Overview of Data Set

Here you can find the data source: https://data.world/the-pudding/gayborhoods. The data set scores 15 US cities based on the **four factors**: same sex married joint tax filings, same sex households, gay bar and pride parade route with each row as a unique zip code. Data set was published in 2018. Though it is somehow outdated, it provides the research framework when new data are available. The goal of my research is to provide recommendations for LGBTQs who plan to relocate to more queer populated areas.

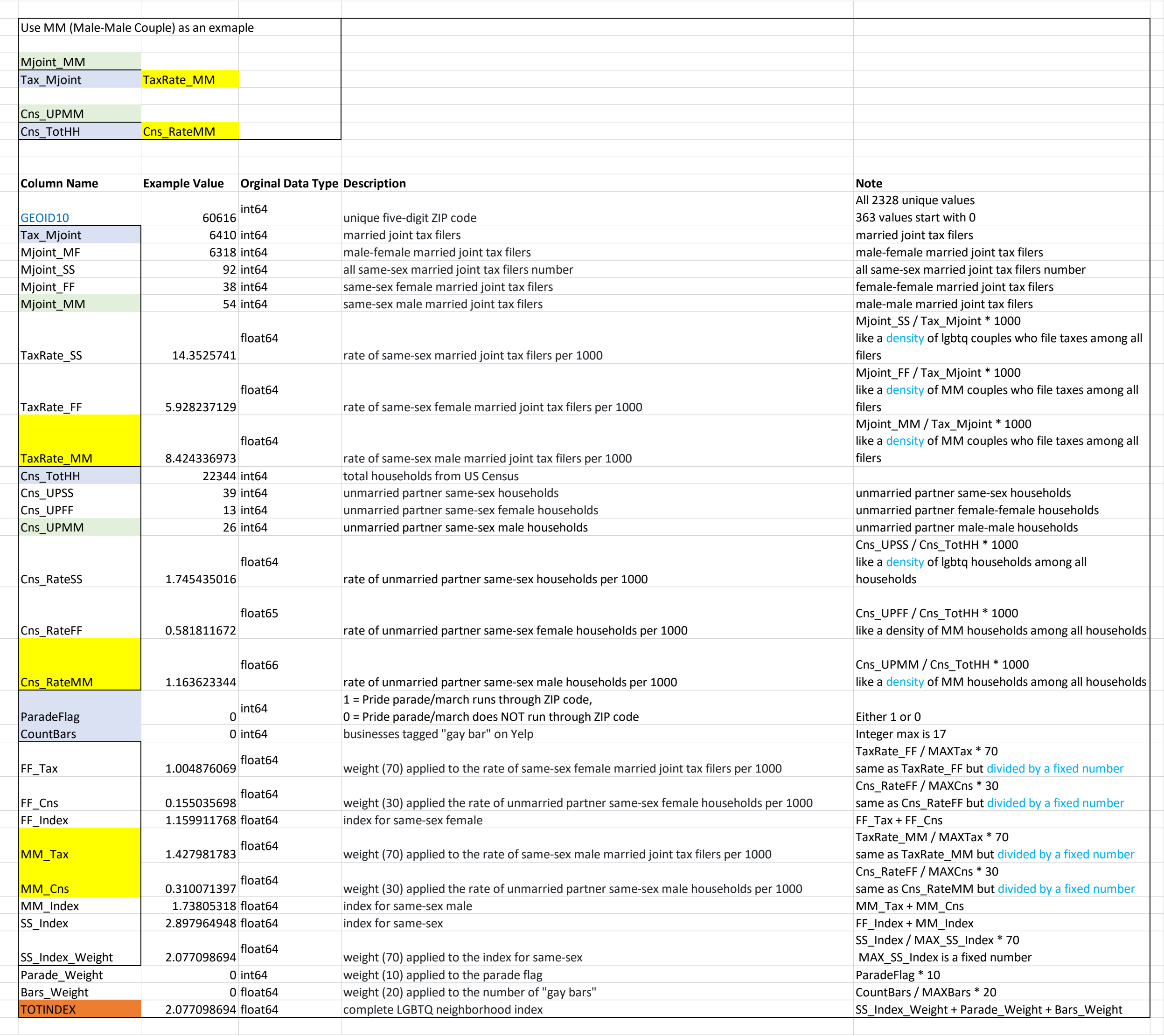
There are 2328 rows and 29 columns in the data set. I focused on cleaning the 'geoid10' – zip code column by ensuring the values are strings with five characters starting with none zero. I performed visualizations on the most zip-code related concerns for the **four factors** above. At the end I investigated correlations for the column I am interested to prepare for stage 2 research.

