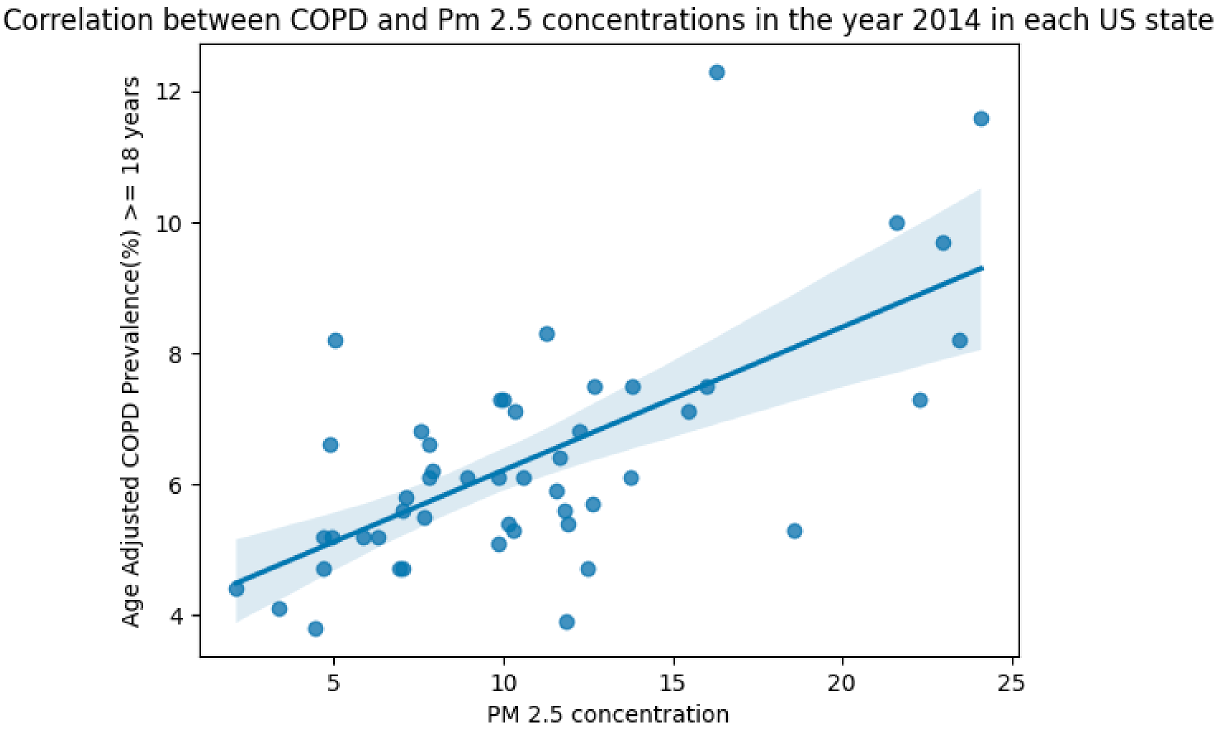
Checkpoint 1: EDA Section

1. **Scatterplot of PM 2.5 concentration and COPD prevalence**

This scatter plot visualizes the relationship between two quantitative variables: Pm2.5 concentration and COPD prevalence. The unit of observation is each US state in the year 2014. The data sources are CDC's census tract level pm2.5 concentrations which we aggregated for each state and CDC's Chronic Disease Indicator dataset.

We observe a positive correlation between PM2.5 levels and COPD prevalence, that is, as the concentration of particulate matter increases, the COPD prevalence percentage also increases.
We notice a couple of states that are outliers — one state has the highest COPD prevalence but the PM2.5 is around 15-16. There is also a state with relatively high pm2.5 with low COPD prevalence.

This strong positive correlation is in line with our causal DAG in the research proposal where we predicted that Pm2.5 concentrations would have an effect on COPD prevalence. This plot is an indication that we could potentially infer the causal effect of pm2.5 on COPD after accounting for confounding factors.

