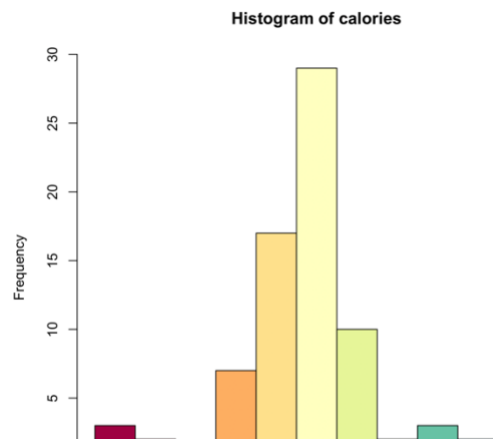


Cereal.csv – Overview from Data Science Perspective

For this project, I decided to use the Kaggle dataset from <https://www.kaggle.com/crawford/80-cereals>, a dataset of 77 different cereal and their qualities as explained on the website as well:

- **name:** Name of cereal
- **mfr:** Manufacturer of cereal
 - A = American Home Food Products;
 - G = General Mills
 - K = Kelloggs
 - N = Nabisco
 - P = Post
 - Q = Quaker Oats
 - R = Ralston Purina
- **type:**
 - cold
 - hot
- **calories:** calories per serving
- **protein:** grams of protein
- **fat:** grams of fat
- **sodium:** milligrams of sodium
- **fiber:** grams of dietary fiber
- **carbo:** grams of complex carbohydrates
- **sugars:** grams of sugars
- **potass:** milligrams of potassium
- **vitamins:** vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
- **shelf:** display shelf (1, 2, or 3, counting from the floor)
- **weight:** weight in ounces of one serving
- **cups:** number of cups in one serving
- **rating:** a rating of the cereals (Possibly from Consumer Reports)

For each variable, I'm going to do some exploration to see what is going on with the data before asking any questions about how the variables could be related to each other and what conclusions we can make about different types of cereal. In this dataset we can see visually that every name is a feature/row with attributes to analyze so I'm not going to analyze the names for this project.



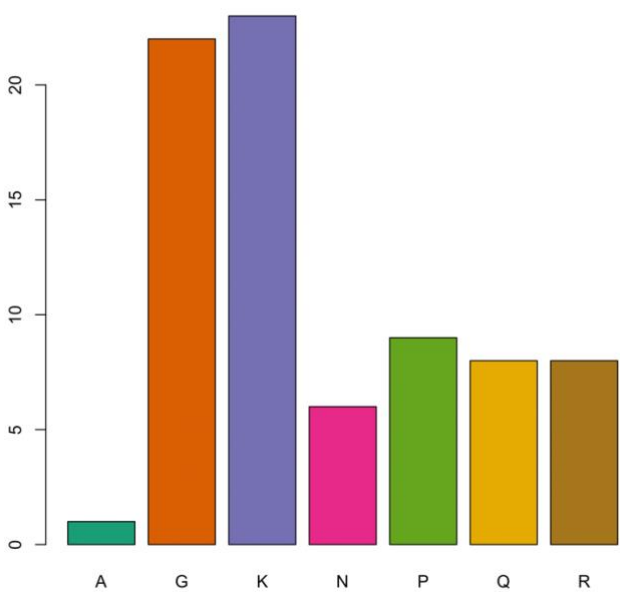
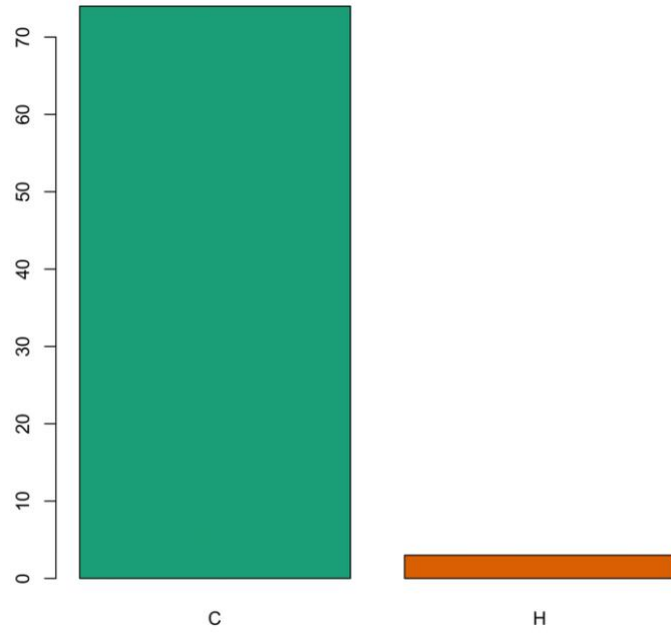
calories		n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
		77	0	11	0.933	106.9	19.86	70	90	100	110	110	124	140

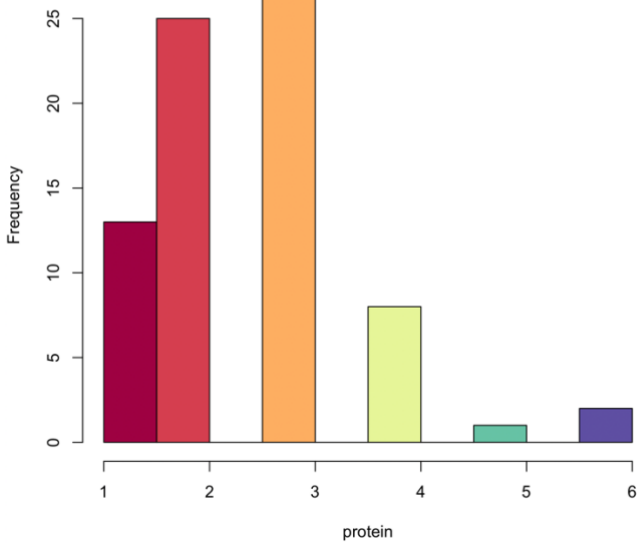
lowest : 50 70 80 90 100, highest: 120 130 140 150 160

Value	50	70	80	90	100	110	120	130	140	150	160
Frequency	3	2	1	7	17	29	10	2	3	2	1
Proportion	0.039	0.026	0.013	0.091	0.221	0.377	0.130	0.026	0.039	0.026	0.013

I'm going to be taking a look at the distribution of the each of the variables, starting with an example, calories. As we see above, from the histogram distribution of the calories per serving, a majority of cereals lie between 80-120 calories per serving. This makes sense as when calculated using R, the mean are 106.8831 (as seen above as well from the describe() in R). This gives us a good idea of how the entire data set looks like if we were to see it through the number of calories per serving.

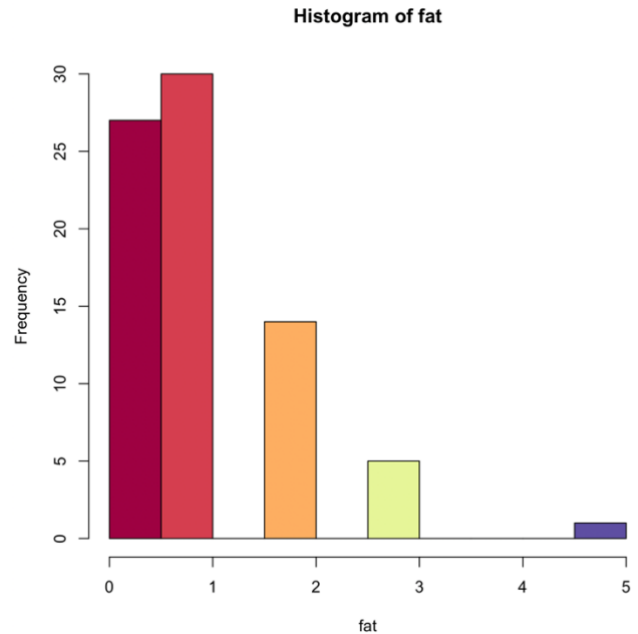
Just like I did above, I'm now going to repeat the same process for the rest of the variables and insert them in a table below to see them all together as seen below. Because not all of them are continuous variables, I will be using histograms and bar charts for all of them to see the distribution visually for each variable. On the right I wrote a description of what I learned about the variable through the histogram/bar chart and on the left would be the name of the variable and how I categorized the variable for data analysis for part 2.

