CA Transportation

Roger Wilson

08/13/22

Contents

Introduction and Data

This data set contains data on modes of transportation to work for Californian Residents aged 16 years and older. Data is from the U.S. Census Bureau, Decennial Census and American Community Survey. The data can be found at the California Health and Human Services Open Data Website.

According to the source, "Commute trips to work represent 19% of travel miles in the United States," with cars being by far the most used mode of transportation. While cars are convenient, they are one of the deadliest modes of transportation and release harmful emissions into the atmosphere. According to the California Air Resources Board, vehicles and cars account for approximately one-third of all of California's CO2 emissions and air pollutants.

Besides driving, commuters can take public transportation, bike, walk, work from home, or even carpool. All of these methods help take cars off of our roads, which can significantly cut emissions and save money.

For this project, we will look at the distributions of Californians that prefer each mode of transportation for work on the county and city level.

It's important to note that this data only includes Californians aged 16 years or older that commute to paid work.

```
# loading data
library(httr)
library(readxl)
GET("https://query.data.world/s/zolw2w2d7avcya3eqyhbtprcscpwjm",
    write_disk(tf <- tempfile(fileext = ".xlsx")))
```

```
## Response [https://download.data.world/file_download/chhs/b1008bb6-2f54-4b49-a0e0-c4d1bc9440ff/transp
##   Date: 2022-08-14 03:11
##   Status: 200
##   Content-Type: application/vnd.openxmlformats-officedocument.spreadsheetml.sheet
##   Size: 22.8 MB
## <ON DISK>  C:\Users\rswil\AppData\Local\Temp\RtmpCIRTF3\file46c6b4d6989.xlsx
```

```
df <- read_excel(tf)
```

```
## Warning: Expecting numeric in A202204 / R202204C1: got 'END OF TABLE'
```

```
names(df)
```

```
##  [1] "ind_id"         "ind_definition" "reportyear"     "race_eth_code"
##  [5] "race_eth_name"  "geotype"        "geotypevalue"   "geoname"
##  [9] "county_name"    "county_fips"    "region_name"    "region_code"
## [13] "mode"           "mode_name"      "pop_total"      "pop_mode"
## [17] "percent"        "LL95CI_percent" "UL95CI_percent" "percent_se"
## [21] "percent_rse"    "CA_decile"      "CA_RR"          "version"
```

```
unique(df$reportyear)
```

```
## [1] "2000"      "2005-2007" "2006-2010" "2008-2010" NA
```

```
# removing rows entirely NA
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
which(is.na(df$race_eth_code), arr.ind=TRUE)
```

```
## [1] 202203
```

```
df <- df %>% slice(-202203)
```

We don't need every column, so let's remove some of the redundant ones:

```
# removing unneeded columns
df = subset(df, select = !(names(df) %in%
                             c("ind_id", "ind_definition", "version",
                               "race_eth_code", "geotypevalue",
                               "region_code", "CA_decile")))
```

We will be working with the following columns:

column	description
reportyear	Year(s) that the indicator was reported
race_eth_name	Name of race/ethnic group
geotype	Type of geographic unit
geoname	Name of geographic unit
county_name	Name of county that geotype is in
county_fips	FIPS code of county that geotype is in
region_name	Metopolitan Planning Organization (MPO)-based region name
mode	Mode of transportation short name
mode_name	Mode of transportation long name

column	description
pop_mode	numerator, number of workers (16 years or older) by mode of transportation
pop_total	denominator, number of workers (16 years or older)
percent	Percent of Residents Mode of Transportation to Work, Population Aged 16 Years and Older
LL_95CI	Lower limit of 95% confidence interval
UL_95CI	Upper limit of 95% confidence interval
percent_se	Standard error of percent
percent_rse	Relative standard error (se/percent * 100) expressed as a percent
CA_RR	Rate ratio to California rate

More information can be found at this data dictionary.

