Análise Descritiva, Correlativa e Preditiva da qualidade do Ar da China

Thiago Siriaki Valentini

thiagovalentini@hotmail.com (mailto:thiagovalentini@hotmail.com)

Curitiba, 25 de Dezembro de 2019.

A análise a seguir utiliza como referência os dados coletados de hora em hora pela Escola de Administração de Guanghua, Centro de Ciência Estatística, da Universidade de Pequim, que foram coletadas entre 01/01/2010 à 31/12/2014 na Embaixada dos EUA em Pequim, utilizando a base: **PM2.5** (https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data)

Dados

Abaixo podemos ver as 15 primeiras observações de *PM2.5* das 43.824 coletadas durante o período do estudo. A PM 2.5, ou “Partículas minúsculas”, são assim chamadas por medirem menos de 2,5 micrômetros (milionésimo de metro), e se formam a partir de gases como dióxido de enxofre, óxido de nitrogênio e compostos voláteis, liberados durante as atividades de combustão.

No	year	month	day	hour	pm2.5	DEWP	TEMP	PRES	cbwd	Iws	Is	Ir
1	2010	1	1	0	NA	-21	-11	1021	NW	1.79	0	0
2	2010	1	1	1	NA	-21	-12	1020	NW	4.92	0	0
3	2010	1	1	2	NA	-21	-11	1019	NW	6.71	0	0
4	2010	1	1	3	NA	-21	-14	1019	NW	9.84	0	0
5	2010	1	1	4	NA	-20	-12	1018	NW	12.97	0	0
6	2010	1	1	5	NA	-19	-10	1017	NW	16.10	0	0
7	2010	1	1	6	NA	-19	-9	1017	NW	19.23	0	0
8	2010	1	1	7	NA	-19	-9	1017	NW	21.02	0	0
9	2010	1	1	8	NA	-19	-9	1017	NW	24.15	0	0
10	2010	1	1	9	NA	-20	-8	1017	NW	27.28	0	0
11	2010	1	1	10	NA	-19	-7	1017	NW	31.30	0	0
12	2010	1	1	11	NA	-18	-5	1017	NW	34.43	0	0
13	2010	1	1	12	NA	-19	-5	1015	NW	37.56	0	0
14	2010	1	1	13	NA	-18	-3	1015	NW	40.69	0	0
15	2010	1	1	14	NA	-18	-2	1014	NW	43.82	0	0

