Tanmay Kapse: Explore effects of Risk Factors on Disease Prevalence across Demographic groups.
Unit of Analysis: Disease Cases, Colab Link: Tanmay Kapse- Homework7.ipynb

For this part of the project, I decided to test out two models, but I will only discuss one, the XGBoost gradient descent model. Before deciding on a type of model, I decided to conduct some data pre-processing and wrangling. I split and pivoted our dataset so that each risk factor and demographic data (age, gender and race) was its own column. This way, I could see the percentages for each year and state, and for each cardiovascular disease.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Year	LocationD	DiseaseTy	OverallPre	Disease_A	Disease_A	Disease_A	Disease_A	Gender_Fe	Gender_M	Race_Hisp	Race_Non	Race_Non	Race_Non	Race_Oth	Cholesterc	Diabetes	Hypertens	Nutrition
2	2011	Arizona	Acute Myo	17.8		14.899999	24.5	25.2	14.15	20.25	2.5	18.5				67.85	9.5	49.8	78.5
3	2011	Arizona	Coronary H	3.9		4.1	12.3	14.2	3.4	4.4		5				67.85	9.5	49.8	78.5
4	2011	Arizona	Major Carc	8.5	2.5	9.6	23	28.5	7.2	9.7	6	9.6			8.4	67.85	9.5	49.8	78.5
5	2011	Arizona	Stroke	3		3.1	8	10.7	2.7	3.4		3.3				67.85	9.5	49.8	78.5
6	2011	Arkansas	Acute Myo	6.5		8	18.1	19.5	5.1	7.9		7.1	5.3			68.2	11.2	58.2	86.4
7	2011	Arkansas	Coronary H	5.7		6.8	15.5	17.1	5	6.4		6.2	2.9			68.2	11.2	58.2	86.4
8	2011	Arkansas	Major Carc	11		13.1	29.9	33.2	9.7	12.4		11.6	8.3		11.4	68.2	11.2	58.2	86.4
9	2011	Arkansas	Stroke	4		5.1	10	12.6	4.2	3.9		4.1	4			68.2	11.2	58.2	86.4
10	2011	California	Acute Myo	3.3	0.9	4.1	10.7	12.3	2.3	4.4	2.5	3.8	5.2	2.5		66.3	8.9	49.75	75.6
11	2011	California	Coronary H	3.5	1	4.3	11	12.4	2.7	4.4	2.6	4.3	4.5	2.5	5.8	66.3	8.9	49.75	75.6
12	2011	California	Major Carc	6.6	2	7.7	20.9	23.2	5.6	7.7	4.9	7.8	8.9	5.1	13	66.3	8.9	49.75	75.6
13	2011	California	Stroke	2.2	0.7	2.2	7.8	8.8	2.2	2.2	1.6	2.7	2.4	1.8		66.3	8.9	49.75	75.6
14	2011	Colorado	Acute Myo	2.7	0.5	2.8	9.9	13.4	1.8	3.5	2.2	2.8				64.35	6.7	47.400000	80.8
15	2011	Colorado	Coronary H	2.5		2.8	9.2	12.6	2	2.9	1.1	2.7				64.35	6.7	47.400000	80.8
16	2011	Colorado	Major Carc	5.3	1.4	5.7	18.2	25.6	4.4	6.2	3.8	5.5	6.6		8.7	64.35	6.7	47.400000	80.8
17	2011	Colorado	Stroke	2	0.8	2.2	5.7	8.9	2	1.9	1.9	1.8				64.35	6.7	47.400000	80.8
18	2011	Connectic	Acute Myo	23.8		24.45	28	28.2	17.6	27.75	3	26.05				66.5	9.3	54.45	79.1
19	2011	Connectic	Coronary H	3.3		3.3	11.1	12.4	2.8	3.9	3.1	3.7				66.5	9.3	54.45	79.1
20	2011	Connectic	Major Carc	7	1.7	6.6	21.5	25.8	6.1	8	6	7.4	5.8			66.5	9.3	54.45	79.1

