

Topic: Crosswalks for Data Classification

Portfolio Projects
with
Pizzazz



Welcome to Today's Lecture!

Today's Lecture

- What do I mean by “crosswalks”?
- **Chi-square analysis:** We don't want too many levels of a classification variable
- **Cardinality:** Term from informatics. We don't want a categorical variable with too high of a cardinality.

Explore our online learning resources!

Visit us: <https://buff.ly/45amjaQ>

- Online tutorials
- Video demonstration
- Online courses
- Mentoring program
- ...and more!



Reasons to Make Classification Crosswalks

- If you program, you have probably programmed a “recoded variable” before (in SAS, SPSS, R, Python, etc.)
- Age → Age groups
- I use crosswalks all the time in data warehousing. Common use: Hospital ID crosswalks.
- When doing a portfolio project, if you have to reduce the number of levels (cardinality) in an important categorical variable, it’s a good time to keep a classification crosswalk

Resources

Example of how to make a simple classification crosswalk

<https://dethwench.com/classification-crosswalks-in-data-science/>

Example of simple classification that reveals findings in a descriptive analysis:

<https://dethwench.com/descriptive-analysis-for-data-science-portfolio-projects/>

Book “Mastering SAS Programming for Data Warehousing” by Monika Wahi – Chapter 10 lays out a detailed guide on developing crosswalks: <https://buff.ly/3lHz1mg>



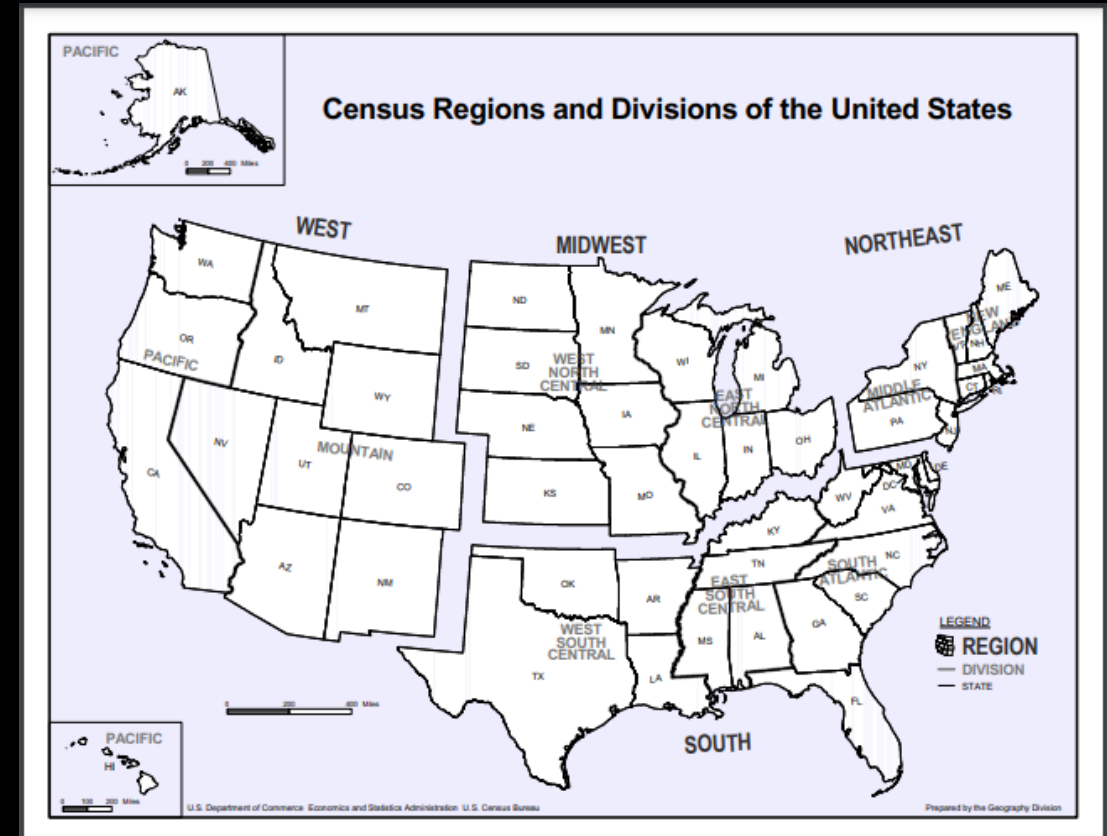
How do you make a crosswalk – literally?

- A crosswalk is just a table of documentation to guide you when you make code. I use Excel.
- In one column, it says the old value, and in the other column, it says the new value.
- The trick is – in the new column, there are repeats. That shows how you are reducing levels for the new variable.
- By designing the crosswalk before programming it, you get to make intelligent decisions.
- Design decisions should be based on what you think will help reveal knowledge pertinent to your research question or aim
- You can make more than set of classifications (xw1k1, xw1k2, etc.)
- You can always add keys (PK, FK) to help with the data processing – or even import the crosswalk!
- You can base your classifications on empirical data (e.g., frequencies of occurrence)



Why Crosswalks for Portfolio Projects?

- The goal is to keep documentation about “why” you made the classification decisions you did
- **Example:** You could use this map to classify 50 states into just a few regions or divisions
- But what if that didn't make sense for your research question?



Census map from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf



Questions?

Facilitator: Monika M.
Wahi, MPH, CPH

Email:
dethwench@gmail.com

Blog:
<https://dethwench.com/blog/>

LinkedIn:
<https://www.linkedin.com/in/dethwench/>

Explore our online learning resources!

Visit us: <https://buff.ly/45amjaQ>

- Online tutorials
- Video demonstration
- Online courses
- Mentoring program
- ...and more!