County Level Analysis

Let's first look at the data from a county level. Intuitively, If a county is not that densely populated, its residents will be more dependent on cars to commute longer distances. On the other hand, if a county is densely populated, residents may be more willing to bike, walk, or take public transit because their commute is shorter. Let's create a subset of data just with the most recent percentages for all residents in each county.

```
# focusing just on countries
df_counties <- df %>%
  filter(race_eth_name == "Total",
         geotype == "CO" | geotype == "CA")
```

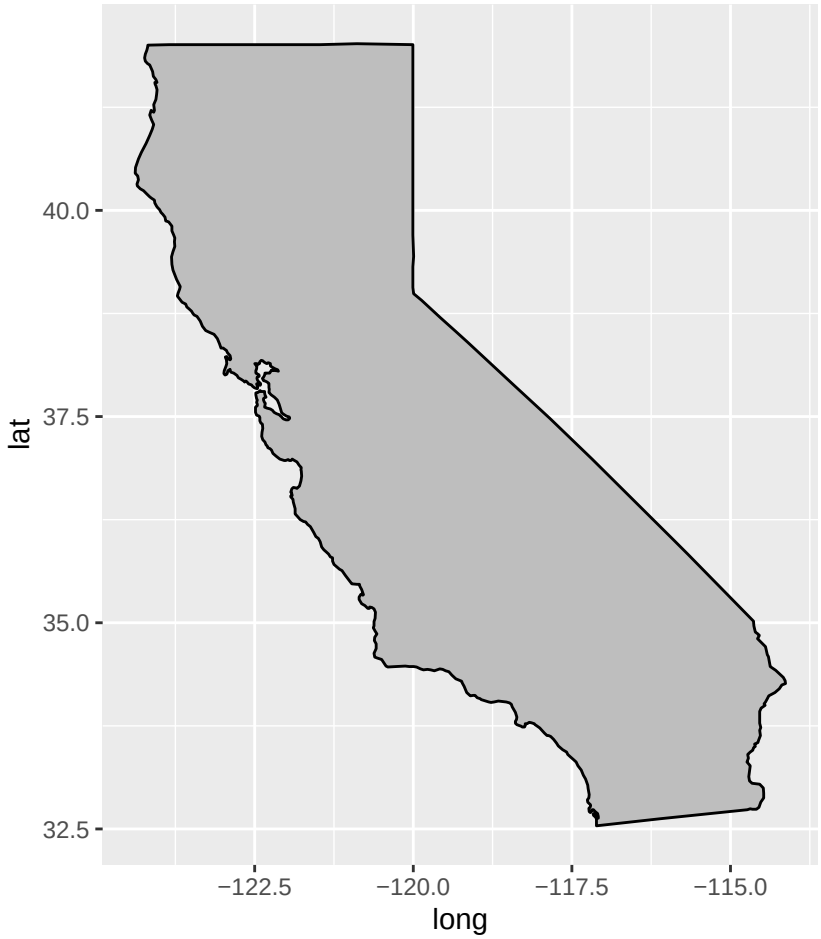
```
head(df_counties)
```

```
library(ggplot2)
library(maps)
library(mapdata)
library(stringr)
```

```
# getting borders
us_states <- map_data("state")
ca <- subset(us_states, region == "california")
us_counties <- map_data("county")
ca_counties <- subset(us_counties, region == "california")

rm(us_states, us_counties)
```

```
# creating base map
ca_base <- ggplot(data = ca,
                  mapping = aes(x = long, y = lat, group = group)) +
                  coord_fixed(1.3) +
                  geom_polygon(color = "black", fill = "gray")
ca_base
```



Looks good! Let's add some border FIPs do our data set so that we can graph our county borders.

```
# adding county borders to df_counties
ca_counties$subregion <- str_to_title(ca_counties$subregion)
df_counties <- left_join(df_counties, ca_counties,
                         by = c("county_name" = "subregion"))
rm(ca_counties)
```

From this Wikipedia on Californian Counties, we can get information on the total area of each county. This will help us graph population densities.

```
library(XML)
url <- "https://en.wikipedia.org/wiki/List_of_counties_in_California"
r <- GET(url)
doc <- readHTMLTable(doc = content(r, "text"), header = TRUE)
areas = doc[2]
```

```
# getting the correct format
areas <- data.frame(areas)
areas <- areas %>%
  select("X..County", "X..Area.6.")
areas <- areas %>%
  rename("County" = "X..County",
         "Area_sqm" = "X..Area.6.") %>%
  mutate_at("County", str_replace, " County", "") %>%
```

