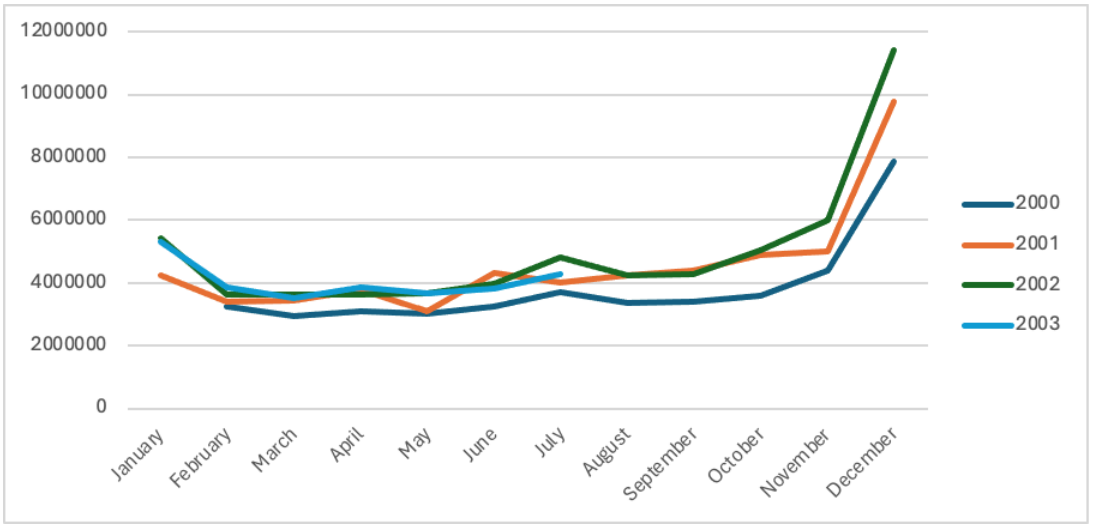
Marketing Analytics Homework #2 - Zihao, Yiting

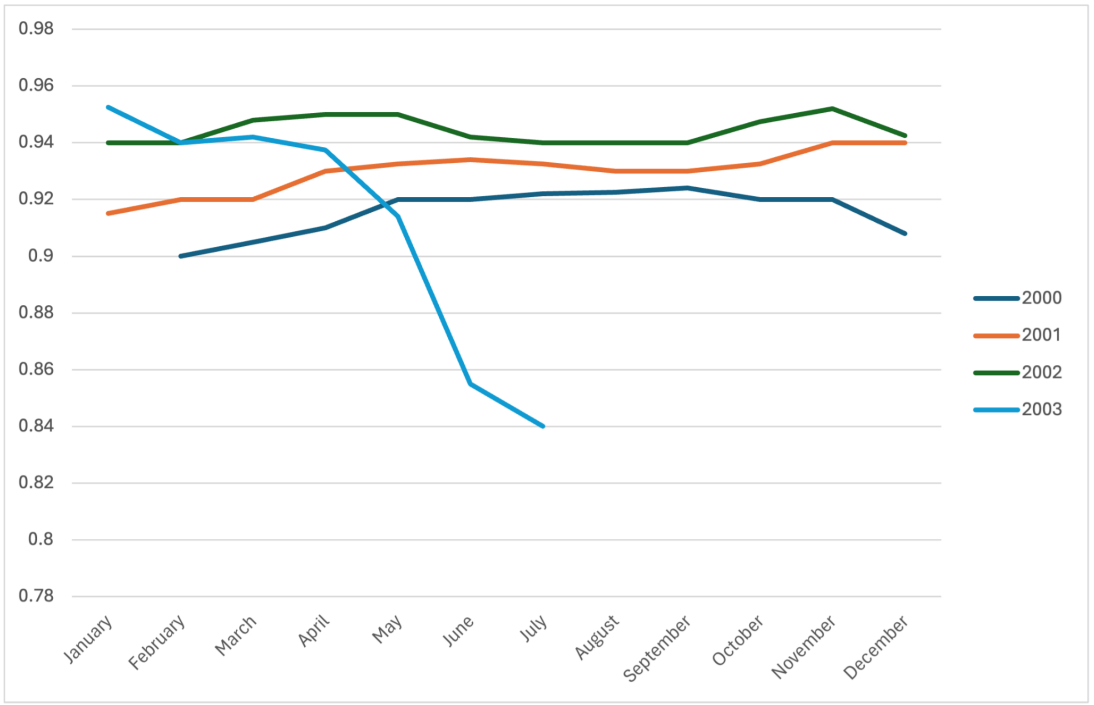
1. Univariate Analysis:

	eq_volum	disacv_c	bonusacv	price_c	price_e	price_p	tvgrp_c	tvgrp_u	trustad	fsi_holi	fsi_non	fsi_comp	itemstor	walmart
MEAN	4430421.47	15.14	33.07	0.93	1.02	0.71	48.14	36.62	0.26	1014.83	3216.59	7239.80	9.24	0.32
SD	1982087.85	6.13	22.36	0.02	0.05	0.04	82.01	81.79	0.44	6180.89	10883.52	17776.87	0.38	0.06
RANGE	12670855.00	27.87	69.00	0.12	0.42	0.25	346.00	398.00	1.00	41590.00	41676.00	92896.00	1.50	0.23

