U.S. CENSUS TRANSPORTATION DATA

Years: 2012 - 2016

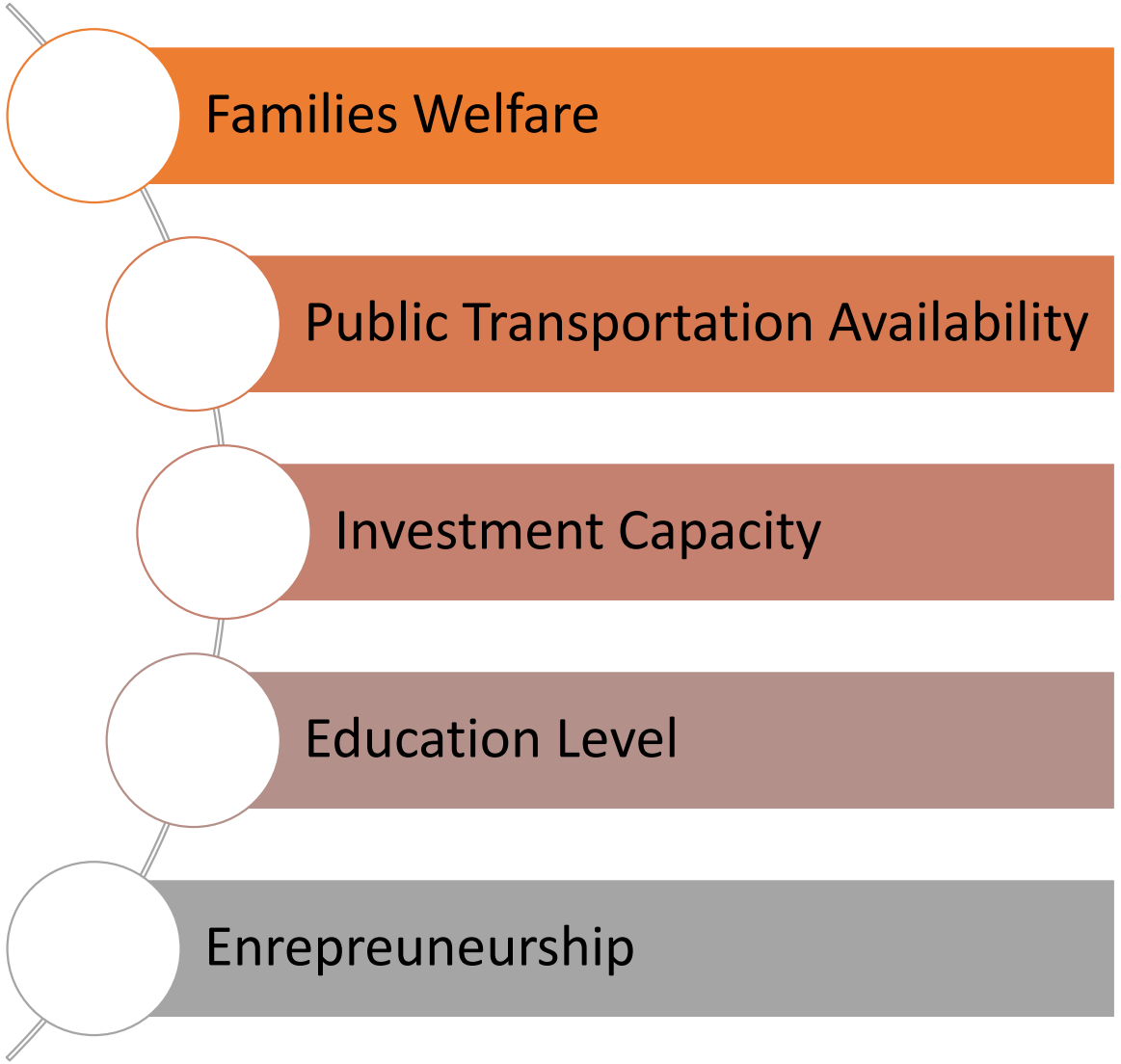
Samira Karimi
Apr, 2020

DATA EXPLANATION

THE **CTPP DATA PRODUCT BASED ON 2012 – 2016 5-YEAR AMERICAN COMMUNITY SURVEY (ACS) DATA** IS DESIGNED TO HELP TRANSPORTATION ANALYSTS AND PLANNERS UNDERSTAND WHERE PEOPLE ARE COMMUTING TO AND FROM, AND HOW THEY GET THERE. THE INFORMATION IS ORGANIZED BY RESIDENCE, WORKPLACE, AND BY THE COMMUTE FROM HOME TO WORK.

