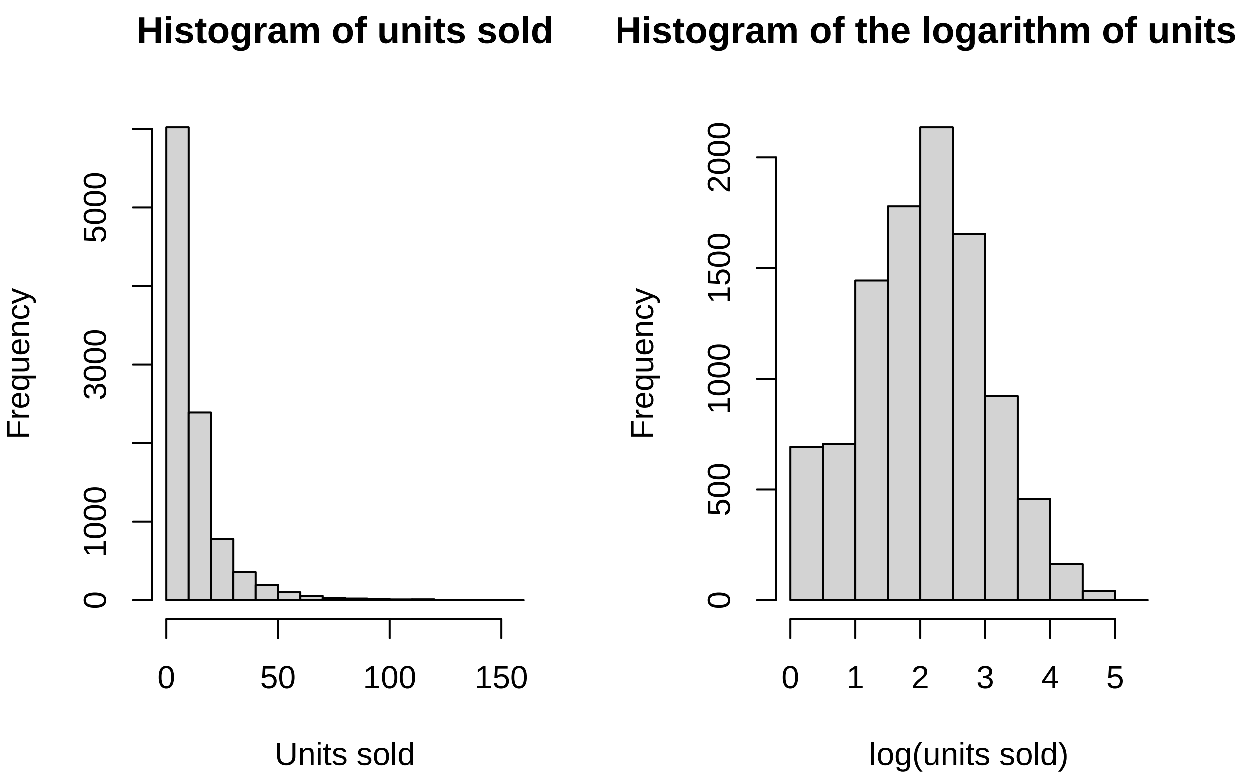
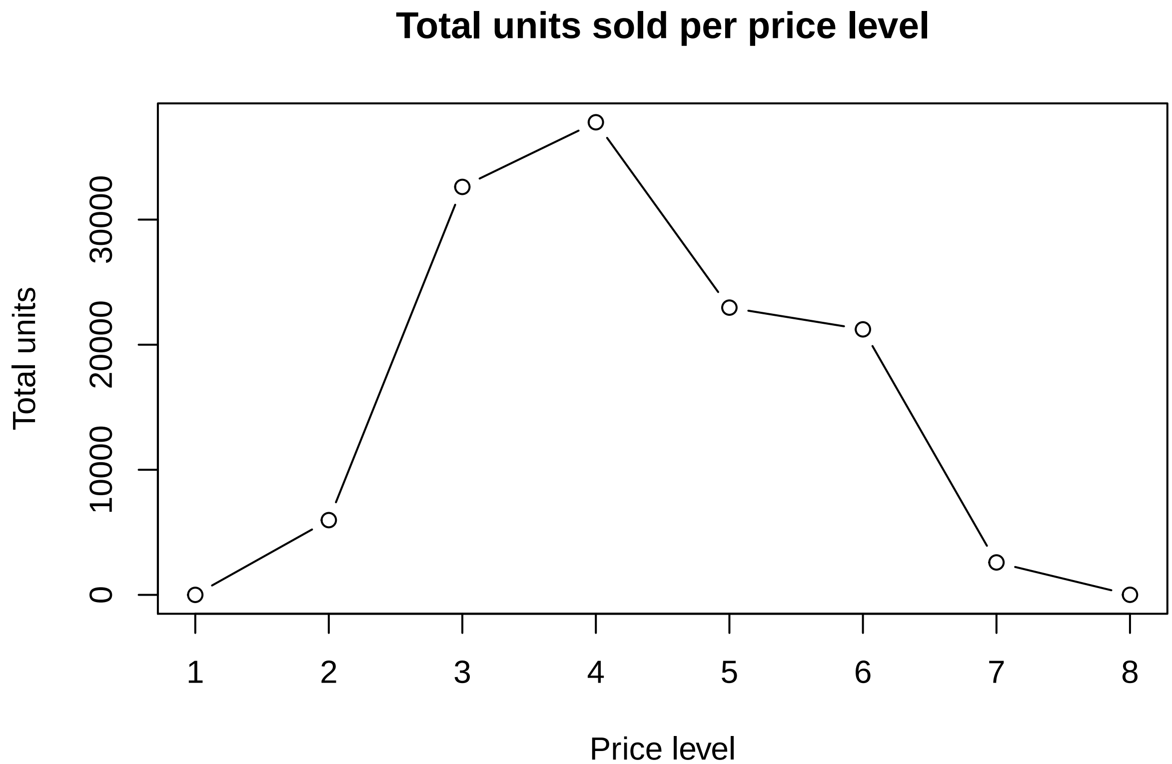
16096156

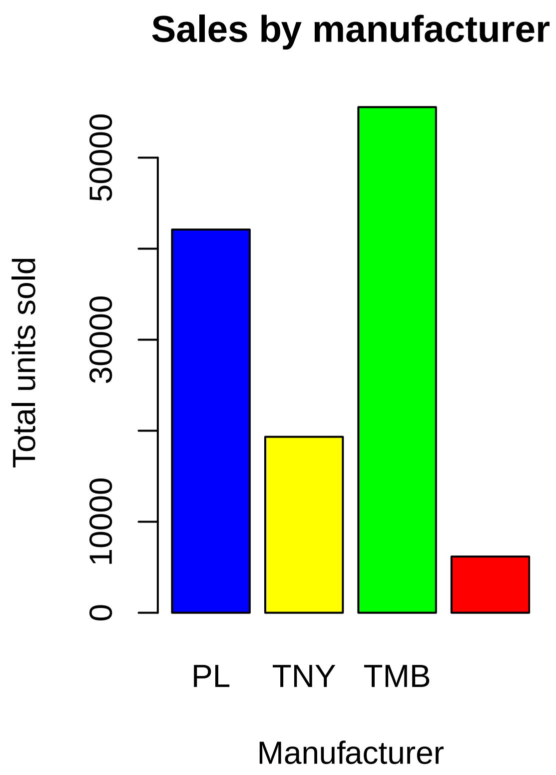
Section 1. Exploratory analysis

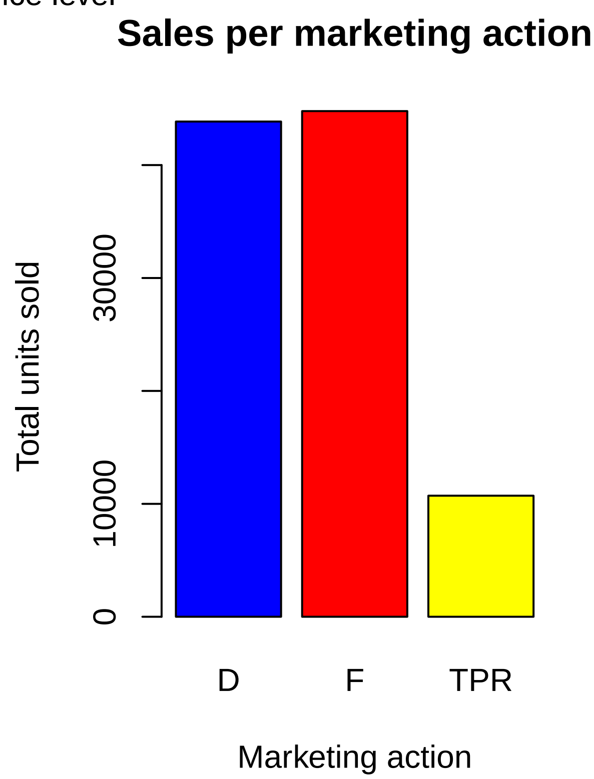
```
## The histogram of the response variable shows a clear positive skew in the data. That,
## combined with the fact that the data is bounded by 0, suggests it might be sensible to apply a
## logarithmic transformation to the response variable to normalize its distribution (represented
## below).
```




```
## We would like to explore first the relationship between price and sales. However, the
## basic units vs price plot is not particularly helpful, therefore we show below the total units
## sold at each price level (rounded to integers). It seems that pizza that is priced more
## moderately has sold significantly more units compared to what can be described as cheap(<3)
## or expensive pizza(>7). Secondly, as it relates to manufacturers, Private Label(PL) and
## Tombstone(TMB) are the best sellers, whereas Tonys(TNY) and King(KNG) are lagging behind
## on units sold. This means that certain manufacturers may bring