Intersections of Internet Access, Income, and Population Density

University of Oregon Data Analytics Bootcamp
Capstone Group Project
March 2022

Presentation Rubric Requirements

- Selected topic
- Reason why they selected their topic
- Description of their source of data
- Questions they hope to answer with the data
- Description of the data exploration phase of the project
- Description of the analysis phase of the project
- Technologies, languages, tools, and algorithms used throughout the project
- Result of analysis
- Recommendation for future analysis
- Anything the team would have done differently

Live Presentation Rubric Requirements

- All team members present in equal proportions
- The team demonstrates interactivity of dashboard in real time
- The presentation falls within any time limits provided by instructor
- Submission includes speaker notes, flashcards, or a video of the presentation rehearsal

Initial Data Exploration

Topic Selection: What We are Exploring + Why

- Original Question:
    - Are there discernable patterns connecting Internet Access, Income, and Population Density?
- Why this matters:
    - Internet access is a necessity today and can have detrimental effects in different facets of life. We wanted to see what patterns may exist between access, or lack thereof, to internet and other factors such as income and population density.

Data Sourcing Phase: What + Where

- Description of Census Source Data:
  - US and Puerto Rico Internet Access per Household
  - Broadband/Dial-up/No Access
  - Income (Ratio to Poverty Level)
  - Urban vs. Rural (based off total population of county)
  - Presence of a Computer and Type of Internet Subscription in Household
  - Ratio of Income to Poverty Level of Families in the Past 12 Months

Topic Selection: What Types of Data + Where

Maybe combine w previous slide

- US Census Data Tables Available:
    - B28003: Presence of a Computer and Type of Internet Subscription in Household
    - B17026: Ratio of Income to Poverty Level of Families in the Past 12 Months
    - S1902: Mean Income in the Past 12 Months (in 2018 Inflation-Adjusted Dollars)
- Supplementary Information:
    - Oregon Urban vs Rural Classifications from USDA

Data Exploration Phase: Initial Planning + Discussion

From initial README.md:

- Data Cleaning and Analysis: Pandas will be used to clean the data and perform an exploratory analysis. Further analysis will be completed using Jupyter Notebook, python and Rstudio.
- Database Storage: Heroku is the database we intend to use, and we will integrate Flask to display the data.
- Machine Learning: Jupyter Notebook is the ML library we'll be using to create a classifier. Our training and testing setup is to test the correlation between variables. Visualize the difference of Ratio of Income to Poverty Level of Families in the Past 12 Months and Presence of a Computer and Type of Internet Subscription in Household.
- Dashboard: In addition to using a Flask template, we will also integrate D3.js for a fully functioning and interactive dashboard. It will be hosted on Tableau.
- Model: Logistic Regression, Unsupervised Learning
- Communication Protocols
  - Plan out weekly tasks and timelines during in class Zoom meeting
  - Use Slack messages for additional questions during the week
  - Use Issues feature in Git Hub to address "issues"

Data Exploration Phase: Adaptations

What else to include here?

- Roadblocks + challenges
  - Data gaps:
    - Data from 839 counties across the United States is not every county, but at least a few counties in every state are represented in our dataset.
    - Solution: Case Studies for Urban / Rural counties
    - Urban / Rural definition doc
  - Connectivity:
    - Trial Tableau accounts, shared visibility
  - Additional data needed:
    - Avg household income dataset

Primary Question: Revised

- Identify if there is an income threshold above which two criteria are met: broadband internet access and a desktop/laptop computer.
    - Primary necessities for sustainable remote work and school
    - One without the other does not accomplish the goal; both are necessary to have useful access level to support higher level education or remote work

Data Analysis Process

Data Analysis Phase: Overview + Technology Stack

- Database: Sketch + Setup
  - pgAdmin, SQL, SQLAlchemy, Jupyter Notebook, Python
  - SQLAlchemy, Heroku
- Machine Learning Model: Outline + Build
  - Python, Jupyter Notebook
- Visualization: Connect + Display
  - Heroku, Tableau, Google Slides

