# Data Scientist - Take Home Exercise NHAI

---

FASTag has allowed for centralized collection of toll data which has opened up various avenues of analytics for overall improvement of highway infrastructure in the country.

In this take-home exercise you've been given two datasets:

1. For a particular highway which has 4 toll plazas, you've been given the all toll transactions for an approximate 24-hour period.
2. You have been given Network Survey Vehicle's data (2 files) for the same highway stretch. NSV are specialised vehicles which run on highways to collect data on the condition of the pavement and furniture on the road. NHAI collects similar information for all the highways under NHAI on a routine basis.

Locations on a highway are identified through chainage, direction and lane depending on the data being collected. A chainage marks the location in meters from a reference point in a highway, the direction is in reference to the starting point. The LHS or RHS and Lane decides the location within a direction.

For the problem statement, you need to identify the following things from the data:

1. Visualize the highway alignment within the Jupyter Notebook along with the location and labels of the Toll Plazas.

2. There have been reports that people have been able bypass toll-plazas on entry and it is leading to losses of revenue. There are two toll-able segments in the data (AB/BA) and (CD/DC). The toll is to be a deducted at the entry to a segment. Sometimes the toll-plaza lets some people go without deducting their entry or people have been able to find an unofficial bypass allowing them to skip and entry gate.
    a. Investigate if there is any toll-leakage, if so at which segment/toll-plaza?
    b. Number of trips which were possibly bypassed toll?
    c. What is the possible revenue loss that is happening pertaining to those trips?

3. The speed of a particular vehicle on a particular stretch in a particular direction depends on multiple factors such as the traffic, vehicle type, condition of the road, direction etc.
    a. From the data available, identify average/median/quartile speeds of vehicles on AB and CD stretches in each direction.
    b. Does the speed vary by type of vehicle anywhere?

4. You have been provided data on the condition of the road LaneIRI values for each direction on the highway. Does the average speed calculated for all vehicles

combined depend on the LaneIRI in that segment and direction? Higher the IRI, the poorer the quality of the road.

5. [**BONUS**] Google and/or Bing has Traffic APIs/ which are able to give estimate travel time between sets of origin and destination points based on estimate traffic. For their APIs they may have a python library to help you call those APIs or you may call the APIs through request module. Within the same notebook, call either Google Traffic API or Bing Maps Traffic API and get the estimated travel time for AB, BA, CD and DC for anytime of the day and any day and compare it with the time as calculated from the toll-data.

**In any place if there is lack of clarity, take the best assumption, document the reasoning behind the assumption and proceed.**

**Please submit the following file on the** https://vacancy.nhai.org/ictpmu/ portal under the same login as used before for application**:**

1. Single Python notebook where you conduct the above analysis with clear documentation, visualizations etc on the approach taken for each question, code documentation and visualizations where necessary. If you have any doubts, please take the best assumption, document your thought process and move ahead. The IPYNB should be self-contained and well-documented and read in one flow.

You will be tested on:

1. Research/Understanding of the domain from reading the relevant documents
2. Python code's documentation and readability
3. All programming questions equally
4. Presentation skills within the notebook.