more sales, which can be
## further explored through more advanced analysis. Lastly, when it comes to marketing efforts,
## it appears as though putting the product on in-store promotional display (D) or featuring it
## in the store leaflet (F) generates significantly more selling activity than just a price
## reduction (TPR). Again,this relationship should be further explored through more advanced
## analysis.
```





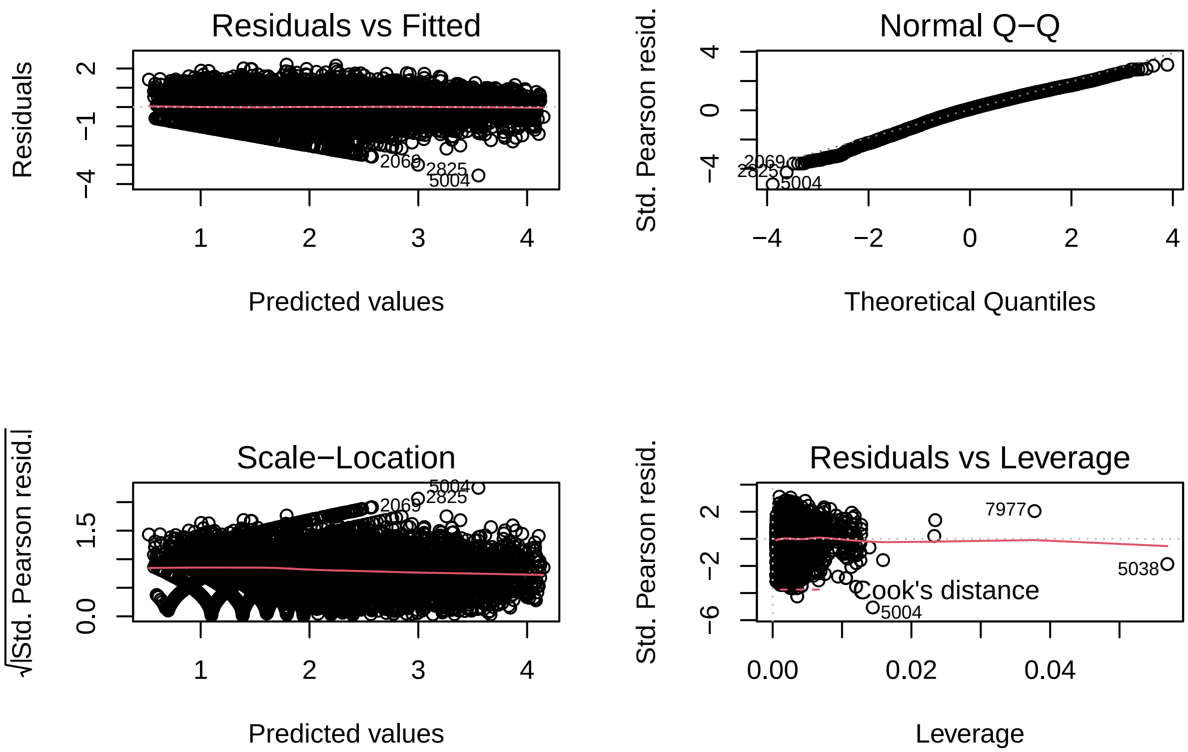



```
## Lastly, before we look towards the regression analysis, we should check if there is any
## correlation between some of the variables in the model. Based on the corelation matrix of the
## grocery data frame, several covariates appear to have a pairwise correlation higher than 0.5
## (or lower than -0.5 in the case of negative correlation), as follows: BASE_PRICE with PRICE,
## UPC, MANUFACTURER; PRICE with UPC; UPC with MANUFACTURER; FEATURE with DISPLAY. In regression
```

```
## analysis, this will generate colinearity, which can really affect the robustness of the
## regression coefficients. Although colinearity does not negate the validity of estimation in
