BAN 602
Quantitative Fundamentals for Analytics

Case Assignment-2

Team Members

Aarthi Sivashankar
Nuzhat Fatima
Yukti Richharia
Dharmi Kanth Cigiri

Introduction

- CSU sells game-day magazines at all home football games during the fall semester.
- **Objective:** Analyze past 9 years of sales data to predict magazine sales for the upcoming season (Year 10).
- **Variables to be predicted**: Magazine Sales, Kickoff Temperature, Game Day Weather.
- **Financials:**
  - Magazine sold at $30 each.
  - Purchased at $10.
  - Unsold magazines disposed of at $5 each.
- **Goal:** Forecast magazine sales and optimize order quantity before the season starts.

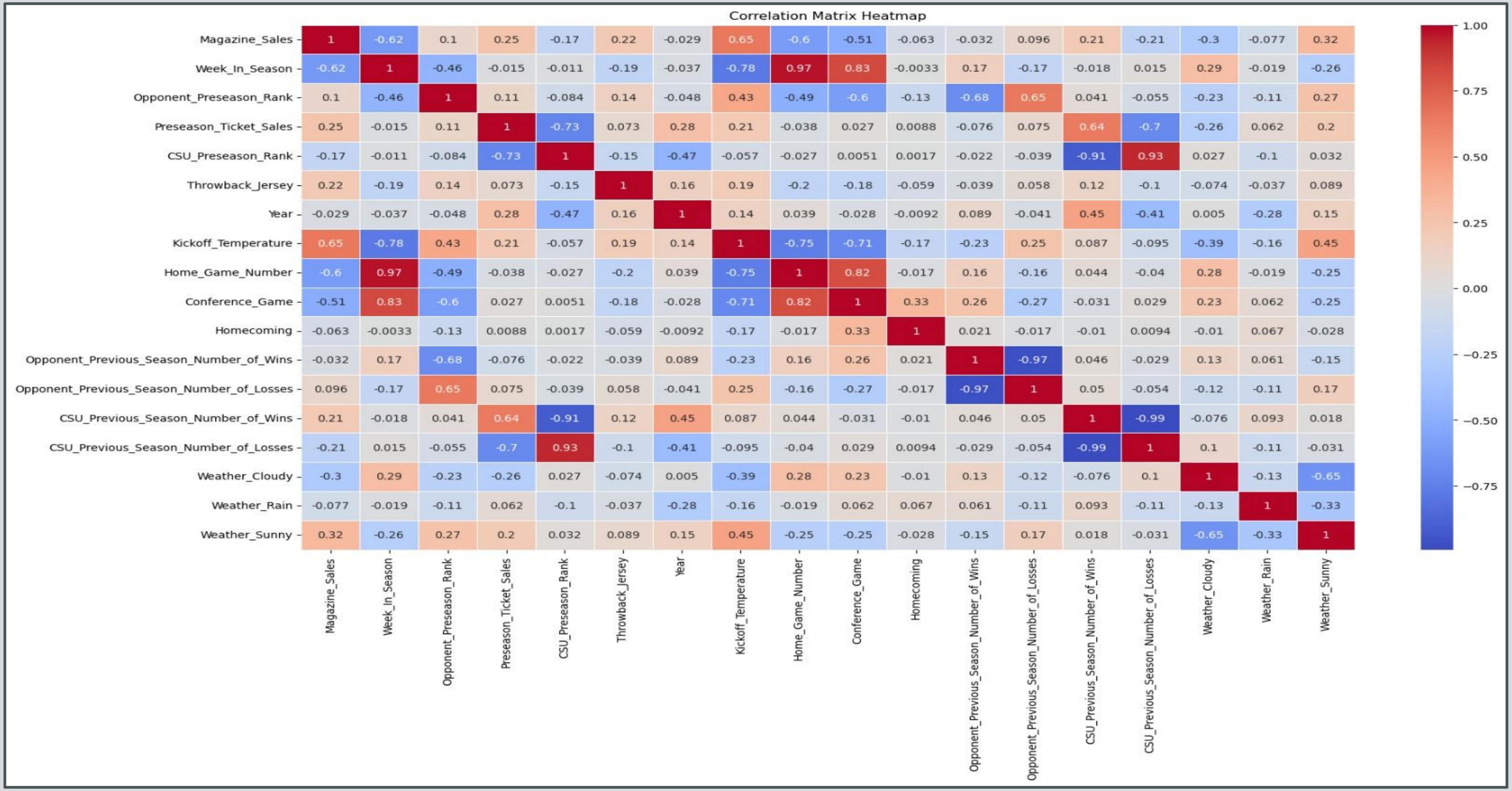
About the Data

- We have derived the data from:
  - https://csueb.instructure.com/courses/44065/files/4762252?wrap=1
- **Response Variable:**
  - Magazine Sales
- **Predictor variables:**
  - Week in season,Opponent Preseason Rank, Preseason Ticket Sales, CSU Preseason Rank, Throwback Jersey, Year, Kickoff Temperature, Home Game Number, Conference Game, Homecoming, Game Day Weather, Opponent’s Previous Season Number of Wins, Opponent’s Previous Season Number of Losses, CSU’s Previous Season Number of Wins, CSU’s Previous Season Number of Losses
- **Experimental Units:**
  - 952 Observations
- **Test and train data:**
  - **Test data set:** Year 1 to Year 8 data
  - **Train data set:** Year 9 data
- **Predicted data:** Year 10 data

Data Dictionary

Opponent - CSU's oponnent
Magazine Sales (Units) - The dependent variable, how many magazies were sold
Year - The year of the sales
Week In Season - The week of the football season for the sales
Opponent Preseason Rank - The preseason polling rank of the opponent
Preseason Ticket Sales - The number of tickes sold for that year's season
Total Game Attendance - The number of fans who attended that game
CSU Preseason Rank - The preseason polling rank of CSU
Home Game Number - The index number for home games
Conference Game (1 = Yes; 0 = No) - Dummy variable indicating whether the game is against a conference oppoent
Homecoming (1 = Yes; 0 = No) - Dummy variable indicating whether the game is the homecoming game
Game Day Weather - Sunny, Rain, or Cloudy
Sunny - Dummy variable indicating Sunny weather for that game
Rain - Dummy variable indicating Rainy weather for that game
Kickoff Temperature - temperature observed at the beginning of the game
Opponent's Previous Season Number of Wins - Number of wins opponent had in most recent season
Opponent's Previous Season Number of Losses - Number of losses opponent had in most recent season
CSU's Previous Season Number of Wins - Number of wins CSU had in most recent season
CSU's Previous Season Number of Losses - Number of losses CSU had in most recent season

