Background description

- The manufacturer of a consumer goods brand would like to know how much extra sales a brand gained that can be related to the marketing activities.
- The sales model with three marketing variations TV, online banner, promotion.
- Except for marketing variables, there are five non-marketing variables: price, time, product, region, month.
- Dataset: JellyBeans_3

Variation	Dummy variable
time	year1, year2, year3
region	North, South, West, East, Capital
month	Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec

Method and challenge

Method

1. Linear regression – to build models.
2. Dummy variables – to check which variables influence sales a lot.
3. Forward, backward, and both stepwise regression – to find the significant variables.
4. Cross-Validation – to check which variables can create a model with the smallest root mean squared error.
5. Assumption test – to test whether applying the test dataset to the model can get the lowest mean squared error.

Challenge

1. Small dataset – The dataset may be too small and cause under-fitted.
2. Variable selection difficulties – Some unobserved variables may affect sales but did not be considered.
3. Linear model limitations – The model may not have a linear relationship between x and y.
4. Multicollinearity – Some variables in the model may have multicollinearity.

Building the initial model

- Multiple regression model:
  sales = TV + banner + promotion + price + time + region + month + random error

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9721.628    328.289  29.613  < 2e-16 ***
price      -1224.724    143.625  -8.527 1.14e-14 ***
TV             8.455      1.477   5.726 5.04e-08 ***
banner        27.443      5.453   5.032 1.31e-06 ***
prom          46.158      1.217  37.937  < 2e-16 ***
North1     -3939.071    108.902 -36.171  < 2e-16 ***
South1     -2364.397    108.914 -21.709  < 2e-16 ***
West1      -4652.894    108.876 -42.736  < 2e-16 ***
East1      -4789.536    109.064 -43.915  < 2e-16 ***
Jan1         -147.327    176.832  -0.833 0.406021
Feb1          178.221    182.873   0.975 0.331267
Mar1          474.928    189.862   2.501 0.013387 *
Apr1         4534.576    175.229  25.878  < 2e-16 ***
May1         1307.823    176.302   7.418 6.85e-12 ***
Jun1          687.337    173.800   3.955 0.000115 ***
Jul1          308.750    170.691   1.809 0.072379 .
Aug1         -120.249    172.905  -0.695 0.487789
Sep1          167.875    174.749   0.961 0.338189
Oct1          151.849    181.016   0.839 0.402812
Nov1          216.622    173.827   1.246 0.214538
year11       -371.932     86.387  -4.305 2.91e-05 ***
year21         -1.003     84.868  -0.012 0.990585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 461.9 on 158 degrees of freedom
Multiple R-squared:  0.9713,    Adjusted R-squared:  0.9675
F-statistic: 254.7 on 21 and 158 DF,  p-value: < 2.2e-16
```

If there is no marketing activity, sales will be 9721.628 packs.

