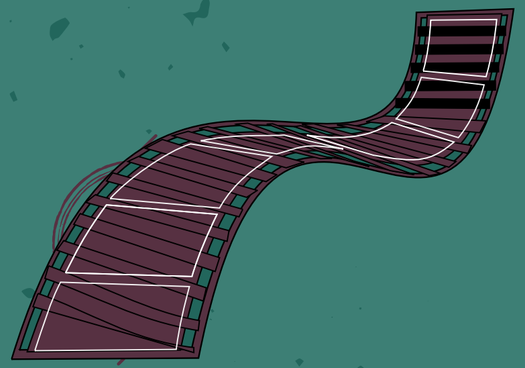


Analysis of Visitors in Italian Museums

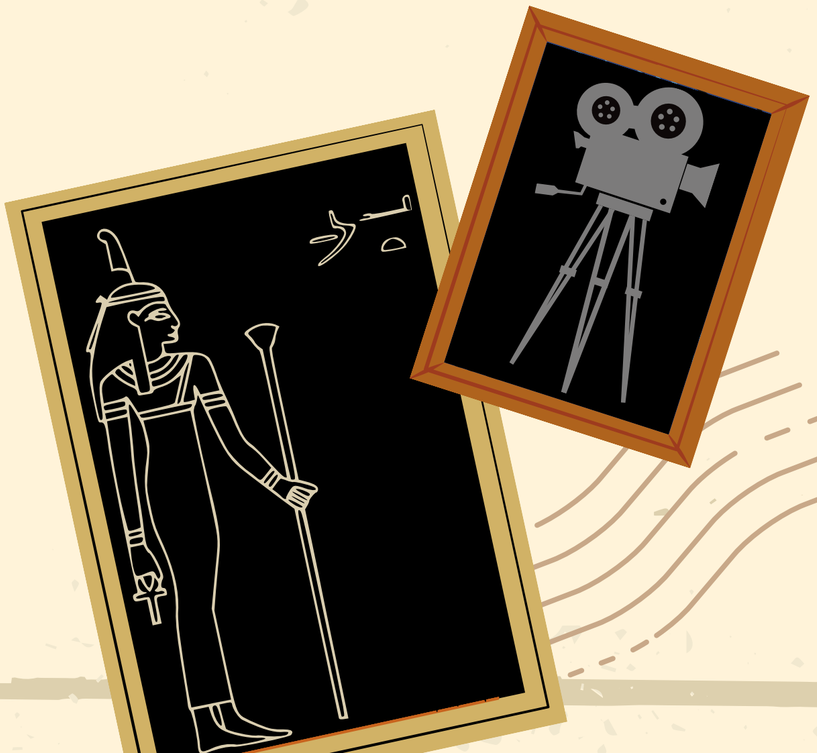
A TIME SERIES PERSPECTIVE

Graziana Capurso
Anna Cerbaro
Dejan Dichoski

12th January 2023

TABLE OF Contents

PROJECT *Overview*



01 Objective

Predict monthly visitors for two Italian museums by leveraging time series forecasting techniques.

- Explore the potential of incorporating **external variables** such as **Google Trends** and others, to enhance prediction accuracy.

02 Applications

- Individuals: **Plan your travel intelligently**.
- Optimize marketing efforts and **promotions** (ticket pricing).
- **Venue Management**: Improved queue control, staff allocation, and resource management.

Inspiration

Botta et al. *EPJ Data Science* (2020) 9:14
https://doi.org/10.1140/epjds/s13688-020-00232-z

EPJ.org

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

In search of art: rapid estimates of gallery and museum visits using Google Trends

Federico Botta[1,2*], Tobias Preis[1,3] and Helen Susannah Moat[1,3]

https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-020-00232-z

Dataset

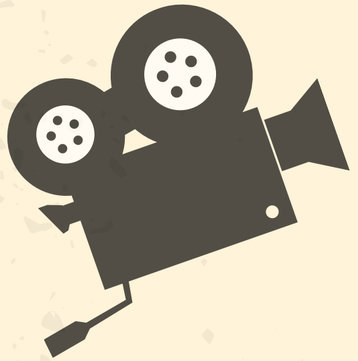
Visitors	→ Target
Date	
Google trends	} Continuous
Tourist arrivals	
Average temperature	
Raining days	
School holidays	
COVID closures	Binary
Renovation	

iL Meteo

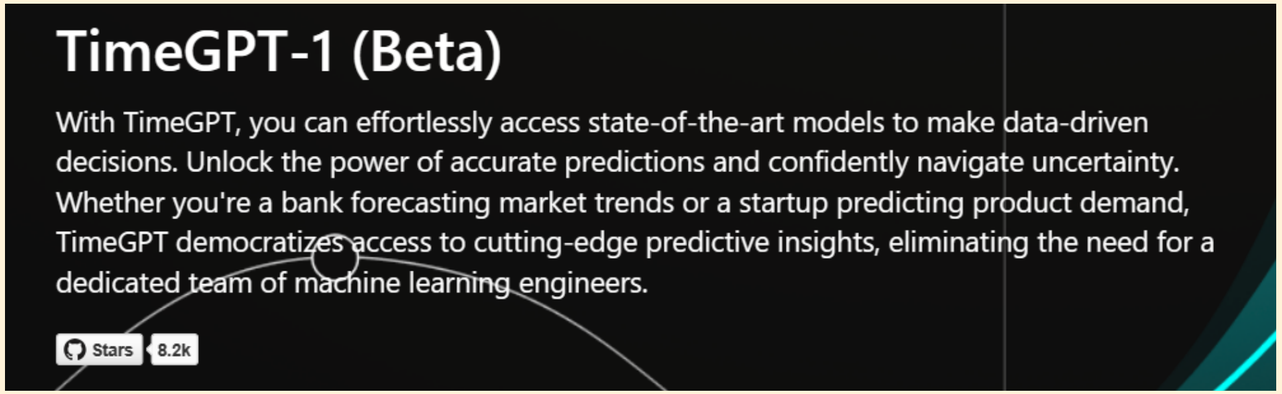
visit Piemonte
REGIONAL MARKETING AND PROMOTION

aperTO
Gli Open Data della Città di Torino

Baseline models

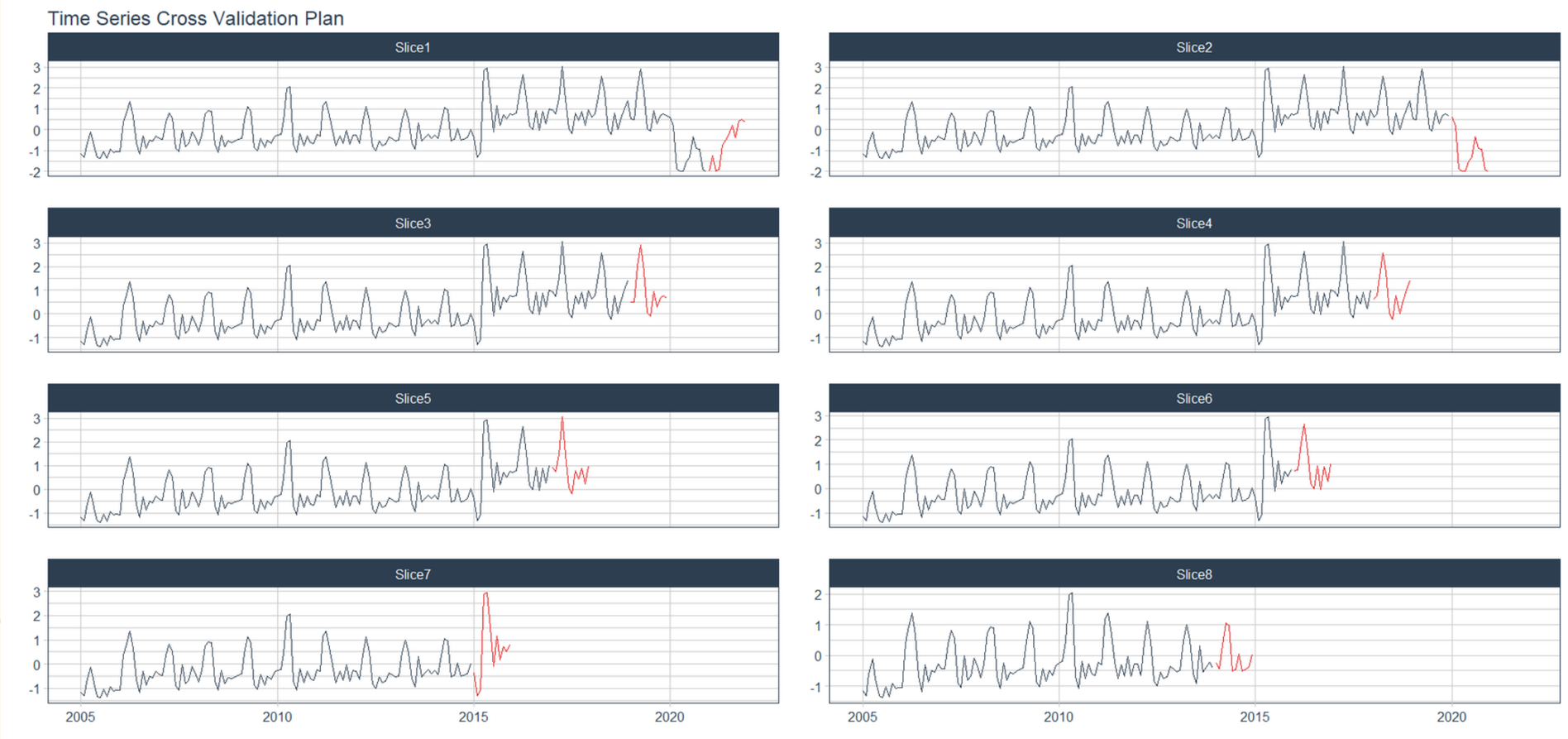


1. **Basic Baseline**: Predicts using the **mean** OR the **same value as the last year**, serving as a simple benchmark.
2. **Advanced Baseline** (SoTA): Utilizes **TimeGPT**, representing the cutting-edge model for time series forecasting.



NIXTLA

Time series CV

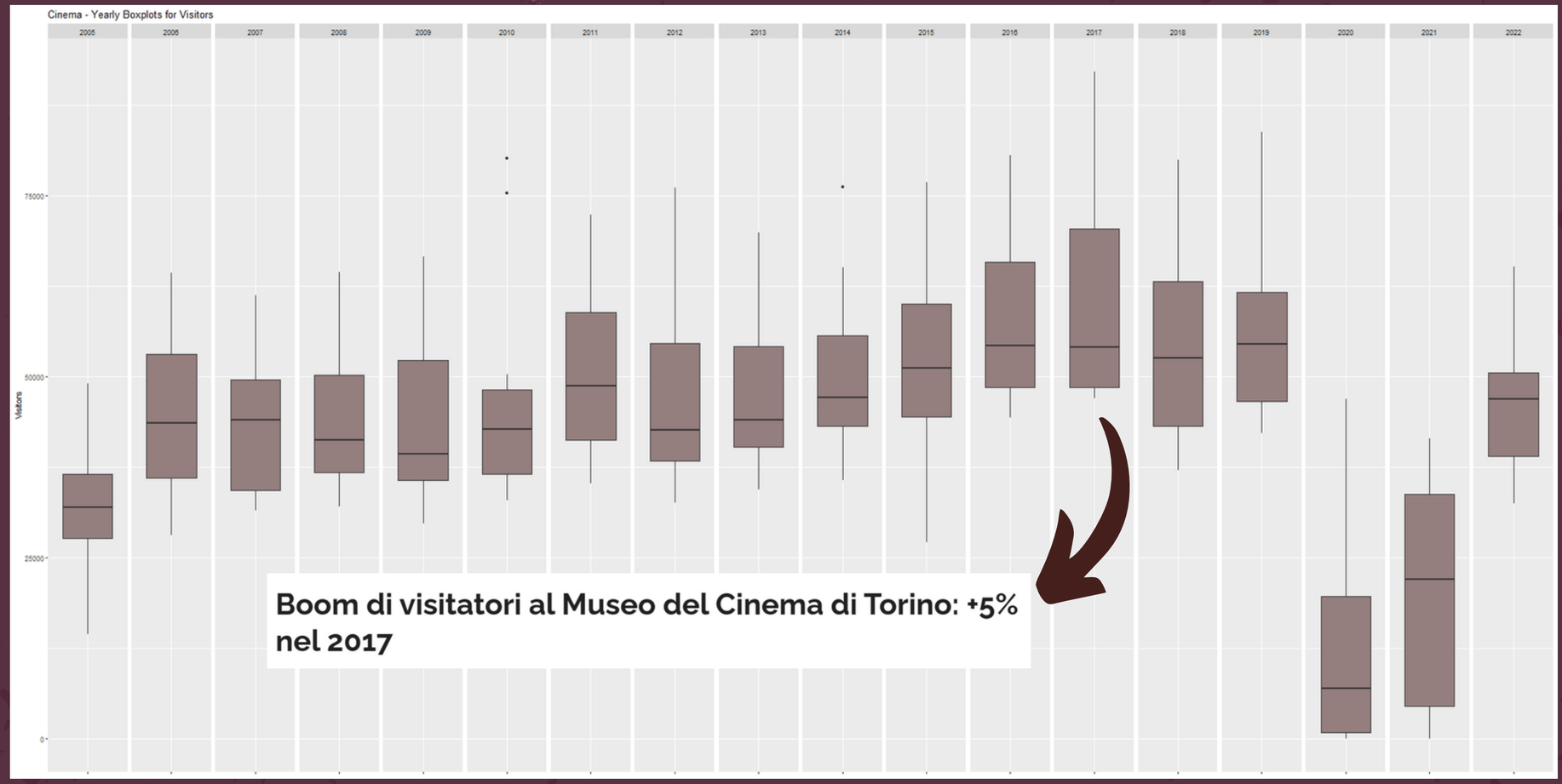


National Museum of Cinema

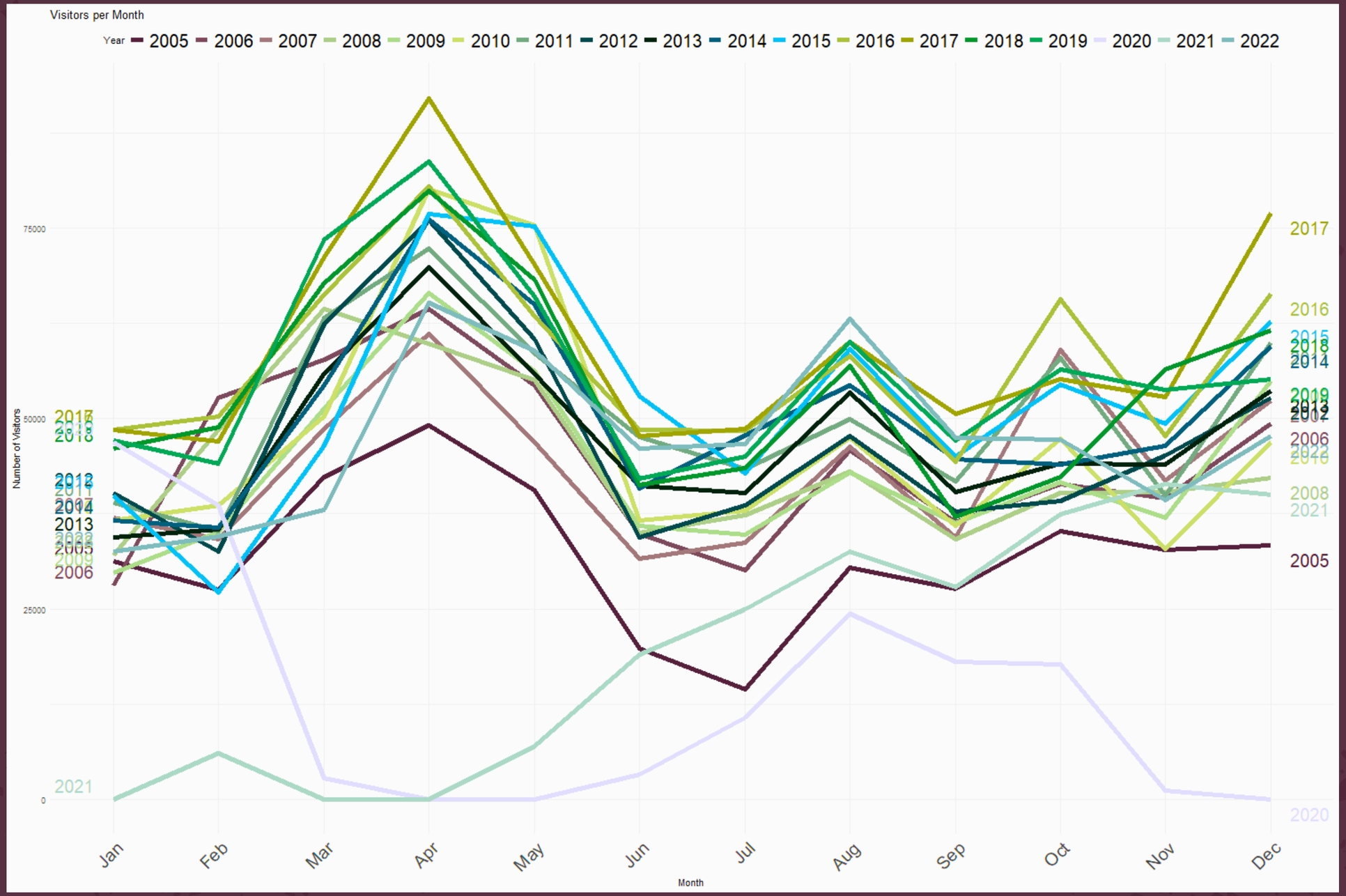
MUSEO NAZIONALE DEL CINEMA TORINO

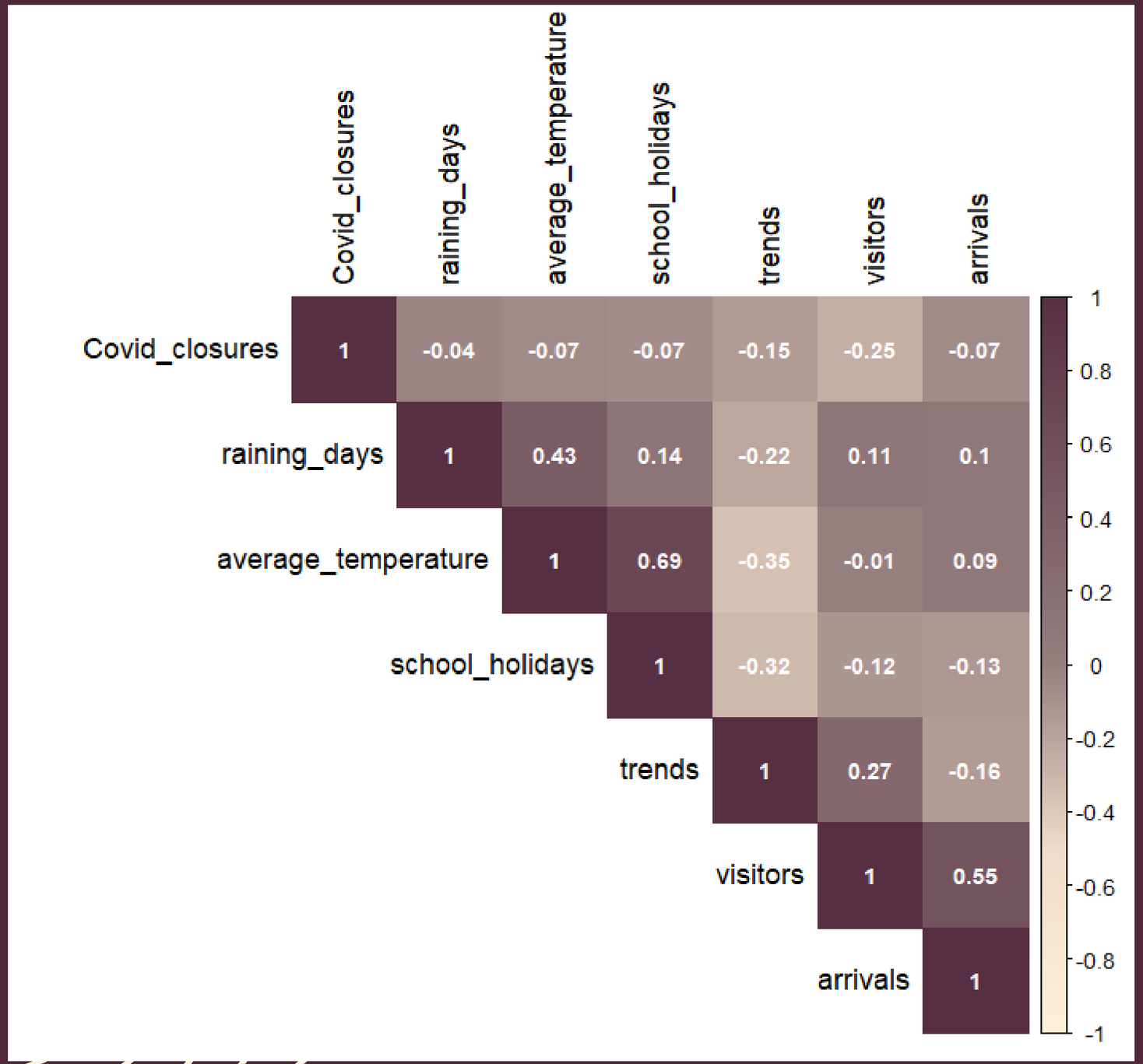
DATA COLLECTED FROM: https://ocp.piemonte.it

