avocado-e4

April 5, 2024

1 Avocado Project (Group E4)

1.0.1 Introduction and Motivation

Dataset: Avocado Prices

The source we are using is the clean and dense dataset provided by **Kaggle**
- Link: https://www.kaggle.com/datasets/neuromusic/avocado-prices

Raw Avocado dataset was provided to Kaggle by **Hass Avocado Board**
- Link: https://hassavocadoboard.com/

In this project, we examine a rich dataset from the Hass Avocado Board, which is sourced from retailer scan data representing national retail volume and price, reflecting actual retail sales for avocados in the United States from 2015 to 2018. Our team aims to decipher the patterns that govern avocado sales and pricing across the United States. Our focus lies in the distinct variables of conventional and organic types of avocados, average pricing, and regional differences of the U.S. Our objective is to employ statistical models to discern patterns and provide insights that can aid in market understanding and decision-making. Through our analysis, we aim to bridge theoretical knowledge with practical market applications, offering a nuanced perspective on the dynamics at play within the avocado industry. Our analysis is motivated by the potential to facilitate better financial planning for avocado enthusiasts everywhere, ensuring that the enjoyment of this beloved fruit continues unabated.

1.0.2 Data Preparation:

A. Data Cleaning We first import the raw data downloaded from kaggle:

A data.frame: 2 x 6

	Date <chr>	AveragePrice <dbl>	Total.Volume <dbl>	type <chr>	year <int>	region <chr>
1	2015-12-27	1.33	64236.62	conventional	2015	Albany
2	2015-12-20	1.35	54876.98	conventional	2015	Albany

We verify that the data is nice and dense:

[123]:
```
# Check for null values in dataframe
na_count = sum(is.na(df))
sprintf("The dataset contains %d na values", na_count)
```

'The dataset contains 0 na values'

We then factorize all categorical variables in our dataset: (justification)

A data.frame: 2 x 6

	Date <chr>	AveragePrice <dbl>	Total.Volume <dbl>	type <fct>	year <fct>	region <fct>
1	2015-12-27	1.33	64236.62	conventional	2015	Albany
2	2015-12-20	1.35	54876.98	conventional	2015	Albany

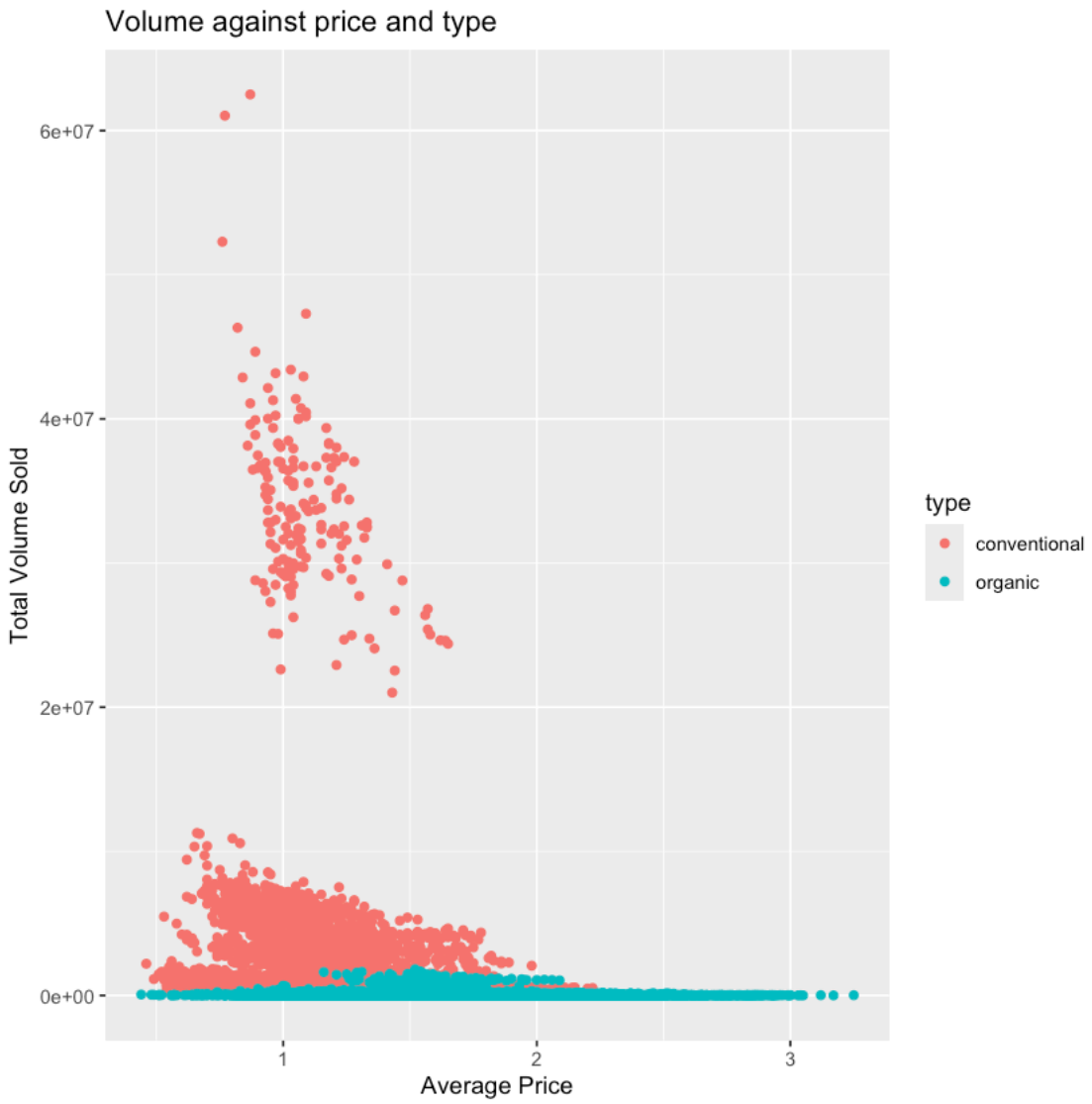
B. Feature Transformation As discussed in the proposal and proposal feedbacks, we will extract month information from the Data Column, as a categorical explanatory variable in our model. By extracting month feature from date, we are making the dimension of features much smaller (from 365 to 12), while perserving most information. This will result in an easier to interpret model and the model will be less prone to overfitting.

