Marketing Data Analysis and Sales Prediction Using Multiple Linear Regression

By

Shivaji Reddy Sama

F454W858

Wichita State University

INTRODUCTION

- Advertising is a technique and practice used to bring products, services, opinions, or causes to public notice for the purpose of persuading the public to respond in a certain way toward what is advertised.
- In order to maximize sales, publishers post ads in multiple advertising medias.
- But with numerous advertising platforms available , the publisher should know on which platforms he/she should publish the ads to increase the chance of selling the product and get maximum ROI.
- The aim is to predict the number of sales by advertising from different platforms by using Multiple Linear Regression.

DATASET INFORMATION

- Dataset is downloaded from Kaggle.

DATA DESCRIPTION:

- This dataset contains the impact of three advertising medias (youtube, facebook and newspaper) on sales. Data are the advertising budget in thousands of dollars along with the sales (in thousands of units). The advertising experiment has been repeated 200 times.
- The aim is to predict the number of sales by advertising from different platforms.
- Independent variables: youtube, facebook, newspaper.
- Dependent variable : sales

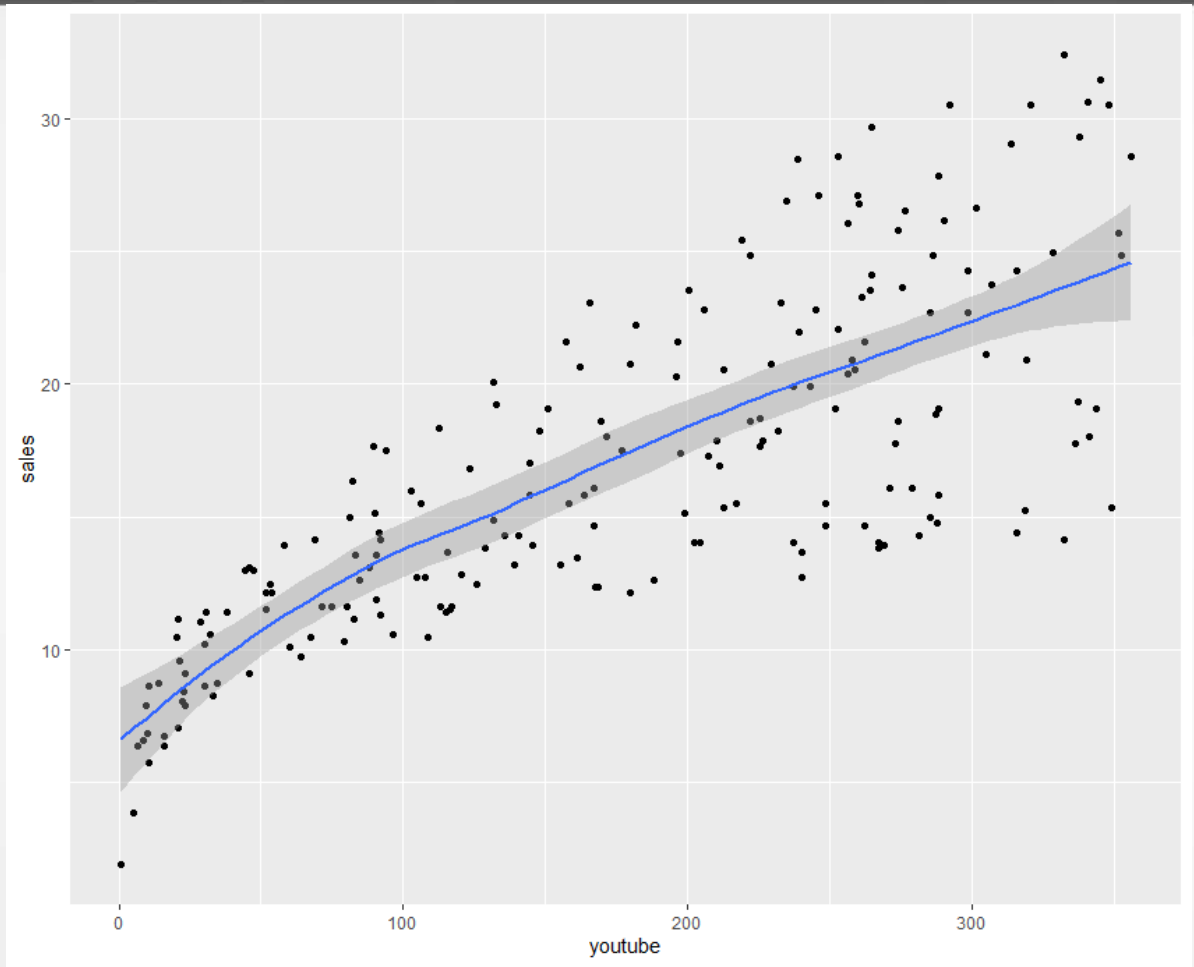
Wichita State University

SUMMARY OF THE DATA

```
> head(marketing)
  youtube facebook newspaper sales
1  276.12    45.36     83.04 26.52
2   53.40    47.16     54.12 12.48
3   20.64    55.08     83.16 11.16
4  181.80    49.56     70.20 22.20
5  216.96    12.96     70.08 15.48
6   10.44    58.68     90.00  8.64
```

