STATISTICAL STRUCTURES IN DATA NUMERICAL ASSIGNMENT

PRANJAL CHAKRABORTY **Date**: 07-12-2024
24BM6JP41

Dataset 1: Carseats (R library: ISLR2)

Univariate Analysis

1. Data Overview

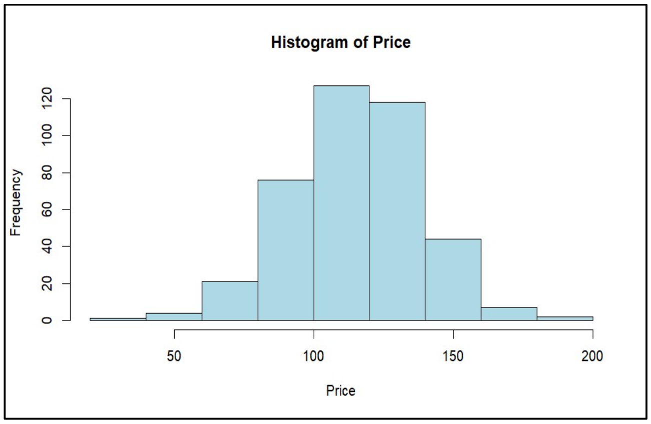
Column	Type	First 5 Observations				
Sales	numeric	9.5	11.22	10.06	7.4	4.15
CompPrice	numeric	138	111	113	117	141
Income	numeric	73	48	35	100	64
Advertising	numeric	11	16	10	4	3
Population	numeric	276	260	269	466	340
Price	numeric	120	83	80	97	128
ShelveLoc	Factor w/ 3 levels “Bad”, “Good”, “Medium”	1	2	3	3	1
Age	numeric	42	65	59	55	38
Education	numeric	17	10	12	14	13
Urban	Factor x/ 2 levels “Yes”,“No"	2	2	2	2	2
US	Factor x/ 2 levels “Yes”,“No"	2	2	2	2	1

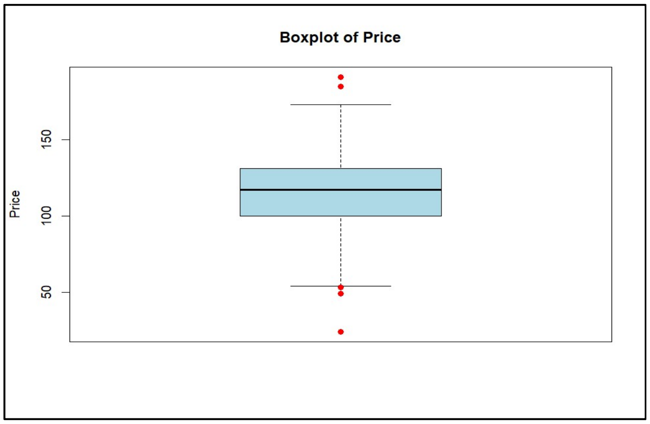
No.of observations = 400 No. of variables = 11

2. Summary Statistics (*Numerical variable: Price (in $)*)

Statistic	**Mean**	**Median**	**Std. Dev**	**Minimum**	**Maximum**
Value	115.795	117.0	23.676	24.0	191.0

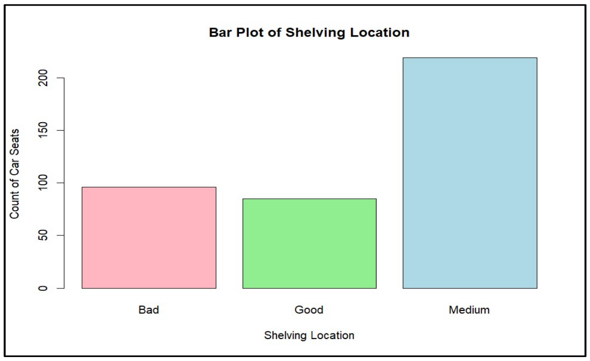
3. Distribution Visualization (*Price*)





The histogram shows that the shape of the distribution is **fairly symmetric and unimodal**, with a peak near the center. It indicates a **normal distribution** with a slight tendency toward **left skewness**.
The boxplot indicates five visible **outliers**, with P values extending beyond 175 on the upper end and below 50 on the lower end. The **interquartile range** (IQR = Q3 – Q1) is 30 (130 – 100).

4. Categorical Variable Analysis (*Categorical variable: ShelveLoc*)



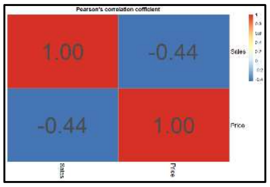
ShelveLoc: Represents the quality of the location of the shelving for the car seats in a store indicating how the placement/visibility of the product influences its sales.

Most frequent shelving location is "**Medium**" while "**Good**" has the **smallest proportion**. This could imply that stores may prefer to standardize shelving to an average quality level rather than optimizing for "Good" locations. Effort should be made to reduce the no. of "Bad" shelving to improve Sales.

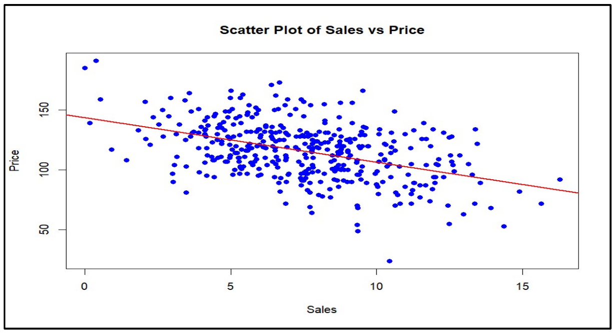
Multivariate Analysis

5. Correlation Analysis (*Sales vs Price*)

Pearson correlation coefficient between Sales and Price: **-0.44.**
It indicates a **moderate negative linear association**. Although correlation does not suggest causation, it hints that as Price increases, Sales volume tends to decrease.

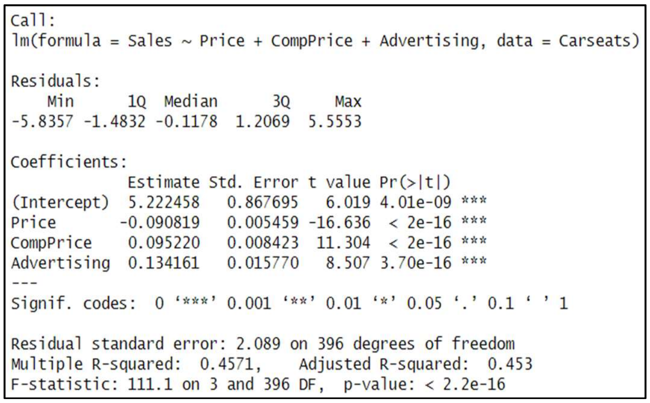


6. Scatter Plot Visualization

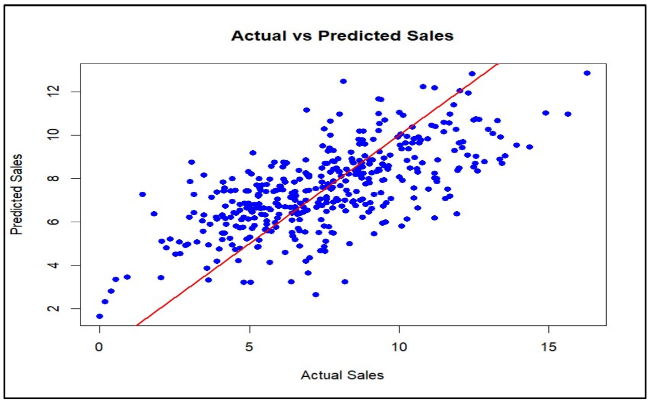


The scatter plot reveals a **moderate negative relationship** between Sales and Price, indicating that Price and Sales are inversely related.
The **downward sloping trend** line aligns with this observation, reinforcing the Pearson correlation value of -0.44.

7. Multiple Regression (*Target: Sales*)



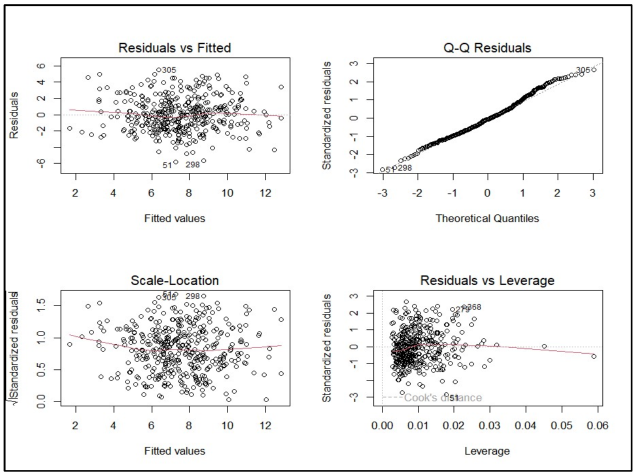
```
Call:
lm(formula = Sales ~ Price + CompPrice + Advertising, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.8357 -1.4832 -0.1178  1.2069  5.5553 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.222458   0.867695   6.019 4.01e-09 ***
Price       -0.090819   0.005459 -16.636  < 2e-16 ***
CompPrice    0.095220   0.008423  11.304  < 2e-16 ***
Advertising  0.134161   0.015770   8.507 3.70e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.089 on 396 degrees of freedom
Multiple R-squared:  0.4571,	Adjusted R-squared:  0.453 
F-statistic: 111.1 on 3 and 396 DF,  p-value: < 2.2e-16
```



Intercept: The baseline predicted value of dependent variable Sales when the predictor variables are set to 0 is 5.22. The p-value (4.01 e-09) suggests it is statistically significant.
Price: For every unit increase in Price, Sales is expected to decrease by 0.09 units, assuming other variables remain constant. This is the **most important predictor** with highest absolute t value (16.63).
CompPrice: For every 1-unit increase in the competitive price, the dependent variable (Sales) is expected to increase by approximately 0.095 units. The p-value (< 2e-16) indicates it is highly statistically significant.
Advertising: It has positive relationship with dependent variable (Sales), indicating that higher advertising leads to more sales. The p-value suggests is statistically significant in the prediction model.

The fitted plot shows that the model effectively predicts Sales, although a linear model is not the best fit. Adjusted for the number of predictors, about **45.71% (Adj R^2 = 0.457)** of the variation in Sales is explained by the model's predictors. **High F-statistic** and **low p-value** suggests model is statistically significant.

8. Model Diagnostics



Homoscedasticity

Residuals vs. Fitted Plot: Residuals are randomly scattered around zero with no clear pattern, suggesting homoscedasticity (constant variance of residuals).

Scale-Location Plot: The red line shows a slight bend, suggesting mild non-constant variance in residuals.

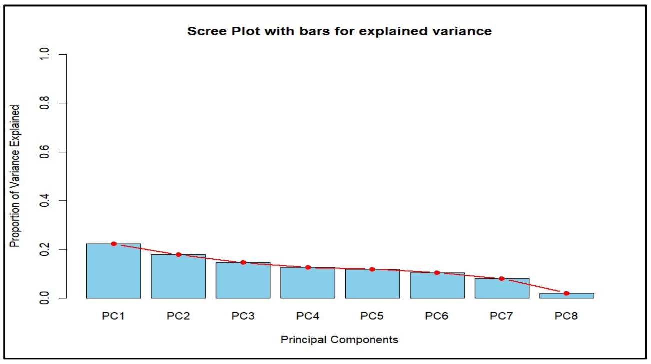
Normality of Residuals

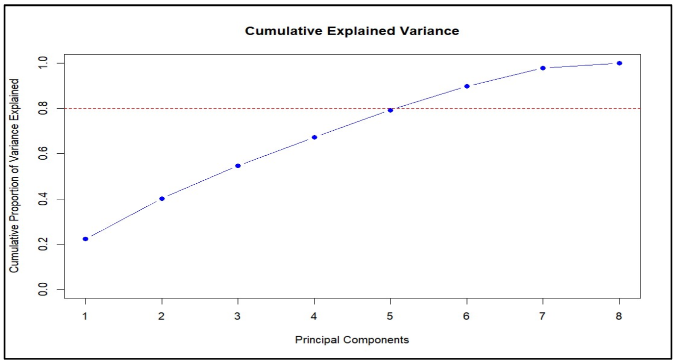
Q-Q Plot: Residuals mostly follow the diagonal, with slight deviations at the tails, suggesting near-normality.

Residuals vs. Leverage Plot: A few points have moderate leverage but do not strongly influence the model fit.

__Advance Analysis__

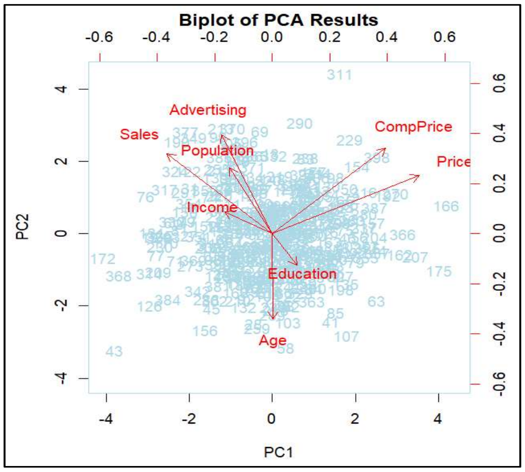
9. Principal Component Analysis



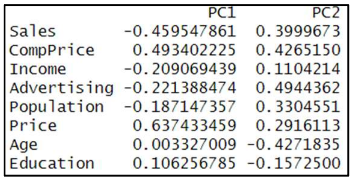


No. of PCs chosen = 5, as they collectively explain **80% of the variation** in the data. While the scree plot has **no clear elbow point**, the slope flattens after the fifth PC, suggesting diminishing returns beyond it.

10. PCA Interpretation



The biplot loadings reveals that **Price** and **CompPrice** strongly **influence PC1**, while **Advertising** and **Population dominate PC2**.



```
                   PC1        PC2
Sales       -0.459547861  0.3999673
CompPrice    0.493402225  0.4265150
Income      -0.209069439  0.1104214
Advertising -0.221388474  0.4944362
Population  -0.187147357  0.3304551
Price        0.637433459  0.2916113
Age          0.003327009 -0.4271835
Education    0.106256785 -0.1572500
```

Price and CompPrice (**Pricing group**) are positively correlated, as seen by their closely aligned arrows, while Age and Education (**demographic group**) negatively correlate with Sales and Advertising (**marketing group**).

Most points cluster near the center, but **outliers** such '311' (high PC2) and '43' (low PC1 & PC2) indicate distinct behaviours.

PC1 separates **pricing** features from sales and advertising, while **PC2** contrasts **marketing** variables with **demographic** factors.

11. Conclusion

Univariate analysis revealed Price has a symmetric distribution with some outliers, while ShelveLoc shows medium-quality shelving as most common. Multivariate analysis indicates Price has moderate negative correlation with Sales. Regression highlights Price as a major driver of Sales, with competitive pricing and advertising boosting performance. Model diagnostics suggest homoscedasticity and near-normality of residuals. PCA indicates 5 PCs explain 80% of variation and biplots suggests PCA separates pricing strategies from sales influencers and marketing variables from demographic factors. Insights suggest focusing on improving advertising and shelving quality while implementing effective pricing strategies to boost sales.

Dataset 2: Boston Housing (R library: MASS)

Univariate Analysis

1. Data Overview

Name	**crim**	**Zn**	**indus**	**chas**	**Nox**	**rm**	**age**	**dis**	**rad**	**tax**	**ptratio**	**lstat**	**medv**
Type	num	num	num	int	Num	num	num	num	int	num	num	num	Target
First 3 Obs.	0.006	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	*4.98*	*24*
	0.027	0	7.07	0	0.469	6.421	78.9	4.96	2	242	17.8	*9.14*	*21.6*
	0.027	0	7.07	0	0.469	7.185	61.1	4.96	2	242	17.8	*4.03*	*34.7*

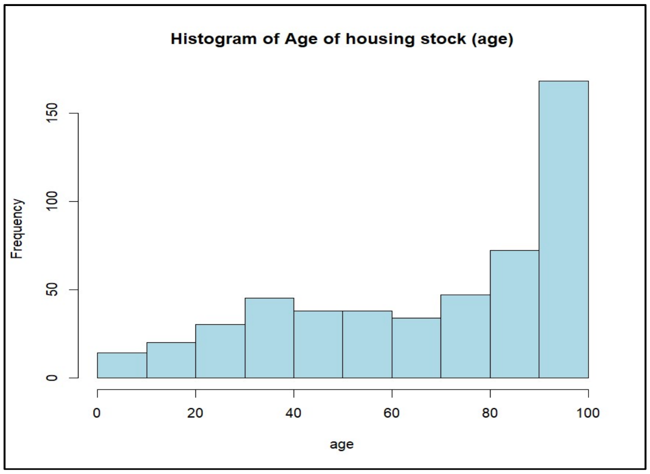
No. of observations = 506

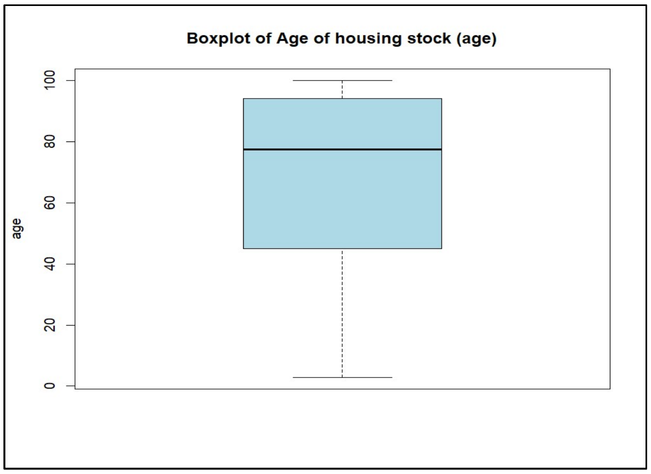
No. of variables = 13

2. Summary Statistics (*Numerical variable: age*)

Statistic	**Mean**	**Median**	**Std. Dev**	**Minimum**	**Maximum**
Value	68.57	77.5	28.15	2.9	100.0

3. Distribution Visualization (*age*)


Histogram of Age of housing stock (age)

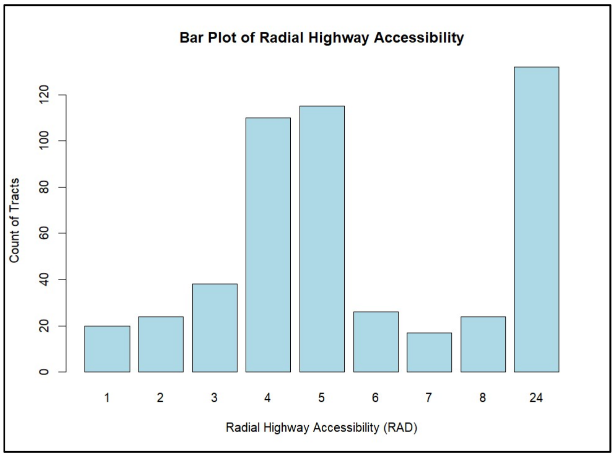

Boxplot of Age of housing stock (age)

The histogram shows that the distribution is **negatively-skewed**, with a higher concentration of data points in the older age ranges (80–100). Significant proportion of housing stock is relatively old (100 years).
The boxplot reveals that the data is **not symmetric**, with the median closer to the upper quartile, indicating a left-skewed distribution. There are no visible outliers. The **interquartile range** (IQR = Q3 – Q1) appears to be around **50** (95 – 45). The whiskers extend to the minimum and maximum ages without exceeding 100.

4. Categorical Variable Analysis (*rad*)

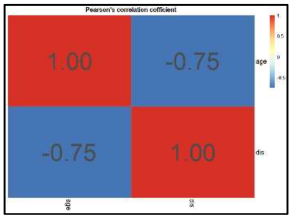
Radial Highway Accessibility (RAD) is a measure of how accessible a location is to highways that radiate outward from a central point.

The distribution of RAD values is **non-uniform** which suggests that many areas are less accessible to radial highways. RAD values **4,5, and 24** have significantly **high count** of tracts. The value of 24 stands out as an **outlier** which could represent a very unique location with exceptional highway access or an error.


Bar Plot of Radial Highway Accessibility

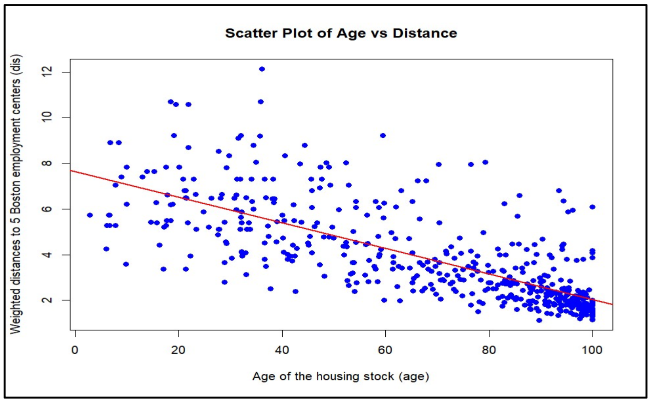
Multivariate Analysis

5. Correlation Analysis (age vs dis)



Pearson correlation coefficient between age and dis (Weighted distance to employment centers)**: - 0.748**. It indicates a **strong negative linear association**. Although correlation does not suggest causation, it hints that as *new housing stocks are being built close to employment centers while older housings are away*.

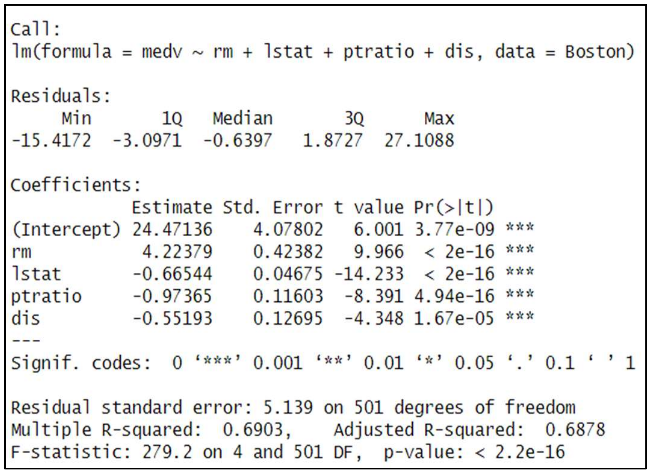
6. Scatter Plot Visualization



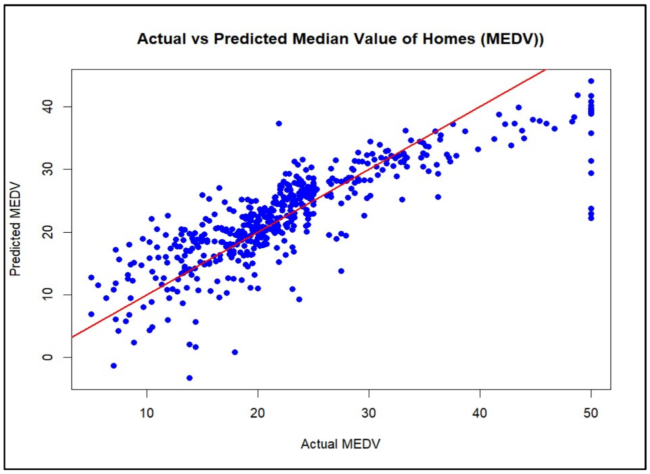
The scatter plot reveals a **fairly strong negative relationship** between 'age' and 'dis', indicating that age of housing and distance to employment centers are inversely related, consistent with the above analysis.

The **downward sloping trend** line aligns with this observation, reinforcing the Pearson correlation value of -0.748.

7. Multiple Regression (*Target: MEDV (Median value of owner-occupied homes in $1000s)*)



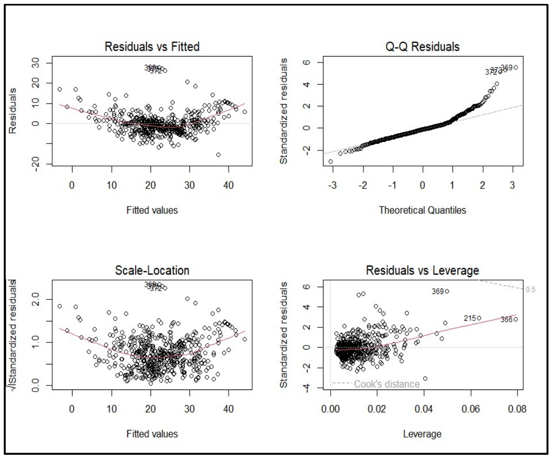
```
Call:
lm(formula = medv ~ rm + lstat + ptratio + dis, data = Boston)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.4172  -3.0971  -0.6397   1.8727  27.1088 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.47136    4.07802   6.001 3.77e-09 ***
rm           4.22379    0.42382   9.966  < 2e-16 ***
lstat       -0.66544    0.04675 -14.233  < 2e-16 ***
ptratio     -0.97365    0.11603  -8.391 4.94e-16 ***
dis         -0.55193    0.12695  -4.348 1.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.139 on 501 degrees of freedom
Multiple R-squared:  0.6903,	Adjusted R-squared:  0.6878 
F-statistic: 279.2 on 4 and 501 DF,  p-value: < 2.2e-16
```



Intercept: The baseline predicted value of dependent variable 'medv' when the predictor variables are set to 0 is 24.47. The p-value (3.77 e-09) suggests it is statistically significant.
rm: For every unit increase in 'rm' (Average number of rooms per dwelling), 'medv' is expected to increase by 4.22 units, assuming other variables remain constant. . This is a very significant predictor (absolute t value = 9.96) and the positive relationship indicates increase in number of rooms raises the housing stock value.
lstat: For every 1-unit increase in the % of lower-status population, the 'medv' is expected to decrease by 0.067 units. The p-value (<2e-16) and t-value (-14.23) indicates it is **most statistically significant predictor**.
ptratio: Pupil-teacher ratio by town has negative relationship with 'medv', indicating that lower ratio (more teachers for every student) leads to rise in the median housing values.
dis: Proximity to employment centers has a statistically significant (p-value = 1.67 e-05) impact on the 'medv' i.e. with 1 unit decrease the medv rises by 0.55 units.

The fitted plot shows that the model effectively predicts MEDV except for a few wrong predictions in the high MEDV range, although a linear model is not the best fit. Adjusted for the number of predictors, about **68.78% (Adj R^2 = 0.68)** of the variation in MEDV is explained by the model's predictors. **High F-statistic** (292.2) and **low p-value** (< 2.2 e-16) suggests that the MLR model is statistically significant.

8. Model Diagnostics



Homoscedasticity

Residuals vs. Fitted Plot: Residuals are randomly scattered around zero, supporting homoscedasticity, though there's slight widening at higher fitted values, suggesting mild heteroscedasticity.

Scale-Location Plot: The red line shows slight curvature, indicating mild non-constant variance in residuals.

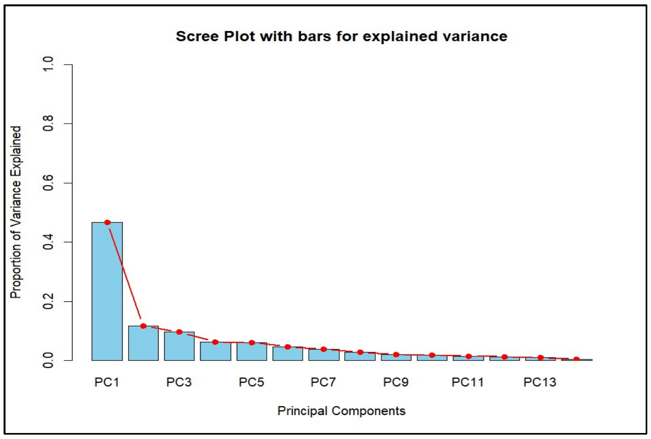
Normality of Residuals

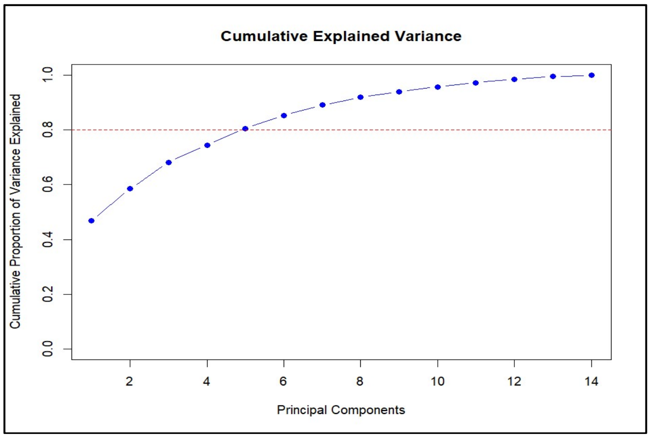
Q-Q Plot: Residuals mostly fall on the diagonal but have significant deviations towards the right tail (Not Normal).

Residuals vs. Leverage Plot: A few points (e.g., 369, 215, 366) have moderate leverage but low Cook's distance, showing they don't strongly influence the model fit.

Advance Analysis

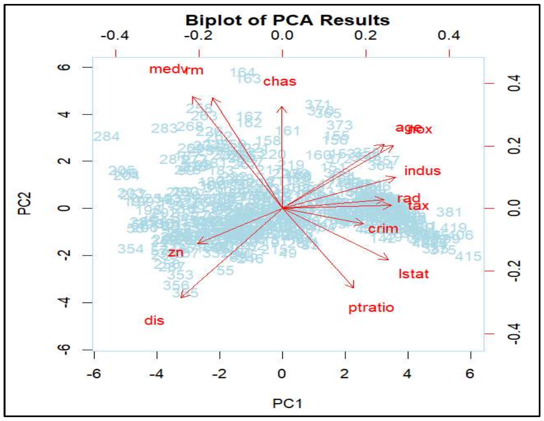
9. Principal Component Analysis





No. of PCs chosen = 5, as they collectively explain **80% of the variation** in the data. While the scree plot has an elbow at 2nd PC but the slope completely flattens after the fifth PC, suggesting that the rest of the PCs have a low eigen values or proportion of variance explained values.

10. PCA Interpretation



PC1: Contrasts **industrial** variables (nox, indus, rad) with **residential** ones (dis, rm), as indicated by opposing arrow directions. PC1 also influenced by black and zn but in opposing direction to industrial variables.

PC2: Highlights housing-related features like chas and medv, distinguishing neighborhoods by **proximity** to the river and **housing prices**.

Correlations: nox, indus, and rad are positively correlated, while dis, zn, rm are negatively correlated with them.

Clusters: The data points appear to form two overlapping clusters along the PC1 potentially reflecting neighborhoods with different residential and industrial characteristics.

11. Conclusion

Univariate analysis reveals that the age distribution is left-skewed, indicating a high concentration of older housing stocks, while few RAD (highway access) values show a relatively high count of tracts. Multivariate analysis highlights older housing is farther from employment centers. Regression identifies rooms and % of lower status population as key predictors of housing value. Regression diagnostics reveal homoscedasticity of residuals which are not normally distributed. PCA indicates 5 PCs explain 80% of variation in data and separates industrial and residential features, emphasizing the importance of housing quality and proximity to urban centers. Low pupil-teacher ratio, and closeness to the Chas River are significant drivers of medv.

