Initial Work Log:

Week 5 (Sep 23-29)

- Finalize research questions and objectives
- Draft project outline

Week 6 (Sep 30-Oct 6)

- Conduct a literature review on autoimmune diseases and natural therapies
- Summarize key findings and relevant studies

Week 7 (Oct 7-13)

- Identify and gather datasets
- Create a list of data sources and begin preliminary data collection

Week 8 (Oct 14-20)

- Complete data collection
- Prepare a consolidated dataset for cleaning

Week 9 (Oct 21-27)

- Begin data cleaning and preprocessing
- Produce a cleaned and standardized dataset

Week 10 (Oct 28-Nov 3)

- Conduct exploratory data analysis (EDA)
- Generate a report summarizing key insights and trends

Week 11 (Nov 4-10)

- Develop predictive models using regression and classification techniques
- Create initial model prototypes

Week 12 (Nov 11-17)

- Train and tune the models
- Prepare optimized models for testing

Week 13 (Nov 18-24)

- Test the models and evaluate their performance
- Compile an evaluation report with metrics

Week 14 (Nov 25-Dec 1)

- Analyze results and draw conclusions
- Draft the final report outlining findings and recommendations

Week 15 (Dec 2-8)

- Finalize the project report and prepare presentation materials
  Ensure everything is ready for submission

New Work Log:

Week 2 (Sep 2-8)

- Brainstorm on what interests me and what I would like to do for my research

Week 3 (Sep 9-15)

- Met with Professor Teresa Amsbury, came up with the topic "The Prevalence of Autoimmune Diseases in Women and Their Turn to Natural Therapies Due to Dissatisfaction with Conventional Medical Care"

Week 4 (Sep 16-22)

- Completed Initial Research Proposal on September 19th
- Sent proposal to Dr. Erisman

Week 5 (Sep 23-29)

- Reach out to Professor Teresa Amsbury to discuss the proposal and databases datasets
- Added Teresa as another advisor
- Scheduled to meet with Professor Christine Borden-King-Jones on Friday 9/27 to explore research methodology for the capstone project.
- Conducted additional research on natural therapies
- Connected concepts from SOC 278 to strengthen my capstone project.
- Planned to search for relevant datasets next week to support the research analysis.

Week 6 (Sep 30- Oct 6)

- Met with Professor Amsbury, gave me advice on my proposal such as:
  1) Be concise. Research papers should not be wordy. For example, your abstract should be about 200 or less words.
  2) You make excellent statements for your argument. These should be supported with citations. Use citations for all scientific claims.
  3) Avoid speaking in 1st person. Keep the paper neutral.
  4) Your writing is excellent. However, just as every initial paper does, there are some minor grammatical and structural errors. I can touch more on that tomorrow.
- Professor Amsbury sent me additional articles to research to solve my problem statement
- Met with Dr. Borden-King-Jones (SOC 278) – gave me advice on how to search in a database for best results
- Mentioned using other words like inequality, sexism, inequity, not just gender bias for example

- She said “I think there is actually just very little on your specific overlap of interests (which is good—all the more reason to research it! –but also frustrating because there isn't as much out there)”
- Research which datasets are good for my research topic (ImmPort, UK Biobank, GEO, ADGC)
- Proposal – Change

Week 7 (Oct 7 – Oct 13)

- Met with my advisor to discuss my research.
- She provided databases where I can find datasets related to my topic (CDC, WHO) also suggested to look at (EU, UN, UK)
- .
- She mentioned she would send more resources and would connect me with someone to see if I can access hospital records related to autoimmune diseases.
- Worked with a dataset this week showing statistics on women with autoimmune diseases.
- The dataset has a higher number of female patients than male, aligning with the higher prevalence of autoimmune diseases in women.
- Satisfaction was rated on a scale from 1 to 5 in the dataset.
- Patients not receiving conventional treatment had a slightly higher average satisfaction (3.15) compared to those receiving conventional treatment (3.02).
- Made some changes to my proposal, but saved the first one

