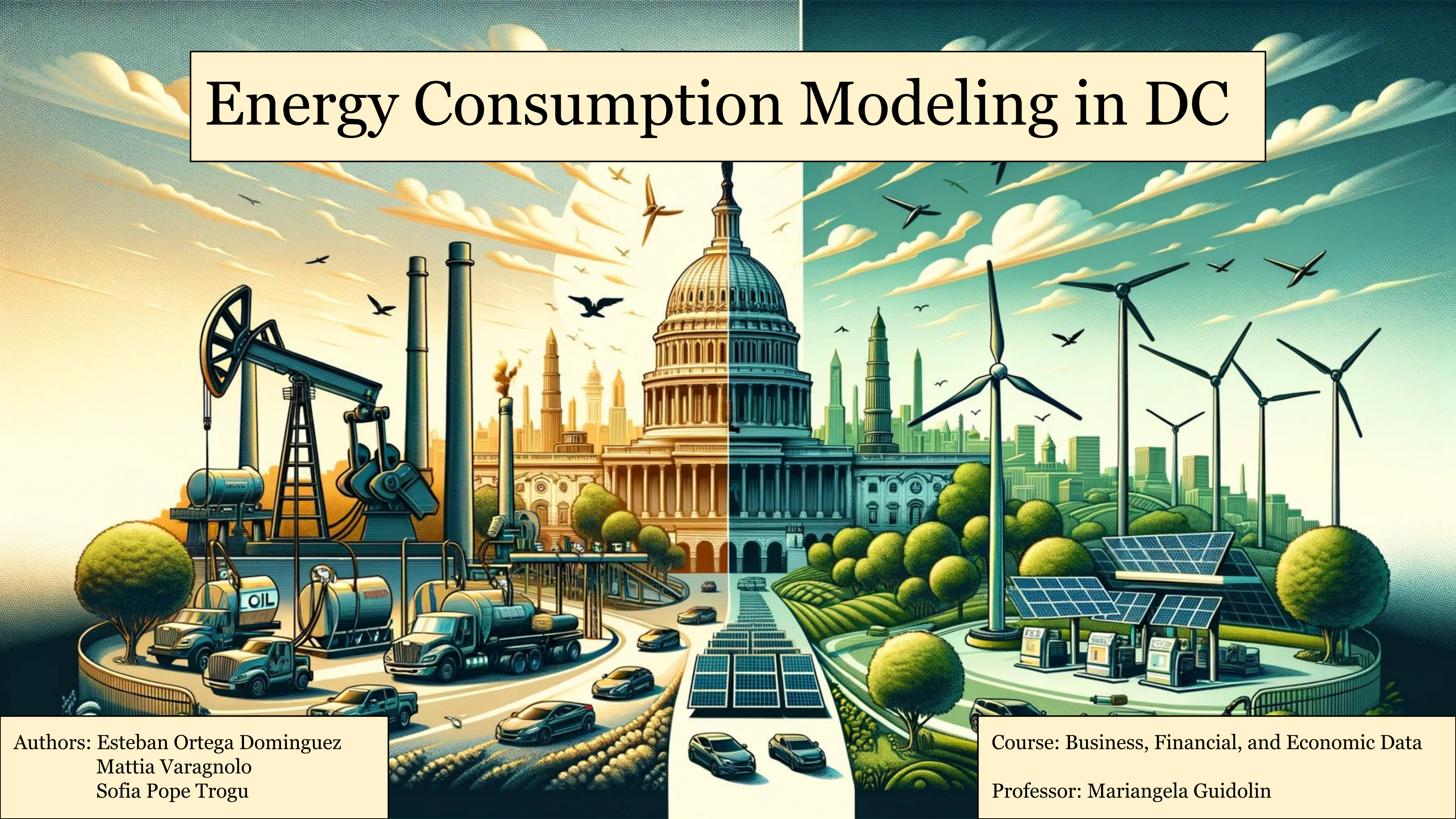
Energy Consumption Modeling in DC



Authors: Esteban Ortega Dominguez
Mattia Varagnolo
Sofia Pope Trogu

Course: Business, Financial, and Economic Data

Professor: Mariangela Guidolin

Contents

Introduction

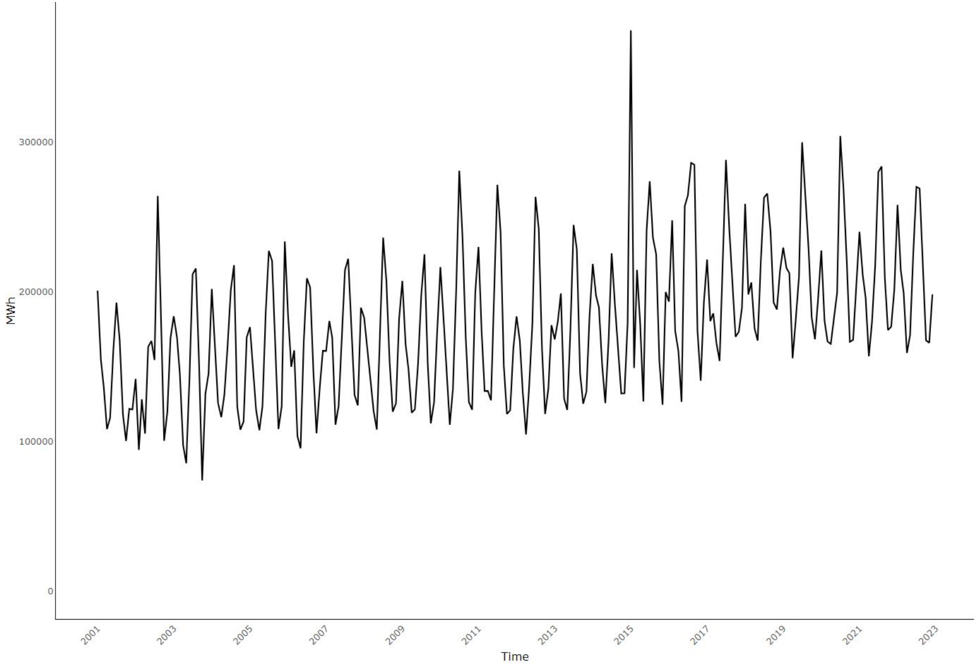
Why Washington DC?

- It’s a small state which enabled us to use regional data such as weather data
- The city experiences a range of climate conditions, from hot summers to cold winters
- Complex energy data due to the recent change in energy sources
- The city has shown a commitment to sustainable energy and various energy-related policies

Objective

Understand and forecast residential electricity sales (MWh) in Washington D.C.

Residential sales (MWh) time series



Source: EIA, U.S Energy Information Administration

Data sources

DATASET			
	Variables		**Explanation**
Energy Data	Heat and Power Commercial		Cogeneration: efficient method of energy/thermal generation from a single fuel source.
	Heat and Power Power Plants		
	Electric Generators		An independent power producer (IPP) is a non public entity that owns facilities to generate electric power for sale to utilities and end users.
	Electric Generators from IPP		
	Natural Gas		Energy sources
	Petroleum		
	Biomass		
	Solar, Thermal, Photovoltaic		
Market Data	Revenue	Residential	Thousand dollars
		Customers	
		Industrial	
		Transportation	
	Sales	Same	MWh
	Price	Same	Cents/KWh
	Customers	Same	Number of customers for each category
Weather Data	Avg temperature		°C
	Min/Max temperature		°C
	Precipitations		mm
	Wind Speed		km/h
	Pressure		hPa

Energy Data: EIA, Energy Information Administration, an independent organization doing statistics and analysis.

Market Data: EIA, Energy Information Administration

eia
Independent Statistics and Analysis
U.S. Energy Information Administration

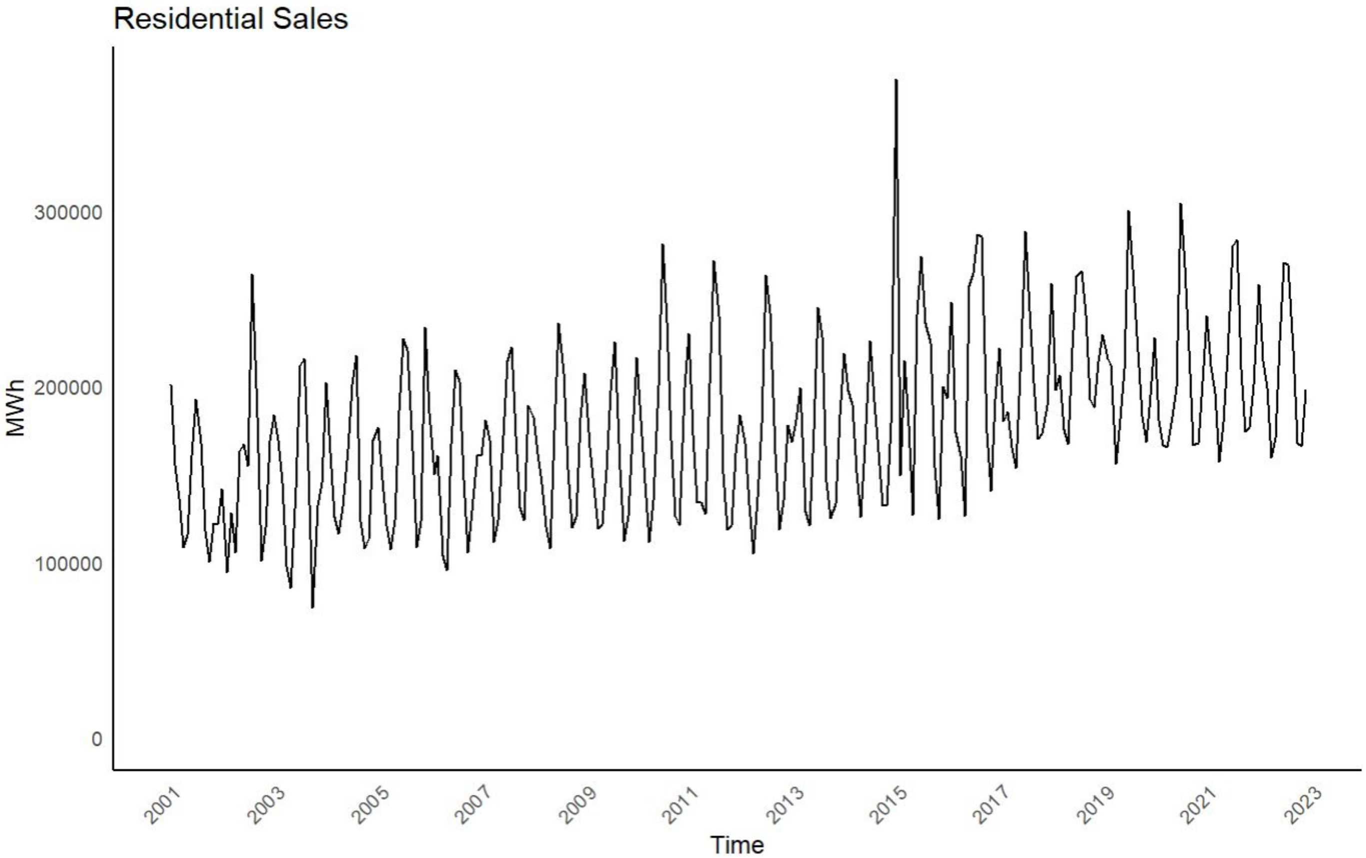
Weather Data: *Meteostat* Python package

How it works?

