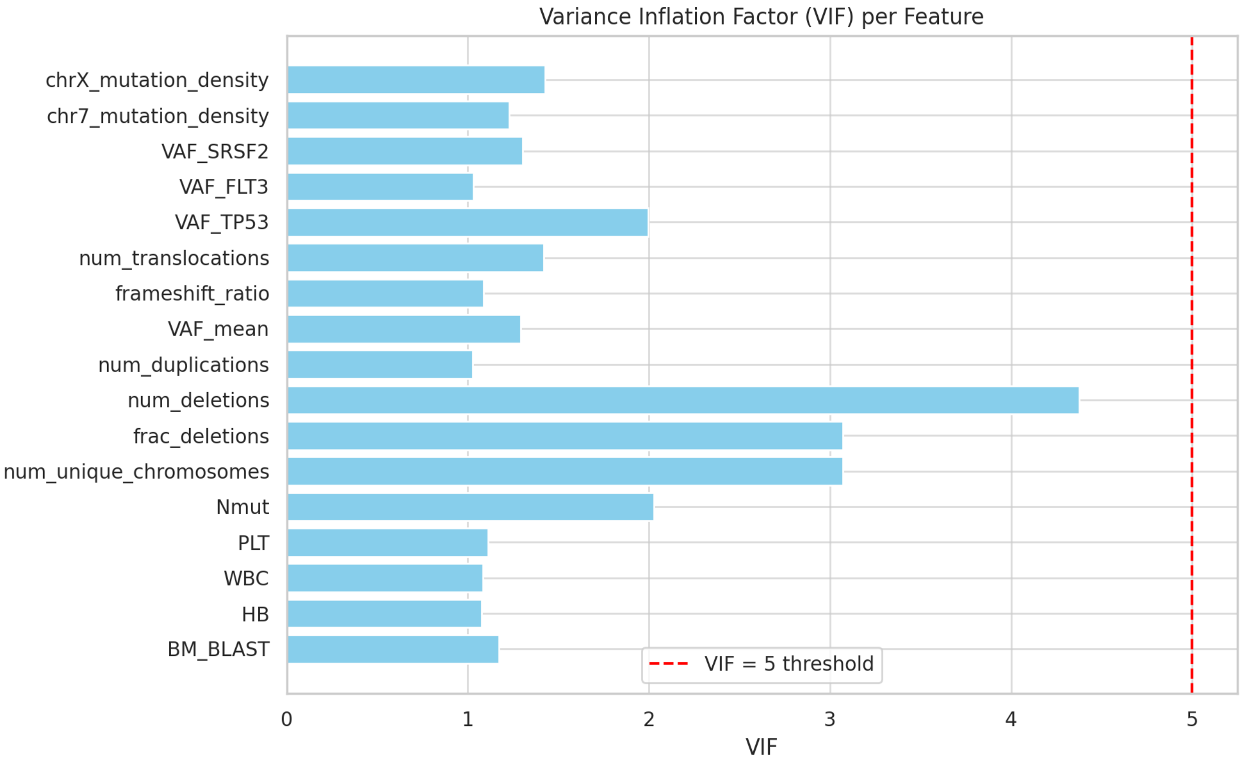
Feature Engineering Report
Survival Prediction for Adult Myeloid Leukemia

This report presents an in-depth exploratory analysis and feature engineering process conducted as part of the 2025 QRT Data Challenge in collaboration with Institut Gustave Roussy. The challenge aimed to build predictive models for overall survival (OS) in patients diagnosed with adult myeloid leukemia, using a dataset composed of clinical and molecular data. The dataset includes: - Clinical variables (e.g., blood counts, bone marrow blasts, karyotype abnormalities) - Molecular data from somatic mutations (e.g., affected genes, variant allele frequency) From these raw data, several informative features were derived to capture disease severity and progression patterns. Notable features include: - Mutation burden: number of mutations, unique genes, average VAF - Key gene-specific VAFs (TP53, FLT3, SRSF2) based on their relevance in hematologic malignancies - Chromosome-specific mutation densities (chr7 and chrX) - Cytogenetic abnormalities: deletions, duplications, translocations, and affected chromosomes The goal of this analysis is to visualize these features, assess their predictive potential, and ensure they are suitable for use in survival models.

