**Introduction**

We will be trying to predict the price of used cars based on their features. As it would help people to decide whether the used car is worth the posted price by different online used-car sites. It would also help people when they plan selling their cars.

**Motivation**

The main motivation behind this project was to help in better decision making in case of buying a used car. A used car has important features like condition, mileage, fuel-type, kilometer-s driven ,these feature’s determine the market price. The price of used cars is very unstable, it changes very frequently depending on its features. There’s also big differences in used car selling platforms. There’s no way of comparing this platforms. This project will help the user’s predict the value of a used car based on its attributes and make the best possible decision in term of buying a used car.

**Problem Statement**

Predicting the price of used cars given the features.

Steps to understand the problem and the approaches used:

Step 1: What is the problem/task?

What is the price of a car given its features.

Step 2: Why does the problem need to be solved?

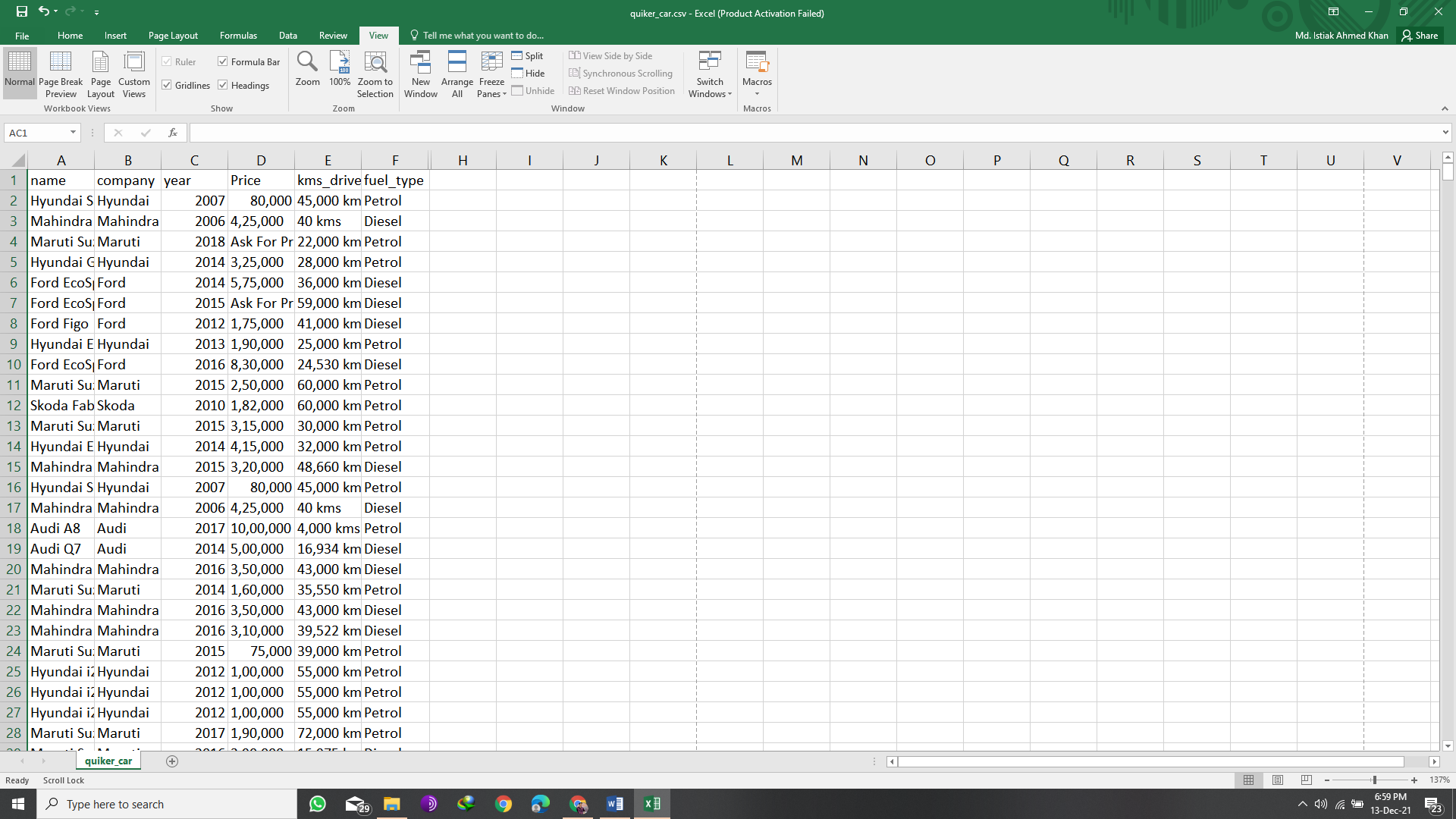
It would help peoples to determine the best prices against different online used-car sites.

Step 3: How would we solve the problem?

The problem is solved using machine learning and data mining techniques/models.

**Data Set**

The dataset that is uses for this project is taken from Quiker Cars. Quiker cars is a online car selling platform. The dataset has 6 rows and 892 columns . Initially data types are object and some of them has nan values.