<div>mfr</div> <div>Categorical variable</div>	<div></div> <div><div>mfr</div><table><tr><td>n</td><td>missing</td><td>distinct</td></tr><tr><td>77</td><td>0</td><td>7</td></tr></table><div>lowest : A G K N P, highest: K N P Q R</div><table><tr><td>Value</td><td>A</td><td>G</td><td>K</td><td>N</td><td>P</td><td>Q</td><td>R</td></tr><tr><td>Frequency</td><td>1</td><td>22</td><td>23</td><td>6</td><td>9</td><td>8</td><td>8</td></tr><tr><td>Proportion</td><td>0.013</td><td>0.286</td><td>0.299</td><td>0.078</td><td>0.117</td><td>0.104</td><td>0.104</td></tr></table></div>	n	missing	distinct	77	0	7	Value	A	G	K	N	P	Q	R	Frequency	1	22	23	6	9	8	8	Proportion	0.013	0.286	0.299	0.078	0.117	0.104	0.104	<div>Mean: N/A</div> <div><p>This is a categorical variable, so I created a bar graph using the bar plot function. General Mills and Kelloggs cereals make up approximately 58% of all the cereals here (definitely a majority).</p></div>
n	missing	distinct																														
77	0	7																														
Value	A	G	K	N	P	Q	R																									
Frequency	1	22	23	6	9	8	8																									
Proportion	0.013	0.286	0.299	0.078	0.117	0.104	0.104																									
<div>type (Cold vs Hot)</div> <div>Categorical variable</div>	<div></div>	<div>Mean: N/A</div> <div><p>Again, this is a categorical variable and we can see her how a majority of the cereals are cold (74/77) cereals. I feel that we don't have enough data for hot cereals in order to compare and contrast cold vs hot cereals in any way but this is good to know before making comparisons of cereals with variables.</p></div>																														

	<pre>type n missing distinct 77 0 2 Value C H Frequency 74 3 Proportion 0.961 0.039</pre>																																											
<div>protein</div> <div>Continuous variable</div>	<div><div>Histogram of protein</div><table><tr><th colspan="7">protein</th></tr><tr><th>n</th><th>missing</th><th>distinct</th><th>Info</th><th>Mean</th><th colspan="2">Gmd</th></tr><tr><td>77</td><td>0</td><td>6</td><td>0.912</td><td>2.545</td><td colspan="2">1.166</td></tr></table><p>lowest : 1 2 3 4 5, highest: 2 3 4 5 6</p><table><tr><th>Value</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th></tr><tr><th>Frequency</th><td>13</td><td>25</td><td>28</td><td>8</td><td>1</td><td>2</td></tr><tr><th>Proportion</th><td>0.169</td><td>0.325</td><td>0.364</td><td>0.104</td><td>0.013</td><td>0.026</td></tr></table></div>	protein							n	missing	distinct	Info	Mean	Gmd		77	0	6	0.912	2.545	1.166		Value	1	2	3	4	5	6	Frequency	13	25	28	8	1	2	Proportion	0.169	0.325	0.364	0.104	0.013	0.026	<p>Protein is measured in grams of protein per serving (and each serving size is different). I'm not exactly sure how reliable using the histogram information is yet since I haven't looked into the oz per serving size. But seeing from here regardless of that, most cereals per serving have between 2-3 grams of protein per serving. It looks like R is detecting this to be a categorical variable vs a quantitative variable due there being many options but I'm going to be treating it like a quantitative/continuous variable .</p>
protein																																												
n	missing	distinct	Info	Mean	Gmd																																							
77	0	6	0.912	2.545	1.166																																							
Value	1	2	3	4	5	6																																						
Frequency	13	25	28	8	1	2																																						
Proportion	0.169	0.325	0.364	0.104	0.013	0.026																																						

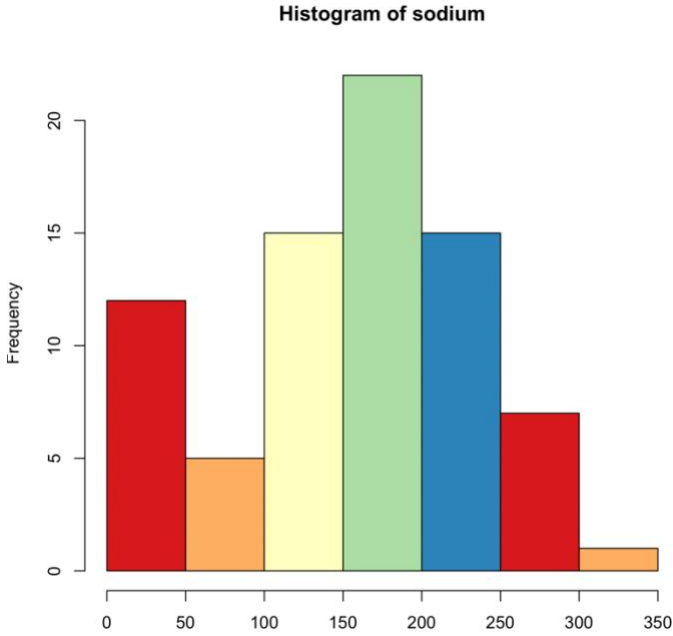
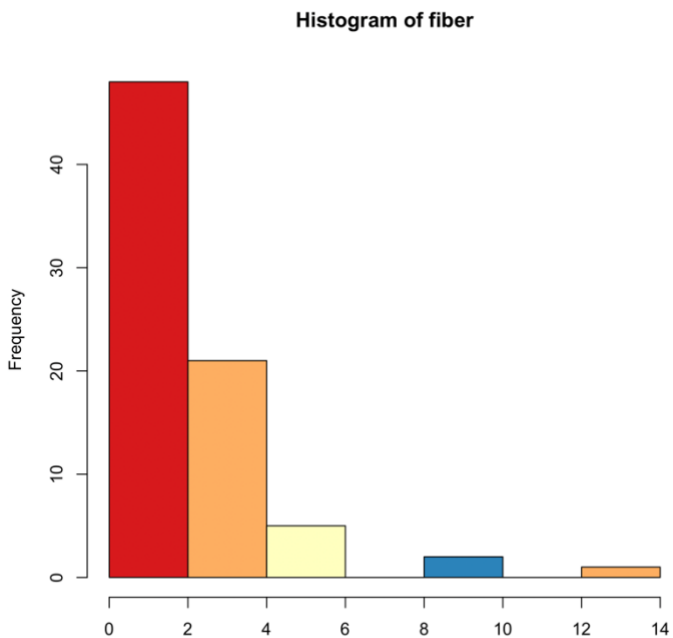
fat

Continuous
variable



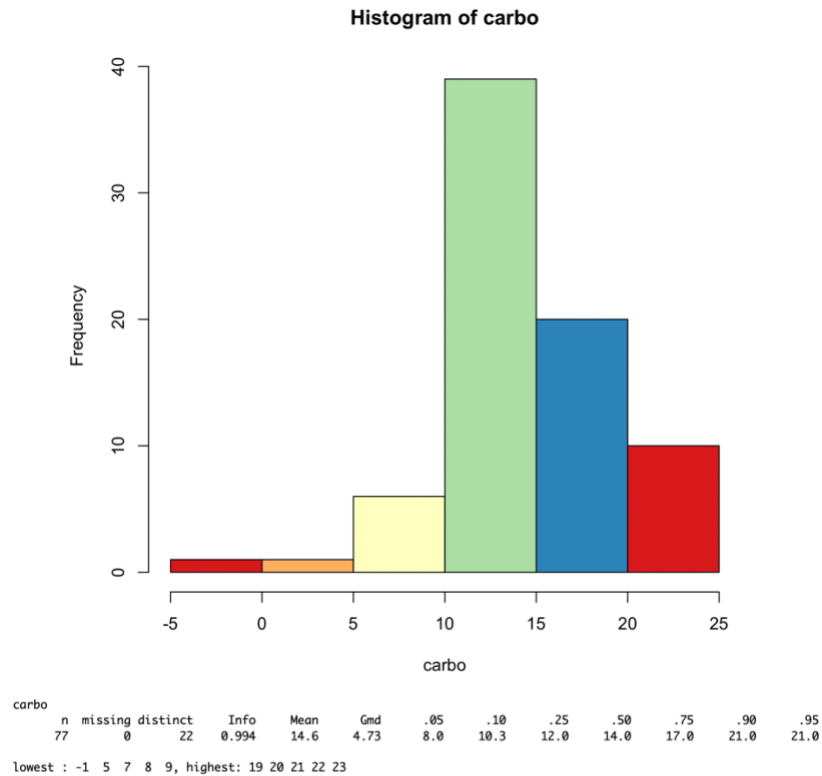
fat						
	n	missing	distinct	Info	Mean	Gmd
	77	0	5	0.892	1.013	1.049
lowest :	0	1	2	3	5	
highest:	0	1	2	3	5	
Value	0	1	2	3	5	
Frequency	27	30	14	5	1	
Proportion	0.351	0.390	0.182	0.065	0.013	