PROJECT APPLICATIONS

- Transportation policy and planning efforts.
- Socioeconomic factors
- Recognizing capacities
- Recognizing needs



DATA PREPROCESSING

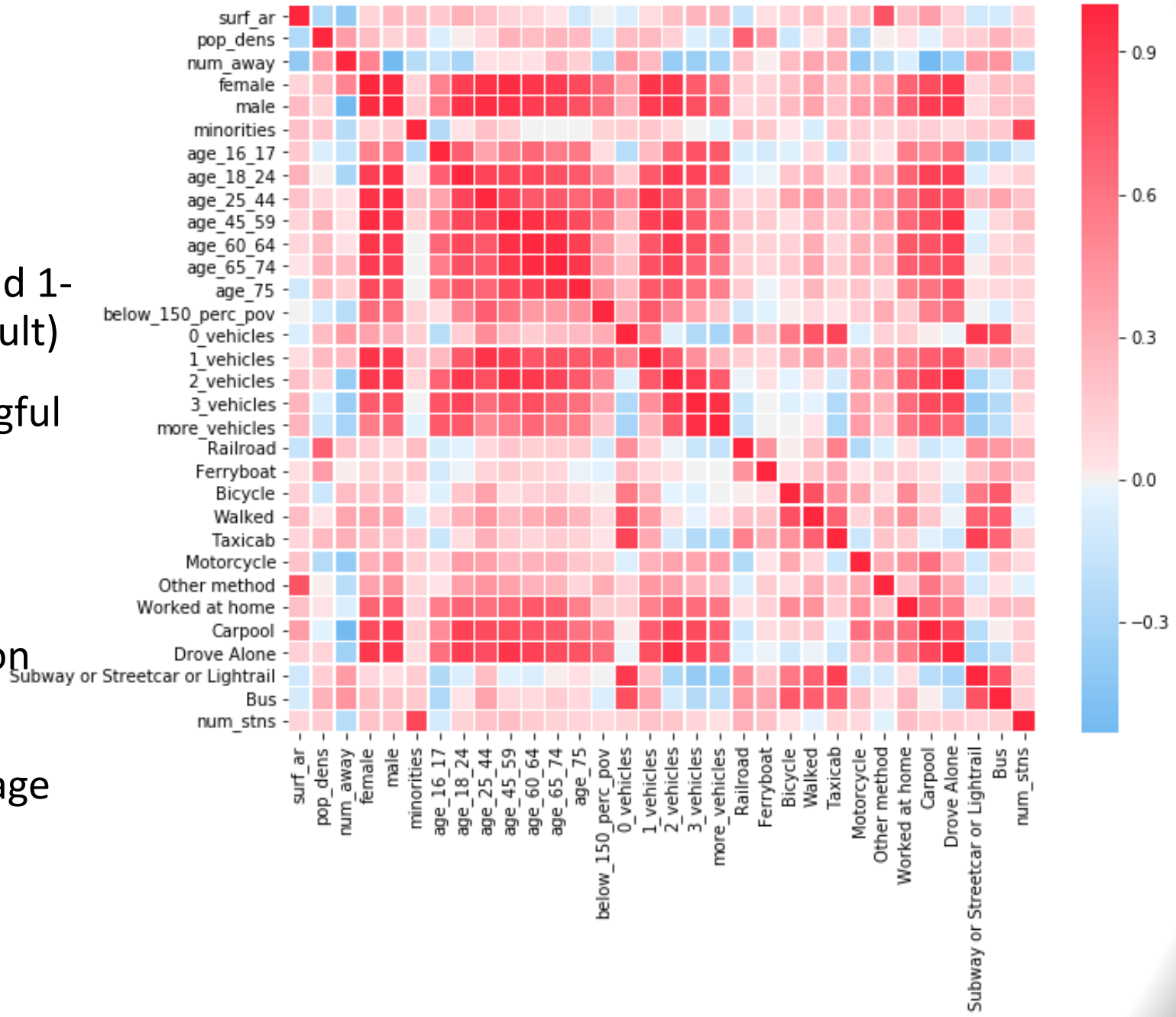
- Data not aimed for one specific task like classification or regression
- No certain response variable
- Required to set a goal for data preprocessing
- Acquired and merged data from many tables
- Initial downloaded data size: 114 Mg

ANALYSIS

- Many columns for a small number of states and districts (52)
- need to explore the relationships between the columns and find the effective ones
- Every factor can be analyzed separately
- As examples, a number of initial guesses has been explored.

