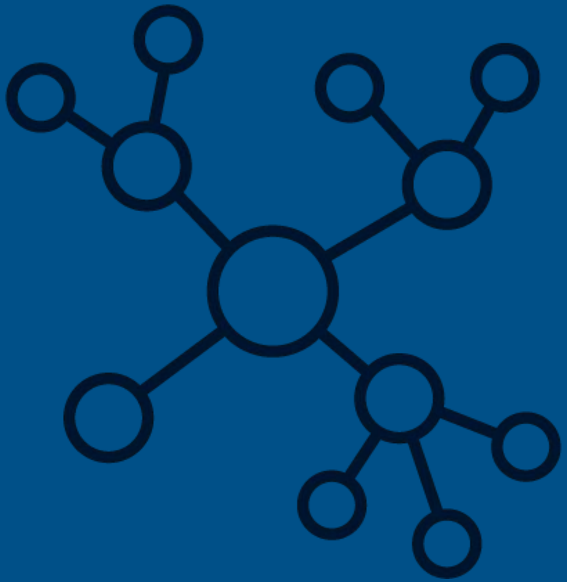
11.19.24

Explainable AI for Estimating Pathogenicity of Genetic Variants Using Large-Scale Knowledge Graphs

Shuya Abe, Shinichiro Tago , Kazuaki Yokoyama and Masaru Fuji



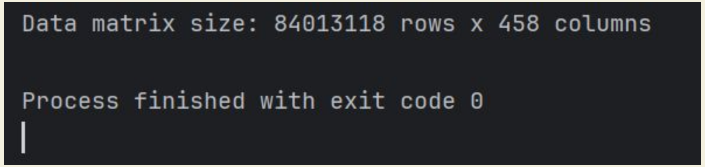
Final Paper Exploratory Data Analysis:
Daniel Gutierrez
CAP5610
Professor Ananda M. Mondal

Omics Data

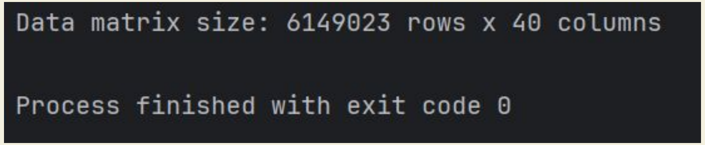
- Omics:
  - ClinVar: Phenomic/Genomic
    - Summarized Genomic data that includes clinical phenotypes, interpretations, and descriptions for SNV's
  - DbNSFP: Genomic
    - Genomic data that includes identification for single nucleotide variants and scores for genomic deletions and mutations
    - Also includes scores for the impact of a variant on protein function and synthesis

Raw Data Matrix

- ClinVar
  - Data Matrix Size: 6149023 rows x 40 columns
  - 273 MB .gz file(compressed 2.35 GB .txt file)
- DbNSFP
  - Multiple variant files separated by chromosome number. Combined into a single dataset
  - Data matrix size: 84013118 rows x 458 columns
  - 36.2 GB .gz file(compressed 202 GB .csv file)



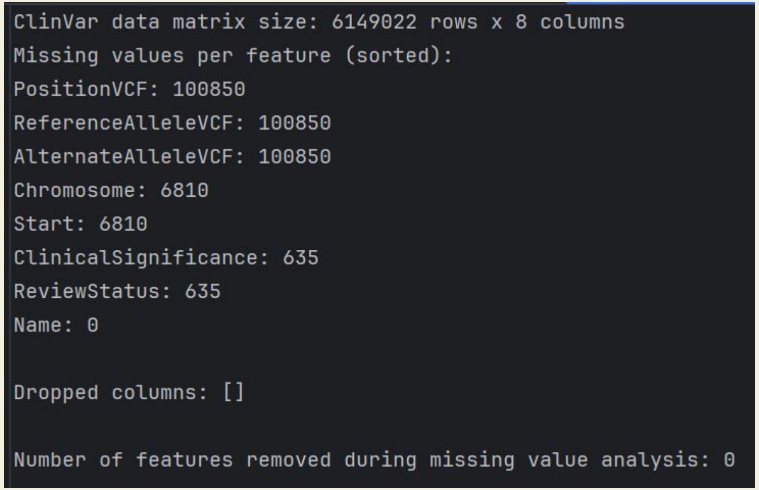
```
Data matrix size: 84013118 rows x 458 columns

Process finished with exit code 0
```