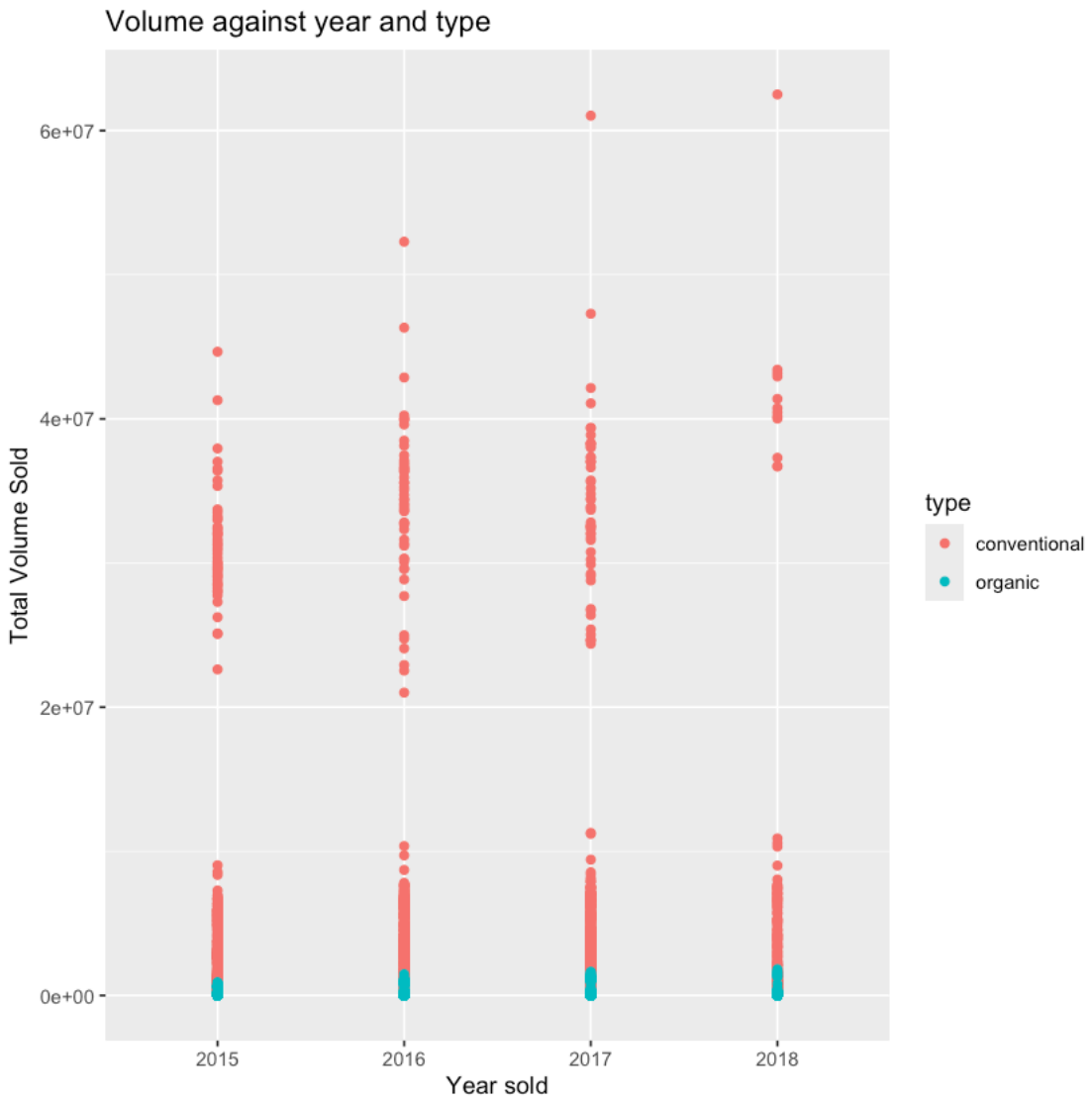
A data.frame: 2 x 7

	Date <date>	AveragePrice <dbl>	Total.Volume <dbl>	type <fct>	year <fct>	region <fct>	Month <fct>
1	2015-12-27	1.33	64236.62	conventional	2015	Albany	12
2	2015-12-20	1.35	54876.98	conventional	2015	Albany	12