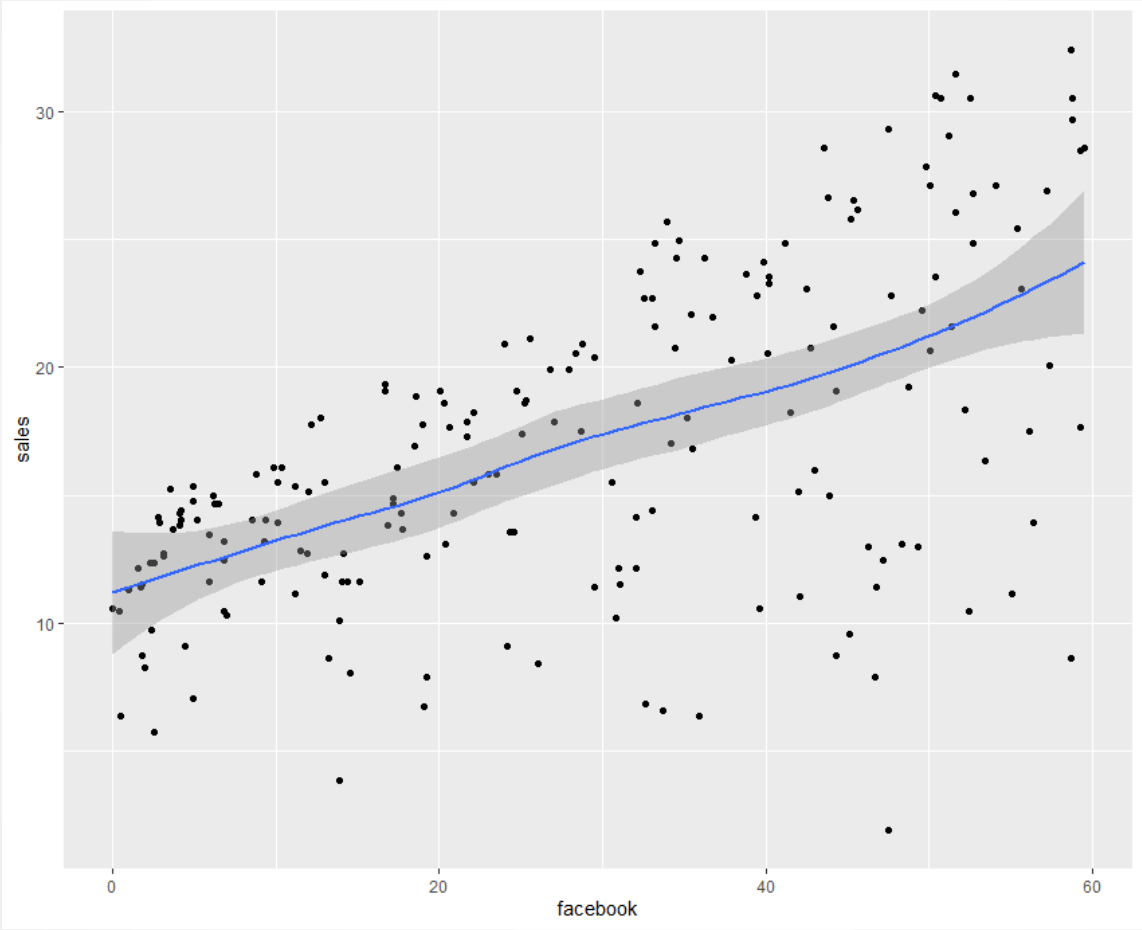
```
> summary(marketing)
    youtube          facebook       newspaper          sales      
 Min.   :  0.84   Min.   : 0.00   Min.   :  0.36   Min.   : 1.92  
 1st Qu.: 89.25   1st Qu.:11.97   1st Qu.: 15.30   1st Qu.:12.45  
 Median :179.70   Median :27.48   Median : 30.90   Median :15.48  
 Mean   :176.45   Mean   :27.92   Mean   : 36.66   Mean   :16.83  
 3rd Qu.:262.59   3rd Qu.:43.83   3rd Qu.: 54.12   3rd Qu.:20.88  
 Max.   :355.68   Max.   :59.52   Max.   :136.80   Max.   :32.40  
>
```

