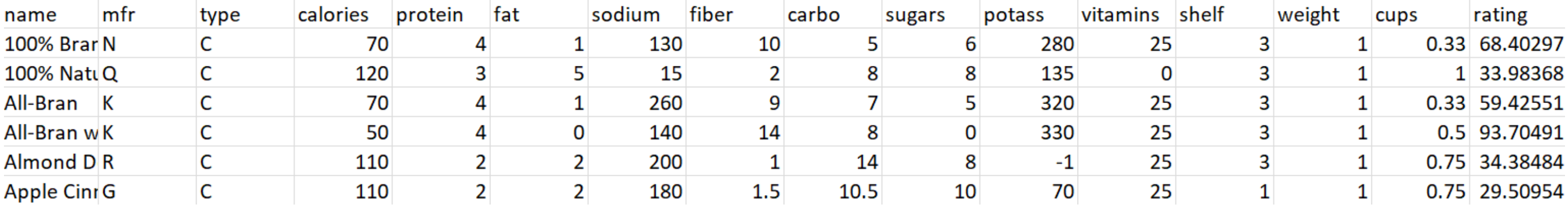
Cereals Dataset

1.Overview of Data Set



name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100% Brar	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100% Natu	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran w	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond D	R	C	110	2	2	200	1	14	8	-1	25	3	1	0.75	34.38484
Apple Cinr	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954

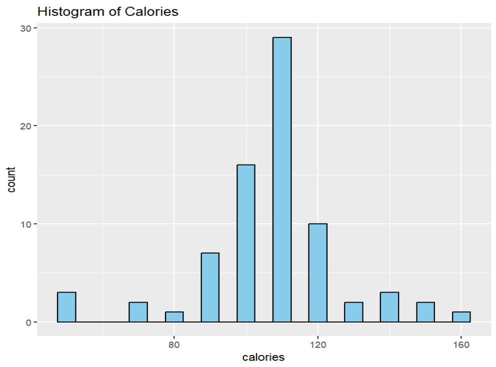
Number of observations: **76** Number of Variables : **16**

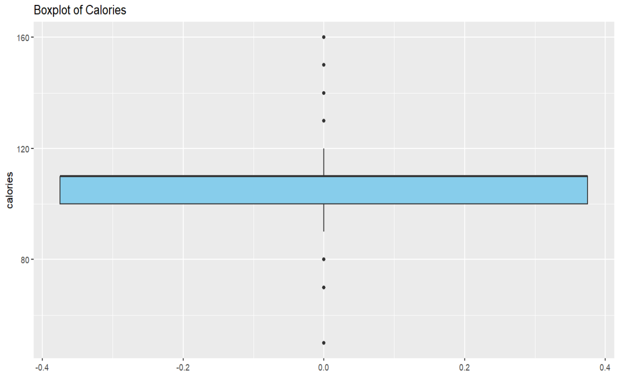
2. Summary Statistics for calories

Mean:107 **Median**:110 **Standard Deviation**:19.60 **Minimum**:50 **Maximum**:160

3. Distribution Visualization for Calories

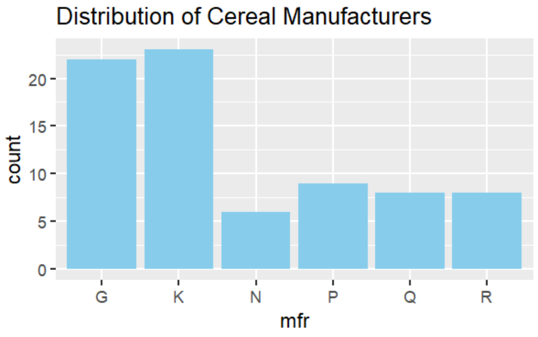
The histogram displays a roughly normal distribution of calories with a peak around 80-90 calories, showing some right-skewed tendency. The boxplot on the right reveals the spread of the data, with a median line clearly visible and what appears to be several outliers above the upper whisker, indicating some unusually high calorie values in the dataset.





4. Categorical Variable Analysis

The bar chart shows the distribution of cereal manufacturers, with letters G and K being the most prominent, each having around 20-22 manufacturers. There is a significant drop to the other categories (N, P, Q, and R) which all have fewer than 10 manufacturers each. This distribution suggests a market dominated by two major groups of manufacturers, while the remaining categories represent smaller players in the cereal manufacturing industry.

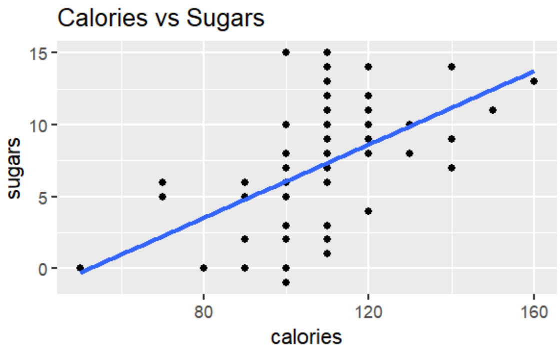


5. Correlation Analysis

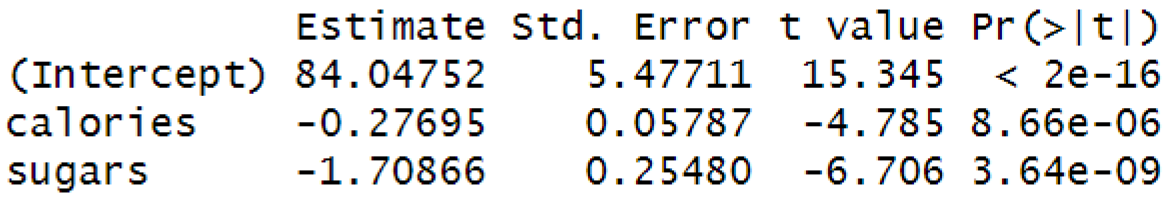
Pearson correlation coefficient between calories and sugar: **0.5615**
Summary: The Pearson correlation suggests a **positive relationship** between calories and total sugar.