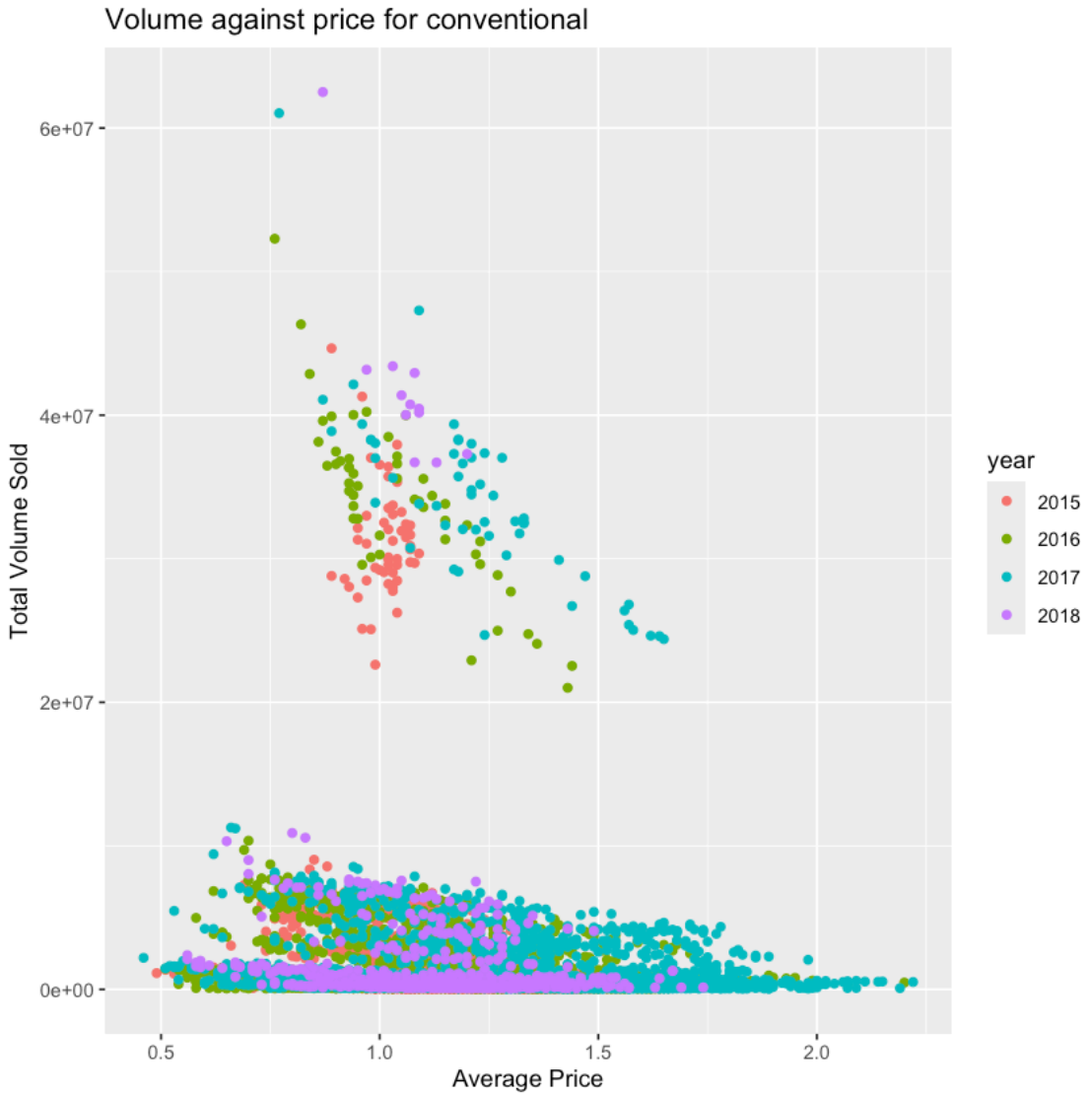
1.0.3 EDA

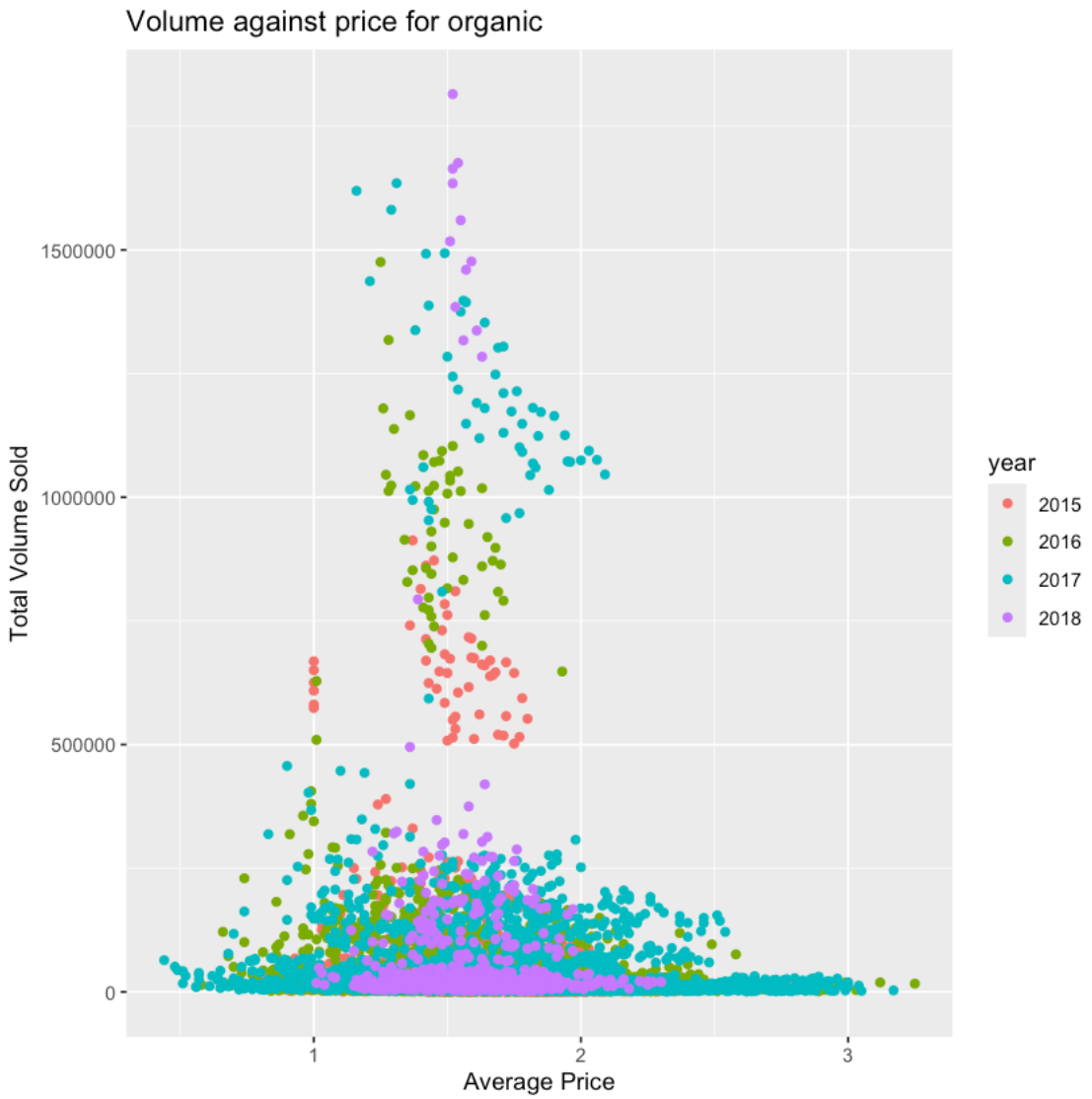
To explore the dataset, we'll begin by creating basic plots to visualize key relationships. We'll plot volume against price and volume against the year sold, with the variable avocado type distinguished by colour in order to see if there's any notable difference between sales of the two.


Volume against price and type
6e+07
4e+07
2e+07
0e+00
Total Volume Sold
1
2
3
Average Price
type
conventional
organic



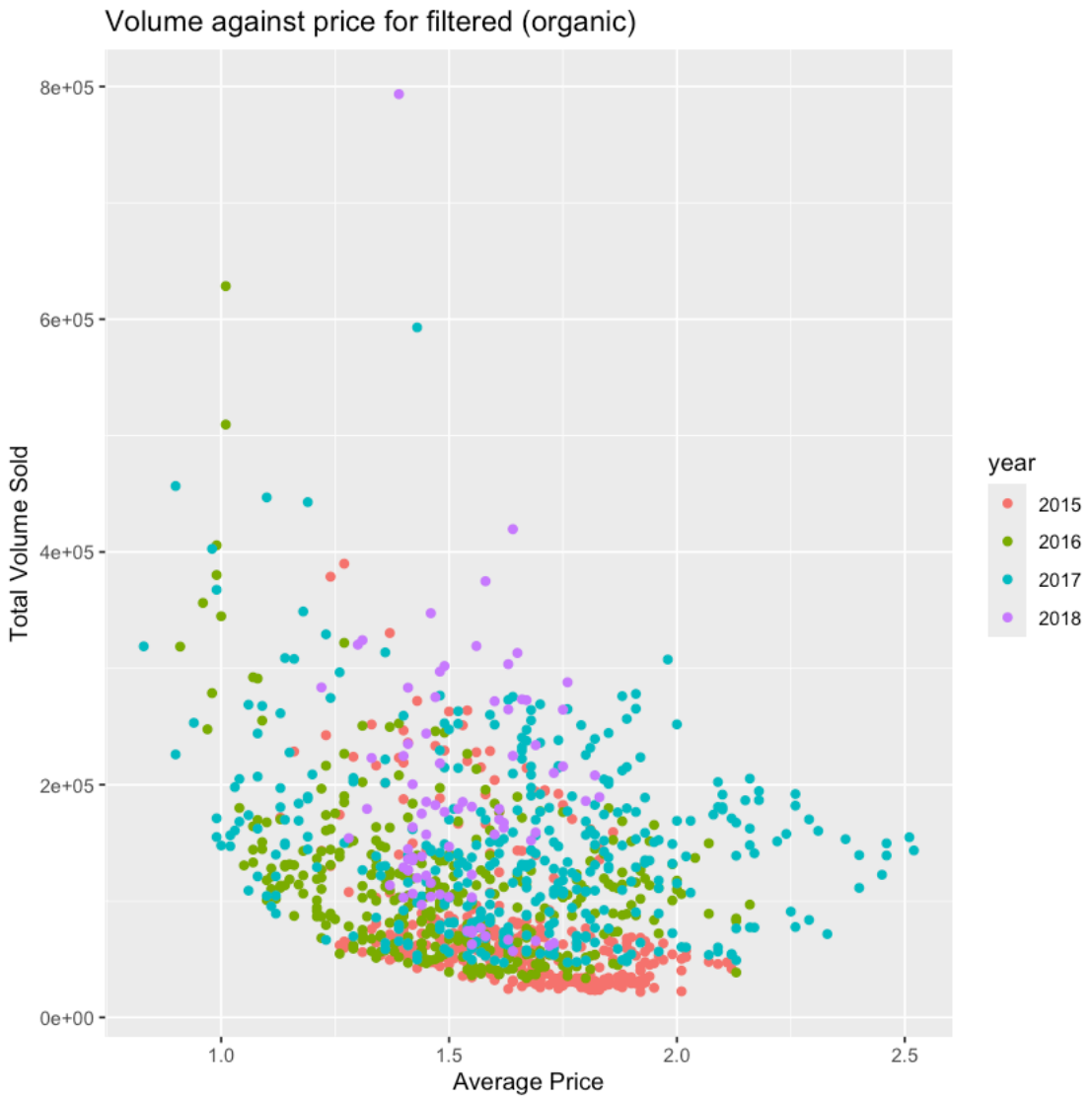
Upon examining these two plots, it becomes evident that there exists a significant disparity in the volume of conventional and organic avocados sold, thus, we will analyze them separately. Furthermore, we will incorporate the year as a variable represented by color to discern any discernible patterns over time.