EXPLORATORY DATA ANALYSIS



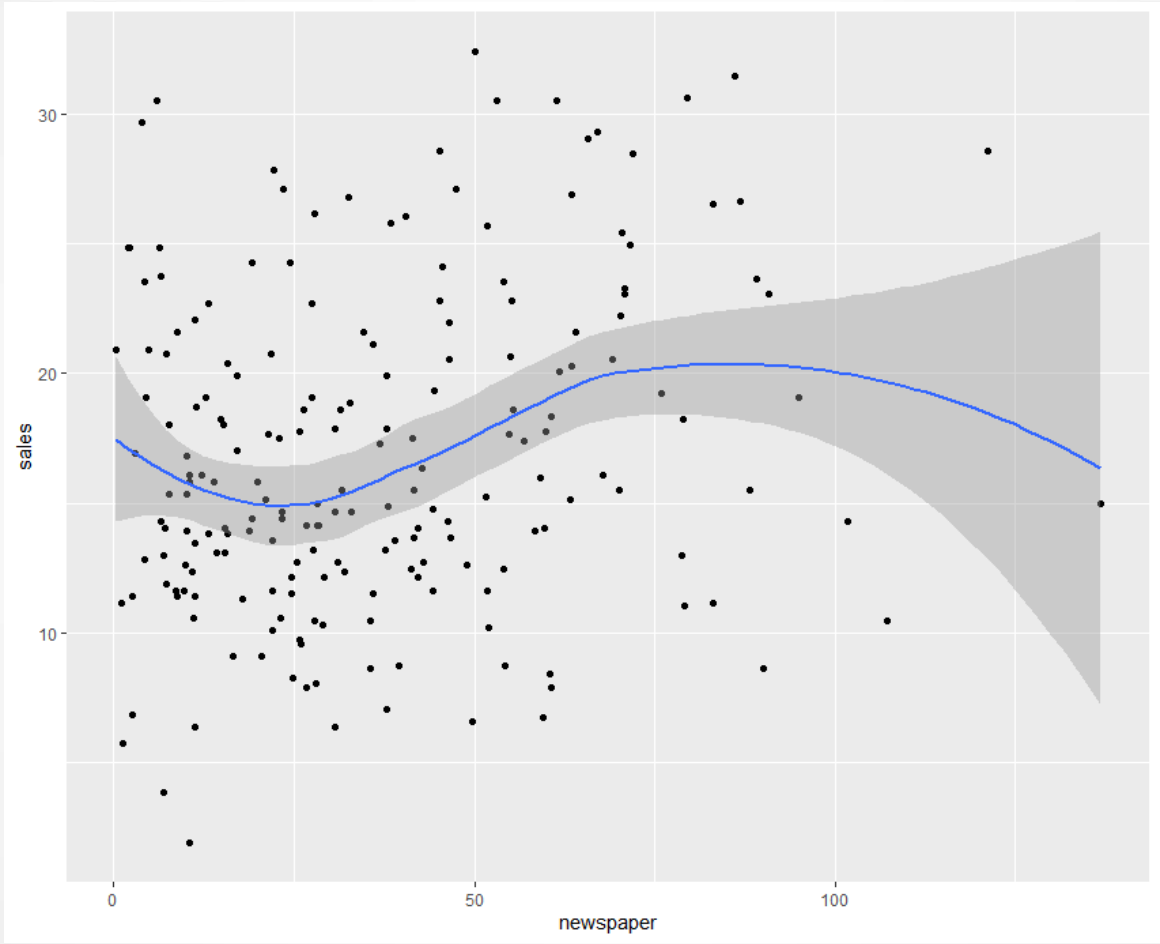
This graph shows that there is a positive relationship between YouTube ads and sales.

Relation between YouTube Ads and Sales



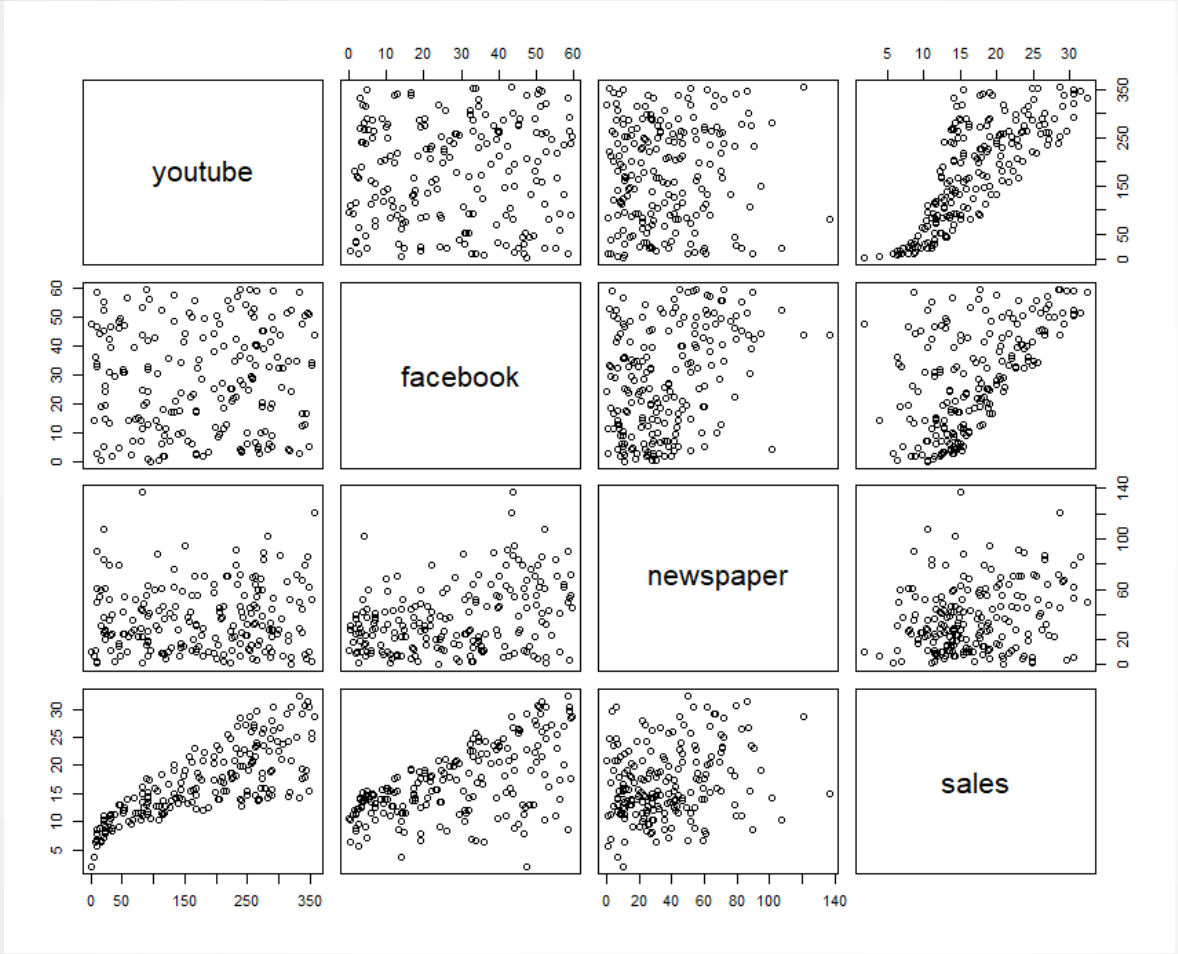
This graph shows that there is a positive relationship between Facebook ads and sales.

