Data Aggregation

Edie Espejo

3/30/2019

I just drove myself into a deep dark pit of data terror, but I was able to find one archive from Kaggle using an internet time machine that gave me an old download link that somehow still worked. 159490 businesses intersect between the 2017 data I found and the 2019 data…

Libraries

```
library(jsonlite)
library(tibble)
library(dplyr)
library(tidyr)
library(stringr)
library(readr)
library(ggplot2)
```

2019 Data

I downloaded this straight off of the Yelp Challenge.

```
yelp_2019 <- "../../../data/yelp-2019/business.json"
yelp_2019 <- stream_in(file(yelp_2019))
```

```
##
 Found 500 records...
 Found 1000 records...
 Found 1500 records...
 Found 2000 records...
 Found 2500 records...
 Found 3000 records...
 Found 3500 records...
 Found 4000 records...
 Found 4500 records...
 Found 5000 records...
 Found 5500 records...
 Found 6000 records...
 Found 6500 records...
 Found 7000 records...
 Found 7500 records...
 Found 8000 records...
 Found 8500 records...
 Found 9000 records...
 Found 9500 records...
 Found 10000 records...
 Found 10500 records...
 Found 11000 records...
 Found 11500 records...
 Found 12000 records...
 Found 12500 records...
 Found 13000 records...
 Found 13500 records...
 Found 14000 records...
 Found 14500 records...
 Found 15000 records...
 Found 15500 records...
 Found 16000 records...
 Found 16500 records...
 Found 17000 records...
 Found 17500 records...
 Found 18000 records...
 Found 18500 records...
 Found 19000 records...
 Found 19500 records...
 Found 20000 records...
 Found 20500 records...
 Found 21000 records...
 Found 21500 records...
 Found 22000 records...
 Found 22500 records...
 Found 23000 records...
 Found 23500 records...
 Found 24000 records...
 Found 24500 records...
 Found 25000 records...
 Found 25500 records...
 Found 26000 records...
 Found 26500 records...
 Found 27000 records...
 Found 27500 records...
 Found 28000 records...
 Found 28500 records...
 Found 29000 records...
 Found 29500 records...
 Found 30000 records...
 Found 30500 records...
 Found 31000 records...
 Found 31500 records...
 Found 32000 records...
 Found 32500 records...
 Found 33000 records...
 Found 33500 records...
 Found 34000 records...
 Found 34500 records...
 Found 35000 records...
 Found 35500 records...
 Found 36000 records...
 Found 36500 records...
 Found 37000 records...
 Found 37500 records...
 Found 38000 records...
 Found 38500 records...
 Found 39000 records...
 Found 39500 records...
 Found 40000 records...
 Found 40500 records...
 Found 41000 records...
 Found 41500 records...
 Found 42000 records...
 Found 42500 records...
 Found 43000 records...
 Found 43500 records...
 Found 44000 records...
 Found 44500 records...
 Found 45000 records...
 Found 45500 records...
 Found 46000 records...
 Found 46500 records...
 Found 47000 records...
 Found 47500 records...
 Found 48000 records...
 Found 48500 records...
 Found 49000 records...
 Found 49500 records...
 Found 50000 records...
 Found 50500 records...
 Found 51000 records...
 Found 51500 records...
 Found 52000 records...
 Found 52500 records...
 Found 53000 records...
 Found 53500 records...
 Found 54000 records...
 Found 54500 records...
 Found 55000 records...
 Found 55500 records...
 Found 56000 records...
 Found 56500 records...
 Found 57000 records...
 Found 57500 records...
 Found 58000 records...
