SECI2143-02 PROBABILITY & STATISTICAL DATA ANALYSIS

SEMESTER 2, 2020/2021

PROJECT 2

CARS CHARACTERISTIC

GROUP MEMBERS:

NO.	**NAME**	**MATRIC NO.**
1.	TERENCE A/L LOORTHANATHAN	A20EC0165
2.	MUHAMMAD DINIE HAZIM BIN AZALI	A20EC0084
3.	RISHMA FATHIMA BINTI BASHER	A20EC0137
4.	NUR IRDINA ALIAH BINTI ABDUL WAHAB	A20EC0115

SUBMITTED TO:

DR. CHAN WENG HOWE

Table of Contents

1.0 Introduction

In a car's system, the characteristic is a very important component, since it is responsible to show the car actual performance. As time goes by, there are various types of cars being created all over the world such as sports cars, minivans, sedans, coupes and many more. Not only that, the cars' characteristics also have different types as it can affect the car performance. It is really important for buyers to know which car is the best and suits their style. Hence, the purpose of this project is to determine whether one of the characteristics can affect the other part of the car or not. The data set is analysed through several statistical tests such as hypothesis testing, correlation, regression and also chi-square test of independence to yield an accurate result. R-studio software has been used to carry out the statistical analysis on the chosen variables in the data set.

2.0 Dataset

Our data is named CarsData. This data is made up of several pieces of information about a group of cars. This data is retrieved from Project datasets. In other words, it is secondary data that is retrieved from another organisation which collected it.

There are 9 variables in the dataset. All of which can be categorised in different levels of measurement as below:

- Nominal scale = {CAR, Model, Origin}
- Ratio scale = {MPG, Cylinders, Displacement, Horsepower, Weight, Acceleration}

Variable	Definition	Key
CAR	Car's Name	
MPG	Miles Per Gallon	
Cylinders	Engine Cylinders	
Displacement	Displacement	
Horsepower	Horsepower Engine	
Weight	Weight of Car	
Acceleration	Acceleration of Car	
Model	Model of Car	
Origin	Orgin Country	US= United States; Europe=A country in the continent of Europe; Japan = Japan

Figure 1: Data Dictionary

CarsData set, there are 8 parameters or variables that have been recorded and can be measured. The data set has been recorded everyday from 14th April 2021 until 23rd April 2021 without missing any single day. Almost all types of car have been recorded in this data set with approximately 132. The variables in this data set are MPG , cylinders , displacement , horsepower , weight , acceleration, model and origin recorded for every single day. After a few meetings with our lecturer, DR. CHAN WENG HOWE we have cut down a few unnecessary parts in the variables list. Hence, our recent data variable for this project has been finalized ,we chose these variables because they are econamale operated and reliable.

3.0 Data Analysis

3.1 Hypothesis Test

The weight from data CarData.slx is selected to do analysis. The test statistics that is going to be used in this case is z-test as the variance of the dataset is unknown, but the sample is large, with more than 30.

It is of interest to determine if there is evidence with significance level of 0.05 to support a claim that the mean weight from data CarData.slx is different from 2871.

According to a website called CreditDonkey, the average weight of a car is 2,871 pounds. We want to check whether this sample has a mean greater than 2,871 pounds or not.

H0: $\mu = 2{,}871$ (weight of the car is equal to 2,871 pounds)

H1: $\mu > 2{,}871$ (weight of the car is not equal to 2,871 pounds)

Significance level, $\alpha = 0.05$

$z_{0.025} = 1.96$

Let: n = sample size

s = sample standard deviation

x-bar = sample mean

mu = population mean

z=test statistic

pval=p-value of test statistic

```
> x=c(3504,3693,3436,3433,3449,4341,4354,4312,4425,3850,309
0,4142,4034,4166,3850,3563,3609,3353,3761,3086,2372,2833,27
74,2587,2130,1835,2672,2430,2375,2234,2648,4615,4376,4382,4
732,2130,2264,2228,2046,1978,2634,3439,3329,3302,3288,4209,
4464,4154,4096,4955)
> n=50
> s=sd(x)
> xbar=mean(x)
> mu=2871
> z=(xbar-mu)/(s/sqrt(n))
> alpha=0.05
> pval=pnorm(z)
> z
[1] 4.180466
> pval
[1] 0.9999855
> |
```

