**Full Data Modelling:**

Input: covid\_restaurant\_integrated\_data

·      Removed first column (no name), ‘id’, ‘Name’

·      Moved ‘is\_closed’ to last column

·      Decided to remove ‘review\_count’ variable because we can’t pass this in as an input variable since restaurant owners can’t predict how many yelp reviews they will get

·      Removed zipcode – should have been an indicator variable, and overlaps with other variables

·      Removed – FIPS\_Code

·      Removed Region Name – duplicates of ‘Area Name’

·      Removed State\_y – duplicates of ‘State\_x’

Output: covid\_restaurant\_full

#Data profiling

·      Removed duplicates

·      Removed ‘sushi’ -  high correlation with ‘japanese’

Output: covid\_restaurant\_full\_post\_profiling.xlsx, full\_data\_profiling.html

#outlier detection

·      Removed record 12159 - outlier that its sfh\_sales\_price is 2,405,436 where the next highest values are 1Mil and below

Output: covid\_restaurant\_full\_post\_outlier.xlsx

#correlation / multicollinearity checks:

·      Removed covid\_substantial - -0.95 corr with covid\_moderate

·      Removed redfin\_sale\_price – 0.99 corr with zillow\_sfh\_sale\_price

·      Removed redfin\_homes\_sold – 0.77 corr with Zillow\_sfh\_sale\_price

·      Removed Urban\_influence\_code\_2013 – 0.94 corr with Rural\_urban\_continuum\_code\_2013 ·      Removed Med\_HH\_Income\_Percent\_of\_State\_Total\_2019 – 0.94 corr with Median\_Household\_Income\_2019

·      From Corr extract:

* Removed Median\_Household\_Income\_2019 – correlation with unemployment\_rate\_2020, covid\_moderate, and zillow\_sfh\_sale\_price
* When calculating VIF, noticed Completeness\_pct had a really high VIF, removed and numbers looked better across the board.
* In second part of analysis, noticed high correlation between redfin days on market and Zillow\_sfh\_sale\_price, so removed redfin\_days\_on\_market

 ·      Removed all those observations with grocery = 1 or convenience = 1, and dropped the grocery and convenience columns

Output: covid\_restaurant\_full\_post\_VIF.xlsx

\*Due to the way we’re taking inputs, went back and made a version with FIPS Code/Zip Code, State, Area Name indicators up to the point where we have the model input dataset

**Feature Importance for final XGB Model**

{'delivery': 118.22736358642578,

'hotdogs': 23.691598892211914,

'chicken\_wings': 11.445265769958496,

'mexican': 10.166533470153809,

'japanese': 9.716278076171875,

'newamerican': 8.87838077545166,

'italian': 8.13509464263916,

'delis': 7.225705146789551,

'sandwiches': 7.189199924468994,

'chinese': 6.735775470733643,

'steak': 6.703718662261963,

'pizza': 6.320083141326904,

'mediterranean': 6.236205577850342,

'Total\_Reviews\_Zip': 6.148074150085449,

'coffee': 6.118718147277832,

'restaurant\_reservation': 6.083000659942627,

'covid\_high': 5.842723846435547,

'asianfusion': 5.827420234680176,

'cafes': 5.826242923736572,

'vegetarian': 5.372157573699951,

'icecream': 5.344324111938477,

'vegan': 5.254461765289307,

'Rural\_urban\_continuum\_code\_2013': 4.935182571411133,

'bbq': 4.916090488433838,

'tradamerican': 4.868152141571045,

'juicebars': 4.867558002471924,

'sportsbars': 4.6455793380737305,

'rating': 4.194856643676758,

'breakfast\_brunch': 4.0786919593811035,

'price': 4.0316572189331055,

'burgers': 3.9390032291412354,

'desserts': 3.8247158527374268,

'bars': 3.68076229095459,

'salad': 3.4744248390197754,

'seafood': 3.4693443775177,

'business\_ratio': 3.3579416275024414,

'pickup': 3.290992498397827,

'Average\_Rating\_Zip': 3.1401796340942383,

'zillow\_sfh\_sale\_price': 3.1173629760742188,

'total\_restaurants\_zip': 2.980626106262207,

'Unemployment\_rate\_2020': 2.755488634109497,

'covid\_moderate': 2.6175262928009033,

'bakeries': 2.616180896759033,

'Average\_Price\_Zip': 2.592618942260742,

'SizeRank': 2.5601718425750732}

**Arizona Only Modelling:**

Removed these features:

·      Removed first column (no name), ‘id’, ‘Name’

·      Region Name – duplicates of ‘Area Name’

·      State\_y – duplicates of ‘State\_x’

·      FIPS\_code – represented by area\_name

Filter down to AZ data and did data profiling:

·      Removed ‘State\_x’ – constant value “AZ”

·      Removed ‘covid\_high’ - has constant value "0"

·      Removed ‘covid\_low’ – has constant value “0”

·      Removed ‘completeness\_pct’ – has constant value “0”

·      Removed 16 duplicated rows

·      Removed ‘bars’ – high correlation with ‘sportsbars’

·      Removed ‘sushi’ -  high correlation with ‘japanese’

Output: covid\_restaurant\_AZ\_post\_profiling.xlsx, covid\_restaurant\_AZ\_post\_profiling.html

#outlier detection

·      Decided to remove ‘review\_count’ variable because we can’t pass this in as an input variable since restaurant owners can’t predict how many yelp reviews they will get

·      Removed ‘area name’ – direct mapping with FIPS\_Code

#correlation / multicollinearity checks:

·      redfin\_homes\_sold pretty much accounted for everything else besides Rural\_urban\_continuum\_code\_2013, Urban\_influence\_code\_2013, Metro\_2013, and Unemployment\_rate\_2020

o   so only keep these + redfin\_homes\_sold

·      Rural\_urban\_continuum\_code\_2013 0.91 corr with Urban\_influence\_code\_2013 so kept only Rural\_urban\_continuum\_code\_2013

·      Unemployment\_rate\_2020 -0.93 corr with Metro\_2013 so kept only unemployment\_rate\_2020

·      Redfin\_homes\_sold is 0.98 corr with business\_ratio so removed business\_ratio

·      Total\_resturants\_zip is 0.94 corr with total\_reviews\_zip so removed total\_reviews\_zip

·      Removed zipcode – should have been an indicator variable, and overlaps with other variables

·      Removed urban\_continuum\_code\_2013 – highly correlated (0.71) with umemployment\_rate\_2020