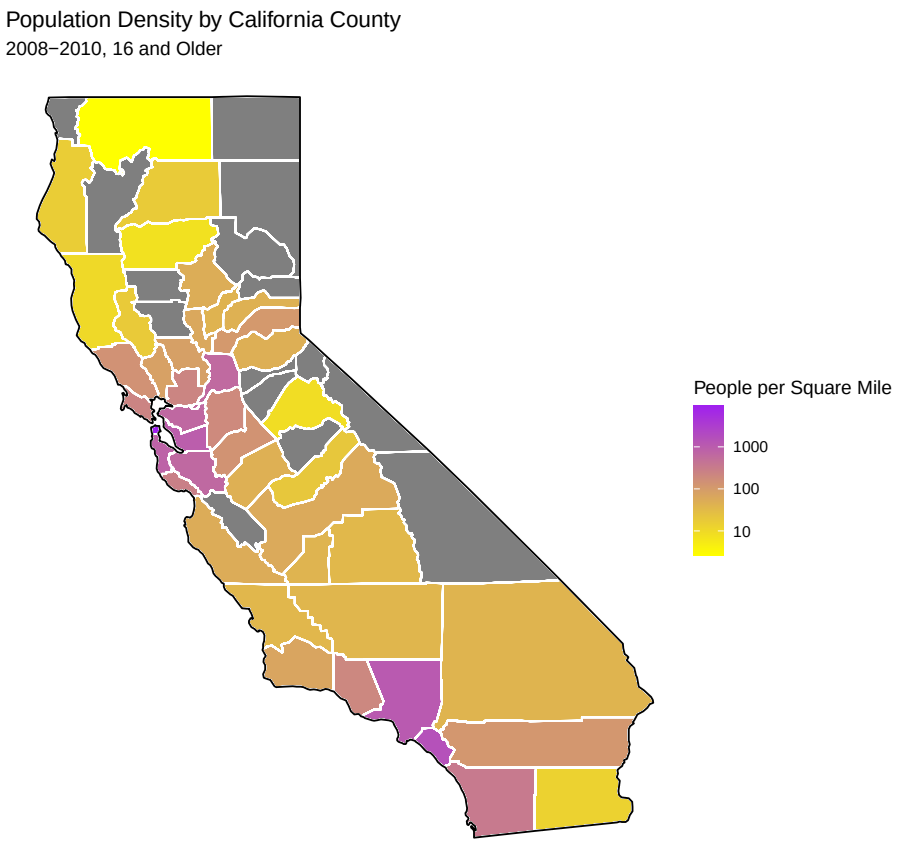
```
  mutate_at("Area_sqm", str_replace, ",", "") %>%
  mutate(Area_sqm = str_extract(Area_sqm, "^\\w+"))

areas$Area_sqm <- as.numeric(areas$Area_sqm)
head(areas)
```

```
df_counties <- left_join(df_counties, areas,
                         by = c("county_name" = "County"))
```

```
# color palette
palette <- colorRampPalette(c("yellow", "purple"))(20)

# adding data onto our CA base map
ca_base +
      geom_polygon(data = df_counties %>% filter(reportyear == "2008-2010",
                                                 geotype == "CO"),
                   aes(fill = pop_total / Area_sqm), color = "white") +
      geom_polygon(color = "black", fill = NA) +
      scale_fill_gradientn(colours = palette, trans = "log10",
                           name = "People per Square Mile") +
  ggtitle(label = "Population Density by California County",
          subtitle = "2008-2010, 16 and Older") +
  theme_void()
```



We can see that Southern California, particularly Los Angeles, as well as the Bay Area, are the most densely populated areas of the state. We'll keep this in mind as we now visualize car dependency for the most

populous cities in California.

