Centene Daily Closing Price

Takahiro Wada

2023-04-16

Introduction

Our goal is to identify the most effective multiple variable model for predicting the daily closing price of Centene (CNC). To achieve this objective, we will leverage the predictive power of four key stocks - Anthem (ANTM), Cigna (CI), United Health Group (UNH), and Humana (HUM) - and examine their relationship with CNC.

Once we have established the most effective predictive model, we will be able to use it to inform our investment decisions and optimize our portfolio. By leveraging data-driven insights, we can make informed investment choices that lead to greater returns and success in the marketplace.

Overall, our analysis will provide valuable insights into the relationship between these stocks and CNC, and help us make more informed decisions when it comes to investing in the healthcare sector.

The Data

```
importdata <- read_excel("ClosingPrices.xlsx")
mydata <- importdata[,c(19,9,22,34,57)]
sum(is.na(mydata))
```

```
[1] 0
```

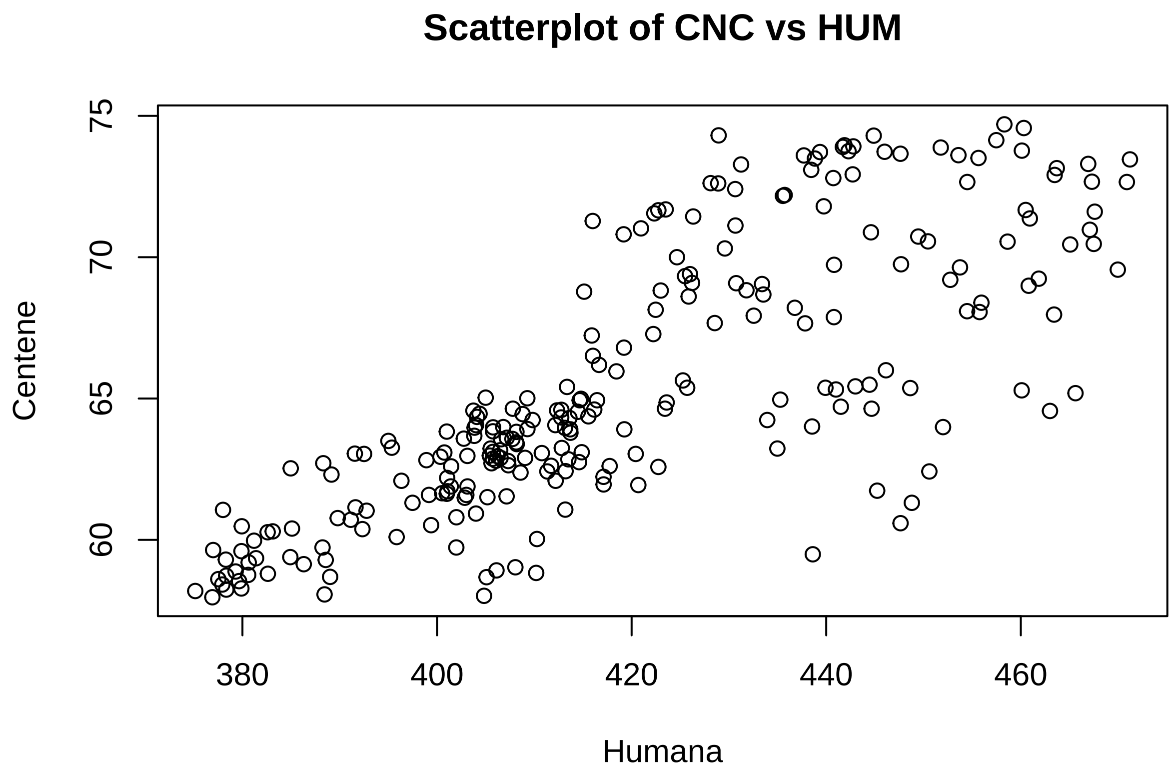
I loaded in the data into R from the Excel file. The file contains 58 different healthcare stocks daily closing price, but I chose only the 5 stocks that are in the same GICS sub-industry category as Centene, Managed Healthcare. I checked to see any missing values or error in the data set, but there is none. This data set was collected by the NASDAQ historical quotes from November 2, 2020 to October 28, 2021.

Analyzing the Variables

I made a scatter plot with each predictor variable to visually analyze the graph and identify any relationship with Centene's daily closing price. This will give us an idea of which predictor variable will be most significant towards our final model.

