STAT632 Project: Advertisement Sales Modeling

```
# --- 1. Load Required Libraries ---
library(MASS)          # Box-Cox transformation
library(glmnet)        # LASSO regression
```

```
Loading required package: Matrix
```

```
Loaded glmnet 4.1-8
```

```
library(randomForest)  # Random Forest
```

```
randomForest 4.7-1.2
```

```
Type rfNews() to see new features/changes/bug fixes.
```

```
library(car)           # Multicollinearity (VIF)
```

```
Loading required package: carData
```

```
library(ggplot2)       # Data visualization
```

```
Attaching package: 'ggplot2'
```

```
The following object is masked from 'package:randomForest':
```

```
    margin
```

```r
library(caret)          # Train/test split and validation
```

```
Loading required package: lattice
```

```r
library(dplyr)          # Data manipulation
```

```
Attaching package: 'dplyr'

The following object is masked from 'package:car':

    recode

The following object is masked from 'package:randomForest':

    combine

The following object is masked from 'package:MASS':

    select

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
# --- 2. Load and Explore Data ---
adver <- read.csv("Advertising And Sales.csv")
summary(adver)
```

```
       ID               TV              Radio          Newspaper     
 Min.   :  1.00   Min.   :  0.70   Min.   : 0.00   Min.   :  0.30  
 1st Qu.: 50.75   1st Qu.: 74.38   1st Qu.:10.07   1st Qu.: 12.75  
 Median :100.50   Median :149.75   Median :22.90   Median : 25.75  
 Mean   :100.50   Mean   :147.03   Mean   :23.29   Mean   : 30.55  
```

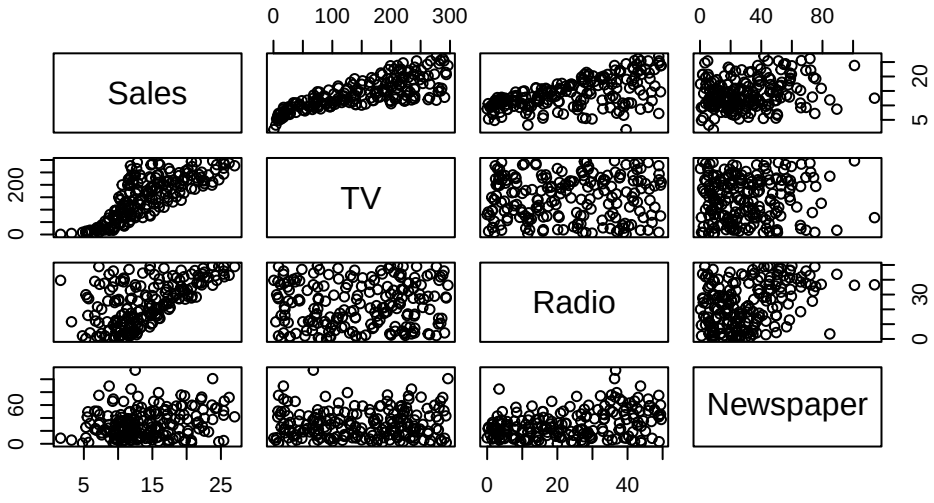
```
 3rd Qu.:150.25   3rd Qu.:218.82   3rd Qu.:36.52   3rd Qu.: 45.10
 Max.   :200.00   Max.   :296.40   Max.   :49.60   Max.   :114.00
     Sales
 Min.   : 1.60
 1st Qu.:10.40
 Median :12.90
 Mean   :14.04
 3rd Qu.:17.40
 Max.   :27.00
```

```r
str(adver)
```

```
'data.frame':   200 obs. of  5 variables:
 $ ID       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ Radio    : num  37.8 39.3 45.9 41.3 12.8 48.9 32.8 19.6 2.1 2.6 ...
 $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ Sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

```r
# --- 3. Pairwise Scatterplot ---
pairs(Sales ~ TV + Radio + Newspaper, data = adver,
      main = "Pairwise Scatterplot of Sales and Advertising Channels")
```

Pairwise Scatterplot of Sales and Advertising Channels



```r
# --- 4. Remove Outliers ---
adver <- adver[-c(131, 156, 99, 108, 200), ]
```

```r
# --- 5. Check Missing Values ---
colSums(is.na(adver))
```

```
       ID        TV     Radio Newspaper     Sales
        0         0         0         0         0
```

```r
# --- 6. Base Linear Model (TV, Radio, Newspaper) ---
lm1 <- lm(Sales ~ TV + Radio + Newspaper, data = adver)
summary(lm1)
```

```
Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = adver)

Residuals:
    Min      1Q  Median      3Q     Max
-5.4236 -0.9315  0.1920  1.1625  2.7446

