**Air pollution**

Author: Jose Manuel Alvarez Gonzalez

Email Id: jmanuela.ag89@gmail.com

Date: 14/2/2022

**Introduction**

For this project we are going to be analysing a data set for the air pollution of 80 different cities in the USA. The main propose of this project is to understand the total mortality rate (TMR) of the data set, we have a total of 11 other variables that relate to measurements of air pollution and the population distributions of each city.  
  
What we are going to be doing on this project, is to identify how the different variables of each city, affect the TMR and identify any relation to it. The way we are going to be performing our analysis, is with the use of different analytical techniques to help us identifying the relationship of the data and also make predictions with it.

**Motivation and Methodology**

What we want to archive on this project, is a full understanding of the data set we have, we want to know if any of the variables we have has a strong correlation with the TMR. Next goal is to use the variables found in the data and be able to create accurate predictions for other cities. With the data set, we have only 80 cities, but if we can manage to create accurate predictions, it will be possible to make an efficient assumption of TMR for each city in the country.

First we are going to look at each variable and find the correlation to the TMR, here we could identify the main correlated variables, then we will be running a principal component analysis, this will help us to create a model and work any prediction of TMR from the different variables. We will also do a gradient descent optimisation and find the lowest coefficient cost and the relation with the TMR.

**Results**

The first stage of the analysis is to see how the variables are correlated. For this stage what we are going to do is to run a correlation analysis with each variable. Looking at the information we get, we can find that is some existent relation between the sulphate readings and TMR, the coefficients are on the .35 range. This is a strong relation but not a majorly significant.  
  
The main variable with a high coefficient ratio of .86 is the GE65 (population over 65), this is a really strong relationship and also shows that they are the most vulnerable portion of the population. To showcase this finding, we have a plot of the increase of GE65 and how affects the TMR

**Chart, scatter chart

Description automatically generated**

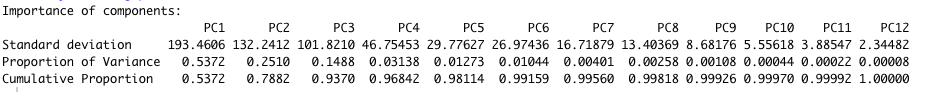
Here we can find a linear relation and we are able to fit a regression line to showcase the relationship. We know this is a significant model by the small P-values and high R-squared values found on the analysis.

**Text, letter

Description automatically generated**

The next step of the analysis is to be able to create predictions from the variables we have, for that we are going to split the data on a 4:1 ratio, we will call them training and testing respectively. For the training data set, we will do a principal component analysis. This will help us reducing the amount of the information and hopefully keep a high value of variance with the first two Principal Components.  
  
After doing the analysis, we can find that with the first two PC, we are able to keep a total of 78.82% of the total information. That’s a good result and we will be able to do some accurate predictions with the information. In the following plot we can see how the variance is spread around the first 10 PC:

**Chart, line chart, scatter chart

Description automatically generated**

In the following plot, we can see the first two PC, I decided to categorise the values according to TMR levels: high for above mean and low for the below values. We can also identify the TMR ~ GE65 relationship:

**Chart, bubble chart

Description automatically generated**

If we decide to include the third PC we could end up with a 93.7% of the total variance, but we will focus only with the first two. The next step to follow, is to create a linear model that predicts the TMR from the 1PC + 2PC, for that we obtain the following formula:

**Text

Description automatically generated**

In the model, the slopes and intercept are significant by their minimal P-values, the intercept of 908.375 equals to the mean value of TMR for the training data set. This means that the PCA helped cluster the information and is not extremely spread out, we can see that is also reflected by the smaller values of the slopes.  
  
Now we can run a prediction of the TMR with the previous model and we will use the testing data set for that. Here we can see the results of our prediction and compare it with the original.  
  
**Text, table

Description automatically generated**

This is an accurate result with no more than a maximum error of 10%

The last part of the project is to minimise the cost function we have, this function uses two variables from the data set PMAX and SMAX. With the power of calculus to derivate the equation for each variable and the use of our gradient descent algorithm to find the minimum cost value, we get to the result that the cost function gets minimized when the values:

PMAX 🡪 120.7

SMAX 🡪 70

The values give the cost function a minimum result of 901.6. Looking at the previous values, we can find that Miami has a similar PMAX and SMAX with a TMR value of 897.

**Conclusion**

After doing some analysis with our data set, we were able to identify some correlation of sulphate levels with the TMR but not sufficient enough to create a model for it. At the same time with the correlation analysis we were able to identify the most vulnerable portion of the population, being the group of 65 and over.

The major goal of this project was successfully accomplish by being able to create a prediction model of the TMR from the two principal components of the data set. We were able make a reliable prediction of TMR value for 20% of the cities with a small margin of error. I will say that it will be better two include the third PC to have a highly effective model.

At the same time we were able to identify the minimum cost of our function and the equivalent TMR value for it, from here we can use the values as a benchmark to reduce the TMR in other cities and keep cost at minimum.