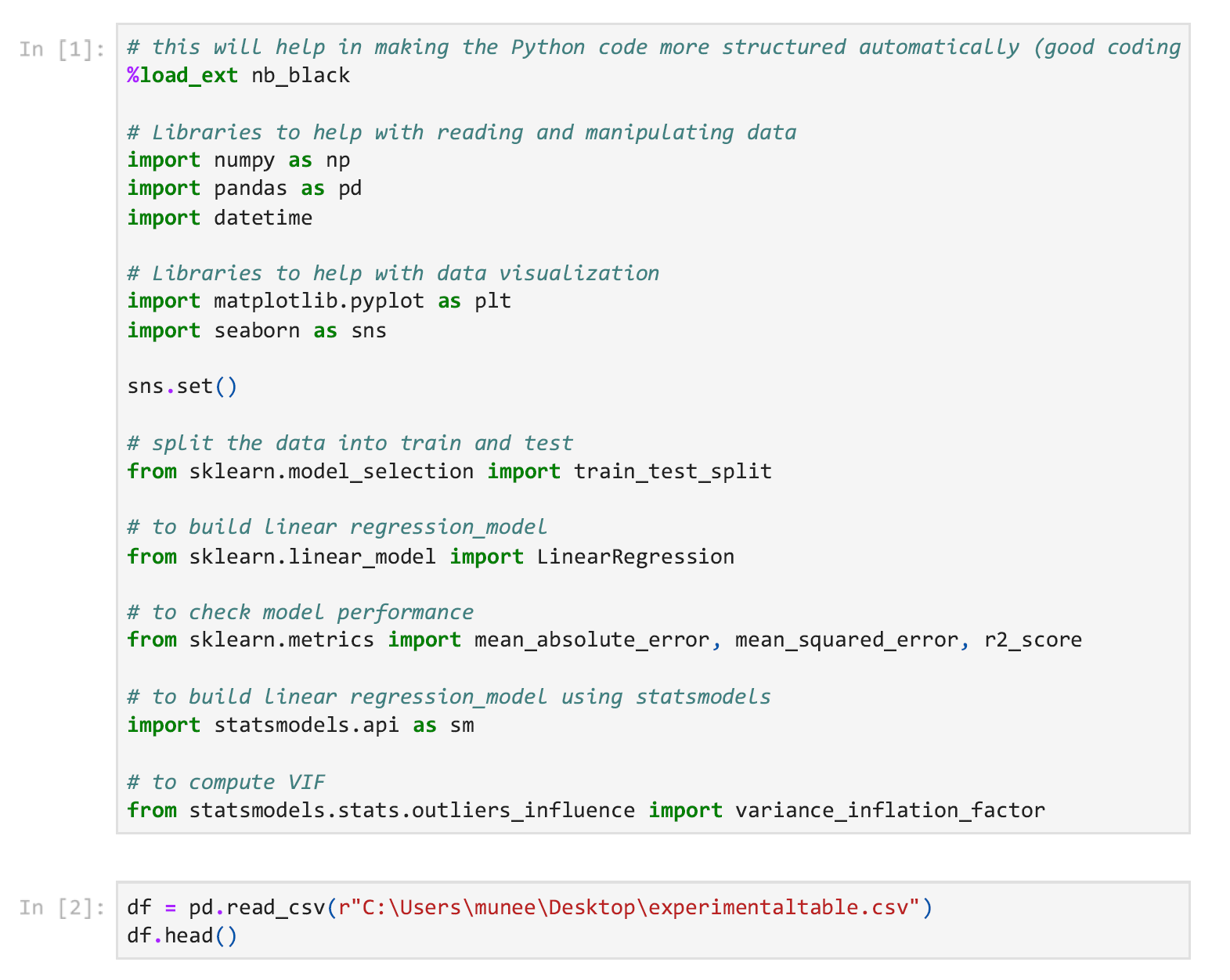
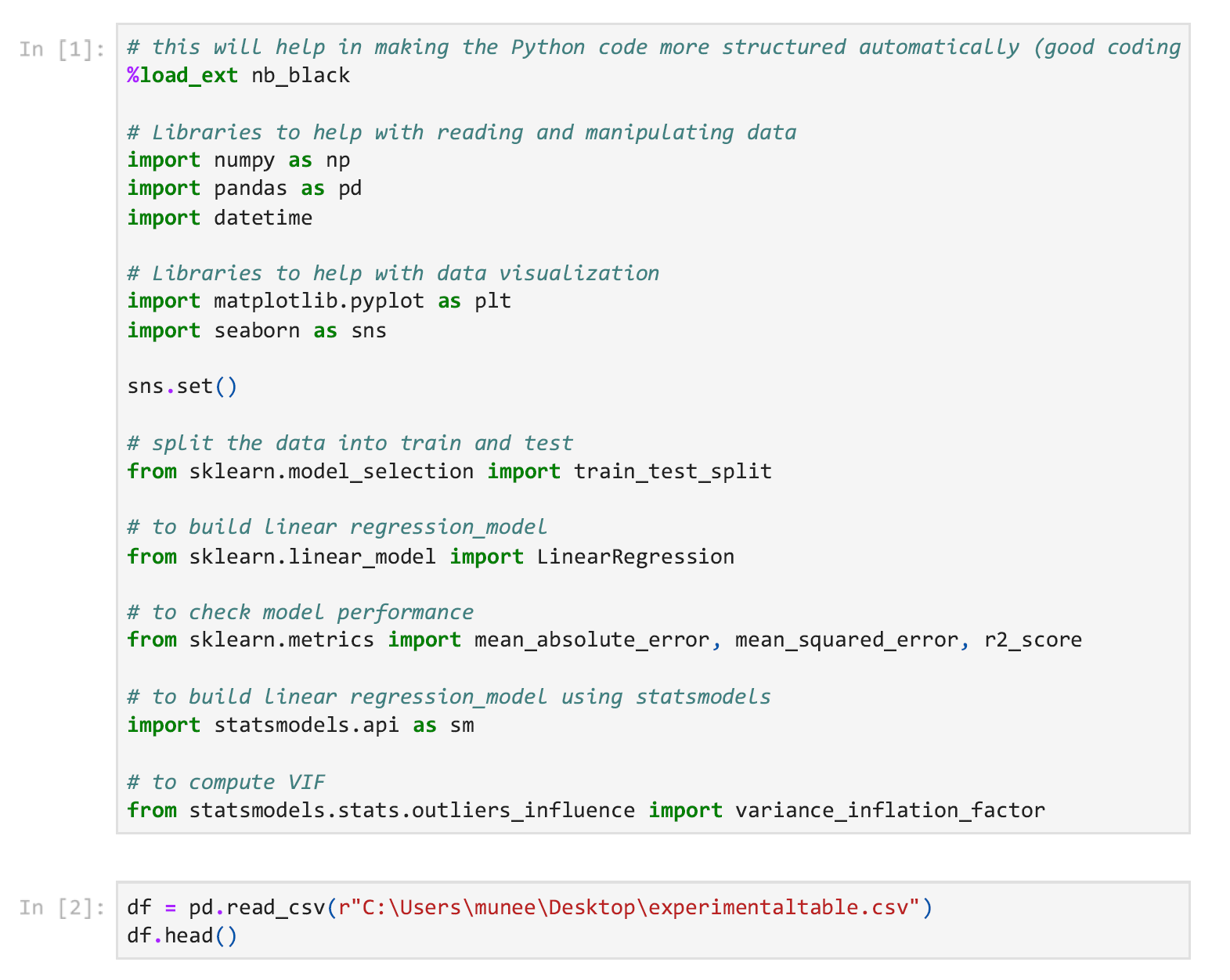
SENSITIVITY AND POSITIVE PREDICTIVE VALUES OF ICD-10 CODES AMONG COVID-19 PATIENTS FROM OPTUM DATABASE IN THE UNITED STATES: A RETROSPECTIVE COHORT STUDY