Variance Inflation Factor (VIF) Analysis

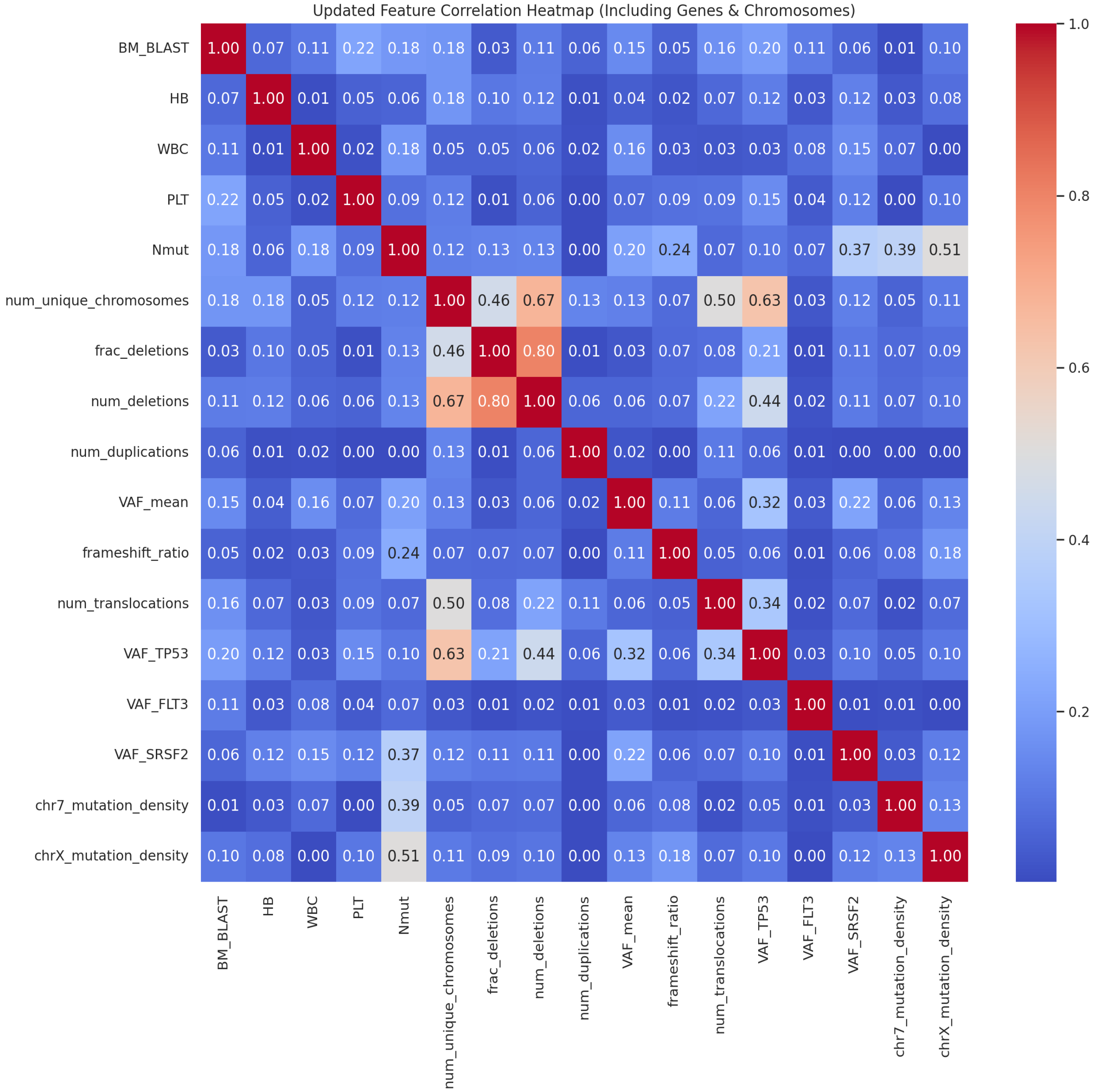
Feature	VIF
BM_BLAST	1.17
HB	1.08
WBC	1.08
PLT	1.11
Nmut	2.03
num_unique_chromosomes	3.07
frac_deletions	3.07
num_deletions	4.38
num_duplications	1.03
VAF_mean	1.29
frameshift_ratio	1.09
num_translocations	1.42
VAF_TP53	1.99
VAF_FLT3	1.03
VAF_SRSF2	1.3
chr7_mutation_density	1.23
chrX_mutation_density	1.43

Features with VIF < 5 are considered acceptable, indicating low multicollinearity.