Among all the 29 columns, these 14 columns [ ] are concentrated on since they compose the core of the scoring system in the data set and contributed to the final total index. All the other columns are derivatives of the 14 columns.

14 core columns [ ]
['geoid10', 'tax_mjoint', 'mjoint_mm', 'taxrate_mm', 'mjoint_ff', 'taxrate_ff', 'cns_tothh', 'cns_upmm', 'cns_ratemm', 'cns_upff', 'cns_rateff', 'paradeflag', 'countbars', 'totindex']

Data Table Schema

Original data set with 29 columns



Use MM (Male-Male Couple) as an exmaple	
Mjoint_MM	
Tax_Mjoint	TaxRate_MM
Cns_UPMM	
Cns_TotHH	Cns_RateMM

Column Name	Example Value	Orginal Data Type	Description	Note
GEOID10	60616	int64	unique five-digit ZIP code	All 2328 unique values 363 values start with 0
Tax_Mjoint	6410	int64	married joint tax filers	married joint tax filers
Mjoint_MF	6318	int64	male-female married joint tax filers	male-female married joint tax filers
Mjoint_SS	92	int64	all same-sex married joint tax filers number	all same-sex married joint tax filers number
Mjoint_FF	38	int64	same-sex female married joint tax filers	female-female married joint tax filers
Mjoint_MM	54	int64	same-sex male married joint tax filers	male-male married joint tax filers
TaxRate_SS	14.3525741	float64	rate of same-sex married joint tax filers per 1000	Mjoint_SS / Tax_Mjoint * 1000 like a density of lgbtq couples who file taxes among all filers
TaxRate_FF	5.928237129	float64	rate of same-sex female married joint tax filers per 1000	Mjoint_FF / Tax_Mjoint * 1000 like a density of MM couples who file taxes among all filers
TaxRate_MM	8.424336973	float64	rate of same-sex male married joint tax filers per 1000	Mjoint_MM / Tax_Mjoint * 1000 like a density of MM couples who file taxes among all filers
Cns_TotHH	22344	int64	total households from US Census	
Cns_UPSS	39	int64	unmarried partner same-sex households	unmarried partner same-sex households
Cns_UPFF	13	int64	unmarried partner same-sex female households	unmarried partner female-female households
Cns_UPMM	26	int64	unmarried partner same-sex male households	unmarried partner male-male households
Cns_RateSS	1.745435016	float64	rate of unmarried partner same-sex households per 1000	Cns_UPSS / Cns_TotHH * 1000 like a density of lgbtq households among all households
Cns_RateFF	0.581811672	float65	rate of unmarried partner same-sex female households per 1000	Cns_UPFF / Cns_TotHH * 1000 like a density of MM households among all households
Cns_RateMM	1.163623344	float66	rate of unmarried partner same-sex male households per 1000	Cns_UPMM / Cns_TotHH * 1000 like a density of MM households among all households
ParadeFlag	0	int64	1 = Pride parade/march runs through ZIP code, 0 = Pride parade/march does NOT run through ZIP code	Either 1 or 0
CountBars	0	int64	businesses tagged "gay bar" on Yelp	Integer max is 17
FF_Tax	1.004876069	float64	weight (70) applied to the rate of same-sex female married joint tax filers per 1000	TaxRate_FF / MAXTax * 70 same as TaxRate_FF but divided by a fixed number
FF_Cns	0.155035698	float64	weight (30) applied the rate of unmarried partner same-sex female households per 1000	Cns_RateFF / MAXCns * 30 same as Cns_RateFF but divided by a fixed number
FF_Index	1.159911768	float64	index for same-sex female	FF_Tax + FF_Cns
MM_Tax	1.427981783	float64	weight (70) applied to the rate of same-sex male married joint tax filers per 1000	TaxRate_MM / MAXTax * 70 same as TaxRate_MM but divided by a fixed number
MM_Cns	0.310071397	float64	weight (30) applied the rate of unmarried partner same-sex male households per 1000	Cns_RateFF / MAXCns * 30 same as Cns_RateMM but divided by a fixed number
MM_Index	1.73805318	float64	index for same-sex male	MM_Tax + MM_Cns
SS_Index	2.897964948	float64	index for same-sex	FF_Index + MM_Index
SS_Index_Weight	2.077098694	float64	weight (70) applied to the index for same-sex	SS_Index / MAX_SS_Index * 70 MAX_SS_Index is a fixed number
Parade_Weight	0	int64	weight (10) applied to the parade flag	ParadeFlag * 10
Bars_Weight	0	float64	weight (20) applied to the number of "gay bars"	CountBars / MAXBars * 20
TOTINDEX	2.077098694	float64	complete LGBTQ neighborhood index	SS_Index_Weight + Parade_Weight + Bars_Weight

The relationships of the focused 14 core columns are listed in the chart below for easier understanding.