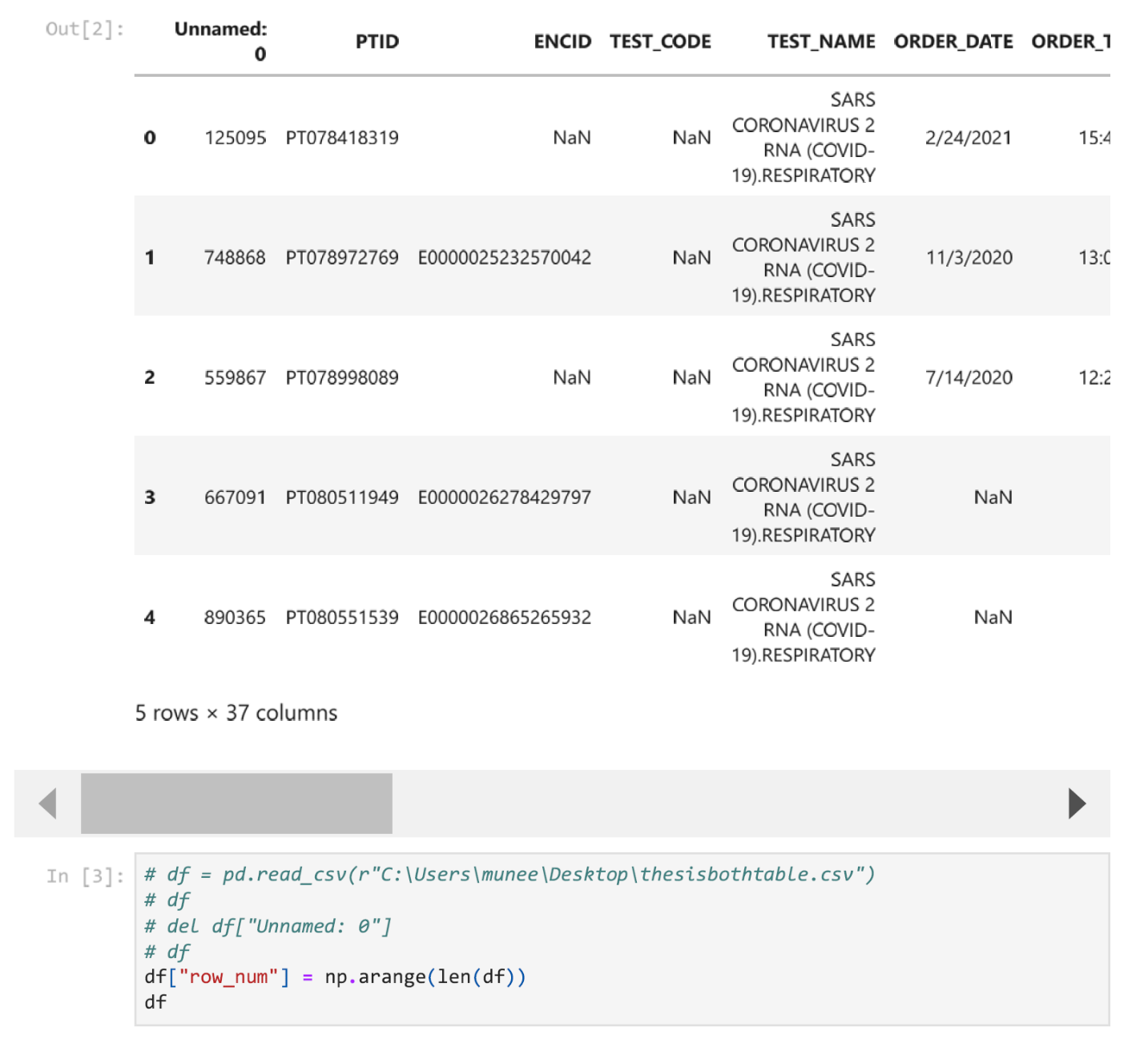
***Programming Component***

Step 1: Import libraries

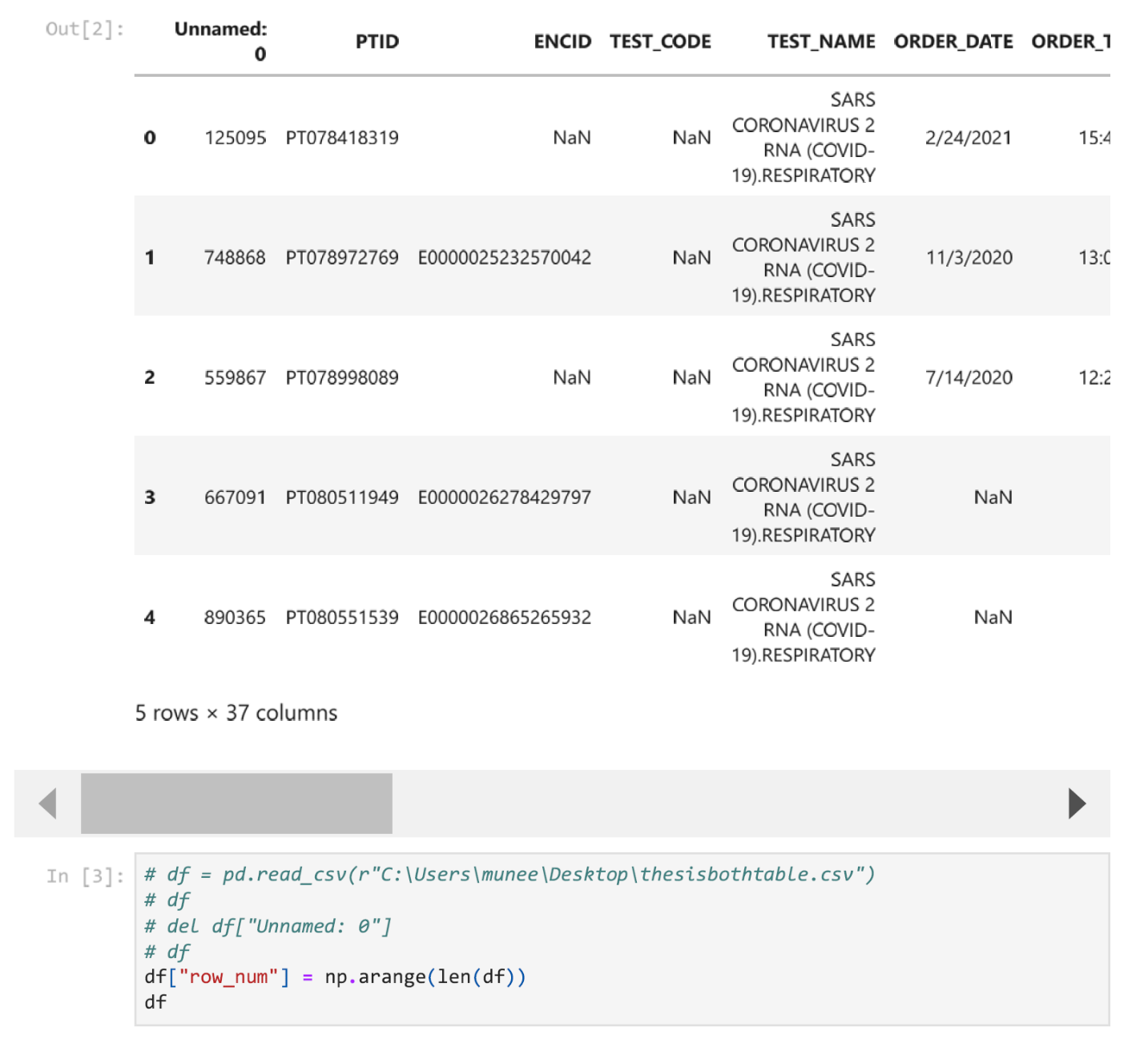


