PROJECT #2 SUMMARY

TO: Christopher Llop

FROM: Bobby Calzaretta, Steve Dille, Ellie Huang, Justin Wu

DATE: December 11, 2018

RE: Project #2: Yelp Data Analysis - Do Yelp Reviews Matter?

I. Overview

Yelp is an online business directory that crowd-sources data from users on business attributes and quality through user reviews. The directory is widely relied upon with tens of millions of unique monthly visitors and over 171 million reviews to date.[1] Some businesses covered by Yelp are registered by the businesses themselves while others are added to Yelp by reviewers. Yelp makes a sample of their data available to researchers in its regular "Yelp Dataset Challenge." The data include json-formatted datasets on businesses, users, reviews, photos, check-ins, and tips for a variety of geographies.[2]

This project relies on Yelp business data from the Yelp Dataset Challenge as well as data from the U.S. Census Bureau's 2010 Decennial Census Survey and 2016 ZIP Code Business Patterns Survey to answer the question: Do Yelp Reviews Matter? That is, can distinctions be identified between the attributes of open businesses and those of closed businesses? We analyze these attributes across different categories of businesses and geography types (*i.e.*, urban, suburban, or rural) and competitive landscapes. Other questions addressed by our analysis include:

- How representative is Yelp's Dataset Challenge data?
- Are there any trends in reviews?
- Are there any trends in check-ins?

Our results show that there is likely a positive association between Yelp reviews and whether or not a business is operational. However, other factors such as the population or number of establishments in a business' area may have a larger impact on the business' success. Ultimately, we conclude that more in-depth research and more sophisticated econometric modeling is required before one can adequately determine the magnitude of these effects.

[1] "About Us," Yelp.com, accessed December 2, 2018, available at https://www.Yelp.com/about.
[2] See, "Yelp Dataset JSON," Yelp.com, accessed December 2, 2018, available at https://www.Yelp.com/dataset/documentation/main.

II. About the Data

We rely upon four data files for this analysis. The primary data we use are derived from the "yelp_academic_dataset_business.json" file. This is a dataset from Yelp covering details on businesses in the Yelp Dataset Challenge sample. It contains 15 columns and 188,593 rows. A record in this dataset is an individual business and its attributes. The full list of fields in this dataset are: 'address', 'attributes', 'business_id', 'categories', 'city', 'hours','is_open', 'latitude', 'longitude', 'name', 'neighborhood', 'postal_code', 'review_count', 'stars', and 'state'. Our analysis primarily focuses on the location fields, 'review_count', 'stars', and 'is_open' field which allow analysts to identify the location of a business, the frequency with which it is reviewed, the overall sentiment towards the business, and whether the business is active or inactive (*i.e.*, no longer operational). The other Yelp dataset we rely on is the "yelp_academic_dataset_checkin.json" file. It contains 2 columns and 157,075 records. A record in this dataset is an individual business. The fields in this dataset are 'business_id', and 'time', corresponding to a business in the business file discussed above and a dictionary containing keys of day-of-week and hour-of-day with values that constitute the counts of check-ins.

The "sf12010860us1.csv" file contains a dataset of Census data by ZIP Code Tabulation Area ("ZCTA") made available by the National Bureau of Economic Research.[3] It contains 18 columns and 33,120 rows. A record in this dataset is an individual ZCTA and its characteristics. These data include the following fields: 'fileid', 'stusab', 'chariter', 'cifsn', 'logrecno', 'p0010001',' 'sumlev', 'geocomp', 'zcta5', 'arealand', 'areawatr', 'name', ''funcstat', 'pop100', 'hu100', 'intptlat', 'intptlon', and 'lsadc'. For the purposes of our analysis, we rely only upon the 'name', 'pop100', and 'arealand' fields which contain information on the ZIP Code Tabulation Area name, total population, and land area in square meters, respectively. As ZIP Codes are a proprietary geography maintained by the US Postal Service and are liable to change over time according to the needs of the Postal Service, the U.S. Census Bureau developed an approximation of ZIP Code geographies, formally referred to a ZIP Code Tabulation Areas.[4]

The file "BP_2016_00CZ1.csv" contains a dataset of Census data by actual ZIP Code on business presence and select characteristics.[5] It contains 12 columns and 38,722 rows. A record in this dataset is

[3] "Census 2010 Zip Code Data -- 5-Digit Zip Code Tabulation Area (ZCTA) Data from Summary File 1", the National Bureau of Economic Research, accessed December 2, 2018, available at http://www.nber.org/data/census-2010-zip-code-data.html.

[4] For more details on this geography, see, "ZIP Code Tabulation Area (ZCTA)", U.S. Census Bureau Fact Finder, accessed on December 2, 2018, available at https://factfinder.census.gov/help/en/zip_code_tabulation_area_zcta.htm.

[5] The data are provided by the U.S. Census Bureau's American FactFinder data portal, available at https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=BP_2016_00CZ1&prodType=table.

an individual USPS ZIP Code and its business characteristics. The data include the following fields: 'GEO.id', 'GEO.id2', 'GEO.display-label', 'GEO.annotation.id', \'NAICS.id', 'NAICS.display-label', 'NAICS.annotation.id', 'YEAR.id', 'ESTAB', 'EMP', 'PAYQTR1', 'PAYANN'. Our analysis focuses on the 'GEO.display-label', 'ESTAB', ['EMP', and 'PAYANN'] fields which contain information on the ZIP Code name, number of businesses, the total employment of those businesses, and the total annual payroll of those business, respectively.

III. Requirements

An important aspect of our analysis is evaluating the geographic distribution of the data provided by Yelp. We employ cartographic maps to visuals these distributions and inform our decisions on how to analyze the data. In order to map data, the "CartoPy" package needs to be installed. This can be done via the terminal using the following command:

```
conda install -c conda-forge cartopy
```

This package allows for the handling of geospatial information and allows for the plotting of the latitude and longitude information provided in the Yelp dataset.[6] More information on this package can be found at scitools.org.uk.

IV. Data Cleaning

Several modifications to the data are needed before they can be used for the proposed analysis. First, given the Yelp data lacks a geographical classification, we rely on population density statistics to classify business locations as rural, suburban, or urban.[7] Population density is derived from the U.S. Census data; therefore, we must first limit the Yelp business data to U.S. states using the field "state." In this process, we throw out any records where the "state" field does not correspond to one of the fifty U.S. states or Washington, D.C. This results in a dataset of 139,182 records (or about 74% of the original data sample).

[6] See "Introduction," CartoPy, accessed on December 3, 2018, available at https://scitools.org.uk/cartopy/docs/latest/.

[7] While the Census Bureau maintains definitions of the "rural" and "urban" areas, there is no official "suburban" distinction and the Census delineations are maintained at larger geographic aggregation than ZIP Code. Furthermore, the Census Bureau only updates their classification of these geographies every ten years. In order to make these distinctions at a more granular level in our data, we relied on the distribution of population density by ZCTA instead of the official Census Bureau definitions.

Next, we merge on data from the U.S. Census by ZIP Code and calculate population density as the total population of a given ZCTA divided by the land area of the ZCTA. Given the geographic approximation, some U.S. ZIP Codes are not represented in the U.S. Census Bureau ZCTA data; therefore, when analyzing data at the level of "rural", "suburban", or "urban" classification, the data are limited to those areas for which the U.S. Census Bureau maintains ZCTAs. Only 1,962 of the US State records from the Yelp Business dataset do not match to any Census ZCTA data. While ideally these discrepancies would be resolved using Geographic Information System software such as ArcGIS to create a crosswalk of ZIP Code and ZCTA based on the physical space each occupies, such software is outside the purview of this project. Moreover, the literature indicates that it is not uncommon for researchers to handle this data limitation in a similar manner to how we have done in this analysis.[8] Hereinafter, we will consider ZCTA synonymous with ZIP Code for the purposes of this summary.