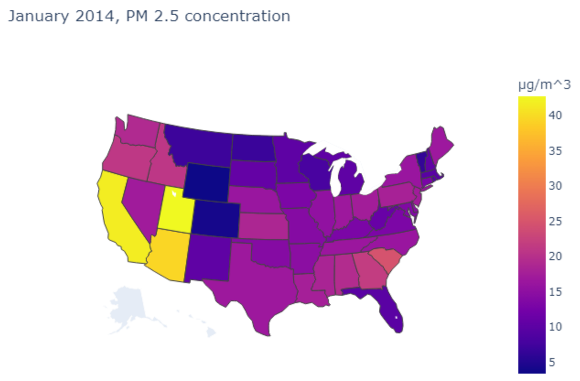


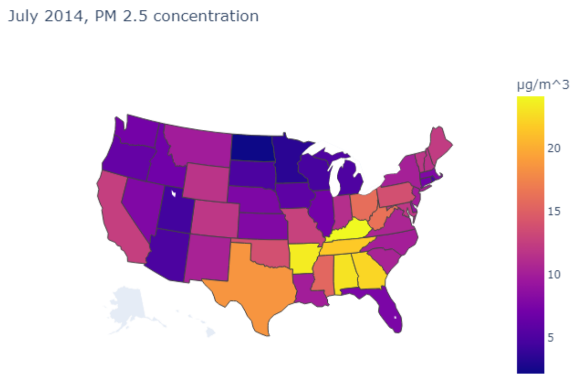
2. Heatmaps of PM 2.5 concentrations at various times in the year

These heatmaps visualize the PM 2.5 concentrations across the 48 states in the contiguous US for January and July 2014. This data was generated from the census tract-level measurements of PM 2.5 published by the CDC. We aggregated this into state-level data by taking a weighted average of each county based on each county's population.

We observe that the states in the South-Central region seem to have the highest PM 2.5 concentrations in the summer time (around July), and states in the Pacific region have higher PM 2.5 levels in the winter (around January). We were originally considering just using data from a single day of the year, but this suggests that it would be better to use an average over the entire year. It could be useful to follow up on the trends for these states over a year (or multiple years) to determine the best way to aggregate the daily data.

One concern with our original problem statement was that one state might affect adjacent states' PM 2.5 concentrations, which could cause our treatment to violate SUTVA due to spillover effects. The heatmaps seem to present some evidence to the contrary, since in the January heatmap we see that Utah has one of the highest PM 2.5 concentration, but its neighbors Colorado and Wyoming have some of the lowest PM 2.5 concentrations. There are still other factors to consider, such as interstate travel or spillover in smaller states such as the New England region, but we would like to see how the results might change if we assume no spillover compared to when we consider only non-adjacent states, as we were originally planning to do.


January 2014, PM 2.5 concentration

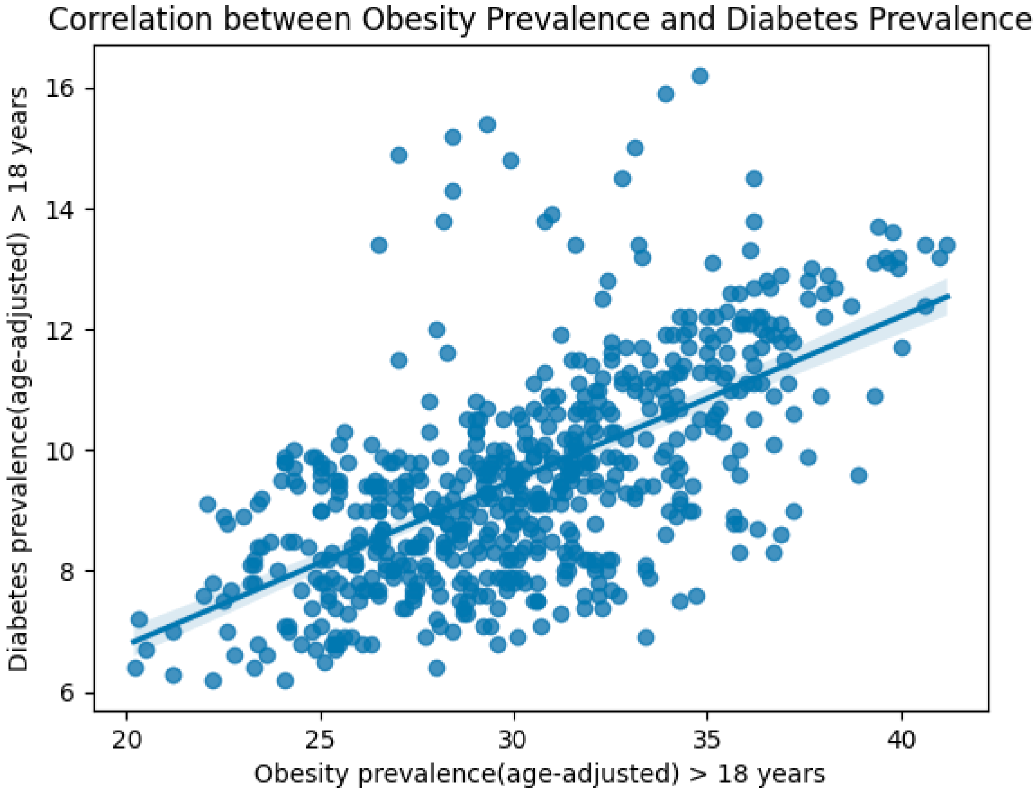

July 2014, PM 2.5 concentration

3. Scatterplot of Obesity Prevalence and Diabetes Prevalence

This scatter plot visualizes the relationship between two quantitative variables: Age Adjusted Diabetes prevalence and Age Adjusted Obesity prevalence. The unit of observation is each US state in each of the years 2011-2014. The data source is the CDC's Chronic Disease Indicator dataset.