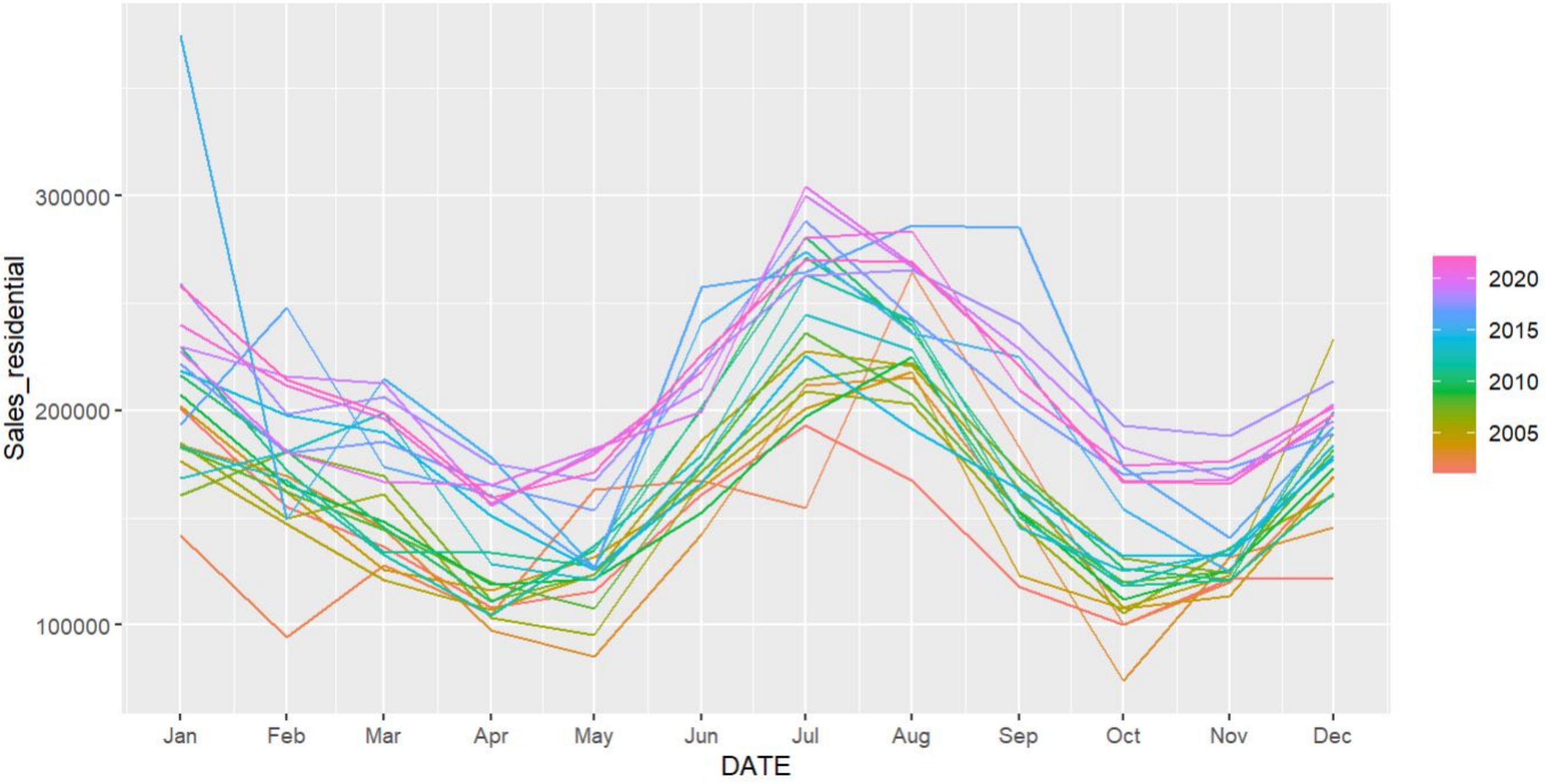
1. Select location (coordinates),
2. Range
3. start/end time
4. It localize the nearest weather stations
5. Collect data