Increasing TV by 1 GRP, increase sales by 8.455 packs.

Increasing banner by 1 GRP, increase sales by 27.443 packs.

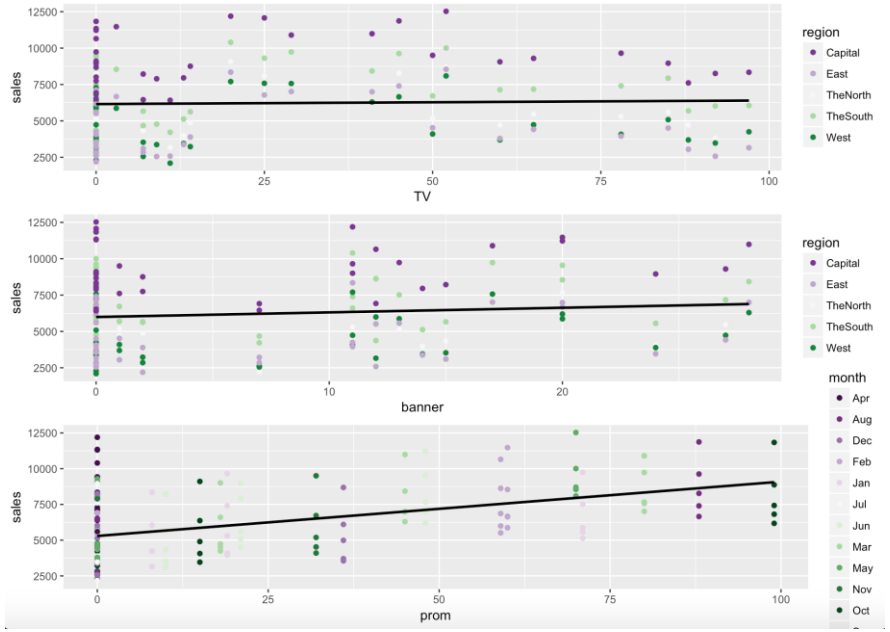
Increasing promotion by 1 GRP, increase sales by 46.158 packs.

high percentage of explained variance

- Correlation between sales and marketing variables

Correlation	TV	Online banner	Promotion
sales	0.0304	0.1085	0.4784

- The scatter plots present that sales does not have a significant relationship with TV, online banner, and promotion.



Finding the optimal model and the association

- Apply the variable selection process on the initial model to decide which variables are important in explaining variables in sales.
- Build the model with selected variables:
  sales = prom + Apr + West + East + North + South + May + banner + price + TV + year1 + Jun + Mar + Jan + Aug + random error

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9927.649    293.388  33.838  < 2e-16 ***
prom           45.965      1.135  40.499  < 2e-16 ***
Apr1         4359.263    132.087  33.003  < 2e-16 ***
East1       -4790.292    108.219 -44.265  < 2e-16 ***
West1       -4653.028    108.041 -43.067  < 2e-16 ***
North1      -3939.382    108.066 -36.454  < 2e-16 ***
South1      -2364.759    108.077 -21.880  < 2e-16 ***
May1         1139.807    143.564   7.939 3.04e-13 ***
price       -1241.425    139.520  -8.898 1.01e-15 ***
banner         28.066      4.718   5.948 1.59e-08 ***
TV              8.354      1.368   6.106 7.16e-09 ***
year11       -372.427     75.370  -4.941 1.90e-06 ***
Jun1          514.529    134.261   3.832 0.000181 ***
Mar1          301.531    148.152   2.035 0.043432 *
Aug1          -286.593   129.001  -2.222 0.027677 *
Jan1          -311.829   143.869  -2.167 0.031643 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 458.4 on 164 degrees of freedom
Multiple R-squared:  0.9707,    Adjusted R-squared:  0.968
F-statistic: 361.8 on 15 and 164 DF,  p-value: < 2.2e-16
```

Increasing promotion by 1 GRP, increase sales by **45.965** packs.

All regional variables are significant variables.

Increasing online banner by 1 GRP, increase sales by **28.066** packs.

Increasing TV by 1 GRP, increase sales by **8.354** packs.

Some are not very significant.

Goodness of fit increased by an insignificant level. (initial Adjusted R-squared is 0.9675)

- How many packs did we sell associated to the advertisements and promotions? And by type of marketing activity?

(total GRP)

```
tv_grp = sum(df$TV) #4455
banner_grp = sum(df$banner) #1275
prom_grp = sum(df$prom) #4380
```

TV: $4455 \times 8.354 = 37216.49$ packs

Online banner: $1275 \times 28.066 = 35783.99$ packs

Promotion: $4380 \times 45.965 = 201328.1$ packs

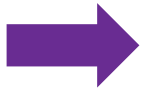
The number of packs associated with ads and promotion
37216.49 + 35783.99 + 201328.1 = **274328.6 packs**

Marketing efficiency and non-marketing sources

- Our TV ads cost us 2 million Pounds, our Banners 500,000 Pounds. Which one is more efficient?

