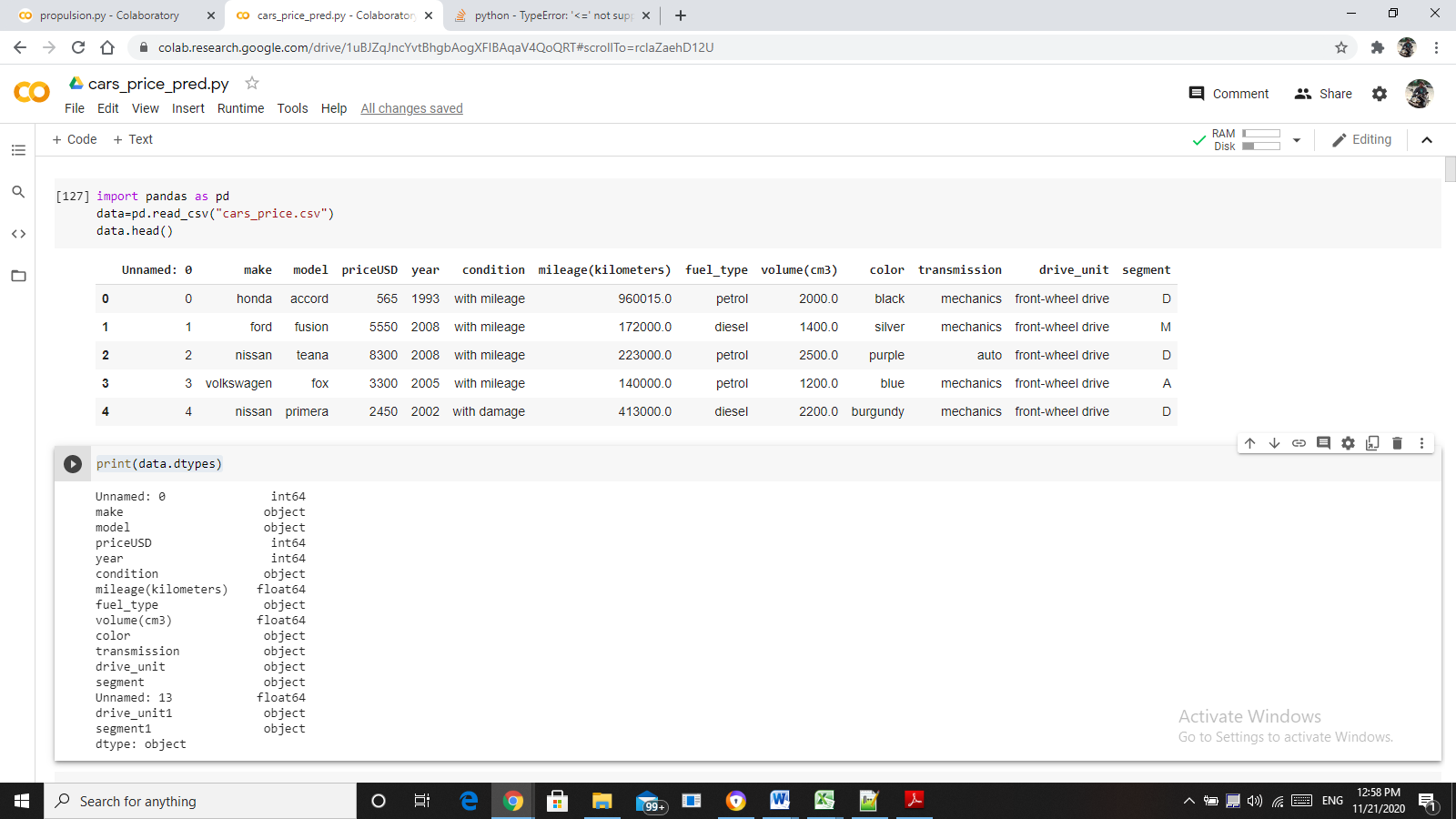
**Assignment 1** - Used Cars Price Prediction and Evaluation

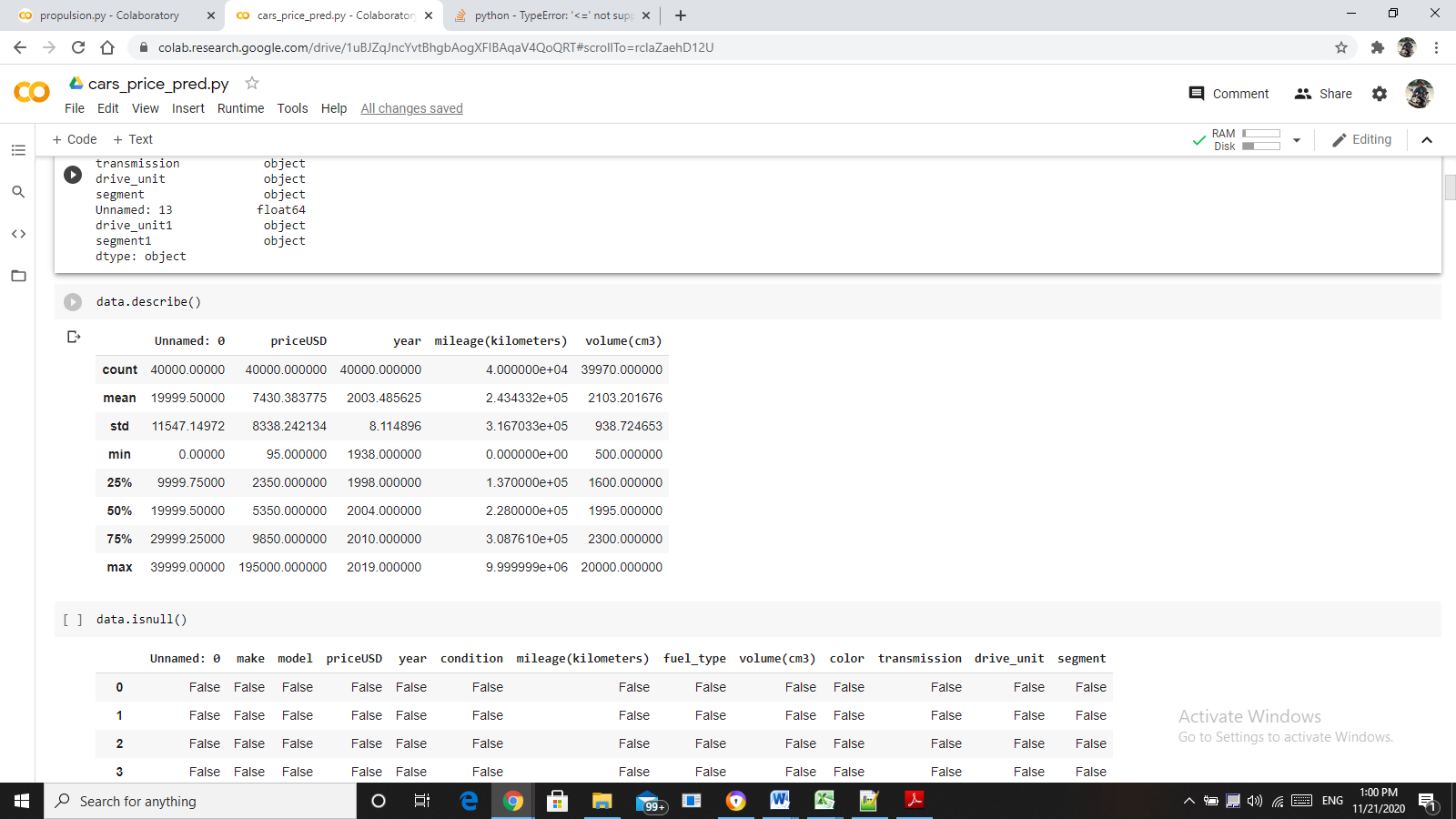
**PART-A**

DATASET

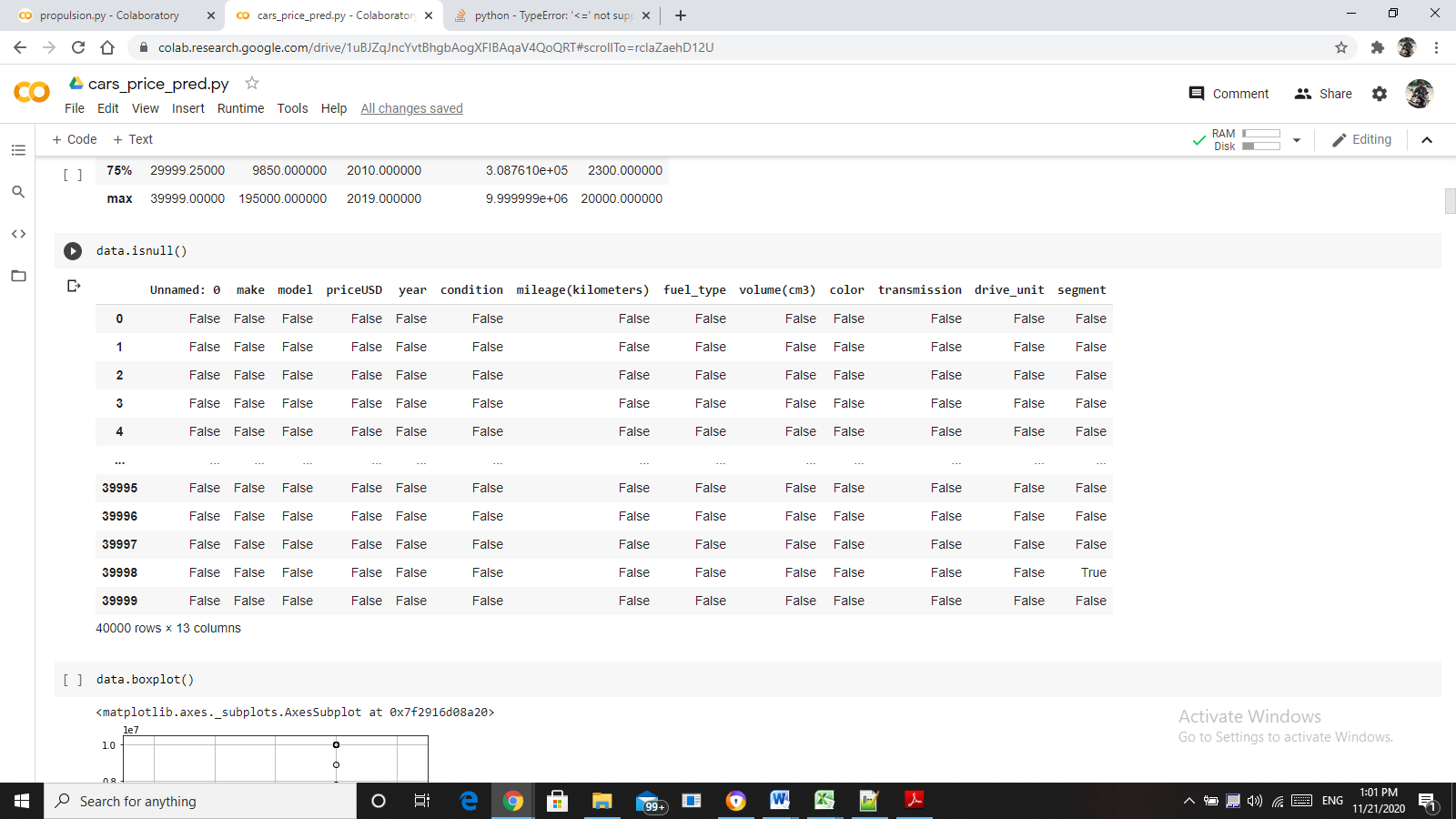


DATA PRE-PROCESSING:

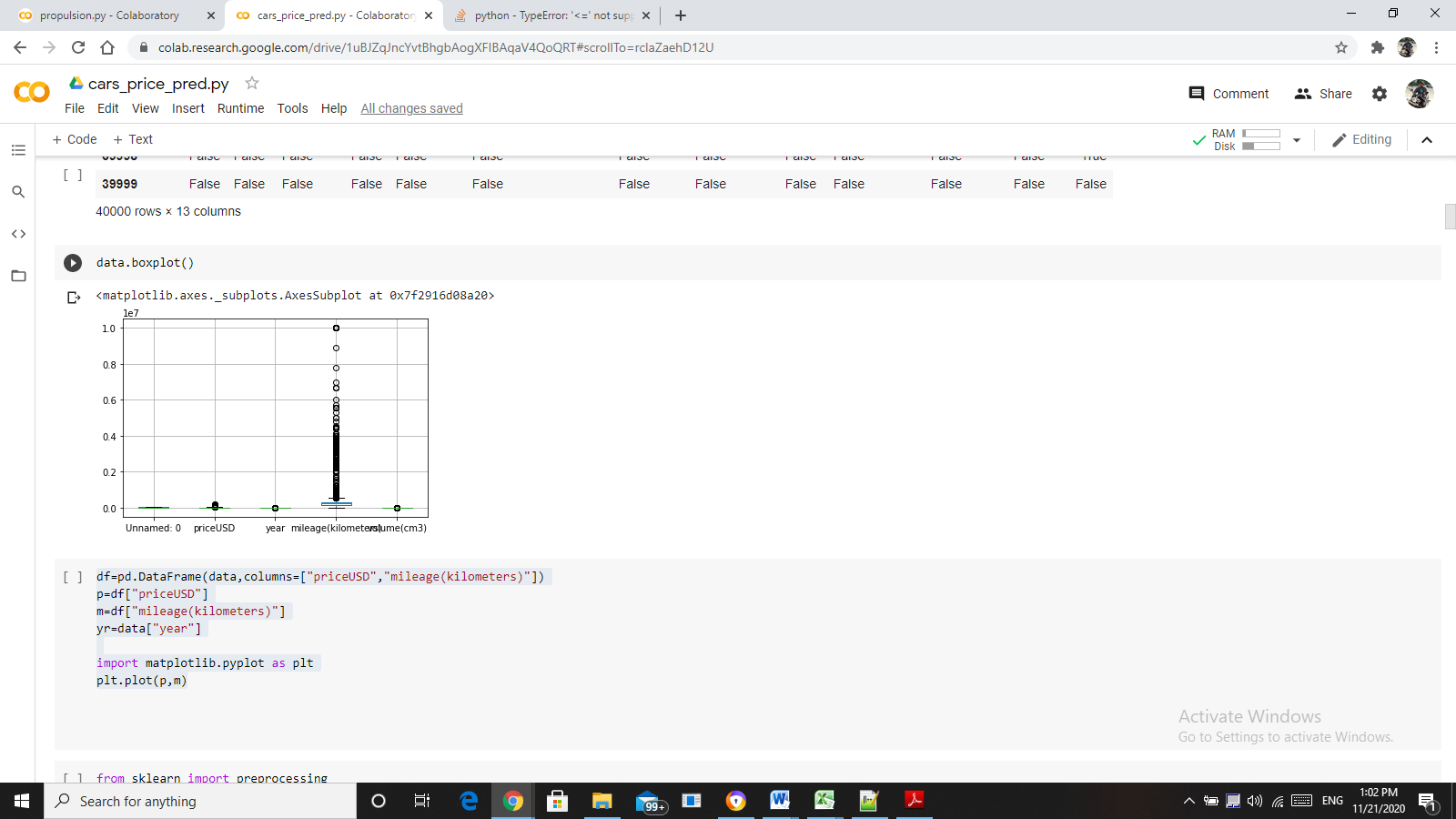
(i) DATA DESCRIPTION



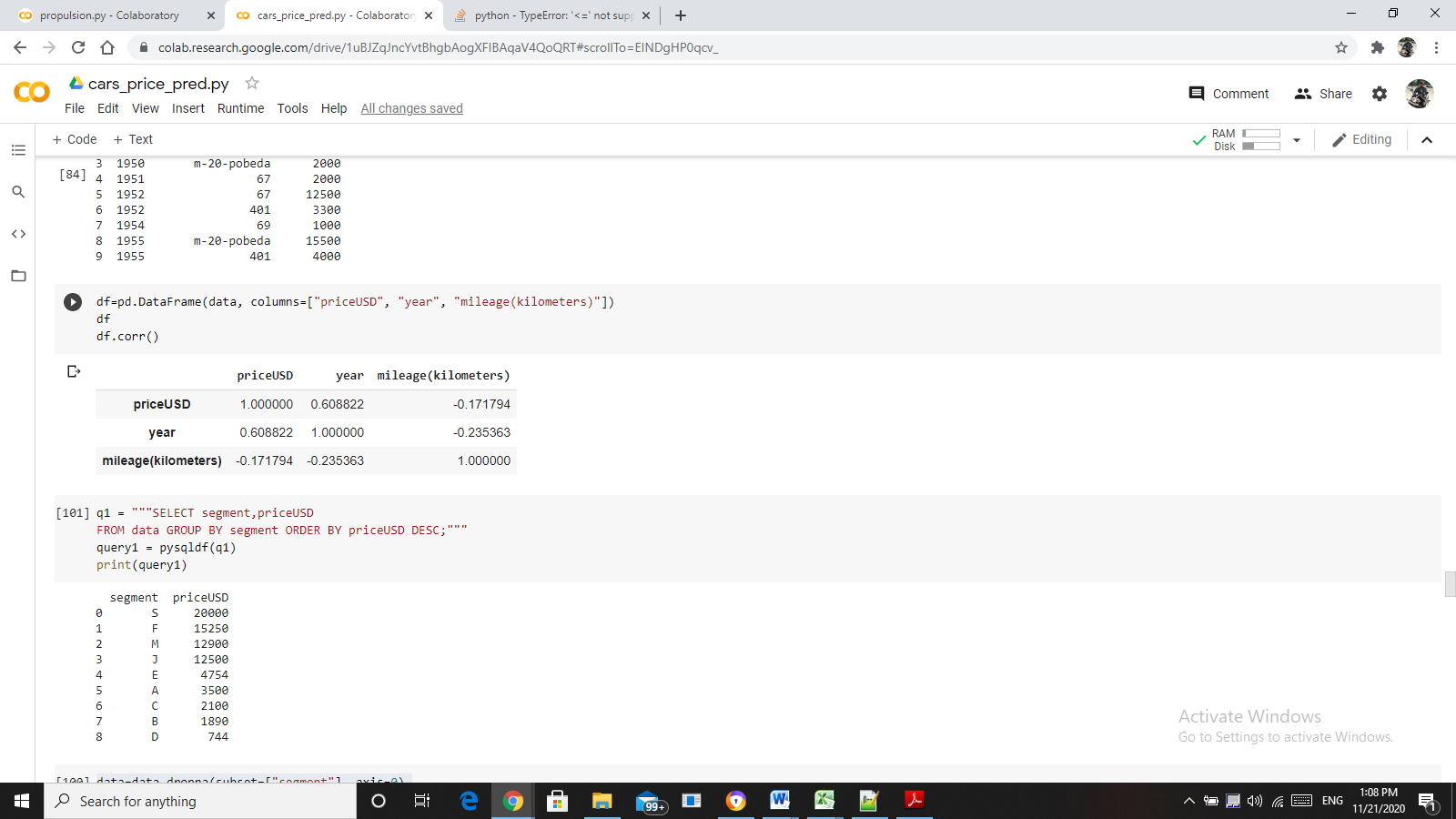
(ii) CECKING FOR NULL VALUES



(iii) DETECTION OF OUTLIERS

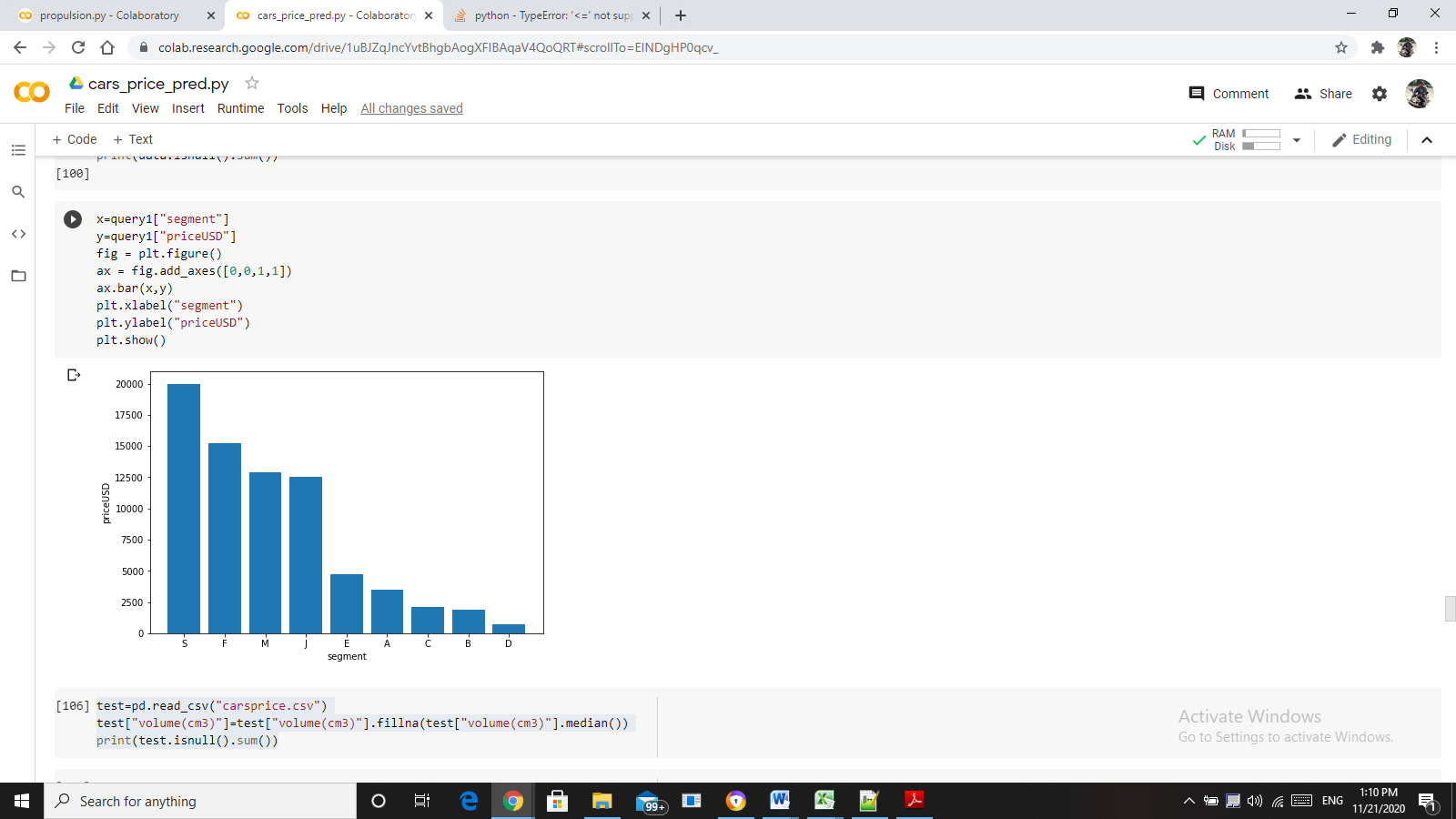


CONCLUSION – FOR PRICESUSD, MILEAGE, YEAR



From the above results, the variables are not co-related. Therefore there is no relationship between them.

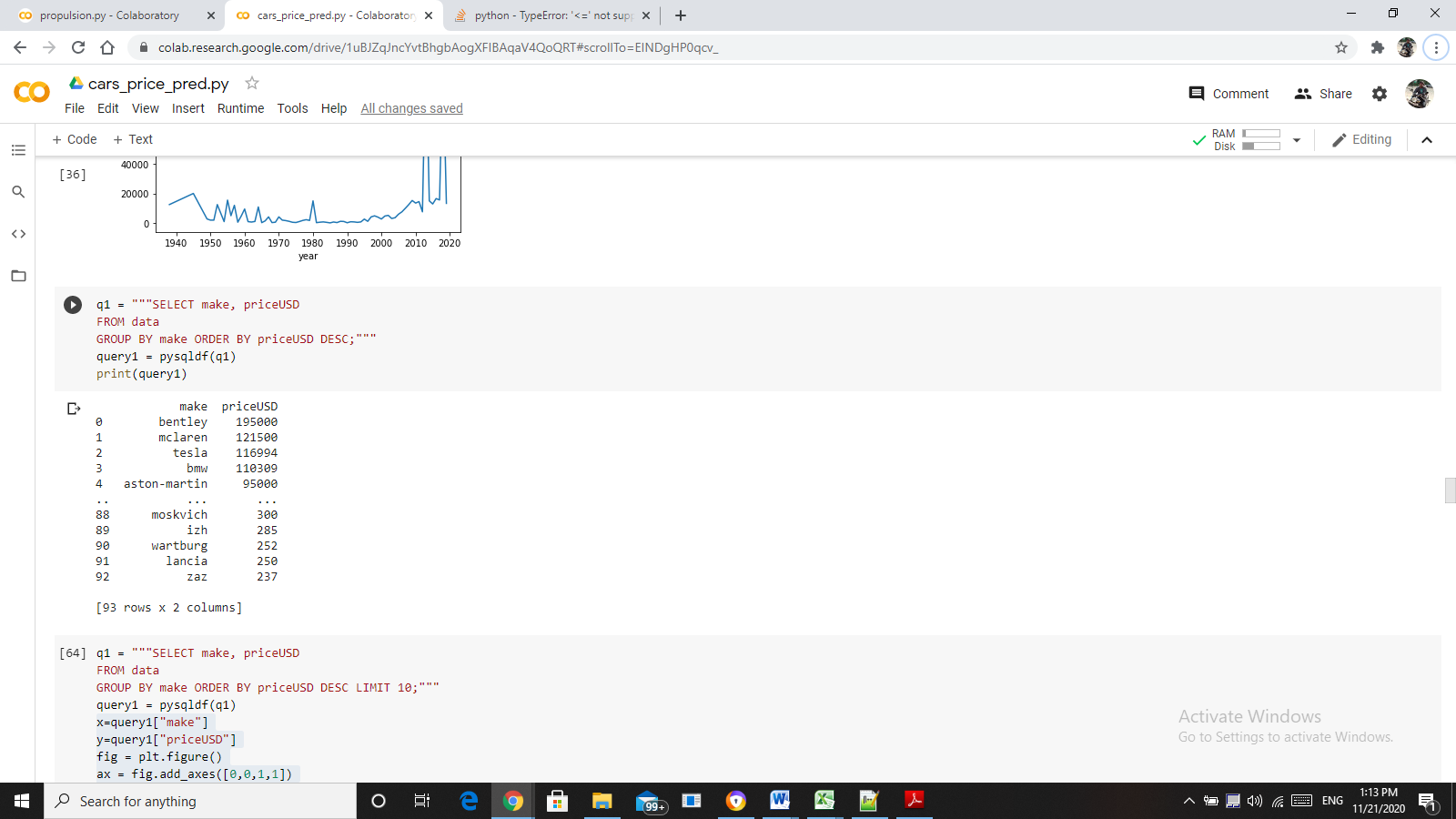
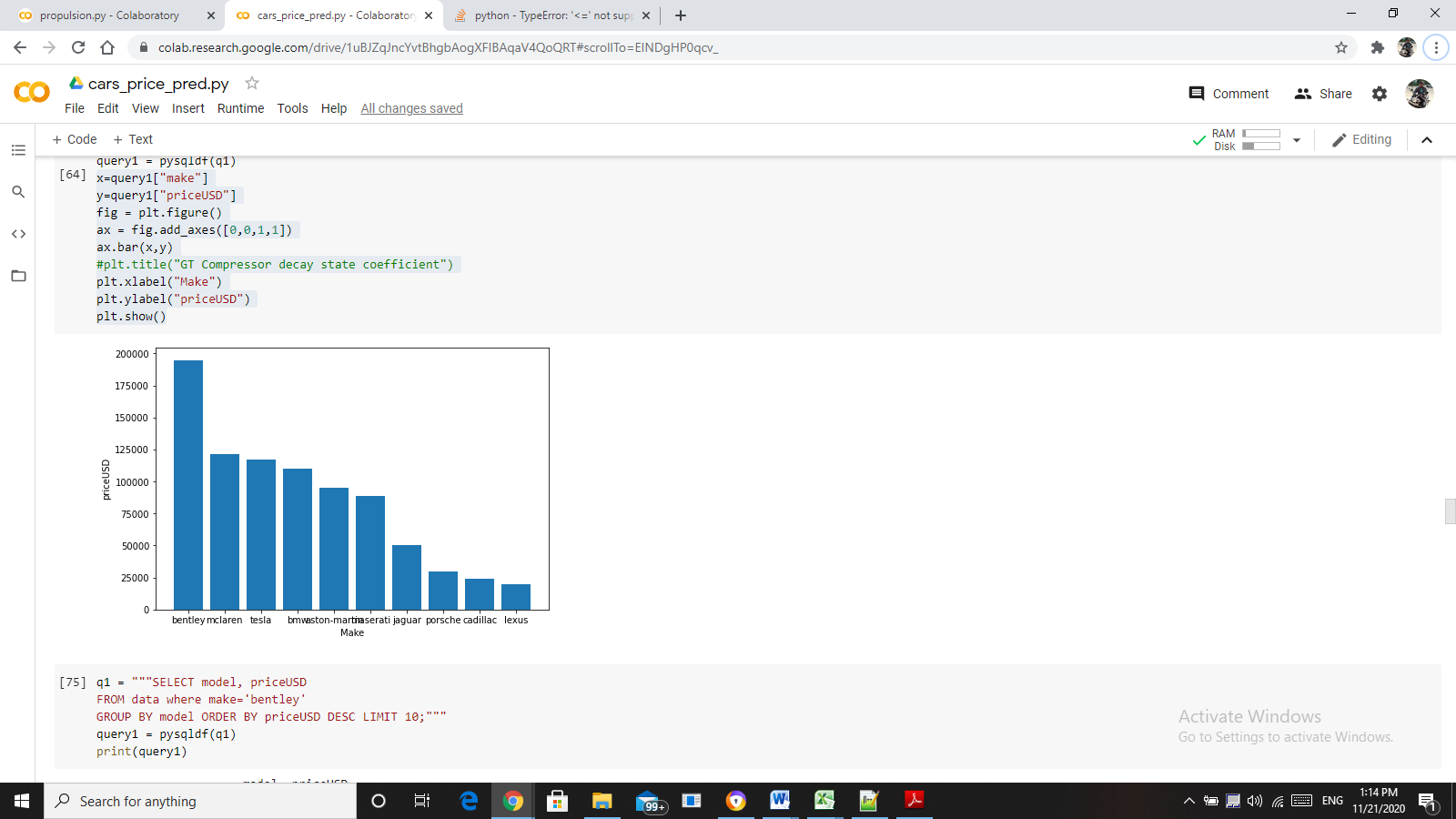
CONCLUSION – FOR SEGMENT, PRICEUSD



From the above graph, segment **S** has the maximum sales.

**GROUP BY QUERY RESULTS**

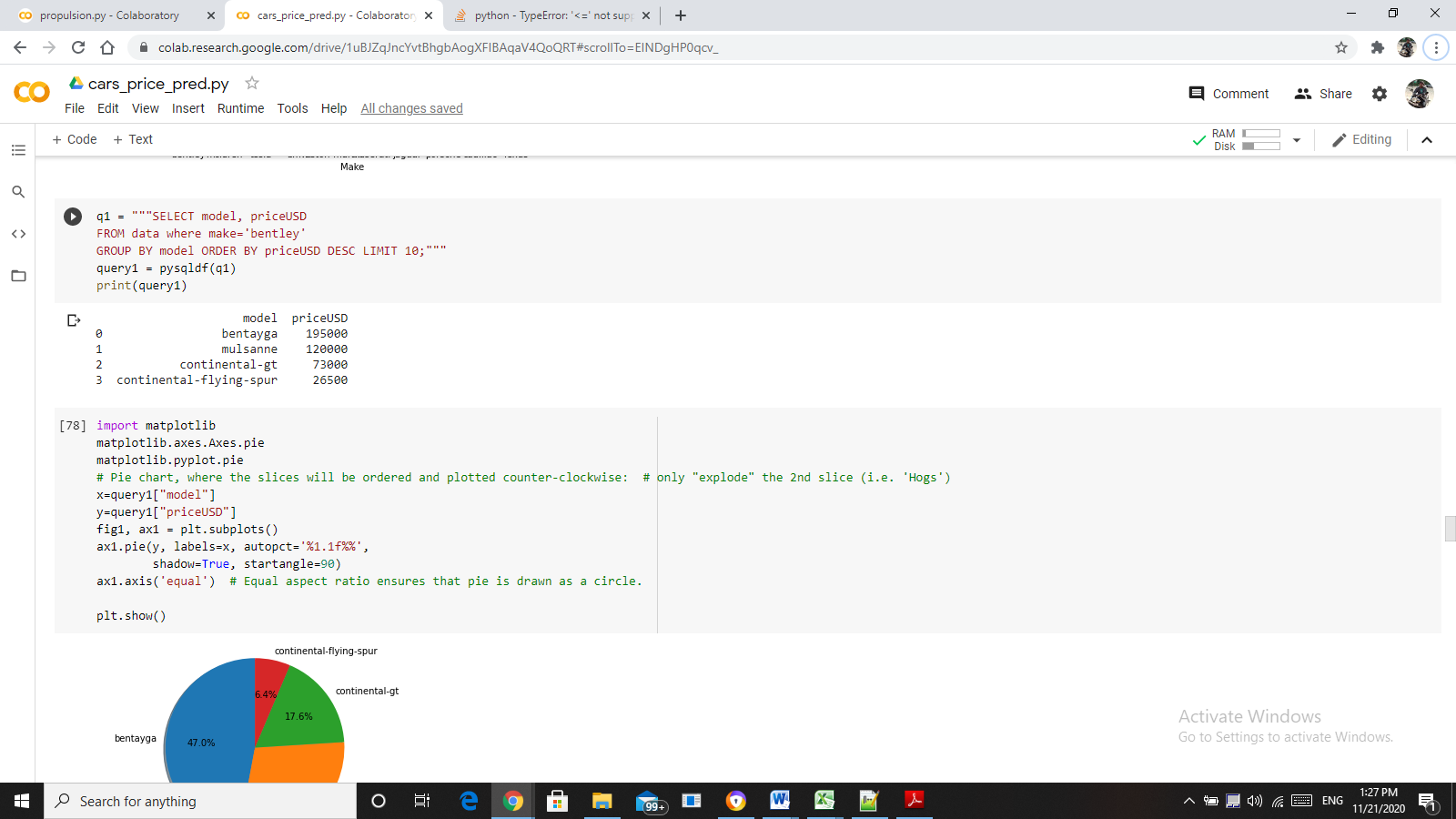
DISPLAYING MAKE, PRICEUSD GROUP BY MAKE ORDER BY PRICEUSD

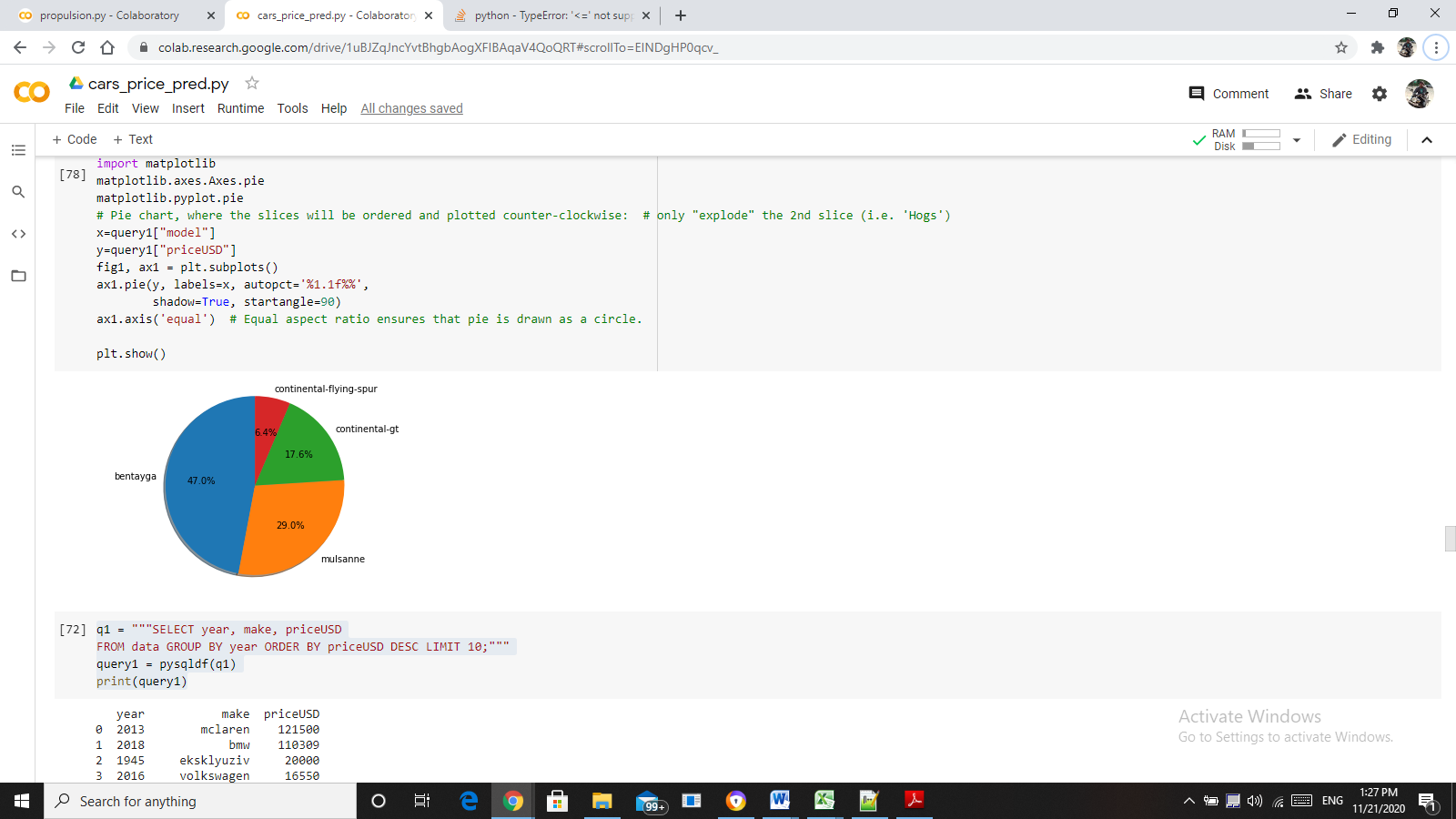
 

* As per the graph, the company **bentley** has the highest sales.

DISPLAYING MODEL FOR A PARTICULAR COMPANY GROUP BY MODEL

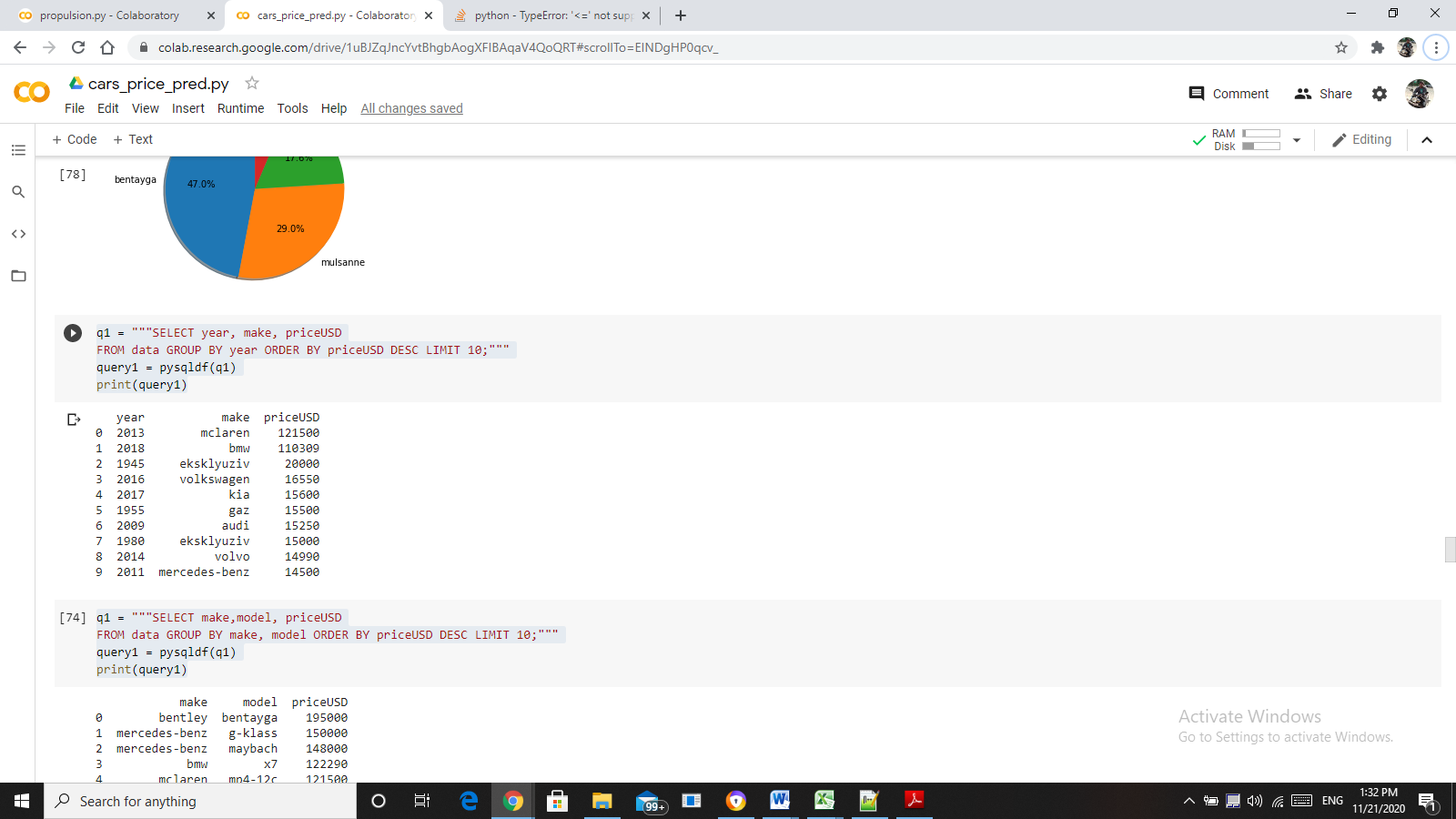
* Since the Bentley Company has the highest sales, the query is grouped by the Bentley models.





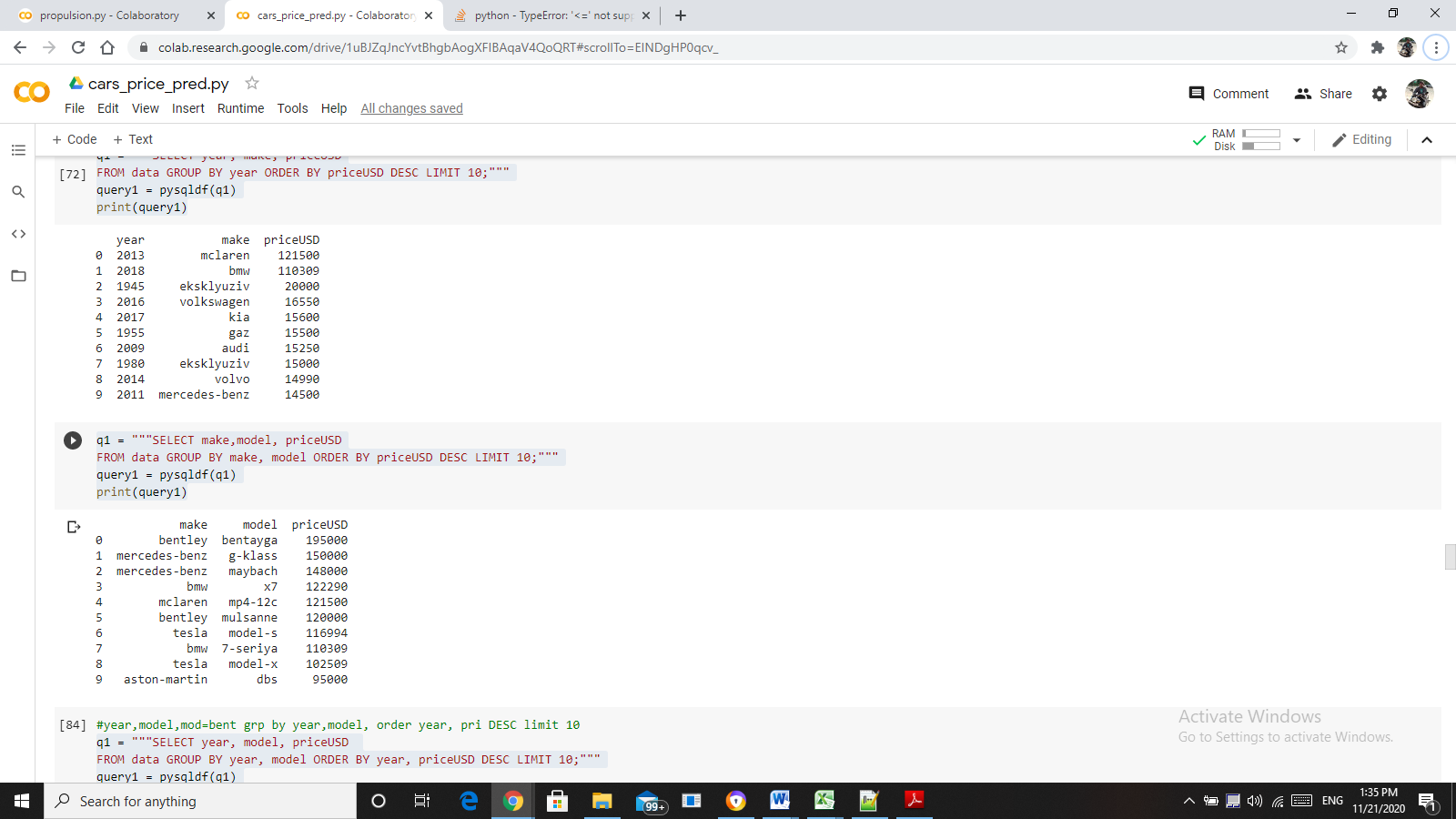
* Based on the pie-chart, the model **bentayga** from the company Bentley has the highest sales.

DISPLAYING YEAR, MAKE, PRICEUSD GROUP BY YEAR ORDER BY PRICEUSD (DESCNDING)



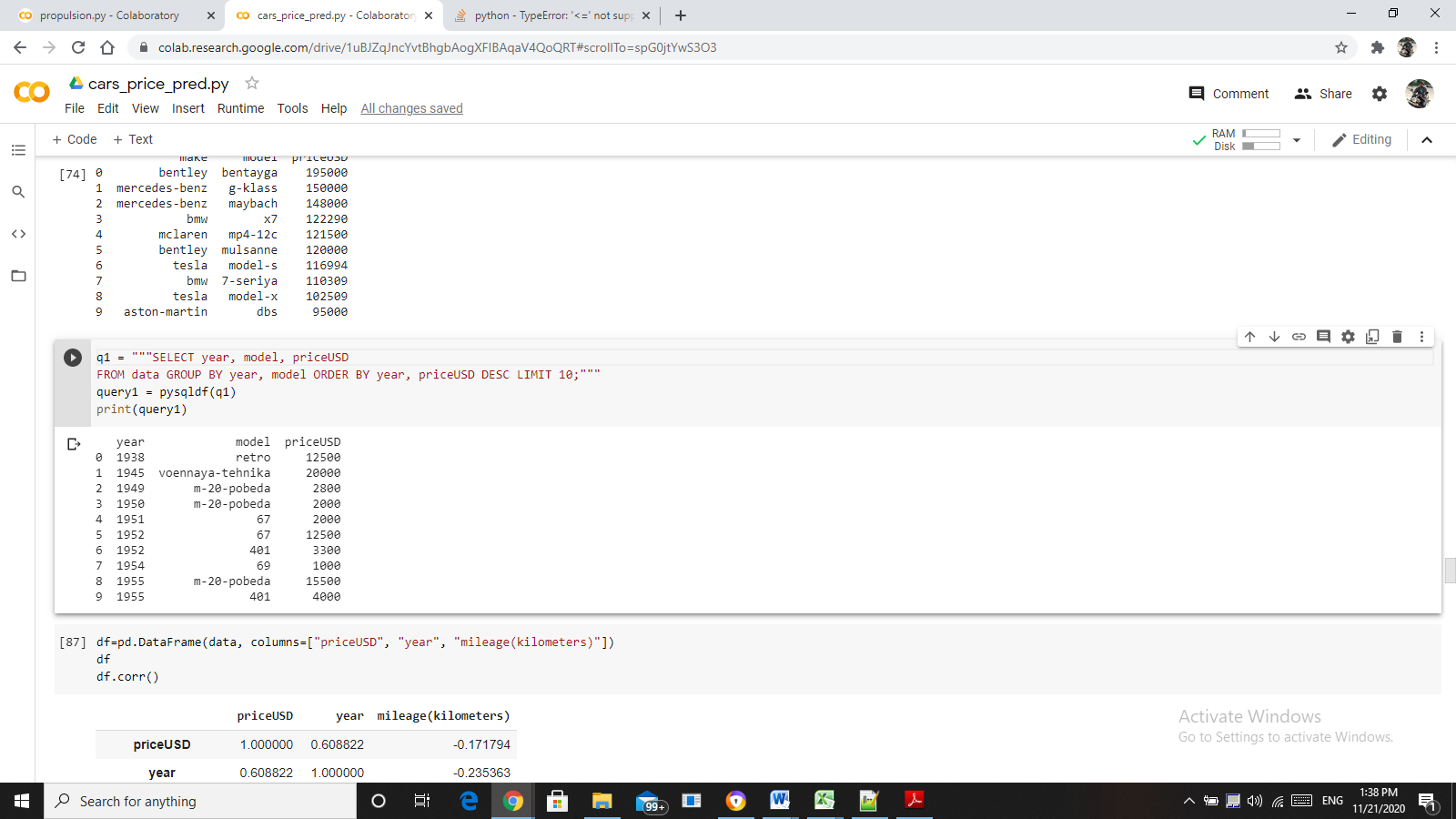
From the result, in the year 2013 the company **McLaren** has the highest sales

DISPLAYING MODEL, MAKE, PRICEUSD GROUP BY MAKE, MODEL ORDER BY PRICEUSD (DESCNDING)



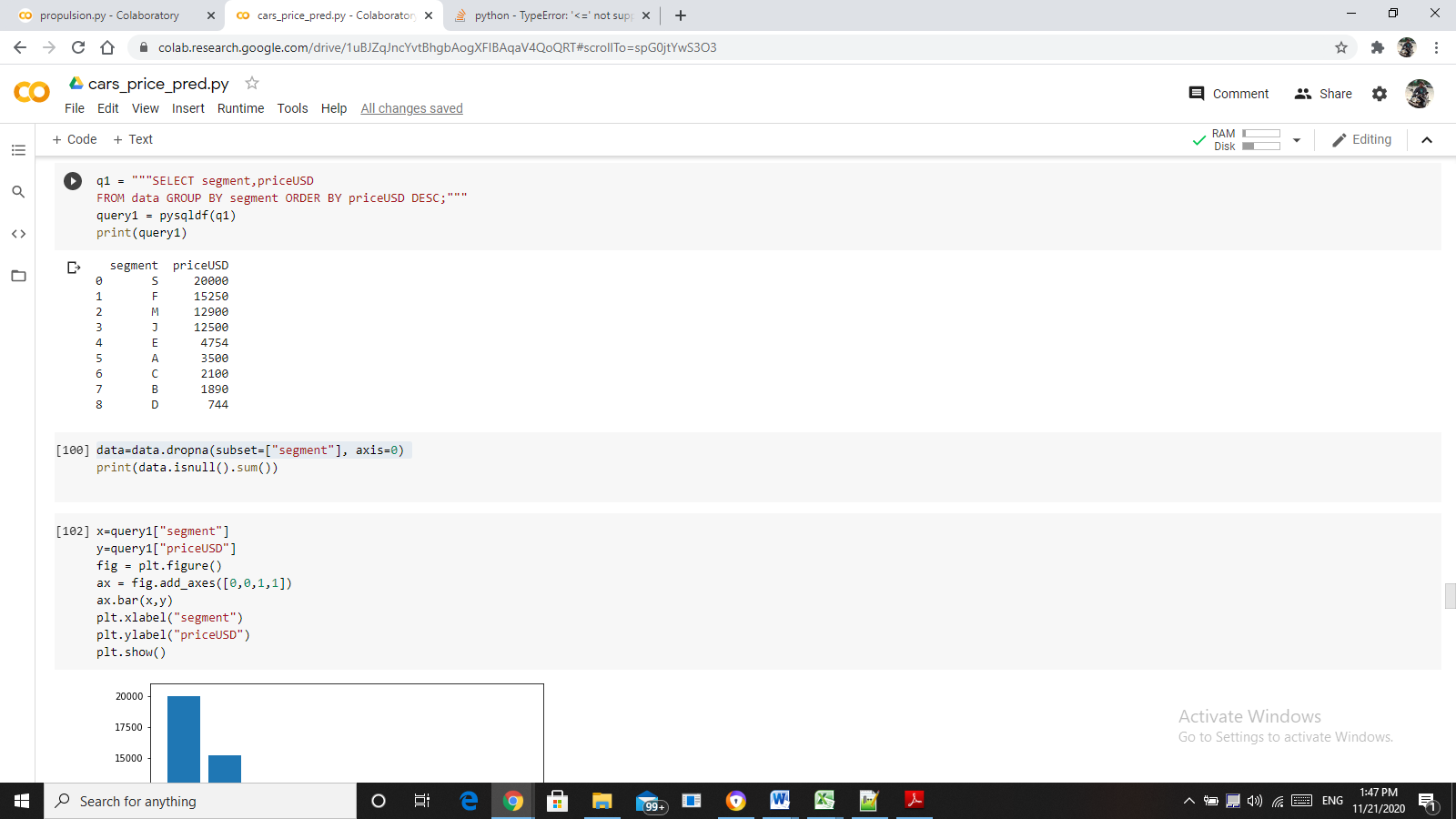
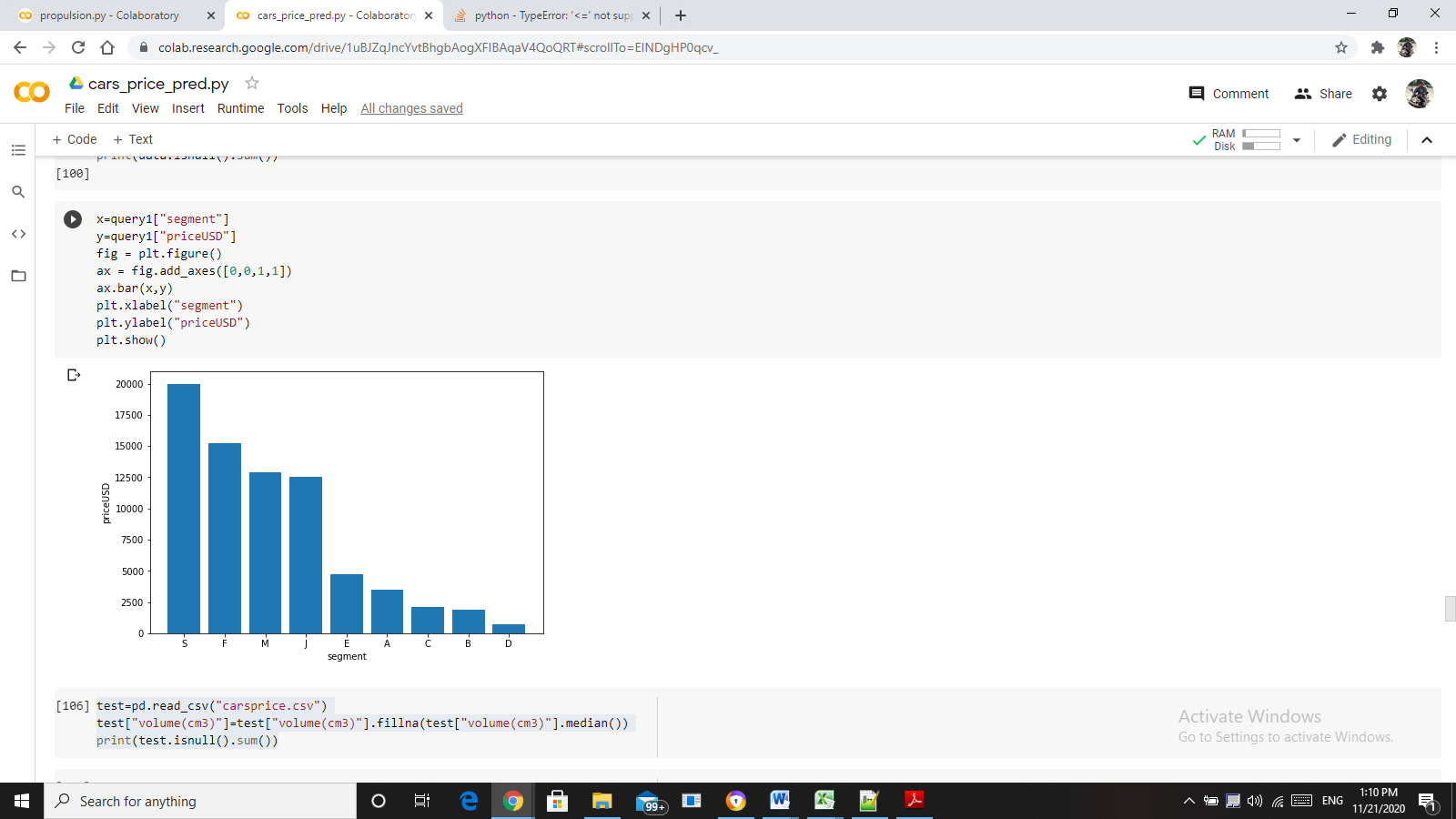
From the result, the **bentayga** of the company Bentley has the highest price.

DISPLAYING YEAR, MODEL, PRICEUSD GROUP BY YEAR, MODEL ORDER BY PRICEUSD (DESCNDING)



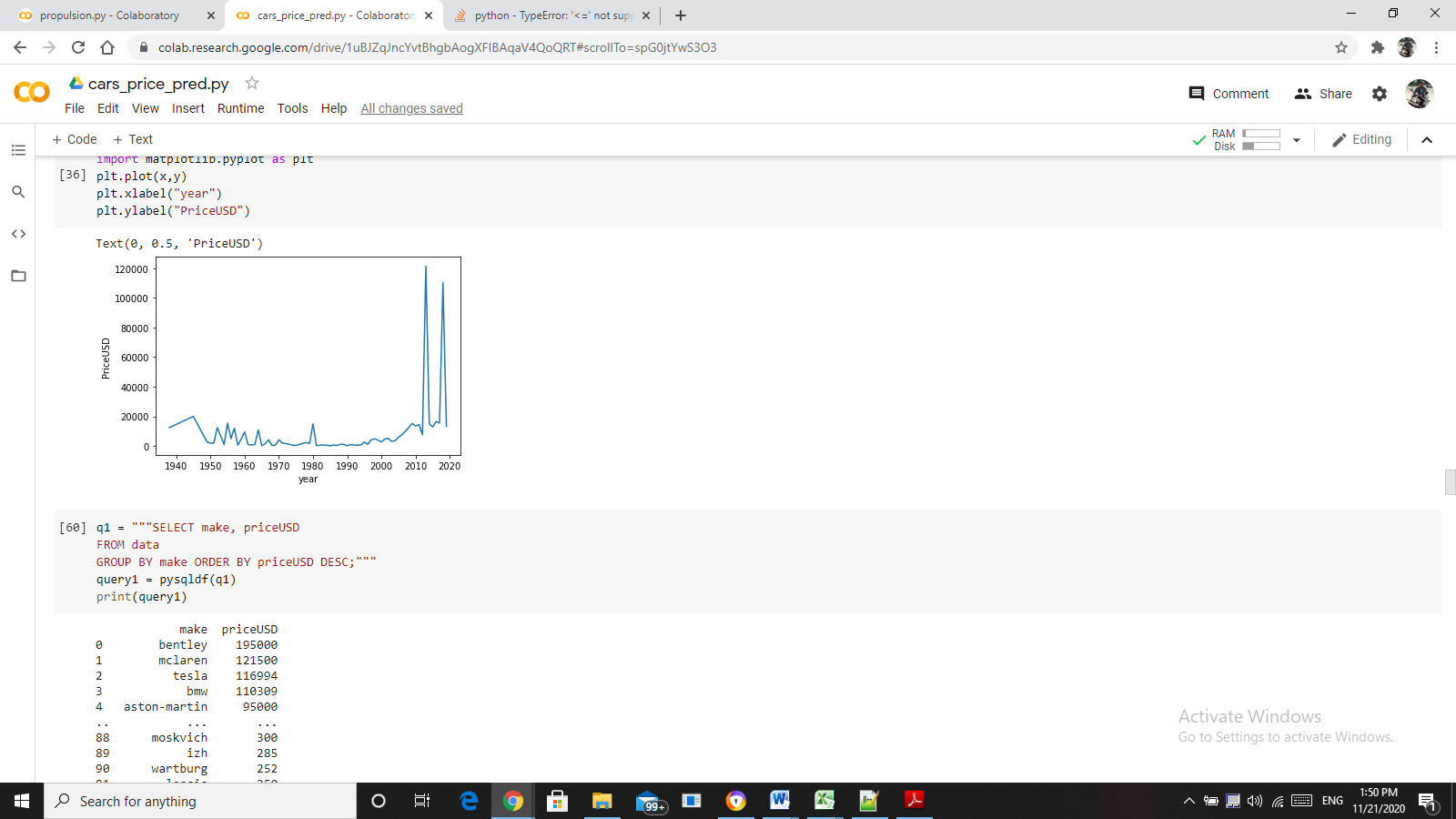
Form the result, in the year 1938 the price of the model **retro** is the highest.

DISPLAYING THE RELATION BETWEEN SEGMENT AND PRICEUSD TO DRAW CONCLUSION

From the above results, the segment **S** has the maximum sales.

**TREND IN DATA**



As per the above graph, the trend at its peak in between the years 2010 and 2020.

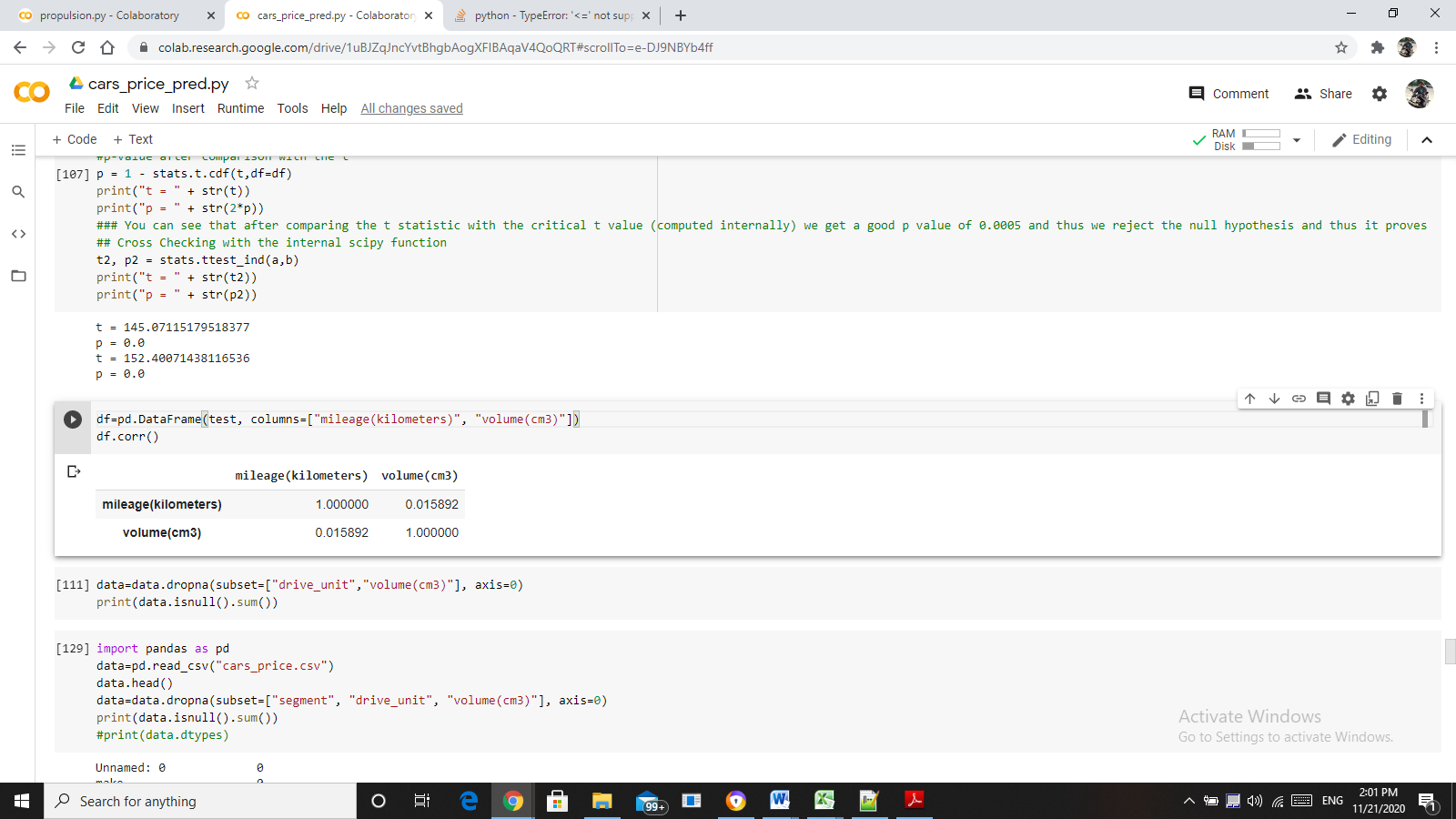
**t-TEST RESULTS**

t-Test is done for the variables mileage (kilometers) and volume (cm3)

t = 145.07115179518377

p = 0.0

Since p value is 0, there is significant difference in mean values of the variables mileage (kilometers) and volume (cm3).



From the above results, the variables are not co-related. Therefore there is no relationship between them.

**CHI-SQUARE TEST RESULTS**

Chi-square test is done for the 3 variables. Fuel type, transmission, drive unit.

RESULT FOR FUELTYPE, TRANSMISSION

alpha = 0.05

p value is 1.0

RESULT FOR TRANSMISSION, DRIVE\_UNIT

alpha = 0.05

p value is 1.0

RESULT FOR FUELTYPE, DRIVE\_UNIT

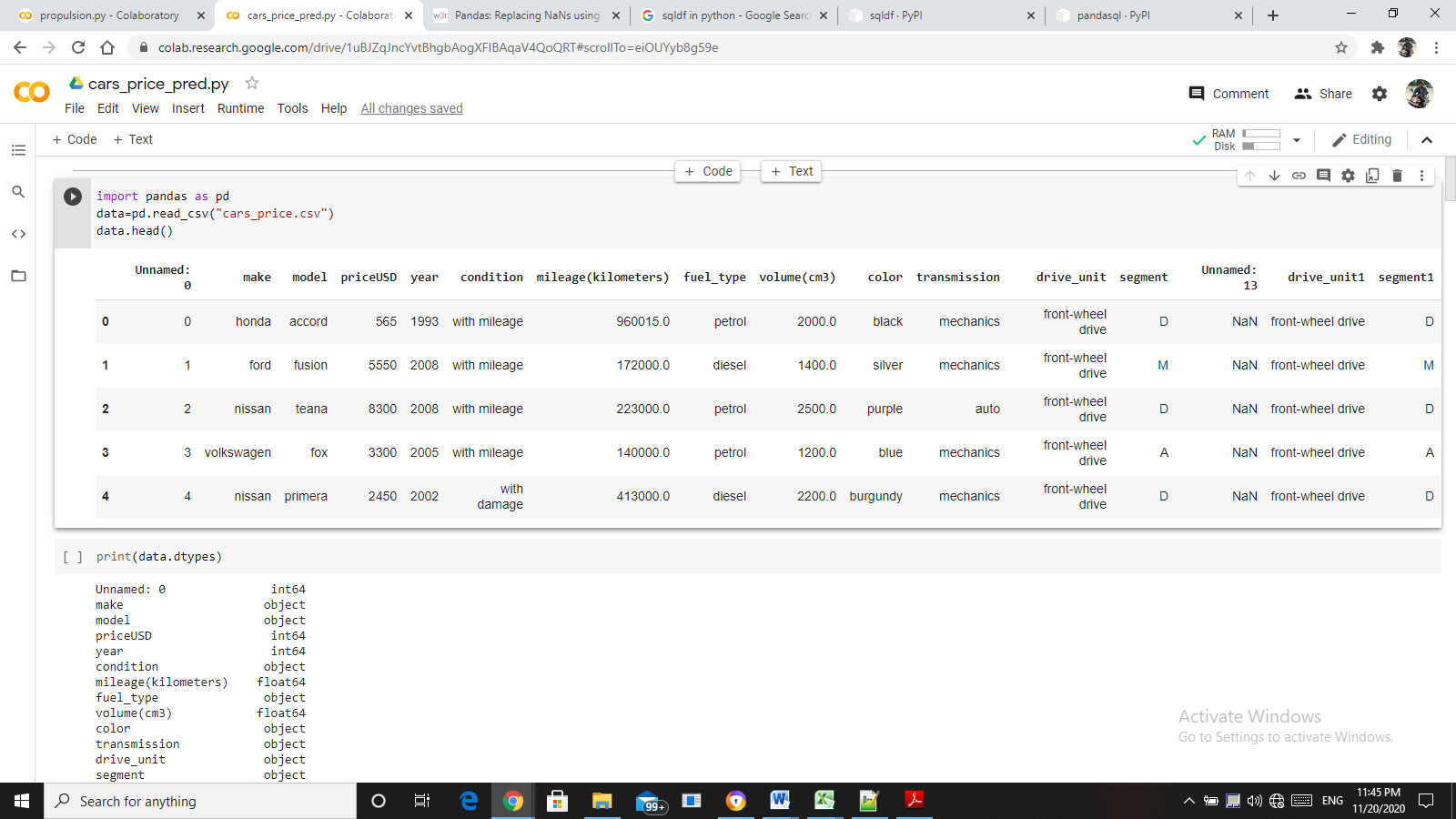
alpha = 0.05

p value is 1.0

Since p value is greater than alpha value, the variables are independent to each other.

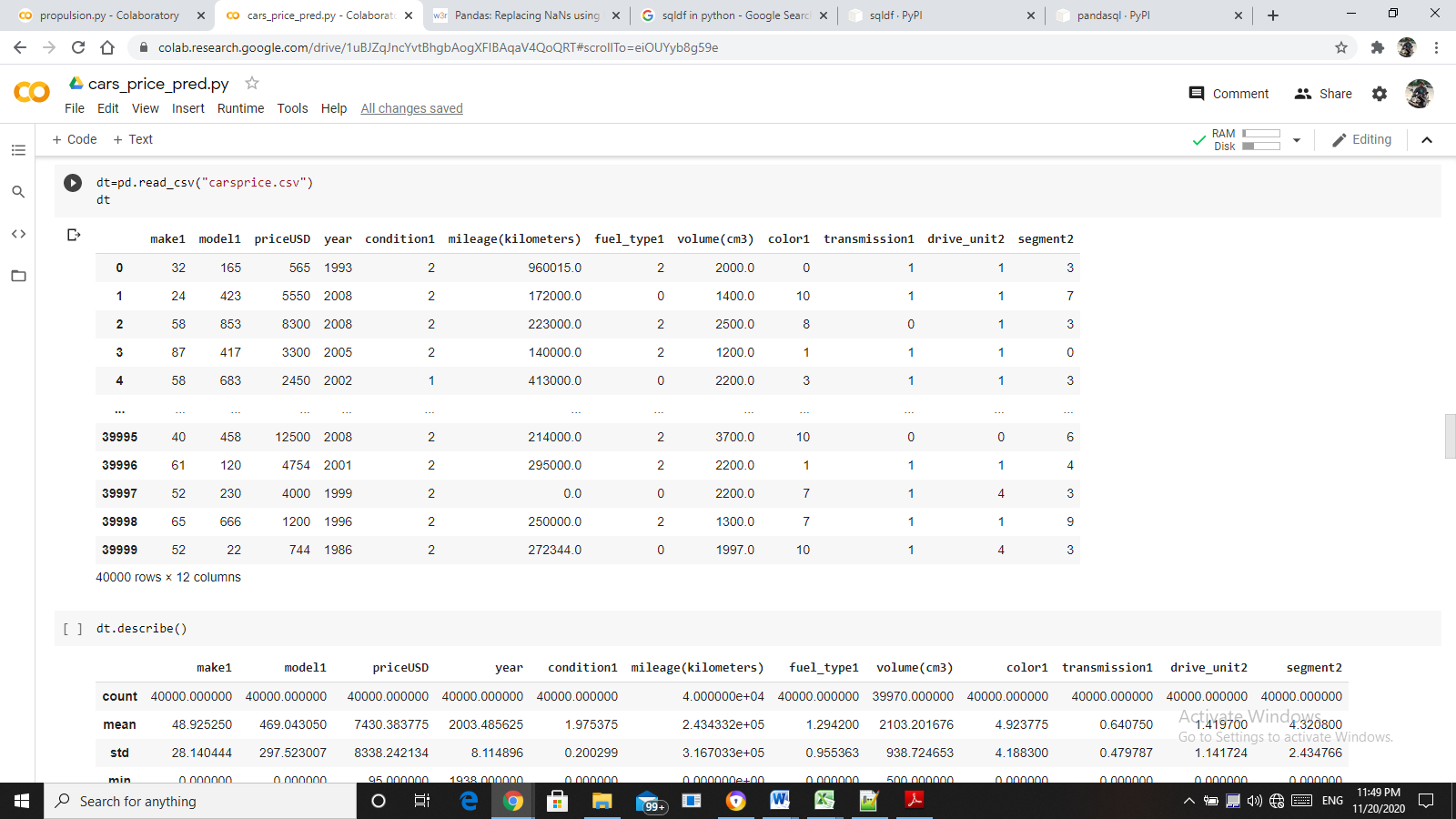
**PART-B**

READ THE DATA:

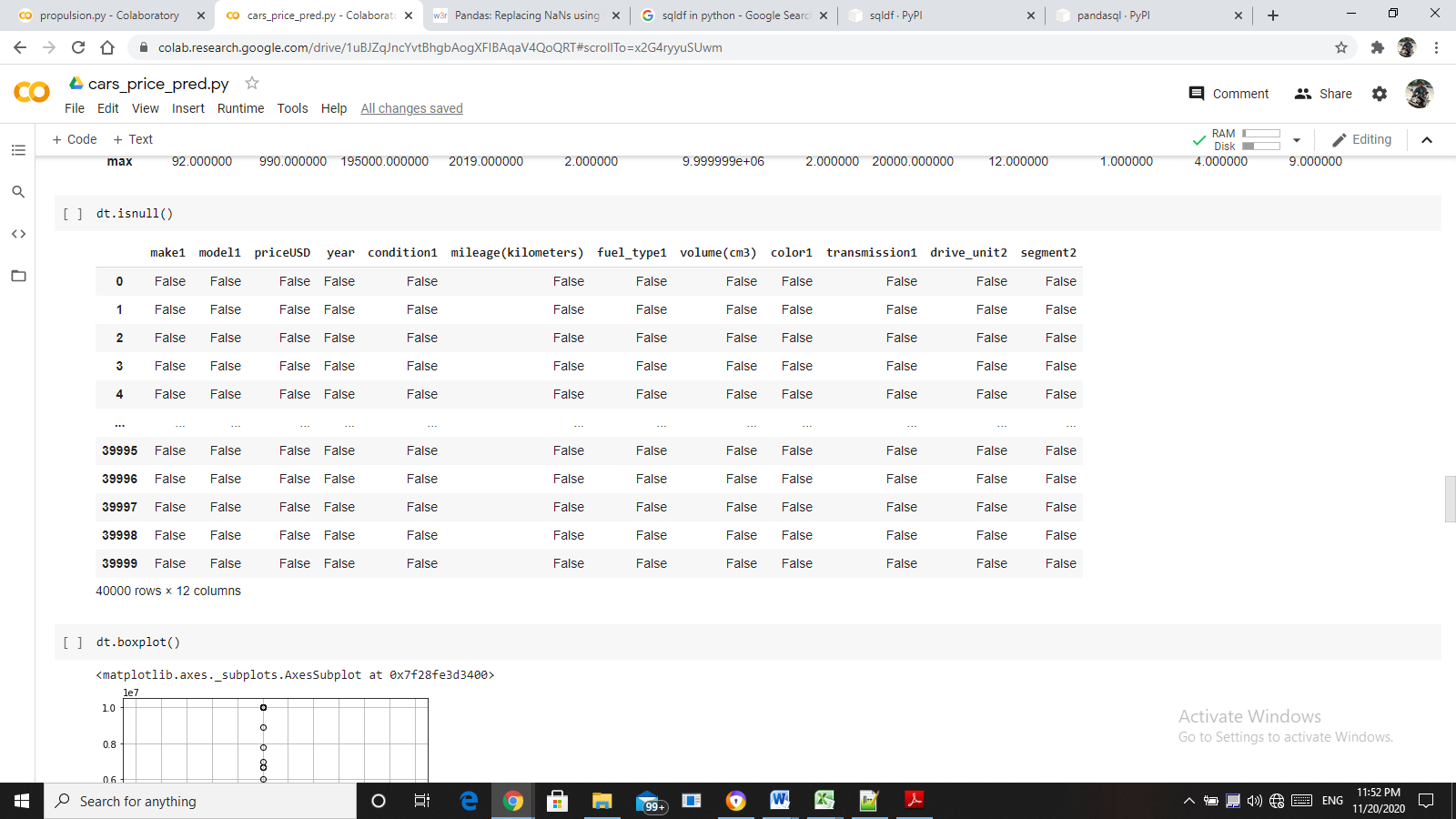


To train the model, the data should be encoded.

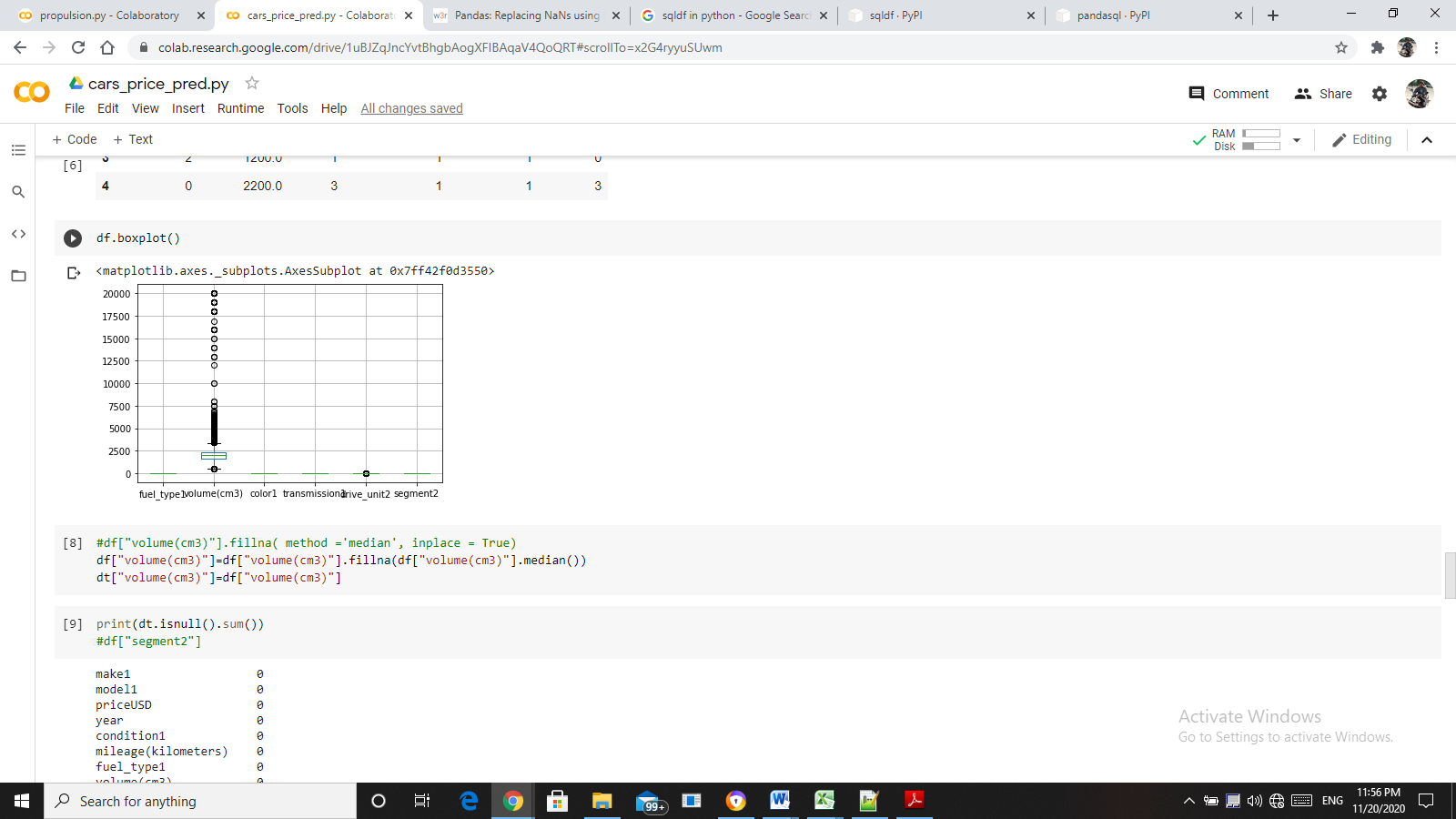
LABEL ENCODING THE DATA:



CHECKING FOR NULL VALUES:

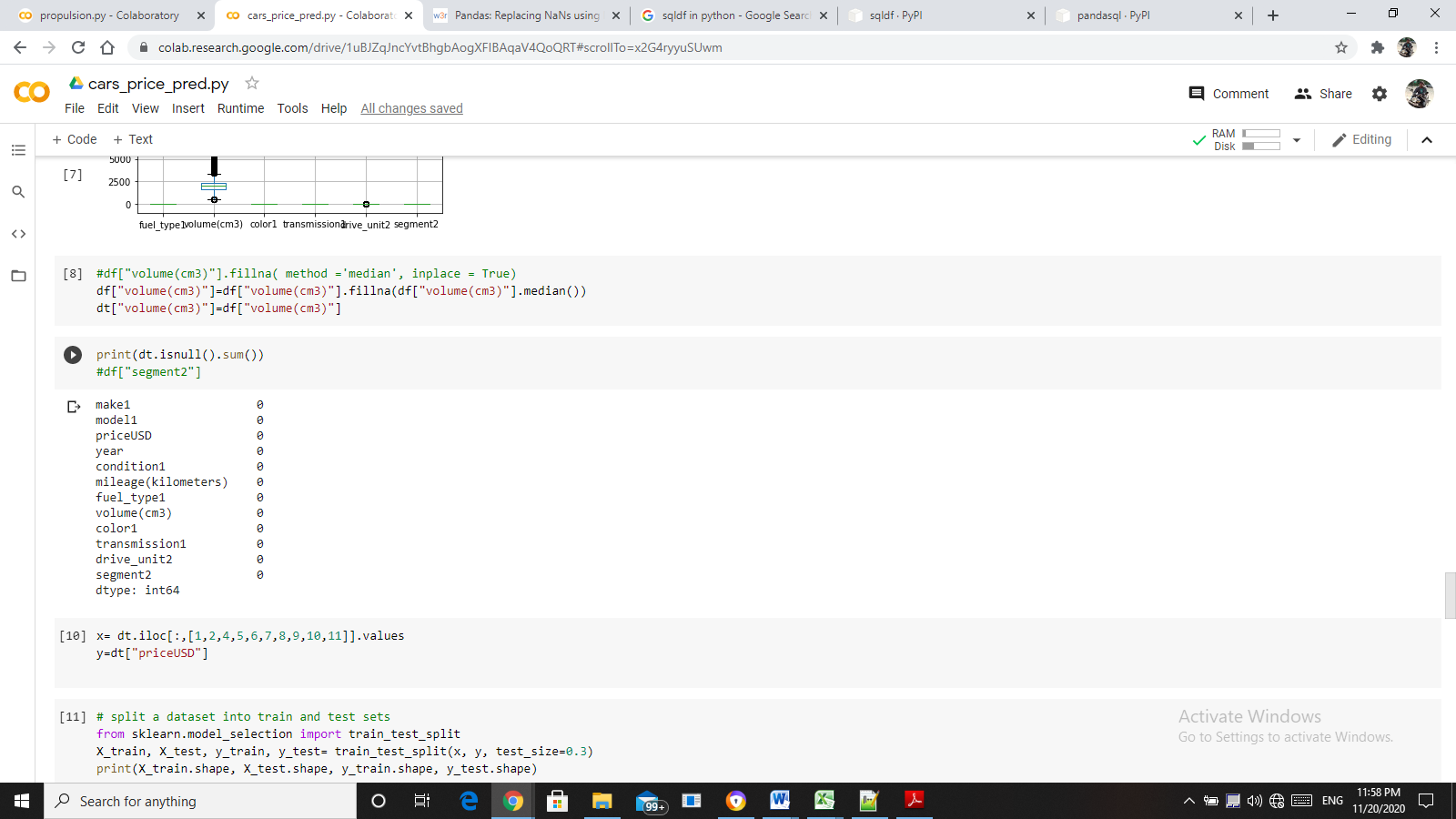


In the column volume, there are 30 null values. It should be treated. So first check for outliers so that the null values could be treated.



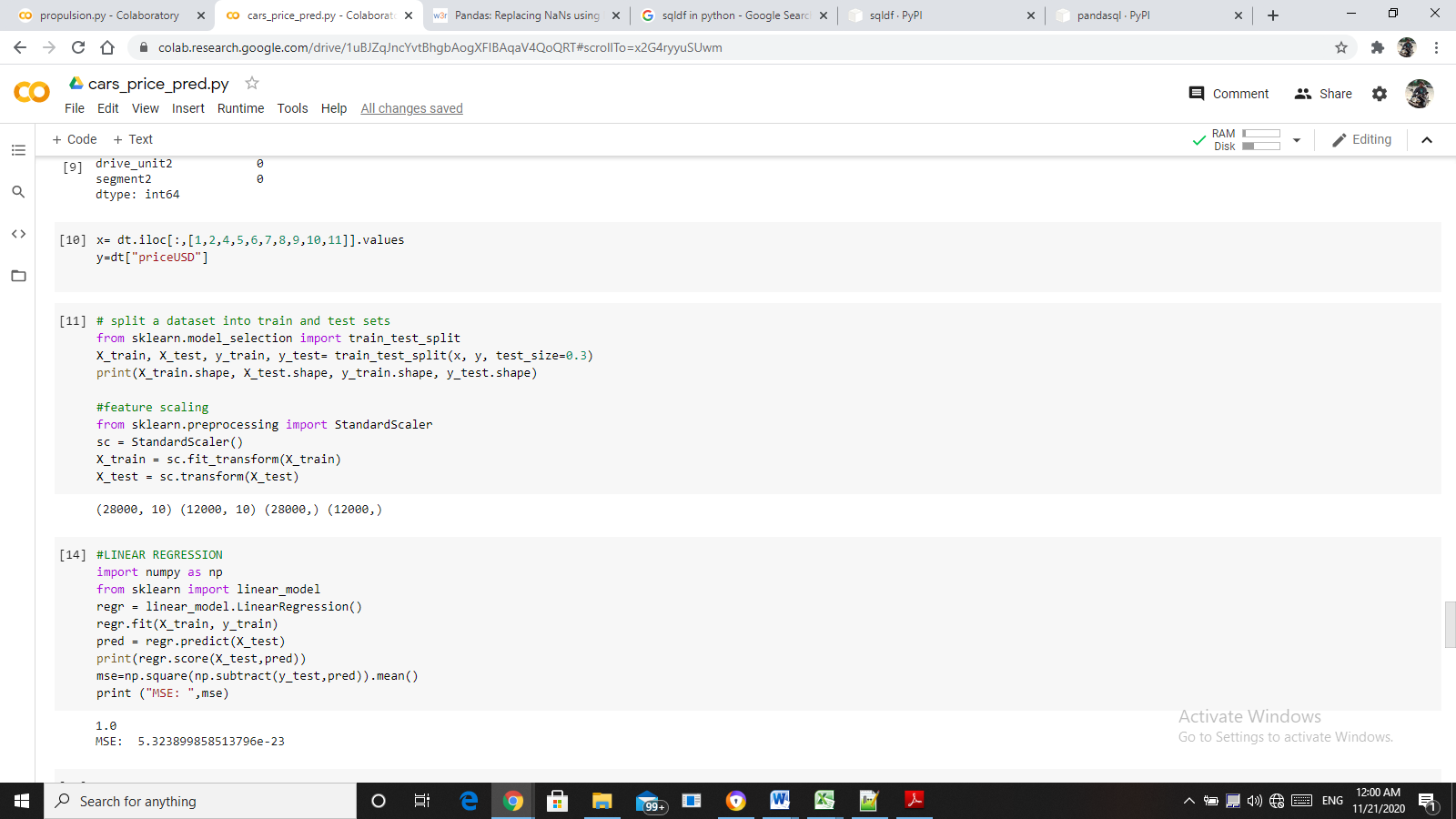
The outliers are more in the column volume. So it can’t be treated with mean. It could be treated with median.

After treating the null values,



As per image, there are no null values.

SPLITING THE DATA INTO TRANING AND TESTING SETS



The data is divided into 70% for training and 30 for testing.

**RESULTS FOR GT COMPRESSOR DECAY STATE CO-EFFICIENT**

MODEL NAME: LINEAR REGRESSION

R-square value: 1.0

Mean Square Error: 5.323899858513796e-23

MODEL NAME: DECISION TREE REGRESSOR

R-square value: 1.0

Mean Square Error: 3273.967833333333

MODEL NAME: RANDOM FOREST REGRESSOR

R-square value: 1.0

Mean Square Error: 9456773.989658201

MODEL NAME: LINEAR SUPPORT VECTOR REGRESSOR

R-square value: 1.0

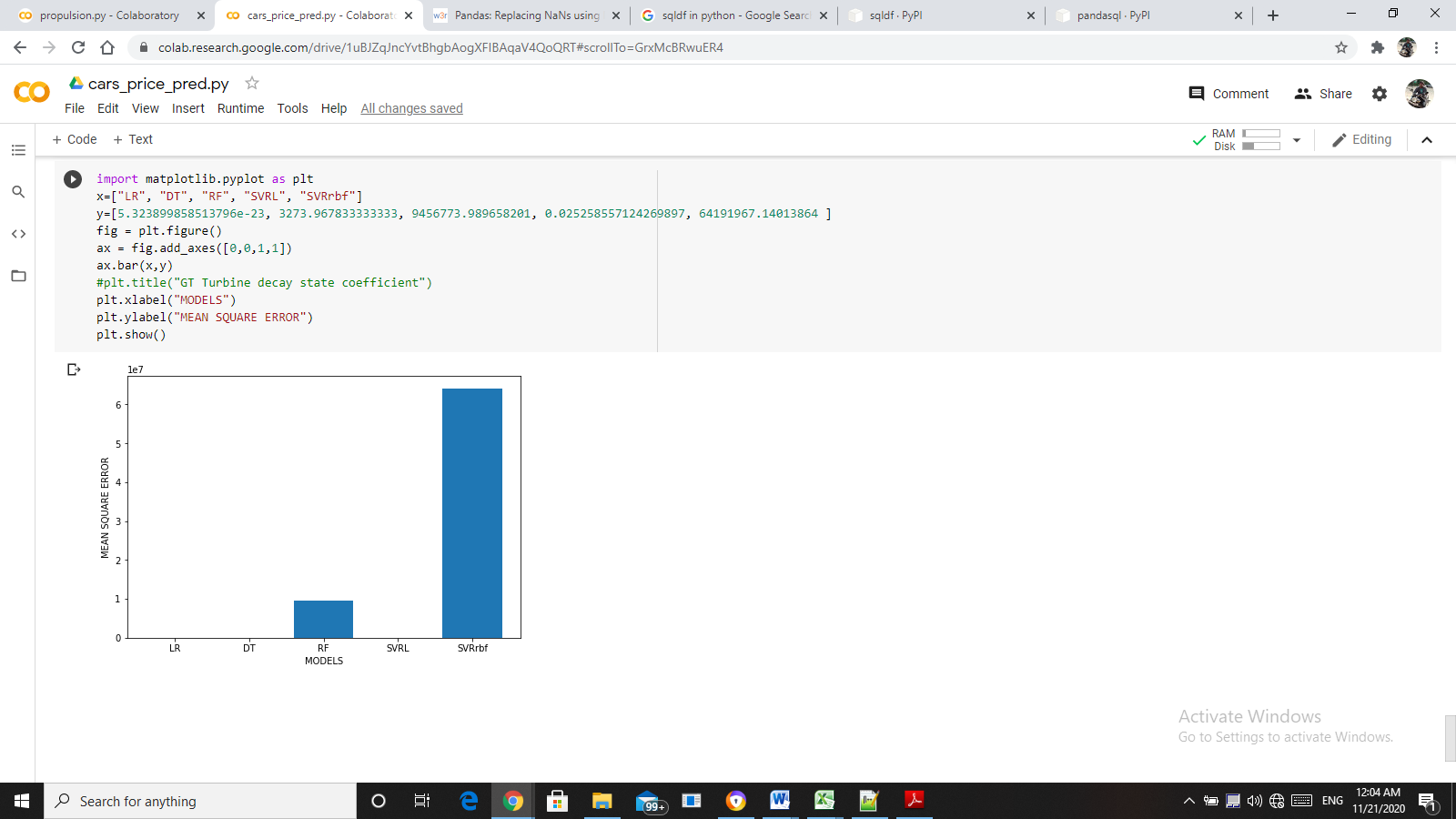
Mean Square Error: 0.025258557124269897

MODEL NAME: rbf SUPPORT VECTOR REGRESSOR

R-square value: 1.0

Mean Square Error: 64191967.14013864

**PLOTTING THE MSE VALUES:**



**INFERENCE:**

While comparing the mean square errors of the models, the mean square error of the LINEAR SUPPORT VECTOR REGRESSOR is less. Hence, the linear support vector regression model is the best fit.