CS210 Project Phase 2 Bartu Sisman 28038

Datasets:
2019-2024 Stock market

url:

https://www.kaggle.com/datasets/saketk511/2019-2024-us-stock-market-data?resource=download

1980-2024 World GDP Growth

url:

https://www.kaggle.com/datasets/sazidthe1/world-gdp-growth

Explanation:
First I manually deleted all the columns of World GDP Growth dataset that correspond to years between 1980-2018 to align it with my Stock Market dataset, In Stock Market dataset I deleted all columns except for "Number", "Date", "Natural_Gas_Price", and " S&P_500_Price" . So, the datasets are clearer.

My basic purpose is to learn about how did change in gas price affected the buying power of all people (by comparing GDP with Natural_Gas_Price) and how did the change of Natural_Gas_Price affected the big businesses or the economy(S&P500 with Natural_Gas_Price)

1. Stock Market Dataset

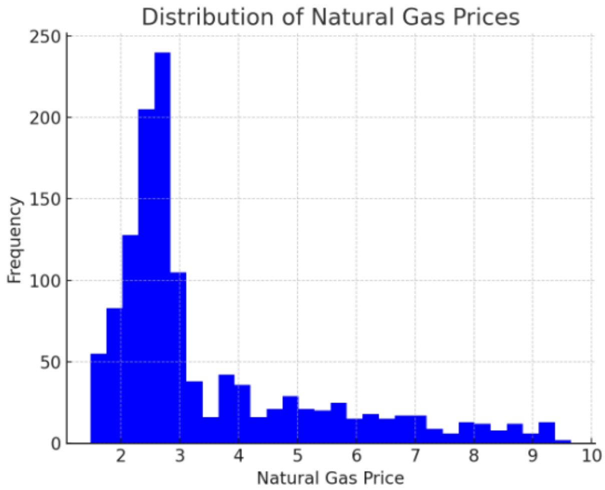
Data Types:

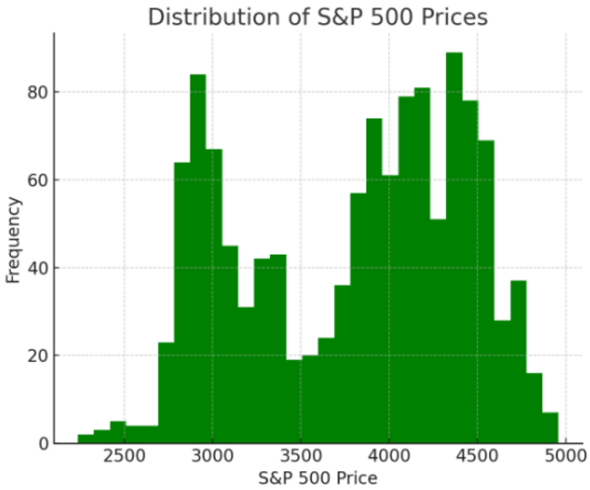
- We have identified the data types as follows:
    - **Number**: Integer
    - **Date**: String (needs conversion for time-series analysis)
    - **Natural_Gas_Price**: Float
    - **S&P_500_Price**: String (needs conversion to float)

Natural_Gas_Price has a mean of 3.495 and a variance of approximately 3.321

Correlations:

- I convert the "S&P_500_Price" to numeric value and checked for Pearson correlation between "Natural_Gas_Price" and "S&P_500_Price" to analyze the linear relationships between them.