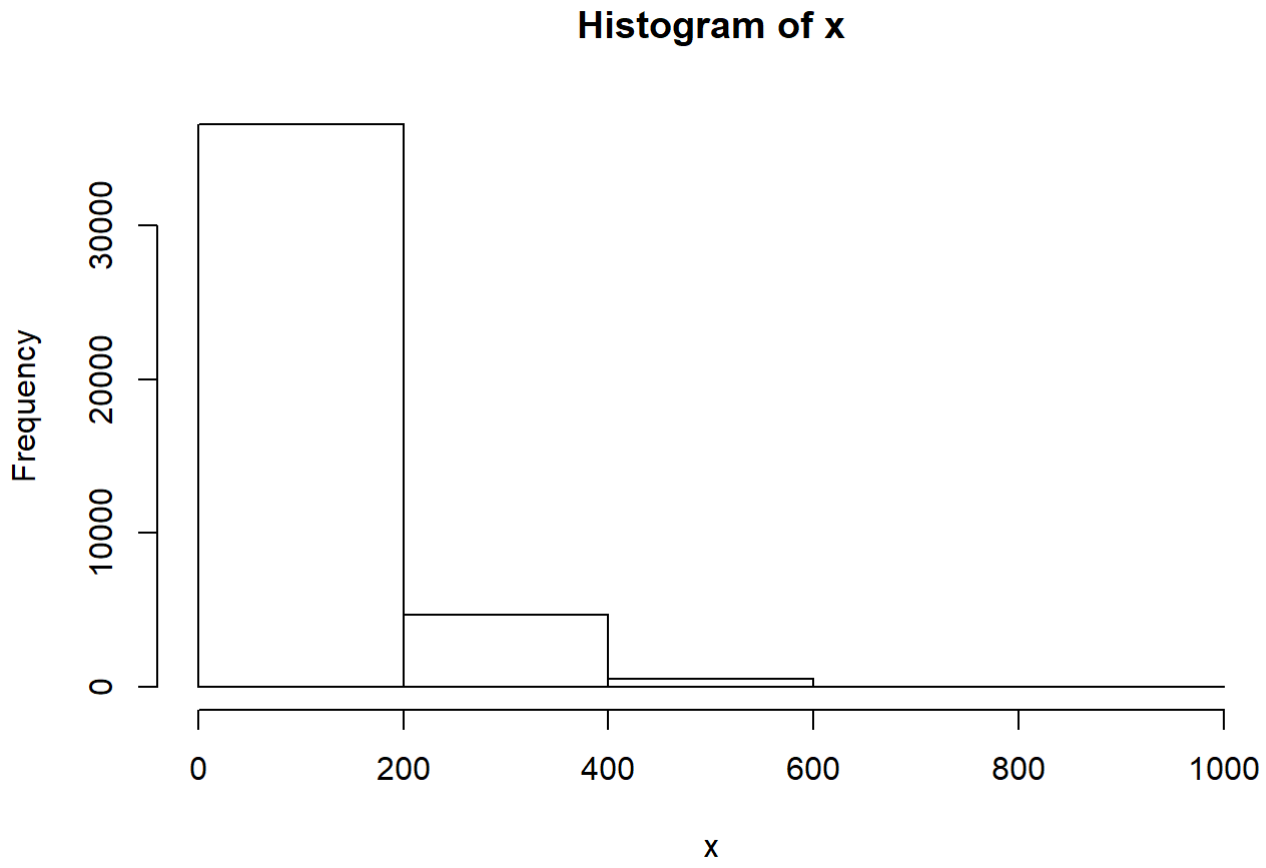
Análise de Quartis

Através dos Quartis, dividiremos de maneira ordenada o conjunto em 4 partes iguais e desta forma, poderemos analisar o grau de dispersão dos dados.

	PM2.5
0%	0
25%	29
50%	72
75%	137
100%	994

Os dados apresentados mostram uma grande concentração dos valores em um lado da curva. Aproximadamente 81% deles possuem concentração menor ou igual à 200ug/m^3 de PM2.5.

Através do histograma abaixo pode-se ter uma melhor visualização dessa análise:



Análise Correlativa

A Análise Correlativa ou Análise de Correlação é utilizada para determinar, de forma estatística, a relação entre duas variáveis. Seu resultado varia entre 1 e -1, onde -1 significa sem nenhuma relação e 1 com total relação.

Para testarmos as Correlações entre os dados, iremos utilizar os três principais métodos correlativos: ***Pearson***, ***Spearman*** e ***Kendall***.

Método de Pearson:

	PM2.5
hour	-0.0231164
DEWP	0.1714233
TEMP	-0.0905340
PRES	-0.0472823
Iws	-0.2477844
Is	0.0192656
Ir	-0.0513687

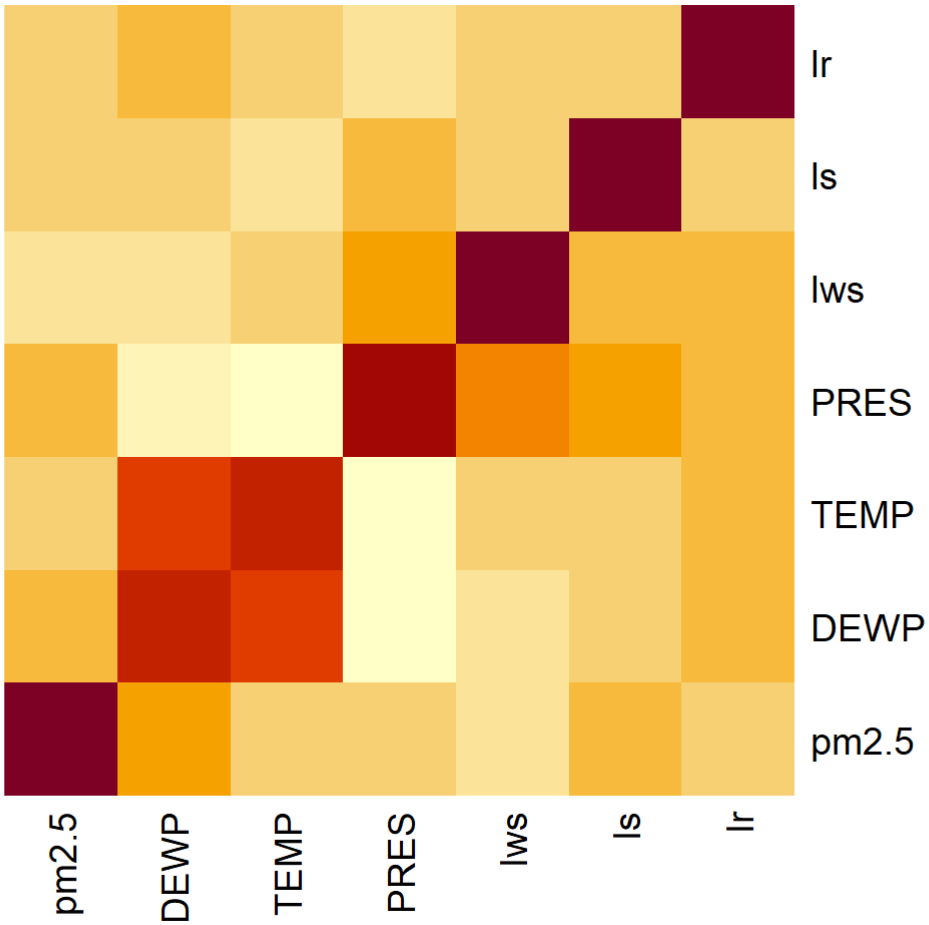
Método de Spearman:

	PM2.5
hour	-0.0256669
DEWP	0.2997061
TEMP	0.0107836
PRES	-0.1416678
Iws	-0.3600188
Is	0.0469782
Ir	-0.0023592

Método de Kendall:

	PM2.5
hour	-0.0168581
DEWP	0.2014732
TEMP	0.0073185
PRES	-0.0918995
Iws	-0.2496211
Is	0.0383712
Ir	-0.0021094

Correlações mostradas pelo heatmap:



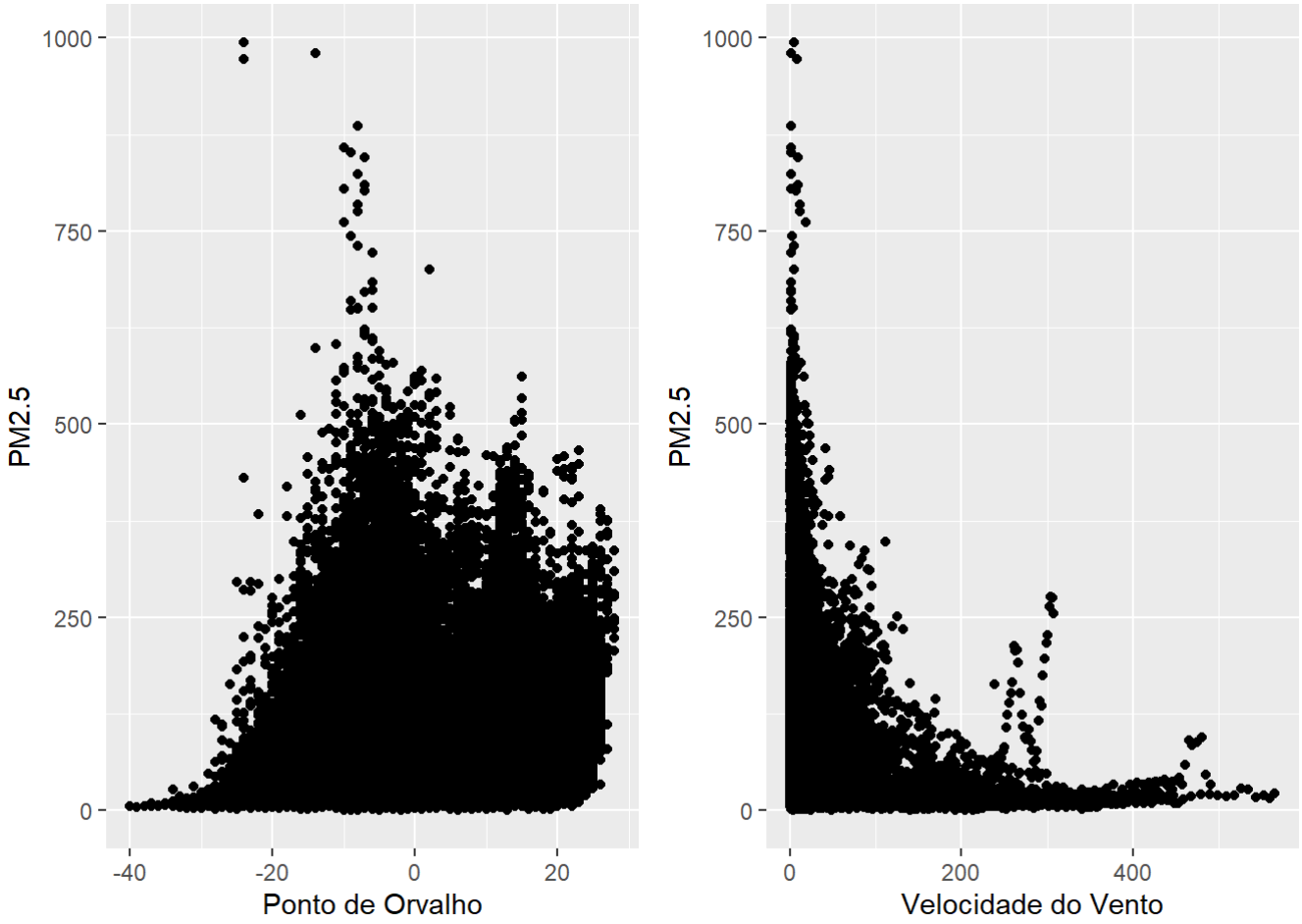
Quanto mais forte for a cor, maior será a correlação, enquanto que, quanto mais fraca a cor, menor a correlação.