6. Scatterplot Visualization:

The scatter plot reveals a positive correlation between calories (80-160) and sugar content (0-15 units), with data points clustering around a linear trend line. Despite the general upward trend, there is notable variability in sugar content, especially at higher calorie levels.

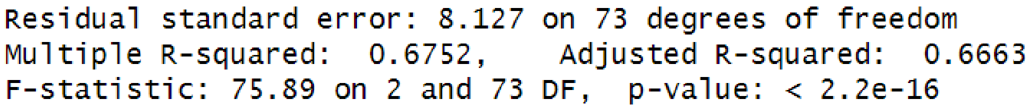


7.Multiple Linear Regression.



```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.04752    5.47711  15.345  < 2e-16
calories    -0.27695    0.05787  -4.785 8.66e-06
sugars      -1.70866    0.25480  -6.706 3.64e-09
```

```
model <- lm(rating ~ calories + sugars, data = cereals)
```



```
Residual standard error: 8.127 on 73 degrees of freedom
Multiple R-squared:  0.6752,    Adjusted R-squared:  0.6663
F-statistic: 75.89 on 2 and 73 DF,  p-value: < 2.2e-16
```

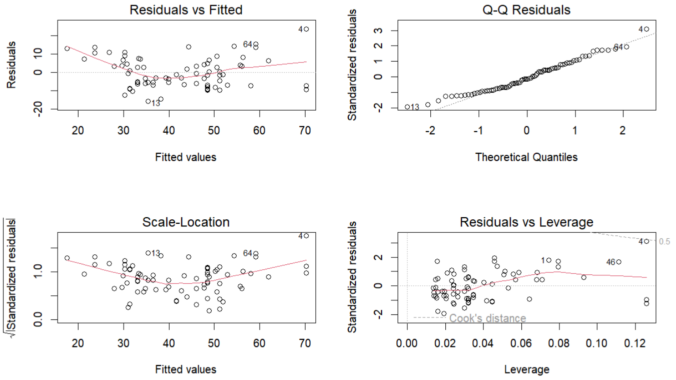
Key Interpretations:

The dependent variable is **rating**, predicted using **calories** and **sugar**. Key insights:

- Intercept = 84.05: Base value when all predictors are zero, highly significant (p<2e-16)
- Calories = -0.277: Each calorie unit decreases response by 0.277, very significant (p<8.66e-06)
- Sugars = -1.709: Each sugar unit decreases response by 1.709, extremely significant (p<3.64e-09)
- RSE = 8.127: Average prediction error is about 8.13 units with 73 df
- R² = 0.6663: Model explains 66.63% of data variation
- F = 75.89: Overall model is highly significant (p<2.2e-16)

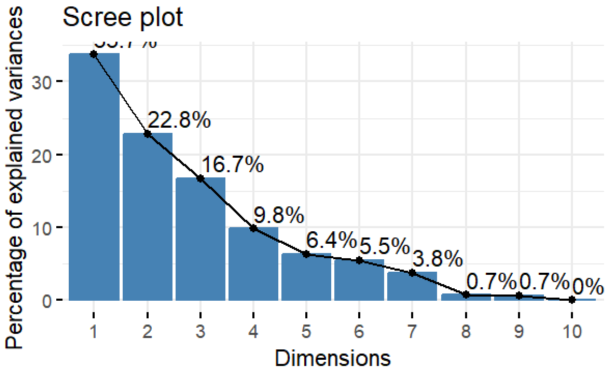
8. Model Diagnostics.

The diagnostic plots show a regression analysis with four key visualizations: residuals vs. fitted values, Q-Q plot of residuals scale-location plot, and residuals vs. leverage. The residuals vs. fitted plot and scale-location plot suggest relatively constant variance with a slight pattern, while the Q-Q plot indicates the residuals follow a roughly normal distribution with some deviation at the tails. There are a few potential outliers and influential points identified, particularly observations 4, 13, and 64, as shown in multiple plots.



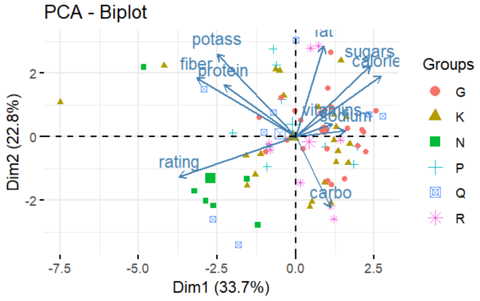
9. Principal Component Analysis

The scree plot shows a steep decline in explained variance across dimensions, with the first dimension accounting for 33.7% and the second for 22.8% of the total variance. The sharp elbow in the plot after the second or third dimension suggests that retaining 2-3 dimensions would capture the most significant patterns in the data while maintaining parsimony.



