Analysis of DoorDash Restaurant Ratings in Canada

Bryan Seo, Junbo Rao, Wendy Phung, Jag Tang

17th November 2024

Cover Page

Title: Analysis of DoorDash Restaurant Ratings in Canada
Contributors:
Wendy Phung – 76185974 (Leader)
Junbo (David) Rao - 55832919
Jag Tang - 89478895
Bryan Seo - 13749536
Date: 17th November 2024

Distribution of Work

Task	Contributor
Choose sample size for SRS and Stratified Sampling	Bryan
Data Cleaning and Exploration	Junbo/Wendy
Project Overview and Cover Page	Jag
R Code Implementation for SRS and Stratified Sampling	Wendy
Analyze Results for SRS and Stratified Sampling	Bryan/Jag
Proportion Code and Analysis	Jag
Appendix and References	Group

Project Overview

Food delivery apps such as UberEats and DoorDash are used extensively throughout Canada, mostly by students and working adults. Food delivery is a great convenience, it saves people time and effort either from cooking or going to get takeout but it comes with drawbacks. Firstly, it is difficult to accurately judge the quality of a restaurant on the app with just photos, so many people usually can only refer to the rating of the restaurant which ranges from 0 to 5.0 stars. Secondly, food delivery is generally quite expensive, and the quality is diminished when compared to eating it at restaurants. The population of interest for this study is the restaurants across Canada

This project aims to determine the quality of food from DoorDash restaurants across Canada. In order to do this, we will perform a simple random sample (SRS) and a stratified sample using proportional allocation

on a dataset of DoorDash restaurants in Canada to estimate the mean rating of restaurants in Canada on DoorDash, and for our secondary parameter, we have defined it as the proportion of restaurants with a rating greater than 4.5. We hope to accurately estimate the mean rating of restaurants on DoorDash in Canada and the proportion of restaurants with a rating above 4.5 to use as a proxy for the quality of food.

Population Overview

- **Target Population:** All DoorDash-listed restaurants across Canada. N = 2160
- **Stratification:** The population is stratified by cities based on third-party Google Maps data.
- **Source of Data:** Kaggle

```
# Data cleaning and data preparation
# Ensure dataset is in current Working Directory using "getwd()"
full_data <- read.csv("cleaned_full_data.csv", header = T)
data <- full_data[,c("X", "restaurant", "star", "num_reviews", "city")] %>%
  rename(id = X)%>%
  drop_na()

dim(data)
```

```
## [1] 2620    5
```

```
summary(data)
```

```
##        id          restaurant             star        num_reviews
##  Min.   :   0.0   Length:2620        Min.   :2.800   Min.   :   10.0
##  1st Qu.: 773.8   Class :character   1st Qu.:4.400   1st Qu.:   57.0
##  Median :1630.5   Mode  :character   Median :4.500   Median :  182.0
##  Mean   :1610.8                      Mean   :4.479   Mean   :  570.4
##  3rd Qu.:2506.2                      3rd Qu.:4.700   3rd Qu.:  560.0
##  Max.   :3287.0                      Max.   :5.000   Max.   :21000.0
##      city
##  Length:2620
##  Class :character
##  Mode  :character
##
##
##
```

For additional analysis of population, please see appendix A.

Parameter of Interest Overview

- **Continuous Parameter:** Average restaurant ratings across major cities in Canada
- **Binary Parameter:** Proportion of restaurants with ratings higher than 4.5

Method Overview

- **Method 1:** Simple Random Sampling
- **Method 2:** Stratified Sampling

Choosing Sample Size

First, we will use the binary form (ie. higher than 4.5, lower than 4.5) of our continuous dataset to estimate our sample size. This is because the binary form allows us to attain a conservative estimate (ie. assume maximum variance of 0.25). Intuitively, we know that binary data typically requires larger samples than continuous data to achieve the same precision as it loses information through discretization. Therefore, the sample size calculated from the binary dataset will be sufficient for the continuous case as well.

