Final Project: Effects of Air Pollution on Countries

DataSci Warriors: Group 5

Our Motivation

Midterm: To explore how air pollution affect countries differently.

Does air pollution affect underdeveloped countries disproportionately?

Final: To dive deeper into our variables to determine if there are interesting models that expand on our findings from the Midterm to give us a more complete picture of how air pollution affect countries.

Data Sources & Variables

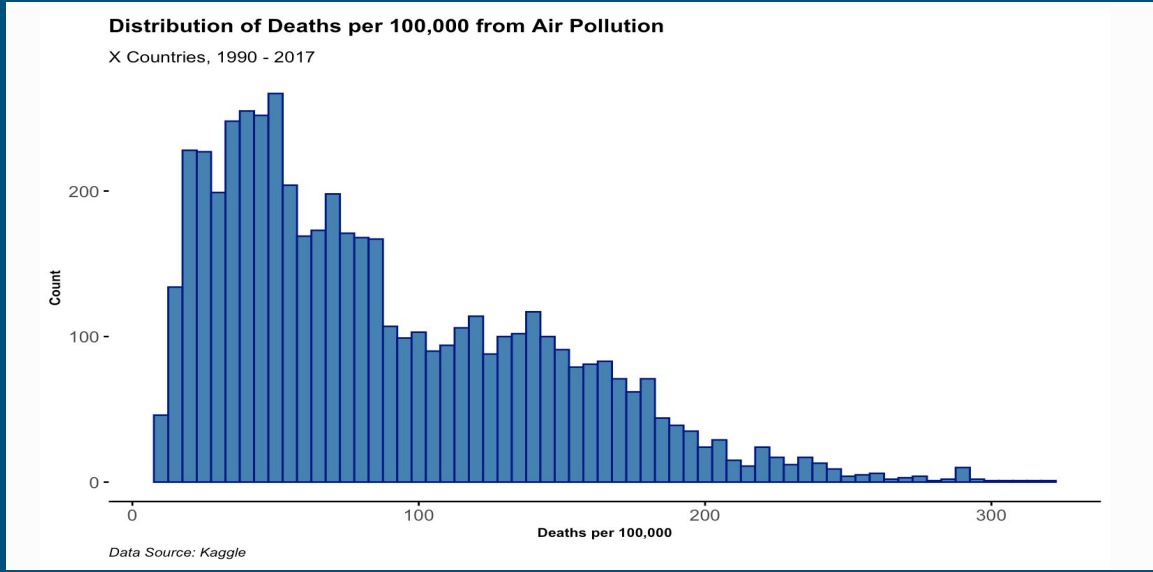
Data	Source
Deaths Due to Air Pollution of Countries from 1990 - 2017	Kaggle
GDP Annual Growth of Countries from 1960 - 2020	Kaggle via WorldBank
United Nations Population and Region Data	United Nations
United Nations ISO-alpha3 code	United Nations
spData for Map Geometries	spData for Mapping

1. GDP : Numerical, Continuous
2. Population Size : Numerical, Continuous
3. Deaths due to Air Pollution : Numerical, Continuous
4. Country : Qualitative, Categorical
5. SDG Region : Qualitative, Categorical
6. Sub Region : Qualitative, Categorical
7. ISO- alpha3 Country Code : Qualitative, Categorical
8. ISO- alpha2 Country Code : Qualitative, Categorical
9. M49 Country Code : Numerical, Categorical
10. Year : Numerical, Categorical
11. GDP per Capita : Numerical, Continuous

E.D.A. Summary

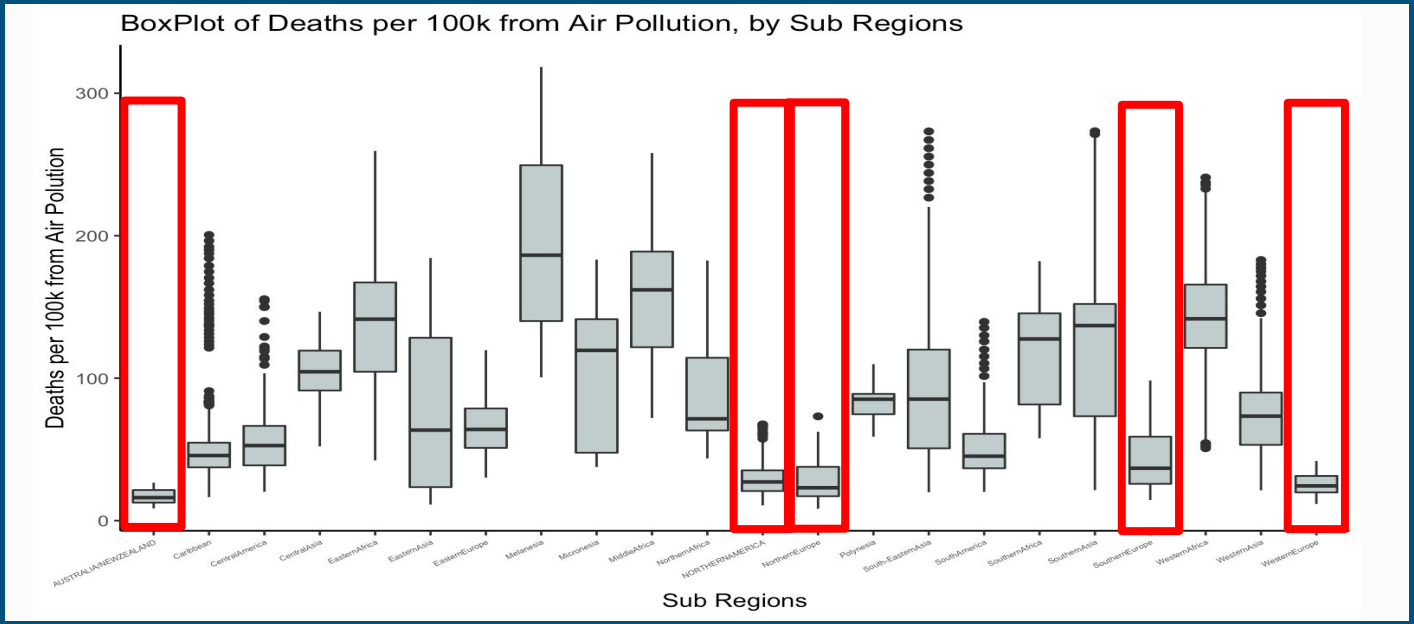
EDA Summary

- Looked at distributions of key numerical features in histograms



EDA Summary

- Looked at boxplots of these key numerical features by SDGRegions and SubRegions



EDA Summary

- Looked at choropleth maps to better visualize our data

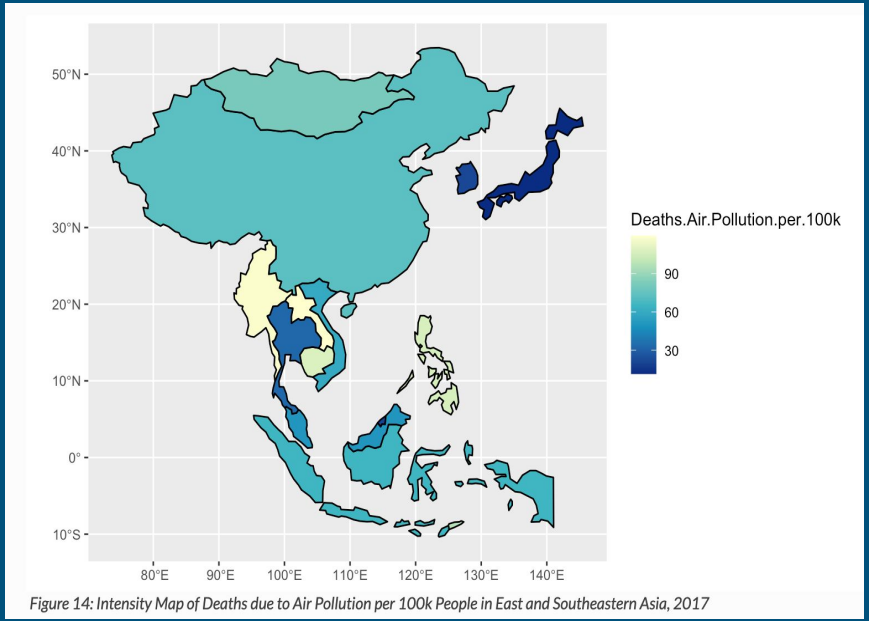
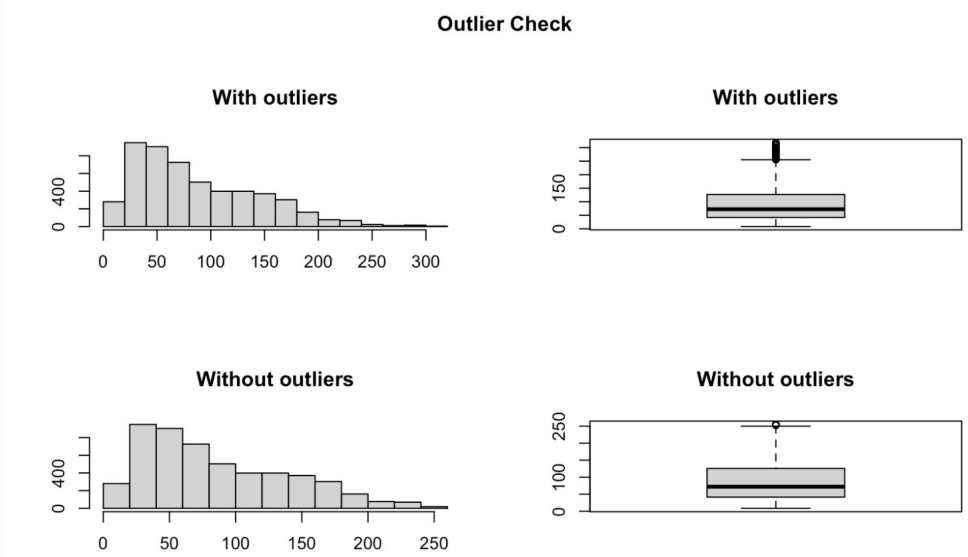


