Marsha Curtis
Michela Ziemer
Erin Parker
Tuyet Do

_ August 2022

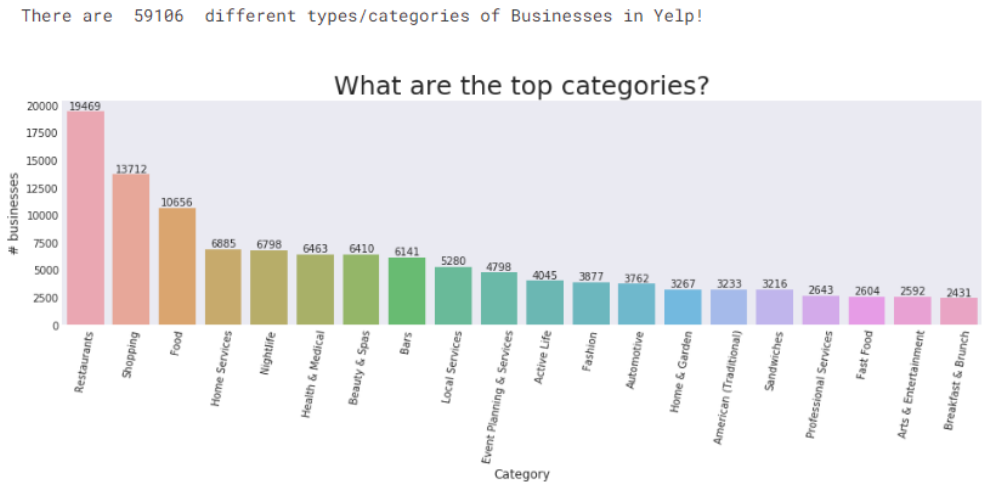
SnowBird Taco Tour

Our initial goal and plan was to do a “Great North American Taco Tour” utilizing Yelp data found on Kaggle in order to analyze taco and burrito locations and help others find them too in the US and Canada. Our inspiration was rooted in the deep and unabiding love nearly all Americans have for the Mexican (or maybe not-so-Mexican depending on how close you are to Canada) favorite. The plan was to provide a resource for others to input their desired location(s) and attribute(s) and then provide them with recommendations. We wanted to use supervised machine learning to predict restaurant ratings based on various business attributes provided in the Kaggle dataset as well. These attributes included elements spanning a wide range of topics like parking availability, ambiance, music and dancing, menu availability for special diets, and whether our four-legged friends are welcome as well.

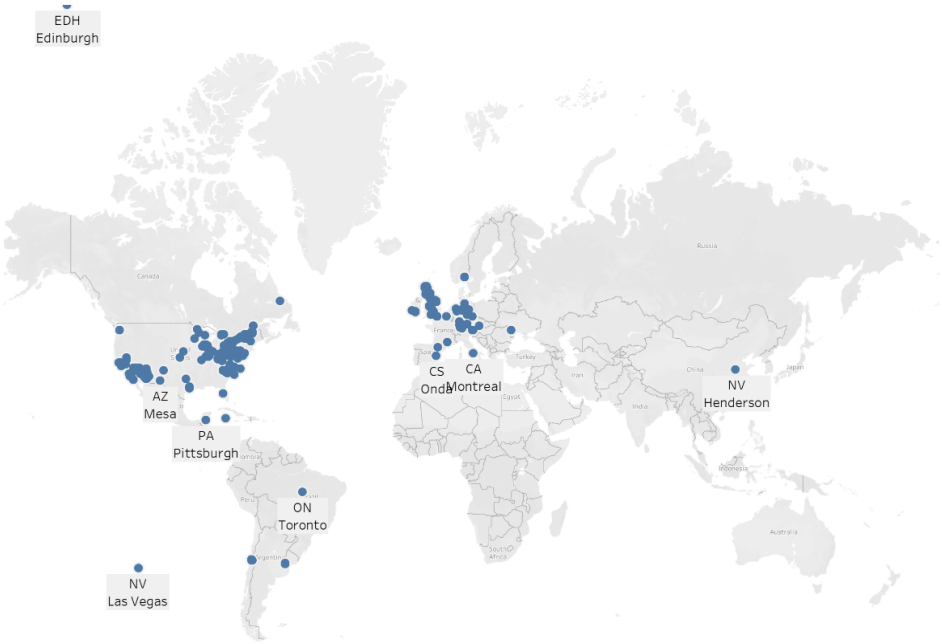
We felt like this was a fun topic because of its relatability, and also we could put a really spicy twist on our dashboards. Our initial expectations were that given the popularity of tacos and burritos, there would be a substantial amount of data and perhaps some concentrations along the coastlines and states bordering Mexico. But with all visions, we found that revisions needed to be made after we did a deep dive into our data. Turning what we were hoping as the “Great North American Taco Tour” into the “SnowBird Taco Tour”.

Exploratory Data Analysis (EDA)

We decided on the data filters to first be only restaurants (19,469) and then to those that serve tacos and burritos (1,083).



