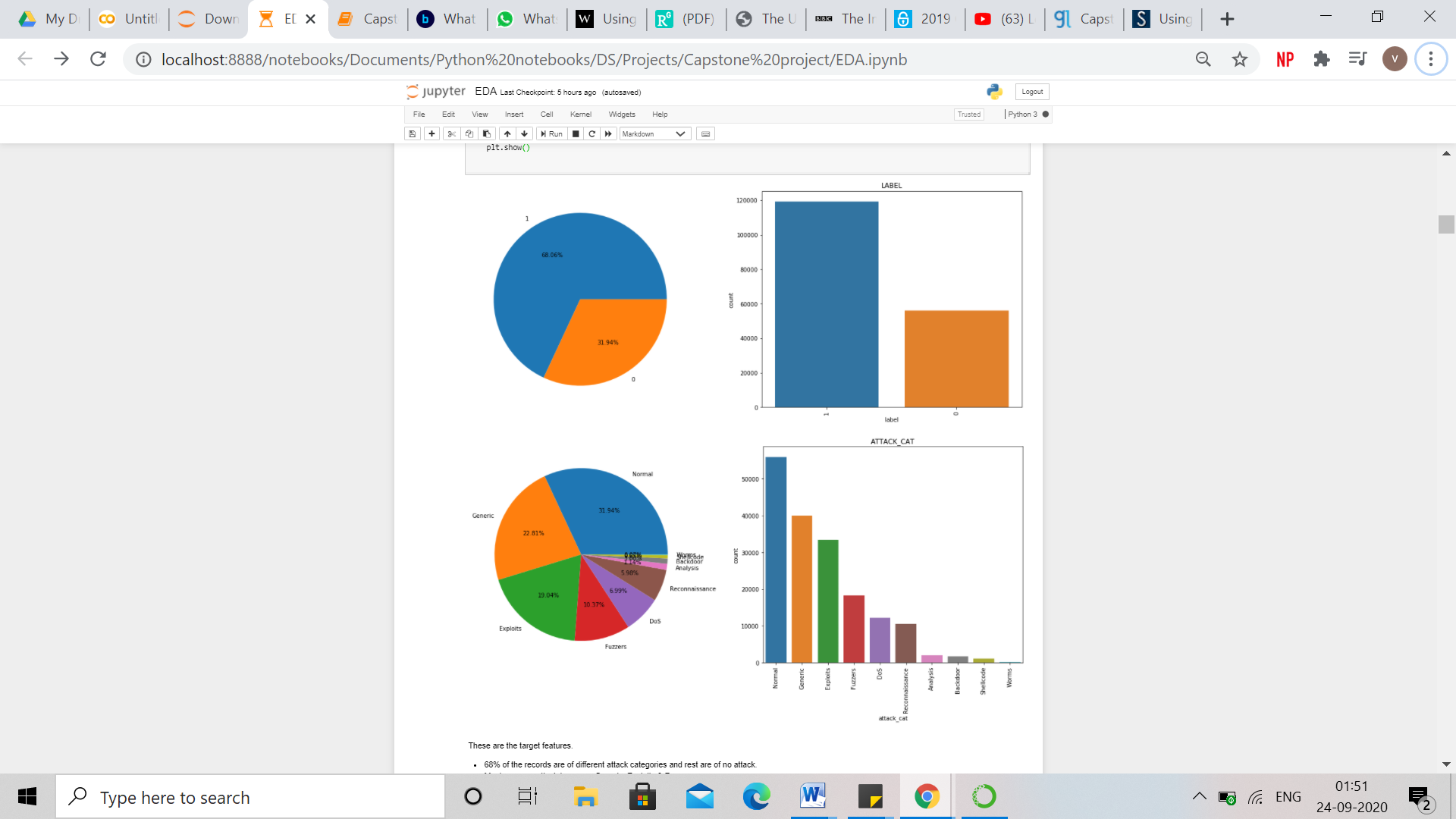
* **Distribution of Variables**

**Target Features**

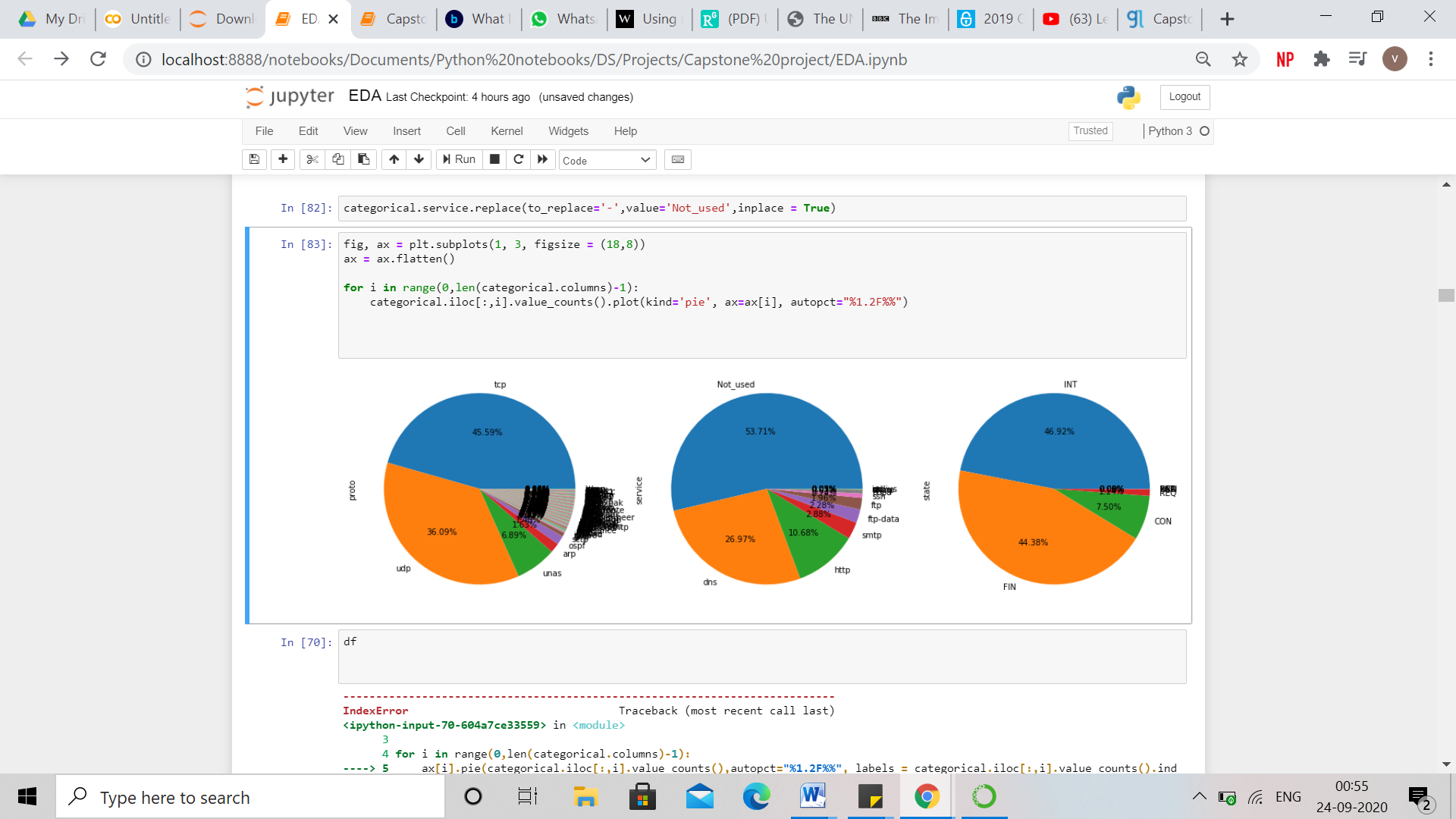
Thereare 2 target features, first is label - which tells whether an attack has taken place (1) or not (0). Second is the attack category feature – it lists 9 different types of attacks and the 10th is normal or no attack.