 Found 58500 records...
 Found 59000 records...
 Found 59500 records...
 Found 60000 records...
 Found 60500 records...
 Found 61000 records...
 Found 61500 records...
 Found 62000 records...
 Found 62500 records...
 Found 63000 records...
 Found 63500 records...
 Found 64000 records...
 Found 64500 records...
 Found 65000 records...
 Found 65500 records...
 Found 66000 records...
 Found 66500 records...
 Found 67000 records...
 Found 67500 records...
 Found 68000 records...
 Found 68500 records...
 Found 69000 records...
 Found 69500 records...
 Found 70000 records...
 Found 70500 records...
 Found 71000 records...
 Found 71500 records...
 Found 72000 records...
 Found 72500 records...
 Found 73000 records...
 Found 73500 records...
 Found 74000 records...
 Found 74500 records...
 Found 75000 records...
 Found 75500 records...
 Found 76000 records...
 Found 76500 records...
 Found 77000 records...
 Found 77500 records...
 Found 78000 records...
 Found 78500 records...
 Found 79000 records...
 Found 79500 records...
 Found 80000 records...
 Found 80500 records...
 Found 81000 records...
 Found 81500 records...
 Found 82000 records...
 Found 82500 records...
 Found 83000 records...
 Found 83500 records...
 Found 84000 records...
 Found 84500 records...
 Found 85000 records...
 Found 85500 records...
 Found 86000 records...
 Found 86500 records...
 Found 87000 records...
 Found 87500 records...
 Found 88000 records...
 Found 88500 records...
 Found 89000 records...
 Found 89500 records...
 Found 90000 records...
 Found 90500 records...
 Found 91000 records...
 Found 91500 records...
 Found 92000 records...
 Found 92500 records...
 Found 93000 records...
 Found 93500 records...
 Found 94000 records...
 Found 94500 records...
 Found 95000 records...
 Found 95500 records...
 Found 96000 records...
 Found 96500 records...
 Found 97000 records...
 Found 97500 records...
 Found 98000 records...
 Found 98500 records...
 Found 99000 records...
 Found 99500 records...
 Found 1e+05 records...
 Found 100500 records...
 Found 101000 records...
 Found 101500 records...
 Found 102000 records...
 Found 102500 records...
 Found 103000 records...
 Found 103500 records...
 Found 104000 records...
 Found 104500 records...
 Found 105000 records...
 Found 105500 records...
 Found 106000 records...
 Found 106500 records...
 Found 107000 records...
 Found 107500 records...
 Found 108000 records...
 Found 108500 records...
 Found 109000 records...
 Found 109500 records...
 Found 110000 records...
 Found 110500 records...
 Found 111000 records...
 Found 111500 records...
 Found 112000 records...
 Found 112500 records...
 Found 113000 records...
 Found 113500 records...
 Found 114000 records...
 Found 114500 records...
 Found 115000 records...
 Found 115500 records...
 Found 116000 records...
 Found 116500 records...
 Found 117000 records...
 Found 117500 records...
 Found 118000 records...
 Found 118500 records...
 Found 119000 records...
 Found 119500 records...
 Found 120000 records...
 Found 120500 records...
 Found 121000 records...
 Found 121500 records...
 Found 122000 records...
 Found 122500 records...
 Found 123000 records...
 Found 123500 records...
 Found 124000 records...
 Found 124500 records...
 Found 125000 records...
 Found 125500 records...
 Found 126000 records...
 Found 126500 records...
 Found 127000 records...
 Found 127500 records...