Looking at most of the cereals though per serving, they all have between 0-1 grams of fat. Like protein again, we see that R is making fat a categorical variable with only 5 possible values across the 77 options, but again I'm going to treat this as a continuous variable as it is a measurement for the specific type of cereal.

<div>sodium</div> <div>Continuous variable</div>	<div><div>Histogram of sodium</div><div>sodium</div><table><tr><th>n</th><th>missing</th><th>distinct</th><th>Info</th><th>Mean</th><th>Gmd</th><th>.05</th><th>.10</th><th>.25</th><th>.50</th><th>.75</th><th>.90</th><th>.95</th></tr><tr><td>77</td><td>0</td><td>27</td><td>0.995</td><td>159.7</td><td>93.51</td><td>0</td><td>0</td><td>130</td><td>180</td><td>210</td><td>254</td><td>282</td></tr></table><div>lowest : 0 15 45 70 75, highest: 250 260 280 290 320</div></div>	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	77	0	27	0.995	159.7	93.51	0	0	130	180	210	254	282	<div>Sodium is again a measurement like the others fat and protein as we've see so far in this table above. Unlike how R made categories within the variable for the last two, there were too many types of values so R determined it as a continuous variable as well.</div>
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95																
77	0	27	0.995	159.7	93.51	0	0	130	180	210	254	282																
<div>fiber</div> <div>Continuous variable</div>	<div><div>Histogram of fiber</div><div>fiber</div><table><tr><th>n</th><th>missing</th><th>distinct</th><th>Info</th><th>Mean</th><th>Gmd</th><th>.05</th><th>.10</th><th>.25</th><th>.50</th><th>.75</th><th>.90</th><th>.95</th></tr><tr><td>77</td><td>0</td><td>13</td><td>0.966</td><td>2.152</td><td>2.289</td><td>0.0</td><td>0.0</td><td>1.0</td><td>2.0</td><td>3.0</td><td>4.4</td><td>5.2</td></tr></table><div>lowest : 0.0 1.0 1.5 2.0 2.5, highest: 5.0 6.0 9.0 10.0 14.0</div><div>Value0.01.01.52.02.52.73.04.05.06.09.010.014.0</div><div>Frequency19163101115441111</div><div>Proportion0.2470.2080.0390.1300.0130.0130.1950.0520.0520.0130.0130.0130.013</div></div>	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	77	0	13	0.966	2.152	2.289	0.0	0.0	1.0	2.0	3.0	4.4	5.2	<div>Just as the previous ones, R made this variable a categorical variable but seeing as it's a measurement I'm going to be treating like a continuous one.</div>
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95																
77	0	13	0.966	2.152	2.289	0.0	0.0	1.0	2.0	3.0	4.4	5.2																

carbo

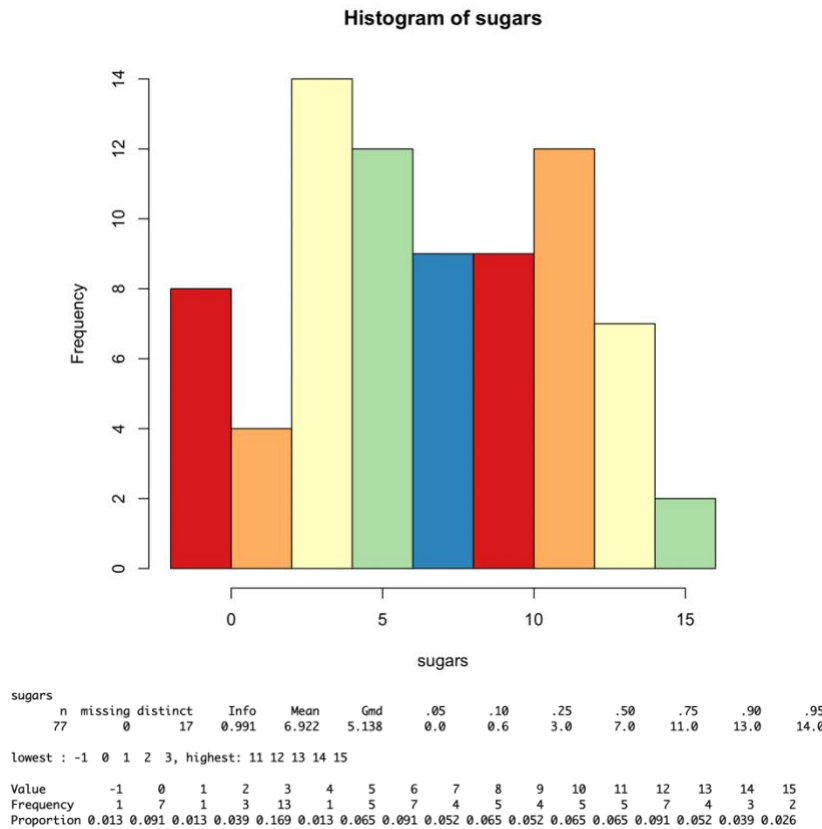
Continuous
variable



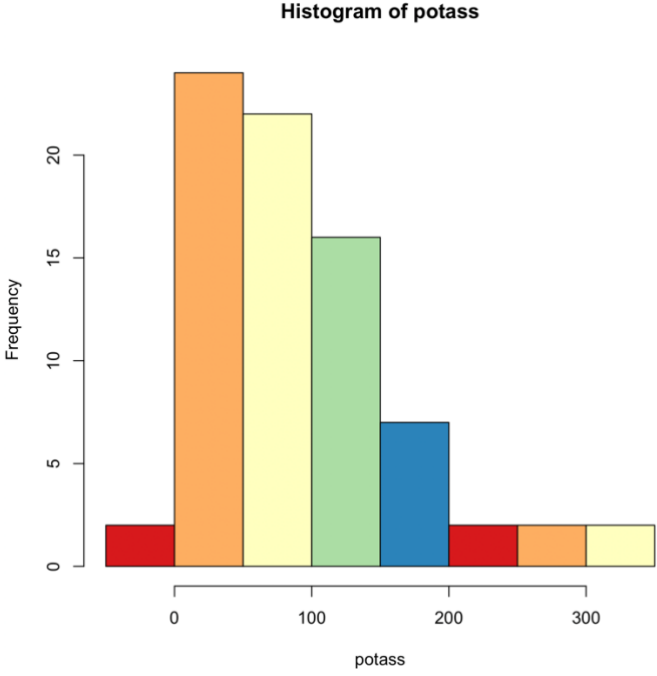
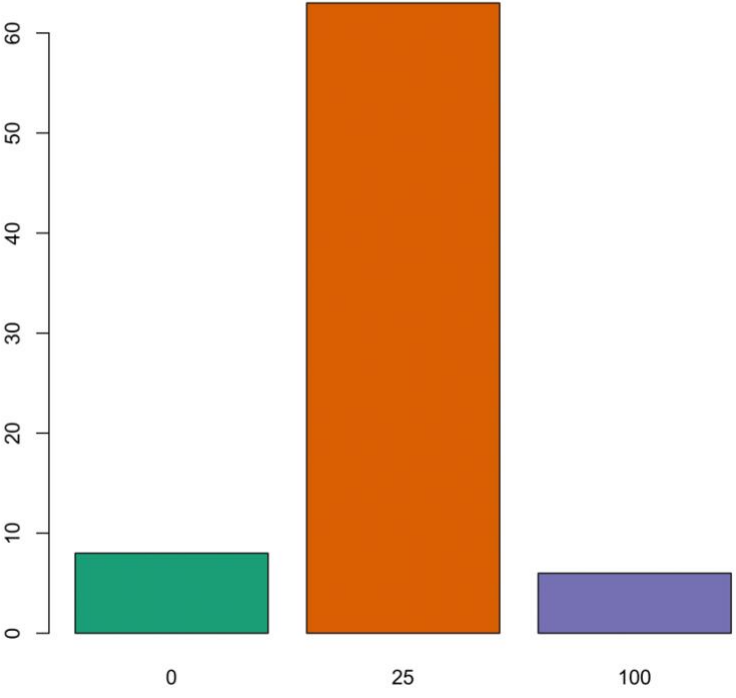
Carbohydrates are like sodium and definitely are a continuous variable, R also makes it a point to show us that as well when you find the summary.