Additionally, the Census Bureau dataset on overall businesses in a given ZIP Code requires some cleaning to merge onto the Yelp business data. Specifically, the field containing the ZIP Code name needs to be limited to the five-digit postal code using string parsing techniques. Ultimately, the merge results in only 562 of the U.S.-limited Yelp business records having no Census Bureau data on business establishments. Records without Census Bureau data are not explicitly excluded from our ultimate dataset, but the values for the corresponding Census records are left missing.

Next, the Yelp field "categories," which classifies businesses according to the services they provide, is too specific in its native form to be useful for comparing businesses by the services provided. Specifically, there are roughly 88,000 categories in the Yelp data due to redundancies in combinations of native categories. This column can be reduced to have 1,305 unique categories for [188,593] businesses. Therefore, a large aspect of the data preparation requires cleaning these classifications into broader categories. Ultimately, we identify 11 generalized types of businesses and used string matching techniques to re-categorize each business into these new groupings by identifying keywords in the native "categories" field. These include: automotive, bars/food/coffee shops, business/government services, education, entertainment, home/personal services, medical services, shopping, sports/fitness, travel, and other. The last category, other, is a catch-all classification for businesses we were unable to identify a clear provided service that could be grouped with other Yelp native categories.

Lastly, focusing our analysis on a specific city requires additional cleaning of the "city" field native to the Yelp business data. The Las Vegas region is commonly referred to as Las Vegas, but is comprised of three cities: Las Vegas, North Las Vegas, and Henderson. We found that the field could be

[8] See, for example, Lopez P. Russ, "Neighborhood Risk Factors for Obesity," *Obesity*, Vol. 15 (8), September 2012, available at https://onlinelibrary.wiley.com/doi/abs/10.1038/oby.2007.251.

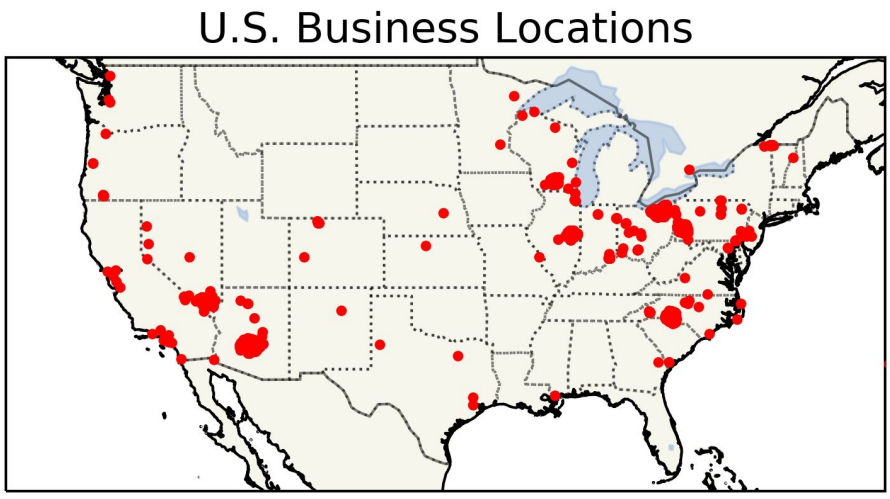
reported with spelling errors or extra whitespace characters. For example, conducting a count of values for the city field where the string "vegas" appears, we find the city of Las Vegas, Nevada can appear as 'las vegas', 'Vegas', 'Las Vegass' or 'las Vegas'. To correct these disparities we employ regular expressions and a manual review process to standardize the city names for our three focus cities Las Vegas, North Las Vegas, and Henderson, that comprise the Las Vegas metropolitan area.

As described in more detail below, our dataset has additional limitations which are not particular to a given field, but require additional filters to draw meaningful conclusions. We details these additional limitations and the specific filters we apply to address them in the next section.

V. Exploratory Analyses

Data coverage - Is the sample representative of what we would expect the population of businesses to be?

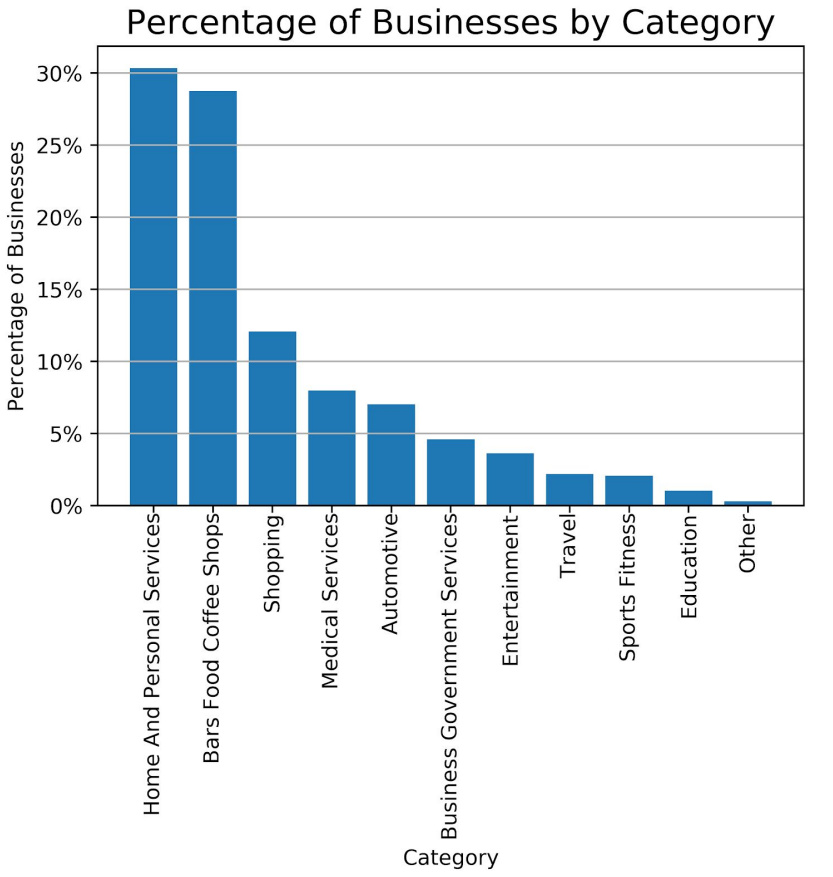
A major question to examine when analyzing data of this sort is whether the data sample is representative of what one would expect of the population. Yelp's documentation does not indicate that their dataset is limited to a specific region, business type, or location type. However, as discussed, in Section IV, the data appear to be predominantly focused on U.S. businesses. Therefore, we proceed to examine whether — within the U.S. — the Yelp business data are representative of what one would expect of the population of U.S. businesses.