68% of the labels are 1 which means signifies that an attack has taken place. Generic and Exploits are the common attacks whereas Worms and Shellcode are the rare attacks.



There are 39 numerical features and 3 categorical features.

**Categorical Features**



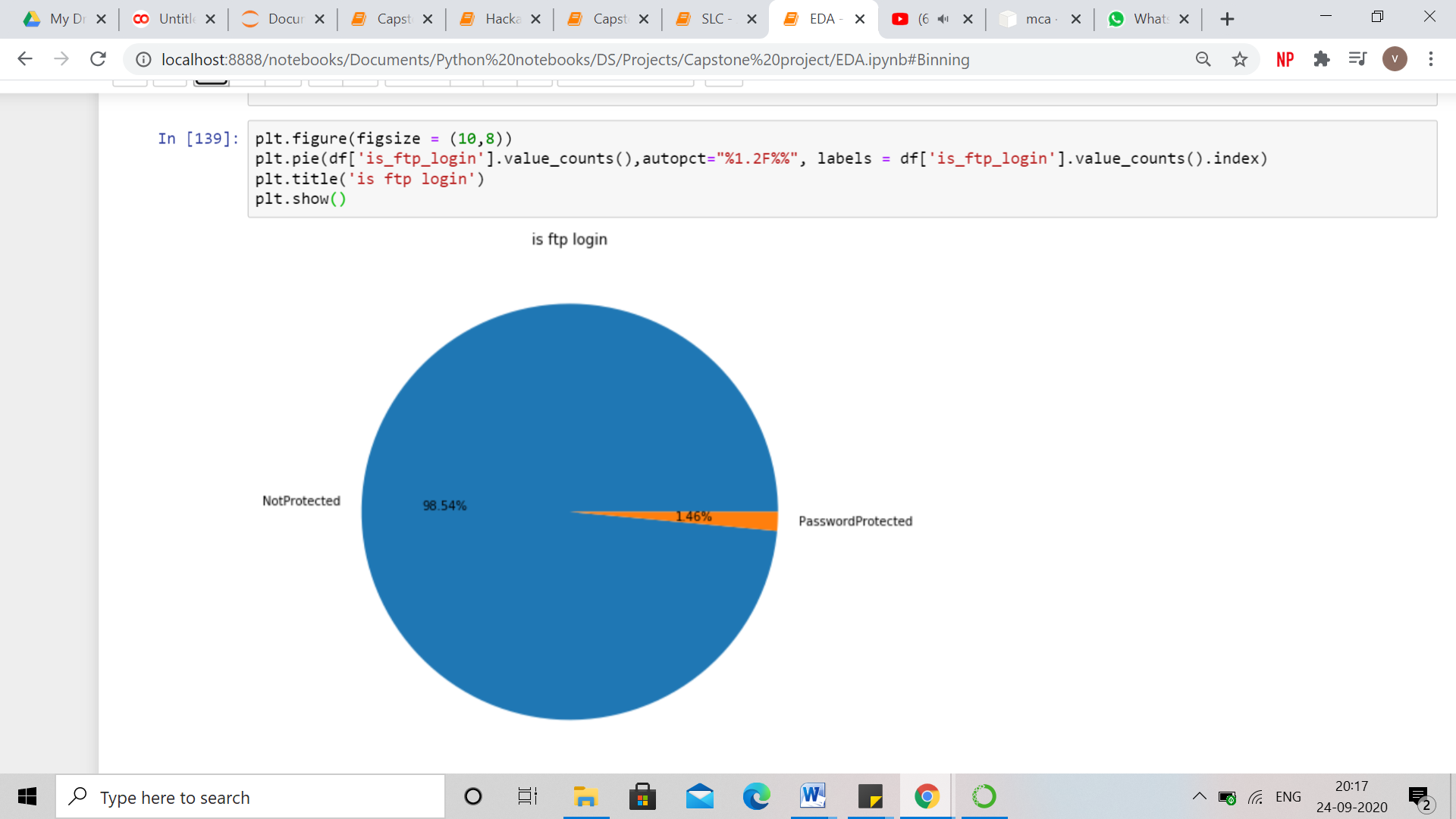
There are 3 categorical columns – proto, service and state.

There are 3 major categories in proto – tcp, udp and unas which represents 88% of the proto data.

In service category, service not used, dns and http makes almost 91% of the service feature.

In state, int, fin and con makes almost 98 % of the state feature.