 Found 128000 records...
 Found 128500 records...
 Found 129000 records...
 Found 129500 records...
 Found 130000 records...
 Found 130500 records...
 Found 131000 records...
 Found 131500 records...
 Found 132000 records...
 Found 132500 records...
 Found 133000 records...
 Found 133500 records...
 Found 134000 records...
 Found 134500 records...
 Found 135000 records...
 Found 135500 records...
 Found 136000 records...
 Found 136500 records...
 Found 137000 records...
 Found 137500 records...
 Found 138000 records...
 Found 138500 records...
 Found 139000 records...
 Found 139500 records...
 Found 140000 records...
 Found 140500 records...
 Found 141000 records...
 Found 141500 records...
 Found 142000 records...
 Found 142500 records...
 Found 143000 records...
 Found 143500 records...
 Found 144000 records...
 Found 144500 records...
 Found 145000 records...
 Found 145500 records...
 Found 146000 records...
 Found 146500 records...
 Found 147000 records...
 Found 147500 records...
 Found 148000 records...
 Found 148500 records...
 Found 149000 records...
 Found 149500 records...
 Found 150000 records...
 Found 150500 records...
 Found 151000 records...
 Found 151500 records...
 Found 152000 records...
 Found 152500 records...
 Found 153000 records...
 Found 153500 records...
 Found 154000 records...
 Found 154500 records...
 Found 155000 records...
 Found 155500 records...
 Found 156000 records...
 Found 156500 records...
 Found 157000 records...
 Found 157500 records...
 Found 158000 records...
 Found 158500 records...
 Found 159000 records...
 Found 159500 records...
 Found 160000 records...
 Found 160500 records...
 Found 161000 records...
 Found 161500 records...
 Found 162000 records...
 Found 162500 records...
 Found 163000 records...
 Found 163500 records...
 Found 164000 records...
 Found 164500 records...
 Found 165000 records...
 Found 165500 records...
 Found 166000 records...
 Found 166500 records...
 Found 167000 records...
 Found 167500 records...
 Found 168000 records...
 Found 168500 records...
 Found 169000 records...
 Found 169500 records...
 Found 170000 records...
 Found 170500 records...
 Found 171000 records...
 Found 171500 records...
 Found 172000 records...
 Found 172500 records...
 Found 173000 records...
 Found 173500 records...
```

```
 Found 174000 records...
 Found 174500 records...
 Found 175000 records...
 Found 175500 records...
 Found 176000 records...
 Found 176500 records...
 Found 177000 records...
 Found 177500 records...
 Found 178000 records...
 Found 178500 records...
 Found 179000 records...
 Found 179500 records...
 Found 180000 records...
 Found 180500 records...
 Found 181000 records...
 Found 181500 records...
 Found 182000 records...
 Found 182500 records...
 Found 183000 records...
 Found 183500 records...
 Found 184000 records...
 Found 184500 records...
 Found 185000 records...
 Found 185500 records...
 Found 186000 records...
 Found 186500 records...
 Found 187000 records...
 Found 187500 records...
 Found 188000 records...
 Found 188500 records...
 Found 189000 records...
 Found 189500 records...
 Found 190000 records...
 Found 190500 records...
 Found 191000 records...
 Found 191500 records...
 Found 192000 records...
 Found 192500 records...
 Found 192609 records...
 Imported 192609 records. Simplifying...
```

```r
yelp_2019 <- flatten(yelp_2019)
yelp_2019 <- as_tibble(yelp_2019)
```

