Projekt M9121

Červenka Michal, Heger Martin, Husa Štěpán

22. listopadu 2025

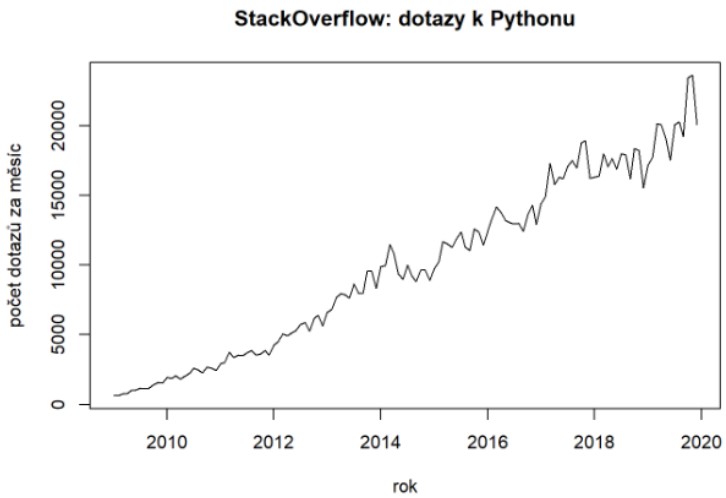
Obsah

Data a cíl projektu

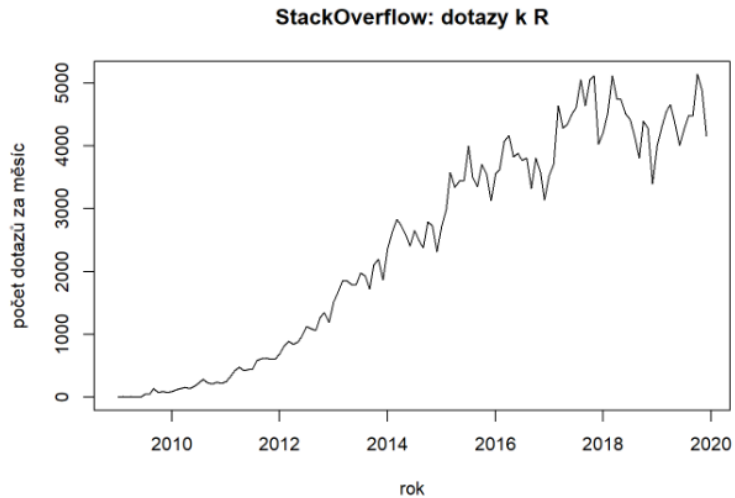
- **Popis dat**:
  - Použitý dataset dostupný zde.
  - Měsíční počty dotazů na StackOverflow od roku 2009 do roku 2020 pro různé programovací knihovny a technologie.

