**Registration ID: 2110387**  
**Analysis of mtcars**

**Introduction:**

Climate change is included in one of the global issues. Organisations are working conscientiously to overcome the problem of fossil fuel and make cars more environmentally friendly to mitigate the problem of carbon emission.

To mitigate the problem of fossil fuel, the car industry is facing the biggest challenge in Engineering. This report uses the data of Mtcars from the dsEssex library, and the data consists of the design, performance, and fuel economy for 32 automobiles from 1973 - to 1974. All of the data therein was extracted from the 1974 Motor Trend US magazine.

Before moving forward with the analysis of the selected data, first, we checked the variables of the data, structure of data and missing values in the data. The data consists of 11 different categorical variables. Mpg - Miles/(US) gallon, cyl Number of cylinders, disp- Displacement (cu.in.), hp - Gross horsepower, drat- Rear axle ratio, wt - Weight (1000 lbs), qsec- 1/4 mile time, vs - Engine (0 = V-shaped, 1 = straight), am - Transmission (0 = automatic, 1 = manual), gear- Number of forwarding gears, carb- Number of carburetors

From the outcomes of the colsums technique, we say that there is no missing value in the data, and the data appears in tidy foam. We perform the descriptive analysis through the summary of the data; we can analyse the minimum value (Min), The value of the first quartile (25th percentile or 1st Qu, median value, mean value, The value of the third quartile (75th percentile or 3rd Qu) and maximum value.

**Exploratory Data Analysis**

The data of mtcars included the various factors that may influence fuel economy or miles per gallon. Some of the variables that are appearing dominant variables with Mpg are horsepower, cylinder, transmission type and gear.

In the exploratory analysis, we explore the data through visualisation and transformation in a systematic way. Therefore, first, we check the distribution of the mpg (Miles/(US) gallon) in a histogram graph because it gives a clear picture of groups of the values into continuous ranges. The below-represented graph shows that the distribution is normal (approximately); our data set is also discrete or continuous.

Chart, histogram

Description automatically generated

For better visualisation, we plot an mpg graph concerning transmission type. The below chart shows that the transmission type=1 ( Manual ) gives better miles per gallon than the type=0(Automatic). The thick black line indicates that the mpg for automatic transmission cars is highly concentrated around the median and the first quantile for manual.

Chart, box and whisker chart

Description automatically generated

From the exploratory analysis of the data, we analyse that most cars have 3 and 4 gears, automatic cars mostly have three gears, and manual cars have mostly five gears. Moreover, cars with low weight mostly have high Mpg. All cars have an average horsepower of 147; those cars with eight cylinders represent high hp.

**Multiple Regression Analysis:**

Multiple linear regression is an extension of [simple linear regression](http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/) used to predict an outcome variable (y) based on multiple distinct predictor variables (x). In mtcars data, mpg is a dependent variable and cycl, disp, hp, drat, wt, qsec, vs, am, gear, and carb are the independent variables. Combinedly, the foam linear equation is.

**Mpg=β0+ β1cyl+ β2disp+ β3hp+ β4drat+ β5wt+ β6qsec+β7vs+β8am+β9gear+β10carb+µ**

Call:

lm(formula = mpg ~ ., data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-3.4506 -1.6044 -0.1196 1.2193 4.6271

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.30337 18.71788 0.657 0.5181

cyl -0.11144 1.04502 -0.107 0.9161

disp 0.01334 0.01786 0.747 0.4635

hp -0.02148 0.02177 -0.987 0.3350

drat 0.78711 1.63537 0.481 0.6353

wt -3.71530 1.89441 -1.961 0.0633 .

qsec 0.82104 0.73084 1.123 0.2739

vs 0.31776 2.10451 0.151 0.8814

am 2.52023 2.05665 1.225 0.2340

gear 0.65541 1.49326 0.439 0.6652

carb -0.19942 0.82875 -0.241 0.8122

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom

Multiple R-squared: 0.869, Adjusted R-squared: 0.8066

F-statistic: 13.93 on 10 and 21 DF, p-value: 3.793e-07

The regression analysis of the multiple (unrestricted)regression model of the mtcars shows that the adjusted R^2 of the model is 0.8066, meaning that this model can explain 80% of the variance in mpg. The probability of the independent variables in the model shows that none of the p-values of the variables is significant with Mpg. Also, the value of F-stats is minimal. Therefore, we must omit some variables to make a model significant with Mpg.

The hypothesis testing of the above model failed to reject H0 that there is no statistically significant relationship between the predictor variable, x, and the response variable, y.

**Correlation:**

The correlation explains the coordination of the variables, or they can change together at a constant rate; below, Table 01 shows that disp, hp and wt and drat have a robust correlation ( > 0.5 ) with mpg.

We observe that the wt and disp are highly positively correlated from the correlation plot. Wt and Mpg, disp, cyl, hp, and mpg are more dependent on the other variable than the rest.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 01: Correlation of Variables | | | | | |
| mpg | **cyl** | **disp** | **hp** | **drat** |  |
| 1.0000000 | -0.852162 | -0.8475514 | -0.7761684 | 0.6811719 |  |
| wt | **qsec** | **vs** | **am** | **gear** | **carb** |
| -0.8676594 | 0.418684 | 0.6640389 | 0.5998324 | 0.4802848 | -0.5509251 |

After running the correlation Analysis and multiple linear regression on different variables, we can observe that the heavier cars and more horsepower cars have lower Mpg. Therefore, we are including only hp and wt as a parameter in the restricted model.

Call:

lm(formula = mtcars$mpg ~ hp + wt, data = mtcars)

Residuals:

Min 1Q Median 3Q Max

-3.941 -1.600 -0.182 1.050 5.854

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 37.22727 1.59879 23.285 < 2e-16 \*\*\*

hp -0.03177 0.00903 -3.519 0.00145 \*\*

wt -3.87783 0.63273 -6.129 1.12e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom

Multiple R-squared: 0.8268, Adjusted R-squared: 0.8148

F-statistic: 69.21 on 2 and 29 DF, p-value: 9.109e-12

The restricted multiple regression model R-squared values is 0.8268, which is relatively high. It represents that this model can explain 82% of the outcomes. Hp and wt p-values show a highly significant but negative result, which means an inverse relationship; the independent variable decreases if the dependent variables increase. The value of F-stats also improved from the unrestricted model. The variable inflation factor shows values less than 5, which indicates that the model does not have multicollinearity.

The hypothesis testing of the above model is rejected H0 and accepted H1. There *is* a statistically significant relationship between x and y.

**Heteroskedasticity:**

**Heteroskedasticity** occurs when the variance for all observations in a data set is not the same, which means unequal variance present in the model. We want to check if the model built can explain some pattern in the response variable (Y) that eventually shows up in the residuals.

Chart, scatter chart

Description automatically generated

We can observe no evidence of heteroskedasticity present in the restricted model from the above graph**.**

**Conclusion:**

To execute the analysis of the mtcars data, we use R programming; our analysis comprises the summary, exploratory analysis, and multiple regression model of the mtcars data.

The analyses show that Mpg is related to vehicle weight and the number of cylinders, and the manual transmission is better than the automatic transmission. Moreover, we also analyse that the car having more cylinders negatively impacts Mpg. Horsepower (hp) and cylinders have a negative relationship with the Mpg, and if one increases, the other one decreases.

The car’s weight (wt) shows a negative intercept and an indirect relationship with the Mpg, but the p-value represents high significance. The regression analysis between Mpg(Y) and wt(X) has the R-value of 0.7528, which is relatively high. We can conclude that only weight gives us 75% efficiency, so if the company is designing a new model, they can save the company cost by investing in only one variable, i.e., the car’s weight.

**References:**

* <https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/mtcars>
* http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/
* Upton, G & Brown, D 2022 *A gentle introduction to data analysis and Statistics*, Oxford University Press.
* James, G, Witten, D, Hastie, T, Tibshirani, R. (2013) *An introduction to statistical learning: with applications in R*, New York: Springer. Vol. Springer texts in statistics, Chapter 3

**Appendix:**

#============Analysis of mtcars R coading=============

library(dsEssex)

data("mtcars") #calling data of mtcars

names(mtcars)

head(mtcars) #checking the variables of the data

str(mtcars) #structure of the data

colSums(is.na(mtcars))# checking the missiing values

summary(mtcars2)# to get the summary of the mtcars data

library(dplyr)

#=======Exploratory Analysis========

library(explore)

mtcars %>%

explore\_all()

hist(mtcars$mpg, breaks=10, xlab = "Miles per gallon",#create a histogram of mpg

main = "Historgram of Miles per gallon of cars",

xlim = range(10:35))

plot(mtcars$mpg ~ as.factor(mtcars$am), mtcars, #create boxplot of values for mpg with respect to transmission

xlab = "Transmission type", ylab="Miles per gallon", col="steelblue",

main="Histogram of MPG by transmission type")

ggplot(mtcars, aes(cyl)) + #create line graph of values for cars with respect to cylinder

geom\_histogram(binwidth=1) + xlab('Cylinders') + ylab('Number of Cars') +

ggtitle('Distribution of Cars by Cylinders')

boxplot(mpg ~ gear, data = mtcars)

cov(mtcars$mpg, mtcars$gear) #to get covariance between mpg and gear

mtcars %>%

select(gear, mpg, hp, cyl, am) %>%

explore\_all(target = gear)

ggplot(mtcars, aes(x = hp, y = mpg, color = vs)) + geom\_point() + geom\_smooth()

#=======Multiple regression========

mtcars.lm=lm(mpg~.,data=mtcars)

summary(mtcars.lm)

mtcars.1m=lm(mpg~wt,data=mtcars)

summary (mtcars.1m)

Model1 <- lm( mtcars$mpg ~ hp + wt + as.factor(am), data=mtcars)

summary(Model1)

Model2 <- lm( mtcars$mpg ~ hp + wt + cyl+carb, data=mtcars)

summary(Model2)

Model3 <- lm( mtcars$mpg ~ hp + wt +cyl, data=mtcars)

summary(Model3)

#correlation

pp <- cor(mtcars) #checkimg correlation of the different columns of mtcars

pp[1,] #checking correlation of mpg to other variables

pp[4,] #checking correlation of hp to other variables

pp[5,] #checking correlation of drat to other variables

pp[6,] #checking correlation of wt to other variables

corr<-select(mtcars,mpg,cyl,hp,drat,disp,wt,qsec,vs,gear,am,carb)

pairs(corr)

library(corrplot)

M<-cor(mtcars) #correlation plot

corrplot.mixed(M,lower="circle",upper="number")

fit1 <- lm( mtcars$mpg ~ hp + wt, data=mtcars)

summary(fit1) #restricted model

plot(resid(fit1)) #checking the heteroskedasticity

abline(h=0)