But shortly after mapping the data based on geocodes in Tableau, it was apparent that the geocodes were not the most accurate and needed to be replaced.



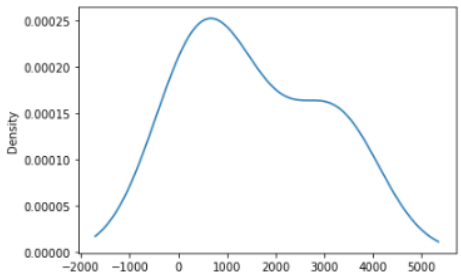
As seen above and to our surprise, the geocodes were not being plotted correctly. For example, there is a plot point for Toronto, Canada in Brazil, South America and even Henderson, Nevada in China to name only a couple. An additional EDA we found was during the machine learning process and it was looking at the visualized density/shape of the ratings. Although it is a fairly even distribution, there could be possible speculations of it being bimodal.

```
# Visualize the value counts of stars
stars_counts = restaurants_df.stars.value_counts()
print(stars_counts)

stars_counts.plot.density()
```

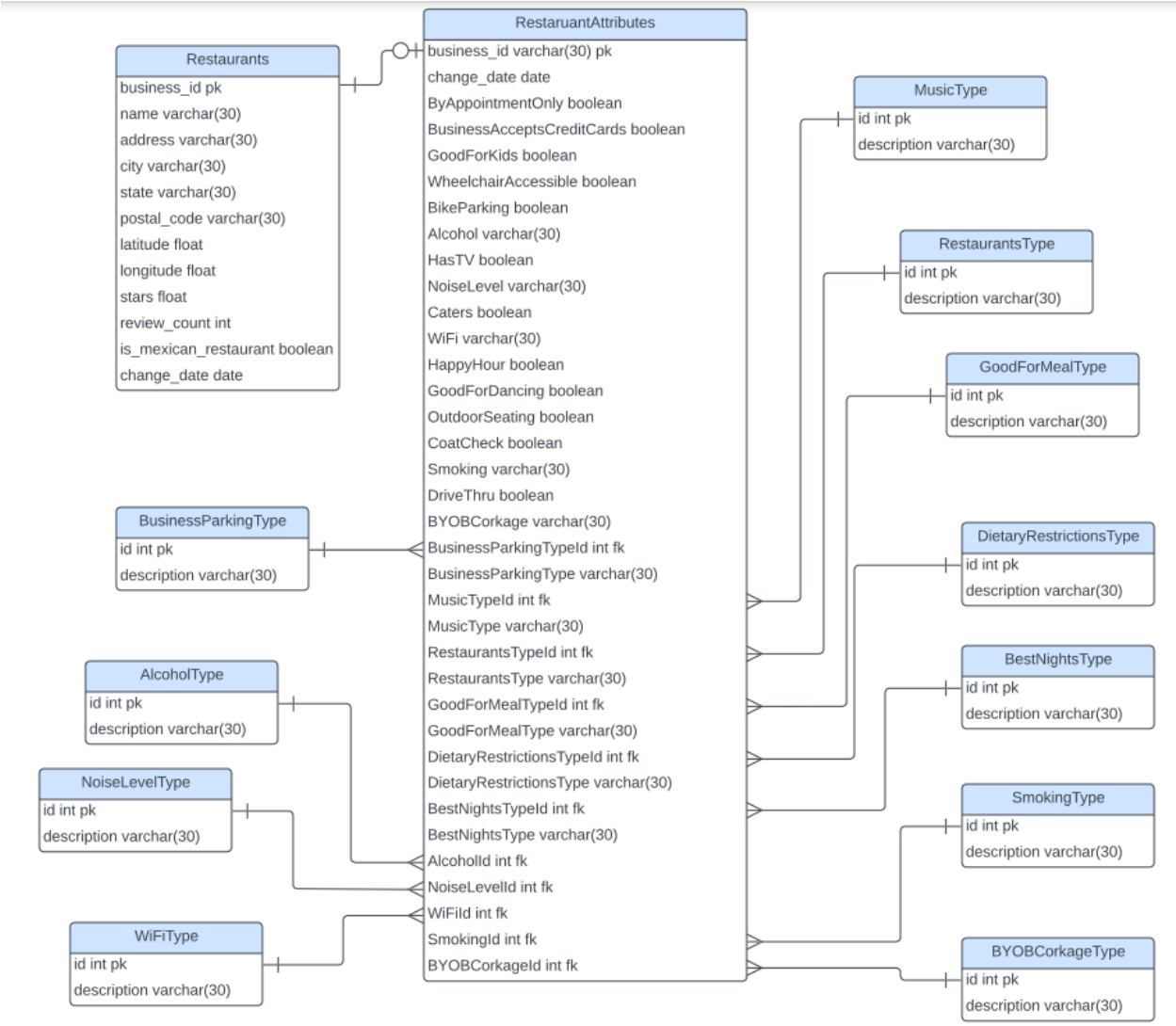
```
3.5    3581
4.0    3194
3.0    2950
2.5    1485
4.5    1363
2.0     678
5.0     333
1.5     209
1.0      55
Name: stars, dtype: int64

<AxesSubplot:ylabel='Density'>
```



Database Section

We considered both BI and ML use cases when developing our data model as well as our cleaning strategy. Restaurant demographic data is contained in a dimension table called “Restaurants” and attributes are contained in a separate table called “RestaurantAttributes” We developed 11 additional small categorical tables which could be used as lookups in BI as well as to reduce the number of dimensions for machine learning. These categorical tables contain derived distinct values from categorical columns or from ranges of boolean columns containing similar attributes.



