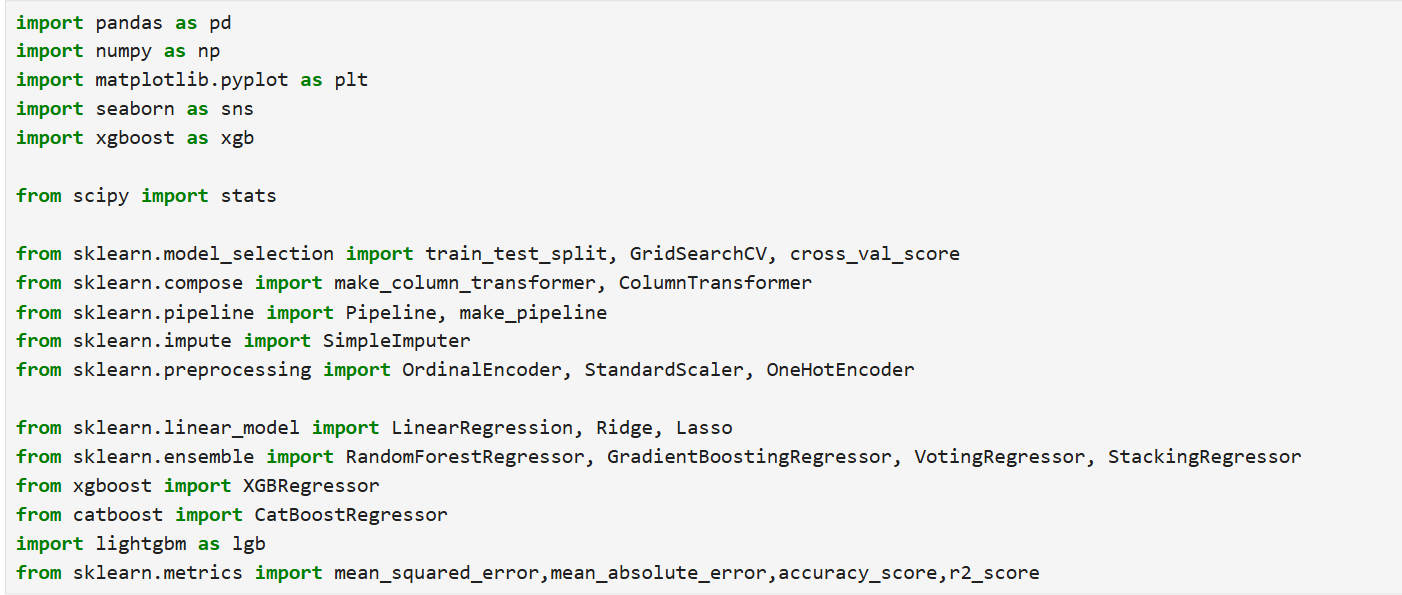
**GARMENT WORKER PRODUCTIVITY PREDICTION**

