Stat108_FinalProject

Dylan Scoble and Aisha Lakshman

2/4/2022

«««< HEAD:Analysis.Rmd

Introduction

In the 2004 documentary Super Size Me, writer and director Morgan Spurlock took on a month-long challenge to only eat McDonalds food. Spurlock experienced a multitude of health issues, including weight gain, cholesterol spike, and negative impacts on his energy and mood, demonstrating the fast-food chains' instrumental role in America's obesity epidemic (Stossel 2006). Spurlock's film not only emphasized the consequences of caloric intake, but also brought light to the nutritional attributes of McDonalds menu items that caused adverse health effects. There are many factors that impact the quality and quantity of calories, such as levels of fat, protein, and carbohydrates, which is why many dieticians support the notion that "not all calories are created equal" (Tolar-Peterson, 2021). Spurlock's documentary and existing literature inspired an investigation of McDonalds menu items' caloric and nutritional records. Our research will address the following question: What nutritional attribute is the best predictor of calories for the McDonalds menu items? Since drink and food items have quite different caloric makeups, we are dividing our dataset between food menu items (Breakfast, Beef & Pork, Chicken & Fish, Salads, Snacks & Sides, Desserts) and drink menu items (Coffee and Tea, Smoothies and Shakes, Beverages). For food menu items, we hypothesize total carbohydrates (grams) is the most accurate predictor of calories. For drink menu items, we hypothesize that total sugar (grams) is the most accurate predictor of calories We will analyze a 2018 dataset from Kaggle titled "Nutritional Facts for McDonald's Menu" to answer our research question. Our chosen dataset provides nutritional information for all of McDonald's menu items, including calories, saturated fat, and cholesterol levels. We will create a linear model for each nutritional attribute with calories as the response variable for each predictor. A linear model for regression analysis is useful in answering our question because it will allow us to confidently determine what nutritional attributes matter the most for calories and predict an item's calorie count based on its predictors.

References

Tolar-Peterson, Terezie. 2021. "Not all calories are created equal - a dietician explains the different ways the kinds of foods you eat matter to your body". The Conversation. Retrieved Febuary 8th, 2022. https://theconversation.com/not-all-calories-are-equal-a-dietitian-explains-the-different-ways-the-kinds-of-foods-you-eat-matter-to-your-body-156900

Stossel, John. 2006. " 'Super Size Me' Carries Weight With Critics". ABC News. Retrieved Febuary 8th, 2022. https://docs.google.com/document/d/1XB-22QylvnbasBKe7n_DfkkENcZWRvIR5X8vZgOK6LY/edit#

Our Data

```
data <- read.csv("data/menu 2.csv")
glimpse(data)
```

```
## Rows: 260
## Columns: 24
## $ Category                      <chr> "Breakfast", "Breakfast", "Breakfast", "~
## $ Item                          <chr> "Egg McMuffin", "Egg White Delight", "Sa~
## $ Serving.Size                  <chr> "4.8 oz (136 g)", "4.8 oz (135 g)", "3.9~
## $ Calories                      <int> 300, 250, 370, 450, 400, 430, 460, 520, ~
## $ Calories.from.Fat             <int> 120, 70, 200, 250, 210, 210, 230, 270, 1~
## $ Total.Fat                     <dbl> 13, 8, 23, 28, 23, 23, 26, 30, 20, 25, 2~
## $ Total.Fat....Daily.Value.     <int> 20, 12, 35, 43, 35, 36, 40, 47, 32, 38, ~
## $ Saturated.Fat                 <dbl> 5, 3, 8, 10, 8, 9, 13, 14, 11, 12, 12, 1~
## $ Saturated.Fat....Daily.Value. <int> 25, 15, 42, 52, 42, 46, 65, 68, 56, 59, ~
## $ Trans.Fat                     <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, ~
## $ Cholesterol                   <int> 260, 25, 45, 285, 50, 300, 250, 250, 35,~
## $ Cholesterol....Daily.Value.   <int> 87, 8, 15, 95, 16, 100, 83, 83, 11, 11, ~
## $ Sodium                        <int> 750, 770, 780, 860, 880, 960, 1300, 1410~
## $ Sodium....Daily.Value.        <int> 31, 32, 33, 36, 37, 40, 54, 59, 54, 59, ~
## $ Carbohydrates                 <int> 31, 30, 29, 30, 30, 31, 38, 43, 36, 42, ~
## $ Carbohydrates....Daily.Value. <int> 10, 10, 10, 10, 10, 10, 13, 14, 12, 14, ~
## $ Dietary.Fiber                 <int> 4, 4, 4, 4, 4, 4, 2, 3, 2, 3, 2, 3, 2, 3~
## $ Dietary.Fiber....Daily.Value. <int> 17, 17, 17, 17, 17, 18, 7, 12, 7, 12, 6,~
## $ Sugars                        <int> 3, 3, 2, 2, 2, 3, 3, 4, 3, 4, 2, 3, 2, 3~
## $ Protein                       <int> 17, 18, 14, 21, 21, 26, 19, 19, 20, 20, ~
## $ Vitamin.A....Daily.Value.     <int> 10, 6, 8, 15, 6, 15, 10, 15, 2, 6, 0, 4,~
## $ Vitamin.C....Daily.Value.     <int> 0, 0, 0, 0, 0, 2, 8, 8, 8, 8, 0, 0, 0, 0~
## $ Calcium....Daily.Value.       <int> 25, 25, 25, 30, 25, 30, 15, 20, 15, 15, ~
## $ Iron....Daily.Value.          <int> 15, 8, 10, 15, 10, 20, 15, 20, 10, 15, 1~
```

