MAD 3 projekt
Regrese koncentrace pevných látek v Pekingu

Bc. Moravec Vojtěch

ZS 2019/2020

Obsah

1 Popis problému **3**

2 Popis datasetu **3**

2.1 Časové atributy 4

2.2 Distribuce hodnot a odlehlá pozorování numerických atributů 6

3 Předzpracování datasetu **9**

4 Regrese **10**

5 Přílohy **11**

1 Popis problému

Námi vybraný dataset [1] se zaměřuje na znečištění vzduchu v Pekingu od 1. Ledna 2010 do 31. Prosince 2014. Tento dataset jsme získali z UCI Machine Learning Repository [2]. Znečištěním rozumíme koncentraci v mikrogramech na metr krychlový ($\mu g/m^3$) pevných částic ve vzduchu. V našem případě se jedná o částice $PM_{2.5}$, jejíchž průměr je maximálně 2.5 µm. Našim cílem je tedy provést explorační analýzu datasetu a následně provést předpověď, regresi koncentrace znečištění vzhledem k času a přiloženým meteorologickým datům.

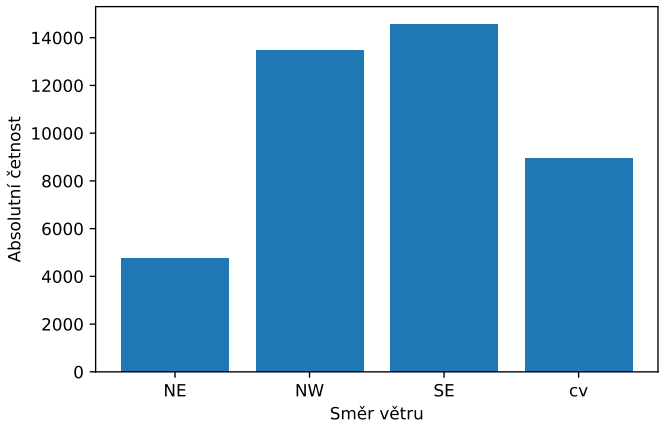
2 Popis datasetu

Dataset obsahuje celkem 43 824 záznamů a 12 atributů, nepočítáme-li číslo řádku. Ve 2 067 řádcích chybí cílová koncentrace a proto byly tyto řádky ihned odstraněny. Všechny záznamy jsou snímany v čase, tedy známe datum a čas měření, z něhož jsme vytvořili další atribut den v týdnu. Pro regresi máme tedy k dispozici 12 atributů, kde vetšina je numerická. Tyto numerické atributy, spolu s cílovým atributem jsou shrnuty v Tabulce 1.

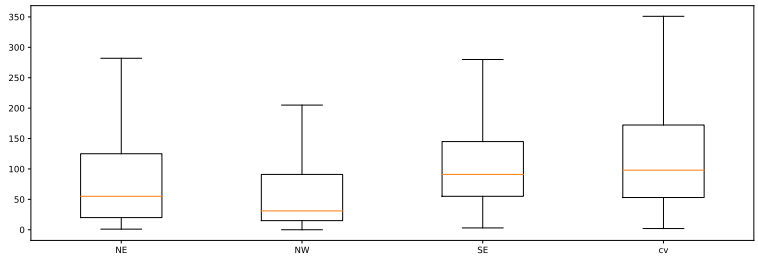
Název atributu	Průměr	σ	Min	Max
Rosný bod (°C)	1,7502	14,4337	-40,0000	28,0000
Teplota (°C)	12,4016	12,1752	-19,0000	42,0000
Tlak (hPa)	1016,4429	10,3007	991,0000	1046,0000
Rychlost větru (m/s)	23,8667	49,6175	0,4500	565,4900
Doba sněžení (hod.)	0,0553	0,7789	0,0000	27,0000
Doba deště (hod.)	0,1949	1,4182	0,0000	36,0000
Koncentrace $PM_{2.5}$ ($\mu g/m^3$)	98,6132	92.0504	0,0000	994,0000

