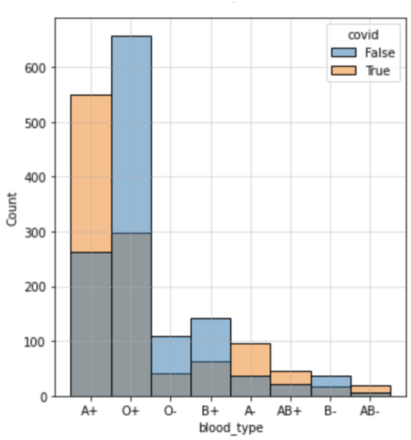


Figure 15: Histogram of blood_type according to covid

blood_type: From the histogram in figure 15 we noticed a relationship between certain blood types and the number of patients with covid: For blood types A+, A-, AB+, and AB-, we perceived that approximately at least 75% percent of patients tested positive for covid, whereas the opposite was true for the remaining blood types. Therefore, we found it appropriate to create a new binary features out of the existing blood_type feature: blood_A_AB, where a '1' in blood_A_AB indicates that the patient has a blood type of A+, A-, AB+, or AB-, and a '0' otherwise.

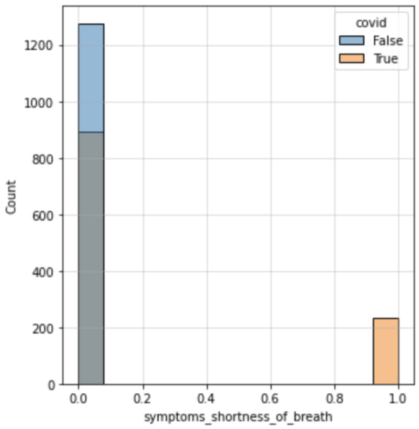


Figure 16: Histogram of symptoms_shortness_of_breath according to covid

breath: In the histogram in figure 16 we can see that 100% of patients with shortness of breath tested positive or covid. However, in the population of those patients without that symptom, although there where more patients that tested negative than positive, the ratio was still too close to consider this feature useful for predicting presence or absence of covid in a patient.

Interesting Findings from Bivariate Analysis:

We noticed that the Pearson covariance between a number of PCR features and other features decreased (in terms of distance from 0), after the global outlier cleaning was applied to them. On the surface it seemed that perhaps the cleaning performed was too harsh. However on close inspection of the scatter plots of the PCR features that were affected, we noticed that sparsely dispersed outliers were the main factors in the existence of the initial high covariance.

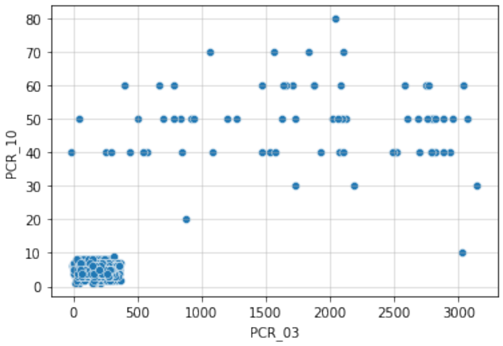


Figure 17: Scatter plot of PCR_03 versus PCR_10 before application of global outlier cleaning

For instance, in the scatter plot in figure 17 we noticed that the cause of the high Pearson correlation between PCR_10 and PCR_03 (0.8) was due to the existence of a seeming linear relationship of the the points. However this relationship is mainly due to a relatively small group of sparse outlying points as opposed to the cluster in the lower left hand portion of the graph. Therefore, although the decrease in correlation post cleaning may seem radical, it is in-fact justified.

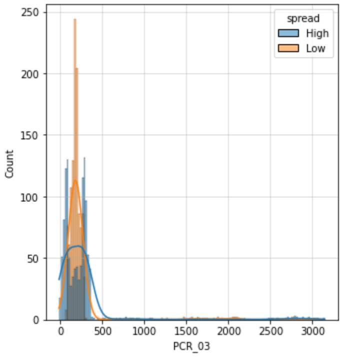


Figure 18: Histogram of PCR_03 with respect to spread

From the histogram plot in figure 18 we noticed a high concentration of data points with low spread relative to high spread in the range of PCR_03 levels of approximately 100 to 300, and vice-versa everyhwere else. From this we concluded that PCR_03 was a seperatable feature with regards to the spread target, and therefore beneficial to maintain as a feature.

Furthermore, from the jointplot of PCR_01 and PCR_02 in figure 14 as previously mentioned in question 22, we can conclude that the features PCR_01 and PCR_02 will be beneficial to maintain as well.