Week 8 (Oct 24 – Oct 20)

- Searched clinicaltrials.gov and found the VITamin D and OmegA-3 TriaL (VITAL; NCT 01169259), a randomized clinical trial involving 25,871 U.S. men and women. The study investigates whether daily supplements of vitamin D3 (2000 IU) or omega-3 fatty acids (1 gram of Omacor® fish oil) reduce the risk of cancer, heart disease, stroke, and autoimmune diseases.
- Participant Demographics:
- 25,871 participants were enrolled and followed for a median of 5.3 years.
- 18,046 self-identified as non-Hispanic white, 5106 as Black, and 2152 as other racial and ethnic groups.
- The mean age of participants was 67.1 years.
- Study Results:
- In the vitamin D arm, 123 participants in the treatment group and 155 in the placebo group had confirmed autoimmune disease (HR 0.78, 95% CI 0.61 to 0.99, P=0.05).
- In the omega-3 fatty acids arm, 130 participants in the treatment group and 148 in the placebo group had confirmed autoimmune disease (HR 0.85, 95% CI 0.67 to 1.08, P=0.19).
- Compared to the placebo group (88 confirmed cases of autoimmune disease), results showed reduced cases in those receiving both vitamin D and omega-3 (HR 0.69, 95% CI 0.49 to 0.96), vitamin D alone (HR 0.68, 95% CI 0.48 to 0.94), and omega-3 alone (HR 0.74, 95% CI 0.54 to 1.03).

- Conclusions:
    - Vitamin D supplementation for five years, with or without omega-3 fatty acids, reduced autoimmune disease risk by 22%.
    - Omega-3 fatty acids reduced autoimmune disease by 15%, though the results were not statistically significant.
    - Both treatment arms showed a larger effect compared to the reference arm (placebo for both supplements).
- Reviewed the study details but could not locate any CSV files related to the trial data on clinicaltrials.gov.
- Found WebPlotDigitizer (https://automeris.io/wpd/), an AI-based tool that can transform images or PDF files of graphs and charts into CSV files
- Continued exploring datasets on BIOGPS, considering potential correlations between gene expression data and the clinical trial.
- Compiled findings and drafted an outline for the next steps, including shifting focus towards more relevant datasets for further analysis.

Week 9 (Oct 21 – Oct 27)

· Met with Stephen on Friday, he showed me step-by-step process on how he used AI to find datasets and explore them
· Used ChatGPT to find publicly available datasets, results: CDC, GHDx, NHANES, SWAN Data, Kaggle Datasets (found celiac_disease_lab_data.csv), WHO, NIH Data Sharing, Autoimmune Registry Datasets, ICPSR
· Topic too broad, narrow it down (come from different angles)
· Take approach where I explore different autoimmune diseases (mainly gender focused), draw conclusions there on prevalence
· Next step to search natural therapies, conventional medical care, and dissatisfaction
· Question (all the research articles, could I use that data to create my own dataset?)
· CELIAC_DISEASE_LAB_DATA.CSV (2207 rows, 15 columns – BEFORE CLEANING AND PREPROCESSING)
· Using Python
· Data Loading and Initial Exploration
    o import necessary libraries and load the data
· Data Cleaning and Preprocessing
    o handle missing values (if any) CODE DID NOT WORK, save it and see what the problem might be
    o issue – “you're trying to modify a copy of a DataFrame, which might not affect the original data as intended” – code needed to change
    o missing Values - 418 missing values in the 'Diabetes Type' column (people who are not affected), after handling missing values, there are no missing values left in any column
    o before handling missing values data shape: (2207,15), data shape after preprocessing: (1788, 34))
    o Columns after preprocessing: Index(['Age', 'Gender', 'Diabetes', 'IgA', 'IgG', 'IgM', 'Disease_Diagnose', 'Diabetes Type_Type 1', 'Diabetes

