stackoverflow

Time Series Project:

Understanding the Growth of Data Science StackOverflow Questions

Alana Willis, Clare Cruz, Dan Nason, & Megan Christy

April 21, 2022

Outline

● Data Description	3
● Research Questions	4
● Exploratory Data Analysis	5-8
● SARIMA	
○ Methods	9
○ Results	10-12
● VAR	
○ Methods	13
○ Results	14-17
● Conclusion	18
● Discussion	19
● Limitations	20

Data Description

- StackOverflow is a public question-forum for programming questions
- Raw counts of StackOverflow questions relating to 82 data science topics from Kaggle
  - R
  - Python
  - Machine learning
  - Classification
  - Regression
  - Clustering
  - Time Series
- Monthly data from January 2011 to December 2019

Research Questions

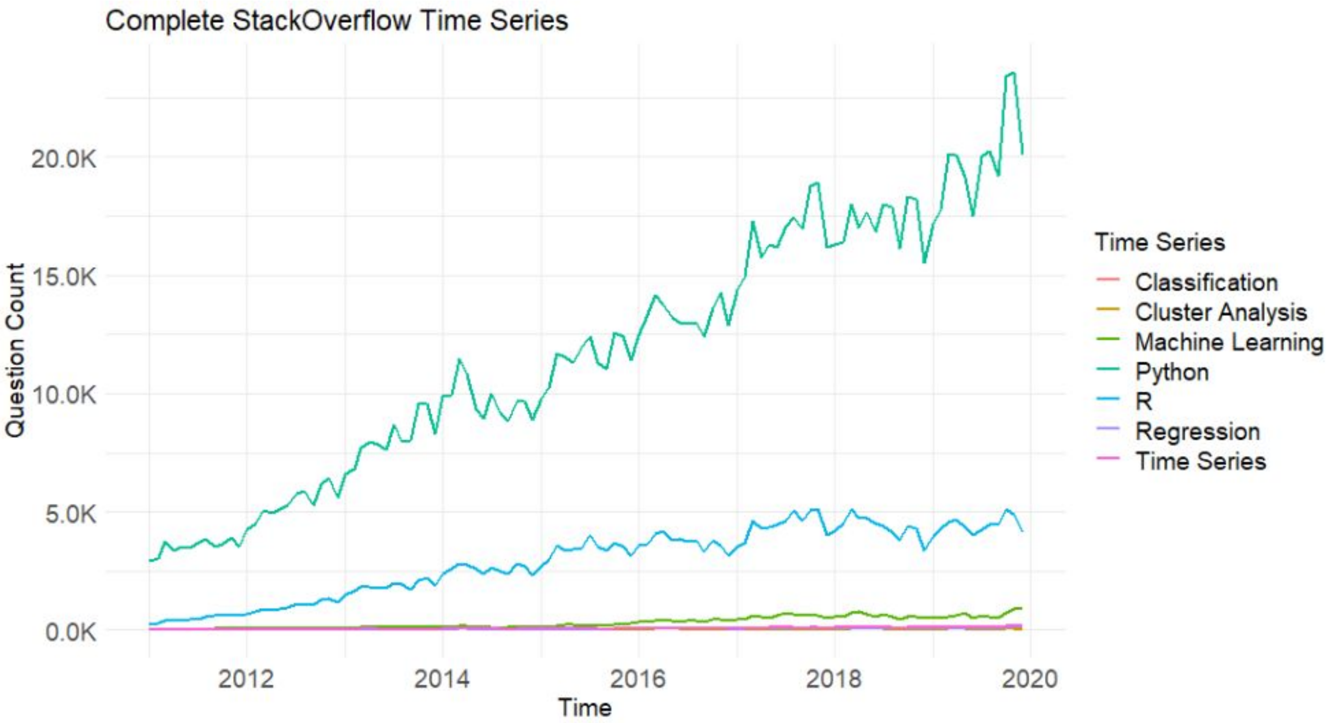
Prediction

Which data science tools between R and Python have the highest predicted growth rate from 2019 to 2021?