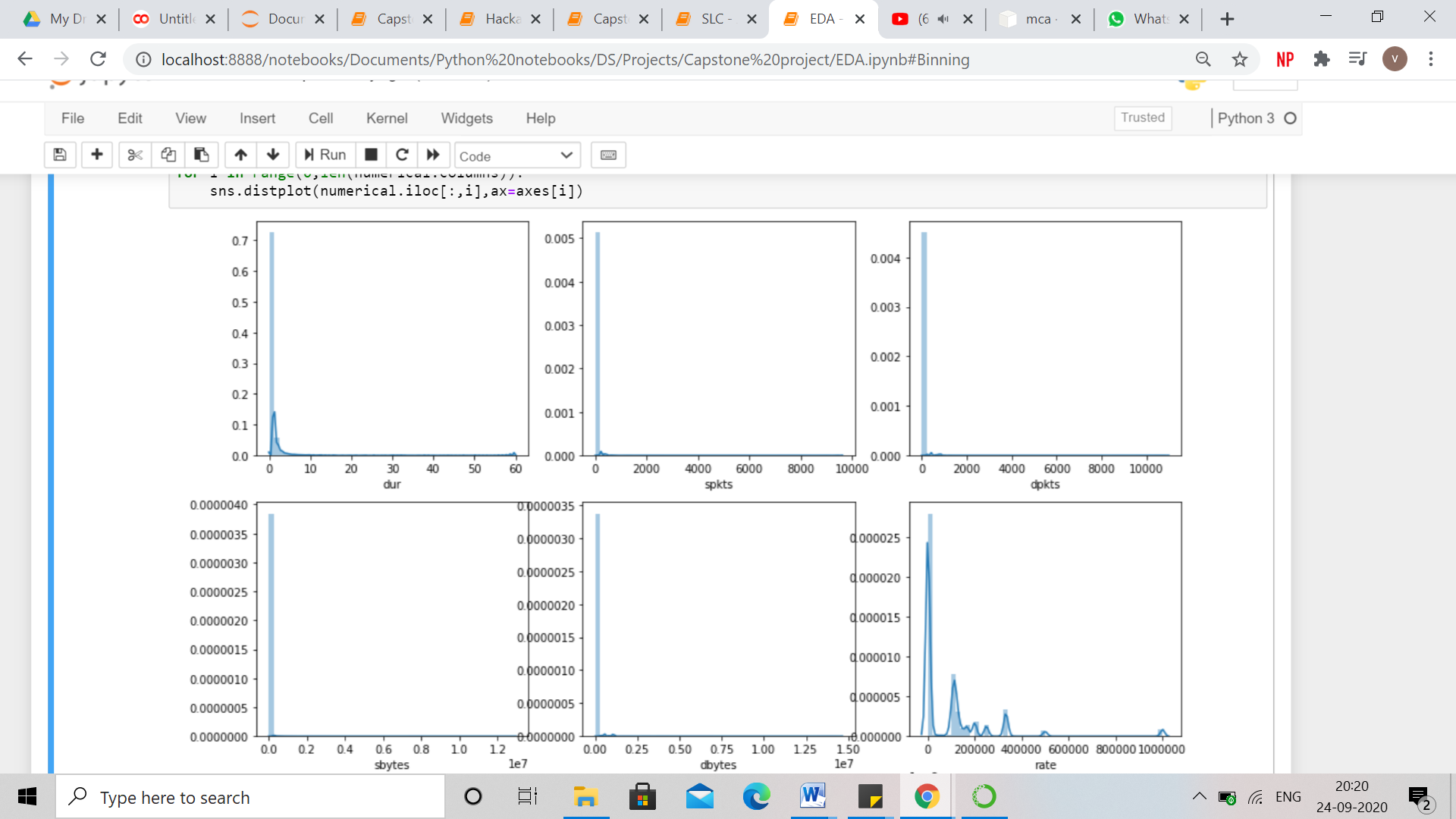
After, EDA it was observed that is ftp login had only 2 values and therefore, it was converted into a categorical column. 0 was replaced as ‘NotProtected’ and 1 was replaced as ‘PasswordProtected’. Most of the data isnot protected by password.