Type_Type 2', 'Diarrhoea_fatty ', 'Diarrhoea_inflammatory', 'Diarrhoea_watery', 'Abdominal_no', 'Abdominal_yes', 'Short_Stature_DSS', 'Short_Stature_PSS', 'Short_Stature_Variant ', 'Sticky_Stool_no', 'Sticky_Stool_yes', 'Weight_loss_no', 'Weight_loss_yes', 'Marsh_marsh type 0', 'Marsh_marsh type 1', 'Marsh_marsh type 2', 'Marsh_marsh type 3a', 'Marsh_marsh type 3b', 'Marsh_marsh type 3c', 'Marsh_none', 'cd_type_atypical', 'cd_type_latent', 'cd_type_none', 'cd_type_potential', 'cd_type_silent', 'cd_type_typical'], dtype='object')

- o one-hot-encoding - Categorical variables like 'Diabetes Type', 'Diarrhoea', 'Abdominal', 'Short_Stature', 'Sticky_Stool', 'Weight_loss', 'Marsh', and 'cd_type' (f.e. 'Diabetes Type_Type 1', 'Diabetes Type_Type 2') – this explains rows going from 15 to 34)

· Exploratory Data Analysis

- o Prevalence of celiac disease by gender

Week 10 (Oct 28 – Nov 3)

- Prepared Presentation for Friday
- Prepared speech for presentation
- Explored more datasets (arthritis.csv, conversion_predictors_of_clinically_isolated_syndrome_to_multiple_sclerosis.csv)
- ARTHRITIS.CSV (675 rows, 8 columns – BEFORE CLEANING AND PREPROCESSING)
- 18 missing values in column ‘y’ – no description of data so it is not known what this column means
- Women more prevalent for this disease – 72.5% compared to men 27.5%
- 382 patients, of which 219 are women and 83 are men
- CONVERSION_PREDICTORS_OF_CLINICALLY_ISOLATED_SYNDROME_TO_MULTIPLE_SCLEROSIS.CSV (273 rows, 20 columns – BEFORE CLEANING AND PREPROCESSING)
- Initial_EDSS (148 missing values) and Final EDSS (148 missing values)
- After Handling all columns – no missing values
- Women more prevalent for this disease – 61.5% compared to men 38.5%
- 273 patients, of which 168 are women and 105 are men

WEEK 11 (Nov 4 – Nov 10)

- **Milena**
  - **Should specify expertise of people referenced.**
  - **Web scraping**
  - **What's the one thing you've learned the last 10 weeks?**

- **Incredibly ambitious**
- **Narrow the topic**
- **Datasets: Look at one thing**
    - **Start with the simplest thing first**
    - **Find one variable that depends upon another variable**
    - **What kind of relationship am I seeing?**
    - **Build your way out.**
    - **Gets you into the data.**
    - **What are two things**
        - **Age and choice to turn to natural therapy**
        - **Younger people more open.**
        - **Demographic groups.**
- **Datasets**
    - **Dig into how dataset was created.**
    - **What motivates the study.**
    - **Rheumatoid arthritis**
        - **Women tend to live longer than men**
        - **And arthritis is an older person's disease**
        - **Was this study done on a comparable group.**
        - **Hidden variables.**
            - **Could be influencing it.**
            - **Could be age that is driving this.**
- **Kaggle Database**
    - **Already has an immense amount of knowledge around each dataset.**
- **Give background information of what the autoimmune diseases are.**

- **Reach out to experts and professional organizations**
    - **Homeopathic**
    - **What are regional resources?**

    - **What is that influenced people towards homeopathic**

- Carefully went through the feedback from both Professor Nigel and Mr. Seifert.
- **Key takeaways:**
    - Professor Nigel emphasized narrowing down the topic and starting with a simple focus on a single variable relationship.
    - Mr. Seifert suggested focusing on demographic groups and investigating specific influences on choices around natural therapies.

