In [6]:
```
library(multcomp)
```

```
Loading required package: mvtnorm

Loading required package: survival

Loading required package: TH.data

Loading required package: MASS


Attaching package: 'TH.data'


The following object is masked from 'package:MASS':

    geyser
```

In [7]:
```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------------------------
---------------------- tidyverse 1.3.1 --

v ggplot2 3.3.5     v purrr   0.3.4
v tibble  3.1.5     v dplyr   1.0.7
v tidyr   1.1.4     v stringr 1.4.0
v readr   2.0.2     v forcats 0.5.1

-- Conflicts -----------------------------------------------------------------
---------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
x dplyr::recode() masks car::recode()
x dplyr::select() masks MASS::select()
x purrr::some()   masks car::some()
```

In [8]:
```
library(cowplot)
library(VIM)
```

```
Loading required package: colorspace

Loading required package: grid

VIM is ready to use.


Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues (https://github.com/statistikat/VIM/issues)


Attaching package: 'VIM'


The following object is masked from 'package:datasets':

    sleep
```

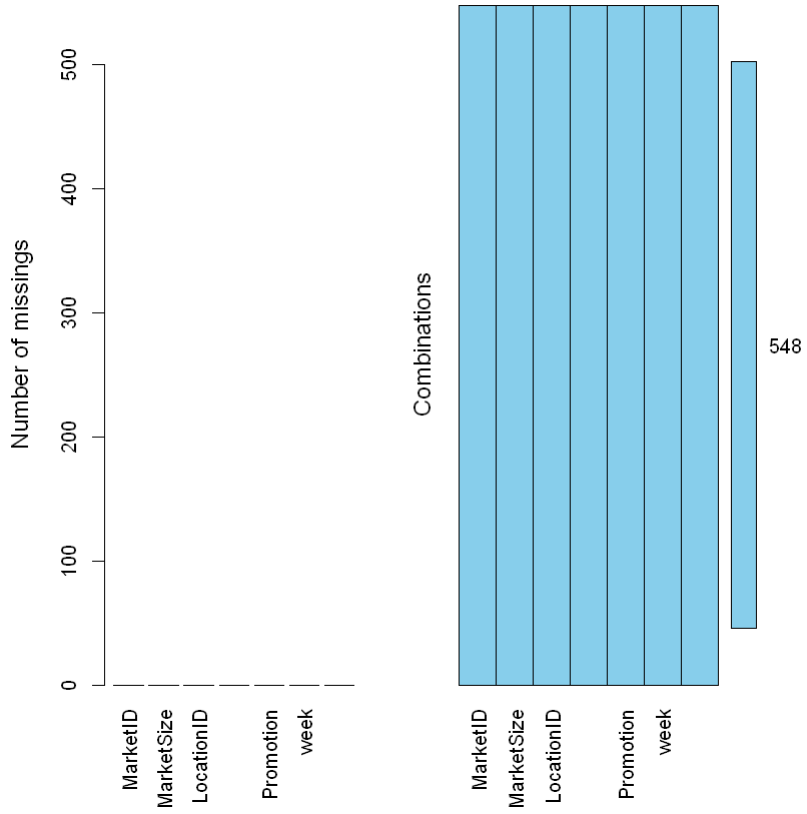
Import and inspect data

In [9]:
```
df<-read.csv("Desktop/DataMining-master/WA_Fn-UseC_-Marketing-Campaign-Eff-UseC_-

#check results
head(df)
```

A data.frame: 6 × 7

	MarketID	MarketSize	LocationID	AgeOfStore	Promotion	week	SalesInThousands
	<int>	<chr>	<int>	<int>	<int>	<int>	<dbl>
1	1	Medium	1	4	3	1	33.73
2	1	Medium	1	4	3	2	35.67
3	1	Medium	1	4	3	3	29.03
4	1	Medium	1	4	3	4	39.25
5	1	Medium	2	5	2	1	27.81
6	1	Medium	2	5	2	2	34.67

In [10]:
```
#check for missing data using VIM package
aggr(df, prop = F, numbers = T) # no red - no missing values
```



In [11]:
```
#summary sales statistics
(grouped.df <- df %>%
  group_by(Promotion) %>%
  summarize(
    count = n(),
    totalSales = sum(SalesInThousands),
    meanSales = mean(SalesInThousands),
    sd = sd(SalesInThousands)))
```

