Assigned Task - Epidemiology Internship Track (Python)

Prepared by: Melanie Atieno Omondi.
Date: 26/04/2025.

- As an initial task at the Center for Epidemiological Modeling and Analysis, you are presented with a dataset from an epidemiological research project, "Analyzing Regional Trends in Influenza-Like Illness (ILI) in Kenya: A Quantitative Epidemiology Case Study."
- The study collected data on year (2023-2024), the epidemiological week (epi_week), county, age categories (age_group), percentage of outpatient visits due to ILI (ili_percentage), and the estimated population for that age group in that county (population).
- The objective is to evaluate temporal and county-specific ILI trends and interpret the findings to inform public health decisions

In [3]:
```python
# Importing Pandas library for querying
import pandas as pd

# Importing visualization libraries for visualizing the data.
import matplotlib.pyplot as plt
import seaborn as sns

# Defining the dataframe.
df = pd.read_csv('Epi_Task_Data.csv')

# Checking for null values to ensure the data is complete.
print(df.isnull().sum())
```

```
year              0
epi_week          0
county            0
age_group         0
ili_percentage    0
population        0
dtype: int64
```

Summary: The dataset is complete (no missing values) and ready for analysis.

In [79]:
```python
# Reading the first three rows.
print (df.head(3))
```

```
   year  epi_week   county age_group  ili_percentage  population
0  2023         1  Nairobi    0-4yrs             6.1        3452
1  2024         1  Nairobi    0-4yrs             3.2        3658
2  2023         2  Nairobi    0-4yrs             4.5        3043
```

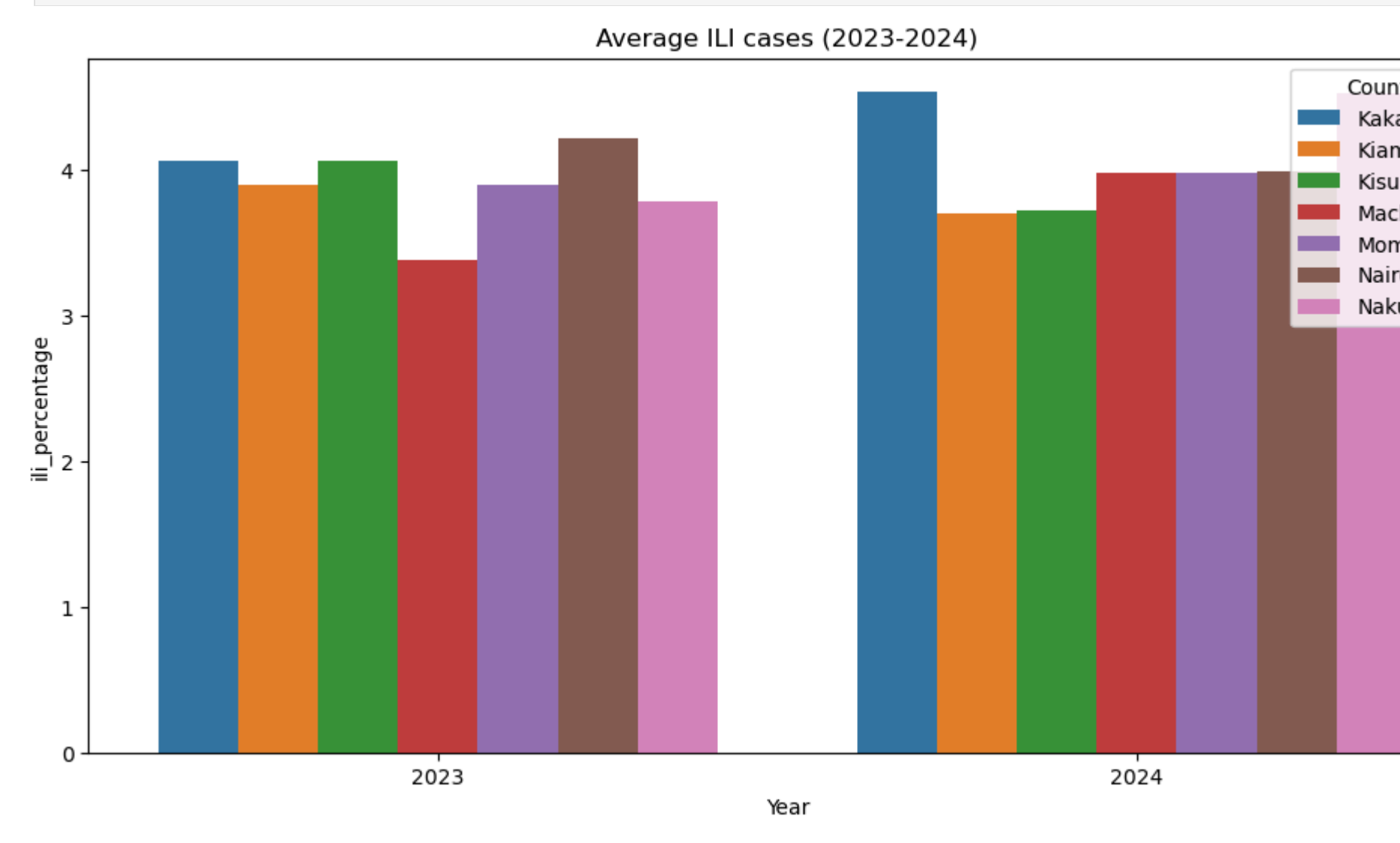
Section A: Descriptive analysis

i. Compute a table showing the mean ILI percentage per county per year.

In [7]:
```python
# Grouping the data by county and year, then calculating the average influenza-like illness cases(%)
avg_ili = df.groupby(['county', 'year'])['ili_percentage'].mean().reset_index()

print(avg_ili[['county', 'year', 'ili_percentage']])
```

```
      county  year  ili_percentage
0   Kakamega  2023           4.060
1   Kakamega  2024           4.540
2     Kiambu  2023           3.895
3     Kiambu  2024           3.705
4     Kisumu  2023           4.065
5     Kisumu  2024           3.725
6   Machakos  2023           3.380
7   Machakos  2024           3.975
8    Mombasa  2023           3.895
9    Mombasa  2024           3.985
10   Nairobi  2023           4.220
11   Nairobi  2024           3.990
12    Nakuru  2023           3.780
13    Nakuru  2024           4.530
```

In [90]:
```python
# Visualizing the data
plt.figure(figsize = (12, 6))
sns.barplot(data = avg_ili, x = 'year', y = 'ili_percentage', hue = 'county')

# Customizations
plt.title('Average ILI cases (2023-2024)')
plt.xlabel('Year')
plt.ylabel('ili_percentage')
plt.legend(title = 'County')
plt.savefig('barplot.png')
plt.show()
```