City Level Analysis

```
# dataset of CA cities
df_cities <- df %>%
  filter(grepl('city', geoname) | geoname == "California",
         geotype == "CA" | geotype == "PL",
         race_eth_name == "Total")

df_cities <- df_cities %>%
      mutate_at("geoname", str_replace, " city", "")
```

```
# longitude and latitude of CA cities
ca_cities <- us.cities %>%
  filter(country.etc == "CA")
```

```
ca_cities <- ca_cities %>%
      mutate_at("name", str_replace, " CA", "")

ca_cities = subset(ca_cities, select = !(names(ca_cities) %in% c("country.etc", "pop", "capital")))

ca_cities = rbind(ca_cities, c("California", NA, NA))
ca_cities$lat <- as.numeric(ca_cities$lat)
ca_cities$long <- as.numeric(ca_cities$long)
```

```
head(ca_cities)
```

```
# joining
df_cities <- inner_join(df_cities, ca_cities,
                        by = c("geoname" = "name"))

rm(ca_cities)
```

Our new data set has information on the 193 most populous cities in CA. Let's visualize the percent of people that drive a car to work.

```
names(df_cities)
```

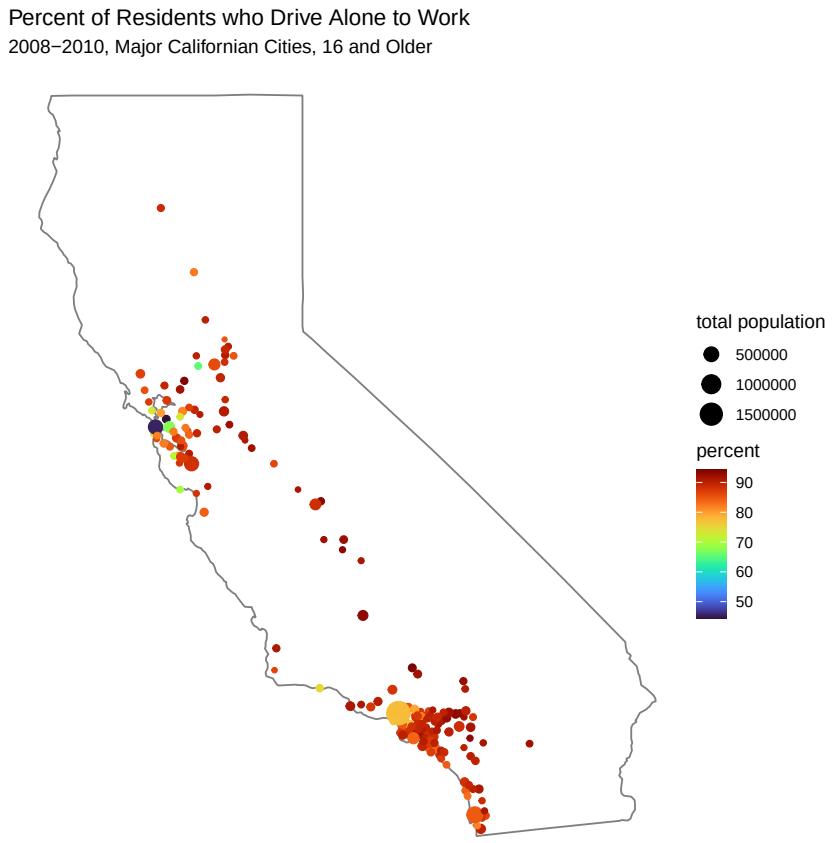
```
##  [1] "reportyear"     "race_eth_name"  "geotype"        "geoname"
##  [5] "county_name"    "county_fips"    "region_name"    "mode"
##  [9] "mode_name"      "pop_total"      "pop_mode"       "percent"
## [13] "LL95CI_percent" "UL95CI_percent" "percent_se"     "percent_rse"
## [17] "CA_RR"          "lat"            "long"
```

```
library(viridisLite)

ggplot(data = df_cities %>% filter(mode == "CARTOTAL",
                                   geotype == "PL",
                                   reportyear == "2008-2010"),
       aes(x = long, y = lat)) +
  borders("state", "California") +
  geom_point(aes(col = percent, size = pop_total)) +
  coord_fixed(1.3) +
  scale_color_gradientn(colors = viridis(20, option = "turbo")) +
  scale_size_continuous(name = "total population") +
```

```
  ggtitle(label = "Percent of Residents who Drive Alone to Work",
          subtitle = "2008-2010, Major Californian Cities, 16 and Older") +
  theme_void()
```

```
## Warning: Removed 17 rows containing missing values (geom_point).
```