- Reached out to the **Vitaly Study Group** for insights or potential resources. Awaiting their response.
- Connected with a **homeopathy professional** back home for regional insights on factors influencing homeopathic choices. She mentioned she would respond by the end of the week.
- Located **additional datasets**, primarily focused on women with autoimmune diseases
- Validated Professor Nigel's point on age being a factor by checking the rheumatoid arthritis dataset: the average age is **50.38**.
- Plan to do web scraping next week to gather more data.
- Initial analysis plan:
    - Identify a single, clear relationship between variables to start (e.g., age and the choice to use natural therapies).
    - Focus on demographic segmentation as suggested by Mr. Seifert, to gain more targeted insights.

Week 12 (Nov 11 – Nov 17)

- Contacted Jacob and am waiting for his response.
- Met with Professor Nigel and discussed my progress.
- Found some datasets on complementary medicine, focusing on it as a potential keyword.
- Talked with Professor Nigel about possibly continuing with this topic, exploring the probability that women with autoimmune diseases turn to natural therapies, given that they are dissatisfied with conventional medical care.
- Learned that more research is needed on this topic; while there are articles discussing it, I couldn't find datasets to work with.
- Spoke with Dr. Erisman about potential next steps.
- Plan to search for connections this weekend; if unsuccessful, I'll shift to focusing on lifestyle factors contributing to obesity, as I already have relevant datasets for that.
- With only 12 weeks left, I need a topic that is "tangible" (something I can make concrete progress on).

- Dryad, Figshare databases
- Dataset **on fasting (1693 rows, 3 columns)**
    - atientVisitKey: 48, Class: derm, Diagnosis: psoriasis
    - PatientVisitKey: 53, Class: GI, Diagnosis: autoimmune
    - PatientVisitKey: 190, Class: autoimmune, Diagnosis: lupus
    - PatientVisitKey: 191, Class: autoimmune, Diagnosis: lupus
    - PatientVisitKey: 254, Class: GI, Diagnosis: autoimmune
    - PatientVisitKey: 401, Class: GI, Diagnosis: autoimmune
    - PatientVisitKey: 453, Class: derm, Diagnosis: psoriasis
    - PatientVisitKey: 470, Class: GI, Diagnosis: autoimmune
    - PatientVisitKey: 520, Class: derm, Diagnosis: psoriasis
    - PatientVisitKey: 657, Class: autoimmune, Diagnosis: lupus
    - PatientVisitKey: 721, Class: derm, Diagnosis: psoriasis
    - PatientVisitKey: 993, Class: GI, Diagnosis: autoimmune
- **Results**
    - GI (gastrointestinal): 5 cases
    - Autoimmune: 3 cases (all lupus)
    - Psoriasis: 4 cases
- **Dataset on obesity (104273 rows, 31 columns) – Nutrition, Physical Activity and Obesity**

Week 13 (Nov 18- Nov 24)

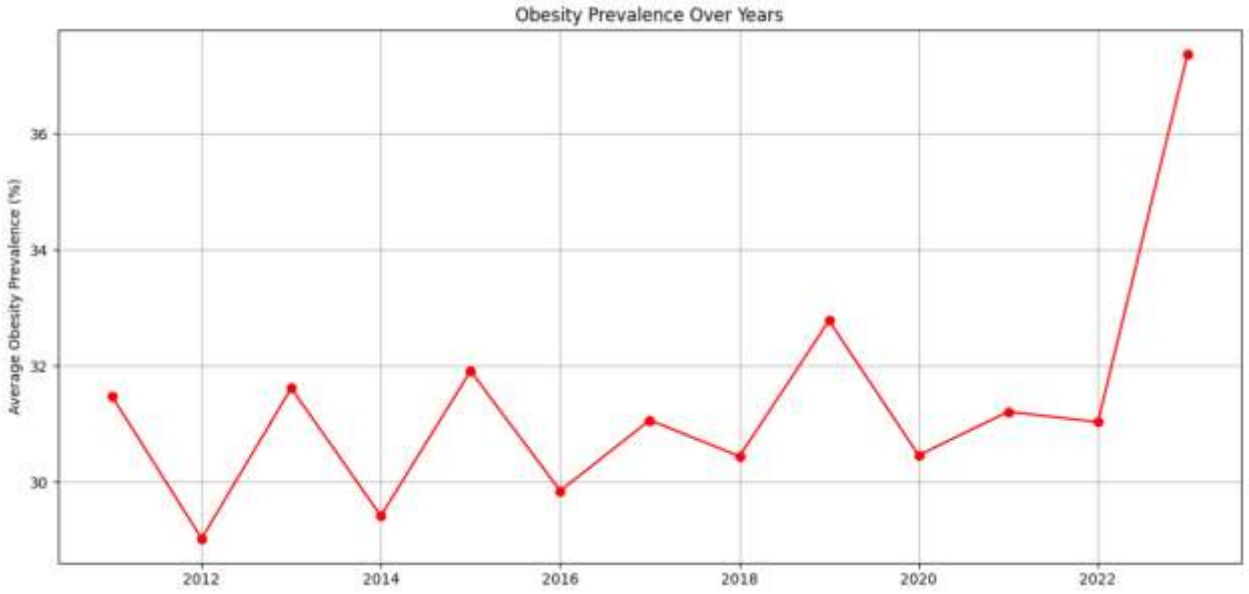
- Got an email from Dr. Schillo
    - Hi Milena,