MM = Male-Male Couple, FF = Female-Female Couple	
For each GEOID10(zip code): 4 Categories (**black bold**) contribute to its final index (**blue bold**)	
Tax_Mjoint	
MM	FF
Mjoint_MM	Mjoint_FF
TaxRate_MM	TaxRate_FF
Mjoint_MM/Tax_Mjoint = TaxRate_MM	Mjoint_FF/Tax_Mjoint = TaxRate_FF
Cns_TotHH	
MM	FF
Cns_UPMM	Cns_UPFF
Cns_RateMM	Cns_RateFF
Cns_UPMM/Cns_TotHH = Cns_RateMM	Cns_UPFF/Cns_TotHH = Cns_RateFF
ParadeFlag	
CountBars	
TOTINDEX	

Data Analytics and Visualization

Overall 'totindex', the total index ranking are based on the **four factors**: same sex married joint tax filings, same sex households, gay bar and pride parade route for each zip code.

ranking	**zipcode**	**city**
No.1	90069	WEST HOLLYWOOD CA
No.2	94114	SAN FRANCISCO CA
No.3	10011	NEW YORK NY
No.4	10014	NEW YORK NY
No.5	94103	SAN FRANCISCO CA
No.6	70116	NEW ORLEANS LA
No.7	20009	WASHINGTON DC
No.8	98122	SEATTLE WA
No.9	30309	ATLANTA GA
No.10	90046	LOS ANGELES CA

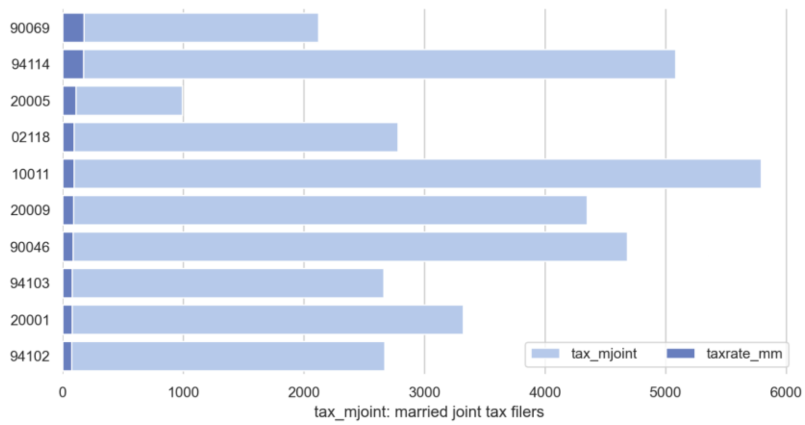
Tax and households rate for mm and ff couples

	taxrate_mm	**taxrate_ff**	**cns_ratemm**	**cns_rateff**
No.1	# 90069-WEST HOLLYWOOD CA	# 02130-JAMAICA PLAIN MA	# 94104-SAN FRANCISCO CA	# 78742-AUSTIN TX
No.2	# 94114-SAN FRANCISCO CA	# 94702-BERKELEY CA	# 94114-SAN FRANCISCO CA	# 20762-ANDREWS AIR FORCE BASE MD
No.3	# 20005-WASHINGTON DC	# 94609-OAKLAND CA	# 90069-WEST HOLLYWOOD CA	# 30317-ATLANTA GA

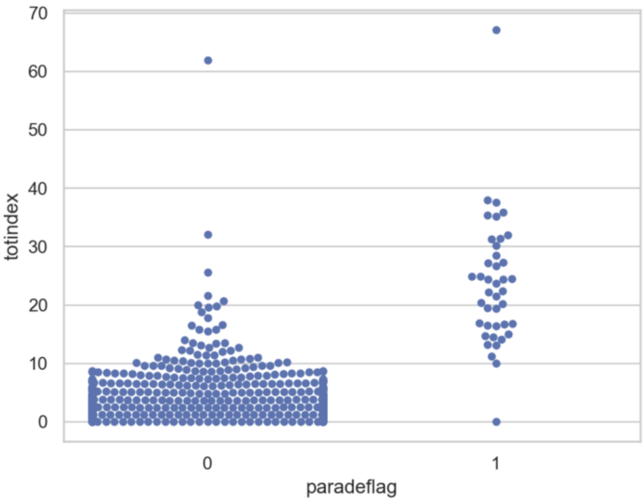
Based on the top ranked zip codes, it shows that **mm and ff couples do not really reside in the same areas**. Also, the rate itself may not represent the whole picture since the total number (denominator) also matters. Below I am looking at the rate and its denominator together to pick up the top ranked zip codes.