Exploratory Data Analysis

For this project, we understand that foods and beverages may have different predictors for their number of calories. Therefore, we will be splitting our dataset into two different dataframes: one for foods, and one for beverages.

We will also be removing all predictors that have "as % of Daily Value" attached at the end, since our purpose is not focused the daily values of the nutrients. These predictors add no value to our dataset or models.

```
data <- data %>%
  select(Category, Item, Serving.Size, Calories, Calories.from.Fat, Total.Fat,
         Saturated.Fat, Trans.Fat, Cholesterol, Sodium, Carbohydrates,
         Dietary.Fiber, Sugars, Protein, Vitamin.A....Daily.Value.,
         Vitamin.C....Daily.Value., Calcium....Daily.Value., Iron....Daily.Value.)

glimpse(data)
```

```
## Rows: 260
## Columns: 18
## $ Category                  <chr> "Breakfast", "Breakfast", "Breakfast", "Brea~
## $ Item                      <chr> "Egg McMuffin", "Egg White Delight", "Sausag~
## $ Serving.Size              <chr> "4.8 oz (136 g)", "4.8 oz (135 g)", "3.9 oz ~
## $ Calories                  <int> 300, 250, 370, 450, 400, 430, 460, 520, 410,~
## $ Calories.from.Fat         <int> 120, 70, 200, 250, 210, 210, 230, 270, 180, ~
## $ Total.Fat                 <dbl> 13, 8, 23, 28, 23, 23, 26, 30, 20, 25, 27, 3~
## $ Saturated.Fat             <dbl> 5, 3, 8, 10, 8, 9, 13, 14, 11, 12, 12, 13, 1~
## $ Trans.Fat                 <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0,~
## $ Cholesterol               <int> 260, 25, 45, 285, 50, 300, 250, 250, 35, 35,~
## $ Sodium                    <int> 750, 770, 780, 860, 880, 960, 1300, 1410, 13~
## $ Carbohydrates             <int> 31, 30, 29, 30, 30, 31, 38, 43, 36, 42, 34, ~
## $ Dietary.Fiber             <int> 4, 4, 4, 4, 4, 4, 2, 3, 2, 3, 2, 3, 2, 3, 2,~
## $ Sugars                    <int> 3, 3, 2, 2, 2, 3, 3, 4, 3, 4, 2, 3, 2, 3, 3,~
## $ Protein                   <int> 17, 18, 14, 21, 21, 26, 19, 19, 20, 20, 11, ~
## $ Vitamin.A....Daily.Value. <int> 10, 6, 8, 15, 6, 15, 10, 15, 2, 6, 0, 4, 6, ~
## $ Vitamin.C....Daily.Value. <int> 0, 0, 0, 0, 0, 2, 8, 8, 8, 8, 0, 0, 0, 0, 0,~
## $ Calcium....Daily.Value.   <int> 25, 25, 25, 30, 25, 30, 15, 20, 15, 15, 6, 8~
## $ Iron....Daily.Value.      <int> 15, 8, 10, 15, 10, 20, 15, 20, 10, 15, 15, 1~
```

```
count(data, Category)
```

```
##             Category  n
## 1        Beef & Pork 15
## 2          Beverages 27
## 3          Breakfast 42
## 4     Chicken & Fish 27
## 5       Coffee & Tea 95
## 6           Desserts  7
## 7             Salads  6
## 8 Smoothies & Shakes 28
## 9     Snacks & Sides 13
```

