Success Predictors in Vegetarian Restaurants: A Yelp Review Sentiment Study

Can we predict vegetarian restaurants' success based on Yelp reviews' sentiment?

Topic motivation

This project aims to investigate how the sentiment of Yelp's reviews about vegetarian restaurants can predict their future success.

- Climate change is becoming a pressing issue with more people being concerned about the welfare of the planet and how the meat industry harms it. Studies(Fresán & Sabaté, 2019) reveal that vegetarian diet reduces greenhouse gas emissions (around -35% for ovo-lacto vegetarians and up to -50% for vegans), land use (-40/50%), and often water consumption compared with omnivorous diets.
- Additionally, studies (Marsh et al., 2011) reveal that vegetarian diet has a positive impact on people's health in terms of lower body weight, better cholesterol levels, reduced heart disease risk and other chronic diseases (e.g diabetes, hypertension, gallstones, kidney stones).
- Therefore people are starting to open up their mind about eating vegetarian options. Not only vegetarian people eat at vegetarian restaurants now, so it is more important than ever to understand what qualities drive success in this sector.
- A survey from Ipsos (2022) has shown that over half of the British population between 16 and 75 years of age have started to use plant-based alternatives in their diets. These findings suggest that there is an increase in demand for vegetarian food.

Setup

Our Setup block is currently structured like this: 1. Set the working directory to the current file location, to make sure the relative paths will work. 2. Load required packages. 3. Check if TinyTex is installed for **pdf** conversion later 4. Check if /raw_data is in the working directory, and create if needed 5. Check if the raw data is in the /raw_data folder and only download if needed

```
#1. set working directory actively
invisible(rstudioapi::documentSave())
invisible(setwd(dirname(rstudioapi::getActiveDocumentContext()$path)))

#2. ckeck if all dependencies are installed, and install if needed
required_packages <- c("tidyverse", "googledrive")

for (pkg in required_packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    suppressMessages(install.packages(pkg))
  }
}
    #load all the dependencies
    invisible(lapply(required_packages, function(pkg) {
  suppressPackageStartupMessages(library(pkg, character.only = TRUE))
}))


#3. check if TinyTex is installed
if (!tinytex::is_tinytex()) {
  message("TinyTeX not installed. Please run tinytex::install_tinytex() in the console.")}

#4. check if raw_data directory is present in the working directory and create if needed
if (!dir.exists("raw_data")) dir.create("raw_data", recursive = TRUE)

#5. check if the raw data is in the /raw_data folder and only download if needed
#raw datasets
datasets <- c(
  business = "business",
  checkin  = "checkin",
  tip      = "tip",
  user     = "user",
  review   = "review"
)

googledrive::drive_deauth() #get into drive as anonymous user
folder_id <- "1WHSh8ZQYzQ3IQI8tJX90cYGR4bDy13v3" #folder id of .csv yelp files
drive_folder <- drive_ls(as_id(folder_id))      #google drive disk image

    #loop to check if datasets are present in /raw_data folder
for (dataset in datasets) {
  rds_path <- file.path("raw_data", paste0(dataset, ".rds"))
  if (file.exists(rds_path)) {
    message("Found: ", rds_path)
    assign(dataset, readRDS(rds_path)) }
  else {
    message("missing: ",rds_path, ". Checking if .csv file is present." )
    csv_path <- file.path("raw_data", paste0("yelp_academic_dataset_",dataset, ".csv"))
    if(file.exists(csv_path)){
       message("Found CSV: ", csv_path, " → reading and saving as .RDS")
      dat <- readr::read_csv(csv_path, show_col_types = FALSE)
      saveRDS(dat, rds_path)
      assign(dataset, dat)
      rm(dat)
      message("Saved as .RDS for speed at: ", rds_path )  }
    else {
      csv_path <- file.path("raw_data", paste0("yelp_academic_dataset_",dataset, ".csv"))
      message("both .CSV and .RDS are missing. Downloading ",dataset," from Google Drive.")
      file <- drive_folder[str_detect(drive_folder$name, dataset), ]
      size_bytes <- as.numeric(file$drive_resource[[1]]$size)
      size_mb <- round(size_bytes / (1024^2), 2)
      message("Download size: ", size_mb, " MB")
      googledrive::drive_download(as_id(file$id), path = csv_path, overwrite = TRUE)
      dat <- readr::read_csv(csv_path, show_col_types = FALSE)
      saveRDS(dat, rds_path)
      assign(dataset, dat)
      rm(dat, file)
    }
  }
  }
```

```
## Found: raw_data/business.rds
```

```
## Found: raw_data/checkin.rds
```

```
## Found: raw_data/tip.rds
```

```
## Found: raw_data/user.rds
```

```
## Found: raw_data/review.rds
```

```
print(Sys.time())
```

```
## [1] "2025-09-04 21:33:19 CEST"
```

