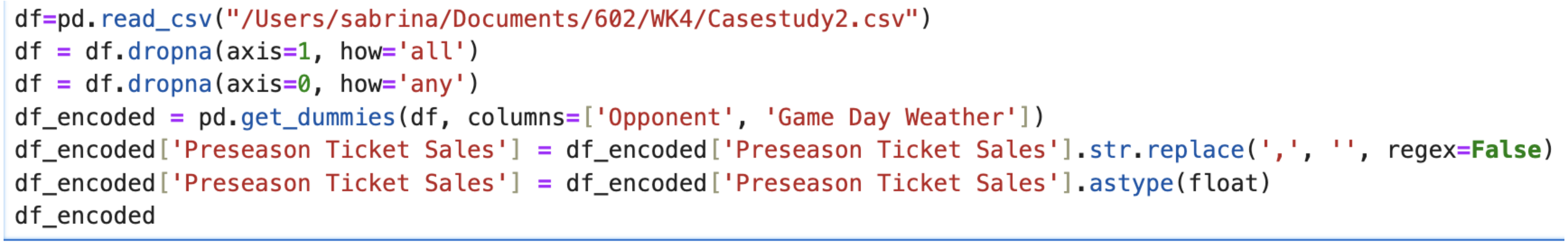
MAGAZINE SALES ANALYSIS

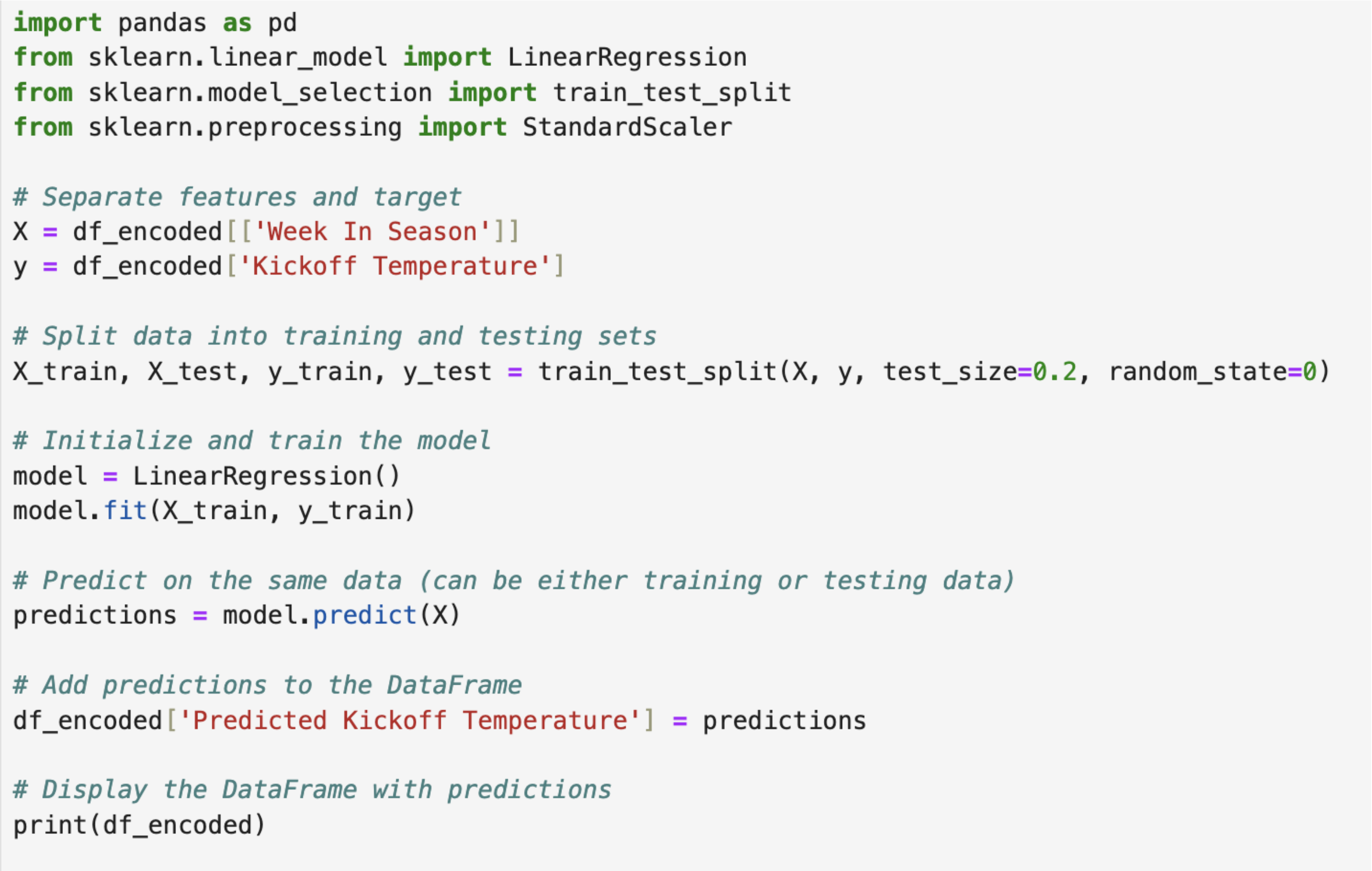
CLEAN DATA-

REMOVE MISSING VALUES
ENCODE QUALITATIVE DATA



```
df=pd.read_csv("/Users/sabrina/Documents/602/WK4/Casestudy2.csv")
df = df.dropna(axis=1, how='all')
df = df.dropna(axis=0, how='any')
df_encoded = pd.get_dummies(df, columns=['Opponent', 'Game Day Weather'])
df_encoded['Preseason Ticket Sales'] = df_encoded['Preseason Ticket Sales'].str.replace(',', '', regex=False)
df_encoded['Preseason Ticket Sales'] = df_encoded['Preseason Ticket Sales'].astype(float)
df_encoded
```