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

Dotazy k R a Pythonu



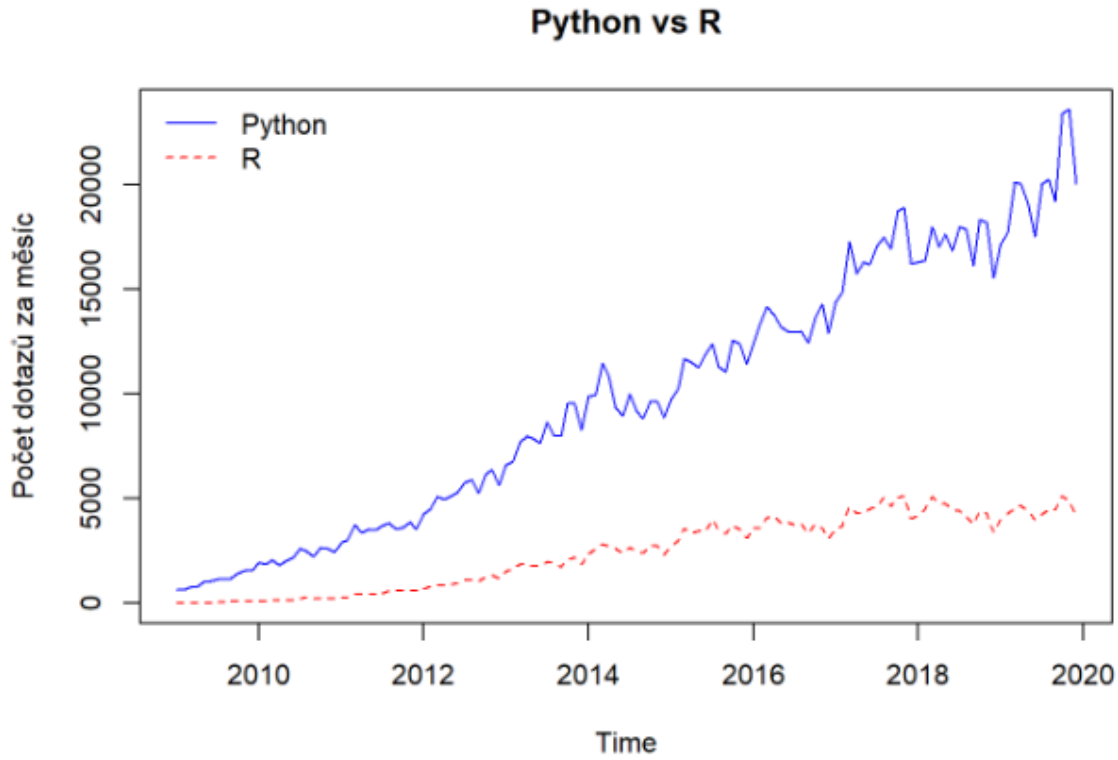
Dotazy k Pythonu



Dotazy k R

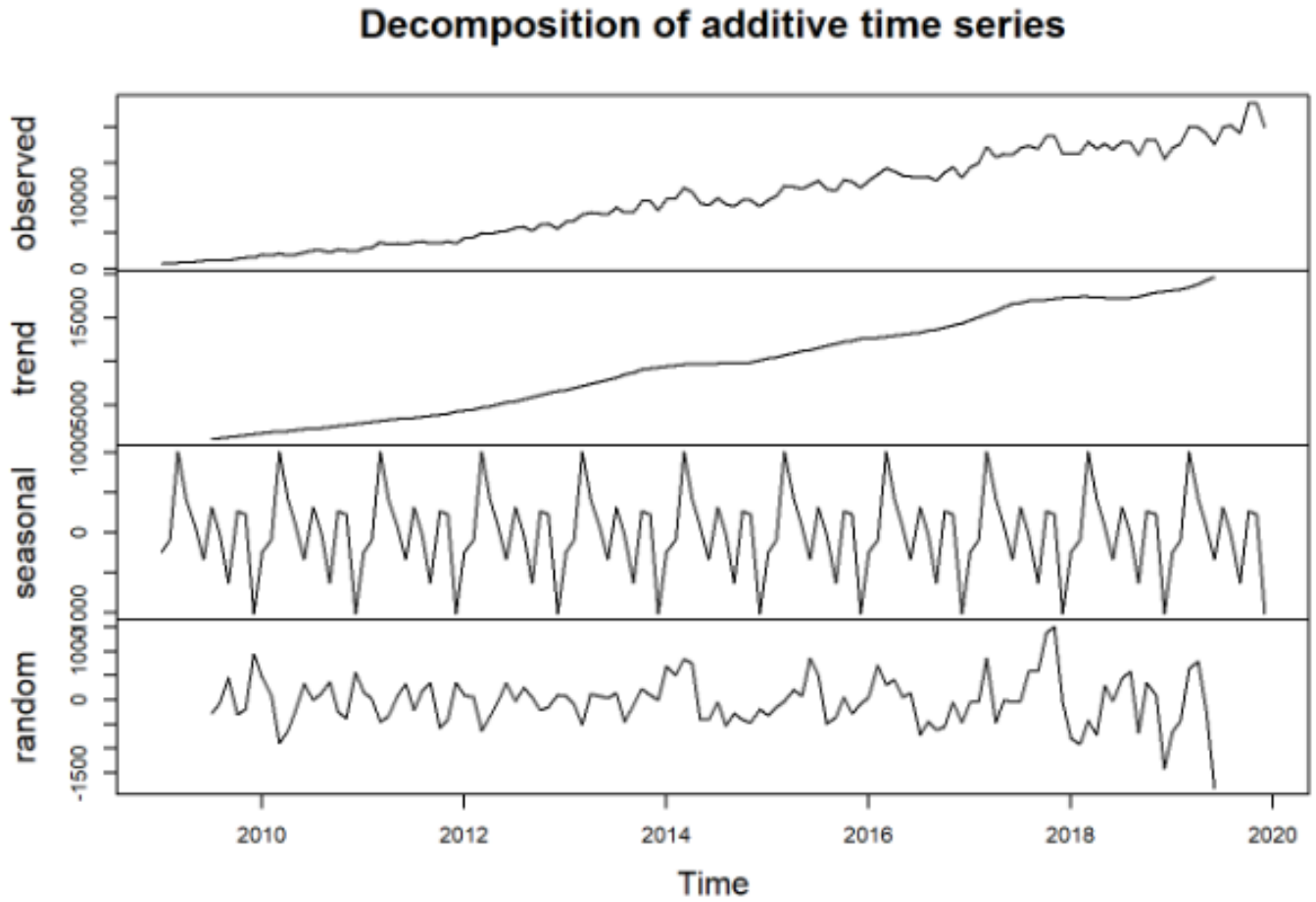
- Jasný rostoucí trend od roku 2009 do 2020.
- Postupem času narůstá i variabilita počtu dotazů.
- Viditelné pravidelné sezónní kolísání.
- **R:** Trend také roste, ale pomaleji a méně výrazně než u Pythonu.
- **R:** Amplituda sezónnosti je menší.
- **R:** Po roce 2017 se růst zpomaluje.

