Abstract. There are few things that drive the business of selling used cars in Indonesia continues to increase, namely new car prices are higher. So the prospective buyers the ability to purchase power not too strong, buying a used car is one option. Linear regression is a statistical analysis of several variables relationship model according to the form of the relationship of linear equations explicitly. The purpose of this research is to apply the method of Linear Regression on selling used car Toyota Innova, and Honda CRV in Indonesia, especially in the cities of Jakarta, Bandung, Surabaya, and Semarang. Used car sales prices are influenced by the age factor car, car mileage, color car, car transmission, and car type. The results showed, an analysis of age factor car sharing and car has a mileage rating of 62.6% level of confidence. By adding some other variable that is the color of the car, the transmission and the cities of car sales has a value of more than 75% accuracy.

1. Introduction
In line with technological advancements, causing more and more types of four-wheeled vehicles to be offered on the market to meet people's needs. This will provide a wider opportunity for the community to choose and buy four-wheeled vehicles that are in accordance with their wants and needs, because in this life people who are also human beings are often faced with various kinds of problems. One of the problems faced by humans is their life needs, namely the need for cars.

Cars that are generally used as a means of transportation to transport passengers, goods and travel both long and near. For that we need a car with a strong engine, a good model, an affordable price so that the car is sold in the market. In the sale of used cars today, to be able to get a price that is in accordance with the conditions of the car there are several factors such as the age of use, distance, and the color of the car that is very influential to determine the selling price of the used car. Therefore not a few of the people who buy used cars are what they want. One way that can be used in selling used car prices so that buyers do not experience losses is by linear regression.

Regression method is one of the statistical analysis techniques used to describe the relationship between one response variable with one or more explanatory variables [1]. Linear regression is divided into 2 namely simple linear regression and multiple linear regression. Simple linear regression requires only one variable while multiple linear regression uses> 1 variable. In this study using multiple linear regression methods

Used car data taken from an online site for buying and selling used cars, namely www.mobil123.com. Many variations in prices offered on the site in each car based on the year. In this study a data on the sales of used Toyota Kijang Innova, and Honda CRV in Jakarta, Surabaya, Bandung and Semarang in
the period of 2008 - 2017. To find out what percentage of the influence of age, distance, and color of the car can affect the increase and decrease in used car sales prices need to be done with linear regression analysis.

2. Literature review

2.1. Multiple linear regression

Multiple Regression Analysis is an analysis that has more than one independent variable. Multiple linear regression techniques are used to determine whether there are significant effects of two or more independent variables \((X_1, X_2, X_3, \ldots, X_k)\) on the dependent variable \((Y)\). Multiple linear regression models for populations can be shown as follows \([2,3]\):

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + e
\]  

(1)

Multiple linear regression models for the population above can be estimated by multiple linear regression models for samples, namely:

\[
\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k
\]  

(2)

where:

\[
\hat{Y} = \text{Estimator value for variable } Y
\]

\[
b_0, b_1, b_2, \ldots, b_k = \text{Allegations for constant parameters } \beta_0, \beta_1, \beta_2, \ldots, \beta_k
\]

\[
X = \text{Independent variable}
\]

The purpose of linear regression analysis is to measure the intensity of the relationship between two or more variables and contain predictions / estimates of the value of \(Y\) and the value of \(X\). The general form of multiple linear regression equations includes two or more variables \([4]\).

2.2. Coefficient of determination \((R^2)\)

\(R^2\) or the determination coefficient is how much the ability of all independent variables to explain the variance of the dependent variable. Tool to measure the level of compatibility / perfection of the regression model or to state the proportion of total diversity of the variable \(Y\) values which can be explained by the values of variable \(X\) through that linear relationship \([5]\). In simple terms the coefficient of determination is calculated by squaring the Correlation Coefficient \((R)\). For example, if the value of \(R\) is 0.80 then the coefficient of determination \((R^2)\) is 0.80 x 0.80 = 0.64. It means that the ability of the independent variable to explain the variance of the dependent variable is 64.0%. Means there are 36% (100% - 64%) variance of the dependent variable which is explained by other factors. Based on these interpretations, it appears that value \(R^2\) is between 0 until 1. \(R^2 = 0\) if all \(b_k = 0\). \(R^2 = 1\) if all \(Y\) observation is right on the surface of the response, in other words \(Y_i = \hat{Y}_i\) \([6]\). Formula of \(R^2\) to measure the proportion of the total diversity of \(Y\) observation values around the average which can be explained by the regression line or the independent variable used will be shown below \([7]\):

\[
R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
\]  

(3)

2.3. Adjusted R-square

There are weaknesses in calculations of \(R^2\), so that many researchers suggest using Adjusted R-Square. The interpretation is the same as R-Square, but the value of Adjusted \(R^2\) can go up or down with the addition of new variables, depending on the correlation between the additional independent variables with the dependent variable \([7]\). The value of Adjusted R-Square can be negative, so if the value is negative, then the value is considered 0, or the independent variable is not at all able to explain the variance of the dependent variable. Indeed, this adjusted \(R^2\) will also increase with increasing number
of variables, but the increase is relatively small. It is often recommended, if more than two independent variables, should use the adjusted R-square [7].

3. Methodology
The data used is used car data taken from an online site for buying and selling used cars, namely www.mobil123.com. From the online site sales data taken are sales of used cars of Toyota Kijang Innova, and Honda CRV in the cities of Jakarta, Surabaya, Bandung, and Semarang in the period of 2008 - 2017 with 30 cars from each city.

3.1. Variable identification and measurement
In this study, the variables used are:

- Car Age ($X_1$) = Age of use of Toyota Innova and Honda CRV cars measured in years.
- Car Mileage ($X_2$) = Mileage from Toyota Innova and Honda CRV cars measured in km.
- Car Color ($X_3$) = The color of each Toyota Innova and Honda CRV car.
- Car Transmission ($X_4$) = Transmission from Toyota Innova and Honda CRV cars which are manual or automatic.
- Car Type ($X_5$) = Type of car that is the Toyota Innova or Honda CRV.
- City ($X_6$) = Where to sell used cars for Toyota Innova and Honda CRV, namely Jakarta, Surabaya, Bandung, and Semarang.
- Car Selling Price ($Y$) = Selling Price of Toyota Innova and Honda CRV cars in the period of 2008 - 2017.

3.2. Data analysis
The data analysis used in this study is Multiple Linear Regression and processed using Minitab 16 software. This model was chosen to determine how much influence the independent variables have on the dependent variables either partially or together. The data used to find linear equations have independent variables and dependent variables. The independent variables are age, distance, color, transmission, type, and city. While the dependent variable is the selling price of used cars.

4. Results and discussion
In this section, a regression analysis is performed on used car sales data on Toyota Kijang Innova and Honda CRV cars sold in Jakarta, Surabaya, Bandung, and Semarang.

4.1. Regression analysis of car prices against age using cars
The following are the results of the regression analysis of the price of the car for the age of the car shown in Figure 1 below:

![Figure 1. Fitted line plot of car prices against the age of using a car.](image-url)
In Figure 1, the prediction of the price of a car using the age variable of the car has an accuracy rate of 63.2%. And the multiple linear regression equation is obtained as follows:

$$Y = 3.41E + 08 - 20551055X_1$$

The equation means: (1) The new price for CRV cars is Rp. 341,000,008 and (2) The price of Innova cars per year will decrease by Rp 20,551,055. There are 15 unusual car prices, namely at numbers 4, 10, 42, 48, 56, 78, 84, 91, 149, 221, 224, 230, 233 and 234. Based on Figure 2, No. 4, cars are sold lower than the predicted selling price (Fit Value) of Rp 238,154,190. At No. 10, cars are sold higher than the predicted selling price (Fit Value) of Rp 73,745,749. And so on for another example.

**Figure 2.** Test results of car prices against the age of using a car.

### 4.2 Regression analysis of car prices against car mileage

The following is the result of a regression analysis of the price of the car for the car mileage shown in Figure 3 below:

In Figure 3, car price predictions using the variable mileage of car use have an accuracy of 33.3%. And the multiple linear regression equation is obtained as follows:

$$Y = 2.77E + 08 - 883223X_2$$
The equation means: (1) The new price for a CRV car is Rp. 277,000,000 and (2) The price of an Innova car for every 1000 km increase in distance will decrease by Rp. 883,223. Then there are 20 unusual car prices, namely at numbers 10, 26, 45, 54, 61, 78, 96, 121, 130, 149, 158, 168, 181, 185, 190, 221, 230, 233, 234, 240. Based on Figure 4, No. 2, the car with the largest km is sold at a price higher than the predicted value (Fit Value), and so on, except at No. 24, cars are sold higher than the predicted value (Fit Value).

4.3. Regression analysis of car prices against age of use and car mileage

The following are the results of the regression analysis of car prices for car mileage shown in Table 1 below:

| Predictor | Coef | SE Coef | T    | P    | R-Sq | R-Sq (adj) | S  |
|-----------|------|---------|------|------|------|------------|----|
| Constant  | 341475953 | 6956529 | 49.09 | 0.000 | 63.6% | 63.3% | 38380623 |
| X1 (Usia) | -19063634 | 1358446 | -14.03 | 0.000 | 63.6% | 63.3% | 38380623 |
| X2 (Jarak) | -132197 | 80383 | -1.64 | 0.101 | - | - | - |

In Table 1, the prediction of car prices with service life and distance traveled by car has an accuracy of 63.6%. And the multiple linear regression equation is obtained as follows:

\[ Y = 3.41E + 08 - 19063634X_1 - 132197X_2 \]

The equation means: (1) The new price for a CRV car is Rp. 341,475,953, (2) The price of Innova cars per year will decrease by Rp. 19,063,634, and (3) The price of Innova cars per 1000 km increase down by Rp 132,197. But the X_2 variable has no significant effect because the p-value> 0.05 is 0.101 which is shown in Table 1 above. There are 21 unusual car prices, namely at numbers 4,10, 26, 45, 54, 61, 78, 96, 121, 130, 149, 158, 168, 181, 185, 190, 221, 230, 233, 234, 240 sold higher and lower than the predictive value (Fit Value) shown in Figure 5 below.
6

Figure 5. Car price test results against age of use and car mileage.

4.4. Regression analysis of car prices on all variables

The following are the results of the regression analysis of car prices shown in Table 2 below:

Table 2. Multiple linear regression test results.

| Predictor         | Coef  | SE Coef | T     | P    | R-Sq | R-Sq (adj) | S    |
|-------------------|-------|---------|-------|------|------|------------|------|
| Constant          | 347878770 | 33844232 | 10.28 | 0.000 | 76.2% | 74.6%     | 31912019 |
| X1 (Age)          | -20326517 | 1212826  | -16.76 | 0.000 |      |            |      |
| X2 (Mileage)      | -60089   | 69168   | -0.87  | 0.386 |      |            |      |
| X3 (Grey_COLOR)   | 12045688 | 32461293| 0.37   | 0.711 |      |            |      |
| X4 (Brown_COLOR)  | -19833778| 36197529| -0.55  | 0.584 |      |            |      |
| X5 (Gold_COLOR)   | 11747973 | 39706202| 0.30   | 0.768 |      |            |      |
| X6 (Green_COLOR)  | 28569331 | 37543125| 0.76   | 0.447 |      |            |      |
| X7 (Black_COLOR)  | 11522971 | 32523054| 0.35   | 0.723 |      |            |      |
| X8 (Maroon_COLOR) | -7241507 | 39665946| -0.18  | 0.855 |      |            |      |
| X9 (White_COLOR)  | 20976756 | 32795692| 0.64   | 0.523 |      |            |      |
| X10 (Automatic)   | 15388159 | 4565496 | 3.37   | 0.001 |      |            |      |
| X11 (Innova)      | -36578615| 4386245 | -8.34  | 0.000 |      |            |      |
| X12 (Jakarta)     | -20938125| 6135338 | -3.41  | 0.001 |      |            |      |
| X13 (Bandung)     | -16195184| 6023067 | -2.69  | 0.008 |      |            |      |
| X14 (Semarang)    | -621227  | 604981  | -1.01  | 0.198 |      |            |      |

