



Policy Attribute Scraper Pipeline

Danh Nguyen
Okeefe Niemann
USF Practicum

Motivations

1. Generate a resource to ease the burden of manually updating values for Zoning Attribute tables on Socrata.
2. Provide insights into the written structure of city zoning policies



Process Overview

1. Retrieve scraped documents from s3

- s3 bucket:
"mtc-redshift-upload-test_kjfshar"

- scraped policies from jurisdictions in the bay area:

- municode.com
- codepublishing.com
- qcode.us
- amlegal.com

2. Extract information by attribute

- scan through policy contents, identifying lines referring to attributes:

- max floor area ratio
- max dwellings unit area
- building height
- units per lot
- minimum lot square feet

3. Organize found attributes

- Populate a single dataframe for all cities with the following columns:

- City
- Attribute
- Values
- Context
- Policy Subsection
- Line No.

4. Identify potential zone context

- Find any potential zone candidates for found attributes

- Finalize the DataFrame with this information under "Zone Candidates"

1. Retrieve scraped documents from s3

- Documents obtained from web scraper created by previous practicum students
- Saved as raw text files
- Only policies with relevant titles scraped
 - Ex: Zoning, Buildings and Construction
- **Most recent scraped** policy used:
 - Ex: 4-21-20

```
'test_kjafshar/berkeley/01-21-20/Title 5 .txt',  
'test_kjafshar/berkeley/01-21-20/Title 6 PUBLIC PROPERTY*.txt',  
'test_kjafshar/berkeley/01-21-20/Title 7 FINANCE, REVENUE AND TAXES*.txt',  
'test_kjafshar/berkeley/01-21-20/Title 8 .txt',  
'test_kjafshar/berkeley/01-21-20/Title 9 BUSINESS LICENSES AND REGULATIONS*.txt',  
'test_kjafshar/berkeley/04-21-20/CHARTER OF THE CITY OF BERKELEY.txt',  
'test_kjafshar/berkeley/04-21-20/Title 1 GENERAL PROVISIONS.txt',  
'test_kjafshar/berkeley/04-21-20/Title 10 ANIMALS*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 11 ENVIRONMENTAL HEALTH*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 12 HEALTH AND SAFETY*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 13 PUBLIC PEACE, MORALS AND WELFARE*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 14 VEHICLES AND TRAFFIC*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 15 HAZARDOUS MATERIALS*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 16 STREETS, SIDEWALKS AND OTHER PUBLIC PROPERTY*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 17 WATER AND SEWERS.txt',  
'test_kjafshar/berkeley/04-21-20/Title 19 BUILDINGS AND CONSTRUCTION*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 2 ADMINISTRATION.txt',  
'test_kjafshar/berkeley/04-21-20/Title 20 SIGNS*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 21 SUBDIVISIONS*.txt',  
'test_kjafshar/berkeley/04-21-20/Title 22 PLANNING AND DEVELOPMENT PROCEDURES.txt',
```

2. Extract information by attribute

- Create a data structure for each city:
 - Policy subsection
 - Values found
 - Context in which the value was found
 - Line Number of context in raw text file
- Attributes identified via:
 - Regular expression list for each attribute
 - Manual inspection
 - Range for each attribute
 - EDA of existing values in “Regional Zoning Code 2018” (Socrata)

```
attr_regex_dict = {'max_far': [r'(?i)floor.+area.+ratio',  
                               r'(?i)max.+far'],  
                  'max_dua': [r'(?i)dua',  
                              r'(?i)dwelling.+unit.+acre',  
                              r'(?i)lots.+acres',  
                              r'(?i)acres.+dwelling.+unit',  
                              r'(?i)dwelling.+(acre|unit)',  
                              r'(?i)unit.+acres?'],  
                  'building_height' : [r'(?i)building.+height',  
                                         r'(?i)maximum.+height',  
                                         r'(?i)height.+structures?'],  
                  'units_per_lot' : [r'(?i)units?.+per.+lot',  
                                     r'(?i)units?.+sq(are)*.feet'],  
                  'minimum_lot_sqft': [r'(?i)unit.+per.+feet',  
                                       r'(?i)mini?m?u?m?.+lot.+sqare?r?e?.+feet',  
                                       r'(?i)sq(uare)*.+fo*et',  
                                       r'(?i)gross.+lot.+area']  
                  }  
  
attr_range_dict = {'max_far': (0, 5),  
                  'max_dua': (0, 100),  
                  'building_height' : (15, 250),  
                  'units_per_lot' : (0.5, 5),  
                  'minimum_lot_sqft': (600, 1e8)  
                  }
```

3. Organize found attributes

- Create single dataframe for all cities with the relevant data found in (2)

	City	Attribute	Values	Context	Policy Subsection	Line No.
13387	los_altos	minimum_lot_sqft	750.0	An approved automatic sprinkler system shall be provided throughout all existing buildings, when additions are made that exceed fifty (50) percent and/or seven hundred and fifty (750) square feet of existing floor areas (area calculations shall not include existing basement floor areas).	Title 12 - BUILDINGS AND CONSTRUCTION	543
13388	los_altos	minimum_lot_sqft	750.0	It is the intent of the city to ensure that all new utility services and relocated existing utility services are placed underground, including additions exceeding fifty (50) percent of floor area and/or seven hundred and fifty (750) square feet or more, excluding basements. For the purpose of this section, removal of roof framing with associated exterior walls down to, or below the subfloor/slab shall be included in the above calculations. Therefore, the following shall apply:	Title 12 - BUILDINGS AND CONSTRUCTION	1608
13389	los_altos	minimum_lot_sqft	5000.0, 3999.0	Apartments, two family units or more per building with not more than one dwelling unit for each five thousand (5,000) square feet of lot area; provided, however, if after dividing the area of the site by five thousand (5,000), a remainder of less than five thousand (5,000) square feet but more than three thousand nine hundred ninety-nine (3,999) square feet is obtained, one additional dwelling unit may be located on the site;	Title 14 - ZONING	1385
13390	los_altos	minimum_lot_sqft	20000.0, 1000.0, 43560.0, 10000.0, 15000.0	Density. The maximum number of permitted dwelling units shall be calculated by subtracting twenty (20) percent of the gross area of the parcel and dividing the remainder by the required lot area per dwelling unit in the appropriate R1 District (ten thousand (10,000) square feet in the R1-10 District, fifteen thousand (15,000) square feet in the R1-H District, twenty thousand (20,000) square feet in the R1-20 District, and forty-three thousand five hundred sixty (43,560) square feet in the R1-40 District).	Title 14 - ZONING	4781
13391	los_altos	minimum_lot_sqft	7000.0	"Personal service" means a use, not conducted within an office, providing services for the personal care of an individual or the fitting, cleaning, repair, or maintenance of personal effects and not primarily for the sale of goods or merchandise. Personal services include beauty shops, barber shops and nail salons, pet grooming, shoe repair, laundry and cleaning services, repair and fitting of clothes, and other similar services. Personal services also include art, dance, music, tutoring centers, fitness studios and health clubs that do not exceed seven thousand (7,000) gross square feet.	Title 14 - ZONING	174
13392	los_altos	minimum_lot_sqft	10000.0, 11000.0, 15000.0	The minimum site area shall be ten thousand (10,000) square feet, except that on corner lots the minimum site area shall be eleven thousand (11,000) square feet and on flag lots the minimum site area shall be fifteen thousand (15,000) square feet.	Title 14 - ZONING	344
13393	los_altos	minimum_lot_sqft	20000.0, 21000.0	The minimum site area shall be twenty thousand (20,000) square feet, except that on corner lots the minimum site area shall be twenty-one thousand (21,000) square feet.	Title 14 - ZONING	510

3 (cont). Greater Context

```
print_Context(df=df_all_cities, line_num=13392, context_window=10)
```

```
City: los_altos  
Attribute: minimum_lot_sqft  
Policy Section: Title 14 - ZONING  
Line Number: 343
```

```
Line: The minimum site area shall be ten thousand (10,000) square feet, except that on corner lots the minimum site area shall be eleven thousand (11,000) square feet and on flag lots the minimum site area shall be fifteen thousand (15,000) square feet.
```

```
Context:
```

```
334: A.
```

```
335: Flag lots;
```

```
336: B.
```

```
337: Large family day care; and
```

```
338: C.
```

```
339: Pre-existing community facilities as provided in Chapter 14.70 of this title. New community facilities are prohibited.
```

```
340: (Ord. 05-285 § 2 (part): Ord. 05-271 § 2 (part): Ord. 04-267 § 2 (part))
```

```
341: 14.06.040 - Site area (R1-10).
```

```
342: COMPARE VERSIONS
```

```
343: The minimum site area shall be ten thousand (10,000) square feet, except that on corner lots the minimum site area shall be eleven thousand (11,000) square feet and on flag lots the minimum site area shall be fifteen thousand (15,000) square feet.
```

```
344: (Ord. 04-267 § 2 (part))
```

```
345: 14.06.050 - Site frontage, width and depth (R1-10).
```

```
346: COMPARE VERSIONS
```

```
347: A.
```

```
348: The minimum site frontage and width shall be eighty (80) feet and the minimum site depth shall be one hundred (100) feet, except that the minimum site width for a corner lot shall be ninety (90) feet and the minimum site frontage on a cul-de-sac turnaround shall be sixty (60) feet.
```

```
349: B.
```

```
350: The minimum width of the access corridor for each flag lot shall be twenty (20) feet, and shall access directly to a public or private street. The access corridor shall not connect to any portion of the turnaround space of a cul-de-sac.
```

```
351: (Ord. 04-267 § 2 (part))
```

```
352: 14.06.060 - Coverage (R1-10).
```

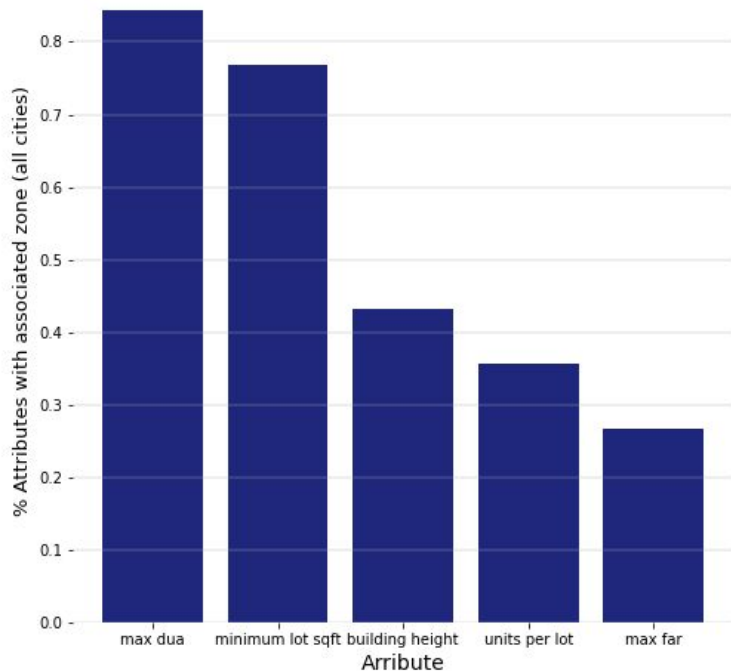
4. Identify potential zone context

- In a context window of 20 lines above and below, identify any potential zone candidates via regular expressions:
 - Candidates with no zones found within greater context registered as “Nan”

