Time-Series Analysis Prediction of Avocado-prices in the US

Ökonometrie 2
SS 2019

Group-members

Dornigg, Thomas (01551056)

Kafenda, Lukas (01607734)

Sunk, David (01605328)

Content

1. Project description
2. Exploratory Data Analysis of the Dataset
3. Testing for Stationarity
4. Correcting for Stationarity
5. Fitting appropriate model
    a. Traditional approach (ARMA, ARIMA, SARIMA)
    b. Facebook-Prophet Library
    c. Comparison
6. Conclusion

1. Project and Dataset Description

- Kaggle dataset from 2018[1]
- represents weekly retail scan data for national retail volume (units) and price
- based on actual retail sales of Hass avocados
- time-range: 04-01-2015 : 25-03-2018
- Research Questions:
    - Was the Avocadopocalypse of 2017 real? (taken from Kaggle)
    - What is the potential price which customers would have to pay for Conventional-Avocados in early 2018 respectively 2019?

2. Exploratory Data Analysis

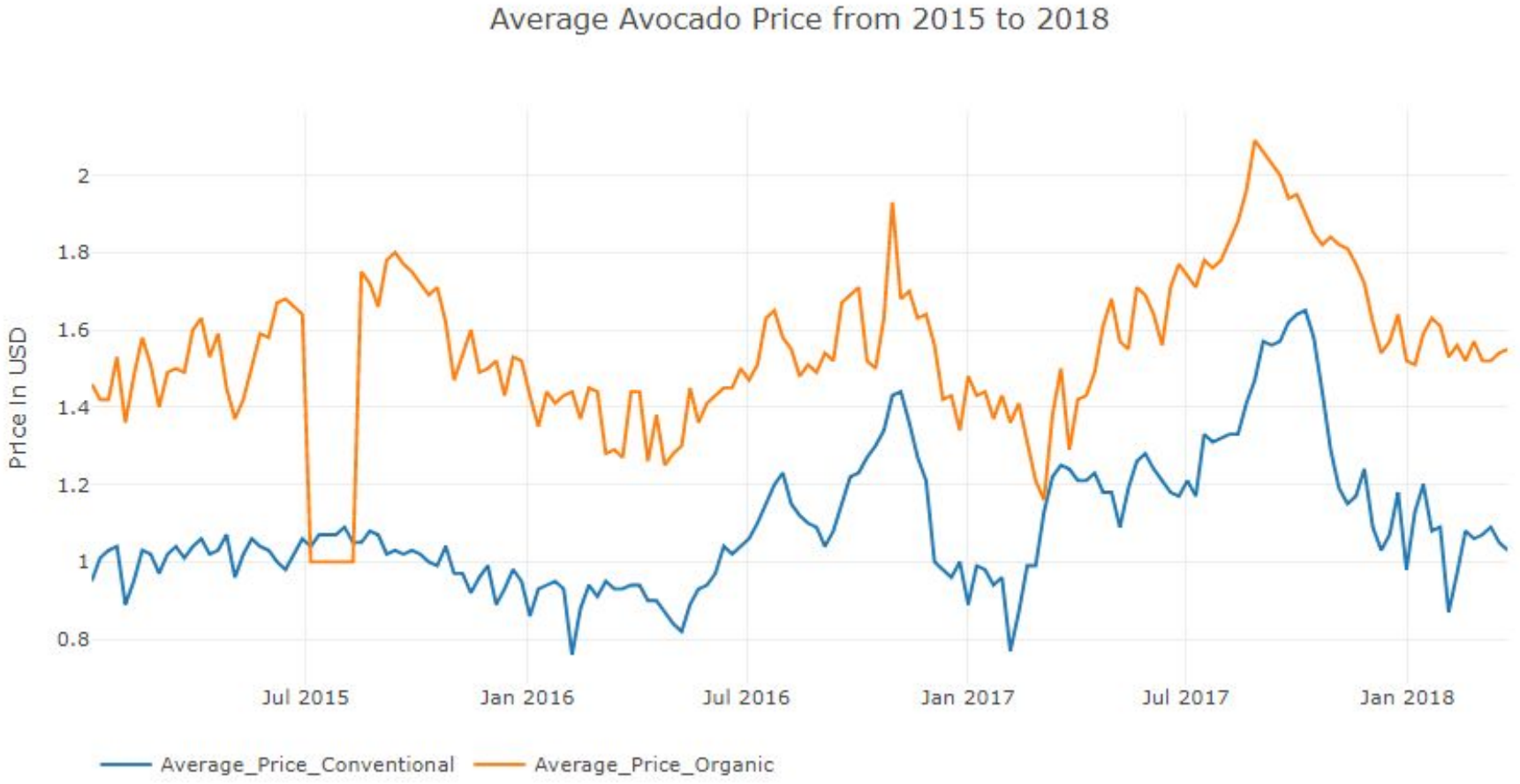
- dataset consists of following features:
  - Date
  - Average Price
  - Type (organic/conventional)
  - Year
  - Total Volume
- avocado-type comparison (descriptive statistics, pie chart)
- evolution of avocado prices (line graph)
- comparison of avocado prices (density plot, boxplots, line graph)
- volume of sold avocados grouped by regions (frequency distribution bar chart)

Conventional

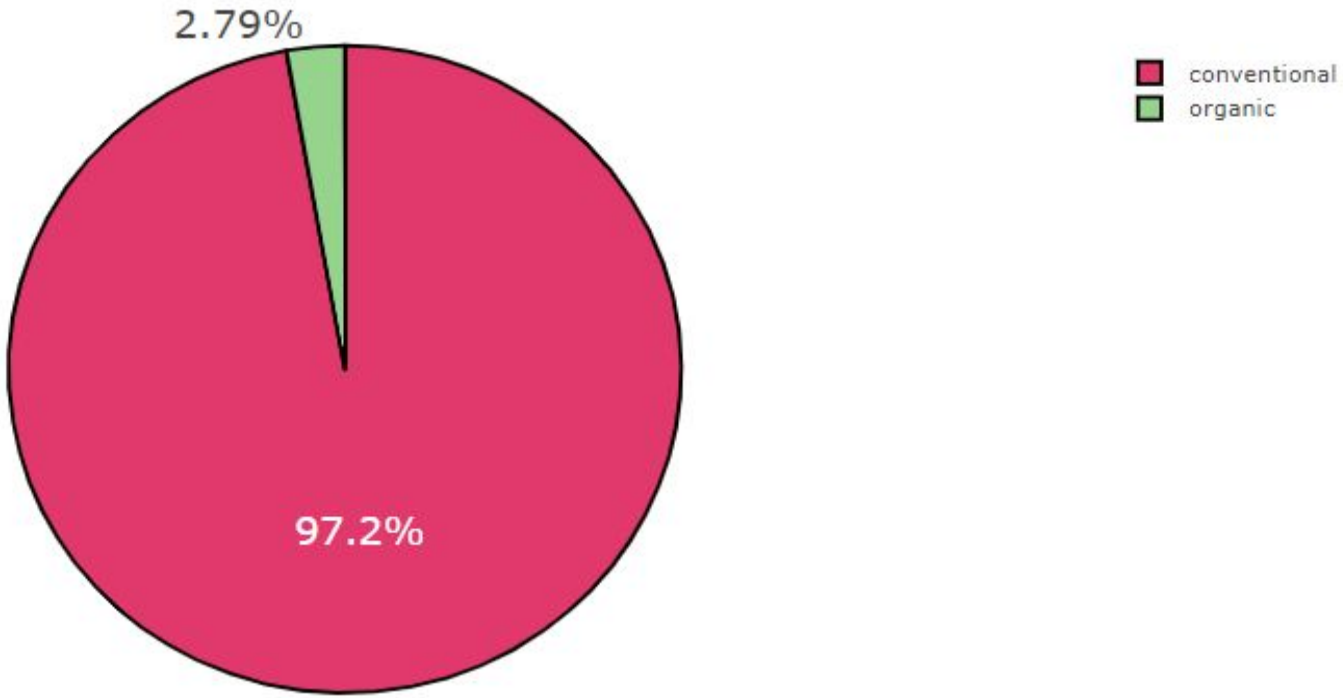
	Average Price	Total Volume	Region	Type
Types	numeric	numeric	constant	constant
Counts	169	169	169	169
Uniques	61	169	1	1
Missing	0	0	0	0
Missing %	0%	0%	0%	0%
Mean	1.09	33735038.97	NaN	NaN
StD	0.17	6118091.85	NaN	NaN
Minimum	0.76	21009730.21	NaN	NaN
Maximum	1.65	62505646.52	NaN	NaN
Mean abs. Deviation	0.13	4532699.23	NaN	NaN
25%	0.97	29761638.48	NaN	NaN
50%	1.04	32994014.16	NaN	NaN
75%	1.19	37026085.75	NaN	NaN
IQR	0.22	7264447.27	NaN	NaN
Skewness	1.13	1.23	NaN	NaN
Kurtosis	1.33	4.31	NaN	NaN

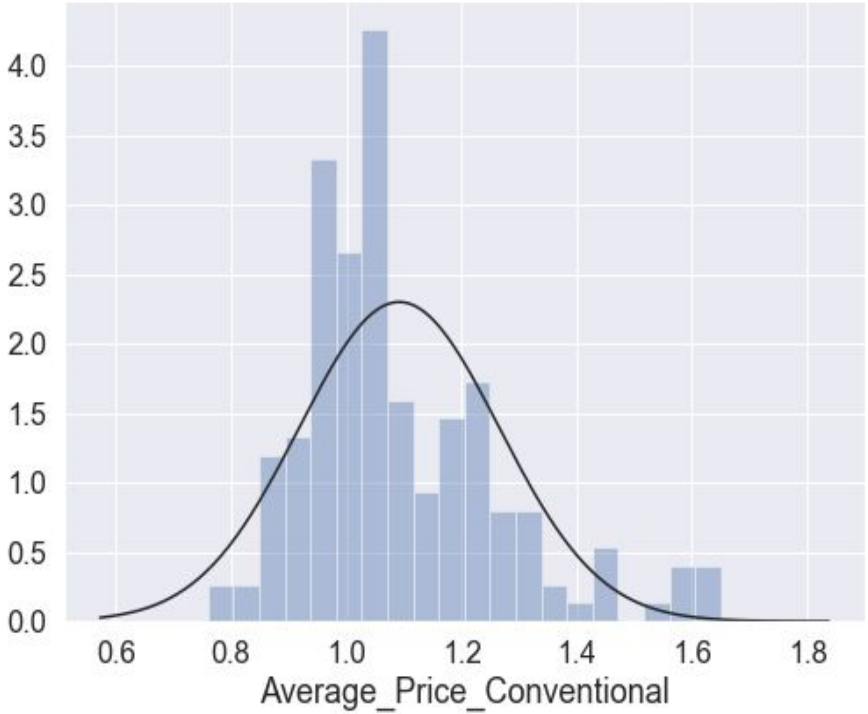
Organic

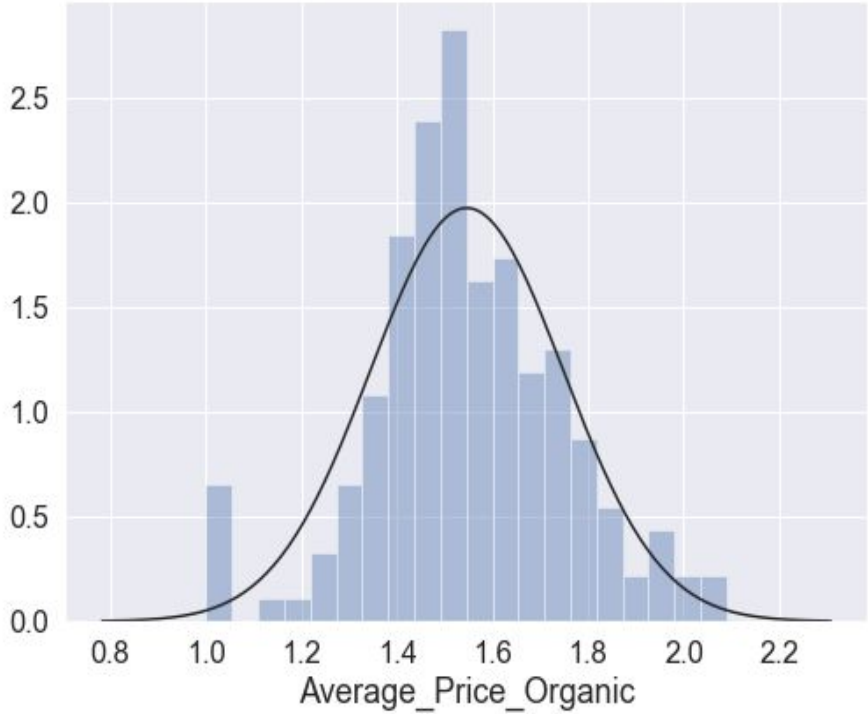
	Average Price	Total Volume	Region	Type
Types	numeric	numeric	constant	constant
Counts	169	169	169	169
Uniques	69	169	1	1
Missing	0	0	0	0
Missing %	0%	0%	0%	0%
Mean	1.55	967565.65	NaN	NaN
StD	0.20	302482.33	NaN	NaN
Minimum	1.00	501814.87	NaN	NaN
Maximum	2.09	1814929.97	NaN	NaN
Mean abs. Deviation	0.15	249361.63	NaN	NaN
25%	1.43	699763.35	NaN	NaN
50%	1.53	967886.13	NaN	NaN
75%	1.67	1148617.16	NaN	NaN
IQR	0.24	448853.81	NaN	NaN
Skewness	-0.13	0.47	NaN	NaN
Kurtosis	0.93	-0.44	NaN	NaN

