Fruit Prices 2020

`~~~~~~~~~~~~~~~~~~~~`

Exploratory Analysis

Shay Korhorn, skorhorn@bellarmine.edu
Alexander James Rallo, ajrallo@bellarmine.edu
Anthony Streib, astreib@bellarmine.edu

I. INTRODUCTION

Short description of the data set including a reference to where it can be found and why you chose it.

The data set we chose is about fruit prices and our goal for this project was to compare different fruits, their prices, yields, etc. Also, some fruits had different types of them such as normal or frozen, so we also compared the prices of those as well. We chose this dataset because all of us are healthy eaters who think diving into this dataset will help us pick the smartest product to purchase and consume.

II. DATA SET DESCRIPTION

Narrative summary of the data set: e.g. this data set contains 398 samples with 7 columns with various data types. A complete listing is shown in **Table 1**. For data types you want to indicate two things (nominal, ordinal, interval, or ratio) and the Pandas data type. For example, age might be ratio/int32. For missing data, indicate what percentage of data from that column are missing. Ensure you check to for NaN, NA, or any other indicators that actually mean missing data.

This data set contains 496 samples with 8 columns with various data types such as Nominal, Ordinal, Interval, and Ratio. There is no missing data from any column.

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Fruit	Nominal	0%
Form	Ordinal	0%
Retail Price	Interval	0%
Retail Price Unit	Ordinal	0%
Yield	Interval	0%
Cup Equivalent Size	Interval	0%
Cup Equivalent Unit	Ordinal	0%
Cup Equivalent Price	Ratio	0%

III. Data Set Summary Statistics

Narrative introduction to the section.

Table 2: Summary Statistics for XXX (name of dataset)

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Retail Price	*62*	2.61605	2.06812895	0.3604	1.15585	1.8684	3.577925	*10.5527*
Yield	*62*	0.87612903	0.17497873	0.46	0.7225	0.98	*1*	*1*

Cup Equivalent Size	*62*	1.704983871	2.949261842	0.1232	0.32245	0.3638	*0.5401*	*8*
Cup Equivalent Price	*62*	0.91974032	0.49196014	0.2292	0.5793	0.89525	*1.150475*	*3.07*

There should be a table for **EACH** categorical variable.

Table 3: Proportions for XXX (n=yyy)

Category	*Frequency*	*Proportion (%)*
Fruit		
Form		
Retail Price Unit		
Cup Equivalent Unit		

Below this are the correct proportions for each category.

|
|
|
|
\/

Here is the Proportions for Fruit:

Apples	1	1.61%
Apples, applesauce	1	1.61%
Apples, frozen concentrate	1	1.61%
Apples, ready-to-drink	1	1.61%
Apricots	2	3.23%
Apricots, packed in juice	1	1.61%
Apricots, packed in syrup or water	1	1.61%
Bananas	1	1.61%
Berries, mixed	1	1.61%
Blackberries	2	3.23%
Blueberries	2	3.23%
Cantaloupe	1	1.61%
Cherries	1	1.61%
Cherries, packed in syrup or water	1	1.61%
Clementines	1	1.61%
Cranberries	1	1.61%
Dates	1	1.61%
Figs	1	1.61%
Fruit cocktail, packed in juice	1	1.61%
Fruit cocktail, packed in syrup or water	1	1.61%
Grapefruit	1	1.61%

Grapefruit, ready-to-drink	1	1.61%
Grapes	1	1.61%
Grapes (raisins)	1	1.61%
Grapes, frozen concentrate	1	1.61%
Grapes, ready-to-drink	1	1.61%
Honeydew	1	1.61%
Kiwi	1	1.61%
Mangoes	2	3.23%
Nectarines	1	1.61%
Oranges	1	1.61%
Oranges, frozen concentrate	1	1.61%
Oranges, ready-to-drink	1	1.61%
Papaya	2	3.23%
Peaches	2	3.23%
Peaches, packed in juice	1	1.61%
Peaches, packed in syrup or water	1	1.61%
Pears	1	1.61%
Pears, packed in juice	1	1.61%
Pears, packed in syrup or water	1	1.61%
Pineapple	2	3.23%
Pineapple, frozen concentrate	1	1.61%
Pineapple, packed in juice	1	1.61%
Pineapple, packed in syrup or water	1	1.61%
Pineapple, ready-to-drink	1	1.61%
Plum	1	1.61%
Plum (prune), ready-to-drink	1	1.61%
Plum (prunes)	1	1.61%
Pomegranate	1	1.61%
Pomegranate, ready-to-drink	1	1.61%
Raspberries	2	3.23%
Strawberries	2	3.23%
Watermelon	1	1.61%

