**Data Mining – Advanced Statistical Modelling Report**

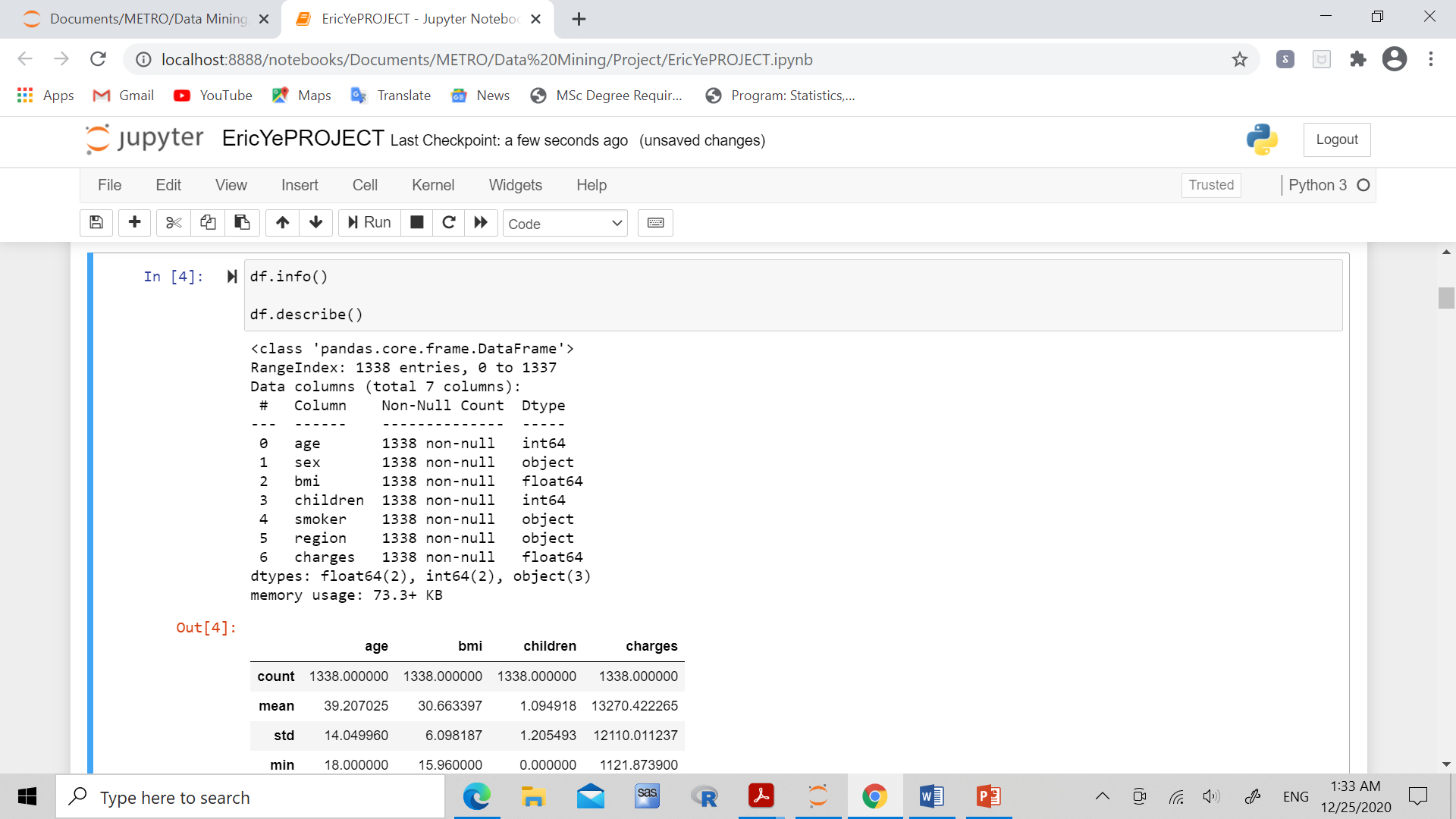
**Eric Ye**

**Gitimoni Saikia**

**December 25, 2020**

**Linear Regression**

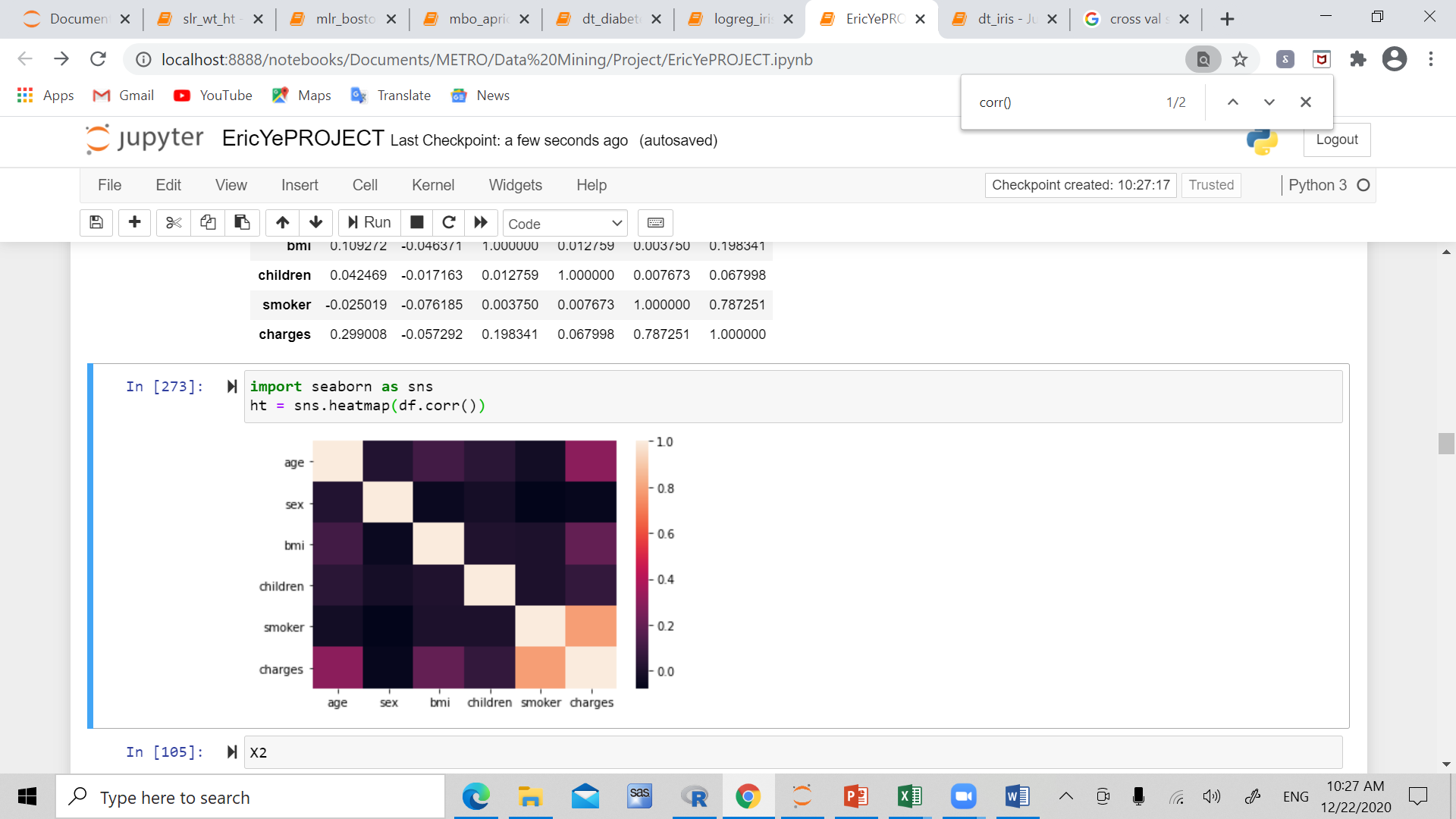
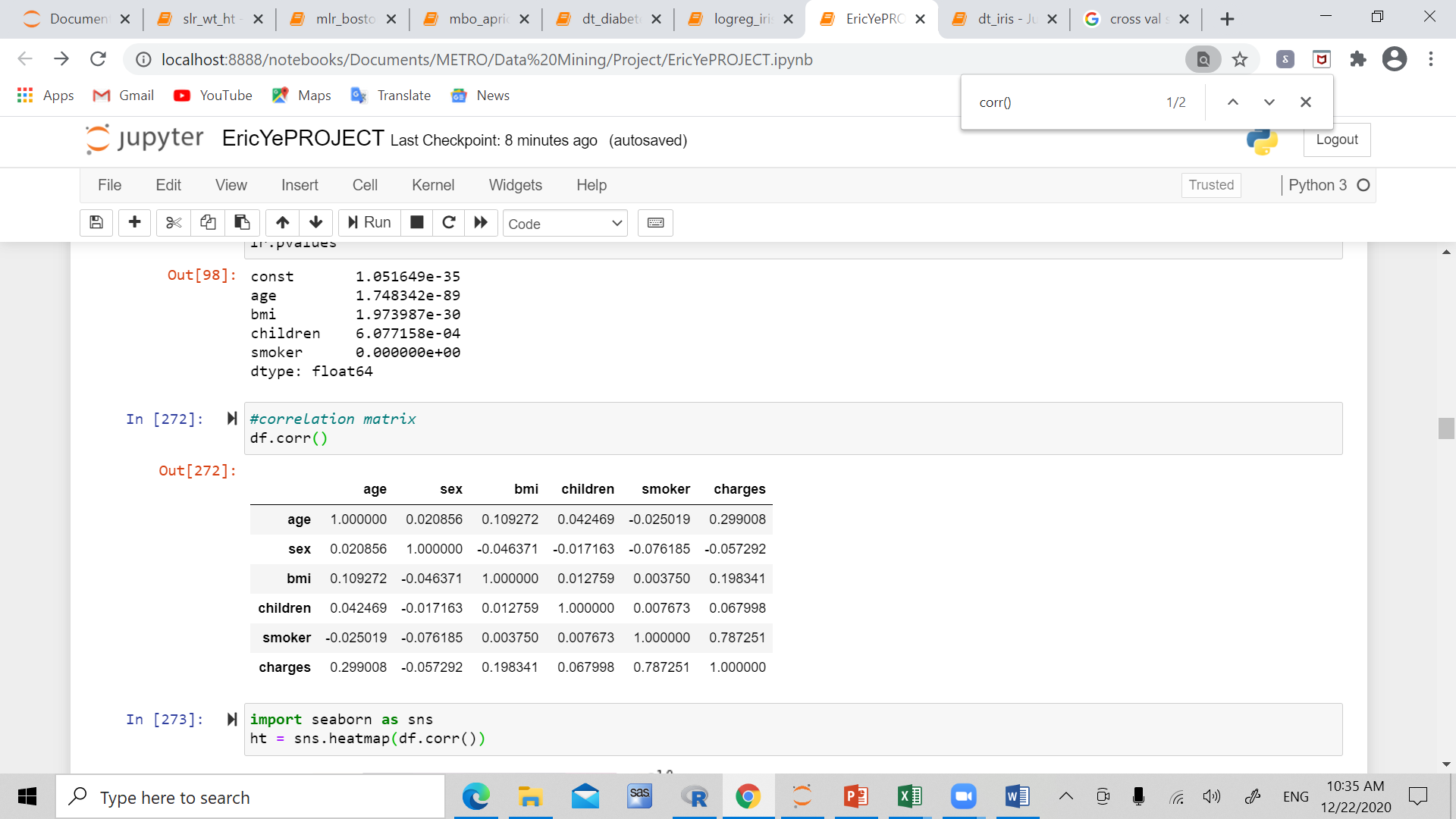
*Properties of dataset*: The dataset sourced from Kaggle, is based on medical insurance data for families. It contains 1338 observations and 7 variables. The variables include age, sex, BMI, children, smoker, region, and insurance charges.



1. **Data cleaning:**

The dataset was first imported, checked for missing values, and encoded for categorical variables