Relation between Facebook Ads and Sales



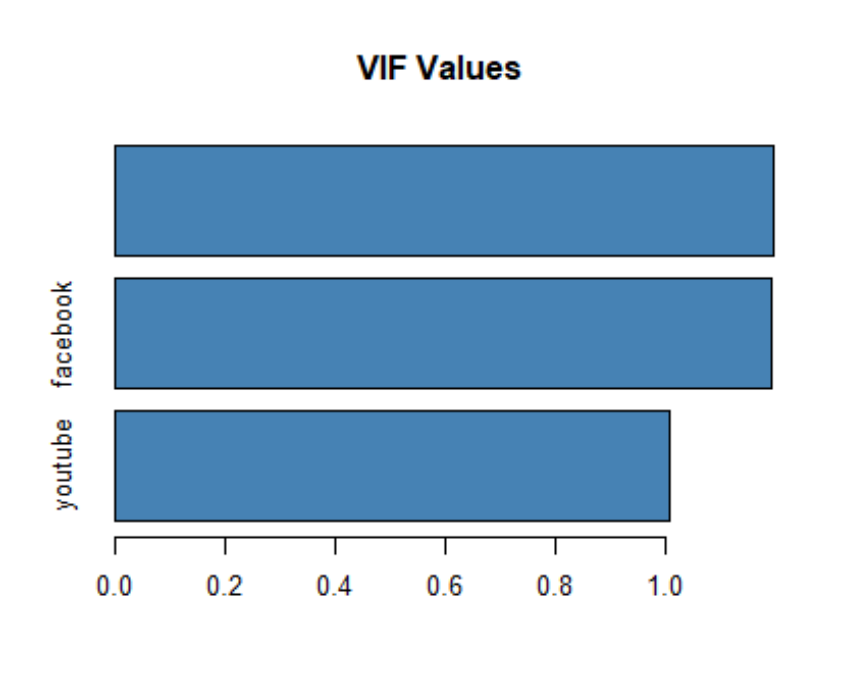
This graph does not show any notable relationship between newspaper and sales.

