Horse Race Betting: Time Series Analysis and Modelling of Winning Odds on Placing Odds

Ronald Schwalb and Samuel Devdas

May 24, 2024

Contents

1 Introduction

Betting markets on betting exchanges are well suited for testing market efficiency, human rationality (including bias formation and classification), and time series-related algorithms(Lewis & Magee, 2011). Horse racing markets are especially useful for such analyses, as they are in abundance, liquid and fast to get results. The vast availability of live and historical data, such as prices (odds and the size of placed bets), provides a rich source for technical analysis. Additionally, other information, such as a horse's recent form, genetics, and race conditions, can be used for fundamental analysis. Each bet has a specified termination point when its asset value is determined(Hausch & Ziemba, 2008).

This study analyses the price data, specifically the odds and their implied probability on win and place markets for horse races. It focusses on the movement during the final minutes before a race starts. Actual trades were considered as the market price. To generate as many data points as possible, we chose the markets and times of the races with the highest liquidity. Data from Betfair, the world's biggest betting exchange, is used, where the markets are much more efficient than at traditional bookmakers (Franck et al., 2013).

2 Literature Review

Models that utilize win-market probabilities to predict place-market probabilities in horse racing have been developed and refined over the years. Two of the most notable models are Harville's model, introduced in 1973, and Henery's model, proposed in 1981. The Harville model is the simplest model, and have been successfully implemented I the past (Hausch & Ziemba, 2008). It does however have a systemic bias of overestimating the placing probabilities of the favourite horses.

Time study analyses on Horse racing markets have been conducted and published. During this literature review, no publications were found that focusses on the causal relationships between different markets e.g. the win and place markets. (Tondapu, 2024) concluded that autocorrelation decays quickly, indicating high information efficiency however some volatility clustering are present indicating weaker efficiency.

3 Methods

The observation that place-markets are less liquid than win markets in horse racing has led to the hypothesis that win markets are more efficient, have a causal relationship with place markets, and can therefore be used for price discovery and market creation. The reasoning behind this hypothesis is based on the following points:

1. There are bigger spreads present in the place markets and desperate punters will bet at these spreads.

2. Information is assumed to reflect faster in liquid markets as more participants are trading.

The authors were so confident in this hypothesis that they developed a model in R, which connects to Betfair's API to stream live win-market odds. They applied the Harville method to place bets in the place markets. The betting amounts were determined using the Kelly criterion, which balances risk and reward to optimize wealth.

This was an example of a model being applied in the "danger zone" of Drew Conway's data science Venn diagram (Conway, Drew, 2010). A contributing factor for "putting the cart before the horses", is that to live stream data through the Betfair API, is that one should place bets to have access. It cannot be used only to retrieve data, required for the analysis.

The approach was changed, and the authors decided to perform time series analysis on historic data. The methods used were:

3.1 Data Collection and Preprocessing

Placeholder for data collection and preprocessing content.

3.2 Descriptive Analysis and Visualization

Data was segmented by horse and market type (WIN, PLACE) to compute time-bucketed mean prices. Each horse's data was summarized over defined time buckets, and visualized to display price trends over time.

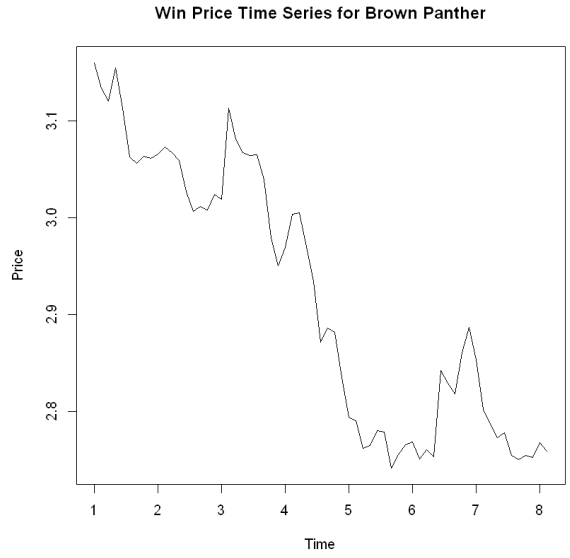

Figure 1: Win Price Time Series for Brown Panther

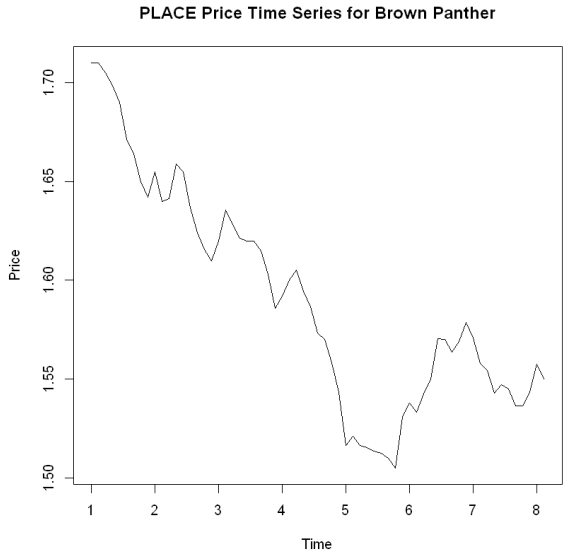

Figure 2: Place Price Time Series for Brown Panther

3.3 Time Series Analysis

Data timestamps were standardized using `POSIXct`, structured to ensure no time gaps for continuous analysis. Time series for WIN and PLACE markets were created with uniform frequency settings, filled with NA for missing values.

3.4 Stationarity

The stationarity of WIN and PLACE price time series for Brown Panther was evaluated using the Augmented Dickey-Fuller (ADF) test, confirming non-stationarity as initial tests indicated p-values greater than 0.05. Subsequent decomposition using STL revealed seasonal components, depicted in plots for both markets. To achieve stationarity, first and second differencing were applied, followed by repeated ADF tests to assess each step's effectiveness in stabilizing the series.

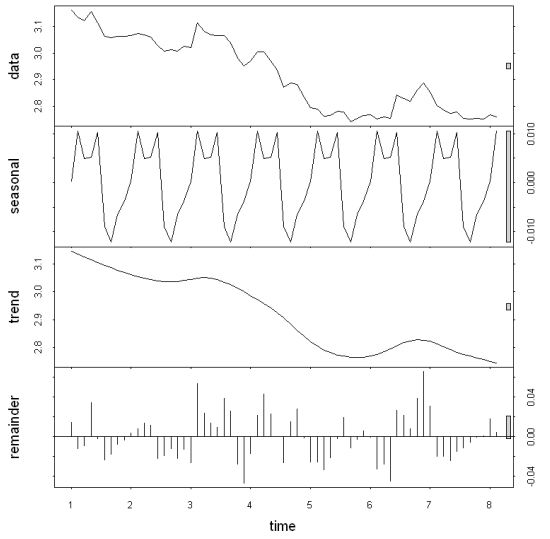

Figure 3: STL Win Price Brown Panther

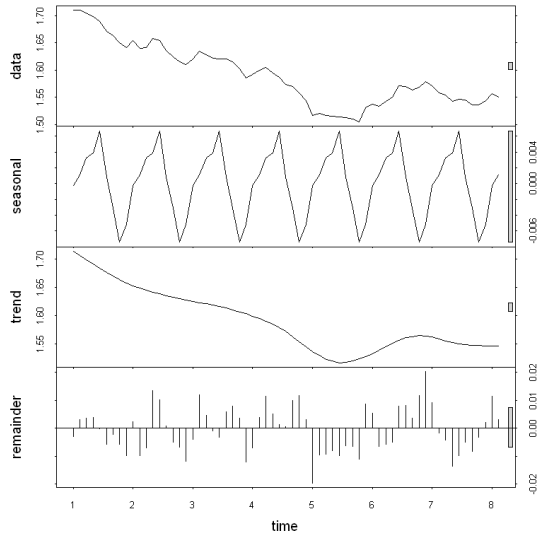

Figure 4: STL Place Price Brown Panther

4 VAR-Model and Granger Causality

The Vector Autoregression (VAR) Model and Granger causality tests were conducted to explore relationships between the second differenced WIN and PLACE price time series of Brown Panther. The VAR model was estimated considering up to three lags based on the Akaike Information Criterion (AIC). Granger causality tests were applied to determine the directional influences between the series.

5 Results

5.1 VAR Model Estimation Results

Model Specifications: Vector Autoregression (VAR) modeled the interdependencies between changes in winning and placing odds for Brown Panther. Analyzed variables were differences of winning odds (W_{diff2}) and placing odds (P_{diff2}), with a constant included.

Estimated Equations: Winning Odds: W_diff2 = -0.662(W_diff2.l1) - 0.540(W_diff2.l2) - 0.346(W_diff2.l3) + 0.198(P_diff2.l1) + 0.022(P_diff2.l2) - 0.175(P_diff2.l3) + 0.00054(const) Placing Odds: P_diff2 = 0.077(W_diff2.l1) + 0.037(W_diff2.l2) - 0.004(W_diff2.l3) - 0.669(P_diff2.l1) - 0.564(P_diff2.l2) - 0.251(P_diff2.l3) + 0.00038(const)

Statistical Insights: Significant predictors included $W_{diff2.l1}$, $W_{diff2.l2}$, $P_{diff2.l1}$ for winning odds, and $P_{diff2.l1}$, $P_{diff2.l2}$ for placing odds, indicating a strong past influence on present odds.

Model Performance: Multiple R-squared values were 35.41% for winning odds and 31.36% for placing odds. Both models' F-statistics confirmed statistical significance, validating the model's predictive reliability.

Interpretation: Significant negative coefficients suggest a mean-reverting dynamic in odds behavior, influenced by historical data. The positive correlation (0.6115) between the models' residuals underscores interconnected market movements.

6 References

7 Figures

8 Appendix

1. Packages used in this report:

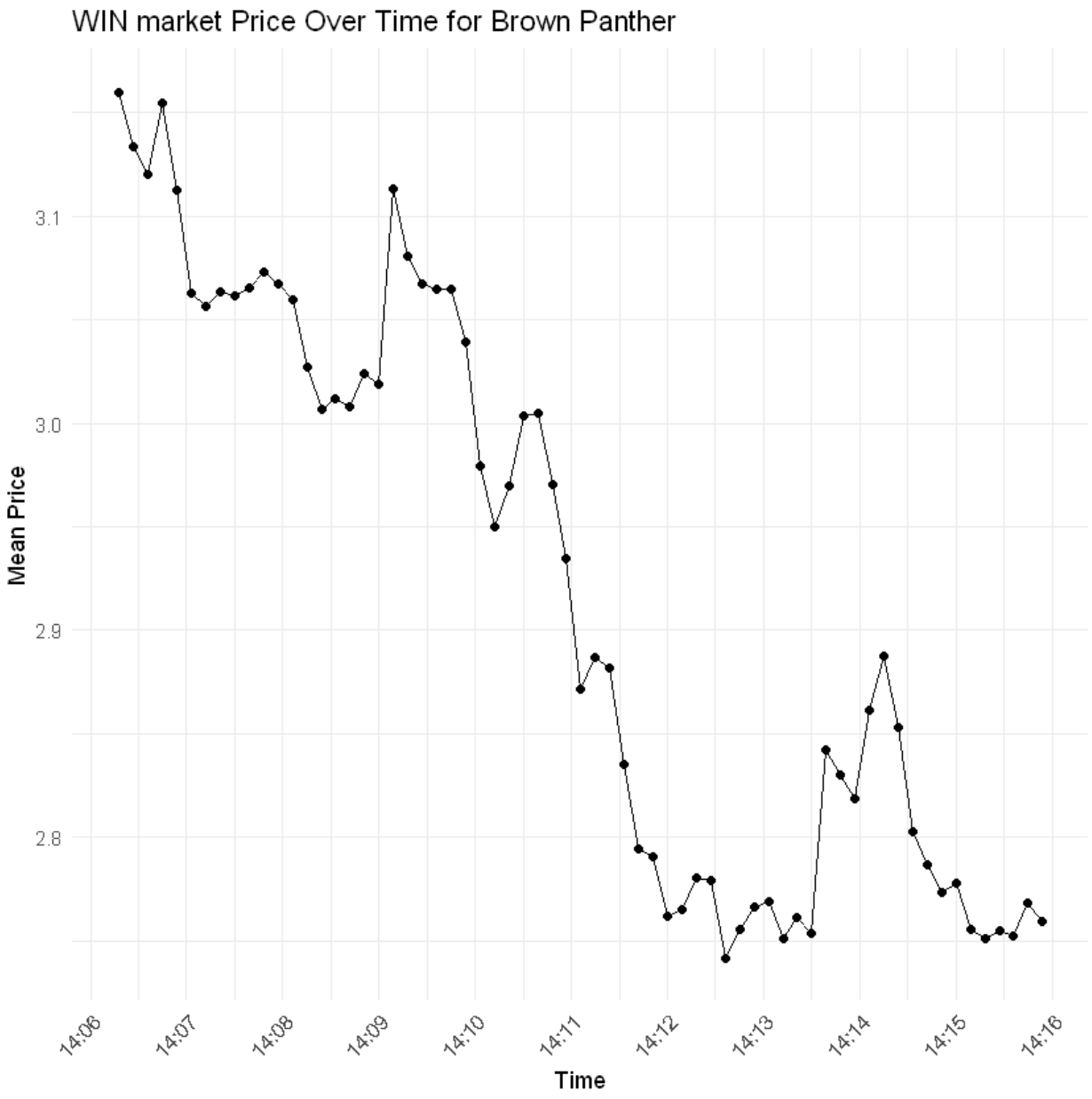

Figure 5: Image 0

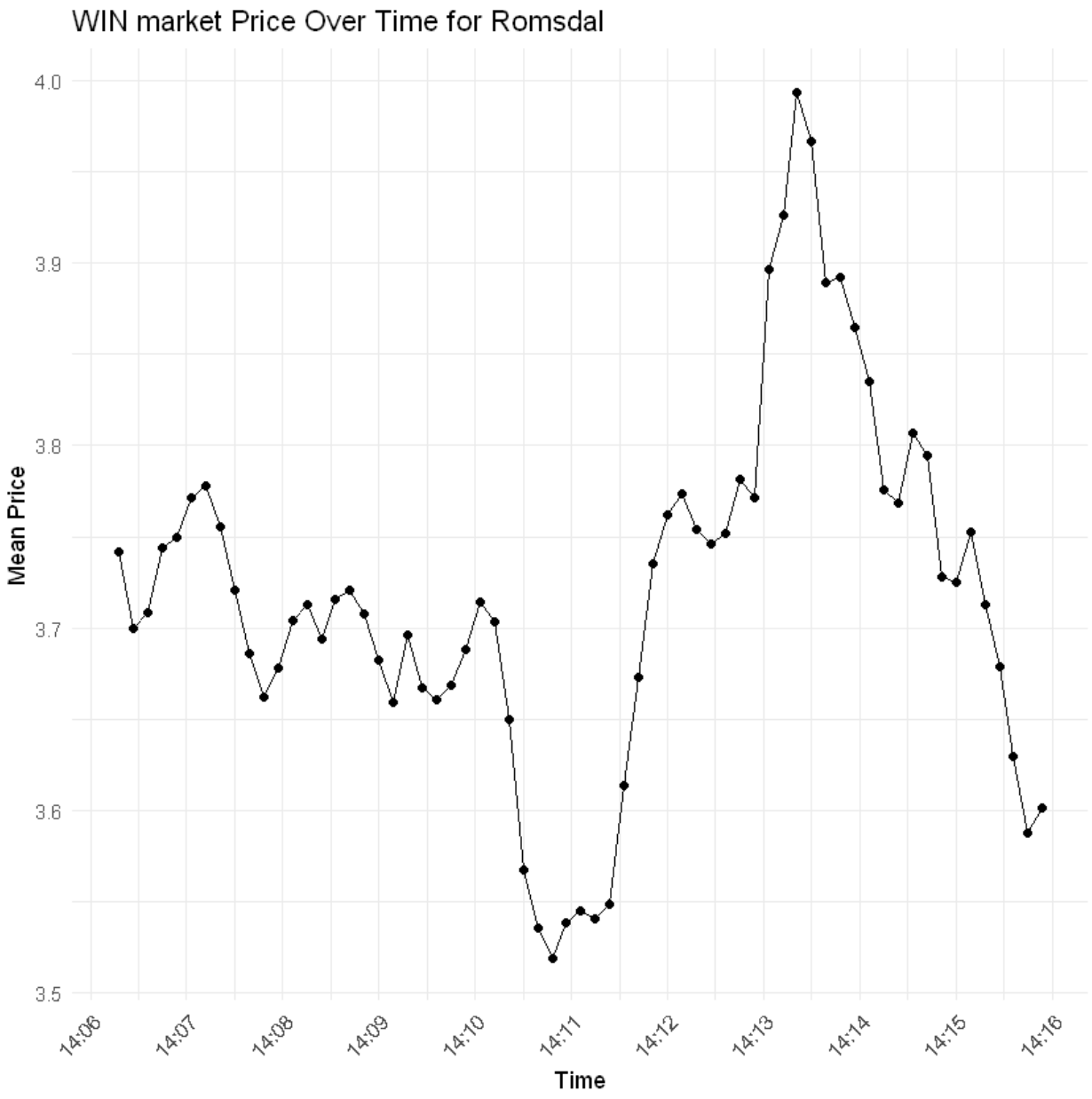

Figure 6: Image 1

Figure 7: Image 2

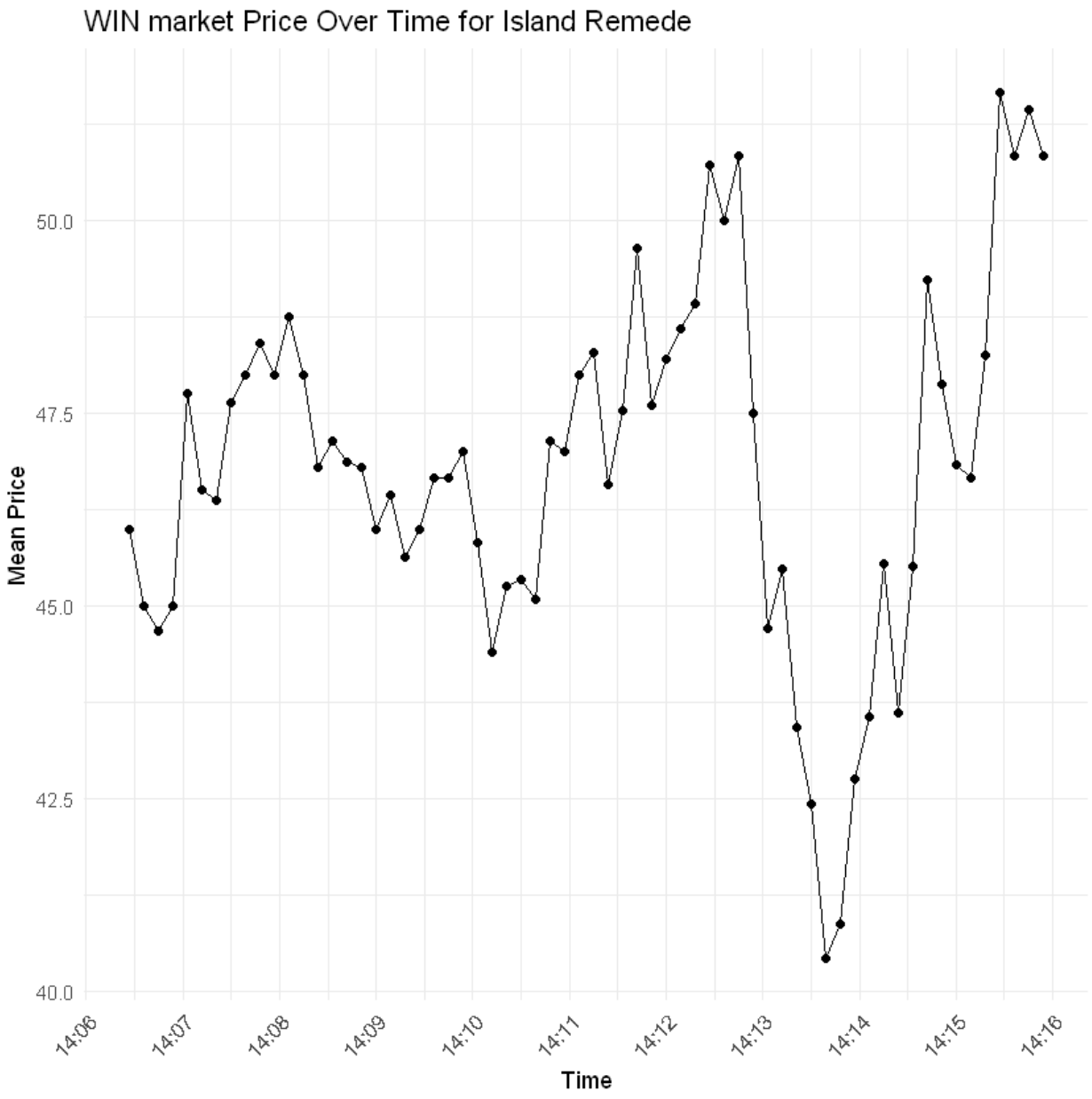

Figure 8: Image 3

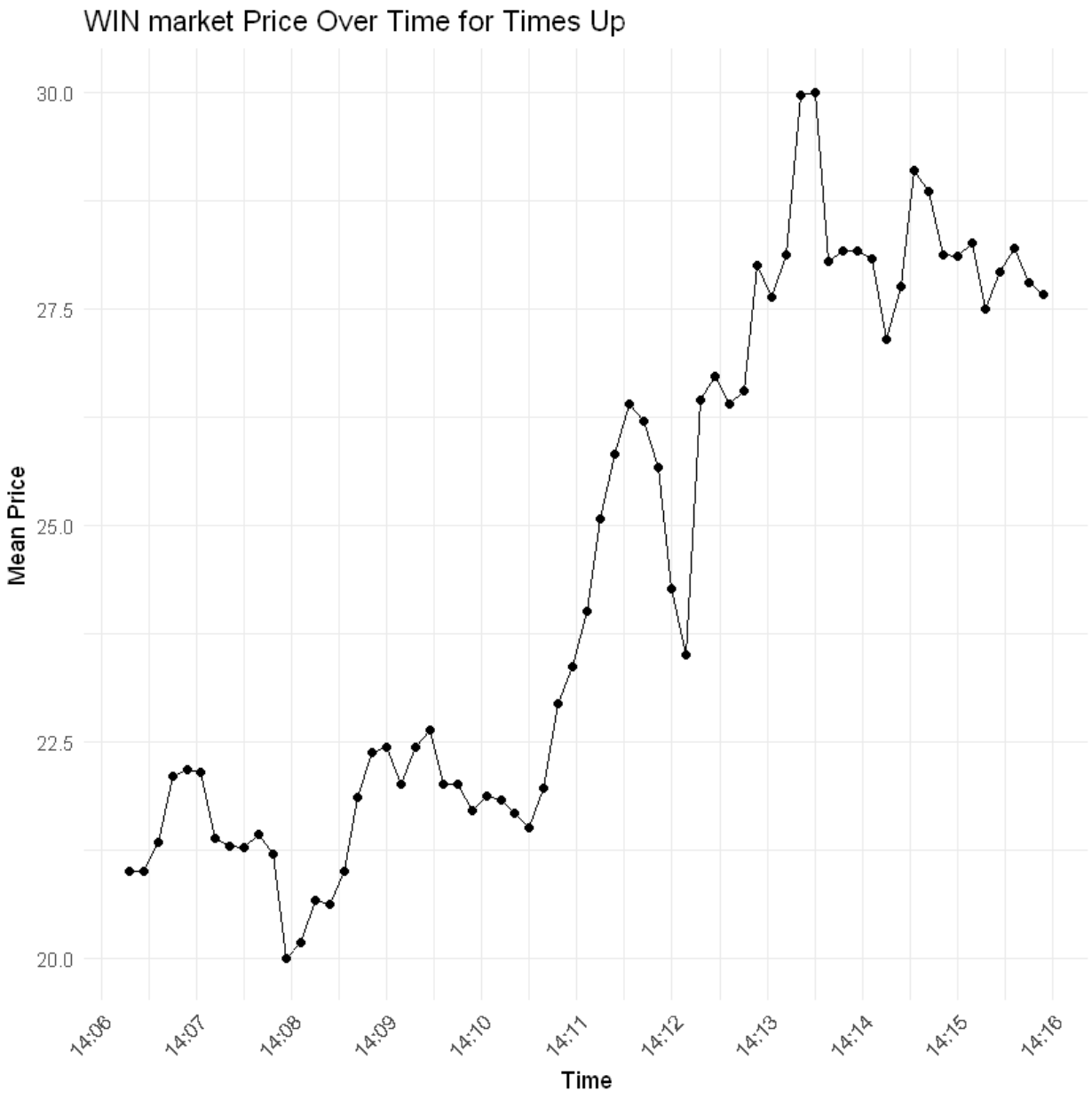

Figure 9: Image 4

Figure 10: Image 5

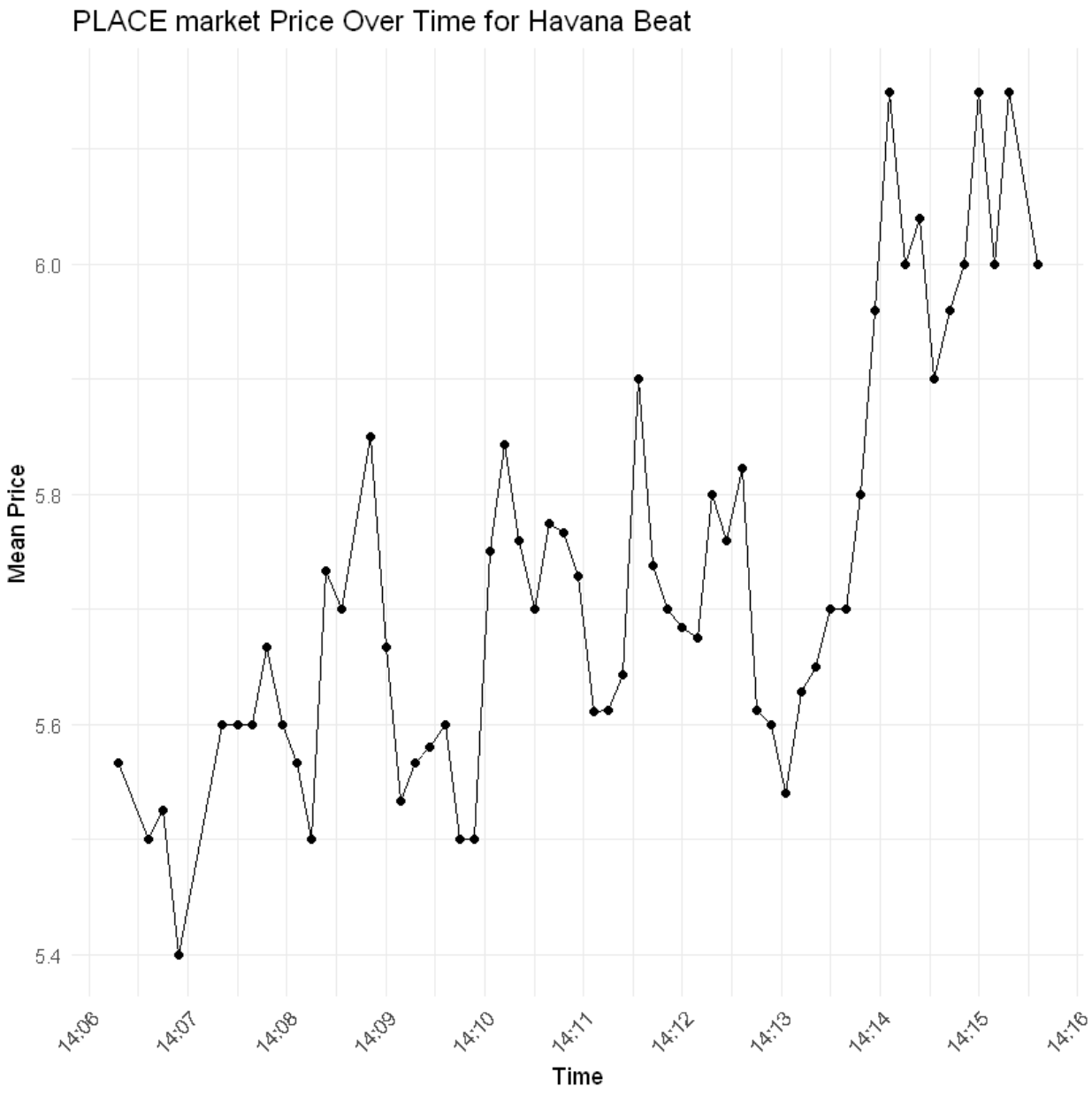

Figure 11: Image 6

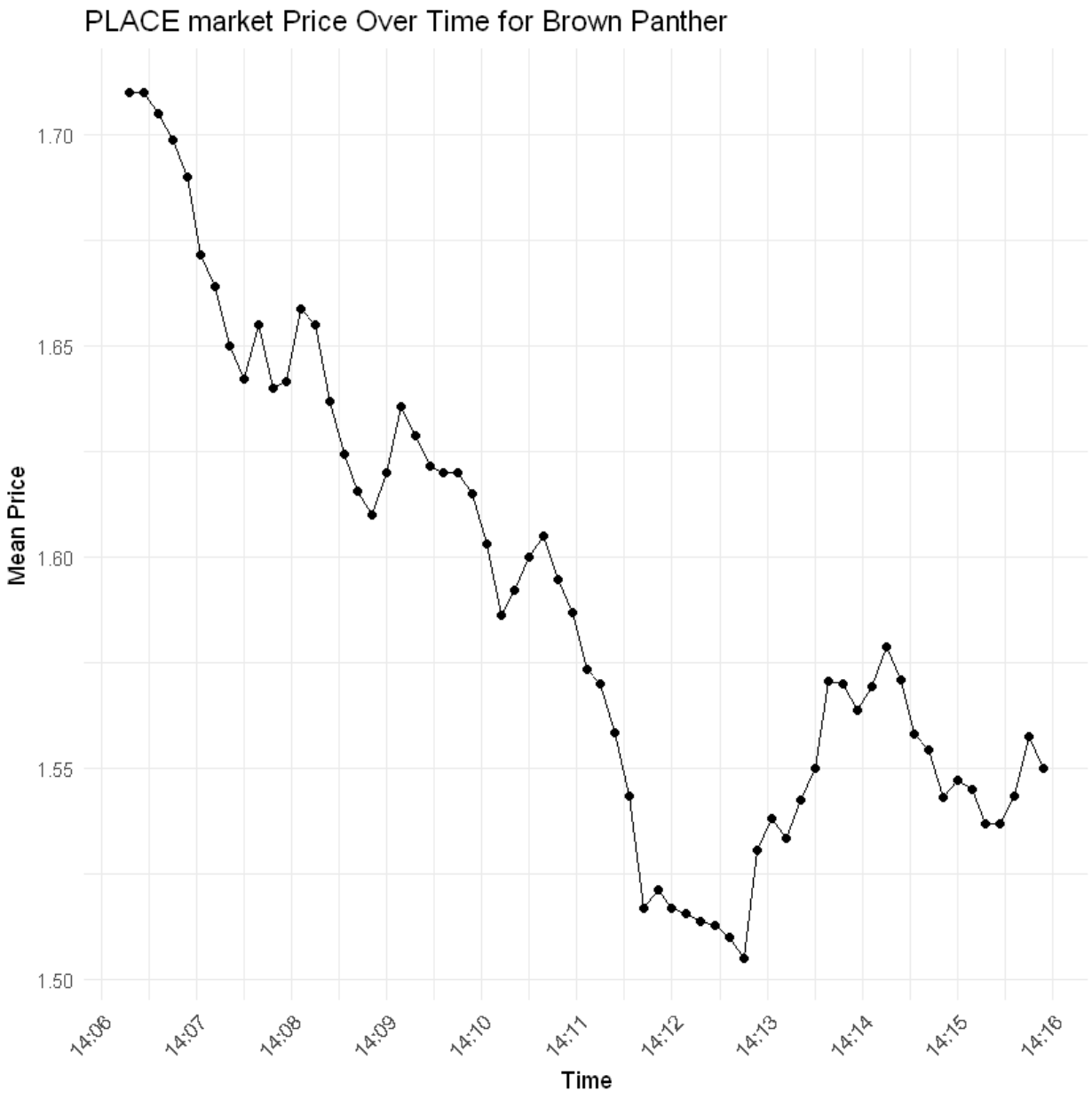

Figure 12: Image 7

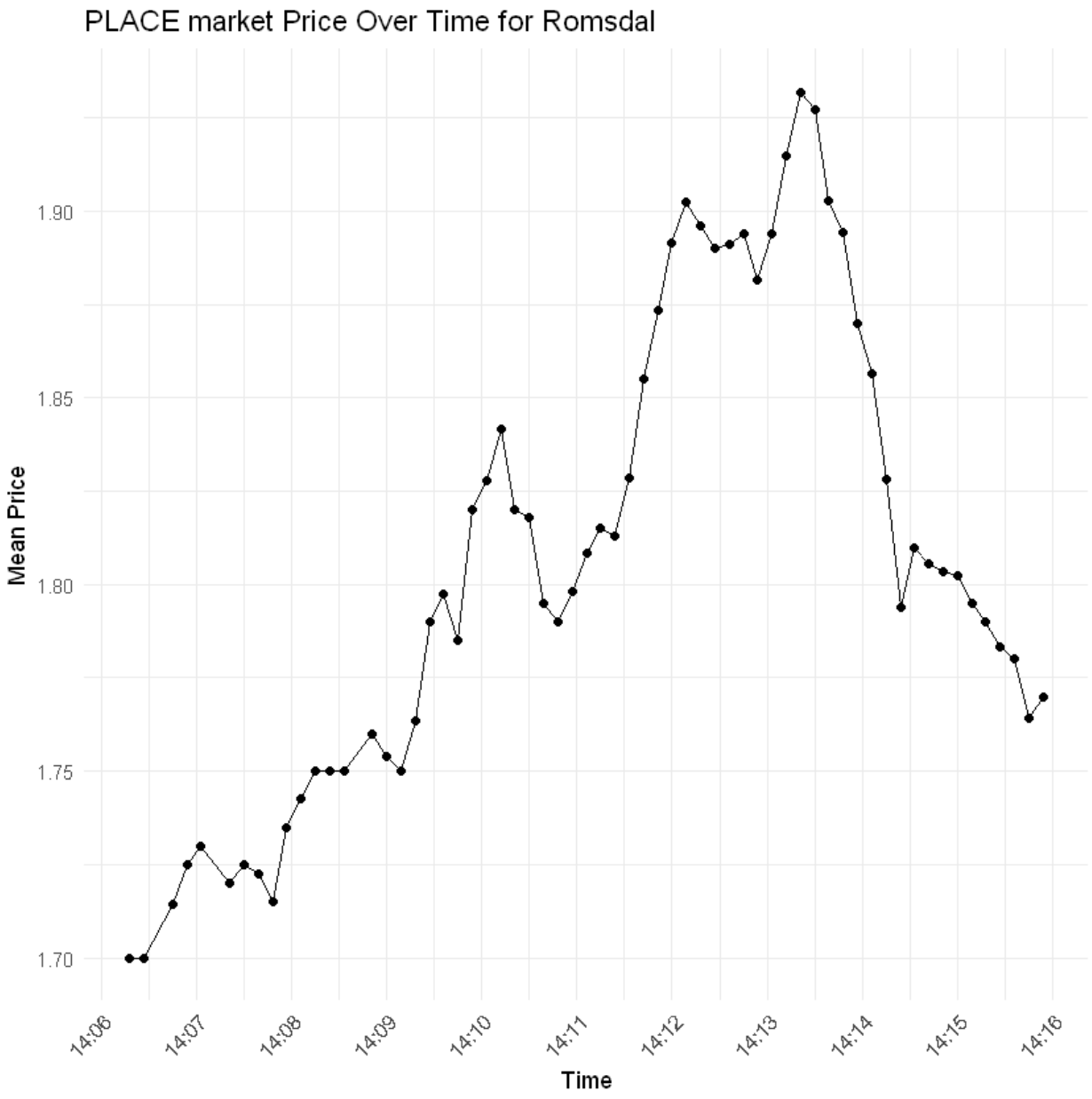

Figure 13: Image 8

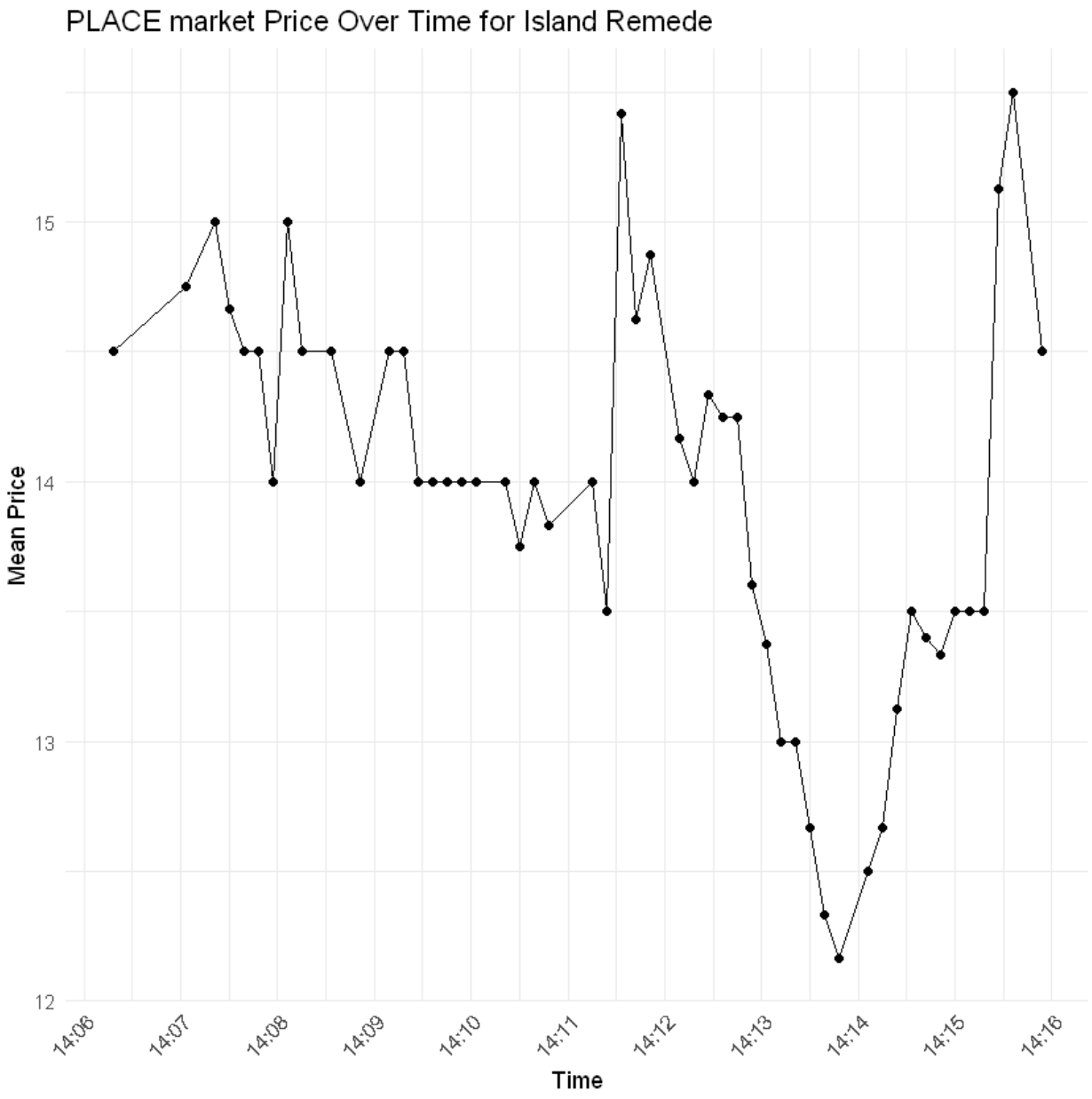

Figure 14: Image 9

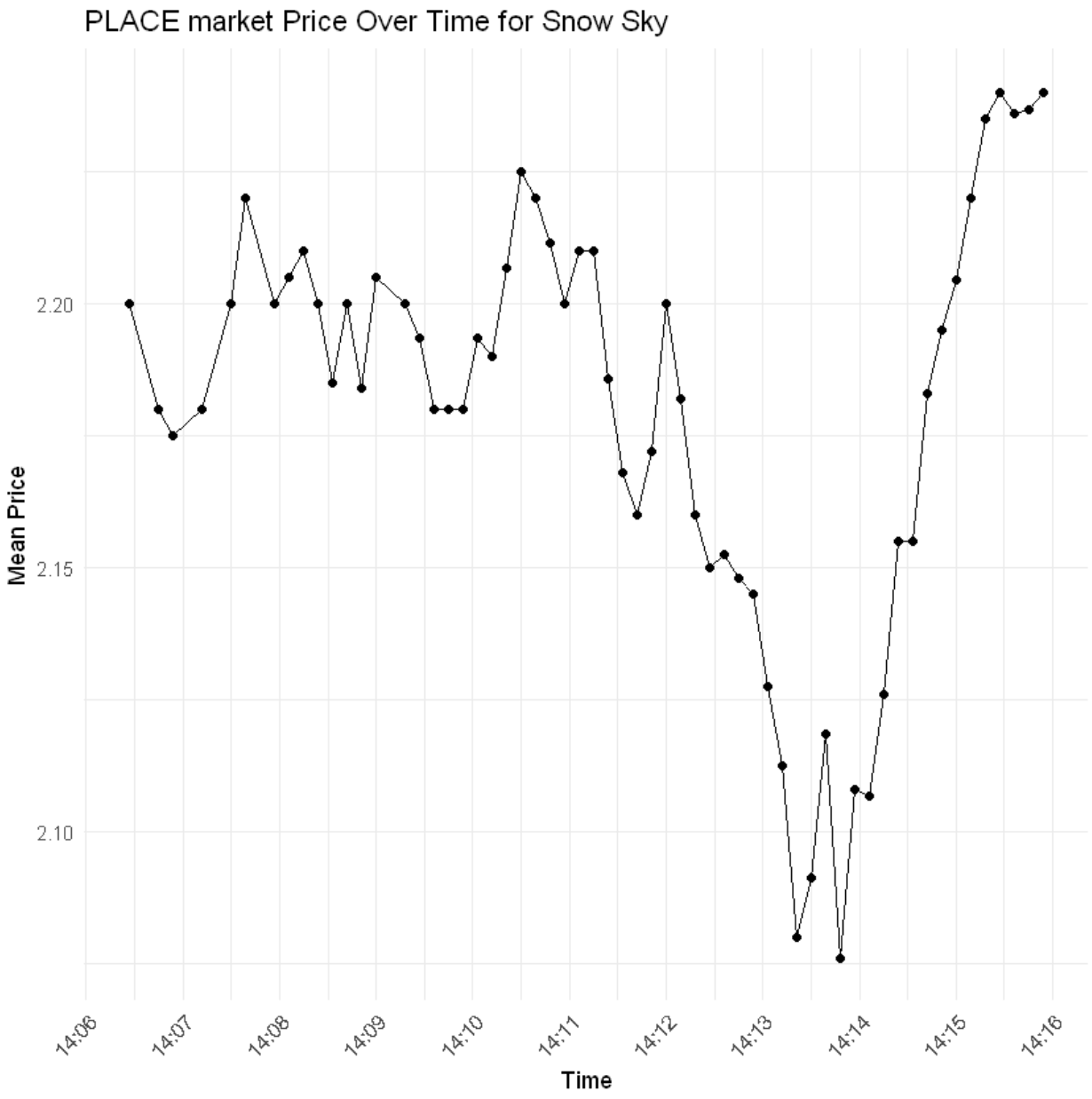

Figure 15: Image 10

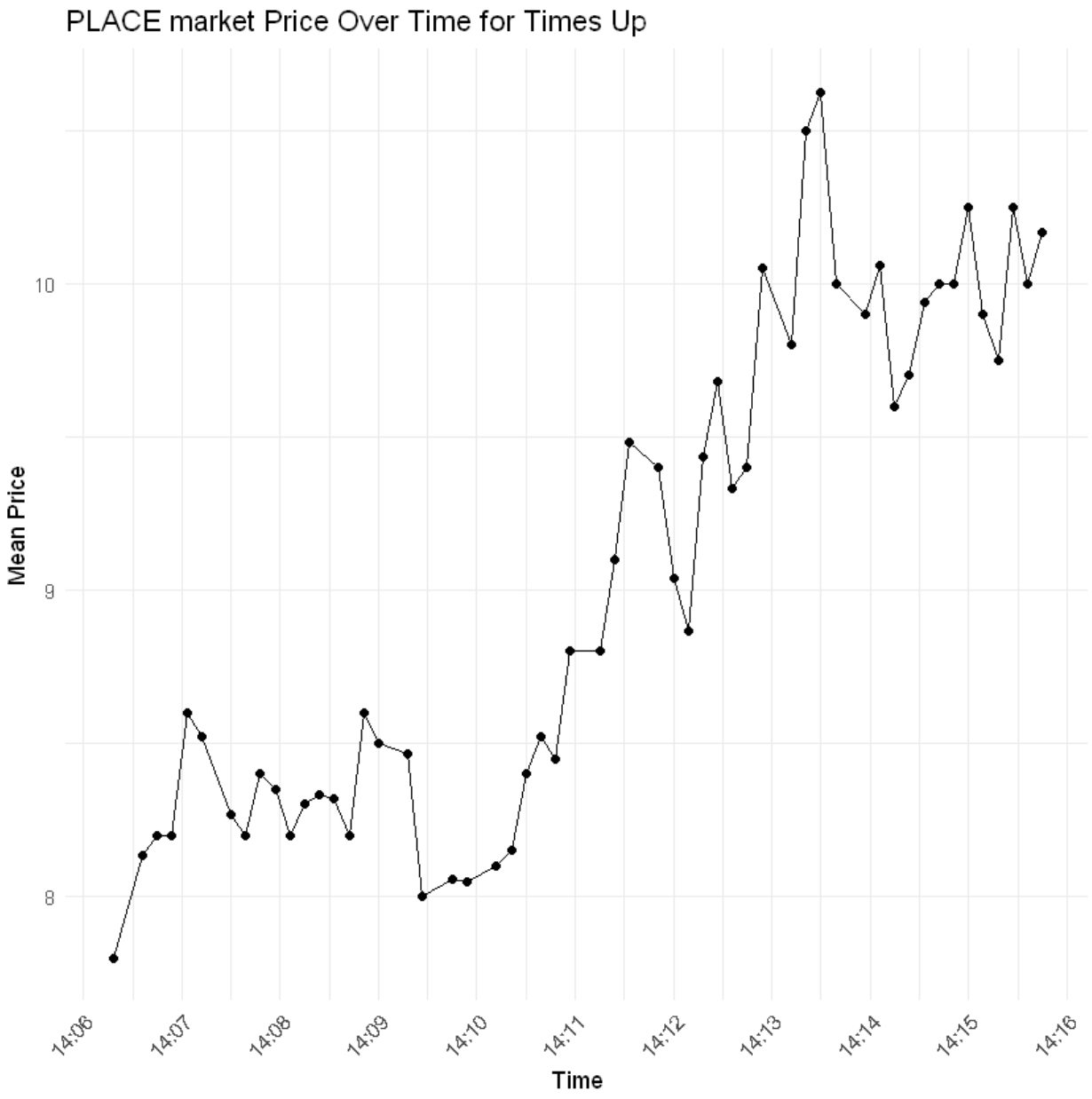

Figure 16: Image 11

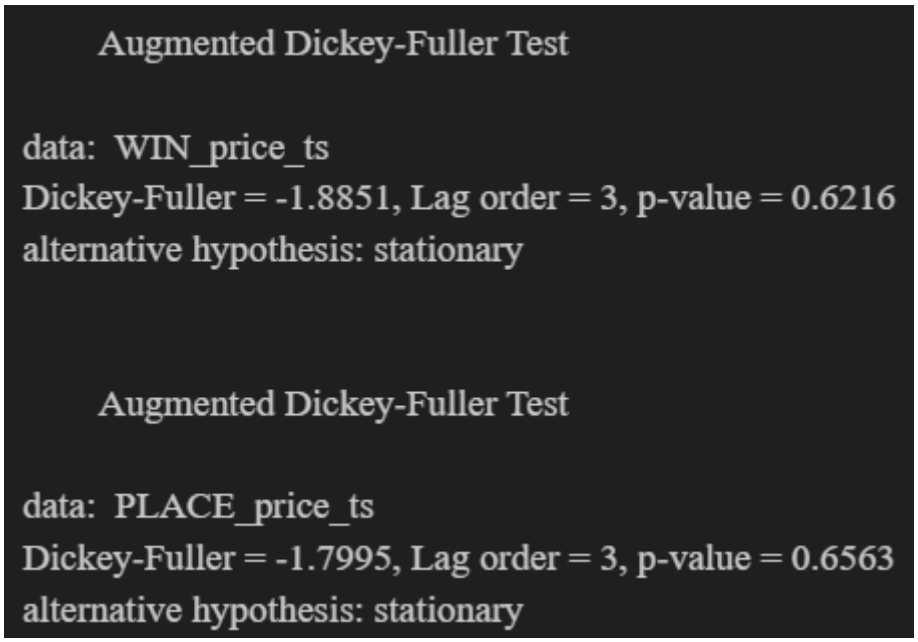

Figure 17: adf raw ts

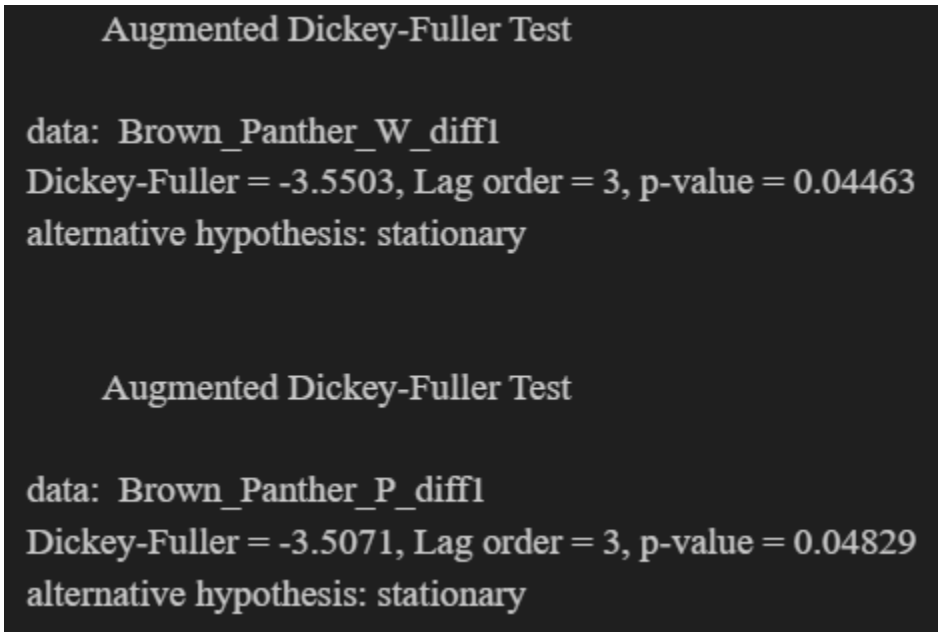

Figure 18: adf diff 1

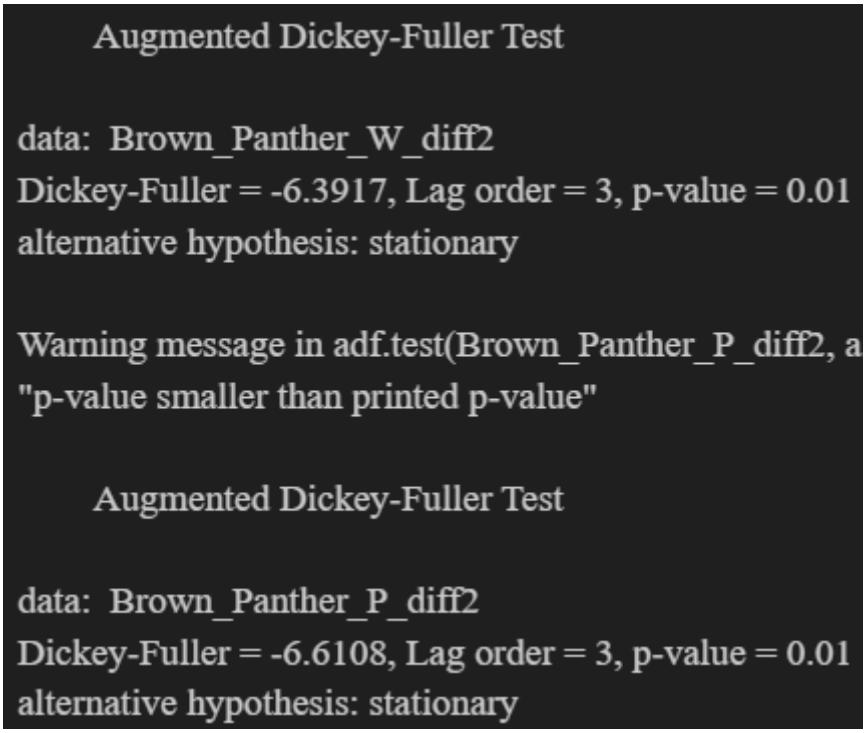

Figure 19: adf diff 2