Table of Contents

Abalone Dataset:

Univariate Analysis:

1. Data Overview

The Abalone dataset contains 4177 observations and 9 variables related to the physical measurements of abalones. The primary goal is to predict the age of an abalone based on these measurements, which serve as easier alternatives to the traditional method of cutting the shell and counting the rings. Summary Statistics

```
'data.frame':   4177 obs. of  9 variables:
$ Sex           : chr  "M" "M" "F" "M" ...
$ Length        : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...
$ Diameter      : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...
$ Height        : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...
$ Whole.weight  : num  0.514 0.226 0.677 0.516 0.205 ...
$ Shucked.weight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
$ Viscera.weight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
$ Shell.weight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
$ Rings         : int  15 7 9 10 7 8 20 16 9 19 ...
```

Figure 1: Structure of the Data

2. Summary Statistics:

Summary statistics (Table 1) provide an overview of the central tendency and spread of each variable.
The mean and median are close, suggesting a roughly symmetric distribution. The standard deviation of 0.099 indicates moderate variability around the mean. The data ranges from 0.055 to 0.65

3. Distribution Visualization

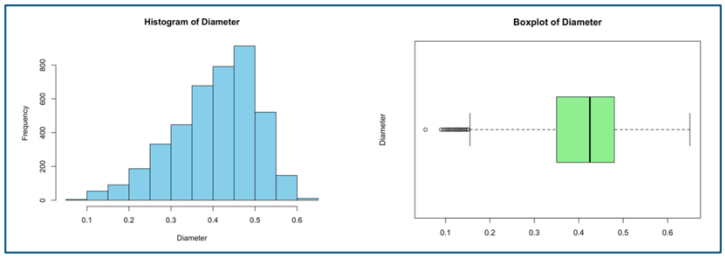

Figure 2: Histogram and Boxplot

The histogram of Diameter exhibits a unimodal, right-skewed distribution, with most observations concentrated between 0.4 and 0.5. The boxplot confirms the skewness and reveals potential outliers on the lower end of the range, below approximately 0.2. The central tendency is evident around the interquartile range (0.4 to 0.5), with a median close to 0.45. These outliers suggest variability among smaller diameter measurements.

4. Categorical Variable(Sex) Analysis:

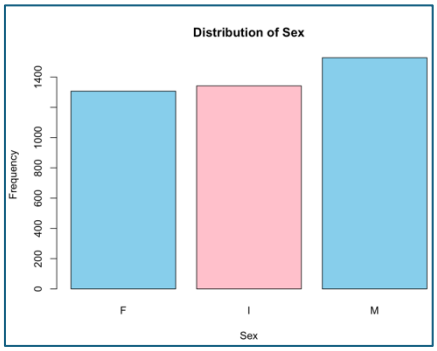

Figure 3: Distribution of Sex

Categories F and M have a similar distribution, with a small difference in frequency. Category I(Infant) is relatively less frequent compared to F(Female) and M(Male). The distribution is not perfectly balanced but shows that all three categories are relatively comparable in frequency.

Multivariate Analysis

5. Correlation Analysis:

Choosen variables for the correlation analysis are Height and Viscera.Weight and the pearson correlation coefficient is 0.7983193 which indicates that the two variables have strong positive correlation

Correlation of each numerical variable w.r.t the other is shown in the Table 2

6. Scatterplot Visualisation:

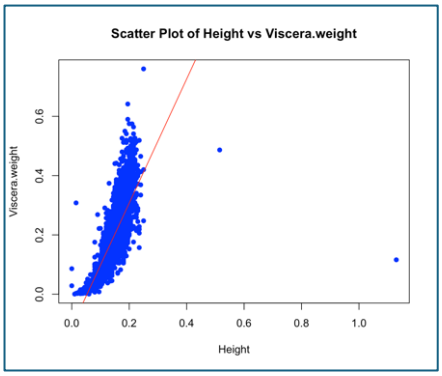

Figure 4: Scatter Plot

The scatter plot of Height versus Viscera.Weight shows a clustered distribution with some outlying points. While there is a general upward trend, the relationship is not strictly linear, as evidenced by the uneven spread of points and deviations from the trendline. The concentration of points at lower values indicates a possible non-linear or heteroscedastic relationship, where Height may not predict Viscera.Weight effectively using a simple linear model.

7. Multiple Regression:

Four models were used to predict Rings. Model 1, using Height and Viscera.weight, explained 32% of variability (R^2=0.3203). Model 2, with all predictors, improved R^2 to 0.5275 but showed heteroscedasticity. Model 3 added polynomial terms, raising R^2 to 0.5634. Model 4, a log-transformed polynomial regression, achieved the best performance with R^2=0.6402 and stable variance.