A tibble: 3 × 5

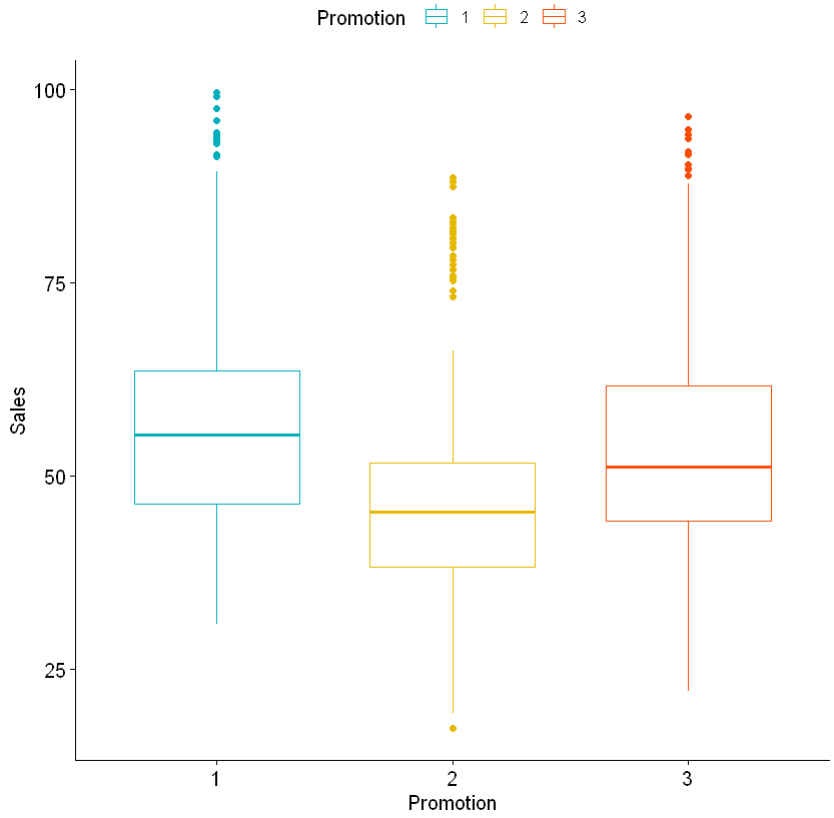
Promotion	count	totalSales	meanSales	sd
<int>	<int>	<dbl>	<dbl>	<dbl>
1	172	9993.03	58.09901	16.55378
2	188	8897.93	47.32941	15.10895
3	188	10408.52	55.36447	16.76623

-We can see that group 3 created the most sales followed by groups 1 & 2 -We can also see that there were 172 stores that were in promotion 1 while there were 188 stores in promotion 2. This is technically not balanced, but nearly-balanced. -As long as we have equal variances in our groups, this shouldn't be a problem.

In [12]:
```
library("ggpubr")
```

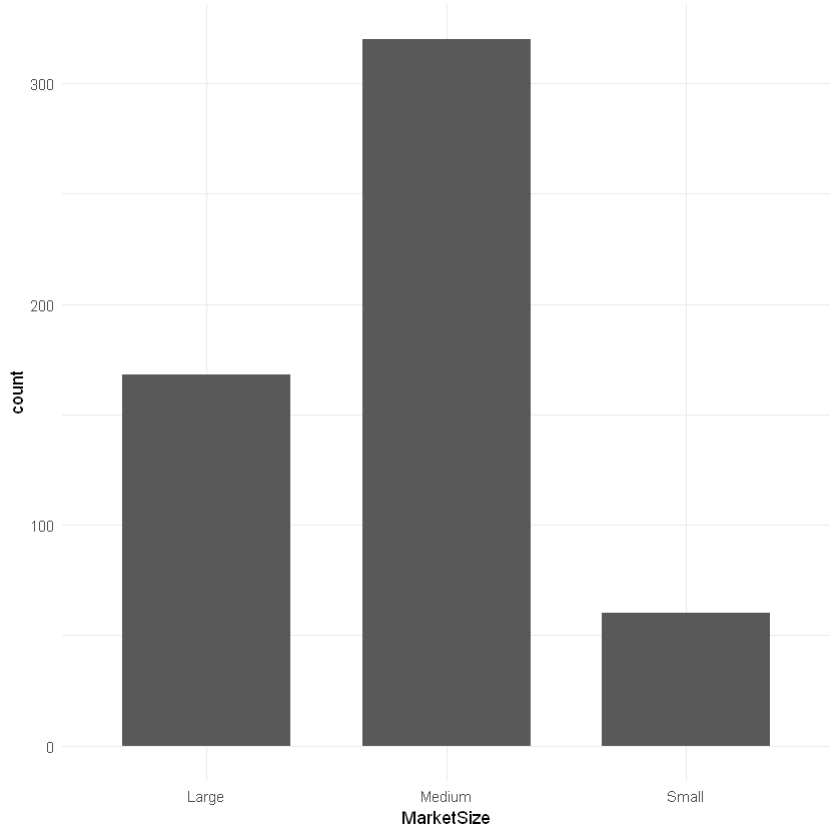
```
Attaching package: 'ggpubr'


The following object is masked from 'package:cowplot':

    get_legend
```

