Ankan Roy - MSML602 - UID: 117527476

Project 3

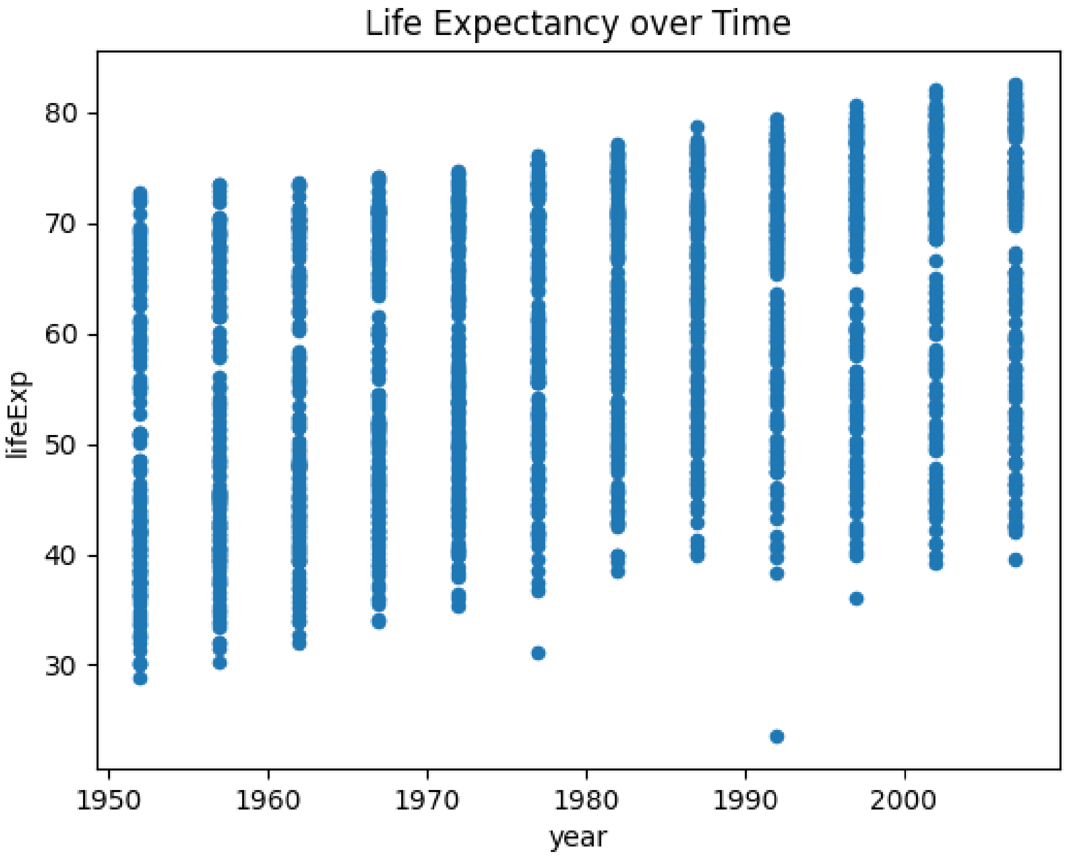
```
import pandas as pd
import numpy as np

data = pd.read_csv("./08_gap-every-five-years.tsv", sep='\t')

data.head()
```

```
       country continent  year  lifeExp       pop   gdpPercap
0  Afghanistan      Asia  1952   28.801   8425333  779.445314
1  Afghanistan      Asia  1957   30.332   9240934  820.853030
2  Afghanistan      Asia  1962   31.997  10267083  853.100710
3  Afghanistan      Asia  1967   34.020  11537966  836.197138
4  Afghanistan      Asia  1972   36.088  13079460  739.981106
```

Exercise 1: *Make a scatter plot of life expectancy across time.*

```
data.plot.scatter(x="year", y="lifeExp", title="Life Expectancy over Time")
```

```
<Axes: title={'center': 'Life Expectancy over Time'}, xlabel='year',
ylabel='lifeExp'>
```

