

# Projeto Data Mining I

*Lucas Parada, Ana Catarina Monteiro, Lucas de Paula*

*11/16/2019*

---

## Introdução

O presente relatório passa pela previsão no nível de poluição do ar, AQI, no município de Pequim, província de Hebei, China, tendo como base um conjunto de dados que caracterizam a qualidade do ar na cidade. Os respetivos dados são recolhidos de 12 locais de monitorização do ambiente, em diferentes localidades, e posteriormente comparados com a estação meteorológica mais próxima, a Administração meteorológica da China. Os mesmos, são medidos desde primeiro de março de 2013 a 28 de fevereiro de 2017 e descrevem temporalmente alguns fatores climáticos (temperatura, pressão, *dew point*, precipitação, direção e velocidade do vento), assim como a concentração de poluentes relevantes para este estudo. Numa primeira fase foi realizada a importação dos dados, assim como a sua limpeza e pré-processamento. Posto isto, procedeu-se à análise dos mesmos, visando encontrar relações entre as diferentes variáveis e de que modo poderiam influenciar a previsão pretendida. Por fim, foram aplicados modelos distintos de previsão que intentaram prever o índice de qualidade do ar segundo os valores dos fatores climáticos e os AQI's anteriormente analisados.

## Definição do problema

O índice de qualidade do ar (AQI) versa o quanto poluído o mesmo se encontra ou o quanto pode vir a estar e consequentemente de que modo influencia o risco de saúde, que tende a aumentar à medida que o AQI aumenta. De acordo com a Organização Mundial de Saúde (OMS), os níveis de poluição em Pequim ultrapassaram níveis considerados perigosos [1] colocando em risco a qualidade de vida da população, e ocasionando prejuízos no ambiente a longo prazo. Por conseguinte, desde janeiro de 2013, [2] a China continental decidiu adotar o AQI para medir os seus níveis de poluição baseados na concentração de seis poluentes atmosféricos, nomeadamente partículas em suspensão menores que 10 micrometro de diâmetro aerodinâmico (PM 10 ) e menores que 2.5 micrometro (PM2.5), dióxido de enxofre (SO<sub>2</sub>), dióxido de azoto (NO<sub>2</sub>), monóxido de carbono (CO) e ozono (O<sub>3</sub>). Este índice é calculado por dia e varia de acordo com seis categorias. Inicialmente é atribuída uma pontuação singular a cada poluente e o valor final no AQI corresponde à mais alta entre as calculadas.

## Pré-processamento dos dados

Inicialmente realizámos a importação dos dados para o formato que considerámos mais apropriado visando facilitar a sua limpeza, pré-processamento e posterior análise.

## Importação e análise dos dados

Numa primeira fase, escolhemos analisar o dataset relativo à estação de Aotizhongxin e elegemos o DataFrame do R para manipular os dados uma vez que este formato facilita a manipulação de várias variáveis de classes diferentes e possui diversas ferramentas que auxiliam este processo. Posto isto, recorrendo à função 'summary'[Anexo 1.1], conseguimos observar algumas informações genéricas relativas aos nossos dados, nomeadamente dados estatísticos e a existência de *missing values*.

Considerando valores referentes a quatro anos completos, e tendo em conta que os valores são obtidos de hora em hora  $(4 * 365 + 1) * 24 = 35064$ , valor este igual ao número de linhas totais), concluímos que não existia nenhuma linha no dataset, tendo em conta, ainda, a ausência de linhas duplicadas.

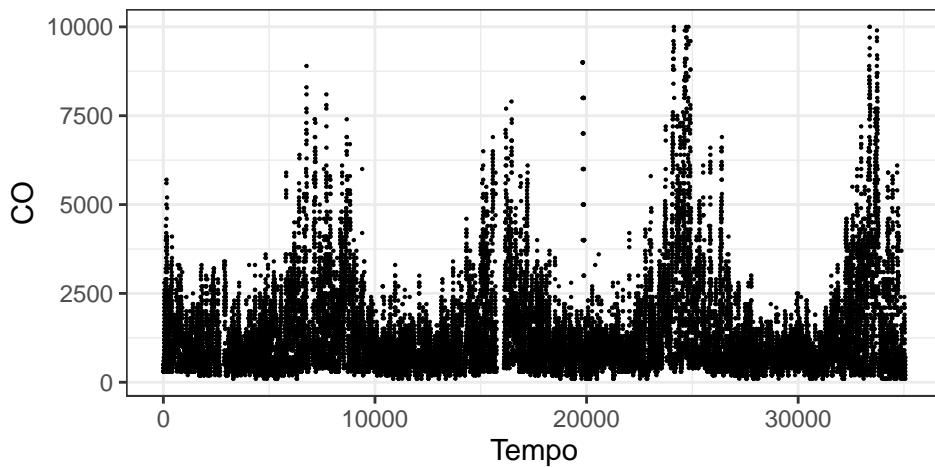
## Data-Cleaning

Passámos, portanto, ao tratamento de *missing values*, considerando primeiro a existência de *outliers*, uma vez que os mesmos poderiam influenciar o preenchimento dos valores referidos anteriormente. Sendo os *outliers* dados com características consideravelmente diferentes da maioria dos outros objetos, isto é, ruído que pode

interferir na análise dos dados, decidimos identificá-los para proceder ao seu tratamento. Para esse efeito, observámos o comportamento das variáveis numéricas, recorrendo a gráficos como o boxplot, comparando três escalas temporais diferentes: todo o dataset, por estação do ano e por mês. Ao verificar a sua distribuição por todo o dataset percebemos que os valores apresentavam padrões correspondentes a quatro estações (figura 1), ainda assim, concluímos que ao visualizar os dados por mês a variação entre os mesmos não era tão significativa, identificando melhor os *outliers*.

**Figura 1**

Variação do CO ao longo do dataset



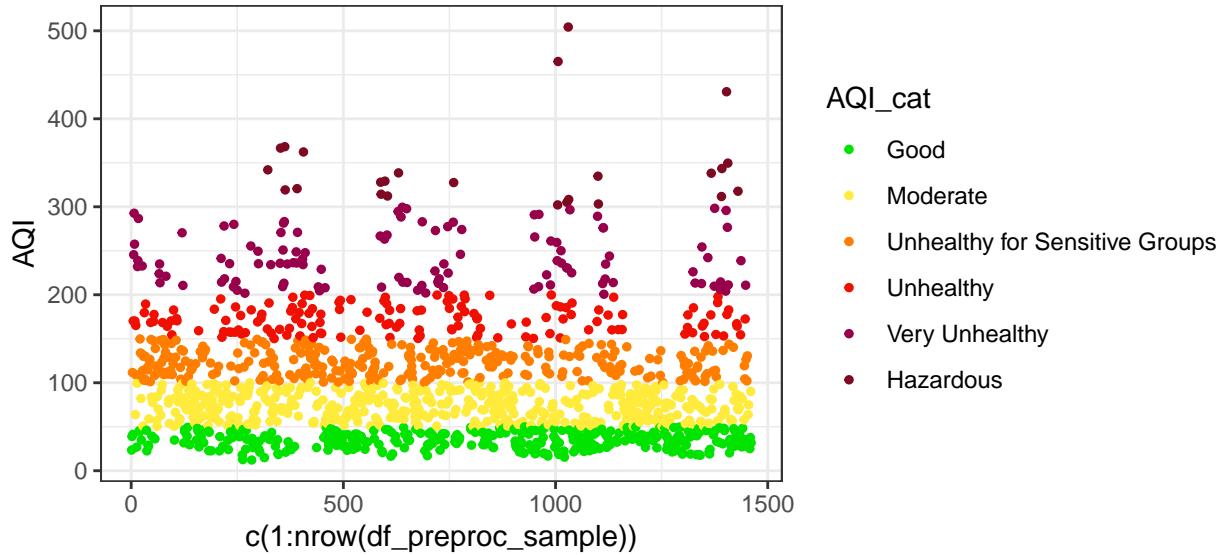
Depois de identificados estes casos, optámos por tratá-los de modos diferentes consoante o seu significado real. Isto é, no caso da variável precipitação (RAIN), os *outliers* observados tratavam-se de casos sensíveis, ou seja, depois de algum tempo na ausência de precipitação qualquer alteração aos valores da mesma foram entendidos como *outliers* e por esse motivo decidimos não alterar os valores desta variável. No que diz respeito aos restantes, alterámos a temperatura (TEMP) para kelvin[4] para trabalhar com valores positivos, e substituímo-la, assim como as restantes variáveis, por 'NA' (valores em falta) em caso de haverem *outliers*. Foram Experimentados dois métodos diferentes para o tratamento dos *outliers*: se estes fossem superiores a quatro vezes a média[3] para o respetivo ano e mês. ou o método que a técnica do boxplot aplica para a eliminação de outliers[5]  $Q1 - 1.5 * IQR < x < Q3 + 1.5 * IQR$ . No final destas operações, utilizámos um gráfico de 'geom\_points' para analisar o comportamento das variáveis após o tratamento dos *outliers* e concluímos que o método do boxplot eliminava dados em demasia pelo que optamos por tratar esses valores recorrendo à primeira estratégia. Ulteriormente a estas operações, visando o tratamento de *missing values*, observámos graficamente de que modo as diferentes variáveis se relacionavam, e encontrámos explicitamente uma relação positiva entre a temperatura e o *dew point* e uma negativa entre a pressão atmosférica (PRES) e o *dew point*, tal como podemos observar nos gráficos do anexo xxxxxxxx

Para preencher os valores que ainda se encontravam em falta, no caso da chuva optámos por substituir pela mediana, quanto à variável categórica, escolhemos o valor correspondente à hora anterior e relativamente às restantes elegemos a interpolação linear como a melhor escolha, uma vez que se tratam de valores temporais com alguma tendência, procurando deste modo aproximar os *missing values* o mais possível aos valores reais.

## Data Pre-Processing

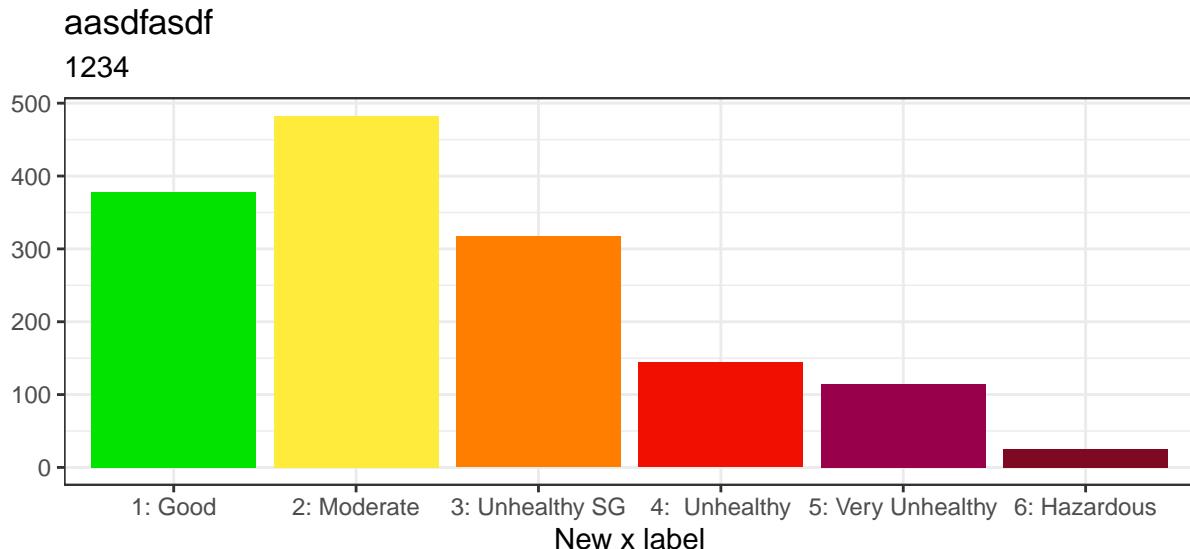
No que concerne ao pré-processamento dos dados, começámos por corrigir as unidades de medida de alguns poluentes [6] de acordo com o pedido para calcular o AQI[7], e seguidamente, calculámos a média diária de todas as variáveis numéricas e a moda da categórica (direção do vento (WD)). Posto isto, visto que cada linha do dataset corresponde a uma hora de cada dia, removemos as colunas relativas à hora e ao número de linhas, e eliminámos as linhas duplicadas. Decidimos ainda calcular as estações do ano numérica e categoricamente, aplicámos one-hot-encoding à variável categórica direção do vento, visando facilitar a manipulação dos dados, e ainda criámos uma nova variável relativa à precipitação, transformando-a noutra, categórica, para mais

tarde auxiliar a previsão pretendida. Procedemos então ao cálculo do AQI para cada variável [7], em seguida, para cada dia, conseguindo finalmente obter uma visão acerca do comportamento do índice de poluição calculado para todo o dataset.