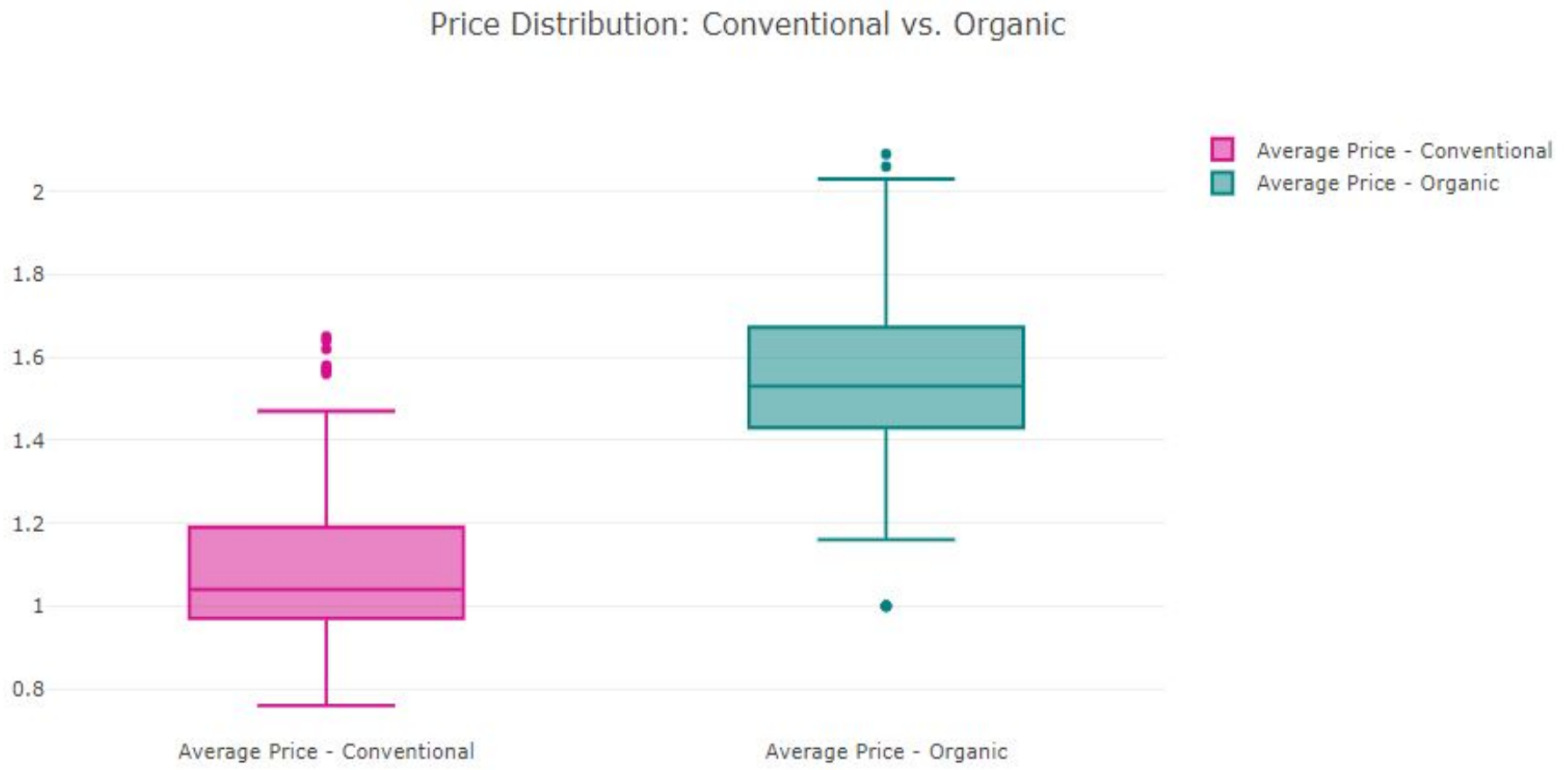

Average Avocado Price from 2015 to 2018
Price In USD
2
1.8
1.6
1.4
1.2
1
0.8
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018
Average_Price_Conventional
Average_Price_Organic

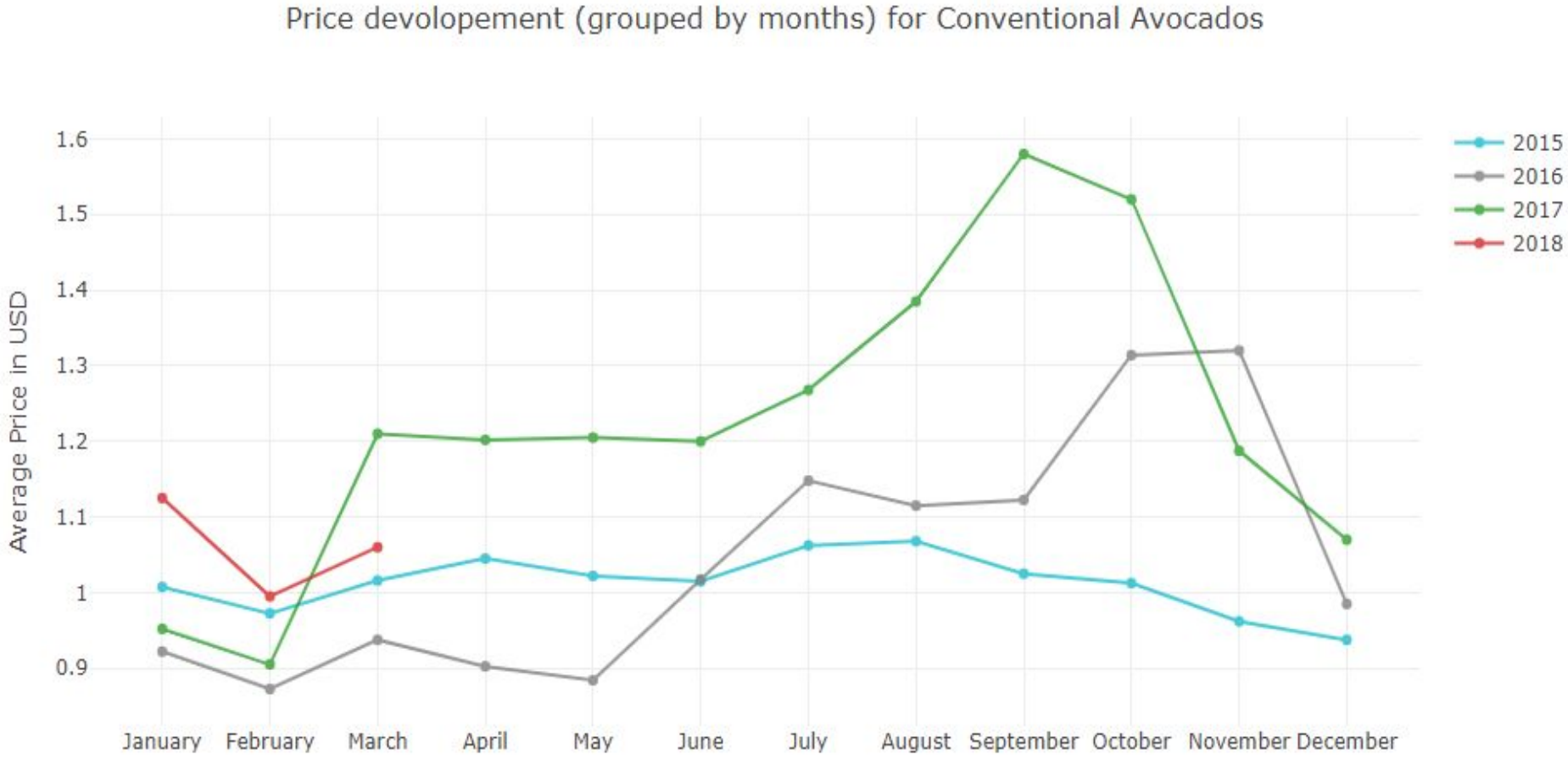
Conventional vs. Organic Avocados grouped by Total Volume sold

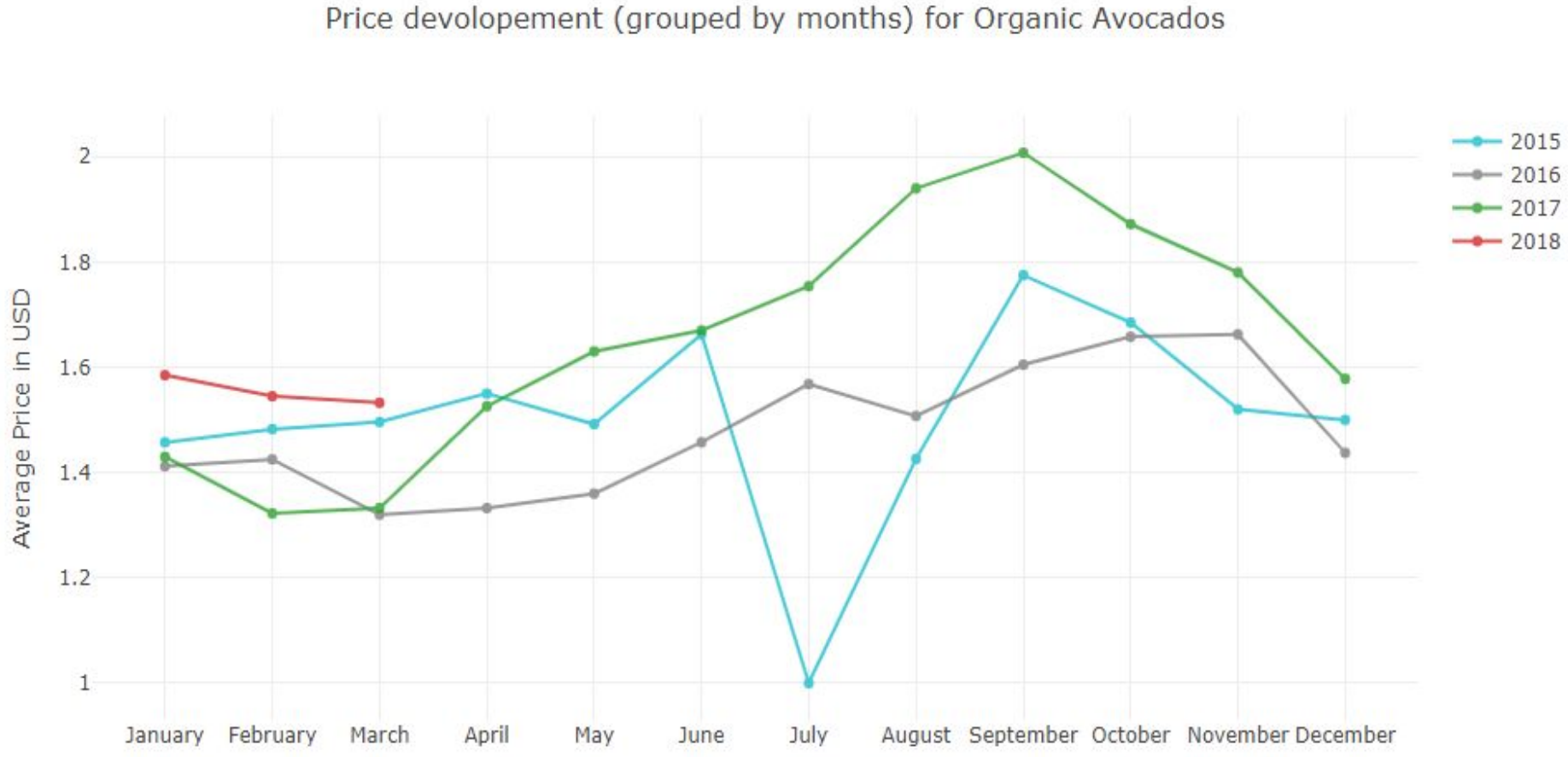


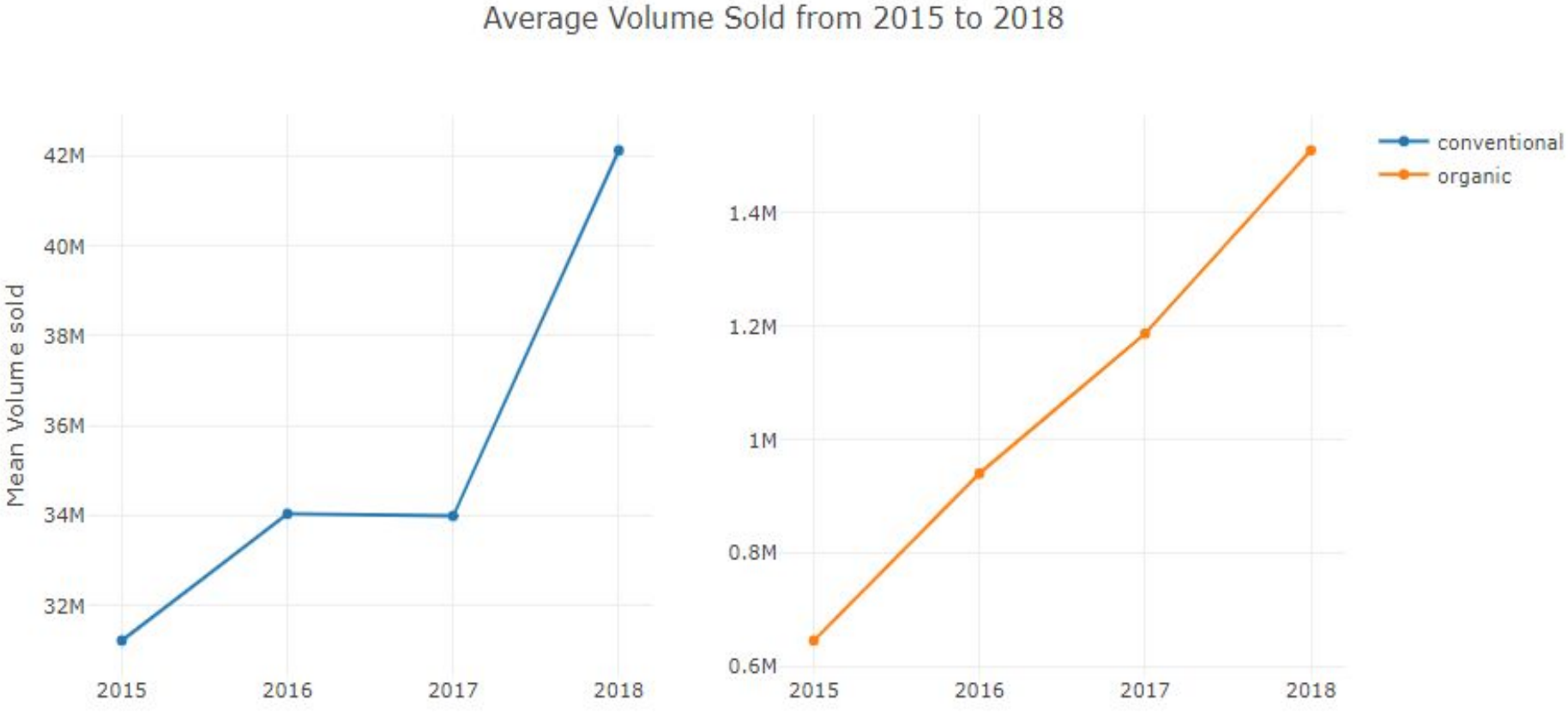

4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Average_Price_Conventional

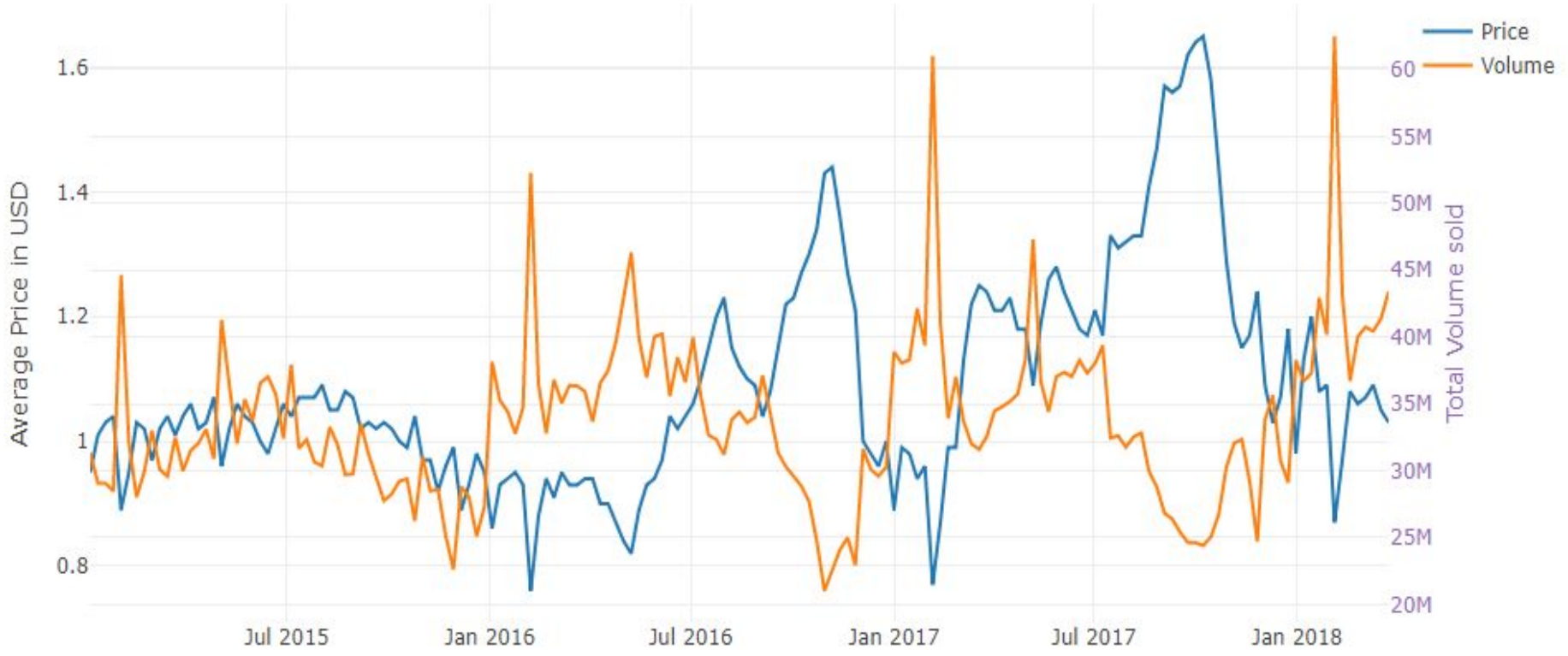

2.5
2.0
1.5
1.0
0.5
0.0
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
Average_Price_Organic

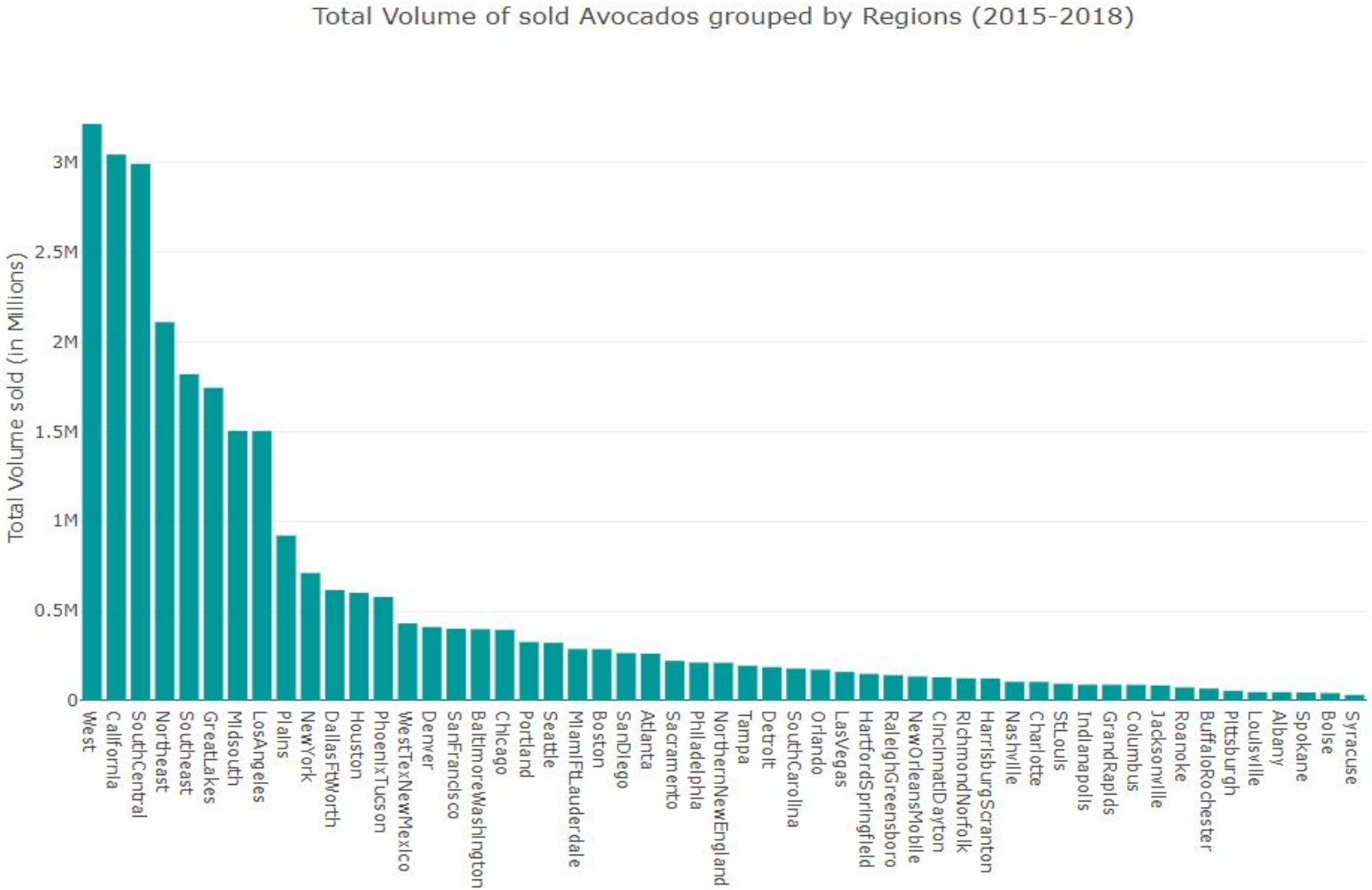

Price Distribution: Conventional vs. Organic
Average Price - Conventional
Average Price - Organic
2
1.8
1.6
1.4
1.2
1
0.8
Average Price - Conventional
Average Price - Organic


Price devolopement (grouped by months) for Conventional Avocados
Average Price in USD
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
January
February
March
April
May
June
July
August
September
October
November
December
2015
2016
2017
2018


Price devolopement (grouped by months) for Organic Avocados
Average Price in USD
2
1.8
1.6
1.4
1.2
1
January
February
March
April
May
June
July
August
September
October
November
December
2015
2016
2017
2018


Average Volume Sold from 2015 to 2018
Mean Volume sold
42M
40M
38M
36M
34M
32M
2015
2016
2017
2018
1.4M
1.2M
1M
0.8M
0.6M
conventional
organic


Average Price vs. Total Volume sold from 2015 to 2018
Price
Volume
Average Price in USD
Total Volume sold
1.6
1.4
1.2
1
0.8
60
55M
50M
45M
40M
35M
30M
25M
20M
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018

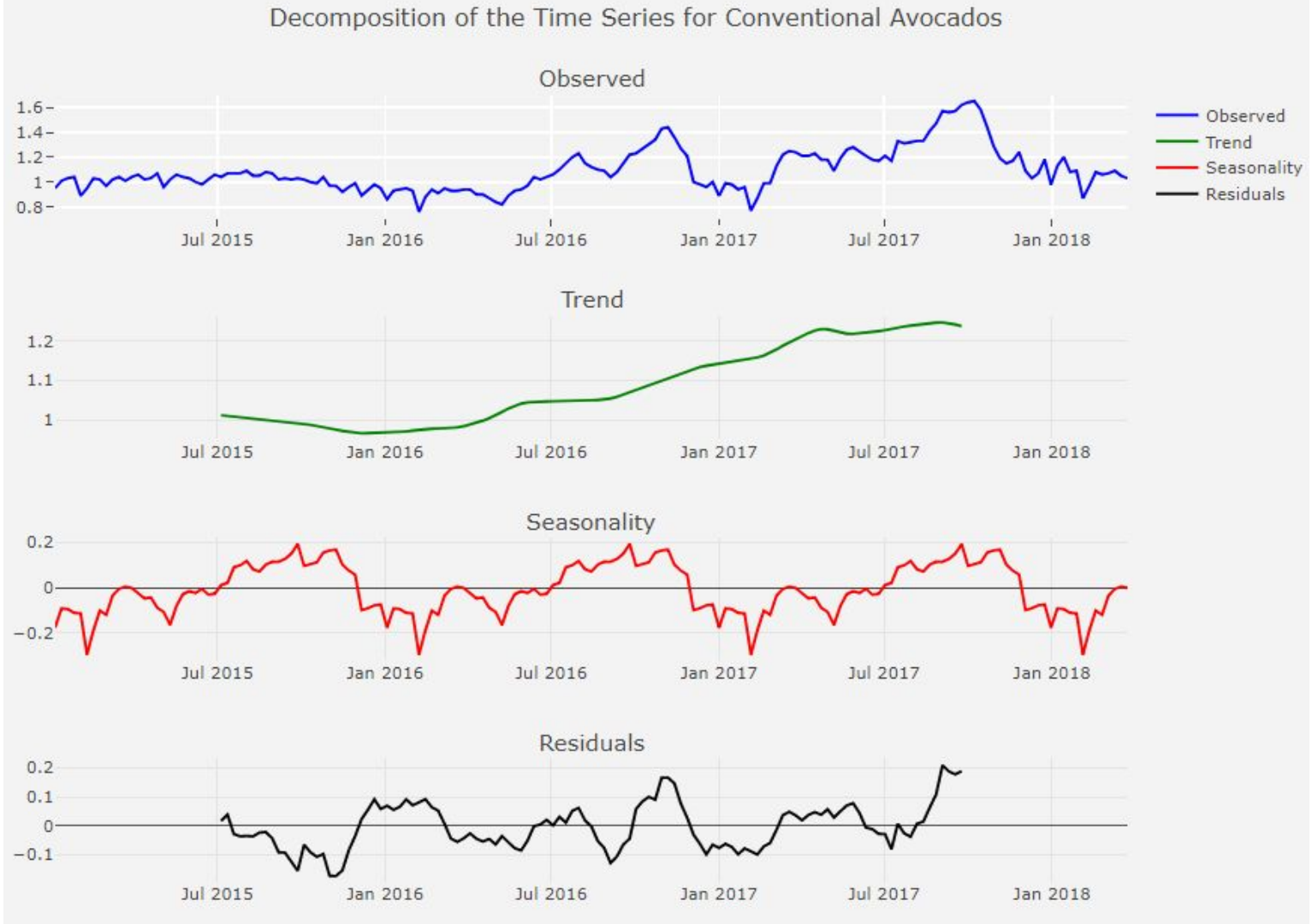

Total Volume of sold Avocados grouped by Regions (2015-2018)
Total Volume sold (in Millions)
3M
2.5M
2M
1.5M
1M
0.5M
0
West
California
SouthCentral
Northeast
Southeast
GreatLakes
Midsouth
LosAngeles
Plains
NewYork
DallasFtWorth
Houston
PhoenixTucson
WestTexNewMexico
Denver
SanFrancisco
BaltimoreWashington
Chicago
Portland
Seattle
MiamiFtLauderdale
Boston
SanDiego
Atlanta
Sacramento
Philadelphia
NorthernNewEngland
Tampa
Detroit
SouthCarolina
Orlando
LasVegas
HartfordSpringfield
RaleighGreensboro
NewOrleansMobile
CincinnatiDayton
RichmondNorfolk
HarrisburgScranton
Nashville
Charlotte
StLouis
Indianapolis
GrandRapids
Columbus
Jacksonville
Roanoke
BuffaloRochester
Pittsburgh
Louisville
Albany
Spokane
Boise
Syracuse

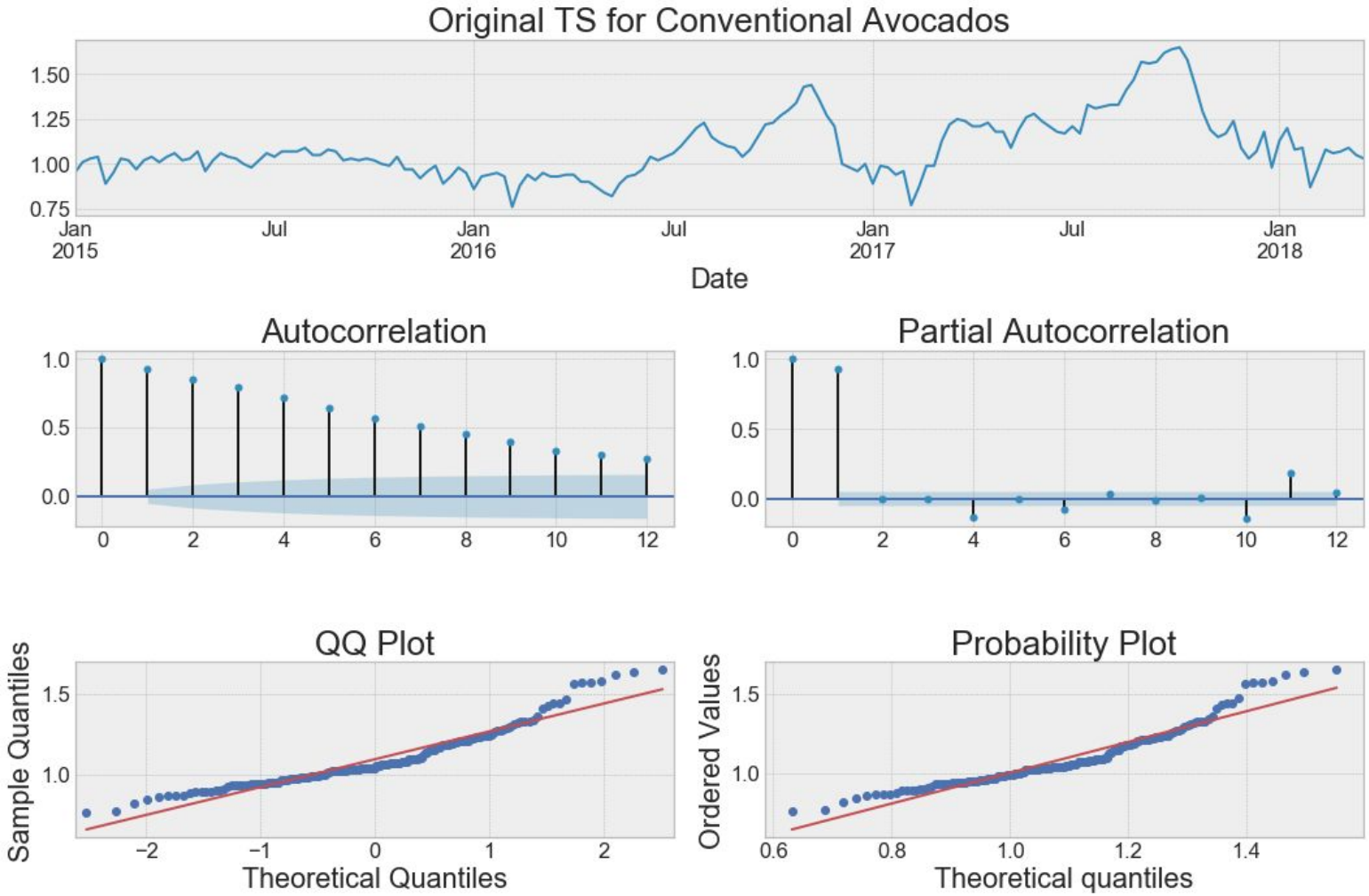
3. Test for Stationarity

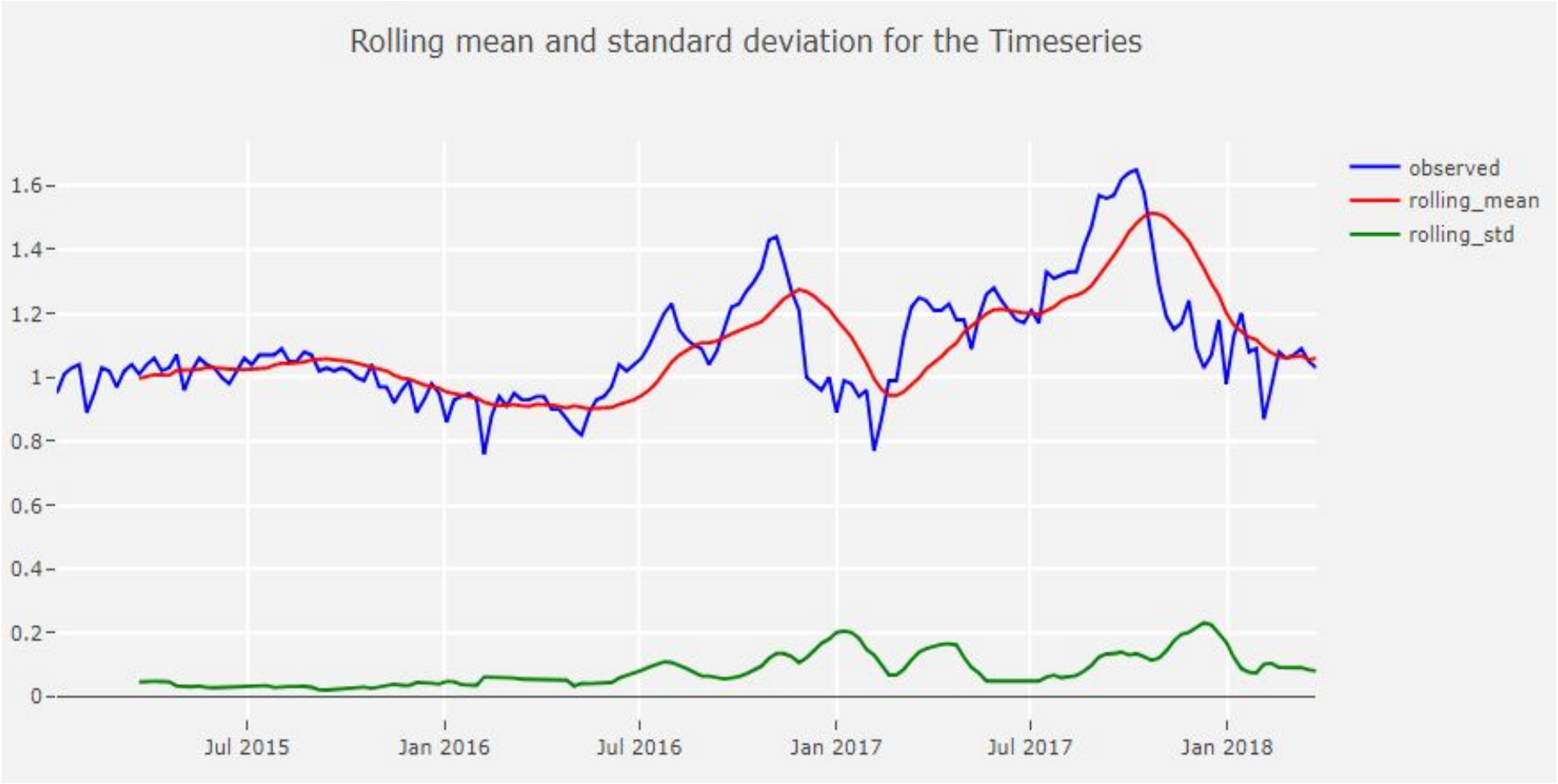
Stationarity is a statistical assumption that a time-series has:

- Constant mean
- Constant variance
- Autocovariance does not depend on time

⇒ Testing both graphically and with statistical-tests


Decomposition of the Time Series for Conventional Avocados
Observed
Trend
Seasonality
Residuals
1.6
1.4
1.2
1
0.8
1.2
1.1
1
0.2
0
−0.2
0.2
0.1
0
−0.1
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018


Original TS for Conventional Avocados
1.50
1.25
1.00
0.75
Jan 2015
Jul
Jan 2016
Jul
Jan 2017
Jul
Jan 2018
Date
Autocorrelation
1.0
0.5
0.0
0
2
4
6
8
10
12
Partial Autocorrelation
1.0
0.5
0.0
0
2
4
6
8
10
12
QQ Plot
Sample Quantiles
1.5
1.0
−2
−1
0
1
2
Theoretical Quantiles
Probability Plot
Ordered Values
1.5
1.0
0.6
0.8
1.0
1.2
1.4
Theoretical quantiles


Rolling mean and standard deviation for the Timeseries
observed
rolling_mean
rolling_std
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018

Tests for Stationarity

1) **Augmented Dickey-Fuller Test**

H_0: The Time-Series is not stationary

H_A: The Time-Series is stationary

2) **KPSS-Test**

H_0: The Time-Series is stationary

H_A: The Time-Series is not stationary

Results for the conducted tests

Test-Statistic	p-value
Augmented-Dickey Fuller Test	0.098633
KPSS	0.034961

Assumption: Time-Series is not-stationary

4. Correcting for Stationarity

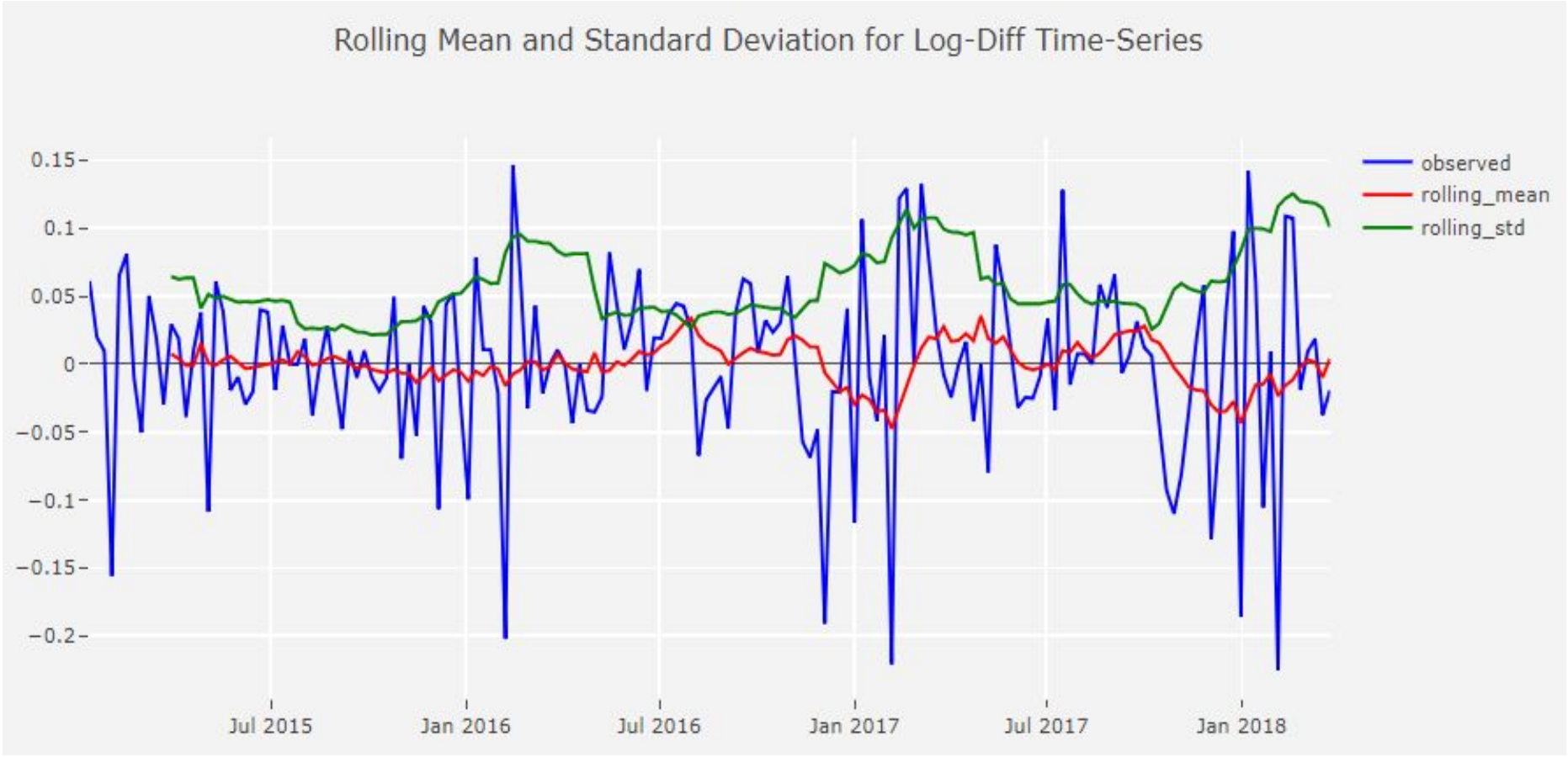
As seen from the graph and test-statistics before, the TS seems to be not stationary:

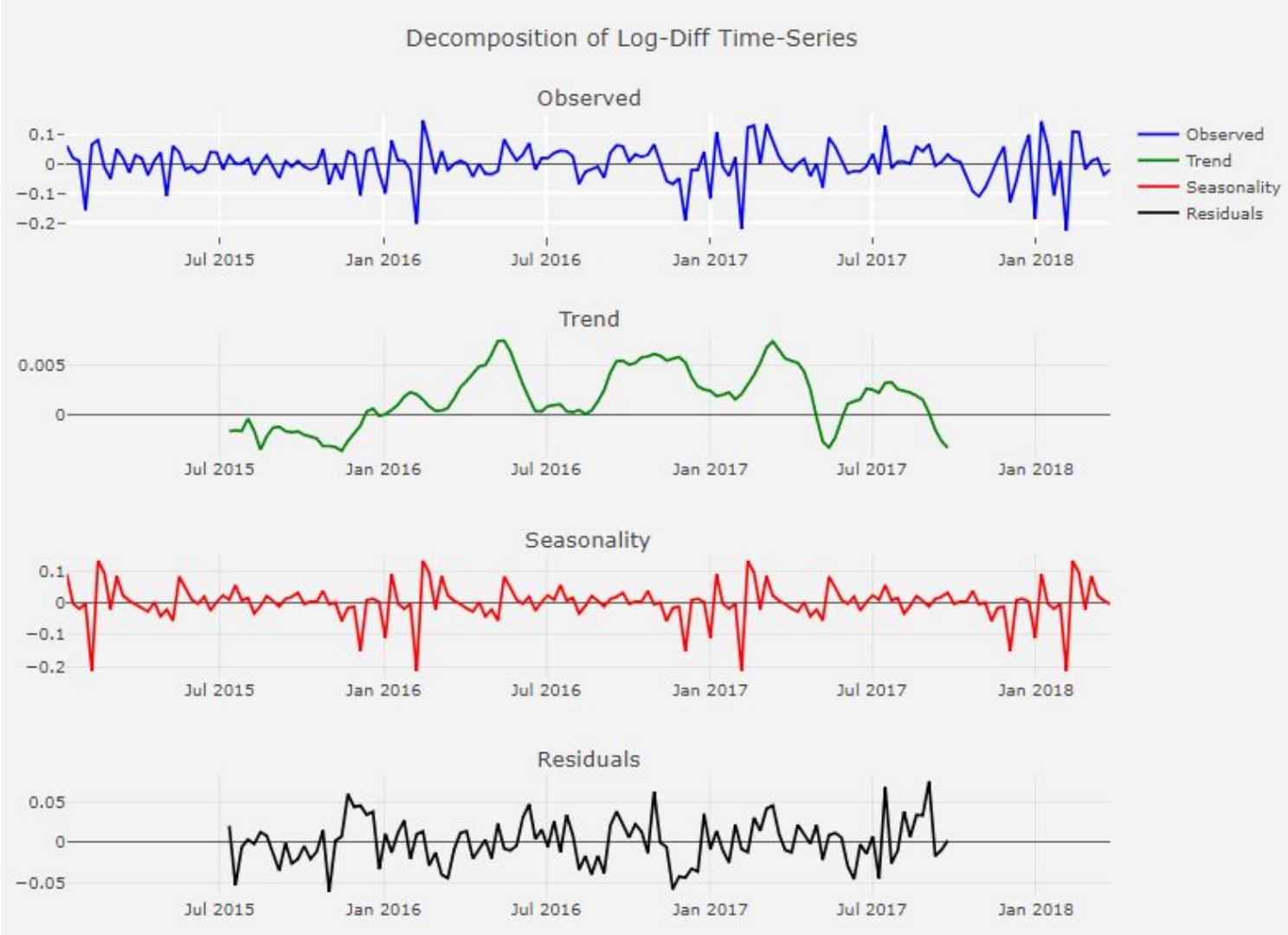
- Trend: mean is not constant over time
- Seasonality: variance is not constant over time

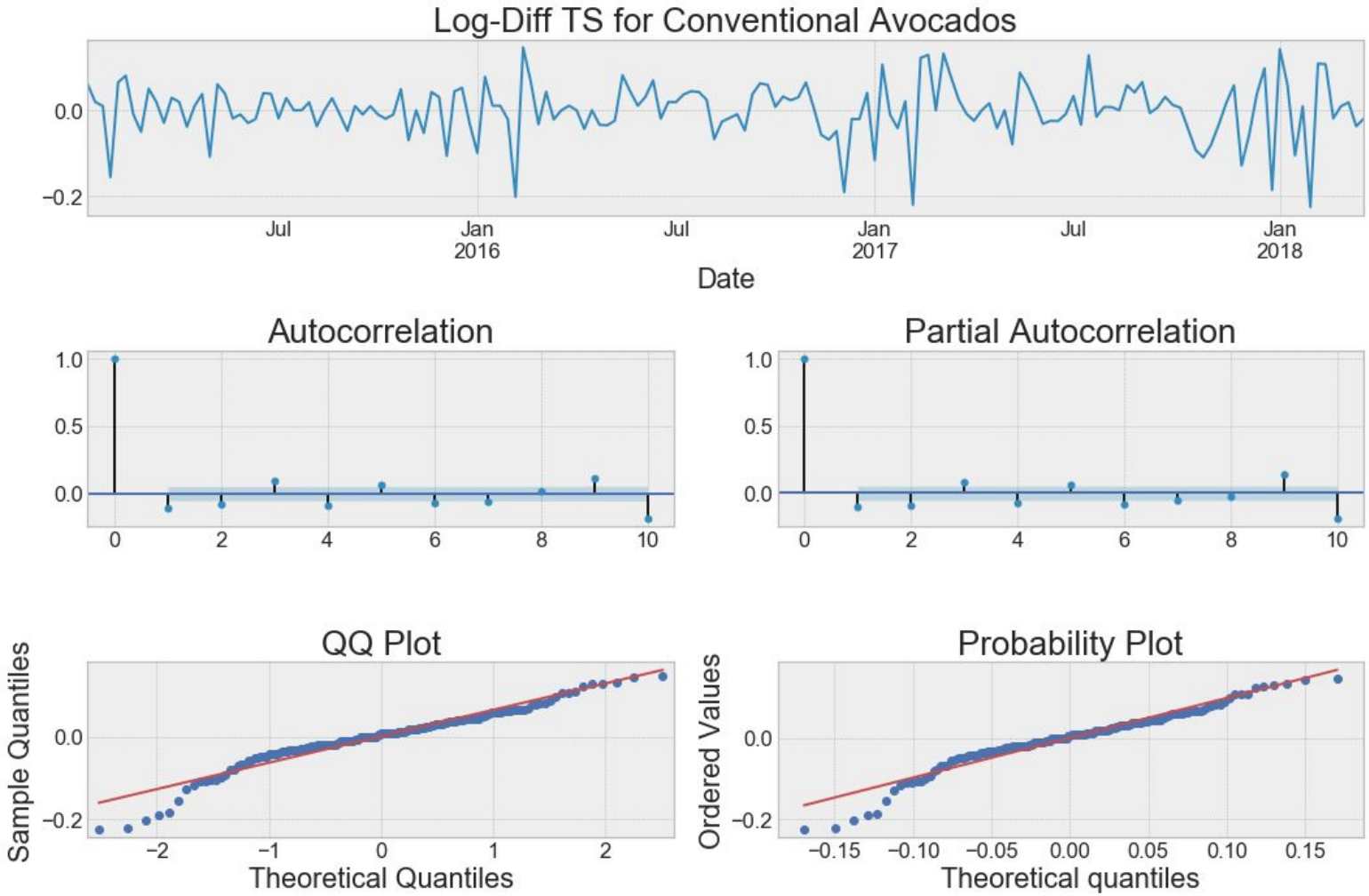
⇨ Correct for Trend and Seasonality in order to make TS stationary

Possible solutions:

- Taking the Log
- Smoothing (rolling averages, e.g. weekly, monthly, …)
- Differencing
- etc.


Rolling Mean and Standard Deviation for Log-Diff Time-Series
0.15
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018
observed
rolling_mean
rolling_std

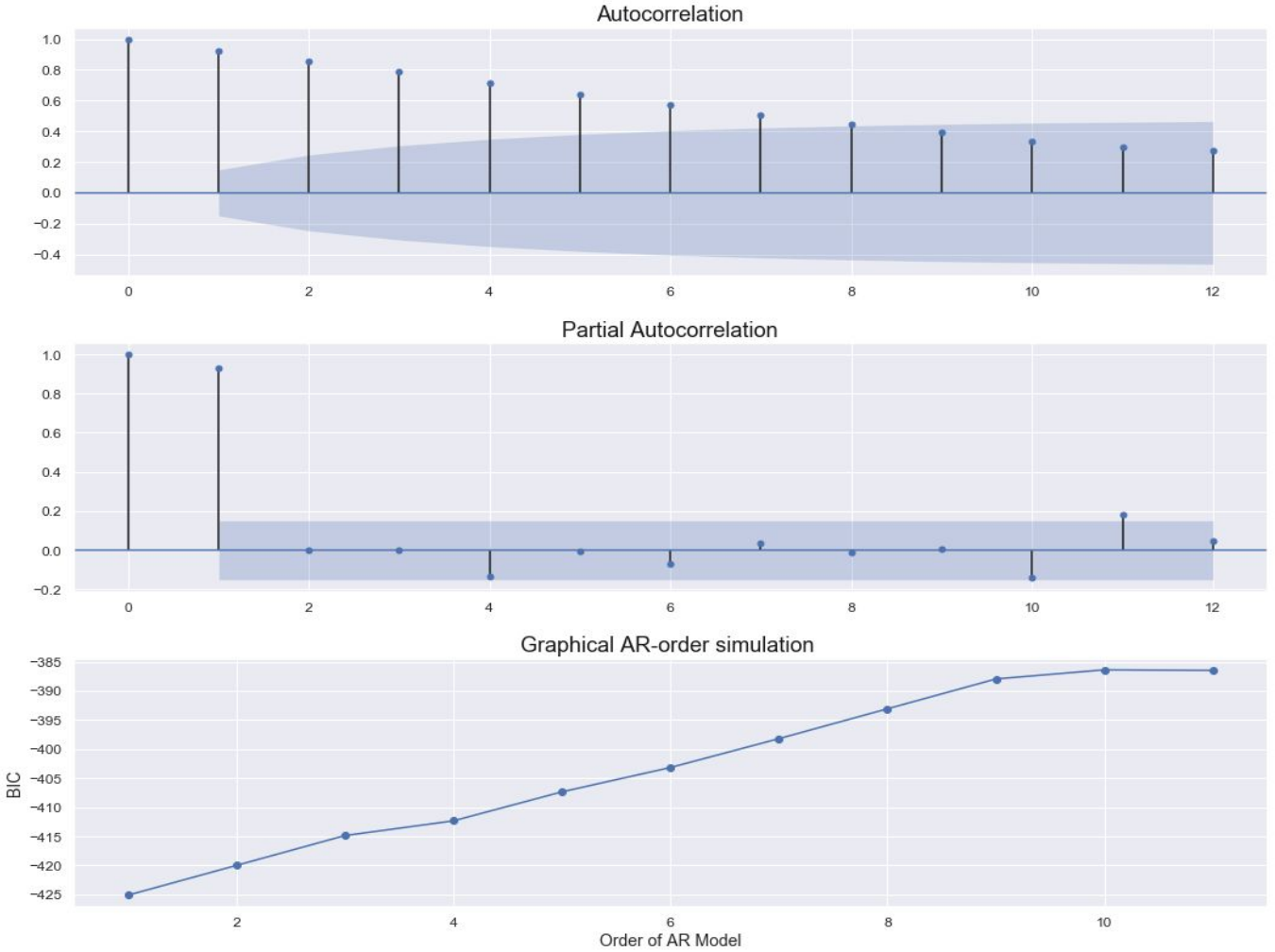

Decomposition of Log-Diff Time-Series
Observed
Trend
Seasonality
Residuals
0.1
0
−0.1
−0.2
0.005
0.05
−0.05
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018

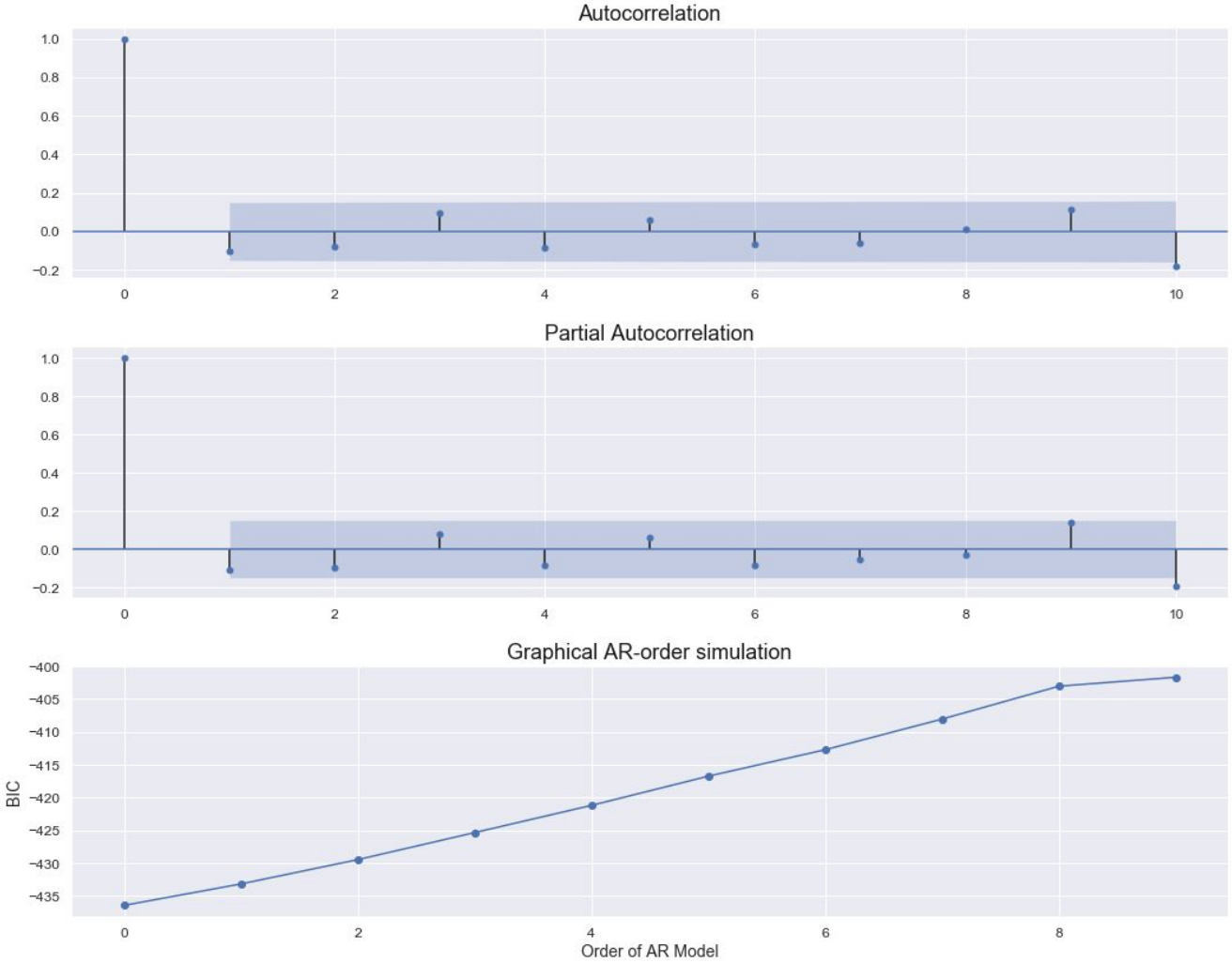

Log-Diff TS for Conventional Avocados
0.0
−0.2
Jul
Jan
2016
Jul
Jan
2017
Jul
Jan
2018
Date
Autocorrelation
1.0
0.5
0.0
0
2
4
6
8
10
Partial Autocorrelation
1.0
0.5
0.0
0
2
4
6
8
10
QQ Plot
Sample Quantiles
0.0
−0.2
−2
−1
0
1
2
Theoretical Quantiles
Probability Plot
Ordered Values
0.0
−0.2
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
Theoretical quantiles