After this step, I began to explore different models. I noticed that there were many missing values for each of the columns, something that I did not initially realize. Dropping the rows with missing values was not feasible either, as almost every row had at least 1 missing value. I researched, and found a new sort of model building technique that utilized some of the decision tree work that we have been doing in class, and combining that with a gradient boosting algorithm. How XGBoost works is that it will create multiple trees, and each one is additive from the previous, and each one stronger in predictive ability. I also utilized hyperparameter tuning to get the best model, without overfitting. I decided to split my model by using risk factors as one set of predictors, and then demographics as a separate set of predictors. I had tried to combine the two, but it gave me an unrealistic r^2 of .99. Theoretically, this seems perfect, but in the real world, we are rarely if ever, getting an r^2 of .99. When I looked at risk factors alone as predictors using XGboost, I got an r^2 value of .31. Though this performance is much worse than the initial model, it is more realistic, and allows for a more detailed analysis in future work. I was also able to see that the initial risk factor that the tree split on was hypertension, suggesting high importance in my ensemble of trees. This was followed by branches on cholesterol abnormalities and nutrition, working its way down to suggesting high predictive power.

I will clarify that I only visualized up to 10 trees in the model, this could be the reason for discrepancies in my tree splits and overall feature importance, where smoking and diabetes got the highest importance value. Since XGBoost works in a way where later trees correct mistakes and have more specific splits, the overall feature importance may have incorporated many more trees to gain a larger cumulative effect. When looking at demographic data, I saw that the White population suffers the most overall from cardiovascular diseases. This is followed by anyone who identifies as a male. The model tells a different story than my EDA, where I saw that other races suffered more overall from cardiovascular diseases. However, in both my tree model and in the EDA, males suffered more overall from cardiovascular diseases. Overall, my model told a slightly different story than my EDA did in answering my sub-goal, which could be due to model selection, data pre-processing or an unknown reason, but I did get insights that were useful to answering my question, even if they were different from my expectations from EDA.

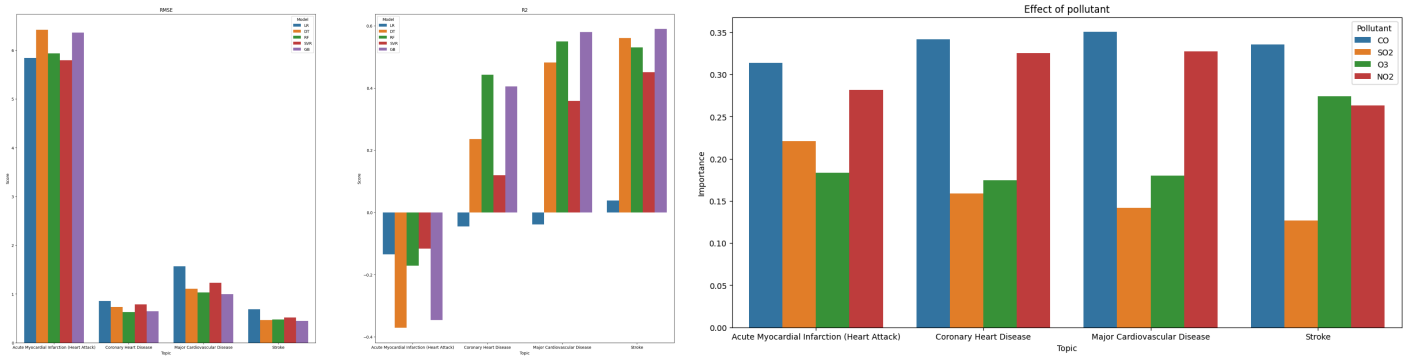
Galav Sharma: Explore effects of pollutants levels on cardiovascular diseases
Unit of Analysis: State, Colab Notebook: [Galav Sharma - Homework 7]

Our analysis revolves around finding connections between cardiovascular health and pollution. My query looks at the effect of pollutants on the rate of cardiovascular diseases. Getting results of models on this subset will help us understand the impact of pollutants on the four different cardiovascular diseases.

