Intersections of Internet Access, Income, and Population Density

University of Oregon Data Analytics Bootcamp
Capstone Group Project
March 2022

Initial Data Exploration

- **Initial Data Exploration**
- Data Analysis Process
- Results + Visualizations
- Wrap Up + Next Steps

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 lack of access can have detrimental effects in different facets of life. We were interested in 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

- Desired Source Data per Household:
  - **Internet**: Broadband/Dial-up/No Access
  - **Computer**: Desktop/Laptop, Tablet, Smartphone, No Computer
  - **Income:** Ratio to Poverty Level
  - **Urban / Rural:** Based on Total Population of County
- 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:
  - USDA: Oregon Urban vs Rural Classifications

Data Exploration Phase: Initial Planning + Discussion

From initial discussions:

- Data Cleaning and Analysis:
  - Pandas was used to clean data and perform an exploratory analysis. Further analysis using Jupyter Notebook, Python and Rstudio.
- Database Storage:
  - SQL, hosted on Heroku
- Machine Learning:
  - Testing correlation between variables: ratio of internet access to income over 12 Months (2018) and Presence of a desktop/laptop.
  - Models: Logistic Regression, Unsupervised Learning
- Dashboard:
  - Tableau
- Communication Protocols:
  - Zoom, Slack, GitHub

Data Exploration Phase: Challenges + Adaptations

- Data gaps:
  - Data from 839 counties across the United States and Puerto Rico (not every county, but at least a few counties in every state are represented in our dataset).
    - Solution: Case Studies for Urban / Rural counties and Urban / Rural definition doc
    - Avg household income dataset
- Data formatting:
  - Income level brackets as separate columns for each county
- Software tools + connectivity:
  - Trial Tableau accounts + shared visibility

Primary Question: Revised

- **Identify if there is an income threshold above which two criteria are both met: presence of Internet Access and a Desktop/Laptop Computer.**
- Why this matters:
  - 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

- ➔ Initial Data Exploration
- ➔ **Data Analysis Process**
- ➔ Results + Visualizations
- ➔ Wrap Up + Next Steps

Data Analysis Phase: Overview + Technology Stack

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

Data Analysis Phase: Data Cleaning

- Preprocessing for Machine Learning:
  - Raw data files from the Census Bureau

	A	B	C	D	E	F
1	B28003_0(	B28003_0(	B28003_0(	B28003_0(	B28003_0(	B28003_0(
2	Estimate!	Margin of	Estimate!	Margin of	Estimate!	Margin of
3	83501	2931	77443	2976	305	332
4	44264	1834	38768	1871	229	205
5	30323	1437	26269	1408	204	216
6	26462	1458	22675	1395	0	201
7	30155	1094	26120	1194	41	68
8	38625	1544	34041	1772	0	201

  - Cleaned with Pandas and SQL

	total integer	one_or_more integer	desktop_laptop integer	only_desktop_laptop integer
1	83501	77443	62729	4403
2	44264	38768	30768	3054
3	30323	26269	19117	654
4	26462	22675	15879	1440
5	30155	26120	21528	599
6	38625	34041	23824	1682

- Preparing data for Visualization:
  - Tableau Calculated Fields

Data Analysis Phase: Database

- Static data:
  - Used pgAdmin to create tables for Machine Learning Models and Visualization
- Tables:
  - Create starter tables from cleaned raw data
- Join:
  - Used SQL to join tables, created inc_int_comp

```sql
-- Join comp_types and income_internet tables
SELECT
    ct.total, ct.one_or_more, ct.desktop_laptop,
    ct.only_tablet_portable, ct.other, ct.only_o
    ii.less_10000_total, ii.less_10000_dialup, i
    ii._10000_19900_dialup, ii._10000_19900_broa
    ii._20000_34999_broadband, ii._20000_34999_n
    ii._35000_49999_none, ii._50000_74900_total,
    ii.greater_75000_total, ii.greater_75000_dia
    ct.id, ct.geographic_area_name
INTO inc_int_comp
FROM comp_types as ct
LEFT JOIN income_internet as ii
ON ct.id = ii.id;
```

```sql
-- Creating tables for access-analysis-heroku
CREATE TABLE comp_types (
    total INT,
    one_or_more INT,
    desktop_laptop INT,
    only_desktop_laptop INT,
    smartphone INT,
    only_smartphone INT,
    tablet_portable INT,
    only_tablet_portable INT,
    other INT,
    only_other INT,
    no_computer INT,
    id INT,
    geographic_area_name VARCHAR,
    PRIMARY KEY (id),
    UNIQUE (geographic_area_name)
);
```

Data Analysis Phase: Database

- Interfaces:
  - Connected Heroku to pgAdmin, to make local database available in the cloud
- Connection String:
  - Used SQLAlchemy in Jupyter Notebook to connect Heroku to Jupyter Notebook

Export Tables to PostgreSQL

```
# Import dependencies

import sqlalchemy
from sqlalchemy.ext.automap import automap_base
from sqlalchemy.orm import Session
from sqlalchemy import create_engine, func, MetaData


from config import password
```

```
connection_string = f"postgresql://yhcaekagzcfjez:{password}@ec2-3-211-240-42.compute-1.amazonaws.com/d2hgqg2homreih"
```

```
# Create engine

engine = create_engine(connection_string)
```

```
# Create Session Link to the Database
session = Session(engine)
```

```
#To add a Jupyter DataFrame to a SQL Table, use the following format

county_data_df.to_sql(name="county_data", con=engine)

# To confirm that the table imported correctly:
# right-click the new table in pgAdmin and select Properties
# click the Columns tab to make sure all columns have an appropriate data type
# Close the Properties window and right-click the new table again
# Select "View/Edit data" followed by "First 100 Rows"
# Right click new table again and select Query Tool
# Inside query tool, run SELECT COUNT(*) FROM {the new table} to make sure all rows were imported
```

```
state_data_df.to_sql(name="state_data", con=engine)

```

```
county_data_df
```

	id	county	state	total_pop	households	avg_household_income_dollars
0	1003	Baldwin County	Alabama	218022	83501	80251
1	1015	Calhoun County	Alabama	114277	44264	61117
2	1043	Cullman County	Alabama	83442	30323	67585
3	1049	DeKalb County	Alabama	71385	26462	57354
4	1051	Elmore County	Alabama	81887	30155	70961

	index bigint	id bigint	county text	state text	total_pop bigint	households bigint
1	0	1003	Baldwin County	Alabama	218022	83
2	1	1015	Calhoun County	Alabama	114277	44
3	2	1043	Cullman County	Alabama	83442	30
4	3	1049	DeKalb County	Alabama	71385	26
5	4	1051	Elmore County	Alabama	81887	30
6	5	1055	Etowah County	Alabama	102501	38
7	6	1069	Houston County	Alabama	104722	38

Data Analysis Phase: Machine Learning Model

Objective: predictions of correlation between internet access and income using Jupyter notebook with Python 3.7 in mlenv.

- First Model:
  - Logistic regression ML with internet access represented as: with/without (1/0). Variables for income computed in brackets for amounts: under 10k to over 75k.
- Second Model:
  - Unsupervised ML evaluating all variables and determining if any patterns arise. Goal to be able to see any unexpected patterns/trends.
- Additional possibilities:
  - Integration of other variables with potential impact, i.e. education and other economic indicators..

Data Analysis Phase: Machine Learning Model Outcome

	precision	recall	f1-score	support
0.0	0.97	1.00	0.98	202
1.0	0.00	0.00	0.00	7
accuracy			0.97	209
macro avg	0.48	0.50	0.49	209
weighted avg	0.93	0.97	0.95	209

>10k

	precision	recall	f1-score	support
0.0	0.00	0.00	0.00	106
1.0	0.49	1.00	0.66	103
accuracy			0.49	209
macro avg	0.25	0.50	0.33	209
weighted avg	0.24	0.49	0.33	209

35k to 50k

	precision	recall	f1-score	support
0.0	0.97	1.00	0.98	202
1.0	0.00	0.00	0.00	7
accuracy			0.97	209
macro avg	0.48	0.50	0.49	209
weighted avg	0.93	0.97	0.95	209

10k to 20k

	precision	recall	f1-score	support
0	0.00	0.00	0.00	24
1	0.89	1.00	0.94	185
accuracy			0.89	209
macro avg	0.44	0.50	0.47	209
weighted avg	0.78	0.89	0.83	209

50k to 75k

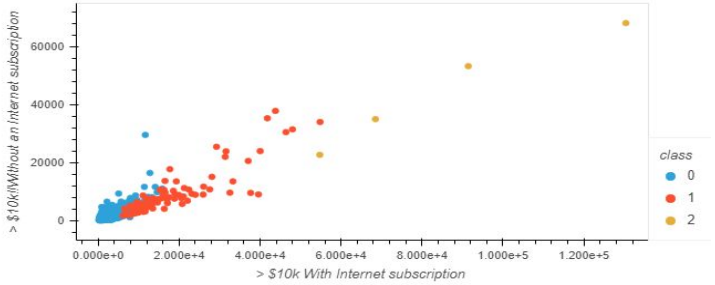
	precision	recall	f1-score	support
0.0	0.90	1.00	0.95	188
1.0	0.00	0.00	0.00	21
accuracy			0.90	209
macro avg	0.45	0.50	0.47	209
weighted avg	0.81	0.90	0.85	209

20k to 35k

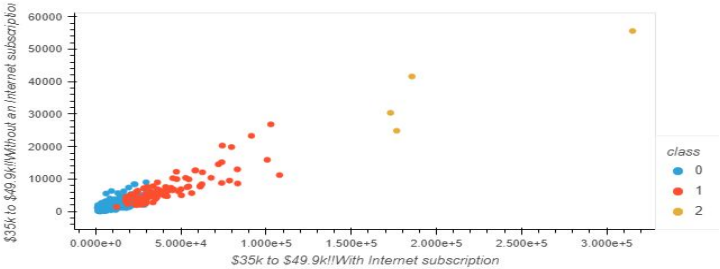
	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	1.00	1.00	1.00	208
accuracy			1.00	209
macro avg	0.50	0.50	0.50	209
weighted avg	0.99	1.00	0.99	209

<75k

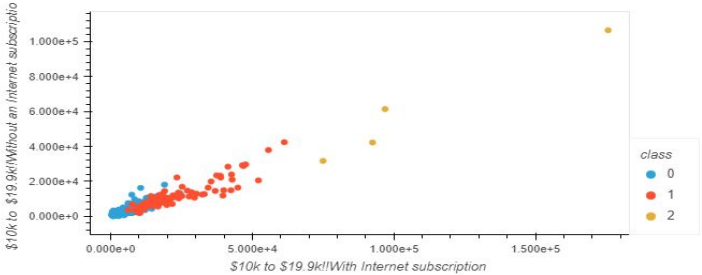
Data Analysis Phase: Machine Learning Model Outcome



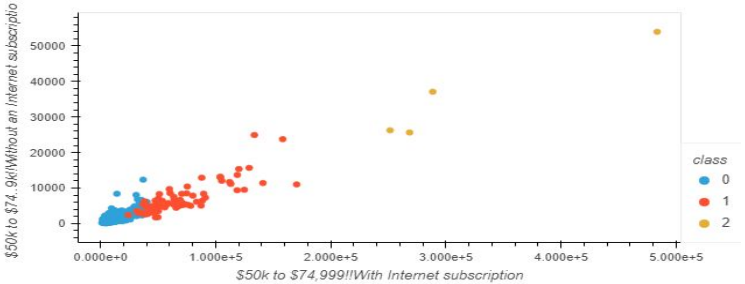
>10k



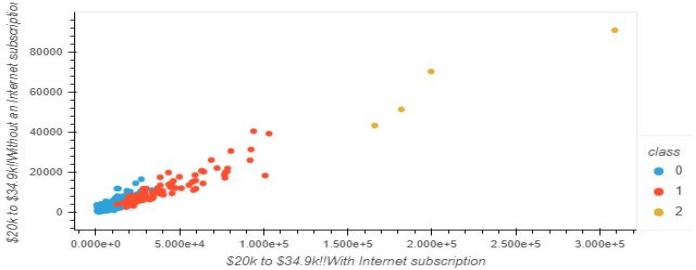
35k
to
50k



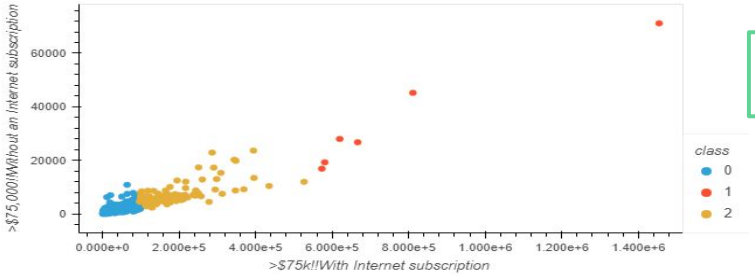
10k
to
20k



50k
to
75k



20k
to
35k



<75k

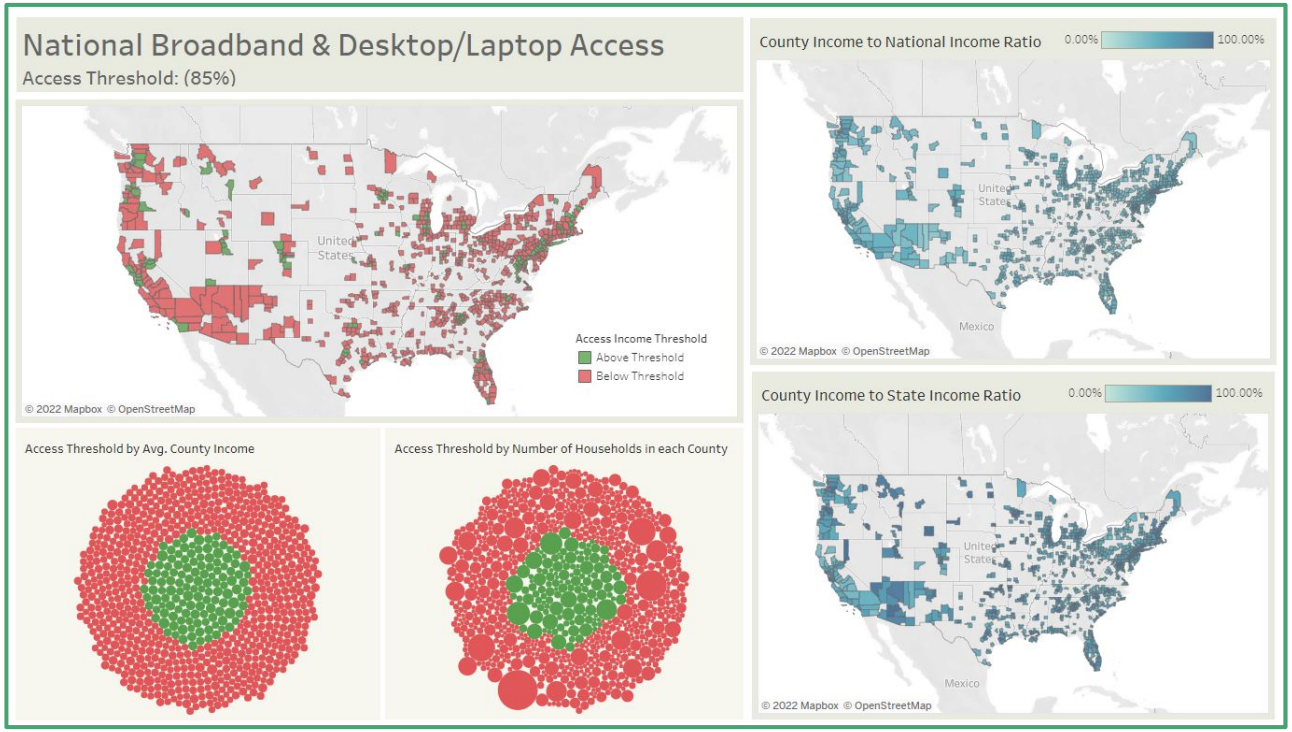
Results + Demonstration

- ➔ Initial Data Exploration
- ➔ Data Analysis Process
- ➔ **Results + Visualizations**
- ➔ Wrap Up + Next Steps

Data Analysis Results

- Results of analysis + interactive demo of dashboard
  - Dashboard built in Tableau + connected through Heroku
  - National Data:
    - Interactive Maps
  - Oregon Data:
    - County Case Studies

Results: Dashboard + Visualizations Demo



Wrap Up + Next Steps

- ➔ Initial Data Exploration
- ➔ Data Analysis Process
- ➔ Results + Visualizations
- ➔ **Wrap Up + Next Steps**

Results Recap

Identify if there is an income threshold above which two criteria are met: broadband internet access and a desktop/laptop computer.

- Results:
  - The regression modeling shows **internet access increases as income increases**, with $35-50k income bracket looking to be a shifting point across the county data.
    - However, there are still higher income counties with lower access levels and vice versa, indicating that there are **likely other contributing factors in addition to income**.
    - For example, in Oregon: $83k is the lowest county income above 85% access threshold, $96k is the highest county income below 85% access threshold.
  - Urban / Rural was not as strong of an indicator as predicted.
    - Often counties are a mix of Urban + Rural areas (i.e. Clackamas County)
    - Urban county does not necessarily indicate higher level of access (i.e. Multnomah County)

One Sentence Walk-Away

Broadband Internet Access and Access to a Desktop/Laptop computer – while fairly widespread – are still not ubiquitous or guaranteed across the United States.

- Out of 15 sampled counties in Oregon, only 4 counties have more than 85% of respondents with both Broadband Internet Access and a Desktop/Laptop computer.
- Limited access to Broadband Internet Access and a Desktop/Laptop computer – and therefore limited access to remote education and work opportunities – are still important factors for a significant portion of Americans today.

Future Recommendations

- Recommendations for future analysis
  - Adjusting threshold from 85% to other options, (i.e. 60%) would allow deeper insights and more granular details for lower income brackets.
    - <10k income bracket averages 64% access across all counties, 35k bracket averages 84% access.
  - Analyzing how average education levels (high school, some college, graduate degree, etc.) of counties correlate – or do not correlate – with levels of access. Does one influence the other?
  - Similarly, identifying whether there are other specific socio-economic or infrastructure factors that have a measurable impact on technology access could inform discussions of expanding access to more Americans through future programs or policies.

Future Recommendations

- Work is being done through various national and regional programs to increase equity of access.
- Additional information on existing programs / initiatives / research:
  - 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 of current initiatives

End

University of Oregon Data Analytics Bootcamp
Capstone Group Project
March 2022

Susan Friesen, Sam Mosher, Daisy Zhao, Kim Conrad

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

Dashboard Wireframe

Detailed outline of Tableau viz ideas, 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