R Assignment - Pizza EDA

Alexander Thiersch, GWU Intro to Data Science DATS 6101

2022-02-18

Contents

```
# 1. Answer each question using in words/paragraph.
# 2. DO NOT use comments inside code blocks (like here) to answer anything. Those are for notes between
# coders/yourself. They will be ignored, and not counted as answers.
# 3. Keep the line/option    knitr::opts_chunk$set(warning = F, results = "hide", message = F)
# You can temporarily turn this on/off and use other option while you are working on the homework if it
# The submitted work should have this option selected instead.
# 4. All charts/graphs/tables should have appropriate captions.
# 5. You may want to use the ezids::outlierKD2 function to handle outliers
# 6. Your grade is also determined by the style. Even if you answers everything correctly, but the html
```

HW Assignment - EDA

This pizza ingredient dataset is from data.world (@sdhilip) The variables are:

Variable	Definition
brand	Pizza brand
id	ID
mois	Amount of water per 100 grams in the sample
prot	Amount of protein per 100 grams in the sample
fat	Amount of fat per 100 grams in the sample
ash	Amount of ash per 100 grams in the sample
sodium	Amount of sodium per 100 grams in the sample
carb	Amount of carbohydrates per 100 grams in the sample
cal	Amount of calories per 100 grams in the sample

As with all your work in this class, knit the RMD file into HTML, zip it with the RMD, and submit the zip file on Blackboard.

Table 2: Table: Statistics summary.

	brand	id	mois	prot	fat	ash	sodium
Min	Length:300	Min. :14003	Min. :25.0	Min. : 6.98	Min. : 4.4	Min. :1.17	Min. :0.250
Q1	Class :character	1st Qu.:14094	1st Qu.:30.9	1st Qu.: 8.06	1st Qu.:14.8	1st Qu.:1.45	1st Qu.:0.45
Median	Mode :character	Median :24020	Median :43.3	Median :10.44	Median :17.1	Median :2.22	Median :0.4
Mean	NA	Mean :20841	Mean :40.9	Mean :13.37	Mean :20.2	Mean :2.63	Mean :0.669
Q3	NA	3rd Qu.:24110	3rd Qu.:49.1	3rd Qu.:20.02	3rd Qu.:21.4	3rd Qu.:3.59	3rd Qu.:0.70
Max	NA	Max. :34045	Max. :57.2	Max. :28.48	Max. :47.2	Max. :5.43	Max. :1.790

Compose your answers using inline R code instead of using the code-block output as much as you can. Coder's comments inside code blocks are never graded.

Pizza Ingredient Dataset

Question 1: Import Dataset

Import the Pizza.csv dataset into R.

```
pizza = data.frame(read.csv("Pizza.csv"))
```

Question 2: Total Number of Observations

How many data points are there?

There are a total of 300 data points.

You can use the `length()` or `nrow()` function.

```
nrow(pizza)
```

[1] 300

Question 3: Dataset Summary Statistics

Look at the summary statistics of the dataset.
Use the `xkablesummary()` function?

