DATA 606 Data Project Proposal

Sung Lee

Data Preparation

```
# load data
library(tidyverse)
library(psych)
library(summarytools)
library(ggplot2)

licenses_df <- read_csv("https://raw.githubusercontent.com/logicalschema/Data606Project/master/License_

names(licenses_df)<-str_replace_all(names(licenses_df), c(" " = "_"))

head(licenses_df)
```

```
## # A tibble: 6 x 25
##   Application_ID License_Number License_Type Application_or_~ Business_Name
##   <chr>          <chr>          <chr>        <chr>            <chr>
## 1 1066-2017-RHIC 1294131-DCA    Business     Renewal          PEYKO TZENOV
## 2 1164-2019-RDPD 1472251-DCA    Business     Renewal          AMERICAN EAS~
## 3 33701-2016-RE~ 2025971-DCA    Business     Renewal          LUCAS ELECTR~
## 4 34278-2018-RE~ 2047043-DCA    Business     Renewal          ELITE WIRELE~
## 5 3891-2019-ALAU 2085269-DCA    Business     Application      A-LIN LAUNDR~
## 6 12418-2017-AE~ 2057790-DCA    Business     Application      RIDGE WIRELE~
## # ... with 20 more variables: Status <chr>, Start_Date <chr>, End_Date <chr>,
## #   Temp_Op_Letter_Issued <date>, Temp_Op_Letter_Expiration <chr>,
## #   License_Category <chr>, Application_Category <chr>, Building_Number <chr>,
## #   Street <chr>, Street_2 <chr>, Unit_Type <chr>, Unit <chr>,
## #   Description <chr>, City <chr>, State <chr>, Zip <chr>, Contact_Phone <chr>,
## #   Longitude <dbl>, Latitude <dbl>, Active_Vehicles <lgl>
```

```
names(licenses_df)
```

```
##  [1] "Application_ID"            "License_Number"
##  [3] "License_Type"              "Application_or_Renewal"
##  [5] "Business_Name"             "Status"
##  [7] "Start_Date"                "End_Date"
##  [9] "Temp_Op_Letter_Issued"     "Temp_Op_Letter_Expiration"
## [11] "License_Category"          "Application_Category"
## [13] "Building_Number"           "Street"
## [15] "Street_2"                  "Unit_Type"
## [17] "Unit"                      "Description"
## [19] "City"                      "State"
```

```
## [21] "Zip"                          "Contact_Phone"
## [23] "Longitude"                    "Latitude"
## [25] "Active_Vehicles"
```

Research question

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

Do economic downturns impact the number of tobacco and liquor license applications in NYC?

Cases

What are the cases, and how many are there?

There are 368,017 applications and renewals for NYC business licenses in varying license categories from 4/10/98 to 4/16/20.

Data collection

Describe the method of data collection.

Data is collected by NYC and recorded on their OpenData platform: https://data.cityofnewyork.us/. This is the original page of the data: https://data.cityofnewyork.us/Business/License-Applications/ptev-4hud/data

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.

This is the link for the data: https://data.cityofnewyork.us/Business/License-Applications/ptev-4hud

Additional information about the data: https://data.cityofnewyork.us/api/views/ptev-4hud/files/e9ee6ec0-796d-4273-853f-0ccd05920c2f?download=true&filename=DCA%20License%20Applications%20Data%20Dictionary.pdf

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The response variable is the *License Category* and it is qualitative.

Independent Variable

You should have two independent variables, one quantitative and one qualitative.

The two independent variables are: *Start Date* (quantitative) and *Application or Renewal* (qualitative).

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```
dfSummary(licenses_df$License_Category)
```

```
## licenses_df$License_Category was converted to a data frame

## Data Frame Summary
## licenses_df
## Dimensions: 368017 x 1
## Duplicates: 367957
##
## ------------------------------------------------------------------------------------------------------------
## No   Variable            Stats / Values                  Freqs (% of Valid)   Graph
## ---- ------------------- ------------------------------- -------------------- ----------------------
## 1    License_Category    1. Amusement Arcade                81 ( 0.0%)
##      [factor]            2. Amusement Device Permanen      793 ( 0.2%)
##                          3. Amusement Device Portable     3849 ( 1.0%)
##                          4. Amusement Device Temporar      497 ( 0.1%)
##                          5. Auction House Premises         140 ( 0.0%)
##                          6. Auctioneer                    1138 ( 0.3%)
##                          7. Bingo Game Operator             79 ( 0.0%)
##                          8. Booting Company                 15 ( 0.0%)
##                          9. Cabaret                        298 ( 0.1%)
##                          10. Car Wash                      364 ( 0.1%)
##                          [ 50 others ]                  360763 (98.0%)       IIIIIIIIIIIIIIIIII
## ------------------------------------------------------------------------------------------------------------
```