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

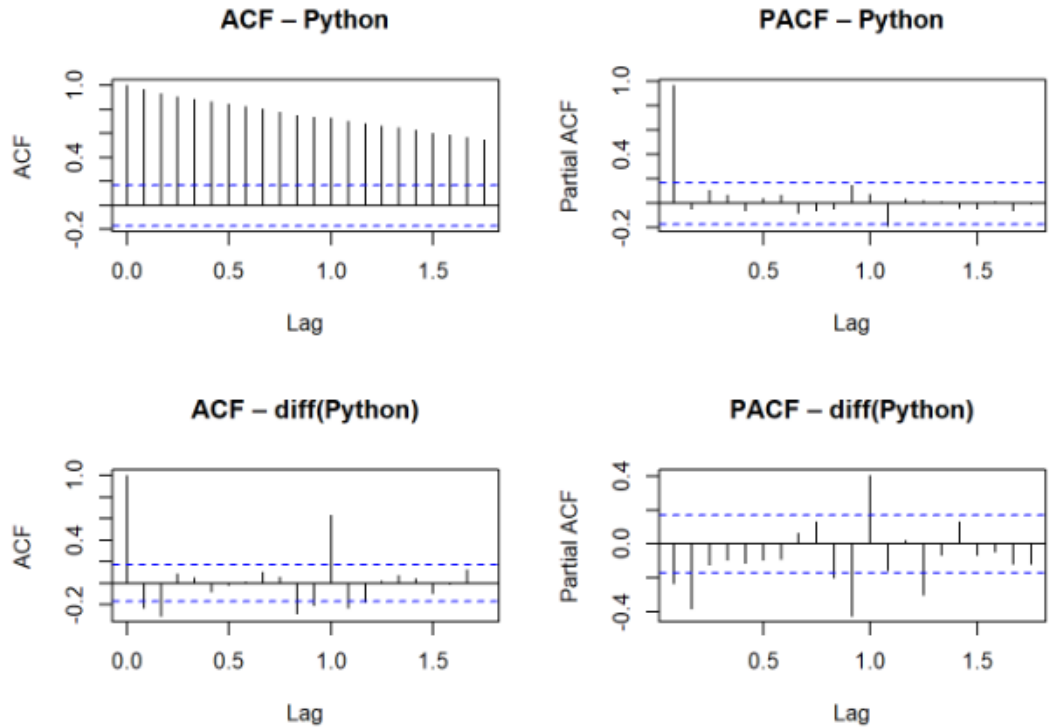


- **Trend:** silně rostoucí, plynulý.
- **Sezónnost:** pravidelná, opakující se ročně.
- **Remainder:** náhodné fluktuace bez jasné struktury.

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.

Manuální volba modelu: Identifikace

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

- **Transformace:** Logaritmus pro stabilizaci rozptylu.
- **Sezónnost** $(s = 12)$**:**
  - $D = 1$: Sezónní diference kvůli roční cykličnosti.
  - $P = 1$: PACF na lagu 12 naznačuje AR(1) v sezónní části.
- **Trend a nesezónní část:**
  - $d = 1$: První diference pro odstranění trendu.
  - $q = 2$: ACF ukazuje významné lagy 1 a 2.
  - $p = 1$: Přidáno pro robustnost (smíšený model).
- **Výchozí model:** $SARIMA(1, 1, 2)(1, 1, 0)_{12}$

Problém redundance parametrů

Po odhadu výchozího modelu jsme narazili na problém:

```
Series: python_log
ARIMA(1,1,2)(1,1,0)[12]

Coefficients:
          ar1      ma1     ma2      sar1
       0.9882  -0.9791  0.0140  -0.3541
s.e.   0.0264   0.1031  0.1013   0.0964
```

- Koeficienty $ar1 \approx 0.99$ a $ma1 \approx -0.98$ mají opačná znaménka a téměř stejnou velikost.
- **Interpretace:** Kořeny se vzájemně vykrátí. AR člen se snaží suplovat nedostatečné diferencování.
- **Řešení:** Nutno zvýšit řád diference na $d = 2$.

Manuální volba modelu: Respecifikace

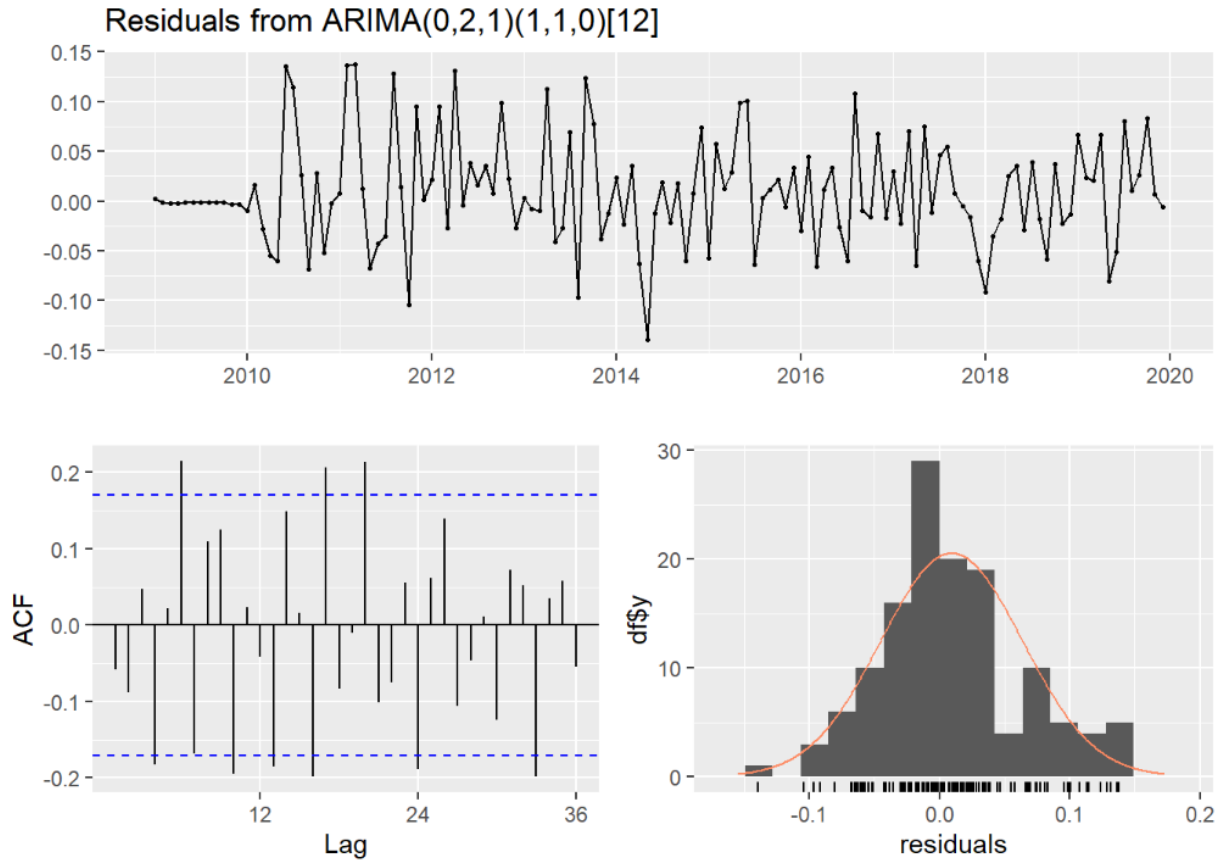
- **Nový model:** $SARIMA(0,2,2)(1,1,0)_{12}$ (bez AR členu, $d = 2$).
- Test významnosti koeficientů (t-test):

	Koeficient	t_statistika	Vyznamny
ma1	ma1	-9.8301991	TRUE
ma2	ma2	0.2512766	FALSE
sar1	sar1	-3.6594020	TRUE

- **Závěr:** Parametr $ma2$ je statisticky nevýznamný ($t < 1.96$).
- $\rightarrow$ Redukujeme model odstraněním $ma2$.

Sumarizace manuální volby modelu

Finální manuální model: $SARIMA(0,2,1)(1,1,0)_{12}$



- **Diagnostika:** Ljung-Boxův test zamítá nezávislost reziduí p-value $= 4.19 \times 10^{-6} < 0.05$.
- Rezidua vykazují autokorelaci, ale jde o nejlepší manuální výsledek.

Automatická volba modelu

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

$$SARIMA(0,2,2)(2,0,0)_{12}$$

Manuální model:

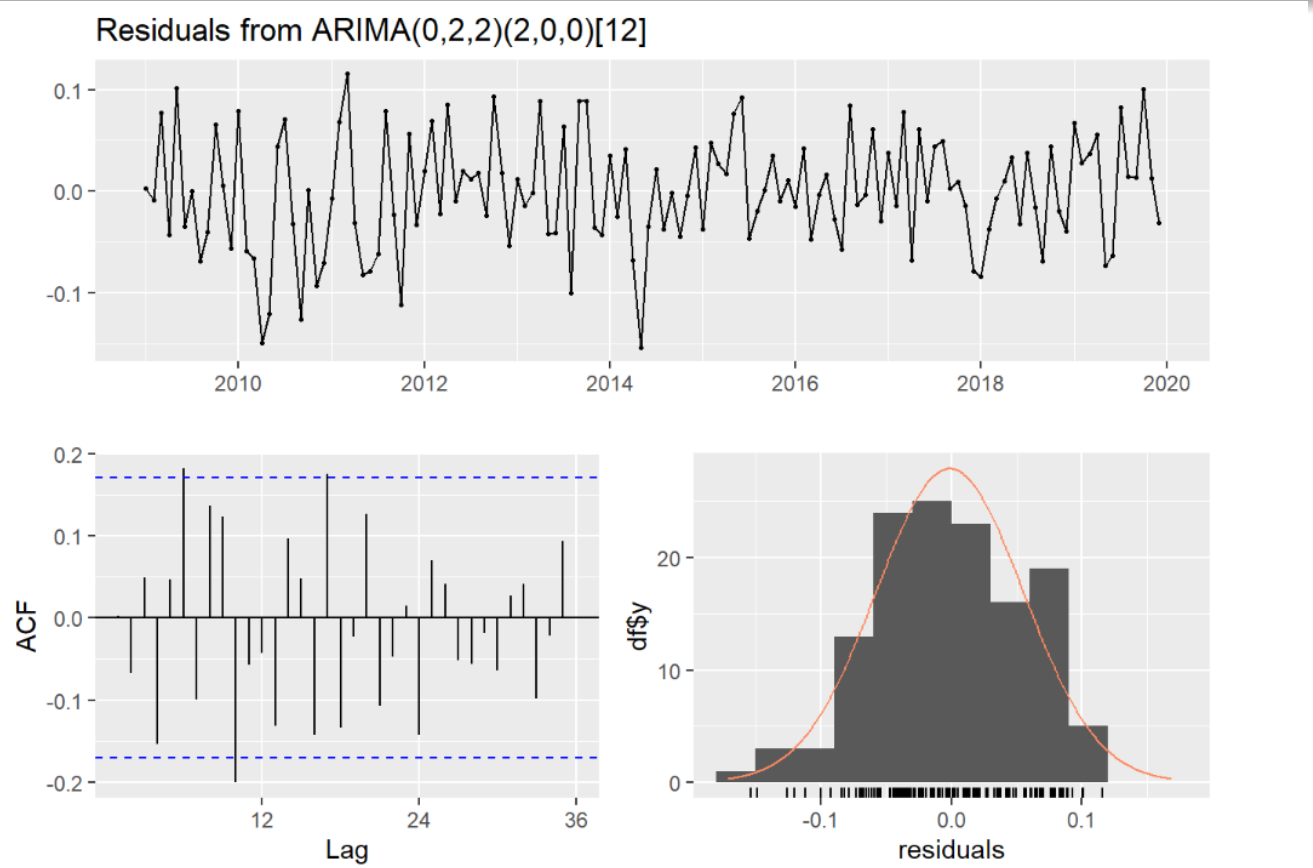
- $(0,2,1)(1,1,0)_{12}$
- AICc: **-319.15**