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      2.24568    0.00297 756.202  < 2e-16
poly(Diameter, 2)1               4.59796    0.84630   5.433 5.86e-08
poly(Diameter, 2)2              -4.70629    0.30923 -15.220  < 2e-16
poly(Height, 2)1                 3.26418    0.49521   6.592 4.90e-11
poly(Height, 2)2                -1.33149    0.27623  -4.820 1.48e-06
poly(Whole.weight, 2)1          31.35541    2.35857  13.294  < 2e-16
poly(Whole.weight, 2)2          -5.74583    1.08532  -5.294 1.26e-07
poly(Shucked.weight, 2)1 -28.82723    1.16484 -24.748  < 2e-16
poly(Shucked.weight, 2)2  6.98124    0.61970  11.266  < 2e-16
poly(Viscera.weight, 2)1  -5.66554    0.94843  -5.974 2.52e-09
poly(Viscera.weight, 2)2   0.95825    0.52252   1.834  0.06674
poly(Shell.weight, 2)1     9.07709    1.09028   8.325  < 2e-16
poly(Shell.weight, 2)2    -1.50321    0.50025  -3.005  0.00267
```

Figure 5:Regression Parameters

The model shows strong linear and quadratic effects for most variables. Whole.weight and Shell.weight have the strongest positive linear impacts (**p < 2e-16**), while Shucked.weight has a significant negative relationship. Non-linear effects are evident for Diameter, Height, and Viscera.weight, with quadratic terms like poly(Whole.weight, 2)_2 and poly(Shucked.weight, 2)_2 also highly significant, reflecting complex relationships with log(Rings).

Other parameters of the different models fitted can be seen in Table 3

Equation of the Linear regression:

log(Rings)=2.24568+(4.598·poly(Diameter, 2)1)−(4.706·poly(Diameter, 2)2)+(3.264·poly(Height, 2)1)−(1.331·poly(Height, 2)2)+(31.355·poly(Whole.weight, 2)1)−(5.746·poly(Whole.weight, 2)2)−(28.827·poly(Shucked.weight, 2)1)+(6.981·poly(Shucked.weight, 2)2)−(5.666·poly(Viscera.weight, 2)1)+(0.958·poly(Viscera.weight, 2)2)+(9.077·poly(Shell.weight, 2)1)−(1.503·poly(Shell.weight, 2)2)

8. Model Diagnostics:

The residuals vs. fitted plot shows random scatter, indicating homoscedasticity, with minor non-linearity or outliers. The Q-Q plot confirms approximate normality, suggesting the model fits well while leaving room for slight improvements.

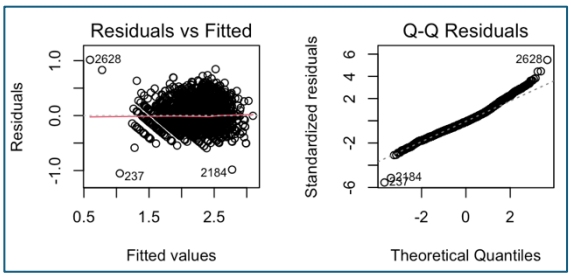

Figure 6: Residual Plots

Advanced Analysis

9. Principal Component Analysis(PCA):

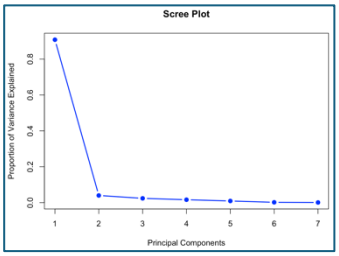

Figure 7: Scree Plot

Based on the scree plot, I would select the first principal component, which explains over 80% of the variance. This significant variance indicates that the first component effectively captures the dataset's structure for dimensionality reduction.

10. PCA Interpretation:

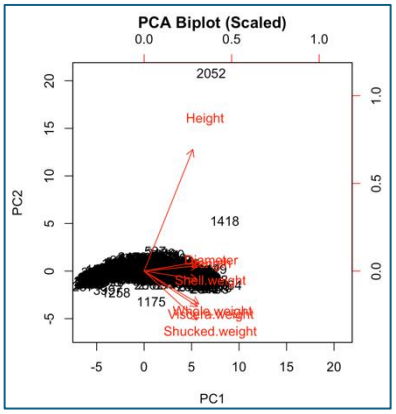

Figure 8: Biplot

The biplot reveals that **PC1** is heavily influenced by weight-related variables (Shell Weight, Viscera Weight, Shucked Weight), indicating their strong correlation. **PC2** is driven primarily by Height, which separates from the weight variables. This pattern highlights that weight characteristics dominate overall variance, while height provides unique structural variance in the data.

Conclusion:

The Abalone dataset analysis reveals physical measurements effectively predict age. Diameter is symmetric but has lower-end outliers, while Height shows a strong correlation with Viscera.Weight (r = 0.798). Multivariate models improved R^2 to 0.64 using log-transformed polynomials, capturing non-linearity. PCA highlights weight variables as dominant drivers of variance, with Height providing unique insights.

Boston Housing Dataset:

Univariate Analysis

1. Data Overview:

The Boston Housing dataset provides information on housing values in various suburbs of Boston, Massachusetts, collected during the 1970s. It contains 506 observations and 13 variables related to socio-economic, demographic, and geographical factors. The goal is to predict the median value of

```
'data.frame':   506 obs. of  15 variables:
 $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ crim    : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn      : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus   : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox     : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm      : num  6.58 6.42 7.18 7 7.15 ...
 $ age     : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis     : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad     : int  1 2 2 3 3 3 5 5 5 5 ...
 $ tax     : int  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black   : num  397 397 393 395 397 ...
 $ lstat   : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv    : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

Figure 9: Structure of the Data

owner-occupied homes, represented as MEDV in $1000s.

2. Summary Statistics:

Summary statistics (Table 4) provide an overview of the central tendency and spread of each variable.

Indus Variable:

The feature exhibits considerable variability, with a mean of 11.14 and a median of 9.69, indicating a slightly skewed distribution towards lower values. The range spans from 0.46 to 27.74, reflecting stark contrasts across regions, while the standard deviation of 6.86 highlights significant dispersion around the mean. This suggests diverse conditions or influences within the dataset.

3. Distribution Visualization:

The boxplot of indus shows a median around 10 and an interquartile range (IQR) spanning from approximately 5 to 15, capturing the middle 50% of the data. The whiskers extend from about 0.5 to 27, indicating a wide range without significant outliers. The symmetry of the box suggests a fairly balanced distribution within the IQR.

The histogram of indus shows a bimodal distribution, with peaks near 5 and 20. This indicates a clear divide, where some towns have low proportions of non-retail business areas, while others have high proportions. The highest frequency is around 20, suggesting many towns are industrially focused, with fewer in intermediate ranges. The data shows a wide spread.

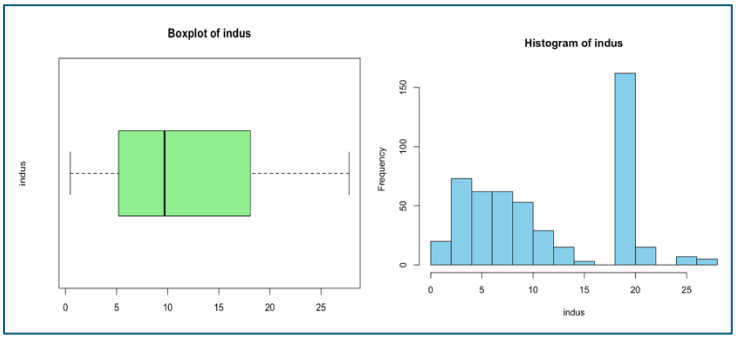

Figure 10:Box plot and Histogram

4. Categorical Variable Analysis:

The CHAS variable, indicating proximity to the Charles River, is highly imbalanced, with the majority of tracts (CHAS = 0) not bounding the river and only a small fraction (CHAS = 1) near it. This rarity suggests that river-adjacent tracts may have distinct characteristics, potentially influencing factors like property values or environmental conditions.