Tabulka 1: Souhrn numerických atributů

Směr větru je kategoriální atribut, nabývající 4 různých hodnot, jejich distribuci můžeme vidět v grafu na Obrázku 1. Originální měřená data obsahovala celkem 16 různých směrů větru, ale autoři je seskupili do 5, kde 4 nalezeneme v našem datasetu. Hodnota CV znamená klidný, proměnný směr. Na Obrázku 2 můžeme vidět závislost cílové proměnné právě na směru větru, všimneme si, že největší koncentrace pevných látek ve vzduchu je právě při klidném větru a naopak nejmenší, když vítr fouká severozápadně.



Obrázek 1: Absolutní četnosti směru větru



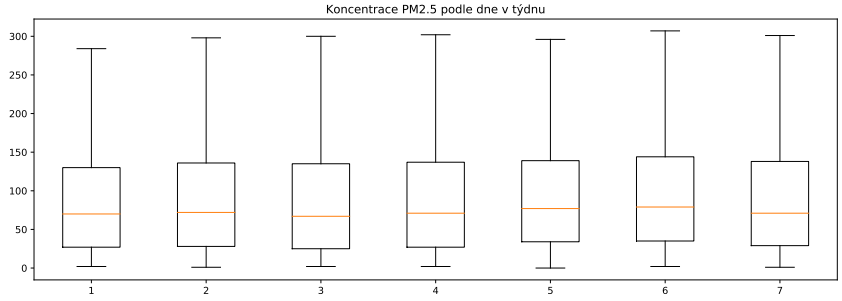
Obrázek 2: Koncentrace $PM_{2.5}$vzhledem ke směru větru

2.1 Časové atributy

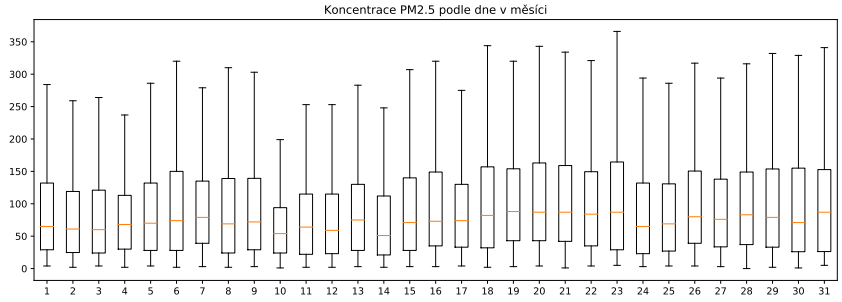
Každý záznam v datasetu obsahuje informaci o dnu, měsící, roku a hodině, kdy byl vytvořen. My jsme dále přidali den v týdnu. Všechny tyto atributy mohou být reprezentovány jako numerické nebo jako kategoriální. Aby jsme se mohli rozhodnout jak budeme s těmity atributy pracovat podíváme se na krabicové grafy cílového atributu v závislosti na dnu, měsící atd. Nejprve se podíváme na závislost pro dny v týdnu, tu můžeme vidět na Obrázku 3.

Zde nepozorujeme žádnou velkou závislost. Podobný výsledek můžeme vidět u dnu měsíce 4, měsíců 5 i roků 6. Koncentrace cílového atributu kolísá a nemůžeme najít žádnou závislost, že by se vzrůstajícím dnem, měsícem koncentrace $PM_{2.5}$rostla či klesala. Rozhodli jsme se tedy, že s atributy budeme pracovat jako s kategoriálními. Navíc vytvoříme další dataset, kde den v měsící zcela odstraníme.

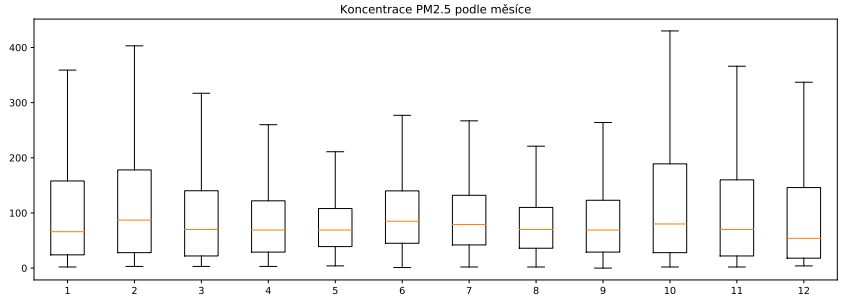
V prílohách ještě najdeme koncentrace vzhledem ke dnu v týdnu 13 a



Obrázek 3: Koncentrace $PM_{2.5}$vzhledem ke dnu v týdnu

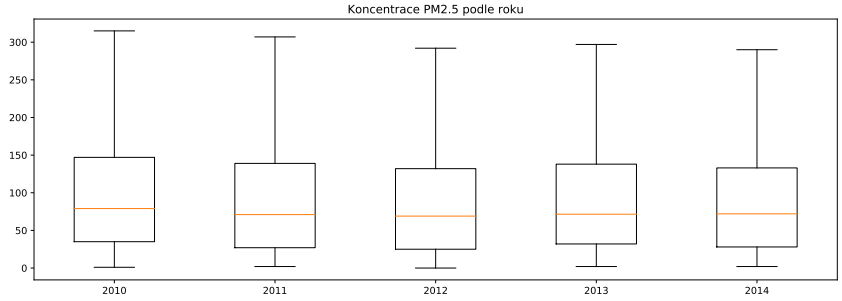


Obrázek 4: Koncentrace $PM_{2.5}$vzhledem ke dnu v měsící



Obrázek 5: Koncentrace $PM_{2.5}$vzhledem k měsící

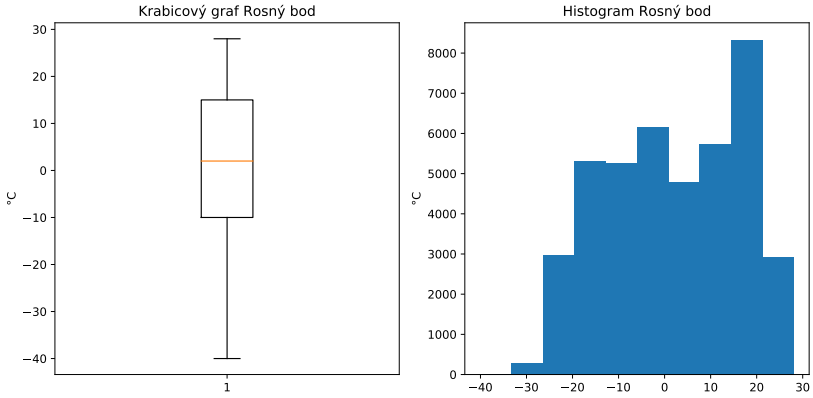
měsící 14 zvlášť pro každý rok.



Obrázek 6: Koncentrace $PM_{2.5}$vzhledem k roku

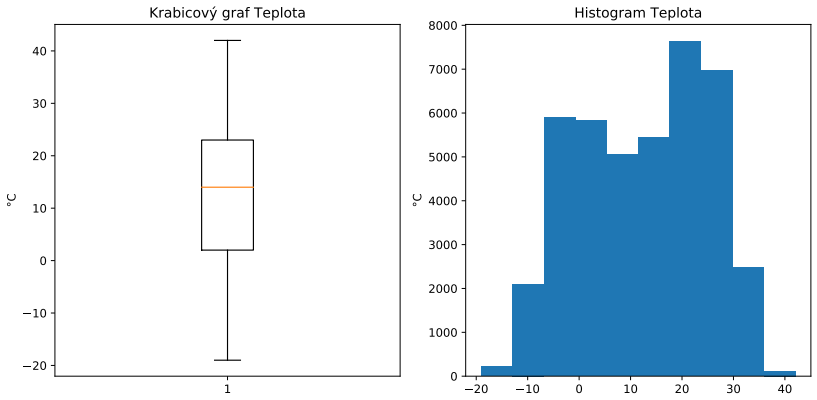
2.2 Distribuce hodnot a odlehlá pozorování numerických atributů