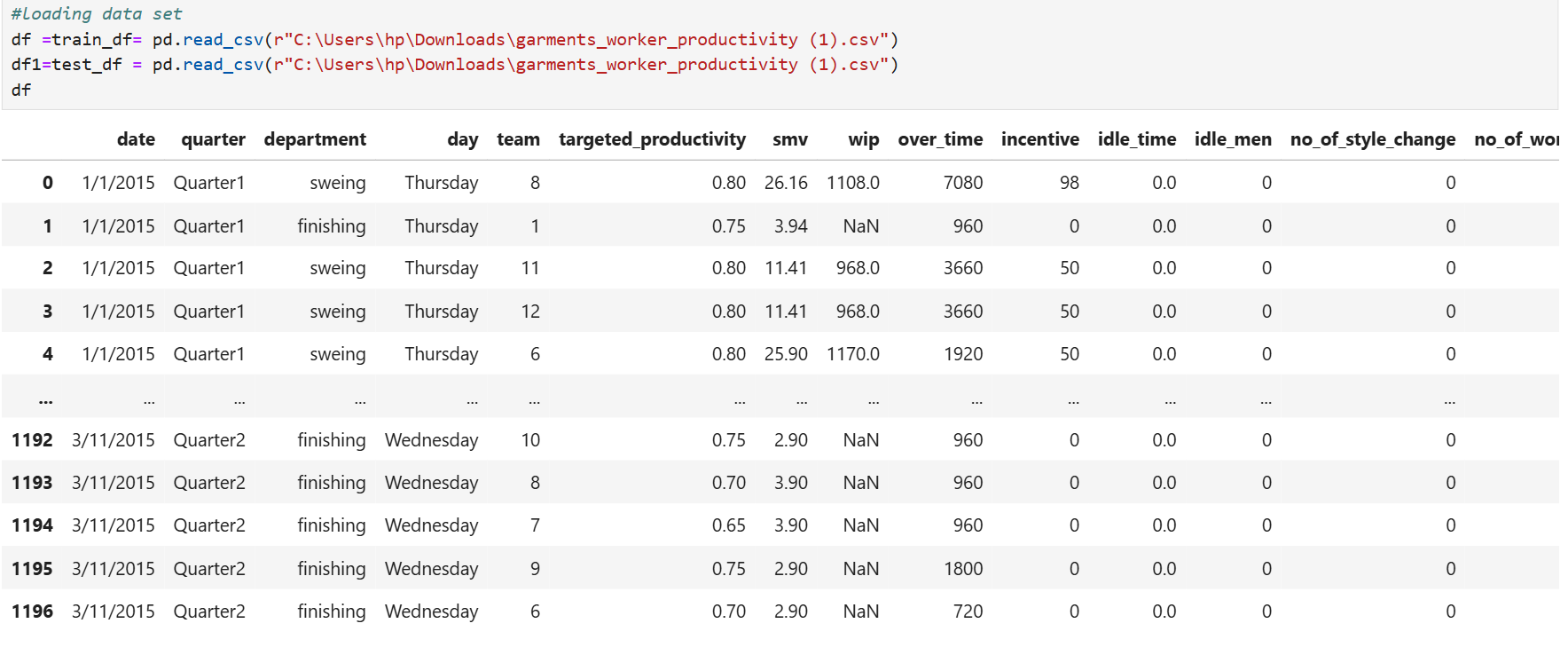
**Introduction**

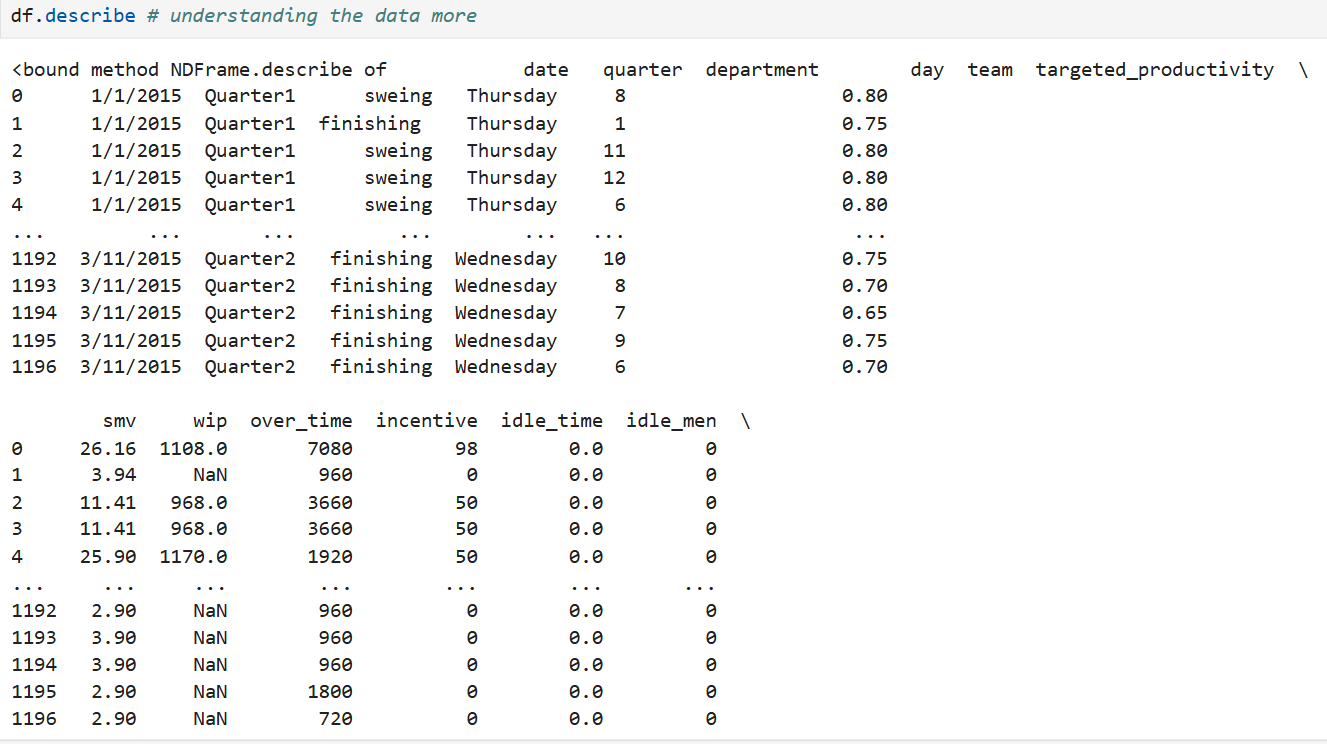
This project focuses on predicting productivity levels in the garment manufacturing industry. Using operational data such as work-in-progress (WIP), overtime, and incentives, you will develop a machine-learning model to forecast productivity. Your work will directly contribute to optimizing resource allocation and improving operational efficiency. Essential tasks include data preparation, analysis, and building models using logistic regression, random forests, and gradient-boosting techniques.

