



HOLY GUACAMOLE

Name

Location

1. Personal Information Examples
2. Major Biology/Chemistry
3. Orientation Trainer
4. Trainer on Compliance and Policy
5. Military Officer

CONTENTS



01 Data Collection

02 Data Wrangling

03 Exploration

04 ANOVAs

05 Logistic Regression



Avocado Consumption

Comparing Prices and Volume of Avocados in three different US Regions

1. *Buffalo/Rochester, NY*
2. *Dallas/Ft Worth, TX*
3. *San Diego, CA*



Methods - Data Collection

- Downloaded from the Woz-U Github repo
- CSV file
- Imported using Pandas library
- Original source is from Kaggle



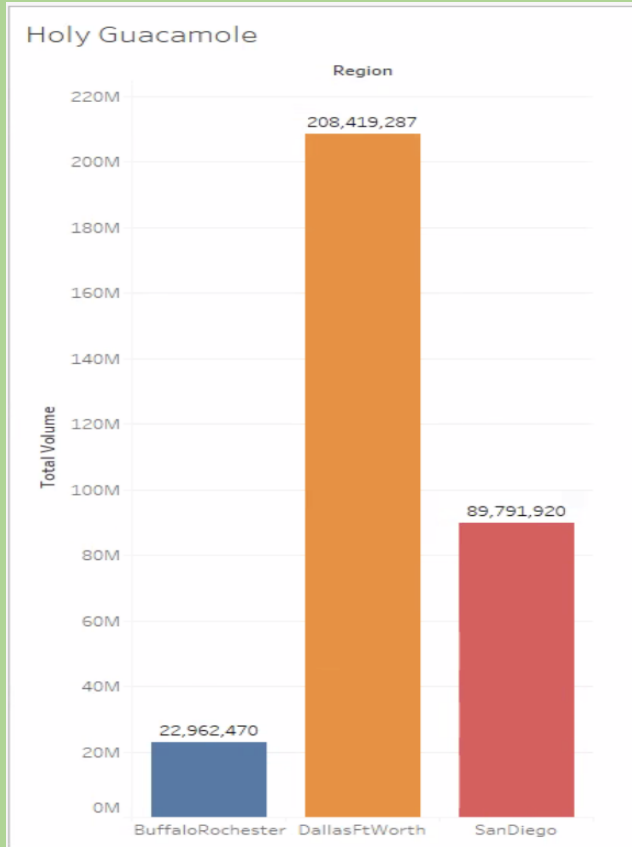
Data Wrangling

- Original dataset contained 18249 rows
- Dataset includes 54 regions
- Over 4 years (2015, 2016, 2017, 2018)
- A subset of the three regions was used in this analysis
- Sample sizes were the same for each region (338)

```
In [27]: df1.region.value_counts()
```

```
Out[27]: BuffaloRochester      338  
DallasFtWorth      338  
SanDiego      338  
Name: region, dtype: int64
```

Avocado Consumption per Region



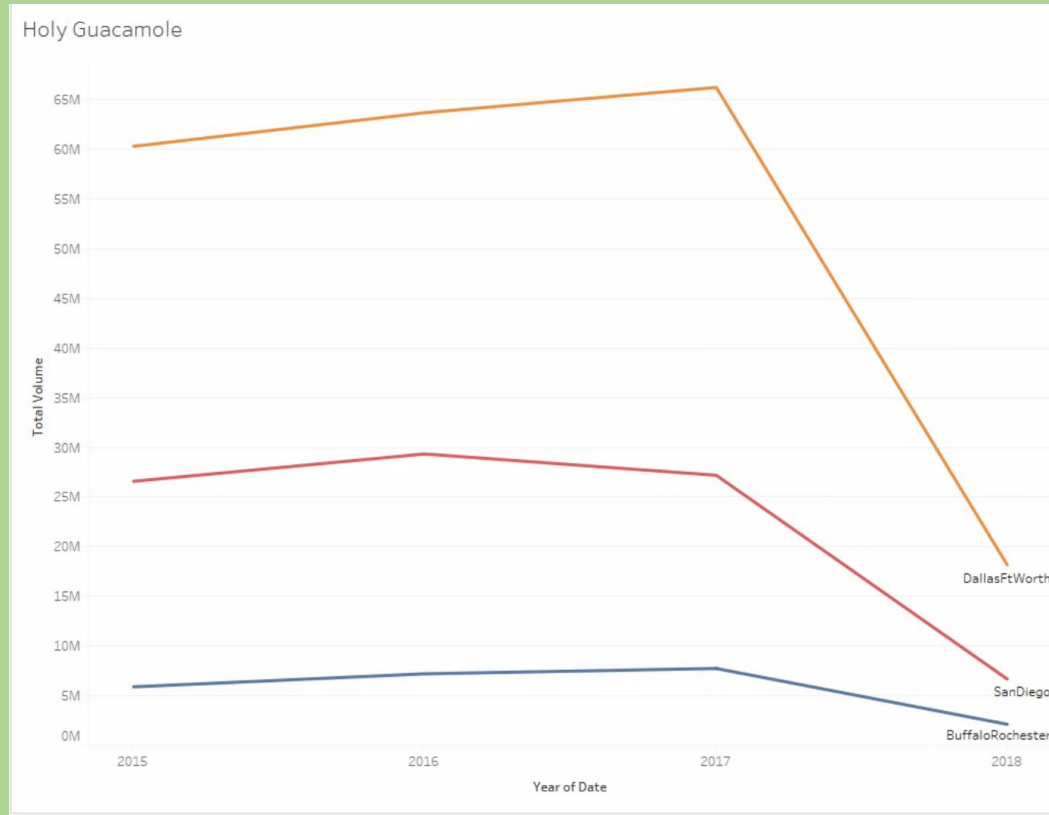
Avocado Price Over Time



**Average Avocado Price of
all 54 regions and all 4
years was \$1.41**

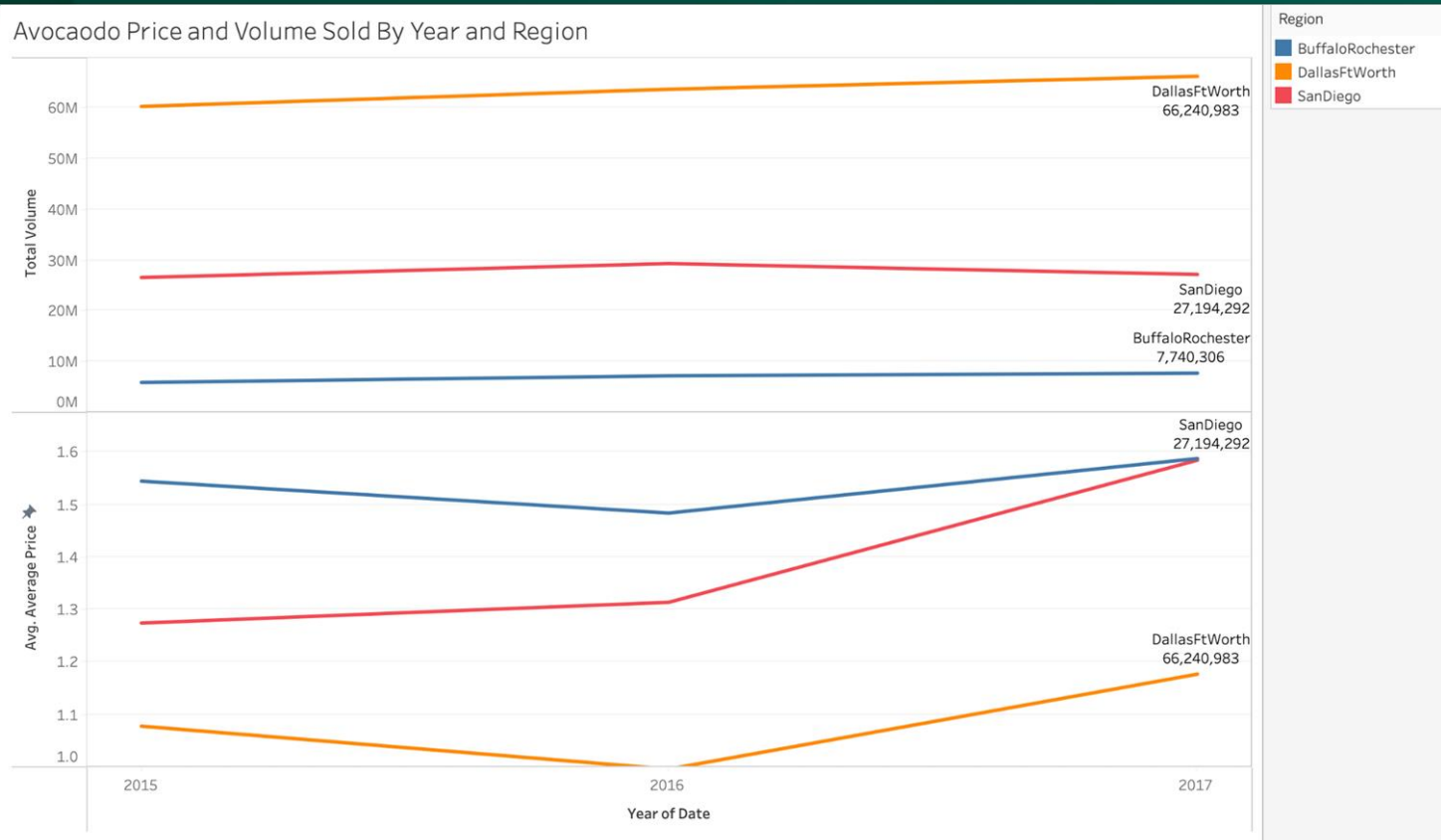


Avocado Consumption Over Time



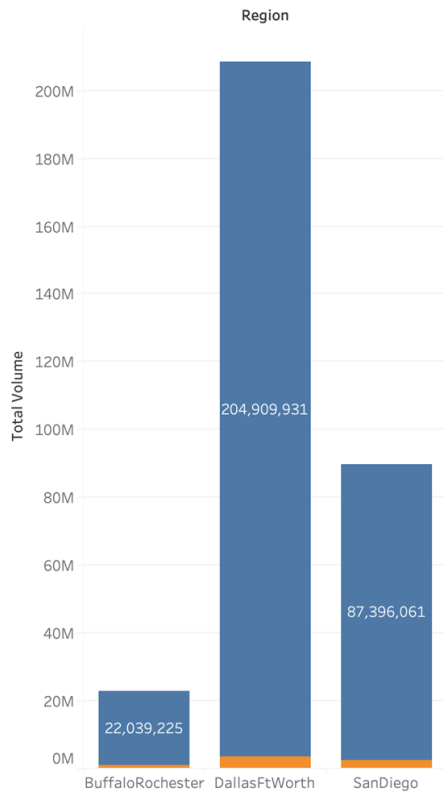
Data Exploration

Avocaado Price and Volume Sold By Year and Region



Data Exploration

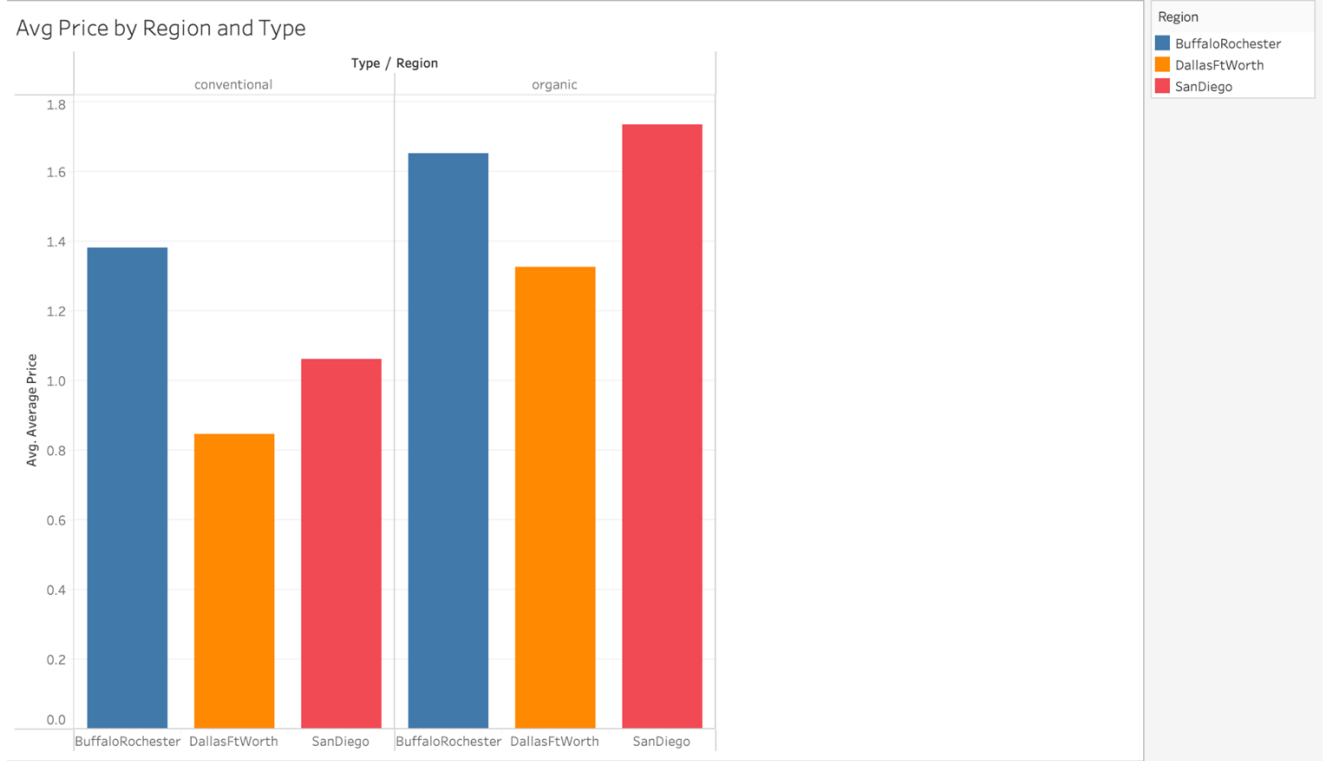
Total Volume by Region



Type

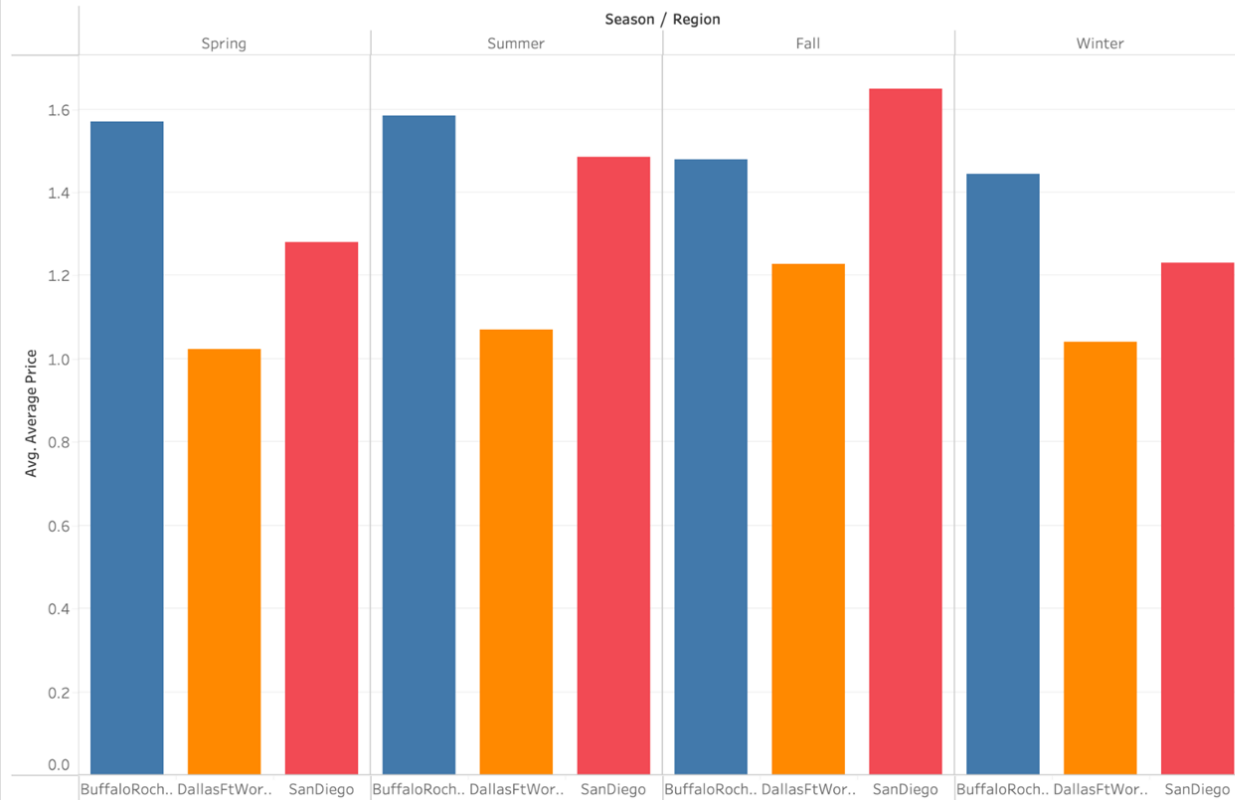
- conventional
- organic

Data Exploration



Data Exploration

Average Price by Season and Region



Region

BuffaloRochester

