DATA SCIENCE FOR ALL

Closing the Digital Divide

COHORT 3

Team 13

Geri Harding

linkedin.com/in/geriharding
@geridoesdata

Millie Symns

linkedin.com/in/millie-symns
@millieosymns

Jayuan Ruiz

linkedin.com/in/jayuan-ruiz-5ba17917a/

Table of Contents

Problem and Solution

Addressing the Digital Divide means determing the need

Problem

Assessing which areas in the United States have the most need for access to internet and technology

Solution

Creating an interactive dashboard that provides a comprehensive score of needs using key metrics

Problem

With the advancement of technology, the entire world is undergoing an age of digital transformation. As the world moves online, from schooling, banking, and even telehealth, class inequities, discriminatory practices, and systemic injustices persist in today's reality. With roughly 1 in 4 households (about 27.6 million Americans) not having any access to internet services, there is an inequitable and widening gap for access to the internet.[1] In a recent study, with median household incomes below $25,000, about 51% did not have internet because they said it was too expensive.[2] Additionally, this gap unfairly affects Black and Brown communities even when accounting for income disparities because of the history of systemic racism.[2]

The problem only worsened as the COVID-19 pandemic ravaged the nation in March 2020. With people mandated to stay home in quarantine, internet access became more of a need than a luxury than ever before. While schools and workplaces were forced to transition to remote, individuals who either had no or limited devices and subpar internet speeds struggled to keep up with the new norm.

Additionally, before the pandemic, there had been a strong effort in upskilling people choosing to continue their education or learn new skills in technology to have more career opportunities. With initiatives such as those offered by large tech companies and organizations, people need the resources of quality technology and the internet. Beyond the pandemic, companies, schools, and all levels of government must keep their eyes open to accessibility disadvantages their members may face.

Therefore, the question of our project became: **"How do we help close the internet accessibility gap to ensure no one is left behind?"**

Solution

To help close the internet accessibility gap, our first task was to identify and assess which areas/ communities were at the most risk of lack of internet access and determine that level of risk. By creating a comprehensive internet accessibility score and mapping the scores into an interactive dashboard, we hoped to illustrate to the government and other third parties that additional funding for infrastructure investment, utility subsidies, and/or device vouchers as a part of the new Bipartisan Infrastructure Deal should be allocated.

Additionally, after understanding where the areas of needs were across the country, we sought to predict two key outcomes: 1) percent of households with broadband internet and 2) percent of households with desktop devices. Our hope in predicting these outcomes was to identify the key factors that have the most influence.

Data

Dataset #1

County Internet Access
Census ACS (from alt data source)

Dataset #2

Income in the past 12 months
Census ACS

Dataset #3

Types of computers and internet subscriptions
Census ACS

Dataset #4

Population/ Population Classification
Census ACS

Dataset #5

Employment Status
Census ACS

Addressing internet and technology accessibility is a complex issue because there are many aspects that leaders should take into consideration. From how much income one has, the education system to the cost of infrastructure. We started with various datasets and then paired them down to what we felt were the essential datasets for exploration.

We settled on five primary datasets. The three Census Bureau datasets were from the American Community Survey (ACS). They had variables related to household types of computer devices and internet subscriptions, household income, household and family size, and employment status, all at the county level. There was also one overall population dataset from the Census Bureau.

We needed to get the broadest picture possible to access the problem for the entire country. While the Census Bureau did have information on internet availability, it was limited to the surveyed counties, just above 800 counties. We found a separate data source on broadband internet on the county level from the Center of Technology & Data out of Arizona State University that included complete information on the percentage per household with broadband internet by county from 2010 to 2018. Since this dataset was the most complete and imperative to our data exploration, we decided to keep this data source.

We extracted the Census Bureau datasets from the website around January 2022. We downloaded each dataset from 2018 with one-year estimates. We chose 2018 data because that was the latest data available from the source, so we wanted the other variables to match the same timeframe for consistency. We assume that population statistics do not vary too much between years, so they should still provide a relatively active picture of what is happening in the country.

Data Cleaning + Preparation

Each dataset came with a set of notes and table descriptions to describe the column names from the Census Bureau. Most datasets had several levels of details with estimates and margins of details with values by race, gender, and age. We kept the Census naming conventions for simplicity and referred to the documents on the columns' names.

With our raw datasets now cleaned and columns selected, the next step would be to combine them into our final analytical dataset. Each file included an “ID” column which contained data for the FIPS Code (Federal Information Processing Standard), the unique identifier. This allowed our team to easily merge all the datasets with a left outer join to the population dataset utilizing the FIPS Code as the primary key.

STEP 1
Remove first row

STEP 2
Split county and state information into two columns

STEP 3
Replaced all characters of “N” and “*****” or blanks with NaN values

STEP 4
Format all numeric values to float data types

STEP 5
Selected specific estimates from each dataset that provided information of the broadest level for the entire population or average households by county

STEP 6
Export each dataset as a newly formatted dataset into CSV files

Core Datasets

The table below highlights the datasets, their structure prior to and after cleaning, the key variables used from each file, and the size after we selected our final variables. As well as the resulting use of each dataset with Core representing data used for modeling/ scoring and supplemental for data used for visualizations and dashboarding purposes.