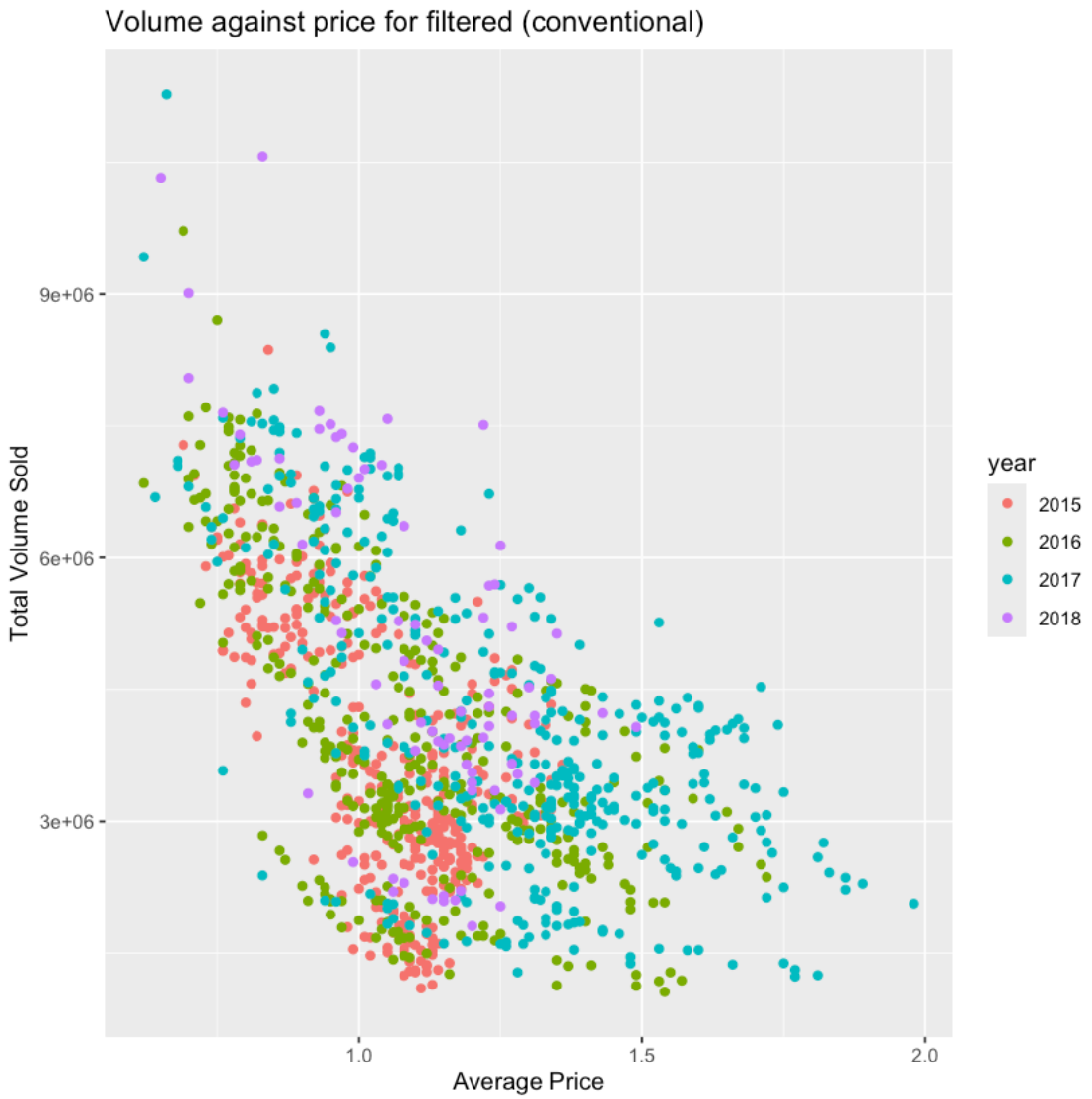

Volume against price for conventional
Total Volume Sold
6e+07
4e+07
2e+07
0e+00
0.5
1.0
1.5
2.0
Average Price
year
2015
2016
2017
2018



In both the conventional and the organic avocado plots, it can be seen that there are two distinct groups, to see why there's this difference, we will group by the other variables to see if we can identify the reason.

In addition, by inspecting the data more closely, it can be seen that many of the top results are for total US/ regions of the US, aka west, east, and central. In order to account for this, we can remove these points and visualize them separately while doing this, we will also make two dataframes, one for conventional, and one for organic.