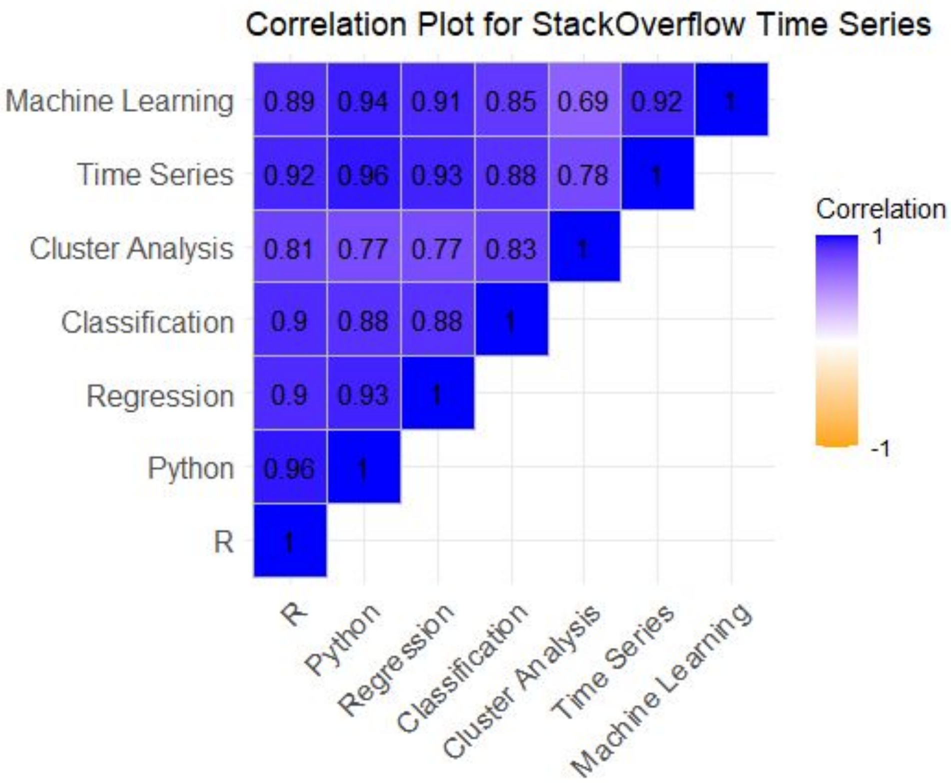
Inference

Which data science topics significantly contributed to the question count for R and Python?

Python has the most StackOverflow questions per month, followed by R



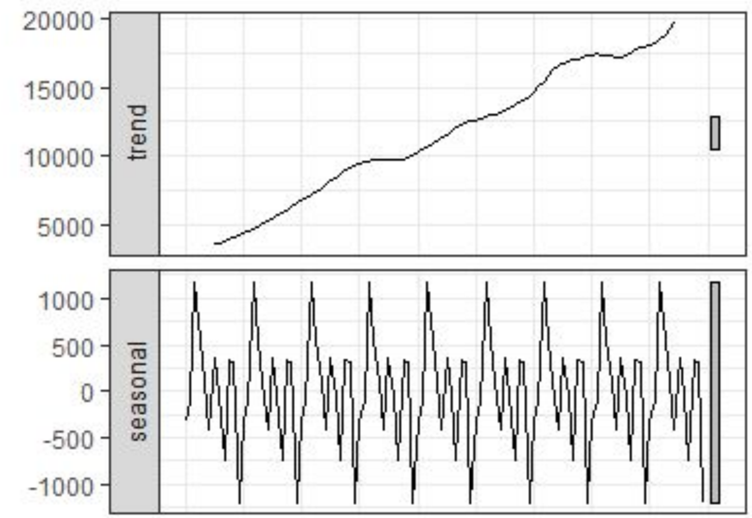
There is a strong correlation between all the time series of interest



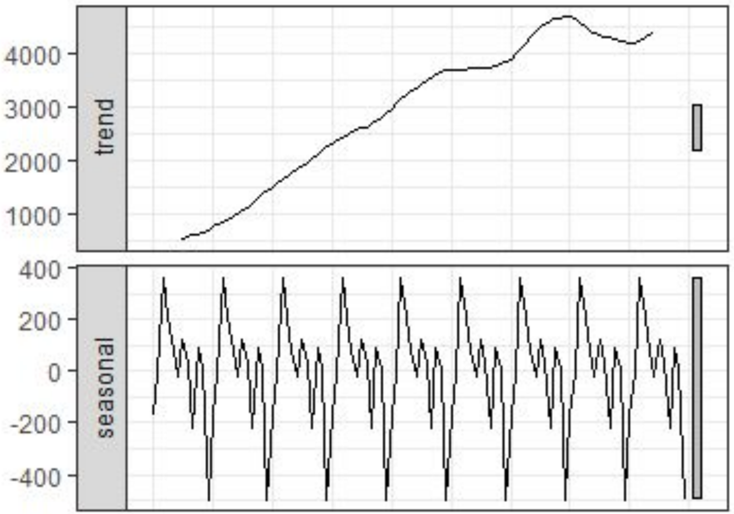
	R	Python	Regression	Classification	Cluster Analysis	Time Series	Machine Learning
Machine Learning	0.89	0.94	0.91	0.85	0.69	0.92	1
Time Series	0.92	0.96	0.93	0.88	0.78	1	
Cluster Analysis	0.81	0.77	0.77	0.83	1		
Classification	0.9	0.88	0.88	1			
Regression	0.9	0.93	1				
Python	0.96	1					
R	1						

There is a trend and a seasonal effect in both the Python and R time series

Decomposition of Python Series

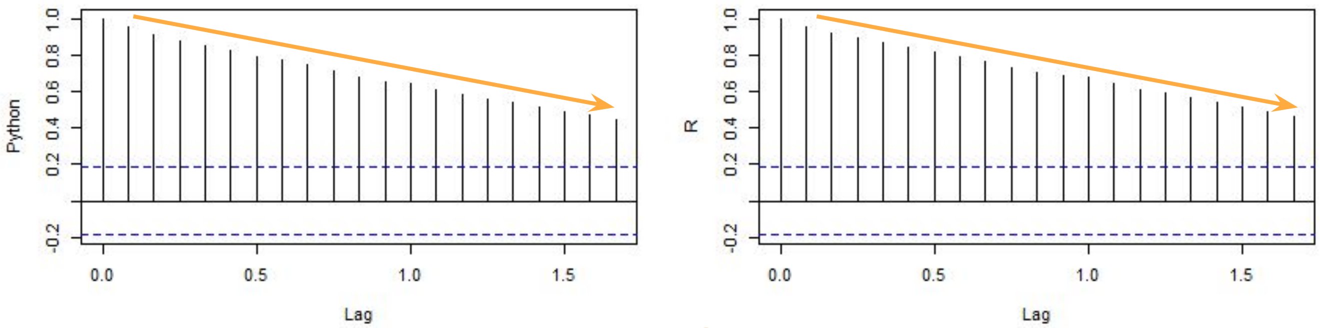


Decomposition on R Series

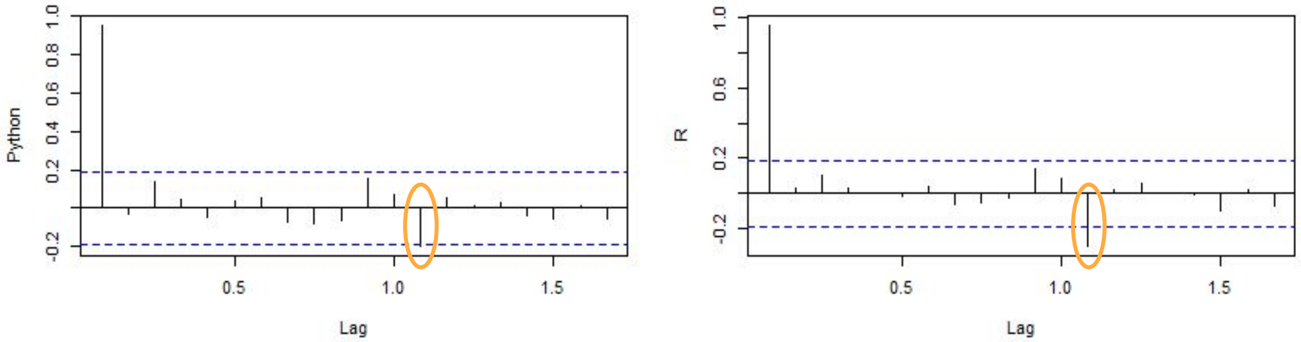


ACF/PACF plots suggest an AR process with seasonal effects

ACF



PACF



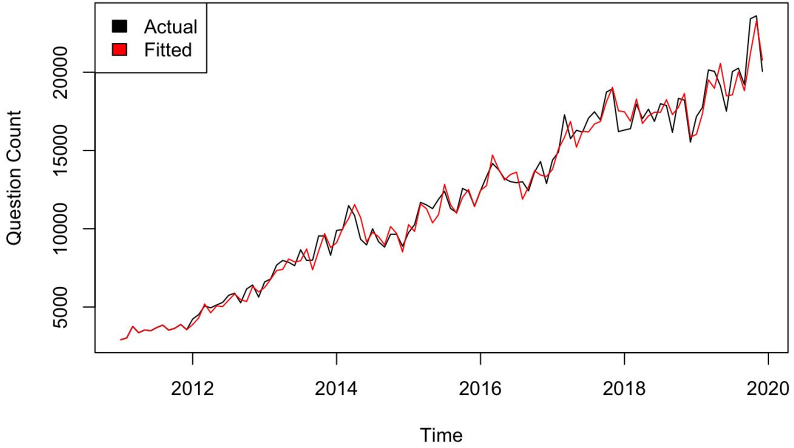
Methods for Modeling (SARIMA)

It's clear that our time series have complex behaviors, so SARIMA models were fit to the R and Python series to gain a better understanding of our data:

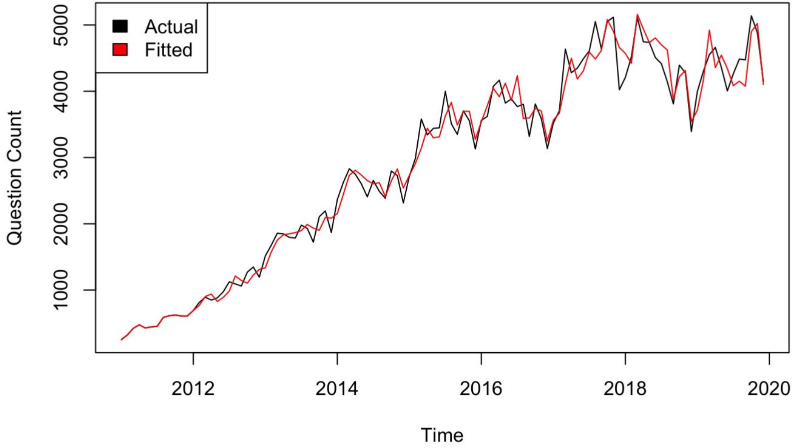
1. **Manually fit a SARIMA**(1,0,0) (1,1,0) [12] for Python and R based on ACF and PACF plots
2. **Verify the initial model** using auto.arima with varying parameters
   a. For Python, auto.arima found the same model we originally fit
   b. For R, auto.arima found an SARIMA(1,1,0) (0,1,1) [12] model
3. **Check model diagnostic plots** and forecasts
   a. Auto.arima model chosen over our original model due to lower AIC and RMSE/MAE