10. PCA Interpretation

This PCA biplot reveals nutritional relationships across different food groups (G, K, N, P, Q, R), with the first two dimensions explaining 56.5% of the total variance (Dim1: 33.7%, Dim2: 22.8%). The plot shows strong positive correlations between calories, sugars, and fat content (clustered on the right), while fiber, protein, and potassium form another correlated group (upper left), suggesting distinct nutritional profiles among the food items.



MtCars Dataset

1. Overview of Data Set

model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1

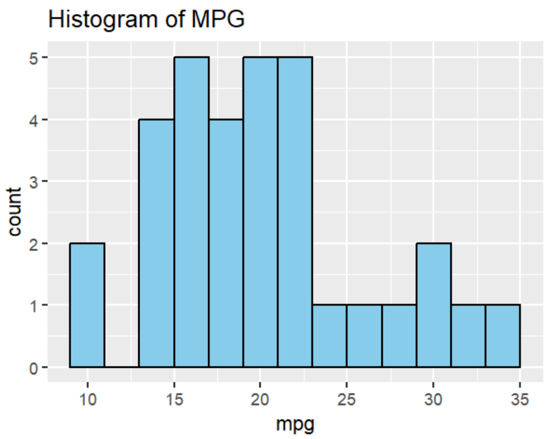
Number of observations: **32** Number of Variables : **11**

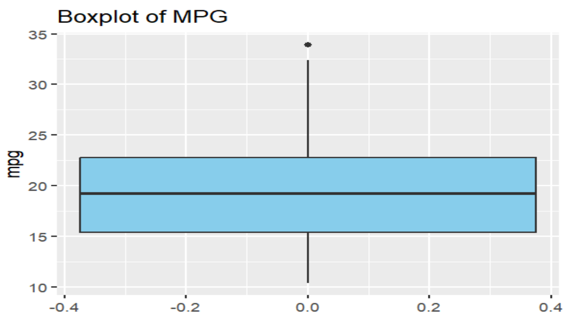
2. Summary Statistics for MPG

Mean:20.09 **Median**:19.20 **Standard Deviation**:6.03 **Minimum**:10.40 **Maximum**:33.90

3. Distribution Visualization for Calories

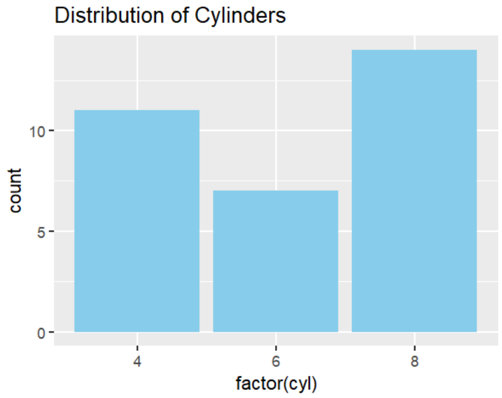
The histogram of MPG shows a roughly normal distribution with a slight right skew, centered around 20 miles per gallon, with most vehicles falling between 15-25 MPG. The boxplot reveals the median MPG is approximately 20, with the interquartile range spanning from about 15 to 22.5 MPG, and there appears to be one notable outlier above 30 MPG. These visualizations suggest that while most cars in the dataset have relatively similar fuel efficiency, there are a few highly efficient vehicles that stand out from the typical range





4. ## Categorical Variable Analysis

The bar graph shows the distribution of cylinders across different vehicle categories. The data indicates that 8-cylinder vehicles are the most common with approximately 14 units, followed by 4-cylinder vehicles with about 11 units, while 6-cylinder vehicles are the least common with roughly 7 units. This distribution suggests a polarization in the market between vehicles with high cylinder counts (8) and those with lower counts (4), with mid-range options being less prevalent.



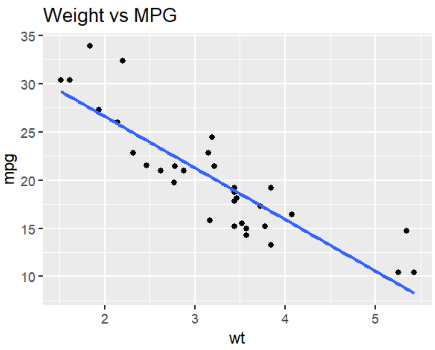
5. ## Correlation Analysis

Pearson correlation coefficient between mpg and weight: **-0.8676594**

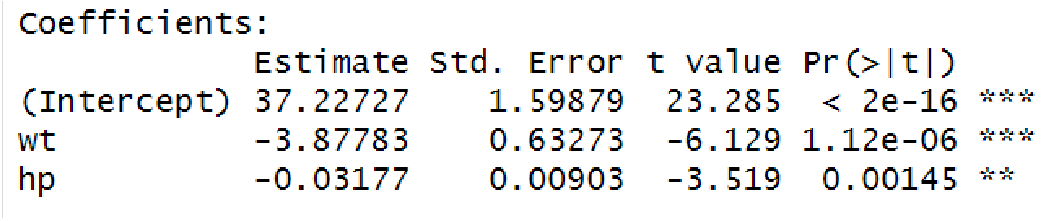
Summary: The **strong negative** correlation coefficient of -0.87 indicates that as a vehicle's weight increases, its fuel efficiency (mpg) tends to significantly decrease.

6. ## Scatterplot Visualization:

The scatter plot clearly demonstrates a strong negative linear relationship between vehicle weight and fuel efficiency (MPG), with the blue regression line showing a consistent downward trend. The data points are reasonably well distributed around the regression line, suggesting that the linear model is appropriate for describing this relationship. While there are a few outliers, particularly in the higher MPG range around 30-35, the overall pattern strongly supports the correlation coefficient of -0.87 and indicates that heavier vehicles consistently achieve lower miles per gallon.