Correlation Matrix



Outliers in the Data

Index	Opponent	Magazine_Sales	Week_In_Season	Opponent_Preseason_Rank	Preseason_Ticket_Sales	CSU_Preseason_Rank	Throwback_Jersey	Year
13	Lincoln University	6463	1	6	44211	73	0	3

Kickoff_Temperature	Home_Game_Number	Conference_Game	Homecoming	Game_Day_Weather	Opponent_Previous_Season_Number_of_Wins
90	1	0	0	Sunny	9

|Studentized residual| > 3: Often considered as a rule of thumb for identifying severe outliers.
|Studentized residual| between 2 and 3: Indicates potential outliers but might not always be problematic, especially with larger datasets.

Kickoff-Temperature Prediction

Kickoff-Temperature Prediction Models

Model 1 - Model with **all** the predictors in the data (Excluding Game_Day_Weather and Magazine Sales)

R-squared:	0.750
Adj. R-squared:	0.613

- P-Value for all the predictors is greater than 0.05

Model 2 - '**Home_Game_Number**' ,'**Conference_Game**' - Chosen based on Correlation Matrix

R-squared:	0.600
Adj. R-squared:	0.583

	coef	std err	t	P>|t|
Intercept	82.2173	3.458	23.774	0.000
Home_Game_Number	-3.7481	1.405	-2.668	0.011
Conference_Game	-12.8827	5.273	-2.443	0.018

Model 3 - '**Week_In_Season**' - Chosen Based on Correlation Matrix

R-squared:	0.588
Adj. R-squared:	0.580

	coef	std err	t	P>|t|
Intercept	82.3311	2.803	29.372	0.000
Week_In_Season	-3.3244	0.382	-8.693	0.000

Cross Validation - 4-Fold

```
Results for all models:
Model: Week_In_Season + Opponent_Preseason_Rank + Preseason_Ticket_Sales + CSU_Preseason_Rank + Throwback_Jers
Average MSE: 171.1617, Standard Error: 29.3317

Model: Week_In_Season
Average MSE: 113.8233, Standard Error: 14.9704

Model: Home_Game_Number + Conference_Game
Average MSE: 115.9441, Standard Error: 8.5762

Best Model (Lowest MSE): Week_In_Season with MSE: 113.8233
```

Kickoff Temperature Prediction - Year 10

```
Index
57    79.597267
58    72.798308
59    65.999349
60    62.599870
61    52.401431
62    49.001952
63    45.602472
Name: Kickoff_Temperature, dtype: float64
```

Game Day Weather Prediction

Model Selection and Prediction

Logistic Regression - Label-Encoding:

- Sunny - 0
- Rainy - 1
- Cloudy -2

Model 1 - Model with **all** the predictors in the data (Excluding Magazine Sales)

```
Accuracy Score: 0.6666666666666666

Predicted Weather for Test Set:
['Sunny' 'Sunny' 'Sunny' 'Sunny' 'Cloudy' 'Rain']
```

Game Day Weather Prediction for Year 10

```
Predicted Weather for year 10
['Sunny' 'Sunny' 'Sunny' 'Sunny' 'Sunny' 'Sunny' 'Sunny']
```

Model 2 - **'Week_In_Season'** , **'Kickoff_Temperature'** , **'Year'**

```
Accuracy Score: 0.8333333333333334

Predicted Weather for Test Set:
['Sunny' 'Sunny' 'Sunny' 'Sunny' 'Sunny' 'Cloudy']
```

Magazine Sales Prediction

Subset Selection

Best Subset Selection

Best subset selection involves fitting separate least squares regression models for every possible combination of p predictors. This includes all models with one predictor, two predictors, and so on. The goal is to compare and evaluate all resulting models to identify the best one.

With 2^p potential combinations, selecting the optimal model can be computationally challenging. So if $p = 10$, then there are approximately 1,000 possible models to be considered, and if $p = 20$, then there are over one million possibilities!

Stepwise Selection

Forward stepwise selection is a feature selection technique used in machine learning and statistics to build a model by adding one feature at a time. The goal is to improve the model's predictive performance while keeping it as simple as possible by including only the most important features.

This selection method will generate **p-squared** models. P is the number of independent variables in the dataset. The threshold used for selection is p-values.

```
Best predictors: ['Kickoff_Temperature']
                            OLS Regression Results
==============================================================================
Dep. Variable:          Magazine_Sales   R-squared:                       0.424
Model:                             OLS   Adj. R-squared:                  0.414
Method:                  Least Squares   F-statistic:                     39.82
Date:                 Mon, 30 Sep 2024   Prob (F-statistic):           5.40e-08
Time:                         14:49:08   Log-Likelihood:                -449.73
No. Observations:                   56   AIC:                             903.5
Df Residuals:                       54   BIC:                             907.5
Df Model:                            1
Covariance Type:             nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept             374.4434    398.456      0.940      0.352    -424.413    1173.299
Kickoff_Temperature    39.3821      6.241      6.310      0.000      26.870      51.895
==============================================================================
Omnibus:                        11.959   Durbin-Watson:                   1.884
Prob(Omnibus):                   0.003   Jarque-Bera (JB):               12.248
Skew:                            0.996   Prob(JB):                      0.00219
Kurtosis:                        4.132   Cond. No.                         251.
==============================================================================
```

