Problem Statement

Feature Engineering to predict road accidents on a highway.

Datasets:

1. Traffic Database: The government has set up traffic sensors on 5 locations on the highway which tell you the number of cars that passed over the sensor for every 30 minutes. The sensor malfunctions from time to time and there may be missing values (both missing measurements and missing rows). Missing values can be imputed by replacing the missing row with the most recently/previously available traffic for that location.

   A sample row:

H2	2nd January	00:30	43

   This means that on H2 location on 2nd January, 43 cars moved on the sensor from 00:30 to 1:00.

2. Accident Database: Whenever an accident happens, it is recorded in this database along with the data, time, location of the accident.

3. Data Engineering frame [DEF]: We need to combine this information to create a dataset which amenable to applying machine learning to. We are tasked to predict whether an accident will happen in the next 15 minutes; this prediction will be updated every 30 minutes. The training of the model is being done only for accident data between 00:00 1st January 2021 to 00:00 10th January, 2021 ie 9 days. There will be 48 30-min periods in a day, so 432 in the period of 9 days. Let's call this number S. The total size of the DEF will be S x unique_locations in our database.
4. The DEF has 4 features for every location L, timestamp T pair:
   a. Flow Diff Same Location: That is change in traffic on the same location L from (T-30, T) period to (T-60, T-30). These capture sudden changes in traffic.
   b. Max Last 120 Same Location: This is max traffic on the same location L between T-30, T-60, T-90 and T-120.
   c. Is T part of rushhour [9-11 AM and 5-6 PM on weekdays only.]
   d. Is T part of weekday
5. The predicted label for each Location L, timestamp T pair is the number of accidents on the location L between T to T+15.

An example row would be the following based on real data in the database:

Location	Time Interval	Is Weekday?	Is Rushhour?	Flow Diff Same Location	Max Last 120 Same Location	Accidents_Next_15
H1	1/1/2021 2:30	**1** (It was a Friday)	**0** (2:30 AM is not rushour)	**66 – 57 = 9** (The difference in traffic on	**99 max{3, 99, 57, 66}** The max traffic	**1** (Number of accidents on H1 on

				H1 in period (2:00-2:30) minus 1:30-2:00)	among between 0:30 to 2:30 on Location H1	1/1/2021 from 2:30 to **3:00\ 2.45**
H1	1/1/2021 3:00	..	..	..	..	(Number of accidents on 1/1/2021 on H1 from 3:00 to 3:30)
…	…					
…	…					
H2	1/1/2021 2:30					
…	**…**					
…	**…**					
…	**…**					
H5	9/1/2021 23:30	**…**				

Total rows: 5 * 24 * 2 * 9

Technical Setup:

Given the remaining machine learning pipeline is also written in python, its suited that the data engineering is also completed in python [preferably pandas]

Create a **well documented** jupyter notebook which takes as input the 2 datasets and generated the final DEF and saves it an excel. Keep the code well commented and documented so we can walk through it.

Think modularity: the more modular the code, the better; Eg. maybe it is later decided, instead of next_15 accident prediction, we want to do next_30 OR instead of max in last 120 minutes, it is decided to do average in the last 120 minutes.

End to end, there should be no manual process or work done in excel.

Take assumptions where needed, but write them down. If there are contradictions, please don’t be afraid to ask.