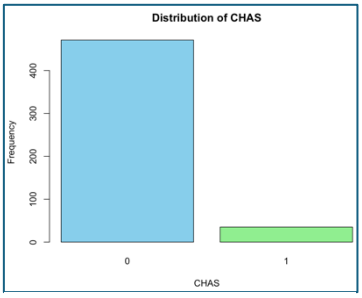

Figure 11: Distribution of CHAS

Multivariate Analysis

5. Correlation Analysis:

The **Pearson correlation coefficient** between age (proportion of older homes) and medv (median house prices) is **0.695**, indicating a strong positive linear relationship. This suggests that as the proportion of older homes increases in a region, the median house prices also tend to rise.

The positive correlation highlights the potential appeal or higher value associated with areas containing older, perhaps historic homes.
Correlation of each numerical variable w.r.t the other is shown in the Table 5

6. Scatter Plot Visualization:

The scatter plot illustrates a strong positive relationship between the average number of rooms (RM) and the median house value (MEDV). As RM increases, MEDV also rises, indicating that larger homes are generally more valuable. Most data points are clustered between 5 to 7 rooms, with house values ranging from $20,000 to $40,000, suggesting this is the typical range for homes. The red trendline confirms a linear trend, making RM a significant predictor of house prices. A few outliers, such as houses with more than 8 rooms or MEDV near 50, may represent luxury properties or unique market conditions.

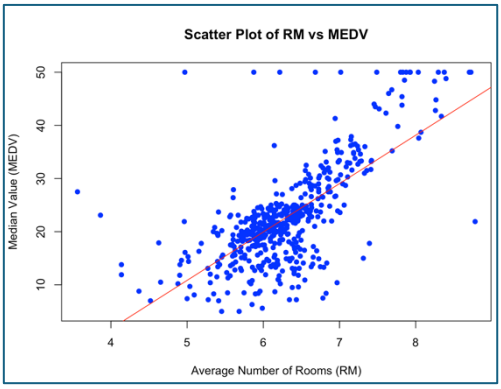

Figure 12:Scatter Plot

7. Multiple Regression:

A linear regression model with rm and lstat predicted medv with an R2 of 0.637. Adding ptratio improved R2 to 0.689. Including all variables, treating chas as categorical, increased R2 to 0.74, explaining 74% of the variability in medv.

Other parameters of the different models fitted can be seen in Table 6

Based on the model, significant predictors include rm (average rooms), lstat (lower status population), ptratio (pupil-teacher ratio), nox (air pollution), and dis(distance to employment centers), all with $p<0.05$. These variables have a substantial impact on house prices.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.341145   5.067492   7.171 2.73e-12
crim         -0.108413   0.032779  -3.307 0.001010
zn            0.045845   0.013523   3.390 0.000754
chas1         2.718716   0.854240   3.183 0.001551
nox         -17.376023   3.535243  -4.915 1.21e-06
rm            3.801579   0.406316   9.356  < 2e-16
dis          -1.492711   0.185731  -8.037 6.84e-15
rad           0.299608   0.063402   4.726 3.00e-06
tax          -0.011778   0.003372  -3.493 0.000521
ptratio      -0.946525   0.129066  -7.334 9.24e-13
black         0.009291   0.002674   3.475 0.000557
lstat        -0.522553   0.047424 -11.019  < 2e-16
```

Figure 13: Regression Parameters

Equation of the Linear regression:
$medv=36.34-0.1084(crim)+0.0458(zn)+2.7187(chas1)-17.3760(nox)+3.8016(rm)-1.4927(dis)+0.2996(rad)-0.0118(tax)-0.9465(ptratio)+0.0093(black)-0.5226(lstat)$

8. Model diagnostics:

The Residuals vs. Fitted plot suggests minor non-linearity and slight heteroscedasticity, while the Q-Q plot shows mild deviations from normality at the tails. Despite these issues, the model achieves a strong R2, indicating good explanatory power.

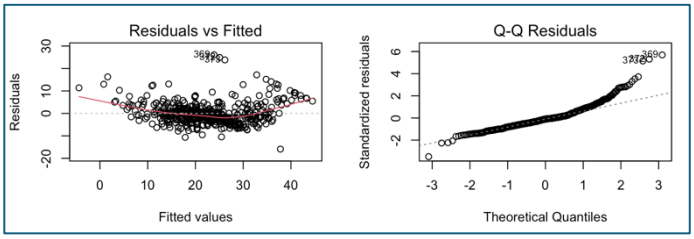

Figure 14: Residual Plot

Advanced Analysis

9. Principal Component Analysis:

Based on the scree plot, the first two components should be selected. The first component explains approximately **40%** of the variance, while the second explains around **20%**, capturing a cumulative **60% of the total variance**. Beyond the second component, additional components contribute minimal variance. Selecting these two components balances simplicity and explanatory power, effectively capturing the key structure of the data while avoiding unnecessary complexity.

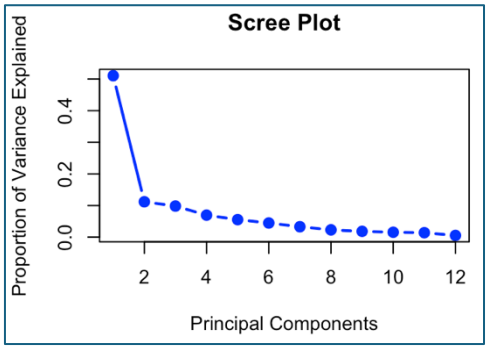

Figure 15: Scree Plot

10. PCA Interpretation:

The PCA biplot reveals that the first principal component (PC1) is strongly influenced by lstat, nox, indus, and ptratio, which are positively correlated, while rm and black are negatively correlated. PC2 is dominated by age, nox, and indus, with dis and zn contributing negatively. Variables like lstat and nox represent socio-economic and environmental factors, while rm and black indicate residential characteristics. Data points spread along PC1 suggest significant socio-economic variation, while PC2 captures additional spatial patterns. These components effectively group areas by housing, environmental, and demographic factors, highlighting key relationships and variance in the data.

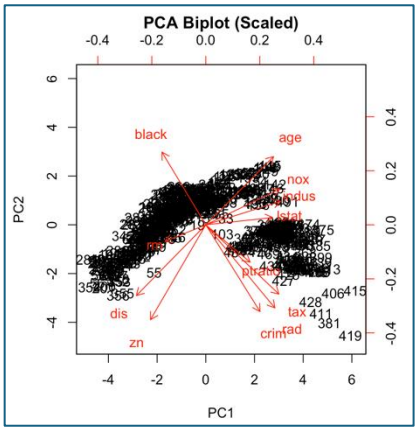

Figure 16: Biplot

Conclusion:

The analysis of the Boston Housing dataset identifies key predictors of house prices, including rm, lstat, ptratio, nox, and dis, with the final regression model explaining 74% of the variability in medv. PCA reveals that the first two components explain 60% of the variance (PC1: 40%, PC2: 20%), dominated by socio-economic (lstat, nox) and spatial factors (age, dis), effectively grouping regions by housing and environmental characteristics.

CarSeats Dataset:

Univariate Analysis:

1. Data Overview:

The **Carseats dataset** contains sales data for child car seats across 400 different stores. It includes information on various socio-economic, demographic, and competitive factors influencing sales. The data is structured

