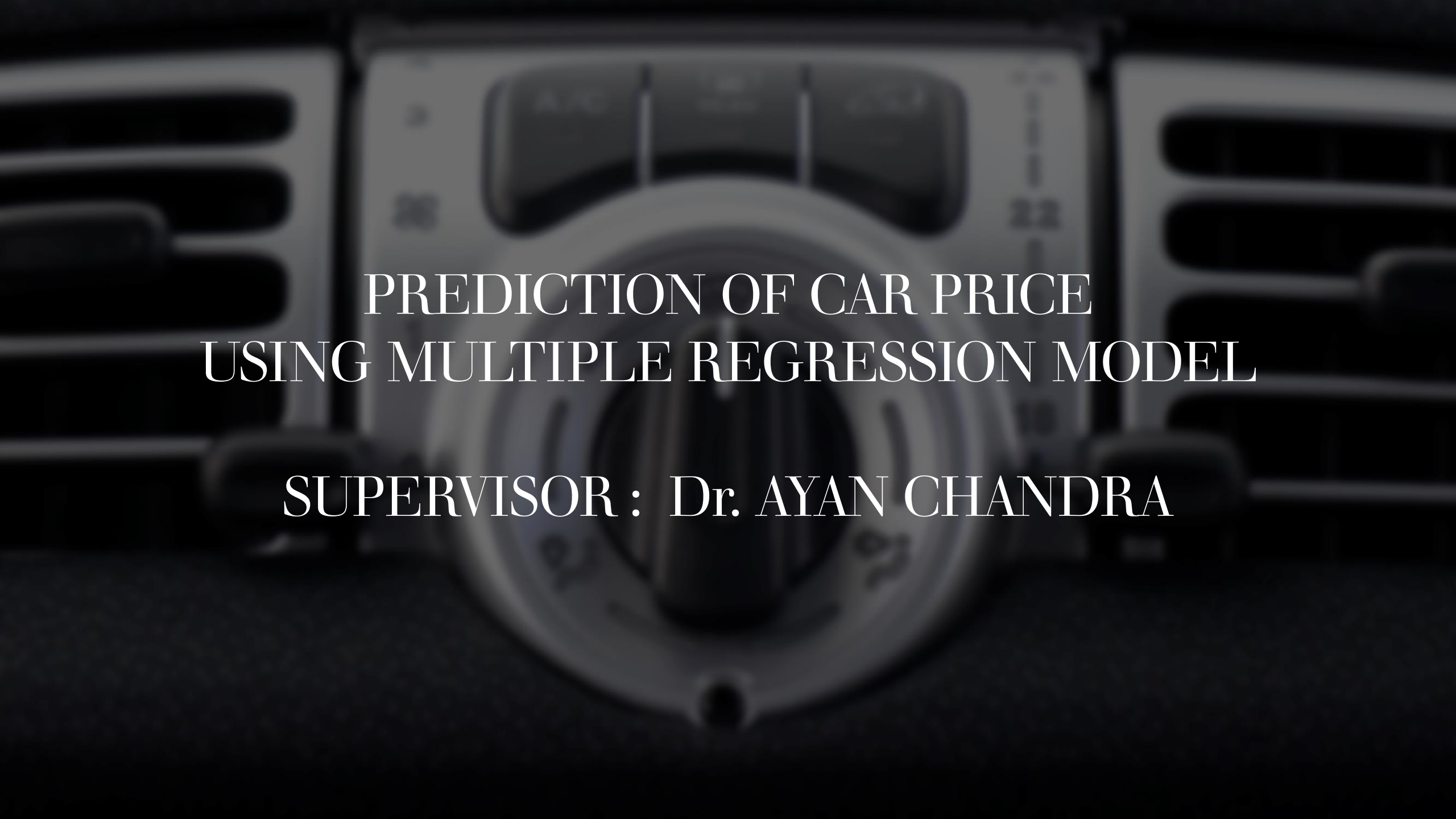
DISSERTATION

ARPAN SAMANTA
SEM 6, 425
REG. NO. A01-1112-0849-20
DEPARTMENT OF STATISTICS
ST. XAVIER'S COLLEGE, KOLKATA

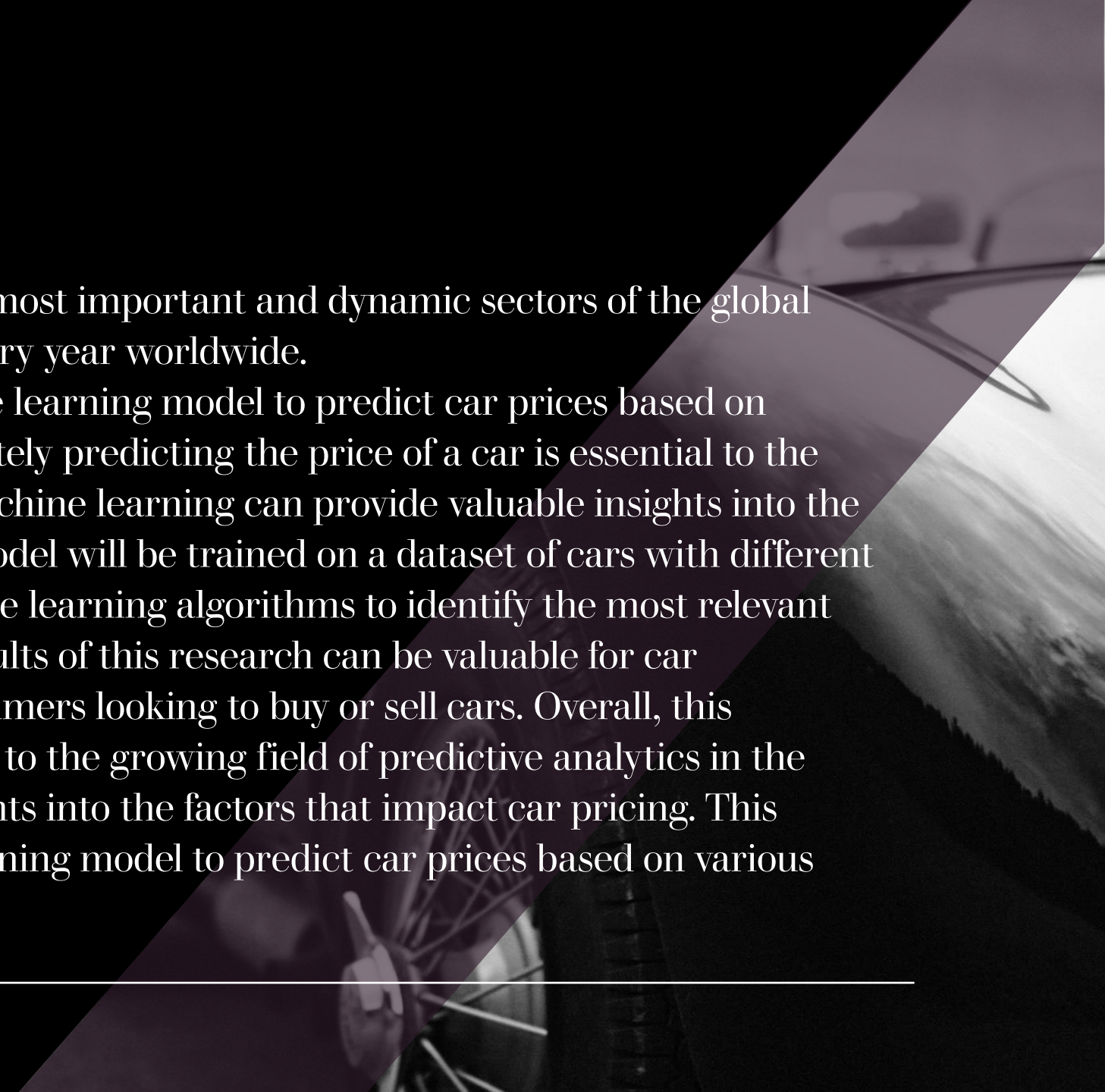


PREDICTION OF CAR PRICE USING MULTIPLE REGRESSION MODEL

SUPERVISOR : Dr. AYAN CHANDRA

INTRODUCTION

The automotive industry is one of the most important and dynamic sectors of the global economy, with millions of cars sold every year worldwide.
This project aims to develop a machine learning model to predict car prices based on various numerical parameters. Accurately predicting the price of a car is essential to the automotive industry's success, and machine learning can provide valuable insights into the factors that impact car pricing. The model will be trained on a dataset of cars with different specifications and prices using machine learning algorithms to identify the most relevant features that affect car prices. The results of this research can be valuable for car manufacturers, dealerships, and consumers looking to buy or sell cars. Overall, this dissertation project aims to contribute to the growing field of predictive analytics in the automotive industry and provide insights into the factors that impact car pricing. This project aims to develop a machine learning model to predict car prices based on various numerical parameters.



OBJECTIVES

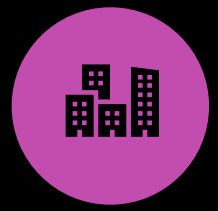
My project deals with a data set on 205 cars and their different numerical characteristics. We will see those characteristics in the following.

Sl no.	wheelbase	carlength	carwidth	carheight	curbweight	enginesize	boreratio	stroke	compressionratio	horsepower	peakrpm	citympg	highwaympg	Car price
1	88.6	168.8	64.1	48.8	2548	130	3.47	2.68	9	111	5000	21	27	13495
2	88.6	168.8	64.1	48.8	2548	130	3.47	2.68	9	111	5000	21	27	16500
3	94.5	171.2	65.5	52.4	2823	152	2.68	3.47	9	154	5000	19	26	16500
4	99.8	176.6	66.2	54.3	2337	109	3.19	3.4	10	102	5500	24	30	13950