DallasFtWorth

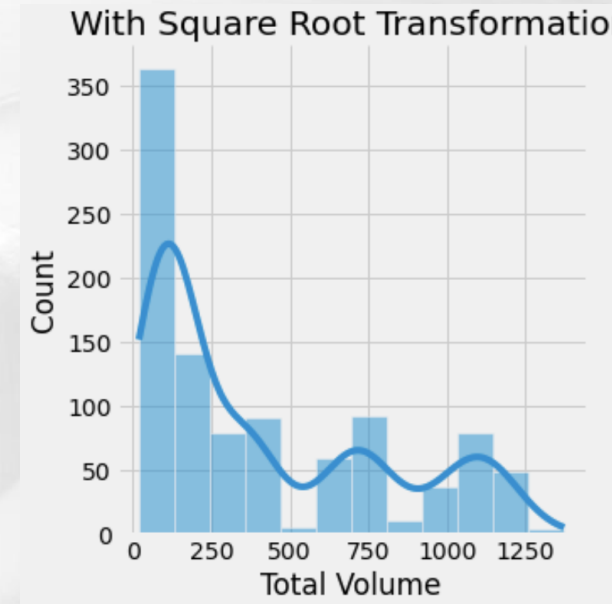
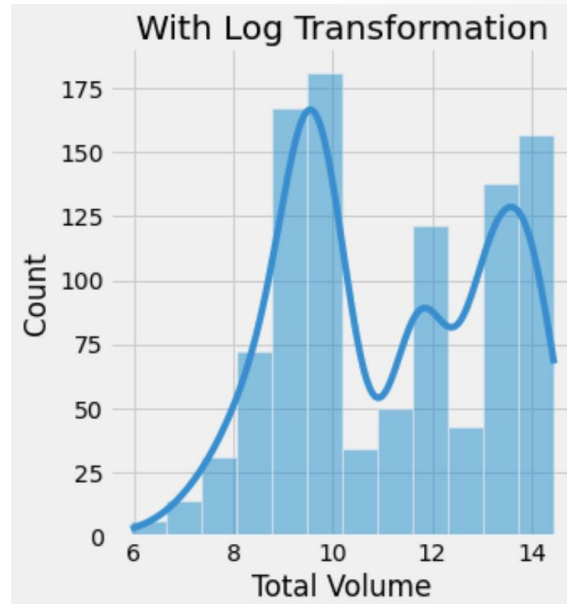
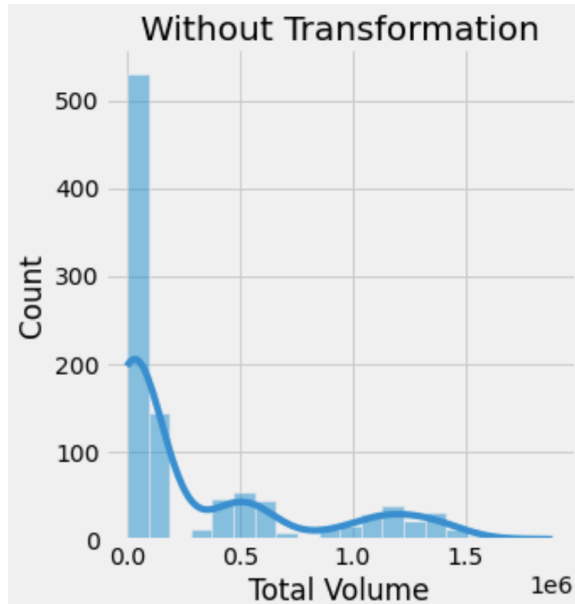
SanDiego

Were the average avocado prices different between the regions?