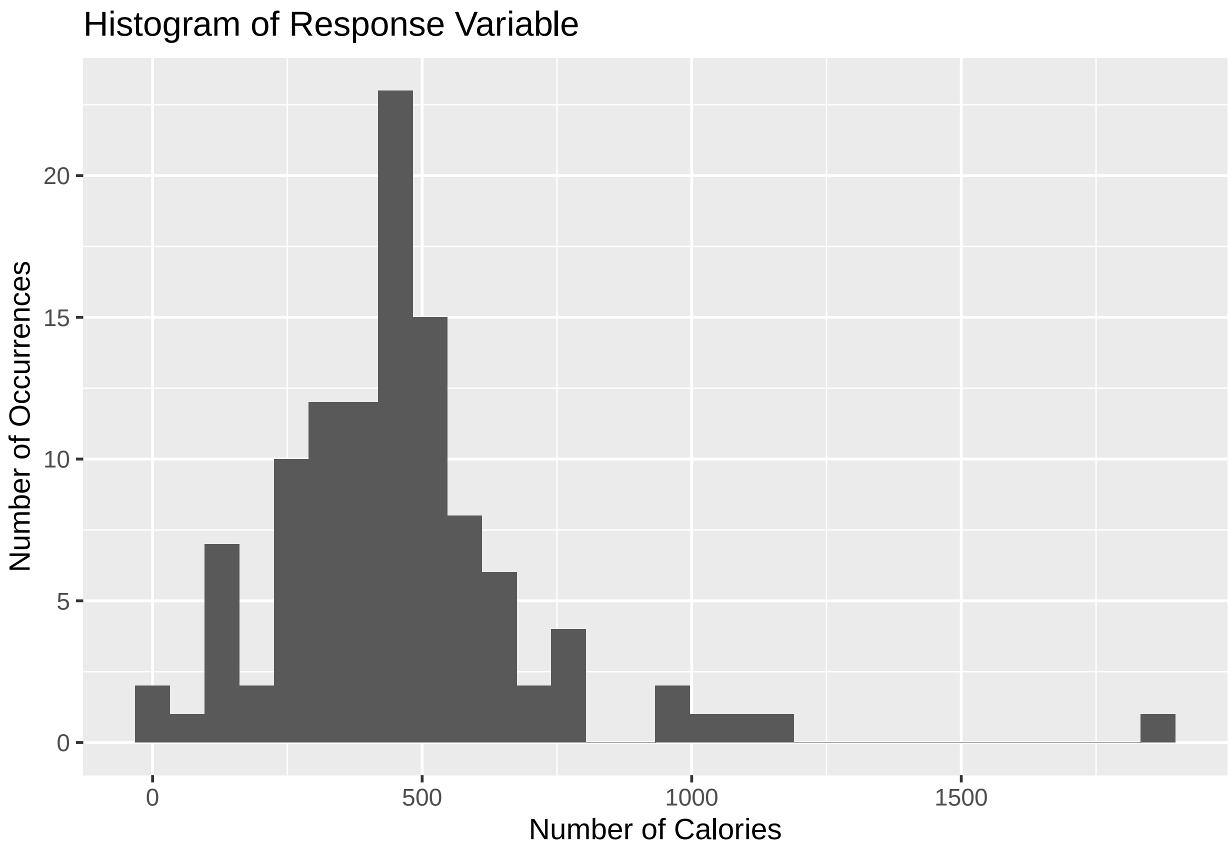
```
food_data <- data %>%
  filter(Category == "Beef & Pork" |
         Category == "Breakfast" |
         Category == "Chicken & Fish" |
         Category == "Desserts" |
         Category == "Salads" |
         Category == "Snacks & Sides")

bev_data <- data %>%
  filter(Category == "Beverages" |
         Category == "Coffee & Tea" |
         Category == "Smoothies & Shakes")
```