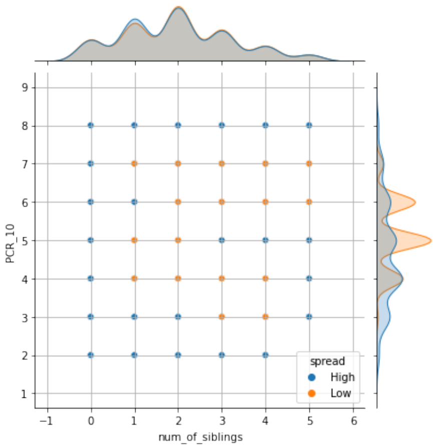


Figure 19: Joint plot of PCR_10 versus siblings with respect to spread

In the joint plot in figure 19 we noticed a mostly separable form involving an outer square edge of mostly high spread and an inner square of mostly low spread. Therefore, we concluded that also the features PCR_10 and num_of_siblings will be beneficial to keep.

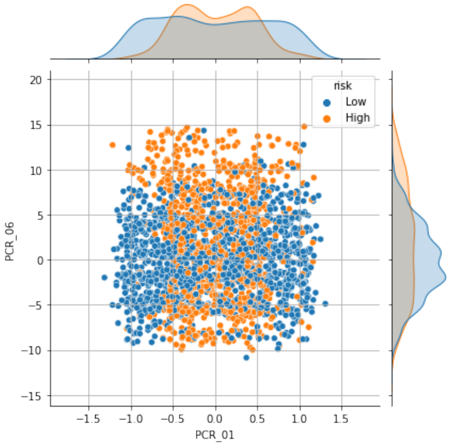


Figure 20: Joint plot of PCR_06 versus PCR_01 with respect to risk

In the joint plot in figure 20 we noticed a mostly separable form similar to the letter "H" where the "H" itself is made up of a high proportion of points with low risk surrounded by clusters of points of high risk. Therefore, we concluded that also the features PCR_06 and PCR_01 will be beneficial to keep.

Q23:

Feature Name	Keep	New	Explanation
patient_id	X	X	Patient's ID has no inherent meaning
age	X	X	Age has a high Pearson correlation with the weight feature and therefore is being dropped
sex	X	X	Applied OHE to this feature because it contains strings
weight	X	X	No correlation with target features
address	X	X	'other' type feature with no discernable derivable features
current_location	X	X	'other' type feature with no discernable derivable features
job	X	X	'other' type feature with no discernable derivable features
num_of_siblings	V	X	Separable graph in conjunction with PCR_10 with respect to spread target
happiness_score	X	X	No correlation with target features
household_income	X	X	No correlation with target features
pcr_date	X	X	'other' type feature with no discernable derivable features
symptoms	X	X	'other' type feature, which on its own has no significance
conversations_per_day	X	X	No correlation with target features
sugar_levels	X	X	No correlation with target features
sport_activity	X	X	No correlation with target features

Feature Name	Keep	New	Explanation
PCR_01	V	X	Separable graph in conjunction with PCR_02 with respect to risk target
PCR_02	V	X	Separable graph in conjunction with PCR_01 with respect to risk target
PCR_03	V	X	Separable graph with respect to spread target
PCR_04	X	X	No correlation with target features
PCR_05	X	X	High correlation with PCR_06
PCR_06	V	X	Separable graph in conjunction with PCR_01 with respect to risk target
PCR_07	X	X	No correlation with target features
PCR_08	X	X	No correlation with target features
PCR_09	X	X	No correlation with target features
PCR_10	V	X	Separable graph in conjunction with num_of_siblings with respect to spread target
blood_type	X	X	Applied OHE to this feature because it contains strings
Male	X	V	No correlation with target features

Feature Name	Keep	New	Explanation
blood_type_A+	X	V	No correlation with target features
blood_type_A-	X	V	No correlation with target features
blood_type_B+	X	V	No correlation with target features
blood_type_B-	X	V	No correlation with target features
blood_type_AB+	X	V	No correlation with target features
blood_type_AB-	X	V	No correlation with target features
blood_type_O+	X	V	No correlation with target features
blood_type_O-	X	V	No correlation with target features
symptoms_cough	X	V	No correlation with target features
symptoms_fever	X	V	No correlation with target features
symptoms_headache	X	V	No correlation with target features
symptoms_low_appetite	X	V	No correlation with target features
symptoms_shortness_of_bre	X	V	No correlation with target features
address_states_**	X	V	No correlation with target features
blood_A_AB	V	V	Correlation found to covid target, as explained in question 22