

Morris-Proposal

Sharon Morris

October 16, 2016

```
# load data
library(jsonlite)
data <- fromJSON("https://data.cityofnewyork.us/resource/xx67-kt59.json")
head(data)

##   cuisine_description          dba record_date boro
## 1 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 2 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 3 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 4 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 5 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 6 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
##   inspection_date building zipcode score    phone      street
## 1 2016-02-18T00:00:00     1007    10462    10 7188924968 MORRIS PARK AVE
## 2 2016-02-18T00:00:00     1007    10462    10 7188924968 MORRIS PARK AVE
## 3 2015-02-09T00:00:00     1007    10462     6 7188924968 MORRIS PARK AVE
## 4 2014-03-03T00:00:00     1007    10462     2 7188924968 MORRIS PARK AVE
## 5 2013-10-10T00:00:00     1007    10462    <NA> 7188924968 MORRIS PARK AVE
## 6 2013-09-11T00:00:00     1007    10462     6 7188924968 MORRIS PARK AVE
##   grade critical_flag camis
## 1     A     Critical 30075445
## 2     A Not Critical 30075445
## 3     A     Critical 30075445
## 4     A Not Critical 30075445
## 5 <NA> Not Applicable 30075445
## 6     A     Critical 30075445
##                               action
## 1 Violations were cited in the following area(s).
## 2 Violations were cited in the following area(s).
## 3 Violations were cited in the following area(s).
## 4 Violations were cited in the following area(s).
## 5 No violations were recorded at the time of this inspection.
## 6 Violations were cited in the following area(s).
##   violation_code
## 1     04L
## 2     08A
## 3     06C
## 4     10F
## 5    <NA>
## 6     04L
##
## 1
## 2
## 3
## 4 Non-food contact surface improperly constructed. Unacceptable material used. Non-food contact surf
## 5
## 6
```

```

##          grade_date           inspection_type
## 1 2016-02-18T00:00:00 Cycle Inspection / Initial Inspection
## 2 2016-02-18T00:00:00 Cycle Inspection / Initial Inspection
## 3 2015-02-09T00:00:00 Cycle Inspection / Initial Inspection
## 4 2014-03-03T00:00:00 Cycle Inspection / Initial Inspection
## 5             <NA> Trans Fat / Second Compliance Inspection
## 6 2013-09-11T00:00:00     Cycle Inspection / Re-inspection

# Convert dataframe to table
require(dplyr)

## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

dataTbl <- tbl_df(data)
head(dataTbl)

## # A tibble: 6 × 18
##   cuisine_description      dba    record_date   boro
##   <chr>                  <chr>        <chr> <chr>
## 1 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 2 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 3 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 4 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 5 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## 6 Bakery MORRIS PARK BAKE SHOP 2016-10-14T06:02:39 BRONX
## # ... with 14 more variables: inspection_date <chr>, building <chr>,
## #   zipcode <chr>, score <chr>, phone <chr>, street <chr>, grade <chr>,
## #   critical_flag <chr>, camis <chr>, action <chr>, violation_code <chr>,
## #   violation_description <chr>, grade_date <chr>, inspection_type <chr>

```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

What is the grade rating most frequently received by New York City food establishments during the initial inspection by the health department? Is there a correlation between the type of grade received during the initial inspection and the type of food establishment (e.g., bakery vs. cafe vs. restaurant)

Cases

What are the cases, and how many are there?

Each row represent a inspection with 18 variables about the inspection. There are 1,000 cases.

Data collection

Describe the method of data collection. The data were compiled from several New York City Department of Health and Mental Hygiene administrative systems.

The data set contains every sustained or not yet adjudicated violation citation from every full or special program inspection conducted up to three years prior to the most recent inspection for restaurants and college cafeterias

Type of study

What type of study is this (observational/experiment)? This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

NYC OpenData, Department of Health and Mental Hygiene, DOHM, New York City Restaurant Inspection Results', <https://nycopendata.socrata.com/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/xx67-kt59>.

Response

What is the response variable, and what type is it (numerical/categorical)?

The response variable is the grade the food establishment receives. The response variable is categorical.

Explanatory

What is the explanatory variable, and what type is it (numerical/categorical)? The explanatory variable is the inspection type, whether the inspection is initial or a re-inspection. The inspection type is a categorical variable.

Relevant summary statistics

Provide summary statistics relevant to your research question. For example, if you're comparing means across groups provide means, SDs, sample sizes of each group. This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

Some data cleaning using tydr will be required before analysis can occur. Tidyr separate function will be used to separate and to place one inspection type into a column. Inspection type variable will be recorded to a numerical value, before running analysis to determine whether initial or re-inspection has a statistical effect on the type of grade received and the type of food establishment.

The cuisine_type variable will be recorded to a numerical value to support the analysis.

Additionally, the missing values for grade will be assessed to determine whether the grade score can be used to calculate a grade to address some missing values. Missing values will be recorded to zeros instead of NAs to support the analysis.

```
# Select variable for analysis
dataSel <- select(dataTbl, cuisine_description, record_date, inspection_date, score,
                    grade, critical_flag, camis, action, violation_code,
                    violation_description, grade_date, inspection_type)
```

```
# Summary statistics will not run since the majority of the data are not numeric  
# sum(datasel)
```

```
# Filter food establishments by inspection type  
#dataFil <- filter(dataSel, cuisine_description %in% c('bakery', 'hamburgers'))
```