## regression, it is something that we should account/ verify for and will do so through the
## introduction of pairwise interactions between variables in the chosen model.
```

Section 2.1. Regression models

```
## Firstly, we ran a regression within the Poisson family, with a log link function and all
## given variables included. It is clear from the diagnostic plots that this model does not fit
## the data, as both the assumption of normality of residuals and that of constant variance of
## residuals are violated. This could be due to the fact that the response variable is
## overdispersed, meaning that its variance is higher than its mean. To account for this, we also
## run a regression model with a quasi-poisson distribution, however there is no evident
## improvement in either the diagnostic plots of the model or in the deviance measure resulting
## from this model vs. the initial attempt. Thirdly, given the shape of the distribution of units
## sold, we also consider a negative binomial model. This is often considered an alternative to
## the Poisson distribution for overdispersed data.The diagnostic plots point again towards a
## model that is not fit to our data (heteroskedasticity, non-normally distributed residuals).
## Violations of the constant variance and normal distribution of residuals invalidate the use
## of AIC for comparisons. The deviance of this model is 10132, lower than that of previous models.

## Lastly, we will explore the log transformation of the response variables. The histogram
## of the data resembles a normal distribution, therefore the last model considered in this
## analysis will be a generalized linear model with a family of normal distributions and a log
## function corresponding to the identity. While this model is a better fit for the data than
## the previous ones, it is crucial to incorporate the effects of the interactions between
## variables, given our findings in Section 1. We thus run the following regression model,
## with the corresponding diagnostic plots below. The assumptions of homoskedasticity and
## normal distribution of residuals appear to be more robust. This model also results in
## lower AIC and deviance than the previous model. We thus conclude this model is the
## optimal choice.