```
'data.frame':   400 obs. of  11 variables:
 $ Sales       : num  9.5 11.22 10.06 7.4 4.15 ...
 $ CompPrice   : num  138 111 113 117 141 124 115 136 132 132 ...
 $ Income      : num  73 48 35 100 64 113 105 81 110 113 ...
 $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
 $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
 $ Price       : num  120 83 80 97 128 72 108 120 124 124 ...
 $ ShelveLoc   : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
 $ Age         : num  42 65 59 55 38 78 71 67 76 76 ...
 $ Education   : num  17 10 12 14 13 16 15 10 10 17 ...
 $ Urban       : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
 $ US          : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

Figure 17: Structure of the Data

with 11 variables, both numeric and categorical, allowing for diverse analysis.

2. Summary Statistics:

Summary statistics (Table 7) provide an overview of the central tendency and spread of each variable.

Price Variable:

The Price variable ranges from $24 to $191, with an average of $115.80, a median of $117, and a standard deviation of $23.68, indicating moderate variability across stores. The wide range reflects significant pricing differences, influenced by competition or regional factors, making it a key variable for understanding its impact on car seat sales.

3. Distribution Visualization:

The histogram of Price exhibits a roughly symmetric, unimodal distribution centered around $115. Most prices fall within the range of $100 to $130, indicating a concentration near the mean. The distribution has a slight tail on both sides, suggesting potential outliers.

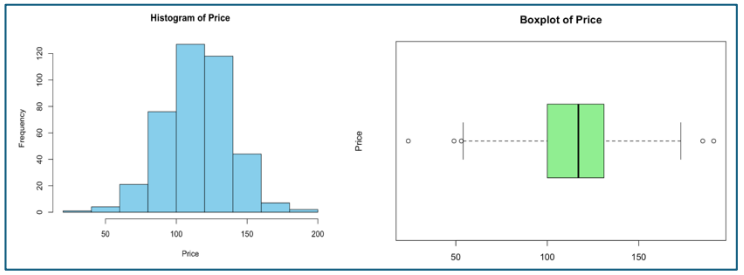

Figure 18:Histogram and Box plot

The boxplot of Price shows a fairly symmetric distribution centered around the median (~$117). The interquartile range (IQR) spans from approximately $100 to $130, capturing the middle 50% of the data. The whiskers extend beyond the IQR, indicating a range of typical values without significant skewness. However, a few outliers are present on both ends.

4. Categorical Variable Analysis:

The Urban variable shows most stores are located in urban areas, with over 250 stores categorized as Yes, compared to fewer than 150 in rural areas (No). This imbalance suggests a strong urban focus, potentially impacting sales trends and marketing strategies differently for urban and rural markets.

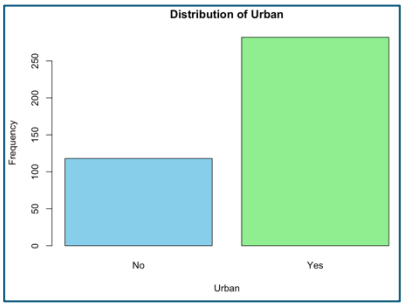

Figure 19:Distribution of Urban

Multivariate Analysis

5. Correlation Analysis:

The correlation between Price and Sales is **-0.445**, indicating a moderate negative relationship. As prices increase, sales tend to decline, reflecting consumer sensitivity to pricing. This highlights Price as a key factor influencing demand
Correlation of each numerical variable w.r.t the other is shown in the Table 8

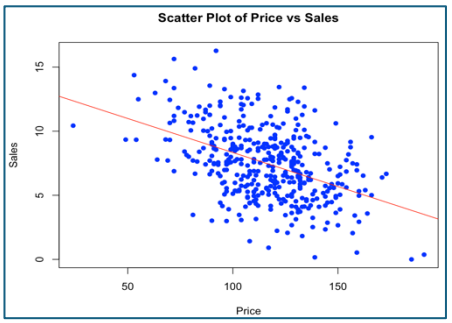

Figure 20: Scatter Plot

6. Scatter Plot Visualization:

The scatter plot shows a moderate negative relationship between Price and Sales, where higher prices correspond to lower sales, as confirmed by the trendline and correlation of -0.445. While most data points cluster around mid-range prices, the scatter indicates other factors may influence sales. This highlights the sensitivity of sales to price changes.