sugars

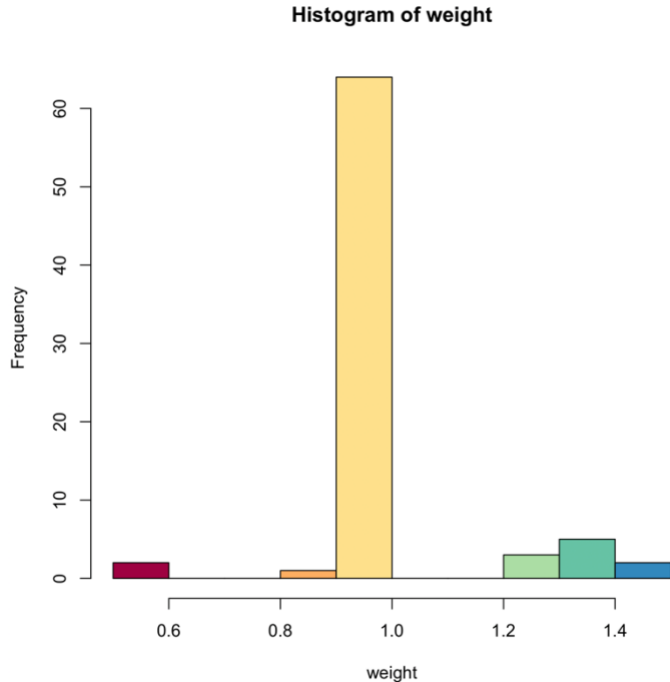
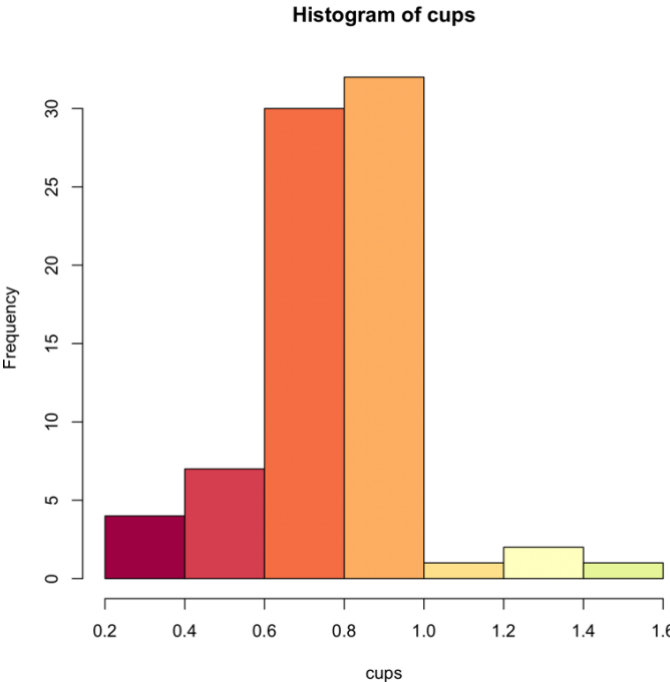
Continuous
variable



This histogram actually seems decently distributed in terms of the data here not leaning directly to only two or three ranges in the histogram but in this scenario the majority are between 6-7 grams of sugars per serving with the average of 6.92.

<p>potassium</p> <p>Continuous variable</p>	<p>Histogram of potass</p>  <pre> potass n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95 77 0 36 0.998 96.08 75.57 .24 .28 .40 .90 120 190 244 lowest : -1 15 20 25 30, highest: 240 260 280 320 330 </pre>	<p>For potassium, we definitely see that a majority of the values are between 0-200 with a mostly evident right skewed graph for potassium. R detects this as a continuous variable which is what we want as potassium is a measurement for each cereal in grams per serving.</p>
<p>vitamins</p> <p>Categorical variable</p>		<p>This is most definitely a categorical variable as there are evidently only three options and the Kaggle description explains how these are categories for FDA percentages of vitamins per cereal. Most of them do have 25% so it's unlikely we'll use this as a variable to compare and contrast cereals as we don't have enough variety of cereals that have different FDA percentages.</p>

	<div>vitamins</div> <table><tr><th>n</th><th>missing</th><th>distinct</th><th>Info</th><th>Mean</th><th>Gmd</th></tr><tr><td>77</td><td>0</td><td>3</td><td>0.451</td><td>28.25</td><td>15.64</td></tr></table> <div>Value025100 Frequency8636 Proportion0.1040.8180.078</div>	n	missing	distinct	Info	Mean	Gmd	77	0	3	0.451	28.25	15.64			
n	missing	distinct	Info	Mean	Gmd											
77	0	3	0.451	28.25	15.64											
<div>shelf</div> <div>Categorical variable</div>	<div><table><tr><th>shelf</th><th>n</th><th>missing</th><th>distinct</th><th>Info</th><th>Mean</th><th>Gmd</th></tr><tr><td></td><td>77</td><td>0</td><td>3</td><td>0.86</td><td>2.208</td><td>0.8941</td></tr></table><div>Value123 Frequency202136 Proportion0.2600.2730.468</div></div>	shelf	n	missing	distinct	Info	Mean	Gmd		77	0	3	0.86	2.208	0.8941	<div>Clearly the shelf variable is a categorical variable as there are only three shelves, we know that the mean and median aren't going to play as much of a role. A majority of these cereals are put on shelf 3 but we have enough information to compare and contrast using this variable if we wanted to.</div>
shelf	n	missing	distinct	Info	Mean	Gmd										
	77	0	3	0.86	2.208	0.8941										

<div>weight</div> <div>Continuous variable</div>	<div><div><div>Histogram of weight</div></div><div><div>weight</div><table><tr><th>n</th><th>missing</th><th>distinct</th><th>Info</th><th>Mean</th><th>Gmd</th></tr><tr><td>77</td><td>0</td><td>7</td><td>0.426</td><td>1.03</td><td>0.1102</td></tr></table><div>lowest : 0.50 0.83 1.00 1.25 1.30, highest: 1.00 1.25 1.30 1.33 1.50</div><table><tr><th>Value</th><th>0.50</th><th>0.83</th><th>1.00</th><th>1.25</th><th>1.30</th><th>1.33</th><th>1.50</th></tr><tr><td>Frequency</td><td>2</td><td>1</td><td>64</td><td>2</td><td>1</td><td>5</td><td>2</td></tr><tr><td>Proportion</td><td>0.026</td><td>0.013</td><td>0.831</td><td>0.026</td><td>0.013</td><td>0.065</td><td>0.026</td></tr></table></div></div>	n	missing	distinct	Info	Mean	Gmd	77	0	7	0.426	1.03	0.1102	Value	0.50	0.83	1.00	1.25	1.30	1.33	1.50	Frequency	2	1	64	2	1	5	2	Proportion	0.026	0.013	0.831	0.026	0.013	0.065	0.026	<div>We can see that most of the weights are approximately 1 oz per serving but some of them are not. A majority of cereals are 1 oz as the serving size and to compare we might subset the ones only with that weight as the percentage vitamins and other continuous variables would be harder to compare.</div>
n	missing	distinct	Info	Mean	Gmd																																	
77	0	7	0.426	1.03	0.1102																																	
Value	0.50	0.83	1.00	1.25	1.30	1.33	1.50																															
Frequency	2	1	64	2	1	5	2																															
Proportion	0.026	0.013	0.831	0.026	0.013	0.065	0.026																															
<div>cups</div> <div>Continuous variable</div>	<div><div><div>Histogram of cups</div></div></div>	<div>This is an approximation of the number of cups of cereal in the serving size, I don't think this would be a good metric to use as it's not a scaled metric such as ounces or grams. While R thinks this is categorical, I see this as a measure of serving size so we're going to be treating this like a continuous variable.</div>																																				