ANALYSIS: CORRELATION HEATMAP

- High correlation of public transportation with 0-vehicle and 1-vehicle households (obvious result)
- The effect of age not as meaningful
- Manipulation of the columns
- Making more general columns
- Example: all public transportation means in one column,
- Dividing ages to young, middle age and over 65



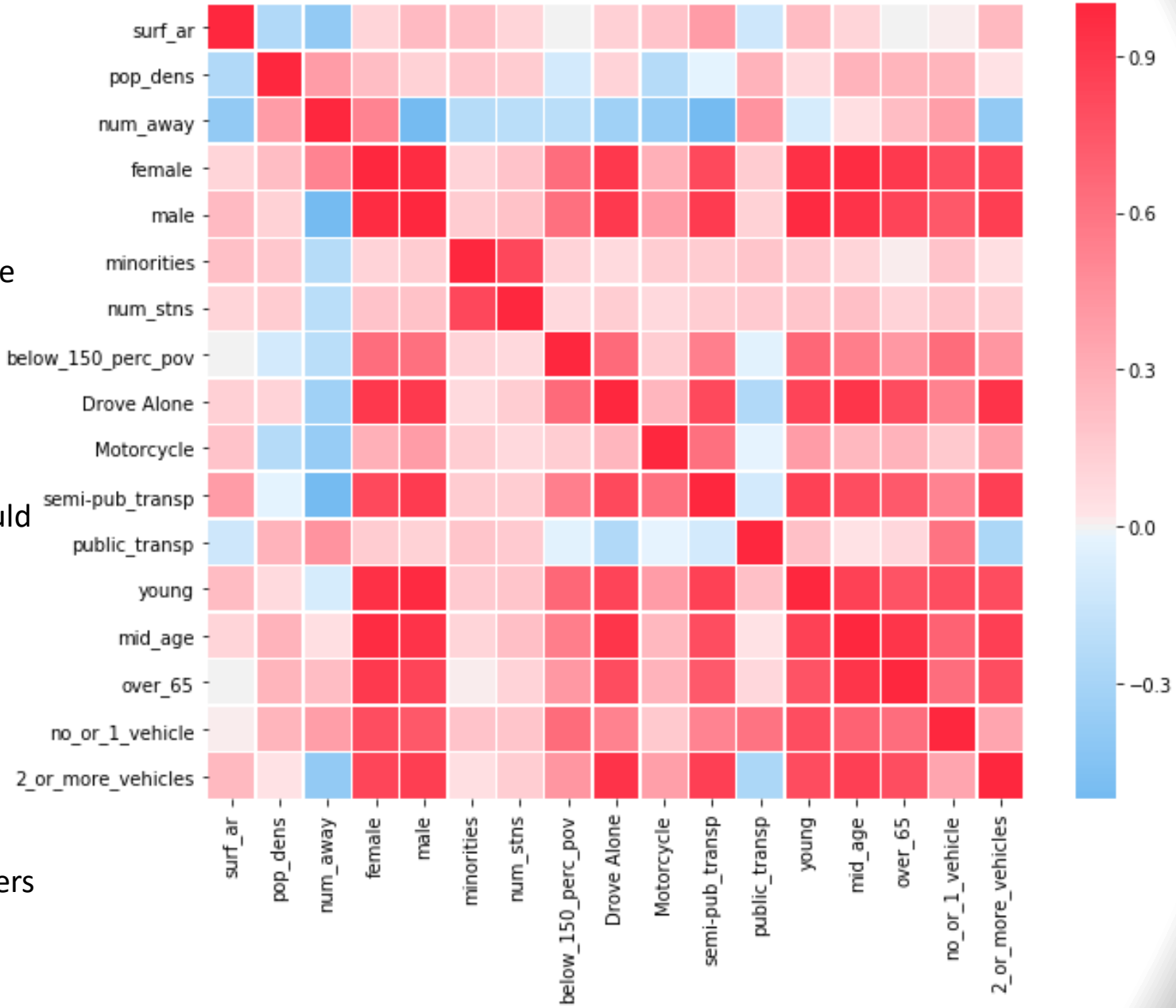
ANALYSIS: CORRELATION HEATMAP

1. The number of vehicles in a household: 2 or more vehicles , a high positive correlation with driving alone and semi-public transportation , negative correlation with public transportation, One or no vehicles highly and positively correlated with public transportation