      Thank you for reaching out. **I don't have the background** to give a direct answer. I do have **some colleagues that work with autoimmune diseases**. I will reach out to them for guidance and **get back to you in the next week.**

      All the best,
      Jake
- Dataset on obesity (104273 rows, 33 columns) – Nutrition, Physical Activity and Obesity
- **Key Columns and Their Descriptions**:
    - YearStart / YearEnd: The range of years for which the data was collected.
    - LocationAbbr / LocationDesc:
    - LocationAbbr: Abbreviation of the location (e.g., AK for Alaska).
    - LocationDesc: Full name of the location.
    - Datasource: The data source (e.g., BRFSS - Behavioral Risk Factor Surveillance System).
- Class / Topic:
    - Class: High-level category (e.g., "Obesity / Weight Status").
    - Topic: More specific focus within the class (e.g., "Obesity / Weight Status").

- Question: Describes the specific measure or survey question (e.g., "Percent of adults aged 18 years and older who have obesity").
    - Data_Value / Data_Value_Alt:
    - Data_Value: The primary data value (e.g., prevalence percentage or rate).
    - Data_Value_Alt: Alternative representation of the data value (often **identical** to Data_Value).
- Low_Confidence_Limit / High_Confidence_Limit: The confidence interval for the data value.
- Sample_Size: Number of survey participants for the specific data point.
- Stratification Information: Provides additional demographic breakdowns, such as:
    - Age(years): Age groups (e.g., "35 - 44").
    - Education: Educational attainment levels.
    - Gender: Male/Female.
    - Income: Income ranges (e.g., "$15,000 - $24,999").
    - Race/Ethnicity: Categories like "White", "Hispanic", etc.
- GeoLocation: Latitude and longitude coordinates for the location.
- Stratification and Category IDs: Unique IDs associated with demographic stratifications.
- Load and Handle Missing Data Values:
    - Missing Values in Dataset:
    - YearStart 0
    - YearEnd 0
    - LocationAbbr 0
    - LocationDesc 0
    - Datasource 0
    - Class 0
    - Topic 0
    - Question 0
    - Data_Value_Unit 15400
    - Data_Value_Type 0
    - Data_Value 10767
    - Data_Value_Alt 10767
    - Data_Value_Footnote_Symbol 93505
    - Data_Value_Footnote 93505
    - Low_Confidence_Limit 10767
    - High_Confidence_Limit 10767
    - Sample_Size 10767
    - Total 100548
    - Age(years) 81928
    - Education 89376
    - Gender 96824
    - Income 78204
    - Race/Ethnicity 74480
    - GeoLocation 1932