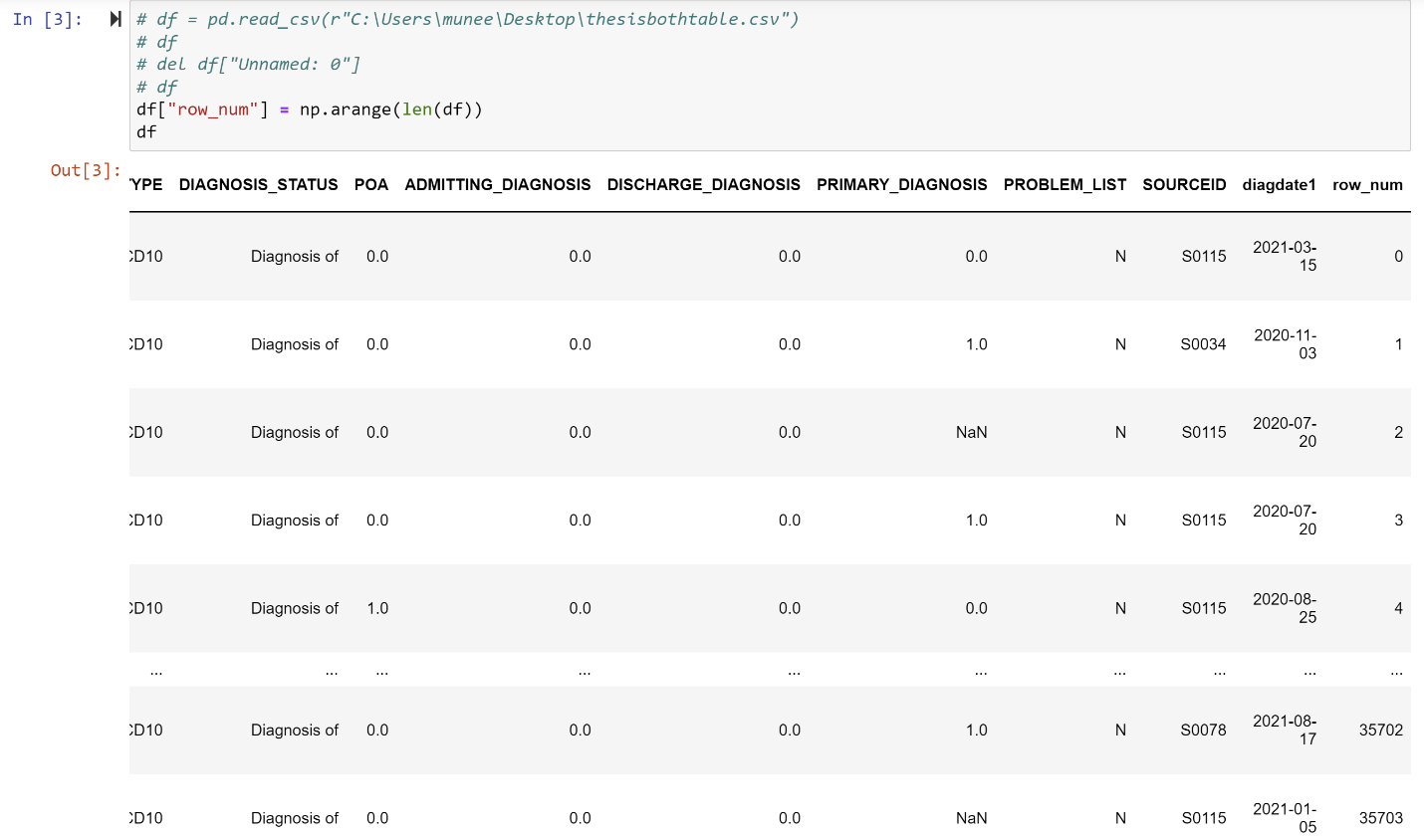
Step 2: Import csv file. This csv file is from UTHealth COVID Optum database.