Figure 2: Test statistic calculation for z-test

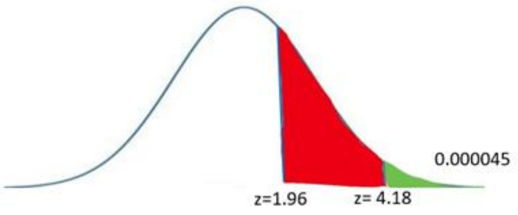


Figure 3: graph for test statistics

test statistic = 4.18 p value = 0.000045 Since test statistic, z-test = 4.18 is less than $z_{0.05}$ = 1.96 thus, null hypothesis is rejected. To conclude, there is strong evidence to conclude that the average weight of the cars is more than 2,871 pounds at the 0.05 significance level.

3.2 Correlation Analysis

The type of correlation coefficient that is used in this case study is Pearson's product moment correlation coefficient. A few variables are selected in conducting this correlation analysis in order to determine the strength of linear relationship between two variables. Hence, more than one correlation analysis has been done by using different variables.

First Correlation

The first correlation is between MPG and Weight, we want to discover, is the car weight affects the MPG (miles per gallon) of a car?

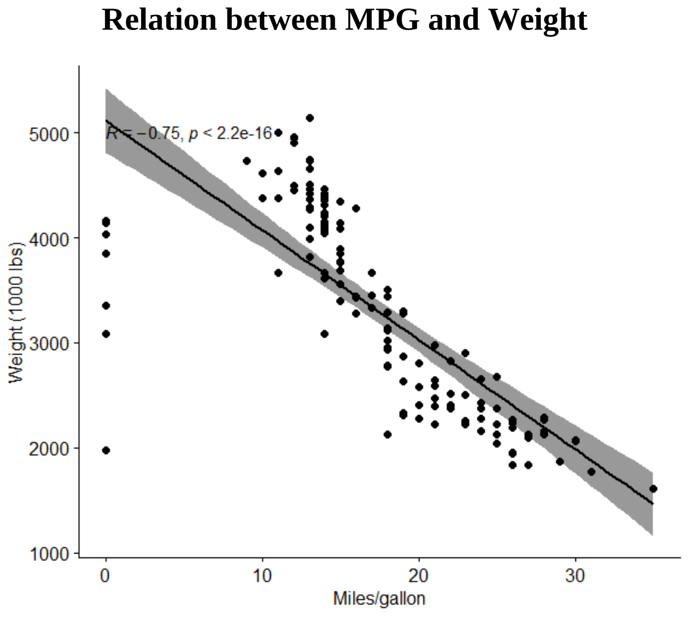


Figure 4: Correlation between variable Miles/gallon and Weight(1000lbs)

Sample correlation coefficient, r = -0.75

It can be seen that the Weight(1000lbs) decreases as the Miles/gallon increases. A scatter plot and correlation analysis of the data indicates that there is a negative relationship between the miles per gallon and weight of the car.

Significance Test for First Correlation

In order to collect more evidence in supporting the relationship between MPG and Weight, a significance test is conducted at $\alpha = 0.05$.

- Hypothesis
  $H_0: \rho = 0$ (car weight does not affects MPG of a car)
  $H_1: \rho \neq 0$ (car weight affects MPG of a car)
- $\alpha = 0.05$, df = 132 – 2 = 130
  critical value: $\pm t_{0.025,130} = \pm 1.9784$
- Test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = -12.9284$$

- p-value = 2.2×10^{-6}

Since the p-value is less than the significance level of 0.05, we rejected the null hypothesis. There is sufficient evidence of a linear relationship between Miles/gallon and Weight(1000lbs) at the 5% level of significance.

Second Correlation

For the second correlation, we want to calculate, is there any relation between the number of cylinders with the displacement?