We observe a positive correlation between obesity prevalence and diabetes prevalence, that is, as the prevalence of obesity increases, the diabetes prevalence also increases.

We use this plot to motivate our first research question on Multiple hypothesis testing (Is there a significant association between Diabetes and Obesity)

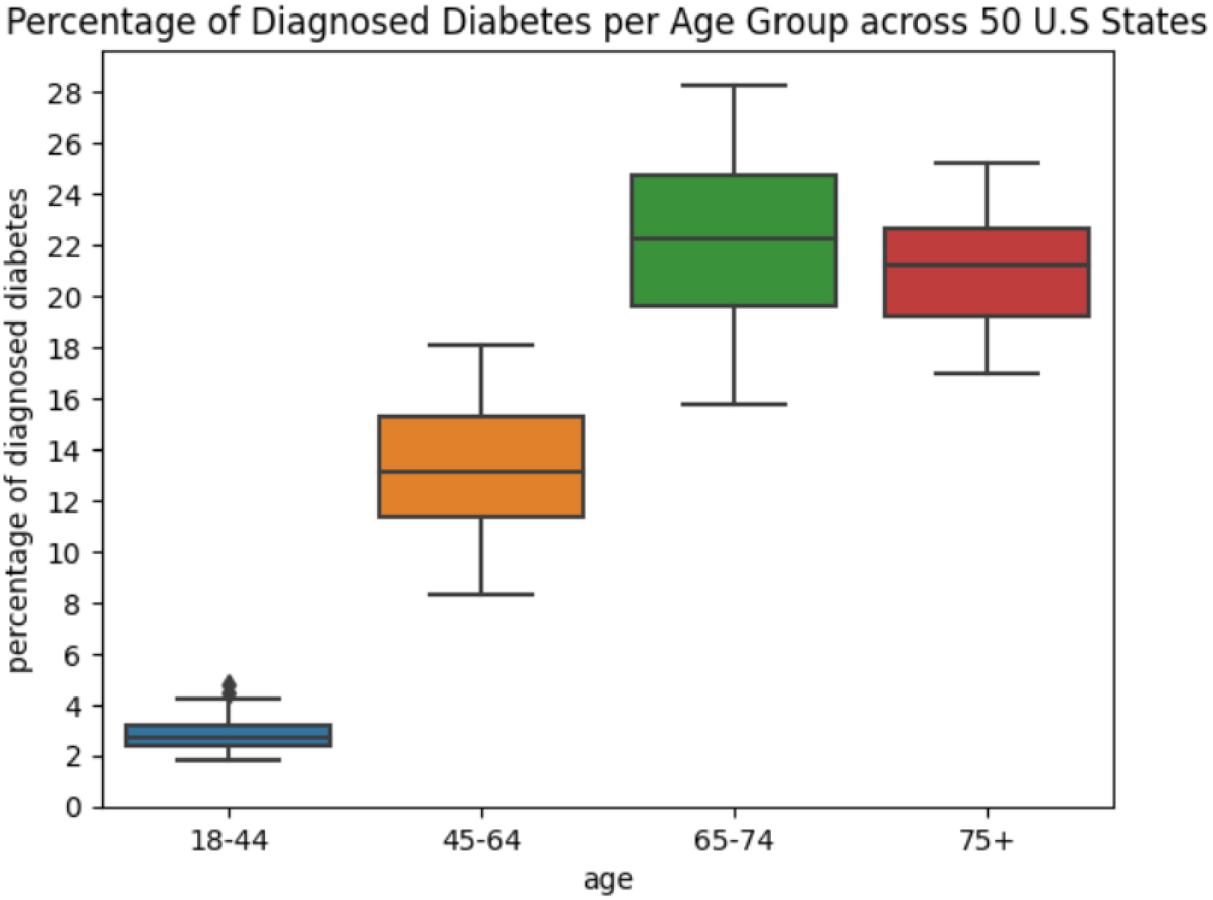


4. Percentage of Diagnosed Diabetes per Age Group (across 50 US States)

This boxplot visualizes the relationship between the categorical variable ‘Age’ and the quantitative variable ‘Percentage of Diagnosed Diabetes’. ‘Age’ is divided into 4 categories which are ranges of ages: ‘18-44’, ‘45-64’, ‘65-74’, and ‘75+’. Each age range has a boxplot that visualizes the distribution of ‘Percentage of Diagnosed Diabetes’ across all 50 states. Each data point is a percentage of people diagnosed with diabetes belonging to an age range and a state.