Backward Selection is a feature selection method that starts with all available predictors in a model and iteratively removes the least significant features based on p-values until only statistically significant predictors remain. This technique helps to simplify models and improve interpretability while maintaining predictive performance.
This selection method searches through 1+p(p+1)/2 models.

```
Best R-squared: 0.6326
Best predictors: ['Week_In_Season', 'Opponent_Preseason_Rank', 'Preseason_Ticket_Sales', 'Year', 'Kickoff_Temperature', 'Opponent_Previous_Season_Wins',
'Opponent_Previous_Season_Losses']
                            OLS Regression Results
==============================================================================
Dep. Variable:        Magazine_Sales   R-squared:                       0.633
Model:                           OLS   Adj. R-squared:                  0.579
Method:                Least Squares   F-statistic:                     11.81
Date:               Mon, 30 Sep 2024   Prob (F-statistic):           1.25e-08
Time:                       14:49:42   Log-Likelihood:                -437.16
No. Observations:                 56   AIC:                             890.3
Df Residuals:                     48   BIC:                             906.5
Df Model:                          7
Covariance Type:           nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Intercept                       -3326.7893   1996.064     -1.667      0.102   -7340.146     686.567
Week_In_Season                   -130.3665     40.280     -3.236      0.002    -211.355     -49.378
Opponent_Preseason_Rank            -9.2078      3.771     -2.442      0.018     -16.790      -1.625
Preseason_Ticket_Sales              0.0570      0.021      2.761      0.008       0.015       0.098
Year                              -86.7381     36.203     -2.396      0.021    -159.529     -13.947
Kickoff_Temperature                20.8580      9.093      2.294      0.026       2.576      39.140
Opponent_Previous_Season_Wins     291.5108    134.661      2.165      0.035      20.756     562.266
Opponent_Previous_Season_Losses   396.3917    158.299      2.504      0.016      78.109     714.674
```

	feature	VIF
0	Intercept	541.386136
1	Week_In_Season	3.025491
2	Opponent_Preseason_Rank	2.507084
3	Preseason_Ticket_Sales	1.203735
4	Year	1.160272
5	Kickoff_Temperature	2.955971
6	Opponent_Previous_Season_Wins	21.333902
7	Opponent_Previous_Season_Losses	19.395252

Choosing the Optimal Model

Indirect way to estimate the test error-Unlike AIC, and BIC, for which a small value indicates a model with a low test error, a large value of adjusted R2 indicates a model with a small test error.

Models	AIC	BIC	Adjusted R-Squared
Forward	903	907	0.41
Backward	890	906	0.579
'Opponent_Preseason_Rank + Kickoff_Temperature+ CSU_Preseason_Rank+ Year'	896	906	0.510
‘Opponent_Preseason_Rank + Week_In_Season+ CSU_Preseason_Rank+ Year’	901	911	0.464
All Predictors	895	919	0.565

Cross Validation

Directly estimating for the test errors in our models.

```
Model: Week_In_Season+Opponent_Preseason_Rank+Preseason_Ticket_Sales+CSU_Preseason_Rank+Throwback_Jersey+Year+Kickoff_Temperature+Homecoming+Opponent_Pr
evious_Season_Wins+Opponent_Previous_Season_Losses+CSU_Previous_Season_Wins
Average MSE: 585897.4131

Model: Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Year
Average MSE: 568218.8713

Model: Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Year+Week_In_Season
Average MSE: 495904.0281

Model: Week_In_Season+Opponent_Preseason_Rank+Preseason_Ticket_Sales+Year+Kickoff_Temperature+Opponent_Previous_Season_Wins+Opponent_Previous_Season_Los
ses
Average MSE: 495301.9240

Model: Kickoff_Temperature+CSU_Preseason_Rank+Home_Game_Number
Average MSE: 600665.8879

Model: Week_In_Season+Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Home_Game_Number+Throwback_Jersey+Homecoming+Game_Day_Weather+Oppon
ent_Previous_Season_Losses
Average MSE: 635311.5536

Model: Opponent_Preseason_Rank+CSU_Preseason_Rank+Kickoff_Temperature
Average MSE: 631972.0912

Model: Opponent_Preseason_Rank+Week_In_Season+CSU_Preseason_Rank+Year
Average MSE: 578293.9385
```

From this, we can observe the lowest Avg. MSE is for the fourth model which we created using backward selection.

Variability in Selection:
The choice of the "best" model could vary depending on how the training and validation sets are split, or the folds in cross-validation. This means the specific model chosen as "best" may change with different splits of the data.

One-Standard-Error Rule: To address this variability, the *one-standard-error rule* is applied. The idea is to:

- Compute the test error for each model size (number of predictors).
- Calculate the standard error of the test error for each model size.
- Select the smallest model (fewest predictors) whose test error is within one standard error of the lowest point on the error curve.

5-Fold Cross Validation

```
Results for all models:
Model: Week_In_Season+Opponent_Preseason_Rank+Preseason_Ticket_Sales+CSU_Preseason_Rank+Throwback_Jersey+Year+Kickoff_Temperature+Homecoming+Opponent_Pr
evious_Season_Wins+Opponent_Previous_Season_Losses+CSU_Previous_Season_Wins
Average MSE: 580537.0610, Standard Error: 122725.1831

Model: Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Year
Average MSE: 520869.5248, Standard Error: 81319.9939

Model: Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Year+Week_In_Season
Average MSE: 460224.0382, Standard Error: 72740.5563

Model: Week_In_Season+Opponent_Preseason_Rank+Preseason_Ticket_Sales+Year+Kickoff_Temperature+Opponent_Previous_Season_Wins+Opponent_Previous_Season_Los
ses
Average MSE: 468888.3329, Standard Error: 82024.8089

Model: Kickoff_Temperature+CSU_Preseason_Rank+Home_Game_Number
Average MSE: 580510.2448, Standard Error: 109693.8551

Model: Week_In_Season+Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Home_Game_Number+Throwback_Jersey+Homecoming+Game_Day_Weather+Oppon
ent_Previous_Season_Losses
Average MSE: 537600.3674, Standard Error: 118374.3367

Model: Opponent_Preseason_Rank+Week_In_Season+CSU_Preseason_Rank+Year
Average MSE: 546591.5202, Standard Error: 122688.7978

Best Model (Lowest MSE): Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Year+Week_In_Season with MSE: 460224.0382
Best Model (One-Standard-Error Rule): Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Year with MSE: 520869.5248
```