Data exploration

The review dataset from yelp includes the following variables: **funny, cool, review_id, text, stars, useful, business_id, date, user_id**

The `$text` variable holds the actual review. The `$review_id`, `$business_id` and `$user_id` variables can be used to merge the dataset with the datasets about the corresponding businesses and users to obtain additional information about the review. The `$funny`, `$cool` and `$useful` variables indicate the amount of votes the review got from other users. The `$stars` variable indicates the star rating (1-5) assigned by the reviewer. The `$date` variable holds the timestamp of the review.

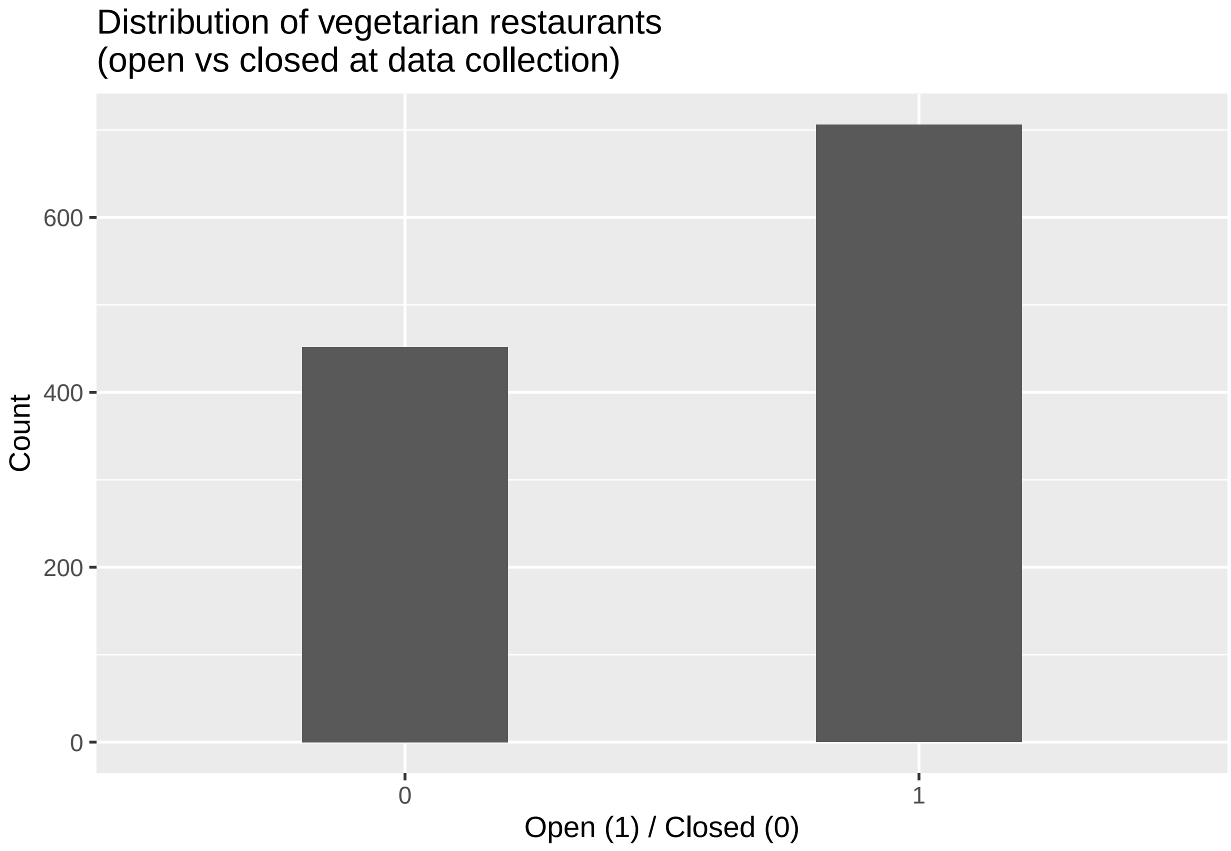
The total # of reviews in the entire Yelp-dataset is 6990280.