Figure 14: Intensity Map of Deaths due to Air Pollution per 100k People in East and Southeastern Asia, 2017

EDA: Outlier Check
Deaths due to Air Pollution per 100k

Feature 2: Deaths due to Air Pollution per 100k

Figure 23: Inspecting Outliers for Deaths.Air.Pollution.per.100k

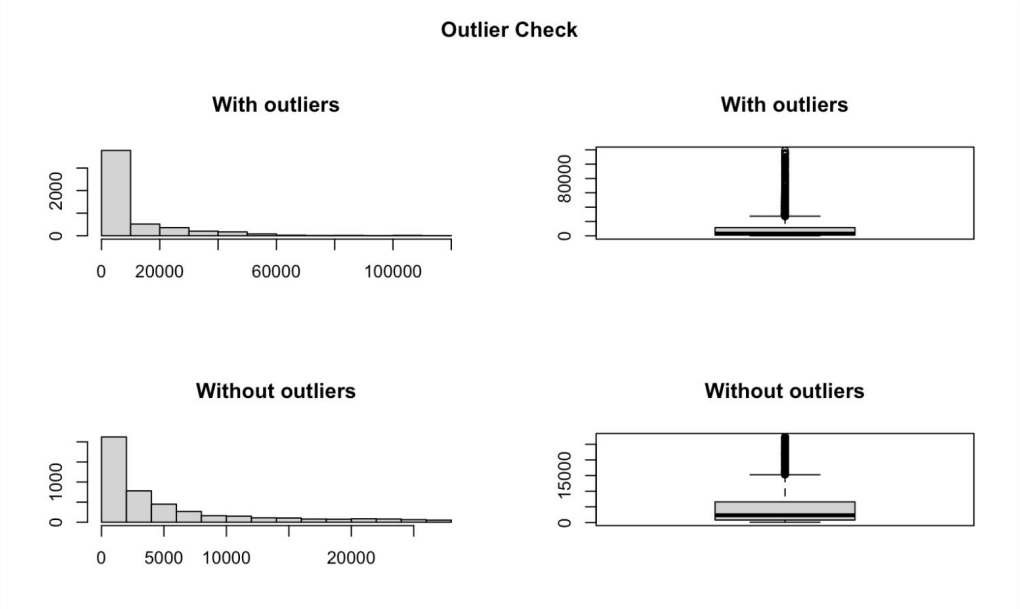



```
## Outliers identified: 36
## Propotion (%) of outliers: 0.7
## Mean of the outliers: 282
## Mean without removing outliers: 87.2
## Mean if we remove outliers: 85.8
## Nothing changed
```

EDA: Outlier Check
GDP per Capita

Feature 1: GDP per Capita

Figure 22: Inspecting Outliers for gdp.per.capita




```
## Outliers identified: 619
## Propotion (%) of outliers: 13.5
## Mean of the outliers: 45676
## Mean without removing outliers: 9961
## Mean if we remove outliers: 5132
## Nothing changed
```

New SMART Questions

1. What are the impacts of population size from GDP and Deaths due to Air Pollution globally?
2. Which countries in Sub-Saharan Africa are more likely to have higher deaths due to air pollution?
3. Can we let the data guide us to the types of groupings that exist in our dataset?
4. Can we make a prediction of the future GDP by considering the indoor and outdoor deaths due to air pollution ?

SMART Q1: What are the impacts of population size from GDP and Deaths due to Air Pollution globally?

Categorized Population

Low(0): Population.thousands >= 0 & Population.thousands <= 50000 ~ 0

Medium(1): Population.thousands >= 100000 & Population.thousands <= 50001 ~ 1

High(2): Population.thousands >= 100001 & Population.thousands <= 1000000000 ~ 2

Logit Model GDP as a Predictor on Population

```
##
## Call:
## glm(formula = population_grouping ~ groupings_pop$gdp.per.capita,
##     data = groupings_pop)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.246  -0.185  -0.181  -0.180   1.820
##
## Coefficients:
##                                  Estimate  Std. Error t value
## (Intercept)                   0.180177461 0.008492388   21.22
## groupings_pop$gdp.per.capita 0.000000553 0.000000456    1.21
##                                          Pr(>|t|)
## (Intercept)                   <0.0000000000000002 ***
## groupings_pop$gdp.per.capita                0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.267)
##
##     Null deviance: 1389.8  on 5196  degrees of freedom
## Residual deviance: 1389.4  on 5195  degrees of freedom
## AIC: 7899
##
## Number of Fisher Scoring iterations: 2
```

Logit Model GDP + Deaths due to Air Pollution as a Predictor on Population

```
##
## Call:
## glm(formula = population_grouping ~ groupings_pop$gdp.per.capita +
##     groupings_pop$Deaths.Air.Pollution.per.100k, data = groupings_pop)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.215  -0.194  -0.186  -0.169   1.842
##
## Coefficients:
##                                                 Estimate   Std. Error t value
## (Intercept)                                  0.202910224  0.018188095   11.16
## groupings_pop$gdp.per.capita                 0.000000136  0.000000543    0.25
## groupings_pop$Deaths.Air.Pollution.per.100k -0.000213150  0.000150811   -1.41
##                                                        Pr(>|t|)
## (Intercept)                                  <0.0000000000000002 ***
## groupings_pop$gdp.per.capita                                0.80
## groupings_pop$Deaths.Air.Pollution.per.100k                 0.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.267)
##
##     Null deviance: 1389.8  on 5196  degrees of freedom
## Residual deviance: 1388.9  on 5194  degrees of freedom
## AIC: 7899
##
## Number of Fisher Scoring iterations: 2
```

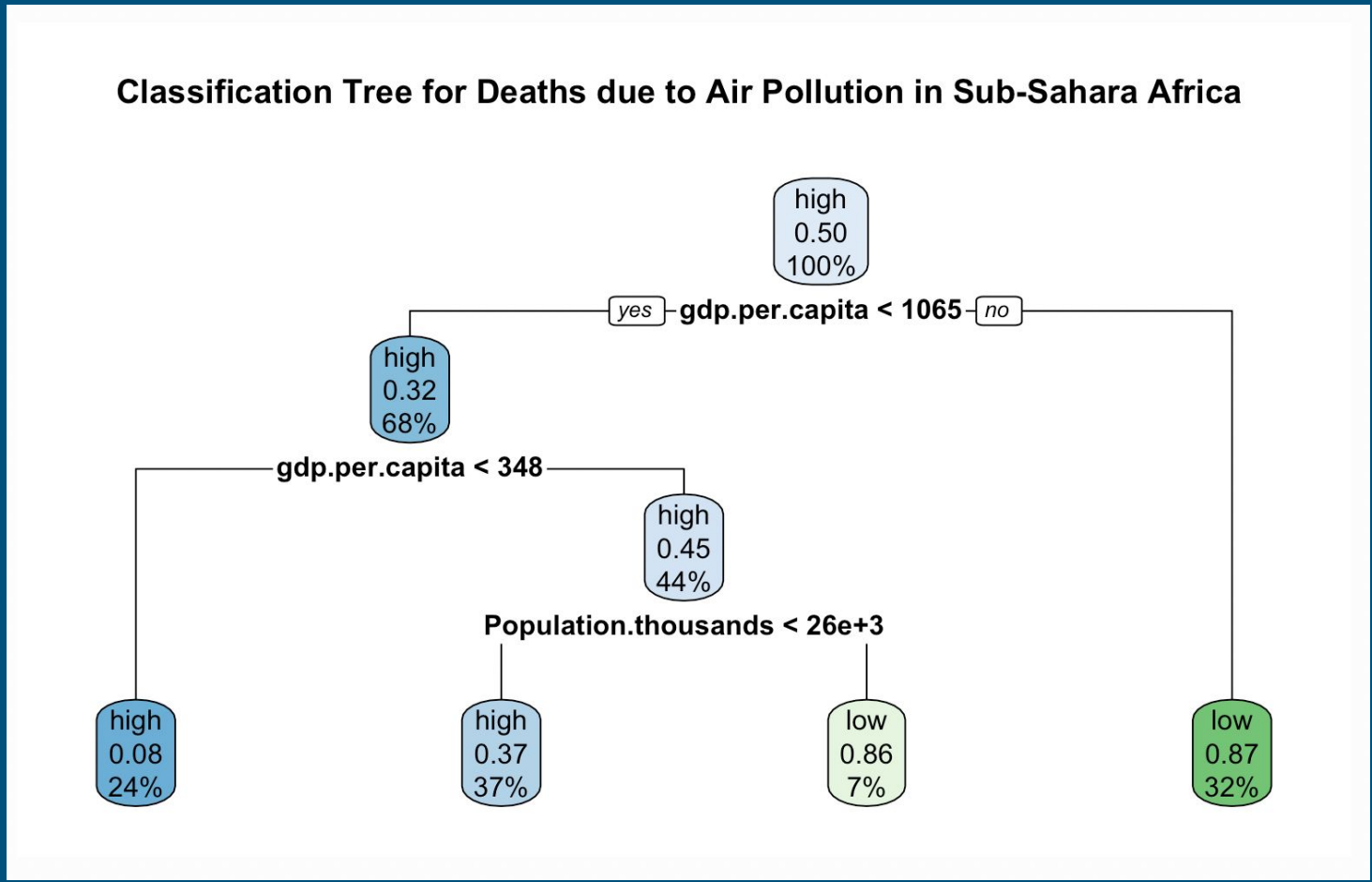
SMART Q2: Which countries in Sub-Saharan Africa are more likely to have higher deaths due to air pollution?