10-Fold Cross Validation

```
Results for all models:
Model: Week_In_Season+Opponent_Preseason_Rank+Preseason_Ticket_Sales+CSU_Preseason_Rank+Throwback_Jersey+Year+Kickoff_Temperature+Homecoming+Opponent_Pr
evious_Season_Wins+Opponent_Previous_Season_Losses+CSU_Previous_Season_Wins
Average MSE: 585897.4131, Standard Error: 105373.3877

Model: Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Year
Average MSE: 568218.8713, Standard Error: 103705.9038

Model: Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Year+Week_In_Season
Average MSE: 495904.0281, Standard Error: 104273.4534

Model: Week_In_Season+Opponent_Preseason_Rank+Preseason_Ticket_Sales+Year+Kickoff_Temperature+Opponent_Previous_Season_Wins+Opponent_Previous_Season_Los
ses
Average MSE: 495301.9240, Standard Error: 89378.7248

Model: Kickoff_Temperature+CSU_Preseason_Rank+Home_Game_Number
Average MSE: 600665.8879, Standard Error: 154138.9543

Model: Week_In_Season+Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Home_Game_Number+Throwback_Jersey+Homecoming+Game_Day_Weather+Oppon
ent_Previous_Season_Losses
Average MSE: 635311.5536, Standard Error: 120410.5335

Model: Opponent_Preseason_Rank+Week_In_Season+CSU_Preseason_Rank+Year
Average MSE: 578293.9385, Standard Error: 160050.8268

Best Model (Lowest MSE): Week_In_Season+Opponent_Preseason_Rank+Preseason_Ticket_Sales+Year+Kickoff_Temperature+Opponent_Previous_Season_Wins+Opponent_P
revious_Season_Losses with MSE: 495301.9240
Best Model (One-Standard-Error Rule): Opponent_Preseason_Rank+Kickoff_Temperature+CSU_Preseason_Rank+Year with MSE: 568218.8713
```

Collinearity of the Predictors

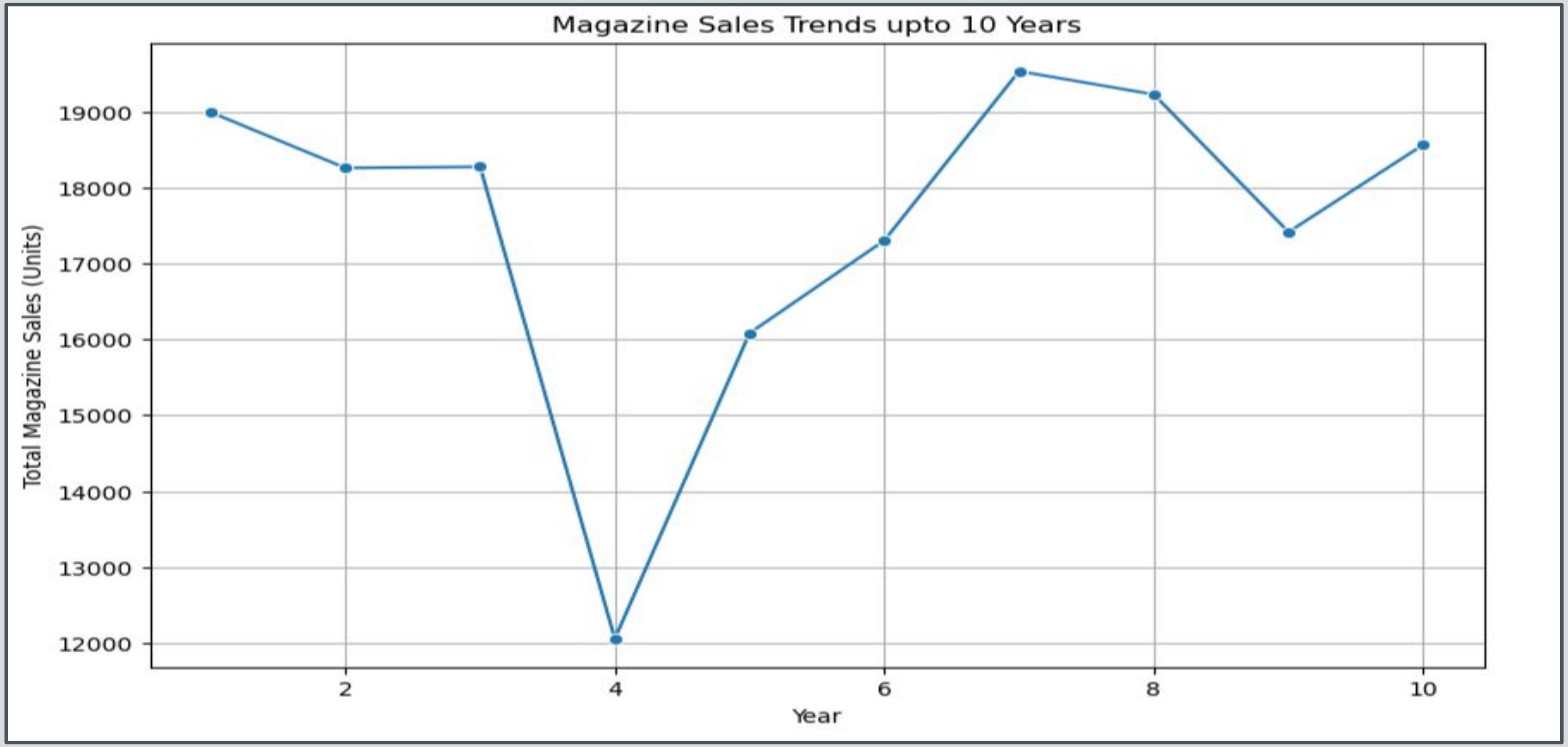
	feature	VIF
0	Intercept	28.097953
1	Opponent_Preseason_Rank	1.263905
2	CSU_Preseason_Rank	1.308712
3	Year	1.344575
4	Kickoff_Temperature	1.268036

VIF = 1: No multicollinearity. VIF between 1 and 5: Moderate correlation that is typically not problematic.
VIF > 5: High multicollinearity, which may affect the model's reliability.
VIF > 10: Extreme multicollinearity, and the model should be re-examined, perhaps by removing or combining predictors

Magazine Sales Prediction for the 10th Year

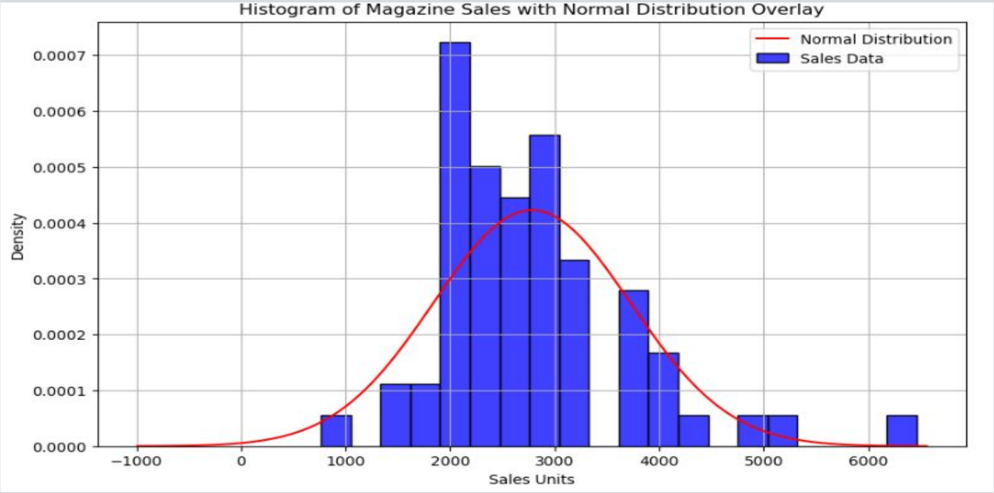
Index	Opponent	Magazine_Sales	Week_In_Season	Opponent_Preseason_Rank	Preseason_Ticket_Sales	CSU_Preseason_Rank	Throwback_Jersey	Year	Kickoff_Temperat
57	University of Missoula	2991.454523	1	120	54584	16	1	10	79.597
58	University of Ames	3211.606836	3	46	54584	16	0	10	72.798
59	Columbus University	3195.963172	5	4	54584	16	0	10	65.999
60	Indiana A&M	2841.815997	6	30	54584	16	0	10	62.599
61	DeKalb College	2162.542938	9	56	54584	16	0	10	52.401
62	Evanston University	1933.662377	10	65	54584	16	0	10	49.001
63	Madison University	1903.734672	11	47	54584	16	0	10	45.602

Magazine Sales Trend upto 10 years



We see an increase in the Magazine Sales for the 10th year as compared to previous year.

Magazine Sales - Order Quantity



We have the following data:

Predicted magazine sales for each game of Year 10 :
[2991,3211,3195,2841,2162,1933,1903]
Selling price (p): $30
Cost price (c): $10
Salvage price (s): $5

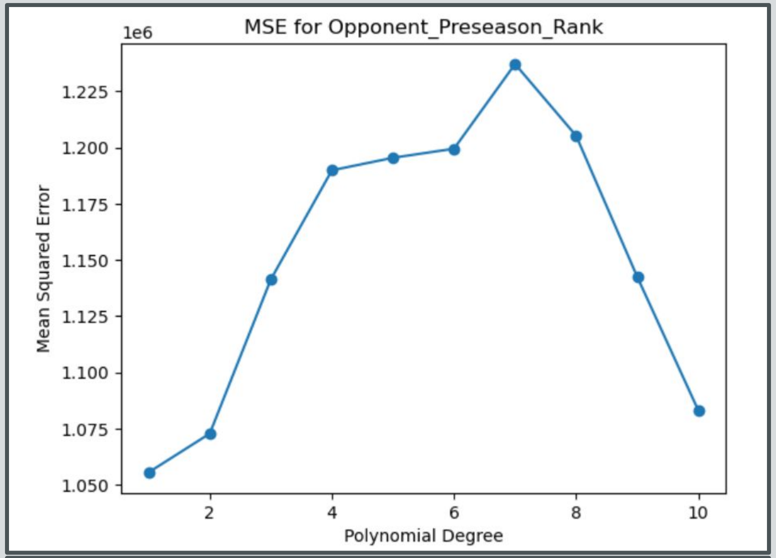
Critical Ratio: $\dfrac{\text{Cost of under-ordering}}{\text{Cost of over-ordering + Cost of under-ordering}}$ = 20/(20+5) = 0.8

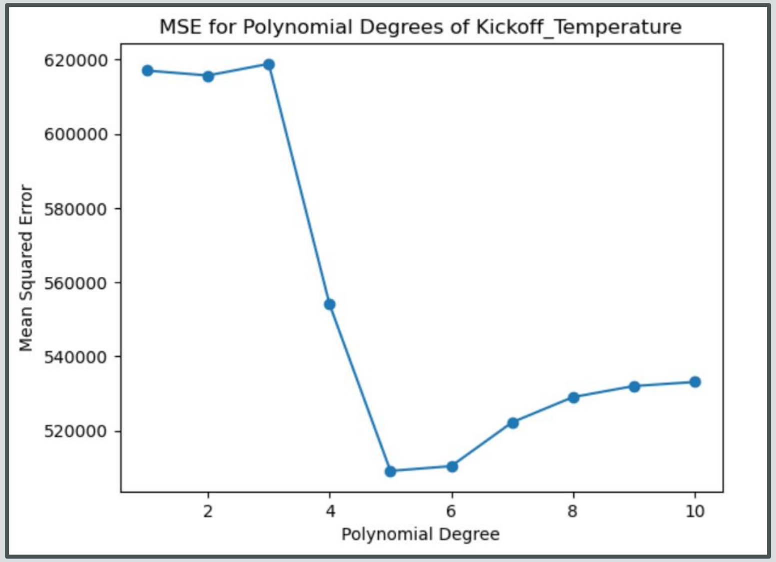
Z Score for the above Critical Ratio is **0.842**