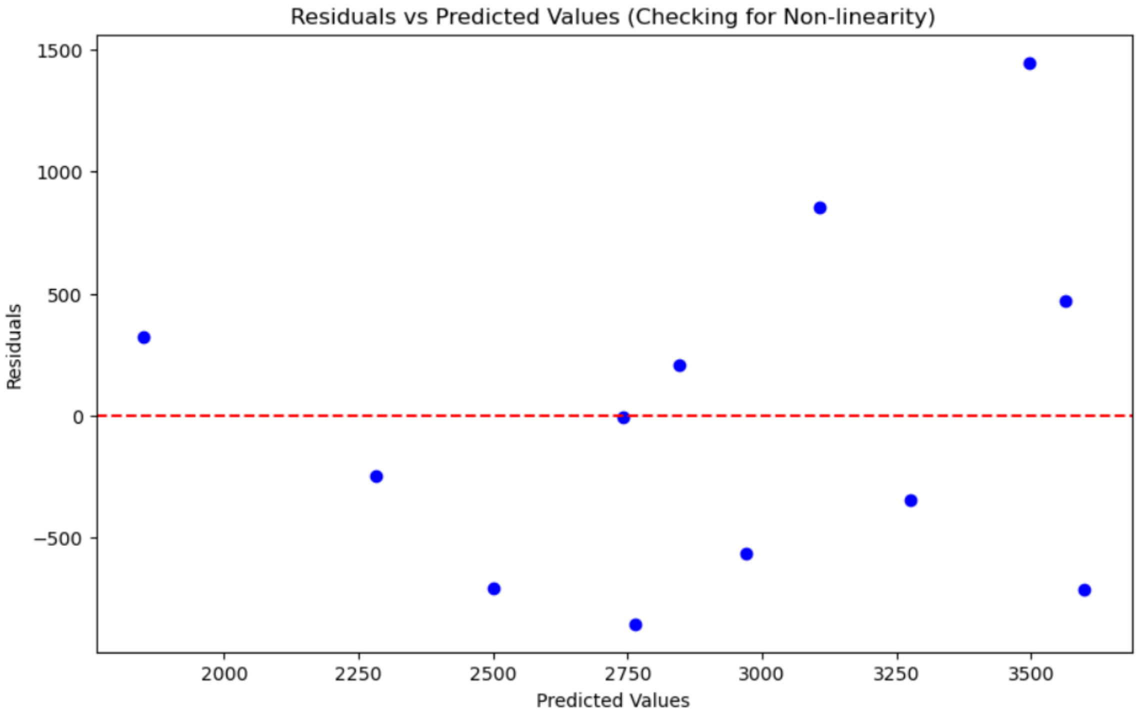
MODEL FIT



```python
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Separate features and target
X = df_encoded[['Week In Season']]
y = df_encoded['Kickoff Temperature']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)

# Initialize and train the model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict on the same data (can be either training or testing data)
predictions = model.predict(X)

# Add predictions to the DataFrame
df_encoded['Predicted Kickoff Temperature'] = predictions

# Display the DataFrame with predictions
print(df_encoded)
```

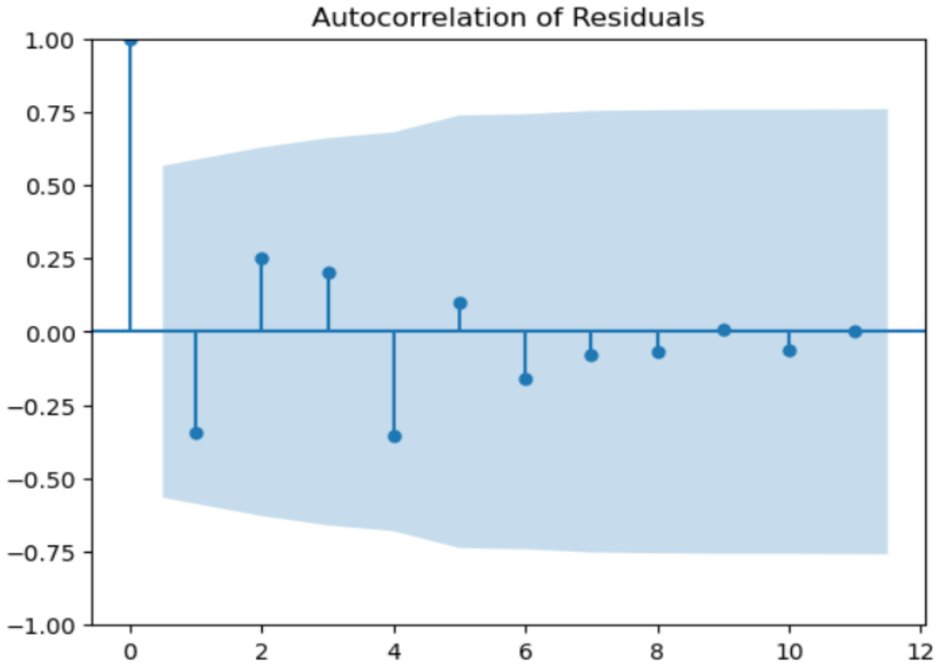
NON-LINEARITY ANALYSIS

- Pattern in Residuals: Residuals are mostly random but show some spread.
- Non-linear Pattern: No clear curved pattern, suggesting the linear assumption may hold. Larger residuals at the extremes might indicate underfitting.
- Variance of Residuals: Wider spread at higher values suggests potential heteroscedasticity.



CORRELATION OF ERROR TERMS.

- Durbin-Watson (2.64): Slight negative autocorrelation, but not severe.
- Autocorrelation Plot: Minimal autocorrelation, except at lag 1—no major concern.
- Model Performance:
  - GLS & HAC: MSE: 452733.99, R^2: 0.48—both models perform similarly, explaining 48% of the variance.
  - Lag Model: MSE: 1819242.55, R^2: 0.37—lagged variable didn't improve performance.
- Conclusion:
  - GLS/HAC are preferable.
  - Lagged model adds complexity without benefit.
  - Focus on refining predictors and fine-tuning.

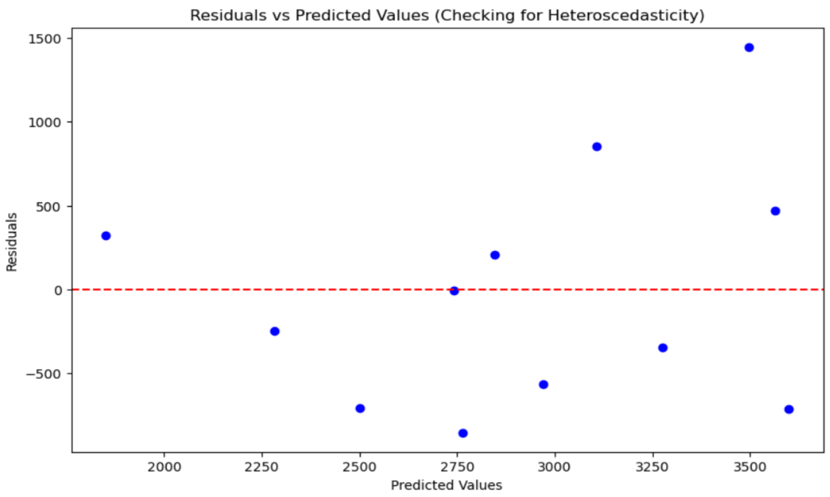


NON-CONSTANT VARIANCE OF ERROR TERMS

- No Clear Pattern: Residuals are mostly randomly scattered around zero with no discernible pattern, suggesting constant variance (homoscedasticity).

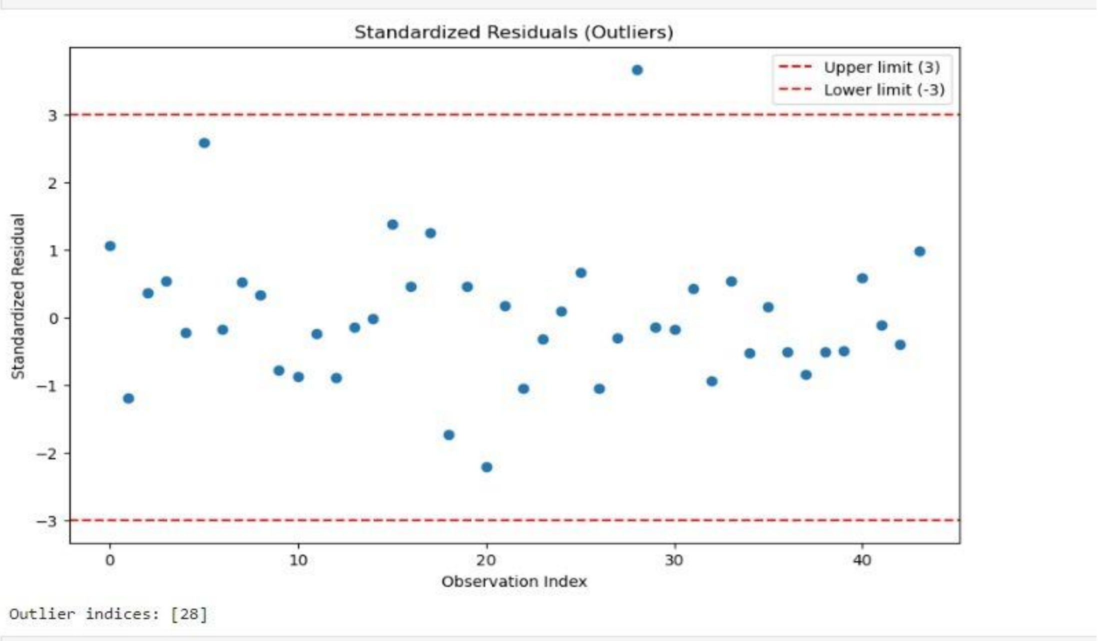
- Mild Variation: Slight spread in residuals at higher predicted values (3000-3500), but no strong cone shape, indicating no significant heteroscedasticity.

- Conclusion: No strong evidence of heteroscedasticity; the model likely satisfies the constant variance assumption.



OUTLIERS

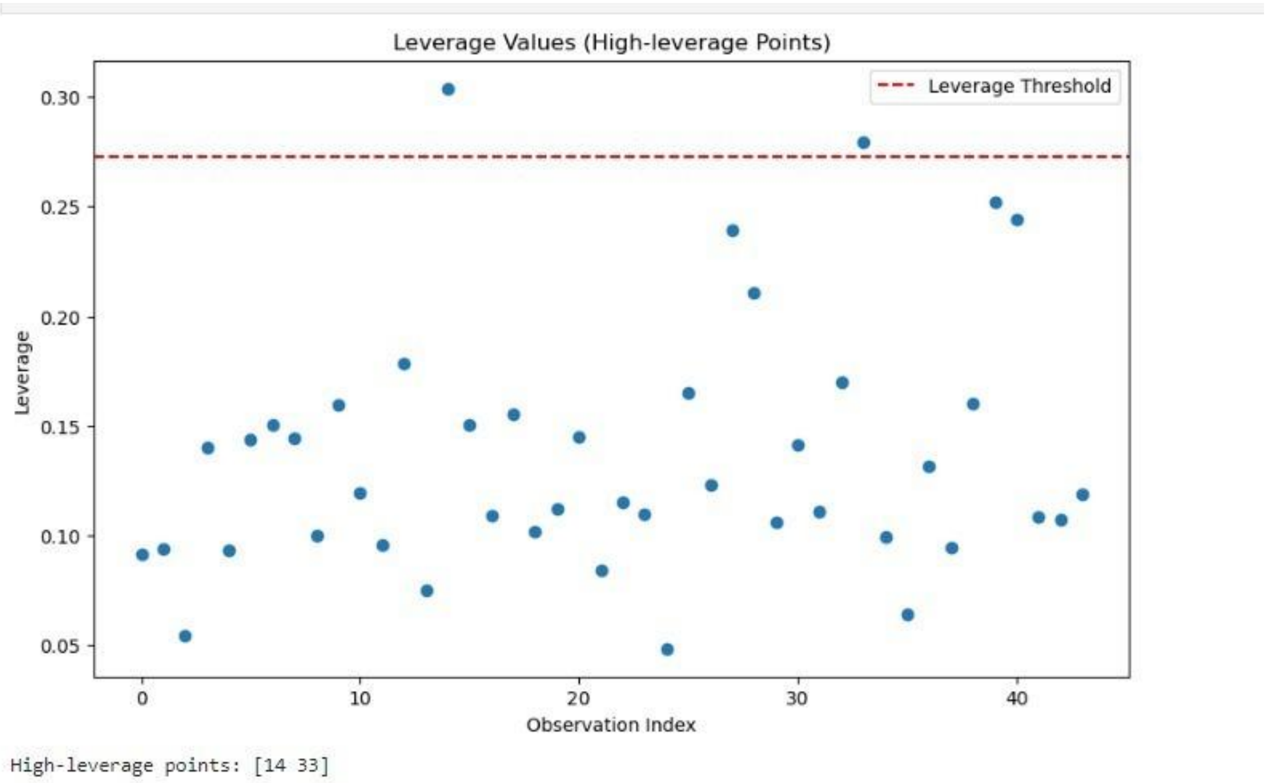
- Observation 28 behaves as an outlier and should be investigated for potential data entry errors or unusual events.
- Potential solutions: Remove the outlier if invalid or use robust regression methods (e.g., RANSAC, Huber) to minimize its influence.
- Impact: Compare model performance with and without the outlier to assess its influence; also check Cook’s distance to see if it’s a high-leverage point affecting the model.



```
Outlier indices: [28]
```

HIGH-LEVERAGE POINTS

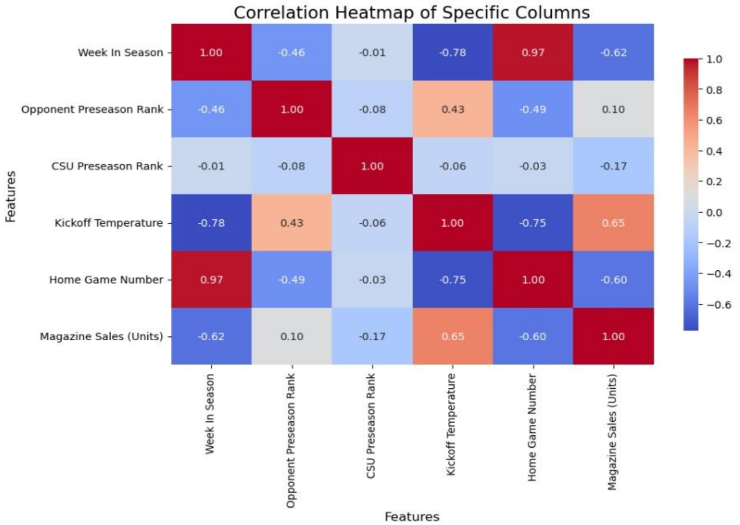
- High-Leverage Points: Observations 14 and 33 exceed this threshold, meaning they could disproportionately influence the regression model.



```
High-leverage points: [14 33]
```

COLLINEARITY

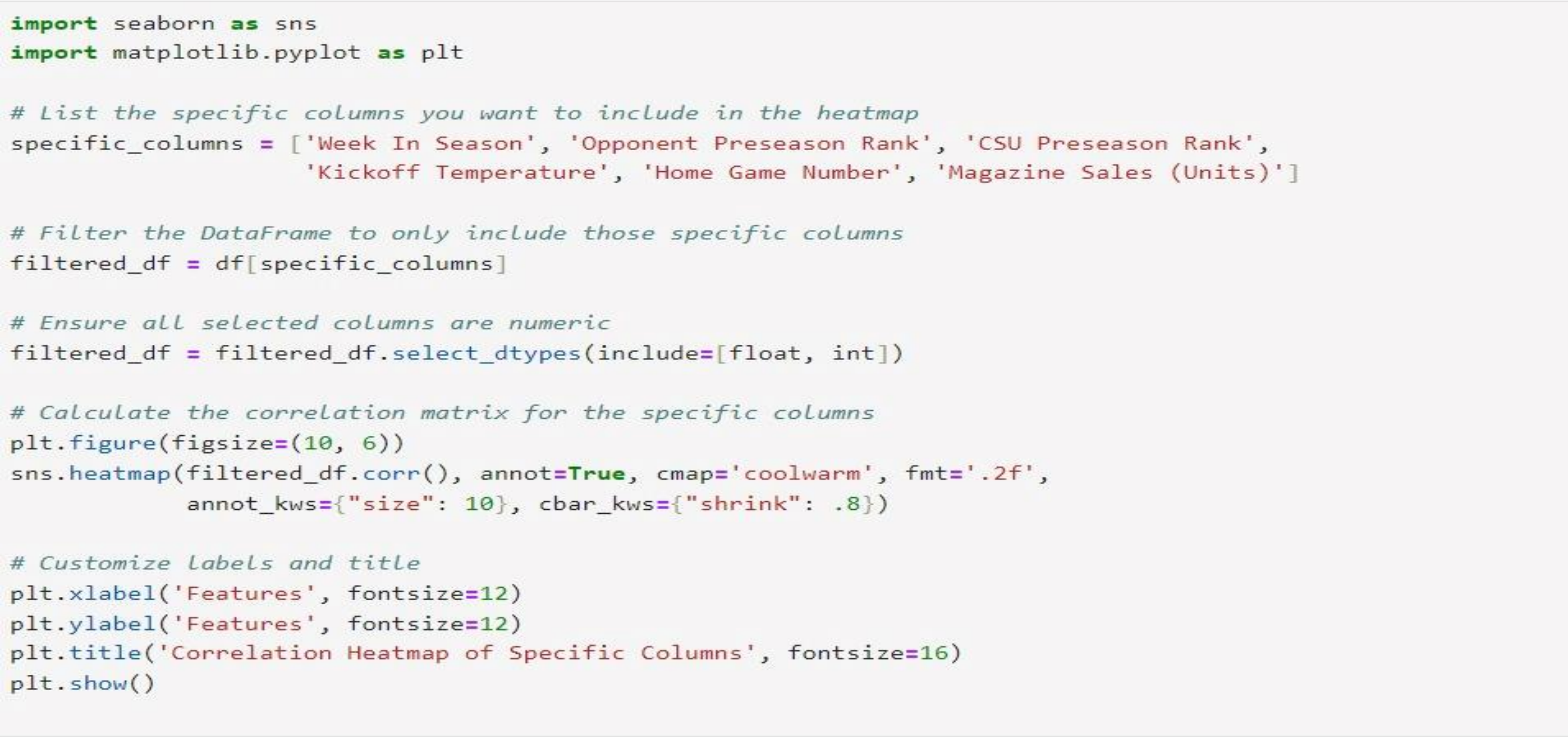
- Collinearity:
- Week In Season (VIF = 19.76) and Home Game Number (VIF = 20.32) show high multicollinearity.
- Other variables (VIF < 5) have no significant multicollinearity issues.
- Recommendations:
  - Remove or combine Week In Season and Home Game Number to reduce multicollinearity.
  - Rerun the model and evaluate performance changes.
  - Consider using PCA to retain variables while addressing multicollinearity.



Correlation Heatmap of Specific Columns

	Week In Season	Opponent Preseason Rank	CSU Preseason Rank	Kickoff Temperature	Home Game Number	Magazine Sales (Units)
Week In Season	1.00	-0.46	-0.01	-0.78	0.97	-0.62
Opponent Preseason Rank	-0.46	1.00	-0.08	0.43	-0.49	0.10
CSU Preseason Rank	-0.01	-0.08	1.00	-0.06	-0.03	-0.17
Kickoff Temperature	-0.78	0.43	-0.06	1.00	-0.75	0.65
Home Game Number	0.97	-0.49	-0.03	-0.75	1.00	-0.60
Magazine Sales (Units)	-0.62	0.10	-0.17	0.65	-0.60	1.00

CORRELATION ANALYSIS

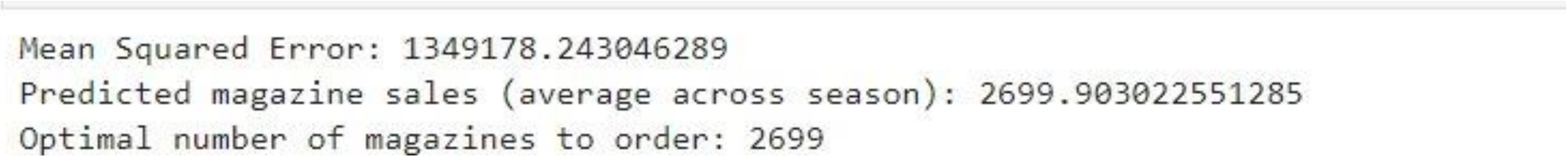


```python
import seaborn as sns
import matplotlib.pyplot as plt

# List the specific columns you want to include in the heatmap
specific_columns = ['Week In Season', 'Opponent Preseason Rank', 'CSU Preseason Rank',
                    'Kickoff Temperature', 'Home Game Number', 'Magazine Sales (Units)']

# Filter the DataFrame to only include those specific columns
filtered_df = df[specific_columns]

# Ensure all selected columns are numeric
filtered_df = filtered_df.select_dtypes(include=[float, int])

# Calculate the correlation matrix for the specific columns
plt.figure(figsize=(10, 6))
sns.heatmap(filtered_df.corr(), annot=True, cmap='coolwarm', fmt='.2f',
            annot_kws={"size": 10}, cbar_kws={"shrink": .8})

# Customize labels and title
plt.xlabel('Features', fontsize=12)
plt.ylabel('Features', fontsize=12)
plt.title('Correlation Heatmap of Specific Columns', fontsize=16)
plt.show()
```

PREDICTION RESULT

Assume that CSU sells each magazine for $30, buys it for $10, and can dispose of unsold magazines for $5 after the season. How many magazines should CSU order? Magazines can only be ordered before the season begins.

Answer:



```
Mean Squared Error: 1349178.243046289
Predicted magazine sales (average across season): 2699.903022551285
Optimal number of magazines to order: 2699
```