Project Proposal: Burrito Cost Analysis

Preston Dunton, Trevor Overton, Jasmine DeMeyer

2022-04-12

Ingredient Model Proposal

We would like to use the burrito dataset that Dr. Wilson made available on RStudio Cloud. One interesting direction of study in this dataset is the cost of burritos.

We can imagine that both the ingredients and the restaurant's margins contribute to cost. To analyze these relationships, we would a model similar to:

$\text{Cost} = \beta_0 + \beta_1 * \text{hasPork} + \beta_2 * \text{hasChicken} + \beta_3 * \text{hasCheese} + ...$

or

$\text{Cost} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$

$$\text{where } \mathbf{X}_{(n\times p)} = \begin{bmatrix} 1 & \text{hasPork}_0 & \text{hasChicken}_0 & \text{hasCheese}_0 & ... \\ 1 & \text{hasPork}_1 & \text{hasChicken}_1 & \text{hasCheese}_1 & ... \\ ... & ... & ... & ... & ... \\ 1 & \text{hasPork}_n & \text{hasChicken}_n & \text{hasCheese}_n & ... \end{bmatrix}$$

$p = \text{Number of ingredients} + 1$

$$\boldsymbol{\beta}_{(p\times 1)} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_p \end{bmatrix}$$

$\boldsymbol{\beta} \sim MVN(\mathbf{0}, \tau^2\mathbf{I})$

$\sigma^2 \sim \text{Gamma}(a, b)$

All variables except Cost are indicator variables for the ingredients taking on values 0 and 1. This model uses Cost as a response variable. It also has parameters that can be interpreted as the marginal cost of ingredients.

Because Cost should be a positive number, we expect Cost and all the $\boldsymbol{\beta}$ parameters to be positive. This means we could model this situation as

$\text{Cost} \sim \text{TruncatedNormal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$ where the normal is truncated at $\text{Cost} >= 0$.

$\boldsymbol{\beta} \sim \text{TruncatedMVN}(\mathbf{0}, \tau^2\mathbf{I})$ where the normal is truncated at $\boldsymbol{\beta} >= \mathbf{0}$.

Choosing Priors

Let $\sigma = 1.5$, and $\text{E}[\sigma^2] = 2.25$, this means that $a/b = 2.25$. Let $a/b^2 = 1$, this means that $a = 2.25^2 = 5.0625$ and $b = 2.25$ This means that we expect burritos with the same ingredients to vary by \$1.50.

$\boldsymbol{\beta} \sim MVN(\mathbf{0}, \tau^2\mathbf{I})$

Let $\tau = 2$ so that $\tau^2 = 4$. This means that we expect the marginal cost of an ingredient to vary by \$2.00.

Restaurant Model Proposal

If we wanted to find more information about individual restaurants, and their effect on cost, we would use a model like

$\text{Cost} = \beta_0 + \beta_1 * fromOscars + \beta_2 * fromDonatos + \beta_3 * fromChipotle + ...$

which would follow almost the exact same distributions and definitions as the ingredients model presented above.

Questions

These models could answer questions like:

- What's the probability that pork costs more on a burrito than chicken?
- What is the average marginal cost of toppings on a burrito?
- How is cost affected by the restaurant from which the burrito was purchased?
- Are certain restaurants significantly more expensive than others?