```
dfSummary(licenses_df$Start_Date)
```

```
## licenses_df$Start_Date was converted to a data frame

## Data Frame Summary
## licenses_df
## Dimensions: 368017 x 1
## Duplicates: 362564
##
## ------------------------------------------------------------------------------------------------------------
## No   Variable      Stats / Values     Freqs (% of Valid)   Graph                Valid     Missing
## ---- ------------- ------------------ -------------------- -------------------- --------- ---------
## 1    Start_Date    1. 01/01/2014          2 ( 0.0%)                            368017    0
##      [factor]      2. 01/01/2015          3 ( 0.0%)                            (100%)    (0%)
##                    3. 01/01/2017         11 ( 0.0%)
##                    4. 01/01/2018          9 ( 0.0%)
##                    5. 01/01/2019          9 ( 0.0%)
##                    6. 01/01/2020          1 ( 0.0%)
##                    7. 01/02/2001        260 ( 0.1%)
##                    8. 01/02/2002         23 ( 0.0%)
##                    9. 01/02/2003         13 ( 0.0%)
```

```
##                     10. 01/02/2004         14 ( 0.0%)
##                     [ 5443 others ]    367672 (99.9%)       IIIIIIIIIIIIIIIIIII
## ---------------------------------------------------------------------------------------------------
```

```
dfSummary(licenses_df$Application_or_Renewal)
```

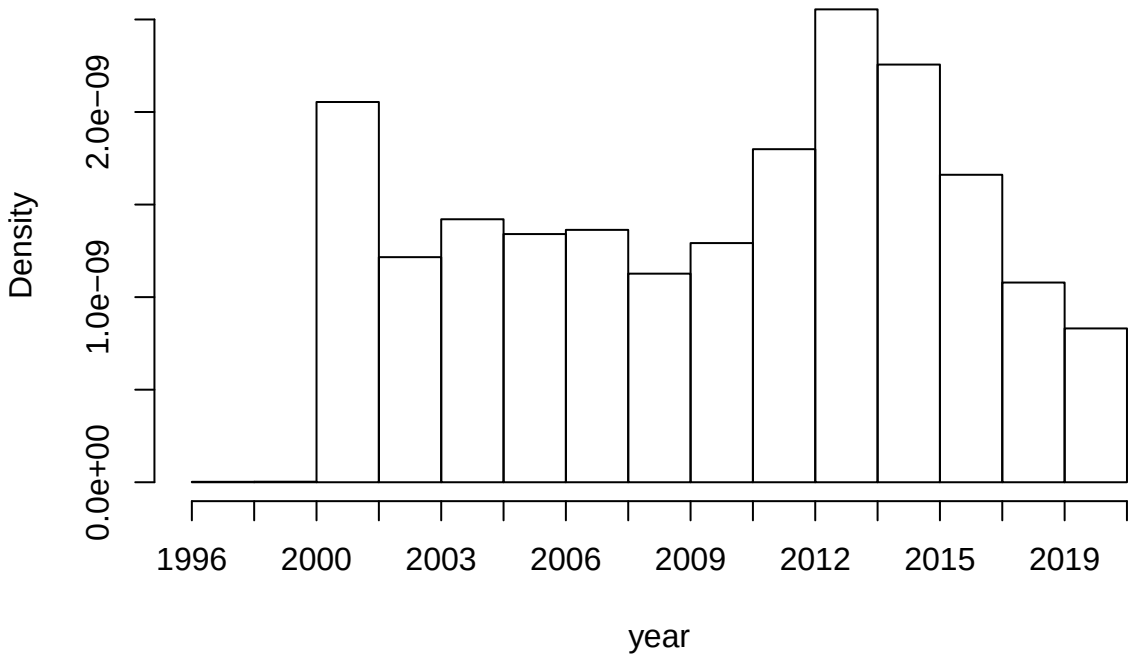
```
## licenses_df$Application_or_Renewal was converted to a data frame

## Data Frame Summary
## licenses_df
## Dimensions: 368017 x 1
## Duplicates: 368015
##
## ---------------------------------------------------------------------------------------------------
## No   Variable                  Stats / Values     Freqs (% of Valid)   Graph          Valid     Miss
## ---- ------------------------- ------------------ -------------------- -------------- --------- -----
## 1    Application_or_Renewal    1. Application     134978 (36.7%)       IIIIIII        368017    0
##      [factor]                  2. Renewal         233039 (63.3%)       IIIIIIIIIII    (100%)    (0%)
## ---------------------------------------------------------------------------------------------------
```

```
tobacco <- filter(licenses_df, License_Category == "Tobacco Retail Dealer")


formatDate <- strptime(as.character(tobacco$Start_Date),format="%m/%d/%Y")
hist(formatDate,breaks=10,xlab="year")
```

```
## Warning in breaks[-1L] + breaks[-nB]: NAs produced by integer overflow
```


Histogram of year
Density
0.0e+00
1.0e−09
2.0e−09
1996
2000
2003
2006
2009
2012
2015
2019
year