Dataset 3: NYC Flights (R library: nycflights13)

Univariate Analysis

1. Data Overview (original # of observations = 3,36,776)

Name	*year*	*month*	*day*	*dep_time*	*sched_dep_time*	*dep_delay*	*arr_time*	*sched_arr_time*
Type	int	int	int	int	int	Num	int	Int
1st obs.	2013	1	1	533	529	4	850	830

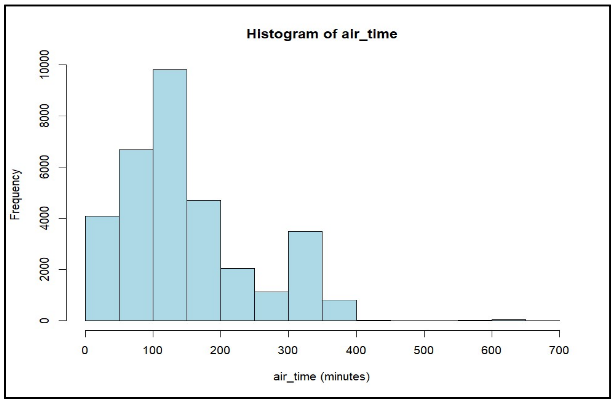
Name	*arr_delay*	*carrier*	*flight*	*tailnum*	*origin*	*dest*	*air_time*	*distance*	*hour*	*minute*	*time_hour*
Type	num	chr	int	int	chr	chr	int	int	int	int	POSIX_ct
1st obs.	20	"UA"	1714	"N24211"	"LGA"	"IAH"	227	1416	5	29	"2013-01-01 05:00:00"

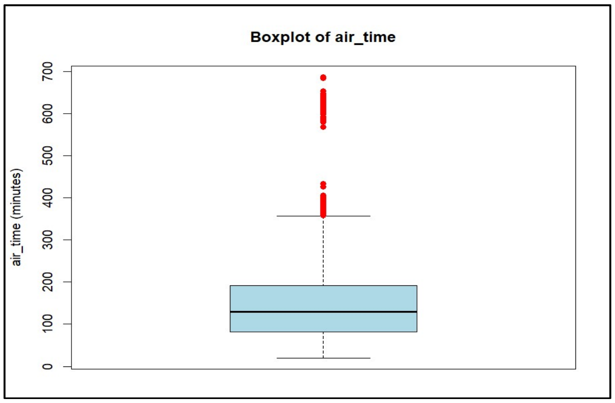
No. of observations after pre-processing = 32735 No. of variables = 19

2. Summary Statistics (*Numerical variable: air_time – Time spent in air in min*)

Statistic	**Mean**	**Median**	**Std. Dev**	**Minimum**	**Maximum**
Value	150.818	129.0	93.595	20.0	686.0

3. Distribution Visualization (air_time)



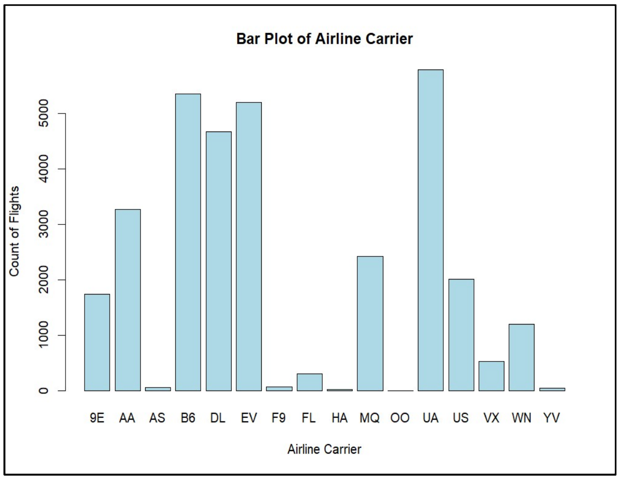


The histogram shows that the distribution of air_time is **right-skewed**, with most flights having air times concentrated between 100 to 200 minutes. A smaller proportion of flights have significantly longer air times, as indicated by the tail stretching towards higher values.
The boxplot confirms the right-skewness, with the **median closer to the lower quartile**. It also reveals potential **outliers beyond 350 minutes**. The interquartile range (IQR = Q3 – Q1) appears to be around 100 minutes, with the whiskers extending to approximately 20 and 380 minutes, capturing most of the data.

4. Categorical Variable Analysis (*carrier*)

Carrier: Two-letter airline code (e.g., "AA" for American Airlines)

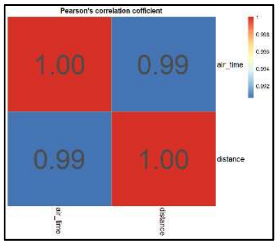
The distribution of carrier values is **non-uniform**, with a few airlines like UA, EV, and B6 dominating the majority of flights, while carriers like AS, HA, and OO have comparatively low counts. This suggests that a few airlines are significantly more frequent and likely preferred by passengers, possibly due to larger networks, better coverage, or operational efficiencies



Multivariate Analysis

5. Correlation Analysis (air_time vs distance)

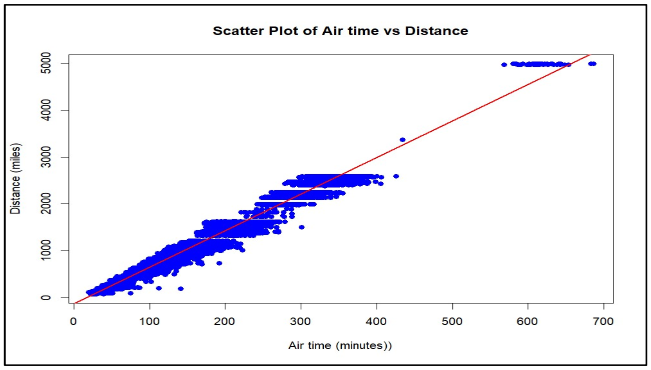
Pearson correlation coefficient between air_timr and distance = **0.99**. It indicates a **very strong positive correlation**. it is obvious to note that as distance between origin and destination increases, time of flight in air increases.



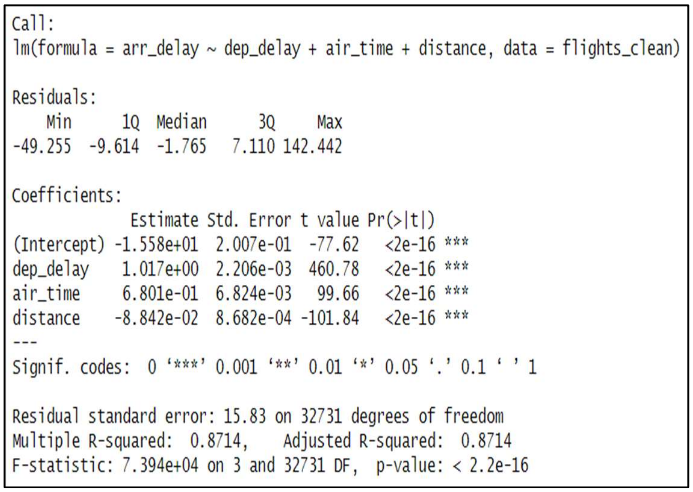
6. Scatter Plot Visualization

The scatter plot reveals a **strong positive relationship** between ‘air_time and ‘distance’ which suggests that the features have a strong positive linear association. Increase in air_time implies more distance between origin and destination.

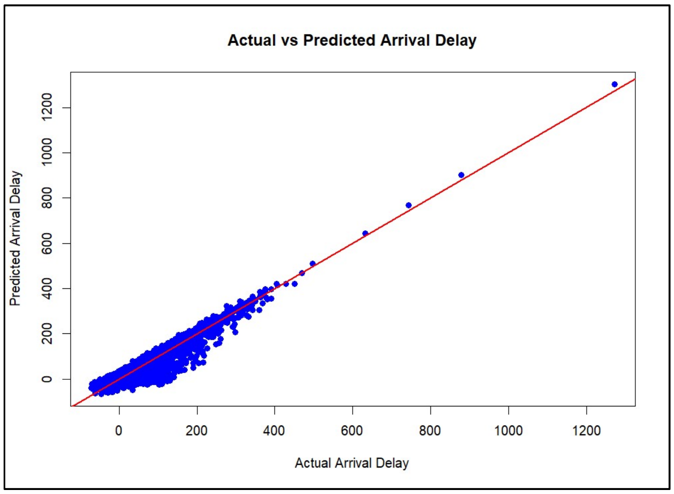
The **upward trend** line aligns with this observation, reinforcing the Pearson correlation value of 0.99.



7. Multiple Regression (*Target: arr_delay (Arrival Delay)*)



```
Call:
lm(formula = arr_delay ~ dep_delay + air_time + distance, data = flights_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.255  -9.614  -1.765   7.110 142.442 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.558e+01  2.007e-01  -77.62   <2e-16 ***
dep_delay    1.017e+00  2.206e-03  460.78   <2e-16 ***
air_time     6.801e-01  6.824e-03   99.66   <2e-16 ***
distance    -8.842e-02  8.682e-04 -101.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.83 on 32731 degrees of freedom
Multiple R-squared:  0.8714,	Adjusted R-squared:  0.8714 
F-statistic: 7.394e+04 on 3 and 32731 DF,  p-value: < 2.2e-16
```

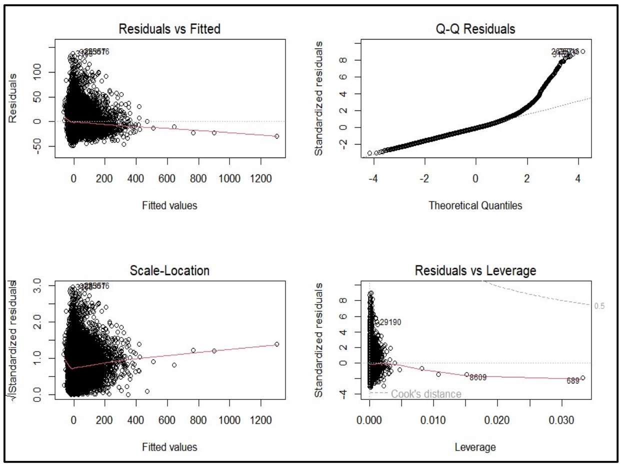


Intercept: The baseline predicted value of dependent variable ‘arr_delay’ when the predictor variables are set to 0 is -15.8 min i.e. arrives 15 min early. The p-value (< 2e-16) suggests it is statistically significant.
dep_delay: For every unit increase in ‘dep_delay’ (delay in departure), ‘arr_delay’ is expected to increase by 1 unit which is an obvious inference making it **most significant predictor** (abs t- value = 460.78).
air_time: For every 1-unit increase in the time of flight in air, the ‘arr_delay’ is expected to increase by 0.68 units. The p-value (< 2e-16) indicates it is highly statistically significant.
distance: The distance between origin and destination appears to have a minor negative relationship with arrival delay, suggesting that longer flights may be more consistent and punctual.

Both air_time and distance estimated could indicate that airlines prioritize careful planning and scheduling for long-haul journeys to maintain reliability.

The fitted plot shows that the model effectively predicts arr_delay even in cases of huge delay values, although few predictions in the low delay range looks to be off. Adjusted for the number of predictors, about **87.14% (Adj R^2 = 0.87)** of the variation in MEDV is explained by the model's predictors. **High F-statistic** (7.39e+04) and **low p-value** (< 2.2 e-16) suggests that the MLR model is statistically significant.

8. Model Diagnostics



Homoscedasticity

Residuals vs. Fitted Plot: Residuals show a cone-shaped pattern with increasing variance at higher fitted values, indicating **heteroscedasticity**.

Scale-Location Plot: The red line curves upward at higher fitted values, further confirming non-constant variance.

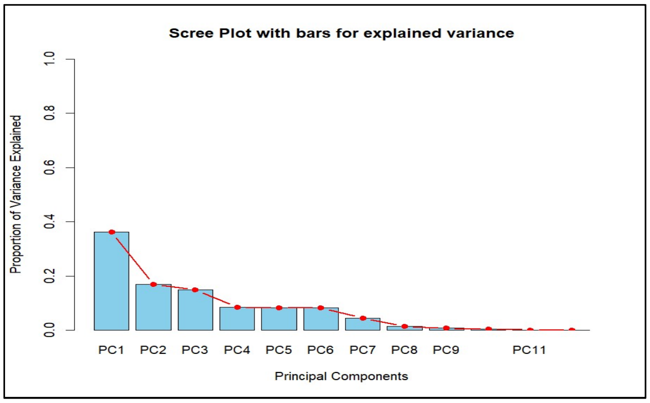
Normality of Residuals

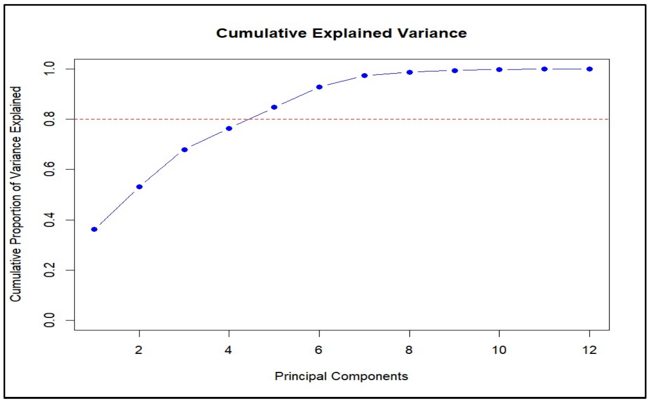
Q-Q Plot: Residuals deviate significantly from the diagonal line, especially at the right tail, indicating non-normality.

Residuals vs. Leverage Plot: Points like 369, 8609 show moderate leverage but low Cook's distance, suggesting limited influence on the model fit.

Advance Analysis

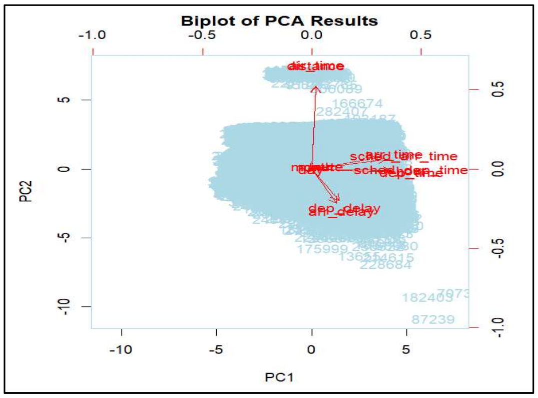
9. Principal Component Analysis





No. of PCs chosen = 4, as they collectively explain about **80% of the variation** in the data. The scree plot has an elbow at PC4 with a flat line until PC6 which completely falls beyond that, suggesting that the rest of the PCs have low eigen values or proportion of variance explained values.

10. PCA Interpretation



PC1: Strongly associated "sched_arr_time," "sched_dep_time," "air_time," as indicated by the direction & length of their vectors.

PC2: Dominated by "**distance**," with its vector pointing almost vertically upward, indicating that it contributes predominantly to PC2 and is relatively independent of PC1.

Two clusters are visible, most data points clustered near the origin and 2nd cluster of points having high 'distance' values.

Variables related to **scheduling and delays** form a cohesive group.

The separation of **travelled distance** from the other variables suggests a distinct dimension of variability while the other variables are more interrelated.C

11. Conclusion

Air time distribution is right-skewed, with most flights between 100-200 minutes, while few carriers (UA, EV, B6) dominate the market. Multivariate analysis reveals a strong linear relationship between air time and distance (Pearson = 0.99). Regression shows departure delays significantly impact arrival delays, with long-haul flights more punctual. Residual analysis implies heteroscedasticity and non-normality indicating that the regression model was unable to capture all patterns. 4 PCs explain 80% variation in data and PCA highlights scheduling variables as key influencers, with distance forming an independent cluster. The analysis revealed that addressing departure delays is essential for improving operational efficiency and punctuality.