A partir dos dados mostrados podemos identificar duas correlações semelhantes nos três métodos:

- O Ponto de Orvalho *DEWP* apresenta a maior correlação com a PM2.5, ou seja, podemos supor que quanto maior for o Ponto de Orvalho, maior será a concentração destas partículas presentes no ar nas gotas formadas durante a noite, aumentando assim sua concentração.
- Já a Velocidade do Vento *IWS* apresenta a menor correlação com PM2.5, ou seja, podemos novamente supor que quanto maior for a velocidade do vento, menor será a concentração desta partícula.

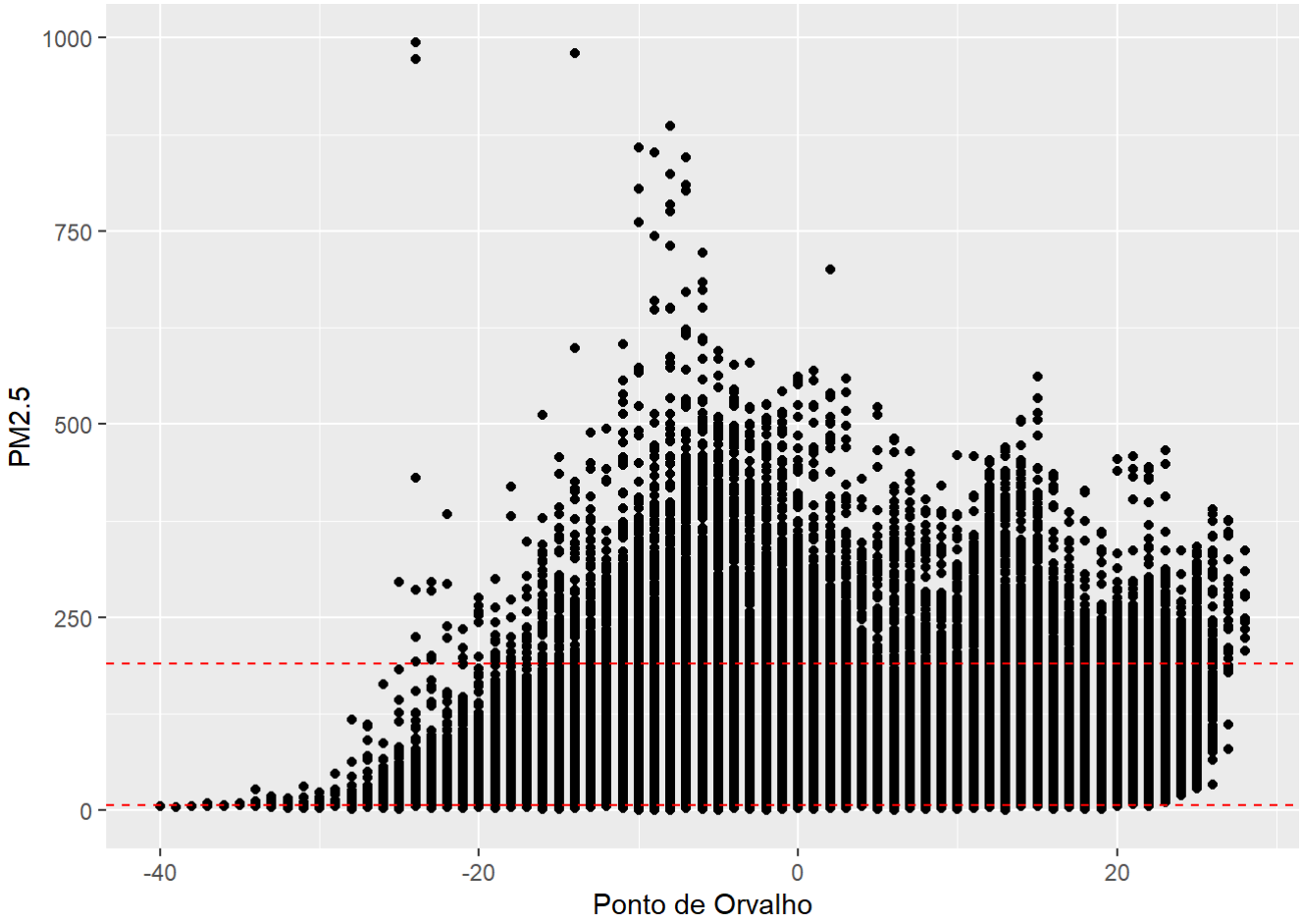
Análise de dispersão

Abaixo podemos observar o comportamento das duas principais variáveis identificadas, com relação à partícula PM2.5.



Utilizamos a análise de dispersão para identificar os comportamentos normais e anormais entre as variáveis.

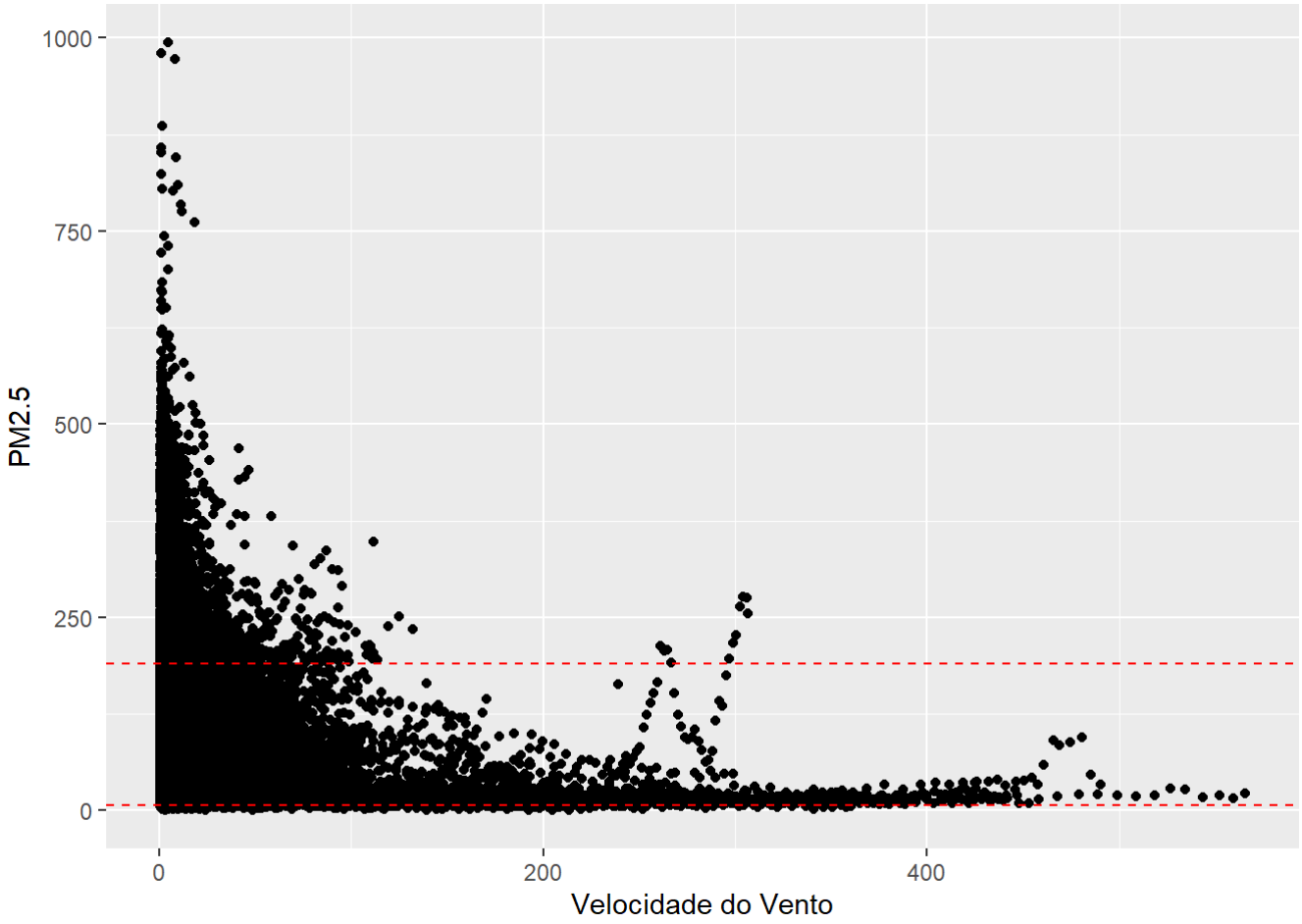
Ponto de Orvalho



Analisando o Gráfico, podemos notar que a concentração de PM2.5 é extremamente baixa quando o Ponto de Orvalho está entre -40 e -30, aumentando a partir daí e começando a reduzir a partir de 10.