In our data engineering notebook we used several libraries: pandas for creating and manipulating dataframes, numpy mainly for null handling, SQLAlchemy for interacting with sqlite database, googlemaps and requests for geocoding, datetime and dateutil.parser for date type conversions, pathlib and csv for interacting with CSV files.

We developed a number of functions for the ETL process. For the restaurants table:

- Function takes input parameters dbg (Y/N indicator) and an api key to pass through to the geocoding functions
- Import data from yelp_business.csv
- Filter to only open businesses containing the category “Restaurants”
- Set a bit flag “is_mexican_restaurant” to indicate if the restaurant has Mexican or Tex-Mex cuisine
- Drop unneeded columns
- Replace nulls with an empty string
- Tag rows with a “change_date” value

```
# Run Restaurants ETL
business_df = etl_restaurants(dbg,gmaps_key)

# Validate success
verify_inserts('Restaurants')
```

```
index BIGINT
business_id TEXT
name TEXT
address TEXT
city TEXT
state TEXT
postal_code TEXT
latitude FLOAT
longitude FLOAT
stars FLOAT
review_count BIGINT
is_mexican_restaurant BOOLEAN
change_date DATE
Row Count: 14225
```

2]:

	index	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_mexican_re
0	10	XOSRcvtaKc_Q5H1SAzN20A	"East Coast Coffee"	"737 West Pike St"	Houston	PA	15342	40.241548	-80.212815	4.5	3	
1	15	l09JfMeQ6ynYs5MCJtrcmQ	"Alize Catering"	"2459 Yonge St"	Toronto	ON	M4P 2H6	43.711399	-79.399339	3.0	12	
2	29	gAy4LYpsScrj8POnCW6btQ	"Toast Cafe"	"2429 Hwy 160 W"	Fort Mill	SC	29708	35.047287	-80.990559	3.5	6	
3	32	1_3nOM7s9WqnJWTNu2-i8Q	"Le Bistro Balmoral"	"305 Rue Sainte-Catherine O"	Montreal	QC	H2X 2A1	45.506772	-73.566725	3.0	8	
4	44	BnuzcebyB1AfxH0kjNWqSg	"Carrabba's Italian Grill"	"245 Lancaster Ave"	Frazer	PA	19355	40.041003	-75.542497	3.5	25	
5	64	EJFdWX908N8Yc2XG0Lky8A	"River Moon Cafe"	"104 43rd St"	Pittsburgh	PA	15201	40.472735	-79.963265	4.0	5	
6	91	F0fEKpTk7gAmuSFI0KW1eQ	"Cafe Mastrioni"	"4250 S Rainbow Blvd, Ste 1007"	Las Vegas	NV	89103	36.111057	-115.241688	1.5	3	
7	97	HAX1zec191t7QkT2sBZ76A	"La Isla Cuban Restaurant"	"1816 Galerea Blvd, Ste D"	Charlotte	NC	28270	35.137223	-80.734594	3.0	4	
8	104	1nhf9BPXOBFBkbRkpsFaxA	"Mirage Grill & Lounge"	"117 Eglinton Avenue E"	Toronto	ON	M4P 1H4	43.707465	-79.394285	2.0	6	
9	118	T5CdfrZWw-uW9Y5L_sddqQ	"Police Station Pizza"	"7235 Steubenville Pike"	Oakdale	PA	15071	40.442828	-80.186293	3.0	9	

To prepare and load data into the Reviews table we performed the following tasks:

- Function takes input value of restaurants as a dataframe
- Import data from yelp_tip.csv

- Drop rows with null values
- Filter to include only restaurant reviews
- Create a primary key by applying index as a column and reorder columns so that index is the first column
- Convert “date” string value to date data type
- Tag rows with a “change_date” value
- Write to Reviews table in SQLite database

```
# Run Reviews ETL
etl_reviews(business_df)

# Validate success
verify_inserts('Reviews')
```

```
business_id TEXT
review_id TEXT
text TEXT
date DATETIME
likes BIGINT
user_id TEXT
change_date DATE
Row Count: 1098322
```

