Abstract: The paper herein will analyze the sale of iced products affected by variation of temperature.

Firstly, we will collect the data of the forecast temperature last year and the sale of iced products and then

conduct data compilation and cleansing. Finally, we will set up the mathematical regression analysis model

based on the cleansed data by means of data mining theory. Regression analysis refers to the method of

studying the relationship between independent variable and dependent variable. Linear regression model

that corresponds to the practical situation is proposed in the paper, which is to set up simple linear

regression model based on practical problem and then to implement the following with the help of the latest

and most popular Python3.6. Python3.6 boasts the features of pure object-oriented, platform independence

and concise and elegant language. So we will call the corresponding library function to predict the sale of

iced products according to the variation of temperature, which will provide the foundation for the company

to adjust its production each month, or even each week and each day. As a result, the situation of over-

production can be avoided. Moreover, the other situation as the profit will be affected by the lack of

production since the rise of temperature will also be avoided. So the regression model also has reference

value for the other fields of marketing.

Abstract: The paper herein will analyze the sale of iced products affected by variation of temperature.

Firstly, we will collect the data of the forecast temperature last year and the sale of iced products and then

conduct data compilation and cleansing. Finally, we will set up the mathematical regression analysis model

based on the cleansed data by means of data mining theory. Regression analysis refers to the method of

studying the relationship between independent variable and dependent variable. Linear regression model

that corresponds to the practical situation is proposed in the paper, which is to set up simple linear

regression model based on practical problem and then to implement the following with the help of the latest

and most popular Python3.6. Python3.6 boasts the features of pure object-oriented, platform independence

and concise and elegant language. So we will call the corresponding library function to predict the sale of

iced products according to the variation of temperature, which will provide the foundation for the company

to adjust its production each month, or even each week and each day. As a result, the situation of over-

production can be avoided. Moreover, the other situation as the profit will be affected by the lack of

production since the rise of temperature will also be avoided. So the regression model also has reference

value for the other fields of marketing.

**E-BIKE & E-SCOOTER PRICE PREDICTION MODEL**

*An internship project report submitted by*

# ROHAN BENJAMIN VARGHESE

# &

# VISHWA PRASAD

# *From*

*SSN School of Advanced Career Education*

*Post Graduate Program in Business Analytics & Data Science*

*Jul 2022 – Dec 2023*

*under the supervision of*

**JOYAL FRANCIS**

**

*SSN Incubation Foundation,*

*OMR, Kalavakkam,*

*Chennai, Tamil Nadu 603110.*

1. **Abstract**

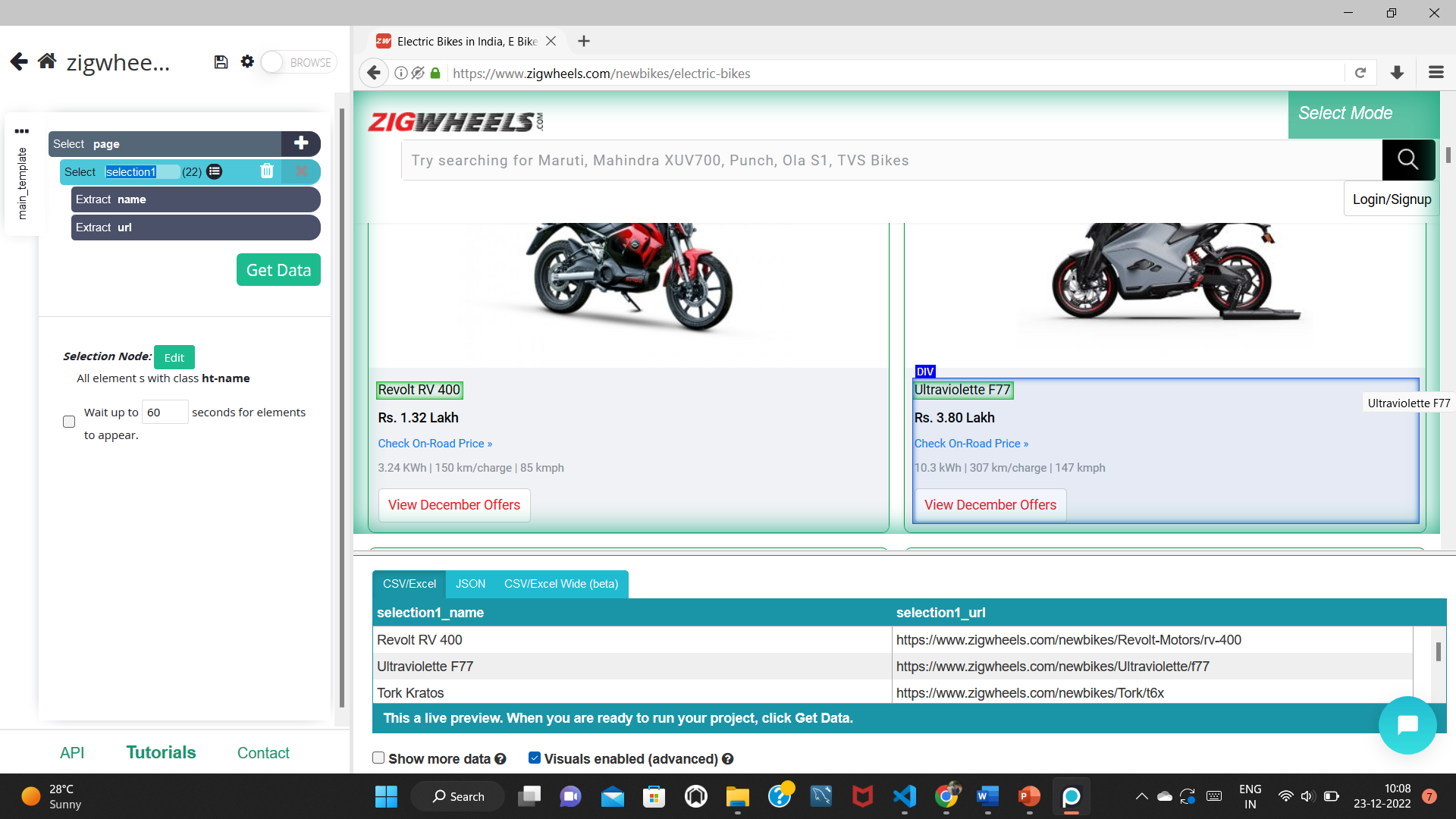
This project aims to collect e-bike, and e-scooter-related data such as speed, battery capacity, charge time, price, etc, from the Indian market and thereby analyze the data to find patterns followed by fitting a model to predict the price of a newly launched product. Firstly, the data is collected from various e-commerce sites using web scrapping techniques. The retrieved data is cleaned and exploratory data analysis is performed with the help of the latest and most popular Python3. Python3 boasts the features of pure object-oriented, platform independence, and numerous libraries such as pandas, NumPy, sci-kit-learn, etc. Finally, the mathematical regression analysis model based on the cleansed data is set up. Regression analysis refers to the method of studying the relationship between the independent variable and the dependent variable. As a result of the implemented model, the project aims at predicting an estimated price for any product that has to be launched in the market.

