Hannah Keim

Chi-Square Analyses of New York City Business Licenses

Analyzing Feature Relationships Using the Chi-Square Test

January 13, 2022

Research Question

When opening a new business, one of the first major decisions made is the location. In New York City, home of over 200,000 businesses, this includes the choice of a borough for an operation (City of New York). An aspiring business owner in NYC might ask: is there a relationship between the boroughs and the number of licensed businesses in that borough? The NYC Department of Consumer Affairs (DCA) keeps records of every business and individual holding a DCA operations license (City of New York, 2022). These records are published publicly by the city on the website NYC OpenData, allowing analysis.

This analysis assesses if there is a statistically significant relationship between a business's address borough and its business license status, and thus add statistical insight to the above asked question. The null hypothesis for this analysis is that there is no significant relationship between borough and license status. The alternative hypothesis is that there is a significant relationship between borough and license status.

Data Collection

The original Legally Operating Businesses dataset consists of 272,157 rows and twenty-seven columns. Features of the dataset include license numbers and information, business address fields, contact information, and details of the business's industry. Data is recorded in categorical and continuous features. The data was acquired from NYC OpenData and had no stated restrictions on its use (City of New York, 2022).

The advantage to using this data is that it is well organized and maintained by the DCA, including weekly updates. Its ability to be publicly used is also advantageous. One disadvantage of the data set is that there are 93,262 rows with null values for the independent feature. The cleaning of these values will delimit the study. Redundant features were a challenge in this data,

which was overcome by selecting the more descriptive feature containing the same information to preserve value.

Data Extraction and Preparation

The data extraction began by downloading the Legally Operating Businesses dataset from NYC OpenData in a CSV file. It was opened in a Jupyter Lab notebook environment using Python 3. The data was inspected, and many features were determined to be irrelevant to this study. Redundant features dropped include address information not identifying the borough, the borough code, and other location related community information. Irrelevant features dropped include license numbers, expiration and creation dates, contact information, and the industry type. Only three features were kept for subsequent analysis. Those features were renamed to type, status, and borough for ease of analysis. The steps to complete this part of the preparation are below.

[215]:
```
# Import libraries and data set
import pandas as pd
import numpy as np

df = pd.read_csv("C:/Users/hkeim/OneDrive/Documents/School/Capstone/Legally_Operating_Businesses.csv", low_memory = False)
```

[216]:
```
# Find redundant/unnecessary features
df.head()
```

[216]:

	DCA License Number	License Type	License Expiration Date	License Status	License Creation Date	Industry	Business Name	Business Name 2	Address Building	Address Street Name	...	Community Board	Council District	BIN	BBL	NTA	Census Tract	Detail	Longitude	Latitude	Location
0	1412954-DCA	Individual	9/30/2012	Inactive	11/4/2011	General Vendor	LATELLA, SALVATORE F.	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	General Vendor Type: General Vendor (White)	NaN	NaN	NaN
1	0967332-DCA	Business	2/28/2017	Inactive	2/4/2010	Home Improvement Contractor	BARBARINO, JOHN JR.	JOHN BARBARINO JR_HOME IMPROVEMENT	239	MEDFORD CT	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	1057563-DCA	Business	2/28/2023	Active	7/27/2000	Home Improvement Contractor	HITE CONSTRUCTION, INC.	NaN	60	WHITNEY RD	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	1108528-DCA	Individual	3/31/2004	Inactive	5/10/2002	Sightseeing Guide	STEWART, KATHLEEN J	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	1400778-DCA	Individual	4/30/2014	Inactive	5/31/2013	Pedicab Driver	DURASINOVIC, MILOS	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

5 rows × 27 columns

[217]:
```
# Drop redundant/unnecessary features
df = df.filter(items = ['License Type', 'License Status', 'Address Borough'])
df = df.rename(columns={'License Type': 'type', 'license Status': 'status', 'Address Borough': 'borough'})
```

[218]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 272157 entries, 0 to 272156
Data columns (total 3 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   type     272157 non-null  object
 1   status   272157 non-null  object
 2   borough  178895 non-null  object
dtypes: object(3)
memory usage: 6.2+ MB
```

Next, unique and null values in each column were investigated. Because borough was the independent feature for this analysis, and missing values compromise the reliability of a model's results, all rows with null values for this feature were dropped (Kwak & Kim, 2017).

```
[5]: # Determine how many unique values each feature has
counts = df.nunique()
print(counts)
```

```
type        2
status      2
borough    10
dtype: int64
```

```
[6]: # Determine the number of null values in each column
df.isnull().sum()
```

```
[6]: type          0
status        0
borough    93262
dtype: int64
```

```
[7]: # Drop all rows with a null value for borough
df.dropna(subset=['borough'], how='all', inplace = True)
```

```
[8]: # Check null count again
df.isnull().sum()
```

```
[8]: type       0
status     0
borough    0
dtype: int64
```

After cleaning the null values, the consistency of the conventions within each column were investigated by displaying each feature's unique values. This step discovered that as the nulls were removed, so were any rows that had a type value of 'individual.' Because a Chi-Square test cannot be run accurately on univariate data, this feature was not included in further analysis (The SciPy Community, 2008-2021). This step also discovered that the borough feature had redundant values, which were cleaned using the .replace() method. The all-capital boroughs were renamed to match their properly capitalized counterparts. At this point, the data was ready for analysis.

```
[9]: # Investigate the consistancy of the feature values
print(df['type'].unique())
print(df['status'].unique())
print(df['borough'].unique())
```

```
['Business']
['Inactive' 'Active']
['Outside NYC' 'Bronx' 'Queens' 'Brooklyn' 'Manhattan' 'Staten Island'
 'BROOKLYN' 'MANHATTAN' 'QUEENS' 'BRONX']
```

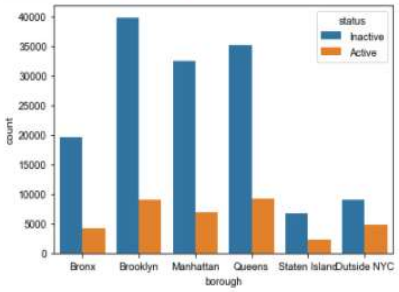
```
[10]: # Clean borough feature values
df['borough'].replace({'BRONX': 'Bronx', 'QUEENS': 'Queens', 'BROOKLYN': 'Brooklyn', 'MANHATTAN': 'Manhattan'}, inplace=True)
print(df['borough'].unique())
```

```
['Outside NYC' 'Bronx' 'Queens' 'Brooklyn' 'Manhattan' 'Staten Island']
```

Analysis

The analysis began by visualizing the data in a count plot to get an idea how the counts of active and inactive licenses were distributed through the boroughs. The count plot shows more inactive than active licenses in every borough, but particularly outside of Staten Island. The visually appears to have a normal distribution but does not appear to be very equally distributed between features.

```
[11]: # Visualize data counts
      import seaborn as sns
      ax = sns.countplot(data=df, x='borough', hue='status', order=['Bronx', 'Brooklyn', 'Manhattan', 'Queens', 'Staten Island', 'Outside NYC'])
      sns.set(rc={'figure.figsize':(10, 10)})
```



Using the Pandas crosstab method, the features were arranged into a contingency table, both to prepare the data for the Chi-Square test, and to visualize the data numerically. In the table, it is important to note that the data fits the assumptions for a Chi-Squared test. The frequencies are all greater than five, both features are categorical and nominal, observations are independent, and observations are frequency counts as opposed to a different aggregation (McHugh, 2013).

```
[12]: # Create contingency table
      statusborough =pd.crosstab(df.status, df.borough)
      print(statusborough)
```

borough	Bronx	Brooklyn	Manhattan	Outside NYC	Queens	Staten Island
status						
Active	4269	8897	6937	4819	9207	2189
Inactive	19653	39792	32462	8922	35043	6705

One advantage of using the Chi-Square test of independence to determine if there is a statistically significant relationship between borough and license status is the ability of Chi-Square to analyze categorical and nominal data. Another is the availability of SciPy's chi2_contingency method to easily run the test and display the resulting statistics (The SciPy Community, 2008-2021). One disadvantage is that while the Chi-Square test can identify a relationship between features, it cannot inform the causality in any way (University of Utah Department of Sociology, 2022). In other words, the test can tell us there is or is not a relationship, with no insight as to why.

Using the contingency table and SciPy's chi2_contingency method, the Chi-Square test of independence was run. The test resulted in a Chi-Square statistic of 2351.3, a p-value of 0.0, and a degree of freedom of five.

```
[14]: # Calculate Chi2 score, p-value, degree of freedom
      from scipy.stats import chi2_contingency
      print(chi2_contingency(statusborough)[0:3])
      print()

      (2351.2799267728856, 0.0, 5)
```

Data Summary and Implications

This analysis has determined that the borough in which a business operates does have a statistically significant relationship with the status of the business's license. For this analysis, the standard p-value threshold of 0.05 is sufficient (University of Southhampton, 2022). As the p-value from the above test is less than the threshold, the null hypothesis is rejected. The alternative hypothesis is accepted. Based on this, I recommend considering Manhattan as a business location. In addition to showing a statistically significant relationship between borough and license status, it also shows that Manhattan has a high number of active licenses compared to the Bronx, Staten Island, and outside NYC. Yet, it has room to grow in active license compared

to Brooklyn and Queens. This makes it a good candidate location for a new active business license (Heaslip, 2020).

While the relationship between city borough and license status is statistically significant, further study is needed to determine a cause for that relationship. One recommendation for further study would be to include more features in further analysis to determine potential relational causes. For instance, how many of the licenses labeled inactive in each borough are also expired? A second recommendation for further study is to use the Chi-Square test findings as potential feature selection for a classifier model to gain further business insights.

References

City of New York. (2022, January 7). *Legally Operating Businesses*. Retrieved from NYC OpenData: https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh

City of New York. (n.d.). *Small Business First.* Retrieved January 11, 2022, from https://www1.nyc.gov/assets/smallbizfirst/downloads/pdf/small-business-first-report.pdf

Heaslip, E. (2020, January 10). *Location, Location, Location: How to Decide Where to Start Your Business*. Retrieved from uschamber.com: https://www.uschamber.com/co/start/strategy/where-to-start-a-business

Kwak, S. K., & Kim, J. H. (2017, August). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, 407-411.

McHugh, M. L. (2013, June 23). The Chi-square test of independence. *Biochemia Medica*, 143-149.

The SciPy Community. (2008-2021). *scipy.stats.chi2_contingency.* Retrieved from SciPy documentation: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

University of Southhampton. (2022). *Chi Square*. Retrieved from University of Southhampton: https://www.southampton.ac.uk/passs/full_time_education/bivariate_analysis/chi_square.page

University of Utah Department of Sociology. (2022). The Chi-Square Test for Independence. Salt Lake City, Utah, USA.