**Fig: Raw Dataset**

**Data Visualization**

Before cleaning

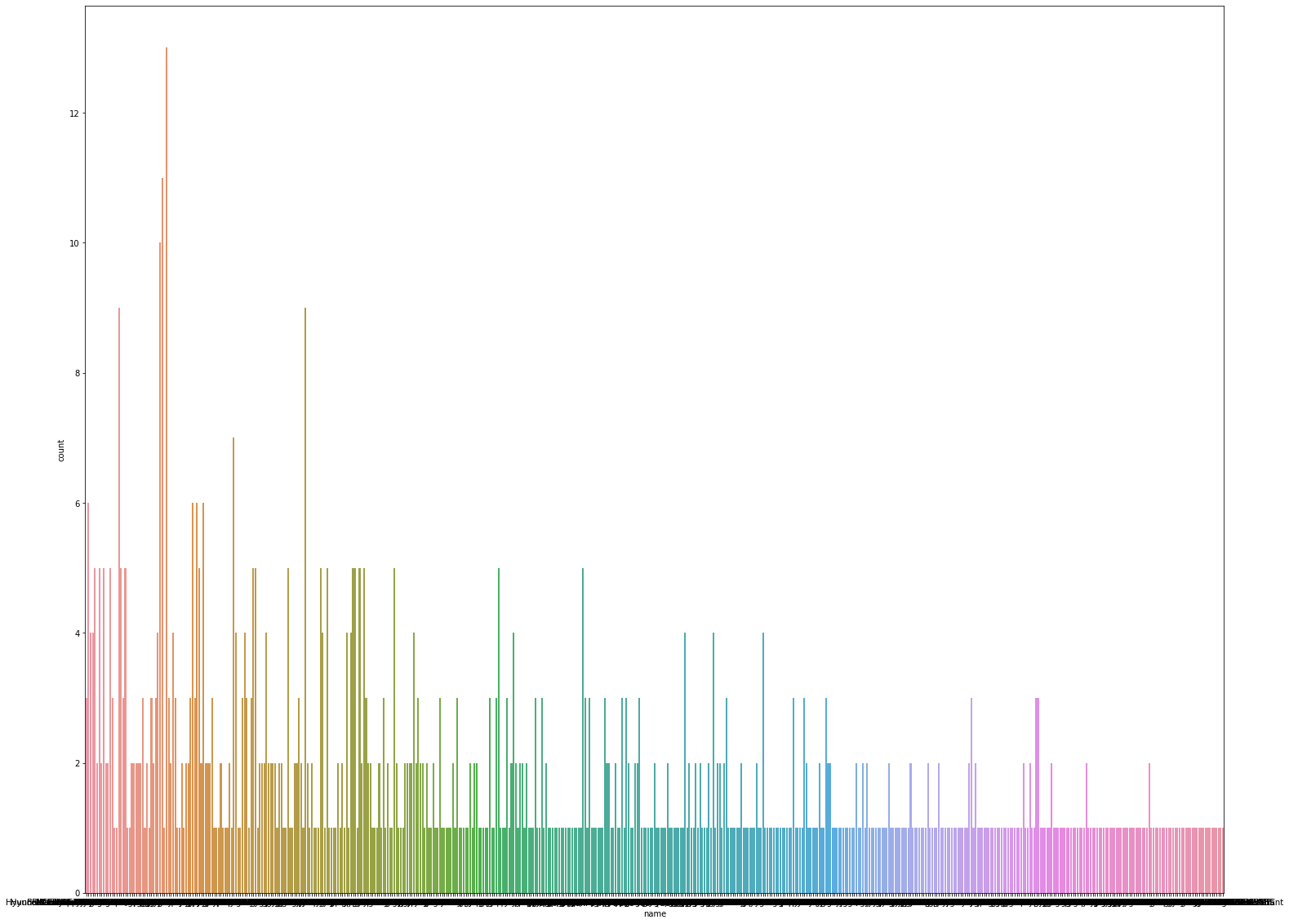