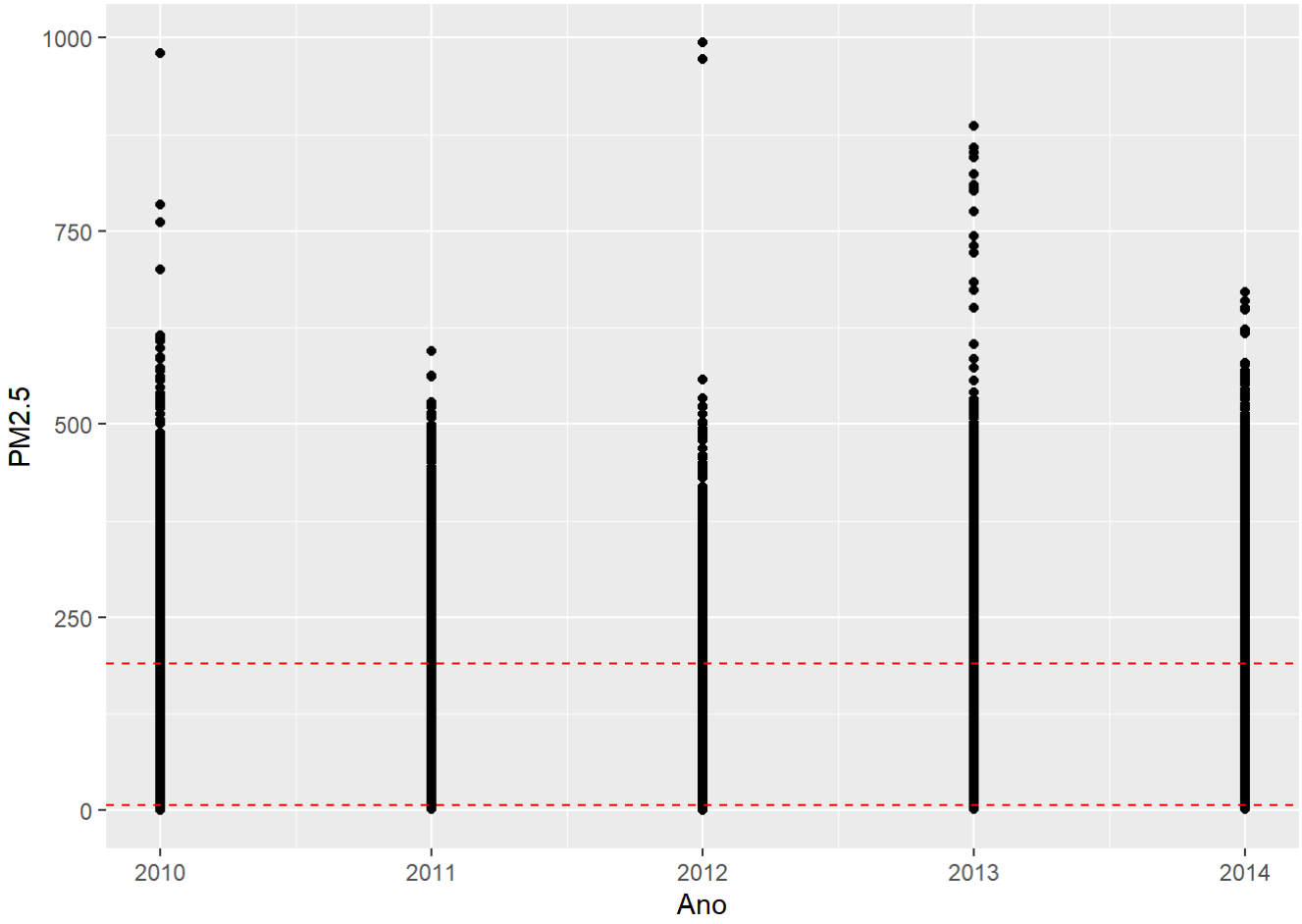
Diferente do que foi anteriormente suposto, podemos notar uma forte concentração da PM2.5 quando o **Ponto de Orvalho** está entre -20 e 0, e não quando este valor está chegando em seu máximo.

Velocidade do Vento



Analisando o Gráfico podemos constatar o que foi suposto anteriormente: Quanto maior for a velocidade do vento, menor será a concentração da Partícula PM2.5.

Concentração de PM2.5 por ano:



Através do Gráfico, podemos identificar um forte aumento na concentração de PM2.5 no ano de 2013, embora que em 2010 e 2012 também apresentem, mesmo que com menor frequência, os maiores valores medidos. O ano com menor concentração de PM2.5 foi 2011 seguido de 2014.

Diferente do que se poderia esperar, os dados não mostraram necessariamente que com o passar dos anos a concentração de PM2.5 aumentaria, mas sim que ela é alta e está presente na vida de milhares de Chineses.

Análise Preditiva através da Regressão Linear Múltipla

Utilizando a Regressão Linear Múltipla, tentaremos demonstrar a partir das variáveis mostradas nos dados, o comportamento da Partícula PM2.5 através de uma equação. Abaixo seguem os coeficientes desta equação sem ajuste:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1728.4313347	72.9899758	23.680393	0.0e+00
base3$DEWP	4.2822083	0.0534548	80.108929	0.0e+00
base3$TEMP	-6.0681011	0.0683620	-88.764250	0.0e+00
base3$PRES	-1.5291410	0.0713510	-21.431247	0.0e+00
base3$Iws	-0.2616415	0.0084361	-31.014553	0.0e+00
base3$Is	-2.2668997	0.5096877	-4.447625	8.7e-06

	Estimate	Std. Error	t value	Pr(>|t|)
base3$Ir	-7.2063495	0.2815797	-25.592575	0.0e+00

O modelo apresentou um R^2 (coeficiente de determinação) ajustado de 0.236, ou seja, ele é capaz de explicar apenas 23.6% das variações da PM2.5.

Ainda existem outros pontos para serem analisados, sendo o primeiro deles o intercepto da função, que expressa um número caso o valor das outras variáveis seja zero.

Excluindo o intercepto da função, obtemos os coeficientes abaixo:

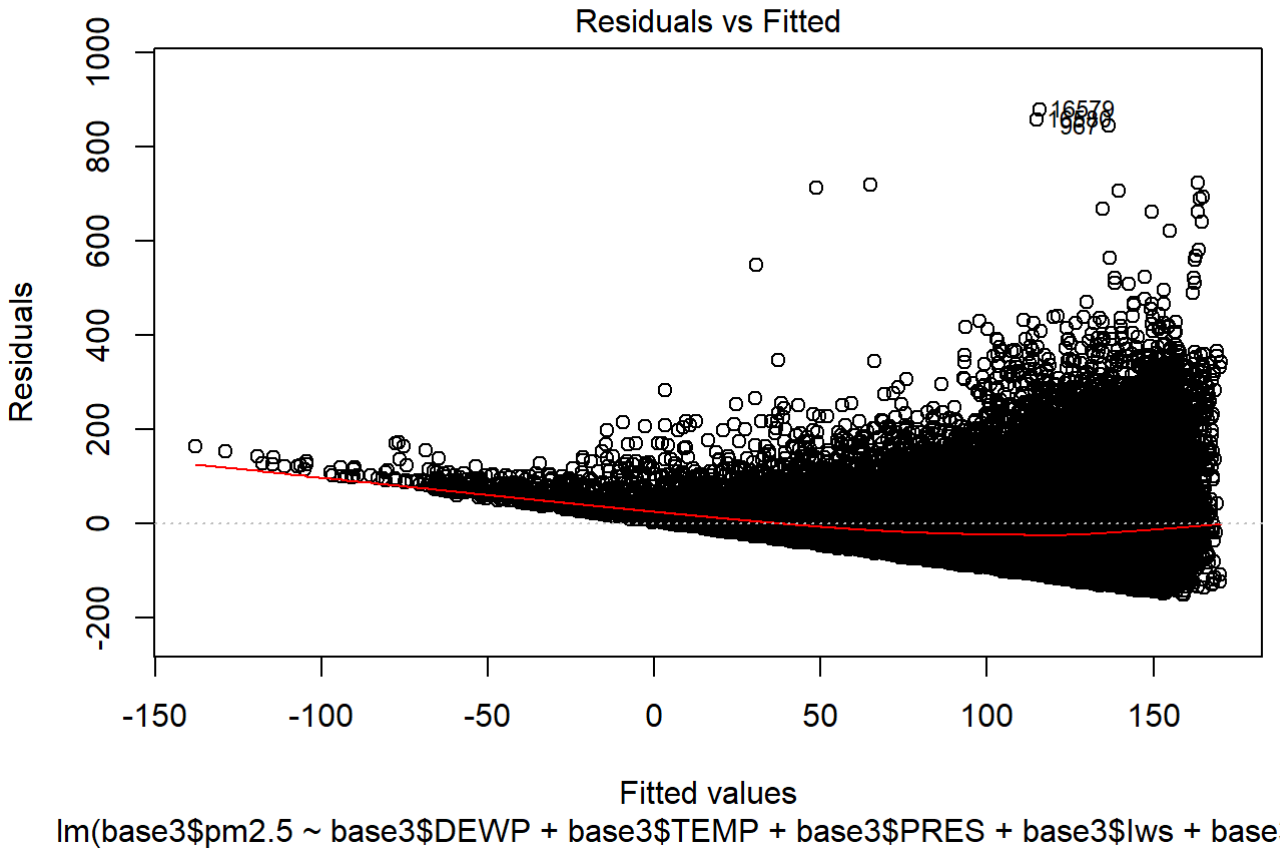
	Estimate	Std. Error	t value	Pr(>|t|)
base3$DEWP	4.6294523	0.0517476	89.462099	0.0e+00
base3$TEMP	-5.2236362	0.0587134	-88.968340	0.0e+00
base3$PRES	0.1603841	0.0007607	210.848203	0.0e+00
base3$Iws	-0.2632018	0.0084922	-30.993395	0.0e+00
base3$Is	-2.3568321	0.5130788	-4.593509	4.4e-06
base3$Ir	-7.0163529	0.2833459	-24.762500	0.0e+00

Com este novo modelo conseguimos melhorar o R^2 de 0.236, para 0.64.