Decision Tree Classification Algorithm

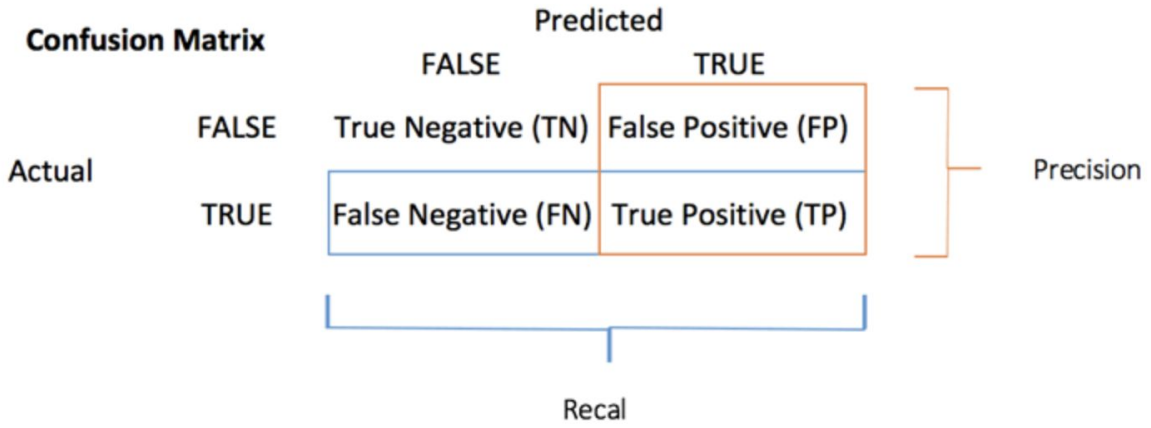
SMART Q2: Data

- Subset the data
- New variable - “deaths”
- Train and test set - 80/20
- Verify Randomization process (low deaths in both sets are about 49%)

SMART Q2: Train



SMART Q2: Confusion Matrix and Accuracy Test I



Confusion Table

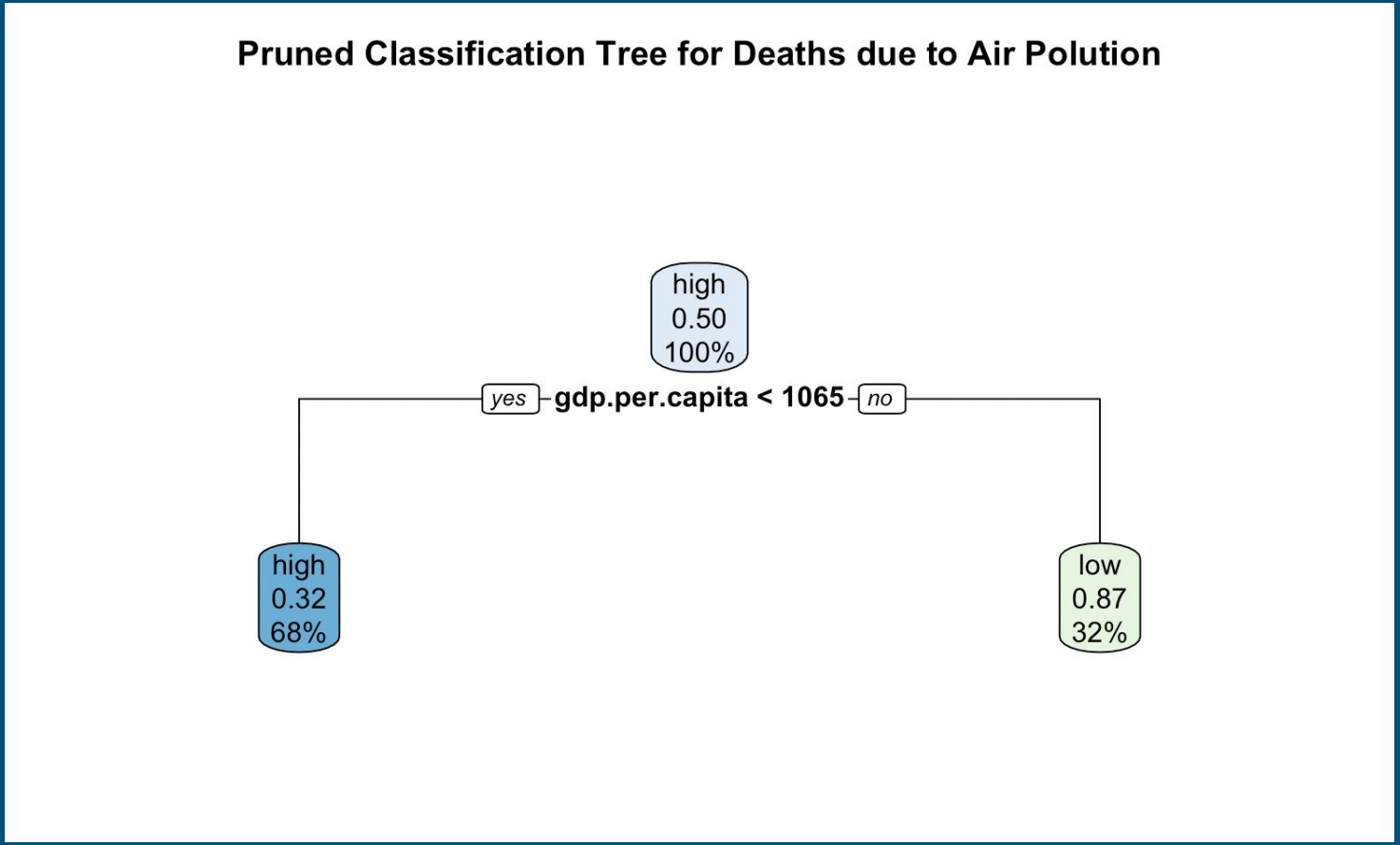
Confusion Matrix

	high	low
high	110	49
low	15	74

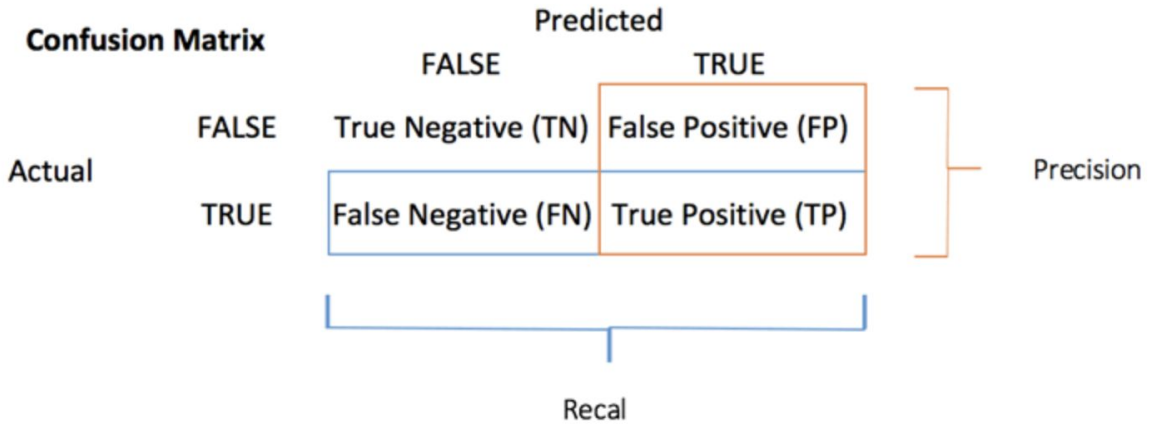
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