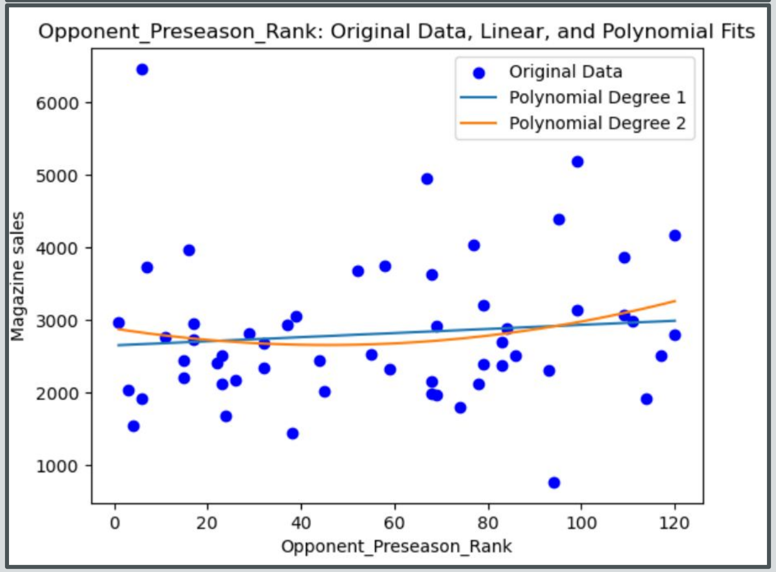
Order Quantity (Q): Total Predicted Sales + (Z Score* Std Dev) = 18236 + (0.842*542.568)

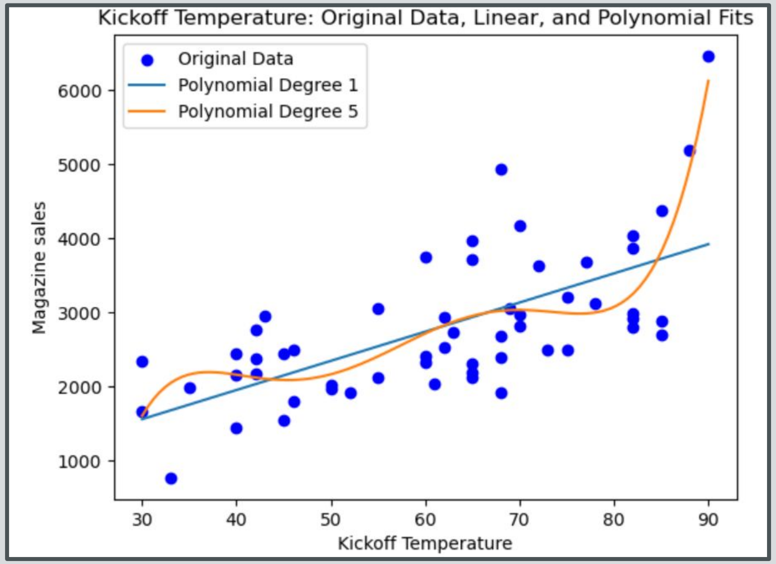
Order Quantity is **18,692** Units of Magazines