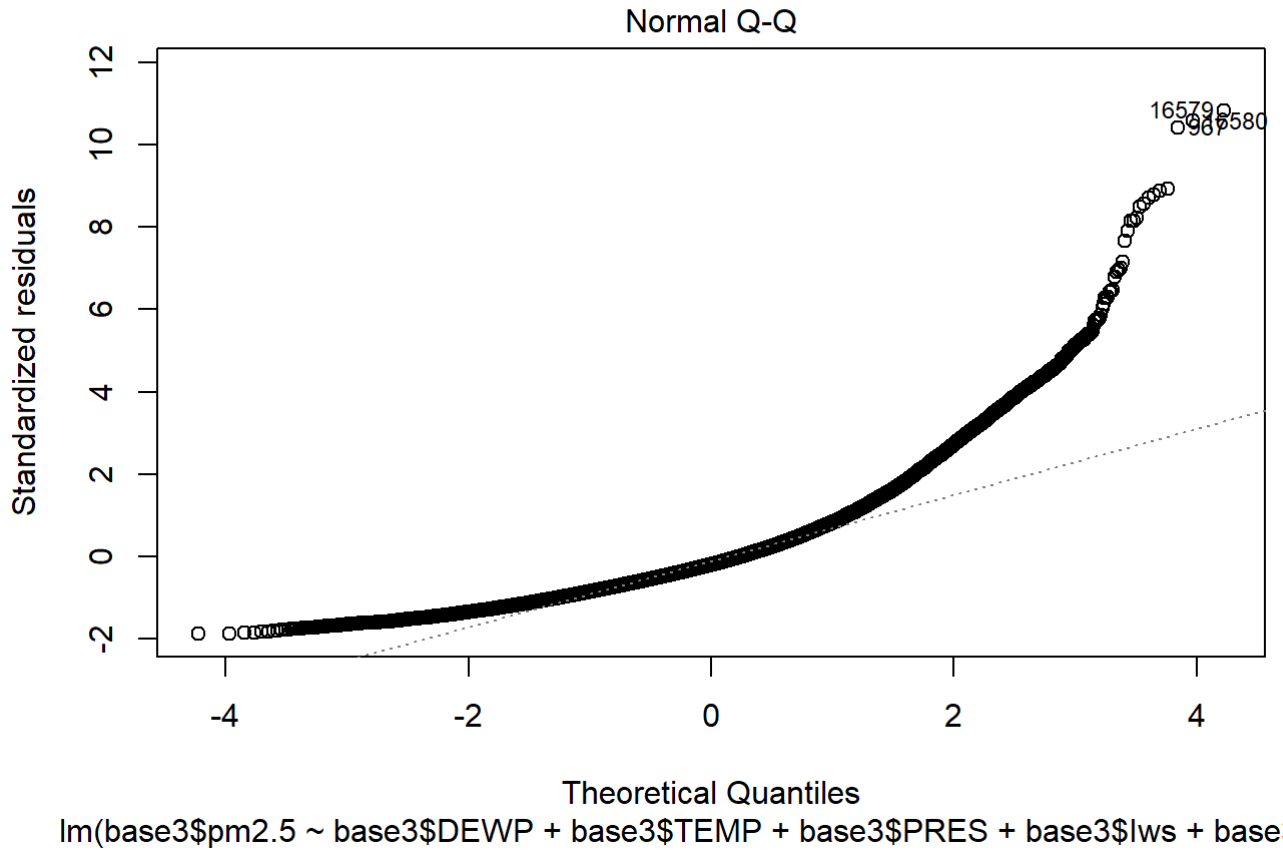
O Segundo Ponto a ser analisado seria o valor de P ou Probabilidade de Significância, onde levando em conta uma significância de 5% para o modelo, valores iguais ou abaixo a isso são estatisticamente significativos, enquanto que os valores acima não são. Para este caso, *aceitamos a hipótese de que a variável é significativa*.

O Terceiro Ponto é analisar se o modelo atende as três principais premissas da regressão linear: 1 - y Possa ser expressado de forma linear ou ter sua função “linearizada”; 2 - Hipótese da Homocedasticidades, onde a variação dos resíduos das observações são constantes em torno da reta; 3 - Os dados seguem uma distribuição normal de probabilidade;

Para realizar estas análises, utilizaremos os gráficos abaixo:



Podemos perceber que a reta se comporta de forma exponencial e os resíduos aumentam ao longo dela.



Neste caso vemos que os dados não seguem próximos a reta, e isso indica que sua distribuição não é boa para atender às premissas do modelo.

O Modelo pode ser ajustado ou ter sua função linearizada para verificar se ocorrerá melhora em suas premissas ou não.

A equação linear (não ajustada)

$$y_{PM2.5} = 4.62x_{DEWP} - 5.22_{TEMP} + 0.16x_{PRES} - 0.26x_{Iws} - 2.35x_{Is} - 7.01x_{Ir}$$

x_{DEWP}: Com o aumento de 1 grau na temperatura do Ponto de Orvalho, o resultado seria o aumento de 4.26 ug/m^3 de PM2.5.

x_{TEMP}: Com o aumento de 1 grau na temperatura, o resultado seria uma redução de 5.22 ug/m^3 de PM2.5.

x_{PRES}: Com o aumento de 1 hPa na pressão, o resultado seria o aumento de 0.16 ug/m^3 de PM2.5.

x_{Iws}: Com o aumento de 1m/s na velocidade do vento, o resultado seria uma redução de 0.26 ug/m^3 de PM2.5.

x_{Ir}: Com o aumento de 1 hora de neve, o resultado seria uma redução de 2.35 ug/m^3 de PM2.5.

x_{TEMP}: Com o aumento de 1 hora de chuva, o resultado seria uma redução de 7.01 ug/m^3 de PM2.5.