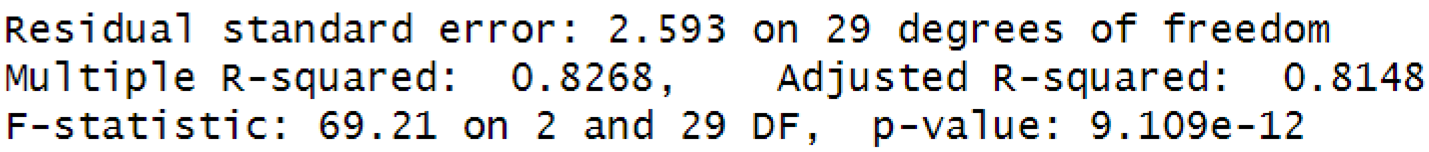


7. ## Multiple Linear Regression



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

```
model <- lm(mpg ~ wt + hp, data = mtcars)
```



```
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
```

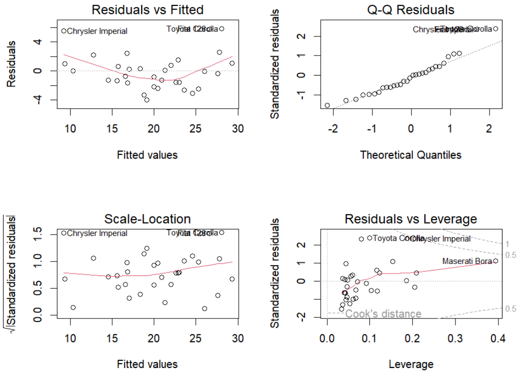
Key Interpretations:

The dependent variable is **mpg**, predicted using **weight** and **horsepower**. Key insights:

- Model explains 81.48% of variance in mpg (R² = 0.8148)
- Weight has strong negative impact on mpg (coefficient = -3.87783, p < 1.12e-06)
- Horsepower has smaller negative impact on mpg (coefficient = -0.03177, p = 0.00145)
- Overall model is highly significant (F-stat = 69.21, p = 9.109e-12)

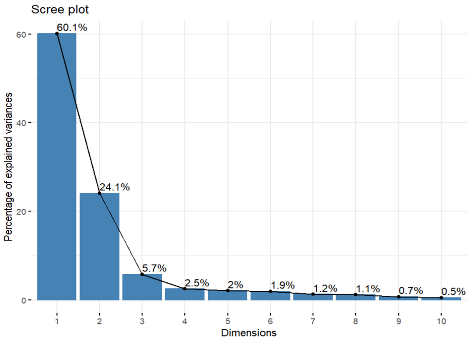
9. Model Diagnostics.

The residual plots reveal potential concerns with the model's assumptions. The Residuals vs Fitted plot shows a curved pattern (U-shape), suggesting a possible non-linear relationship that isn't captured by the current model. The Q-Q plot indicates reasonably normal distribution of residuals with some deviation at the extremes, while the Residuals vs Leverage plot identifies three influential points (Chrysler Imperial, Toyota Corolla, and Maserati Bora) that may be affecting the model's performance.



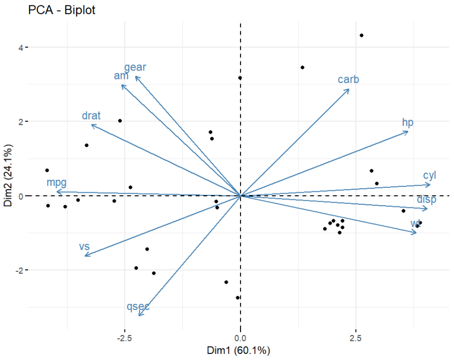
9. Principal Component Analysis

This scree plot shows the percentage of explained variance across 10 dimensions, with a sharp decline after the first dimension. The first dimension accounts for 60.1% of the variance, followed by a significant drop to 24.1% for the second dimension, while subsequent dimensions contribute minimally (less than 6% each). This pattern suggests that a two-dimensional solution might be sufficient to capture most of the meaningful variation in the data, as these two dimensions together explain approximately 84.2% of the total variance.



10. PCA Interpretation

This PCA biplot visualizes the relationships between various car characteristics, with the first two principal components explaining 84.2% of the total variance (PC1: 60.1%, PC2: 24.1%). The plot reveals strong positive correlations between horsepower (hp), cylinder count (cyl), displacement (disp), and carburetor (carb) on the right side, while fuel efficiency (mpg) shows negative correlations with these performance-related variables, suggesting a clear trade-off between power and fuel economy. The variables gear, am (transmission type), and drat (rear axle ratio) are grouped together on the left side, indicating their interrelated nature in the mechanical aspects of the vehicles.