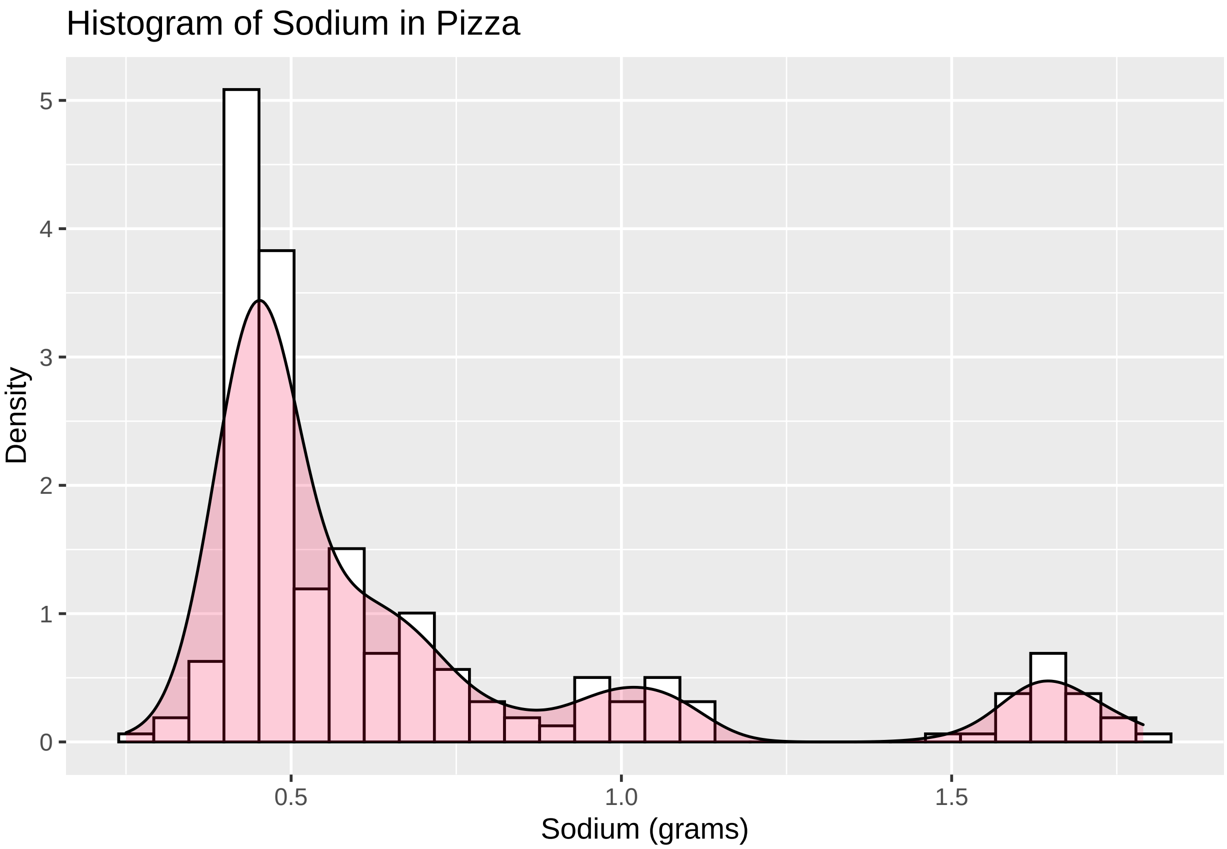
```
ezids::xkablesummary(pizza)
```

Question 4: Histograms, Boxplots, and QQ-Plots

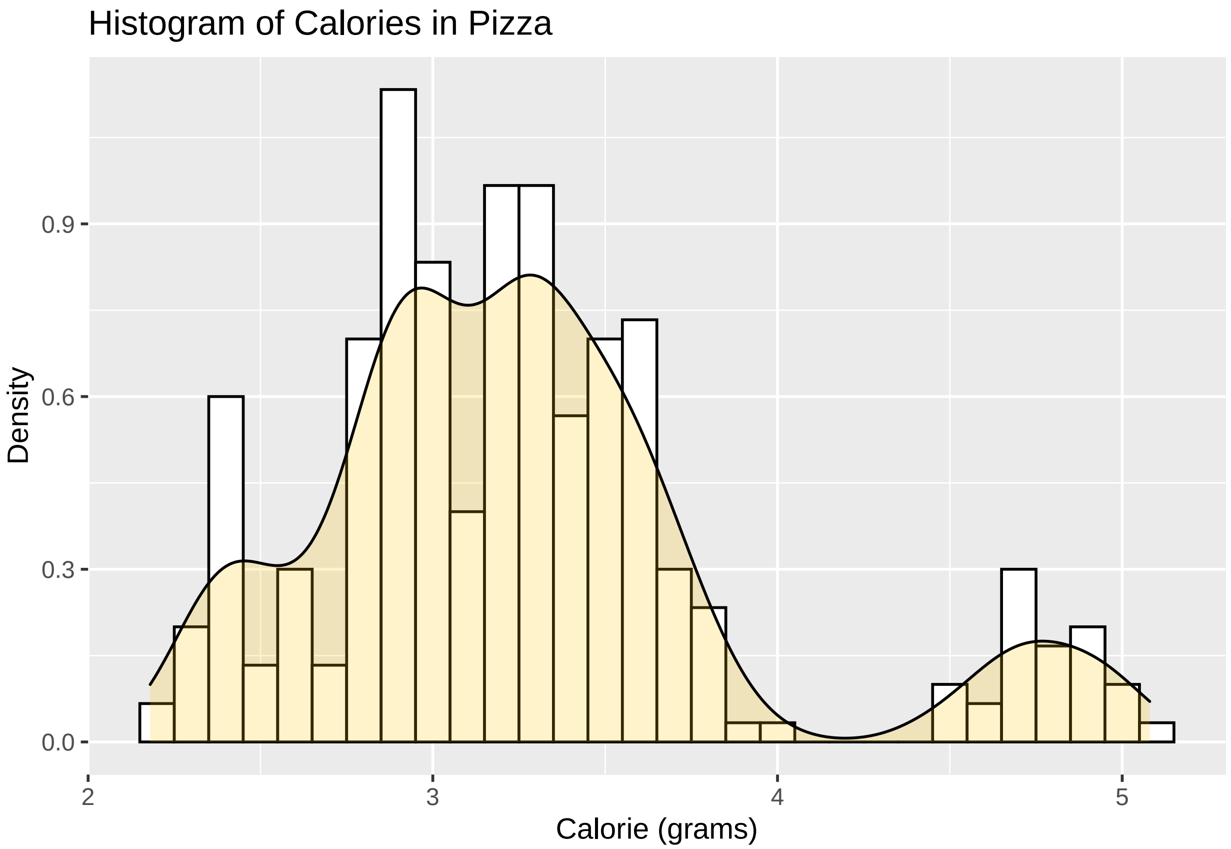
Make Plots - 1
For sodium and calorie, make histograms, boxplots (using `ggplot()`), and QQ-plots (just regular `qqnorm()` function). Make sure all plots have appropriate titles, x- and y- labels, units on the axes if applicable. It is also much nicer to add some color to your charts instead of plain old black and white. For chart titles, if no appropriate title you can think of, just use y vs x. Don't get mixed up (somehow I find up to half of the presentations have the wrong ones). It is NEVER x vs y. **Always y vs x**.

#Histogram Plots for Sodium and Calories

```
#Histogram for Sodium
ggplot(pizza, aes(x=sodium)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#f50041") +
 ggtitle("Histogram of Sodium in Pizza") + xlab("Sodium (grams)") + ylab("Density")
```