74%

SMART Q2: Prune the Tree



SMART Q2: Confusion Matrix and Accuracy Test II



Confusion Table

Confusion Matrix II

	high	low
high	111	63
low	14	60

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

74%

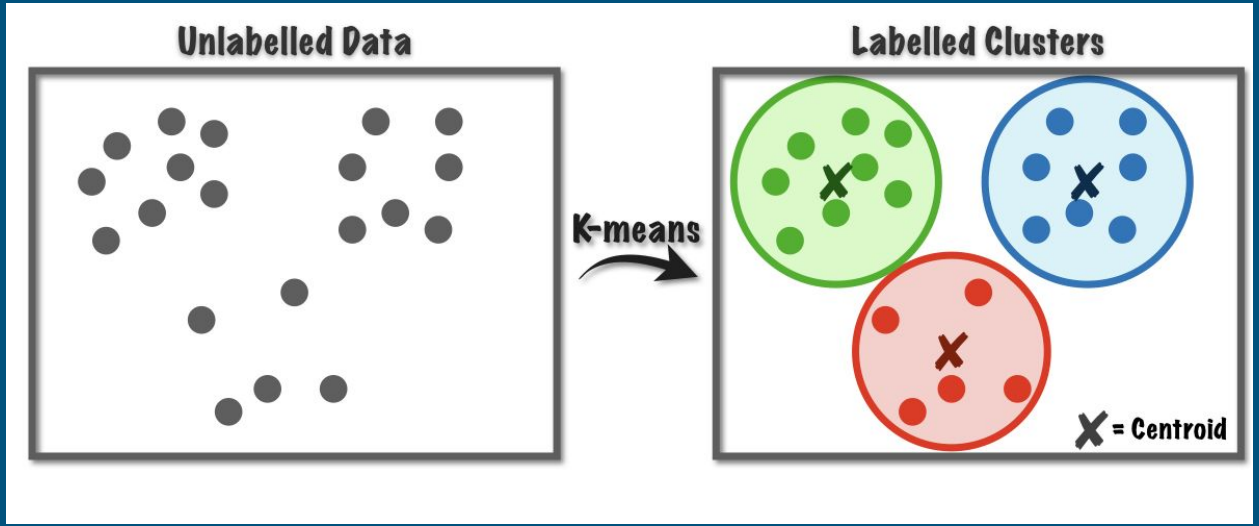
SMART Q2: How Can We Improve this Model?

Instead of of an individual tree, we can employ other techniques such as bagging, random forests, and boosting, which will significantly improve our predictive performance!

SMART Q3: Can we let the data guide us to the types of groupings that exist in our dataset?

SMART Q3: Can we let the data tell us what type of groupings exist in our dataset?

Let's try a K-Means Clustering Algorithm!



Source: Alan Jeffares, TowardsDataScience

SMART Q3: Finding Optimal K

Elbow Method with Total Within Sum of Squares

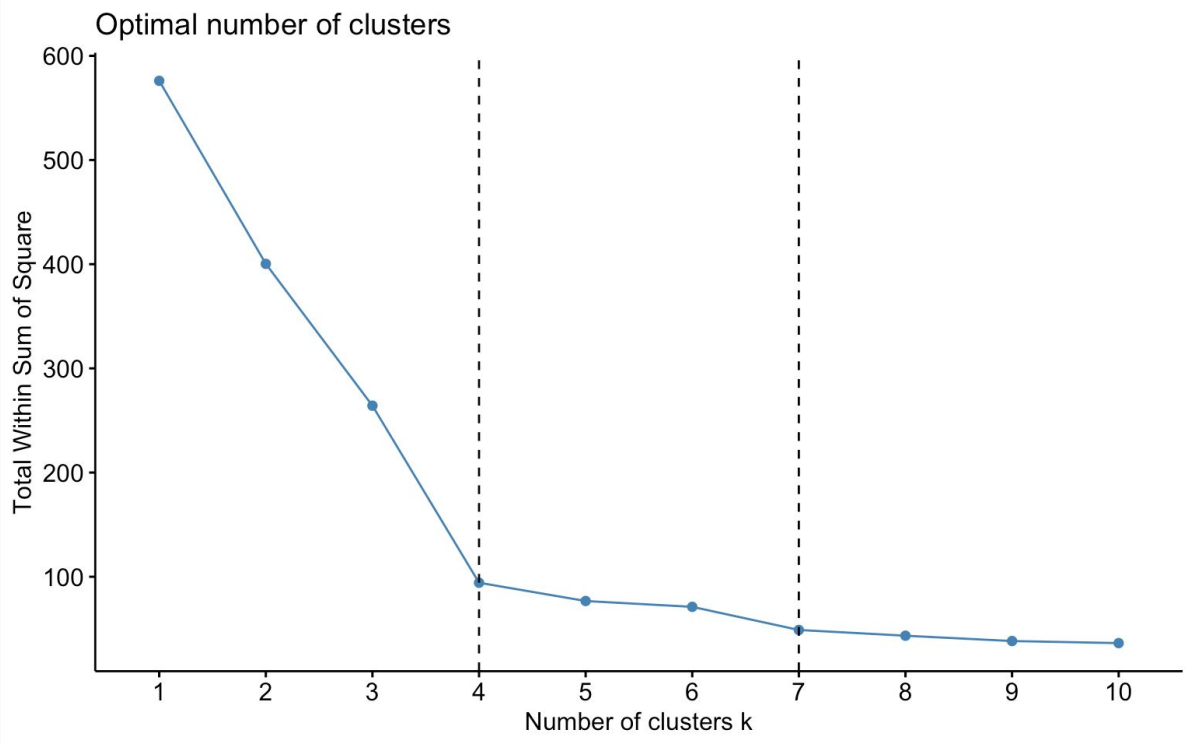


Figure 6.3-2: Finding the optimal number of clusters using the Within Sum of Squares Method.

SMART Q3: Finding Optimal K

Looking at the Gap Statistics

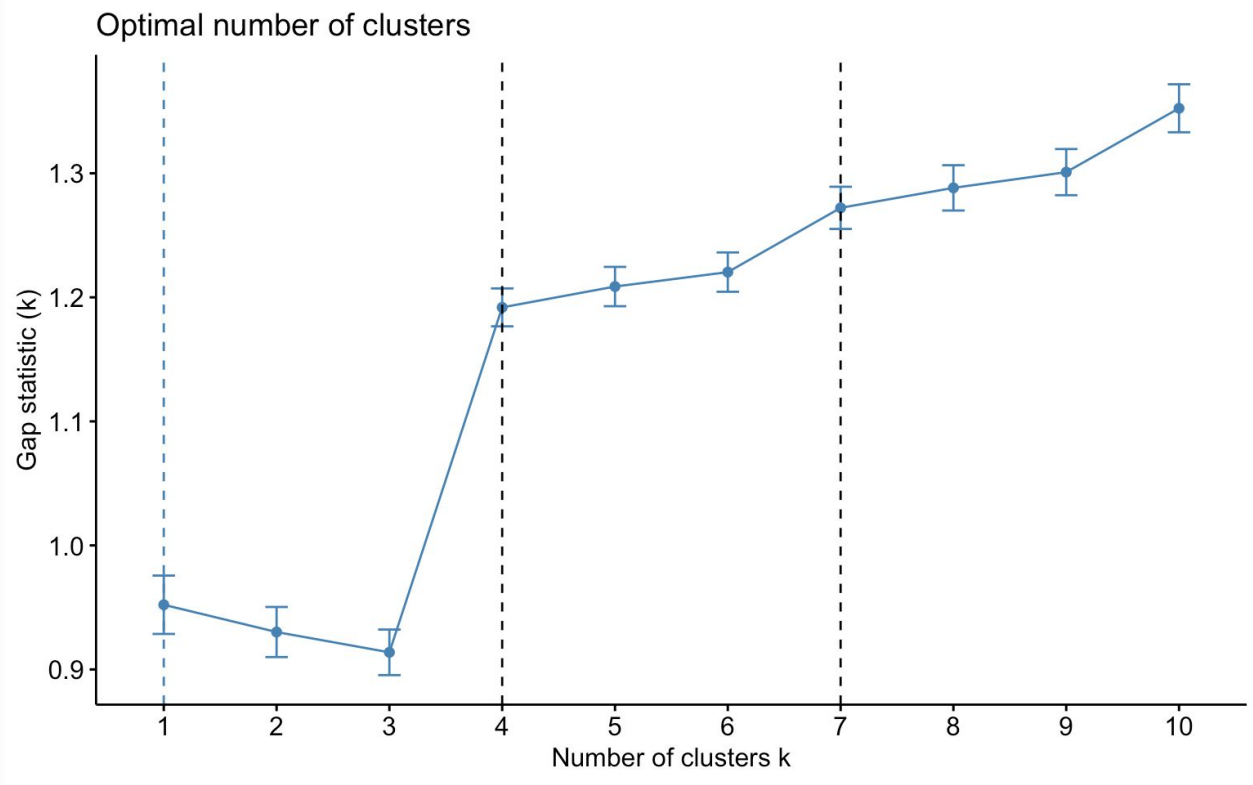


Figure 6.3-3: Finding the optimal number of clusters using the Gap Statistic.

