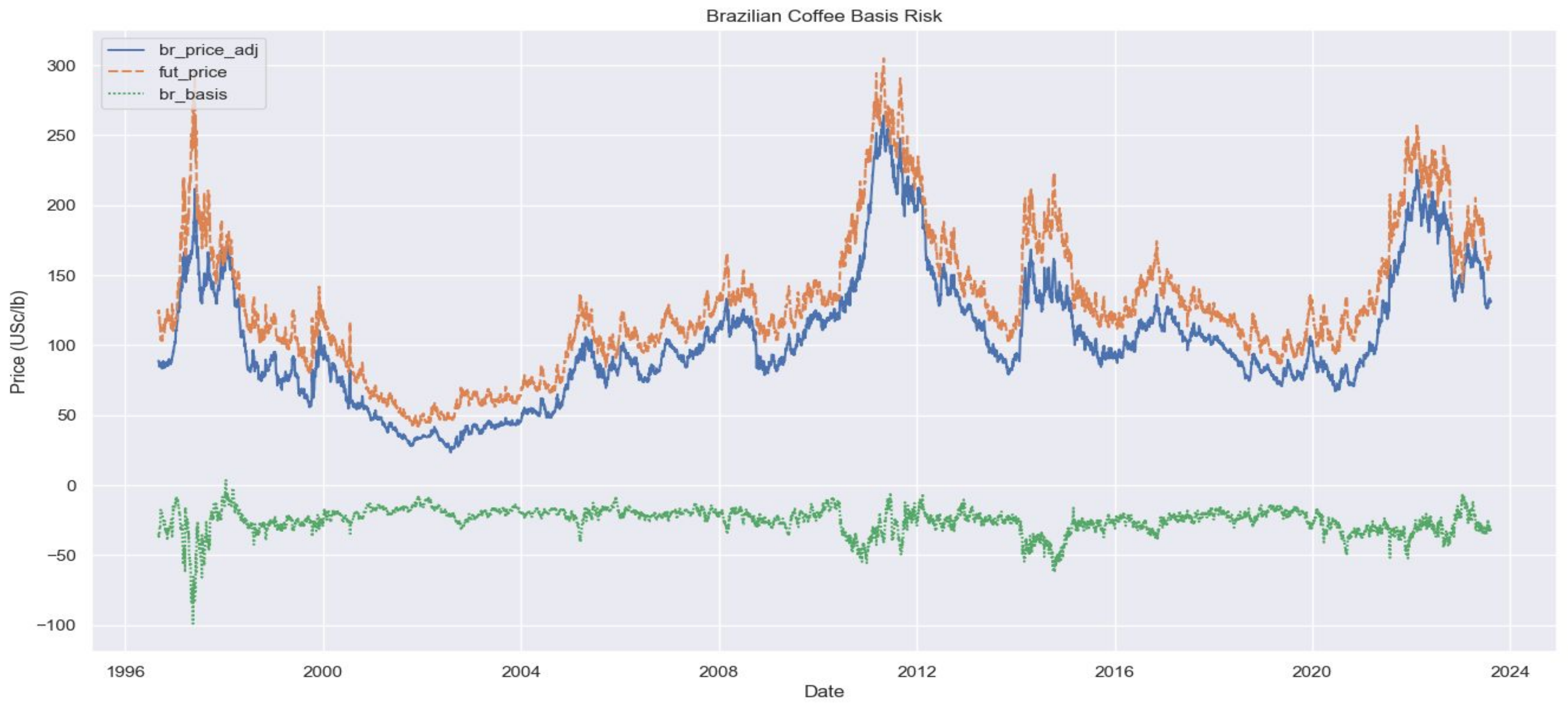
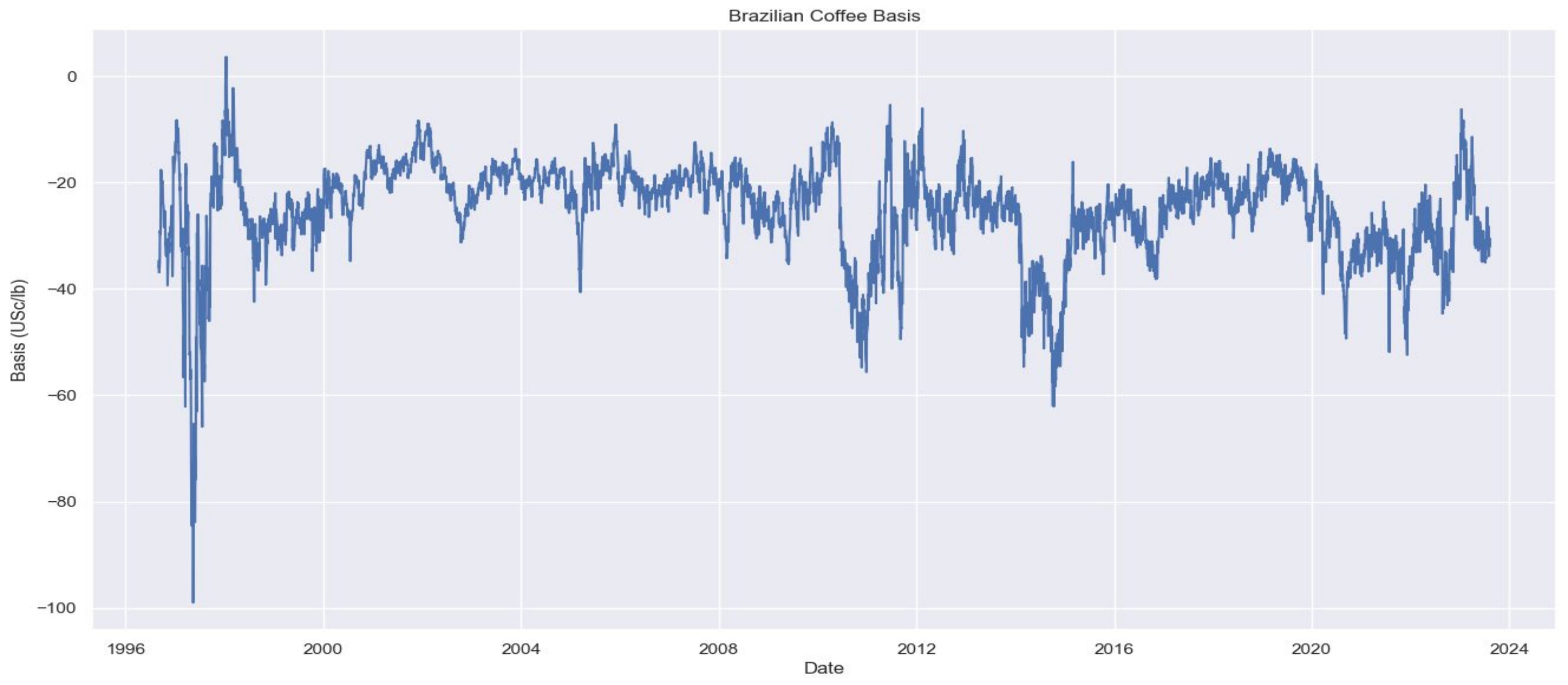
Predictive Modeling of Brazilian Basis Risk: Insights from Time Series and ML Analysis

3rd Capstone project - Data Scientist Career Path

Coffee Series and Brazilian Coffee Basis Risk



Basis Risk



Brazilian Coffee Basis
Basis (USc/lb)
0
−20
−40
−60
−80
−100
1996
2000
2004
2008
2012
2016
2020
2024
Date

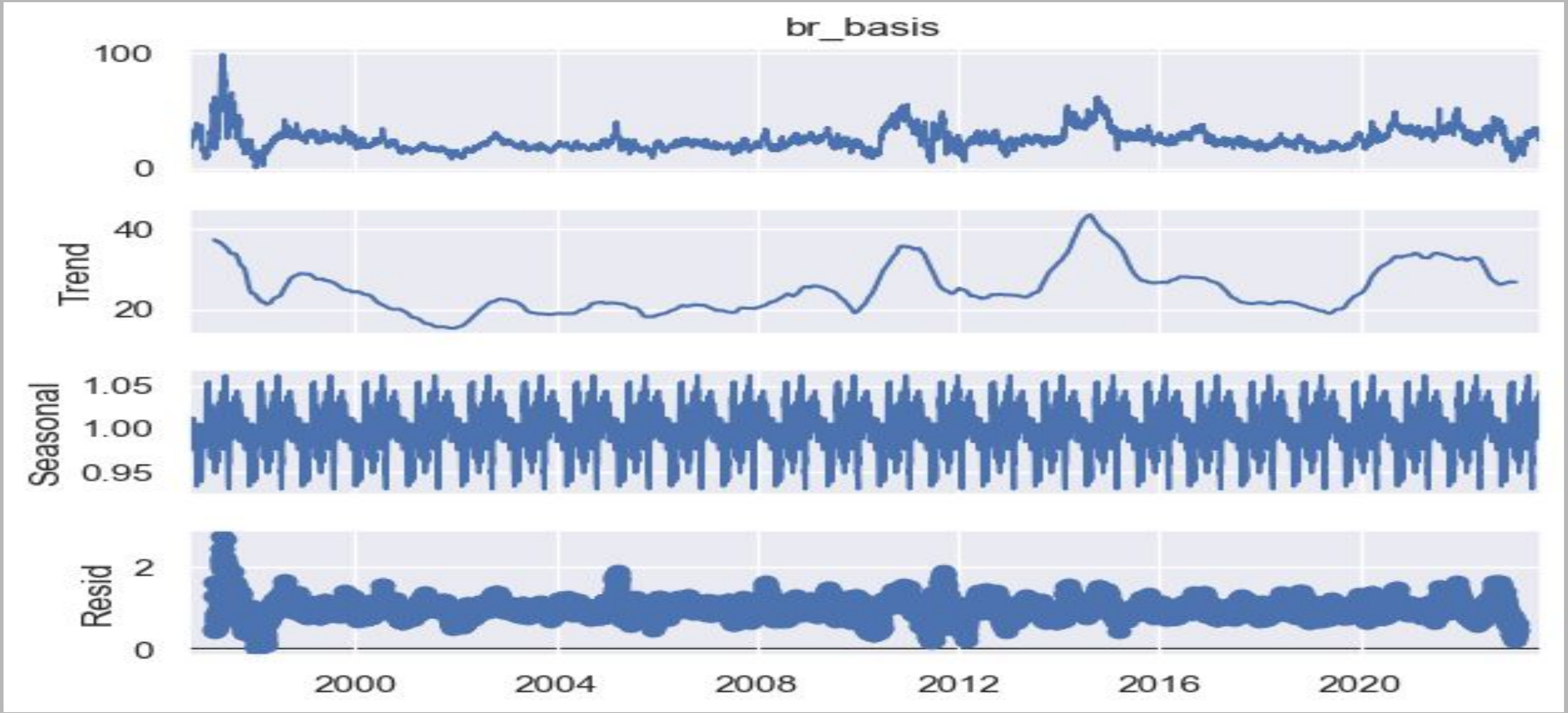
Coffee Series Descriptive Statistics

	br_price	br_price_adj	fut_price	br_basis
count	6515.000000	6515.000000	6515.000000	6515.000000
mean	138.433088	104.653779	129.811604	-25.157825
std	59.825587	45.227437	49.486644	9.009344
min	30.920000	23.375155	41.500000	-98.992103
25%	103.840000	78.501813	101.875000	-29.039687
50%	127.380000	96.297775	121.100000	-23.507018
75%	169.150000	127.875401	153.350000	-19.279366
max	349.390000	264.134711	304.900000	3.641727

Time Series used in the study

```
<class 'pandas.core.frame.DataFrame'>
PeriodIndex: 4911 entries, 1996-09-03 to 2016-12-29
Freq: D
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   br_price      4911 non-null   float64
 1   br_price_adj  4911 non-null   float64
 2   fut_price     4911 non-null   float64
 3   br_basis      4911 non-null   float64
 4   fx            4911 non-null   float64
dtypes: float64(5)
memory usage: 230.2 KB
```

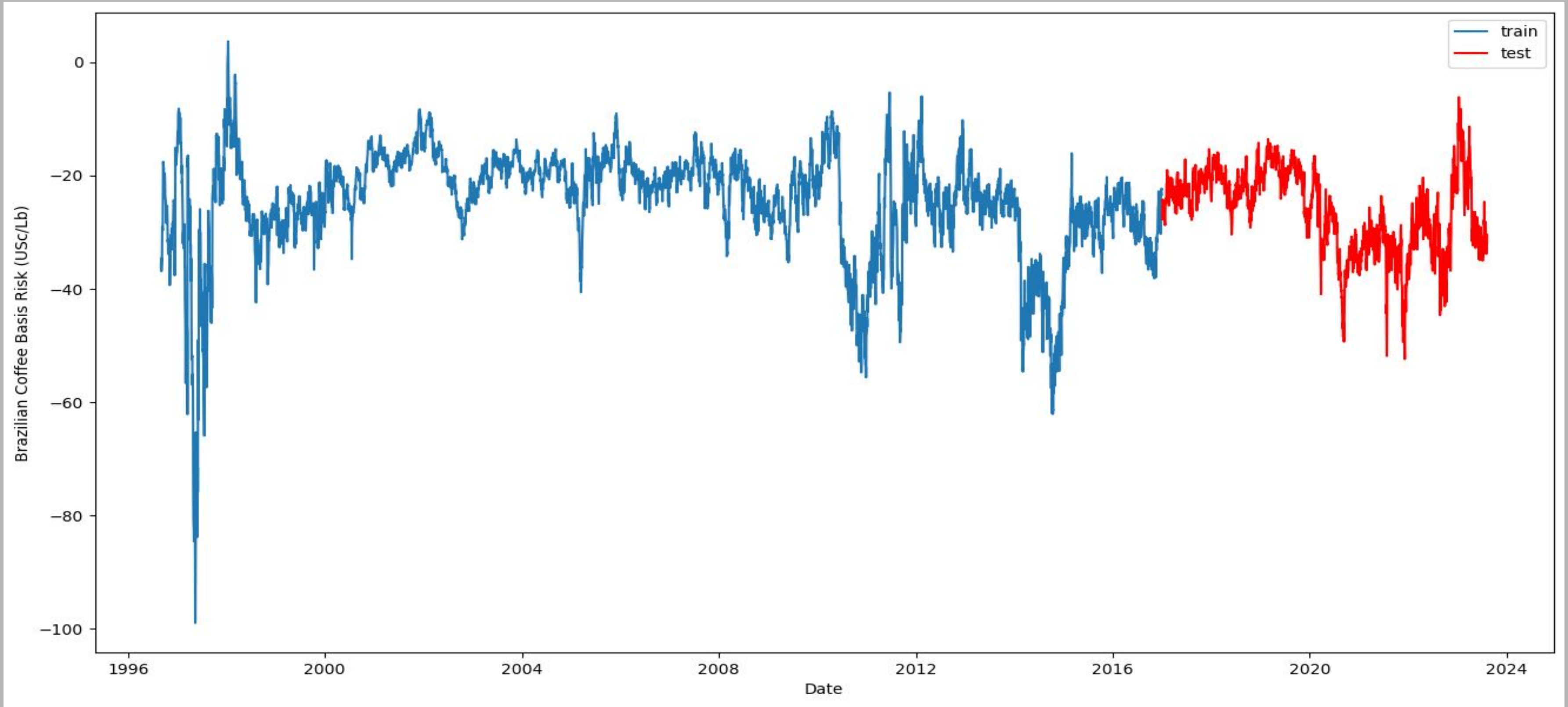
Brazilian coffee Basis Decomposition



Brazilian coffee Basis Correlation

	br_price	br_price_adj	fut_price	br_basis
br_price	1.000000	1.000000	0.778723	-0.132908
br_price_adj	1.000000	1.000000	0.778723	-0.132908
fut_price	0.778723	0.778723	1.000000	-0.725300
br_basis	-0.132908	-0.132908	-0.725300	1.000000

Train Test Split



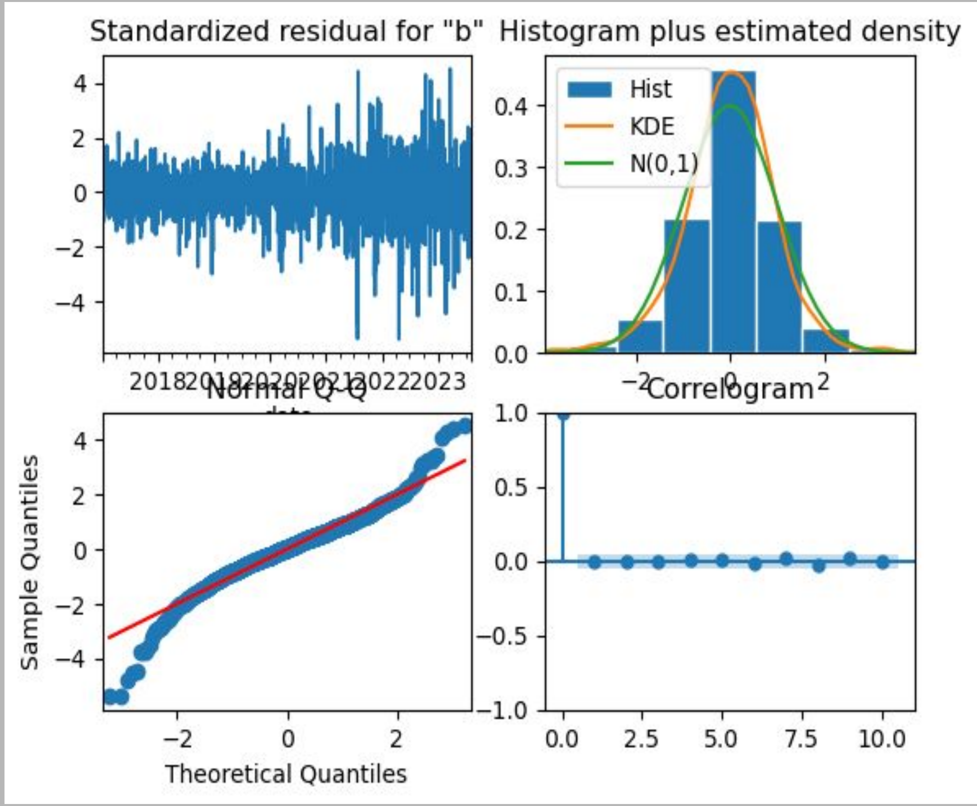
Base Model - ARIMA(1,1,3)

SARIMAX Results

Dep. Variable:	br_basis	No. Observations:	1604
Model:	ARIMA(1, 1, 3)	Log Likelihood	-3309.003
Date:	Fri, 01 Dec 2023	AIC	6628.007
Time:	08:58:29	BIC	6654.905
Sample:	01-03-2017	HQIC	6637.994
	- 08-08-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7644	0.065	11.682	0.000	0.636	0.893
ma.L1	-0.8820	0.067	-13.091	0.000	-1.014	-0.750
ma.L2	0.0041	0.028	0.143	0.886	-0.052	0.060
ma.L3	-0.0072	0.027	-0.265	0.791	-0.060	0.046
sigma2	3.6347	0.083	43.573	0.000	3.471	3.798

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	671.30
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	3.64	Skew:	-0.37
Prob(H) (two-sided):	0.00	Kurtosis:	6.08



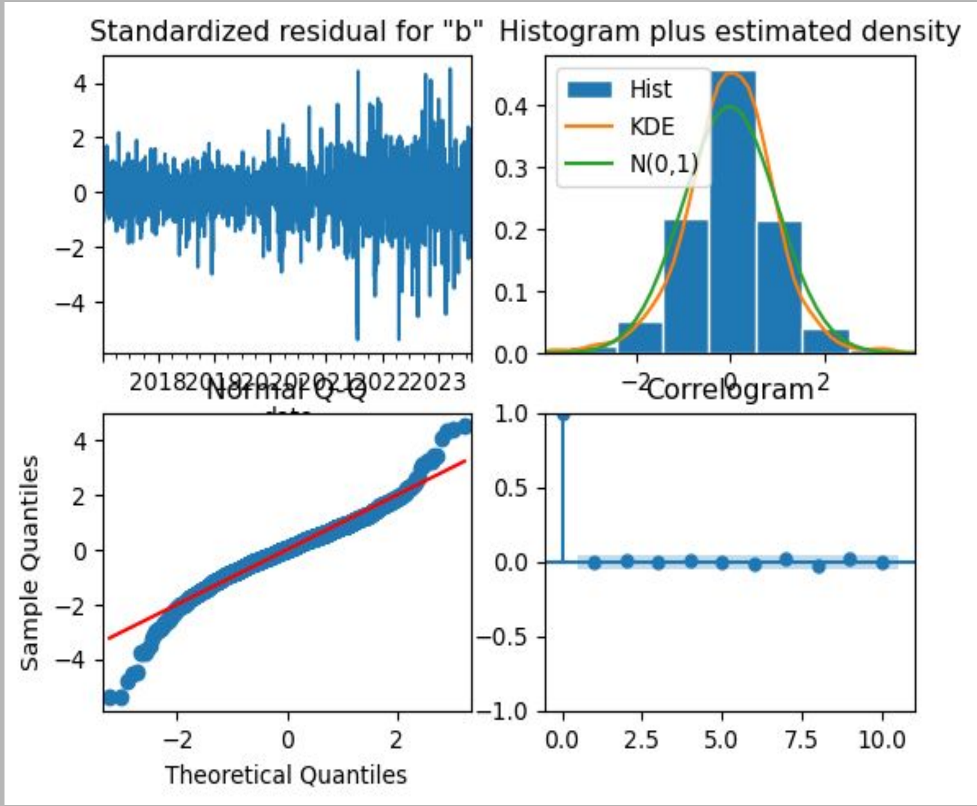
ARIMA(1,1,1)

SARIMAX Results

Dep. Variable:	br_basis	No. Observations:	1604
Model:	ARIMA(1, 1, 1)	Log Likelihood	-3309.027
Date:	Fri, 01 Dec 2023	AIC	6624.055
Time:	08:59:26	BIC	6640.193
Sample:	01-03-2017	HQIC	6630.047
	- 08-08-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7743	0.037	20.806	0.000	0.701	0.847
ma.L1	-0.8908	0.027	-33.557	0.000	-0.943	-0.839
sigma2	3.6349	0.082	44.480	0.000	3.475	3.795

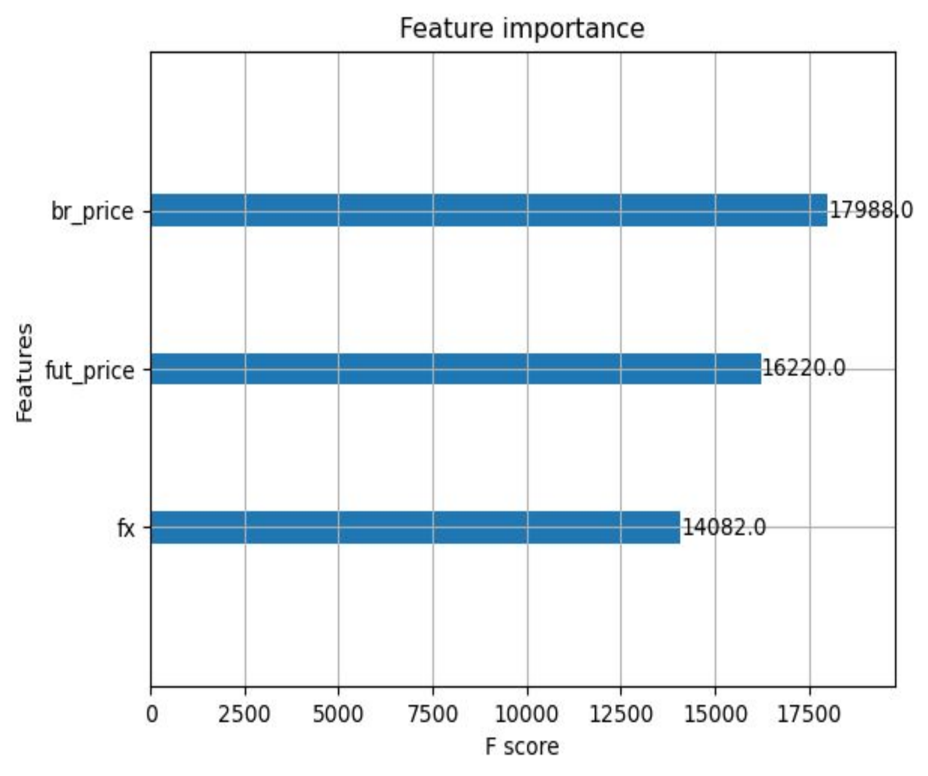
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	676.45
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	3.65	Skew:	-0.37
Prob(H) (two-sided):	0.00	Kurtosis:	6.10



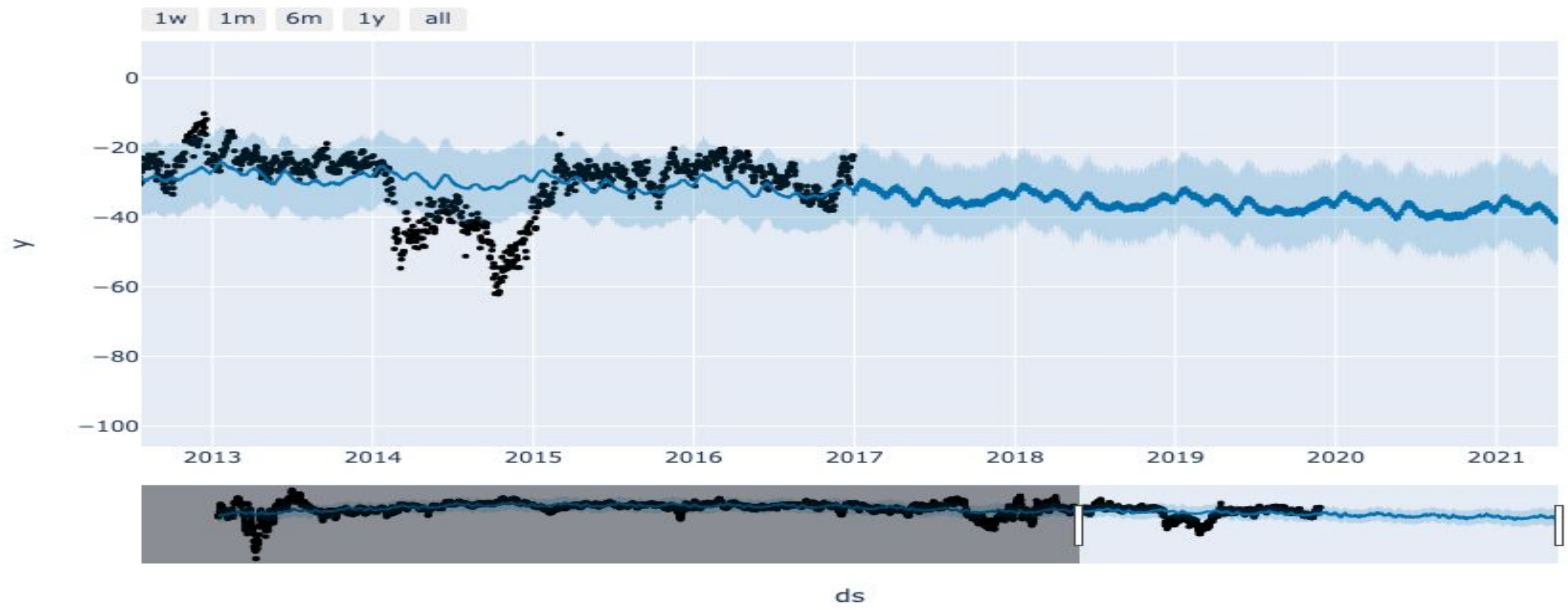
XGBoost Model

```
XGBRegressor

XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=None, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=0.005, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=0.05, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=8, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=400, n_jobs=None,
             num_parallel_tree=None, random_state=42, ...)
```