Data Inspection

Loading Data

```
load('./burritodata.Rda')
head(burrito)
```

```
##               Location Cost Hunger Length Circum Volume Tortilla Temp Meat
## 1   Donato's taco shop 6.49    3.0     NA     NA     NA        3  5.0  3.0
## 2 Oscar's Mexican food 5.45    3.5     NA     NA     NA        2  3.5  2.5
## 3 Oscar's Mexican food 4.85    1.5     NA     NA     NA        3  2.0  2.5
## 4 Oscar's Mexican food 5.25    2.0     NA     NA     NA        3  2.0  3.5
## 5         Pollos Maria 6.59    4.0     NA     NA     NA        4  5.0  4.0
## 6         Pollos Maria 6.99    4.0     NA     NA     NA        3  4.0  5.0
##   Fillings Meat_filling Uniformity Salsa Synergy Wrap Reviewer overall Beef
## 1      3.5          4.0        4.0   4.0     4.0    4    Scott    3.80    1
## 2      2.5          2.0        4.0   3.5     2.5    5    Scott    3.00    1
## 3      3.0          4.5        4.0   3.0     3.0    5    Emily    3.00    0
## 4      3.0          4.0        5.0   4.0     4.0    5  Ricardo    3.75    1
## 5      3.5          4.5        5.0   2.5     4.5    4    Scott    4.20    1
## 6      3.5          2.5        2.5   2.5     4.0    1    Emily    3.20    0
##   Pico Guac Cheese Fries Sour_cream Pork Chicken Shrimp Fish Rice Beans Lettuce
## 1    1    1      1     1          0    0       0      0    0    0     0       0
## 2    1    1      1     1          0    0       0      0    0    0     0       0
## 3    1    1      0     0          0    1       0      0    0    0     0       0
## 4    1    1      0     0          0    0       0      0    0    0     0       0
## 5    1    0      1     1          0    0       0      0    0    0     0       0
## 6    0    1      1     0          1    0       1      0    0    1     1       1
##   Tomato Bell_peper Carrots Cabbage Sauce Cilantro Onion Taquito Pineapple Ham
## 1      0          0       0       0     0        0     0       0         0   0
## 2      0          0       0       0     0        0     0       0         0   0
```

```
## 3      0          0       0       0     0        0     0       0         0   0
## 4      0          0       0       0     0        0     0       0         0   0
## 5      0          0       0       0     0        0     0       0         0   0
## 6      1          0       0       0     0        0     0       0         0   0
##   Chile_relleno Nopales Lobster Egg Mushroom Bacon Sushi Avocado Corn Zucchini
## 1             0       0       0   0        0     0     0       0    0        0
## 2             0       0       0   0        0     0     0       0    0        0
## 3             0       0       0   0        0     0     0       0    0        0
## 4             0       0       0   0        0     0     0       0    0        0
## 5             0       0       0   0        0     0     0       0    0        0
## 6             0       0       0   0        0     0     0       0    0        0
```

There are 239 observations in this dataset.

Here are the different columns in the dataset:

```
colnames(burrito)
```

```
##  [1] "Location"      "Cost"          "Hunger"        "Length"
##  [5] "Circum"        "Volume"        "Tortilla"      "Temp"
##  [9] "Meat"          "Fillings"      "Meat_filling"  "Uniformity"
## [13] "Salsa"         "Synergy"       "Wrap"          "Reviewer"
## [17] "overall"       "Beef"          "Pico"          "Guac"
## [21] "Cheese"        "Fries"         "Sour_cream"    "Pork"
## [25] "Chicken"       "Shrimp"        "Fish"          "Rice"
## [29] "Beans"         "Lettuce"       "Tomato"        "Bell_peper"
## [33] "Carrots"       "Cabbage"       "Sauce"         "Cilantro"
## [37] "Onion"         "Taquito"       "Pineapple"     "Ham"
## [41] "Chile_relleno" "Nopales"       "Lobster"       "Egg"
## [45] "Mushroom"      "Bacon"         "Sushi"         "Avocado"
## [49] "Corn"          "Zucchini"
```

Cost

We are interested in cost. Let's see if there are any missing cost values, and then look at the distribution of costs.

```
which(is.na(burrito$Cost))
```

