CS-2704-Data Analytics with Python

Final Report

Topic: *Global Education completion Rate vs Unemployment*

1)Brief explanation of Data:

The world's educational systems are complex and diverse, reflecting the social, economic, and cultural contexts of different countries and regions. This study aims to explore the various dimensions of educational outcomes in the context of global education. These dimensions include gender differences in educational achievement, degrees of educational completion, and reading and writing ability. This data set also emphasizes on the employment rate around the globe and the rate of drop outs from primary school to higher education.

Dataset: ***https://www.kaggle.com/datasets/nelgiriyewithana/world-educational-data/data***

2)Hypothesis:

- My first hypothesis was that there is a negative correlation between the level of education and the unemployment rate globally. In other words, as the level of education increases, the unemployment rate decreases.
- This theory is predicated on the idea that people with greater educational attainment are more qualified and have a wider range of abilities, which increases their employability. It's crucial to remember, though, that a wide

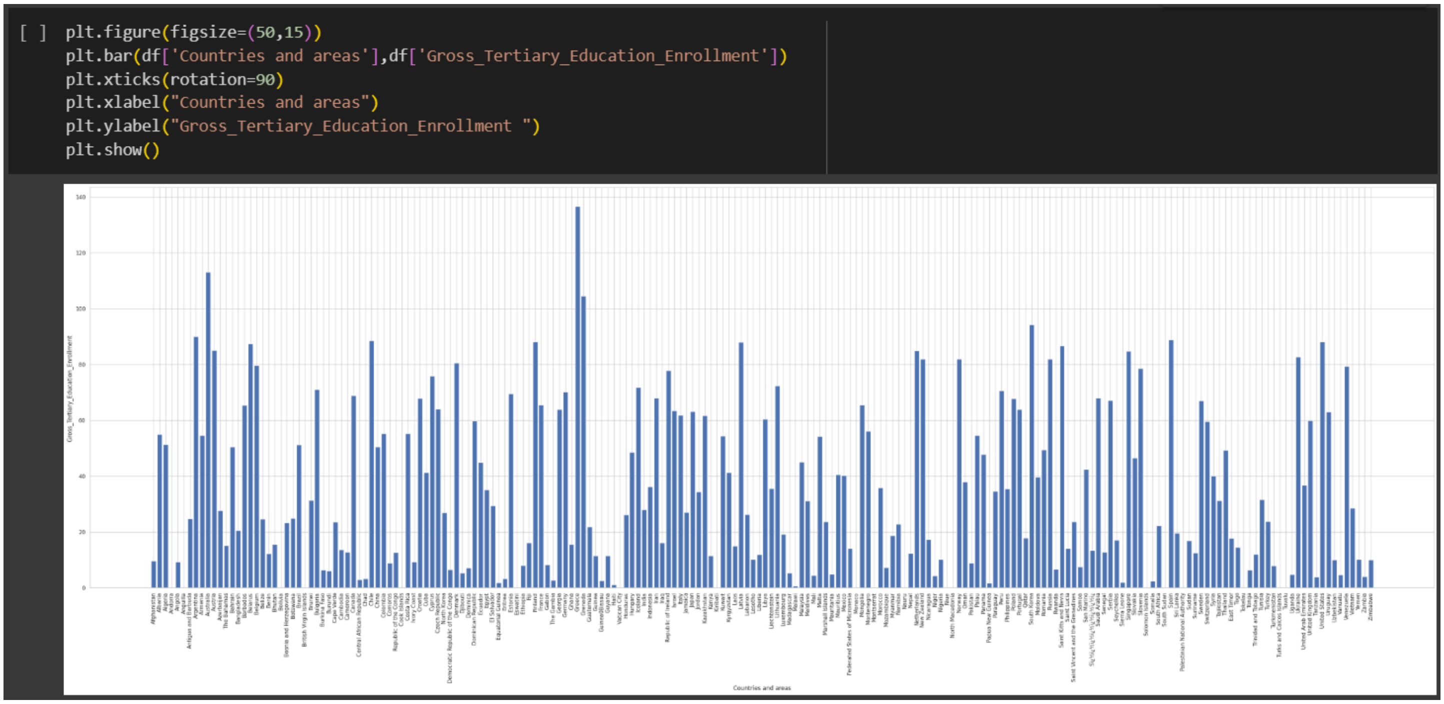
range of other factors, such as governmental regulations, industry growth, and economic conditions, can affect unemployment rates. Consequently, while education can have a big impact, unemployment rates are not only determined by it. To test this theory and ascertain the degree and importance of the association, more investigation and data analysis would be required.

3)The analysis and the implication:

I have analyzed different aspects of the of the data set. There are different columns in the data set about various dimensions of global education. I took few key aspects and compared them. My key aspects include the completion rate of both male and female in primary level, lower secondary level, and higher secondary level. I have also taken the drop out rates of male and female into consideration. My main aim was to compare it with the unemployment problem of the world. There are also important parts where it shows the total global enrollment in primary schools in different regions of the world. I had to trim off some of the data for making my analysis easier. Some parts could have been useful like the proficiency of people at different subjects/aspects like maths, reading & writing but I choose not to make it complex and education vs unemployment was my main target. There were also longitude and latitude of the areas which I didn't take in consideration. I have compared Male vs Female completion rate, made graphs of the total enrollment around the world. I worked in also the dropout rates in different regions of the world. Lastly the graph of total unemployment rate of different regions and then comparing it with the total male and female education completion rate.