Plots illustrate that the SARIMA models fit the data relatively well

Observed vs. Fitted SARIMA(1,0,0)(1,1,0)[12] values for Python

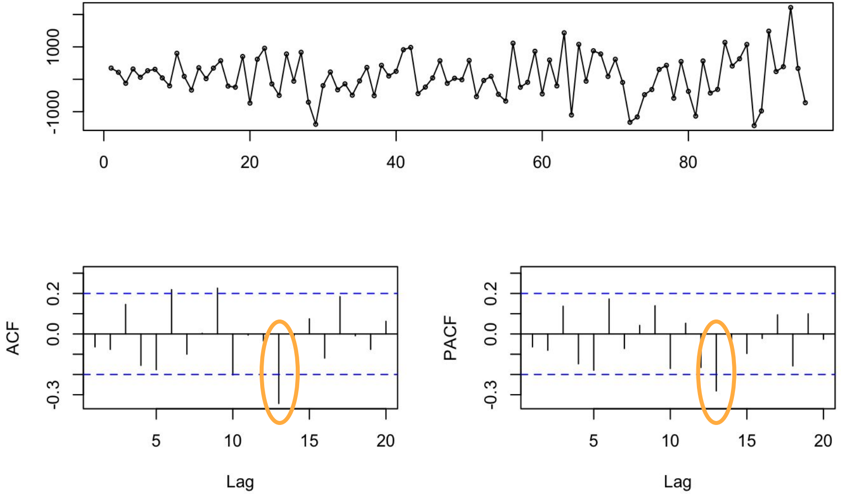


Observed vs. Fitted SARIMA(1,1,0)(0,1,1)[12] values for R

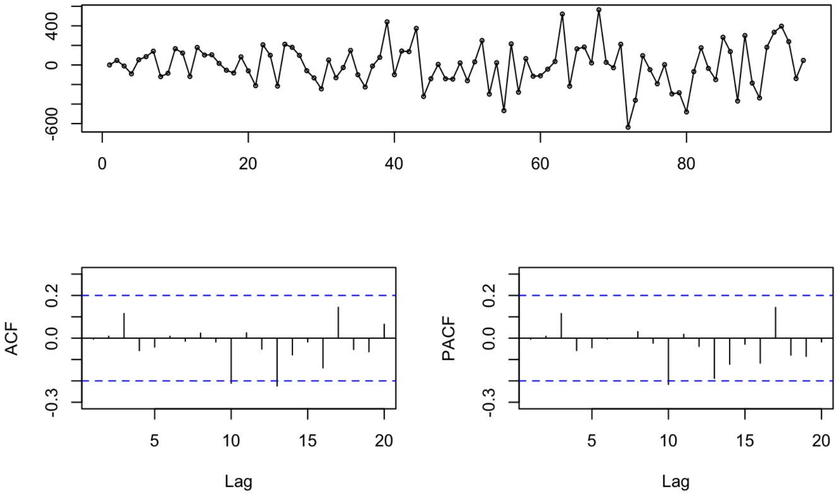


Model diagnostics suggest that the residuals are white noise, but a seasonal pattern remains present in Python