Fig: Raw Name

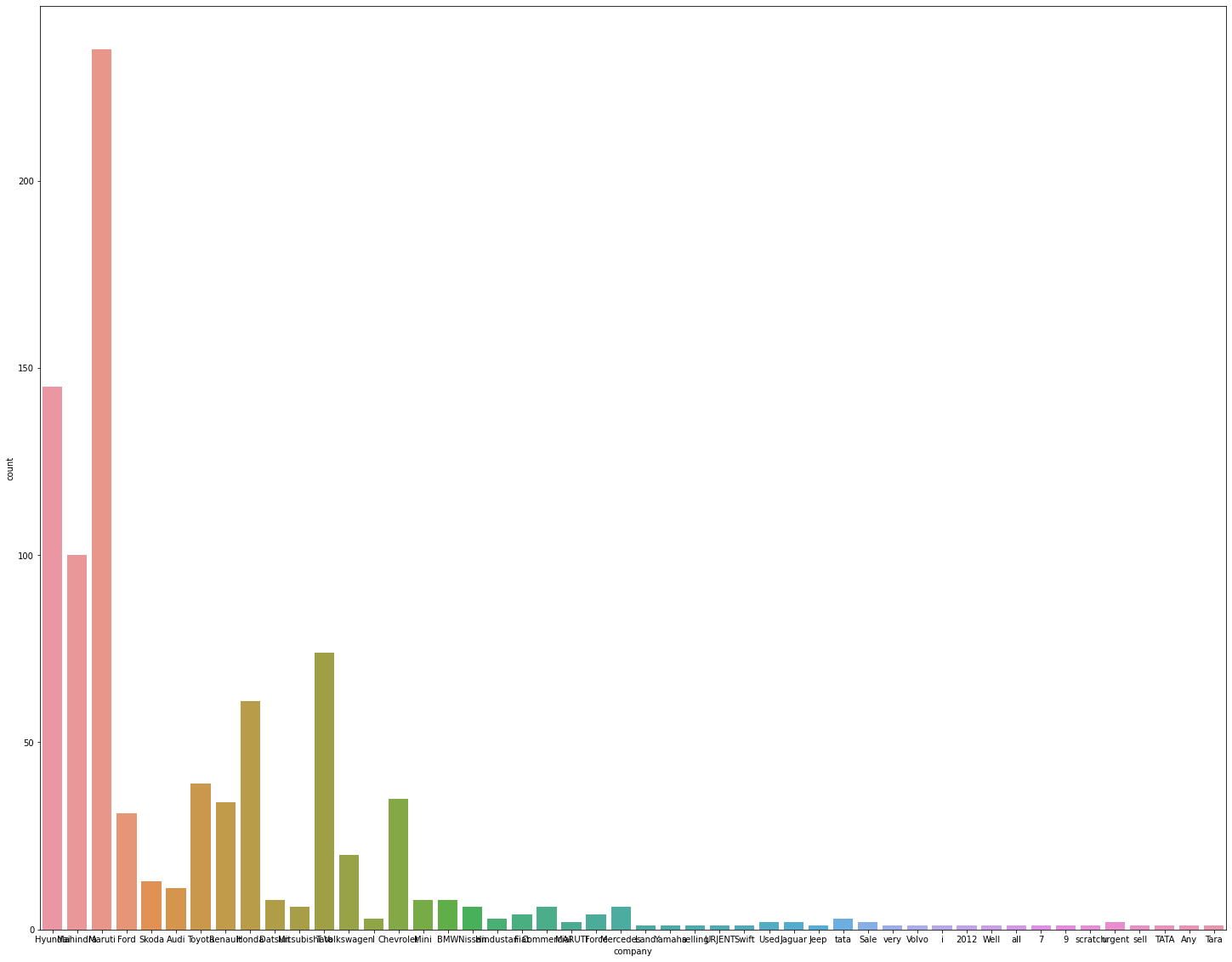


Fig: Raw Company

Fig: Raw Year

