

Projekt M9121

Červenka Michal, Heger Martin, Husa Štěpán

3. prosince 2025

- 1 Úvod
- 2 Náhled na časové řady
- 3 Testování stacionarity
- 4 Volba vhodných modelů
- 5 Diagnostika finálního modelu
- 6 Diagnostika finálního modelu
- 7 Predikce
- 8 Závěr

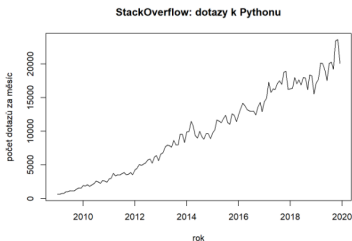
- **Popis dat:**

- Použitý dataset dostupný zde.
- Měsíční počty dotazů na StackOverflow od roku 2009 do roku 2020 pro různé programovací knihovny a technologie.

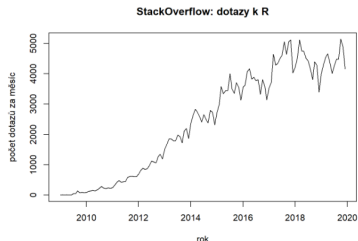
- **Cíl projektu:**

- Zaměříme se pouze na dvě proměnné – Python a R.
- Porovnání počtu dotazů k těmto programovacím jazykům v čase.
- Následně budeme modelovat a analyzovat časovou řadu, která popisuje počet dotazů k Pythonu.

Dotazy k R a Pythonu



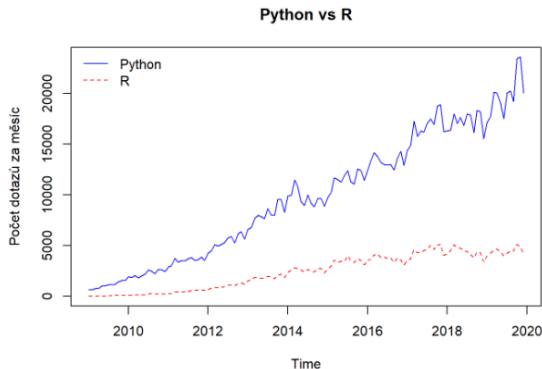
Dotazy k Pythonu



Dotazy k R

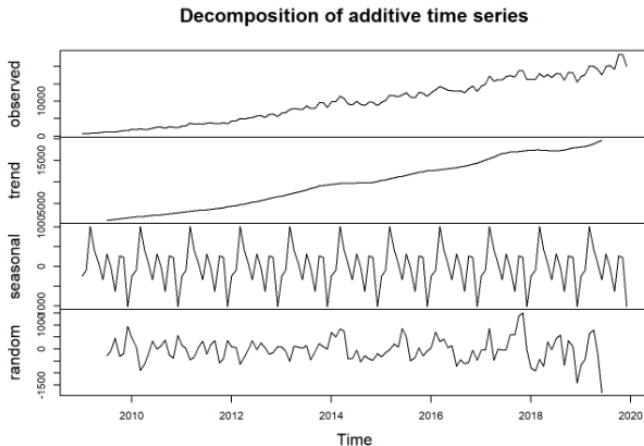
- Jasný rostoucí trend od roku 2009 do 2020.
- Postupem času narůstá i variabilita počtu dotazů.
- Viditelné pravidelné sezónní kolísání.
- **R:** Trend také roste, ale pomaleji a méně výrazně než u Pythonu.
- **R:** Amplituda sezónnosti je menší.
- **R:** Po roce 2017 se růst zpomaluje.

Společný graf a popisné statistiky



	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Python	631	3744	9652	9857	15591	23602
R	2.0	608.8	2613.5	2411.9	4000.5	5138.0

Aditivní dekompozice: Python

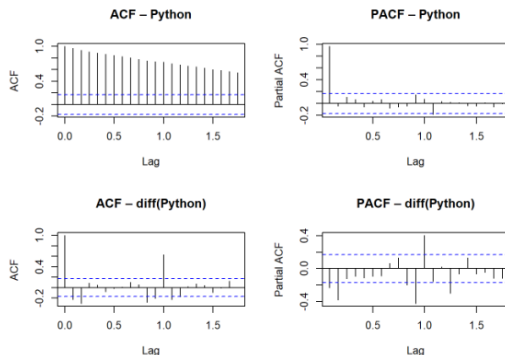


- **Trend:** silně rostoucí, plynulý.
- **Sezónnost:** pravidelná, opakující se ročně.
- **Remainder:** náhodné fluktuace bez jasné struktury.

Korelogramy: Python

Python

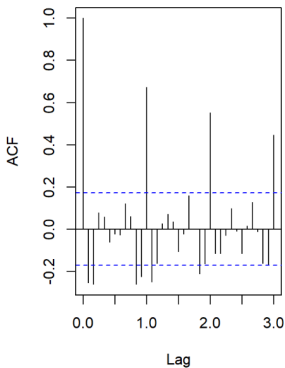
► Kód



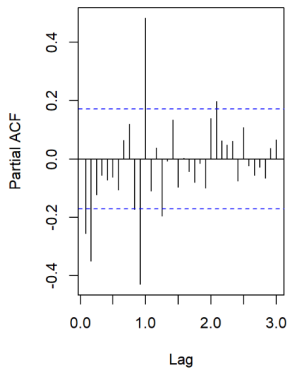
- Původní řada: ACF pomalu klesá → nestacionární, potvrzuje trend.
- Po diferenciaci: ACF má výrazný lag 1 a 2, PACF také → vhodné modely ARIMA(1,1,2) nebo blízké alternativy.

Analýza stacionarity: Srovnání diferencií

ACF – diff(Python)



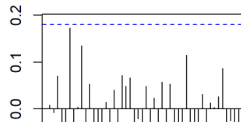
PACF – diff(Python)



ACF – diff(Python)



PACF – diff(Python)



Vycházíme z modelu SARIMA $(p, d, q)(P, D, Q)_s$:

- **Transformace:** Odmocnina pro stabilizaci rozptylu(box.cox-lambda= 0, 42).
- **Sezónnost ($s = 12$):**
 - $D = 1$: Sezónní difference kvůli roční cykličnosti.
 - $Q = 1$: ACF na lagu 12 naznačuje SMA(1) v sezónní části.
- **Trend a nesezónní část:**
 - $d = 1$: První difference pro odstranění trendu.
 - $q = 1$: ACF ukazuje významné lag 13.
- **Výchozí model:** $SARIMA(0, 1, 1)(0, 1, 1)_{12}$

Manuální volba modelu: Respecifikace

- **Model:** SARIMA(0, 1, 1)(0, 1, 1)₁₂ (t-test):

Coefficients:

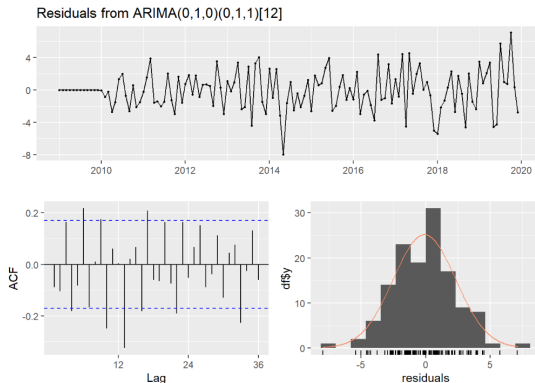
	ma1	sma1
	-0.1103	-0.4844
s.e.	0.1001	0.1031

sigma^2 = 6.513: log likelihood = -280.94
AIC=567.89 AICc=568.1 BIC=576.22

► Kód

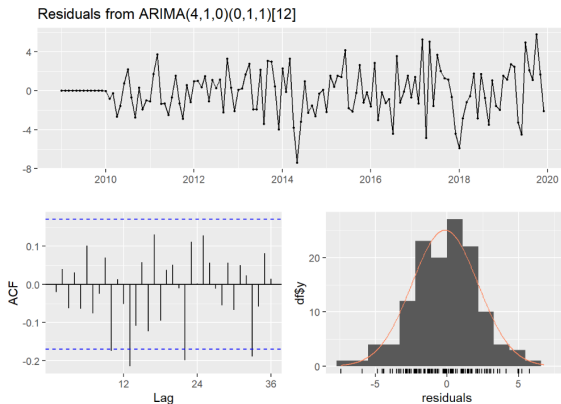
	Koeficient	t_statistika	Vyznamny
ma1	ma1	-1.102125	FALSE
sma1	sma1	-4.699447	TRUE

Finální manuální model: $SARIMA(0, 1, 0)(0, 1, 1)_{12}$



- **Diagnostika:** Ljung-Boxův test zamítá nezávislost reziduí
 $p\text{-value} = 1.745e - 08 < 0.05$.
- Rezidua vykazují autokorelaci, ale jde o nejlepší manuální výsledek.

Auto.arima model: $SARIMA(4, 1, 0)(0, 1, 1)_{12}$



- **Diagnostika:** Ljung-Boxův test zamítá nezávislost reziduí
 $p\text{-value} = 0.0185 < 0.05$.
- Rezidua vykazují autokorelaci.

Algoritmus `auto.arima` (s `stepwise=FALSE`) našel model:

$$SARIMA(4, 1, 0)(0, 1, 1)_{12}$$

Auto.arima model:

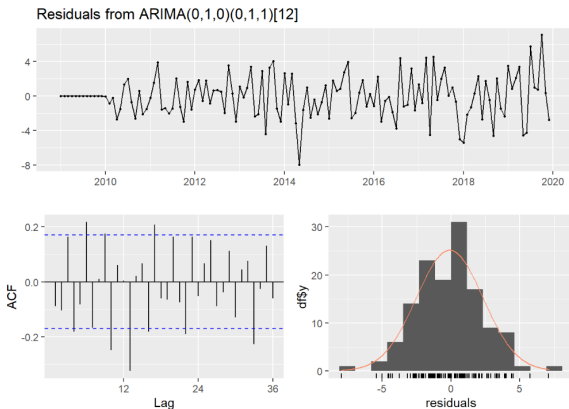
- $(4, 1, 0)(0, 1, 1)_{12}$
- AICc: **566.1312**

Manuální model (Vítěz):

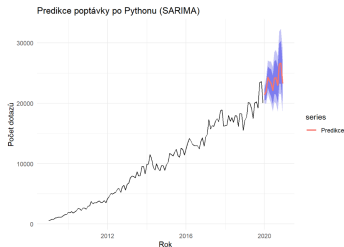
- $(0, 1, 0)(0, 1, 1)_{12}$
- AICc: **567.1595**

- Automatický model je nevýrazně lepší dle informačních kritérií, zato složitější.
- **Validace:** Oba přístupy měli shodu v sezonní části.

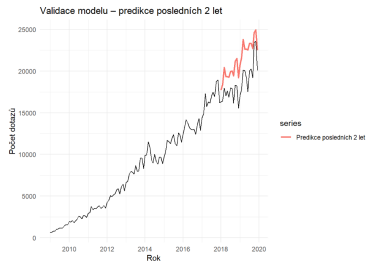
Diagnostika finálního modelu $((0, 1, 0)(0, 1, 1)_{12})$



- **Ljung-Boxův test:** $p\text{-value} = 1.745e - 08 < 0.05 \rightarrow$ zamítáme H_0 .
- **Interpretace:** Model nezachytil veškerou strukturu (pravděpodobně kvůli strukturálnímu zlomu po r. 2017), ale pro predikci trendu je dostačující.



Predikce poptávky na 12 měsíců dopředu



Validace: predikce posledních 2 let

- Analyzovali jsme vývoj popularity jazyka Python na StackOverflow.
- Zjistili jsme silný rostoucí trend a pravidelnou sezónnost.
- Odmocninová transformace a diferencování pomohly data stacionarizovat.
- Ruční analýza koeficientů a porovnání s automatická selekcí identifikovaly jako nejvhodnější model SARIMA(0, 1, 0)(0, 1, 1)[12].
- Byl použit pro predikci, která očekává další růst zájmu s typickými sezónními výkyvy.

- **Omezení modelu:**

- Hlavním limitem modelu je zamítnutí hypotézy o nezávislosti reziduí (Ljung–Boxův test).
- To naznačuje, že model nedokázal plně absorbovat veškerou strukturu dat, pravděpodobně v důsledku strukturálního zlomu v popularitě Pythonu kolem roku 2017.
- Dále je přítomna heteroskedasticita (změny rozptylu), kterou odmocninová transformace odstranila jen částečně.
- Model je vhodný pro bodovou predikci trendu, ale intervaly spolehlivosti mohou být mírně podhodnocené.

- **Možnosti vylepšení:**

- Přesnost modelu by mohla být zvýšena zkrácením časové řady (např. použitím dat pouze od roku 2015), čímž by se eliminoval vliv staršího, odlišného chování trendu.
- Mohlo by být vhodné využít pokročilejší modely odolné vůči strukturálním zlomům, jako např. Prophet.
- Lze také zahrnout exogenní proměnné (např. počet pracovních dnů v měsíci).

Děkujeme za pozornost!