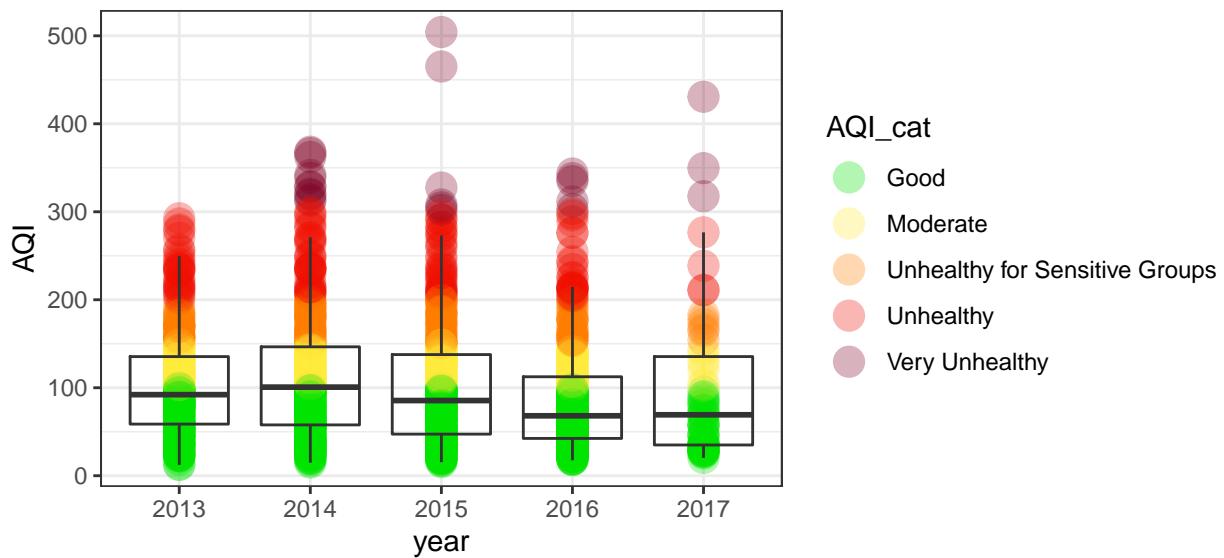


Com base na análise realizada, futuramente apresentada, e tendo em conta os gráficos que estudámos, constatámos uma soberania do poluente PM2.5, o que tornava quase impossível aferir relações e conclusões com as restantes variáveis. Por este motivo, tornou-se imperativo para nós desconsiderar os valores do poluente nos nossos cálculos, ainda que tenhamos entendido o porquê da China considerar o mesmo como o maior e mais alarmante poluente na cidade. Posto isto, para efeitos do cálculo do índice de poluição, tomámos como referência a tabela que calcula o AQI nos Estados Unidos, adotando os seus intervalos e unidades de medida como modelo.[7] ##### Análise dos dados Fizemos uma análise os dados, agora já limpos, e procurámos estabelecer relações que se mostrassem relevantes, tanto para extrair informações úteis, como para a previsão pretendida.

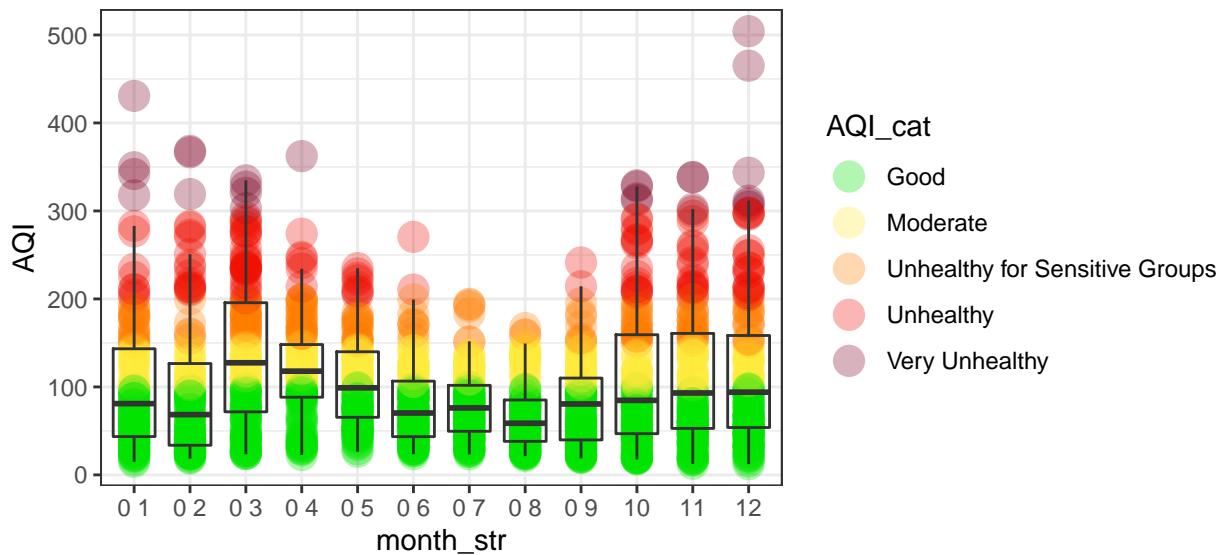
Começámos por relacionar o AQI ao longo de diferentes frações de tempo, comparando graficamente os seus valores, numéricos e categóricos, e averiguando de que modo conseguiríamos lograr mais conclusões. Inicialmente analisámos o comportamento do AQI face a todo o dataset, através de um gráfico de barras, e concluímos que a grande maioria dos dias definia o seu índice de poluição como ‘Moderate’ e, ‘Hazardous’ verificou-se, comparativamente aos restantes, como o mais baixo.

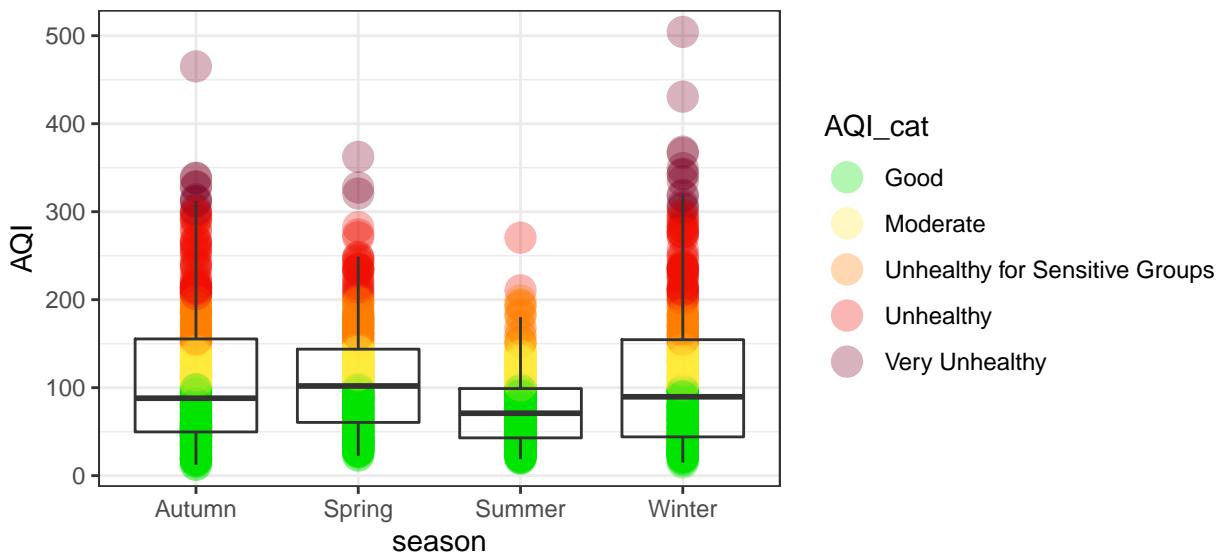


Comparando, depois, o mesmo ao longo dos anos que o dataset inclui, demo-nos conta, (atendendo a um gráfico de ‘geom\_point’), que 2017 apresentava os valores mais baixos de poluição, mas uma vez que somente conhecemos valores referentes a dois meses desse ano, desconsiderámo-lo da nossa análise, desse modo, é perfeitável que desde 2014 os níveis de poluição diminuíram nos anos consecutivos. Correspondentemente, esta conclusão veio corroborar algumas pesquisas, que sugerem a adoção de políticas contra a poluição desde 2014, “quando no dia 4 de março o primeiro-ministro Li Keqiang anunciou uma mudança nos rumos do país:”Vamos declarar guerra à poluição assim como declaramos guerra à pobreza“” [8]

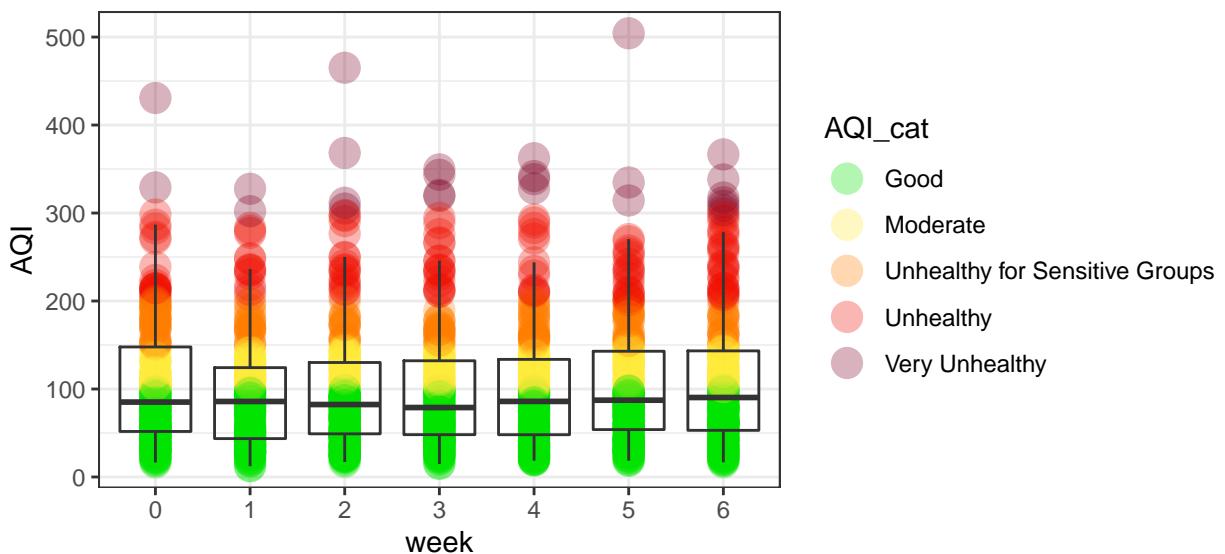


Continuamente, comparámos o AQI face ao mês e à estação do ano, os quais mostraram um aumento de poluição nos meses/estações mais frias e uma redução nos mais quentes, nomeadamente no mês de agosto, que não ultrapassou o índice de 150(‘Moderate’).



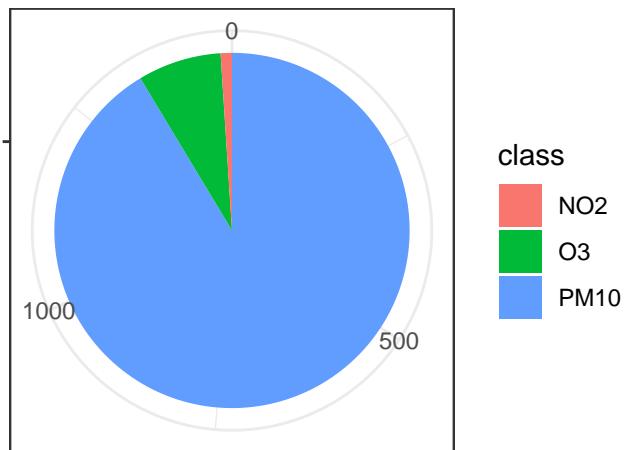


De seguida, realizámos a análise comparativamente aos dias da semana, na expectativa de encontrar níveis mais altos ou mais baixos durante determinados dias. No entanto, observámos que as alterações do AQI não eram significativas, descartando por isso, a hipótese deste fator influenciar o modelo de previsão.



Posto isto, procurámos saber, recorrendo ao gráfico ‘pie chart’ quais os poluentes que mais influenciavam o AQI, e, de acordo com o mesmo, verificámos que, dos cinco poluentes considerados, os que mais caracterizavam o índice de poluição, de modo crescente, eram: NO<sub>2</sub>, O<sub>3</sub> e em maior quantidade, PM10. De modo consequente, observámos o comportamento deste último poluente, por hora, considerando a possibilidade da sua concentração ser superior em algum momento do dia, no entanto, o mesmo apresentou diferenças irrelevantes, entendendo que esta análise não influenciaria os modelos de previsão, uma vez que o AQI é calculado por dia [Anexo xxxxx].

Pie Chart of class

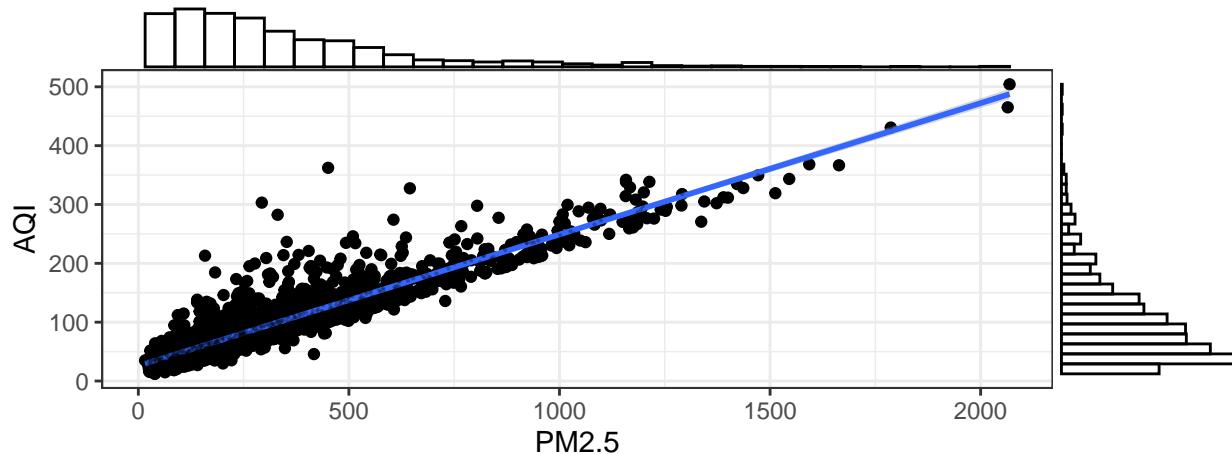


Source: mpg

Depois de realizada esta comparação, analisámos, de novo recorrendo a um gráfico ‘seom\_point’ e a um ‘histograma’, a possível correlação entre os vários poluentes e o AQI, e demo-nos conta que existe uma clara relação linear entre os poluentes PM10 e PM2.5 tal como podemos observar na figura 5 enquanto que os restantes não mostraram uma relação tão forte[Anexo xxxxx].

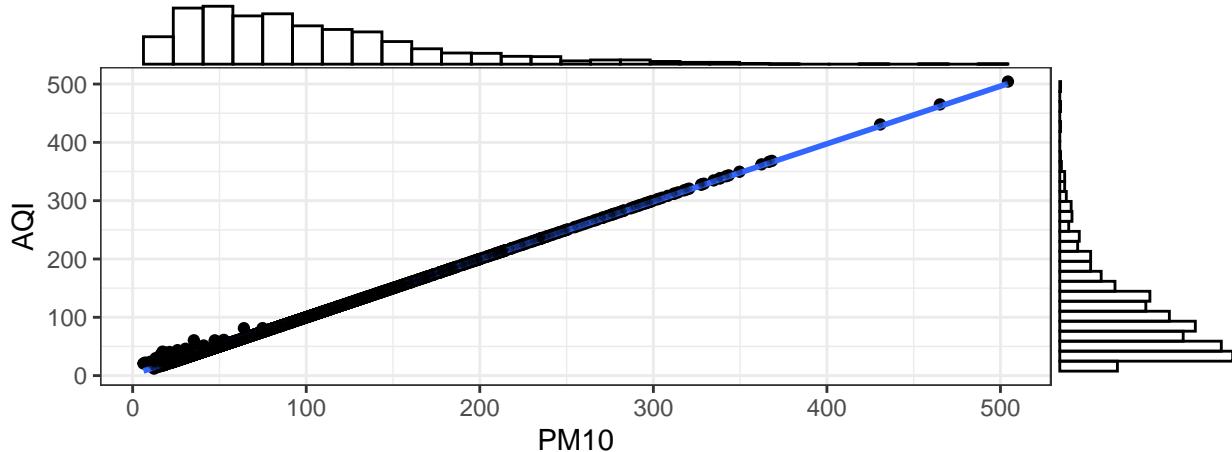
## PM2.5

### Relação entre PM2.5 e AQI



## PM10

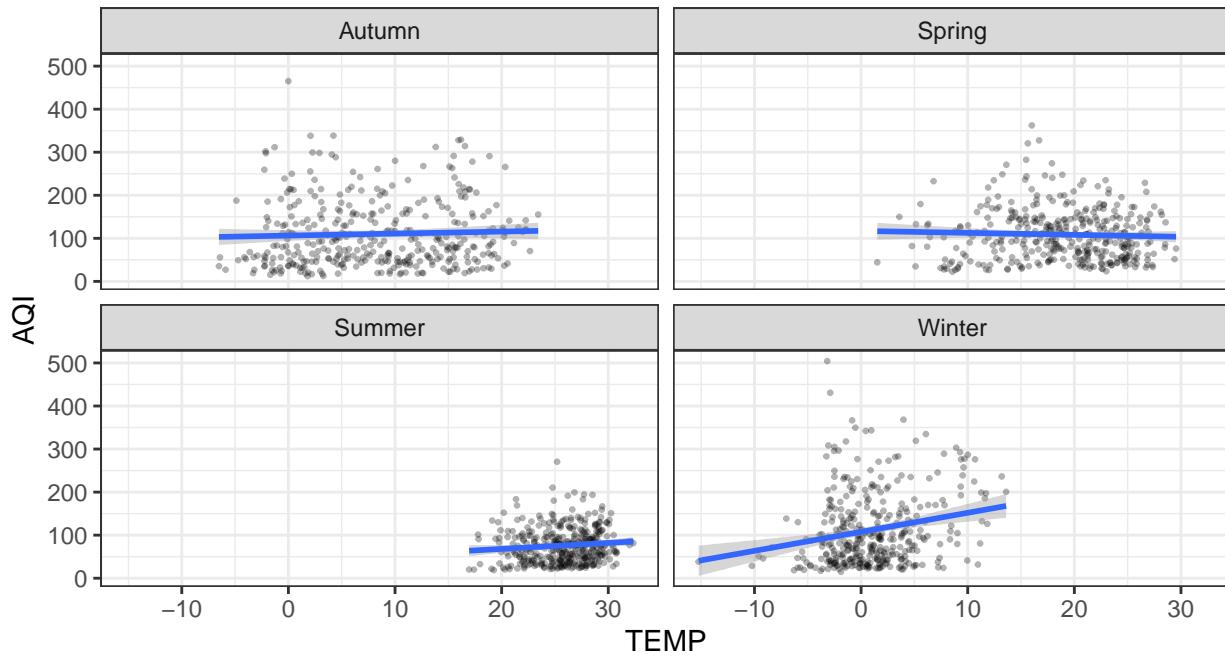
### Relação entre PM10 e AQI



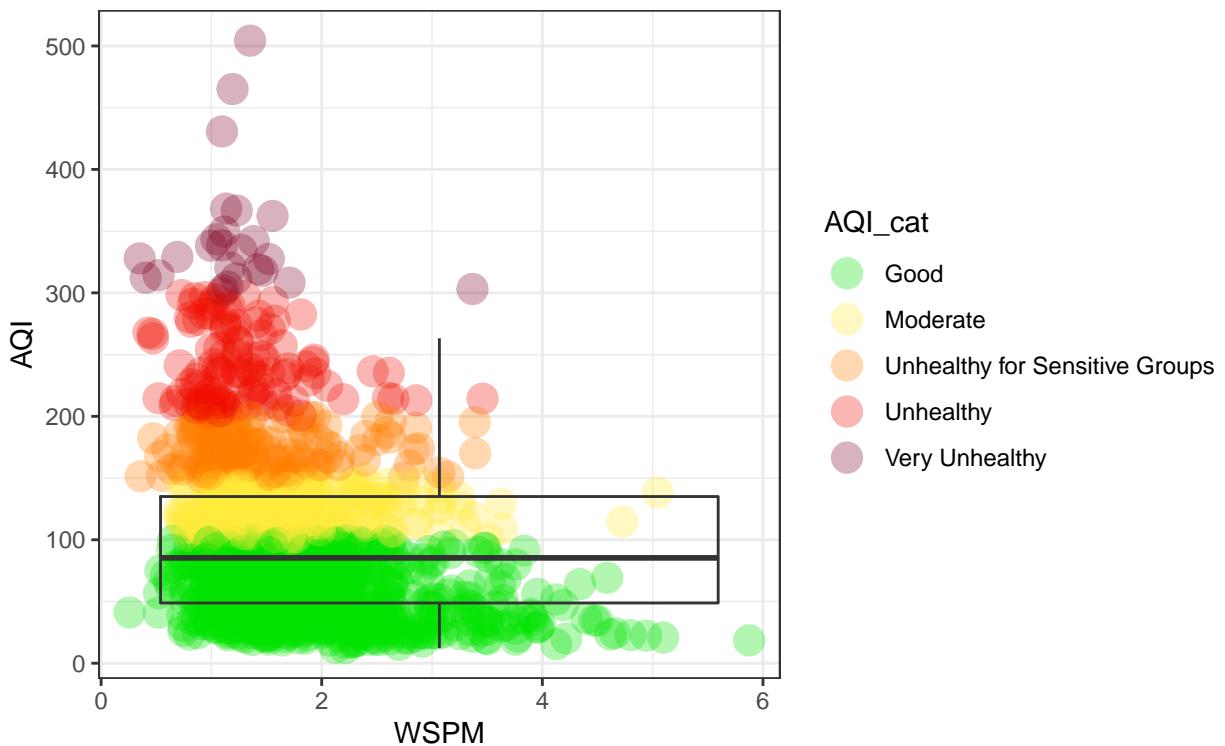
No que respeita aos fatores climáticos, realizámos ainda algumas comparações com o AQI que se mostraram muito relevantes. Comparado à fatores como a temperatura, precipitação, pressão e *dew point* percebemos que as mesmas dificilmente se relacionavam. Ainda assim, tentámos visualizar estes valores ao longo das quatro estações do ano, mas os resultados anteriores prevaleceram: a relação entre estas variáveis e o AQI era muito pobre, como demonstradas na imagem a seguir e no Anexo xxxxx.