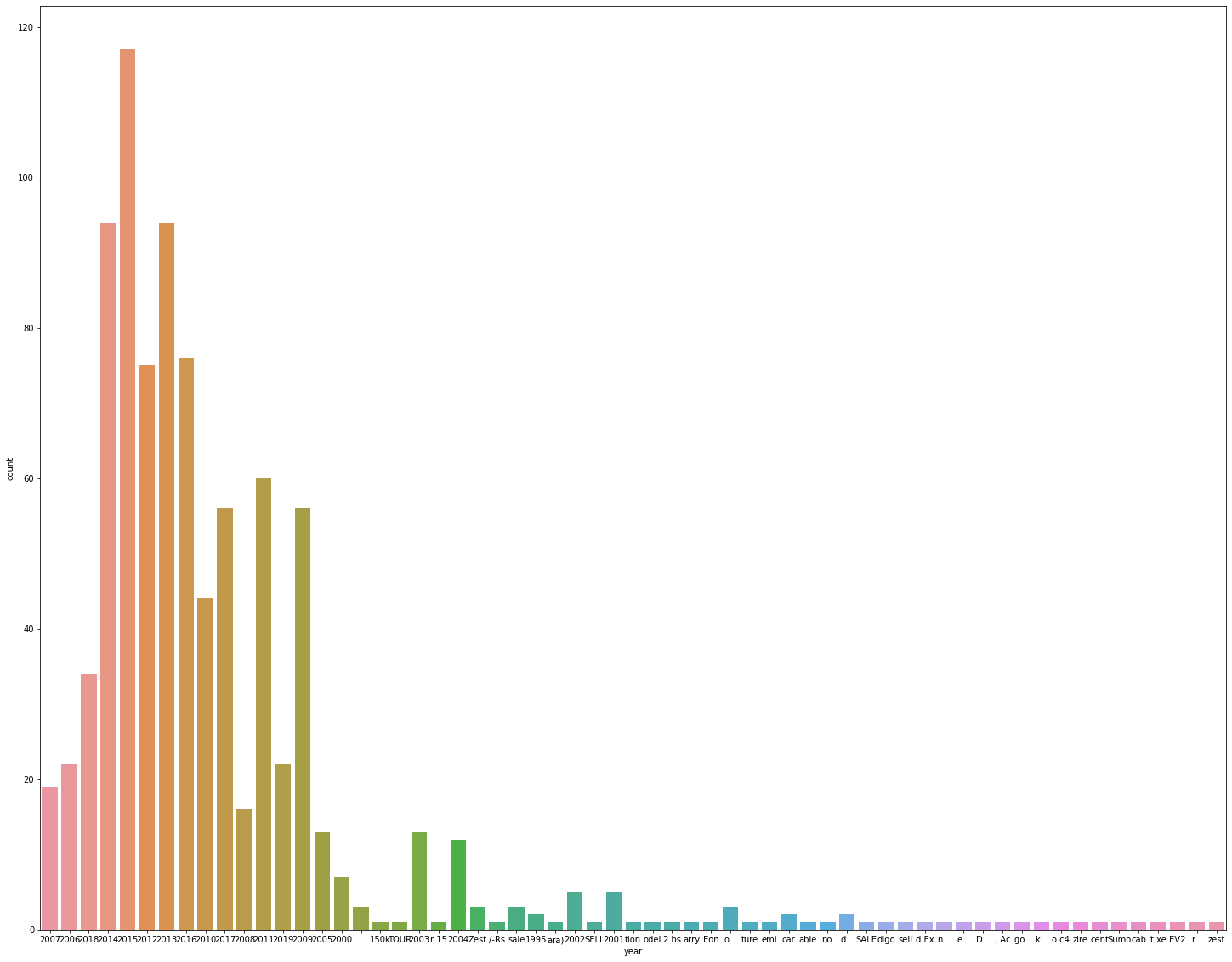
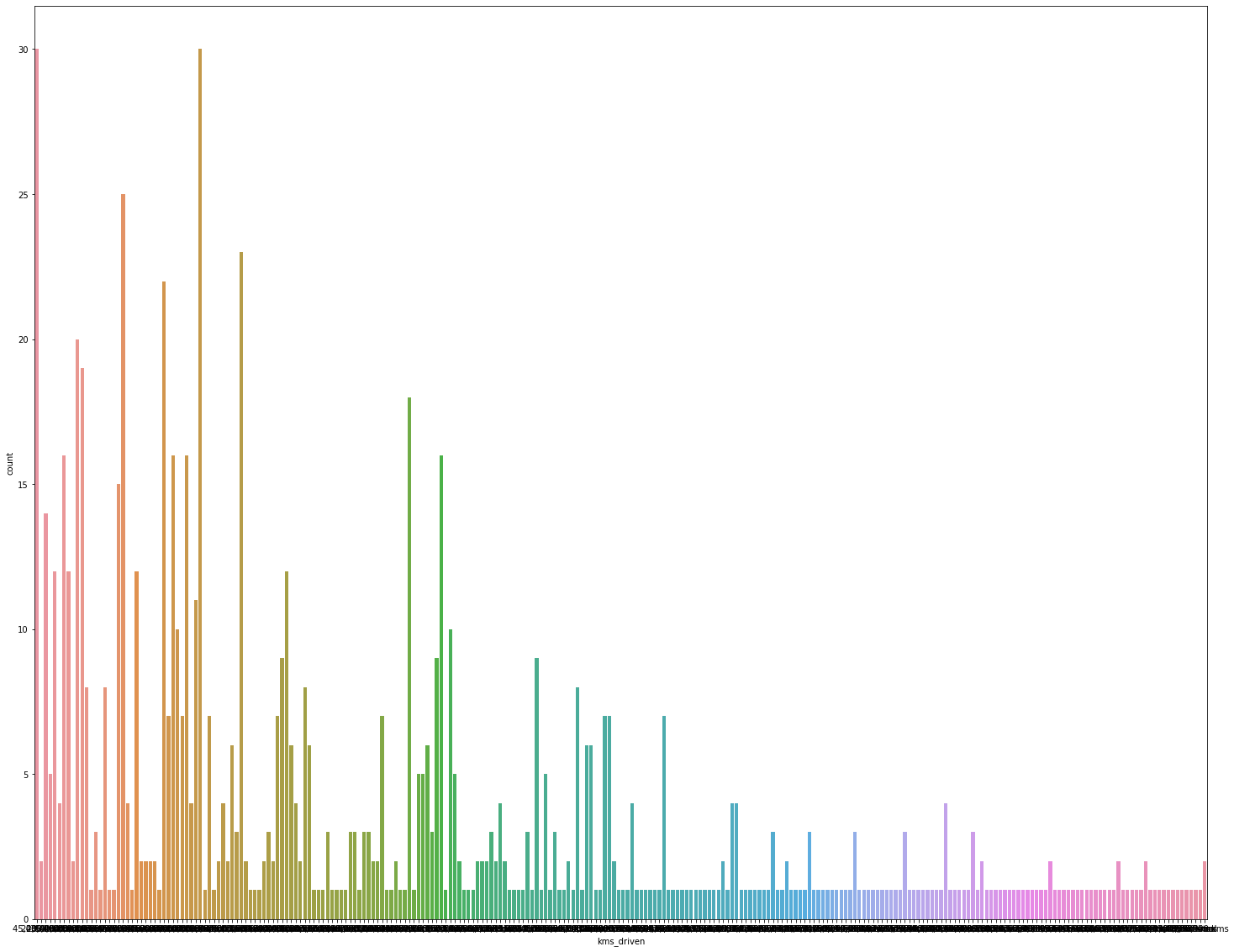
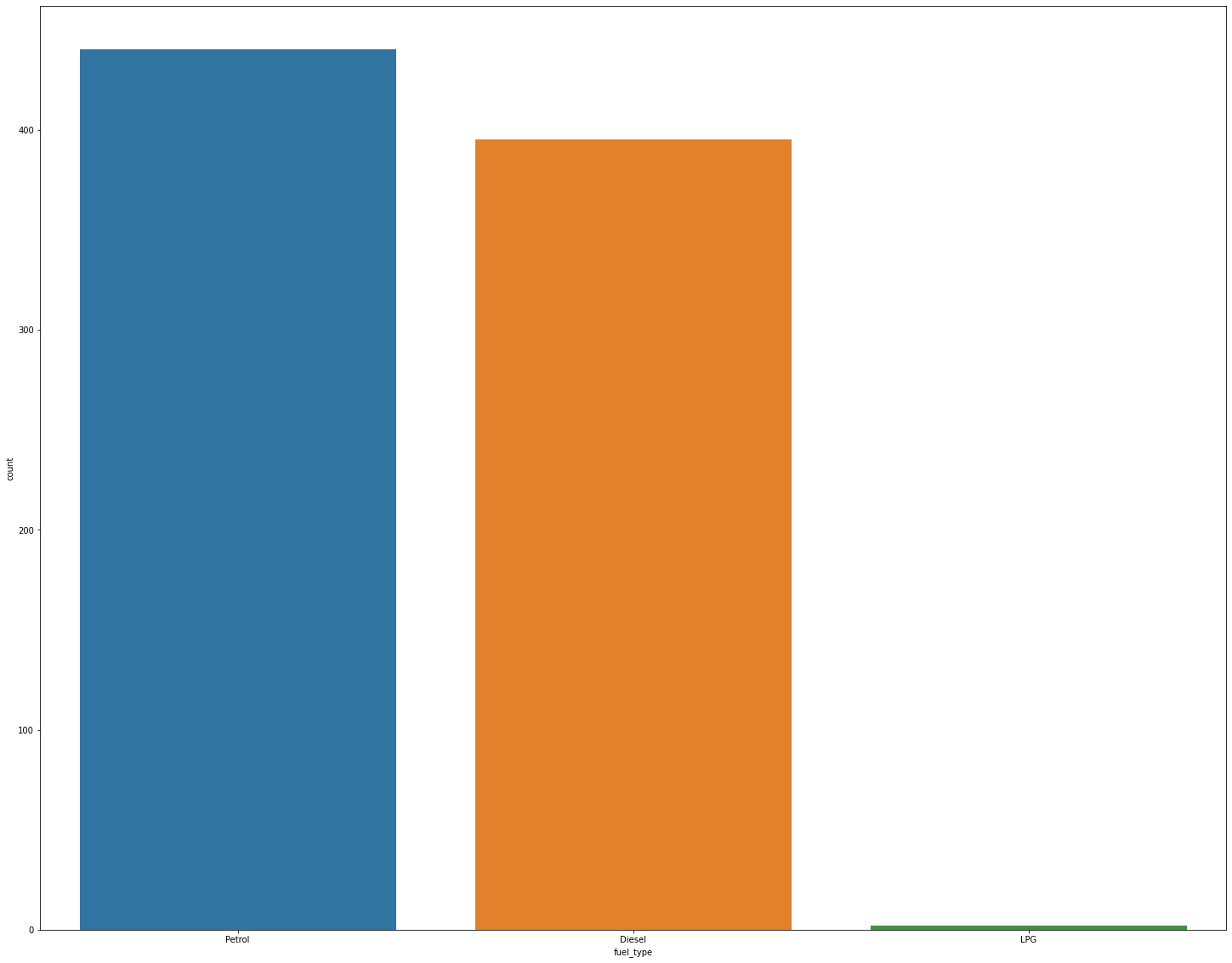