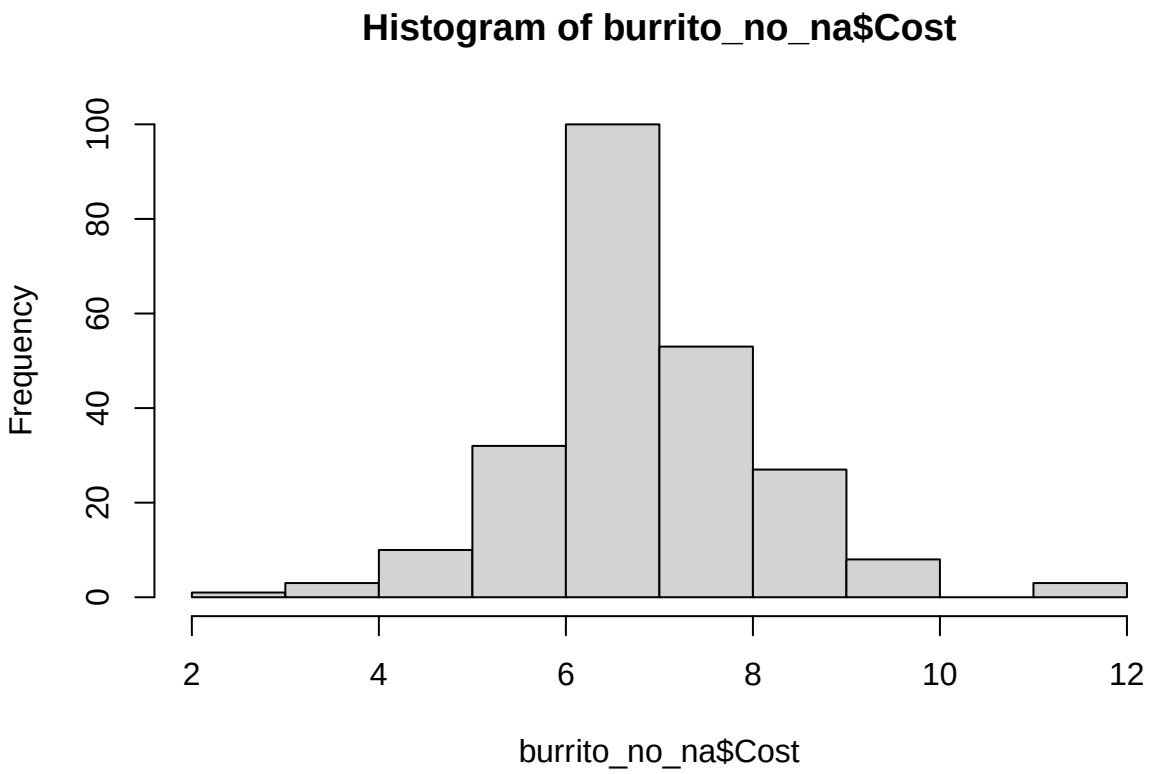
```
## [1] 113 135
```

There are two observations with `NA` costs. Lets remove these from the dataset and continue using it.

```
burrito_no_na = burrito[!is.na(burrito$Cost),]
nrow(burrito_no_na)
```

```
## [1] 237
```

```
hist(burrito_no_na$Cost)
```



Burrito costs appear to be somewhat normally distributed. This is good for our linear regression models.

Ingredients

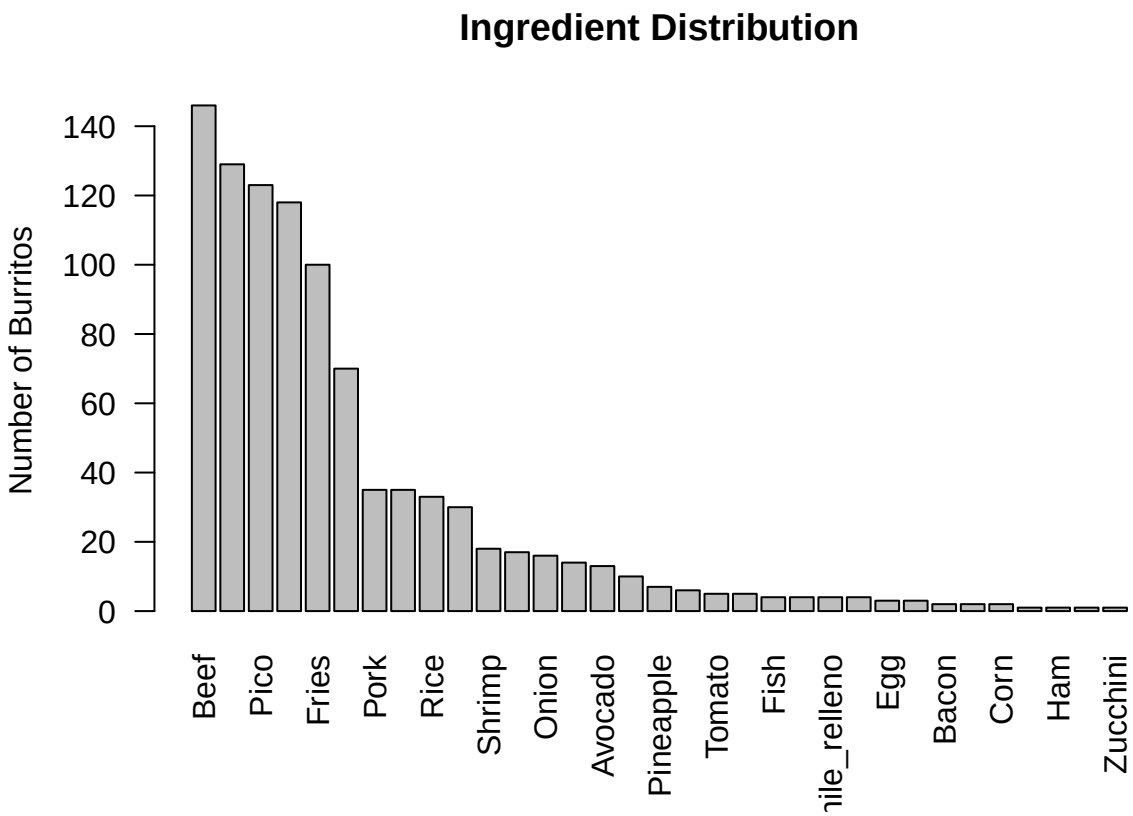
We now should now investigate the ingredients, see if there are any missing values, and then see if any ingredients need to be combined into an "Other" category.

```
ingredient_names = colnames(burrito_no_na)[18:50]
num_burrito_ingredients = c()
for (ingredient in ingredient_names) {
  num_burrito_ingredients = c(num_burrito_ingredients,
                              sum(burrito_no_na[ingredient]))
}
ingredient_counts_df = data.frame(ingredient=ingredient_names,
                                  count=num_burrito_ingredients)

# sort by count
ingredient_counts_df = ingredient_counts_df[order(ingredient_counts_df$count, decreasing=TRUE),]

kable(ingredient_counts_df, row.names=FALSE)
```

