Here is a very brief sketch of what I am thinking for my thesis project.

Please note that I did the code and graphs very quickly just to have something to show and is by no means final. I attached an R file that has the code I used to generate the graphs and output for this proof of concept. I am excited to get specific feedback on the parameters of the project and produce the best thesis I can with your help.

Based on feedback received, I think I am going to research the predictive accuracy of ARIMA models and their variants (ARMA, SARIMA, MA, AR) for the NASDAQ 100 Technology Sector Index (^NDXT). It is an index composed of the top 100 public tech companies in the US, so it seemed to be a better fit for an IT project than a general stock market. It is not really the subject of the study as much as it is relevant, interesting, and high-quality time series data.

Generally, the data from January 3, 2012 to today looks like this:

A close up of text on a white background

Description automatically generated

I just quickly did an automatically generated ARIMA model to do a tiny prediction of 20 days just to make sure it works:

A screenshot of a cell phone

Description automatically generated

The little blue line and gray shaded area show the predicted values with a confidence interval at 80% and 95% by default.

I also just quickly tested the assumptions of normality with a QQ plot, which is an important predicate of building the models.

A close up of a map

Description automatically generated

What I find more interesting is that the forecast model that is built gives the prediction numbers:

> forecastAutoARIMA

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

2096 5277.223 5215.066 5339.380 5182.161 5372.285

2097 5270.375 5191.543 5349.206 5149.812 5390.937

2098 5272.248 5174.090 5370.406 5122.129 5422.368

2099 5274.122 5159.861 5388.382 5099.375 5448.868

2100 5275.995 5147.636 5404.354 5079.687 5472.303

2101 5277.868 5136.813 5418.923 5062.143 5493.594

2102 5279.742 5127.043 5432.441 5046.208 5513.275

2103 5281.615 5118.099 5445.132 5031.539 5531.692

2104 5283.489 5109.828 5457.150 5017.897 5549.080

2105 5285.362 5102.117 5468.607 5005.113 5565.611

2106 5287.236 5094.884 5479.587 4993.059 5581.412

2107 5289.109 5088.062 5490.155 4981.635 5596.583

2108 5290.982 5081.602 5500.363 4970.762 5611.202

2109 5292.856 5075.460 5510.251 4960.378 5625.333

2110 5294.729 5069.604 5519.854 4950.430 5639.028

2111 5296.603 5064.005 5529.201 4940.875 5652.330

2112 5298.476 5058.638 5538.314 4931.675 5665.277

2113 5300.349 5053.483 5547.216 4922.800 5677.899

2114 5302.223 5048.523 5555.922 4914.223 5690.223

2115 5304.096 5043.743 5564.450 4905.920 5702.272

While I did not bother much with formatting making it pretty, the fundamental proof of concept is that I can change the parameters of models, predict for a period of time such as a week, a month, or a quarter in this tech index. In practicality, I would probably take Q1’s data and try to predict the values for the month of April due to time constraints. It would be fairly straightforward to take the prediction values for April and compare them to reality.

This would give real accuracy where I could do any number of comparisons like t-tests, ANOVA/f-tests, and plot the prediction against reality. The specifics here are up for debate, and I am happy to accept feedback about what tests should be done for each model.

Overall, I want to make some models, test them against reality, and compare the results. The timeframe to analyzed and predicted, specific tests to run, and finer details of the project definitely need refinement. I do feel confident that I am able to do the actual programming once the objectives are clearly defined. I also have seen a ton of research relating to predictive models, ARIMA models, seasonal decomposition, and a ton of other related topics to support my research.

Here is how I see the research questions:

Null Hypothesis: There is no significant difference between any of the models created and the automatically generated ARIMA model for the month of April 2020 in the NDXT Index.

Alternative Hypothesis: The automatically generated ARIMA model is the best predictor of the real values of the NDXT Index for the month of April 2020.

Please let me know what you think about this idea and any suggestions you have. I am excited to do this because it is at the intersection of a bunch of academic and professional interests.