Většina hodnot rosného bodu se pohybuje v rozmezí od -20 °C do 20 °C. V rosném bodě nepozorujeme žádné odlehlé pozorování. Krabicový graf a histrogram můžeme vidět na Obrázku 7. Dále můžeme vidět grafy pro teplotu a to na Obrázku 8. Pro teplotu jsme také nenalezli odlehlé pozorování, to stejné platí pro atmosferický tlak.



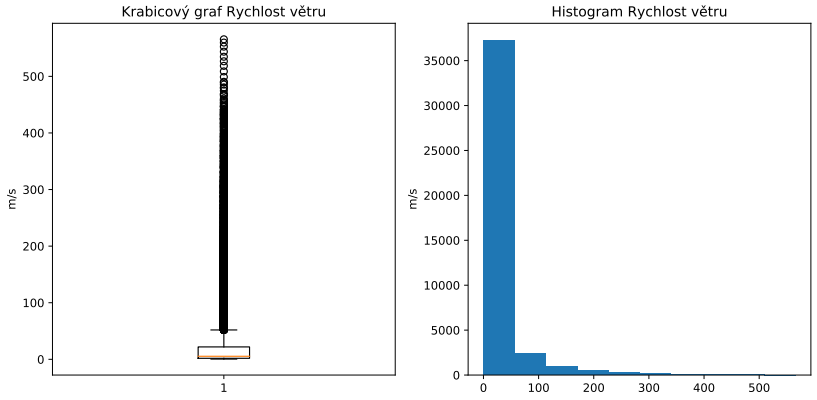
Obrázek 7: Krabicový graf, histrogram rosného bodu

Nejvíce odlehlých pozorování nalezneme pro rychlost větru, jedná se o 4 893 záznamů z celkových 41 757. Tyto pozorování můžeme vidět na grafu v Obrázku 9. Na histogramu vedle si všimneme, že rychlost větru je většinou do 100 m/s. Odlehlé pozorování tedy odpovídají extrému, kdy se do Pekingu dostal silný vítr. Tak stejně odlehlé pozorování najdeme u doby sněžení a deště. Obrázky 10 a 11. Drtivou většinu času v Pekingu nesněží ani neprší více jak hodinu v kuse, proto jsou hodnoty od 1 - 2 hodin označeny za



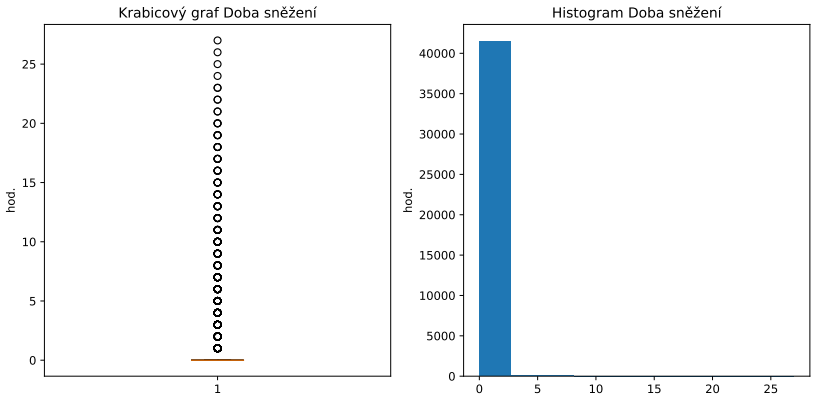
Obrázek 8: Krabicový graf, histrogram teploty

odlehlé pozorování. Dlé mého názoru by bylo špatné rozhodnutí tyto hodnoty odstranit, neboť právě ony mohou značit určitý pokles či vzrůst koncentrace $PM_{2.5}$

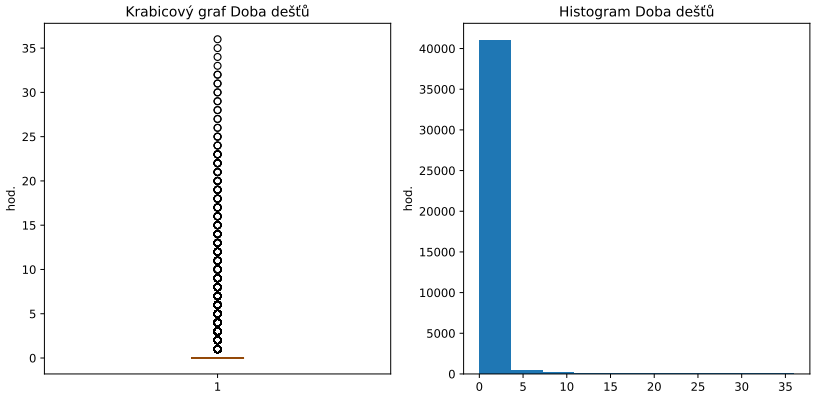


Obrázek 9: Krabicový graf, histrogram rychlost větru

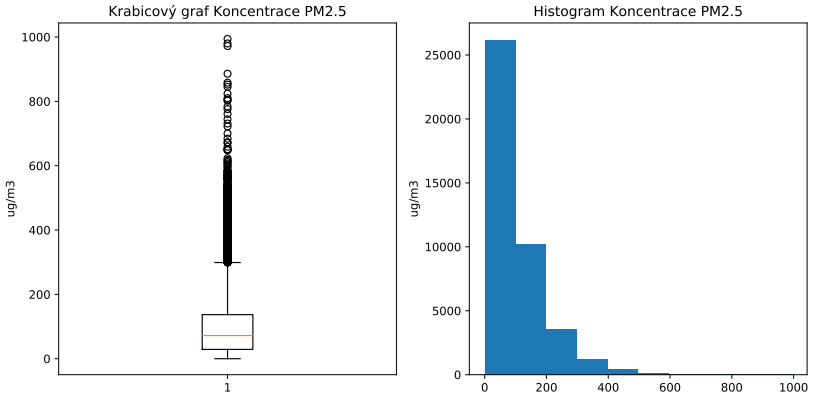
Pro cílový atribut koncentrace $PM_{2.5}$bylo nalezeneo 1773 odlehlých pozorování, jak můžeme vidět v Obrázku 12. Tyto odlehlé pozorování budou odstraněny pro všechny datasety, pro které budeme zkoušet regresi. Odlehlé pozorování z předchozích atributů odstraníme v nově vytvořeném datasetu a vyzkoušíme jaký vliv bude mít jejich přitomnost či absence.



Obrázek 10: Krabicový graf, histrogram doba sněžení



Obrázek 11: Krabicový graf, histrogram doba deště



Obrázek 12: Krabicový graf, histrogram koncentrace $PM_{2.5}$

3 Předzpracování datasetu

Zde si popíšeme operace, které jsme provedli s originálním datasetem, aby jsme získali naše datasety, nad kterými budeme zkoušet regresi. První dataset jsme získali tak, že jsme odstranili odlehlé pozorování cílového atributu $PM_{2.5}$a provedli binarizaci atributů den v týdnu, den v měsící, měsíc, rok a hodina. Výsledný dataset jsme následně normalizovali do rozmezí 0,0 až 1,0. Pro druhý dataset jsme provedli to stejné ale odstranili jsme den měření, následně jsme provedli normalizaci. Pro třetí dataset jsme odstranili všechny odlehlé pozorování, provedli binarizaci atributů a následnou normalizaci. Z tohoto datasetu jsme následně pomocí redukce dimenze vytvořili ještě další 2 datasety, kterým jsme nechali 20 resp. 10 nejdůležitějších atributů. Souhrn datasetu nalezeneme v Tabulce 2.

Jméno	Počet transakcí	Počet attributů
df_binAll	39984	89
df_binNoDay	39984	58
df_noOut_binNoDay	33512	58
df_noOut_binNoDay_20attr	33512	20
df_noOut_binNoDay_10attr	33512	10

Tabulka 2: Souhrn datasetů pro regresi

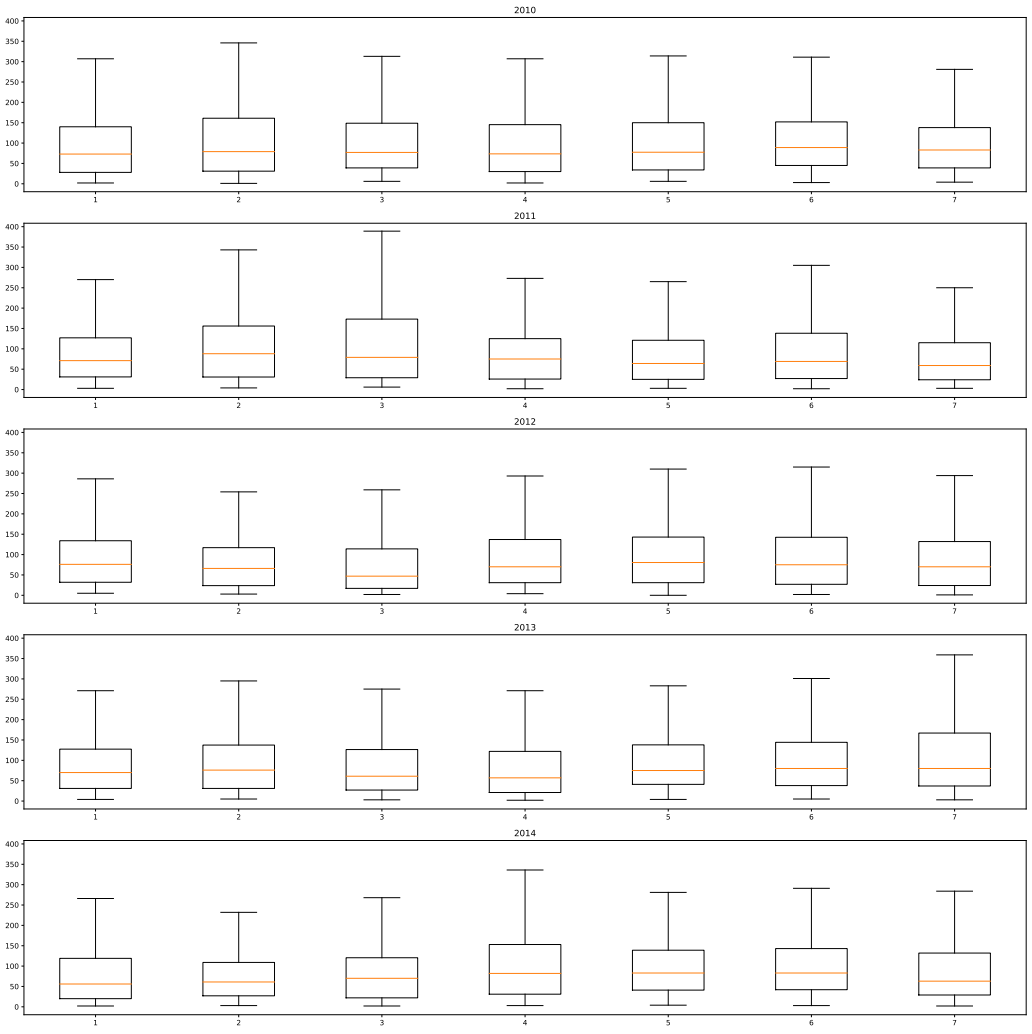
4 Regrese

Reference

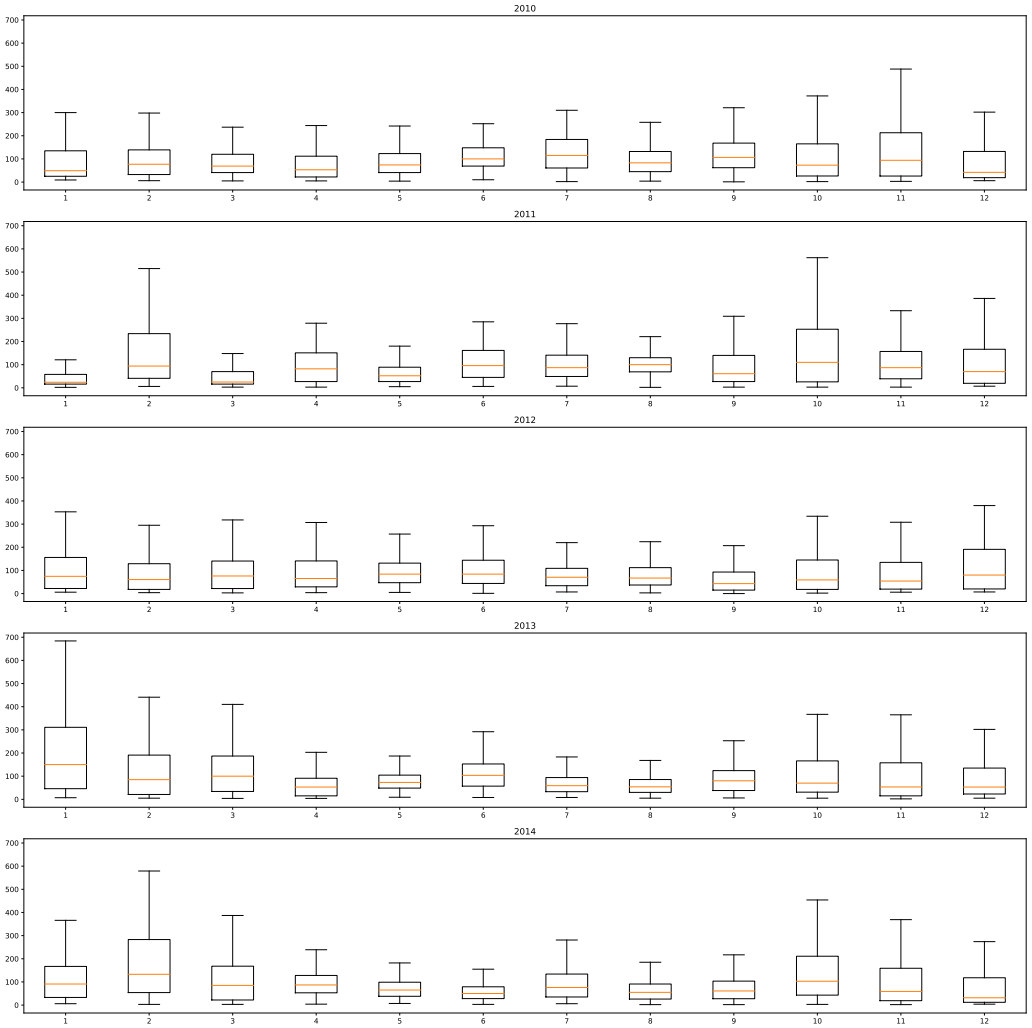
[1] Liang X., Zou T., Guo B., Li S., Zhang H., Zhang S., Huang H., and Chen S. X., "Assessing beijing's pm2.5 pollution: severity, weather impact, apec and winter heating," 2015.

[2] D. Dua and C. Graff, "UCI machine learning repository," 2017.

5 Přílohy



Obrázek 13: Koncentrace $PM_{2.5}$vzhledem ke dnu v týdnu pro každý rok



Obrázek 14: Koncentrace $PM_{2.5}$vzhledem k měsíci pro každý rok