**CSC 435 LAB 02 Points: 20**

**Data Cleansing and Preparation**

**Z-score and Null Hypothesis**

Your Team Member names:

1. Stryder R. Coleman

2. Padraic Darby

**Part A: Data Cleansing and Preparation**

# Background

Data on Candy bars is the sample data of the name of the candy and its nutrition facts. The data file possibly has many errors due to manual data entry and needs to be cleansed before using it to build the model.

# The Task

The task is to evaluate the quality of the data and cleanse the raw data before it can be used to build models. This requires **exploring data, recoding impure data, identifying missing data, handling outliers, and deleting duplicate data**.

# Purpose:

This lab is designed for you to demonstrate the data cleansing & data preparation tasks. After completing the tasks in this lab, you should be able to:

1. Create Summary statistics of data
2. Recode inaccurate data
3. Identify outliers if any and decide whether to exclude it or keep it
4. Identify and remove duplicate data
5. Assess and impute missing data
6. Z-score and Null hypothesis

# Data File:

Lab2\_Candy Bars.csv

# Perform the following tasks.

1. Understand the data by Filling the following summary table of each variable. Briefly explain the summary statistics.

| **Columns** | **N** | **N Missing** | **N Categories** | **Min** | **Max** | **Mean** | **Std Dev** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Brand | 96 | 0 | 26 | . | . | . | . |
| Name | 96 | 0 | 75 | . | . | . | . |
| Serving/pkg | 96 | 0 | Na | 1 | 4.5 | 1.25 | 0.785 |
| Oz/pkg | 96 | 0 | Na | 1.2 | 6 | 2.09 | 1.05 |
| Calories | 93 | 3 | Na | 125 | 27000 | 535 | 2775 |
| Total fat g | 95 | 1 | Na | 0 | 29 | 12.3 | 6.02 |
| Saturated fat g | 96 | 0 | Na | 0 | 15 | 6.15 | 3.55 |
| Cholesterol g | 95 | 1 | Na | 0 | 20 | 5.2 | 5.18 |
| Sodium mg | 93 | 3 | Na | 5 | 210 | 76.2 | 48.3 |
| Carbohydrate g | 96 | 0 | Na | 17 | 1500 | 45.7 | 150 |
| Dietary fiber g | 96 | 0 | Na | 0 | 4 | 1.11 | 0.893 |
| Sugars g | 92 | 4 | Na | 1 | 2450 | 50 | 253 |
| Protein g | 96 | 0 | Na | 0 | 9 | 3.67 | 2.24 |
| Vitamin A %RDI | 92 | 4 | Na | 0 | 6 | .25 | 0.979 |
| Vitamin C %RDI | 96 | 0 | Na | 0 | 50 | .635 | 5.19 |
| Calcium %RDI | 94 | 2 | Na | 0 | 20 | 4.77 | 3.84 |
| Iron %RDI | 96 | 0 | na | 0 | 15 | 2.88 | 2.8 |

Most of the data has an accurate mean and std dev because the min and max are reasonable. However, there are several variables with outliers that skew the mean and mess up the Std Dev.

1. Cleanse the data in column “**Brand**” by Recoding the data. Copy and paste your code and briefly explain the result.

Lab2\_Candy\_Bars$Brand <- str\_to\_title(Lab2\_Candy\_Bars$Brand)

The code above will capitalize the first character in every word and it will make every other character in the string into a lowercase character..

1. Identify and remove duplicate data records. How many duplicate records were there? What are the records and Identify the number of duplications for each record?

There are 21 duplicates. Here is a chart to show the number of duplicates for each duplicate and their identity.

Chart

Description automatically generated

1. Identify outliers among the continuous variables. Use **Box plots** to identify outliers. Copy and paste 2 charts of your choice below and interpret the charts. Discuss whether these outliers should be excluded or kept. Why or why not?

Graphical user interface, application, table

Description automatically generatedGraphical user interface, application

Description automatically generated

The 26000-calorie outlier and the 1500 carbohydrate outlier should be excluded. Since both values are so much greater than the rest, they would negatively affect the maximum value, the mean, and the standard deviation of the graph.

1. Finally, provide a summary of missing values in the data. How many missing values for each variable? How do you think the missing values should be handled? Just explain your reasons.

The chart below shows the number of missing values per variable. I think the missing data should be removed since the data is missing completely at random therefore it is safe to remove the observations with NAs.

|  |  |
| --- | --- |
| Variable | # of missing values |
| Brand | 0 |
| Name | 0 |
| Serving/pkg | 0 |
| Oz/pkg | 0 |
| Calories | 3 |
| Total fat g | 1 |
| Saturated fat g | 0 |
| Cholesterol g | 1 |
| Sodium mg | 3 |
| Carbohydrate g | 0 |
| Dietary fiber g | 0 |
| Sugars g | 4 |
| Protein g | 0 |
| Vitamin A %RDI | 4 |
| Vitamin C %RDI | 0 |
| Calcium %RDI | 2 |
| Iron %RDI | 0 |

**Part B: Z-score and Null Hypothesis**

1. A mean for a normal distributed test of 45 students is 70 with a standard deviation of 10.

Z-score = (X-Mean)/Standard Deviation

1. If a student scored 65% on the test how well did, he or she do as compared to other students?

m = 70 n = 45 SD = 10 x = 65

(65 – 70)/10= -5/10 = -.5 SD

1–0.30854 = 0.69146

They scored better than 30.9% of students but lower than 69.1% of the students.

1. What is the probability of a student scoring more than 90?

m = 70 n = 45 SD = 10 x = 90

1. – 70)/10 = 20/10 = 2 SD

1–0.97725= 0.02275

They scored better than 97.7% of students but lower than 2.3% of the students.

1. A recent survey of college campuses across SUNY claims that students spend an average of 2.7 hours a day using their cell phones. A random sample of 35 College at Brockport students showed an average use of 2.9 hours a day, with a standard deviation of 0.4 hours. Do College at Brockport students use their cell phones more than the typical SUNY Colleges student?

Write the Null & Alternative hypothesis for the above problem?

n = 35 s = 2.9 sd = 0.4 m = 2.7

Null Hypothesis: Brockport students using cell phones more than the typical SUNY Colleges student.

Alternative Hypothesis: Brockport students using cell phones less or equal to the typical SUNY Colleges student.