We are only interested in reviews on vegetarian restaurants.

```r
business_vr <- business%>%
  filter (str_detect(categories, "Restaurants")) %>%
  filter (str_detect(categories, "Vegetarian"));rm(business)
```

The `$is_open` variable in the business dataframe shows whether the business is still running. Lets check the distribution of vegetarian restaurants that were still running at the time of Yelp's data collection:

```r
business_vr %>% ggplot (aes(x = factor(is_open, levels = c(0,1)))) +
  geom_bar(width = 0.4) +
  labs(title = "Distribution of vegetarian restaurants\n(open vs closed at data collection)",
       x = "Open (1) / Closed (0)",
       y = "Count")
```



We can see that the majority of vegetarian restaurants (61%) is still active.

Let's keep only reviews that are from Vegetarian Restaurants:

```
review_sub <- review %>%
  semi_join(business_vr %>% select(business_id), by = "business_id")
```

The total # of reviews for Vegetarian Restaurants is 184004. We are only interested in recent reviews. Therefore we will only use reviews that were posted after 2018-01-01 00:00:00 UTC. The number of reviews then shrinks to 71948. This is still way too much data to analyze. Therefore, we will only focus on Vegetarian Restaurants that have at least **200 reviews**.

```
rm(user);
review_sub_2018 <- review_sub %>% filter (date >= "2018-01-01 00:00:00 UTC")
review_sub_2018 <- merge(review_sub_2018, business_vr,
                         by.x = "business_id",
                         by.y = "business_id",
                         all.x = TRUE)

reviews_sub_2018_over200 <- review_sub_2018 %>%
  group_by(name) %>%
  mutate(review_n = n()) %>%
  arrange(desc(review_n)) %>% filter(review_n >= 200)
```

The following table shows the 10 vegetarian restaurants that have the most reviews between the period *2017-12-31 23:25:33* and *2017-12-31 23:25:33*.

```
  reviews_sub_2018_over200 %>% select (name, review_n) %>% distinct() %>% head(.,10)%>%print()
```

```
## # A tibble: 10 x 2
## # Groups:   name [10]
##    name                review_n
##    <chr>                  <int>
##  1 Sabrina's Café          1699
##  2 Yard House              1502
##  3 GW Fins                 1029
##  4 Zahav                   1025
##  5 True Food Kitchen        859
##  6 Seasons 52               816
##  7 honeygrow                786
##  8 Fresh Kitchen            680
##  9 Jacques-Imo's Cafe       651
## 10 Tumerico                 637
```