5	99.4	176.6	66.4	54.3	2824	136	3.19	3.4	8	115	5500	18	22	17450
6	99.8	177.3	66.3	53.1	2507	136	3.19	3.4	8.5	110	5500	19	25	15250
7	105.8	192.7	71.4	55.7	2844	136	3.19	3.4	8.5	110	5500	19	25	17710
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
200	104.3	188.8	67.2	57.5	3157	130	3.62	3.15	7.5	162	5100	17	22	18950
201	109.1	188.8	68.9	55.5	2952	141	3.78	3.15	9.5	114	5400	23	28	16845
202	109.1	188.8	68.8	55.5	3049	141	3.78	3.15	8.7	160	5300	19	25	19045
203	109.1	188.8	68.9	55.5	3012	173	3.58	2.87	8.8	134	5500	18	23	21485
204	109.1	188.8	68.9	55.5	3217	145	3.01	3.4	23	106	4800	26	27	22470
205	109.1	188.8	68.9	55.5	3062	141	3.78	3.15	9.5	114	5400	19	25	22625

Working Dataset

CHARACTERISTICS

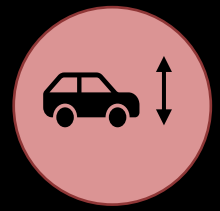
CAR PEAK RPM



MILEAGE IN CITY

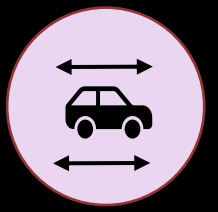
MILEAGE ON HIGHWAY

PRICE OF CAR



HEIGHT OF CAR

WIDTH OF CAR

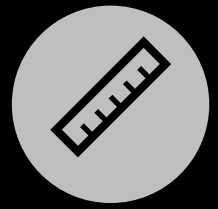


LENGTH OF CAR

WEIGHT OF CAR

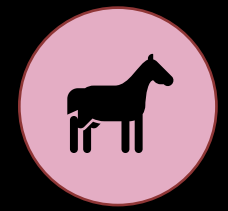
SIZE OF CAR

BORERATIO OF CAR

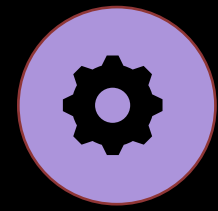


STROKE INSIDE THE ENGINE

COMPRESSION RATIO OF CAR



HORSE POWER



WHEELBASE OF CAR

My primary objective is to determine if there is a correlation between the price of a car and its other parameters. I will also evaluate the goodness of fit, identify any outliers, assess the individual impact of each parameter on the car price, and examine the residual plot to determine it’s randomness. Ultimately, I aim to create a regression equation that can predict the price of a car based on its useful characteristics.

LIMITATIONS

Note that, it is not necessary that only these parameters will affect the price of the car

There exist some categorical parameters such as fuel type, number of doors, drive wheel location, etc. in the data set that has the ability to put some effect on the car price

However, since our response is a continuous random variable so in order to simplify our work we're only working with the numerical characteristics

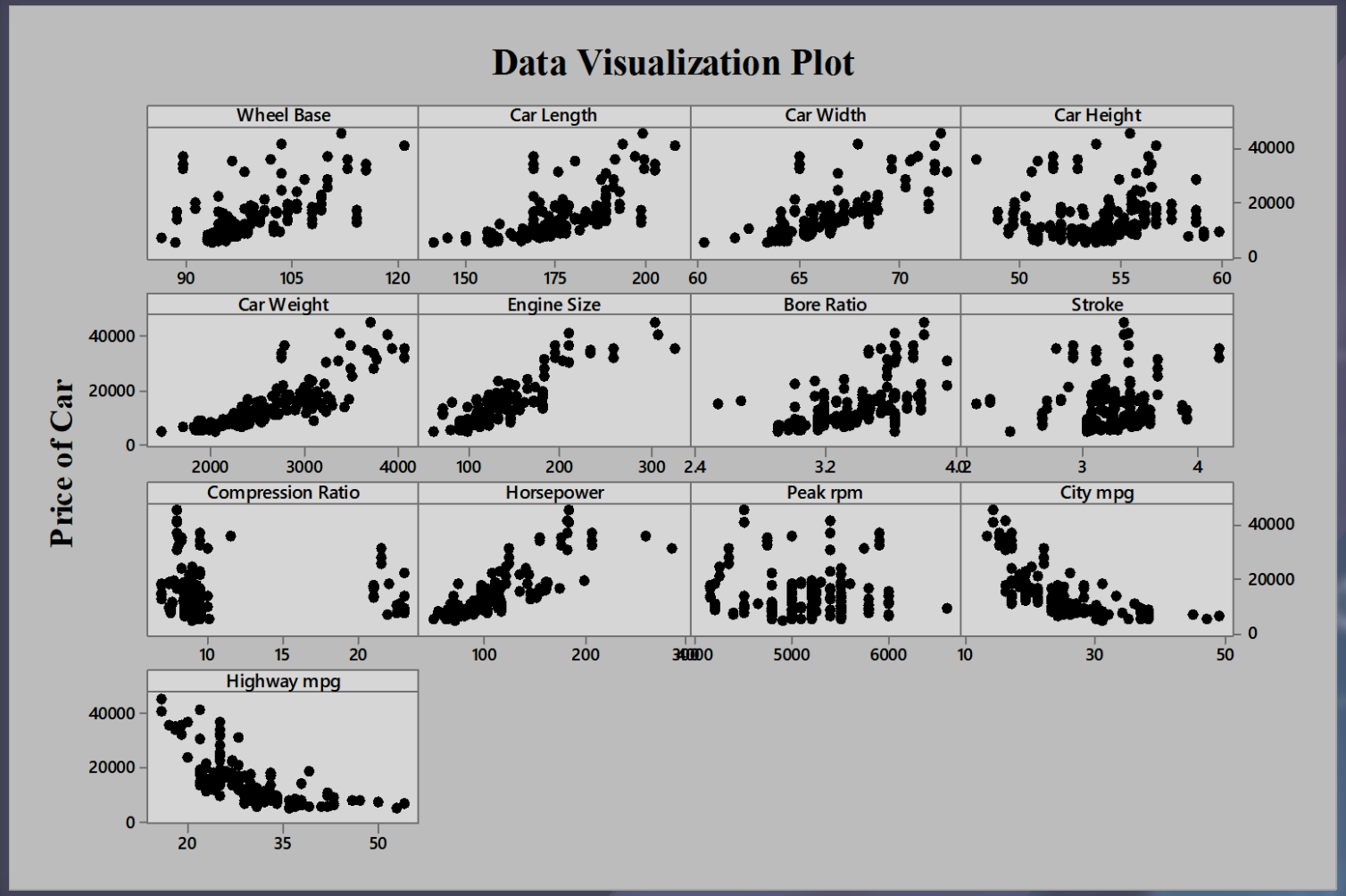
DATA VISUALIZATION

Data visualization is a technique where we represent the initial data in the form of some common graphics, charts, plots, animation, etc. and roughly try to analyze

Since, we're working with multivariate data, a singular plot won't be enough to analyze

Since there are 13 numeric characteristics in the data set apart from the car price, we plot 13 scatter plots.

Graph | 1



Observing these plots we see that almost every parameter is more or less affecting the car price. The influence of wheelbase, car length, car width, car weight, bore ratio, and horsepower on the price individually is indicating a positive correlation. Whereas, city mpg and highway mpg are indicating a negative correlation. However, the plot of car height, Stroke, compression ratio, and peak rpm are not showing any significant pattern properly. So, we drop them and continue our analysis with the other parameter for the shake of simplicity.

DATA DESCRIPTION

The data set I am working with is provided by Manish Kumar on Kaggle.com. The data set provides information on 205 cars and their various characteristics. Among these characteristics some are numeric and some are categorical in nature. However, as previously discussed I am working with numeric characteristics only. And using those characteristics as predictors that are showing some significant effect on the car price in the data visualization section.

In order to fulfill my purpose of simplifying the model mathematically, I assign different variables to each of those chosen car parameters.

X_1 = Price of car

X_2 = Weelbase of car

X_3 = Length of car

X_4 = Width of car

X_5 = Height of car

X_6 = The weight of a car without occupants or baggage.

X_7 = Horsepower

X_8 = Boreratio of car

X_9 = Mileage in city

X_{10} = Mileage on highway

From now on, instead of calling a particular characteristic by it's name, we'll simply denote it by it's corresponding variable.

Moreover, for multiple linear regression analysis, we'll call X_1 i.e. Price of car as *the response* and the rest of the X_j's, j = 2 (1) 10 as our *predictors*. Clearly, there're 9 predictors corresponding to a single response, for each of which data on 205 cars are present.

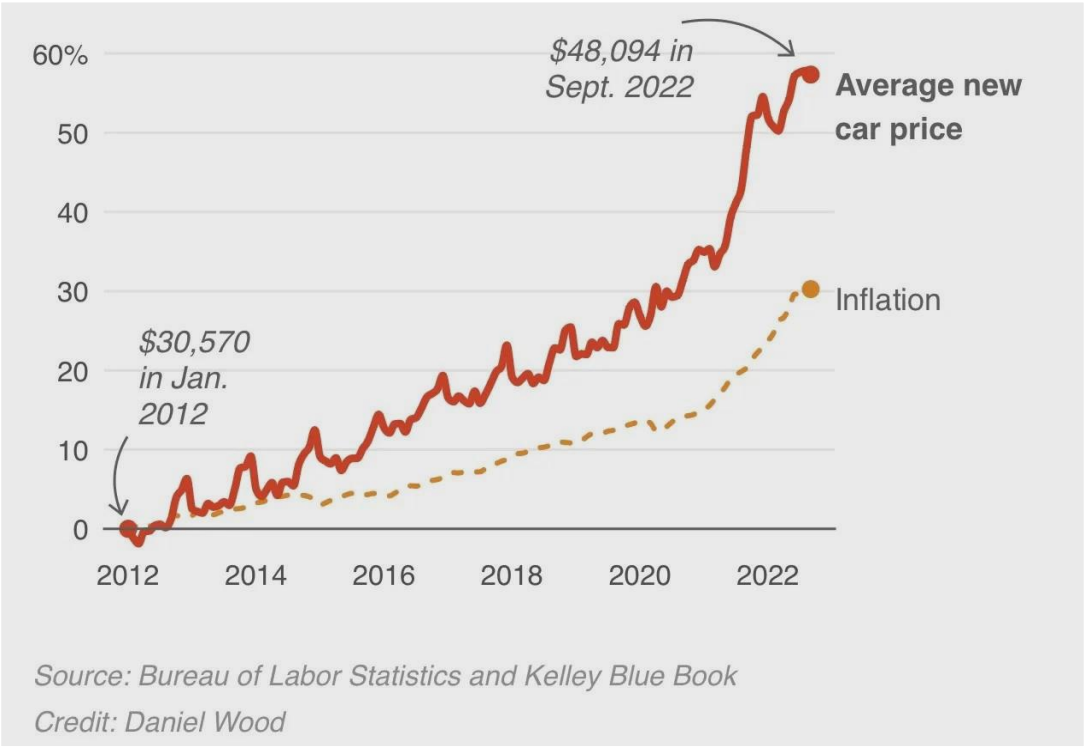
CHOICE OF RESPONSE AND ITS SIGNIFICANCE

Selecting the response variable in a multiple regression analysis is crucial, and it should be chosen with care. The response variable should have some potential effects of individual predictors on it. The information suggests that from data visualization, it is evident that the 9 predictors are influencing the car price individually, and they are expected to impact the price jointly. Additionally, practical knowledge also indicates that the price of a car is determined by several characteristics.

Car prices can significantly impact the global economy. The price of cars can directly affect consumer spending, inflation, and economic growth as they are one of the most significant purchases in the common population. Increasing car prices can lead to a higher cost of transportation, which can result in higher prices of goods and services, leading to reduced consumer spending and lower sales for businesses. The automotive industry is a crucial source of income for many countries, such as Japan, Germany, the United States, South Korea, China. Any harm to the sector can cause job losses, reduced tax revenue, and a slowdown in these economic growths.


Let’s see the Kelley Blue Book and the Bureau Of Labor Statistics analysis that states that states that the average new car price at the beginning of 2023 is $49,388 in the U.S

Graph | 2



The graph shows the average price of a new car in the US from January 2012 to September 2022. The vertical axis shows the price of the car, while the horizontal axis shows the year and month.

As we can see from the graph, the average new car price has been steadily increasing over the years. In January 2012, the average new car price was around $30,570. By September 2022, the average price had increased to around $48,094.

We can also see that the rate of increase in new car prices has been much faster than inflation since 2014. The graph shows a steep upward trend in prices from around mid-2014, with a noticeable acceleration in 2021 and 2022. This acceleration was due to a decline in public transportation and supply shortages.

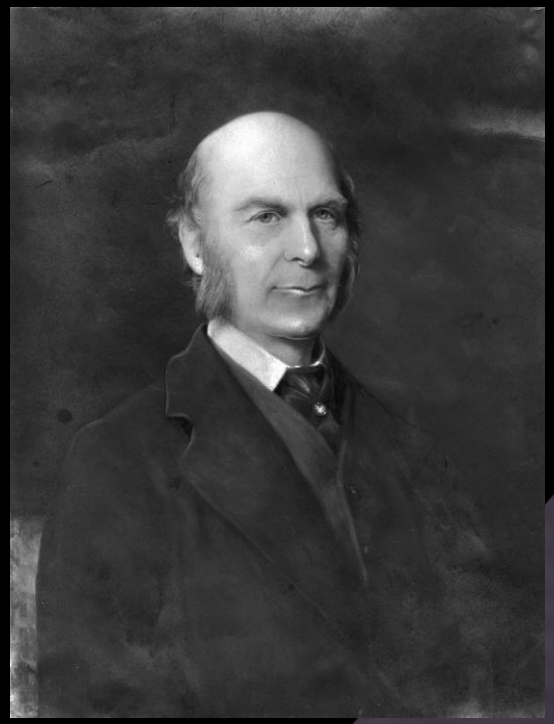
Overall, the graph shows a clear and steady increase in the average price of a new car in the US over the past decade, with significant spikes in recent years.

HISTORY OF MULTIPLE LINEAR REGRESSION

The history of multiple linear regression dates back to the early 19th century when French mathematician Adrien-Marie Legendre proposed a method to fit a line to a set of data points in 1805. This method was later used by the British scientist Sir Francis Galton for least squares multiple regression.

- In the early 20th century, pioneers of modern statistics such as Karl Pearson and Ronald Fisher developed the statistical theory of multiple regression, which included the concepts of regression coefficients, standard errors, and hypothesis testing. With the introduction of computer technology in the mid-20th century, multiple linear regression analysis became possible for large data sets. This has led to the widespread use of regression analysis in various fields.
- In recent years, the combination of multiple linear regression with other statistical and computational techniques has led to the development of advanced predictive models capable of handling large amounts of data and making accurate predictions. It is worth noting that in addition to simple linear and multiple linear regression, more advanced regression techniques have been developed in response to increasing data and computing power.

Adrien-Marie Legendre



Sir Francis Galton

Karl Pearson

Ronald Fisher

A statistical model which is linear in it's parameters is called a linear model. In multiple linear regression the relation between response and multiple predictors is represented by a suitable linear equation, say,

- $X_1 = \beta_1 + \beta_2 * X_2 + \beta_3 * X_3 + \ldots\ldots\ldots + \beta_p * X_p + \varepsilon$

Where, X_1 is the response, X_j, j = 2 (1) p are the (p-1) predictors. Whereas ε is the error term and β_j , j = 2 (1) p are the (p-1) regression coefficients and β_1 is the intercept term.

The goal of the model is to determine the value of the coefficients that provide best fit to the data i.e. gives minimum error. This is typically done using a method called the method of least squares in which we minimize the sum of squares of errors with respect to the coefficients and thus solving various equations we get the values of βj's.

Multiple linear regression can be used for a variety of applications, such as predicting sales based on advertising spend, analyzing the relationship between multiple demographic variables and health outcomes, or predicting crop yields based on weather data.

However, it is important to note that MLR assumes that the relationship between the dependent and independent variables is linear and that there is no multicollinearity (high correlation) between the independent variables. If these assumptions are not met, the results of the analysis may be inaccurate.

MODEL AND ASSUMPTIONS

Let us consider the following linear regression model with n (= 205) sample observations and (p-1) (= 9) explanatory variables.

$$X_{1i} = b_1 + b_2X_{2i} + b_3X_{3i} \ldots\ldots\ldots\ldots\ldots\ldots + b_pX_{pi} + u_i \qquad i = 1\,(1)\,n$$

We re-write it in the matrix format as,

- $\underline{\boldsymbol{Y}} = \mathbf{X}\underline{\boldsymbol{\beta}} + \underline{\boldsymbol{u}}$

With the assumptions,

- $\underline{u} \sim N_n(\underline{0}, \sigma^2 I_n)$
- **Design matrix X is non – stochastic.**
- **Rank (X) = p < n**

Where $\boldsymbol{Y}$ is the response vector of order $n\times1$, X is the design matrix of order $n\times p$, $\boldsymbol{\beta}$ is the parameter vector of order $p\times1$ and $\boldsymbol{u}$ is the error vector of order $n\times1$.

SOFTWARES USED

We will use Minitab – 17 and R – programming for our necessary computation purpose

REGRESSION ANALYSIS

Analyzing the data on 205 cars in the above-mentioned statistical softwares we get, the coefficient of determination as,

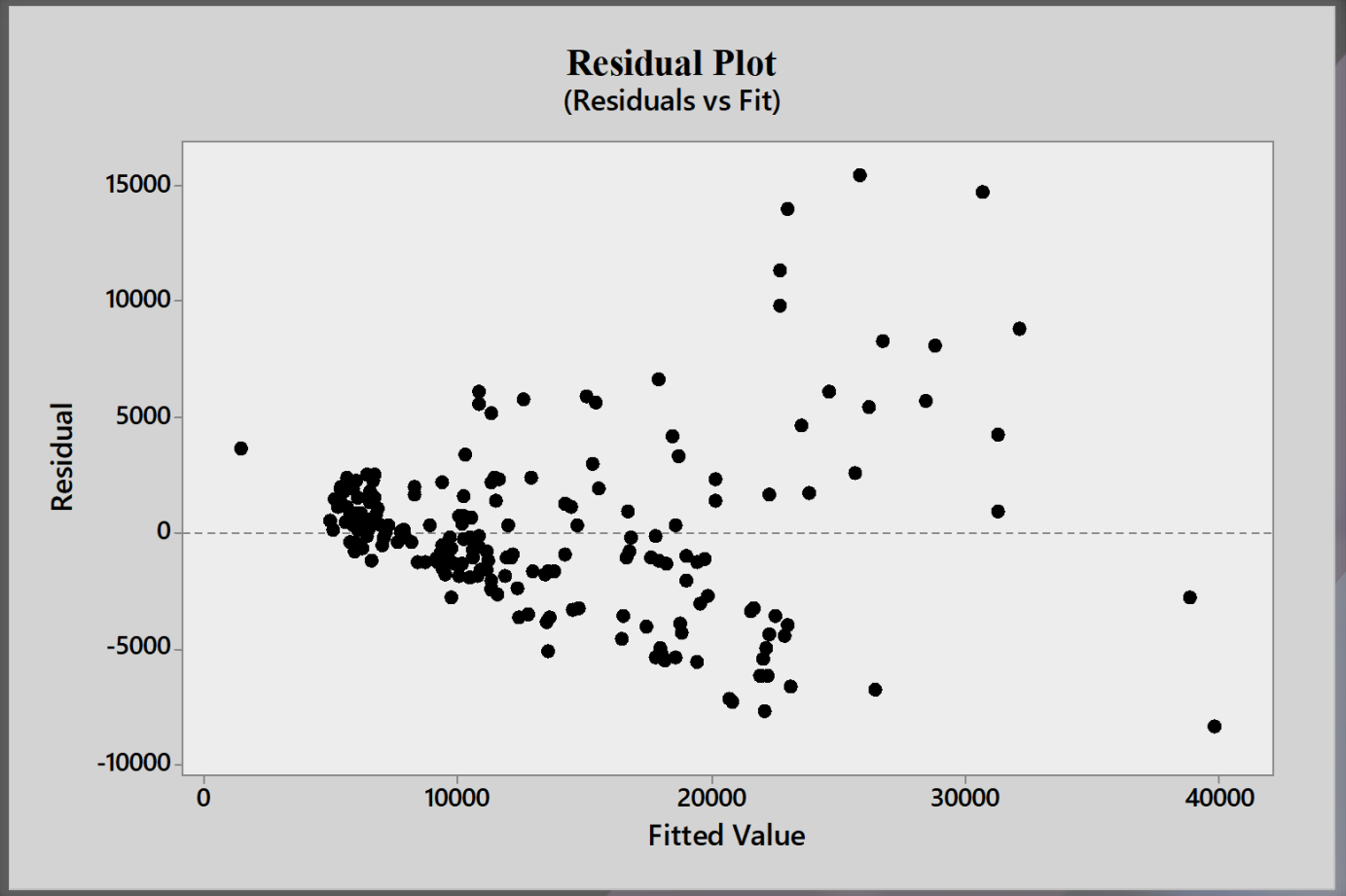
$$R^2 = 78.72\%$$

And, the multiple regression model after estimating the model parameters is given by,

$$X_{1.2345678910} = -54517 + 99 * X_2 - 71.6 * X_3 + 568 * X_4 + 58 * X_5 + 7.15 * X_6 - 839 * X_7 + 102.2 * X_8 + 90 * X_9 + 51 * X_{10}$$

Where, $X_{1.2345678910}$ is the predictor of X_1 based on the predictors X_j, j = 2 (1) 10.

Graph | 3



We would also like to check if there is any pattern in the residual plot. For that, we plot the residual values on y – axis taking the fitted response values of X_1 on the x – axis , thus we get the residual plot.

From the residual plot, it is evident that the residuals are scattered around 0 and not much of a pattern can be observed. However, there may exist some outlier / influencial points that has to be omitted.

PRESENCE OF OUTLIERS

Dennis Cook (1977) introduced a distance measure for commonly used estimates to study the influence of a data point when performing least squares regression analysis. Data points with large residuals and/or high leverage may distort the outcome and accuracy of a regression. **Cook's distance** measures the effect of deleting a given observation. Points with a large Cook's distance are considered to merit closer examination of the analysis.

Cook's distance Di of observation i (i = 1,2.....205) is defined as the sum of all the changes in the regression model when observation i is removed from it.

$$D_i = \sum_{i=0}^{n} \frac{\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{ps^2}$$

Where,

$\hat{y}_{j(i)}$ is the fitted response value obtained when excluding i ,

p is the number of terms in the model + 1

s^2 is the means square error of the regression model.

CALCULATIONS

Using MINITAB, the Cook's Distance have been obtained for all the observations. The observations with comparatively high value (close to 1) are considered as influential points which distorts the regression equation. We remove such observations and re – fit the model.

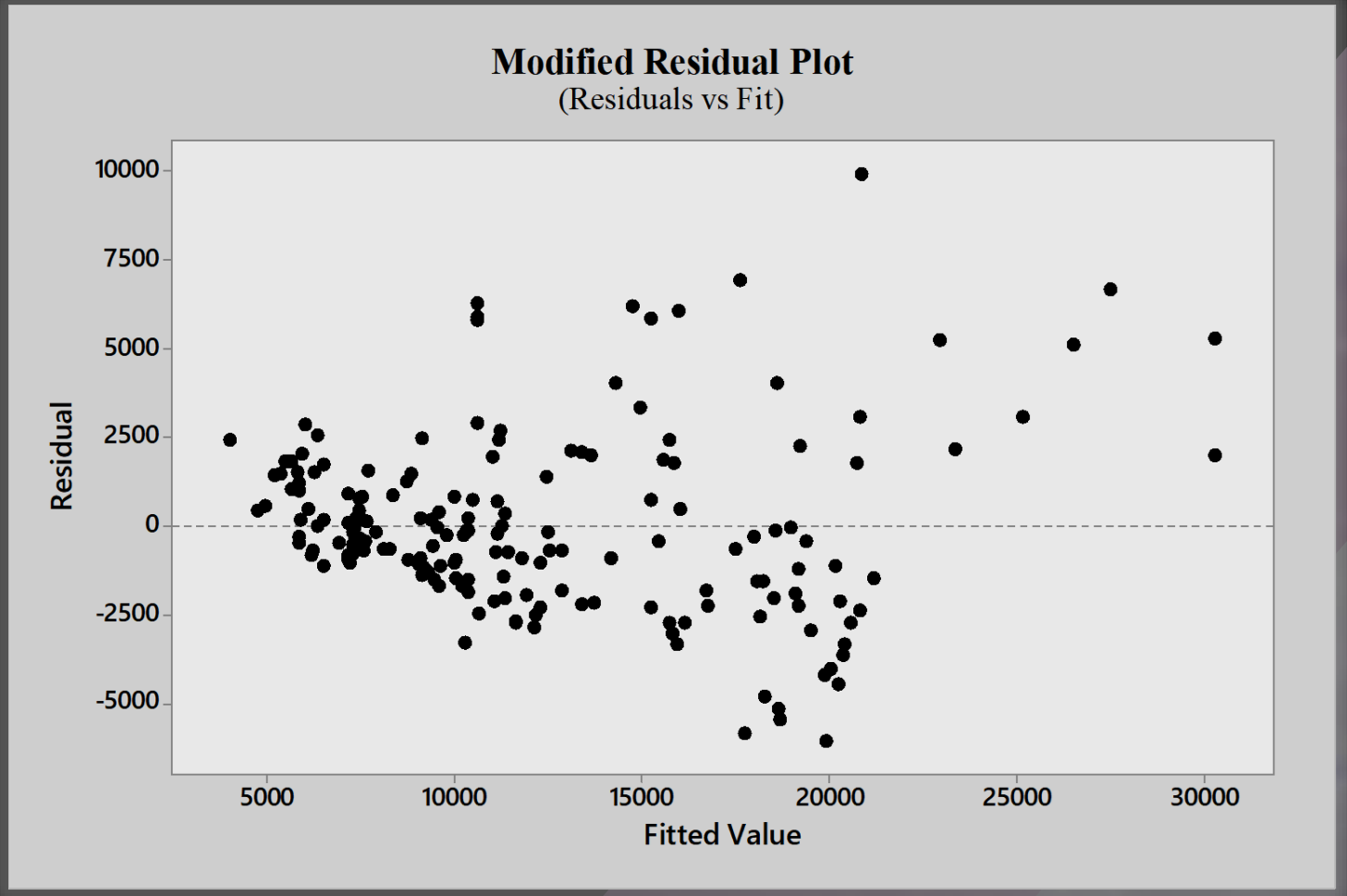
After removing the unusual points we get our coefficient of determination as,

$R^2 = 81.81\ \%$

Corresponding to the model,

$$X_{1\ 234567891} = -37924.685 + 177.460 * X_2 - 96.192 * X_3 + 462.614 * X_4 - 48.681 * X_5 + 8.932 * X_6 - 2284.362 * X_7 \_ 44.842 * X_8 - 127.425 * X_9 + 172.465 * X_{10}$$

Graph | 4



Also, the new residual plot is given by,

We see that the scatteredness of the plot in the y – direction is much less in the new plot than in the previous. This is because we've omitted many influential points. However, the plot doesn't show any pattern that is observable.

SIGNIFICANT PREDICTORS

- We'll now try to rebuild the model with the most useful predictors, in order to make the model more smooth
- For this, we need the different p-values corresponding to different predictors that we've obtained while fitting the outlier-free model in R

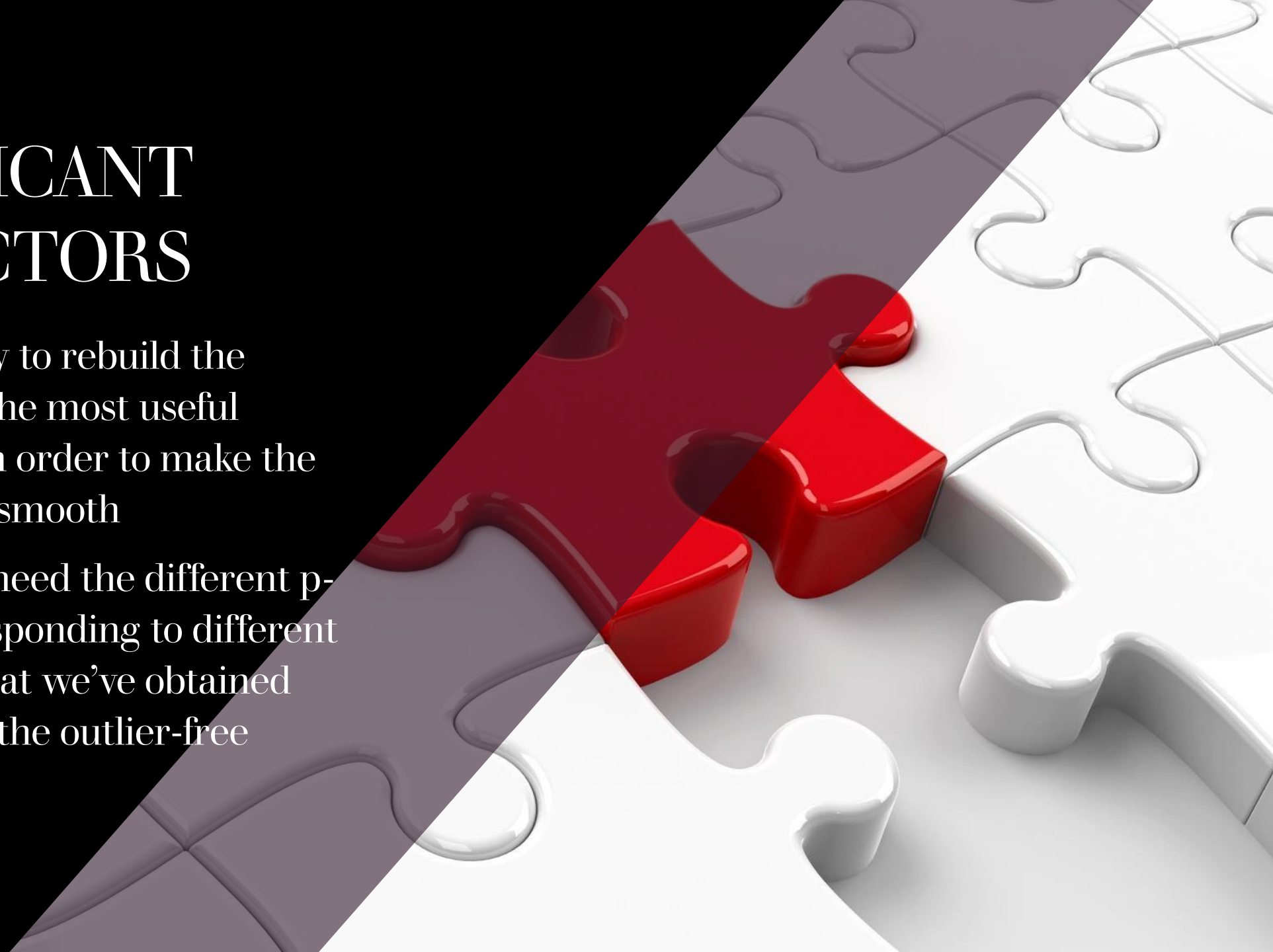


Table | 1

Predictor	p - value
X2	0.05065
X3	0.0442
X4	0.03697
X5	0.6767
X6	$4.88 * 10^{-10}$
X7	0.01782
X8	0.00108
X9	0.37822
X10	0.21114

Lesser the p–value more significant the predictor. At 5% level of signicance, we will choose those predictors whose p–value is lesser than 0.05. That is our significant predictors are X_3 , X_4 ,X_6 , X_7 , X_8 . Among these, X_6 is the best significant predictor.

MULTICOLLINEARITY

Multicollinearity is a very common problem in multiple linear regression analysis. It might appear among the predictor variables that there exists a high correlation (more than 0.7 or 0.8) between one or more pairs of independent variables. It leads to unreliable or unstable estimates of regression coefficients affecting the accuracy of the model's prediction. In order to check for multicollinearity we've to first check the correlation in between the independent predictors. We give the coefficient of determination, r^2 by the following table.

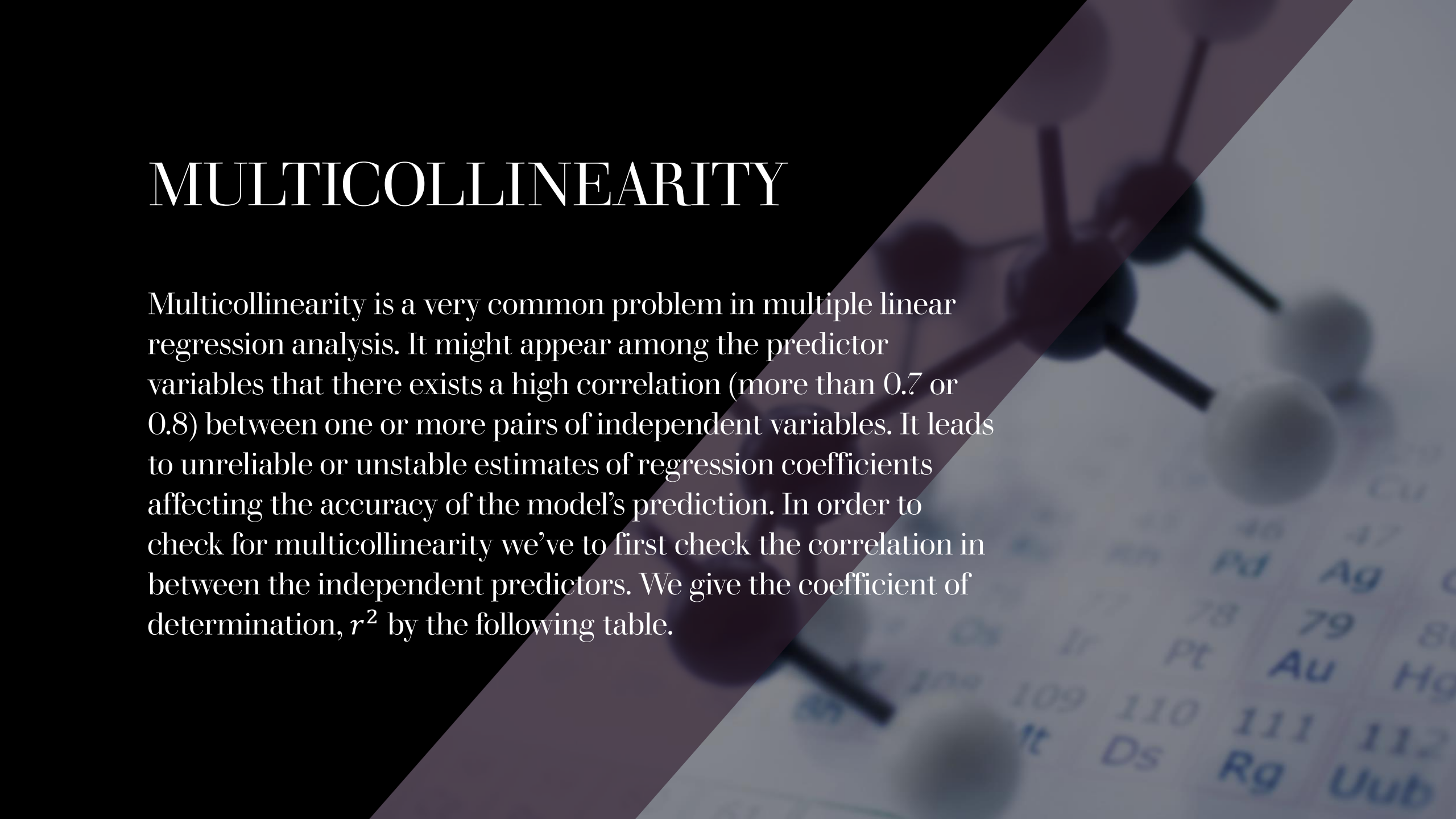


Table | 2

r^2	X_3	X_4	X_6	X_7	X_8
X_3	1	0.8524	0.8899	0.6096	0.6014
X_4	0.8524	1	0.8696	0.5344	0.6239
X_6	0.8899	0.8696	1	0.6398	0.7676
X_7	0.6096	0.5344	0.6398	1	0.5139
X_8	0.6014	0.6239	0.7676	0.5139	1

We see there is a high correlation between X_3 and X_4 as well as between X_3, X_6 and X_4, X_6. So in order to overcome this problem of multicollinearity we omit the predictors X_3 and X_4.

And thus we get 3 predictors, X_6, X_7, X_8 corresponding to a single response X_1 in our ultimate model. With these information we form our ultimate modified model.

MODIFIED REGRESSION MODEL

The coefficient of determination for our ultimate model is,

$R^2 = 80.43\ \%$

This value may look a negligible amount of smaller than the previous coefficient of determination, but the model is much more stable due to the absence of multicollinearity. The model is now given by,

$$X_{1.678} = -7627.2234 + 10.2214 * X_6 - 2786.2447 * X_7 + 33.3958 * X_8$$

Where, $X_{1.678}$ is the predictor of X_1 in the new model based on X_6, X_7, X_8.

i.e. the model is expressed as,

$$Price\ of\ Car = -7627.2234 + 10.2214 * Car\ weight - 2786.2447 * Bore\ ratio + 33.3958 * Horse\ power$$

The mean absolute percentage error is coming out to be,

MAPE = 15.09 %

The MAPE value is showing a moderately lower value i.e percentage error between the predicted value and the actual value of the response variable is moderately low. This indicates a better fit of the model as it means that the predicted values are closer to the actual values

The mean sum of square of errors is coming out to be,

MSE = 6984000

Whereas,

Root MSE = 2642.726

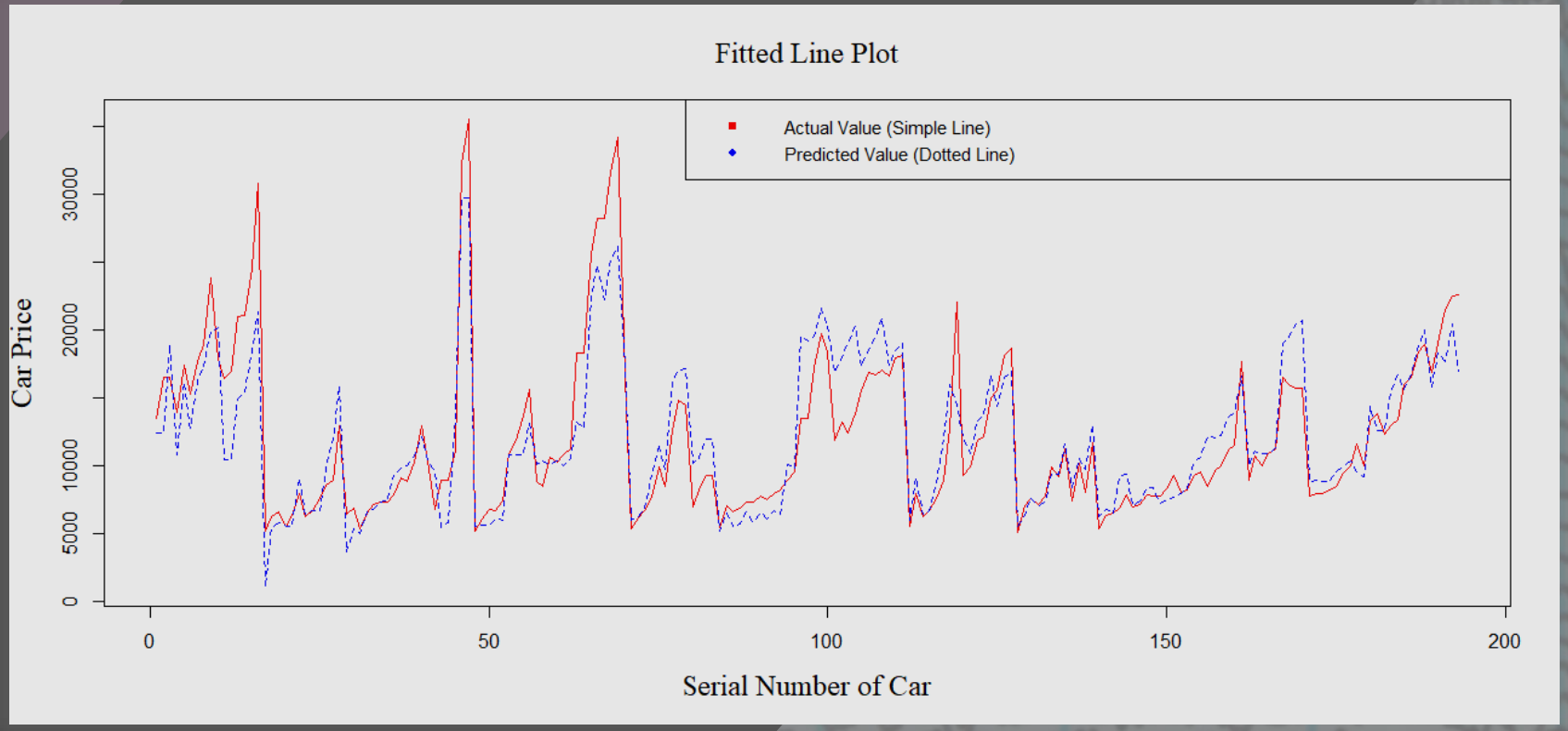
Considering, the price of a car can vary in a high scale of numbers, we can consider it as a moderately good fit

Measure of Accuracy

PREDICTION

- Putting the value of the corresponding predictors we can now easily predict the price of a given car
- For prediction purposes we'ld like to see the fitted line plot where we've fitted the actual value of the car and its predicted value using this model in the same graph taking the car serial number on x – axis and car price on y – axis

Graph 5



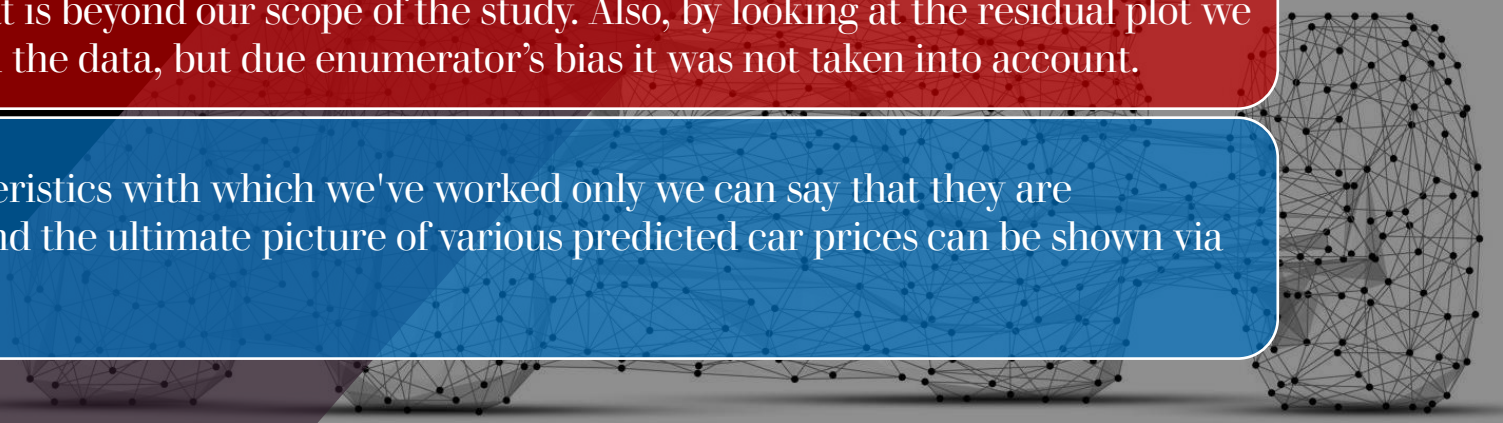
From the plot, it is observable that though the fit isn't completely accurate, both plots are showing a similar kind of pattern. In the long run data analysis, this information might be helpful.

CONCLUSION

Apparently, it may look that the value of the coefficient of determination along with the other measure of accuracy should have expressed much better values, but we have to remember that we've only chosen the numerical characteristics of a car as our independent variables through limitations.

There may exist some categorical parameters that affect the price of car. Working with those parameters alongside the numerical can explain the remaining correlation.However, that is a study of deeper machine learning algorithms and software that is beyond our scope of the study. Also, by looking at the residual plot we see heteroscedasticity was present in the data, but due enumerator's bias it was not taken into account.

Talking about the numerical characteristics with which we've worked only we can say that they are successfully affecting the car price and the ultimate picture of various predicted car prices can be shown via this model



ACKNOWLEDGEMENT

I would like to thank Dr. Ayan Chandra, my supervisor and dissertation guide, for his guidance during the course of my dissertation. In addition, I would like to show all my appreciation to all my respected teachers of St. Xavier's College Statistics faculty, who have instilled in me a strong research mindset, a sense of curiosity, and the ability to continue in this field. Lastly, I would like to extend my gratitude to St. Xavier's College for the opportunity to present a dissertation project paper on a topic of my choice. I would also like to thank them for helping me to develop a research mindset in me.

REFERENCE

- *“Fundamentals of Statistics – Vol 2” by A.M. Goon , M.K. Gupta and B. Dasgupta*
- *"Introduction to Linear Regression Analysis" by Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining.*
- *“Regression Analysis and Linear Models: Concepts, Applications, and Implementation” by Richard B. Darlington and Andrew F. Hayes*
- *Keggle.com*
- *Youtube.com*
- *financialsamurai.com*

THANK YOU