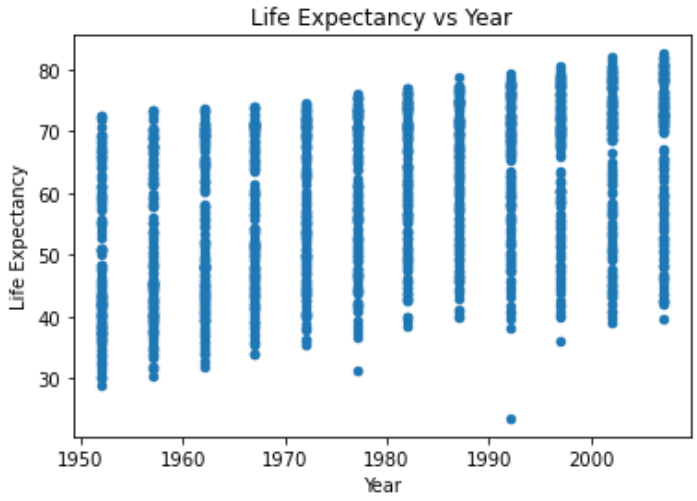
CMSC320 Project 3

Dmitri Kontchaev

Exercise 1

In this part we will simply create a scatter plot of the average of life expectancy in each country across time.

In [ ]:
```
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv("gap.tsv", sep='\t')
data.plot.scatter(x='year',y='lifeExp')
plt.xlabel("Year")
plt.ylabel("Life Expectancy")
plt.title("Life Expectancy vs Year")
plt.show()
```



Question 1

The general trend seems to be that life expectancy is increasing over the years since both the lower bound and upper bound of life expectancy generally tend to increase each year. Simply by looking at the graph, the relationship does appear to be linear.

Question 2

The distribution of life expectancy across countries for individual years changes with the years. At first, the distribution is skewed towards more countries having lower life expectancies with a unimodal shape. As time goes on, the distribution becomes skewed towards more countries having higher life expectancies. For a range of years (around 1957-1977) the distribution is slightly bimodal. At no point does the distribution appear entirely symmetrical.

Question 3

I would reject the null hypothesis as there does appear to be a correlation between year and life expectancy.

Question 4

A violin plot of the residuals would likely look similar to the original violin plot with more negative valued residuals in the earlier years and more positive valued residuals in the later years.

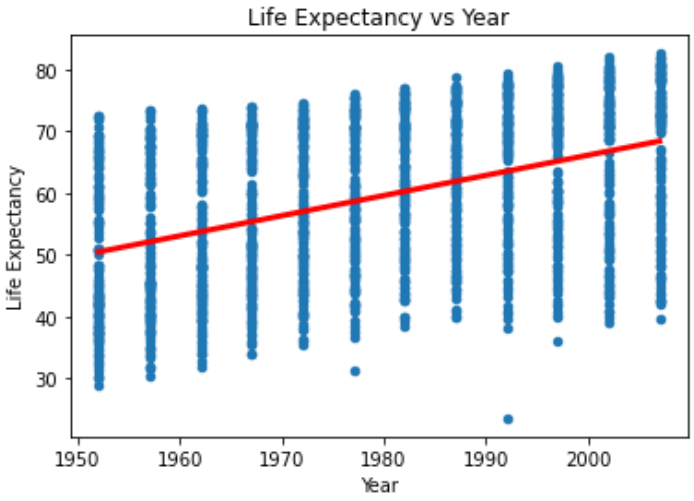
Question 5

The assumptions of the linear regression model suggest that the violin plot of the residuals would show symmetry around the mean residual value of 0. The residual violin plot will also show constant variance of the residuals as the years go on.