ingredient	count
Beef	146
Cheese	129

ingredient	count
Pico	123
Guac	118
Fries	100
Sour_cream	70
Pork	35
Sauce	35
Rice	33
Beans	30
Shrimp	18
Chicken	17
Onion	16
Cilantro	14
Avocado	13
Lettuce	10
Pineapple	7
Bell_peper	6
Tomato	5
Cabbage	5
Fish	4
Taquito	4
Chile_relleno	4
Nopales	4
Egg	3
Mushroom	3
Bacon	2
Sushi	2
Corn	2
Carrots	1
Ham	1
Lobster	1
Zucchini	1

```
barplot(ingredient_counts_df$count, ylab='Number of Burritos',
        main='Ingredient Distribution',
        names.arg=ingredient_counts_df$ingredient, las=2)
```



It looks like there are many ingredients where there are few burritos with them. These are good ingredients to group into an "Other" type category. Let's decide a cutoff:

```
kable(ingredient_counts_df[ingredient_counts_df$count < 10,], row.names=FALSE)
```

ingredient	count
Pineapple	7
Bell_peper	6
Tomato	5
Cabbage	5
Fish	4
Taquito	4
Chile_relleno	4
Nopales	4
Egg	3
Mushroom	3
Bacon	2
Sushi	2
Corn	2
Carrots	1
Ham	1
Lobster	1
Zucchini	1

Just by luck, it looks like all ingredients with more than 10 burritos are quite normal (Avocado, Cilantro,

Onion, ...), but all ingredients with fewer than 10 burritos are quite rare (Pineapple, Bell Pepper, Fish, Lobster, ...). Let's use 10 as our cutoff, and now define some categories to group these ingredients into.

Maybe some groups like this:

- Vegetables = (Pineapple, Bell Pepper, Tomato, Cabbage, Mushroom, Corn, Carrots, Zucchini)
- Breakfast = (Egg, Bacon, Ham)
- Other = (Fish, Taquito, Chille Relleno, Nopales, Sushi, Lobster)

These groups will be turned into new indicator variables that we can use with the other 15 ingredients (Beef through Lettuce). Note, we might need to change the Breakfast category because it still only adds up to only 6 burritos, not over 10.

