# Part 1: Data Exploration and Evaluation

*Create an exploratory data analysis project. Load the data and perform any necessary cleaning and aggregations to explore and better understand the dataset. Based on your exploration, please describe your high level findings in a few sentences. Please include two data visualizations and two summary statistics to support these findings.*

Using Tableau, I found two things interesting:

1. By creating a symbol view of Home Ownership type and Income level it validated assumptions I had made. States like NY, CA and TX had larger income levels as well as a higher number of individuals renting and having a mortgage than owning a home. Figure 1 is the data visualization of noted observations.

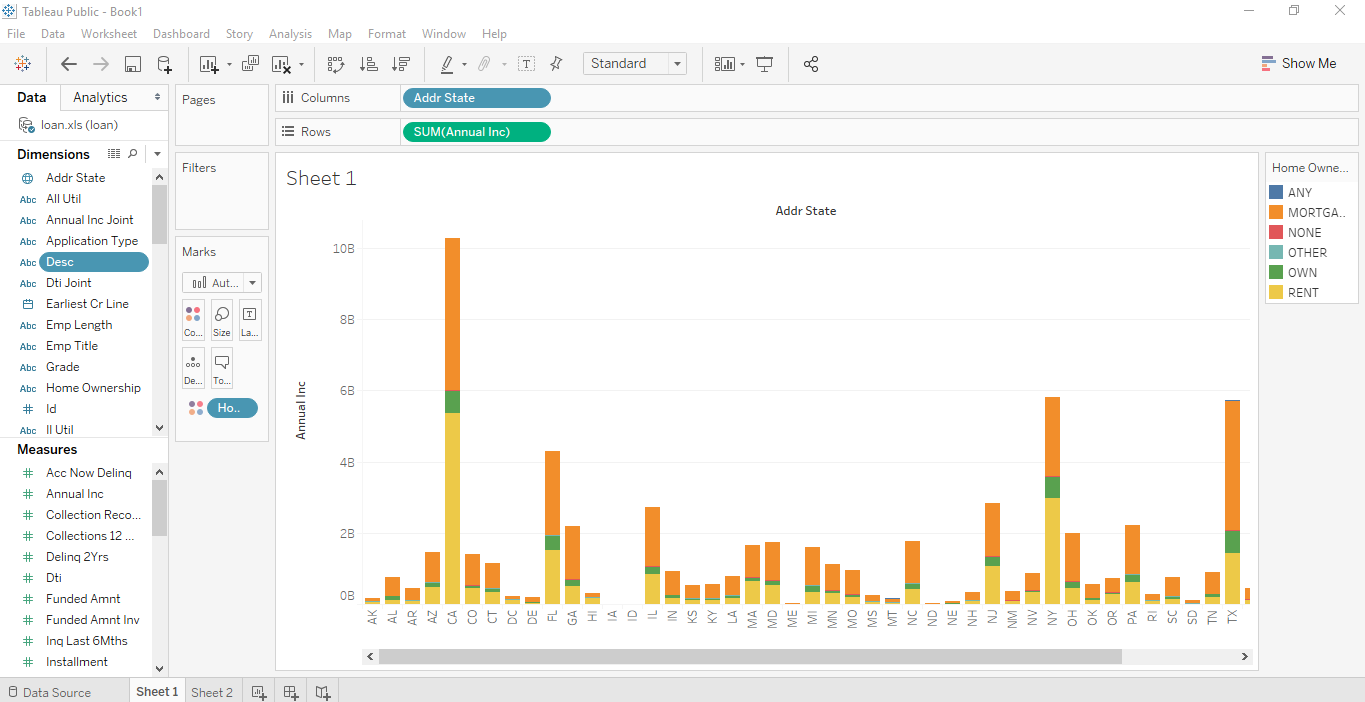
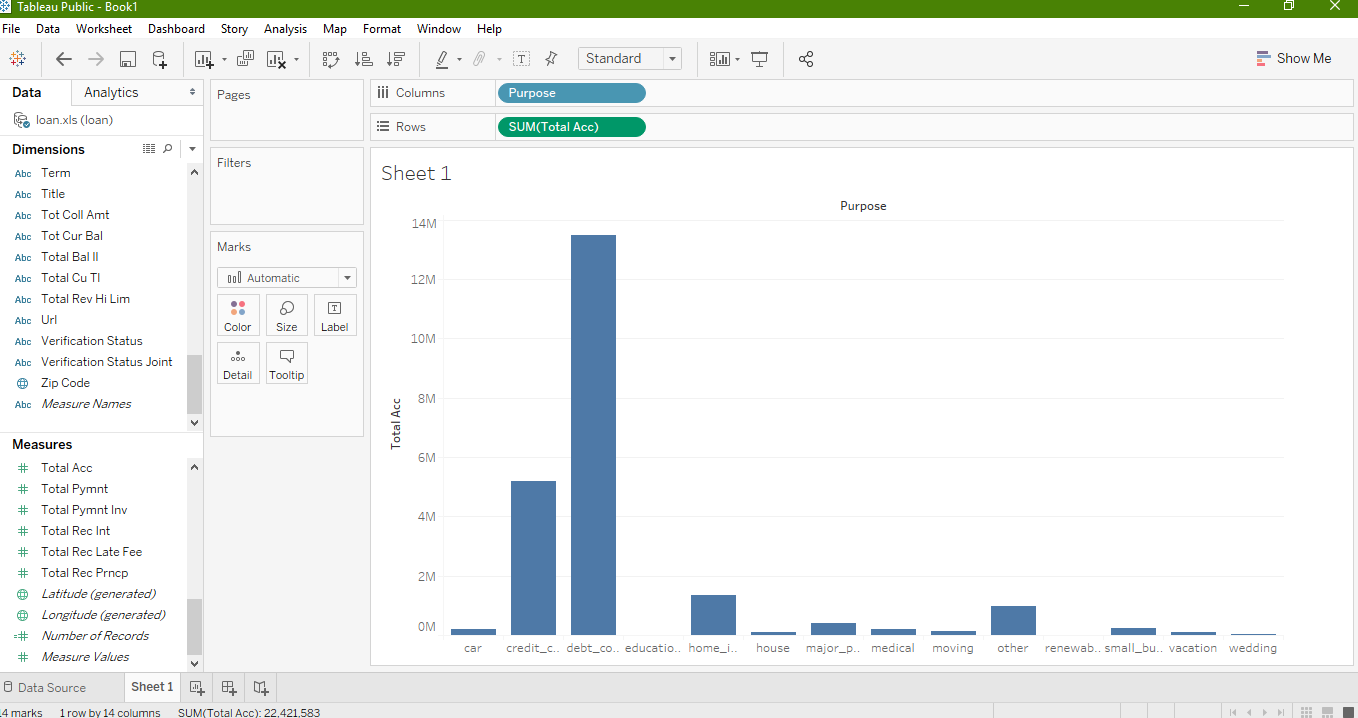