	business_id	review_id	text	date	likes	user_id	change_date
0	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	Combo A: Roast duck, roast pork, Singapore noo...	2015-10-12	0	6tbXpUIU6upoeqWNDo9k_A	2022-08-11
1	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	Make reservation on weekend	2013-01-27	0	CxDOIDnH8gp9KXzpBHJYXw	2022-08-11
2	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	Great place for couple has $7.99 dish	2013-01-27	0	CxDOIDnH8gp9KXzpBHJYXw	2022-08-11
3	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	King of bbq pork for $22	2013-01-27	0	CxDOIDnH8gp9KXzpBHJYXw	2022-08-11
4	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	Their lunch combos for small groups is a decen...	2013-01-29	0	Tc3GAQdAfOW542ROdyCZPg	2022-08-11
5	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	Make sure to request the delicious house soup....	2015-01-04	0	mFwRTTDW0Yr-rFkTF2cFsw	2022-08-11
6	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	$7.50 lunch special, dish of rice\/noodles wit...	2017-01-15	0	0cUzu82KJiE5_xZA0lu3ZQ	2022-08-11
7	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	$5 lunch special	2014-07-11	0	2oMkzQcRL7-d7URt3Xo_Xg	2022-08-11
8	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	$6 lunch special. A lot of selection on the lu...	2015-02-19	0	3yMtpQ_wV4ZGg6E69uE1PQ	2022-08-11
9	-6MefnULPED_I942VcFNA	-6MefnULPED_I942VcFNA	BBQ pork is sold out early on Saturday	2013-03-23	0	EiP1OFgs-XGcKZux0OKWIA	2022-08-11

To clean data for and load RestaurantAttributes table, we performed the following tasks:

- Function takes input value of restaurants as a dataframe
- Import from yelp_business_attributes.csv
- Correct issue with column naming (all cols after business_id were shifted 1 left)
- Filter to only include restaurants in data set
- For numeric categories: Convert “Na” to 0
- For string categories: Convert “Na” to “no”
- For boolean columns: Convert Null to False; Convert “Na”, Null, 0, and “False” to False; and Convert “True” and 1 to True
- Drop columns with only one value - these won’t be meaningful
- Write to CSV file for BI to use directly and for ML to import for further processing
- Tag rows with a “change_date” value
- Write data to RestaurantAttributes table in SQLite database

```
# Run RestaurantAttributes ETL
etl_restaurant_attributes(business_df)

# Validate success
verify_inserts('RestaurantAttributes')
```

```
business_id TEXT
change_date DATE
ByAppointmentOnly BOOLEAN
BusinessAcceptsCreditCards BOOLEAN
BusinessParking_garage BOOLEAN
BusinessParking_street BOOLEAN
BusinessParking_validated BOOLEAN
BusinessParking_lot BOOLEAN
BusinessParking_valet BOOLEAN
RestaurantsPriceRange2 TEXT
GoodForKids BOOLEAN
WheelchairAccessible BOOLEAN
BikeParking BOOLEAN
Alcohol TEXT
HasTV BOOLEAN
NoiseLevel TEXT
RestaurantsAttire TEXT
Music_dj BOOLEAN
Music_karaoke BOOLEAN
RestaurantsGoodForGroups BOOLEAN
Caters BOOLEAN
WiFi TEXT
RestaurantsReservations BOOLEAN
RestaurantsTakeOut BOOLEAN
HappyHour BOOLEAN
GoodForDancing BOOLEAN
RestaurantsTableService BOOLEAN
OutdoorSeating BOOLEAN
RestaurantsDelivery BOOLEAN
BestNights_monday BOOLEAN
BestNights_friday BOOLEAN
BestNights_wednesday BOOLEAN
BestNights_thursday BOOLEAN
BestNights_sunday BOOLEAN
BestNights_saturday BOOLEAN
GoodForMeal_dessert BOOLEAN
GoodForMeal_latenight BOOLEAN
GoodForMeal_lunch BOOLEAN
GoodForMeal_dinner BOOLEAN
GoodForMeal_breakfast BOOLEAN
GoodForMeal_brunch BOOLEAN
CoatCheck BOOLEAN
Smoking TEXT
DriveThru BOOLEAN
BYOBCorkage TEXT
DietaryRestrictions_dairy-free BOOLEAN
DietaryRestrictions_gluten-free BOOLEAN
DietaryRestrictions_vegan BOOLEAN
DietaryRestrictions_kosher BOOLEAN
DietaryRestrictions_halal BOOLEAN
DietaryRestrictions_soy-free BOOLEAN
DietaryRestrictions_vegetarian BOOLEAN
Row Count: 13848
```

	business_id	change_date	ByAppointmentOnly	BusinessAcceptsCreditCards	BusinessParking_garage	BusinessParking_street	BusinessP
0	XOSRcvtaKc_Q5H1SAzN20A	2022-08-11	False	False	False	False	
1	I09JfMeQ6ynYs5MCJtrcmQ	2022-08-11	False	False	False	True	
2	gAy4LYpsScrj8POnCW6btQ	2022-08-11	False	False	False	False	
3	1_3nOM7s9WqnJWTNu2-i8Q	2022-08-11	False	False	False	False	
4	BnuzcebyB1AfxH0kjNWqSg	2022-08-11	False	False	False	False	
5	EJFdWX908N8Yc2XG0Lky8A	2022-08-11	False	False	False	False	
6	F0fEKpTk7gAmuSFI0KW1eQ	2022-08-11	False	False	False	False	
7	HAX1zec191t7QkT2sBZ76A	2022-08-11	False	False	False	False	
8	1nhf9BPXOBFBkbRkpsFaxA	2022-08-11	False	False	False	False	
9	T5CdfrZWw-uW9Y5L_sddqQ	2022-08-11	False	False	False	False	