Exploratory Data Analysis

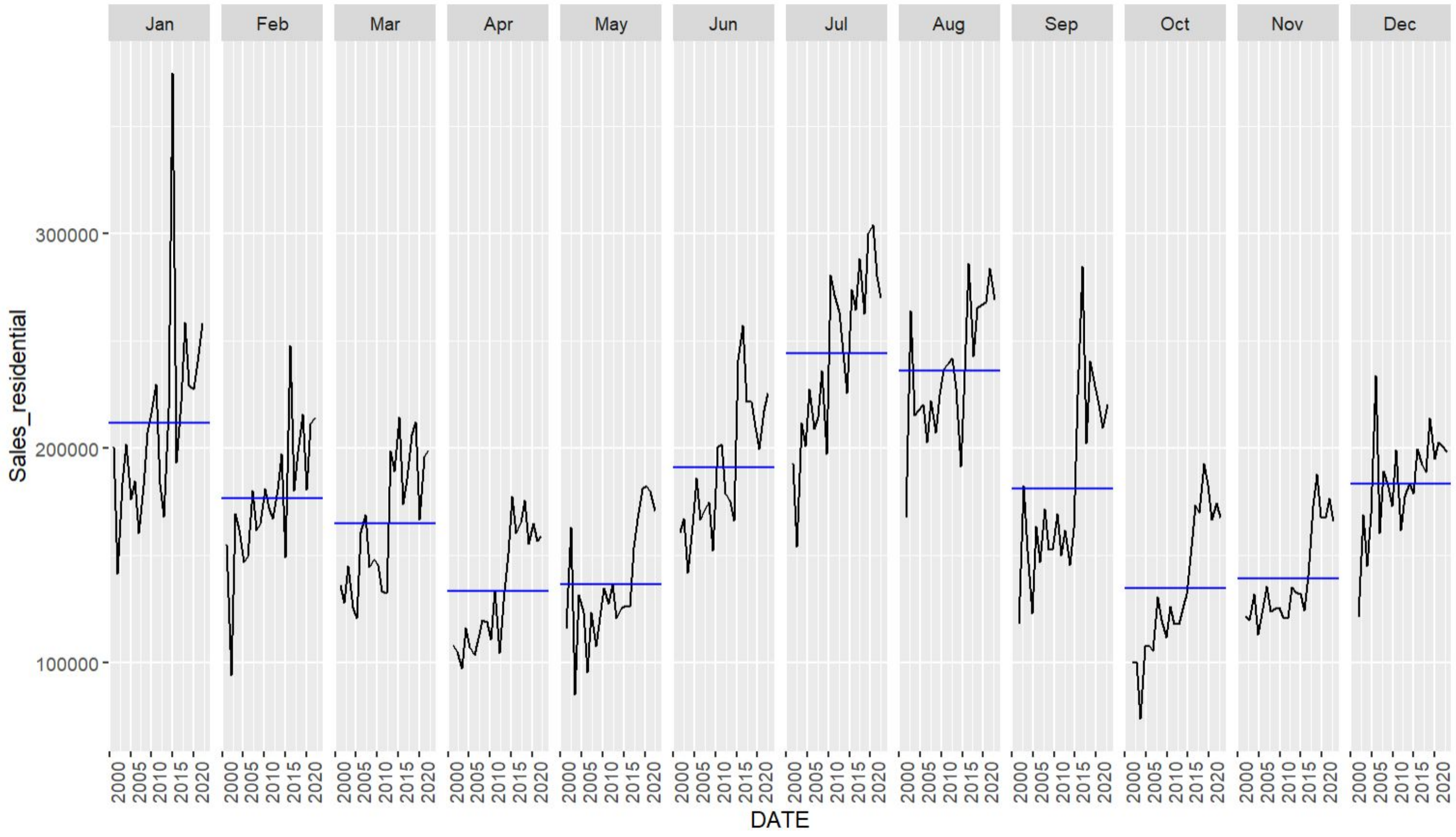
Residential sales time series



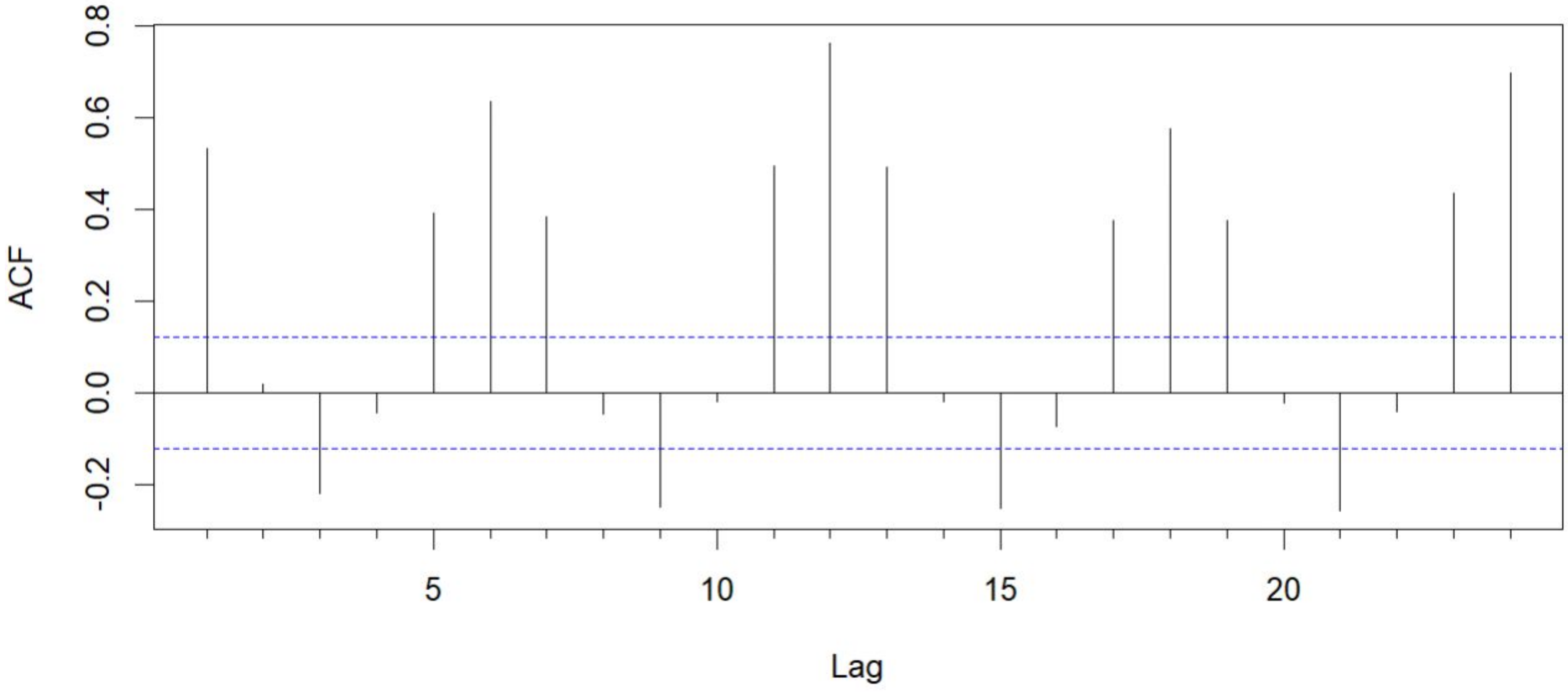
Seasonal plot



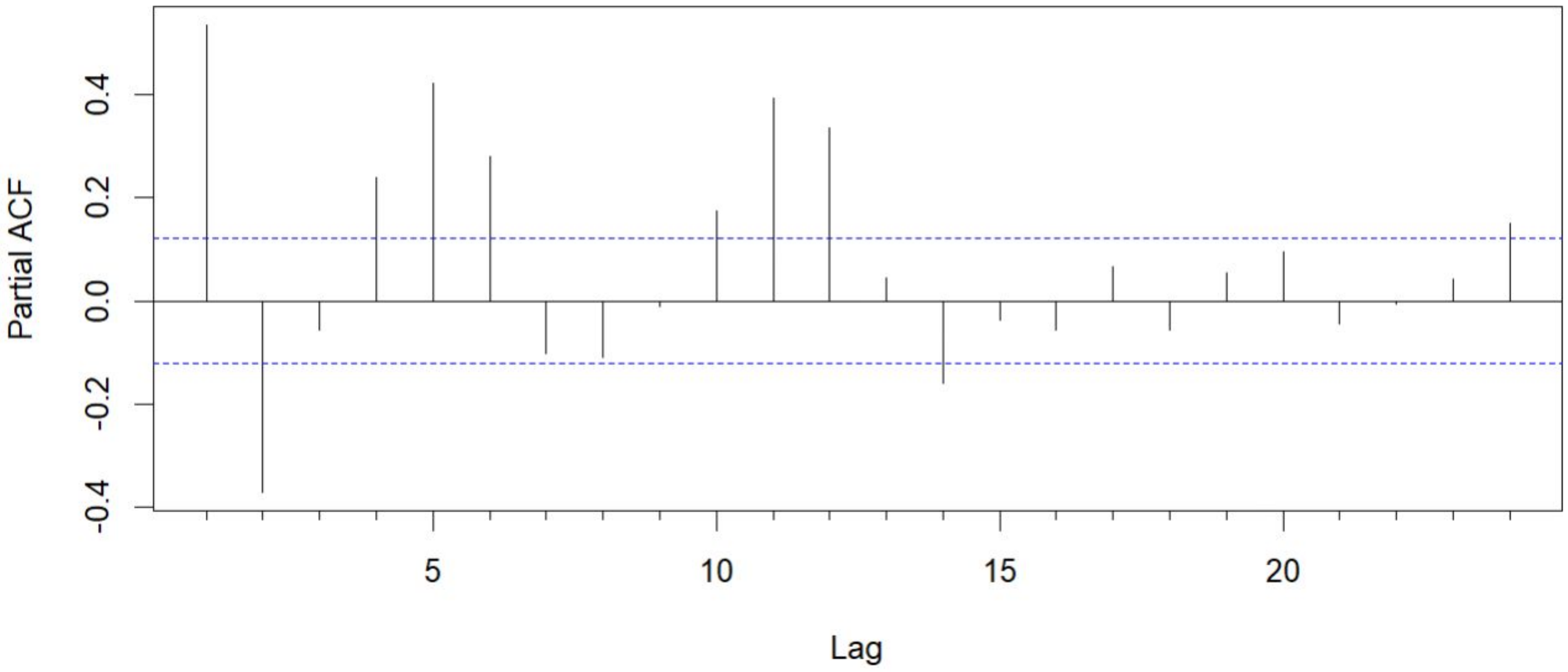
Seasonal subseries



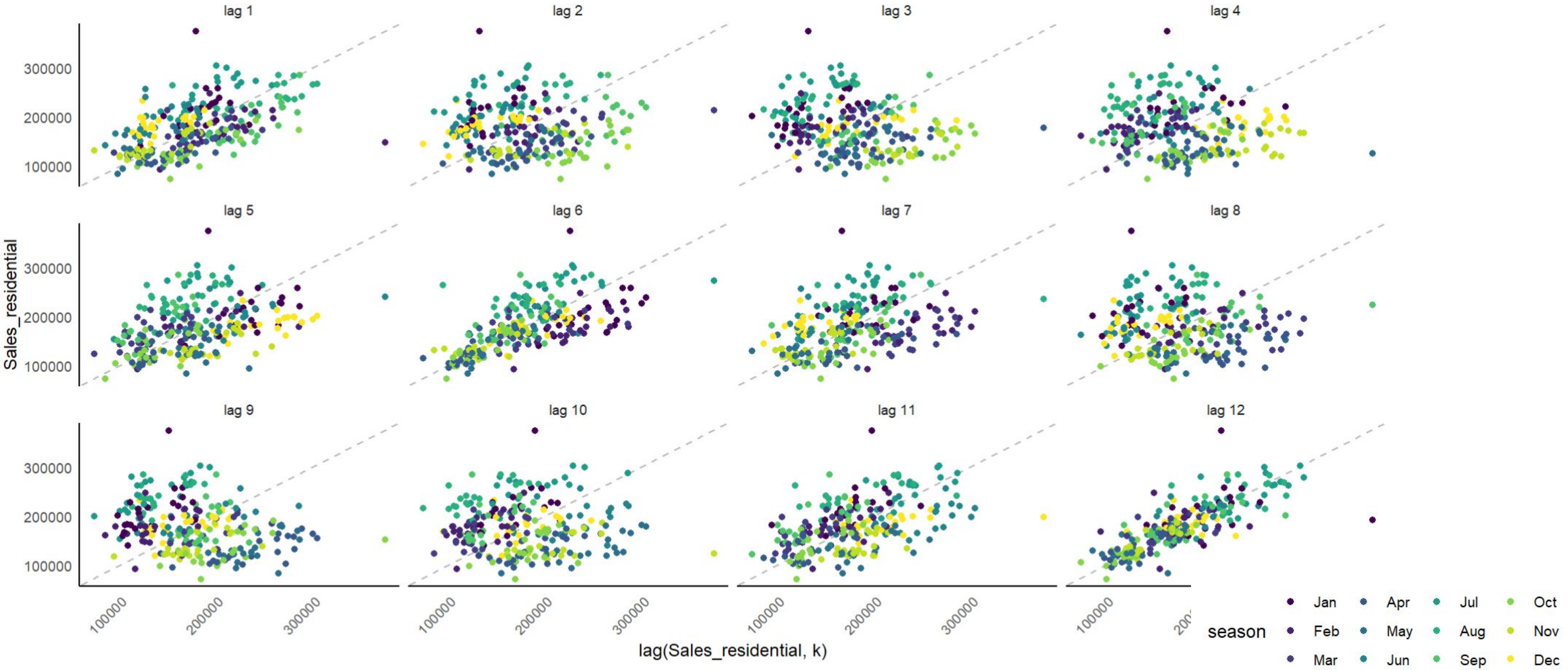
Autocorrelation for residential sales



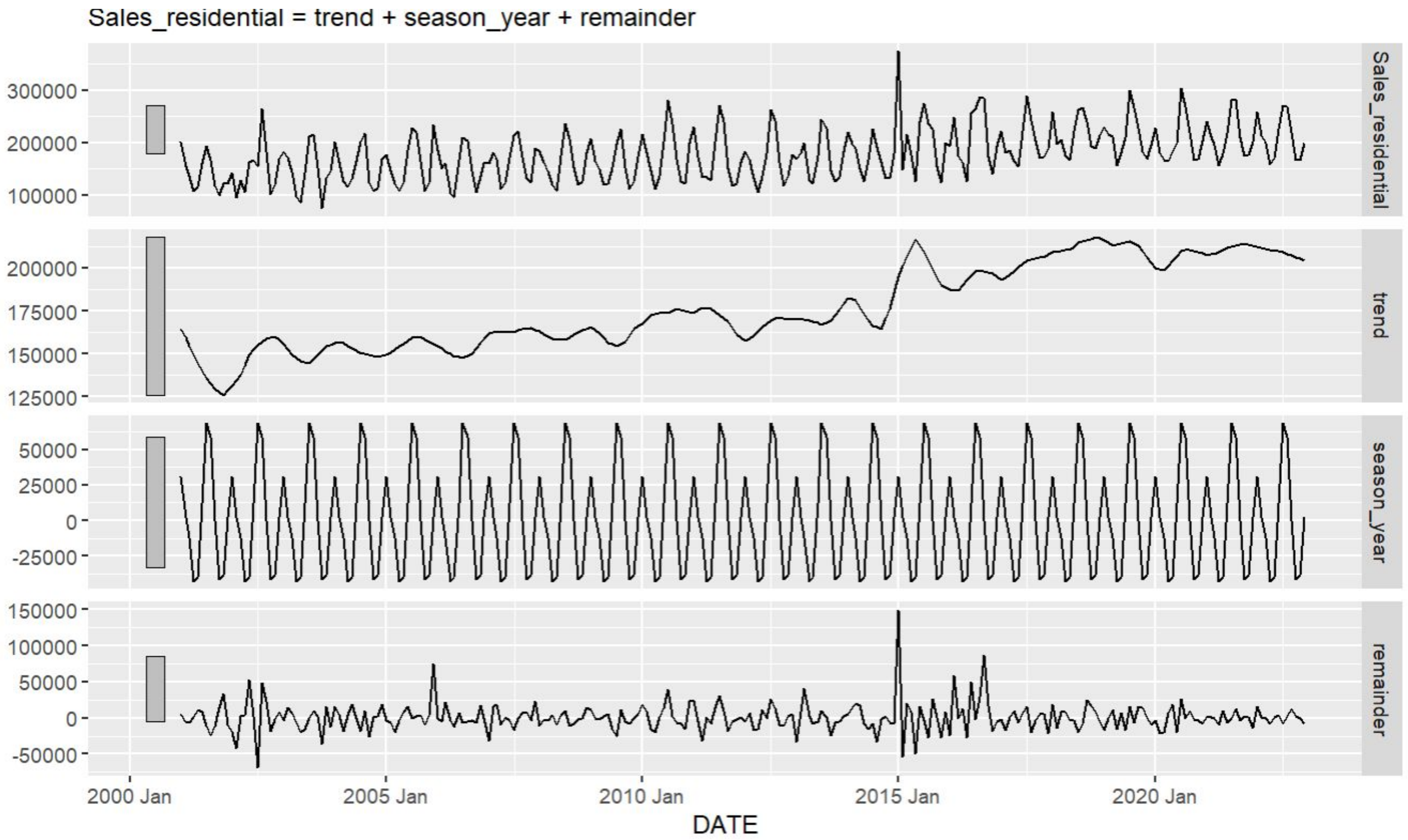
Partial autocorrelation for residential sales



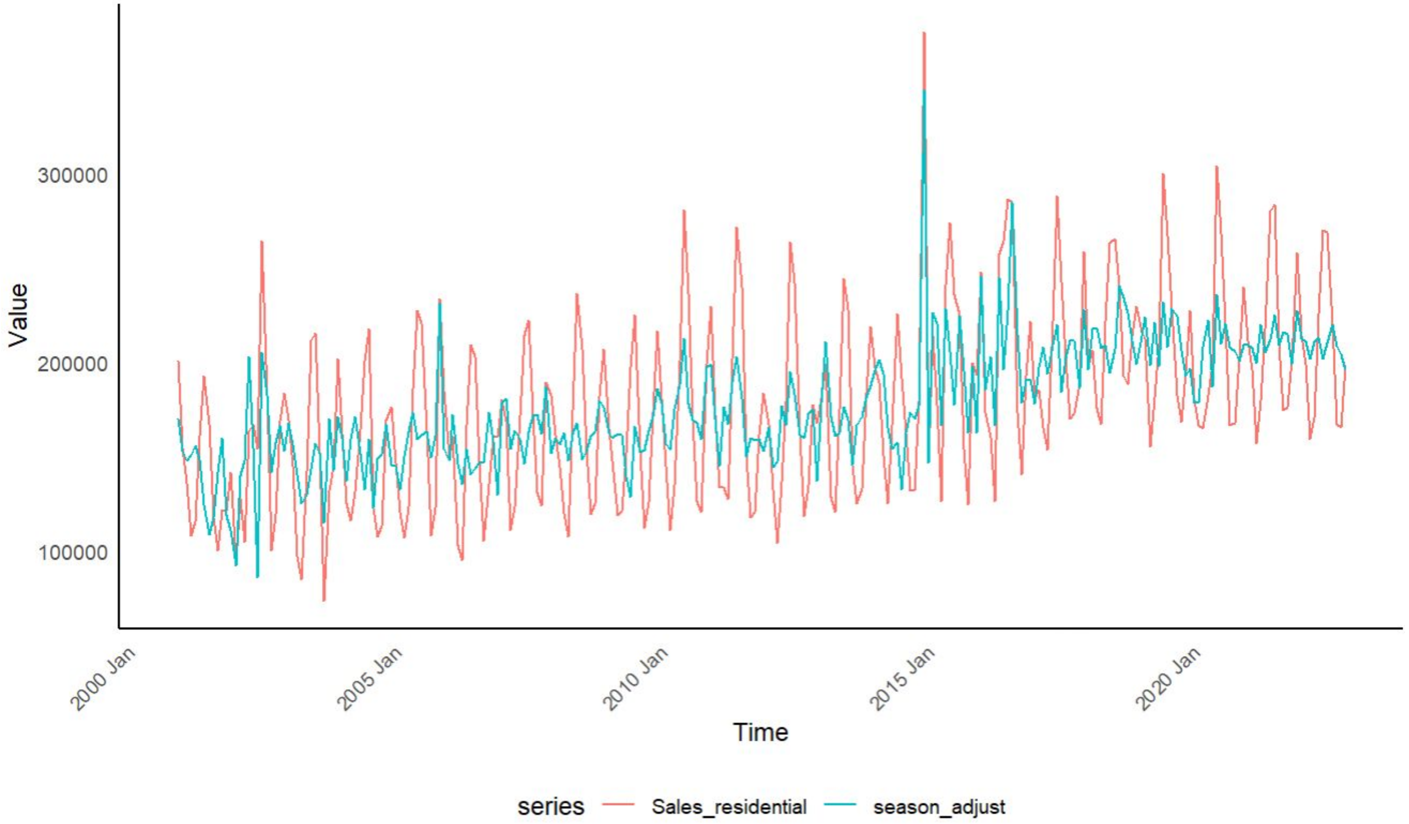
Lags plot


lag 1
lag 2
lag 3
lag 4
lag 5
lag 6
lag 7
lag 8
lag 9
lag 10
lag 11
lag 12
Sales_residential
100000
200000
300000
lag(Sales_residential, k)
season
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

