Coursera IBM Data Science Capstone Project

New City Same Hood

Authors:
William Ngana
195801950

Toronto, ON

Date Started: Feb,2020

1 Introduction

1.1 Background

Toronto is Canada's largest city, I has many different neighbourhoods with many communities. As a prospective business owner you when establishing your business you need to choose an area that would be optimal for you. The area that a business is built in can make or break it. Build in the wrong place a business can fail, but built in the right place the business will thrive and grow. Having a proper understanding of the community and the business that are already there is of great importance to prospective business owners.

1.2 Problem

Some data that can be looked at to figure out the best place to build you new business would be, the restaurants that already exist in the area, the demographics of the area, for example building an Italian restaurant in Lil' Italy would probably be a good idea. As well since the rent price in Toronto is so high we can look at price range. If you are looking to start a high end restaurant you wouldn't want to build in an with a concentration of fast food. That will be the aim of the project to locate the best areas in the City of Toronto to build a business by trying to groups the different post codes in Toronto and seeing which areas are similar.

1.3 Interest

This problem would be of interests to anyone looking to start a restaurant or getting into the restaurant business in Toronto and trying to identify a place to put it.

2 Data Acquisition and Cleaning

2.1 Sources of Data

The main source of data that I'm going to use in this project comes from here this is a table of all the boroughs and neighbourhoods in Toronto. I will also be making calls to ForSquare api and from there be able to extract relevant venue data for each borough. The api call will then return the a list of all the top 50 venues in the area and the type of venue it is.

2.2 Data Cleaning and Feature Selection

The data that was scrapped from wiki and the data that was retrieved from the api call have been combined to form one table. This was done in several steps that I will outline below

The data frame that was the result of scraping the wiki page had various miss in data points, so I am needed to clean up the data a bit. I will do this by dropping the rows in which the value of the borough is "Not assinged". Then I will combine all the rows with the same postcode value and combine the neighbourhoods separated by a comma. Using data from online I then added the Longitude and Latitude for each postcode.

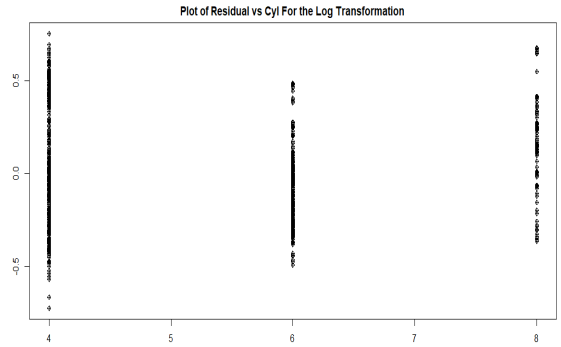
The next step was to take the data frame I had and for each Post Code determine the top 50 venues that were around it. This was done with the getNearbyVenues method which returned a dataframe with the post code the latitude and longitude of the neigbourhood the venue and the longitude and latitude of the venue as well as the venue category. For this particular project I didn't require all of the columns so I'm going to create a new dataframe where the Neighborhood and venue Latitude/Longitude is dropped. For this project what we are look for most is the kinds of venues in each area. So our features will be the Postal Code of the area and the number of each kind of venue of each possible categories. To do this we will first need to extract the unique categories from the dataframe then sum up how many of each are in a particular postal code.

[PostalCode,Venue category 1,Venue category 2,....] and so on will be the columns of the data frame. The final step was to determine how many of those occure in each neighbourhood. To do this we first need to group by the neighbourhood then count the number of occurrences of a particular type in a neighborhood.

3 Analysis

Appendices

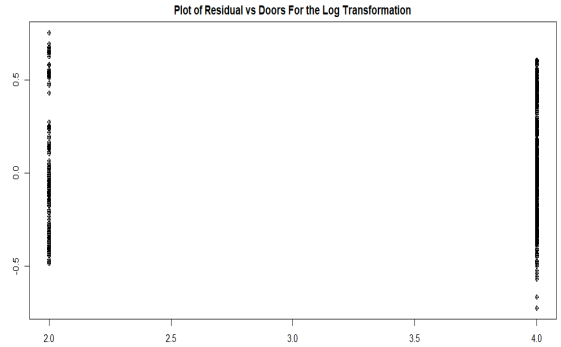
A Log Transformation



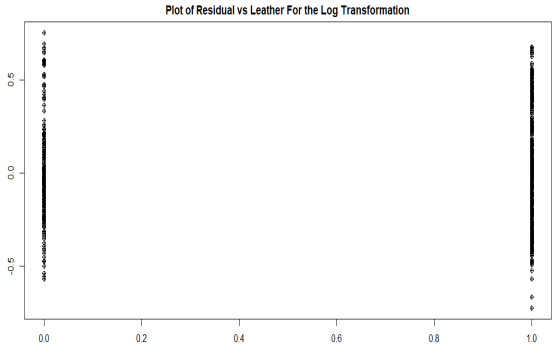
(a) Residual plot of the variable Cyl with the log transformed price