TV advertisement
2000000/ 37216.49 = 53.73962 TV cost for increase sales by 1
37216.49/ 2000000 = 0.018608 sales gained from investing 1 unit in TV

Online banner
500000/ 35783.99 = 13.97273 online banner cost for increase sales by 1
35783.99/ 500000 = 0.071568 sales gained from investing 1 unit in online banner



Online Banner
is more efficient

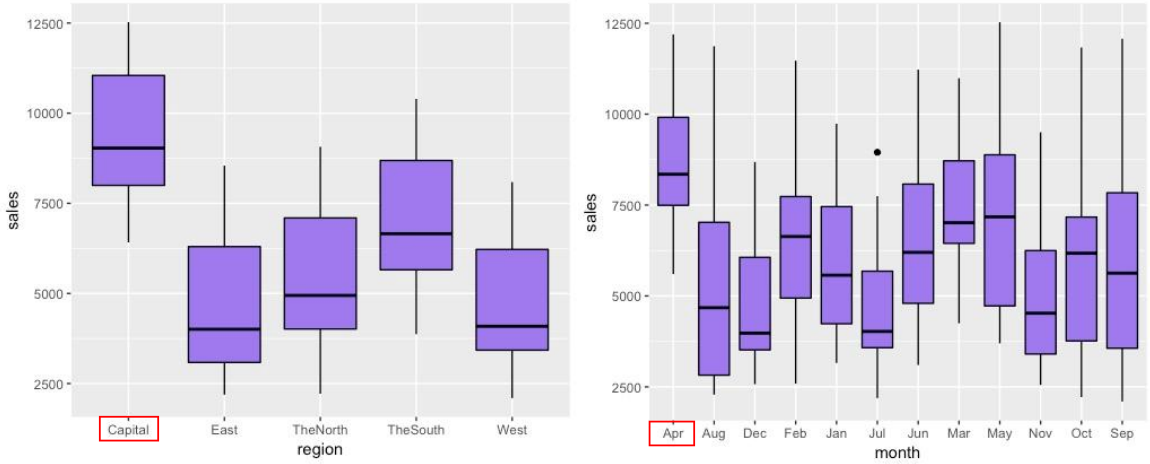
- Can you explain to possible sources of the variation, other than our marketing activities?

```
> summary(aov(sales ~ month, data = df))
             Df    Sum Sq  Mean Sq F value   Pr(>F)
month        11 219515379 19955944    3.51 0.000192 ***
Residuals   168 955106771  5685159

> summary(aov(sales ~ region, data = df))
             Df    Sum Sq   Mean Sq F value Pr(>F)
region        4 572575783 143143946   41.61 <2e-16 ***
Residuals   175 602046367   3440265

> summary(aov(sales ~ time, data = df))
             Df    Sum Sq Mean Sq F value Pr(>F)
time          1 1.847e+04   18470   0.003  0.958
Residuals   178 1.175e+09 6598897
```