Here is the Proportions for Form:

Canned	12	19.35%
Dried	9	14.52%
Fresh	24	38.71%
Frozen	6	9.68%

Juice	11	17.74%

Here is the Proportions for Retail Price Unit:

per pint	11	17.74%
per pound	51	82.26%

Here is the Proportions for Cup Equivalent Unit:

fluid ounces	11	17.74%
pounds	51	82.26%

After you summarize the categorical variables, generate a correlation matrix for all continuous variables (not categorical – this doesn't make sense)

Table 4: Correlation Table/Tables

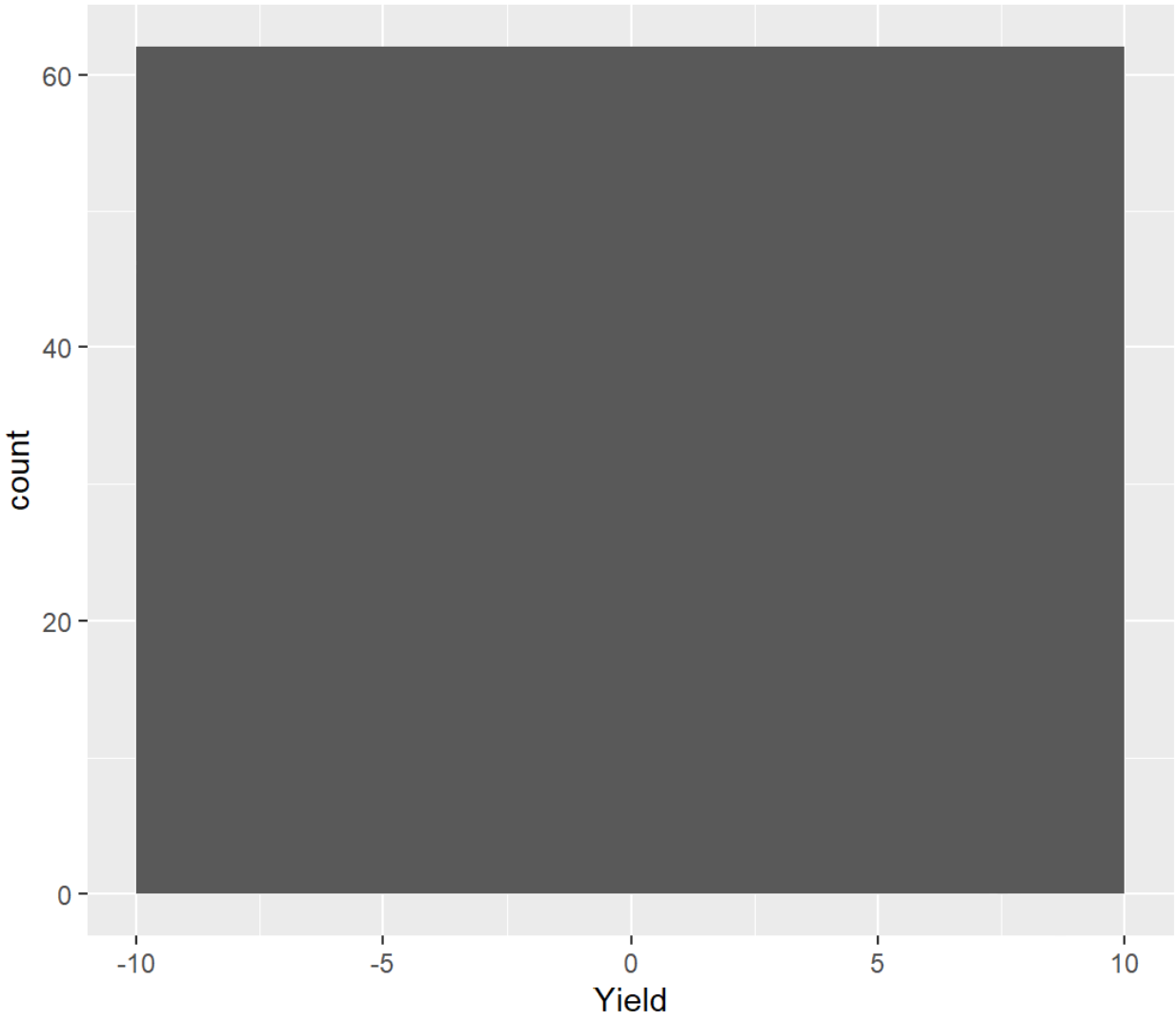
	Retail Price	*Yield*
Retail Price	*1.0000000*	*0.3633054*
Yield	*0.3633054*	*1.0000000*