Relation between Newspaper Ads and Sales



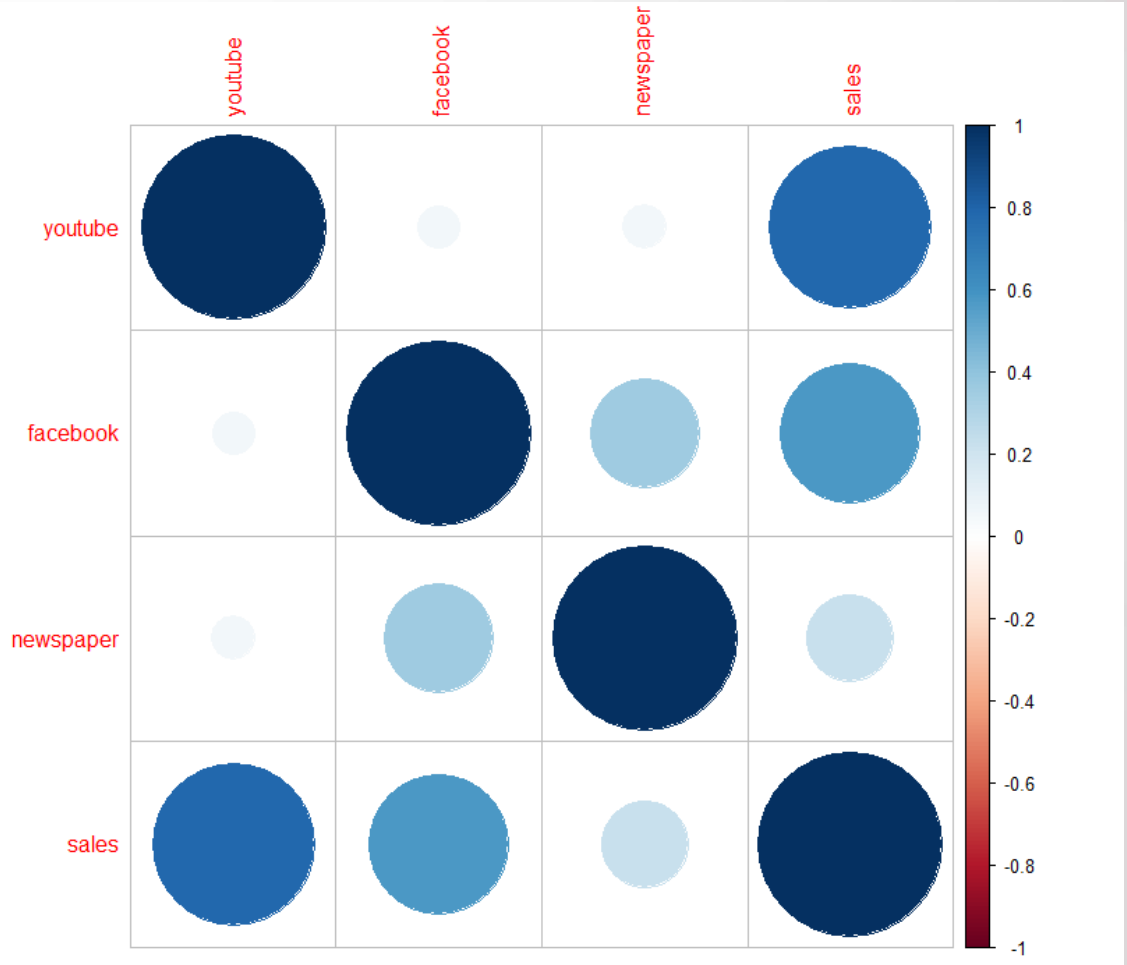
SCATTER PLOT MATRIX

CHECKING MULTI-COLLINEARITY



Vif values are Low

```
> cor(cbind(marketing))
             youtube   facebook  newspaper     sales
youtube   1.00000000 0.05480866 0.05664787 0.7822244
facebook  0.05480866 1.00000000 0.35410375 0.5762226
newspaper 0.05664787 0.35410375 1.00000000 0.2282990
sales     0.78222442 0.57622257 0.22829903 1.0000000
```



Wichita State University

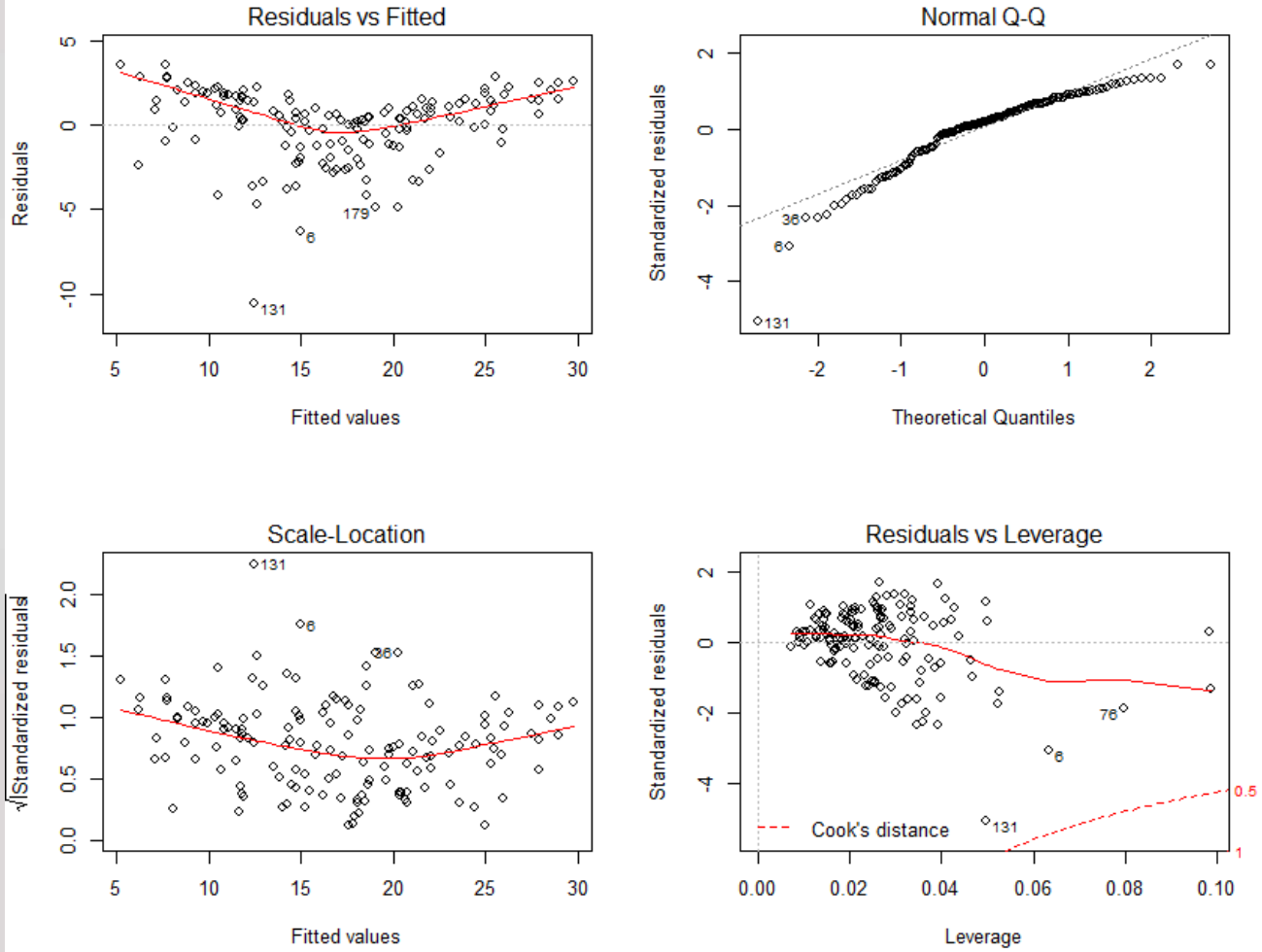
MODEL BUILDING AND EVALUATION

Wichita State University

MODEL1 BUILDING

```
> model<-lm(sales~.,data=training_dataset)
> summary(model)

Call:
lm(formula = sales ~ ., data = training_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5420  -1.0925   0.4029   1.4426   3.5736

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.3289079  0.4658098   7.146 3.93e-11 ***
youtube      0.0458732  0.0017169  26.719  < 2e-16 ***
facebook     0.1915325  0.0104145  18.391  < 2e-16 ***
newspaper   -0.0006754  0.0075475  -0.089    0.929
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.131 on 146 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.8821
F-statistic: 372.7 on 3 and 146 DF,  p-value: < 2.2e-16
```