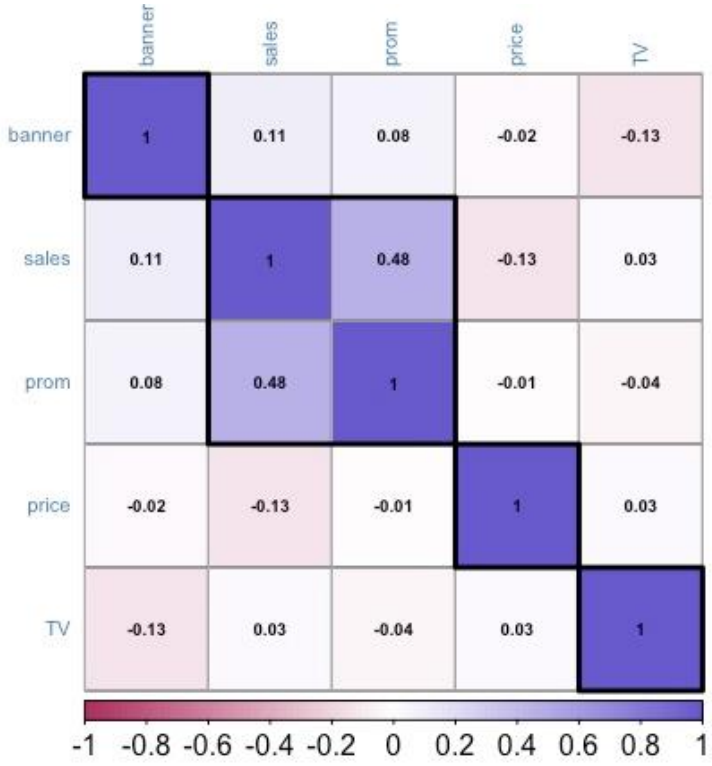
- Use ANOVA to find whether the variables are significant.
- Since p-value 0.000192 < 0.05, there are significant variables in month for sales.
- Since p-value 2e-16 < 0.05, there are significant variables in region for sales.
- Since p-value 0.958 > 0.05, there is no significant variable in time for sales.

Interpretation of results and conclusion

- The boxplots below present that most of the sales are concentrated in April and Capital.



Conclusion
If the manufacturer of a consumer goods brand wants to do marketing activities, promotion is the best choice to improve a little sales.



- There are various variables associated with sales in marketing.
- Although three marketing activities do not have a significant relationship with sales, we found that promotion is highlighted in the association problem. The correlation between sales and promotion is 0.48, which is higher than the online banner and TV.
- Price is less correlated with sales, but it is a significant variable for explaining sales in the model.

Validation and robustness checks

Jelly Beans Dataset | Train set (70%) | Test set (30%)

- Use the train set to create a model with all variables and use leave-one-out cross-validation to get root mean squared error.

RMSE	Rsquared	MAE
491.6287	0.9631174	389.7364

- Use forward, backward, and both stepwise selections to find significant variables.
  (Forward/ Both: prom, Apr, West, East, North, South, May, banner, price, TV, year1, Jun, Mar, Aug, Jan)
  (Backward: Jan, Aug, year2)
- Use selected variables to do leave-one-out cross validation and find RMSE.

RMSE	Rsquared	MAE
471.4586	0.9660586	379.736

[Forward stepwise selection]

RMSE	Rsquared	MAE
2073.14	0.3533114	1767.946

[Backward stepwise selection]

- Use forward selection to build the model and to predict sales in the test dataset because it gives us the lowest RMSE.
- The optimal model does not have multicollinearity.

prom	Apr1	East1	West1	North1	South1	May1	price	banner	TV	year11	Jun1
1.1412	1.1419	1.6055	1.6002	1.6009	1.6013	1.3489	1.0270	1.4415	1.6105	1.0816	1.1798

Mar1	Aug1	Jan1	
1.4365	1.0892	1.3547	→ all < 6

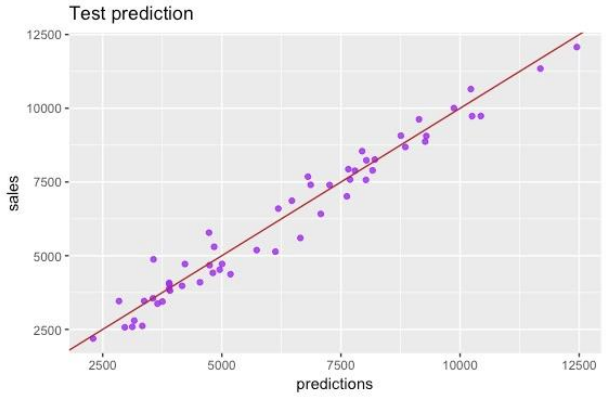
- On average, the predictions error of sales in the test set is around 499 packs and in train set is around 416 packs.

```
> rmse(test$sales, test$predictions)
[1] 498.9963
> rmse(train$sales, train$predictions)
[1] 416.1977
```

