R Notebook

Recently purchased of a instant pot and had made me curious me hungry turned me into a cooking monster with goal of making different dishes that I haven't try making before. When I came across this Epicurious dataset, I thought it would be interesting to fin

This dataset has 2,0052 observations and 680 variables. Some interesting features are: * ratings * carlories * protein * fat * sodium * reamaining varibles: - state names - drink name - meal type: breakfast, dinner, lunch - festival: halloween - cooking method: contained in the title variable - ingredient: chicken

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(dplyr)
```

Want to understand if the ratings corresponse to carlories, fat and protein.

```
library(readr)
epi_r <- read_csv("~/GitHub/Opendata/epi_r.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   title = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

Data Exploration

A few big values. Let't check to see if that make sense- Max carlories, sodium and fat dishe comes from Pear-Cranberry Mincemeat Lattice Pie. Looks like the nutrient value in this dish is an entry mistake. Also, other max data points are hugely out of proportion.

```
summary(epi_r$rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.750   4.375   3.714   4.375   5.000
```

```
summary(epi_r$calories)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##        0      198      331     6323      586 30110000     4117
```

```
summary(epi_r$protein)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##      0.0      3.0      8.0    100.2     27.0 236500.0     4162
```

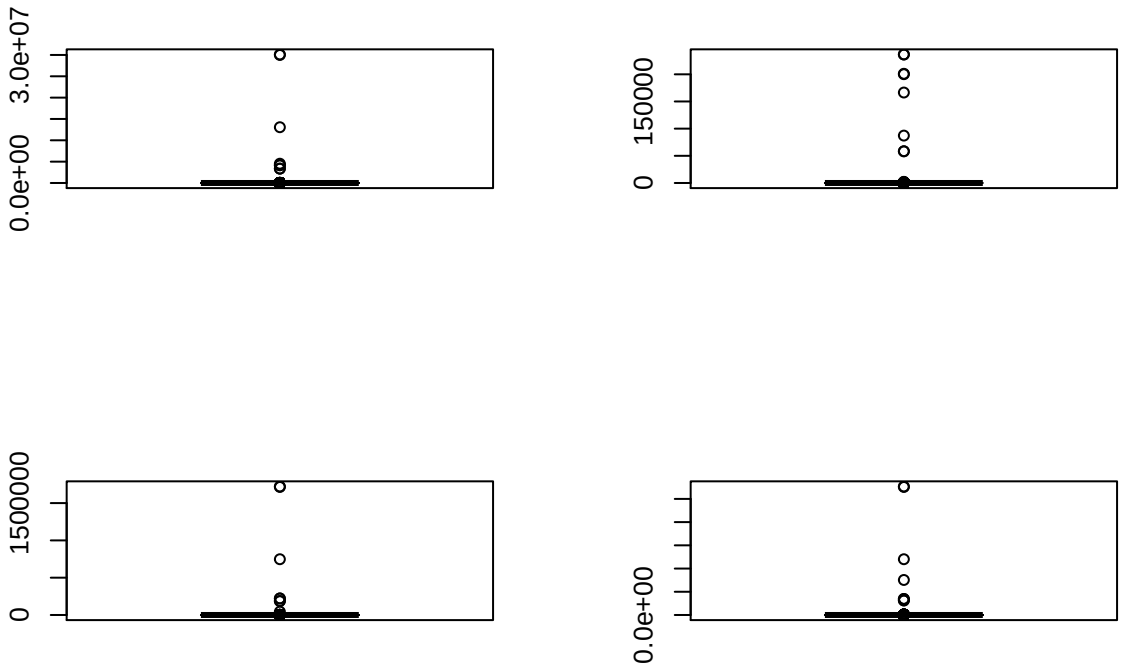
```
summary(epi_r$fat)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##       0.0       7.0      17.0     346.9      33.0 1723000.0      4183
```

```
summary(epi_r$sodium)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##        0       80      294     6226      711 27680000     4119
```

```
par(mfrow=c(2,2))
boxplot(epi_r$calories)
boxplot(epi_r$protein)
boxplot(epi_r$fat)
boxplot(epi_r$sodium)
```




```
#which.max(epi_r$fat)
#which.max(epi_r$sodium)
#hich.max(epi_r$carlories)
#epi_r[11392,1]
```

