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.

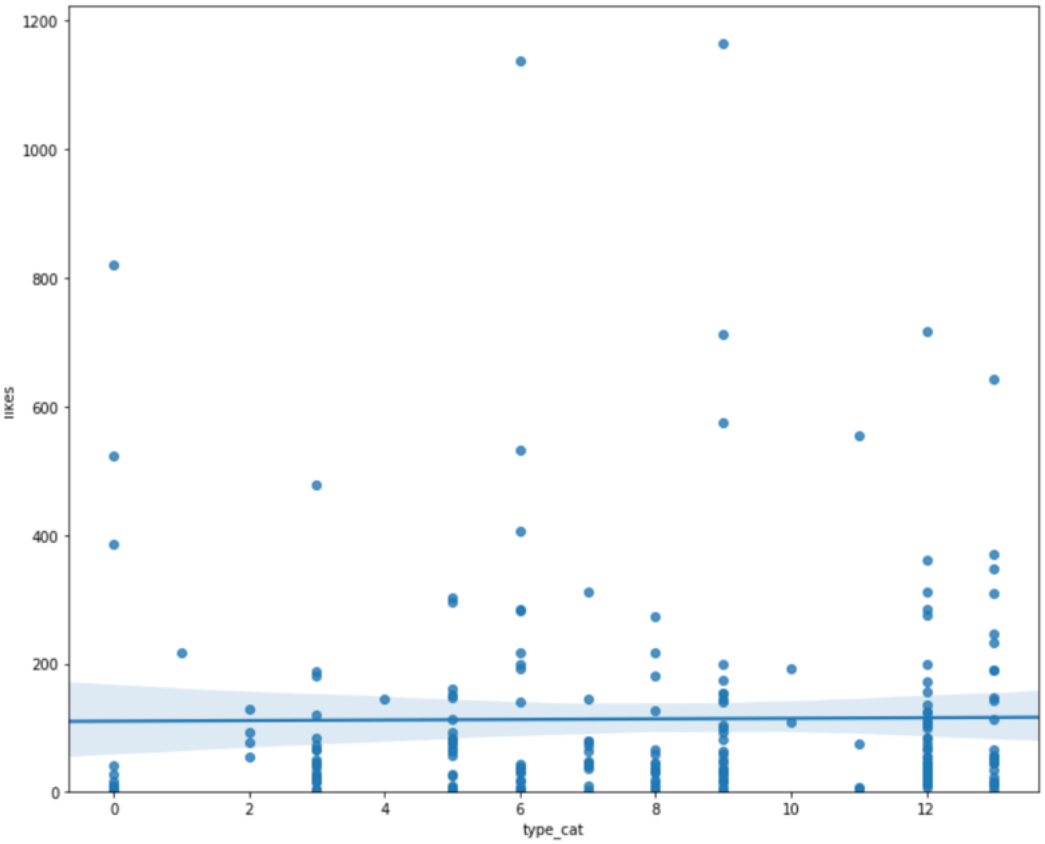
Results

The correlation of the feature variables for San Francisco is indicated by the following:

	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

No strong correlation exists between any category is present. Rating and likes show a moderate correlation, however rating is removed from the analysis. The insignificant outcomes are also prevalent with the regression models.

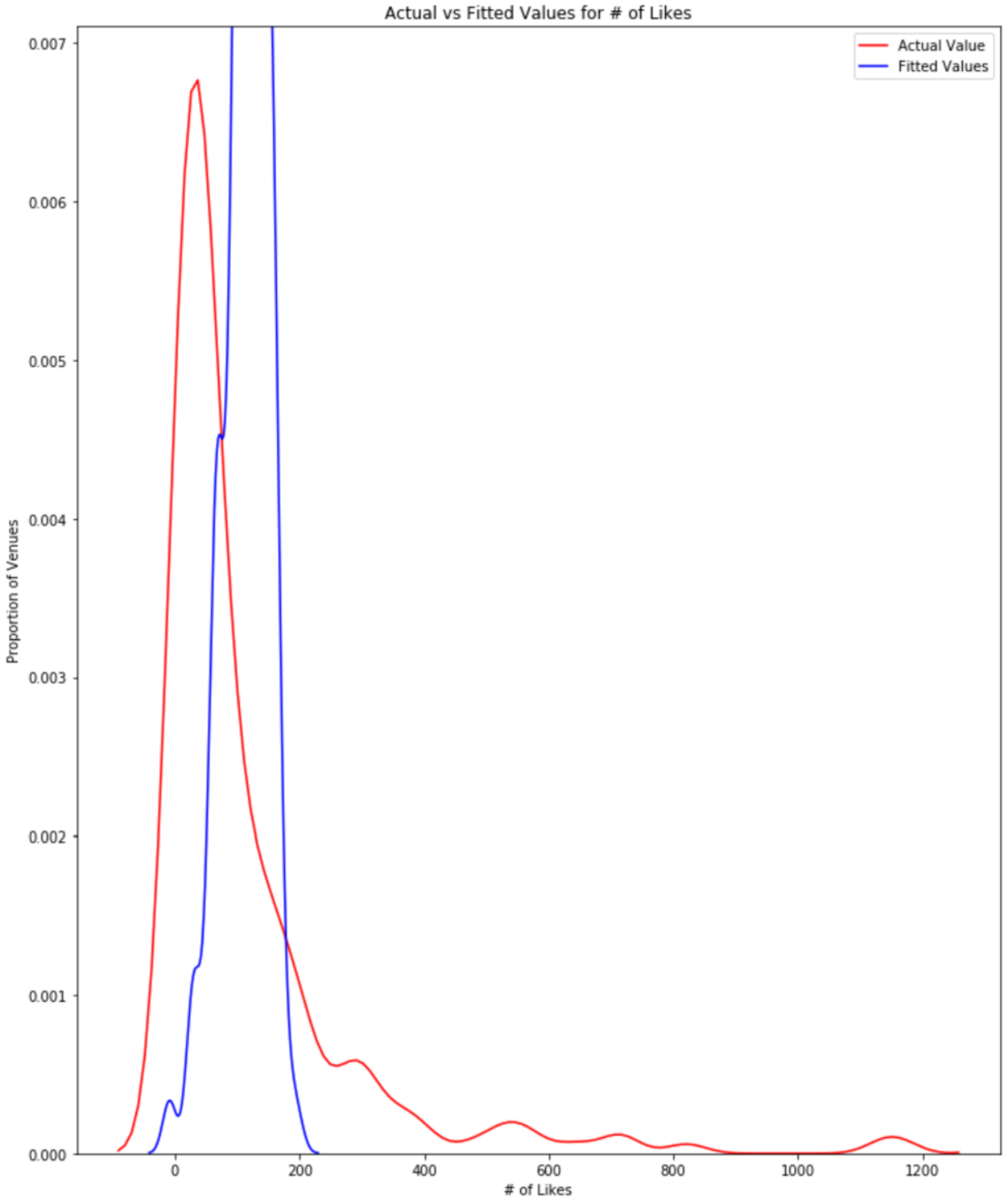
Simple Linear Regression – San Francisco



Model Setup
Predictor Variable: type_cat
Response Variable: likes

Model Evaluation
R-square: 0.000108211807773
MSE: 28670.8825173

Multiple Linear Regression – San Francisco



Model Setup

Predictor Variables: type_cat, price_tier, zipcode

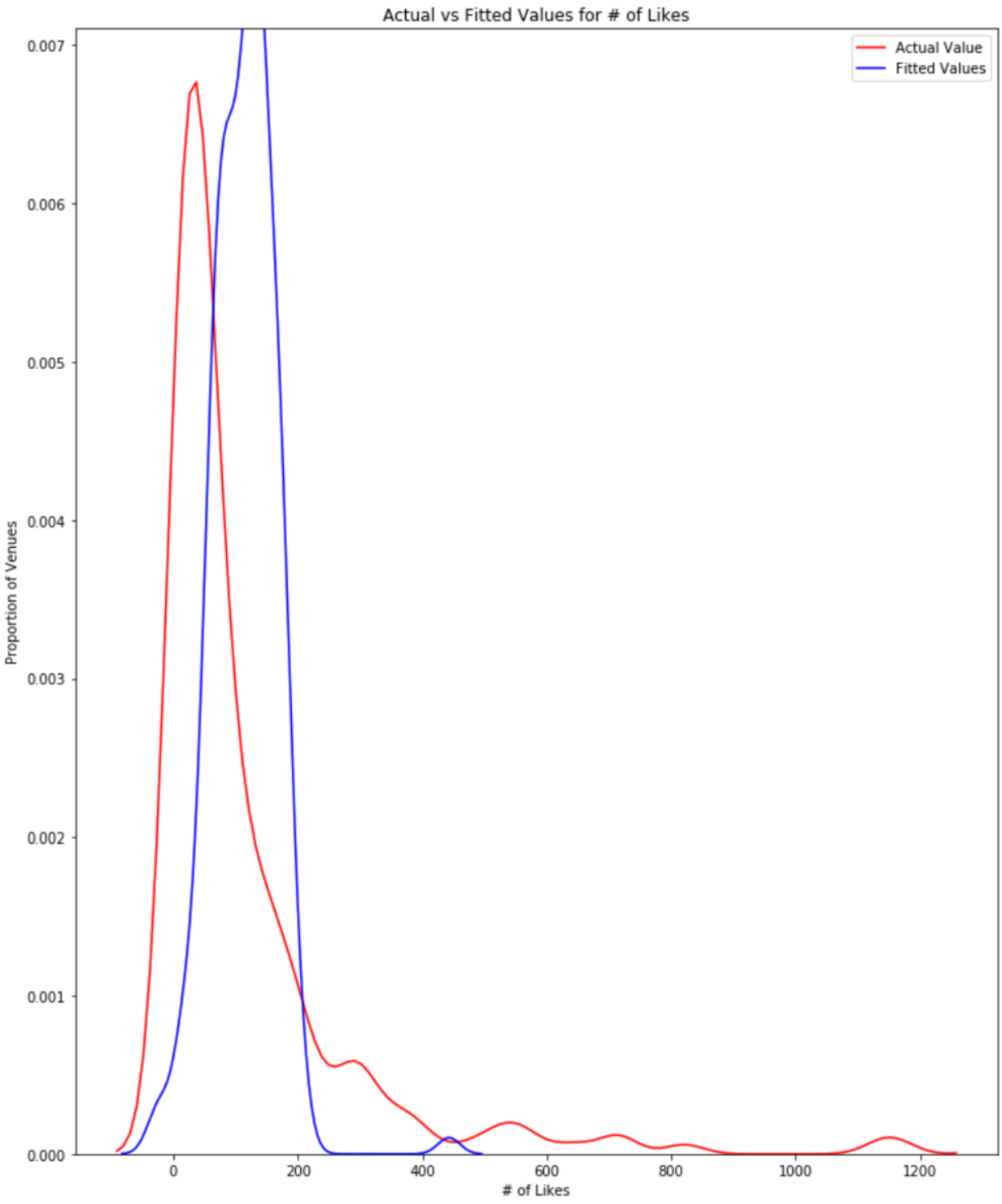
Response Variable: likes

Model Evaluation

R-square: 0.0410841174926

MSE: 27495.9399968

Polynomial Regression (4th order) – San Francisco



Model Setup

Predictor Variables: type_cat, price_tier, zipcode

Response Variable: likes

Model Evaluation

R-square: 0.0894506597658

MSE: 26109.0784707

The decision-tree accuracy for the San Francisco dataset is 0.0222, indicating a weak classifier.

The Los Angeles dataset shows a similar pattern for the modeling.

Los Angeles Venue Data – Correlations:

	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

Los Angeles Venue Data - Model Evaluations:

Model	Predictor Variables	Response Variable	R-square	MSE
Simple Linear Regression	type_cat	likes	0.001185327	3832.541985
Mutiple Linear Regression	type_cat, price_tier, zipcode	likes	0.045158255	3663.813898
Polynomial Regression	type_cat, price_tier, zipcode	likes	0.26553054	2818.225564

Los Angeles Venue Data - Decision Tree accuracy: 0

The aggregate summary examining the # of Likes by cuisine type for each city is shown below:

San Francisco

	likes		price_tier
	count	**mean**	**mean**
type			
Chinese	14	133.357143	1.714286
Cuban	1	218.000000	2.000000
Ethiopian	4	88.500000	1.750000
Indian	22	73.681818	1.909091
Indonesian	1	145.000000	2.000000
Italian	21	88.809524	2.095238
Japanese	19	190.789474	1.947368
Korean	14	73.357143	2.000000
Mediterranean	16	74.375000	1.687500
Mexican	29	148.724138	1.310345
Middle Eastern	2	150.500000	1.500000
Peruvian	4	161.250000	1.750000
Thai	46	96.891304	1.826087
Vietnamese	28	126.142857	1.321429

Los Angeles

	likes		price_tier
	count	**mean**	**mean**
type			
Chinese	10	15.000000	1.600000
Cuban	6	81.833333	1.833333
Ethiopian	7	29.000000	2.000000
Indian	6	18.666667	1.500000
Indonesian	1	182.000000	2.000000
Italian	14	37.571429	1.500000
Japanese	20	53.950000	1.850000
Korean	15	85.200000	2.066667
Mediterranean	4	44.000000	1.750000
Mexican	17	39.705882	1.235294
Middle Eastern	3	5.333333	2.000000
Mongolian	2	7.000000	1.500000
Peruvian	6	32.666667	1.833333
Thai	31	48.000000	1.709677
Vietnamese	5	61.400000	2.000000

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.

Strengthening this analysis, to better determine the venue's responsiveness in this attribute, would require a greater number of data sources. For example, a larger database of venues would be beneficial, as Foursquare limits the number of records to 50 per search criteria. Additional structured data would aide in this analysis as well: 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.