IV. DATA SET GRAPHICAL EXPLORATION

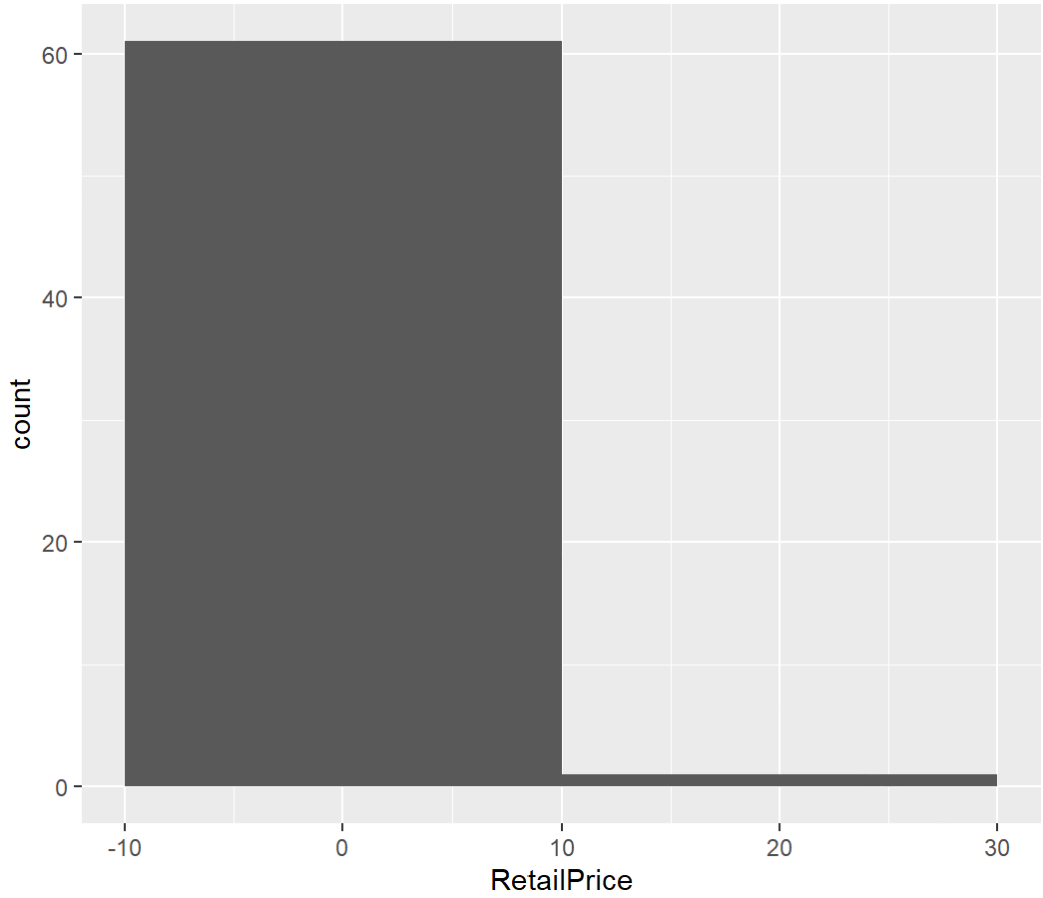
Narrative introduction to the section. In each section below, indicate any interesting distributions, anomalies, imbalance, etc. that you notice.

A. Distributions
B. ScatterPlots / Pairwise Plots (continuous variables)
C. Barcharts (categorical variables)
D. Other Plots - don't skimp – there are likely other plots that would be useful that I haven't already specified. Include those in this section.