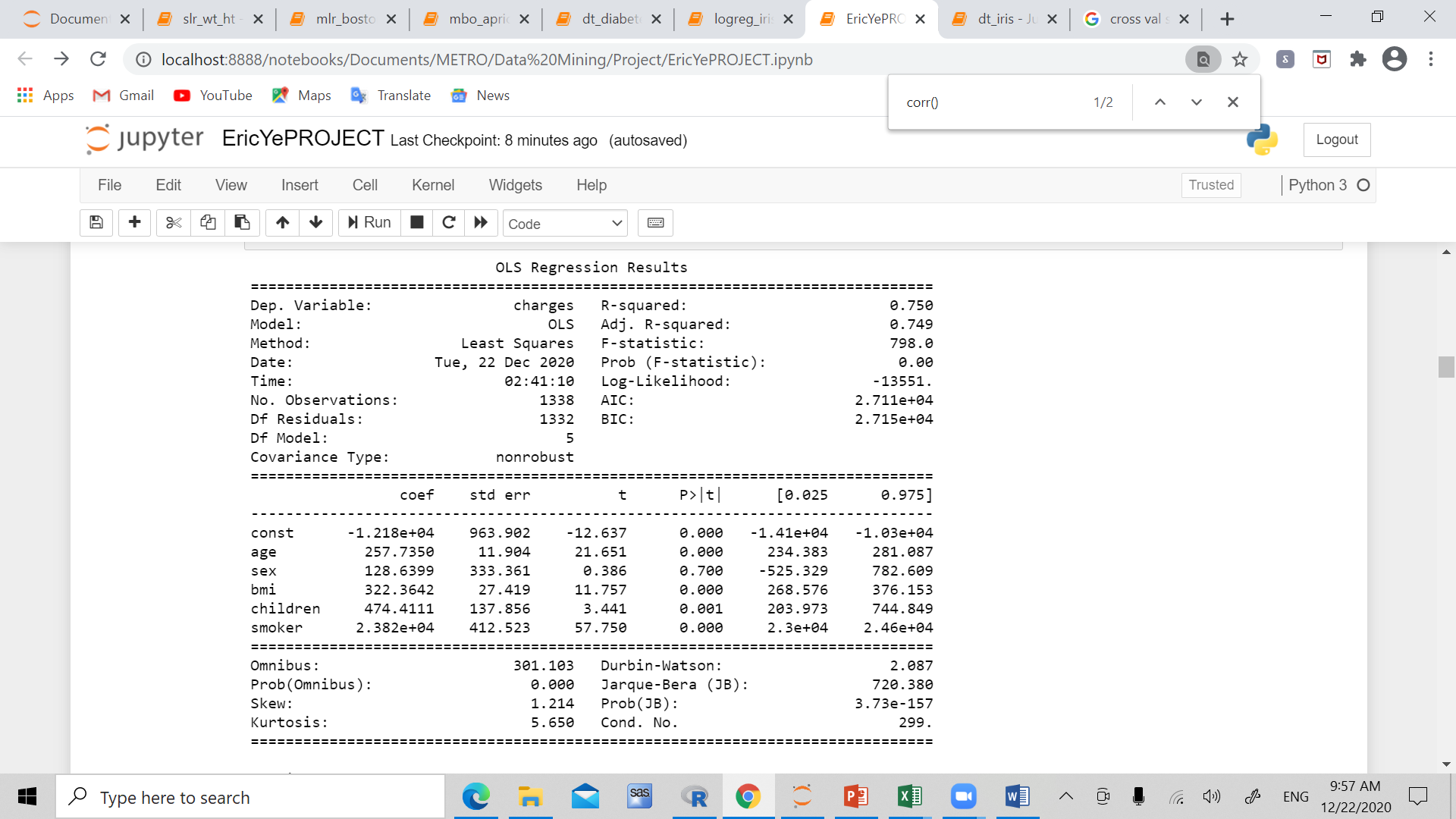
1. **Correlation matrix/ Heat map:**

****

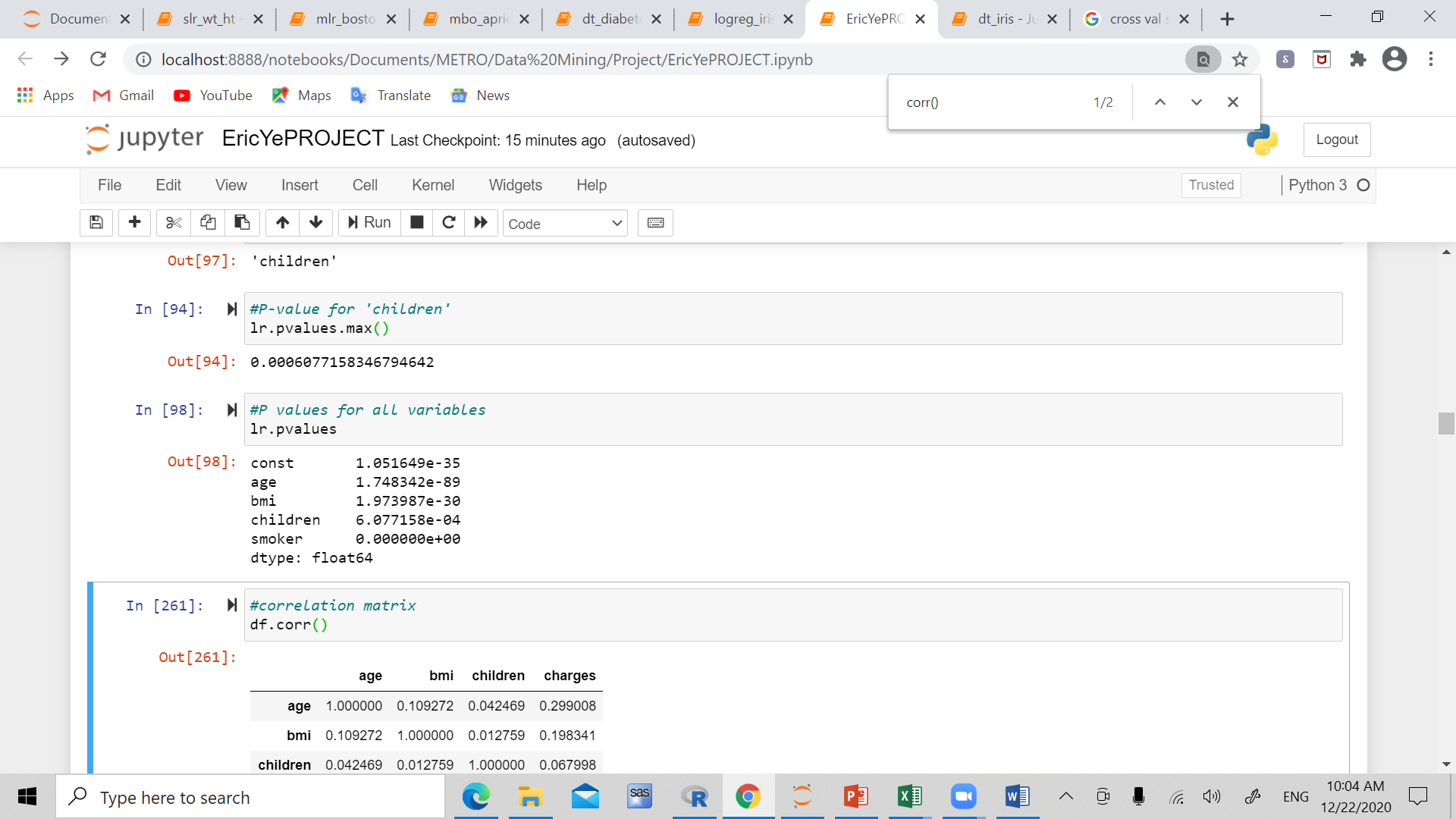
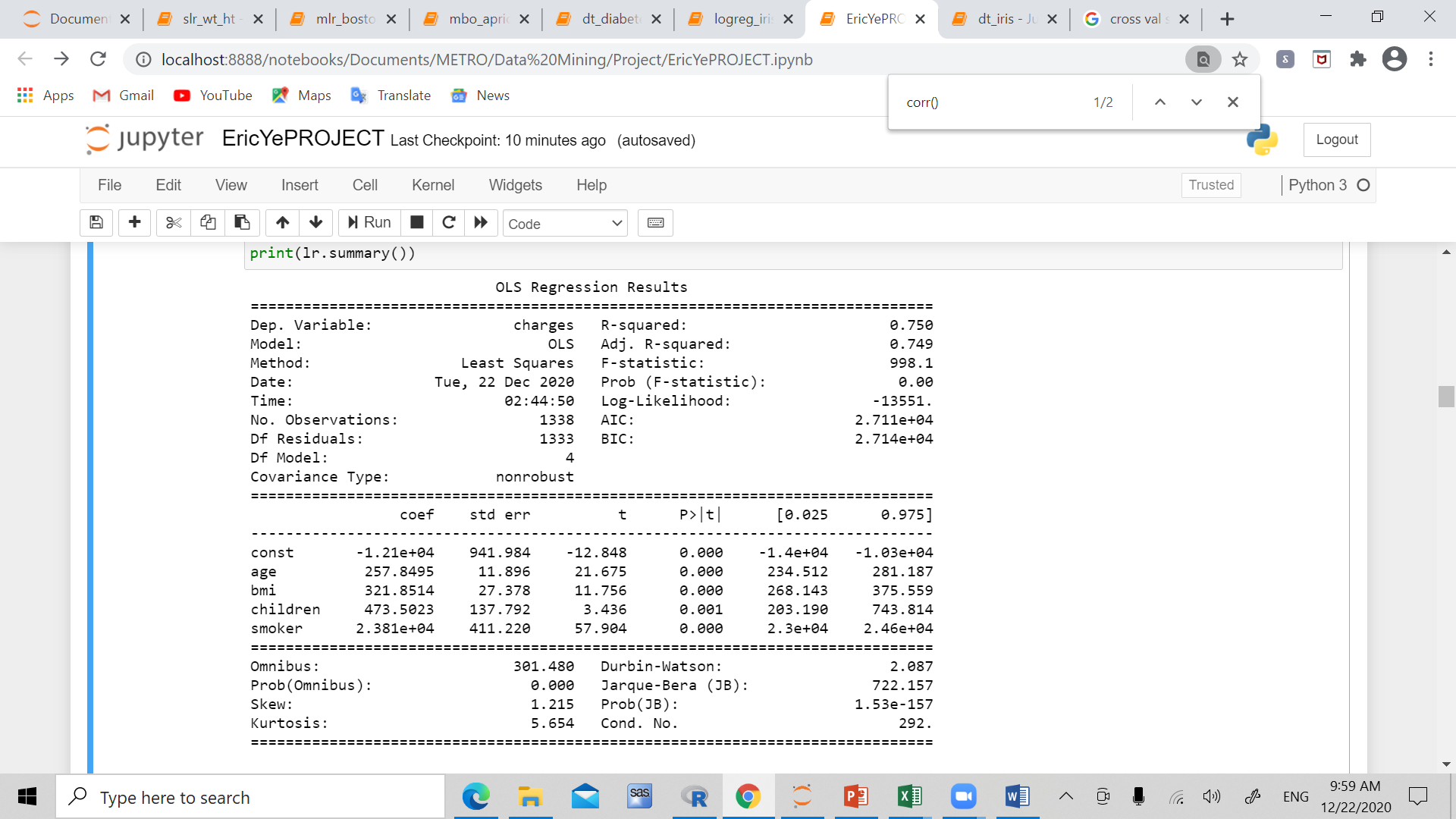
This suggests that smoker, BMI, and