Non-linearity of the response-predictor relationships





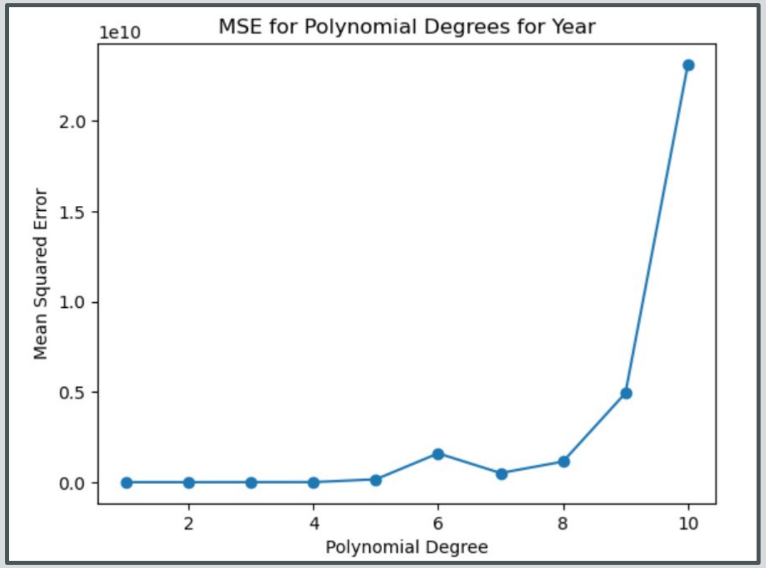


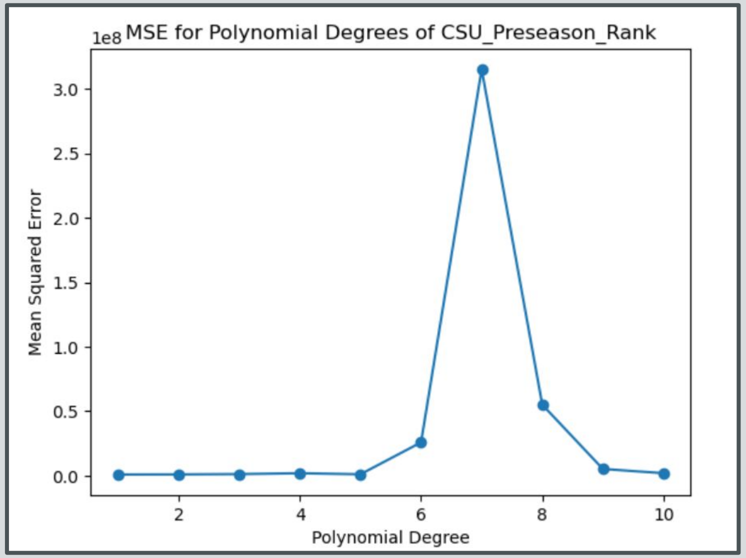


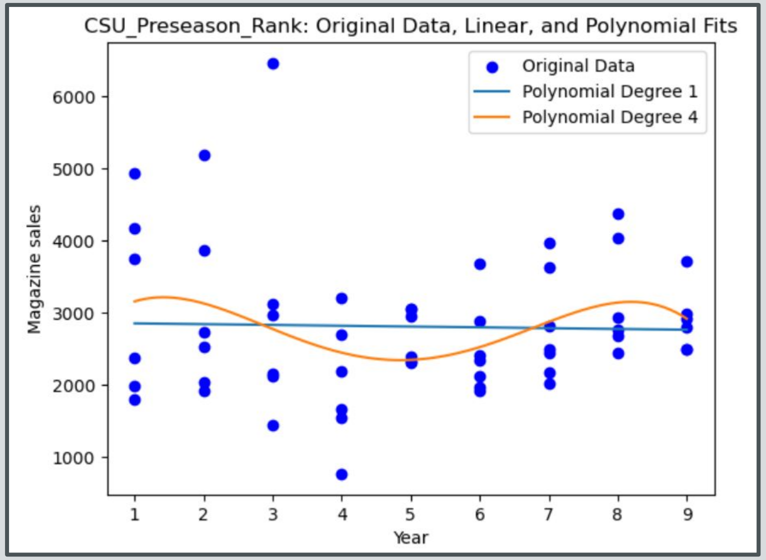
- From the Line graph, we can get the polynomial degree with lowest MSE for each Predictor Variable in our Model.

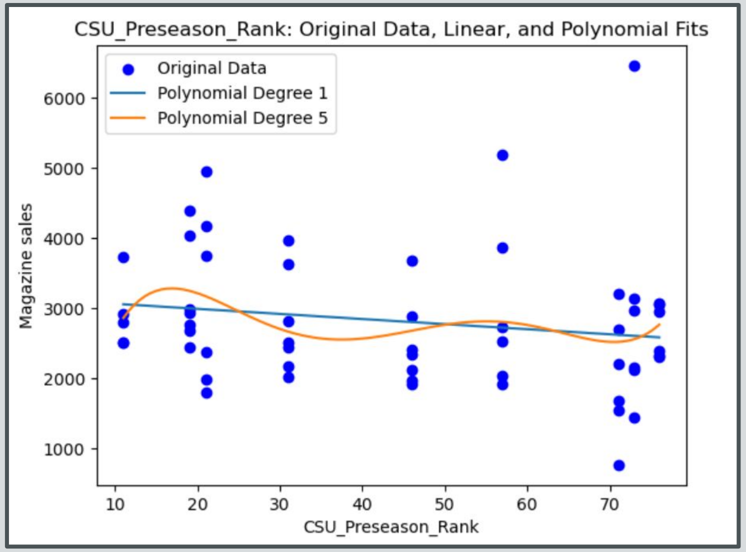
- The Scatter Plot shows that the pattern in the Original data is captured well using Polynomial model more than the linear model. This proves the Non Linearity in each predictor Variable.