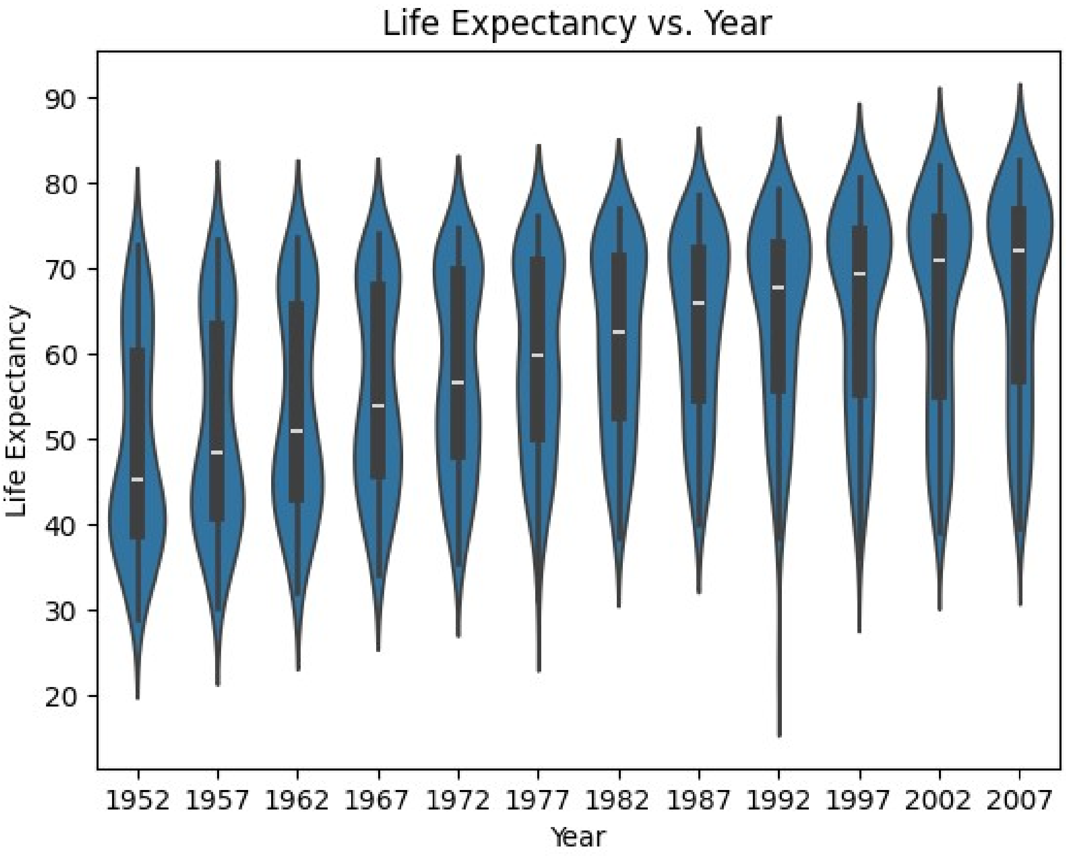


Question 1: *Is there a general trend (e.g., increasing or decreasing) for life expectancy across time? Is this trend linear? (answering this qualitatively from the plot, you will do a statistical analysis of this question shortly)*

After generating the average life expectancy over time across all countries, there is a clear increasing linear trend for life expectancy as the years go on.

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

sns.violinplot(x ="year",
             y ="lifeExp",
             data = data)

plt.xlabel("Year")
plt.ylabel("Life Expectancy")
plt.title("Life Expectancy vs. Year")
```

```
Text(0.5, 1.0, 'Life Expectancy vs. Year')
```



Question 2: How would you describe the distribution of life expectancy across countries for individual years? Is it skewed, or not? Unimodal or not? Symmetric around it's center?

According to the generated violin plot, near the beginning (year 1952), the life expectancy had two peaks, particularly ~30-40 years and ~65-75 yeras. However, as time progressed, the life expectancy skewed more to ages ~70-80. This shows that life expectancy has increased over time.

Question 3: Suppose I fit a linear regression model of life expectancy vs. year (treating it as a continuous variable), and test for a relationship between year and life expectancy, will you reject the null hypothesis of no relationship? (do this without actually writing the code and fitting the model yet. I am testing your intuition.)

n_0 = there is no relationship between life expectancy and year.

Considering I fit a linear regression model of life expectancy vs. year, I would reject the null hypothesis, as there is a relationship between life expectancy and year. We can see that life expectancy increases over time.

Question 4: What would a violin plot of residuals (errors) from the linear model in Question 3 vs. year look like? (Again, don't do the analysis yet, answer this intuitively). You would answer this question in the context of the variables in Question 3.

Considering we have a linear plot of residuals, a violin plot of residuals should be clustered around 0, meaning that the predictions being made are generaly accurate.

Question 5: According to the assumptions of the linear regression model, what should that violin plot look like? You would answer this question as a general property of the residuals of the linear regression. We are looking to see if the Question 3 model fits the general assumptions.

Considering we have a violin plot of residuals, the distribution would be relatively narrow, meaning that the predictions being made are generally accurate.

Exercise 2: Fit a linear regression model using, e.g., the `LinearRegression` function from Scikit-Learn or the closed-form solution we derived in class, for life expectancy vs. year (as a continuous variable).

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

X = data[['year']]
y = data['lifeExp']

X_train, X_test, y_train, y_test =
train_test_split(X,y,test_size=0.2,random_state=101)

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

distance_pred = lm.predict(X)
distance_pred

array([50.63854468, 52.22083481, 53.80312494, ..., 64.87915585,
       66.46144598, 68.04373611])
```

