- Are there variables that are particularly significant in terms of explaining the answer to your project questions?

  The variables at play are all similar in nature, if not the same. They are all prices. To say something about the variety, they are high, low, adjusted close, close and open prices. These are the prices picked up from the yahoo finance api. There is a multitude of ways to explore and model prices. At first I considered predicting prices of some industry based on the prices of another industry. This concept was interesting due to the relationship across certain industries and their potential dependencies. Most importantly, prices alone are sufficient for studying price movement and creating models but in theory and practice, it is more than likely the case that prediction may have more accurate outcomes, if you take into account macroeconomic factors and certain exogenous events. Adding these factors doesn't necessarily complicate the task of prediction but it does affect time to completion.

- Are there significant differences between subgroups in your data that may be relevant to your project aim?

  To my knowledge, there are no subgroups in the price data we're working with. However, exploring the nature of groups of values by correlation may be a fruitful path to research. The final set of tickers used for modeling after portfolio selection had normal price movement, in the statistical sense. Outside of that, the prices were at different levels. Exploring levels is another path for research in the future. I believe using fundamentals would be the simplest path to a solution, though an expensive one.

- Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

  After careful consideration, the ultimate time series approach to modelling was a classical one, the autoregressive approach. There is much literature to read there about the varieties of autoregression models but the decision was final to take a univariate time dependent approach. So I took a careful look at the autocorrelation and partial autocorrelation methods to help tune the autoregression models. For instance, arima model. The 3 parameters p,d,q used to implement the arima method are crucial, where p represents the number of lags, d the number times to difference and q the number of lag errors to use in the moving average process. The equation is linear. In arima models, the order of any p,d,q generally doesn't exceed 3. Returning to the issue, an automated model determines on its own, using the acf and pacf, the best parameters for prediction.

When correlations are found in the lags, lags will be used in the model. They are useful for prediction and potentially forecasting of errors.

- What are the most appropriate tests to use to analyze these relationships?

  As previously noted, autocorrelation test and partial autocorrelation tests were used.