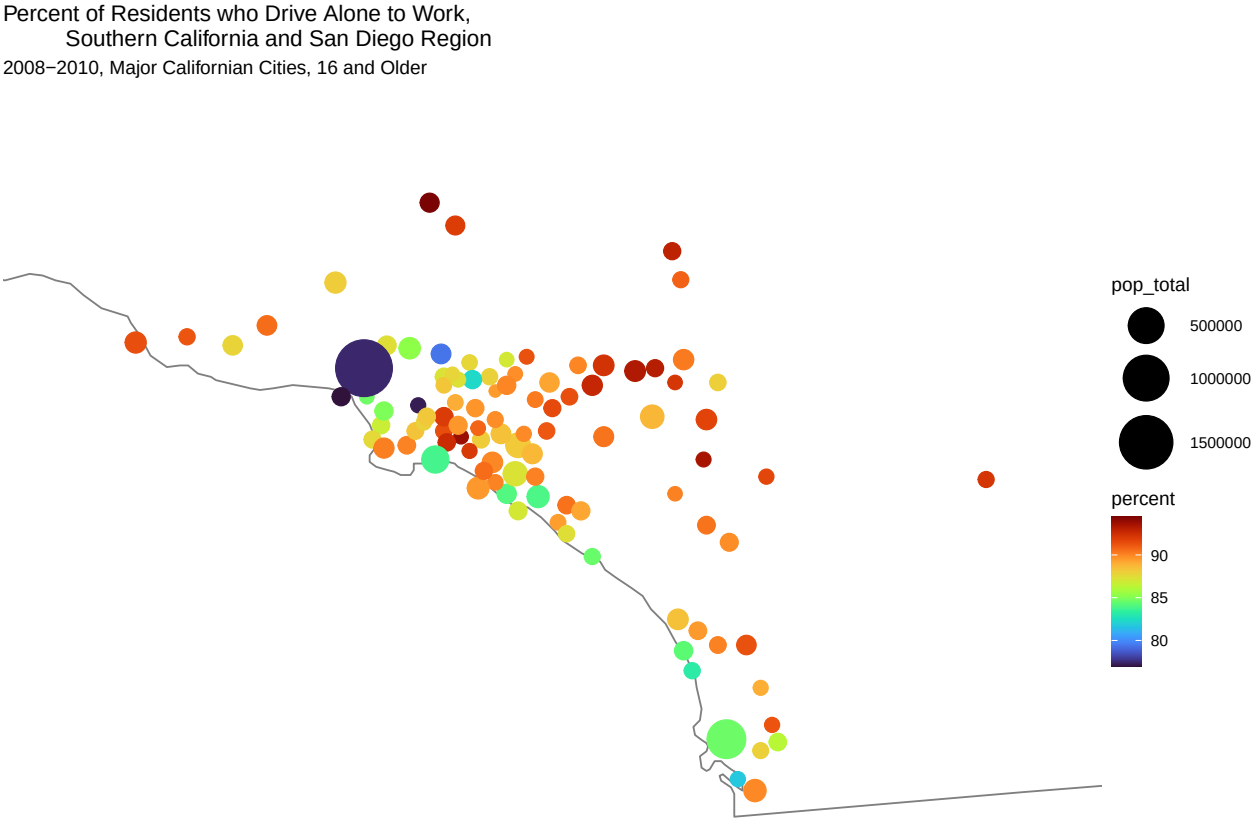


The Bay Area, specifically San Francisco, has the lowest percentage of Californians that drive to work. Los Angeles is also not as car dependent as the rest of Southern California. Let's zoom in to Southern California so that the area is more visible.

```
ggplot(data = df_cities %>% filter(region_name == "Southern California" |
                                     region_name == "San Diego",
                                   mode == "CARTOTAL",
                                   geotype == "PL",
                                   reportyear == "2008-2010"),
       aes(x = long, y = lat)) +
  borders("state", "California") +
  geom_point(aes(col = percent, size = pop_total)) +
  scale_color_gradientn(colors = viridis(20, option = "turbo")) +
  coord_fixed(xlim = c(-119.5  , -116.0),  ylim = c(32.5, 35)) +
  scale_size_continuous(trans = "log10", name = "total population") +
  scale_size(range = c(3, 15)) +
  ggtitle(label = "Percent of Residents who Drive Alone to Work,
          Southern California and San Diego Region",
          subtitle = "2008-2010, Major Californian Cities, 16 and Older") +
  theme_void()
```

```
## Scale for 'size' is already present. Adding another scale for 'size', which
## will replace the existing scale.
```

```
## Warning: Removed 13 rows containing missing values (geom_point).
```