Figure 1: Relationship between Total Income, Ownership Type and Region

1. Another thing I noted about the data was the relationship loan type and number of individuals. As shown in Figure 2, individuals taking loans for debt consolidation was the highest followed by credit card payments.



# Figure 2: Purpose for loans and Total Account

# Part 2: Data Pipeline Engineering

*Please build a prototype of a production data pipeline that will feed an analysis system (data warehouse) based on this dataset. This system will allow data scientists and data analysts to interactively query and explore the data, and will also be used for machine learning model training and evaluation. Assume that the system will receive periodic updates of this dataset over time, and that these updates will need to be processed in a robust, efficient way.*

Steps taken to complete this:

* Using Xampp, I created a local instance of a mysql database **data\_analysis\_db**
* Database scripts can be seen in **create\_schema.sql**
  + Schema is meant to represent a subset of models seen in the data set. Tables like members, loans, payments and transactions would be a starting point to building a data warehouse where all data seen in the loans dataset would be stored.
* Python script **python\_funcs.py** written in Notepad++ editor with dependencies on pymysql and pandas.
  + Script contains sample functions that can be run to query or update the data set.
    - get\_overdue\_payments()
    - update\_int\_rate\_amnt(type, amount)
* Summary: Tables described are base tables that can be used in the data warehouse. Keeping the relational nature of a data warehouse, data used for reporting can be stored and analyzed here. Metrics such as payments, regions, transactions, late accounts could be measured using helper functions like those listed in python\_funcs.py. Additional features this data warehouse would need can be connectors to data feeds such as Salesforce and connectors to visualization/reporting/analysis tools.