Data Analysis Phase: Data Cleaning

- Data preprocessing was done with cleaning the .csv files downloaded from the Census bureau website. I used both Pandas and eyes-on scrolling in Excel to clean the data. There were a few rows with null/other values creating float and object variables that were easier to just eliminate (two in total). This brought my data to fully Integer. Also, the model did not care for 0 values, it popped an error with less than 2, so there were two values, that were zero that I changed to 2.
- Cleaning round 2 for visualization needs

Update w details

Data Analysis Phase: Database

- Static data: Used pgAdmin to create four tables to use in the machine learning: comp_types, income_internet, inc_int_comp, inc_int_no_int
- Interfaces: Connected Heroku to pgAdmin, to make local database available in the cloud
- Tables: Used pgAdmin and SQL to create tables with the cleaned raw data, started with two tables: comp_types and income_internet
- Join: Used SQL to join tables, created inc_int_comp
- Connection String: Used SQLAlchemy in Jupyter Notebook to connect Heroku to Jupyter Notebook

Data Analysis Phase: Machine Learning Model Initial Plan

- Predictions of whether there is a correlation between Internet Access and Income/Population Density
- First Model:
  - Logistic regression ML with Internet Access represented as y/n (0/1). Variables for Income from 0.5 to >= 5.0 (10 discrete) and variables for Population size in US/Puerto Rico counties 0k-100k to >= 1M at 100k Intervals.
- Second Model:
  - Unsupervised ML with plugging in all variables and determining if any patterns arise. I'm hoping our supervised ML will be able to aid in targeting more accurately our unsupervised ML.
- Additional:
  - If there is time, we would also like to look at other variables with potential impact i.e. education.

Data Analysis Phase: Machine Learning Model Outcome

Possible split two slides or edit down slightly

- Jupyter Notebook is the ML library we'll be using to create a classifier.
- Our modeling setup is to visualize the correlation between variables. Differences between various incomes (under 10K to over 75k) using presence of a Computing device (desktop/laptop/smartphone/tablet/other) and internet subscription or no internet subscription.
- Feature selection was based on the data that we have. I have spent a good bit of time learning that there is no way to correlate data with an unequal amount of variables, so that was somewhat limiting in feature selection I will discuss this further later on.
- The data was split into training and testing sets and errored out. For logistic regression, the outcome seems to be binary, when I attempted to use ".(income)without internet access" as my y, it threw a value error and said "ValueError: The least populated class in y has only 1 member, which is too few. The minimum number of groups for any class cannot be less than 2." The lowest value was 18.
- I ended up using unsupervised learning, which was quite happy with the data and performed as expected. However, this data does not seem to allow confusion matrix and accuracy analysis like LR. Therefore I will use PCA and covariance.

Results + Demonstration

Data Analysis Results

High level overview of results + outline of Tableau viz – can build out or combine w following slide

- Results of analysis
  - Questions answered
- Live interactive demonstration of dashboard
  - Dashboard built in Tableau
  - Map – National Data
  - Charts / Graphs – Oregon Data