Relation between Number of Cylinders and Displacement

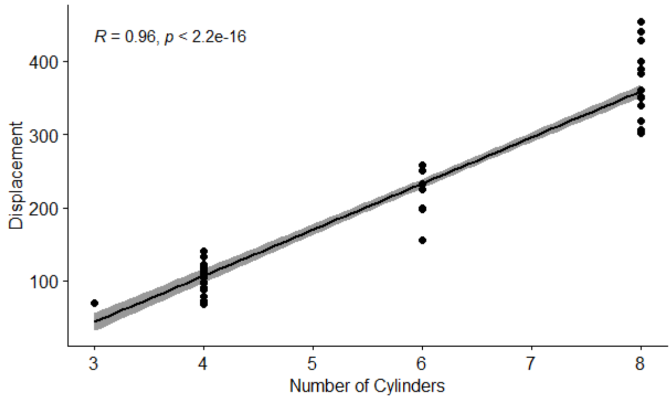


Figure 5: Correlation between variable Number of Cylinders and Displacement

Sample correlation coefficient, r = 0.96

It can be seen that the Displacement increases as the Number of Cylinders increases. A scatter plot and correlation analysis of the data indicates that there is a positive relationship between the cylinders and displacement of the car.

Significance Test for Second Correlation

In providing evidence of a linear relationship between the Number of Cylinders and Displacement at 0.1 level of significance, significance test has been conducted.

- Hypothesis

  $H_0: \rho = 0$ (number of cylinders does not affects car displacement)

  $H_1: \rho \neq 0$ (number of cylinders affects car displacement)
- $\alpha = 0.1$, df = 132 – 2 = 130

  critical value: $\pm t_{0.05,130} = \pm 1.6567$
- Test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = 39.0917$$

- p-value = 2.2×10^{-6}

Since the p-value is less than the significance level of 0.1, we rejected the null hypothesis. There is sufficient evidence of a linear relationship between Number of Cylinders and Displacement at the 10% level of significance.

At last, from the correlation analysis that have be done, we can conclude that there is a strong negative correlation between Miles/gallon and Weight(1000lbs) and there is also a strong positive correlation between Number of Cylinders and Displacement.

3.3 Regression Analysis

Regression test is used to predict the value of a dependent variable based on the value of at least one independent variable. Because of that, this test can explain impacts of changes in an independent variable on the dependent variable.

In this project's Regression Test:

- Dependent Variable (Y) = Acceleration
- Independent Variable (X) = Horsepower

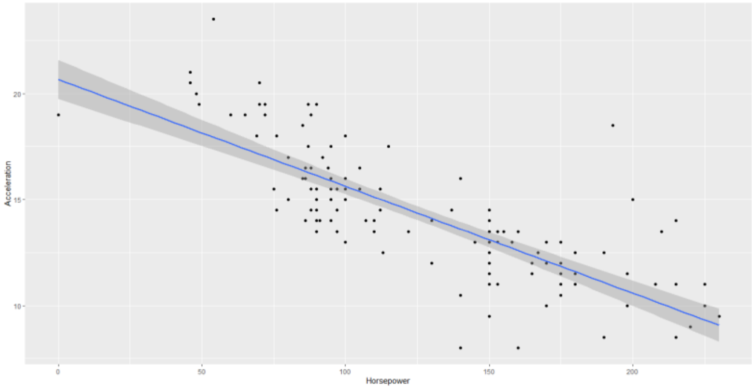


Figure 6: Negative Linear Relationship Model of CarsData

There are many different types of regression models, based on CarsData the regression model is a Negative Linear Relationship Model.

From the code segment : [summary(datalinear)],

The equation is: y = 20.66819 + -0.05039x

From the equation, b_0 = +20.668190 is the estimated average value of acceleration when the value of horsepower is zero. Therefore, this indicates that based on the horsepower of cars observed, 20.66819 is a value of acceleration that cannot be explained by horsepower.

From the equation, b_1 = -0.05039 is the measure of estimated change in the average value of acceleration because of a one-unit change in horsepower. Therefore, this value indicates that acceleration decreases by 0.05039, on average for every addition of one unit of horsepower.

<u>Coefficient of determination:</u>

Coefficient of determination is the portion of the total variation in the acceleration that is explained by variation in the horsepower.

$$R^2 = 0.623976$$

Since $0 < R^2 < 1$, linear relationship is proved to be weaker between horsepower and acceleration. Some but not all the variation in acceleration is explained by variation in horsepower. According to R^2 obtained, only 62% of the variation is explained.