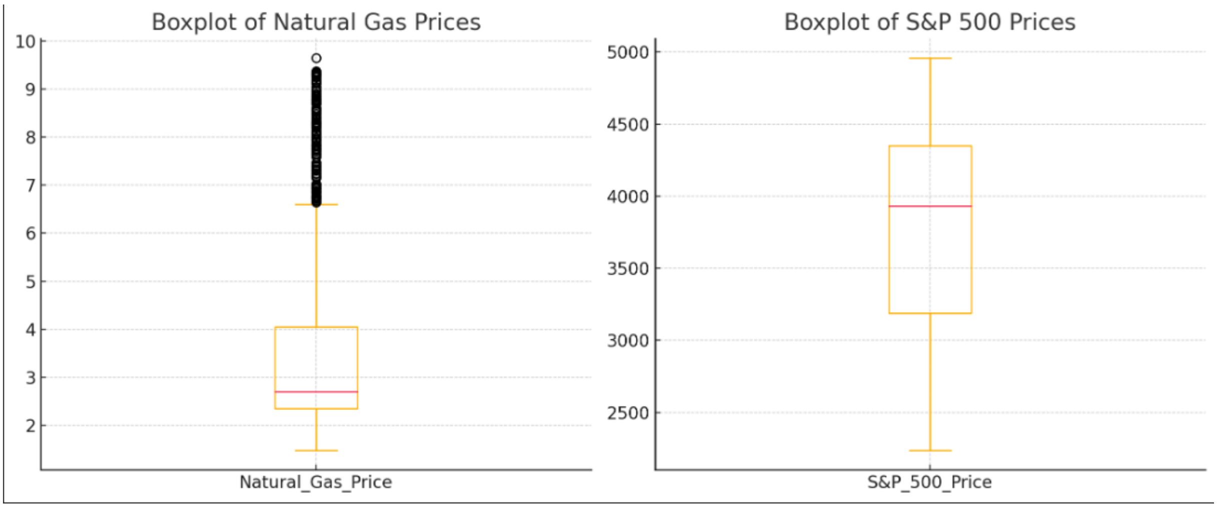




Histograms:

Natural Gas Prices: The histogram of natural gas prices is slightly right-skewed, meaning most of the prices are on the lower side with a few exceptions that are much higher. This kind of distribution suggests that while natural gas prices are usually stable, they sometimes spike due to specific conditions or events.

S&P 500 Prices: The histogram for the S&P 500 prices shows a right-skewed distribution as well, which tells us that the prices are generally trending upwards over time, though there are periods of volatility where prices fluctuate more significantly..



Boxplots:

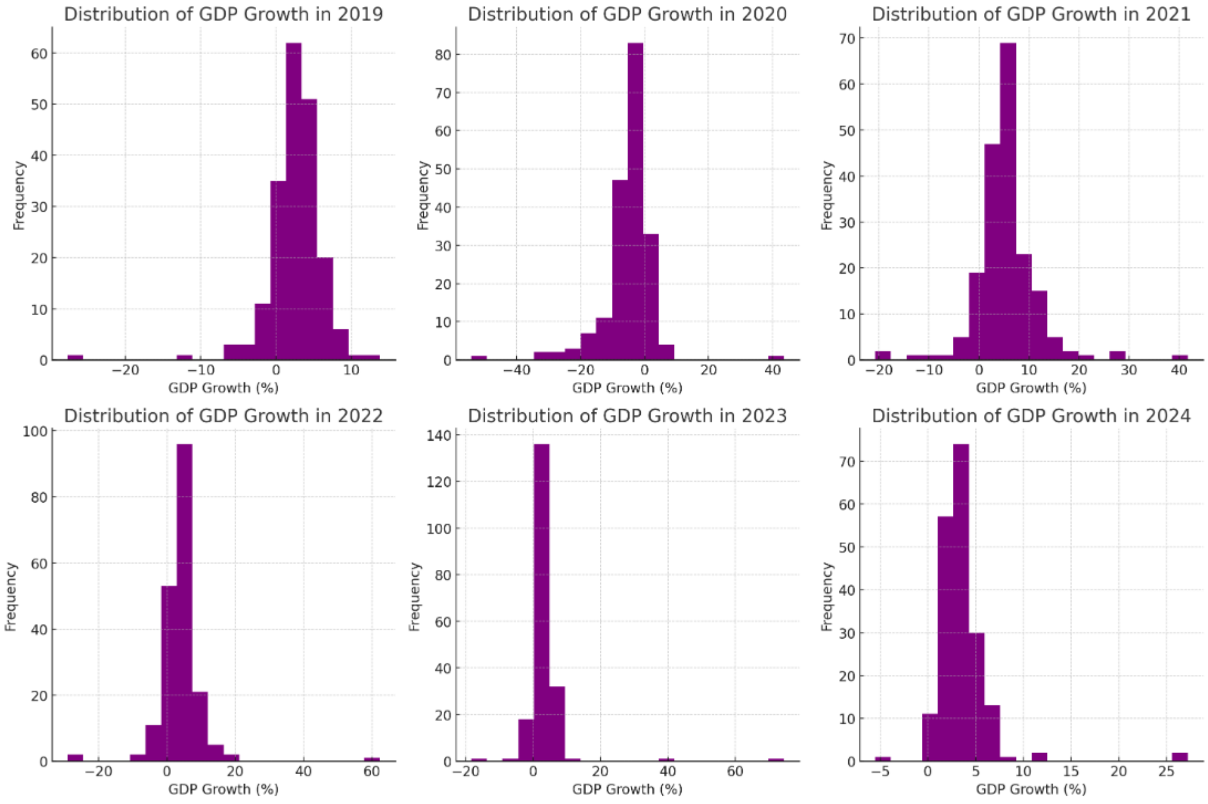
Natural Gas Prices: The boxplot for natural gas prices highlights several outliers, particularly on the higher end of the price range. These outliers could be related to moments of high market volatility or particular external factors that drove prices up temporarily.