SMART Q3: K-Means Clustering Results

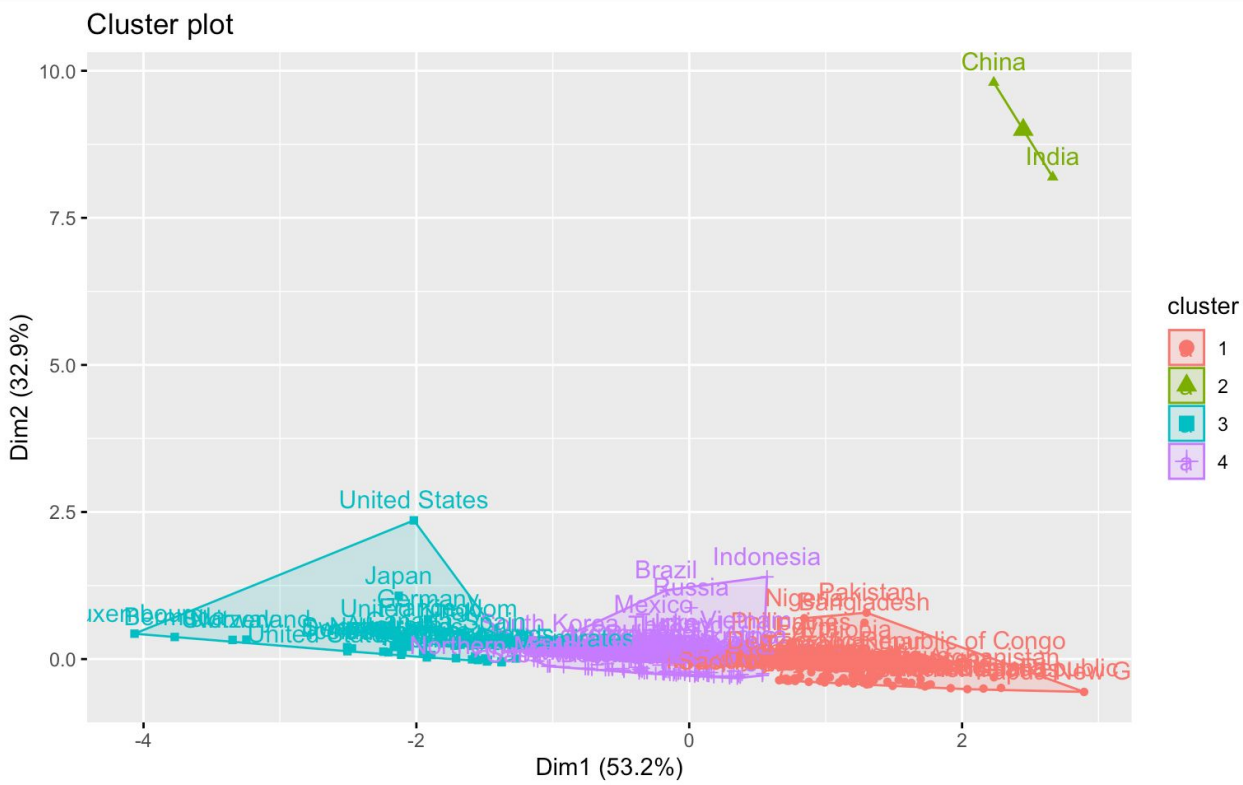


Figure 3.6-7: Visualing K-Means Clustering Model.

SMART Q3: K-Means Clustering Results

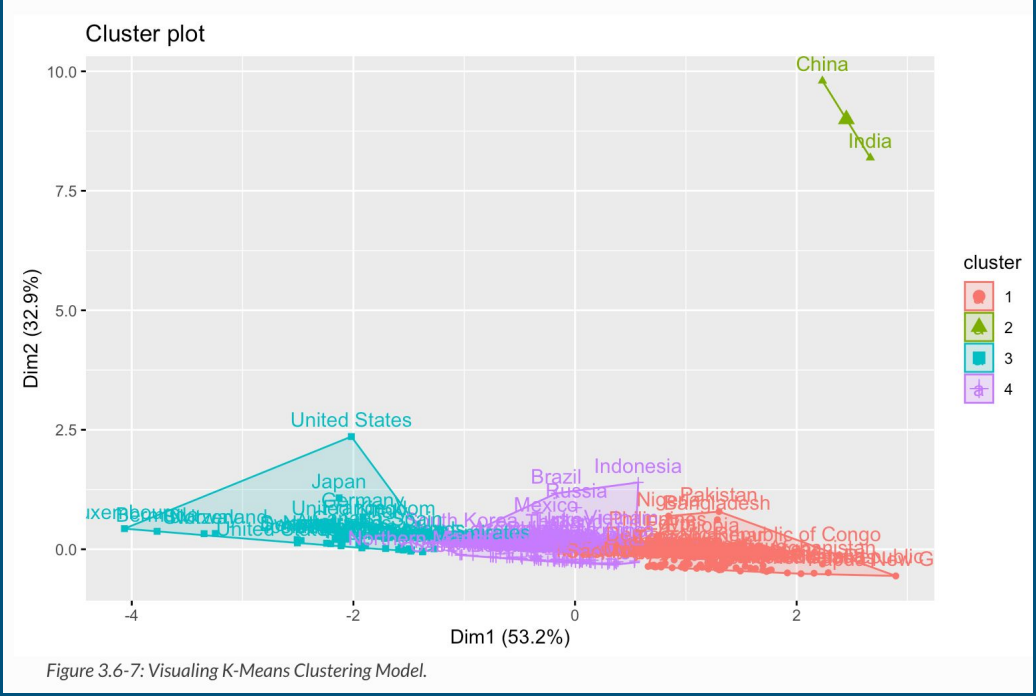


Figure 3.6-7: Visualing K-Means Clustering Model.

Figure 6.3-4: K-Means Clustering Average Values per Cluster

Cluster	Deaths.Air.Pollution.per.100k	gdp.per.capita	Population.thousands
1	1.16234711051454	-0.61738173258503	-0.101796957671015
2	1.01091758337588	-0.564333841794921	9.25163623841697
3	-1.04092828247631	1.92017020504228	-0.0578237235004342
4	-0.494528110744794	-0.253250480051675	-0.107965219042902

SMART Q3: SDGRegions in Clusters

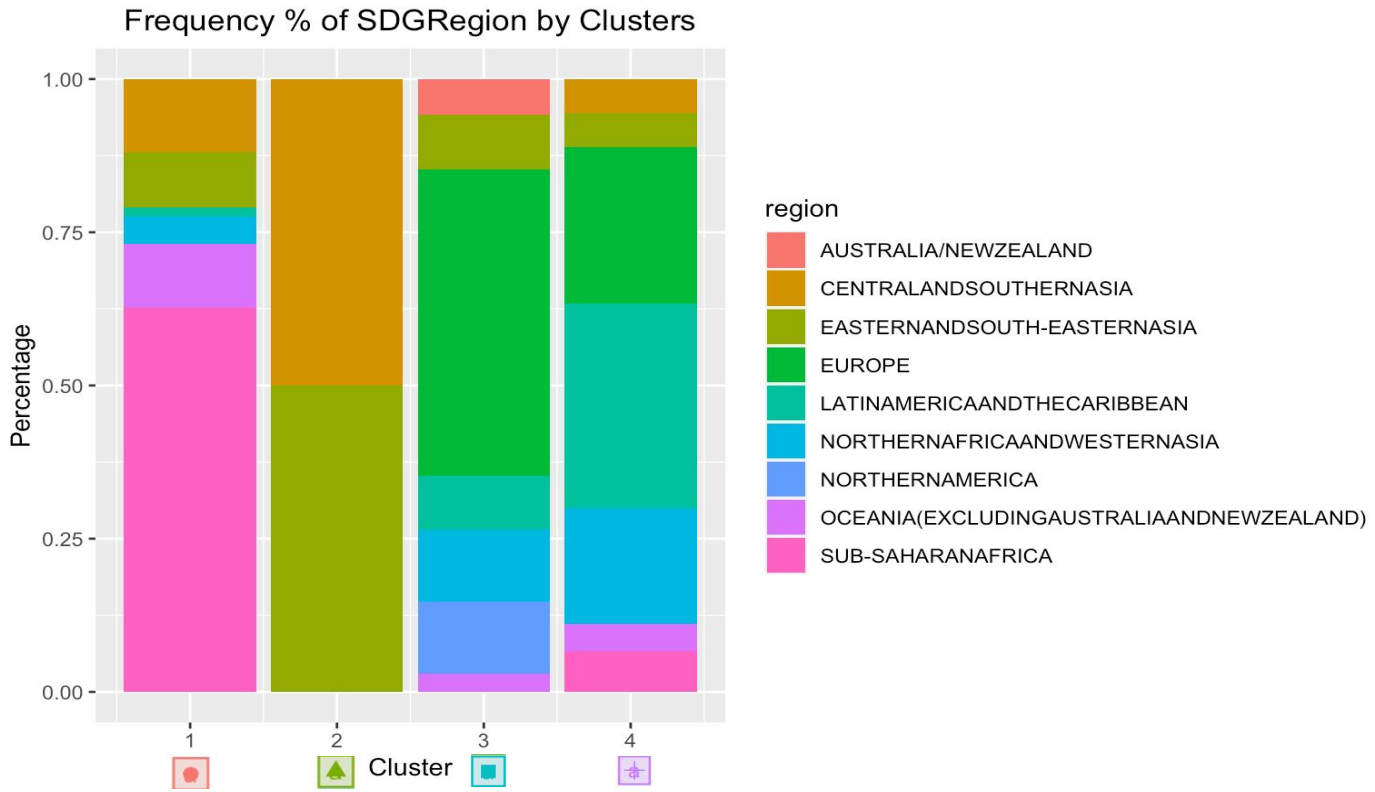


Figure 6.3-8: Frequency % of SDGRegion by Clusters.

