Instructions - R assignment

1. There is a dataset (~ 100 rows) with unstructured / partial address information. You will write R code that identifies, matches / predicts, standardizes the address to

(a) house/street nbr

(b) locality/colony,

(c) area/ward,

(d) city

(e) pincode

2. You will need to complete each address based on hierarchy of information, and use heuristics to fill in missing information. For example, if colony is known, we will typically know the area/city and pincode - even if not specified in the address. For example if the original address string only has "jwalamukhi hostel" - the output should resolve to "IIT Delhi, Hauz Khas, New Delhi, 110016". However if only house number is known (eg. house# 1212), and no other fields are known, then we can't determine any subsequent fields (colony / area / city / pincode), so they will be null.

So output that we expect is a csv file with the following fields:

- Original input string
- Best predicted full address
- Best predicted address part 1 - House / Street nbr
- Best predicted address part 2 - Locality / Colony
- Best predicted address part 3 - Area / Ward
- Best predicted address part 4 - City
- Best predicted address part 5 - Pincode

3. Here are the metrics that we will use to evaluate the performance of the code:
If any of the address parts 1 to 5 is not null, we will assume that the code / algorithm has determined it can predict something - so it will be a positive prediction. If all the address parts parts 1 to 5 are null, we will assume that the code / algorithm has determined it cannot predict - so it will be a negative prediction. Below are the examples:

Address string	Predicted House / Street nbr (part 1)	Predicted locality / colony (part 2)	Predicted area / ward (part 3)	Predicted City (part 4)	Predicted Pincode (part 5)	Outcome
jwalamukhi hostel, delhi	Jwalamukhi Hostel	IIT Delhi	Hauz Khas	New Delhi	110016	True Positive
house nbr 1202, delhi	House Nbr 1022			New Delhi		True Positive
house nbr 1202, delhi						False Negative
hauz nbr 1202, delhi			Hauz Khas	New Delhi		False Positive
zxdasd						True Negative

a) We need high precision / very low false discovery rate (i.e FP / (FP + TP)) > 95% (Type I error)

b) We need reasonable false omission rates (i.e. FN / (FN + TN)) <= 25% (Type II error)

4. You can use any publicly available api / data, or r libs etc in their code to help determine the correct address predictions.

5. You will get 3 days to do this assignment. You will need to submit the following by Wednesday, 12th October:

a) R code

b) Any public data set used, with URL of that data

c) the code output as defined in #2

d) The analysis of their output (they should classify their predictions as True Positive, False Positive, True Negative, False Negative) and submit their precision and omission rates.