7. Multiple Regression:

Model 1, using Age and Price, explains only 27.6% of the variability in Sales (R2=0.27) with a higher residual error of 2.41. In contrast, Model 2 includes additional predictors like CompPrice, Income, Advertising, and ShelveLoc, significantly improving explanatory power (R2=0.872) and reducing residual error to 1.019. So we are going with model 2 for better predicting of sales

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.475226   0.505005   10.84   <2e-16
CompPrice       0.092571   0.004123   22.45   <2e-16
Income          0.015785   0.001838    8.59   <2e-16
Advertising     0.115903   0.007724   15.01   <2e-16
Price          -0.095319   0.002670  -35.70   <2e-16
ShelveLocGood   4.835675   0.152499   31.71   <2e-16
ShelveLocMedium 1.951993   0.125375   15.57   <2e-16
Age            -0.046128   0.003177  -14.52   <2e-16
```
Figure 21: Regression Parameters

Other parameters of both models fitted can be seen in Table 9

All parameters are highly significant ($p<0.001$), indicating their strong impact on Sales. Positive contributors include CompPrice, Income, Advertising, and shelving quality (ShelveLoc), highlighting their importance in driving sales. Conversely, Price and Age negatively affect sales, emphasizing the sensitivity of demand to pricing and population demographics.

Sales=5.4752+0.0926(CompPrice)+0.0158(Income)+0.1159(Advertising)−0.0953(Price)+4.8357 (ShelveLocGood)+1.9520(ShelveLocMedium)−0.0461(Age)

8. Model diagnostics:

The residuals vs. fitted plot shows random scatter around the horizontal line, confirming homoscedasticity and no strong non-linearity, indicating the model captures the linear relationship effectively. The Q-Q plot reveals residuals are approximately normally distributed, with minor deviations at the tails likely due to outliers. These deviations are minimal and do not compromise the model.

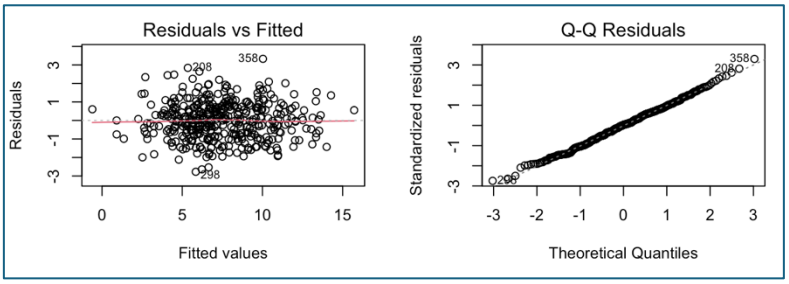

Figure 22: Residual Plot

Overall, the diagnostic plots suggest the assumptions of linear regression are reasonably met, supporting a good model fit.

Advanced Analysis

9. Principal Component Analysis:

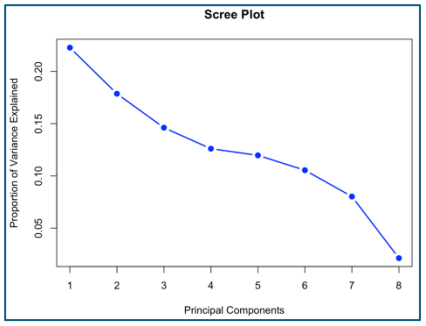

Figure 23: Scree Plot

I would choose 3 components based on the scree plot, as the "elbow point" occurs after the third component. These components explain 20%, 18%, and 15% of the variance, cumulatively capturing 53%. Subsequent components add minimal variance, making them less informative. This selection balances dimensionality reduction with retaining most of the data's structure.

10. PCA Interpretation:

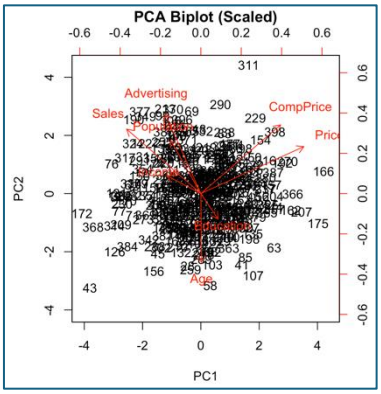

Figure 24: Biplot

The PCA biplot shows that PC1 is dominated by CompPrice and Price, capturing pricing-related variance, while PC2 highlights Advertising, Sales, and Population, reflecting marketing and demographic influences. CompPrice and Price cluster closely, indicating strong correlation, while Advertising and Sales align, emphasizing their marketing impact. These components effectively group observations based on competitive pricing and advertising dynamics, revealing key patterns in the data structure.

Conclusion:

The analysis of the Carseats dataset revealed several key insights. Univariate analysis highlighted that Price exhibits moderate variability and significantly influences sales, as shown by its negative correlation (-0.445) with Sales. Categorical analysis showed that urban stores dominate, suggesting urban-focused sales strategies. Multivariate analysis, through regression, identified CompPrice, Advertising, Income, and ShelveLoc as positive contributors to sales, while Price and Age negatively impact sales. Model diagnostics confirmed a strong fit for the regression model (R^2=0.872). PCA further revealed key relationships, with PC1 capturing pricing-related variance and PC2 emphasizing marketing influences, providing a comprehensive understanding of sales drivers.

Student Performance Dataset:

Univariate Analysis:

1. Data Overview:

The Student Performance Dataset is a dataset designed to examine the factors influencing academic student performance. The dataset consists of 10,000 student records

