

CAPSTONE: FIND THE BEST PLACE IN OAKLAND, EMERYVILLE , AND SAN DIEGO TO OPEN A RESTAURANT

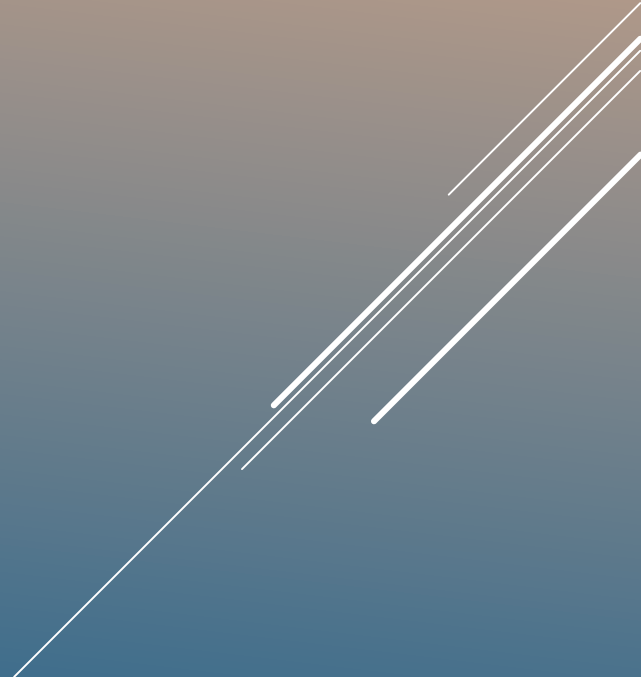
Applied Data Science Capstone

IBM Data Science Professional Certificate

INTRODUCTION:

A decision must be made in order to select the city to place a restaurant. Oakland, Emeryville, and San Diego have been selected for the search. Very good reasons must be exposed in order to attract potential investors in the food business. The type of the restaurant, the city, and the location in the city, are key factors to get success in the business.

The scope of this project is to accurately predict the acceptance that a new restaurant can expect based on the type of cuisine and the location in the cities selected. linear and logistic regressions are used to find which method is better for the prediction.

Several white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, extending from the right edge towards the center.

DATA DESCRIPTION:

Initial dataframe used is named 'raw_dataset', it has the information required to make the analysis.

Foursquare is used to retrieve city coordinates, and to obtain the URLs with the raw data in JSON. From each URL: 'name', 'categories', 'latitude', 'longitude', and 'id' columns are taken for each city, so a city column is also included.

Restaurants in a 1000km radius from the coordinates provided by the geolocator, will be analyzed. Cleaning is performed to remove noisy data and getting only restaurant valid data provided by Foursquare. 'likes' data is important for decision making. Only valuable information is pulled in order to have a strong analysis.

'id' column is used in order to pull the 'likes' and include the information in the dataframe.



Data:

Capstone Notebook:

https://github.com/grodriguece/Coursera_Capstone/blob/master/AppliedDataScienceCapstoneW5.ipynb

Foursquare City Coordinates and venues info:

1. https://api.foursquare.com/v2/venues/explore?&client_id=R3JTQTGYXBW0HQCG5BYPSW3AAOLL3KOUTOUATGMPZSQ01LXB&client_secret=K1YBPNGZ1NJIYKS2ILN41SHNVSK4SGGSL3IXCD0SUA4Y3SQ&v=20201212&ll=37.8044557,-122.2713563&radius=1000&limit=100
2. https://api.foursquare.com/v2/venues/explore?&client_id=R3JTQTGYXBW0HQCG5BYPSW3AAOLL3KOUTOUATGMPZSQ01LXB&client_secret=K1YBPNGZ1NJIYKS2ILN41SHNVSK4SGGSL3IXCD0SUA4Y3SQ&v=20201212&ll=37.8314089,-122.2865266&radius=1000&limit=100
3. https://api.foursquare.com/v2/venues/explore?&client_id=R3JTQTGYXBW0HQCG5BYPSW3AAOLL3KOUTOUATGMPZSQ01LXB&client_secret=K1YBPNGZ1NJIYKS2ILN41SHNVSK4SGGSL3IXCD0SUA4Y3SQ&v=20201212&ll=32.7174202,-117.1627728&radius=1000&limit=100

100 venues for Oakland, California were returned by Foursquare.

100 venues for Emeryville, California were returned by Foursquare.

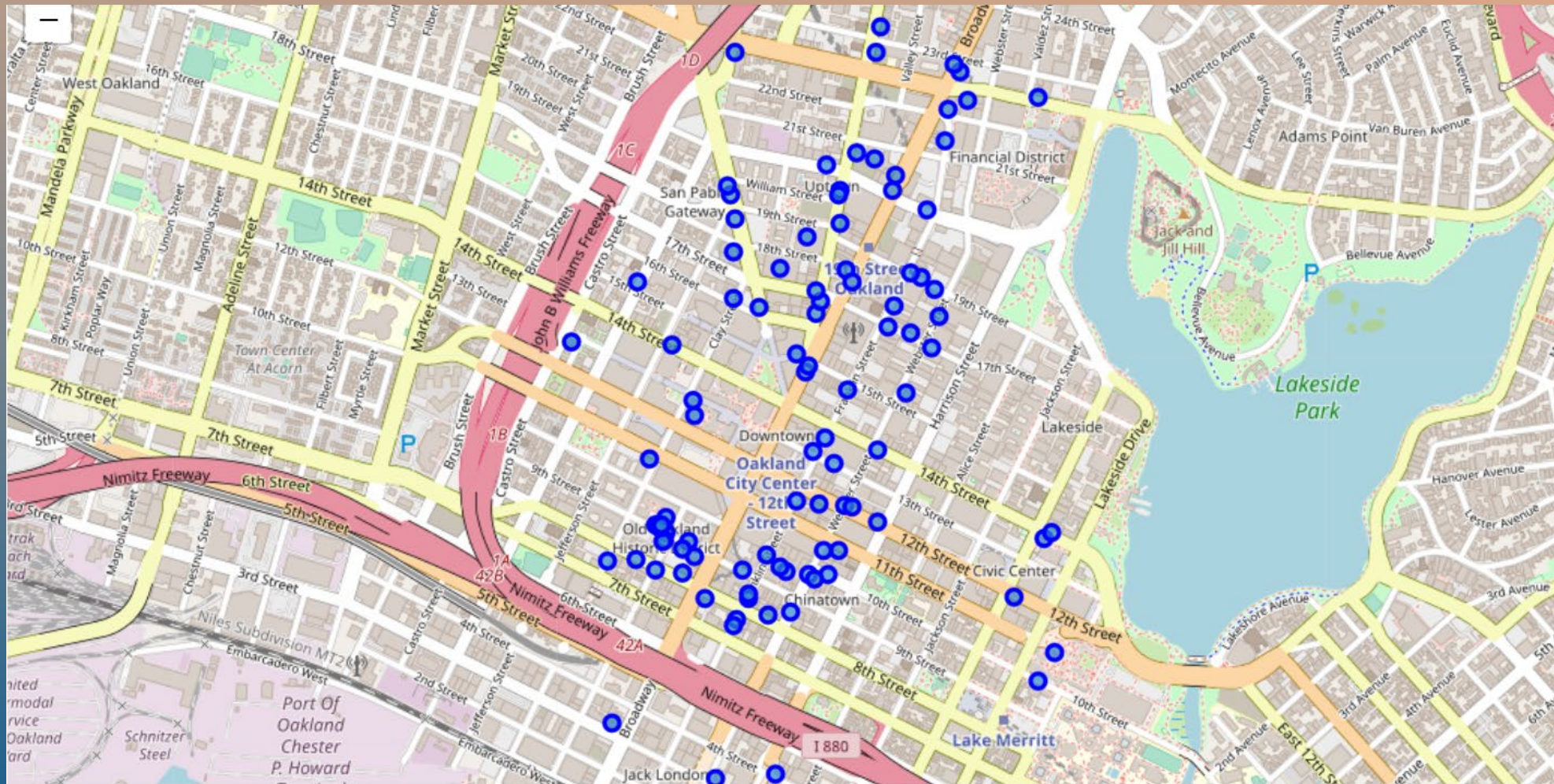
100 venues for San Diego, California were returned by Foursquare.

Combined dataset for three cities: `nearby_venues`

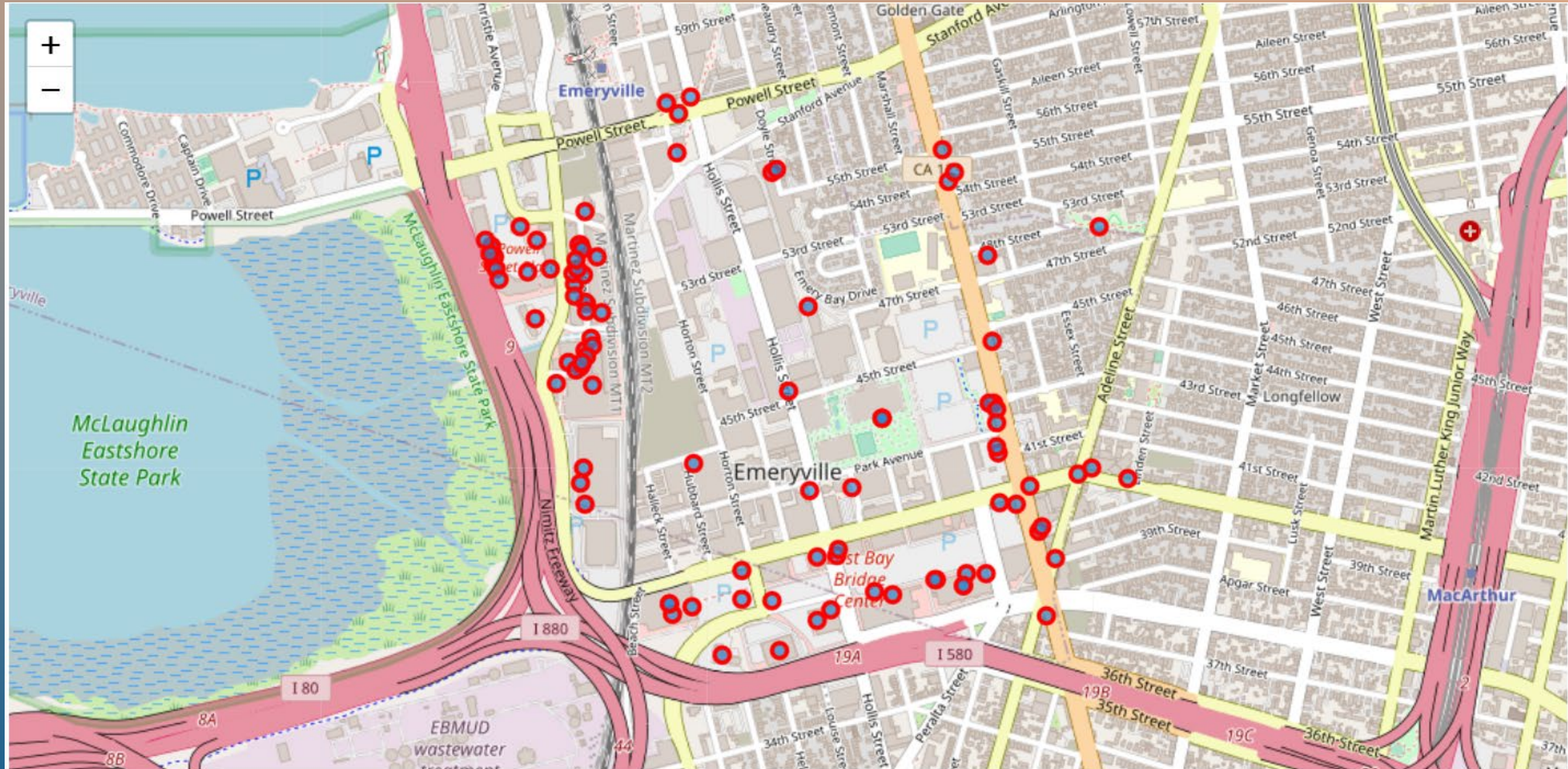
Dataframe prepared for Machine Learning: `raw_dataset`

Linear Regression dataframe: `reg_dataset`

Oakland venues locations:



Emeryville venues locations:



San Diego venues locations:



nearby_venues

```
nearby_venues.head()
```

	name	categories	lat	lng	id	city
0	Oaklandish	Clothing Store	37.805075	-122.270726	4dfb9c2c1f6eeef806ab898c	Oakland
1	Golden Lotus Vegetarian Restaurant	Vegetarian / Vegan Restaurant	37.803290	-122.270473	49cebb1bf964a520785a1fe3	Oakland
2	Bar Shiru	Bar	37.806378	-122.270393	5c5b9abdf870fd002c35d291	Oakland
3	Cafe Van Kleeef	Bar	37.806660	-122.270273	46884818f964a52056481fe3	Oakland
4	Cape & Cowl	Comic Shop	37.806725	-122.272747	56562410498ea43ab630819a	Oakland

```
nearby_venues['likes'] = like_list  
nearby_venues.head()
```

```
[77, 22, 156, 14, 9, 71, 33, 202, 65, 51, 104, 369, 61, 177, 93, 188, 39, 40, 39, 23, 43, 56, 24, 102, 13, 259, 11, 73, 239, 25, 36, 45, 69, 5, 43, 52, 229, 33, 120, 99, 247, 30, 332, 133, 35, 24, 31, 61, 13, 18, 56, 92, 16, 62, 68, 4, 65, 17, 17, 5, 1, 22, 41, 0, 3, 31, 15, 0, 156, 76, 9, 2, 78, 131, 171, 142, 26, 41, 105, 34, 94, 18, 24, 19, 35, 296, 31, 128, 21, 320, 104, 30, 31, 480, 540, 489, 174, 132, 185, 103, 54, 204]
```

	name	categories	lat	lng	id	city	likes
1	Golden Lotus Vegetarian Restaurant	Vegetarian / Vegan Restaurant	37.803290	-122.270473	49cebb1bf964a520785a1fe3	Oakland	77
6	Beauty's Bagel Shop	Bagel Shop	37.806082	-122.268356	5bd0959cf1fdaf002ce03e11	Oakland	22
7	Abura-Ya	Japanese Restaurant	37.805959	-122.267693	539a69a7498ee67090b2b285	Oakland	156
11	Anula's Cafe	Sandwich Place	37.803583	-122.270151	4b50d22df964a520a73327e3	Oakland	14
13	World Famous Hotboys	Fried Chicken Joint	37.806526	-122.272040	5e0a805333617d00086cd498	Oakland	9

Methodology:


Both linear and logistic regression are used to train and test the data.

Linear regression is used to predict the number of 'likes' a new restaurant in this region will acquire. Sci-Kit Learn is used for this stage.

Logistic regression is used as the classification method.

Since binning is used when classifying by number of 'likes', multinomial logistic regression is used to perform the analysis.

Although the ranges are discrete categories, they can be considered ordinal in nature. The logistic regression is specified as being both multinomial and ordinal.

Several white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, extending from the right edge towards the center.

raw_dataset

```
raw_dataset.head()
```

	name	categories	lat	lng	id	city	likes	categories_classified
1	Golden Lotus Vegetarian Restaurant	Vegetarian / Vegan Restaurant	37.803290	-122.270473	49cebb1bf964a520785a1fe3	Oakland	77	american
6	Beauty's Bagel Shop	Bagel Shop	37.806082	-122.268356	5bd0959cf1daf002ce03e11	Oakland	22	casual
7	Abura-Ya	Japanese Restaurant	37.805959	-122.267693	539a69a7498ee67090b2b285	Oakland	156	asian
11	Anula's Cafe	Sandwich Place	37.803583	-122.270151	4b50d22df964a520a73327e3	Oakland	14	casual
13	World Famous Hotboys	Fried Chicken Joint	37.806526	-122.272040	5e0a805333617d00086cd498	Oakland	9	casual

col_0	count
categories_classified	
american	21
asian	26
casual	18
european	17
latin	19

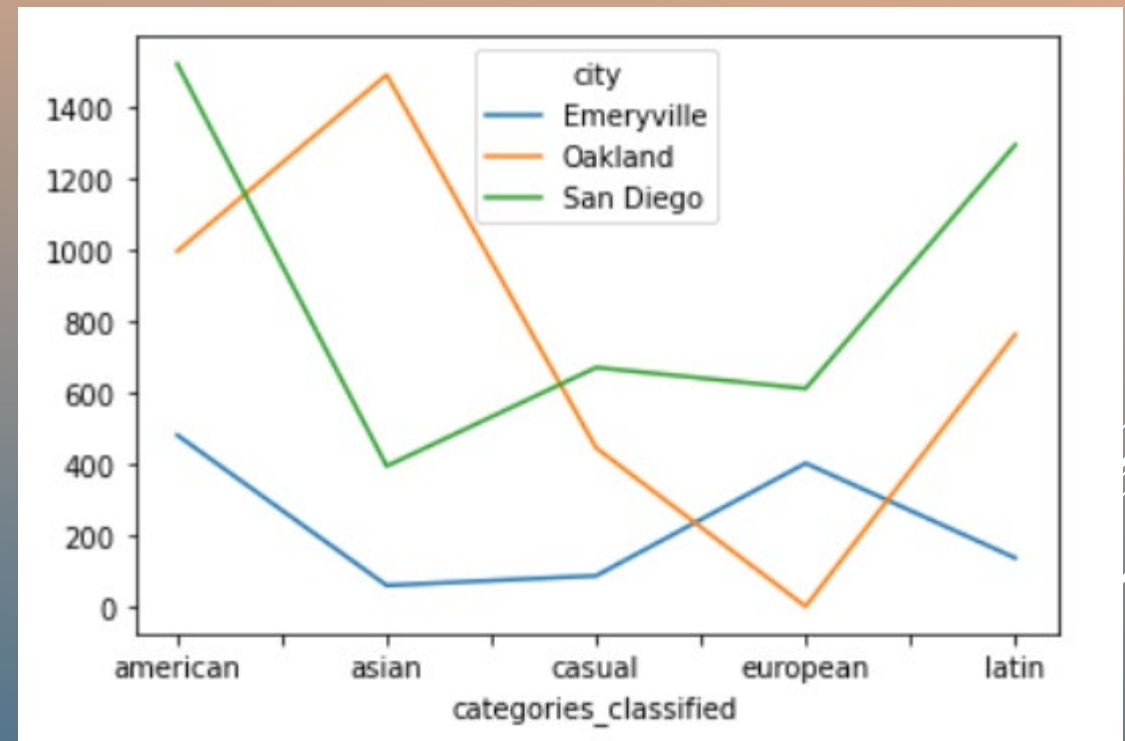
Linear Regression:

```
reg_dataset.head()
```

	name	american	asian	casual	european	latin	Emeryville	Oakland	San Diego	ranking	likes
1	Golden Lotus Vegetarian Restaurant	1	0	0	0	0	0	1	0	2	77
6	Beauty's Bagel Shop	0	0	1	0	0	0	1	0	3	22
7	Abura-Ya	0	1	0	0	0	0	1	0	1	156
11	Anula's Cafe	0	0	1	0	0	0	1	0	3	14
13	World Famous Hotboys	0	0	1	0	0	0	1	0	3	9

Likes distribution:

city	Emeryville	Oakland	San Diego
categories_classified			
american	480.0	994.0	1519.0
asian	59.0	1488.0	393.0
casual	86.0	444.0	669.0
european	401.0	NaN	610.0
latin	136.0	761.0	1292.0



American type cuisine is the one with the highest likes in San Diego and Oakland.

Multiple linear Regression:

```
Coefficients: [[ 78.40365032  37.13520878  13.58336968  30.80867748  91.69528923  
 8.97113452 -32.99602203  24.02488751]]
```

```
Residual sum of squares: 15630.49  
Variance score: 0.07
```

A linear regression model was trained on a random subsample of 80% and then the other 20% was used for testing purposes. In order to evaluate if the model is reasonable, the residual sum of squares and variance score were both calculated (15630.49, 0.07). The variance score is quite low, which means that is not a good way of modeling the data.

LogisticRegression:

Multinomial Ordinal Logistic Regression

```
coef = mul_ordinal.coef_[0]  
print (coef)
```

```
[ 0.25944851 -0.37299808 -0.27256256  0.1043218   0.38656054  0.08339691  
 -0.52329456  0.43989866]
```

Multinomial Ordinal Logistic Regression Prediction Capabilities

```
jaccard_score(y_test, yhat, average='weighted')
```

```
0.3612836438923396
```

```
log_loss(y_test, yhat_prob)
```

```
1.0331877615384777
```

The multinomial ordinal logistic regression model was also trained on a random subsample of 80% and then tested on the remaining 20%. The jaccard score and log-loss were both calculated (36.13% and 1.033 respectively). A jaccard score of 36.13% is quite reasonable.

Logistic Regression:

Exploration of Coefficient Magnitudes of Full Dataset

```
print(coef)
[ 0.44665774 -0.20762377 -0.19235879 -0.03626791  0.0821337  0.06140855
 -0.74041853  0.67901042]
```

```
print(classification_report(y_test, yhat))
```

	precision	recall	f1-score	support
1	0.67	0.40	0.50	10
2	0.00	0.00	0.00	5
3	0.55	0.92	0.69	13
accuracy			0.57	28
macro avg	0.40	0.44	0.40	28
weighted avg	0.49	0.57	0.50	28

The coefficients show that opening a restaurant in San Diego (0.679), or serving American cuisine (0.447) are positively associated with 'likes'.

The results showed that the precision score for classifying whether the new restaurant would fall into classes 1, 2, or 3 (highest, medium, lowest) were 67%, 0%, and 55%. Therefore, the model is better at predicting if a restaurant will fall into the best or worst percentile of likes.

Results:

A linear regression model was trained on a random subsample of 80% and then the other 20% was used for testing purposes. In order to evaluate if the model is reasonable, the residual sum of squares and variance score were both calculated (15630.49, 0.07). The variance score is quite low, which means that is not a good way of modeling the data. So logistic regression was selected for the analysis.

The multinomial ordinal logistic regression model was also trained on a random subsample of 80% and then tested on the remaining 20%. The jaccard score and log-loss were both calculated (36.13% and 1.033 respectively). A jaccard score of 36.13% is quite reasonable. The classification report is included in the analysis.

Given the modestly accurate ability of this mode, we have the ability to run the model on the complete dataset. The coefficients show that opening a restaurant in San Diego (0.679), or serving American cuisine (0.447) are positively associated with 'likes'.

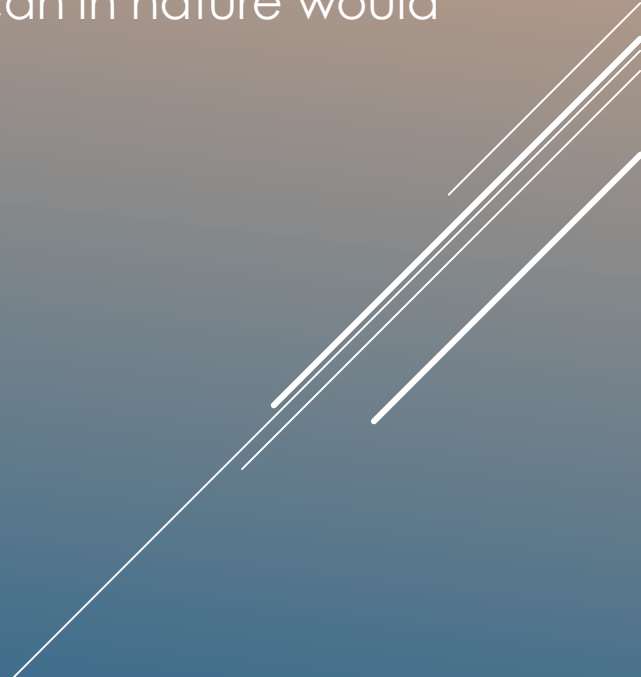
Discussion:

The first thing to note is that given the data, logistic regression presents a better fit for the data over linear regression. Using logistic regression, we were able to obtain a Jaccard Score of 36.13%, which although not perfect, is more reasonable than the low variance score obtained from the linear regression (0.07). As stated before, please note that for the purposes of this project, we are assuming that likes are a good proxy for how well a new restaurant will do in terms of brand, image and by extension how well the restaurant will perform business-wise. Whether or not these assumptions hold up in a real-life scenario is up for discussion, but this project does contain limitations in scope due to the amount of data that can be fetched from the FourSquare API.

As such, to obtain insights into this data, we can proceed with breaking down the results of the logistic regression model. The results showed that the precision score for classifying whether the new restaurant would fall into classes 1, 2, or 3 (highest, medium, lowest) were 67%, 0%, and 55%. Therefore, the model is better at predicting if a restaurant will fall into the best or worst percentile of likes. This is good as we are mostly concerned with whether the restaurant will perform well or not so the high accuracy of predictions for the two extremum is a welcome feature. This allows us to fairly accurately predict the general performance of the business opportunity. Different binning methods for the classes were attempted, but the use of 3 bins by far yielded the best Jaccard Score.

Discussion:

Additionally, not only are we attempting to predict the general business performance but also pull insights to inform on business strategy. In this case strategy insight can be gleamed from the coefficient values from running the logistic regression on the full dataset. As such, we can see that opening a restaurant in Emeryville, or serving cuisine that is asian or casual in nature, are associated negatively with "likes." This suggests that the business opportunity should be opening a restaurant in either Oakland or San Diego, with a cuisine that is Latin or American in nature would be the best approach for maximizing likes.

Several white lines of varying lengths and angles are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

Conclusion:

In conclusion, after analyzing restaurant 'likes' in California from the 300 restaurants, it can be concluded that the approach to best take when looking to maximize business performance (as measured by 'likes') is to open a restaurant that is either Latin, or American and that opening the venue in either Oakland or San Diego rather than Emeryville would be the best approach. Additionally, the predictive capabilities of the logistic regression prediction model proved to be the most accurate for classifying whether a restaurant fell in either the best or worst classes when the data was binned into their 3 respective classes.