In [1]:
```
# libraries
import numpy as np
import pandas as pd
import altair as alt
alt.renderers.enable('default')
# warnings
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
# display settings
pd.options.display.max_columns = None
```

PSTAT 100 Project plan

Group members:

Contributions:

1. Kabir Snell: Found the data set, wrote the background, and came up with questions for future work
2. Member 2 worked on tidying the dataset.
3. Member 3 worked on variable summaries and explanatory plots
4. Member 4 worked on

---

0. Background

California has a well-developed public transportation system that includes buses, trains, light rail, and subways. The state's largest public transportation agency is the Los Angeles County Metropolitan Transportation Authority (Metro), which serves the Los Angeles metropolitan area with an extensive network of bus and rail lines. Other major public transportation agencies in California include the San Francisco Municipal Transportation Agency (SFMTA) and the San Diego Metropolitan Transit System (MTS). In addition to these larger agencies, many smaller cities and towns throughout California have their own public transportation systems, which may include buses, shuttles, or other types of services. These systems often provide connections to regional or statewide transportation networks, making it possible for people to travel throughout the state using public transportation. While many areas have well-established public transportation systems, there may be gaps or limitations in service in some parts of the state, particularly in more rural or remote areas.

The Walkable Distance to Public Transit dataset, available on data.ca.gov, provides information on the number of households in California that are within a certain distance of public transit stops. The data is based on estimates of walking distance from households to the nearest transit stop, as well as information on the types of transit services available at each stop.

The dataset is organized by county, and includes information on the number and percentage of households within various walking distance ranges of transit stops. For example, the data might indicate how many households are within a 5-minute walk of a transit stop, as well as how many households are within a 10-minute or 15-minute walk.

The dataset also provides information on the types of transit services available at each stop, including whether the stop serves buses, trains, or other modes of transportation. This information can be useful for understanding the accessibility of public transportation in different areas of the state, and for identifying opportunities to improve transit service or expand transit infrastructure.

Overall, the Walkable Distance to Public Transit dataset is a valuable resource for policymakers, transportation planners, and researchers who are interested in understanding the availability and accessibility of public transportation in California.

---

1. Data description

Basic Information

The data is the percent of the population that resides within 1/2 mile of a major transit location (bus/rail/ferry stop) in four California regions, and whose waiting time is less than 15 minutes during peak commute hours. The data is stratified by 8 race/ethnicity groups and includes both geographic information and statistical reliability measurements.

The data was collected by the California Department of Public Health, as part of the "Health Communities Data and Indicators Project (HCI). The goal of the project was to evaluate how city plans and policies affect community health. The data includes 2012 Transit Stops from the San Diego and Southern California Association of Governments, as well as the Metropolitan Transportation Commission. It also includes 2008 Transit Stops from the Sacramento Council of Government and 2010 block-level population data from the U.S. Census Bureau. The data is updated decennially. The four California regions are defined as the following:

- Southern California (SCAG): Imperial, Los Angeles, Orange, Riverside, San Bernardino, and Ventura
- Sacramento (SACOG): Placer, Sacramento, and Yolo
- Bay Area (MTC): Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, Sonoma
- San Diego County

Data values were obtained using automated methods to download information from various public websites. One important data set was from the 2010-2012 California Household Travel Survey. Multiple data collection methods were used in this survey, including computer-assisted telephone interviews and online/mail surveys. To identify census blocks inside 1/2 mile of the transit stops, geospatial software was used. In order to compile the data into one data set, the census blocks from the 2010 U.S. Census were merged with the blocks from the transit data, and population counts were aggregated by census tract, cities/towns, county, and region. The data was processed into Excel files with standard formats.

The population is adults aged 18 years and over, who reside in the four California regions. The sampling frame includes adults in these four regions, with access to telephone or mail services. The sampling mechanism for the respective year (2008 or 2012) is a probability sample because the surveys downloaded by the HCI project were sent to randomly selected adults However, the scope of inference has limitations. The data is from the year 2012 for the SCAG, MTC, and San Diego regions; while, 2008 for the SACOG region. Some transit stops and services may have changed during that time period. As well, the population data was collected from the 2010 U.S. Census, which is a different time period than the transit data (2008, 2012). Therefore, some variation may exist if demographics changed.

Data semantics and structure

Name	Variable description	Type	Units of measurement
year	year when data was reported	Numeric	Calendar year
race_eth_name	name of the different races/ethnicities ('AfricanAm', 'AIAN', 'Asian', 'Latino', 'Multiple', 'NHOPI', 'Other', 'Total', 'White')	Object	Name
geotype	describes the level of geography for data in that row ('RE'=region, 'CT'=census tract, 'PL'=place/town/city, 'CO'=county)	Object	Name
geoname	name of the city/town	Object	Name
county_name	name of the county	Object	Name
region_name	name of the region ('Sacramento Area', 'Bay Area', 'San Diego', 'Southern California')	Object	Name
pop_trans_acc	number of residents that live within 1/2 mile of public transportation	Numeric	Integer
pop2010	total number of residents that reside in that county	Numeric	Integer
p_trans_acc	the percent of residents that live within 1/2 mile of public transportation	Numeric	Float
LL_95CI	lower limit of the 95th confidence interval for p_trans_acc	Numeric	Float
UL_95CI	upper limit of the 95th confidence interval for p_trans_acc	Numeric	Float
se	standard errror	Numeric	Float
rse	relative standard error	Numeric	Float

Data Overview

In [2]:
```
# load tidied data and print rows
data = pd.read_csv(
    'tidy-data',
    dtype = {'pop_trans_acc':'Int64',
             'county_fips': 'Int64'},
    index_col = 0
)

data.head()
```

Out[2]:

	year	race_eth_code	race_eth_name	geotype	geotypevalue	geoname	county_name	county_fips	region_name	region_code	pop_trans_acc	pop2010	p_trans_acc	LL_95CI
0	2008	3	AfricanAm	CO	6061	Placer	Placer	6061	Sacramento Area	8	55	4427	0.012424	0.009161
1	2008	1	AIAN	CO	6061	Placer	Placer	6061	Sacramento Area	8	51	2080	0.024519	0.017873
2	2008	2	Asian	CO	6061	Placer	Placer	6061	Sacramento Area	8	117	19963	0.005861	0.004802
3	2008	4	Latino	CO	6061	Placer	Placer	6061	Sacramento Area	8	1835	44710	0.041042	0.039203
4	2008	7	Multiple	CO	6061	Placer	Placer	6061	Sacramento Area	8	241	10658	0.022612	0.019790

---

2. Initial Explorations

Basic properties of the dataset

(a) Dimensions of the data

In [3]:
```
data.shape
```

Out[3]:
```
(66006, 17)
```

The data set contains 66,006 rows (66,005 observations / 1 header) and 17 columns (variables).

(b) Missing values

In [4]:
```
pd.DataFrame(data.isna().sum()).transpose().rename(index = {0: 'missing'})
```

Out[4]:

	year	race_eth_code	race_eth_name	geotype	geotypevalue	geoname	county_name	county_fips	region_name	region_code	pop_trans_acc	pop2010	p_trans_acc	LL_9
missing	0	0	0	0	0	54	63	63	0	0	1182	0	1557	11

Yes, there are missing values in the data set. The variables missing are 'geoname,' 'county_name,' 'county_fips,' 'pop_trans_acc,' 'p_trans_acc' and the statsitical reliability measurements: ('LL_95CI,' 'UL_95CI,' 'se,' and 'rse.') The majority of variables missing are from these statistical measurements. In addition, none of the variables are missing more than 2.5% of the time, except 'rse' (22.3%). Some of the values are missing because the 'geotype' is stratified by four levels: ('RE'=region, 'CT'=census tract, 'PL'=place/town/city, 'CO'=county). Therefore, 'geoname' or 'county_name' may be missing if the row observation is for the overall region ('RE'). 'pop_trans_acc' is missing when the 'pop2010' (2010 population) was 0. And as a result of this, the stastical reliability measurements are missing for these rows too.

(c) Variable summaries

Value counts of the race/ethnicity groups by year

In [5]:
```
pd.DataFrame(data.groupby(['year']).race_eth_name.value_counts()).transpose().rename(index = {'race_eth_name': 'count'})
```

Out[5]:

year	2008									2012							
race_eth_name	AIAN	AfricanAm	Asian	Latino	Multiple	NHOPI	Other	Total	White	AIAN	AfricanAm	Asian	Latino	Multiple	NHOPI	Other	Total	White
count	518	518	518	518	518	518	518	518	518	6816	6816	6816	6816	6816	6816	6816	6816	6816

- Since the 2008 study was confined to the Sacramento region, there are less observations for each race/ethnicity group. In the year 2012, the study incorporated three regions (Southern California, Bay Area, San Diego) and therefore, had more observations for each group.

Summary statistics of the percent of residents that reside within 1/2 mile of public transportation. Sorted by year, region, and race/ethnicity

In [6]:
```
# grouping the data by the level = 'RE' (Sacramento, Bay Area, and Southern California) to provide a regional overview
# the San Diego region has no level 'RE,' so group by 'CO' and this gives same results

data_region = data[(data.geotype == 'RE') | ((data.region_name == 'San Diego') & (data.geotype == 'CO'))]
data_region.groupby(['year', 'region_name', 'race_eth_name']).p_trans_acc.describe().drop(columns = {'std'}).head()
```

Out[6]:

year	region_name	race_eth_name	count	mean	min	25%	50%	75%	max
2008	Sacramento Area	AIAN	1.0	0.191476	0.191476	0.191476	0.191476	0.191476	0.191476
		AfricanAm	1.0	0.212514	0.212514	0.212514	0.212514	0.212514	0.212514
		Asian	1.0	0.211100	0.211100	0.211100	0.211100	0.211100	0.211100
		Latino	1.0	0.189178	0.189178	0.189178	0.189178	0.189178	0.189178
		Multiple	1.0	0.186008	0.186008	0.186008	0.186008	0.186008	0.186008

(Chart continues below with the other regions...)

- This chart shows how access to public transportation varies across each region. Across all race/ethnicity groups, the Bay Area has the highest mean access to public transportation. The Sacramento Area has the lowest mean access. There are similar public transportation rates between the San Diego and Southern California regions.

Value counts of the counties sorted by year and region

In [7]:
```
pd.DataFrame(data.groupby(['year', 'region_name']).county_name.value_counts()).transpose().rename(index = {'county_name': 'count'})
```

Out[7]:

year	2008																		
region_name	Sacramento Area			Bay Area									San Diego	Southern Cal					
county_name	Sacramento	Placer	Yolo	Santa Clara	Alameda	Contra Costa	San Francisco	San Mateo	Sonoma	Solano	Marin	Napa	San Diego	Los Angeles	Orange	Riverside	San Bernardino	Ventura	In
count	3177	954	495	3573	3438	2358	1791	1719	1242	972	792	468	6138	22392	5625	4797	3789	1782	

- This chart lists the counties in each of the four regions and shows the number of observations per each county. It is important to note that the San Diego region only has one county named 'San Diego'. Out of the counties, Los Angeles had the most observations at 22,392 almost 4x higher than the next highest county of San Diego.

Value counts of the level of geography sorted by year and region

In [8]:
```
pd.DataFrame(data.groupby(['year', 'region_name']).geotype.value_counts()).transpose().rename(index = {'geotype': 'count'})
```

Out[8]:

year	2008				2012										
region_name	Sacramento Area				Bay Area				San Diego			Southern California			
geotype	CT	PL	CO	RE	CT	PL	CO	RE	CT	PL	CO	CT	PL	CO	RE
count	3987	639	27	9	14292	1989	81	9	5652	477	9	35604	3168	54	9

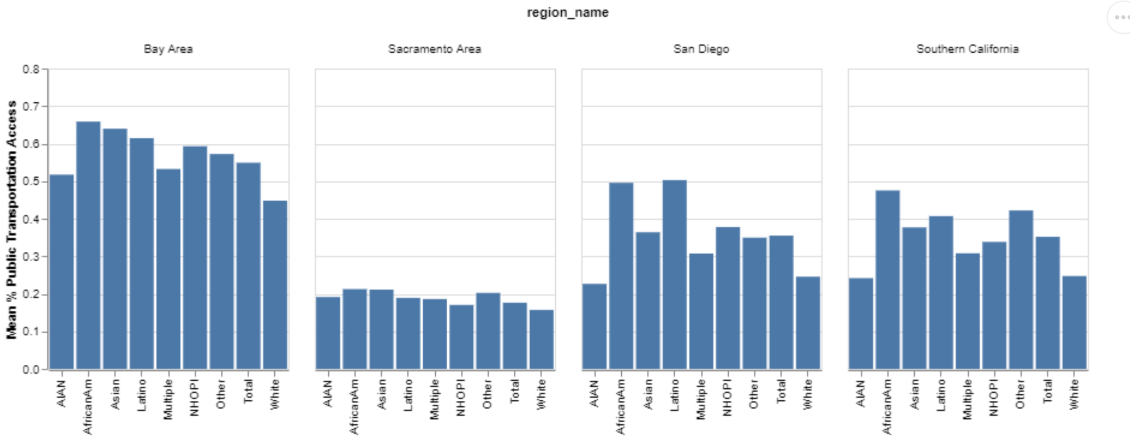
- 'geotype' describes the level of geography for data in that row *('RE'=region, 'CT'=census tract, 'PL'=place/town/city, 'CO'=county)*. Having more values for 'PL' means we have more granular data, as we can observe public transportation rates in a specific city. One example where the geotype is PL is from row 4015: 'Auburn city (PL) / Placer (CO) / Sacramento Area (RE). This allows us to examine a specific city, compared to the geotype being 'CO' or 'RE,' where we could only examine a county or region.

Exploratory analysis

(a) Access to public transportation by region and race/ethnicity

In [9]:
```
alt.Chart(data_region).mark_bar().encode(
    x = alt.X('race_eth_name', title = ''),
    y = alt.Y('mean(p_trans_acc)', scale=alt.Scale(domain=[0, 0.8]), title = 'Mean % Public Transportation Access')
).properties(width = 200, height = 250).facet(column = 'region_name')
```

Out[9]:

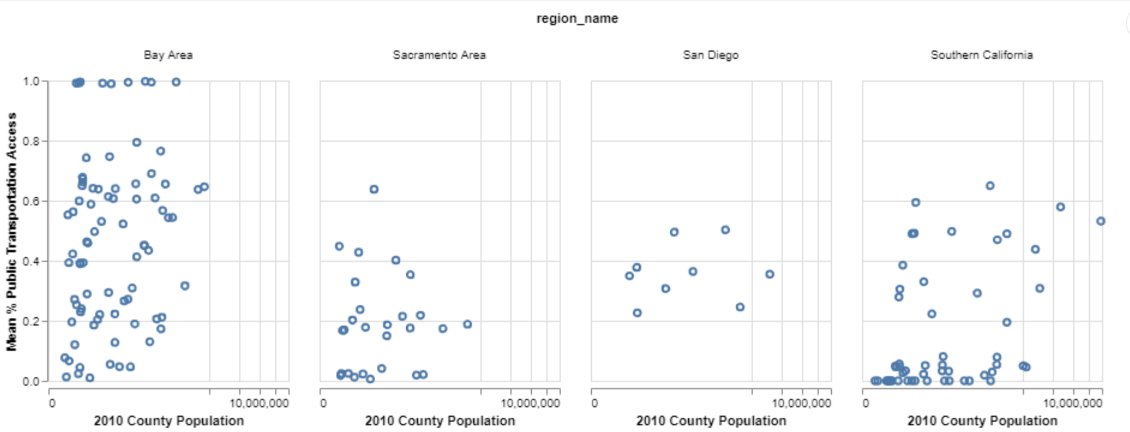


The bar charts compare the mean percentage of residents that live within 1/2 mile of public transportation across different race/ethnicity groups. It also separates each bar chart accoding to the different region and shows that region is a more significant indicator in access to public transportation, compared to race/ethnicity; however, differences are still present between each race/ethnicity.

(b) Population's relationship with access to public transportation

In [10]:
```
alt.Chart(data[(data.geotype == 'CO')]).mark_point().encode(
    x = alt.X('pop2010', scale = alt.Scale(zero = False, type = 'pow', exponent = 0.25), title = '2010 County Population'),
    y = alt.Y('mean(p_trans_acc)', title = 'Mean % Public Transportation Access')
).properties(width = 200, height = 250).facet(column = 'region_name')
```

Out[10]:



The scatter plots shows how access to public transportation varies with the population in each **county** in the region (the points represent each county). Counties with larger populations, specifically Southern California, have higher access to public transportation, but this relationship is less apparent in the other regions.

---

3. Planned work

Further Exploration Required:

(1) How does access to public transportation differ across race/ethnicity groups?

- Merge the regions together and sort them by race/ethnicity groups. Examine if the differences between groups exist at both the region level ('plot a' in Part 2) and statewide (California) level. Determine a threshold value where the difference between race/ethnicity groups is considered "significant."

(2) How does access to public transportation vary by region and the populations of the counties within each region?

- Sort counties in each region by their population. Determine if a linear or multiple regression model can fit mean access to public transport by population. Do outside research on certain counties, specifically those with large populations, to determine why they have high, or low rates for access to public transportation.