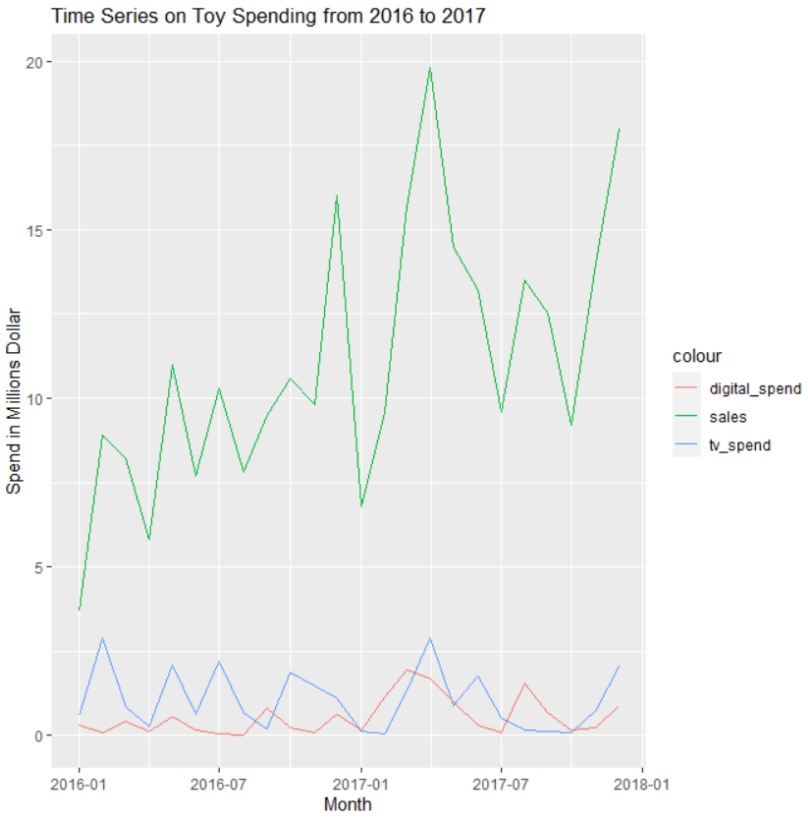
Q2




Q3

| | sales | tv_spend | digital_spend |
|---|---|---|---|
| sales | 1.0000000 | 0.4406862 | 0.6647654 |
| tv_spend | 0.4406862 | 1.0000000 | 0.0720594 |
| digital_spend | 0.6647654 | 0.0720594 | 1.0000000 |

There is a relatively high correlation between **sales** and *digital_spend*, followed by a moderate relationship between **sales** and *tv_spend*. However, the correlation between *tv_spend* and *digital_spend* is not very obvious.

Q 4

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

Residuals:
     Min       1Q   Median       3Q      Max
-2154673  -784491    88873   744202  2098775

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.665e+09  7.557e+09   1.014  0.32393
month         -5.287e+00  5.215e+00  -1.014  0.32414
tv_spend       1.947e+00  3.168e-01   6.145 8.39e-06 ***
digital_spend  2.872e+00  5.153e-01   5.573 2.74e-05 ***
trend          1.419e+07  1.372e+07   1.035  0.31455
xmas           3.097e+06  9.854e+05   3.143  0.00563 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1244000 on 18 degrees of freedom
Multiple R-squared:  0.9206,    Adjusted R-squared:  0.8985
F-statistic: 41.74 on 5 and 18 DF,  p-value: 2.771e-09
```

a) Adjusted R-squired:
The adjusted R-squared of this regression model is around at 0.90 with a well linear fit of the predictor on **sales**. It shows that 89.5% of the variances in **sales** can be interpreted by the independent variables.

b) P-Value:
A strong relationship between the dependent variable **sales** and *tv-spend* and *digital_spend* indicates the statistical significance with both $p < .001$.

*xmas*, on the other hand, is at a relative statistical significance of $p < .05$.
The remaining independent variables, including *month* and *trend*, show the non-significance, and it is likely the null hypothesis is true with no relationship between the **sales** predictor.

Q5

```
> contribution_percentage_tv
[1] 90.41275
```

The contribution margin ratio from TV spend to sales is 90.4%, with the contribution in absolute dollar value of 240M. It shows that the 25M spending in the tv area brings the extra 240M benefits in the total final sales.

```
> contribution_abs_tv
[1] 240227126
> sum(toyData$tv_spend)
[1] 25473374
```

Q6

```
> roi
[1] 943.0519
```

The tv return on investment is at 943.1% with a high return rate.

Q 7

```
> pred_the_next_year
       1        2        3
11750228 12760264 15656009
```

The linear predictions of the beginning three months sales amount are as shown in the able.

Q8

1. Fix (promotion) cost

Introducing the fixed cost in the promotion campaign activities will give a better insight into the sales margin and return. Further, it will also help better understand the single variable's contribution towards the final sales amount.

2. Full variables data for the initial data cleaning

With all the variables settled in the initial data cleaning, it is easier to analyze and determine whether there exists multicollinearity between independent variables on the feature selection to improve the model's performance.

3. Past few years historical data

A more historical dataset means it will be easier and quicker to evaluate the residual, bias and other basic features. By filtering out the biased data, the performance and accuracy of the model will enhance significantly.