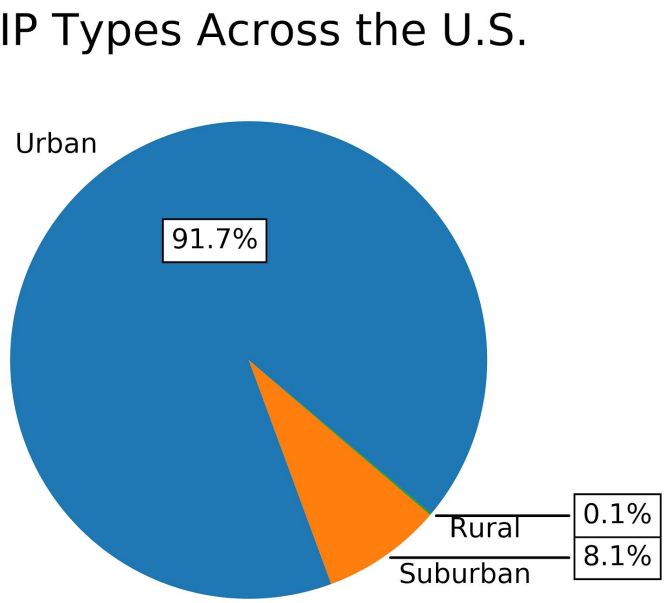
First, we examine whether the data cover the U.S. fully, or are limited to certain regions. As demonstrated in the map above, the data do not have wide geographic coverage.[9] Many states are not in

[9] Note, Alaska and Hawaii are excluded from this map, but, as demonstrated below, the dataset contains no records for these states.

the data at all, and businesses in the center of the country are clearly under-represented. Other states have a paucity of businesses in the Yelp data. This lack of coverage suggests that additional data filters are appropriate before making conclusions about patterns in businesses success or viability relative to the Census data we have collected.



Next, we consider whether the relative number of businesses by type fit our expectations in the Yelp data. The above chart demonstrates that the Yelp data is dominated by the Home and Personal Services and Bars/Food/Coffee Shop categories. While this may not be representative of the relative categorical composition of the population of U.S. Businesses, it does seem to be consistent with what would expect of business review data of the sort that Yelp aggregates. That is, one would anticipate there to be more people using and/or reviewing services in the relative order they appear in this chart, especially given the breadth of services that can be classified under Home and Personal Services.



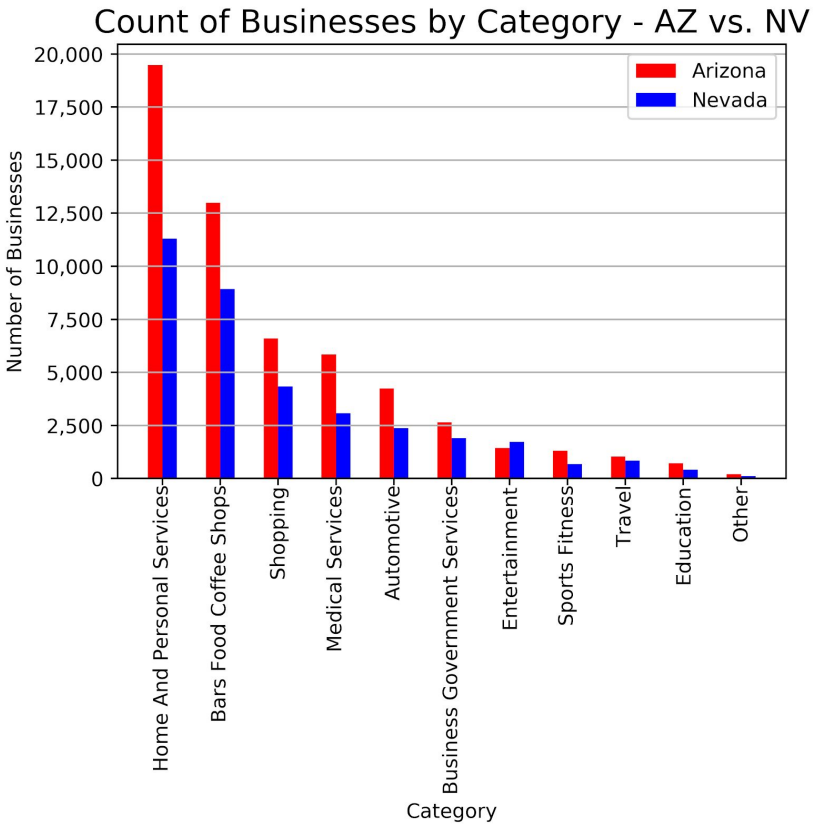
Lastly, we examine whether there are disparities in terms of location type (*i.e.*, our urban-suburban-rural distinction). One would expect the data to primarily be composed of businesses in urban areas, followed by suburban and rural areas, respectively given the relative population densities of these area types. The data indicate the expected pattern persists in the Yelp sample. However, it is apparent that businesses in suburban and rural areas are underrepresented in the sample given these areas' relative share of businesses. For instance, the across all ZIP Codes in the Census Bureau's population data, 61% of the population lives in areas that we classify as urban, compared to 37% living in suburban areas, and 3% living in rural areas. While it might be incorrect to assume the ratio of businesses by type of area would mirror the ratio of people, the disparity presented in the Yelp data suggests its sample is biased in favor of covering urban regions.

Data coverage - Is there a region suitable for analyzing factors of business success?

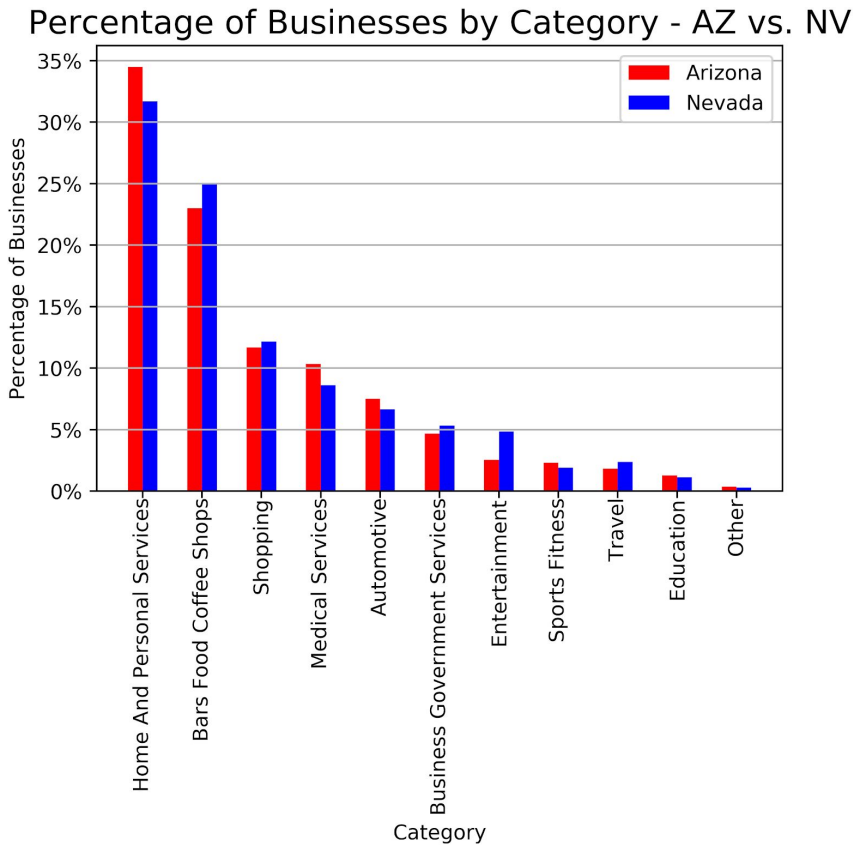
Given the deficiencies of the geographic distribution of businesses on a national level, we consider the viability of an analysis at the state-level.

	Number of Businesses
AZ	56495
NV	35688
NC	14359
OH	13664
PA	10966
WI	5042
IL	1937
SC	770
IN	101
OR	72
CO	43
NY	19
CA	8
AR	2
MO	2
WA	2
VT	2
AL	2
MA	1
MN	1
GA	1
FL	1
DE	1
MT	1
NE	1
VA	1

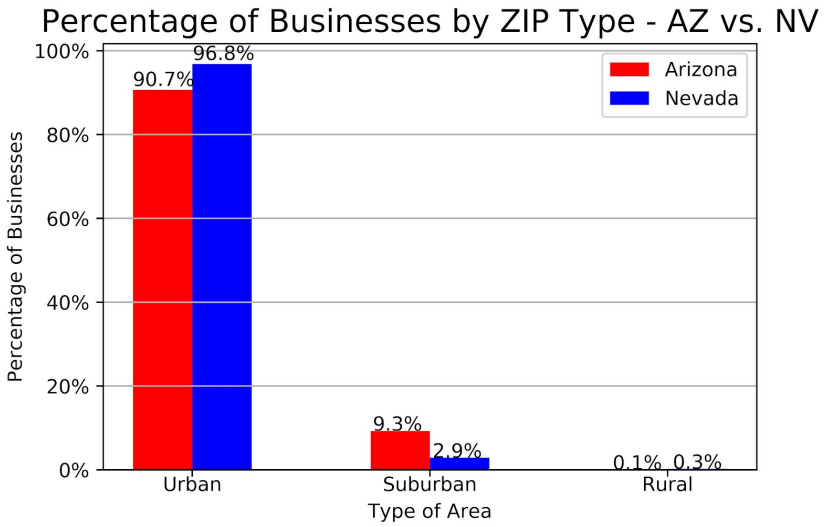
The table above presents a tabulation of business counts by state. It demonstrates that Arizona and Nevada have the largest concentration of businesses in the Yelp data, with more than double the next state's count of businesses. Indeed, most states have fewer than 100 businesses in total. Therefore, we proceed by analyzing the types of businesses and areas covered within the top two states in order to ultimately select one to focus our analysis.



The chart above demonstrates that both states cover the same categories of businesses and every category of business in the dataset is represented in each of these states. With no category clearly under-represented in these states, we require a view of the relative share of businesses by category in each state to determine if one state's data overweights or underweights certain categories.



The chart above examines this issue by presenting the percentages of businesses in each category for Arizona and Nevada. The chart demonstrates that the two states are fairly similar in the relative composition of business categories (as well as being very similar to the breakdown of categories at the national level). Home and Personal Services as well as Bars/Food/Coffee Shops are most abundant in the data, while Education and Travel businesses appear least frequently (besides those grouped under Other). These findings are in line with expectations. One would anticipate that there are few businesses that could be classified under Education or Travel (*e.g.*, schools and tutoring services, or airports and bus stations), while one would also expect the most frequently occuring categories to consist of businesses that reviewers would be more likely to add to Yelp for reviewing. Surprisingly, the Entertainment category has similar representation between the two states. Given the prominence of the casino, gaming, and sports-betting industry in Nevada, this suggests Yelp's sample is under-representative of such businesses. However, since these industries are banned in most states, this underrepresentation likely improves a researcher's ability to generalize any findings from analyzing Nevada to a broader context.



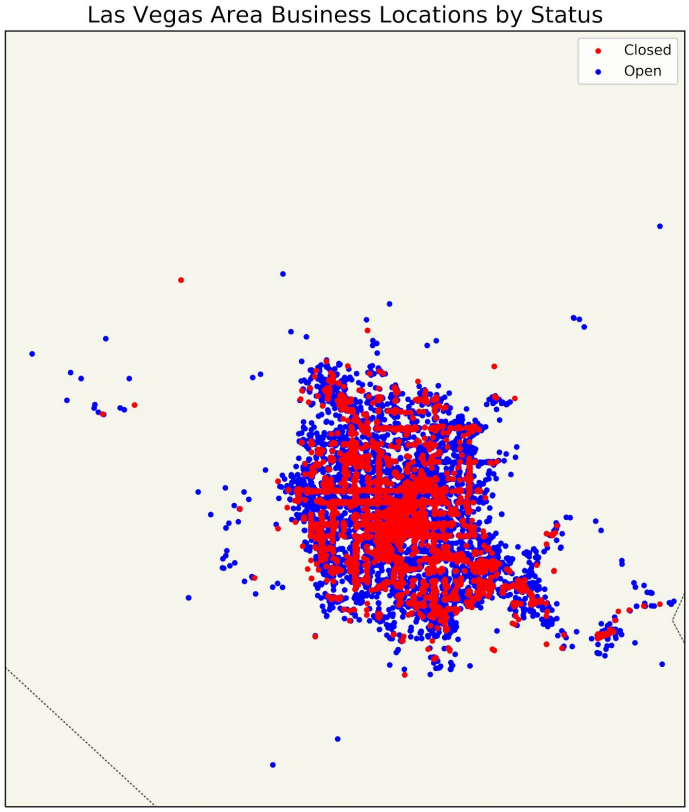
One final comparison between the two states that our data allow is the relative representation of businesses by urban, suburban, or rural classification. The chart above demonstrates that, not surprisingly, businesses in both states are primarily in more densely populated, urban areas. This finding is consistent with the geographic distribution of businesses in these two states being clustered around the cities of Las Vegas and Phoenix as can be gleaned from the map above. This view also shows that while Arizona has more businesses in suburban areas (and in a similar percentage as in the nation-wide dataset), businesses in rural areas in both states are negligible enough where it would be difficult to draw conclusions about differences between businesses in rural vs. urban or suburban areas. Therefore, we by-and-large abandon analyzing the data from this perspective.

While Arizona has the larger number of businesses, we conclude that the composition of business types is comparable within the two states. Given Prof. Laskowski's outstanding lecture "at" Circus Circus, we were inspired us to take the project Vegas-bound. Thus, we focus our core analyses below on Nevada, and, in particular, Las Vegas.

Data coverage - Are there enough closed businesses in Las Vegas for a meaningful comparative analysis?

Given the nature of the Yelp data as a database of crowd-sourced reviews and information on businesses, it is not unlikely that shuttered businesses may be underrepresented in the dataset. This bias can be attributed to the fact that shuttered businesses would generally not be reviewed after closing and those that closed before being reviewed would not appear in the data. However, it is Yelp's policy not to delete business pages; therefore, any underrepresentation of closed businesses should not be exacerbated

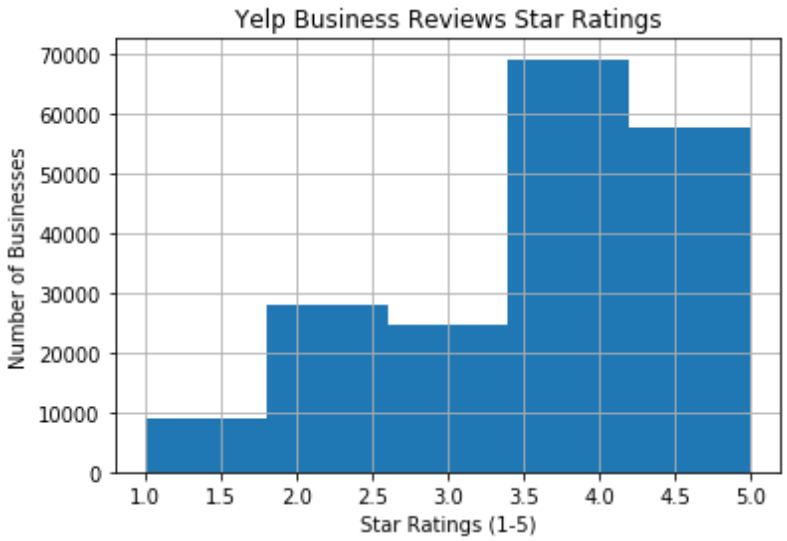
by any culling of data on these locations.[10] Indeed, there are 6,410 closed businesses compared to 28,954 open businesses in Las Vegas, Nevada. This, coupled with the map below showing the plots of closed businesses overlaid on top of open businesses in a fairly uniform fashion all over the city suggests that while closed businesses may be underrepresented, the deficiency should not prohibit researchers from drawing meaningful conclusions about the differences between operational and shuttered businesses.



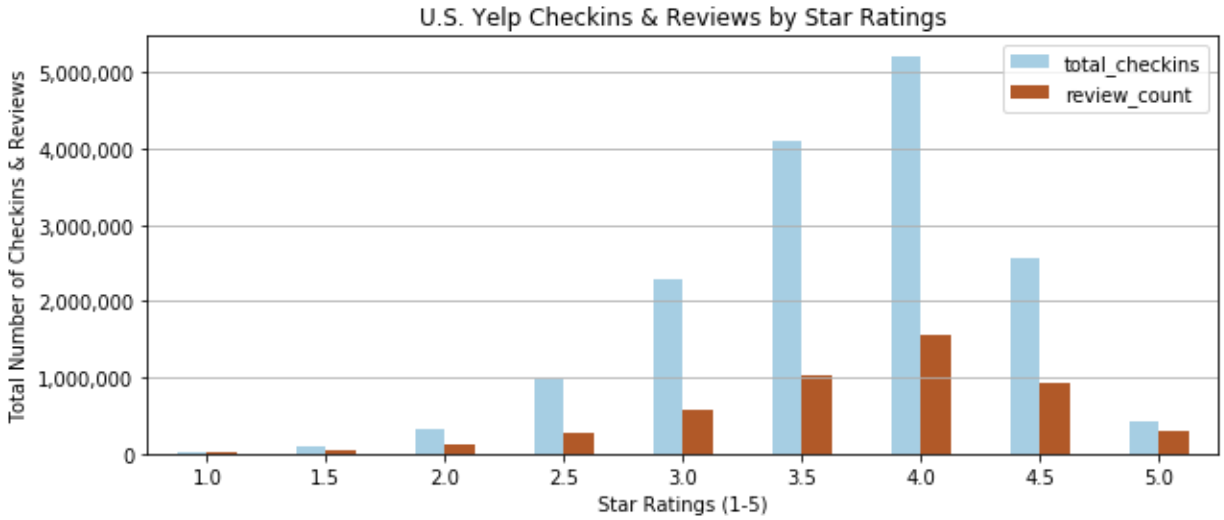
Understanding Yelp businesses - what is the distribution of ratings?

We started out by seeking to understand the business data from the perspective of star ratings given in user reviews. First, we explored overall if the star ratings given are normally distributed. From our collective personal experience, we hypothesized the distribution might be U-shaped, as we imagined that users would provide stars only if the businesses were outstanding or severely underperforming. To our surprise, we learned that Yelp reviewers have a strong bias towards positive reviews as can be seen in the below figure. The mean rating across all businesses was 3.63.

[10] See, "Why do businesses that are permanently closed sometimes still appear in Yelp search results?", Yelp.com, accessed December 10, 2018, available at https://www.yelp-support.com/article/Why-do-businesses-that-are-permanently-closed-sometimes-still-appear-in-Yelp-search-results?l=en_US.



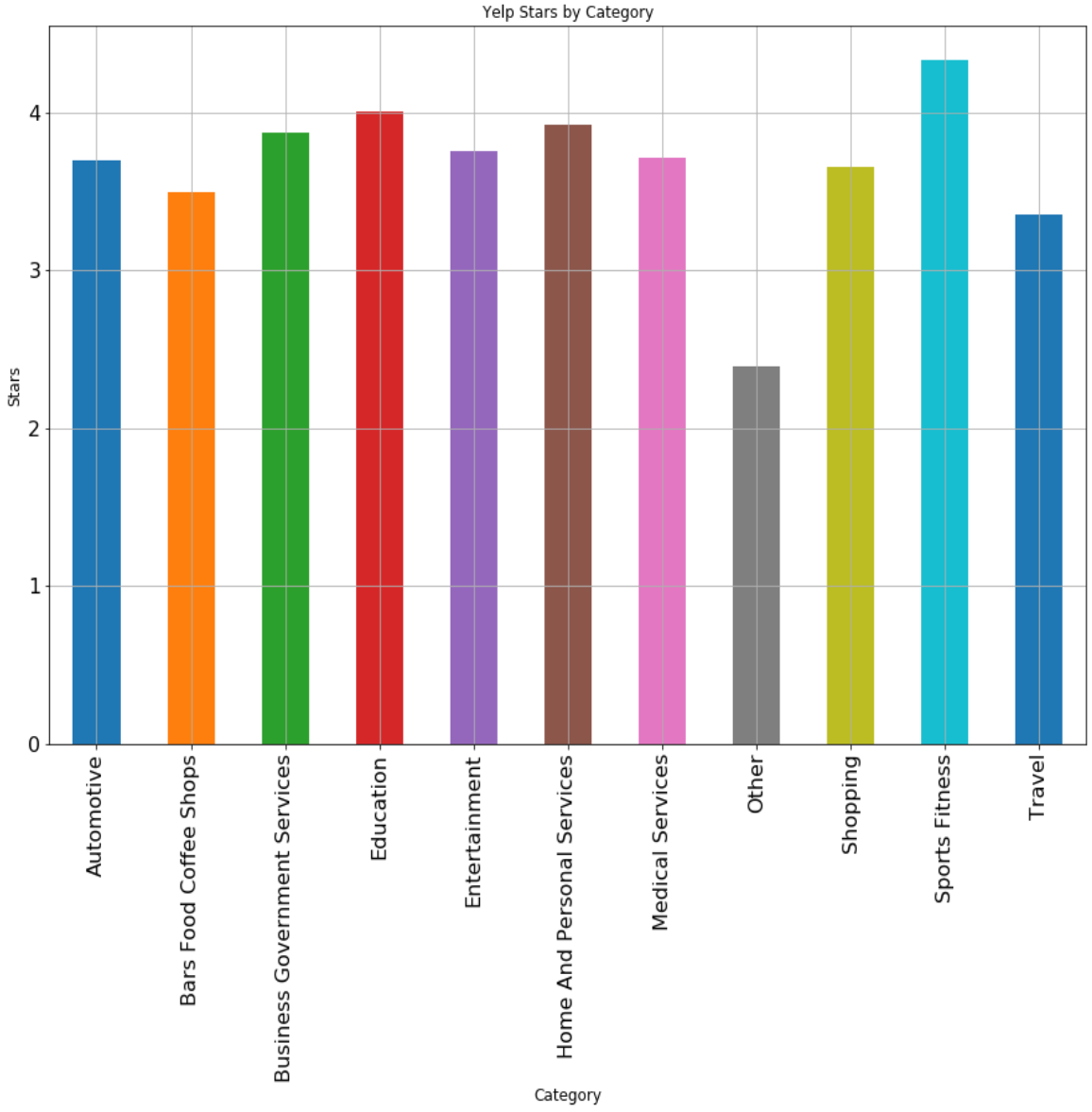
Understanding Yelp users - what is the distribution of ratings by check-ins and reviews?



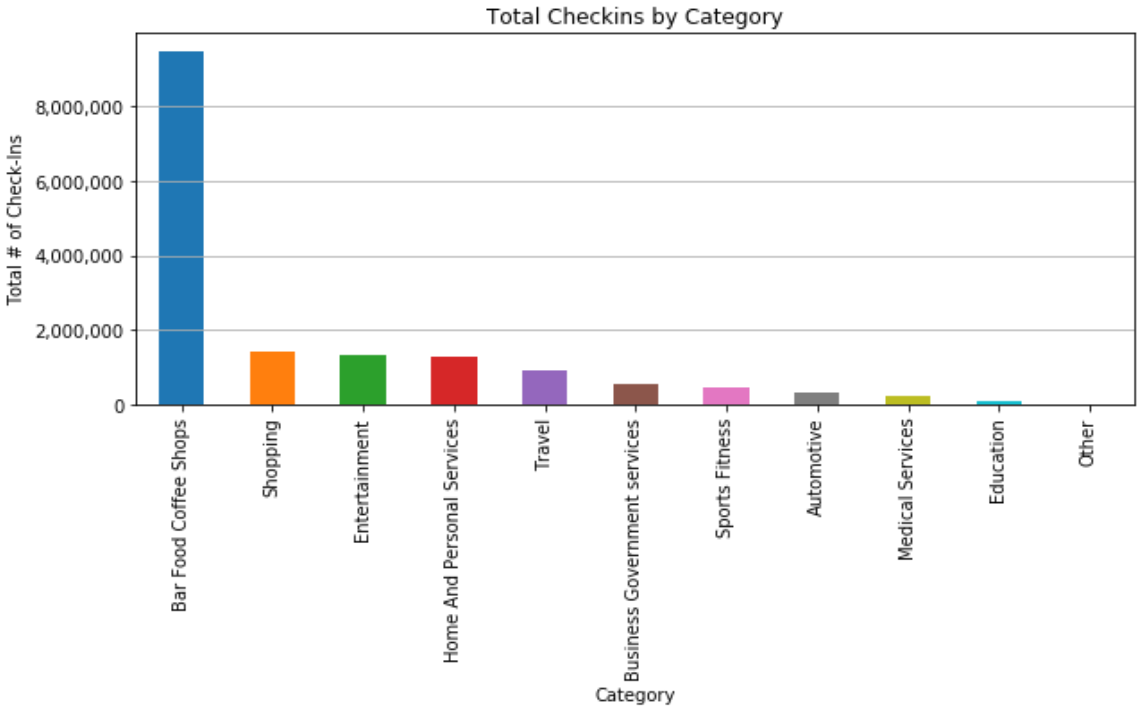
The chart above compares total number of user's check-ins to businesses with user reviews of businesses. Compared to the histogram of Yelp ratings distribution, we see they trend in similar ways. The majority of Yelp users review restaurants within the 3.5 - 4.5 stars range, and these ranges are also the restaurants most frequented by Yelp users. It is also evident that Yelp users are far more likely to check-in to businesses than to write a review, likely due to a lower effort on the part of the user. And though we do not know the percent overlap between users who review and users who check-in to restaurants, we can reasonably assume that the large population presented by the Yelp dataset gives us a good idea of general user behavior.

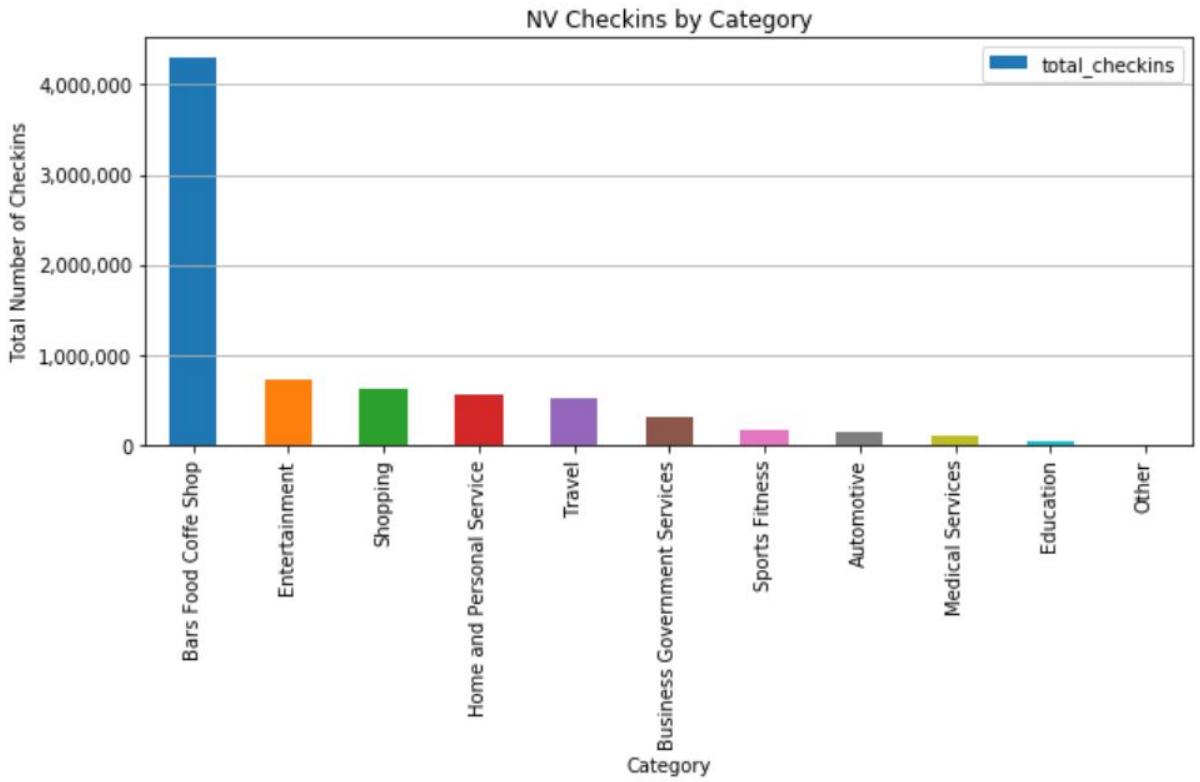
Understanding user behavior through categories - what is the distribution across ratings?

The chart below examines the star ratings given across the 11 categories of businesses we identified. From this, we find that the Sports and Fitness category receives the highest star ratings while the lowest star ratings are given to the Travel category and the Bars/Food/Coffee Shops category being second from the bottom. However, from a number of reviews perspective in Nevada, reviewers were much more likely to rate establishments in the Bars/Food/Coffee Shops category (25% of all reviews) even though they rated them 2nd from last.



What is the distribution of categories by check-ins, both nationally and in Nevada specifically?





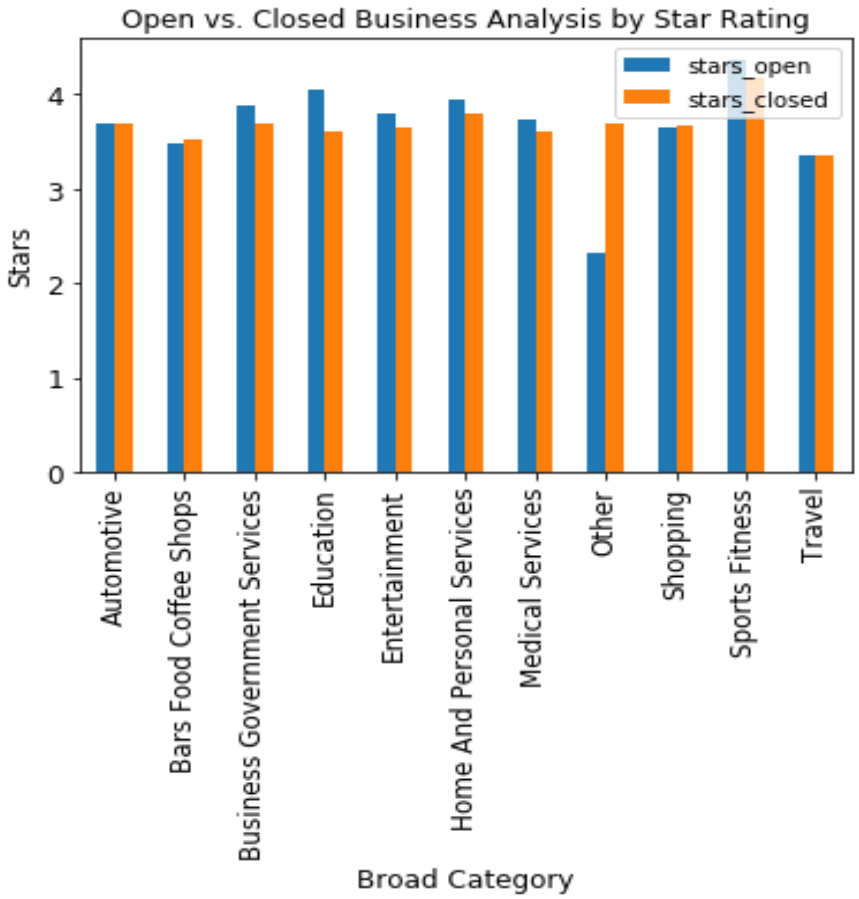
The charts above depict number of check-ins by categories on a national level and in Nevada specifically. Despite having a lower than average star rating of 3.48, the Bars/Food/Coffee Shops category is by far the most frequented by Yelp users. And despite having the highest star rating of 4.1, Sports and Fitness category is not one of the top 5 categories of businesses that Yelp users check-in to. Even when we focus our data just on Nevada, the same trend holds true. This could suggest that as Yelp users frequent restaurants more, they may be more critical of the service that they receive. Alternatively, it could suggest that Yelp users are simply responding to the environment around them. And unsurprisingly, in Nevada, users checked-in to entertainment businesses more often users on a national level, given the amount of entertainment businesses centralized in Nevada.

VI. Analysis: Do Reviews Matter?

From a comparison of values, it would appear that stars have some form of impact on businesses' open-status. If we compare the star ratings of businesses that are closed to those that are open, we see that the star ratings of closed businesses are marginally lower than those of open businesses.

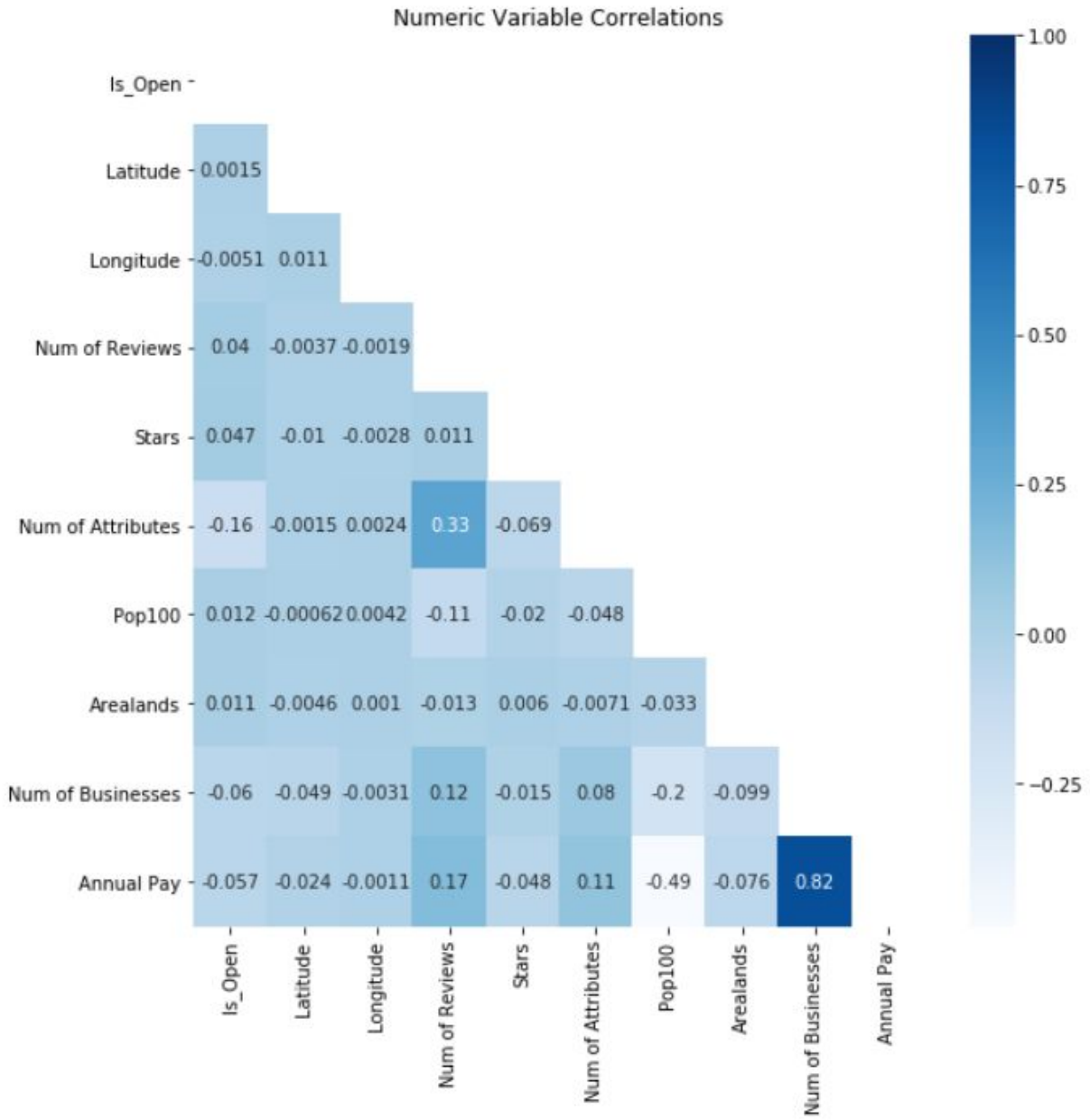
Business Still Open	Star Rating
No	3.514756
Yes	3.655406

In the following barplot, the marginal difference in star-ratings of closed vs. open businesses is generally true across the categories of business as well. While there is a sharper contrast in the category 'Other', this category captures the businesses where the category information was unknown so it is premature to draw any conclusions about this difference in ratings without further inquiry into those businesses.



To further investigate the relationship between the available data's potential independent variables against the dependent variable of a business' open status, we create a correlation plot which denotes the strength of correlation between two variables. In the correlation plot below, we see relatively weak correlation amongst most of the predictors. Of particular relevance to the previous insights, we see that star rating and whether or not a business is open is weakly correlated (0.047). While these results are slightly counterintuitive, the data shows (in both the bar plots and correlation plot) that star rating is at

least weakly correlated with the open-status of a business. However, given the intuition behind the purpose of stars, further analysis must be conducted before ruling out that the number of a stars a Yelp business has is a weak predictor. For example, it may be that there is an interaction between stars and another variable that has a significant impact on a whether or not a business will stay opened or close.



Which Variables are Statistically Significant?

To address this question, we decided to use logistic regression classification models to determine statistical significance. Given that the dependent variable is binary, it would have been inappropriate to use Ordinary Least Squares Regression. While there are basic assumptions about homoscedasticity, independence and normality that are necessarily to determine the validity and applicability of using such a model for predictive analysis, we simply wanted to gain insights to predictor importance, but not make any strong affirmative statements about the relationships of these variables. As such, the first approach was to create a logistic regression model containing all the information available. Categorical variables were transformed into binary dummy variables. The following table is a sample and portion of the initial results. Given the amount cities, neighborhoods, ZIP Code types (rural, suburban, urban), and broad categories, there was a large amount of variation. This complicated the interpretability of the results, which ended up detracting from modeling the overall effect and gleaning any meaningful insights.

Dep. Variable:	is_open	**No. Observations:**	34791
Model:	Logit	**Df Residuals:**	34750
Method:	MLE	**Df Model:**	40
Date:	Mon, 10 Dec 2018	**Pseudo R-squ.:**	0.08857
Time:	17:23:11	**Log-Likelihood:**	-14990.
converged:	False	**LL-Null:**	-16447.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
latitude	-0.0678	0.187	-0.362	0.717	-0.435	0.299
longitude	-0.0044	0.007	-0.676	0.499	-0.017	0.008
review_count	0.0035	0.000	17.211	0.000	0.003	0.004
stars	0.0186	0.015	1.235	0.217	-0.011	0.048
pop100	2.288e-07	1.36e-06	0.168	0.866	-2.43e-06	2.89e-06
arealand	-1.479e-10	2.48e-10	-0.598	0.550	-6.33e-10	3.37e-10
ESTAB	-0.0001	5.8e-05	-2.318	0.020	-0.000	-2.07e-05
PAYANN	1.752e-08	2.83e-08	0.619	0.536	-3.8e-08	7.3e-08
attribute_len	0.0030	0.003	0.943	0.346	-0.003	0.009
city_Las Vegas	0.0096	0.064	0.151	0.880	-0.115	0.135
city_North Las Vegas	0.2304	0.099	2.317	0.021	0.035	0.425
neighborhood_	4.8499	6.862	0.707	0.480	-8.600	18.300
neighborhood_Anthem	4.5935	6.856	0.670	0.503	-8.843	18.030
neighborhood_Centennial	4.9645	6.888	0.721	0.471	-8.536	18.465
neighborhood_Chinatown	4.4023	6.862	0.642	0.521	-9.047	17.852

The next iteration of the logistic model was then limited to non-binary information likely tied more directly to the businesses. This version of the model only looked at the review counts (defined as the number of reviews for the business), star rating (the over average amount of stars for the business) and attribute length (the number of attributes, such as bike parking and dog-friendliness, associated with the business). These results, which are printed below, were much clearer, and we saw high statistical significance for all three fields. With all p-values at 0.00, which is less than an alpha set at 0.05 or 0.01 (95% and 99% confidence), we can reject the null hypothesis (no relationship between the predictor and outcome variable) for each of these predictors and conclude that the number of reviews, average star rating, and number of attributes are statistically significant. Again, given the need for more in-depth analysis to determine if this type of model is appropriate for the data, we do not propose to interpret these coefficients strictly, but instead consider their sign and relative magnitudes as an indication of what factors might impact a business' operating status and in what way more generally. These results suggest that a business with a greater number of reviews and higher star ratings is more likely to be operational in the Yelp data.

Logit Regression Results

Dep. Variable:	is_open	**No. Observations:**	35364
Model:	Logit	**Df Residuals:**	35361
Method:	MLE	**Df Model:**	2
Date:	Mon, 10 Dec 2018	**Pseudo R-squ.:**	0.01412
Time:	22:20:35	**Log-Likelihood:**	-16501.
converged:	True	**LL-Null:**	-16738.
		LLR p-value:	2.329e-103

	coef	std err	z	P>|z|	[0.025	0.975]
review_count	0.0034	0.000	16.720	0.000	0.003	0.004
stars	0.4541	0.005	87.615	0.000	0.444	0.464
attribute_len	-0.0551	0.002	-28.064	0.000	-0.059	-0.051

Lastly, to examine whether data external to the Yelp business dataset might be relevant to businesses' operating status, we conducted another iteration of a logistic model that includes Census data. These results are shown below. Pop100 represents the total population in a Las Vegas ZIP Codes ESTAB represents the number of establishments in the ZIP Code, and PAYANN represents the total annual payroll of those establishments. Given the magnitude of these fields, we performed log transformations on the Census data. The results of the following logistic regression suggest all the predictors, from the Yelp and Census data, are highly significant. Since the p-values for all predictors are 0.000 and less than

alpha 0.05 and 0.01 (95% and 99% confidence), we reject the null hypothesis that these fields do not affect a business' open vs. closed-down status, and conclude these fields likely have a statistically significant relationship with the outcome variable. Moreover, the relative signs and magnitudes of these controls tell an interesting story. They suggest that while review counts and star ratings may improve the likelihood that a business in operational in Las Vegas, factors such as population and workers' wages in the ZIP Codes of these businesses may substantially improve this likelihood. Moreover, the more complex a business' offerings (*i.e.*, the longer the attribute list) and the greater the overall number of businesses in the same ZIP Code, the less likely a business will be operational. Notably, however, these regressions have a low R-squared suggesting the models do not fit the data very well.

Logit Regression Results

Dep. Variable:	is_open	**No. Observations:**	34793
Model:	Logit	**Df Residuals:**	34787
Method:	MLE	**Df Model:**	5
Date:	Mon, 10 Dec 2018	**Pseudo R-squ.:**	0.04734
Time:	22:52:34	**Log-Likelihood:**	-15668.
converged:	True	**LL-Null:**	-16447.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
review_count	0.0040	0.000	18.978	0.000	0.004	0.004
stars	0.0800	0.014	5.789	0.000	0.053	0.107
attribute_len	-0.0731	0.002	-34.859	0.000	-0.077	-0.069
log_pop100	0.2457	0.012	19.727	0.000	0.221	0.270
log_ESTAB	-0.5885	0.053	-11.151	0.000	-0.692	-0.485
log_PAYANN	0.2318	0.026	8.943	0.000	0.181	0.283

VII. Conclusions

The Yelp Data Challenge data contain a bevy of information of potential interest to researchers. We have demonstrated, however, that the business dataset is limited by poor geographic coverage and an urban-business bias that restricts researchers from determining broad patterns from supplemental data such as population and business density at the ZIP Code-level. Despite these limitations, our research has found that some potentially meaningful conclusions can be drawn by focusing on areas with the best coverage in the Yelp data. These conclusions come with the caveat that the results may differ geographically, especially outside of urban areas.

Specifically, we find that the overall distribution of star ratings and check-ins are both skewed towards more positive values. Additionally, the distribution of these metrics are not uniform among categories of businesses. That is, reviewers are more inclined to give positive reviews to certain types of businesses such as sports and fitness than others. These observations hold both in terms of the overall U.S. coverage of the Yelp data, and in Nevada where the data have particularly good coverage.

To examine the relevance of Yelp reviews and its impact on businesses, we first view the distribution for the number of stars for open vs. closed businesses. Considering the average star rating as well as the visual trend within these bar plots, it appears that there is only a marginal difference in stars via the opened and closed businesses. The correlation plot we created confirms this, where the correlation between the two fields is weak at 0.047.

To make more scientifically sound conclusions, we conduct multiple iterations of logistic classification models to draw preliminary insights about statistical significance. We acknowledge many of the assumptions to implement logistic regression may not be met, and thus further pursuing logistic regression will require more analyses and transformations to ensure proper usage. However, from careful comparison between model iterations and methodical variable testing, we are able to demonstrate at both 95% and 99% confidence levels that the number of reviews and average star rating likely have a positive impact on Las Vegas businesses' open status, while the number of attributes associated with a business likely has a negative effect on this status. Furthermore, Census data detailing the total population in the Las Vegas ZIP Codes, number of business establishments in the area, and the total annual payroll of these businesses are also statistically significant at 95% and 99% confidence levels, indicating all listed fields are likely relevant to Las Vegas businesses' operation status. Next steps entail investigating the interaction effect between the statistically significant numeric data and the removed categorical variables such as neighborhood and the featured engineered business categories. Additionally, researching and employing non-parametric machine learning algorithms such as K nearest neighbor or boosted tree algorithms may better capture and predict the behavior and results of whether or not a Las Vegas Yelp business will remain open or close down.