1. **Introduction:**

In most cities, the dominant mode of transport is still a car or public transport, with a smaller number of travellers walking and cycling. A signiﬁcant percentage of cars in the overall modal share makes cities less sustainable and less liveable. The spatial expansion of cities, with the inevitable increase in congestion on the one hand, and the growing need to travel on the other, has conditioned the emergence of different modes of transport with the idea of balancing user requirements and trafﬁc system supply. The two main representatives of a relatively new wave of mobility are electric bicycles, which have been active on the market for almost two decades, and electric scooters, which have experienced growth in the past few years. These two types of transport belonging to the micro-mobility subsystem, which can be deﬁned as the use of small mobility devices, designed to carry one or two people or “last mile” deliveries. The main advantages of these types, primarily e-scooters, are reﬂected in their characteristics: dimensions, weight, speed, manoeuvrability, ﬂexibility, eco-friendly, etc. This project aims at understanding the performance of the current e-bike and e-scooter market. The data which includes features such as motor type, claimed range, battery capacity, charging time, top speed, and price is extracted by scrapping popular e-bike and e-scooter websites. Once the raw data is cleaned and processed, logical regression is used to predict the price of any new product being launched in the market. The entire workflow is explained below in a sequential manner.

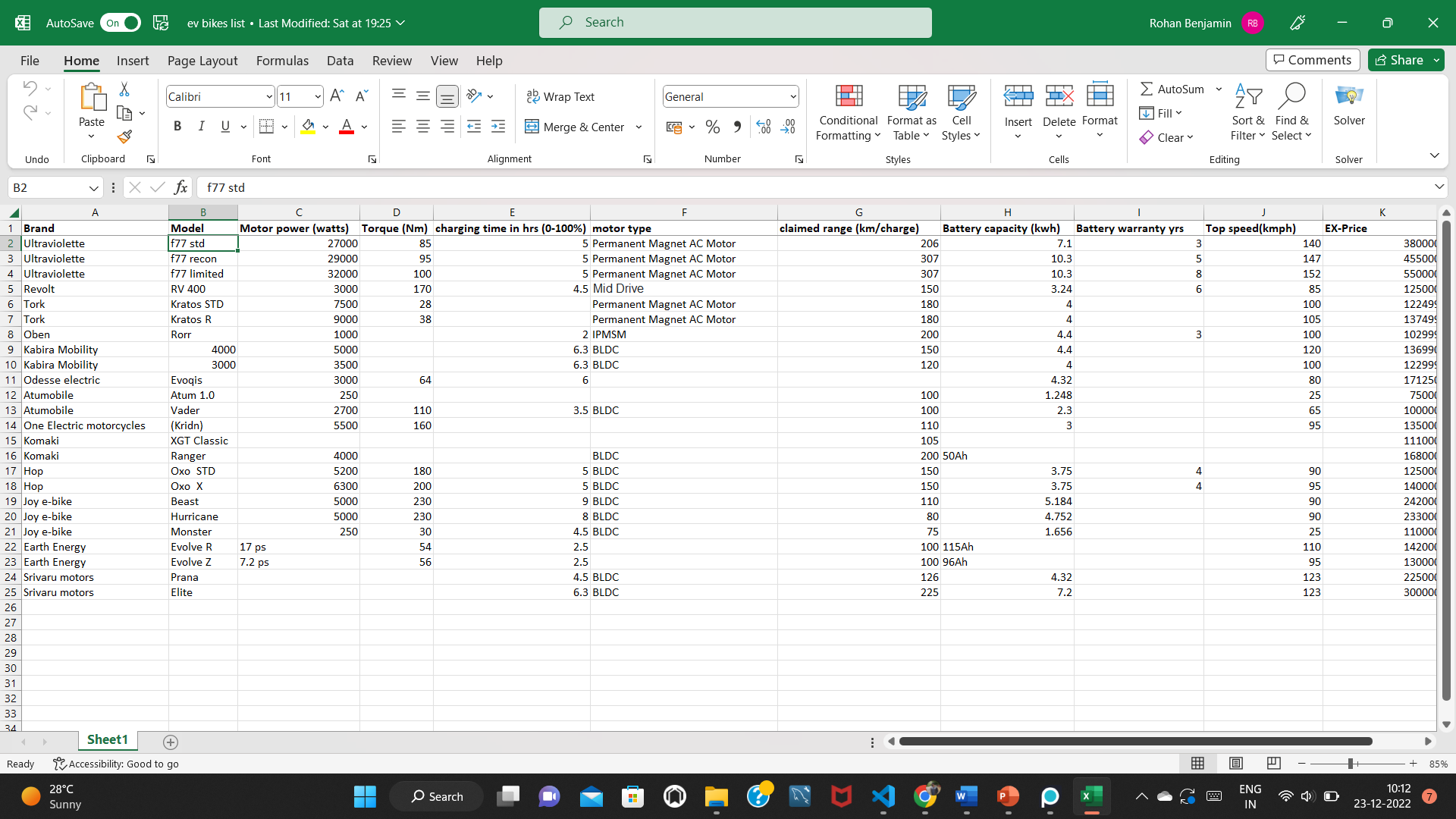
1. **Data collection and Web Scrapping**

Before leveraging data into a successful strategy to meet the requirement the first task is collecting the data. It is extremely difficult to collect performance-related data for specific components of a vehicle as companies do not easily provide such information. Hence the data associated with the overall performance of a vehicle is taken into consideration from various e-commerce sites. The features selected include the name of the vehicle, motor type, claimed range, battery capacity, charging time, and top speed as all these features contribute towards the price of a product. Here data is scrapped from the Zigwheels website which is an automotive and motorcycling website based in India that provides industry news, reviews, and advice to consumers. The data is scrapped using the ParseHub tool which is a free and powerful web scraping tool. Parsehub enables easily getting data from millions of web pages, entering thousands of links, using their REST API, and downloading the extracted data in Excel and JSON. The obtained excel containing the raw data still contains missing values and incomplete data which is treated in the next step.



**Fig.1.1** Parsehub web scraping

Parsehub enables us to select all similar features simultaneously on a page by selecting one of them as seen in Fig.11. The ‘click’ option helps in accessing data from other pages as well. A preview of the data can also be seen at the bottom of the page.