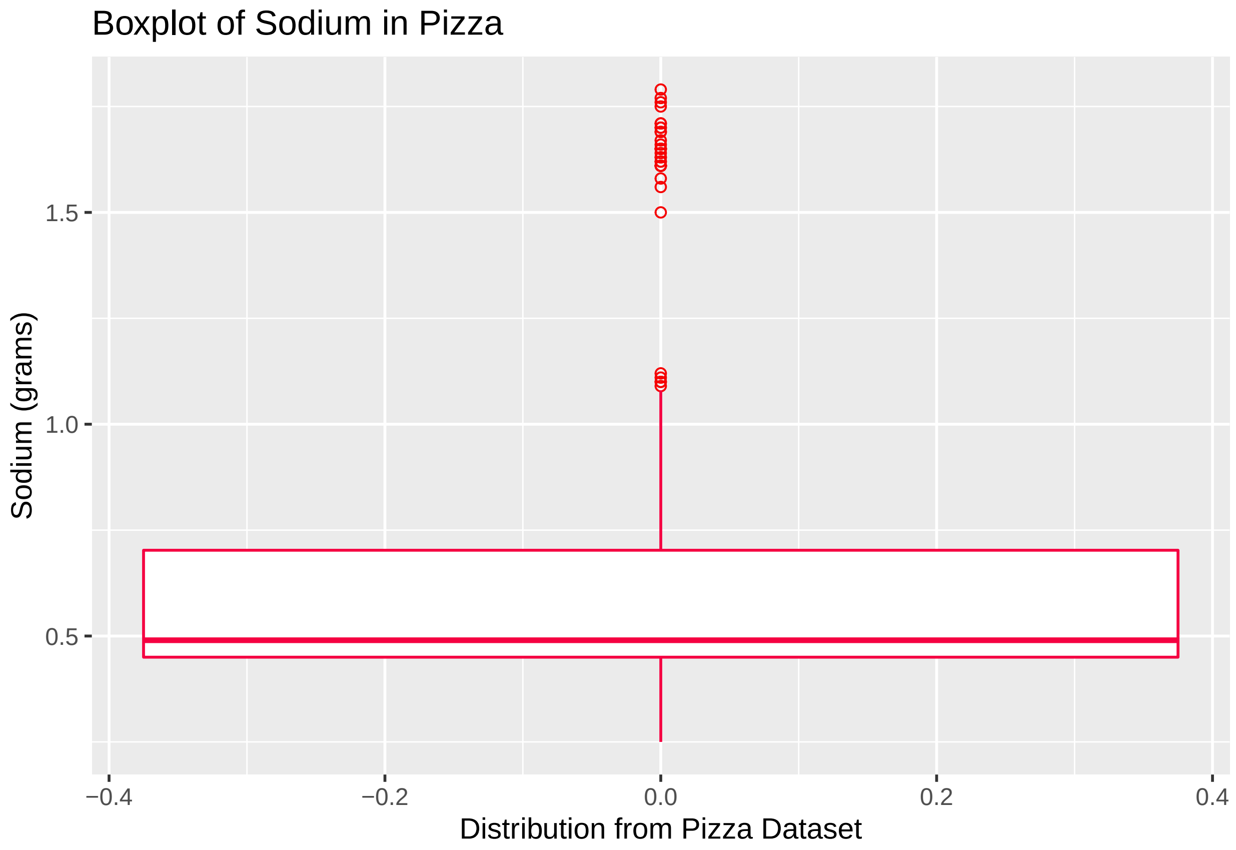


```
#Histogram for Calorie
ggplot(pizza, aes(x=cal)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#ffc400") +
 ggtitle("Histogram of Calories in Pizza") + xlab("Calorie (grams)") + ylab("Density")
```

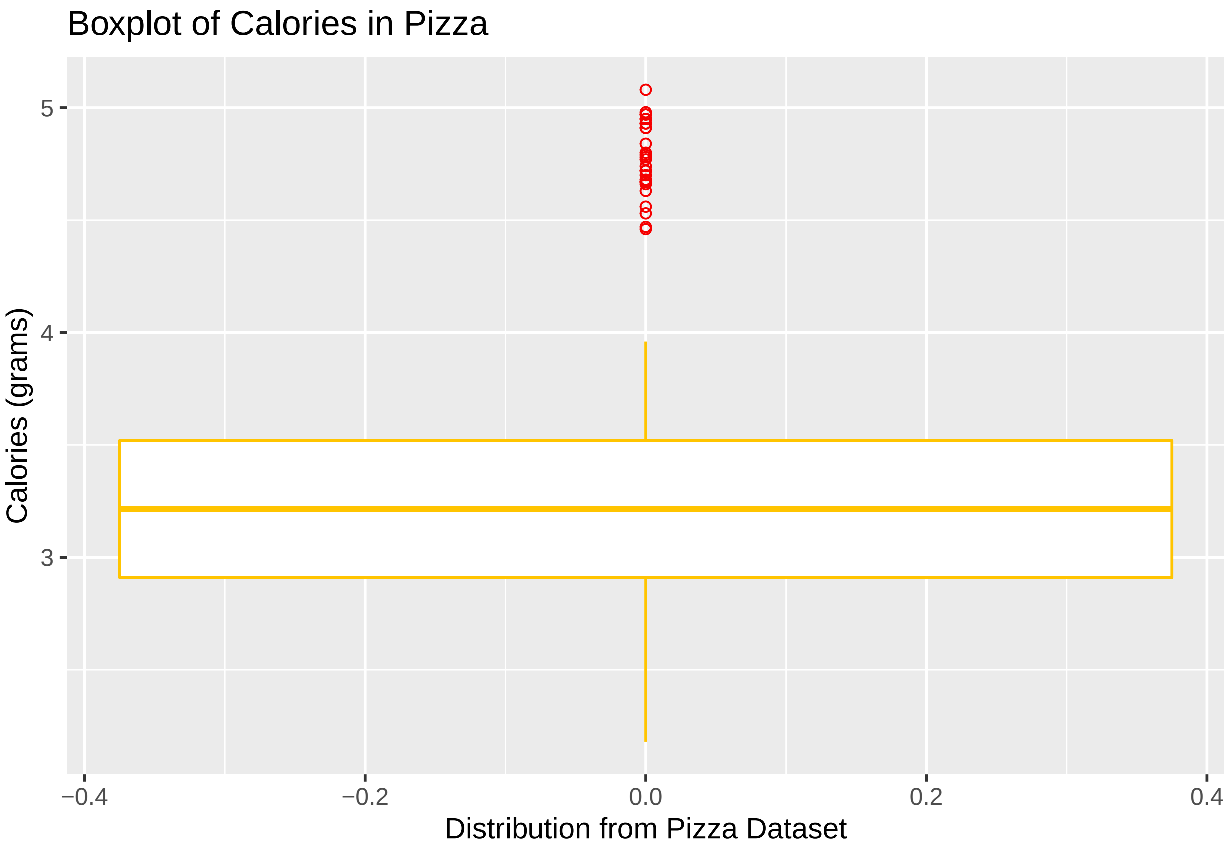


#Boxplots for Sodium and Calories

```
#Boxplot of Sodium
ggplot(pizza, aes(y=sodium))+
  geom_boxplot(fill='white', color="#f50041", outlier.colour="#f50000", outlier.shape=1) +
  ggtitle("Boxplot of Sodium in Pizza") + xlab("Distribution from Pizza Dataset") + ylab("Sodium (grams
```