Eq_Volume time series



Price_C



Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
price_c	179	0.93	0.022	0.84	0.92	0.94	0.96
price_e	179	1	0.054	0.7	0.99	1.1	1.1
price_p	179	0.71	0.038	0.54	0.69	0.74	0.79

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
tvgrp_c	179	48	82	0	0	84	346
tvgrp_u	179	37	82	0	0	2.5	398

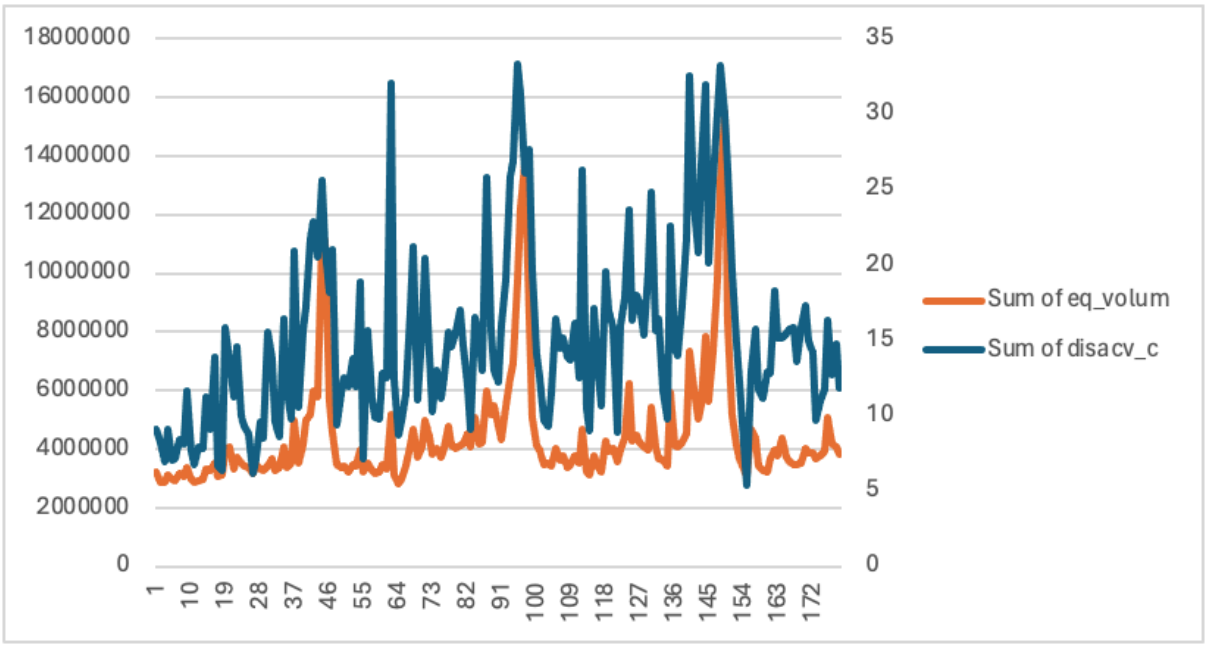
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
fsi_comp	179	7240	17777	0	0	0	92896
fsi_holi	179	1015	6181	0	0	0	41590
fsi_non	179	3217	10884	0	0	0	41676