2017 Data

I used the wayback machine to get me this download link.

```r
wayback <- "../../../data/yelp-2017/yelp_business.csv"
wayback <- read_csv(wayback)
head(wayback)
```

```
## # A tibble: 6 x 13
##   business_id name  neighborhood address city  state postal_code latitude
##   <chr>       <chr> <chr>        <chr>   <chr> <chr> <chr>          <dbl>
## 1 FYWN1wneV1~ "\"D~ <NA>         "\"485~ Ahwa~ AZ    85044           33.3
## 2 He-G7vWjzV~ "\"S~ <NA>         "\"310~ McMu~ PA    15317           40.3
## 3 KQPW81Ff1y~ "\"W~ <NA>         "\"602~ Phoe~ AZ    85017           33.5
## 4 8DShNS-LuF~ "\"S~ <NA>         "\"500~ Tempe AZ    85282           33.4
## 5 PfOCPjBrlQ~ "\"B~ <NA>         "\"581~ Cuya~ OH    44221           41.1
## 6 o9eMRCWt5P~ "\"M~ <NA>         "\"Ric~ Stut~ BW    70567           48.7
## # ... with 5 more variables: longitude <dbl>, stars <dbl>,
## #   review_count <dbl>, is_open <dbl>, categories <chr>
```

Combining the datasets

I want to get choose to use businesses that are in both the datasets only.

```
intersecting_businesses <- intersect(wayback$business_id, yelp_2019$business_id)
length(intersecting_businesses)
```

```
## [1] 159490
```

Sanity checks…

```
wayback_subset   <- wayback %>% filter(business_id %in% intersecting_businesses)
yelp_2019_subset <- yelp_2019 %>% filter(business_id %in% intersecting_businesses)
c(nrow(wayback_subset), nrow(yelp_2019_subset))
```

```
## [1] 159490 159490
```

I'm going to just collect the 0's and 1's.

```
yelp_2018_subberset <- yelp_2019_subset %>% select(business_id, is_open) %>% rename(open_2019=is_open)
head(yelp_2018_subberset)
```

```
## # A tibble: 6 x 2
##   business_id            open_2019
##   <chr>                      <int>
## 1 1SWheh84yJXfytovILXOAQ         0
## 2 QXAEGFB4oINsVuTFxEYKFQ         1
## 3 gnKjwL_1w79qoiV3IC_xQQ         1
## 4 68dUKd8_8liJ7in4aWOSEA         1
## 5 gbQN7vr_caG_A1ugSmGhWg         1
## 6 Y6iyemLX_oylRpnr38vgMA         0
```

```
wayback_subset <- wayback_subset %>% rename(open_2017=is_open)
causal_set <- merge(wayback_subset, yelp_2018_subberset, by="business_id")
head(causal_set)
```

```
##              business_id                             name
## 1 __1uG7MLxWGFIv2fCGPiQQ       "SpinalWorks Chiropractic"
## 2 __3I-DDkqM9XjLH1cJl3VA        "Montallegro Barber Shop"
## 3 __3qOwWFBUE8mdOToI7YrQ                   "Custom Kings"
## 4 __47_7H-yK3HChO5vyut_Q "Instant Muffler and Autorepair"
## 5 __6jYJ6Hm-Qq8XQEGDrOGQ               "Winfield Gene DO"
## 6 __8j8yhsmE98wNWHJNyAgw                    "Urawa Sushi"
##                            neighborhood                  address      city
## 1                                  <NA> "15640 N 7th St, Ste A3"   Phoenix
## 2 Villeray-Saint-Michel-Parc-Extension     "7244 Rue Hutchison"  Montreal
## 3                             Southeast                       "" Las Vegas
## 4                                 <NA>       "1295 Weston Road"      York
## 5                                 <NA>        "2121 S Mill Ave"     Tempe
## 6               Entertainment District  "254 Adelaide Street W"   Toronto
##   state postal_code latitude  longitude stars review_count open_2017
## 1    AZ       85022 33.62885 -112.06598     5           26         1
## 2    QC     H3N 1Z1 45.52986  -73.62373     5           13         1
## 3    NV       88901 36.05566 -115.16942     1           12         1
## 4    ON     M6M 4R2 43.68924  -79.49529     1            3         1
## 5    AZ       85282 33.40559 -111.93944     4            4         1
## 6    ON     M5H 1X6 43.64823  -79.38926     3           73         1
##                                           categories open_2019
## 1 Physical Therapy;Chiropractors;Health & Medical         1
## 2               Hair Salons;Barbers;Beauty & Spas         1
## 3 Screen Printing/T-Shirt Printing;Local Services         1
## 4                          Auto Repair;Automotive         1
## 5                        Doctors;Health & Medical         1
## 6                 Restaurants;Japanese;Sushi Bars         1
```

The businesses we look at should have been open in 2017.

```
causal_set <- causal_set[which(causal_set$open_2017==1),]
dim(causal_set)
```

```
## [1] 133828     14
```

