Data Analysis

Feature selection:

Lasso and random forest

Correlation: person, find some have strong positive correlation.

|  |
| --- |
| the population and GVA pearsonr correlation coefficient is -0.236652, p-value is 0.032308  the population and best\_Uni\_ranking pearsonr correlation coefficient is 0.573248, p-value is 0.000000  the population and hospitals\_number pearsonr correlation coefficient is 0.744543, p-value is 0.000000  the population and house\_price pearsonr correlation coefficient is -0.264770, p-value is 0.016223  the population and population pearsonr correlation coefficient is 1.000000, p-value is 0.000000  the population and pubs\_number pearsonr correlation coefficient is 0.541882, p-value is 0.000000  the population and road\_traffic\_2015 pearsonr correlation coefficient is 0.523504, p-value is 0.000000  the population and road\_traffic\_2016 pearsonr correlation coefficient is 0.513857, p-value is 0.000001  the population and school\_number pearsonr correlation coefficient is 0.901054, p-value is 0.000000  the population and stations\_number pearsonr correlation coefficient is 0.684171, p-value is 0.000000  the population and total\_jobs pearsonr correlation coefficient is 0.985716, p-value is 0.000000  the population and traffic\_noise pearsonr correlation coefficient is 0.776772, p-value is 0.000000  the population and unemployment pearsonr correlation coefficient is 0.376342, p-value is 0.000493 |

Compare with survey: find some difference. Why?

Application:

Prediction, 相关性乘权重, limitation analysis.