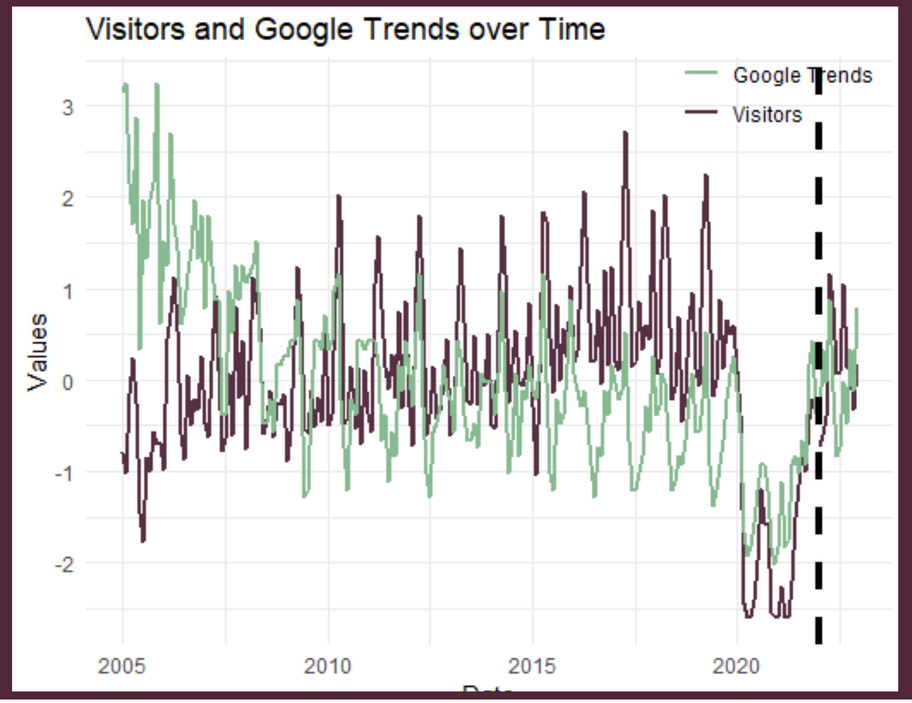
EDA

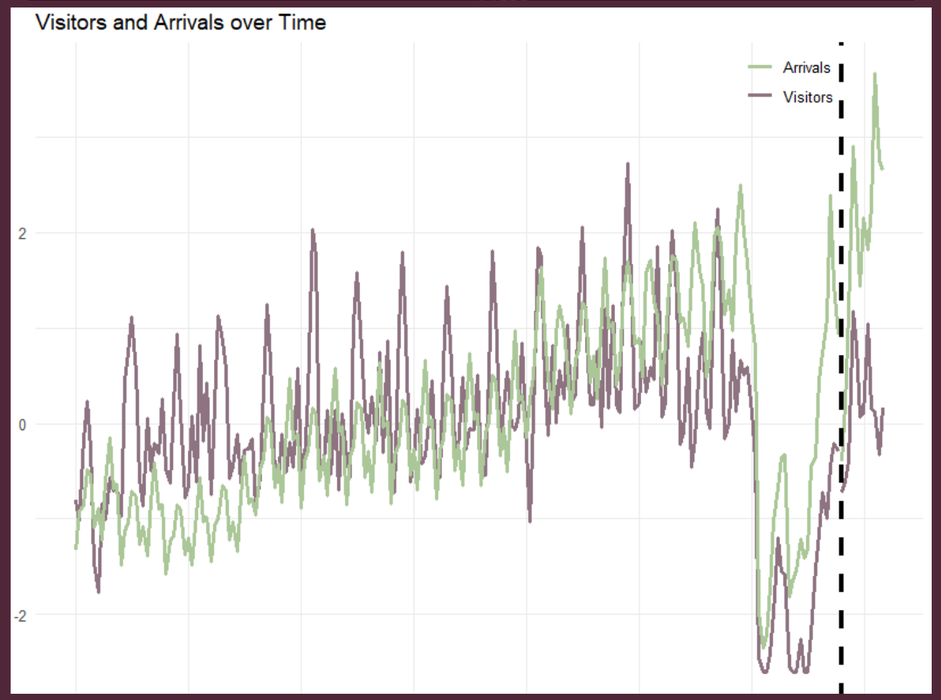


Yearly Boxplots for Visitors


Visitors per Month
Year 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
Number of Visitors
75000
50000
25000
0
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
2017
2016
2015
2018
2014
2019
2007
2006
2010
2008
2021
2005
2020

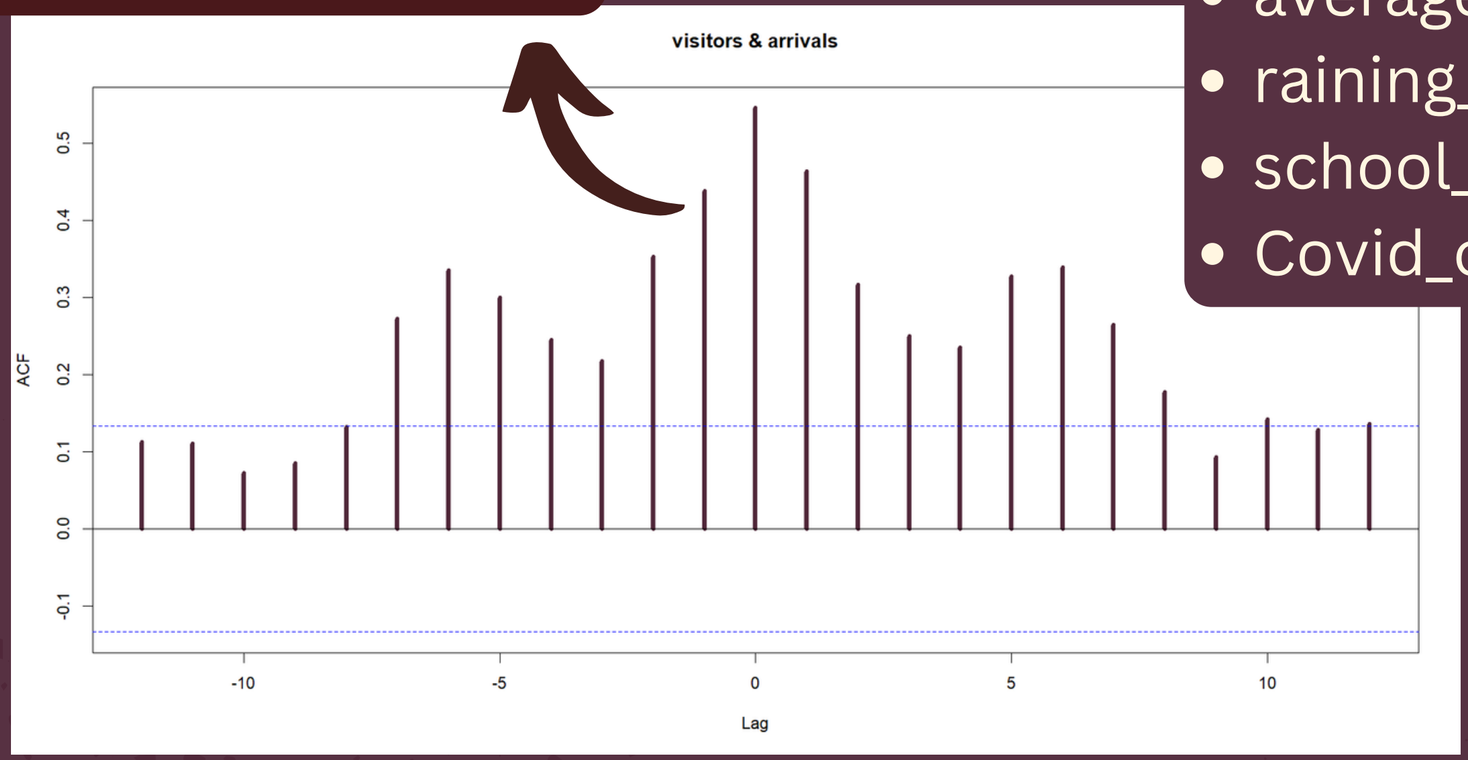

Covid_closures
raining_days
average_temperature
school_holidays
trends
visitors
arrivals
Covid_closures 1 -0.04 -0.07 -0.07 -0.15 -0.25 -0.07
raining_days 1 0.43 0.14 -0.22 0.11 0.1
average_temperature 1 0.69 -0.35 -0.01 0.09
school_holidays 1 -0.32 -0.12 -0.13
trends 1 0.27 -0.16
visitors 1 0.55
arrivals 1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Visitors and Google Trends over Time
Google Trends
Visitors
Values
3
2
1
0
-1
-2
2005
2010
2015
2020


Visitors and Arrivals over Time
Arrivals
Visitors
2
0
-2

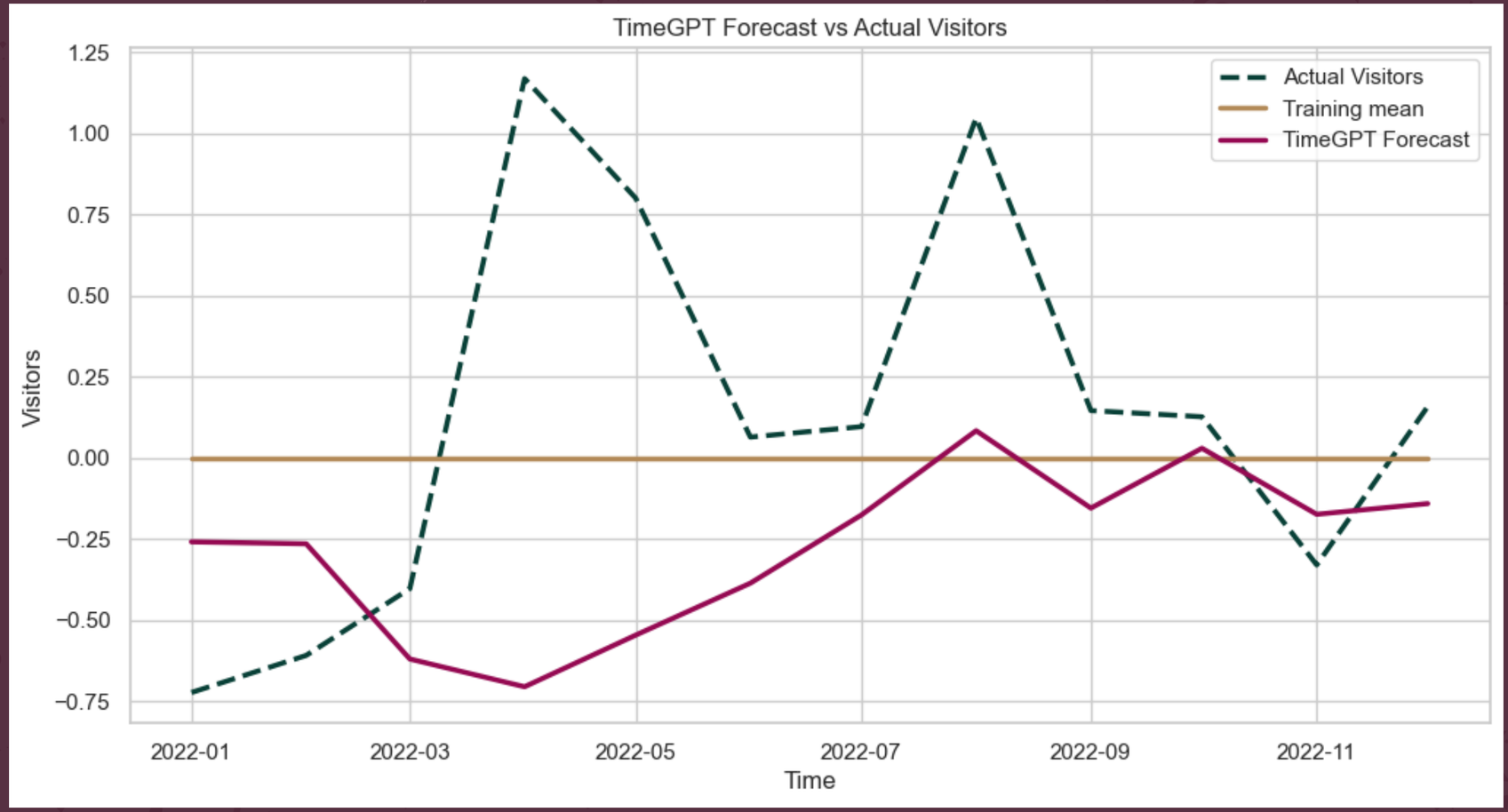
Lagged variables

visitors[t] ~ arrivals[t-1]



- google_trends[t-3]
- average_temperature[t-3]
- raining_days[t-1]
- school_holidays[t-3]
- Covid_closures[t-1]

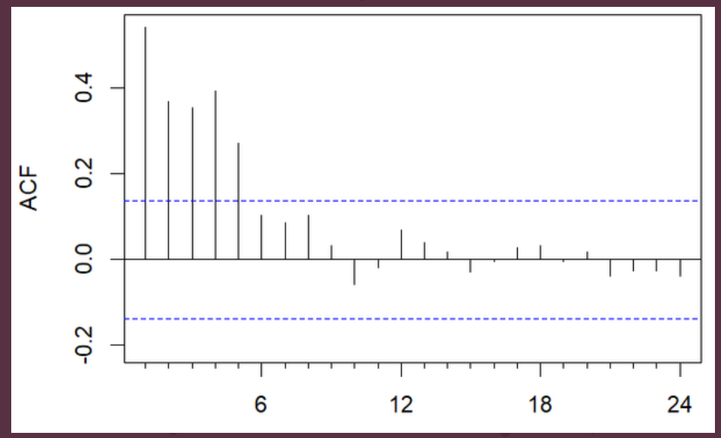
Baselines

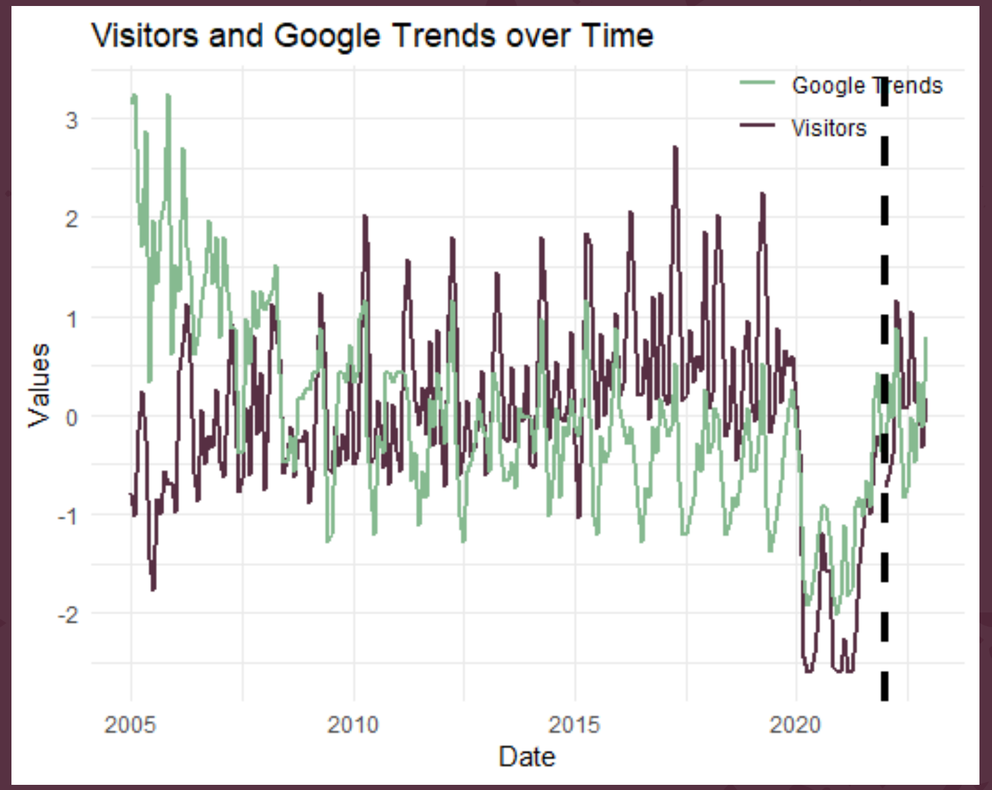

TimeGPT Forecast vs Actual Visitors
Actual Visitors
Training mean
TimeGPT Forecast
Visitors
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
2022-01
2022-03
2022-05
2022-07
2022-09
2022-11
Time