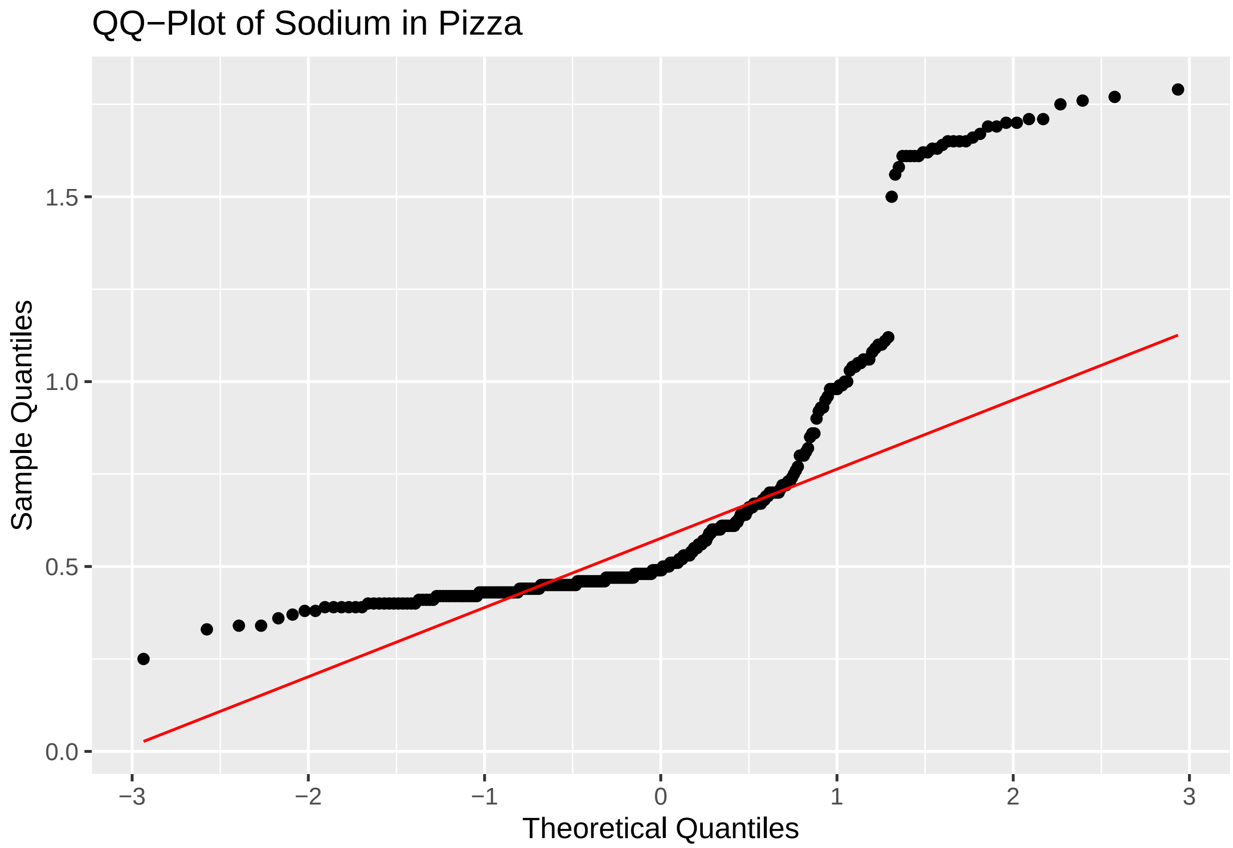


```
#Boxplot of Calories
ggplot(pizza, aes(y=cal))+
  geom_boxplot(fill='white', color="#ffc400", outlier.colour="#f50000", outlier.shape=1) +
  ggtitle("Boxplot of Calories in Pizza") + xlab("Distribution from Pizza Dataset") + ylab("Calories (g
```

