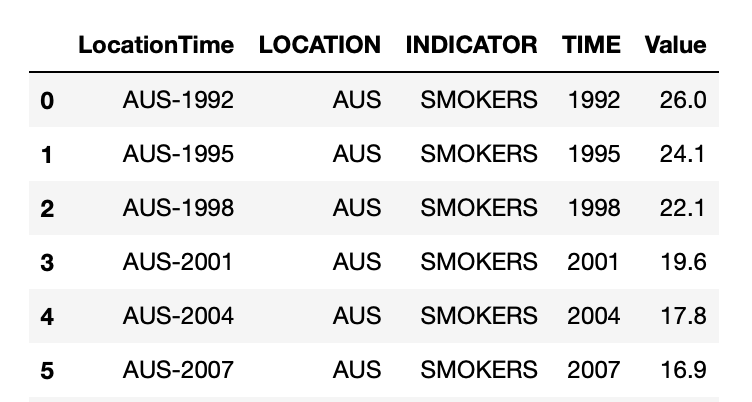
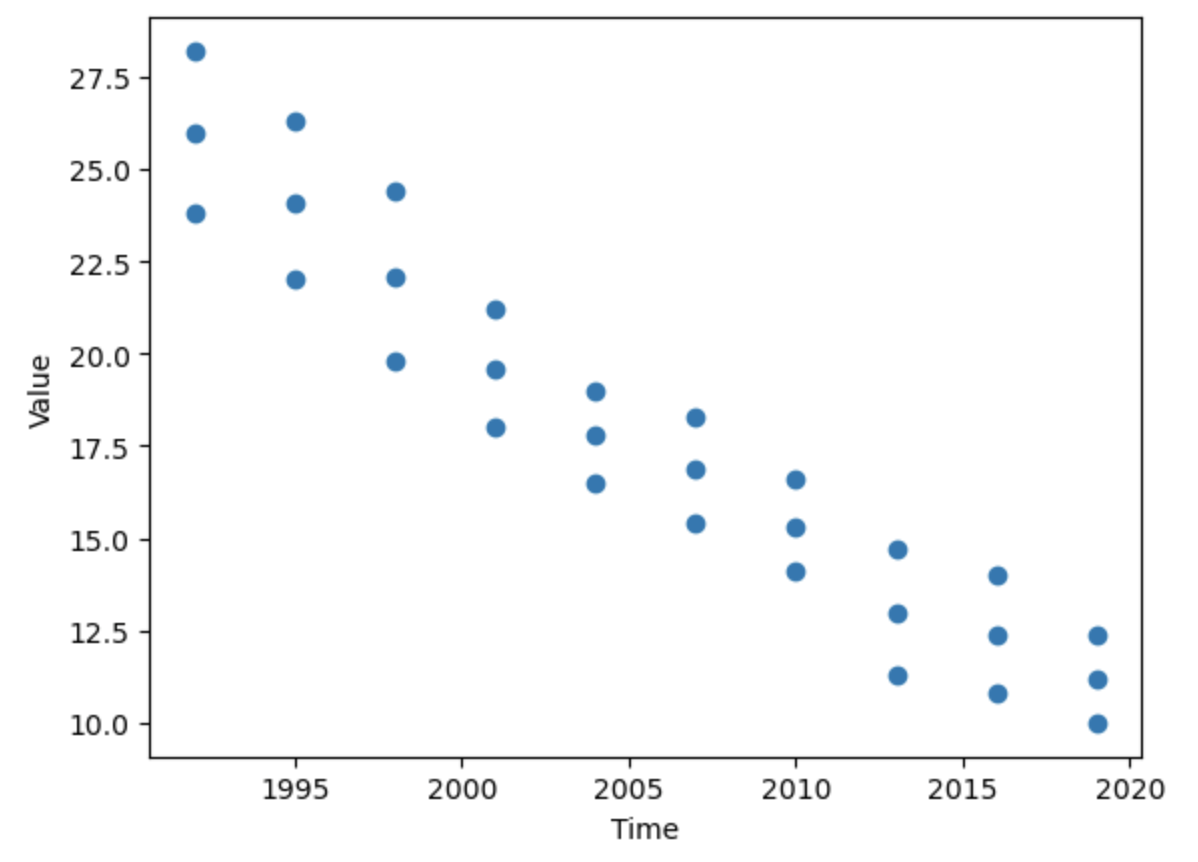
The aim of our project is to develop a model to predict suicide rates in 1, 3, and 5 years Organisation for Economic Co-operation and Development (OECD) countries while also identifying determinants with the most impact. The purpose is to hopefully gain insight on major determinants that heavily increase or decrease suicide rates in these countries. To achieve this goal, our team has decided to utilise a Supervised Learning Multiple Linear Regression Model. A Supervised Learning Regression Model strengths exhibiting the relationships between variables while also being able to predict values. We feel this type of model is the best fit to accomplish our goal of creating a model for predicting suicide risk in the next 1, 3, and 5 years across OECD/select countries. In addition to our Regression Model our team has decided to add a Random Forest component to find the importance of each variable our team is testing for. This will not only add analytical depth, but also make our model useful when identifying role players in suicide rates. This will make our model a valuable tool in not only forecasting a trends of suicide rates, but also spotlighting the potential characters that will have a lasting effect on people’s health.

**Training the model:**

At first we start with an exploratory data analysis of our datasets. We used our smoking\_clean.csv dataset as an example. After loading the data we filtered it on a specific country. In this case Australia. AUS.



With the filtered data on a country we plotted a graph with the TIME as independent variable and the VALUE as dependent variable.



Here we can see that in Australia the number of smokers is decreasing in our chosen time between 1990 and 2020.

In our further analysis we want to do a similar exploratory analysis for our other datasets. (Education, Alcohol, Healthy Spending, …)

After the analysis we know and understand the trends in the countries. Now we can start our supervised machine learning regression model and use the data to train it. The supervised model will give us an outlook and will predict until the year 2025 how the features like alcohol or smoking are impacting the society.

With this trends discovered and predicted and also the suicide rate predicted we can merge the charts and see which factors are correlating with an increasing or decreasing suicide rate in the countries.

Independent Variable:

* Time

Dependent Variables:

* Alcohol
* Smoking
* Education
* Employment
* Healthy\_Spending
* Suicide\_Rate
* Obesity
* Poverty\_Rate
* Social\_Support
* Social\_Spending

**Technology: Jupyter Notebook / Google Colab**

Mock up Code of Multiple Linear Regression Model:

#Importing dependencies/library

Import pandas as pd

from pathlib import Path

import matplotlib.pyplot as plt

From sklearn.linear\_model import LinearRegression

import statsmodels.api as sm

#Load Data

Df = pd.read\_csv(“ csv path”)

#Split data

X = df.([‘time\_column’]).values.reshape(-1,1)

y = df.([‘suicide\_column’]), df.([‘alcohol\_column’]), …

#Model

model = LinearRegression()

model.fit(X,y)

y\_pred = model.predict(X)

#Plotting

plt.scatter(X, y)

plt.plot(X, y\_pred, color='red')

plt.show()

#Slope and intercept

print(model.coef\_)

print(model.intercept\_)

#summary

print\_model = model.summary()

print(print\_model)

#statsmodels

X = sm.add\_constant(X)   
model = sm.OLS(y, X).fit()  
predictions = model.predict(X)

print\_model = model.summary()  
print(print\_model)

Random Forest Model:

# import dependencies

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.ensemble import RandomForestClassifer

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import confusion\_matrix, accuracy\_score, classification\_report

# Loading the data

suicide\_data\_df = pd.read\_csv('suicide\_data.csv')

alcohol\_data\_df = pd.read\_csv('alcohol\_data.csv')

# Merging the data

data\_df = pd.merge(suicide\_data, alcohol\_data, on='country')

# Splitting the data into features and target

X = data.drop(['suicide\_rate'], axis=1)

y = data['suicide\_rate']

# Splitting the data into training and test sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=0)

# Creating the model and training on the data

rf = RandomForestRegressor(n\_estimators=100, random\_state=0)

rf.fit(X\_train, y\_train)

# Making predictions on the test set

y\_pred = rf.predict(X\_test)

# Evaluating the model's performance

mae = mean\_absolute\_error(y\_test, y\_pred)

print('Mean Absolute Error:', mae)

# Scatter plot to visualize relationship between suicide rate and alcohol consumption

plt.scatter(data['alcohol\_consumption'], data['suicide\_rate'])

plt.xlabel('Alcohol Consumption')

plt.ylabel('Suicide Rate')

plt.show()