Child Care Deserts and Venue Data in the State of Colorado

Applied Data Science Capstone Project

Michelle Salvador

A.Introduction

A child care desert (CCD)is a place where child care for children under five is not readily available. This project focuses on exploring CCDs in Colorado and their relationship with businesses in the same location. Providing child care itself is a business and learning what variables and predictors affect their site (if any) such as location and proximity to other business types is beneficial. For people living in CCDs this can become a problem as parents will encounter difficulties finding someone to look after their child while they have to go to work. In addition to the high cost of child care, having to commute (possibly by public transport) can worsen a person's economic hardship by having to choose between work and caring for their child.

B. Background

The problems Child Care Deserts present

The following section outlines the problems CCDs cause to parents, guardians, caregivers, and policy makers.

B.1. Problems for Parents or Guardians (and Children)

Living in a **CCD** can present a problem to families with children under the age of 5. Without someone to care for their children, a 2015 poll conducted by the Washington Post showed that over 75% of mothers and 50% of fathers all over the United States had to decline a work opportunity and change or quit a job because of the scarcity of child care or paid leave.

According to CAP, 51% of people in the US live in a CCD, with low income people being at 55%. A report by the New America think tank found that child care costs on average $9,589 - more than the average in-state college tuition of $9,410! This leaves many parents unable to pay anything greater and places an additional economic stress to many families. To make the situation worse, for those who live in CCDs this means having to travel much longer distances to find childcare they can afford on top of the struggle to find any at all. Research by [5 in 3] shows that when early childcare is available the child's social, economic, health, and education outcome improves.

B.2. Problems for Child Care Providers

In most situations the demand for a service increases supply of that service in the market. However, childcare presents a series of problems to providers in terms of business economics and liability. Since parents are often unable to afford higher childcare costs, providers make low profits. Caring for infants is also highly liable as young children are fragile and mistakes in care can have mortal consequences. One example being Sudden Infant Death Syndrome (SIDS) where babies can die without warning if they are placed to sleep incorrectly. Infant-care guidelines and complex regulations are put in place by the government to reduce these dangers but they cause an uninviting business situation along with the low profit margins. The report by CAP [3] has seen an increase of childcare providers in affluent suburbs where parents are willing to pay greater fees, but demand continues to outpace childcare options in all other places.

B.3. Problem for Policy Makers

With the problems outlined above for parents, guardians, and childcare providers, CCDs are a nonpartisan problem for policy thinkers [4]. As roads and bridges are thought of as infrastructure to support the American economy, investing in childcare is a necessary service that is needed by citizens that is not being met.. Further understanding the cause for lack of childcare can provide insight into policy solutions and funding in areas most needed by the population.

C. Project Goal Description

The aim of this project is to study the CCD found in Colorado and how they relate to the types of businesses that are present in the corresponding ZIP code. Learning more about the characteristics of CCD when it comes to why this highly demanded business is not thriving can be of interest to parents, childcare business providers, and policy makers wishing to improve the circumstances.

I choose Colorado since this is the state I reside in and according to a CAP report [5], 45% of residents live in a CCD using 2014 census estimate data, by the ZIP code definition.

Using foursquare, census data, and a childcare database provided by the Colorado government I wish to explore the following:

1. Are there certain types of businesses that occur with higher frequency in Child Care Deserts? If so, what type? For example I hypothesise that ZIP codes with high amounts of adult businesses (liquor stores, marijuana dispensaries, adult venues etc) will have a higher likelihood to exist in CCDs. However, I believe it is out of the scope of this project knowing if the existence of these types of businesses cause CCD.

2. Is there a correlation with low business density to CCD? This could indicate as to how easy it is to start a business at that location or the population density that works in that ZIP code. I hypothesise that ZIP codes with low business density will be more likely to be a CCD.
3. Can a model be created using foursquare business data to predict if a census tract will be a CCD? The various types of machine learning techniques learned in this course can be applied to attempt the highest accuracy. I predict a model can be built with accuracy greater than random coin toss (50%).

D. Data

In order to tackle the goals of this project, first we must define what CCD is as there two definitions available [5][3].One using ZIP code and another census tract. As the data for Colorado and geo-location was most readily available and up-to-date for the ZIP code definition, I will use the ZIP code to define CCD.

D.1. Child Care Deserts: A definition

According to the Center for American Progress (CAP) a **Child Care Desert** (**CCD**) [5] is a **ZIP code** defined as having **both** the following characteristics:

1. At least **30 children** under the age of 5
   AND
2. Has either **no child care centers** or **so few centers** that there are **more than three times as many children** under age 5 as there are spaces in centers.

This definition is given since children under the age of five are usually not enrolled in school and according to census averages, one third of children are “regularly in nonrelative care.” [8 lynda]

D.2. Data Sources and Cleansing

The Colorado Information Marketplace was a wealth of data. From it, I got the ***Census Zip Codes in Colorado 2017*** (*Census*) [7] data for ***Colorado Licensed Child Care Facilities Report*** (*Facilities*) [8]**.** Using **Foursquare** I gathered the venues within an approximate radius equal to the ZIP code area. In an effort to use the most recent data I choose the datasets available with the most recent versions available.

D.2.i. Census Zip Codes in Colorado 2017

This dataset contains demographic information of each ZIP code in colorado. The relevant information I used was the ZIP code (**zipcode**), total population (**pop**), the population of children under 5 (**ageless5**), and multipolygon data. All other columns were dropped. ZIPs with no population were dropped. These are P.O. box ZIPs.

Using the multipolygon data and ArcGIS [9], an online geographic information system software, I calculated the centroid of each polygon and found the **latitude** and **longitude** for each ZIP as these were unavailable online.

Learning how to use the geojson area package, I found the total area of the polygon contained by the ZIP and calculated the radius (**radius_meters**) of a circle that would be enclosed in a square having the same area of the ZIP polygon to the lowest hundred. Though overlap would be present with other ZIPs, this radius variable solved the problem of need for an area variable to calculate business density. Since there are ZIPs with areas as small as a few hundred meters and as large as over 20,000 meters - using a constant radius in Foursquare would not be acceptable.

The resulting pandas data frame looked like this (508 rows x 6 columns):

	zipcode	ageless5	pop	Latitude	Longitude	radius_meters
0	80476	11	196	39.695974	-105.731550	6100
1	80477	0	146	40.348242	-106.926910	100
2	80478	27	1625	40.001670	-105.868600	5200
3	80479	0	5	40.041288	-106.855700	8500
4	80480	89	1342	40.621621	-106.244570	31400
5	80481	6	601	40.106130	-105.480440	6800

D.2.ii. Colorado Licensed Child Care Facilities Report

This dataset contained the names of all licensed child care facilities in Colorado with their address, type of service provided, and child capacity. Facilities that only provided 'School-Age Child Care Center' were dropped since we are targeting children under 5 who are not school aged. I then grouped the facilities by ZIP and found the sum of their capacity (**child capacity**). I merged this data with the previous data frame and using the definition of CCDs labeled each ZIP as **Desert (label 1)** (met definition of CCD), **Few_Child (label 2)**(if there were less than 30 children), and **Not_Desert (label 3)** (over 30 children and enough capacity requirements).

The resulting data frame was as follows (376 rows x 11 rows)

zipcode	child capacity	ageless5	pop	Latitude	Longitude	radius_meters	Desert	Few_Child	Not_Desert	Classification
80477	203.0	0	146	40.348242	-106.92691	100	0	2	0	2
80478	39.0	27	1625	40.001670	-105.86860	5200	0	2	0	2
80480	15.0	89	1342	40.621621	-106.24457	31400	1	0	0	1
80481	486.0	6	601	40.106130	-105.48044	6800	0	2	0	2
80483	36.0	56	660	40.149432	-106.90681	7500	0	0	3	3

Interestingly there were ZIPs with no children yet a large capacity for childcare.
To visualize the amount of children and total population who lived in CCD I created the following pie charts.