STL decomposition



Seasonally adjusted values comparison

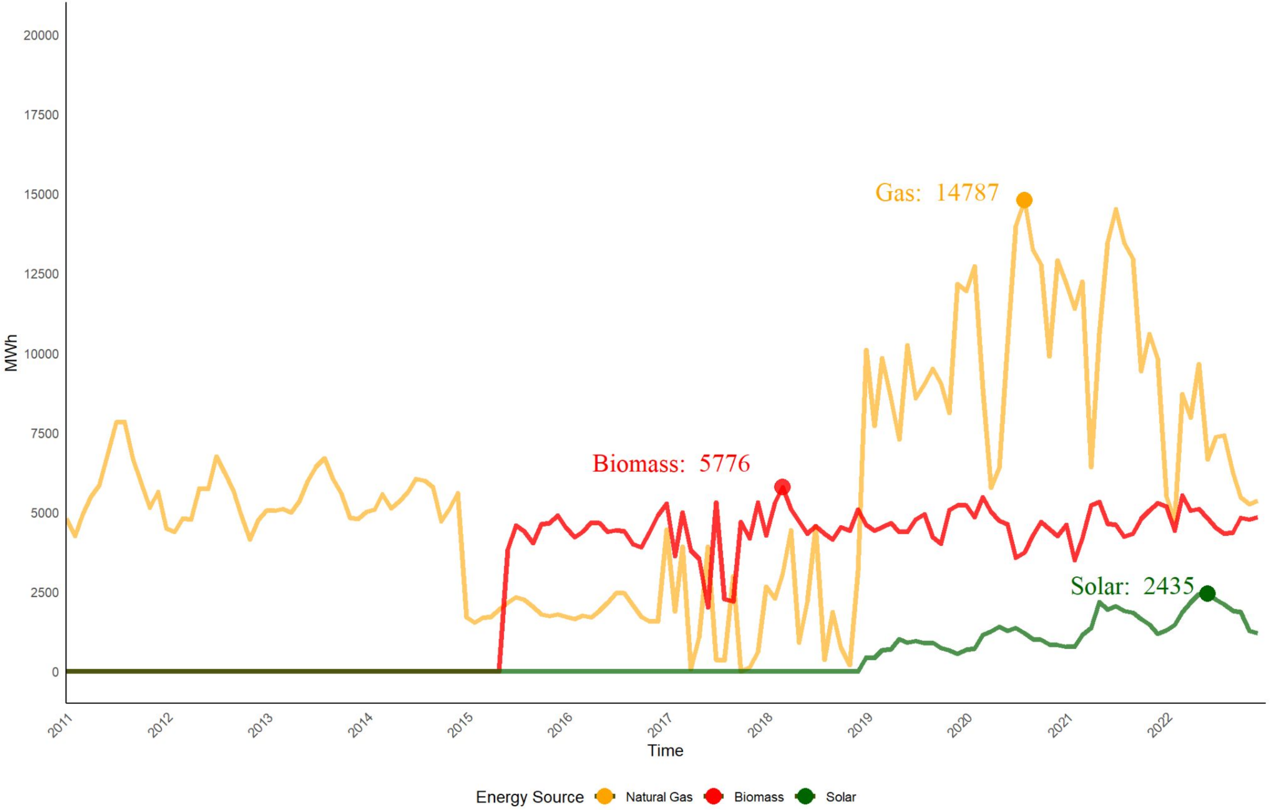


An insight on our dataset

Current generation

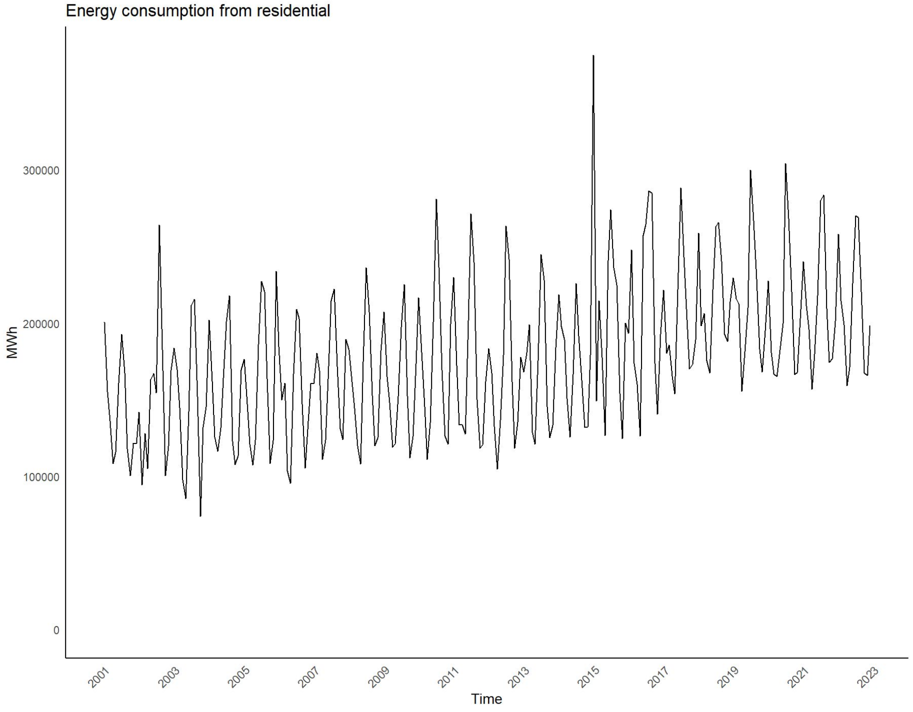
In 2019, the District claimed that 100% of the city's electricity come from renewable sources by 2032, including at least 5.5% from solar energy.

Target: Sales Residential	
Variables	Correlation
Biomass	0.433
Gas	0.361
Solar, thermal	0.289



An insight on our dataset

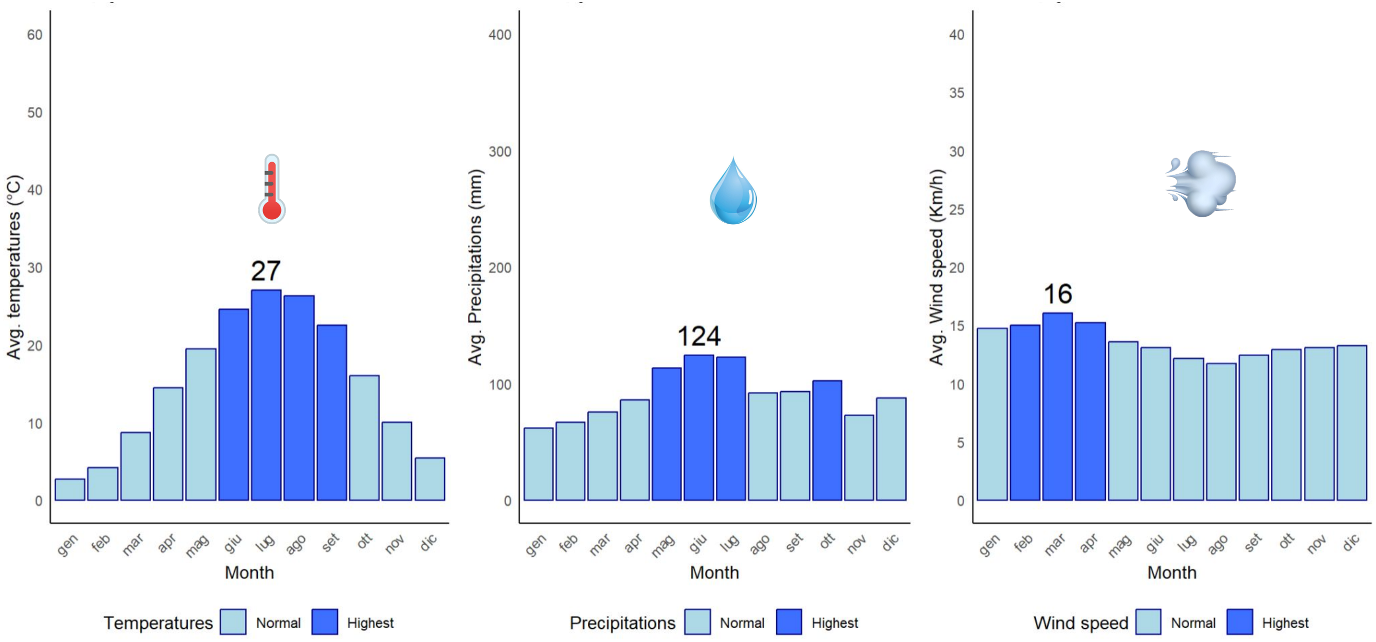
Market Data



Target: Sales Residential	
Variables	Correlation
Customers transportation	0.457
Biomass	0.433
Customers residential	0.401
Heat and power (commercial) cogeneration	0.396
Gas	0.361
Price Residential	0.339
Customers commercial	0.336
Solar, thermal, and photovoltaic	0.289
tmin	0.284

An insight on our dataset

Weather data



Target: Sales Residential	
Variables	Correlation
tmin	0.284
tavg	0.257
tmax	0.240
wspd	-0.204
prcp	0.105
pres	-0.098

Modelling

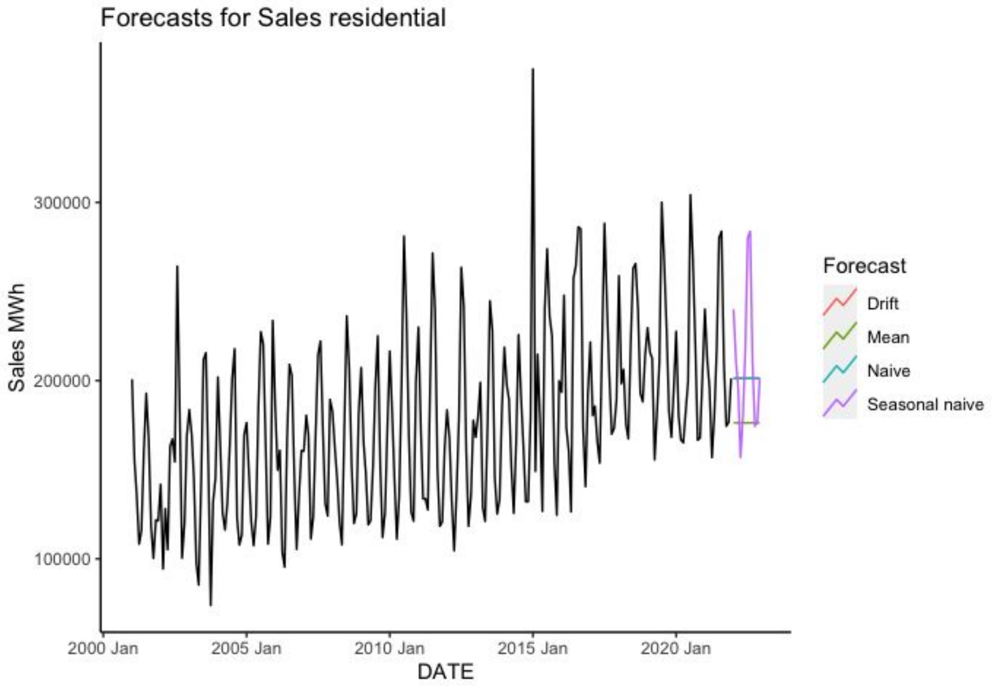
1. Benchmark models
2. Linear regression
3. Multiple linear regression
4. Holt-Winters exponential smoothing
5. ARIMA
6. KNN
7. Gradient Boosting
8. GAM

Train and test set

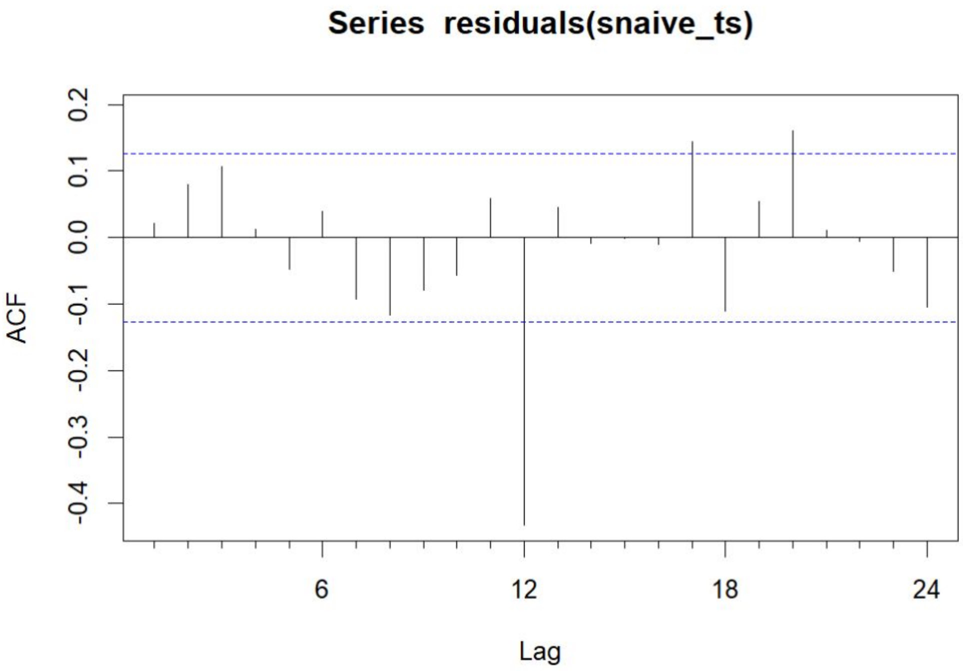
- Model fitting on pre 2022 and forecasting and testing on 2022

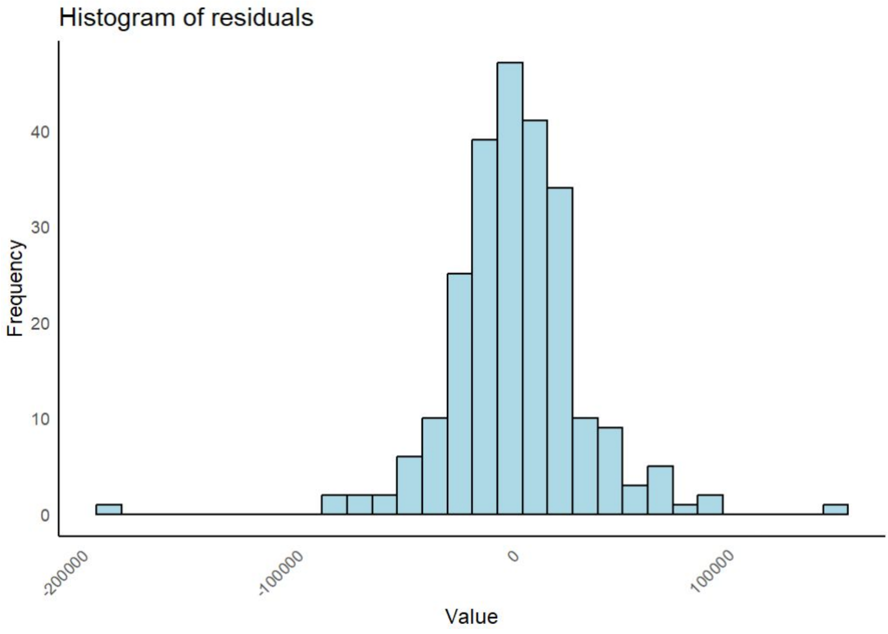
Benchmarks: drift, mean, naive, seasonal naive

