Exploratory Data Analysis

Emmett Powers

11/20/2019

Project Update and Overview:

In this project, we're exploring a section of the Yelp dataset, which includes dozens of variables regarding hundreds of thousands of businesses in 10 metropolitan areas. We'd like to eventually construct a classification model that can predict the price category of a restaraunt (1, 2, 3, 4 are Yelp's categories) based off of some number of variables. Before we can do this, however, we want to play with our data. For one, we want to restrict our data frame to include only restaraunts. Additionally, we want to get an understanding of our variables and their covariance, both in order to make the best model and also just to understand interesting trends in the data. For example, maybe there are differences in variables depending on region. The goal of our EDA is to get a big picture of our data and its nuances, and hone in on variables we want to include in our model.

Provenance

Public Data Set from Yelp - Link: https://www.yelp.com/dataset/challenge Unit of observation: Price categories 1-4 (1 being cheapest and 4 most expensive). There are a total of 6,685,900 reviews each comprised of a varietyo of variables from price range, to category of location, through the actual start rating (1-5 scale).

Exploration

Data Prep

Here we upload the business section of the Yelp dataset. Note: we are not using the text review or photos parts of the Yelp data set.

```
library(jsonlite)
business <- stream_in(file("business.json"))
```

We want to restrict the data frame to include only restauraunts. We do this by including only rows that mention the category "restauraunts" in the categories column. "'{r message = FALSE} library(rlist) library(dplyr) library(tidyverse)

clean.business <- business clean.business$Restaurant < -0new.business < -mutate(clean.business, Restaurant = grepl("Restaurants", clean.businesscategories)) clean.yelp <- filter(new.business, Restaurant == "TRUE") # not sure how this worked but if it ain't broke don't fix it

Removing attribute label from data to allow further cleaning

attribute <- clean.yelp$attributes$restaurant_data < -clean.yelp$restaurant_data$attributes <- NULL restaurant_data <- cbind(restaurant_data, attribute)

Now that we have a solid dataset of restaurants, we want to get rid of variables that we don't care to examine or ## that mostly contain N/As.

Variables of interest

voi <- c("name", "city", "state", "latitude", "longitude", "stars", "review_count", "RestaurantsTakeOut", "RestaurantsPriceRange2", "OutdoorSeating", "Alcohol", "categories")

Final Cleaned Data (Subject to Change)

rdata <- subset(restaurant_data, select=voi) summary(rdata)

"'

Missingness

In order to illustrate missingness we used the Amelia package's "missmap" function. Here we see we are missing about 6% of our observations, which is not too bad and thus moving forward we will likely simply remove all observations which are missing since: 1. We have no reason to believe there exists any bias in the missing observations, this seems to be rather random, 2. We would still have 94% of the data set available. Of note: Most of the missingness appears to be centered around more niche variables such as Outdoor Seating or Alcohol.

```
library(dplyr)
library(Amelia)
missmap(rdata)

## Cleaning Data
#rdata_clean <- rdata[complete.cases(rdata),]
```

Univariate Analysis of Response

```
rdata$RestaurantsPriceRange2 <- strtoi(rdata$RestaurantsPriceRange2)

summary(rdata$RestaurantsPriceRange2)
hist(rdata$RestaurantsPriceRange2)
boxplot(rdata$RestaurantsPriceRange2)
```

A use of the summary function on our response function yields the following: our Mean value is 1.672, and our median is 2. Further, if we look at our histogram we see that our observations are clustered around 1 & 2, with few around 3 & 4. Our boxplot highlights the same point.

Bi-trivariate analyses

```
## Missing Matrix & Elaboration on different X variables

plot(rdata$stars, rdata$RestaurantsPriceRange2,)
plot(rdata$review_count, rdata$RestaurantsPriceRange2)
```

Here we have graphed two potentially important predictors: Star rating and number of reviews. At first glance, the scatterplots do not appear to reveal any meaningful relationship.