(b) Residual plot of the variable Cruise with the log transformed price

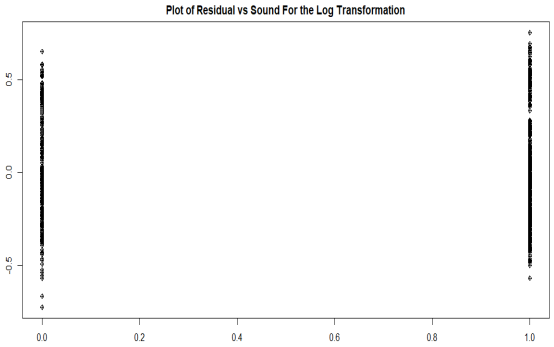
Figure 1: Explanatory Variable Log Transformation



(a) Residual plot of the variable Door with the log transformed price



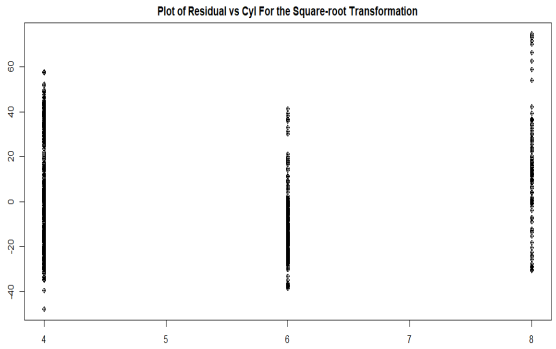
(b) Residual plot of the variable Leather with the log transformed price



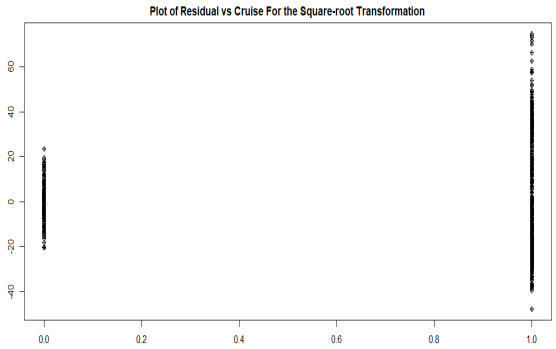
(a) Residual plot of the variable Cyl with the log transformed price

Figure 2: Explanatory Variable Log Transformation

B Square-root Transformation

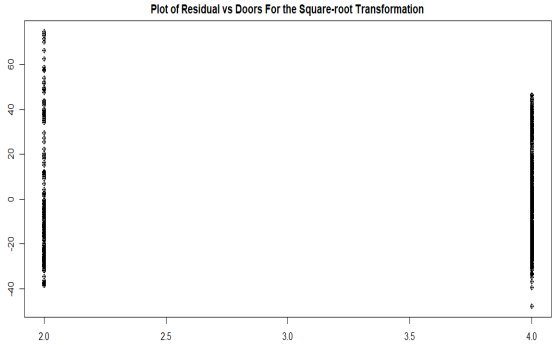


(a) Residual plot of the variable Cyl with the Square-root transformed price

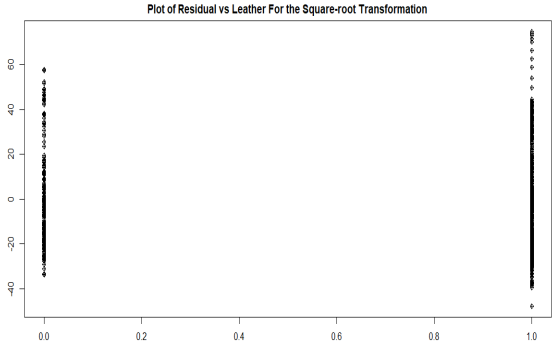


(b) Residual plot of the variable Cruise with the Square-root transformed price

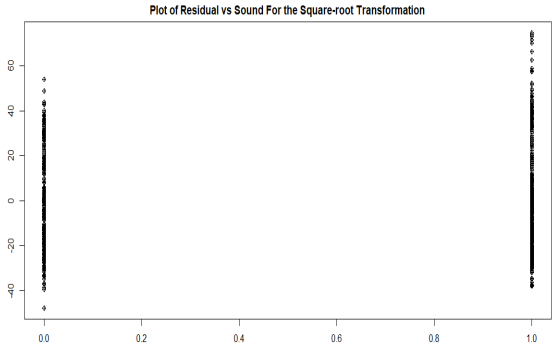
Figure 4: Explanatory Variable Square-root Transformation



(a) Residual plot of the variable Door with the Square-root transformed price



(b) Residual plot of the variable Leather with the Square-root transformed price



(c) Residual plot of the variable Sound with the Square-root transformed price

Figure 5: Explanatory Variable sqrt Transformation