.model	ME	RMSE	MAE	MPE	MAPE	ACF1
Drift	8,666.49	39,612.47	33,123.80	0.86	15.57	0.51
Mean	33,614.06	51,223.47	40,570.13	13.15	17.39	0.51
Naive	8,678.17	39,613.76	33,122.00	0.86	15.57	0.51
Seasonal naive	-747.67	9,587.53	8,307.67	-0.55	3.87	-0.09

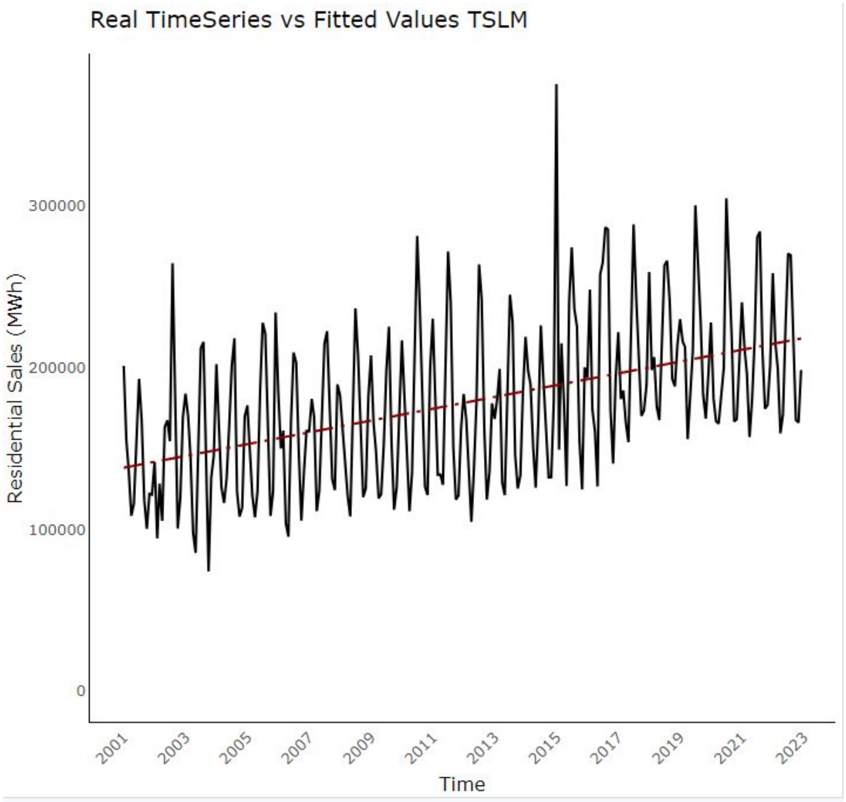


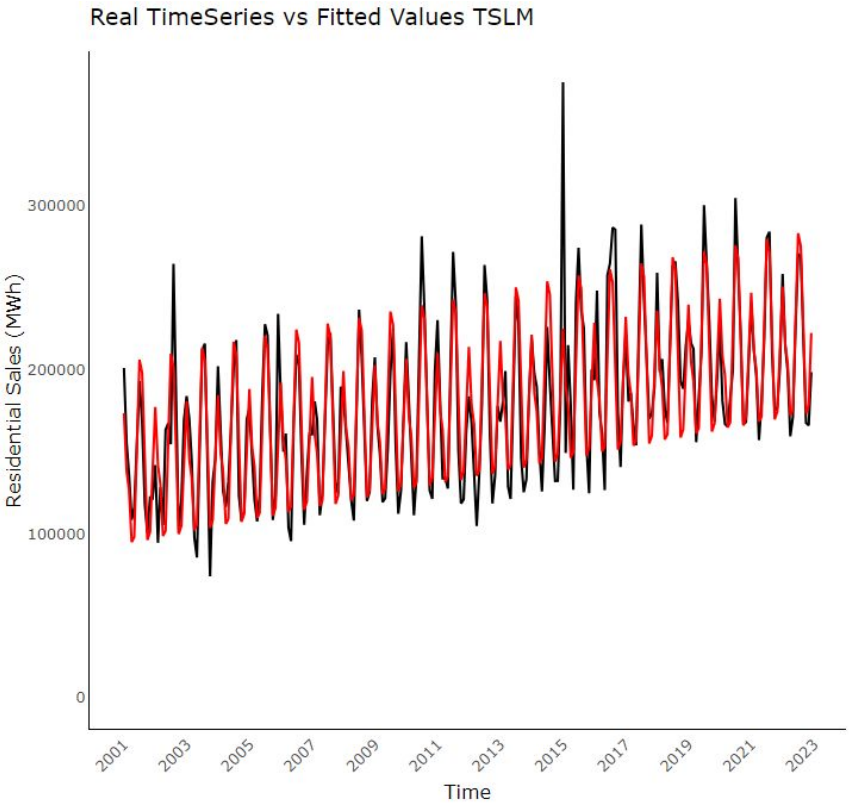
SNAIVE residuals analysis



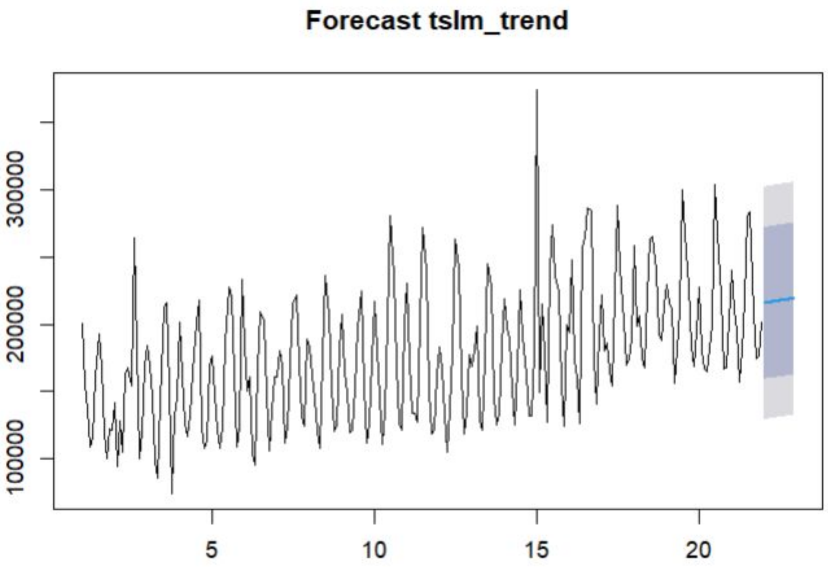


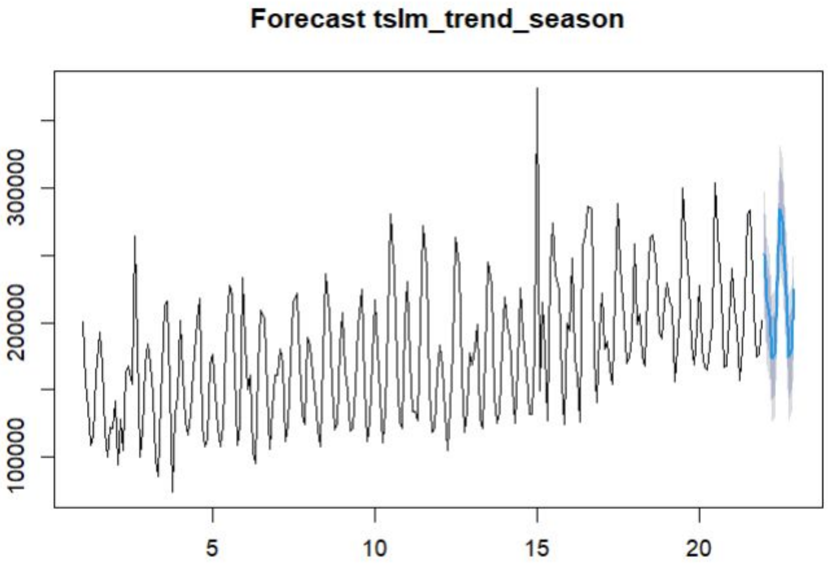
Linear models: trend and trend + season



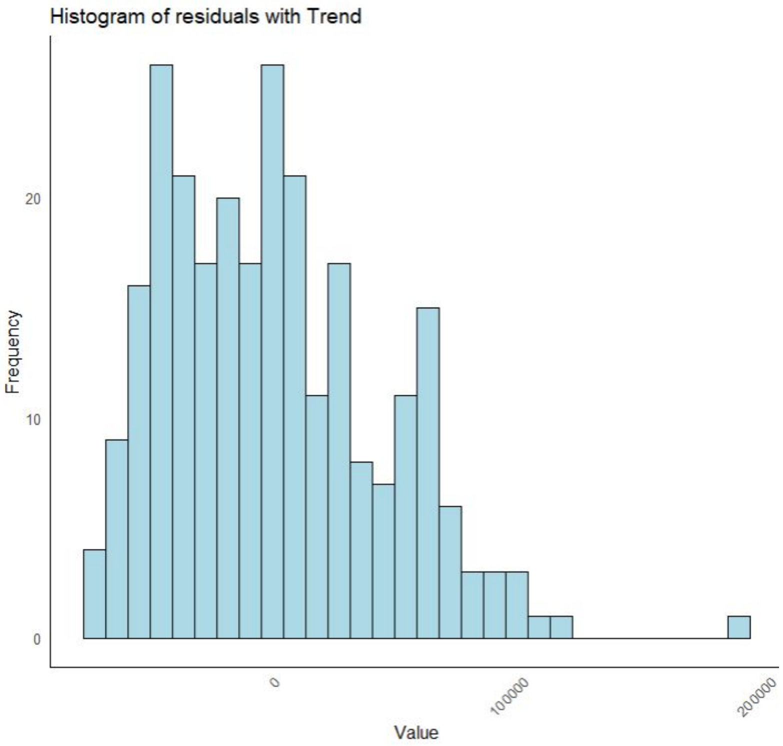


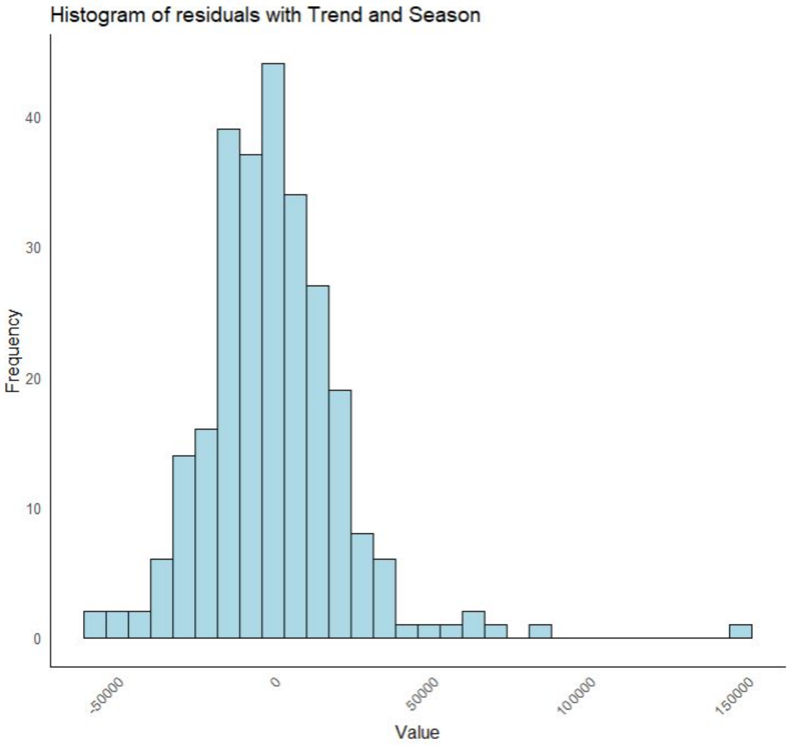
Forecast of linear model: trend and trend + season