Data cleaning

Dataset reduced from 2,0052 obs. to 1,8251 obs.

```
#compute number of dulicates
epi_duplicates<- sum(duplicated(epi_r))
#remove duplicate rows
epi_r<-distinct(epi_r)
```

Create new variables "rating_new" and assigned "good","ok","bad" ranking per rating scores.

```
epi_new<- epi_r %>%
  mutate(rating_round = round(epi_r$rating)) %>%
  mutate(rating_new = ifelse(rating_round >=4,"good",
                             ifelse(rating_round >=3,"ok","bad")))
```

```
## Warning: Mangling the following names: bon app<U+FFFD><U+FFFD>tit -> bon
## app<U+FFFD><U+FFFD>tit, cr<U+FFFD><U+FFFD>me de cacao -> cr<U+FFFD><U
## +FFFD>me de cacao. Use enc2native() to avoid the warning.

## Warning: Mangling the following names: bon app<U+FFFD><U+FFFD>tit -> bon
## app<U+FFFD><U+FFFD>tit, cr<U+FFFD><U+FFFD>me de cacao -> cr<U+FFFD><U
## +FFFD>me de cacao. Use enc2native() to avoid the warning.
```

```
head(epi_r$rating)
```

```
## [1] 2.500 4.375 3.750 5.000 3.125 4.375
```

```
head(epi_new$rating_new)
```

```
## [1] "bad"  "good" "good" "good" "ok"   "good"
```

The 25th, 50th, 70th, and 95th percentiles of the recipe carlories are 198, 331, 586, 1316 calories respectively.

```
quantile(epi_new$calories,c(.25,.50,.75,.95),na.rm=TRUE)
```

```
##    25%    50%    75%    95%
##  205.0  345.0  599.0 1323.9
```

```
quantile(epi_new$sodium,c(.25,.50,.75,.95),na.rm=TRUE)
```

```
##  25%  50%  75%  95%
##   88  304  732 2058
```

Outliers in this dataset lie in the upper tail of distribution so I decided to remove data geater than 95th percentile.

```
epi_new <- epi_new %>% filter(
              epi_new$fat <quantile(epi_new$fat,.95, na.rm=TRUE),
              epi_new$calories<quantile(epi_new$calories,.95,na.rm=TRUE),
              epi_new$protein < quantile(epi_new$protein,.95,na.rm=TRUE),
              epi_new$sodium  < quantile(epi_new$sodium, .95,na.rm=TRUE)) %>% na.omit()
```

```
## Warning: Mangling the following names: bon app<U+FFFD><U+FFFD>tit -> bon
## app<U+FFFD><U+FFFD>tit, cr<U+FFFD><U+FFFD>me de cacao -> cr<U+FFFD><U
## +FFFD>me de cacao. Use enc2native() to avoid the warning.
```

Create a new dataset with the five variables of interest.

```
epi_new_graph<- select(epi_new, c(calories:sodium), rating_new) %>% gather(key, value, -rating_new)
```

Collapse columnes into rows except rating variable.

```
epi_new_graph %>% head(10)
```

```
## # A tibble: 10 x 3
##    rating_new      key value
##         <chr>    <chr> <dbl>
##  1        bad calories   426
##  2       good calories   403
##  3       good calories   165
##  4         ok calories   547
##  5       good calories   948
##  6       good calories   170
##  7       good calories   602
##  8       good calories   256
##  9       good calories   766
## 10       good calories   174
```