Dataset 4: mtcars

Univariate Analysis

1. Data Overview

Name	*mpg*	*cyl*	*disp*	*hp*	*drat*	*wt*	*qsec*	*vs*	*am*	*gear*	*carb*
Type	num	cat	num	num	num	num	num	num	cat	cat	cat
First 3 Obs.	21	6	160	110	3.9	2.62	16.5	0	1	4	4
	21	6	160	110	3.9	2.88	17	0	1	4	4
	22	4	108	93	3.85	2.32	18.6	1	1	4	1

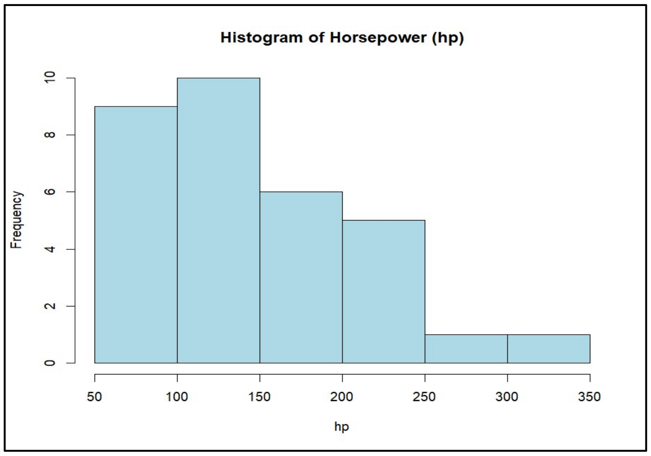
No. of observations = 32

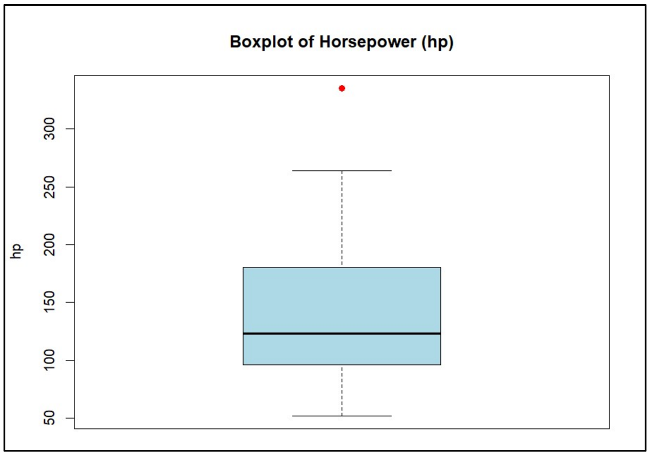
No. of variables = 11

2. Summary Statistics (*Numerical variable: hp (Gross HorsePower)*)

Statistic	**Mean**	**Median**	**Std. Dev**	**Minimum**	**Maximum**
Value	146.69	123.0	68.56	52.0	335.0

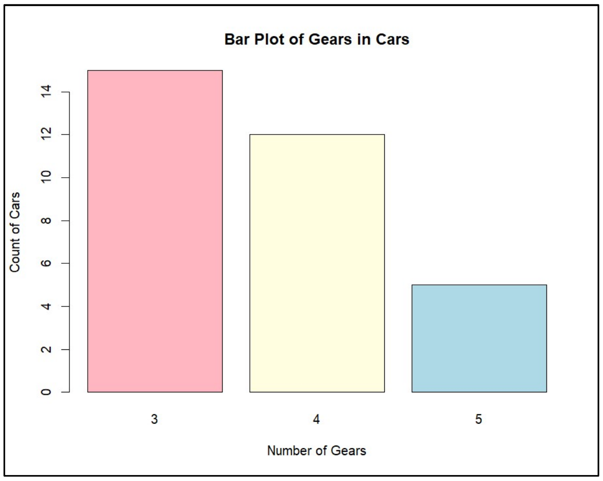
3. Distribution Visualization (*hp*)





The histogram shows that the distribution of hp is **right-skewed**, with most cars having gross horsepower concentrated between 50 to 150 hp. A smaller proportion of cars have significantly high horsepower, as indicated by the tail stretching towards higher values.
The boxplot confirms the right-skewness, with the **median closer to the lower quartile**. It also reveals potential **outliers beyond 260 hp**. The interquartile range (IQR = Q3 – Q1) appears to be around 75 (175 – 100), with the whiskers extending to approximately 50 and 260 units, capturing most of the data.

4. Categorical Variable Analysis (gear)



Gear: The number of forward gears in the transmission.

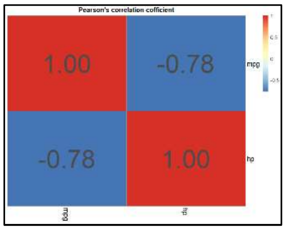
The **majority** of cars in the mtcars dataset have **3 forward gears** (might be reflecting older or simpler vehicle designs), indicating their prevalence in the sample. In contrast, cars with **5 forward gears** (representing high-performance models) are the **least** common, highlighting their rarity.

Cars with 4 gears fall in the middle in terms of frequency, suggesting they strike a balance between simplicity and performance.

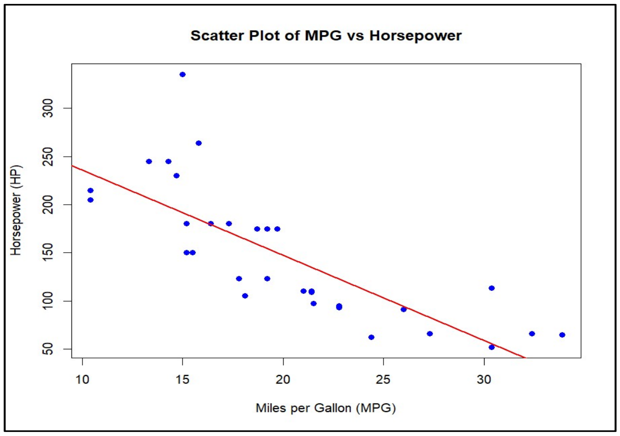
Multivariate Analysis

5. Correlation Analysis (mpg vs hp)

Pearson correlation coefficient between mpg (miles per gallon) and hp (horsepower) = **-0.776**. It indicates a **strong negative correlation**. This indicates that cars with higher horsepower tend to have lower fuel efficiency.



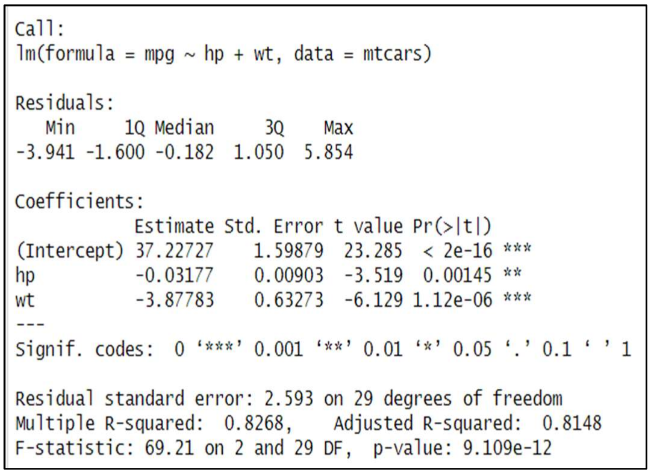
6. Scatter Plot Visualization



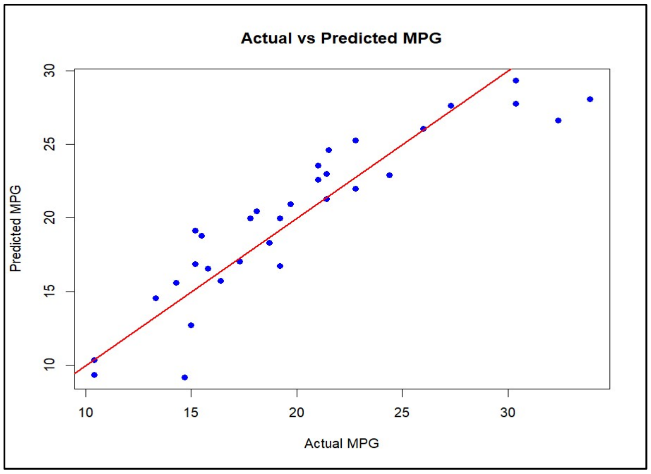
The scatter plot reveals a **strong negative relationship** between hp and mpg which suggests that the features have a negative linear association. Increase in hp leads to decrease in efficiency in mpg.

The **downward trend** line aligns with this observation, reinforcing the Pearson correlation value of -0.78.

7. Multiple Regression (*Target: mpg*)



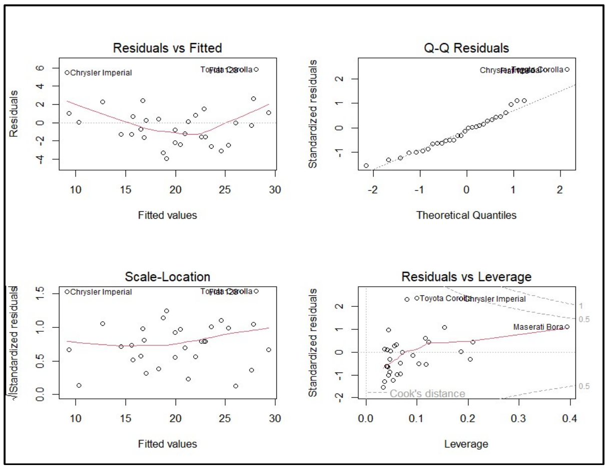
```
Call:
lm(formula = mpg ~ hp + wt, data = mtcars)

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
```



Intercept (37.227): The baseline predicted value of dependent variable ‘mpg’ when the predictor variables are set to 37.23 miles per gallon. The p-value (< 2e-16) suggests it is statistically significant.
hp (-0.0317): For every unit increase in ‘hp’ (horsepower), ‘mpg’ is expected to decrease by 0.03 units consistent with the pearson correlation and scatter plot observations earlier.
Wt (-3.878): For every 1-unit increase in the weight of car (in 1000 lbs), the ‘mpg’ is expected to decrease by 3.88 units. It has the highest coefficient value among the predictors and its p-value (< 2e-16) indicates it is highly statistically significant, making it **most important predictor** variable in this case (absolute t-value = 6.129)

The fitted plot shows that the model effectively predicts ‘mpg’. The model explains **82.68%** of the variation in mpg (Multiple R^2=0.8268), with an adjusted R^2 of 0.8148. The F-statistic (69.21) and p-value (< 0.001) indicate the model is statistically significant overall. The scatter plot of actual vs. predicted mpg shows a **strong linear relationship**, indicating the model fits well. However, some deviations from the line suggest minor prediction errors.

8. Model Diagnostics



Homoscedasticity

Residuals vs. Fitted Plot: The residuals do not show a clear cone-shaped pattern but exhibit slight non-linearity, indicating mild heteroscedasticity.

Scale-Location Plot: The red line remains relatively flat with minor curvature, suggesting near-constant variance overall.

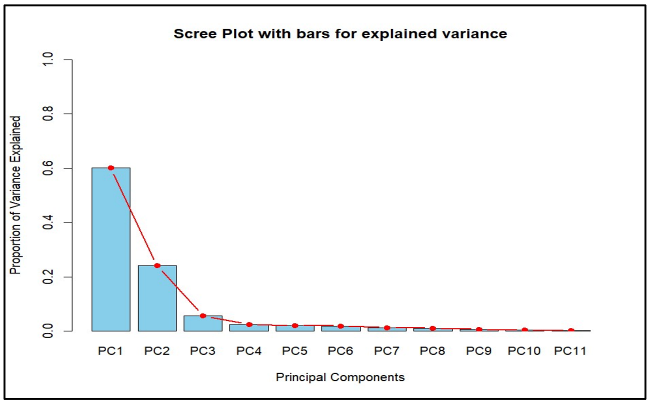
Normality of Residuals

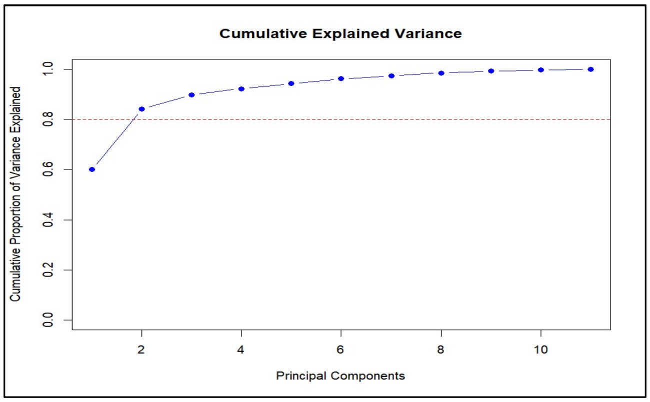
Q-Q Plot: Residuals mostly align with the diagonal, but some deviation is noticeable at the tails, indicating slight non-normality.

Residuals vs. Leverage Plot: Points like Chrysler Imperial and Maserati Bora exhibit moderate leverage, with low Cook's distance, suggesting limited influence on the overall model fit.

Advance Analysis

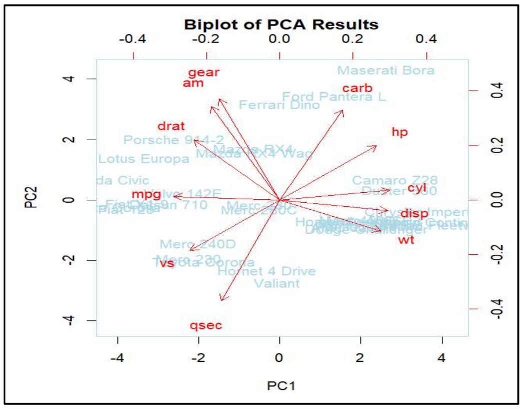
9. Principal Component Analysis

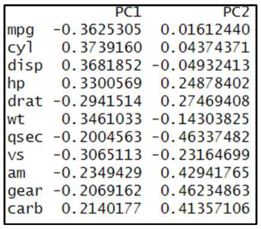




No. of PCs chosen = 3, as they collectively explain about **90% of the variation** in the data. The scree plot has an elbow at PC3 which tails off for the subsequent PCs, suggesting that the rest of the PCs have a low eigen values or proportion of variance explained values.

10. PCA Interpretation





```
              PC1         PC2
mpg  -0.3625305  0.01612440
cyl   0.3739160  0.04374371
disp  0.3681852 -0.04932413
hp    0.3300569  0.24878402
drat -0.2941514  0.27469408
wt    0.3461033 -0.14303825
qsec -0.2004563 -0.46337482
vs   -0.3065113 -0.23164699
am   -0.2349429  0.42941765
gear -0.2069162  0.46234863
carb  0.2140177  0.41357106
```

PC1: Primarily separates variables associated with **engine power and car weight** from others. wt, hp, and disp, cyl strongly influence PC1. They are directly opposed by mpg & vs.

PC2: gear, qsec and am dominate PC2. It contrasts **transmission-related** features (gear, am) and **acceleration** (qsec).

The biplot reveals a somewhat even distribution of car models across the features. PC1 differentiates heavy, high-power cars from light, efficient models, while PC2 captures distinctions in acceleration and transmission.

11. Conclusion

Univariate analysis shows horsepower (hp) is right-skewed having few potential outliers. 3-gear models are most prevalent while high-performance models are comparatively rare. Multivariate analysis reveals a strong negative correlation between mpg (fuel efficiency) and hp. Regression identifies weight (wt) as the most significant predictor of mpg, with heavier cars being less efficient. Model diagnostics reveal mild heteroscedasticity and non-normality indicating that the model needs further improvements. PCA indicates that 3 PCs explain 90% of the variation in the data. The biplot distinguish high-power, heavy vehicles from lightweight, fuel-efficient ones. These insights highlight the importance of balancing performance and efficiency in automotive design.