# PLANNING CASE TRACKING SYSTEM

## TRANSIT AND SOCIOECONOMIC ANALYSIS

CASE STUDY 1: TRANSIT-ORIENTED COMMUNITIES ENTITLEMENTS
CASE STUDY 2: ALCOHOL PERMIT ENTITLEMENTS
CREATING A DATA PIPELINE

# Executive Summary

Processing entitlements, or land use exemptions, is one of the Department of City Planning's core functions. The Performance Management Unit was particularly interested in diving into the last decade's worth of entitlement data, stored in the Planning Case Tracking System (PCTS) database, to glean key geographic and socioeconomic insights.

The ITA Data Science team took on this project and provided key geographic and socioeconomic analyses around entitlement activity. We also provided a standardized data processing pipeline for PCTS (**Section IV**), which provides the main framework for moving forward and exploring new analyses in the future.  All the data sources, data pipeline, and technical analyses were set up in a completely reproducible manner and documented in our [GitHub repository](GitHub repository).

## TOC Summary Findings

The first part of analysis focused on the Transit Oriented Communities (TOC) suffix (**Section II**). In setting forth the Transit Oriented Communities Guidelines, one planning / policy goal was to increase affordable housing and housing density near transit. This blanket policy did not affect all rail lines, TOC Tiers, or zoning equally.

- 58% of TOC entitlements were in Tier 3.
- 90% of TOC entitlements occurred in R3, R4, and C2 zone classes.
- 76% of the TOC entitlements occurred along the B - Red and/or D - Purple Lines.
- The Wilshire / Western rail station (Koreatown) had the most TOC entitlements.
- 40% of the TOC entitlements occur along seven bus routes, and 60% of the entitlements occur along 13 bus routes. Five of the top seven bus routes are Rapid lines.
- Zoning is a more important predictor of TOC entitlements than socioeconomic characteristics.
    - A 10% increase in favorable zoning (R3, R4, C2) in a tract was associated with a 23% increase in TOC entitlements, holding all other variables constant.
    - A 10% increase in the population of renters was associated with an 8% increase in TOC entitlements, holding all other variables constant.

## TOC Recommendations

DCP's role is to understand what makes R3, R4, and C2 zoning particularly attractive for developers. Ultimately, the city's planning land use and planning process must work in conjunction with the significant capital outlay and investment of new light rail and bus rapid transit lines. The dual goals of increased transit ridership and increased housing supply can be reached when the land use planning supports it. More transit riders mean more residents who don't use a car for all their trips;  the only residents that can afford to have fewer cars than the average are ones who live near transit. This analysis has shown some areas need further planning and policy to encourage this positive feedback loop.

## Socioeconomic Analysis Summary Findings

The second part of the analysis focused on all entitlement suffixes (**Section III**). The socioeconomic characteristics at the tract-level came from the American Community Survey 5-year estimates. In particular, the characteristics included were: % workers from zero-vehicle households, % renter population, % non-Hispanic white population, median household income, and total population.

We used a Poisson regression to look at how the total number of entitlements at the tract-level for the entire decade were related to these socioeconomic characteristics. Poisson regressions are often used for count data; in this case, we are looking at counts of entitlements per tract. For the 30 highest (citywide count over the last five years) entitlement suffixes, we provided a broad summary of whether each socioeconomic characteristic is positively or negatively related to the number of entitlements.

However, our TOC analysis showed the importance of including other important factors that are related to specific entitlements. Socioeconomic characteristics may not be as highly correlated to the entitlements once zoning is considered. We included an in-depth Poisson regression for stand-alone alcohol sale permits (CUB).

Our CUB findings:

- Zoning was a more important predictor of CUB entitlements than other socioeconomic characteristics.
  - An increase of 100 additional eligible parcels (commercial and manufacturing), there is a 20.5% increase in CUB entitlements, holding other variables constant.
  - A 10% increase in population of renters, there is a 17.2% increase in CUB entitlements, holding other variables constant. *Note:* Renters and zero-vehicle workers were highly collinear.
  - A 10% increase in the population of white non-Hispanic residents, there is a 9.5% increase in CUB entitlements, holding other variables constant.

## Socioeconomic Analysis Recommendations

We have developed much of the framework and data pipeline for answering future socioeconomic-related analyses. However, we found that one Poisson regression model fitted for all entitlement suffixes was insufficient. It showed some patterns related to socioeconomic characteristics, but with the omitted variable bias present, we do not interpret the coefficients without doing an in-depth Poisson regression.

Our two in-depth Poisson regression for the TOC and CUB suffixes showed that the most important predictor of entitlements was zoning. That said, socioeconomic characteristics are still important because of their equity implications. Higher renter populations were associated with more TOC entitlements; higher renter (or zero-vehicle workers), and higher white non-Hispanic populations were associated with more CUB entitlements.

## Report Author: Tiffany Chu (ITA)

## Project Team (ITA Data Science)

Hunter Owens *Project Manager, Programmer Analyst*
Tiffany Chu *Data Analyst*
Ian Rose *Data Engineer*

## Project Team (DCP)

### Performance Management Unit

Ashley Atkinson *Project Manager, City Planner*
Roy Samaan *Planning Assistant*
Jonathan Chiu *Planning Assistant*

### GIS

Betty Dong *GIS Chief*
Tom Tran *GIS Specialist*



# Acknowledgements

Our special thanks to Maryam Abbassi (ITA) and _____ (DCP) for establishing this partnership between our two departments. We recognize Hunter Owens for his project management and exploratory analysis, and Ian Rose for his data engineering, data pipeline work, dashboard deployment, and regression analysis.

We gratefully acknowledge the Performance Management Unit's time and resources in scoping and guiding this project through its various stages. In particular, we thank Ashley Atkinson (DCP / HCIDLA), Roy Samaan (DCP), and Jonathan Chiu (DCP) for providing feedback and lending their subject matter expertise. We also appreciate DCP lending its technical resources and implementing this work, specifically Mony Patel (DCP), Betty Dong (DCP) and Tom Tran (DCP). We also thank David Terukina (DCP) for providing us with PCTS database access.

# I. Introduction

The Department of City Planning (DCP) Performance Management Unit (PMU) submitted a project proposal to ITA's Data Science Federation (DSF) call for projects. DSF projects are typically paired with universities. Due to the complexity of the database and a project timeline beyond what students can reasonably finish in one semester, the ITA Data Science team took on this project internally.

Entitlements are requests from a property owner for an exemption or exception to what is permitted under current land use regulations. The entitlement is then subject to planning review before being allowed to proceed. Tracking and reviewing these entitlements is one of DCP's core functions; PMU was interested in diving into the last decade's worth of data and gleaning key geographic and socioeconomic insights. DCP records and tracks its entitlement applications  in the Planning Case Tracking System (PCTS). PMU wanted to leverage data analytics and GIS tools to provide insights into whether various types of entitlements submitted were correlated with neighborhood or location characteristics.

Once the ITA Data team started cleaning and processing the PCTS data, it was clear that this project needed a standardized data processing pipeline. Data analytics projects often suffer from reproducibility issues; it is hard to reproduce the same counts, trends, or results when different people answer the same question with the same dataset. It is imperative to have both a canonical data source (PCTS) and a standardized data processing pipeline. For this project, we had a canonical data source, but the project scope changed to accommodate a standardized data processing pipeline as one of the deliverables.

## A. Project Scope

The analysis portion of the project answered the following questions:

(1) *Transit / geographic analysis for TOC suffix only*

What kind of entitlement activity occurred near high-quality transit?
Are TOC entitlements spatially concentrated near certain rail stations, rail lines, or bus corridors?

This is covered in **Section II** of the report.

(2) *Socioeconomic analysis across all entitlement suffixes*
How are socioeconomic characteristics correlated with various entitlement suffixes?

(a) Transit Oriented Communities (TOC) in-depth regression analysis

(b) Stand-alone alcohol sale (CUB) in-depth regression analysis.

This is covered in **Section III** of the report.

Establishing a data pipeline for this project meant familiarizing ourselves with how DCP generates its internal SQL queries for its internal PCTS reports. We also developed several tools that make this project reproducible and lay the groundwork for future data analytics projects DCP would undertake using the PCTS data. We created a Python package called `laplan` and use the Civis Analytics platform to automate the data cleaning pipeline and deploy dashboards. The data pipeline process and tools developed and used are covered in **Section IV**.

## B. Data Sources

**Table 1** lists all the data sources used to conduct the analysis. The vast majority of data came from open data portals, with the exception of DCP's PCTS database. All the data sources are documented in our data catalog and are stored in an S3 bucket.[1]

**Table 1 Data Sources**

| Data | Source | Years Available | Data Used In |
|---|---|---|---|
| Entitlements | DCP - Planning Case Tracking System (PCTS) | 2010-2019 | All analyses |
| PCTS documentation | DCP website for prefixes, suffixes[2], DCP emailed PCTS codebook[3] | | All analyses |
| Parcels | County of LA, Open Data Portal, Tax Assessor Parcels[4] | 2006-2019 | All analyses |
| Zoning (ZIMAS) | City of LA, DCP, GeoHub (geospatial open data portal)[5] | 2019 | All analyses |
| Zoning documentation | City of LA, DCP (emailed)[6] [7] | | All analyses |
| TOC tiers and TOC parcels | City of LA, DCP (emailed) | 2017 | Geographic / transit analysis |
| Metro existing rail stations and rail lines | LA Metro Developer GIS Data[8], `Metro_Rail_Lines`, `Metro_Rail_Stations` | 2019 | Geographic / transit analysis |

[1] To ensure that we have the data from open data portals even if URLs break, we save versioned copies in our S3 bucket. The `make mirror` command in our Makefile creates a versioned copy of the open data sources in the S3 bucket.
[2] https://planning.lacity.org/resources/prefix-suffix-report
[3] https://github.com/CityOfLosAngeles/planning-entitlements/blob/master/references/PCTS_TABLE_FIELDS.xlsx
[4] https://data.lacounty.gov/Parcel-/Assessor-Parcels-Data-2006-thru-2019/9trm-uz8i
[5] https://geohub.lacity.org/datasets/49ad06a6b8c945debbbea865b1832ee2_0
[6] https://github.com/CityOfLosAngeles/planning-entitlements/blob/master/references/Zoning_Types_Master_Table_092518.xlsx
[7] https://github.com/CityOfLosAngeles/planning-entitlements/blob/master/references/Zoning_Code_Summary_March2020.pdf
[8] https://developer.metro.net/bus-rail-gis-data/

| Data | Source | Years Available | Data Used In |
|---|---|---|---|
| Metro planned rail stations | LA Times, Data Desk, LA Metro Maps[9], `planned-crenshaw-stops`, `planned-purple-line-extension-stops`, `planned-regional-connector-stops` | 2015 | Geographic / transit analysis |
| Metro bus stations and bus lines | LA Metro Developer GIS Data[10], `BusStopServingLines1219`, `BusLineServingStops1219`, `RapidBRT1219`, | 2019 | Geographic / transit analysis |
| Metrolink rail stations | City of LA, GeoHub[11] | 2019 | Geographic / transit analysis |
| Metrolink rail lines | Southern California Association of Governments, Open Data Portal[12] | 2019 | Geographic / transit analysis |
| General Transit Feed Specification (GTFS) for LA Metro | LA Metro, GTFS[13], `gtfs_bus` | 2020 | Geographic / transit analysis |
| GTFS for Big Blue Bus | Big Blue Bus, GTFS Service, `current` | 2020 | Geographic / transit analysis |
| GTFS for Culver City Bus | Open Mobility Data, January 2020[14] | 2020 | Geographic / transit analysis |
| City boundary | City of LA, GeoHub[15] | 2020 | All analyses |
| Census tract | City of LA, GeoHub[16] | 2010 | All analyses |
| American Community Survey, 5-year tables | US Census Bureau[17], tables downloaded: B01001, B02001, B19001, B25008, S1903, S0801, S0802 | 2010-2018 (used 2018 only) | Socioeconomic analysis |

[9] https://github.com/datadesk/lametro-maps
[10] https://developer.metro.net/bus-rail-gis-data/
[11] https://geohub.lacity.org/datasets/6f6c4677365b4418bd585db2ef8e201f_186
[12] https://gisdata-scag.opendata.arcgis.com/datasets/ac623114be664c7488836db5c22e3566_0
[13] https://gitlab.com/LACMTA/gtfs_bus/-/raw/aef844c79c31c40ef751a3472b3882406307b05a/gtfs_bus.zip?inline=false
[14] http://transitfeeds.com/p/culver-city-bus/1057
[15] https://geohub.lacity.org/datasets/09f503229d37414a8e67a7b6ceb9ec43_7
[16] https://geohub.lacity.org/datasets/census-tracts-2010-population
[17] https://data.census.gov/cedsci/

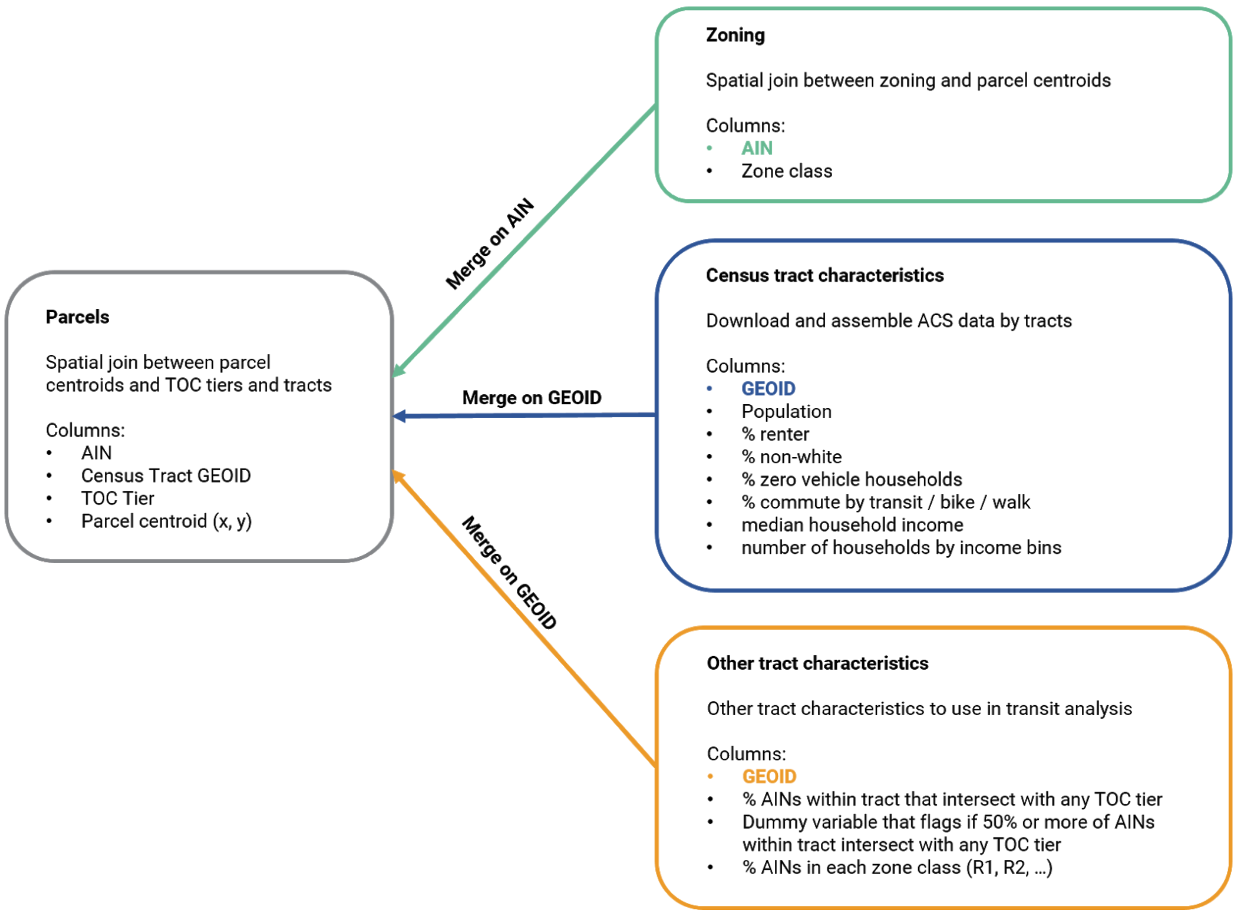
## C. Methodology for Parcel Base File

Since parcels are the most granular level of data we have, we used it as the base file. We created several *crosswalks* that link parcels with other characteristics we want, such as, including census tract, zone class, and entitlements. These crosswalks were key "puzzle pieces" in constructing the analysis table.

The process for compiling the PCTS data, which comes from four tables in the PCTS database, is discussed more in detail in **Section IV - Data Pipeline**.

**Figure 1** shows our approach in bringing all these various data sources together at the parcel-level. Parcel with tier information, parcel with tract GEOID, zoning, and other tract characteristics were constructed through spatial joins. The census tract characteristics table was constructed by downloading relevant American Community Survey (ACS) 5-year tables and merging them together.

**Figure 1 Creating the Parcel Base File**




*Source: Analysis by ITA Data Science*

## LA County Assessor Parcels

We used LA County parcel data from 2006-2019 to construct our full list of unique parcels. When we used just the 2019 parcel file, about 10% of the entitlements were unable to be linked to a parcel. Including all the historical AINs allowed us to attach all the entitlements to parcel IDs. The trade-off in using the 2006-2019 AINs is that we only have parcel centroids, whereas the 2019 AIN file came with parcel polygons. Doing spatial joins between two sets of polygons is computationally expensive. Our spatial joins always involved finding a point-in-polygon, such as which parcel centroid fell in which tract polygon, making parcel centroids a reasonable choice.

Duplicate Assessor Identifier Numbers (AIN) were removed, but duplicates based on parcel centroids were kept. This means that the same parcel (based on the same X, Y point) can be linked to different AINs over time. Unsure of which AIN is actually used in PCTS, we kept them all. Parcels were then spatially joined to census tracts and TOC tiers using parcel centroids. Counting the number of unique parcels within a census tract was done by counting unique parcel centroids, not AINs; the same physical boundary of a parcel would only be counted once, regardless of how many AINs were associated with it.

## Zoning

Zoning information is stored in ZIMAS. We used the version provided on GeoHub, with fewer columns than what DCP typically can access. DCP's [Guide to Zoning String] showed that the zoning string is made up of component parts. The zoning string contains information about prefix on (Q)ualified or (T)entative zone classifications, zone class, the height district, (D)evelopment limits, and specific plans and overlays applicable.

We parsed the zoning string to get the relevant component parts, and used a spatial join to attach zoning information to parcels. We also counted the total number of unique parcels (based on parcel centroid) within a tract and calculated the percentages of parcels within the tract belonging to each zone class. There were 63 zone classes captured. Keeping the zone classes as disaggregated as possible provided the flexibility needed for the crosswalk to be used across multiple analyses.

Zone classes included:

'A1', 'A2', 'A2P',

'C1', 'C1.5', 'C2', 'C4', 'C5', 'CM', 'CR',

'GW',

'HJ', 'HR',

'M', 'M1', 'M2', 'M3', 'MR1', 'MR2',

'NI',

'OS',

'P', 'PB', 'PF',

'R1', 'R1H1', 'R1P', 'R1R3', 'R1V1', 'R1V2', 'R1V3', 'R2', 'R2P', 'R3', 'R3P', 'R4', 'R4P', 'R5', 'R5P',

'RA', 'RAS3', 'RAS4',

'RD1.5', 'RD2', 'RD3', 'RD4', 'RD5', 'RD6',
'RE', 'RE11', 'RE15', 'RE20', 'RE40', 'RE9',
'RMP', 'RS', 'RU', 'RW1', 'RW2',
'RZ2.5', 'RZ3', 'RZ4', 'RZ5'

See **Section IV - Data Pipeline** and **Appendix A - Zoning** for more detail.

## Census Tract Characteristics

The American Community Survey (ACS) 5-year estimates provided census tract-level socioeconomic characteristics. We downloaded all the tables using the Census API and processed all the data to fit our analysis needs in scripts (download data, clean data part 1, clean data part 2, subset for analysis, and assemble as a crosswalk). We were interested in socioeconomic characteristics such as the population, race/ethnicity, median income, zero vehicle households, renter-occupied households, and commute mode.

**Table 2** lists the ACS tables that were downloaded and stored.[18] Although we downloaded 2010-2018 data, we only used the 2018 data in the analysis. However, we cleaned the data for all years and resolved any major data changes that may have occurred in 2014.

**Table 2 ACS Tables**

| Outcome | ACS Table Name | Main Dataset | Years |
|---|---|---|---|
| median income | HOUSEHOLD INCOME IN THE PAST 12 MONTHS (INFLATION-ADJUSTED) (B19001) by income ranges and has supplemental tables for each race | ACS 5-year | 2010-2018 |
| median income | MEDIAN INCOME IN THE PAST 12 MONTHS (INFLATION-ADJUSTED) (S1903) for race categories | ACS 5-year Subject | 2010-2018 |
| % zero vehicle workers | MEANS OF TRANSPORTATION TO WORK BY SELECTED CHARACTERISTICS (S0802) | ACS 5-year Subject | 2010-2018 |
| commute mode | COMMUTING CHARACTERISTICS BY SEX (S0801) | ACS 5-year Subject | 2010-2018 |
| % renter-occupied | TOTAL POPULATION IN OCCUPIED HOUSING UNITS BY TENURE (B25008) | ACS 5-year | 2010-2018 |
| % white or % non-white | RACE (B02001) | ACS 5-year | 2010-2018 |
| % white, non-Hispanic or converse | SEX BY AGE (B01001) for race and ethnicity categories | ACS 5-year | 2010-2018 |

*Source: US Census Bureau, American Community Survey*

[18] A more comprehensive version of Table 2 is available at: https://github.com/CityOfLosAngeles/planning-entitlements/blob/master/references/Census_Tables.xlsx

See **Section IV - Data Pipeline** and **Appendix A - Census** for more detail.

## Other Tract Characteristics

We created a crosswalk containing other tract characteristics further derived from the zoning and Census data already gathered. This enabled us to use these tract characteristics across multiple analyses without generating it each time. This crosswalk was adapted and expanded through iterations; we sliced the data in various ways and found the characteristics that could be used across multiple analyses. The main categories of characteristics involved understanding the relationship between census tracts and TOC tiers and zoning.

Knowing how many parcels within a tract fell into one of the TOC tiers was one piece of information, but a further step was grouping a tract as being a TOC tract or not. A TOC tract was defined as 50% or more of the parcels intersecting *any* one of the TOC tiers.

Through our descriptive charts, we found that the majority of TOC entitlements occurred in just a handful of zone classes. Even though a much larger list of zone classes were eligible for the TOC entitlement, these “favorable” zone classes were a proxy for unobserved incentives that developers faced or operated under. Including these favorable zone classes as a predictor in our regression analysis proved extremely helpful.

# II. Geographic / Transit Analysis

In 2016, Measure JJJ passed and a Transit Oriented Communities (TOC) program was included in the zoning code. DCP "create[d] TOC guidelines for all housing developments within a half-mile radius of a major transit stop", providing "a package of new incentives for building affordable housing near public transit (City of Los Angeles, Department of City Planning, Transit Oriented Communities Incentive Program)."[19] The goal of that policy was to encourage affordable housing equally across all bus / rail corridors. The entitlement activity in these corridors would be a proxy for increased affordable housing or denser housing in transit-rich areas.

One major question of interest was: **what kind of entitlement activity occurred near high-quality transit?** Of course, the TOC suffix could occur near high-quality transit, but perhaps, relative to other suffixes, TOC entitlements were only a small part of entitlement activity in those corridors. Another question to explore is **where *are* the entitlements spatially concentrated and where was it lacking?** Which rail stations or bus corridors are they concentrated along? Those lacking entitlement would benefit from other policy and planning programs to encourage increased affordable housing.

## A. Methodology

DCP published [TOC guidelines] in November 2017 in response to the passage of Measure JJJ. The guidelines defined what was considered being near high-quality transit. Four tiers were established, with Tier 4 being the closest to the intersection of Metro rail stations and Rapid bus lines, and Tier 1 being much further out from Metro rail stations, the intersection of Regular and Rapid bus lines, or the intersection of two Regular bus lines. **Figure 2** shows how each tier was defined based on proximity to various types of transit.

Each tier was associated with base incentives, which allow for increased residential density (higher density for higher tiers), increased floor area ratios, and the decreased parking minimum requirements. There were also additional incentives, which further adjust setbacks, open space, lot coverage, and building heights. Each tier was also bundled with certain affordable housing requirements, which set forth certain percentages of units to be made available for extremely low or very low households. Taken together, the goal was to increase housing density near transit while making sure a portion of that increase was set aside as affordable units.

[19] https://planning.lacity.org/plans-policies/transit-oriented-communities-incentive-program

**Figure 2 TOC Affordable Housing Incentive Area Tiers**

| Type of Major Transit Stop | Tier 1 (Low) | Tier 2 (Medium) | Tier 3 (High) | Tier 4 (Regional) |
|---|---|---|---|---|
| | Distance to Major Transit Stop | | | |
| **Two Regular Buses** (intersection of 2 non Rapid Bus* lines, each w/ at least 15 min. average peak headways) | 750 - 2640 ft. | < 750 ft. | - | - |
| **Regular plus Rapid Bus*** (intersection of a Regular Bus and Rapid Bus line) | 1500 – 2640 ft. | 750 – <1500 ft. | < 750 ft. | - |
| **Two Rapid Buses*** (intersection of two Rapid Bus lines) | - | 1500-2640 ft. | < 1500 ft. | - |
| **Metrolink Rail Stations** | 1500 – 2640 ft. | 750 – <1500 ft. | < 750 ft. | - |
| **Metro Rail Stations** | - | - | ≤ 2640 ft. | < 750 ft. from intersection with another rail line or a Rapid Bus* |

*Source: DCP, Transit Oriented Communities Guidelines, Chart 1*

In order to quantify how much of the entitlement activity occurred near certain rail stations or bus lines, we needed to reconstruct the TOC tiers using the General Transit Feed Specification (GTFS) data for LA Metro, Santa Monica Big Blue Bus, and Culver CityBus routes. GTFS is a standard format for transit agencies to report the bus schedules for each route. The guidelines state that only bus lines with at least an average of 15 minute headway should be counted. We allow for an extra 20% buffer and err on the side of selecting ineligible bus routes, to make sure we caught all bus routes. We noted that the DASH bus service is also included in DCP's list of transit agencies, but was excluded from our analysis. However, DASH buses that served the downtown core are often co-located near other major bus routes, and the entitlements would still be otherwise captured. The GTFS data allowed us to calculate AM and PM weekday peak frequencies for bus routes, and selected only the bus routes that met the 15 minute cut-off.[20]

Next, buffers around the bus stops and rail stations were drawn. DCP did have the boundaries of the four TOC tiers as well as the list of parcels that are TOC-eligible, which was compiled with additional verification. We used spatial joins to attach the TOC tiers to each parcel. Parcels that were attached to multiple tiers were assigned to the higher tier (with more attractive incentives). Then, we attached the bus stop(s), bus line(s), and rail stations(s) to each parcel. At this point,

[20] This GTFS work is documented in this script and this notebook.

the same parcel would appear multiple times if it was near multiple bus stop(s) and rail station(s).[21]

Only parcels that fell within the TOC tiers are TOC-eligible, and we focused on the entitlement cases that took place in those areas. In terms of PCTS entitlements, we included cases from October 2017 through March 2020. Cases with the prefixes ENV (environmental), PAR (pre-application review), and ADM (administrative) were excluded. If a case contained a “TOC” suffix, it was counted as a TOC entitlement, even if other suffixes were also requested. Entitlements without the “TOC” suffix were counted as Non-TOC entitlements.

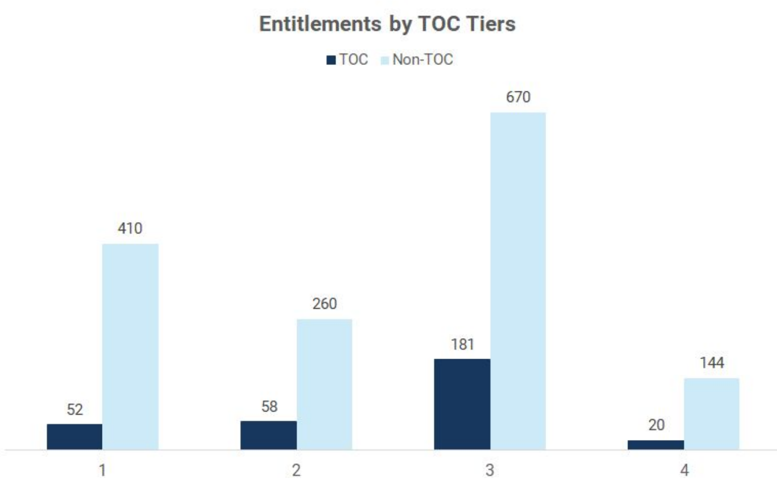
# B. Analysis Findings

## By Tier and Zone Class

First, we looked at the distribution of TOC and Non-TOC entitlements across the TOC tiers. Of course, TOC entitlements only occurred in one of the TOC tiers. Non-TOC entitlements can occur inside or outside the TOC tiers. Entitlements asking for TOC-eligibility verification were excluded from the analysis.

**Figure 3** shows that 58% of the TOC entitlements across the four eligible tiers occurred in Tier 3. Being <750 ft of a Regular + Rapid, <1,500 ft of two Rapid lines, <750 ft of Metrolink rail station, or <=2,640 ft of a Metro rail station all qualified as Tier 3. Tier 3 had the biggest area in terms of eligible zones than other tiers, resulting in the tier capturing the majority of Non-TOC entitlements.  In terms of the fraction of TOC entitlements *within* a tier, Tier 3 (21%) and Tier 2 (18%) had the highest proportions of TOC to Non-TOC entitlements.

**Figure 3 Entitlements by TOC Tiers**




*Source: PCTS, TOC Tiers; analysis by ITA Data Science*

[21] An individual bus stop that has several bus routes passing through would be listed as the same bus stop ID. But, bus stops even 20 feet apart would have separate IDs. For example, there are two unique bus stops on Flower & 7th St. One bus stop is for Metro bus routes 438, 448, and 534. The second bus stop, 50 feet away, is for the Metro Silver Line and Metro route 460.

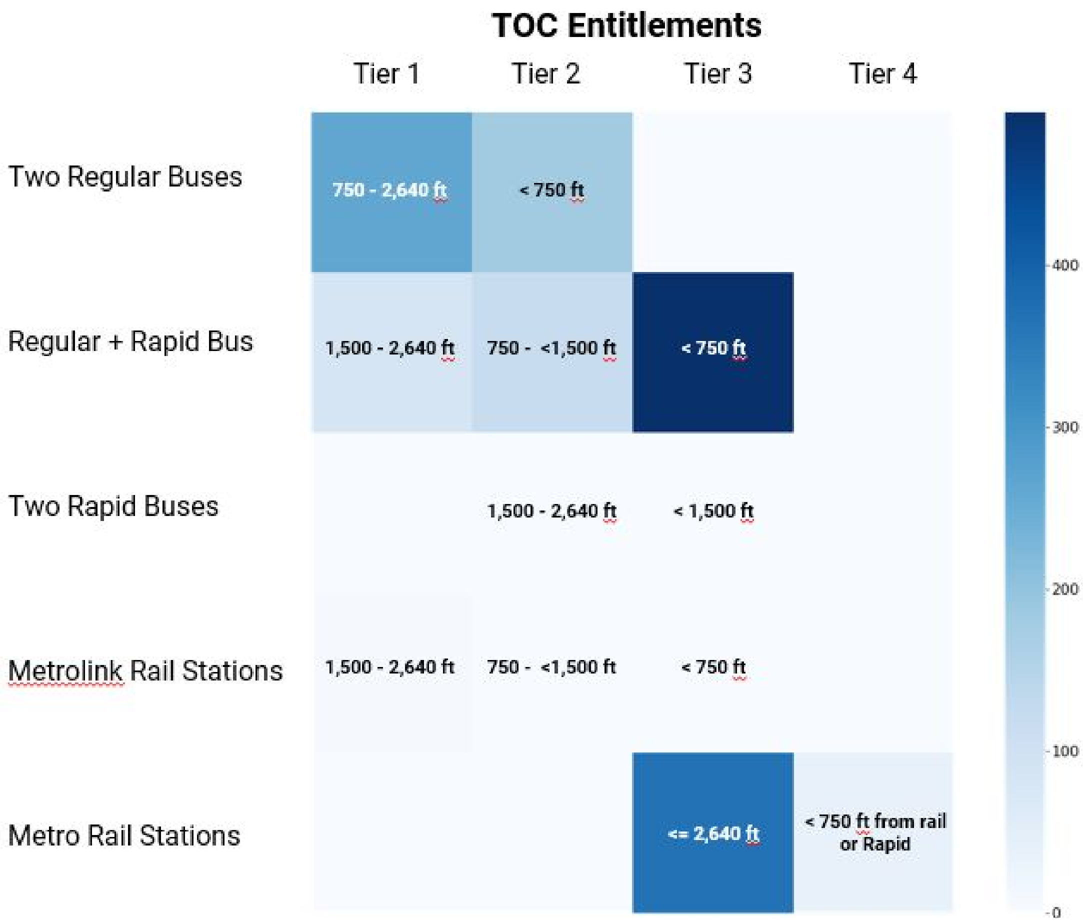
**Figure 4** shows the number of TOC entitlements according to which rule was used to determine the tier assignment. Rather than only knowing that an entitlement was Tier 3, we identified which rule(s) with which the parcel is associated. That is, was the parcel was <750 ft of a Regular + Rapid and/or <1,500 ft of two Rapid lines and/or <750 ft of Metrolink rail station and/or <=2,640 ft of a Metro rail station? Certainly, multiple rules could be applied. For example, a parcel in the downtown core near a Metro rail station would likely also be near the intersection of two Rapid bus lines, since bus stops tend to be co-located near rail stations for better connectivity. In that case, if the parcel had a TOC entitlement, it would be counted twice, once for being near a Metro rail station, and a second time for being near the intersection of two Rapid lines.

The rules that generated the most TOC entitlements were being located within 750 ft of the intersection of Regular + Rapid bus (Tier 3), within a 0.5 mile of a Metro rail station (Tier 3), and between 750 ft - 2,640 ft of the intersection of two Regular buses (Tier 1). Tier 3 had the most number of TOC entitlements, but most of these occurred near where a Regular + Rapid bus intersected and near Metro rail stations. Unsurprisingly, TOC entitlements within 0.5 mile of Metro rail stations were popular, as much of the public transit literature agrees that people will comfortably undertake a quarter to half mile walk to reach transit, and are even willing to walk up to 0.7 mi to access heavy rail.[22] [23] [24]

Interestingly, few occurred near where two Rapid buses intersected, which could be due to the relative scarcity of that relative to Rapid buses intersecting with Regular buses. Rapid buses tend to run along major corridors, such as Wilshire Blvd, Santa Monica Blvd, Olympic Blvd, Venice Blvd, Sepulveda Blvd, Vermont Av, and Western Av. Many of these are parallel routes, with only a few intersections between the east-west and north-south routes.

[22] https://safety.fhwa.dot.gov/ped_bike/ped_transit/ped_transguide/ch4.cfm
[23] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4941821/
[24] https://media.metro.net/images/service_changes_transit_service_policy.pdf

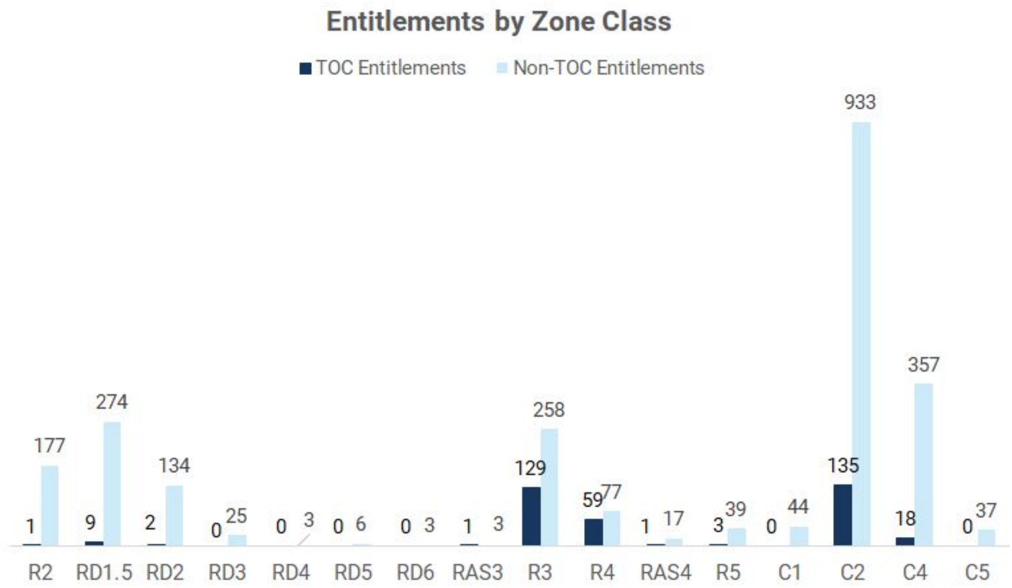
**Figure 4 Number of TOC Entitlements by Tier Rule**




*Source: PCTS, TOC Tiers, GTFS; analysis by ITA Data Science*

In terms of zoning, even though there were theoretically many zone classes that qualify for TOC entitlements, 90% of TOC entitlements occurred in the R3, R4, and C2 zone classes (**Figure 5**). Clearly, developers viewed the R3, R4, and C2 zone classes as being financially viable for this type of denser housing construction with affordable housing requirements sought after in the TOC guidelines. The Non-TOC entitlements that are occurring on TOC-eligible parcels were mostly taking place in RD1.5, R2, and R3 zone classes. That a significant portion of TOC and Non-TOC entitlements occurred in the R3 zone areas could mean that there are certain attractive or particular characteristics about this zoning that generates a lot of entitlements. It could be a sweet spot for certain types of development, and warrants additional exploration and planning domain knowledge to determine those factors.

**Figure 5 Entitlements by Zone Classes**



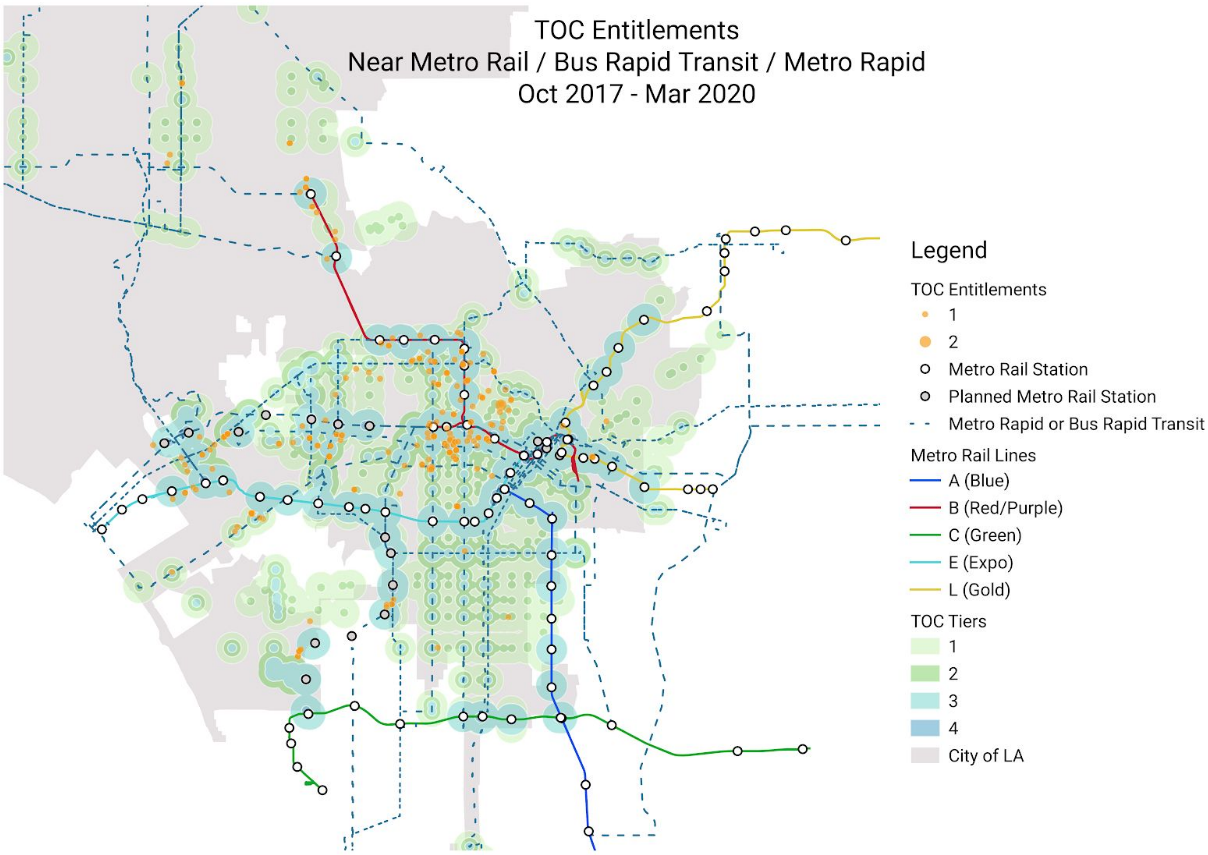

*Source: PCTS, ZIMAS zoning; analysis by ITA Data Science*

Putting all the TOC entitlements onto a map with the Metro rail, bus rapid transit, and Rapid lines showed that most of the TOC entitlements appeared to be concentrated around the B - Red and D - Purple Lines (**Figure 6**). Along the planned Metro lines, there was some activity around the Purple Line Extension near Century City and the Crenshaw Line around Crenshaw Blvd / Slauson Av.

Where there is little or zero TOC entitlement activity was also notable; there were few TOC entitlements around the existing light rail lines, including the A - Blue Line, C - Green Line, and E - Expo Line, L - Gold Line. Areas near the A - Blue Line have long been documented as areas of disinvestment in the urban planning literature. Particularly the A - Blue Line, the first rail line to open in Los Angeles, consistently high ridership, and lower rates of car ownership, improving development near those stations would have significant equity implications.[25] However, identifying whether other light rail lines are similarly missing those development antecedents is crucial to improving and encouraging TOC development near all light rail lines. Most of the rail lines converge on the downtown core, where specific plans and overlays provided better incentives for developers than what is listed under the TOC guidelines. However, there are plenty of areas that could and should sustain TOC development outside of the downtown core and along the B - Red / D - Purple Lines.

[25] Loukaitou-Sideris, Anastasia and Tridib Banerjee (2000). The Blue Line Blues: Why the Vision of Transit Village May Not Materialize Despite Impressive Growth in Transit Ridership. https://escholarship.org/content/qt8jd663ht/qt8jd663ht.pdf

**Figure 6 Map of TOC Entitlements**

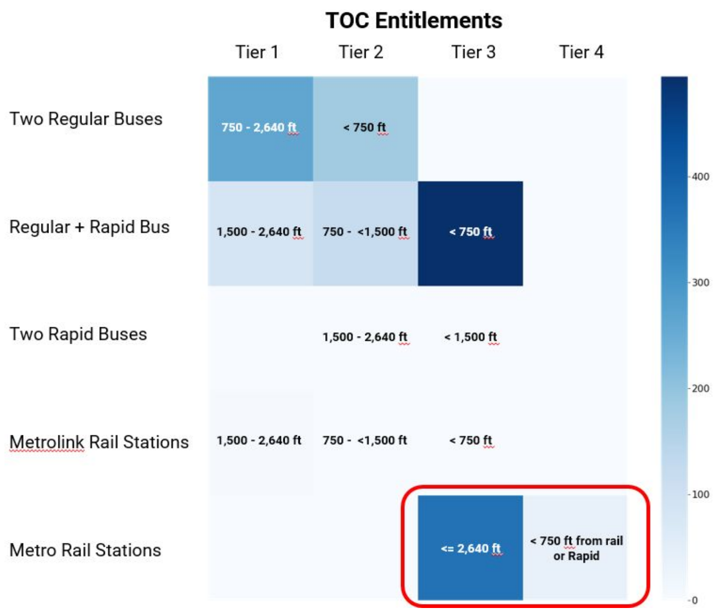



Note: The B - Red Line and D - Purple Line are shown as one line. Prior to LA Metro's 2020 renaming of the lines, these two lines were considered as one heavy rail line, sharing a significant portion of its stops. In light of the Purple Line Extension connecting between Koreatown to West LA, LA Metro renamed the Purple Line the D line.

*Source: PCTS, LA Metro Developer GIS Data; TOC Tiers; analysis by ITA Data Science*

## By Rail Station

After establishing that TOC entitlements occurred mostly in Tier 3 with R3, R4, and C2 zoning, we quantified TOC entitlement activity by Metro rail stations (circled in red below). We look across Tiers 3 and 4 to see where TOC entitlements were occurring.




Half of the TOC entitlements associated with Metro Rail (Tiers 3 and 4) occurred near seven stations; these seven stations are all along the B - Red and/or D - Purple Lines (**Table 3**). The Wilshire / Western Station in Koreatown is associated with the most TOC entitlements, followed by the Vermont / Santa Monica Station in East Hollywood. The full table is available in **Appendix B**.

**Table 3 TOC Entitlements by Rail Station (Top Rail Stations)**

| Station | Line | # TOC | Station's % TOC Entitlements |
|---|---|---|---|
| Wilshire / Western Station | D - Purple | 51 | 12.7% |
| Vermont / Santa Monica Station | B - Red | 34 | 8.5% |
| Wilshire / Vermont Station | B - Red / D - Purple | 28 | 7.0% |
| Wilshire / Normandie Station | D - Purple | 23 | 5.7% |
| Hollywood / Vine Station | B - Red | 22 | 5.5% |
| Vermont / Beverly Station | B - Red | 21 | 5.2% |
| Westlake / MacArthur Park Station | B - Red / D - Purple | 20 | 5.0% |

*Source: PCTS, LA Metro Developer GIS Data; analysis by ITA Data Science*

## By Rail Line

We also explored how TOC entitlements were concentrated by rail lines and bus routes. By aggregating station-level data to line-level, we could identify the specific corridors that were benefiting from this policy.

**Table 4** shows that 76% of the TOC entitlements occurred along the B - Red, D - Purple, or B - Red / D - Purple Lines. The B - Red Line and D - Purple Line share six stations, between Union Station and Wilshire / Vermont. The existing D - Purple Line runs between Koreatown and Downtown LA; the future D - Purple Line extends to Westwood (UCLA). TOC entitlements were eligible and did occur along the planned extension of the D - Purple Line. The existing B - Red Line runs between North Hollywood and Downtown LA.

There was little activity around the existing E - Expo Line and L - Gold Line. These light rail lines have been around for quite some time. The E - Expo Line is the newest; its first phase opened in 2012, and its second phase to Santa Monica completed in 2016. Even the Crenshaw / LAX Line (possibly named K - Line), still under construction with its first phase opening in 2021/2022, had the same level of TOC entitlement activity as the E - Expo and L - Gold Lines.

In addition to looking at how many TOC entitlements occurred along a rail line, we also counted the number of stations with TOC entitlements and calculated the average number of entitlements per station. This helps us understand whether TOC entitlements were dispersed around across all the stations along a rail line, or if they were fairly concentrated around a few stations. For example, the B - Red / D - Purple segment had 53 TOC entitlements over 3 stations, and its 17.7 average TOC entitlements per station showed a high concentration of TOC entitlements around those 3 stations. That differed from the E - Expo Line, which had 49 TOC entitlements, roughly similar to the B - Red / D - Purple, but these entitlements were spread across 10 stations, resulting in 4.9 average TOC entitlements per station, much lower than the 17.7 for the B - Red / D - Purple.

The average TOC entitlements per station metric doesn't speak to whether concentration is desirable or not, but simply helps us understand whether the TOC entitlements were occurring at punctuated intervals throughout a line or heavily concentrated in a few neighborhoods.

**Table 4 TOC Entitlements by Rail Line**

| Rail Line | # TOC | # Stations | Line's %TOC Entitlements | Avg Entitlements per Station (# TOC / # stations w/ TOC) |
|---|---|---|---|---|
| D - Purple | 129 | 7 | 32.1% | 18.4 |
| B - Red | 125 | 8 | 31.1% | 15.6 |
| B - Red / D -Purple | 53 | 3 | 13.2% | 17.7 |
| E - Expo | 49 | 10 | 12.2% | 4.9 |
| K - Crenshaw | 26 | 4 | 6.5% | 6.5 |
| L - Gold | 16 | 6 | 4.0% | 2.7 |
| A - Blue | 2 | 1 | 0.5% | 2.0 |
| A - Blue / E- Expo | 2 | 1 | 0.5% | 2.0 |

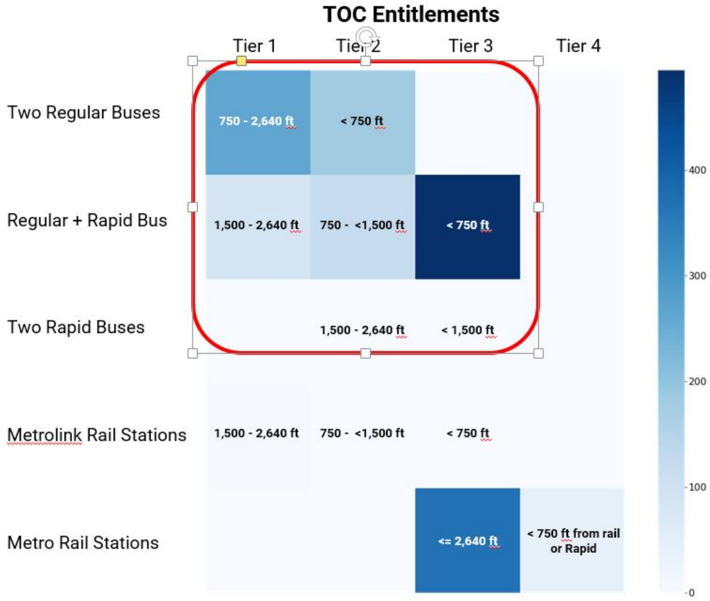
*Source: PCTS, LA Metro Developer GIS Data; analysis by ITA Data Science*

## By Metrolink Station

Metrolink is considered commuter rail, but very few of the TOC entitlements were associated with being near any Metrolink station. In fact, we found only four TOC entitlements associated with the Van Nuys Metrolink Station in Tier 1, located on parcels 1,500 - 2,640 ft away from the station. We conclude that although Metrolink is included in the guidelines for TOC development, it is hardly a major contributor of TOC entitlements.

## By Bus Route

We looked across multiple tiers to see which bus routes had concentrations of TOC entitlements. This is equivalent to looking across tiers by transit mode to see which stations or lines had concentrations of TOC entitlements (circled in red below).




In terms of bus routes, 40% of the TOC entitlements occurred along seven bus routes, and 60% of the entitlements occurred along 13 bus routes (**Table 5**). The full table is in **Appendix B**. Of these 13 bus routes, only one is from Santa Monica Big Blue Bus (R12), and all others are LA Metro bus routes. The R12 connects UCLA / Westwood to the Expo Line.

Of the top seven routes, five are Rapid lines:

- 754 (Hollywood - 105 Fwy via Vermont)
- 757 (Hollywood - Hawthorne via Western)
- 720 (Commerce - DTLA - Santa Monica via Wilshire)
- 780 (Pasadena - Mid City via Fairfax / Hollywood / Colorado)
- 710 (Hollywood / Vine - South Bay Galleria via Crenshaw)

**Table 5 TOC Entitlements by Bus Route**

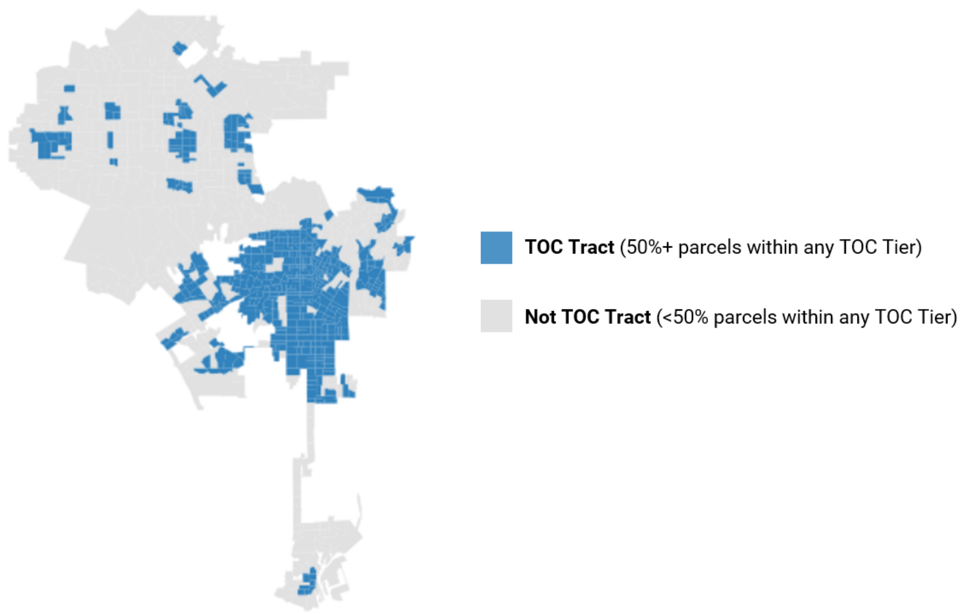
| Bus Route | # TOC | Route's % TOC Entitlements |
|---|---|---|
| 754 | 141 | 9.3% |
| 757 | 125 | 8.2% |
| 720 | 102 | 6.7% |
| 780 | 86 | 5.7% |
| 207 | 63 | 4.2% |
| 710 | 61 | 4.0% |
| 10/48 | 57 | 3.8% |
| 33 | 53 | 3.5% |
| 704 | 49 | 3.2% |
| 204 | 47 | 3.1% |
| 603 | 43 | 2.8% |
| R12 | 42 | 2.8% |
| 206 | 41 | 2.7% |

*Source: PCTS, GTFS, LA Metro Developer GIS Data; analysis by ITA Data Science*

## Socioeconomic Characteristics of TOC Tracts

Bringing in socioeconomic characteristics from ACS data, we can see whether TOC-eligible tracts were similar or different compared to Non-TOC-eligible tracts. A tract is defined as a TOC tract if 50% or more of the parcels within the tract fell within any one of the TOC Tiers. Given how TOC Tiers were drawn as concentric rings of circular buffers around rail stations and bus stops, these tiers divided tracts into various shapes. The simplest way to categorize tracts is to look across all the tiers and see if half or more of the parcels fall into one of the tiers. **Figure 6** plots the TOC tracts.

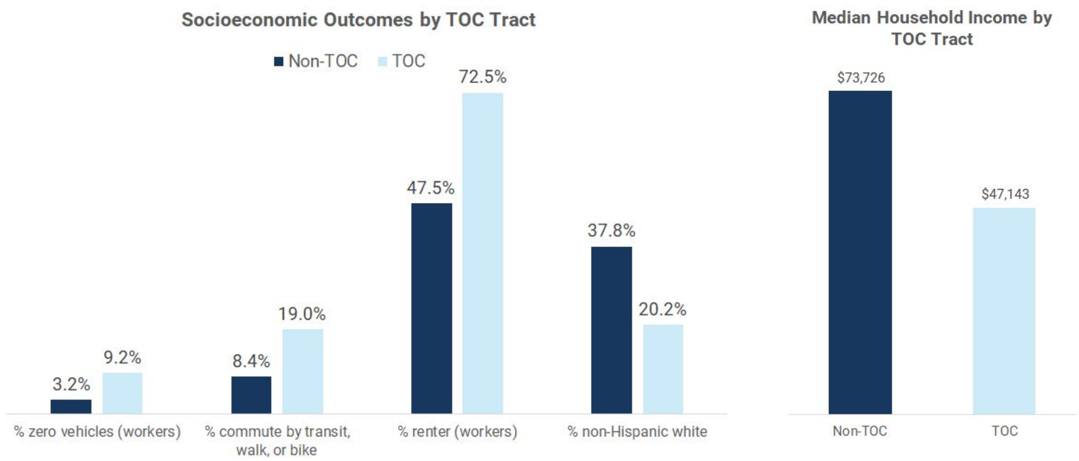
### Figure 6 TOC Tracts Map




*Source: TOC Tiers, census tracts; analysis by ITA Data Science*

**Figure 7** compares socioeconomic outcomes between TOC vs Non-TOC tracts. TOC tracts were associated with more workers from zero-vehicle households, more workers who commute by transit, walking, or biking, more renters, lower proportions of non-Hispanic white populations, and lower median household incomes.

### Figure 7 Socioeconomic Outcomes by TOC vs Non-TOC Tract



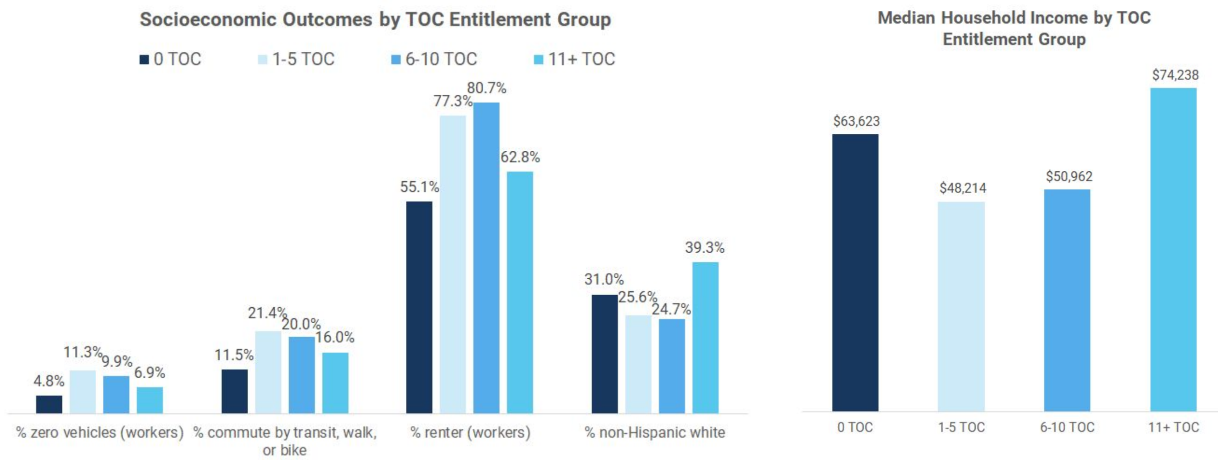
| | % zero vehicles (workers) | % commute by transit, walk, or bike | % renter (workers) | % non-Hispanic white | median household income | # tracts |
|---|---|---|---|---|---|---|
| **Non-TOC** | 3.2% | 8.4% | 47.5% | 37.8% | $73,726 | 639 |
| **TOC** | 9.2% | 19.0% | 72.5% | 20.2% | $47,143 | 510 |

*Source: PCTS, ACS 2018 5-year estimates, TOC Tiers; analysis by ITA Data Science*

Instead of grouping tracts into a dichotomy of being a TOC tract or not, we also use a gradient approach, grouping tracts by the number of TOC entitlements that actually occurred. Tracts can have zero TOC entitlements, 1 - 5, 6 - 10, or 11 or more. This categorizes tracts according to the number of entitlements it *actually* had, not whether the tract was composed of many parcels that were *eligible*.

**Figure 8** compares the socioeconomic outcomes. Except for tracts with 11 or more TOC entitlements, tracts that have between 1 and 10 TOC entitlements are associated with more workers from zero-vehicle households, commute by transit / walking / biking, more workers from renter-occupied households, fewer non-Hispanic whites, and lower median household income. The tracts with 11 or more TOC entitlements (only 3 tracts), look much more similar to tracts with no TOC entitlements.

## Figure 8 Socioeconomic Outcomes by TOC Entitlements Group



| | % zero vehicles (workers) | % commute by transit, walk, or bike | % renter (workers) | % non-Hispanic white | median household income | # tracts |
|---|---|---|---|---|---|---|
| **0 TOC** | 4.8% | 11.5% | 55.1% | 31.0% | $63,623 | 973 |
| **1-5 TOC** | 11.3% | 21.4% | 77.3% | 25.6% | $48,214 | 149 |
| **6-10 TOC** | 9.9% | 20.0% | 80.7% | 24.7% | 50,962 | 24 |
| **11+ TOC** | 6.9% | 16.0% | 62.8% | 39.3% | $74,238 | 3 |

*Source: PCTS, ACS 2018 5-year estimates, TOC Tiers; analysis by ITA Data Science*

We rely on **Section III - Socioeconomic Analysis** and particularly the TOC regressions to confirm what we found in the descriptive findings, and to look for other factors that are associated with where TOC entitlements are located.

# III. Socioeconomic Analysis

Our socioeconomic analysis looks at the neighborhood socioeconomic characteristics that were associated with certain types of entitlements. This broadens the scope to all suffixes in PCTS. We used the same tract-level characteristics used in the transit / geographic analysis, except for percent of workers who commute by transit / biking / walking, since that is more transit-specific.

These characteristics were:

- percent of workers from zero-vehicle households
- percent of population from renter-occupied households
- percent of non-Hispanic white population
- median household income
- total population (to derive population density)

We used a Poisson regression to look at how the total number of entitlements at the tract-level for the entire decade were related to these socioeconomic characteristics. Poisson regressions are often used for count data; in this case, we are looking at counts of entitlements per tract.

The regression focused on the 30 suffixes with the most number of entitlements in the entire City over the decade. Additionally, we did two in-depth regressions for the TOC (Transit Oriented Communities) and CUB (stand-alone alcohol sale) suffixes. These in-depth regressions added zoning as a predictor variable, as zoning may capture the situation on-the-ground that developers face, but otherwise not captured by the socioeconomic characteristics.

## A. Methodology

We started with all entitlements between January 2010 and March 2020. We further filtered out certain cases in consultation with the PMU team (listed alphabetically below). Child cases were dropped, but the child case's prefixes and suffixes were stored with the parent case.

Prefixes removed: [“ENV”]
Suffixes removed:

> ["CA", "CATEX", "CPIO", "CPU", "CRA”,
>
> “EIR”, "FH", "G",
>
> "HD", "HPOZ",
>
> "ICO", "IPRO", "K",
>
> "LCP", "NSO", "RFA",
>
> "S", "SN", "SP",
>
> "ZAI"]

The entitlements were then spatially joined to tracts with their socioeconomic characteristics. Any specific case that touched more than 20 parcels was treated as an outlier and removed; those are likely community plan updates or a one-time entitlement applied to an entire neighborhood. The counts were aggregated across years to get total counts for each entitlement suffix and tract. The 30 entitlement suffixes with the most entitlements were kept for the Poisson regression.

These 30 entitlement suffixes are (in descending order by total count by suffix):

['CWC', 'SPP', 'CUB', 'CEX', 'TOC',

'OVR', 'ZAA', 'ZV', 'CWNC', 'VSO',

'1A', 'DRB', 'CU', 'MSP', 'CDP',

'ZAD', 'MEL', 'SPR', 'DB', 'PMEX',

'PMLA', 'SL', 'MA', 'CPIOC', 'ADU',

'CUW', 'CDO', 'UDU', 'COA', 'HCM']

Regression coefficients describe the relationship between various predictor variables (X's), with our outcome variable (Y). It isolates the relationship between each predictor variable by holding all the other predictors constant. Regressions are useful because the results may show us that certain predictors were not as significant as originally thought and/or give us an estimated "impact" of each predictor on the outcome variable.

For our two in-depth regressions, we first fit a random forest regression model. Random forest regressions are very good at fitting the current data, but a major drawback is its difficulty in interpreting its coefficients. We go back to using a Poisson regression because it's a regression model that is used for count data and the regression coefficients are much more interpretable. We go a further step of interpreting the regression coefficients to get a statement of impact such as: "A 10% increase in [predictor variable A] is associated with an x% increase in the number [this particular type of] entitlements."

## B. General Regression Findings

The Poisson regression is fitted, and by running the same regression over 100 samples, we are able to understand the distribution of the coefficients (see **Appendix C**). **Figure 9** summarizes whether the coefficients were positively or negatively related to the number of entitlements for that suffix.

**Figure 9 Direction of Regression Coefficient for 30 Largest Suffixes**

- Being positively related with median household income means that in areas where income is higher, there are more entitlements of that suffix.
- Being positively related with % renter means that as the renter-occupied population increases, there are more entitlements of that suffix.
- Being positively related with % zero-vehicle workers means that as the number of zero-vehicle workers increases, there are more entitlements of that suffix.
- Being positively related with population density means that as the population density increases, there are more entitlements of that suffix.
- Being positively related with % white non-Hispanic means that as the number of white non-Hispanic people increases, there are more entitlements of that suffix.

**Positively related** **Negatively related**

| **Suffix** descending order by # entitlements | **median household income** | **% renter** | **% zero-vehicle workers** | **population density** | **% white non-Hispanic** |
|---|---|---|---|---|---|
| CWC | | | | Negatively related | |
| SPP | | | | Negatively related | Positively related |
| CUB | | Positively related | Positively related | | Positively related |
| CEX | Positively related | Positively related | | | Positively related |
| TOC | Negatively related | Positively related | | Positively related | |
| OVR | Positively related | | | Negatively related | Positively related |
| ZAA | Positively related | | | Negatively related | Positively related |
| ZV | Positively related | Positively related | Positively related | Negatively related | Positively related |
| CWNC | | | | Negatively related | |
| VSO | Positively related | Positively related | | | Positively related |
| 1A | Positively related | Positively related | Positively related | | Positively related |
| DRB | Positively related | | | | Positively related |
| CU | | Positively related | | Negatively related | |
| MSP | Positively related | Negatively related | | Negatively related | Positively related |
| CDP | Positively related | | | Negatively related | Positively related |
| ZAD | Positively related | Negatively related | | Negatively related | Positively related |

| | median household income | % renter | % zero-vehicle workers | population density | % white non-Hispanic |
|---|---|---|---|---|---|
| MEL | | | | | |
| SPR | | | | | |
| DB | | | | | |
| PMEX | | | | | |
| PMLA | | | | | |
| SL | | | | | |
| MA | | | | | |
| CPIOC | | | | | |
| ADU | | | | | |
| CUW | | | | | |
| CDO | | | | | |
| UDU | | | | | |
| COA | | | | | |
| HCM | | | | | |

*Source: PCTS; ACS 2018 5-year estimates; analysis by ITA Data Science*

No coefficients are reported in **Figure 9**, but only the direction of the relationship between the predictor (socioeconomic variables) and the outcome. For each suffix, the regression was run 100 times, giving us the distribution of the coefficients. The code for this regression is here.

- If the minimum estimate of the coefficient was positive (and zero is not in the interval at all), then the predictor was **positively related** to the number of entitlements. This means that **as the value in the predictor variable increases** (more renter-occupied, higher population density, etc), there was an **associated increase in entitlements**, holding other variables constant.

- If the maximum estimate of the coefficient was negative (and zero is not in the interval at all), then the predictor was **negatively related** to the number of entitlements. This means that **as the value of the predictor variable increases** (more zero-vehicle workers, higher income), there was an **associated decrease in entitlements**, holding other variables constant.

We use our in-depth regressions to do a deep-dive for the TOC and CUB suffixes. We were able to look more closely at which predictor variables were associated with the number of entitlements for that suffix, beyond just the socioeconomic characteristics included here. We also interpret the coefficients in the in-depth regressions.

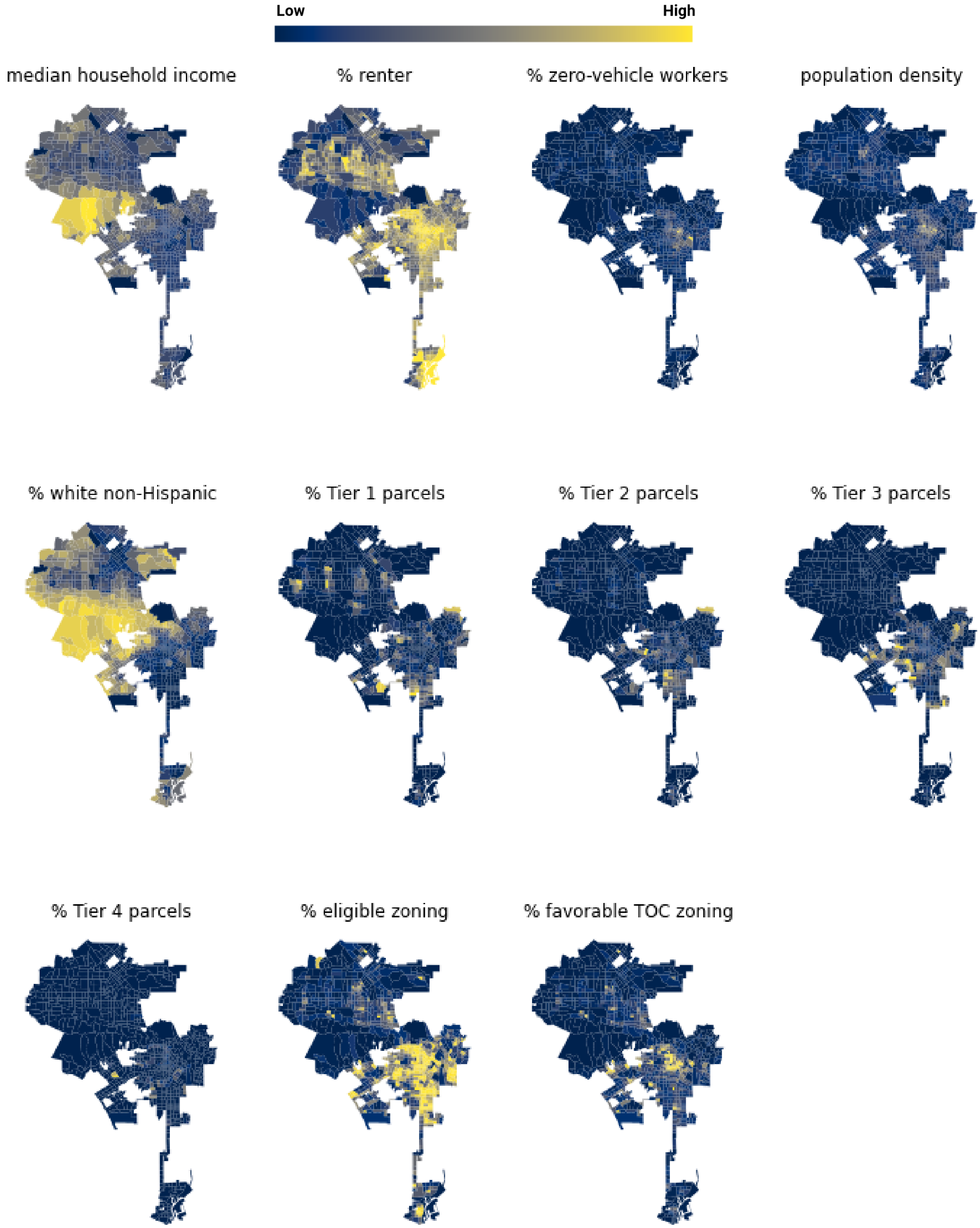
## C. In-depth TOC Regression

In our earlier transit / geographic analysis, we found that Tier 3 had the most TOC entitlements, containing nearly 58% of all TOC entitlements. In terms of proportion of TOC entitlements out of total entitlements within each tier, Tier 3 (21%) and Tier 2 (18%) had the highest (**Figure 3**). We also found that there were three zone classes (R3, R4, C2) were associated with 90% of the TOC entitlements (**Figure 5**).

Our in-depth regression included additional factors that were strongly correlated in our descriptive analysis. Our outcome variable is the number of TOC entitlements per parcel (to adjust for differences in the number of parcels within each census tract, we use this average value). The code for this regression is here.

**Figure 10** is a visual inspection to make sure all predictors look as expected. **Table 6** provides summary statistics of these predictor variables for the TOC vs Non-TOC tracts. Descriptively, TOC tracts appeared to be different from Non-TOC tracts. TOC tracts have higher commute shares by non-car modes, more workers from zero-vehicle households, more renters, lower median household incomes, higher proportions of parcels with eligible and favorable TOC zoning, and more parcels in each TOC tier. The regression analysis would determine whether all or some of these characteristics, and which ones, impact how many TOC entitlements are observed in the tract.

## Figure 10 All TOC Predictor Variables Mapped




*Source: ACS 2018 5-year estimates, TOC Tiers; ZIMAS zoning; analysis by ITA Data Science*

**Table 6 Summary Statistics for Predictors by TOC / Non-TOC Tract**

| | Non-TOC Tract (< 50% parcels in any TOC Tier) | TOC Tract (50%+ parcels in any TOC Tier) |
|---|---|---|
| % commute by transit / biking / walking | 4.2% | 9.5% |
| % zero-vehicle workers | 1.6% | 4.6% |
| % renter | 47.5% | 72.5% |
| % white non-Hispanic | 37.8% | 20.2% |
| median household income | $73,726 | $47,143 |
| % eligible zoning | 10.2% | 56.8% |
| % favorable TOC zoning | 4.1% | 21.6% |
| % Tier 1 parcels | 3.9% | 35.6% |
| % Tier 2 parcels | 0.6% | 19.2% |
| % Tier 3 parcels | 1.8% | 26.7% |
| % Tier 4 parcels | 0.1% | 1.2% |

*Source: ACS 2018 5-year estimates, TOC Tiers; ZIMAS zoning; analysis by ITA Data Science*

To fit the regression, only tracts with some non-zero parcels in any tier (at least one parcel in any of the tiers) were included. The random forest regression did a good job at predicting the number of TOC entitlements per parcel, but the results are hard to interpret. However, with random forest regression, we were able to learn which predictors were more or less important in predicting TOC entitlements per parcel. After looking at the “feature importance” and “permutation importance”, we narrowed down our list of predictors to a smaller set of variables (see **Appendix D** for random forest results, feature importance, and permutation importance results).

**Table 7** lists the full set of predictors included and the smaller list of the most important predictors. The random forest regression was used to narrow down our predictors to the most important, and the smaller set of predictor variables were used in the Poisson regression.

Table 7 TOC List of Predictors

| | Full List | Smaller List |
|---|---|---|
| **ACS socioeconomic characteristics** | | |
| median household income | X | |
| % renter | X | X |
| % zero-vehicle workers | X | |
| % white non-Hispanic | X | |
| population density | X | |
| **TOC-specific characteristics**[26] | | |
| % Tier 1 parcels | X | |
| % Tier 2 parcels | X | X |
| % Tier 3 parcels | X | X |
| % Tier 4 parcels | X | |
| **Zoning characteristics** | | |
| % parcels with eligible zoning[27] (16 zone classes) | X | |
| % parcels with favorable zoning (R3, R4, C2) | X | X |

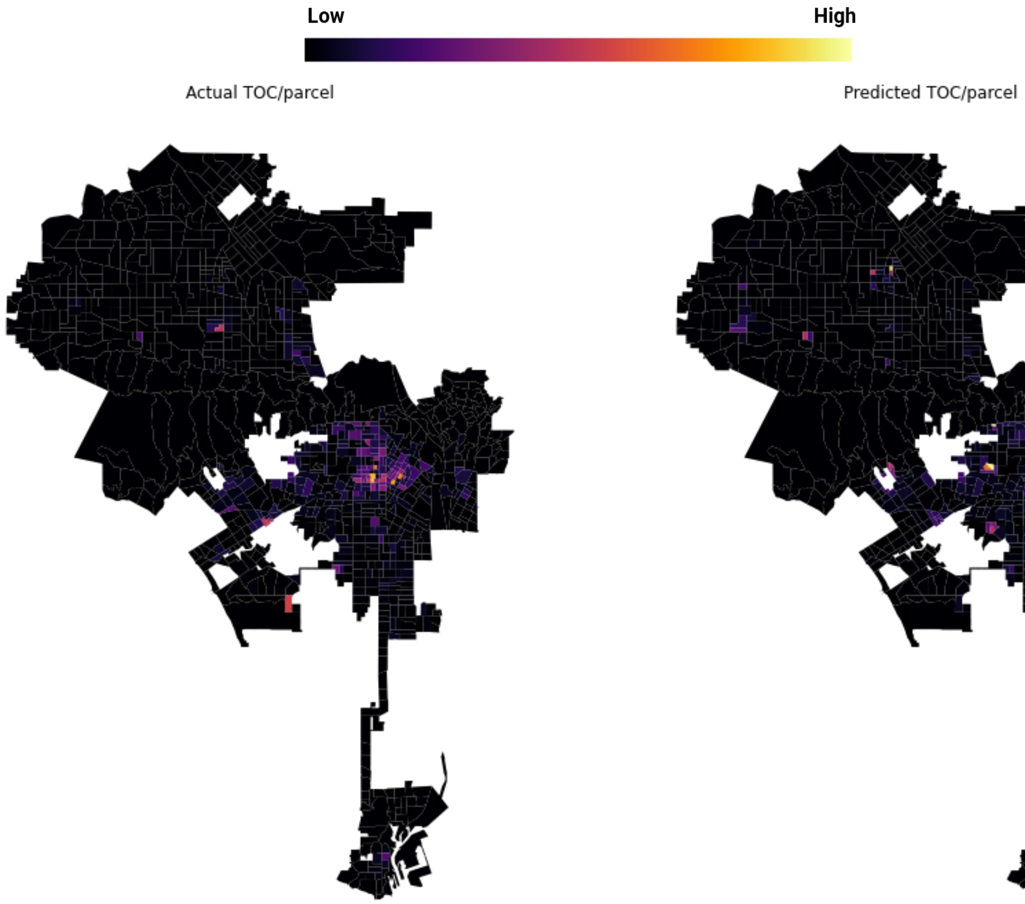
*Source: ACS 2018 5-year estimates, TOC Tiers; ZIMAS zoning; analysis by ITA Data Science*

We estimated a Poisson regression with the smaller list of predictors. **Figure 11** shows that the Poisson regression with the four predictors also did a fairly good job at identifying where TOC entitlements would be, similar to the random forest model.

[26] These are not perfectly collinear because the excluded group is the % parcels in no tier.
[27] This was the list of eligible zones (16 zone classes).

```
[
    'R2', 'R3', 'RAS3', 'R4', 'RAS4', 'R5',
    'RD1.5', 'RD2', 'RD3', 'RD4', 'RD5', 'RD6',
    'C1', 'C2', 'C4', 'C5'
]
```

**Figure 11 TOC Poisson Regression Results**

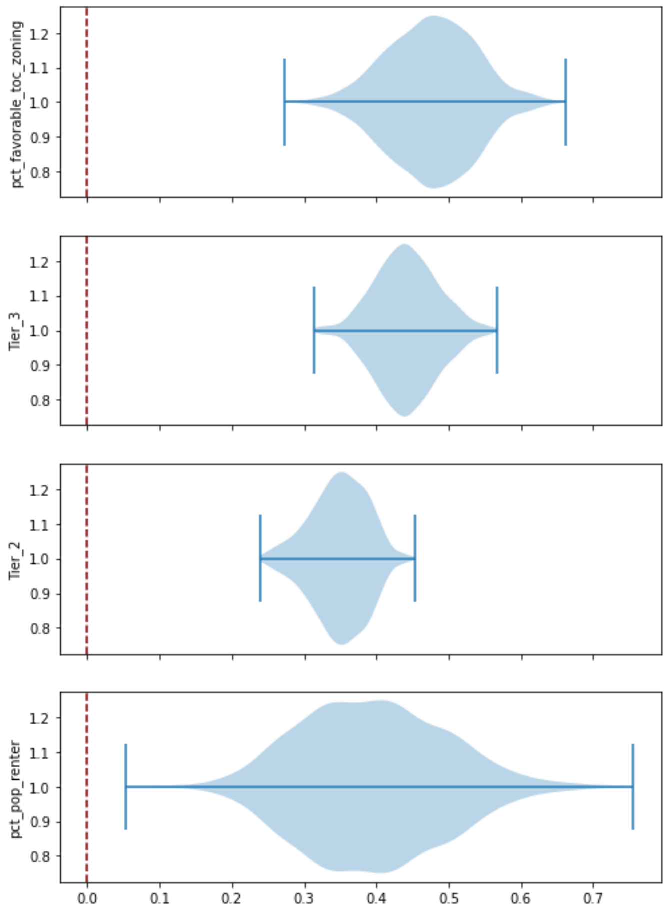



*Source: ACS 2018 5-year estimates, TOC Tiers; ZIMAS zoning; analysis by ITA Data Science*

The coefficients coming from a Poisson regression are much more interpretable. The Poisson regression is estimated over 100 samples to produce confidence intervals for each of the estimated coefficients. Coefficients that fall to the right of the dotted red line are positively related to TOC entitlements; coefficients that fall to the left of the dotted red line are negatively related. If the 95% confidence intervals overlap with zero (dotted red line), then the effect is not statistically different from zero. The less overlap between the confidence interval (blue curve) to the dotted red line, the more significant the effect the predictor has on the outcome. All the coefficients have a positive and significant relationship with TOC entitlements per parcel (**Figure 12**). To interpret the coefficients, we use this formula[28]: Given a unit change in $x_j$, the fitted $\hat{y}$ changes by $\hat{y}(e^{bj} - 1)$.

[28] How to interpret parameter estimates in Poisson GLM results: https://stats.stackexchange.com/questions/128926/how-to-interpret-parameter-estimates-in-poisson-glm-results

**Figure 12 TOC Poisson Regression Coefficients**




*Source: ACS 2018 5-year estimates, TOC Tiers; ZIMAS zoning; analysis by ITA Data Science*

Even though TOC tracts looked fairly different than Non-TOC tracts on a list of predictor variables, only four of those characteristics were important in impacting how many TOC entitlements occurred in that tract. We interpret the coefficients:

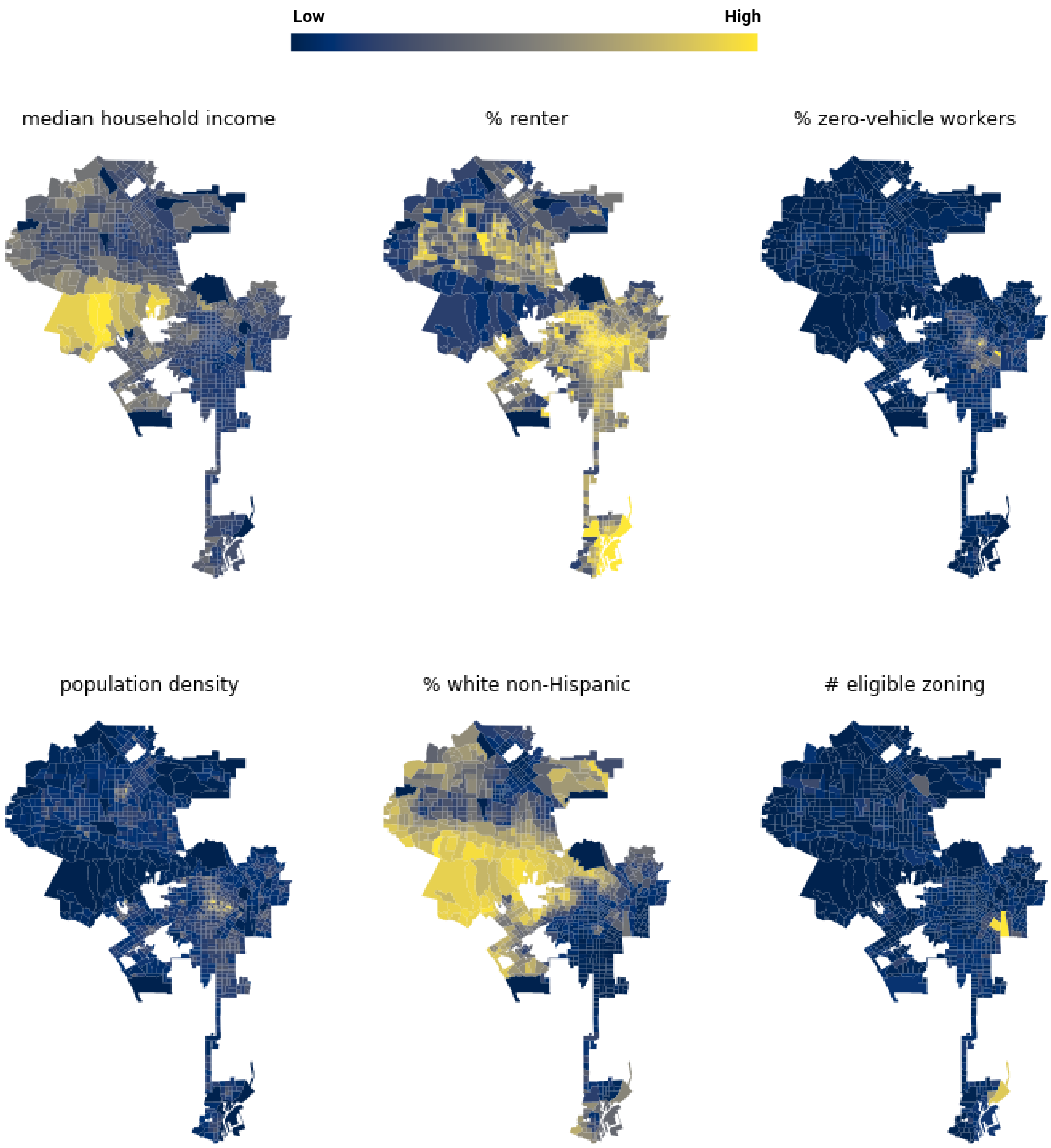
- For every 10% increase in favorable TOC (R3, R4, C2) zoning, there is a 23% increase in TOC entitlements, holding other variables constant.
- For every increase of 100 Tier 3 parcels, there is a 33% increase in TOC entitlements, holding other variables constant.
- For every increase of 100 Tier 2 parcels, there is a 36% increase in TOC entitlements, holding other variables constant.
- For every 10% increase in population of renters, there is a 8% increase in TOC entitlements, holding other variables constant.

## D. In-depth CUB Regression

Our second in-depth regression is on where stand-alone alcohol sale permits (CUB) are. Only commercial and manufacturing parcels were eligible for this type of entitlement. The code for this regression is here.

**Figure 13** is a visual inspection to make sure all predictors look as expected.

**Figure 13 All CUB Predictor Variables Mapped**




*Source: ACS 2018 5-year estimates; ZIMAS zoning; analysis by ITA Data Science*

The random forest regression does a good job at predicting the number of CUB entitlements per parcel, but the results are hard to interpret. However, with random forest regression, we were able to learn which predictors were more or less important in predicting CUB entitlements per parcel. After looking at the “feature importance” and “permutation importance”, we narrowed down our list of predictors to a smaller set of variables (see **Appendix E** for random forest results, feature importance, and permutation importance results).

In sifting through feature importance and permutation importance, we found that certain predictor variables were highly correlated with one another. The variables % renter and % zero-vehicle workers basically moved together; whichever variable was included in the regression would soak up all the “effect”. Due to the highly collinear nature of % renter and % zero-vehicle workers, we ran the Poisson regression twice, including just one of them each time, along with the other predictor variables, to see which predictor did a better job at fitting the actual CUB entitlements per tract. We chose % renter in the end. Population density, when included, gave us a coefficient with a very small effect size; it was excluded as a result.

**Table 8** lists the full set of predictors included and the smaller list of the most important predictors. The random forest regression was used to narrow down our predictors to the most important, and the smaller set of predictor variables were used in the Poisson regression.

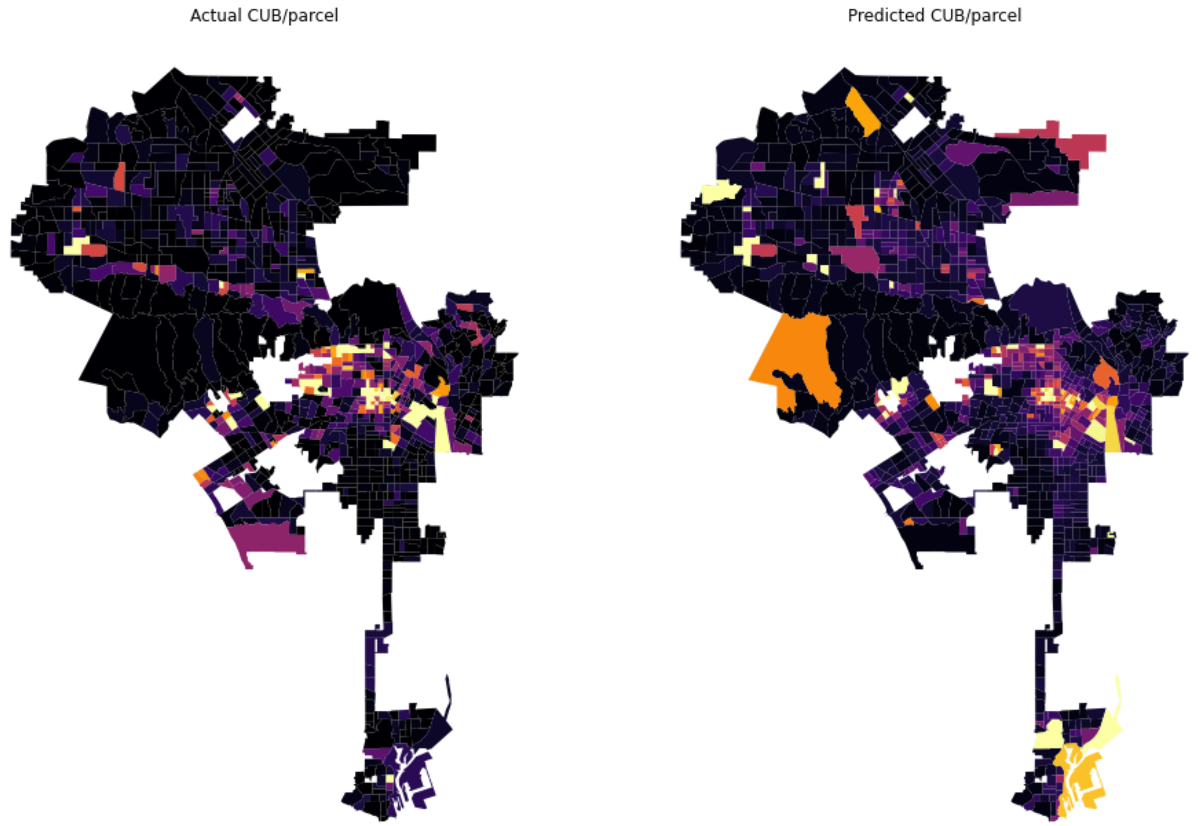
**Table 8 CUB List of Predictors**

| | Full List | Smaller List |
|---|---|---|
| **ACS socioeconomic characteristics** | | |
| median household income | X | |
| % renter | X | X |
| % zero-vehicle workers | X | |
| % white non-Hispanic | X | X |
| population density | X | |
| **Zoning characteristics** | | |
| # parcels with eligible zoning[29] (12 zone classes) | X | X |

*Source: ACS 2018 5-year estimates; ZIMAS zoning; analysis by ITA Data Science*

[29] This was the list of eligible zones (12 zone classes).
[
'CR', 'C1', 'C1.5', 'C2', ‘C4’, 'C5’ ‘CM’,
‘MR1', 'MR2', 'M1', 'M2', 'M3',
]

**Figure 14 CUB Poisson Regression Results**

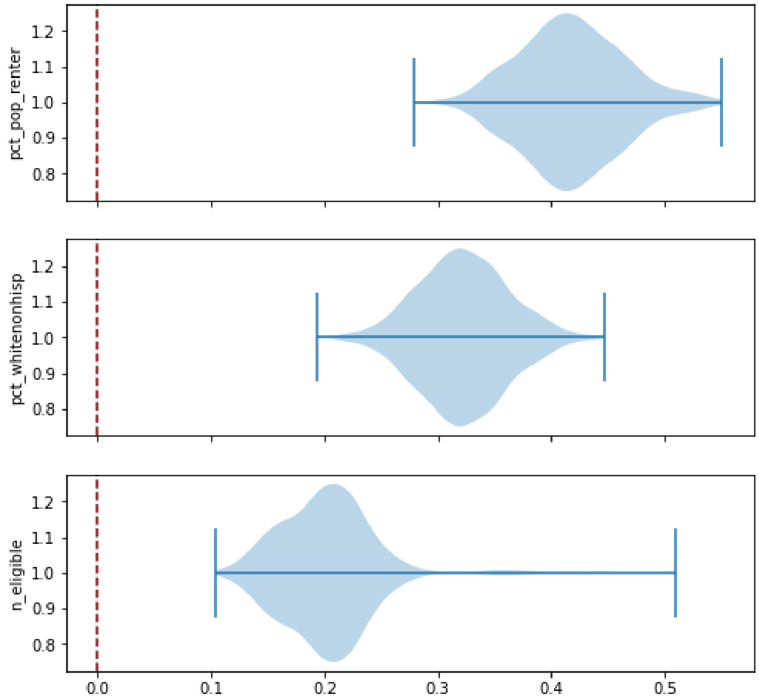



*Source: ACS 2018 5-year estimates; ZIMAS zoning; analysis by ITA Data Science*

The coefficients coming from a Poisson regression are much more interpretable. The Poisson regression is estimated over 100 samples to produce confidence intervals for each of the estimated coefficients. Coefficients that fall to the right of the dotted red line are positively related to CUB entitlements; coefficients that fall to the left of the dotted red line are negatively related. If the 95% confidence intervals overlap with zero (dotted red line), then the effect is not statistically different from zero. The less overlap between the confidence interval (blue curve) to the dotted red line, the more significant the effect the predictor has on the outcome. All the coefficients have a positive and significant relationship with CUB entitlements per parcel (**Figure 15**). To interpret the coefficients, we use this formula[30]: Given a unit change in $x_j$, the fitted $\widehat{y}$ changes by $\widehat{y}(e^{bj} - 1)$.

[30] How to interpret parameter estimates in Poisson GLM results: https://stats.stackexchange.com/questions/128926/how-to-interpret-parameter-estimates-in-poisson-glm-results

**Figure 15 CUB Poisson Regression Coefficients**




*Source: ACS 2018 5-year estimates; ZIMAS zoning; analysis by ITA Data Science*

In interpreting the coefficients, again, we stress that the population of renters and zero-vehicle workers are collinear, but that we included the renter population as one of our predictors. We interpret the coefficients:

- For every 10% increase in population of renters, there is a 17.2% increase in CUB entitlements, holding other variables constant.
- For every 10% increase in the population of white non-Hispanic residents, there is a 9.5% increase in CUB entitlements, holding other variables constant.
- For every 100 additional eligible parcels (commercial and manufacturing), there is a 20.5% increase in CUB entitlements, holding other variables constant.

# IV. Data Pipeline

Reproducibility is one of the key tenets of completing data analytics projects on behalf of City departments. To this end, we've documented all our code in this GitHub repository, saved all the datasets in a versioned S3 bucket, created the `laplan` Python package, and automated the data cleaning process in Civis to produce ongoing processed data.

One major hurdle in data science is reproducibility. Two people might get slightly different answers from answering the same question with the same data source. Two things are needed to improve reproducibility: (1) canonical data source, in this case, the PCTS data, and (2) standardized data cleaning pipeline.

This section describes not only specific decisions made within the data cleaning pipeline, but covers some of the tools the team leveraged. To make the entirety of our analysis fully reproducible from beginning to end, all the code was stored in GitHub (version control), the data is stored in an S3 bucket, and we also created the `laplan` package to generalize the data cleaning process. We also use Civis Analytics to create some dashboards as intermediary work products.

## A. Data Cleaning Issues Resolved for PCTS

Getting the canonical data source was simple, as DCP owns and maintains the PCTS data, and gave us access to the data. The standardized data cleaning pipeline was less simple to construct. Some of the issues we ran into and resolved were:

- **How should we count parent and child cases? Should we drop child cases and count parent cases only?**

  We built a flexible and open-ended option. Within the `laplan` package, the function has an optional argument to keep or drop child cases. Depending on the question answered, the user should choose whether only parent cases are needed or parent and child cases are needed.

- **How should prefixes and suffixes be stored for each case?**

  We store prefixes and suffixes associated with each case in dummy variables. The dummy variable is True if that prefix or suffix appears in the case string, and False otherwise.

- **How do we reconcile the prefixes and suffixes that appear in parent and child cases? What if they are slightly different?**

  We built a flexible and open-ended option. Within the `laplan` package, the function has an optional argument to keep the prefixes / suffixes in the child cases and "roll up" the history into the parent case's history or not.

For example, the parent case might list suffixes A, B, and C, but child case 1 lists suffixes A, B, and D. Instead of losing suffix D when we drop the child case, we store suffix D as part of the parent case. As mentioned above, the prefixes / suffixes are stored in dummy variables. Therefore, suffixes A, B and C are already flagged through the parent case, but now suffix D gets flagged through the child case, and is stored in the row as the parent case. Suffixes A and B appear in both the parent and child cases, but don't get double-counted, as our dummy variable is True/False.

Parent / child case starts out as:

| CASE_NUMBER | parent_or_child | DIR | TOC | 1A |
|---|---|---|---|---|
| DIR-2018-4336-TOC | parent | True | True | False |
| DIR-2018-4336-TOC-1A | child | True | True | True |

Parent / child case ends up as (if we only want to keep parent case, and keep child's history, look at how 1A changes without changing anything else):

| CASE_NUMBER | parent_or_child | DIR | TOC | 1A |
|---|---|---|---|---|
| DIR-2018-4336-TOC | parent | True | True | **True** |

- **Can we subset by start or end date?**

  Yes, start and end dates are optional arguments, and are based on the CASE_FILE_DATE. If no start or end dates are given, it defaults to Jan 1, 2010 as the start date, and today's date as the end date.

- **What if we want to remove prefixes or suffixes?**

  A list of prefixes and/or suffixes to include is optional. If no lists are given, then all the valid prefixes and suffixes are returned.

  If we define a smaller list of prefixes as the "allow list", then only cases with prefixes on the "allow list" are returned. Cases only have one prefix; by excluding some prefixes, we automatically exclude those cases, without looking at what suffixes those cases contained.