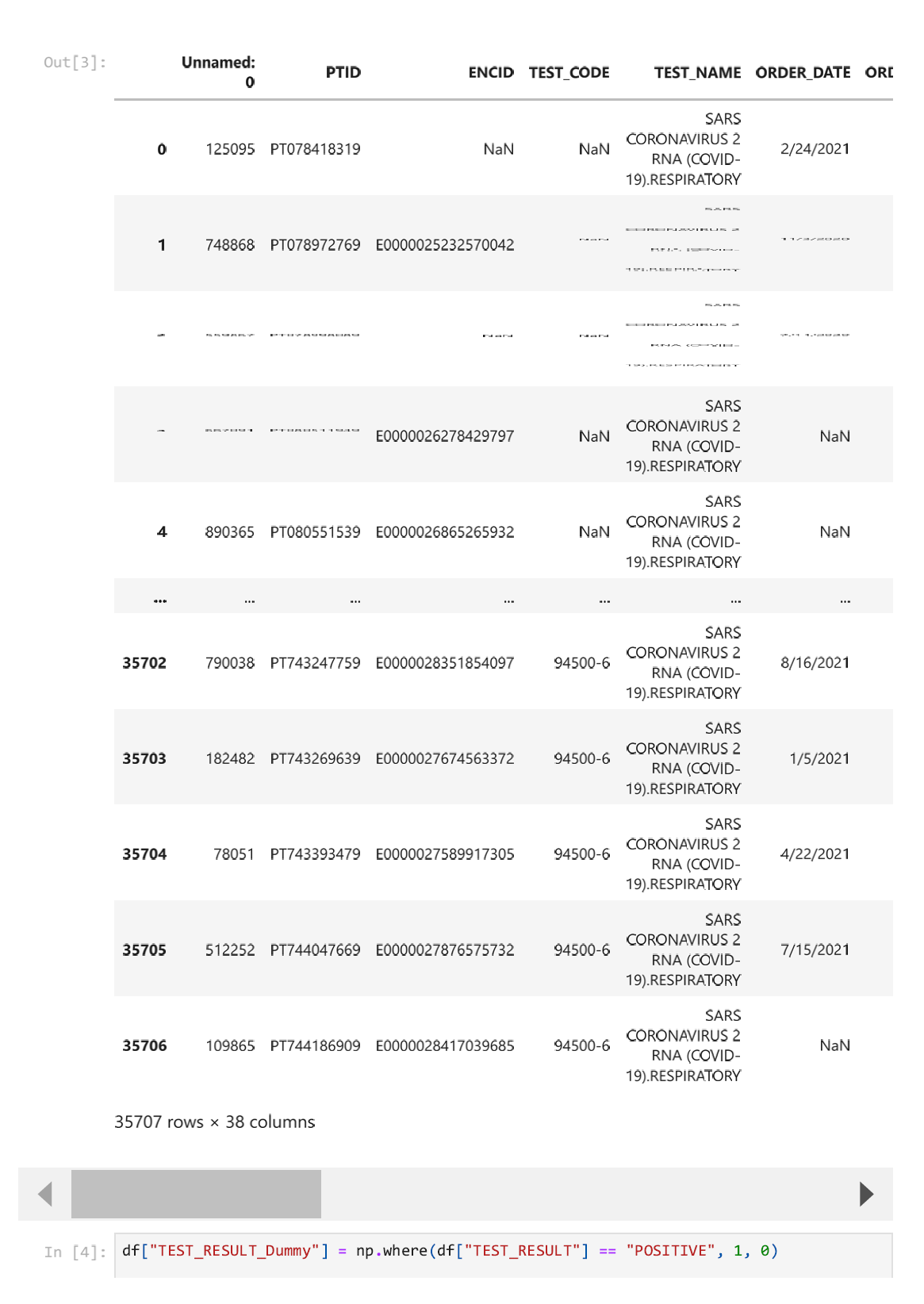


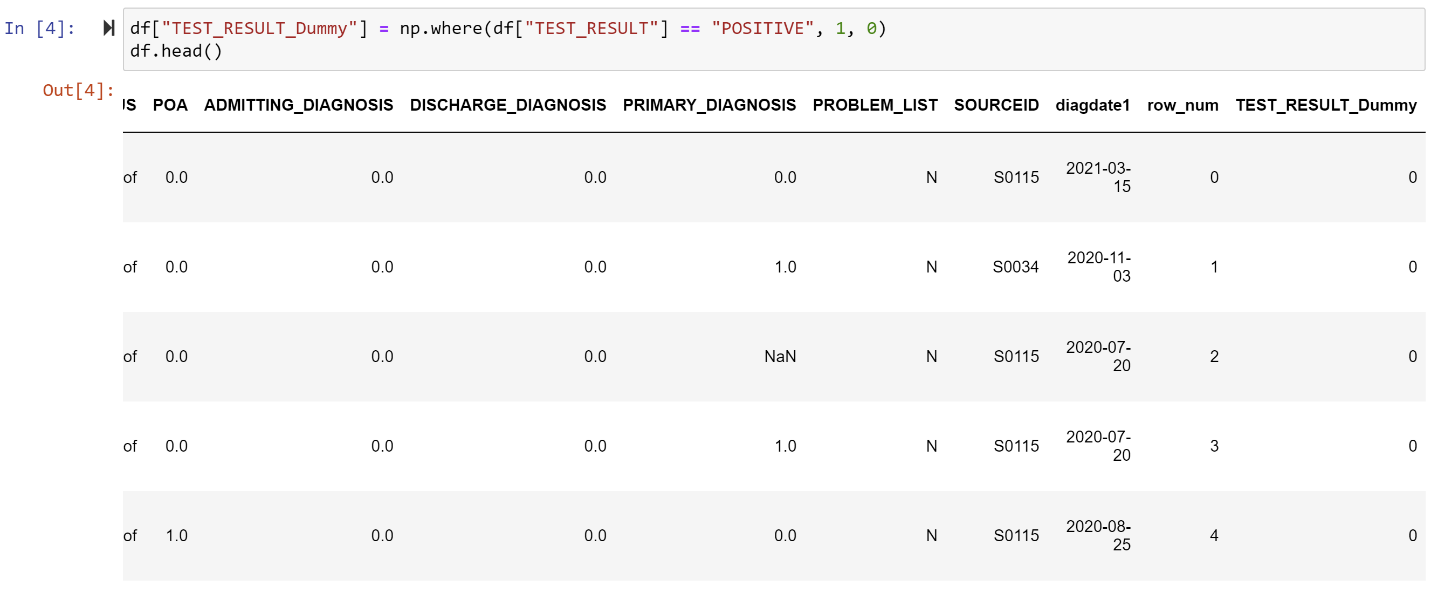
Step 3: Assign PTID as numbers



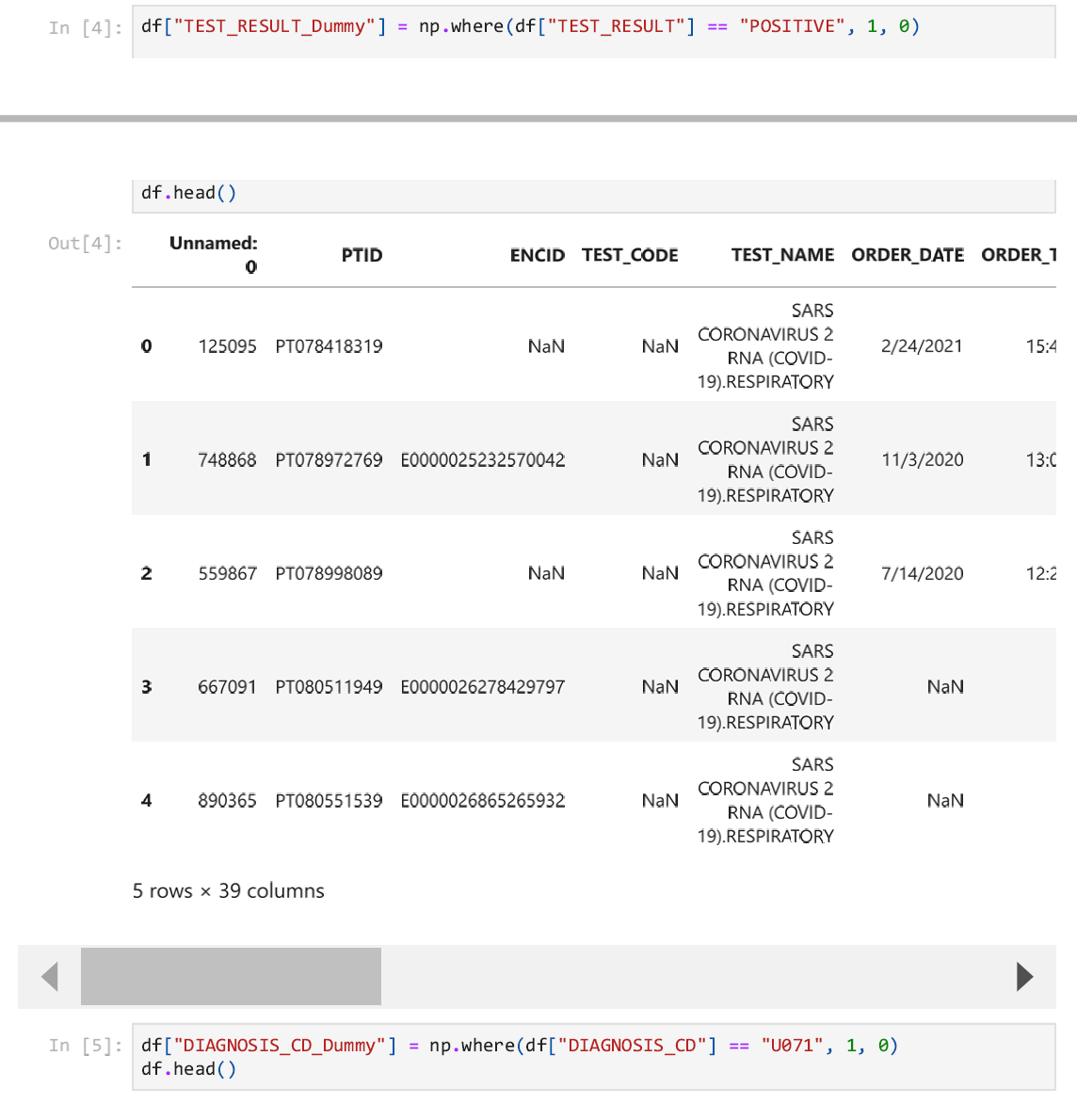


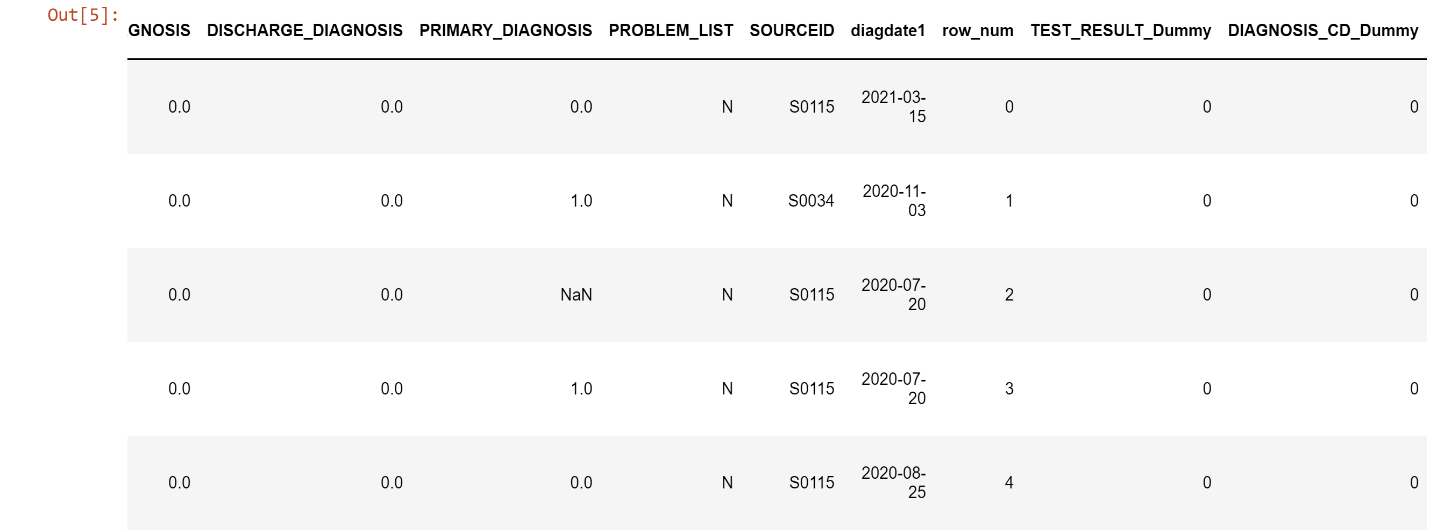
Step 4: Recode “TEST\_RESULT” as 1 for Positive and 0 for negative in a new column called “TEST\_RESULT\_Dummy”





Step 5: Recode “DIAGNOSIS\_CD” as 1 for U071 and 0 for codes that are not U071 in a new column “DIAGNOSIS\_CD\_Dummy”