Exercise 2

In this part we create a linear regression model using Scikit-Learn and Statsmodels

In [ ]:
```
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
import statsmodels.formula.api as smf
import numpy as np

reg = LinearRegression()
X = np.array(data['year']).reshape(-1, 1)
y = np.array(data['lifeExp']).reshape(-1, 1)

reg.fit(X, y)
y_pred = reg.predict(X)

data.plot.scatter(x='year',y='lifeExp')
plt.plot(X, y_pred, color='red',
         linewidth=3)
plt.xlabel("Year")
plt.ylabel("Life Expectancy")
plt.title("Life Expectancy vs Year")
plt.show()

coeff = reg.coef_[0]
intercept = reg.intercept_

model = smf.ols(formula='lifeExp ~ year', data=data).fit()
print(model.summary())
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                lifeExp   R-squared:                       0.190
Model:                            OLS   Adj. R-squared:                  0.189
Method:                 Least Squares   F-statistic:                     398.6
Date:                Sun, 30 Apr 2023   Prob (F-statistic):           7.55e-80
Time:                        23:12:21   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
year           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 6

The average yearly increase in life expectancy would be the coefficient of regression of our model since the independent variable is years and the dependent variable is life expectancy. This value is 0.3259, so on average life expectancy increase by 0.3259 years around the world every year.

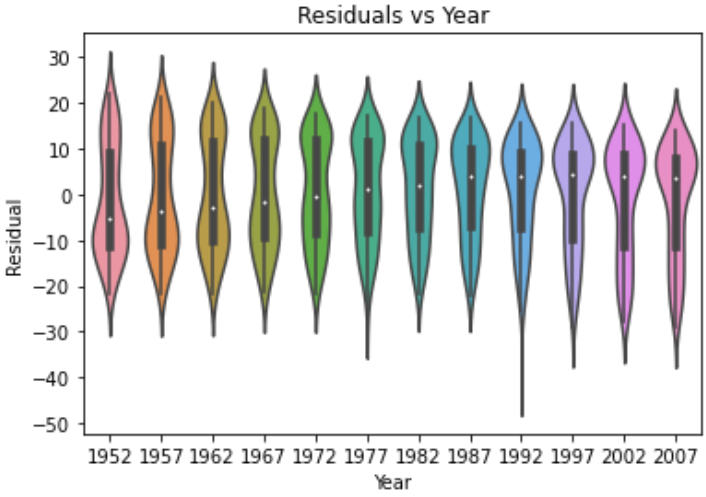
Question 7

Yes, I do reject the null hypothesis of no relationship between year and life expectancy. I reject the null hypothesis because our linear model shows there is a linear relationship between year and life expectancy and the p-value of this coefficient of linear regression is incredibly low at 7.55e-80, so with a significance level of 0.05, we can reject the null hypothesis.

Exercise 3

In this part we create a violin plot of the residuals from the model in the previous section