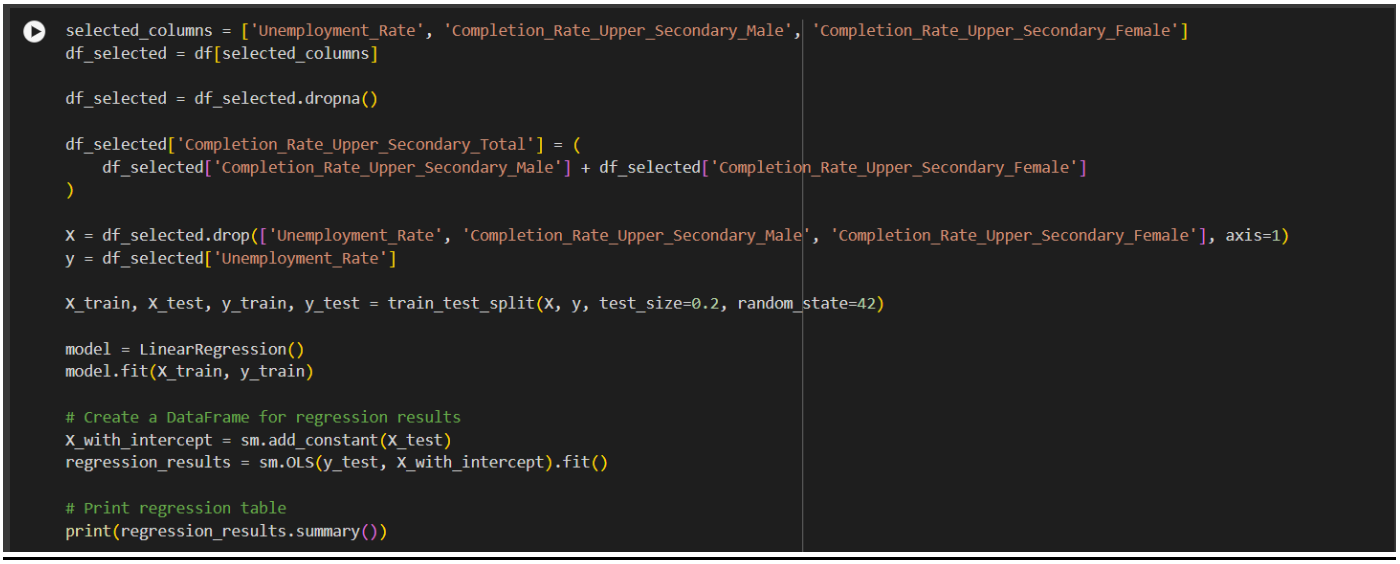
Github: https://github.com/YasinRahman18/World-Education

Global Enrollment Date:

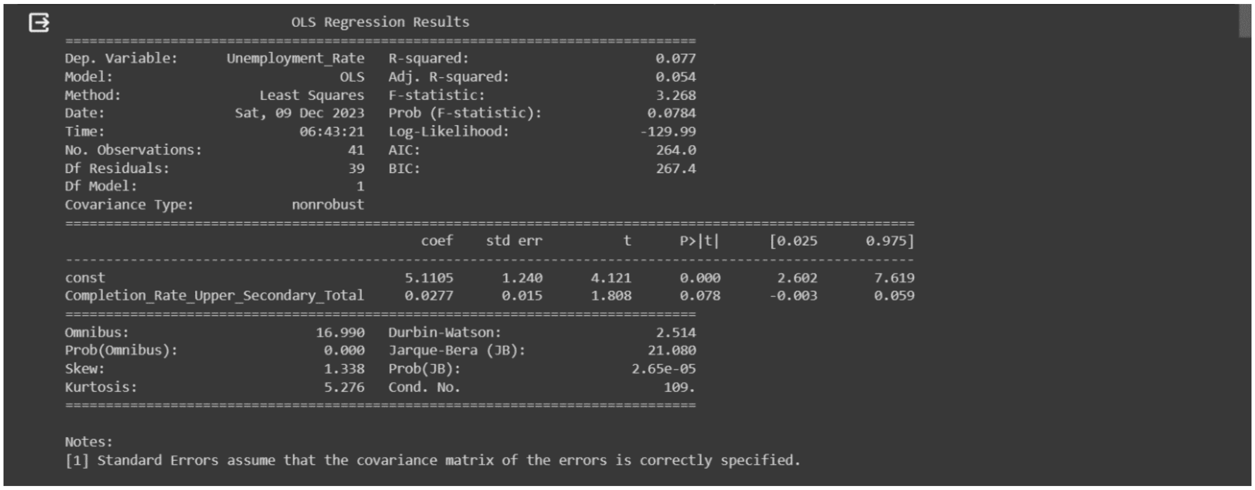


```
[ ] plt.figure(figsize=(50,15))
    plt.bar(df['Countries and areas'],df['Gross_Tertiary_Education_Enrollment'])
    plt.xticks(rotation=90)
    plt.xlabel("Countries and areas")
    plt.ylabel("Gross_Tertiary_Education_Enrollment ")
    plt.show()
```