	taxrate_mm	**taxrate_ff**	**cns_ratemm**	**cns_rateff**
No.1	# 90069-WEST HOLLYWOOD CA	# 94702-BERKELEY CA	# 94104-SAN FRANCISCO CA	# 78742-AUSTIN TX
No.2	# 20005-WASHINGTON DC	# 94609-OAKLAND CA	# 90069-WEST HOLLYWOOD CA	# 20762-ANDREWS AIR FORCE BASE MD
No.3	# 94114-SAN FRANCISCO CA	# 02130-JAMAICA PLAIN MA	# 94114-SAN FRANCISCO CA	# 30317-ATLANTA GA

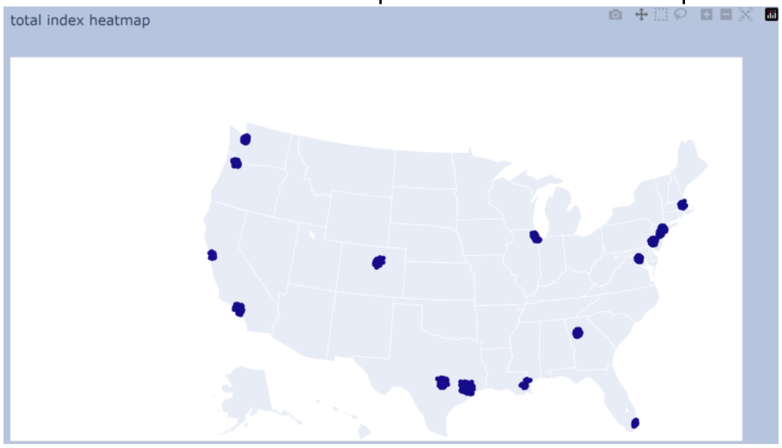
Some of the rankings have been changed which are highlighted in grey above. For example, I am not only looking at taxrate_mm (dark blue), but also consider their base (light blue), the total married joint tax filers. Even though zip code 20005 has 110 rate compared to zip code 94114's 172, 20005 has fewer than one fifth of 94114's total married joint tax filers. I rank 20005 higher since I would prefer to live in a less populated area but still have crowds of LGBTQs exist.



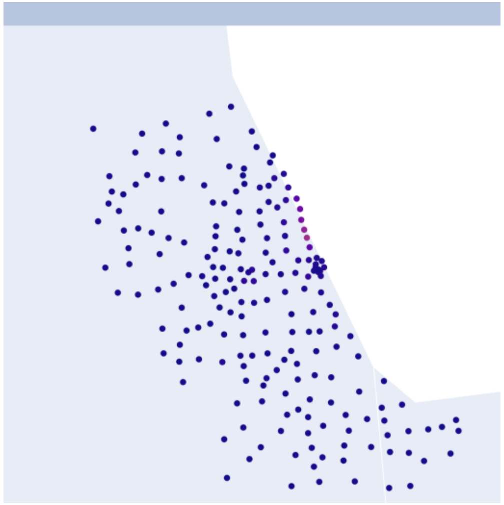
I wanted to look at the distribution of zip codes' totindex depending on if parade flag goes through the zip code or not.
The left upmost single point is # 94114-SAN FRANCISCO CA
The right upmost single point is # 90069-WEST HOLLYWOOD CA



I created an interactive heatmap in Pandas-Bokeh. The zip codes' totindex are represented by the colors of the dots.



Zoom in Chicago below



Column 'taxrate_mm''s correlations with others, from the lowest to the highest.

tax_mjoint	geoid10	cns_tothh	cns_rateff	cns_upff	mjoint_ff	paradeflag	taxrate_ff	countbars	cns_ratemm	cns_upmm	mjoint_mm	totindex	taxrate_mm
-0.032716	0.064971	0.202179	0.248544	0.368448	0.426547	0.450557	0.57647	0.650955	0.674153	0.769398	0.839537	0.899079	1

For more details and visualizations, please check out the Jupyter notebook "queerhood.ipynb".
Queers, get ready to party!