Homework_2_wi19b004

Michael Scheidl

19 9 2021

Contents

1 Assignment

The `Violations` data set in the `mdsr` package contains information regarding the outcome of health inspections of restaurants in New York City. Use these data to calculate the median violation score by zip code for zip codes in Manhattan with 50 or more inspections. What pattern do you see between the number of inspections and the median score.

2 Solution

2.1 Load Libraries

```
library(tidyverse)
library(mdsr)
```

This code block loads the library `tidyverse` for using `dplyr`, which contains tibbles and functions for transformations of dataframes. It also loads the `mdsr` package for the `Violations` data used in this exercise.

2.2 Create Tibble

```
data <- as_tibble(Violations)
```

Load the `Violations` data and save it in the variable `data` as tibble.

2.3 Filter and Group desired Data

This codeblock filters the `Violations`data, to continue only with valid inspection data (inspection data after 1900.01.01) from Manhattan. After that the result is grouped by zipcode, camis (unique restaurant id), inspection date, inspection type and the score with summarise. This is done, to get only one entry per inspection, not per violation, because the violation score of the inspection is recorded with each violation. After that the inspections are grouped per zip code, the number of inspections and the meadian violation score per zip code is calculated. The result is filtered to only contain zip codes with 50 or more inspections. Information for the data set are from: https://data.cityofnewyork.us/api/views/43nn-pn8j/files/3016a624-55c0-4bd0-bfb4-95c6b9ea6ba4?download=true&filename=About_NYC_Restaurant_Inspection_Data_on_NYC_OpenData_092418.docx

```
filtered_data <- data %>%
  filter(boro == "MANHATTAN" & inspection_date > as.Date("1900-01-01")) %>% # Filter data for MANHATTAN
  group_by(zipcode, camis,inspection_date, inspection_type, score) %>% # Group data to get unique inspe
  summarise() %>% # Summarize data to get inspections
  group_by(zipcode) %>% # Only Group by zipcode
  summarise(number_of_inspections = n(), med_score = median(score, na.rm=TRUE)) %>% # Get number of ins
  filter(number_of_inspections >=50) # Only show zipcodes with more than 50 inspections
filtered_data
```

```
## # A tibble: 47 x 3
##    zipcode number_of_inspections med_score
##      <int>                 <int>     <dbl>
##  1   10001                  3318        12
##  2   10002                  3383        12
##  3   10003                  5076        12
##  4   10004                   927        12
##  5   10005                   463        12
##  6   10006                   359        12
##  7   10007                   937        12
##  8   10009                  2364        12
##  9   10010                  1763        12
## 10   10011                  3371        12
## # ... with 37 more rows
```

We can observe, that the median score is consistent at 12 after enough inspections are done, with only a few deviation zip codes. It is mostly lower when only 300 or less inspections are done.

```
print(filtered_data %>% filter(number_of_inspections < 500), n = 12)
```

```
## # A tibble: 12 x 3
##    zipcode number_of_inspections med_score
##      <int>                 <int>     <dbl>
##  1   10005                   463      12
##  2   10006                   359      12
##  3   10020                   289      10
##  4   10030                   254      11
##  5   10037                   231      13
##  6   10039                   207      12
##  7   10112                    93       9
##  8   10119                   108      12
##  9   10121                    86       7
## 10   10280                    67      12
## 11   10281                    99       9
## 12   10282                    93       9.5
```