VIT®

Vellore Institute of Technology

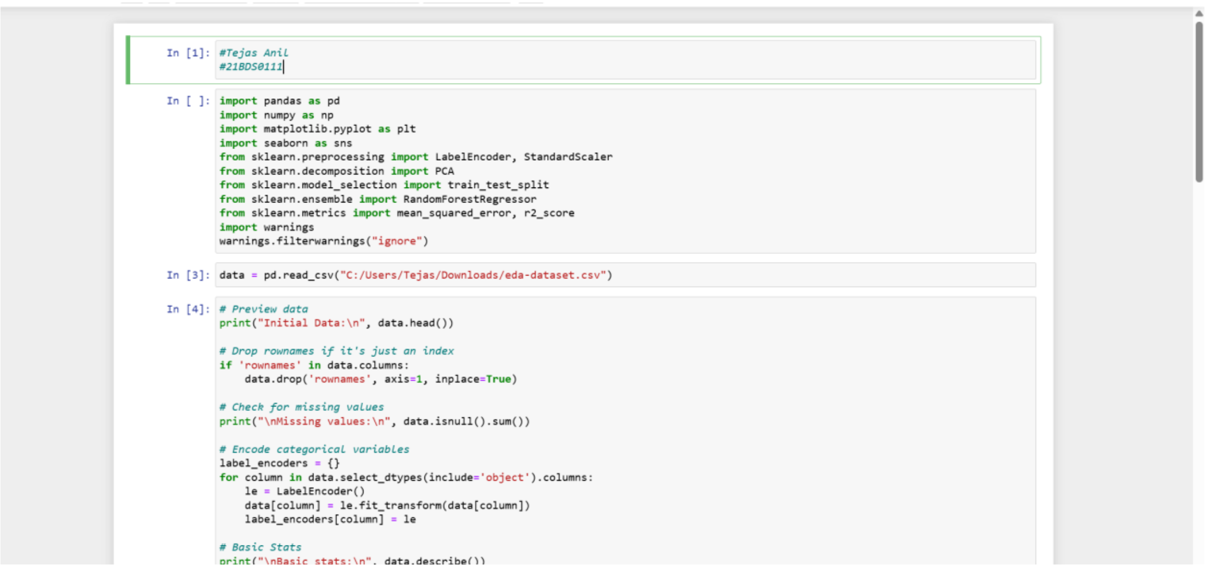
(Deemed to be University under section 3 of UGC Act, 1956)

EDA

DA

TEJAS ANIL

21BDS0111

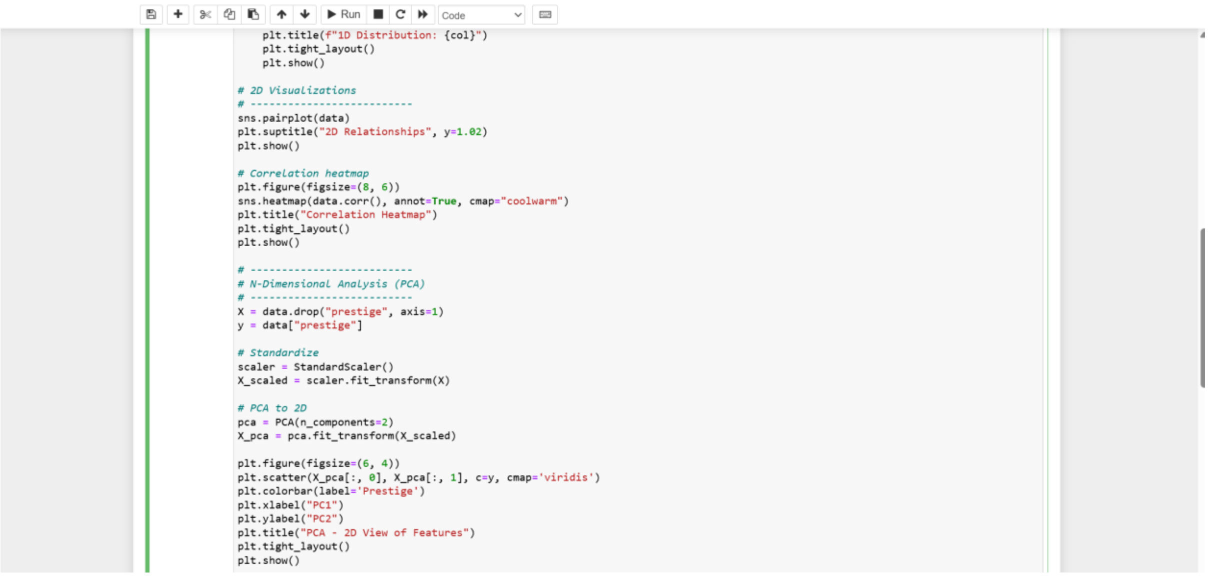


```
In [1]: #Tejas Anil
        #21BDS0111
```

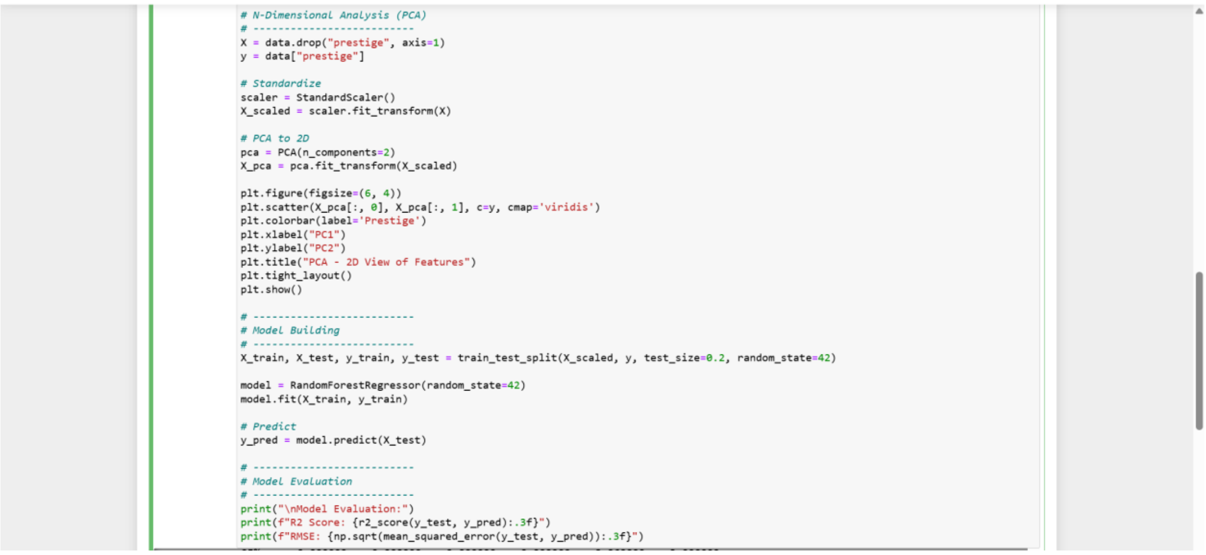
```
In [ ]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        from sklearn.preprocessing import LabelEncoder, StandardScaler
        from sklearn.decomposition import PCA
        from sklearn.model_selection import train_test_split
        from sklearn.ensemble import RandomForestRegressor
        from sklearn.metrics import mean_squared_error, r2_score
        import warnings
        warnings.filterwarnings("ignore")
```

```
In [3]: data = pd.read_csv("C:/Users/Tejas/Downloads/eda-dataset.csv")
```

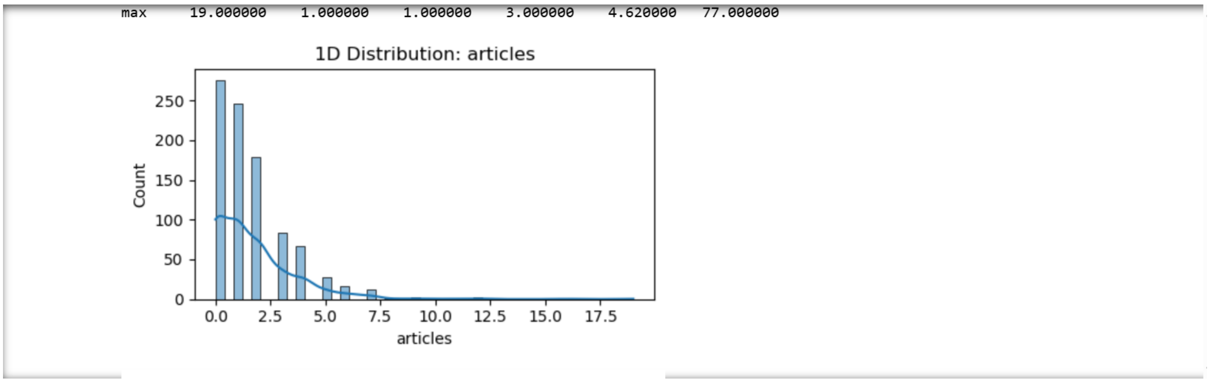
```
In [4]: # Preview data
        print("Initial Data:\n", data.head())

        # Drop rownames if it's just an index
        if 'rownames' in data.columns:
            data.drop('rownames', axis=1, inplace=True)

        # Check for missing values
        print("\nMissing values:\n", data.isnull().sum())

        # Encode categorical variables
        label_encoders = {}
        for column in data.select_dtypes(include='object').columns:
            le = LabelEncoder()
            data[column] = le.fit_transform(data[column])
            label_encoders[column] = le

        # Basic Stats
        print("\nBasic stats:\n", data.describe())
```



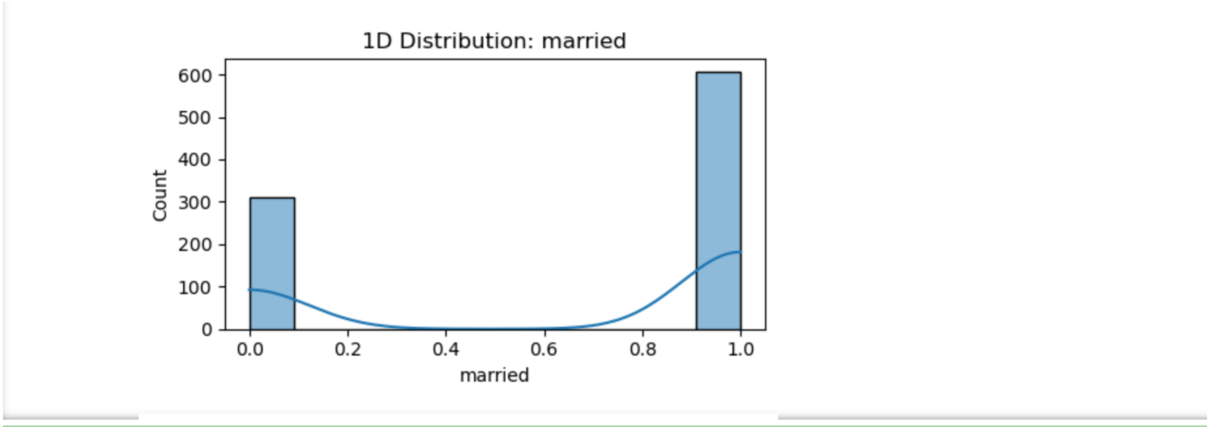
```
    plt.title(f"1D Distribution: {col}")
    plt.tight_layout()
    plt.show()

# 2D Visualizations
# -------------------------
sns.pairplot(data)
plt.suptitle("2D Relationships", y=1.02)
plt.show()

# Correlation heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(data.corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.tight_layout()
plt.show()

# -------------------------
# N-Dimensional Analysis (PCA)
# -------------------------
X = data.drop("prestige", axis=1)
y = data["prestige"]

# Standardize
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# PCA to 2D
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

plt.figure(figsize=(6, 4))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=y, cmap='viridis')
plt.colorbar(label='Prestige')
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("PCA - 2D View of Features")
plt.tight_layout()
plt.show()
```

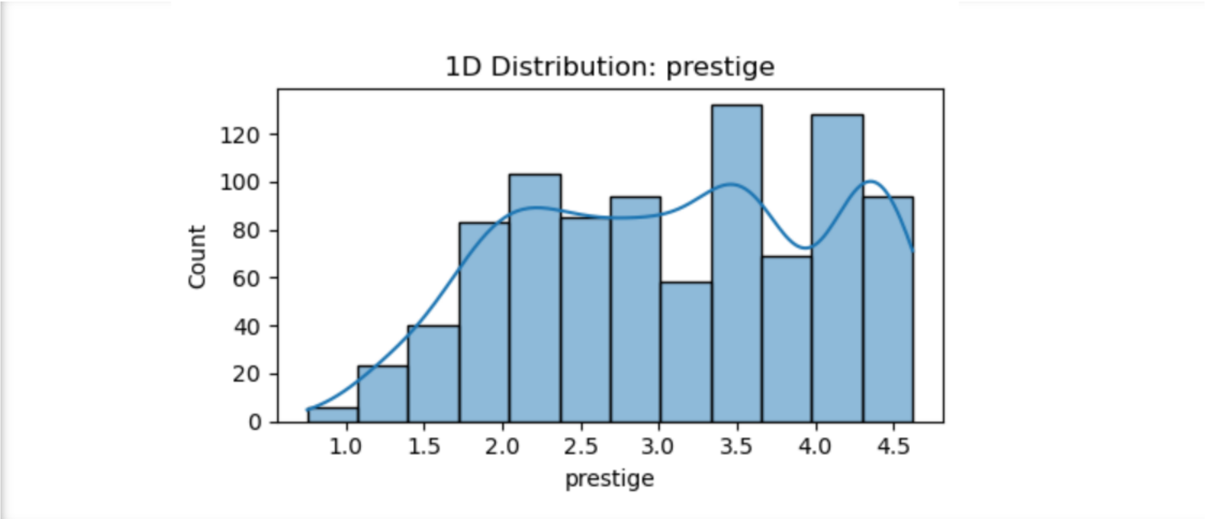


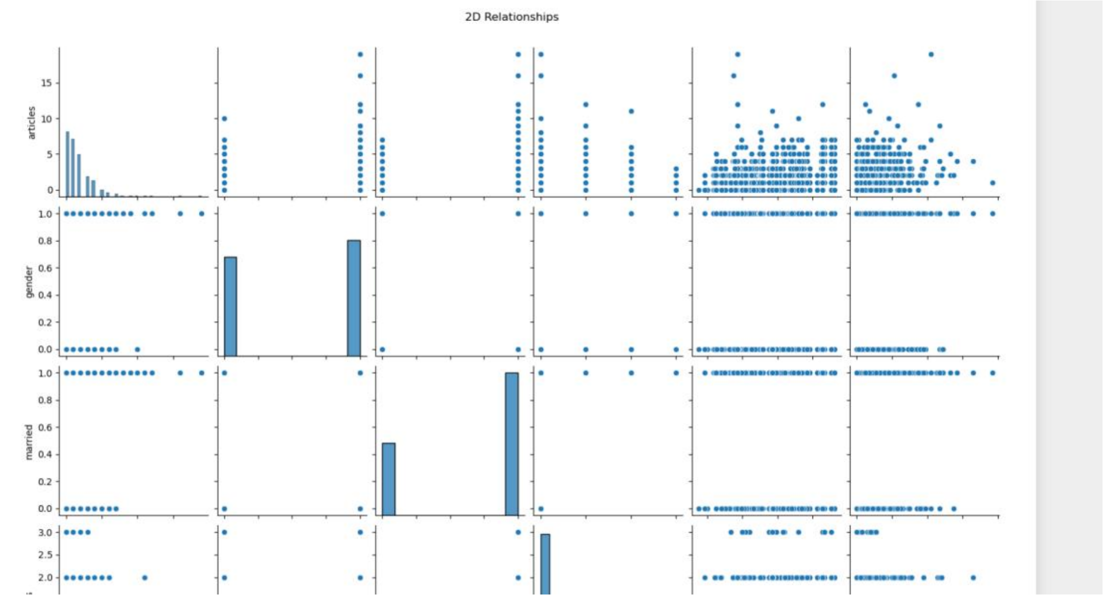
```
# N-Dimensional Analysis (PCA)
# -------------------------
X = data.drop("prestige", axis=1)
y = data["prestige"]

# Standardize
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# PCA to 2D
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

plt.figure(figsize=(6, 4))
plt.scatter(X_pca[:, 0], X_pca[:, 1], c=y, cmap='viridis')
plt.colorbar(label='Prestige')
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("PCA - 2D View of Features")
plt.tight_layout()
plt.show()

# -------------------------
# Model Building
# -------------------------
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

model = RandomForestRegressor(random_state=42)
model.fit(X_train, y_train)

# Predict
y_pred = model.predict(X_test)

# -------------------------
# Model Evaluation
# -------------------------
print("\nModel Evaluation:")
print(f"R2 Score: {r2_score(y_test, y_pred):.3f}")
print(f"RMSE: {np.sqrt(mean_squared_error(y_test, y_pred)):.3f}")
```

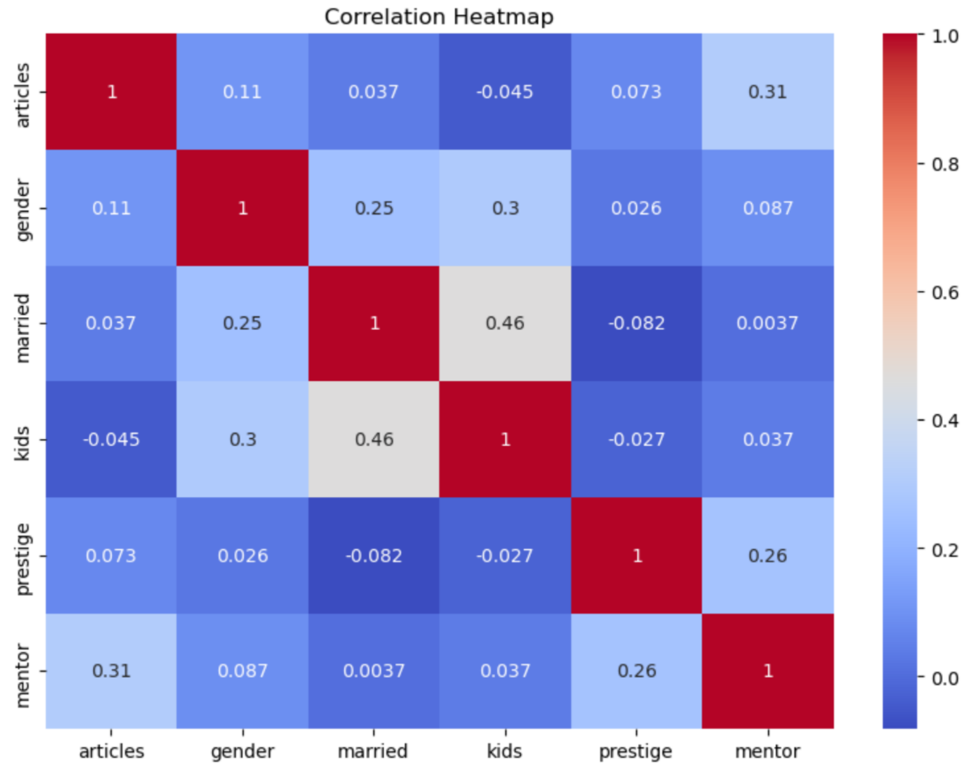


```
max      19.000000    1.000000    1.000000    3.000000    4.620000   77.000000
```

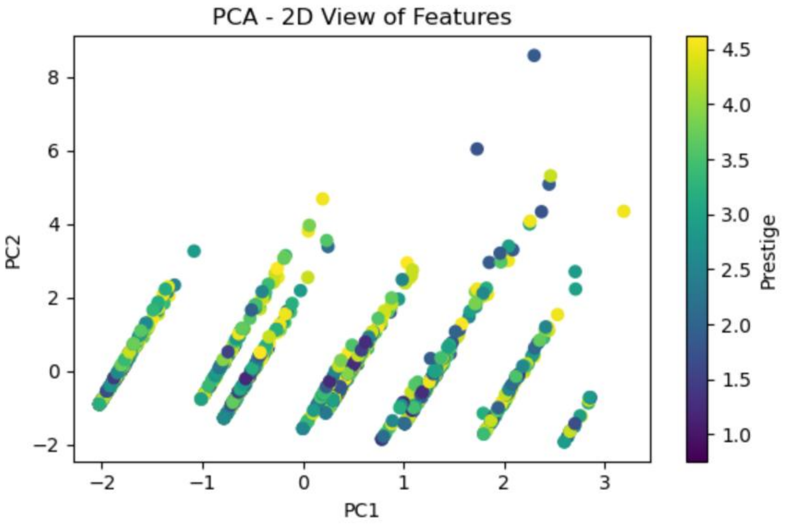



1D Distribution: prestige
Count
0
20
40
60
80
100
120
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
prestige


2D Relationships
articles
0
5
10
15
gender
0.0
0.2
0.4
0.6
0.8
1.0
married
0.0
0.2
0.4
0.6
0.8
1.0
2.0
2.5
3.0



| | articles | gender | married | kids | prestige | mentor |
|---|---|---|---|---|---|---|
| articles | 1 | 0.11 | 0.037 | -0.045 | 0.073 | 0.31 |
| gender | 0.11 | 1 | 0.25 | 0.3 | 0.026 | 0.087 |
| married | 0.037 | 0.25 | 1 | 0.46 | -0.082 | 0.0037 |
| kids | -0.045 | 0.3 | 0.46 | 1 | -0.027 | 0.037 |
| prestige | 0.073 | 0.026 | -0.082 | -0.027 | 1 | 0.26 |
| mentor | 0.31 | 0.087 | 0.0037 | 0.037 | 0.26 | 1 |



```
Model Evaluation:
R2 Score: -0.234
RMSE: 1.085
```