```
'data.frame':   10000 obs. of  6 variables:
 $ Hours.Studied                  : int  7 4 8 5 7 3 7 8 5 4 ...
 $ Previous.Scores                : int  99 82 51 52 75 78 73 45 77 89 ...
 $ Extracurricular.Activities     : chr  "Yes" "No" "Yes" "Yes" ...
 $ Sleep.Hours                    : int  9 4 7 5 8 9 5 4 8 4 ...
 $ Sample.Question.Papers.Practiced: int  1 2 2 2 5 6 6 6 2 0 ...
 $ Performance.Index              : num  91 65 45 36 66 61 63 42 61 69 ...
```

Figure 25:Structure of the Data

and 6 variables, with each record containing information about various predictors and a performance index.

2. Summary statistics:

Summary statistics (Table 10) provide an overview of the central tendency and spread of each variable.

Scores
The numeric variable scores has a mean of 69.45 and a median of 69, indicating a symmetric distribution. The standard deviation of 17.34 reflects moderate variability. The range spans from a minimum of 40 to a maximum of 99, showing a widespread. This suggests that the data is evenly distributed around the central value, with notable variability across observations.

3. Distribution Visualization:

The histogram of Previous.Scores exhibits a relatively uniform distribution, with frequencies evenly spread across the score range of 40 to 100. There are no clear peaks or significant skewness, suggesting the data is uniformly distributed. The absence of extreme bars at the edges indicates no obvious outliers. This distribution suggests a balanced representation of scores across the range.

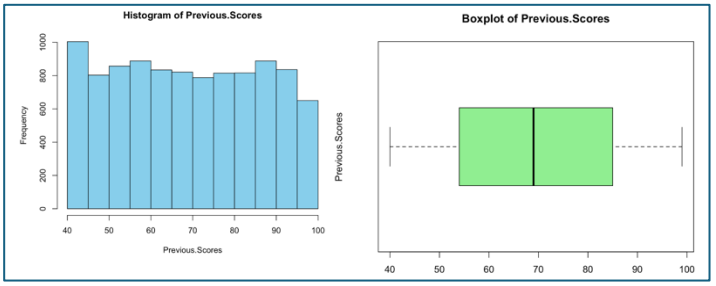

Figure 26:Histogram and Box plot

The boxplot of Previous.Scores shows a symmetric distribution, with the median centered between the interquartile range (IQR). The data spans from approximately 40 to 100, with no visible outliers as all data points fall within the whiskers.

4. Categorical Variable Analysis:

The distribution of Sleep.Hours shows a relatively uniform spread across the range of 4 to 9 hours, with a slight peak at 8 hours. This indicates that most individuals tend to sleep around 8 hours, while other durations are evenly distributed. There are no extreme peaks or dips, suggesting consistent sleeping patterns across the population. The data does not indicate any potential outliers.

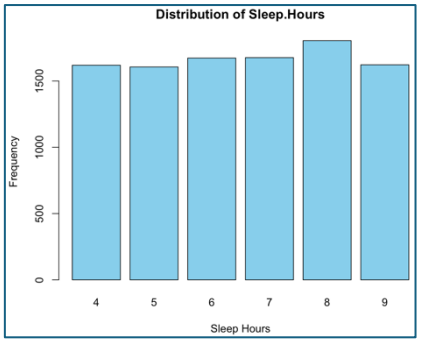

Figure 27:Distribution of Sleep Hours

Multivariate Analysis

5. Correlation Analysis:

The Pearson correlation coefficient of 0.915 indicates a very strong positive relationship between Previous.Scores and Performance.Index. This suggests that higher previous scores are strongly associated with better performance, highlighting the predictive value of prior academic

achievement.
Correlation of each numerical variable w.r.t the other is shown in the Table 11

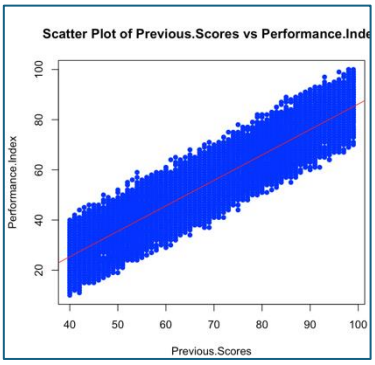

Figure 28:Scatter Plot

6. Scatterplot Visualization:

The scatter plot of Previous.Scores vs. Performance.Index shows a strong positive linear relationship, supported by the red trendline. As Previous.Scores increase, the Performance.Index also rises consistently, indicating that students with higher prior scores tend to perform better. The tightly clustered points around the trendline and minimal scatter suggest a strong correlation, confirming that Previous.Scores is a reliable predictor of performance.

7. Multiple Regression:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -29.816790   0.104529  -285.2   <2e-16
Previous.Scores  1.019123   0.001317   773.8   <2e-16
Hours.Studied    2.857637   0.008821   323.9   <2e-16
```

Figure 29:Regression Parameters

Model 1, using only **Previous.Scores**, explains 83.8% of the variability in **Performance.Index** (R^2=0.838) with a residual standard error of 7.744. The residuals show approximate normality, but the higher residual error indicates room for improvement. Model 2, incorporating **Hours.Studied** as an additional predictor, significantly improves explanatory power (R^2=0.986) and reduces residual error to 2.284. Moreover, the residuals in Model 2 exhibit near-perfect normality, confirming a better fit.

Other parameters of both models fitted can be seen in Table 12

Both predictors are highly significant ($p<0.001$), with **Previous.Scores** and **Hours.Studied** positively influencing **Performance.Index**. The regression equation for Model 2 is:

Performance.Index = -29.817 + 1.019(Previous.Scores) + 2.858(Hours.Studied)

8. Model Diagnostics:

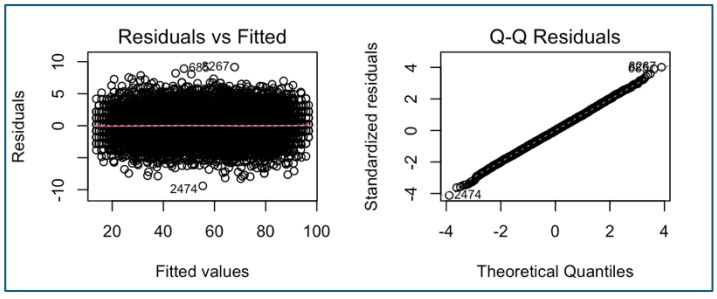

Figure 30:Residual Plot

The residuals vs. fitted plot shows a random scatter, confirming homoscedasticity and no significant non-linearity, validating the linear model's appropriateness. The Q-Q plot reveals residuals aligning closely with the theoretical quantile line, indicating approximate normality. These diagnostics confirm that the model satisfies the assumptions of linear regression, ensuring reliable predictions and valid statistical inferences. Overall, the model demonstrates a good fit to the data.

Advanced Analysis:

9. Principal Component Analysis:

Based on the scree plot, I would choose the first two components. The "elbow" occurs after the second component, where the rate of variance explained significantly decreases. These two components collectively explain approximately 60% of the total variance, striking a balance between dimensionality reduction and retaining essential information. Additional components contribute minimal variance, making them less impactful for the analysis.

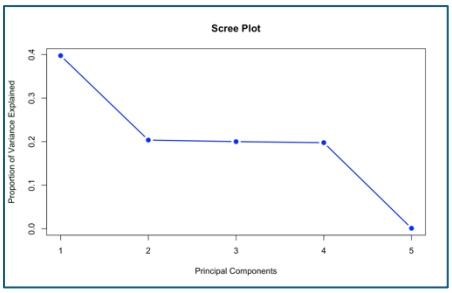

Figure 31:Scree Plot

10. PCA Interpretation:

The PCA biplot reveals the loadings of the first two principal components (PC1 and PC2). **PC1** is primarily influenced by "Previous Scores" and "Performance Index," both contributing strongly in the negative direction, capturing trends in academic performance. "Sleep Hours" shows a moderate positive influence on PC1. **PC2** is dominated by "Hours Studied" and "Sample Question Papers Practiced," which align positively, reflecting study-related behaviors. "Sleep Hours" also contributes positively but to a lesser extent. The plot demonstrates distinct groupings, with academic performance clustering along PC1 and study preparation aligning with PC2, highlighting the relationships between performance and preparation efforts.

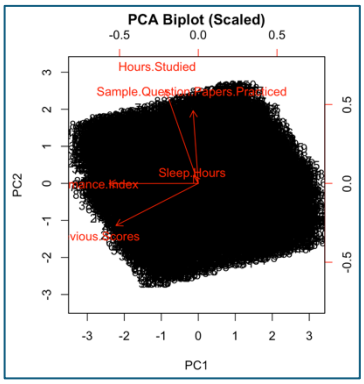

Figure 32:BiPlot

Conclusion:

The Student Performance dataset highlights key insights through univariate and multivariate analysis. **Previous.Scores** shows a symmetric, uniform distribution, while **Sleep.Hours** peaks around 8 hours, reflecting balanced habits. **Performance.Index** reveals moderate variability, indicating diverse academic outcomes. Multivariate analysis identifies a strong linear relationship between **Previous.Scores** and **Performance.Index** (correlation = 0.915), with higher scores predicting better performance. Adding **Hours.Studied** to regression models improves explanatory power (R^2=0.986), emphasizing its importance alongside prior scores. Principal Component Analysis identifies two components explaining 60% of the variance, one capturing academic performance and the other study-related behaviors. Overall, the dataset emphasizes preparation and past performance in predicting success.

Appendix:

Abalone Dataset:

Table 1:

return

Variable	Min	Max	Mean	Median	StdDev
Length	0.075	0.815	0.5239921	0.545	0.1200929
Diameter	0.055	0.65	0.4078813	0.425	0.0992399
Height	0	1.13	0.1395164	0.14	0.0418271
Whole.weigh	0.002	2.8255	0.8287422	0.7995	0.490389
Shucked.wei	0.001	1.488	0.3593675	0.336	0.2219629
Viscera.weig	0.0005	0.76	0.1805936	0.171	0.1096143
Shell.weight	0.0015	1.005	0.2388309	0.234	0.1392027
Rings	1	29	9.9336845	9	3.224169

Table 2:

return

	Length	Diameter	Height	Whole.weight	Shucked.weight	Viscera.weight	Shell.weight	Rings
Length	1.00	0.99	0.83	0.93	0.90	0.90	0.90	0.56
Diameter	0.99	1.00	0.83	0.93	0.89	0.90	0.91	0.57
Height	0.83	0.83	1.00	0.82	0.77	0.80	0.82	0.56
Whole.weigh	0.93	0.93	0.82	1.00	0.97	0.97	0.96	0.54
Shucked.wei	0.90	0.89	0.77	0.97	1.00	0.93	0.88	0.42
Viscera.weig	0.90	0.90	0.80	0.97	0.93	1.00	0.91	0.50
Shell.weight	0.90	0.91	0.82	0.96	0.88	0.91	1.00	0.63
Rings	0.56	0.57	0.56	0.54	0.42	0.50	0.63	1.00

Table 3:

return

Metric	Model 1 (Height + Viscera.weight)	Model 2 (All variables)	Model 3 (Polynomial)	Model 4 (Log-transformed Polynomial)
Residual Std. Error	2.659	2.218	2.133	0.1919
R-squared	0.3203	0.5275	0.5634	0.6402
Adjusted R-squared	0.32	0.5269	0.5621	0.6392
F-statistic	983.5	776	447.8	617.5
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Boston Data:

Table 4:

return

Variable	Min	Max	Mean	Median	StdDev
X	1	506	253.5	253.5	146.213884
crim	0.00632	88.9762	3.6135236	0.25651	8.6015451
zn	0	100	11.3636364	0	23.322453
indus	0.46	27.74	11.1367787	9.69	6.8603529
nox	0.385	0.871	0.5546951	0.538	0.1158777
rm	3.561	8.78	6.2846344	6.2085	0.7026171
age	2.9	100	68.5749012	77.5	28.1488614
dis	1.1296	12.1265	3.7950427	3.20745	2.1057101
rad	1	24	9.5494071	5	8.7072594
tax	187	711	408.237154	330	168.537116
ptratio	12.6	22	18.4555336	19.05	2.1649455
black	0.32	396.9	356.674032	391.44	91.2948644
lstat	1.73	37.97	12.6530632	11.36	7.1410615
medv	5	50	22.5328063	21.2	9.1971041

Table 5:

return

	X	crim	zn	indus	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
X	1	0.4074072	-0.1033934	0.3994389	0.3987362	-0.0799712	0.2037835	-0.302211	0.686002	0.6666259	0.2910742	-0.2950412	0.2584648	-0.2266036
crim	0.40740717	1	-0.2004692	0.4065834	0.4209717	-0.2192467	0.3527343	-0.3796701	0.6255051	0.5827643	0.2899456	-0.3850639	0.4556215	-0.3883046
zn	-0.1033934	-0.2004692	1	-0.5338282	-0.5166037	0.31199059	-0.5695373	0.6644082	-0.3119478	-0.3145633	-0.3916785	0.1755203	-0.4129946	0.3604453
indus	0.39943885	0.4065834	-0.5338282	1	0.7636514	-0.3916759	0.6447785	-0.708027	0.5951293	0.7207602	0.3832476	-0.3569765	0.6037997	-0.4837252
nox	0.39873617	0.4209717	-0.5166037	0.7636514	1	-0.3021882	0.7314701	-0.7692301	0.6114406	0.6680232	0.1889327	-0.3800506	0.5908789	-0.4273208
rm	-0.0799712	-0.2192467	0.3119906	-0.3916759	-0.3021882	1	-0.2402649	0.2052462	-0.2098467	-0.2920478	-0.3555015	0.1280686	-0.6138083	0.6953599
age	0.20378351	0.3527343	-0.5695373	0.6447785	0.7314701	-0.2402649	1	-0.7478805	0.4560225	0.5064556	0.261515	-0.273534	0.6023385	-0.3769546
dis	-0.302211	-0.3796701	0.6644082	-0.708027	-0.7692301	0.20524621	-0.7478805	1	-0.4945879	-0.5344316	-0.2324705	0.2915117	-0.4969958	0.2499287
rad	0.68600198	0.6255051	-0.3119478	0.5951293	0.6114406	-0.2098467	0.4560225	-0.4945879	1	0.9102282	0.4647412	-0.4444128	0.4886763	-0.3816262
tax	0.66662592	0.5827643	-0.3145633	0.7207602	0.6680232	-0.2920478	0.5064556	-0.5344316	0.9102282	1	0.460853	-0.441808	0.5439934	-0.4685359
ptratio	0.29107423	0.2899456	-0.3916785	0.3832476	0.1889327	-0.3555015	0.261515	-0.2324705	0.4647412	0.460853	1	-0.1773833	0.3740443	-0.5077867
black	-0.2950412	-0.3850639	0.1755203	-0.3569765	-0.3800506	0.12806864	-0.273534	0.2915117	-0.4444128	-0.441808	-0.1773833	1	-0.3660869	0.3334608
lstat	0.25846477	0.4556215	-0.4129946	0.6037997	0.5908789	-0.6138083	0.6023385	-0.4969958	0.4886763	0.5439934	0.3740443	-0.3660869	1	-0.7376627
medv	-0.2266036	-0.3883046	0.3604453	-0.4837252	-0.4273208	0.69535995	-0.3769546	0.2499287	-0.3816262	-0.4685359	-0.5077867	0.3334608	-0.7376627	1

Table 6:

return

Model	Intercept	R-squared	Adjusted R-squared	Residual Std. Error	F-statistic	Degrees of Freedom
rm + lstat	-1.35827	0.6386	0.6371	5.54	444.3	503
rm + lstat + ptratio	18.56711	0.6786	0.6767	5.229	353.3	502
All Variables	36.461352	0.7414	0.734	4.743	100.6	491

Car Seats data:

Table 7:

return

Variable	Min	Max	Mean	Median	StdDev
Sales	0	16.27	7.496325	7.49	2.824115
CompPrice	77	175	124.975	125	15.334511
Income	21	120	68.6575	69	27.986037
Advertising	0	29	6.635	5	6.650364
Population	10	509	264.84	272	147.376436
Price	24	191	115.795	117	23.676664
Age	25	80	53.3225	54.5	16.200297
Education	10	18	13.9	14	2.620528

Table 8:

return

	Sales	CompPrice	Income	Advertising	Population	Price	Age	Educatior
Sales	1.00	0.06	0.15	0.27	0.05	-0.44	-0.23	-0.05
CompPrice	0.06	1.00	-0.08	-0.02	-0.09	0.58	-0.10	0.03
Income	0.15	-0.08	1.00	0.06	-0.01	-0.06	0.00	-0.06
Advertising	0.27	-0.02	0.06	1.00	0.27	0.04	0.00	-0.03
Population	0.05	-0.09	-0.01	0.27	1.00	-0.01	-0.04	-0.11
Price	-0.44	0.58	-0.06	0.04	-0.01	1.00	-0.10	0.01
Age	-0.23	-0.10	0.00	0.00	-0.04	-0.10	1.00	0.01
Education	-0.05	0.03	-0.06	-0.03	-0.11	0.01	0.01	1.00

Table 9:

return

Metric	Model 1 (Age + Price)	Model 2 (CompPrice, Income, Advertising, Price, ShelveLoc, Age)
Residual Std.	2.41	1.019
R-squared	0.2757	0.872
Adjusted R-s	0.272	0.8697
F-statistic	75.55	381.4
p-value	< 2.2e-16	< 2.2e-16

Student Performance data:

Table 10:

return

Variable	Min	Max	Mean	Median	StdDev
Hours.Studied	1	9	4.9929	5	2.589309
Previous.Scores	40	99	69.4457	69	17.343152
Sleep.Hours	4	9	6.5306	7	1.695863
Sample.Question.Papers.Practiced	0	9	4.5833	5	2.867348
Performance.Index	10	100	55.2248	55	19.212558

Table 11:

return

	Hours.Studie	Previous.Sco	Sleep.Hours	Sample.Ques	Performance
Hours.Studie	1.00	-0.01	0.00	0.02	0.37
Previous.Sco	-0.01	1.00	0.01	0.01	0.92
Sleep.Hours	0.00	0.01	1.00	0.00	0.05
Sample.Ques	0.02	0.01	0.00	1.00	0.04
Performance	0.37	0.92	0.05	0.04	1.00

Table 12:

return

Metric	Model 1 (Previous.Score	Model 2 (Previous.Scores + Hours.Studied)
Residual Std.	7.744	2.284
R-squared	0.8376	0.9859
Adjusted R-sc	0.8376	0.9859
F-statistic	51560	348800
p-value	< 2.2e-16	< 2.2e-16