	City	Attribute	Values	Context	Policy Subsection	Line No.	Zone Candidates
13381	los_altos	minimum_lot_sqft	2019.0, 750.0	R313.1 Townhouse automatic fire sprinkler systems. An automatic residential fire sprinkler system shall be installed in accordance with National Fire Protection Association's (NFPA) Standard 13D in all new townhouses and in existing townhouses, when additions are made that increase the building area to more than the allowable Fire-Flow Appendix B, Tables B105.1(1) and B105.1(2) of the 2019 California Fire Code, and/or additions exceeding fifty (50) percent of the existing living area (existing square foot calculations shall not include existing basement) and/or additions exceeding seven hundred and fifty (750) square feet. When automatic fire sprinkler systems are required by this section, all associated attached garages shall be included. Additions over fifty (50) percent and/or seven hundred and fifty (750) square feet as referenced above, shall be treated as a new structure regarding installation of fire sprinkler systems. For the purpose of this section, removal of roof framing with associated exterior walls down to, or below the subfloor/slab shall be included in the above calculations. Therefore, the following shall apply:	Title 12 - BUILDINGS AND CONSTRUCTION	86	D2, R301, R313, D0, B105, D1
13382	los_altos	minimum_lot_sqft	2019.0, 750.0	R313.2 One and two-family dwellings automatic fire sprinklers systems. An automatic residential fire sprinkler system shall be installed in accordance with National Fire Protection Association's (NFPA) Standard 13D in all new one and two-family dwellings and in existing dwellings, when additions are made that increase the building area to more than the allowable Fire-Flow Appendix Tables B105.1(1) and B105.1(2) of the 2019 California Fire Code, and/or additions exceeding fifty (50) percent of the existing living area (existing square foot calculations shall not include existing basement) and/or additions exceeding seven hundred and fifty (750) square feet. When automatic fire sprinkler systems are required by this section, all associated garages shall be included. Additions over fifty (50) percent and/or seven hundred and (750) square feet as referenced above, shall be treated as a new structure regarding installation of fire sprinkler systems. For the purpose of this section, removal of roof framing with associated exterior walls down to, or below the subfloor/slab shall be included in the above calculations. Therefore, the following shall apply:	Title 12 - BUILDINGS AND CONSTRUCTION	89	D2, R602, D0, R313, B105, D1
13389	los_altos	minimum_lot_sqft	5000.0, 3999.0	Apartments, two family units or more per building with not more than one dwelling unit for each five thousand (5,000) square feet of lot area; provided, however, if after dividing the area of the site by five thousand (5,000), a remainder of less than five thousand (5,000) square feet but more than three thousand nine hundred ninety-nine (3,999) square feet is obtained, one additional dwelling unit may be located on the site;	Title 14 - ZONING	1385	R3-5
13390	los_altos	minimum_lot_sqft	20000.0, 1000.0, 43560.0, 10000.0, 15000.0	Density. The maximum number of permitted dwelling units shall be calculated by subtracting twenty (20) percent of the gross area of the parcel and dividing the remainder by the required lot area per dwelling unit in the appropriate R1 District (ten thousand (10,000) square feet in the R1-10 District, fifteen thousand (15,000) square feet in the R1-H District, twenty thousand (20,000) square feet in the R1-20 District, and forty-three thousand five hundred sixty (43,560) square feet in the R1-40 District).	Title 14 - ZONING	4781	R1-40, R1-, R1-10, R1-20, R1
13392	los_altos	minimum_lot_sqft	10000.0, 11000.0, 15000.0	The minimum site area shall be ten thousand (10,000) square feet, except that on corner lots the minimum site area shall be eleven thousand (11,000) square feet and on flag lots the minimum site area shall be fifteen thousand (15,000) square feet.	Title 14 - ZONING	344	R1-10
13393	los_altos	minimum_lot_sqft	20000.0, 21000.0	The minimum site area shall be twenty thousand (20,000) square feet, except that on corner lots the minimum site area shall be twenty-one thousand (21,000) square feet.	Title 14 - ZONING	510	R1-

Efficiency of zone context collection

How many zones were found?