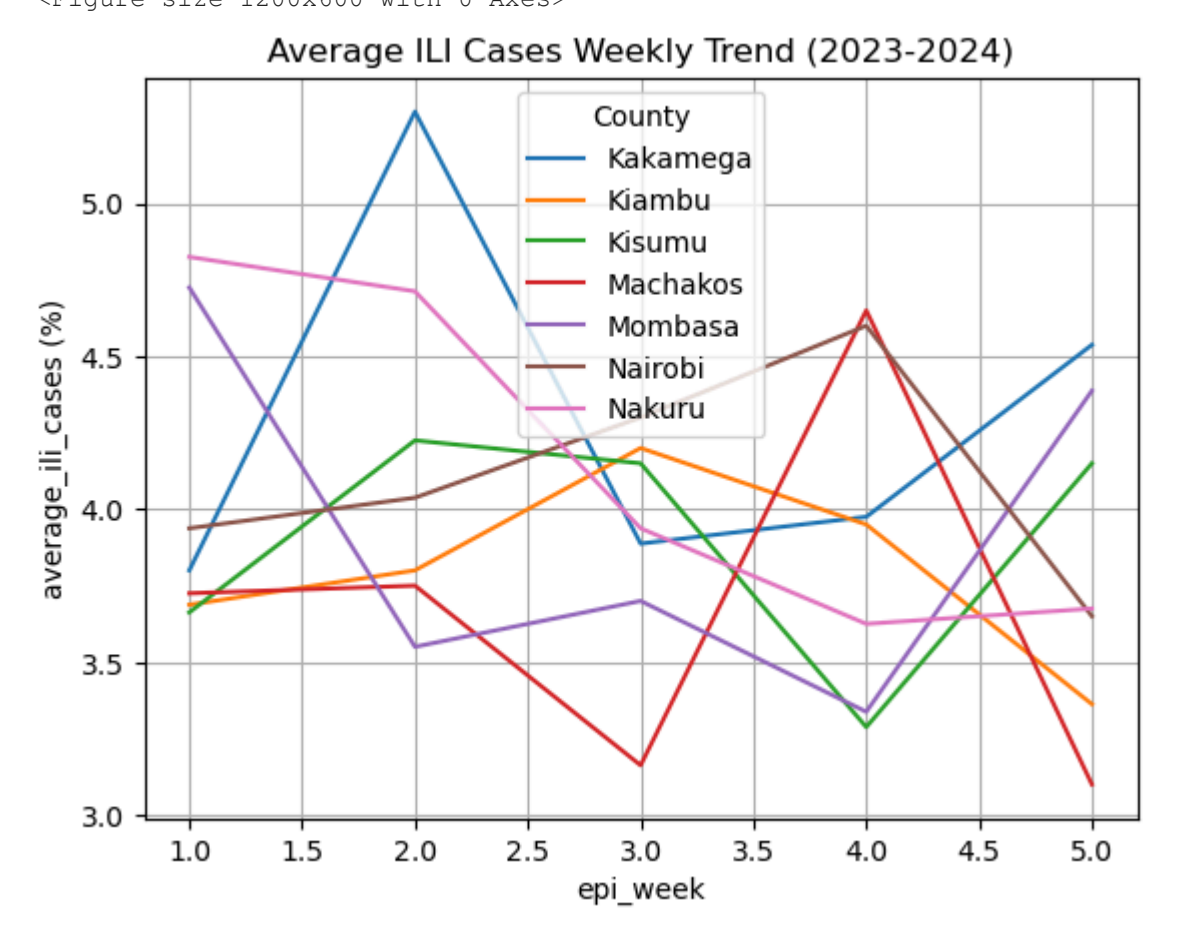


Summary: In 2023, the highest average ili cases were reported in Nairobi at 4.2%, while the lowest were in Machakos at 3.4%. In 2024, Kakamega led, experiencing a surge from 4% to 4.5%, while the lowest cases were in Kiambu at 3.7%. - There was an observed increase in ili cases in Kakamega, Machakos, Mombasa and Nakuru in 2024 from 2023, while cases in Kiambu, Kisumu and Nairobi seemed to have dropped slightly, this could possibly be due to improved healthcare access in these areas or adequate public sensitization.

ii. A plot of ILI weekly trends identifying the peak ILI weeks across counties and describe in 3-5 sentences the seasonal pattern of ILI cases.

In [92]:
```python
# Grouping the weekly trend data by county and epi_week to aggregate ili_percentage (average).
wt = df.groupby(['county', 'epi_week'])['ili_percentage'].mean().reset_index()

print (wt)

# Pivoting the data to reshape it for plotting (creating a pivot table)
pivot_data = wt.pivot(index = 'epi_week', columns = 'county', values = 'ili_percentage')

# Visualizing the data
plt.figure(figsize = (12, 6))
pivot_data.plot()

# Customizations
plt.title('Average ILI Cases Weekly Trend (2023-2024)')
plt.xlabel('epi_week')
plt.ylabel('average_ili_cases (%)')
plt.legend(title='County')
plt.grid(True)
plt.savefig('wk_trends.png')

plt.show()
```

```
      county  epi_week  ili_percentage
0   Kakamega         1          3.8000
1   Kakamega         2          5.3000
2   Kakamega         3          3.8875
3   Kakamega         4          3.9750
4   Kakamega         5          4.5375
5     Kiambu         1          3.6875
6     Kiambu         2          3.8000
7     Kiambu         3          4.2000
8     Kiambu         4          3.9500
9     Kiambu         5          3.3625
10    Kisumu         1          3.6625
11    Kisumu         2          4.2250
12    Kisumu         3          4.1500
13    Kisumu         4          3.2875
14    Kisumu         5          4.1500
15  Machakos         1          3.7250
16  Machakos         2          3.7500
17  Machakos         3          3.1625
18  Machakos         4          4.6500
19  Machakos         5          3.1000
20   Mombasa         1          4.7250
21   Mombasa         2          3.5500
22   Mombasa         3          3.7000
23   Mombasa         4          3.3375
24   Mombasa         5          4.3875
25   Nairobi         1          3.9375
26   Nairobi         2          4.0375
27   Nairobi         3          4.3000
28   Nairobi         4          4.6000
29   Nairobi         5          3.6500
30    Nakuru         1          4.8250
31    Nakuru         2          4.7125
32    Nakuru         3          3.9375
33    Nakuru         4          3.6250
34    Nakuru         5          3.6750
<Figure size 1200x600 with 0 Axes>
```



