Supplemental materials for *How Suburban are Big American Cities?*

blogpost published on May 22, 2015, available here on the FiveThirtyEight website

supplemental material posted available here on jedkolko.com; updated May 27, 2015

please let me know of errors here on jedkolko.com

Additional data

There are two downloadable datasets associated with the blogpost.

The first is the full urban/suburban/rural classification for ZIP Code Tabulation Areas (ZCTAs), the Census equivalent of U.S. Postal Service ZIP codes. Only ZCTAs with at least 100 households are included. Please keep in mind that this is based on a national predictive model of the local characteristics associated with whether people say they live in an urban, suburban, or rural area. As with any predictive model, there will be some misclassifications of individual ZCTAs.

The full ZCTA classification file includes three columns:

- **ZCTA**: note that this is stored as a numeric variable for easy use in other applications, so those beginning with 0 are shown as four-digit numbers (e.g., 02116 in Boston is listed as 2116)
- **density**: this is households per square mile (land area only, not water area), according to the 2010 decennial Census
- **classification**, based on the model described below: 1=urban, 2=suburban, and 3=rural

The second is the list of all cities with at least 500,000 population, along with Census estimates of 2014 population and 2013-2014 population growth and the % of households in each city in ZCTAs classified as urban.

Additional methodological detail

The urban/suburban/rural classification presented in the blogpost is based three data inputs: (1) data on 2,008 responses to an online survey, conducted by Trulia, asking people to describe the area where they live as urban, suburban, or rural; (2) the ZIP code of those respondents; (3) Census and other government data on all ZCTAs in the U.S.

From a candidate set of more than 20 characteristics at various levels of geography, we used classification trees (the rpart package in R) to identify the characteristics that best predict whether respondents describe the area where they live as urban, suburban, or rural.

The most important characteristic in predicting how people describe where they live is ZCTA density – that is, households (equivalently, occupied housing units) per square mile (land area only; water area was excluded). The simplest classification tree assigned ZCTAs as follows:

- urban: households per square mile >=2213.2
- suburban: households per square mile >=101.6 and < 2213.2
- rural: households per square mile <101.6

The final model incorporated other important characteristics. Residents of very small cities and towns rarely said they lived in an urban area, even if their neighborhood was quite dense. Residents of lower-income neighborhoods with older housing stock often said they lived in an urban area, even if it was lower-density. Residents of lower-density ZIP codes with lots of businesses sometimes called their neighborhoods urban; so did residents of lower-density, higher-income ZIP codes that are next to higher-density ZIP codes. These additional characteristics were all part of the final classification; in many cases, lower density ZIP codes were classified as urban.

The model had some trouble with a small number of downtown ZIP codes that have low household density because they are primarily business districts. To handle this, we created additional univariate models and identified the level of four characteristics -- establishment density, employment density, non-car commuting, and non-single-family housing -- associated with a survey response of urban. ZCTAs that met urban levels of *all four* of those characteristics were reclassified manually as urban. This applied to 25 ZCTAs nationally, mostly non-residential central business districts of mid-size and large cities.

Measures of ethnic and racial diversity and income inequality, as well as metro-level characteristics, were tested but did not add enough explanatory power to make it into the final model. And many characteristics highly correlated with density, like share of residents commuting by car, added little explanatory power to the model.

ZIP codes / ZCTAs were the smallest level of geography we could use because that is what respondents self-reported. While the characteristics of even smaller geographies, like tracts or block groups, could help predict how people describe where they live, the analysis was limited by our knowing only respondent ZIP codes.