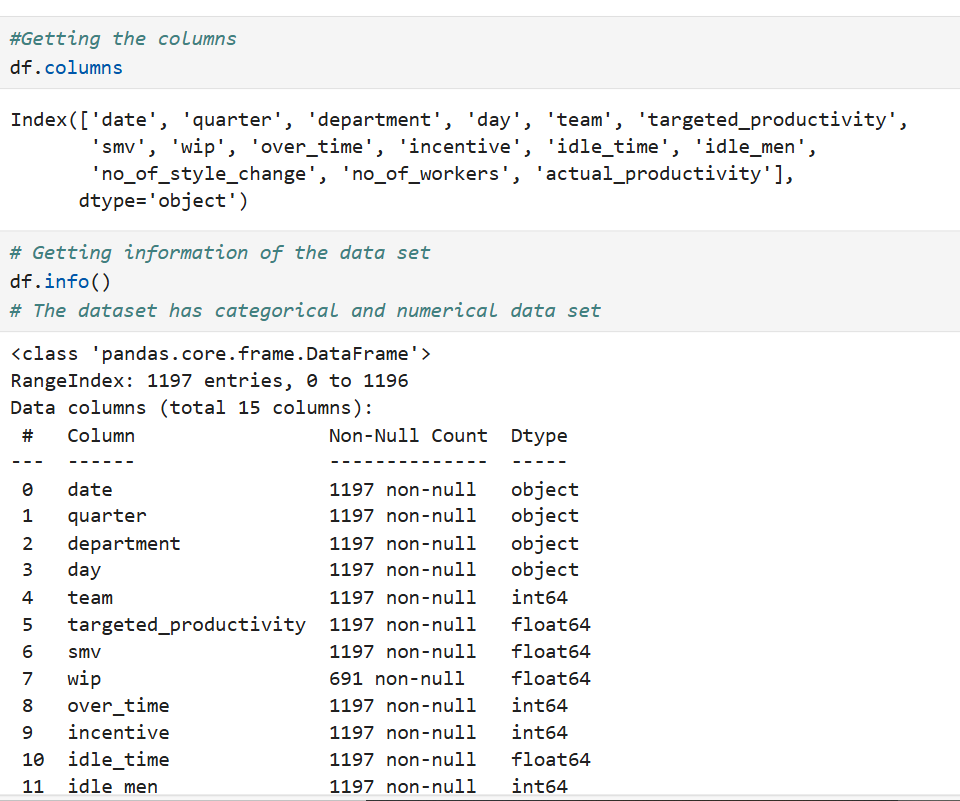
**Loading the libraries**

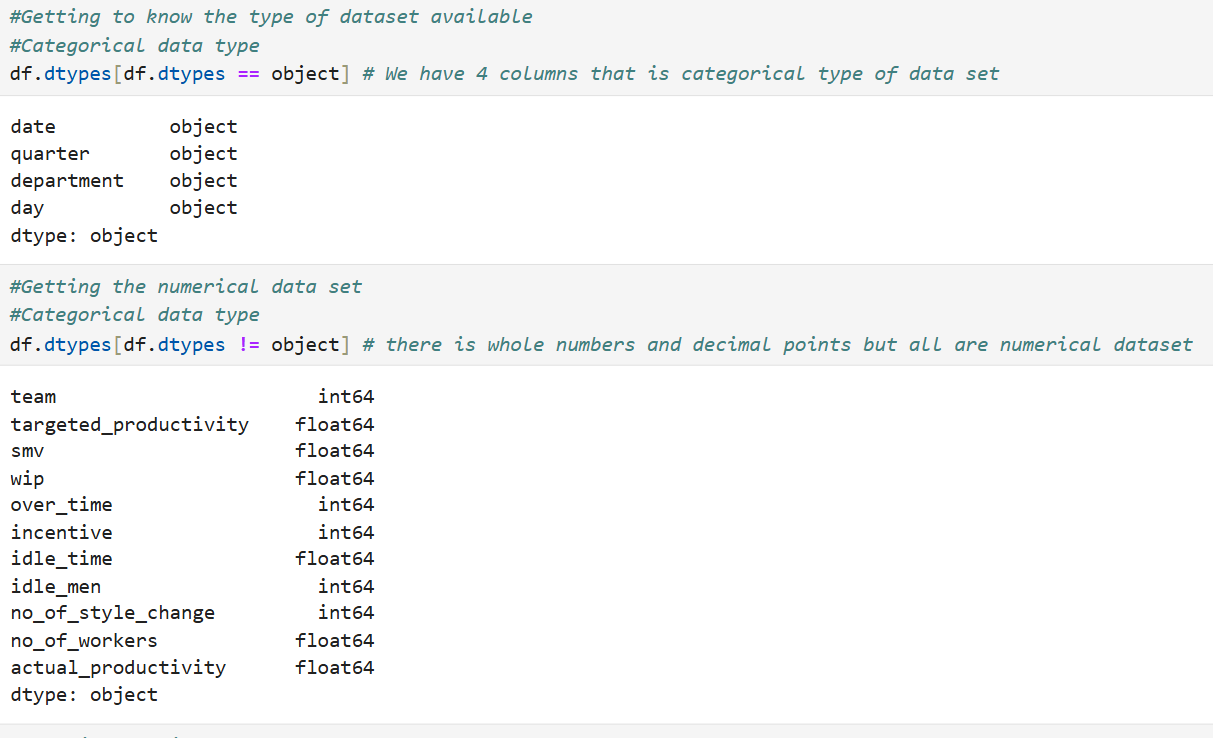