SMART Q3: SDGRegions in Clusters

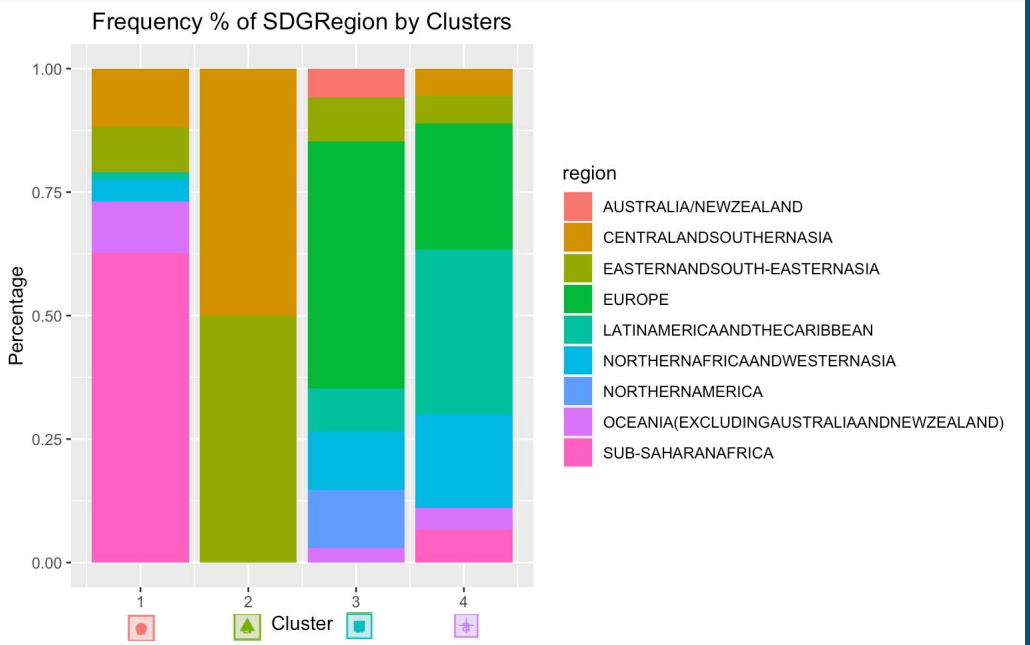


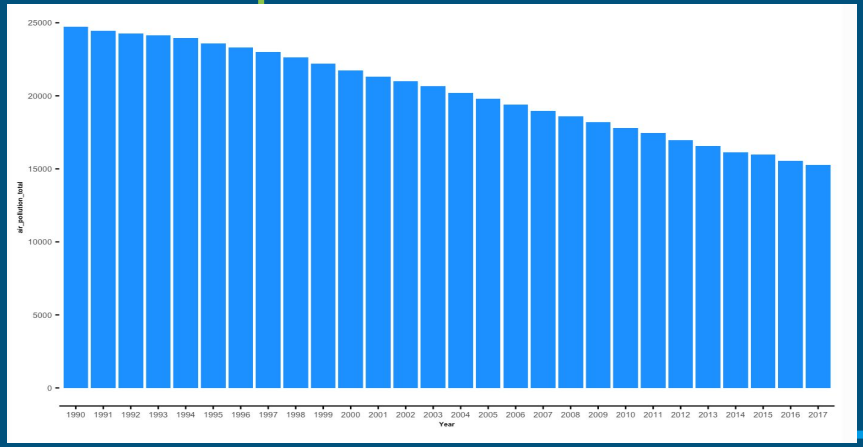
Figure 6.3-8: Frequency % of SDGRegion by Clusters.

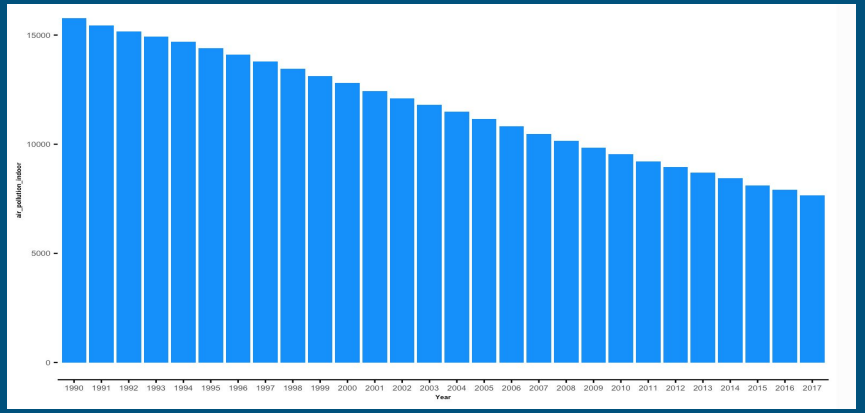
Figure 6.3-4: K-Means Clustering Average Values per Cluster

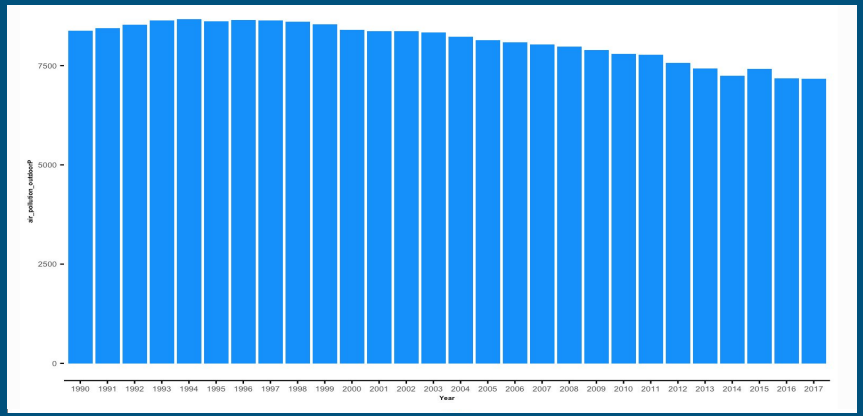
Cluster	Deaths.Air.Pollution.per.100k	gdp.per.capita	Population.thousands
1	1.16234711051454	-0.61738173258503	-0.101796957671015
2	1.01091758337588	-0.564333841794921	9.25163623841697
3	-1.04092828247631	1.92017020504228	-0.0578237235004342
4	-0.494528110744794	-0.253250480051675	-0.107965219042902

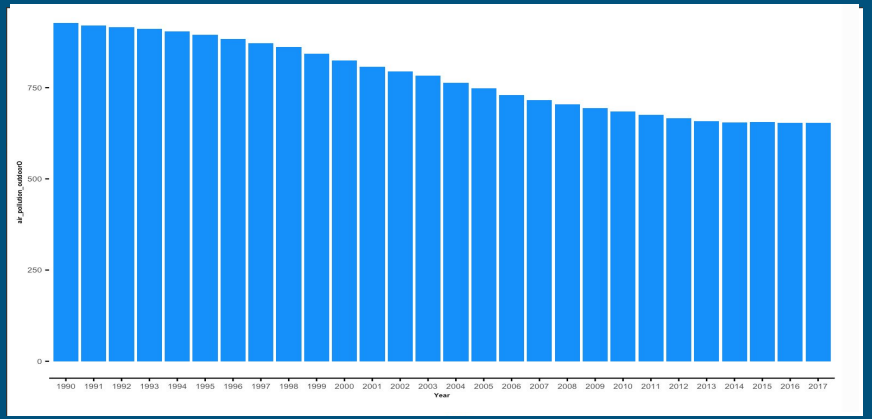
SMART Q4: Can we make a prediction of the future GDP by considering the indoor and outdoor deaths due to air pollution?

Distributions of pollution deaths per year









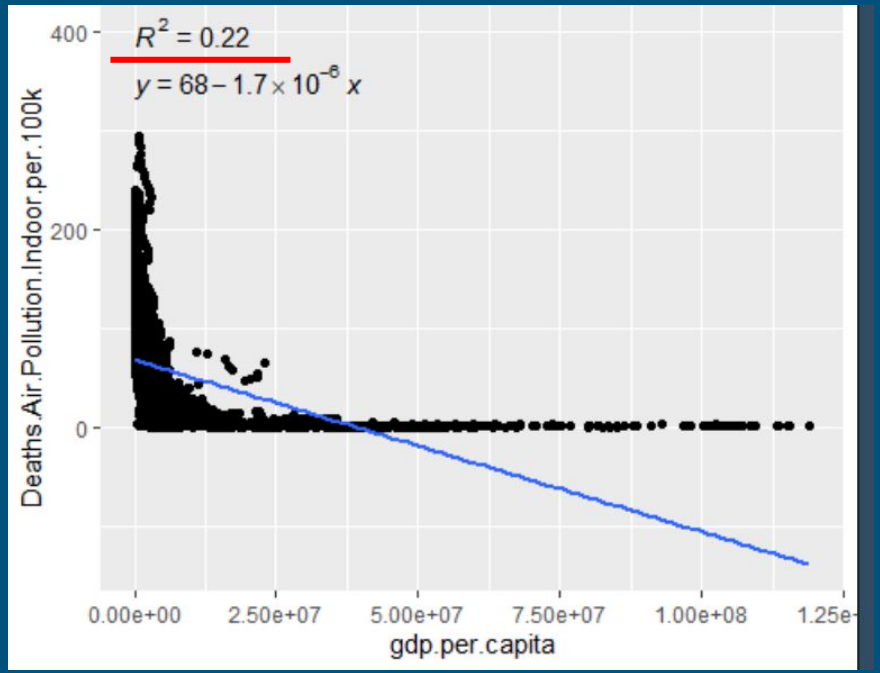
Indoor air pollution

```
Call:
lm(formula = gdp.per.capita ~ Deaths.Air.Pollution.Indoor.per.100k,
    data = final_df_sf)

Residuals:
      Min        1Q    Median        3Q       Max
-15507992  -8789500  -3652179   3500158 102785728

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       16339638     254687    64.2   <2e-16 ***
Deaths.Air.Pollution.Indoor.per.100k  -126644       3308   -38.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13900000 on 5195 degrees of freedom
Multiple R-squared:  0.22,      Adjusted R-squared:  0.22
F-statistic: 1.47e+03 on 1 and 5195 DF,  p-value: <2e-16
```



Indoor air pollution (log)

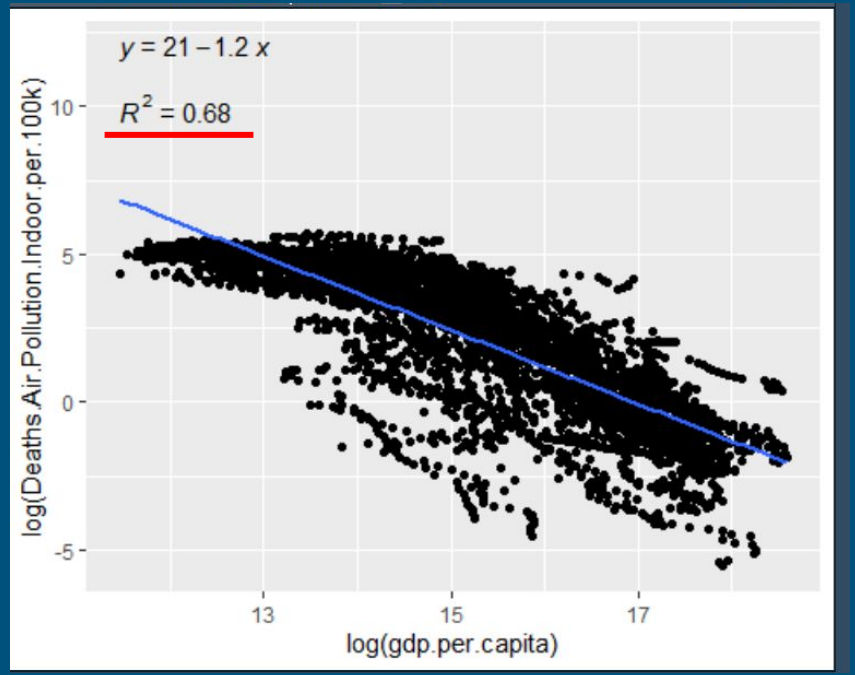
```
Call:
lm(formula = log(gdp.per.capita) ~ log(Deaths.Air.Pollution.Indoor.per.100k),
    data = final_df_st)

Residuals:
   Min     1Q Median     3Q    Max
-3.353 -0.599  0.098  0.649  2.892

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         16.33254    0.01763     927   <2e-16
log(Deaths.Air.Pollution.Indoor.per.100k) -0.54760    0.00517    -106   <2e-16

(Intercept)                         ***
log(Deaths.Air.Pollution.Indoor.per.100k) ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.891 on 5195 degrees of freedom
Multiple R-squared:  0.684,     Adjusted R-squared:  0.684
F-statistic: 1.12e+04 on 1 and 5195 DF,  p-value: <2e-16
```



Outdoor air pollution

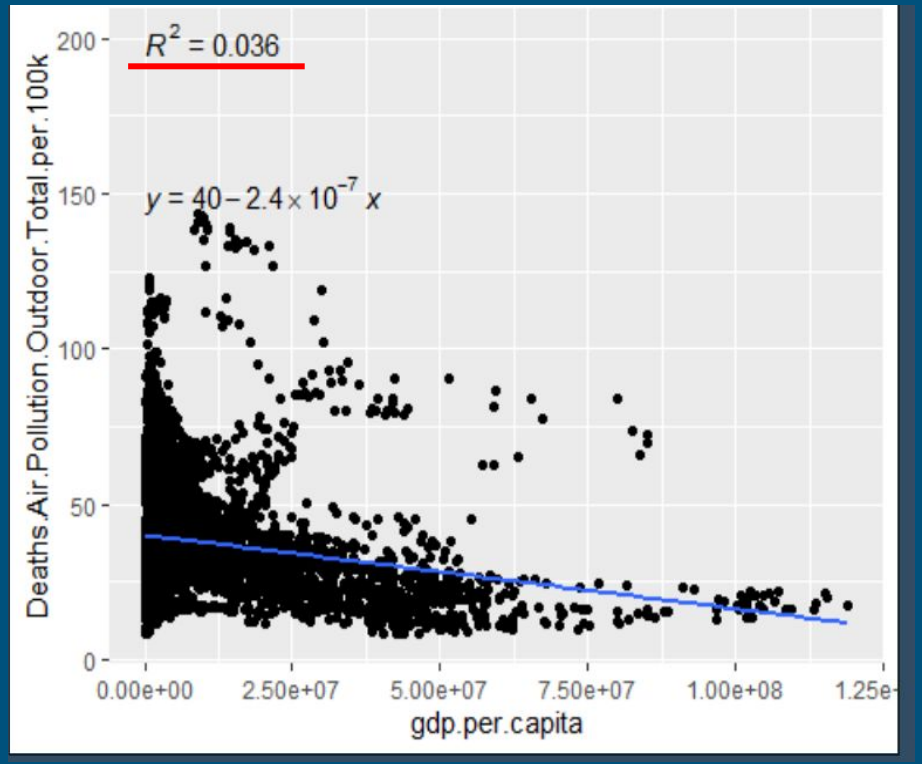
```
Call:
lm(formula = gdp.per.capita ~ Deaths.Air.Pollution.Outdoor.Total.per.100k,
    data = final_df_sf)

Residuals:
      Min        1Q    Median        3Q       Max
-14051002  -9537666  -5336112   2524943 106063848

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               15642957     461887    33.9   <2e-16
Deaths.Air.Pollution.Outdoor.Total.per.100k  -150380      10831   -13.9   <2e-16

(Intercept)                               ***
Deaths.Air.Pollution.Outdoor.Total.per.100k ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15400000 on 5195 degrees of freedom
Multiple R-squared:  0.0358,    Adjusted R-squared:  0.0356
F-statistic:  193 on 1 and 5195 DF,  p-value: <2e-16
```



Outdoor air pollution (log)

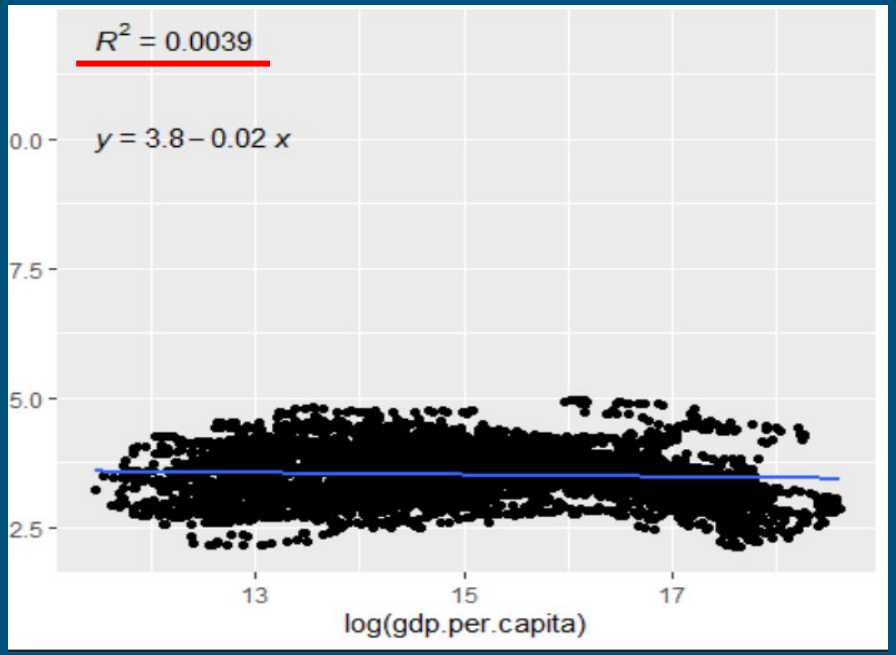
```
Call:
lm(formula = log(gdp.per.capita) ~ log(Deaths.Air.Pollution.Outdoor.Total.per.100k),

    data = final_df_sf)

Residuals:
   Min     1Q Median     3Q    Max
-3.591 -1.272 -0.027  1.297  3.471

Coefficients:
                                              Estimate Std. Error t value
(Intercept)                                    15.6988     0.1565  100.28
log(Deaths.Air.Pollution.Outdoor.Total.per.100k)  -0.1991     0.0442   -4.51
                                              Pr(>|t|)
(Intercept)                                    < 2e-16 ***
log(Deaths.Air.Pollution.Outdoor.Total.per.100k)  6.8e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.58 on 5195 degrees of freedom
Multiple R-squared:  0.00389,   Adjusted R-squared:  0.0037
F-statistic: 20.3 on 1 and 5195 DF,  p-value: 6.75e-06
```



Indoor and outdoor air pollution

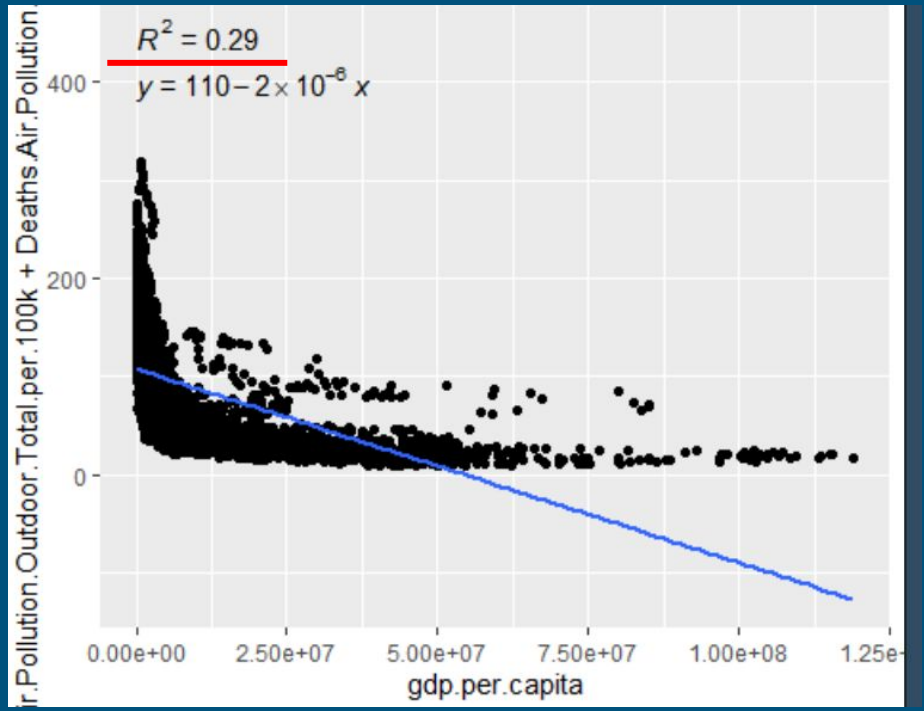
```
Call:
lm(formula = gdp.per.capita ~ Deaths.Air.Pollution.Indoor.per.100k +
    Deaths.Air.Pollution.Outdoor.Total.per.100k, data = final_df_sf)

Residuals:
      Min        1Q    Median        3Q       Max
-18198393  -7914162  -2888919   3421808  96920299

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               26402616     457591    57.7   <2e-16
Deaths.Air.Pollution.Indoor.per.100k       -144486       3190   -45.3   <2e-16
Deaths.Air.Pollution.Outdoor.Total.per.100k  -242550       9394   -25.8   <2e-16

(Intercept)                               ***
Deaths.Air.Pollution.Indoor.per.100k      ***
Deaths.Air.Pollution.Outdoor.Total.per.100k ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13100000 on 5194 degrees of freedom
Multiple R-squared:  0.309,     Adjusted R-squared:  0.309
F-statistic: 1.16e+03 on 2 and 5194 DF,  p-value: <2e-16
```



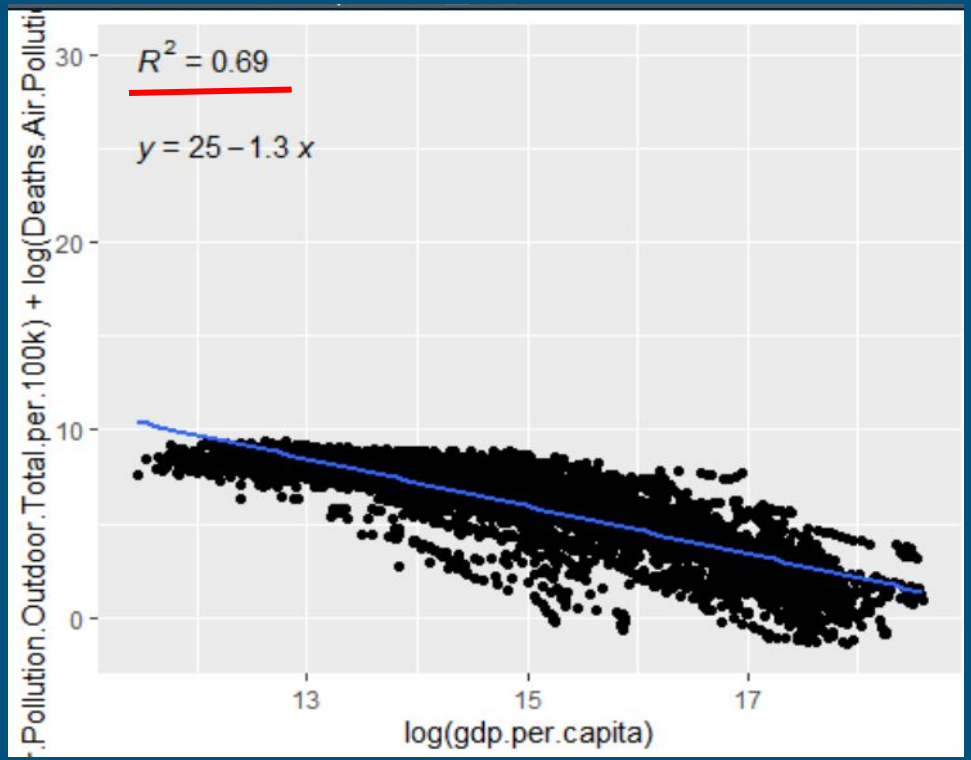
Indoor and outdoor air pollution (log)

```
Call:
lm(formula = log(gdp.per.capita) ~ log(Deaths.Air.Pollution.Outdoor.Total.per.100k)
    +
    log(Deaths.Air.Pollution.Indoor.per.100k), data = final_df_sf)

Residuals:
   Min     1Q Median     3Q    Max
-3.200 -0.594  0.056  0.621  2.901

Coefficients:
                                            Estimate Std. Error t value
(Intercept)                                 17.60602    0.08820   199.6
log(Deaths.Air.Pollution.Outdoor.Total.per.100k) -0.35985    0.02444   -14.7
log(Deaths.Air.Pollution.Indoor.per.100k)   -0.55212    0.00507  -108.8
                                            Pr(>|t|)
(Intercept)                                   <2e-16 ***
log(Deaths.Air.Pollution.Outdoor.Total.per.100k)   <2e-16 ***
log(Deaths.Air.Pollution.Indoor.per.100k)     <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.873 on 5194 degrees of freedom
Multiple R-squared:  0.696,     Adjusted R-squared:  0.696
F-statistic: 5.96e+03 on 2 and 5194 DF,  p-value: <2e-16
```



Conclusion

1) For the Logit Regression, we observed small p-values indicating that all the coefficients are found to be significant. GDP has a positive effect on population, but deaths due to air pollution have a negative effect on population.
2) As shown above, post-pruning, our accuracy is still 74%. So, pruning had little effect although the tree has become simpler. Therefore, we have a model that can predict high deaths due to air pollution with reasonable (74%) accuracy.
3) For clustering, observed that clusters have large discrepancies in regional representation while the data used did not have any explicit geospatial components.
4) Although death caused by indoor air pollution and outdoor air pollution seems to have a strong relationship with total death caused by air pollution, but we can't accurately predict GDP through death caused by indoor and outdoor air pollution.

Thank You!