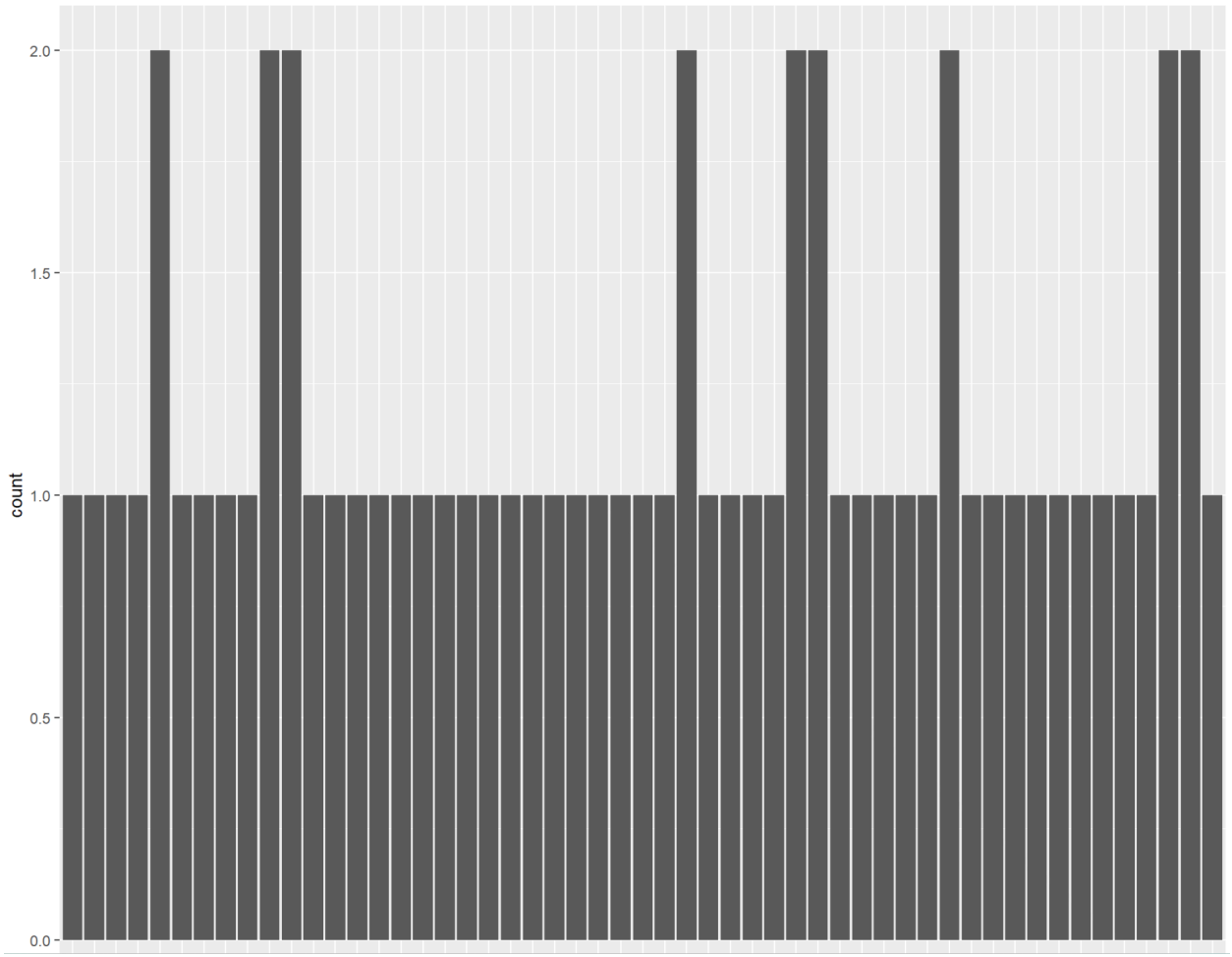
All figures should be cited formatted like this and mentioned in the text.