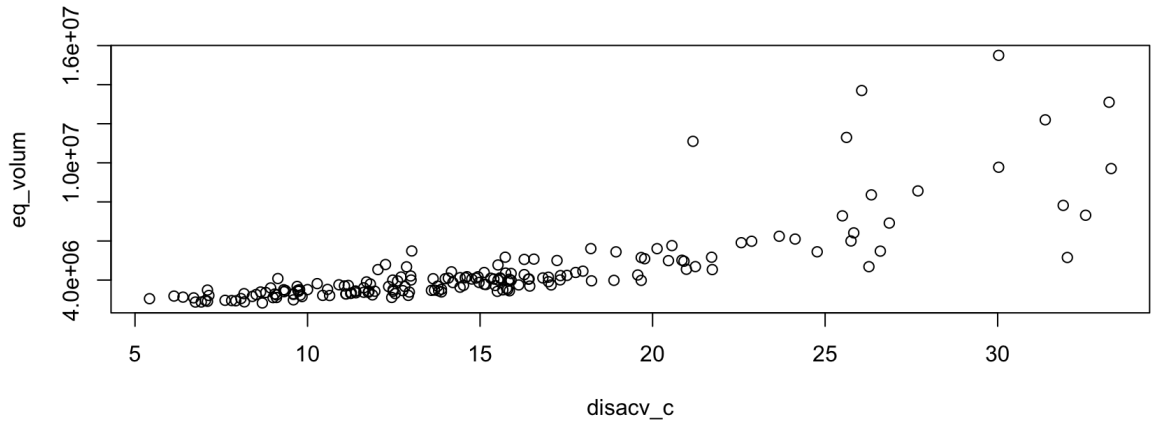
1. Selling equivalent volume will have seasonal change for each year, especially will have a huge amount of increase around December.
2. Prices for product C also have seasonal changes that have decreased in the price around December.
3. To compare the price part with other competitors, the price of brand c obtains relatively small standard deviation within three brands, which indicates that there's relatively low price fluctuation on brand C.
4. In the TV advertising, the impact on brand C is much higher than the competitors on average.
5. For the coupon part, the FSIs of brand C are relatively lower than competitors both in holiday and non-holiday.

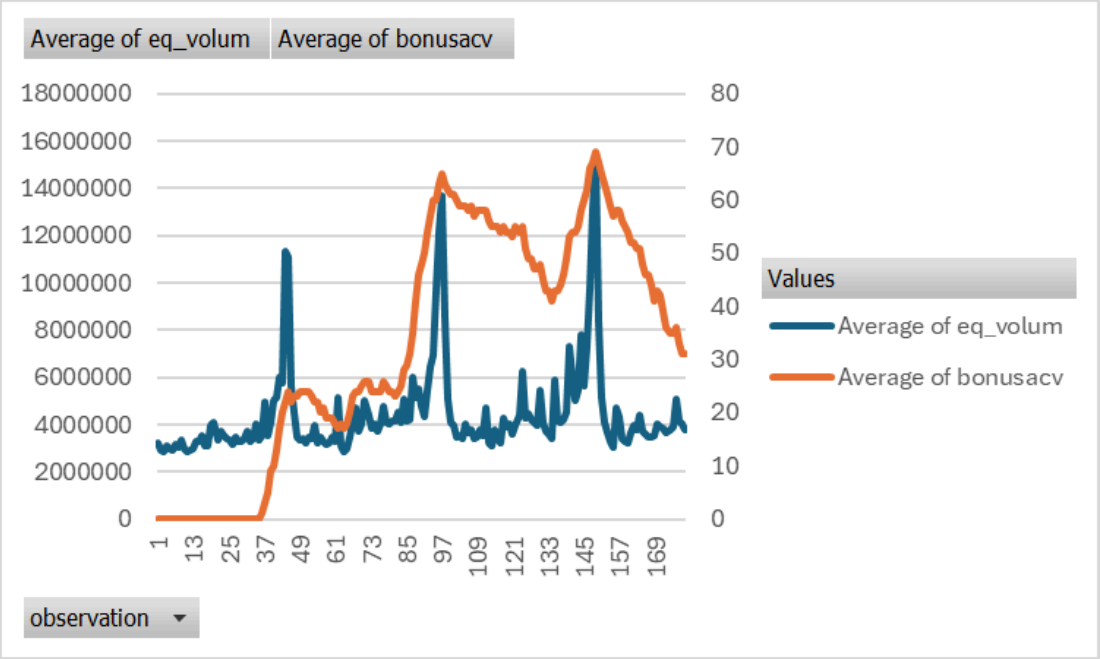
2. Bivariate Analysis:

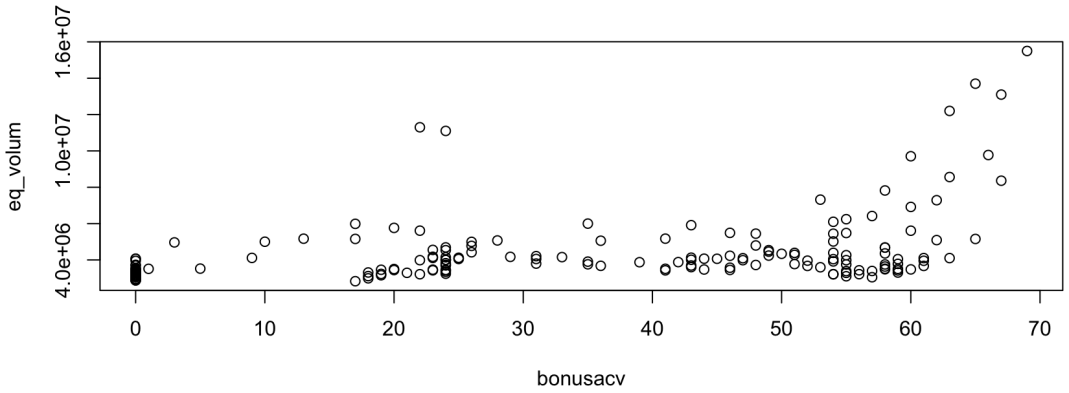
	eq_volum
eq_volum	1
disacv_c	0.785449
tvgrp_c	0.477333
bonusacv	0.391879
fsi_holi	0.359082
trustad	0.160452
price_p	0.157502
walmart	0.152933
price_c	0.126065
fsi_comp	0.100522
fsi_non	0.072088
tvgrp_u	0.042259
itemstor	0.022647
price_e	-0.044737

```
> cor(price_c, eq_volum)
[1] 0.1260649
> cor(price_p, eq_volum)
[1] 0.1575022
> cor(price_e, eq_volum)
[1] -0.04473694
```









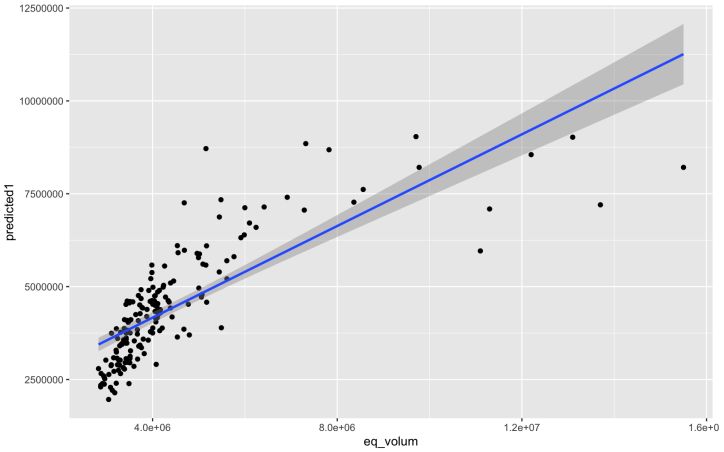
1. When the increasing in the disacv_c or bonusacv which in effect eq_volum causes increasing.
2. Disacv_c has the highest correlation the eq_volum and also higher than 0.75 to show a strong relationship.

Model Explenation:

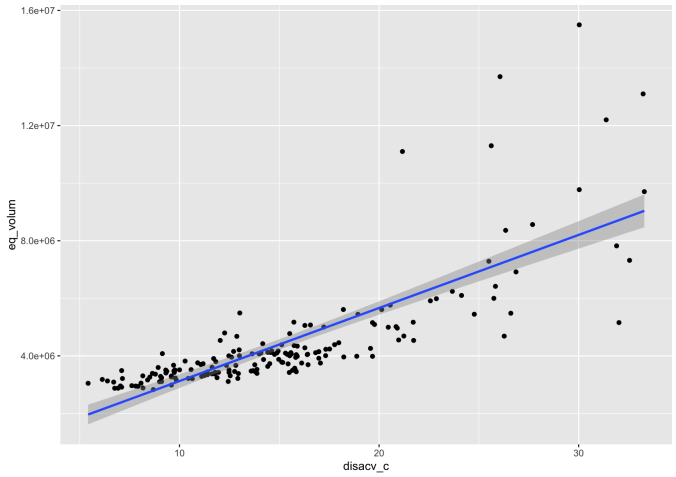
In the simplest model, I chose disacv_c as the independent variable (x-axis) because it exhibits the highest correlation with unit sales volume. We could see the adjusted R-squared is 0.6.

```
Call:
lm(formula = eq_volum ~ disacv_c)

Residuals:
     Min       1Q   Median       3Q      Max
-3558591  -635629  -119297   386524  7289804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   586756     245525    2.39   0.0179 *
disacv_c      253861      15036   16.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1230000 on 177 degrees of freedom
Multiple R-squared:  0.6169,    Adjusted R-squared:  0.6148
F-statistic: 285.1 on 1 and 177 DF,  p-value: < 2.2e-16

> round(summary(mod1)$coeff,3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 586756.0  245524.95   2.390    0.018
disacv_c    253860.8   15035.92  16.884    0.000
```

After evaluating the model's fit to the data, it appears that the model did not fit well. Therefore, our next step is to assess the linear regression assumptions.