Scatter Plots of CNC VS HUM

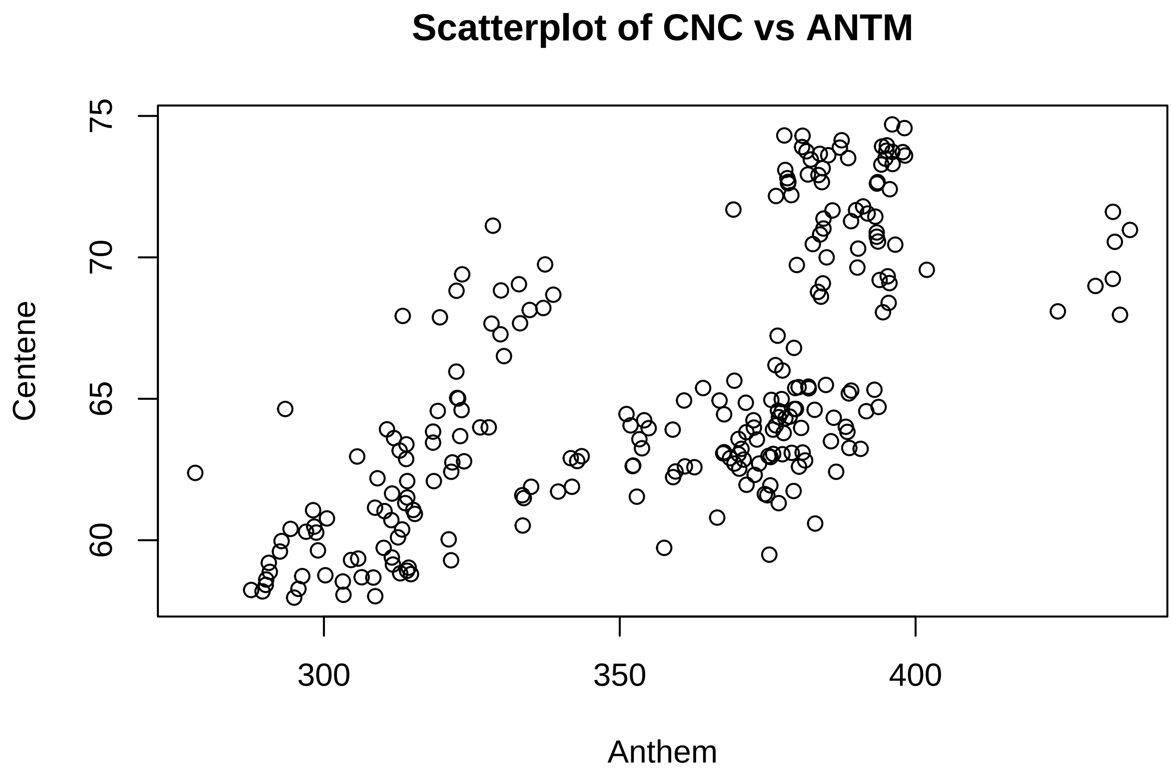
```
plot(x = mydata$HUM, y = mydata$CNC, main = "Scatterplot of CNC vs HUM",
     xlab = "Humana", ylab = "Centene")
```



We see a moderate linear relationship between Centene and Humana with moderate dispersion in the upper right region.