- Apply filters and geocode cleaning algorithms identified in EDA phase
    - Filter to Mexican and Tex Mex restaurants only to limit API calls
    - Convert address fields to a full address string to be used as an url input parameter

    - Apply function to rows in dataframe
    - Output to csv file for use in BI (Tableau dashboards)

We developed a number of helper functions to aid in this process. These are small functions that represent tasks that needed to be performed multiple times, and helped reduce the repetition in our code.

- Validation function
    - Input parameter: table name
    - Prints column list with data types and row counts from SQLite table using sqlalchemy ORM query
    - Returns dataframe containing the first 10 rows in the table using Pandas query
- Geocoding function
    - Need to install: pip install -U googlemaps
    - Input parameter: address
    - Make API call and returns list containing latitude and longitude
- Type conversion: string to date
    - Input parameter: date column with string data type
    - Outputs values cast to date data type
- Type conversion: string to boolean
    - Input parameters: dataframe, column
    - Maps a set of strings to boolean values
- Adding change date:
    - Input parameter: dataframe
    - Outputs dataframe with a column “change_date” appended with the value set to the current date
- Dimensionality reduction
        Inputs search string and dataframe
    - Generates a new dimension table based on column names matching the search string; attributes are id and description; description = col names based on search string
    - Applies a new Id and Type column containing encoded values vs bit flags to help with BI; otherwise this would only apply the ID column
    - Returns the updated dataframe
- Encoding
    - Inputs column name and dataframe
    - Generates a new dimension table based on input column name; attributes are id and description; description = distinct values in input column
    - Applies a new Id column containing encoded values
    - Returns the updated dataframe

There are a number of things where, if given unlimited time, we could do better. For example, we would build the attributes table without string versions for the categories so that the table functions as a standard relational table. This was done so BI queries could use a flat

table, but is not best practice. Another example is that there is a bug in the reduce_dims function where it isn't stripping the search string from the description correctly. This leads to values like "BusinessParkingTypevalet" and "BusinessParkingTypelot" rather than having clean categorical labels like "valet" or "lot" and takes more space than is truly required. Finally, we would like to have spent more time on address cleansing to correct typos. For example we see Phoenix spelled as Pheonix or listed as Phoenix, AZ, which affects the functionality of our dropdown menus.

Machine Learning

Our goal with our machine learning project was to use restaurant attributes to predict how a restaurant would be rated by customers. Most of our data cleaning work was handled in ETL so that data could be shared between Machine Learning and the BI dashboards. Kicking off the machine learning process, our modeled data was split into training and testing sets. After testing multiple models: RandomForestClassifier, AdaBoostClassifier, ExtraTreesClassifier, GradientBoostingClassifier, and XGBoost with weighting: scale_pos_score of 14 and 15 based on formula, we saw that all models were extremely overfit in the training phase, and XGBoost RandomForest were the worst in that regard. All models yielded accuracy scores around 40-50% in the testing set. Also, in all models we observed that 4-star ratings had the highest F1 scores around 60-65%, 1 and 5-star reviews were the lowest with F1 scores between 0 and 3%. In the end we dropped all attributes except the 9 most complete attributes and re-trained the model, which yielded very similar results. Finally, we selected the LightGBM model as our final model as it seems least overfit and performed similarly to the others.

XGBoost example:

```
# Scale pos weight formula = (row count - count of least po
# XGBoost Regression
xgb = XGBClassifier(random_state=42, scale_pos_weight=15)
xgb = evaluateModel(xgb, X_train, y_train, X_test, y_test)
```

```
TRAINING SET
              precision    recall  f1-score   support

           2       0.00      0.00      0.00       665
           3       0.56      0.15      0.24      3326
           4       0.52      0.96      0.67      5081
           5       0.88      0.01      0.01      1272

    accuracy                           0.52     10344
   macro avg       0.49      0.28      0.23     10344
weighted avg       0.54      0.52      0.41     10344

[[   0  100  565    0]
 [   0  515 2810    1]
 [   0  227 4854    0]
 [   0   73 1192    7]]

Testing SET
              precision    recall  f1-score   support

           2       0.00      0.00      0.00       222
           3       0.39      0.11      0.17      1109
           4       0.50      0.93      0.65      1694
           5       0.00      0.00      0.00       424

    accuracy                           0.49      3449
   macro avg       0.22      0.26      0.20      3449
weighted avg       0.37      0.49      0.37      3449
```