Linear models: Residuals histogram

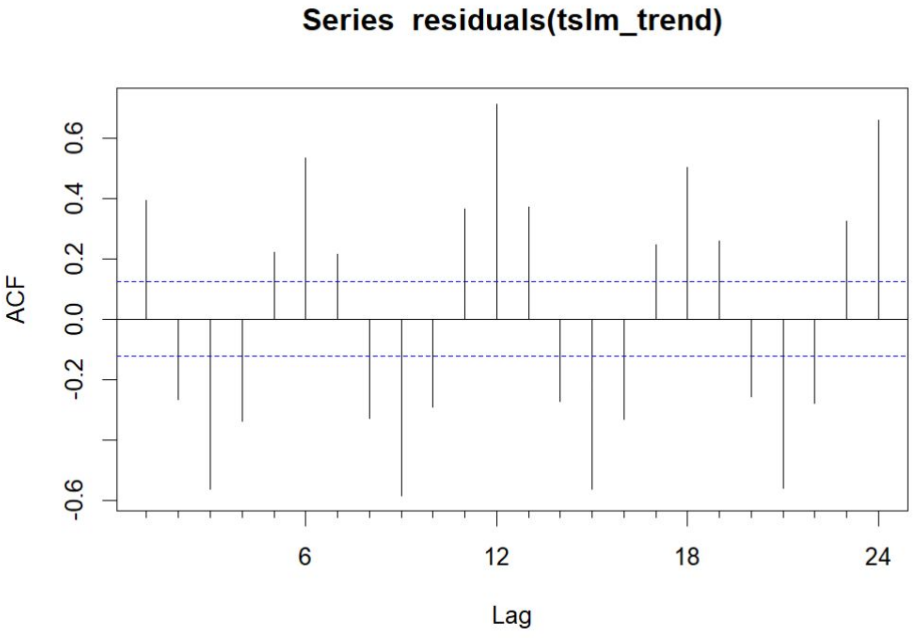


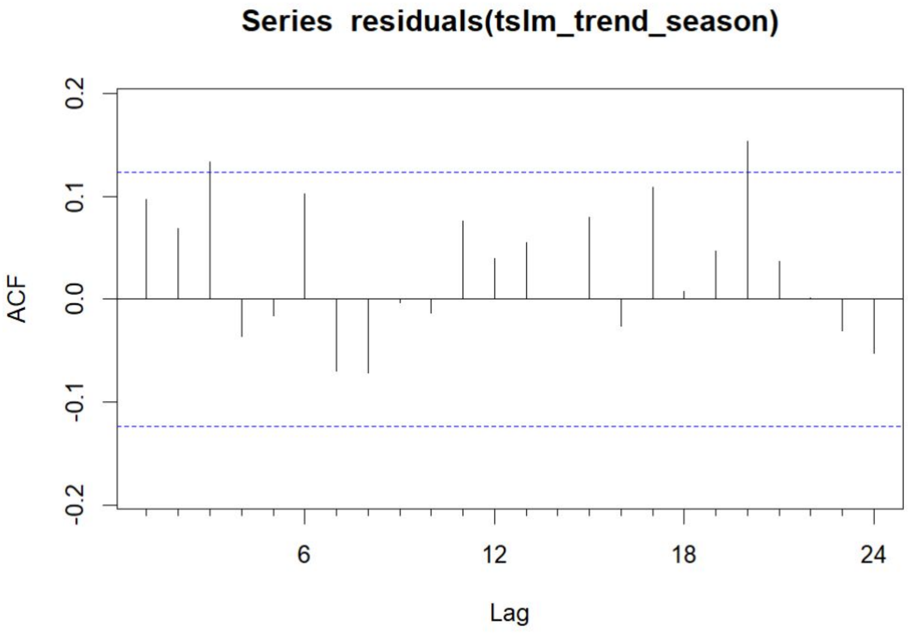


Durbin Watson = 1.199, p-value = 0.00000000002131

Durbin Watson = 1.7889, p-value = 0.04598

Linear models: ACF residuals





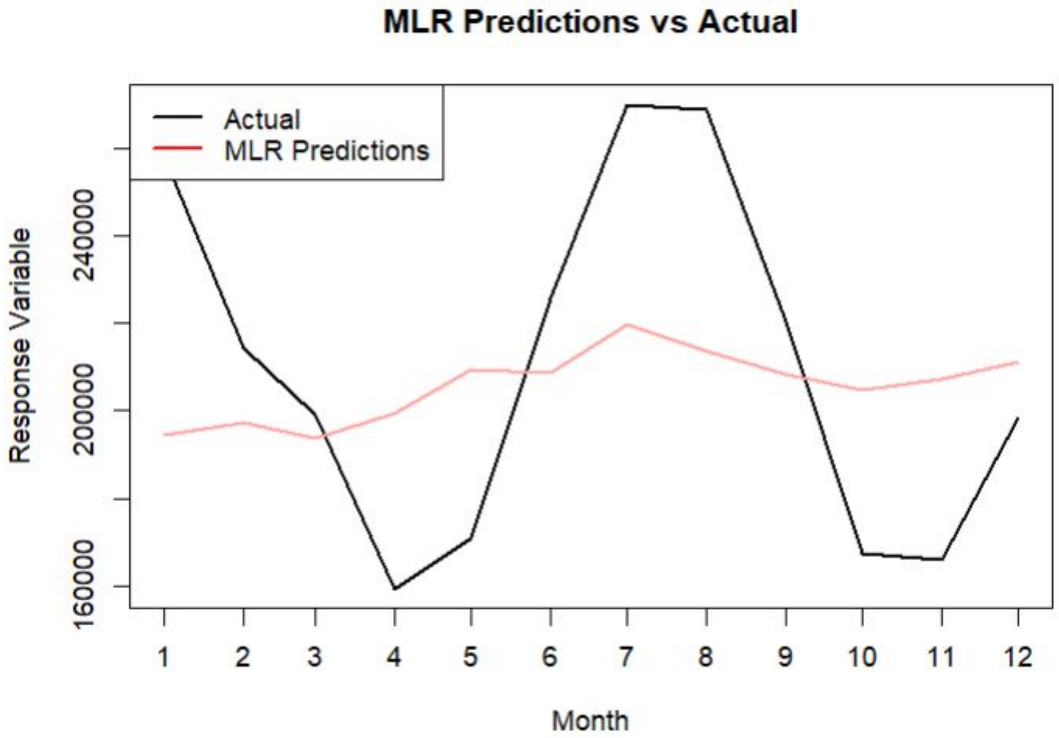
Multiple linear regression variable selection

- ➢ Variable selection through AIC stepwise selection
- ➢ Multiple collinearity reduction through max VIF variable suppression (VIF = 7 threshold)

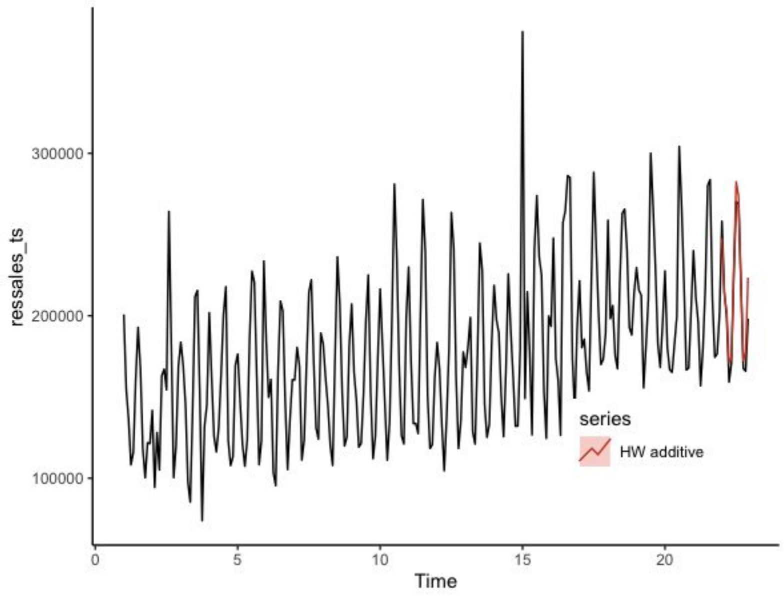
Variable	P_Value
(Intercept)	0.000001
Electric.Generators..Independent.Power.Producers	0.000040
Solar.Thermal.and.Photovoltaic	0.894031
Price_commercial	0.300827
Price_industrial	0.135612
Customers_transportation	0.000000
Price_total	0.654119
tavg	0.443434
prcp	0.472792
wspd	0.062141

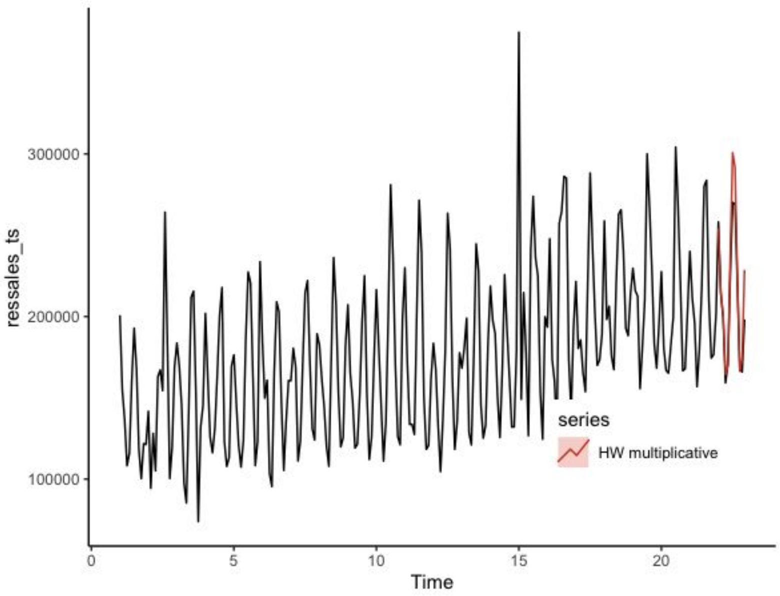
* Full train data

Multiple linear regression forecast



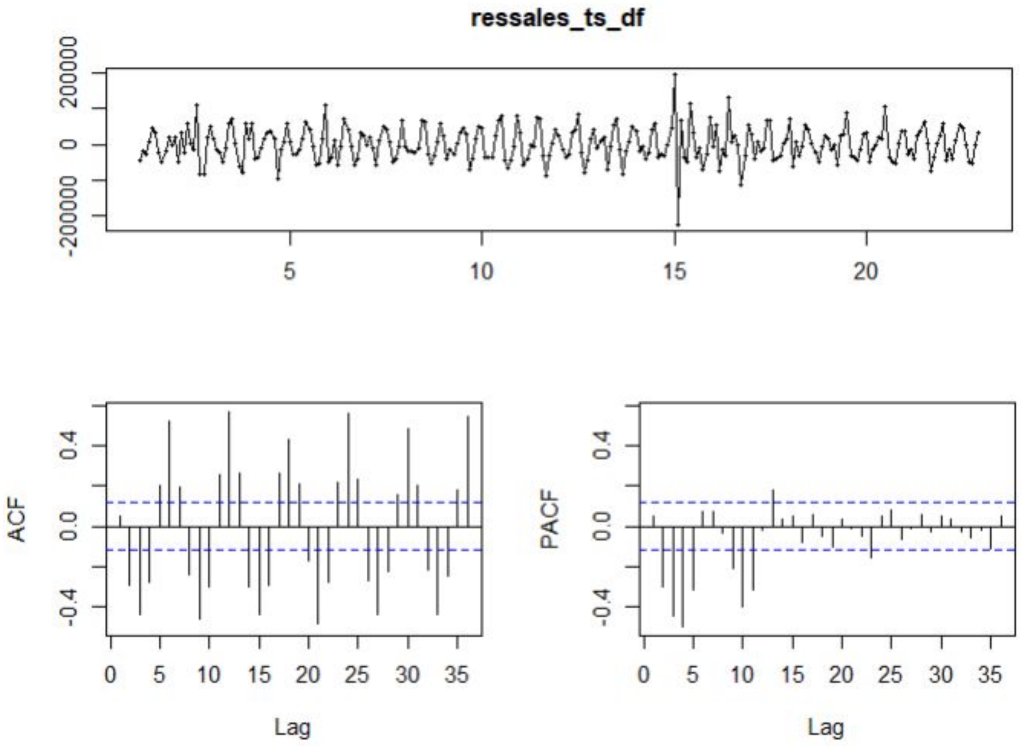
Holt-Winters exponential smoothing method