Scatter Plots of CNC VS ANTM

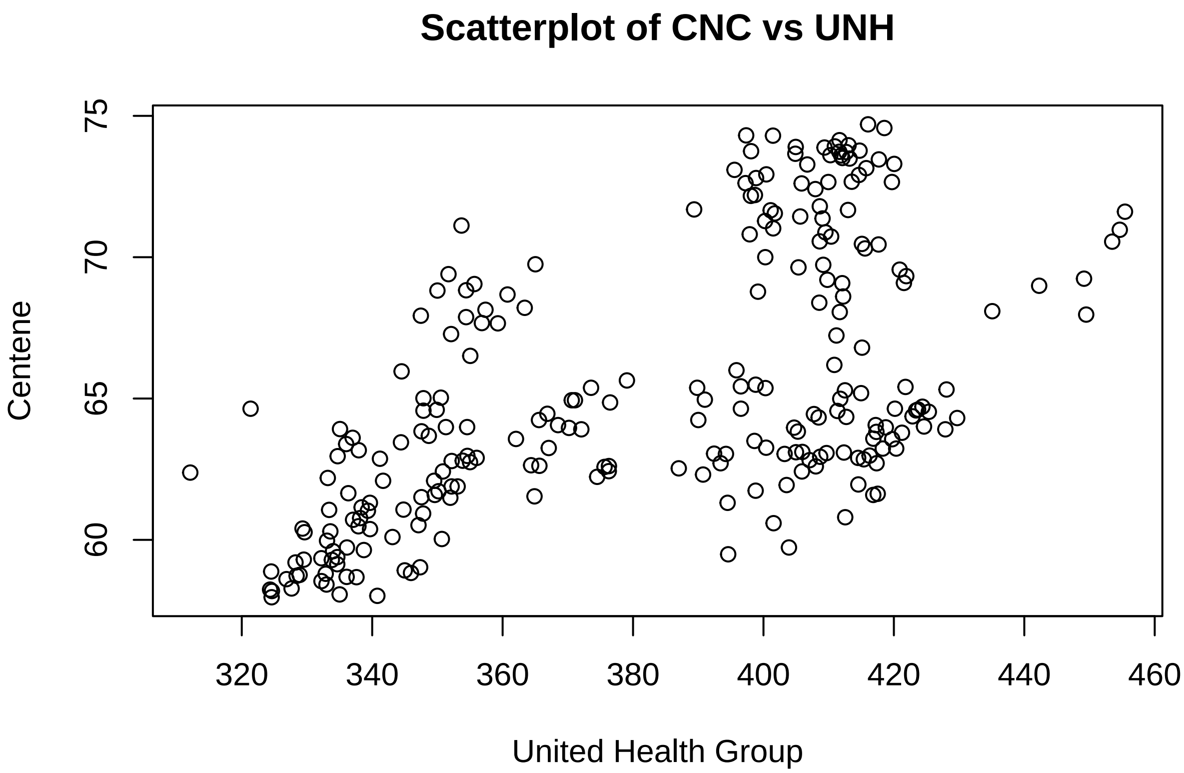
```
plot(x = mydata$ANTM, y = mydata$CNC, main = "Scatterplot of CNC vs ANTM",
     xlab = "Anthem", ylab = "Centene")
```



We see a moderate linear relationship between Centene and Anthem with moderate dispersion in the upper right region.

Scatter Plots of CNC VS UNH

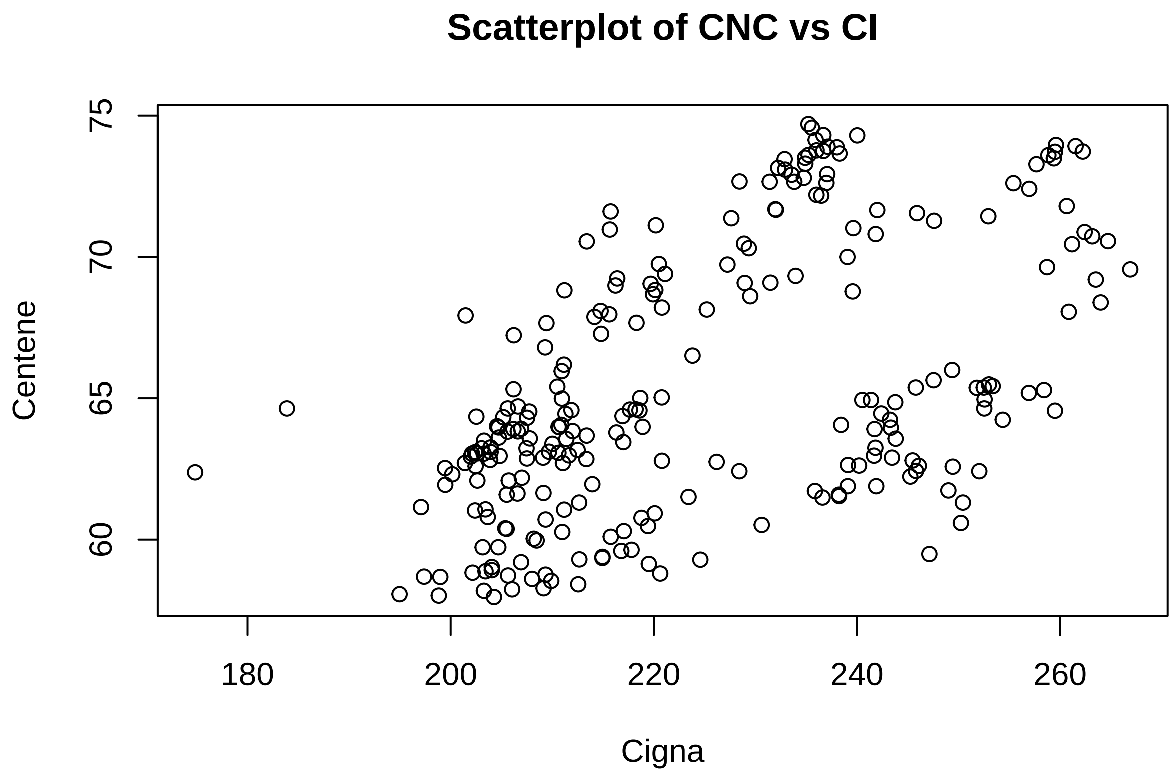
```
plot(x = mydata$UNH, y = mydata$CNC, main = "Scatterplot of CNC vs UNH",
     xlab = "United Health Group", ylab = "Centene")
```