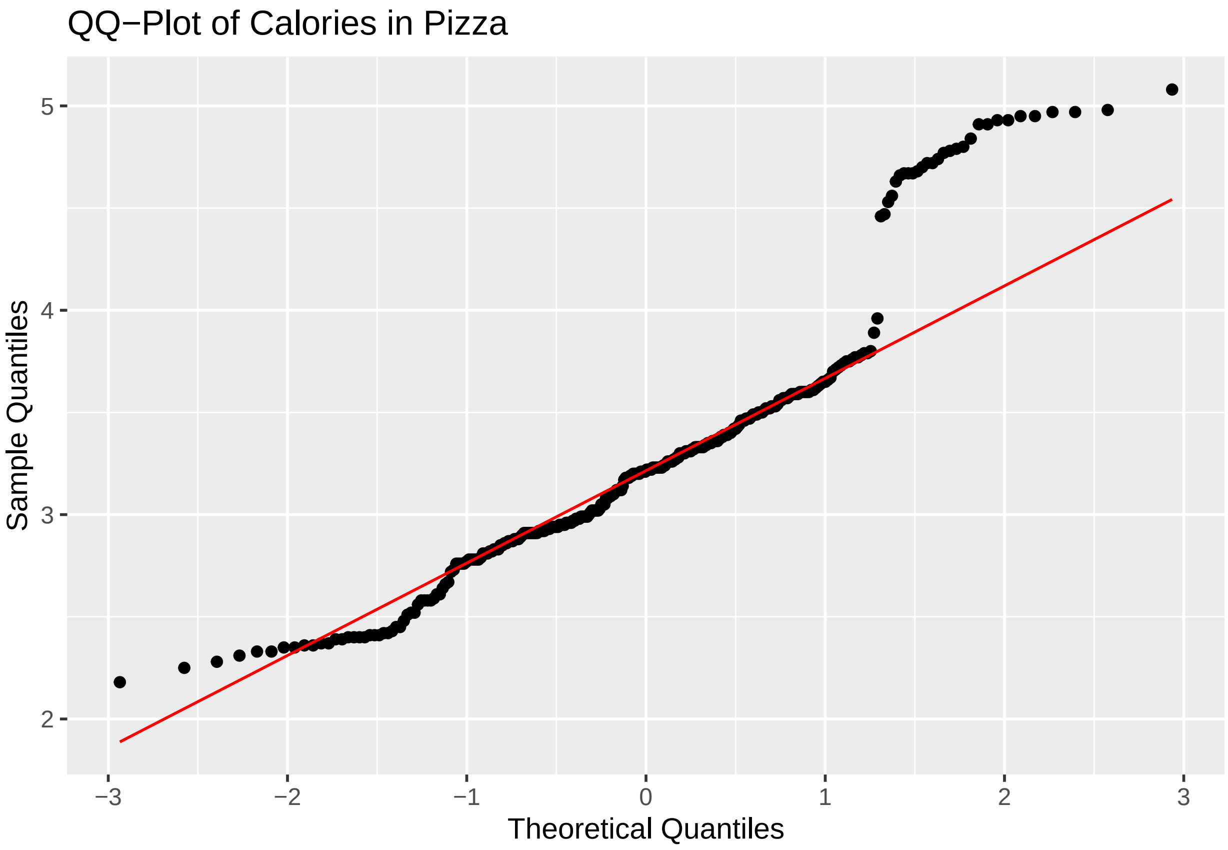


#QQ-Plots for Sodium and Calories

```
#Sodium QQ-Plot
ggplot(pizza, aes(sample = sodium)) +
  stat_qq() +
  stat_qq_line(col = "red") +
  ggtitle("QQ-Plot of Sodium in Pizza") + xlab("Theoretical Quantiles") + ylab("Sample Quantiles")
```



```
#Calorie QQ-Plot
ggplot(pizza, aes(sample = cal)) +
  stat_qq() +
  stat_qq_line(col = "red") +
  ggtitle("QQ-Plot of Calories in Pizza") + xlab("Theoretical Quantiles") + ylab("Sample Quantiles")
```

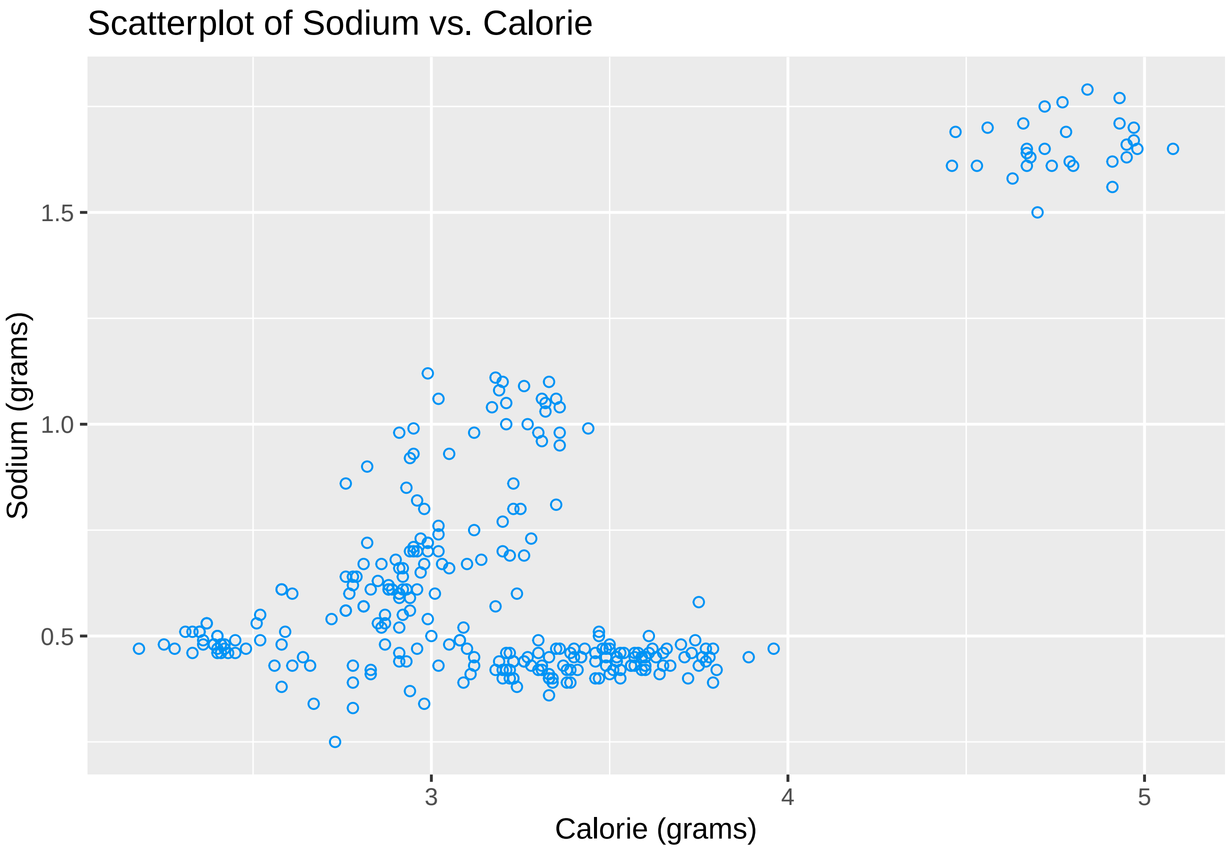


Question 5: Scatterplot - Sodium vs. Calorie

Make Plots - 2

Making a scatterplot (using `ggplot()`), between sodium and calorie, color by the brand. As always, give the plot appropriate title, axis labels, and make it look good.

```
ggplot(pizza, aes(x=cal, y=sodium)) + geom_point(color="#0093f5", shape=1) + ggtitle("Scatterplot of So
```