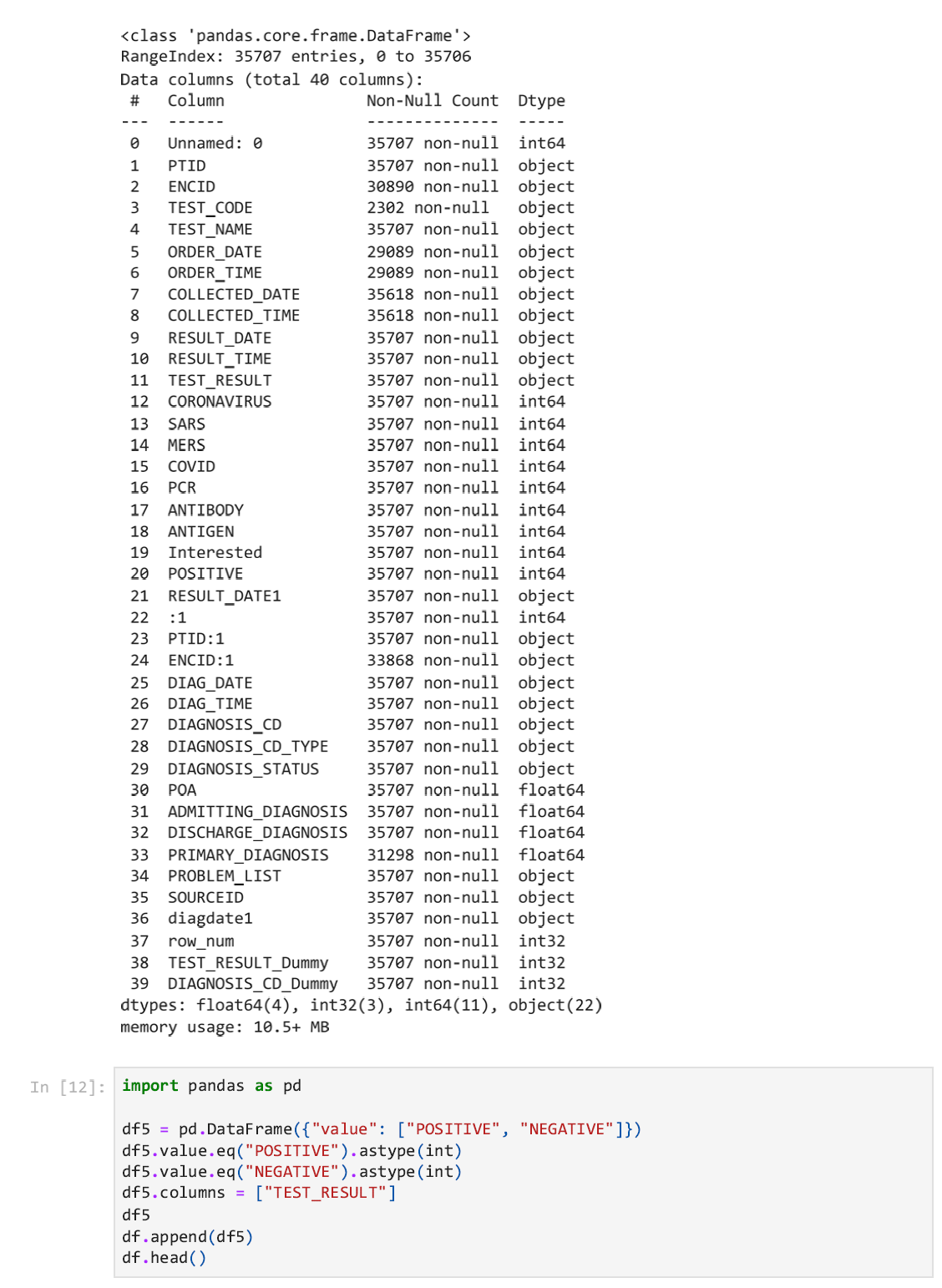


Step 6: Calculate sensitivity, PPV, NPV, and specificity from df by counting the rows

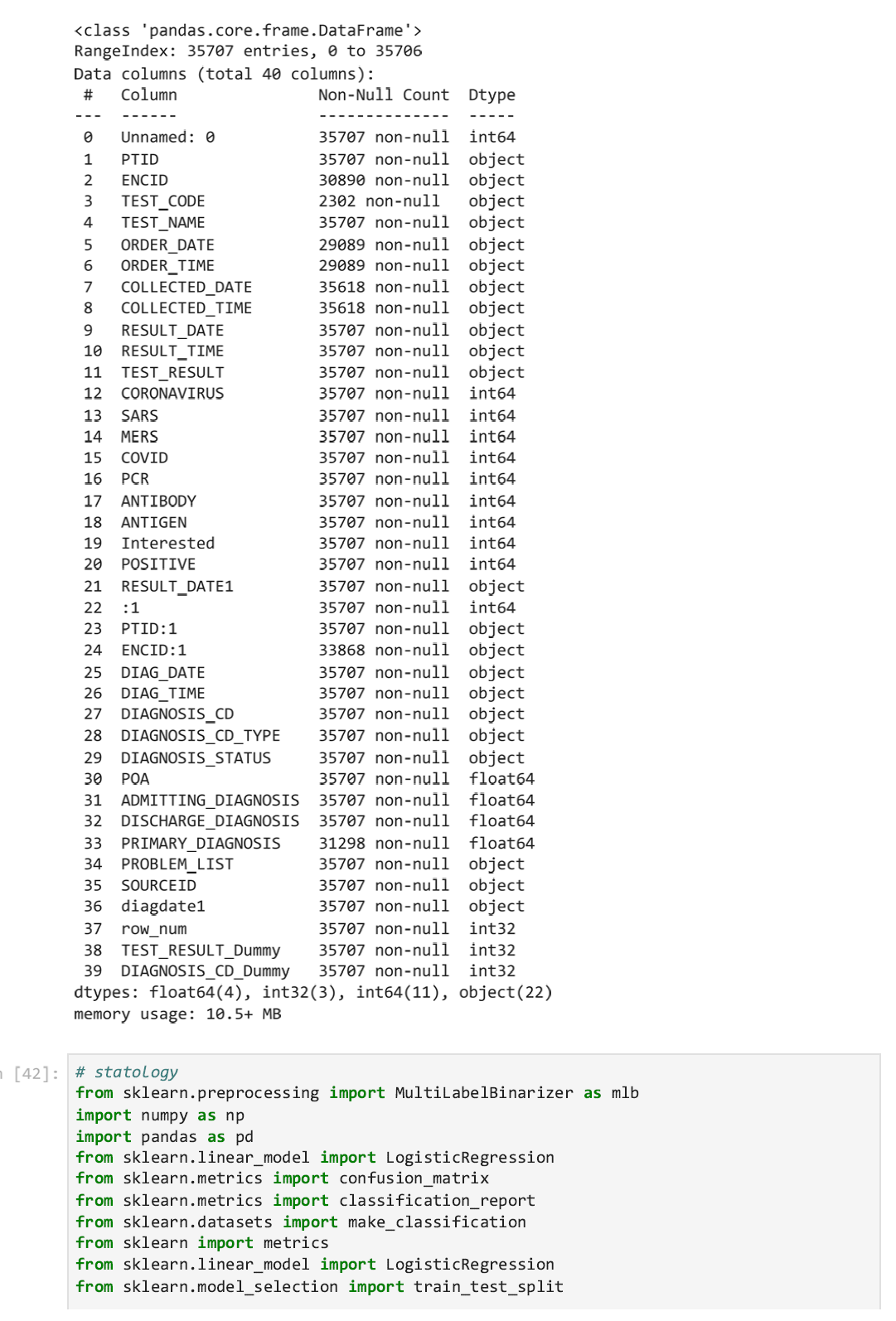




Step 7: Used the code df.info() to see dtype for each variable



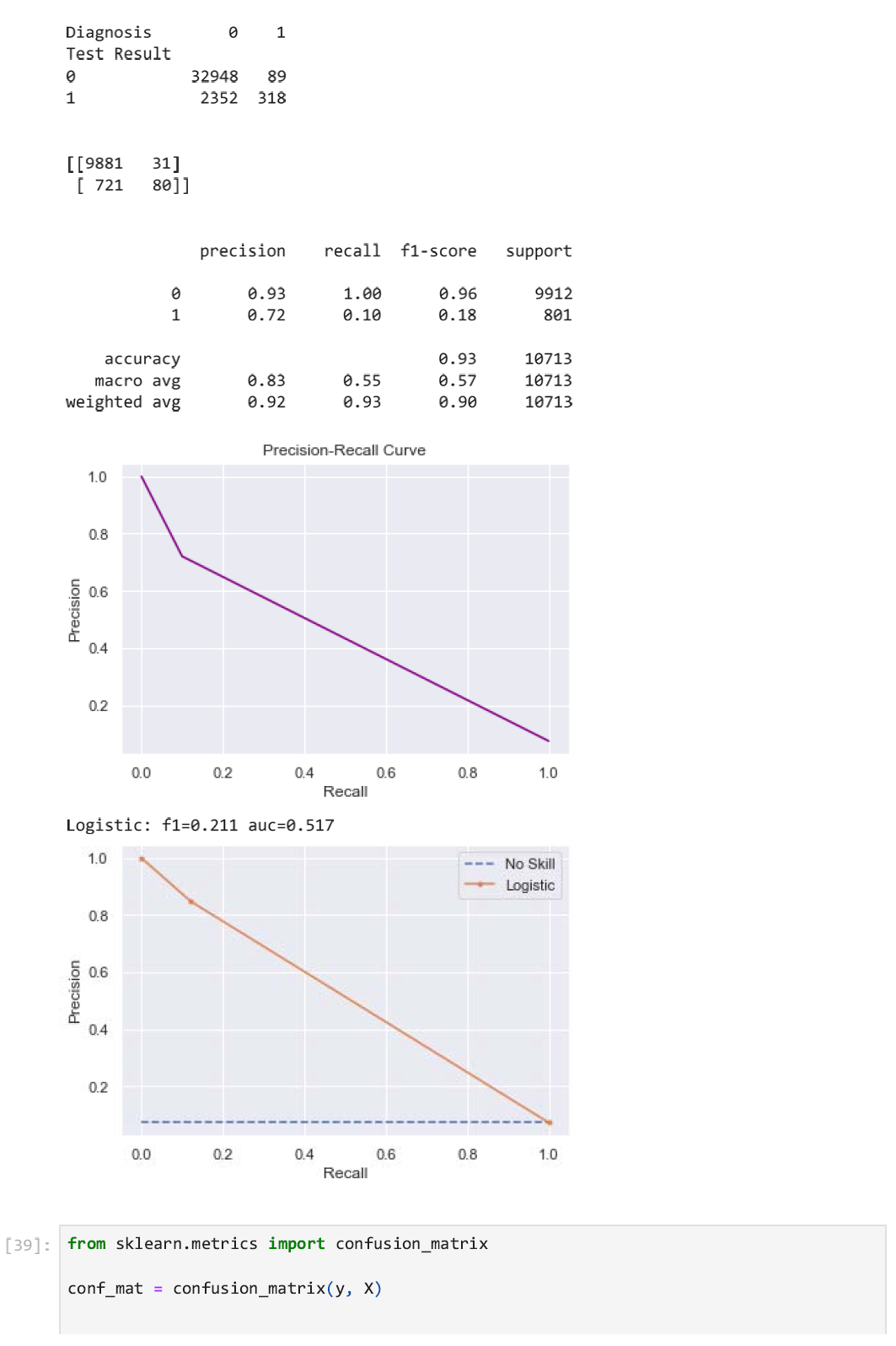
Step 8: Import libraries to create confusion matrix