2. percentage of minorities: high positive correlation with the number of amtrak stations ('num_stns'). Could be due to the culture and pupulation combination of many northern and west-side states

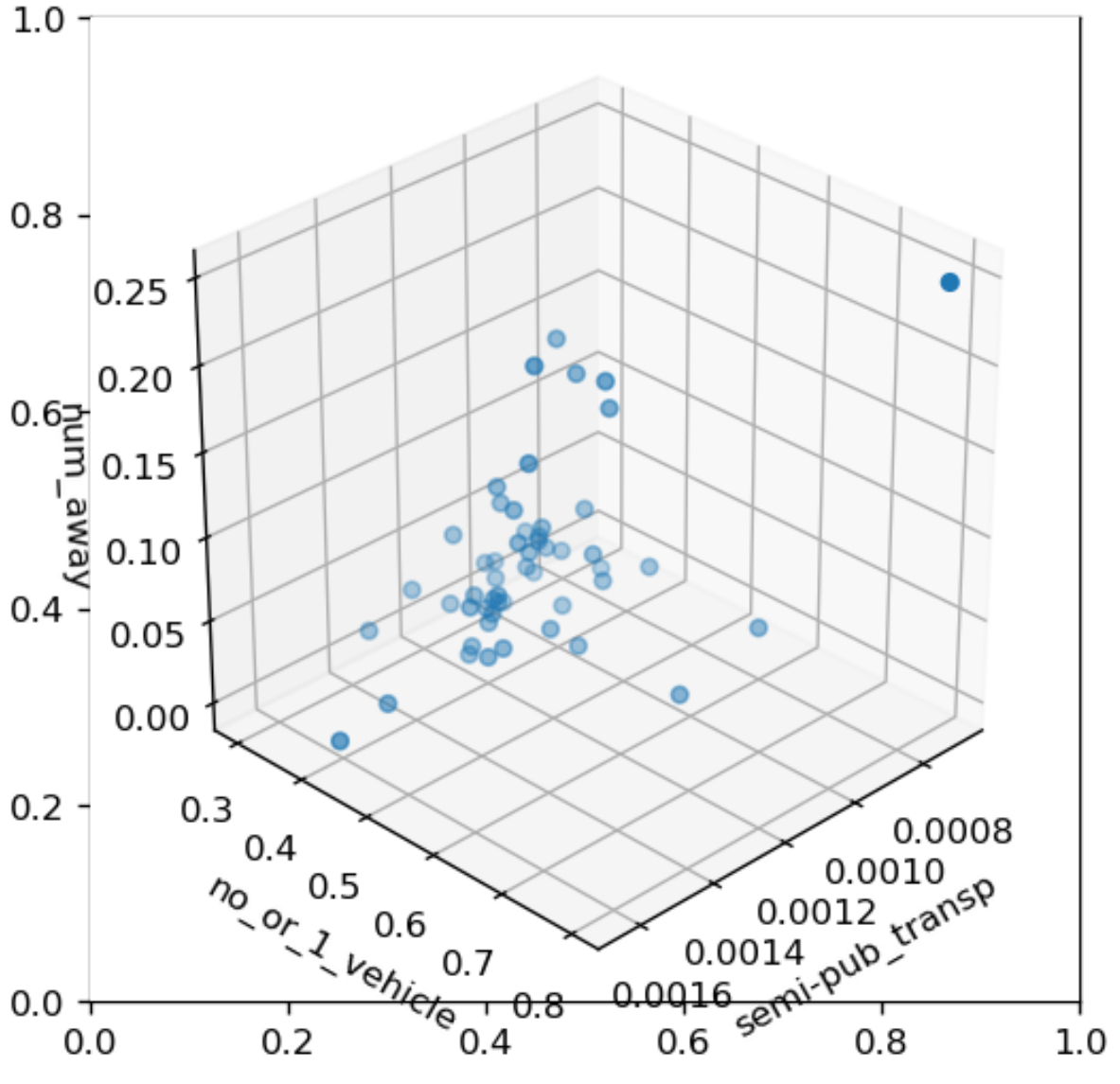
3. The number of Amtrak stations:
No positive correlation with the percent of away workers. A better metric:
number_of_stations/surface_area

