EECS-731

TRAVELING WITH CONFIDENCE?

Dana Almansour, Michael Bechtel, Brandon Wheat

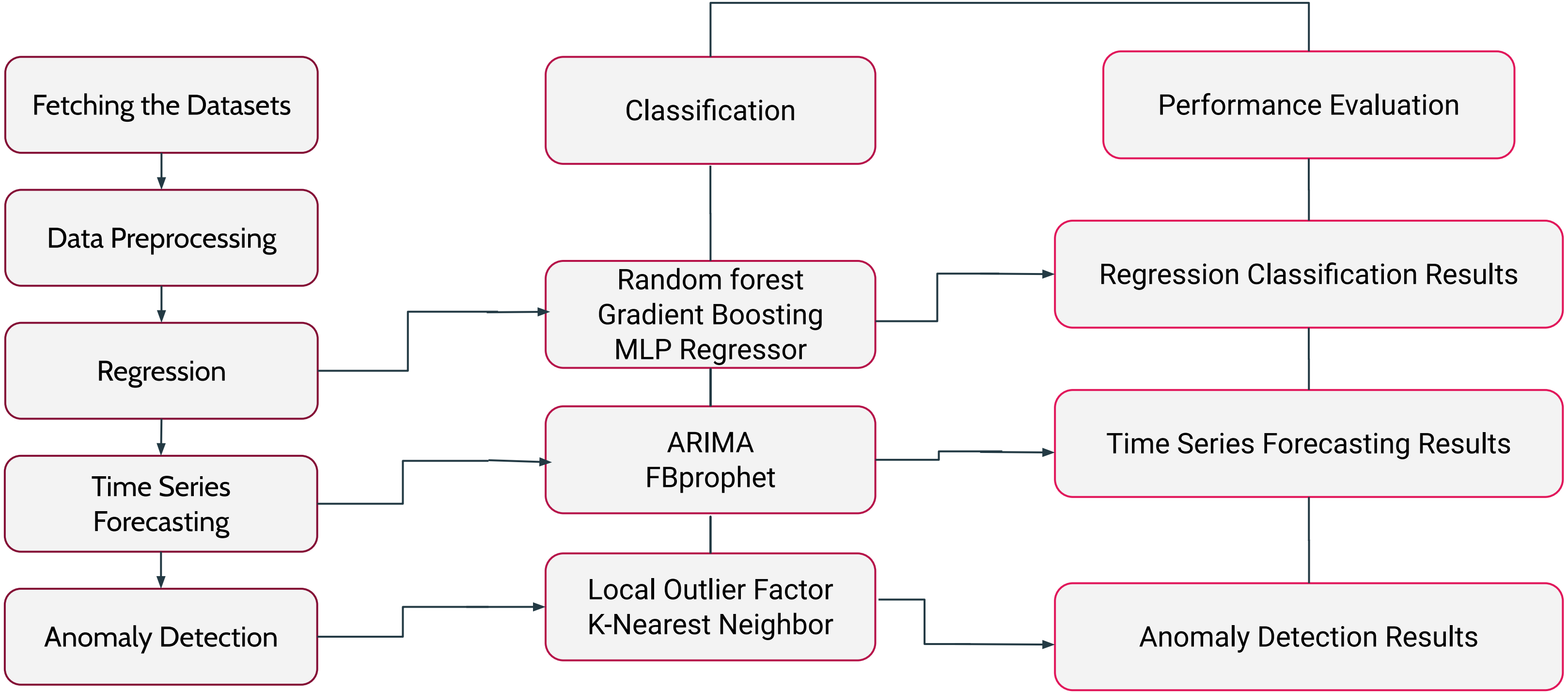
Outline

- Motivation
- Overview
- Data Sets
- Data Exploration
- Regression
- Time Series Forecasting
- Anomaly Detection
- Learning Outcomes

Motivation

- Industry: Public transportation (bus, train, etc.)
- International usage of public transportation has plummeted due to COVID
  - For example, usage in the US was down ~60% nationally in March (FTA)
- There have been recent rebounds, but no full recoveries
- We wanted to potentially answer the following questions:
  - Do COVID cases correlate to transit usage?
  - Given previous trends, what will future trends look like?
  - Are there any usage outliers in a given time period?

Overview



Fetching the Datasets

Data Preprocessing

Regression

Time Series Forecasting

Anomaly Detection

Classification

Random forest
Gradient Boosting
MLP Regressor

ARIMA
FBprophet

Local Outlier Factor
K-Nearest Neighbor

Performance Evaluation

Regression Classification Results

Time Series Forecasting Results

Anomaly Detection Results

Data Sources

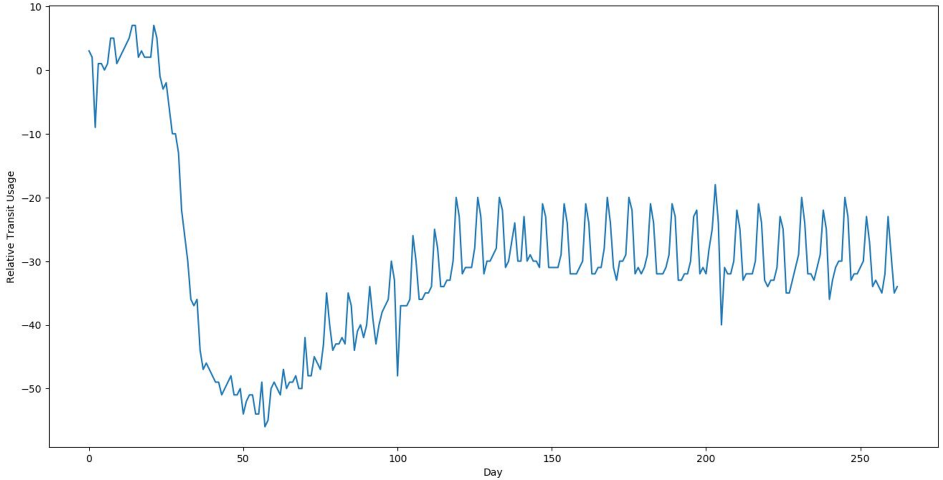
- Daily transit hub usage: Google's Community Mobility Reports
  - Provides several per-country datasets (135 countries total)
  - Each set contains data for several mobility metrics (grocery stores, retail, parks, etc.)
  - We use the transit hub metric provided in these reports
  - Source: https://www.google.com/covid19/mobility/
- Daily COVID cases: Our-World-in-Data COVID-19 Dataset
  - Provides general COVID trends on a per-country basis
  - We use the number of new cases per day and the overall totals for each country
  - Source: https://ourworldindata.org/coronavirus-source-data

Exploratory Data Analysis

- Looked at data from 10 countries:
    - CA, DE, ES, FR, GB, IT, JP, KR, MX, US
    - When showing results, we anonymize the countries (A - J)
- All transit hub usage data had the same date range: 2/15/20 - 11/3/20
- Take COVID case data in the same range
    - All tested countries started reporting prior to 2/15

Example Transit Hub Usage

- Usage for Country A



- All values are percentages of transit hub usage relative to a "pre-COVID" baseline (1/3/20 - 2/6/20):
- Each tested country displays similar usage trends

Regression

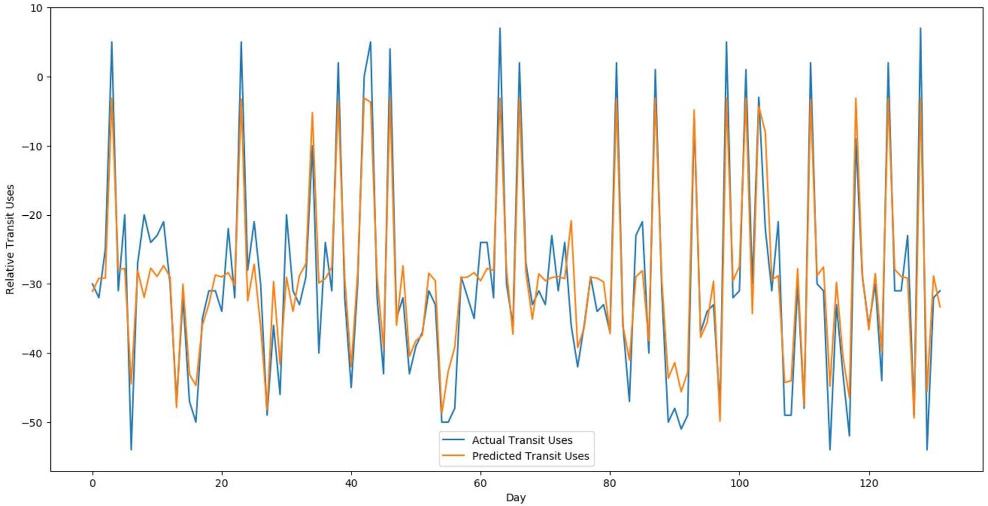
- How do COVID cases correlate to transit hub usage?
- Tested three Regression models on all country datasets
  - Random Forest
  - Gradient Boosting
  - MLP Regressor (Neural Network)
- For the MLP Regressor, better results were achieved by preprocessing COVID case data (MinMax)
  - The Random Forest and Gradient Boosting models performed better with the raw data.
- Approach: shuffle and split data into train and test sets (50/50)
- Use root mean squared error (RMSE) to evaluate performance

Regression Comparison

Country	Random Forest	Gradient Boosting	MLP Regressor
A	5.51	5.89	**4.86**
B	8.38	8.79	**6.37**
C	**5.62**	6.40	5.71
D	**10.45**	11.74	16.15
E	**9.23**	8.65	8.86

- Example: countries A - E
- No single model performed the best across all countries
  - While not shown, Gradient Boosting did perform the best for some countries (G, I)

Regression

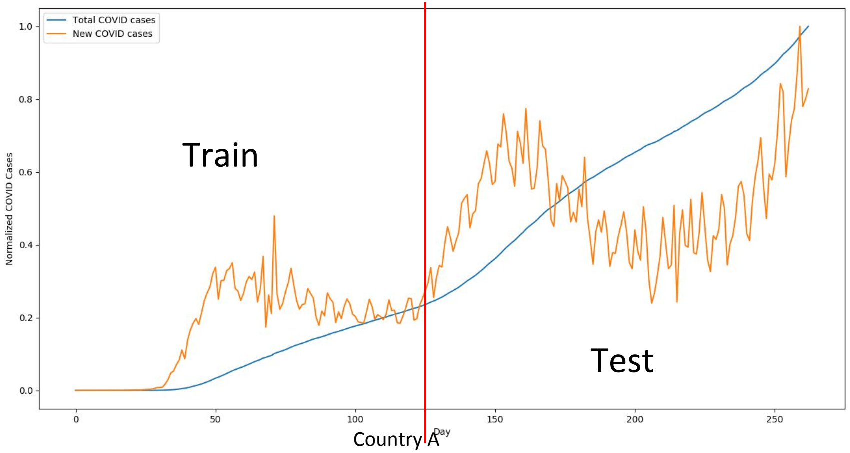


- Example: Country A
- Model: MLP Regressor
  - RMSE = 5.07

- COVID cases do roughly correlate to transit hub usage in given datasets.
  - This was evident across all tested countries.
- Are COVID cases an accurate indicator for transit hub usage?
  - Can this be used for forecasting?

Regression/Time Series

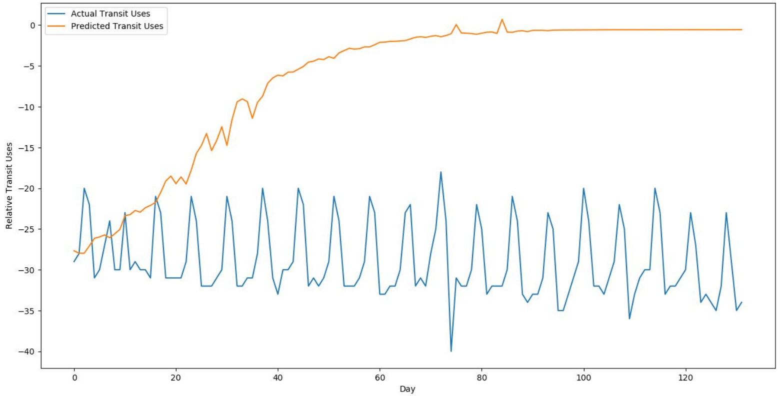
- Use the same approach, but don't shuffle the data → simulate future predictions



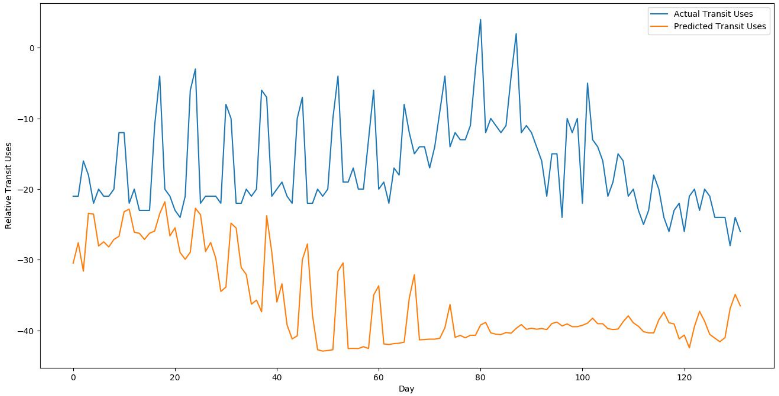
- For each country, use the regression model that performed the best on the shuffled data.
- Can the models still accurately predict transit hub usage?

Regression/Time Series

- Examples: MLP Regressor for countries A and B



Country A



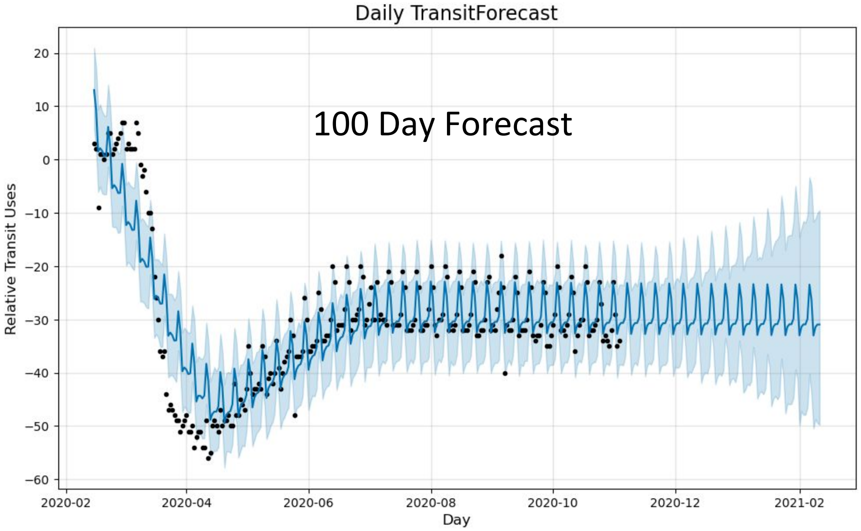
Country B

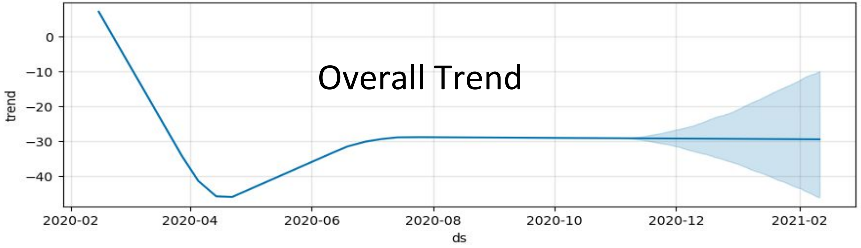
- No, the models can't use COVID cases to predict future transit hub trends
  - Predictions were different for each country, but none were ever close to the actual trends
- For each country, the same results occurred with other train/test splits (70/30, 90/10, etc.)

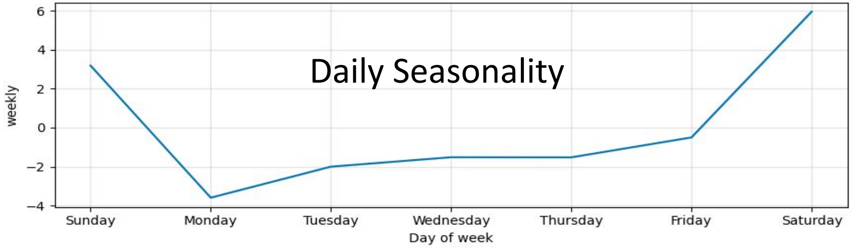
Time Series Forecasting

- Tested two time series forecasting models
  - FBProphet
  - ARIMA
- FBProphet is a tool developed by Facebook for time series forecasting, it considers three main components:
  - Seasonality
  - Trends
  - Events
- ARIMA has three main features:
  - Auto Regression - regresses a variable on past values of itself
  - Integration - reduces seasonality from a time series
  - Moving Average - removes random movements from a time series
- We use both models to predict 100 days in the future (11/4/20 - 2/12/21)

FBProphet

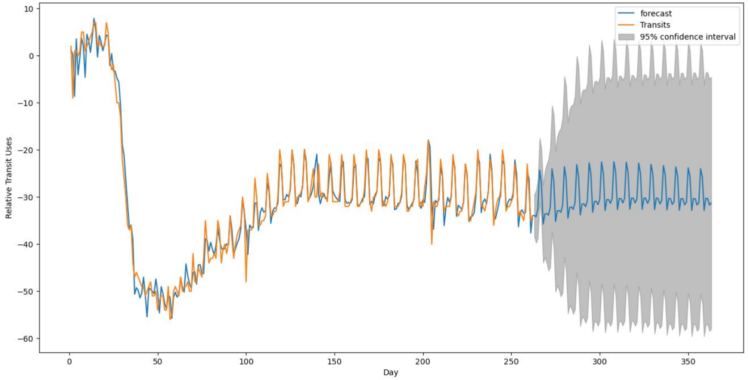




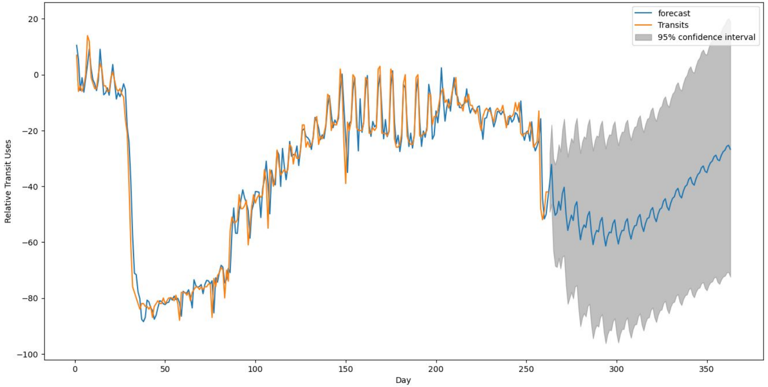


- For country A there is no negative or positive trend, which is reflected in the forecast
- For country A Transit is greatest on Saturdays and lowest on Mondays

ARIMA



Country A



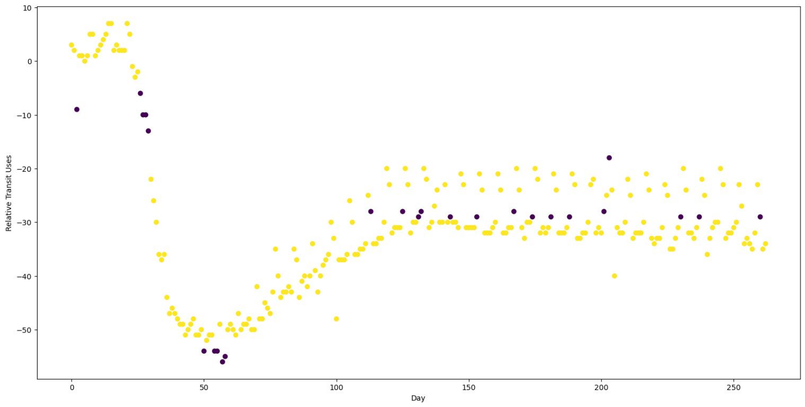
Country D

- For country A, ARIMA also predicts that usage will remain stagnant
  - For the most part, FBProphet and ARIMA had similar predictions
- All countries had different prediction patterns but stayed around the average
  - The model never predicted that a country would get back to pre-COVID usage

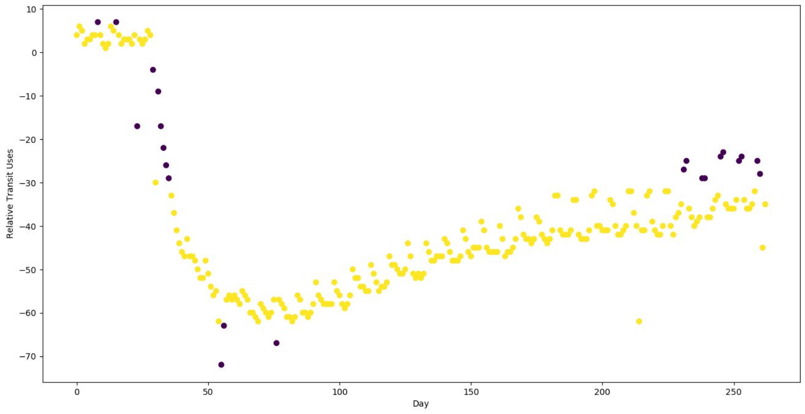
Anomaly Detection

- Tested two anomaly detection models on all our country datasets
    - Local Outlier Factor
    - K-Nearest Neighbor
- Local Outlier detection computes the local density deviation of a given point with respect to is neighbors
    - Data with much lower densities were considered outliers
- K-Nearest Neighbor measures a points distance from a clusters center
    - Data with much larger distances from a center were considered anomalies

Local Outlier Factor



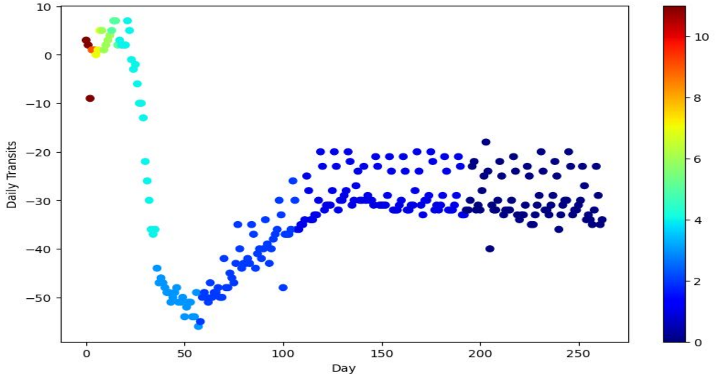
Country A



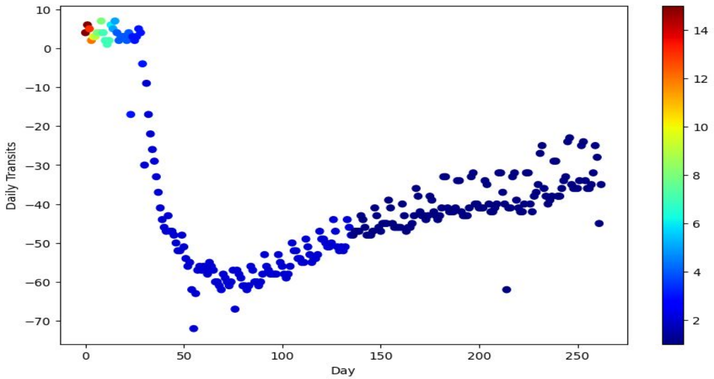
Country J

- The outliers for each country are represented by the purple data points
  - Since LOF >> 1, we can deduce that these are outliers present in this dataset which we are analysing.
- Since Local Outlier Factor takes into account the knn and ldr in a very efficient manner, we can deduce outlier here more efficiently than Knn clustering anomaly detection outlier.
- Some countries produce more usable results (country J)

K-Nearest Neighbors



Country A



Country J

- The points with the highest anomaly scores are colored red and the lowest are colored blue
- The value of k is to be kept optimal
  - High k values cause larger anomaly scores (overfitting)
  - Low k values cause lower anomaly scores (underfitting)
- In both cases, the anomalies occurred at the beginning of the timeline
  - This also happened in the other tested countries
- Which model is better at detecting anomalies for our data?

Learning Outcomes

- Regression
    - COVID cases and transit hub usage due correlate for a given time period
    - COVID cases can't be used to reliably predict future usage trends
- Time Series
    - Usage trends are likely to remain at their current levels (at least for the next 100 days)
    - Usage may increase in some countries but not by a significant margin
- Anomaly Detection
    - There are outliers present in the given time period and we have used two methods for its analysis
    - Local Outlier Factor is more efficient than Knn clustering for detecting outliers.
    - Some outlier trends are more helpful for addressing transit hub usage
- For many models, all countries displayed similar results and behaviors
    - No one country really stood out from the rest
- All results and visualizations can be found at: https://github.com/mbechtel2/EECS731-GroupProject