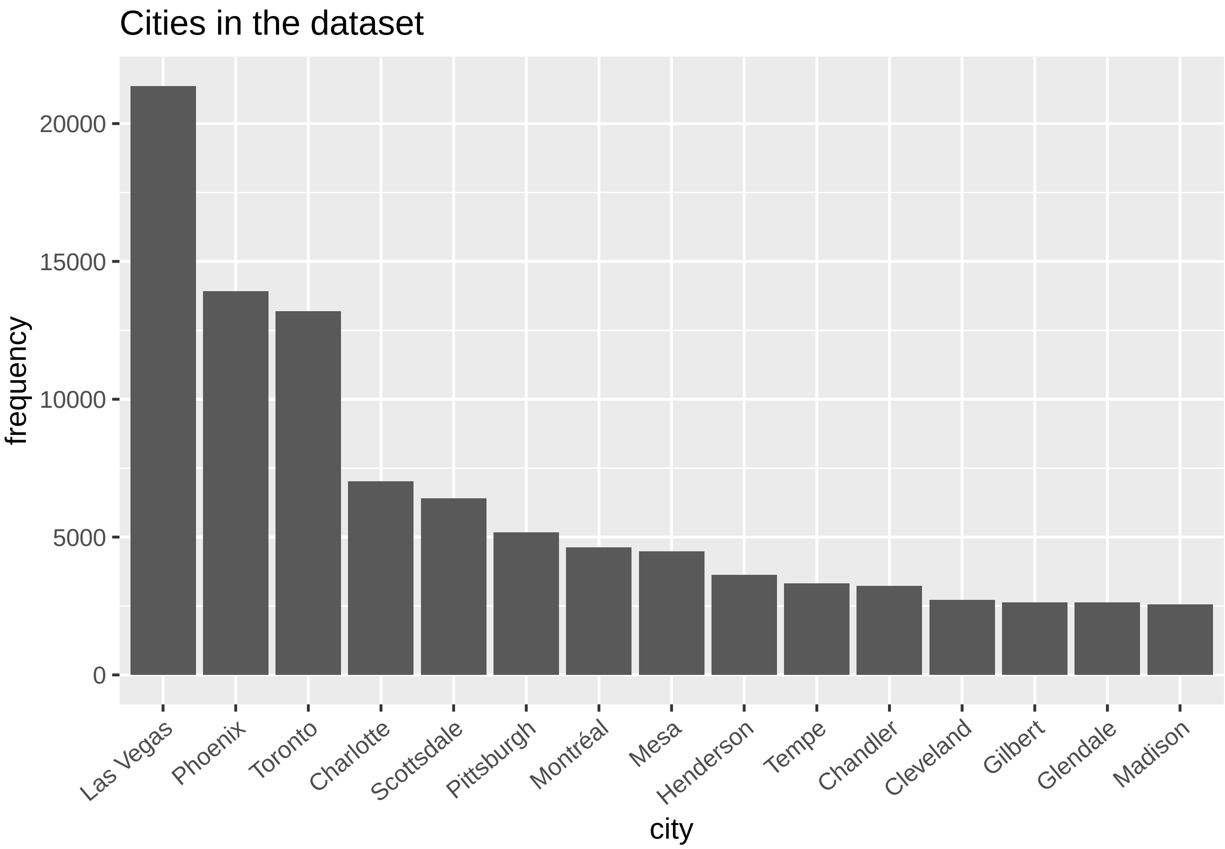
According to this, 6,267 closed in two years. The other 127,561 carried on.

```
table(causal_set$open_2019)
```