It seems that Californians are very dependent on cars for commuting. Los Angeles, the largest circle, is the least dependent city. It also appears that most of the coastal cities don't commute as much with cars to work. We can find the percentage of residents that drive to work:

```
df %>%
  filter(reportyear == "2008-2010",
         race_eth_name == "Total",
         geoname == "California",
         mode == "CARTOTAL") %>%
  select(reportyear, percent)
```

84.6% of Californians drive for their commute. Now, let's see which California cities are the most and least dependent on cars.

A Closer Look

```
df_cities %>%
    filter(mode == "CARTOTAL",
           geotype == "PL",
           reportyear == "2008-2010") %>%
  select(geoname, region_name, mode, pop_total, percent) %>%
```

```
  arrange(desc(percent)) %>%
  head(10)
```

Lancaster has the highest proportion of residents aged 16 and older that commute to work. And nine out of the ten most car dependent cities are in the Southern California or San Joaquin Valley regions.

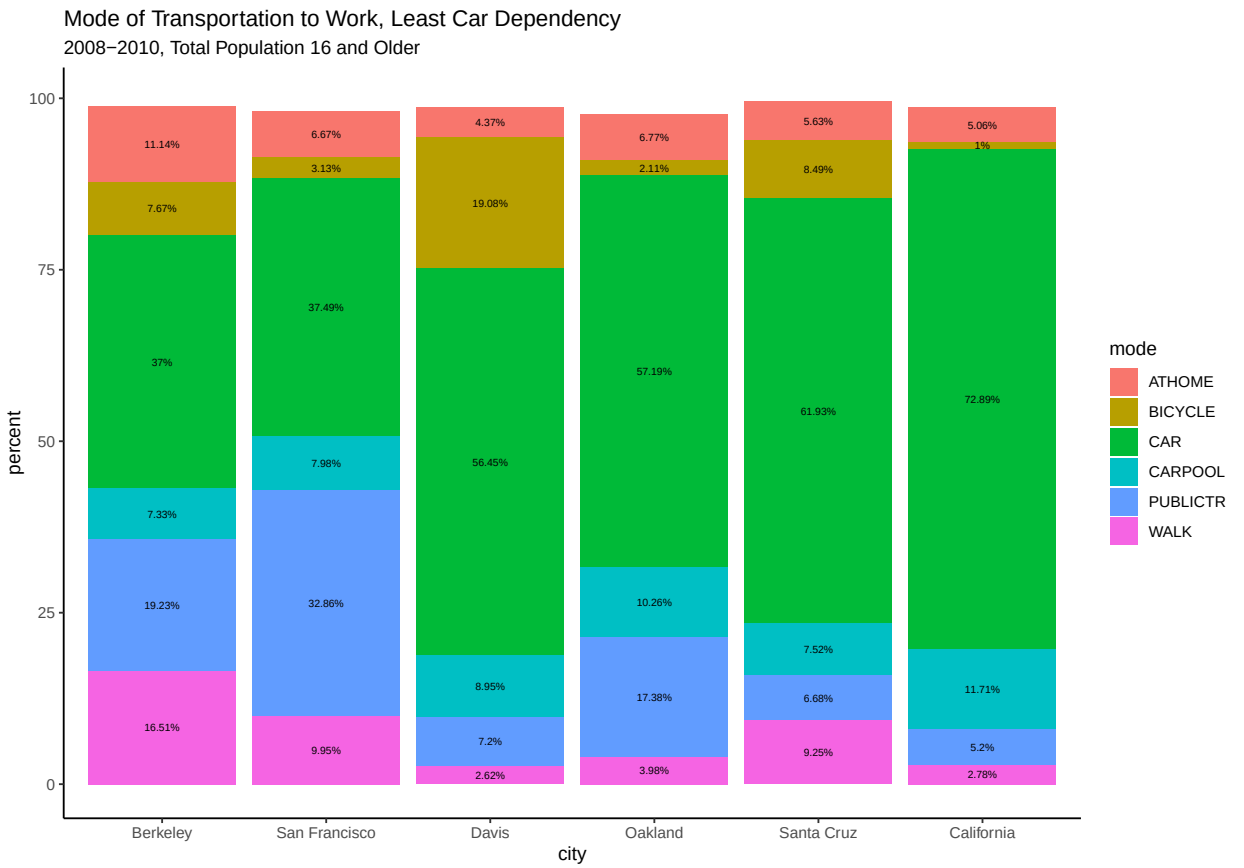
```
df_cities %>%
  filter(mode == "CARTOTAL",
         geotype == "PL",
         reportyear == "2008-2010") %>%
  select(geoname, region_name, mode, pop_total, percent) %>%
  arrange(percent) %>%
  head(10)
```