Fig:Raw kms\_driven



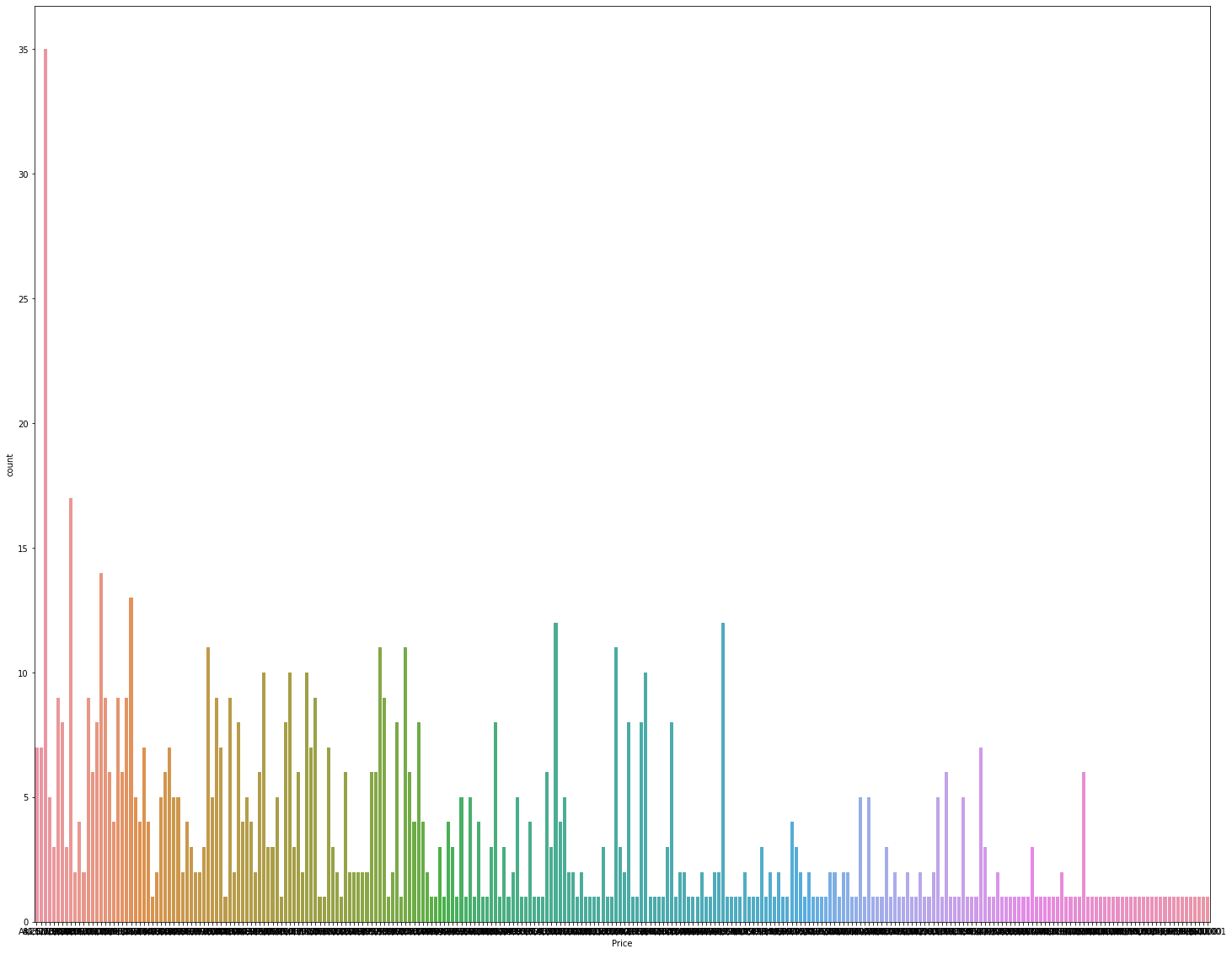
Fig: Raw Fuel Type

Fig: Raw Price

**Data Cleaning Steps**

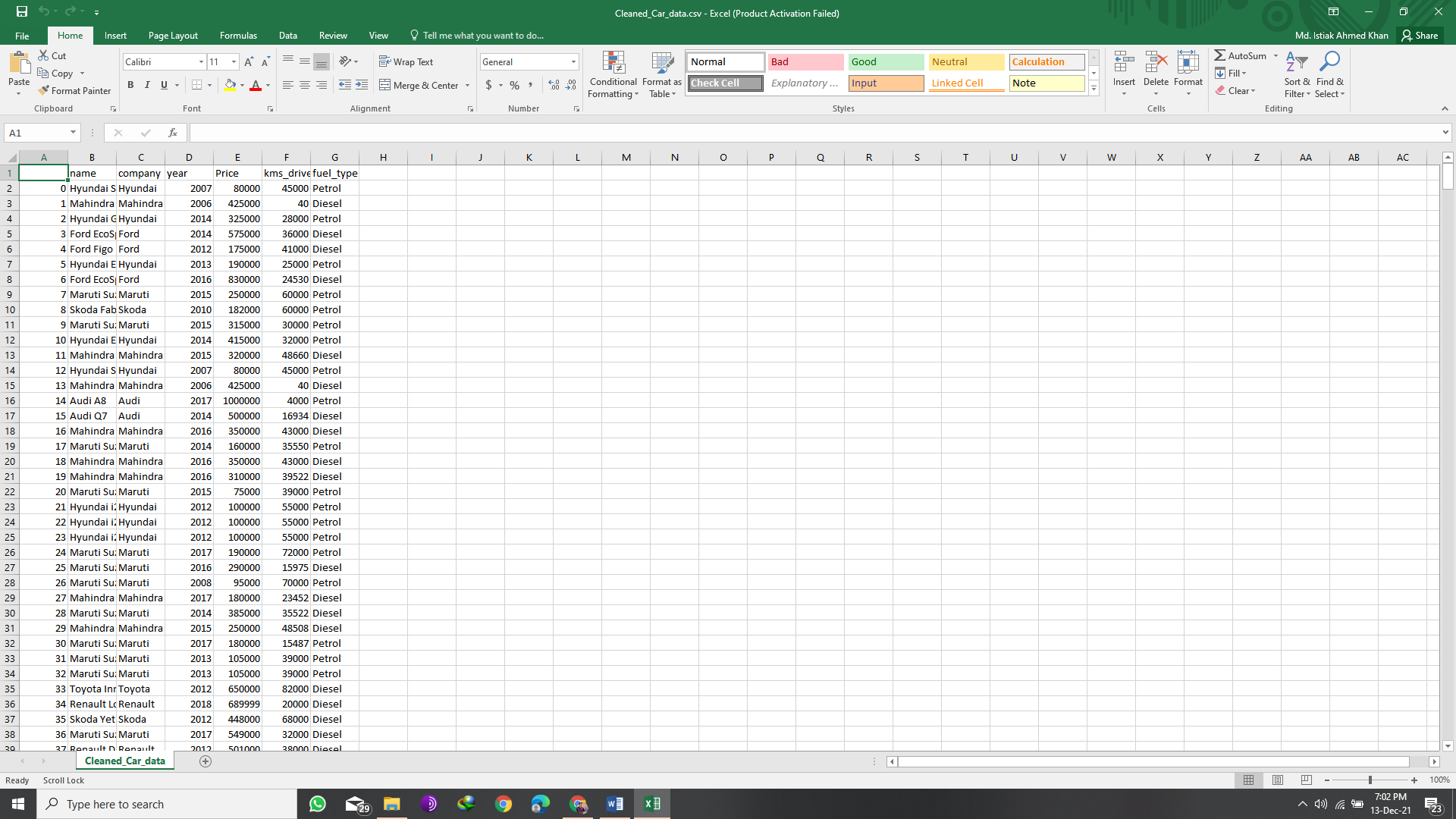
* We removed the string values form column ‘year ’ and changed its data type to int.
* We removed the line “Ask For Price” from column ‘ Price’ and removed the comma’s between the integer numbers. Then we changed its datatype to int.
* In the column ‘kms\_driven’ we remove the non-numeric values and change its datatype to int.
* From the column ‘fuel\_type’ we change remove the nan values.
* In the ‘name’ column the string values in some cases are too long. So, we had the split the string, sliced and joined the first the word of the strings.
* Due to the these changes in the dataset ,index points in the datasets has shifted to.So we reset the index point.

Fig: Cleaned Dataset

**Methods and Model**

Linear Regression Model:

The simple linear regression model is represented by the equation:

**y = α + βX**

By mathematical convention, the two factors that are involved in a simple linear regression analysis are designated X and y. The equation that describes how y is related to x is known as the regression model [. Here in the equation α is the y intercept of the regression line and β is the slope. A regression line can show a positive linear relationship, a negative linear relationship, or no relationship. If the graphed line in a simple linear regression is flat (not sloped), there is no relationship between the two variables. If the regression line slopes upward with the lower end of the line at the y intercept (axis) of the graph, and the upper end of line extending upward into the graph field, away from the x