

# Regrese koncentrace $\text{PM}_{2.5}$ v Pekingu

## MAD 3 projekt

---

Bc. Moravec Vojtěch

ZS 2019/2020

Vysoká škola báňská – Technická univerzita Ostrava

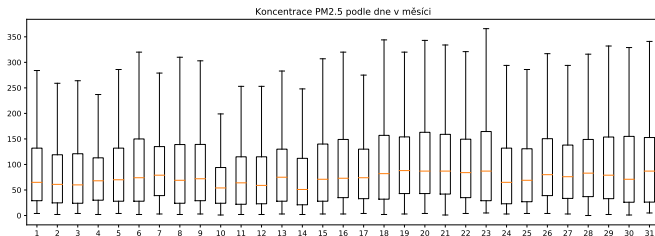
# Informace ohledně datasetu

- 43 824 záznamů a 12 atributů
- Data od 1.1.2010 až do 31.12.2014
- Záznamy o měření koncentrace pevných částí PM<sub>2.5</sub> ve vzduchu ( $\mu\text{g}/\text{m}^3$ )
- 6 numerických atributů + 1 cílový numerický
- 1 kategoriální atribut
- zbylé atributy reprezentují datum a čas
  - vytvořili jsme nový atribut "den v týdnu"

- Jak pracovat s atributy dne, měsíce, roku, hodiny, dne v týdnu?
- Numerický nebo kategoriální atribut?
- Podíváme se na závislost cílového atributu, vzhledem k těmto atributům
  - Nalezneme-li závislost (např. koncentrace roste s měsícem) - numerický
  - Jinak kategoriální a provedeme jejich binarizaci

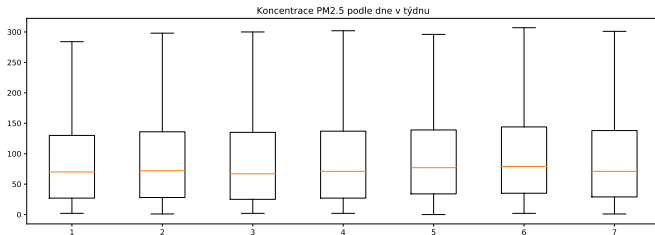
# Koncentrace vzhledem k měsíci

Nevidíme žádnou závislost, že koncentrace klesá nebo roste spolu s měsícem.



# Koncentrace vzhledem k dnu týdne

Koncentrace  $\text{PM}_{2.5}$  je nezávislá na dnu týdne.



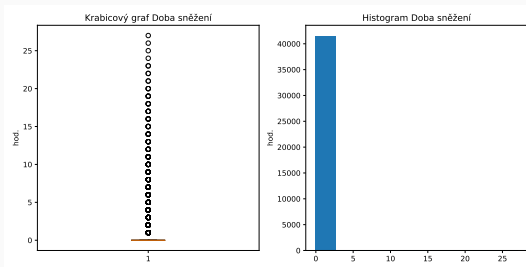
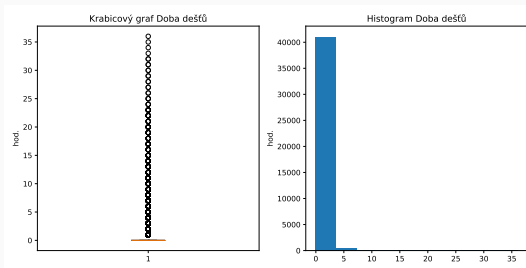
Odlehlé pozorování nalezené pro:

- Rychlost větru (4 893) <sup>1</sup>
- Doba deště (1 739) <sup>1</sup>
- Doba sněžení (368) <sup>1</sup>
- Koncentrace PM<sub>2.5</sub> (1 773) - všechny odstraněny

---

<sup>1</sup>Odlehlé pozorování byly ponechány v některých datasetech

# Histogram doba sněžení, deště

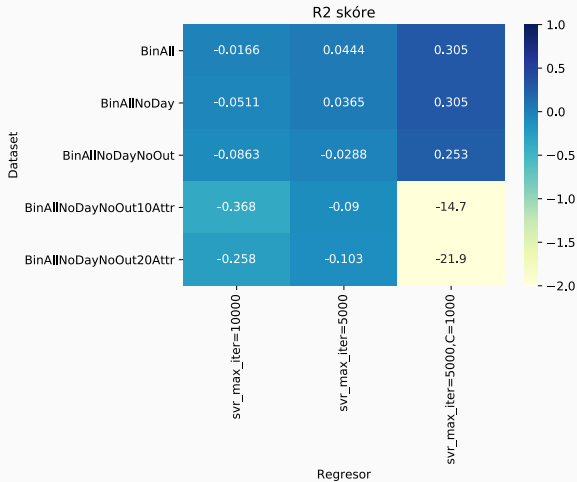


- Ve všech datasetech byla provedena normalizace hodnot do rozmezí 0,0 až 1,0

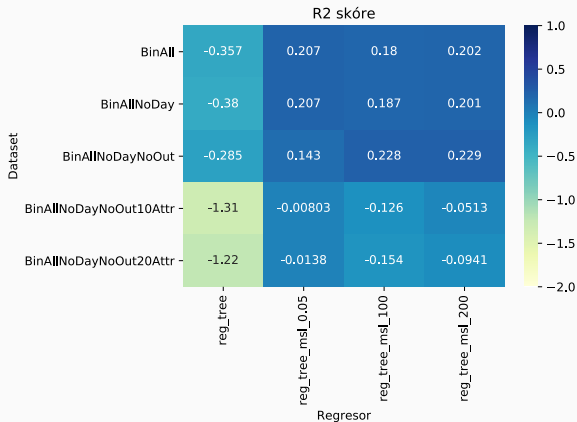
Dataset	Počet transakcí	Počet atributů
df_binAll	39984	89
df_binNoDay	39984	58
df_noOut_binNoDay	33512	58
df_noOut_binNoDay_20attr	33512	20
df_noOut_binNoDay_10attr	33512	10



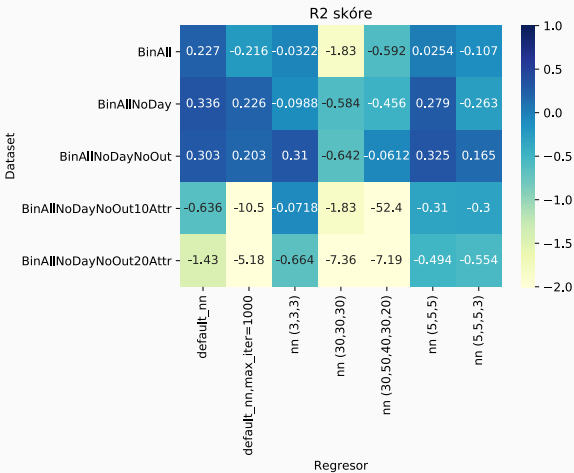
# Regrese pomocí SVR



# Regrese pomocí DecisionTreeRegressor



# Regrese pomocí MLPRegressor

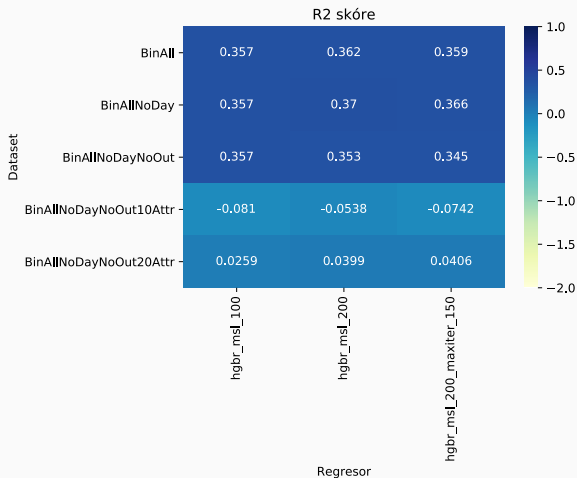


## Sourhn výsledků

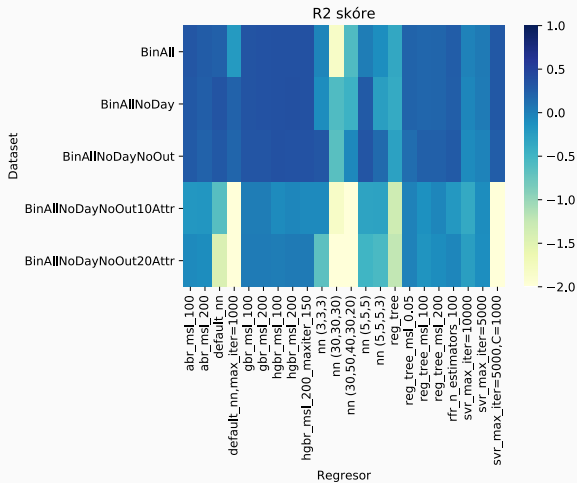
Algoritmus	Čas učení (s)	$R^2$	MSE
SVR	27,209	0,305	3326,722
DecisionTreeRegressor	0,078	0,229	3731,770
MLPRegressor	18,880	0,336	3172,276

- RandomForestRegressor vedl pouze k minimálnímu zlepšení
- Nejlepší boostovací metody seřazeny podle zlepšení (od nejlepšího):
  1. HistGradientBoostingRegressor
  2. GradientBoostingRegressor
  3. AdaBoostRegressor

# Boostovací metody - HistGradientBoostingRegressor



# Souhrn všech regresorů



- Provedli jsme explorační analýzu
- Vytvořili jsme datasety
- Otestovali jsme regresory a zhodnotili jsme je
- Výsledky regresorů jsou podprůměrné
- Koncentrace  $\text{PM}_{2.5}$  je těžko předpověditelná



**Děkuji za pozornost**  
**Otázky?**

---