****

**Loading the data set**

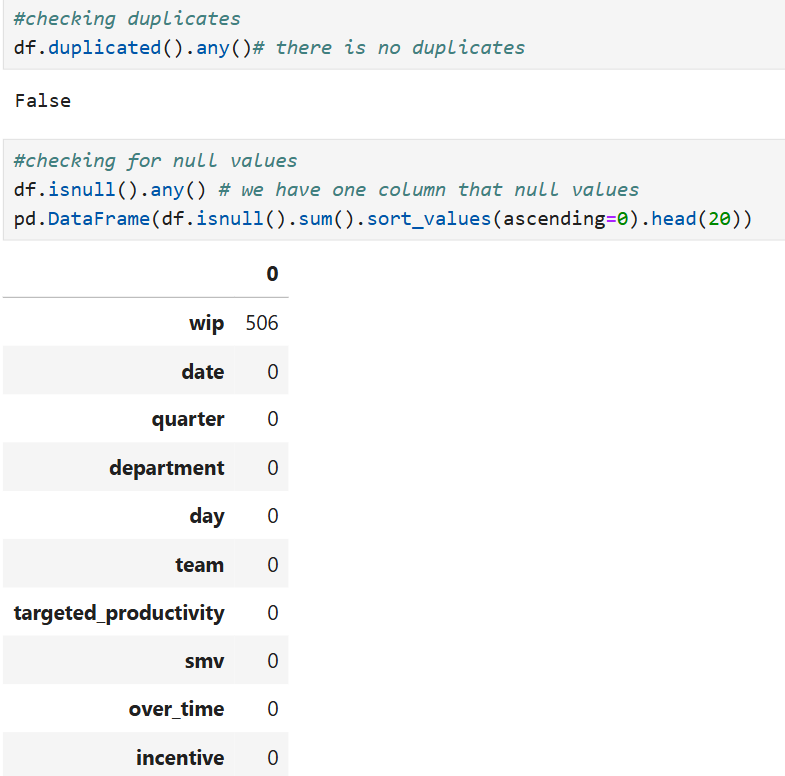
****

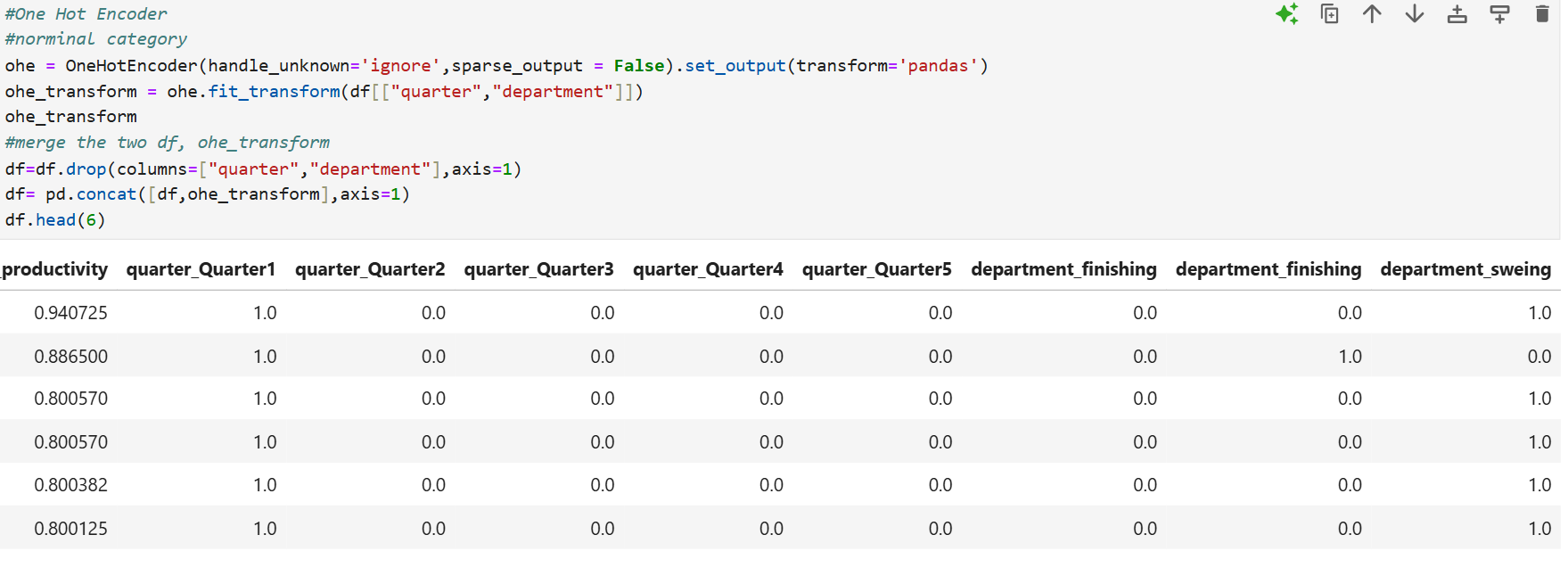
