Exploring Chicago Park District Movies in the Parks 2018 Dataset

Jianghua Yang Group#9

Dataset Introduction

- Try to choose Real-world data, but not like TPCH.
- Chicago Park District Movies in the Parks 2018](https://data.cityofchicago.org/Events/ Chicago-Park-District-Movies-in-the-Parks-2018/e2v8-k3us)



Dataset Details

Column Name	Description	Туре
Day		Plain Text T
Date		Date & Time 🖽
Park		Plain Text T
Park Phone		Plain Text T
Title		Plain Text T
CC		Plain Text T
Rating		Plain Text T
Underwriter		Plain Text T
Park Address		Plain Text T
Location		Location

Explore the datasets and caveats

- Compact view: Show the dataset as a simple table
- Detail view: this view shows data distributions for each column
- Column view: this view shows more statistics about each column, but not the actual data in the spreadsheet
- Plot data: Try plot Location on OpenStreetMap
- Check FD: write python code to check FD columns.

Project pdf: https://github.com/IITTeaching/cs520-f23-group-9/blob/main/vizier_Final_Assignment/Vizier-final-pdm.pdf

Data Quality Problems

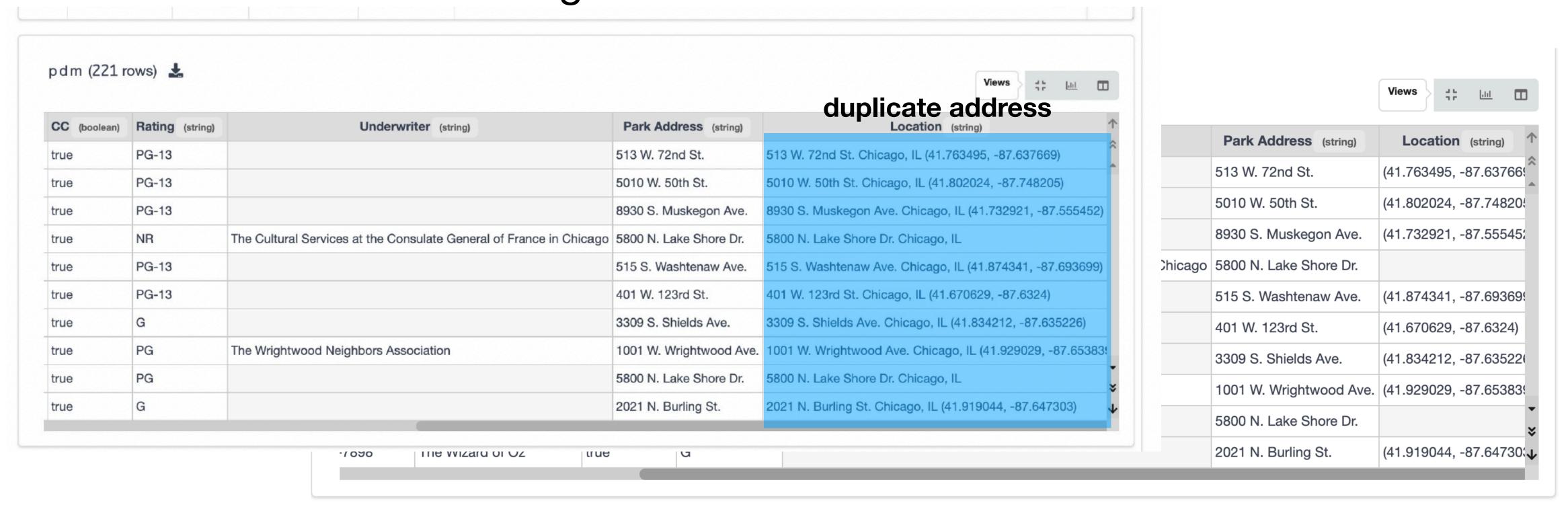
- Missing values
- Duplicate text
- Useless columns

• Issue-1 : remove useless columns {K, L, M, N, O}

Remove

I	J	K	L	M	N	0
Park Address	Location	Boundaries - ZIP Cod	Community Area	Zip Codes	Census Tracts	Wards
513 W. 72nd St.	513 W. 72nd St. Chicago, IL (41.763495, -87.637669)	11	66	21559	511	32
5010 W. 50th St.	5010 W. 50th St. Chicago, IL (41.802024, -87.748205)	7	53	22268	605	28
8930 S. Muskegon Ave.	8930 S. Muskegon Ave. Chicago, IL (41.732921, -87.555452)	25	42	21202	175	47
5800 N. Lake Shore Dr.	5800 N. Lake Shore Dr. Chicago, IL					
515 S. Washtenaw Ave.	515 S. Washtenaw Ave. Chicago, IL (41.874341, -87.693699)	28	28	21184	38	23
401 W. 123rd St.	401 W. 123rd St. Chicago, IL (41.670629, -87.6324)	19	50	21861	7	22
3309 S. Shields Ave.	3309 S. Shields Ave. Chicago, IL (41.834212, -87.635226)	40	35	21194	376	48
1001 W. Wrightwood Ave.	1001 W. Wrightwood Ave. Chicago, IL (41.929029, -87.653839)	16	68	21190	795	34
5800 N. Lake Shore Dr.	5800 N. Lake Shore Dr. Chicago, IL					
2021 N. Burling St.	2021 N. Burling St. Chicago, IL (41.919044, -87.647303)	16	68	21190	798	34

 Issue-2: Remove the duplicate information in the Location column and retain the latitude and longitude.