In [13]:
```
ggboxplot(df, x = "Promotion", y = "SalesInThousands",
          color = "Promotion", palette = c("#00AFBB", "#E7B800", "#FC4E07"),
          ylab = "Sales", xlab = "Promotion")
```



Data visualization and exploration

In [14]:
```
(viz_1 <- ggplot(df, aes(x=MarketSize))+
  geom_bar(stat="count", width=0.7)+
  theme_minimal())
```



Data cleaning

In [15]:
```
#check the promotion variable
str(df$Promotion) #an integer object, we need to change this, changing numerical
```

```
 int [1:548] 3 3 3 3 2 2 2 2 1 1 ...
```

In [16]:
```
#factor the promotion variable before we model it
df$Promotion <- as.factor(df$Promotion)

#check results
str(df$Promotion)
```

```
 Factor w/ 3 levels "1","2","3": 3 3 3 3 2 2 2 2 1 1 ...
```

Data question & hypothesis test

Does store sales differ by promotion?

In [17]:
```
aggregate(SalesInThousands ~ Promotion, df, mean)
```

A data.frame: 3 × 2

Promotion <fct>	SalesInThousands <dbl>
1	58.09901
2	47.32941
3	55.36447

In [18]:
```
#promotion 1 has the highest level of sales, but
# is it statistically significant?
```

Significance Testing - ANOVA

Promotion 1 has the highest mean of sales, but is it statistically significant?

In [19]:
```
#We plot the ANOVA model to visualize confidence
#intervals for mean sales by promotion
df.anova <- aov(SalesInThousands ~  Promotion, data = df)
summary(df.anova)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
Promotion     2  11449    5725   21.95 6.77e-10 ***
Residuals   545 142114     261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions and interpretation: We see that the sales differs by Promotion, and the model is statistically significant but we don't know which pair groups are significant

How can we change this?? We need to perform additional testing

Post hoc testing

In [20]:
```
#Use glht() to perform multiple pairwise-comparisons for
# a one-way ANOVA: (with confidence interval and p-values)
summary(glht(df.anova, linfct = mcp(Promotion = "Tukey")))
```

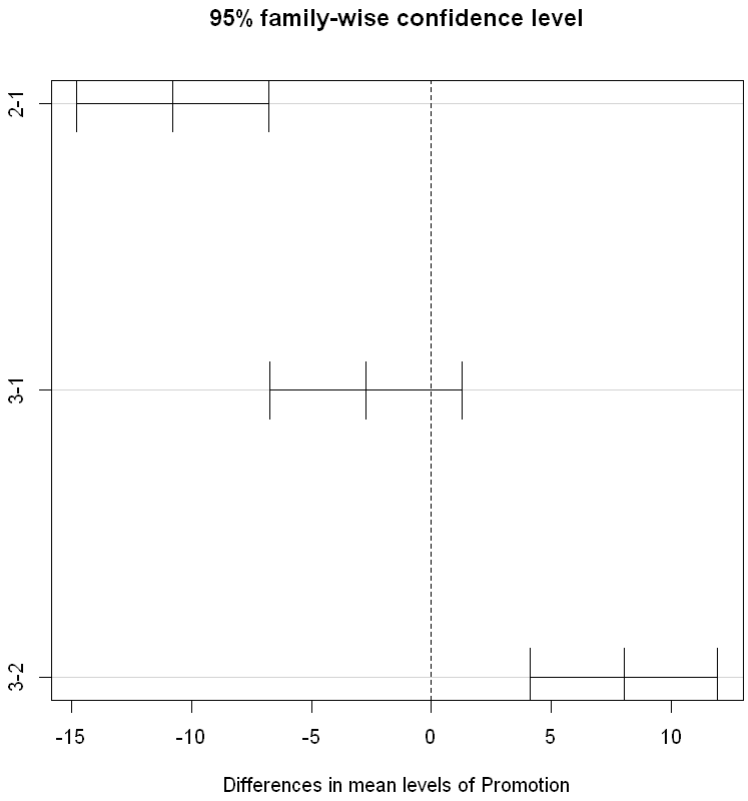
```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = SalesInThousands ~ Promotion, data = df)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0  -10.770      1.704  -6.321   <1e-04 ***
3 - 1 == 0   -2.735      1.704  -1.605    0.244
3 - 2 == 0    8.035      1.666   4.824   <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

In [21]:
```
#group 2 is significant against group 1
#group 3 is significant against group 2

TukeyHSD(aov(df.anova), "Promotion") #does same as glht function but includes the
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = df.anova)

$Promotion
          diff        lwr       upr     p adj
2-1 -10.769597 -14.773842 -6.765351 0.0000000
3-1  -2.734544  -6.738789  1.269702 0.2443878
3-2   8.035053   4.120802 11.949304 0.0000055
```

In [22]:
```
#diff: difference between means of the two groups
#lwr, upr: the lower and the upper end point of the confidence interval at 95% (d
#p adj: p-value after adjustment for the multiple comparisons.
```

In [23]:
```
# plot difference in mean levels of promotion
plot(TukeyHSD(df.anova))
```



In [24]:
```
#Post hoc testing
posthoc <- TukeyHSD(x=a1, conf.level = 0.95)
posthoc
```

```
Error in TukeyHSD(x = a1, conf.level = 0.95): object 'a1' not found
Traceback:

1. TukeyHSD(x = a1, conf.level = 0.95)
```

In [25]:
```
#diff: difference between means of the two groups
#lwr, upr: the lower and the upper end point of the confidence interval at 95% (d
#p adj: p-value after adjustment for the multiple comparisons.
```

With all 3 plotted with confidence intervals, Promo 2 is significantly worse than Promo 1 and 3, but we cannot say that Promo 1 and 3 are significant as their confidence intervals overlap.

Non Parametric Tests

In [26]:
```
#1. Kruskal-Wallis
#Non-parametric alternative to ANOVA
# It's recommended when the assumptions of one-way ANOVA test are not
# met. One of those assumptions are that the residuals are normally
# distributed

kruskal.test(SalesInThousands ~ Promotion, data = df)
```

```
        Kruskal-Wallis rank sum test

data:  SalesInThousands by Promotion
Kruskal-Wallis chi-squared = 53.295, df = 2, p-value = 2.674e-12
```

In [27]:
```
#The p-value is tiny; therefore, we can reject null hypothesis that there
# are no differences in group means, but we don't know which groups.

#2. We do pairwise comparisons and adjust for multiple groups
pairwise.wilcox.test(df$SalesInThousands, df$Promotion,
                 p.adjust.method = "bonferroni", paired = FALSE)
```

```
        Pairwise comparisons using Wilcoxon rank sum test with continuity corre
ction

data:  df$SalesInThousands and df$Promotion

  1       2
2 1.8e-11 -
3 0.11    3.6e-07

P value adjustment method: bonferroni
```

In [28]:
```
#3. Levene's test for non-normal distribution - we check due to skew in residuals
library(car)
leveneTest(SalesInThousands ~ Promotion, data = df)
```

A anova: 2 × 3

	Df	F value	Pr(>F)
	<int>	<dbl>	<dbl>
group	2	1.269679	0.2817515
	545	NA	NA

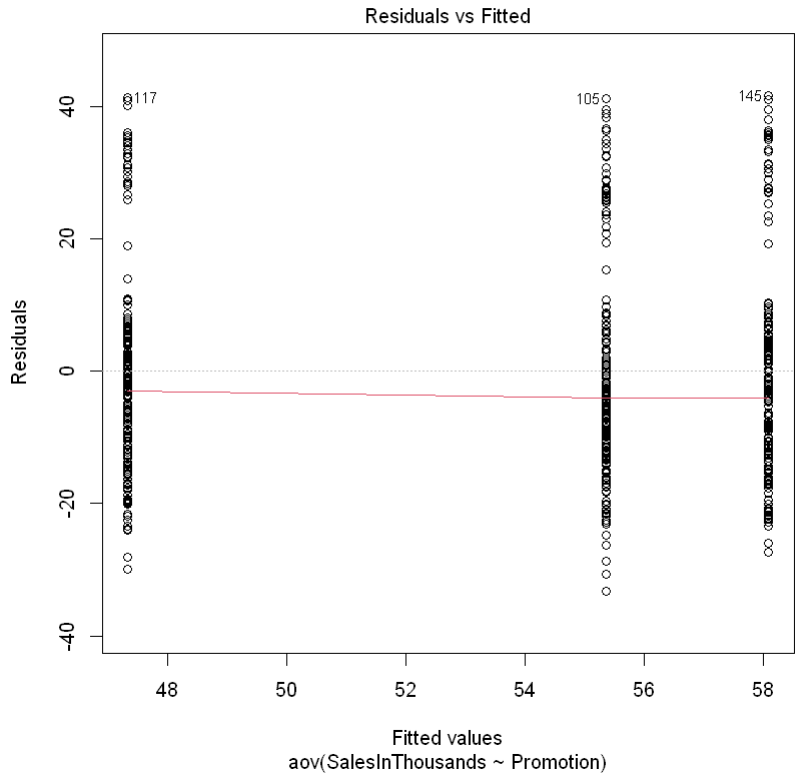
In [29]:
```
#We see that the p-value for Promotion 2 is large and therefore not significant.
```

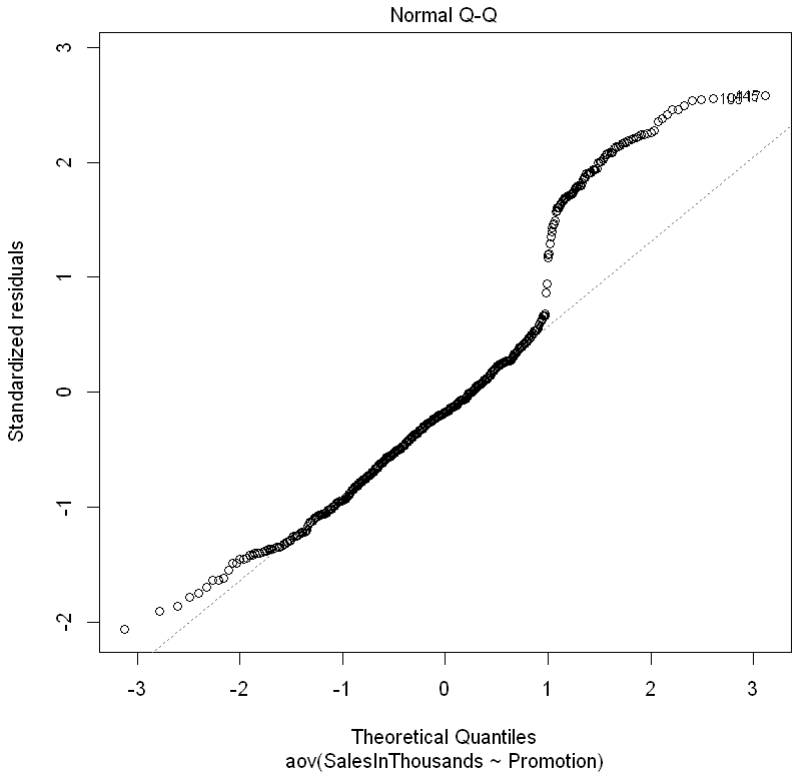
In [30]:
```
#4. Shapiro-Wilk (Has better power than K-S test) A Shapiro-Wilk test is the test

# Extract the residuals
aov_residuals <- residuals(object = df.anova)
# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals )
```

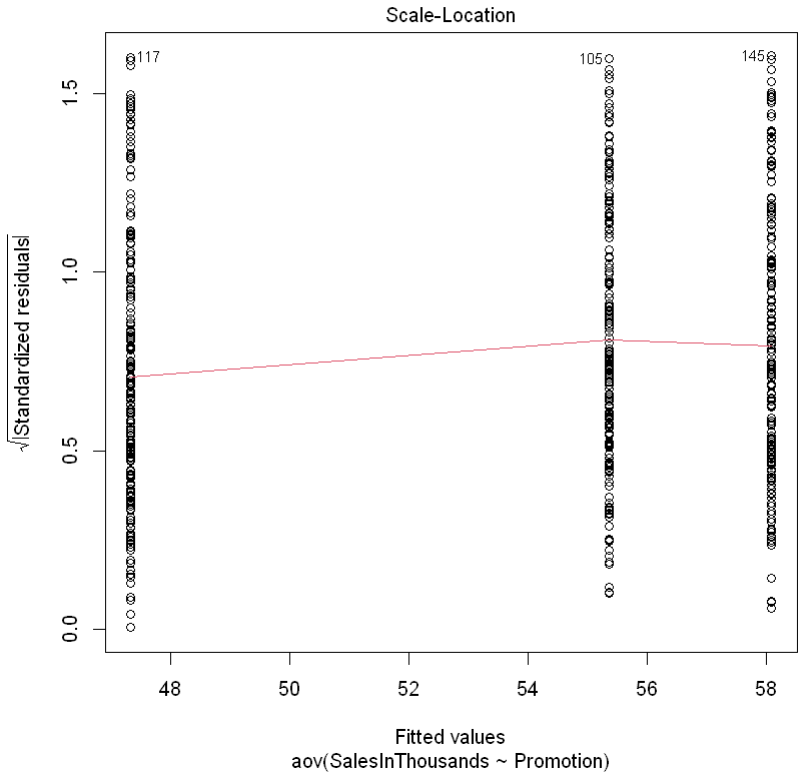
```
        Shapiro-Wilk normality test

data:  aov_residuals
W = 0.92208, p-value = 3.155e-16
```

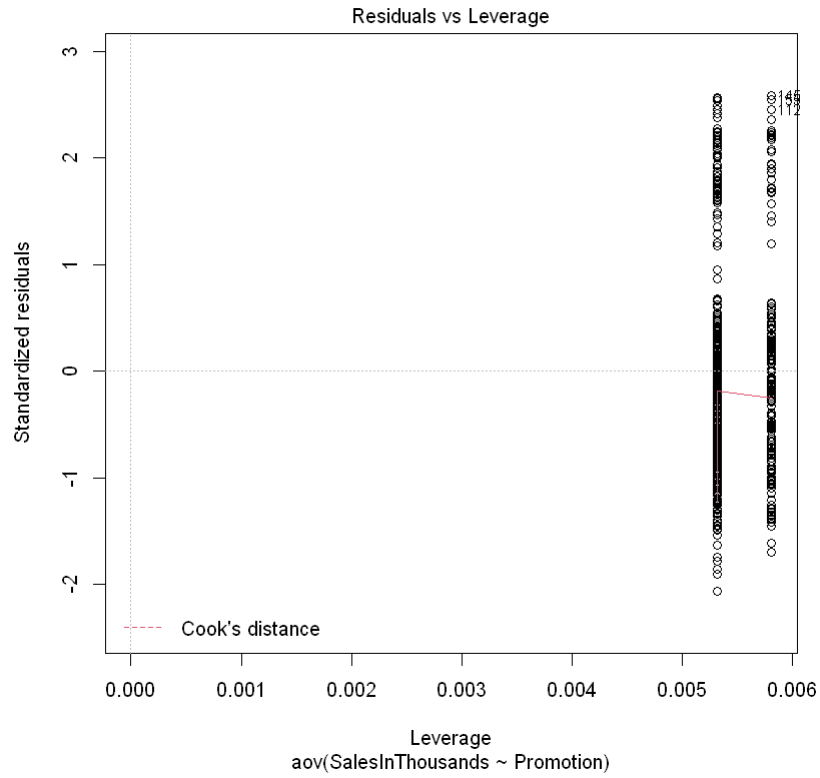
In [31]:
```
#We reject null hypothesis that residuals are normally distributed
```

In [32]:
```
#5. Homogeneity of variances
plot(df.anova) #first anova model -- Looks good
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
aov(SalesInThousands ~ Promotion)
-3
-2
-1
0
1
2
3


Scale-Location
117
105
145
√|Standardized residuals|
0.0
0.5
1.0
1.5
48
50
52
54
56
58
Fitted values
aov(SalesInThousands ~ Promotion)


Residuals vs Leverage
145
112
Standardized residuals
-2
-1
0
1
2
3
Cook's distance
0.000
0.001
0.002
0.003
0.004
0.005
0.006
Leverage
aov(SalesInThousands ~ Promotion)

This validates what we have done above with original anova model. Our conclusions from are original findings are still valid most likely due to having a very large sample size to make the group comparisons.

Inference : What should you tell the marketing & sales team?

Let's run again with just promotion 1 & 3 to see if we can get a significant result. The test should not take as long to run as we only have 2 groups to compare so we could see significant results quite fast.

Having a proper control group for comparison to be able to calculate the impact of the promotions

It appeared in group 1 there were some stores that were slightly younger than those in Group 3 it may not have made a difference but we should try to control for this.