4. Surface area: Negative correlation with away workers and public transportation (expected)



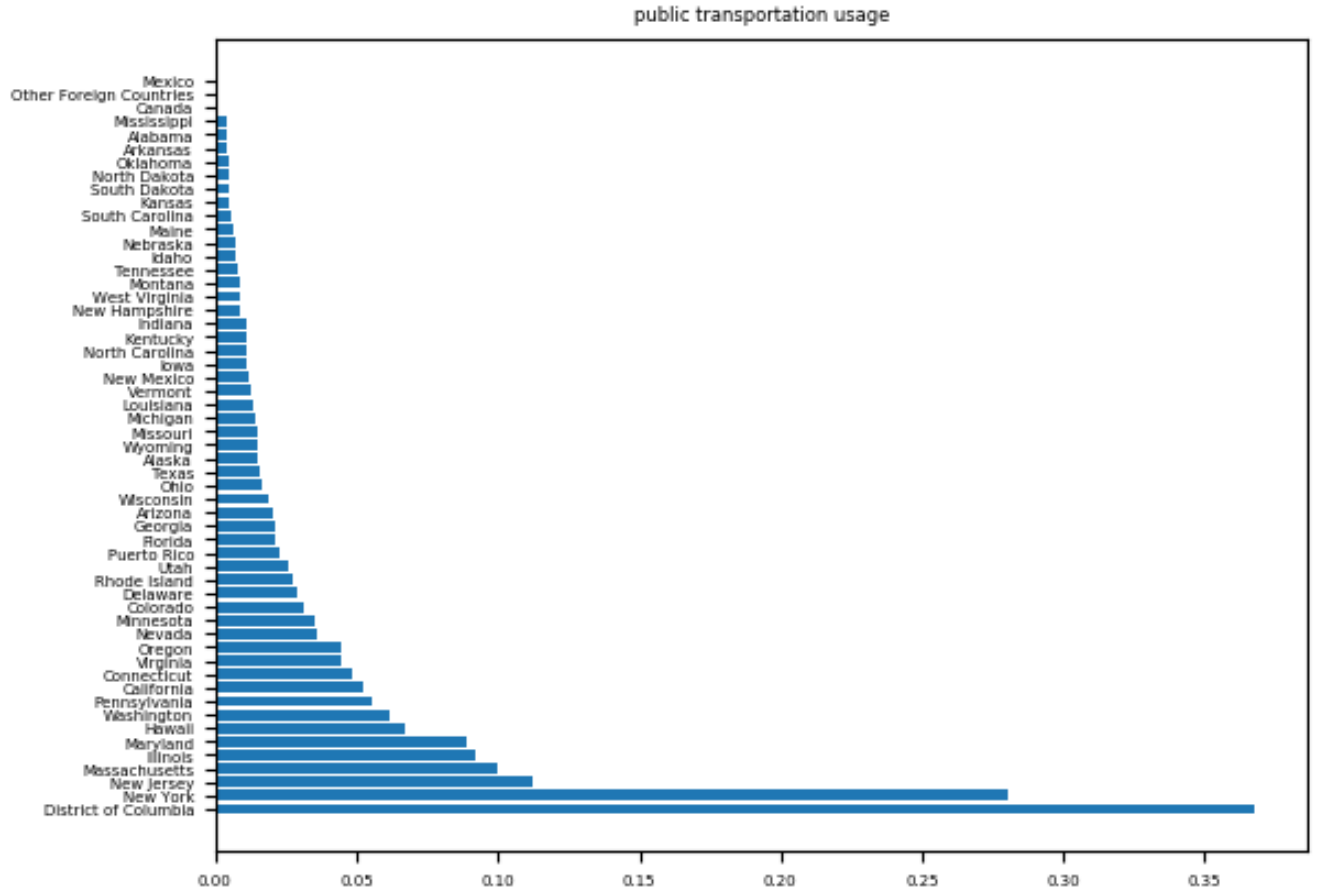
ANALYSIS: BAR CHART

- A somewhat linear relationship between the number of away workers with semi-public transportation users