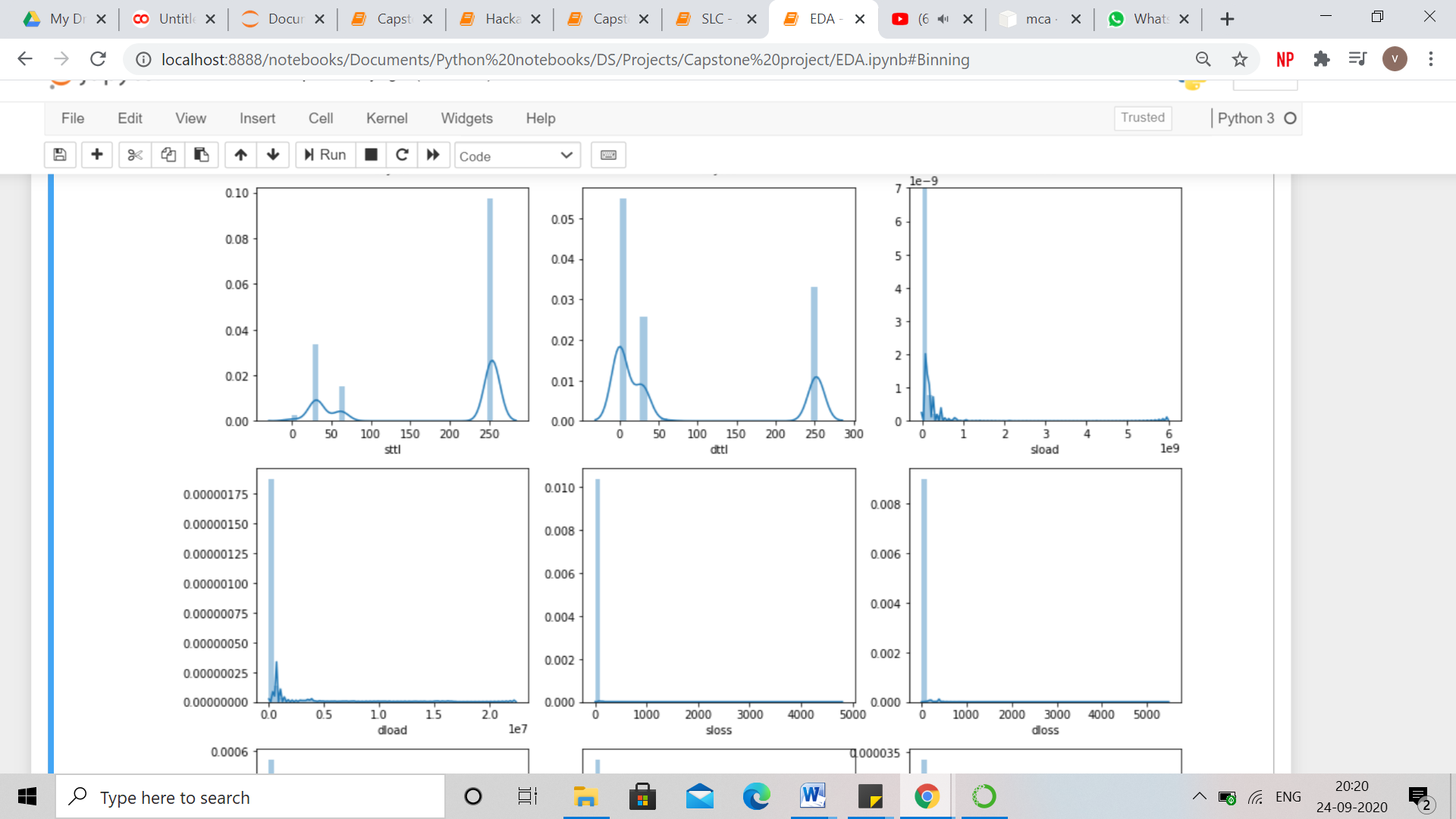


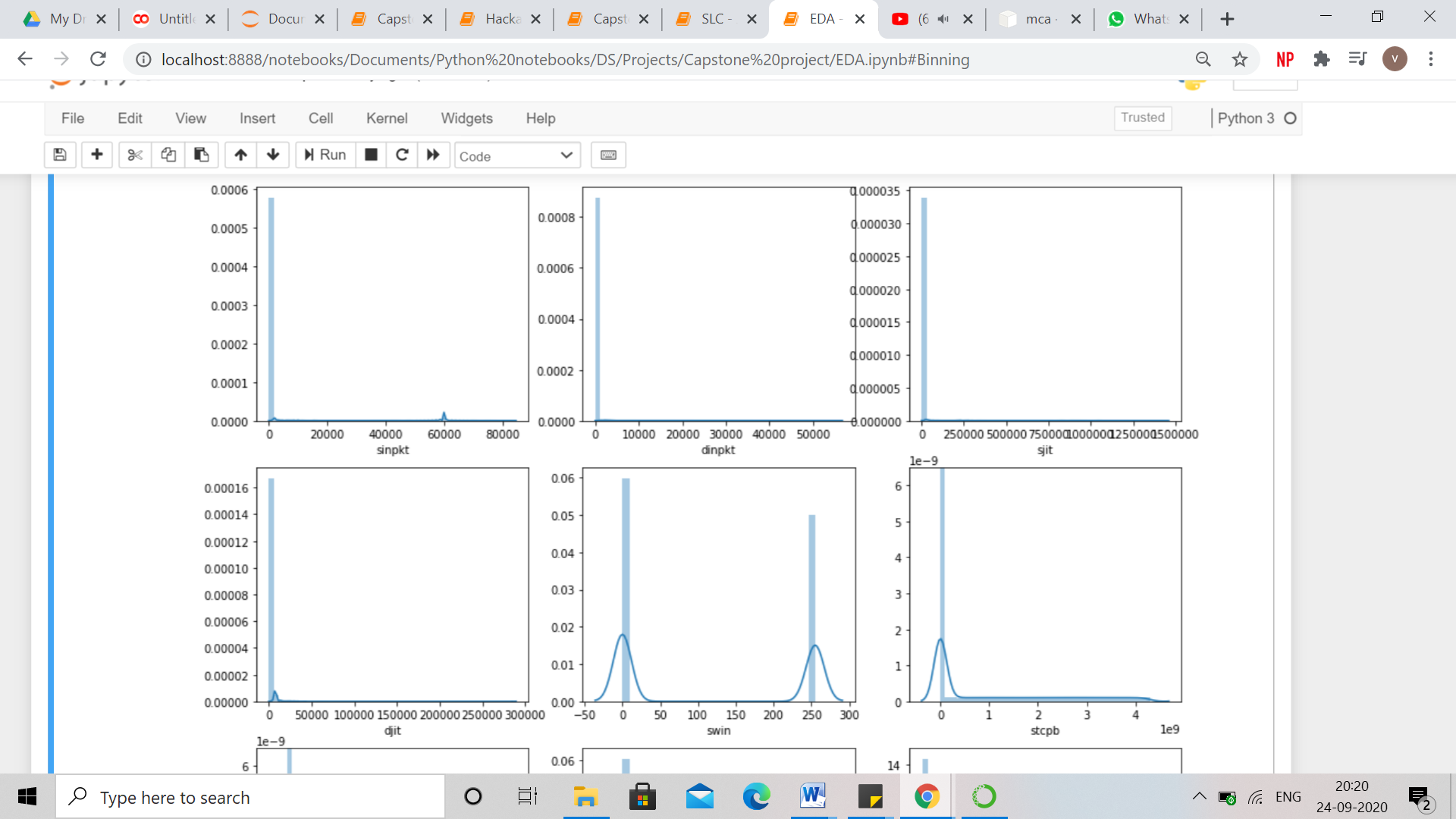
**Numerical Features**

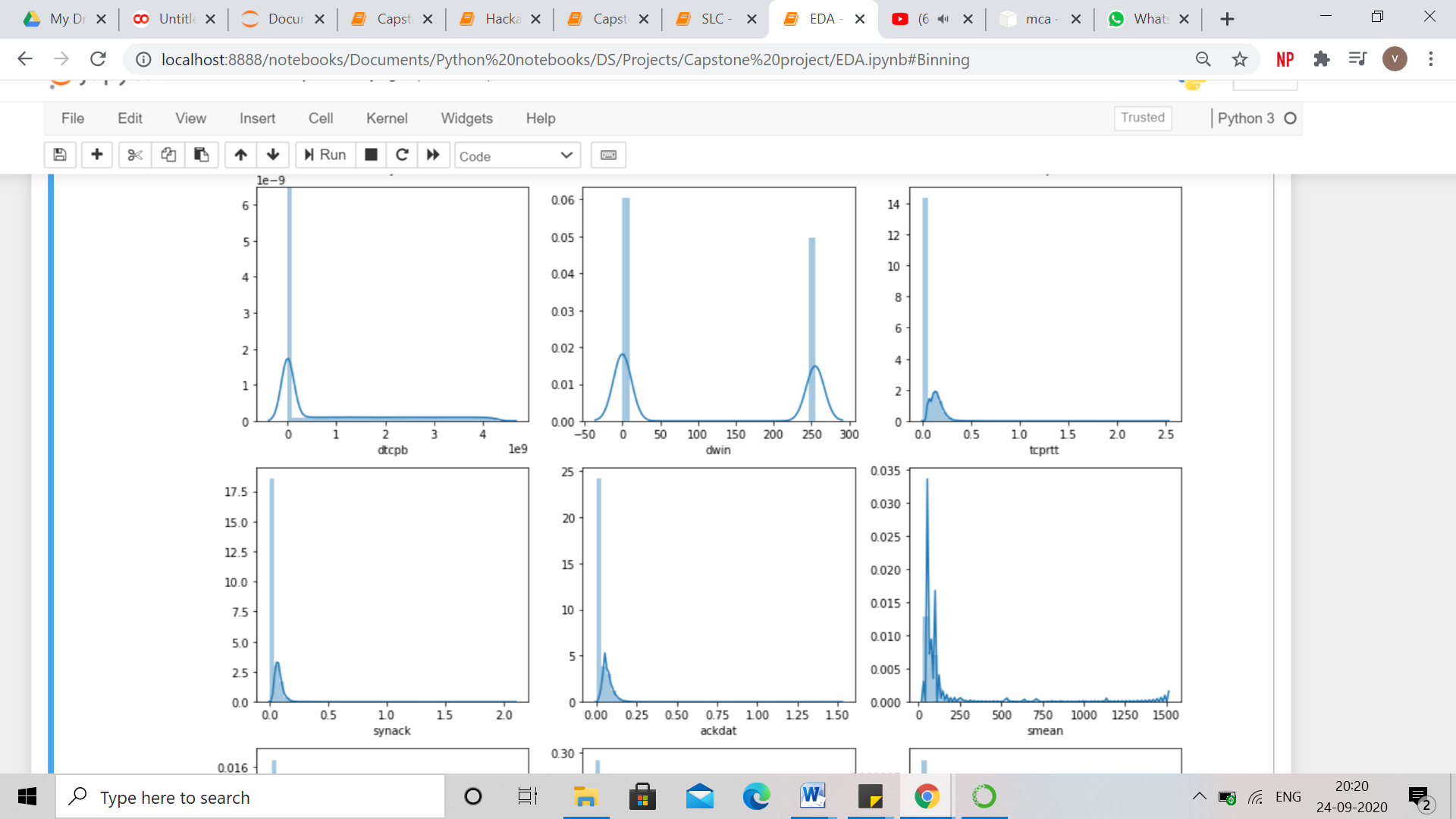
There are 39 numerical features present in the dataset.

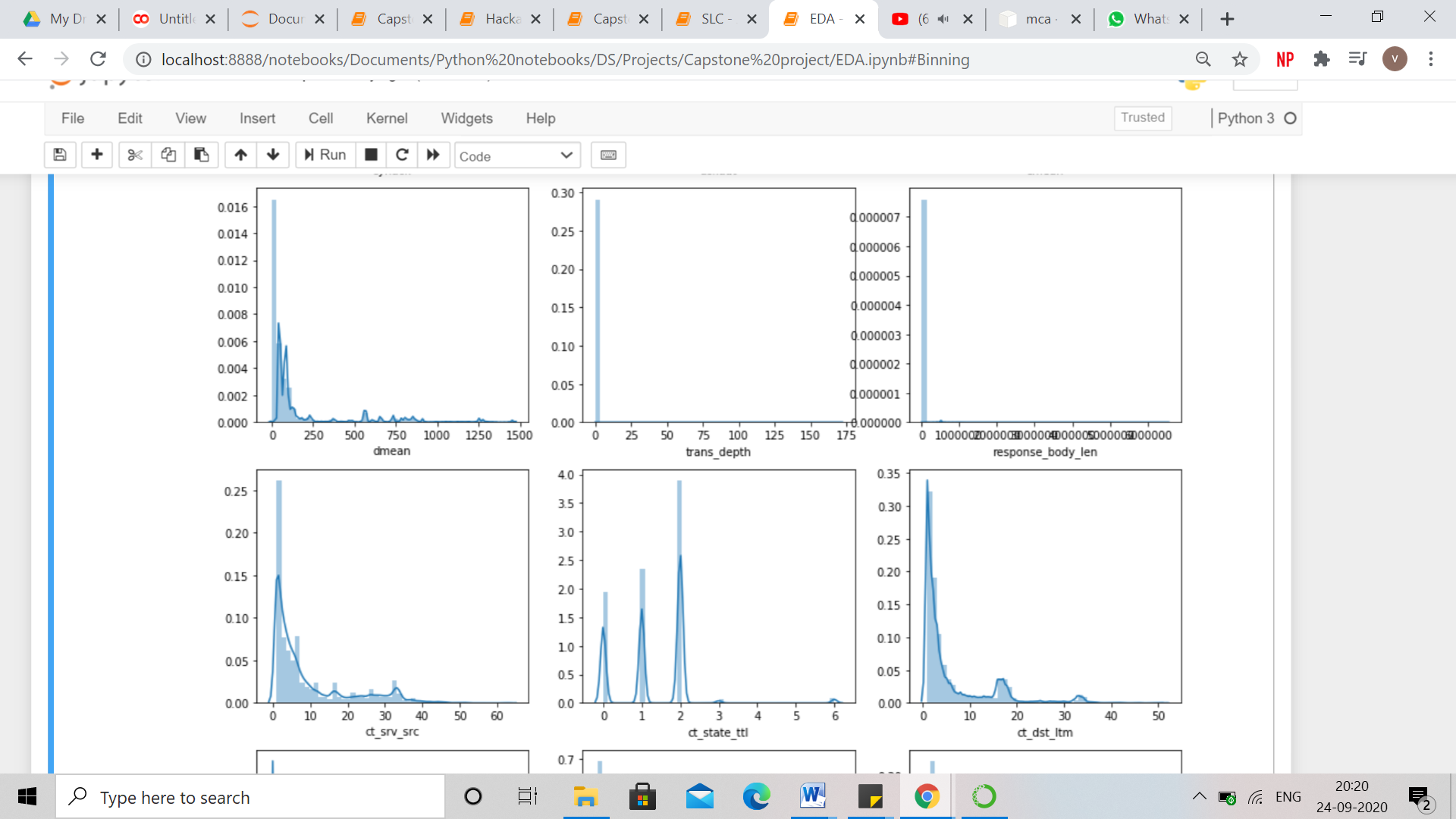
It can be observed that almost all the numerical features are right skewed and not normally distributed.

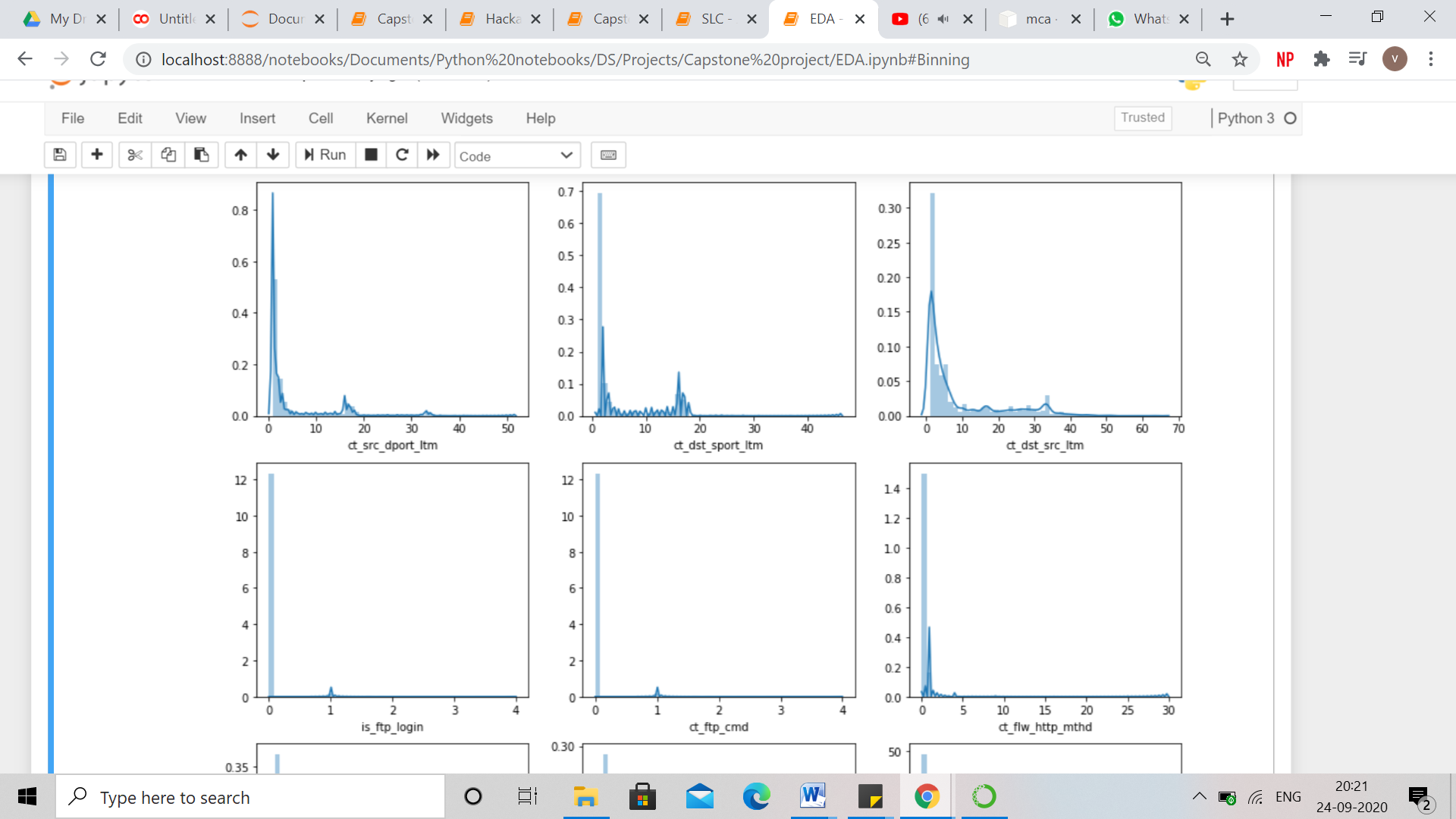


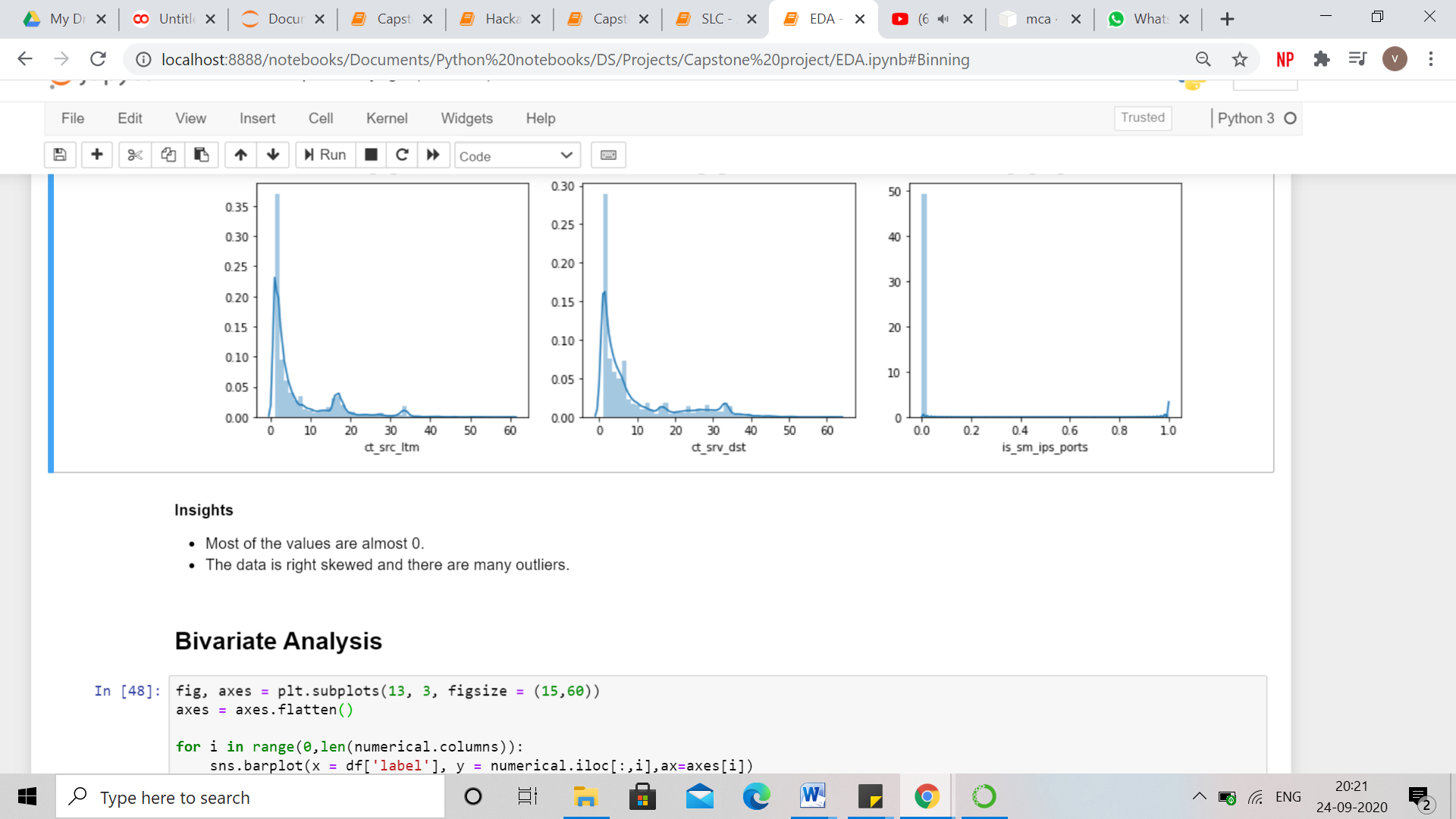












* **Presence of Outliers and its treatment**

Outliers are stragglers — extremely high or extremely low values — in a data set that can throw off your stats. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

