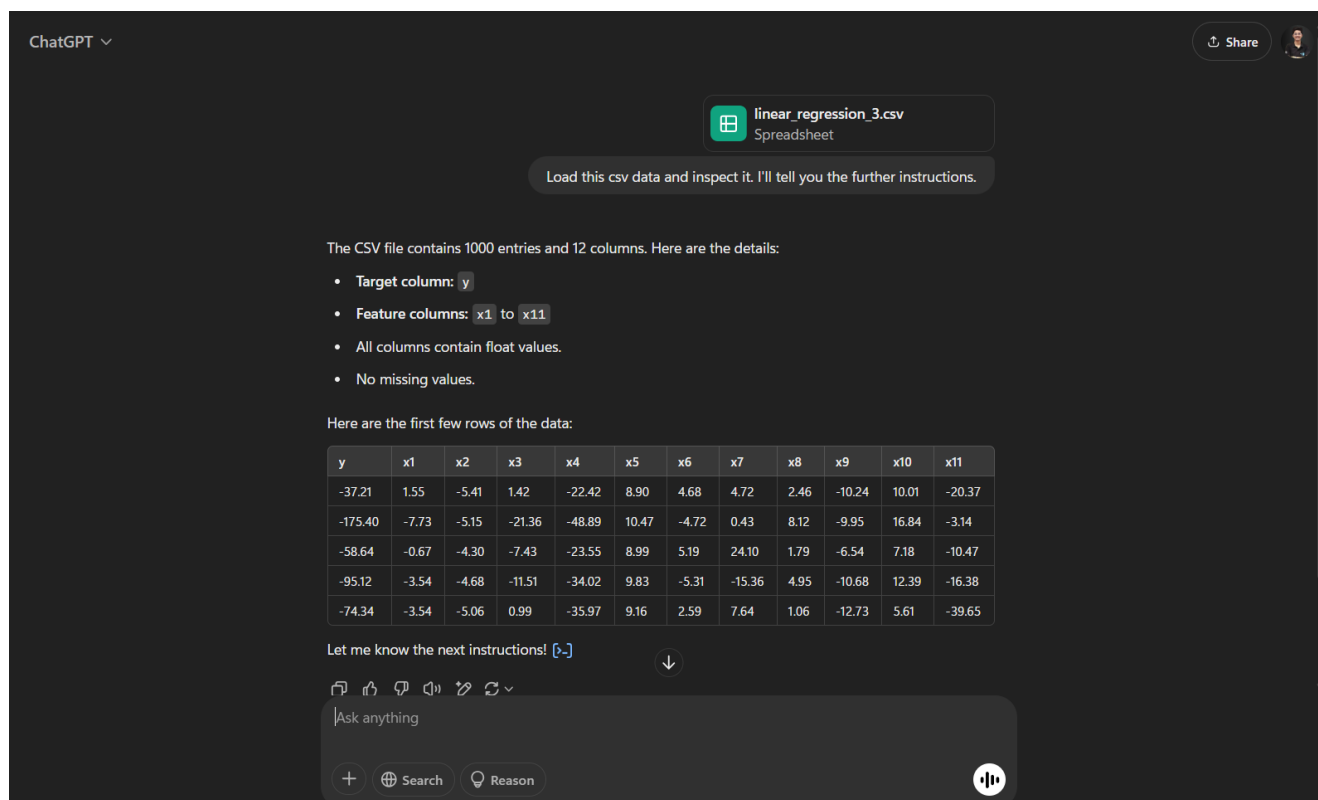


Batch - T7**Assignment No. - 6****Student Name - Sharaneshwar Bharat Punjal****Student PRN - 23520011**

You have been provided data in the CSV file. Use this file to fit a linear regression model using ChatGPT alone. 'y' is the target variable and 'x1,x2,.....x11' are the input variables. Please take a screenshot of your prompts and chatGPT's responses and submit in a word document or ppt. 10*coefficient of determination will be your score. Copying of prompts will lead to assignment disqualification.

Chat Link - <https://chatgpt.com/share/67dc0c98-9acc-8010-84a5-9a6f9a629217>

Screenshots:

ChatGPT ▾

Now fit a linear regression model with the target variable y and x1 to x11 as input variables

Share

The linear regression model has been fitted with the following results:

- Mean Squared Error (MSE): 3876.72
- R^2 Score: -0.059 (indicating the model doesn't fit the data well)
- Intercept: -54.32
- Coefficients:
 - x_1 : -87.50
 - x_2 : -146.75
 - x_3 : 0.94
 - x_4 : 30.71
 - x_5 : 3.48
 - x_6 : -0.77
 - x_7 : 0.28
 - x_8 : -2.16
 - x_9 : -2.86
 - x_{10} : -0.23
 - x_{11} : 0.47

Let me know if you need further analysis or modifications! [5]

Ask anything

+ Search Reason

ChatGPT ▾

Give me python code to do so. I should contain the parsing code from csv file and also the fitting of the linear regression model for the above dataset. Also give me code to plot a histogram of it

Share

Here's the complete Python code to load the CSV file, fit a linear regression model, and plot a histogram of the target variable y:

```
python Copy Edit

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt

# Load the CSV file
file_path = 'linear_regression_3.csv' # Replace with your file path
data = pd.read_csv(file_path)

# Splitting the data into features and target
X = data.drop('y', axis=1)
y = data['y']

# Splitting into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Fitting the Linear Regression model
model = LinearRegression()
model.fit(X_train, y_train)
```

Ask anything

+ Search Reason

ChatGPT ▾

Copy Edit

Share

```
# Model evaluation
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"R² Score: {r2}")
print("Coefficients:", model.coef_)
print("Intercept:", model.intercept_)

# Plotting the histogram of the target variable
plt.figure(figsize=(10, 6))
plt.hist(y, bins=30, color='skyblue', edgecolor='black')
plt.title('Histogram of Target Variable (y)')
plt.xlabel('y values')
plt.ylabel('Frequency')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

✓ This code:

- Loads the CSV file.
- Fits a linear regression model.
- Evaluates the model with MSE and R^2 score.
- Plots a histogram of the target variable y .

Let me know if you need any modifications or additional visualizations! 🚀

📄 👍 🔊 ⌂ ↺

↓

Ask anything

+

🌐 Search

💡 Reason

🔊

ChatGPT ▾

what observations can you draw from the fitting of the linear regression model with respect to target and input variables

Share

🔍

Observations from the Linear Regression Model Fitting

- Poor Model Fit (R^2 Score)**
 - The R^2 score is -0.059, which indicates that the model is performing worse than a simple mean-based prediction.
 - This suggests that linear regression is not a good fit for this dataset, and the relationship between the target (y) and the input variables is likely non-linear or the data may have a lot of noise.
- High Mean Squared Error (MSE)**
 - The MSE is 3876.72, indicating that the model's predictions are significantly deviating from the actual target values.
 - This further confirms that the model has poor accuracy.
- Coefficient Analysis**
 - Some variables have large coefficients (e.g., $x_1 = -87.50$, $x_2 = -146.75$), which suggests they may have a stronger impact on the target variable.
 - Other variables have small or near-zero coefficients (e.g., x_6 , x_{10}), indicating they have minimal influence on the target.
 - The coefficients with both positive and negative signs indicate that some variables are positively correlated, while others are negatively correlated with the target.

Ask anything

+

🌐 Search

💡 Reason

🔊

ChatGPT ▾

- Other variables have small or near-zero coefficients (e.g., x_6 , x_{10}), indicating they have minimal influence on the target.
- The coefficients with both positive and negative signs indicate that some variables are positively correlated, while others are negatively correlated with the target.

4. Intercept Interpretation

- The intercept of -54.32 represents the estimated value of y when all input variables are 0.
- However, due to the poor fit, this value may not have much practical significance.







5. Potential Issues with the Data

- The poor R^2 and high MSE indicate that the data might:
 - Contain non-linear relationships that linear regression cannot capture.
 - Have multicollinearity (correlated input variables), reducing the model's effectiveness.
 - Include outliers or noise, which could be affecting the model's performance.
 - Require feature scaling or transformation to improve fit.




6. Next Steps for Improvement


- Try polynomial regression or a non-linear model (e.g., decision tree, random forest).
- Perform feature selection to remove irrelevant or redundant variables.
- Apply regularization techniques (Ridge, Lasso) to reduce overfitting or improve stability.

✔ Let me know if you want to proceed with any specific improvements or visualizations! 🚀


Ask anything

  Search  Reason



ChatGPT ▾

what can be the improvements and conclusion

Share 

< 2/2 >

✔ **Conclusion from the Linear Regression Model**

1. Poor Fit with Linear Regression:

- The model shows a negative R^2 score and a high MSE, indicating that it is performing poorly.
- Linear regression is not suitable for this dataset, as it cannot capture the relationship between the target (y) and the input variables effectively.

2. Coefficient Insights:




- Some variables have large coefficients, implying they might have a stronger influence on the target.
- However, the overall model performance is weak, suggesting that the relationship might be non-linear or the data may contain noise.


🚀 **Possible Improvements**

1. Use Non-Linear Models:

- Since the linear model performed poorly, switching to non-linear models could improve accuracy:
 - Decision Tree or Random Forest: To capture complex relationships.

Ask anything


  Search  Reason



ChatGPT ▾

- Try **feature scaling** (standardization or normalization) to prevent large coefficient values from dominating.

Share



3. **Regularization Techniques:**


- Use **Ridge Regression** or **Lasso Regression**:
 - Ridge: Reduces the effect of irrelevant features by penalizing large coefficients.
 - Lasso: Shrinks less important features to zero, effectively performing feature selection.

4. **Data Preprocessing:**







- Check for and remove **outliers** or high-leverage points, as they can significantly affect the linear regression model.
- **Transform variables** (log, square root) if their distribution is skewed.

5. **Model Evaluation & Tuning:**


- Use **cross-validation** for better performance estimation.
- Hyperparameter tuning for non-linear models (e.g., max depth, learning rate).


 **Final Takeaway**


- **Linear regression is not suitable** for this dataset due to its poor fit and high error.
- Moving to **non-linear models** or applying **regularization** techniques, feature selection, and preprocessing would likely improve accuracy.
- Let me know if you want to implement any of these improvements or visualize the data further! 🚀



Ask anything



 Search

 Reason