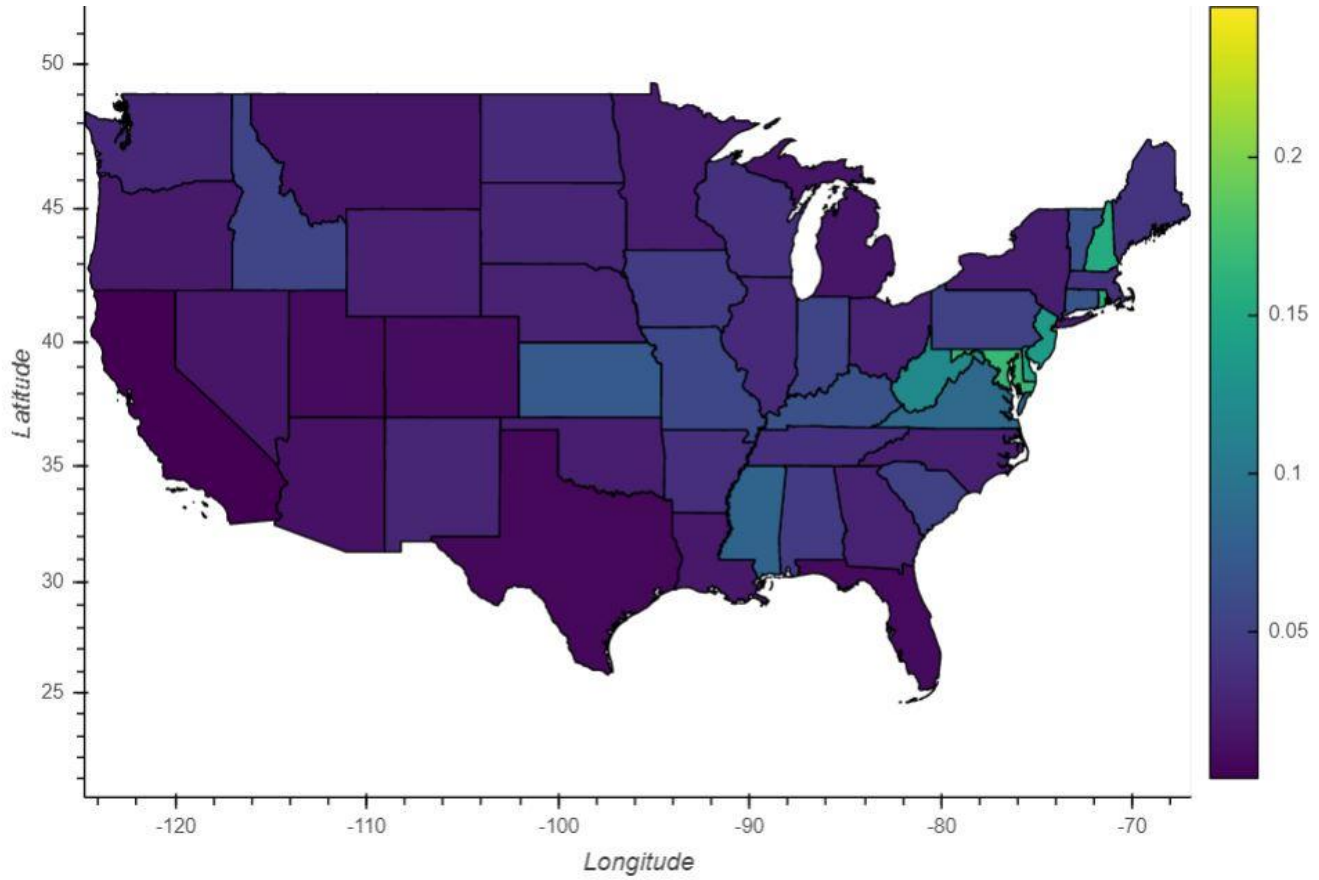
ANALYSIS: BARCHART

- Considerably higher usage of public transportation in New York and District of Columbia



ANALYSIS: HEATMAP

- Texas and : lowest percentage of away workers
- Some north-eastern states: highest percentage of away workers (could be due to the low surface area)

