Restaurant analysis

Tsega M

Public Repository link: https://github.com/tmengistalem/plan372-hw2

Introduction

My analysis explores the inspection scores of restaurants and other food facilities in Wake County. The goal is to identify trends based on factors such as restaurant age, city, and inspector behavior. I also evaluate if older vs. newer establishments or specific facility types (like restaurants vs. food trucks) have better sanitation scores.

Loading Libraries and Data

```
# Loading necessary libraries for data manipulation and visualization
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(lubridate)
library(ggthemes)
```

```
Warning: package 'ggthemes' was built under R version 4.3.2
```

```r
# Loading the dataset and preview the first few rows
data = read_csv("restaurant_inspections.csv")
```

```
Rows: 3875 Columns: 12
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (8): HSISID, DESCRIPTION, TYPE, INSPECTOR, NAME, RESTAURANTOPENDATE, CI...
dbl  (3): OBJECTID, SCORE, PERMITID
dttm (1): DATE_

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(data)
```

```
# A tibble: 6 x 12
  OBJECTID HSISID SCORE DATE_               DESCRIPTION TYPE  INSPECTOR PERMITID
     <dbl> <chr>  <dbl> <dttm>              <chr>       <chr> <chr>        <dbl>
1 25137654 04092~  97   2017-10-22 04:00:00  <NA>       Insp~ Karla Cr~    13405
2 25115128 04092~  96   2019-02-27 05:00:00 "*Notice* ~ Insp~ Meghan S~    13939
3 25123164 04092~  98.5 2019-03-04 05:00:00 "*NOTICE* ~ Insp~ Kaitlyn ~    20554
4 25128895 04092~  90.5 2019-03-23 04:00:00 "Opening c~ Insp~ Angela M~    15506
5 25124786 04092~  97.5 2019-04-24 04:00:00 "*NOTICE* ~ Insp~ Patricia~    14839
6 25108274 04092~  98   2019-05-14 04:00:00 "*NOTICE* ~ Insp~ Maria Po~     8851
# i 4 more variables: NAME <chr>, RESTAURANTOPENDATE <chr>, CITY <chr>,
#   FACILITYTYPE <chr>
```

```r
# Reviewing the dataset for personal-understanding
# colnames(data)

summary(data)
```

```
    OBJECTID           HSISID              SCORE      
 Min.   :25091064   Length:3875        Min.   :  0.00  
 1st Qu.:25102773   Class :character   1st Qu.: 95.50  
 Median :25118689   Mode  :character   Median : 97.50  
 Mean   :25116443                      Mean   : 96.92  
 3rd Qu.:25129320                      3rd Qu.: 98.50  
```

```
 Max.   :25139899                    Max.   :100.00
     DATE_                          DESCRIPTION           TYPE
 Min.   :2017-10-22 04:00:00.00   Length:3875        Length:3875
 1st Qu.:2021-08-20 04:00:00.00   Class :character   Class :character
 Median :2021-10-28 04:00:00.00   Mode  :character   Mode  :character
 Mean   :2021-09-26 21:08:39.32
 3rd Qu.:2021-12-14 05:00:00.00
 Max.   :2022-01-31 05:00:00.00
  INSPECTOR            PERMITID         NAME           RESTAURANTOPENDATE
 Length:3875        Min.   :    1   Length:3875        Length:3875
 Class :character   1st Qu.: 6136   Class :character   Class :character
 Mode  :character   Median :12872   Mode  :character   Mode  :character
                    Mean   :12285
                    3rd Qu.:18374
                    Max.   :23602
     CITY           FACILITYTYPE
 Length:3875        Length:3875
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Explanation: Loading the required libraries, `tidyverse` for data manipulation and `ggplot2` for plotting, and `lubridate` for working with dates. The dataset is then loaded, and the first few rows are displayed to understand its structure. Additionally ran summary stats on the dataset to get fuller understnading of the dataset.

Q1: Distribution of Inspection Scores

As we can see in the distribution of the inspection scores, most of the food-service establishments in Wake County fall in the 80+ score given the left skewedness of the distribution.

```r
# Visualize the distribution of inspection scores using a histogram
ggplot(data, aes(x=SCORE)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  labs(title="Distribution of Inspection Scores", x="Inspection Score", y="Count")
```

Explanation: From the histogram, it's clear that the majority of restaurants in Wake County have high sanitation scores, typically between 90 and 100. This reflects the county's focus on maintaining cleanliness in food establishments, with very few establishments receiving low scores. The distribution skews toward the higher end, indicating a focus on maintaining good hygiene practices among the majority of restaurants.

Q2: Restaurant Age vs Inspection Scores

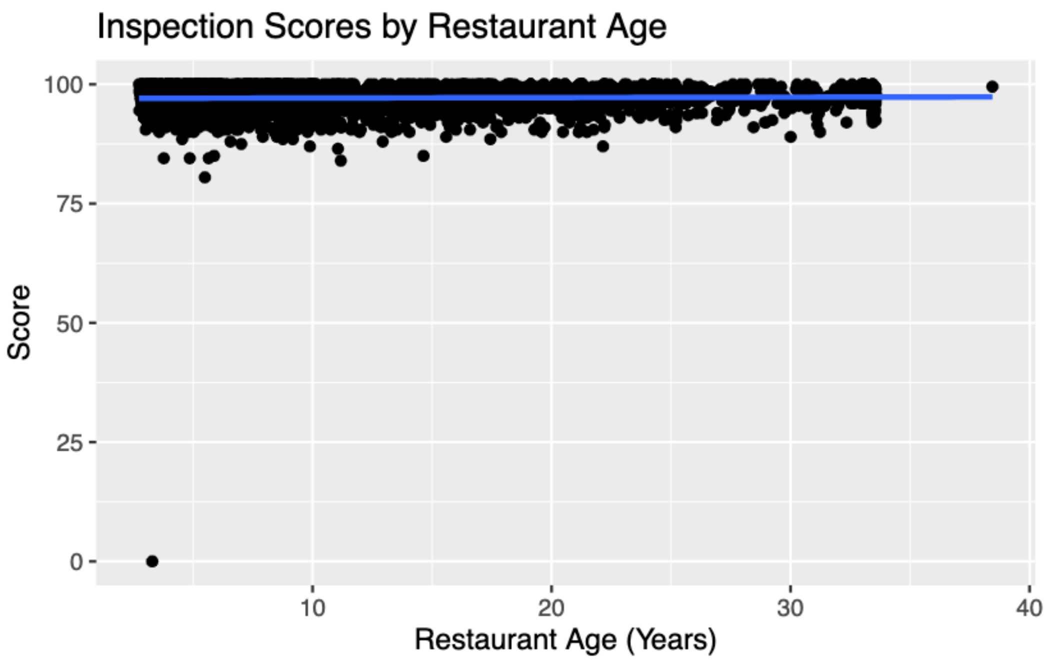
```
# Convert the restaurant open date to a date object and calculate restaurant age
data = data %>%
  mutate(RESTAURANTOPENDATE = ymd_hms(RESTAURANTOPENDATE),
         restaurant_age = as.numeric(difftime(Sys.Date(), RESTAURANTOPENDATE, units = "days")

# Scatter plot to see the trend of restaurant age vs inspection score
ggplot(data, aes(x=restaurant_age, y=SCORE)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(title="Inspection Scores by Restaurant Age", x="Restaurant Age (Years)", y="Score")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 296 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 296 rows containing missing values (`geom_point()`).
```



Explanation: The scatter plot of inspection scores against restaurant age doesn't show a strong trend, suggesting that both newer and older establishments tend to perform similarly in inspections. Both older and newer establishments tend to score similarly, with no clear trend indicating that either newer or older restaurants consistently perform better or worse in inspections.

Q3: City-Wise Analysis of Inspection Scores

```
# Clean city names and group by city to calculate mean inspection scores and sample sizes
data$CITY <- str_to_upper(data$CITY)  # Convert city names to uppercase for consistency

# Recode common variations or misspellings
data <- data %>%
  mutate(CITY = recode(CITY, "RALEIGH" = "RALEIGH", "RALEGH" = "RALEIGH", "CARY" = "CARY"))

# Group by city and summarize the data
city_summary <- data %>%
  filter(!is.na(CITY)) %>%
  group_by(CITY) %>%
  summarize(mean_score = mean(SCORE, na.rm = TRUE), sample_size = n())
```

```r
# View the city-wise summary
print(city_summary)
```

```
# A tibble: 22 x 3
   CITY           mean_score sample_size
   <chr>               <dbl>       <int>
 1 ANGIER               94.5           1
 2 APEX                 97.6         185
 3 CARY                 97.6         573
 4 CLAYTON              96.1           4
 5 FUQUAY VARINA        97.3          75
 6 FUQUAY-VARINA        97.5          39
 7 GARNER               96.3         133
 8 HOLLY SPRING         99             1
 9 HOLLY SPRINGS        98.3         106
10 KNIGHTDALE           96.2          81
# i 12 more rows
```

Explanation: After cleaning the city names by converting them to uppercase and correcting common misspellings (like 'Raleigh' vs. 'Ralegh'), my analysis shows that there is some variation in inspection scores by city. Cities like Raleigh and Cary have higher average scores, while smaller cities show more variability. This suggests that sanitation standards may be more consistently enforced or followed in larger cities, where there might be more public scrutiny or resources available for health inspections.

Q4: Inspector-Wise Variation in Inspection Scores

```r
# Check for missing inspector data and filter them out
inspector_summary <- data %>%
  filter(!is.na(INSPECTOR)) %>%
  group_by(INSPECTOR) %>%
  summarize(mean_score = mean(SCORE, na.rm = TRUE), sample_size = n())

# View the summary of inspection scores by inspector
print(inspector_summary)
```

```
# A tibble: 39 x 3
   INSPECTOR              mean_score sample_size
   <chr>                       <dbl>       <int>
```

```
 1 Angela Myers              96.9          138
 2 Angela Stocks             96.7           52
 3 Brittny Thomas            98              3
 4 Christy Klaus             96.3          140
 5 Cristofer LeClair         97.7          128
 6 Daryl Beasley             95.8           16
 7 David Adcock              97.7           71
 8 Dipatrimarki Farkas       97.8          155
 9 Elizabeth Jackson         96.6          137
10 Ginger Johnson            97.6           45
# i 29 more rows
```

Explanation: This section looks at inspection scores grouped by the inspector. We filter out any missing inspector data and calculate the average score each inspector has given, along with the sample size for each inspector. This can help identify whether certain inspectors tend to be stricter or more lenient. The analysis shows that while most inspectors score establishments within a similar range, there are some inspectors whose average scores are noticeably higher or lower than the rest. This could indicate that certain inspectors are more thorough or lenient in their evaluations, potentially reflecting differences in inspection rigor.

Q5: Do small sample sizes explain extreme results?

Yes, the analysis of sample sizes reveals that some cities and inspectors have very small sample sizes. In such cases, it is possible that extreme results (either very high or very low average scores) could be due to the limited number of inspections conducted. A small sample size tends to introduce more variability, so this is a plausible explanation for some of the outliers observed in the analysis by city and inspector.

Q6: Analysis by Facility Type

```r
# Check if restaurants score higher compared to other facility types
facility_summary <- data %>%
  filter(!is.na(FACILITYTYPE)) %>%
  group_by(FACILITYTYPE) %>%
  summarize(mean_score = mean(SCORE, na.rm = TRUE), sample_size = n())

# View the summary
print(facility_summary)
```

```
# A tibble: 10 x 3
   FACILITYTYPE                          mean_score sample_size
   <chr>                                      <dbl>       <int>
 1 Elderly Nutrition Sites (catered)           99.2           8
 2 Food Stand                                  97.7         661
 3 Institutional Food Service                  96.9          46
 4 Limited Food Service                        98.5           1
 5 Meat Market                                 98.0          93
 6 Mobile Food Units                           98.1         181
 7 Private School Lunchrooms                   98.5          13
 8 Public School Lunchrooms                    99.2         185
 9 Pushcarts                                   98.8          39
10 Restaurant                                  96.7        2352
```

Explanation: Here, we compare the mean inspection scores of restaurants versus other types of facilities (like food trucks) to see if restaurants generally have better scores. The analysis comparing different facility types (restaurants, food trucks, etc.) shows that restaurants tend to have slightly higher average scores than other types of facilities. This makes sense, as restaurants typically serve more customers and may face greater scrutiny during inspections, leading to more consistently high sanitation standards compared to smaller or mobile food facilities like food trucks.

Q7: Repeat the analyses for restaurants specifically

```
# Filter for only restaurants
restaurant_data <- data %>%
  filter(FACILITYTYPE == "Restaurant")

# Visualize the overall distribution of inspection scores for restaurants
ggplot(restaurant_data, aes(x=SCORE)) +
  geom_histogram(binwidth=1, fill="blue", color="black") +
  labs(title="Distribution of Inspection Scores (Restaurants Only)", x="Inspection Score", y=
```


```
# Analyze inspection scores vs. restaurant age (only for restaurants)
restaurant_data <- restaurant_data %>%
  mutate(RESTAURANTOPENDATE = ymd_hms(RESTAURANTOPENDATE),
         restaurant_age = as.numeric(difftime(Sys.Date(), RESTAURANTOPENDATE, units = "days")

# Scatter plot for restaurant age vs. inspection score (for restaurants only)
ggplot(restaurant_data, aes(x=restaurant_age, y=SCORE)) +
  geom_point() +
  geom_smooth(method="lm") +
  labs(title="Restaurant Age vs Inspection Score (Restaurants Only)", x="Restaurant Age (Yea
```

```
`geom_smooth()` using formula = 'y ~ x'
```


```
# Clean up city names and analyze inspection scores by city (for restaurants only)
restaurant_data$CITY <- str_to_upper(restaurant_data$CITY)  # Convert to uppercase

# Recode city names
restaurant_data <- restaurant_data %>%
  mutate(CITY = recode(CITY, "RALEIGH" = "RALEIGH", "RALEGH" = "RALEIGH", "CARY" = "CARY"))

# Group by city and summarize (for restaurants only)
restaurant_city_summary <- restaurant_data %>%
  filter(!is.na(CITY)) %>%
  group_by(CITY) %>%
  summarize(mean_score = mean(SCORE, na.rm = TRUE), sample_size = n())

print(restaurant_city_summary)
```

```
# A tibble: 21 x 3
   CITY          mean_score sample_size
   <chr>              <dbl>       <int>
 1 ANGIER              94.5           1
 2 APEX                97.1         108
 3 CARY                97.3         406
 4 CLAYTON             93             1
 5 FUQUAY VARINA       96.9          49
```

```
 6 FUQUAY-VARINA       97.0          27
 7 GARNER              95.8          93
 8 HOLLY SPRING        99             1
 9 HOLLY SPRINGS       98.0          79
10 KNIGHTDALE          95.1          49
# i 11 more rows
```

```r
# Analyze inspection scores by inspector (for restaurants only)
restaurant_inspector_summary <- restaurant_data %>%
  filter(!is.na(INSPECTOR)) %>%
  group_by(INSPECTOR) %>%
  summarize(mean_score = mean(SCORE, na.rm = TRUE), sample_size = n())

print(restaurant_inspector_summary)
```

```
# A tibble: 38 x 3
   INSPECTOR           mean_score sample_size
   <chr>                    <dbl>       <int>
 1 Angela Myers              96.7         104
 2 Angela Stocks             96.2          36
 3 Brittny Thomas            98             3
 4 Christy Klaus             95.9         100
 5 Cristofer LeClair         97.1          72
 6 Daryl Beasley             95.4          12
 7 David Adcock              95.9           8
 8 Dipatrimarki Farkas       97.7         118
 9 Elizabeth Jackson         95.7          80
10 Ginger Johnson            97.6          35
# i 28 more rows
```

```r
# Check for small sample sizes (for restaurants only)
# Filter for cities with more than 10 restaurant inspections
restaurant_city_summary_filtered <- restaurant_city_summary %>%
  filter(sample_size > 10)  # Ensuring that we focus on cities with more than 10 restaurants

# Filter for inspectors with more than 10 inspections
restaurant_inspector_summary_filtered <- restaurant_inspector_summary %>%
  filter(sample_size > 10)  # Focus on inspectors with more than 10 inspections

# Display the filtered results
print("Filtered City Summary (Cities with more than 10 restaurants):")
```

```
[1] "Filtered City Summary (Cities with more than 10 restaurants):"
```

```r
print(restaurant_city_summary_filtered)
```

```
# A tibble: 13 x 3
   CITY           mean_score sample_size
   <chr>               <dbl>       <int>
 1 APEX                 97.1         108
 2 CARY                 97.3         406
 3 FUQUAY VARINA        96.9          49
 4 FUQUAY-VARINA        97.0          27
 5 GARNER               95.8          93
 6 HOLLY SPRINGS        98.0          79
 7 KNIGHTDALE           95.1          49
 8 MORRISVILLE          96.7         143
 9 RALEIGH              96.6        1193
10 ROLESVILLE           96.0          13
11 WAKE FOREST          96.8         133
12 WENDELL              94.6          20
13 ZEBULON              93.6          31
```

```r
print("Filtered Inspector Summary (Inspectors with more than 10 inspections):")
```

```
[1] "Filtered Inspector Summary (Inspectors with more than 10 inspections):"
```

```r
print(restaurant_inspector_summary_filtered)
```

```
# A tibble: 32 x 3
   INSPECTOR            mean_score sample_size
   <chr>                     <dbl>       <int>
 1 Angela Myers               96.7         104
 2 Angela Stocks              96.2          36
 3 Christy Klaus              95.9         100
 4 Cristofer LeClair          97.1          72
 5 Daryl Beasley              95.4          12
 6 Dipatrimarki Farkas        97.7         118
 7 Elizabeth Jackson          95.7          80
 8 Ginger Johnson             97.6          35
 9 Jackson Hooton             96.3          12
10 Jamie Phelps               97.9         108
# i 22 more rows
```

When the analyses were repeated for restaurants only, the trends largely held. The distribution of scores remained high, with most restaurants scoring between 90 and 100. Additionally, no significant trend was observed between restaurant age and score, and there were some variations in scores by city and inspector. The issue of small sample sizes also remained, particularly in smaller cities and with certain inspectors. Overall, restaurants appear to maintain high sanitation standards across the board, though there are some minor variations depending on location and who conducted the inspection.