In [ ]:
```
import seaborn as sns
data['residuals'] = data['lifeExp'] - (intercept + coeff*data['year'] )
sns.violinplot(x='year', y='residuals', data=data)
plt.xlabel("Year")
plt.ylabel("Residual")
plt.title("Residuals vs Year")
plt.show()
```



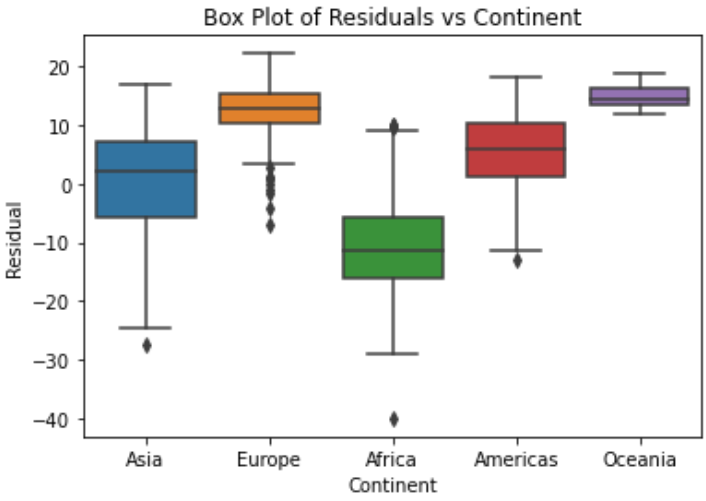
Question 8

Yes, the violin plot of the residuals matches my expecations in that it looks similar to the original violin plot, and there were more negative residuals in the earlier years and more positive residuals in the later years.

Exercise 4

In this part we create a boxplot of the model residuals for each continent

In [ ]:
```
sns.boxplot(y='residuals', x='continent', data=data)
plt.xlabel("Continent")
plt.ylabel("Residual")
plt.title("Box Plot of Residuals vs Continent")
plt.show()
```



Question 9

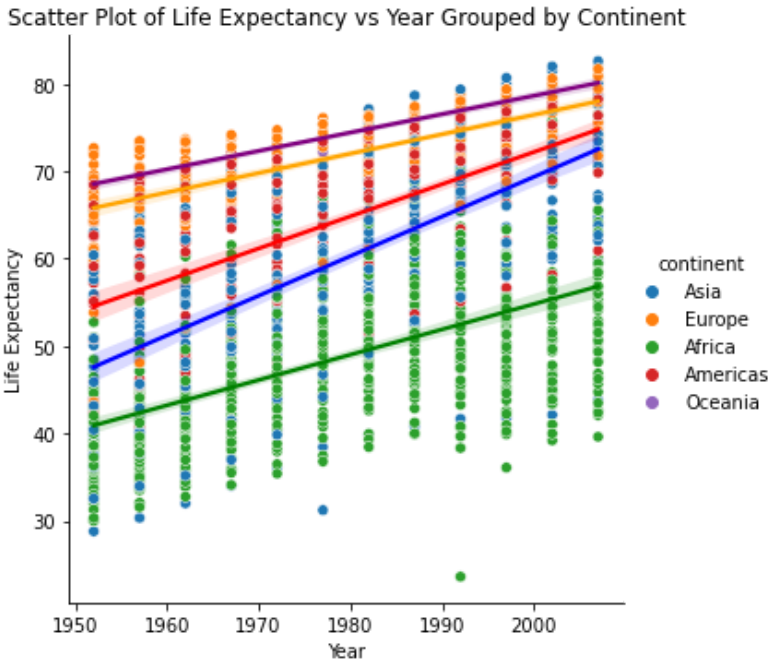
Yes, there appears to be a dependence between model residual and continent since the box plots for different continents have different shapes and lie at different heights on the residual axis. This suggests that the year is not the only variable that affects life expectancy, and continent/location also likely plays a role.

Exercise 5

In this part we recreate the scatter plot we had before of life expectancy vs year, but this time, the dots are colored by their respective continent, and each continent has a regression line shown.

In [ ]:
```
cont_groups = data.groupby('continent')
sns.relplot(data=data, x='year', y='lifeExp', hue='continent')
sns.regplot(data=cont_groups.get_group('Africa'), x='year', y='lifeExp', scatter=False, color='gr
sns.regplot(data=cont_groups.get_group('Asia'), x='year', y='lifeExp', scatter=False, color='blue
sns.regplot(data=cont_groups.get_group('Europe'), x='year', y='lifeExp', scatter=False, color='or
sns.regplot(data=cont_groups.get_group('Americas'), x='year', y='lifeExp', scatter=False, color='
sns.regplot(data=cont_groups.get_group('Oceania'), x='year', y='lifeExp', scatter=False, color='p
```