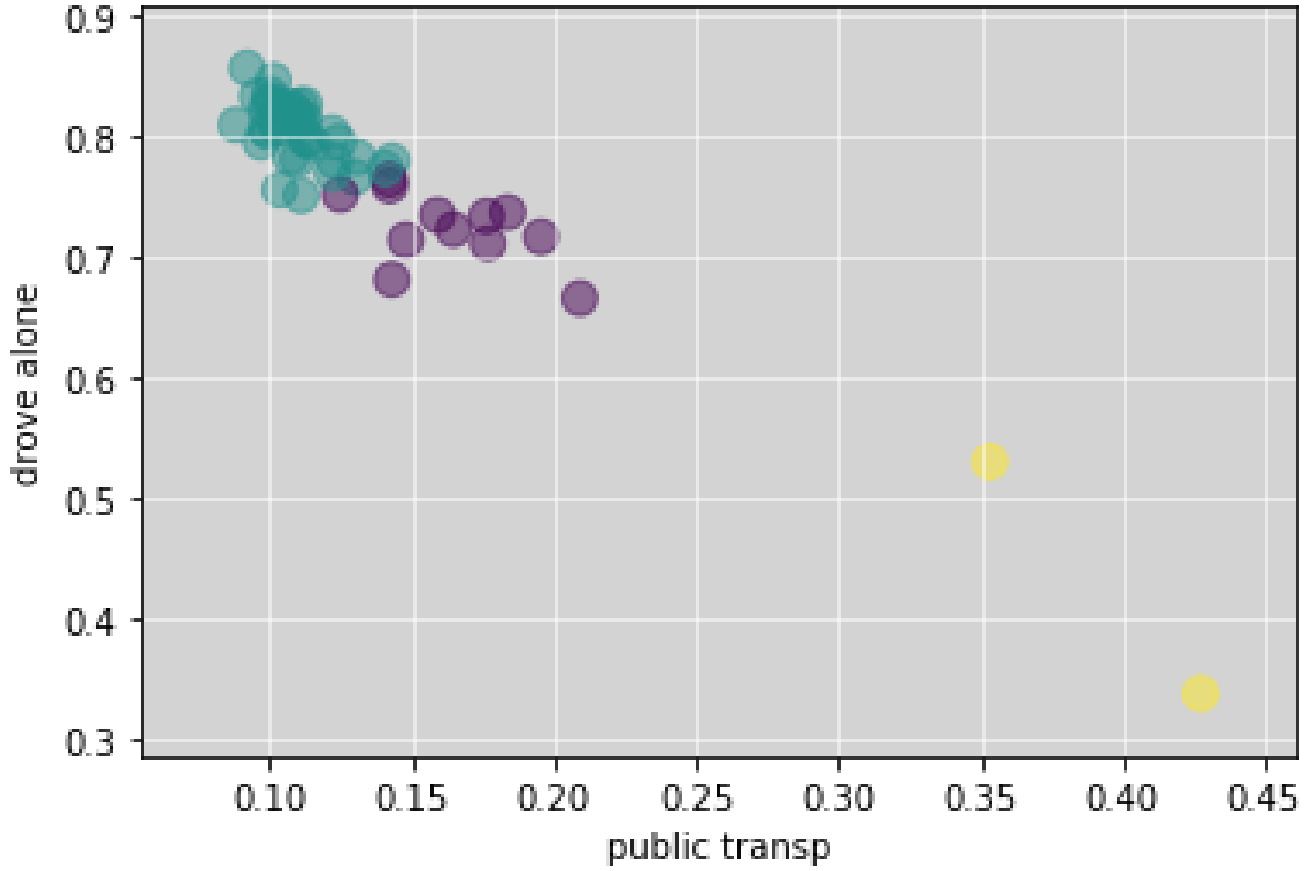


FORECAST

- Clustering (Healthy Transportation)
- Ridge Regression
- Lasso Regression
- Different Response Variables

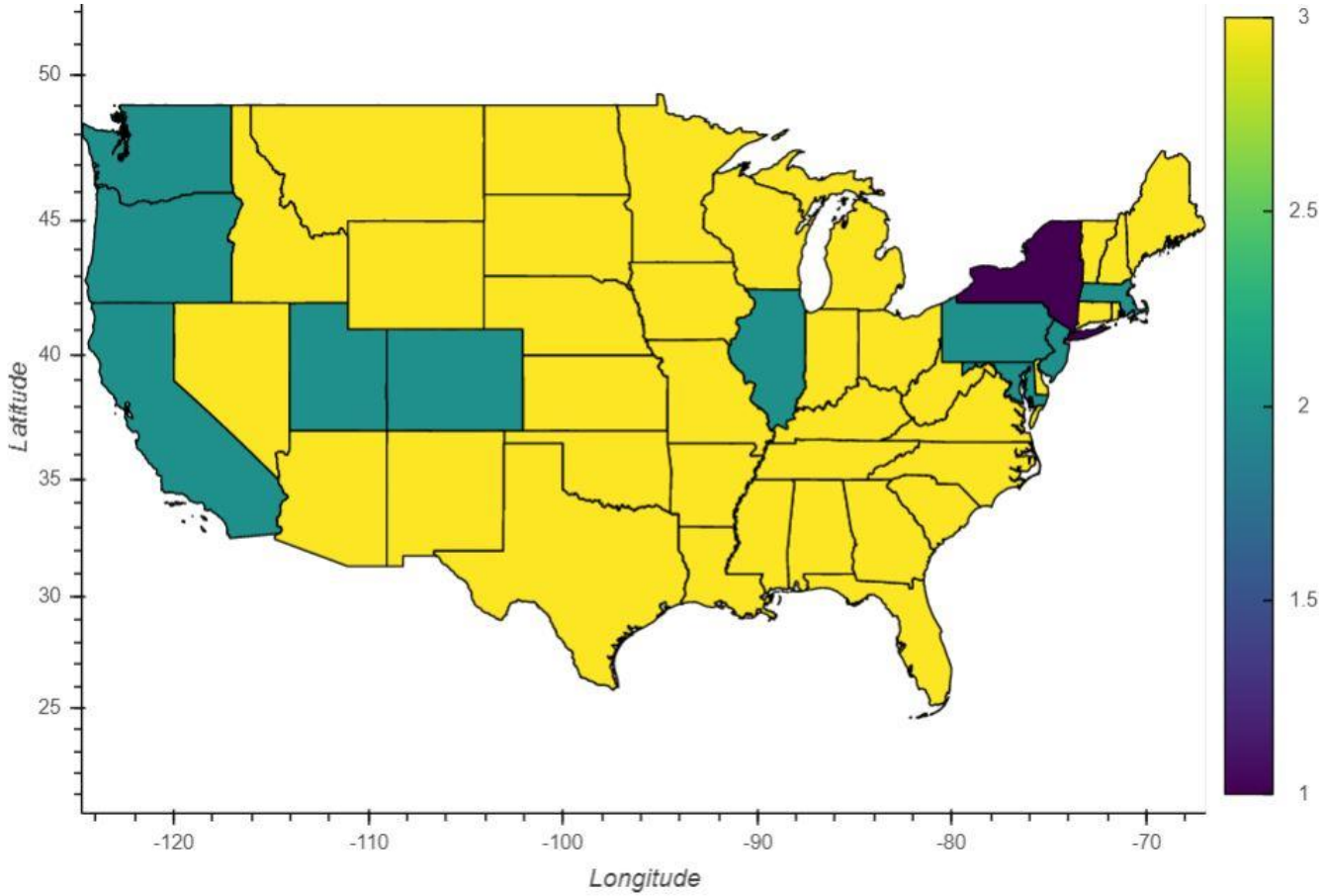
CLUSTERING: PUBLIC TRANSPORTATION USAGE

- The two yellow dots 'New York' and 'District of Columbia‘: Remarkably healthier transportation method than all other states



CLUSTERING: HEATMAP

- The two yellow dots 'New York' and 'District of Columbia‘: Remarkably healthier transportation method than all other states



REGRESSION: PUBLIC TRANSPORTATION

- Independent var's: 'surf_ar','pop_dens','young','minorities','num_stns','no_or_1_vehicle','num_away','semi-pub_transp'
- Lasso: Public transportation usage can be explained through a linear regression based on the number of away workers.

MSE: 0.0071, Coef = [-0., 0. , 0. , 0. , -0. , 0. , 0.183, -0.]

- Ridge: Also gives importance to having no or one vehicle and semi-public transportation.

MSE: 0.0065, Coef = [-6.63e-09, 4.23e-06, 6.85e-03, 7.40e-11, -3.95e-05, 2.14e-02, 6.44e-02, -1.62e-02]

REGRESSION: NUMBER OF AWAY WORKERS

- Independent var's: 'surf_ar','pop_dens','young','minorities','num_stns','no_or_1_vehicle','public_transp','semi-pub_transp'
- Highly correlated data; Ridge seems a better choice
- Usefulness of Lasso variable selection
- Lasso: number of away workers explained mostly by public transportation.

MSE: 0.0040, Coef = [-2.53e-08, 3.78e-05, 0.0, -3.04e-10, -0.0, 1.47e-03, 4.48e-01, -0.0]

- Ridge: Also gives importance to semi-public transportation. Reason: Correlation

MSE: 0.0022, Coef = [-5.16e-08, 3.80e-05, 2.64e-02, -3.25e-09, -2.45e-04, 5.47e-02, 2.38e-01, -1.28e-01]