intercept (axis) a positive linear relationship exists. If the regression line slopes downward with the upper end of the line at the y intercept (axis) of the graph, and the lower end of line extending downward into the graph field, toward the x intercept (axis) a negative linear relationship exists.

Formulation of Linear Regression Model:

Linear Regression model consist of random variable Y (called as a response variable) as a linear function of another random variable X (called as a predictor variable) that is represented by the equation

Y= α + βX (eq 1)

α & β are regression coefficients specifying the Y intercept and slope of the line respectively. The regression coefficient α & β are solved by the method of least squares, which minimize the error between the actual data & the estimate of the line[9] .Given s sample of data or data points of the form

(x1,y1),(x2,y2)…….( xs, ys)

than the regression coefficients α & β are given by

β =∑( xi - x) (yi -y)/ ∑( xi -x) (eq 2) α =y- βx (eq 3)

These values of regression coefficients α and β calculated in equation (2) & (3) are substituted in equation (1) so as to obtain the relationship between the response variable X and the target variable Y.

Algorithm of Linear Regression Technique:

The linear regression technique works on the following algorithm .

Step 1: Take the values of variable Xi and Yi

Step 2: Calculate the average for variable Xi such that average is x= (X1 +X2 +……..+ Xi )/ Xi

Step 3: Calculate the average for variable Yi such that average is y= (Y1 +Y2 +……..+ Yi )/ Yi

Step4: Calculate the value of regression coefficient β by substituting the values of Xi , Yi average of Xi and average of Yi in the equation 2 Step

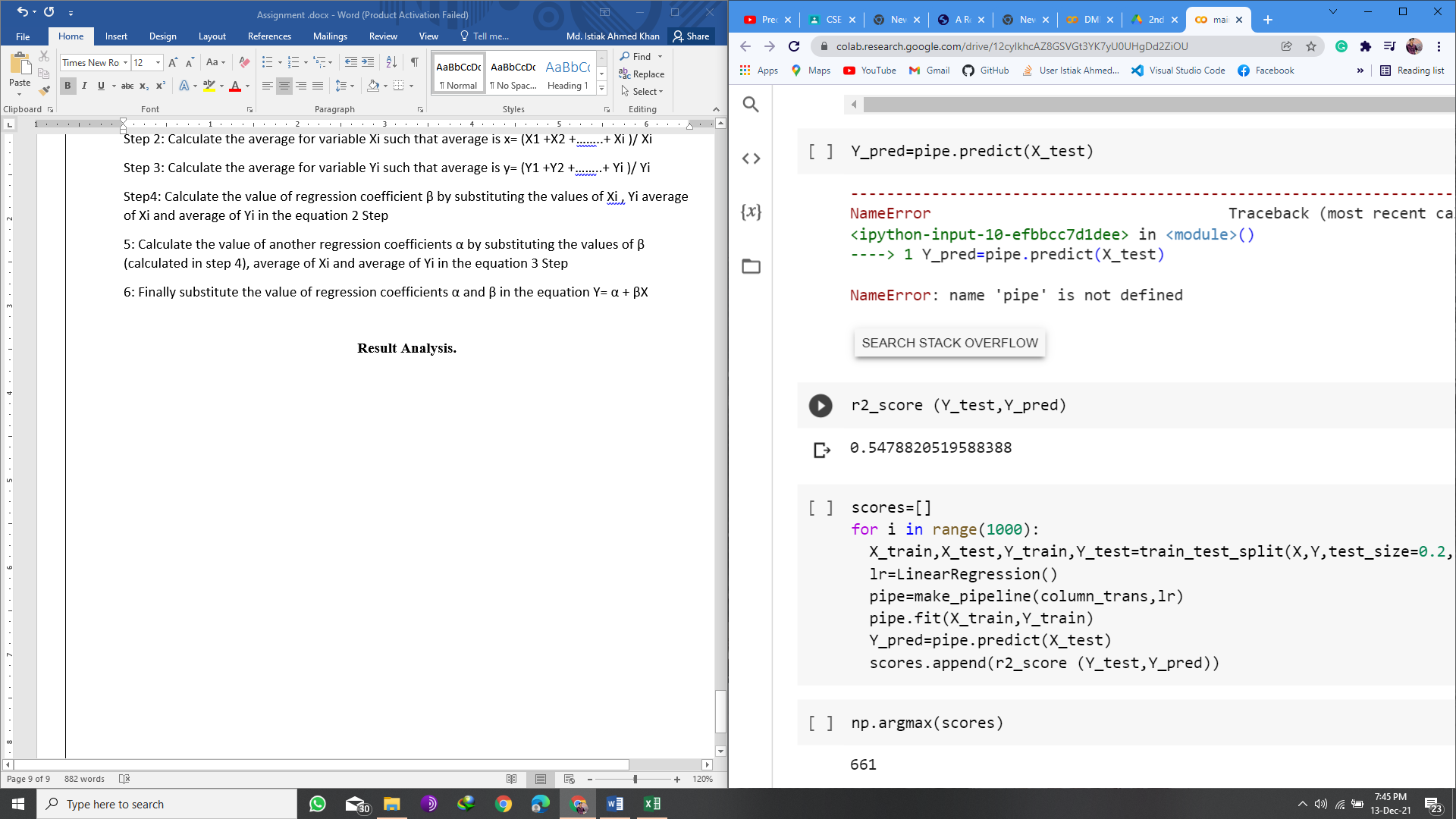
5: Calculate the value of another regression coefficients α by substituting the values of β (calculated in step 4), average of Xi and average of Yi in the equation 3 Step

6: Finally substitute the value of regression coefficients α and β in the equation Y= α + βX

**Result Analysis.**

Coefficient of determination also called as R2 score is used to evaluate the performance of a linear regression model. R2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points

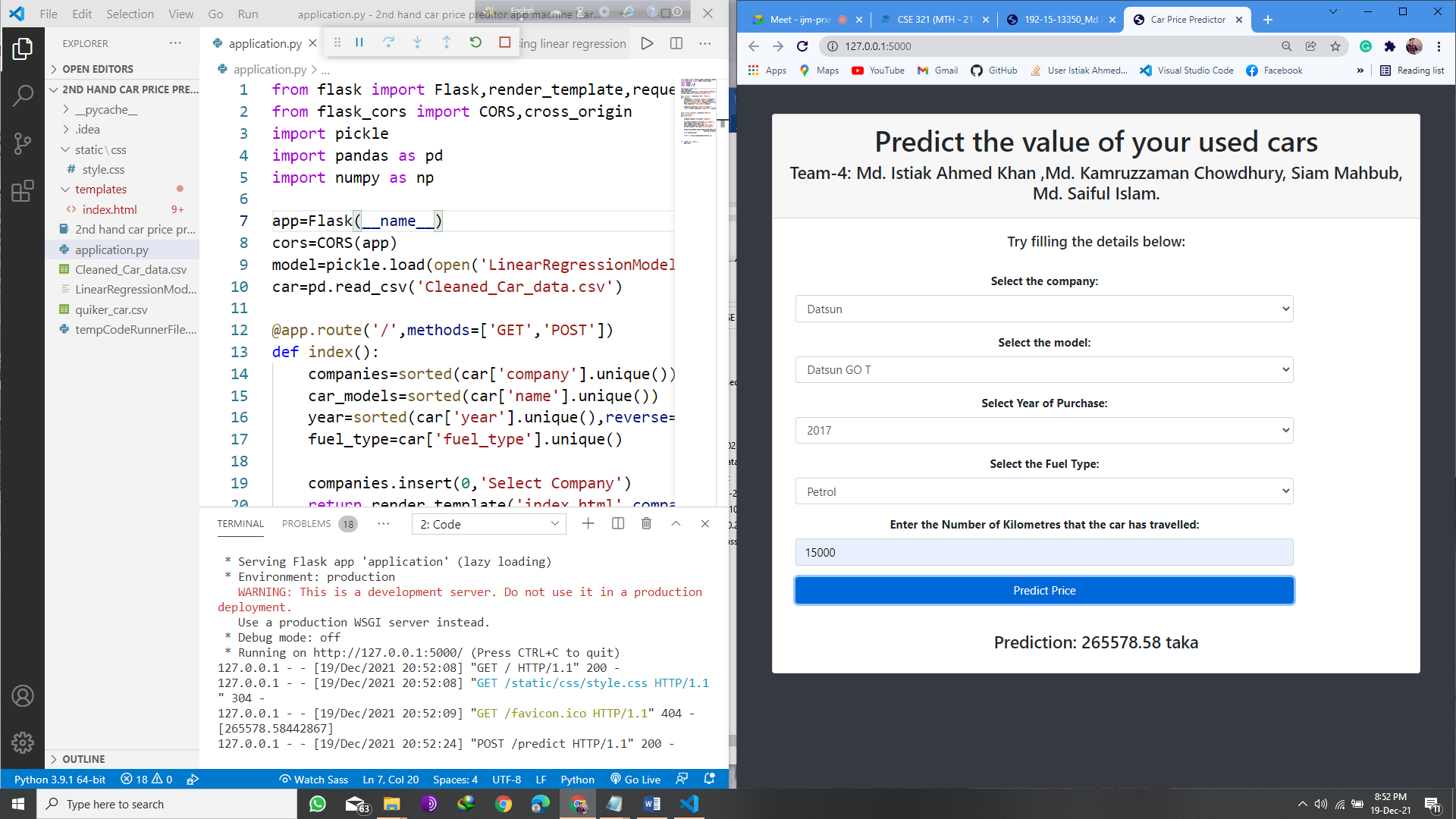
In this project the score is = 0.5478820519588388

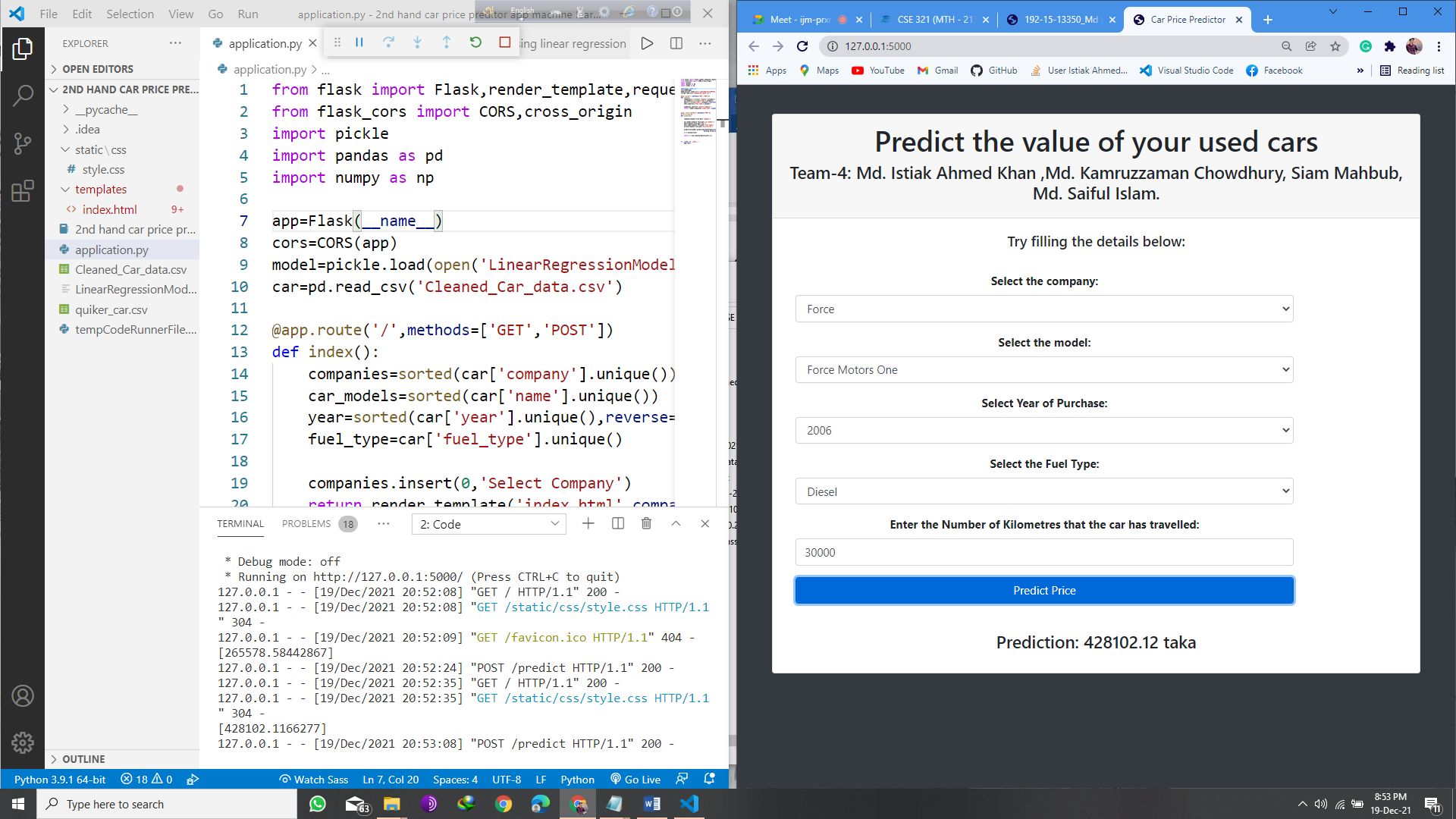


**Application**

It can be used to predict the value of a used car to compare the value between different market places and to estimate the value of a car based on its model ,purchase year, fuel type , kilometers driven .

**Simulation & Experimentation**

****

****

**Timeline For Implementation**

Activities:

* Data collection -ET ,2 day.
* Cleaning the data-ET, 1 day.
* Splitting train and test for best r2 score-ET,1 day.
* Fit the algorithm and generate a model-ET, 2 day.
* Front-end website design with-ET, 1 day.
* Connecting front end with the algorithm for input output validation-ET 3 day.

**Conclusion and Future Work**

This project can help a lot of people in making their dream of buying a car true. The user’s can make the best possible decision in term of buying a car. In future, this price determination technique can be used for various things like determining the value of used laptops, mobile, motorcycles. There’s a possibly a thousand use of this in the relatable fields.