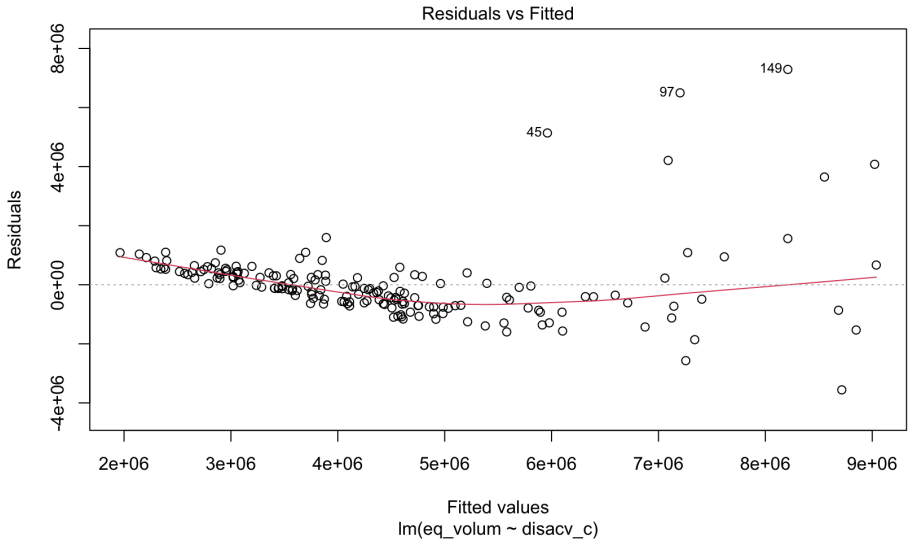
Assumption 1 - Check the linearity



We've noticed two main issues: first, the data points don't align well with the regression line, despite a slight trend. Second, there's considerable dispersion among data points, especially for disacv_c values exceeding 20.

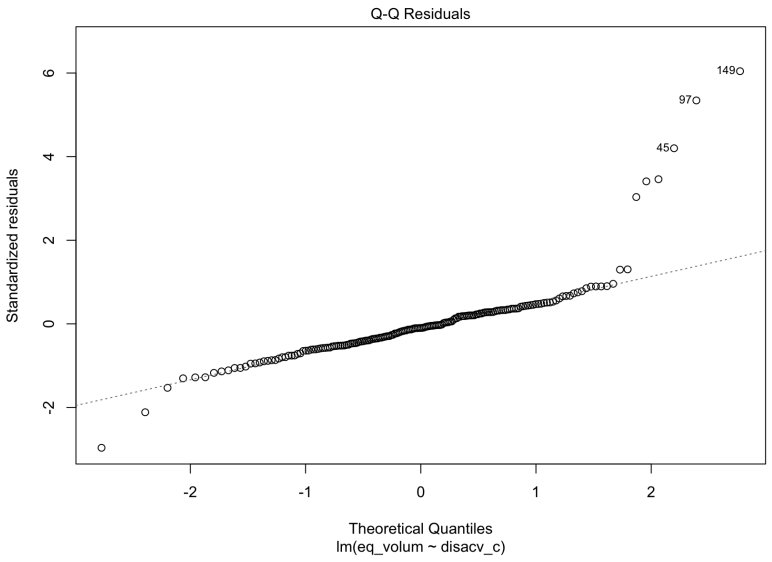
Assumption 2 - Check the residuals

```
> head(eq_volum)
[1] 3240042 2885233 2877506 3107180 2954494 2913908
> head(predicted1)
      1       2       3       4       5       6
2896889 2660799 2343473 2896889 2373936 2389168
> head (residuals1)
       1        2        3        4        5        6
343152.6 224434.1 534033.1 210290.6 580557.8 524740.2
```

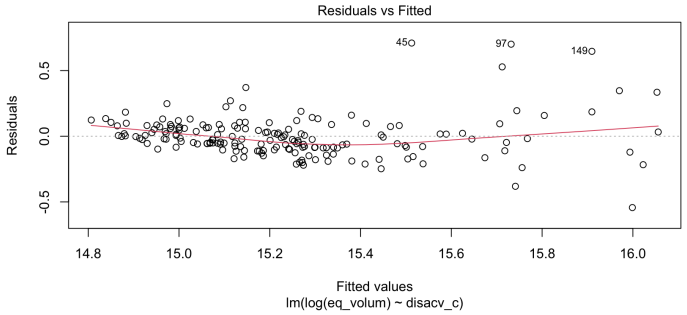


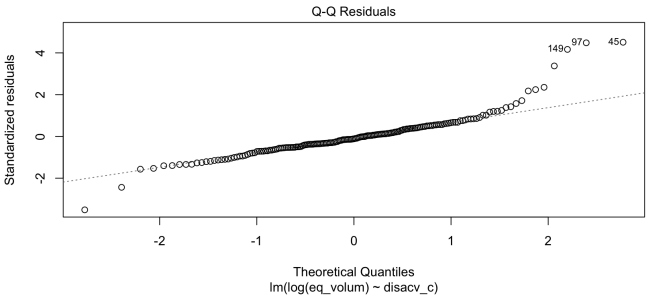
The residuals didn’t align with the horizontal dotted line, which needed to be dealt with further.