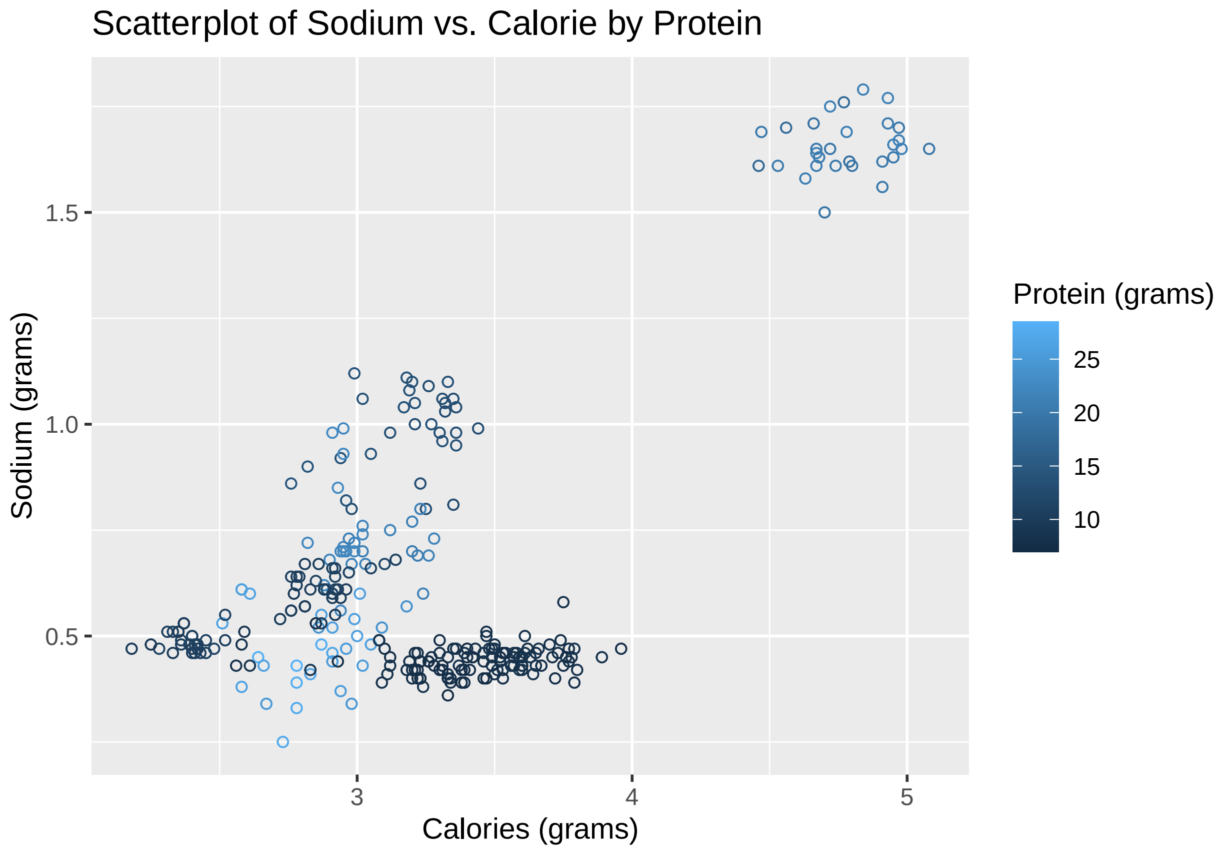
Question 6: Scatterplot - Sodium vs. Calorie by Protein

Make Plots - 3

Making another scatterplot (using `ggplot()`), between sodium and calorie, color by protein. What major difference do you see between this and the previous scatterplot?

The major difference between the two scatterplots is that the graph in Question 5 displays the relationship between two variables (sodium and calories) and the graph in Question 6 displays the relationship between three variables (sodium, calorie, and protein)

```
ggplot(pizza, aes(x=cal, y=sodium)) + geom_point(aes(color=prot), shape=1) + ggtitle("Scatterplot of Sod
```