Non-linearity of the response-predictor relationships

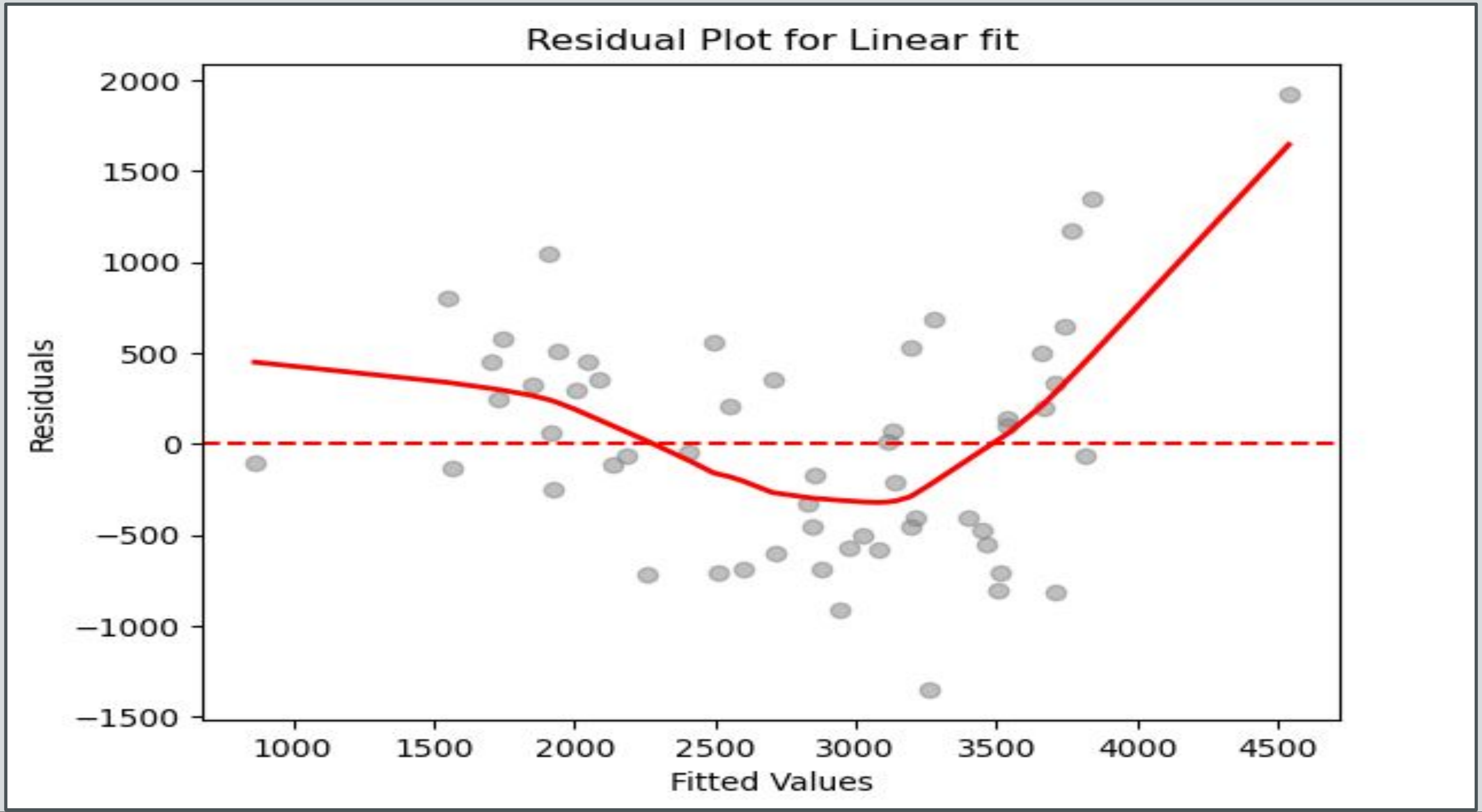




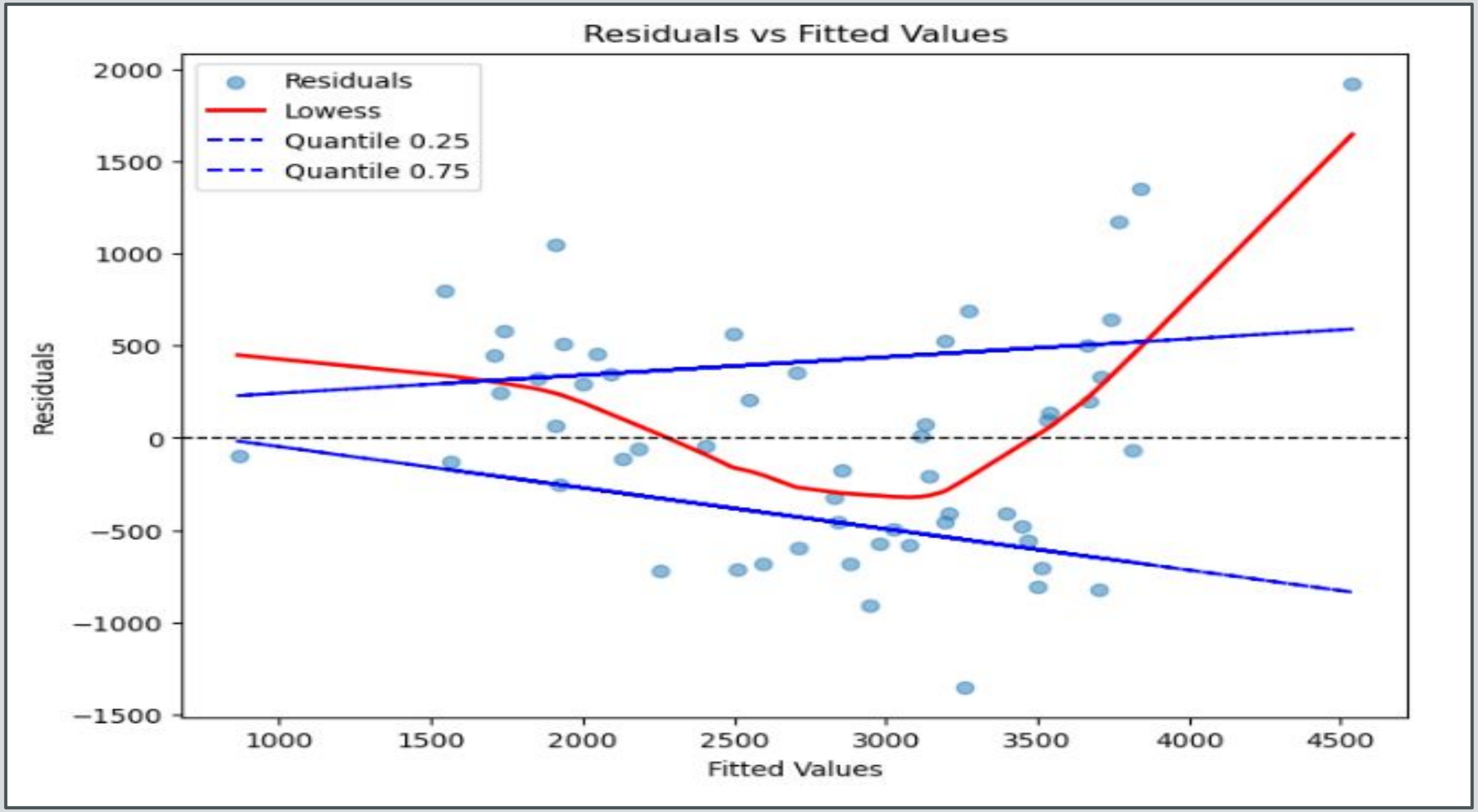




Non-Linearity of the Response-Predictor Relationships



Non-Constant Variance of error terms



Funnel Shape: The red Lowess line shows a clear curve, indicating that the variance of the residuals is not constant across all fitted values.
Expanding Residual Spread: As the fitted values increase, the spread of residuals also increases (especially on the right side of the plot).

Leverage Points

	Opponent_Preseason_Rank	Kickoff_Temperature	CSU_Preseason_Rank	Year	prediction	studentized_residuals	leverage
0	6	90.0	73	3	4420.191439	3.062033	0.210929
1	3	61.0	57	2	3358.297121	1.985033	0.101888
2	99	88.0	57	2	3942.047859	1.876592	0.104081
3	67	68.0	21	1	3764.246053	1.766873	0.143763
4	17	43.0	76	5	1834.075256	1.681690	0.086146
5	6	52.0	57	2	2905.811692	1.494154	0.083617
6	32	30.0	46	6	1351.160706	1.489698	0.096708
7	15	65.0	71	4	3071.617235	1.315489	0.071636
8	74	46.0	21	1	2660.626767	1.296018	0.149341

High leverage threshold=2×(p+1)/n
HLT for the model is (2(5+1))/56 LT= 0.2142

Thank You!
Questions?