Question 6: On average, by how much does life expectancy increase every year around the world?

```
import statsmodels.formula.api as smf

model = smf.ols(formula="y ~ X", data=data).fit()
print(model.summary())

                            OLS Regression Results

======================================================================
========
Dep. Variable:                      y   R-squared:
0.190
Model:                            OLS   Adj. R-squared:
0.189
Method:                 Least Squares   F-statistic:
398.6
Date:                Sun, 09 Nov 2025   Prob (F-statistic):
7.55e-80
Time:                        17:23:22   Log-Likelihood:
-6597.9
No. Observations:                1704   AIC:
```

```
1.320e+04
Df Residuals:                    1702   BIC:
1.321e+04
Df Model:                           1

Covariance Type:            nonrobust

======================================================================
========
                 coef    std err          t      P>|t|      [0.025
0.975]
----------------------------------------------------------------------
--------
Intercept   -585.6522     32.314    -18.124      0.000    -649.031
-522.273
X              0.3259      0.016     19.965      0.000       0.294
0.358
======================================================================
========
Omnibus:                      386.124   Durbin-Watson:
0.197
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
90.750
Skew:                          -0.268   Prob(JB):
1.97e-20
Kurtosis:                       2.004   Cond. No.
2.27e+05
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 2.27e+05. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

Question 7: Do you reject the null hypothesis of no relationship between year and life expectancy? Why?

According to the OLS Regression Results, we have the following:

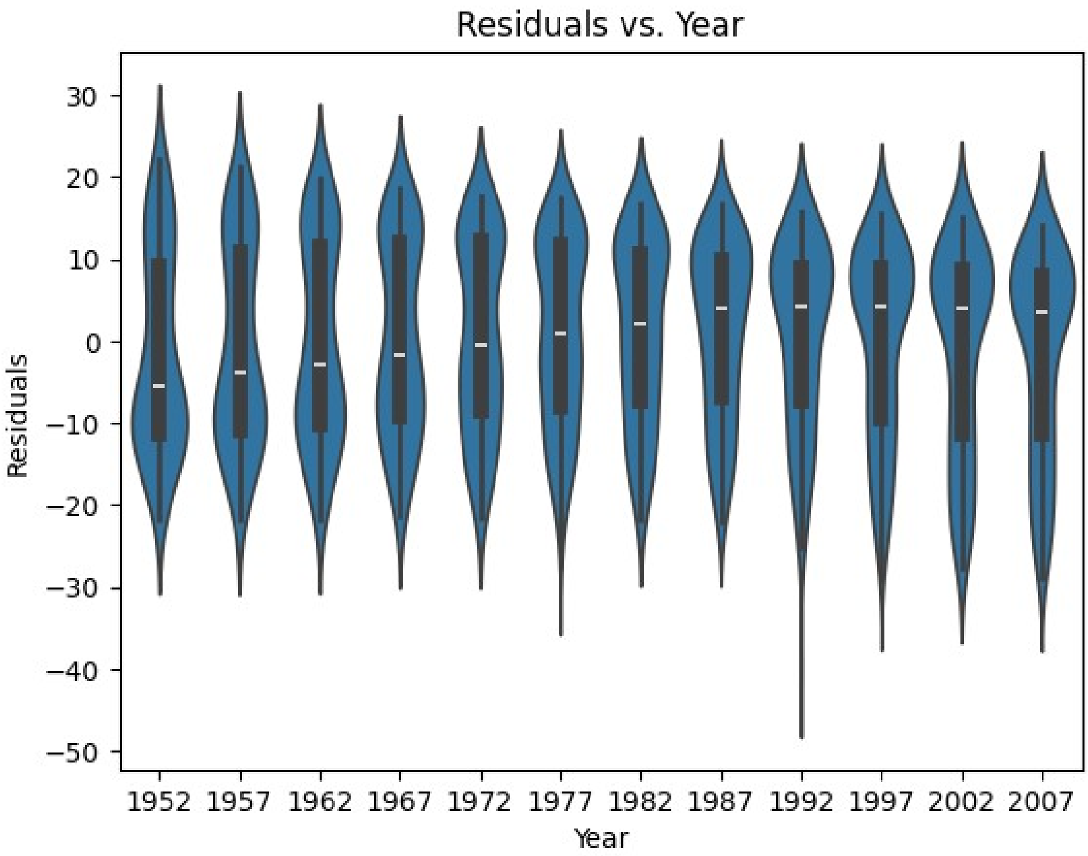
- coef = 0.3259
- std err = 0.016
- t = 19.965
- p-value (P>|t|) = 0.000

According to the results, life expectancy increases on average by ~0.33 years per year globally. This increase is statistically significant with a P-Value of 0.000. The model explains about 19% of the variation in life expectancy. With this in mind, we would reject the null hypothesis.

Exercise 3: Make a violin plot of residuals vs. year for the linear model from Exercise 2.