Distribution of nutrients

```
library(ggplot2)
library(ggjoy)
plot_density = function (dataset, rating) {
  ggplot(dataset, aes(x=value,y=rating, fill =rating)) +
  geom_joy2(scale = 0.9)  + facet_wrap(~key, scales="free") +
  labs(title='Epicurious recipe rating vs nutrients', caption='source: Epicurious')
}
plot_density(epi_new_graph, epi_new_graph$rating_new)
```

```
## Picking joint bandwidth of 40.4
```

```
## Picking joint bandwidth of 2.79
```

```
## Picking joint bandwidth of 2.08
```

```
## Picking joint bandwidth of 62.5
```

Average calories intake for an adult is 2000 to 2500kcal per day No more than 1,500 milligrams (mgs) of sodium a day *Recommended daily fat Intakes is 23 grams* 46 - 56 grams of protein per day for an average weight adult (this number varies per body weight) *Observation* Recipes that have Good and Ok rating have similiar distribution and spread while bad reciepts have narrower spread yet most recipes are fall within the recommended nutrient intake.

Correlation between carlories, sodium, protein and fat

```
#pairs(~calories+sodium+protein+fat,data=epi_new, main="Simple Scatterplot Matrix")
```

Use Scatterplot Matrices from car package to make scatterplot matrix with regression line and qqplot along the histogram.

```
library(car)
```

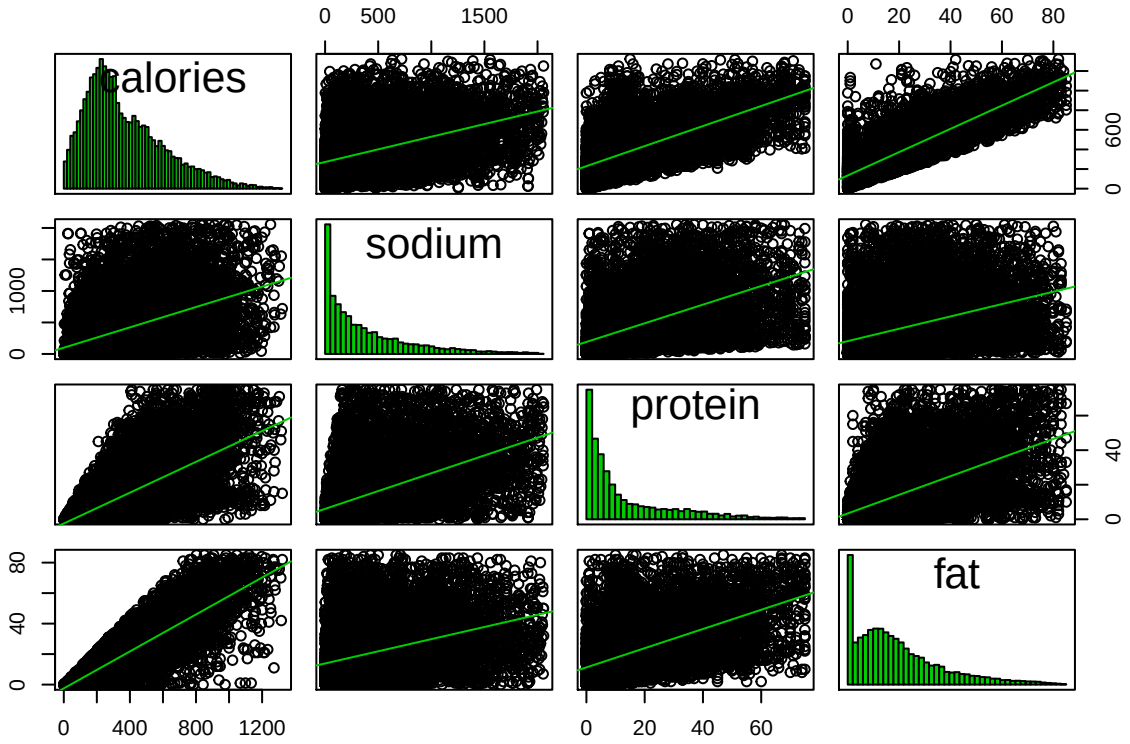
```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
scatterplot.matrix(~calories+sodium+protein+fat,data=epi_new, diagonal="histogram", smoother="FALSE" )
```

```
## Warning: 'scatterplot.matrix' is deprecated.
## Use 'scatterplotMatrix' instead.
## See help("Deprecated") and help("car-deprecated").
```




```
detach(package:car)
```

Observation calories vs fat and carlories vs protein seem to have a linear correlation. Let's find out the variables correlation coefficient.