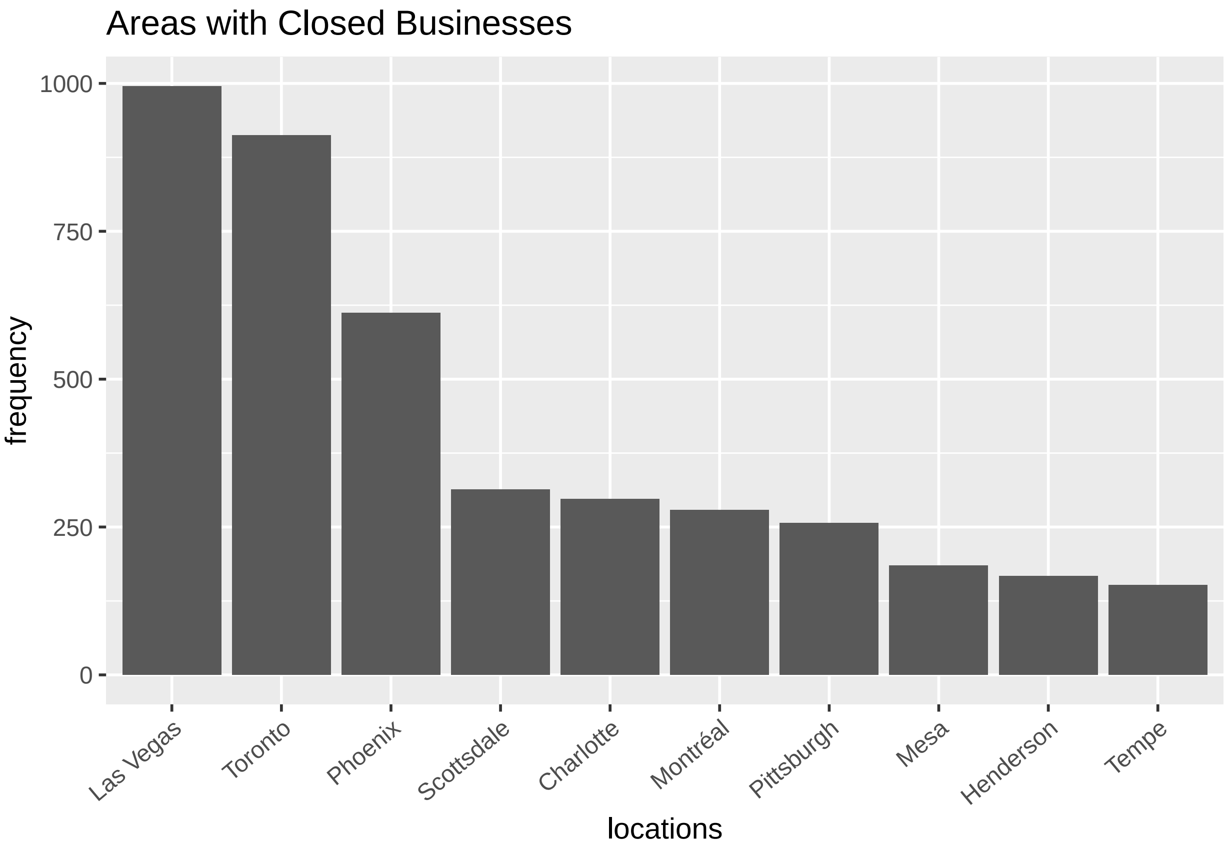
```
## 
##      0      1 
##   6267 127561
```

```
causal_set %>% filter(open_2019==0)
```

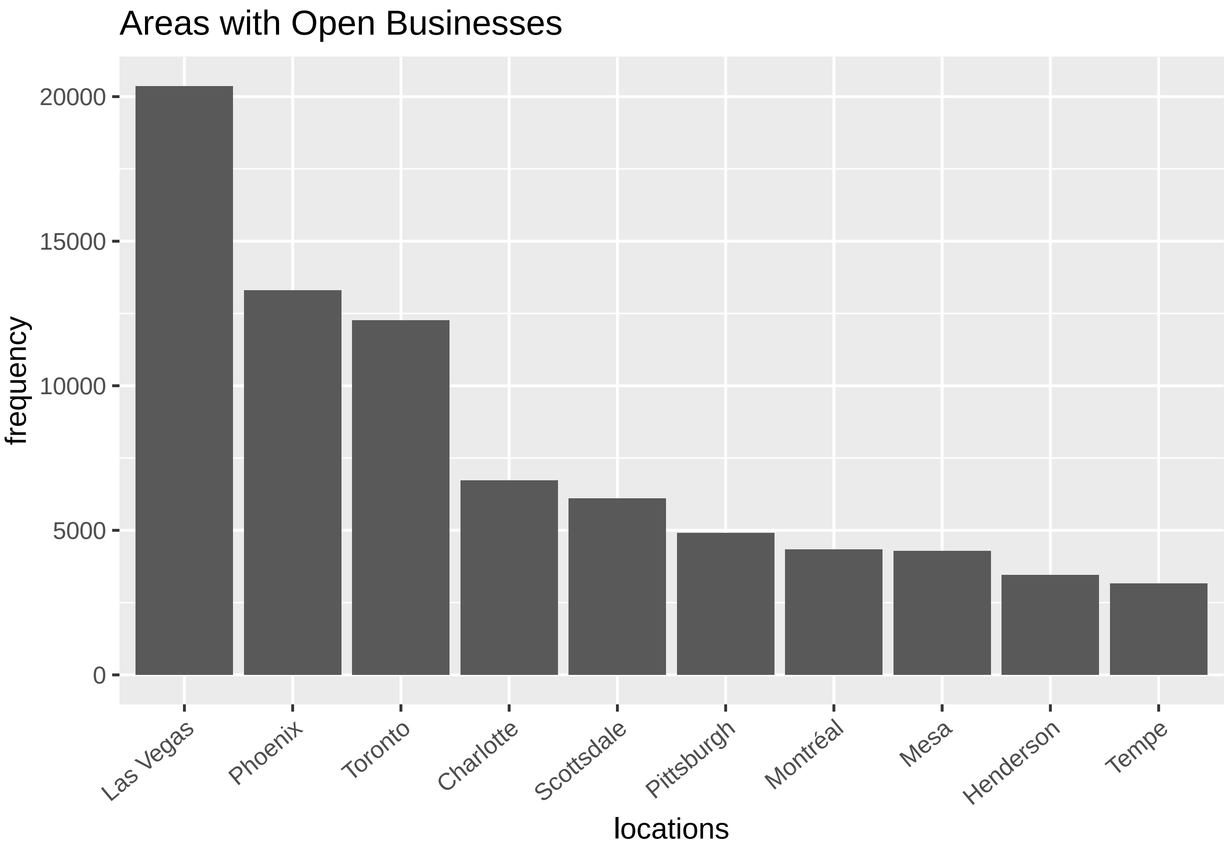
```
causal_set_cities <- data.frame(sort(table(causal_set$city), decreasing=TRUE))
names(causal_set_cities) <- c("city", "frequency")
ggplot(head(causal_set_cities, 15), aes(x=city, y=frequency)) + geom_bar(stat="identity") + theme(axis.
```



```
closed_businesses <- data.frame(sort(table(causal_set %>% filter(open_2019==0) %>% pull(city)), decreas
names(closed_businesses) <- c("locations", "frequency")
ggplot(head(closed_businesses, 10), aes(x=locations, y=frequency)) + geom_bar(stat="identity") + theme(
```



```r
open_businesses <- data.frame(sort(table(causal_set %>% filter(open_2019==1) %>% pull(city)), decreasing
names(open_businesses) <- c("locations", "frequency")
ggplot(head(open_businesses, 10), aes(x=locations, y=frequency)) + geom_bar(stat="identity") + theme(ax
```


Areas with Open Businesses
frequency
20000
15000
10000
5000
0
Las Vegas
Phoenix
Toronto
Charlotte
Scottsdale
Pittsburgh
Montréal
Mesa
Henderson
Tempe
locations