

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

Ronald Schwalb and Samuel Devdas

May 24, 2024

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Literature Review</b>	<b>2</b>
<b>3</b>	<b>Methods</b>	<b>2</b>
3.1	Data Collection and Preprocessing . . . . .	3
3.2	Descriptive Analysis and Visualization . . . . .	3
3.3	Time Series Analysis and Stationarity . . . . .	4
<b>4</b>	<b>VAR-Model and Granger Causality</b>	<b>4</b>
4.1	VAR Model Estimation Results . . . . .	5
<b>5</b>	<b>Appendix</b>	<b>6</b>

## 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 in 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 a historic race that took place in York on 15 May 2015. The price data of all six participating horses were analysed, however only the detailed analysis of one specific horse, Brown Panther, is described in this report:

### 3.1 Data Collection and Preprocessing

The data source for this study was Betfair’s historical data, which is available through three pricing plans: Basic (free, 1-minute time frequency), Advanced (£69/month, 1-second time frequency), and Pro (£230/month, 50-millisecond time frequency). Pro plan data for May 2015 is available for free and was chosen for this study.

The selected race, held on May 15, 2015, met the criteria of having six runners, high liquidity (£2.43 million traded on the win market and £0.2 million on the place market), and no hurdles.

Each market is stored in a JSON file. The files for the selected race were retrieved and converted to a CSV file for analysis. The relevant fields included the UNIX timestamp, market identification, horse identification, last traded price, traded volume, an “in play” indicator, and a tuple containing the traded price and cumulative volume traded at that price. The study focused on analysing the actual trades over time, using fields like last traded price and volume only to verify and validate the data

### 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.

The data hygiene and distributions were thoroughly explored and investigated for the two markets for each of the six horses. It was determined that the most critical processing step at this stage was to obtain as many equally spaced intervals as possible, with each interval containing at least one data point, particularly for the less liquid place markets. Key decisions needed to be made regarding the triggers influencing this process, specifically the chosen horse, the selected time period used, and the bin size:

The chosen horse was Brown Panther, the horse with the most trades (£1.19 mil in total). The bin size and time period were more difficult to determine, as the closer the time gets to the start of the race, the more liquid the markets become, hence shorter or narrower bins can be used which implies more bins per time period, however shorter time periods. A trail and error approach was used to decide on 65 bins with a width of 9 seconds, hence a time period from 9.75 minutes before race start until race start.

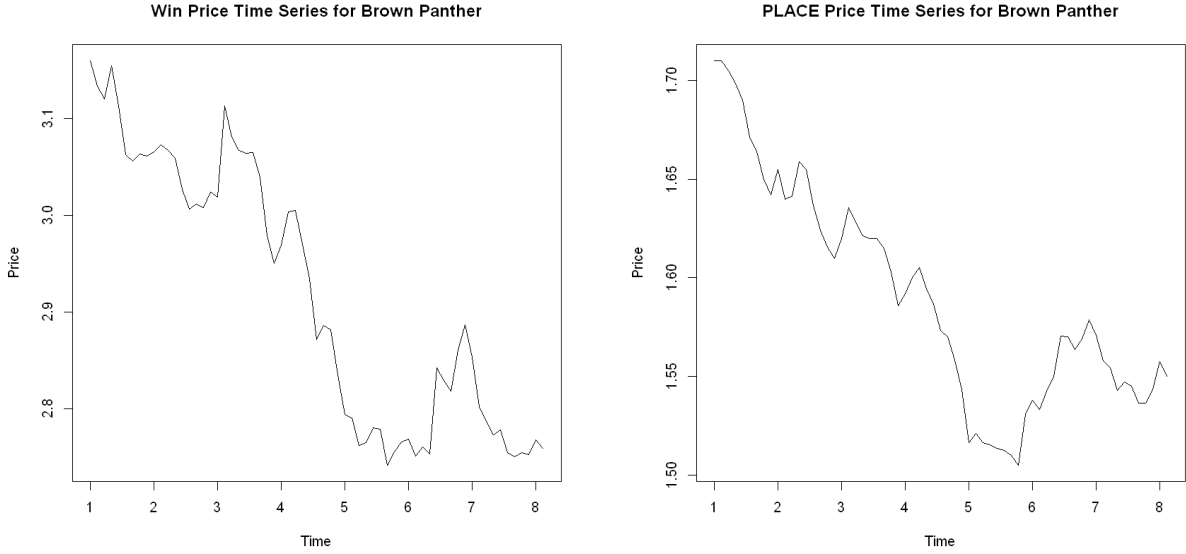


Figure 1: Win Price Time Series for Brown Panther  
Figure 2: Place Price Time Series for Brown Panther

### 3.3 Time Series Analysis and Stationarity

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.

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.

After the data were explored and transformed into a timeseries with regular intervals by applying equal width binning, the Augmented Dicky-Fuller test was applied to the two series. The hypothesis could not be rejected that the series are non-stationary, and hence differencing was applied.

## 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.

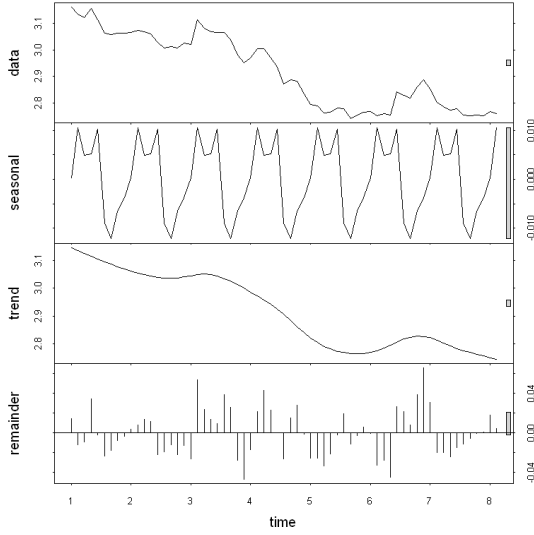


Figure 3: STL Win Price Brown Panther

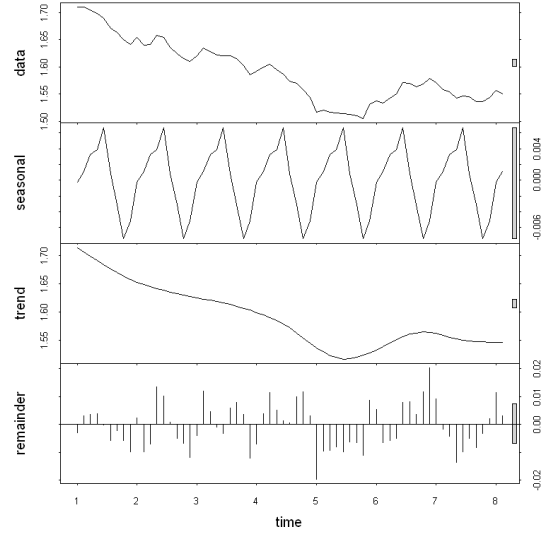


Figure 4: STL Place Price Brown Panther

## 4.1 VAR Model Estimation Results

A VAR model was applied to Brown Panther's differenced WIN and PLACE price data across 60 observations, considering up to 6 lags as suggested by the Akaike Information Criterion (AIC). The analysis revealed significant lagged interactions within and between the two markets, offering predictive insights into price movements.

### 4.1.1 Key Findings:

- **WIN Prices Dynamics:**

- Significant negative coefficients at lags 1 and 2 demonstrate a mean-reverting behavior in WIN prices.
- PLACE prices showed limited influence on WIN prices, indicating minimal cross-market effects.

- **PLACE Prices Dynamics:**

- Strong self-reversion in PLACE prices evident from significant negative coefficients at lags 1 and 2.
- Minimal influence from WIN prices, emphasizing market-specific dynamics.

### 4.1.2 Model Diagnostics and Statistical Significance:

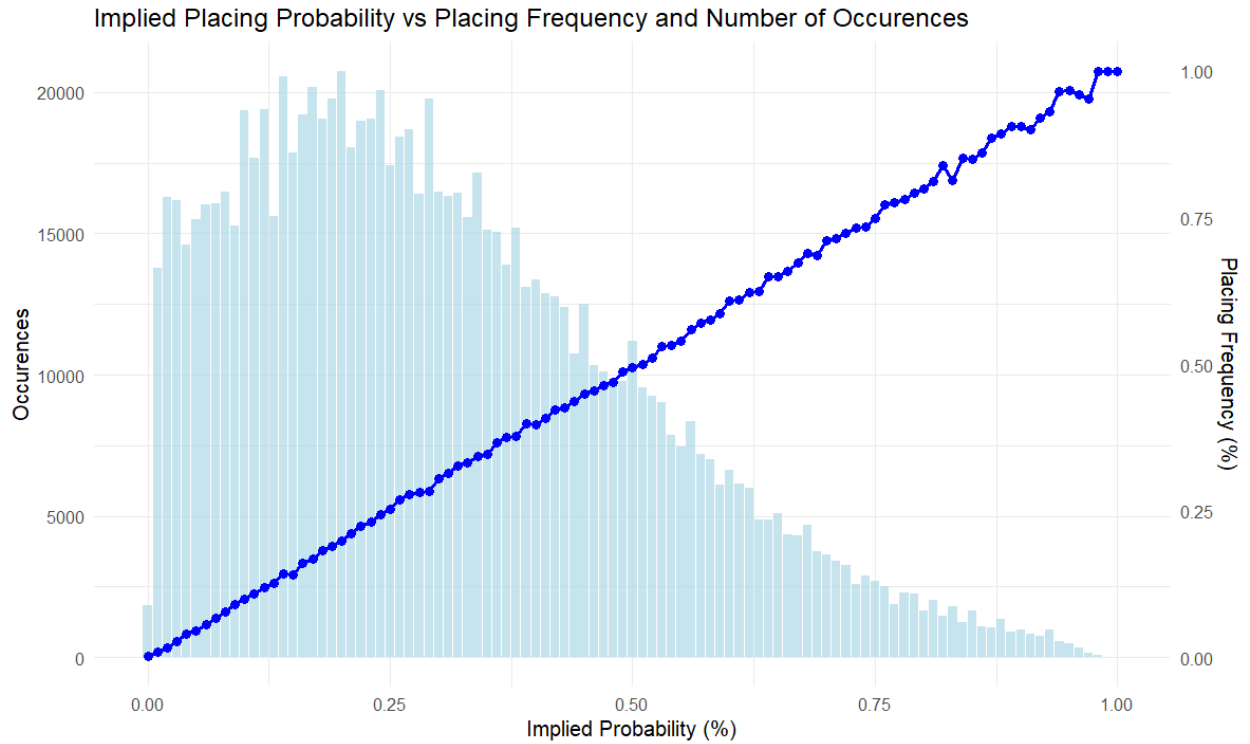
- **Model Stability:** Confirmed by roots of the characteristic polynomial all lying within the unit circle.

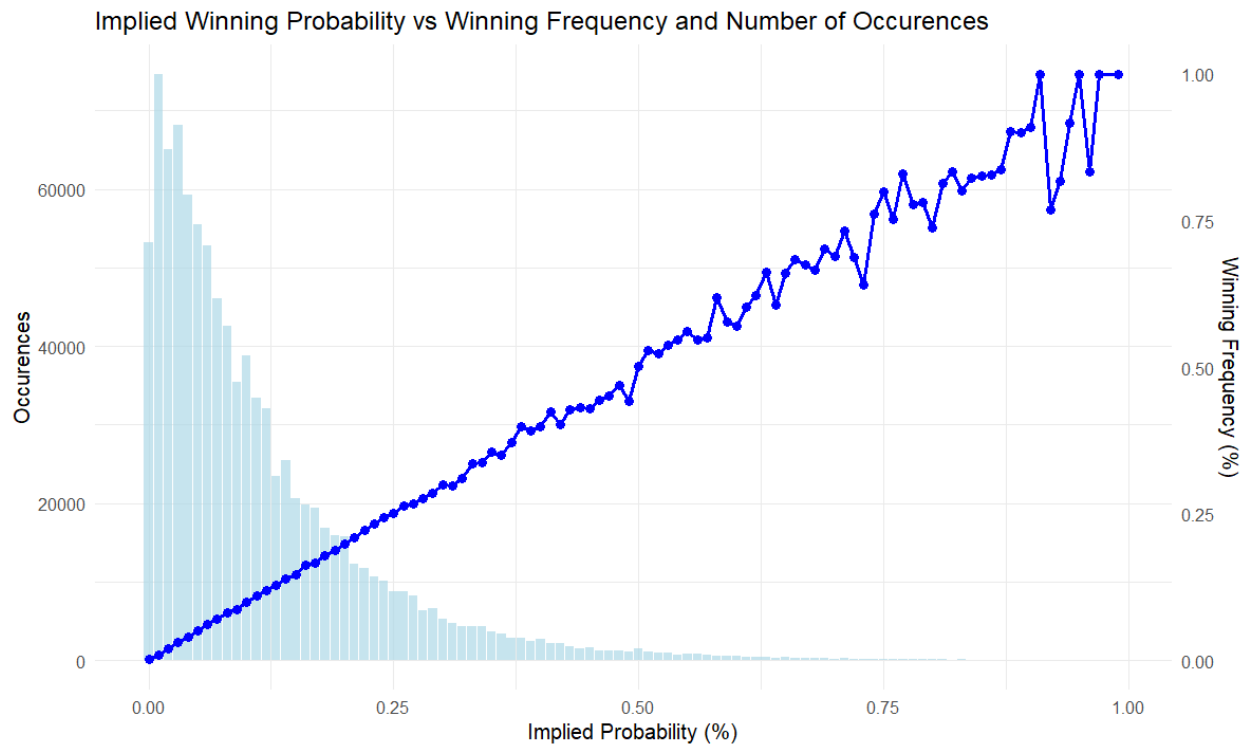
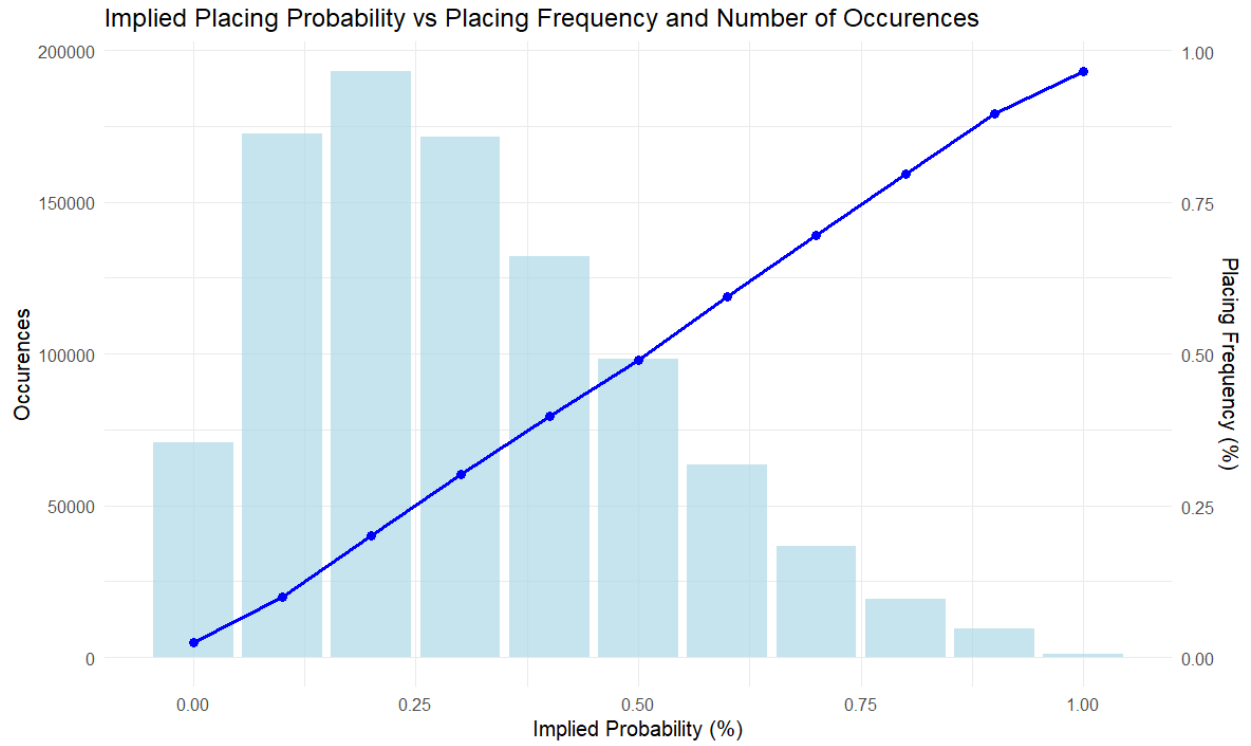
- **Residuals Analysis:** Low covariance and a moderate correlation of 0.6115 between residuals suggest a good model fit but highlight some unexplained interactions.
- **Adjusted R-squared:** Values of 0.281 for WIN and 0.2359 for PLACE indicate moderate explanatory power.
- **F-statistics:** Statistically significant across both equations, validating the model's predictive reliability.

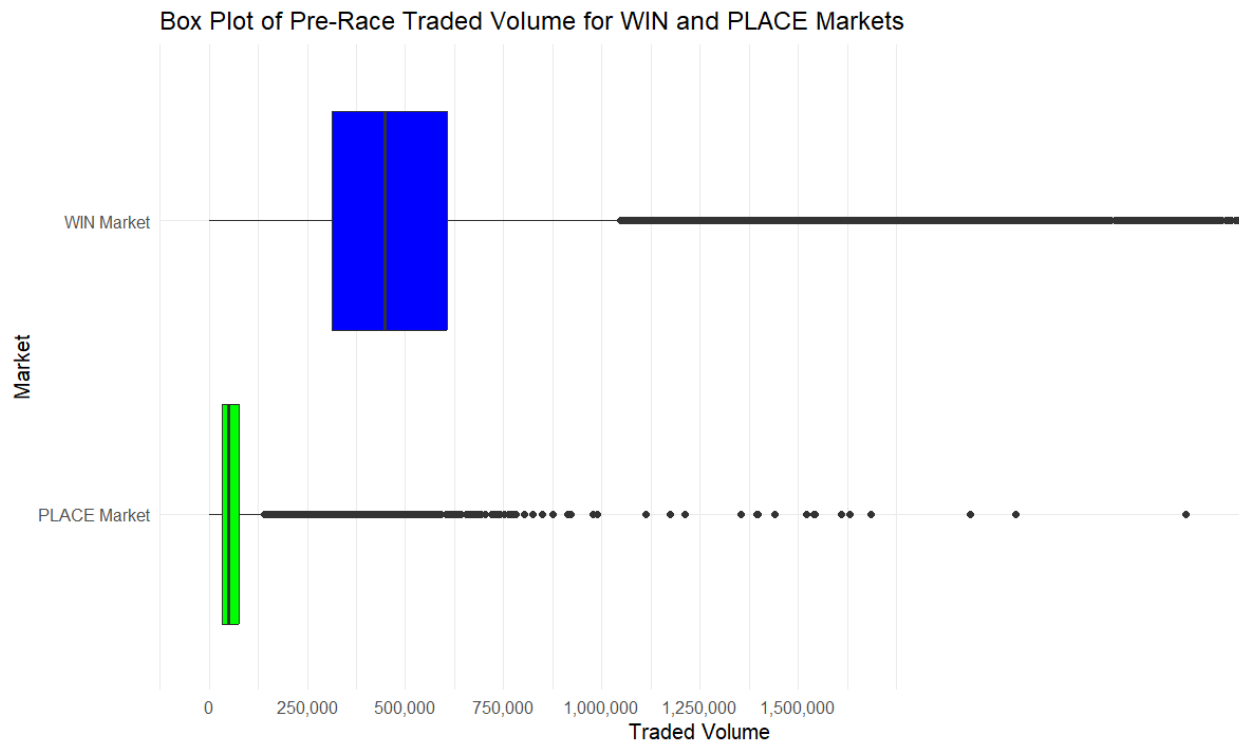
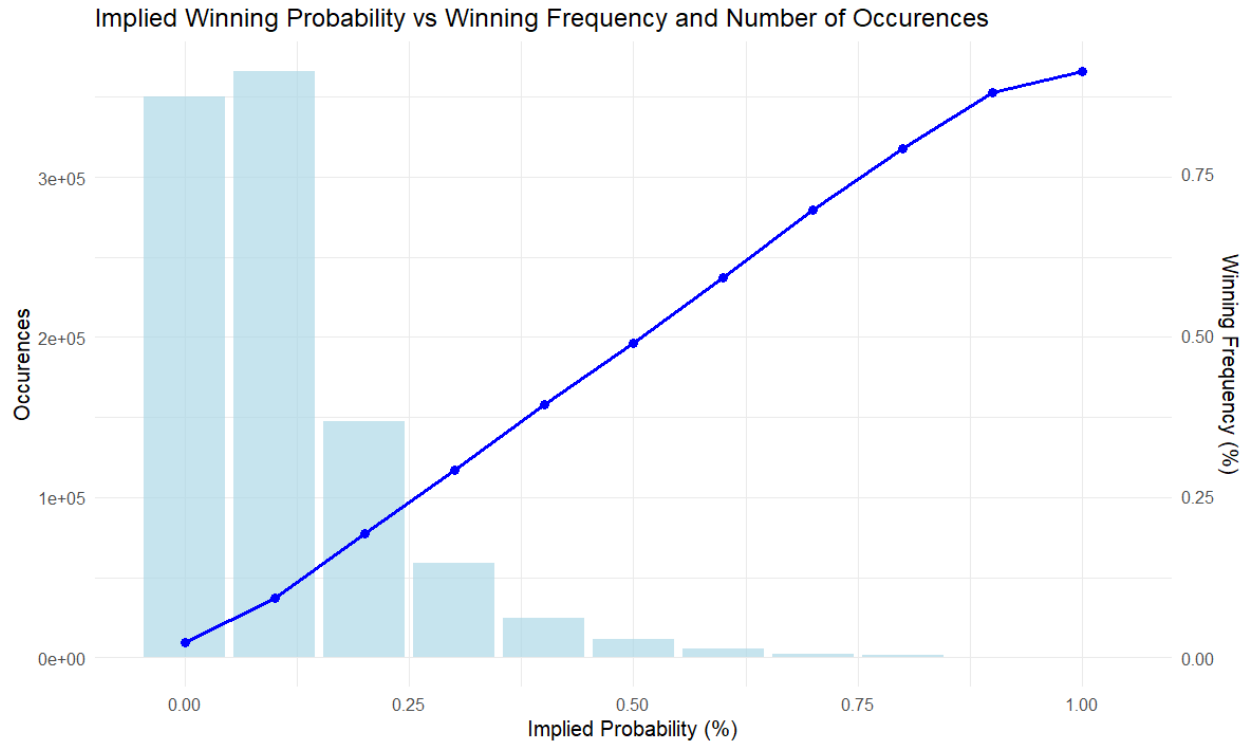
### 4.1.3 Interpretation

The VAR analysis underscores the mean-reverting nature of both WIN and PLACE markets, with past price changes serving as strong predictors for future prices. The observed dynamics are predominantly market-specific with limited inter-market dependencies. This understanding enhances strategic betting decisions, leveraging historical data to predict future odds movements effectively. 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 is rejected based on this analysis. The win-price and place-price for Brown Panther has a correlation of 0.95 however as mentioned very little causality. This is yet another example that demonstrates that correlation does not imply causation.

## 5 Appendix







Packages used in this report:



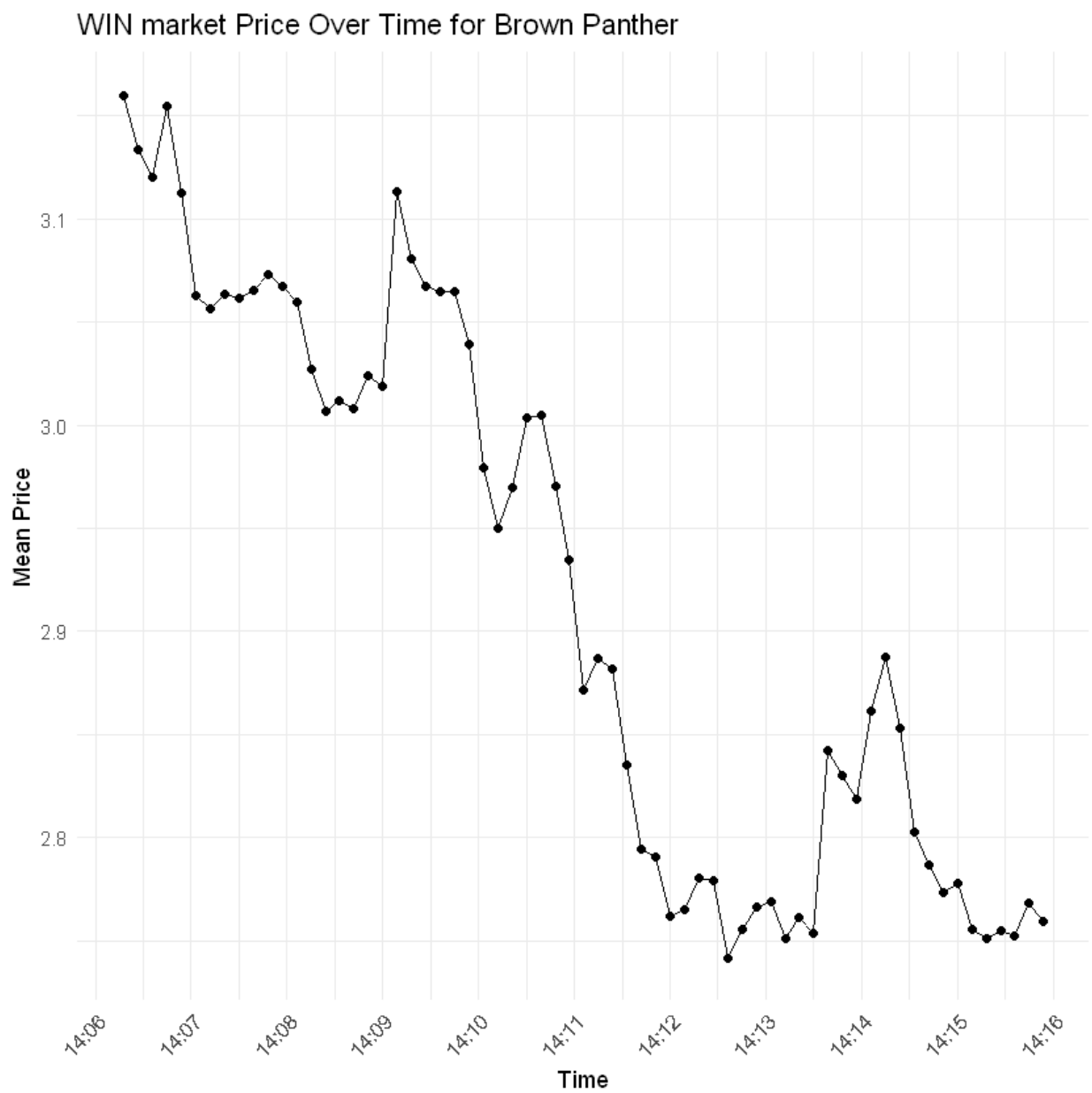


Figure 5: Image 0

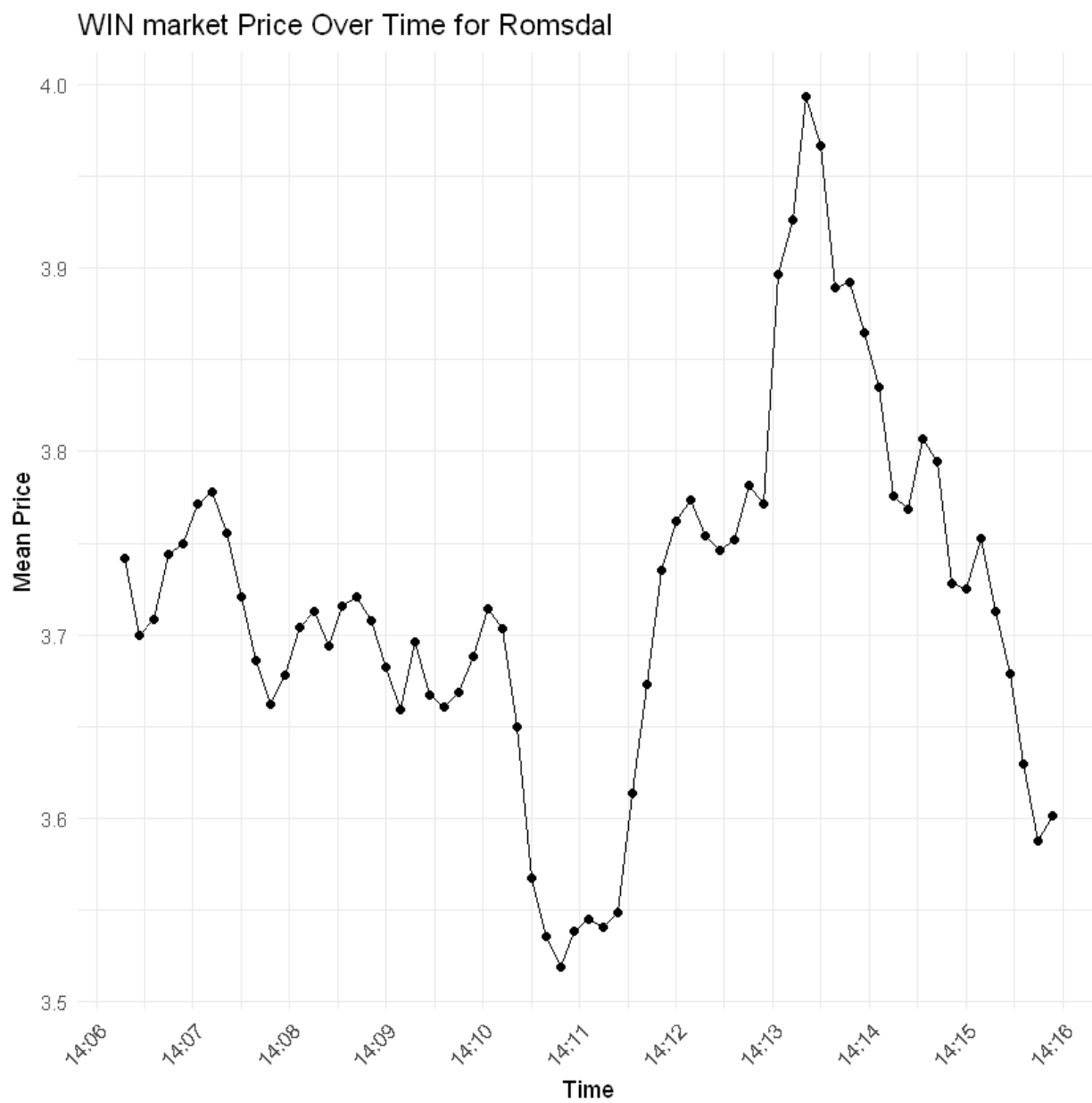


Figure 6: Image 1



Figure 7: Image 2

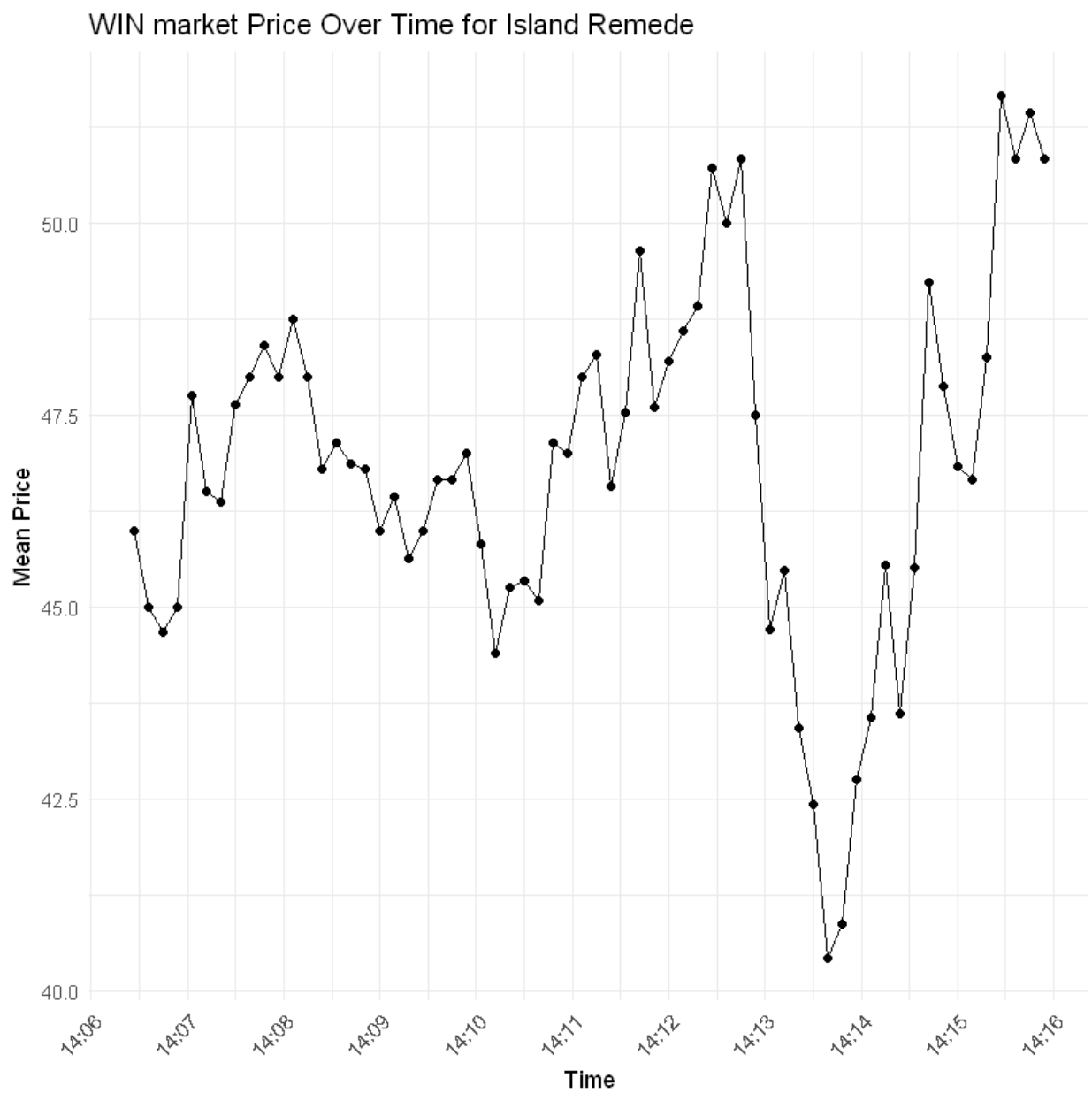


Figure 8: Image 3

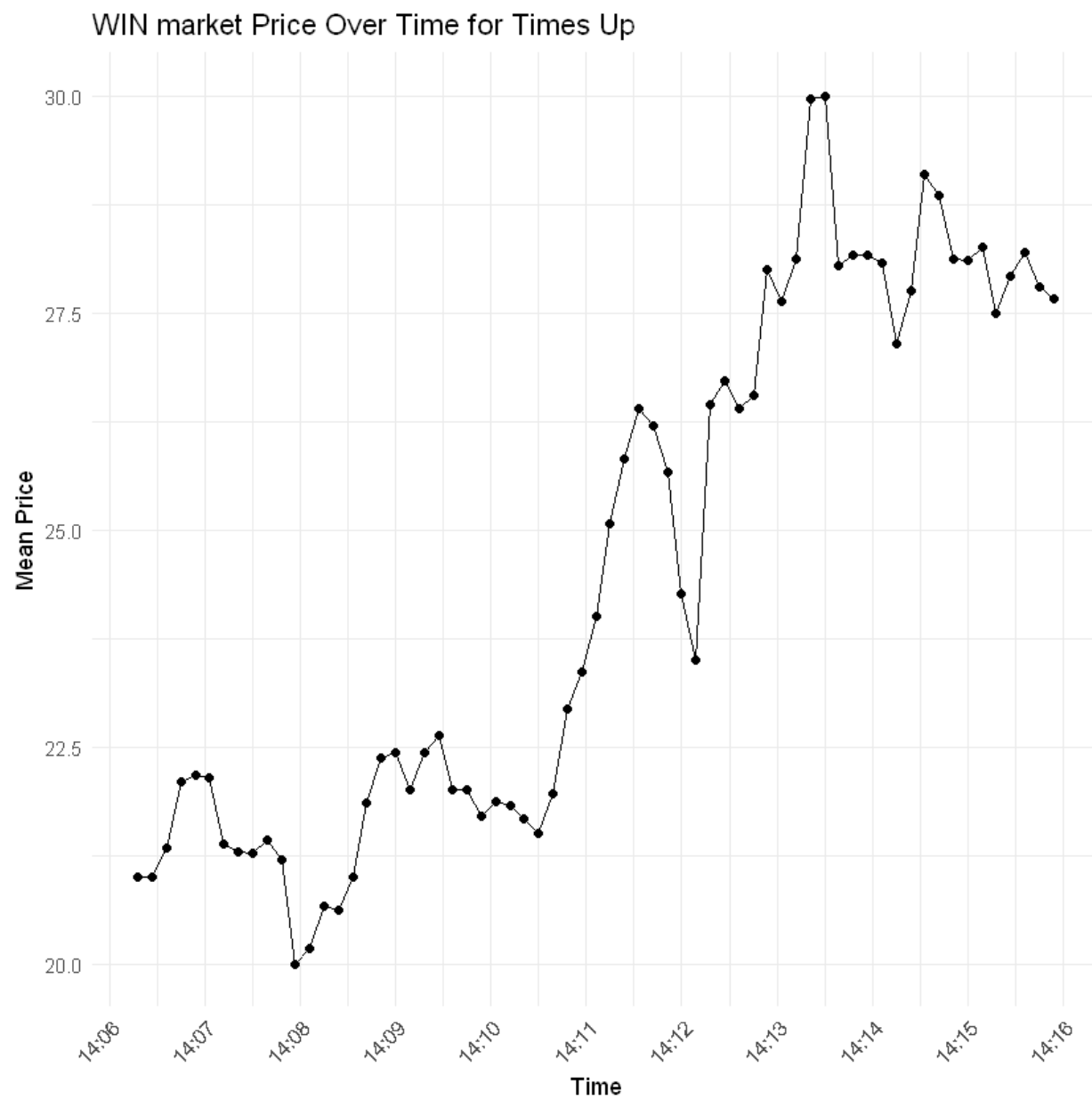


Figure 9: Image 4

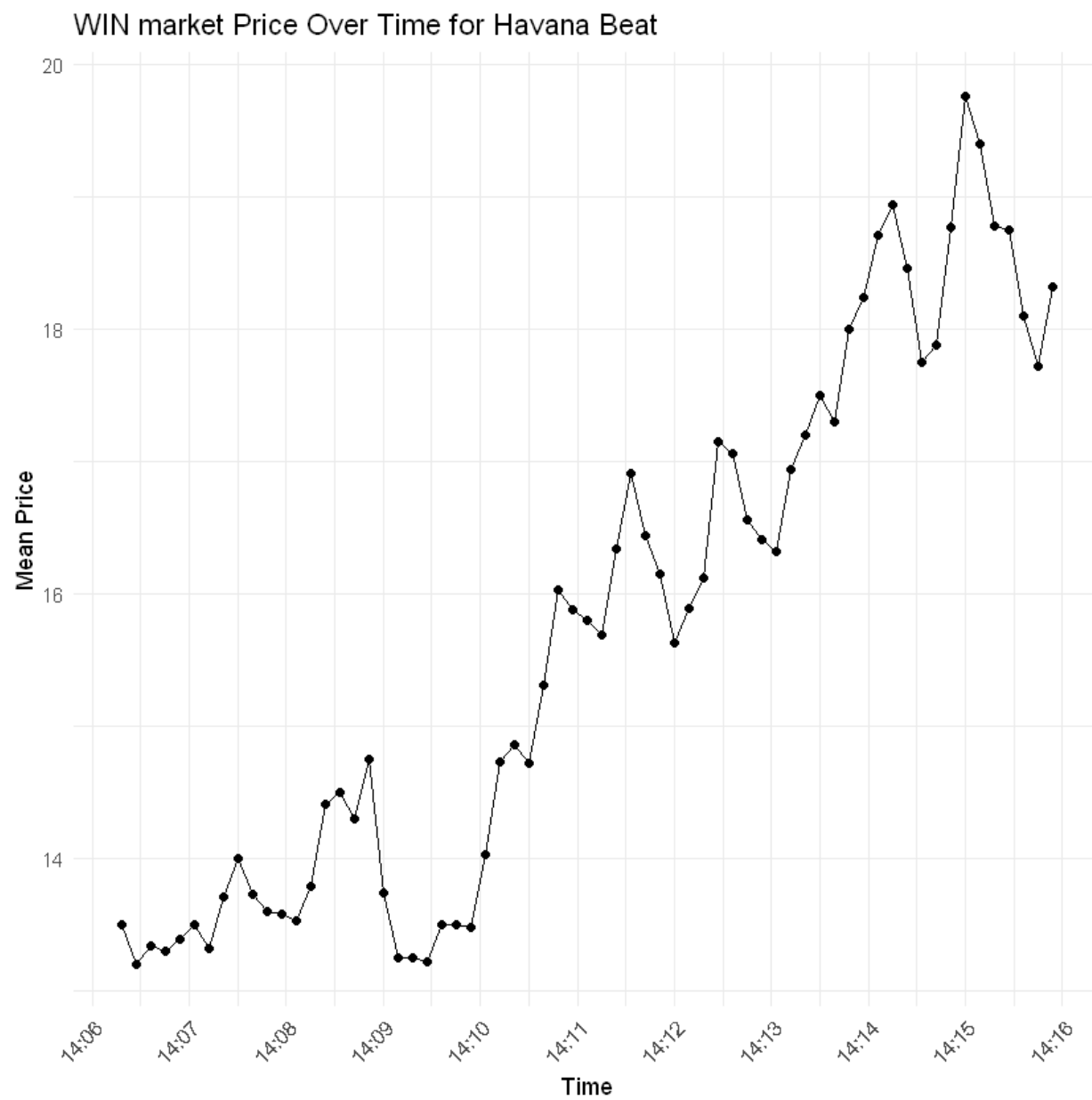


Figure 10: Image 5

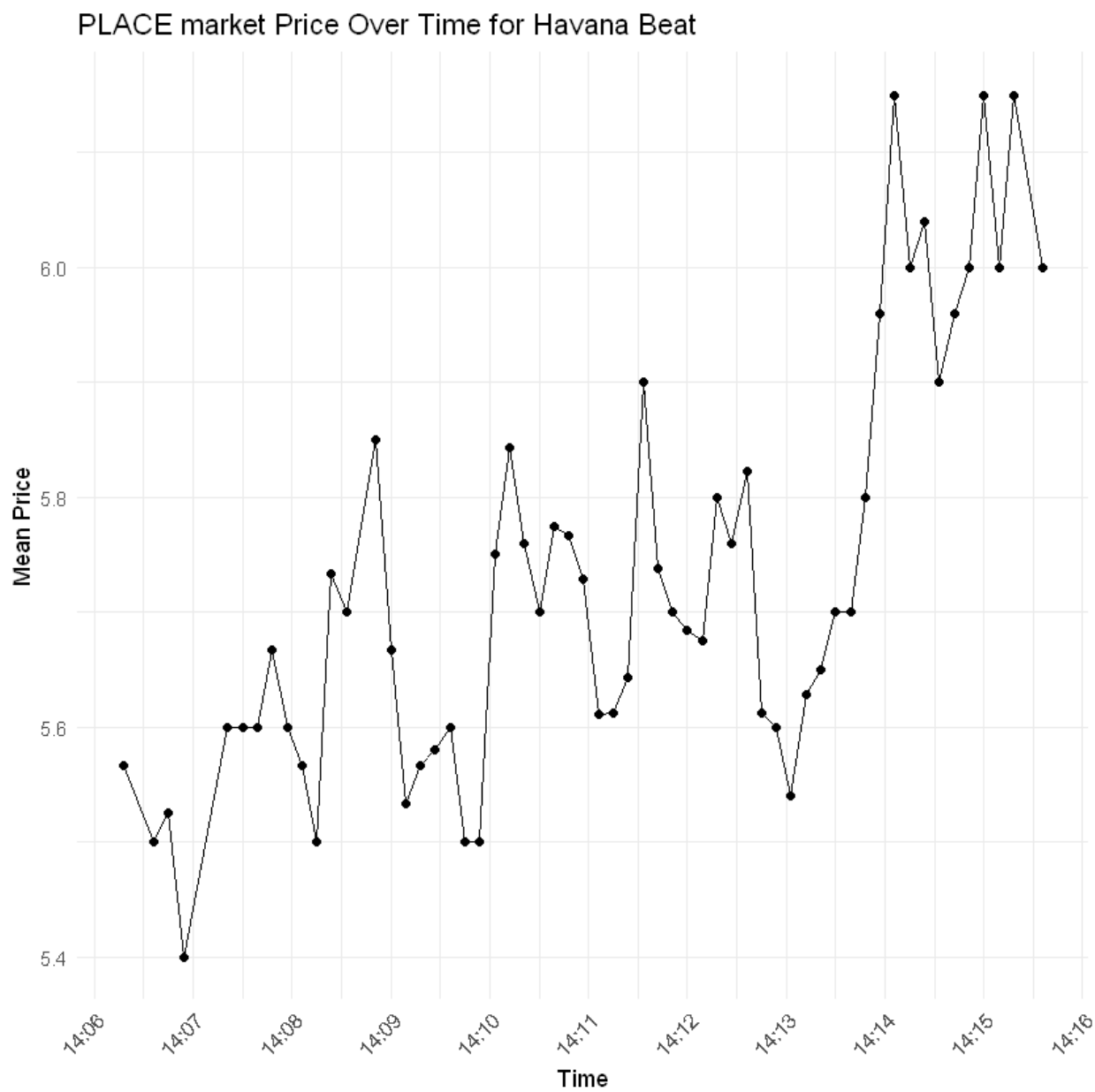


Figure 11: Image 6

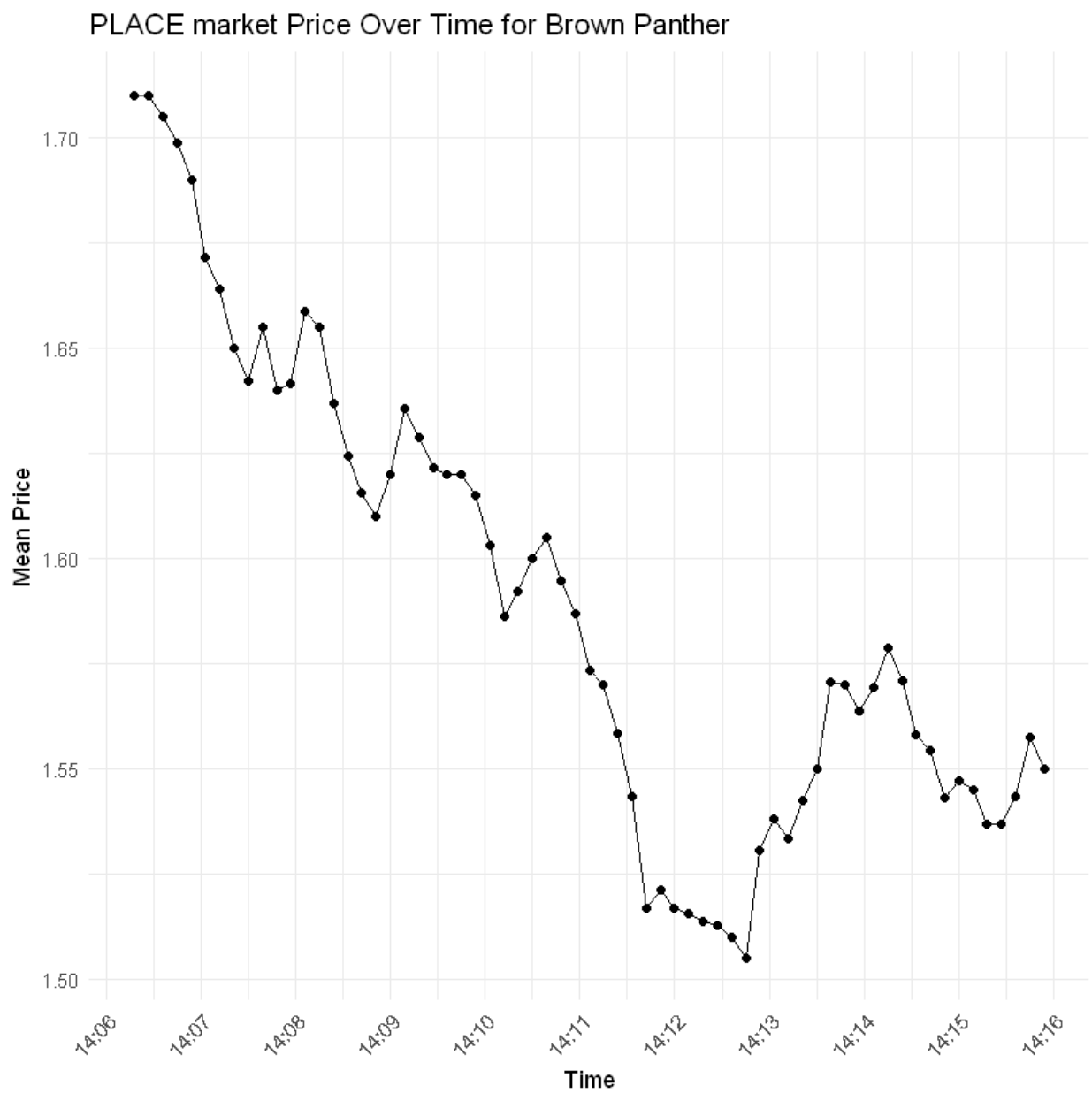


Figure 12: Image 7



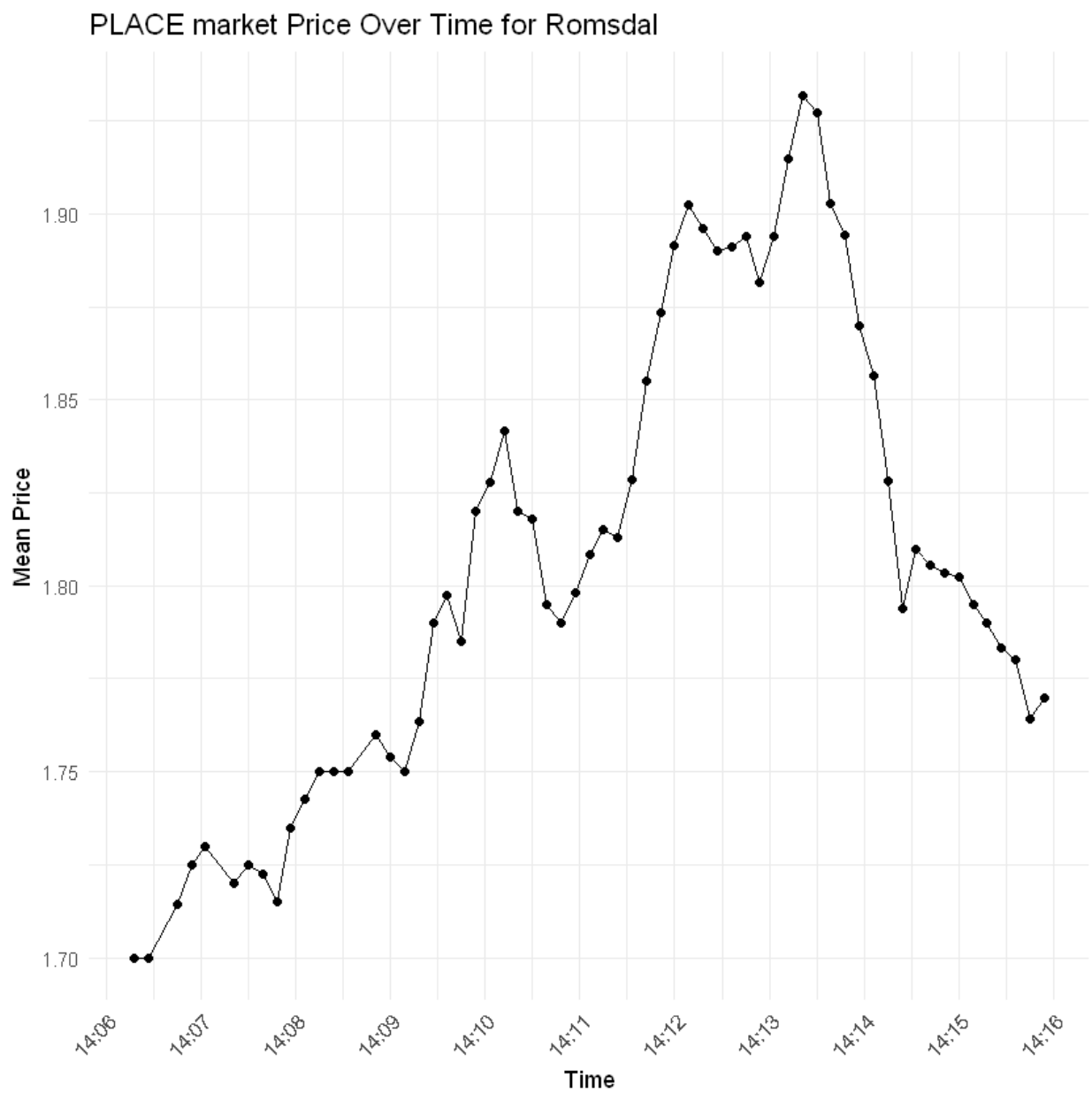


Figure 13: Image 8

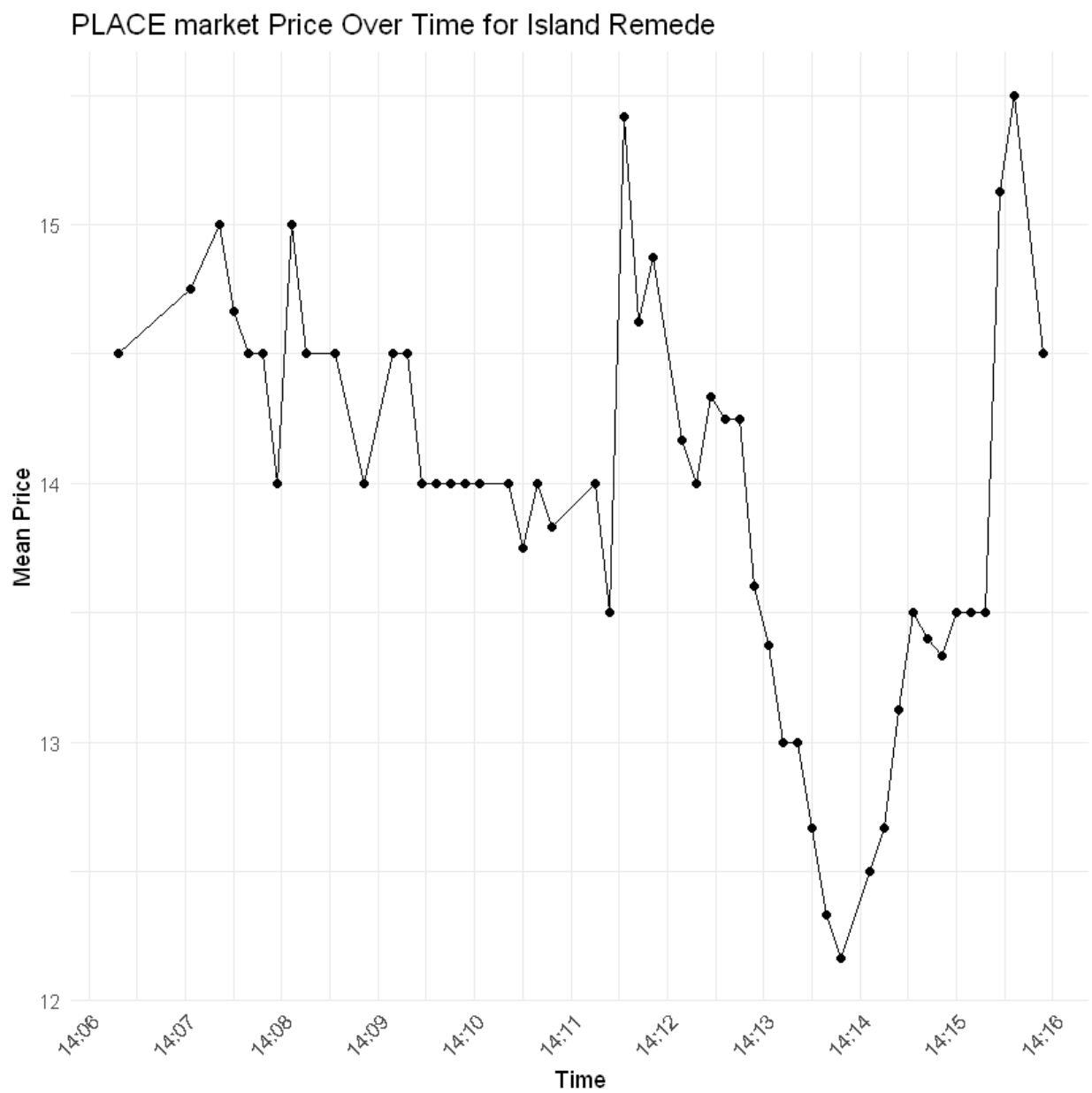


Figure 14: Image 9

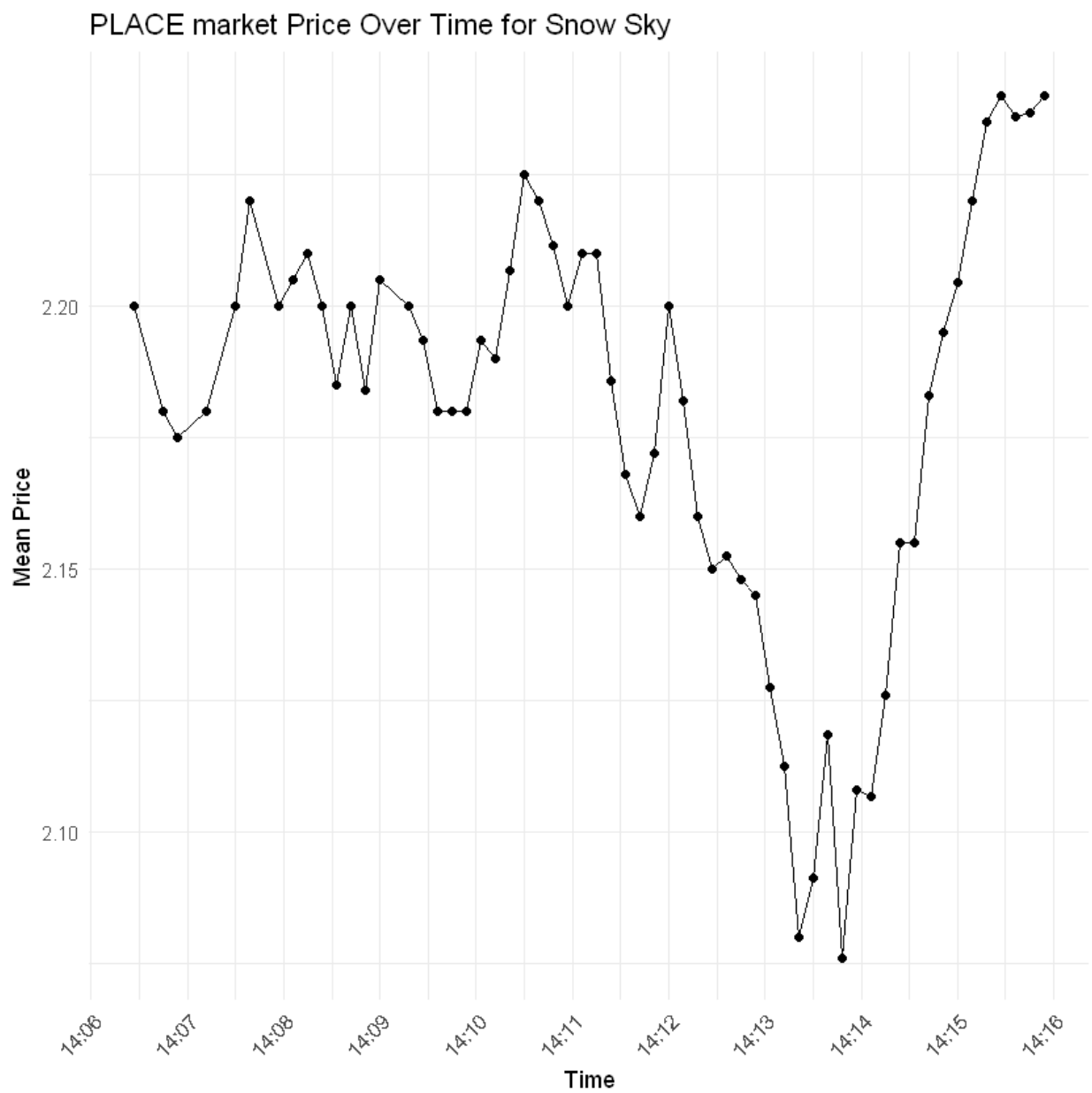


Figure 15: Image 10

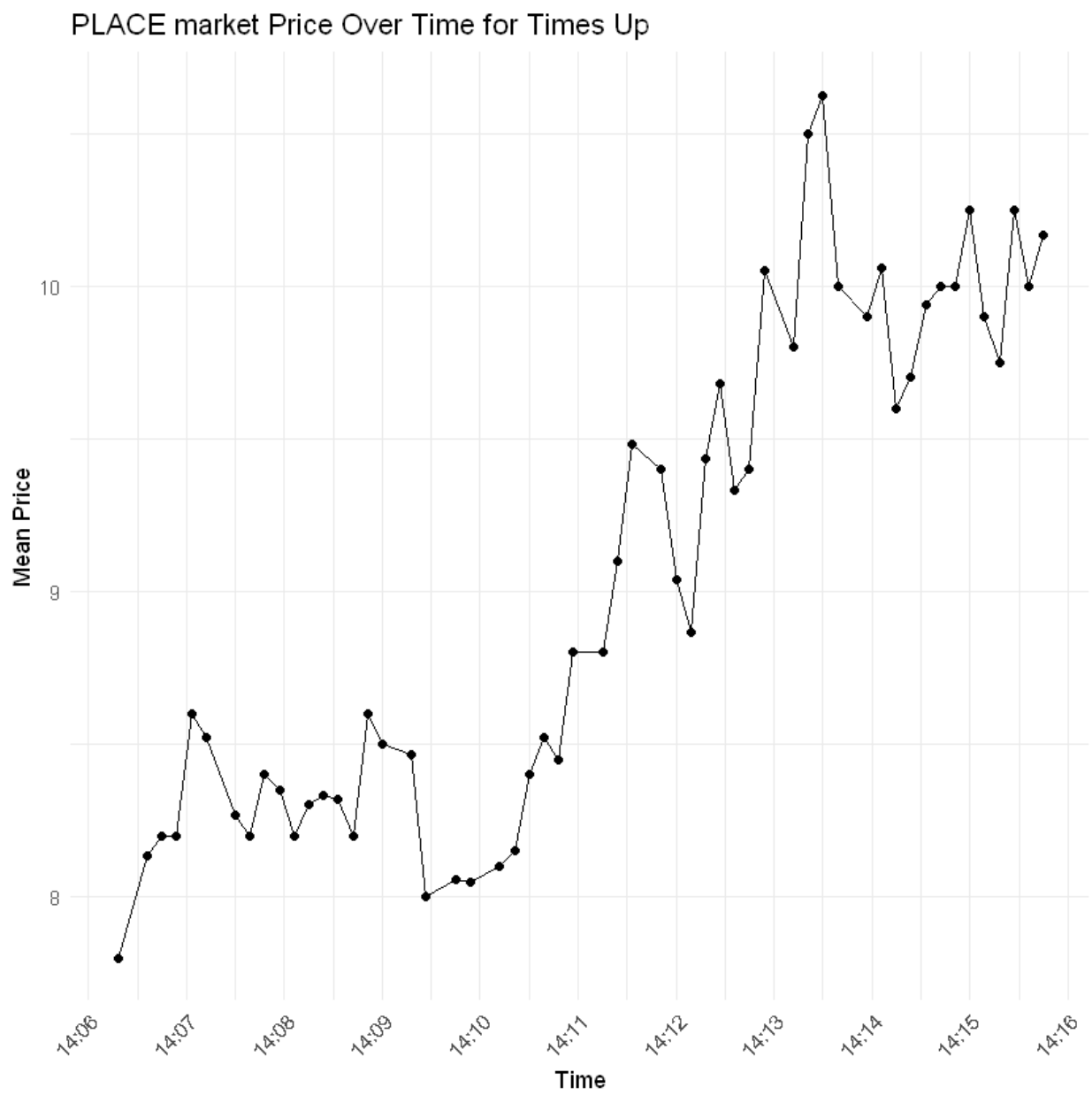


Figure 16: Image 11

```
Augmented Dickey-Fuller Test

data: WIN_price_ts
Dickey-Fuller = -1.8851, Lag order = 3, p-value = 0.6216
alternative hypothesis: stationary

Augmented Dickey-Fuller Test

data: PLACE_price_ts
Dickey-Fuller = -1.7995, Lag order = 3, p-value = 0.6563
alternative hypothesis: stationary
```

Figure 17: adf raw ts

```
Augmented Dickey-Fuller Test

data: Brown_Panther_W_diff1
Dickey-Fuller = -3.5503, Lag order = 3, p-value = 0.04463
alternative hypothesis: stationary

Augmented Dickey-Fuller Test

data: Brown_Panther_P_diff1
Dickey-Fuller = -3.5071, Lag order = 3, p-value = 0.04829
alternative hypothesis: stationary
```

Figure 18: adf diff 1

```
Augmented Dickey-Fuller Test

data: Brown_Panther_W_diff2
Dickey-Fuller = -6.3917, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary

Warning message in adf.test(Brown_Panther_P_diff2, a
"p-value smaller than printed p-value"

Augmented Dickey-Fuller Test

data: Brown_Panther_P_diff2
Dickey-Fuller = -6.6108, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary
```

Figure 19: adf diff 2

```

VAR Estimation Results:
=====
Endogenous variables: Brown_Panther_W_diff2, Brown_Panther_P_diff2
Deterministic variables: const
Sample size: 60
Log Likelihood: 326.085
Roots of the characteristic polynomial:
0.751 0.751 0.6799 0.6799 0.6462 0.5122
Call:
VAR(y = cbind(Brown_Panther_W_diff2, Brown_Panther_P_diff2),
    lag.max = 6, ic = "AIC")

Estimation results for equation Brown_Panther_W_diff2:
=====
Brown_Panther_W_diff2 = Brown_Panther_W_diff2.l1 + Brown_Panther_P_diff2.l1
                        + Brown_Panther_W_diff2.l2 + Brown_Panther_P_diff2.l2
                        + Brown_Panther_W_diff2.l3 + Brown_Panther_P_diff2.l3
                        + const

                        Estimate Std. Error t value Pr(>|t|)
Brown_Panther_W_diff2.l1 -0.6624451  0.1598545  -4.144 0.000123 ***
Brown_Panther_P_diff2.l1  0.1984273  0.4977605   0.399 0.691760
Brown_Panther_W_diff2.l2 -0.5395946  0.1822832  -2.960 0.004590 **
Brown_Panther_P_diff2.l2  0.0224809  0.5350652   0.042 0.966644
Brown_Panther_W_diff2.l3 -0.3461828  0.1568178  -2.208 0.031627 *