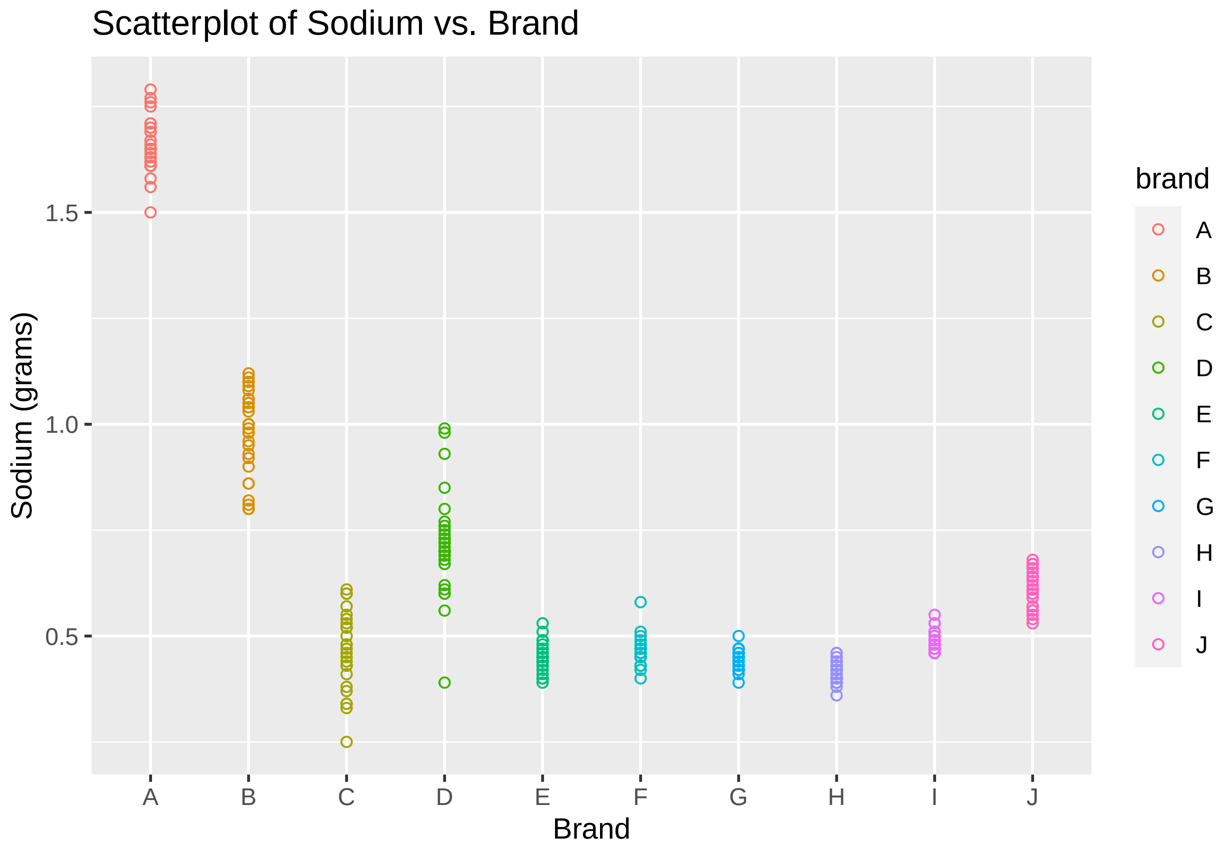
Question 7: Scatterplot - Sodium vs. Brand

Make Plots - 4

This time, make a plot with sodium as y, and brand as x. (What kind of plot would you choose?)

I would have chosen to plot the data using a boxplot. This is because the x-axis contains a categorical variable and the y-axiz contians numeric variable. A boxplot would be able to plot the distribution of sodium by brand of pizza.

```
ggplot(pizza, aes(x=brand, y=sodium)) + geom_point(aes(color=brand), shape=1) + ggtitle("Scatterplot of
```



Question 8: Removing Outliers

Outliers
Use the `ezids::outlierKD2()` function to remove the outliers for sodium, then run the function again to remove outliers on calories. Re-do the QQ-plots for these two variables. Do you see much improvements?

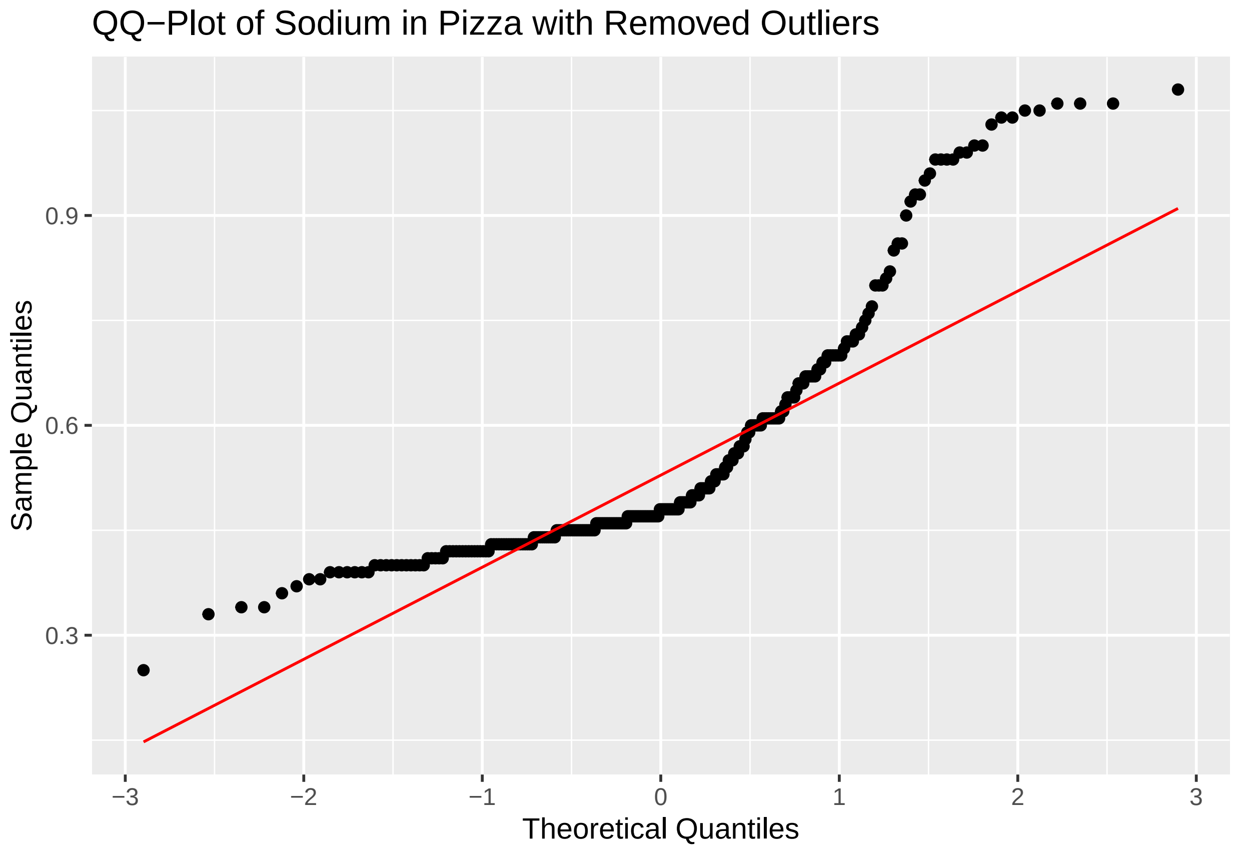
After removing the outliers from the sodium and calorie variables, there is some improvement in the QQ-plots for each variable. The QQ-plot of the Calories in Pizza exhibits the most improvement. This is indicated by the data points aligning more evenly with the standard normal variate, which is represented by the red line.