Step 9: Plot the original confusion matrix without any classifier

Step 11: Plot confusion matrix without classifier

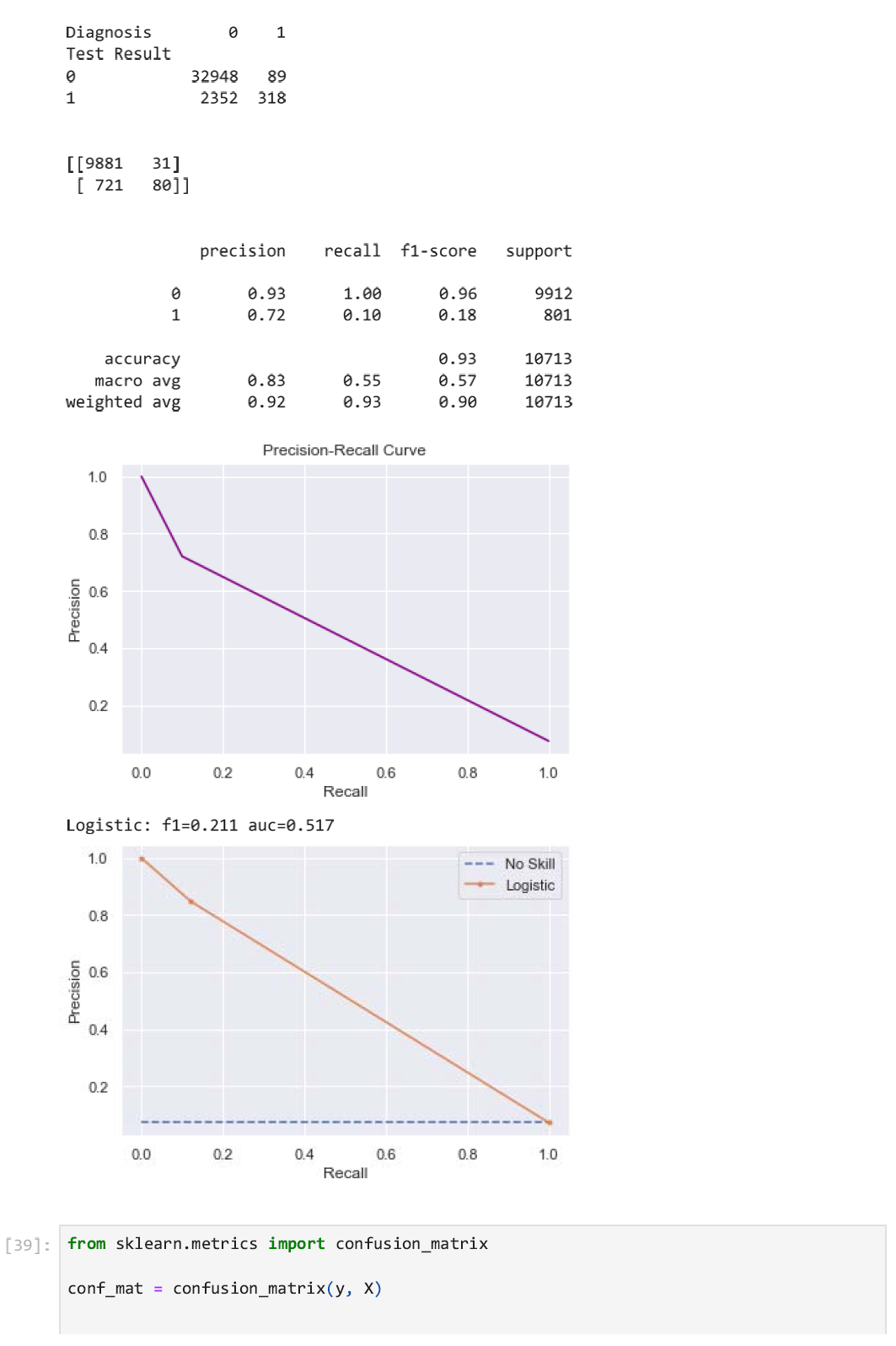






Step 10: Create Logistic Regression Confusion Matrix to fit the model

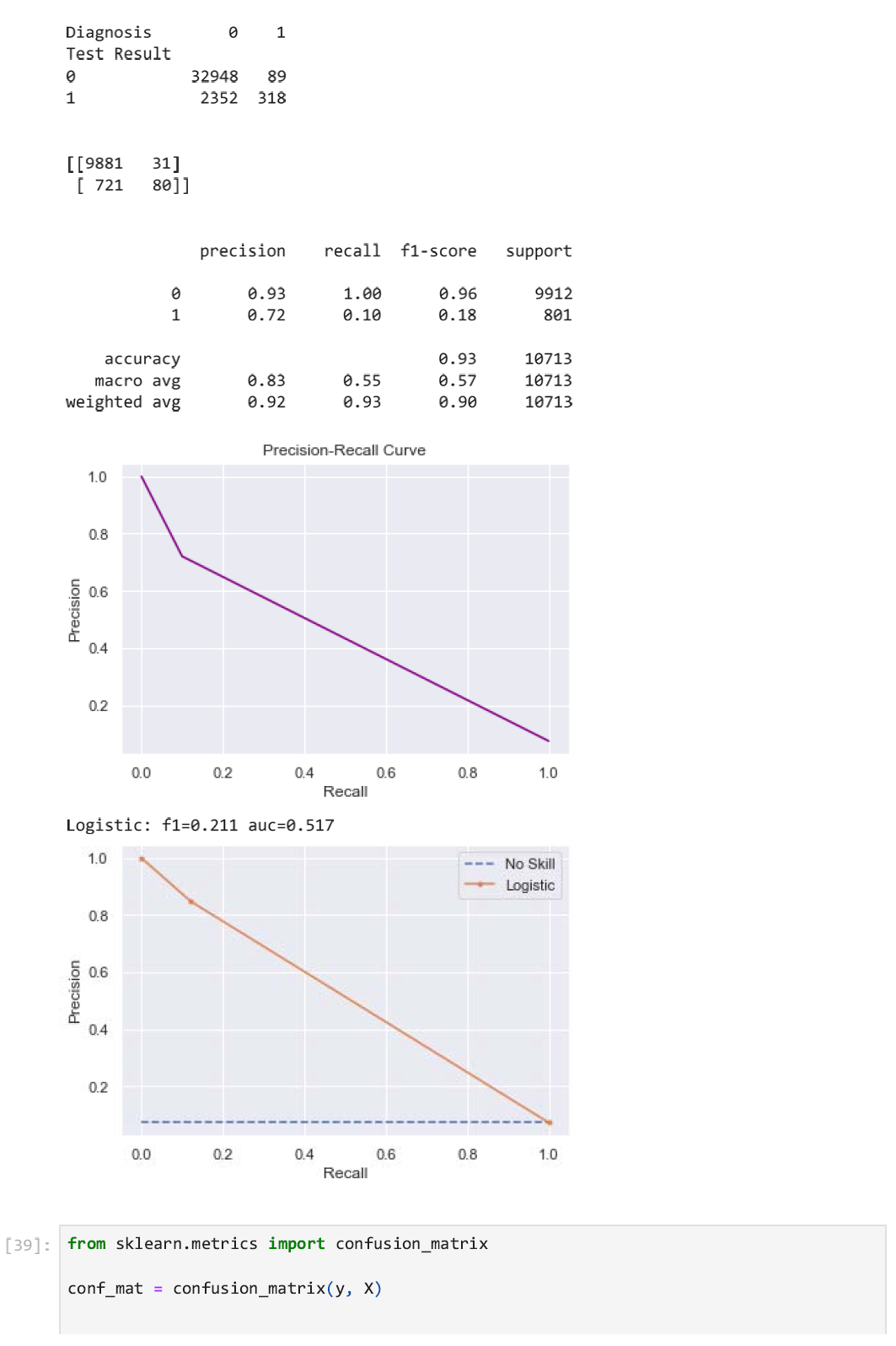




Step 11: Calculate precision, recall, threshold, plot Precision-Recall Curve, and plot Logistic Regression Model with F1 score and AUC







***Proposal***

The purpose of this study is to estimate the sensitivity and positive predictive values of the ICD-10 codes that are used to classify COVID-19 cases compared to a positive PCR test result. The public health significance is the risk of COVID-19. However, to date, no studies have determined the sensitivity and PPV of ICD-10 codes among COVID-19 cases from the Optum database. The findings of the Wu et al. study showed the estimated validity of one ICD-10 code (U07.1) ranges from 49% to 98% in a Canadian suburban hospital. This range is concerning as the validity should be high (above 80%) for the code to be properly used.1 Yet, there are few studies that investigated the sensitivity and PPV from a large, public patient population database from a non-medical setting, such as the Optum database. This present study aims to fill that gap in knowledge.

*Input Data*

The original dataset, COVID Optum database, is from UTHealth. This dataset has provided many important de-identified patient’s information, including diagnosis code, test result, test date, order date, test name, and other medical information. The dataset has around 1 million observations, which is a big dataset for the final project. Therefore, I only use around 0.04% of the whole dataset by randomly sampling. This study includes almost 40,000 observations with 8 columns. The SARS-COVID-19 test (1 for positive and 0 for negative) as the response variable belongs to the binary distribution. In addition, 2 explainable variables are categorical variables.

*Expected Output Data*

​​ The confusion matrix without classifier will be calculated. The response or “y” variables would be the positive and negative SARS-COVID19 test. The predictor variables or “X” variables would be the diagnosis code of U071 and other diagnosis codes. To balance the confusion matrix, the logistic regression confusion matrix and Precision-Recall Curve will be used to observe the relationship between the predictor variables and response variables. The overall model fit is evaluated with logistic regression model with F1 score and AUC score.

References:

1. Wu G, D'Souza A, Quan H, Southern D, Youngson E, Williamson T, Eastwood C, and Xu Y. Validity of ICD-10 codes for COVID-19 patients with hospital admissions or ED visits in Canada: a retrospective cohort study. BMJ Open. 2022;12(1):1-7.