S&P 500 Prices: Similarly, the boxplot for the S&P 500 prices also shows outliers, mainly on the higher side. These could represent times when the market performed exceptionally well, potentially during economic recoveries or when investor confidence was particularly high.

2.World GDP Dataset

Data Types:

- **country_name**: Object (String) - This column contains country names
- **indicator_name**: Object (String) - This column details the economic indicator, typically GDP growth rates.
- **2019 to 2024**: Float64 - These columns represent the GDP growth rates for each respective year. These are numerical data suitable for various statistical analyses and trend assessments. Each value describes a percentage change in GDP, allowing for decimal precision.



Distribution of GDP Growth in 2019

Distribution of GDP Growth in 2020

Distribution of GDP Growth in 2021

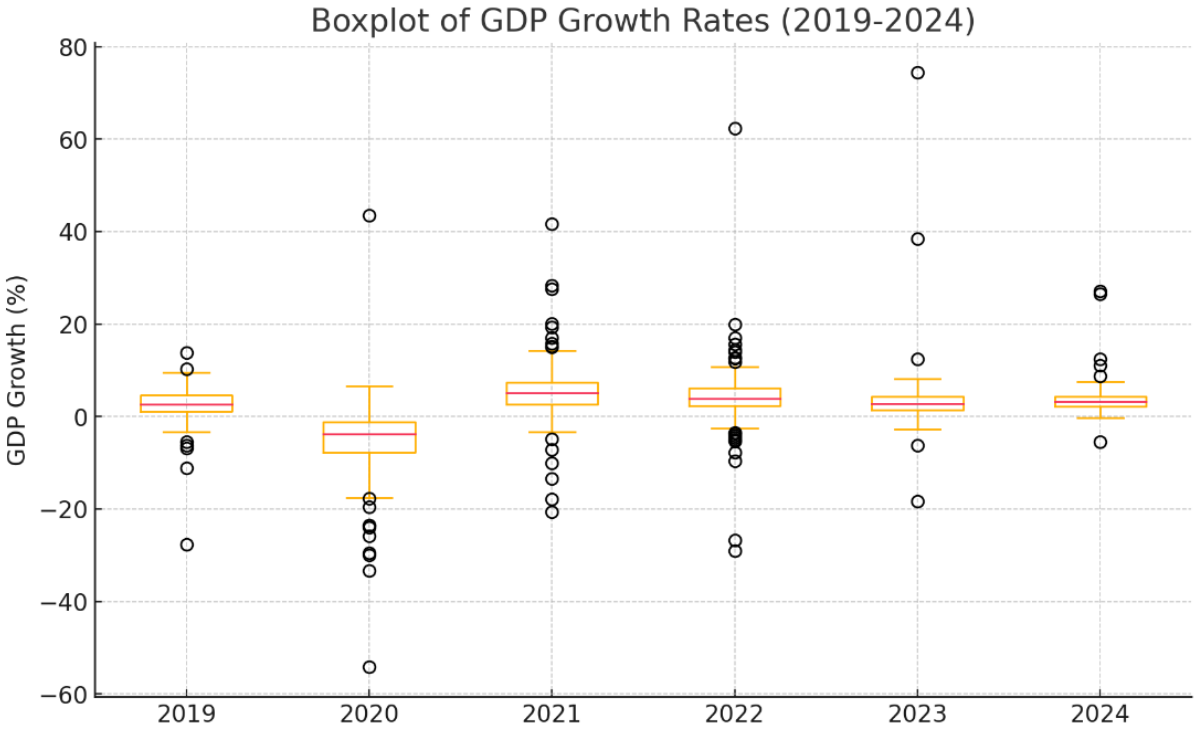
Distribution of GDP Growth in 2022

Distribution of GDP Growth in 2023

Distribution of GDP Growth in 2024

Histograms:

The histogram for 2020 is really skewed to the left, showing that a lot of countries saw their economies shrink because of the COVID-19 pandemic. It's pretty clear just from looking at the chart. The other years are more balanced, with most of the growth rates hanging around the lower single digits, which seems more normal.