Response Variable

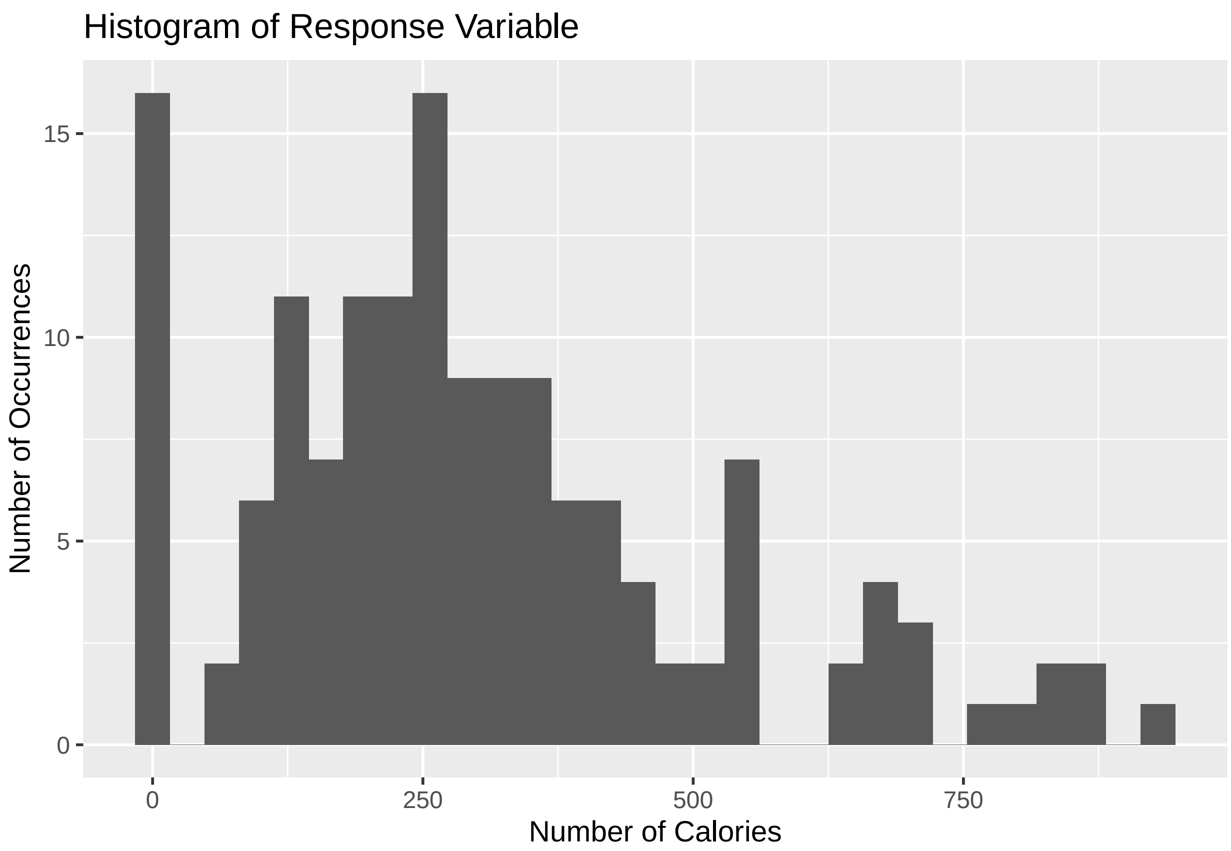
```
ggplot(data=food_data, aes(x=Calories)) +
  geom_histogram() +
  labs(title="Histogram of Response Variable",
       x="Number of Calories",
       y="Number of Occurrences")
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



```
ggplot(data=bev_data, aes(x=Calories)) +
  geom_histogram() +
  labs(title="Histogram of Response Variable",
       x="Number of Calories",
       y="Number of Occurrences")
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



```
food_data %>%
  summarise(mean = mean(Calories),
            median = median(Calories),
            std_dev = sd(Calories),
            iqr = IQR(Calories))
```

```
##       mean median  std_dev iqr
## 1 462.0909    445 249.3343 210
```

```
bev_data %>%
  summarise(mean = mean(Calories),
            median = median(Calories),
            std_dev = sd(Calories),
            iqr = IQR(Calories))
```

```
##       mean median  std_dev iqr
## 1 299.4667    270 208.8215 235
```