Volume against price for filtered (organic)
Total Volume Sold
8e+05
6e+05
4e+05
2e+05
0e+00
1.0
1.5
2.0
2.5
Average Price
year
2015
2016
2017
2018



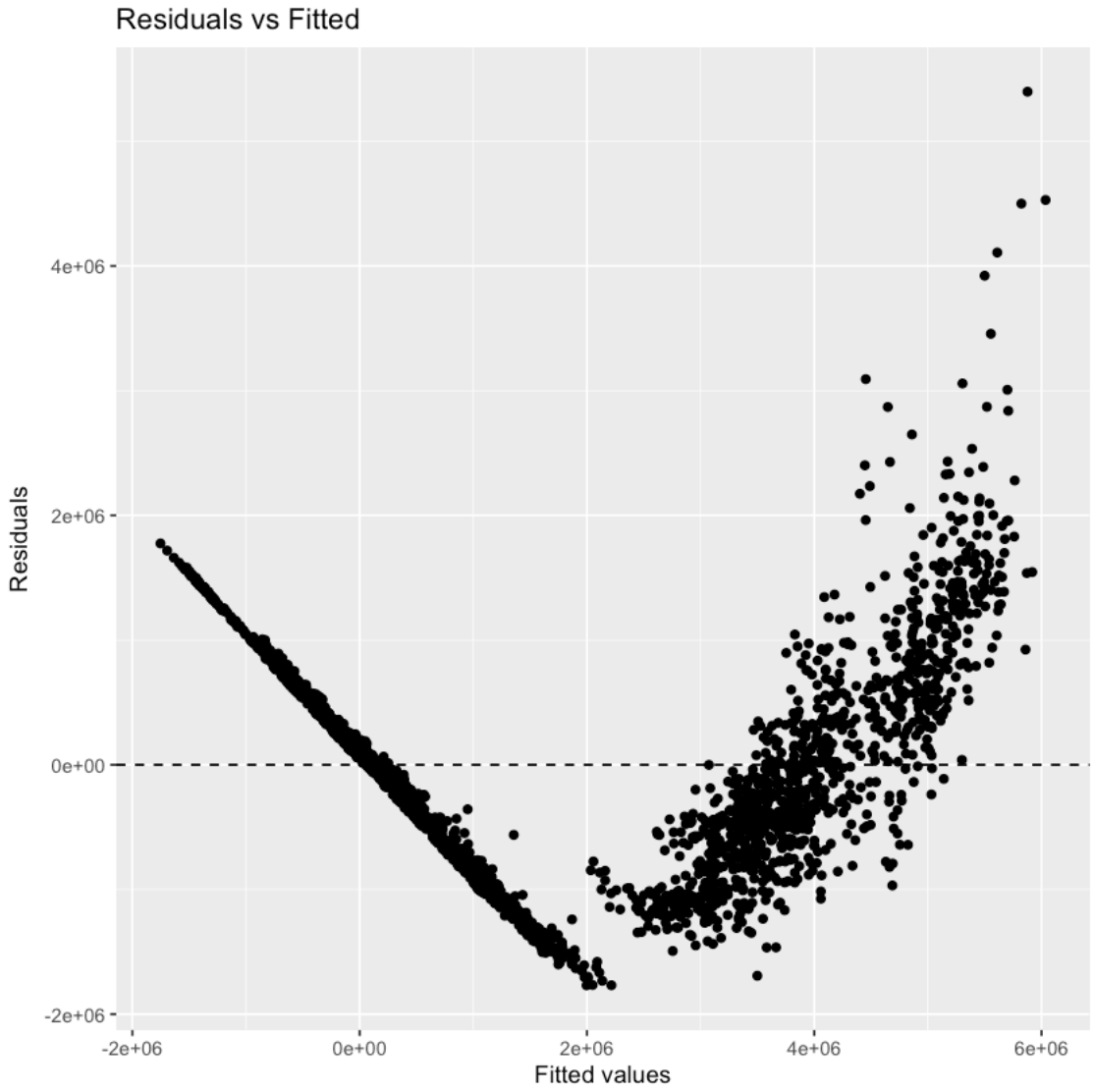
After removing the data points for individual cities, significantly clearer patterns for both conventional and organic avocados can be noticed, and as such, we will continue to use this refined dataset moving forward. The rationale for why we have decided to use this dataset as opposed to the dataset with individual cities lies in this dataset's ability to perform much more advanced model fitting. With this dataset, we can efficiently fit a greater variety of models within a reasonable timeframe owing to the lowered amount of data compared to the alternative, enabling more comprehensive analysis and insights.