    - ClassID 0
    - TopicID 0
    - QuestionID 0
    - DataValueTypeID 0
    - LocationID 0
    - StratificationCategory1 0
    - Stratification1 0
    - StratificationCategoryId1 0
    - StratificationID1
    -
- Unique values in 'LocationDesc':
  ['**Alaska**' 'Alabama' **'Arkansas'** 'Arizona' 'California' 'Colorado'
  'Connecticut' 'District of Columbia' 'Delaware' 'Florida' 'Georgia'
  'Hawaii' 'Iowa' 'Idaho' 'Illinois' 'Indiana' 'Kansas' 'Kentucky'
  'Louisiana' 'Massachusetts' 'Maryland' 'Maine' 'Michigan' 'Minnesota'
  'Missouri' **'Mississippi'** 'Montana' 'North Carolina' 'North Dakota'
  'Nebraska' 'New Hampshire' 'New Jersey' **'New Mexico'** 'Nevada' 'New York'
  'Ohio' 'Oklahoma' 'Oregon' 'Pennsylvania' 'Rhode Island' 'South Carolina'
  'South Dakota' 'Tennessee' 'Texas' 'National' 'Utah' 'Virginia' 'Vermont'
  'Washington' **'Wisconsin'** 'West Virginia' **'Wyoming' 'Puerto Rico'** '**Guam'**
  'Virgin Islands']
- Unique values in 'Gender': [nan 'Female' 'Male']
    - Prevalence
        - **Female 47.30458**
        - **Male 52.69542**
- Unique values in 'Age(years)': [nan '35 - 44' '25 - 34' '18 - 24' '55 - 64' **'65 or older'** '45 - 54']
-
- Unique values in 'Income': [nan **'$15,000 - $24,999' '$50,000 - $74,999' '$75,000 or greater' 'Less than $15,000' '$25,000 - $34,999**' 'Data not reported' '**$35,000 - $49,999'**]
- Unique values in 'Race/Ethnicity': ['2 or more races' 'Other' nan 'Non-Hispanic White' 'Asian'
  'Hawaiian/Pacific Islander' 'American Indian/Alaska Native' 'Hispanic' 'Non-Hispanic Black']
- Prevalence over the years:



- Key Insights:
  State with highest obesity prevalence: Virgin Islands (33.81%)
  Age group with highest obesity prevalence: 65 or older (33.11%)
  Gender with highest obesity prevalence: Male (52.69%)
  Income level with lowest obesity prevalence: Data not reported (30.69%)
  Year with highest average obesity prevalence: 2023 (37.36%)

Week 14-15 (Nov 25 – Dec 8)