TSLM

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.295638   0.178720   7.250 1.04e-11 ***
google_trends -0.131613   0.060819  -2.164  0.03172 *
arrivals       1.009978   0.045963  21.974  < 2e-16 ***
trend         -0.011853   0.001114 -10.636  < 2e-16 ***
season2       -0.320944   0.159375  -2.014  0.04545 *
season3        0.477704   0.159451   2.996  0.00310 **
season4        0.868805   0.161798   5.370 2.30e-07 ***
season5        0.214780   0.164641   1.305  0.19364
season6       -0.358607   0.170656  -2.101  0.03694 *
season7       -0.551964   0.171685  -3.215  0.00153 **
season8        0.475880   0.160921   2.957  0.00350 **
season9       -0.835009   0.167911  -4.973 1.48e-06 ***
season10      -0.788912   0.169493  -4.655 6.10e-06 ***
season11      -0.387321   0.161297  -2.401  0.01731 *
season12       0.237504   0.160821   1.477  0.14139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.462 on 189 degrees of freedom
Multiple R-squared:  0.8012,    Adjusted R-squared:  0.7865
F-statistic: 54.42 on 14 and 189 DF,  p-value: < 2.2e-16
```

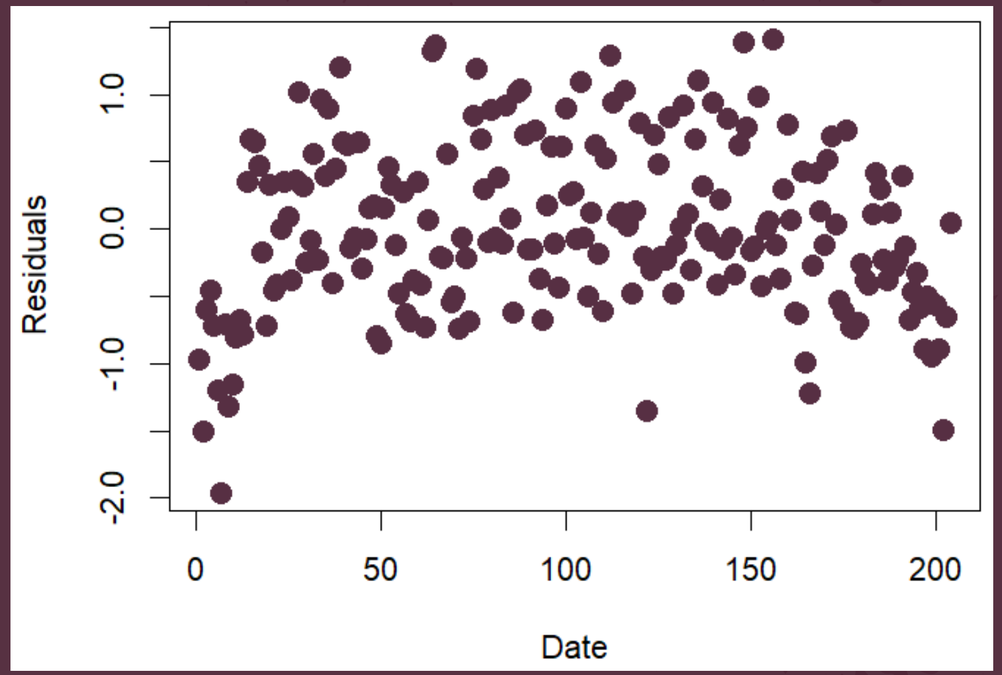




L1/L2 Regularized Regression

Top performer: **RIDGE**.

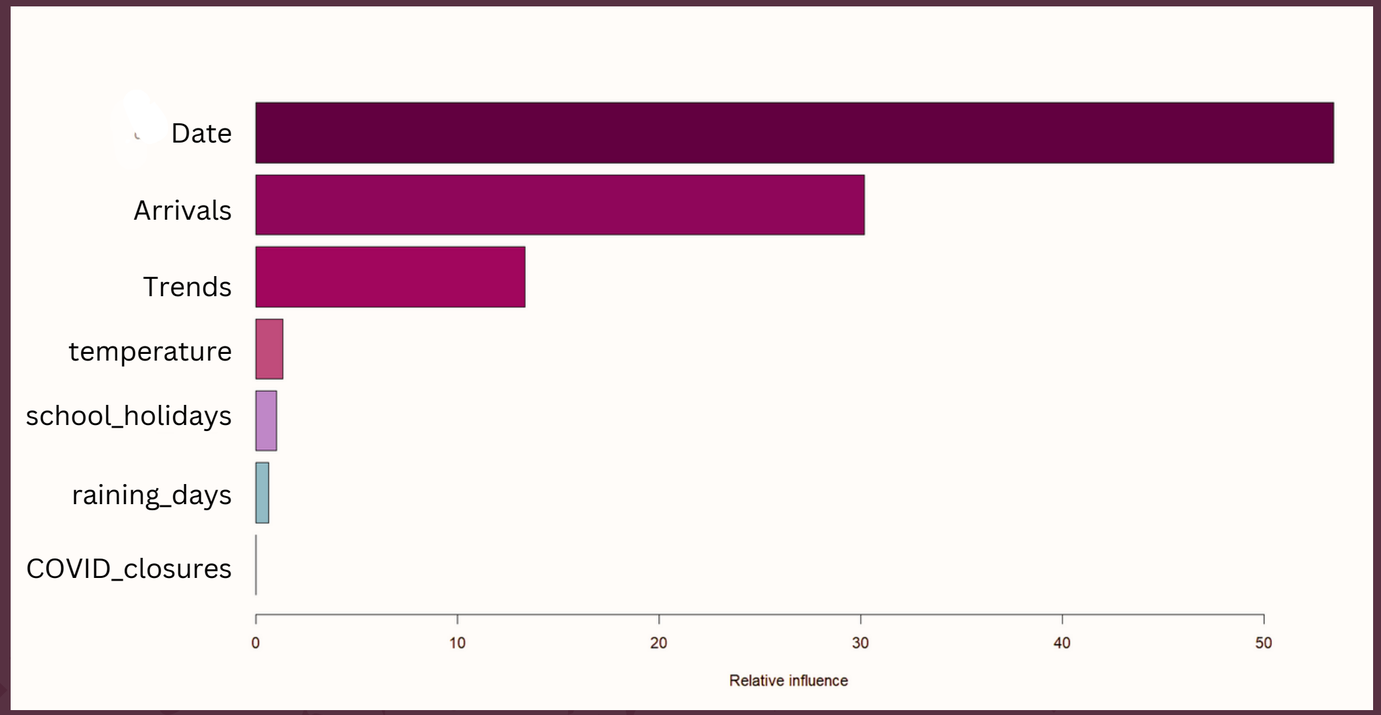
(Intercept)	0.299390532
year	-0.239334982
month	-0.046015655
date_numeric	-0.227478854
trends	0.121826512
average_temperature	-0.120760058
raining_days	0.102305064
school_holidays	0.137463598
arrivals	0.903604713
Covid_closures	-0.009818287

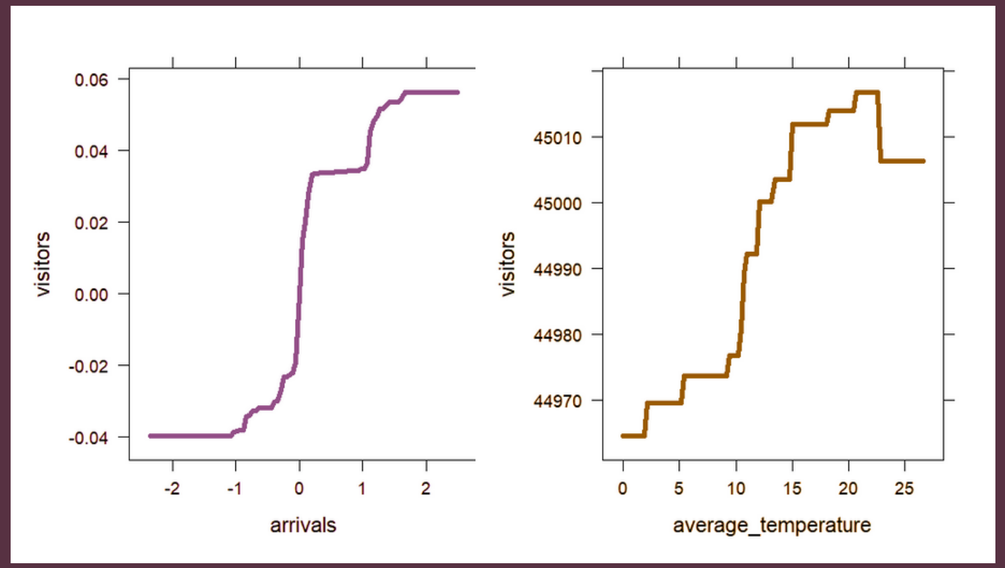


actual values - predictions

Boosting: GBM and XGBoost

BEST PREDICTORS

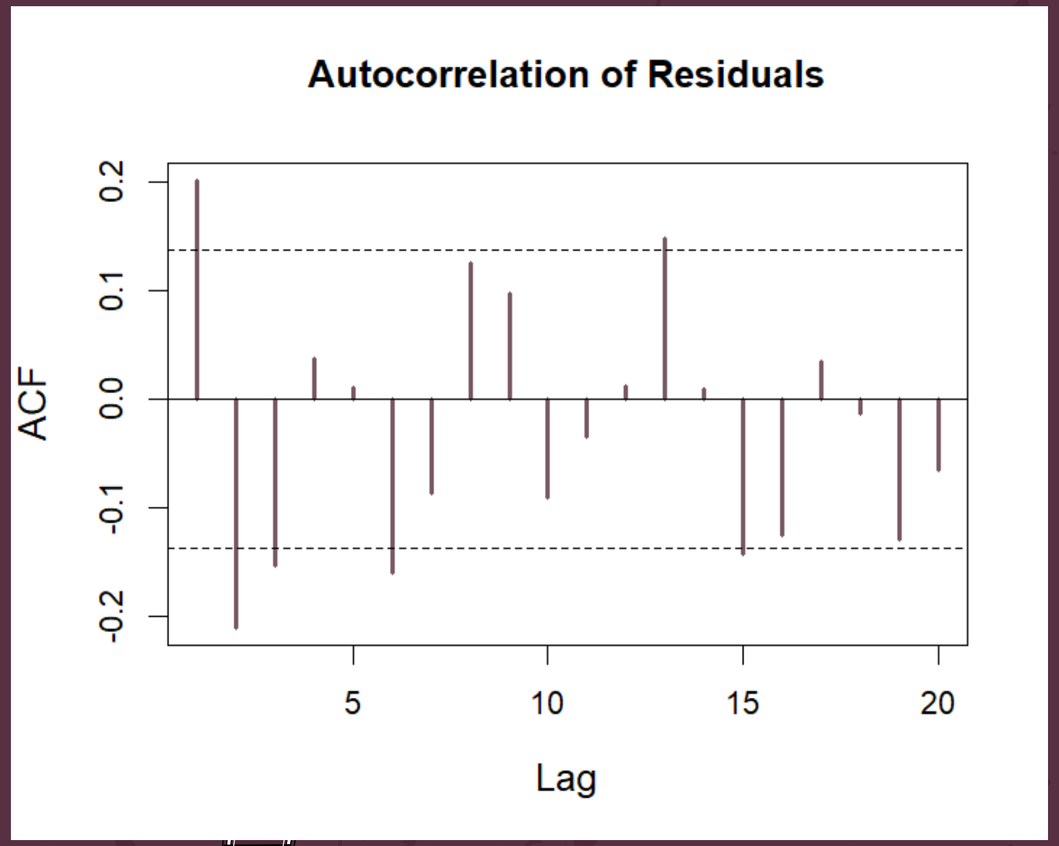


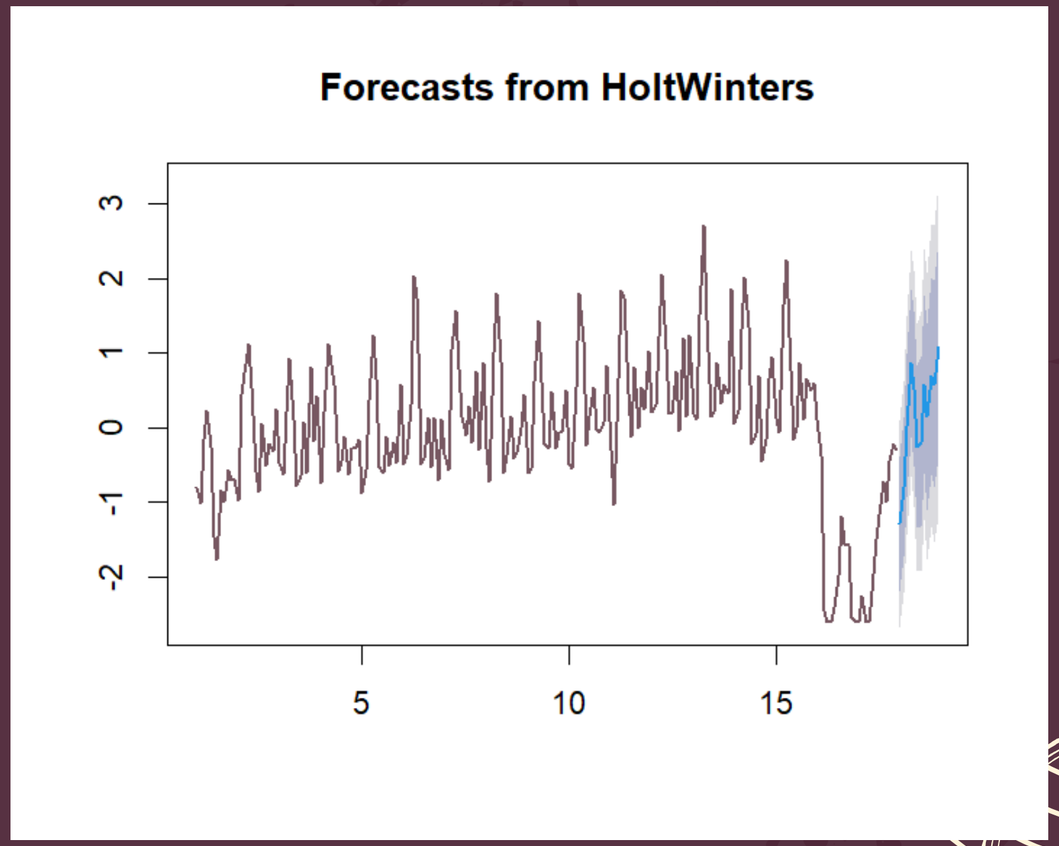


PDPs (all the other features constant)

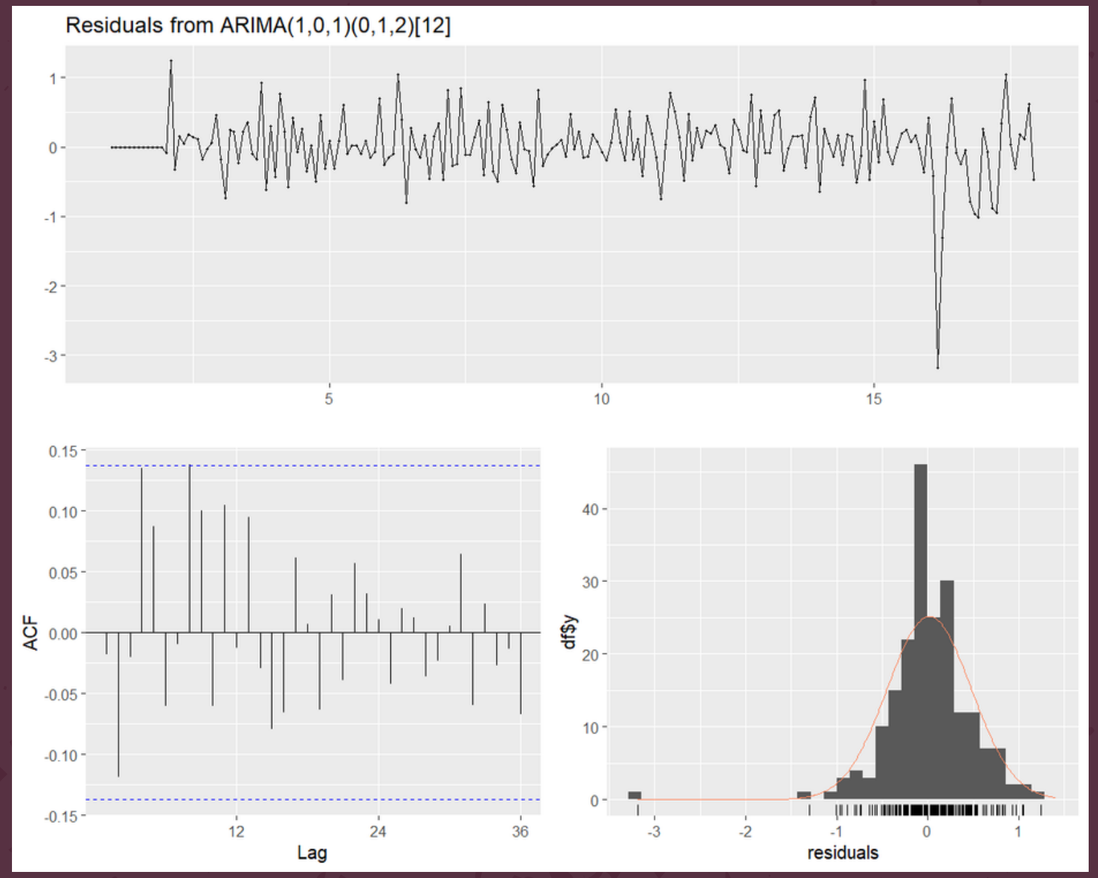
Holt-Winters' Exponential smoothing

Extends simple exponential smoothing to account for seasonality and trends in the data.





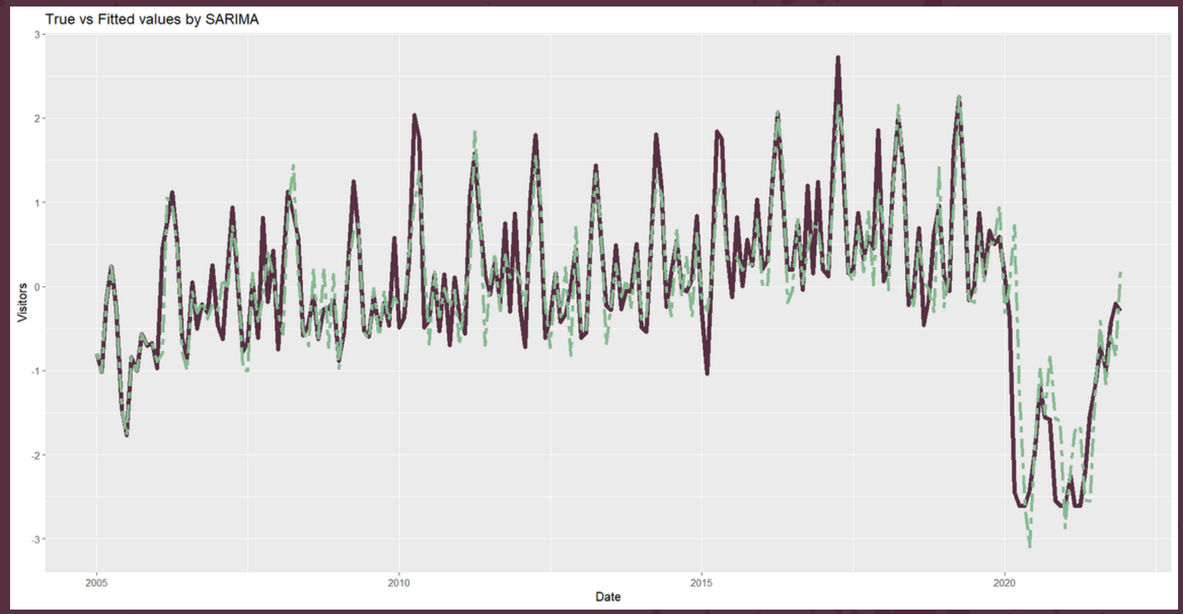
SARIMA



Pro: Effective for capturing linear trends and seasonality in time series data.

Cons:

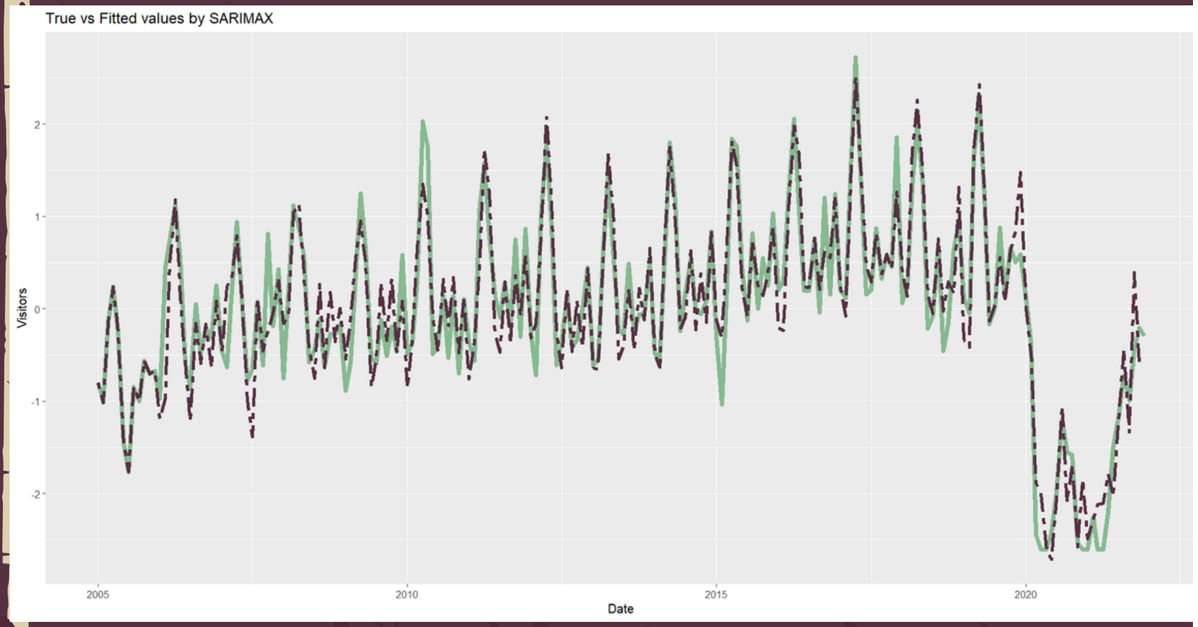
- Assumes linearity and stationarity;
- Sensitivity to outliers;
- Requires careful tuning of hyperparameters for optimal performance.

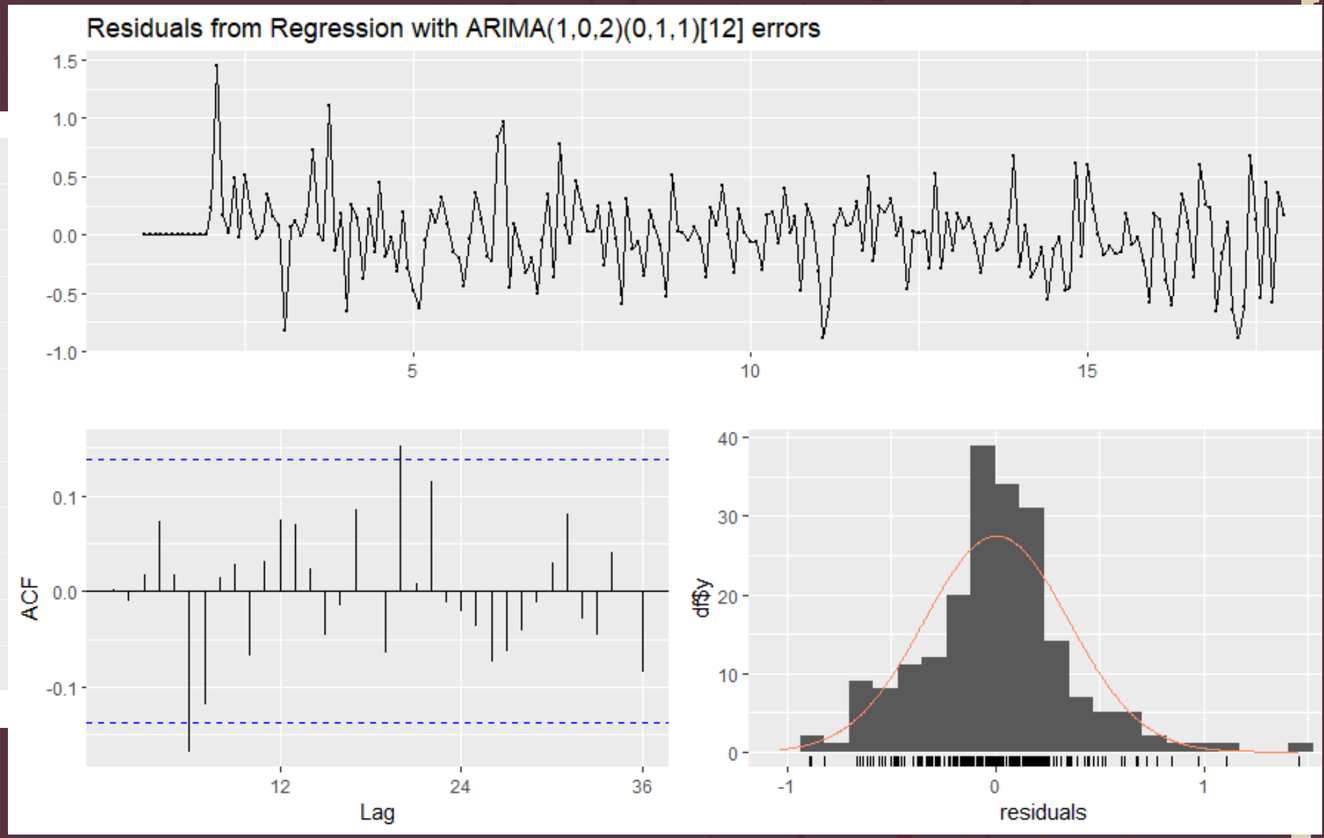


SARIMAX

External regressors:

- School holidays
- Covid closures
- Tourist arrivals
- Google trends

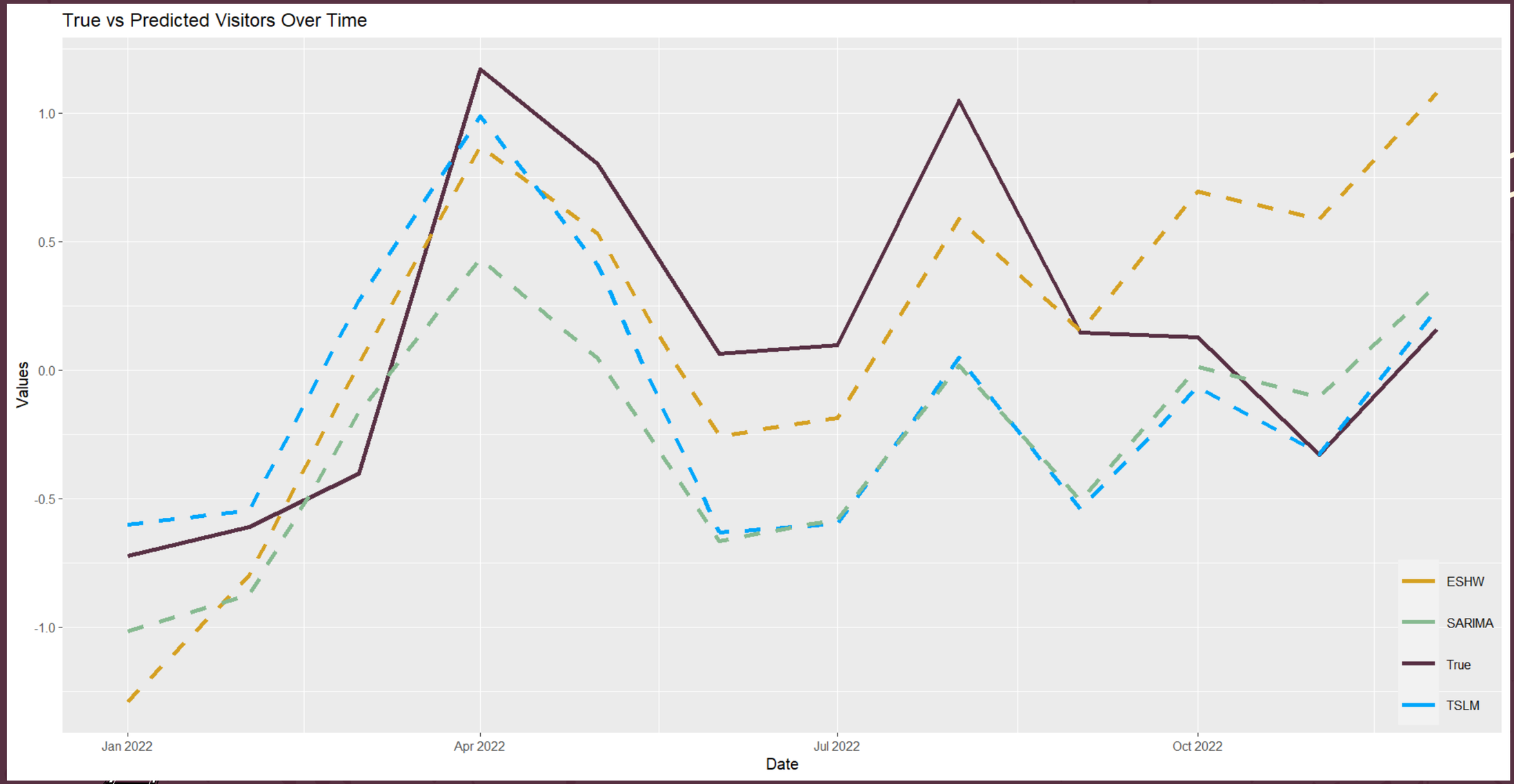




Forecasting results

Model	Adj R^2	RMSE	MAPE	AIC
Exp. smoothing Holt Winters	/	0,509	2,025	/
TSLM – Trend and Seasonality	0,203	0,512	2,615	547,121
SARIMA	0.783	0,57	2,469	300,266
XGBoost – TSCV	0,877	0,589	0,803	/
Boosting – TSCV	0.737	0,613	1,248	/
SARIMAX	0.880	0,654	2,608	188,719
Multiple LR Stepwise Both	0.806	0,845	2,531	262,314
Multiple LR Manual Features	0.664	0,883	3,6	365,472
GBM + TLSM + Boosting	/	0,904	3,499	/
GAM Stepwise	/	0,932	3,299	227,286
TSLM – Manual Features	0.787	0,933	1	280,321
L1/L2 Regularization TSCV	/	0,953	4,332	/
Generalized Bass Model – 2R	/	0,983	4,26	/
Auto ARIMA	0.770	1,248	6,092	298,947

Baseline	RMSE	MAE	MAPE
Train mean	0,602	0,473	1
TimeGPT	0,768	0,652	1,73


True vs Predicted Visitors Over Time
Values
1.0
0.5
0.0
-0.5
-1.0
Jan 2022
Apr 2022
Jul 2022
Oct 2022
Date
ESHW
SARIMA
True
TSLM

Best model Error Analysis

TRUE	PREDICTED	ERROR (%)	MONTH
47181	45217	**4.16**	October
47727	50742	6.32	December
39288	43205	9.97	November

Best predictions

TRUE	PREDICTED	ERROR (%)	MONTH
63103	45299	**28.21**	August
46099	33508	27.31	June
46655	34971	25.04	July

Worst predictions

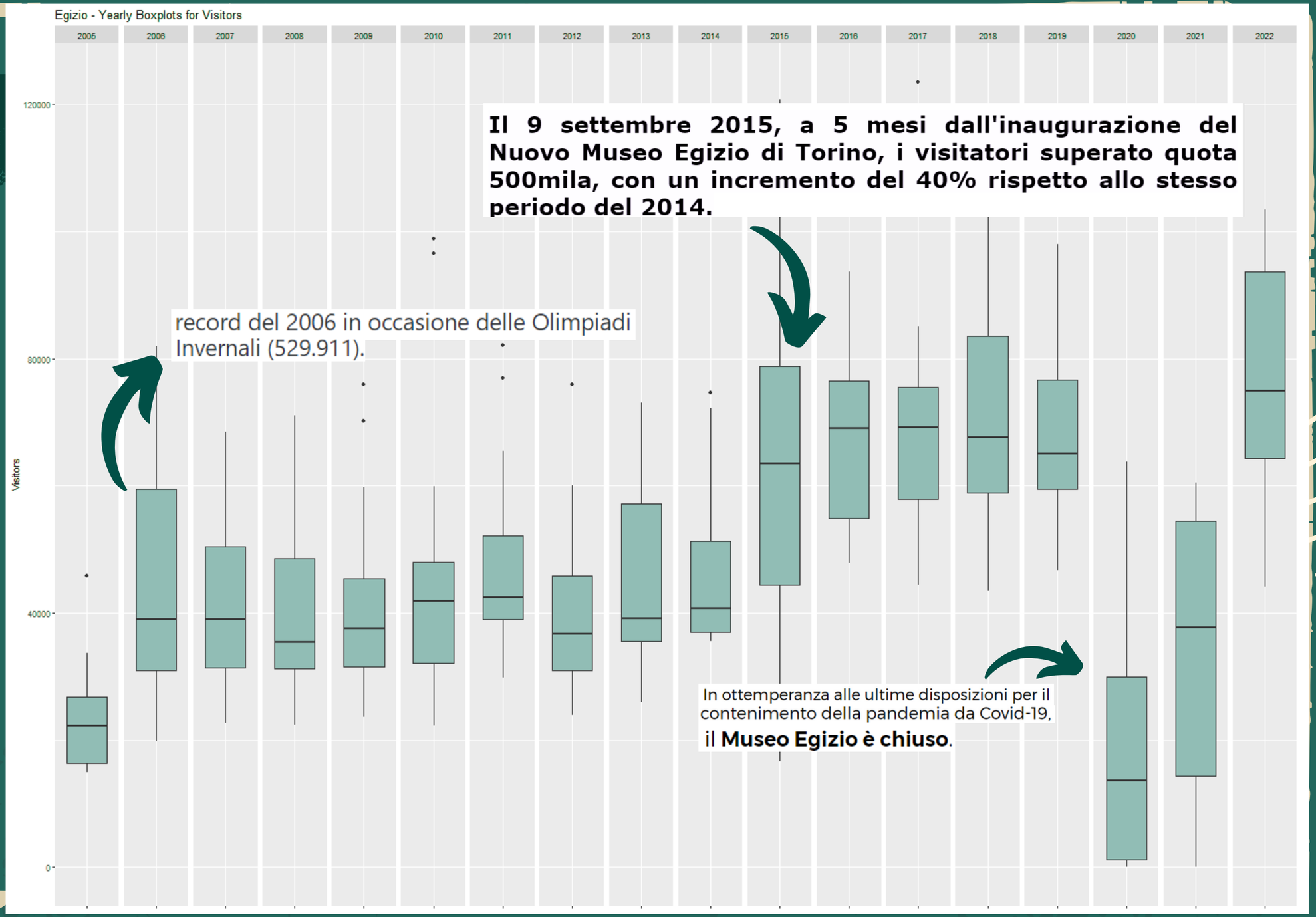
Museo Egizio (Egyptian Museum)

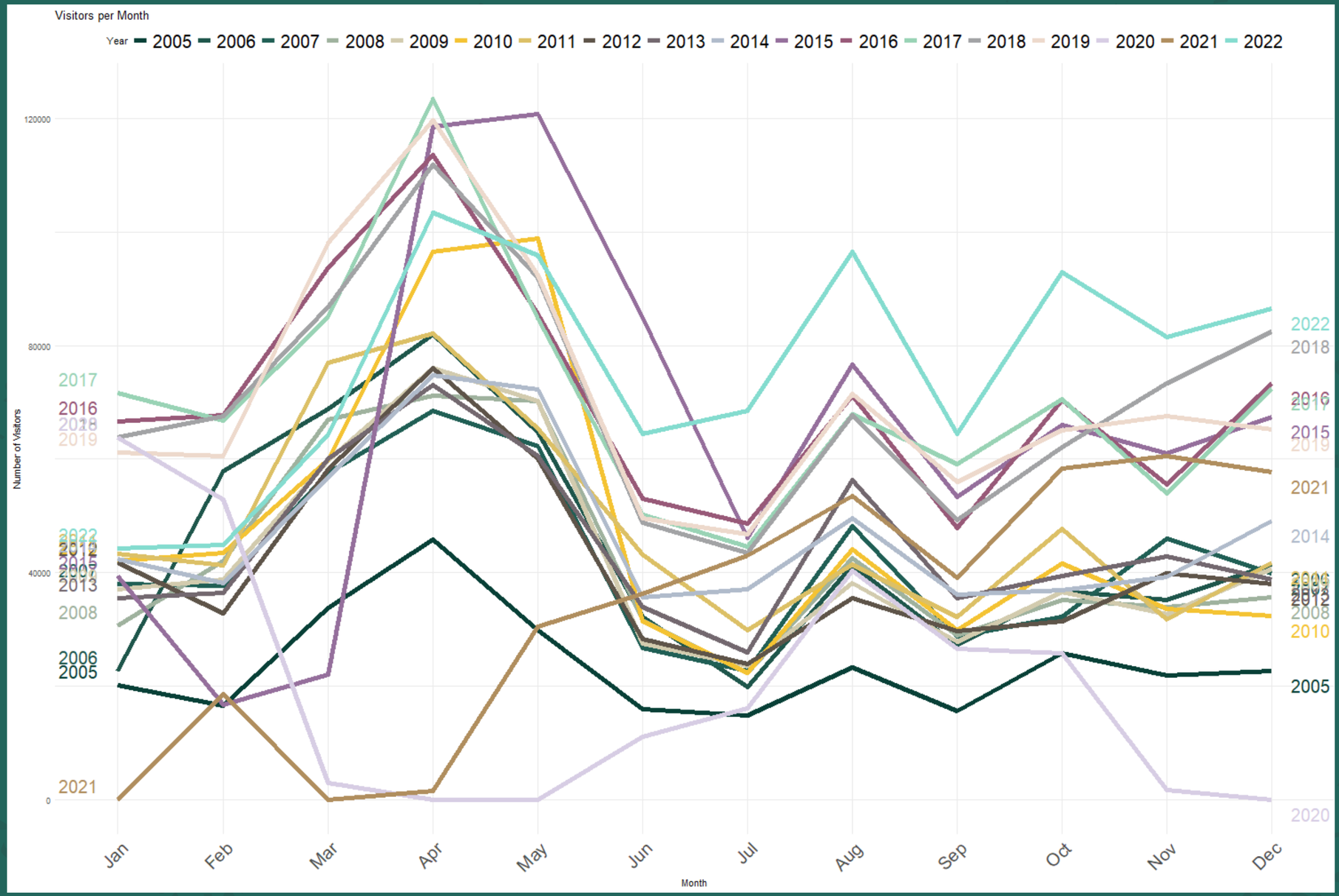
MUSEO
EGIZIO

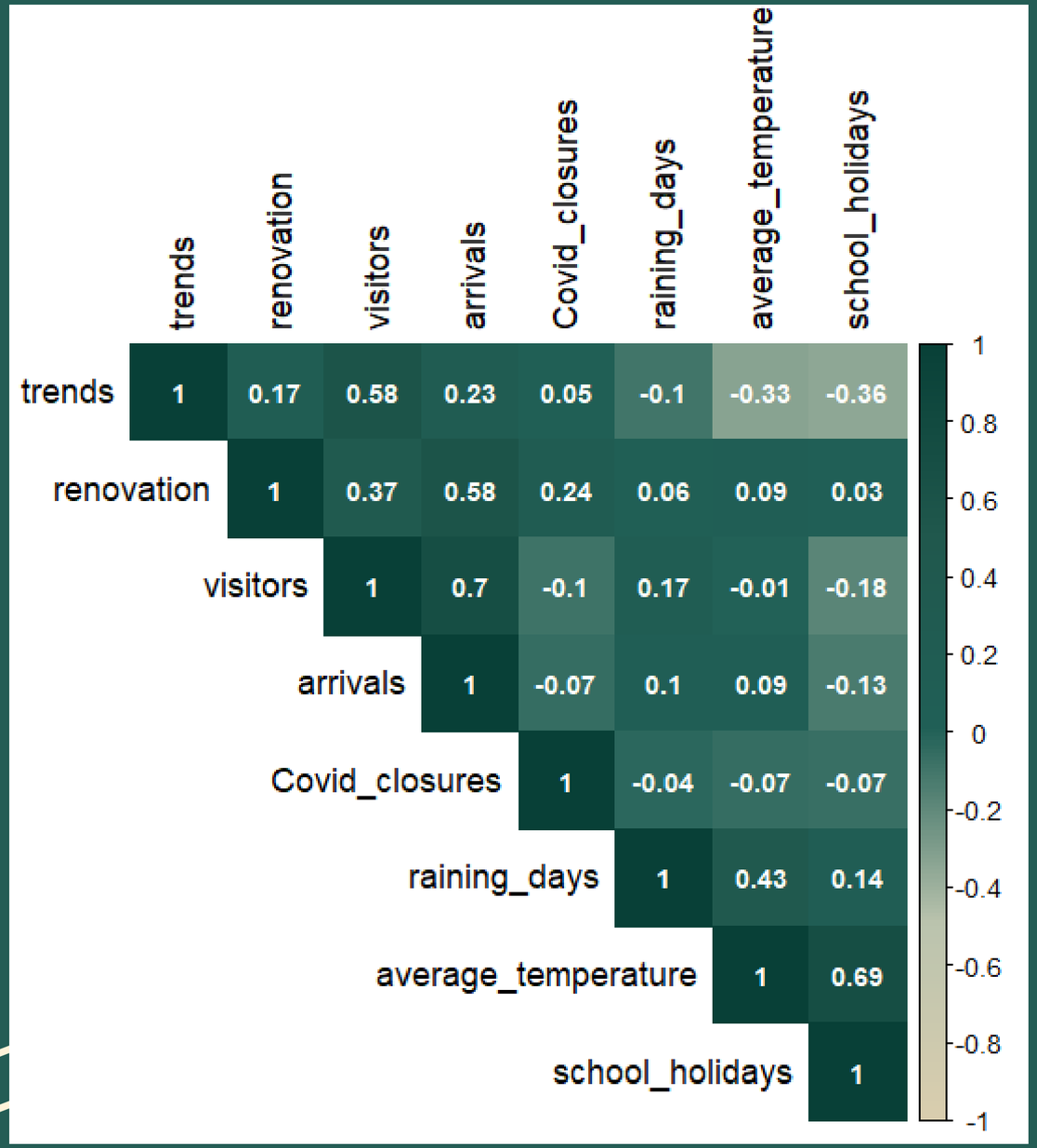
DATA COLLECTED FROM: https://ocp.piemonte.it

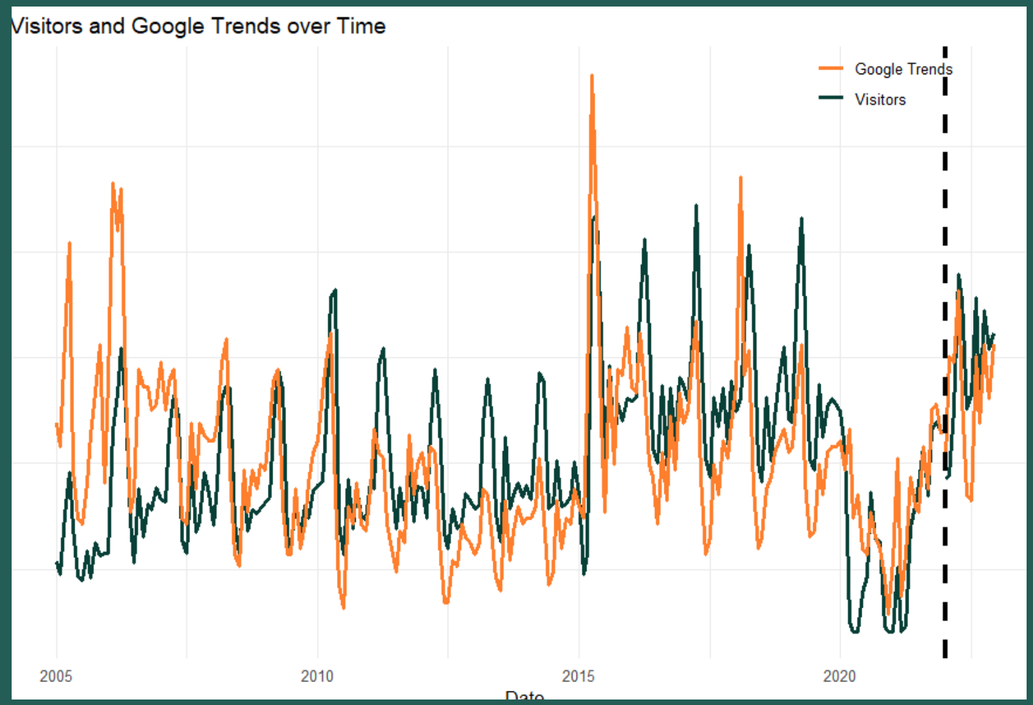
EDA

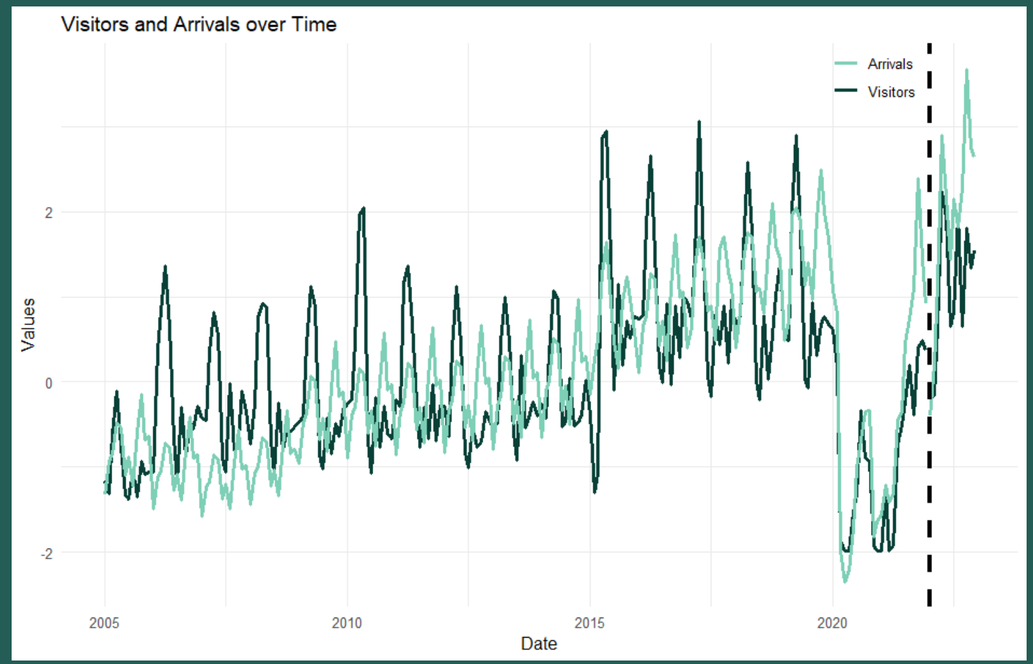
Yearly Boxplots for Visitors




Visitors per Month
Year 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
120000
80000
40000
0
Number of Visitors
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

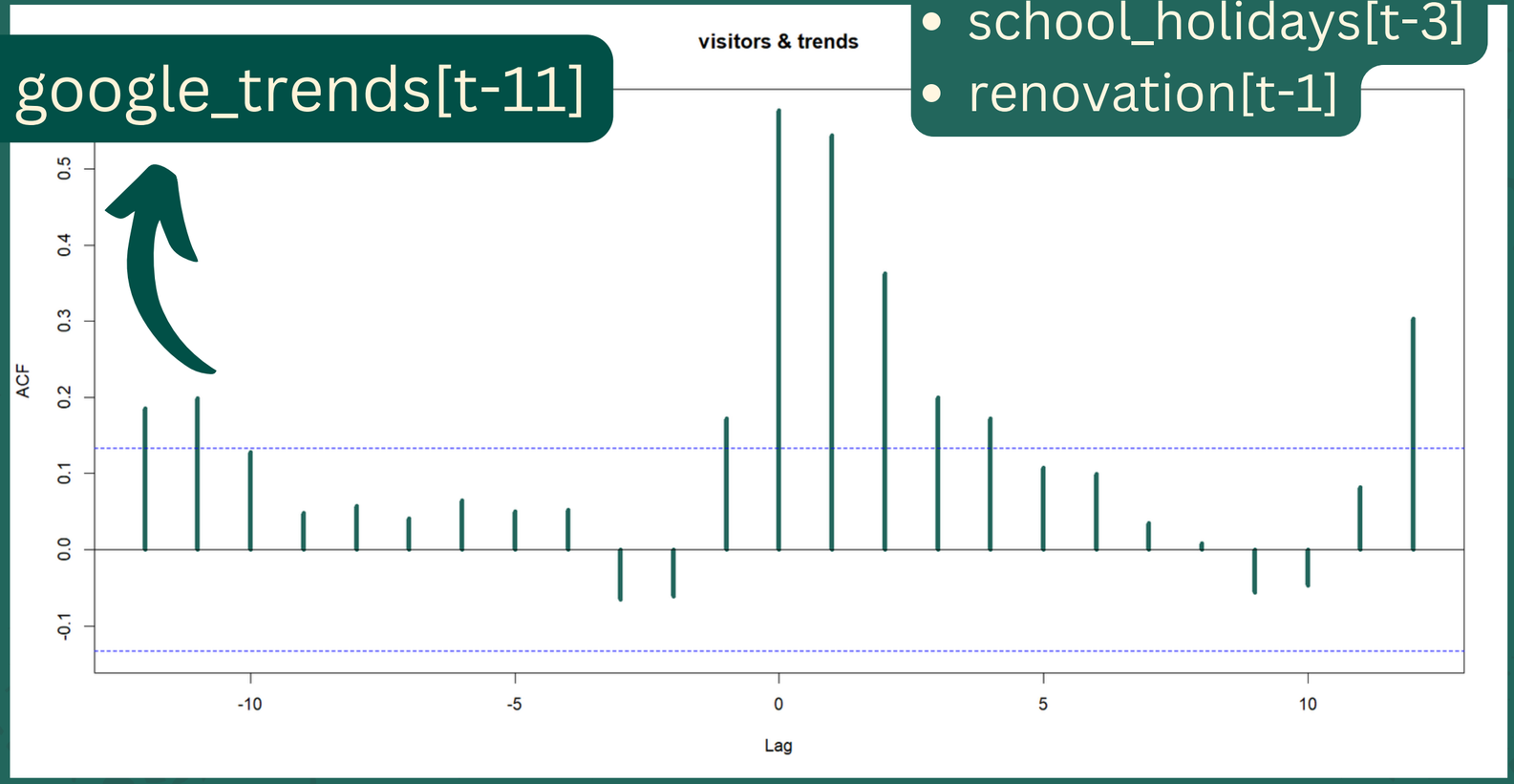

trends
renovation
visitors
arrivals
Covid_closures
raining_days
average_temperature
school_holidays
trends 1 0.17 0.58 0.23 0.05 -0.1 -0.33 -0.36
renovation 1 0.37 0.58 0.24 0.06 0.09 0.03
visitors 1 0.7 -0.1 0.17 -0.01 -0.18
arrivals 1 -0.07 0.1 0.09 -0.13
Covid_closures 1 -0.04 -0.07 -0.07
raining_days 1 0.43 0.14
average_temperature 1 0.69
school_holidays 1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


Visitors and Google Trends over Time
Google Trends
Visitors
2005
2010
2015
2020
Date


Visitors and Arrivals over Time
Arrivals
Visitors
Values
2
0
-2
2005
2010
2015
2020
Date

Lagged variables

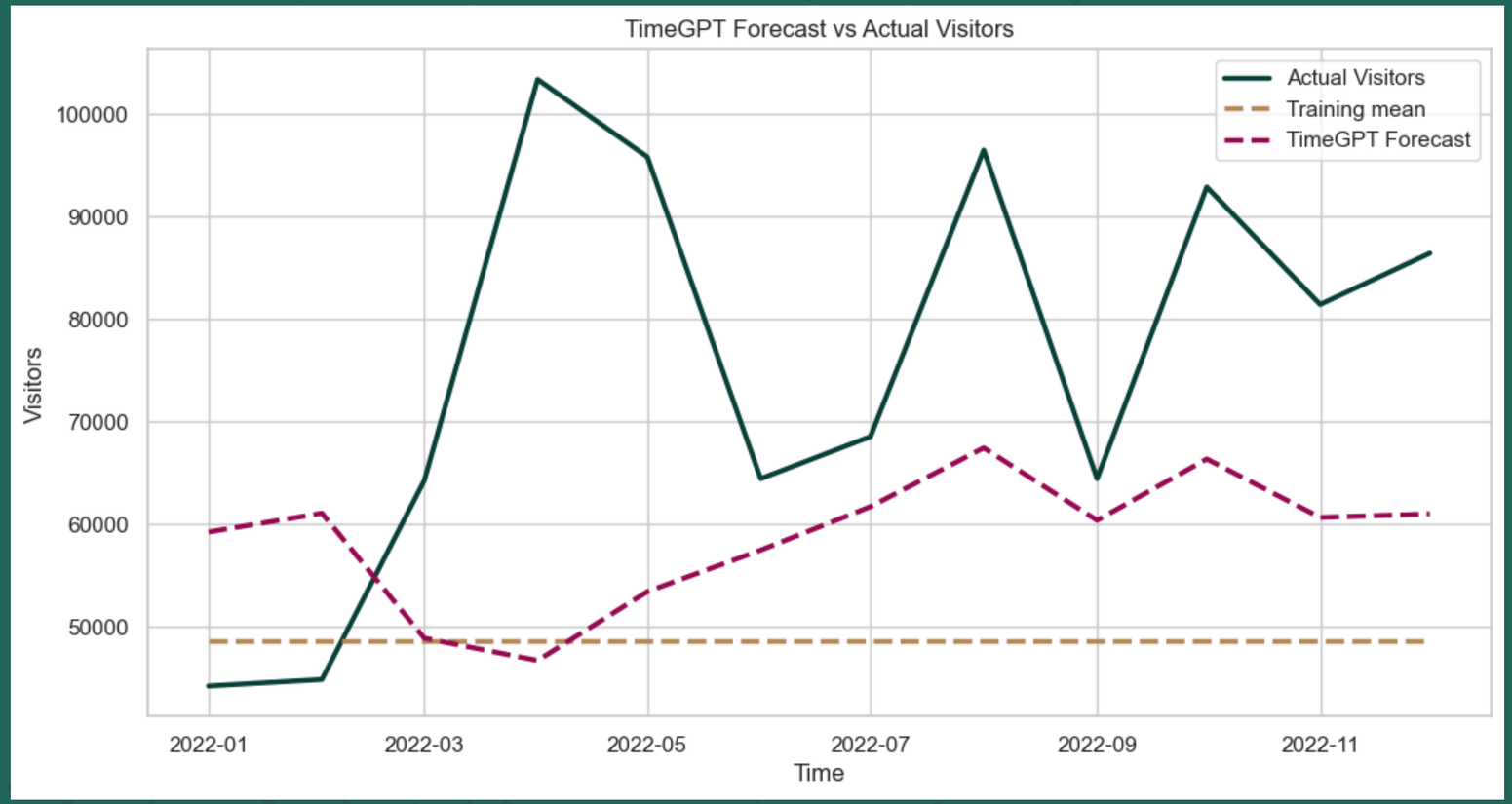
visitors[t] ~ google_trends[t-11]



- COVID: There were no significant lags.
- arrivals[t-6]
- average_temperature[t-3]
- raining_days[t-1]
- school_holidays[t-3]
- renovation[t-1]

Baselines

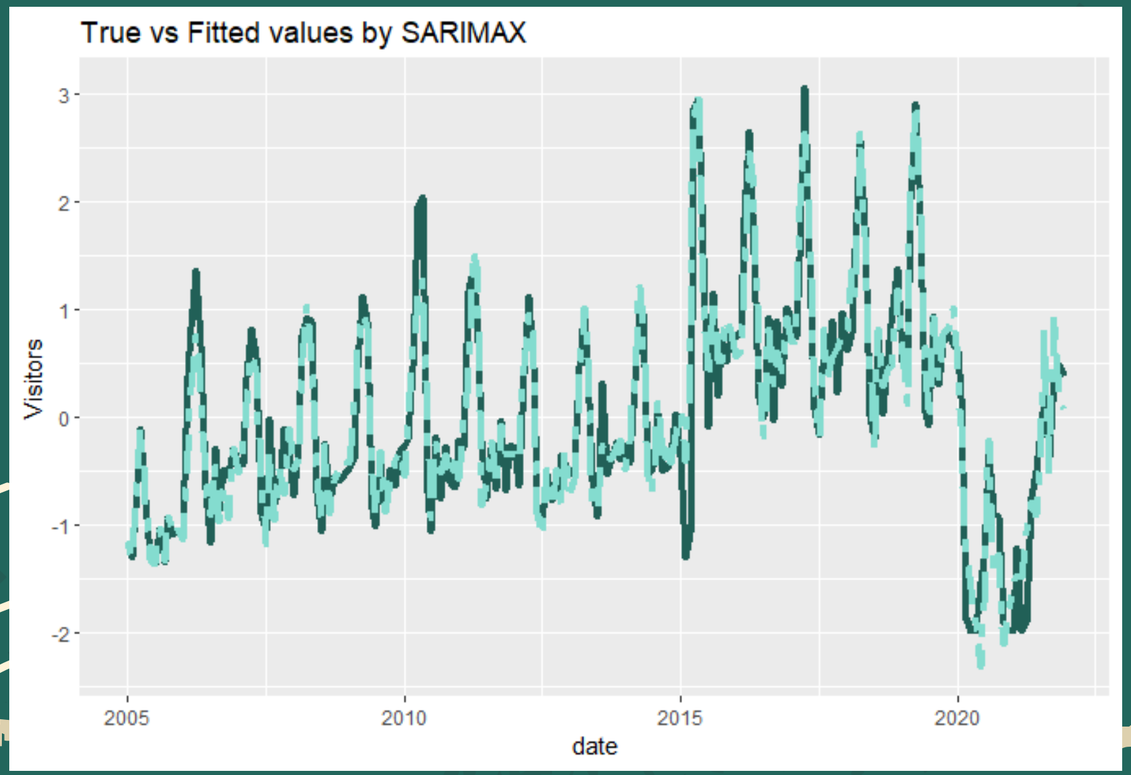
Baseline	RMSE	MAPE
Mean	1.353	1.0
TimeGPT	1.085	1.185

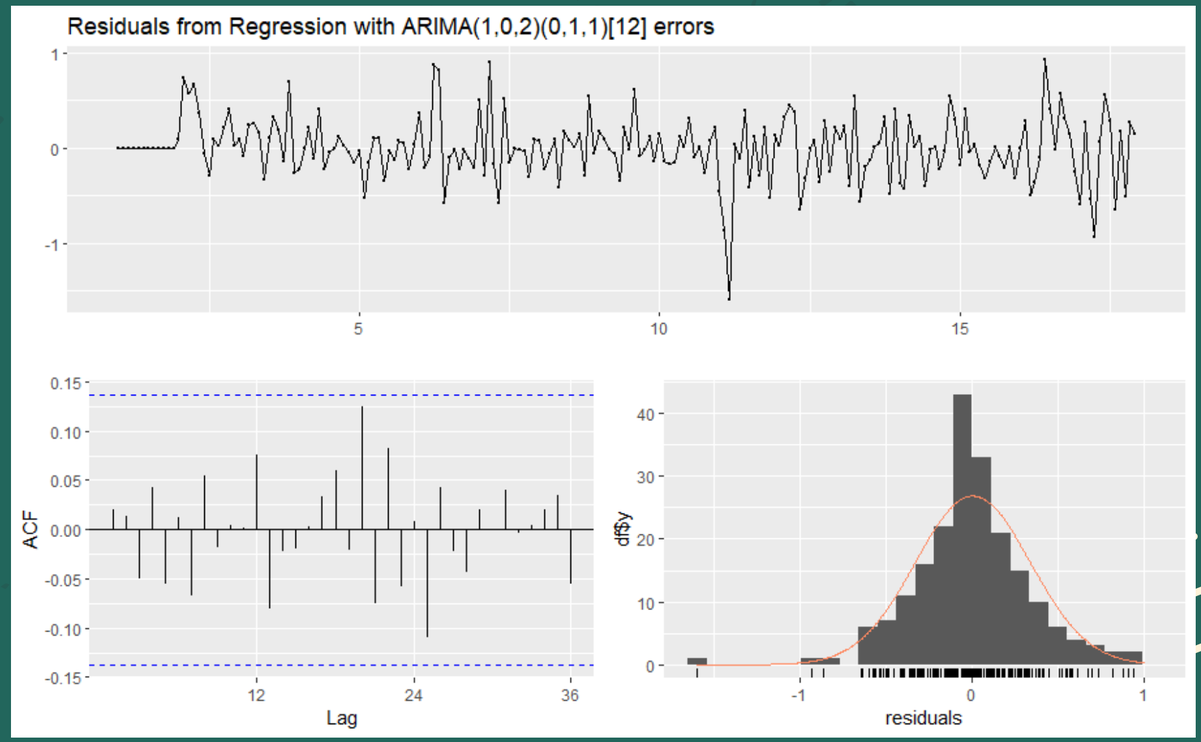


SARIMAX

External regressors:

- Renovation
- Tourist arrivals
- Google trends





RMSE	MAPE	R^2	AIC
0.278	0.386	0.901	314.330

Stepwise GAM

```
Call: gam(formula = visitors ~ year + month + s(date_numeric, df = 4) +
    trends + s(raining_days, df = 3) + school_holidays + s(arrivals,
    df = 4) + lagged_trends + lagged_average_temperature + lagged_renovation,
    data = egizio_train_df, trace = FALSE)
Deviance Residuals:
      Min        1Q    Median        3Q       Max
-1.678836 -0.169421  0.003474  0.184206  1.149119

(Dispersion Parameter for gaussian family taken to be 0.1144)

    Null Deviance: 203 on 203 degrees of freedom
Residual Deviance: 20.026 on 174.9999 degrees of freedom
AIC: 165.4248
```

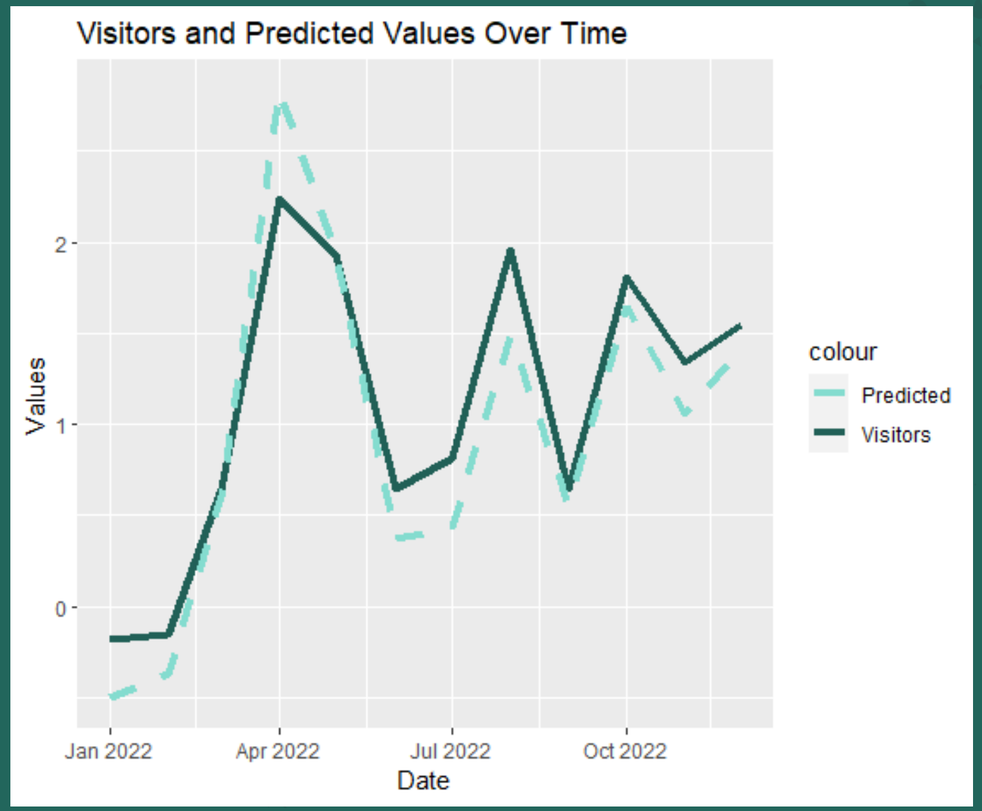
Anova for Parametric Effects

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
year	1	12.109	12.109	105.8185	< 2.2e-16	***
month	11	59.912	5.447	47.5952	< 2.2e-16	***
s(date_numeric, df = 4)	1	11.174	11.174	97.6473	< 2.2e-16	***
trends	1	48.077	48.077	420.1284	< 2.2e-16	***
s(raining_days, df = 3)	1	0.091	0.091	0.7977	0.373019	
school_holidays	1	0.121	0.121	1.0574	0.305232	
s(arrivals, df = 4)	1	36.197	36.197	316.3158	< 2.2e-16	***
lagged_trends	1	2.214	2.214	19.3499	1.886e-05	***
lagged_average_temperature	1	0.327	0.327	2.8555	0.092845	.
lagged_renovation	1	1.141	1.141	9.9677	0.001876	**
Residuals	175	20.026	0.114			

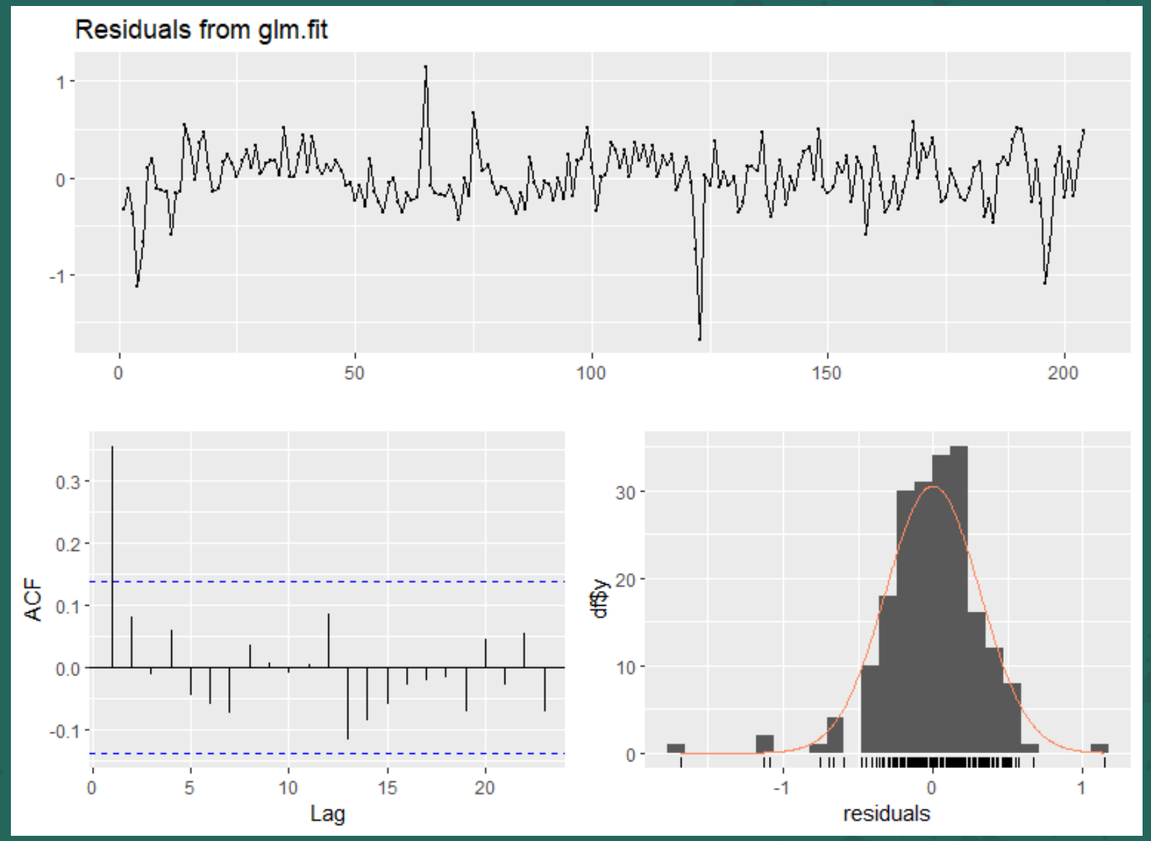
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Anova for Nonparametric Effects