Auto ARIMA (Vítěz):

- $(0,2,2)(2,0,0)_{12}$
- AICc: **-351.24**

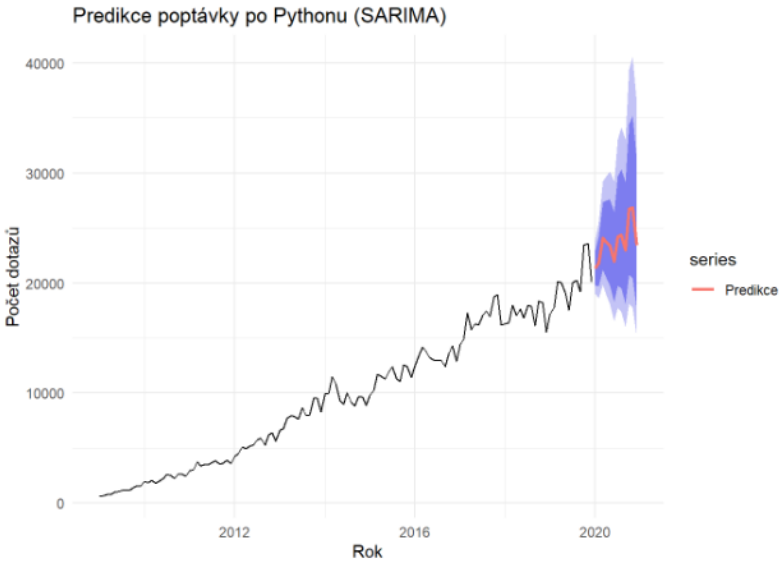
- Automatický model je výrazně lepší dle informačních kritérií.
- **Validace:** Oba přístupy potvrdily nutnost druhé diference ($d = 2$).

Diagnostika finálního modelu (Auto ARIMA)

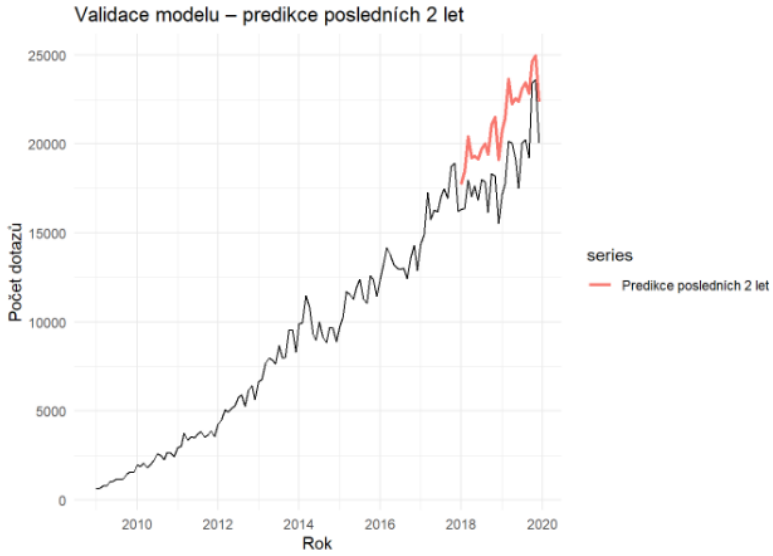


- **Ljung-Boxův test:** p-value $= 0.0011 < 0.05 \rightarrow$ zamítáme H_0.
- **Interpretace:** Model nezachytil veškerou strukturu (pravděpodobně kvůli strukturálnímu zlomu po r. 2017), ale pro predikci trendu je dostačující.

Predikce



Predikce poptávky na 12 měsíců dopředu



Validace: predikce posledních 2 let

Závěr

- Analyzovali jsme vývoj popularity jazyka Python na StackOverflow.
- Zjistili jsme silný rostoucí trend a pravidelnou sezónnost.
- Logaritmická transformace a diferencování pomohly data stacionarizovat.
- Ruční analýza koeficientů a porovnání s automatická selekceí identifikovaly jako nejvhodnější model SARIMA(0,2,2)(2,0,0)[12].
- Byl použit pro predikci, která očekává další růst zájmu s typickými sezónními výkyvy.

Diskuze

- **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 logaritmická 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!