Greeting from The Gene Box!

The assignment consists of two problems.

You should send the solution with the code in a separate folder for each problem.

Code clearly: Try to write code that is easy to read. We value this skill higher than cleverness or using the least number of lines/characters possible. Reading through your code and comments, we should be able to figure out how your code is organized and why it works. If we can't, we reserve the right to deduct points.

Make assumptions whenever necessary but should be mentioned in the comments.

Any questions regarding the assignment can be directed to bhushan@thegenebox.com

Use R or python as your coding language.

PROBLEM 1

The file was generated from GenomeStudio software. The data fields in the attached files are:

rsID Chromosome Location Genotype

rsID: The unique ID assigned to a change in single nucleotide in the DNA at specific location of the chromosome. May be repeated for quality control. (Look out for weird cases).

Chromosome: chromosome no on which the variation is located.

Location: Location of the variation on the chromosome.

Genotype: Humans have two sets of each chromosomes. Genotype (say GG) tells us about what nucleotides are present on each chromosome for the given variant location (In this case, G is present on both copies). The second part is the genotype call score which is calculated with specified threshold for the genotypes.

A. Find unique rsIDs and chromosome and location with genotype keeping in mind the following things:
   1) While removing duplication consider keeping observations with higher call rates
   2) Remove no calls represented as '--' but if all repeated observation are no calls keep one observation with no call (to imply that rsID has no call in the data).

B. Prepare a summary of the duplications. (Is there any relation with the duplicates?)

C. How will you report a final version of the data? Do not forget to submit the csv/txt file of it.

D. Visualize the data according to your understanding and along with the code, include its image file with the solution.

Related topics you might want to read: Genotype, Allele, Single Nucleotide Polymorphism,

PROBLEM 2

The attached file contains the following fields.
Note that the file doesn’t have the header fields as mentioned here.

SNP	Allele1	Allele2	S1			S2			…………………….		
	A	B	AA	AB	BB	AA	AB	BB			
Rs123	T	C	0	0	1	1	0	0	1	0	0
Rs124	A	G	0	0	1	0	1	0	0	0	1

Transform the data to obtain the following output:

SNP	S1	S2	………..
Rs123	CC	TT	
Rs124	GG	AG	

Your submission should contain the following files:

1) Output file – must be a excel sheet, should be in the format as specified above
2) Code file – txt file containing the code with comments whenever necessary.