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.171631   0.296987  10.679   <2e-16 ***
TV           0.044623   0.001321  33.778   <2e-16 ***
Radio        0.193681   0.008097  23.921   <2e-16 ***
Newspaper   -0.005430   0.005479  -0.991    0.323
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.558 on 191 degrees of freedom
Multiple R-squared:  0.907, Adjusted R-squared:  0.9055
F-statistic: 620.8 on 3 and 191 DF,  p-value: < 2.2e-16
```

```r
# --- 7. Reduced Linear Model (TV and Radio only) ---
lm2 <- lm(Sales ~ TV + Radio, data = adver)
summary(lm2)
```

```
Call:
```

```
lm(formula = Sales ~ TV + Radio, data = adver)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5951 -0.8560  0.2263  1.1425  2.7664

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.073673   0.280043   10.98   <2e-16 ***
TV          0.044597   0.001321   33.77   <2e-16 ***
Radio       0.190869   0.007583   25.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.558 on 192 degrees of freedom
Multiple R-squared:  0.9065,    Adjusted R-squared:  0.9055
F-statistic: 930.8 on 2 and 192 DF,  p-value: < 2.2e-16
```

```r
# --- 8. Interaction Model (TV * Radio * Newspaper) ---
lm_interaction <- lm(Sales ~ TV * Radio * Newspaper, data = adver)
summary(lm_interaction)
```

```
Call:
lm(formula = Sales ~ TV * Radio * Newspaper, data = adver)

Residuals:
   Min     1Q Median     3Q    Max
-2.820 -0.371  0.144  0.488  1.458

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.818e+00  4.028e-01  16.926  < 2e-16 ***
TV                  1.844e-02  2.346e-03   7.859 2.96e-13 ***
Radio               3.435e-02  1.425e-02   2.410   0.0169 *
Newspaper           1.293e-03  1.462e-02   0.088   0.9297
TV:Radio            1.089e-03  8.411e-05  12.950  < 2e-16 ***
TV:Newspaper       -3.634e-06  7.919e-05  -0.046   0.9635
Radio:Newspaper    -4.164e-05  4.108e-04  -0.101   0.9194
TV:Radio:Newspaper -3.037e-07  2.294e-06  -0.132   0.8948
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.7803 on 187 degrees of freedom
Multiple R-squared:  0.9772,    Adjusted R-squared:  0.9763
F-statistic:  1143 on 7 and 187 DF,  p-value: < 2.2e-16
```

```r
# --- 9. Full Polynomial + Interaction Model ---
lm_poly <- lm(Sales ~ TV + I(TV^2) + Radio + I(Radio^2) +
              Newspaper + I(Newspaper^2) + I(TV^3) + I(TV^4) + I(TV^5) +
              TV * Radio * Newspaper, data = adver)
summary(lm_poly)
```

```
Call:
lm(formula = Sales ~ TV + I(TV^2) + Radio + I(Radio^2) + Newspaper +
    I(Newspaper^2) + I(TV^3) + I(TV^4) + I(TV^5) + TV * Radio *
    Newspaper, data = adver)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84050 -0.20301 -0.01117  0.18327  0.84042

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.594e+00  2.256e-01  15.930  < 2e-16 ***
TV                  1.500e-01  1.175e-02  12.768  < 2e-16 ***
I(TV^2)            -1.677e-03  2.420e-04  -6.931 7.10e-11 ***
Radio               3.178e-02  7.723e-03   4.114 5.89e-05 ***
I(Radio^2)          1.954e-04  1.333e-04   1.466 0.144304
Newspaper          -1.477e-04  6.455e-03  -0.023 0.981768
I(Newspaper^2)     -5.305e-08  4.129e-05  -0.001 0.998976
I(TV^3)             1.020e-05  2.033e-06   5.017 1.25e-06 ***
I(TV^4)            -3.028e-08  7.478e-09  -4.049 7.61e-05 ***
I(TV^5)             3.453e-11  9.984e-12   3.459 0.000677 ***
TV:Radio            1.057e-03  3.816e-05  27.706  < 2e-16 ***
TV:Newspaper        4.602e-06  3.615e-05   0.127 0.898857
Radio:Newspaper     7.962e-05  2.004e-04   0.397 0.691651
TV:Radio:Newspaper -6.359e-07  1.081e-06  -0.588 0.557261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3267 on 181 degrees of freedom
Multiple R-squared:  0.9961,    Adjusted R-squared:  0.9958
```