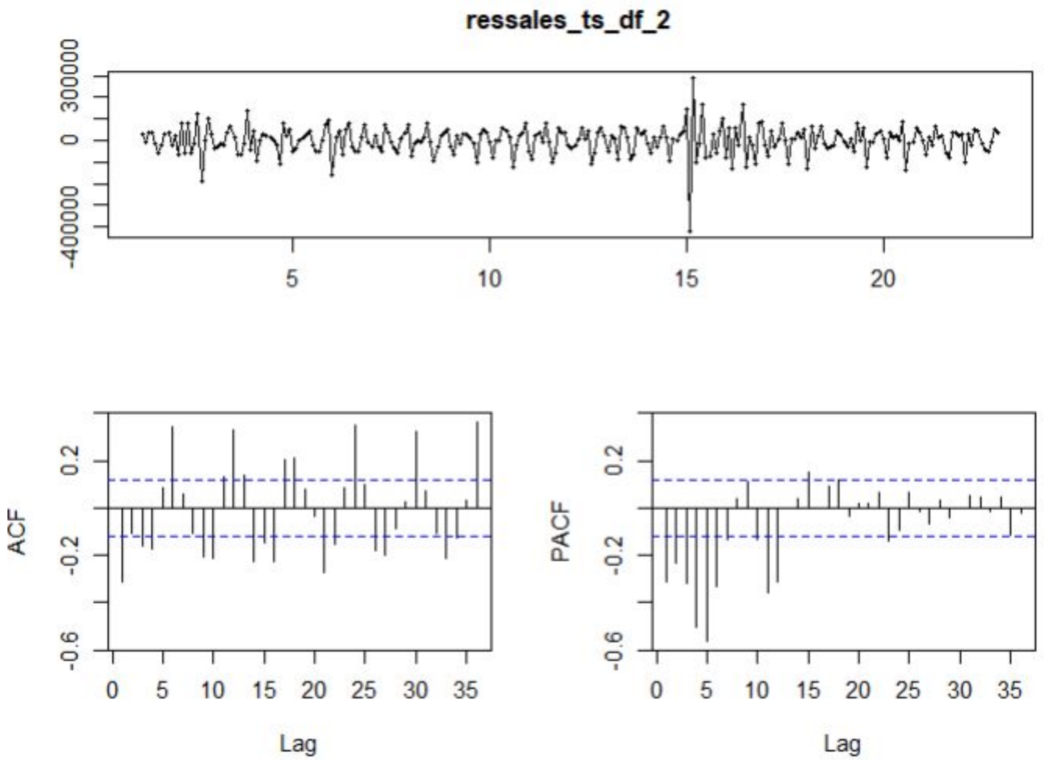


.model	ME	RMSE	MAE	MPE	MAPE	ACF1
HW additive	-5,623.42	10,236.63	7,739.02	-3.02	3.87	0.05
HW multiplicative	-8,836.07	14,754.12	9,883.12	-3.94	4.43	0.08

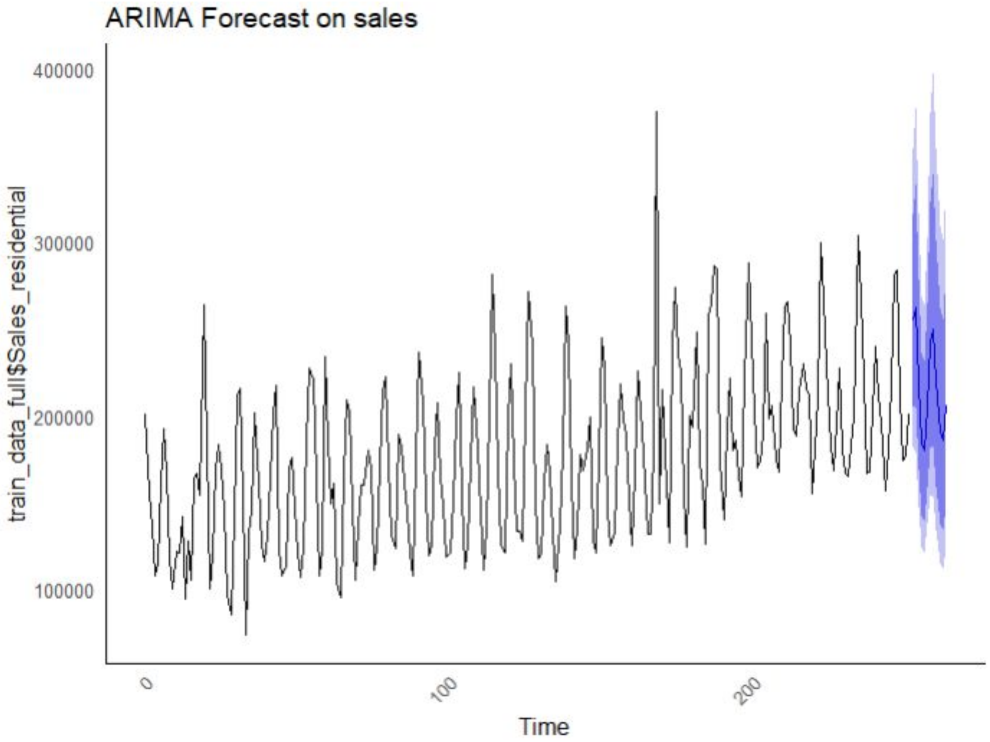
ARIMA: differencing

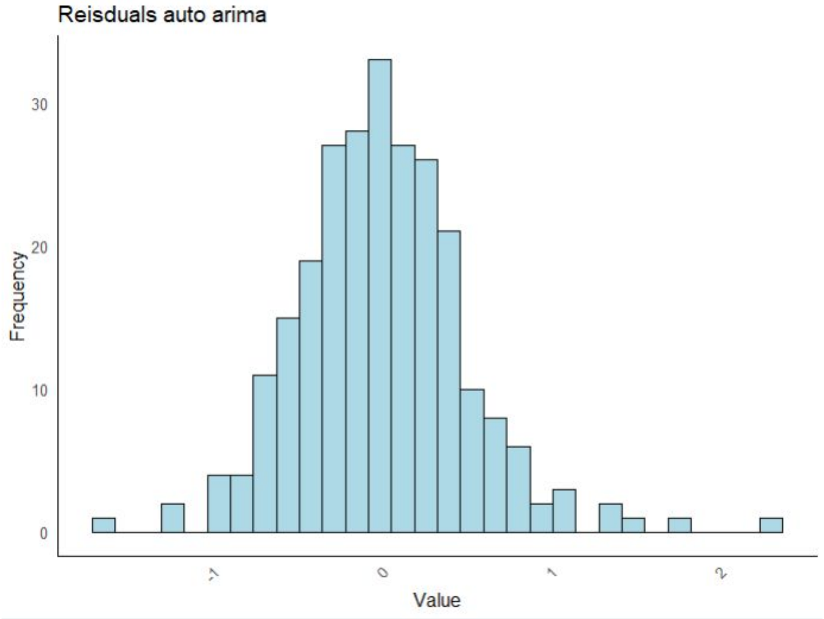


ARIMA: second differencing



ARIMA forecasting (auto arima)



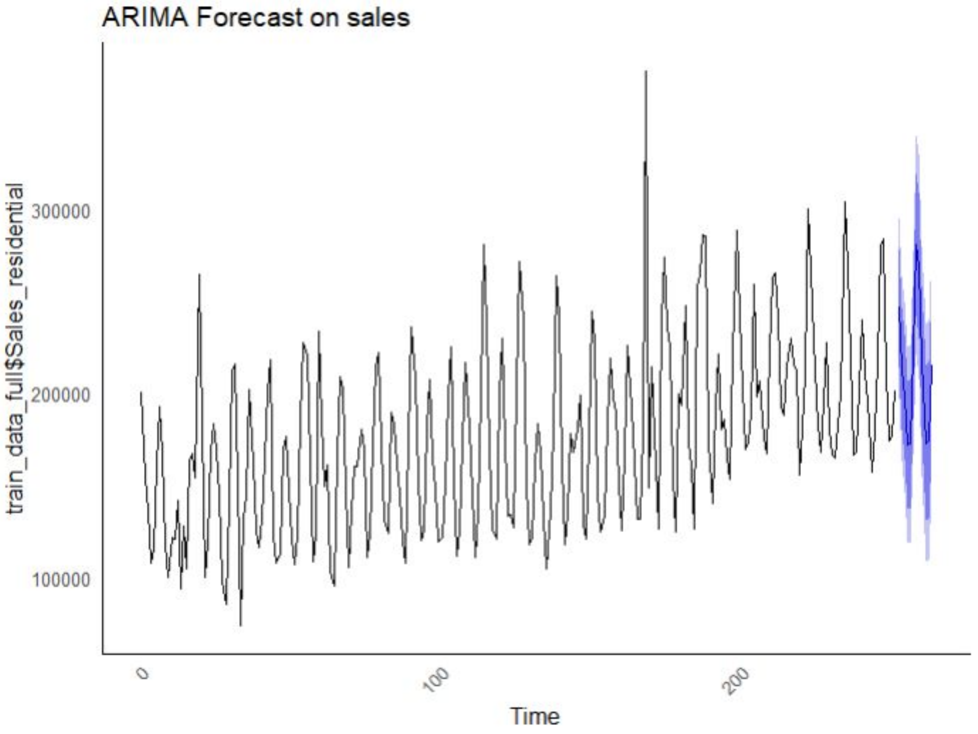


.model	ME	RMSE	MAE	MPE	MAPE	ACF1
ARIMA	-6,853.93	22,223.89	18,591.54	-4.63	9.24	0.02

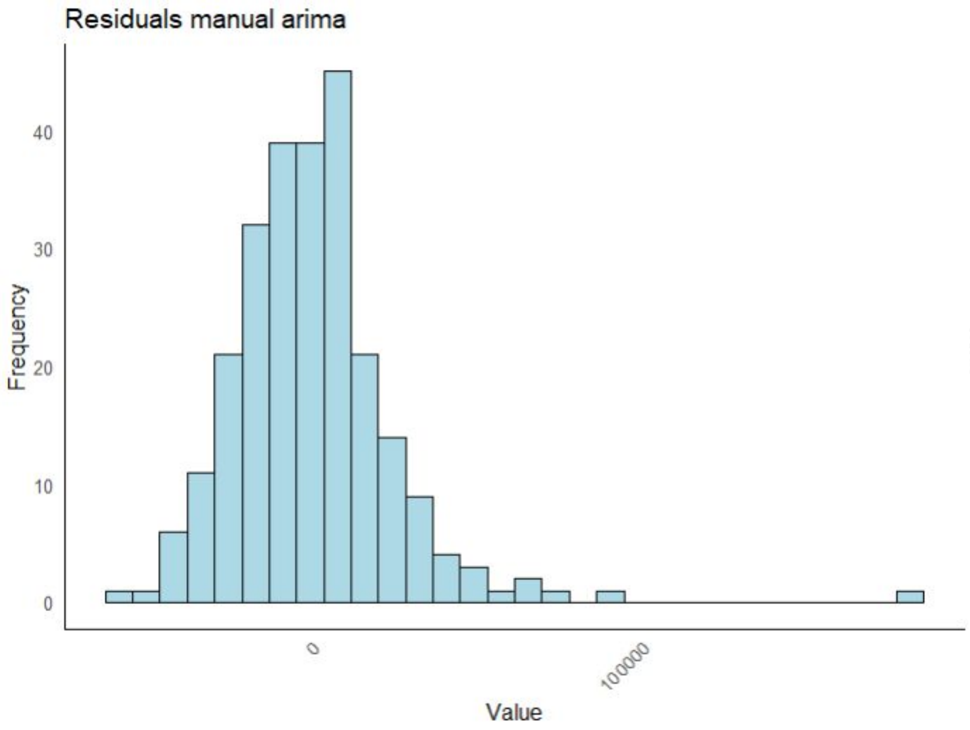
SARIMA forecasting (manual arima)

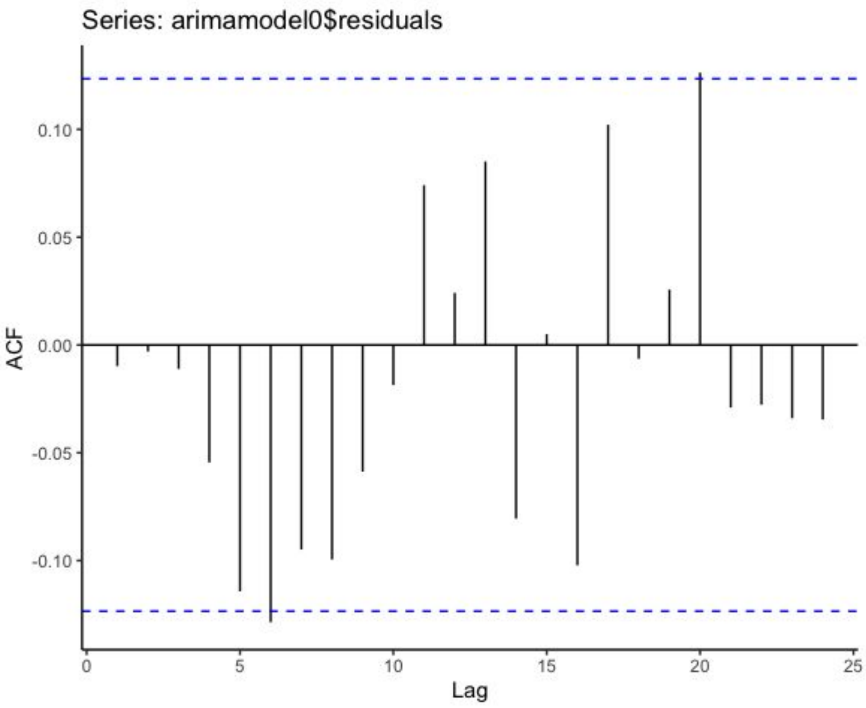
Formula ≈ ARIMA(5,1,0)(1,0,1)[12]

.model	ME	RMSE	MAE	MPE	MAPE	ACF1
ARIMA_manual	-2,920.62	8,273.20	6,367.99	-1.76	3.19	-0.01