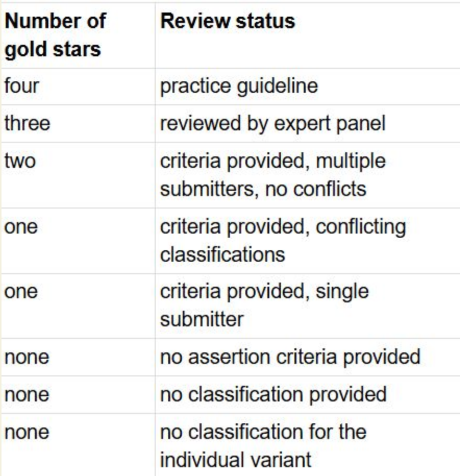
```
Data matrix size: 6149023 rows x 40 columns

Process finished with exit code 0
```

ClinVar EDA

- Most of the columns in Clinvar are index or metadata features. Those not used to link Clinvar samples to Dbnsfp samples were removed (32 columns).
  - ex) LastEvaluated, GeneID, OtherID, etc.
- Features Removed (>=80% with ‘-’,’na’,’-1’): none
- The features used to link ClinVar to Dbnsfp:
  - PositionVCF, ReferenceAlleleVCF, AlternativeAlleleVCF, Chromosome
- ClinicalSignificance will serve as the class labels for Pathogenicity
  - Pathogenic/Benign
  - Exclude “Likely”,”Uncertain” Samples
- ReviewStatus will be considered later on for further sample selection criteria

