ARIMA Model Predicting Future Delta of Tickets at Face Value vs Secondary Market at Scotiabank Arena

## Difference Between Stationary and Non-Stationary Variables

In the context of time series analysis, the terms "stationary" and "non-stationary" refer to the behavior of a variable over time. Understanding this distinction is crucial when using models like ARIMA. Now, why is this important for building ARIMA model? The "I" in ARIMA represents the integrated component, accounting for differencing to transform non-stationary series into stationary ones. Therefore, we need to know if our variable is stationary or non-stationary so we know if we need to take the first difference or not.

The differencing operation in ARIMA removes the trend or other non-stationary patterns, making the series stationary. ARIMA models assume that the underlying time series is stationary or approximately stationary after differencing. By making the series stationary, ARIMA models can capture the autocorrelation and predict future values based on the past behavior of the series.

### Stationary Variables

* Frequently returns to its mean
* Temporary impact doesn't last forever

### Non-Stationary Variables

* Does not frequently return to its mean
* Temporary shock has a permanent impact

## What is ARIMA

ARIMA stands for Autoregressive Integrated Moving Average. The model combines three components:

* Autoregression (AR) - captures the relationship between an observation and a lagged value
* Differencing (I) - transforms non-stationary series into stationary ones
* Moving Average (MA) - Considers the error terms and their relationship to past observations

ARIMA models use the historical behavior of a time series to predict future values, making them valuable tools for forecasting in various domains.

ARIMA models are denoted by (P,D,Q). Our first step is to find D, which is the integrated part of the model.

D = the number of times a series must be differenced to be made stationary

If D = 0, the series is stationary (I(0)).

D can also be a fraction

***Example: ARIMA (1,0,1)***

**Yt​=ϕ1​⋅Yt−1​+θ1​⋅εt−1​+εt​**  
​

*Y\_t represents the value of the time series at time 't'*

*φ₁ is the autoregressive coefficient*

*Y\_{t-1} represents the value of the time series at the previous time step*

*θ₁ is the moving average coefficient*

*ε\_{t-1} represents the error term at the previous time step*

*ε\_t represents the current error term*

**P tells you how many Y\_t variables to include**

**Q tells you how many error terms to include**

### My Approach to Constructing Models

With respect to manifest optimization, there are many sections of the arena we are trying to re-price. Thus, my approach will be as follows:

#### Individual ARIMA Models:

In this approach, I would build separate ARIMA models for each seat section. Each model would be trained and fitted specifically for the historical data of that particular section. This approach allows for customization and fine-tuning of the models based on the characteristics and trends observed in each section. You can analyze and forecast prices separately for each section, considering their unique dynamics

#### Composite ARIMA Model:

I will also create a single composite ARIMA model that incorporates data from all seat sections. In this case, I would combine the data from all sections into a single time series dataset and build an ARIMA model based on the aggregated data. This approach assumes that there are common trends and patterns across different sections, and the model captures the overall behavior of ticket prices. I can then generate forecasts for each section based on the composite model, considering any variations or adjustments specific to individual sections.

### Composite ARIMA

First, I will construct a composite ARIMA model that analyzes ticketing pricing from a holistic view. This can be especially useful if there are limited data points for some seat sections or if certain sections exhibit erratic or unpredictable price movements. A composite model may provide more reliable forecasts by leveraging the combined information.

A picture containing text, screenshot, plot, diagram

Description automatically generatedTo start, we are going to look at a time-series plot of the re-sale prices of Maple Leaf tickets from 2016-2023, grouped by day.

Off first glance, it seems that there is no trend in the data, and that the variables seems to regress to its mean. Thus, I could conclude that this dataset is a stationary, and has no unit root. However, an eye test is not enough to come to this conclusion, and we must conduct the necessary statistical testing to come to a real conclusion.

We can also check if there's a trend in this time-series data by doing an OLS regression of a constant and a time trend with % Price Delta as our dependent variable. We will also use robust standard errors. Robust standard errors provide a more reliable estimation of the standard errors of the regression coefficients in the presence of heteroscedasticity. Heteroscedasticity refers to the condition where the variance of the error term in a regression model is not constant across all levels of the independent variables.

A picture containing text, screenshot, font, number

Description automatically generated

## Testing for Heteroskedasticity

Heteroscedasticity refers to a situation in which the variability or spread of data points is not consistent across the range of values of a variable. In simpler terms, it means that the amount of variation in the data is not the same for all values of the variable being considered.