Overall Summary: - There are notable peaks in weeks 2 (Kakamega: 5.3% - highest) and 4 (Machakos and Nairobi), with drastic drops noted in week 3 (Machakos), week 4 (Mombasa and Kisumu), and week 5 (Machakos: 3.1% - lowest). For the first three weeks, Nakuru recorded a continuous pattern of decreasing average cases from 4.8% to 3.6% until a slow increase of up to 3.7% in the last surveillance week. Nairobi and Kiambu have had a steady increase in average cases within the first two weeks, with a decrease experienced in the 3rd week (Kiambu) and the last week (Nairobi). Detailed Summary: Epi-Week 1-2: Kakamega recorded a rapid increase in cases peaking in week 2, compared to Nairobi and Kiambu, which had slow, steady increases, and Machakos had the fewest recorded cases, while Mombasa experienced a notable drop compared to all other counties. Epi-Week 2-3: Kakamega and Machakos had a drastic drop in cases, while cases in Nairobi and Kiambu increased significantly from the previous week. There is an observed steady drop in cases in Kisumu and a continuous drop in Nakuru from the previous week. Epi-Week 3-4: Machakos experienced a surge in cases from the previous week, peaking in week 4, as well as a continuous increase in cases in Nairobi, with its highest peak at week 4 compared to the previous two weeks. Kakamega experienced a slow and steady increase despite the drastic drop in cases within the previous week. Kisumu and Mombasa experience a drop in cases, with Nakuru having an even lower count of cases compared to the previous two weeks. Epi-Week 4-5: Kakamega, Mombasa, and Kisumu experience a rapid increase in cases, and Nakuru experiences a slow rise in cases compared to the previous three weeks. Machakos, Nairobi, and Kiambu experience a rapid drop in cases.

Section B: Computing epidemiological measures

i. Calculate the incidence rates per 100,000 population across any three counties.

In [31]:
```python
# Grouping data by county to aggregate population and ILI percentage (totals)
grouped = df.groupby(['county']).sum()

# Filtering for Nairobi, Kisumu, and Mombasa
counties = grouped.loc[['Nairobi', 'Kisumu', 'Mombasa']]

# Calculating the incidence rate for the selected counties
# Incidence rate = Total new cases per county (2023-2024) / Total population (estimated to be 100,000)
counties['incidence_rate'] = counties['ili_percentage'] * counties['population'] / 100000

print(counties[['incidence_rate']])
```