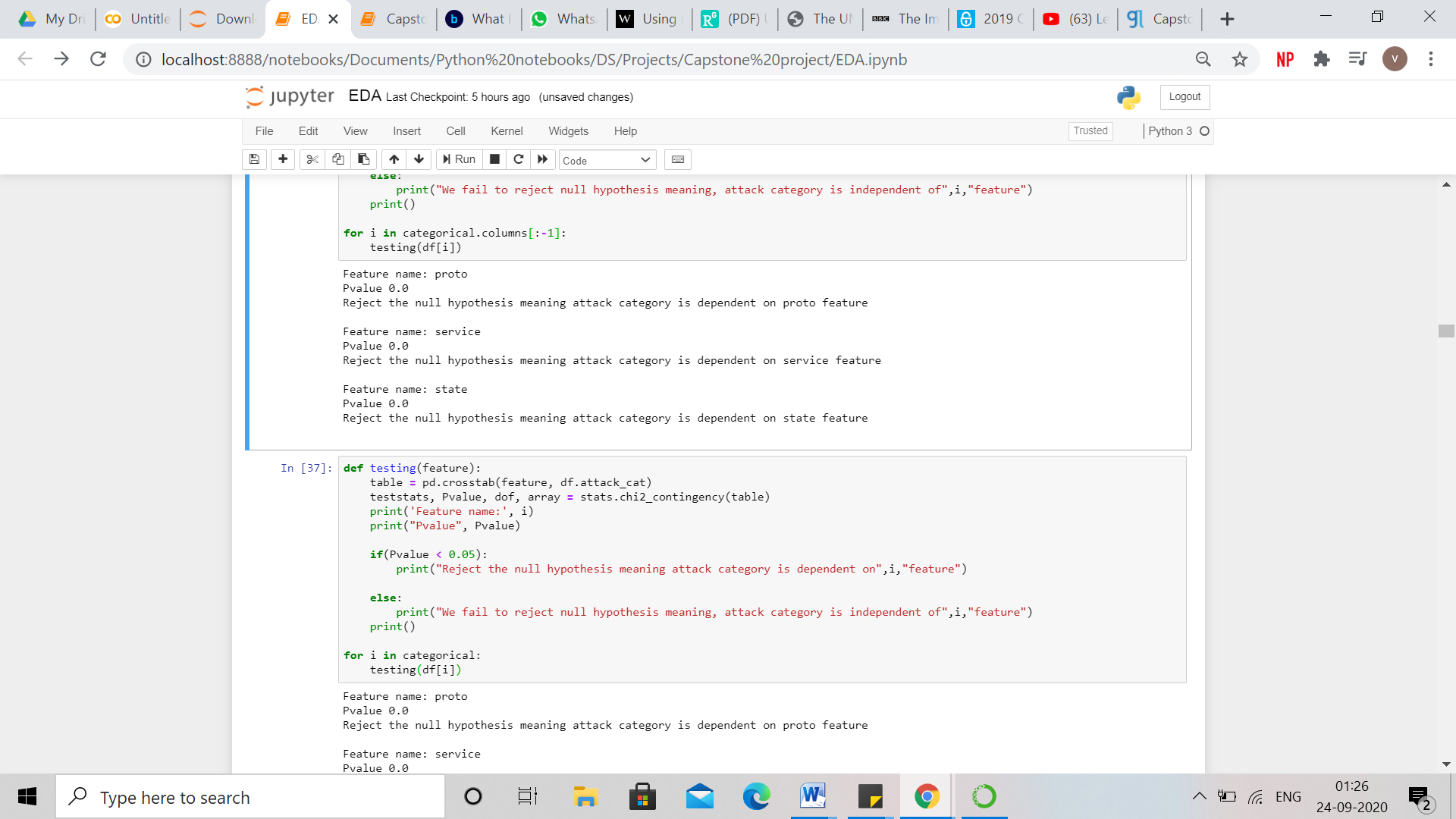
From the distribution plot of numerical features, it can be seen that the features are right skewed and hence there are outliers present at the right or higher side of the data. Higher values have lower frequency.

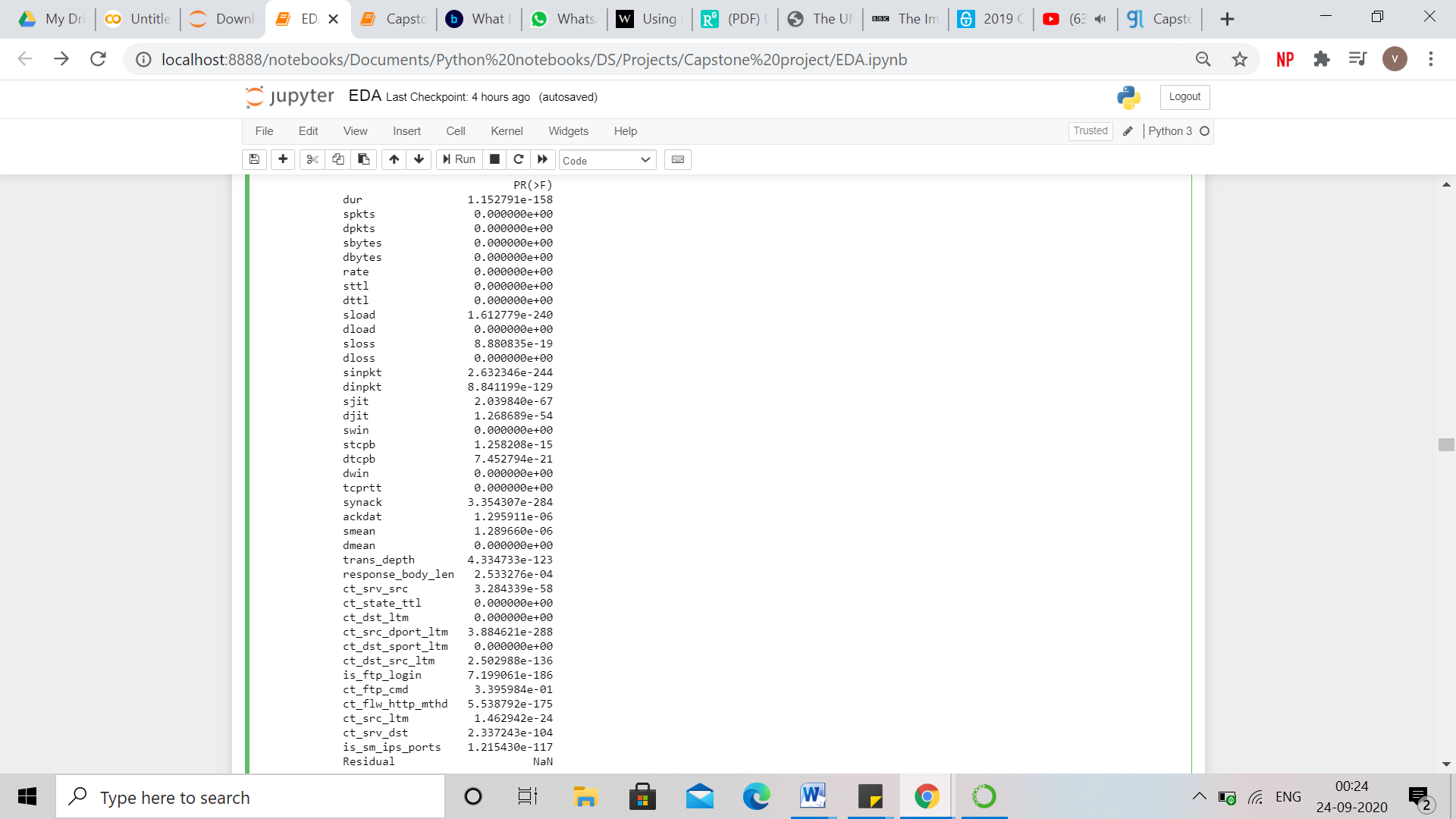
These outliers are a significant part of the data and help in understanding the patterns and making clusters and not a noise in the data. Therefore, these outliers are not treated and left as it is.