Boxplots:

These plots for 2020 are full of outliers on the low end, pointing to countries that really struggled economically during the pandemic. It's quite a contrast compared to other years, where you see outliers on both the high and low ends, suggesting that economic performance varied a lot more from country to country.

Hypothesis Formulation

Null Hypothesis (H0): Changes in natural gas prices are more closely correlated with changes in S&P 500 prices than with changes in GDP growth rates.

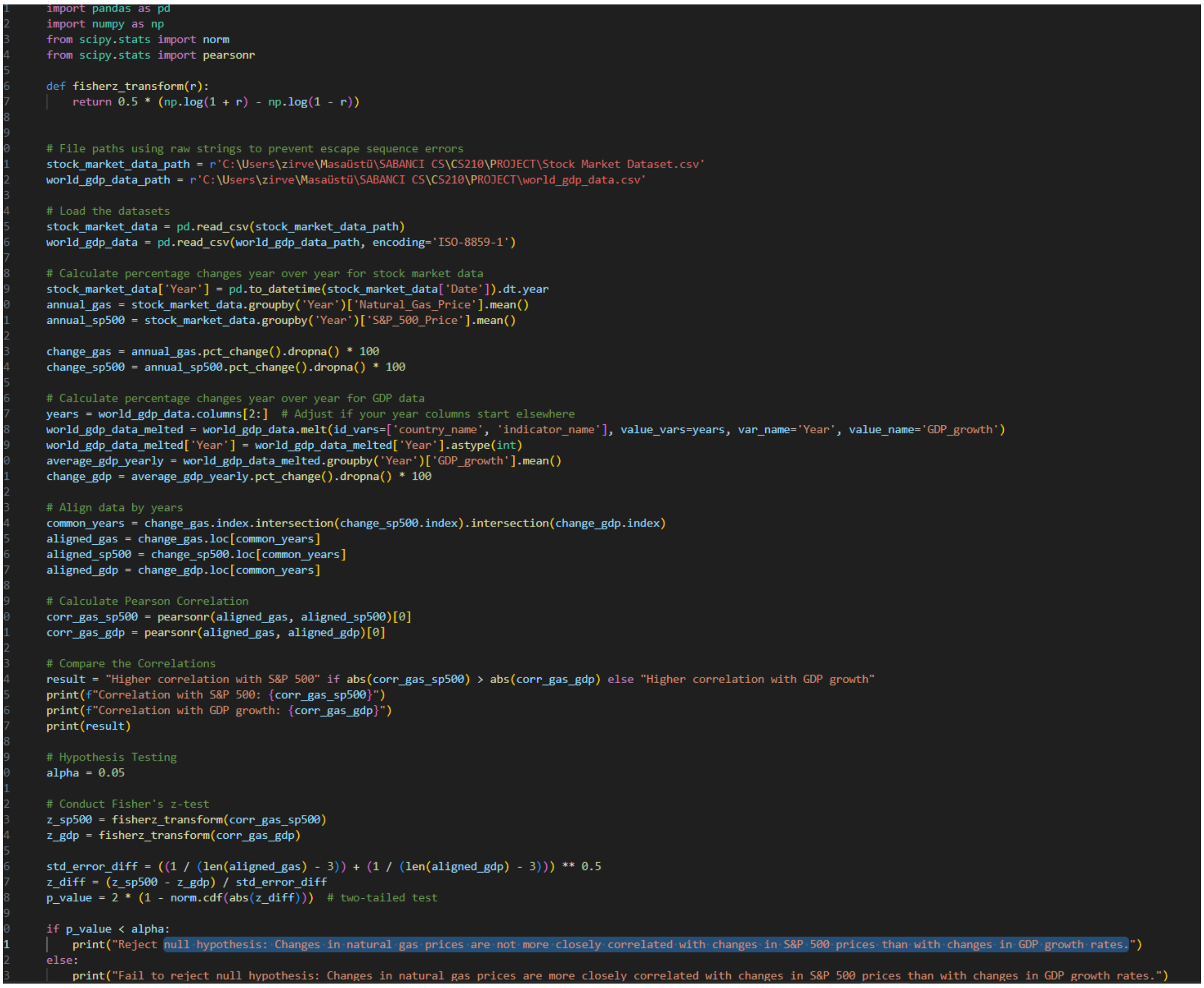
Alternative Hypothesis (HA): Changes in natural gas prices are not more closely correlated with changes in S&P 500 prices than with changes in GDP growth rates.

Hypothesis Testing

To test this hypothesis, we need to:

1. Calculate the correlation between changes (year-to-year differences) in natural gas prices and changes in S&P 500 prices.
2. Calculate the correlation between changes in natural gas prices and changes in GDP growth rates.
3. Compare these correlations to see if the correlation with S&P 500 prices is indeed stronger.

Since our datasets are time-series, we will first compute the yearly changes for natural gas prices, S&P 500 prices, and GDP growth rates. Then we'll calculate the Pearson correlation coefficients for these changes and use statistical tests to evaluate the significance of the differences between these correlations.



```python
import pandas as pd
import numpy as np
from scipy.stats import norm
from scipy.stats import pearsonr

def fisherz_transform(r):
    return 0.5 * (np.log(1 + r) - np.log(1 - r))


# File paths using raw strings to prevent escape sequence errors
stock_market_data_path = r'C:\Users\zirve\Masaüstü\SABANCI CS\CS210\PROJECT\Stock Market Dataset.csv'
world_gdp_data_path = r'C:\Users\zirve\Masaüstü\SABANCI CS\CS210\PROJECT\world_gdp_data.csv'

# Load the datasets
stock_market_data = pd.read_csv(stock_market_data_path)
world_gdp_data = pd.read_csv(world_gdp_data_path, encoding='ISO-8859-1')

# Calculate percentage changes year over year for stock market data
stock_market_data['Year'] = pd.to_datetime(stock_market_data['Date']).dt.year
annual_gas = stock_market_data.groupby('Year')['Natural_Gas_Price'].mean()
annual_sp500 = stock_market_data.groupby('Year')['S&P_500_Price'].mean()

change_gas = annual_gas.pct_change().dropna() * 100
change_sp500 = annual_sp500.pct_change().dropna() * 100

# Calculate percentage changes year over year for GDP data
years = world_gdp_data.columns[2:]  # Adjust if your year columns start elsewhere
world_gdp_data_melted = world_gdp_data.melt(id_vars=['country_name', 'indicator_name'], value_vars=years, var_name='Year', value_name='GDP_growth')
world_gdp_data_melted['Year'] = world_gdp_data_melted['Year'].astype(int)
average_gdp_yearly = world_gdp_data_melted.groupby('Year')['GDP_growth'].mean()
change_gdp = average_gdp_yearly.pct_change().dropna() * 100

# Align data by years
common_years = change_gas.index.intersection(change_sp500.index).intersection(change_gdp.index)
aligned_gas = change_gas.loc[common_years]
aligned_sp500 = change_sp500.loc[common_years]
aligned_gdp = change_gdp.loc[common_years]

# Calculate Pearson Correlation
corr_gas_sp500 = pearsonr(aligned_gas, aligned_sp500)[0]
corr_gas_gdp = pearsonr(aligned_gas, aligned_gdp)[0]

# Compare the Correlations
result = "Higher correlation with S&P 500" if abs(corr_gas_sp500) > abs(corr_gas_gdp) else "Higher correlation with GDP growth"
print(f"Correlation with S&P 500: {corr_gas_sp500}")
print(f"Correlation with GDP growth: {corr_gas_gdp}")
print(result)

# Hypothesis Testing
alpha = 0.05

# Conduct Fisher's z-test
z_sp500 = fisherz_transform(corr_gas_sp500)
z_gdp = fisherz_transform(corr_gas_gdp)

std_error_diff = ((1 / (len(aligned_gas) - 3)) + (1 / (len(aligned_gdp) - 3))) ** 0.5
z_diff = (z_sp500 - z_gdp) / std_error_diff
p_value = 2 * (1 - norm.cdf(abs(z_diff)))  # two-tailed test

if p_value < alpha:
    print("Reject null hypothesis: Changes in natural gas prices are not more closely correlated with changes in S&P 500 prices than with changes in GDP growth rates.")
else:
    print("Fail to reject null hypothesis: Changes in natural gas prices are more closely correlated with changes in S&P 500 prices than with changes in GDP growth rates.")
```

According to the code we accept the hypothesis.

Plot Interpretation and Trends

The scatter plot visualizing S&P 500 prices relative to natural gas prices reveals a mix of trends and dispersion. The actual data, marked by red dots, displays a wide scatter especially at lower natural gas prices, suggesting no clear linear relationship across the range. The predictive model, shown by the blue line, however, indicates a positive slope, implying that an increase in natural gas prices could lead to a rise in S&P 500 prices, particularly noticeable at higher natural gas price points.

Model Fit and Implications

The linear regression model appears to better predict S&P 500 prices at higher natural gas prices, where the data points are less scattered. This suggests that the model might be more effective under conditions of elevated natural gas prices. The spread around the predictive line, however, indicates significant variability unaccounted for by natural gas prices alone, hinting at other economic or market factors influencing the S&P 500.

Conclusion

In my analysis, the linear regression model suggests a generally positive correlation between natural gas prices and S&P 500 prices, which aligns with the economic theory that higher energy prices may indicate robust economic activity. This relationship is captured in the model's tendency to predict rising S&P 500 prices as natural gas prices increase.

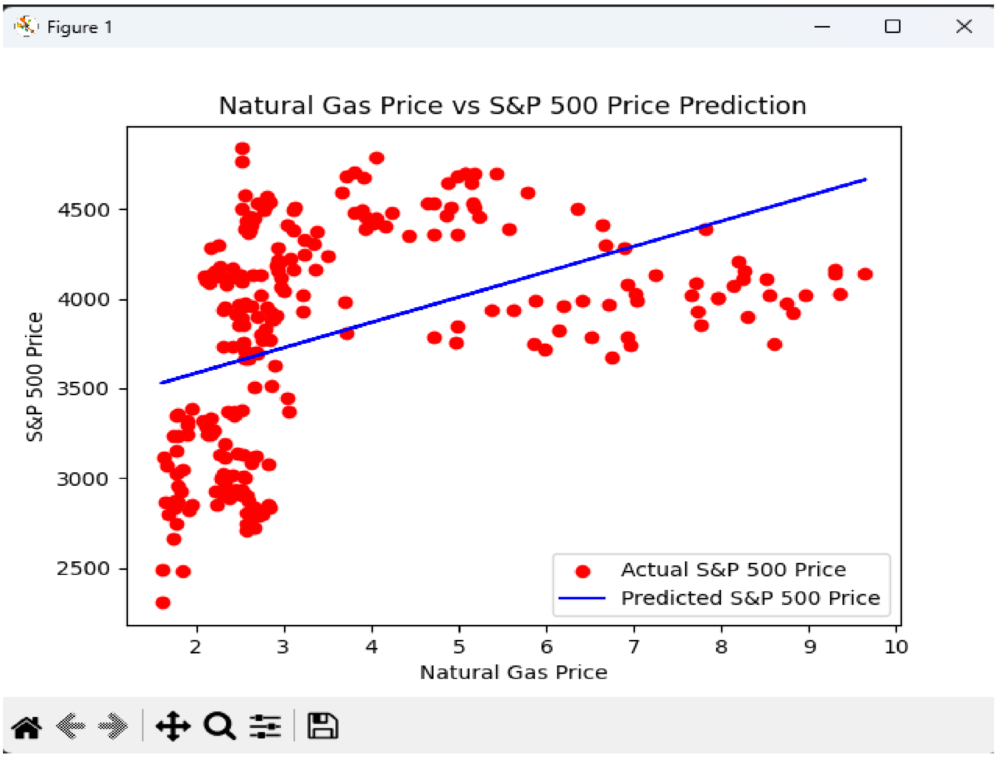


Figure 1
Natural Gas Price vs S&P 500 Price Prediction
S&P 500 Price
4500
4000
3500
3000
2500
2
3
4
5
6
7
8
9
10
Natural Gas Price
● Actual S&P 500 Price
Predicted S&P 500 Price