Assignment 2 - Processing VCF file

1. Download the vcf file from https://spliceatlas.s3.amazonaws.com/clinvar_20220227_10K.vcf
2. File contains 10,000 lines
3. This is a kind of tsv file (tab separated values)
4. Line containing single # is the header. This contains the column headers
5. File will contain the following 8 columns (tab separated).
    a. CHROM
    b. POS
    c. ID
    d. REF
    e. ALT
    f. QUAL
    g. FILTER
    h. INFO
6. 2 lines from file is given here as a sample
    a. #CHROM POS ID REF ALT QUAL FILTER INFO
    b. 1 861332 1019397 G A . . ALLELEID=1003021;CLNDISDB=MedGen:CN517202;CLNDN=not_provided;CLNHGVS=NC_000001.10:g.861332G>A;CLNREVSTAT=criteria_provided,_single_submitter;CLNSIG=Uncertain_significance;CLNVC=single_nucleotide_variant;CLNVCSO=SO:0001483;GENEINFO=SAMD11:148398;MC=SO:0001583|missense_variant;ORIGIN=1;RS=1640863258
    c. 1 865519 1125147 C T . . ALLELEID=1110865;CLNDISDB=MedGen:CN517202;CLNDN=not_provided;CLNHGVS=NC_000001.10:g.865519C>T;CLNREVSTAT=criteria_provided,_single_submitter;CLNSIG=Likely_benign;CLNVC=single_nucleotide_variant;CLNVCSO=SO:0001483;GENEINFO=SAMD11:148398;MC=SO:0001627|intron_variant;ORIGIN=1
7. Produce a output/result file (filename_of_ur_choice.csv - comma separated values) containing 14 columns by processing the downloaded file
    a. Result file should contain 14 columns which are given as follows CHROM,POS,ID,REF,ALT,ALLELEID,CLNHGVS,CLNSIG,CLNVC,ORIGIN,RS,Gene_ID,Gene_symbol,Consequence
    b. CHROM,POS,ID,REF,ALT can be collected directly from the first 5 columns of downloaded vcf file
    c. To collect remaining data please use the last/8th col (INFO)
        i. 8th col values are separated by ‘;’
        ii. 8th col will have Attribute=value;Attribute=value;Attribute=value; and so on

iii. ALLELEID,CLNHGVS,CLNSIG,CLNVC,ORIGIN,RS can be collected directly from attributes of 8th col
    1. ALLELEID=
    2. CLNHGVS=
    3. CLNSIG=
    4. CLNVC=
    5. ORIGIN=
    6. RS=
iv. Gene_ID & Gene_Symbol can be collected from GENEINFO attribute in 8th column
    1. GENEINFO=Gene_Symbol:Gene_ID
v. Consequence can be collected from MC attribute of 8th col
    1. MC=SO:SO_ID|Consequence
vi. If any of the attribute isn't available put '-' in the result file

d. Expected o/p for the first 2 lines are
    i. 1,861332,1019397,G,A,1003021,NC_000001.10:g.861332G>A,Uncertain_significance,single_nucleotide_variant,1,1640863258,148398,SAMD11,missense_variant
    ii. 1,865519,1125147,C,T,1110865,NC_000001.10:g.865519C>T,Likely_benign,single_nucleotide_variant,1,-,148398,SAMD11,intron_variant

8. From the .csv file created in the above step, count the different type of origin (10th col)
    a. Values of Origin col have the following meaning.
        i. 0 - unknown;
        ii. 1 - germline;
        iii. 2 - somatic;
        iv. 4 - inherited;
        v. 8 - paternal;
        vi. 16 - maternal;
        vii. 32 - de-novo;
        viii. 64 - biparental;
        ix. 128 - uniparental;
        x. 256 - not-tested;
        xi. 512 - tested-inconclusive;
    b. If you get number other than listed above, classify under 'Others'
    c. Make a data structure (dictionary) that has Origin type as key and the count of them as value
        i. Expected o/p

```
'Others' => 95,
'de-novo' => 38,
'inherited' => 32,
'somatic' => 18,
'maternal' => 19,
'paternal' => 14,
```

'germline' => 6648,
'uniparental' => 6,
'unknown' => 145

d. Make a data structure of the same as mentioned in (c) into a json object. Display the results as a table using HTML