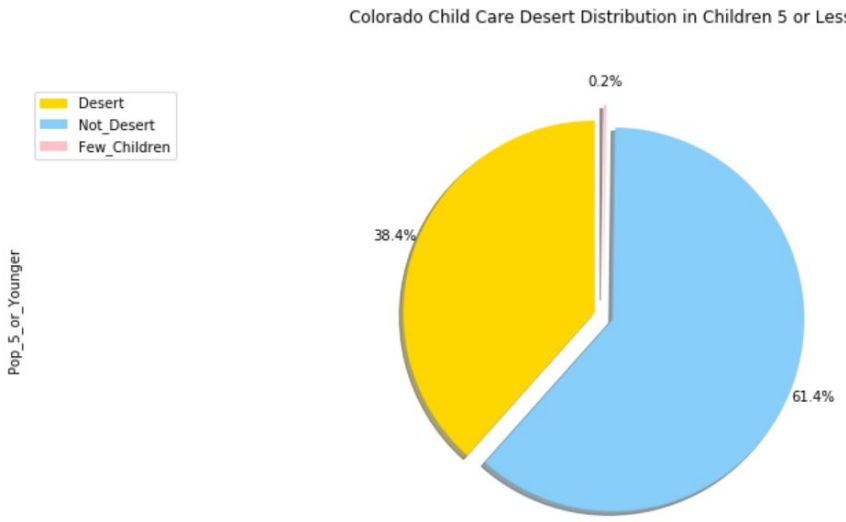

Fig.1

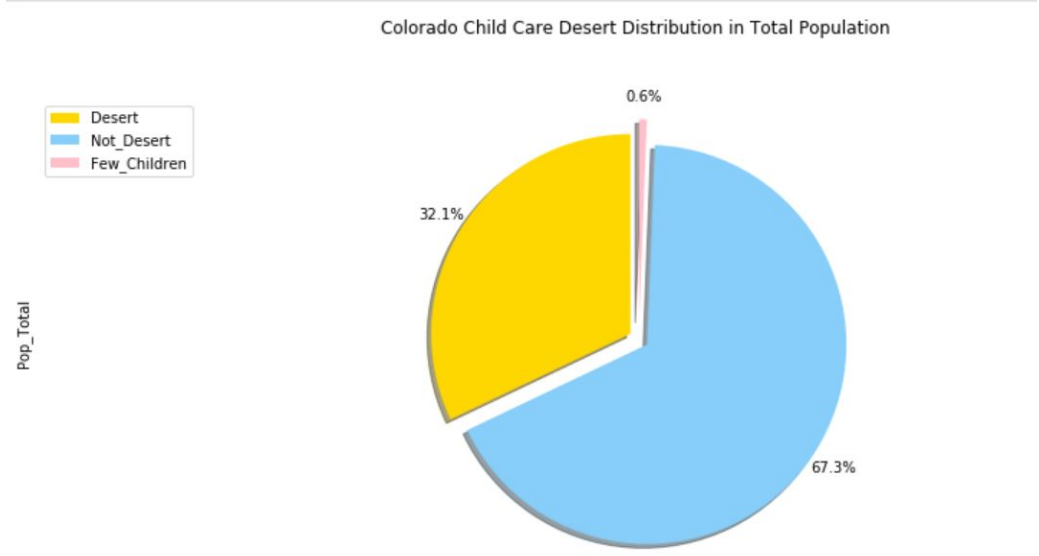

Fig.2

D.2.iii. Foursquare API

Using the Foursquare API and the radius value of each ZIP I calculated from the census data, I collected the number of venus in a ZIP and the top categories in each. Though I did place an upper limit of 100 venues, I did not see much “clipping”. By this I mean that if the venue count for the ZIPs where 100 for most, then this means I would need to increase the venue limit count. I did not see this.

80033	42	42	42	42	42	42
80045	13	13	13	13	13	13
80102	20	20	20	20	20	20
80103	9	9	9	9	9	9
80104	82	82	82	82	82	82
80105	7	7	7	7	7	7
80106	6	6	6	6	6	6
80107	23	23	23	23	23	23
...	...	...	...	...	...	...
81503	100	100	100	100	100	100
81504	41	41	41	41	41	41

Fig. 3

432 unique categories where found. A wide range. 364 ZIP in total. Just to make sure not too many ZIPs were maxing at 100, I created a Histogram showing the number of ZIPs that had a certain number of venue counts.

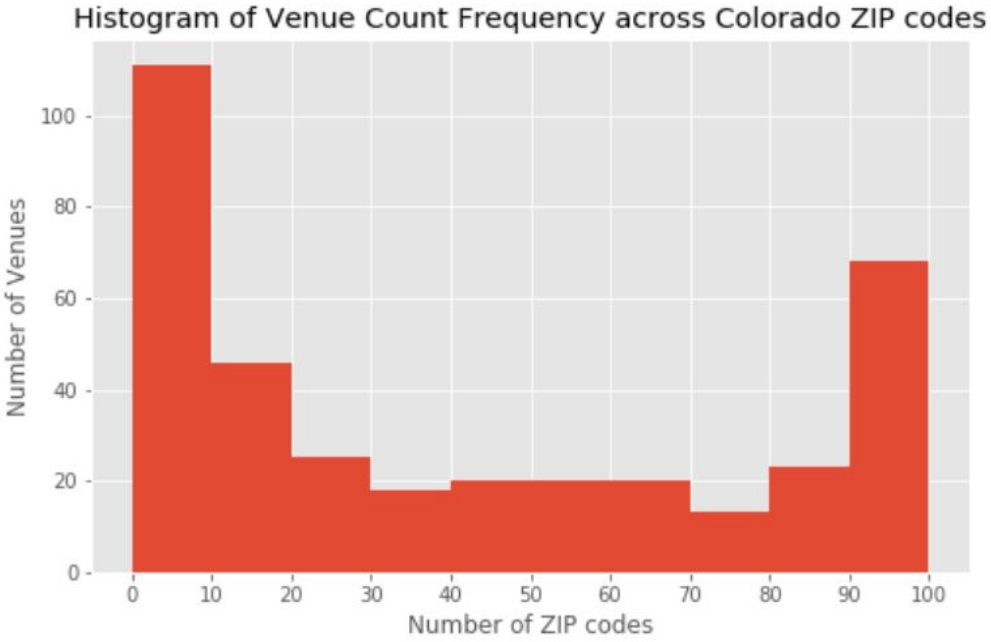

Fig.4

Interestingly when I ran a distortion score using the elbow method to decide how many K clusters the graph showed no elbow. Also when I plotted the radii on the map I noticed how there was much overlap in the areas being sampled for each ZIP. Therefore I decided to divide the radius in half to reduce this overlap but still sample a larger area if the ZIP was larger.

The histogram changes as follows. Because of the reduction of the radius used, the frequency of ZIPs with lower venue counts increased.

E. Data Exploration

E.1 Venue Count vs Zip Code Variables

I want to see if there are any venues that occur more in relation to the following variables:
population count (**pop**), area (**radius_meters**), or number of children under 5 (**ageless5**). First I want to see how these are distributed across the CCD classifications.

Here I use a box plot to show the average and the outliers of how many venues are found in each zip code according to classification. The figure shows the venue count with the same axis of 100 as the max.