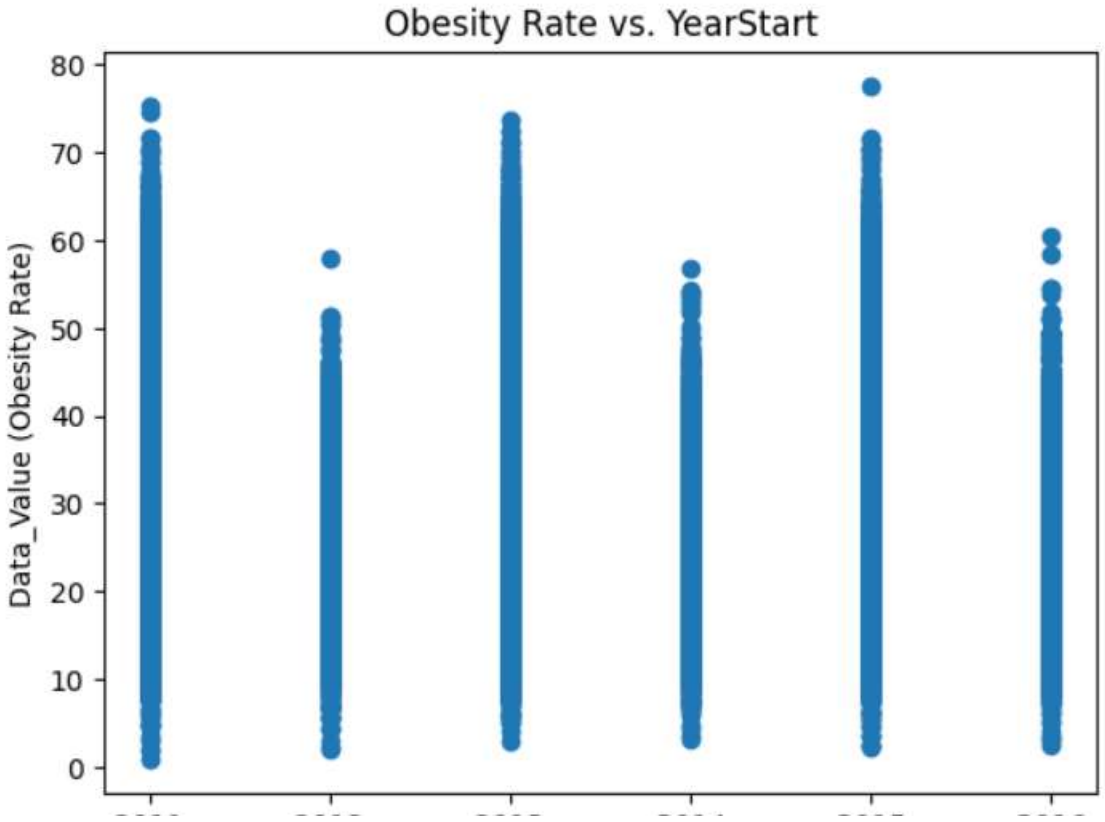
- **Meeting with Dr. Erisman**
  - scikit
  - if i don't see linear, it might be classification
  - any types of correlations
  - clustering
  - random forest
  - testing columns one against the others
  - removing the data to see relationships
  - scaling
  - based on sklear library what model would you suggest
  - see what does the column label mean
- **Meeting #2 with Dr. Erisman**
  - breaking up dataset based on certain values (OWS, PA, FV)
  - compare those per state - f.e.
  - think about that this dataset is comprised of multiple datasets, is there something easy
  - is there couple fields that are always field in to see if i have the set that is good
  - **location - visualize**
  - gender - new small dataset

- how do surveys differ that make up the data for brfss
- group them in coherent way, take separate things and try to find correlation
-
- talk about process of finding and grouping data
- charts, correlations
-
- model - problem for now

About the Division of Nutrition, Physical Activity, and Obesity

AT A GLANCE

CDC's Division of Nutrition, Physical Activity, and Obesity (DNPAO) invests in efforts to support healthy eating, active living, and healthy weight for all people. These investments advance public health strategies that prevent chronic diseases related to diet and inactivity to protect the health of people across the nation.