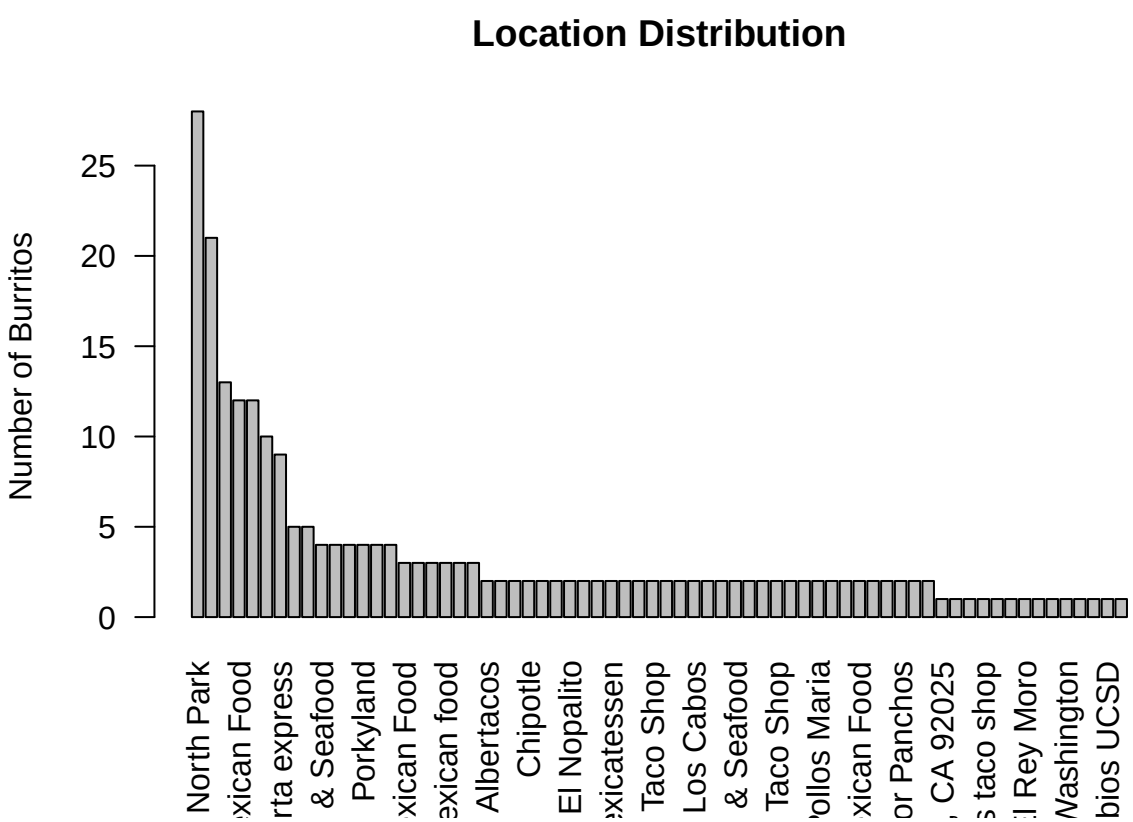
Restaurant Analysis

```
location_counts_df = aggregate(data.frame(count = burrito_no_na$Location),
                               list(location = burrito_no_na$Location), length)
# sort by count
location_counts_df = location_counts_df[order(location_counts_df$count, decreasing=TRUE),]

kable(location_counts_df, row.names=FALSE)
```

location	count
Lucha Libre North Park	28
California Burritos	21
Rigoberto's Taco Shop	13
Los Primos Mexican Food	12
Taco stand	12
Taco Stand	10
Vallarta express	9
El Zarape	5
Lolita's taco shop	5
Cancun Mexican & Seafood	4
Goody's	4
Lolita's Taco Shop	4
Porkyland	4
Roberto's Taco Shop Clairemont	4
Tony's Fresh Mexican Food	4
Colima's Mexican Food	3
La Perla Cocina	3
Lolita's Taco shop	3
Oscar's Mexican food	3
Taco Surf PB	3
Tacos por favor	3
Albertacos	2
Burros and Fries	2
Carmen's Mexican Food	2
Chipotle	2
El Cuervo	2
El dorado Mexican food	2

location	count
El Nopalito	2
El Torrito Foods	2
Graciela's Taco Shop	2
Jorge's Mexicatessen	2
Juanita's Taco Shop	2
JV's Mexican Food	2
Karina's Taco Shop	2
King Burrito	2
Lola's 7 Up Market & Deli	2
Los Cabos	2
Los tacos	2
Los Tacos	2
Mi Asador Mexican & Seafood	2
Mikes Taco Club	2
Netos Mexican Food	2
Nico's Taco Shop	2
Papa Chito's Mexican Food	2
Pokirrito	2
Pollos Maria	2
Raul's Mexican food	2
Rigoberto's Taco Shop La Jolla	2
Roberto's Very Mexican Food	2
Rudy's Taco Shop	2
Senor Grubby's	2
Senor Panchos	2
Sotos Mexican Food	2
Tacos La Bala	2
Alberto's 623 N Escondido Blvd, Escondido, CA 92025	1
Chili Peppers	1
Colima's	1
Donato's taco shop	1
El Indio	1
El Pueblo Mexican Food	1
El Rey Moro	1
Humbertos	1
Kotija Jr.	1
MXN on Washington	1
Pedro's Tacos	1
Qdoba Mexican Grill, Seatac Airport	1
Rubios UCSD	1
Saguaro's	1

```
barplot(location_counts_df$count, ylab='Number of Burritos',
        main='Location Distribution', names.arg=location_counts_df$location, las=2)
```



We see that the vast majority of locations in the dataset are only represented less than 5 times. We could possibly group these by region on a map (e.x. Downtown San Diego, ...) or we might only do analysis on the top restaurants. We'd have to check that subsetting on the top restaurants reduces the number of ingredients in the analysis however. Do you have any suggestions?