**Charissa Tagupa**

**Week 6**

**GitHub:** [charissa58/Stateside-Data-Searchers: We will explore and analyze the "New York Health Rankings" factors that influence longevity and premature deaths. (github.com)](https://github.com/charissa58/Stateside-Data-Searchers)

Summary Document:

Scrum Master for next week: We all were.

List at least 5 things the team did well and will continue doing:

Coding, Analysis, Meeting, Edits and give feedback.

List at least 3 things the team did poorly and how you will mitigate them next sprint:

Practice more.

List shout-outs to any team members for excelling in any way:

Although we were all so busy with our personal and professional life, we did our best to meet and complete this final project.

What did you learn as a team this week?

That we work well as a team and that these skills will help develop our confidence in data science.

What did you learn as an individual this week?

That data science can be flawlessly done as a team, as long as we make time for it.

**Summary to Final Project:**

**Selected Topic:**

Longevity, and influencing factors, in New York State

**Question for this analysis:**

Which of the investigated factors, exhibit the greatest influence on longevity, on a county basis?

**Reason For Topic Selection:**

The topic for this analysis was an interesting one to apply a machine learning model to investigate clustering and corelated factors that affect lifespan in different areas.

**Data Source Description:**

Data was sourced from this location: <https://www.countyhealthrankings.org/app/new-york/2022/downloads>

The dataset contains data for each county in New York State, and provides annual data for a variety of factors linked to longevity (e.g. obesity rate, food environment index etc..). Datasets are available from 2011 to 2022,

Of the listed data, the following variables are to be analyzed:

* Deaths
* % Poor Health
* % Low Birthweight
* Food Environment Index
* Number of Driving Deaths
* Number of Uninsured People
* Mental Health Provider Ratio

**Question We Hope To Answer With The Data:**

Which of the investigated factors, exhibit the greatest influence on longevity, on a county basis?

Hypothesis: We predict that in urban areas, the death ratio is much lower than other areas.

Second Hypothesis: It is predicted that rural areas exhibit a greater excess death ratio, due to lower population densities and reduced availability of services.

**Data Exploration Phase:**

A variety of machine learning models were tested against the available data to assess the best model fit. Supervised Learning and Unsupervised Learning were both explored as candidates; the final selection and classification was completed using Primary Component Analysis and Hierarchical Clustering

**Machine Learning Model:**

**Pre-processing Data**

* For our data pre-processing we had to do some cleaning before using in any ML model so it could run without error. We first join all the data collected from the web into one CSV file and then clean up the columns and rows with null or zero values. Finally, we grouped data by county and took the mean to then scale out the data.

**Feature engineering and Selection**

* Our preliminary feature are the factors that may affect a premature death, we want to find out which of the features (factors) has a higher effect on the cause of premature deaths. The data collected broke down by NYS county's, we were curious to analyze which feature had a stronger correlation with the death rate. We wanted to analyze over the years but data would not work well with any machine learning models. We applied a linear and logistic regression model,that didn't not give us much information about the data and the accuracy on the model was 0%. After applying the LR model we grouped the data by county with the mean, then scaled data points to ratios. We then ran a PCA and applied the K-mean model.

**Spliting and Training data**

* For our ML model selection the linear regression didn’t call for a split. When we applied a Logistic regression model, we applied a split by making "X: Food environment index"" and "y: death total." We stratified the data to 60 training 40 test.

**Machine Learning Model Decision**

* We originally thought the linear regression model would work with our data set but it didn't work well without data. There was no correlation or patterns observed with this model. The data point was all scattered and the linear regression was almost flat for all input labels. This led us to apply a PCA and then do a K-mean model to instead cluster our data point for each feature. After applying this model, we did observed patterns among our data points.
* When we reference our question "Which of the investigated factors, exhibit the greatest influence on longevity, on a county basis?" the ML model does indeed answer it. We can see per county which of the factors affects the most on the death ratio.

**Output feature:**

* death ratio

**Input features:**

* Fair or Poor Health %;
* Low birthweight %;
* Food Environment Index;
* Driving Deaths ratio;
* Uninsured ratio;
* Mental Health Provider ratio

**Database**

After cleaning and merging the County Health Ranking of NY from past 10 years on Jupyter Notebook, we used the data to calculate the ratios as comparing the health factors to the census population of 2020. These dataset and tables are stored on PostgreSQL and connected with Machine Learning model via SQLAlchemy.

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Graphical user interface, application

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**Analysis Summary:**

The parameters under investigation were divided into quartiles (Low, Below Average, Above Average, High), and sorted by Death Ratio. This provided a high-level view of class commonalities. These groupings are shown on in our dashboard and slide deck.

Dashboard and Visualizations:

Tableau shall be used for visualizations, this may be used in a standalone manner or embedded in a webpage, pending final design. This may read from the DB or via results files generated.

Additionally, excel may be used for initial data exploration

**Technology Flow:**

Diagram

Description automatically generated