Correlation bw fat and carlories

```
cor.test(epi_new$calories, epi_new$fat,
                    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  epi_new$calories and epi_new$fat
## t = 182.28, df = 12900, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8438256 0.8534793
## sample estimates:
##       cor
## 0.8487231
```

Observation p-value of the test is 2.2**-16, which is less than the significance level ??=0.05. `Carlories` and `fat` are highly correlated.

Correlation bw fat and protein

```
cor.test(epi_new$calories, epi_new$protein,
                    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  epi_new$calories and epi_new$protein
## t = 104.03, df = 12900, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6659251 0.6846950
## sample estimates:
##       cor
## 0.6754195
```

Observation `Carlories` and `protein` is also found correlated.

Correlation bw fat and sodium (extra)

```
cor.test(epi_new$calories, epi_new$sodium,
                    method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  epi_new$calories and epi_new$sodium
## t = 58.199, df = 12900, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4422541 0.4695898
## sample estimates:
##       cor
## 0.4560295
```

Observation cor.coeff is 0.45 that tells me there is a moderate correlation between `calories` and `sodium`.

Note: The observations in this data is very large so shapiro.test() for normal distribution is not needed in this case.

5-star rating and healthy dishes

Now- if I want a healthy yet delicious 5-star ratings recipes for dinner, what dishes can I make?

```
healthy_dish<- epi_new %>% group_by(rating, calories, fat, dinner) %>%
  filter(rating ==5, calories < 1000, fat < 10, dinner==1)
```

```
## Warning: Mangling the following names: bon app<U+FFFD><U+FFFD>tit -> bon
## app<U+FFFD><U+FFFD>tit, cr<U+FFFD><U+FFFD>me de cacao -> cr<U+FFFD><U
## +FFFD>me de cacao. Use enc2native() to avoid the warning.
```

```
healthy_dish %>% head(10)
```

