ISYE: Course Project

I selected the Boston School Buses project for analysis due to my prior experience as a high school teacher, which provided me with firsthand insights into the dynamics of school buses and the experiences of students who rely on them. This project aimed to, and successfully achieved, the optimization of school bus routes to enhance efficiency, reduce costs, and improve the overall student experience in transit. The analysis models I chose to reverse engineer for this study were K-Means, ARIMA, and SVM, each offering unique perspectives in addressing the complexities of route optimization. Together, these models provide a comprehensive approach: K-Means optimizes bus stop locations, ARIMA forecasts ridership changes, and SVM assists in student safety and experience, ensuring a holistic and efficient optimization of the school bus system.

The collection of data for the School Buses project would involve collaborating with various sources. Historical ridership and schedule data can be acquired from local transportation or school district authorities. Route and special events information is accessible through city event planners or transit databases. Weather, demographic, economic, and traffic data require sourcing from governmental databases and forecasting services. Data on competing transportation modes, road networks, and public transit details can be obtained from city planning offices and GIS platforms. Student home locations, community feedback, and legal regulations need collaboration with school districts, community surveys, and local government offices, ensuring compliance with privacy laws.

The refresh frequency for the School Buses project's data varies: Ridership and bus schedule data should be updated each school year or semester. Weather and traffic data require more frequent, ideally real-time or daily updates, particularly during variable weather seasons or in areas with changing traffic conditions. Updates for special events and public transportation network data should be made as needed, especially before major events or transit changes. Demographic and economic data can be refreshed annually, while student home locations and community feedback might be best updated at the start of each school year to reflect demographic shifts and community needs.

An annual review and update of the models, ideally aligned with each new school year, would be beneficial. More frequent updates may be necessary in response to significant changes in student population, road infrastructure, or based on user feedback. Additionally, the model should be capable of real-time adjustments to accommodate unforeseen events such as road closures, extreme weather conditions, or extraordinary circumstances like the school closures during the COVID-19 pandemic.

Geospatial Optimization Model:

Data:

- ➔ *Student Home Locations:* This is the primary dataset for clustering. It should include precise geographical coordinates (latitude and longitude) of each student's home.
- ➔ *Road Network Data:* Detailed maps of the road network, including types of roads, their conditions, speed limits, and any roadblocks or one-way streets. This information is crucial for route optimization.
- ➔ *Traffic Pattern Data:* Information about typical traffic conditions, including peak hours, average vehicle speeds, and congestion points. This data helps in planning routes that avoid traffic bottlenecks.
- ➔ *Public Transportation Network Data:* Existing bus routes, stops, and schedules should be considered to integrate the new stops efficiently into the current system.
- ➔ *Pedestrian Infrastructure Data:* Data on sidewalks, crosswalks, and pedestrian safety measures. This is vital for ensuring that the walking paths to and from the bus stops are safe for students.
- ➔ *Demographic and Socioeconomic Data:* Information about the student population, including age groups, socioeconomic status, and any special needs. This data can influence decisions about stop locations to serve the community better.
- ➔ *Weather and Environmental Data:* Long-term weather patterns and environmental conditions can affect route planning and stop placement, especially in regions with extreme weather.
- ➔ *Terrain and Topographical Data:* Understanding the local terrain is important, especially in hilly or mountainous areas, as it impacts the feasibility of certain routes and stop locations.
- ➔ *Legal and Regulatory Information:* Local regulations or laws that might affect where bus stops can be located or how buses can operate in certain areas.
- ➔ *Community Feedback and Input*: Data gathered from surveys or community meetings can provide insights into the needs and preferences of students and parents.
- ➔ *Historical Transportation Data:* Past data on bus usage, efficiency, and issues faced can provide valuable lessons for improving the system.

Model:

Clustering algorithm, K-Means, to group students based on their home locations. Each cluster can represent a potential consolidated bus stop location. This model focuses on optimizing the location of bus stops based on geographical information.

The ***K-Means algorithm*** partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean. The formula or process can be summarized in the following steps:

Initialization: Randomly select k distinct data points as the initial centroids.
Assignment: Assign each data point to the closest centroid based on the Euclidean distance.
The distance d between two points x and y in a plane with coordinates:
d(x, y) = sqrt((x_2 - x_1)^2 + (y_2 - y_1)^2)

Update: Recalculate the centroids as the mean of all points assigned to that cluster.
Repeat: Repeat steps 2 and 3 until the centroids no longer change significantly.

Result:

Ensure that the consolidated bus stops are optimized to save time and money:

- Cluster Centroids: The primary output will be the coordinates of the centroids that represent the optimal central points for the consolidated bus stops. Each centroid is calculated to be geographically central to the students' home locations within that cluster, optimizing for proximity.
- Cluster Assignments: Each student's home location will be assigned to one of the clusters. This information shows which students would use each proposed bus stop.
- Number of Clusters (Bus Stops): The number of clusters determined (like by the Elbow Method) indicates the number of bus stops proposed by the model.
- Cluster Sizes: The size of each cluster, i.e., the number of students assigned to each cluster, will be an output. This helps in understanding the demand at each bus stop and planning the bus capacity accordingly.
- Visual Representation: Often, the results are presented in a graphical format, such as a map showing the locations of all students with different clusters marked in different colors. The centroids (proposed bus stop locations) can also be marked on this map.
- Distance Metrics: information on the distances between each student's home and their assigned bus stop. This can be used to ensure that all students have reasonable access to their designated stops.
- Potential Savings Analysis: Depending on the sophistication of your model, it might include an analysis of the time and cost savings achieved by the proposed bus stop locations compared to the current system.

Demand Prediction Model:

Data:

- ➔ *Historical Ridership Data:* This is the most crucial dataset. It should include the number of passengers boarding and alighting at each stop, ideally broken down by time of day, day of the week, and season. This helps in identifying daily, weekly, and seasonal patterns in bus usage.
- ➔ *Bus Schedule Data:* Information about bus schedules, including frequency, timing, and changes over time. This helps in understanding how schedule changes impact ridership.
- ➔ *Route Information:* Data on the routes, including length, stops, and any changes made to the routes over time.
- ➔ *Special Events Data:* Information about local events (like sports games, concerts, or festivals) that could cause spikes in ridership.
- ➔ *Weather Data:* Historical weather data, as weather conditions can significantly impact public transportation usage.

- ➔ *School and Public Holiday Calendars:* School schedules and public holidays can significantly affect ridership patterns, especially in areas with a lot of student commuters.
- ➔ *Economic and Demographic Data:* Changes in the local economy, population demographics, and employment rates can influence public transportation usage trends.
- ➔ *Traffic Data:* Data on road traffic conditions, which can impact bus travel times and reliability.
- ➔ *Competing Transportation Modes Data:* Data on usage of other modes of transportation like subways, bicycles, rideshare services, which can offer insights into competition and complementary aspects affecting bus ridership.

Model:

The **ARIMA model** uses three components: AutoRegression (AR), Integration (I), and Moving Average (MA):

AutoRegressive (AR) Part: This part models the changing variable as a linear combination of its own previous values.
Integrated (I) Part: This represents the differencing of raw observations to make the time series stationary (i.e., mean, variance, and covariance are constant over periods).
Moving Average (MA) Part: This models the error of the model as a linear combination of error terms that occurred contemporaneously and at various times in the past.
Determine the order of ARIMA parameters (p, d, q) using:
Information criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion).

Result:

- Predict the demand for bus services at different locations and times, considering the number of students, school start times, and the presence of special needs children.
- Predict the number of students likely to use the bus at specific times and locations.
- Predicted Values: The primary output will be the forecasted values for each time point in the future. For instance, the model will provide a predicted number of riders for each day in the forecast period.
- Confidence Intervals: Confidence interval giving an estimated range of values which is likely to include the unknown population parameter, the estimated range being calculated from a given set of sample data.
- Graphical Representation: Time series plot showing historical data points, the predicted values, and possibly the confidence intervals around these predictions.

<u>Safety Assessment Models</u>

Data:

- ➔ *Crime Statistics:* Data on local crime rates, types of crimes, and their frequency. This information is crucial for assessing the safety of potential bus stop locations.

- ➔ *Traffic Data:* Information about traffic volume, speed limits, accident rates, and pedestrian incidents. High traffic areas might be riskier for bus stops, especially for children.
- ➔ *Street Lighting and Infrastructure:* Data on street lighting, presence of sidewalks, pedestrian crossings, and other infrastructure that contribute to the safety of a location.
- ➔ *Public Surveillance and Police Presence:* Information about areas under surveillance and the presence of law enforcement can be indicative of safety.
- ➔ *Environmental Hazards:* Data on environmental risks like proximity to industrial areas, pollution levels, or areas prone to natural disasters.
- ➔ *Accessibility Data:* Information on the accessibility of areas for individuals with special needs, which is important for inclusive planning.
- ➔ *Community Feedback:* Insights from surveys or community meetings about perceived safety and concerns of local residents.
- ➔ *Historical Accident Data:* Records of past accidents near potential bus stop locations can be a significant indicator of safety.
- ➔ *Land Use and Zoning Information:* Data on the surrounding land use (residential, commercial, industrial) and zoning regulations which might influence the safety and suitability of a location.
- ➔ *Public Amenities and Facilities:* Proximity to public amenities like parks, schools, and hospitals, which can impact the foot traffic and safety of an area.
- ➔ *School District Boundaries and Policies:* Information about school district boundaries and their policies regarding transportation can impact where stops should be located.
- ➔ *Weather and Climate Data:* In areas with extreme weather conditions, considering climate data can help in assessing the safety of a location across different seasons.

Model:

Support Vector Machines (SVM): Find a plane that best separates the classes in the feature space.
Hyperplane Equation:
$w_1x_1 + w_2x_2 + ... + w_nx_n + b = 0$,
where w is the weight vector and b is the bias.
Objective Function: SVM maximizes the margin between the two classes. The margin is defined by the distance between the hyperplane and the nearest data points of each class.

Result:

SVMs can be used for binary classification tasks, making them suitable for predicting safe or unsafe locations.

- Labels: each location will be labeled as 'safe' or 'unsafe'. This can be presented as a list or a column in a dataset.
- Support Vectors: Vectors that define the decision boundary. Understanding which data points are support vectors can give insights into the model's decision-making process.
- Decision Function Values: The decision function value indicates on which side of the plane the data point lies and the distance from it. Higher absolute values of the decision function indicate points that are farther from the decision boundary, implying a more confident classification.

- Performance Metrics: metrics like accuracy and the area under the ROC curve (AUC-ROC) can be used to assess its performance. These metrics provide insight into how well the model is performing overall.
- Confusion Matrix: This matrix shows the number of true positives, false positives, true negatives, and false negatives. It's a helpful tool to understand the model's performance in each class.

References

SAS Institute Inc. (n.d.). Optimizing bus routes helps funnel cost savings back to schools. Retrieved [2023-11-28], from https://www.sas.com/en_us/customers/boston-public-schools.html