We see a moderate linear relationship between Centene and United Health Group with moderate dispersion in the upper region.

Scatter Plots of CNC VS CI

```
plot(x = mydata$CI, y = mydata$CNC, main = "Scatterplot of CNC vs CI",
     xlab = "Cigna", ylab = "Centene")
```



We see a moderate linear relationship between Centene and Cigna with moderate dispersion in the upper region.

Verifying Error Assumptions

We must confirm that our data set is fit to be used in our model by analyzing the error.

Sum of Residuals

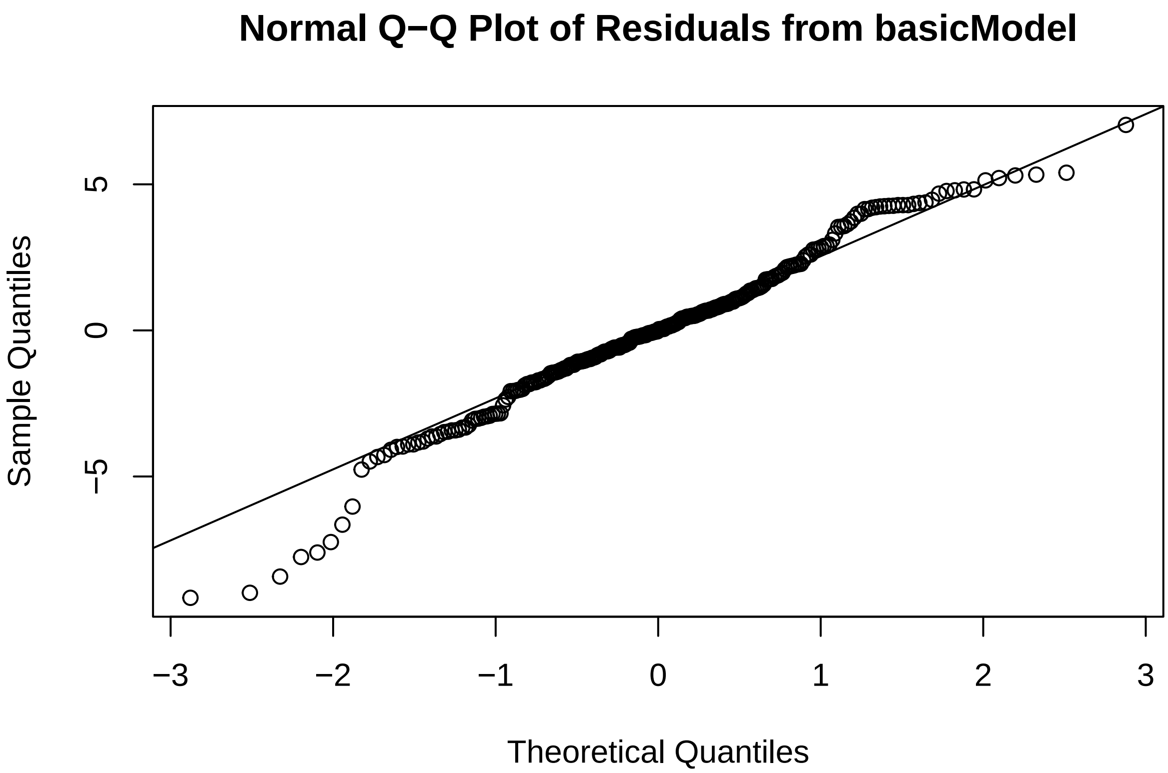
```
basicModel <- lm(mydata$CNC ~ mydata$ANTM + mydata$CI + mydata$HUM + mydata$UNH, data = mydata)
sum(resid(basicModel))
```

```
[1] 1.464107e-15
```

As we can see, the sum of errors is very close to 0. This indicates that the model is a good fit for the data, and that the predicted values are similar to the actual values. However, simply adding up residuals is not enough to determine the overall goodness of fit of the model. We will next plot a normal Q-Q plot to visually see if the data is normally distributed.

Normal Q-Q plot

```r
residual <- basicModel$residuals
qqnorm(residual, main="Normal Q-Q Plot of Residuals from basicModel")
qqline(residual)
```



As we can see from the graph, the data is not perfectly linear with the Q-Q line. Based off the curvature on both ends of the graph, a quadratic terms may help the normality of the error terms.

Correlation Matrix

```r
mydata.rcorr = rcorr(as.matrix(mydata))
mydata.rcorr
```

```
      CNC ANTM   CI  HUM  UNH
CNC  1.00 0.66 0.52 0.78 0.59
ANTM 0.66 1.00 0.47 0.72 0.96
CI   0.52 0.47 1.00 0.57 0.31
HUM  0.78 0.72 0.57 1.00 0.65
UNH  0.59 0.96 0.31 0.65 1.00

n= 250


P
```

```
     CNC ANTM CI HUM UNH
CNC        0    0  0   0
ANTM  0         0  0   0
CI    0    0       0   0
HUM   0    0    0      0
UNH   0    0    0  0
```

We want to determine numerically which predictor variables are linearly correlated with Centene. Scatter plots alone is not enough to determine which variables to use. Looking at the correlation matrix, each variable had a linear coefficient value greater than 0.50 and a p-value less than 0.05 making each predictor variable significant. We can see that Humana has the highest linear coefficient value.

Variable Selection Process

Based off the scatter plots and correlation matrix, each predictor variable seems significant enough to be included towards the model. However, to make sure we have selected the right variables to create our model, we will run a Stepwise Akaike Information Criterion (AIC) regression. It involves a step-by-step process of adding or removing variables from a model based on their statistical significance and their contribution to the overall goodness of fit of the model.

The AIC is a measure of the relative quality of a statistical model, and it is based on the likelihood function of the model and the number of parameters included in the model. The lower the AIC value, the better the model fits the data.

Stepwise AIC Variable Selection

```
model <- lm(mydata$CNC ~ mydata$ANTM + mydata$CI + mydata$HUM + mydata$UNH, data = mydata)
ols_step_both_aic(model, details = TRUE)
```

```
Stepwise Selection Method
-------------------------

Candidate Terms:

1 . mydata$ANTM
2 . mydata$CI
3 . mydata$HUM
4 . mydata$UNH

 Step 0: AIC = 1482.59
 mydata$CNC ~ 1


Variables Entered/Removed:

                             Enter New Variables
---------------------------------------------------------------------------
Variable        DF       AIC         Sum Sq         RSS       R-Sq     Adj. R-Sq
---------------------------------------------------------------------------
mydata$HUM       1     1248.159     3315.184     2105.311     0.612         0.610
mydata$ANTM      1     1339.435     2387.440     3033.054     0.440         0.438
```

```
mydata$UNH      1    1375.523    1916.437    3504.058    0.354         0.351
mydata$CI       1    1404.172    1490.972    3929.522    0.275         0.272
----------------------------------------------------------------------------

- mydata$HUM added


 Step 1 : AIC = 1248.159
 mydata$CNC ~ mydata$HUM

                             Enter New Variables
----------------------------------------------------------------------------
Variable       DF      AIC        Sum Sq        RSS       R-Sq     Adj. R-Sq
----------------------------------------------------------------------------
mydata$ANTM     1    1235.461    3435.386    1985.108    0.634         0.631
mydata$UNH      1    1242.199    3381.156    2039.338    0.624         0.621
mydata$CI       1    1243.631    3369.443    2051.051    0.622         0.619
----------------------------------------------------------------------------

- mydata$ANTM added


 Step 2 : AIC = 1235.461
 mydata$CNC ~ mydata$HUM + mydata$ANTM

                         Remove Existing Variables
----------------------------------------------------------------------------
Variable       DF      AIC        Sum Sq        RSS       R-Sq     Adj. R-Sq
----------------------------------------------------------------------------
mydata$ANTM     1    1248.159    3315.184    2105.311    0.612         0.610
mydata$HUM      1    1339.435    2387.440    3033.054    0.440         0.438
----------------------------------------------------------------------------

                            Enter New Variables
---------------------------------------------------------------------------
Variable      DF      AIC        Sum Sq        RSS       R-Sq     Adj. R-Sq
---------------------------------------------------------------------------
mydata$CI      1    1232.748    3472.460    1948.034    0.641         0.636
mydata$UNH     1    1233.387    3467.478    1953.016    0.640         0.635
---------------------------------------------------------------------------

- mydata$CI added


 Step 3 : AIC = 1232.748
 mydata$CNC ~ mydata$HUM + mydata$ANTM + mydata$CI

                         Remove Existing Variables
----------------------------------------------------------------------------
Variable       DF      AIC        Sum Sq        RSS       R-Sq     Adj. R-Sq
----------------------------------------------------------------------------
mydata$CI       1    1235.461    3435.386    1985.108    0.634         0.631
mydata$ANTM     1    1243.631    3369.443    2051.051    0.622         0.619
mydata$HUM      1    1314.644    2695.663    2724.831    0.497         0.493
```