One-Way ANOVA

ANOVA Assumption of Normality

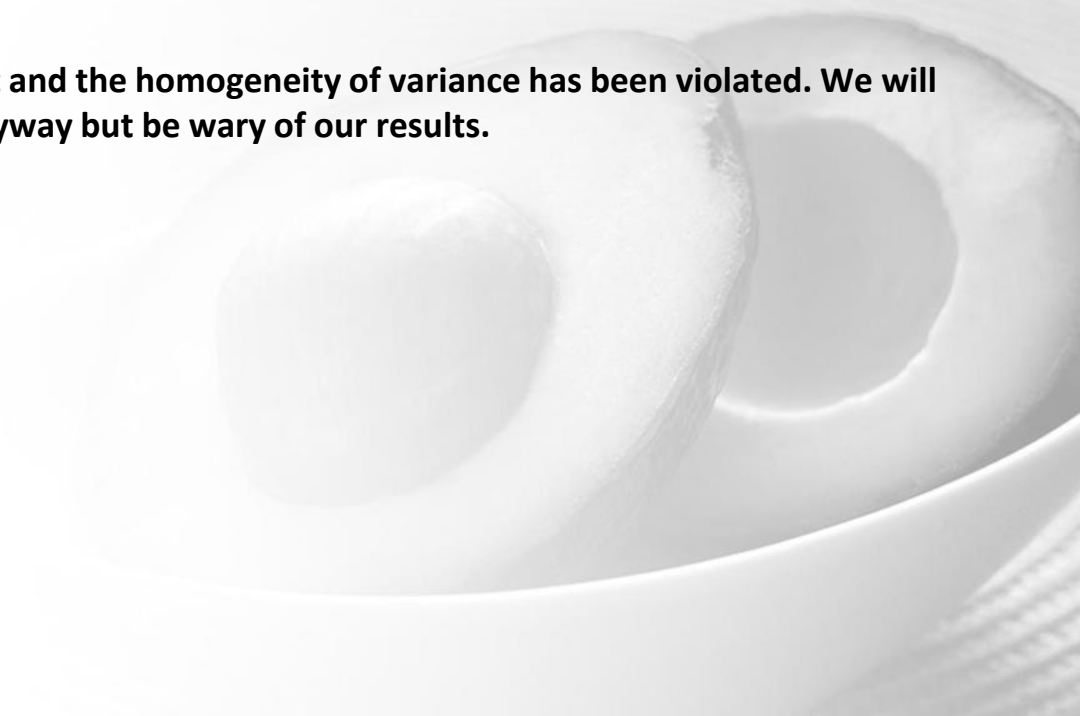


Log Transformation seems to normalize the most.

One-Way ANOVA

ANOVA Assumption of Normality - CONT

However, a Bartlett's Test result is significant and the homogeneity of variance has been violated. We will continue anyway but be wary of our results.



One-Way ANOVA

ANOVA Assumptions

- Sample size has a least 20 cases per independent variable
- No overlap between groups (regions) and region testings are not related
- Sphericity is not applicable because it's not a between subjects design

One-Way ANOVA Comparing Region and Volume Sold

ANOVA Results

- ANOVA result was significant serving as evidence that there was a difference between average avocado consumption (volume sold) of the three regions
- There was also a significant difference in consumption BETWEEN each region

Total Volume	
region	
BuffaloRochester	67936.302959
DallasFtWorth	616625.108136
SanDiego	265656.567130

One-Way ANOVA Comparing Season and Price

ANOVA Results

- Assumption of normality was met
- Homogeneity of Variance was violated, so made corrections using R software
- ANOVA was significant serving as evidence prices differ per season

	SeasonR <chr>	Mean <dbl>
1	Fall	1.45
2	Spring	1.29
3	Summer	1.38
4	Winter	1.24

	Fall	Spring	Summer
Spring	3.0e-05	-	-
Summer	0.33582	0.09003	-
Winter	1.4e-09	0.59198	0.00032

Predicting Organic or Non-Organic with Logistic Regression

Logistic Regression Results

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	506	0
1	1	507

Accuracy : 0.999

95% CI : (0.9945, 1)

No Information Rate : 0.5

P-Value [Acc > NIR] : <2e-16

Kappa : 0.998

McNemar's Test P-Value : 1

Sensitivity : 0.998

Specificity : 1.000

- The model resulted in 99% accuracy.
- There was only one false-positive resulting in a sensitivity of 99.8%.
- The average price of organic avocados was \$1.57.
- The median price of organic avocados was \$1.52.
- The average price of conventional avocados was \$1.10.
- The median price of conventional avocados was \$1.07.



THANK
YOU