	<pre>cups n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95 77 0 12 0.926 0.821 0.2522 0.466 0.500 0.670 0.750 1.000 1.000 1.026 lowest : 0.25 0.33 0.50 0.67 0.75, highest: 1.00 1.13 1.25 1.33 1.50 Value 0.25 0.33 0.50 0.67 0.75 0.80 0.88 1.00 1.13 1.25 1.33 1.50 Frequency 1 3 7 13 16 1 2 30 1 1 1 1 Proportion 0.013 0.039 0.091 0.169 0.208 0.013 0.026 0.390 0.013 0.013 0.013 0.013</pre>																																					
rating	<pre>rating n missing distinct Info Mean Gmd .05 .10 .25 .50 .75 .90 .95 77 0 77 1.000 42.67 15.48 22.67 27.92 33.17 40.40 50.83 60.09 68.27 lowest : 18.04285 19.82357 21.87129 22.39651 22.73645, highest: 68.23588 68.40297 72.80179 74.47295 93.70491</pre> <p style="text-align: center;">Histogram of rating</p> <table><caption>Histogram Data (Estimated)</caption><thead><tr><th>Rating Range</th><th>Frequency</th></tr></thead><tbody><tr><td>15 - 20</td><td>2</td></tr><tr><td>20 - 25</td><td>11</td></tr><tr><td>25 - 30</td><td>20</td></tr><tr><td>30 - 35</td><td>19</td></tr><tr><td>35 - 40</td><td>13</td></tr><tr><td>40 - 45</td><td>5</td></tr><tr><td>45 - 50</td><td>2</td></tr><tr><td>50 - 55</td><td>1</td></tr><tr><td>55 - 60</td><td>1</td></tr><tr><td>60 - 65</td><td>1</td></tr><tr><td>65 - 70</td><td>1</td></tr><tr><td>70 - 75</td><td>1</td></tr><tr><td>75 - 80</td><td>1</td></tr><tr><td>80 - 85</td><td>1</td></tr><tr><td>85 - 90</td><td>1</td></tr><tr><td>90 - 95</td><td>1</td></tr><tr><td>95 - 100</td><td>1</td></tr></tbody></table>	Rating Range	Frequency	15 - 20	2	20 - 25	11	25 - 30	20	30 - 35	19	35 - 40	13	40 - 45	5	45 - 50	2	50 - 55	1	55 - 60	1	60 - 65	1	65 - 70	1	70 - 75	1	75 - 80	1	80 - 85	1	85 - 90	1	90 - 95	1	95 - 100	1	Looks like a lot of these cereals were rated pretty low out of a 100 (40??) seeing that the majority was 30 and 50, I wished they had more cereals to compare. This is clearly a continuous variable as it's the rating of the cereal given by other people.
Rating Range	Frequency																																					
15 - 20	2																																					
20 - 25	11																																					
25 - 30	20																																					
30 - 35	19																																					
35 - 40	13																																					
40 - 45	5																																					
45 - 50	2																																					
50 - 55	1																																					
55 - 60	1																																					
60 - 65	1																																					
65 - 70	1																																					
70 - 75	1																																					
75 - 80	1																																					
80 - 85	1																																					
85 - 90	1																																					
90 - 95	1																																					
95 - 100	1																																					

To conclude, there were a couple of things I learned just by going through each variable and not only creating a histogram but also understanding how R described each variable. While some variables can be categorized and seen in R as categorical variables, there are some that are actually continuous because they can be values in the middle and are measure of some quality of the cereal. This data set has an abundance of certain types of cereal which are cold (type), Kelloggs or General Mills (manufacturer), 25% of FDA recommended vitamins, and located on shelf 3 (shelf). I think this definitely gives an idea to me for where to start in terms of using these categorical variables to understand how likely some type of cereal is to be based off a certain brand or maybe even the shelf it's put on in the store. I'm not exactly sure how the majority of one type in comparison to other will play out when we run hypothesis tests/graph the data, but we shall see in part 2!

Hypothesis #1: Do the categorical variables individually effect the mean number of calories a cereal serving has?

To begin, I'm going to be exploring the effect of the categorical variables on the difference in calories per serving for each cereal. I think this an appropriate approach as instead of seeing something I already

have common knowledge and affirming it (i.e., that more sugary cereals have more calories), I thought it would be interesting to see if any of the categorical variables individually or also with interaction would have any effect on the difference in means for the calorific values of each cereal. I'm using one-way ANOVA as I believe it shows best how individually each categorical variable on its own affects the means of the calories. For this one-way ANOVA test, we're going to assume that each of the samples of types of cereal were taken from a normally distributed population, the samples were drawn independently, and the variance of the data is the same. We also know that the dependent variable, calories, is a continuous variable.

- 1) Which of the four categorical variables could cause a significant difference in means for calories per serving? Set the $\alpha = 0.05$ to compare the results. Use Tukey's HSD approach to see which of the pairs in the category have the most difference.

One-way ANOVA is used to see if one independent variable has an effect on the means of the categories from the independent variable in the dependent variable. The independent variable is a categorical variable as ANOVA determines if the different types within that category make a difference on the dependence variable with the one-way ANOVA test. In this scenario, we're going to be testing each of the categorical variables against the calories' variable in this data set and use ANOVA to determine if any of the types in the each of the categorical variables have an effect on the calories or not.

For each of these one-way ANOVA tests, the null hypothesis will be that all means for each (insert categorical variable) would be equal calories wise. The alternate hypothesis is that the means aren't equal.

1. Manufacturer vs Calories – Does the manufacturer cause a difference between the calorific value of a cereal per serving?

```
> summary(manufac)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
mfr	6	4710	785.0	2.276	0.0459 *
Residuals	70	24142	344.9		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

After running ANOVA, we can see here that the p-value is definitely close to 0.05 but still less than 0.05 and we can reject the null hypothesis and say that the means are different and so the manufacturer is a cause of the calories per serving being different.

2. Type vs Calories – Personally, I'm not sure if this is a good one to test as most of these are cold cereals and there isn't data for hot cereals to compare. But just to see what ANOVA says, we're going to run it anyway. Does the type (hot vs cold) cause a difference between the calorific value of a cereal per serving?

```
> summary(typ)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	1	148	147.9	0.386	0.536
Residuals	75	28704	382.7		

As we see here, the p-value given was 0.5 which is much greater than 0.05 and we cannot reject the null hypothesis so we accept it. From here we can conclude that the type of cereal is not a variable that causes a difference in calories per serving for a cereal.

3. Vitamins vs Calories -- Does the percentage of vitamins determined by the FDA (0,25,100) cause a difference between the calorific value of a cereal per serving?

```
> summary(vit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(vitamins)	2	4620	2310.1	7.055	0.00157 **
Residuals	74	24232	327.5		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Here we can see that the p-value is smaller enough than the given alpha 0.05 and therefore we can reject the null hypothesis and conclude that the cereal in fact do differ based off of the FDA percentage they have.

4. Shelf vs Calories -- Does the shelf (1,2,3) cause a difference between the calorific value of a cereal per serving?

```
> summary(shel)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(shelf)	2	559	279.7	0.732	0.485
Residuals	74	28292	382.3		

We can see here that the since the p-value is much greater than the alpha of 0.05 and accept the null hypothesis. Using this, we can conclude that there is no difference between the calories of each cereal depending on what shelf they are on.