Enter New Variables

Variable	DF	AIC	Sum Sq	RSS	R-Sq	Adj. R-Sq
mydata$UNH	1	1233.872	3479.275	1941.219	0.642	0.636

No more variables to be added or removed.

Final Model Output

Model Summary

R	0.800	RMSE	2.814
R-Squared	0.641	Coef. Var	4.305
Adj. R-Squared	0.636	MSE	7.919
Pred R-Squared	0.630	MAE	2.133

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	3472.460	3	1157.487	146.169	0.0000
Residual	1948.034	246	7.919		
Total	5420.494	249			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	4.332	3.148		1.376	0.170	-1.868	10.533
mydata$HUM	0.111	0.011	0.583	9.904	0.000	0.089	0.133
mydata$ANTM	0.026	0.007	0.199	3.607	0.000	0.012	0.040
mydata$CI	0.024	0.011	0.101	2.164	0.031	0.002	0.047

Stepwise Summary

Variable	Method	AIC	RSS	Sum Sq	R-Sq	Adj. R-Sq
mydata$HUM	addition	1248.159	2105.311	3315.184	0.61160	0.61004
mydata$ANTM	addition	1235.461	1985.108	3435.386	0.63378	0.63081
mydata$CI	addition	1232.748	1948.034	3472.460	0.64062	0.63623

```
----------------------------------------------------------------------------------
```

From the four variables, the selection process did not include United Health Group into the final model as it determines it was not significant towards the model. We will use Anthem (ANTM), Cigna (CI), and Humana (HUM) in our model. Currently, the model has an Adjusted R-Squared of 0.636 and RMSE of 2.814. Hopefully adding more complex terms will improve the overall model.

Interaction Terms

Since all the stocks are in the same sub-industry of healthcare, it is logical to think that each stock may be related to one another. We will add the interaction terms into the model and analyze each interaction variables p-values in hopes of improvement.

```
InteractionModel <- lm(mydata$CNC ~ mydata$ANTM + mydata$CI +
                         mydata$HUM + mydata$ANTM*mydata$CI +
                         mydata$ANTM*mydata$HUM + mydata$CI*mydata$HUM,
                          data = mydata)
summary(InteractionModel)
```

```
Call:
lm(formula = mydata$CNC ~ mydata$ANTM + mydata$CI + mydata$HUM +
    mydata$ANTM * mydata$CI + mydata$ANTM * mydata$HUM + mydata$CI *
    mydata$HUM, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7221 -1.4866  0.2449  1.6766  6.4482

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -1.221e+02  4.428e+01  -2.759  0.00625 **
mydata$ANTM            -3.171e-01  1.209e-01  -2.623  0.00928 **
mydata$CI               3.942e-01  1.963e-01   2.008  0.04572 *
mydata$HUM              8.114e-01  1.373e-01   5.909 1.16e-08 ***
mydata$ANTM:mydata$CI   2.179e-03  4.493e-04   4.850 2.20e-06 ***
mydata$ANTM:mydata$HUM -2.892e-04  1.991e-04  -1.452  0.14766
mydata$CI:mydata$HUM   -2.754e-03  6.019e-04  -4.575 7.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.669 on 243 degrees of freedom
Multiple R-squared:  0.6806,    Adjusted R-squared:  0.6727
F-statistic: 86.28 on 6 and 243 DF,  p-value: < 2.2e-16
```

From the summary, we can see that the interaction term ANTMxHUM is not significant and will be removed from the model. We also see that adding the interaction terms have increased the Adjusted R-Squared value.

