Trabajo Académico final

Roberto Alvarado

May 28, 2025

UTPL

Dataset y presentación de datos

Creación de base de datos

SQLITE, es una herramienta de manejo de bases de datos sql

```
sqlite3 cancer.db
sqlite3> .mode csv
sqlite3> .import global.csv
sqlite3> .exit
```

Limpieza de datos

```
//Eliminar columnas que no sirven
database_csv.loc[:,database_csv.columns != ”Patient_ID”]
database_clean.loc[:,database_clean.columns != ”Country_Region”]

//Conseguir el atributo objectivo
database_clean.loc[:,database_clean.columns == ”Target_Severity_Score”]
```

Analizar data

- Explorar data

  `https://github.com/Kanaries/pygwalker`

- Hacerse preguntas
- Hacer diagramas para responder estas preguntas con diagramas

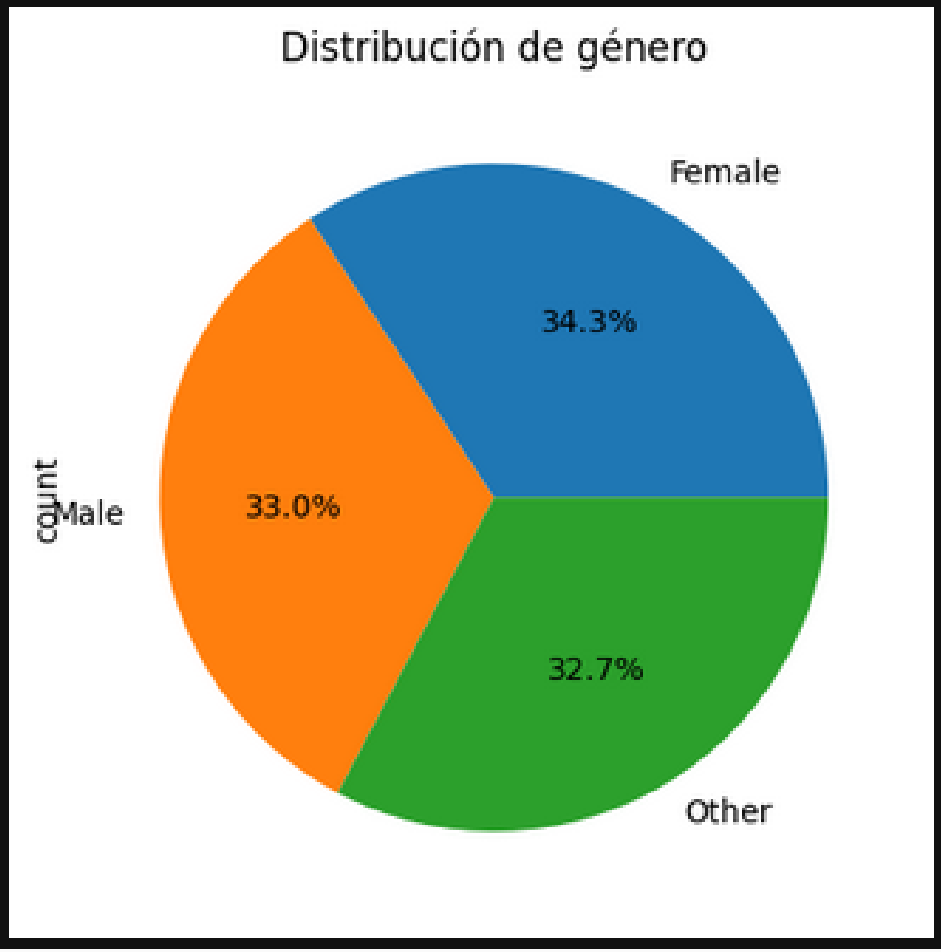
Global Cancer dataset 2019-2025

Attributes [’Patient_ID’,’Age’, ’Gender’, ’Year’, ’Genetic_Risk’, ’Air_Pollution’, ’Alcohol_Use’, ’Smoking’, ’Obesity_Level’, ’Cancer_Type’, ’Cancer_Stage’, ’Treatment_Cost_USD’, ’Survival_Years’,’Target_Severity_Rate’]

¿Comó es la distribución de genéros de esta base datos?

```
#Para hacer los gráficos, voy a responder preguntas
#1. Como esta dividida la base de datos entre generos
plt.title("Distribución de género")
database_csv['Gender'].value_counts().plot.pie(autopct='%1.1f%%')
```

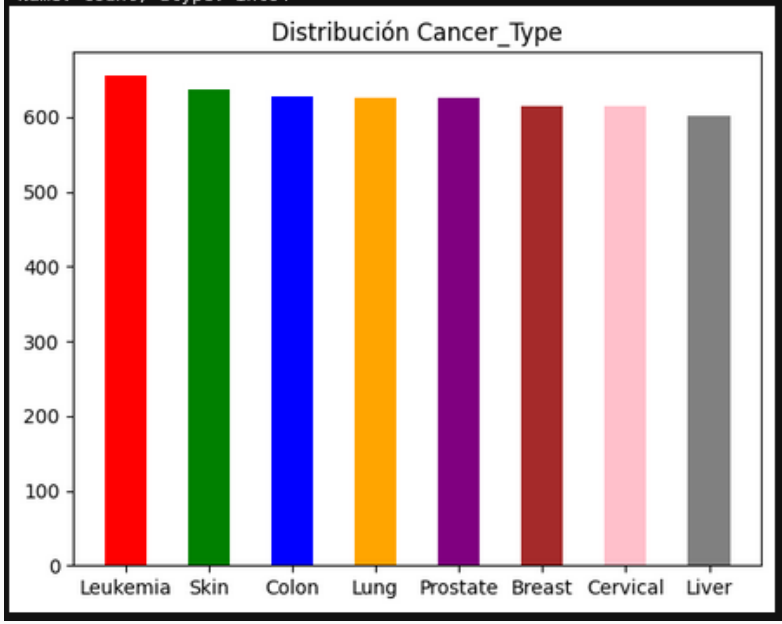
```
<Axes: title={'center': 'Distribución de género'}, ylabel='count'>
```



¿Comó es la distribución de los tipos de cancer de esta base datos?

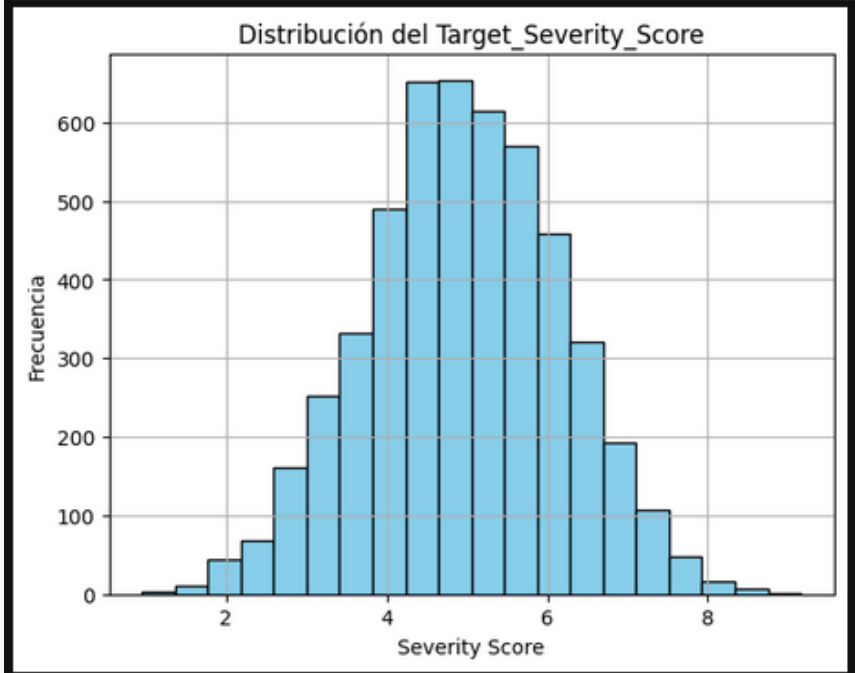
```python
#2. Y los tipos de cancer?
def count_values(col):
    plt.title(f"Distribución {col}")
    data = database_csv[col]
    categories = data.value_counts().index
    counts = data.value_counts().values
    plt.bar(categories, counts, width=0.5,color=get_colors(len(categories)))
    print(data.value_counts())
count_values("Cancer_Type")
```

```
Cancer_Type
Leukemia    655
Skin        636
Colon       627
Lung        626
Prostate    625
Breast      615
Cervical    614
Liver       602
Name: count, dtype: int64
```



¿Comó es la distribución de los resultados?

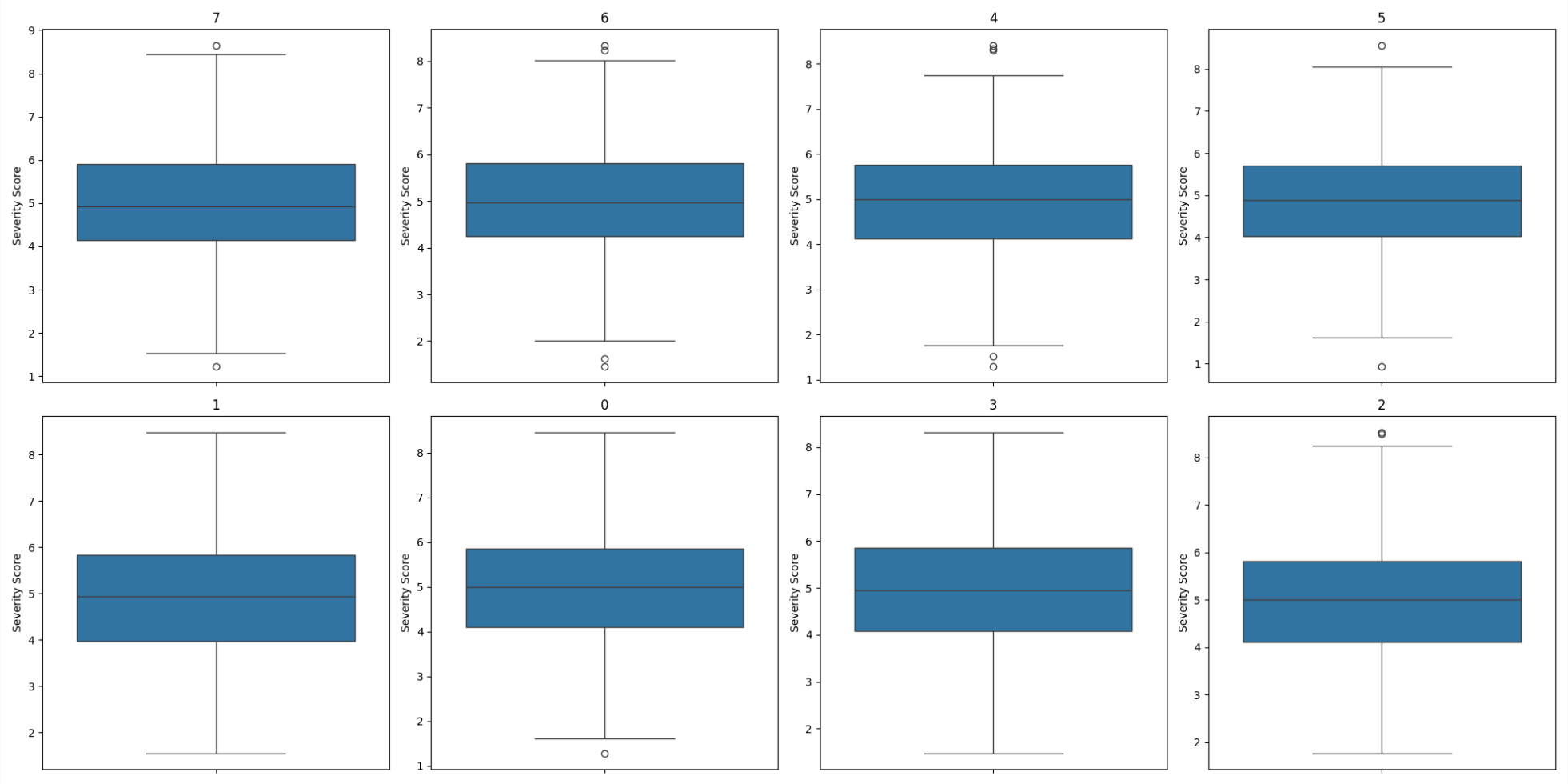
```
#La severidad del cancer como esta distribuida?
plt.hist(database_clean['Target_Severity_Score'], bins=20, color='skyblue', edgecolor='black')
plt.title('Distribución del Target_Severity_Score')
plt.xlabel('Severity Score')
plt.ylabel('Frecuencia')
plt.grid(True)
plt.show()
```



¿Existen tipos de cancer en general más severo el uno de otro?

```
cancer_types = database_clean['Cancer_Type'].unique()
cols = 4
rows = 2
fig, axes = plt.subplots(rows, cols, figsize=(20, 10))
for i, cancer in enumerate(cancer_types):
    ax = axes[i // cols, i % cols]
    sns.boxplot(data=database_clean[database_clean['Cancer_Type'] == cancer],
                y='Target_Severity_Score', ax=ax)
    ax.set_title(f'{cancer}')
    ax.set_ylabel("Severity␣Score")
plt.tight_layout()
plt.show()
```

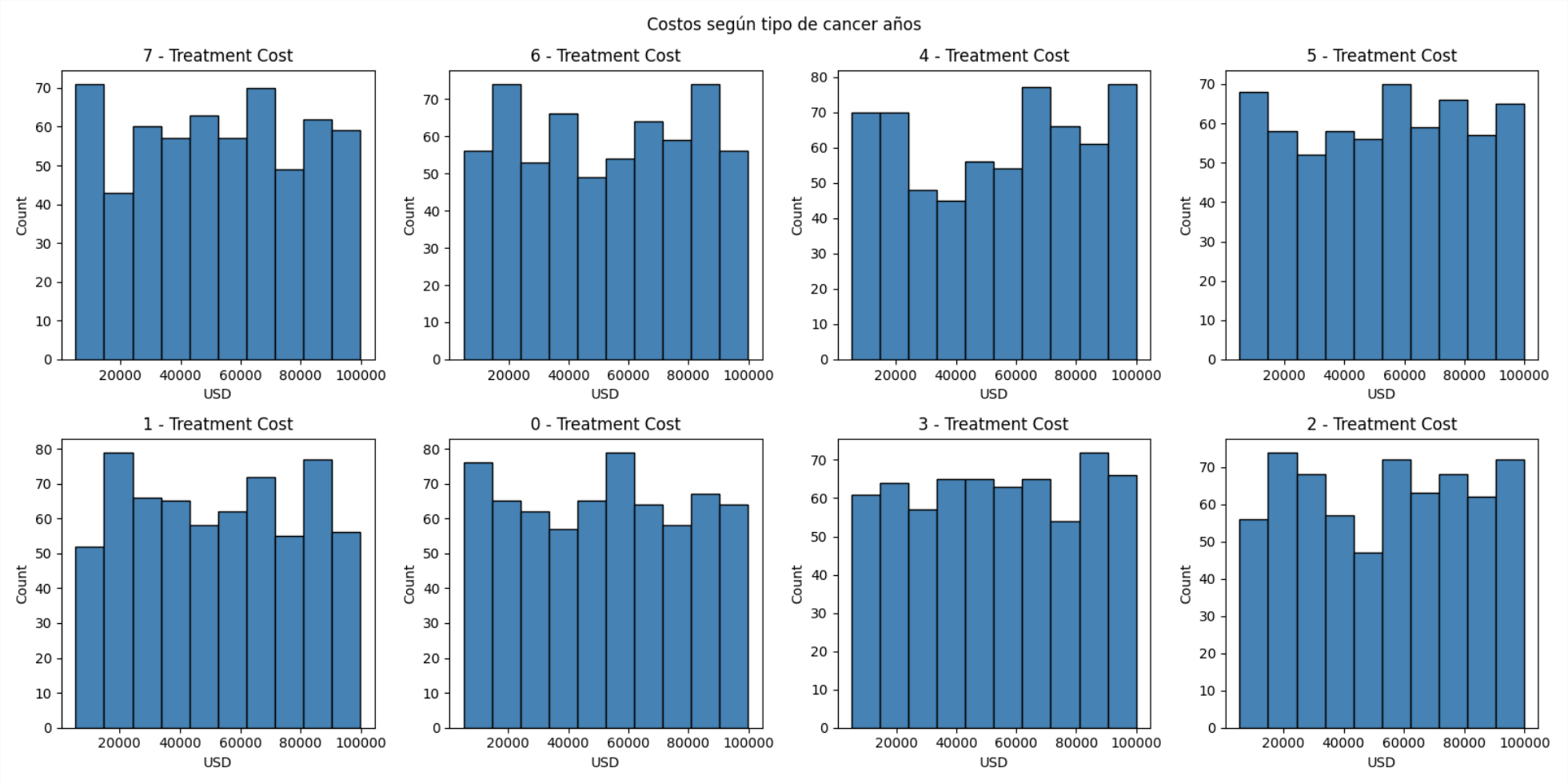
¿Existen tipos de cancer en general más severo el uno de otro?



La severidad según la edad como se presenta

```
cols = 4
rows = 2 fig, axes = plt.subplots(rows, cols, figsize=(16, 4 * rows))
fig.suptitle("Costos␣según␣tipo␣de␣cancer␣")
for i, cancer in enumerate(cancer_types):
    ax = axes[i // cols, i % cols]
    subset = database_clean[database_clean['Cancer_Type'] == cancer]
    ax.hist(subset['Treatment_Cost_USD'], bins=10, color='steelblue', edgecolor='black')
    ax.set_title(f"{cancer}␣-␣Treatment␣Cost")
    ax.set_xlabel('USD')
    ax.set_ylabel('Count')

plt.tight_layout()
plt.savefig("./figures/img6.png")
plt.show()
```

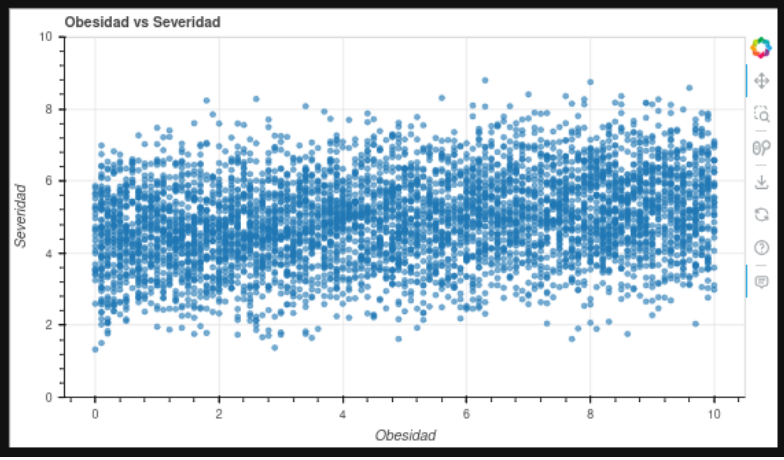
La severidad según la edad como se presenta



¿Cuál es la relación entre obesidad y la severidad?

[58]:
```
from bokeh.plotting import figure, show, output_notebook
from bokeh.models import ColumnDataSource, HoverTool

output_notebook()

source = ColumnDataSource(small_database)
p = figure(title="Obesidad vs Severidad", x_axis_label="Obesidad", y_axis_label="Severidad", width=700, height=400,y_range=(0, 10),)
p.circle(x='Obesity_Level', y='Target_Severity_Score', source=source, size=5, alpha=0.6)

hover = HoverTool(tooltips=[
    ("Obesidad", "@Obesity_Level"),
    ("Género", "@Gender"),
    ("Tipo Cáncer", "@Cancer_Type"),
    ("Severidad", "@Target_Severity_Score"),
])
p.add_tools(hover)

show(p)
```

BokehJS 3.1.1 successfully loaded.



¿Cuál es la relación entre el alcoholismo y la severidad?

```python
import matplotlib.pyplot as plt
import numpy as np

x = database_clean['Smoking']
y = database_clean['Target_Severity_Score']

m, y_i = np.polyfit(x, y, 1)
y_pred = m * x + y_i
plt.figure(figsize=(8, 5))
plt.scatter(x, y, alpha=0.3, label='Data')
plt.plot(x, y_pred, color='red', label=f'y = {slope:.2f}x + {intercept:.2f}')
plt.title('Linear Regression: Smoking vs Severity')
plt.xlabel('Smoking')
plt.ylabel('Target Severity Score')
plt.legend()
plt.grid(True)
plt.show()
```