Brown_Panther_P_diff2.l3 -0.1749295  0.4751129  -0.368 0.714202
const                    0.0005434  0.0042112   0.129 0.897819
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03257 on 53 degrees of freedom
Multiple R-Squared: 0.3541, Adjusted R-squared: 0.281
F-statistic: 4.842 on 6 and 53 DF, p-value: 0.0005228

```

Figure 20: VAR Summary 1

```

Estimation results for equation Brown_Panther_P_diff2:
=====
Brown_Panther_P_diff2 = Brown_Panther_W_diff2.11 + Brown_Panther_P_diff2.11
                        + Brown_Panther_W_diff2.12 + Brown_Panther_P_diff2.12
                        + Brown_Panther_W_diff2.12 + Brown_Panther_P_diff2.12
                        + Brown_Panther_W_diff2.13 + Brown_Panther_P_diff2.13
                        + const
                        Estimate Std. Error t value Pr(>|t|)
Brown_Panther_W_diff2.11  0.0772935  0.0550730  1.403 0.166308
Brown_Panther_P_diff2.11 -0.6688187  0.1714884 -3.900 0.000273 ***
Brown_Panther_W_diff2.12  0.0368209  0.0628002  0.586 0.560151
Brown_Panther_P_diff2.12 -0.5640814  0.1843406 -3.060 0.003469 **
Brown_Panther_W_diff2.13 -0.0043635  0.0540268 -0.081 0.935932
Brown_Panther_P_diff2.13 -0.2514280  0.1636858 -1.536 0.130477
const                      0.0003771  0.0014508  0.260 0.795937
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01122 on 53 degrees of freedom
Multiple R-Squared: 0.3136, Adjusted R-squared: 0.2359
F-statistic: 4.035 on 6 and 53 DF, p-value: 0.002109

Covariance matrix of residuals:
      Brown_Panther_W_diff2 Brown_Panther_P_diff2
Brown_Panther_W_diff2      0.0010607      0.0002235
Brown_Panther_P_diff2      0.0002235      0.0001259

Correlation matrix of residuals:
      Brown_Panther_W_diff2 Brown_Panther_P_diff2
Brown_Panther_W_diff2      1.0000      0.6115
Brown_Panther_P_diff2      0.6115      1.0000

```

Figure 21: VAR Summary 2