In Table 2, the prediction of car prices with age, the distance of car use and color has an accuracy of 76.2%. And the multiple linear regression equation is obtained as follows:

\[ Y = -34.8E + 0.08 - 203.26517X_1 - 60.089X_2 + 120.45688X_3 - 198.33778X_4 + 117.47973X_5 + 28.569331X_6 + 11.522971X_7 - 72.41507X_8 + 20.976756X_9 + 15.388159X_{10} - 36.578615X_{11} - 62.1227X_{12} \]

The equation means: (1) The new price for CRV cars is Rp. 347,888,770, and (2) The price of Innova cars per year will decrease by Rp. 20,326,517. For the variable distance, Semarang city and all color variables do not have a significant effect because p-value < 0.05. There are 26 unusual car prices, namely at number 4, 10, 27, 45, 46, 48, 54, 56, 73, 78, 80, 84, 91, 106, 127, 129, 130, 141, 149, 171, 179, 197,
At number 4, 45, 48, 54, 56, 73, 84, 91, 127, 129, 171 and 179 the selling price is below the predicted value while the others are sold higher than the value the prediction (Value Fit) and only number 197 whose value is the same as the car selling price prediction value shown in Figure 6 below.

![Figure 6. Car price test results against the age of use, mileage and color of the car.](image)

### 4.5. Best model of multiple linear regression

In choosing a model, of course I will choose the best. The model itself is influenced by the variables taken. To find out the best model, the calculation is done to see the best R-Sq value based on what and what variables are taken. The following are the results of the Best Subsets Regression that have been done:

| Table 3. Multiple linear regression test result. |
|-----------------------------------------------|
| 
| Vars | R-Sq | R-sq (adj) | Mallows Cp | S | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 |
|------|------|------------|-------------|---|----|----|----|----|----|----|----|----|----|-----|-----|
| 1    | 65.2 | 63         | 110.7       |   |    |    |    |    |    |    |    |    |    |     |     |
| 2    | 33.3 | 33         | 391.7       |   |    |    |    |    |    |    |    |    |    |     |     |
| 2    | 72.1 | 71.9       | 28.8        |   |    |    |    |    |    |    |    |    |    |     |     |
| 2    | 66.1 | 65.8       | 85.4        |   |    |    |    |    |    |    |    |    |    |     |     |
| 3    | 73.9 | 72.9       | 19.7        |   |    |    |    |    |    |    |    |    |    |     |     |
| 3    | 72.9 | 72.6       | 22.3        |   |    |    |    |    |    |    |    |    |    |     |     |
| 4    | 73.6 | 73.6       | 13.9        |   |    |    |    |    |    |    |    |    |    |     |     |
| 4    | 73.6 | 73.6       | 14.7        |   |    |    |    |    |    |    |    |    |    |     |     |
| 5    | 75.2 | 74.7       | 5.6         |   |    |    |    |    |    |    |    |    |    |     |     |
| 5    | 75.3 | 74.8       | 13.5        |   |    |    |    |    |    |    |    |    |    |     |     |
| 6    | 75.6 | 75.3       | 3.3         |   |    |    |    |    |    |    |    |    |    |     |     |