Assumption 3 - Check the residuals normal distribution.



The left and most right parts are not aligned with the dotted line. Thus, I conducted log transformation in the eq_volum column.

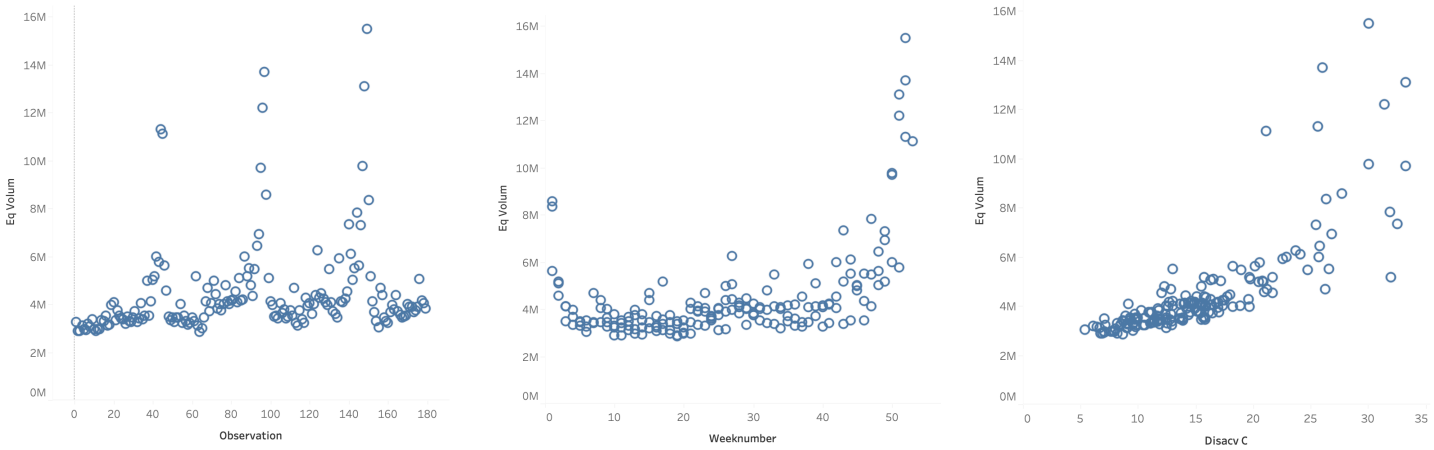




I expanded mod1 to mod2 by adding bonusacv, tvgrp, itemstor, and prices, which show a strong correlation with eq_volum. However, I omitted the price of brand E due to its negative correlation with sales volume, which could distort the model's interpretation. As we could see, the **Adjusted R-squared jumped from 0.6 to 0.86** after adding these variables.

```
Call:
lm(formula = log(eq_volum) ~ disacv_c + +bonusacv + tvgrp_c +
    tvgrp_u + itemstor + price_c + price_p)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25732 -0.07336 -0.01710  0.05575  0.52105

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.9247209  0.5532593  23.361  < 2e-16 ***
disacv_c     0.0316611  0.0019683  16.086  < 2e-16 ***
bonusacv     0.0080536  0.0009286   8.672 3.17e-15 ***
tvgrp_c      0.0006804  0.0001223   5.562 1.01e-07 ***
tvgrp_u      0.0005784  0.0001301   4.446 1.57e-05 ***
itemstor     0.3388989  0.0475757   7.123 2.81e-11 ***
price_c     -2.1542653  0.4547245  -4.738 4.53e-06 ***
price_p      0.5407531  0.2684401   2.014   0.0455 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1162 on 171 degrees of freedom
Multiple R-squared:  0.8709,    Adjusted R-squared:  0.8656
F-statistic: 164.8 on 7 and 171 DF,  p-value: < 2.2e-16
```

For the seasonality factor, there’s few observations.