After performing Exploratory Data Analysis on this subset, I could not find any conclusive relationships between pollutants and their effects on cardiovascular diseases, and thus fit several models (Linear Regression, Decision Tree, Random Forest, Support Vector Regression, and Gradient Boost), each with one of the cardiovascular diseases as the response variable and the mean pollutant values and AQIs as the feature columns. These were chosen to understand the relation between the pollutants and cardiovascular diseases. Single hyperparameter tuning was then conducted for all types of models to get the best fits. I created a parameter grid and checked which parameters yielded the best results. The models were evaluated based on the average R^2 score for 'Coronary Heart Disease', 'Major Cardiovascular Disease', and 'Stroke', as the models did not fit well for 'Acute Myocardial Infarction'. Of all the models, the Gradient Boosting Regressor Model yielded the best results. For this, I tested various hyperparameter values for n_estimators, learning_rate, max_depth, min_samples_split, min_samples_leaf, etc. For each value of the parameters, the models were trained and evaluated on the average R^2 values for the models for three cardiovascular diseases. From this, I achieved best results for the models with min_samples_leaf=2.

For the models, not much data wrangling had to be performed. Since there are four different cardiovascular diseases, the data fed to the models had to be split on the 'Topic' column to only have data for a single disease for every model. The 'Overall' column was then used as the response variable and the pollutant columns were used as the features. These models were then evaluated based on the R^2 and RMSE scores obtained. From the analysis, I found that the 'CO Mean' and 'NO_2 AQI' columns had the most impact on the Gradient Boosting Regressor.

From this analysis and the results of the models, I found that Tree-based models perform the best, with Gradient Boosting Regression outperforming the rest. As seen in the figure below, tree-based models perform well in predicting three of the four cardiovascular diseases. However, despite various attempts, none of the models produced satisfactory results for 'Acute Myocardial Infarction', with each of them yielding negative or low R^2 values.



From the figure above, we can see that Carbon Monoxide (CO) has the highest influence across diseases with around 30-35% feature importances for the Gradient Boosting Models. It can also be seen that different pollutants impact different diseases. CO plays an important role towards all cardiovascular diseases, NO_2 has higher influence on 'Major Cardiovascular Disease' and 'Coronary Heart Disease', and O_3 gains importance for the model predicting 'Stroke'. This highlights the nature of the effect of pollutants on cardiovascular diseases.

Pratyush Joshi: Explore the different trends over time for key pollutants.
Unit of Analysis: Year, Colab Notebook: Pratyush_Joshi_Homework_7.ipynb

For my analysis, the method I chose was a linear regression. The reason I chose this method is because I thought it might help identify trends in our data more easily. By generating a line of best fit, I could compare the current results that the data generated and see if there might be some relationship between my two variables. I ran this method by using a simple linear regression using the sklearn.linear_model module, and imported LinearRegression from it. I used the variable I generated from part 2, that grouped our merged dataset by 'Year', while taking the mean of each column. I initialized the parameters, X and y, with X the independent variable being set to the year and also reshaping the parameter to become a 2d array instead of 1d to work with sklearn. The y variable or dependent variable was set to be the different mean values of the different pollutants such as 'CO_Mean' and 'SO2_Mean'. Other than that no other fine-tuning of parameters was required as I used simple LinearRegression() to fit the data, as the default method already works well enough for simple regression.

While fitting the model, wrangling was not required as the data had already been cleaned prior to merging and any missing values in the pollutant data had been removed. And since I focused only on using two columns for each linear regression, no scaling was also needed. Instead I kept the data the same and focused on the two columns I would need for this analysis, being the 'Year' column and the target column I would analyze upon, being one of the different mean columns for the pollutants.

After evaluation it seems that a majority of the means seem to have weak trend fits, with only 1 out of the 4 models I analyzed having an actual strong performance. For my evaluation I used MSE as well as R-squared values to evaluate the trends between the different models. Out of the 4 models only one had an r-squared of above 0.7 being the O3 mean with an r-squared of 0.802, and also had the lowest MSE score 8.25e-08, while the others had MSE scores a lot greater. The only other model with a decent r-squared score, being NO2 mean with an r–squared of 0.638, had the highest MSE score among the other models at 0.0406, indicating that while the model had less variance than average, it still was far off from what it was expected to be.