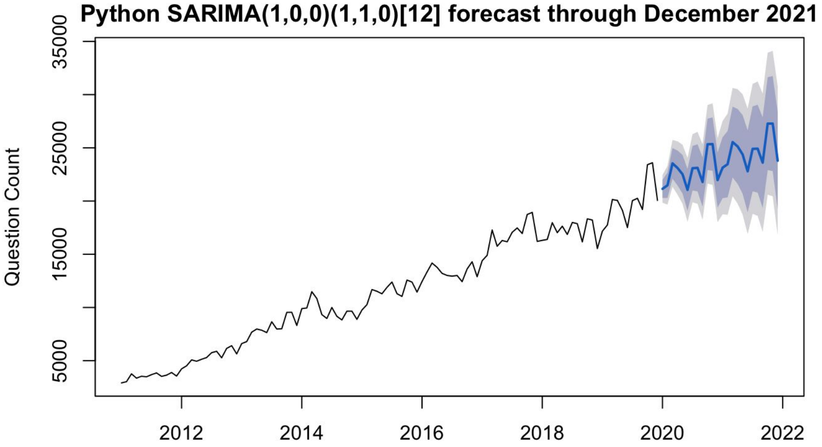
Python Residuals

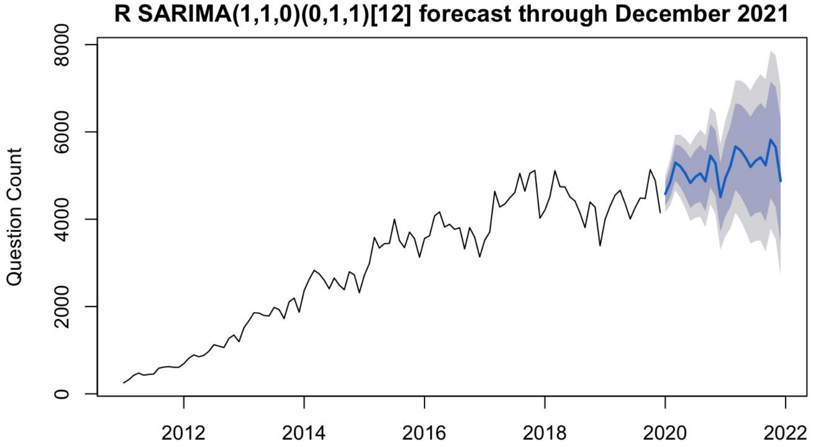


R Residuals



Python & R forecasts suggest an increasing trend consistent with prior data



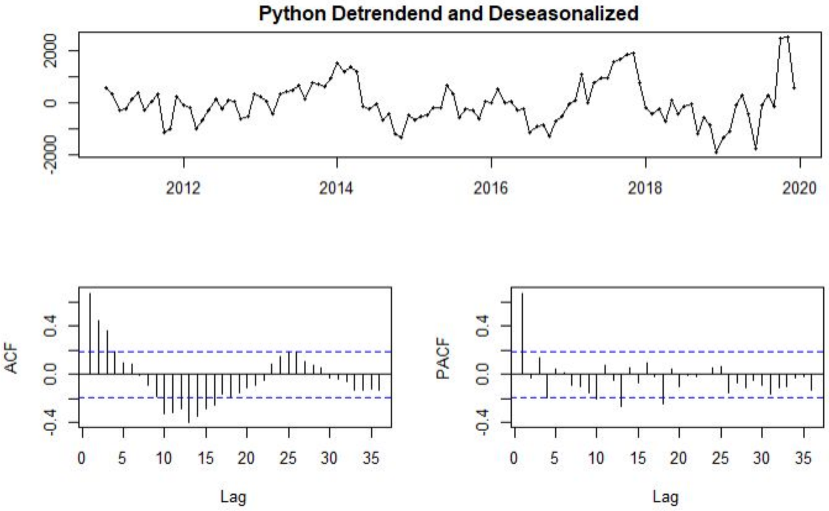


Methods for Modeling (VAR)

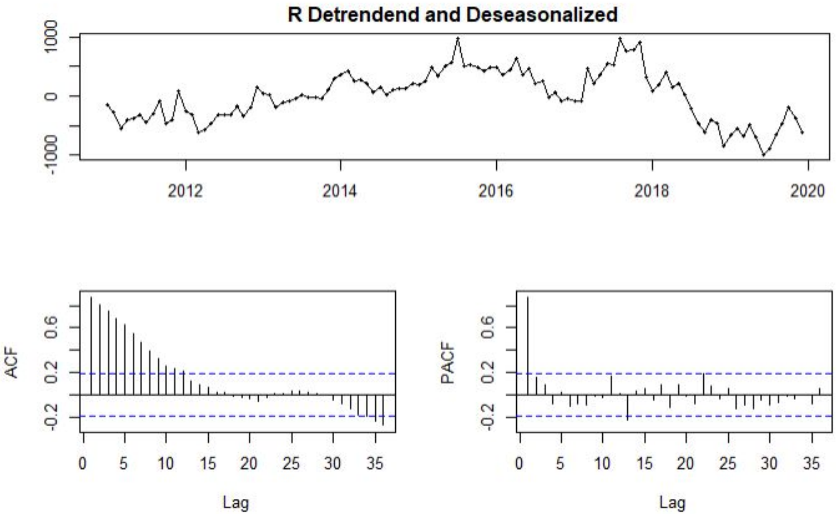
A VAR model was fitted to the data to answer our research questions:

1. **Detrend and deseasonalize** the Python and R series based on findings from the SARIMA model and in the exploratory data analysis
2. **Create two VAR models** using the VARselect function to select an appropriate order for the Python and R series
3. **Check model diagnostic** plots and forecasts
4. **Evaluate the summary output** for the VAR models
5. **Forecast** question counts for the next two years in the original data scale

After transforming the data, plots suggest Python and R follow an AR process



Python Detrendend and Deseasonalized



R Detrendend and Deseasonalized

The R model has more significant predictors and a higher adjusted R^2

Python		
Variable (Lag 1)	***Coefficient Est.***	***Std. Error***
Python	0.67***	0.07
Machine Learning	0.39	0.69
Classification	4.87	6.86
Regression	-5.43	5.81
Time Series	0.24	4.29
Cluster Analysis	-2.65	7.35

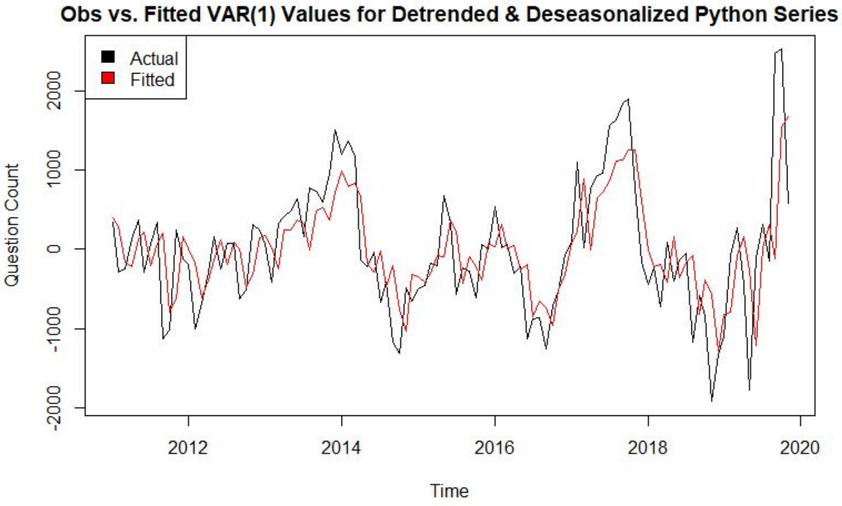
Adj. R^2: 0.44

R		
Variable (Lag 1)	***Coefficient Est.***	***Std. Error***
R	0.87***	0.05
Machine Learning	-0.24	0.21
Classification	7.79***	2.15
Regression	-2.57	1.78
Time Series	0.73	1.36
Cluster Analysis	-6.18***	2.25

