The Dining Experience – San Francisco and Los Angeles

Ambar Mukherjea

5/31/2019

Introduction

Los Angeles and San Francisco are two of the most significant metropolitan cities in the state of California. The two cities cater to the extensive needs for residents, tourists, business travelers, and entrepreneurs in one of the most thriving states in the country. With the varied culinary tastes of the cities' residents and visitors, it would be beneficial for diners and potential restauranteurs to know: what are the most enjoyable types of cuisines in each city? What factors contribute to the appeal of each of the cities' restaurants? Does Los Angeles share similar characteristics to San Francisco in culinary tastes? These components are essential for a great dining experience, and for potential restaurant owners to know where to open new venues so their ventures are lucrative.

Data

Sources

The restaurant venue data is sourced from the Foursquare API. Location data that will be used to identify food category venues, in datasets for San Francisco and Los Angeles are:

Variable	Description
name	Venue name
type	Search criteria that identifies venue cuisine type
Id	Venue's unique Foursquare identifier
city	City of venue's location
latitude	Venue's latitude coordinate
longitude	Venue's longitude coordinate
zipcode	Venue's zip code
likes	Total number of like recorded by users for the venue
rating	Venue's rating (graded on a floating scale from 0 to 10)
price_tier	Venue's price tier (1 = least pricey to 4 = most pricey)

Utilizing the Foursquare API calls, data is collected in a structured, tabular form, maintaining a maximum of 50 records for each search criteria term for the cuisine type. Distinct search criteria are derived from the popular ethnic cuisines documented on yelp.com for the two cities.

Data Cleaning

Using Foursquare's *near* operator, the above components are collected not only for the cities of interest (San Francisco, Los Angeles), but also include data points for nearby cities (ex: Oakland). Because it is intended to show the appeal of restaurants exclusively in the two cities, records with data from cities occurring outside of San Francisco and Los Angeles are dropped. Records that contain missing values that occur in key variables zip codes, ratings, and price tiers are dropped as well. Imputing such values presents challenges due to the ambiguity of data profile in each city. The categorical variable for venue cuisine type is encoded numerically with a new variable, type_cat, for appropriate modeling and analysis.

Feature Selection

After data is sourced and cleaned for each city, the following features are used for Model Evaluation and Machine Learning:

- type_cat
- zipcode
- rating
- price_tier
- likes

221 samples are present for the San Francisco dataset, while 146 samples are available for Los Angeles.

Methodology

To further examine how a venue's reputation can be shaped by its feature, San Francisco and Los Angeles data sets are analyzed through multiple models to identify the best relationship. The models are designed as the following:

1. Correlation of Variables to seek any strong relationships with the # of Likes for each venue.
2. Simple Linear Regression Modeling and Evaluation to examine how # of Likes is influenced by cuisine type.
3. Multiple Linear Regression Modeling and Evaluation to examine how # of Likes is influenced by cuisine type, zip code, and price tier.
4. 4^{th}-order Multivariate Polynomial Regression Modeling and Evaluation to examine how # of Likes is influenced by cuisine type, zip code, and price tier.
5. Decision tree classification and evaluation to predict the # of Likes from cuisine type, zip code, and price tier.
6. Grouped aggregate metrics of # of Likes and price tier and by Cuisine types to identify trends.

Foursquare uses the # of Likes, for a venue, as an input to determine its rating. Hence, the rating has an inherent positive correlation with the # of Likes. Rating is removed as a variable for modeling.

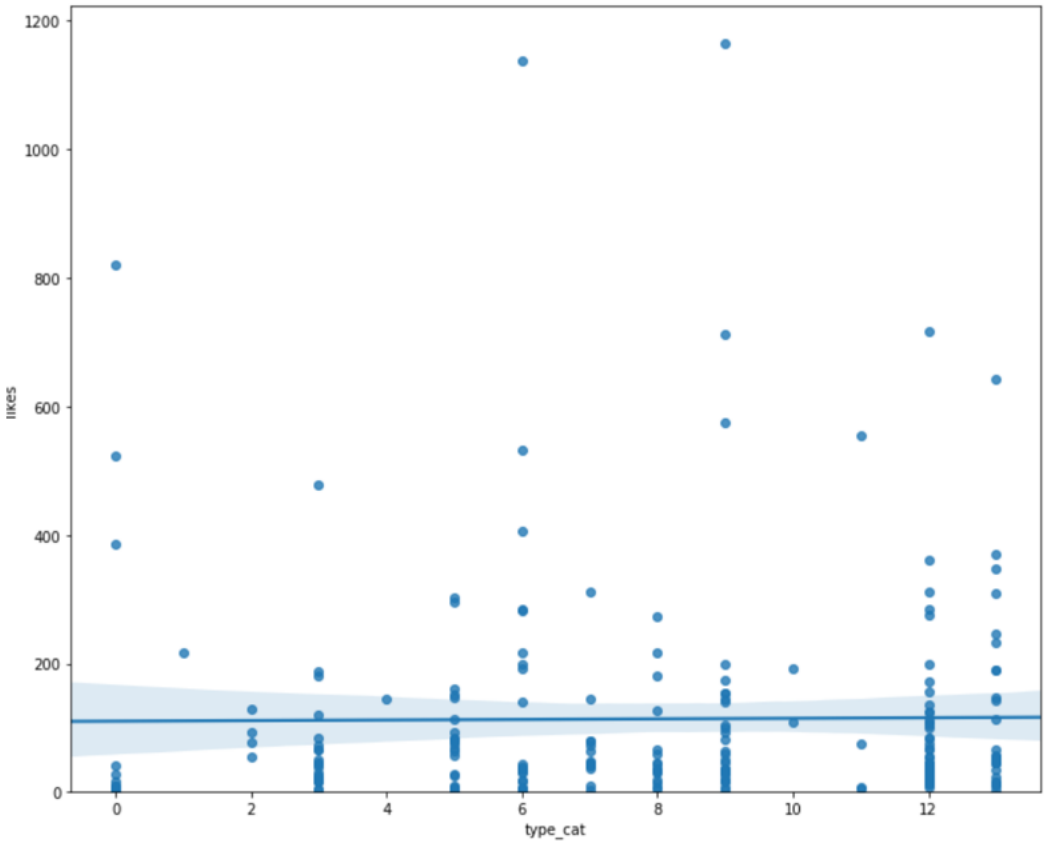
Results

Below is the analysis of the models applied to the datasets from San Francisco and Los Angeles venues.

Figure 1. Correlation of Feature Variables – San Francisco Venues

	type_cat	zipcode	rating	price_tier	likes
type_cat	1.000000	-0.207785	0.113514	-0.218993	0.010402
zipcode	-0.207785	1.000000	-0.009801	0.105350	-0.158288
rating	0.113514	-0.009801	1.000000	0.175463	0.524607
price_tier	-0.218993	0.105350	0.175463	1.000000	0.109196
likes	0.010402	-0.158288	0.524607	0.109196	1.000000

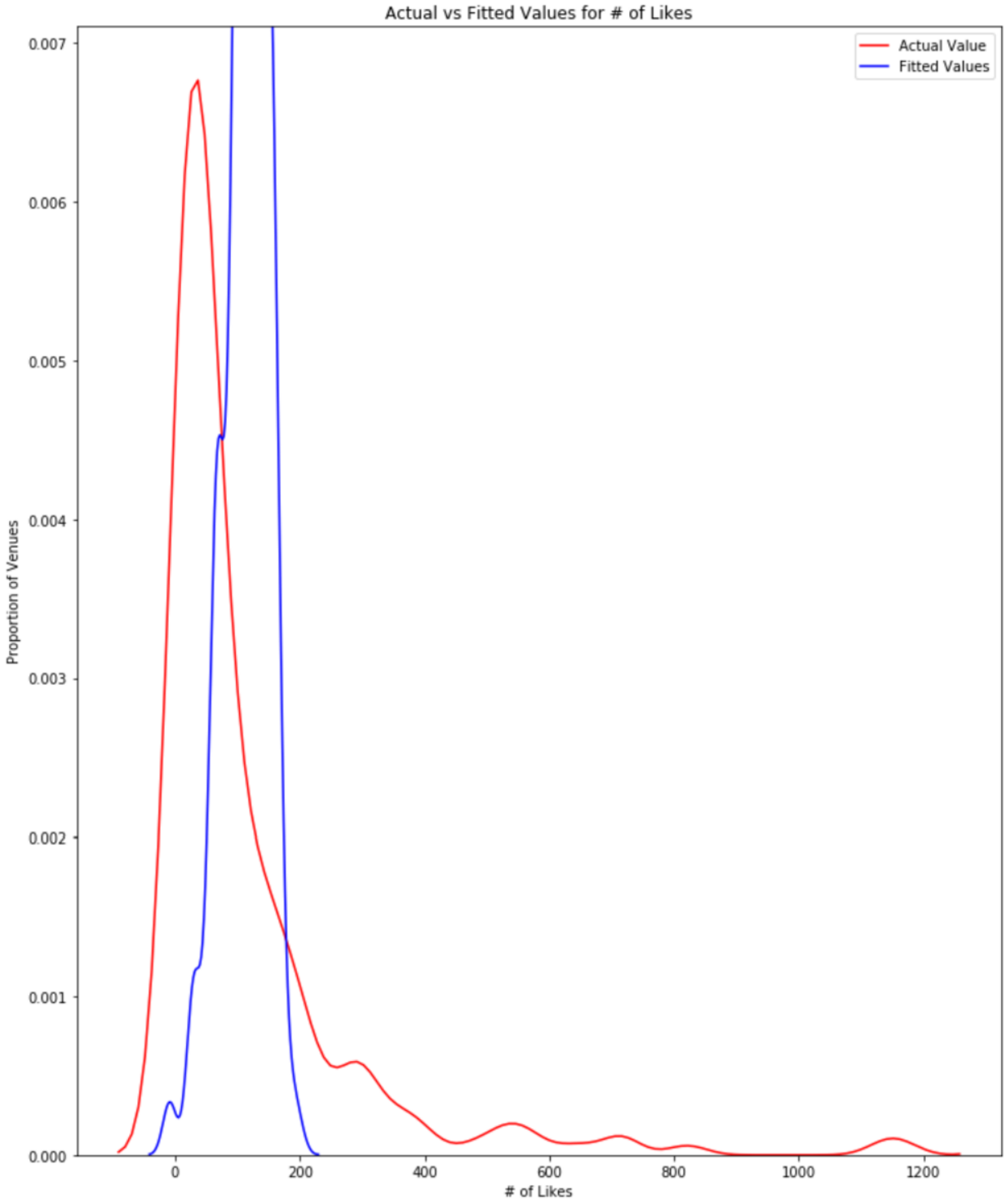
Figure 2. Simple Linear Regression Model Evaluation – San Francisco Venues



Model Setup
Predictor Variable: type_cat
Response Variable: likes

Model Evaluation
R-square: 0.000108211807773
MSE: 28670.8825173

Figure 3. Multiple Linear Regression Model Evaluation – San Francisco Venues



Model Setup

Predictor Variables: type_cat, price_tier, zipcode

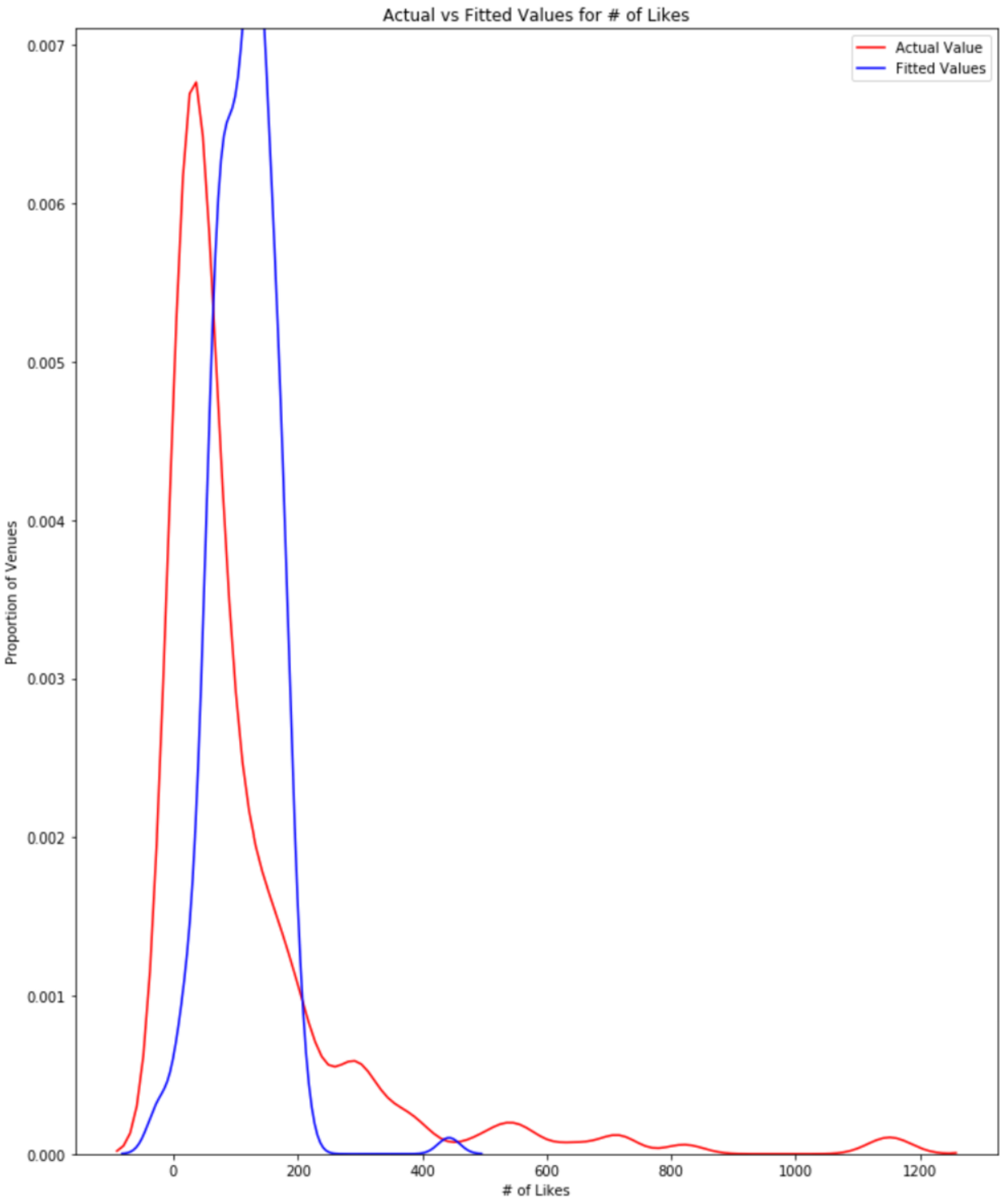
Response Variable: likes

Model Evaluation

R-square: 0.0410841174926

MSE: 27495.9399968

Figure 4. Polynomial Regression (4th order) – San Francisco Venues



Model Setup

Predictor Variables: type_cat, price_tier, zipcode

Response Variable: likes

Model Evaluation

R-square: 0.0894506597658

MSE: 26109.0784707

Figure 5. Correlation of Feature Variables – Los Angeles Venues

	type_cat	zipcode	rating	price_tier	likes
type_cat	1.000000	0.012061	0.034609	-0.003766	0.034429
zipcode	0.012061	1.000000	-0.089493	-0.154334	-0.101145
rating	0.034609	-0.089493	1.000000	0.099353	0.556762
price_tier	-0.003766	-0.154334	0.099353	1.000000	0.196802
likes	0.034429	-0.101145	0.556762	0.196802	1.000000

Figure 6. Summary of Model Evaluations for San Francisco and Los Angeles Venues

			San Francisco Venues		**Los Angeles Venues**	
Model	**Predictor Variables**	**Response Variable**	**R-square / Accuracy Score**	**MSE**	**R-square / Accuracy Score**	**MSE**
Simple Linear Regression	Venue Type Category	# of Likes	0	28670.883	0.001	3832.542
Multiple Linear Regression	Venue Type Category, Zip Code, and Price Tier	# of Likes	0.041	27495.94	0.045	3663.814
Polynomial Regression	Venue Type Category, Zip Code, and Price Tier	# of Likes	0.089	26109.078	0.266	2818.226
Decision Tree Classifier	Venue Type Category, Zip Code, and Price Tier	# of Likes	0.022	-	0	-

Figure 7. Summary of Aggregate Data by Cuisine Types for San Francisco and Los Angeles Venues

	San Francisco Venues			Los Angeles Venues		
Cuisine Type	**# of Venues**	**Average # of Likes**	**Average Price Tier**	**# of Venues**	**Average # of Likes**	**Average Price Tier**
Chinese	14	133.357	1.714	10	15	1.6
Cuban	1	218	2	6	81.833	1.833
Ethiopian	4	85.5	1.75	7	29	2
Indian	22	73.682	1.909	6	18.667	1.5
Indonesian	1	145	2	1	182	2
Italian	21	88.81	2.095	14	37.571	1.5
Japanese	19	190.789	1.947	20	53.95	1.85
Korean	14	73.357	2	15	85.2	2.067
Mediterranean	16	74.375	1.688	4	44	1.75
Mexican	29	148.724	1.31	17	39.705	1.235
Middle Eastern	2	150.5	1.5	3	5.333	2
Mongolian	0	0	0	2	7	1.5
Peruvian	4	161.25	1.75	6	32.667	1.833
Thai	46	96.891	1.826	31	48	1.71
Vietnamese	28	126.143	1.321	5	61.4	2

Discussion

It is recommended that restaurant-goers, in San Francisco and Los Angeles, do not follow the bulk of the analysis derived from Foursquare data, to determine the types of food venues they would visit. The type of cuisine, location, and price tier of each venue is shown not to be heavily correlated with, or significantly influence, how enjoyable the venue is. The recorded number of likes also shows a bias to be recorded om venues that are cheap or moderately priced. Restaurant patrons should rely on other methods to persuade their culinary experiences. Moreover, the Foursquare data is definitely not recommended for potential restauranteurs to guide them in opening a business in either of the two cities.

However, the Foursquare data can be used as a general guide for to gage the popularity of cuisine types in the two locations. San Francisco has a larger amount of restaurant data, where many cuisine types have 20 or more samples. Cuisine such as Thai, Mexican, Chinese, and Vietnamese experience a large number of likes per venue. By contrast, Los Angeles does not have as much data present, but it can be seen that Cuban and Korean restaurants are enjoyable in the city.

Conclusion

This analysis observed, through Foursquare location data, the relationship between the quantity of patrons liking a food venue and the venue's type of cuisine. This enjoyability factor was also tested against other variables, such as the venue's location within the city and its price tier. After multiple regression and machine learning models, it is definitive that the venues' number of likes is not influenced by any other variables, and diners and entrepreneurs should not use this data to lead them in enjoying or providing strong culinary experiences.

Foursquare data appears constrained, including the deprecation of calling on useful features (ex: # of Check-ins), a 50-record limit per search criteria, and outdated usage of the platform by vendors and customers. Strengthening the analysis, to better determine the venue's responsiveness from key attributes, would require a greater number of data sources. For example, a larger database of venues would be beneficial, so more sampled restaurants could be included in the models. Additional structured data would aide in this analysis as well: location features such as population density, median age, median income, and cultural diversity would result in a more pronounced model. Unstructured data, such as venues' menus, tips, and recommendations, would further bolster the predictability of the venues' appeal.