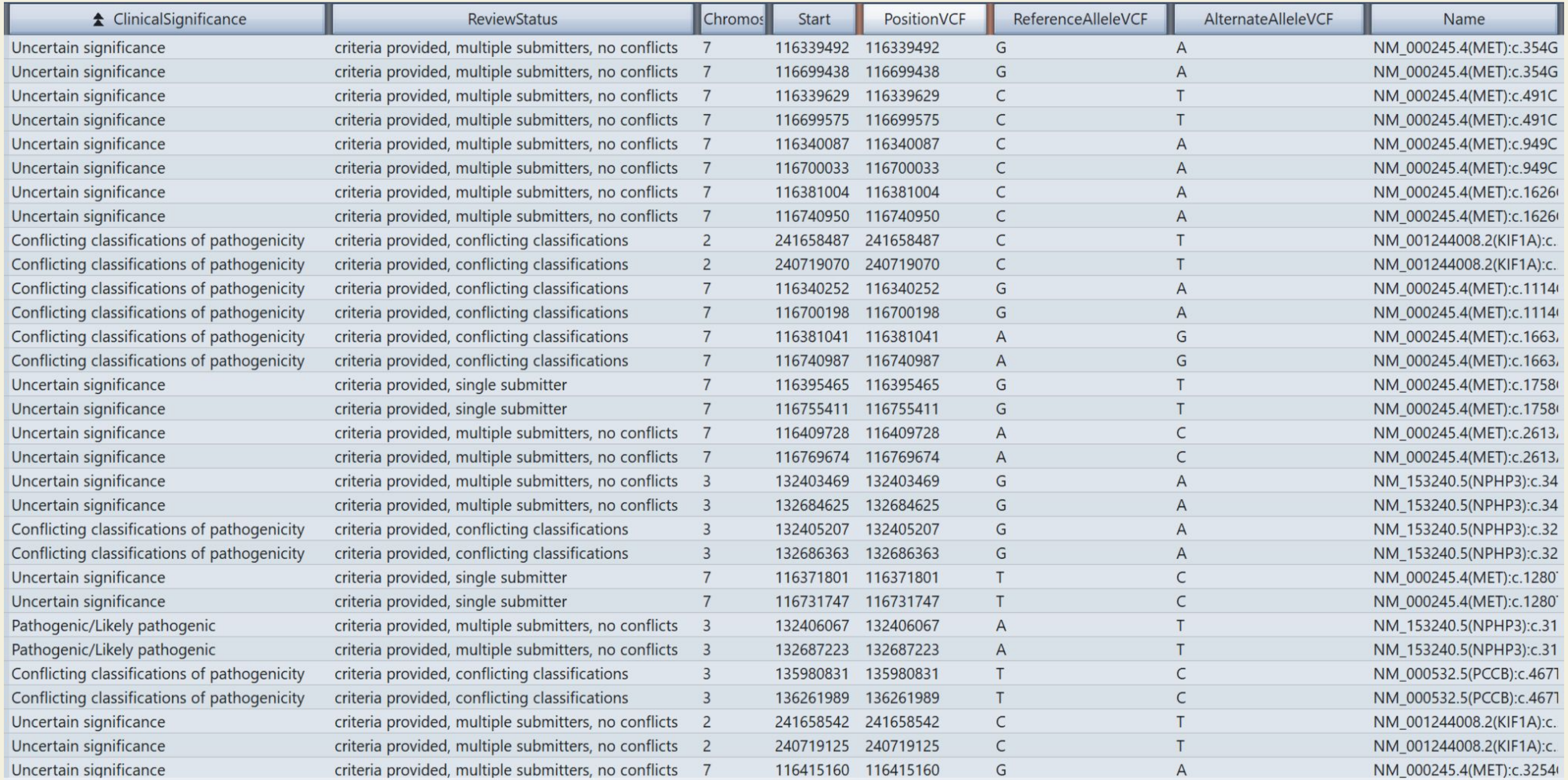


```
ClinVar data matrix size: 6149022 rows x 8 columns
Missing values per feature (sorted):
PositionVCF: 100850
ReferenceAlleleVCF: 100850
AlternateAlleleVCF: 100850
Chromosome: 6810
Start: 6810
ClinicalSignificance: 635
ReviewStatus: 635
Name: 0

Dropped columns: []

Number of features removed during missing value analysis: 0
```



Number of gold stars	Review status
four	practice guideline
three	reviewed by expert panel
two	criteria provided, multiple submitters, no conflicts
one	criteria provided, conflicting classifications
one	criteria provided, single submitter
none	no assertion criteria provided
none	no classification provided
none	no classification for the individual variant