## From this model, a number of covariates seem to be negatively related to an increase
## in sales, namely BASE_PRICE, PRICE, MANUFACTURERS (PL, TMB, TONYS). Equally, the model
## also suggests that the interactions between BASE_PRICE and the three previously
## mentioned manufacturers might have a significant impact on sales, however their
## coefficients are not statistically significant at a 5% significance level.

## Regression formula:

## UNITS_LOG ~ BASE_PRICE + PRICE + WEEK_END_DATE + STORE_NUM +
##     UPC + MANUFACTURER + DISPLAY + FEATURE + TPR_ONLY + BASE_PRICE *
##     PRICE + BASE_PRICE * UPC + BASE_PRICE * MANUFACTURER + PRICE *
##     UPC + UPC * MANUFACTURER + FEATURE * DISPLAY

## Diagnostic plots:
```



Section 2.2. More sophisticated regression analysis

```
## Despite the greater interpretability of splines/ additive models, it is generally believed
## that methods like decision trees or gradient boosting are more suitable for prediction in
## moderate to high dimensions (which is the case of our data here) and therefore, I have elected
## to focus on CART trees, random forests and gradient boosting in this question.
```

```
## Judging by both the mean absolute error and the mean squared error, gradient boosting
## appears to be the optimal model as it generates the smallest differences between predicted
## values and actual values (based on the test set), on average. A table summarizing the MAE
## and MSE for each model is rendered below:
```

```
##                names      MAE       MSE
## 1               CART 6.728891 118.38710
## 2      Random forest 5.047458  68.93190
## 3 Gradient boosting 4.557631  55.22597
```

```
## A possible interpretation of the gradient boosting model is that 3 variables appear to
## contribute the most to the prediction, namely DISPLAY, PRICE and BASE_PRICE. Intuitively, it
## is reasonable to believe that the price of a product and the fact that it is part of a
## promotional display determine sales of that product and even more so, that the actual price
## charged (PRICE) plays a more significant role than the product’s base price (BASE_PRICE).
## Somewhat unexpectedly, the product’s manufacturer does not influence sales significantly, and
## neither does having the product’s price reduced without marketing that action in any way.
```

Section 3. Final model selection

```
## The means of the RMSEs corresponding to the two models elected in Section 2 are:
##  17.03 (regression model),  7.26 (gradient boosting).

## We performed a 10-fold cross validation computation to generate 10 different values for
## the RMSEs of each the regression model and the gradient boosting method. On the back of the
## t-tests performed on the two RMSE data sets, we can conclude that the mean RMSE resulting from
## the regression model is higher.

## The main advantages of the linear regression model are interpretability and ease of
## implementation in practice. Given that prediction is our focus in this scenario however, the
## disadvantages of linear regression weigh heavier on the decision of which model to choose.
## Its robustness is questioned by 1. existing colinearity between some of the explanatory
## variables and 2. the absence of a linear relationship between the response variable and some
## of the covariates (e.g. price vs units sold follows a clearly non-linear curve). The latter
## impacts the model's prediction power significantly, as can be seen by the higher RMSE compared
## to that of the gradient boosting model. On the other hand, the main advantage of gradient
## boosting is its increased predictive accuracy. This method is particularly effective in
## high dimensions and with large datasets, which is the case in our current scenario.
## However, gradient boosting can be slower to implement and more expensive when compared to
## regression models as it requires large datasets that can be difficult/ expensive to acquire.
## On the basis of its superior predictive accuracy, we conclude that gradient boosting with
## hyperparamter tuning is the final elected model to use. Below we analyze its output and aim
## to interpret its results.
```

Section 4. Prediction estimate

```
## A brief interpretation of the gradient boosting method suggests that DISPLAY, PRICE AND
## BASE_PRICE have the largest contribution to the reduction in RMSE , while it is interesting to
## see that certain variables, such as the UPC, TPR_ONLY, MANUFACTURER contribute very little in
## that regard. Additionally, variables such as PRICE, BASE_PRICE and DISPLAY also have a
## significant impact on the prediction of sales, all of which appear to be positively correlated
## to an increase in sales. This suggests, for instance, that displaying the product in store
## as part of a promotion could generate an increase in sales of just over 3%.

## Similarly, our analysis suggests that the impact of a 10% reduction in price, keeping
## all else constant is:  -0.29 , corresponding to a reduction of  -29 % in units sold.
```