Dataset	Source	Raw File Size	Key Variables	Raw File Size
Population/ Population Classification	Census	3223 rows × 47 columns	ID, Population Total, County, State, Population, Classification	3221 rows × 5 columns
County Internet Access	Center of Technology & Data out of Arizona State University	59528 rows × 6 columns	ID, CFIPS, County, State, Year, Broadband Percentage	3133 rows × 6 columns
Income in the past 12 months	Census	842 rows × 104 columns	ID, County, State, Estimated Total Households per Household, Estimated Median Income, Estimated Total Families, Estimated Total Income per Family	838 rows × 7 columns
Types of computers and internet subscriptions	Census	839 rows × 126 columns	ID, County, State, Estimated Total Households with Desktops, Smartphone, Portable, Other, None, Estimated Percent Households with Desktops, Smartphone, Portable, Other, None	838 rows × 34 columns

Additional Datasets

Additionally, we used other datasets as references to make decisions later for our analysis. The household size information helped us determine the largest household size to apply the poverty level as a reference point for our scoring on income. The regions and divisions provided other categories for exploring the data besides the county level and the whole country. A community member on Github saved a CSV file for researchers equivalent to the information in a PDF file from the Census Bureau.

Dataset	Source	Raw File Size	Key Variables	Raw File Size
Employment Status	Census	840 rows × 256 columns	ID, County, State, Estimated Total Employed 16+, Estimated Ratio Employed 16+, Estimated Unemployed Ratio 16+	838 rows × 6 columns
Household and Family Size	Census	840 rows × 176 columns	ID, Official Total Households Estimate, Estimate Average Household Size, Official Total Families Estimate, Estimate Average Family Size	838 rows × 5 columns
Census Bureau Regions and Divisions	Census & Github	52 rows × 4 columns	State, State code, Region, Division	-- --

Exploratory Data Analysis

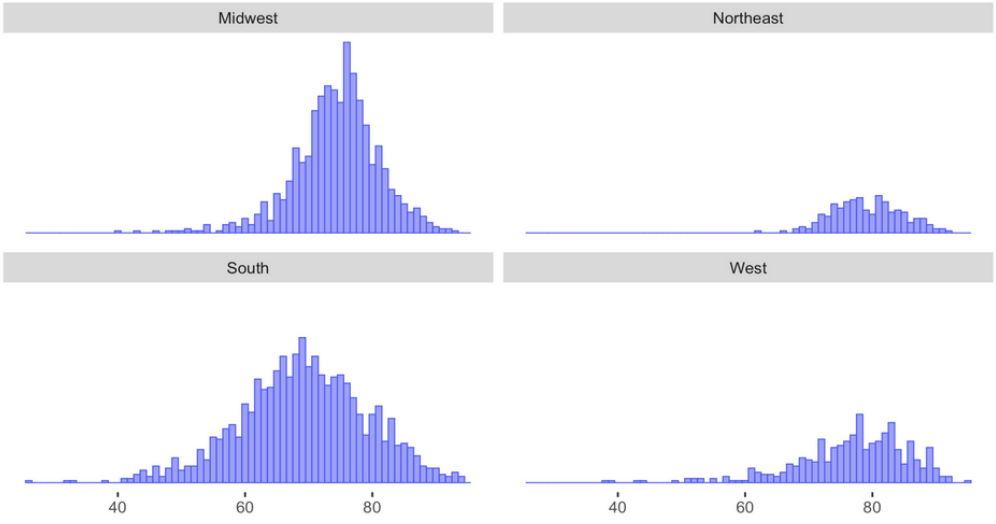
Because our objective was to assess varying communities' internet accessibility and the factors leading to that, the EDA for our project mainly focused on three areas: 1) Percentage of households with broadband internet, 2) Median income levels, and 3) Percentage of households with desktops, smartphones or portable devices. We looked at the distribution of each variable on its own and the correlation with other variables. We also observed distributions by region. Below are some examples of our findings in our exploration.

01

BROADBAND INTERNET

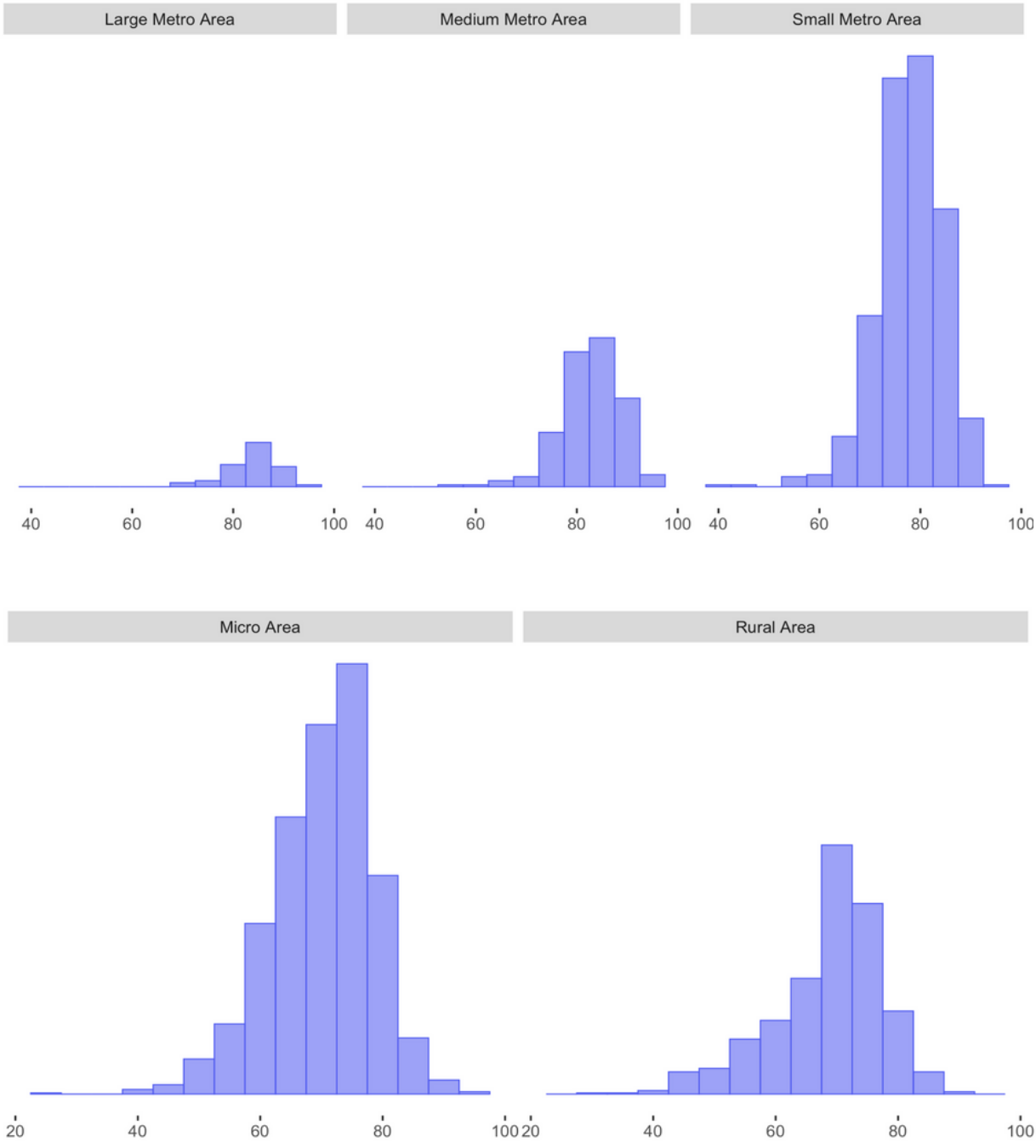
When initially looking at the distribution of the percentage of households with broadband internet, we see that most of the data is between 70%- 80%, with a left skew.

When segmenting the percentage of households with broadband data by region, you see the differences in resources. Below we can see that the frequency by region and the data peaks around 80% for all regions. However, there is a significantly larger spread for the south and west regions as these were the only regions to go below 40%. The south region also has a large range of the percentage of households to have broadband internet. This is an example of the variability of internet access in the regions.



When looking at broadband by population class, we see a similar confirmation with medium and large metro hovering around 80% in most counties while small metro, micro, and rural areas skew lower around 70%.

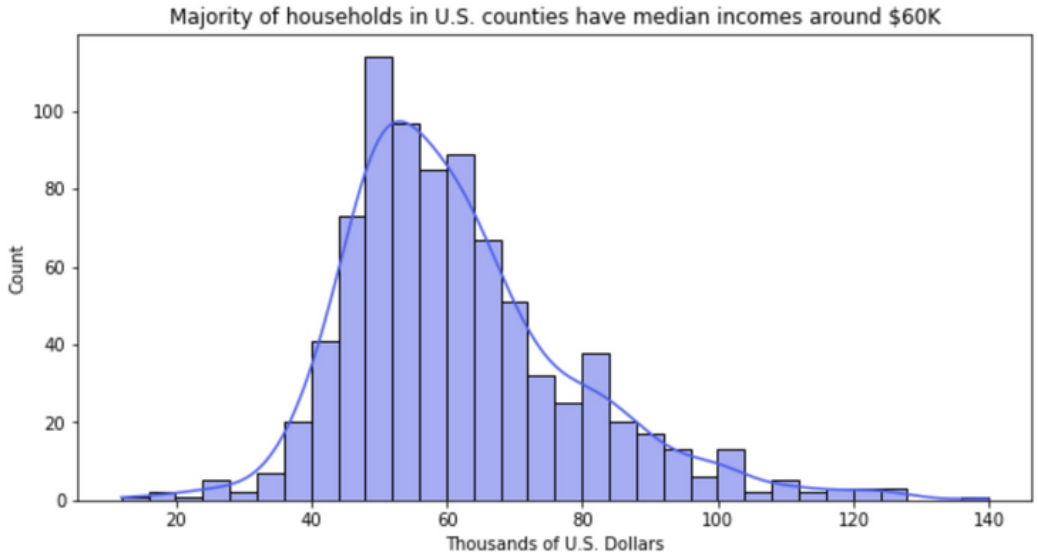
This was a significant observation and distinction for our potential audience because we are being to see where discrepancies in recourses may take place.



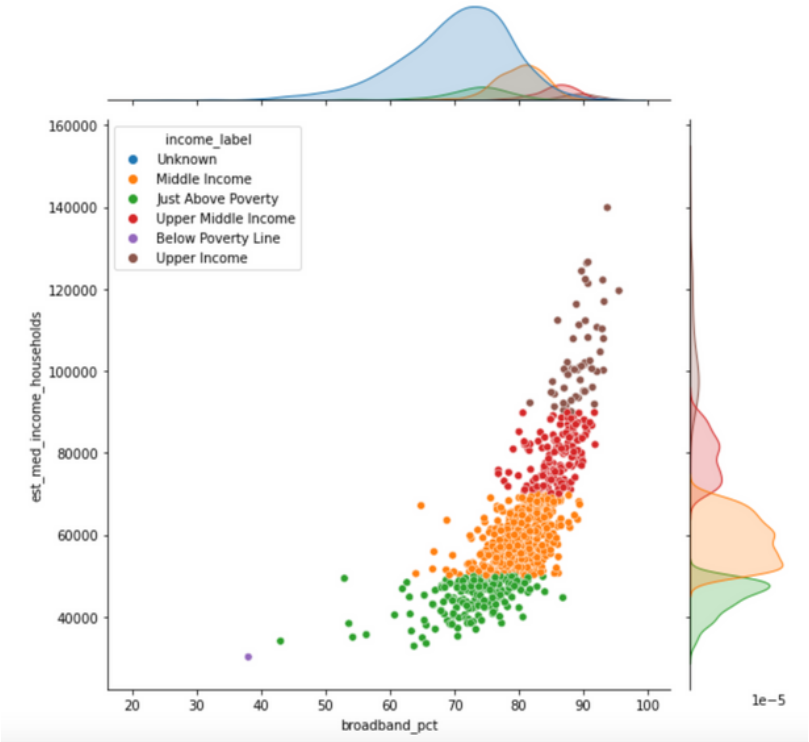
02

MEDIAN HOUSEHOLD INCOME

Income alone showed us that most households across U.S. counties had a median household income of $60K. When we look at the median income by region, it did not appear to differ too much from the overall distribution.



However, when we started to look at the relationship between income and broadband internet. As median income goes up, so does the percentage of households with broadband internet. This highlights that the communities most affected by lack of internet access are those with a lower median income.



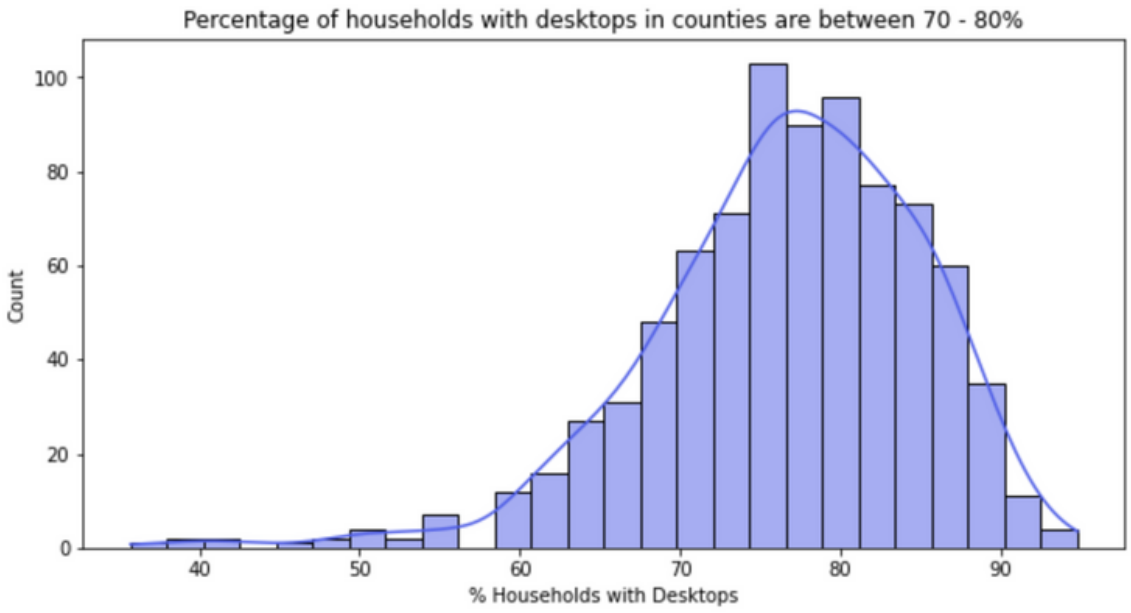
03

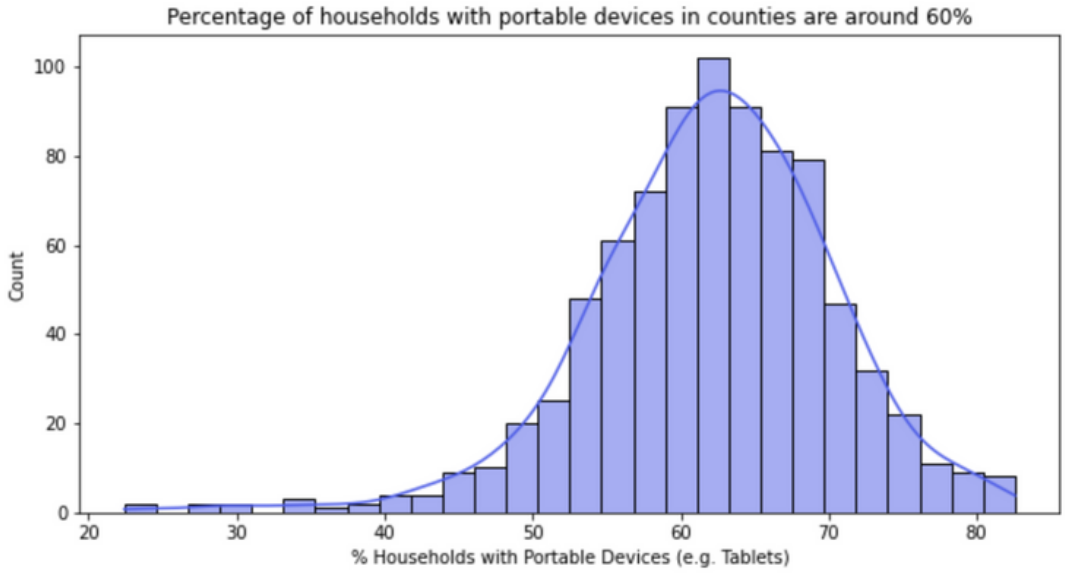
DEVICE TYPES

The dataset provided information on three different devices in the households: 1) Desktops, 2) Smartphones, and 3) Portables (e.g., tablets or iPads).

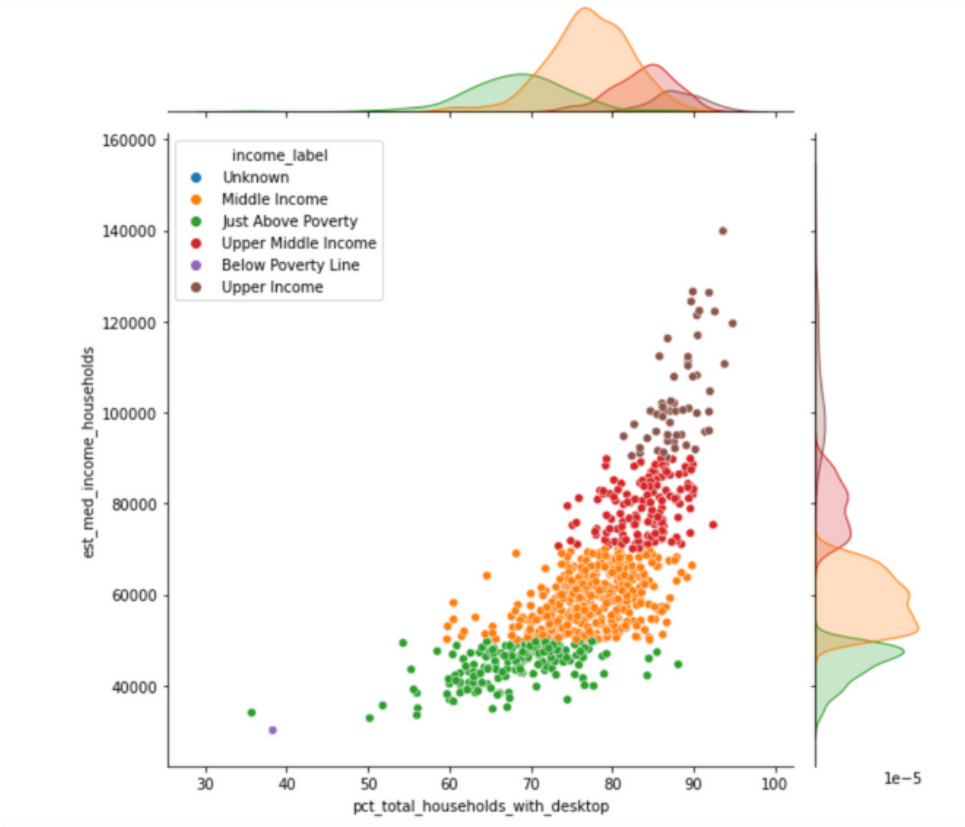
There was also information on households that did not have any devices in the home. We felt it was essential to prioritize desktops and portable devices, given that they have the most flexibility regarding what a person can do when accessing the internet.

Most counties across the countries have households where 70 - 80% have desktops in the home. Portable devices such as tablets come at a close second, around 60%.





Our device access exploration echoed what we saw when looking at other variables. In this case, the percentage of households with desktops increases as the average median household income increases.



Methodology + Analysis

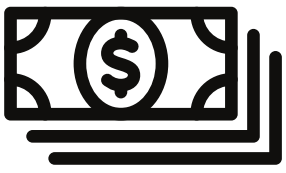
For our first approach, we created a multivariate comprehensive score to highlight the areas of need for resources. Based on our exploratory analysis, we felt it would be best to include the internet, technology, and income variables. We did not include unemployment status since we felt median income would account for the employment information.

Internet Accessibility Risk Score =
Broadband Internet Score + Income Score
+ Desktop Score + Portable Score + Smartphone Score

It is important to note that our total score measure has imperfections. The main being the missingness of survey data for rural and micro areas. Because of this, it makes it hard to accurately and fairly compare scores between all counties. Over half of the counties only had a score from the broadband internet scoring because that was all the information available. Our total score is the sum of all scores, with the lower number indicating a higher need. This way, we can at least indicate that even with a lower score, we can indicate both areas of need and areas that need more research conducted simultaneously.

In the future, we would like to develop an equation for each element differently. For example, a weighted score would consider the differences between a desktop versus a smartphone, where the desktop can be weighted more because of its versatility.

Broadband Score



Income Score

Device Scores

Broadband Internet Scoring

Broadband Percentage Score was assigned using percentage ranges defined by Census Bureau Data. Scores were assigned with 5 being the lowest percentage and 1 being the highest percentage.

- 1 = Broadband Percentage between 25-50
- 2 = Broadband Percentage between 50-60
- 3 = Broadband Percentage between 60-70
- 4 = Broadband Percentage between 70-80

Income Scoring

Income scoring is based on the U.S. Census poverty indicators. For our dataset, the county's largest average household size (non-family) was 4. We used the highest poverty line for a household of four and applied it to the entire dataset. The poverty line was for a household of 4 in Alaska.

- 1 (Below Poverty Line) = less than $31,380
- 2 (Just Above Poverty) = between $31,380 - $50,000
- 3 (Middle Income) = between $50,000 - $70,000
- 4 (Upper Middle Income) = between $70,000 - $90,000
- 5 (Upper Income) = greater than $90,000 + (not including NaN)

Device Scoring

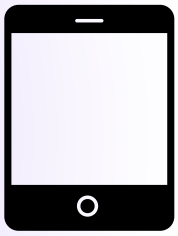
When scoring for devices, each device type had its own distribution and mean. Since these distributions were relatively normal, with some devices having a left skew, we generated the score to capture all reported values rather than treat those values that skewed the data left as outliers. We separated each scoring bucket by the number of standard deviations from the mean. Starting with values less than two standard deviations from the mean as a score of one, and if above two standard deviations away as a score of 5. We included smartphones in our score to have the most information for total score values.

Device Scoring

Device Scoring consists of four separate scores based on device ownership

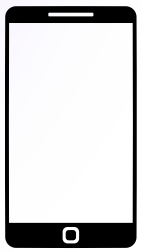
Desktop Scoring

1 = Desktop percentage between 50 - 59.5
2 = Desktop percentage between 59.5 - 68.0
3 = Desktop percentage between 68.0 - 85.1
4 = Desktop Percentage between 85.1 - 93.6
5 = Desktop Percentage greater than 93.6



Portable Scoring

1 = Portable percentage between 36.7 - 45.0
2 = Portable percentage between 45.0 - 53.3
3 = Portable percentage between 55.3 - 69.9
4 = Portable Percentage between 69.9 - 78.2
5 = Portable Percentage greater than 78.2



Smartphone Scoring

1 = Smartphone percentage between 66.2 - 71.9
2 = Smartphone percentage between 71.9 - 77.6
3 = Smartphone percentage between 77.6 - 89.1
4 = Smartphone Percentage between 89.1 - 94.8
5 = Smartphone Percentage greater than 94.8

No Device Scoring

1 = No Device percentage between (-4.192) - 0.096
2 = No Device percentage between 0.096 - 4.384
3 = No Device percentage between 4.384 - 12.96
4 = No Device Percentage between 12.96 - 17.248
5 = No Device Percentage greater than 17.248

Statistical Analysis & Predictive Modeling

Second, we built two linear regression models using factors like income, employment status, and demographic breakdown, to predict the percentage of households with broadband internet and desktops. Since there was a strong correlation between our variables, we chose a linear regression model as our method of choice for predictions. Additionally, we felt that the percentage of households with broadband internet and desktops were the best variables for our outcome predictions because of the importance in our scoring analysis.

Since the data on broadband in the household varied by income, region, and population class, we were curious if the location would be a significant factor in predicting our outcomes. There were similar thoughts with the income differences seeming to correlate with the percentage of internet and devices. We included these variables in our model to test our theories on their impact.

Given the level of missingness in our datasets with variables from the Census Bureau only accounting for a little over 800 counties out of the over 3,000 across the country, the outcomes of our predictions are not as generalizable as they could be for the country.

For our final models, both models are able to explain above 80% of the variance in the data. Additionally, both models had the estimated median income for households as a significant factor.

Predicting the percentage of households with broadband internet

It was surprising that neither region (except the South) nor pop class were not significant factors. The detail is good to know that area is not as important, however, we do believe the level representation of regions has had an effect on the significance in the model since there are limited regions and areas represented in the dataset.

OLS Regression Results

Dep. Variable:	broadband_pct	R-squared:	0.821
Model:	OLS	Adj. R-squared:	0.819
Method:	Least Squares	F-statistic:	526.5
Date:	Mon, 28 Mar 2022	Prob (F-statistic):	3.29e-295
Time:	09:32:34	Log-Likelihood:	-1967.8
No. Observations:	812	AIC:	3952.
Df Residuals:	804	BIC:	3989.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	28.4962	1.360	20.952	0.000	25.827	31.166
region[T.Northeast]	-0.5199	0.316	-1.646	0.100	-1.140	0.100
region[T.South]	-0.5574	0.249	-2.237	0.026	-1.047	-0.068
region[T.West]	-0.2768	0.325	-0.852	0.394	-0.914	0.361
est_med_income_households	6.313e-05	9.77e-06	6.460	0.000	4.39e-05	8.23e-05
est_unemp_pop_ratio_16_over	0.0625	0.063	0.989	0.323	-0.062	0.187
pct_total_households_with_desktop	0.4449	0.027	16.544	0.000	0.392	0.498
pct_total_households_with_portable	0.2167	0.027	7.904	0.000	0.163	0.271

Omnibus:	216.157	Durbin-Watson:	1.940
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1143.655
Skew:	-1.102	Prob(JB):	4.56e-249
Kurtosis:	8.380	Cond. No.	9.10e+05

Predicting the percentage of households with desktops

When predicting the percentage of households with desktops, the unemployment rate became a significant factor in addition to median income and internet in the household.

OLS Regression Results

Dep. Variable:	pct_total_households_with_desktop	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.839
Method:	Least Squares	F-statistic:	704.3
Date:	Tue, 15 Mar 2022	Prob (F-statistic):	2.67e-316
Time:	12:33:58	Log-Likelihood:	-2083.1
No. Observations:	812	AIC:	4180.
Df Residuals:	805	BIC:	4213.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.6002	2.034	3.244	0.001	2.607	10.594
pop_class[T.Medium Metro Area]	-0.1838	0.500	-0.368	0.713	-1.165	0.797
pop_class[T.Small Metro Area]	-0.4307	0.489	-0.881	0.379	-1.391	0.529
est_med_income_households	2.985e-05	1.14e-05	2.614	0.009	7.44e-06	5.23e-05
est_unemp_pop_ratio_16_over	-0.3719	0.069	-5.364	0.000	-0.508	-0.236
broadband_pct	0.5866	0.035	16.719	0.000	0.518	0.655
pct_total_households_with_portable	0.3795	0.029	12.922	0.000	0.322	0.437

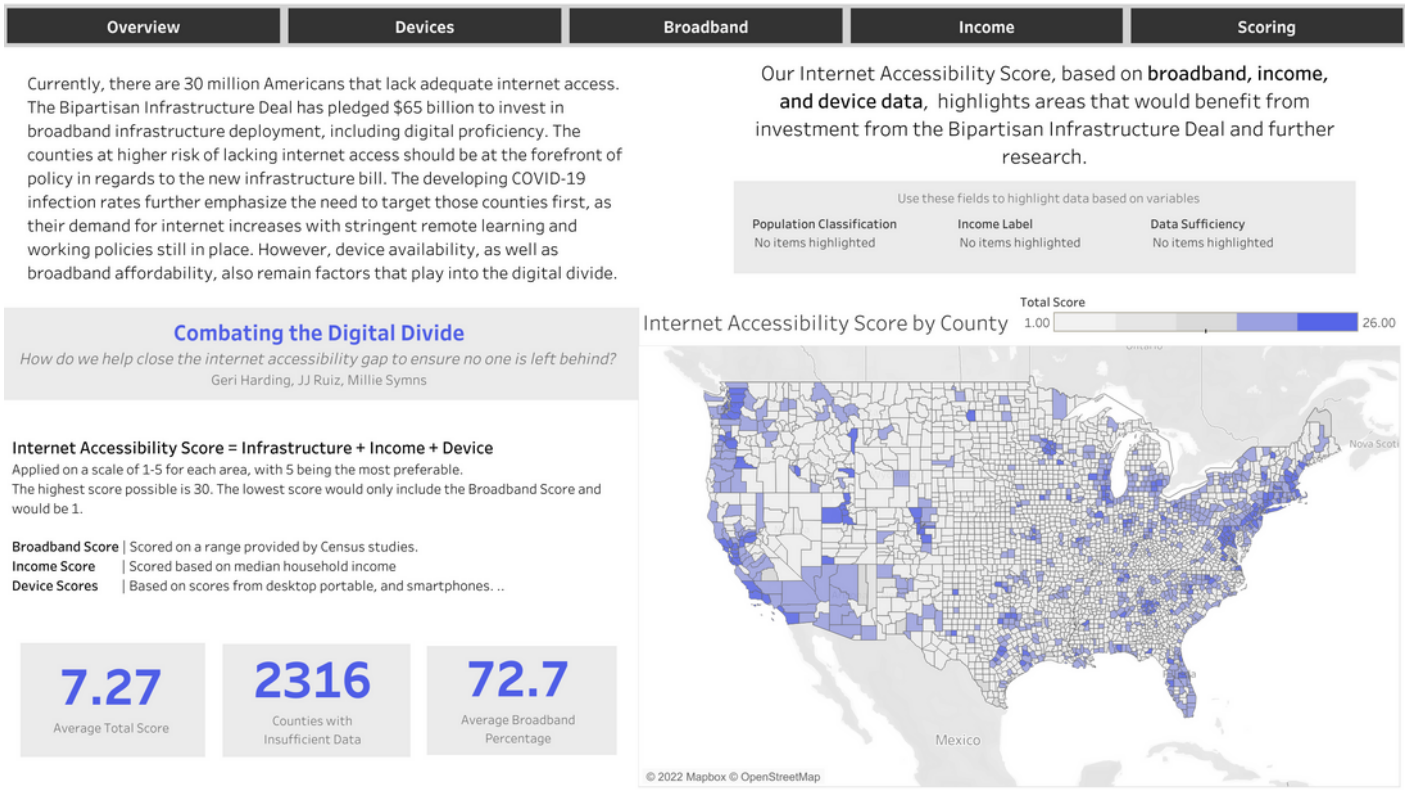
Omnibus:	59.995	Durbin-Watson:	1.679
Prob(Omnibus):	0.000	Jarque-Bera (JB):	179.093
Skew:	-0.331	Prob(JB):	1.29e-39
Kurtosis:	5.203	Cond. No.	1.19e+06

Dashboard

Using our datasets and scoring analysis, we created an interactive dashboard for users to explore the information by region, population classification, state, and county.

Our dashboard can be helpful for businesses like Apple to distribute device donations to communities in need, schools and organizations planning to implement e-learning opportunities, and local and federal governments to survey the state of internet accessibility in their constituency. It allows the user to interact with the dashboard map in overview to view each county's scorecard.

In the dashboard, there are five pages the user can explore: 1) Overview, 2) Broadband, 3) Devices, 4) Income, and 5) Scoring.



Click here to visit the interactive dashboard

1. **Overview:** Provides an overview of the project and the internet accessibility risk score on a map by highlighting population classification, income label, and data sufficiency.
2. **Broadband**: Focuses on the intersection between income, population, and region while allowing the user to filter the data by either state or region.
3. **Devices:** Provides a breakdown of desktop, portable, or smartphone device ownership levels or lack of across population classification and income labels.
4. **Income:** Highlights the disparity between device ownership and broadband levels.
5. **Scoring:** Provides visualizations to observe the distribution of the internet accessibility score, focusing on population classification. Score distribution is highly varied based on income level.

Potential Audiences

- Policymakers in the U.S. Government
- State and local governments preparing plans for Infrastructure Bill allotment
- The general population who is interested in the issue of access
- Companies relating to internet and technology who are looking for corporate social responsibility initiatives to start related to the topic of access
- Companies and/or nonprofits interested in upskilling workers from underprivileged backgrounds
- The Department of Education representatives interested in seeing how internet access could relate to their system

Conclusions

After assessing the needs of each county and ranking their internet accessibility, through data on the percentage of households with broadband internet, device access, and average median income, we now have a better sense of what the digital divide looks like across the country.

Our dashboard provides insights for all interested parties to visualize and interpret where leaders can do work to minimize the digital divide. Furthermore, our exploratory data analysis and predictive modeling also provide insights into critical factors that influence this issue.

Some overall takeaways from our project:

- **Conduct more research** – There was not enough data in rural and micro areas related to computer devices and other important demographic factors. Based on our research, rural areas are often trailing behind other regions in the country regarding resources.

- **Income influences access to resources** – In our predictive models, the average median household income plays a significant factor in predicting the percentage of households with broadband internet and desktops. With the internet being a necessity for work, schooling, telehealth, and other resources, the goal is to move away from income being a driving factor in having access to basic needs. This information further supports the need for the infrastructure bill to prioritize areas where the median income is low.

- **Funding priorities should go to the south and midwest regions** - While there were limitations to our internet accessibility scoring method, with the data we do have at hand, we recommend priority should be given to counties in the south and midwest regions for infrastructure funding. The south region was a significant factor when predicting the percentage of households with broadband internet, meaning that this region is significantly different from the rest. Additionally, the south and midwest regions also have a proportional majority of the rural and micro areas, which fall behind based on broadband internet connectivity compared to other regions.

Future Work

While we are confident in our findings we believe future work can always be done to build upon our solution.

One idea is to include the cost of infrastructure that enables internet and pricing models from companies. With this level of data, we would be able to identify what factors or inefficiencies drive the cost of the internet to reduce the cost of setup. Likewise, with data on pricing models, we would see which companies behave with predatory price inflation, ideally helping users of our tool find better and more affordable options.

Another would be including data on connectivity speeds from internet providers as a factor in our scoring. Regarding internet speeds as a necessity, the FCC's current definition of the minimum required speed is 25/3 Mbps. While this may be acceptable for a household consisting of only one person, the requirements to support home broadband vary drastically depending on the number of people in the household and how residents use bandwidth. For example, in a family of five, people video conferencing for work or school throughout the day may require far more downstream and upstream bandwidth than the current 25/3 Mbps threshold.[3]

We would also explore weighting in our scoring so that each element is not of equal standing. We can have a heavier weight on the percentage of households with broadband internet and desktops in the future.

Lastly, we would have a more in-depth data analysis to include demographic background to see what communities are being affected by the digital divide. In other systematic structures, not all groups are affected equally. Looking into any potential disparities by race, gender and age would allow those most afflicted by the digital divide concrete evidence and a voice in this uphill battle.

References

[Footnote 2] Chao , Becky, et al. “The Cost of Connectivity 2020.” *New America*, July 2020, https://www.newamerica.org/oti/reports/cost-connectivity-2020/focus-on-the-united-states/.

[Footnote 3] Fritz, Jack, and Dan Littmann. *Broadband for All: Charting a Path to Economic ... - Deloitte*. Apr. 2021, https://www2.deloitte.com/content/dam/Deloitte/us/Documents/process-and-operations/us-broadband-for-all-economic-growth.pdf.

[Footnoote 1] McNally, Catherine. “Nearly 1 in 4 Households Don't Have Internet-and a Quarter Million Still Use Dial-Up.” *Reviews.org*, 18 Oct. 2021, https://www.reviews.org/internet-service/how-many-us-households-are-without-internet-connection/.

Reach out with questions