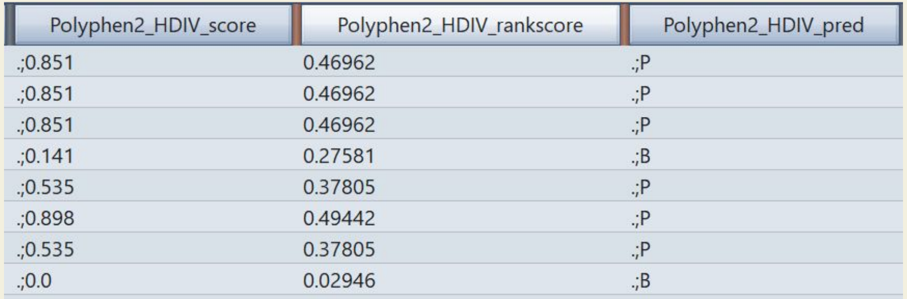
Filtered ClinVar Sample



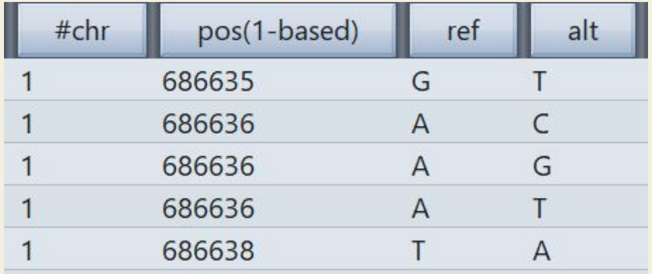
ClinicalSignificance	ReviewStatus	Chromos	Start	PositionVCF	ReferenceAlleleVCF	AlternateAlleleVCF	Name
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116339492	116339492	G	A	NM_000245.4(MET):c.354G
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116699438	116699438	G	A	NM_000245.4(MET):c.354G
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116339629	116339629	C	T	NM_000245.4(MET):c.491C
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116699575	116699575	C	T	NM_000245.4(MET):c.491C
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116340087	116340087	C	A	NM_000245.4(MET):c.949C
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116700033	116700033	C	A	NM_000245.4(MET):c.949C
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116381004	116381004	C	A	NM_000245.4(MET):c.1626
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116740950	116740950	C	A	NM_000245.4(MET):c.1626
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	2	241658487	241658487	C	T	NM_001244008.2(KIF1A):c.
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	2	240719070	240719070	C	T	NM_001244008.2(KIF1A):c.
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	7	116340252	116340252	G	A	NM_000245.4(MET):c.1114
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	7	116700198	116700198	G	A	NM_000245.4(MET):c.1114
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	7	116381041	116381041	A	G	NM_000245.4(MET):c.1663
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	7	116740987	116740987	A	G	NM_000245.4(MET):c.1663
Uncertain significance	criteria provided, single submitter	7	116395465	116395465	G	T	NM_000245.4(MET):c.1758
Uncertain significance	criteria provided, single submitter	7	116755411	116755411	G	T	NM_000245.4(MET):c.1758
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116409728	116409728	A	C	NM_000245.4(MET):c.2613
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116769674	116769674	A	C	NM_000245.4(MET):c.2613
Uncertain significance	criteria provided, multiple submitters, no conflicts	3	132403469	132403469	G	A	NM_153240.5(NPHP3):c.34
Uncertain significance	criteria provided, multiple submitters, no conflicts	3	132684625	132684625	G	A	NM_153240.5(NPHP3):c.34
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	3	132405207	132405207	G	A	NM_153240.5(NPHP3):c.32
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	3	132686363	132686363	G	A	NM_153240.5(NPHP3):c.32
Uncertain significance	criteria provided, single submitter	7	116371801	116371801	T	C	NM_000245.4(MET):c.1280
Uncertain significance	criteria provided, single submitter	7	116731747	116731747	T	C	NM_000245.4(MET):c.1280
Pathogenic/Likely pathogenic	criteria provided, multiple submitters, no conflicts	3	132406067	132406067	A	T	NM_153240.5(NPHP3):c.31
Pathogenic/Likely pathogenic	criteria provided, multiple submitters, no conflicts	3	132687223	132687223	A	T	NM_153240.5(NPHP3):c.31
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	3	135980831	135980831	T	C	NM_000532.5(PCCB):c.467T
Conflicting classifications of pathogenicity	criteria provided, conflicting classifications	3	136261989	136261989	T	C	NM_000532.5(PCCB):c.467T
Uncertain significance	criteria provided, multiple submitters, no conflicts	2	241658542	241658542	C	T	NM_001244008.2(KIF1A):c.
Uncertain significance	criteria provided, multiple submitters, no conflicts	2	240719125	240719125	C	T	NM_001244008.2(KIF1A):c.
Uncertain significance	criteria provided, multiple submitters, no conflicts	7	116415160	116415160	G	A	NM_000245.4(MET):c.3254

DbNSFP Filtering

- Similarly to ClinVar, metadata and index columns were filtered out based on context from the readme
- For each algorithm represented in the dataset, there was at least a raw score, rank score, and prediction
  - Rank Score is a normalized output that computes the ratio of the sample's rank over the entire dataset for that algorithm
  - Rank Score usually has a scale of [0,1] with 1 being closest to a behavior that can indicate pathogenicity
- Problems with Parsing Raw Scores:
  - Each column uses different indicators for missing values
  - Many samples have multiple entries for one column, usually surrounded by a semicolon
- The features used to link ClinVar to Dbnsfp:
  - Chr, pos(1-based), ref, alt

