

SML2010 Data Set

On-line learning of indoor temperature forecasting model towards energy efficiency

PREPARED BY:

SHAYAN PAUL

Machine learning Intern| AIROBOTICA Service Pvt Ltd

INTRODUCTION

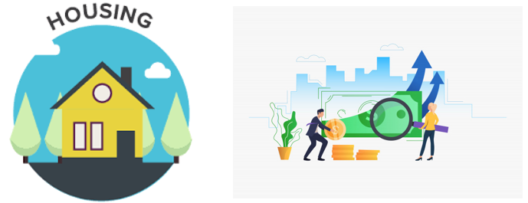
* [What is Regression?](https://www.edureka.co/blog/linear-regression-for-machine-learning/#regression)
* [Types of Regression](https://www.edureka.co/blog/linear-regression-for-machine-learning/#types)
* [What is Linear Regression?](https://www.edureka.co/blog/linear-regression-for-machine-learning/#linear)
* [Linear Regression Terminologies](https://www.edureka.co/blog/linear-regression-for-machine-learning/#terms)
* [Advantages And Disadvantages Of Linear Regression](https://www.edureka.co/blog/linear-regression-for-machine-learning/#advantages)
* [Linear Regression Use Cases](https://www.edureka.co/blog/linear-regression-for-machine-learning/#usecases)

## [Use case – Linear Regression Implementation](https://www.edureka.co/blog/linear-regression-for-machine-learning/" \l "example)

## What is regression?

The main goal of regression is the construction of an efficient model to predict the dependent attributes from a bunch of attribute variables. A regression problem is when the output variable is either real or a continuous value i.e salary, weight, area, etc.

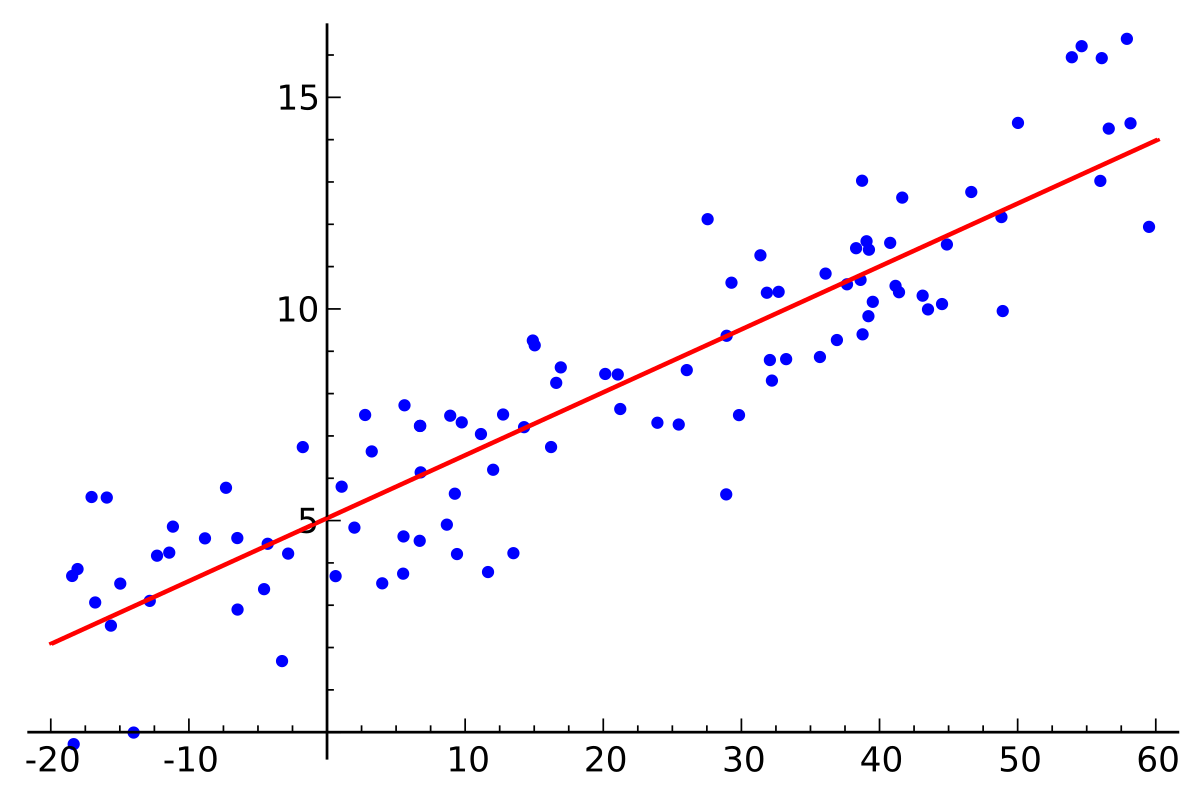
We can also define regression as a statistical means that is used in applications like housing, investing, etc. It is used to predict the relationship between a dependent variable and a bunch of independent variables. Let us take a look at various types of regression techniques.



## What is linear regression?

The main goal of regression is the construction of an efficient model to predict the dependent attributes from a bunch of attribute variables. A regression problem is when the output variable is either real or a continuous value i.e salary, weight, area, etc.

We can also define regression as a statistical means that is used in applications like housing, investing, etc. It is used to predict the relationship between a dependent variable and a bunch of independent variables. Let us take a look at various types of regression techniques.



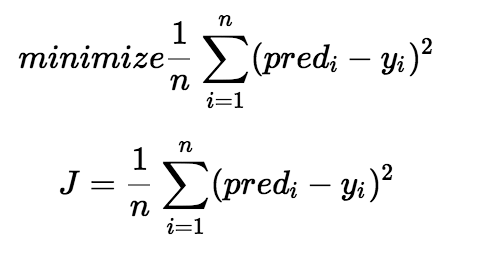
Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below.

y = a\_0 + a\_1 \* x ## Linear Equation

The motive of the linear regression algorithm is to find the best values for a\_0 and a\_1. Before moving on to the algorithm, let’s have a look at two important concepts you must know to better understand linear regression.

## Cost Function

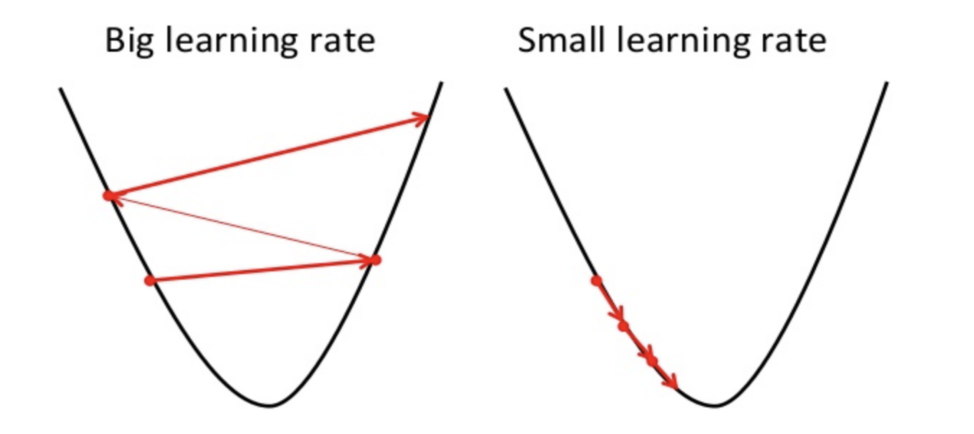
The cost function helps us to figure out the best possible values for a\_0 and a\_1 which would provide the best fit line for the data points. Since we want the best values for a\_0 and a\_1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.



We choose the above function to minimize. The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error(MSE) function. Now, using this MSE function we are going to change the values of a\_0 and a\_1 such that the MSE value settles at the minima.

## Gradient Descent

The next important concept needed to understand linear regression is gradient descent. Gradient descent is a method of updating a\_0 and a\_1 to reduce the cost function(MSE). The idea is that we start with some values for a\_0 and a\_1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.



To draw an analogy, imagine a pit in the shape of U and you are standing at the topmost point in the pit and your objective is to reach the bottom of the pit. There is a catch, you can only take a discrete number of steps to reach the bottom. If you decide to take one step at a time you would eventually reach the bottom of the pit but this would take a longer time. If you choose to take longer steps each time, you would reach sooner but, there is a chance that you could overshoot the bottom of the pit and not exactly at the bottom. In the gradient descent algorithm, the number of steps you take is the learning rate. This decides on how fast the algorithm converges to the minima.

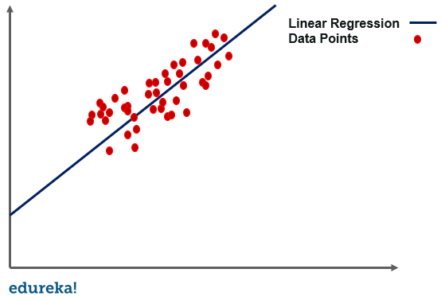
## **Types Of Regression**

The following are types of regression.

1. **Simple Linear Regression**
2. **Polynomial Regression**
3. **Support Vector Regression**
4. **Decision Tree Regression**
5. **Random Forest Regression**

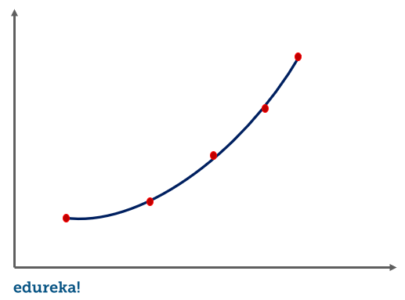
### **Simple Linear Regression**

One of the most interesting and common regression technique is simple linear regression. In this, we predict the outcome of a dependent variable based on the independent variables, the relationship between the variables is linear. Hence, the word linear regression.



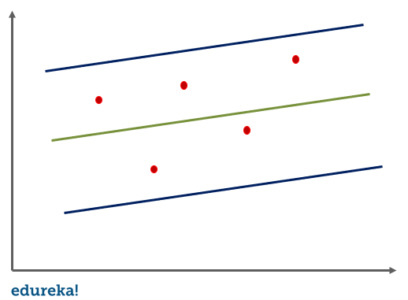
### **Polynomial Regression**

In this regression technique, we transform the original features into polynomial features of a given degree and then perform regression on it.



### **Support Vector Regression**

For [support vector machine](https://www.edureka.co/blog/support-vector-machine-in-python/) regression or SVR, we identify a hyperplane with maximum margin such that the maximum number of data points are within those margins. It is quite similar to the support vector machine classification algorithm.



### **Decision Tree Regression**

A [decision tree](https://www.edureka.co/blog/decision-tree-algorithm/) can be used for both regression and [classification](https://www.edureka.co/blog/classification-in-machine-learning/). In the case of regression, we use the ID3 algorithm(Iterative Dichotomiser 3) to identify the splitting node by reducing the standard deviation.

### **Random Forest Regression**

In random forest regression, we ensemble the predictions of several decision tree regressions. Now that we know about different types of regression let us take a look at simple linear regression in detail.