* **Statistical Significance of variables**

Statistical significance is the likelihood that a relationship between two or more variables is caused by something other than chance.

Statistical significance is used to provide evidence concerning the plausibility of the null hypothesis, which hypothesizes that there is nothing more than random chance at work in the data. Statistical hypothesis testing is used to determine whether the result of a data set is statistically significant.

****



Statistical Significance of numerical features has been tested by ANOVA and for categorical features, chisquare has been used.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

The null hypothesis states that the feature does not have any significant effect on the result/target feature. Whereas, the alternate hypothesis states that at least one of the sample means is different from another and the feature has a significant effect on the target feature.

The Chi Square statistic is commonly used for testing relationships between categorical features. The null hypothesis of the Chi-Square test is that no relationship exists between the categorical features i.e. they are independent of each other and whereas alternate hypothesis states that there is a relationship between the categorical features and independent feature has an effect on the target variable.

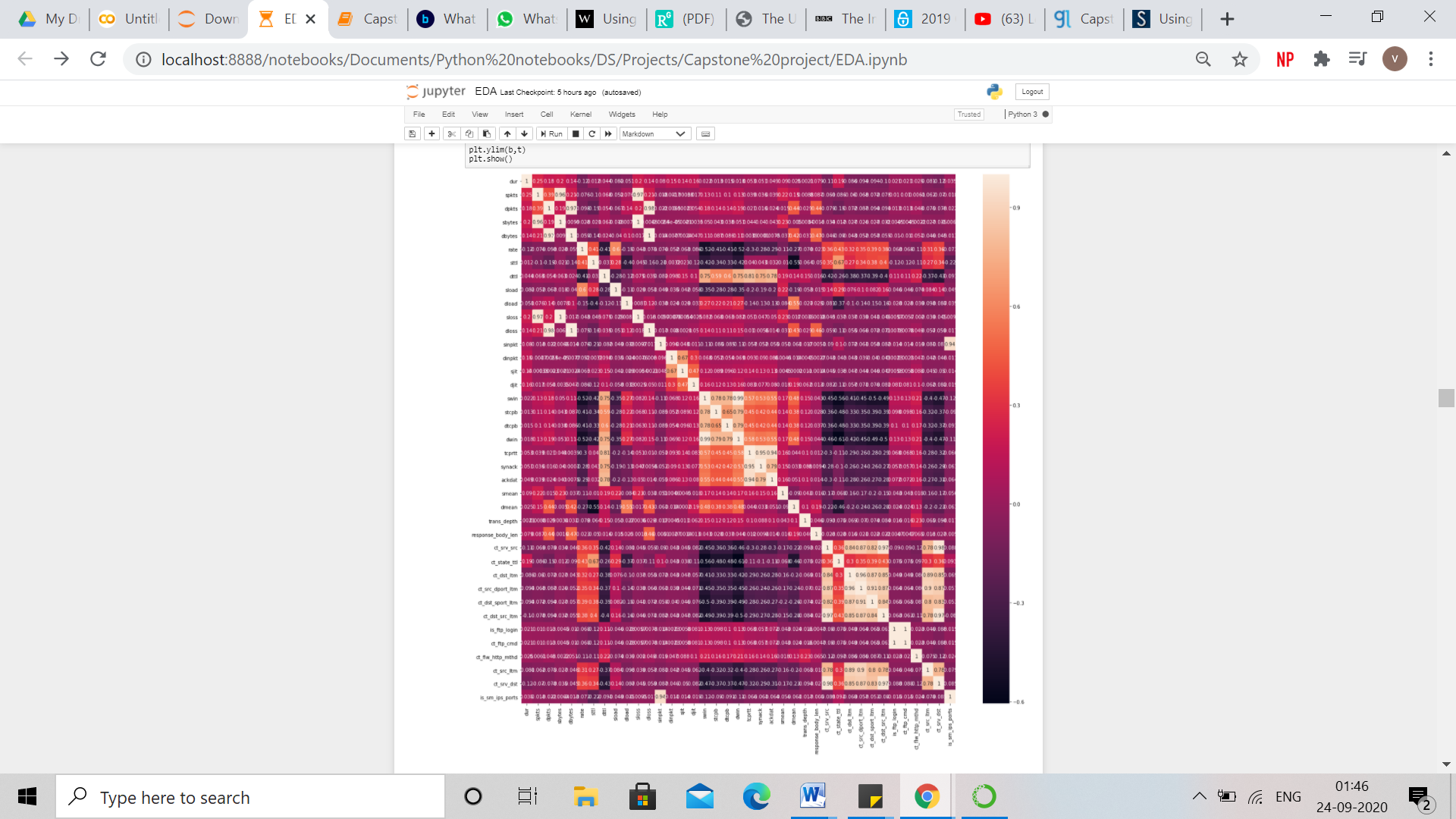
In both the cases, pvalue is less than the alpha( default = 0.05) which states that null hypothesis has been rejected and the features are statistically significant to predict the label - target variable.

* **Multicollinearity**

Multicollinearity occurs when [independent variables](https://statisticsbyjim.com/glossary/predictor-variables/) in a [regression](https://statisticsbyjim.com/glossary/regression-analysis/) model are correlated. This [correlation](https://statisticsbyjim.com/glossary/correlation/) is a problem because independent variables should be *independent*. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

The problem of multicollinearity is different in clustering analysis because there’s no dependent variable or beta coefficient. A certain number of observations measured on a specified number of variables are used for creating segments. Each observation belongs to one segment, and each segment can be defined in terms of all the variables used in the analysis. When variables used in clustering are collinear, some variables get a higher weight than others. If two variables are perfectly correlated, they effectively represent the same concept.

One major use of PCA lies in overcoming the multicollinearity problem. PCA can aptly deal with such situations by excluding some of the low-variance principal components.

****