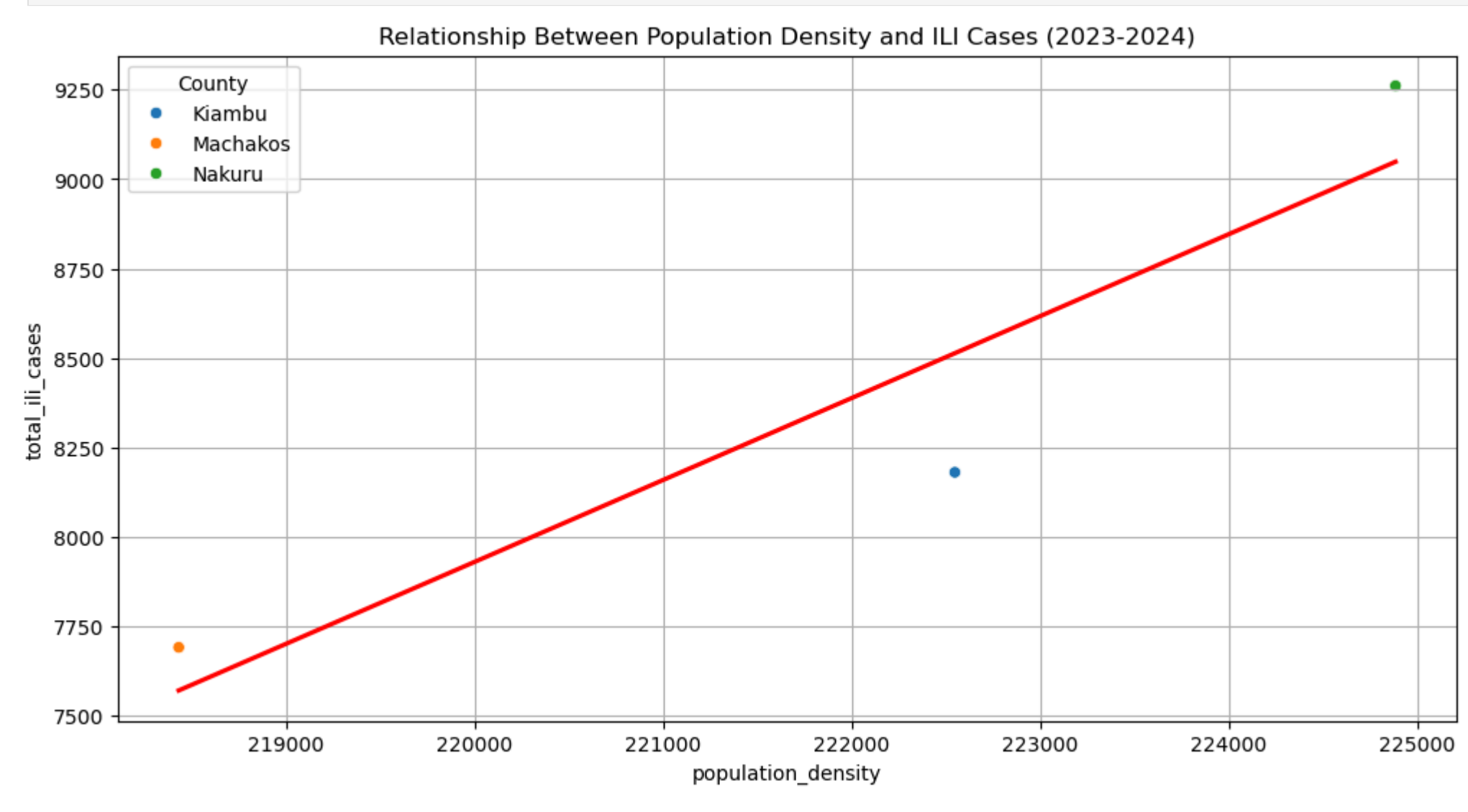
```
         incidence_rate
county
Nairobi      270.558908
Kisumu       306.481970
Mombasa      351.879824
```

Summary: Between 2023-2024, Mombasa had the highest incidence rate, 352 new cases per 100,000 people, suggesting a higher influenza-like illness disease burden compared to Nairobi and Kisumu. This could maybe be caused by various determinants such as population density.

ii. Using any statistical method you are familiar with, compare ILI percentages across any three counties.

My statistical analysis approach: - Chosen counties (Machakos, Nakuru, Kiambu) Step 1. I will use a scatter plot to visually observe the relationship between ili cases and population density within the counties and formulate a hypothesis. Step 2. I will use regression to quantify the relationship observed to validate/reject the hypothesis.

In [94]:
```python
# Step 1: Visualizing the relationship and hypothesis formation.

# Selecting counties and finding total absolute cases
sc = df[df['county'].isin(['Machakos', 'Nakuru', 'Kiambu'])].copy()
sc ['total_cases'] = sc ['ili_percentage'] * sc['population']/100

# Grouping the data by selected counties for aggregation (totals)
new_counties = sc.groupby(['county']).sum()

# Visualizing the scatterplot
plt.figure(figsize = (12,6))
sns.scatterplot(data = new_counties, x = 'population', y = 'total_cases', hue = 'county')

# Adding a trend line
sns.regplot(data = new_counties, x = 'population', y = 'total_cases', scatter = False, color = 'red', ci = None)

# Customizations
plt.title ('Relationship Between Population Density and ILI Cases (2023-2024)')
plt.xlabel ('population_density')
plt.ylabel ('total_ili_cases')
plt.legend (title = 'County')
plt.grid (True)
plt.savefig('scatter_plot.png')

plt.show()
```