Berkeley is the least car dependent city in California, with only 44.3% of its adult residents driving to work. A majority of these cities are also in the Bay Area.

```
# cities with the greatest percentage increase
least_df <- df_cities %>%
         filter(geoname %in% c("Berkeley", "San Francisco", "Davis",
                               "Oakland", "Santa Cruz", "California"),
                (mode != "CARTOTAL"),
                reportyear == "2008-2010")

least_df$geoname = factor(least_df$geoname,
                          levels = c("Berkeley", "San Francisco", "Davis",
                                     "Oakland", "Santa Cruz", "California"))

ggplot(data = least_df,
       aes(x = geoname, y = percent, fill = mode, order = percent)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(percent, 2), "%")),
            position = position_stack(vjust = 0.5), size = 2)  +
  xlab("city") +
  ggtitle(label = "Mode of Transportation to Work, Least Car Dependency", subtitle = "2008-2010, Total
  theme_classic()
```



A majority of commuters in Berkeley and San Francisco take public transportation or carpool to work, far above the state-wide averages. It's also interesting that 19% of commuters in Davis bike to work, then the national average is only 1%!

Probably and equally as intriguing question, however, is which cities have improved their environmentally friendly commuting habits? We can keep track of the cities where the percentage of residents aged 16 and older that bike, walk, take public transportation, or work from home have increased the most from 2000 to 2010.

```
percent_2000 <- df_cities %>%
  filter(race_eth_name == "Total",
         mode == "CARTOTAL",
         reportyear == "2000") %>%
  select(geoname, region_name, percent) %>%
  rename("percent_2000" = "percent")

percent_2010 <- df_cities %>%
  filter(race_eth_name == "Total",
         mode == "CARTOTAL",
         reportyear == "2006-2010") %>%
  select(geoname, percent) %>%
  rename("percent_2010" = "percent")

percent_df <- inner_join(percent_2000, percent_2010, by = "geoname")

percent_df <- percent_df %>%
  mutate(percent_2000 = 100 - percent_2000) %>%
```

```
  mutate(percent_2010 = 100 - percent_2010) %>%
  mutate(percent_change = percent_2010 - percent_2000)
```