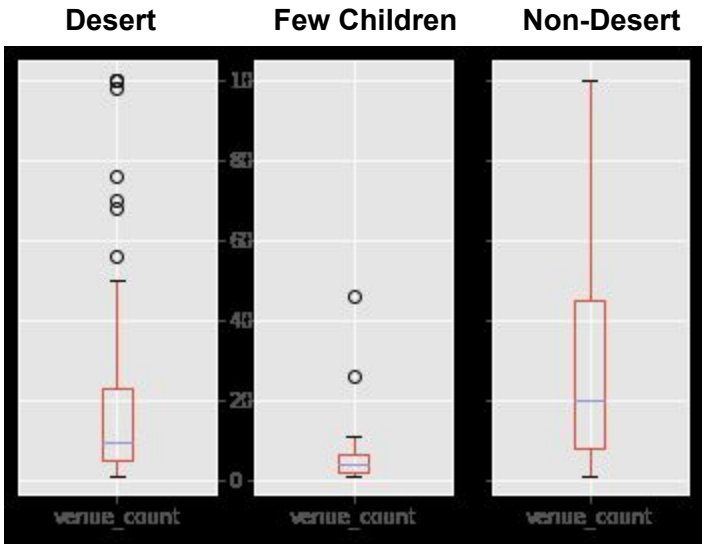

Fig.5 Venues Per Zip Code

Desert	Few Children	Non-Desert
count 104 mean 18.79 std 22.4 min 1.0 25% 5.0 50% 9.5 75% 23 max 100.0	count 39 mean 5.6 std 8.035 min 1.0 25% 2.0 50% 4.0 75% 6.5 max 46.0	count 204 mean 30.245 std 29.22 min 1.0 25% 7.8 50% 20.0 75% 45.0 max 100.0

There are more venues in the ‘Not Desert’ class overall, but this could be because there is a greater area and or more people. Now I’ll explore if this is still true if we factor in population.

The following box plot shows the proportion of venues divided by the population of each zip code within each CCD class - all plotted on the same axis with a max value of 1.

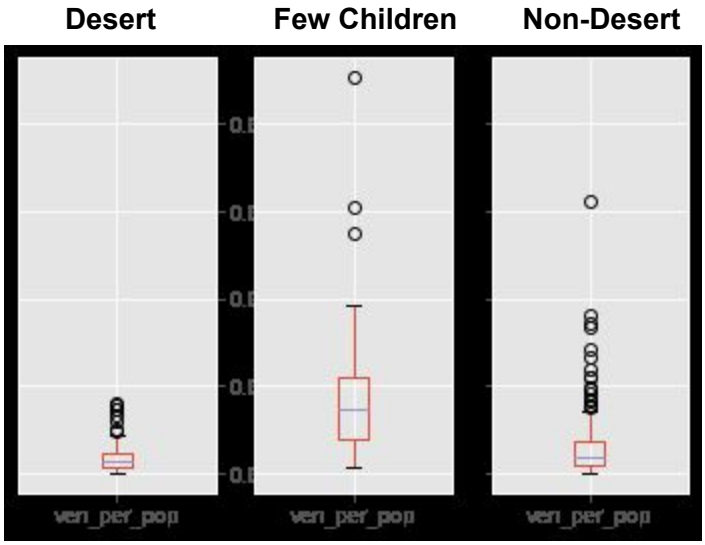

Fig.6 Venues Per Population

Desert	Few Children	Non-Desert
count 104 mean 0.002 std 0.002 min 0.0001 25% 0.0006 50% 0.001 75% 0.002 max 0.008	count 39 mean 0.0093 std 0.009 min 0.0008 25% 0.004 50% 0.0073 75% 0.0109 max 0.045	count 204 mean 0.003 std 0.0036 min 0.00006 25% 0.001 50% 0.002 75% 0.0036 max 0.031

As expected we see that the ‘Few Children’ has a higher ratio of venues per person probably because zip codes with this classification serve as rest stops or attractions for other people than the people who actually live there. We can observe that there are more venues per person in the Non-Desert versus the Desert. **Deserts have 61% of the venues per person than the Non-Deserts** have according to this comparison.
This is comparing across the population. What does the venue distribution look like if I look at the children under 5 population count?

The figure below shows what the ratio of each zipcode for each CCD class is when dividing the number of children per population thus showing the child-under-5 per person-over-5 proportion.

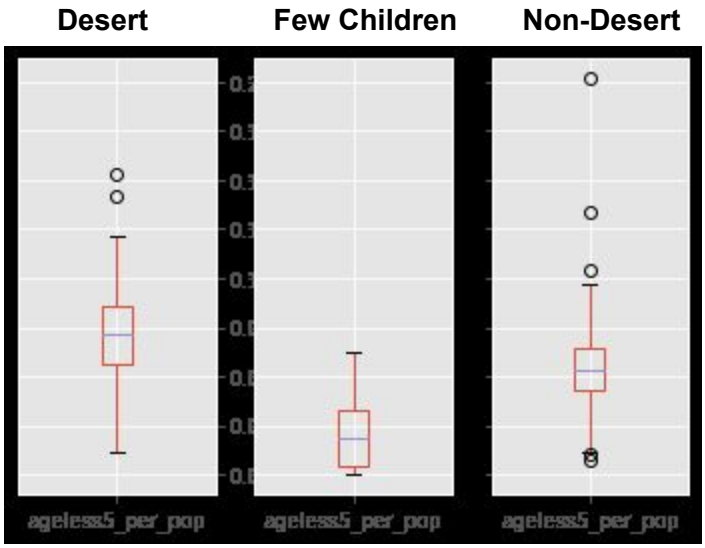

Fig.7 Ageless5 Per Population

Desert	Few Children	Non-Desert
count 104 mean 0.07 std 0.02 min 0.01 25% 0.06 50% 0.07 75% 0.09 max 0.1	Count 39 mean 0.009 std 0.009 min 0.0008 25% 0.004 50% 0.007 75% 0.01 max 0.04	count 204 mean 0.003 std 0.004 min 0.00006 25% 0.001 50% 0.002 75% 0.004 max 0.03

The average ratio in Deserts is about 130% higher than the Non-Desert and still higher than the Few Children class. Although it does not say for certain that there are more children in Deserts than Non-Deserts, because it could just mean that the children in those areas are on average older; there is a higher proportion of babies (younger than 5) to the rest of the population in Deserts than the other classes.

Instead about population now I will look how venues are distributed across land area.

Plotted vs the same axis, the figure below shows how many venues per meter are found in each zip code. The maximum value is 4.9e-05. The values are small as the venues are distributed across a larger land area.

Fig.8 Venues Per Meter

Desert	Few Children	Non-Desert
count 104 mean 7.3e-07 std 1.5e-06 min 1.2e-09 25% 1.5e-08 50% 1.1e-07 75% 9.5e-07 max 1.0e-05	count 39 mean 1.1e-07 std 4.3e-07 min 3.8e-10 25% 2.9e-09 50% 6.66e-09 75% 2.1e-08 max 2.5e-06	count 204 mean 1.1e-06 std 3.8e-06 min 7.78e-10 25% 2.2e-08 50% 1.4e-07 75% 9.8e-07 max 4.9e-05

On average there is 150% more venues per area in the Non-Desert class than Desert. However, this can mean that there is a larger land area in Deserts and hence the distribution. Therefore I will use the venue to population proportion since I am interested in seeing how Colorado serves the population count found in the area and not so much the land since Colorado having a wide range of landscapes such as mountains to plains can vary in the habitability of each meter.

Nonetheless, I am interested in seeing how the population density is distributed in the different CCD classes.

The figure below shows population per meter surveyed in this project plotted with the same axis across for comparison, with a max value at 1.0e-02.