MODEL1 EVALUATION



Model1 Evaluation using Residual plots:

- Linearity
- Normality
- Homoscedasticity
- Influential obs

Wichita State University

MODEL1 EVALUATION

Actual vs Predicted values

```
> head(actuals_preds)
   actuals predicteds
1    26.52  24.627243
5    15.48  15.716486
9     5.76   4.284171
13   11.04  12.652987
17   15.00  15.380867
21   21.60  21.674614
```

Model Evaluation Metrics

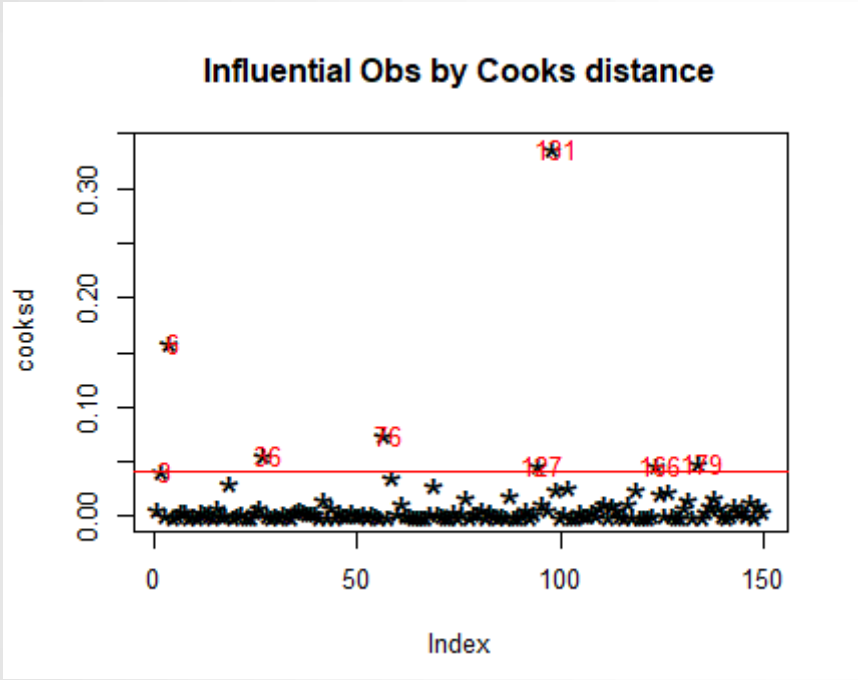
RMSE	MAE	MIN_MAX ACCURACY
1.67	1.33	0.89

Wichita State University

STEPWISE REGRESSION

```
> ols_step_forward_p(model)

                           Selection Summary
--------------------------------------------------------------------------------
        Variable                    Adj.
Step    Entered      R-Square    R-Square      C(p)         AIC         RMSE
--------------------------------------------------------------------------------
   1    youtube        0.5674      0.5645    400.8932    852.6534     4.0958
   2    facebook       0.8845      0.8829      2.0080    656.5646     2.1235
--------------------------------------------------------------------------------

> ols_step_backward_p(model)


                         Elimination Summary
-----------------------------------------------------------------------
        Variable                    Adj.
Step     Removed      R-Square    R-Square      C(p)       AIC         RMSE
-----------------------------------------------------------------------
   1    newspaper       0.8845      0.8829    2.0080    656.5646     2.1235
-----------------------------------------------------------------------

> ols_step_best_subset(model)
         Best Subsets Regression
----------------------------------------
Model Index    Predictors
----------------------------------------
     1         youtube
     2         youtube facebook
     3         youtube facebook newspaper
----------------------------------------

                                                Subsets Regression Summary
-------------------------------------------------------------------------------------------------------------------------------------------
Model    R-Square    Adj.        Pred
                     R-Square    R-Square      C(p)         AIC         SBIC         SBC         MSEP         FPE         HSP         APC
-------------------------------------------------------------------------------------------------------------------------------------------
  1        0.5674      0.5645      0.5547    400.8932    852.6534    423.0156    861.6853    2516.2834    16.9989      0.1141      0.4443
  2        0.8845      0.8829      0.8775      2.0080    656.5646    231.0456    668.6071     676.3916     4.5992      0.0309      0.1202
  3        0.8845      0.8821      0.8752      4.0000    658.5563    233.0925    673.6095     681.0190     4.6608      0.0313      0.1218
-------------------------------------------------------------------------------------------------------------------------------------------
```

OUTLIER DETECTION




```
> car::outlierTest(model)
     rstudent unadjusted p-value Bonferroni p
131 -5.574014         1.1787e-07    1.768e-05
>
```