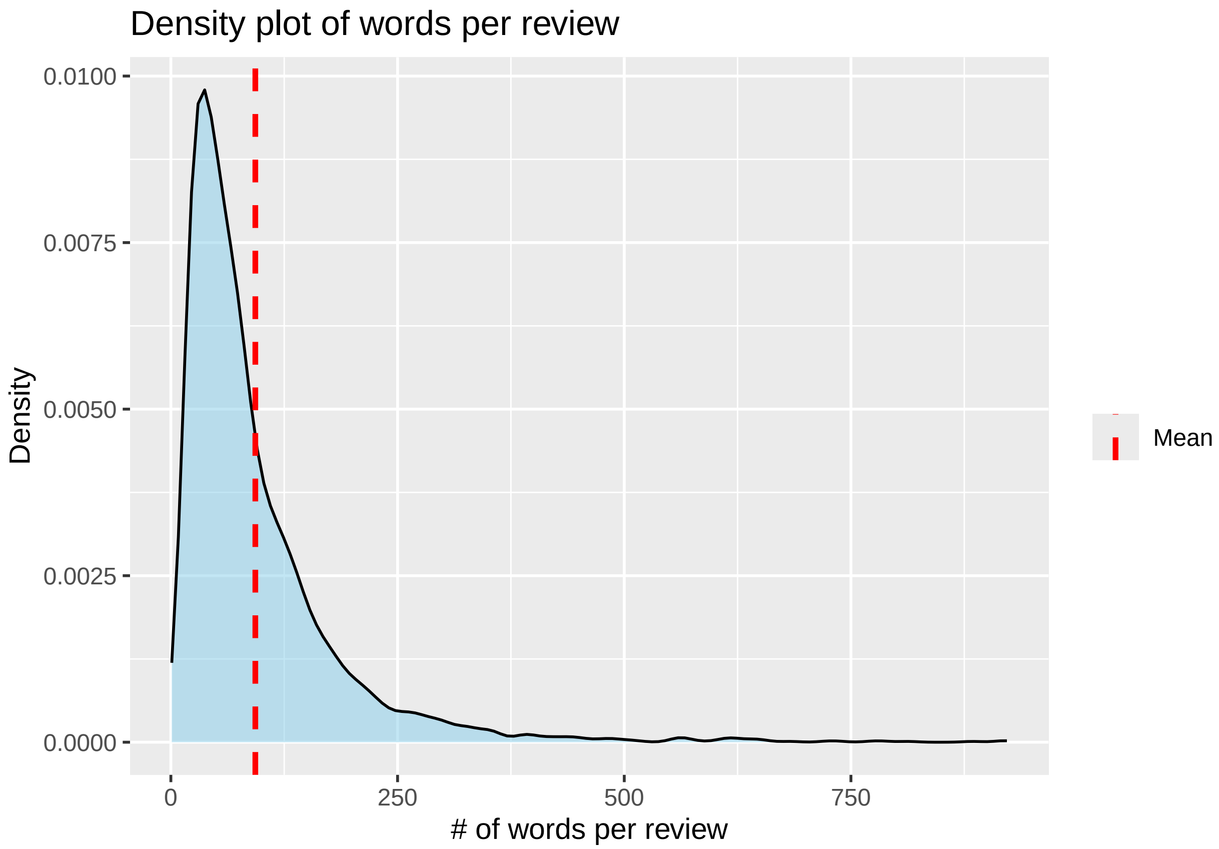
It is also interesting to compute the average length of these reviews. To do that, we will just count the number of *spaces* and add 1 as a proxy variable.

```
reviews_sub_2018_over200 <- reviews_sub_2018_over200 %>% mutate (word_count = str_count(text, " ") + 1)
```

The average number of words in these reviews is **93.1375306**. The density plot looks as following:

```
mean_wc <- mean(reviews_sub_2018_over200%>%pull(word_count))
reviews_sub_2018_over200 %>%
```

```r
dplyr::slice_sample(prop = 0.1) %>%
ggplot(aes(x = word_count)) +
geom_density(fill = "skyblue", alpha = 0.5, n = 128) +
geom_vline(
  aes(xintercept = mean_wc, color = "Mean"),   # mapped -> legend
  linetype = "dashed", linewidth = 1
) +
scale_color_manual(values = c("Mean" = "red")) +
labs(
  title = "Density plot of words per review",
  x = "# of words per review",
  y = "Density",
  color = ""   # legend title
)
```



In the plot, we can see that the # of words is skewed towards 0, likely due to some outliers.

REFERENCES

Chiarelli, N. (2022, March 29). Almost half of UK adults set to cut intake of animal products. Ipsos. Retrieved September 3, 2025, from https://www.ipsos.com/en-uk/almost-half-uk-adults-set-cut-intake-animal-products

Fresán, U., & Sabaté, J. (2019). Vegetarian Diets: Planetary Health and Its Alignment with Human Health. Advances in Nutrition, 10, S380–S388. https://doi.org/10.1093/advances/nmz019

Marsh, K., Zeuschner, C., & Saunders, A. (2011). Health implications of a vegetarian diet. American Journal of Lifestyle Medicine, 6(3), 250–267. https://doi.org/10.1177/1559827611425762