**Ten most suitable counties for mobile broadband technology in U.S**

**Team members**: [Cesar Mosquera](https://app.slack.com/team/ULVMYVBJ8) , Emi Babu, [Jacqueline McBean-Blake](https://app.slack.com/team/ULX4GF25T) and Pratikchhya Upadhyay

**Project Scope:**

This project intends to analyze the most suitable locations to invest in mobile broadband Network Technology in the United States. Our model uses variables such as location, income, population, age, per capita income, and the number of households with cellular data to identify the target locations ideal for building or improvising the current broadband cellular network.

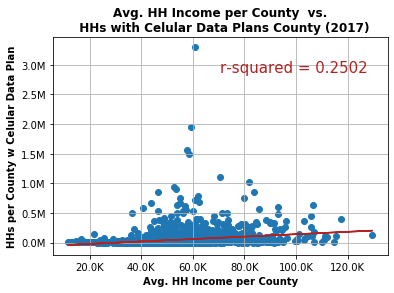
**Data Cleaning Process:**

During our preliminary analysis, we examined the datasets from a number of sources, such as, Census Bureau, U.S. Bureau of Labor Statistics, Goberning.com. Rows with missing values and irrelevant columns were dropped from the master set. During our investigation, we found that the variable, Metropolitan Area was a common variable on all datasets. Our initial intent was to use this variable to join all relevant datasets and extract the relevant columns. Unfortunately, we found that this method could be used to load only 76 rows to the dataframe. Moreover, the resulting plots did not appear significant. Therefore, this approach had to be abandoned.

Our alternate approach was to use the google API to extract the dataset. In this case, we were able to retrieve all relevant dataset at the county level. Further, we used similar analysis and cleansing techniques that we had earlier perfected for our preliminary dataset. First, we renamed the columns, so that their names would reflect more relevancy towards our project. Second, we dropped variables that were irrelevant to our project. Third, we deleted all missing observations.

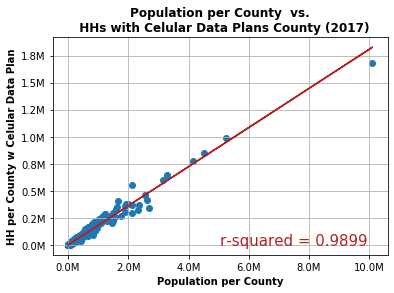
**Analysis:**

Multiple, scatter plots were created to compute the correlation among the variables. Our first analysis was to find, if there is a relationship between household with cellular data plan by county and average household income by county.



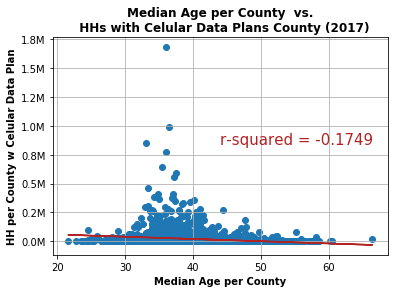
In the above scatter plot, the x-axis is the average household per county and the y-axis is the households with cellular data plans in county. A preliminary regression showed an r-squared value 0.2 indicating a weak correlation between the two variables. N was 3,220(Rows). It could be seen that increase in income has no relationship with households with cellular data plans.

On our second analysis, we wanted to find if there was any relationship between population by county with household per county with cellular data plan.



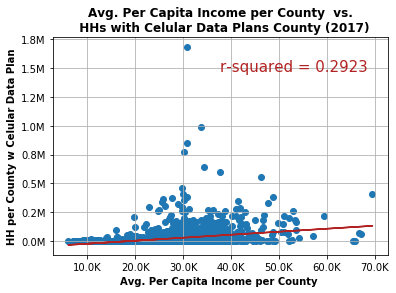
In the scatterplot the x-axis represents population per county and y-axis represents household per county with cellular data plan. A preliminary regression shows a strong correlation between those variables. R-squared value was also significantly high at 0.98, suggesting that the number of cellular data plans increase with population.

Similarly, the third analysis consisted of comparing median age per county with household per county with cellular data plan.



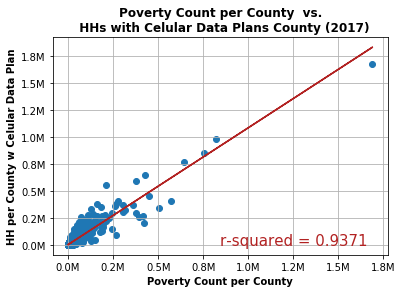
In the scatterplot the x-axis represents the median age per county and y-axis represents household per county with cellular data plan. A preliminary regression shows a negative, and weak correlation between those variables. R-squared value was -0.17, suggesting that households with members that call between the age groups of 30 to 50 have a higher chance of having a cellular data plan.

Our fourth analysis consisted of comparing Average Per Capita Income per county with cellular data plan and household per county with cellular data plan.



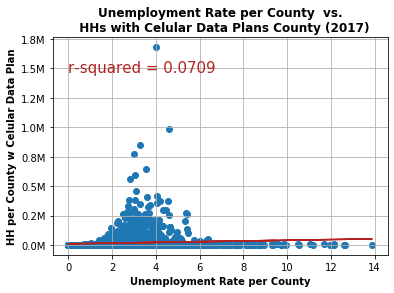
In the scatterplot the x-axis represents average per capita income per county and y-axis represents household per county with cellular data plan. A preliminary regression shows a weak correlation. The R=squared was .0.29 signifying that increase in per capita income is not correlated with increase in number of households with cellular data plans.

