1. Describe the dataset. (ex: Information about Traffic violations in Montgomery County, MD; or The number of microaneurysms found in a patient's eye and whether or not they have diabetic retinopathy.)

2. How many records does the dataset have?

3. How many features does the dataset have? List or describe a few of them.

4. What can you try to predict in this dataset? (ex: We can try using the features, including age, race, gender, car make and model, etc, to predict the type of traffic violation; or We can use the number of microaneurysms measured in the patient's eye to predict whether or not they have diabetic retinopathy.)

5. Is this a **labeled** dataset, appropriate for a supervised learning classification problem? (In other words, if you are trying to predict whether or not someone has a disease, does your dataset contain whether or not each record has the disease?)

6. Provide a link to the dataset, if there is one. If you are getting your data from somewhere other than a link, where are you getting it from?

1. There are two datasets- the smart locations database and Census tract data. Both datasets are for the entire U.S.
2. The dataset has about 200,000 records. Some of the records do not have data for specific features, each record is a census tract (a piece of land).
3. The data set has over 29 features. The features include housing units per acre, people per acre, jobs per acre, jobs per household, and land use diversity. Some of the features use raw data and some are scaled.
4. We can try using the features, which include urban design (street intersection density, land use diversity, etc.), combined with demographic information(working-age population within a 45 minute transit ride, zero car households percentage, etc.) to predict the percent of people who drove alone to work.
5. The dataset will include the percentage of people that drove alone to work for each region/record. This will be our class to predict. As it isn’t a label used in supervised learning classification models, we will use linear regression instead.
6. SocialExplorer (https://www.socialexplorer.com/tables/ACS2012_5yr/R12068878) provides the modeshare, while the EPA provides the smart locations database https://edg.epa.gov/data/PUBLIC/OP/SLD.

Note: We would need to reduce certain features or focus the geographic scope if it is too computationally expensive to analyze.

Search for block groups (uses 11 digits)

Use American Community Survey (ACS) instead of Census data with smart location database