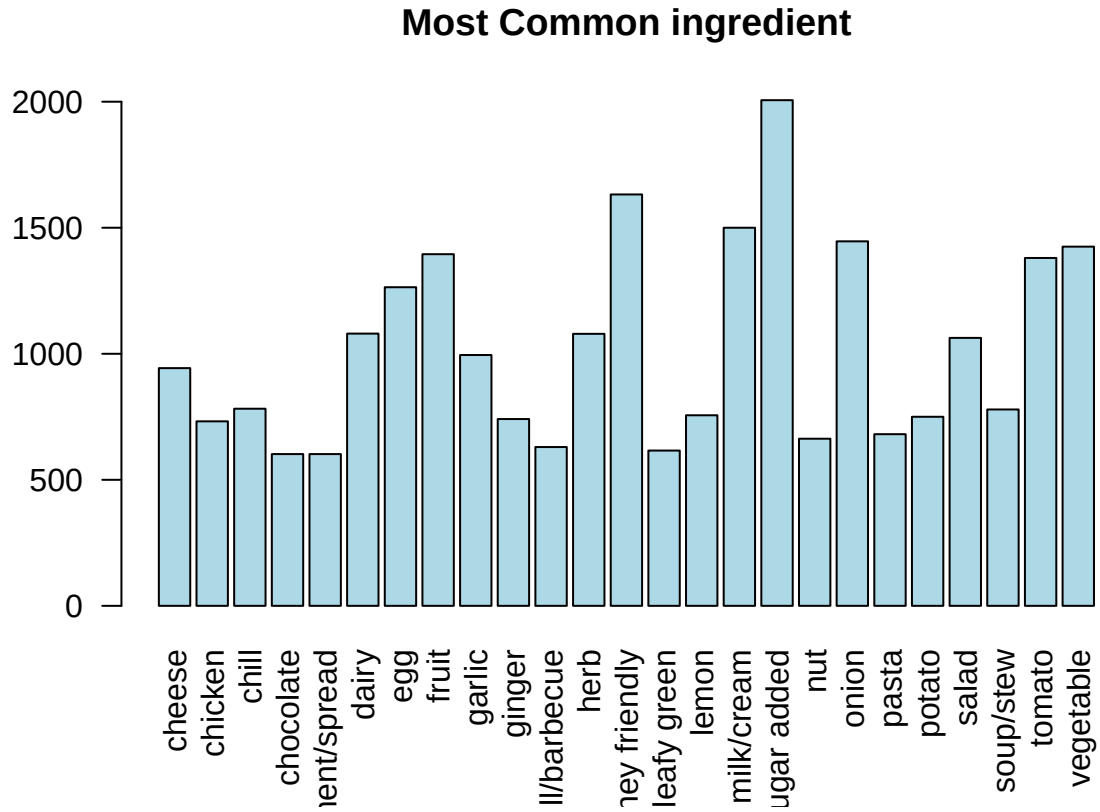
```
## # A tibble: 10 x 682
## # Groups:   rating, calories, fat, dinner [10]
##                                                          title rating
##                                                          <chr>  <dbl>
##  1             Chris Lilly's Flank Steak and Shiitake Yakitori      5
##  2                                       Everyday Yellow Dal      5
##  3                                          Teriyaki Salmon      5
##  4     Chopped Fried-Fish Tacos (Tacos de salpicón de pescado)      5
##  5                         Sautéed Greens with Toasted Walnuts      5
##  6                      Brazilian Fish Stew (Moqueca Capixaba)      5
##  7                    Halibut Ceviche With Tomato and Cucumber      5
##  8     Veal Chops with Sherry Gastrique and Roasted Peperonata      5
##  9 Garnet Yams with Maple Syrup, Walnuts, and Brandied Raisins      5
## 10            Butternut Squash, Kale, and Crunchy Pepitas Taco      5
## # ... with 680 more variables: calories <dbl>, protein <dbl>, fat <dbl>,
## #   sodium <dbl>, `#cakeweek` <dbl>, `#wasteless` <dbl>, `22-minute
## #   meals` <dbl>, `3-ingredient recipes` <dbl>, `30 days of
## #   groceries` <dbl>, `advance prep required` <dbl>, alabama <dbl>,
## #   alaska <dbl>, alcoholic <dbl>, almond <dbl>, amaretto <dbl>,
## #   anchovy <dbl>, anise <dbl>, anniversary <dbl>, `anthony
## #   bourdain` <dbl>, aperitif <dbl>, appetizer <dbl>, apple <dbl>, `apple
## #   juice` <dbl>, apricot <dbl>, arizona <dbl>, artichoke <dbl>,
## #   arugula <dbl>, `asian pear` <dbl>, asparagus <dbl>, aspen <dbl>,
## #   atlanta <dbl>, australia <dbl>, avocado <dbl>, `back to school` <dbl>,
## #   `backyard bbq` <dbl>, bacon <dbl>, bake <dbl>, banana <dbl>,
## #   barley <dbl>, basil <dbl>, bass <dbl>, `bastille day` <dbl>,
## #   bean <dbl>, beef <dbl>, `beef rib` <dbl>, `beef shank` <dbl>, `beef
## #   tenderloin` <dbl>, beer <dbl>, beet <dbl>, `bell pepper` <dbl>,
## #   berry <dbl>, `beverly hills` <dbl>, birthday <dbl>, biscuit <dbl>,
## #   bitters <dbl>, blackberry <dbl>, blender <dbl>, `blue cheese` <dbl>,
## #   blueberry <dbl>, boil <dbl>, `bok choy` <dbl>, `bon appétit` <dbl>,
## #   `bon app<U+FFFD><U+FFFD>tit` <dbl>, boston <dbl>, bourbon <dbl>, braise <dbl>,
## #   bran <dbl>, brandy <dbl>, bread <dbl>, breadcrumbs <dbl>,
## #   breakfast <dbl>, brie <dbl>, brine <dbl>, brisket <dbl>,
## #   broccoli <dbl>, `broccoli rabe` <dbl>, broil <dbl>, brooklyn <dbl>,
## #   `brown rice` <dbl>, brownie <dbl>, brunch <dbl>, `brussel
## #   sprout` <dbl>, buffalo <dbl>, buffet <dbl>, bulgaria <dbl>,
## #   bulgur <dbl>, burrito <dbl>, butter <dbl>, buttermilk <dbl>,
## #   `butternut squash` <dbl>, `butterscotch/caramel` <dbl>, cabbage <dbl>,
## #   cake <dbl>, california <dbl>, calvados <dbl>, cambridge <dbl>,
## #   campari <dbl>, camping <dbl>, canada <dbl>, candy <dbl>, ...
```