Our fifth analysis consisted of comparing poverty count per county and household per county with cellular data plan.



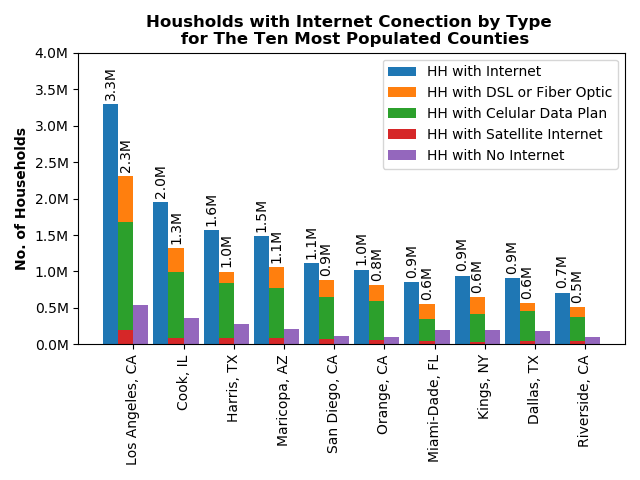
In the scatterplot the x-axis represents poverty count per county and y-axis represents household per county with cellular data plan. A preliminary regression shows a strong correlation. The R-squared was 0.93 signifying that as poverty increases in a given county the number of households with cellular data plan proportionally increases.

Our last analysis consisted of comparing unemployment rate per county and household per county with cellular data plan.



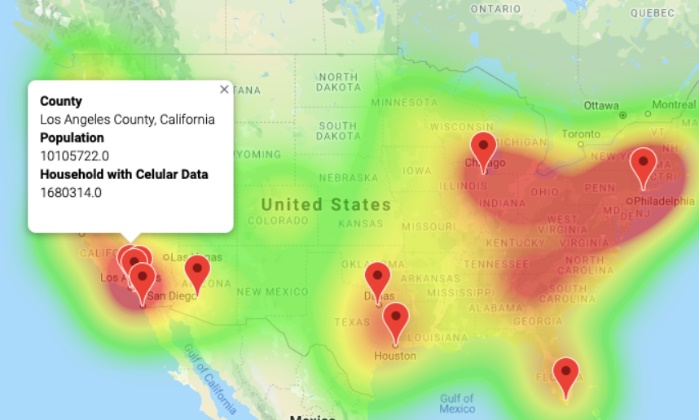
In the scatterplot the x-axis represents unemployment rate per county and y-axis represents household per county with cellular data plan. A preliminary regression shows a poor correlation. The R-squared was 0.07 signifying that unemployment rate in a given county has no relation with the number of households with cellular data plan.

At the end, we created a bar graph showing ten best counties to invest on mobile broadband network based on population and internet use.



The bar graph represents the number of households in y-axis mapped against ten counties that we think are best suited for further investment in mobile broadband technology. Each bar also represents the county population where color codes have highlighted to represent types of available internet.

Finally, we created a heat map to show the ten best counties suitable for investment in mobile broadband technology.



Each marker represents the high population county, where investment in mobile broadband technology would be most suitable.

**Conclusion:**

There seems to be a positive relationship between household income and household with cellular data plan. Similarly, individual relationships between household with cellular data and per capita income, population per county, unemployment rate per county, poverty count per county also seems to be positive. However, the relationship between median age per county and household with cellular data plan per county seems to be negative. Finally, the heat map demonstrates the ten best counties where investment in mobile broadband technology would be most suitable. The ten best counties are:

Los Angeles County, California

Cook County, Illinois

Harris County, Texas

Maricopa County, Arizona

San Diego County, California

Orange County, California

Miami-Dade County, Florida

Kings County, New York

Dallas County, Texas

Riverside County, California

**Reference:**

**Source: U.S. Bureau of Labor Statistics**

[**https://www.bls.gov/web/metro/laummtrk.htm**](https://www.bls.gov/web/metro/laummtrk.htm)

**Source: U.S. Bureau of Labor Statistics**

[**https://www.bls.gov/web/metro/metro\_oty\_change.htm**](https://www.bls.gov/web/metro/metro_oty_change.htm)

**Source: U.S. Bureau of Labor Statistics**

[**https://www.bea.gov/news/2018/gross-domestic-product-metropolitan-area-2017**](https://www.bea.gov/news/2018/gross-domestic-product-metropolitan-area-2017)

**Source: U.S. Census Bureau**

[**https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk**](https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk)

**Source: Goberning.com**

[**https://www.governing.com/gov-data/city-internet-connection-household-adoption-rates-data.html**](https://www.governing.com/gov-data/city-internet-connection-household-adoption-rates-data.html)

**Source: U.S. Bureau of Labor Statistics** [**https://data.bls.gov/cew/apps/table\_maker/v4/table\_maker.htm#type=3&year=2018&qtr=3&own=0&ind=10&supp=0**](https://data.bls.gov/cew/apps/table_maker/v4/table_maker.htm#type=3&year=2018&qtr=3&own=0&ind=10&supp=0)

**Source: Google API, Census API**