LightGBM example:

```
# LGBMClassifier Regression
lgb = LGBMClassifier(random_state=42)
lgb = evaluateModel(lgb, X_train, y_train, X_test, y_test)
```

```
TRAINING SET
              precision    recall  f1-score   support

           2       0.76      0.06      0.12       665
           3       0.61      0.40      0.48      3326
           4       0.58      0.90      0.70      5081
           5       0.75      0.08      0.15      1272

    accuracy                           0.59     10344
   macro avg       0.67      0.36      0.36     10344
weighted avg       0.62      0.59      0.53     10344

[[  42  196  421    6]
 [   6 1325 1981   14]
 [   5  477 4584   15]
 [   2  181  983  106]]

Testing SET
              precision    recall  f1-score   support

           2       0.29      0.02      0.03       222
           3       0.37      0.24      0.29      1109
           4       0.51      0.80      0.62      1694
           5       0.22      0.03      0.05       424

    accuracy                           0.48      3449
   macro avg       0.35      0.27      0.25      3449
weighted avg       0.41      0.48      0.41      3449
```

Our primary issue with regard to machine learning was data completeness. Most business attributes have values like Na, No, or None so our models have weak prediction power. Secondarily we had technical issues related to bugs in recent versions of XGBoost. These issues are documented as follows:

- https://stackoverflow.com/questions/71996617/invalid-classes-inferred-from-unique-values-of-y-expected-0-1-2-3-4-5-got/72084851#72084851
    - Resolved by down-grading XGBoost version
- https://github.com/dmlc/xgboost/issues/2334
    - Resolved by punting and using GradientBoost instead; no mocel really works on this data set anyway.

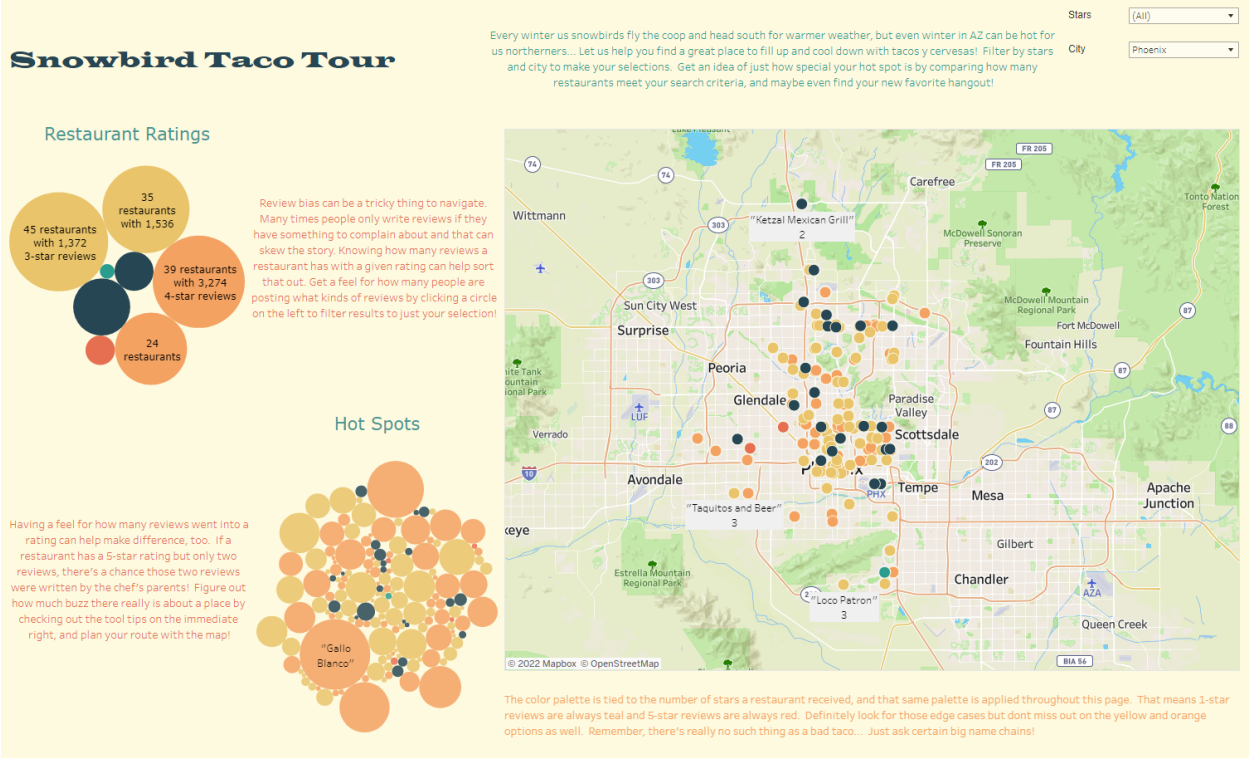
Our most important lesson is that machine learning is a voracious beast when it comes to consuming data. To train an effective model, data must be complete and at volume. Since we didn't have complete enough data to thoroughly train, we decided that as long as all the

functionality is there and works we were going to have fun with it. Check out our predictions for how a restaurant in some hot destinations might perform on our website!

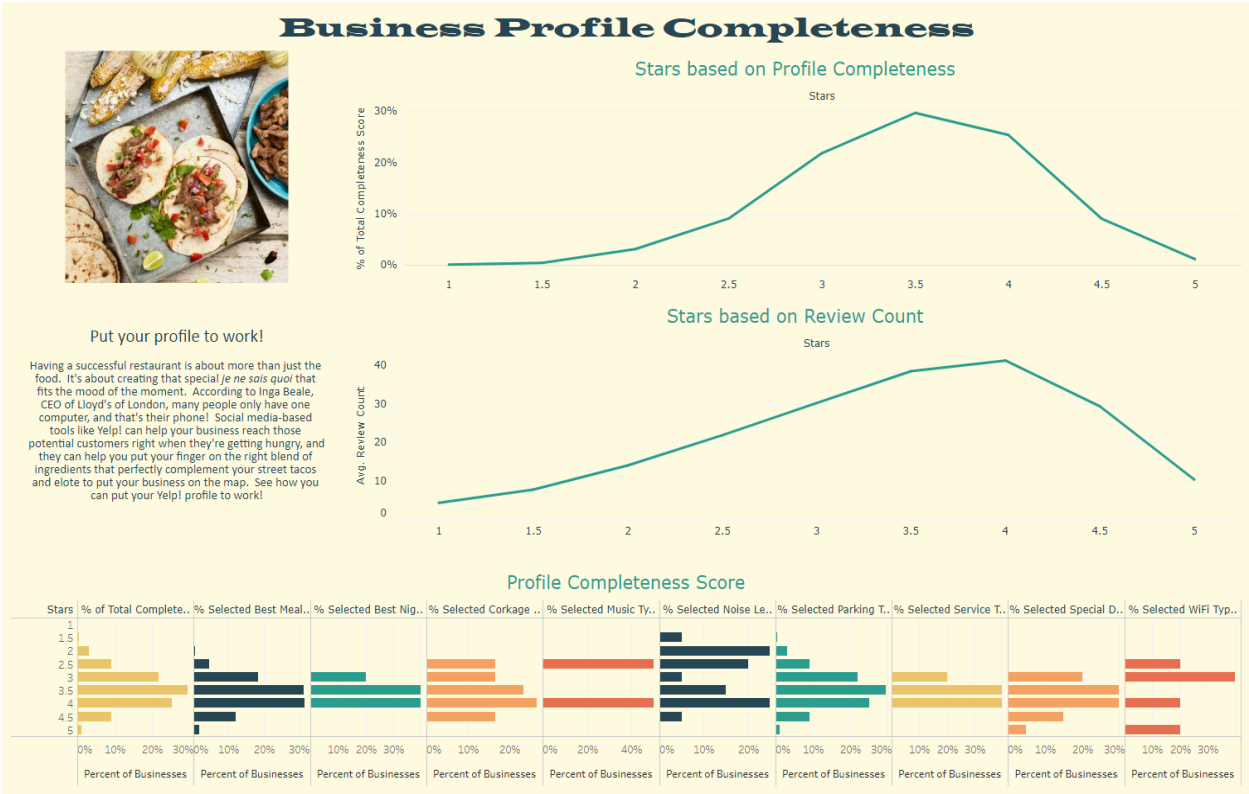
Tableau

Our concept, Snowbird Taco Tour, is based on the notion that it's common for people to flee the cold weather in the northern states, like Minnesota, for warmer weather in Phoenix, Arizona. When we head south we have to find a great spot to eat and hang out with friends and family, and why not make it easy: everybody loves tacos!

We identified a couple inspiration dashboards on Tableau Public, and we attempted to use certain design elements from them such as using a unified and cohesive color palette, page organization and use of whitespace.



Our first dashboard is customer focused, with the goal of making it easy to find a great locale for a tasty south-of-the-border style plate of goodness. Our dashboard is organized to be read in two columns. The left panel groups restaurants by how many stars they've received in the top graph, and in the bottom left panel lets them filter down to a single "hot spot" if they desire. Clicking the dots applies a consistent filter across the board, so to see all 4-star reviews the user clicks a dot on the top panel and all 4-star reviews are revealed on the map as well as in the "hot spots" panel on the bottom left. From there, the user can pick a restaurant based on location on the map or they can further explore options on the Hot Spots pane. In Hot Spots, the dot is sized based on the count of reviews that contributed to that restaurant's scoring.



Our second dashboard is business focused. More customers in the door means more customers to leave reviews. The better the reviews of course the more customers will come in, but that doesn't appear to be the end of the story. Here, we highlight how important a complete profile is to businesses by comparing how complete a business profile is to how many reviews they receive. Correlation does not equal causation, of course, but it seems likely that the users visiting these establishments are doing so based on the information they can find on their phones, and that they're using the same channels to share their experiences. In this digital age, savvy business owners will tap into this information and respond to changing trends.

Web Design

The color palette for the webapp and tableau dashboards was inspired by the colors of popular fiestas like Cinco de Mayo and Dio de los Muerto. We customized a popular color palette to translate our inspiration into something both easy on the eyes and with enough variety to be able to create a good sense of hierarchy.

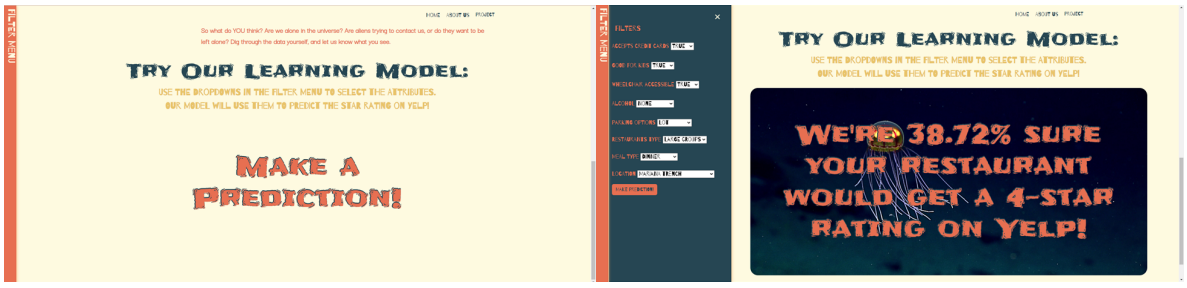
Another way we added visual hierarchy to the webapp was by using google fonts. For titles and major headings we used Frijole, for menus and subheadings Barrio, and for paragraphs Darker Grotesque. These three fonts really brought the fiesta theme to life and make it easier for users to navigate each webpage.

The UI for our webpage was loosely inspired by a webflow template: Delice - Restaurant Website Template, however we decided to not use a template for the HTML or CSS. Bootstraps, W3Schools and stack overflow were referenced often when we had roadblocks. We also used an AOS plugin to add scrolling animations to make information on our webpages more easy to navigate.

For site navigation we used a fixed top navigation bar so users can navigate from to other pages of the site without having to scroll up to the top. Our webapp ended up having seven pages, so to prevent the navigation bar from looking too busy we added a “Project” drop down menu. If users are interested in interacting with the different features of our web app they can access them through the drop down, while users just interested in the basic concepts (the executive summary or information about the developers) aren’t bogged down by the clutter. When a user selects a webpage from the dropdown the jumbotron text changes to show the navigation level of the active web page.

The home page features an executive summary of the project as well as links to our written analysis and github page. The about me page contains basic information about the developers of this project. The tableau pages (Snowbird Taco Tour (Tableau), Business Profile Completeness(Tableau)) were embedded into their respective web pages. We opted to just embed the dashboards into the webpage versus using the API and spent our time working on functionality in the machine learning and SQL sections.

The machine learning and data engineering pages have a side navigation that includes the filters for the user to choose from. Our model takes in latitudes and longitudes as one of the features, but most users don’t just know latitudes and longitudes off the top of their head. To improve the user experience we used interesting locations and had the model use their latitudes and longitudes for the model. On the machine learning page when the user clicks the “make prediction” button after choosing their filters the model predicts a rating. The rating and the probability of its accuracy are displayed on an image that is dynamically chosen based on the location the user chose.



The data engineering page includes a plotly bubble chart of the average ratings of restaurants that shows the distribution of the ratings based on the attributes the user filtered by. There is also a table of the restaurant information based on the query.

Conclusion

Through this experience from concept to reality, we learned that not everything will line up perfectly but it was the journey and overcoming the hurdles that brought our Snow Bird Tour to life. Our main limitation was that we did not have as complete data for our original goal of the Great North America Taco Tour and even as we adjusted to the Snow Bird Tour, we still stumbled into unanticipated problems but did what we could with the limitations of data. We believe that as time and more data is collected through Yelp and other review platforms, the future work on this project will be able to showcase as we originally intended it to be. By placing point A to point B and generating a list of tacos and burritos to eat and experience along the way.

Works Cited:

Data:
https://www.kaggle.com/code/jagangupta/what-s-in-a-review-yelp-ratings-eda/data
Website Inspiration:
https://webflow.com/templates/html/delice-restaurant-website-template
Fiesta image:
https://tampabaydatenightguide.com/wp-content/uploads/sites/2/2019/04/House-of-Mexico-Dancers_Richard-Benton_2014-e1555451057268.jpg
Coolors Palette:
https://coolors.co/264653-2a9d8f-fefae0-efb366-f4a261-e76f51
Fonts:
https://fonts.google.com/specimen/Frijole?query=frij
https://fonts.google.com/specimen/Barrio?query=barrio
https://fonts.google.com/specimen/Darker+Grotesque?query=darker
Scroll Animation:
https://github.com/michalsnik/aos
Side nav:
https://www.w3schools.com/howto/howto_js_sidenav.asp
Dynamically change bg image
https://sebhastian.com/javascript-change-background-image/
Search selection box:
https://makitweb.com/make-a-dropdown-with-search-box-using-jquery/