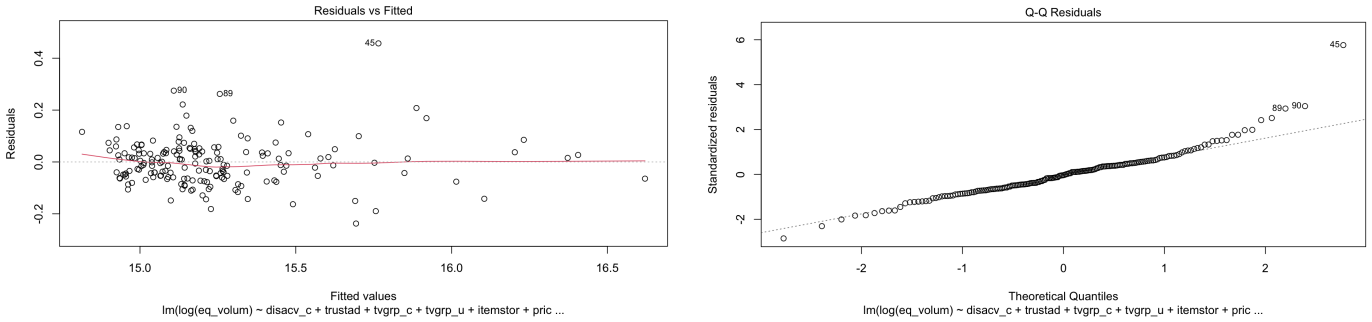
Based on the three graphs above, it's evident that there are extreme peaks in unit sales volume under certain circumstances. To address this issue, I incorporated dummy variables based on these findings.

```r
mod4 <- lm(log(eq_volum) ~ disacv_c + trustad + bonusacv + tvgrp_c + tvgrp_u + itemstor + price_c + price_p + disacv_c_peak, data = dat)
summary(mod4) #0.876

mod5 <- lm(log(eq_volum) ~ disacv_c + trustad +bonusacv + tvgrp_c + tvgrp_u + itemstor + price_c + price_p + disacv_c_peak + bonusacv_peak, data = dat)
summary(mod4) #0.876

mod6 <- lm(log(eq_volum) ~ disacv_c  + trustad +bonusacv + tvgrp_c + tvgrp_u + itemstor + price_c + price_p + disacv_c_peak + bonusacv_peak + weeknum_peak, data = dat)
summary(mod6) #0.93

mod7 <- lm(log(eq_volum) ~ disacv_c  + trustad +bonusacv + tvgrp_c + tvgrp_u + itemstor + price_c + price_p + disacv_c_peak + bonusacv_peak + weeknum_peak + observation_peak, data = dat)
summary(mod7) # 0.94
```

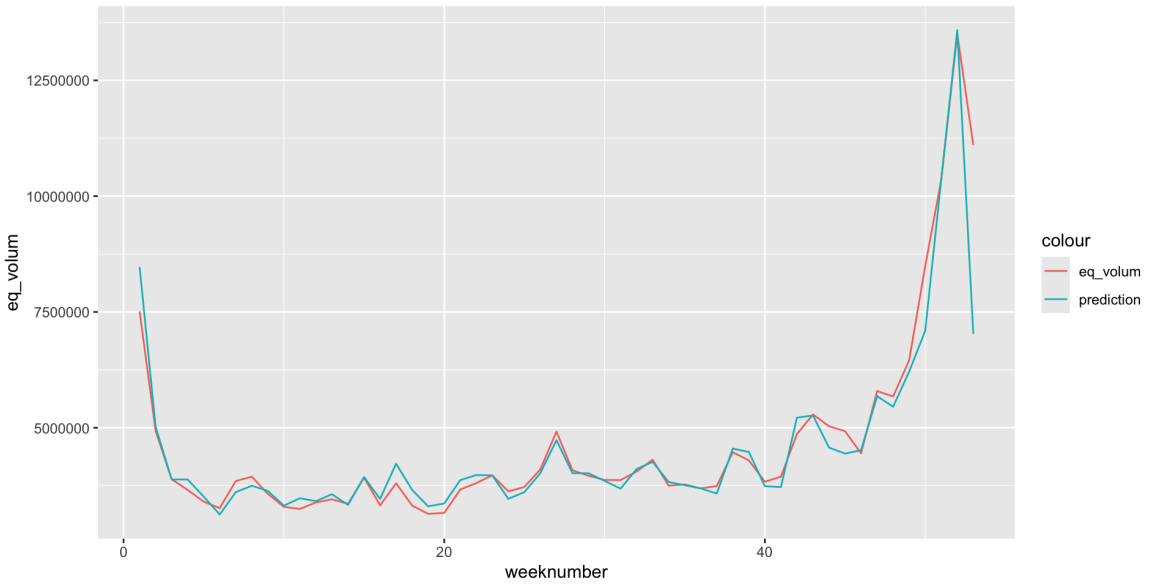
Adding more dummy variables to the model has led to an increase in the adjusted R-squared, indicating that these seasonalities indeed impact the unit sales volume for brand C. However, we've encountered a multicollinearity issue with the bonusacv variable. Furthermore, we also omitted the variables with P-value over 0.05.

```
> vif(mod7)
        disacv_c         bonusacv          tvgrp_c          tvgrp_u         itemstor
        3.261691         6.916245         1.377382         1.653120         4.764370
         price_c          price_p    disacv_c_peak    bonusacv_peak    weeknum_peak
        1.425292         1.353855         2.812142         1.753652         1.953168
observation_peak
        1.502581
```

Now we could check the residuals and normal distribution of model 8.



