In this question we will try to detect anomalies in the bidding behavior. To do that we will analyze two datasets:

- notifications_final - contains information on notifications for three places in Russia;
- protocols_final - contains information on bids for three places in Russia.

1. Write a **function**, function_fields_extract_not_zk, which takes as an argument an absolute path to notificationzk.csv and does the following:
   (a) Reads this file using fread, as a column separator please use "#" symbol, the class of each column should be "character";
   (b) Keeps following fields in the data:
      i. notificationnumber - notification number;
      ii. versionnumber - version of the notification;
      iii. publishdate - publication date;
      iv. customerrequirement_maxprice - starting price in the auction;
      v. notificationcommission_p1date - bid submission start date;
      vi. notificationcommission_p2date - bid submission end date.
2. Write a **function,** function_combine_not_zk, which does the following:
   (a) As input takes absolute path to the folder, notifications_final;
   (b) Applies function_fields_extract_not_zk to all files with the name notificationzk.csv within folder notifications_final (including subfolders) and combines them together into a data.table;
   (c) Orders this data.table by notificationnumber, versionnumber and publishdate in the decreasing order;
   (d) Keeps unique observations by notificationnumber;
   (e) Keeps only those observations where field customerrequirement_maxprice does not contain "&" symbol;
   (f) Removes those observations whose notificationnumber starts with "99";
   (g) Creates a new variable days which equals to the time difference in days between notificationcommission_p2date and notificationcommission_p1date;
   (h) Writes the created data.table to file, notification_zk_clean.csv - this file should be written to the folder notifications_final;

(i) Please perform all these operations in the order which is specified above.

3. Write a **function**, function_fields_extract_prot_zk, which as an argument takes an absolute path to the file protocolzk1.csv and does the following:

   (a) Opens the file (please use fread to read .csv file, use "#" as a column separator, column classes for all variables should be set to be characters);

   (b) Extracts following variables:

      i. protocoldate;

      ii. notificationnumber - notification id;

      iii. protocolprotocolapplications_application_journalnumber - this variable provides an information on unique identifiers of bidders (separated by &&&&), for example, for notification number 0348100035711000002, this variable equals 1&&&&3 - this tells us that there were two bidders in this auction, first had an identifier 1 and the second had an identifier 3, notice that numeric values do not have a meaning here.

      iv. protocolprotocolapplications_application_price - this variable provides us with bids of all bidders;

      v. protocolprotocolapplications_application_appdate - provides you with the information on date and time when a particular bid was submitted, "T" and "Z" here do not have any meaning and should be replaced by empty characters.

4. Write a **function,** function_combine_prot_zk, which does the following:

   (a) As an input takes absolute path to folder protocols_final;

   (b) Applies the function_fields_extract_prot_zk to all files called protocolzk1.csv within folder protocols_final (including subfolders) and combines them together into a data.table;

   (c) Orders this data.table by notificationnumber, protocoldate in the decreasing order;

   (d) Keeps unique observations by notificationnumber;

   (e) Removes those observations numbers whose notificationnumber starts with "99";

   (f) Writes the created data.table to file, protocols_zk_clean.csv - this file should be written to the folder protocols_final;

5. Merge notification_zk_clean.csv with protocols_zk_clean.csv, as a merging variable you should use notificationnumber, additionally do the following:

(a) Keep only those observations where you have at least 2 bidders;

(b) Keep only those observations where all submitted bids are different;

(c) Keep only those observations where all submission times are different;

(d) For each row obtain min bid and second minimum bid, call these variable min_bid, second_min_bid;

(e) Calculate whether the bidder who submitted the minimum bid, bided the last (i.e., had the largest protocolprotocolapplications_application_appdate). Please create a dummy variable, winner_last, which equals 1 if this is the case and 0 otherwise.

6. We next label auction as suspicious if the percentage decrease between first two bids is smaller than 1%. I.e., if :

$$suspicious = \begin{cases} 1 \text{if} \left| \frac{b_1 - b_2}{b_2} \right| \leq 0.01 \\ 0 \text{otherwise} \end{cases}$$

where:

b_1 : the lowest bid

b_2 : second lowest bid

7. Run the following linear probability model:

$$suspicious_i = \alpha + \beta \times 1\{winner = \text{submitted bid the last}\}_i + \gamma \times \text{nr\_bidders}_i + \xi \times \text{starting price}_i + \mu \times \text{nr\_days}_i + \epsilon_i$$

Conclude, whether you find something suspicious?