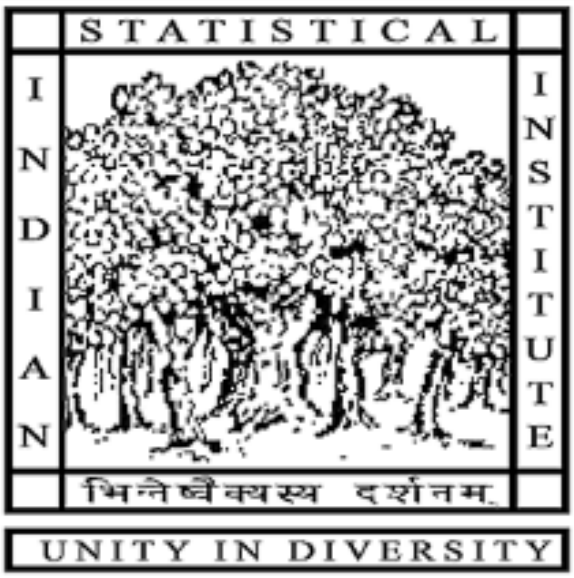


NUMERICAL ASSIGNMENT

STATISTICAL STRUCTURES IN DATA

Instructor:

Dr Subhajit Dutta

Bhavya Arya

24BM6JP14

PGDBA Batch-10

Table of Contents

Dataset: MTCARS

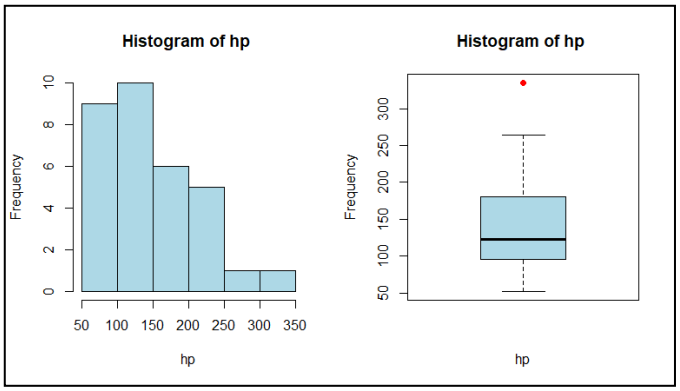
CODE

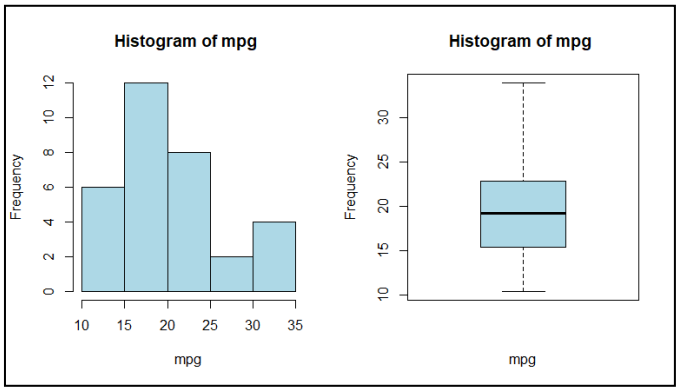
Data Overview

'MT CARS "is a built-in dataset in R that contains information on **32 different car models**. It includes **11 attributes** describing car performance, design, and configuration aspects. While most variables are numerical, some represent categorical information (e.g., the number of cylinders, transmission type, and engine shape).

Univariate Analysis

Numerical Data





	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
mpg	20.09	6.03	19.20	19.70	10.40	33.90	23.50	0.61	-0.37
Hp	146.69	68.56	123.00	141.19	52.00	335.00	283.00	0.73	-0.14
Wt	3.22	0.98	3.33	3.15	1.51	5.42	3.91	0.42	-0.02

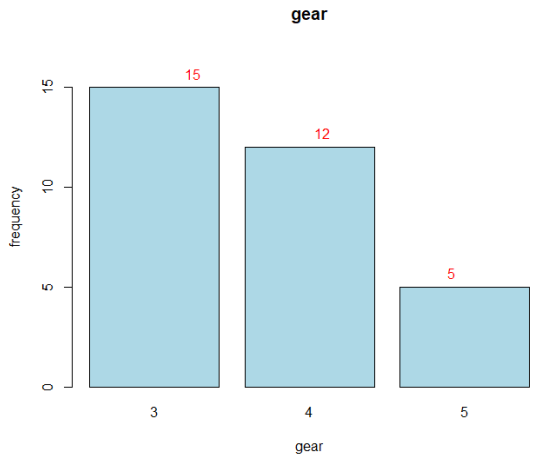
Horse Power (hp):

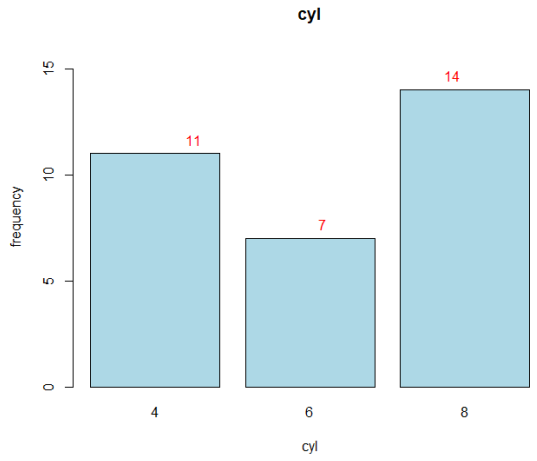
- The data is right skewed as depicted by the histogram and is confirmed by a negative skewness value.
- An outlier is seen in the box plot at the right extreme and the presence of more outliers on the right than on the left can be confirmed by the lower value of the trimmed mean as compared to the actual mean.

Miles per Gallon (mpg):

- The data is right skewed as depicted by the histogram and is confirmed by a negative skewness value.
- There are no outliers in the box plot. However, a small difference in the trimmed mean depict right skewness.

Categorical Data





- **Gears:** *represents the configurations of the gears*

Of the available **[3,4,5]** gear configurations, **3 gears** is the most common with presence in **46.9%** cars while the **5 gears** configuration is the least common with only presence in **15.6%** of vehicles.

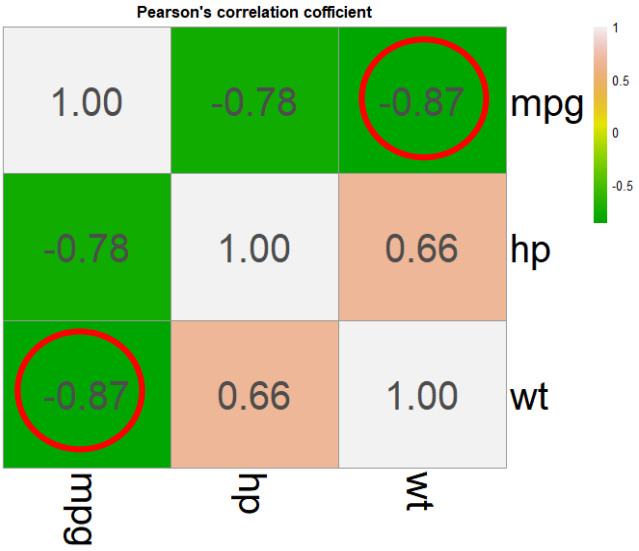
- **Cylinders:** *represents the number of cylinders*

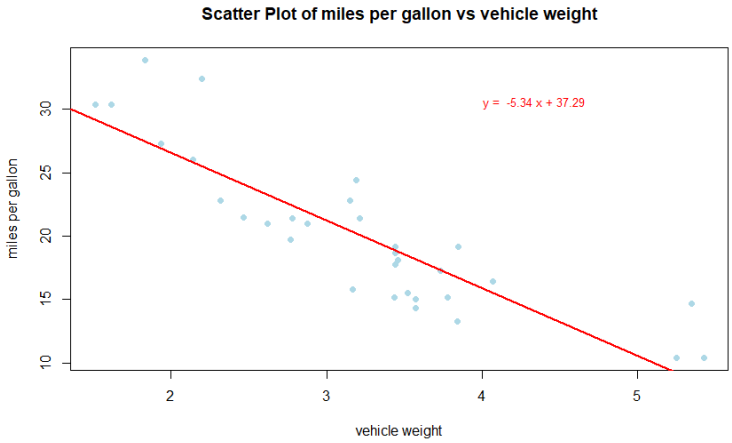
Of the available **[4,6,8] cylinder** configurations, **8 cylinder** is the most common with presence in **43.7%** of cars while the **6 gears** configuration is the least common with only presence in **21.9%** of vehicles.

Dataset: MTCARS

Multivariate Analysis

Correlation matrix based on **Pearson's formula** was plotted for all the numerical variables and plotted using a heat map. Based on the correlation matrix a negative correlation is observed **between miles per gallon(mpg)** and **weight of the vehicle**. Thus a linear regression was carried out for it.



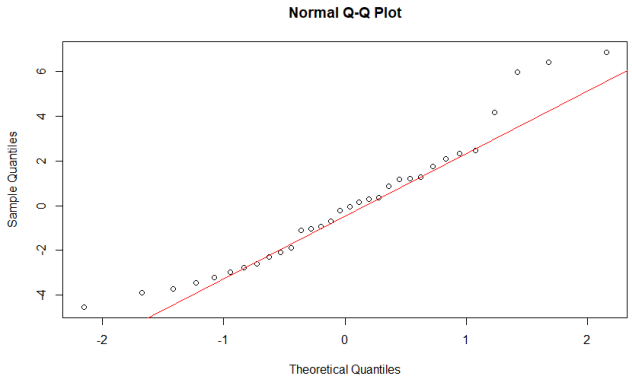


$$mpg = -5.34\,wt + 37.29$$

Observations from the model

- $Residual\ error = 3.05$
- $R^2 = 0.75$
- $F - Statistic = 91.4$
- $p - value = 1.9\,e^{-10}$

Weight of the vehicle significantly covers almost 75% of the variance in mpg. With more the weight, lesser the mpg. Also the QQ plot for residuals is close to the theoretical normal QQ plot except for some deviations in the extreme right.

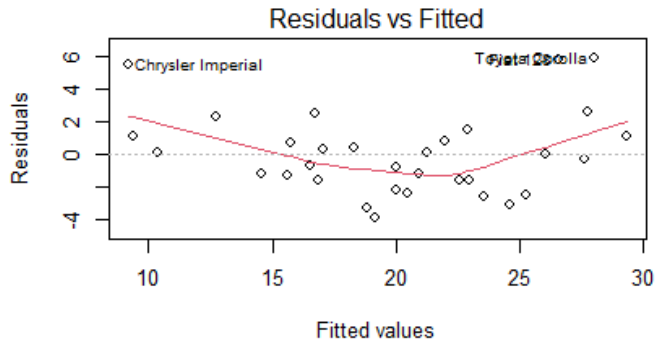


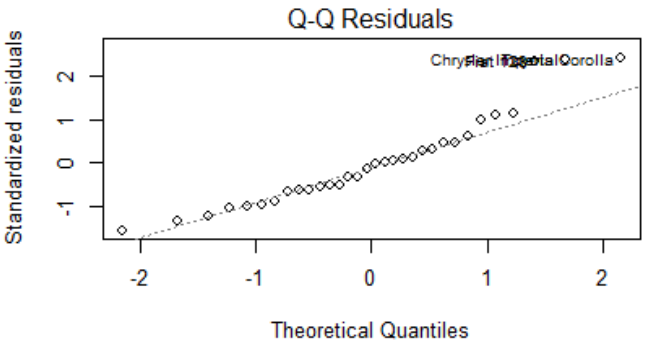
Multiple regression

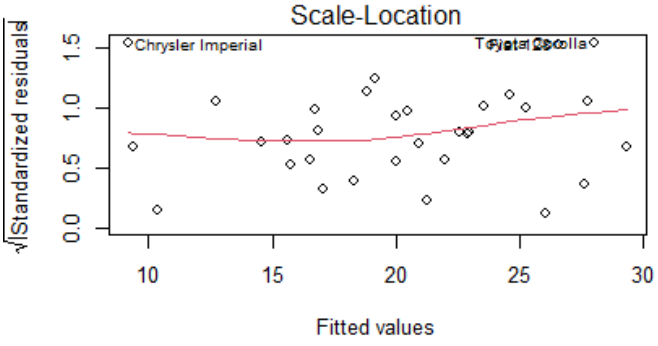
	Estimate	Std.Error	t-value	Pr(>|t|)
(Intercept)	37.22727	1.59879	23.285	<2E-16
hp	-0.03177	0.00903	-3.519	0.00145
wt	-3.87783	0.63273	-6.129	1.12E-06

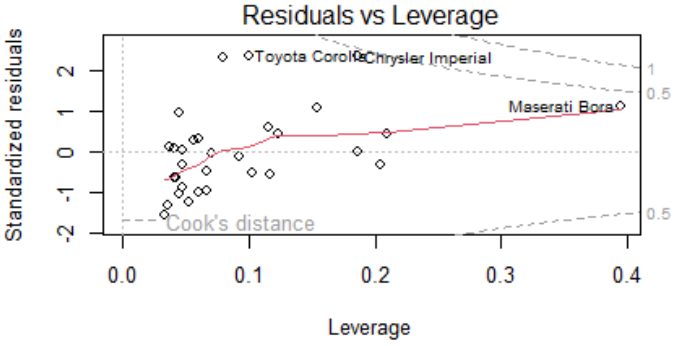
- **'hp'** and **'wt'** have very low **p-values (less than 0.05)**, indicating that both are statistically significant in explaining the variation in '**mpg**'. Also, the **p-value** for the **F-statistic is 9.109e-12,** indicating that the overall regression model is highly significant.
- **Multiple R-squared is 0.8268**, hence about 82.7% of the variation in mpg is explained by the combination of **'hp'** and **'wt'**. This indicates a strong fit of the model to the data.
- Residuals are randomly scattered but show minor deviations from the expected pattern. slight curvature in the residuals suggests potential non-linearity.
- Most residuals align with the diagonal line, indicating approximate normality.
- The spread of residuals remains almost constant, indicating near-homoscedasticity. However, slight upward trend at right values suggests mild heteroscedasticity.
- Points like Chrysler Imperial, Toyota Corolla, and Maserati Bora exhibit high leverage or influence. Maserati Bora lies near Cook's distance contours, so it is a highly influential observation.

Dataset: MTCARS

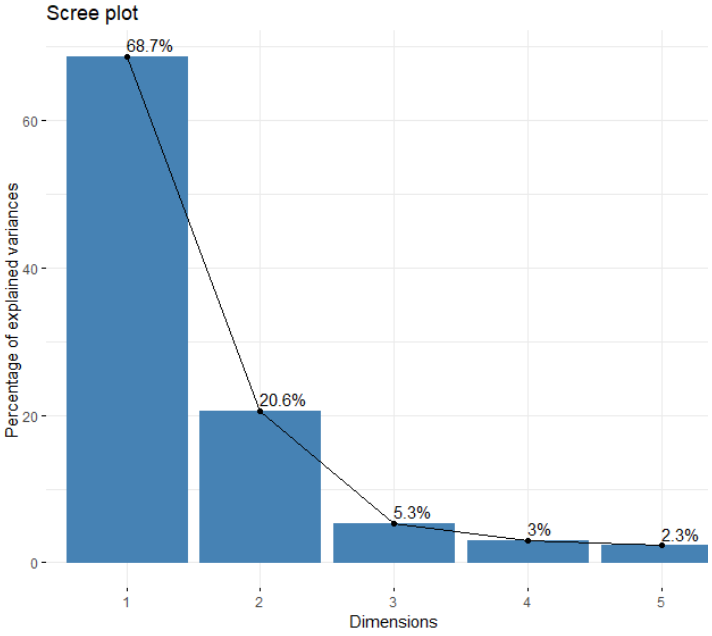


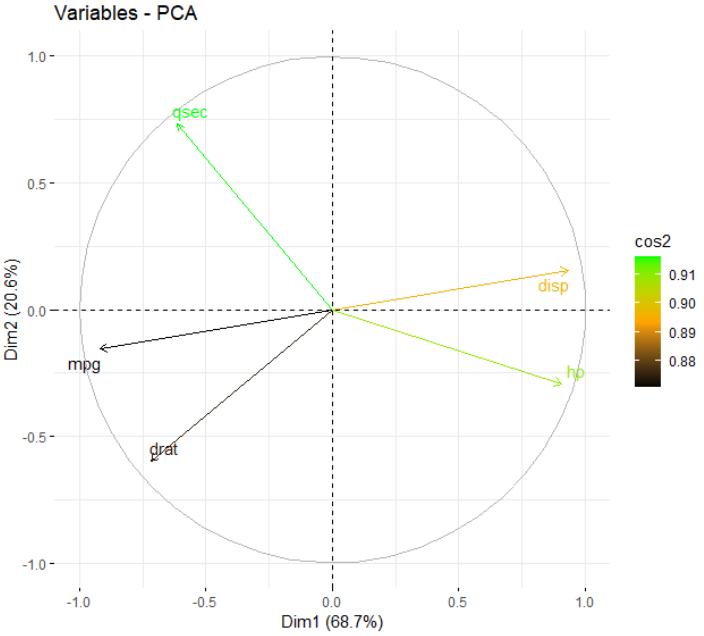






Advanced Analysis (PCA)





- As per the scree plot, first **3 Principal components** cover about **95%** of the variance and should be included for final analysis. However, for the ease of visualization only the first 2 PCs have been shown which covers about **89.3%** of total variation.
- Based on the relative **cos2** values in Bi plot, **'diplacement(disp)'** and **'horsepower(hp)'** contributes greatly to **PC1** while **'miles per gallon(mpg)'** and **'drat'** contribute significantly towards **PC2**

Dataset: Diamonds

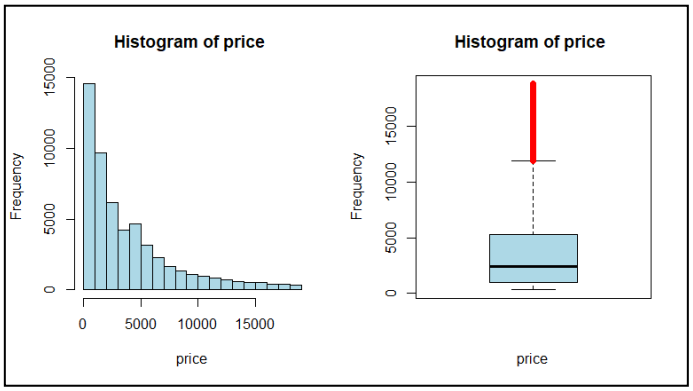
CODE

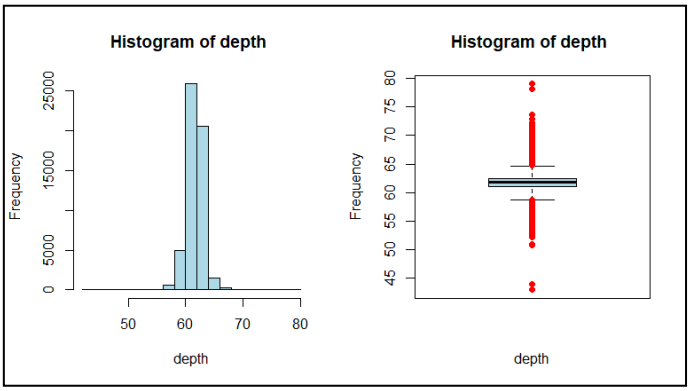
Data Overview

The diamonds dataset is a built-in dataset available in the ggplot2 package in R. It contains information on 53,940 diamonds, with 10 variables describing their physical characteristics and pricing. These variables include both categorical and numerical attributes, such as the carat (weight of the diamond), cut (quality of the cut, e.g., "Ideal," "Premium"), color (grading from "D" to "J"), clarity (grading of diamond clarity, e.g., "IF," "VS1"), price (in US dollars), as well as physical dimensions like x, y, and z (length, width, and depth).

Univariate Analysis

Numerical Data





	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
carat	0.8	0.5	0.7	0.7	0.2	5.0	4.8	1.1	1.3
depth	61.7	1.4	61.8	61.8	43.0	79.0	36.0	-0.1	5.7
table	57.5	2.2	57.0	57.3	43.0	95.0	52.0	0.8	2.8
price	3932.8	3989.4	2401.0	3159.0	326.0	18823.0	18497.0	1.6	2.2
x	5.7	1.1	5.7	5.7	0.0	10.7	10.7	0.4	-0.6
y	5.7	1.1	5.7	5.7	0.0	58.9	58.9	2.4	91.2
z	3.5	0.7	3.5	3.5	0.0	31.8	31.8	1.5	47.1

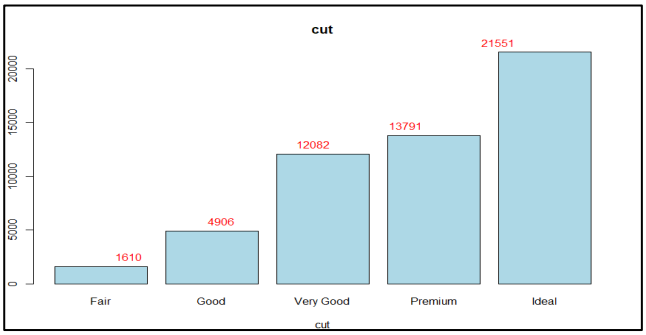
Price:

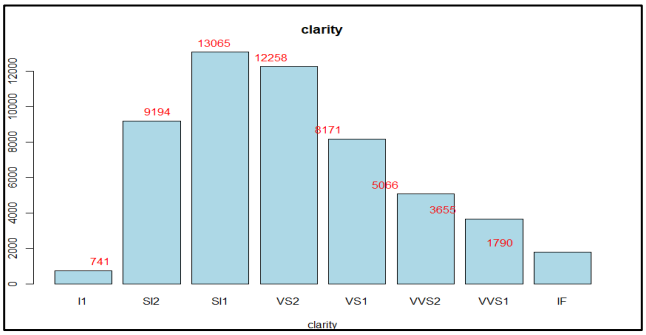
- The data is highly right-skewed, as shown by the histogram, indicating a concentration of lower-priced diamonds and a gradual tapering of higher prices.
- The box plot confirms the presence of numerous outliers at the higher end, suggesting that a small subset of diamonds is priced significantly higher than the majority.

Depth:

- The histogram shows that the depth data is nearly symmetric, with most values clustered around the centre.
- The box plot reveals a few outliers on both extremes

Categorical Data





Dataset: Diamonds

- **Cut:** *quality of the diamond's cut, ranked from Fair (lowest) to Ideal (highest)*

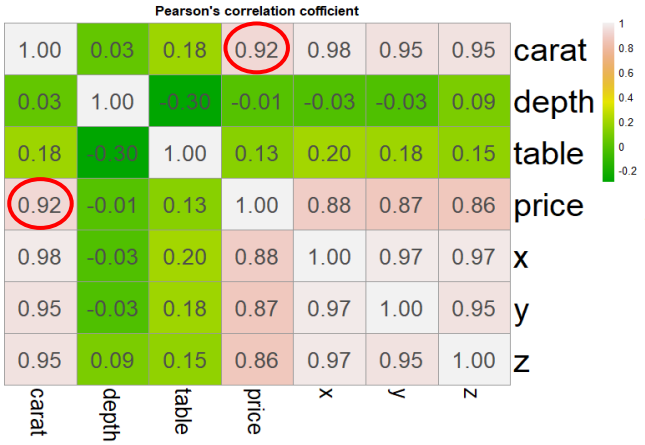
Ideal cut is the most common found cut whereas Fair is the least common cut. This suggest low quality cuts are rare and Ideal quality cuts are preffered

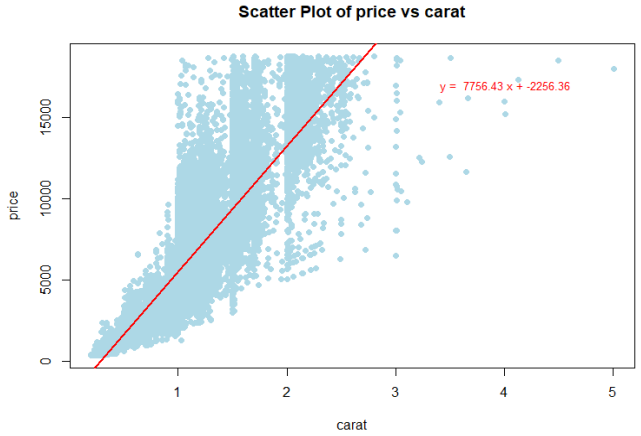
- **Clarity:** *quality of the diamond's clarity, ranked from I1 (lowest) to IF (highest)*

S1 is the most common clarity while I1 is the least common clarity. Indicating very low cut and very high cut diamonds are rare

Multivariate Analysis

Correlation matrix based on pearson's formuls was plotted for all the numerical variables and plotted using a heat map. Based on the correlation matrix a strong positive correlation is observed between **'Price'** of the diamond and it's weight in **'Carats'.** This indicates that as weight of the diamond increases the price increases and vice-versa.



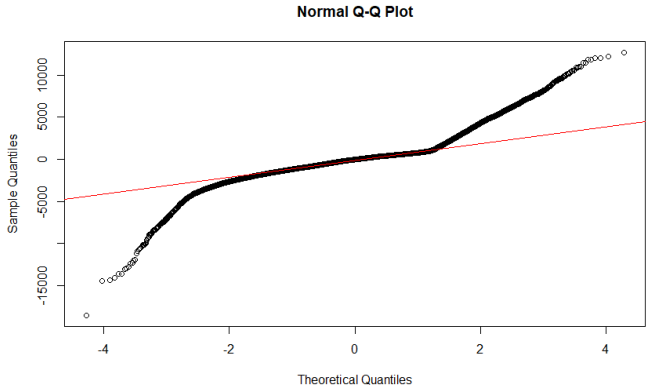


$$price = 7756\,wt + 37.29$$

Observations from the model

- $Residual\ error = 1549$
- $R^2 = 0.85$
- $F - Statistic = 3.04E^5$
- $p - value < 2.2\ E^{-16}$

Weight of the diamond significantly covers almost 85% of the variance in price. With more the weight, higher the price. However, the QQ-plot deviates from the theoretical values at both ends, suggesting some presence of non homoscedasticity



Multiple regression

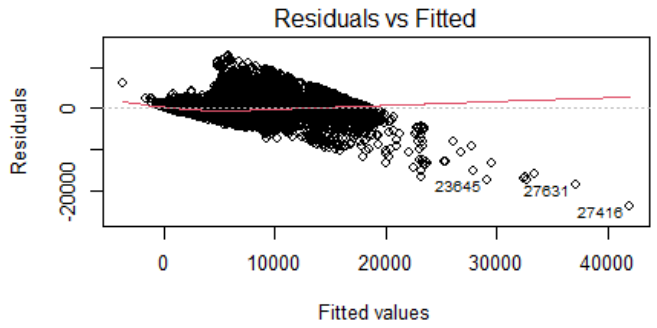
- All the parameters have very low **p-values (less than 0.05)**, indicating that all are statistically significant in explaining the variation in '**price'**. Also, the **p-value** for the **F-statistic is 9.109e-12,** indicating that the overall regression model is highly significant.
- **Multiple R-squared is 0.86**, hence about 86% of the variation in price is explained by the combination of all the variables. This indicates a strong fit of the model to the data.

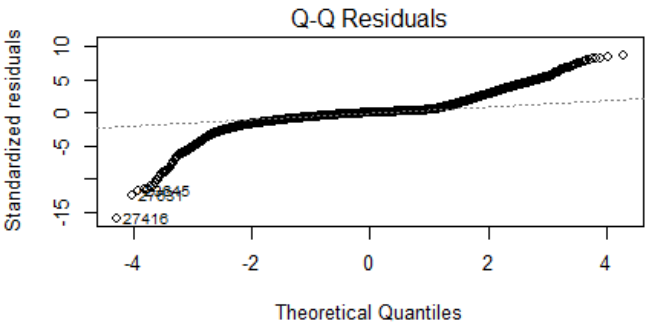
	Estimate	Std.Error	t-value	Pr(>|t|)
(Intercept)	20849.316	447.562	46.584	<2e-16
carat	10686.309	63.201	169.085	<2e-16
depth	-203.154	5.504	-36.910	<2e-16
table	-102.446	3.084	-33.216	<2e-16
x	-1315.668	43.070	-30.547	<2e-16
y	66.322	25.523	2.599	0.00937
z	41.628	44.305	0.940	0.347

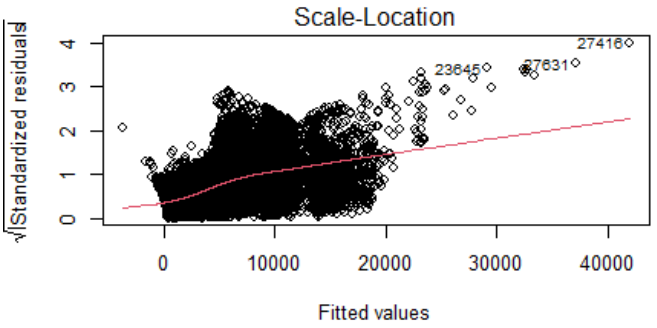
- There is a non-random pattern in the residuals, with a curved trend. Suggesting that the relationship between predictors and response is not entirely linear. and higher fitted values indicate heteroscedasticity.
- The residuals deviate significantly from the diagonal line, especially at the tails. This indicates that the residuals are not normally distributed.

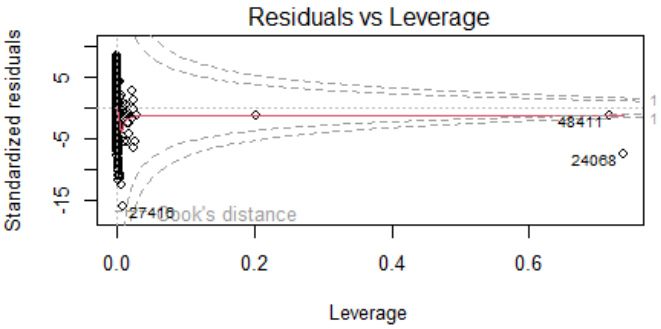
Dataset: Diamonds

- The points are not uniformly scattered around the horizontal line, suggesting that the model's assumptions about constant variance are violated.
- Other points such as **24068** and **48411** are also close to or beyond the Cook's distance lines, indicating that they may be influencing the regression results disproportionately.

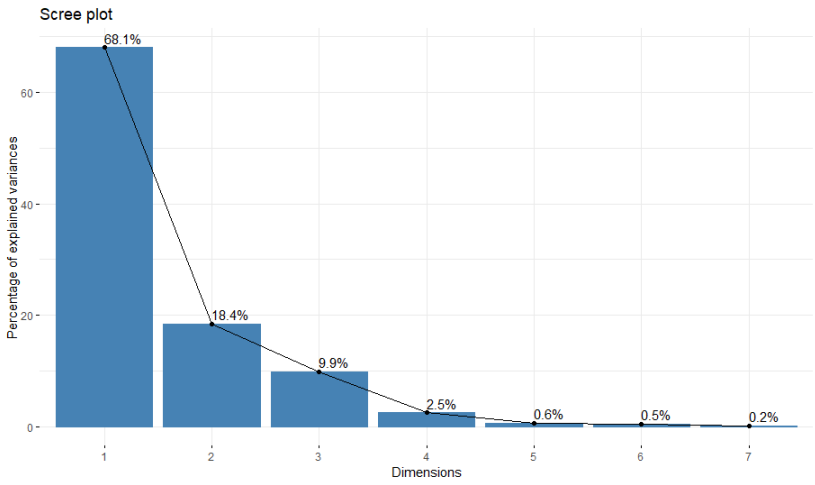


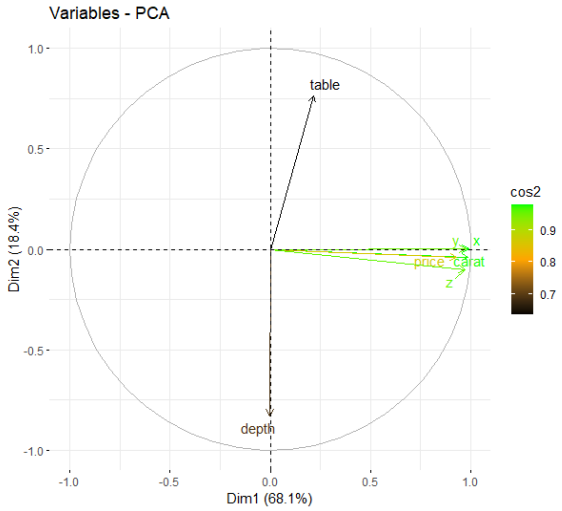






Advanced Analysis (PCA)





- As per the scree plot, first **3 Principal components** cover about **96.4%** of the variance and should be included for final analysis. However, for the ease of visualization only the first 2 PCs have been shown which covers about **86.5%** of total variation.
- Based on the relative **cos2** values in Bi plot, **'depth'** contributes greatly to **PC1** while **' X', ' y'** and **'carat'** contribute significantly towards **PC2**

Dataset: UScereal

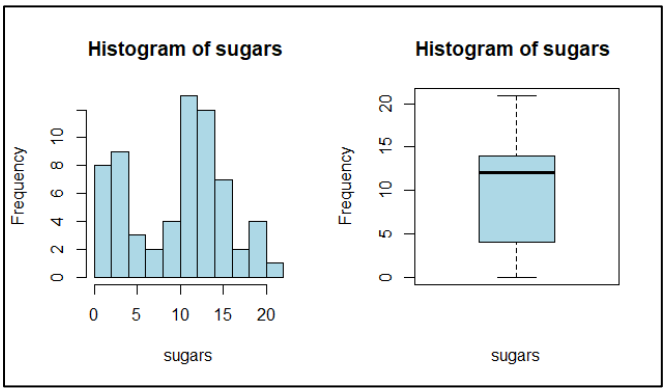
CODE

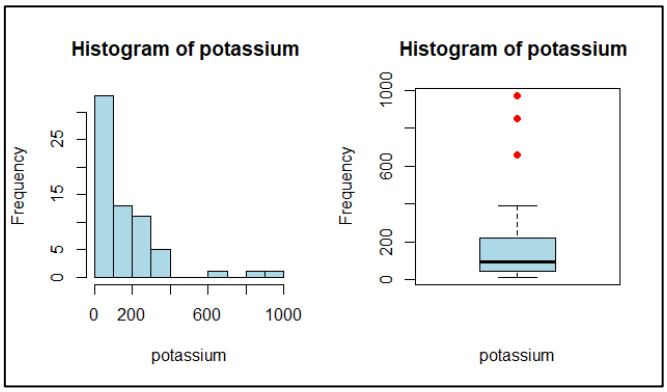
Data Overview

The **UScereal** dataset, available in the **MASS** package, contains nutritional information on 65 breakfast cereals marketed in the U.S. It provides details on **11 attributes** such as calories, protein, fat, sodium, fiber, carbohydrates, sugars, potassium, and vitamins per serving. Additional variables include the manufacturer (mfr), shelf location (shelf), serving weight (weight), and the number of cups per serving (cups).

Univariate Analysis

Numerical Data





	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
calories	149.4083	62.41187	134.3284	141.0889	50	440	390	2.189706	6.941696
protein	3.683705	2.642618	3	3.254163	0.75188	12.12121	11.36933	1.5677	2.476064
sodium	237.8384	130.6296	232	235.3153	0	787.8788	787.8788	1.251926	4.940988
fibre	3.870844	6.133404	2	2.565716	0	30.30303	30.30303	3.004274	9.377083
sugars	10.05084	5.835239	12	10.08196	0	20.89552	20.89552	-0.21041	-1.08465
potassium	159.1197	180.2886	96.59091	125.6354	15	969.697	954.697	2.565113	7.687938
vitamins	1.969231	0.352191	2	2	1	3	2	-0.44561	4.798646

Potassium:

- The histogram of potassium shows a strong right skew, indicating that most cereals have low potassium content, with a few having much higher levels.
- The box plot highlights significant outliers on the upper end, suggesting that some cereals have exceptionally high potassium compared to the majority.

Sugars:

- The histogram of sugars displays a more uniform spread, with a slight concentration around mid-range values.
- The box plot does not indicate any outliers, suggesting a consistent distribution of sugar content across the cereals.

Categorical Data

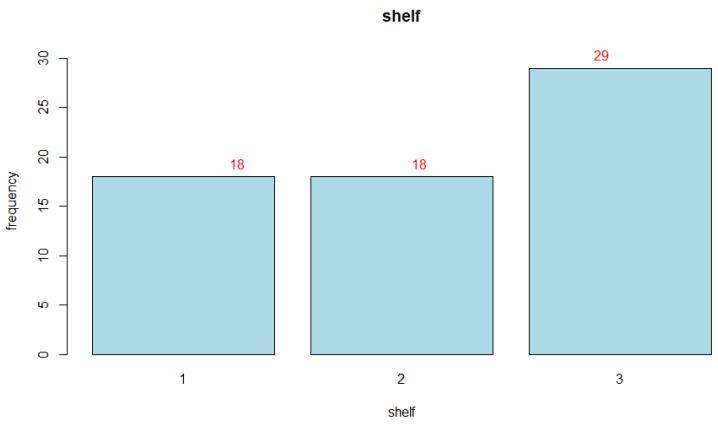
Shelf: *location of the cereal in a grocery store (1 = lower shelf, 2 = middle shelf, 3 = top shelf)*

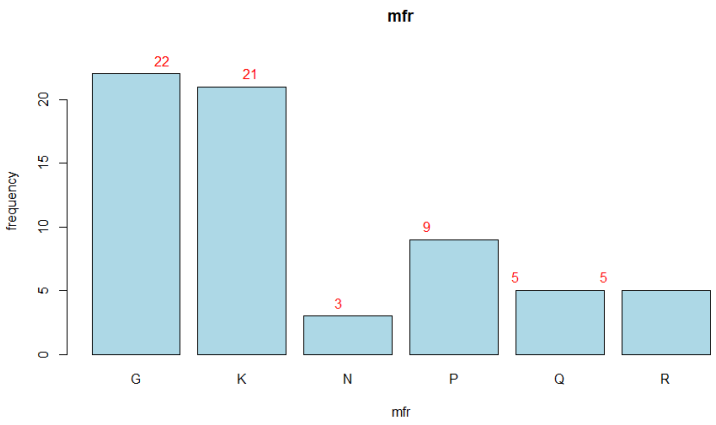
Most of the cereals are kept in the top shelf with rest equally divided between middle and bottom shelves. Suggesting preferred location of placement is the top shelf, with no specific preference among middle and bottom shelves

Manufacturer(mfr): *encoded as a single-letter abbreviation for the company*

There is close competition in the manufacturer with top 2, abbreviated as G & K being separated by a single unit. These two have presence significantly more that any of the other manufacturers.

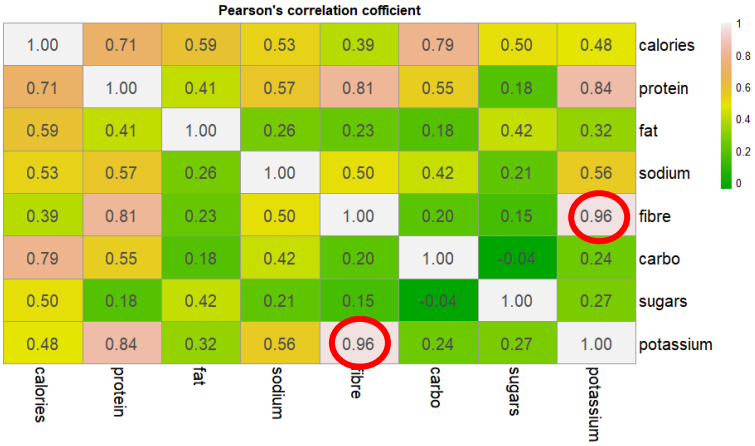
Dataset: UScereal

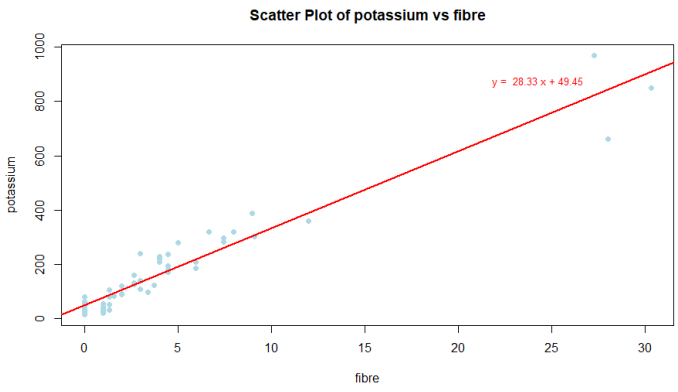




Multivariate Analysis

Correlation matrix based on Pearson's formula was plotted for all the numerical variables and plotted using a heat map. Based on the correlation matrix a strong positive correlation is observed between **'fibre content'** of the cereal and it's **'potassium'.** This indicates that as fibre content of the cereal increases the potassium content increases and vice-versa.



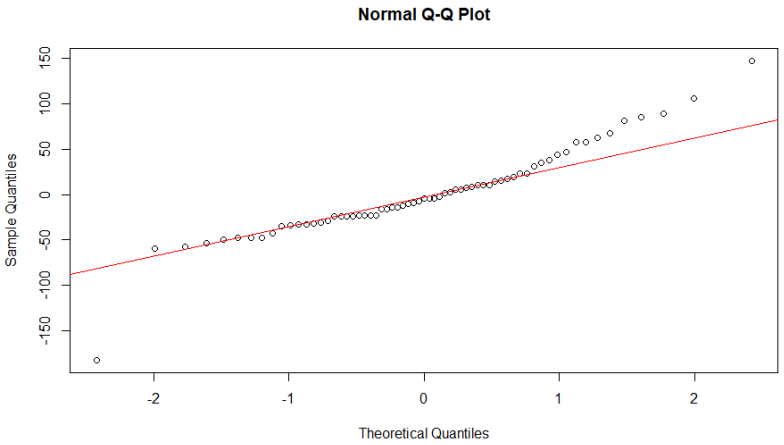


$$potassium = 28.33\ fibre + 49.44$$

Observations from the model

- $Residual\ error = 48.41$
- $R^2 = 0.929$
- $F - Statistic = 824.8$
- $p - value < 2.2\ E^{-16}$

Fibre content of the cereal significantly covers almost 93% of the variance in potassium content. With more the fibre content, higher the potassium. However, the QQ-plot deviates from the theoretical values towards the higher values.



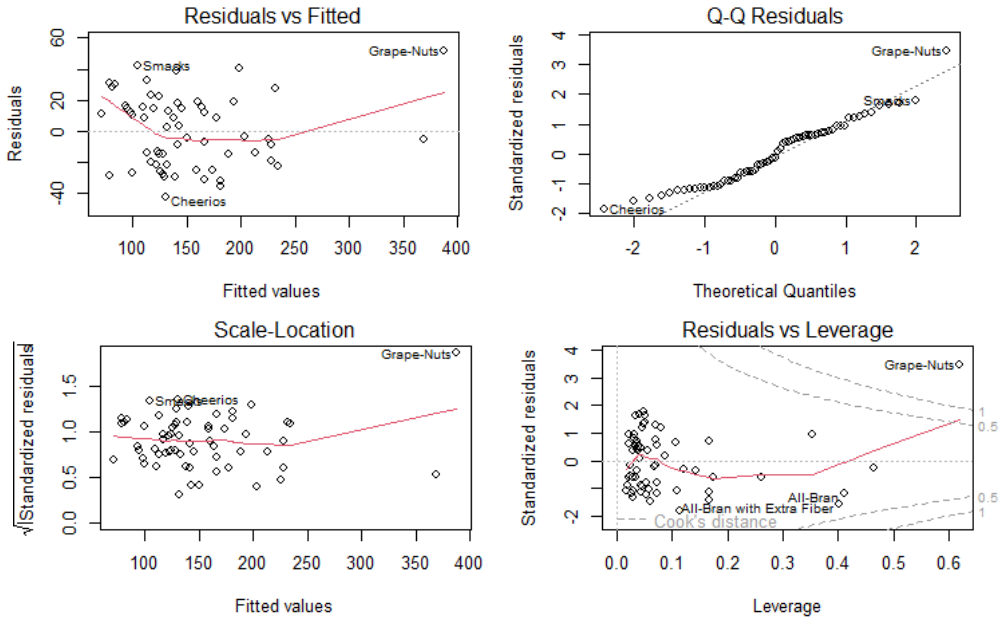
Multiple regression

- **'fat'** and **'carbohydrates'** the parameters have very low **p-values (less than 0.05)**, indicating that they are statistically significant in explaining the variation in '**calories**'. However, rest of the parameters have high p-values and are insignificant

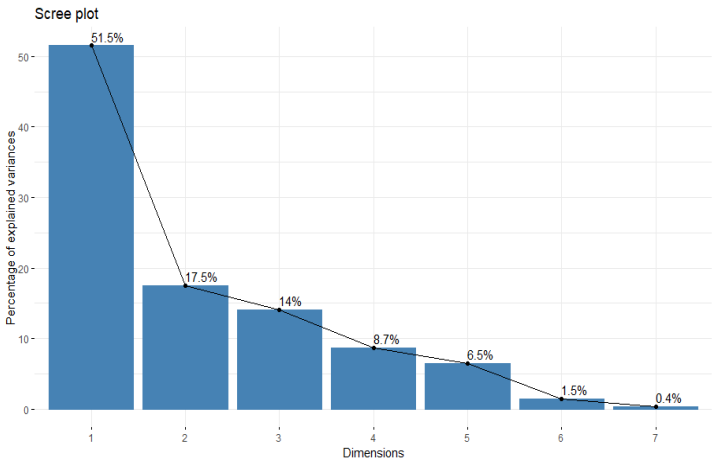
	Estimate	Std.Error	t-value	Pr(>|t|)
(Intercept)	17.84323	8.38812	2.127	0.0376
protein	4.68958	2.88909	1.623	0.109
fat	14.62685	2.10793	6.939	3.43e-09
sodium	0.03667	0.02928	1.252	0.2154
fibre	-0.15947	1.01948	-0.156	0.8762
carbo	4.27582	0.51351	8.327	1.54e-11

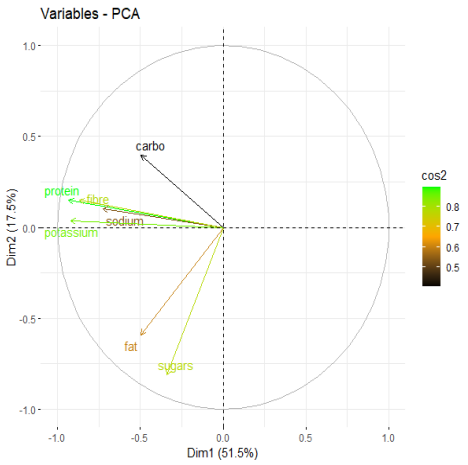
Dataset: UScereal

- **Multiple R-squared is 0.86**, hence about 86% of the variation in price is explained by the combination of all the variables. This indicates a strong fit of the model to the data.
- There is a slight nonlinear pattern, indicating potential model misspecification or omitted variables.
- Most data points align with the diagonal line, indicating approximate normality of residuals.
- Variance seems to increase with fitted values, showing heteroscedasticity.
- Data points such as "Grape-Nuts" and "Smacks" show high leverage, aligning with earlier observations.
- "Grape-Nuts" has high leverage and is influential, as highlighted by Cook's distance.



Advanced Analysis (PCA)





- As per the scree plot, first **4 Principal components** cover about **91.7%** of the variance and should be included for final analysis. However, for the ease of visualization only the first 2 PCs have been shown which covers only about **69%** of total variation.
- Based on the relative **cos2** values in Bi plot, **'fats'** and **'sugars'** contributes greatly to **PC1** while **' fibre', ' potassium'** and **'sodium'** contribute significantly towards **PC2**

Dataset: Boston

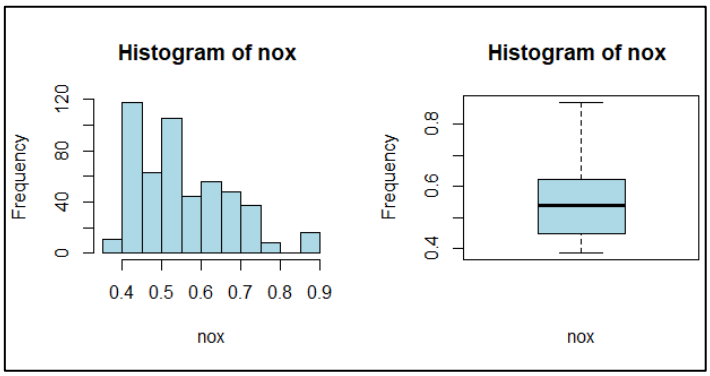
CODE

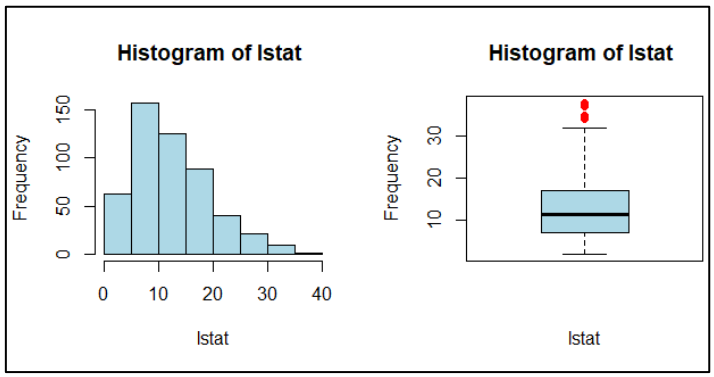
Data Overview

The **Boston Housing dataset** contains **506 observations** (data points) and **14 variables** in total, including 13 numerical attributes and 1 target variable. These attributes describe various socio-economic and housing-related factors for different areas in the Boston region.

Univariate Analysis

Numerical Data





	Mean	Standard Deviation	Median	Trimmed Mean	Minimum	Maximum	Range	Skewness	Kurtosis
nox	0.5546951	0.1158777	0.538	0.5450601	0.385	0.871	0.486	0.7249897	-0.08741064
tax	408.2371542	168.5371161	330	400.044335	187	711	524	0.6659891	-1.15031761
indus	11.1367787	6.8603529	9.69	10.9318719	0.46	27.74	27.28	0.2932747	-1.2401949
ptratio	18.4555336	2.1649455	19.05	18.6625616	12.6	22	9.4	-0.7975743	-0.304801
black	356.6740316	91.2948644	391.44	383.1695074	0.32	396.9	396.58	-2.8732597	7.10371496
lstat	12.6530632	7.1410615	11.36	11.8990394	1.73	37.97	36.24	0.9010929	0.46281705
nox	0.5546951	0.1158777	0.538	0.5450601	0.385	0.871	0.486	0.7249897	-0.08741064

Nitric Oxides (nox):

- The histogram of "nox" exhibits a left-skewed distribution, suggesting that most neighborhoods in the dataset have moderate nitric oxide concentration levels, with fewer areas showing high concentrations.
- The box plot indicates a symmetrical distribution without noticeable outliers, suggesting a relatively consistent range of "nox" values.

Lower Status of the Population (lstat):

- The histogram of "lstat" shows a right-skewed distribution, where most neighborhoods have lower percentages of the lower socioeconomic status population, while a few neighborhoods have significantly higher percentages.
- The box plot highlights the presence of significant outliers on the upper end, indicating that some neighborhoods deviate markedly from the majority in terms of "lstat" values.

Categorical Data

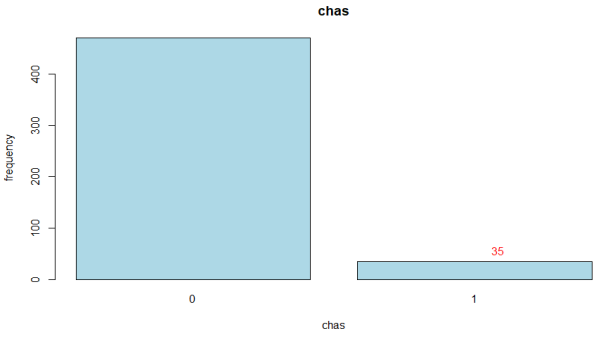
- **Chas:** *binary variable indicating whether the property is adjacent to the Charles River*

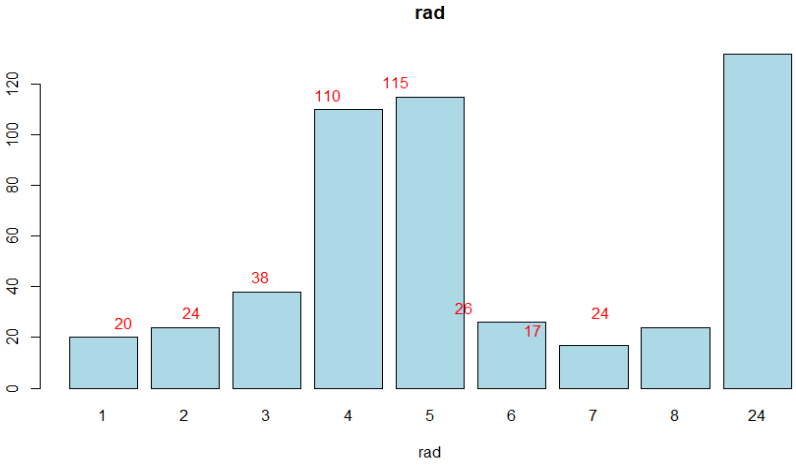
Major proportion of theproperties bound the Charles river and only a handful of them are away from it

- **Rad:** *index indicating accessibility to radial highways*

Maximum number of properties have a good connectivity to the highways as the have an index off 24. However a major chunk of the properties have been indexed 4 & 5 indicating connection to highway is not prominent

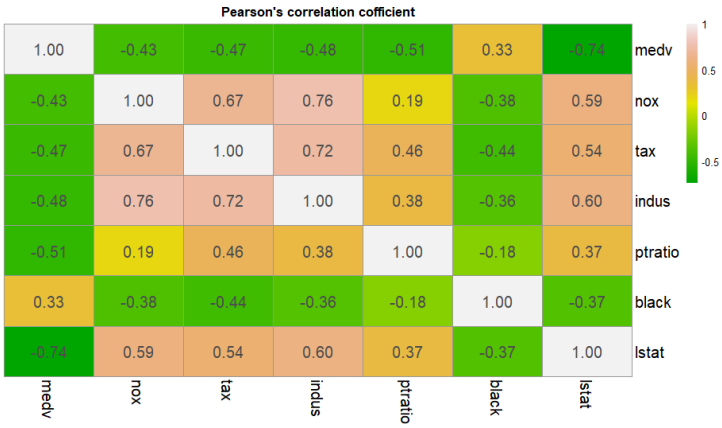
Dataset: Boston

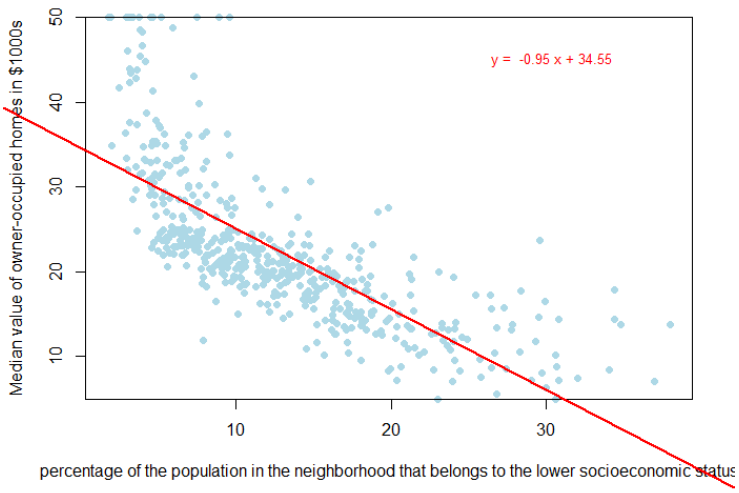




Multivariate Analysis

Correlation matrix based on Pearson's formula was plotted for all the numerical variables and plotted using a heat map. Based on the correlation matrix a negative correlation is observed between **'medv'** and **'lstat'.**

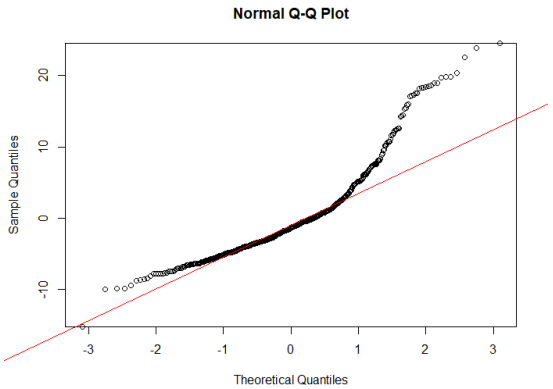




Observations from the model

- $\boldsymbol{Residual\ error = 6.12}$
- $\boldsymbol{R^2 = 0.54}$
- $\boldsymbol{F - Statistic = 601.6}$
- $\boldsymbol{p - value < 2.2\ E^{-16}}$

R squared value is too low indicating only 54% of the variation is explained by the model. Further the QQ plot varies from the theoretical plot suggesting non-homoscedasticity.



Multiple regression

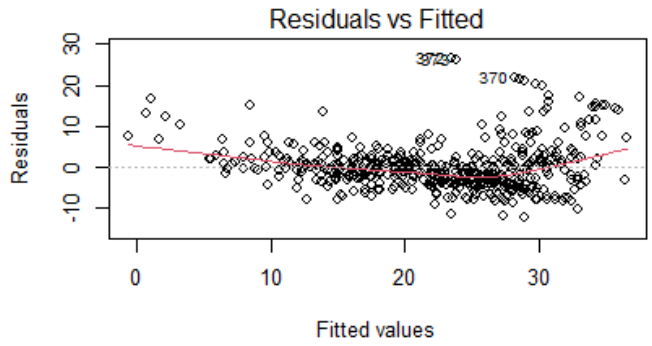
- **'ptratio'** and **'lstat'** have very low **p-values (less than 0.05)**, indicating that they are statistically significant in explaining the variation in **'medv'**. However, rest of the parameters have high p-values and are insignificant
- **Multiple R-squared is 0.62**, hence only 62% of the variation in price is explained by the combination of all the variables. This indicates a strong fit of the model to the data.

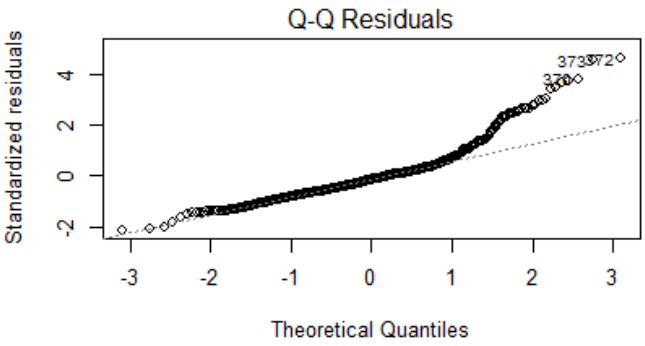
	Estimate	Std.Error	t-value	Pr(>|t|)
(Intercept)	51.28	3.43	14.95	<2e-16
nox	-0.13	3.84	-0.03	0.9728
tax	0.00	0.00	0.54	0.5873
indus	0.01	0.07	0.16	0.8705
ptratio	-1.17	0.14	-8.19	2.25E-15
black	0.01	0.00	2.15	0.0317
lstat	-0.81	0.05	-16.57	<2e-16

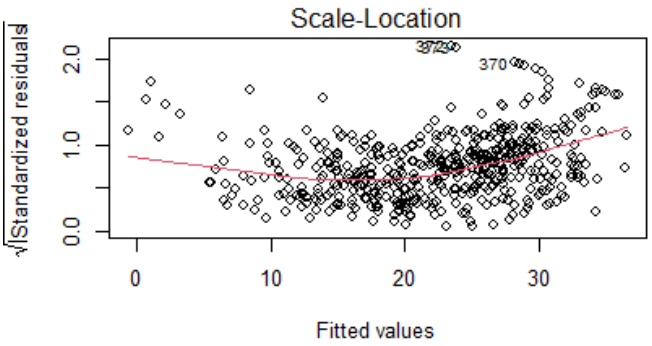
- The residuals show some nonlinear pattern, indicating potential issues with model specification.
- Residuals are more spread out for higher fitted values, suggesting possible heteroscedasticity.

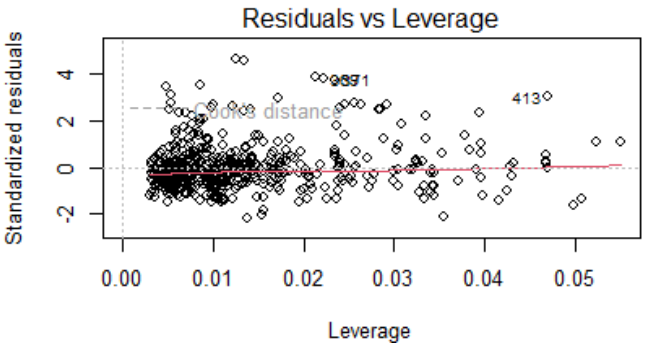
Dataset: Boston

- Most points follow the theoretical quantile line, suggesting residuals are approximately normal. However, Residuals are more spread out for higher fitted values, suggesting possible heteroscedasticity.
- Observations like "370," "372," and "413" show higher leverage, making them influential points.
- Cook's distance flags these observations, especially "370," as having a significant influence on the model.

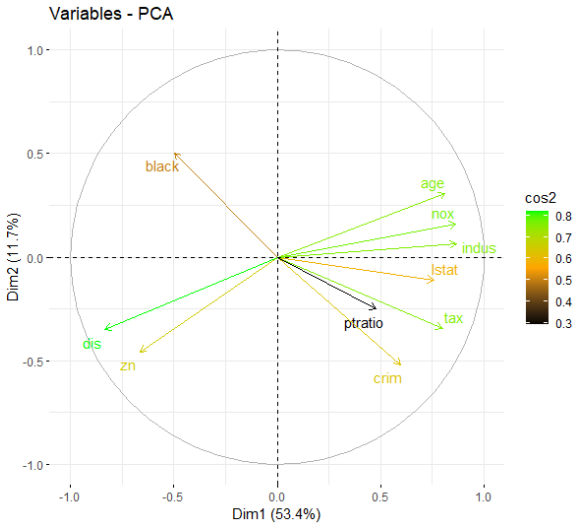


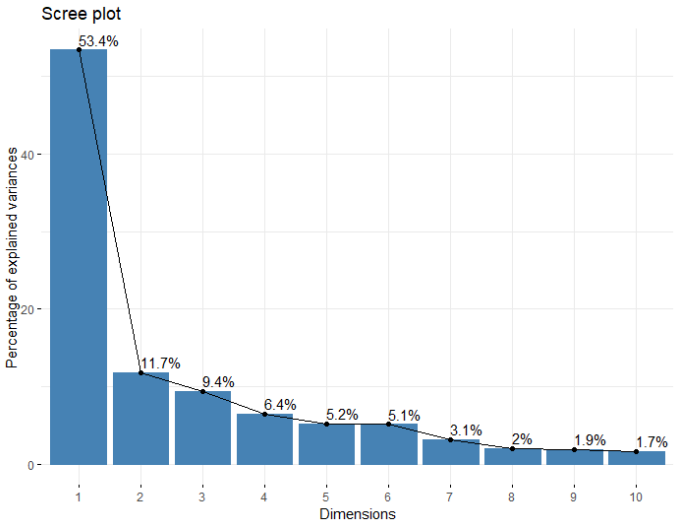






Advanced Analysis (PCA)





- As per the scree plot, first **6 Principal components** cover about **91.2%** of the variance and should be included for final analysis. However, for the ease of visualization only the first 2 PCs have been shown which covers only about **65.1%** of total variation.