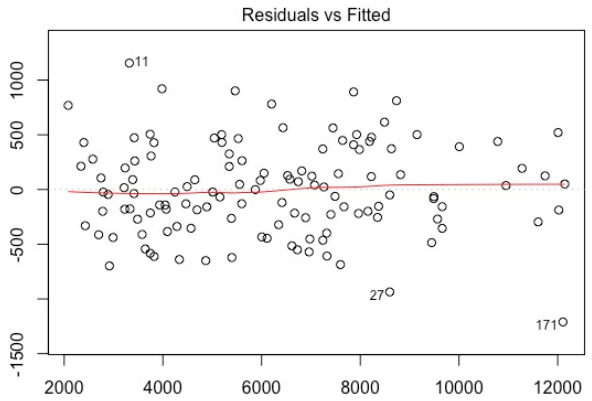
- On average, the model's prediction of sales is wrong by 7.7%.

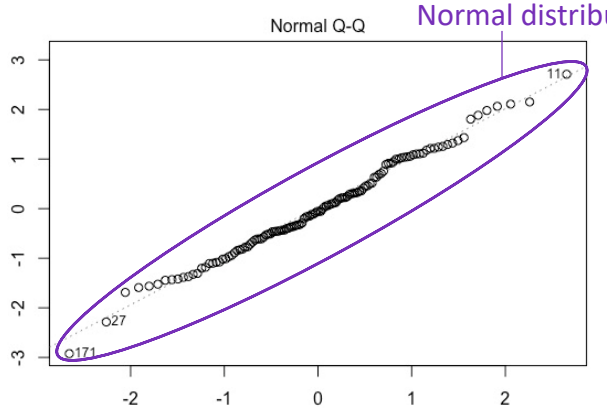
```
> mape(test$sales, test$predictions)
[1] 0.07787375
```

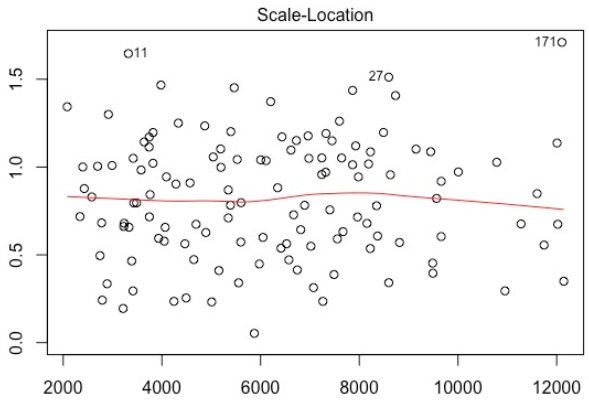
- The residuals are close to the red line as well as RMSE for test data is slightly bigger than train data, the model is **neither over-fitted nor under-fitted**.



Assumption tests and limitations of the model







- These plots are based on the optimal model.
- In Residuals vs Fitted plot, because the red line is close to the horizontal dotted line, sales and variables have a linear relationship. Additionally, the expected value of residuals is approximately equal to zero.
- In the Normal Q-Q plot, most of the residuals closing to line x=y shows that data in this model is a normal distribution.
- In the Scale-Location plot, the result is homoscedasticity.
- Through the Durbin-Watson test, since the p-value is larger than 0.05, the errors are not autocorrelated. It is independence assumption.

```
lag Autocorrelation D-W Statistic p-value
  1       0.02600565      1.920172   0.366
Alternative hypothesis: rho != 0
```

→ 0.366 > 0.05

It meets the common assumption.

Other dangers in the case

- Since the dataset is not big enough, the model may not be extremely accurate.
- Because some factors are not be considered in the model, the model may have some space to be improved.