age have the most correlation with charges.

1. **Linear regression model/ p – values:**

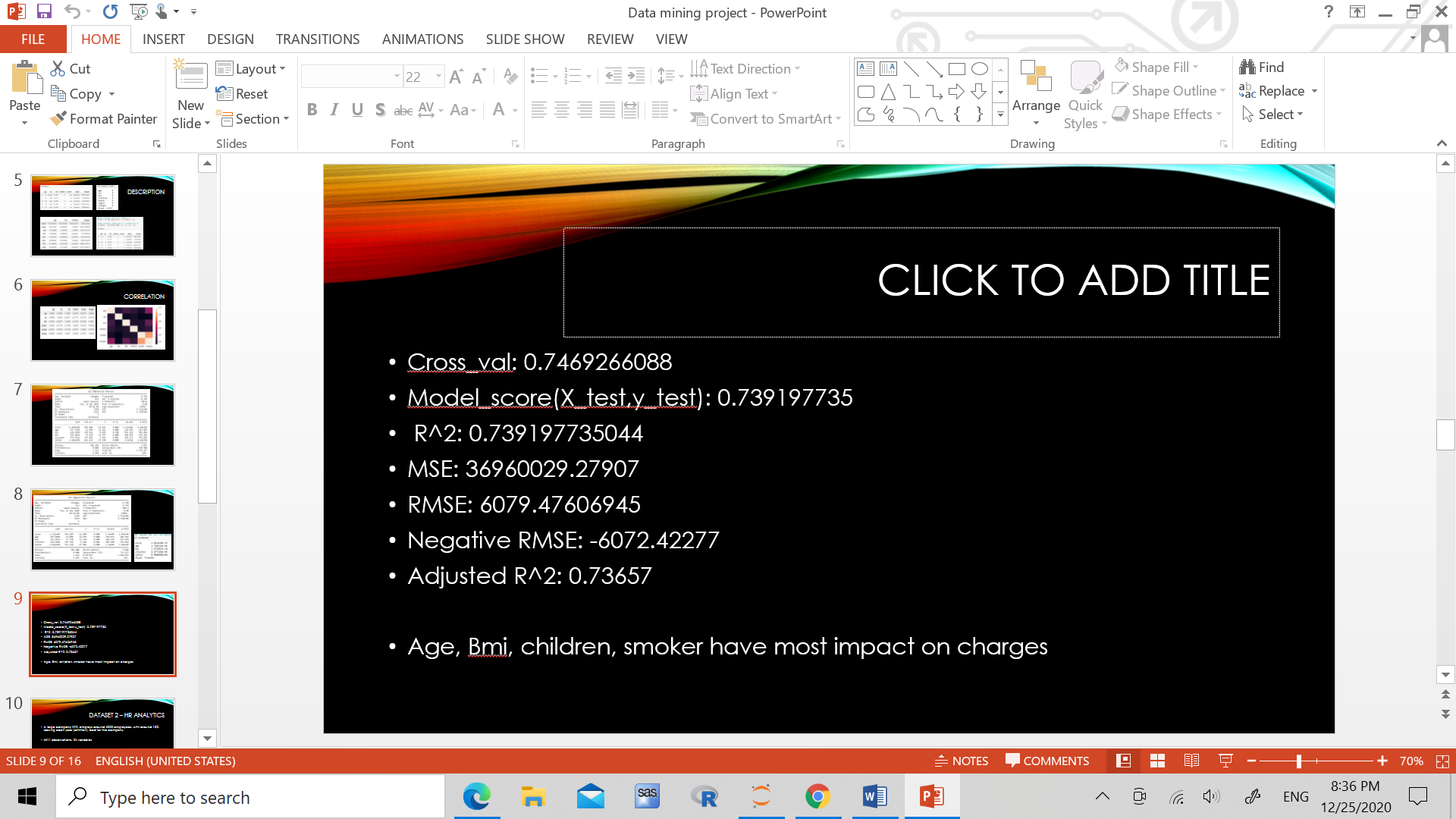
Created first model with OLS method, containing all variables:



Then backwards elimination was used to eliminate and select those significant variables to charges, and print the p-values for each of those significant variables:



1. **R-sq, adjusted R-sq, RMSE:**



**Conclusion:**

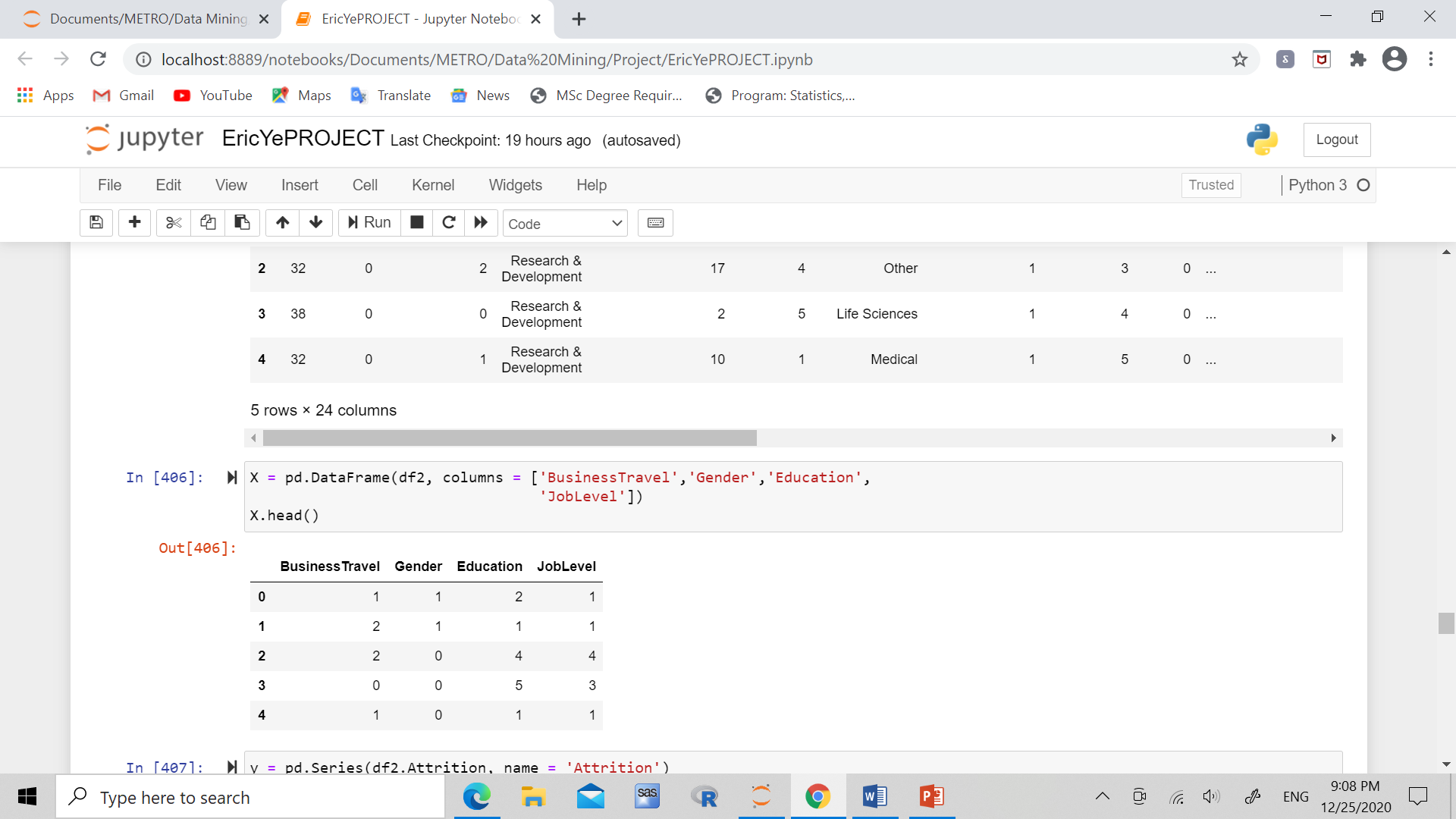
The statistics and results suggest that indeed, smoking, bmi, and age, and number of children are all significant factors in the decision of insurance charges.