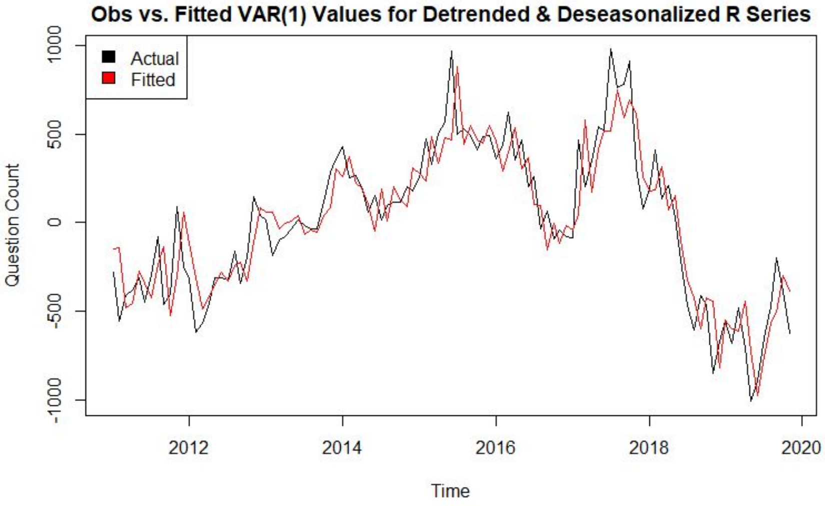
Adj. R^2: 0.81

*** indicates significance at $\alpha < .05$

Plots illustrate that the VAR models also fit the data relatively well

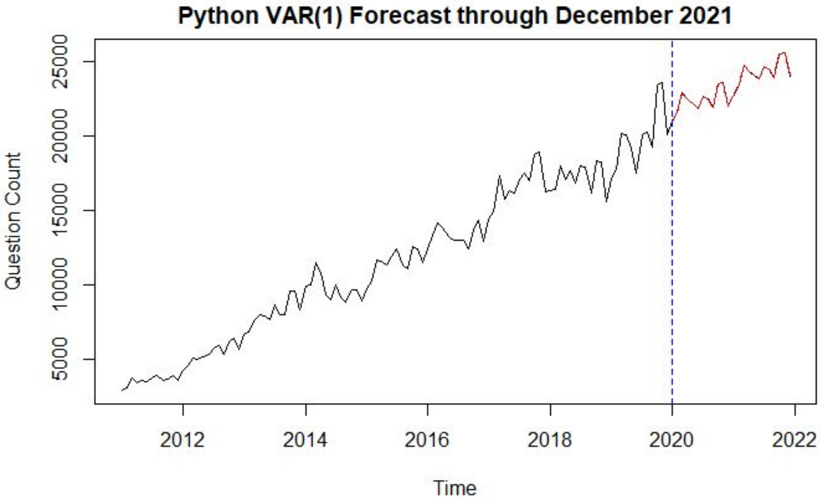


Obs vs. Fitted VAR(1) Values for Detrended & Deseasonalized Python Series

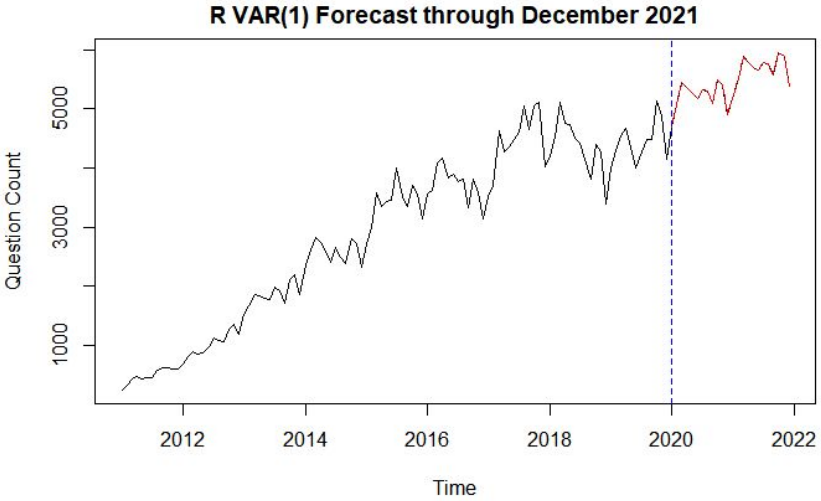


Obs vs. Fitted VAR(1) Values for Detrended & Deseasonalized R Series

Forecasts suggest an increasing trend consistent with prior data



Python VAR(1) Forecast through December 2021



R VAR(1) Forecast through December 2021

Conclusions

Prediction

- The Python model forecasts the number of questions to grow from 20,058 in December 2019 to 24,001 in December 2021, a **19.7%** growth rate
- The R model forecasts the number of questions to grow from 4,150 in December 2019 to 5,380 in December 2021, a **29.7%** growth rate

Inference

- None of the topics significantly contributed to predicting the Python series
- Classification and Clustering were the only topics that significantly contributed to predicting the R series
- R overall has a better fit with machine learning topics than Python

Discussion

- R is primarily used for statistical modeling while Python has many other uses (i.e. software engineering), which explains the differences in:
  - Significant predictors
  - Adjusted R^2
  - Growth rate
- Unsure why clustering and classification are the only significant predictors

Limitations

- R vs. Python may not be a reasonable comparison because R is used for statistical modeling while Python has many other uses
- Limited selection of topics
- Better features may exist
  - Chosen by intuition/research question
- Feature redundancy
  - Double counting likely present in data
  - e.g.) Python vs. Python 3.0
- May be losing important patterns in the data due to monthly aggregation

Next Steps

- Extend time frame of analysis
- Verify results of forecast with new data
- Include more predictors in the models
- Increase frequency of the data (daily, weekly, etc.)
- Add indicator variables for whether the semester is in session

Thank You!

Questions?