```r
# import dataset
clinvar_variants <- read.csv("~/Documents/GitHub/ECL298_2026/Describe_data/clinvar_variants.csv",
                             na.strings = c("", "NA", ".", " "))
```

```r
dim(clinvar_variants)
```

```
## [1] 63446    14
```

This dataset consists of 63,446 variants across 14 variables.

```r
summary(clinvar_variants)
```

```
##     CHROM                POS                REF                ALT
##  Length:63446       Min.   :   138755   Length:63446       Length:63446
##  Class :character   1st Qu.: 32810084   Class :character   Class :character
##  Mode  :character   Median : 57663378   Mode  :character   Mode  :character
##                     Mean   : 77461544
##                     3rd Qu.:112508141
##                     Max.   :247607973
##
##      CLASS        Consequence           IMPACT          Amino_acids
##  Min.   :0.0000   Length:63446       Length:63446       Length:63446
##  1st Qu.:0.0000   Class :character   Class :character   Class :character
##  Median :0.0000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.2514
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
##     AF_EXAC            SIFT             PolyPhen           LoFtool
##  Min.   :0.00000   Length:63446       Length:63446       Min.   :0.000
##  1st Qu.:0.00000   Class :character   Class :character   1st Qu.:0.024
##  Median :0.00004   Mode  :character   Mode  :character   Median :0.159
##  Mean   :0.01460                                         Mean   :0.349
##  3rd Qu.:0.00123                                         3rd Qu.:0.732
##  Max.   :0.49989                                         Max.   :1.000
##                                                          NA's   :3777
##    CADD_PHRED        BLOSUM62
##  Min.   : 0.001   Min.   :-3.0
##  1st Qu.: 7.195   1st Qu.:-2.0
##  Median :14.140   Median :-1.0
##  Mean   :15.707   Mean   :-0.4
##  3rd Qu.:24.100   3rd Qu.: 1.0
##  Max.   :99.000   Max.   : 3.0
##  NA's   :1040     NA's   :38315
```

Missing values were observed in LoFtool, CADD_PHRED, and BLOSUM62.

CHROM: chromosome number. POS: genetic position within the chromosome. REF: reference allele. ALT: alternative allele.

```r
table(clinvar_variants$CLASS)
```

```
##
##     0     1
## 47493 15953
```

CLASS: variant is classified as pathogenic (1) or benign (0).

```r
table(clinvar_variants$Consequence)
```

```
##
##                                                 3_prime_UTR_variant
##                                                                 414
##                                                 5_prime_UTR_variant
##                                                                 607
##                                              downstream_gene_variant
##                                                                  26
##                                                  frameshift_variant
##                                                                1717
##                          frameshift_variant&splice_region_variant
##                                                                  55
##                                       frameshift_variant&start_lost
##                                                                   4
##           frameshift_variant&start_lost&start_retained_variant
##                                                                   1
##                                        frameshift_variant&stop_lost
##                                                                   3
##                          frameshift_variant&stop_retained_variant
##                                                                   1
##                                                   inframe_deletion
##                                                                 549
##                            inframe_deletion&splice_region_variant
##                                                                   9
##                                                   inframe_insertion
##                                                                 179
##                           inframe_insertion&splice_region_variant
##                                                                   1
##                                                     intron_variant
##                                                                4262
##                     intron_variant&non_coding_transcript_variant
##                                                                   1
##                                                   missense_variant
##                                                               30837
##                            missense_variant&splice_region_variant
##                                                                 944
##                                          protein_altering_variant
##                                                                   8
##                                           splice_acceptor_variant
##                                                                 391
##                  splice_acceptor_variant&coding_sequence_variant
##                                                                   6
## splice_acceptor_variant&coding_sequence_variant&intron_variant
```

```
##                                                                 8
##                         splice_acceptor_variant&intron_variant
##                                                                 6
##                                              splice_donor_variant
##                                                               513
##                   splice_donor_variant&coding_sequence_variant
##                                                                 9
##    splice_donor_variant&coding_sequence_variant&intron_variant
##                                                                18
##                             splice_donor_variant&intron_variant
##                                                                17
##                       splice_region_variant&3_prime_UTR_variant
##                                                                 2
##                       splice_region_variant&5_prime_UTR_variant
##                                                                15
##   splice_region_variant&coding_sequence_variant&intron_variant
##                                                                 1
##                            splice_region_variant&intron_variant
##                                                              3286
##                         splice_region_variant&synonymous_variant
##                                                               536
##                                                       start_lost
##                                                                88
##                                  start_lost&5_prime_UTR_variant
##                                                                 1
##                               start_lost&splice_region_variant
##                                                                 2
##                                                     stop_gained
##                                                            1595
##                                  stop_gained&frameshift_variant
##                                                                25
##                                    stop_gained&inframe_deletion
##                                                                 1
##                                   stop_gained&inframe_insertion
##                                                                 1
##                              stop_gained&protein_altering_variant
##                                                                 1
##                                 stop_gained&splice_region_variant
##                                                                64
##                                                      stop_lost
##                                                                10
##                                   stop_lost&3_prime_UTR_variant
##                                                                 3
##                                             stop_retained_variant
##                                                                 9
##                       stop_retained_variant&3_prime_UTR_variant
##                                                                 1
##                                                synonymous_variant
##                                                             17139
##                                             upstream_gene_variant
##                                                                80
```

Consequence: it describes the predicted functional effect of each variant on the gene or protein.

```r
table(clinvar_variants$IMPACT)
```

```
##
##     HIGH      LOW MODERATE MODIFIER
##     4540    20989    32527     5390
```

The IMPACT variable categorizes variants into four functional severity levels (HIGH, MODERATE, LOW, and MODIFIER) based on their predicted effects on gene and protein function.

Amino_acids: Amino acid change.

AF_EXAC: the alternative allele frequency in population.

SIFT and PolyPhen provide categorical functional predictions, while LoFtool and CADD_PHRED generate continuous scores reflecting gene-level intolerance and variant-level deleteriousness, respectively.

BLOSUM62: BLOSUM62 is an evolutionary conservation score that quantifies how frequently a specific amino acid substitution is observed in conserved protein regions. Lower scores indicate less conservative substitutions that are more likely to disrupt protein function.