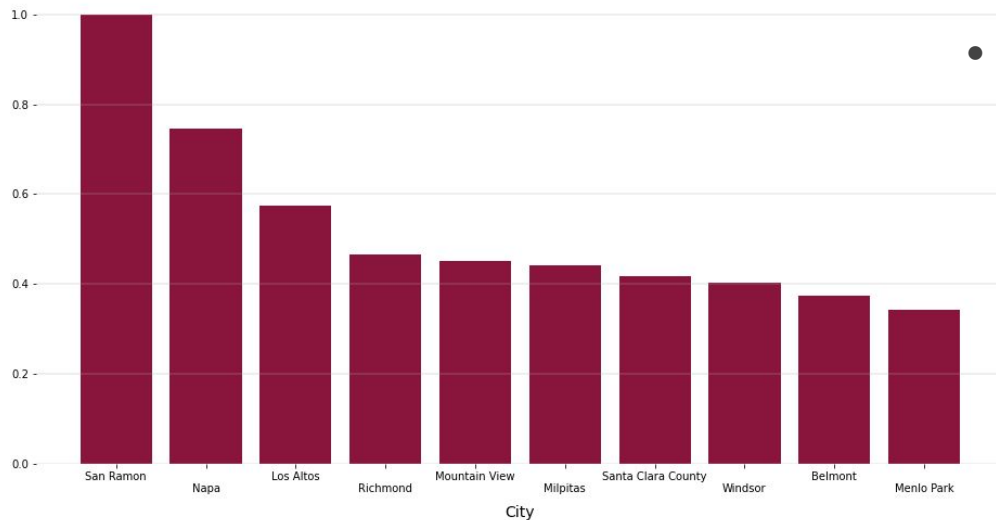
Derived / hidden values
requires more human
interventions

- Taking into account all city find rates:
 - Highest rate: Max dwellings unit area
 - Lowest rate: Max floor area ratio
- Interpretation:
 - Some values more “hidden” in the policy text
 - Values delegated to tables
 - Many tables omitted from original scraper
 - Units per lot / max far calculated rather than cited

Structure of City Policy Layout

Policy Structure (top 10)

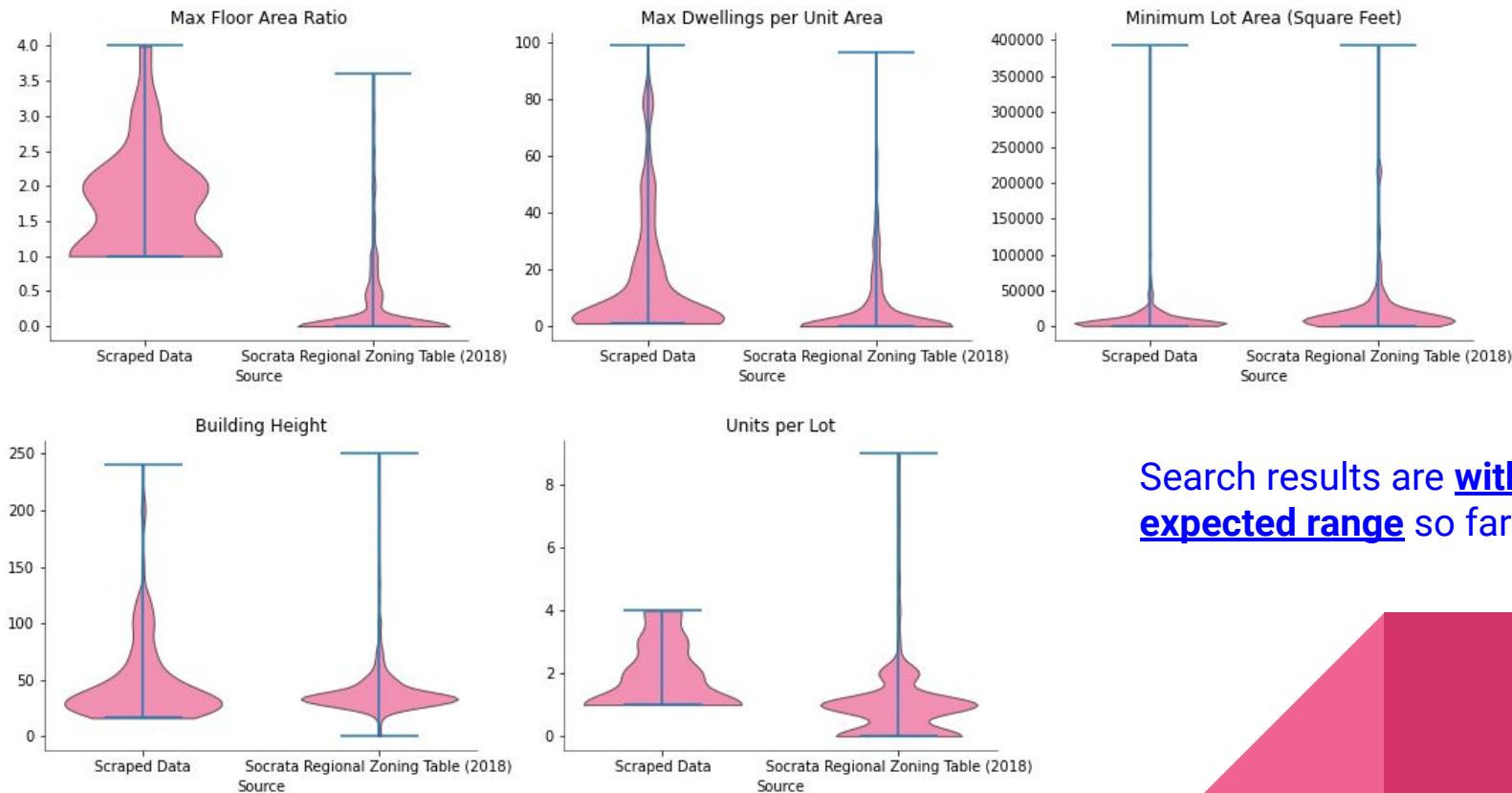
Zone found rates (all attributes)



- Cities attributes with no associated zones found:
 - Livermore, East Palo Alto, Atherton, El Cerrito, San Bruno, Daly City
- Summary statistics over all cities:
 - Mean: 0.125
 - Variance: 0.033
 - Median: 0.054
 - Max: 0.746 (Not including 1.0)
 - Min: 0.004 (Not including 0)

About **12%** of the attributes are found per city

Comparing to Socrata values



Search results are within
expected range so far

Next Steps

- Improve search results via refining regular expressions/attribute ranges (ongoing)
- Use probabilistic models - such as a multi-label classifier:
 - Training set: final table created at the end of pipeline
 - Resulting model could yield more promising results with scraped documents
- Revisit scrapers and update to pull tables from policy websites



Table Example 1 - Danville (AmLegal)

```
pd.read_html('amlegal_danville.html')

[
  0      0      1
  0      Table 32-22.1      Table 32-22.1
  1      Residential Zoning District      Minimum Lot Area,
      0      1
  0      R-6      6,000 square feet
  1      R-7      7,000 square feet
  2      R-10     10,000 square feet
  3      R-12     12,000 square feet
  4      R-15     15,000 square feet
  5      R-20     20,000 square feet
  6      R-40     40,000 square feet
  7      R-65     65,000 square feet
  8      R-100    100,000 square feet,
```

Table 32-22.1	
Residential Zoning District	Minimum Lot Area
R-6	6,000 square feet
R-7	7,000 square feet
R-10	10,000 square feet
R-12	12,000 square feet
R-15	15,000 square feet
R-20	20,000 square feet
R-40	40,000 square feet
R-65	65,000 square feet
R-100	100,000 square feet