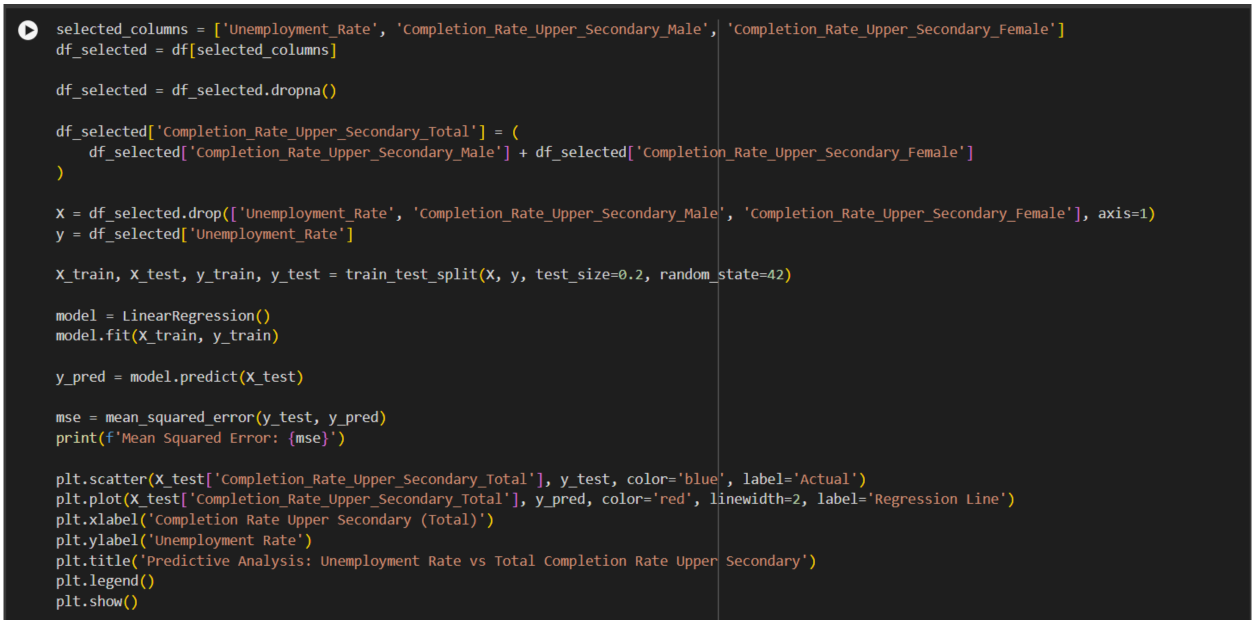
Predictive Model:



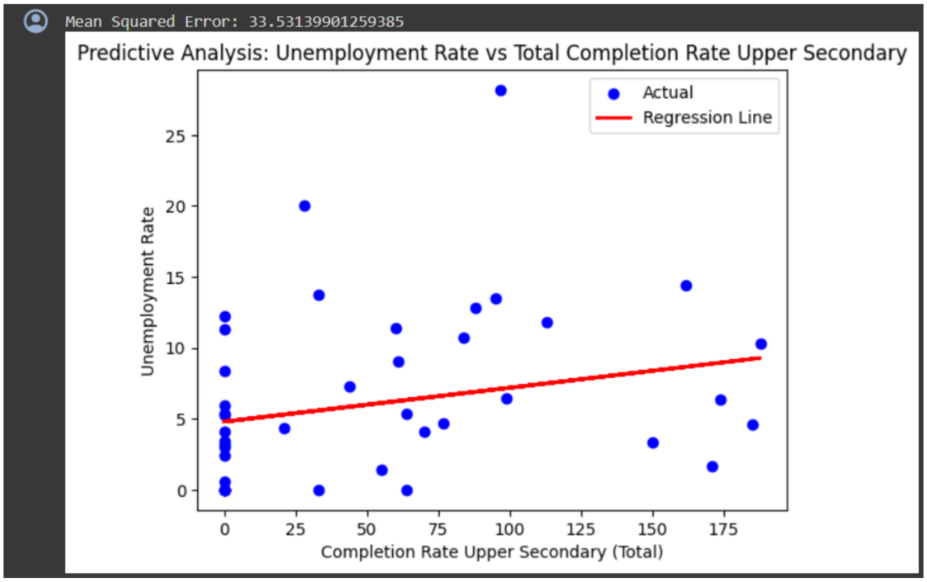
```
selected_columns = ['Unemployment_Rate', 'Completion_Rate_Upper_Secondary_Male', 'Completion_Rate_Upper_Secondary_Female']
df_selected = df[selected_columns]

df_selected = df_selected.dropna()

df_selected['Completion_Rate_Upper_Secondary_Total'] = (
    df_selected['Completion_Rate_Upper_Secondary_Male'] + df_selected['Completion_Rate_Upper_Secondary_Female']
)

X = df_selected.drop(['Unemployment_Rate', 'Completion_Rate_Upper_Secondary_Male', 'Completion_Rate_Upper_Secondary_Female'], axis=1)
y = df_selected['Unemployment_Rate']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

# Create a DataFrame for regression results
X_with_intercept = sm.add_constant(X_test)
regression_results = sm.OLS(y_test, X_with_intercept).fit()

# Print regression table
print(regression_results.summary())
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Unemployment_Rate   R-squared:                       0.077
Model:                            OLS   Adj. R-squared:                  0.054
Method:                 Least Squares   F-statistic:                     3.268
Date:                Sat, 09 Dec 2023   Prob (F-statistic):             0.0784
Time:                        06:43:21   Log-Likelihood:                -129.99
No. Observations:                  41   AIC:                             264.0
Df Residuals:                      39   BIC:                             267.4
Df Model:                           1
Covariance Type:            nonrobust
=====================================================================================================
                                        coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------
const                                 5.1105      1.240      4.121      0.000       2.602       7.619
Completion_Rate_Upper_Secondary_Total 0.0277      0.015      1.808      0.078      -0.003       0.059
==============================================================================
Omnibus:                       16.990   Durbin-Watson:                   2.514
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               21.080
Skew:                           1.338   Prob(JB):                     2.65e-05
Kurtosis:                       5.276   Cond. No.                         109.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

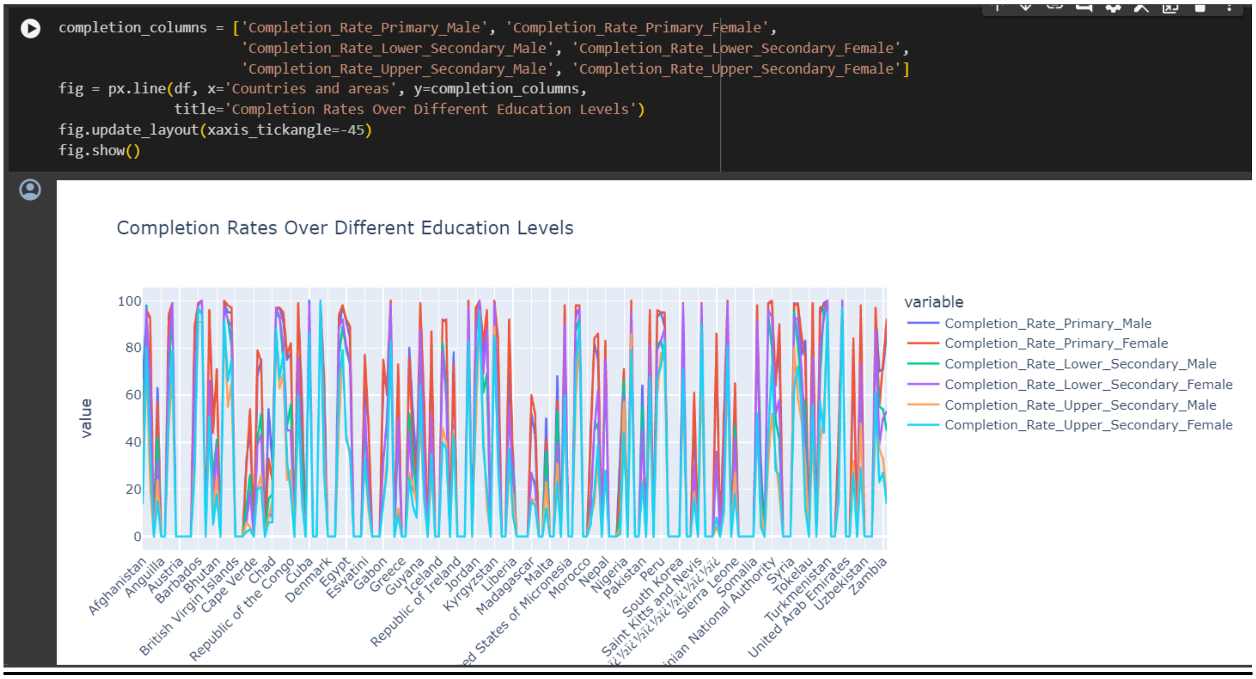