## **Advantages And Disadvantages**

|  |  |
| --- | --- |
| **Advantages** | **Disadvantages** |
| Linear regression performs exceptionally well for linearly separable data | The assumption of linearity between dependent and independent variables |
| Easier to implement, interpret and efficient to train | It is often quite prone to noise and overfitting |
| It handles overfitting pretty well using dimensionally reduction techniques, regularization, and cross-validation | Linear regression is quite sensitive to outliers |
| One more advantage is the extrapolation beyond a specific data set | It is prone to multicollinearity |

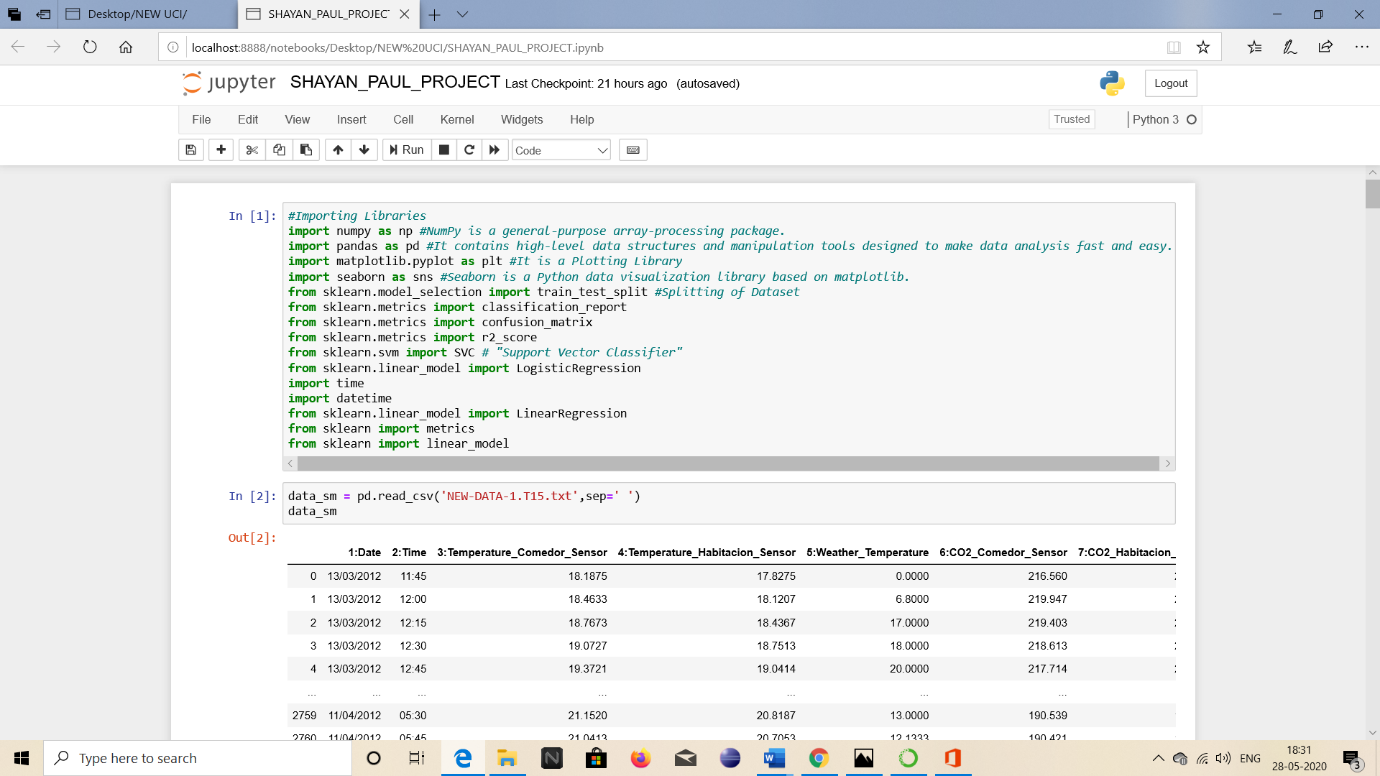
## **Use Case – Implementing Linear Regression**

The process takes place in the following steps:

1. [**Importing Libraries and Packages**](https://www.kaggle.com/oktayozturk/linear-regression-example#p1)
2. [**Loading and Viewing Data Set**](https://www.kaggle.com/oktayozturk/linear-regression-example#p2)
3. [**Plotting and Visualizing Data**](https://www.kaggle.com/oktayozturk/linear-regression-example#p3)
4. **Modeling and Predicting with sklearn**

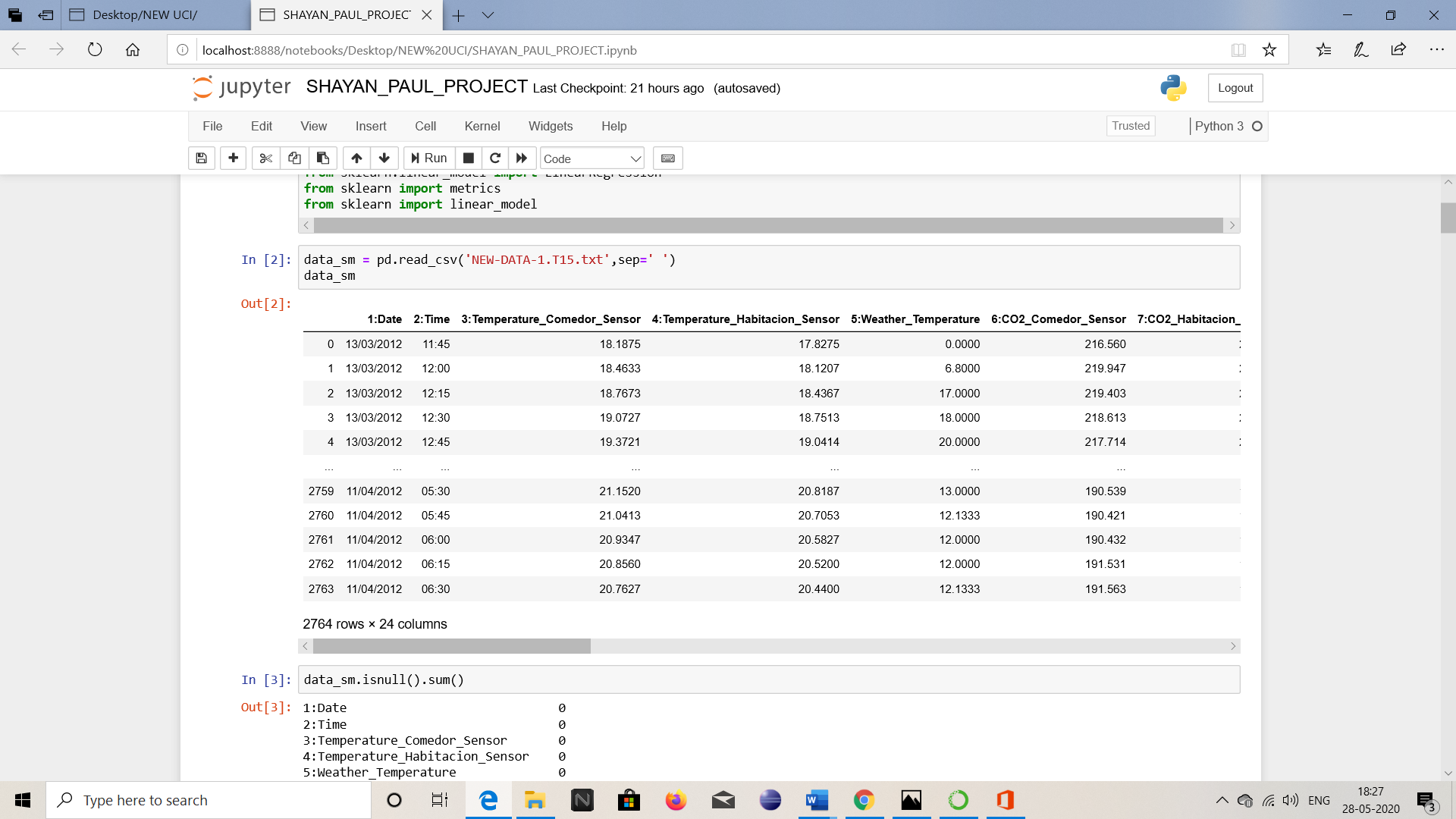
**1. Importing Libraries and Packages**

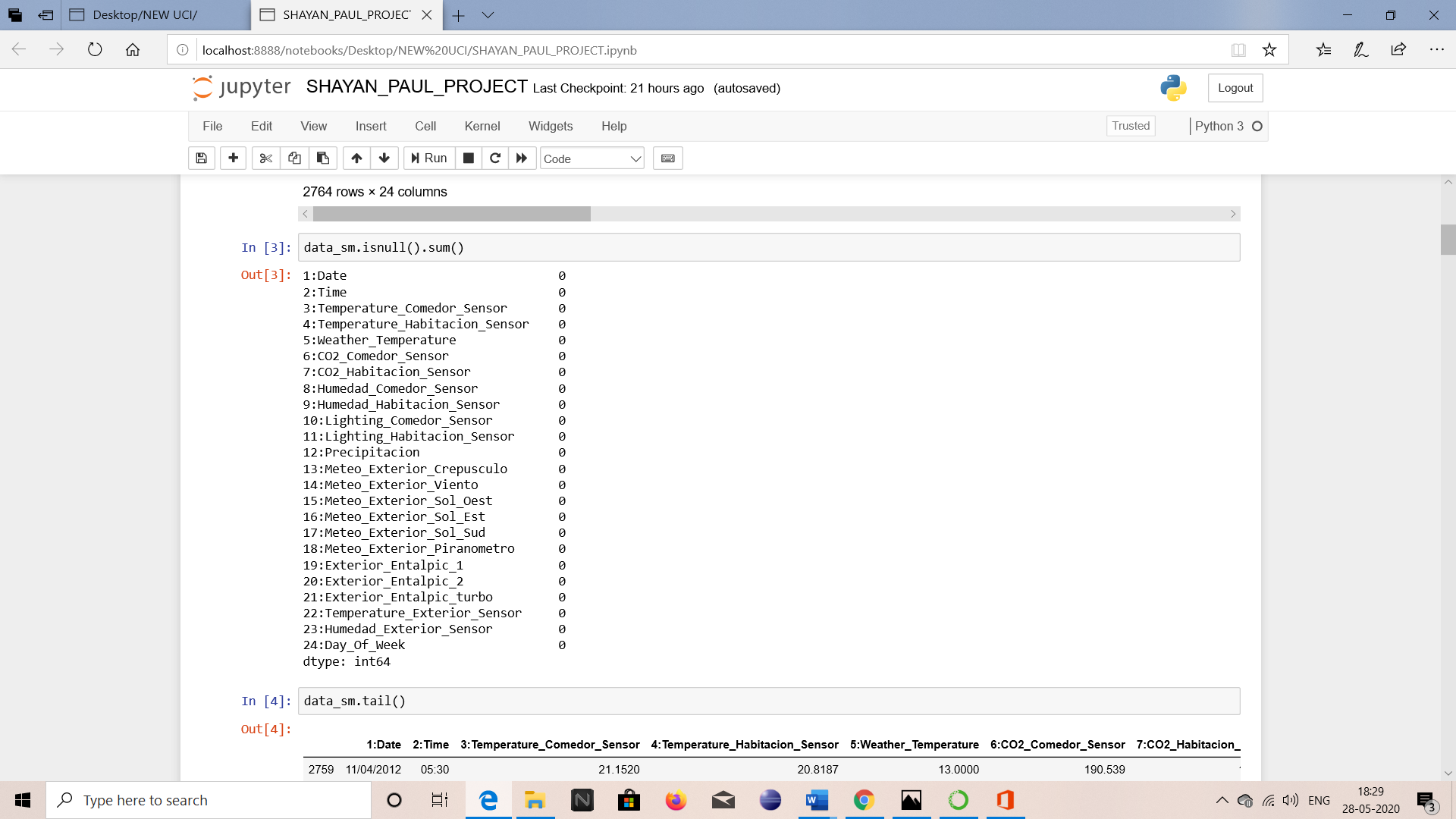
We will use these packages to help us manipulate the data and visualize the features/labels as well as measure how well our model performed. Numpy and Pandas are helpful for manipulating the dataframe and its columns and cells. We will use matplotlib along with Seaborn to visualize our data.



