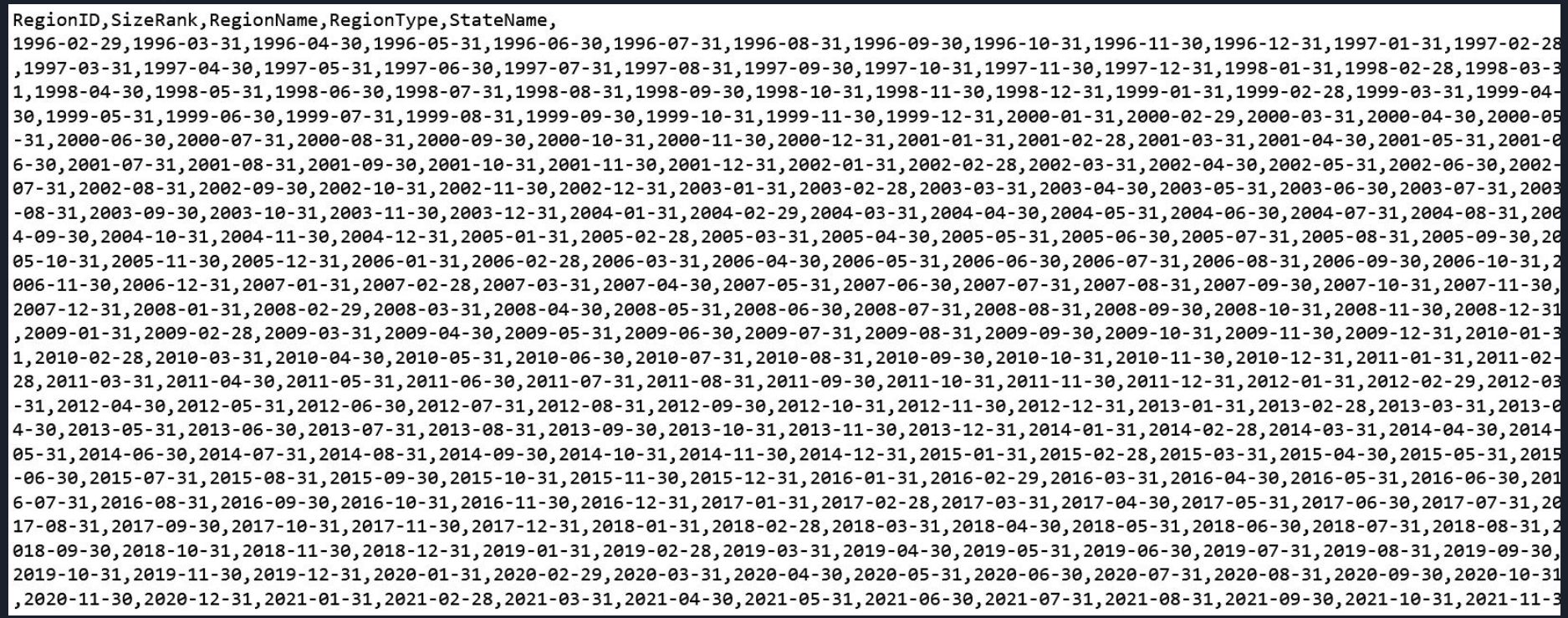
Project 3-Data Engineering

David Czoper, Anna Ramer, Jithu Jacob, Monica Dahl

Raw Data:
Zillow Home Values



```
RegionID,SizeRank,RegionName,RegionType,StateName,
1996-02-29,1996-03-31,1996-04-30,1996-05-31,1996-06-30,1996-07-31,1996-08-31,1996-09-30,1996-10-31,1996-11-30,1996-12-31,1997-01-31,1997-02-28
,1997-03-31,1997-04-30,1997-05-31,1997-06-30,1997-07-31,1997-08-31,1997-09-30,1997-10-31,1997-11-30,1997-12-31,1998-01-31,1998-02-28,1998-03-3
1,1998-04-30,1998-05-31,1998-06-30,1998-07-31,1998-08-31,1998-09-30,1998-10-31,1998-11-30,1998-12-31,1999-01-31,1999-02-28,1999-03-31,1999-04-
30,1999-05-31,1999-06-30,1999-07-31,1999-08-31,1999-09-30,1999-10-31,1999-11-30,1999-12-31,2000-01-31,2000-02-29,2000-03-31,2000-04-30,2000-05
-31,2000-06-30,2000-07-31,2000-08-31,2000-09-30,2000-10-31,2000-11-30,2000-12-31,2001-01-31,2001-02-28,2001-03-31,2001-04-30,2001-05-31,2001-0
6-30,2001-07-31,2001-08-31,2001-09-30,2001-10-31,2001-11-30,2001-12-31,2002-01-31,2002-02-28,2002-03-31,2002-04-30,2002-05-31,2002-06-30,2002-
07-31,2002-08-31,2002-09-30,2002-10-31,2002-11-30,2002-12-31,2003-01-31,2003-02-28,2003-03-31,2003-04-30,2003-05-31,2003-06-30,2003-07-31,2003
-08-31,2003-09-30,2003-10-31,2003-11-30,2003-12-31,2004-01-31,2004-02-29,2004-03-31,2004-04-30,2004-05-31,2004-06-30,2004-07-31,2004-08-31,200
4-09-30,2004-10-31,2004-11-30,2004-12-31,2005-01-31,2005-02-28,2005-03-31,2005-04-30,2005-05-31,2005-06-30,2005-07-31,2005-08-31,2005-09-30,20
05-10-31,2005-11-30,2005-12-31,2006-01-31,2006-02-28,2006-03-31,2006-04-30,2006-05-31,2006-06-30,2006-07-31,2006-08-31,2006-09-30,2006-10-31,2
006-11-30,2006-12-31,2007-01-31,2007-02-28,2007-03-31,2007-04-30,2007-05-31,2007-06-30,2007-07-31,2007-08-31,2007-09-30,2007-10-31,2007-11-30,