```python
selected_columns = ['Unemployment_Rate', 'Completion_Rate_Upper_Secondary_Male', 'Completion_Rate_Upper_Secondary_Female']
df_selected = df[selected_columns]

df_selected = df_selected.dropna()

df_selected['Completion_Rate_Upper_Secondary_Total'] = (
    df_selected['Completion_Rate_Upper_Secondary_Male'] + df_selected['Completion_Rate_Upper_Secondary_Female']
)

X = df_selected.drop(['Unemployment_Rate', 'Completion_Rate_Upper_Secondary_Male', 'Completion_Rate_Upper_Secondary_Female'], axis=1)
y = df_selected['Unemployment_Rate']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
print(f'Mean Squared Error: {mse}')

plt.scatter(X_test['Completion_Rate_Upper_Secondary_Total'], y_test, color='blue', label='Actual')
plt.plot(X_test['Completion_Rate_Upper_Secondary_Total'], y_pred, color='red', linewidth=2, label='Regression Line')
plt.xlabel('Completion Rate Upper Secondary (Total)')
plt.ylabel('Unemployment Rate')
plt.title('Predictive Analysis: Unemployment Rate vs Total Completion Rate Upper Secondary')
plt.legend()
plt.show()
```



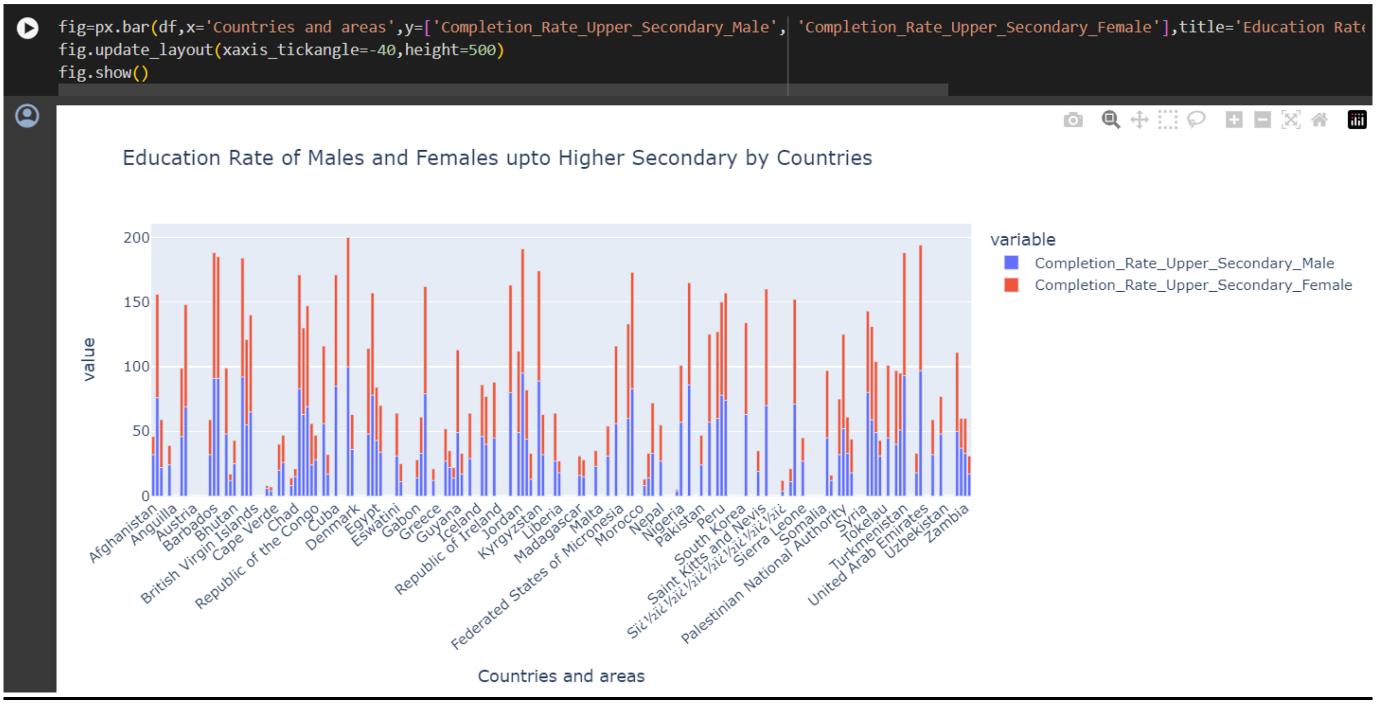
```
Mean Squared Error: 33.53139901259385
```