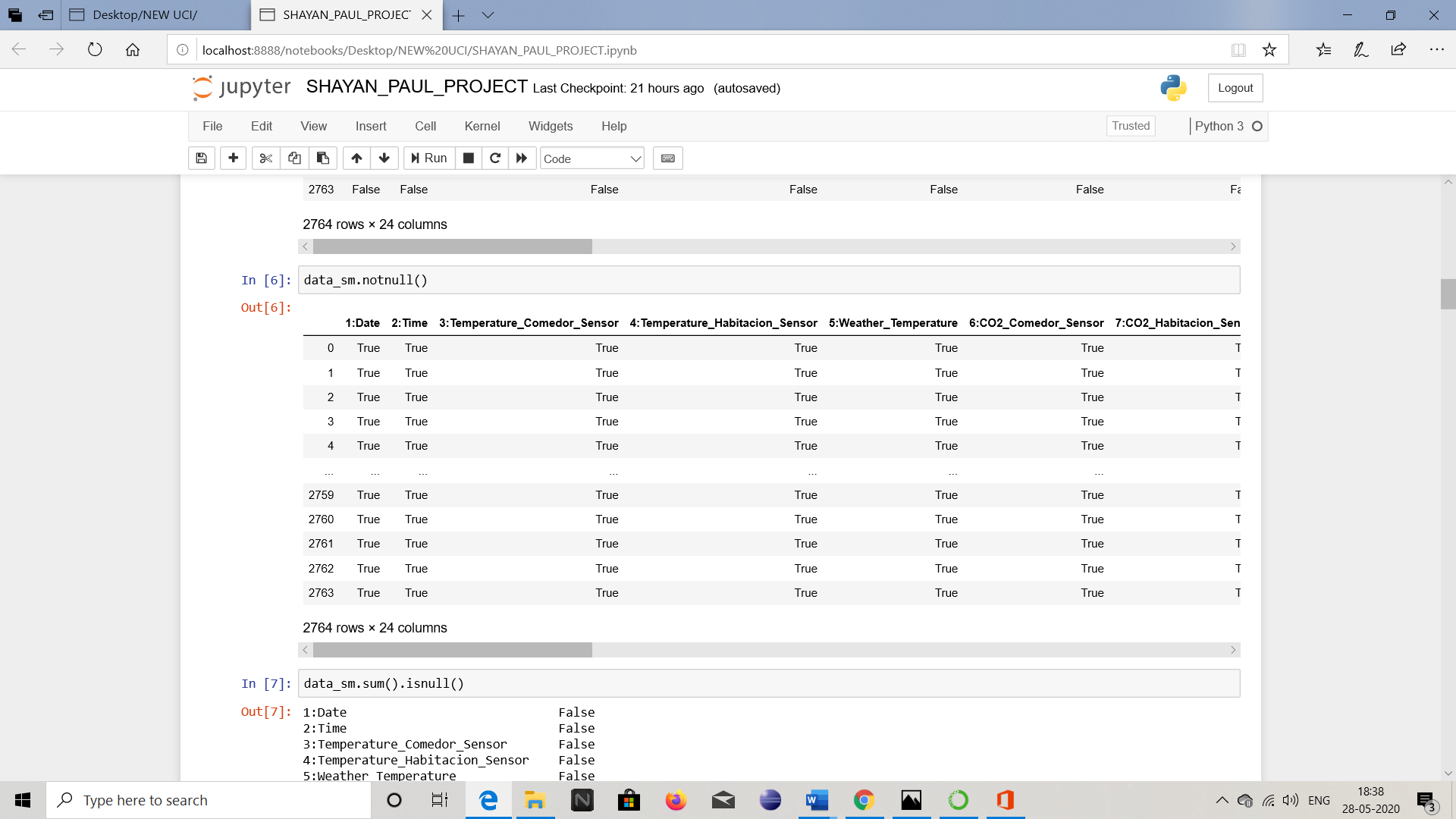
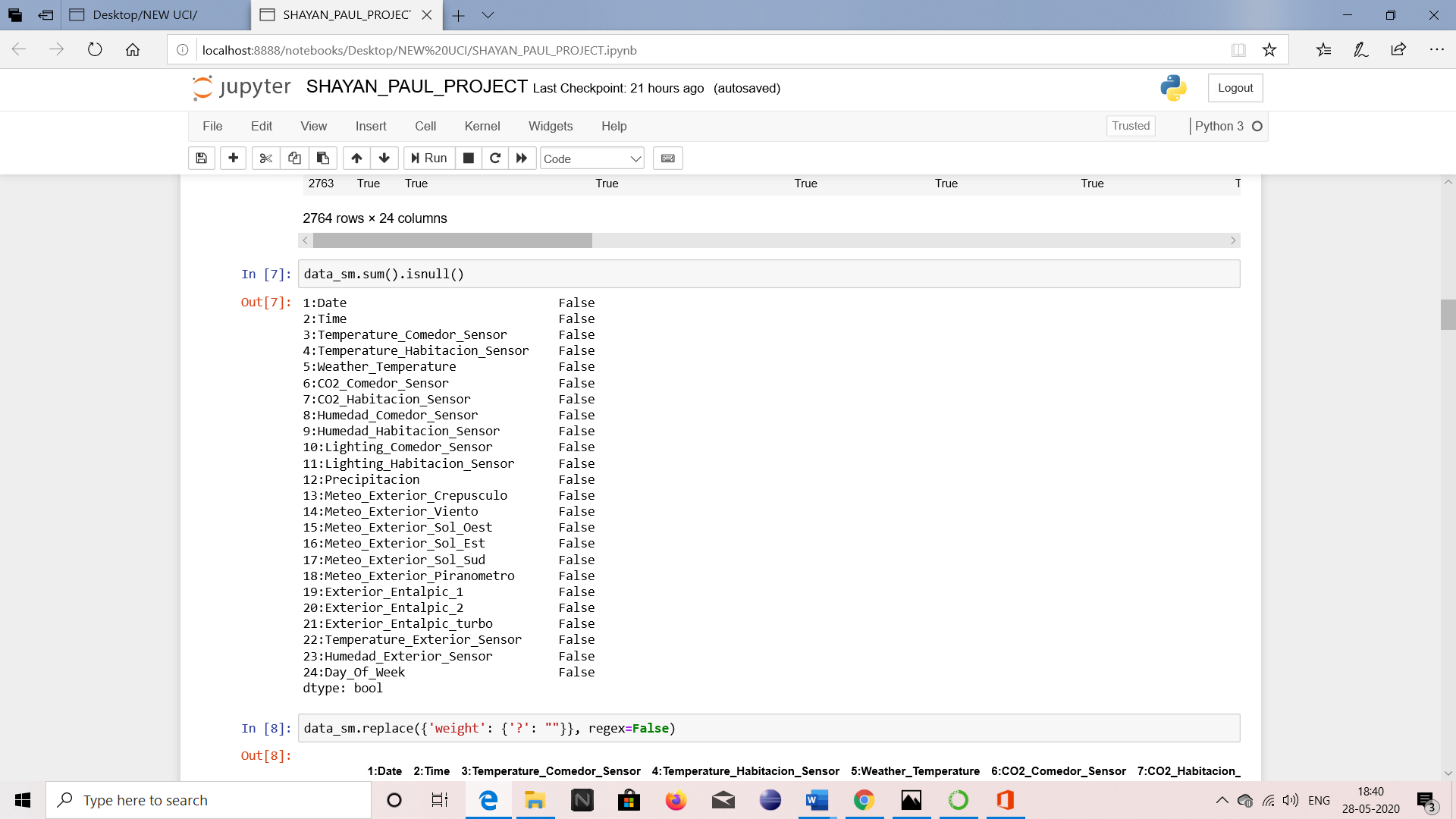
**2. Loading and Viewing Data Set**

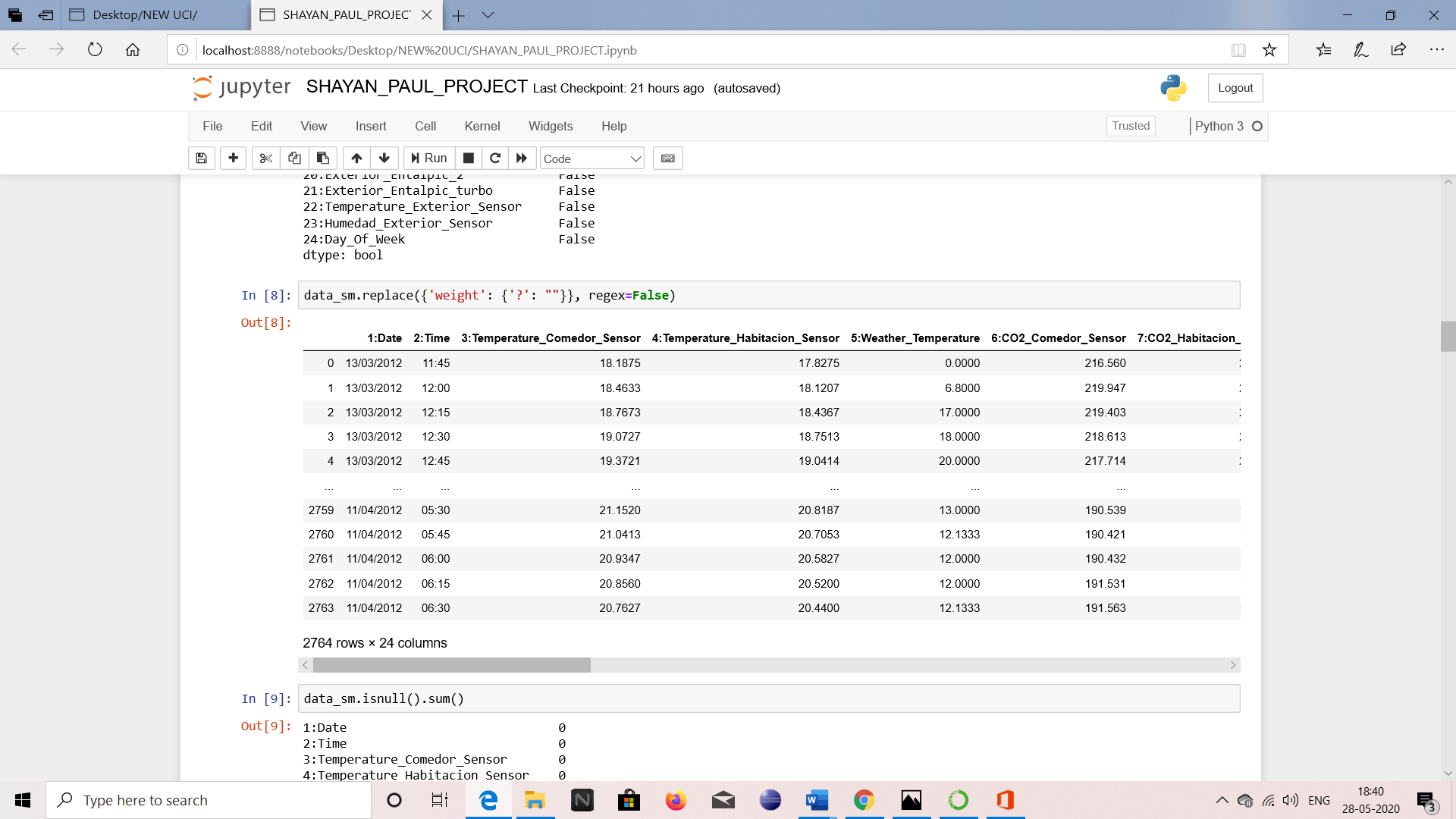
With Pandas, we can load both the training and testing set that we wil later use to train and test our model. Before we begin, we should take a look at our data table to see the values that we'll be working with. We can use the head and describe function to look at some sample data and statistics.

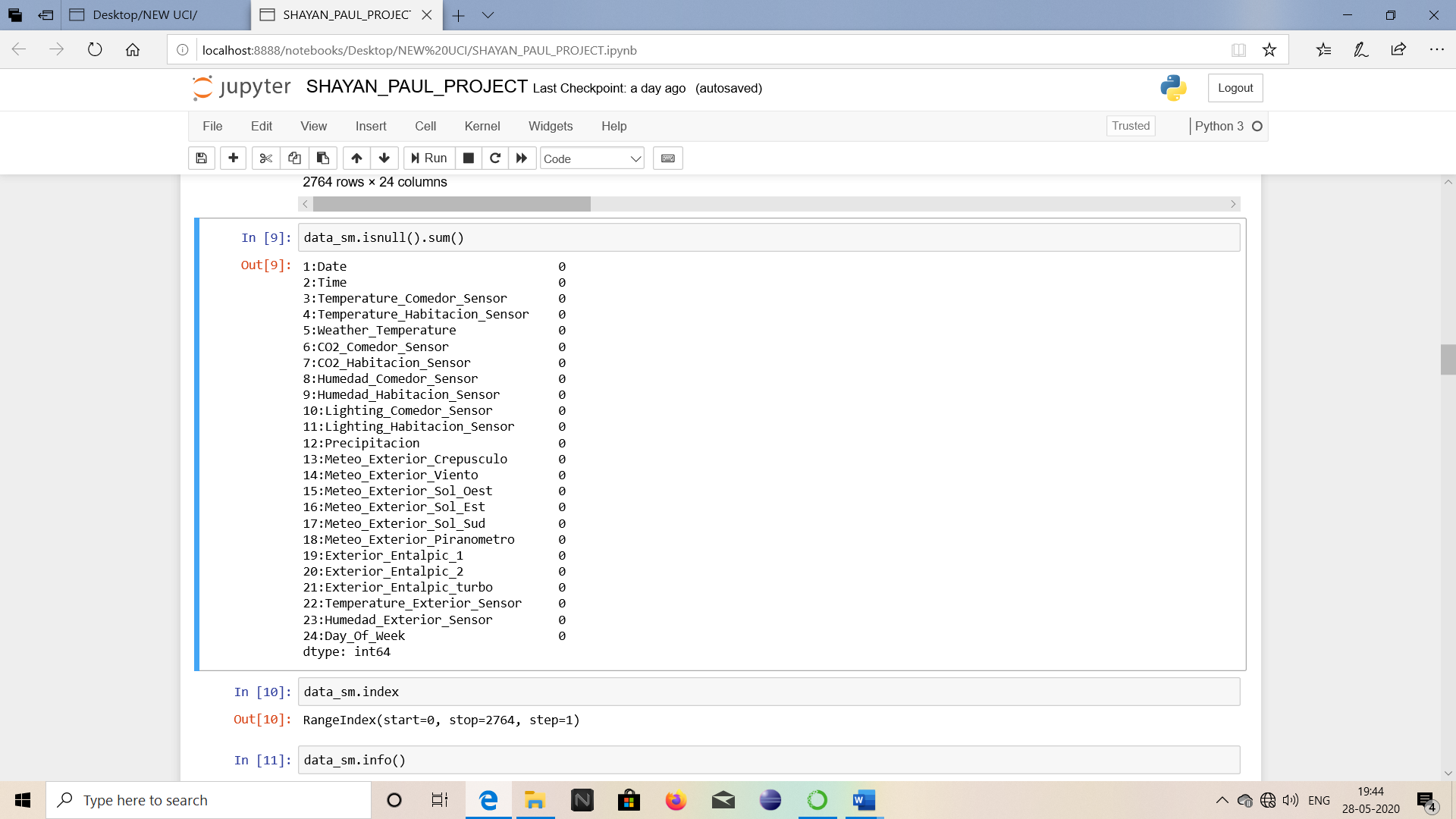


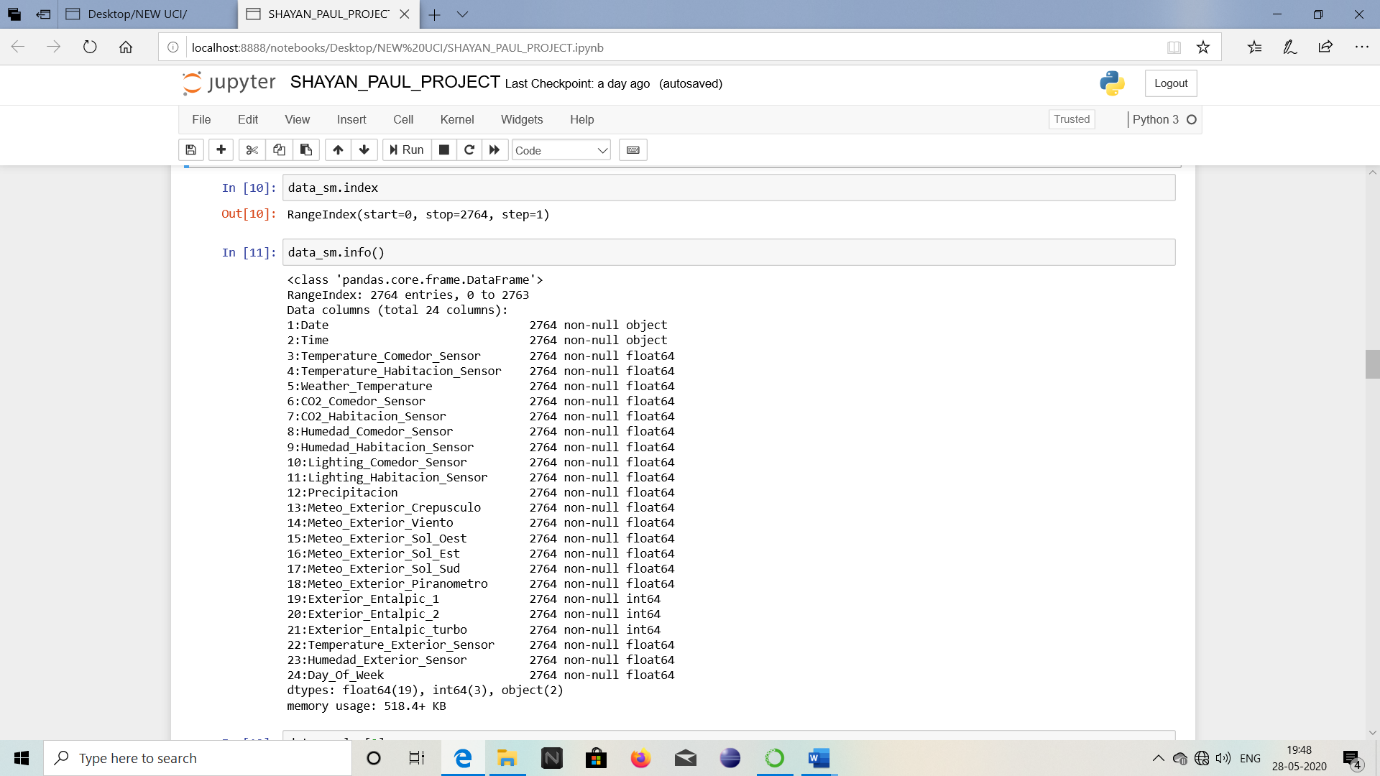


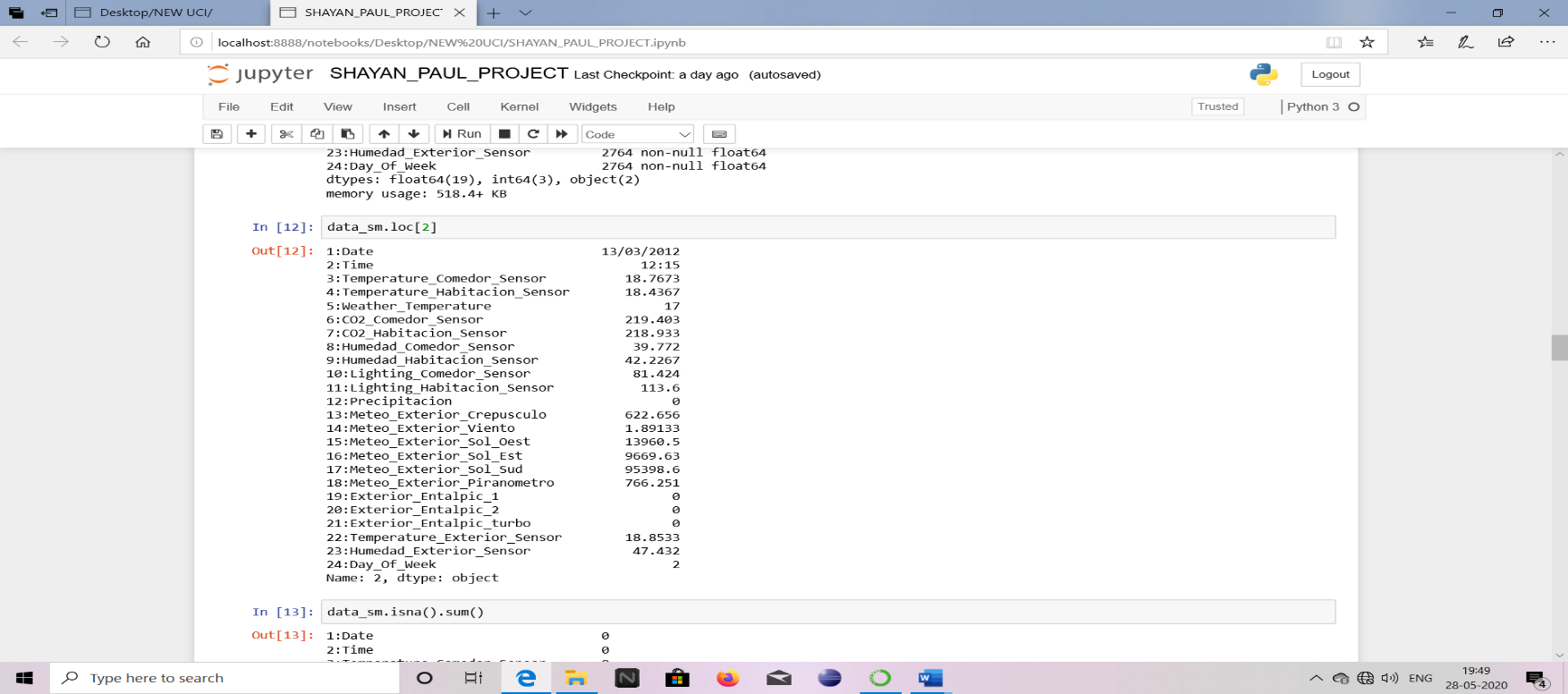




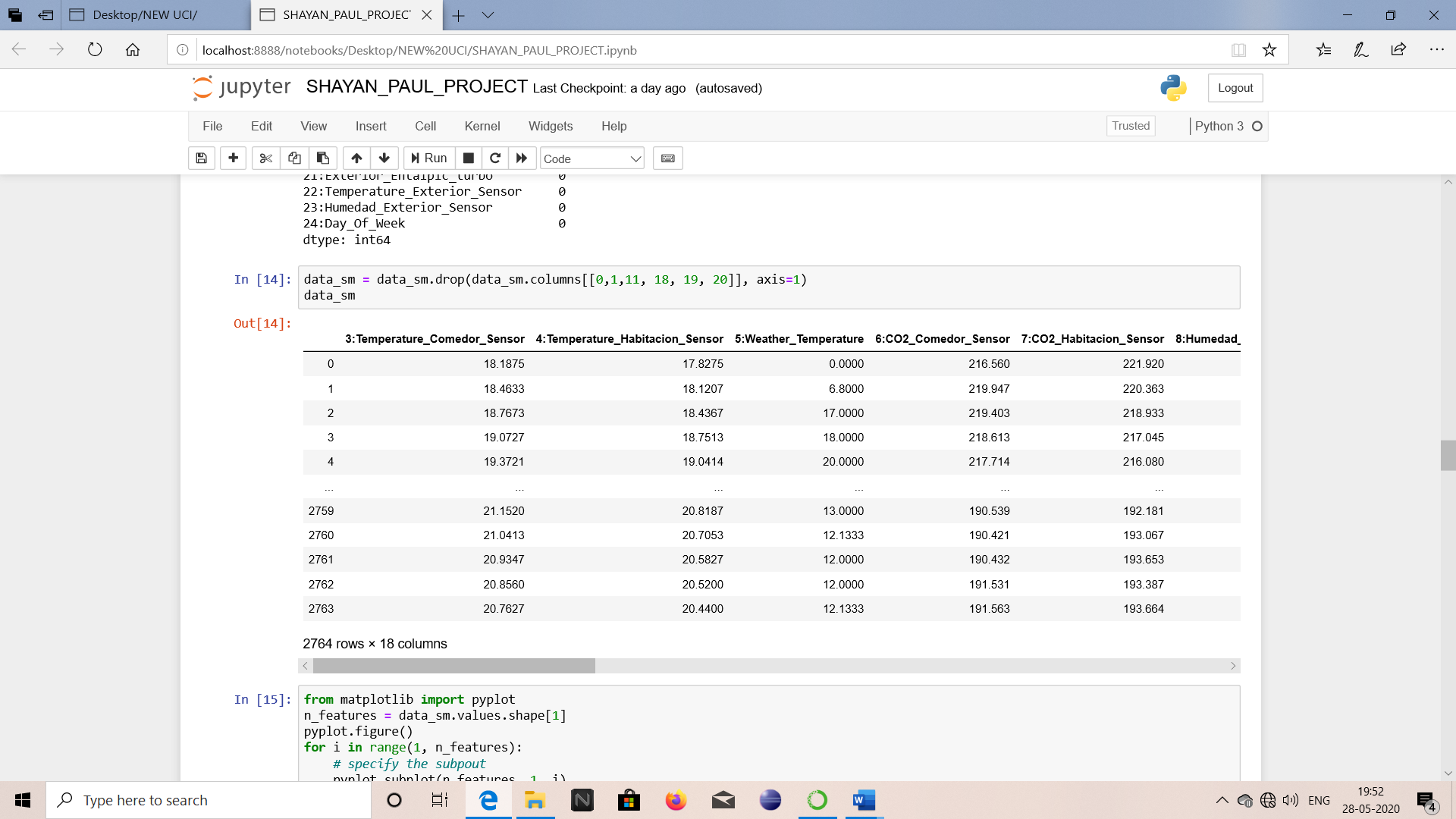




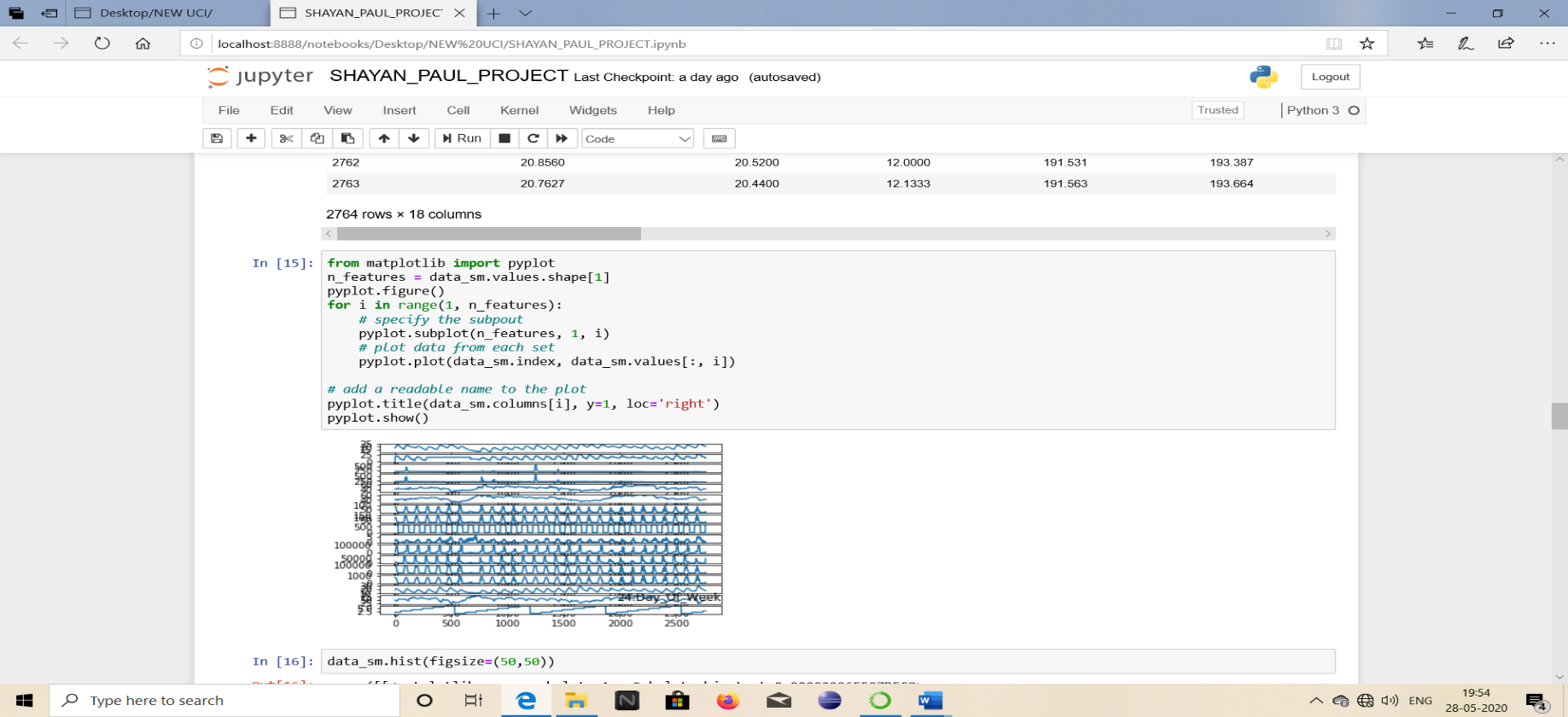


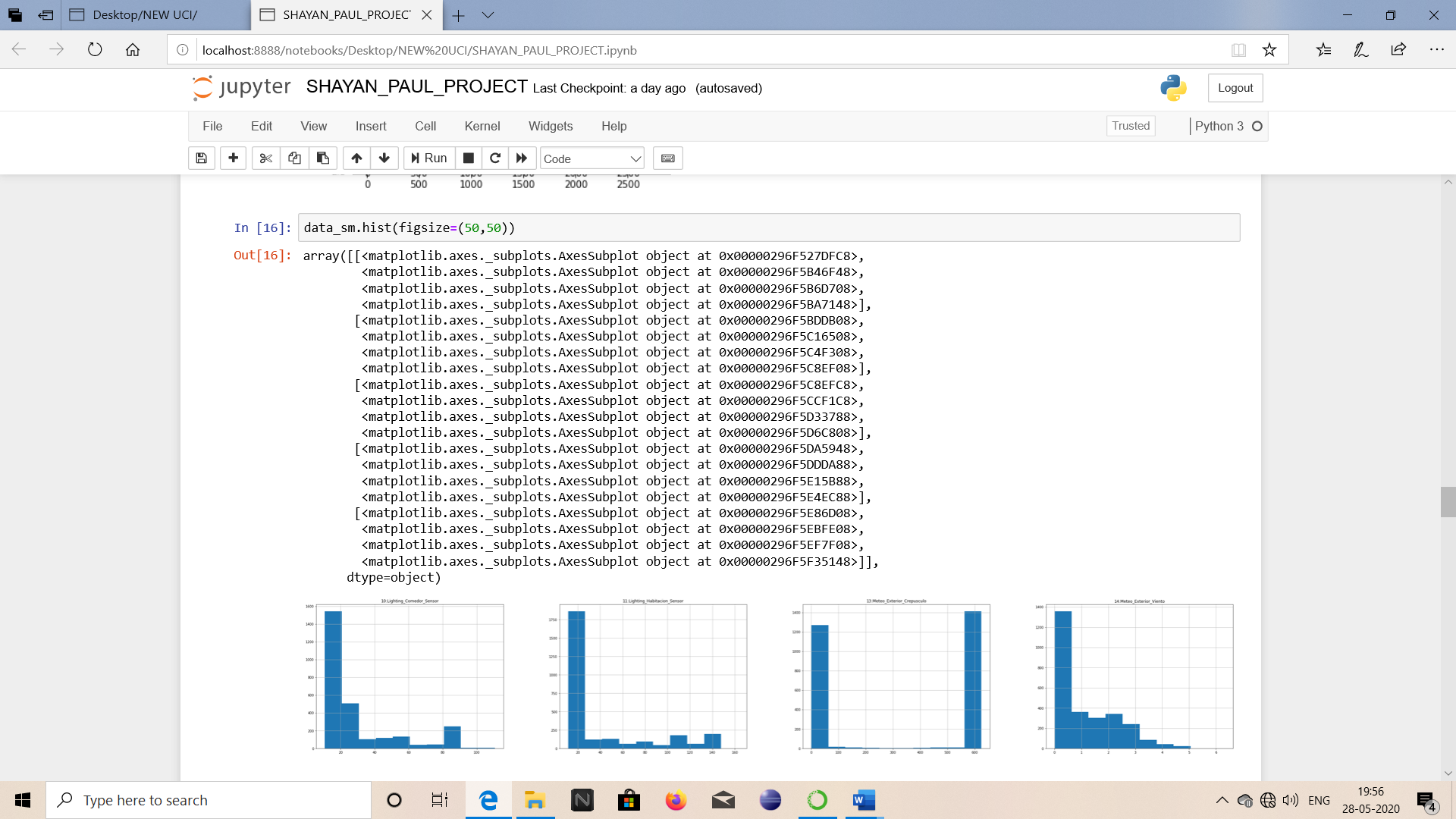


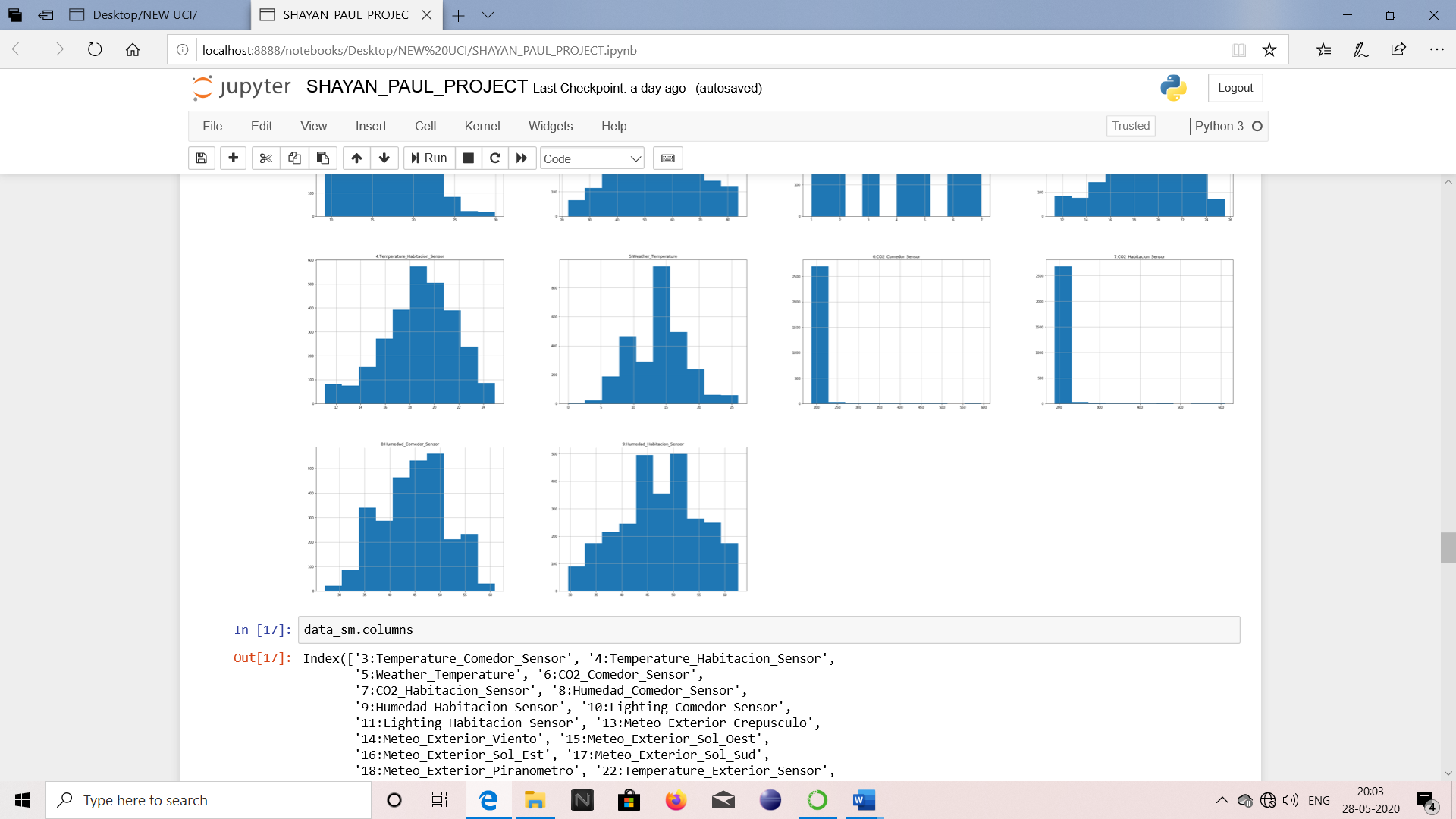


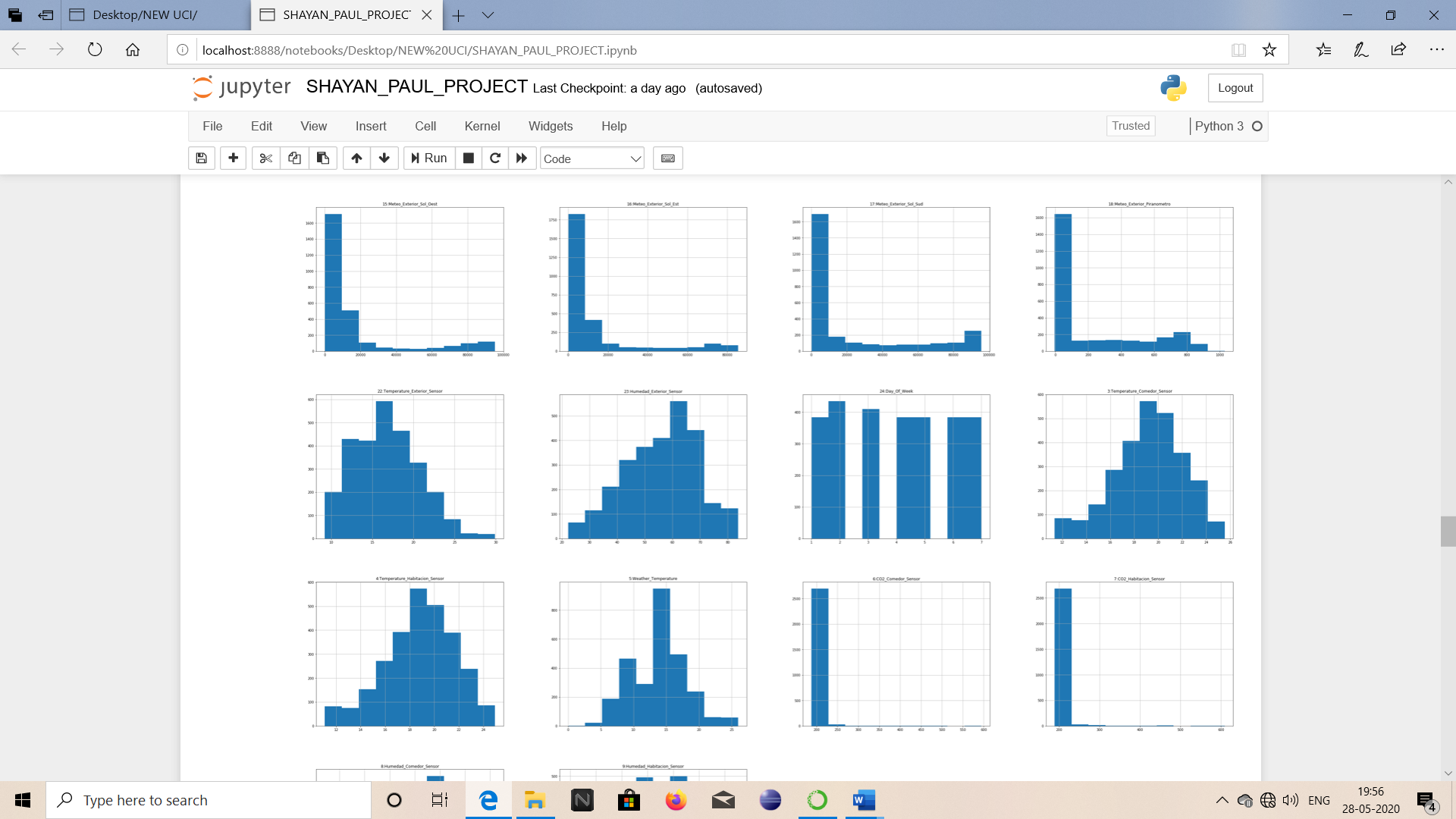


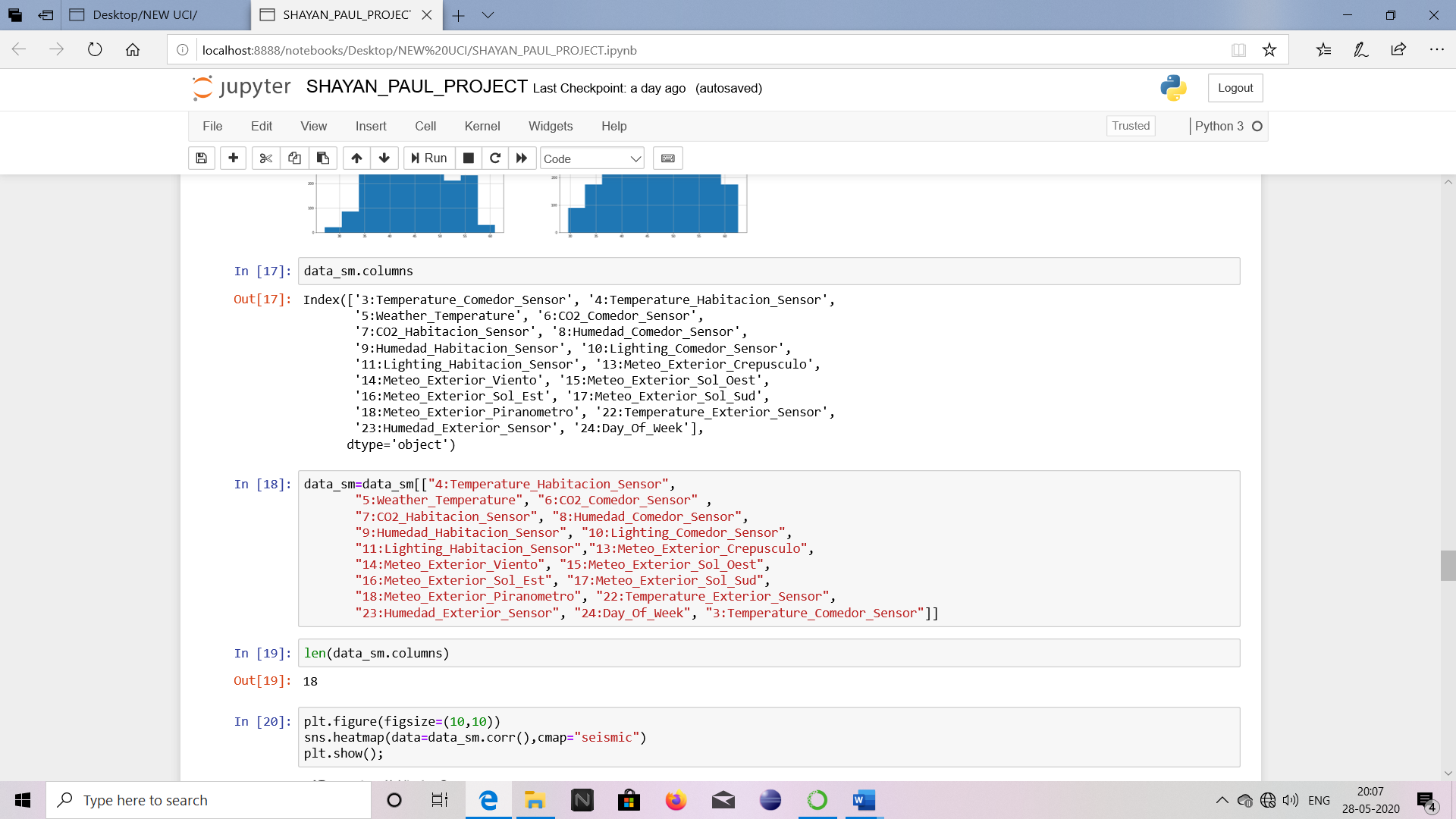
**3. Plotting and Visualizing Data**

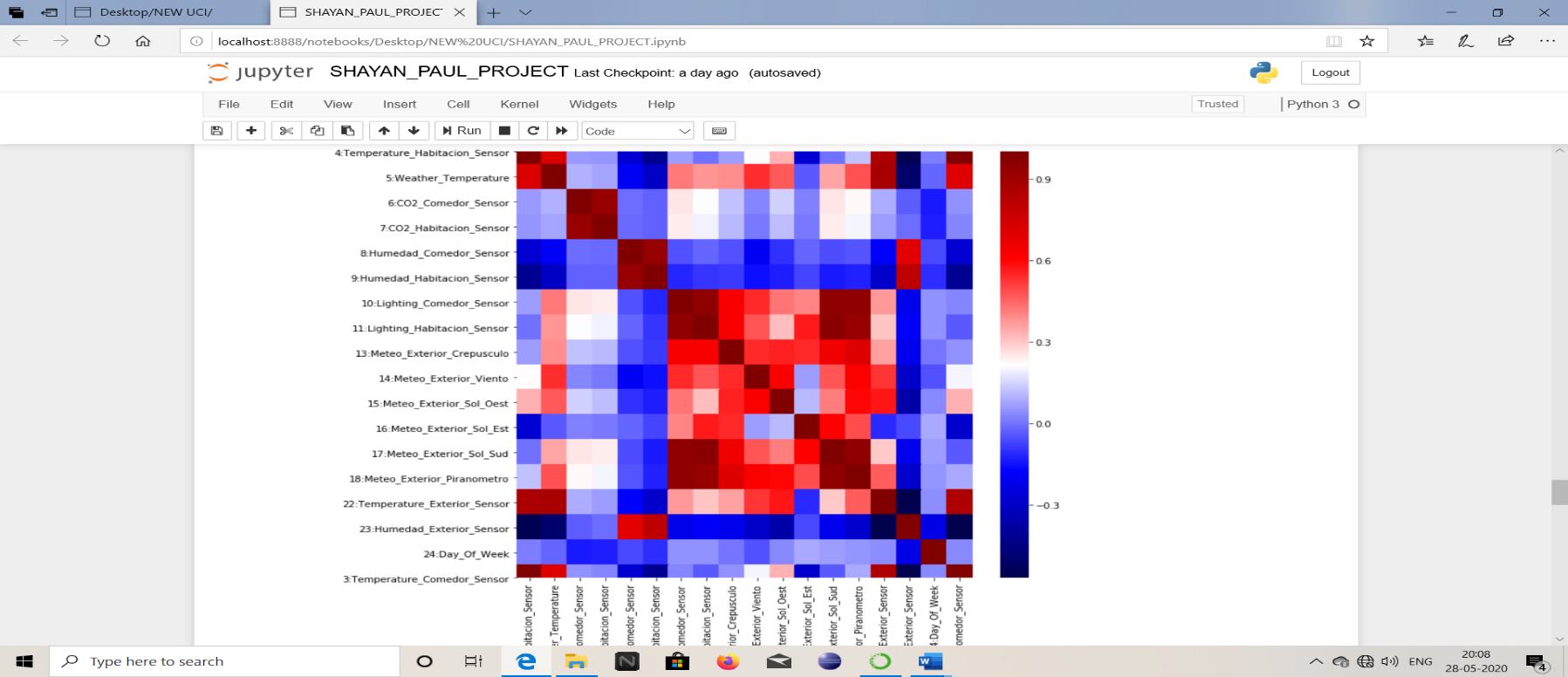
 Plotting the data with matplotlib scatter





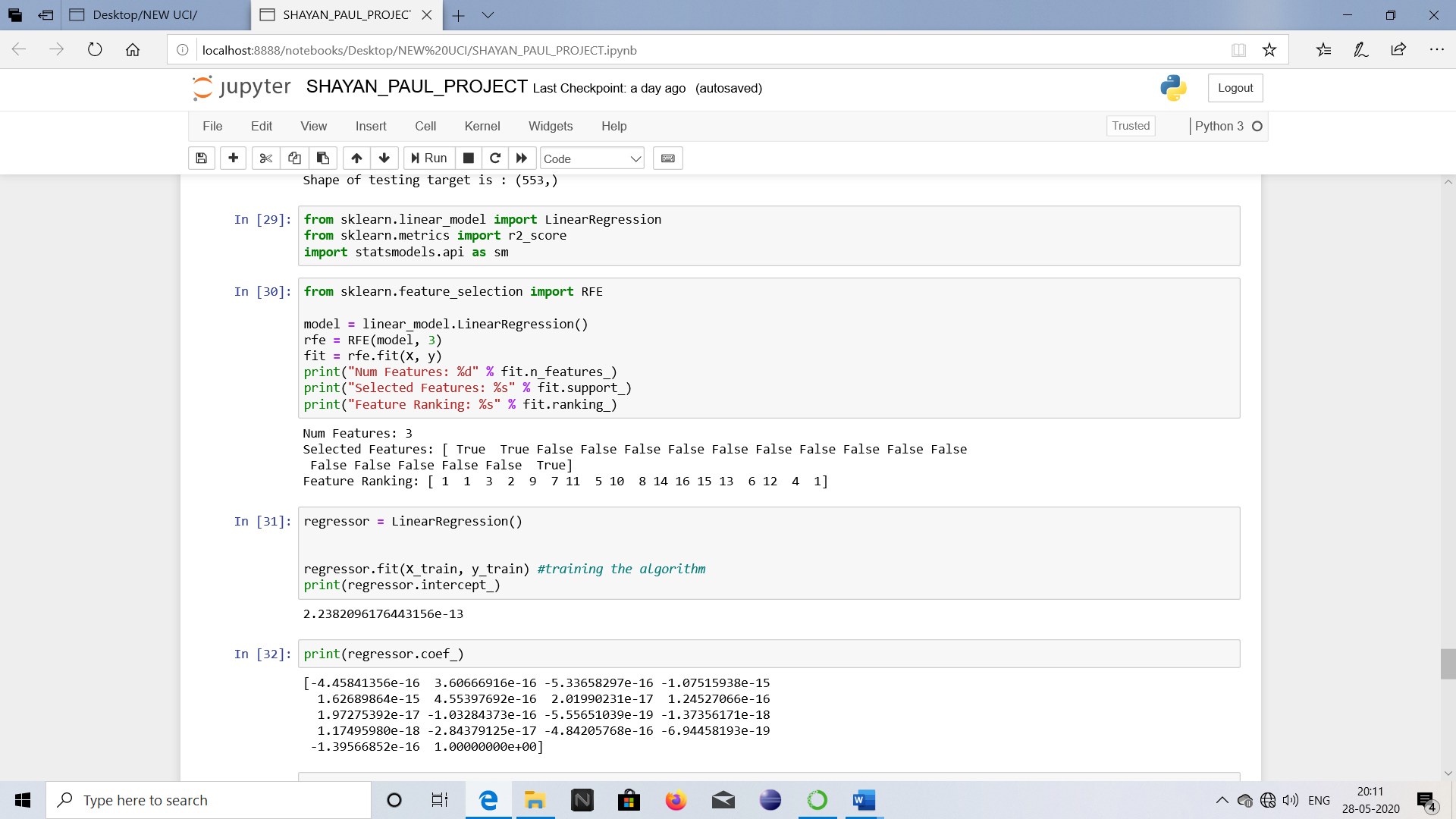


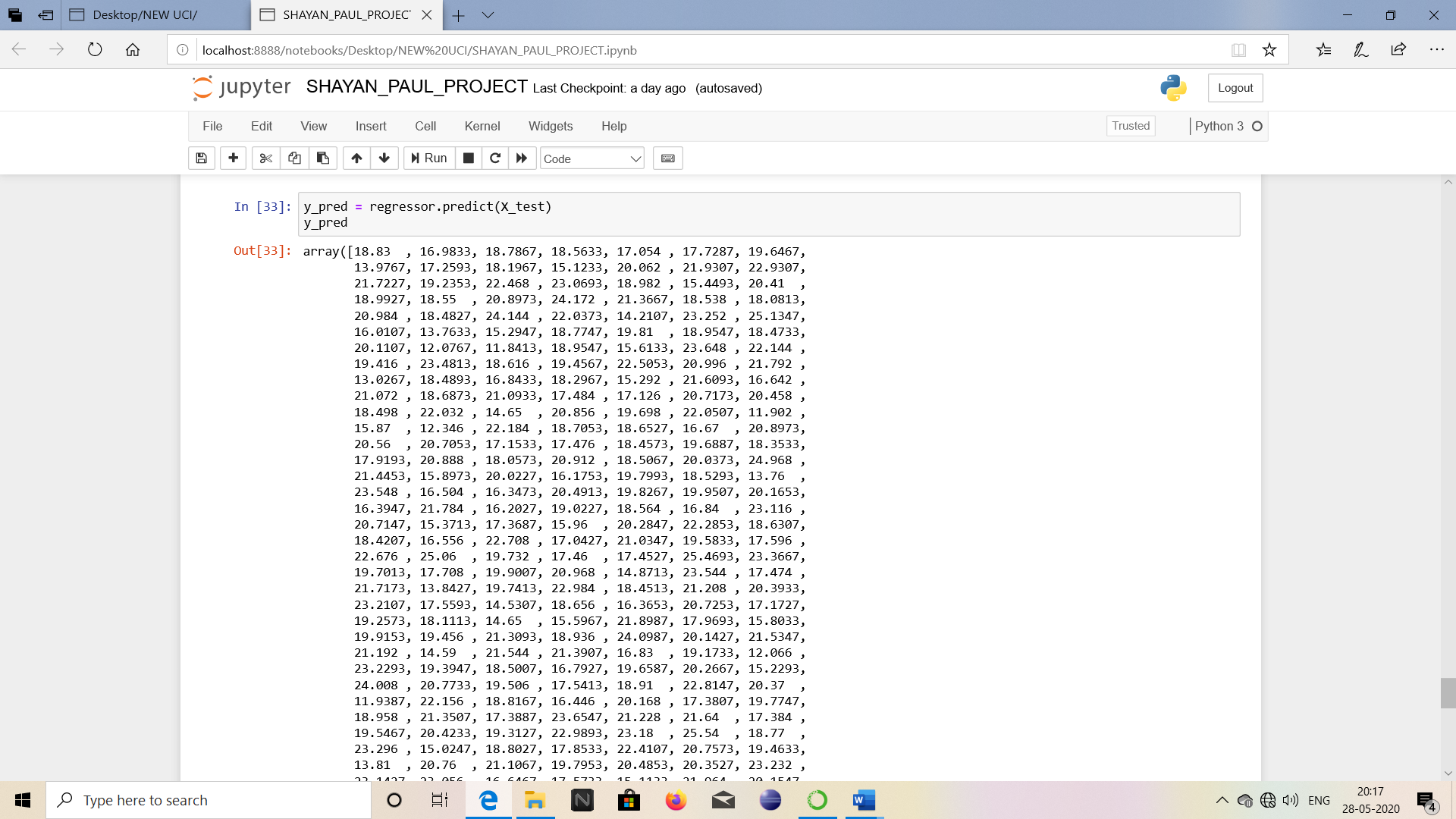


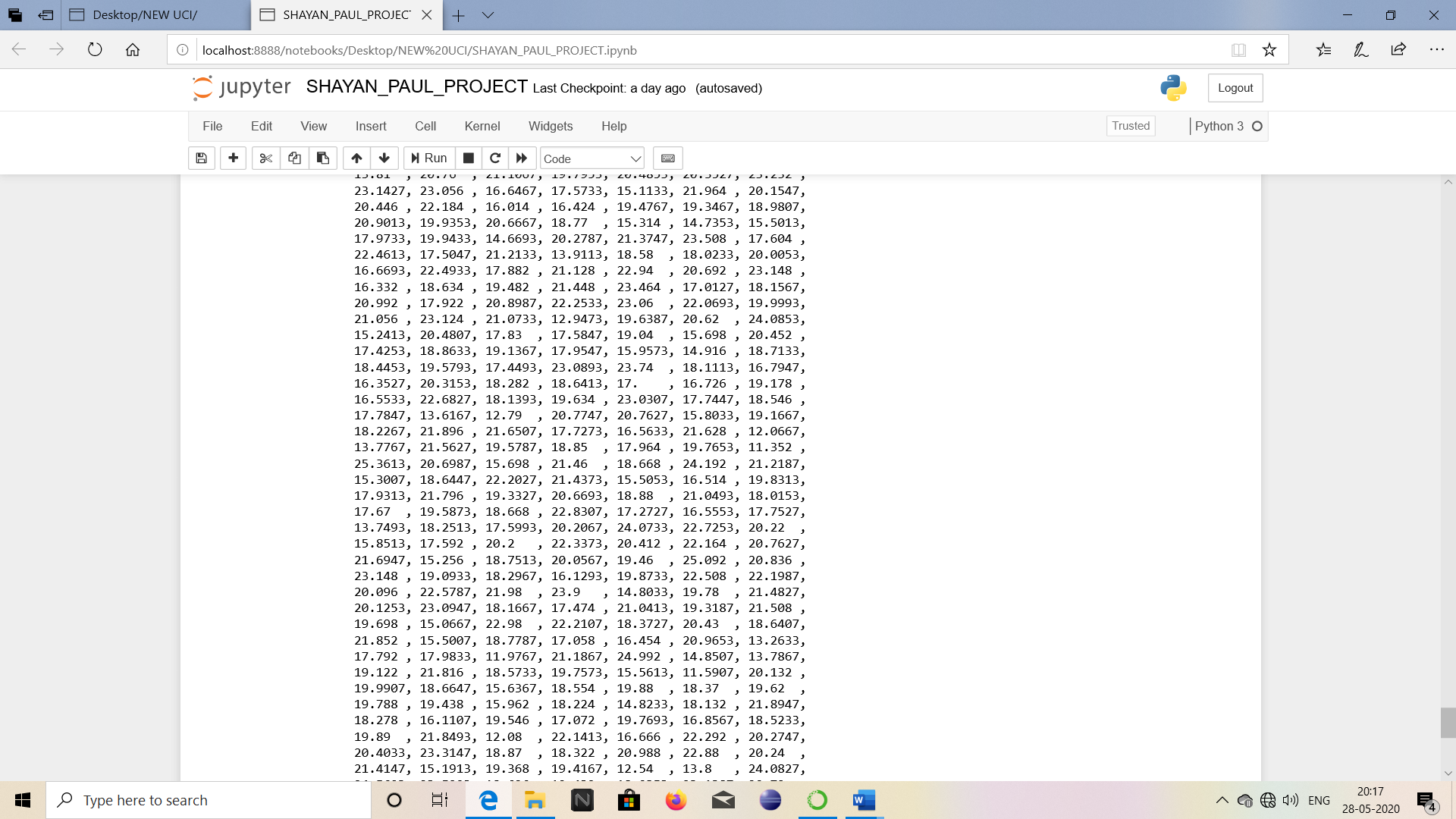


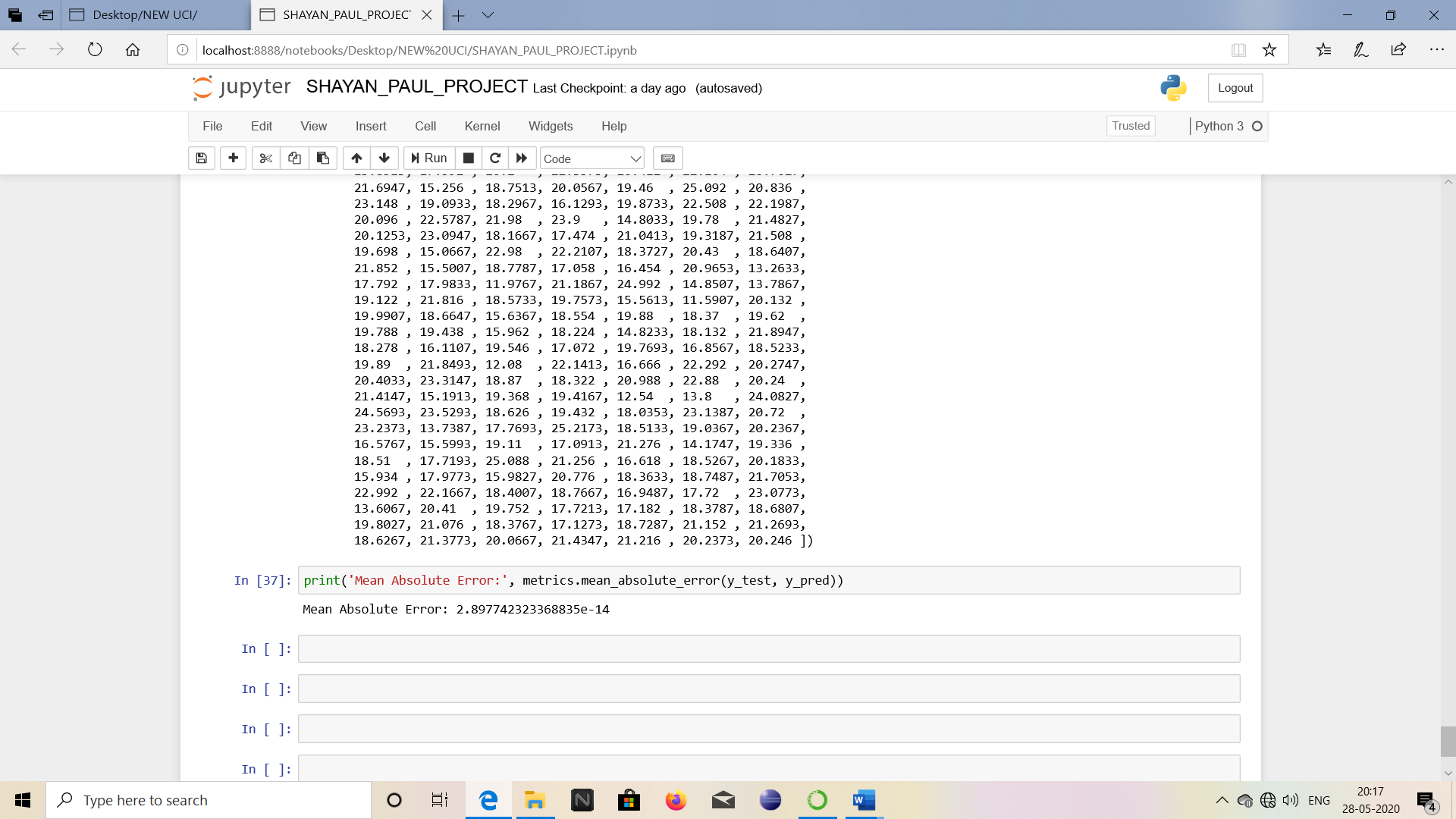
# **4 Model Fitting, Optimizing, and Predicting**

Now that our data has been processed and formmated properly, and that we understand the general data we're working with as well as the trends and associations, we can start to build our model. We can import different classifiers from sklearn.









Here is the final value:

Mean absolute error: 2.897742323368835e-14