Summary: From the scatter plot above, there seems to be a linear positive relationship between population density and influenza-like illness. Nakuru, with the highest population density, recorded higher cases, followed by Kiambu, and Machakos recorded the lowest cases while having the lowest population density. * Hypothesis : An increase in population density increases ili cases in Kiambu, Machakos and Nakuru.

In [83]:
```python
# Step 2: Quantifying the positive linear relationship observed and validating/rejecting the hypothesis.

# Importing regression library
import statsmodels.api as sm

# Defining independent and dependent variables
X = new_counties[['population']]  # Predictor value: Population
y = new_counties['total_cases']   # Target value: Total ILI Cases

# Adding constant (intercept term)
X = sm.add_constant(X)

# Fitting the regression model
model = sm.OLS(y, X).fit()

print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            total_cases   R-squared:                       0.868
Model:                            OLS   Adj. R-squared:                  0.736
Method:                 Least Squares   F-statistic:                     6.571
Date:                Sat, 26 Apr 2025   Prob (F-statistic):              0.237
Time:                        13:28:43   Log-Likelihood:                -20.677
No. Observations:                   3   AIC:                             45.35
Df Residuals:                       1   BIC:                             43.55
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -4.244e+04   1.98e+04     -2.141      0.278   -2.94e+05    2.09e+05
population     0.2290      0.089      2.563      0.237      -0.906       1.364
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.256
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.479
Skew:                          -0.629   Prob(JB):                        0.787
Kurtosis:                       1.500   Cond. No.                     1.85e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.85e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
C:\ProgramData\anaconda3\Lib\site-packages\statsmodels\stats\stattools.py:74: ValueWarning: omni_normtest is not valid with less than 8 observations; 3 samples were given.
  warn("omni_normtest is not valid with less than 8 observations; %i "
```

Summary: - Coefficient (coef = 0.229): This suggests for every additional person in the population, there is a 0.229 increase in total ili cases. - R-squared value (R = 0.868): This suggests that 86.8% of the variability in the total ili cases is attributed to the population density, showing population density is a key driver in influencing ili cases. - P value (P = 0.237): This suggests no strong evidence to validate the hypothesis due to a statistical insignificance of the population since the value is greater than 0.05. * Even though there was no strong evidence to conclude an increase in population density increases total ili cases, there could be a limitation due to the limited sample size of only 3 counties (Machakos, Kiambu, and Nakuru), expanding the sample size might improve significance.

Section C: Communicating results

In [97]:
```python
from IPython.display import Image, display

display(Image(filename='Intro_slide.png'))
display(Image(filename='Slide_1.png'))
display(Image(filename='Slide_2.png'))
display(Image(filename='Slide_3.png'))
display(Image(filename='Slide_4.png'))
```

Analyzing Regional Trends in Influenza-Like Illness (ILI) in Kenya: A Quantitative Epidemiology Case Study.