AQI Dataset

1. Overview of Data Set

Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
Russian Federation	Praskoveya	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
Brazil	Presidente Dutra	41	Good	1	Good	5	Good	1	Good	41	Good
Italy	Priolo Gargallo	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good
France	Punaauia	22	Good	0	Good	22	Good	0	Good	6	Good
United States of America	Punta Gorda	54	Moderate	1	Good	14	Good	11	Good	54	Moderate

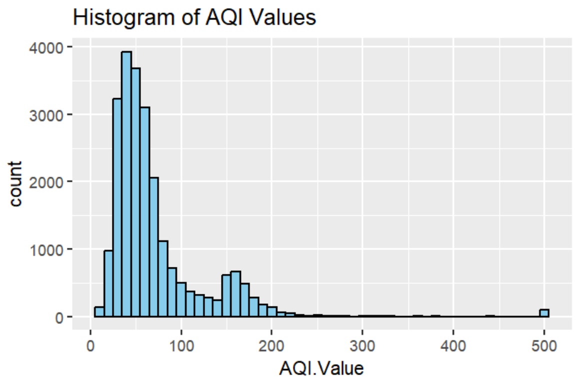
Number of observations: **23463**   Number of Variables : **12**

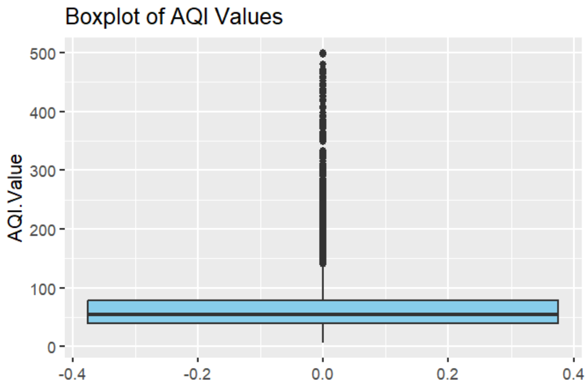
2. Summary Statistics for AQI Value

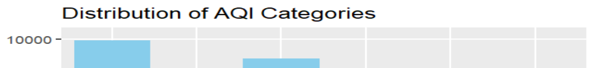
Mean:72.01 **Median**:55.00 **Standard Deviation**:56.05 **Minimum**:6 **Maximum**:500.00

3. Distribution Visualization for AQI Value

The histogram shows a right-skewed distribution of AQI (Air Quality Index) values, with most measurements concentrated between 0-100 and a peak around 25-50. The boxplot reveals the presence of numerous outliers extending up to approximately 500, while the bulk of the data (interquartile range) lies between approximately 25-75. This distribution pattern suggests that while air quality is generally moderate to good, there are occasional episodes of very poor air quality represented by the outliers.







4. Categorical Variable Analysis

The bar chart shows the distribution of Air Quality Index (AQI) categories, with "Good" and "Moderate" being the most frequent classifications, accounting for approximately 10,000 and 9,000 counts respectively. The less desirable categories like "Unhealthy," "Unhealthy for Sensitive Groups," and "Very Unhealthy" occur less frequently, suggesting that air quality conditions are generally favorable in the measured area.

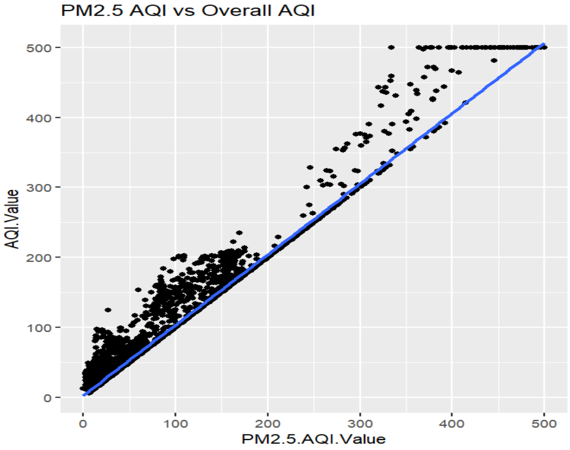
5. Correlation Analysis

Pearson correlation coefficient between aqi value and pm 2.5 value: **0.9843**

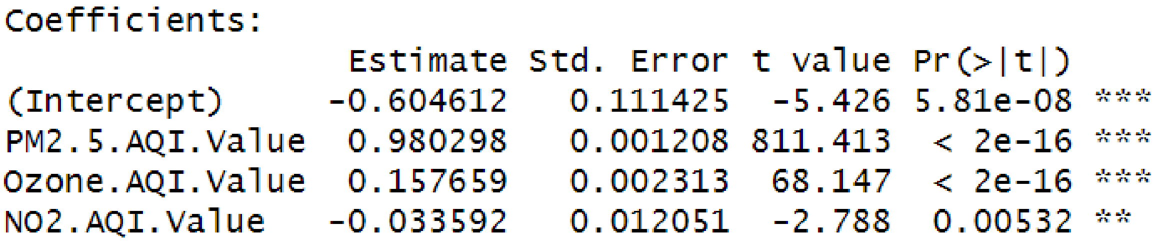
Summary: The **strong positive** correlation coefficient of -0.98 indicates that as pm 2.5 value increases, the aqi value increases as well.

6. Scatterplot Visualization:

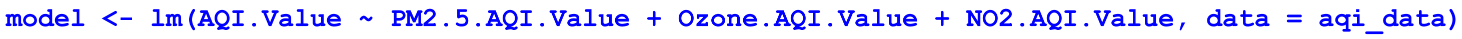
The scatter plot shows the relationship between PM2.5 AQI values and overall AQI values, with a strong positive correlation indicated by the blue trend line. The data points generally cluster around this line at lower AQI values (0-200), but show more dispersion and deviation above the trend line at higher AQI values (300-500), suggesting that PM2.5 becomes a dominant but not exclusive factor in determining overall air quality during severe pollution events.



