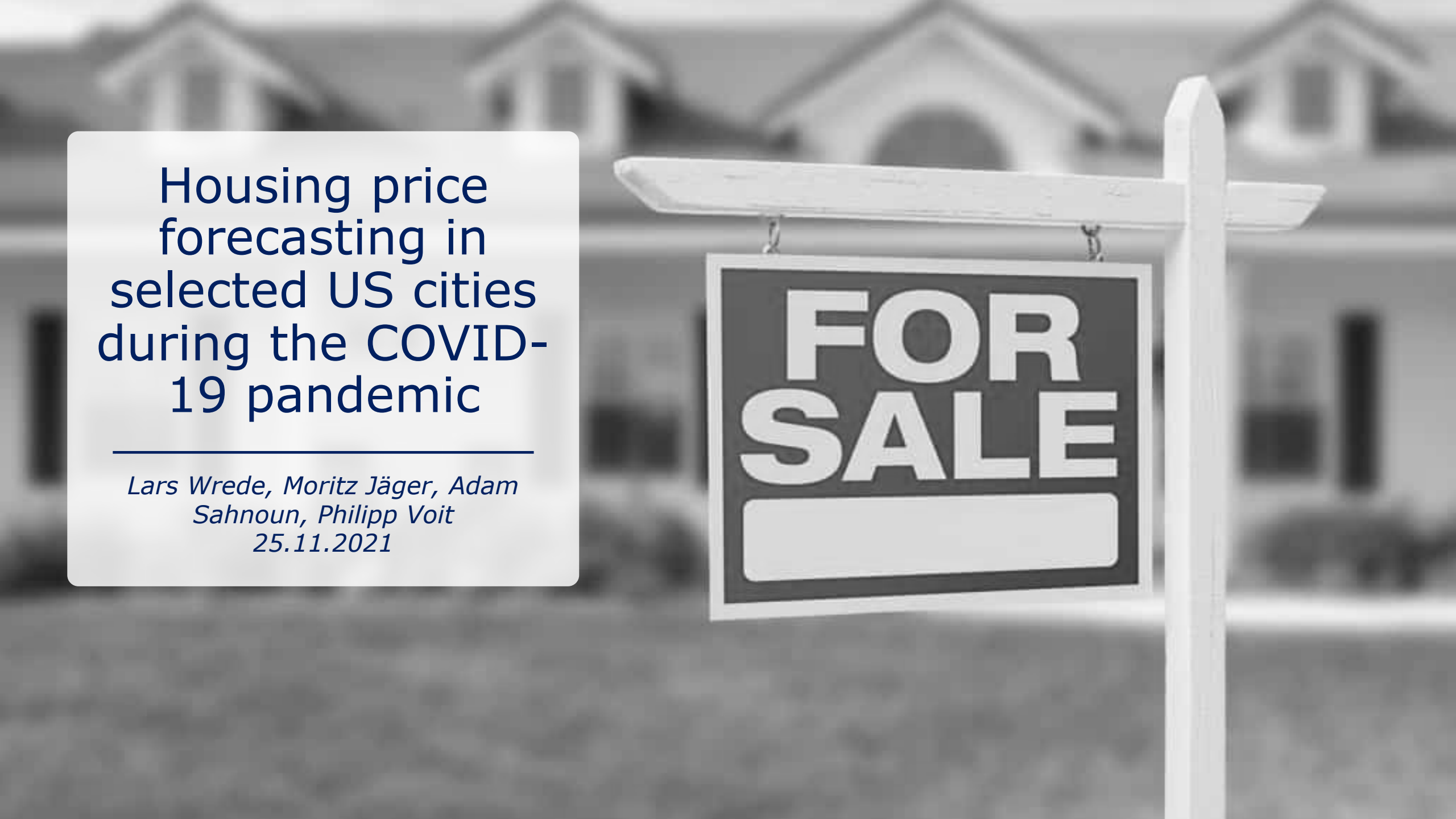
Housing price forecasting in selected US cities during the COVID-19 pandemic

Lars Wrede, Moritz Jäger, Adam Sahnoun, Philipp Voit

25.11.2021

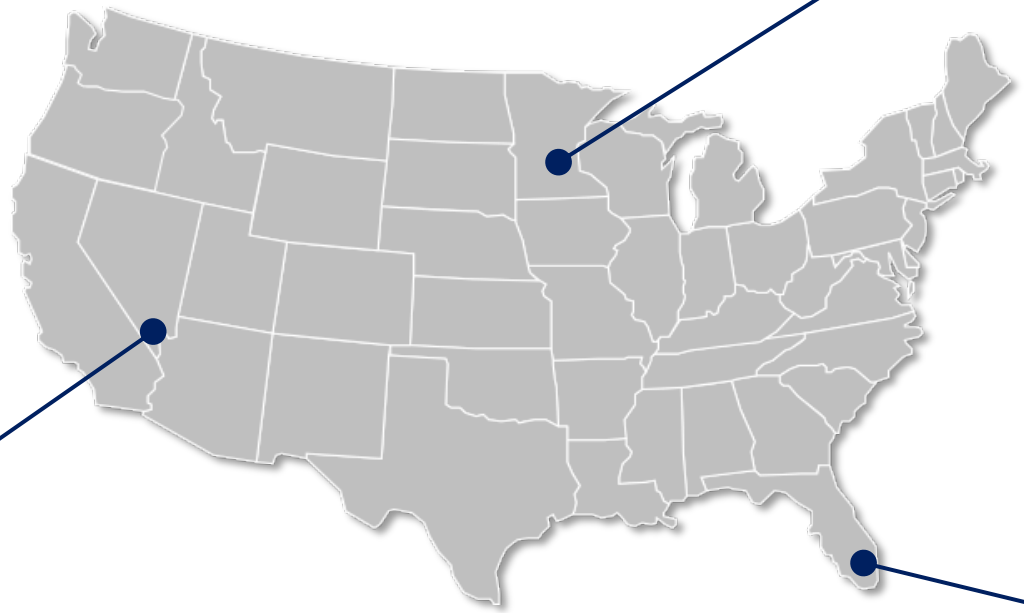


Agenda

1. Introduction (What is examined and data sources)
    1. Research question
    2. Analysis of the data bases
2. Model selection
3. Methodology & Results
    1. Seasonality & Trend
    2. Stationarity (Augmented Dickey Fuller)
    3. Autocorrelation / Partial Autocorrelation
    4. Applying the SARIMA model correctly
    5. Residual Analysis
    6. Forecasting with SARIMA
    7. Verification of the results with the OLS regression
4. Conclusion

1. Which cities are analyzed?

COVID-19 disrupted the way we live. Many of us are forced to **work from home** and generally spent more time in our homes. This disruption lets us believe that the **impact** of COVID-19 **can be seen in the housing prices.**



Minneapolis[1]

- 1,8 mio
- 2.012 km²
- 59.000 $
- 47,2%

Las Vegas

- 2,2 mio
- 20.955 km²
- 44.600 $
- 52,7%

Miami

- 2,7 mio
- 6.297 km²
- 92.800 $
- 36,6%

[1] Numbers for Minneapolis contains the data of both, Minneapolis and St. Paul (twin cities)
Source: https://www.census.gov/

1.1 Research question

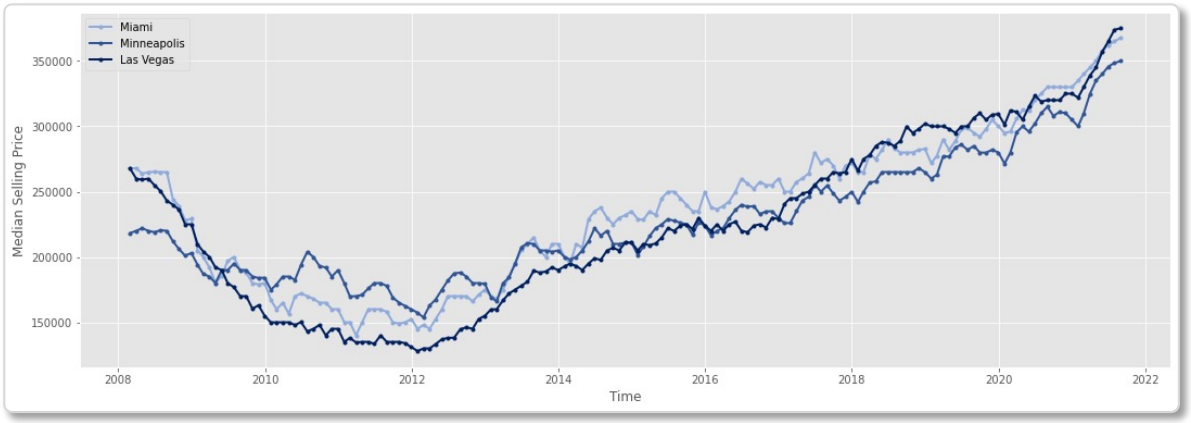
Does COVID-19 have an impact on the housing prices in Minneapolis, Miami and Las Vegas?

1.2 Analysis of the data bases

Housing price data

Database: Zillow

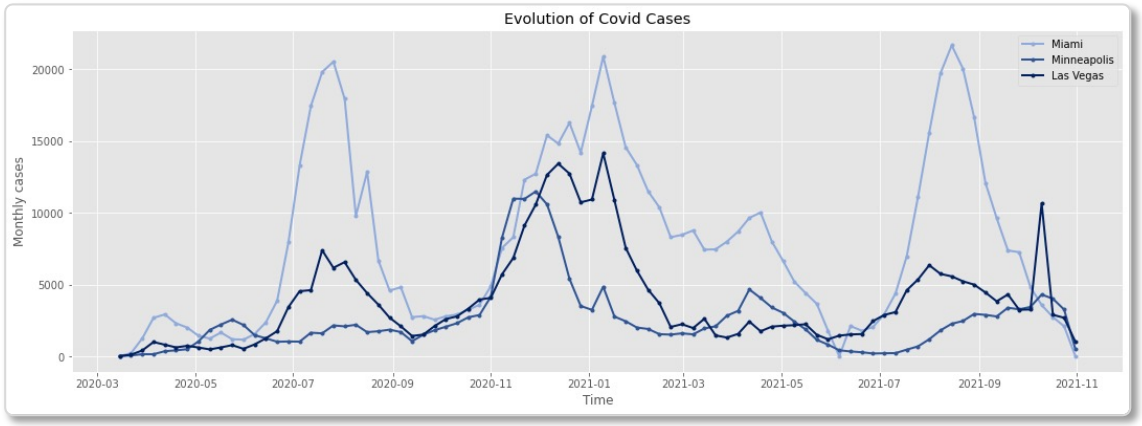
- Real estate agent
- Selling prices of single-family, condominium and co-operative homes with a county record
- Checking for NaN values
- Outlier detection



COVID-19 Data

Database: New York Times

- Github page with publicly available Covid-19 data
- Numbers updated every weekday
- Accumulated
- First data from March 2020



2. ARIMA & OLS for the analysis of time-series

What do we want to model:

Forecast:
- Housing prices w/o impact of COVID-19 and compare it to the actual data
- Time series
- Seasonal data
- Moving average
- < 2 years forecast

model?

(S)ARIMA
- widely used *time series* forecasting approaches
- Fits well our necessities

How can we verify our results:

Ordinary least squares
- By which extend impacts an increasing number of COVID-19 cases the housing prices

COVID-19 cases → **impact?** → **Housing prices $$**

3.1 Seasonality & Trend

Seasonality means the data has:

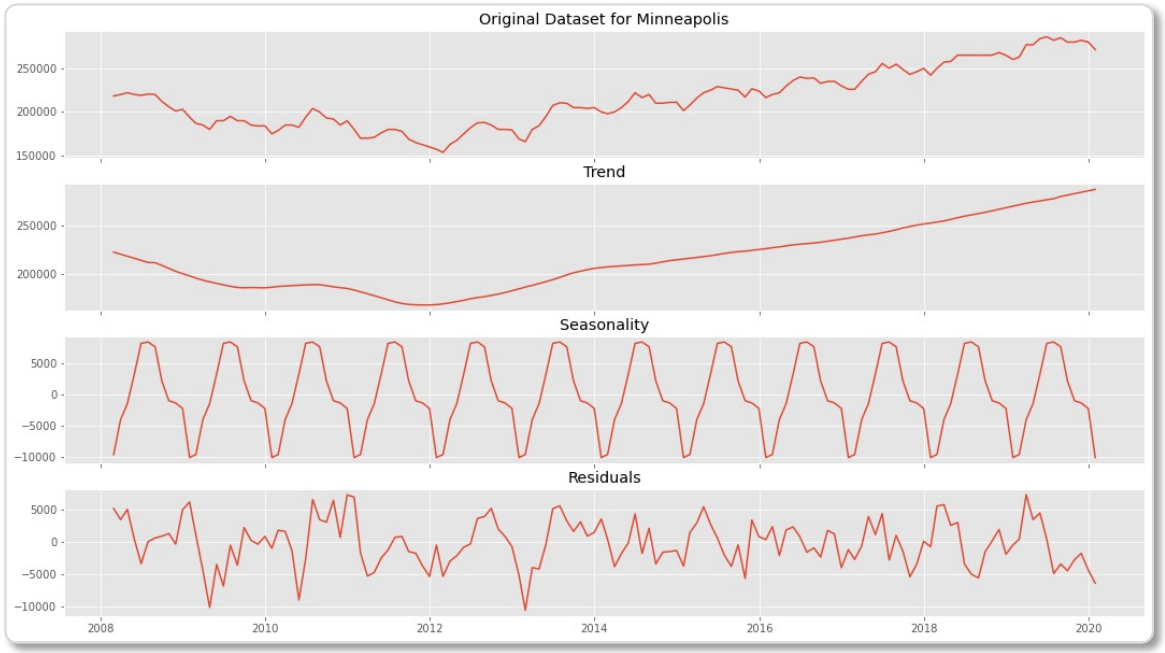
- Predictable and repeated pattern
- Repeats after any amount of time

The original data set can be broken down into:

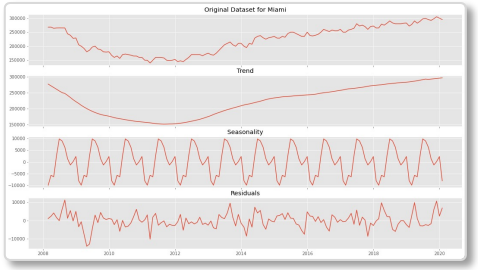
- Trend
- Seasonality
- Residuals

```
from statsmodels.tsa.seasonal import seasonal_decompose
…
def seasonal_decomposition(Column, City):
          seasonal_decom = seasonal_decompose(Column,
          model='additive', period=12,
          extrapolate_trend=12)
```

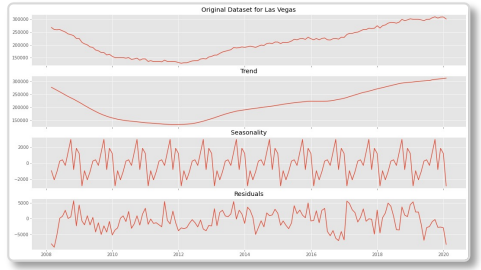
Minneapolis



Miami



Las Vegas



3.2 Stationarity

Augmented Dickey Fuller Test

Tests for trend stationarity. If the data is non-stationary it has to be made that way.

Different ways to make data stationary are:

- Taking the difference between periods
  `df.diff()`
- Taking the log
  `np.log(df)`
- Taking the square root
  `np.sqrt(df)`
- Proportional changes
  `np.shift(1)/df`

```
from statsmodels.tsa.stattools import adfuller
…
checking_stationarity(train.MSP)
```

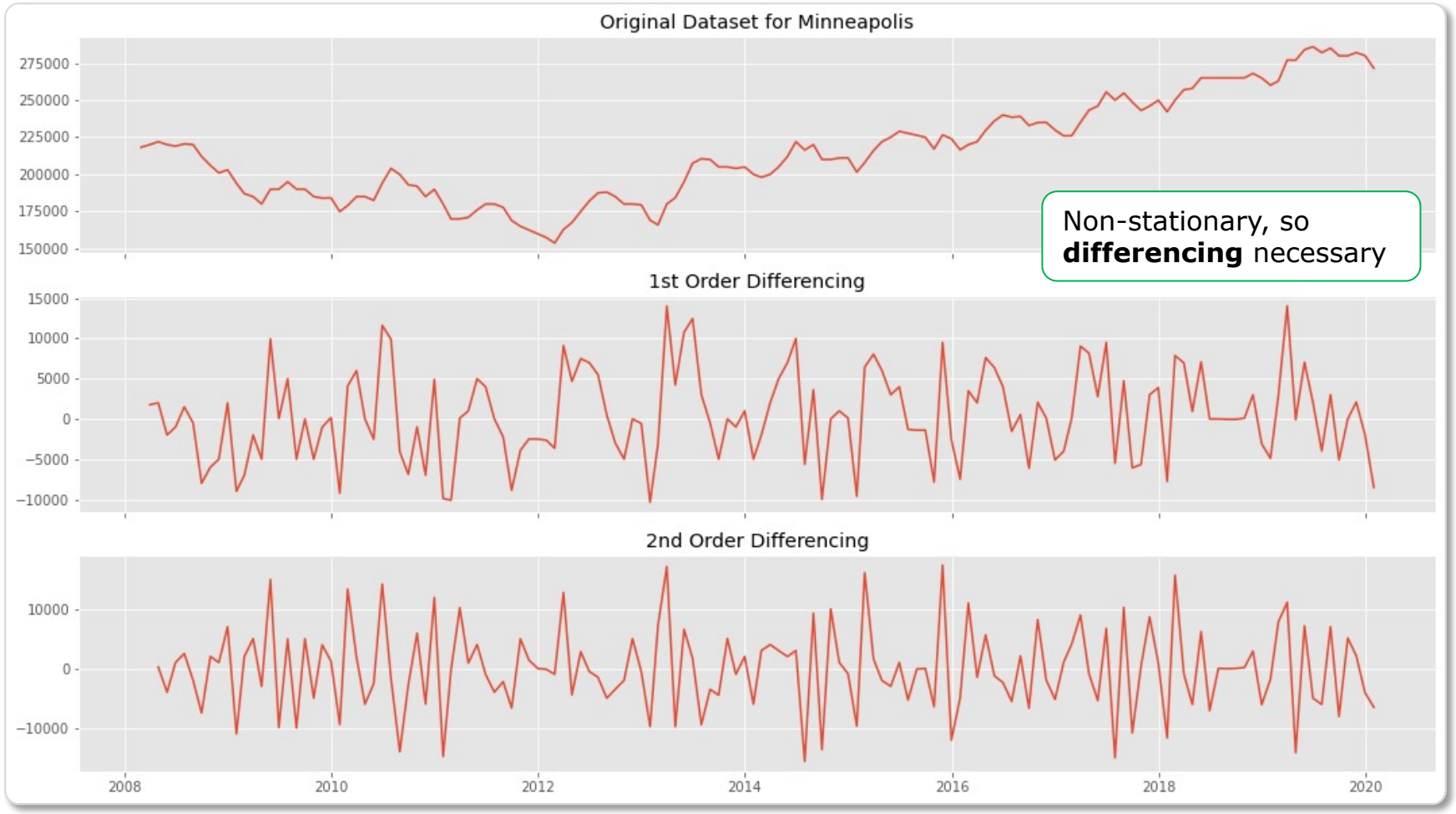
```
Output Minneapolis:
(-0.387770, 0.912121, x, x, {-3.482, -2.884, -2.579})
…
2nd Order:
(-10.148634, 0.00000, x, x, {-3.482, -2.884, -2.579})
```

- 0th element is test-statistics
  - More negative the more liekely the data is stationary
- 1st element is p-value
  - If p is small → reject non-stationarity
- 4th element is the critial test statistics

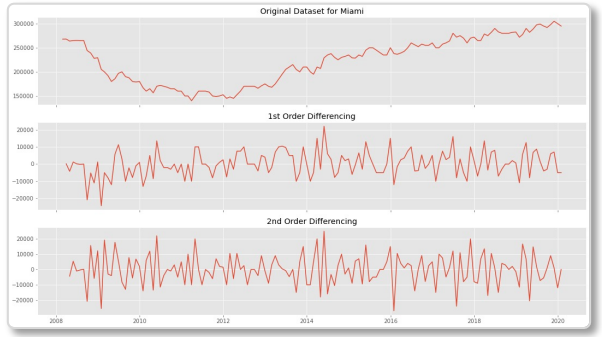
[1] https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.adfuller.html

3.2 Stationarity

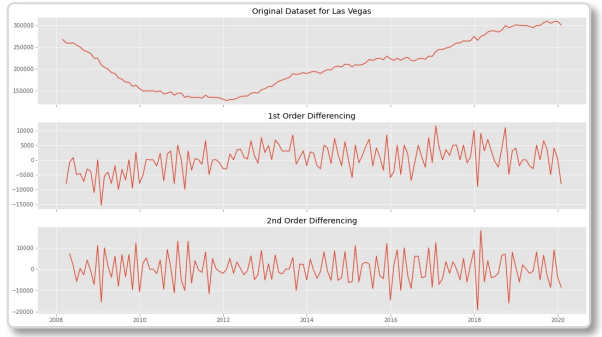
Minneapolis



Miami



Las Vegas



3.3 Autocorrelation / Partial Autocorrelation

Autocorrelation function (ACF)

Is the correlation between a time series and the same time series offset by n steps.

- lag-1 autocorrelation → $corr(y_t, y_{t-1})$
- lag-2 autocorrelation → $corr(y_t, y_{t-2})$
- …
- lag-n autocorrelation → $corr(y_t, y_{t-n})$

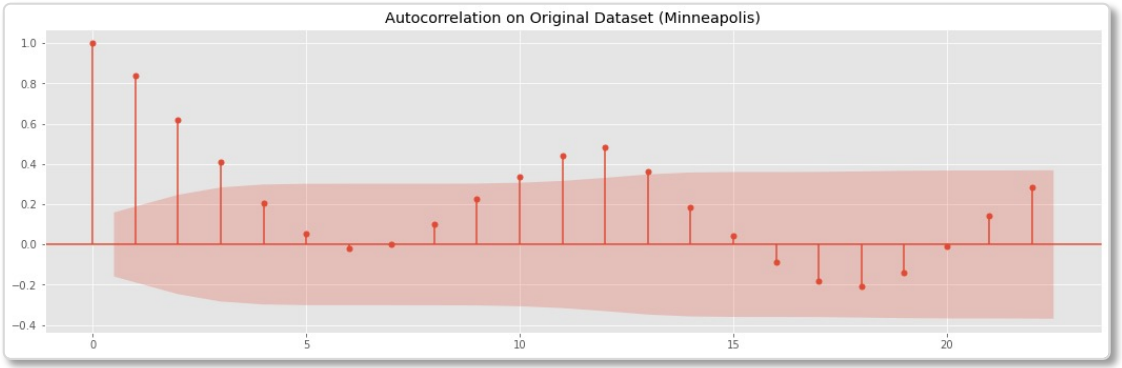
Partial autocorrelation function (PACF)

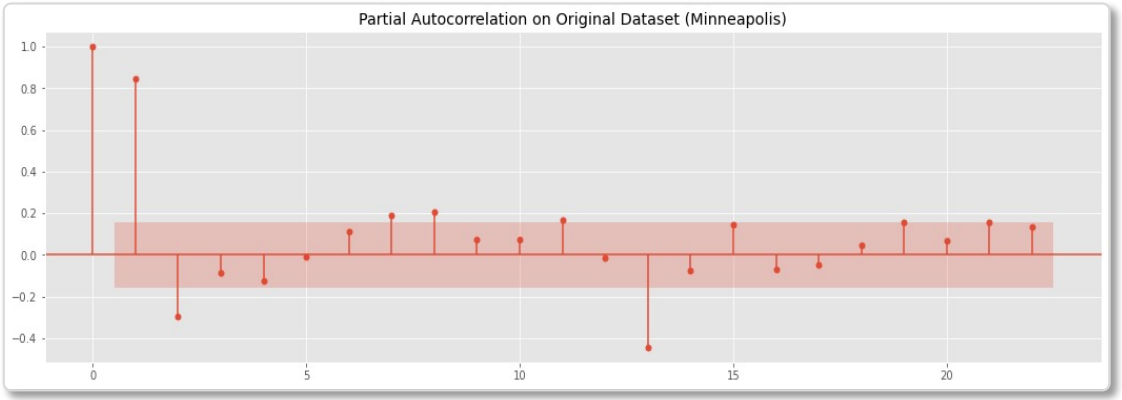
Is the correlation between a time series and the lagged version of itself after subtracting the effect of correlation at smaller lags

Analysing the models with the following table:

	AR(p)	MA(q)	ARMA(p,q)
ACF	Tails off	Cuts off after lag p	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

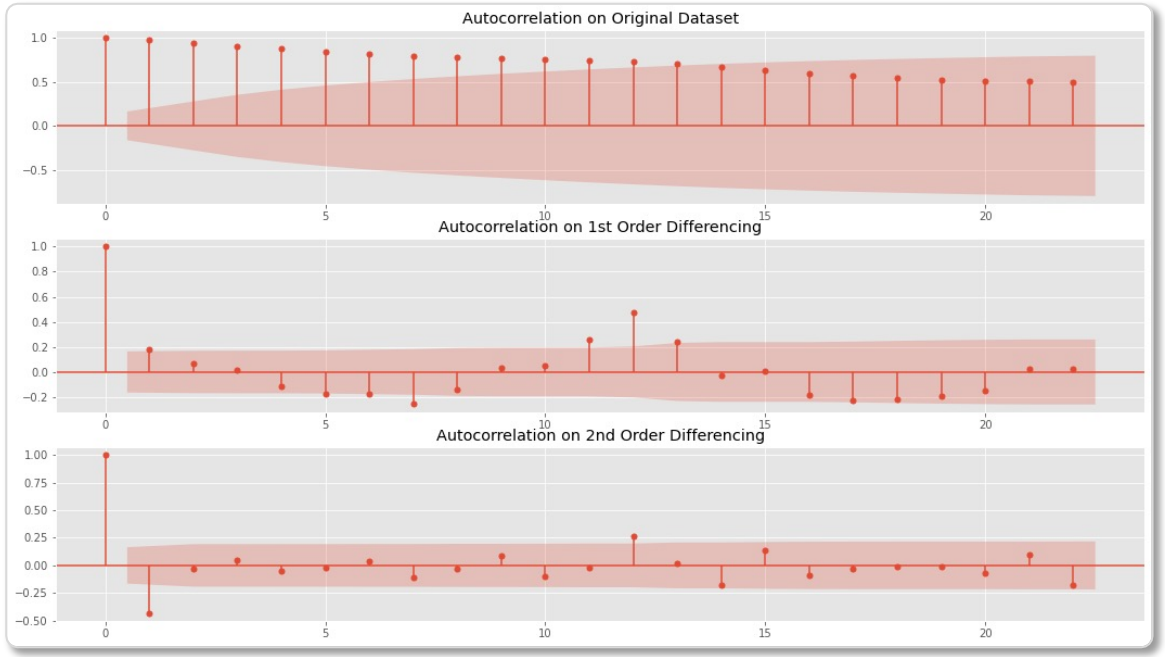
```
from statsmodels.tsa.stattools import acf, pacf
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
…
```



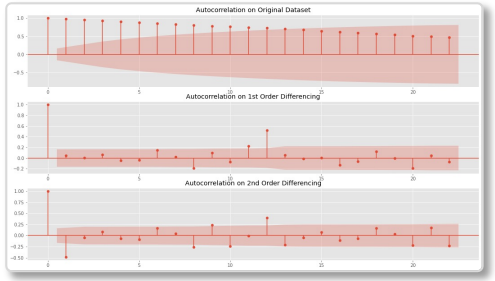


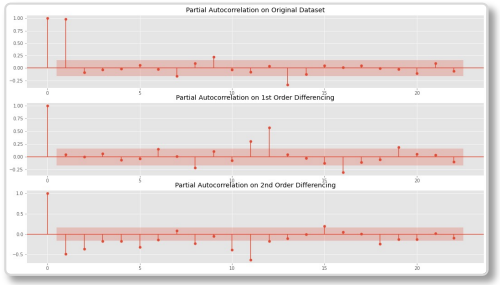
3.3 Autocorrelation & Partial autocorrelation

Minneapolis

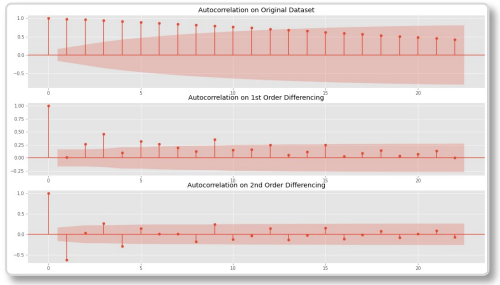


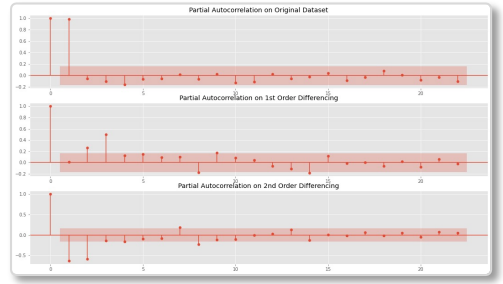
Miami





Las Vegas





3.4 Applying the SARIMA model correctly

Seasonal ARIMA = SARIMA $(p,d,q)(P,D,Q)_M$

- Non seasonal orders:
  - *p: autoregressive order (AR)*
  - *d: differencing order (I)*
  - *q: moving average order (MA)*
- Seasonal orders:
  - *P: seasonal autoregressive order(AR)*
  - *D: seasonal differencing order (I)*
  - *Q: seasonal moving average order*
  - *M: number of time steps per cycle*

3.4 Applying the SARIMA model correctly

Auto-arima procedure

- determine the lowest AICC to find best performing SARIMA model

```
import pmdarima as pmd
…
smodel_MSP = pmd.auto_arima(train.MSP, test='adf’,
          max_p=3, max_q=3, m=12, seasonal=True)
```

Variable	SARIMA-spec	AICC[1]	BIC[2]	HQIC[3]
Minneapolis	***0|1|3 |1|1|1***	***2613.234***	***2630.485***	***2620.244***
	1|1|2 |1|1|1	2613.276	2630.527	2620.286
	2|1|1 |1|1|1	2613.285	2630.536	2620.295
Miami	***0|1|0 |0|1|0***	***2672.600***	***2675.475***	***2673.768***
	0|1|0 |0|1|0 intercept	2673.072	2678.822	2675.408
	1|1|0 |0|1|0	2675.275	2681.026	2677.612
Las Vegas	***2|1|0 |0|1|2***	***2634.187***	***2648.563***	***2640.028***
	2|1|1 |0|1|2	2634.770	2652.021	2641.780
	1|1|1 |0|1|2	2634.866	2649.242	2640.708

[1] corrected AIC
[2] Schwarz Bayesian criterion
[3] Hannan-Quin criterion

3.5 Residual analysis

How far are our prediction from the true values?

If the model fits well the residuals will be white Gaussian noise

1. *Standardized residuals*
   should have no obvious structure

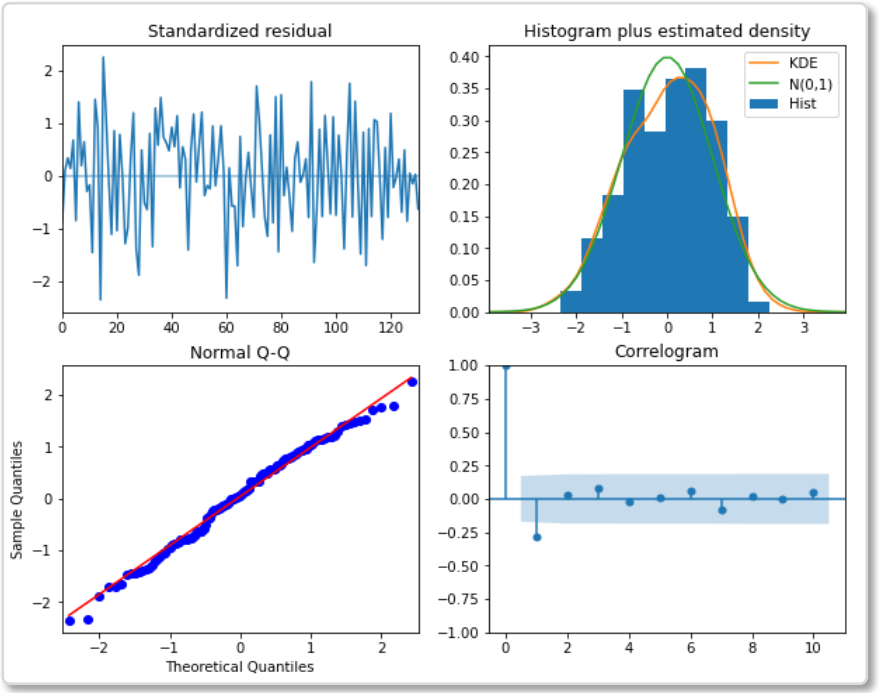
2. *Histogram plus estimated density*
   lines should be almost overlapping

3. *Normal Q-Q*
   all the points should lie along the red line

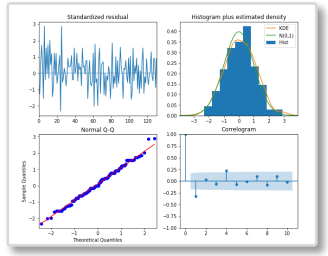
4. *Correlogram*
   95% should not be significant (inside the shaded area)

```
results.plot_diagnostics( )
plt.show( )
```

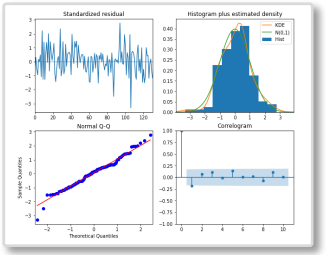
Minneapolis



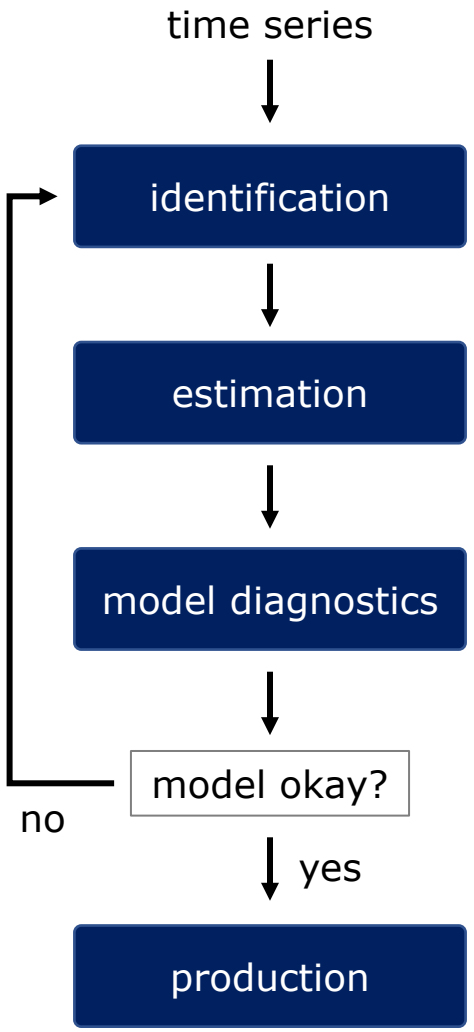
Miami



Las Vegas

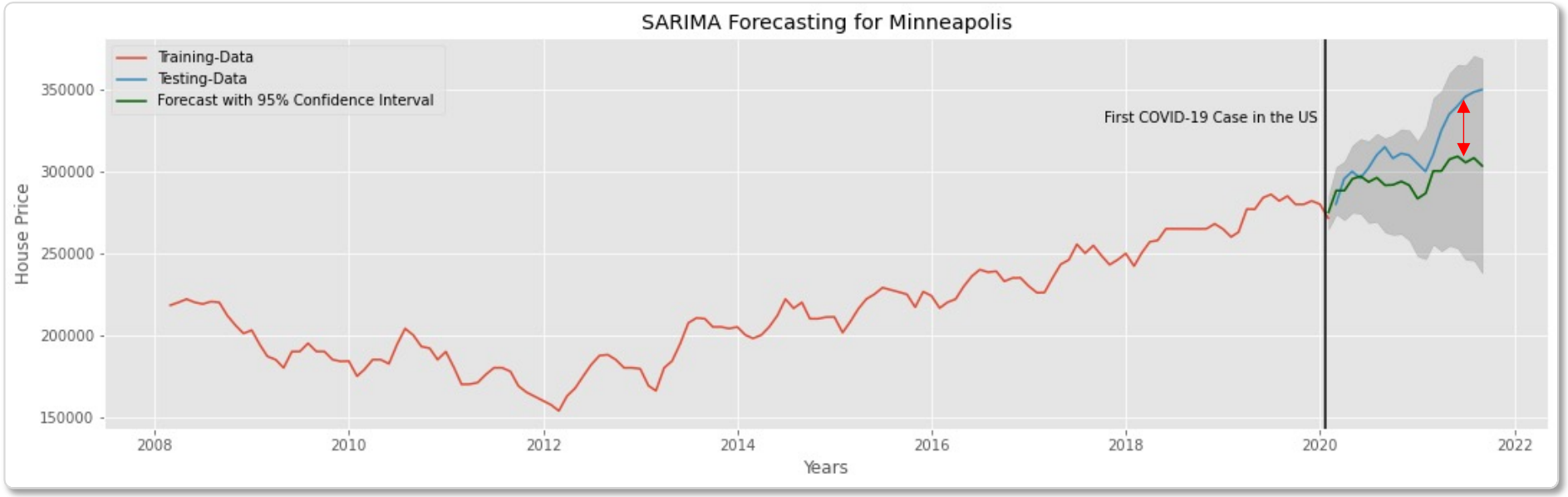


Workflow

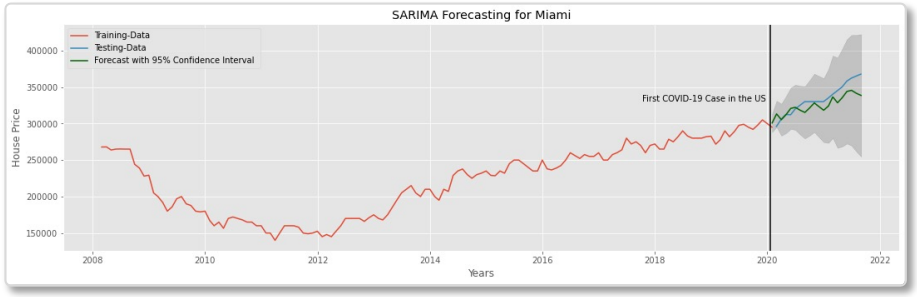

time series
identification
estimation
model diagnostics
model okay?
no
yes
production

3.6 Forecasting with SARIMA

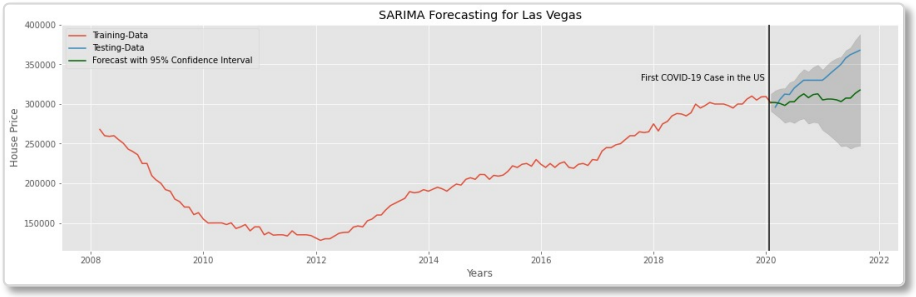
Minneapolis



Miami



Las Vegas



Explanation:

green → forecast of the median selling price w/o COVID-19.

blue → actual median selling price of houses during COVID-19.

Subtracting the green line from the blue line yields the **excess trend**.

3.7 Verification of the results with the OLS regression

Tool to analyze the relationship between a set of independent and dependent variables.

Minneapolis

OLS Regression Results

Dep. Variable:	Residuals	R-squared (uncentered):	0.648
Model:	OLS	Adj. R-squared (uncentered):	0.627
Method:	Least Squares	F-statistic:	31.28
Date:	Mon, 22 Nov 2021	Prob (F-statistic):	3.23e-05
Time:	21:38:31	Log-Likelihood:	-197.83
No. Observations:	18	AIC:	397.7
Df Residuals:	17	BIC:	398.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Covid	14.9213	2.668	5.593	0.000	9.293	20.550

Omnibus:	1.215	Durbin-Watson:	0.337
Prob(Omnibus):	0.545	Jarque-Bera (JB):	1.012
Skew:	-0.520	Prob(JB):	0.603
Kurtosis:	2.484	Cond. No.	1.00

Covid = 14.921

How much the *Residuals* is expected to increase when *Covid* increases by one

R squared = 65%

Percentage of variation in dependent that is explained by independent variables

F-statistic = 0.000

Closer to zero the more meaningfull the regression

Durbin-Watson = 0.337

Value between 1 and 2 preferred. Implies that the variance of errors is constant

Prob (Omnibus) = 0.545

To check if errors are normally distributed. Closer to 1 is better

Notes:
[1] R^2 is computed without centring (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

3.7 Verification of the results with the OLS regression

Miami

OLS Regression Results

Dep. Variable:	Residuals	R-squared (uncentered):	0.510
Model:	OLS	Adj. R-squared (uncentered):	0.482
Method:	Least Squares	F-statistic:	17.72
Date:	Mon, 22 Nov 2021	Prob (F-statistic):	0.000590
Time:	21:38:15	Log-Likelihood:	-191.10
No. Observations:	18	AIC:	384.2
Df Residuals:	17	BIC:	385.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Covid	0.2343	0.056	4.209	0.001	0.117	0.352

Omnibus:	0.328	Durbin-Watson:	0.826
Prob(Omnibus):	0.849	Jarque-Bera (JB):	0.483
Skew:	-0.145	Prob(JB):	0.785
Kurtosis:	2.252	Cond. No.	1.00

Las Vegas

OLS Regression Results

Dep. Variable:	Residuals	R-squared (uncentered):	0.333
Model:	OLS	Adj. R-squared (uncentered):	0.294
Method:	Least Squares	F-statistic:	8.484
Date:	Mon, 22 Nov 2021	Prob (F-statistic):	0.00970
Time:	21:38:19	Log-Likelihood:	-209.18
No. Observations:	18	AIC:	420.4
Df Residuals:	17	BIC:	421.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Covid	0.8619	0.296	2.913	0.010	0.238	1.486

Omnibus:	0.624	Durbin-Watson:	0.233
Prob(Omnibus):	0.732	Jarque-Bera (JB):	0.613
Skew:	0.070	Prob(JB):	0.736
Kurtosis:	2.107	Cond. No.	1.00

Miami with the lowest Covid coef (0.2343) indicating less correlation between number of Covid cases and increase of housing prices.
This is **in contra** of our research question.

Notes:
[1] R^2 is computed without centring (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

4. Conclusion

Critic on the selected model:

- Sensitive algorithm implementation
- Difficult definition of parameters
- High complexity (preparation of data)
- High-danger of overfitting
- Difficult in explanation and interpretation

Critic on the used variable:

- Variable “Corona numbers” was picked randomly
- Other parameters such as unemployment rate, house-owner rate or average salary aren’t regarded in our model
- statistical evaluation of corona cases are potentially misinterpreted (e.g. different measurements in different states)

Results of our investigations:

- Positive correlation between cases & housing prices
- Assumption -> need for add. space increased due to home office
- Trust in assets such as real estate itself has increased
- During lockdown, barely any construction work
- Residuals in the Miami case were the lowest:
  - Lower house-owner rate (appr. 30%) -> lower demand in general for buying properties

Forecast for San Francisco