<u>T-test:</u>

Regression test hypothesis statements:

H_0 : β1 = 0 (There no linear relationship between horsepower and acceleration)

H_1 : β1 ≠ 0 (There exists linear relationship between horsepower and acceleration)

y = 20.66819 + -0.05039x

d.f = 132 - 2 = 130

-t = -1.656659 t = 1.656659

t_0 = 8.22×10^{-14}

p-value = 2.2×10^{-16}

Since the p-value is less than 0.05 significance level, null hypothesis is rejected. From the statistical test above, there is convincing evidence to conclude that Horsepower does affect the Acceleration.

3.4 ANOVA

ANOVA is used to test the quality of all variables by analyzing sample variances. In this project, we want to test all different variables whether they have the same mean or not.

Step 1

To test different variables, have the same mean

H0: µ0 = µ1

H1: there is different mean

Step 2

Category 1 (MPG)

N =132

x' = 17.5455

S= 6.8889

Category 2 (Cylinders)

N = 132

x' = 12.3008

S = 1.8593

Category 3 (Displacement)

N = 132

x' = 245.4508

S = 126.4553

Category 4 (Horsepower)

N = 132

x' = 126.7576

S = 48.1501

Category 5 (Weight)

N = 132

x' = 6522.0000

S = 961.0925

Category 6 (Acceleration)

N = 132

x' = 14.280303

S = 3.0718

Category 7 (Model)

N = 132

x' 142.0301

S = 1.1807

Step 3

Find the variance between samples:

(17.5455 + 12.30075 + 245.4508 + 126.7576 + 6522.0000 + 14.280303 + 142.0301) / 7

= 1011.4807

$S_{x'}$ = 2431.4598

$ns^2_{x'}$ = 132 $(2431.4598)^2$ = 780383573.2

Step 4

s^2_p = $((79.5354)^2 + (1.8593)^2 + (126.4553)^2 + (48.1501)^2 + (961.0925)^2 + (3.0718)^2 + (1.1807)^2)$ /7