The QQ-Plot of Sodium in Pizza exhibits the least amount of improvement. Despite removing outliers, the sodium variable data significantly deviates from the straight line or the standard normal variate. This indicates that the data is highly skewed. However, this is a slight improvement from the original QQ-Plot of Sodium in Pizza in which the data exhibited even more deviation from the red line or standard normal variate.

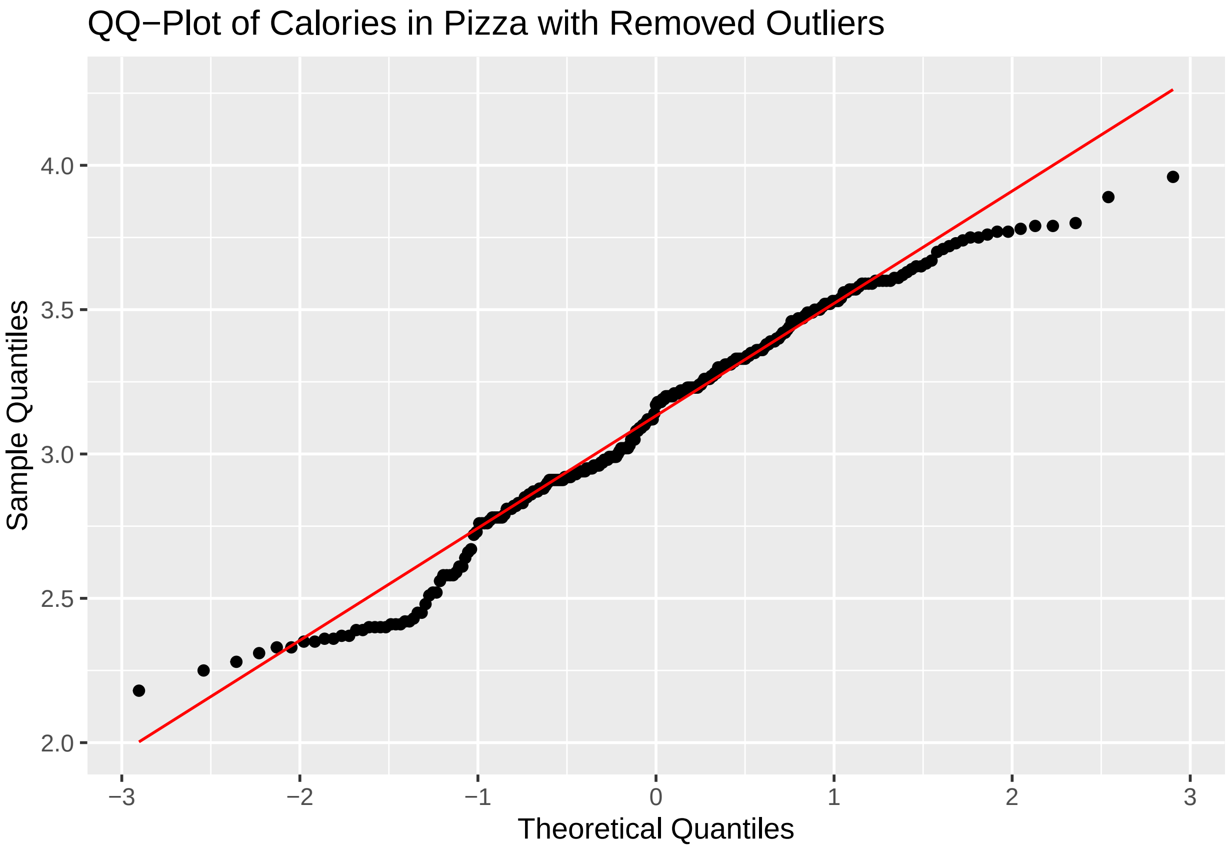
```
mod_sodium<-ezids::outlierKD2(pizza, sodium, rm=TRUE, boxplt=FALSE, histogram=FALSE, qqplt=FALSE)

mod_cal<-ezids::outlierKD2(pizza, cal, rm=TRUE, boxplt=FALSE, histogram=FALSE, qqplt=FALSE)

#Sodium QQ-Plot
ggplot(mod_sodium, aes(sample = sodium)) +
  stat_qq() +
  stat_qq_line(col = "red") +
  ggtitle("QQ-Plot of Sodium in Pizza with Removed Outliers") + xlab("Theoretical Quantiles") + ylab("Sa
```



```
#Calorie QQ-Plot
ggplot(mod_cal, aes(sample = cal)) +
  stat_qq() +
  stat_qq_line(col = "red") +
  ggtitle("QQ-Plot of Calories in Pizza with Removed Outliers") + xlab("Theoretical Quantiles") + ylab(
```


QQ−Plot of Calories in Pizza with Removed Outliers
Sample Quantiles
4.0
3.5
3.0
2.5
2.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles