Exploratory data analysis

Notas fiscais electronica

By Andre Luis

Introduction

This document is going to describe the process to know how much a customer will spend and the sales forecast for the next week.

I am going to use R language to help to understand the dataset behavior, and also to calculate and make the predictions.

The dataset I am using was available by Totvs at this address: https://github.com/TOTVS/MDMStatic/blob/master/code-challenge/TOTVS%20Labs%20-%20AI%20Challenge%20-%20Dataset.zip?raw=true.

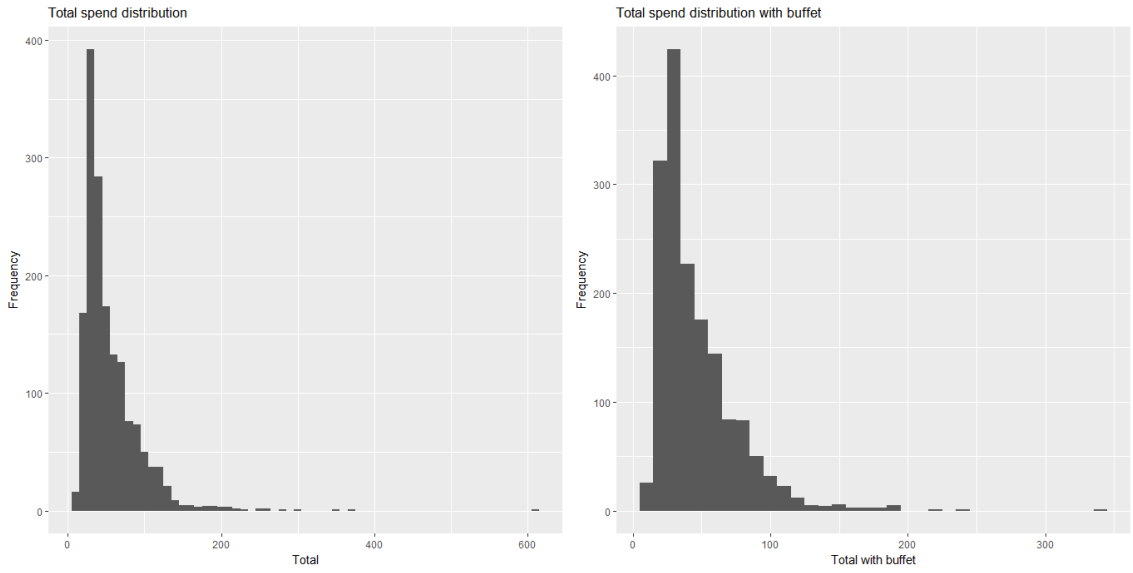
Analysis

I create two dataset. The first one has the nota fiscal data and the second has the products from the nota fiscal data. The nota fiscal data dataset has 1635 and the product dataset has 3547 observations from the nota fiscal electronica. Both dataset have some attributes shared (mesa, dataemissao and week) which allows us to cross data.

Product dataset.

There are 3547 observations in the product dataset. From it, there are 1635 observation where the product is a buffet, what means the amount of buffet is equals the number of customer.

I figured out the price of the Buffet product has a variance by nota fiscal eletronica. So, I plot the next two histogram to compare the data distribution from the nota fiscal total and the distribution from the buffet price.



Both histogram have a similar distribution with a positively skewed distribution with outliers' values.

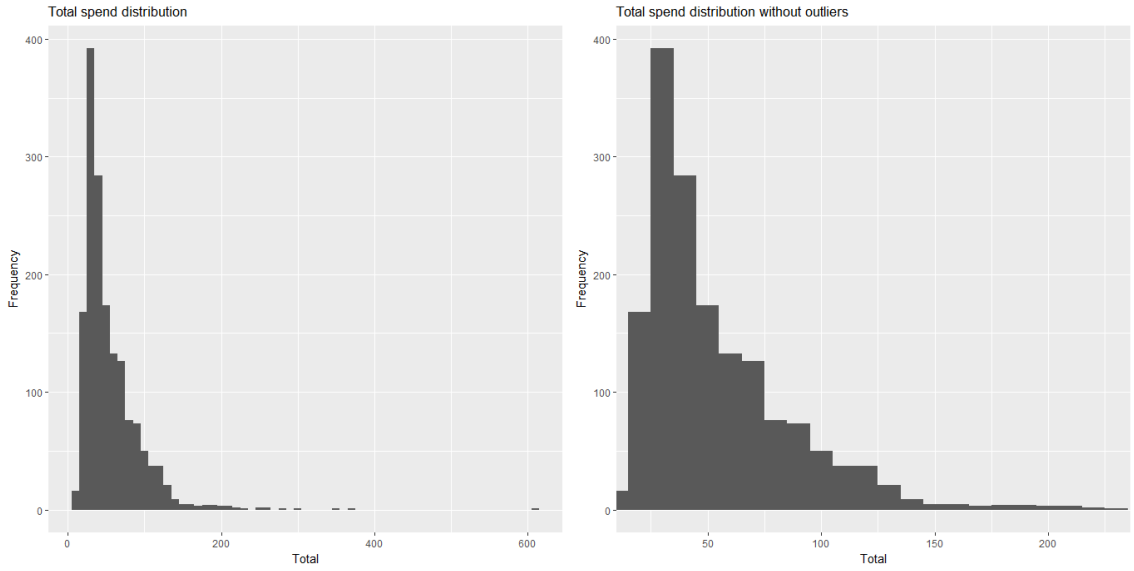
After filtered the products with price above 200, there were 4 observations, 3 buffets and one Sushi Especial. These outliers could be mistyped value, because the values are higher above the price product mean, 25.58, however, this price can also mean a family or group of people together due the amount of product in the nota fiscal where the price was high.

Based on the distribution from the last two histogram, I think the product dataset can be disregarded and we can make the predictions based on the nota fiscal dataset. From now on, I am going to use just the nota fiscal dataset.

Nota fiscal dataset.

Valortotal is the property name from the total spend by nota fiscal in the dataset.

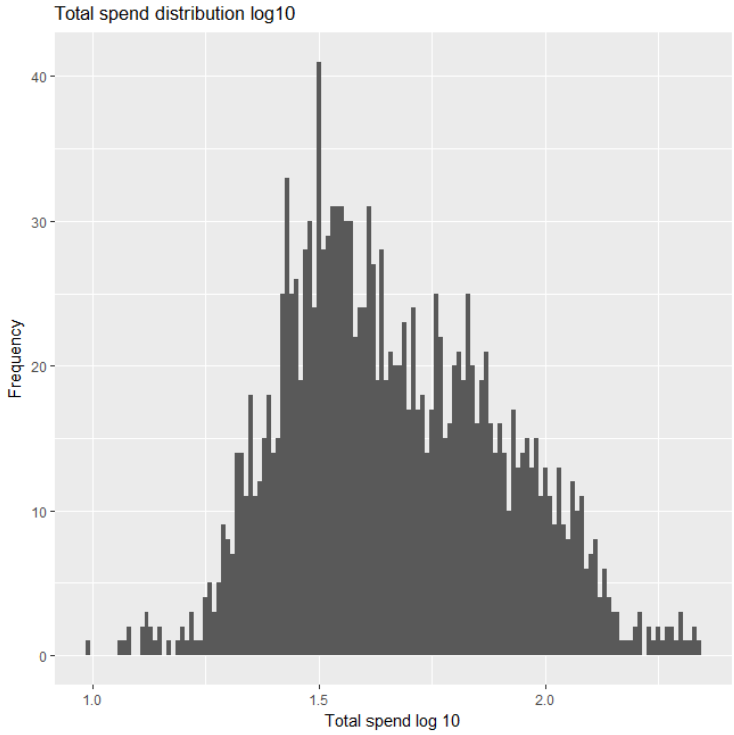
Analyzing the nota fiscal dataset, two histogram was created to describe the valortotal distribution.



They show a distribution positively skewed and the first one shows the data with the outliers and the second disregards them. The nota fiscal total above 226 have just 10 observations, so I considered them outliers.

Predicting the value spend by nota fiscal.

The next histogram chart has the same dataset without outlier and it was applied the log10 in the total value (x-axis), and now it is close a normal distribution.



The rules for normally distributed data says that 2 standard deviation below and above the mean represent around 95% from the population.