= 135478.3336

Step 5

Calculate test statistic, F

$F = ns^2_x / s^2_p$ = 780383573.2000 / 135478.3336 = 5760.2094

Step 6

Calculate numerator and denominator degree of freedom

- Numerator = k -1 = 7-1 = 6
- Denominator = k(n-1) = 7 (132-1) = 917

Step 7

Find critical value of F with $\alpha = 0.05$ from f distribution table

F-critical value = 2.099

Step 8

Test the claim and state the conclusion

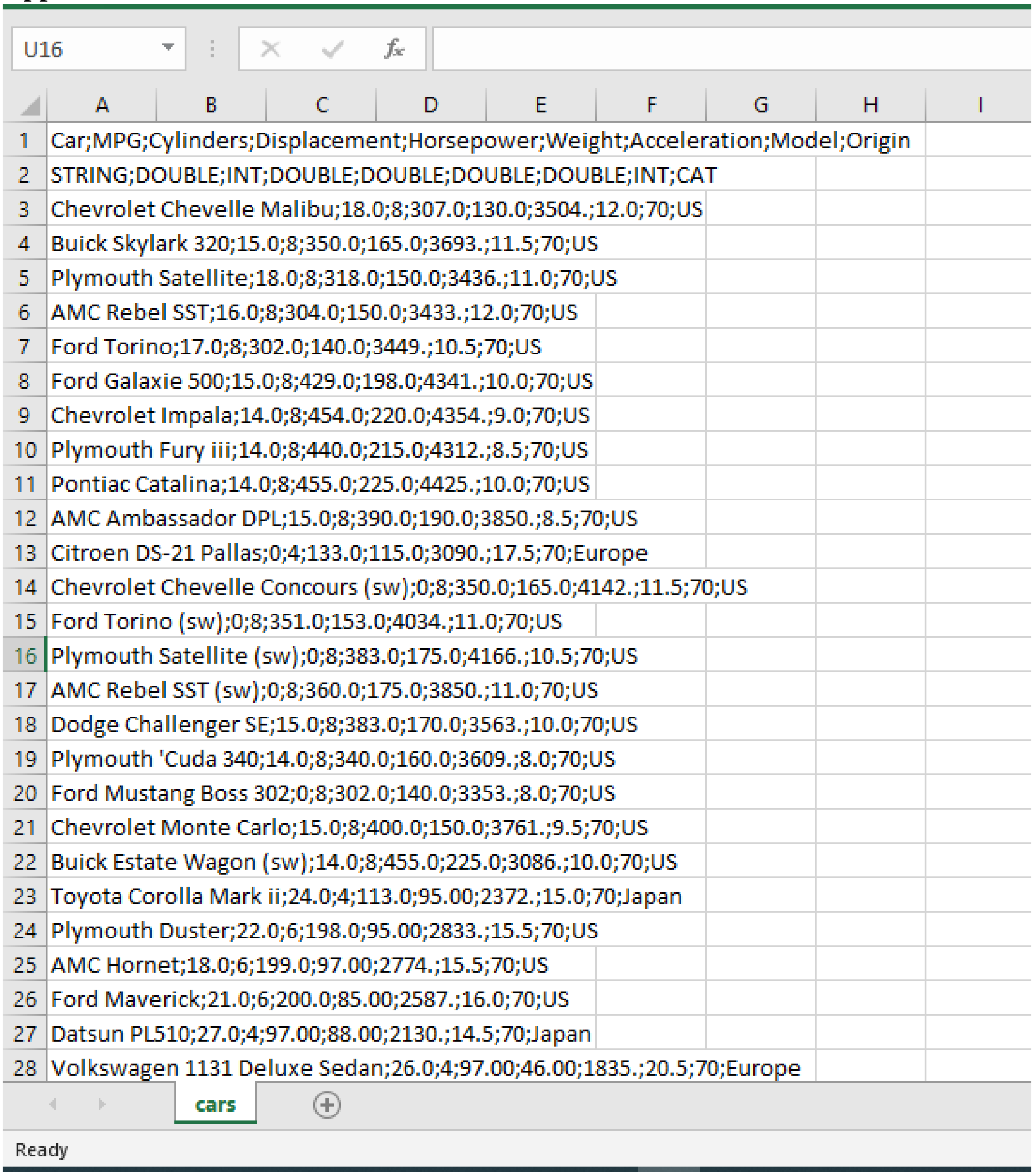
Since F-test statistic < F-critical value (5760.2094< 2.099), we reject the null hypothesis There is sufficient evidence that we can conclude that there is difference in mean among the variables.

4.0 Conclusion

In conclusion, based on the data from the DataCars.xls which has the data about the cars, it can be concluded that the weight of the car is more than 2871 pounds. Besides, negative correlation between Miles/gallon and Weight(1000lbs) and there is also a positive correlation between number of Cylinders and Displacement. Then, there is convincing evidence to conclude that Horsepower does affect the Acceleration. Hence, the relationship between horsepower and acceleration does not appear to be linear. Lastly, based on the ANOVA test we can conclude that there is difference in mean among the variables.

In a nutshell, there are important aspects that need to be considered when buying a car or when we want to get to know about cars. So, when conducting these kinds of tests, we could find out more about the cars in general. All the information gathered from the test will be so much informative for either a car enthusiast or a normal person.