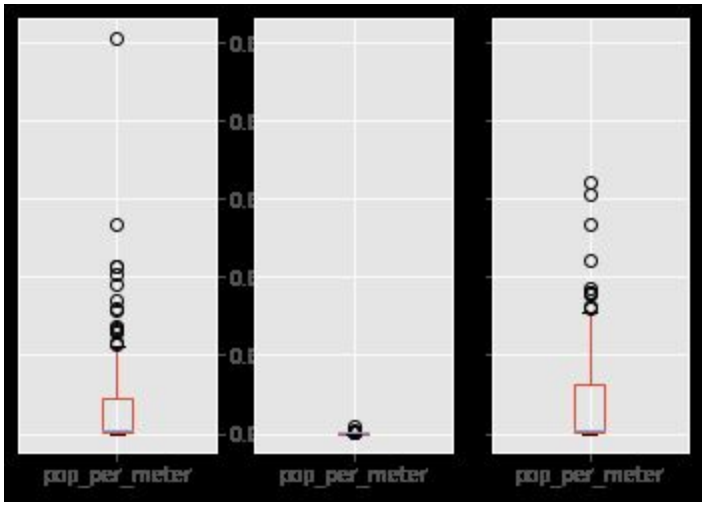

Fig.9 Total Population per Surveyed Square Meter

Desert	Few Children	Non-Desert
count 104 mean 8.2e-04 std 1.5e-03 min 3.7e-07 25% 9.7e-06 50% 8.6e-05 75% 9.0e-04 max 1.0e-02	count 39 mean 9.1e-06 std 3.14e-05 min 1.2e-07 25% 4.2e-07 50% 1.3e-06 75% 3.5e-06 max 1.9e-04	count 204 mean 7.6e-04 std 1.2e-03 min 3.4e-07 25% 7.7e-06 50% 5.3e-05 75% 1.3e-03 max 6.4e-03

There is on average 8.2e-04 people per surveyed meter in Deserts vs 7.6e-04 in Non-Deserts. **Seven percent more people per meter in Deserts than Non-Deserts**, hence a larger density.

Finally in exploring the data, I would like to see how venue count is distributed across babies (**ageless5**). From the definition of what CCD deserts are there fewer Child Care venues serving children under 5 in Deserts. Similarly, I hypothesize that the trend will be similar and there will also be fewer venues of other types ‘serving’ children under 5 in Deserts as well.

As observed from the figure below with the same axis and max value of 1, the Non-Desert class has many more venues per “baby” than any other class. The Few-Child in some cases has no children under 5, therefore in dividing venues per zero shows as a ratio of infinite max.

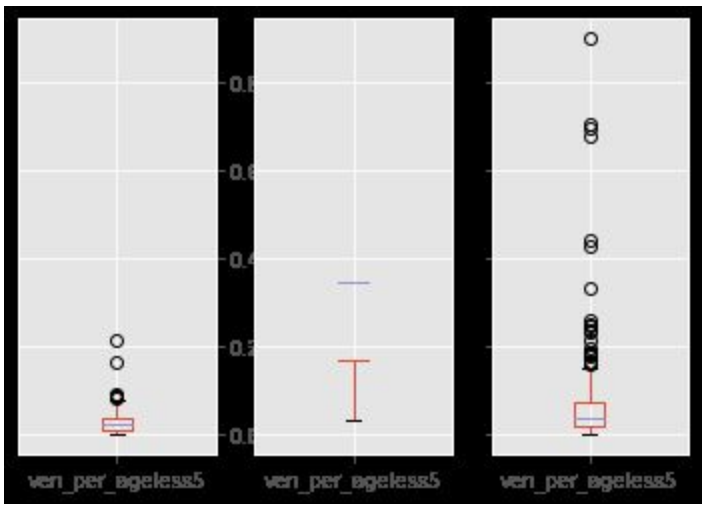

Fig.10 Venues per Child Under 5

Desert	Few Children	Non-Desert
count 104 mean 0.03 std 0.03 min 0.002 25% 0.009 50% 0.02 75% 0.04 max 0.22	count 39 mean inf std NaN min 0.03 25% 0.2 50% 0.3 75% inf max inf	count 204 mean 0.07 std 0.1 min 0.0004 25% 0.02 50% 0.04 75% 0.07 max 0.9

There are 133% more venues per ‘baby’ in Non-Deserts serving them.

Looking at the distribution of venues across each variable, now I am interested in the occurrence of venue types depending on the CCD classification. Are there certain types of venues that occur more frequently in Deserts, Non-Deserts, or Few Children?

E.2. Top Venue Frequency in Each CCD Class

Separating each class I found the top venues in each. Fast Food for Deserts, American Restaurant for Few Children, and Coffee Shop for Non-Deserts.

	Classification	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	1	Fast Food Restaurant	Mexican Restaurant	Pizza Place	Coffee Shop	Convenience Store	Sandwich Place	American Restaurant	Park	Grocery Store	Hotel
1	2	American Restaurant	Trail	Hotel	Campground	Bar	Restaurant	Resort	Coffee Shop	Ski Area	BBQ Joint
2	3	Coffee Shop	Pizza Place	Fast Food Restaurant	Mexican Restaurant	American Restaurant	Hotel	Sandwich Place	Grocery Store	Park	Convenience Store

From all the top 10 I wanted to compare and see how many of these there where in the other classes so I combined the lists to a unique list and plotted the amount in each.

Note CCD Classification in the figure is Desert(0), Few Child (1), and Non-Desert (2).

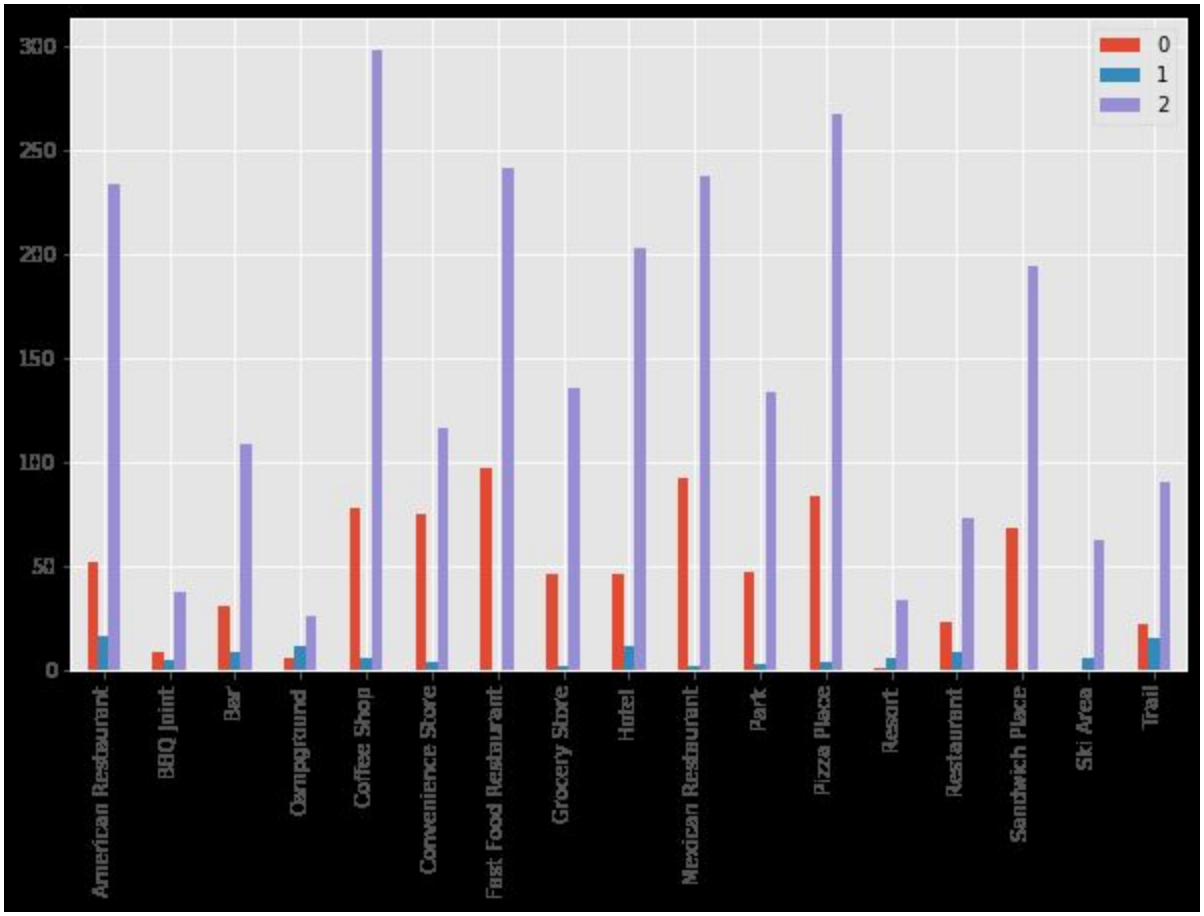

Fig.11 Top Venue Count in each CCD classification

Just plotting the number of venues in each category we can see the Non-Deserts have more of everything. Is this because there are more people? Let's scale for population:

Once again the Legend shows 0 for Desert, 1 for Few Child, and 2 for Non-Desert.

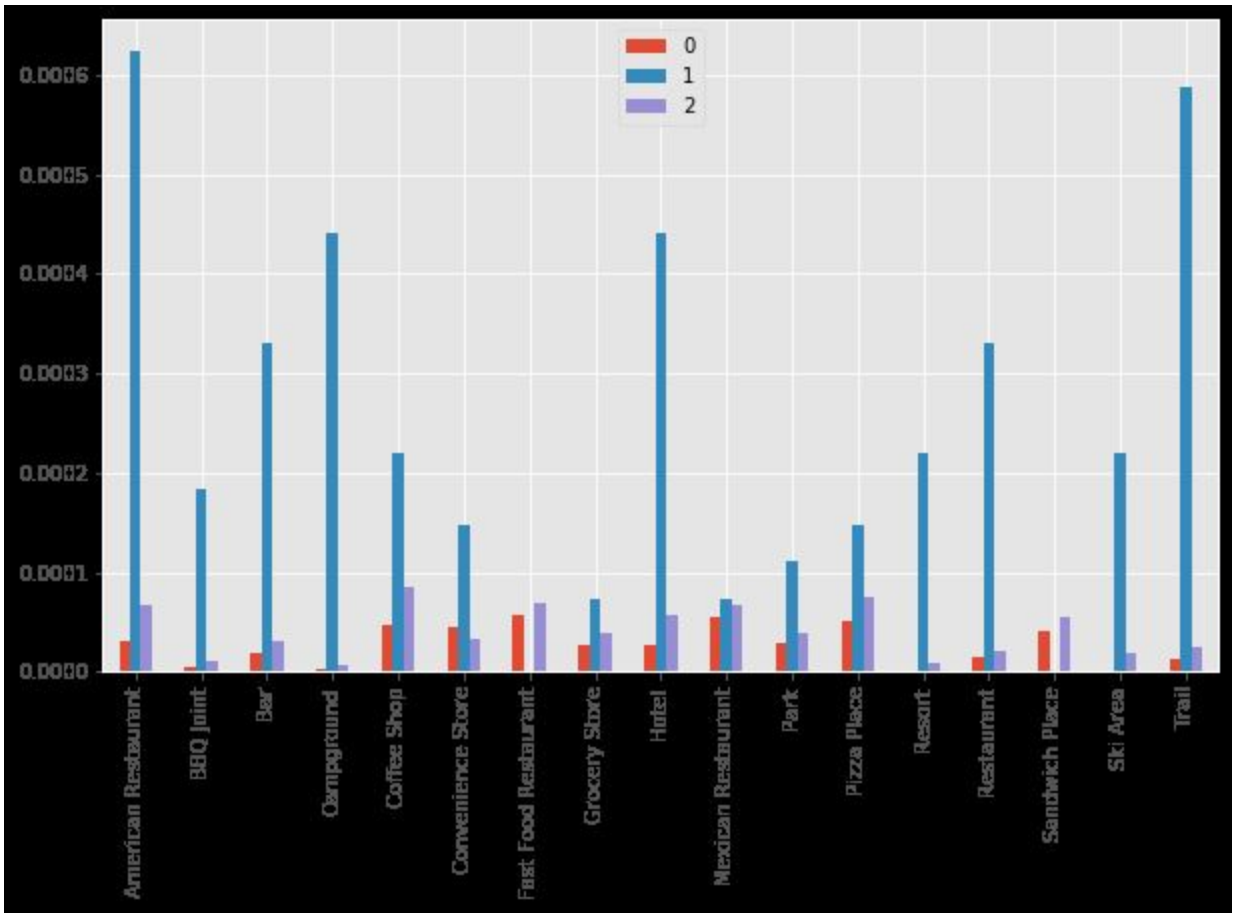

Fig.12 Top Venue Count per person in each CCD classification

Taking population into control we see the graph changes drastically showing that the Few-Child class has a large proportion of top venues per person. Indicating that one there are few people but also that there are homogenous venue types.

I would like to look more closely at what differentiates Deserts and Non-Deserts so I remove Few-Children and plot again.

The legend shows 0 for Desert and 2 for Non-Desert.

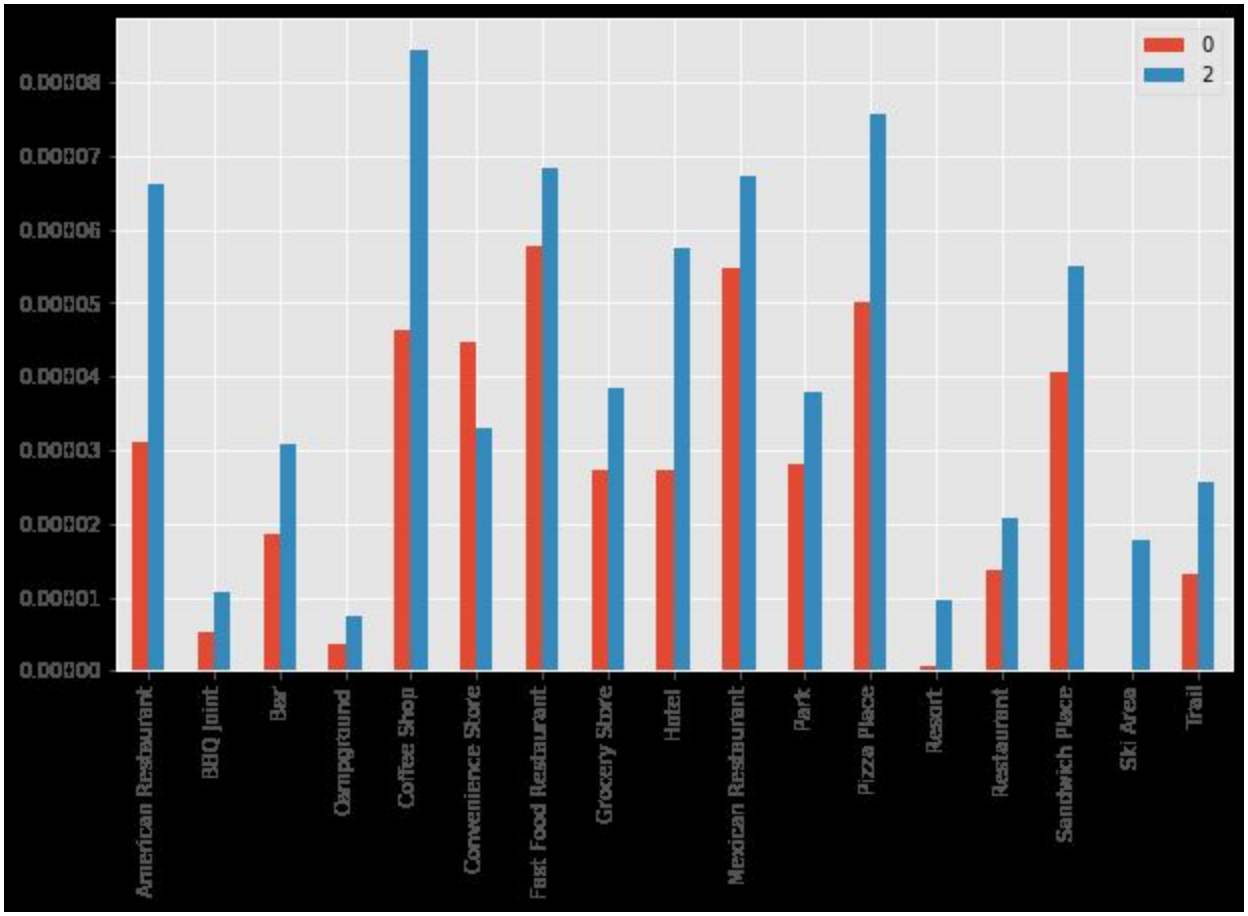

Fig.13 Top Venue Count per person in Desert and Non-Desert Classes

There are more venues per person in the Non-Desert once scaled for population as we saw previously. How many more times of each? I divide the number of venues per person of the Non-Desert by the Desert and plot to see. To be able to visually see the difference between the venues that are more abundant in each class I subtract 1. This causes venues with the same ratio in Desert or non-Desert will be 0 in the plot while those more abundant in the Non-Desert to have a positive value.

The figure below shows the ratio of venus per person of the Non-Desert areas divided by the Desert minus one to see which venues are more abundant in either class. Those with higher count in the Desert have a negative value. Only Convenience Stores are more abundant in the top venues of both classes.

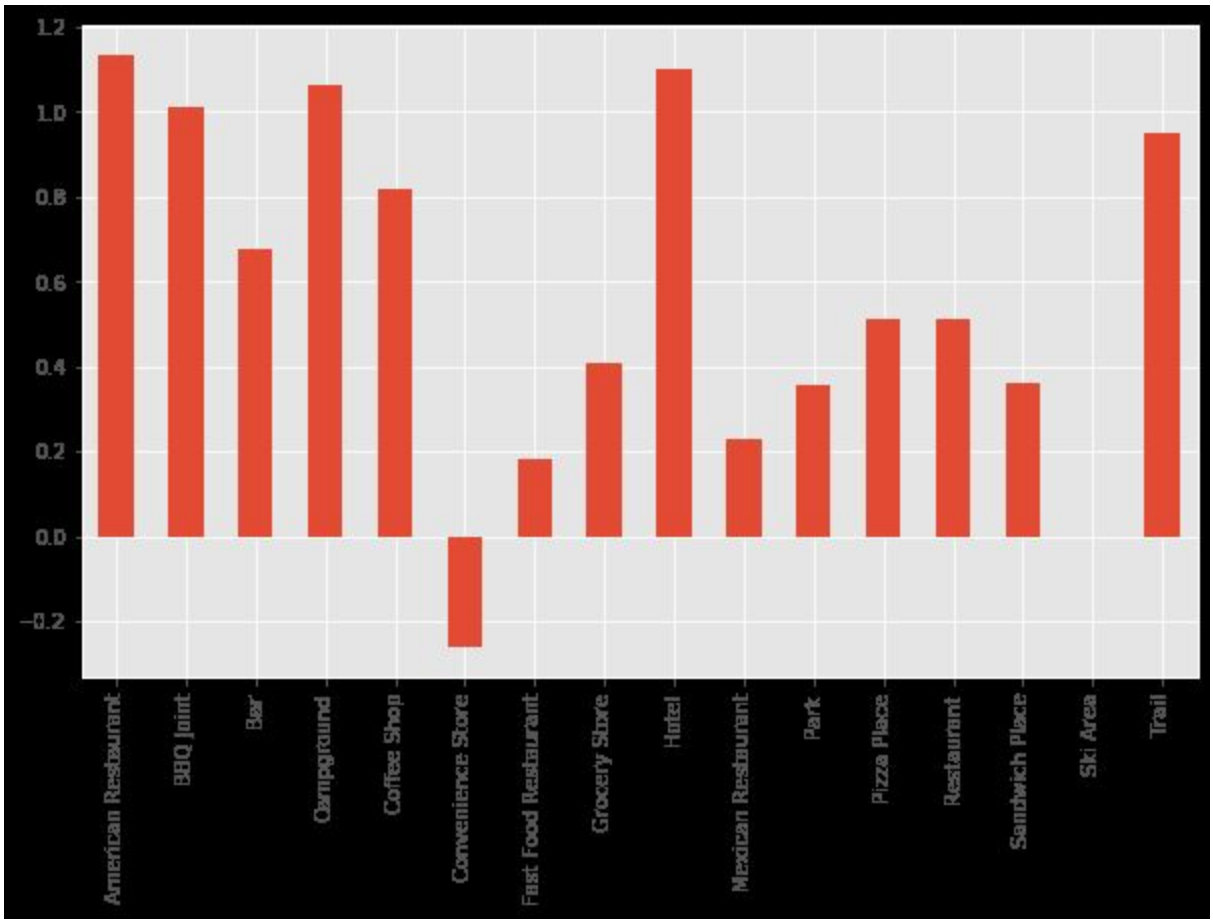

Fig.14 Top Venue Count per person of the Non-Desert divided by the Desert minus one.

The resort venue had 16 times more venues in the Non-Desert vs the Desert so I excluded this from the graph to focus on the relationship of other venues. Also Ski venues only exist in Non-Deserts and Few-Children so these are also not shown. **Hence, people who don't have Child Care readily available don't have resorts or Ski venues readily available either?**

From the graph other observations are:

- Although for example Fast Food Restaurants happen more frequently in 'Deserts', there are still more Fast Food Restaurants per person in 'Non-Deserts'.
- Deserts have more coffee shops per person than any other type of shop.
- The only thing Child Care 'Deserts' have more of are convenience stores amongst the top venues.

Let's look at the Coffee Shop and Convenience store venues since there are more convenience stores on average in the Desert than any other venue and Coffee shops is the venue with the greatest prevalence in the Non-Desert. These two variables seem like a way to make a distinction between the two CCD classes. I will focus mostly on the Desert and Non-Desert for now.

E.3 Visualizing Coffee Shop and Convenience Store Differences in Deserts vs Non-Deserts

Not only will I focus on the Coffee Shop and Convenience Store variables but also on the Venues per Population and Venues per Ageless5. I will use the number of venues per person as the measure to plot instead of the count because I want to see a continuous value rather than a discrete plot.

Red shows Non-Desert while blue points show Deserts.
Figure 15 shows Venue per Ageless5 vs Convenience vs Coffee.

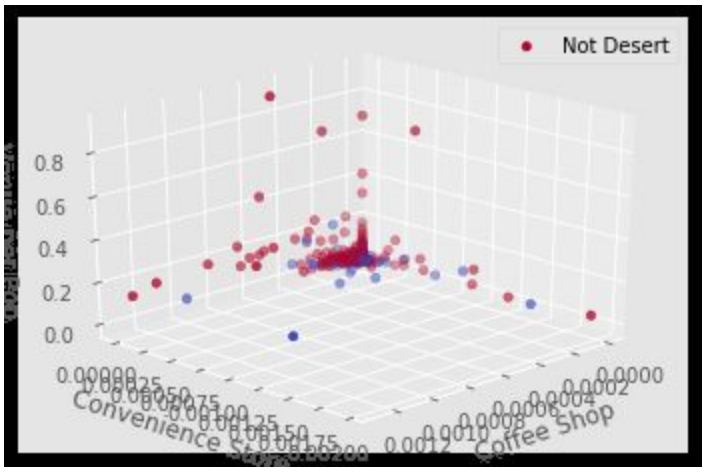

Fig.15

Venue per Population vs Convenience vs Coffee.

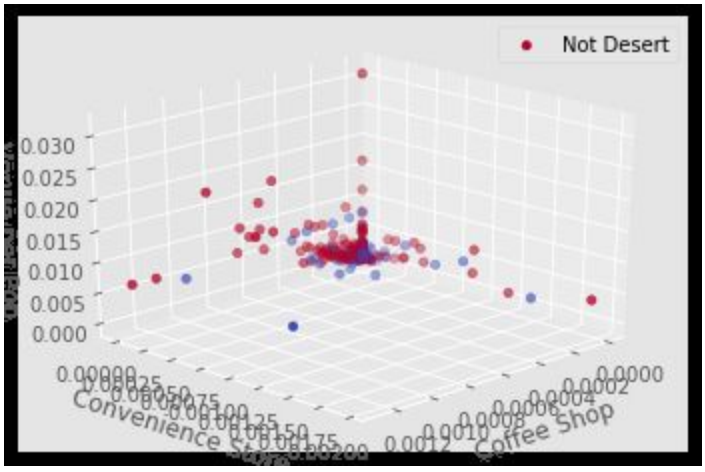

Fig.16

There does seem to be a correlation. Plotting to see just Venue per Pop vs Convenience there seems to be a trend.

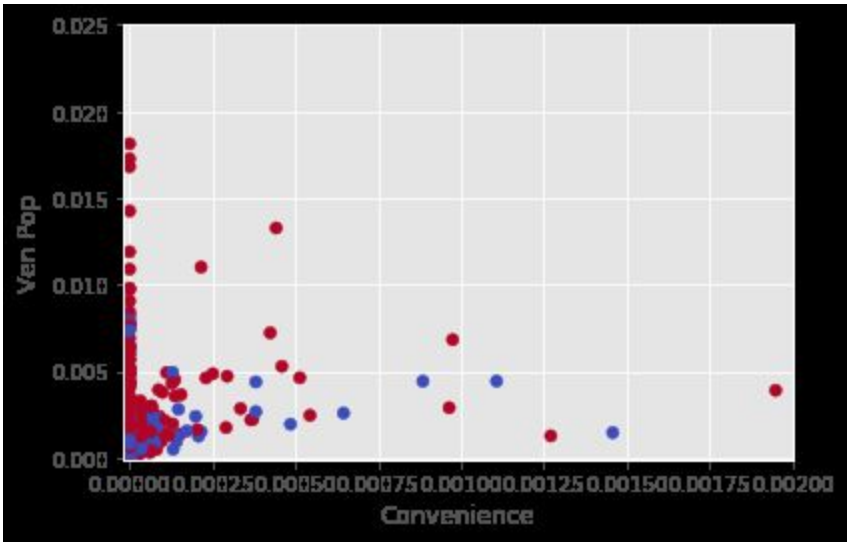

Fig.17

I place a circle around the trends that I notice for each class in the plot. Nonetheless there are many zip codes that overlap in these circles from each class near the origin.

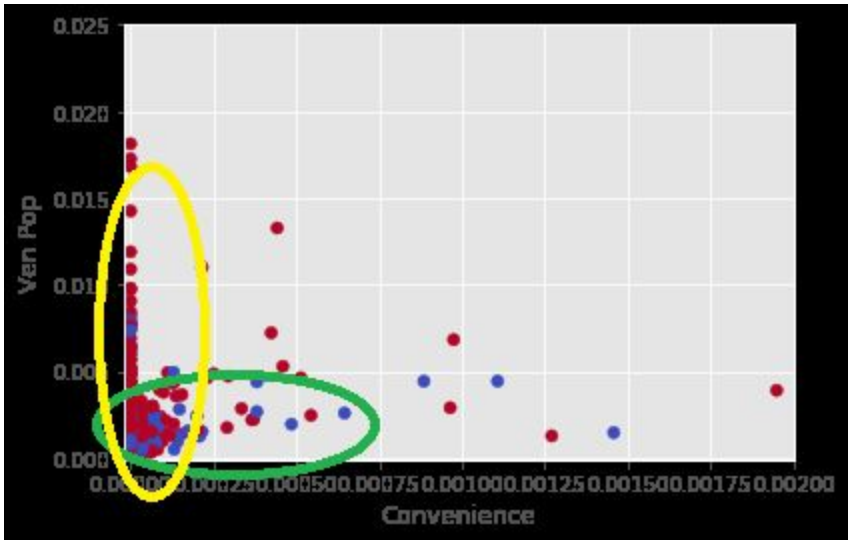

Fig.18

I do the same for Coffee shop.

We can see that there are more coffee shops per person in more zip codes in the Non-Desert (Red), however there are still many zip codes in that class that have a low coffee per person count making it hard to separate these two classes.

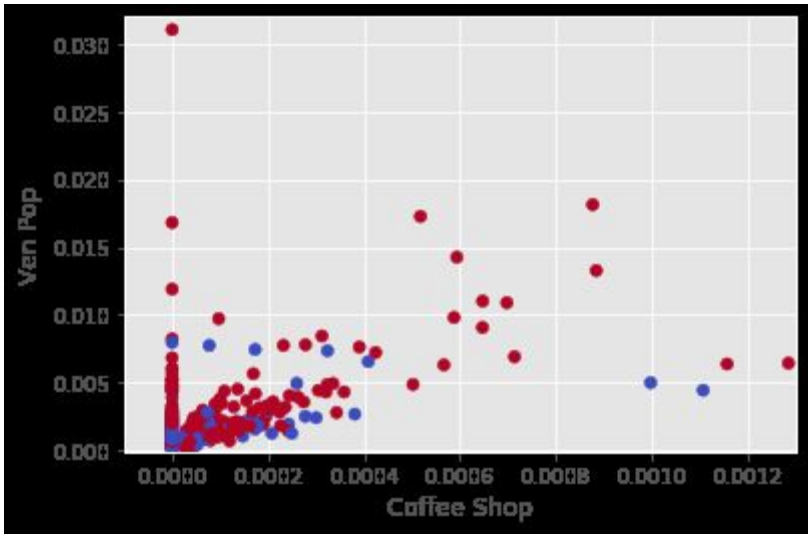

Fig. 19

I place circles around the trends noting Non-Desert has a larger variance yet a similar mean to Desert.

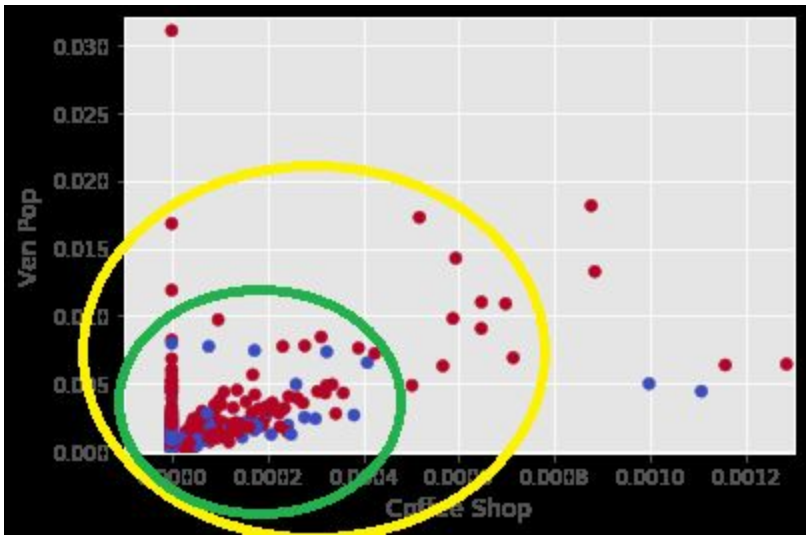

Fig. 20

Seeing that there is some way to differentiate between the two classes having used only two venues this gives me hope that I can I create a Machine Learning model that is better than random using the venue counts.

F. Modeling

I tested using only the coffee and convenience store variables in combination with population and total venue count but the ARI (Adjusted Rand Index) metric was terrible. Hence, I decided to use the total top ten venues as the X in constructing my model and saw an improvement in the ARI metric.

F.1 Unsupervised Clustering

Similar to the K-Means clustering in Labs of this project I was curious to see if K-clustering and DBSCAN could separate the three CCD classes though they are an unsupervised method.

The results are as follows:

```
from sklearn.metrics.cluster import adjusted_rand_score
#k-means performance:
print("ARI =", adjusted_rand_score(y, k_means.labels_))
>>  0.0235

#DBSCAN performance:
print("ARI =", adjusted_rand_score(y, (clusters)))
>> 0.0563
```

Where the closer the ARI score is to 1 the better a model the method was able to find. I searched manually for the best epsilon and min_samples of DBSCAN (eps =.8, min_samples =2). Plotting the top two venues to get a visual representation of the output of DBSCAN clusters below, we can see that similar to the circles in figure ???, that I placed previously for coffee and convenience store vs venues per population, the DBSCAN algorithm clusters two classes, one with a smaller radius clustering at the origin and the other with a wider radius in blue as well.
However similar to my “manual clusters”, there is much overlap and as the ARI score indicates much error

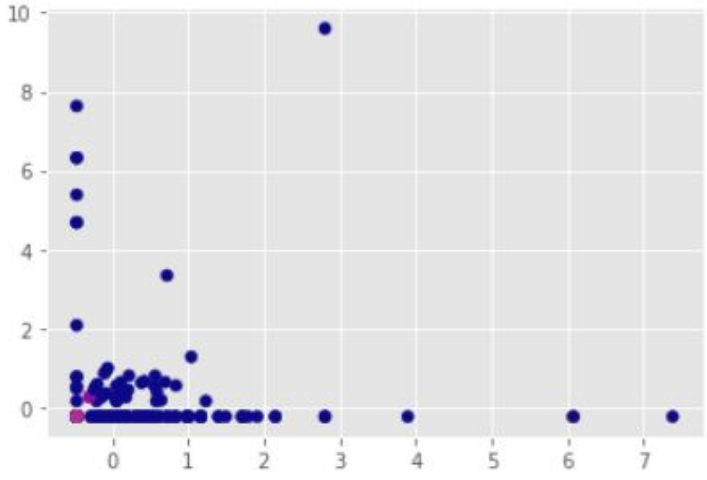

Fig.21 DBSCAN Results of Desert (pink) and Non-Desert (blue) using the top two venues scaled as the axes to plot.

F.2. SVM Modeling

For a supervised model, I chose SVM as there are various kernels within the method I can use to see if I can create a “better than random” model.

The following shows the kernel methods I used from the sklearn library and their ARI score results:

	rbf	**poly**	**sigmoid**	**linear**
C= 10	0.042	0	0.086	0.103
C= 100	0.016	0	0.127	0.105
C=400	0.208	.061	0.100	0.087

C values were chosen here after finding C=400 for rbf gave optimum value. C being the error penalty for training.

Looking further into SVM results with the rbf kernel at C=400, the confusion matrix is as follows:
Repeating the experiment gave similar results.

	precision	recall	f1-score	support
1	0.50	0.05	0.09	21
2	1.00	0.33	0.50	6
3	0.64	0.98	0.77	43
micro avg	0.64	0.64	0.64	70
macro avg	0.71	0.45	0.45	70
weighted avg	0.63	0.64	0.54	70

```
Confusion matrix, without normalization
[[ 1  0 20]
 [ 0  2  4]
 [ 1  0 42]]
```

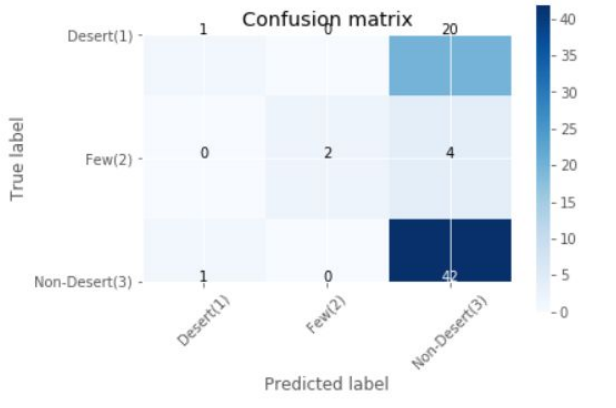

Fig. 22

Testing the various kernels in SVM the best ARI score given with ‘rbf’ shows only slightly better than random at about 0.2. Since there are mostly Non-Desert class samples, the models tend to classify everything as Non-Desert. Even when using class-weights as balanced, where the samples were multiplied to match the number of samples available for Non-Desert, the model suffered. Therefore, I continued testing using only the original number of samples.

G. Conclusion and Discussion

In this project, I explored Child Care Deserts (CCD) in relation to foursquare business venue data available for populated zip codes in Colorado. Since living in a Child Care Desert means there is little availability of Child Care for people living in a zip code finding someone to care for their child under 5 may become a hardship parents or guardians may face. This can mean they must travel or commute further, or sacrifice a job opportunity or salary, in order to care for their child. Given the opportunity to use foursquare business venue data to explore a problem, I decided to see if there was any observations I could make in relation to CCDs and business in Colorado. Realizing that Child Care providers are businesses as well and their owners must make decisions to location, demand, and customers willingness to pay the fees needed to cover expenses and profit, I hypothesised that there would be observations I could make to other businesses in the same area.

After analyzing and exploring the data I observed that

- Although places that lacked child care, Deserts, where on average more densely populated, there where more venues per person in places where there were Few-Children or non-Deserts. This means that there was also a lower amount of other business types available in Deserts, not only Child Care.
- If we looked at the density of babies per person (children under 5 as looked in this project), the greatest density existed in Desert areas. Therefore not only was there a lower amount of businesses of all types serving the Desert areas, but the people in these areas tended to have more babies to find child care for on average. This may seem obvious as Child Care Deserts are defined as places where there are three times the amount of children as there is child care capacity for them. However, the interesting thing is that, on average, all other venues also underserved this population, not only child care.
- Looking at the types of venues that are most frequently available in Non-Deserts, American Restaurants and Coffee shops were the top two venue types. All other top venues where also most frequently available in Non-Deserts in comparison to Deserts. The only top business type most available in Deserts where Convenience Stores.
- Ski-Areas exist only in Non-Deserts and Few-Children. There are none in Deserts.
- Fast food restaurants were the most frequent venue
- Using the one-hot encoding of the top 10 venues as a way to develop an ML model SVM produces a model a little better than random. Doing it manually (comparing vs human) gives similar results where it is difficult to separate the classes because there are a large amount of samples of every CCD class that have a scarce number of the top venues.

- However, since we can produce a model better than random, we can see there is some correlation between the top business types and predicting if the zip code will be a CCD.

H. Future Directions

While this project was limited in scope to foursquare business venue types and count, in the future other data such as population demographics, salary and age can be integrated to study why CCD emerge. Though not studied directly, it seems like the population in Non-Deserts may have higher wages to afford coffee shops and ski-areas with higher frequency. Therefore indicating support to the background research, where child care providers may choose to locate in areas where parents who are able to pay the fees needed to make a profit, even if government subsidies exist. Perhaps this way perpetuating the loop of penalizing economically disadvantaged parents and guardians.

Sources

[1] https://childcaredeserts.org/index.html
[2] https://www.newamerica.org/in-depth/care-report/introduction/
[3]https://www.americanprogress.org/issues/early-childhood/reports/2017/08/30/437988/mapping-americas-child-care-deserts/
[4] https://www.npr.org/sections/health-shots/2017/01/03/506448993/child-care-scarcity-has-very-real-consequences-for-working-families
[5]https://www.americanprogress.org/issues/early-childhood/reports/2016/10/27/225703/child-care-deserts/
[6] https://data.colorado.gov/
[7] https://data.colorado.gov/Demographics/Census-Zip-Codes-in-Colorado-2016/rwak-e74e
[8]https://data.colorado.gov/Early-childhood/Colorado-Licensed-Child-Care-Facilities-Report/a9rr-k8mu
[9] https://www.arcgis.com/home/index.html