Ideas for DCI Project

We should discuss with our clients in city hall: what do they want to do with DCI?

* To look at the disparity among zip codes?
* To see which zip codes would be any easy fix to bring in corporate sponsorship / jobs / entrepreneurs?
* To target one DCI component? Like HS\_Grad? Or Poverty?
* Economic development: What have other cities with similar zip code DCI rankings and component scores done to improve their ranking? If SA does X and Y could it bring more manufacturing jobs (like a new Toyota plant, call center, small business, or high tech start up) in these targeted areas. Could targeting the zip codes around SA Raceway vastly improve/grow the area much like the Formula 1 track in Austin?
* What can be done to encourage an increase in Chg\_Biz and Chg\_Job, not only at the large corporate level but small businesses?
* Education: Could a program that prevents high school drop outs not only be good for DCI (and the zip code), but also greatly improve poverty and median income? Have any other cities instituted programs that have improved their impoverished areas?
* Growth: This city is going to grow regardless of how the city encourages / discourages growth. How can we make certain areas grow, to alleviate infrastructure? How can we bring in more defense contractors to Military City?

We have questions for EIG:

* Tell them which data we downloaded, and ask “Have we used the right data from the Census Bureau?”
* EIG eliminated cities with low population – did it keep all zip codes regardless of size (one zip code has only 15 people)?
* When we remove zip codes that have missing data in one or more of the DCI components, we are left with 24645. How did EIG keep 26000+ zip codes?

We should develop stylistic standards for plots:

* Ask Dr. Kilger whether UTSA has guidelines for fonts, colors, and logos, particularly applying to ggplot
* Ask our city hall contact whether The City of San Antonio has guidelines for fonts, colors, and logos particularly applying to ggplot
* Develop project standards: fonts, colors, logos, ggplot theme, position of legend

Plot the seven DCI component ranks and the rank of the average of the DCI ranks as follows:

* Do it two ways: all U.S. zip codes and Texas-only zip codes (San Antonio might look relatively better among Texas-only)
* X-axis: 24645 (worst position) to 1 (best position) in the country
* X-axis: 1935 to 1 in Texas
* Y-axis: the value of the component (e.g. vacancy percent from 0 to 100) – with the “bad” value at the bottom (high vacancy, low job growth, etc.)
* The line or the scatter plot will go from the bottom left to the top right of the grid
* Have rectangular areas (like the presidential terms in Dr. Malshe’s assignment) colored according to the color code used by OIG (EIG) >> I knew that exercise would come in hand one day
* Highlight relevant zip codes in two ways (on two sets of charts):
  + Have San Antonio’s zip codes shown with large dots (or squares or whatever) – this will require a categorical variable (“SA” = 0 or 1)
  + Color code all zip codes in Texas (Texas vs. U.S. or Texas by region)
* Place a San Antonio logo (The Alamo, Towers of America, or the San Antonio skyline?) in the upper left and a UTSA logo in the lower right (roadrunner?)

Plot the San Antonio zip codes as follow:

* Each zip code has a bar chart with eight bars: the seven DCI rankings and the ranking of the averages
* Color code the bars and use the legend to identify the color of each DCI component
* Taller bars are good; calculate Height = 24645 – ranking
* We’ll have to produce multiple charts to include all zip codes:
  + Sort by zip code and include the zip codes in the chart title
  + All charts should be to the same scale
  + Use notable images for each zip code in San Antonio (Alamo, Tower of Americas, Six Flags, SeaWorld, Toyota, etc.) if there’s room

Test for correlations and clusters:

* Between each pair of raw DCI components (especially No\_HS vs. No\_Job, Chg\_Job vs. Chg\_Biz, and Med\_Inc vs. Poverty)
* Between each pair of DCI component ranks
* Between consecutive zip codes (can we use autoregressive tools?)
* Between zip codes in San Antonio with similar DCI rankings
* Explore clustering with K nearest neighbors and hierarchical cluster analysis

Look at variances among the DCI rankings:

* New variable “Spread” = each zip code’s highest rank minus its lowest rank.
* What is the distribution of “Spread” (range, quartiles, median, mean, standard deviation)? Is it normally distributed?
* What are the spreads for San Antonio zip codes? Do they rank consistently across seven metrics, or is one metric very different from the others? If so, which one? Is 78xxx the “low vacancy” zip code? Or the “low high school” zip code?
* Are any San Antonio zip codes near the border of two categories (e.g. around 19716, the division between “at risk” and “distressed”)?

Study the gradients:

* We’ve ranked each DCI component from 1 to 24645, but there are much fewer possible values in the raw data (vacancy is 0.0 – 100.0; 1000 possible values). There are many zip codes with the same scores on a particular DCI component. For each San Antonio zip code and each DCI component, what are the lowest and highest possible rankings? Does that materially affect the overall DCI score?
* Can we develop an equation (regression?) that predicts DCI ranking from the seven raw DCI components?
* How much variance does each DCI ranking explain? Could we predict whether a zip code is distressed with fewer variables? (Principal Component Analysis or Factor Analysis?)
* Can we create a Shiny app to allow city hall to ask, “If we reduce vacancies by 10 percentage points, what does that buy us in the rankings?” or “What would it take for each zip code to move up by 1% or to move out of the distressed category?”

Study San Antonio’s results:

* How much do our zip codes differ within the city? (Spoiler alert: San Antonio has some of the highest rankings and lowest ranking in the country, but let’s quantify that.)
* Are our zip code differences due to a few of the DCI components or is it across the board?
* How much variance is there in each raw DCI score and each DCI ranking?