Applying the log 10 for valortotal, we have:

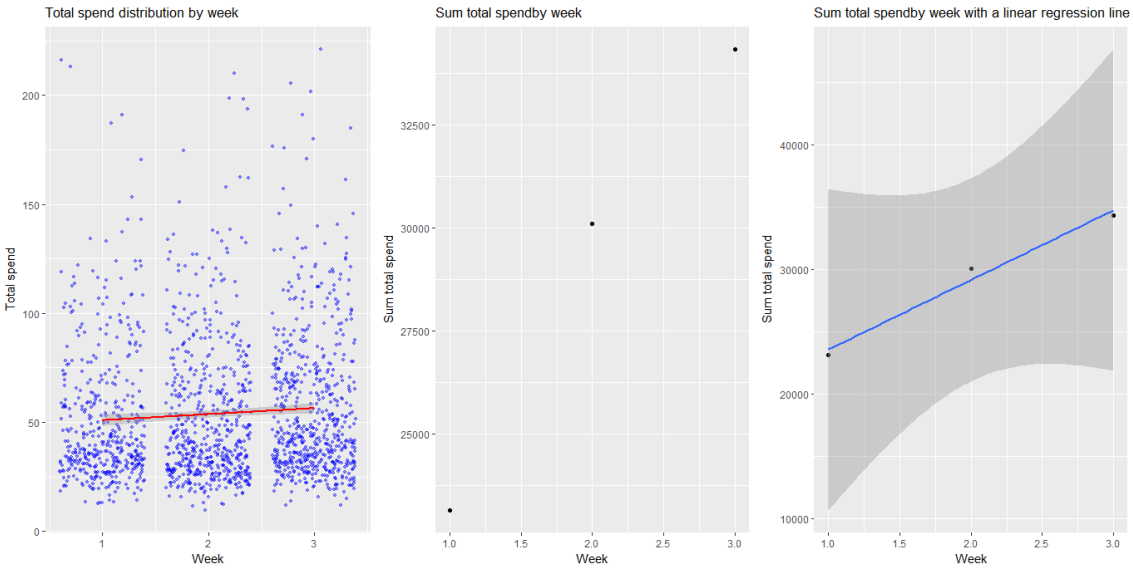
Mean = 1.673;

Standard deviation = 0.243;

95% of the nota fiscal valor total are from 15.66 (1.194) to 136.66 (2.135) with the mean 47.09 (1.673).

Predicting the sales forecast for the next week.

The next image shows 3 scatters diagram from valortotal by week without outliers.



The first diagram shows the distribution from the values and its concentration by week. I figured out a small increase from the first week to the second week and from the second to the third week. The red line better represent this small increase. Then, in the second diagram, the sum of all the values was plotted by week and it is clear a positive linear correlation. In the third diagram, a linear regression line was drawn and the points is very close to the line.

Based on the third chart, I figured out that it is possible to use linear regression to predict the sales forecast for the next week.

Using lm function from r language, I found the r-square, the fit value and the confidence interval.

R-squared is around 98.06% (0.9896), what indicate the variance of the week predicts the sum of the values spend by nota fiscal.

The total sum (fit value) is 40371.57, and the confidence interval is from 14603.75 to 66139.40.

Restaurant's best location

The dataset has an additional information informing the table ("mesa") where the customer sit. The frequency can indicate the best place from the restaurant area.

The table named "mesa 22" has a frequency of 26, followed by the mesa 36, mesa 1 (25), mesa 10 (24), mesa 21 (24), mesa 4 (24), mesa 5 (24), mesa 7 (24). This location is chose by more than 10% of the customer. There isn't enough information if this table are close from each other, but this information can be important to help to distribute the waiters in the restaurant area.

Conclusion.

After analyzing the data set, I figured out there was not fields neither with missing information nor TypeError, therefore, it was not necessary neither create method to clean nor to fix the dataset. The only reason to change the dataset were the outliers.

From the rules for normally distributed data, the mean plus 2 standard deviation above and below, represents 95% of the population. Applying log10 to the total value spend by customer change the distribution to normal.

After calculating the standard deviation and the mean, I conclude for 95% of the customer will spend from 15.66 (1.194) to 136.66 (2.135) and the probably value is around 47.09 (1.673).

I used linear regression to calculate the sales forecast for the next week (week 4). The total sum sales for the next week is 40371.57 (fit value), and the confidence interval is from 14603.75 to 66139.40.