Appendix



	A
1	Car;MPG;Cylinders;Displacement;Horsepower;Weight;Acceleration;Model;Origin
2	STRING;DOUBLE;INT;DOUBLE;DOUBLE;DOUBLE;DOUBLE;INT;CAT
3	Chevrolet Chevelle Malibu;18.0;8;307.0;130.0;3504.;12.0;70;US
4	Buick Skylark 320;15.0;8;350.0;165.0;3693.;11.5;70;US
5	Plymouth Satellite;18.0;8;318.0;150.0;3436.;11.0;70;US
6	AMC Rebel SST;16.0;8;304.0;150.0;3433.;12.0;70;US
7	Ford Torino;17.0;8;302.0;140.0;3449.;10.5;70;US
8	Ford Galaxie 500;15.0;8;429.0;198.0;4341.;10.0;70;US
9	Chevrolet Impala;14.0;8;454.0;220.0;4354.;9.0;70;US
10	Plymouth Fury iii;14.0;8;440.0;215.0;4312.;8.5;70;US
11	Pontiac Catalina;14.0;8;455.0;225.0;4425.;10.0;70;US
12	AMC Ambassador DPL;15.0;8;390.0;190.0;3850.;8.5;70;US
13	Citroen DS-21 Pallas;0;4;133.0;115.0;3090.;17.5;70;Europe
14	Chevrolet Chevelle Concours (sw);0;8;350.0;165.0;4142.;11.5;70;US
15	Ford Torino (sw);0;8;351.0;153.0;4034.;11.0;70;US
16	Plymouth Satellite (sw);0;8;383.0;175.0;4166.;10.5;70;US
17	AMC Rebel SST (sw);0;8;360.0;175.0;3850.;11.0;70;US
18	Dodge Challenger SE;15.0;8;383.0;170.0;3563.;10.0;70;US
19	Plymouth 'Cuda 340;14.0;8;340.0;160.0;3609.;8.0;70;US
20	Ford Mustang Boss 302;0;8;302.0;140.0;3353.;8.0;70;US
21	Chevrolet Monte Carlo;15.0;8;400.0;150.0;3761.;9.5;70;US
22	Buick Estate Wagon (sw);14.0;8;455.0;225.0;3086.;10.0;70;US
23	Toyota Corolla Mark ii;24.0;4;113.0;95.00;2372.;15.0;70;Japan
24	Plymouth Duster;22.0;6;198.0;95.00;2833.;15.5;70;US
25	AMC Hornet;18.0;6;199.0;97.00;2774.;15.5;70;US
26	Ford Maverick;21.0;6;200.0;85.00;2587.;16.0;70;US
27	Datsun PL510;27.0;4;97.00;88.00;2130.;14.5;70;Japan
28	Volkswagen 1131 Deluxe Sedan;26.0;4;97.00;46.00;1835.;20.5;70;Europe

Figure 7: Raw dataset

K131

	A	B	C	D	E	F	G	H	I
1	CAR	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model	Origin
2	Chevrolet Chevelle Malibu	18	8	307	130	3504	12	70	US
3	Buick Skylark 320	15	8	350	165	3693	11.5	70	US
4	Plymouth Satellite	18	8	318	150	3436	11	70	US
5	AMC Rebel SST	16	8	304	150	3433	12	70	US
6	Ford Torino	17	8	302	140	3449	10.5	70	US
7	Ford Galaxie 500	15	8	429	198	4341	10	70	US
8	Chevrolet Impala	14	8	454	220	4354	9	70	US
9	Plymouth Fury iii	14	8	440	215	4312	8.5	70	US
10	Pontiac Catalina	14	8	455	225	4425	10	70	US
11	AMC Ambassador DPL	15	8	390	190	3850	8.5	70	US
12	Citroen DS-21 Pallas	0	4	133	115	3090	17.5	70	EUROPE
13	Chevrolet Chevelle Concours	0	8	350	165	4142	11.5	70	US
14	Ford Torino	0	8	351	153	4034	11	70	US
15	Plymouth Satellite	0	8	383	175	4166	10.5	70	US
16	AMC Rebel SST	0	8	360	175	3850	11	70	US
17	Dodge Challenger SE	15	8	383	170	3563	10	70	US
18	Plymouth 'Cuda 340	14	8	340	160	3609	8	70	US
19	Ford Mustang Boss 302	0	8	302	140	3353	8	70	US
20	Chevrolet Monte Carlo	15	8	400	150	3761	9.5	70	US
21	Buick Estate Wagon	14	8	455	225	3086	10	70	US
22	Toyota Corolla Mark ii	24	4	113	95	2372	15	70	JAPAN
23	Plymouth Duster	22	6	198	95	2833	15.5	70	US
24	AMC Hornet	18	6	199	97	2774	15.5	70	US
25	Ford Maverick	21	6	200	85	2587	16	70	US
26	Datsun PL510	27	4	97	88	2130	14.5	70	JAPAN

Sheet1

Ready

Figure 8: Processed dataset