Research Question 1: Did the 2012 Olympic games increase tourist activity in East London region?

Create the model in time series and make sure to talk about how we overcome cyclicality, trends, and correlated observations by doing good synthetic control.

Data Cleaning:

- Remove duplicates, null values and missing data from UK_international_visits
- Check for and remove outliers
- Normalize numeric columns
- Standardize column names

Feature Engineering:

- Create features representing cultural engagement, museums visits and road lengths
- Log transform visits and spend to normalize distributions

Models:

- Use Regression (? Not sure what methods to try here) to predict visits and spending due to their count nature
- Evaluate using mean absolute error and residual sum of squares
- Fit models on pre-2012 data and generate synthetic control predictions for 2012+

Synthetic Control:

- Create a "synthetic London" using regions without Olympics as controls
- Compare actual post-2012 data to synthetic predictions to estimate causal effect

Placebo Tests:

- Apply synthetic control to other 11 UK regions as placebos
- Compare placebo effects to London's to check for statistical significance

Research Question 2: Were these gains distributed evenly across society?

Data Cleaning:

- Remove duplicates, NA values and outliers from london_earnings_by_borough data
- Normalize wages using UK_inflation data to adjust for inflation
    - Convert string columns to categorical data types
    - Standardize column names

Feature Engineering on borough-level:

- Create features:
    - From london_economic_activity data:
        - Workforce participation rate by gender
    - From london_taxpayer_income data:
        - Average income by borough
        - Median income by borough
    - From london_development_database:
        - Number of new houses completed by borough
    - From land_registry data:

        - Median house prices by borough
    - From voa_average_rent_borough:
        - Average rent prices by borough
    - From london_public_houses data:
        - Number of pubs by borough

Classify boroughs by level economic level

Models:

- Fit separate linear regression models to predict:
    - Women's wages by borough
    - Men's wages by borough
    - Part time workers' wages by borough
    - Full time workers' wages by borough
- Borough-level economic and labor market features as predictors

Synthetic Control:

- For each workforce category, generate synthetic control predictions for Newham post-2012
- Compare actual Newham wage changes to synthetic predictions to estimate causal effect

Placebo Tests:

- Apply synthetic control method to other 31 London boroughs as placebos
- Compare placebo effects to Newham's effects to determine statistical significance
- Only significant effects for particular workforce groups would indicate uneven distribution of gains

Overall, the methodology aims to determine:

- If wage gains from the Olympics varied across workforce categories
- Which groups saw the largest/smallest effects
- If any groups were actually negatively impacted: evaluate East vs West London

Research Question 3: What was the impact of the London Olympic games on London’s transport infrastructure and underground? Which regions were impacted the most?

Data Cleaning:

- Clean and normalize data in london_underground_activity and london_underground_station_info datasets
- Remove duplicates, nulls and outliers
- Standardize column names and data types

Feature Engineering:

- Create features:
    - Population density for each borough
    - Average income per borough
    - Total ticket sales per borough

Models:

- Try linear regression, lasso and random forest to predict:
    - Underground station entries/exits per borough
    - Transport infrastructure spending per borough

Evaluate models using:

- Mean squared error
- R2 score

Tune hyperparameters to optimize model performance

Analyze Results:

- Compare actual vs predicted values for:
    - Underground ridership
    - Transport infrastructure spending
- Identify boroughs with:
    - Largest increase in underground ridership
    - Highest transport infrastructure spending
- Interpret coefficient weights to determine influential features

Check predictions against boroughs known to be impacted for validation

The features will help analyze:

- Change in underground entries/exits before and after the Olympics
- Which lines and local authorities saw the largest impact
- Which stations saw the largest increase in ridership
- Which lines and local authorities were most impacted

Discuss whether there is a high correlation between what we observe on the data and economic factors that were prevalent in that time period.