After seeing how the manufacturer and the vitamin percentage are causes of difference in means for calories per serving for each cereal, I'm going to use Tukey's HSD to see which pairs of manufacturers and vitamin percentages differ the most.

```
> TukeyHSD(manafac)
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = calories ~ mfr)
```

```
$mfr
      diff      lwr      upr      p adj
G-A 11.3636364 -46.280327 69.007600 0.9966669
K-A  8.6956522 -48.893802 66.285106 0.9992557
N-A -13.3333333 -74.227354 47.560687 0.9940842
P-A  8.8888889 -50.537590 68.315368 0.9992948
Q-A -5.0000000 -64.796741 54.796741 0.9999763
R-A 15.0000000 -44.796741 74.796741 0.9877988
K-G -2.6679842 -19.480483 14.144514 0.9990120
N-G -24.6969697 -50.662267  1.268328 0.0727841
P-G -2.4747475 -24.782176 19.832681 0.9998749
Q-G -16.3636364 -39.639486  6.912214 0.3446939
R-G  3.6363636 -19.639486 26.912214 0.9990961
N-K -22.0289855 -47.873044  3.815073 0.1456552
P-K  0.1932367 -21.972954 22.359428 1.0000000
Q-K -13.6956522 -36.836176  9.444872 0.5547480
R-K  6.3043478 -16.836176 29.444872 0.9813032
P-N 22.2222222  -7.491017 51.935462 0.2730000
Q-N  8.3333333 -22.113677 38.780344 0.9808573
```

```
> TukeyHSD(vit)
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = calories ~ as.factor(vitamins))
```

```
$`as.factor(vitamins)`
      diff      lwr      upr      p adj
25-0 23.730159  7.485560 39.97476 0.0023037
100-0 31.666667  8.292374 55.04096 0.0050352
100-25 7.936508 -10.555059 26.42807 0.5625963
```

We

percentage by the FDA is 0. We can also say that although the p-value for manufacturers was less than and close to 0.05, that in fact there wasn't that much change in the means for calories per cereal serving dependent on manufacturers. The shelf and the type had absolutely no effect on the means of cereal calories per serving at all as we saw with the initial one-way ANOVA tests.

On the left, after running TukeyHSD, we can see that when trying to find which pair of brands has a p-value less than 0.05 to reject the null hypothesis and conclude that there is a difference, there actually is none. I find this interesting because when we ran one-way ANOVA, we did get a p-value that made sense but when we look closer, we realize that it's probably because of the one or two outliers here of R-N and N-G which were getting closer to 0.05 for their p-value but not less than 0.05. This would also explain why the p-value was VERY close to 0.05 in the original ANOVA I ran.

Again, on the left, we see the TukeyHSD when run on the vitamins ANOVA and we can see that between 25-0 and 100-0 have a variety in calorific value. This would make sense as logically speaking less or more vitamins in a food, in this scenario.

can overall conclude, that vitamins definitely cause a difference in means but only when the vitamins

Hypothesis #2: Do pairs of categorical variables have significant interactions that effect the rating of the cereal?

In the first hypothesis, we saw how only the vitamins category affected the means on calories through one-way ANOVA. Clearly, the calories factor wasn't really affected by most of the categorical variables (only one and that was vitamins). Because of this, I feel it would be intriguing to see the effect of multiple of the categorical variables together on the means of a different variable not calories.

Instead of looking at a nutritional fact, I'd like to see if the means of the rating of the different cereals gets when seeing different two-way interactions between the four categorical variables we have in this data. I

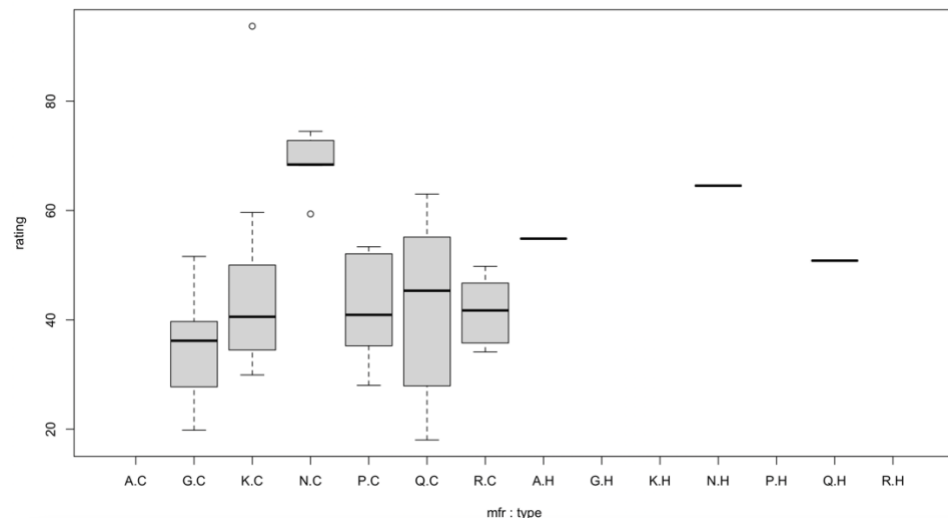
feel because we saw how one-way ANOVA worked with individual categorical variables it might be more useful to see if combinations of categorical variables and if their interactions cause a difference in means for the rating. When using two-way ANOVA, the independent variables are both going to be categorical variables and the dependent variable is going to be the continuous variable in this scenario, which would be the rating of the cereal. We are also going to assume normality (each sample was taken from a normally distributed population), variance equality (the variance of different data being the same) and that each sample was drawn independently from one another.

The null hypothesis for each of these two way ANOVA tests will be that there is no interaction between the two categorical variables. The alternate hypothesis will that there is some interaction between the two categorical variables. The alpha level which is going to be tested is 0.05.

Below is a table where on the right we have the boxplot to see the variability of the data beforehand and there's also a summary table of the two-way ANOVA test run on those variables. Along with the two categorical variables tested with interaction on the left, there is a description of my thoughts on the variability of the boxplot and conclusions from the summary from the two-way ANOVA.

manufacturer and type

As we see, there's definitely a lot of variability between the data that exists. We can see that this doesn't work for all the data though because almost half the data is missing as there weren't enough entries for hot cereals. From the p-value of 0.467, we can conclude that there is no significant interaction between manufacturer and type that effects the rating of the cereal.

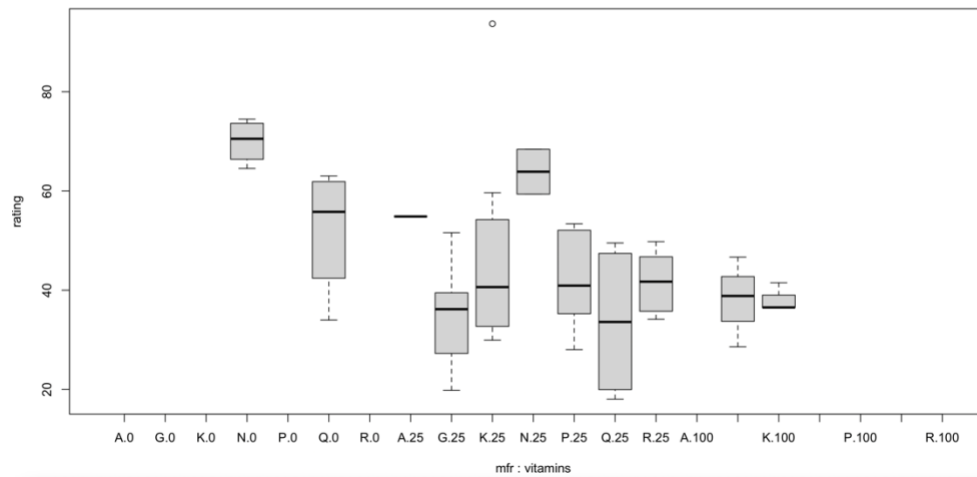


```
> summary(mfrtyp)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mfr	6	5524	920.7	6.670	1.4e-05	***
type	1	12	11.7	0.085	0.771	
mfr:type	1	74	74.0	0.536	0.467	
Residuals	68	9387	138.0			

manufacturer and vitamin

Again here, a decent amount of data is missing because certain manufacturers definitely have more cereals at certain FDA vitamin level while others do not. We can see that between both the mfr and vitamins there's a lot of variability. Again, the p-value of 0.113 tells us that the interaction between the variables is insignificant.



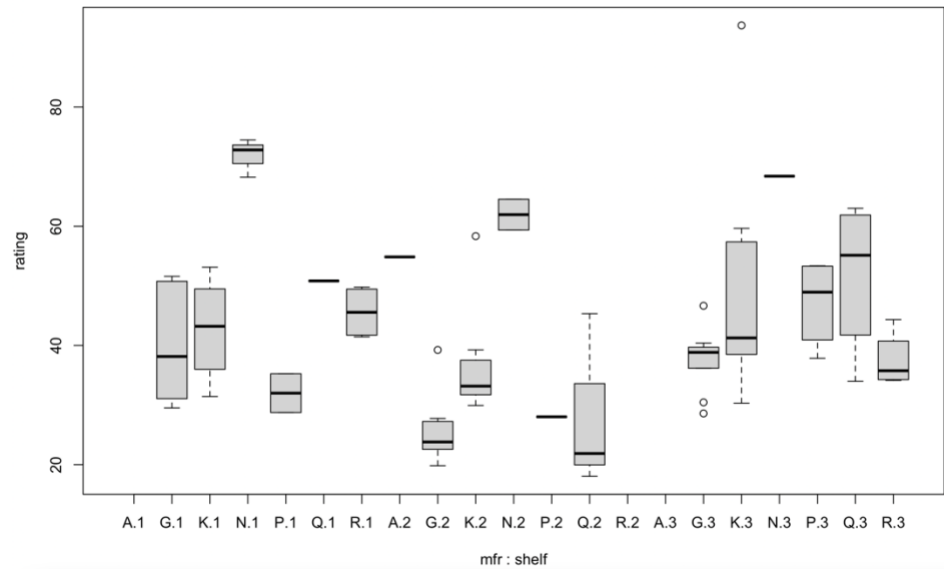
```
> summary(mfrvit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mfr	6	5524	920.7	7.083	7.51e-06	***
vitamins	1	87	87.0	0.669	0.416	
mfr:vitamins	3	807	269.0	2.069	0.113	
Residuals	66	8579	130.0			

```
---
```