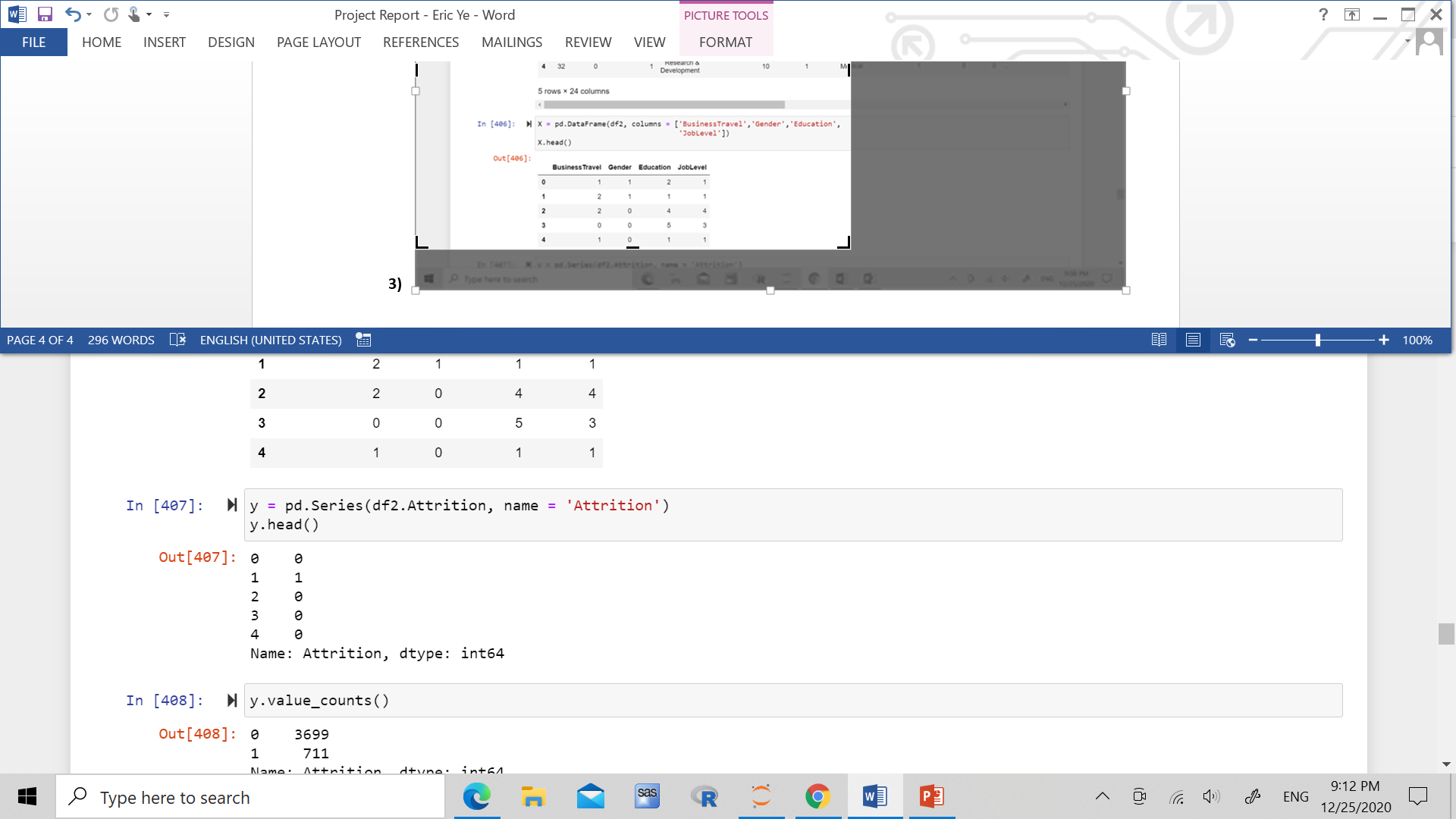
**Logistic Regression:**

*Properties of dataset*: The dataset sourced from Kaggle, is based on HR analytics data for company XYZ. It contains 4411 observations and 24 variables. Among the many variables, the ones that we do analysiss on include age, distance from home, job level, monthly income, education, number of companies worked, years at company, years since last promotion, percent salary hike, and attrition(whether or not the employee has left in the past year). In this analysis, we find the variables that are associated with attrition of the employee.

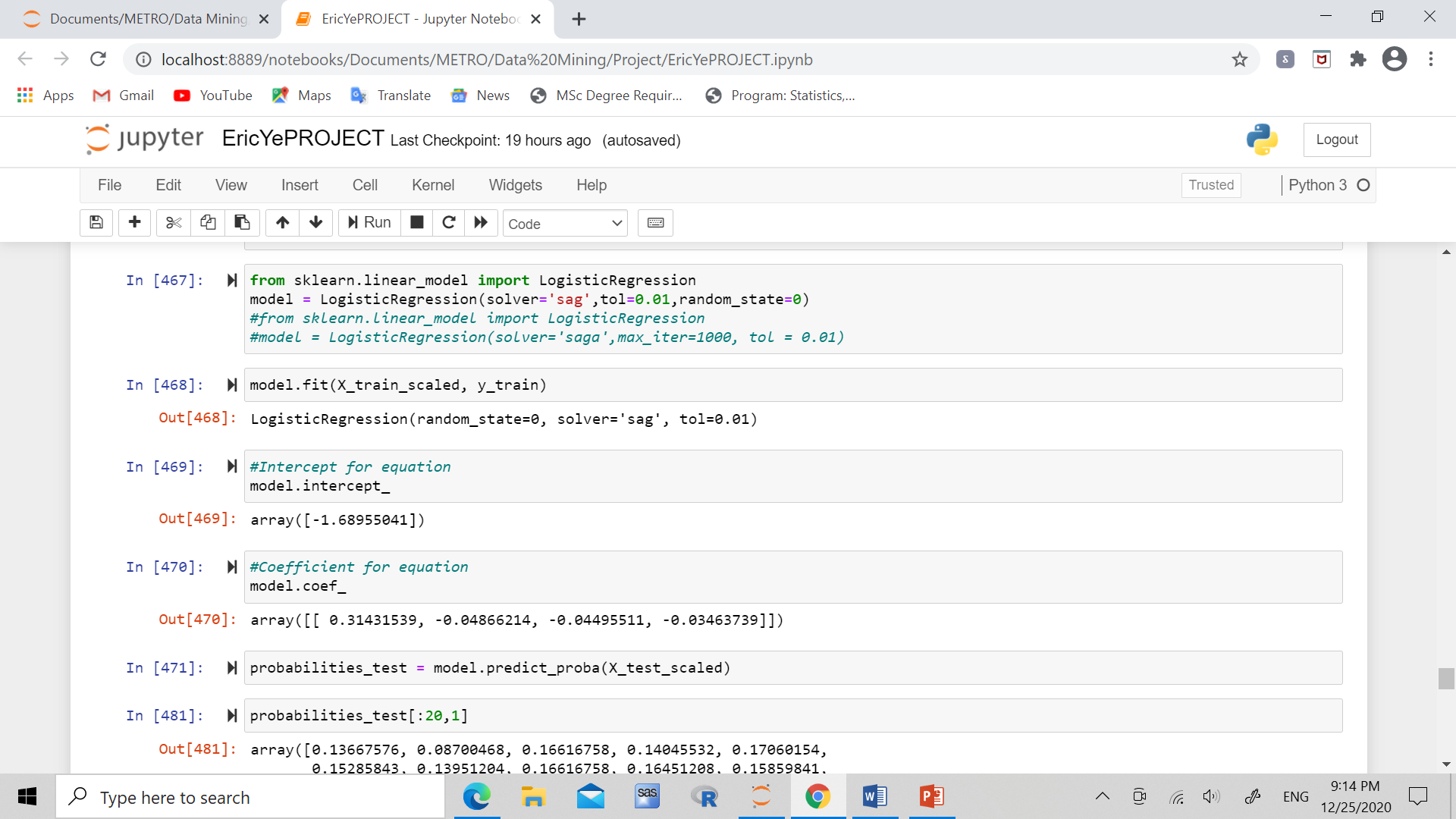
1. **Data cleaning/ Missing value detection/ encoding**