2007-12-31,2008-01-31,2008-02-29,2008-03-31,2008-04-30,2008-05-31,2008-06-30,2008-07-31,2008-08-31,2008-09-30,2008-10-31,2008-11-30,2008-12-31
,2009-01-31,2009-02-28,2009-03-31,2009-04-30,2009-05-31,2009-06-30,2009-07-31,2009-08-31,2009-09-30,2009-10-31,2009-11-30,2009-12-31,2010-01-3
1,2010-02-28,2010-03-31,2010-04-30,2010-05-31,2010-06-30,2010-07-31,2010-08-31,2010-09-30,2010-10-31,2010-11-30,2010-12-31,2011-01-31,2011-02-
28,2011-03-31,2011-04-30,2011-05-31,2011-06-30,2011-07-31,2011-08-31,2011-09-30,2011-10-31,2011-11-30,2011-12-31,2012-01-31,2012-02-29,2012-03
-31,2012-04-30,2012-05-31,2012-06-30,2012-07-31,2012-08-31,2012-09-30,2012-10-31,2012-11-30,2012-12-31,2013-01-31,2013-02-28,2013-03-31,2013-0
4-30,2013-05-31,2013-06-30,2013-07-31,2013-08-31,2013-09-30,2013-10-31,2013-11-30,2013-12-31,2014-01-31,2014-02-28,2014-03-31,2014-04-30,2014-
05-31,2014-06-30,2014-07-31,2014-08-31,2014-09-30,2014-10-31,2014-11-30,2014-12-31,2015-01-31,2015-02-28,2015-03-31,2015-04-30,2015-05-31,2015
-06-30,2015-07-31,2015-08-31,2015-09-30,2015-10-31,2015-11-30,2015-12-31,2016-01-31,2016-02-29,2016-03-31,2016-04-30,2016-05-31,2016-06-30,201
6-07-31,2016-08-31,2016-09-30,2016-10-31,2016-11-30,2016-12-31,2017-01-31,2017-02-28,2017-03-31,2017-04-30,2017-05-31,2017-06-30,2017-07-31,20
17-08-31,2017-09-30,2017-10-31,2017-11-30,2017-12-31,2018-01-31,2018-02-28,2018-03-31,2018-04-30,2018-05-31,2018-06-30,2018-07-31,2018-08-31,2
018-09-30,2018-10-31,2018-11-30,2018-12-31,2019-01-31,2019-02-28,2019-03-31,2019-04-30,2019-05-31,2019-06-30,2019-07-31,2019-08-31,2019-09-30,
2019-10-31,2019-11-30,2019-12-31,2020-01-31,2020-02-29,2020-03-31,2020-04-30,2020-05-31,2020-06-30,2020-07-31,2020-08-31,2020-09-30,2020-10-31
,2020-11-30,2020-12-31,2021-01-31,2021-02-28,2021-03-31,2021-04-30,2021-05-31,2021-06-30,2021-07-31,2021-08-31,2021-09-30,2021-10-31,2021-11-3
```

Raw Data: Population

	A	B	C	D	E	F	G
1	Geographic Area	Estimates Base	**2020**	**2021**	**2022**		
2	Abbeville city, Alabama	2,355	2,356	2,361	2,366		
3	Adamsville city, Alabama	4,372	4,360	4,292	4,224		
4	Addison town, Alabama	661	659	666	669		
5	Akron town, Alabama	227	226	226	221		
6	Alabaster city, Alabama	33,330	33,385	33,741	33,873		
7	Albertville city, Alabama	22,392	22,407	22,544	22,726		
8	Alexander City city, Alabama	14,847	14,822	14,701	14,636		
9	Aliceville city, Alabama	2,172	2,163	2,121	2,075		
10	Allgood town, Alabama	542	541	544	553		
11	Altoona town, Alabama	944	944	942	941		
12	Andalusia city, Alabama	8,842	8,844	8,824	8,790		
13	Anderson town, Alabama	253	253	254	254		
14	Anniston city, Alabama	21,562	21,513	21,343	21,182		
15	Arab city, Alabama	8,443	8,449	8,501	8,623		
16	Ardmore town, Alabama	1,322	1,335	1,370	1,393		
17	Argo town, Alabama	4,361	4,365	4,379	4,364		
18	Ariton town, Alabama	657	655	657	659		
19	Arley town, Alabama	328	327	330	331		

ETL: Extract Zillow

Out[2]:

	RegionID	SizeRank	RegionName	RegionType	StateName	1996-02-29	1996-03-31	1996-04-30	1996-05-31	1996-06-30	...	2023-05-31	
0	102001	0	United States	country	NaN	102066.723398	102325.134506	102872.937891	103617.903465	104422.257114	...	3.467107e+05	
1	394913	1	New York, NY	msa	NY	178474.779852	177739.600114	177569.023789	177793.340930	178655.172637	...	6.167480e+05	
2	753899	2	Los Angeles, CA	msa	CA	185452.585240	186598.596892	187465.936966	187833.585659	188476.217739	...	8.908439e+05	
3	394463	3	Chicago, IL	msa	IL	129644.374948	129051.457766	130466.841005	132148.339940	133709.170452	...	3.007739e+05	
4	394514	4	Dallas, TX	msa	TX	109830.986773	110509.150315	111563.142434	112413.465437	112921.778396	...	3.776917e+05	
5	394692	5	Houston, TX	msa	TX	107407.315102	107679.194312	108359.743412	108805.548804	108952.909112	...	3.059645e+05	
6	395209	6	Washington, DC	msa	VA	168081.911654	168319.414680	168345.566357	169104.548885	169856.379185	...	5.471539e+05	
7	394974	7	Philadelphia, PA	msa	PA	111084.052724	110533.942728	110618.671608	111330.216953	112258.836585	...	3.395340e+05	
8	394856	8	Miami, FL	msa	FL	100666.957547	101124.173555	101396.231444	101791.734710	102079.939430	...	4.597249e+05	
9	394347	9	Atlanta, GA	msa	GA	115696.739391	116132.010240	117127.730483	118187.701768	119315.126000	...	3.726342e+05	
10	394404	10	Boston, MA	msa	MA	153690.647938	154480.890965	155704.967821	157451.759075	160048.115647	...	6.482677e+05	
11	394976	11	Phoenix, AZ	msa	AZ	113566.094298	114380.751736	115065.510771	115762.589655	116416.289061	...	4.440184e+05	
			San										

ETL: Extract Population

Out[3]:

	Geographic Area	Estimates Base	2020	2021	2022
0	Abbeville city, Alabama	2355.0	2356.0	2361.0	2366.0
1	Adamsville city, Alabama	4372.0	4360.0	4292.0	4224.0
2	Addison town, Alabama	661.0	659.0	666.0	669.0
3	Akron town, Alabama	227.0	226.0	226.0	221.0
4	Alabaster city, Alabama	33330.0	33385.0	33741.0	33873.0
5	Albertville city, Alabama	22392.0	22407.0	22544.0	22726.0
6	Alexander City city, Alabama	14847.0	14822.0	14701.0	14636.0
7	Aliceville city, Alabama	2172.0	2163.0	2121.0	2075.0
8	Allgood town, Alabama	542.0	541.0	544.0	553.0
9	Altoona town, Alabama	944.0	944.0	942.0	941.0
10	Andalusia city, Alabama	8842.0	8844.0	8824.0	8790.0
11	Anderson town, Alabama	253.0	253.0	254.0	254.0
12	Anniston city, Alabama	21562.0	21513.0	21343.0	21182.0
13	Arab city, Alabama	8443.0	8449.0	8501.0	8623.0
14	Ardmore town, Alabama	1322.0	1335.0	1370.0	1393.0

Out[2]:

	State	Abbreviation
0	Alabama	AL
1	Alaska	AK
2	Arizona	AZ
3	Arkansas	AR
4	California	CA

ETL: Transform Zillow

- Drop:
  - Unneeded Row
  - Unused Columns
  - Data before 2014
  - Null values
- Split the existing region name column into City and State
- Calculate yearly average price per city
- Clean data of spaces
- Change datatype from float to integer
- Reformat for use in SQL
- Export

ETL: Transform Population

- Created and imported helper spreadsheet of State abbreviations
- Drop:
  - Unused columns
  - Rows with null values
  - Duplicate location/state pairs
- Split:
  - Location into city/state (drop full state name after merge)
  - Location designation from name (drop designation)
- Using a left merge, add the state abbreviation for each state name
- Rearrange and rename columns
- Sort for population and choose top 10 places
- Export

ETL: Transform → Load
Zillow

	City	State	2014_avg	2015_avg	2016_avg	2017_avg	2018_avg	2019_avg	2020_avg	2021_avg	2022_avg	2023_avg	2024_avg
1	New York	NY	402362	416699	435647	458805	480549	495660	514550	564338	607852	621560	638713
2	Los Angeles	CA	494696	516088	543868	593676	651578	661262	702113	800193	897896	898687	923457
3	Chicago	IL	192162	200389	211026	222481	232501	237701	245586	271834	293834	302620	307221
4	Dallas	TX	169951	187043	209094	230764	248263	256268	268992	314521	378230	375329	371673
5	Houston	TX	168793	186431	197103	204476	212859	221167	230612	262498	305275	305278	303308
...	...	...	...	...	...	...	...	...	...	...	...	...	...
890	Zapata	TX	103351	107795	107930	113224	114638	119435	128080	139111	137278	119349	108358
891	Ketchikan	AK	254184	268618	278640	287025	304212	324276	344258	368896	391994	388390	375909
892	Craig	CO	155559	156684	158803	162708	169899	180380	194530	219458	254478	276521	277276
893	Vernon	TX	67850	66962	66663	67127	70610	75248	81507	86875	95918	91386	85823
894	Lamesa	TX	68181	72209	76035	75334	76942	82307	91559	98232	94129	83427	74886

893 rows × 13 columns

ETL: Transform → Load
Zillow

	City	State	Year	Price
0	New York	NY	2014	402362
1	Los Angeles	CA	2014	494696
2	Chicago	IL	2014	192162
3	Dallas	TX	2014	169951
4	Houston	TX	2014	168793
...	...	...	...	...
9818	Zapata	TX	2024	108358
9819	Ketchikan	AK	2024	375909
9820	Craig	CO	2024	277276
9821	Vernon	TX	2024	85823
9822	Lamesa	TX	2024	74886

9823 rows × 4 columns

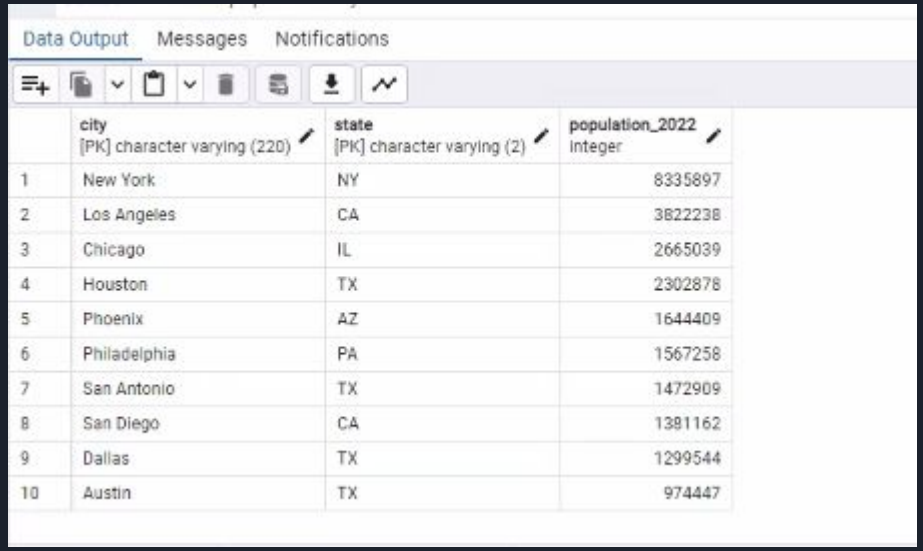
ETL: Load Cerberus

- Cerberus is a library that has validation tools for creating SQL tables from existing data
- Used in this project to check that the tables had no non-matching data

	City	State	Year	Price	population_2022	_merge
55	Washington	DC	2014	370565	NaN	left_only
56	Washington	DC	2015	379705	NaN	left_only
57	Washington	DC	2016	389330	NaN	left_only
58	Washington	DC	2017	403513	NaN	left_only
59	Washington	DC	2018	416829	NaN	left_only
...	...	...	...	...	...	...
9818	Lamesa	TX	2020	91559	NaN	left_only
9819	Lamesa	TX	2021	98232	NaN	left_only
9820	Lamesa	TX	2022	94129	NaN	left_only
9821	Lamesa	TX	2023	83427	NaN	left_only
9822	Lamesa	TX	2024	74886	NaN	left_only

ETL: Load Zillow

	City	State	Year	Price
0	New York	NY	2014	402362
1	Los Angeles	CA	2014	494696
2	Chicago	IL	2014	192162
3	Dallas	TX	2014	169951
4	Houston	TX	2014	168793
5	Philadelphia	PA	2014	204589
6	Phoenix	AZ	2014	202542
7	San Diego	CA	2014	455060



	city [PK] character varying (220)	state [PK] character varying (2)	population_2022 integer
1	New York	NY	8335897
2	Los Angeles	CA	3822238
3	Chicago	IL	2665039
4	Houston	TX	2302878
5	Phoenix	AZ	1644409
6	Philadelphia	PA	1567258
7	San Antonio	TX	1472909
8	San Diego	CA	1381162
9	Dallas	TX	1299544
10	Austin	TX	974447

ETL: Load Population

	City	State	population_2022
11690	New York	NY	8335897
1442	Los Angeles	CA	3822238
3410	Chicago	IL	2665039
17027	Houston	TX	2302878
665	Phoenix	AZ	1644409
15280	Philadelphia	PA	1567258
17502	San Antonio	TX	1472909
1571	San Diego	CA	1381162
16800	Dallas	TX	1299544
16583	Austin	TX	974447

Data Output | Messages | Notifications

	city [PK] character varying (220)	state [PK] character varying (2)	year [PK] integer	price integer
1	New York	NY	2014	402362
2	Los Angeles	CA	2014	494696
3	Chicago	IL	2014	192162
4	Dallas	TX	2014	169951
5	Houston	TX	2014	168793
6	Philadelphia	PA	2014	204589
7	Phoenix	AZ	2014	202542
8	San Diego	CA	2014	455060
9	San Antonio	TX	2014	156013
10	Austin	TX	2014	240550
11	New York	NY	2015	416699

Final Result

Housing DataFrame:

	city	state	year	price
0	New York	NY	2014	402362
1	Los Angeles	CA	2014	494696
2	Chicago	IL	2014	192162
3	Dallas	TX	2014	169951
4	Houston	TX	2014	168793
5	Philadelphia	PA	2014	204589
6	Phoenix	AZ	2014	202542
7	San Diego	CA	2014	455060

Population DataFrame:

	city	state	population_2022
0	New York	NY	8335897
1	Los Angeles	CA	3822238
2	Chicago	IL	2665039
3	Houston	TX	2302878
4	Phoenix	AZ	1644409
5	Philadelphia	PA	1567258
6	San Antonio	TX	1472909
7	San Diego	CA	1381162
8	Dallas	TX	1299544
9	Austin	TX	974447

Thank you!