From this boxplot, we can see that across all 50 states, the age range 18-44 has the least dispersed data with a median between 2-3 percent, far below the other age ranges. However it has a couple of outliers. The age range 65-74 has the most dispersed data with a wide interquartile range, yet its median is around the same as the 75+ age range even though the range for 75+ is smaller.

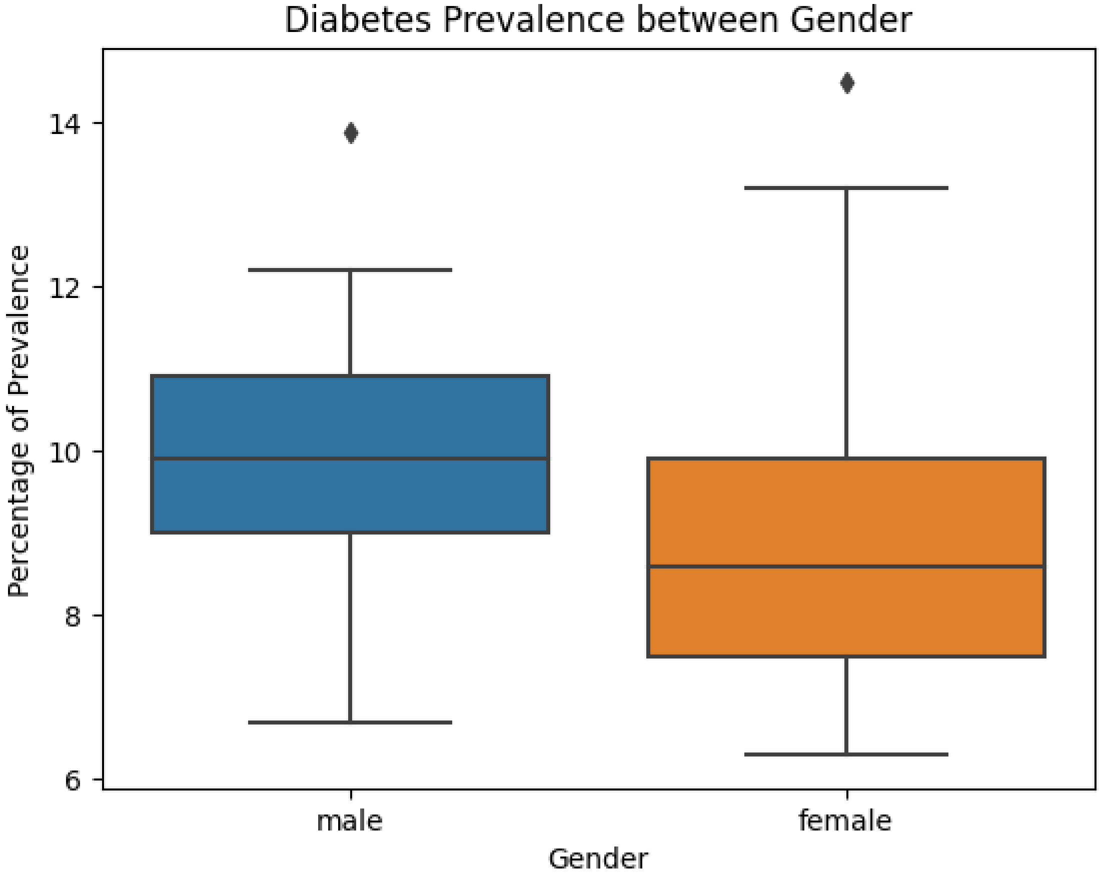
Overall, this boxplot shows a progression in the diagnosis of diabetes over multiple ages. There are lower percentages when people are younger and higher percentages when people are older. This plot will motivate our first research question on Multiple hypothesis testing (Is there a significant association between Diabetes and Age.)



5. Diabetes Prevalence between Gender

The boxplot below depicts the differences of diabetes prevalence between genders. The data is divided by two gender categories, either male or female.

We see that females have a much wider distribution yet the median is lower than the male prevalence. The female distribution has a lower minimum value, 25th percentile, 50th percentile, and 75th percentile. However, the female distribution also has a higher maximum and more outliers; therefore, we would like to further explore this difference and discover if it is due to underlying differences between the genders or if it was a difference due to chance. This will feed into our first research question where we will use multiple hypothesis testing to answer this question.


Diabetes Prevalence between Gender
Percentage of Prevalence
male
female
Gender