Visually, heteroscedasticity can be observed as a pattern in the scatter plot of the data points. Instead of the data points forming a consistent spread around a line or curve, they may fan out or cluster together in certain regions, indicating varying levels of variability.

Heteroscedasticity can cause issues in statistical analysis because it violates the assumption of homoscedasticity (equal variance) that is often assumed in many statistical models. It can affect the reliability of statistical tests, confidence intervals, and predictions. Therefore, it is important to identify and address heteroscedasticity to ensure accurate analysis and interpretation of the data.

If the variable displays increasing or decreasing variance over time (heteroscedasticity), taking the logarithm can help stabilize the variance. By compressing larger values and expanding smaller values, the logarithmic transformation can mitigate the impact of extreme values and make the series more stationary, which aligns with the assumptions of ARIMA models.

Lets take a look at a X-Y scatter plot of our time-series data, with X being time and Y being the resale value of tickets:

A picture containing text, screenshot, diagram, line

Description automatically generated

We can see here that our error terms look to be very spread out and far away from our line of best fit. From solely an eye-test, our variable seems to possess heteroskedasticity.

To be sure, we should conduct the White heteroscedasticity test or White's general test, which is a statistical test used to detect heteroscedasticity.

**White Test – Test for Heteroskedasticity**

* **H0: Homoskedasticity**
* **HA: Heteroskedasticity**

A picture containing text, screenshot, font, number

Description automatically generated

Overall, based on these results, it does not appear that there is heteroscedasticity present in the resale\_price variable.

# Unit Root Tests

A unit root test, also known as a stationarity test, is a statistical test used to determine if a time series has a unit root or not. A unit root indicates that the series is non-stationary, meaning it has a stochastic trend and does not possess a constant mean or variance over time. On the other hand, if the unit root test rejects the presence of a unit root, it suggests that the series is stationary. This is important, because we need to find out if our variable(s) are stationary or not to make our ARIMA model.

Unit root test the null hypothesis that a series needs to be differenced to be made stationary. Many might think that we simply need to test that I = 1 using a t-test. However, this type of thinking breaks down if our variable is not stationary, as it then does not follow the t-distribution. When I = 0 in our equation, our variable is following the Dickey-Fuller distribution.

Each time we change the deterministic in the testing equation, we get a slightly different distribution of the t-test of the null that. These tests are usually referred to as the for, as you might suspect, the no-constant (NC), constant (C), and constant and trend (CT) tests. In practice we recognize that the errors in the testing equation could contain some serial correlation(autocorrelation) and we try to control for that by using what’s called the Augmented Dickey-Fuller (ADF) test.

### Types of Unit Root Tests

There are several types of unit root tests commonly used in econometrics and time series analysis. Here are the following tests we are going to use:

**1. Augmented Dickey-Fuller (ADF) Test**

**2. Augmented Dickey-Fuller Generalized Least Squares (ADF-GLS) Test**

**3. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test**

**4. Phillips-Perron (PP) Test**

**5. Fractional Integration**

**5. LS Unit Root Test**

## Augmented Dickey Fuller

The ADF (Augmented Dickey-Fuller) test is a statistical test used to determine whether a time series has a unit root or is stationary. It is commonly used in econometrics and time series analysis. The test helps to identify the presence of a trend in the data, which is an indication of non-stationarity.

Null Hypothesis (H0): >= 0

Alternative Hypothesis (HA): < 0

##### Lags: ADF Test

\_Number of Observations\_ = 273

\_Number of Lags\_ = `r round((273)^(1/3), 2)`

= 7 lags in ADF test

\* Thus, we start at the 9th observation (F+2). The lag length, denoted by "F," represents the number of lagged differences included in the test. Adding 2 to "F" accounts for the inclusion of the constant and trend terms in the ADF regression model. \*

A screenshot of a computer program

Description automatically generated with low confidence

The results of the Augmented Dickey-Fuller (ADF) test for the resale\_price variable are provided for three different model specifications: without constant, with constant, and with constant and trend. Here's what the results indicate:

1. ***Test without constant:***

The unit-root null hypothesis being tested is that the coefficient (a) in the autoregressive (AR) model is equal to 1, implying the presence of a unit root (non-stationarity) in the resale\_price variable.

The estimated value of (a - 1) is -0.00443847, suggesting that the variable is close to being stationary.

The test statistic (tau\_nc(1)) is -0.237304, and the associated asymptotic p-value is 0.601, which is greater than the typical significance level of 0.05. Therefore, there is insufficient evidence to reject the unit-root null hypothesis, indicating that the resale\_price variable is likely non-stationary.