In order to determine the appropriate sample size, we chose a 95% CI half-width length of 0.05, and assumed the worst-case variance for the binary data (0.25). We also assumed that the sampling distribution will approximate to a normal distribution. Given the guessed variance, intended % for the CI, and the half-width, we were able to calculate the minimum sample size required n=326

Sample Size Calculation for Binary/Continuous Case with FPC

```
N <- 2620 # Total population
z <- 1.96 # z-score for 95% CI
p <- 0.5 # Worst-case variance
e <- 0.05 # Margin of error

n <- round((z^2 * p * (1 - p)) / e^2)
fpc <- n/(1+n/N)
sample_size <- round(fpc)
sample_size
```

```
## [1] 335
```

Sample Size Calculation for Stratified Case

Relying on third-party information based on Google Maps data, we will stratify the data into cities, and perform a proportional allocation based on the number of restaurants in each city.

```
# Stratified Sampling Proportional Allocation
# Calculate n.h
N.h <- tapply(data$star, data$city, length)
cat("Population Strata Sizes:", N.h, "\n")
```

```
## Population Strata Sizes: 100 321 210 432 172 622 591 172
```

```
city_allocation <- round((N.h/N)*sample_size)
city_allocation
```

```
##  Branpton   Calgary  Edmonton  Montreal    Ottawa   Toronto Vancouver  Winnipeg
##        13        41        27        55        22        80        76        22
```

Method 1: Simple Random Sampling

```r
# Population mean of star
p_mean_star <- mean(data$star)
cat("Population Mean:", p_mean_star, "\n")
```

```
## Population Mean: 4.479237
```

SRS for Mean

```r
IDS <- data$id
N <- length(data$id)
n <- sample_size

# Simple Random Sampling
srs_sample <- sample.int(N, n, replace = FALSE)
srs_sample.IDS <- IDS[srs_sample]
data_srs_sample <- subset(data, id %in% srs_sample.IDS)

# Sample Mean
srs_sample.mean <- mean(data_srs_sample$star)
cat("Sample Mean:", srs_sample.mean, "\n")
```

```
## Sample Mean: 4.498209
```

```r
# Sample Variance for Sample Mean
srs_sample.variance <- sum((data_srs_sample$star - srs_sample.mean)^2) /
  (length(data_srs_sample$star) - 1)
cat("Sample Variance:", srs_sample.variance, "\n")
```

```
## Sample Variance: 0.06340996
```

```r
# Sample SD for Sample Mean
srs_sample.sd <- sqrt(srs_sample.variance)
cat("Sample Standard Deviation:", srs_sample.sd, "\n")
```

```
## Sample Standard Deviation: 0.2518133
```

```r
# Calculate 95% CI for population mean
srs_sample.se <- sqrt((1 - n / N) / n) * srs_sample.sd
cat("Sample SE:", srs_sample.se, "\n")
```

```
## Sample SE: 0.01284839
```

```r
srs_CI <- c(srs_sample.mean - 1.96 * srs_sample.se, srs_sample.mean + 1.96 * srs_sample.se)
cat("95% Confidence Interval:", srs_CI, "\n")
```

```
## 95% Confidence Interval: 4.473026 4.523392
```

SRS for Proportion

```r
# Sample proportion of restaurants with ratings greater than 4.5
srs_sample_p <- mean(data_srs_sample$star > 4.5)
cat("Sample Proportion of Restaurants with Ratings > 4.5:", srs_sample_p, "\n")
```

```
## Sample Proportion of Restaurants with Ratings > 4.5: 0.480597
```

```r
# Sample variance of the sample proportion
srs_sample.variance_p <- srs_sample_p * (1 - srs_sample_p) / nrow(data_srs_sample)
cat("Sample Variance of the Proportion:", srs_sample.variance_p, "\n")
```

```
## Sample Variance of the Proportion: 0.0007451448
```

```r
# Sample SD of the sample proportion
srs_sample.sd_p <- sqrt(srs_sample.variance_p)
cat("Sample Standard Deviation of the Proportion:", srs_sample.sd_p, "\n")
```

```
## Sample Standard Deviation of the Proportion: 0.02729734
```

```r
# 95% Confidence Interval for the sample proportion
srs_sample.se_p <- sqrt(1 - (n / N)) * srs_sample.sd_p
srs_CI_p <- c(srs_sample_p - 1.96 * srs_sample.se_p, srs_sample_p + 1.96 * srs_sample.se_p)
cat("Sample Standard Error of the Proportion:", srs_sample.se_p, "\n")
```

```
## Sample Standard Error of the Proportion: 0.02549252
```

```r
cat("95% Confidence Interval for the Proportion:", srs_CI_p, "\n")
```

```
## 95% Confidence Interval for the Proportion: 0.4306317 0.5305624
```

SRS Analysis

SRS Mean Analysis

Using SRS, we estimated the average restaurant rating on DoorDash in Canada. The sample mean rating was 4.47, which is close to the population mean of 4.479. Notably, the standard error of this approach for mean is 0.0148. This suggests that the SRS method provides a reliable estimate of the true average restaurant rating.

SRS Proportion Analysis

For the proportion of restaurants with ratings greater than 4.5, the SRS method yielded a sample proportion of 48.36%, and sample standard error is 0.0255. The 95% confidence interval for this proportion is between 43.36% and 53.36%, indicating that nearly half of the restaurants have ratings above 4.5. This estimate provides a reasonable understanding of the proportion of highly rated restaurants on DoorDash in Canada.

Method 2: Stratified Sampling

Stratified for Mean

```
n.h <- city_allocation

cities <- names(N.h)

set.seed(0)
data_str_sample <- NULL
for (i in 1: length(cities))
{
  row.indices <- which(data$city == cities[i])
  sample.indices <- sample(row.indices, n.h[i], replace = F)
  data_str_sample <- rbind(data_str_sample, data[sample.indices, ])
}

# Stratified Estimation

# Mean rating per city
str_mean.h <- tapply(data_str_sample$star, data_str_sample$city, mean)
cat("Stratified Mean Ratings by City:\n")
```

```
## Stratified Mean Ratings by City:
```

```
print(str_mean.h)
```

```
##  Branpton   Calgary  Edmonton  Montreal    Ottawa   Toronto Vancouver  Winnipeg
##  4.123077  4.504878  4.437037  4.394545  4.277273  4.540000  4.534211  4.386364
```

```
# Variance of ratings per city
str_variance.h <- tapply(data_str_sample$star, data_str_sample$city, var)
cat("Stratified Variance of Ratings by City:\n")
```

```
## Stratified Variance of Ratings by City:
```

```
print(str_variance.h)
```

```
##   Branpton    Calgary   Edmonton   Montreal     Ottawa    Toronto  Vancouver
## 0.12192308 0.06247561 0.08626781 0.09200673 0.17612554 0.06167089 0.05108070
##   Winnipeg
## 0.19647186
```

```
# Standard error of ratings per city
str_se.h <- sqrt((1 - n.h / N.h) * str_variance.h / n.h)
cat("Stratified Standard Errors by City:\n")
```

```
## Stratified Standard Errors by City:
```

```r
print(str_se.h)
```

```
##   Branpton    Calgary   Edmonton   Montreal     Ottawa    Toronto  Vancouver
## 0.09032977 0.03645775 0.05276651 0.03820826 0.08355669 0.02591788 0.02420090
##   Winnipeg
## 0.08825112
```

```r
# Weighted mean for stratified estimate of the mean
str_prop.mean <- sum(N.h / N * str_mean.h)
cat("Stratified Estimate of the Mean Rating (Overall):", str_prop.mean, "\n")
```

```
## Stratified Estimate of the Mean Rating (Overall): 4.458908
```

```r
# Standard error for the stratified mean estimate
str_prop.mean_se <- sqrt(sum((N.h / N)^2 * str_se.h^2))
cat("Stratified Estimate of Standard Error (Overall):", str_prop.mean_se, "\n")
```

```
## Stratified Estimate of Standard Error (Overall): 0.01485707
```

```r
# Calculate 95% CI for population mean
str_prop_mean_CI <- c(str_prop.mean - 1.96 * str_prop.mean_se,
                      str_prop.mean + 1.96 * str_prop.mean_se)
cat("95% Confidence Interval:", str_prop_mean_CI, "\n")
```

```
## 95% Confidence Interval: 4.429788 4.488028
```

Stratified for Proportion

```r
# Proportion:
# Adding a binary indicator for ratings > 4.5 stars
data <- data %>%
  mutate(rating_above_4_5 = ifelse(star > 4.5, 1, 0))

# Stratified Sampling Proportional Allocation
N.h <- tapply(data$rating_above_4_5, data$city, length)
city_allocation <- round((N.h / N) * n)

n.h <- c(13, 41, 27, 55, 22, 80, 76, 22) # Adjust based on city_allocation
cities <- names(N.h)

set.seed(0)
data_str_sample <- NULL
for (i in 1:length(cities)) {
  row.indices <- which(data$city == cities[i])
  sample.indices <- sample(row.indices, n.h[i], replace = FALSE)
  data_str_sample <- rbind(data_str_sample, data[sample.indices, ])
}

# Stratified Estimation for Proportion
```

```r
str_prop.h <- tapply(data_str_sample$rating_above_4_5, data_str_sample$city, mean)
str_variance.h <- tapply(data_str_sample$rating_above_4_5, data_str_sample$city, var)
str_se.h <- sqrt((1 - n.h / N.h) * str_variance.h / n.h)

str_prop.p <- sum(N.h / N * str_prop.h)
cat("Stratified Proportion:", str_prop.p, "\n")
```

```
## Stratified Proportion: 0.4493515
```

```r
str_prop.p_se <- sqrt(sum((N.h / N)^2 * str_se.h^2))
cat("Stratified Proportion SE:", str_prop.p_se, "\n")
```

```
## Stratified Proportion SE: 0.02471124
```

```r
# Calculate 95% CI for population mean
str_prop_p_CI <- c(str_prop.p - 1.96 * str_prop.p_se, str_prop.p + 1.96 * str_prop.p_se)
cat("95% Confidence Interval:", str_prop_p_CI, "\n")
```

```
## 95% Confidence Interval: 0.4009175 0.4977856
```

Stratified Sampling Analysis

Stratified Mean Analysis

Using stratified sampling with proportional allocation based on cities, we estimated the mean restaurant rating to be 4.459. The standard error of this estimate was 0.0149. This standard error is comparable to that obtained from the SRS method, indicating similar precision. The stratified mean estimate is slightly lower than the SRS estimate, which may reflect differences in ratings across cities accounted for by the stratification.

Stratified Proportion Analysis

The stratified sampling method estimated the proportion of restaurants with ratings greater than 4.5 to be 0.4494 (or 44.94%). The standard error for this estimate was 0.0247, which is smaller than the standard error from the SRS method at 0.0255. Based on this result, stratified sampling is a better method to estimate the proportions than SRS.

Conclusion and Discussion

Conclusion:

This project utilized two sampling methods—Simple Random Sampling (SRS) and Stratified Sampling—to estimate two key parameters: the mean restaurant rating on DoorDash in Canada and the proportion of restaurants with ratings greater than 4.5

The SRS method estimated the mean rating as 4.47 with a 95% confidence interval of [4.441, 4.499] and the proportion of restaurants with ratings above 4.5 as 48.36% with a confidence interval of [43.36%, 53.36%]. The estimated standard errors for the mean and proportion were 0.0148 and 0.0255, respectively.

The Stratified Sampling method, using cities as strata, provided a slightly lower estimate for the mean rating at 4.458, with a standard error of 0.0149 and 95% confidence interval of [4.430, 4.488]. The proportion of restaurants rated above 4.5 was estimated to be 44.94%, with a standard error of 0.0247. and 95% confidence interval of [40.09%, 49.78%]. The confidence intervals obtained through this method were narrower compared to SRS, demonstrating the improved precision from stratification.

Overall, both methods produced similar estimates for the mean and proportion, validating the robustness of the results. However, stratified sampling was marginally more efficient, particularly for estimating the proportion, as it accounted for differences across cities.

Discussion:

The results highlight the value of stratified sampling when the population exhibits heterogeneity, as it provided improved precision over SRS by leveraging the differences in restaurant ratings across cities.

The mean ratings across Canadian restaurants on DoorDash indicate generally high satisfaction, with nearly half of the restaurants achieving a rating above 4.5. This suggests that users perceive DoorDash restaurants as offering high-quality service. However, it is also important to consider factors such as rating inflation, which may skew user perceptions.

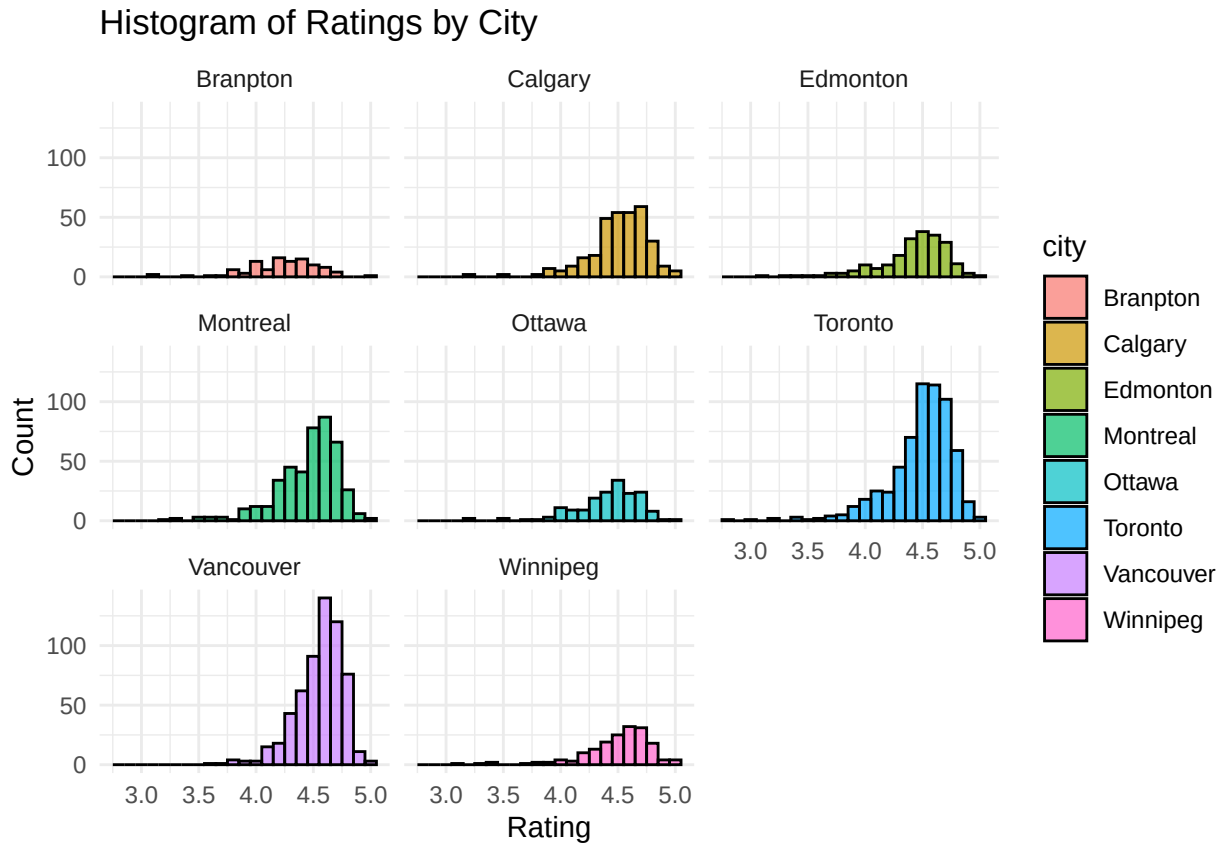
Limitations:

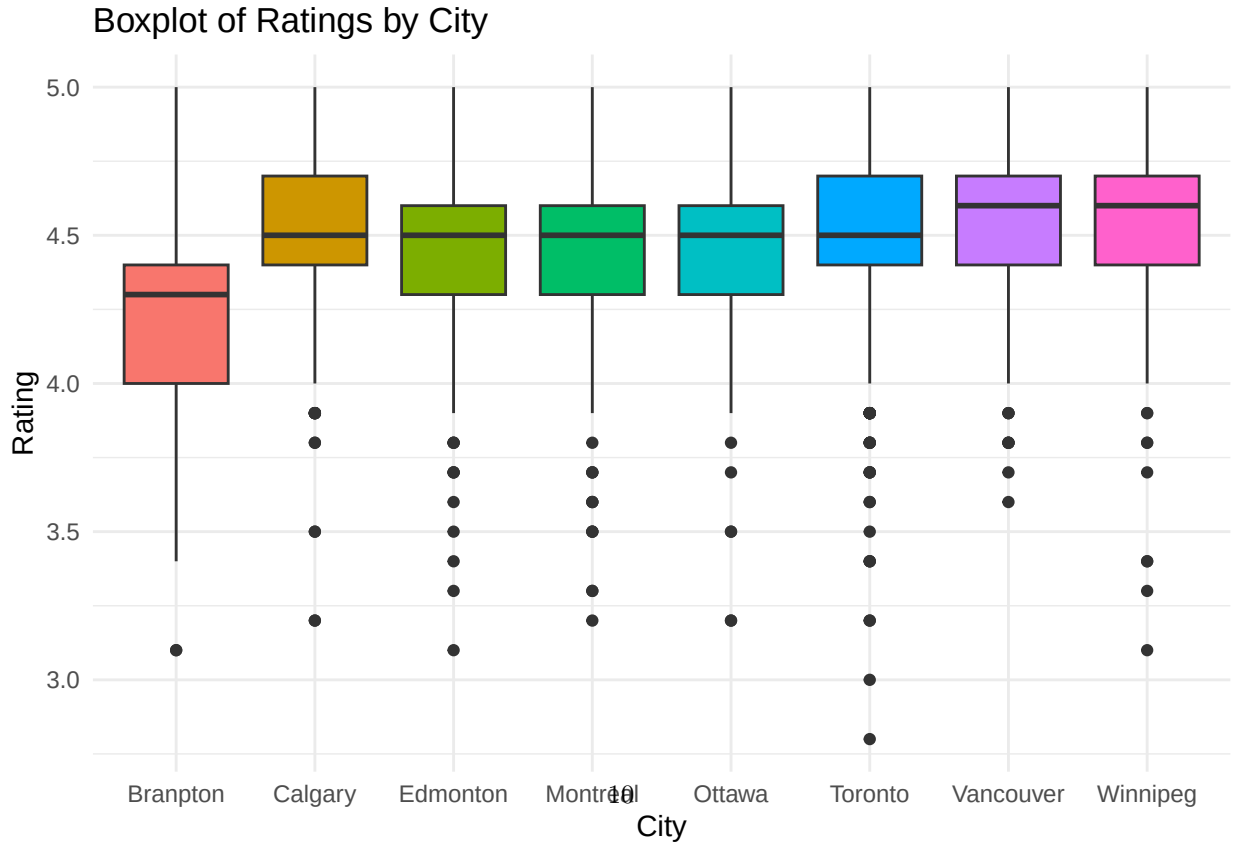
1. **Representation of the Population**: The dataset only includes DoorDash-listed restaurants, which may not fully represent all restaurants in Canada. This limitation could introduce bias, as restaurants available on other food delivery platforms or not listed on DoorDash might differ in quality and customer ratings.
2. **Rating Influences**: The study relies solely on customer ratings as a proxy for food quality. Ratings can be influenced by non-food-related factors such as delivery time, app usability, or customer service, potentially skewing the results.
3. **Data Assumptions**: The analysis assumes that ratings are an unbiased and normally distributed measure across the population. If the distribution of ratings is skewed, particularly in cities with a smaller number of restaurants, this could affect the accuracy of the estimates.

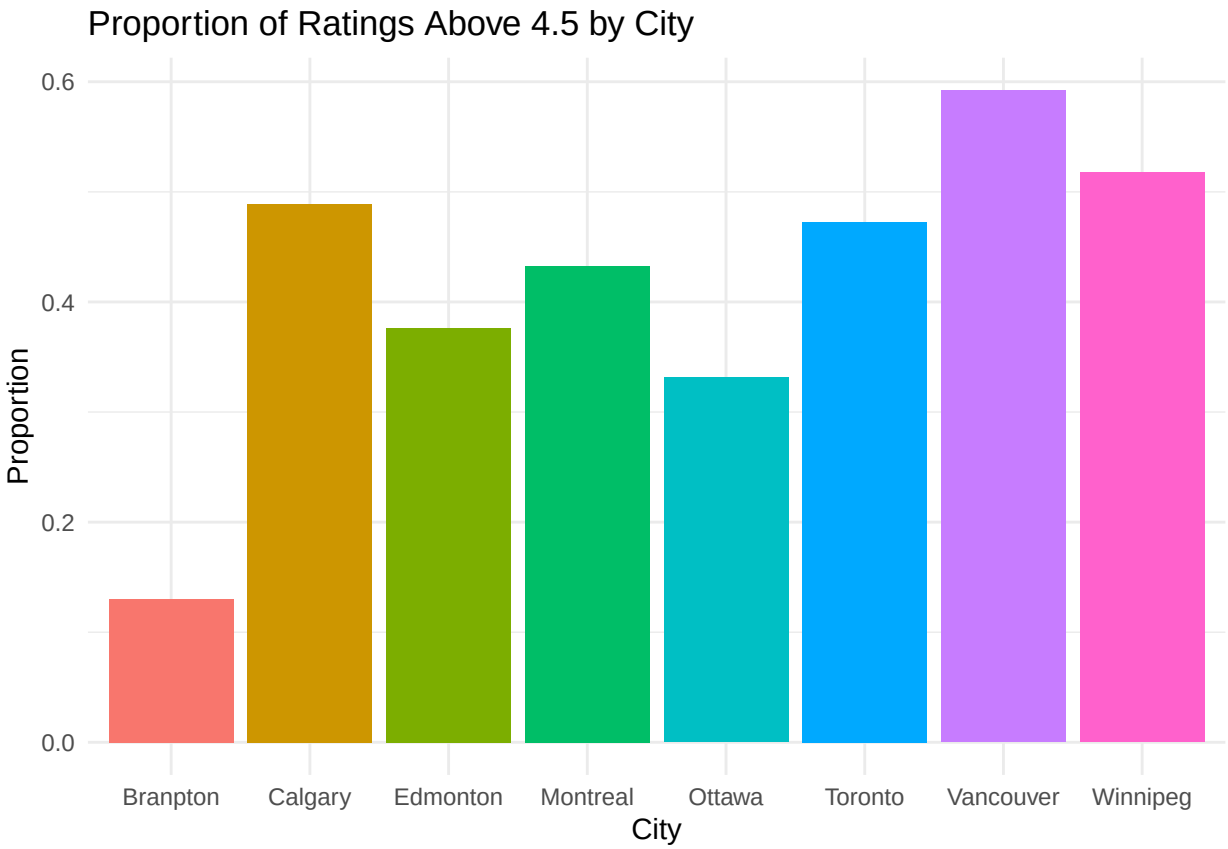
These limitations should be considered when interpreting the findings and could inform improvements in future studies.

Appendix

A


Histogram of Ratings by City
Branpton
Calgary
Edmonton
Montreal
Ottawa
Toronto
Vancouver
Winnipeg
Count
Rating
city
Branpton
Calgary
Edmonton
Montreal
Ottawa
Toronto
Vancouver
Winnipeg


Boxplot of Ratings by City
Rating
City
Branpton
Calgary
Edmonton
Montreal
Ottawa
Toronto
Vancouver
Winnipeg



References

1. DoorDash dataset from Kaggle
2. Google Maps city data: gosnappy.io