Most common ingredient

```
ingredient_count <- epi_new %>%
  select(alcoholic:turkey, -summer, -fall, -spring, -christmas, - `christmas eve`, -thanksgiving, -`qui
  colSums()
```

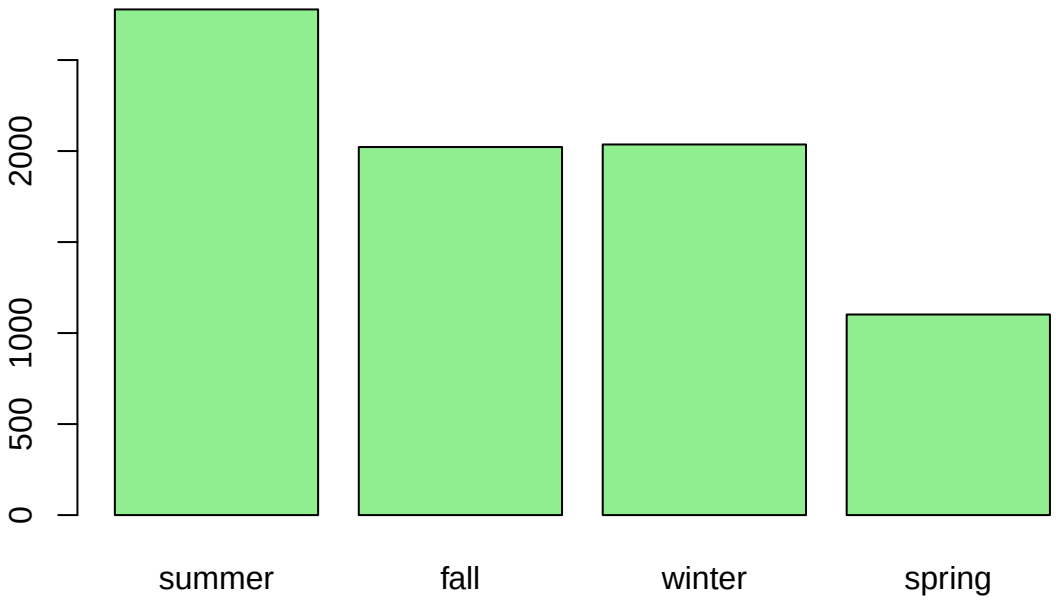
```
ingredient_count2 <- subset ( ingredient_count, ingredient_count >600)
```

```
barplot(ingredient_count2, las=2, col="lightblue", main="Most Common ingredient")
```




```
season<- epi_new %>% select(summer,fall,winter,spring) %>% colSums()
```

```
barplot(season, col="lightgreen")
```


0
500
1000
2000
summer
fall
winter
spring