1. ***Test with constant:***

This model includes a constant term along with lagged values of the resale\_price variable.

The estimated value of (a - 1) is -0.46673, which suggests a stronger indication of non-stationarity compared to the previous model.

The test statistic (tau\_c(1)) is -5.7789, and the associated asymptotic p-value is 4.002e-07 (very low). This indicates strong evidence to reject the unit-root null hypothesis, suggesting that the resale\_price variable is likely stationary when a constant term is included.

1. **Test with constant and trend:**

This model includes both a constant term and a linear trend along with lagged values of the resale\_price variable.

The estimated value of (a - 1) is -0.53996, indicating a stronger indication of non-stationarity compared to the previous models.

The test statistic (tau\_ct(1)) is -6.30313, and the associated asymptotic p-value is 2.014e-07 (very low). This provides strong evidence to reject the unit-root null hypothesis, suggesting that the resale\_price variable is likely stationary when both a constant term and trend are included.

In summary, based on the ADF test results, it appears that the inclusion of a constant term and trend in the model helps to achieve stationarity in the resale\_price variable. This suggests that the variable is likely not affected by a unit root, implying that it is stationary or trend-stationary.

## Augmented Dickey Fuller Generalized Least Squares

Although the ADF test is telling us that our time-series data does not contain a unit root, it is not as very powerful test. We do have a test that is a little more powerful called the ADF-GLS test. Overall, the ADF-GLS test is a more advanced and reliable method for testing the presence of a unit root and determining the stationarity of a time series. It addresses the limitations of the standard ADF test and provides more accurate and robust results, making it a preferred choice in many empirical applications.

We choose the maximum lag length to include in the augmented version of the test just as we did with the regular ADF test, and we can include a constant or a trend in the test to see if the series has a unit root.

The test was conducted in two scenarios: one with only a constant term and another with both a constant and trend term. The test was performed with 7 lags of the first difference of "resale\_price" (denoted as (1-L)resale\_price), which helps to capture any serial correlation in the data.

In both cases, the null hypothesis being tested is that the coefficient of the lagged variable (a) is equal to 1, suggesting the presence of a unit root. If the null hypothesis is rejected, it indicates that the variable is stationary.

1. **Test with constant:**

A picture containing text, screenshot, font, number

Description automatically generated

1. **Test with constant and trend**

The estimated value of (a - 1) is -0.172003, indicating that there is a negative relationship between the lagged variable and the current variable. The test statistic (tau) is -2.52005, and the approximate p-value is 0.011. Since the p-value is less than the significance level (commonly set at 0.05), the null hypothesis of a unit root is rejected. This suggests that the variable "resale\_price" is stationary without a unit root.

1. **Test with constant:**

A screenshot of a computer code

Description automatically generated with low confidence

Similar to the previous case, the estimated value of (a - 1) is negative (-0.219026), indicating a negative relationship between the lagged variable and the current variable. The test statistic (tau) is -2.81244, and the approximate p-value is 0.051. In this case, the p-value is greater than the significance level of 0.05, so we fail to reject the null hypothesis of a unit root. This suggests that including a trend term does not sufficiently remove non-stationarity from the variable "resale\_price."

In both cases, the 1st-order autocorrelation coefficient for the error term (e) is negative, indicating some degree of serial correlation in the data. Additionally, the F-statistics for the lagged differences are significant (p-value < 0.05) in both cases, indicating the presence of lagged autocorrelation.

Overall, based on these results, it can be concluded that the variable "resale\_price" is likely stationary without a unit root, especially when considering the test results without the trend term.

## Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test: Testing for Stationarity

Even though our unit root tests allow us to test the null hypothesis that a series has a unit root, we also need a test with a null hypothesis that a series is stationary. This is the opposite of the unit root hypothesis that a series is not stationary, and has a trend. Hopefully, this type of test would provide confirmatory results leading us to conclude if our data has a unit root or not. If we do not reject the null of a unit root, we should reject the null of stationarity. If we reject the null of a unit root, we should not reject the null of stationarity. However, there are cases in which these tests may contradict each other in which we look to further tests such as fractional integration, which we will discuss later on.

If the series is stationary, this test has a certain distribution – so we compare our calculated values to the critical values, and decide whether to reject the null or not. Just like our ADF test, there is a choice of lag truncation parameter, but instead of taking the cubed root, we take our sample sized raised to the power of 1/4.

##### Lags: KPSS Test

\_Number of Observations\_ = 265 (removed F+2 obsv.)

