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

·      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 Rural\_urban\_continuum\_code\_2013 – 0.94 corr with Urban\_influence\_code\_2013

·      Removed Med\_HH\_Income\_Percent\_of\_State\_Total\_2019 – 0.94 corr with Median\_Household\_Income\_2019

·      From Corr extract:

o   Removed Median\_Household\_Income\_2019 – correlation with unemployment\_rate\_2020, covid\_moderate, and zillow\_sfh\_sale\_price

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

**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