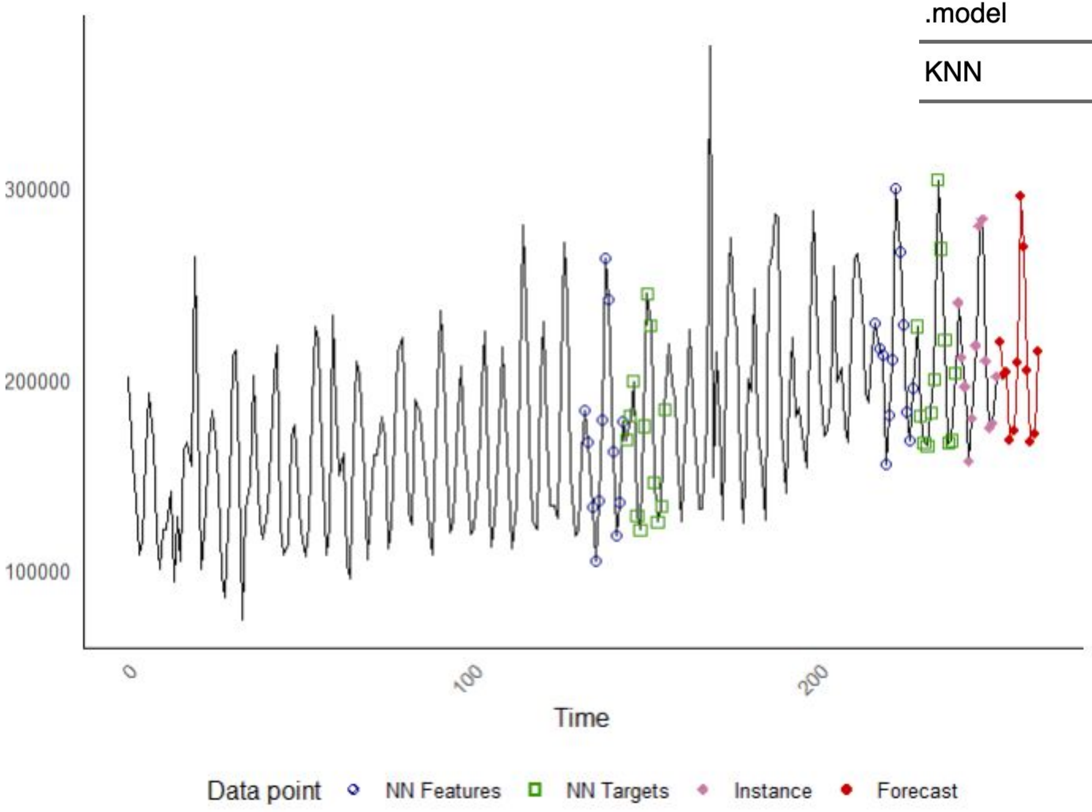


Residuals manual SARIMA



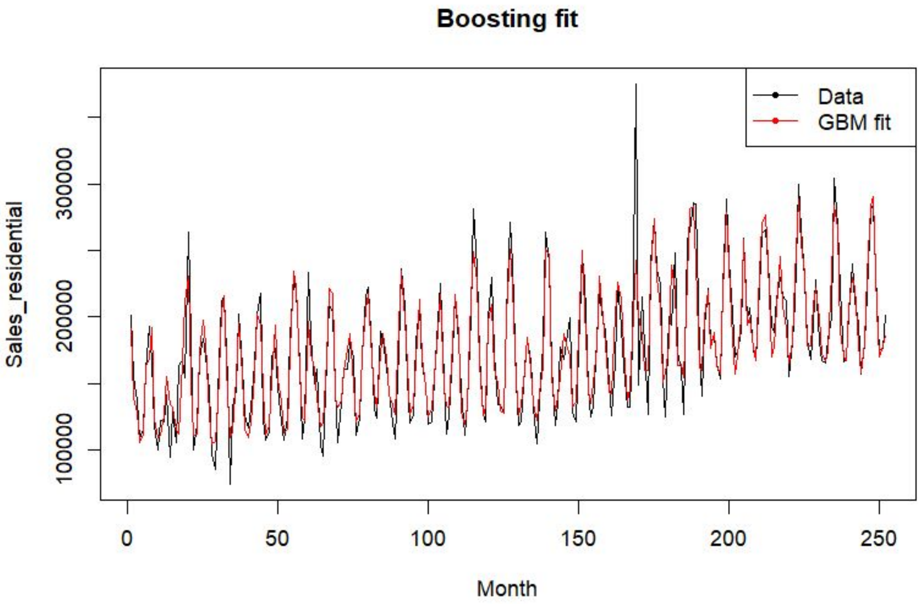


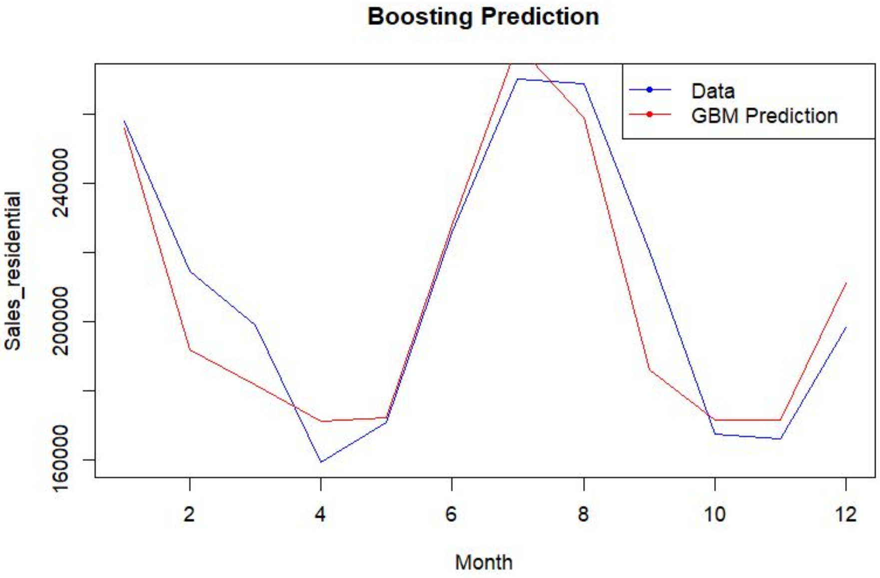
KNN regression



.model	ME	RMSE	MAE	MPE	MAPE
KNN		14,145.19	11,700.54		5.94

Generalized Boosted Regression Modeling (GBM)

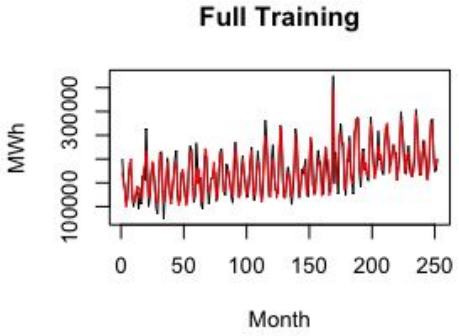




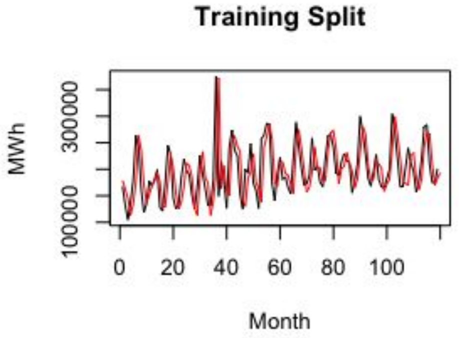
.model	ME	RMSE	MAE	MPE	MAPE
Gradient Boosting	3,270.60	14,611.77	11,150.40	1.21	5.36

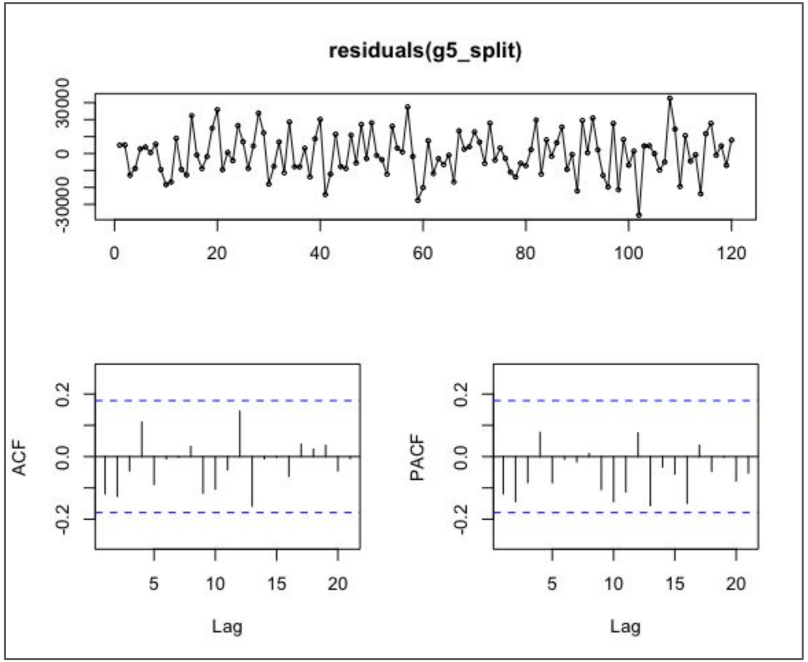
Generalized additive model (GAM)

AIC: 5,687



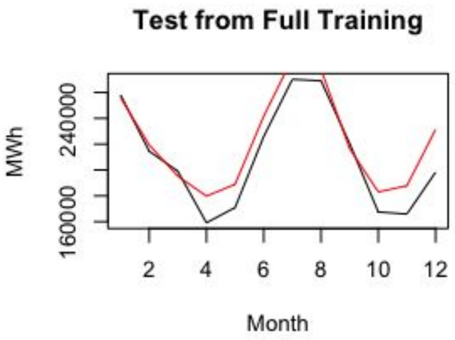
AIC: 2,671

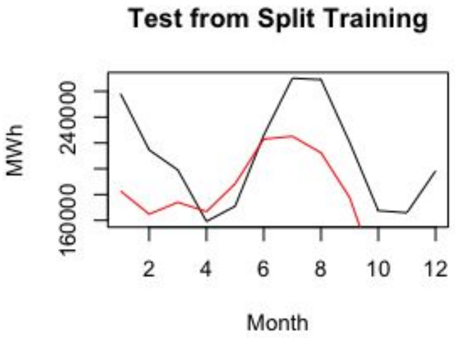




GAM Test Results

.model	ME	RMSE	MAE	MPE	MAPE
GAM complete	-12,077.36	16,358.58	13,710.96	-6.39	7.14
GAM splitted	34,059.33	43,300.14	38,227.46	15.49	17.98





Results and Conclusions

Results

Model	Predictors	RMSE	MAE	MAPE
Benchmarks				
Drift	last 2 values	39,612	33,124	16
Mean	mean	51,223	40,570	17
Naïve	last value	39,614	33,122	16
Seasonal naïve	last value from same s	9,588	8,308	3.87
Linear Regression				
TSLM	t + s	11,233	8,958	4.53
MLR	***	37,464	32,672	15.79
MLR 2012	***	40,805	35,221	16.86
ARIMA				
ARIMA	ARIMA(5,1,0)(1,0,1)[12]	8,273	6,368	3.19
auto ARIMA	ARIMA(5,1,0)	22,224	18,592	9.24
Non-Parametric				
Gradient Boosting	Decision Trees	9,642	8,775	4.22
KNN	2 NN	14,145	11,701	5.94
Exponential Smoothing				
Holt-Winters'(+)	AAA	10,237	7,739	3.87
Blended				
GAM	***	16,358	13,711	7.14
GAM 2012	***	43,300	38,227	17.39

AIC: 5,997

AIC: 5,389

AIC: N/A

AIC: 6,489

***these models are based on numerous predictors and smoothing parameters determined through stepwise regression.

Conclusions

- Best models:
  - Seasonal ARIMA
  - Seasonal Naive
  - TSLM with trend and seasonality
- Very simple methods such as SNAIVE perform very well on forecasting
- Gradient Boosting performs decently, but black box model lowers interpretability
- Holt-Winters appears to capture consistent seasonal variation well, but has a worse AIC than best models

Final Overview and Future Directions

Problem Formulation & Data Collection	Exploratory Data Analysis	Modelling	Model Selection
Modeled over 20 years of electricity sales data in residential homes in DC as a case study for energy consumption in US.	Determined trends and seasonality in data, studied history of energy sources in DC, and performed subjective feature selection.	Ran ~ dozen models on training set (full or post-2012) and tested on last year of data.	After assessing performance metrics, AIC, and behavior of residuals, determined a couple options for future forecasting: **seasonal naive, tslm,** and **SARIMA.**

References

https://www.eia.gov/electricity/data.php

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

Guidolin, M. (2023). Innovation Diffusion Models: Theory and Practice. John Wiley & Sons.

Thank you for your time!

Any questions?