First the data was cleaned, missing values were replaced with 0, and categorical variables were encoded by dummy variables and labelling.

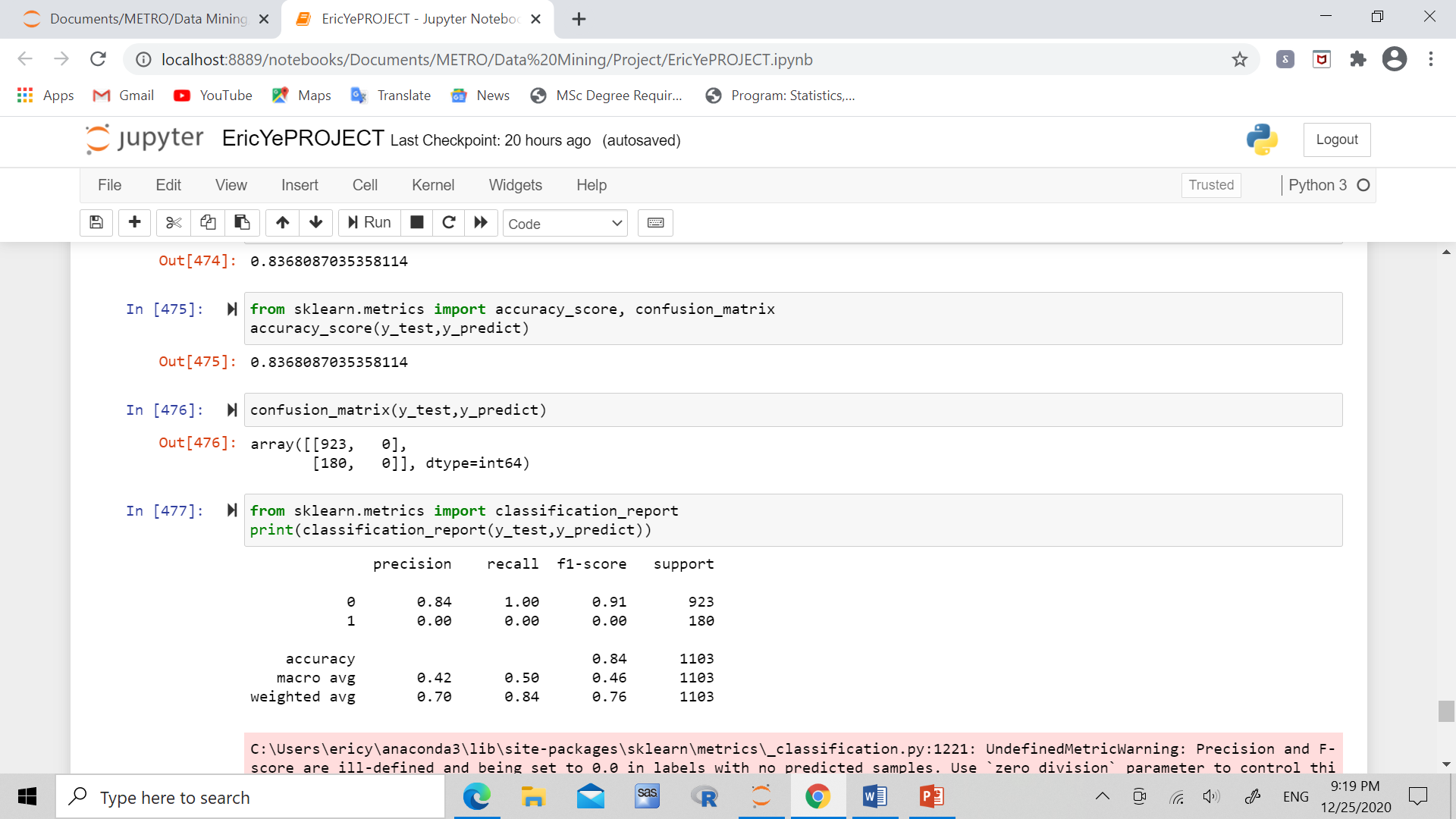
1. **Select variables, split into X and Y, Y being attrition:** 