## TEMP

### Relação entre TEMP e AQI

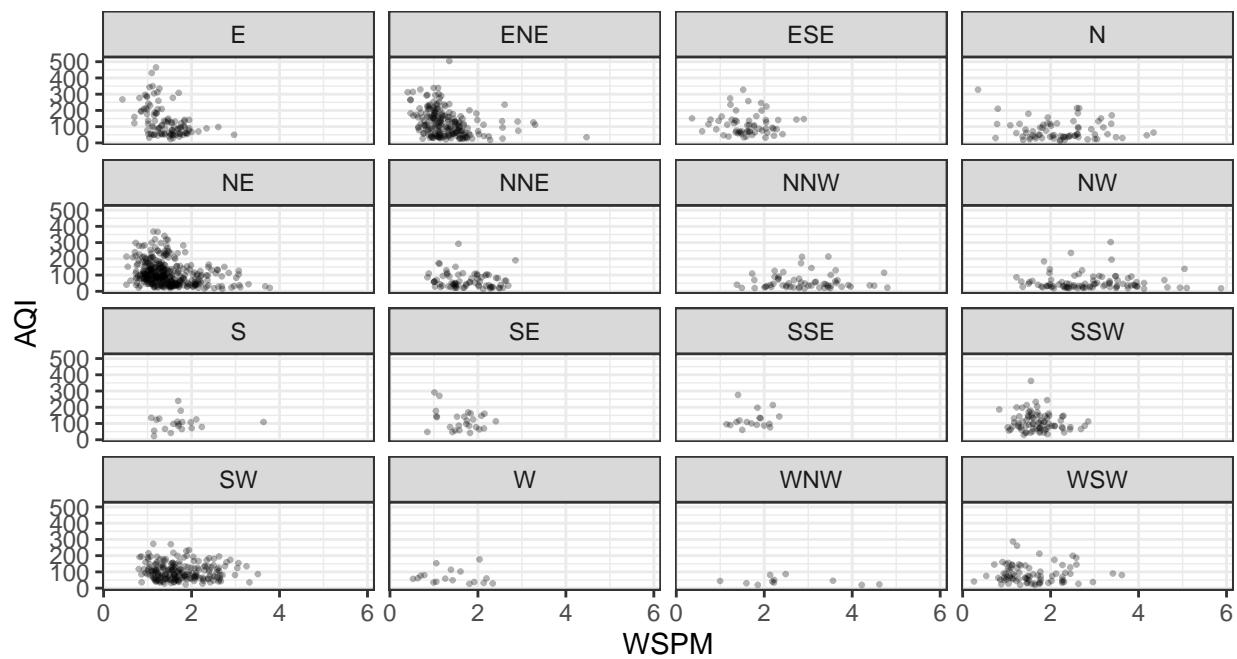


Por outro lado, quando comparado com a velocidade do vento (WSPM) constatámos que o AQI tendia a diminuir à medida que a velocidade do vento aumentava, o que se justifica, uma vez que ventos fortes podem ajudar a limpar o ar em tempos muito curtos [9]. Seguidamente, tornou-se imperativo comparar o AQI com a velocidade do vento tendo em conta as diferentes direções do mesmo, e identificámos a predominância em algumas direções, conforme se comprova com a imagem 6. O que nos levou a crer que este fator climático era o que mais influenciaria o nosso modelo de previsão.



## WSPM

### Relação entre WSPM e AQI



### Modelos de previsão

Por fim concluídas as fases de limpeza e pré-processamento de dados, e levando em consideração os resultados das análises que efetuámos, podemos por fim refletir e decidir que modelos de previsão iríamos utilizar nesta etapa. Ambicionando prever valores o mais próximo possível dos reais, testámos, no total, quatro modelos preditivos, comparando mais tarde os seus resultados. Desses quatro, decidimos adotar dois modelos diferentes de regressão, e outros dois de classificação, os quais nos permitiram aferir algumas conclusões entre o mesmo

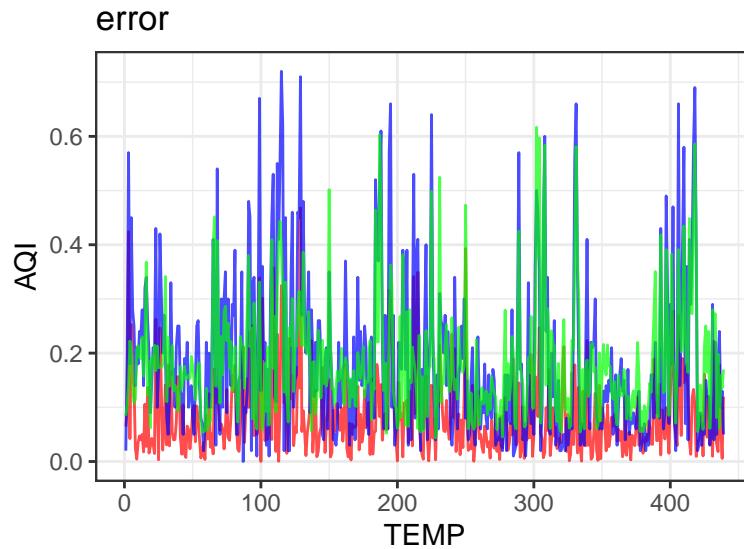
tipo de previsão, e entre diferentes tipos, para prever a mesma variável, neste caso, o AQI. Ainda antes de aplicar qualquer um dos modelos referidos, procedemos à normalização dos dados, e arredondámos depois as variáveis numéricas para duas casas decimais, tendo em vista a obtenção de melhores resultados. Posto isto, dividimos o dataset em treino (70%) e teste (30%), com `set.seed(123)` para que os dados estudados se mantivessem constantes.

Ao executar os modelos de previsão testamos diferentes combinações de parâmetros, para que desta forma pudesse minimizar o erro e aumentar a eficiência do modelo. Assim sendo, no que diz respeito aos modelos de regressão, escolhemos testar, em primeiro lugar, uma rede neuronal.

### **Rede Neuronal.**

Para a criação do modelo foi utilizada a biblioteca ‘neuralnet’, e os parâmetros: direção do vento (fator atmosférico que mais influencia o índice de poluição, como verificado na análise dos dados), a precipitação, velocidade do vento, estação do ano e o AQI do dia anterior. Após efetuar a previsão pretendida, obtivemos os valores relativos à *mean squared error* e à *mean absolute error*, respectivamente, 0.012 e 0.083. No final, utilizando um gráfico de linhas, analisamos a variação entre os valores de AQI reais (azul), os valores previstos (verde) e o erro entre ambos (vermelho).

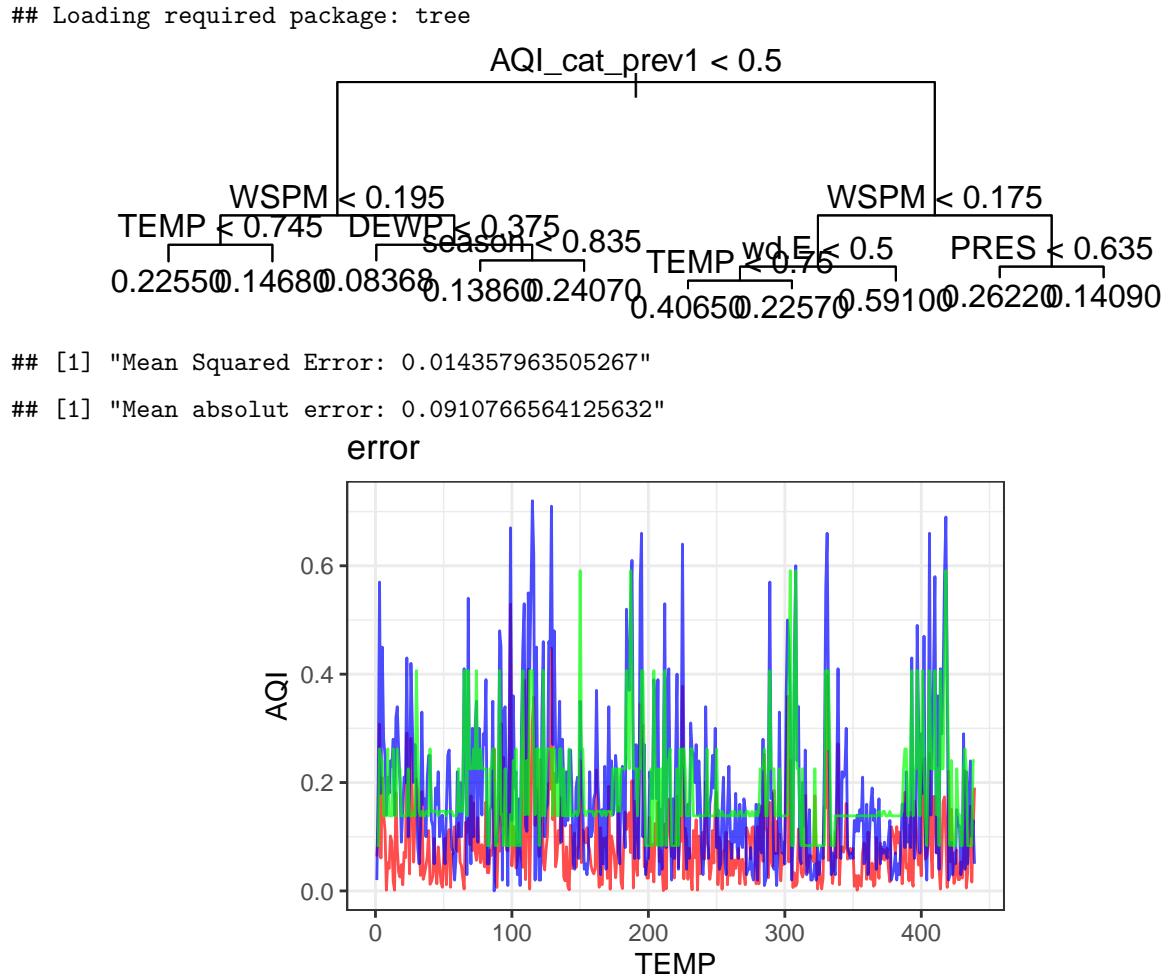
```
## Loading required package: neuralnet
##
## Attaching package: 'neuralnet'
## The following object is masked from 'package:dplyr':
##     compute
```



```
## [1] "Mean Squared Error: 0.0125605913166901"
## [1] "Mean absolute error: 0.0837780862067789"
```

### **Árvore de decisão.**

Ainda acerca de modelos regressivos, elegemos a ‘decision tree’ como segundo modelo. Para a mesma, recorremos à biblioteca ‘tree’ mas desta vez, quase todos os parâmetros se mostraram relevantes, influenciando de modo positivo o resultado final. Ao calcular os valores relativos à *mean squared error* e à *mean absolute error*, obtivemos 0.143, 0.091 na devida ordem. No final, assim como no modelo anterior, utilizando um gráfico de linhas para analisar a variação entre os valores de AQI reais (azul), os valores previstos (verde) e o erro entre ambos (vermelho).



Não obstante, os modelos de regressão não se mostraram excelentes a solucionar o problema inicialmente descrito, pelo que determinamos ser a altura de testar modelos de classificação, procurando resultados diferentes e capazes de perceber qual o melhor método a aplicar. Procedemos assim à criação de uma ‘random forest’, empregando a biblioteca ‘ranger’, e, depois de alguns testes, identificámos os parâmetros que, combinados, apresentam os melhores resultados, sendo estes a direção e velocidade do vento, o *dew point*, o AQI do dia anterior e a precipitação. Relativamente a este último parâmetro, notámos, na secção de análise de dados, que a sua relação com o AQI era muito fraca mas a sua ausência também causava alterações no resultado final. Por esse motivo, substituímos a variável precipitação por uma variável categórica (RAIN\_cat) que se mostrou mais adequada ao modelo, e veio corroborar as nossas deduções, ao mostrar consideráveis melhorias na ‘accuracy’. Calculámos depois a matriz de confusão, com base na classificação categórica do AQI, e ainda o valor da ‘accuracy’ que nos indica o quão próximos nos encontramos dos valores reais, à qual obtivemos uma aproximação de 48,29%.

```

##
##      0   1   2   3   4   5
##  0  66  52   3   1   1   0
##  1  19 101  13   4   1   0
##  2   6  61  13   4   1   0
##  3   3  24   2   5   4   1
##  4   1   8   3   5  26   0
##  5   0   1   1   0   8   1

```

```

## [1] "Accuracy: 48.2915717539863%"

```

Não suficientemente satisfeitos com estes resultados, implementámos o modelo de SMV, também de classificação, desta vez com a biblioteca ‘e1071’ e, pelos mesmos motivos, aplicámos os parâmetros escolhidos para a ‘random forest’. Contudo, neste caso decidimos tratar a variável de classe de modo binário, agrupando desde ‘Good’ até ‘Unhealthy for Sensitive Groups’ (0-150) e outra com as restantes (>150). No final deste modelo os valores de ‘accuracy’ correspondiam a 86,3% o que representa uma visível melhoria de quase 50% face à anterior e provando assim, ser este o modelo que melhor conseguiu prever, ainda que categoricamente, o valor do índice de poluição, (AQI). Para efeitos de visualização do comportamento do SVM podemos encontrar os gráficos no anexo XXX.

## SVM

```
## 
## Attaching package: 'e1071'
## The following object is masked from 'package:mltools':
##   skewness
## The following object is masked from 'package:tidyimpute':
##   impute
##   y_pred
##   0   1
##   0 328 18
##   1  42 51
## [1] "Accuracy: 86.33%"
```

## Conclusão

De acordo com a pesquisa efetuada e tendo como base a análise retirada neste trabalho, uma coisa é clara, a China ainda é um país com elevados índices de poluição, apesar deste relatório resida apenas na interpretação dos valores obtidos pela por uma estação meteorológica específica. Ainda assim confirmando algumas das nossas pesquisas, conseguimos verificar que, em média, o AQI tende a diminuir depois de 2014. Conforme sugerem as nossas análises, os fatores climáticos, tidos em conta neste dataset, apresentaram relações muito fracas quando comparados com o AQI, todavia, aquando dos modelos de previsão,

## Referências Bibliográficas

- [1] [https://www.bbc.com/portuguese/noticias/2013/01/130114\\_poluicao\\_china\\_mm](https://www.bbc.com/portuguese/noticias/2013/01/130114_poluicao_china_mm)
- [2] [https://en.wikipedia.org/wiki/Air\\_quality\\_index](https://en.wikipedia.org/wiki/Air_quality_index)
- [3] <https://www.r-bloggers.com/outlier-detection-and-treatment-with-r/>
- [4] <https://www.metric-conversions.org/pt/temperatura/celsius-em-kelvin.htm>
- [5] [https://www.dcc.fc.up.pt/~rpribeiro/aulas/DMI1920/material/DMI\\_3-DataExpl\\_1920.pdf](https://www.dcc.fc.up.pt/~rpribeiro/aulas/DMI1920/material/DMI_3-DataExpl_1920.pdf)
- [6] <http://www.apis.ac.uk/unit-conversion?fbclid=IwAR1ffhFbcOA9-tPqzGMTcueGweOeNhYauUL1qDYn2C8twtdP83Gm-6DFC9M>
- [7] [https://en.wikipedia.org/wiki/Air\\_quality\\_index](https://en.wikipedia.org/wiki/Air_quality_index)
- [8] <https://revistagalileu.globo.com/Ciencia/noticia/2018/03/guerra-contra-poluicao-comeca-dar-frutos-na-china.html>
- [9] [http://aqicn.org/faq/2015-11-05/a-visual-study-of-wind-impact-on-pm25-concentration](http://aqicn.org/faq/2015-11-05-a-visual-study-of-wind-impact-on-pm25-concentration)

## Anexos

```
## [1] "Anexo 1"
##      No          year        month       day
## Min.   : 1   Min.   :2013   Min.   : 1.000   Min.   : 1.00
## 1st Qu.: 8767 1st Qu.:2014   1st Qu.: 4.000   1st Qu.: 8.00
## Median :17532 Median :2015   Median : 7.000   Median :16.00
## Mean   :17532 Mean   :2015   Mean   : 6.523   Mean   :15.73
## 3rd Qu.:26298 3rd Qu.:2016   3rd Qu.:10.000  3rd Qu.:23.00
## Max.   :35064 Max.   :2017   Max.   :12.000  Max.   :31.00
```

```

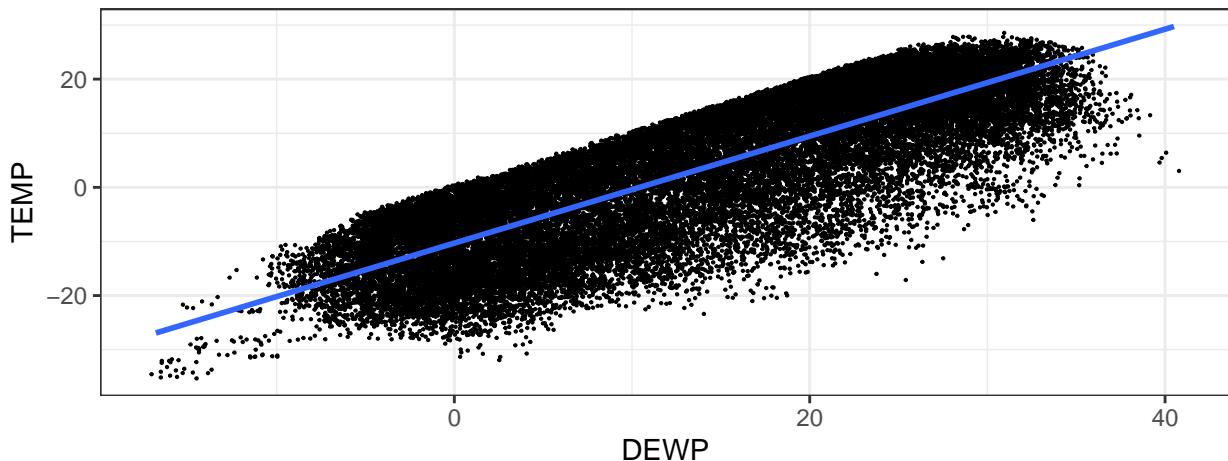
##          hour        PM2.5        PM10        SO2
##  Min.   : 0.00   Min.   : 3.00   Min.   : 2.0   Min.   : 0.2856
##  1st Qu.: 5.75   1st Qu.: 22.00  1st Qu.: 38.0  1st Qu.: 3.0000
##  Median :11.50   Median : 58.00  Median : 87.0  Median : 9.0000
##  Mean   :11.50   Mean   : 82.77  Mean   :110.1  Mean   : 17.3759
##  3rd Qu.:17.25   3rd Qu.:114.00 3rd Qu.:155.0 3rd Qu.: 21.0000
##  Max.   :23.00   Max.   :898.00  Max.   :984.0  Max.   :341.0000
##          NA's    :925       NA's    :718       NA's    :935
##          NO2        CO        O3        TEMP
##  Min.   : 2.00   Min.   :100     Min.   : 0.2142  Min.   :-16.80
##  1st Qu.:30.00   1st Qu.: 500    1st Qu.: 8.0000  1st Qu.: 3.10
##  Median :53.00   Median : 900    Median : 42.0000  Median : 14.50
##  Mean   :59.31   Mean   :1263    Mean   : 56.3534  Mean   : 13.58
##  3rd Qu.:82.00   3rd Qu.:1500    3rd Qu.: 82.0000 3rd Qu.: 23.30
##  Max.   :290.00  Max.   :10000   Max.   :423.0000  Max.   : 40.50
##  NA's   :1023    NA's   :1776    NA's   :1719      NA's   :20
##          PRES        DEWP        RAIN        wd
##  Min.   : 985.9  Min.   :-35.300  Min.   : 0.00000  NE   : 5140
##  1st Qu.:1003.3 1st Qu.: -8.100  1st Qu.: 0.00000  ENE  : 3950
##  Median :1011.4  Median : 3.800  Median : 0.00000  SW   : 3359
##  Mean   :1011.8  Mean   : 3.123  Mean   : 0.06742  E    : 2608
##  3rd Qu.:1020.1 3rd Qu.: 15.600 3rd Qu.: 0.00000  NNE  : 2445
##  Max.   :1042.0  Max.   : 28.500  Max.   :72.50000  (Other):17481
##  NA's   :20       NA's   :20      NA's   :20       NA's   : 81
##          WSPM        station
##  Min.   : 0.000  Aotizhongxin:35064
##  1st Qu.: 0.900
##  Median : 1.400
##  Mean   : 1.708
##  3rd Qu.: 2.200
##  Max.   :11.200
##  NA's   :14

## [1] "Anexo ANALISE POR HORA PM10"

```

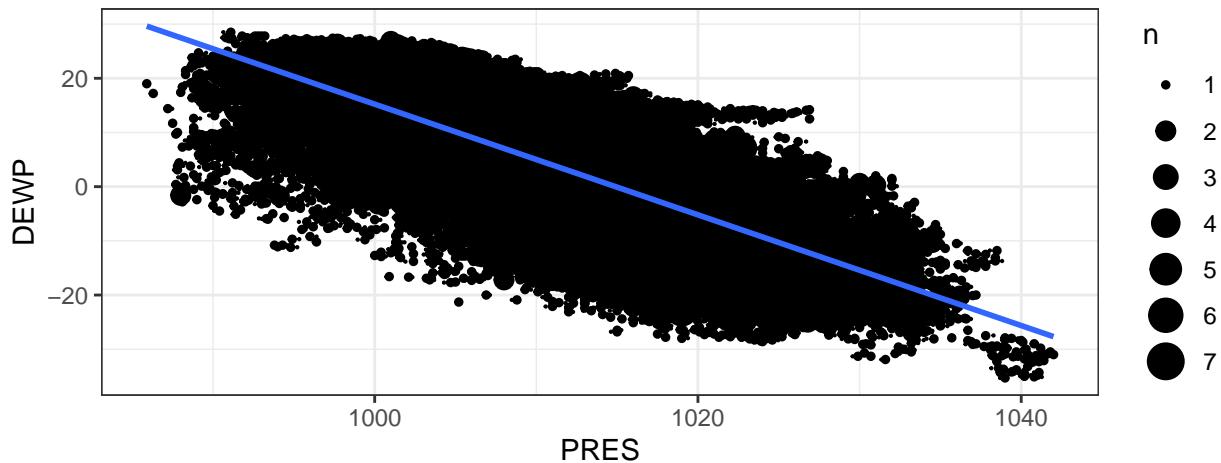
## TEMP

### Relação entre TEMP e DEWP

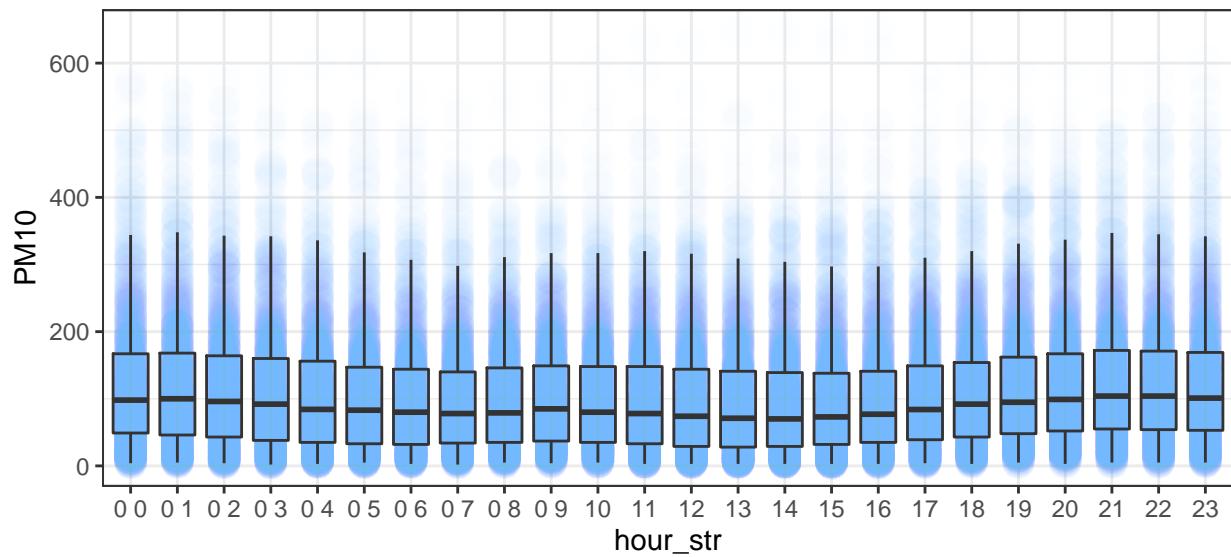


## PRES

### Relação entre PRES e DEWP



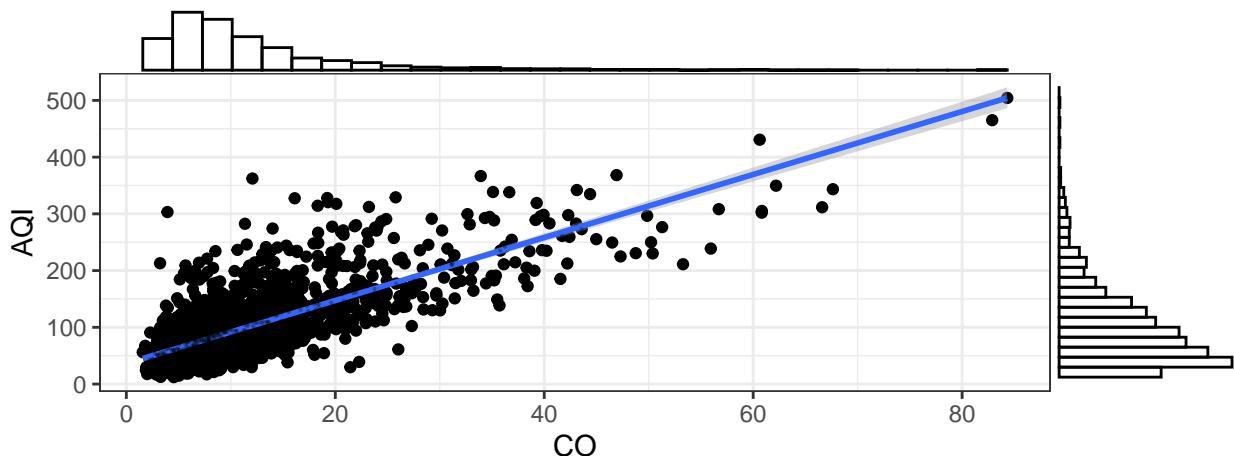
```
## [1] "Anexo 2 ANALISE POR HORA PM10"
```



```
## [1] "Anexo 3 RELAÇÃO AQI E POLUENTES"
```