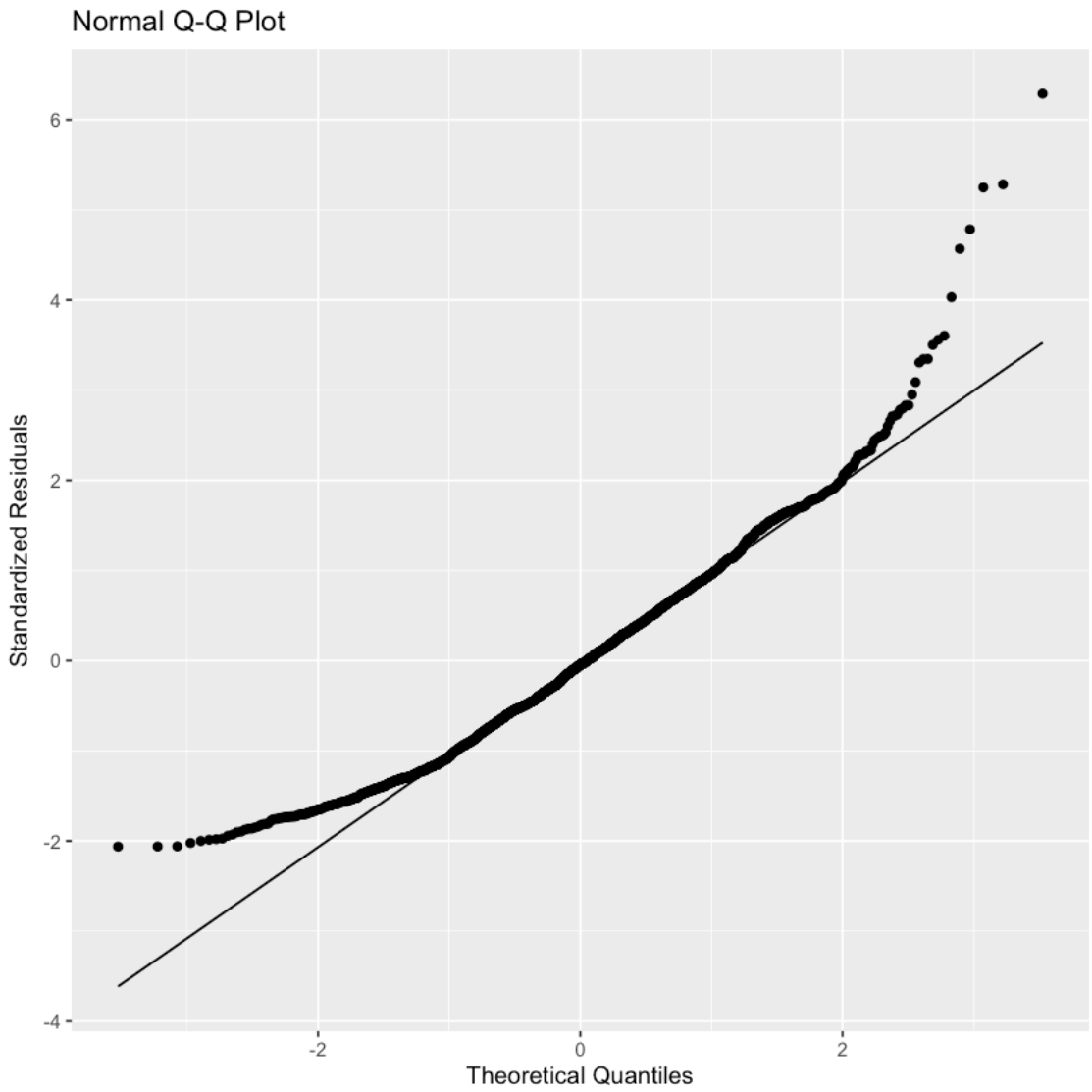
1.0.4 Model Fitting

We first try an additive model with all features and no interaction.

[129]:
```
full_model_wo_int <- lm(Total.Volume ~ type + AveragePrice + year + Month +␣
↪region , data = df_filtered)
```

the model summary shows typeorganic, typeorganic and most of year/month region are significant. And our Adjusted R-squared is 0.8553 which is a high value. We need residual and qq plot to check our model.

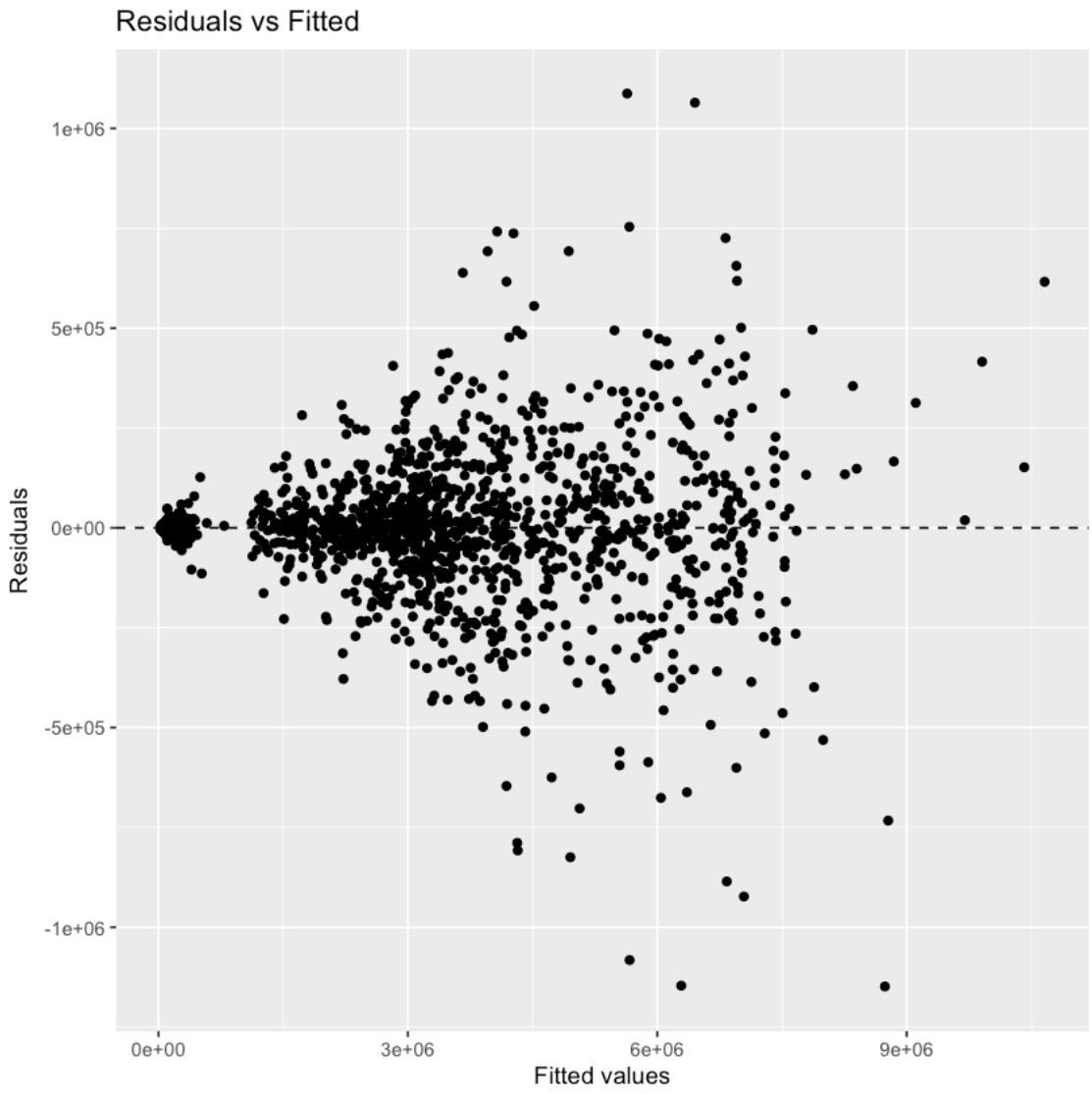


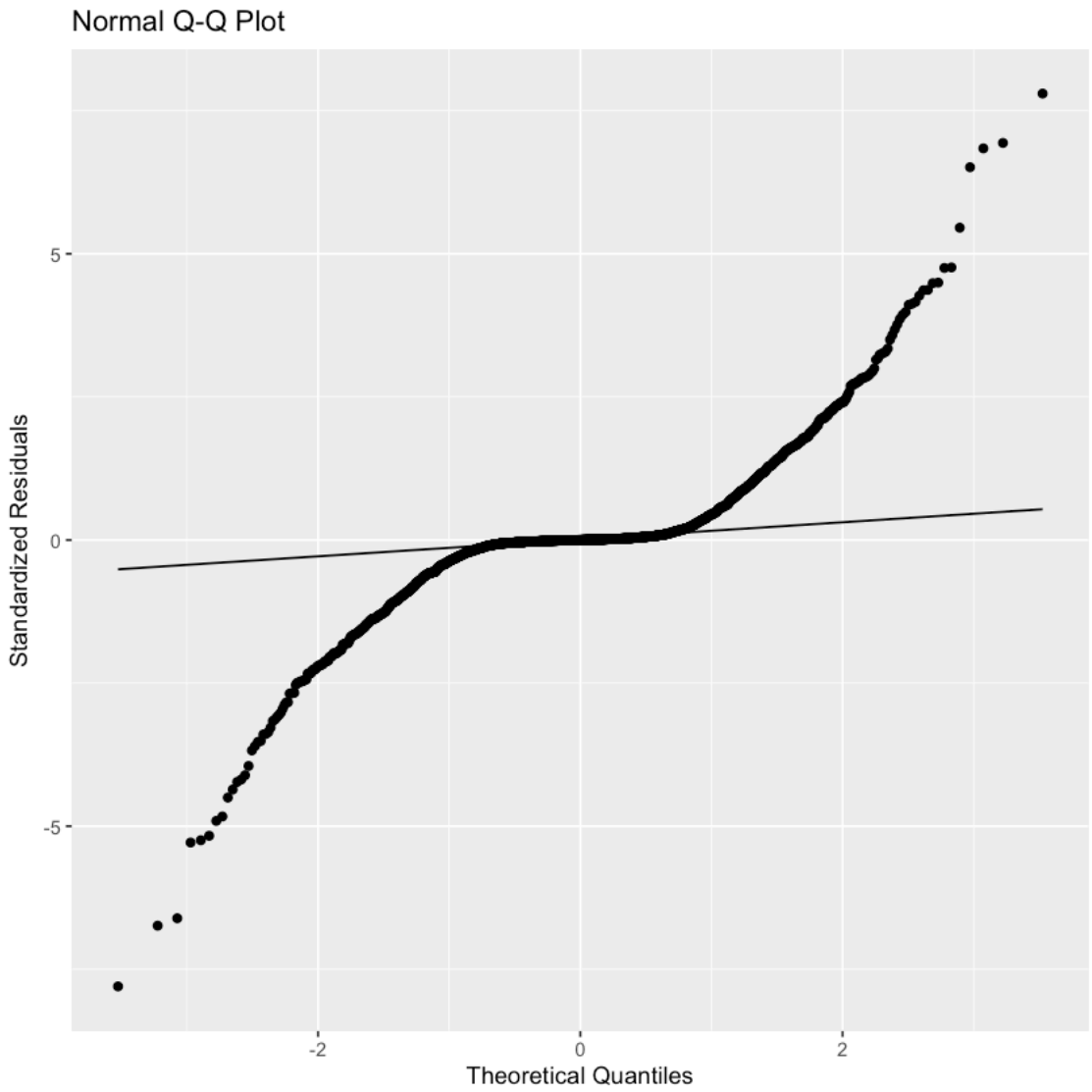


As we can see, when we use a model with all features but no interaction, there is a clear quadratic pattern on the residual plot, suggesting that the relationship between the independent and dependent variables is not linear and we need a more complicated model like adding interaction term. Combined with the findings we discovered in the EDA phase, we decide to try a model with interactions. Also the qq plot shows a right skew pattern, so we wanna a new model with some interaction terms.

[132]:
```
full_model_w_int <- lm(Total.Volume ~ type * AveragePrice * year * Month *␣
↪region , data = df_filtered)
```

The model with interaction give a 0.9922 Adjusted R-squared value which is very high, and we need residual and qq plot to investigate this model.

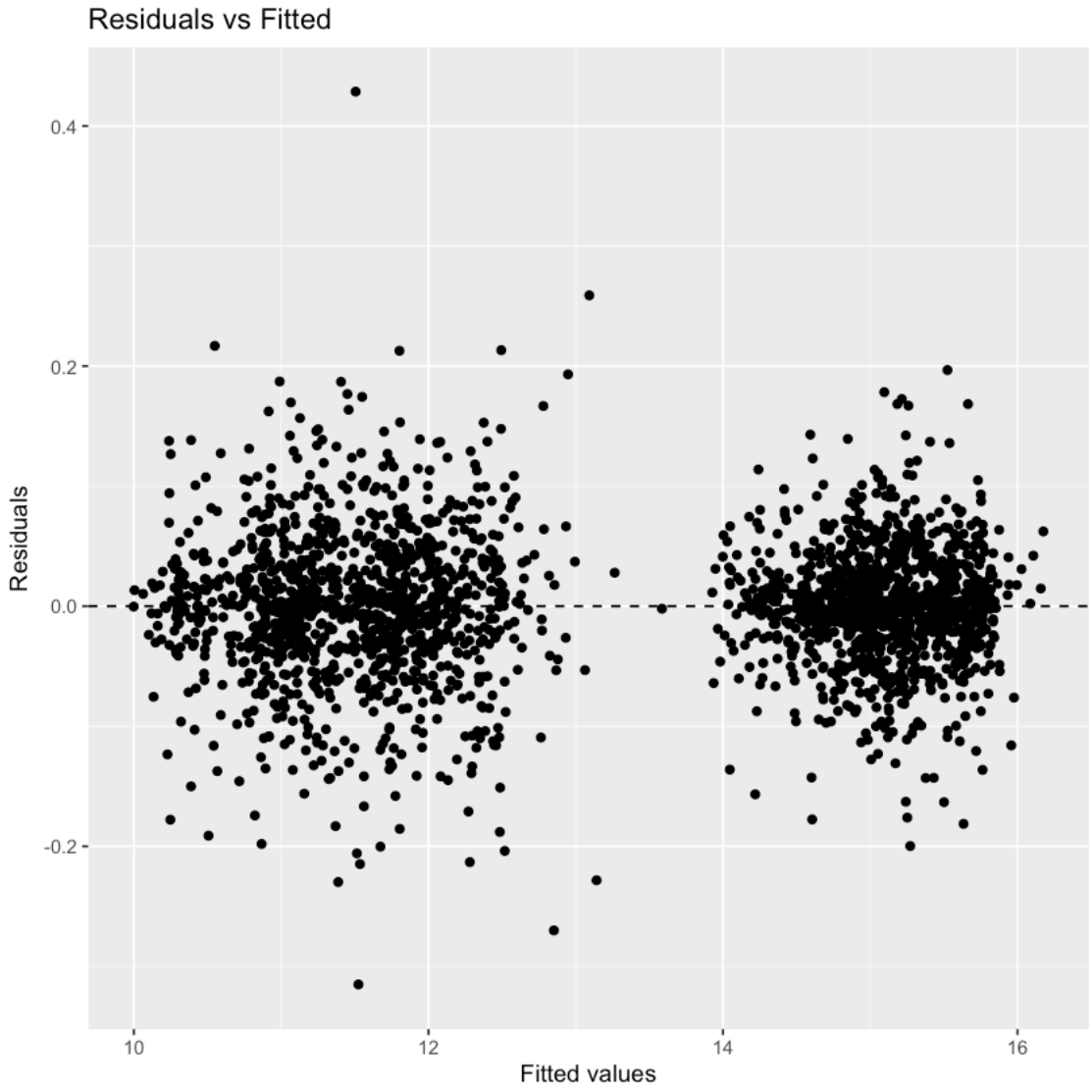

Residuals vs Fitted
1e+06
5e+05
0e+00
-5e+05
-1e+06
Residuals
0e+00
3e+06
6e+06
9e+06
Fitted values

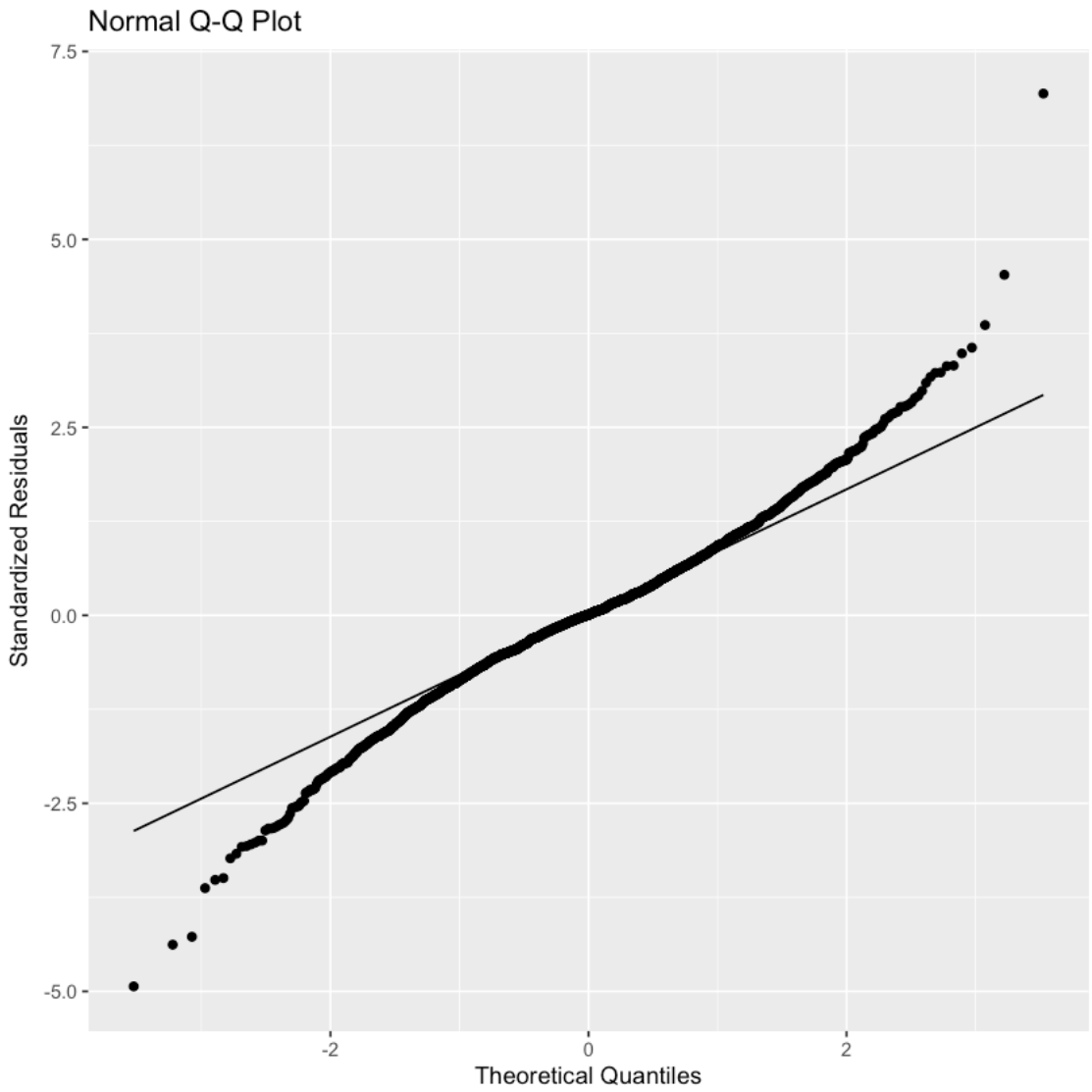


Incorporating interactions into the fitted model effectively eradicates the previously discerned patterns, signaling a marked improvement in fit. Nevertheless, a closer examination of the residual plot reveals a conspicuous cone shape, indicative of growing variance. Furthermore, it's worth noting that the QQ plot exhibits heavy tails, suggesting potential deviations from normality in the data distribution. To mitigate these issue, we propose applying a logarithmic transformation to the response variable, total volume.

[135]:
```
full_model_w_int_log <- lm(log(Total.Volume) ~ type * AveragePrice * year *␣
↪Month * region , data = df_filtered)
```

For our full model, we have Adjusted R-squared: 0.9983 is extremely high and indicates that the model explains almost all of the variability in the response variable. This suggests that the model fits the data very well.


Residuals vs Fitted
0.4
0.2
0.0
-0.2
Residuals
10
12
14
16
Fitted values



After transforming data, it seems like the residual plot doesn't show such obvious pattern which violate the linear regression assumptions like before, and the qq plot also show the model improvement with a little bit heavy tail.

But would it be necessary to have all of these interactions? In this next section we try a backward selection strategy, and fit five models each with one of the feature left out of the interaction. We then compare them based on AIC score and Mallow's Cp metric. The AIC of each model:

[138]:
```
partial_model_w_int_log1 <- lm(log(Total.Volume) ~ type + AveragePrice * region␣
↪* year * Month, data = df_filtered)
partial_model_w_int_log2 <- lm(log(Total.Volume) ~ AveragePrice + type * region␣
↪* year * Month, data = df_filtered)
partial_model_w_int_log3 <- lm(log(Total.Volume) ~ region + type * AveragePrice␣
↪* year * Month, data = df_filtered)
```

```
partial_model_w_int_log4 <- lm(log(Total.Volume) ~ year + type * AveragePrice *␣
 ↪region * Month, data = df_filtered)
partial_model_w_int_log5 <- lm(log(Total.Volume) ~ Month + type * AveragePrice␣
 ↪* region * year, data = df_filtered)
```

```
AIC for model without type interection =  -2467.607
AIC for model without AveragePrice interection =  -3710.834
AIC for model without region interection =  -517.3629
AIC for model without year interection =  -1050.002
AIC for model without Month interection =  -2867.676
AIC for full interactive model =  -4630.334
```

We found that the full model has the lowest AIC value comparing to others, which is consistent with the Adjusted R sqr result, hence we're keeping this as the best model.

1.0.5 Conclusion

In conclusion, in our analysis of the avocado market, it's clear that the interactions between variables—particularly region and type—are critical to understanding sales patterns. Our findings show that different U.S. regions exhibit unique sales trends, indicating that a 'one size fits all' approach to pricing and supply chain management is not optimal. The interaction terms in our model underscore the importance of tailoring strategies to regional market conditions.

The high adjusted R^2 value of our model is a testament to its strength and reliability, suggesting a significant proportion of the variance in avocado sales is explained by the variables we've considered. This robust fit opens up promising avenues for predicting avocado sales, giving suppliers and retailers a powerful tool to forecast market demand based on observable factors.

In practical terms, the implications of our research are substantial. By harnessing the predictive power of our model, businesses can set data-driven prices and adjust supply chain mechanisms to optimize sales and profitability in the dynamic avocado market. The ability to predict sales with a high degree of confidence can be a game-changer for decision-makers within the industry.