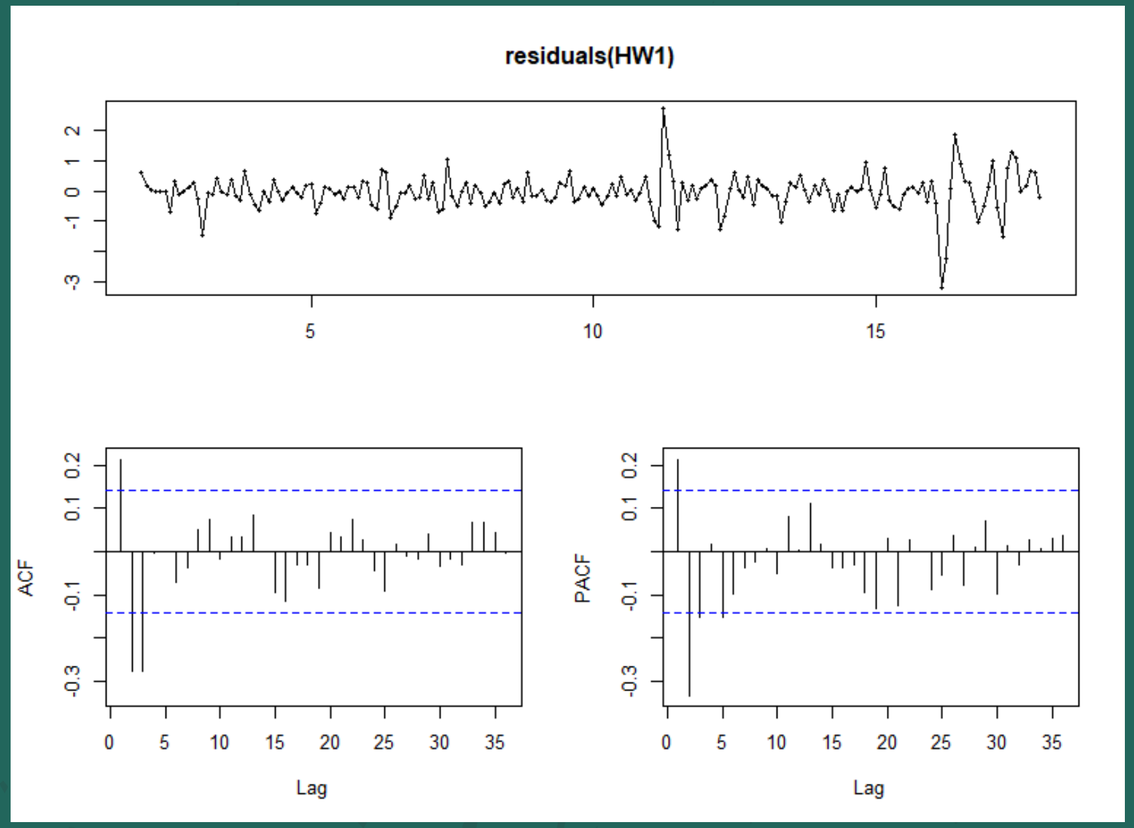
	Npar Df	Npar F	Pr(F)	
(Intercept)				
year				
month				
s(date_numeric, df = 4)	3	29.4449	1.776e-15	***
trends				
s(raining_days, df = 3)	2	2.9279	0.056135	.
school_holidays				
s(arrivals, df = 4)	3	4.8311	0.002954	**
lagged_trends				
lagged_average_temperature				
lagged_renovation				

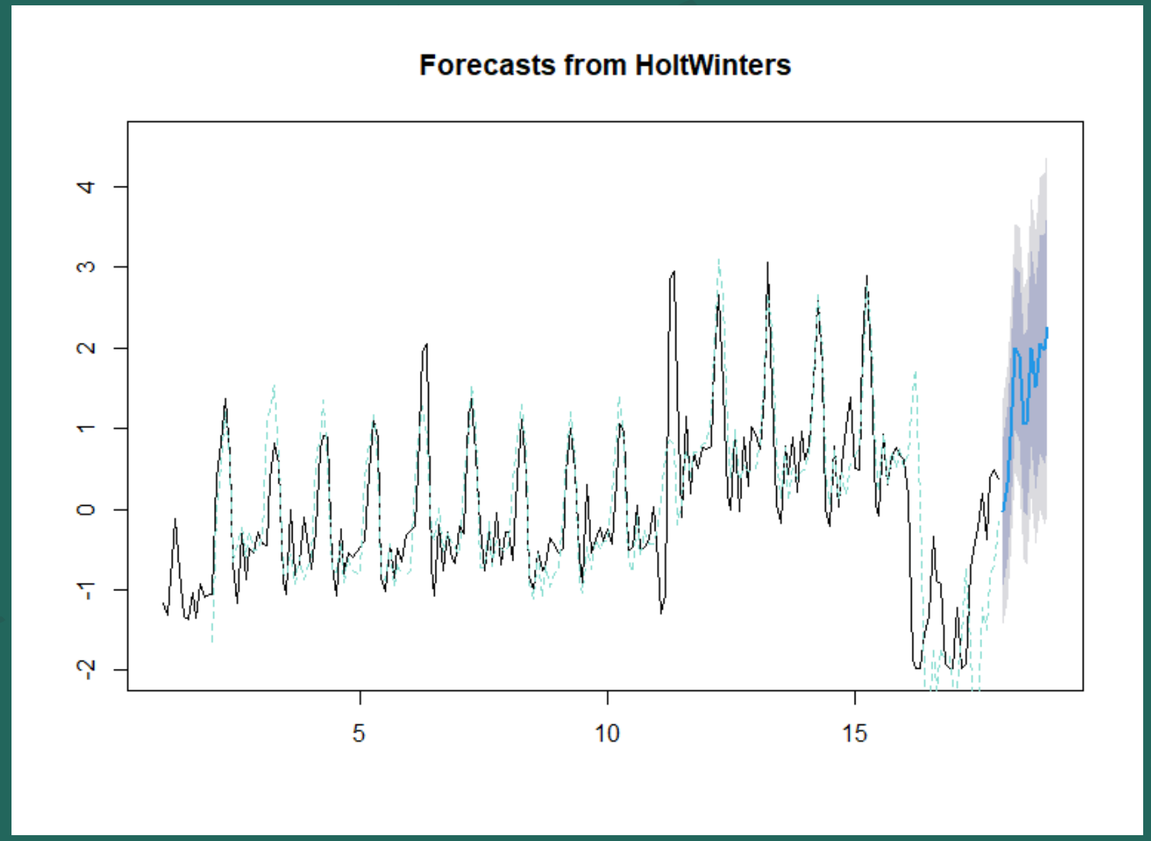


RMSE	MAPE	AIC
0.299	0.246	165.425



Holt-Winters' Exponential Smoothing

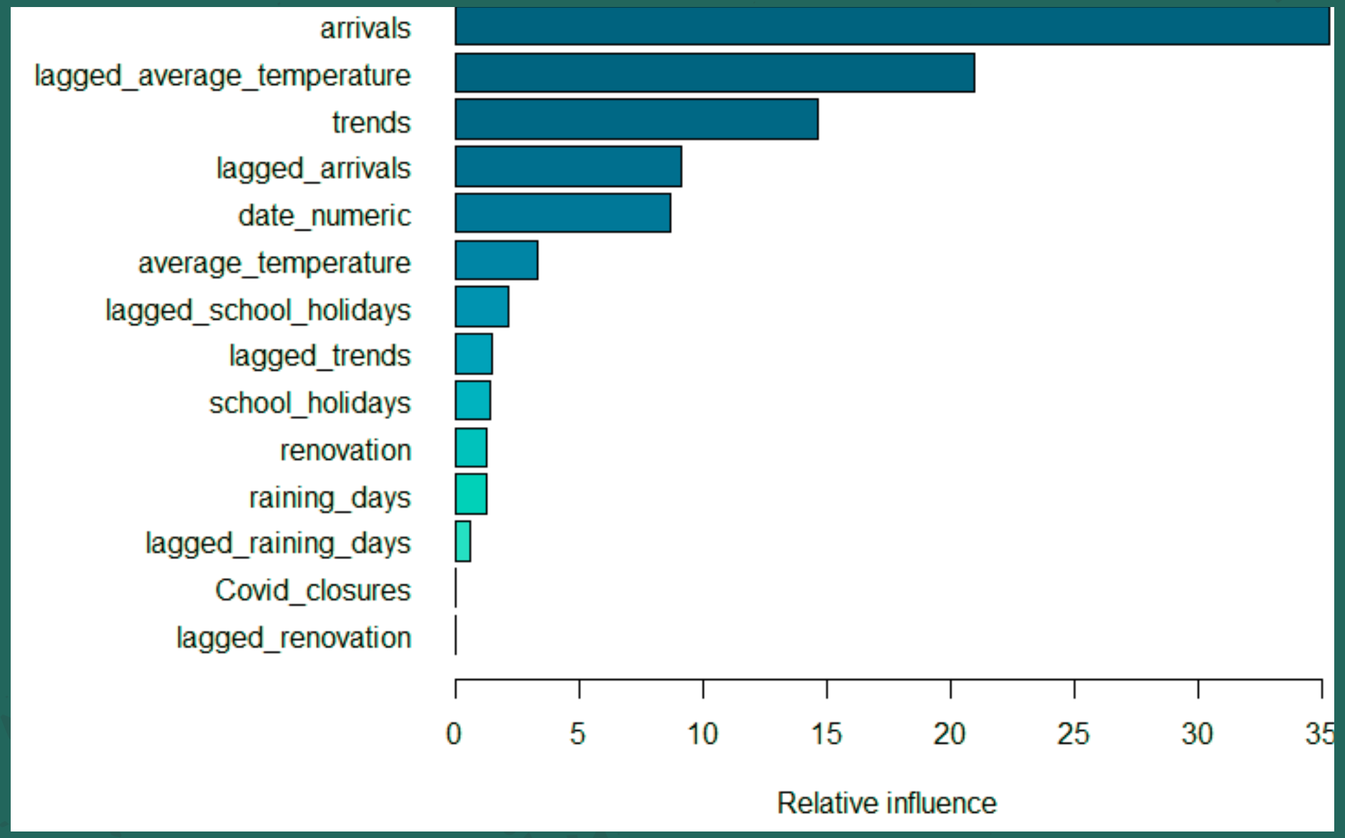




RMSE	MAE	MAPE
0.44	0,359	0.652

Gradient Boosting

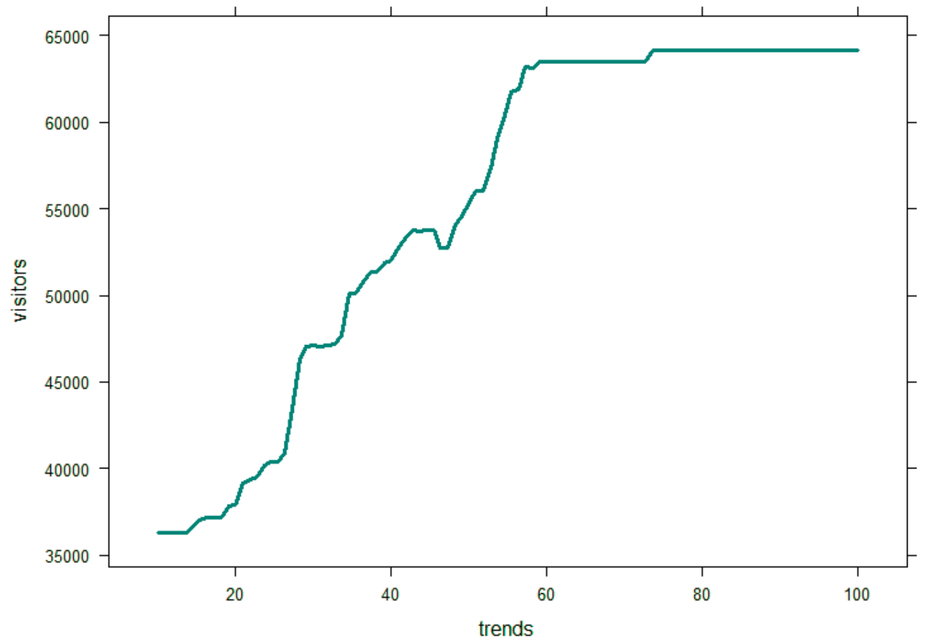
BEST PREDICTORS

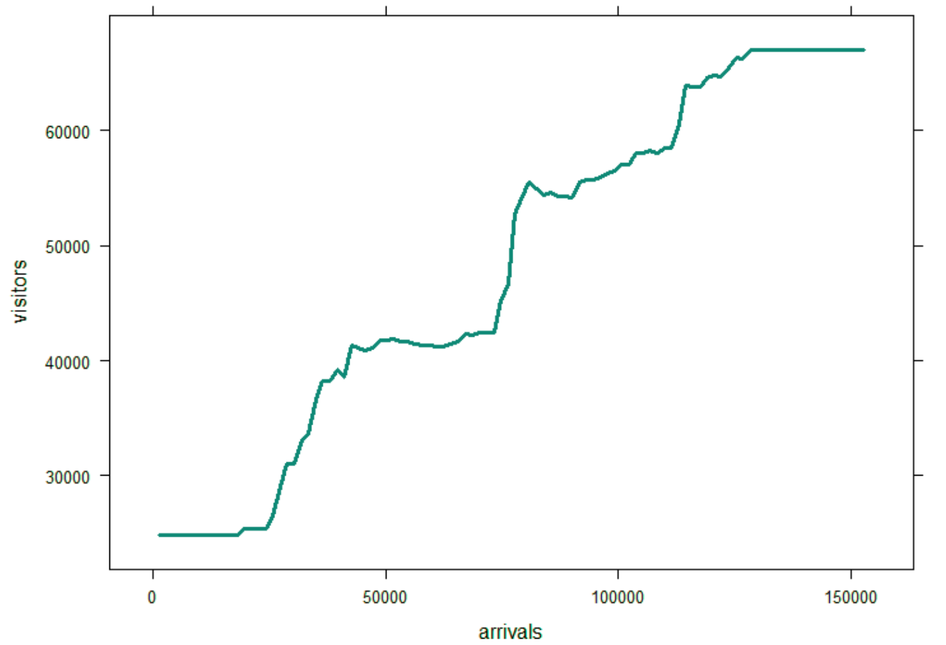


WHY?

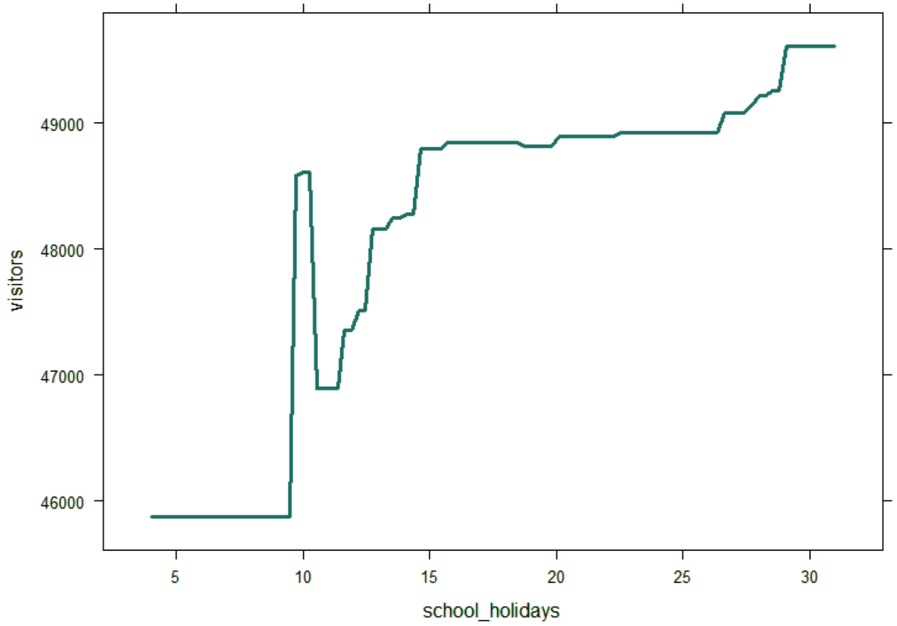
- Interpretability of the results
- Predictive powers

RMSE	MAE	MAPE
0.52	0,469	0.746



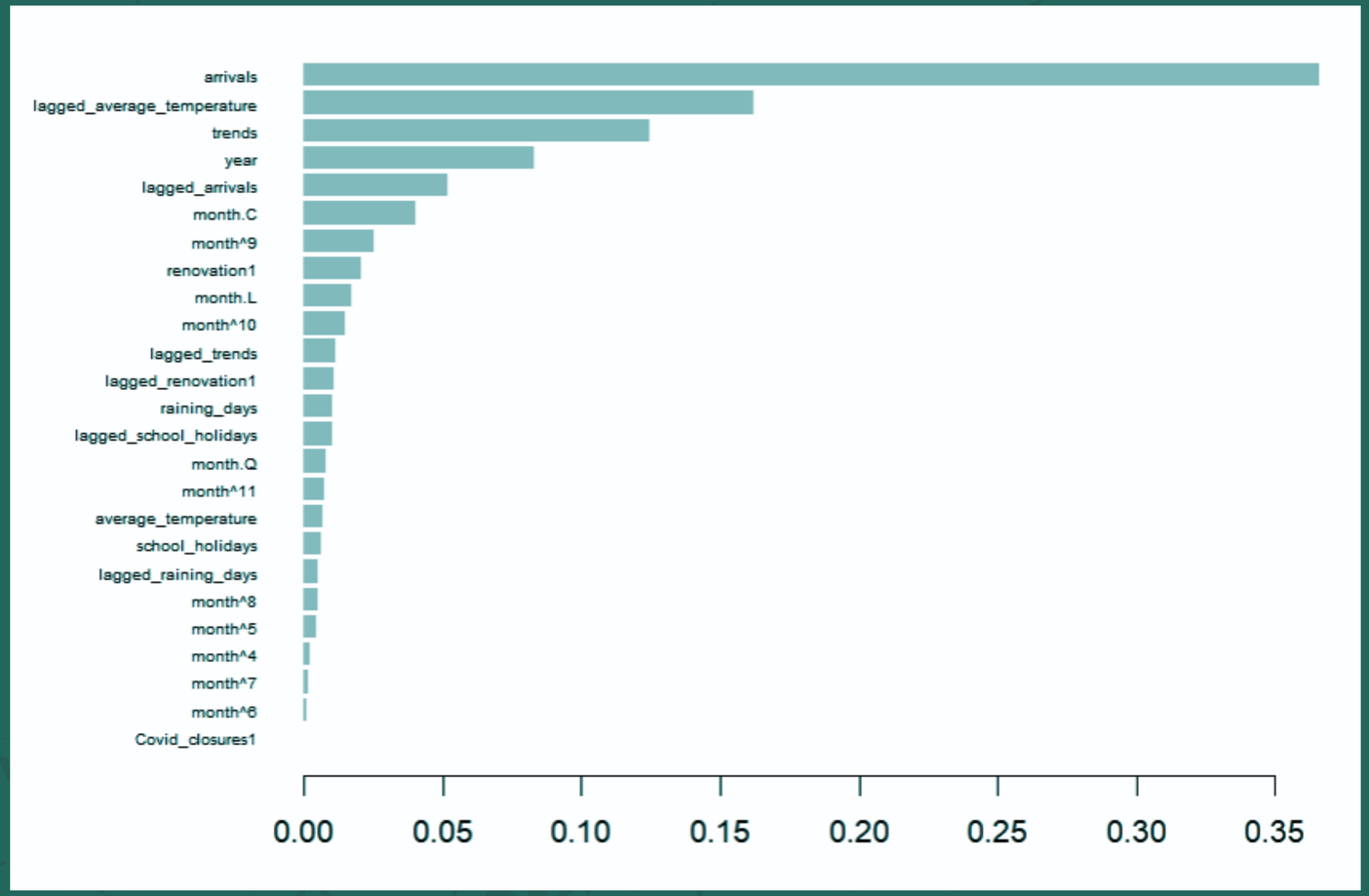


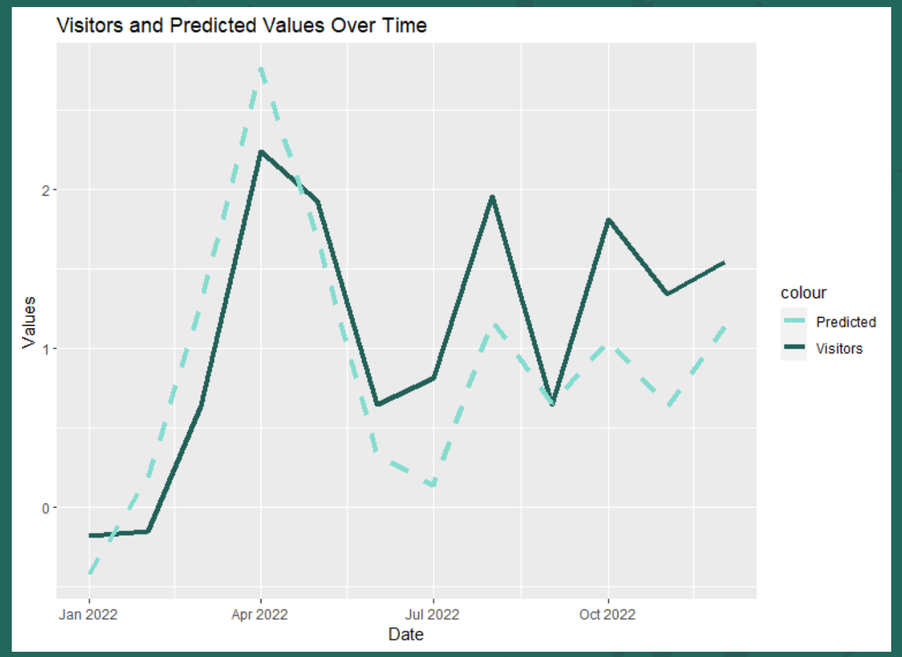
Partial dependence plots



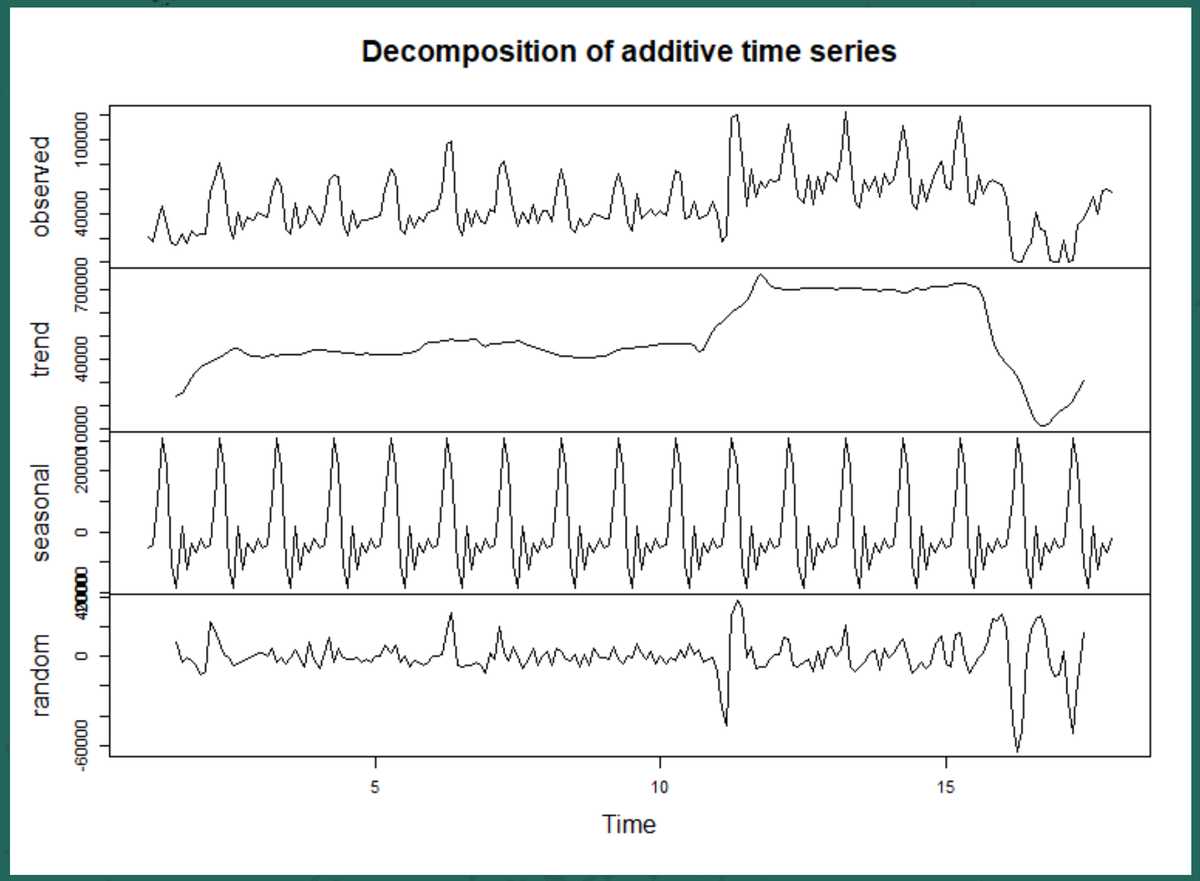
XGboost

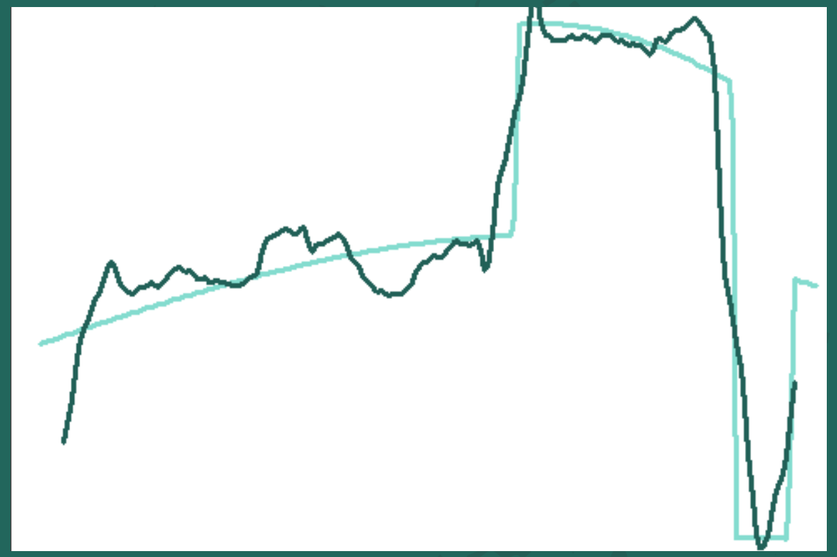
RMSE	MAE	MAPE
0.572	0.502	0.532





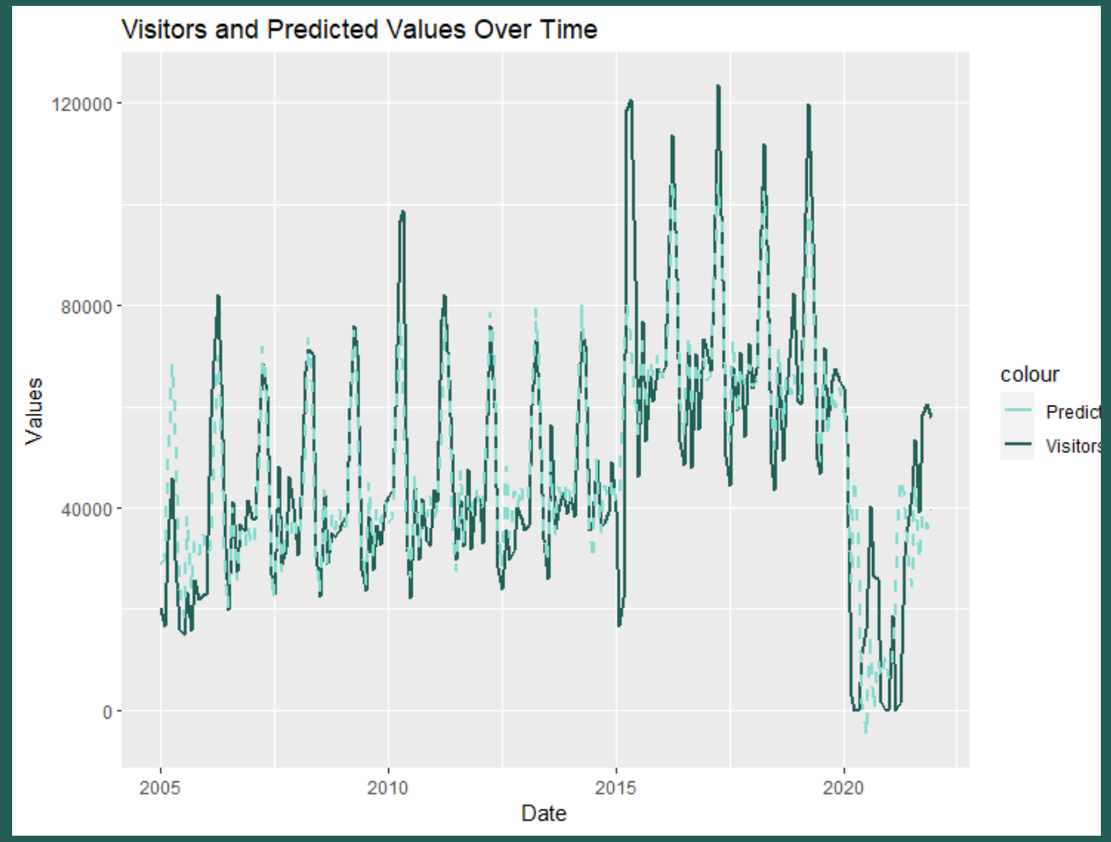
Decomposition: GBM+TLSM+Boosting



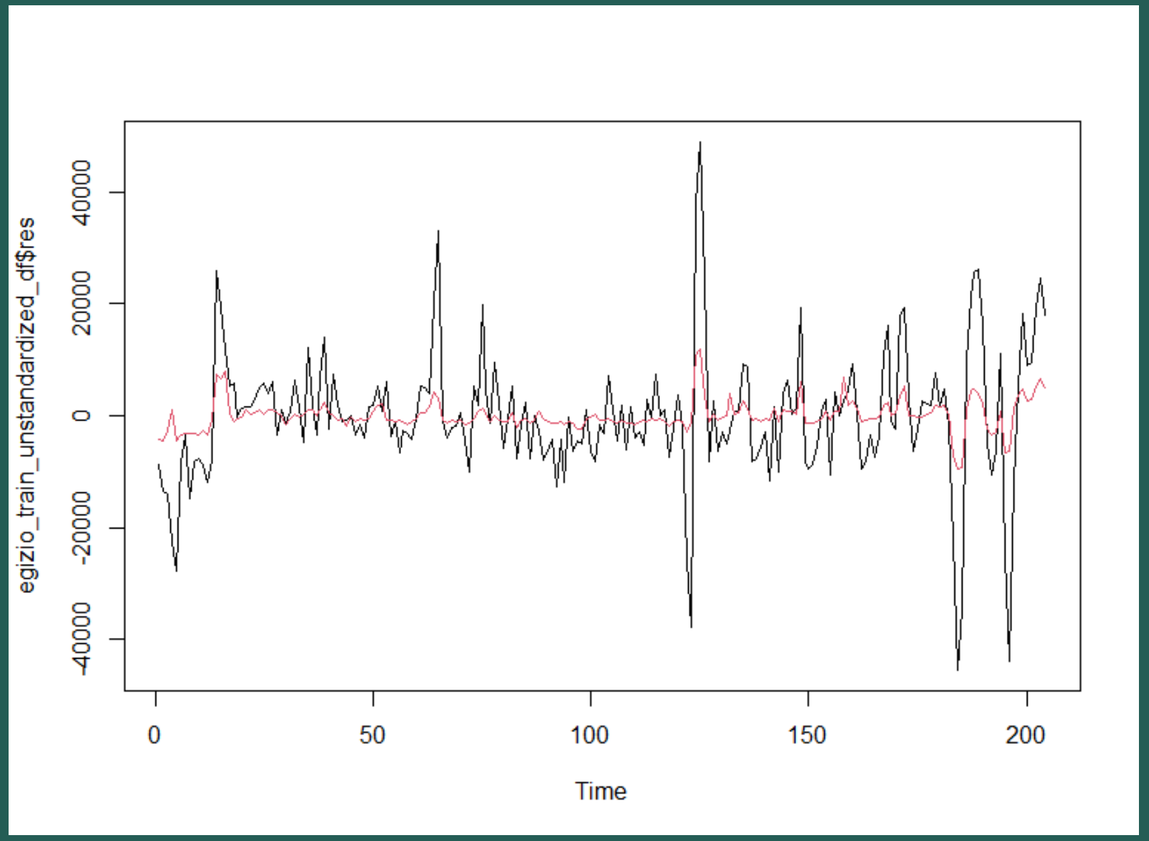


1. **Trend**: Modelled with GBM with 2 rectangular shocks:
   - Renovation 2015
   - Covid 2020

2. **Seasonality** with TSLM



3 . **Residuals** with Boosting




```
final_predictions <- trend_predictions + seasonality_predictions + residual_predictions
```

RMSE	MAE	MAPE
1.38	1.234	1.274

Forecasting results

Model	Adj R^2	RMSE	MAPE	AIC
SARIMAX	0,884	0.271	**0,373**	314,329
GAM Stepwise	/	0,299	0,433	165,425
Exp. smoothing Holt Winters	/	0,44	0,652	/
TSLM – Manual Features	0,838	**0,444**	1,117	242,818
SARIMA Improved	0,769	0,468	0,642	314,323
Multiple LR Stepwise Both	0,832	0,47	0,546	233,697
Boosting – TSCV	0,992	0,513	0,573	/
L1/L2 Regularization TSCV	/	0,545	1,01	/
XGBoost – TSCV	0,988	0,572	0,532	/
Multiple LR Manual Features	0,764	0,597	1,1	296,348
TSLM – Trend and Seasonality	0,29	0,946	4,162	523,408
Auto ARIMA	0,699	1,177	1,233	/
GBM + TLSM + Boosting	/	1,38	1,274	/
Generalized Bass Model – 2R	/	1,615	1,184	/

Baseline	RMSE	MAPE
Train mean	1.353	1
TimeGPT	1.085	1.185

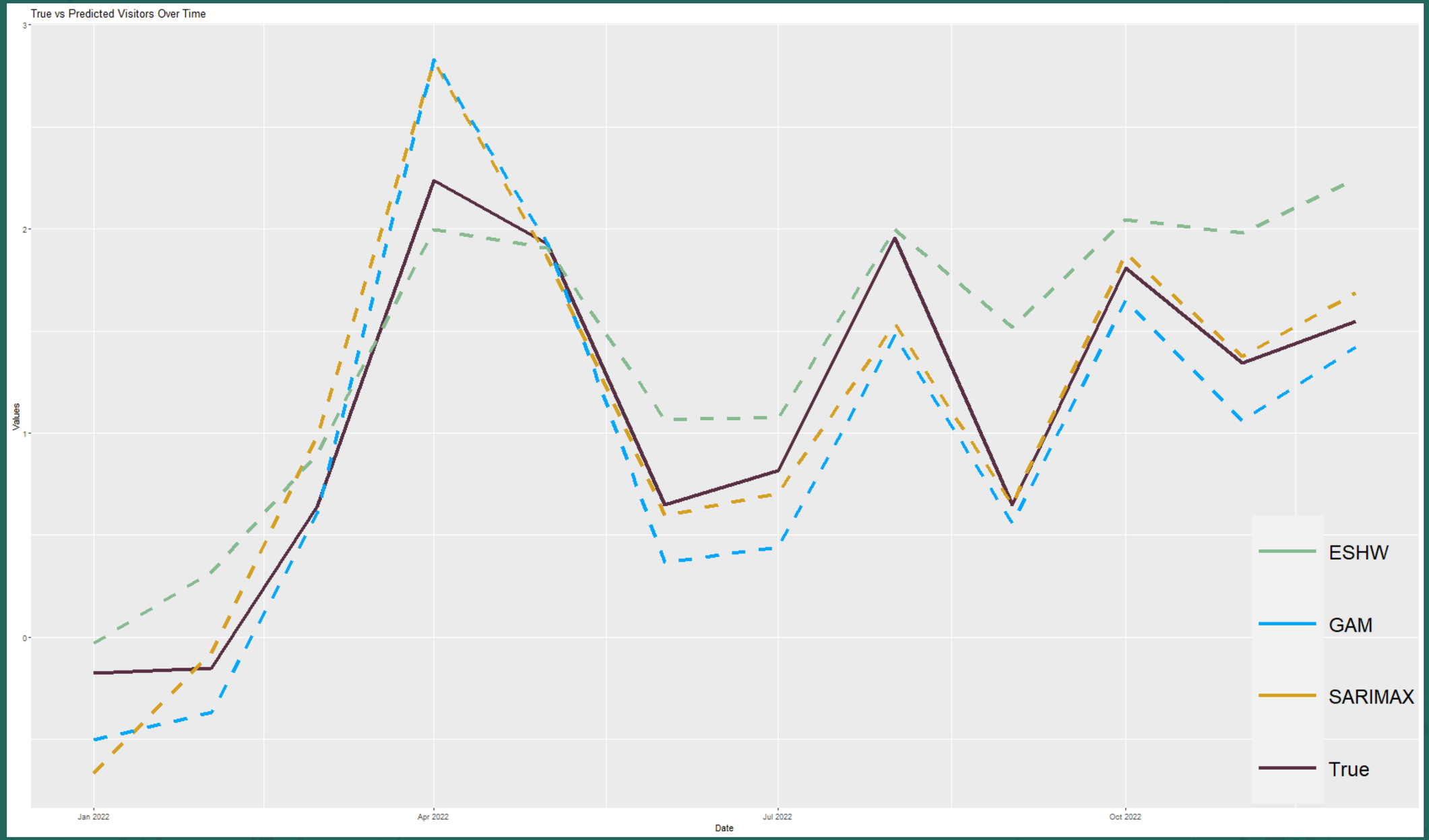
Best model Error Analysis

TRUE	PREDICTED	ERROR (%)	MONTH
64444	64501	**0.08**	September
81450	82311	1.05	November
95850	94215	1.92	May

Best predictions

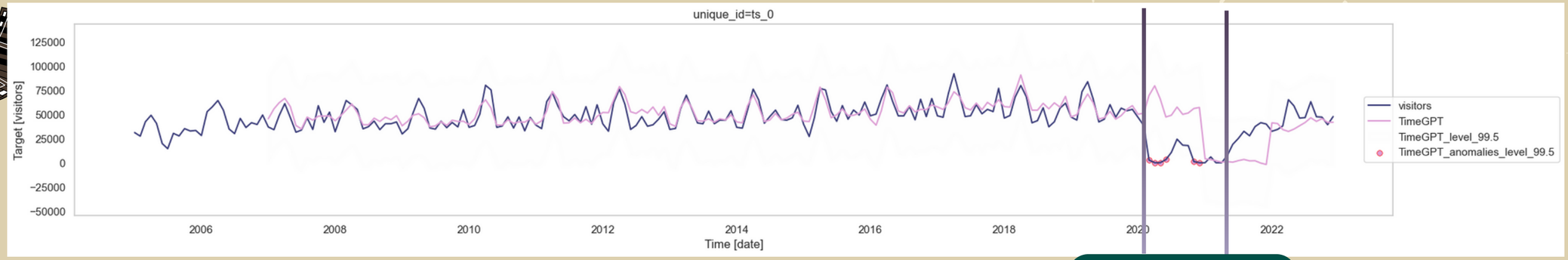
TRUE	PREDICTED	ERROR (%)	MONTH
103418	117733	**26.97**	January
103418	117733	13.84	April
64264	72815	13.31	March

Worst predictions


True vs Predicted Visitors Over Time
Values
3
2
1
0
Jan 2022
Apr 2022
Jul 2022
Oct 2022
Date
ESHW
GAM
SARIMAX
True

COVID lockdown

Effects on forecasting



Outliers

COVID lockdown

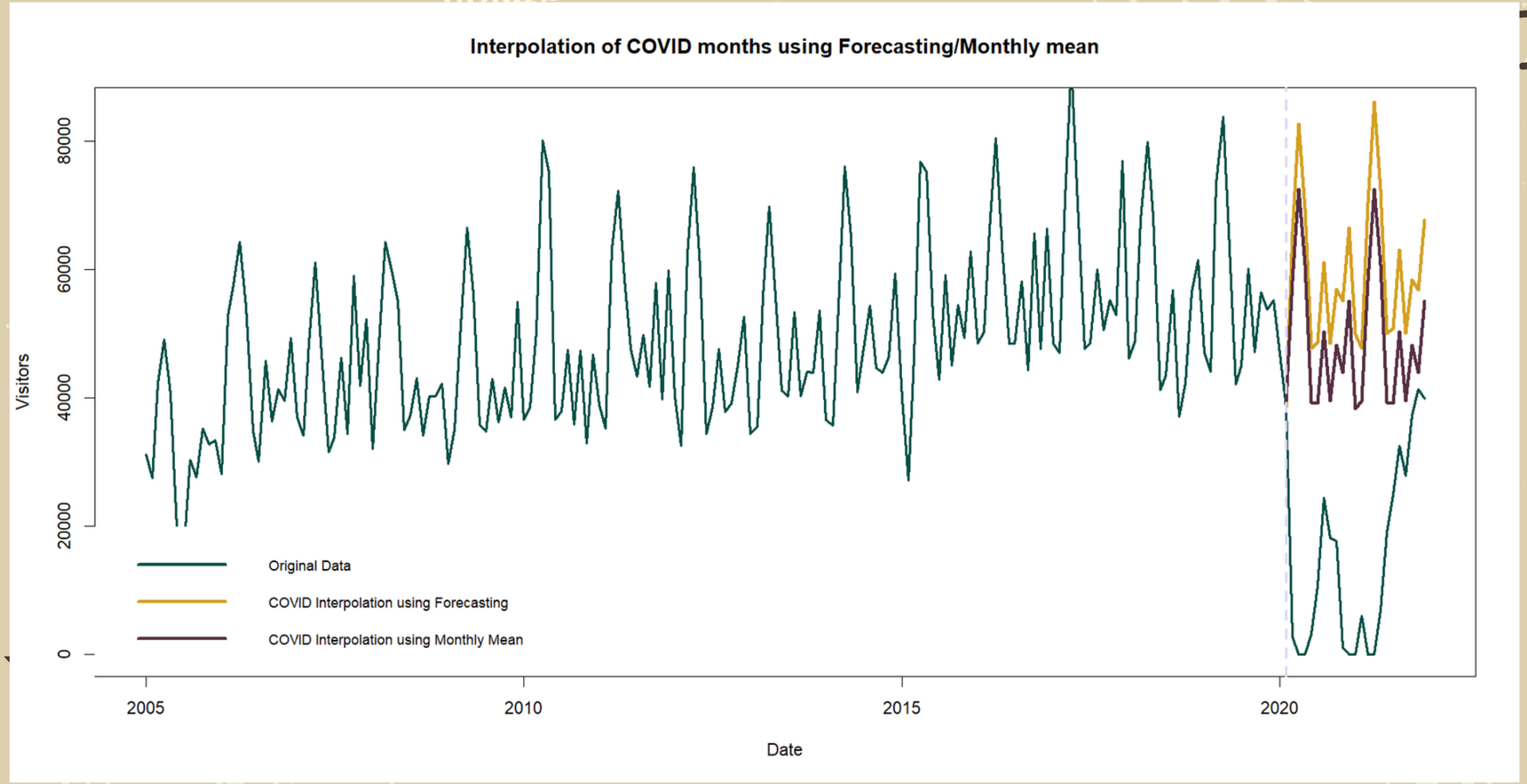
Effects on forecasting

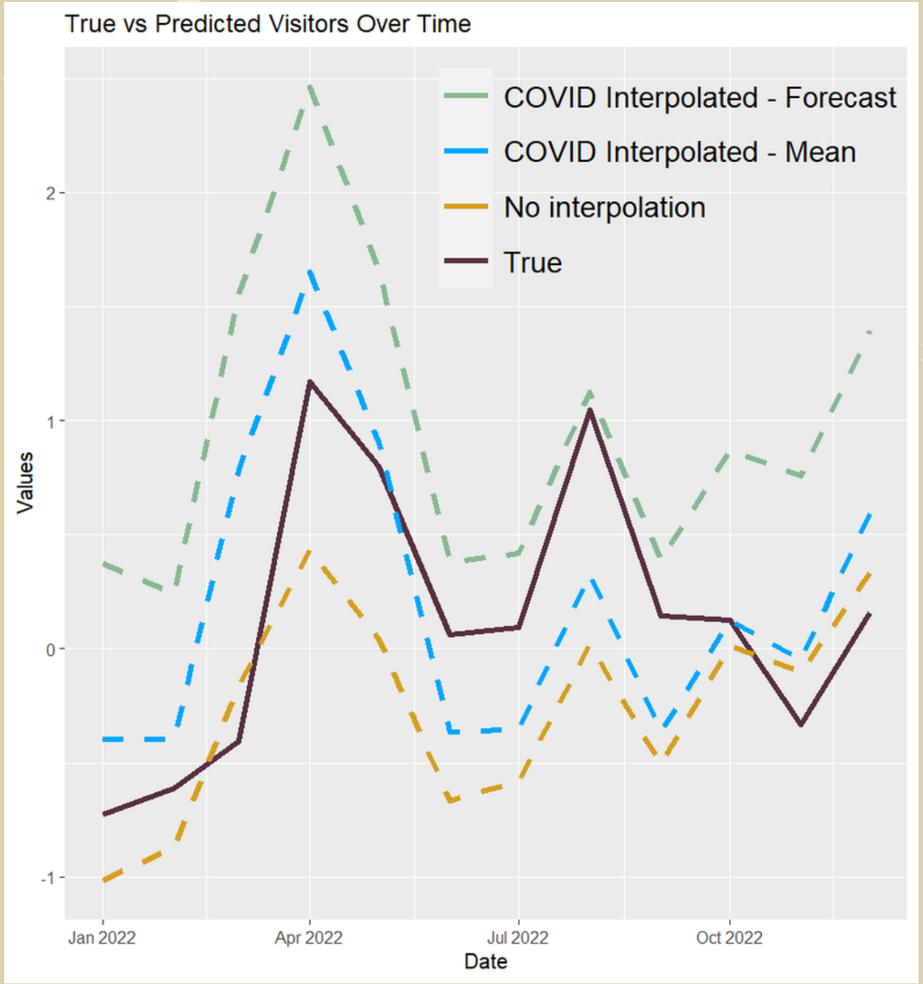
Approach: Replace outliers using **interpolation**.

1. Use a good **forecasting** model;
2. Replace each month using the historical **monthly mean**.

Then refit on the whole artificial dataset and forecast on the test set (2022).

Cinema Museum



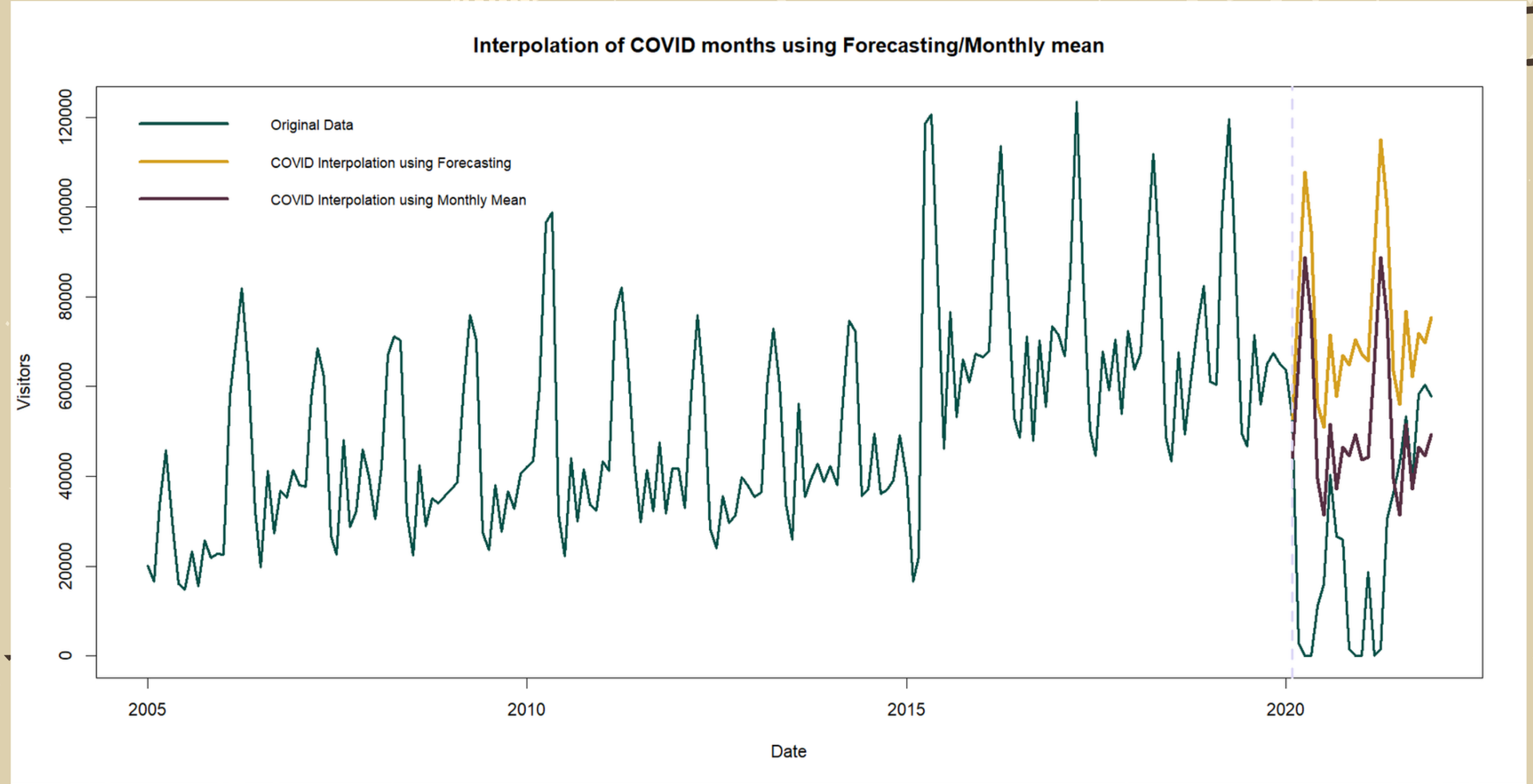


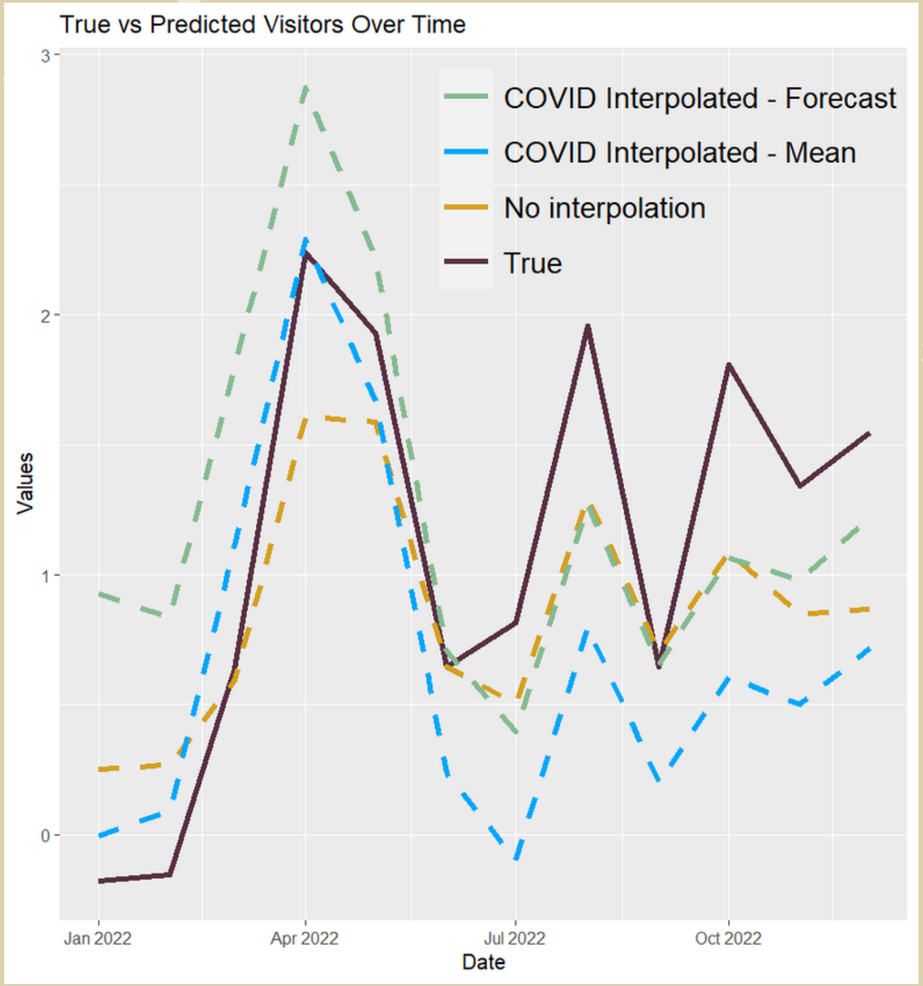
Cinema Museum

MODEL	Adj R^2	RMSE	MAPE	AIC
Best Model	0.783	**0.570**	**2.469**	**300.266**
COVID with forecasting	0.838	0.988	3.083	140.185
COVID with mean	0.780	**0.520**	**1.958**	**145.994**

Results after interpolation

Egyptian Museum





Egyptian Museum

Results after interpolation

MODEL	Adj R^2	RMSE	MAPE	AIC
Best Model	0.769	**0.468**	0.642	**314.330**
COVID with forecasting	0.850	**0.675**	1.396	**199.213**
COVID with mean	0.804	0.693	0.694	218.488

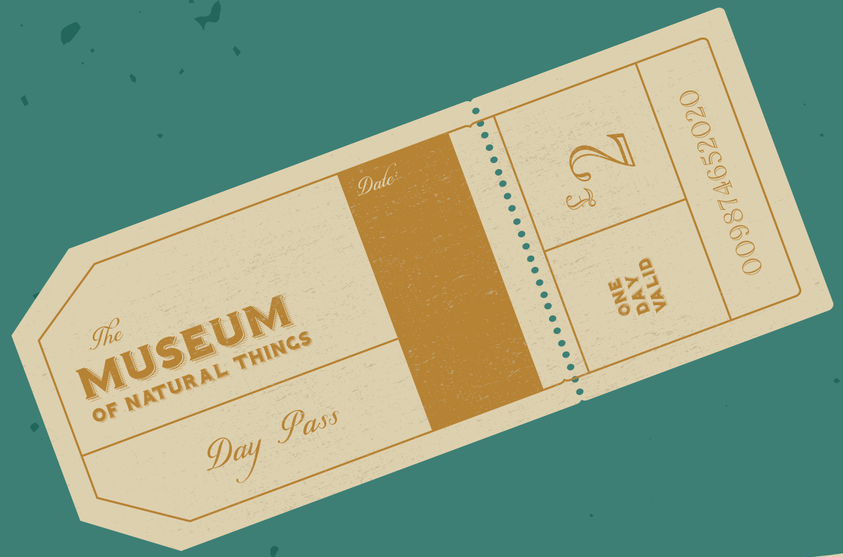
Conclusion

- The best performing models were **Holt-Winters**' exp. smoothing, **SARIMA**, **SARIMAX**.
- Significant variables:
  - **Google Trends** are significant, but not as much as for the British Museum.
  - **Arrivals** were highly significant.
  - Weather data: **average temperature** and lagged raining days.
- The interpolation of COVID period wasn't entirely successful, due to the fact that **post COVID** there was still lingering **hesitancy** among people to visit museums.

- How to better plan your travel, by avoiding queues?
  - **Most visited month**: April.
  - **Least visited months**: January, February, September.

Future work

- **Other informative variables** can be included, such as details about promotional activities, the number of museum cards purchased, etc.
- Adding the **lagged variables** made the models significantly harder to tune.
  - In-depth feature selection is needed.
  - More lags can be utilized.
- Regarding forecasting, SOTA models were utilized. Only **LSTMs** remain to be experimented with.


The
MUSEUM
OF NATURAL THINGS
Day Pass
Date
23
ONE
DAY
VALID
00987465202O

Thank you!