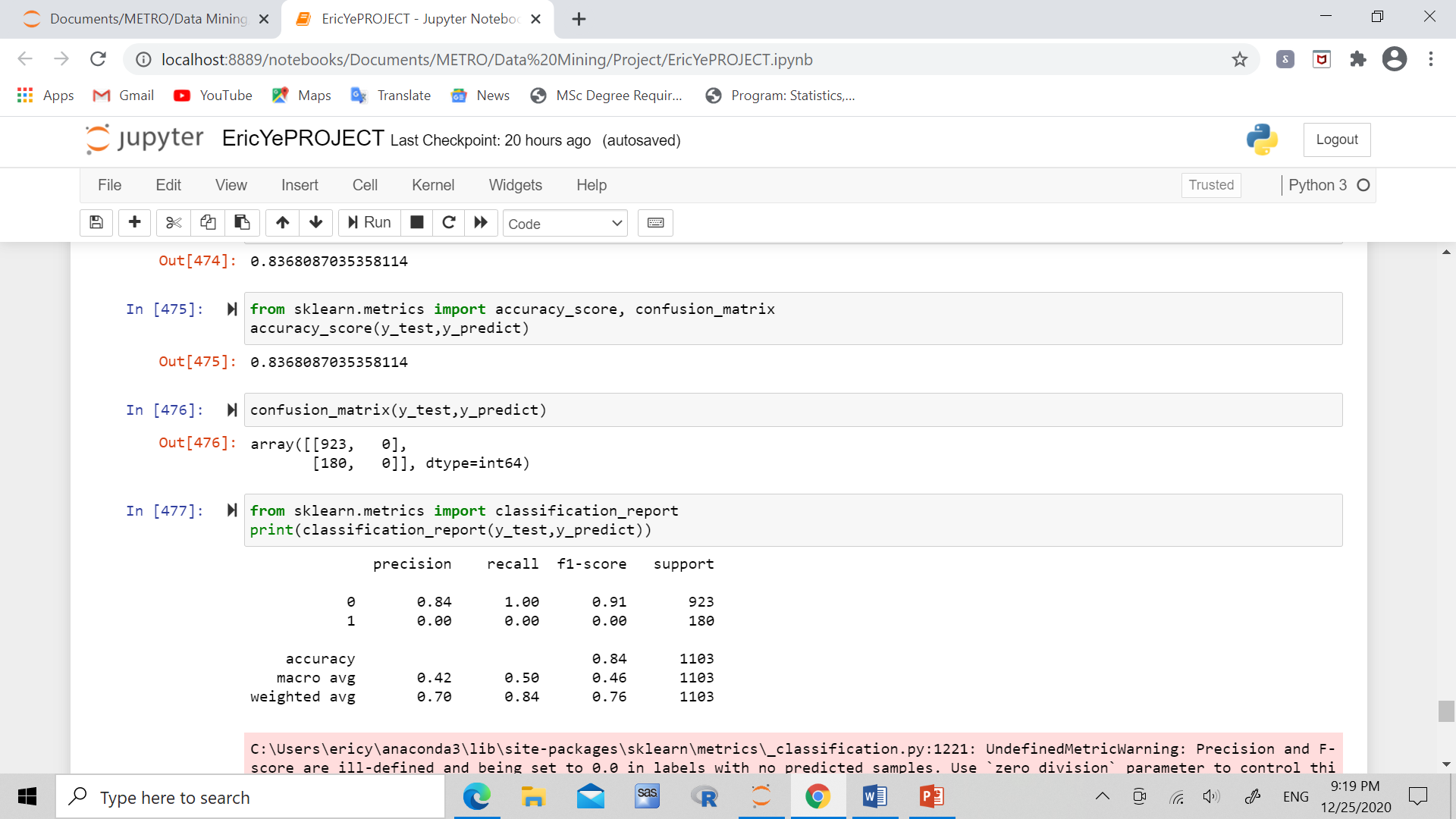
1. **Logistic Equation**



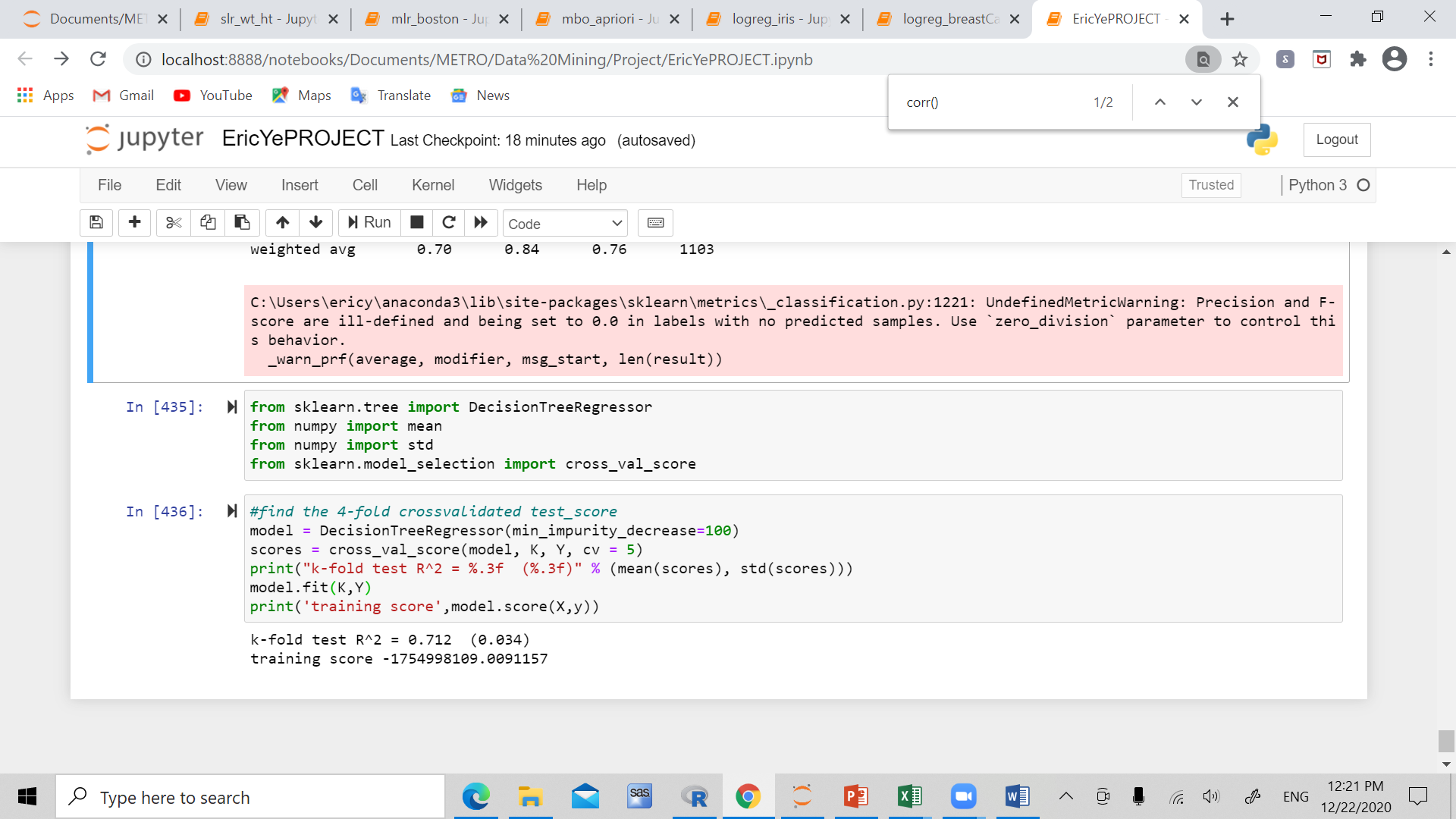
1. **Confusion Matrix**



1. **Accuracy, Recall, Precision**



1. **K – fold test R^2, training score:**

****

**Conclusion:**

Due to the poor precision, recall, and f1-score in the 1 category, we can conclude that the confusion matrix may not be the best type of classification in this case. Ways that we can seek to improve the logistic model is to use a better classification model, such as a decision tree, AUC with ROC score, or by using more observations in the dataset.