```
plt.xlabel("Year")
plt.ylabel("Life Expectancy")
plt.title("Scatter Plot of Life Expectancy vs Year Grouped by Continent")
plt.show()
```



Question 10

Yes, the regression model should include an interaction term for continent AND year because we have already shown there is a relationship between life expectancy and year, but this plot shows there is very likely a relationship between continent and life expectancy, since there are fairly clear boundaries of one continent's life expectancy for a year, and another continent's. Additionally, the trend lines show different slopes with different intercepts.

Exercise 6

In this part we create a new model which includes a term for the interaction between continent and life expectancy

In [ ]:
```
data['continent'].unique()
model2 = smf.ols(formula='lifeExp ~ continent * year', data=data).fit()
print(model2.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, 30 Apr 2023   Prob (F-statistic):               0.00
Time:                        23:12:24   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

The method used here to construct the regression model with an interaction term for continents uses one of the continents as a reference. In our case, Africa is being used as a reference, which means the effect of being in Africa on life expectancy has a regression coefficient of 0 in this model. This is because the continents are a categorical variable and a reference is needed, but none of the other parameters are 0.

Question 12

Code:

In [ ]:
```
inc_africa = model2.params['year']
inc_americas = model2.params['continent[T.Americas]:year'] + inc_africa
inc_asia = model2.params['continent[T.Asia]:year'] + inc_africa
inc_europe = model2.params['continent[T.Europe]:year'] + inc_africa
inc_oceania = model2.params['continent[T.Oceania]:year'] + inc_africa

print("Africa: increase of " + str(inc_africa) + " years in life expectancy per year")
print("Americas: increase of " + str(inc_americas) + " years in life expectancy per year")
print("Asia: increase of " + str(inc_asia) + " years in life expectancy per year")
print("Europe: increase of " + str(inc_europe) + " years in life expectancy per year")
print("Oceania: increase of " + str(inc_oceania) + " years in life expectancy per year")
```

```
Africa: increase of 0.2895292630445184 years in life expectancy per year
Americas: increase of 0.36765093706285346 years in life expectancy per year
Asia: increase of 0.45312240389899094 years in life expectancy per year
Europe: increase of 0.22193214452203344 years in life expectancy per year
Oceania: increase of 0.2102723776221585 years in life expectancy per year
```

Exercise 7

In [ ]:
```
print("F value of model with only year: " + str(model.fvalue))
print("F value of model with year and continent: " + str(model2.fvalue))
```

```
F value of model with only year: 398.6047457117627
F value of model with year and continent: 424.27290234006927
```

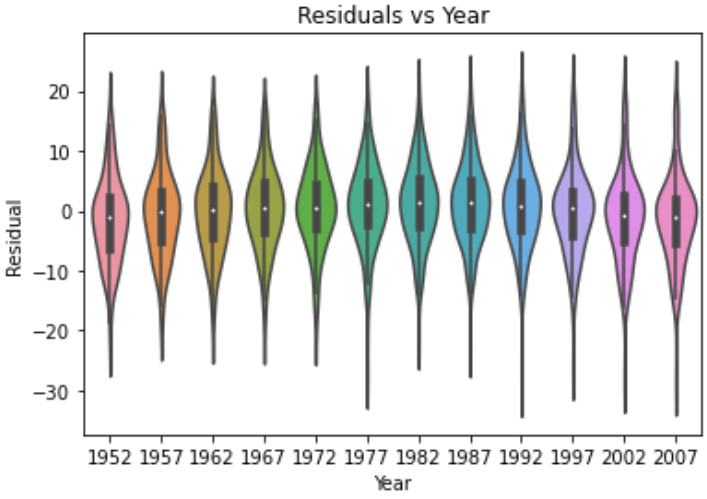
Question 13

Yes, the interaction model is significantly better than the year only model as can be seen by the much larger f value for the second model.

Exercise 8

For this part, we will make a residual vs year violin plot for the interaction model. This will be done by adding the residuals from the interaction model to the original datafram and using seaborn to create a violin plot.

In [ ]:
```
import seaborn as sns

data['residuals2'] = model2.resid
sns.violinplot(x='year', y='residuals2', data=data)
plt.xlabel("Year")
plt.ylabel("Residual")
plt.title("Residuals vs Year")
plt.show()
```



This violin plot of the residuals match the assumptions of the linear regression model well. The residuals seem to be fairly symmetrical around a value of 0, and the variance of the residuals seems to be fairly consistent. This suggests that the fitted values would be more accurate using this model than they would be using the previous year-only model.