**Getting information about the data**

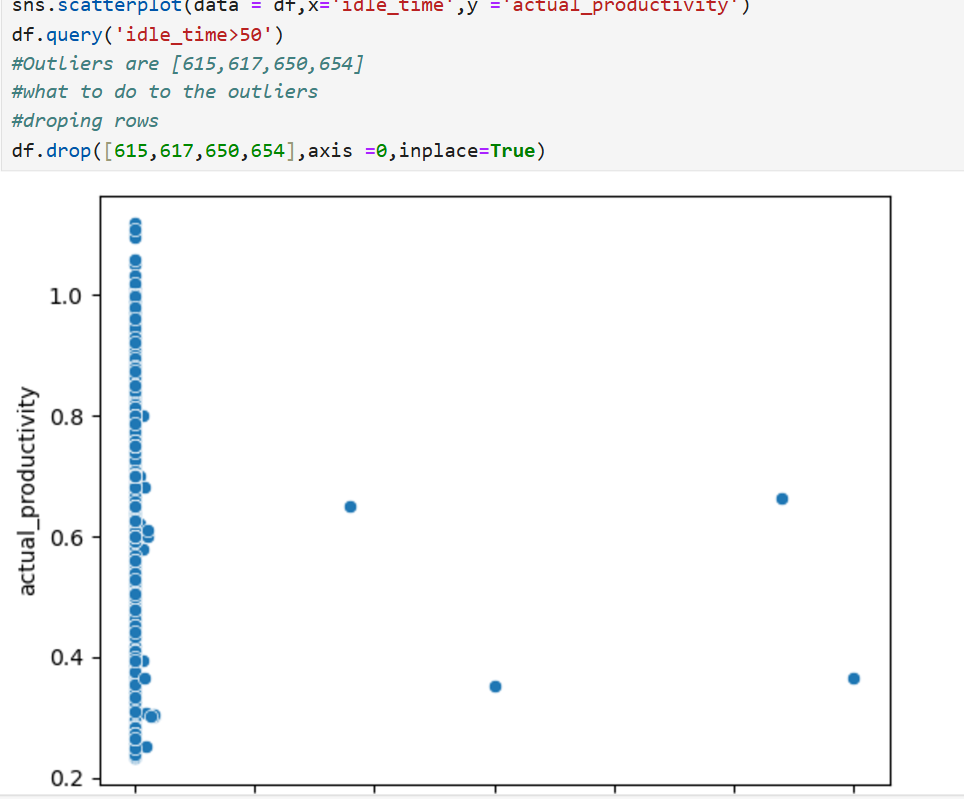
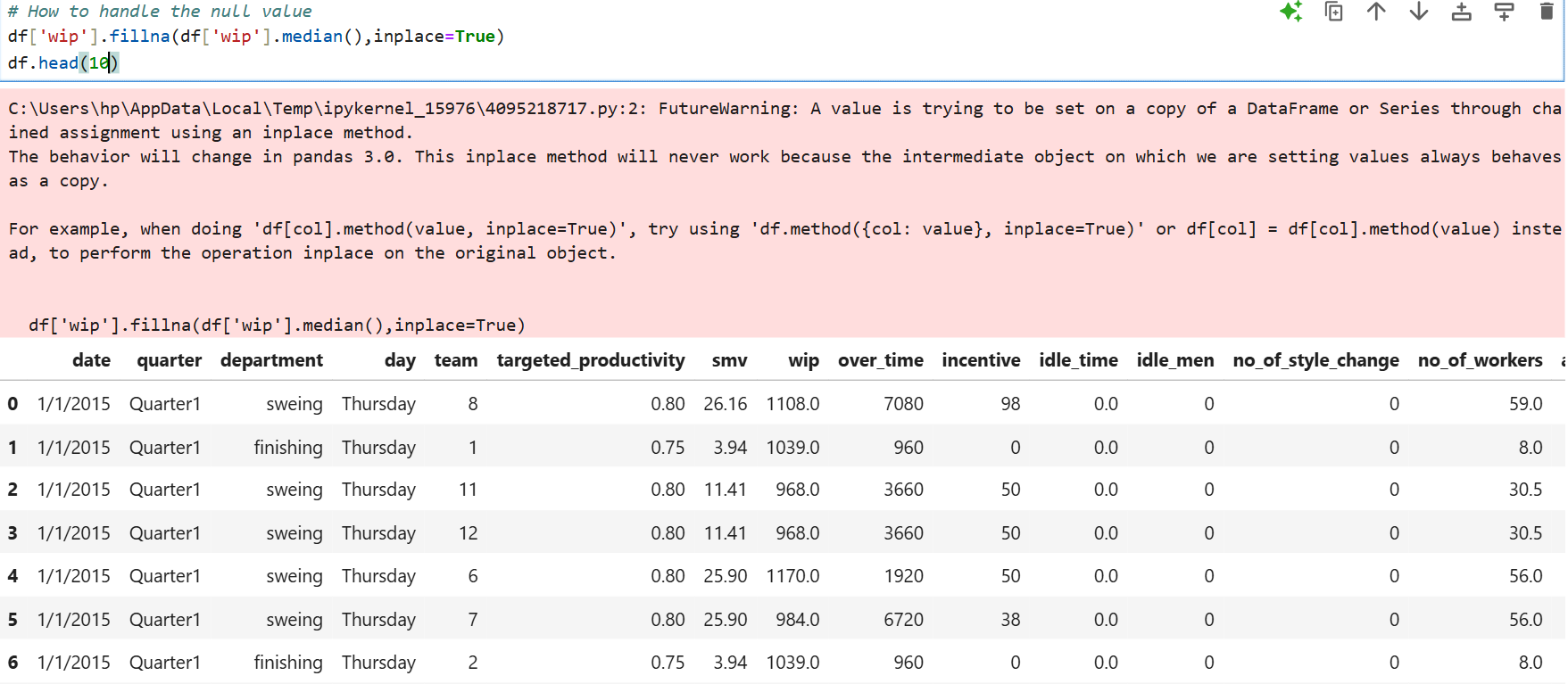
****

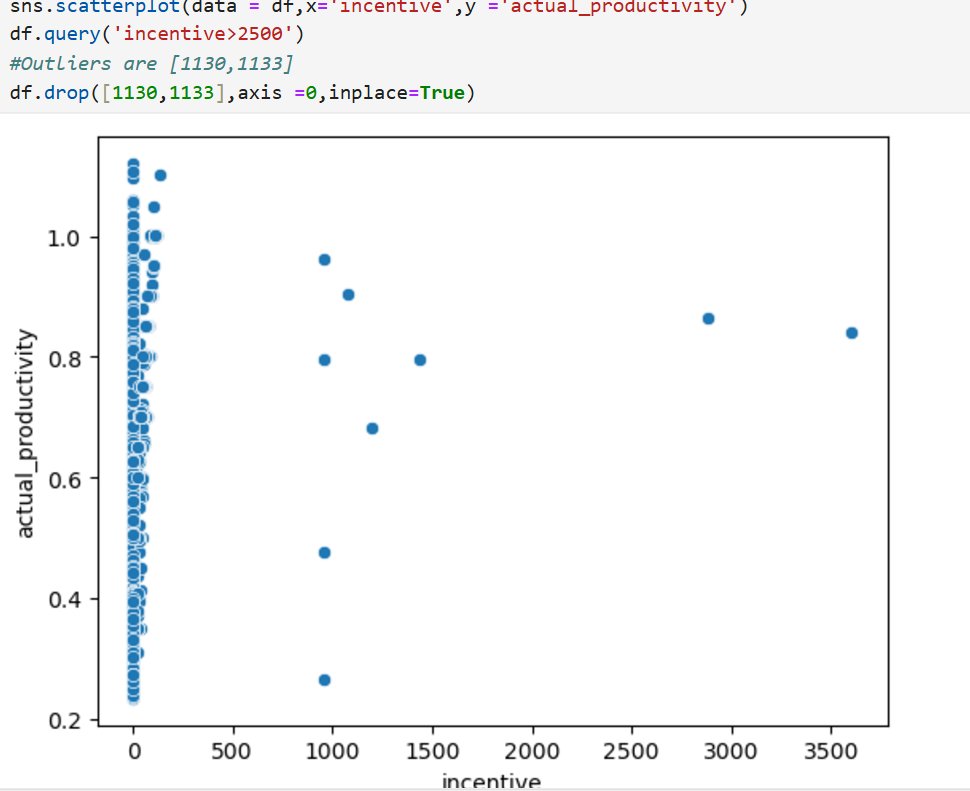
**Type of dataset**

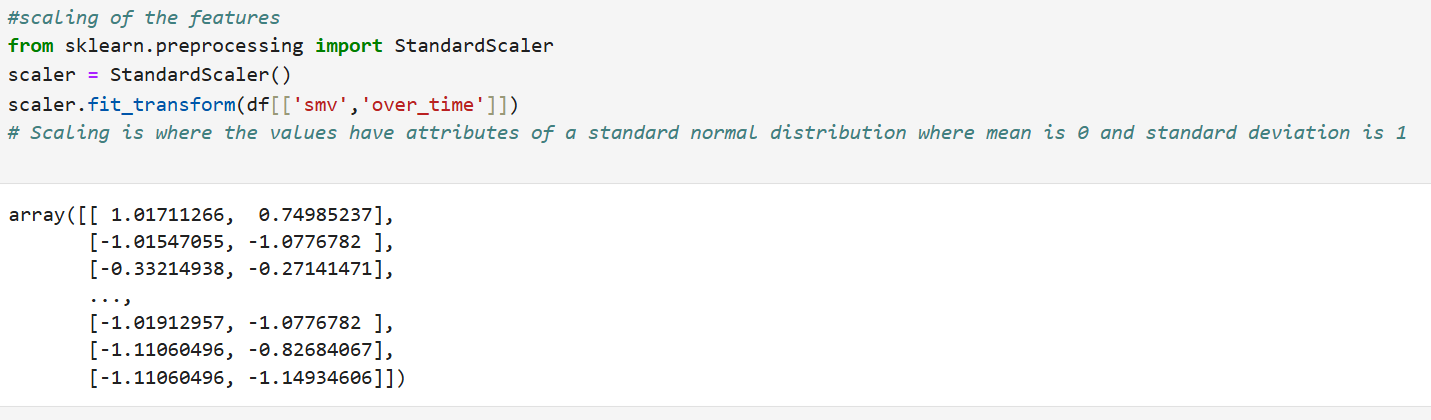
**Cleaning the dataset by looking for duplicates , null values and One Hot Encoding**

****

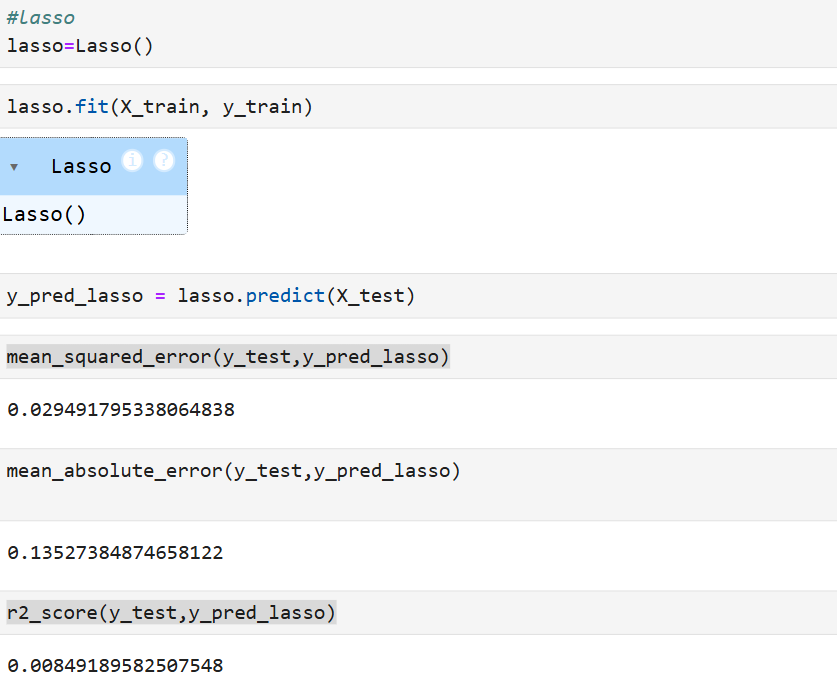
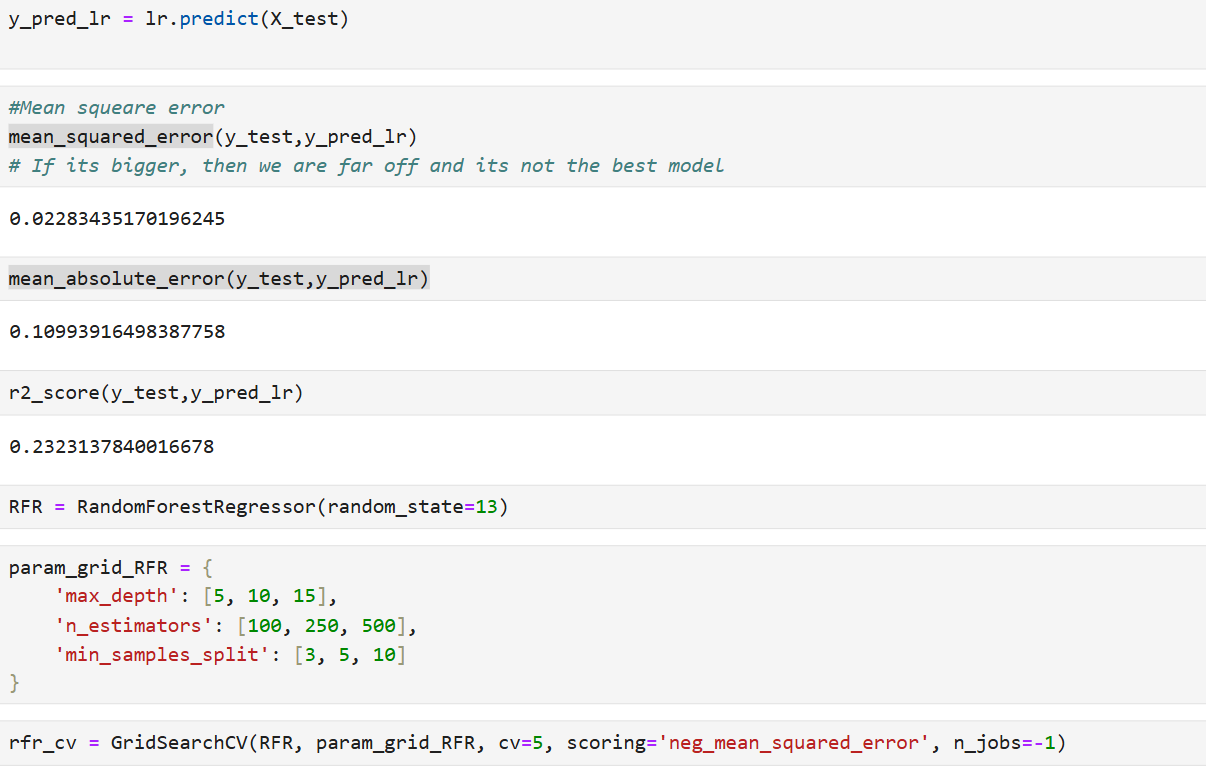
****

**Fixing the null values and checking for outliers using scatter plot**

****

**Standardization**

**Pipeline**

**Heatmap Machine Learning Models**

**Data Understanding and Preprocessing**

This data is in CSV format. I had to import some libraries so that I could perform some analysis of the data. I imported the data into my Jupiter notebook. The data set had 15 columns and 1197 rows, so it's a small dataset.

The data had both categorical and numerical data types. The categorical dataset includes the date, quarter, department, and day, while the numerical data set had the rest of the dataset.

The data provided had many errors, so as a Machine Learning specialist, the first thing you should do to the raw data provided is to clean it.

1. Data cleaning is done by removing any duplicate values that might be there. Luckily for me, this data was not duplicated. The values were also well aligned in the columns, so there was no need to use trim or proper to make the data look presentable.
2. Handling missing values. This is another way of cleaning data values, where you deal with the null values. This is done in different ways depending on the dataset set available. In this dataset, wip had null values, so we used the median to fill the null values. There is another method where you delete the null values. This is where the null values are so many, and the data set is so enormous that deleting the null row will not affect our data.
3. Detecting outliers. The best method to detect outliers is using a boxplot or scatter plot, as you can see the outliers visually. From  idle\_time, the outliers were ID [615,617,650,654], and I dropped them. The outliers were [1130,1133] on incentive, so I did drop them.
4. Scaling . This is where we use another normalization or standardization to run our model. Standardization is best because it makes our data set properties of a typical standard distribution where the mean is usually zero and the standard deviation is 1. I performed standardization on smv 'over\_time,' and the values ranged from 1 to 0. It helps models that are sensitive to features to perform better
5. Encoding. This is done to categorical data that is nominal. I performed it on departs and quarters. The encoding makes the categorical data numerical, and the machine learning models can interpret them.
6. Dropping the data set. I did drop some columns that were not useful in the data set, such as date and day.

**Visualization**

To better understand the relationship between the features and the target value, you need to plot some diagrams. A heat map is the best way to visualize the correlation between all the numerical features and the target.

**Modelling**

A variety of machine models are going to be created to help in the prediction. The models include linear, lasso, random Forest Regressor, and Gradient boosting. We must split the data into X and y to conduct the modeling. Our  X = 'team,' 'targeted\_productivity,' 'smv,' 'over\_time,' 'incentive,' 'idle\_time,' 'idle\_men,' 'no\_of\_style\_change,' 'no\_of\_workers' and y = ['actual\_productivity']. Then, we use this to get the MAE, MSE, and R2