Results for conducted tests on Log-Diff-Time-Series

Test-Statistic	p-value
Augmented-Dickey Fuller Test	0.000144
KPSS	0.10

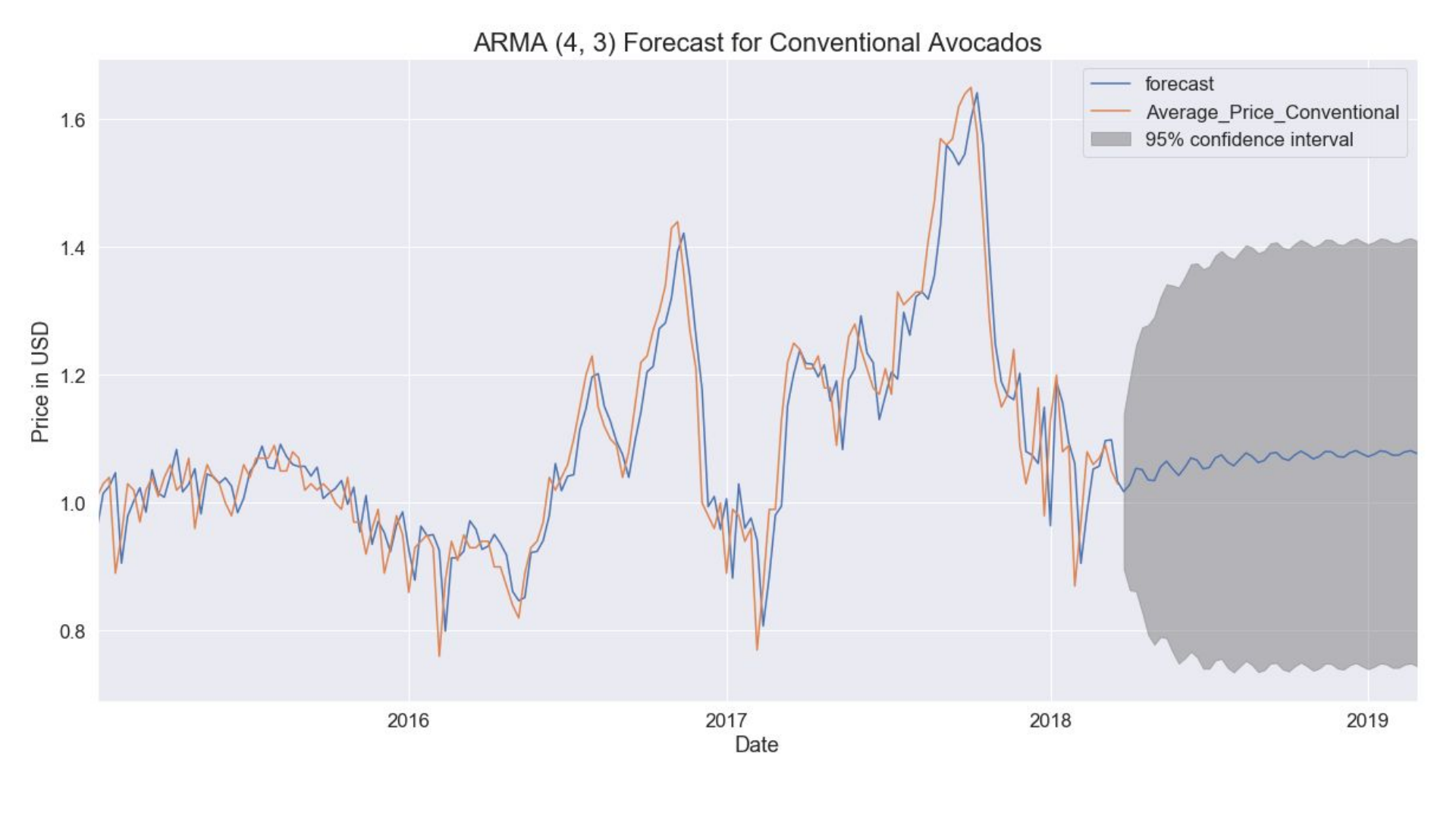
Assumption: Time-Series is stationary

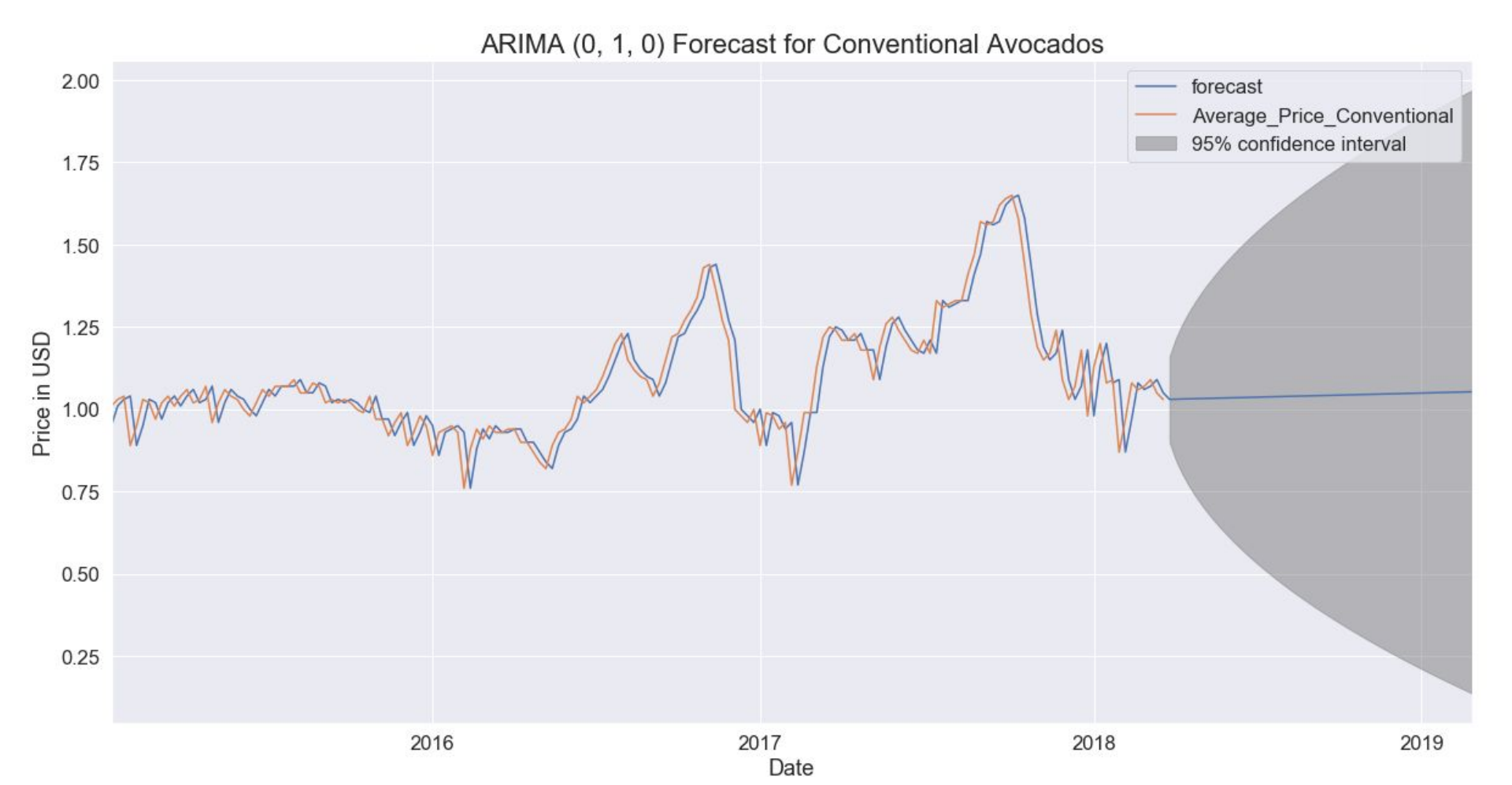

Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
2
4
6
8
10
12
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
2
4
6
8
10
12
Graphical AR-order simulation
BIC
−385
−390
−395
−400
−405
−410
−415
−420
−425
2
4
6
8
10
Order of AR Model


Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
2
4
6
8
10
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
2
4
6
8
10
Graphical AR-order simulation
BIC
−400
−405
−410
−415
−420
−425
−430
−435
0
2
4
6
8
Order of AR Model

5.1 Model-Fitting: Traditional approach

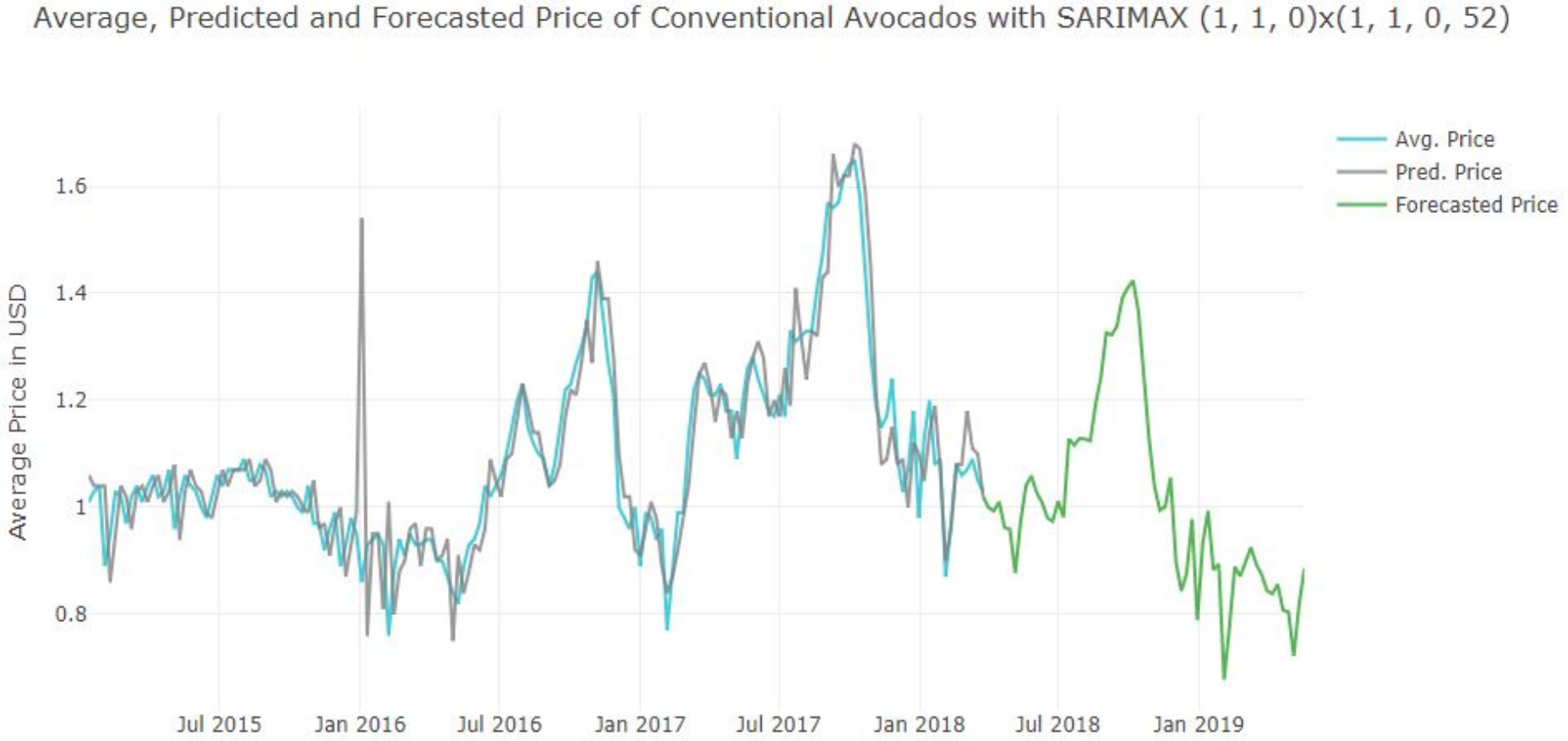
- since model-fits for log-diff TS yielded in bad results, we tried to fit a model for the original TS
- following fits were conducted:
  - ARMA (4, 3)
  - ARIMA (0, 1, 0)
  - SARIMA (1, 1, 0) x (1, 1, 0, 52)
- model evaluation with statistical metrics:
  - RMSE (= Root Mean Square Error)
  - MAPE (= Mean Absolute Percentage Error)
  - MAE (= Mean Absolute Error)

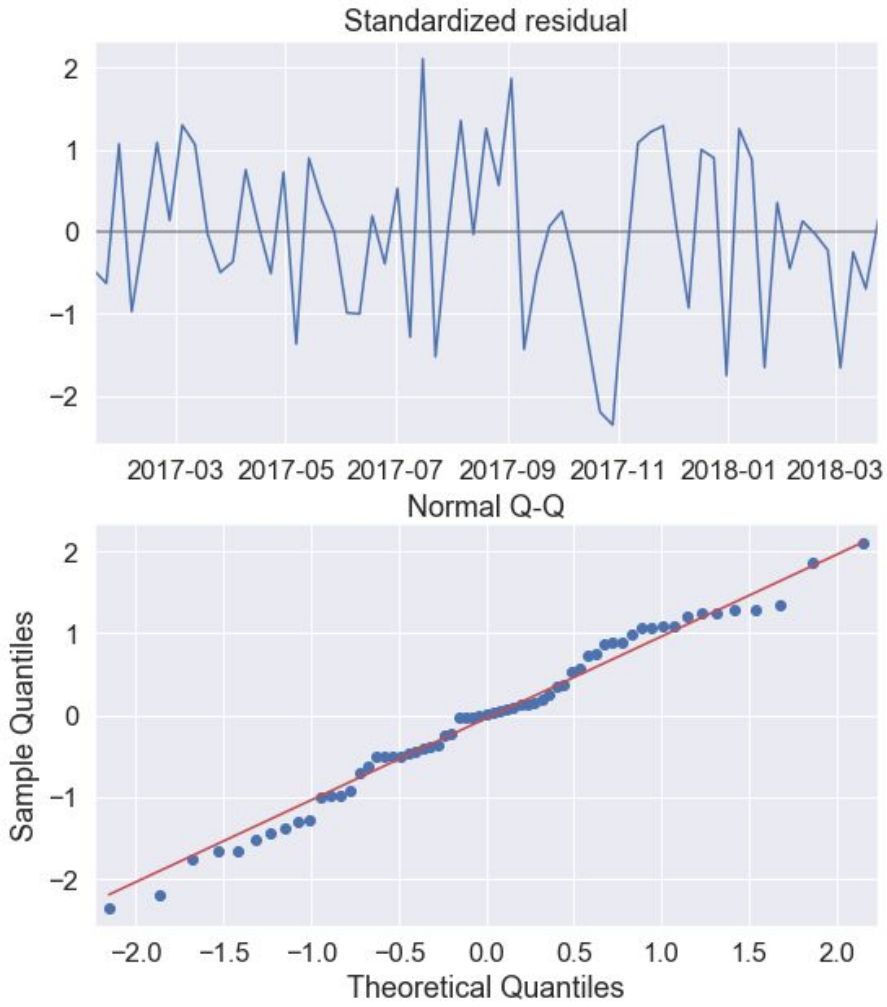

ARMA (4, 3) Forecast for Conventional Avocados
forecast
Average_Price_Conventional
95% confidence interval
Price in USD
1.6
1.4
1.2
1.0
0.8
2016
2017
2018
2019
Date

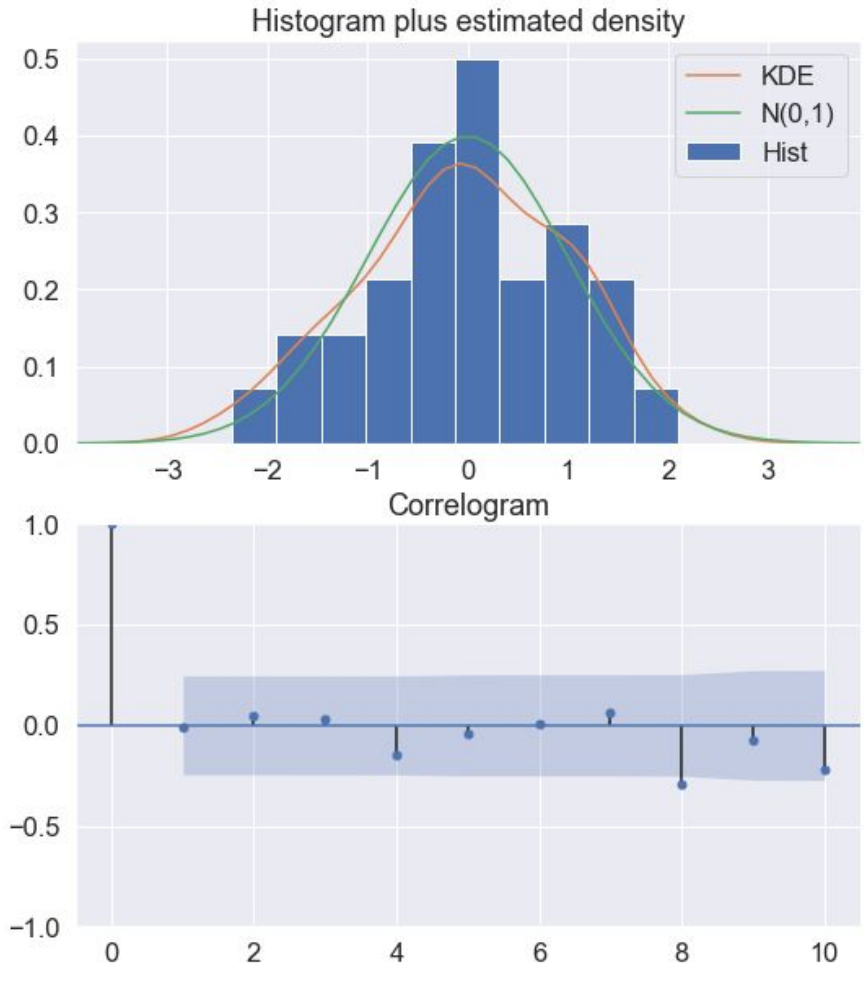

ARIMA (0, 1, 0) Forecast for Conventional Avocados
forecast
Average_Price_Conventional
95% confidence interval
Price in USD
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
2016
2017
2018
2019
Date

Short Detour: SARIMA

- ARIMA can model data with a trend, but not with a seasonal component
- extension of ARIMA: **Seasonal Autoregressive Integrated Moving Average** (= SARIMA)
- configuration of SARIMA:
  - Trend Elements:
    - p: Trend autoregressive order
    - d: Trend difference order
    - q: Trend moving average order
  - Seasonal Elements:
    - P: Seasonal autoregressive order
    - D: Seasonal difference order
    - Q: Seasonal moving average order
    - m: Number of time steps (e.g.: 12 for monthly data, 52 for weekly data etc.)

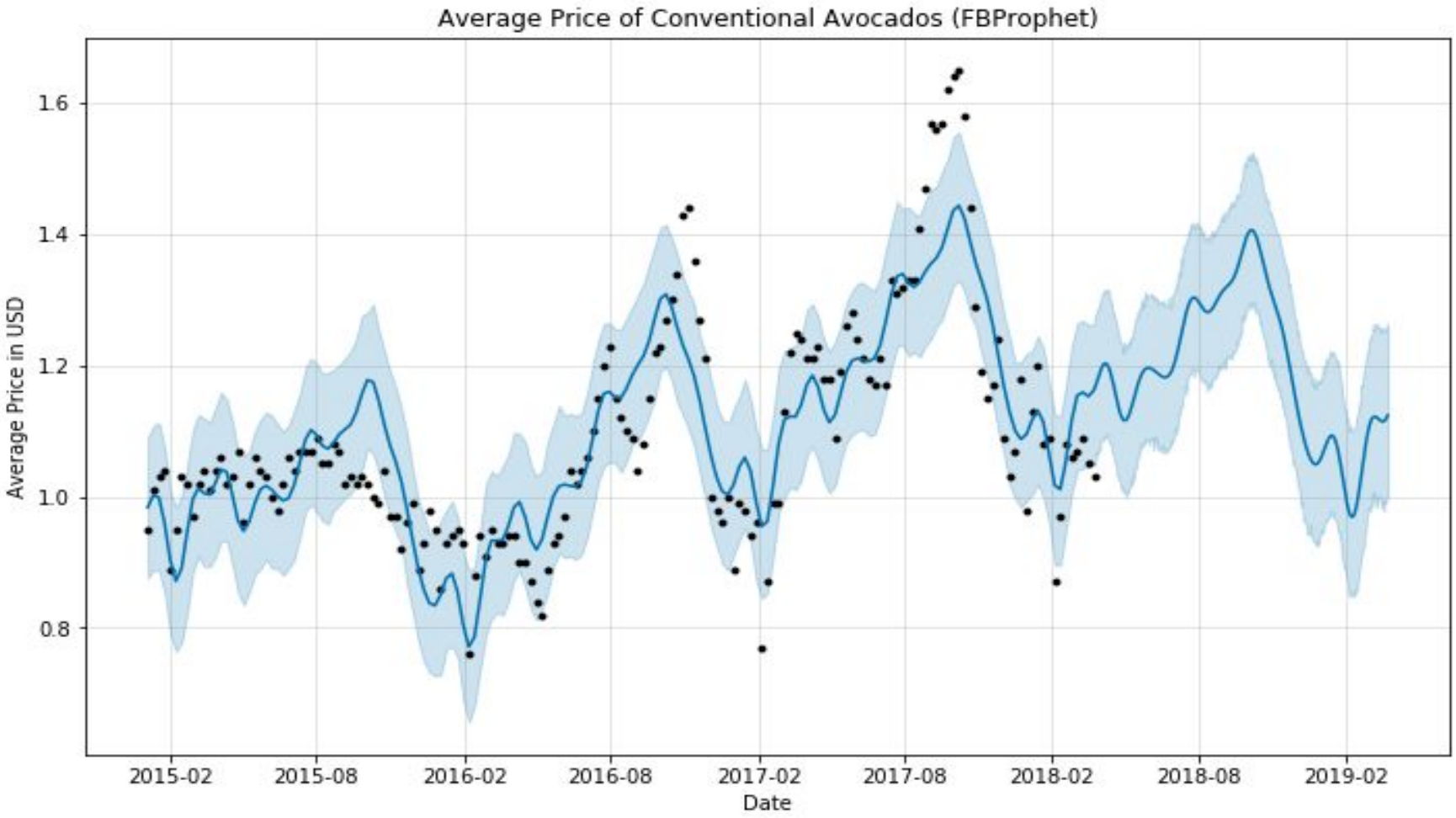

Average, Predicted and Forecasted Price of Conventional Avocados with SARIMAX (1, 1, 0)x(1, 1, 0, 52)
Avg. Price
Pred. Price
Forecasted Price
Average Price in USD
1.6
1.4
1.2
1
0.8
Jul 2015
Jan 2016
Jul 2016
Jan 2017
Jul 2017
Jan 2018
Jul 2018
Jan 2019


Standardized residual
2
1
0
−1
−2
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
Normal Q-Q
Sample Quantiles
Theoretical Quantiles
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Histogram plus estimated density
KDE
N(0,1)
Hist
0.5
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3
Correlogram
1.0
0.5
0.0
−0.5
−1.0
0
2
4
6
8
10

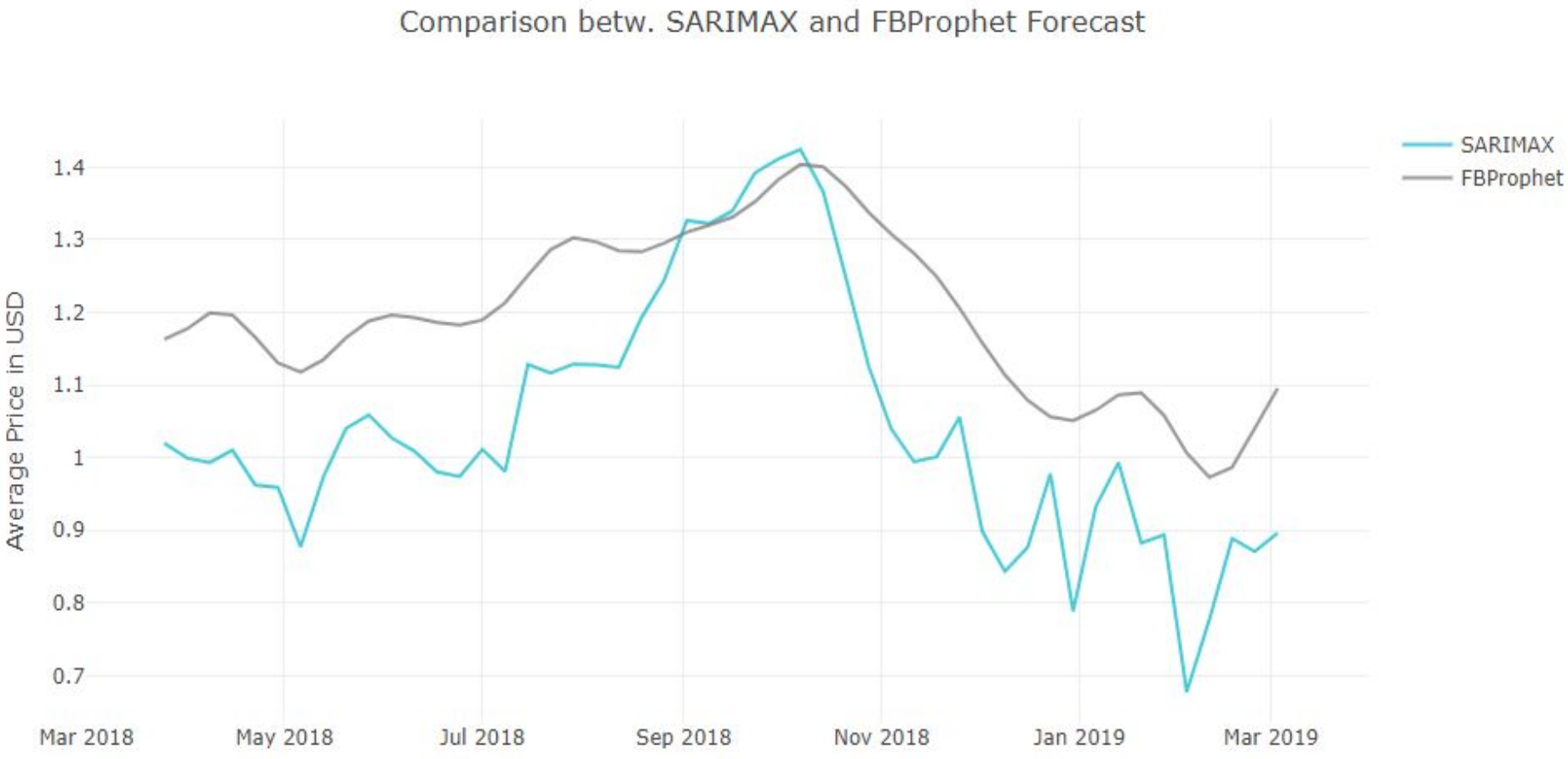
5.2 Forecast with FBProphet

- created by the Core Data Science Team (> 650 Data Scientists) at *Facebook*
- developed for Python and R
- library was open-sourced in 2017
- forecasts for TS-data based on additive models where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects
- **does not require manuel pre-processing steps** (e.g. differencing to make the data stationary)
- works best for TS which has strong seasonal effects and several seasons of historical data


Average Price of Conventional Avocados (FBProphet)
Average Price in USD
1.6
1.4
1.2
1.0
0.8
2015-02
2015-08
2016-02
2016-08
2017-02
2017-08
2018-02
2018-08
2019-02
Date

5.3 Comparison

	RMSE	MAE	MAPE
SARIMA	0.109	0.055	0.057
FBProphet	0.086	0.0675	0.061


Comparison betw. SARIMAX and FBProphet Forecast
SARIMAX
FBProphet
Average Price in USD
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
Mar 2018
May 2018
Jul 2018
Sep 2018
Nov 2018
Jan 2019
Mar 2019

6. Conclusion

- First Research Question:
  - Yes, as clearly seen from the graphs, 2017 was *the year* for avocados
  - Average price and volume of sold avocados rose tremendously
- Second Research Question:
  - Price range for predicted avocados between 2018 and early 2019: 0.80 - 1.10 USD

Sources

URL: https://www.kaggle.com/neuromusic/avocado-prices, Zugriff 27.6.2019

URL: https://hassavocadoboard.com/, Zugriff 27.6.2019