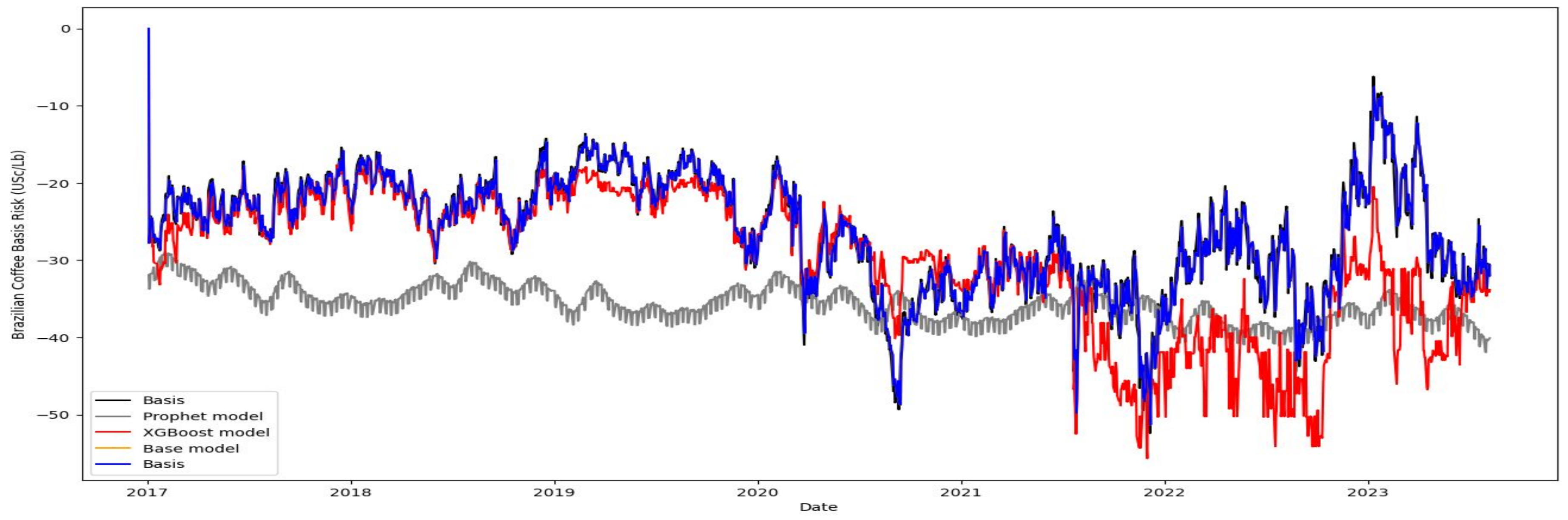


Prophet Model



Models Comparison

	Base model	ARIMA_111	XGBoost	Prophet
MAE	1.419698	1.419602	4.544749	10.098708
RMSE	2.028188	2.028214	7.071423	11.639217



Conclusion

In conclusion, the **ARIMA(1,1,1)** model emerged as the most effective among the studied models, showcasing its robust performance in modeling the Brazilian basis risk time series.

This model not only excelled in short-term predictions but also demonstrated competence in long-term forecasting. On the other hand, **XGBoost** proved to be a valuable tool, particularly well-suited for short to mid-term forecasts, offering reliable insights into the dynamics of the basis risk.

Overall, both the ARIMA and XGBoost models present strong capabilities in capturing and predicting the complexities of the Brazilian basis risk, making them practical and valuable tools for practitioners, hedgers, and investors. The accurate predictions generated by these models can serve as valuable inputs for informed decision-making in real-world applications.