### Unusual Observations

| Obs | X1 | Y : Marge | Fit | SE Fit | Residual | St Dev | L Resid |
|-----|----|-----------|-----|--------|----------|--------|---------|
| 4   | 1490000000 | 28375797 | 700.2867 | -8927877 | 229797 |
| 20  | 1300000000 | 659543335 | 126660999 | 7000897 | 2.69R |
| 27  | 6050000000 | 1200882377 | 24628390 | -88037 | 0.00X |
| 45  | 2000000000 | 2760650000 | 7457931 | 71230592 | 2.33R |
| 46  | 6000000000 | 1335205252 | 32552562 | -6220852 | -2.15R |
| 54  | 1000000000 | 1305282377 | 18852330 | -58037 | 0.00X |
| 66  | 6000000000 | 1832082377 | 13852369 | -628052 | -2.15R |
| 73  | 6000000000 | 2026682627 | 22746932 | -706687 | 0.82X |
| 76  | 2000000000 | 29766289 | 5916290 | 90291712 | 2.99R |
| 80  | 6000000000 | 3455500000 | 6605549 | 6059054 | 5.00R |
| 81  | 9000000000 | 22746807 | 6776654 | -678076 | -2.10R |
| 156 | 1000000000 | 2139329737 | 6036669 | 7069357 | 3.20R |
| 127 | 6000000000 | 1888941800 | 18576450 | -3894180 | -0.14X |
| 128 | 7000000000 | 1943877476 | 1993279 | -2789747 | -0.99X |
| 129 | 8000000000 | 1642380929 | 16970937 | 2279168 | 0.84X |
| 140 | 1000000000 | 1746466356 | 6112650 | 650465 | 0.81X |
| 145 | 3000000000 | 3126063430 | 3246766 | 8571660 | 2.70R |
| 171 | 9000000000 | 239157747 | 23914287 | -4259774 | -1.94X |
| 179 | 5000000000 | 2300025253 | 2254197 | 4219774 | 1.94X |
| 197 | 3000000000 | 1780600000 | 31512015 | 0 | X |
| 212 | 4000000000 | 2481838379 | 5818404 | -6515387 | -3.01R |
| 224 | 3000000000 | 3149878460 | 664978 | 96112440 | 3.09R |
| 230 | 3000000000 | 3152966666 | 6050008 | 09713334 | 2.89R |
| 234 | 7000000000 | 3295567579 | 6072741 | 10074321 | 3.59R |
In the table above, the R-Sq value of each variable can be seen if it is entered into the model. Vars shows how many variables are taken, and the contents of "x" indicate which variables are taken to the model. If you see more and more variables taken the value of more than 75%, even reaching 76.2. This indicates that the variable we choose has a significant influence on the results of the model.

5. Conclusion

Based on the results of data analysis and discussion of sales of used cars for Toyota Kijang Innova and Honda CRV in the cities of Jakarta, Surabaya, Bandung, and Semarang, conclusions can be taken as follows:

- Using the variable age of use of cars in predicting the selling price of cars can be used as a reference to determine the price of used cars sold because it has a trust value of 63.2%.
- If using the car mileage variable only in predicting the selling price of the car only has a trust value of 33.2%.
• The selling price can also be predicted based on the age and distance of the used cars sold because of the trust value of 63.6%.
• To predict the selling price of cars with other additional variables namely age, distance, colors of cars, transmissions and cities car sales have a trust value of 76.2%.
• The P value indicates the value that helps to conclude whether the calculation is correct or not. If the value of P < 0.05 then the value can be said to be good. But conversely if the number P > 0.05 then the results obtained means that it is not used to make an estimate of the conclusion. The best result of P value is if P is below 0.05.
• In the experiment above, the more variables used to predict car prices, the higher the accuracy in predicting prices. The price of a car can be predicted when several variables, namely age, distance, color of the car, transmission and cities of car sales are combined and get an accuracy value above 75%.

References
[1] Nafi’iyah N 2016 “Perbandingan Regresi Linear, Backpropagation Dan Fuzzy Mamdani Dalam Prediksi Harga Emas,” Seminar Nasional Inovasi dan Aplikasi Teknologi di Industri (SENIATI) 2016[ ISSN : 2085-4218]
[2] Supranto J 2004 Analisis Multivariat Arti dan Interpretasi Jakarta: Rineka Cipta
[3] Amrin 2016 “Data Mining dengan Regresi Linier Berganda Untuk Peramalan Tingkat Inflasi,” Jurnal Techno Nusa Mandiri XIII 74
[4] Riduwan 2008 Dasar-dasar Statistika (Bandung: Alfabeta)
[5] Riana D 2012 Statistika Deskriptif itu mudah (Jakarta: Jelajah Nusa)
[6] Hatidja D 2006 Bahan Ajar Analisis Regresi (Manado: Jurusan Matematika FMIPA UNSRAT)
[7] Spiegel M 1961 Theory and Problems of Statistics (New York: Schaum Publishing Co)