Table Example 2 - Foster City (CodePub)

17.12.040 Area, bulk, yard and height regulations.



The following area, bulk, yard and height regulations shall apply in the R-1 district, except as modified by Chapter [17.54](#), Yards:

Minimum Building Site		Minimum Lot Area Per Dwelling Unit		Minimum Yards Required			Maximum Height Permitted	Maximum Coverage Permitted
Minimum		Area						
Width (ft.)	Area (sq. ft.)	(sq. ft.)	Front (ft.)	Side (ft.)	Rear (ft.)	Stories	Ft.	(%)
40	5,000	5,000	20	5	20	2	25	50

(Ord. 515 § 2 Exh. A (part), 2005; Ord. 38 § 1 (part), 1972: prior code § 10-404.014)

```
> MI
foster_city_tables = pd.read_html('FosterCity1712.html')

> MI
foster_city_tables

[ Minimum Building Site \
  Minimum Width (ft.) Minimum Area (sq. ft.)
0          40          5000

  Minimum Lot Area Per Dwelling Unit      Minimum Yards Required \
              (sq. ft.) Front (ft.)      Side (ft.)
0          5000          20          5

  Rear (ft.) Stories      Maximum Height Permitted Maximum Coverage Permitted
0          20      2          25          50 ]
```

Table Difficult cases - AmLegal

Table 19.28.060 Site Development Regulations			
	R1-5	R1-6, 7.5, 8, 10, 20, etc., and R1-6e	R1-a
A. Minimum net lot area ¹	5,000 square feet	the number multiplied by 1,000 square feet	10,000 square feet
B. Minimum lot width (at the front setback line)	50 feet	60 feet	75 feet
C. Landscaping	See Chapter 14.15 , Landscape Ordinance		Landscaping plans are required for all additions or new homes. The purpose of the landscaping is to beautify the property and to achieve partial screening of building

Cupertino

Antioch

HEIGHT, AREA & SETBACK REGULATIONS FOR PRIMARY STRUCTURE																	
Zone	Maximum Height Feet ^a	Minimum Building Site Sq. Ft.	Minimum Lot Width in Feet		Maximum Lot Coverage	Minimum Density Allowed (Units per Gross Developable Acre)	Maximum Density Allowed Units Per Gross Developable Acre ^d	Front Yard Minimum	Minimum Side Yard Required in Feet ^c		Minimum Rear Yard Required in Feet						
			Corner	Interior					Corner	Interior							
RE	TO BE DETERMINED BY CITY COUNCIL THROUGH PLANNED DEVELOPMENT PROCESS																
RR	TO BE DETERMINED BY CITY COUNCIL THROUGH PLANNED DEVELOPMENT PROCESS																
R-4	35	6,000	65	60	40%	NA	4 du/acre	ℓ	ℓ	5 ft.	20 ft.						
R-6	35	6,000	65	60	40%	NA	6 du/acre	ℓ	ℓ	5 ft.	20 ft.						
R-10	45	6,000	65	60	40%	NA	10 du/acre	ℓ	ℓ	5 ft.	10 ft.						
R-20	45	20,000	70	70	40%	NA	20 du/acre	ℓ	ℓ	5 ft.	10 ft.						
R-25	45	20,000	70	70	50%	20 du/acre	25 du/acre	ℓ	ℓ	5 ft.	10 ft. ^m						
R-35	45	20,000	70	70	50%	30 du/acre	35 du/acre	ℓ	ℓ	5 ft.	10 ft. ^m						
PBC	35	20,000	65	60	35%	NA	0	ℓ	ℓ	0 ft.	0 ft.						



Summary

- Policy Attribute Pipeline: Process, Organization, Storage
- Findings within results
- Next proposed steps to further improve pipeline