```
data["residuals"] = model.resid

sns.violinplot(x="year", y="residuals", data=data)

plt.xlabel("Year")
plt.ylabel("Residuals")
plt.title("Residuals vs. Year")
```

```
Text(0.5, 1.0, 'Residuals vs. Year')
```



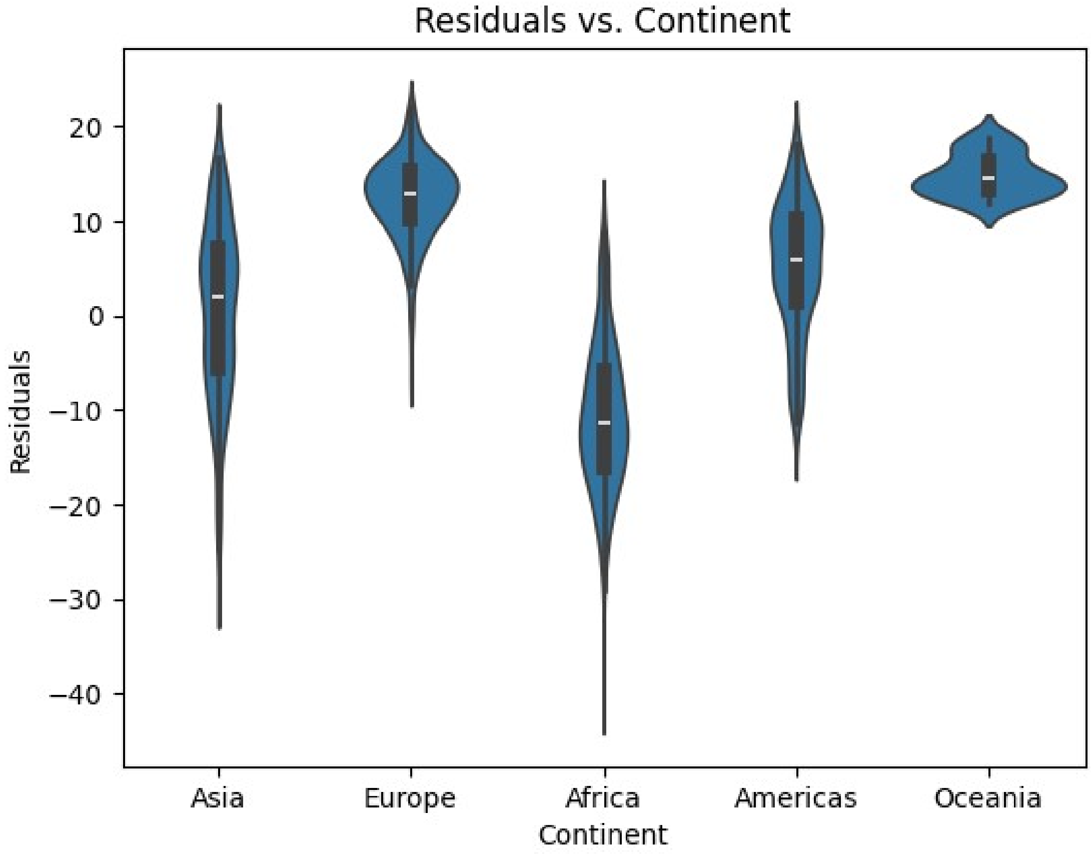
Question 8: Does the plot of Exercise 3 match your expectations (as you answered Question 4)?

The violin plot shows that in earlier years (~ 1952-1972), most residuals are negative, indicating that actual life expectancy was lower than what the linear model predicted. In later years (1982-2007), the residuals shift upward, suggesting that actual life expectancy exceeded the model's predictions. This pattern indicates that the increase in global life expectancy over time is not strictly linear. In contrast, life expectancy improved more rapidly in the later decades. This contradicts my expectation in Question 4, where I expected residuals to be cenetred around 0, showing linear growth in life expectancy.

Exercise 4: Make a boxplot (or violin plot) of model residuals vs. continent.

```
sns.violinplot(x="continent", y="residuals", data=data)

plt.xlabel("Continent")
plt.ylabel("Residuals")
plt.title("Residuals vs. Continent")
```

```
Text(0.5, 1.0, 'Residuals vs. Continent')
```



Question 9: Is there a dependence between model residual and continent? If so, what would that suggest when performing a regression analysis of life expectancy across time?

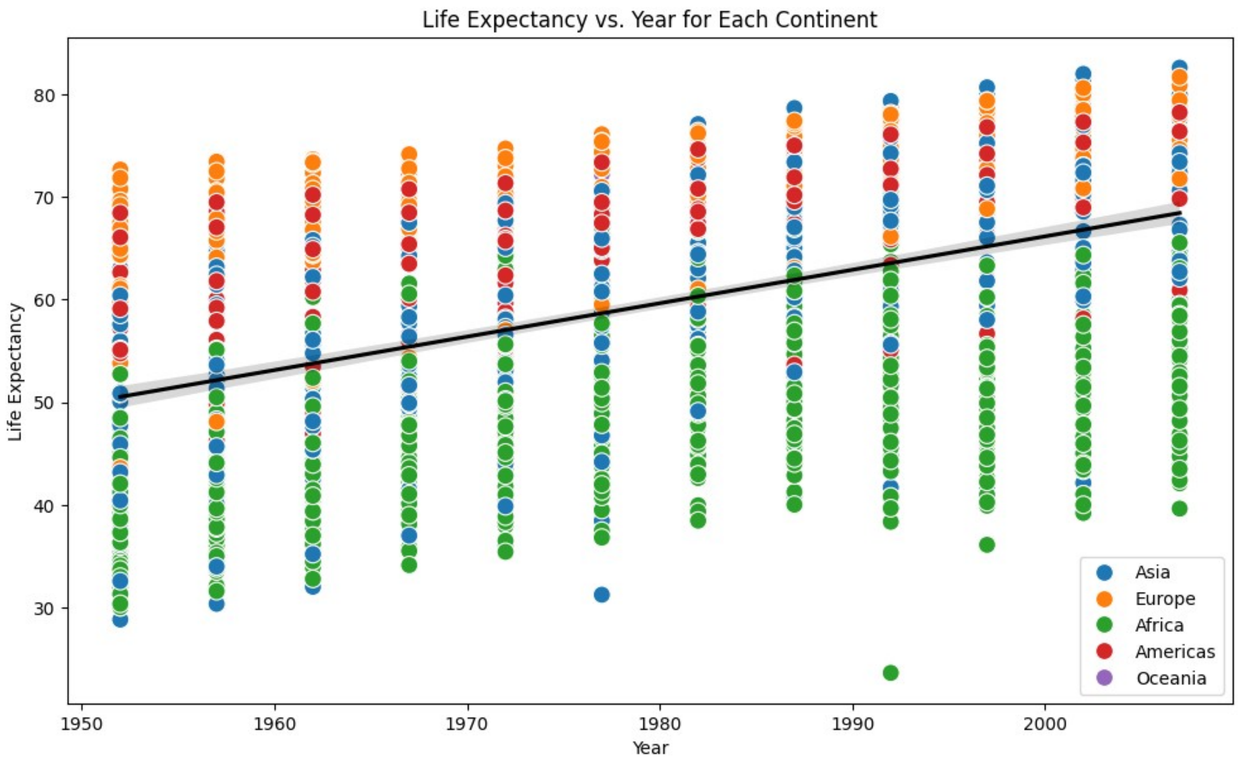
According to the violin plot, we can see a dependence between model residual and continent. Residuals for Africa tend to be negative, Asia seems to be centered near 0, and Europe, Americas, and Oceania tend to be positive. This tells us that the relationship between year and life expectancy is not the same across continents, and that the linear model overlooked continent as a significant varible in this relationship.

Exercise 5: As in the Moneyball project, make a scatter plot of life expectancy vs. year, grouped by continent, and add a regression line.

```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='year', y='lifeExp', hue='continent', data=data,
s=100)
```

```
sns.regplot(data=data, x='year', y='lifeExp', scatter=False,
color='black')
plt.legend(fontsize=10)
plt.tight_layout()

plt.xlabel("Year")
plt.ylabel("Life Expectancy")
plt.title("Life Expectancy vs. Year for Each Continent")
```

```
Text(0.5, 1.0, 'Life Expectancy vs. Year for Each Continent')
```



Question 10: Based on this plot, should your regression model include an interaction term for continent and year? Why?

According to the plot, all continents have different starting life expectancy. However, the life expectancy could als be changing at different rates in each continent. This would warrant the inclusion of an interaction term for continent and year, as it would let the effect of year differ by continent.

Exercise 6: Fit a linear regression model for life expectancy including a term for an interaction between continent and year. You may import statsmodel.formula.api to run ordinary least squares for this.

```
model = smf.ols('lifeExp ~ continent + year + continent:year',
data=data).fit()
```

```
print(model.summary())
```

```
                            OLS Regression Results                            
======================================================================
========
Dep. Variable:                lifeExp   R-squared:
0.693
Model:                            OLS   Adj. R-squared:
0.691
Method:                 Least Squares   F-statistic:
424.3
Date:                Sun, 09 Nov 2025   Prob (F-statistic):
0.00
Time:                        17:23:23   Log-Likelihood:
-5771.9
No. Observations:                1704   AIC:
1.156e+04
Df Residuals:                    1694   BIC:
1.162e+04
Df Model:                           9

Covariance Type:            nonrobust

======================================================================
========================
                                 coef    std err          t      P>|t|
[0.025      0.975]
----------------------------------------------------------------------
------------------------
Intercept                   -524.2578     32.963    -15.904      0.000
-588.911    -459.605
continent[T.Americas]       -138.8484     57.851     -2.400      0.016
-252.315     -25.382
continent[T.Asia]           -312.6330     52.904     -5.909      0.000
-416.396    -208.870
continent[T.Europe]          156.8469     54.498      2.878      0.004
49.957     263.737
continent[T.Oceania]         182.3499    171.283      1.065      0.287
-153.599     518.298
year                           0.2895      0.017     17.387      0.000
0.257       0.322
continent[T.Americas]:year     0.0781      0.029      2.673      0.008
0.021       0.135
continent[T.Asia]:year         0.1636      0.027      6.121      0.000
0.111       0.216
continent[T.Europe]:year      -0.0676      0.028     -2.455      0.014
-0.122      -0.014
continent[T.Oceania]:year     -0.0793      0.087     -0.916      0.360
```

```
-0.249      0.090
======================================================================
========
Omnibus:                       27.121   Durbin-Watson:
0.242
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
44.106
Skew:                          -0.121   Prob(JB):
2.65e-10
Kurtosis:                       3.750   Cond. No.
2.09e+06
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 2.09e+06. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

Question 11: Are all parameters in the model significantly different from zero? If not, which are not significantly different from zero?

According to the OSL Regression Results, all models except Oceania are signifficantly different from zero. Since the P-Values of Oceania are greater than the common alpha value of 0.05, it means that we do not have evidence that Oceania differs in baseline life expectancy from the other continents.

Question 12: On average, by how much does life expectancy increase each year for each continent? (Provide code to answer this question by extracting relevant estimates from model fit).

```
coef = model.params

print(f"Africa: {coef["year"]}")
print(f"Americas: {coef['year'] +
coef["continent[T.Americas]:year"]}")
print(f"Asia: {coef['year'] + coef["continent[T.Asia]:year"]}")
print(f"Europe: {coef['year'] + coef["continent[T.Europe]:year"]}")
print(f"Oceania: {coef['year'] + coef["continent[T.Oceania]:year"]}")

Africa: 0.28952926304617876
Americas: 0.3676509370646101
Asia: 0.45312240390059466
Europe: 0.22193214452372828
Oceania: 0.2102723776239897
```

According to the above calculations, life expectancy increases over time in all continents, but at different rates. Oceania and Europe have the lowest average annual increase in life expectancy at ~0.21 and ~0.22 years respectively. Asia has the fastest growth rate at ~0.45, while Africa and Americas are in the middle.

Exercise 7: Perform an F-test that compares how well two models fit your data:

(a) the linear regression models from Exercise 2 (only including year as a covariate) and

(b) Exercise 6 (including interaction between year and continent).

The F-test is obtained by calling fvalue from the results of exercises 2 and 6.

```
model_e2 = smf.ols(formula="y ~ X", data=data).fit()
model_e6 = smf.ols('lifeExp ~ continent + year + continent:year',
data=data).fit()

f_value_e2 = model_e2.fvalue
f_value_e6 = model_e6.fvalue

comparison = model_e6.compare_f_test(model_e2)
comparison # (f_value, p_value, df_diff)
```

```
(346.5535276625867, 0.0, 8.0)
```

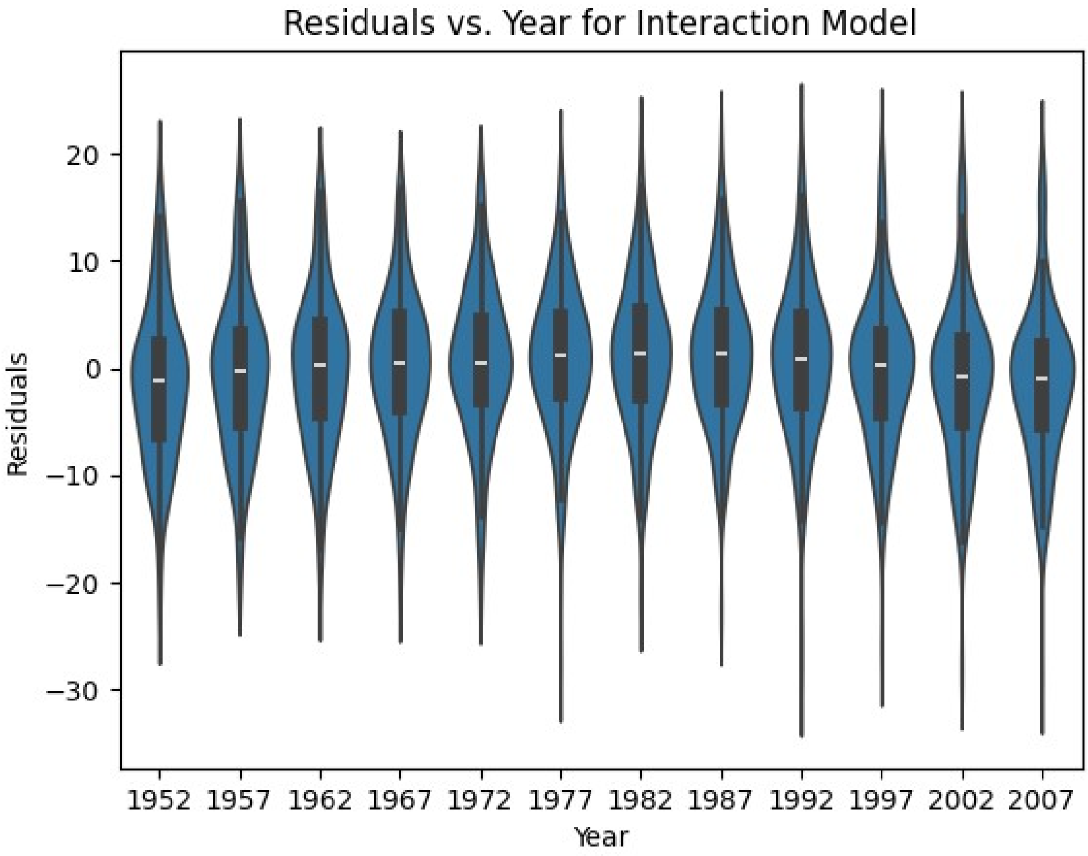
Question 13: Is the interaction model significantly better than the year-only model? Why?

The F-test comparing the two models shows an F-Value of 346.55 and P-Value of 0.0. These statistics indicate that the model including continent and the interaction term for continent and year provides a significantly better fit than the model using year alone.

Exercise 8: Make a residuals vs. year violin plot for the interaction model. Comment on how well it matches assumptions of the linear regression model. Do the same for a residuals vs. fitted values model.

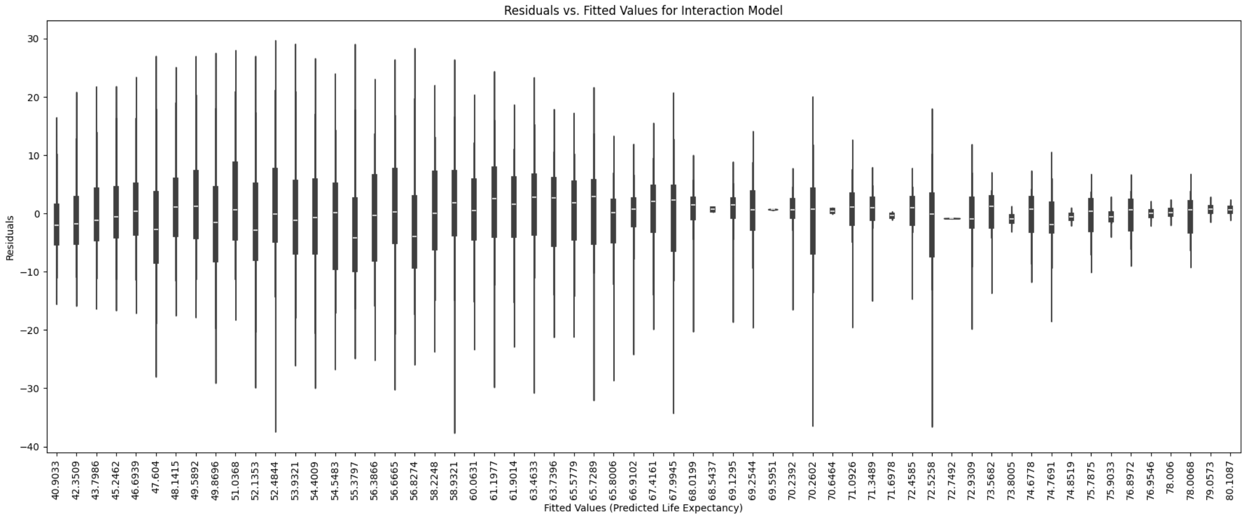
```
data["residuals"] = model_e6.resid

sns.violinplot(x="year", y="residuals", data=data)

plt.xlabel("Year")
plt.ylabel("Residuals")
plt.title("Residuals vs. Year for Interaction Model")
```

```
Text(0.5, 1.0, 'Residuals vs. Year for Interaction Model')
```



```
data["fitted"] = model_e6.fittedvalues.round(4)

plt.figure(figsize=(22, 8))
plt.tight_layout()
sns.violinplot(x="fitted", y="residuals", data=data)
plt.xticks(rotation=90)

plt.xlabel("Fitted Values (Predicted Life Expectancy)")
plt.ylabel("Residuals")
plt.title("Residuals vs. Fitted Values for Interaction Model")
```

```
Text(0.5, 1.0, 'Residuals vs. Fitted Values for Interaction Model')
```



The residuals vs. year plot shows that the residuals are centered around zero, but the spread of the residuals changes over time, indicating that the variance is not constant over time. The residuals vs. fitted values plot shows that the distribution of residuals narrows as fitted values increase, which also indicates variance that is not constant over time. In some years, some countries show more or less prediction error than others. While the models do capture the general trends, these patterns show that the model's errors aren't evenly spread out, and the data points are not completely independent from each other.

How My Code Works

My code first looks at the overall relationship between life expectancy and year. Then, a more complex analysis is done, where a linear regression uses year to predict life expectancy, then checks the model by examining how the errors, or residuals, behave across years and continents. These checks show that different continents follow different patterns, so the code fits an interaction model that allows each continent to have its own trend. The code then compares the two models using an F-Test to show that the interaction model shows how the distribution of life expectancy shifts upward over the years. Finally, the code plots visualizations for residuals vs. year and residuals vs. fitted values for the interaction model, showing that there is an uneven variance of residuals across years and countries.