A lack of linear regression displays that the pollution data might not be suitable to provide concrete correlations towards the heart disease date, and would most likely not help us find a clear answer to our business goal. This considering the fact that when the data was aggregated by 'Year', the data included a range of data from different locations across America. This most likely creates an inconsistency as many locations throughout America often have different levels of pollution rates, so to take the average across all those values would smooth out any strong patterns that might exist in certain areas. As a result it generates a plot that varies a large amount due to the differences in data it receives and as a result would be hard to make any correlations towards the heart disease data, as the aggregates mask any potential of discovering possible local trends that show a link between heart disease and pollution rates.

Group Analysis/Methods, Github link: https://github.com/galav12/CSC442-Project **:**
Our business goal is to identify and explore the effects of specific air pollutants and risk factors over time on cardiovascular disease prevalence across given demographics such as age, gender and race. Since our dataset is very complex, we decided as a team that the best course of action is to split into different subsets to understand different relationships in the dataset. The results gained from these were then used to analyze the overall dataset and meet our business goal. We all saw that for our dataset, a linear regression model is not the best way to represent our data, and predict future values. We have many predictors, and a quite obvious non-linear relationship between these predictors and our target variable which is cardiovascular disease prevalence. We saw that tree-based models achieve the best results. The Gradient Boosted Decision Tree performed the best for our subsets, as this can handle missing values, and also predict non-linear, complex relationships better. By using this modelling technique, we utilized hyperparameter tuning, to get the most specific tree design, and stronger predictive ability.

In the sub-goal discussing risk factor effects on cardiovascular diseases, the XGBoosted Tree model found that Hypertension and Cholesterol Abnormalities are the risk factors that affected disease prevalence the most. Like mentioned above, one of the tuning parameters is the number of trees, for which I used 10. The more trees we get, the more and more specific and accurate predictions we get, while increasing the risk for overfitting. The feature importances show that smoking is the most important predictor, and this is using the best number of trees. Our R^2 value shows that it is a better model than others, with an r^2 of .31, as compared to .03 from the MLR.

While looking at the effects of pollutants on cardiovascular diseases, the GradientBoostingRegressor performed the best across most diseases with R^2 values around 0.4-0.6, indicating pollutants have a decent impact on the cardiovascular diseases. We also found that Acute Myocardial Infarction could not be reliably predicted, indicating that the disease also depends on many other factors and there is not a direct relationship with these pollutants. From the GradientBoostingRegressor, we also found the effect of pollutants on different diseases by looking at the feature importances. We found that CO has high influence across all diseases, NO_2 had a high impact on Coronary Heart Disease and Major Cardiovascular Disease, and O_3 having a high impact on Stroke. This indicates that pollutants have different impacts on different diseases and highlights the nature of the effect of pollutants on cardiovascular diseases.

The analysis related to the pollutant trends over time using linear regression, did not contribute to the overall business goal, as the nonlinear relationships made it difficult to analyze pollution trends over time. Originally it was planned that if pollution trends were positive or negative, the heart diseases statistics would have a positive relationship with the data, indicating a correlation between rates of pollution and heart disease. But because of the nonlinear trends, it made it hard to make that analysis when put next to the heart disease data and as a result our insights towards the business goal were limited because of it.

From this, we concluded that both pollutants and risk factors have an impact on the prevalence of cardiovascular disease, with varying impacts on different diseases. Through Gradient Boosting models, we found that pollutants have varying impacts on different cardiovascular diseases, with CO, NO_2, and O_3 having significant impacts for most diseases. We also found that Hypertension and Cholesterol abnormalities significantly affect cardiovascular disease prevalence.

Timeline

Planning EDA: 3/15
Querying and finishing EDA: 3/20-3/31
Modelling Research: 3/25-4/1
Modelling Building: 4/1-4/10
Model Evaluation/Planning final group deadlines: 4/10-4/14
Finishing results and conclusions for analysis: 4/13-4/14
Group Analysis: 4/14
Presentation: 4/15

Future work

We can work to utilize more advanced modelling techniques to achieve our business goal, and gain a better understanding of reasons for cardiovascular disease. Right now, we use very basic models, with some hyperparameter tuning, but looking at the complexity of our data, we need a more complex model to better represent all our predictors and response variables. These results can be used by policy makers, doctors, medical researchers, sustainability engineers, and anyone who deals with either medical or environmental data. These results can spark changes in policy, or guide focused research to better the care for those who suffer from risk factors. This in turn can reduce the number of people overall who suffer from these diseases.

Limitations

The heart disease data was limiting as it did not include statistics relating to the city that the data was taken from. As a result this limited the potential of our analysis as we could not analyze specific trends that might be occurring in certain regions across the country. We also had outliers in our dataset, which could have impacted our EDA and modelling, even though we took care of them, there is always a possibility for data leakage. Finally, the scope of our data is only from 2011-2015. A larger scope would have better helped understanding any long term trends that may have existed between pollution levels and heart disease rates. A larger scope could also help us understand if the relationship was stable or if there were any large events or seasonal changes that might have caused fluctuations in the data.

Our model as mentioned is not perfect to represent our data, and that is always a limitation. No model is perfect and there definitely is a better model out there to more accurately represent data.

Challenges

Some challenges included choosing what models to use to evaluate our queries, since there were many numbers of models we could have chosen from and we wanted to make sure that the model we chose would be the best towards making analyses from our data. Also the challenge between choosing the different queries because we wanted to make sure that our own original insights would contribute to the group's final business goal. When creating models, we attempted to use many different tuning methods, especially for MLR, and one of those was combining the two separate predictor variable sets into one. This caused an inflation of the r^2 value, especially with the inputted values. When we split the predictor variables, we got a very low r^2, but it was honestly better than seeing a .99, which is unrealistic in almost any real world scenario. When selecting a dataset for the future, one thing we definitely want to make sure of is our confidence in model representation of the data. We want to ask the questions of what actual analysis can be done, and what exactly do we want to pull out, and to not leave those questions to be answered during the EDA process. It can cause a lot of unnecessary complications during model building, as we did experience.

Appendix:

Tanmay Kapse- Tables and Statistics from MLR model that used inputting random values for missing values from datasets, and also looking at only risk factors.

Average CV MSE (Linear Regression): 10.183054936572074
RMSE: 3.2884707939451543
R² value: 0.031324815663193895
Coefficients:
Nutrition: -0.046
Cholesterol_Abnormalities: 0.024
Diabetes: 0.175
Hypertension: 0.017
Obesity: 0.113
Physical_Inactivity: -0.068
Smoking: 0.140

Shows that we do not have a linear relationship between risk factors and disease prevalence. r^2 value is very very low, and I had to choose a different model (XGBoost Trees) to provide a better, if not perfect fit to the data. When I took a look at the model output for demographics, I got an unrealistic value of .99 for r^2, for which overfitting could be an issue. The data inputted for missing values could also be not as random as we thought it would be, leading to over-smoothing.

Average CV MSE (Linear Regression): 0.020256833266854367
RMSE: 0.08454027947255516
R² value: 0.9993597972221234
Coefficients:
Disease_Age25_44: -0.132
Disease_Age45_64: 0.008
Disease_Age65+: -0.018
Disease_Age75+: 0.010
Gender_Female: 0.485
Gender_Male: 0.531
Race_Hispanic: -0.047
Race_Non-Hispanic White: 0.032
Race_Non-Hispanic Black: 0.014
Race_Non-Hispanic Asian: -0.009
Race_Other: -0.000

Data Dictionary for the Pivoted Dataset:
Overall_Pred: Overall percentage of people suffering from a cardiovascular disease
Year: Year in which data was taken
LocationDesc: State in which data was taken
DiseaseType: Type of Cardiovascular Disease
Disease_Age25_44: Percentage of people between the ages of 25 and 44 suffering from a cardiovascular disease
Disease_Age45_64: Percentage of people between ages 45 and 64 suffering from a cardiovascular disease
Disease_Age65+: Percentage of people above the age of 65 suffering from a cardiovascular disease
Disease_Age75+: Percentage of people above the age of 75 suffering from a cardiovascular disease
Gender_Female: Percentage of women who are suffering from a cardiovascular disease
Gender_Male: Percentage of Men who are suffering from a cardiovascular disease
Race_Hispanic: Percentage of those who identify as hispanic suffering from a cardiovascular disease
Race_Non-Hispanic White: Percentage of those who identify as White suffering from a cardiovascular disease
Race_Non-Hispanic Black: Percentage of those who identify as Black suffering from a cardiovascular disease
Race_Non-Hispanic Asian: Percentage of those who identify as Asian suffering from a cardiovascular disease

Nutrition: Risk factor that shows prevalence of poor nutrition in the population
Cholesterol Abnormalities: Risk factor that shows prevalence of Cholesterol Abnormalities in the population.
Diabetes: Risk factor that shows prevalence of Diabetes in the population.
Hypertension: Risk factor that shows prevalence of Hypertension in the population.
Obesity: Risk factor that shows prevalence of Obesity in the population.
Smoking: Risk factor that shows prevalence of smoking in the population.
Physical_Inactivity: Risk factor that shows prevalence of Physical Inactivity in the population.

Galav Sharma - R^2 and RMSE values for the different models:

Topic	LR_RMSE	LR_R2	DT_RMSE	DT_R2	RF_RMSE	RF_R2	SVR_RMSE	SVR_R2	GB_RMSE	GB_R2
Acute Myocardial Infarction (Heart Attack)	5.842933	-0.135112	6.421249	-0.370932	5.935286	-0.171278	5.793335	-0.115923	6.362121	-0.345801
Coronary Heart Disease	0.859307	-0.045140	0.734557	0.236290	0.627704	0.442318	0.788798	0.119339	0.648347	0.405033
Major Cardiovascular Disease	1.565757	-0.038860	1.105223	0.482383	1.031165	0.549427	1.230266	0.358633	0.995762	0.579836
Stroke	0.685589	0.038589	0.463654	0.560288	0.479176	0.530353	0.518239	0.450661	0.447834	0.589782

Data Dictionary of Queried Subset:
Topic - Disease Type
Overall_Overall - Overall percentage of people with disease in state in that year
CO Mean - The yearly average of the mean concentration of CO per day
SO2 Mean - The yearly average of the mean concentration of SO2 per day
O3 Mean - The yearly average of the mean concentration of O3 per day
NO2 Mean - The yearly average of the mean concentration of NO2 per day
CO AQI - The yearly average of the air quality index of CO per day
SO2 AQI - The yearly average of the air quality index of SO2 per day
O3 AQI - The yearly average of the air quality index of O3 per day
NO2 AQI - The yearly average of the air quality index of NO2 per day

Pratyush Joshi

Data Dictionary of Queried Subset:
Year - Year the data took place
CO Mean - The yearly average of the mean concentration of CO per day
SO2 Mean - The yearly average of the mean concentration of SO2 per day
O3 Mean - The yearly average of the mean concentration of O3 per day
NO2 Mean - The yearly average of the mean concentration of NO2 per day

R^2 and MSE for different linear regressions:
NO2 vs Year:

```
Slope: -0.1895723133108948 | Intercept: 390.2632577096096
Mean Squared Error (MSE): 0.040640449575650245
R-squared (R²): 0.6388022025456153
```

CO vs Year:

```
Slope: -0.0020972305534285333 | Intercept: 4.458834604455411
Mean Squared Error (MSE): 2.3977289956890896e-05
R-squared (R²): 0.2684060758552407
```

SO2 vs Year:

```
Slope: -0.005746243207755666 | Intercept: 12.15283492083739
Mean Squared Error (MSE): 0.001110869574651929
R-squared (R²): 0.056111956899390636
```

O3 vs Year:

```
Slope: -0.0004094324180368709 5 | Intercept: 0.8511918117405515
Mean Squared Error (MSE): 8.255071292950233e-08
R-squared (R²): 0.8024254233023131
```