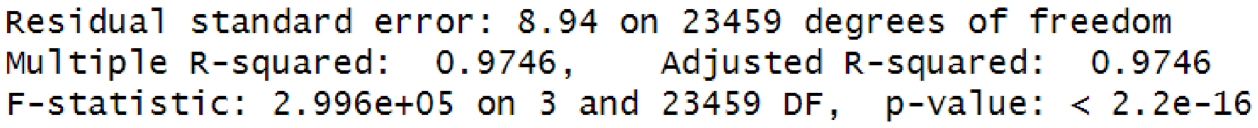
7. Multiple Linear Regression



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -0.604612   0.111425  -5.426 5.81e-08 ***
PM2.5.AQI.Value  0.980298   0.001208 811.413  < 2e-16 ***
Ozone.AQI.Value  0.157659   0.002313  68.147  < 2e-16 ***
NO2.AQI.Value   -0.033592   0.012051  -2.788  0.00532 ** 
```



```
model <- lm(AQI.Value ~ PM2.5.AQI.Value + Ozone.AQI.Value + NO2.AQI.Value, data = aqi_data)
```



```
Residual standard error: 8.94 on 23459 degrees of freedom
Multiple R-squared:  0.9746,    Adjusted R-squared:  0.9746 
F-statistic: 2.996e+05 on 3 and 23459 DF,  p-value: < 2.2e-16
```

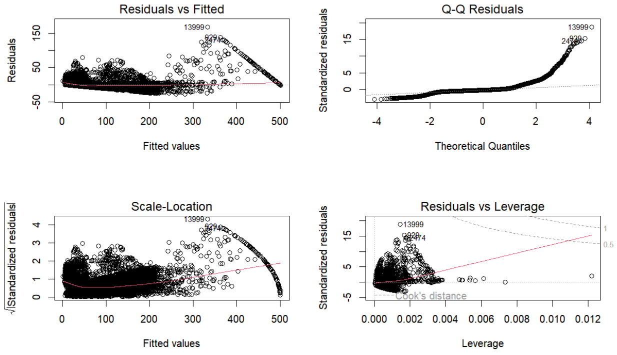
Key Interpretations:

The dependent variable is **aqi value**, predicted using **pm2.5, ozone** and **N02**. Key insights:

- Model explains 97.46% of AQI variance (Adjusted R² = 0.9746)
- PM2.5 shows strongest effect (coef = 0.98), followed by Ozone (0.16)
- NO2 has slight negative impact (coef = -0.03)
- High F-value (2.996e+05) and low p-value confirm model reliability.

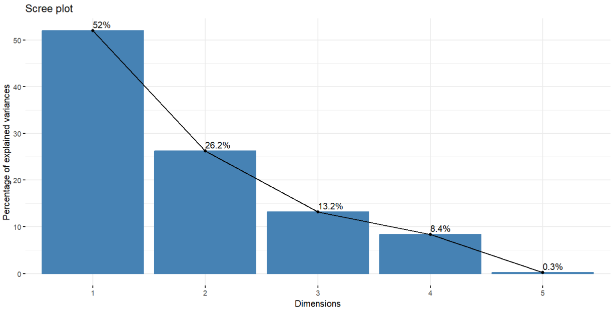
10. Model Diagnostics.

The diagnostic plots reveal potential issues with the model's assumptions. The Residuals vs Fitted plot shows a non-linear pattern and heteroscedasticity (uneven spread of residuals), while the Q-Q plot indicates deviation from normality at the tails. The high leverage points visible in the Residuals vs Leverage plot suggest the presence of influential observations that could be affecting the model's performance.



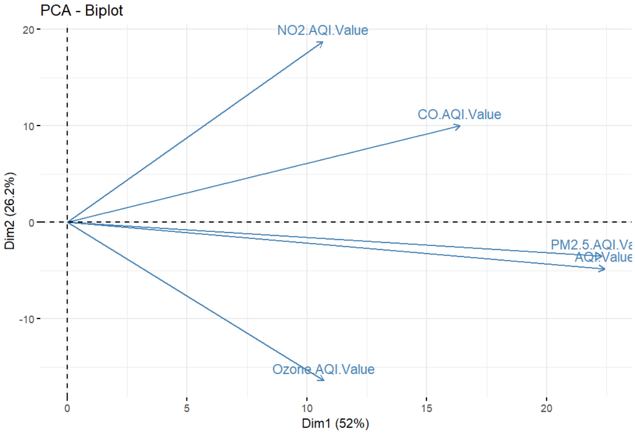
9. Principal Component Analysis

The scree plot shows a steep decline in explained variance across five dimensions. The first dimension accounts for the majority of variance at 52%, followed by a significant drop to 26.2% for the second dimension. The remaining dimensions (3-5) contribute minimally, suggesting that a two-dimensional solution would capture approximately 78.2% of the total variance and would be most appropriate for data analysis.



10. PCA Interpretation

The PCA biplot reveals important relationships between air quality parameters across two principal components that explain 78.2% of total variance (Dim1: 52%, Dim2: 26.2%). NO2 and CO AQI values show positive correlations as indicated by their vectors pointing in similar directions in the upper quadrants. Notably, Ozone AQI values demonstrate an inverse relationship with PM2.5 and AQI values, as shown by their vectors pointing in opposite directions, suggesting these pollutants tend to have opposing patterns in the atmosphere.



California Housing Dataset

1. Overview of Data Set

longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	ocean_proximity	median_house_value
-122.5	37.79	52	8	1	13	1	15.0001	NEAR BAY	500001
-117.79	35.21	4	2	2	6	2	2.375	INLAND	137500
-116.95	33.86	1	6	2	8	2	1.625	INLAND	55000
-117.76	35.22	4	18	3	8	6	1.625	INLAND	275000
-114.62	33.62	26	18	3	5	3	0.536	INLAND	275000
-117.27	34.17	16	30	3	49	8	4.625	INLAND	250000

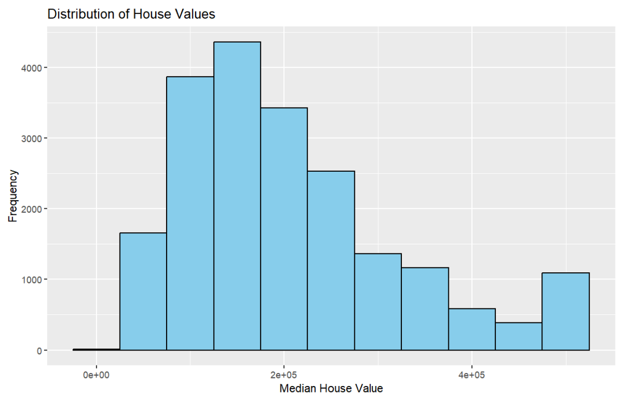
Number of observations: 20433 Number of Variables : **10**

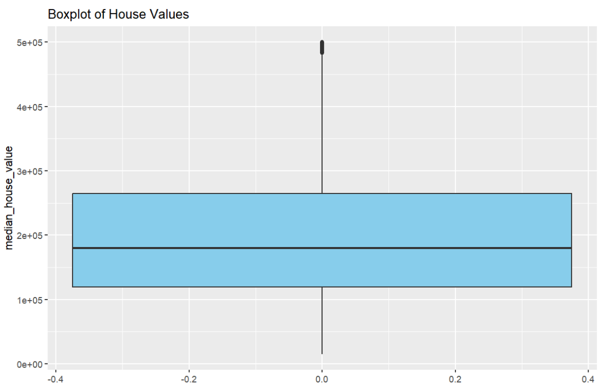
2. Summary Statistics for Median House Value

Mean:206864 **Median**:179700 **Standard Deviation**:115435.7 **Minimum**:14999 **Maximum**:500001

3. Distribution Visualization for AQI Value

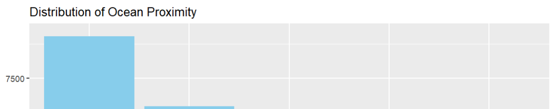
The histogram on the left shows a distribution of house values that appears to be right-skewed, with the peak frequency occurring around the median value range. The boxplot on the right indicates the presence of some outliers, particularly on the higher end of the distribution, as shown by points extending beyond the whiskers. The data suggests that while most house values are clustered around a central range, there are some properties with significantly higher values that pull the distribution to the right.





4. Categorical Variable Analysis

The highest frequency is observed for locations



"<1H OCEAN" with around 8,000 counts, followed by "INLAND" areas with approximately 6,500 counts. The categories "NEAR BAY" and "NEAR OCEAN" have similar lower frequencies of about 2,500 counts each, while "ISLAND" shows minimal representation.

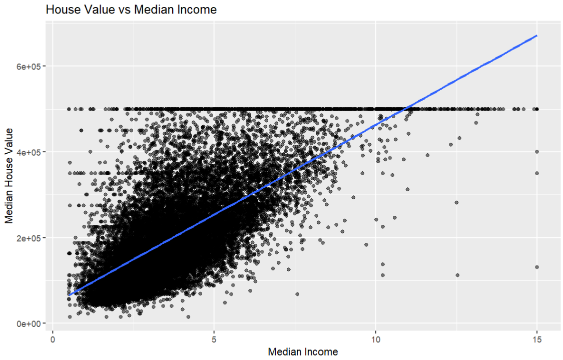
8. Correlation Analysis

Pearson correlation coefficient between medium house value and median income: **0.6884**

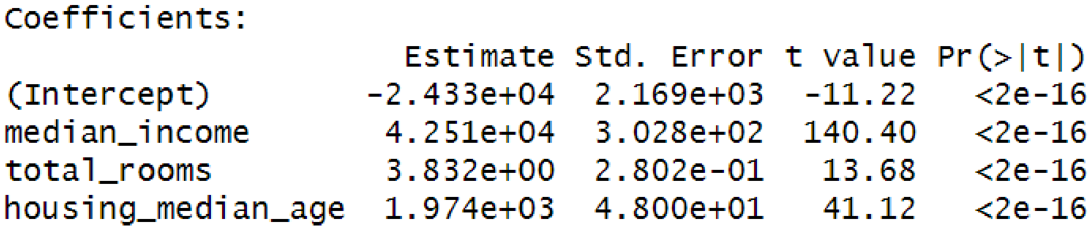
Summary: The **moderately strong positive** correlation coefficient of 0.6884 indicates that as median income increases, there tends to be a corresponding increase in medium house values, though the relationship isn't perfectly linear.

9. Scatterplot Visualization:

The scatter plot shows a clear positive correlation between median income and house values, as indicated by the upward-sloping blue trend line. The data points form a dense cloud with more concentration in the lower to middle ranges of both income and house values, suggesting this represents a typical distribution of housing markets. There is notable dispersion around the trend line, particularly in the middle-income range, indicating that factor beyond income also influence house values, while some outliers exist at higher income and house value levels.

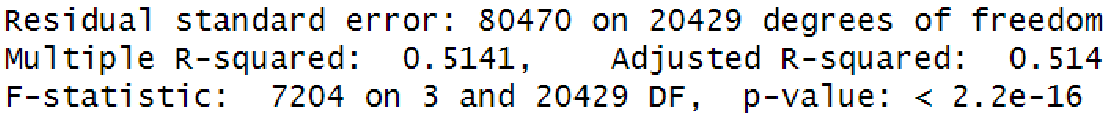


10. Multiple Linear Regression