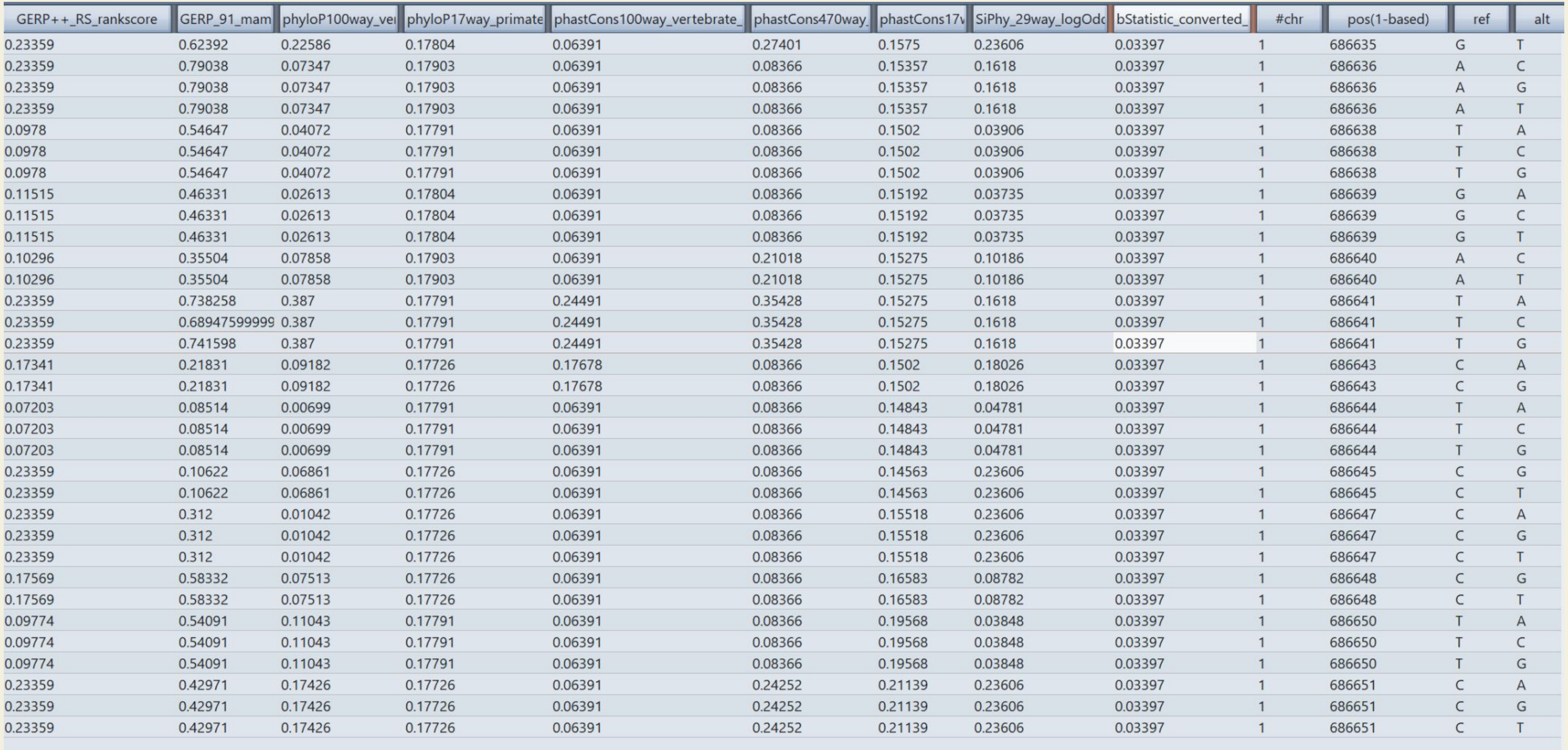


Polyphen2_HDIV_score	Polyphen2_HDIV_rankscore	Polyphen2_HDIV_pred
.;0.851	0.46962	.;P
.;0.851	0.46962	.;P
.;0.851	0.46962	.;P
.;0.141	0.27581	.;B
.;0.535	0.37805	.;P
.;0.898	0.49442	.;P
.;0.535	0.37805	.;P
.;0.0	0.02946	.;B