Model 8 has shown significant improvement with adjusted R-squared of 0.91 over the simplest model. Now, let's compare the predicted sales volume with the true sales volume. Since we applied log transformation, we'll use the exp function to retrieve the true data.




```
Residuals:
     Min       1Q   Median       3Q      Max
-0.19766 -0.05865 -0.00797  0.05296  0.45254

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.363e+01  2.120e-01  64.286  < 2e-16 ***
disacv_c         2.964e-02  1.463e-03  20.260  < 2e-16 ***
tvgrp_c          5.399e-04  9.063e-05   5.958 1.46e-08 ***
itemstor         7.477e-02  2.114e-02   3.537 0.000523 ***
price_p          5.408e-01  1.970e-01   2.745 0.006706 **
bonusacv_peak    9.201e-02  2.609e-02   3.527 0.000542 ***
weeknum_peak     3.945e-01  3.961e-02   9.960  < 2e-16 ***
observation_peak 2.827e-01  6.171e-02   4.581 9.00e-06 ***
trustad          5.622e-02  1.785e-02   3.150 0.001930 **
fsi_non          1.850e-06  6.326e-07   2.925 0.003926 **
fsi_holi         3.695e-06  1.214e-06   3.043 0.002721 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08805 on 168 degrees of freedom
Multiple R-squared:  0.9272,    Adjusted R-squared:  0.9228
F-statistic: 213.9 on 10 and 168 DF,  p-value: < 2.2e-16
```

We created a new binary variable which can help us to change the numeric data into factor data which will be more straightforward for the model.

Based on the final model, we can see trustad(advertising) is 5.6e-02, disacv_c(discounts) is 3.0e-02, fsi_non(coupon) is 1.9e-06 and fsi_holi(coupon) is 3.7e-06. Based on these, we can see advertising works best and then discounts and then coupons. I feel that one of the main reasons for this is because advertising will show to more people and also message trust will cause people to remember and will try to constantly buy Brand C. And for the discounts, because people are already in the shop and trying to find the most competitive brand to purchase, a discount will let the Brand C be more competitive and cause people to purchase it. And for the coupon, people might get it through mail or flyer, which means at that moment, people might be interested in the context but not willing to buy it, and they also need to save that coupon to be able get a discount in the shop. All these processes might cause people to be less willing or might forget to bring or use coupons which all will cause coupons to have a less effect on the Brand C selling amount.

Based on the final model, I still will use all these three ways but will have different ways to use them, each tailored for optimal impact. Advertising, especially focusing on trust, will be our forefront strategy. By harnessing the power of social media and television, we aim to not only expand our reach but also cement a relationship of trust with our customers. This approach is rooted in the model's indication that trust-themed advertising significantly boosts sales. When it comes to discounts, our approach will be dynamic, balancing cost and profit to offer compelling value to customers. Initially, more generous discounts might be used to attract customers, especially in tandem with our trust-building advertising efforts. Over time, as customer loyalty solidifies, we plan to gradually scale back on discounts without compromising customer satisfaction. For coupons, our strategy will become more targeted. By analyzing data to identify demographic clusters most responsive to coupons, we will refine our distribution to maximize redemption rates. Recognizing the potential for market saturation, we're also exploring innovative coupon strategies, such as offering customers a choice between discounts or tangible rewards like small toys. This flexibility could enhance the perceived value of our coupons.

In terms of budget allocation, there will be a deliberate shift towards increased spending on advertising and selective discounts. The investment in coupons will be more strategic, focusing on targeted distribution and innovative offerings to engage specific customer segments. This refined approach reflects a deep understanding of our model's insights, enabling us to make informed decisions that optimize Brand C's marketing mix for sales growth and customer loyalty.

Illustration：Company realizes advertising and discounts will affect selling the most and they want to do less couponing. But they wanna see how it will affect sales if there is an increase in advertising and discounts and a decrease in coupons.

For this question, we need to build a formula for that.

eq_volum = 13.63+0.05622X+0.02964Y+0.00000185Z1+0.000003695Z2.

through this formula we can see coupon really affect so tiny for the overall selling, if company can transfer the spend from coupon the either advertising or discount, it might have positive effect, of course, we need to calculate the rate of cost between those three methods to make sure it is positive which means in the same amount of money, it can affect more in advertising of discounts compared to coupon.