```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.433e+04  2.169e+03  -11.22   <2e-16
median_income       4.251e+04  3.028e+02  140.40   <2e-16
total_rooms         3.832e+00  2.802e-01   13.68   <2e-16
housing_median_age  1.974e+03  4.800e+01   41.12   <2e-16
```

```
model <- lm(median_house_value ~ median_income + total_rooms + housing_median_age,
data = housing)
```



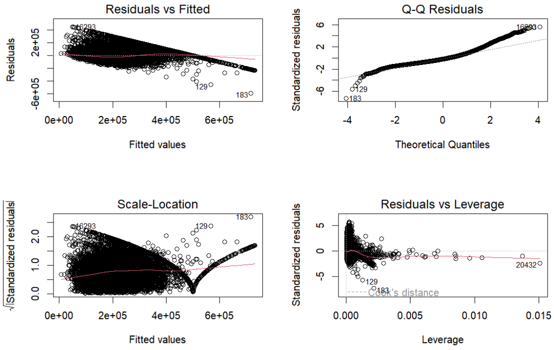
```
Residual standard error: 80470 on 20429 degrees of freedom
Multiple R-squared:  0.5141,    Adjusted R-squared:  0.514
F-statistic:  7204 on 3 and 20429 DF,  p-value: < 2.2e-16
```

The dependent variable is **aqi value**, predicted using **pm2.5, ozone** and **N02**. Key insights:

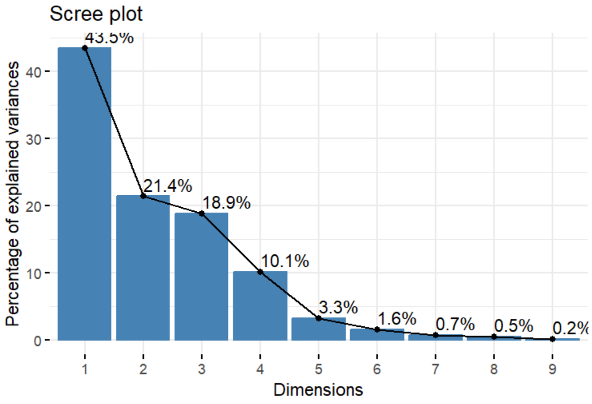
- The model explains 51.4% of the variance in median house values (R-squared = 0.5141), indicating a moderate fit
- All predictors are highly statistically significant (p < 2e-16), with median income having the strongest effect (t = 140.40)
- For every unit increase in median income, house value increases by $42,510, holding other variables constant
- The F-statistic of 7204 with a very low p-value (< 2.2e-16) indicates that the model as a whole is statistically significant

8. Model Diagnostics.

The diagnostic plots reveal potential issues with the model's assumptions. The Residuals vs Fitted plot shows a slight non-linear pattern and heteroscedasticity (uneven spread of residuals), suggesting the model may not fully capture the relationship in the data. The Q-Q plot indicates deviation from normality at the tails, while the Residuals vs Leverage plot highlights several influential points (notably observations 129, 183, and 20432) that could be affecting the model's performance.



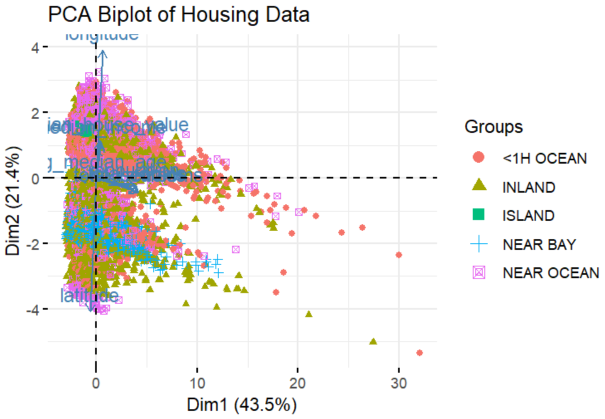
9. Principal Component Analysis



The scree plot shows the percentage of explained variance across 9 dimensions, with a clear elbow pattern. The first dimension accounts for the highest variance at 43.5%, followed by substantial drops to 21.4% and 18.9% for dimensions 2 and 3 respectively. After the third dimension, there is a sharp decline in explained variance (below 10%), suggesting that a three-dimensional solution might be optimal for this dataset as it captures approximately 83.8% of the total variance.

10. PCA Interpretation

This PCA biplot visualizes housing data across five different location categories (INLAND, NEAR BAY, NEAR OCEAN, <1H OCEAN, and ISLAND), with the first two principal components explaining 43.5% and 21.4% of the variance respectively. The plot reveals clear geographic patterns, with longitude and latitude vectors nearly perpendicular to each other, and median house value showing some correlation with geographical coordinates. Most observations cluster around the center, but INLAND properties show the greatest dispersion, particularly along Dimension 1, while coastal properties (NEAR BAY, NEAR OCEAN, and <1H OCEAN) tend to form more concentrated clusters.


PCA Biplot of Housing Data
longitude
ian_house_value
edian_income
g_median_age
latitude
Dim2 (21.4%)
Dim1 (43.5%)
Groups
<1H OCEAN
INLAND
ISLAND
NEAR BAY
NEAR OCEAN