manufacturer and shelf

Unlike the last ones, we can definitely see that all the cereals are on a variety of shelves and are a variety of manufacturers. I see that we definitely have less of data with the mfr of A as there's almost no data for the boxplots to even exist on this boxplot. Within both mfr and shelf though there's a good amount of variability. But, with the p value being much greater than 0.05, so therefore this interaction is insignificant to the rating of a cereal.

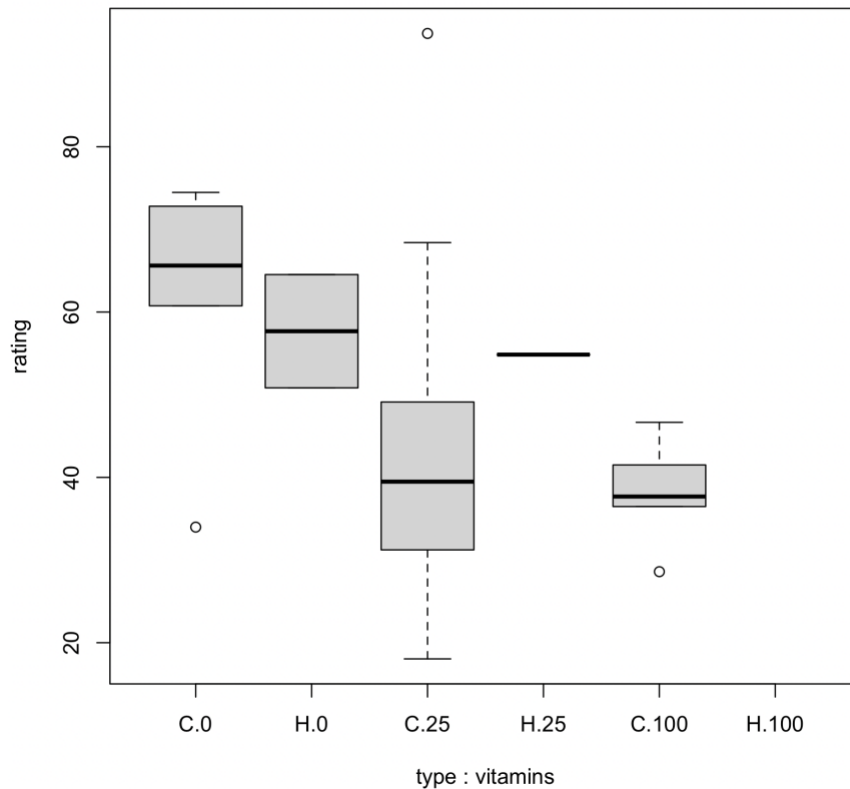


> summary(mfrshelf)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
mfr	6	5524	920.7	7.014	9.18e-06	**
shelf	1	163	163.1	1.243	0.269	
mfr:shelf	5	909	181.8	1.385	0.242	
Residuals	64	8401	131.3			

Type and vitamin

Because there are only two types and two levels of vitamins, we're working with a much smaller amount of boxplots. Again, there's a lot of data lacking for hot types of cereal and therefore it makes sense that the H.100 is missing and that H.25 only has one cereal in it. As for the p-value, it's really big and therefore the interaction between type and vitamin is insignificant to the rating of a cereal.

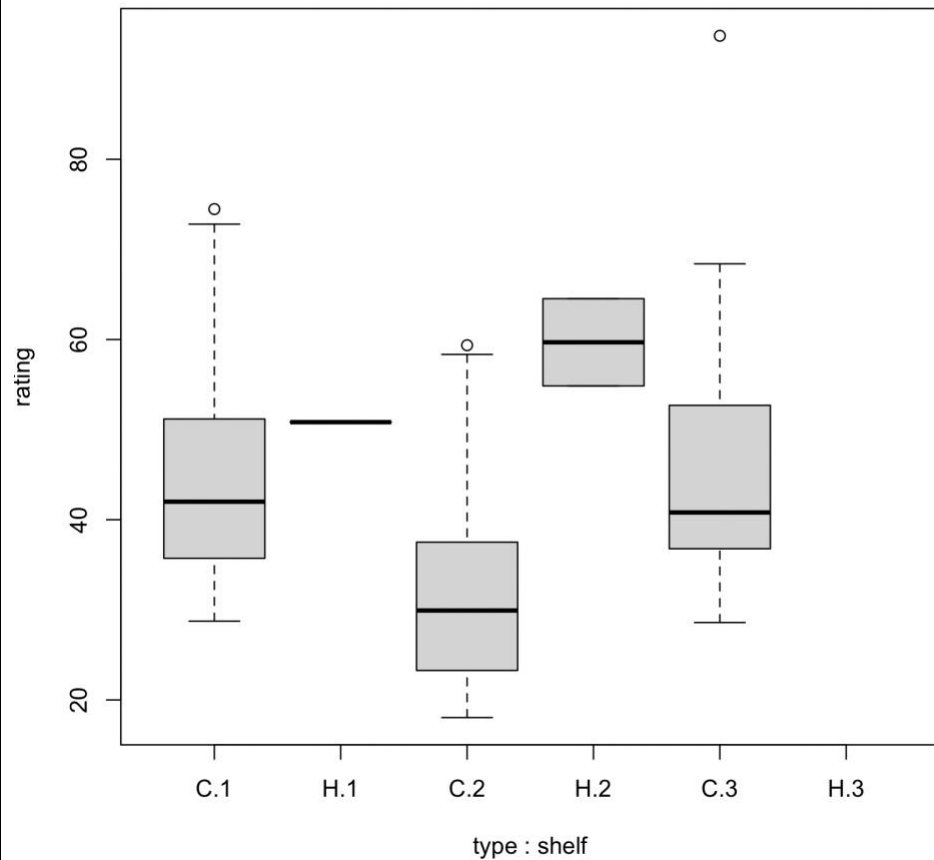


```
> summary(typvit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	1	618	618.1	3.286	0.0740
vitamins	1	644	644.3	3.425	0.0683
type:vitamins	1	0	0.2	0.001	0.9770
Residuals	73	13734	188.1		

type and shelf

Just like the last one, we're lacking data for hot cereals and therefore the hot cereals on shelves 3 and 1 have only one and no data at all. The p-value tells us again that this interaction is insignificant to the rating of a cereal.

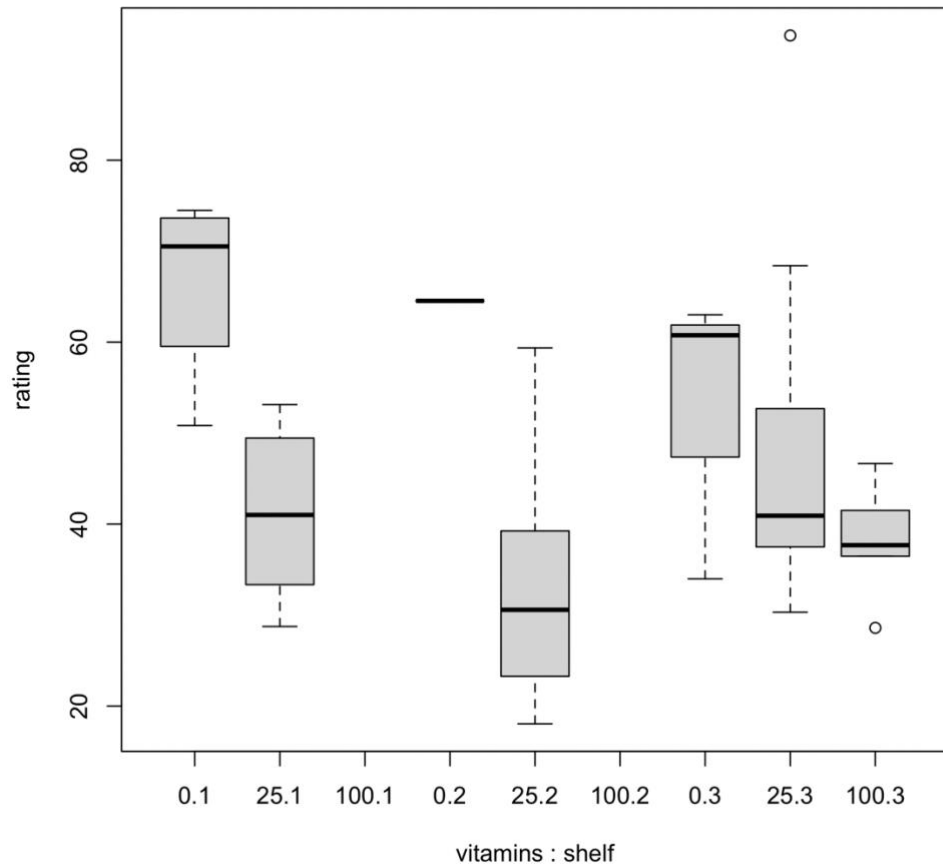


```
> summary(typoshel)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	1	618	618.1	3.157	0.0798
shelf	1	41	41.1	0.210	0.6482
type:shelf	1	43	42.9	0.219	0.6410
Residuals	73	14295	195.8		

vitamin and shelf

In this data set, we can see we're lacking cereals that have a 100% FDA rating for vitamins which isn't that good. We can see though there's variability across the different vitamins and shelves. In this scenario, the p-value is extremely small and therefore we can say that the interaction between vitamins and shelf are significant to the rating of the cereal.



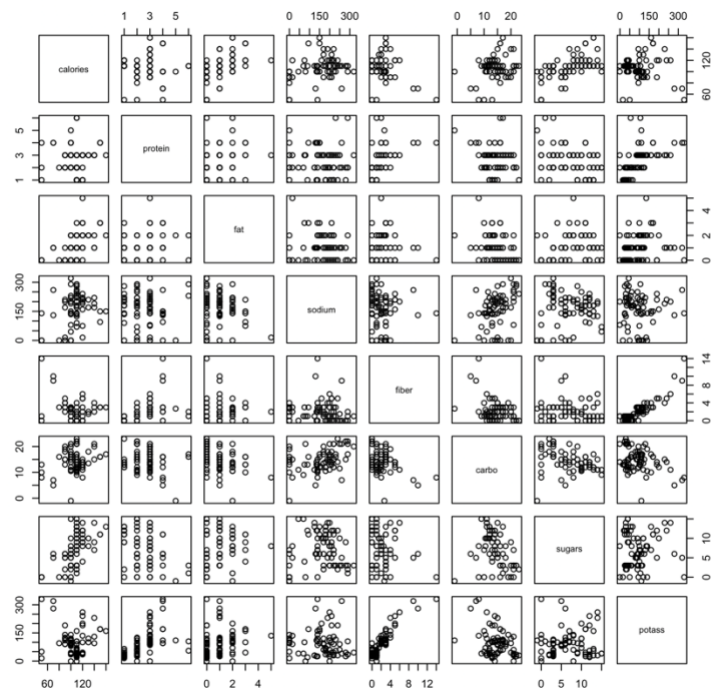
```
> summary(vitshel)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
vitamins	1	868	867.7	5.599	0.0206	*
shelf	1	155	155.4	1.003	0.3199	
vitamins:shelf	1	2659	2659.1	17.156	9.14e-05	***
Residuals	73	11315	155.0			

In conclusion to this hypothesis, we can say that the interaction between the vitamins variable and shelf variable are significant to the rating of the cereal. Essentially, when people were rating the cereals, the shelf the cereal was put on with the percentage of vitamins suggested by the FDA both effected the rating they gave for it. As we look through each of the summaries, we can also see that variable manufacturer had a really small p-value each time it was tested which may tell us that it might want to be something we test with one-way ANOVA in the future to see if it affects the mean rating of the cereal.

Research Question #3: Which of the nutritional facts of these cereals are correlated and which are the most strongly related and the weakest?

After working testing the effects of categorical variables on quantitative variables throughout this entire paper, I was interested to see if any of the nutritional facts of cereals were correlated. I found this curious as usually some nutritional values are correlated normally to each other, but I wanted see how they correlate specifically for cereals. In order to this, I had to create a panel with all the possible scatterplots for each of the nutritional facts to see if there was in fact any visual correlations. As we can see here, we see that some have clusters and others such as the graph for fiber vs potassium does have some linear correlation of some kind as well. Some due to having categorical-like variables (i.e. protein which in this table only had 4 possibilities of 1,2,3 & 4) are portrayed as completely vertical or horizontal lines (i.e. no correlation). To see which of these are most strongly/weakly correlated with each other, we can run a command in R which allows us to see all of them at once.



	calories	protein	fat	sodium	fiber	carbo	sugars	potass
calories	1.00000000	0.01906607	0.498609814	0.300649227	-0.29341275	0.2506809	0.56234029	-0.06660886
protein	0.01906607	1.00000000	0.208430990	-0.054674348	0.50033004	-0.1308636	-0.32914178	0.54940740
fat	0.49860981	0.20843099	1.000000000	-0.005407464	0.01671924	-0.3180435	0.27081918	0.19327860
sodium	0.30064923	-0.05467435	-0.005407464	1.000000000	-0.07067501	0.3559835	0.10145138	-0.03260347
fiber	-0.29341275	0.50033004	0.016719237	-0.070675009	1.000000000	-0.3560827	-0.14120539	0.90337367
carbo	0.25068091	-0.13086365	-0.318043492	0.355983473	-0.35608274	1.00000000	-0.33166538	-0.34968522
sugars	0.56234029	-0.32914178	0.270819175	0.101451381	-0.14120539	-0.3316654	1.00000000	0.02169581
potass	-0.06660886	0.54940740	0.193278602	-0.032603467	0.90337367	-0.3496852	0.02169581	1.00000000

As seen above, this table mirrors the original panel of scatterplots but instead of the scatterplot itself, we can see all the correlations. We can see that fiber and potassium in cereals have the highest correlation with 0.90337 and the weakest correlation would be 0.016 with fiber and fat. We can see above that fiber and potassium definitely seem to be somewhat linear when we see the scatterplot. Overall, we can conclude that most of these nutritional facts aren't correlated like we thought they were and although fiber and fat may be strongly correlated we can't assume any causation as there may also be confounding variables in the cereal that we don't know about and can't find or presume.