• issue-3: Find the rows with missing values and use the Google API query address to populate the latitude and longitude

e (string)	Title (stri	ng)	CC (boolean)	Rating (string)		Underwrit	ter (string)	Park Address (string)	Location (string)			
174	Black Panther		true	PG-13				513 W. 72nd St.	(41.763495, -87.63766		Views	
6022	Justice League		true	PG-13				5010 W. 50th St. sing	(41.802024, -87.74820	Douls Address (111)	Location (atrium)	
6023	Star Wars: The	_ast Jedi	true	PG-13				8930 S. Muskegon Ave.	(41.732921, -87.555452	Park Address (string)	Location (string)	
1134	Donkey Skin (Pe	eau d'ane)	true	NR	The Cultural Se	rvices at the Cons	ulate General of France in Chicago	5800 N. Lake Shore Dr.		i13 W. 72nd St.	(41.763495, -87.637669)	
5001	Justice League		true	PG-13				515 S. Washtenaw Ave.	(41.874341, -87.693699	010 W. 50th St.	(41.802024, -87.748205)	
7090	The Fate of the	Furious	true	PG-13				401 W. 123rd St.	(41.670629, -87.6324)	3930 S. Muskegon Ave.	(41.732921, -87.555452)	
6012	Cars 3		true	G				3309 S. Shields Ave.	(41.834212, -87.635220			
7816	Paddington 2		true	PG	The Wrightwood	d Neighbors Assoc	ciation	1001 W. Wrightwood Ave.	(41.929029, -87.65383	800 N. Lake Shore Dr.	(41.9861626, -87.652908;	
1134	9 to 5		true	PG				58	*	15 S. Washtenaw Ave.	(41.874341, -87.693699)	
7898	The Wizard of Oz true		G				2021 N. Burling St.	(41.919044, -87.64730:4	01 W. 123rd St.	(41.670629, -87.6324)		
)12	Cars 3		true	G					3309 S. Shields Ave.	(41.834212, -87.635226)	
	Paddington 2		true	PG	i	The Wrightwood Neighbors	s Association		1001 W. Wrightwood Ave.	(41.929029, -87.653839)		
	34	34 9 to 5 true		PG	i				5800 N. Lake Shore Dr.	(41.9861626, -87.652908		
	398 The Wizar		zard of Oz	true	G					2021 N. Burling St.	(41.919044, -87.647303)	

- verify time and day of the week column
- verify Location column

```
9 match_all = True
11 for row in ds.rows:
       date = datetime.strptime(row.get_value('Date'), '%m/%d/%Y').strftime('%A')
       day_of_week = date[:3]
      if day_of_week != row.get_value('Day'):
           match_all = False
           print(f"Day and Date mismatch in row {row.get_value('Day')} vs {date}")
18 if match_all:
      print("All rows match")
20
22 issue-5: verify Location column
23
25 all_locations_not_none = True
27 for row in ds.rows:
       location = row.get_value('Location')
      if location is None:
           all_locations_not_none = False
           print(f"None value found in Location column in row {row}")
33 if all_locations_not_none:
       print("All rows have non-None values in the Location column")
```

Challenges

Row 53 run google api:

https://maps.googleapis.com/maps/api/geocode/json?address='8050 S. Chapel.'&key={api_key}

return ZERO_RESULTS.

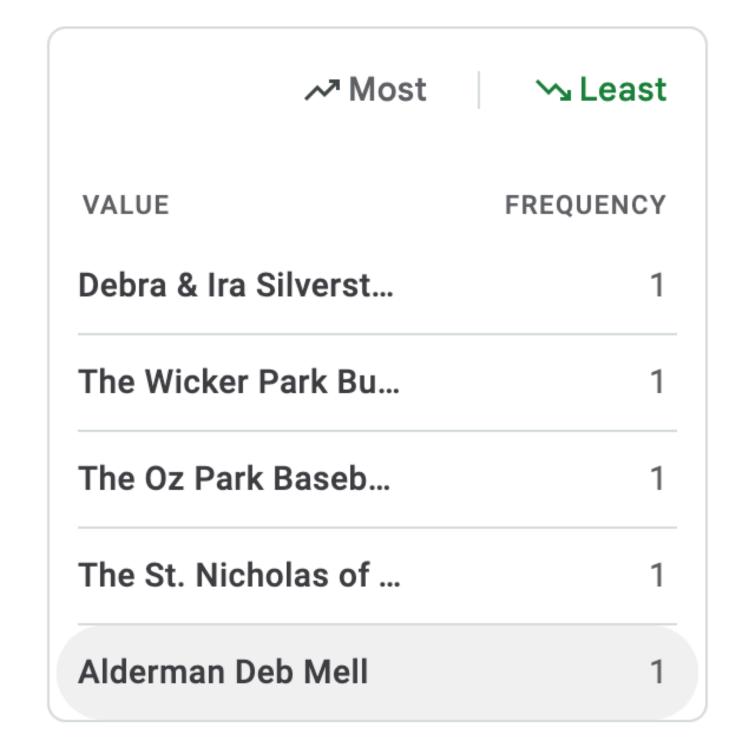
But we already know this place is at Chicago, IL. Then fix Park Address field of this row. Then update request to:

api_url = https://maps.googleapis.com/maps/api/geocode/json?address={park_address}, Chicago, IL&key={api_key}

this get the right result.

Challenges

- Underwriter column:
 - Total 221 rows
 - 140 empty cells
 - 33 unique values
- For Underwriter column: we can not fix missing values by some algorithm, because it's meaningless.
- So leave it as it is.



Total rows	221
Empty cells	140
Unique values	33

All stats include data in hidden or filtered rows

Visualization

After clean data, then can Plot Data on Map