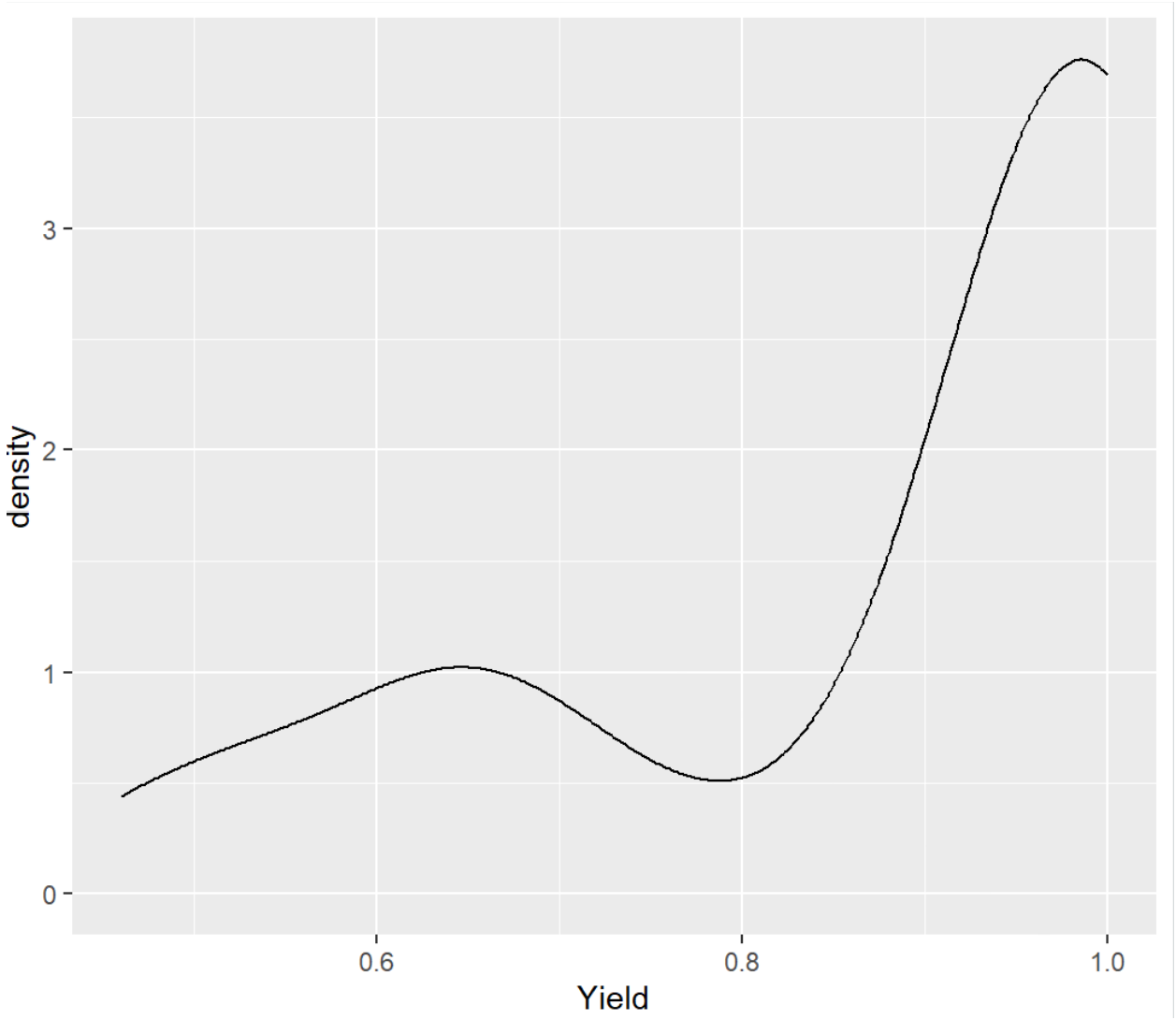
First graph: For our first graph we used R Studio functions to calculate the count of the Yield for each Fruit category. As shown above, you can see that there is a count for each Fruit category as it takes up the entire graph. This proves there is no missing data for both columns.



Second Graph: For our second graph we used R Studio to calculate the count of different Retail Prices compared to count of other columns. As shown above, there were 50% with different Retail Prices than the rest of the other 50%.



Third Graph: For our third graph, we compared count and each Fruit Category with different forms, such as Fresh, Canned, Juice, Dried, and Frozen, from the Form category. Therefore, as shown above, you can see there are 9% of the Fruits with different categories than the rest.



Fourth Graph: For our fourth and final graph, we compared the density versus the yield. Using a density graph we can acquire the most common number for the yield category. As shown above, you can see that 1.0 is the most common number for the yield of each fruit category. We can also see that around 0.65 there is a slight rise in increase for common yield, and then it dips at 0.8 and immediately jumps to 1.0, being the most common.

V. SUMMARY OF FINDINGS

Finish up with a paragraph or two of summarizing your findings about this data set.

To summarize our findings, we have found comparisons from the types of fruit compared to the fruit category, such as more apples will be fresh than any other of the categories. Another find from this dataset is the correlation between the yield and retail price, the higher the yield of the fruit, the higher the retail price will be due to yield meaning the most you will get out of each harvest of fruit, this supported our hypothesis. Overall, we have learned a

lot from this project, such as getting more in depth with R studio, and learning more about fruits, their different categories, and the reasoning behind their prices.