Quadratic Terms

We will next add quadratic terms to the model and determine if they are significant to the model.

```r
sqANTM <- mydata$ANTM^2
sqCI <- mydata$CI^2
sqHUM <- mydata$HUM^2
sqUNH <- mydata$UNH^2
QuadraticModel <- lm(mydata$CNC ~ mydata$ANTM + mydata$CI + mydata$HUM +
                         mydata$ANTM*mydata$CI + mydata$CI*mydata$HUM + sqANTM + sqCI + sqHUM,
                          data = mydata)
summary(QuadraticModel)
```

```

Call:
lm(formula = mydata$CNC ~ mydata$ANTM + mydata$CI + mydata$HUM +
    mydata$ANTM * mydata$CI + mydata$CI * mydata$HUM + sqANTM +
    sqCI + sqHUM, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6761 -1.3554  0.0267  1.5825  5.9083

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1.293e+02  5.072e+01  -2.550   0.0114 *
mydata$ANTM           -8.417e-01  1.316e-01  -6.394 8.32e-10 ***
mydata$CI              1.558e+00  2.263e-01   6.883 5.04e-11 ***
mydata$HUM             6.716e-01  2.690e-01   2.496   0.0132 *
sqANTM                -4.097e-04  2.047e-04  -2.002   0.0464 *
sqCI                  -5.967e-03  7.227e-04  -8.255 1.00e-14 ***
sqHUM                 -1.955e-04  4.410e-04  -0.443   0.6580
mydata$ANTM:mydata$CI  5.481e-03  6.107e-04   8.975  < 2e-16 ***
mydata$CI:mydata$HUM  -1.982e-03  7.957e-04  -2.490   0.0134 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.346 on 241 degrees of freedom
Multiple R-squared:  0.7553,    Adjusted R-squared:  0.7472
F-statistic: 93.01 on 8 and 241 DF,  p-value: < 2.2e-16
```

From the summary, we can see that the quadratic term sqHUM^2 is not significant and will be removed from the model. We also see that adding the quadratic terms have increased the R-Squared value.

Multicollinearity

Since we are using multiple predictor variables in this model, it is important to confirm none of the variables are correlated to one another.

```r
vif(lm(mydata$CNC ~ mydata$ANTM + mydata$CI + mydata$HUM,
                          data = mydata))
```

```
mydata$ANTM   mydata$CI  mydata$HUM
   2.079838    1.490767    2.369435
```

Based on the Variance Inflation factor, we see each predictor variable is less than 10. There is no multicollinearity.

Final Model

```r
FinalModel <- lm(mydata$CNC ~ mydata$ANTM + mydata$CI + mydata$HUM +
                         mydata$ANTM*mydata$CI + mydata$CI*mydata$HUM + sqANTM + sqCI,
                          data = mydata)
summary(FinalModel)
```

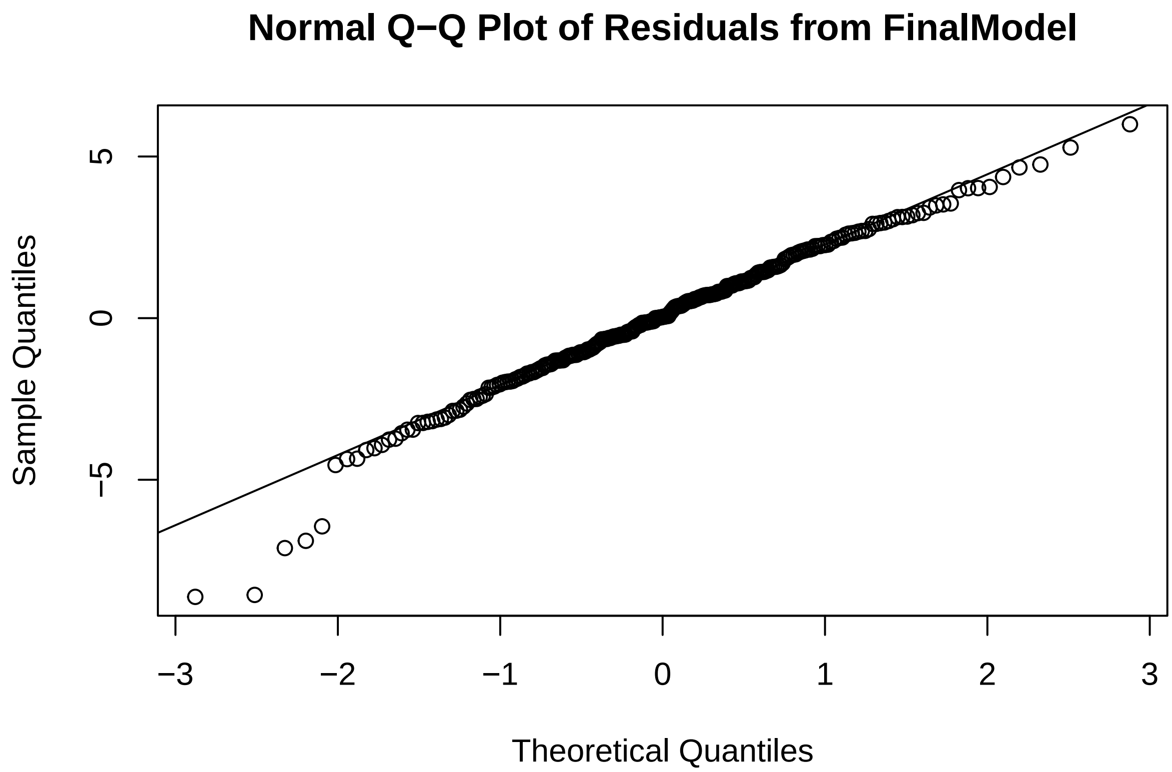
```
Call:
lm(formula = mydata$CNC ~ mydata$ANTM + mydata$CI + mydata$HUM +
    mydata$ANTM * mydata$CI + mydata$CI * mydata$HUM + sqANTM +
    sqCI, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6222 -1.3570  0.0363  1.5715  5.9970

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1.141e+02  3.721e+01  -3.066 0.002414 **
mydata$ANTM           -8.162e-01  1.182e-01  -6.907 4.35e-11 ***
mydata$CI              1.586e+00  2.170e-01   7.308 3.94e-12 ***
mydata$HUM             5.642e-01  1.167e-01   4.835 2.37e-06 ***
sqANTM                -4.764e-04  1.387e-04  -3.435 0.000698 ***
sqCI                  -5.867e-03  6.857e-04  -8.557 1.35e-15 ***
mydata$ANTM:mydata$CI  5.584e-03  5.639e-04   9.903  < 2e-16 ***
mydata$CI:mydata$HUM  -2.245e-03  5.281e-04  -4.251 3.04e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.342 on 242 degrees of freedom
Multiple R-squared:  0.7551,    Adjusted R-squared:  0.7481
F-statistic: 106.6 on 7 and 242 DF,  p-value: < 2.2e-16
```

Our final model is :

$Y = -114.1 - 0.8162(X_1, ANTM) + 1.586(X_2, CI) + 0.5642(X_3, HUM) - 0.0004764(X_4, ANTM^2) - 0.005867(X_5, CI^2) + 0.005584(X_6, ANTMxCI) - 0.002245(X_7, CIxHUM)$

The final model has an R-Square Adjusted = 0.7481 which means that 74.81% of the variation in Centene's daily closing price are explained by the predictor variables. The R-Squared Adjusted is greater than the acceptable value for R-Squared Adjusted of 50%. The F-Statistic = 106.6 and P-Value = <0.0001. I also noticed when we removed the non-significant interaction and quadratic term, our overall model and the independent variables t-test statistics have improved, so it is safe to say we have found our best prediction model.

```r
resids <- FinalModel$residuals
qqnorm(resids, main="Normal Q-Q Plot of Residuals from FinalModel")
qqline(resids)
```



As we can see, the Normal Q-Q plot shows the residuals are normally distibuted with majority of data points on the line with a few outliers on both ends of the line.

Conclusion

I can confidently conclude that we have developed the best model for predicting Cenetene's daily closing price using predictor variables: ANTM, CI, HUM, ANTMˆ2, CIˆ2, ANTMxCI, CIxHUM.

Dataset Citation

"Market Activity Market Activity -> Stocks Options Etfs Mutual Funds Indexes Commodities Cryptocurrency Currencies Futures Fixed Income Global Markets Quick Links Real-Time Quotes after-Hours Quotes Pre-Market Quotes NASDAQ-100 Symbol Screener Online Brokers Glossary Sustainable Bond Network Symbol Change History IPO Performance Ownership Search Dividend History Investing Lists Rulebooks & Regulations Fundinsight Market Events Economic Calendar Earnings IPO Calendar Dividend Calendar Spo Calendar Holiday Schedule Analyst Activity Analyst Recommendations Daily Earnings Surprise Forecast Changes Commodities -> Gold Copper Crude Oil Natural Gas Nasdaq Data Statistical Milestones Total Returns Daily Market Statistics Most Active See All Market Activity ->." Nasdaq, www.nasdaq.com/market-activity/stocks/cnc/historical.