  If we define a smaller list of suffixes as the "allow list", then all cases where those suffixes appear are returned, even if some of those cases also include suffixes that are **not** on the "allow list".

  Excluding by prefixes is more stringent than excluding by suffixes.

  For example, if only one suffix, ["TOC"], is set as the "allow list" for suffixes, a case that is *DIR-2018-4336-TOC-1A* would also be included. "TOC" appears in the case string, which means the case will be returned, even if "1A" is not part of the "allow list".

The exact functions for subsetting the PCTS data is here.

In terms of joining the various tables from PCTS to come up with entitlement data, we used the same SQL query that DCP uses for its PCTS internal report. The code for that is available here. After that, we use our functions in `laplan` to further subset the PCTS data by various start / end dates, a list of prefixes and/or suffixes to include or exclude, and deciding whether child cases should be kept or dropped.

After coming up with the sample of entitlements we want to analyze, we aggregated entitlements up to the parcel or tract-level. Refer back to **Figure 1** for how we would assemble the various pieces of data needed to answer a policy / planning research question.

For the transit / geographic analysis, we created a master dataset that looked at entitlements from October 2017 - March 2020, attached the relevant transit information along with entitlements aggregated to the parcel-level, and kept only parcels that fell within TOC tiers.

For the socioeconomic analysis, we created a master dataset that looked at entitlements from January 2010 - March 2020, removed a smaller list of suffixes, aggregated entitlements to the tract-level, and removed outlier cases that touched 20 or more parcels.

Given that different analyses required different master datasets, we were careful in creating our dataset "puzzle pieces". Each component had to be modular and flexible enough to accommodate various types of analyses easily assembled to build a master dataset.

The challenges and complexities we faced in assembling the "puzzle pieces" for our two analyses led us to propose a desired solution for the data cleaning pipeline to increase reproducibility. Rather than creating datasets that were designed for a specific analysis in mind, we wanted the PCTS data to undergo a similar pipeline process so future users have several options in place. Instead of making the decisions we made over and over, which could have easily been different decisions (not wrong, just different), it was important we laid out clearly what decisions were made in consultation with the PMU team. As a result, users would connect to the same canonical PCTS database, put that raw data through a standardized basic processing pipeline, and receive a *somewhat* processed dataset to use for their analysis.

## B. Python Package: laplan

The `laplan` package was borne out of the desire to generalize the data cleaning and processing pipeline. This is a Python package that can be installed on its own for use elsewhere. It is installed in our Python environment (all documented in our Dockerfile, requirements, and conda-requirements). The functions included are all fairly general and are designed to standardize the initial stages of cleaning and processing. Our analyses all do further processing.

We created our master datasets in an ad-hoc fashion, but after creating the two we used for our analyses, it was obvious that the steps for cleaning and processing the data could be generalized. This was especially needed if we wanted to swap out our static PCTS dataset that ended in March 2020 for a live connection to the PCTS dataset. It would also clearly list the pieces that needed maintenance, such as the ACS data, which is updated once a year.

The `laplan` package is comprised of three cleaning scripts:

1. `pcts`: functions to subset PCTS data by prefixes, suffixes, start / end dates, parents and child cases are all included here.

2. `zoning`: functions to parse the zoning string given in ZIMAS, so we could more easily access characteristics such as zone class, height district, etc.

3. `census`: functions to download a select group of ACS tables using the Census API, and putting those tables through several steps to clean, standardize, reshape, and aggregate so the table easily merges at the census tract level.

The documentation associated with this package (with examples) is included in full in **Appendix A**. However, the most updated documentation will always be in our GitHub repository.

## C. GitHub

Our `laplan` package is stored in this [project's GitHub repository]. GitHub is a platform for source code management and version control. It facilitates working collaboratively on a project, providing a version of the code for all team members to make and track changes. As only code, not any data, is stored in GitHub, the repository for this project is open and public.

In conjunction with a Docker image, which provides a standardized Python environment for users of this project, all the scripts and analyses of this project are completely reproducible.

## D. Civis Analytics

Civis Analytics is a data science platform ITA purchased while this project was ongoing. Civis is a platform designed for solving various environment issues users might face while using Python or R. Many issues arise because environments are hard to standardize, packages may get updated or outdated, or updated packages may break functionalities with other installed packages. The ITA Data Science team uses Docker to solve the environment issue. Civis can both take Docker images users upload, but it also supplies its own Docker images for the user.

In addition, Civis has the ability to deploy dashboards. Jupyter Notebooks can be turned into interactive dashboards using Voila. We created a dashboard to show how many entitlements were filed for various suffixes and years. That dashboard, while not a deliverable in itself, provided us with a way to communicate with PMU on whether the PCTS cleaning scripts were getting the desired counts. It was a means to fine-tune the cleaning script without talking about the code itself, but allowed us to explain what was happening in the background and show how the results differed by using different cleaning methods. This underscores again the importance of a standardized data cleaning pipeline, as steps taken in different orders can lead to different results in the end. Civis makes it fairly easy to deploy, update, start or stop sharing the dashboards.

An area that the ITA Data Science team has not explored much in Civis is connecting the live PCTS database and scheduling a daily data update for processed PCTS dataset stored in a new database or exporting it as a Google Sheet or another format. This would be an area of future phase of work that requires DCP's system support to come up with how best to implement the data cleaning pipeline that is done in Python currently  in conjunction with DCP's PowerBI and ESRI use.

Other areas DCP might want to use Civis for is its Jupyter Notebooks or R Notebooks with Civis's own environment, or to input the same environment that the ITA Data Science team uses.

# V. Conclusion

This section summarizes the recommendations that arose out of our transit / geographic analysis and socioeconomic analysis, lays out areas of future work, and applications of this analysis.

## A. TOC Main Findings

Our in-depth TOC regression combined with our transit / geographic analysis shows the types of locations that have seen TOC entitlement activity. It is not simply enough for areas to be designated as Tier 1, 2, 3, or 4. The tier designation must be combined with zoning characteristics. The zone classes that were favorable toward TOC entitlements are R3, R4, and C2 zoning, with 90% of the TOC entitlements occurring within those zones (**Figure 5**). The random forest regression found that favorable zoning was the most important predictor for TOC entitlements (**Appendix D**). Our Poisson regression found that a 10% increase in favorable zoning in a tract was associated with a 23% increase in TOC entitlements, holding all other variables constant.

Tier designations were also important, particularly Tier 2 and Tier 3 parcels (**Figure 3**). Tier 3 held the most number of TOC entitlements *across* the four tiers, whileTier 2 had the highest proportion of TOC entitlements *within* the tier. Particularly, two of the rules within Tier 3 generated the most TOC entitlements (**Figure 4**), which were being within < 750 ft of the intersection of Regular and Rapid buses and being within 0.5 mile of a Metro Rail station.

Socioeconomic characteristics did play a role, but a small one. Descriptively, TOC tracts and Non-TOC tracts had different socioeconomic characteristics, with TOC tracts having more workers from zero-vehicle households, more workers who commute by transit, walking, or biking, more renters, have lower proportions of non-Hispanic white populations, and lower median household incomes (**Figure 7**). However, in its impact on TOC entitlements, only the percent of renters was included as a predictor. We found that a 10% increase in the population of renters was associated with an 8% increase in TOC entitlements, holding all other variables constant.

Overall, zoning seems to play the biggest role in encouraging TOC entitlements. There is a blanket policy encouraging transit-oriented communities and affordable housing, but its impacts are far from equal. Most of the TOC entitlements that have taken place since October 2017 have occurred along the B - Red / D - Purple Lines. While heavy rail is the most suited for TOC development because of its exclusive right-of-way underground, it should not remain the only transit type that garners TOC development.

Many of the existing light rail lines connect to downtown, a major job center. Even the E - Expo Line has been open for several years. The fact that little affordable housing (or few TOC entitlements) have occurred along all those light rail lines indicates a housing / land use planning and policy gap. Light rail, a significant capital investment, should be paired with the right land use planning to encourage the type of development that is occurring along the B - Red

/ D - Purple Lines. Specifically, planners have a role to bridge the gap in what makes R3, R4, C2 zoning attractive for developers and apply those conditions elsewhere along the existing light rail lines. The R3 zoning is particularly attractive, not just to TOC entitlements, but even to Non-TOC entitlements.

The Metro rail stations downtown don't see a lot of TOC entitlements because there are other overlays and specific plans that are much more attractive to developers than the TOC program. Besides downtown, are there already eligible areas near transit that can be made more attractive to developers with additional policy programs? A future and long-term step would be to convert ineligible zoning near transit to be eligible. Upzoning itself is a ceiling, not a floor, and making zoning changes to allow eligibility near transit can allow for a gradual conversion toward more housing and more affordable housing.

Ultimately, the city's planning land use and planning process must work in conjunction with the significant capital outlay and investment of new light rail and bus rapid transit lines. The dual goals of increased transit ridership and increased housing supply can be reached when the land use planning supports it. More transit riders mean more residents who don't use a car for all their trips;  the only residents that can afford to have fewer cars than the average are ones who live near transit. This analysis has shown some areas need further planning and policy to encourage this positive feedback loop.

## B. Socioeconomic Findings

By combining the past decade's worth of PCTS entitlements and the ACS socioeconomic data at the tract-level, we used a Poisson regression to understand how socioeconomic characteristics were typically correlated with entitlements (**Figure 9 and Appendix C**). However, through our deep-dive regressions into the TOC and CUB entitlements specifically, we found that there were other predictor variables that were also highly correlated with entitlements. There are many differences across tracts, but zoning explained a large portion of those differences, beyond what socioeconomic characteristics explained. For both TOC and CUB suffixes, zoning was the most important predictor.

- For every 10% increase in favorable TOC (R3, R4, C2) zoning, there is a 23% increase in TOC entitlements, holding other variables constant.
- For every 100 additional eligible parcels (commercial and manufacturing), there is a 20.5% increase in CUB entitlements, holding other variables constant.

Socioeconomic characteristics are still important factors in explaining varying levels of entitlements across tracts. However, we are cautious to only interpret their impact for TOC and CUB suffixes, because those regressions included zoning as a predictor.  Our findings indicate besides zoning differences, there are equity implications too.

For TOC entitlements, higher renter populations are associated with slightly more TOC entitlements. With TOC entitlements, their concentrated activity along the B - Red / D - Purple Lines and the lack of activity along most of the other rail lines, could explain this pattern.

- For every 10% increase in population of renters, there is a 8% increase in TOC entitlements, holding other variables constant.

For CUB entitlements, an area with more renters (or more zero-vehicle workers) is associated with more alcohol permits. An area with more white non-Hispanic residents is also associated with more alcohol permits.

- For every 10% increase in population of renters, there is a 17.2% increase in CUB entitlements, holding other variables constant. Renters and zero-vehicle workers were highly correlated with each other. Switching out renters for zero-vehicle workers results in a similar effect.
- For every 10% increase in the population of white non-Hispanic residents, there is a 9.5% increase in CUB entitlements, holding other variables constant.

## C. Future Areas of Work

### Implement Data Cleaning Pipeline

One area of future work would depend on DCP's own implementation of the data cleaning pipeline. It is currently a Python package, but would likely have to be adapted to fit into DCP's existing pipeline of work. Whether DCP wants to convert it to additional SQL scripts or keep it as Python is up to them, but our Python functions serve as a model for what optional arguments are needed to accommodate various analyses.

This could be a short-term project in and of itself, as it requires connecting to the live PCTS database, putting it through a processing pipeline, figuring out where the processed data would be stored and accessed by end users, updating it daily, and connecting it to PowerBI and ESRI for DCP end users.

Some initial challenges would be to figure out where and what language the processing is done, whether Civis would be used for some part of this process, and how PowerBI and ESRI users can access the cleaned data scaled across multiple users with minimal setup work.

### Accommodate End User Needs

PowerBI should primarily be leveraged for its visualization capability, instead of doing the processing in the DAX language. ESRI remains a tool for geospatial work and visualization, which means that data cleaning and processing should be offloaded elsewhere. However, a canonical version of the processed data (with tract characteristics, etc) should be produced as a shapefile or geojson for ESRI users. As much as possible, these joins can be done outside of ESRI for cleaner processed data.

Also, one area of maintenance is that the ACS data needs an annual update. We have cleaned the data from 2010-2018 for the ACS tables we downloaded, even though we only used the 2018 ACS data. Soon, the 2019 5-year estimates will be available and need to be run through the cleaning scripts to the latest available data.

## D. Application of this work

- DCP person-hours allocation across various planning areas, etc
- Get DCP to help flush this section out

# References

GitHub repository: https://github.com/CityOfLosAngeles/planning-entitlements

Various documentation in GitHub repository:

- Repository Overview
- Data Workflow
- Scripts and Notebooks
- laplan package
- Data catalogs: open data portals, catalog, Census Bureau
- Other reference files

City of Los Angeles, Department of City Planning, PCTS Prefix and Suffix report: https://planning.lacity.org/resources/prefix-suffix-report

City of Los Angeles, Department of City Planning, Transit Oriented Communities Guidelines: https://planning.lacity.org/ordinances/docs/toc/TOCGuidelines.pdf

# Appendix A `laplan` documentation

The Python package `laplan` is included in our GitHub repository here, along with the README. Always refer to the GitHub repository for the most updated version.

A copy of the README is shown below.

-----

The `laplan` package is created for the Los Angeles Department of City Planning. There are 3 sub-modules, each of which can be used independently.

The sub-modules that allow users to clean up zoning data from ZIMAS, entitlement data from PCTS, and Census data from the American Community Survey.

1. Getting Started
2. Zoning
3. PCTS
4. Census
    - Cleaning ACS Data
    - Three Types of ACS Tables
    - General Functions
    - Income Functions

## Getting Started

This package is installed in our Docker image.

To install it into another GitHub repo:

```
pip install
"git+https://github.com/CityOfLosAngeles/planning-entitlements#subdirectory=laplan"

# Ways to use within notebook/script
import laplan
import laplan.census
from laplan import census
```

# Zoning

The sub-module is `zoning.py`. Zoning data comes from ZIMAS, and is publicly available on the GeoHub. Planning's Guide to Zoning String shows that the zoning string is made up of component parts.

The zoning string contains information about prefix on (Q)ualified or (T)entative zone classifications, zone class, the height district, (D)evelopment limits, and specific plans and overlays applicable.

The `ZoningInfo` dataclass takes a zoning string and returns any or all of the components as a new dataframe.

Ex 1: Return all the components

```
import laplan

parsed_col_names = ['Q', 'T',
                    'zone_class', 'specific_plan',
                    'height_district', 'D', 'overlay']

# ZONE_CMPLT is the column to be parsed.
def parse_zoning(row):
    try:
        z = laplan.zoning.ZoningInfo(row.ZONE_CMPLT)
        return pd.Series([z.Q, z.T,
                          z.zone_class, z.specific_plan,
                          z.height_district, z.D, z.overlay],
                          index = parsed_col_names)
    # If it can't be parsed, return either a failed or blank string
    except ValueError:
        return pd.Series(['failed', 'failed',
                          'failed', 'failed',
                          'failed', 'failed', ''],
                          index = parsed_col_names)

parsed = df.apply(parse_zoning, axis = 1)
df = pd.concat([df, parsed], axis = 1)
```

| ZONE_CMPLT | Q | T | zone_class | height_district | D | overlay |
|---|---|---|---|---|---|---|
| C2-1-SP | False | False | C2 | 1 | False | [SP] |
| [Q]C1.5-1VLD-RIO | True | False | C1.5 | 1 | True | [RIO] |

Ex 2: Return just one of the components

```
parsed_col_names = ['zone_class']

def parse_zoning(row):
    try:
        z = laplan.zoning.ZoningInfo(row.ZONE_CMPLT)
        return pd.Series([z.zone_class],
                         index = parsed_col_names)
    except ValueError:
        return pd.Series(['failed'],
                         index = parsed_col_names)


parsed = df.apply(parse_zoning, axis = 1)
```

# PCTS

The sub-module is `pcts.py`. PCTS case strings contain prefixes and suffixes. Planning's PCTS Prefix & Suffix Report lists the valid values.

The `PCTSCaseNumber` dataclass takes a string and returns any or all of the components as a new dataframe (note that `year` and `case` are available columns in PCTS, and parsing these may not be necessary). This dataclass is used infrequently.

The function `subset_pcts` can be used once a PCTS connection is made. It standardizes the initial steps in the data cleaning pipeline so that the PCTS data is extracted and parent/child cases are combined in a standardized way before analysis. The function has optional args. `subset_pcts` and `drop_child_cases` should be used in conjunction with one another. The default is that the full dataset is returned.

- pcts: pandas.DataFrame of PCTS data.
- start_date: defaults to "1/1/2010".
- end_date: defaults to present day.
- prefix_list: a list of prefixes of interest, defaults to all prefixes.
- suffix_list: a list of suffixes of interest, defaults to all suffixes.
- get_dummies: bool, defaults to False. True returns columns for all the prefixes/suffixes of interest.
- verbose: bool, defaults to False. True returns some comments for prefixes/suffixes that have no cases.

Ex: Return PCTS entitlement cases between Oct 2017-Dec 2019 for the ADM and DIR prefixes and TOC suffixes.

```
import laplan

prefix_list = ['ADM', 'DIR']
suffix_list = ['TOC']

df = laplan.pcts.subset_pcts(
    pcts,
    start_date = "10/1/17",
    end_date = "12/31/19",
    prefix_list=prefix_list,
    suffix_list=suffix_list,
    get_dummies=True,
    verbose=True,
)
```

| CASE_NBR | CASE_FILE_RCV_DT | ADM | DIR | TOC |
|---|---|---|---|---|
| DIR-2017-81-TOC-SPR | 2018-10-19 | False | True | True |
| ADM-2017-4594-TOC | 2017-11-08 | True | False | True |

The function `drop_child_cases` returns a dataframe of only parent cases.

- df: pandas.DataFrame returned from `subset_pcts`.
- keep_child_entitlements: bool, defaults to True. True means that the parent case should also hold all of the prefixes and suffixes from any child cases. `get_dummies` must be True in `subset_pcts`. False means all the prefix/suffix dummies of the parent case show up, but child cases are dropped, and the prefixes/suffixes of the child cases are not stored. If a child case holds a different suffix not found in the parent case, `keep_child_entitlements = True` would store this information.

```
df2 = laplan.pcts.drop_child_cases(
    df,
    keep_child_entitlements=True
)
```

## Census

The sub-module is `census.py`.

## Cleaning ACS Data

The American Community Survey (ACS) data for various years, topics, and geographies all follow a similar pattern. Browse the Census Data Catalog or Census API to get the tables needed.

The scripts to download clean Census data are provided below. These scripts can be adapted to include other Census tables; our project dealt with a limited subset of ACS tables for census tracts.

1. Download Census data
2. Clean Census data, part 1
3. Clean Census data, part2
4. Subset Census

The resulting table from these scripts has this form. At minimum, the table MUST have columns `['GEOID', 'year', 'table', 'main_var', 'second_var', 'num']`, in order to use the functions in `census.py`. These necessary columns have stars next to the column name in the table below.

| **Column** | **Description** |
|---|---|
| GEOID * | `str` preferable, but `numeric` works, geographic identifier for county, tract, block group, etc. |
| variable | `str`, the original Census variable name, such as `B01001_001` or `S0801_C01_001`. This is tagged as more human-readable columns, `main_var` and `second_var`. |
| year * | `numeric`, year associated with the table |
| table * | `str`, a human-readable name given to the table in `C2_clean_census.py`. Ex: For `S0801_C01_001`, the table is `S0801`, and is `commute`. |
| main_var * | `str`, a human-readable name that captures what the variable is mainly about. Ex: For `S0801_C01_001`, the `C01` portion what tags `main_var` as `workers`. `C02` would be `male`, `C03` would be `female`, etc. |
| last2 | `str`. The last 2 digits of `variable`. |

| | |
|---|---|
| second_var * | `str`, a human-readable name that captures what the last two digits from the variable. Ex: For `S0801_C01_001`, the last 2 digits is `01` and designates `total`. |
| new_var * | `str`, combines `main_var` and `second_var`. Ex: For `S0801_C01_001`, this value is `workers_total`. |
| pct | `numeric`, holds percent values, ranging from 0-1. |
| num * | `numeric`, holds count values. The method of standardizing across tables is to have one column holding counts and one column holding percents, and filling in all the values for all tables. |

## Three Types of ACS Tables

ACS tables always provide a summary statistic with a `total`, the denominator, representing the universe from which this summary statistic is derived. This universe can be the entire population, the population 16 years and up, workers 16 years and up, etc.

When downloading ACS tables through the Census API, remember these things:

- Know the *unit* of the numerator and denominator.
- Does the unit change across years? Sometimes, the table will undergo a change; it will change from reporting count values to percent values after a certain year.
- Are variables are stable in reporting the same information? Particuarly, if the table has undergone a change, new columns might be added, such as `C02`. The *same* information might be found in `C01` from 2010-2013, and then in `C02` from 2014-onward.

We broadly group ACS tables into 3 types in our data cleaning process:

1. Count tables: counts are provided for numerator and denominator. Ex: # households that fall into particular income range, as well as the total # households overall within a census tract (or any other geography)
2. Percent tables: percent for numerator and count for denominator. Ex: 15 for people with less than HS education (which is 15%, not a count of 15 people), and 1,000 households in census tract. These need to be converted to counts from percents using the denominator.
3. Dollar tables: median household income or aggregate income, inflation-adjusted for each year. These tables separate, particularly because median household income is a tricky topic. Users beware! Do not calculate summary statistics from median income values (average median income is meaningless); only report values as is. Users should think carefully about what the ACS is reporting and how to use it meaningfully in analysis.

## General Functions

The main function is `transform_census_percent`, which uses 3 sub-functions, each of which can be used on its own. `transform_census_percent` takes a long, cleaned Census df, grabs one or more columns to aggregate, and reshapes the df to be wide. Example notebook.

`transform_census_percent()`: this function subsets the Census df for a particular table and year, grabs the relevant rows, aggregates them, renames the aggregated row, and then calculates the percent. The specifics of the function are best illustrated in an example.

```
import laplan

commute_group = [
    "workers_transit",
    "workers_walk",
    "workers_bike"
]

# Grab the 2018 commute table for all workers
# Aggregate transit, walk and bike
# Rename aggregated group as "non_car_workers"
# Calculate percent (non_car_workers / workers_total)
# Numerator is non_car_workers
# Denominator is workers_total
# Rename this new column "pct_non_car_workers"

laplan.census.transform_census_percent(
    "commute",
    2018,
    "workers",
    commute_group,
    "non_car_workers",
    "non_car_workers",
    "workers_total"
)
```

Cleaned, long Census df:

| GEOID | variable | year | table | main_var | second_var | new_var | num |
|---|---|---|---|---|---|---|---|
| A | S0801_C01_001 | 2018 | commute | workers | total | workers_ total | 1000 |
| A | S0801_C01_009 | 2018 | commute | workers | transit | workers_ transit | 50 |
| A | S0801_C01_010 | 2018 | commute | workers | walk | workers_ walk | 10 |
| A | S0801_C01_011 | 2018 | commute | workers | bike | workers_ bike | 20 |

`transform_census_percent` returns a wide df that looks like:

| GEOID | non_car_workers | workers_total | pct_non_car_workers |
|---|---|---|---|
| A | 80 | 1000 | 0.08 |

The sub-functions can be used individually, and should be used to construct reshaped income and race/ethnicity tables.

- Median household income: `subset_census_table` will return those values needed. No aggregation should be done!
- Households by income ranges: this table is used in conjunction with the `income_percentiles` function to re-calculate median household incomes.
- Race/ethnicity: use sub-functions because race/ethnicity groups can sum over 100%; race and ethnicity are *not* mutually exclusive.

`subset_census_table(table, table_name, year, main_var)`: subsets our cleaned, long Census df and grabs a particular table, year, and main_var.

```
# census_df is a df of
# cleaned, long Census data described above.

# 2018 commute mode table
subset_census_table(
    census_df,
    "commute",
    2018,
    "workers"
)
```

`aggregate_group(df, aggregate_me, name="aggregated_group")`: this function takes the df from subset_census_table and aggregates several rows into one. If no aggregation is needed, simply provide a list of 1. The list is made up of value(s) from new_var.

```
# To aggregate 3 groups:
# Rename new_var to "non_car_workers"
commute_group = [
    "workers_transit",
    "workers_walk",
    "workers_bike"
]

aggregate_group(
    df,
    commute_group,
    "non_car_workers"
)

# To aggregate 1 group
# Rename new_var to "zero_veh_workers"

vehicle_group = ["workers_veh0"]
aggregate_group(
    df,
    vehicles_group,
    "zero_veh_workers"
```

```
)
```

`make_wide(df, cols)`: this function takes reshapes the df from long to wide, or takes rows and pivots them to be columns. Cols is a list of values in from new_var.

```
reshape_me = ['non_car_workers', 'workers_total']

make_wide(df, reshape_me)
```

## Income Functions

The income functions are `make_income_range_wide` and `income_percentiles`.

`make_income_range_wide(df, year, main_var="total")`: subsets the long, cleaned Census df and grabs the `incomerange` table for a particular year and main_var. The default main_var is `total`, which is all households, rather than a specific race or ethnicity.

```
# 2018 all households by income bins
make_income_range_wide(
    census_df,
    2018,
)

# 2018 white households by income bins
make_income_range_wide(
    census_df,
    2018,
    "white"
)
```

`income_percentiles`: takes a df and returns the estimated income percentiles. This can be used to re-calculate the median household income (50th percentile). The households are aggregated done to a larger geographic area after `make_income_range_wide`, after which the median household income is re-calculated over this larger geographic area. If no aggregation is needed, then using the median household income table is sufficient in itself. The function returns percentiles in thousands of dollars, so multiply by 1,000 to get the result in dollars.

```
# Calculate 25th, 50th, and 75th percentiles.
iqr_df = (df.apply(
        lambda r:
        pd.Series(laplan.census.income_percentiles(
            r, [25,50,75]),
            dtype="float64"),
            axis=1,
    ).rename(
        columns={0: "Q1",
                1: "Q2",
                2: "Q3"})
)
```

`iqr_df` looks like (note: units are thousands of dollars):

| GEOID | Q1 | Q2 | Q3 |
| --- | --- | --- | --- |
| A | 30.5 | 55.7 | 82.6 |
| B | 40.5 | 58.7 | 90.6 |

# Appendix B TOC Entitlements by Rail Station

The full table of number of TOC entitlements by rail station is shown below.

**Table 3 TOC Entitlements by Rail Station**

| Station | Line | # TOC | Station % TOC Entitlements |
|---|---|---|---|
| Wilshire / Western Station | Purple | 51 | 12.7% |
| Vermont / Santa Monica Station | Red | 34 | 8.5% |
| Wilshire / Vermont Station | Red/Purple | 28 | 7.0% |
| Wilshire / Normandie Station | Purple | 23 | 5.7% |
| Hollywood / Vine Station | Red | 22 | 5.5% |
| Vermont / Beverly Station | Red | 21 | 5.2% |
| Westlake / MacArthur Park Station | Red/Purple | 20 | 5.0% |
| Wilshire / Fairfax | Purple | 18 | 4.5% |
| Vermont / Sunset Station | Red | 15 | 3.7% |
| Wilshire / La Cienega | Purple | 12 | 3.0% |
| Westwood / UCLA | Purple | 12 | 3.0% |
| Hollywood / Western Station | Red | 12 | 3.0% |
| North Hollywood Station | Red | 11 | 2.7% |
| Century City / Constellation | Purple | 10 | 2.5% |
| Crenshaw / Slauson | Crenshaw | 10 | 2.5% |
| Hollywood / Highland Station | Red | 9 | 2.2% |
| Palms Station | Expo | 8 | 2.0% |
| Culver City Station | Expo | 7 | 1.7% |
| Westwood / Rancho Park Station | Expo | 7 | 1.7% |
| Expo / Sepulveda Station | Expo | 7 | 1.7% |
| Florence / West | Crenshaw | 6 | 1.5% |
| Florence / Hindry | Crenshaw | 6 | 1.5% |

| | | | |
|---|---|---|---|
| Expo / Bundy Station | Expo | 5 | 1.2% |
| Expo / Crenshaw Station | Expo | 5 | 1.2% |
| Farmdale Station | Expo | 5 | 1.2% |
| Union Station - Metro Red & Purple Lines | Red/Purple | 5 | 1.2% |
| Union Station - Metro Gold Line | Gold | 4 | 1.0% |
| Leimert Park | Crenshaw | 4 | 1.0% |
| Pico / Aliso Station | Gold | 3 | 0.7% |
| Mariachi Plaza / Boyle Heights Station | Gold | 3 | 0.7% |
| Wilshire / La Brea | Purple | 3 | 0.7% |
| Heritage Square / Arroyo Station | Gold | 3 | 0.7% |
| Soto Station | Gold | 2 | 0.5% |
| Pico Station | Blue/Expo | 2 | 0.5% |
| Expo / Vermont Station | Expo | 2 | 0.5% |
| Expo / La Brea Station | Expo | 2 | 0.5% |
| Grand / LATTC Station | Blue | 2 | 0.5% |
| Universal / Studio City Station | Red | 1 | 0.2% |
| Chinatown Station | Gold | 1 | 0.2% |
| La Cienega / Jefferson Station | Expo | 1 | 0.2% |

*Source: PCTS, LA Metro Developer GIS Data; analysis by ITA Data Science*

### Table 5 TOC Entitlements by Bus Route

| Bus Route | # TOC | Route's % TOC Entitlements |
|---|---|---|
| 754 | 141 | 9.3% |
| 757 | 125 | 8.2% |
| 720 | 102 | 6.7% |
| 780 | 86 | 5.7% |
| 207 | 63 | 4.2% |
| 710 | 61 | 4.0% |
| 10/48 | 57 | 3.8% |
| 33 | 53 | 3.5% |
| 704 | 49 | 3.2% |
| 204 | 47 | 3.1% |
| 603 | 43 | 2.8% |
| R12 | 42 | 2.8% |
| 206 | 41 | 2.7% |
| 217 | 41 | 2.7% |
| 200 | 40 | 2.6% |
| 728 | 39 | 2.6% |
| 14/37 | 38 | 2.5% |
| 4 | 33 | 2.2% |
| 224 | 33 | 2.2% |
| 233 | 28 | 1.8% |
| 705 | 27 | 1.8% |
| 901 | 24 | 1.6% |
| 163/162 | 23 | 1.5% |
| 35/38 | 23 | 1.5% |

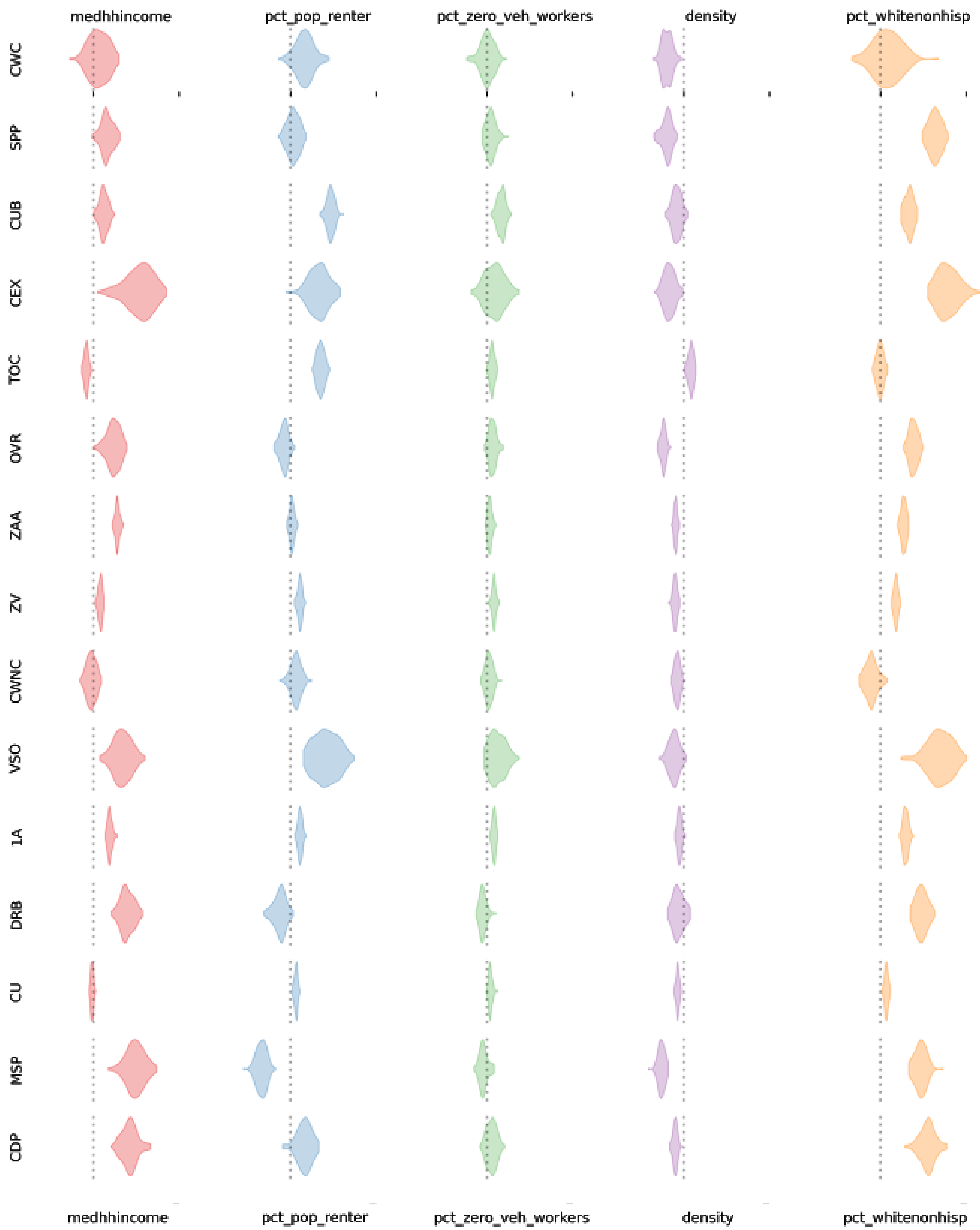
| | | |
|---|---|---|
| 30/330 | 22 | 1.5% |
| 164 | 20 | 1.3% |
| 212/312 | 20 | 1.3% |
| 28 | 15 | 1.0% |
| CC 6R | 15 | 1.0% |
| 16/17/316 | 14 | 0.9% |
| 165 | 13 | 0.9% |
| 105 | 12 | 0.8% |
| CC 1 | 11 | 0.7% |
| 7 | 10 | 0.7% |
| 110 | 10 | 0.7% |
| 501 | 8 | 0.5% |
| 20 | 7 | 0.5% |
| 111 | 7 | 0.5% |
| 3 | 6 | 0.4% |
| 751 | 6 | 0.4% |
| 152/353 | 6 | 0.4% |
| 1 | 5 | 0.3% |
| 18 | 4 | 0.3% |
| 8 | 4 | 0.3% |
| 53 | 4 | 0.3% |
| 2/302 | 4 | 0.3% |
| 770 | 4 | 0.3% |
| 76 | 3 | 0.2% |
| 745 | 3 | 0.2% |
| 180/181 | 3 | 0.2% |

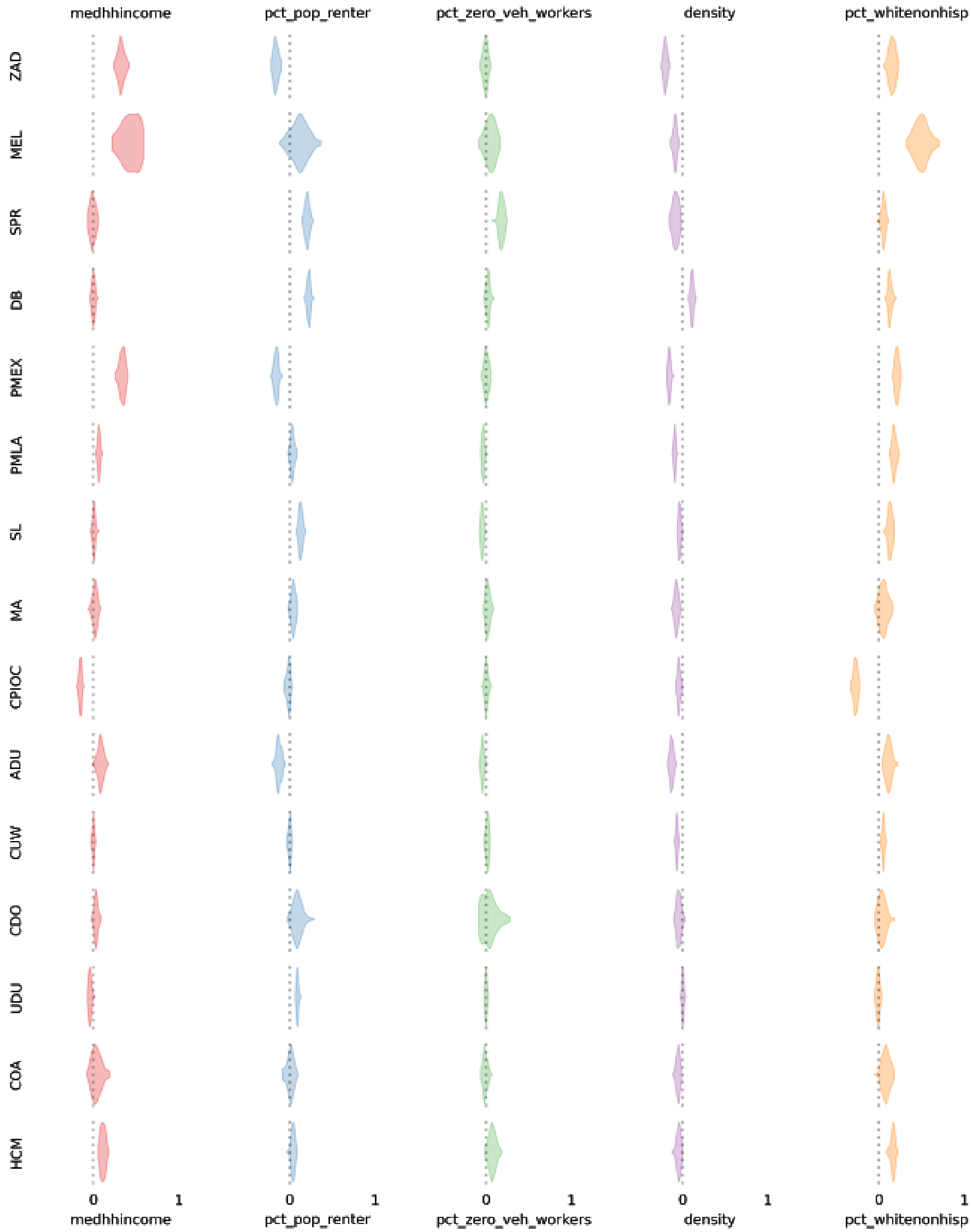
| | | |
|---|---|---|
| 760 | 2 | 0.1% |
| 910/950 | 2 | 0.1% |
| 78/79/378 | 2 | 0.1% |
| 40 | 2 | 0.1% |
| 66 | 2 | 0.1% |
| 750 | 2 | 0.1% |
| 55/355 | 2 | 0.1% |
| 60 | 2 | 0.1% |
| 51/52/351 | 1 | 0.1% |
| 605 | 1 | 0.1% |
| 150/240 | 1 | 0.1% |
| 115 | 1 | 0.1% |
| 68 | 1 | 0.1% |

*Source: PCTS, GTFS, LA Metro Developer GIS Data; analysis by ITA Data Science*

# Appendix C General Regression Results

The figure below shows the results of taking the Poisson regression and estimating the coefficients over 100 samples (to get a distribution of the estimated coefficients).

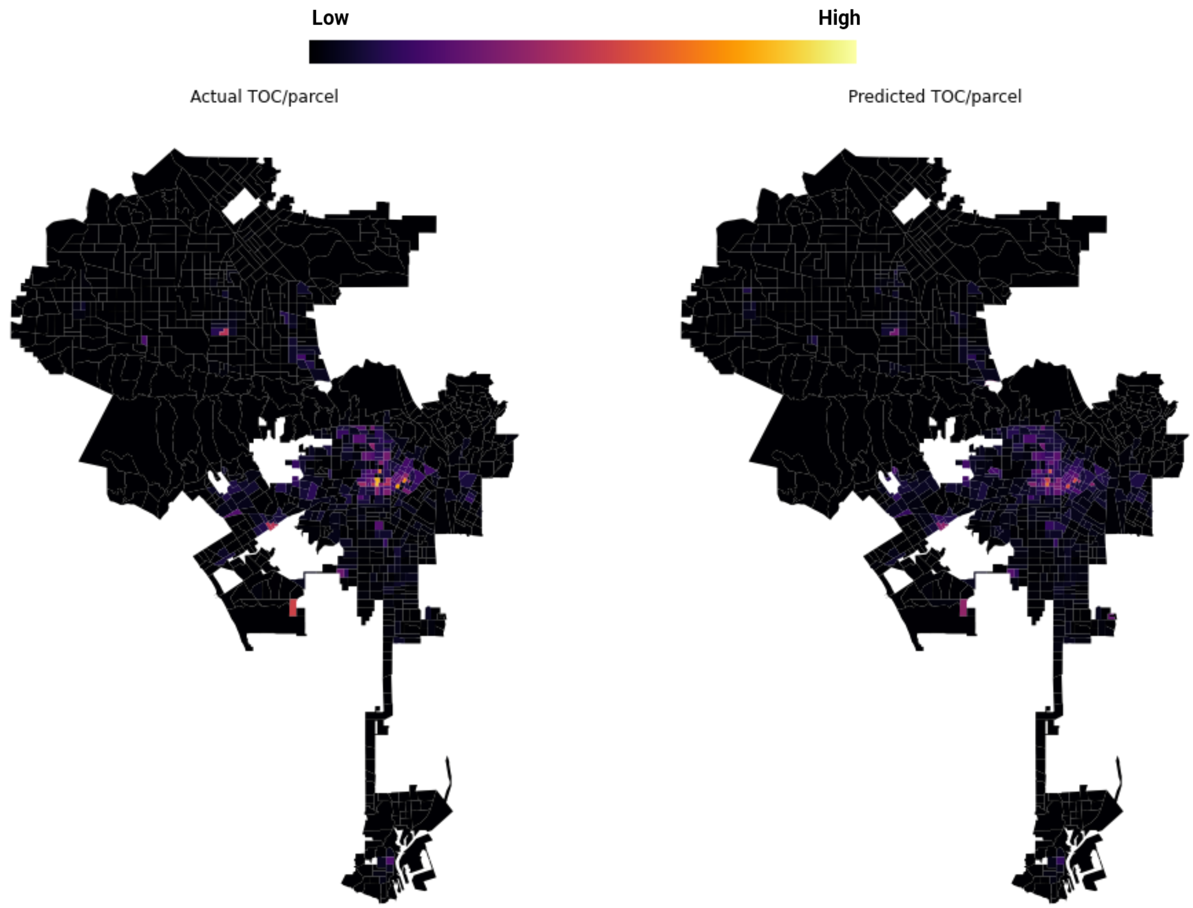



medhhincome
pct_pop_renter
pct_zero_veh_workers
density
pct_whitenonhisp
ZAD
MEL
SPR
DB
PMEX
PMLA
SL
MA
CPIOC
ADU
CUW
CDO
UDU
COA
HCM
0
1
medhhincome
pct_pop_renter
pct_zero_veh_workers
density
pct_whitenonhisp

# Appendix D TOC Regression Results

**Figure 1** shows that using all the predictors does pretty well at predicting where the TOC entitlements are going to be. But, **Figure 2** shows that we can achieve roughly the same result by using a much smaller set of predictors.

**Figure 1 Random Forest Regression: Actual vs Predicted TOC Entitlements per Parcel (All Predictors)**




*Source: PCTS, ACS 5-year estimates; ZIMAS zoning; TOC Tiers; analysis by ITA Data Science*

To get to our smaller set of predictors, we look at both "feature importance" and "permutation importance".

Feature importance sorts each predictor variable into an order of relative importance in predicting the outcome variable. The four most important predictors are % favorable TOC zoning, % Tier 2 parcels, % Tier 3 parcels, and % renter.

```
Feature importance
 pct_favorable_toc_zoning     0.324363
 Tier_2                       0.137294
```

| | |
|---|---|
| Tier_3 | 0.134571 |
| pct_pop_renter | 0.074533 |
| pct_zero_veh_workers | 0.057386 |
| pct_eligible_zoning | 0.054022 |
| medhhincome | 0.052795 |
| pct_whitenonhisp | 0.051502 |
| density | 0.04892 |
| Tier_1 | 0.04262 |
| Tier_4 | 0.021995 |

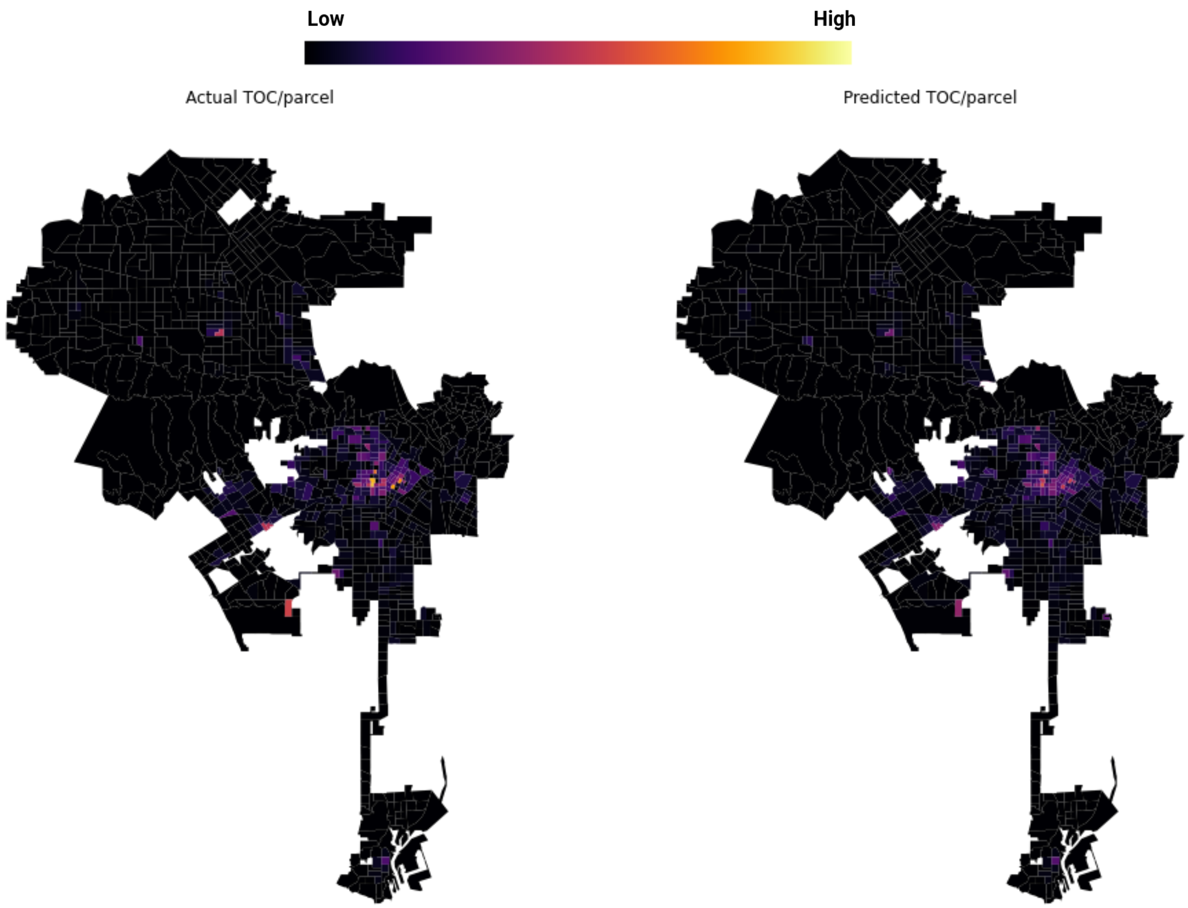
Since random forest starts with taking one variable, splitting up the data into different groups, then moves on and continues to split up the data on a second, third variable, and so forth, which variable it starts splitting the data on is important. The “permutation importance” switches the order. If the results change dramatically, it means the order of the variables was important. If the results don’t change dramatically, the relative importance of a particular predictors is more robust. In other words, permutation importance acts as a sensitivity analysis to double check that the predictors identified as important are not that sensitive to tweaks in the model. Our permutation importance check shows that the same four predictors are still ranked as the four most important. We are more confident that these four predictor variables do a good job at capturing the relationship with TOC entitlements.

Permutation importance

| | |
|---|---|
| pct_favorable_toc_zoning | 1.038712 |
| Tier_3 | 0.409482 |
| Tier_2 | 0.324424 |
| pct_pop_renter | 0.107643 |
| pct_zero_veh_workers | 0.07239 |
| Tier_1 | 0.069896 |
| pct_eligible_zoning | 0.061694 |
| pct_whitenonhisp | 0.056987 |
| medhhincome | 0.054486 |
| density | 0.044569 |
| Tier_4 | 0.036737 |

**Figure 2** shows that the random forest regression performs similarly well, even after narrowing down our full list of predictor variables down to four.

**Figure 2 Random Forest Regression: Actual vs Predicted TOC Entitlements per Parcel (Fewer Predictors)**

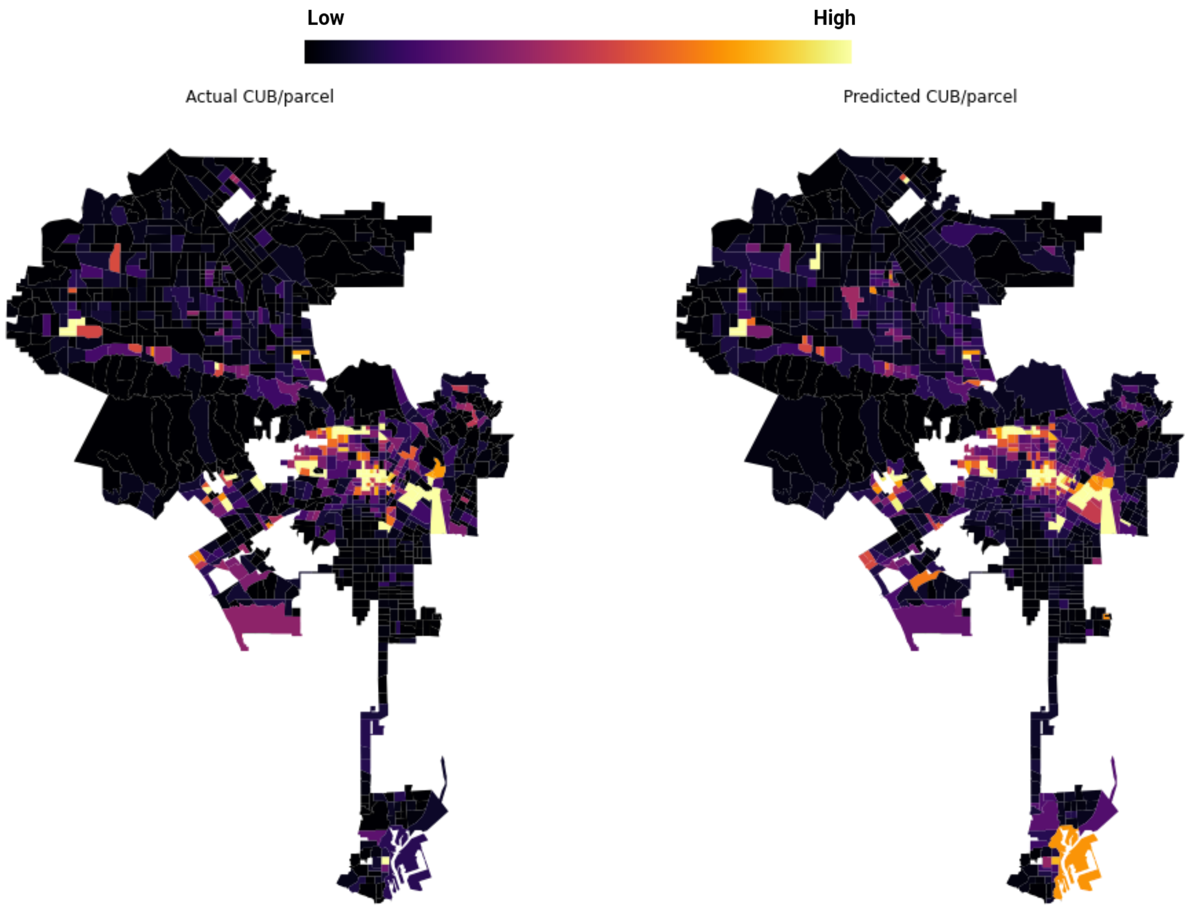



*Source: PCTS, ACS 5-year estimates; ZIMAS zoning; TOC Tiers; analysis by ITA Data Science*

# Appendix E CUB Regression Results

**Figure 1** shows that using all the predictors does pretty well at predicting where the TOC entitlements are going to be. But, **Figure 2** shows that we can achieve roughly the same result by using a much smaller set of predictors.

**Figure 1 Random Forest Regression: Actual vs Predicted CUB Entitlements per Parcel (All Predictors)**




*Source: PCTS, ACS 5-year estimates; ZIMAS zoning; analysis by ITA Data Science*

To get to our smaller set of predictors, we look at both “feature importance” and “permutation importance”.

Feature importance sorts each predictor variable into an order of relative importance in predicting the outcome variable. The four most important predictors are % favorable TOC zoning, % Tier 2 parcels, % Tier 3 parcels, and % renter.

```
Feature importance
 n_eligible                 0.433692

 pct_whitenonhisp           0.183764
```

| | |
|---|---|
| density | 0.118279 |
| pct_zero_veh_workers | 0.107967 |
| pct_pop_renter | 0.079773 |
| medhhincome | 0.076524 |
| n_eligible | 0.433692 |

Since random forest starts with taking one variable, splitting up the data into different groups, then moves on and continues to split up the data on a second, third variable, and so forth, which variable it starts splitting the data on is important. The “permutation importance” switches the order. If the results change dramatically, it means the order of the variables was important. If the results don’t change dramatically, the relative importance of a particular predictors is more robust. In other words, permutation importance acts as a sensitivity analysis to double check that the predictors identified as important are not that sensitive to tweaks in the model.

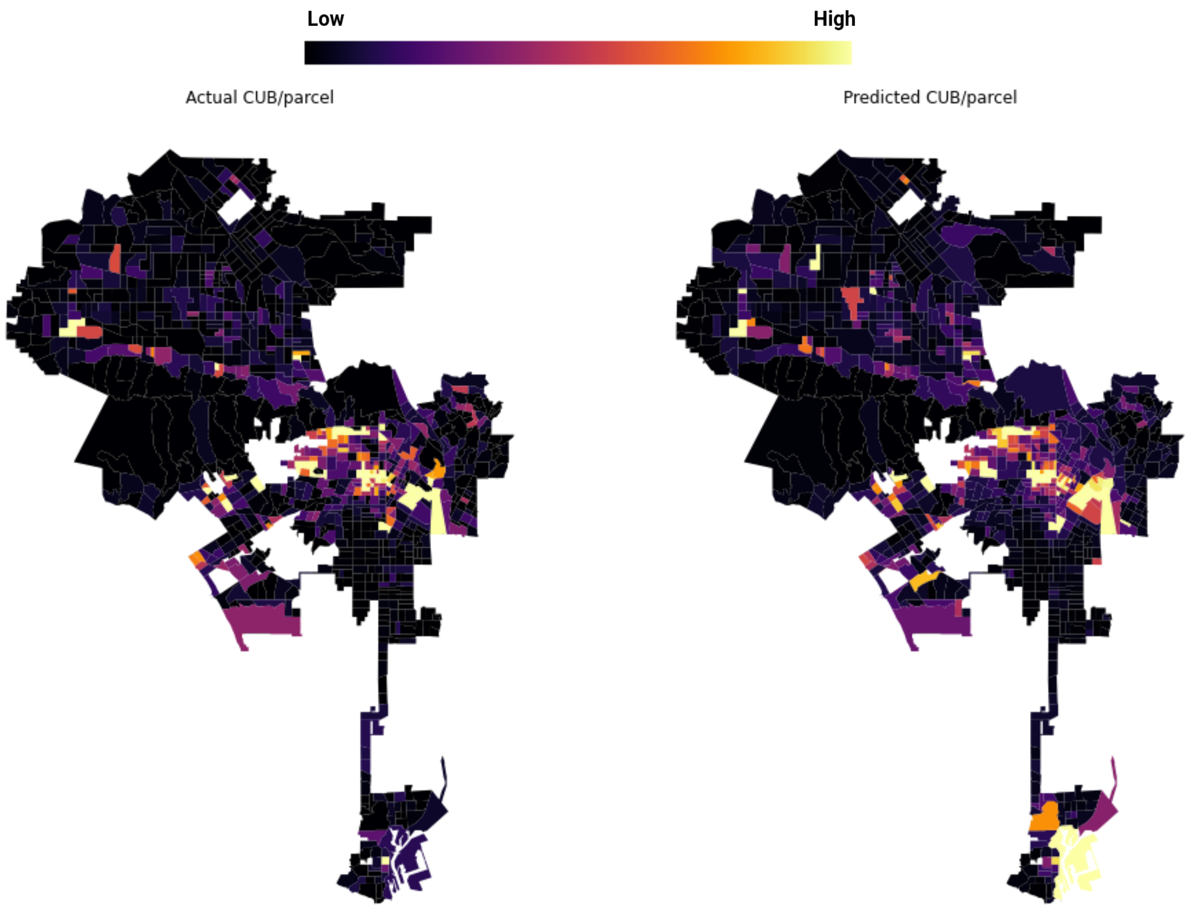
Our permutation importance check shows that the same two predictors are still ranked as the two most important (# parcels with eligible zoning, % white non-Hispanic). However, the subsequent variables, although they do are sorted in the same order as in the feature importance, they are highly correlated with each other.

When population density is included, the coefficient is significant, but its impact is so small, even a 100% increase (doubling population density) would not change the number of CUB entitlements. When either % zero vehicle workers or % renter is included, the coefficient is the same, meaning these two predictors are highly collinear with one another. Since they basically move together, it is hard for the regression to attribute an effect to either one; rather, the effect is attributed to whichever variable is included in the regression.

Permutation importance

| | |
|---|---|
| n_eligible | 0.965982 |
| pct_whitenonhisp | 0.423037 |
| density | 0.179443 |
| pct_zero_veh_workers | 0.139164 |
| pct_pop_renter | 0.093016 |
| medhhincome | 0.08233 |

**Figure 2 Random Forest Regression: Actual vs Predicted CUB Entitlements per Parcel (Fewer Predictors)**




*Source: PCTS, ACS 5-year estimates; ZIMAS zoning; analysis by ITA Data Science*