Completion Rate in different countries:



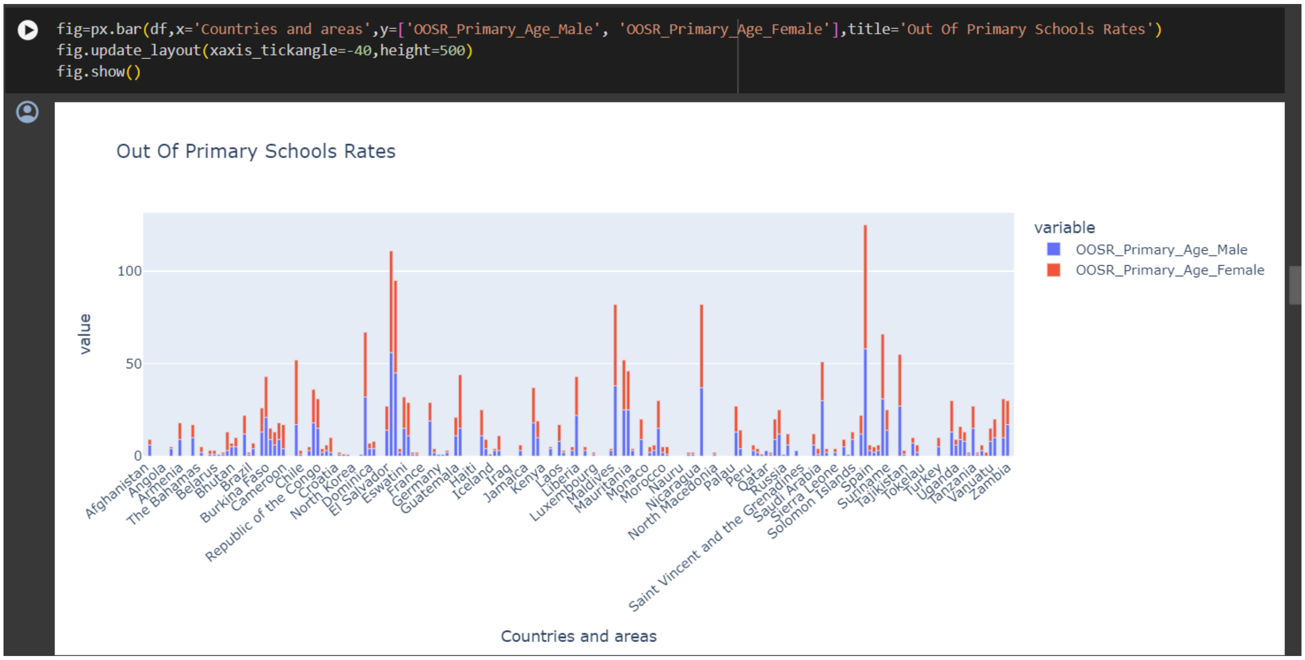
```
completion_columns = ['Completion_Rate_Primary_Male', 'Completion_Rate_Primary_Female',
                      'Completion_Rate_Lower_Secondary_Male', 'Completion_Rate_Lower_Secondary_Female',
                      'Completion_Rate_Upper_Secondary_Male', 'Completion_Rate_Upper_Secondary_Female']
fig = px.line(df, x='Countries and areas', y=completion_columns,
              title='Completion Rates Over Different Education Levels')
fig.update_layout(xaxis_tickangle=-45)
fig.show()
```

Upper Secondary Completion rates:



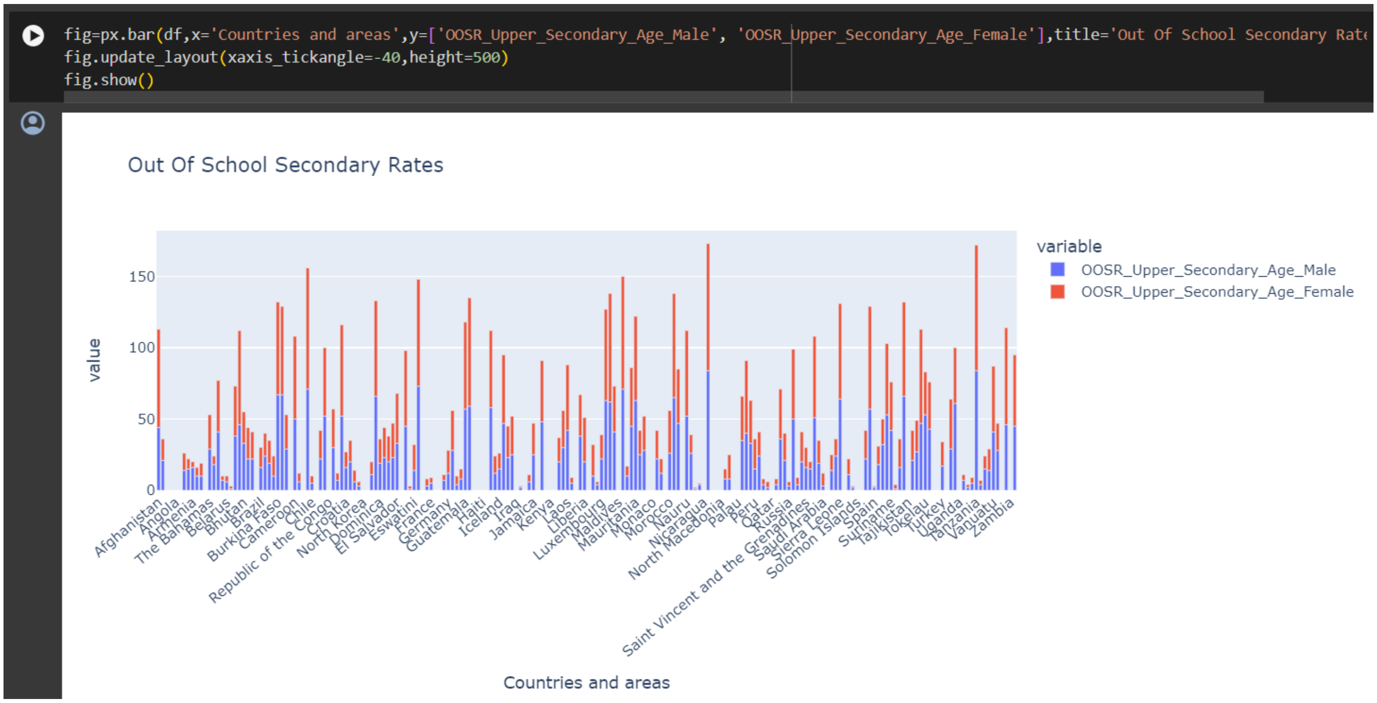
```
fig=px.bar(df,x='Countries and areas',y=['Completion_Rate_Upper_Secondary_Male', 'Completion_Rate_Upper_Secondary_Female'],title='Education Rat
fig.update_layout(xaxis_tickangle=-40,height=500)
fig.show()
```

Dropping out of Primary schools:



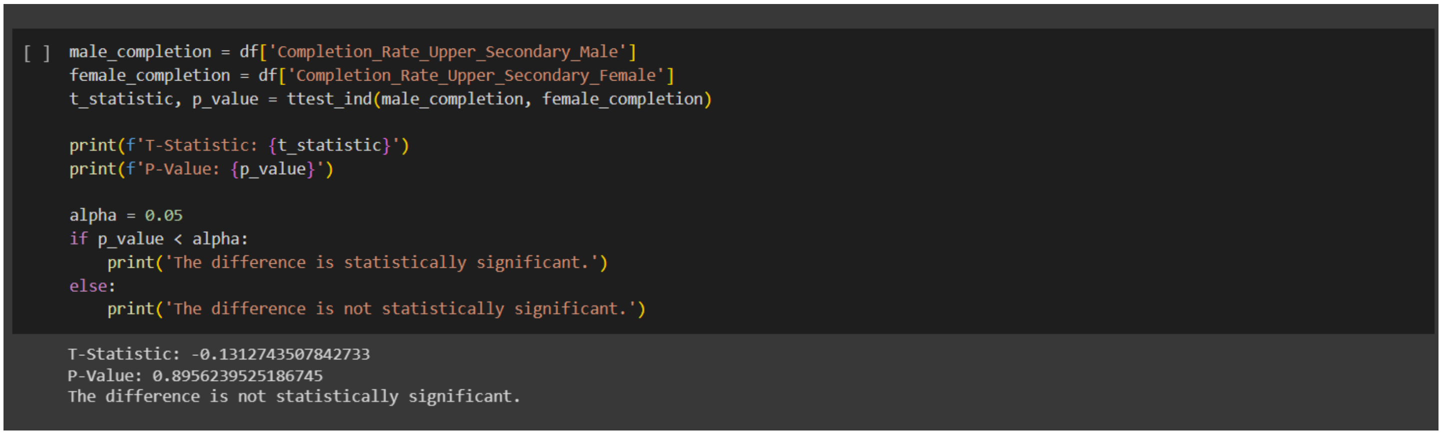
```
fig=px.bar(df,x='Countries and areas',y=['OOSR_Primary_Age_Male', 'OOSR_Primary_Age_Female'],title='Out Of Primary Schools Rates')
fig.update_layout(xaxis_tickangle=-40,height=500)
fig.show()
```

Dropping out of Secondary school:



```
fig=px.bar(df,x='Countries and areas',y=['OOSR_Upper_Secondary_Age_Male', 'OOSR_Upper_Secondary_Age_Female'],title='Out Of School Secondary Rate
fig.update_layout(xaxis_tickangle=-40,height=500)
fig.show()
```

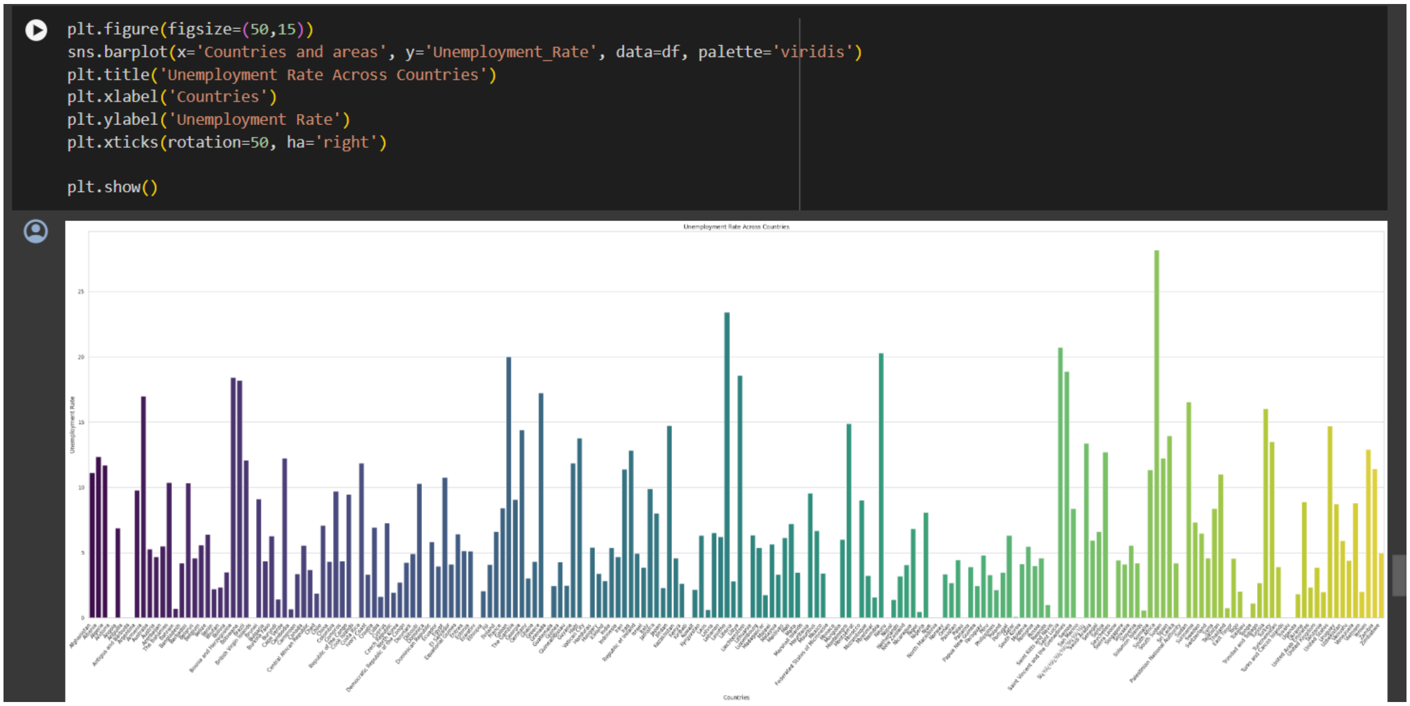
T-test Male vs Female education:



```
[ ] male_completion = df['Completion_Rate_Upper_Secondary_Male']
    female_completion = df['Completion_Rate_Upper_Secondary_Female']
    t_statistic, p_value = ttest_ind(male_completion, female_completion)

    print(f'T-Statistic: {t_statistic}')
    print(f'P-Value: {p_value}')

    alpha = 0.05
    if p_value < alpha:
        print('The difference is statistically significant.')
    else:
        print('The difference is not statistically significant.')
```

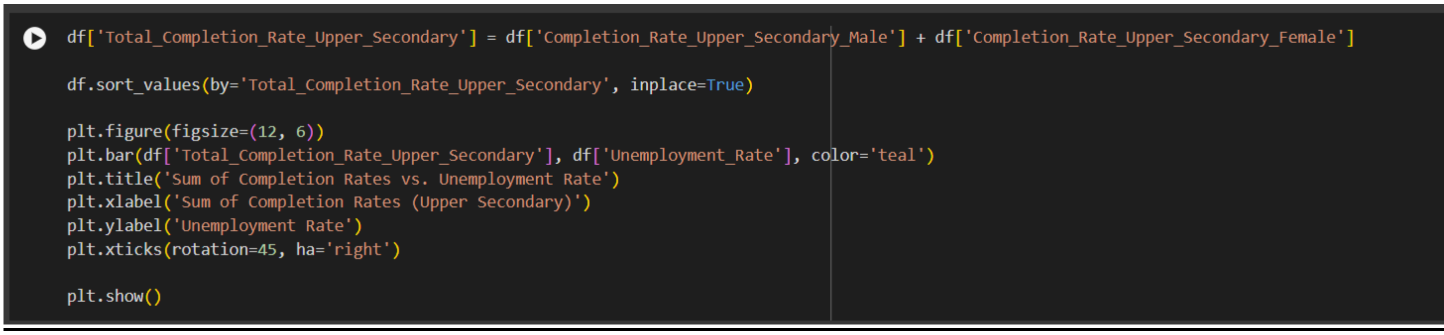
```
T-Statistic: -0.1312743507842733
P-Value: 0.8956239525186745
The difference is not statistically significant.
```

Unemployment by countries:

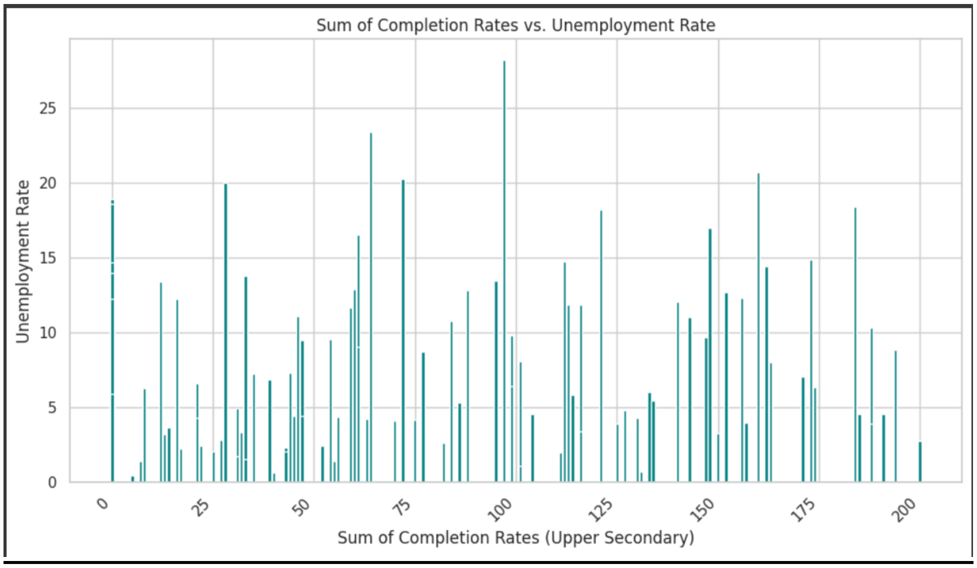


```
plt.figure(figsize=(50,15))
sns.barplot(x='Countries and areas', y='Unemployment_Rate', data=df, palette='viridis')
plt.title('Unemployment Rate Across Countries')
plt.xlabel('Countries')
plt.ylabel('Unemployment Rate')
plt.xticks(rotation=50, ha='right')

plt.show()
```

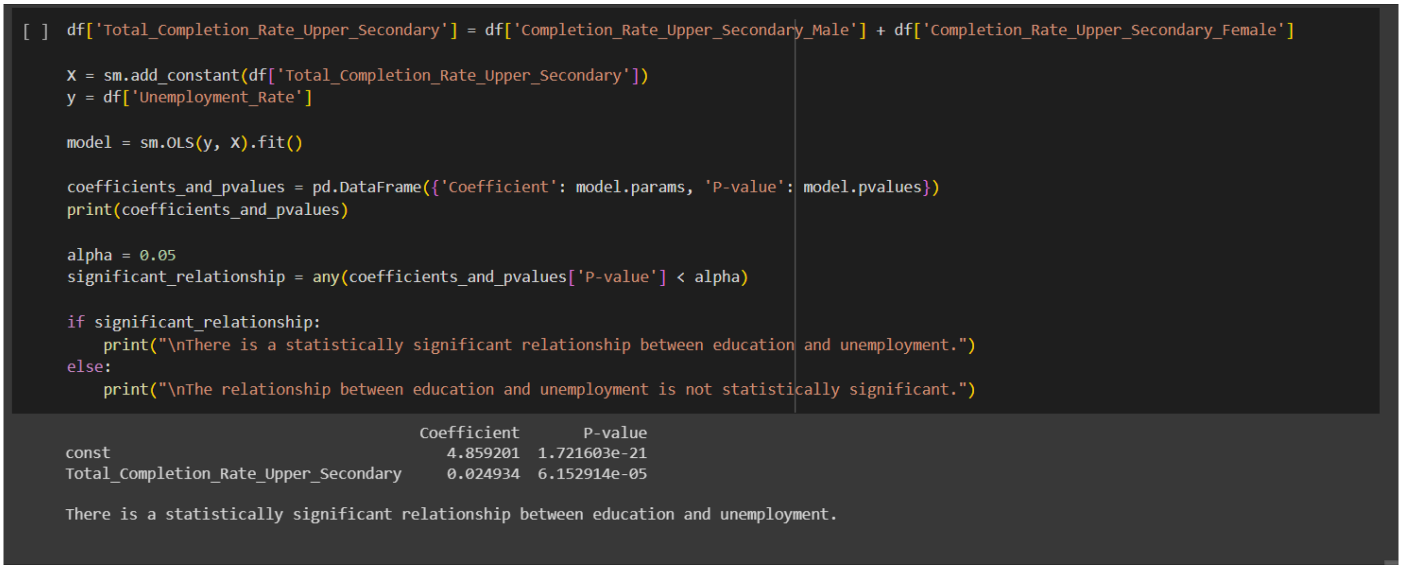
Unemployment vs Education completion:



```
df['Total_Completion_Rate_Upper_Secondary'] = df['Completion_Rate_Upper_Secondary_Male'] + df['Completion_Rate_Upper_Secondary_Female']

df.sort_values(by='Total_Completion_Rate_Upper_Secondary', inplace=True)

plt.figure(figsize=(12, 6))
plt.bar(df['Total_Completion_Rate_Upper_Secondary'], df['Unemployment_Rate'], color='teal')
plt.title('Sum of Completion Rates vs. Unemployment Rate')
plt.xlabel('Sum of Completion Rates (Upper Secondary)')
plt.ylabel('Unemployment Rate')
plt.xticks(rotation=45, ha='right')

plt.show()
```



Relation between Education and Unemployment:



```
[ ] df['Total_Completion_Rate_Upper_Secondary'] = df['Completion_Rate_Upper_Secondary_Male'] + df['Completion_Rate_Upper_Secondary_Female']

    X = sm.add_constant(df['Total_Completion_Rate_Upper_Secondary'])
    y = df['Unemployment_Rate']

    model = sm.OLS(y, X).fit()

    coefficients_and_pvalues = pd.DataFrame({'Coefficient': model.params, 'P-value': model.pvalues})
    print(coefficients_and_pvalues)

    alpha = 0.05
    significant_relationship = any(coefficients_and_pvalues['P-value'] < alpha)

    if significant_relationship:
        print("\nThere is a statistically significant relationship between education and unemployment.")
    else:
        print("\nThe relationship between education and unemployment is not statistically significant.")
```

```
                                       Coefficient       P-value
const                                     4.859201  1.721603e-21
Total_Completion_Rate_Upper_Secondary     0.024934  6.152914e-05

There is a statistically significant relationship between education and unemployment.
```

4) Conclusion:

There is certainly a relationship between education completion rate and unemployment rate in the regions. Based on the testing results, we can conclude that there is a positive linear relationship between unemployment rate and completion rate of upper secondary education in most countries, but this relationship is not very strong or consistent across countries or over time. There may be other factors that influence both variables or confound their effects, such as economic conditions, social policies, demographic trends, etc., that need to be considered when interpreting these results or making predictions based on them. The predictive model doesn’t show much significance at this point with the recourses I have. To improve the model's effectiveness, more data & predictors may be required as it may not be a reliable indicator of the unemployment rate.

5) References:

- Chen Zhongchang and Wu Yongqiu, Author “The relationship between education and employment: A theoretical analysis and empirical test”

- Kaggle , A prefect website for demonstration of different data analytics works “ Kaggle: Your Machine Learning and Data Science Community”

- Hadley Wickham: Chief Scientist at RStudio, known for his contributions to the R programming language and various data science hadley (Hadley Wickham) · GitHub