DSS Datathon

Presentation Speaking Notes:

(5-7 min presentation + ~3 min Q/A):

Intro (Saumya)

- Factors: According to an article which studied healthcare rankings on a county level published by the College of William and Mary, there are 4 primary factors that affect availability

Data & Limitations (Angela)

- We went about answering our problem statement by collecting data for the 4 categories, using PCA and K-means clustering to cluster counties with similar behavior
- Data comes from US census, county level data compiled by a group at UC Berkeley, AQI data from United States Environmental Protection Agency
- Granularity of data (county and why county level)
- We decided to work with public health insurance data since individuals who are on private health insurance are more likely and able to get medical treatment in hospitals. Since we are striving for accessibility, pop-up clinics for uninsured)
- The slide shows the features we used to model the 4 categories

EDA (Tarsus)

- Since we are using public health insurance data, it makes sense that employment is so strongly negatively correlated with coverage. Counties with higher employment rates are likely to have higher rates of private insurance.

K-means model (Arin)

- K means is an unsupervised learning algorithm that partitions data into clusters that share similar characteristics
- We first used PCA to decompose the data into two principal components, allowing us to find the most relevant features that separated the clusters as well as allowing for us to visualize the clusters in two dimensions.
- Used k means to cluster U.S. counties into clusters, each of which share the same “health level” as determined by the four factors that affect availability.

Results & Next Steps (Ruhi)

- This dataframe shows the aggregated median values for each of our 4 clusters, and for all the features included in our model
- We visually compared these to determine the cluster which has the greatest need for a pop-up clinic
- As you can see in the first row, we observed that cluster 0 has the lowest values for employment, hospitals, diabetes percentage, and median AQI

- Thus, based on our 4 criteria factors (health behaviors, clinical care, socioeconomic levels, physical environment), we found that cluster 0 would benefit the most from pop-up clinics as it has the lowest values for employment and hospitals, poorest air quality, and highest values for diabetes percentage
- However, that cluster 0 also appears to have the highest coverage rates.
    - effect may have been caused by the negative correlation between employment rates and coverage
    - we would investigate further in future work

Q: Negative values because the data has been normalized? But there are values > 1 (due to strongly right-tailed dist)

Future work

- Density of hospitals as related to population beyond raw count
- Included private insurance for a more comprehensive picture of coverage
- Additional proxies for health behaviors, such as mortality rates, heart and lung disease prevalence, income, age and other demographic factors
- Environmental factors may even include manufacturing plant locations, county wealth/income measures, weather / precipitation (which may affect the frequency with which residents can exercise outdoors or travel for a medical checkup
- All to create a clearer picture of counties that would most benefit from pop-up clinics

Thank you for listening, we are happy to take any questions now.

Q: The most important feature was employment, but then the next most important was coverage and diabetes percentage basically tied, followed by the others.

Center project around these values (values of DSS):

Accessibility
Diversity
Integrity
Collaboration

Alexandra's Presentation

My Initial big Ideas for Problems to solve:
Public policy decisions
Funding allocation

Tip - look into external data somehow!

Census data is used for:

- infrastructure investment: Where to build new schools, parks,
- journalists/media use it to conduct accurate polls
- help older Americans w housing related stuffs
- businesses use to understand demographics of customers
- economic development needed in more rural areas
- transportation needs: (what time do you leave your home to go to work?)
    - inform planners where stoplights, bus stops should be to make more efficient...

More info:
American community survey: what to look at for trends! (taken every year)
-housing data, etc. all the bread and butter for the population
-emergency and management map to mash data from Dif sources to form one solution

Data collected from business owners every 5 years
-businesses w/o employers

resources:
-Census academy

Alexandra's tips on how to start a project:

-Pull a data profile from some community of interest, might find very unique things
-start w own interest like clean oceans, then figure out how to use census data to solve this problem

Ana-Maria's Presentation

Census infrastructure onboarding:
data.census.gov
-census geography, customize views, multiple searches
-keep in mind geography hierarchy
-rest in the slides

Team Meeting 1: 11/13/2020

Brainstorming:

- Hospital funding over time
- Interesting insights of metrics across counties
    - How does health insurance coverage for people over age 65 across counties?
- Health insurance coverage - "marginalized" communities?
    - Correlation of people not represented by data
- Disability, mental, elderly
- Time series model
- College grad employment?
- Correlation between grocery/retail store with proximity to transportation service or housing
    - To maximize revenue
    - Need to normalize revenue
    - Or vice versa (transportation -> near stores)
    - Accessibility vs. usability
- Given features like SAT score, location, etc, predict college admission
- COVID risk vs healthcare coverage, elderly population, other features? In CA counties or across US
- Sustainable waste management systems - where to set these up? How does census data help solve this?

