JOURN 4432: Advanced Data Journalism

**FINAL PROJECT**

**May 4th, 2022**

**Student: Han Vu**

**Instructor: Liz Lucas**

**Main Topics:**

1. Using the tidycensus package to pull data from the Census’s American Community Survey (acs5) for the last decade (2011 – 2020) to examine how the City of Columbia has changed over the years.
2. Replicate a cool project to calculate Missouri’s Social Vulnerability to the block group level in various indicators.

**Topic 1:**

**Introduction**

The first reason I chose to explore Columbia over the years is that I want to see if there is anything that has dramatically changed (for the good or the bad), but we haven’t noticed. I haven’t seen any project like that about Columbia, and I want to be the first to discover something interesting about the city, which may lead to some policy implications.

The second reason is that exploring the tidycensus package is so fun. I attended three days workshop organized by the University of Michigan (<https://github.com/walkerke/umich-workshop-2022>) and learned how to use tidycensus from Kyle Walker. Moreover, through the workshop, I discovered his book, and it was the most amazing thing that happened that month. Walker’s book is not only about the tidycensus package, but about tigris, tmap, leaflet, ggplot2, all the amazing packages to visualize the data. So, from the first and brief introduction to tidycensus and Kyle Walker in MaryJo Webster’s session at the NICAR conference in Atlanta in March, when I barely knew R, now I have access to many resources that helped me to build this project.

**Methodology**

The variables I choose to examine are: population demographics, median age, median income, education (percent of people 25 years and older hold bachelor's degree or higher), crime rate (not included in census data), poverty rate, and unemployment rate.

Period: 10 years (2011 – 2020)

Data source: The Census Bureau ACS5 data

Tool: R (tidycensus, tidyverse, ggplot2, and other packages)

**Challenges:**

* The amount of variables and how to find out what variables I need to pull out the data.
* My knowledge about America in general, the city in specific, and how the data is collected are big challenges. For journalists, information and experiences accumulated from past works are so important. They help you spot on some unusual things, giving you the confidence to go with your choice.
* The City of Columbia is at “place” or “cbsa” (metro area) geography level, which is not much supported by the tidycensus package (place level can pull the data from 2010 onward) as well as how the Census Bureau collected data as well. If I do Boone County level, it would be much easier. However, the purpose is to find the story, not for easy work. And city level does not allow me to flex my newly learned skills in data visualization. Haha. That’s why I included a bonus script (Boone County exploration) to show that I’ve tried a lot with this project, much more than what is shown in the main script.
* Overcoming all the technical challenges, I faced the most significant challenge (for the journalist): there’s no story.

All the data I have pulled out proved that the city has slightly changed for the better, nothing worth noticing. But at least there’s the answer to my initial question about how the city has changed. So, have I achieved my goal? Yes. I have my answer, and I learn a lot. But somehow, for me, this is still a failed project since I have no interesting story.

**Along with the script, a data folder, and this note, I turned in a graphic that I made from the data that I collected through the process. I think it’s more expressive than writing a (not so interesting) story.**

**Topic 2:**

**Replicate a cool project to make a Missouri’s version (and Boone County’s version which I do not include in the project folder) of Social Vulnerability.**

I explained thoroughly about the project in the script.

**What I have learned through the process:**

* Lots of try and error. You can spend much time digging and finding nothing. That’s the pain of data journalists.
* Data journalism is so details oriented. It’s not about the well-known big projects with 2.94 TB data, 13.4 million files that 600 journalists have worked together to expose 50 politicians, etc., but about a comma, about how you have to pay attention to each and every variable. A minor mistake can twist a whole story and there’s no excuse. One minor mistake can make a big project you have worked on for months collapse.
* I have tried a lot of cool R functions, but I’m still a newbie, and my skill is not that good. Reproducing other’s projects is much easier than starting your own, with no guidance on where to pull the data from, how to ask the question, and which way to go…
* However, it's all good at the end of the day as long as you’re curious and you’re willing to learn. And the most important skill now for a data journalist, to me, is knowing how to ask (mostly Google) questions.
* **Special thanks to:**

**+ Liz Lucas,** who introduces me to R.

**+ MaryJo Webster,** who introduces me to tidycensus and Kyle Walker.

**+ Kyle Walker** and **Anthony Holmes help me** complete this project, even though they don’t know it.