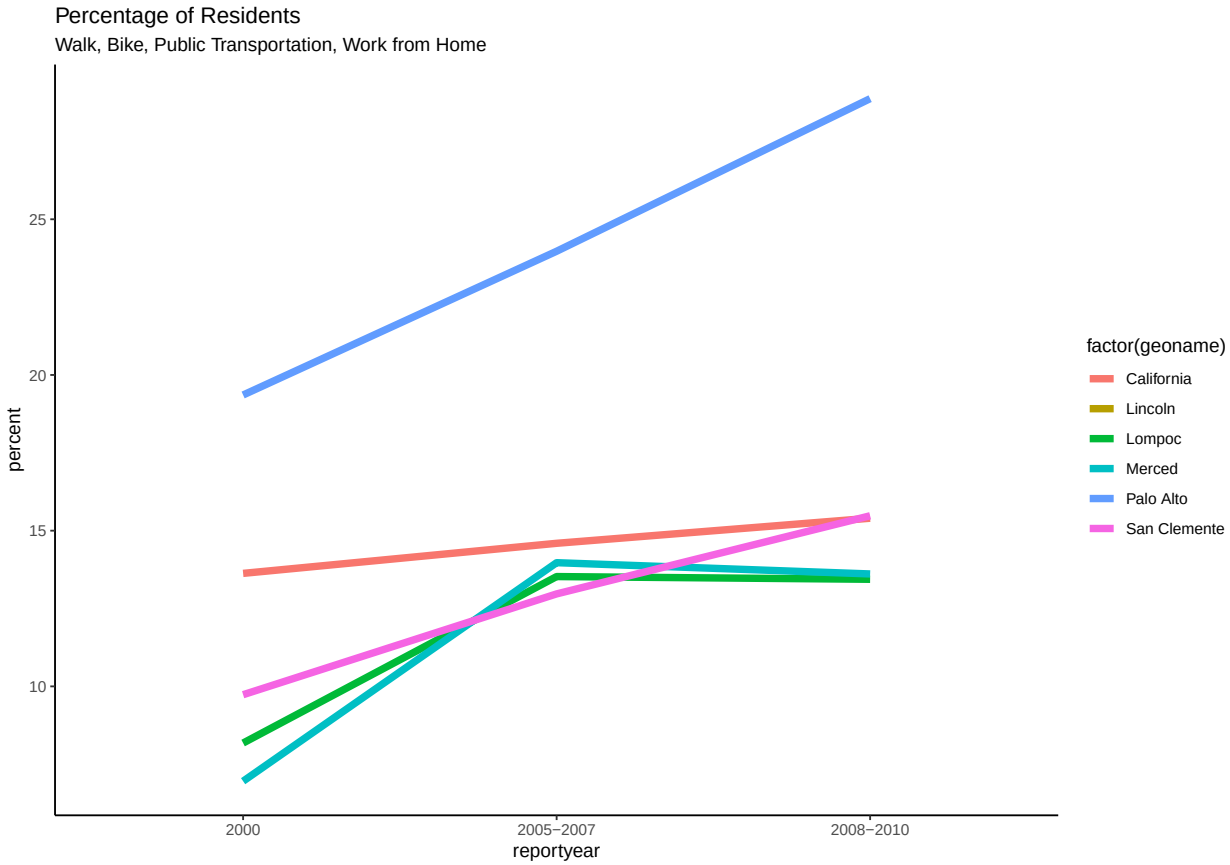
```
percent_df %>%
  arrange(desc(percent_change)) %>%
  head(10)
```

Merced tops our list, with an almost 7% increase in residents ditching cars for work from 2000 to 2010 (In 2005, a new UC campus was built in Merced: could the emergence of a college scene have spurred this?)

We can visualize these trends over time as well.

```
# cities with the greatest percentage increase
top <- c(percent_df[percent_df$percent_change > 5.7, ]$geoname, "California")

ggplot(data = df_cities %>%
         filter(geoname %in% top,
                race_eth_name == "Total",
                mode == "CARTOTAL",
                reportyear != "2006-2010"),
  aes(x = reportyear, y = 100 - percent, group = geoname, color = factor(geoname))) +
  geom_line(size = 2) +
  ylab("percent") +
  ggtitle(label = "Percentage of Residents", subtitle = "Walk, Bike, Public Transportation, Work from H
  theme_classic()
```



Overall, the percent of Californians taking alternative modes of transportation to work had steadily increased

from 2000 to 2010. Merced had the steepest growth, but was already a relatively car-free city. The other cities with the fastest transition interestingly all originally had a percentage lower than the state average.

Wrapping Up

This project was very interesting. I got a lot of practice working with dplyr, graphing locations, organizing data sets, and so much more.

It's important to note that the latest data in this data set is from the 2010 census, which was 12 years ago. It would be fun to find more recent data to see if there are any significant changes (gas prices, for example, are a record high: how does that affect the average commute?) Our data also focuses only on commutes for Californians with paid work aged 16 and older. The data excludes driving for school, errands, vacations, and more.

Our analysis of the 200 most populated Californian cities helps shed some light on how car dependent our state is. Many cities don't a great public bus system. Others have so few sidewalks that walking or biking to work is just impossible. Still, cities have been improving accessibility to more environmentally friendly and economical modes of transportation.

This data set was accessed from the California Health and Human Services Open Data Website. They provided amazing data dictionaries, narrative examples, research sources, and so much more.

Eric C. Anderson's github repository on making maps in R was an incredible source. I used this reference for all of my ggplot maps.

I used this Wikipedia article on Californian counties for more great background information and its tables.