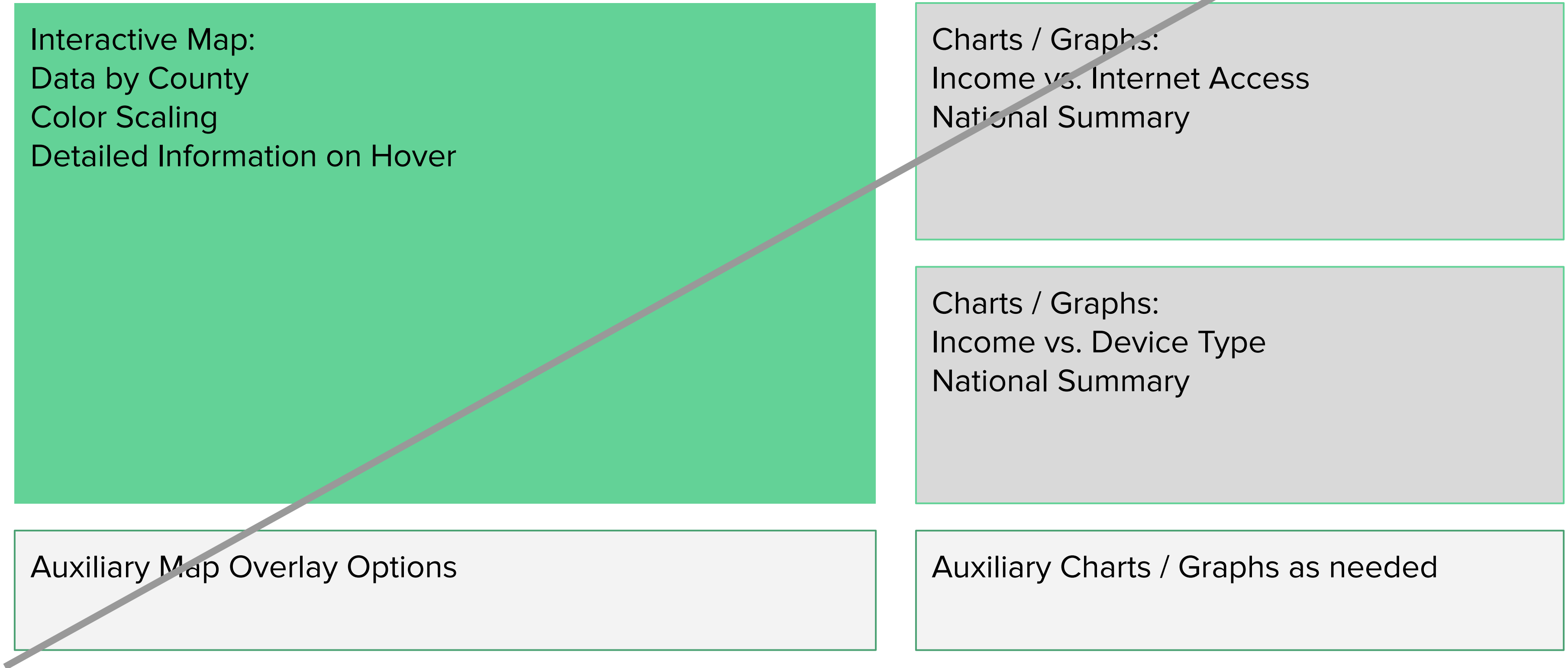
Dashboard Wireframe

Detailed outline of Tableau viz, can skip during actual pres

- Interactive Map with layers
  - Internet Access, Computer/Device, Income, Population
  - Census Map API
- Charts + graphs
  - Income level vs. Internet access (broadband, dialup, none)
  - Income level vs. Device type (desktop/laptop, smartphone, tablet, other, none)
    - Smartphone only vs. Desktop/Laptop
  - Q: income threshold for which Internet / Device becomes feasible
    - Internet + Desktop/Laptop as necessary components for remote school/work
  - Urban case study
  - Rural case study
- Map overlays for auxiliary info
  - Type of internet (broadband / dial-up)
  - Education level

Results: Dashboard + Visualizations

Dashboard Wireframe



Interactive Map:
Data by County
Color Scaling
Detailed Information on Hover

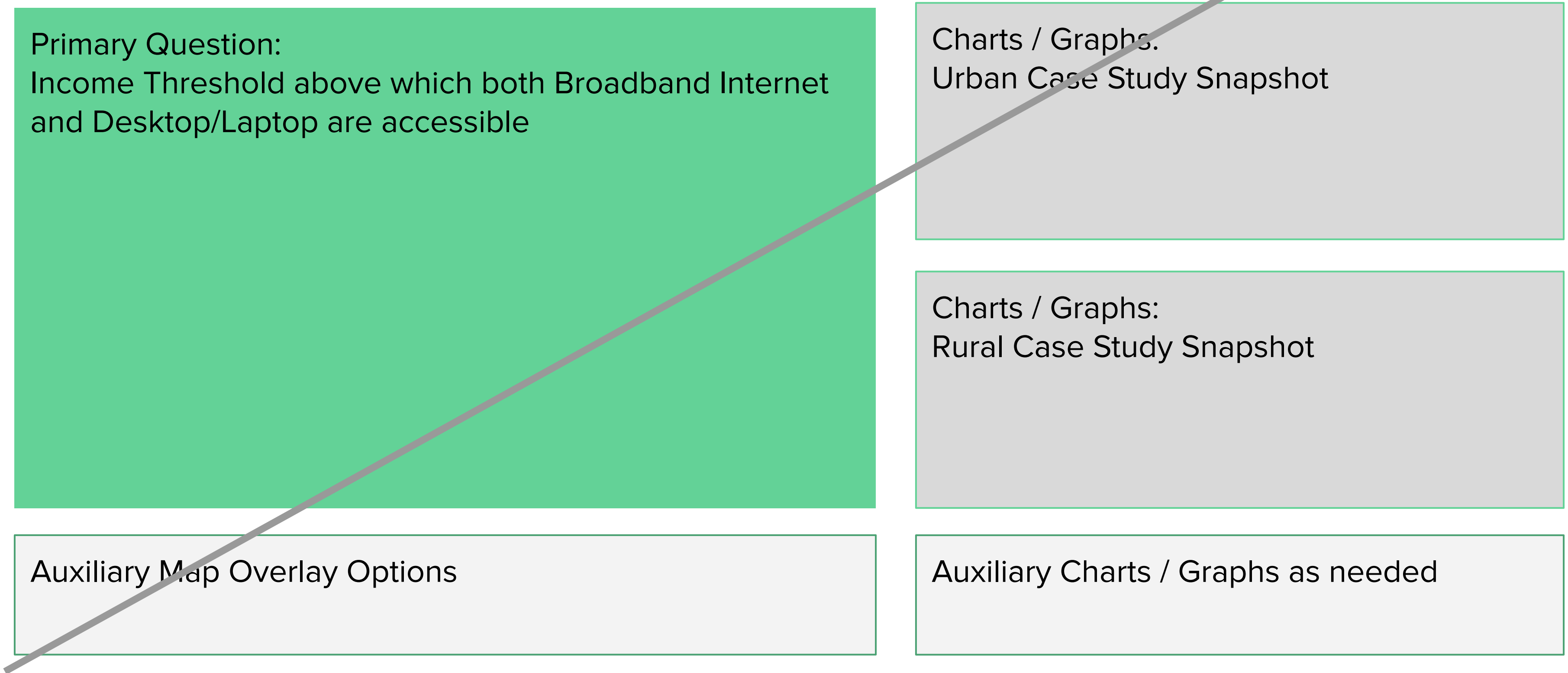
Charts / Graphs:
Income vs. Internet Access
National Summary

Charts / Graphs:
Income vs. Device Type
National Summary

Auxiliary Map Overlay Options

Auxiliary Charts / Graphs as needed

Dashboard Wireframe



Primary Question:
Income Threshold above which both Broadband Internet and Desktop/Laptop are accessible

Charts / Graphs:
Urban Case Study Snapshot

Charts / Graphs:
Rural Case Study Snapshot

Auxiliary Map Overlay Options

Auxiliary Charts / Graphs as needed

Wrap Up

Recap

Any other high level takeaways can be included here

- Tools
- Data Sources
- Analysis Process
- Results

Future Recommendations

Could be two slides: one on future analysis/data + one on team process reflection

- Recommendations for future analysis
- Additional information on existing programs / initiatives in this area
  - Pew Research https://www.pewresearch.org/fact-tank/2021/06/22/digital-divide-persists-even-as-americans-with-lower-incomes-make-gains-in-tech-adoption/
  - USDA Rural Development / Rural Broadband Loans and Grants https://www.rd.usda.gov/programs-services/telecommunications-programs/rural-broadband-loans-loangrant-combinations-and-loan-guarantees
    - Map
- Anything the team would have done differently

End

University of Oregon Data Analytics Bootcamp
Capstone Group Project
March 2022