- California coastal danger, wildfires, Deforestation, CO_2 emission. How does census data help solve these problems?
    - Evacuations, where to set up evacuation centers?
        - Features to consider:
- http://www.pinchofintelligence.com/looking-back-at-my-first-datathon/
    - Inspiration lol
    - They tried to determine which areas are best to set up Marine Protected Areas in order to solve the problem of sustainable fishing. E.g. "A mangrove forest or a coral reef (both good places for food and breeding) are positive influences, while places near cities are bad, as the water is more polluted."
-

Environmental/Health

Current Plan

Problem Statement: Given a fixed budget, which counties in the United States would benefit the most from increased funding towards healthcare accessibility?

Approach: Look at the association between health insurance coverage and air quality and other metrics, and see how strong associations are. Then, use these associations to build a clustering model to determine the optimal counties to provide funding for.

Problem we're running into: how to define the counties that would benefit the most?
-we found that employment as strong negative association to public health insurance coverage, but how do we decide which counties to give funding to? The counties with lots of ppl with low coverage but high employment, or counties with ppl that have high coverage but low employment?

-look at the number of people who are neither publicly nor privately covered instead.
- Came across negative numbers (people who have both public and private insurance coverage) and data from different sources

Problem:

- For people with no health insurance coverage, hospital treatment is very expensive. Pop-up clinics can be a way to receive free treatment. Given a constraint on the number of pop-up clinics we can open in a state, can we recommend locations to open up pop-up clinics based on features such as health insurance coverage, median income, age demographics (larger proportion of seniors?), disability, education level,

Approach:

- Use clustering to determine clusters of counties where each cluster represents counties that have similar level of needs for healthcare accessibility

Sources:
https://github.com/Yu-Group/covid19-severity-prediction/blob/master/data/readme.md

https://www.wm.edu/as/publicpolicy/schroedercenter/for-faculty/Downloadable%20Health%20Datasets/County%20Level%20Downloadable%20Health%20Datasets/index.php

- 4 categories: health behaviors, clinical care, social and economic factors, physical environment
- Decided to work with public health insurance because those with private health insurance are more likely and can afford to go to hospitals
- We tried KMeans Clustering and found 3 clusters to be optimal. However, it was a challenge to interpret the results
    - Counties within a cluster are similar in behavior, while counties across clusters are different
    - Could not find meaningful insights from the 3 clusters
    - https://realpython.com/k-means-clustering-python/
- Supervised versus unsupervised learning?

https://data.census.gov/cedsci/table?text=insurance%20rates&t=Health%20Insurance&g=0100000US.04000.001&tid=ACSSE2017.K202702&hidePreview=false

- Over time, and for public too

https://data.census.gov/cedsci/table?text=insurance%20rates&t=Health%20Insurance&g=0100000US.04000.001&tid=ACSSE2017.K202701&hidePreview=false

- By age

https://data.census.gov/cedsci/table?text=insurance%20rates&t=Health%20Insurance&g=0100000US.04000.001&tid=ACSST1Y2017.S2702&hidePreview=false

- More demographics

https://data.census.gov/cedsci/table?text=insurance%20rates&t=Class%20of%20Worker&g=0100000US.04000.001&tid=ACSDP1Y2019.DP03&hidePreview=false

- Selected economic status

https://data.census.gov/cedsci/table?text=insurance%20rates&t=Housing&g=0100000US.04000.001&tid=ACSDT1Y2019.C27021&hidePreview=false

- Housing and insurance

https://www.ehealthinsurance.com/resources/individual-and-family/what-is-private-health-insurance

- Private versus public health insurance
- We're observing individuals with both private and public insurance
- Categories of jobs with people who have private/public/both/none
-

Brainstorm features instead of water/air quality:

- Access to plumbing
- Want Other accessibility things that indicate the accessibility of a place

- https://data.census.gov/cedsci/table?text=insurance%20rates&g=0400000US06.050000&tid=ACSSE2019.K202702&hidePreview=false (Per county CA)
- https://data.census.gov/cedsci/all?text=insurance%20rates&g=0400000US06.050000
- https://www.kidsdata.org/topic/526/environment-waterquality/table
- https://data.ca.gov/dataset/water-quality-data/resource/5eeaf27e-315d-4b95-9c34-cf63d168e8f5 (water quality)
- https://data.ca.gov/dataset/asthma-hospitalization-rates-by-county (asthma hospitalization rates by county)
- https://data.census.gov/cedsci/table?t=International%20and%20Domestic%20Migration&g=0100000US.050000&tid=ACSST1Y2019.S0701&hidePreview=false (migration in and out of county, how important health care is to people?)
- https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual (AQI per county US)
- https://data.census.gov/cedsci/table?q=income%20and%20poverty&tid=ACSST1Y2019.S2303&hidePreview=false (work status in past 12 months)
- https://data.census.gov/cedsci/table?q=income%20and%20poverty&tid=ACSST1Y2019.S1901&hidePreview=false (income in past 12 months)
- https://data.census.gov/cedsci/table?q=income%20and%20poverty&tid=ACSST1Y2019.S1702&hidePreview=false (poverty status in past 12 months)

Presentation Ideas

Final presentation can be in **any** form:
https://www.notion.so/Group-Presentations-3dee8055a473404984699808 1ec10321

Ideas:

- Show some plots of association between various variables
-

Have a map of US to visualize the counties of interest