\_Number of Lags\_ = `r round((265)^(1/4), 2)`

= 5 lags in KPSS test

**Null Hypothesis (H0):** Stationary

**Alternative Hypothesis (HA):** Non-Stationary

*\*IF THE CALC. KPSS > CRIT. KPSS, SERIES IS STATIONARY\**

The results provided are from the KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test conducted on the variable "resale\_price" with the inclusion of a trend term. The KPSS test is used to assess the stationarity of a time series variable. It tests the null hypothesis that the variable is stationary against the alternative hypothesis of non-stationarity.

Let's analyze the results for each case:

1. **Test without trend**

A picture containing text, screenshot, font, number

Description automatically generated

In this case, the KPSS test is performed without the inclusion of a trend term. The test statistic is 1.00031. The critical values at different significance levels are provided as well.

The test statistic exceeds the critical values at all significance levels, indicating that the null hypothesis of stationarity is rejected. The p-value is also less than 0.01, supporting the rejection of stationarity. These results suggest that the variable "resale\_price" is non-stationary even without a trend term.

1. **Test with trend**

A picture containing text, screenshot, font

Description automatically generated

The test statistic for this case is 0.373628. To evaluate the test, critical values are provided at different significance levels (10%, 5%, and 1%). The critical value represents the threshold beyond which the null hypothesis of stationarity is rejected. In this case, the critical values are 0.120, 0.148, and 0.217, respectively.

Since the test statistic is lower than the critical values at all significance levels, we reject the null hypothesis of stationarity. The p-value is also less than 0.01, which further supports the rejection of stationarity. These results suggest that the variable "resale\_price" is non-stationary when a trend term is included.

In summary, both cases of the KPSS test indicate that the variable "resale\_price" is non-stationary, regardless of the inclusion or exclusion of a trend term. This implies that the mean and variance of the variable change over time, indicating the need for further analysis and consideration of appropriate modeling techniques for non-stationary time series data.

Our ADF and ADF-GLS tests are telling us that our time-series is stationary, but our KPSS test is saying our time-series is non-stationary. This conflicting result suggests that the time series may have some complex characteristics, such as a combination of a stationary and non-stationary component, or a trend that is neither completely stationary nor non-stationary. Thus, we need to do further testing.

## Fractional Integration

The fractional integration test is a statistical test used to determine the level of integration (or differentiation) in a time series variable. This test is especially useful when we have conflicting inferences with our ADF tests and KPSS tests.

**Null Hypothesis (H0):** The alternative hypothesis assumes the presence of a unit root, indicating that the time series variable is integrated of order d, where d is a positive number less than 1 (0 < d < 1).

**Alternative Hypothesis (HA):** The null hypothesis assumes the absence of a unit root, suggesting that the time series variable is stationary or integrated of order 0 (d = 0).

A picture containing text, screenshot, font

Description automatically generated

The Local Whittle Estimator and GPH test results suggest that the "resale\_price" series does have a unit root. A unit root indicates that the series is non-stationary and has a random walk behavior, where shocks or changes have a permanent effect on the series.

The estimated degree of integration from both tests (0.40187 and 0.45729) indicates that the series exhibits some level of persistence or memory. This suggests that past values of the series can have a lasting impact on its future behavior.

The test statistics (4.17636 for the Local Whittle Estimator and 3.33909 for the GPH test) measure the strength of evidence against the null hypothesis of no unit root. The corresponding p-values (0.0000 and 0.0026) are very low, providing strong evidence to reject the null hypothesis.

Therefore, based on these test results, we can conclude that the "resale\_price" series does have a unit root. It is non-stationary and exhibits a random walk behavior, where changes in the series have a permanent impact.

## LS Unit Root Test

A picture containing text, handwriting, line, font

Description automatically generatedWhile the ADF or ADF-GLS tests will tell you whether the series has a unit root or not, sometimes they are not very reliable – in other words, the power of the tests can be very low. This is especially if there is a shift in one or both constant and trends in the series. For example, the series below has 100 observations, and it seems to be random until 2014, cycling around a constant of 0; from that point on, it seems to be random, cycling around a constant value of about 50.

If we perform the ADF test on the series, we will not reject the null of a unit root, with or without a trend, with a p-value of about 67% in the constant test, and over 48% in the test including a constant and trend, even though we can see that the series really does not have a unit root.

A powerful test, we can do is called the LS Unit Root Test for breaks. The LS Unit Root Test for breaks is an extension of the LS Unit Root Test that accounts for potential structural breaks in a time series variable. Structural breaks refer to shifts or changes in the underlying dynamics of the series, such as changes in trend, volatility, or relationship between variables, occurring at specific points in time.