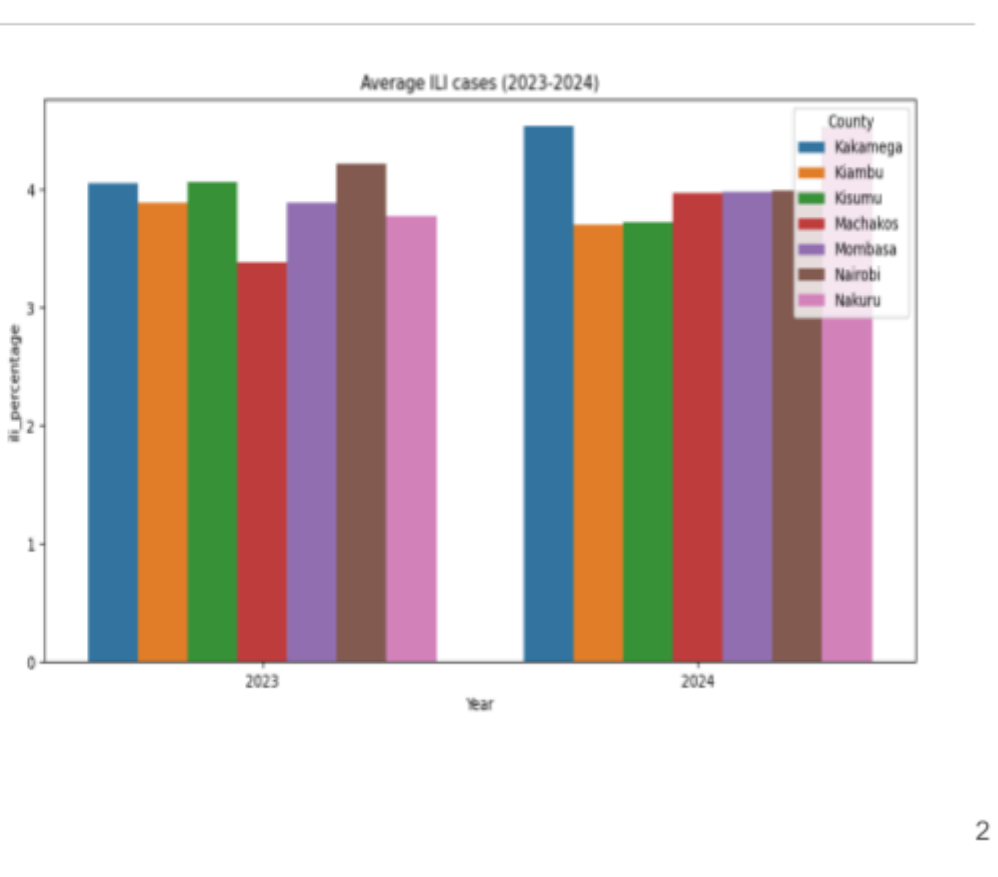
Epidemiological Analysis

Prepared by Melanie Omondi.