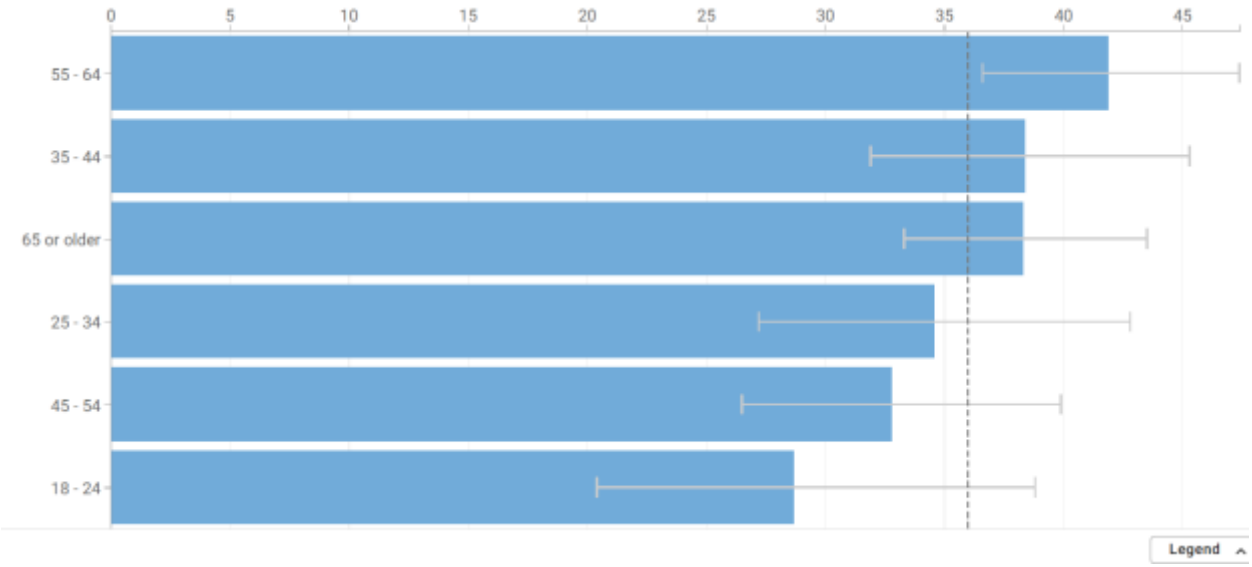


Figure 1 - Alaska

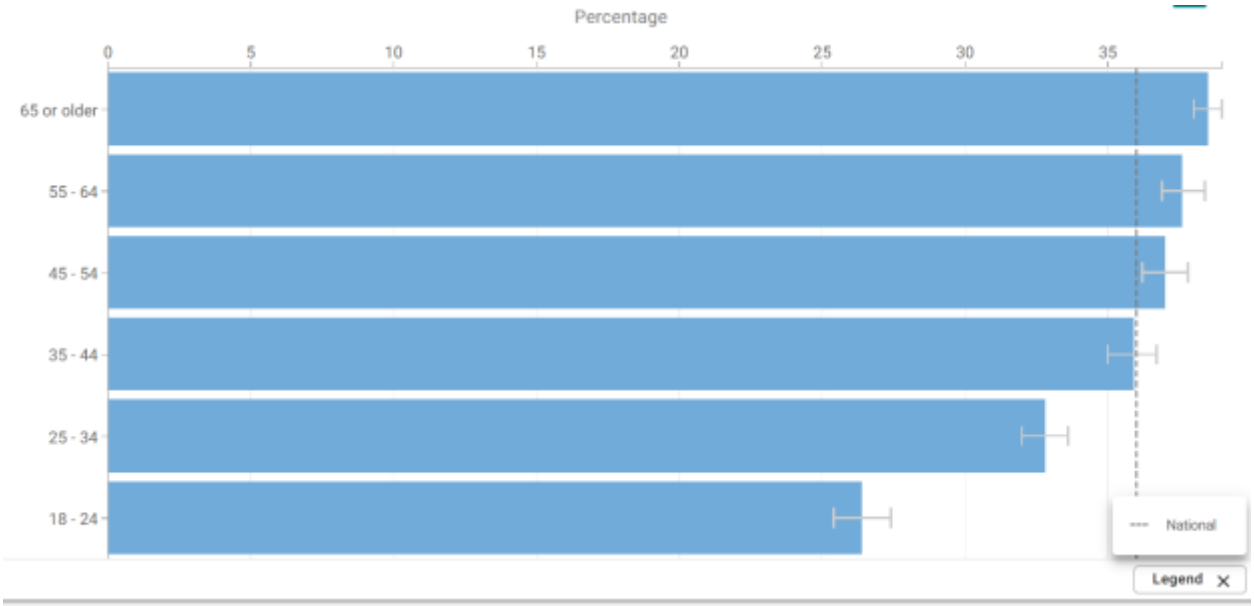


Figure 2 Kansas

- Explore other correlations (gender, age, ethnicity, socioeconomic status etc.
- Try models
    - Predict which locations are likely to have higher or lower obesity rates in the future.
    - Predict the obesity rate for a specific location in a future year
    - Predict whether a location will have high, medium, or low obesity rates in the future based on historical data

Week 16 (Dec 9 – Dec 15)

- Try to find a model
- Finish research proposal for new topic
- Finish and prepare for presentation