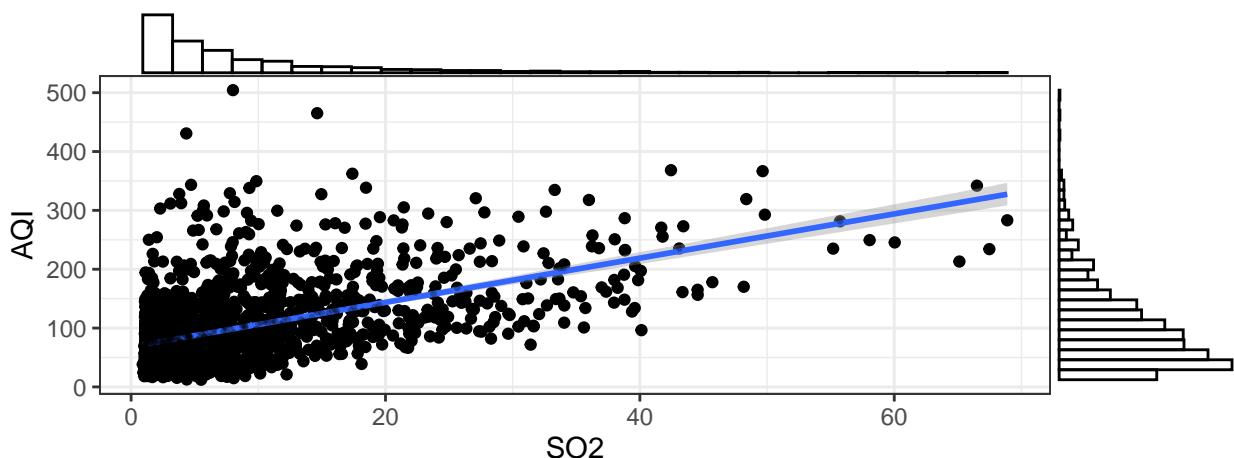
CO

Relação entre CO e AQI



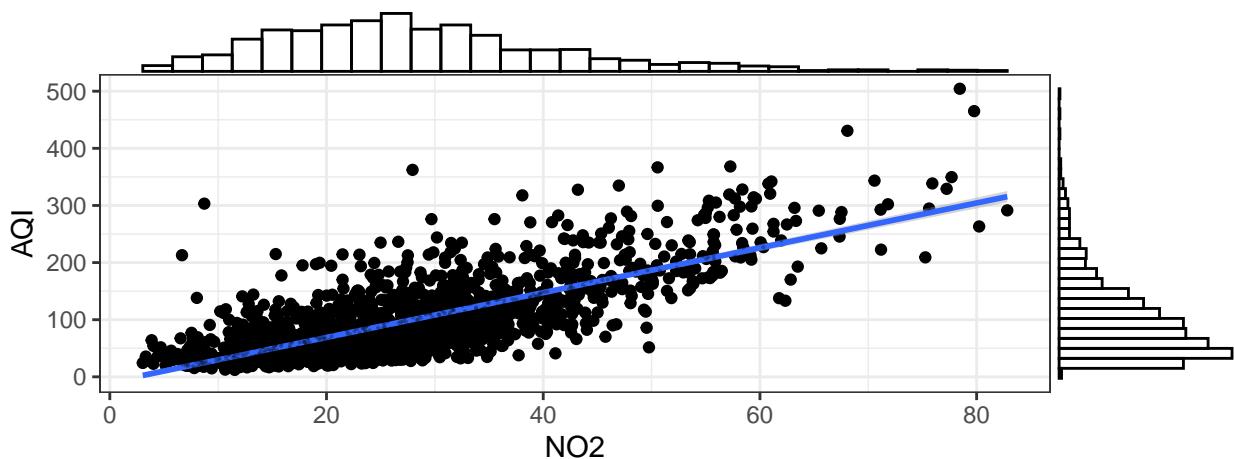
SO<sub>2</sub>

Relação entre SO<sub>2</sub> e AQI



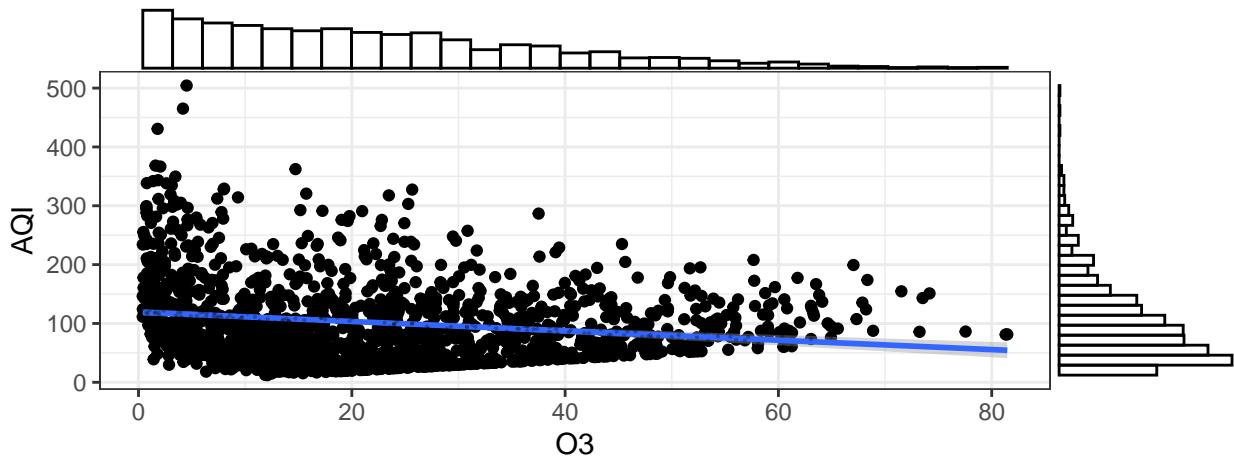
NO<sub>2</sub>

Relação entre NO<sub>2</sub> e AQI



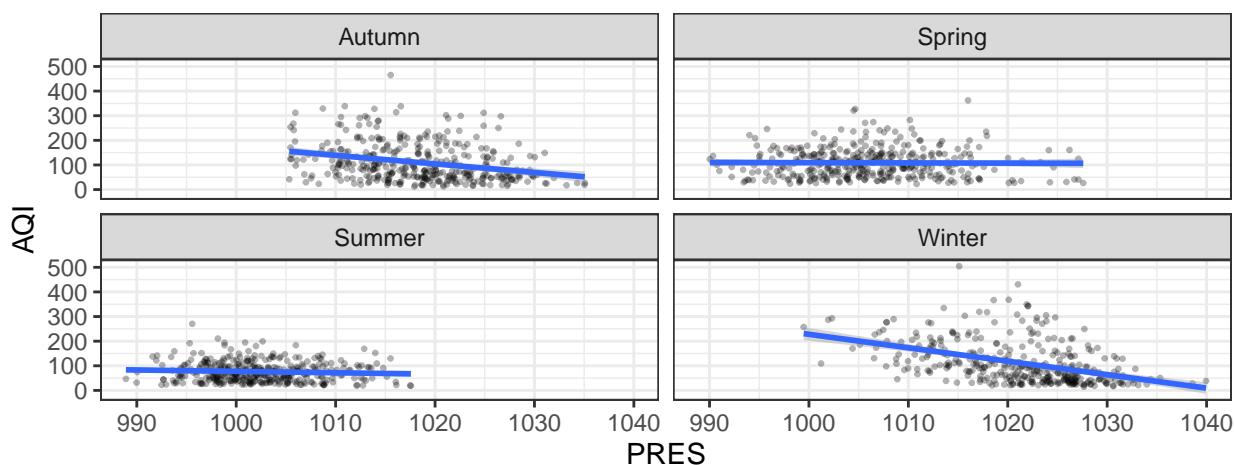
### O3

Relação entre O3 e AQI



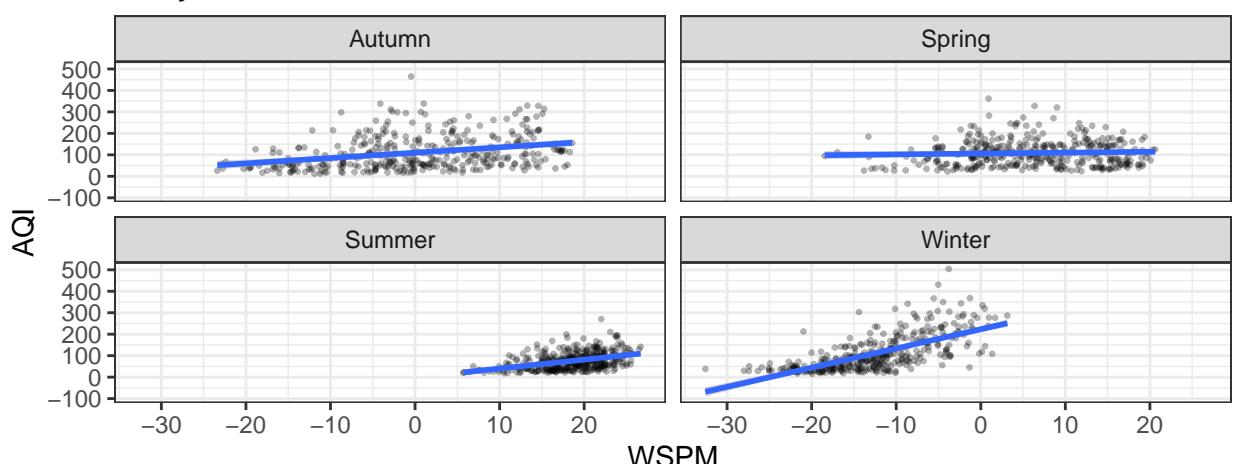
### PRES

Relação entre PRES e AQI



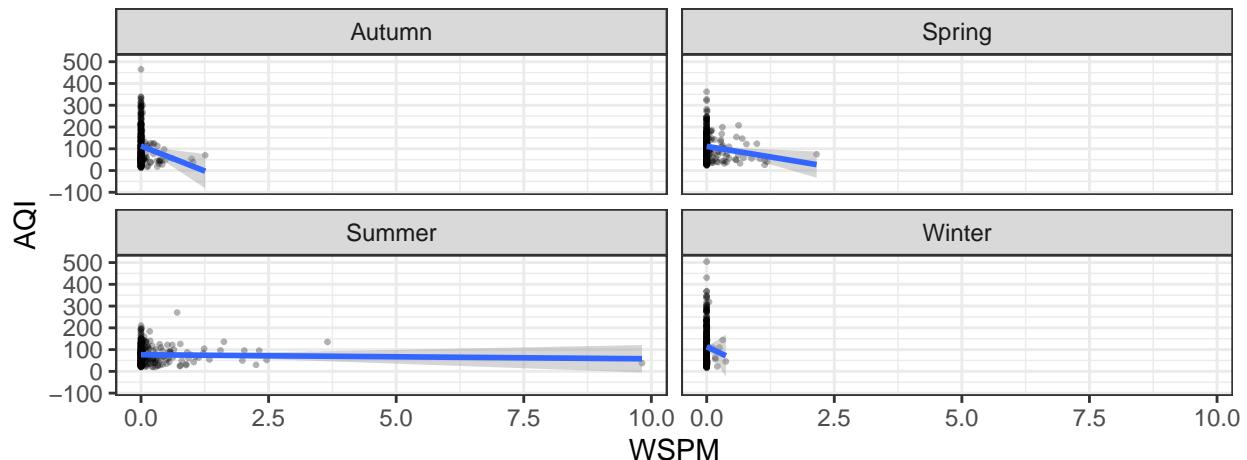
### DEWP

Relação entre DEWP e AQI



## DEWP

### Relação entre DEWP e AQI



Começamos por verificar se existia algum dia em falta no dataframe e vimos que não. Sabe-se que o dataset possui os valores referentes a 4 anos completos especificados por hora então devem existir  $(4365+1)24 = 35064$  rows

## Outliers

Seguimos com a análise da existência de outliers em variáveis numéricas. Começando por ver fazer a análise por variável como um todo, em seguida fizemos a análise por variável tendo em conta a estação do ano e por último por variável por mês.

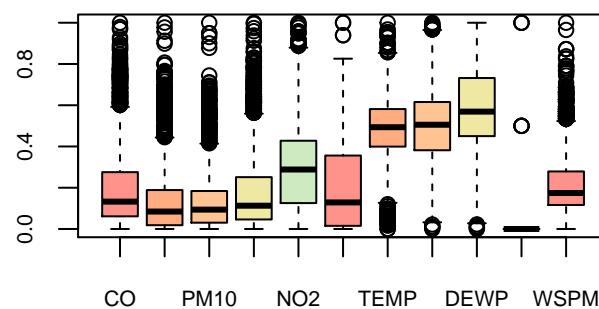
### Como um todo

### Por estação do ano

### Por mês

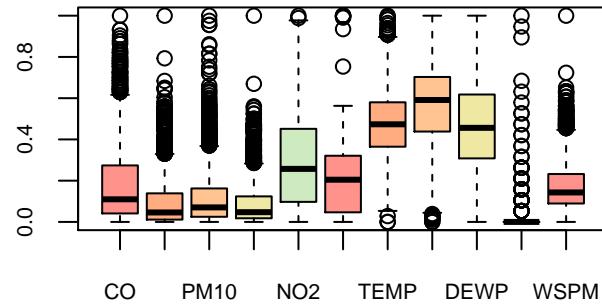
```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
## [1] ""
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

## January



```
## [1] ""
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

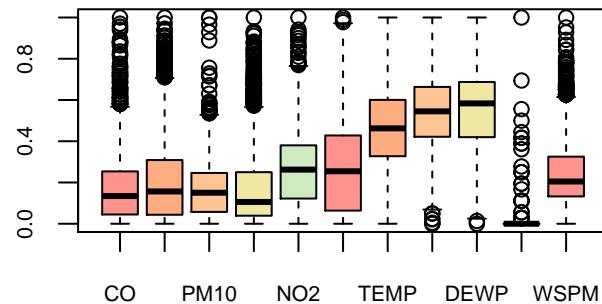
## February



```
## [1] ""
```

```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

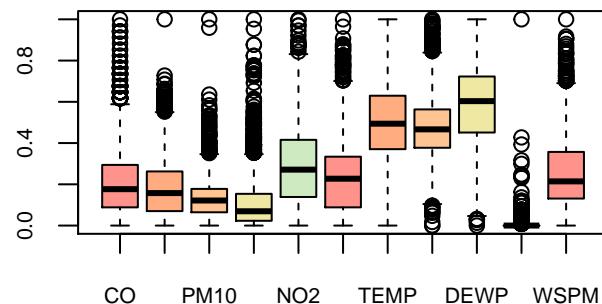
## March



```
## [1] ""
```

```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

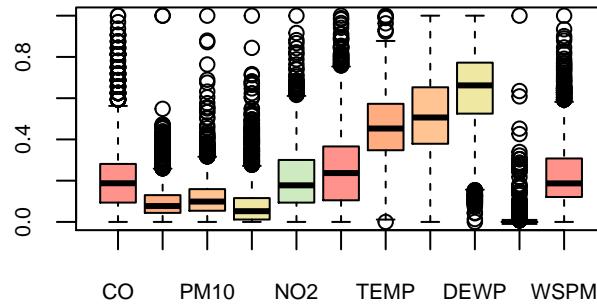
## April



```
## [1] ""
```

```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

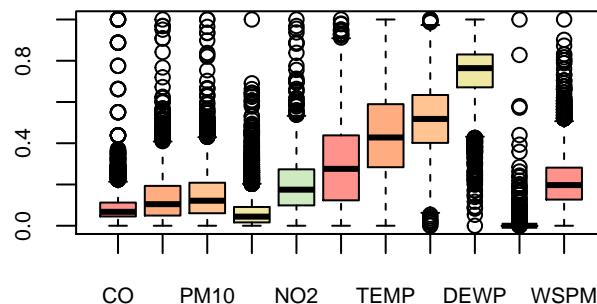
## May



```
## [1] ""
```

```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

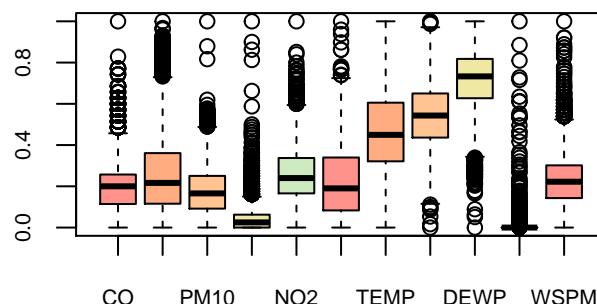
## June



```
## [1] ""
```

```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

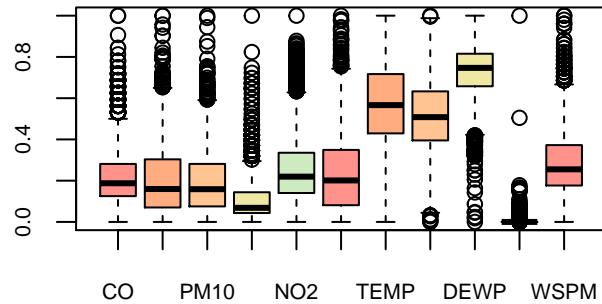
## July



```
## [1] ""
```

```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

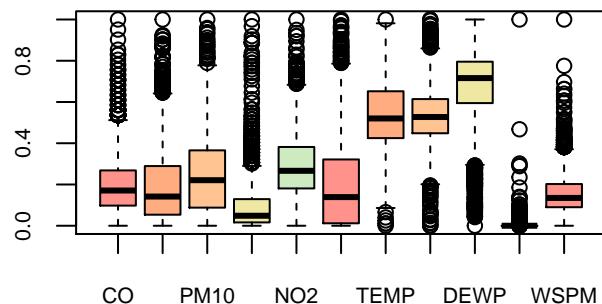
## August



```
## [1] ""
```

```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

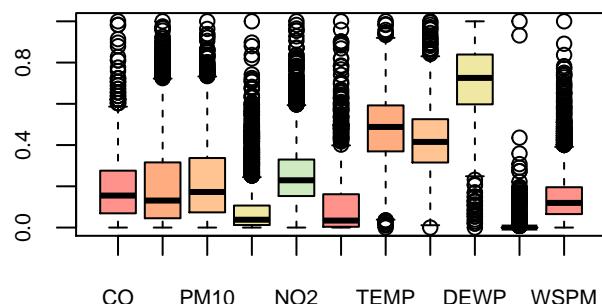
## September



```
## [1] ""
```

```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

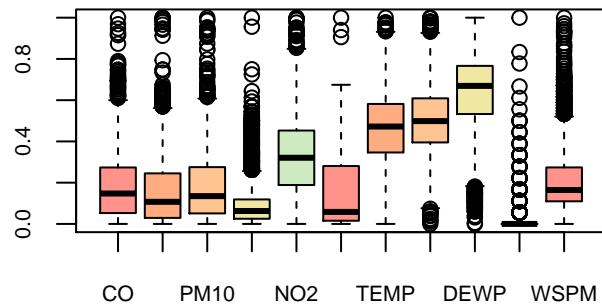
## October



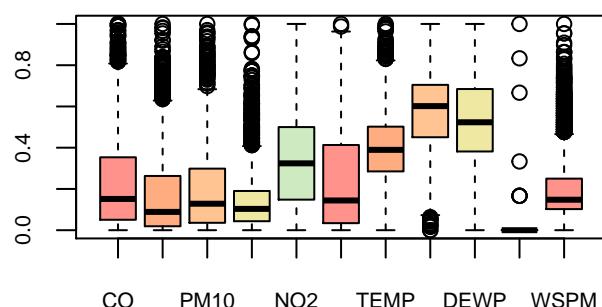
```
## [1] ""
```

```
## Warning in par(cxy = c(0.5, 0.5)): graphical parameter "cxy" cannot be set
```

## November



## December



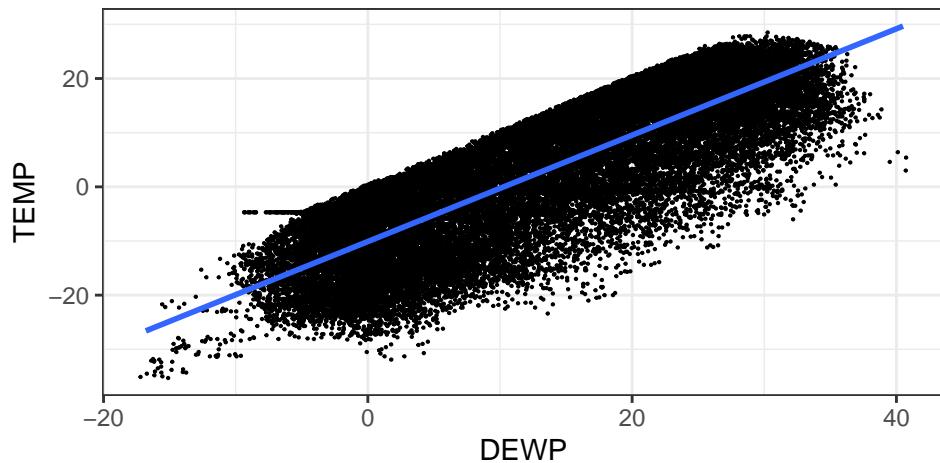
```
## [1] "
```

Apartir da análise mencionada anteriormente foi identificada que as varáveis “PM2.5”, “PM10”, “SO2”, “NO2”, “CO”, “O3”, “TEMP”, “PRES”, “DEWP”, “RAIN”, “wd”, “WSPM” possuem missing values.

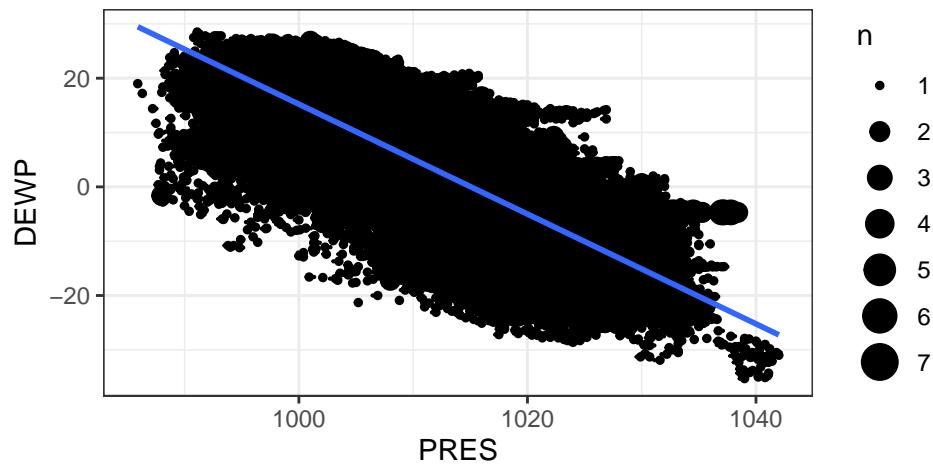
Em seguida foi uma análise e foi verificado que as variáveis “TEMP”, “PRES”, “DEWP”, “RAIN” esta a faltar no mesmo dia atravez do gráfico:

Como a variável DEWP possui muitos missings values fizemos um estudo relacionando a varável DEWP e as outras variáveis climaticas para assim tentarmos identificar alguma relação entre elas. Concluimos que existe uma relação direta com a TEMP e uma relação inversa com A variável PRES.

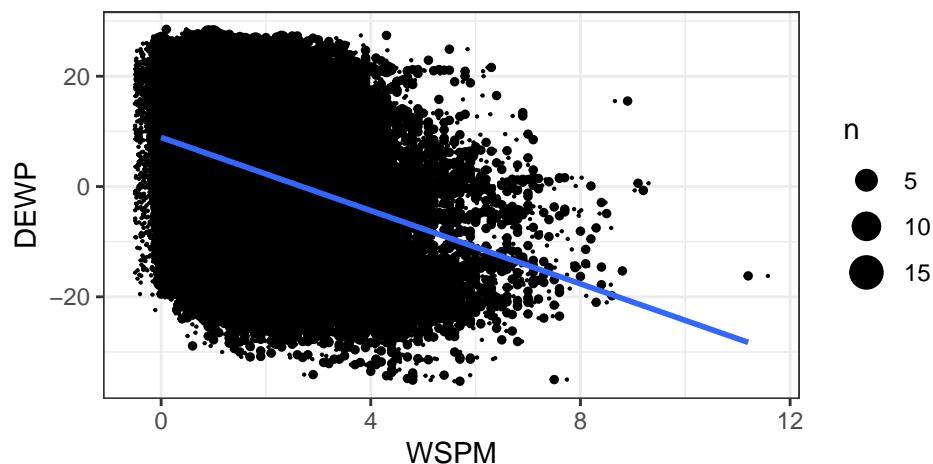
**TEMP**  
Relação entre TEMP e DEWP

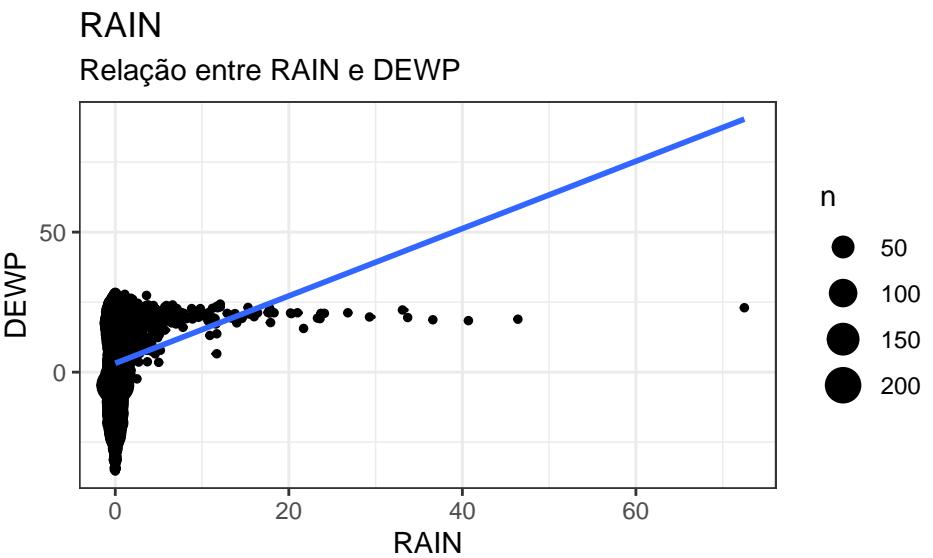


**PRES**  
Relação entre PRES e DEWP



Jittered Points  
Relação entre WSPM e DEWP






---

```
# Analise dos dados
```

---

## Predictive modelling

installs and imports

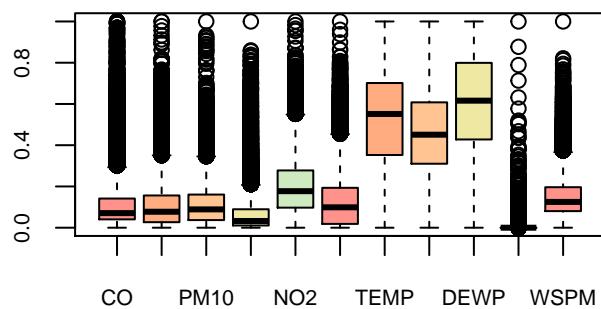
Normalizar os dados

Separar o dataset in training e test

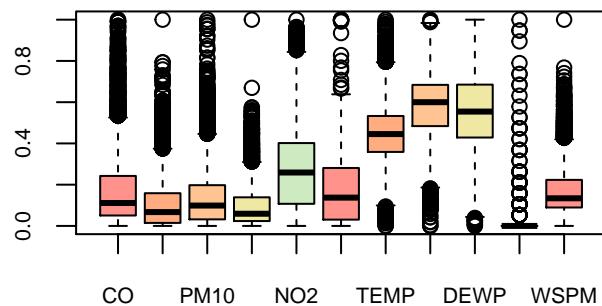
Criar o modelo (rede neuronal)

Utilizando a biblioteca neuralnet

Utilizando a biblioteca tree

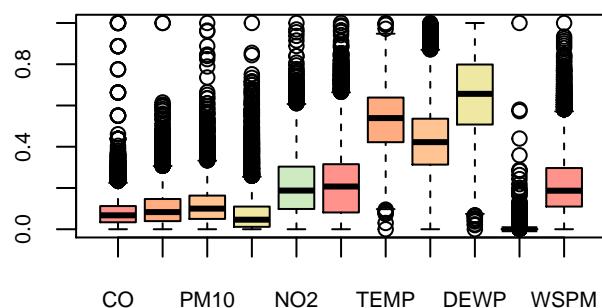


### Winter



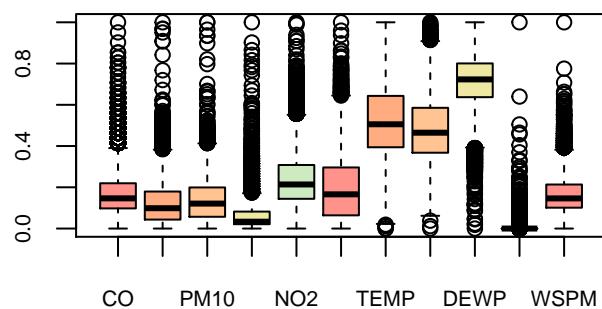
```
## [1] "
```

### Spring



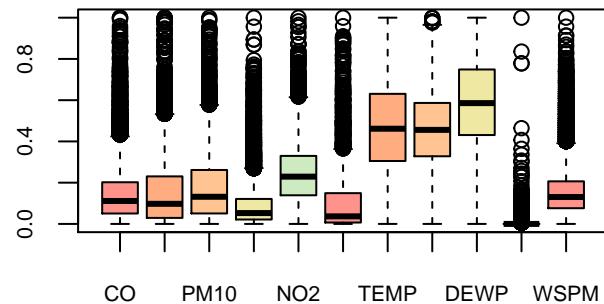
```
## [1] "
```

### Summer



```
## [1] "
```

## Autumn



```
## [1] "
```