#chr	pos(1-based)	ref	alt
1	686635	G	T
1	686636	A	C
1	686636	A	G
1	686636	A	T
1	686638	T	A

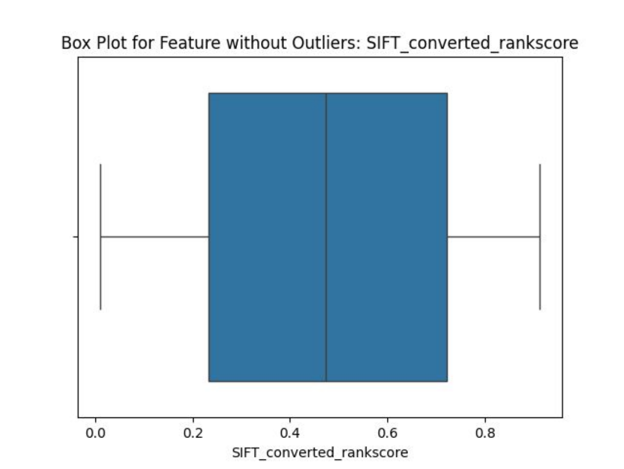
Filtered DbNSFP Sample

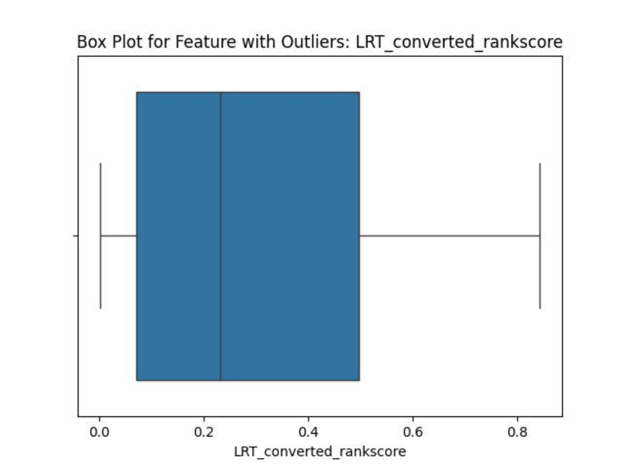


GERP++_RS_rankscore	GERP_91_mam	phyloP100way_ve	phyloP17way_primate	phastCons100way_vertebrate_	phastCons470way_	phastCons17w	SiPhy_29way_logOd	bStatistic_converted_	#chr	pos(1-based)	ref	alt
0.23359	0.62392	0.22586	0.17804	0.06391	0.27401	0.1575	0.23606	0.03397	1	686635	G	T
0.23359	0.79038	0.07347	0.17903	0.06391	0.08366	0.15357	0.1618	0.03397	1	686636	A	C
0.23359	0.79038	0.07347	0.17903	0.06391	0.08366	0.15357	0.1618	0.03397	1	686636	A	G
0.23359	0.79038	0.07347	0.17903	0.06391	0.08366	0.15357	0.1618	0.03397	1	686636	A	T
0.0978	0.54647	0.04072	0.17791	0.06391	0.08366	0.1502	0.03906	0.03397	1	686638	T	A
0.0978	0.54647	0.04072	0.17791	0.06391	0.08366	0.1502	0.03906	0.03397	1	686638	T	C
0.0978	0.54647	0.04072	0.17791	0.06391	0.08366	0.1502	0.03906	0.03397	1	686638	T	G
0.11515	0.46331	0.02613	0.17804	0.06391	0.08366	0.15192	0.03735	0.03397	1	686639	G	A
0.11515	0.46331	0.02613	0.17804	0.06391	0.08366	0.15192	0.03735	0.03397	1	686639	G	C
0.11515	0.46331	0.02613	0.17804	0.06391	0.08366	0.15192	0.03735	0.03397	1	686639	G	T
0.10296	0.35504	0.07858	0.17903	0.06391	0.21018	0.15275	0.10186	0.03397	1	686640	A	C
0.10296	0.35504	0.07858	0.17903	0.06391	0.21018	0.15275	0.10186	0.03397	1	686640	A	T
0.23359	0.738258	0.387	0.17791	0.24491	0.35428	0.15275	0.1618	0.03397	1	686641	T	A
0.23359	0.68947599999	0.387	0.17791	0.24491	0.35428	0.15275	0.1618	0.03397	1	686641	T	C
0.23359	0.741598	0.387	0.17791	0.24491	0.35428	0.15275	0.1618	0.03397	1	686641	T	G
0.17341	0.21831	0.09182	0.17726	0.17678	0.08366	0.1502	0.18026	0.03397	1	686643	C	A
0.17341	0.21831	0.09182	0.17726	0.17678	0.08366	0.1502	0.18026	0.03397	1	686643	C	G
0.07203	0.08514	0.00699	0.17791	0.06391	0.08366	0.14843	0.04781	0.03397	1	686644	T	A
0.07203	0.08514	0.00699	0.17791	0.06391	0.08366	0.14843	0.04781	0.03397	1	686644	T	C
0.07203	0.08514	0.00699	0.17791	0.06391	0.08366	0.14843	0.04781	0.03397	1	686644	T	G
0.23359	0.10622	0.06861	0.17726	0.06391	0.08366	0.14563	0.23606	0.03397	1	686645	C	G
0.23359	0.10622	0.06861	0.17726	0.06391	0.08366	0.14563	0.23606	0.03397	1	686645	C	T
0.23359	0.312	0.01042	0.17726	0.06391	0.08366	0.15518	0.23606	0.03397	1	686647	C	A
0.23359	0.312	0.01042	0.17726	0.06391	0.08366	0.15518	0.23606	0.03397	1	686647	C	G
0.23359	0.312	0.01042	0.17726	0.06391	0.08366	0.15518	0.23606	0.03397	1	686647	C	T
0.17569	0.58332	0.07513	0.17726	0.06391	0.08366	0.16583	0.08782	0.03397	1	686648	C	G
0.17569	0.58332	0.07513	0.17726	0.06391	0.08366	0.16583	0.08782	0.03397	1	686648	C	T
0.09774	0.54091	0.11043	0.17791	0.06391	0.08366	0.19568	0.03848	0.03397	1	686650	T	A
0.09774	0.54091	0.11043	0.17791	0.06391	0.08366	0.19568	0.03848	0.03397	1	686650	T	C
0.09774	0.54091	0.11043	0.17791	0.06391	0.08366	0.19568	0.03848	0.03397	1	686650	T	G
0.23359	0.42971	0.17426	0.17726	0.06391	0.24252	0.21139	0.23606	0.03397	1	686651	C	A
0.23359	0.42971	0.17426	0.17726	0.06391	0.24252	0.21139	0.23606	0.03397	1	686651	C	G
0.23359	0.42971	0.17426	0.17726	0.06391	0.24252	0.21139	0.23606	0.03397	1	686651	C	T

DbNSFP Subset

- There still a need to complete a full EDA on the complete filtered dbnsfp dataset before merging with clinvar and use in ML
- Limited EDA on DbNSFP on first 10,000 samples
  - Number of features removed due to missing values(>=80%): 3
    - EVE_rankscore: 10,000
    - phyloP470way_mammalian_rankscore: 10,000
    - LINSIGHT_rankscore: 8610
  - Number of samples removed during outlier analysis: 7987
    - integrated_fitCons_rankscore: 1989





Clean Data Matrix

- Clinvar
  - Cleaned ClinVar data matrix size: 6149022 rows x 8 columns
- DbNSFP Subset
  - Cleaned DbNSFP subset: 2013 rows x 59 columns