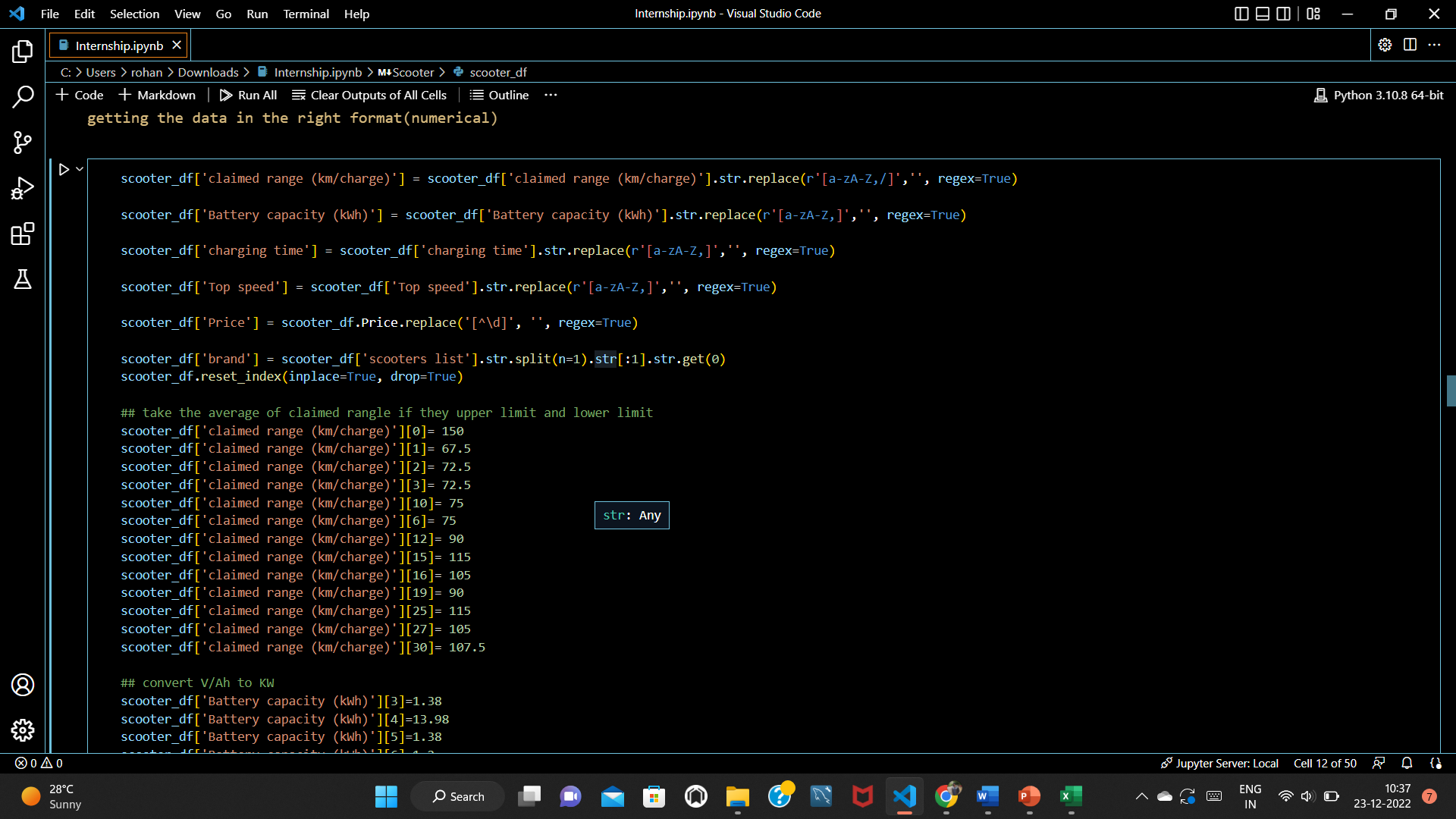


**Fig.1.2** Data retrieved in excel using Parsehub

Once all the necessary features are selected the data can in be downloaded in a CSV format by using the ‘get data’ option followed by clicking on ‘Run’.

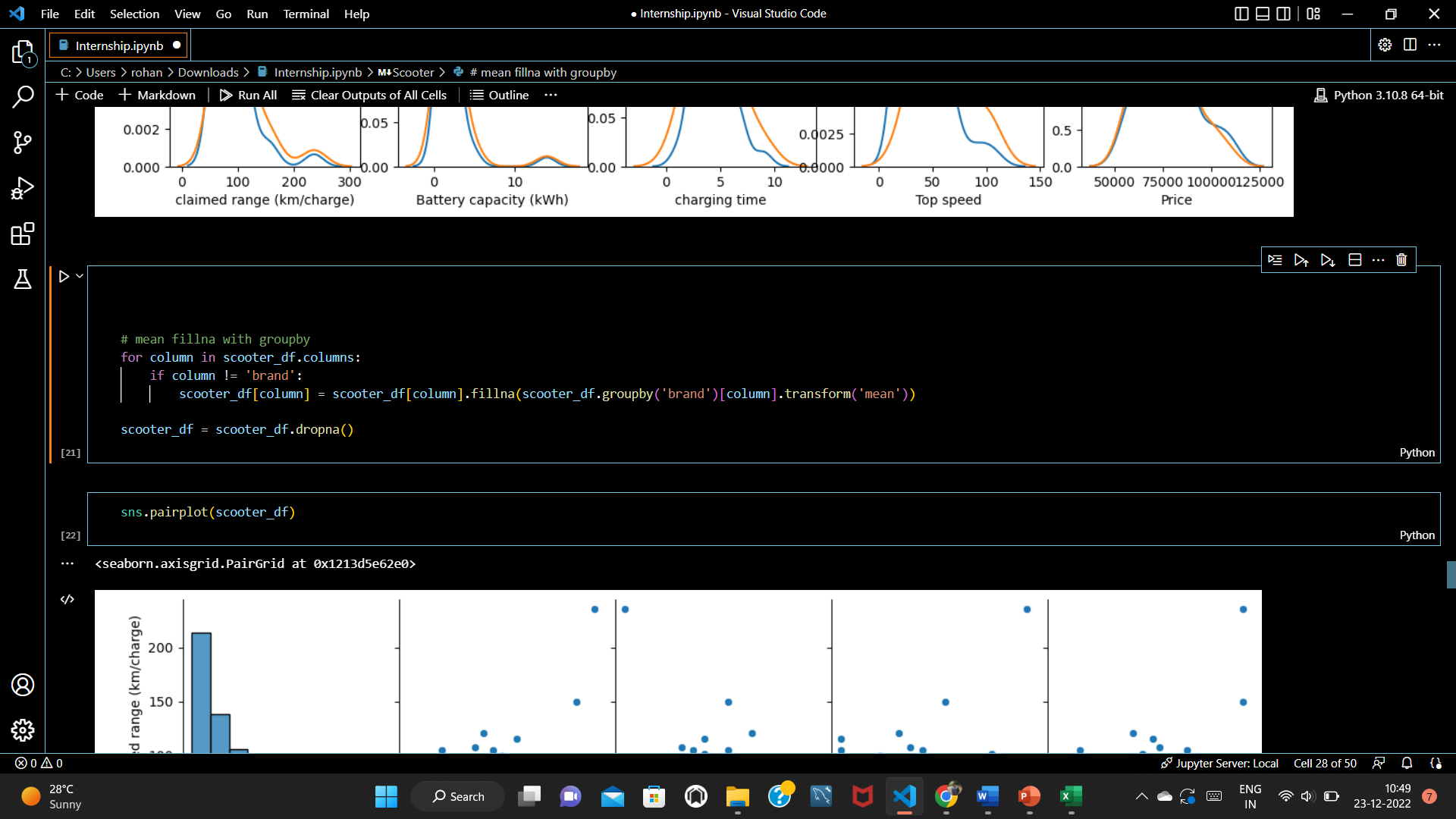
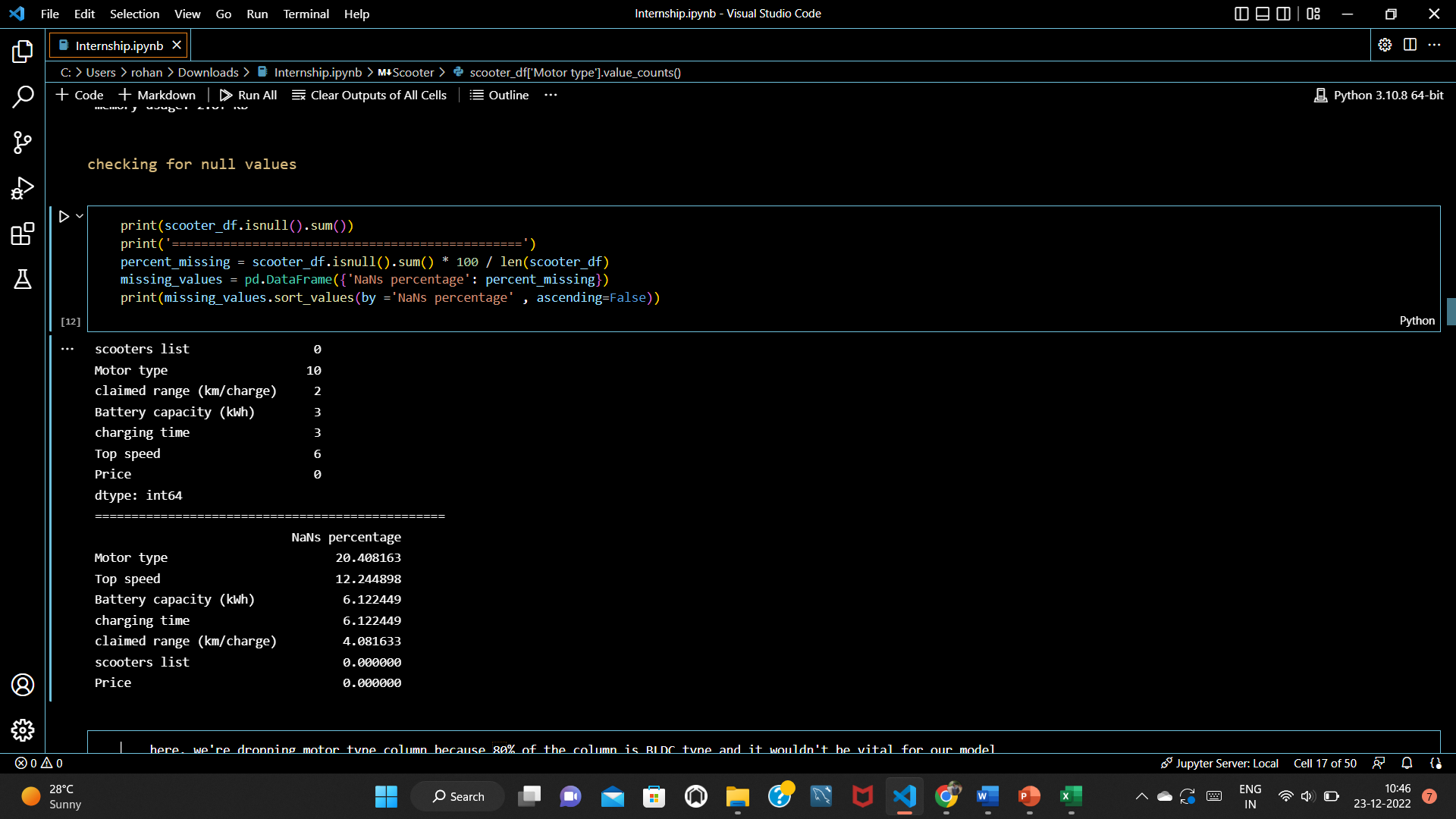
1. **Exploratory data analysis and pre-processing**

The data obtained is still raw and needs to be treated. Python 3 is used here with the pandas and NumPy libraries for this process. By looking at the data it can be seen that some values are in different formats and units. This is corrected and can be seen in Fig.2.1.



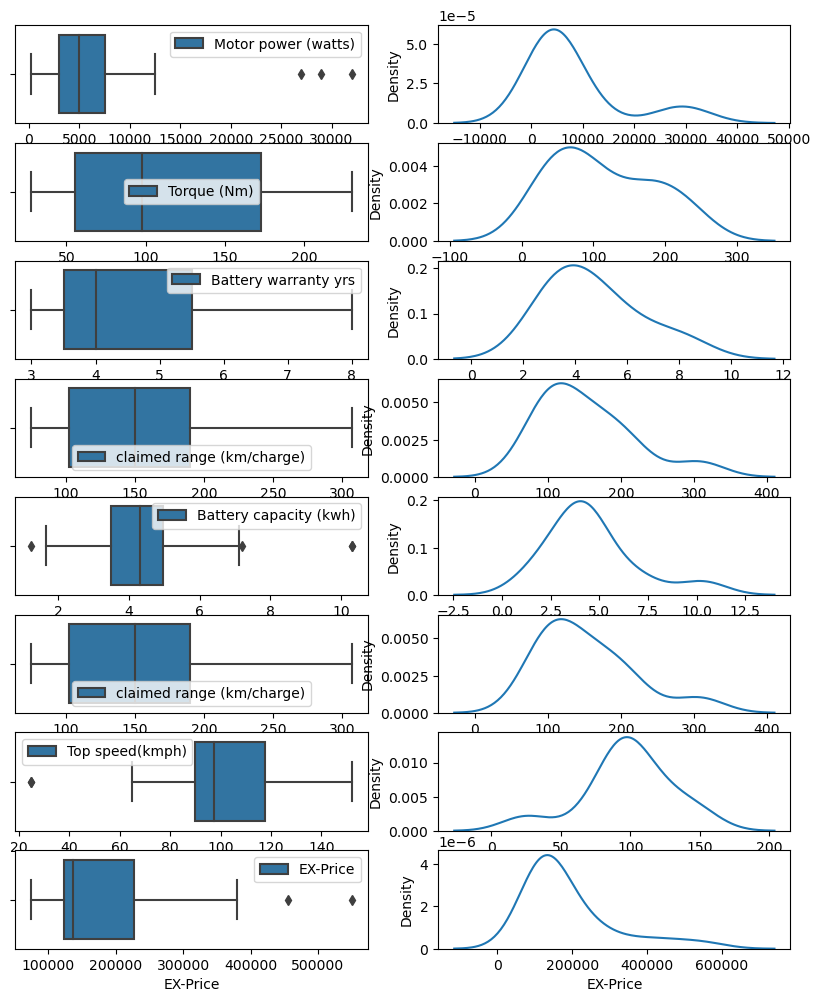
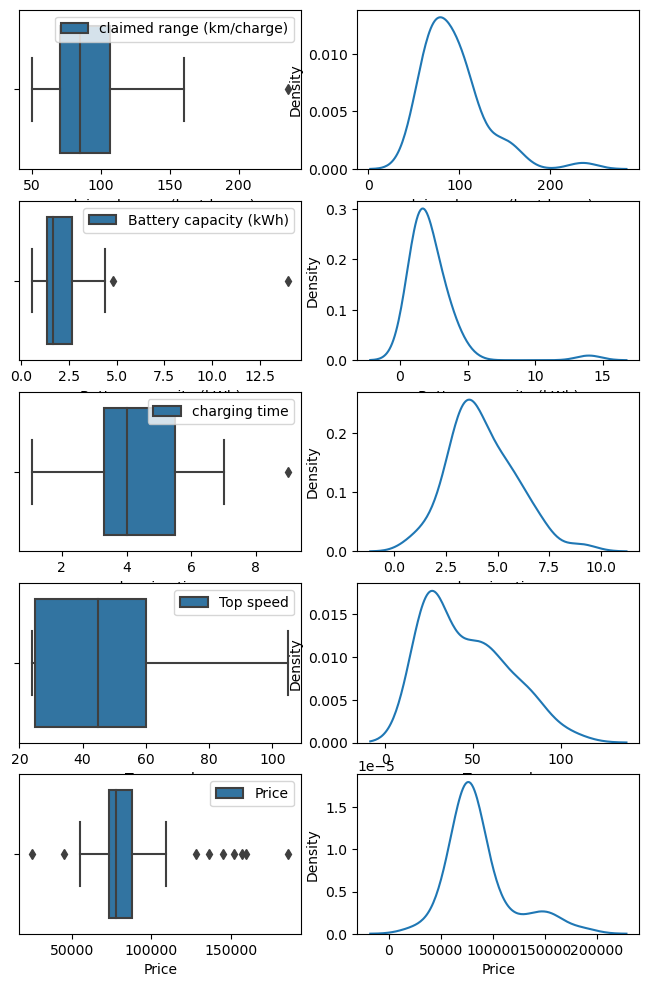
**Fig.2.1** Getting data in the correct format

Once all the data is in the required format, the next step is to describe the data and check for null values and outliers. Outliers are treated by plotting a box plot of the features and removing the outliers if any. Null values are treated by imputing the mean of the column for numerical data by grouping based on the brand as seen in Fig.2.2. The motor type feature is dropped as 80% of the data contains the same value i.e., ‘BLDC’, and imputing the mode for this categorical feature doesn’t make sense.



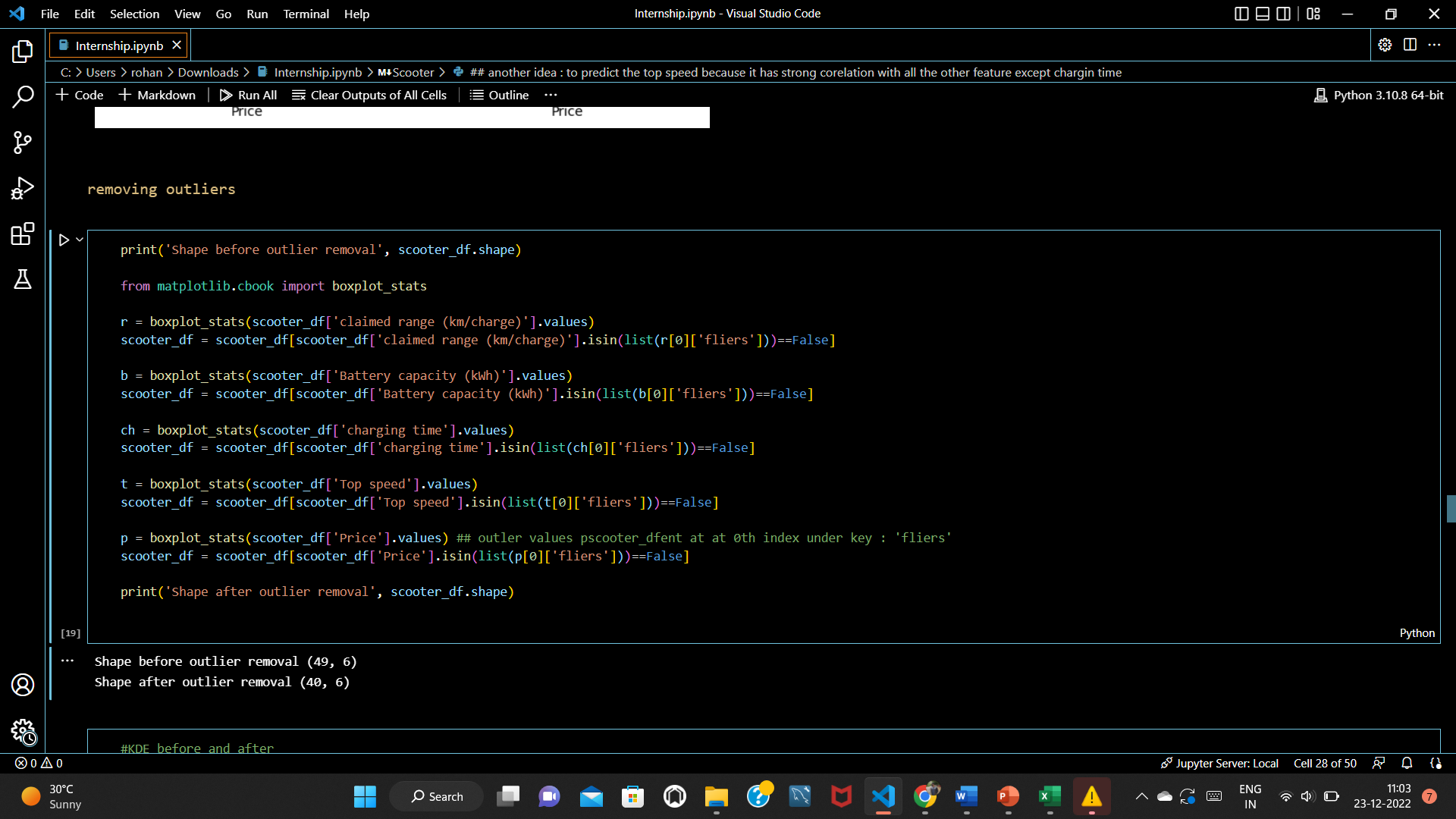
**Fig.2.2** Imputing null values

Once the null values are treated the next step is to treat the outliers. We can identify the outliers by using a box plot which is easily accessible using the seaborn library. Fig.2.3 illustrates the boxplot for each feature and we can see the outliers present.

****

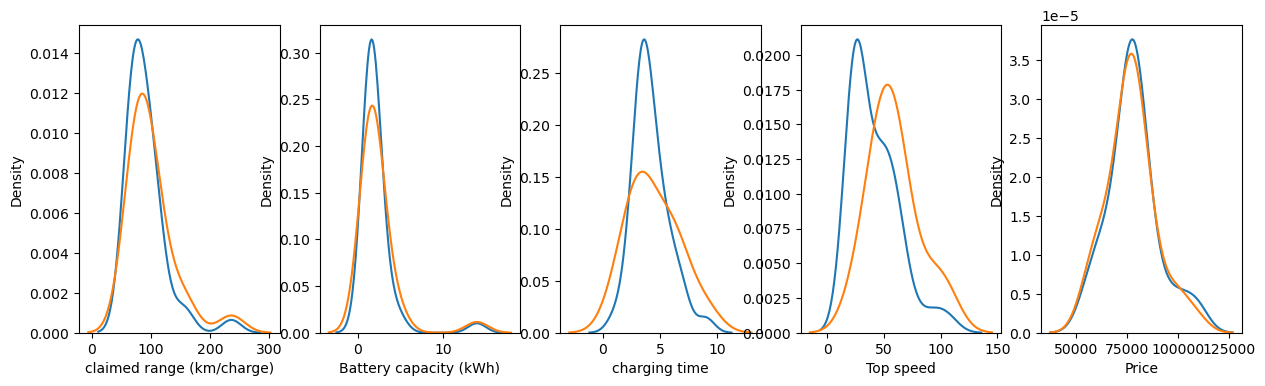
**Fig.2.3** Identifying outliers for e-scooters and e-bikes respectively

Once the outliers are identified they have to be removed and this is visible in Fig.2.4 which contains the code snippet to remove the outliers



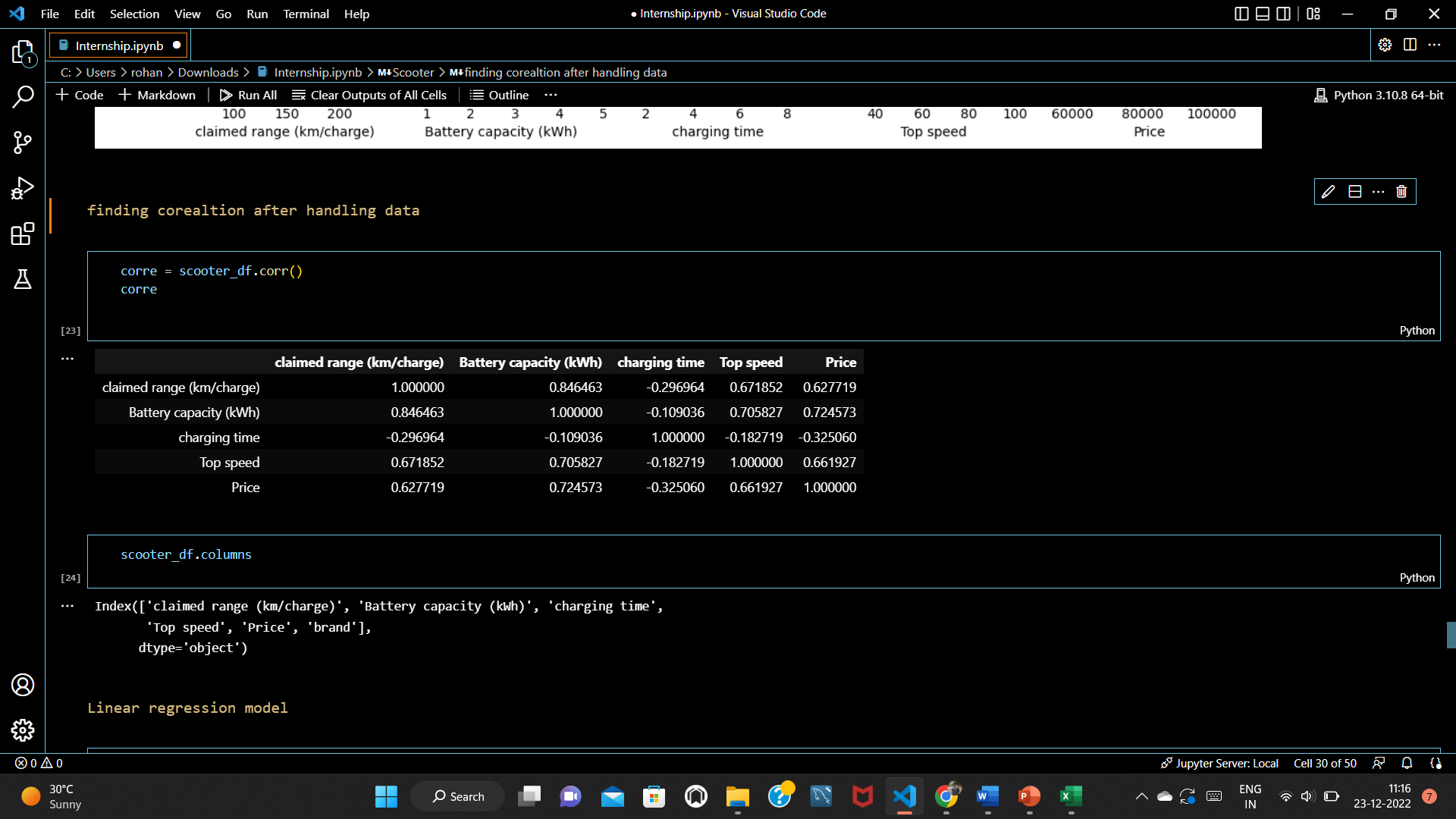
**Fig.2.4** Removing outliers

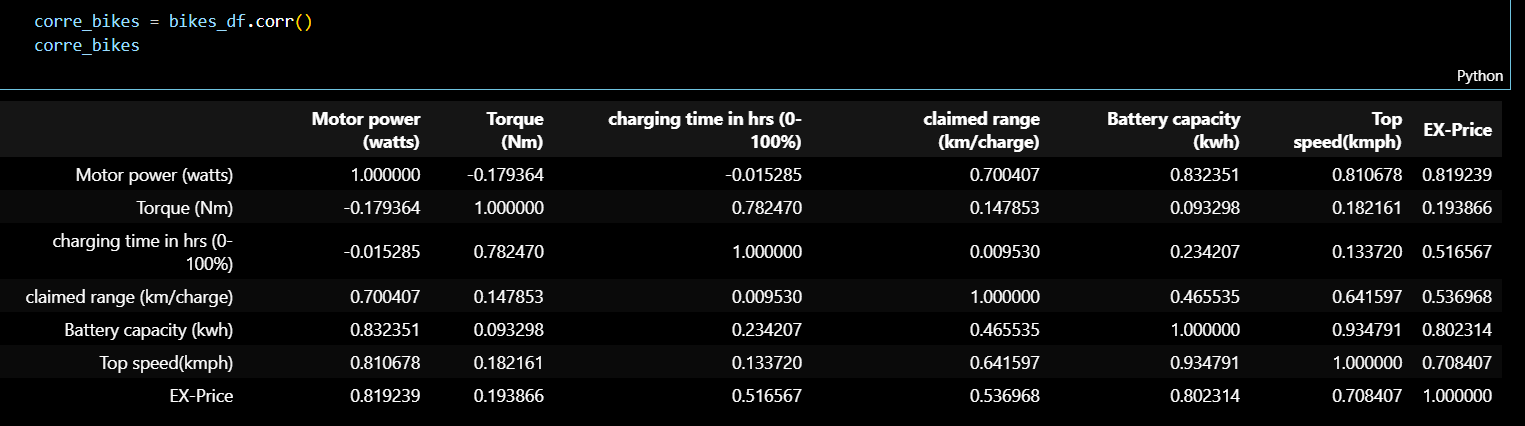
The outliers are removed now and we can see how the data looks now from fig.2.5.

****

**Fig.2.5** Comparison after removing outliers

To select suitable features Python provides an inbuilt correlation function that uses Pearson’s correlation to identify the correlation between different features. here our target value is Price & we can observe that Top Speed, battery capacity, and range are more strongly correlated to price (our target value) than the other features.





**Fig.2.6** Correlation matrix for e-scooters and e-bikes respectively

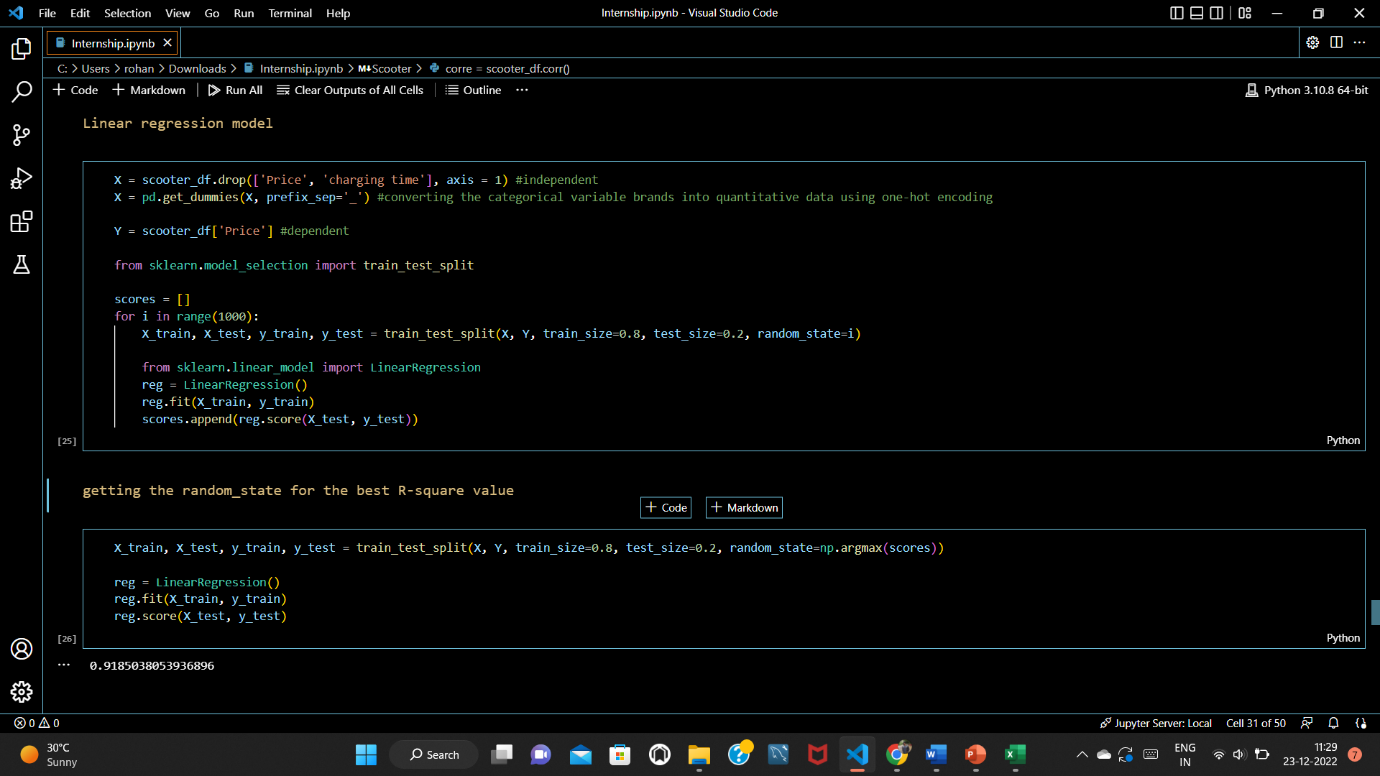
1. **Model fitting**

Linear regression analysis can be divided into simple linear regression and multiple linear regression. The project will mainly analyze the simple linear regression model which is the analysis method of studying the relations between the independent variable and the dependent variable. We will set the model of dependent variable y and the independent variable 𝑥𝑥i (i=1,2,3……) that will influence the variable *y* and predict the development trend of *y* , Simple linear regression model will be expressed as followed:

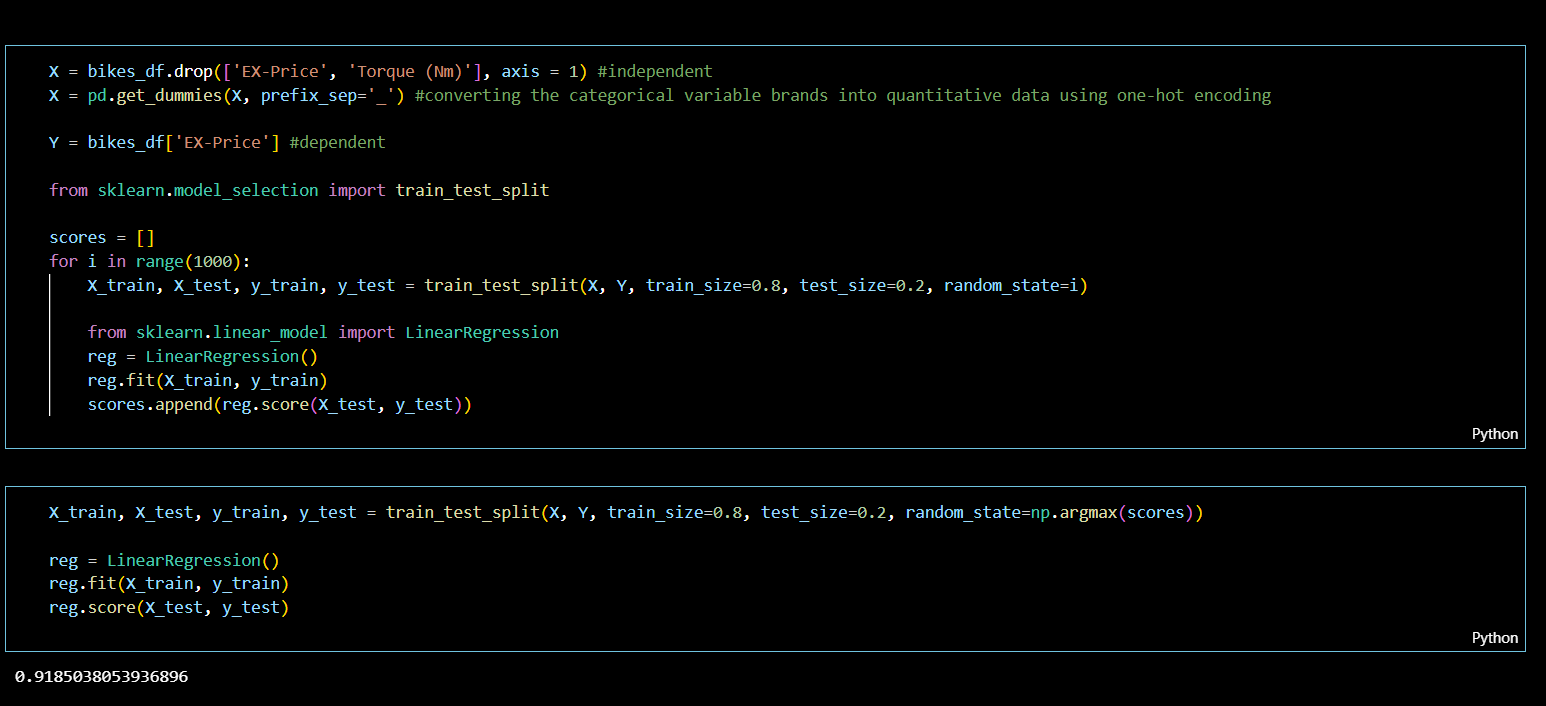
*y*  *a*0  *a* 1 *x*  *e*

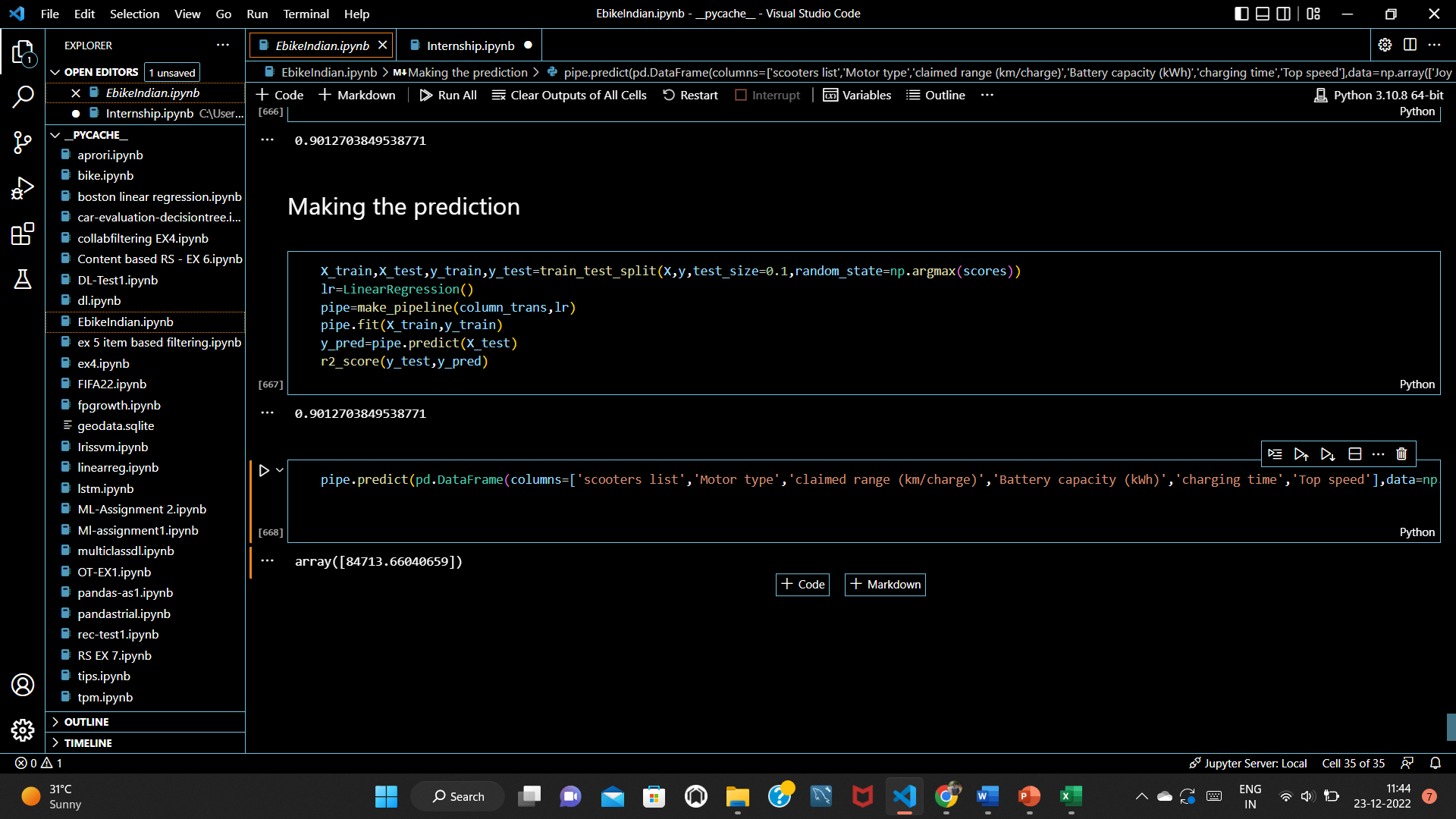
*y* is the dependent variable and *x* is the independent variable. *a*0, the constant term, is the intercept of the regression line on the vertical axis and *a*1 is the regression coefficient that is the slope of the regression line. *e* is the random error that will be used to express the effect of random factors on the dependent variable.

Python’s sklearn library helps in easily splitting the train and test data and also helps in fitting the linear regression model. The below figure shows how the model is fit and accuracy is predicted. The pipeline module from the sklearn library is used to predict as seen in Figure.3.2.



**Fig.3.1** Fitting the model





**Fig.3.2** Making the prediction

The above model has an accuracy of 90% and the Fig.3.2 shows a prediction being made for the given feature values. This process has been implemented for e-scooters and is being implemented for e-bikes as well.

1. **Conclusion**

A linear regression model has been implemented and it can predict the price of newly launched e-scooters with an accuracy of 91.8% for e-scooters and 99.8% for e-bikes with the given features. The same model can be further used for e-cars. We can further create pipelines where users/customers can interact with the API where they can enter the details of the product to get a ballpark of the vehicle’s price