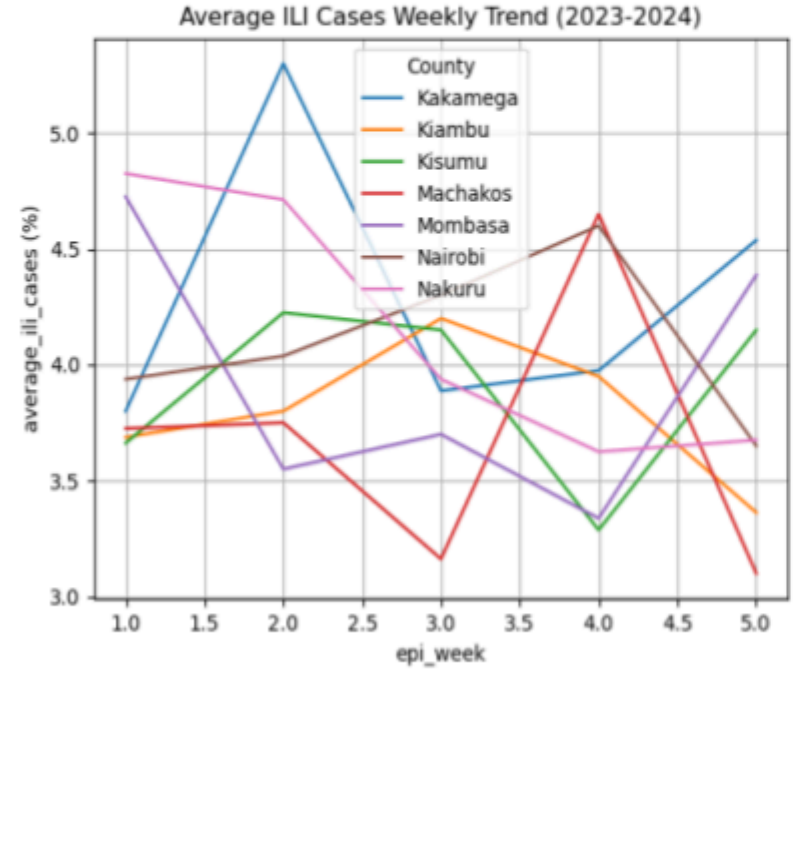
Key Findings

Average ILI cases

- In 2023, the highest average ILI cases were reported in Nairobi at 4.2%, while the lowest were in Machakos at 3.4%.
- In 2024, Kakamega led, experiencing a surge from 4% to 4.5%, while the lowest cases were in Kiambu at 3.7%.
- There was an observed increase in ILI cases in Kakamega, Machakos, Mombasa, and Nakuru in 2024 from 2023, while cases in Kiambu, Kisumu, and Nairobi seemed to have dropped slightly. This could possibly be due to improved healthcare access in these areas or adequate public sensitization.



Key Findings



Weekly ILI Trend in Counties

- There are notable peaks in weeks 2 (Kakamega: 5.3% - highest) and 4 (Machakos and Nairobi), with drastic drops noted in week 3 (Machakos), week 4 (Mombasa and Kisumu), and week 5 (Machakos: 3.1% - lowest).
- For the first three weeks, Nakuru recorded a continuous pattern of decreasing average cases from 4.8% to 3.6% until a slow increase of up to 3.7% in the last surveillance week.
- Nairobi and Kiambu have had a steady increase in average cases within the first two weeks, with a decrease experienced in the 3rd week (Kiambu) and the last week (Nairobi).

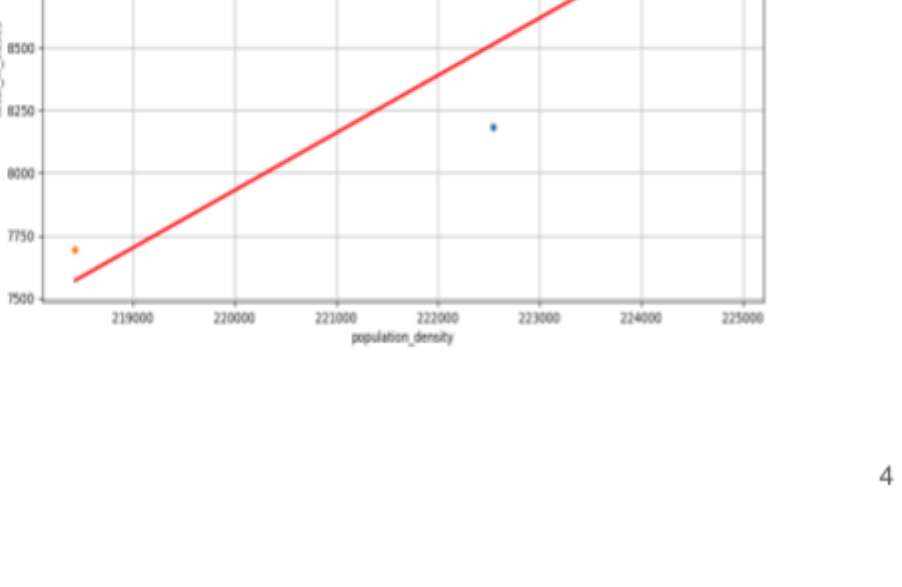
Key Findings

Incidence rates per 100,000.

- Between 2023 and 2024, Mombasa had the highest incidence rate, 352 new cases per 100,000 people, suggesting a higher influenza-like illness disease burden compared to Nairobi and Kisumu. This could maybe be caused by various determinants such as population density.

Relationship between Population Density and ILI Cases.

- Even though there was no strong evidence to conclude an increase in population density increases total ILI cases, there could be a limitation due to the limited sample size of only 3 counties (Machakos, Kiambu, and Nakuru), expanding the sample size might improve significance.



Public Health Recommendations

Suggested Interventions.

- Since population density can be attributed to 87% of ILI cases as observed from sampling Machakos, Nakuru and Kiambu, counties with higher populations should be considered high risk and prioritized in ILI control interventions such as vaccine distribution for equitable and impactful population health outcomes.
- Promote ILI sensitization in counties with low ILI cases to improve prevention measures and lower incidence rates. An example could include community health education outreach programs on hygiene practices in primary schools.
- Establish early detection strategies in observed high-incidence counties to avoid resurgence, such as introducing ILI rapid screening sites across sub-counties.

In [ ]: