

# Data Scientist - Take Home Exercise NHAI

---

## Part One:

You have been given Network Survey Vehicle's data (2 files) for one entire 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.

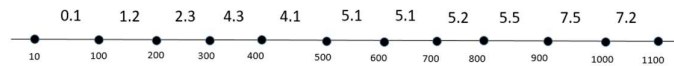
Locations on a highway are identified through chainage, direction and lane depending on the data being collected. A chainage marks the location in meters from a reference point in a highway, the direction is in reference to the starting point and decides LHS or RHS and Lane decides the location within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

For eg.

- Width of the shoulder is described between start chain-age and end chain-age and direction.
- Carriageway furniture such as information boards are having single chainage and direction.

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

1. Visualize the LaneIRI (roughness) as colour points depending on the value of IRI (choose the red-green scale but you'll need to research about IRI to figure what colour to use for low or high numbers. Green should represent a smooth surface) on a map within your Python notebook. Each 100m stretch in a lane should be represented by a single point. You may choose the lat-long of the start chainage of the record. Basemap can be any.
2. List of continuous stretches where the LaneIRI in the lane has the same integer value (0, 1 or 2, floor value). The idea is to identify continuous stretches of similar IRI condition. Once each stretch has been identified:
  - 2.1. Choose appropriate visualization to examine the relationship between the size of the stretches versus the common IRI of the said stretches and document insights if any.
  - 2.2. Tabulate the number of continuous stretches in each IRI bracket (0, 1, 2... so on)



| Stretches | Start | End  | Stretch IRI |
|-----------|-------|------|-------------|
| S1        | 10    | 100  | 0           |
| S2        | 100   | 200  | 1           |
| S3        | 200   | 300  | 2           |
| S4        | 300   | 500  | 4           |
| S5        | 500   | 900  | 5           |
| S6        | 900   | 1100 | 7           |

- Find the number of continuous stretches with 3 or more IRI and length greater than 300m. Within that list, find the longest stretch.
- Number potholes by DrainType in the increasing and decreasing direction separately. Both the datasets are recorded at chainage level but with different granularities. Take and record necessary assumptions to be able to combine the datasets.

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

- Single Python notebook where you conduct the above analysis with clear documentation on the approach taken for each question and code documentation. If you have any doubts, please take the best assumption, document your thought process and move ahead.
- Small presentation deck with the insights and deliverables asked above.

You will be tested on:

- Research/Understanding of the domain from reading the relevant documents
- Python code's documentation and readability
- All 4 programming questions equally
- Presentation

## Part Two:

Imagine, NHAI wants to identify the quantum of tree cover surrounding the highways in a quick manner to get an approximate sense of where on the highway there is green cover and there isn't. You have been given the task as a Data Scientist to pilot that on a particular highway whose GIS alignment is known. What is the cheapest and most effective way to do the same? Please provide steps you would follow to solve this problem in as much detail as possible (not going beyond 3 slides). Please Note: You don't need to actually implement the solution, but provide a roadmap on how you would achieve it if given the task without any external help or expenditure.

Please add these 3 slides to the same presentation used for the exercise above.