```
F-statistic:  3578 on 13 and 181 DF,  p-value: < 2.2e-16
```

```r
# --- 10. Reduced Polynomial + Interaction Model ---
lm_poly1 <- lm(Sales ~ TV + I(TV^2) + Radio + TV:Radio +
               I(TV^3) + I(TV^4) + I(TV^5), data = adver)
summary(lm_poly1)
```

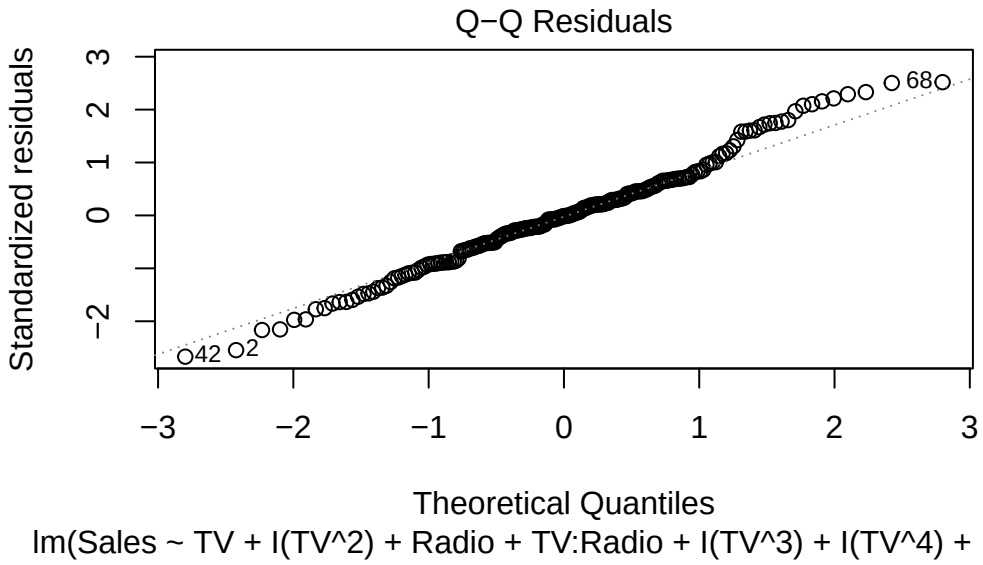
```
Call:
lm(formula = Sales ~ TV + I(TV^2) + Radio + TV:Radio + I(TV^3) +
    I(TV^4) + I(TV^5), data = adver)

Residuals:
     Min       1Q   Median       3Q      Max
-0.85876 -0.19498 -0.00409  0.17912  0.81061

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.541e+00  1.877e-01  18.865  < 2e-16 ***
TV           1.472e-01  1.145e-02  12.854  < 2e-16 ***
I(TV^2)     -1.610e-03  2.327e-04  -6.919 7.01e-11 ***
Radio        4.484e-02  3.191e-03  14.051  < 2e-16 ***
I(TV^3)      9.599e-06  1.952e-06   4.919 1.90e-06 ***
I(TV^4)     -2.795e-08  7.164e-09  -3.901 0.000134 ***
I(TV^5)      3.125e-11  9.536e-12   3.277 0.001250 **
TV:Radio     1.030e-03  1.870e-05  55.094  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3255 on 187 degrees of freedom
Multiple R-squared:  0.996, Adjusted R-squared:  0.9959
F-statistic:  6698 on 7 and 187 DF,  p-value: < 2.2e-16
```

```r
# --- 11. Residual Diagnostics ---
plot(lm_poly1, which = 2)  # Q-Q plot
```




```r
shapiro.test(residuals(lm_poly1))  # Normality test
```

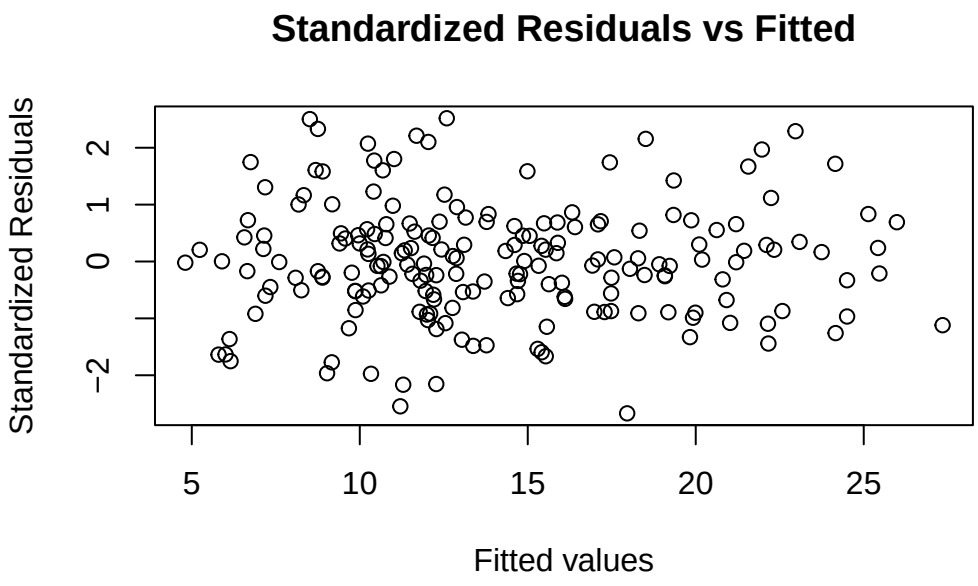
```
    Shapiro-Wilk normality test

data:  residuals(lm_poly1)
W = 0.98932, p-value = 0.1537
```

```r
# Standardized residuals
std_residuals <- rstandard(lm_poly1)
outliers <- which(abs(std_residuals) > 3)
cat("Outliers are at rows:", outliers, "\n")
```

```
Outliers are at rows:
```

```r
# Residuals vs Fitted
plot(lm_poly1$fitted.values, std_residuals,
     main = "Standardized Residuals vs Fitted",
     xlab = "Fitted values", ylab = "Standardized Residuals")
abline(h = c(-3, 3), col = "red", lty = 2)
```




```
# --- 12. Model Comparison Table ---
models <- list(
  "Full Model" = lm1,
  "Polynomial Model" = lm_poly,
  "Interaction Model" = lm_interaction,
  "Reduced Polynomial" = lm_poly1
)

aic_table <- data.frame(
  Model = names(models),
  Adjusted_R2 = sapply(models, function(m) round(summary(m)$adj.r.squared, 4)),
  Residual_Std_Error = sapply(models, function(m) round(summary(m)$sigma, 3)),
  AIC = sapply(models, function(m) round(AIC(m), 2))
)
print(aic_table)
```

```
                              Model Adjusted_R2 Residual_Std_Error    AIC
Full Model                 Full Model      0.9055              1.558 732.33
Polynomial Model     Polynomial Model      0.9958              0.327 132.62
Interaction Model   Interaction Model      0.9763              0.780 466.48
Reduced Polynomial Reduced Polynomial      0.9959              0.325 125.44
```

```
# --- 13. Cross-Validation (Train/Test Split) ---
set.seed(213)
train_index <- createDataPartition(adver$Sales, p = 0.7, list = FALSE)
```

```r
train_data <- adver[train_index, ]
test_data <- adver[-train_index, ]

# Fit the final model
lm_model <- lm(Sales ~ TV + I(TV^2) + Radio + TV:Radio +
               I(TV^3) + I(TV^4) + I(TV^5), data = adver)

# RMSE for training and testing sets
train_rmse <- sqrt(mean((predict(lm_model, train_data) - train_data$Sales)^2))
test_rmse <- sqrt(mean((predict(lm_model, test_data) - test_data$Sales)^2))
cat("Train RMSE:", train_rmse, "Test RMSE:", test_rmse, "\n")
```

```
Train RMSE: 0.3122478 Test RMSE: 0.3338593
```

```r
# --- 14. Random Forest Model and Comparison ---
rf_model <- randomForest(Sales ~ TV + I(TV^2) + Radio + I(TV^3) + I(TV^4) +
                         TV:Radio + I(TV^5), data = adver)

# Compare RMSE
lm_pred <- predict(lm_model, adver)
rf_pred <- predict(rf_model, adver)
cat("Linear Model RMSE:", sqrt(mean((lm_pred - adver$Sales)^2)), "\n")
```

```
Linear Model RMSE: 0.3187166
```

```r
cat("Random Forest RMSE:", sqrt(mean((rf_pred - adver$Sales)^2)), "\n")
```

```
Random Forest RMSE: 0.6452739
```

```r
# --- 15. LASSO Regression ---
x <- model.matrix(Sales ~ TV + I(TV^2) + Radio + TV:Radio +
                  I(TV^3) + I(TV^4), data = adver)[, -1]
y <- adver$Sales

# Cross-validated LASSO
lasso_cv <- cv.glmnet(x, y, alpha = 1)
cat("Best lambda:", lasso_cv$lambda.min, "\n")
```

```
Best lambda: 0.001244067
```

```r
# Final LASSO model
lasso_model <- glmnet(x, y, alpha = 1, lambda = lasso_cv$lambda.min)
lasso_predictions <- predict(lasso_model, newx = x)
head(lasso_predictions)
```

```
         s0
1 21.468248
2 10.798021
3  8.652071
4 18.525758
5 13.474984
6  7.892222
```