DA105.3 Build a Data Visualization Project (Udacity)

Chosen Metadata: US Demographic data

1. Project description

Generate insights from census tract data which covers factors related to transportation, income and poverty rate (disparity) along and across state lines.

2. Submission

PDF of this markdown document.

3. Insights and rationale for design choice of visualizations

Insight 1:

- Question: What is the relationship between income and poverty?
- Summary: Overall, counties exhibit inversely proportional relationship between incomes per capita with poverty rates (disparity). Convex-to-concave nature of Power trendline characterizes the complex system nature of incomes and disparity, often a feature in politics and socio-economic factors in which many actors are involved.
- Design choice:
    - Scatter chart aptly captures with high-res for a granularity needed to study the income/disparity relationship, sans bias created by arbitrary consolidation to state/region etc. It faithfully preserves true ranges/distributions of counties across incomes/disparity frontier. This is important so that the trendline is as representative as possible. Drawing relationships at consolidation at state-levels would have hidden the insights of certain counties pulling their weight in exerting outsized influence on state-level statistics.

    - Gnostic use of color-blind friendly colors also uncover further insights from metro-affiliations (i.e. network effect argument) on incomes and disparity.
    - Interactive filters for polities (state and metro) for audience to further explore interesting and possibly confounding insights.

Insight 2:

- Question: Do counties near the coasts earn more?
- Summary: Counties along the coasts generally earn more particularly across New England, Middle Atlantic, Metro DC and SF Bay Area. However, pockets of high/extreme poverty (e.g. Bronx 31% poverty) sit adjacent with those of the highest incomes (e.g. NY-Manhattan $65K income), thus spatially overlapping and exposing a degree of socio-economic segregation within relatively close proximities.
- Design choice:
    - Map aptly captures geospatial properties of incomes/disparity for counties.
    - Conscious choice made to overlay population-sized circles to demonstrate statistical materiality (else, too much visual weight given to less populated, thus less statistically material areas of Midwest and Mountain division of West). This is done instead of just coloring the counties on the map.
    - Color-blind friendly colors critical to draw geospatial inferences on the spread of incomes and disparity across the country.
    - Interactive filters for polities (state and metro) and bands for incomes/disparity for audience to further explore interesting and possibly confounding insights.

Insight 3:

- Question: What is the biggest employment sector across the states?
- Summary: Biggest employment sector is professional nationally (at 37%) and in every state. Northeast region has highest share in professional vs other regions. Across selected states, higher incomes proportional to larger % share for professional sector. Zooming in for top/bottom 10, top 10 states (except NY) also have below national avg (15%) poverty; while disparities in bottom 10 are significant.
- Design choice:

    - Stacked bar chart aptly captures % share splits by employment sector across regions and the country's total.
    - Packed bubbles could be more aesthetically pleasing but data structure prevents this.
    - Color-blind friendly colors essential to draw visual identification. Placement of 'professional' bar, adjacent to the income per capital bar, aims to encapsulate the inferred relationship between professional employment sector and higher incomes.
    - Interactive filters for polities (region, division and state) and top/bottom parameter to allow audience to focus or expand the scope of self-directed investigation.

4. Resources:

- Dashboard (Tableau Story) https://public.tableau.com/app/profile/alan.kong2051/viz/USCensusDemographicData_16858286856570/Story
- Github https://github.com/coderedstorage/US_Census_Demographic_Data
    - US Census Demographic Data.twbx (Tableau Workbook)
    - acs2015-county-data.csv (original data obtained from https://www.kaggle.com/muonneutrino/us-census-demographic-data/data). Imported into MySQL as table called 'kaggle_data'
    - metro_counties.csv (State, County, Metropolitan and MSA mapping). Imported into MySQL as table called 'metro'
    - state_code.csv (state codes and state regions/divisions used by US Census bureau). Imported into MySQL as table called 'state_code'
    - us_regdiv.pdf (original pdf from US Census bureau) to populate state_code.csv.
    - us_census_data.csv (final dataset committed to Tableau) exported from view called 'us_census_data' on MySQL.

5. Tools used

- MySQL to facilitate enablement of original data (acs2015_county_data.csv) to generate final dataset (us_census_data.csv):
    - Overlay urbanization conglomeration (MSAs) information from metro_counties.csv.

    - Overlay state codes, regions and divisions from state_code.csv.
- Tableau Public to create visualization (requires final dataset us_census_data.csv to be committed to Tableau Public).

6. Condition of dataset

- 37-field dataset at county level for:
    - Populations (total, gender, racial, employed).
    - Income related (total and per capita).
    - Disparity(poverty and child poverty rates).
    - Transportation mode.
    - Unemployment rate.
    - Job types (professional, service, office etc).
- More details, see Appendix.

Appendix: About the dataset

Below is copied from https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data.

Context

This dataset expands on my earlier New York City Census Data dataset. It includes data from the entire country instead of just New York City. The expanded data will allow for much more interesting analyses and will also be much more useful at supporting other data sets.

Content

The data here are taken from the DP03 and DP05 tables of the 2015 American Community Survey 5-year estimates. The full datasets and much more can be found at the American Factfinder website. Currently, I include two data files:

- acs2015_census_tract_data.csv: Data for each census tract in the US, including DC and Puerto Rico.
- acs2015_county_data.csv: Data for each county or county equivalent in the US, including DC and Puerto Rico. The two files have the same structure, with just a

small difference in the name of the id column. Counties are political subdivisions, and the boundaries of some have been set for centuries. Census tracts, however, are defined by the census bureau and will have a much more consistent size. A typical census tract has around 5000 or so residents. The Census Bureau updates the estimates approximately every year. At least some of the 2016 data is already available, so I will likely update this in the near future.

Acknowledgements

The data here were collected by the US Census Bureau. As a product of the US federal government, this is not subject to copyright within the US.

Inspiration

There are many questions that we could try to answer with the data here. Can we predict things such as the state (classification) or household income (regression)? What kinds of clusters can we find in the data? What other datasets can be improved by the addition of census data?