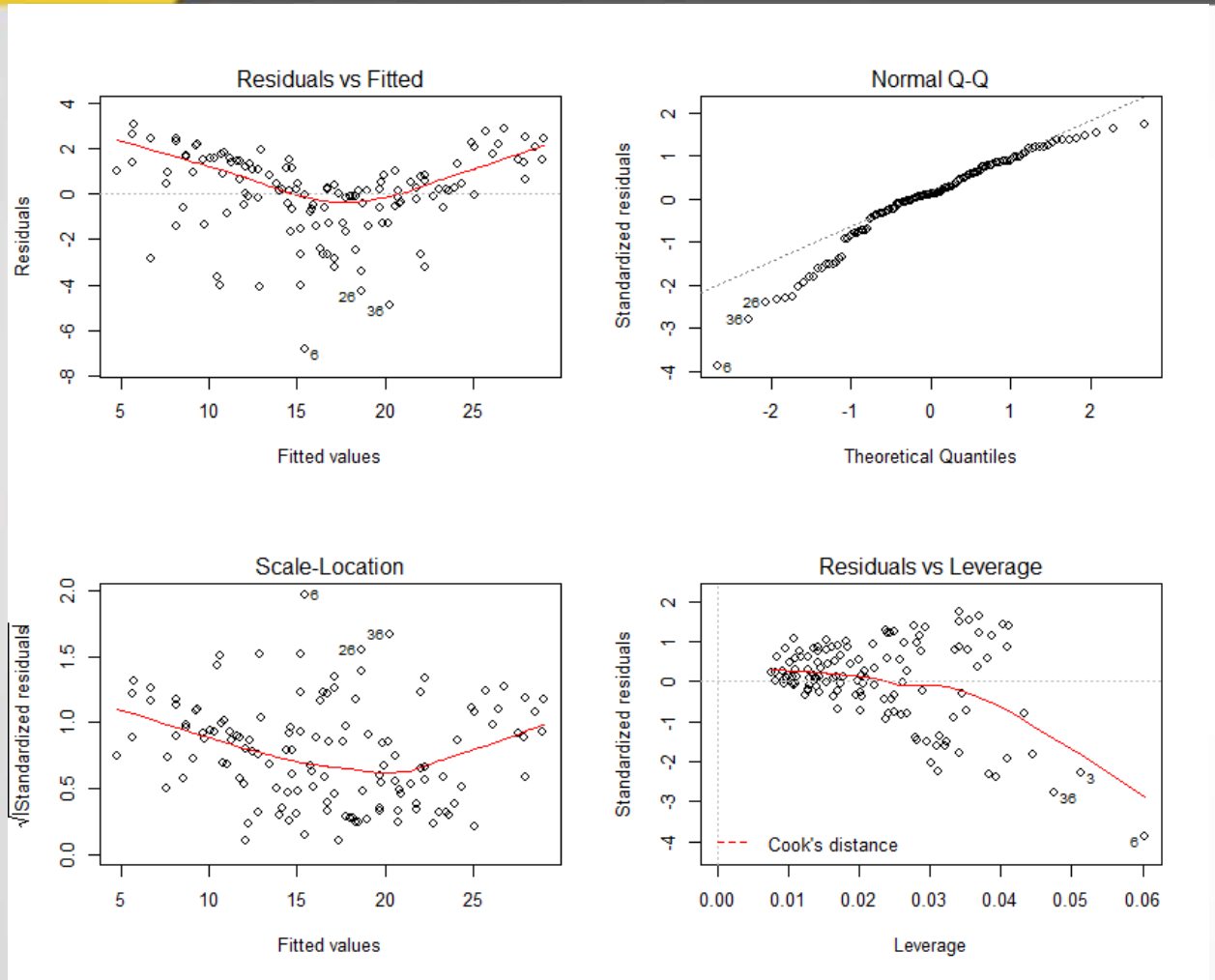
Wichita State University

MODEL2 BUILDING

```
> model2<-lm(sales~.,data=training_dataset)
> summary(model2)

Call:
lm(formula = sales ~ ., data = training_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8065 -0.6449  0.1891  1.3418  3.0624

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.827969   0.376729   10.16   <2e-16 ***
youtube     0.044506   0.001585   28.07   <2e-16 ***
facebook    0.190080   0.009123   20.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.809 on 129 degrees of freedom
Multiple R-squared:  0.9156,    Adjusted R-squared:  0.9143
F-statistic: 699.4 on 2 and 129 DF,  p-value: < 2.2e-16
```

- Adj R-square increased from 0.8821 to 0.9143
- F-statistic increased from 372.7 to 699.4

MODEL2 EVALUATION



Model2 Evaluation using Residual plots:

- Linearity
- Normality
- Homoscedasticity
- Influential obs

MODEL 2

Actual vs Predicted values

```
> head(actuals_preds)
   actuals predicteds
1    26.52   24.73899
4    22.20   21.33951
7    14.16   14.38042
10   12.72   15.09178
13   11.04   13.10522
16   26.88   25.14390
```

Model Evaluation Metrics

RMSE	MAE	MIN_MAX ACCURACY
1.99	1.58	0.89

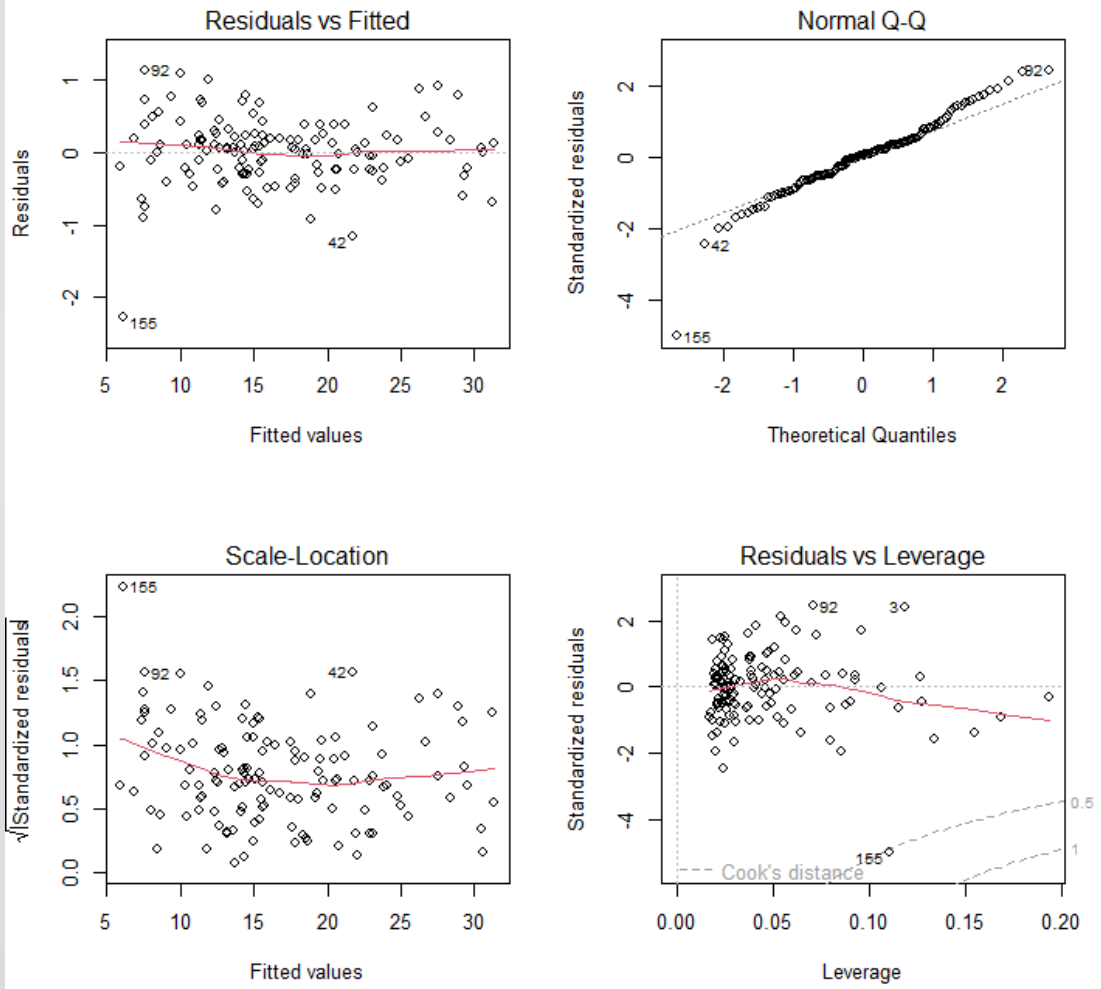
Wichita State University

FINAL MODEL

```
> final_model <- lm(sales ~ facebook + poly(youtube, 3)+ facebook*youtube,
+                 data = training_dataset)
>
> summary(final_model)

Call:
lm(formula = sales ~ facebook + poly(youtube, 3) + facebook *
    youtube, data = training_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-2.26153 -0.25351  0.03006  0.22859  1.12584

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.126e+01  7.687e-02 146.447  < 2e-16 ***
facebook            4.578e-02  4.736e-03   9.665  < 2e-16 ***
poly(youtube, 3)1   2.351e+01  8.958e-01  26.247  < 2e-16 ***
poly(youtube, 3)2  -9.177e+00  4.833e-01 -18.987  < 2e-16 ***
poly(youtube, 3)3   4.246e+00  4.803e-01   8.840 7.08e-15 ***
youtube                    NA         NA      NA       NA
facebook:youtube    8.537e-04  2.334e-05  36.579  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4783 on 126 degrees of freedom
Multiple R-squared:  0.9942,    Adjusted R-squared:  0.994
F-statistic:  4348 on 5 and 126 DF,  p-value: < 2.2e-16
```

- Adj R-square increased from 0.9143 to 0.994
- F-statistic increased from 699.4 to 4348

FINAL MODEL EVALUATION



Final Model Evaluation using Residual plots:

- Linearity
- Normality
- Homoscedasticity
- Influential obs

FINAL MODEL

Actual vs Predicted values

```
> head(actuals_preds)
   actuals predicteds
1    26.52   25.73868
4    22.20   22.25779
7    14.16   13.49771
10   12.72   13.53246
13   11.04   10.08249
16   26.88   26.79786
```

Model Evaluation Metrics

RMSE	MAE	MIN_MAX ACCURACY
0.46	0.37	0.97

Wichita State University

CHECKING OVERFITTING

```
> #view summary of k-fold CV
> print(model)
Linear Regression

199 samples
  2 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 179, 179, 179, 179, 179, 180, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.4664857  0.9947671  0.358404

Tuning parameter 'intercept' was held constant at a value of TRUE
>
```

```
> ldply(list(final_model), model_fit_stats)
  R.squared Adj.R.squared Ratio.Adj.R2.to.R2 Pred.R.squared  PRESS
1     0.994          0.994                  1           0.993 32.959
>
```

Wichita State University

CONCLUSION

MODEL 1

RMSE	MAE	MIN_MAX ACCURACY
1.67	1.33	0.89

MODEL 2

RMSE	MAE	MIN_MAX ACCURACY
1.99	1.58	0.89

FINAL MODEL

RMSE	MAE	MIN_MAX ACCURACY
0.46	0.37	0.97

In conclusion, we can say that the final model is performing much better and satisfies all the assumptions when compared to the other two models.

Therefore, we have successfully built a multiple regression model which can be used to predict the sales using the amount of money spent on the given advertising platforms.

Wichita State University

Wichita State University