This bar chart visualizes the Variance Inflation Factor for each feature. The red dashed line indicates a VIF of 5, which is commonly used as a threshold for multicollinearity concerns.

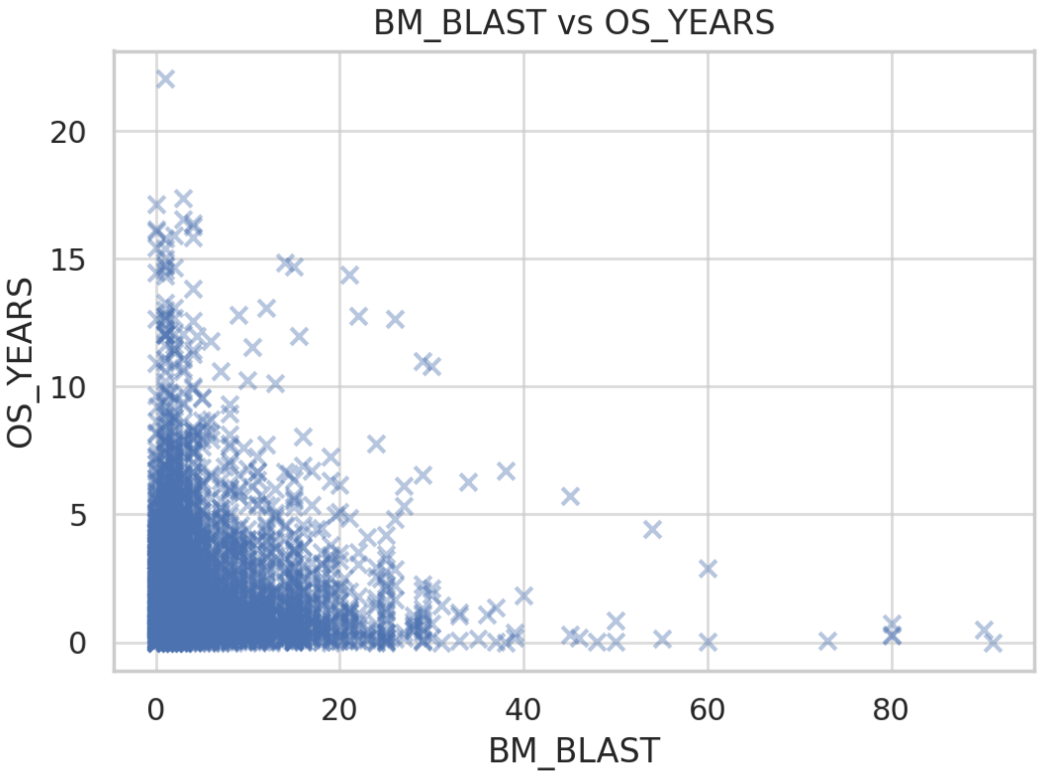
Feature Correlation Heatmap



	BM_BLAST	HB	WBC	PLT	Nmut	num_unique_chromosomes	frac_deletions	num_deletions	num_duplications	VAF_mean	frameshift_ratio	num_translocations	VAF_TP53	VAF_FLT3	VAF_SRSF2	chr7_mutation_density	chrX_mutation_density
BM_BLAST	1.00	0.07	0.11	0.22	0.18	0.18	0.03	0.11	0.06	0.15	0.05	0.16	0.20	0.11	0.06	0.01	0.10
HB	0.07	1.00	0.01	0.05	0.06	0.18	0.10	0.12	0.01	0.04	0.02	0.07	0.12	0.03	0.12	0.03	0.08
WBC	0.11	0.01	1.00	0.02	0.18	0.05	0.05	0.06	0.02	0.16	0.03	0.03	0.03	0.08	0.15	0.07	0.00
PLT	0.22	0.05	0.02	1.00	0.09	0.12	0.01	0.06	0.00	0.07	0.09	0.09	0.15	0.04	0.12	0.00	0.10
Nmut	0.18	0.06	0.18	0.09	1.00	0.12	0.13	0.13	0.00	0.20	0.24	0.07	0.10	0.07	0.37	0.39	0.51
num_unique_chromosomes	0.18	0.18	0.05	0.12	0.12	1.00	0.46	0.67	0.13	0.13	0.07	0.50	0.63	0.03	0.12	0.05	0.11
frac_deletions	0.03	0.10	0.05	0.01	0.13	0.46	1.00	0.80	0.01	0.03	0.07	0.08	0.21	0.01	0.11	0.07	0.09
num_deletions	0.11	0.12	0.06	0.06	0.13	0.67	0.80	1.00	0.06	0.06	0.07	0.22	0.44	0.02	0.11	0.07	0.10
num_duplications	0.06	0.01	0.02	0.00	0.00	0.13	0.01	0.06	1.00	0.02	0.00	0.11	0.06	0.01	0.00	0.00	0.00
VAF_mean	0.15	0.04	0.16	0.07	0.20	0.13	0.03	0.06	0.02	1.00	0.11	0.06	0.32	0.03	0.22	0.06	0.13
frameshift_ratio	0.05	0.02	0.03	0.09	0.24	0.07	0.07	0.07	0.00	0.11	1.00	0.05	0.06	0.01	0.06	0.08	0.18
num_translocations	0.16	0.07	0.03	0.09	0.07	0.50	0.08	0.22	0.11	0.06	0.05	1.00	0.34	0.02	0.07	0.02	0.07
VAF_TP53	0.20	0.12	0.03	0.15	0.10	0.63	0.21	0.44	0.06	0.32	0.06	0.34	1.00	0.03	0.10	0.05	0.10
VAF_FLT3	0.11	0.03	0.08	0.04	0.07	0.03	0.01	0.02	0.01	0.03	0.01	0.02	0.03	1.00	0.01	0.01	0.00
VAF_SRSF2	0.06	0.12	0.15	0.12	0.37	0.12	0.11	0.11	0.00	0.22	0.06	0.07	0.10	0.01	1.00	0.03	0.12
chr7_mutation_density	0.01	0.03	0.07	0.00	0.39	0.05	0.07	0.07	0.00	0.06	0.08	0.02	0.05	0.01	0.03	1.00	0.13
chrX_mutation_density	0.10	0.08	0.00	0.10	0.51	0.11	0.09	0.10	0.00	0.13	0.18	0.07	0.10	0.00	0.12	0.13	1.00

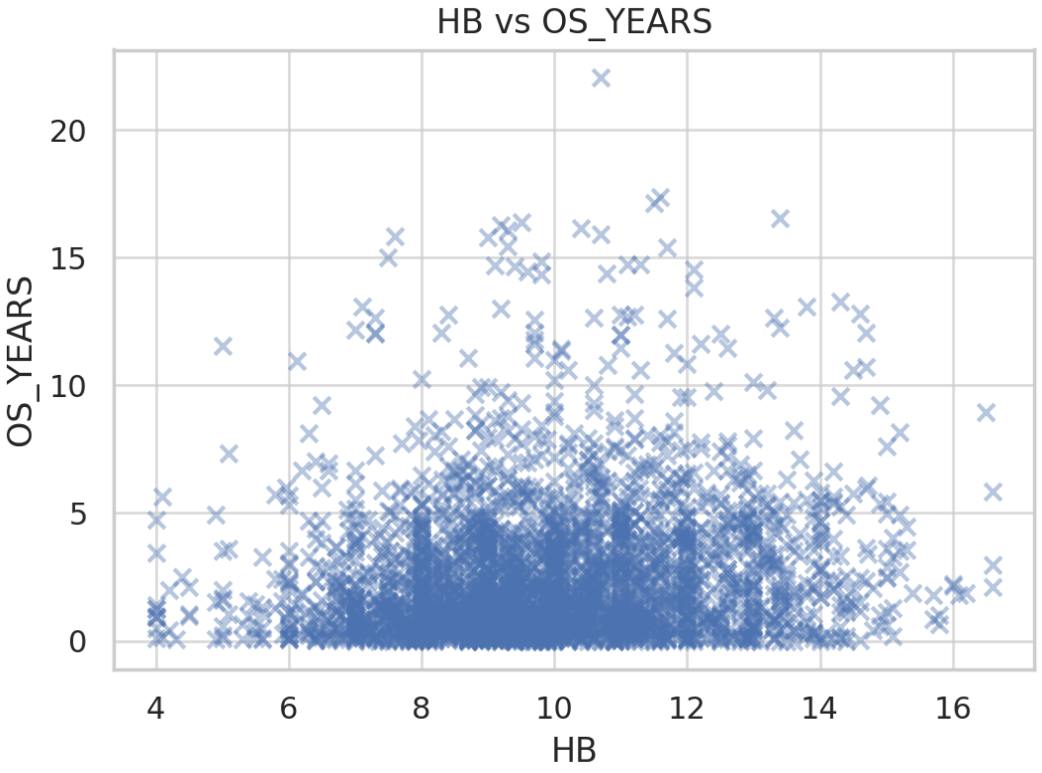
This heatmap shows pairwise correlations between engineered features. A high correlation (> 0.8) was detected between 'num_deletions' and 'frac_deletions', which may affect model stability. All other features show acceptable independence.

BM_BLAST vs OS_YEARS



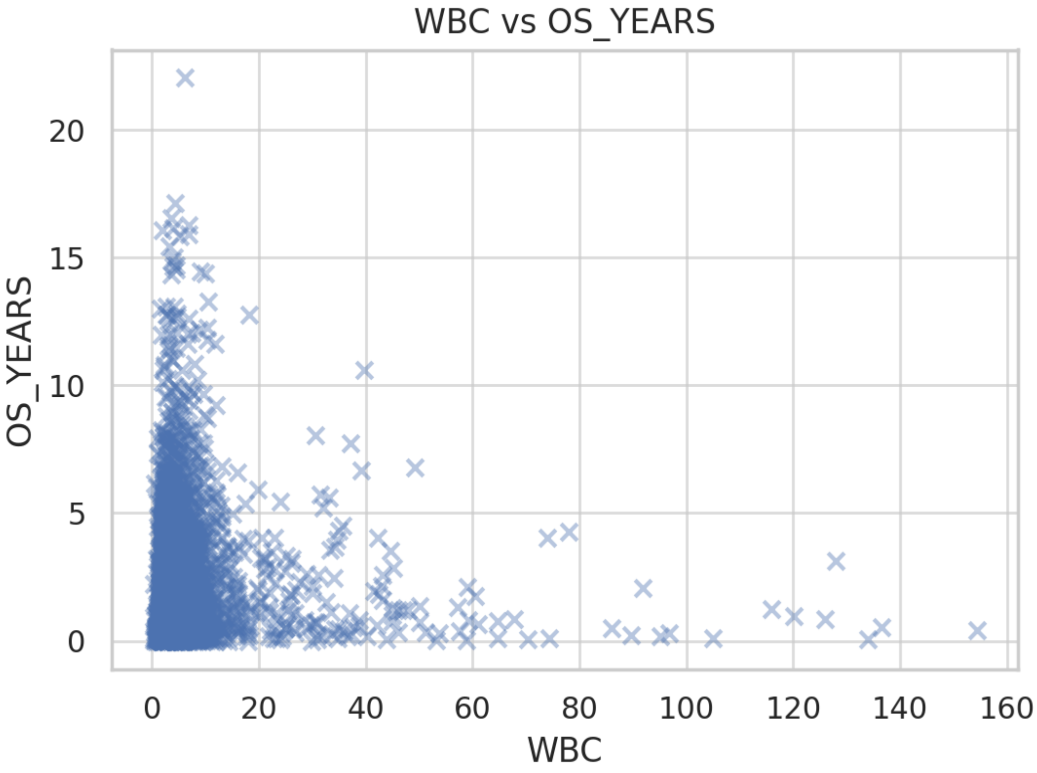
The scatter plot above shows the relationship between BM_BLAST and overall survival (OS_YEARS). This helps assess whether BM_BLAST might influence survival duration.

HB vs OS_YEARS



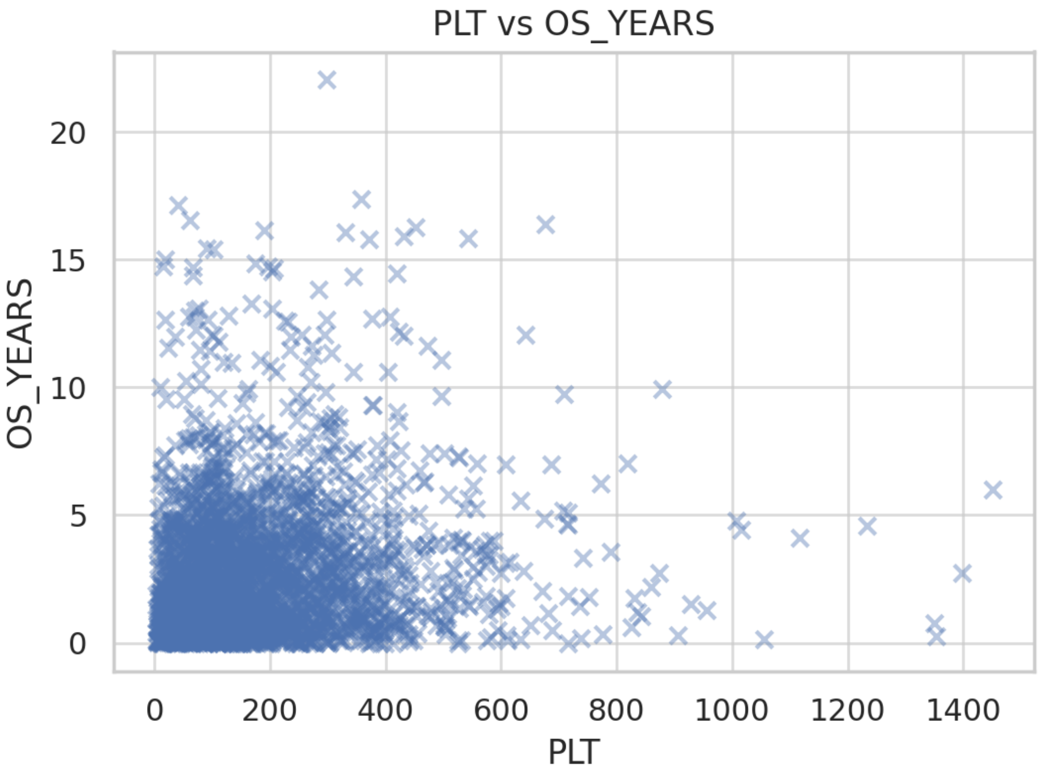
The scatter plot above shows the relationship between HB and overall survival (OS_YEARS). This helps assess whether HB might influence survival duration.

WBC vs OS_YEARS



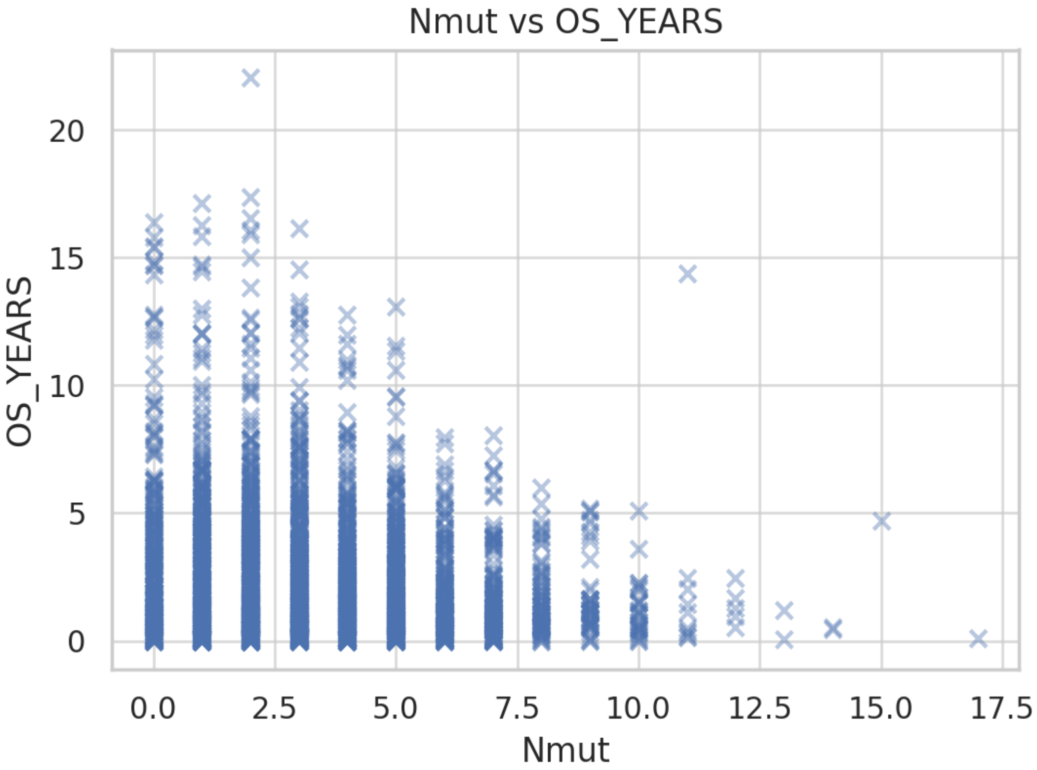
The scatter plot above shows the relationship between WBC and overall survival (OS_YEARS). This helps assess whether WBC might influence survival duration.

PLT vs OS_YEARS



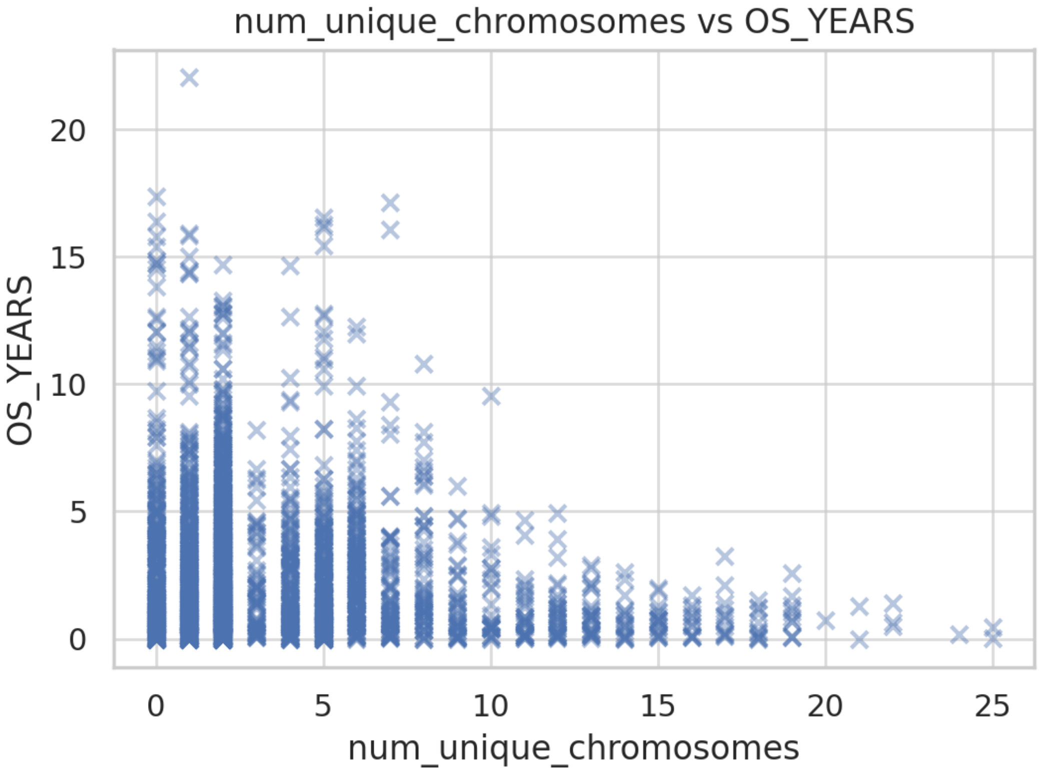
The scatter plot above shows the relationship between PLT and overall survival (OS_YEARS). This helps assess whether PLT might influence survival duration.

Nmut vs OS_YEARS



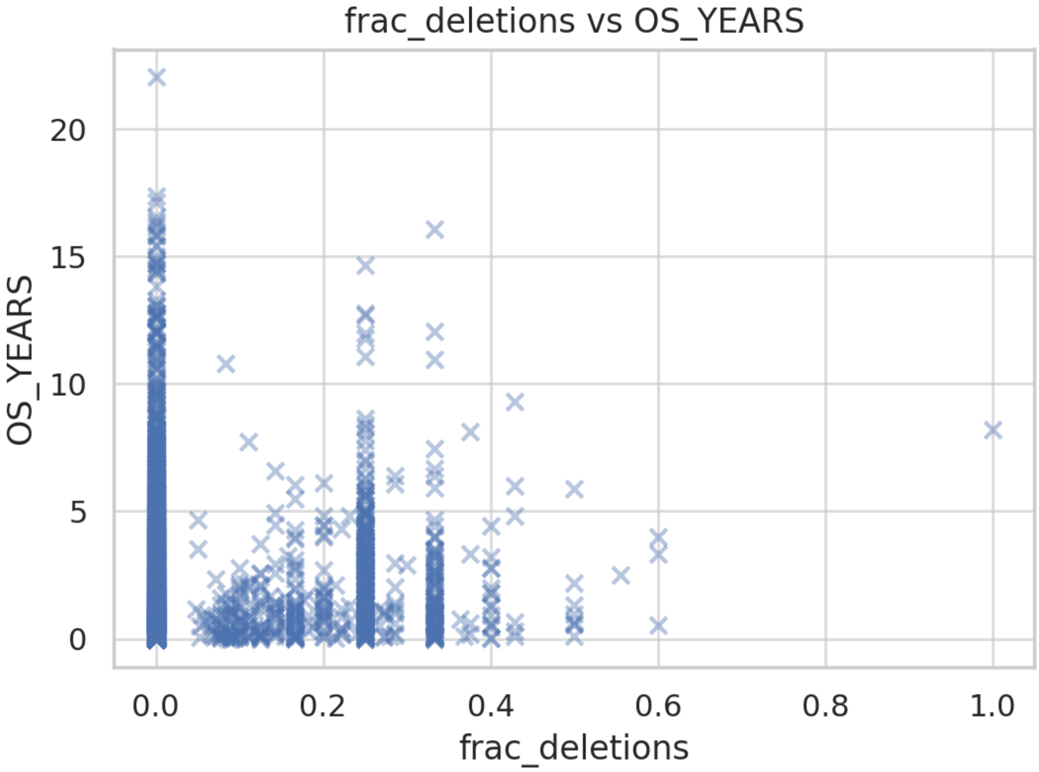
The scatter plot above shows the relationship between Nmut and overall survival (OS_YEARS). This helps assess whether Nmut might influence survival duration.

num_unique_chromosomes vs OS_YEARS



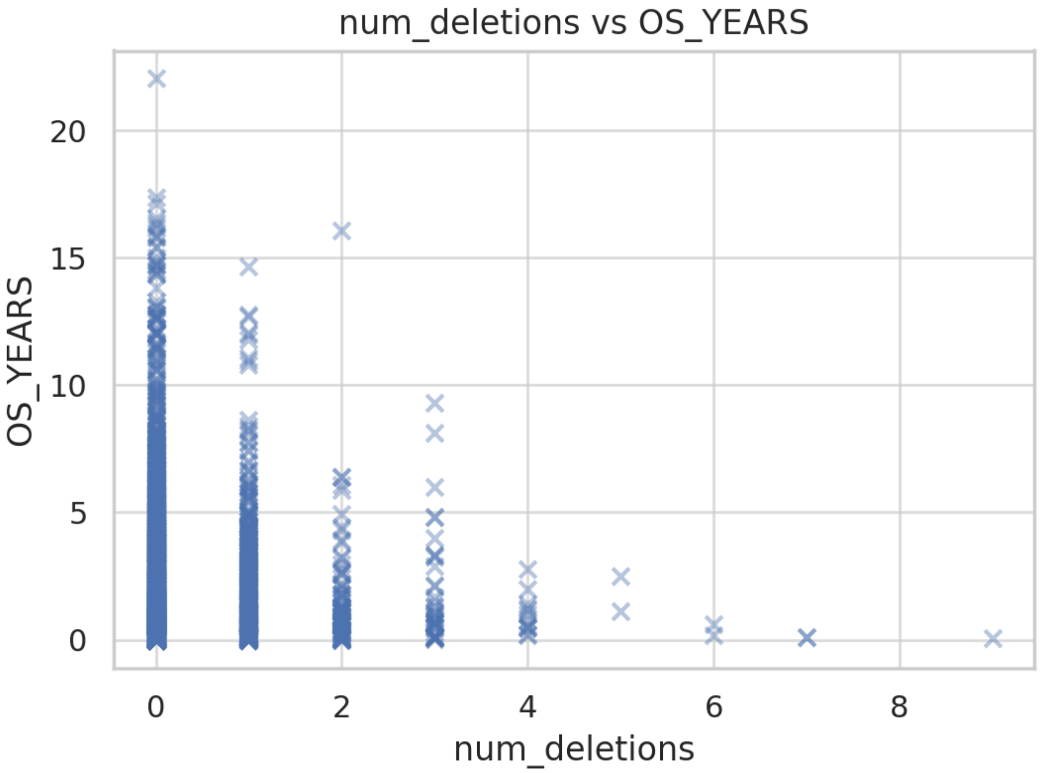
The scatter plot above shows the relationship between num_unique_chromosomes and overall survival (OS_YEARS). This helps assess whether num_unique_chromosomes might influence survival duration.

frac_deletions vs OS_YEARS



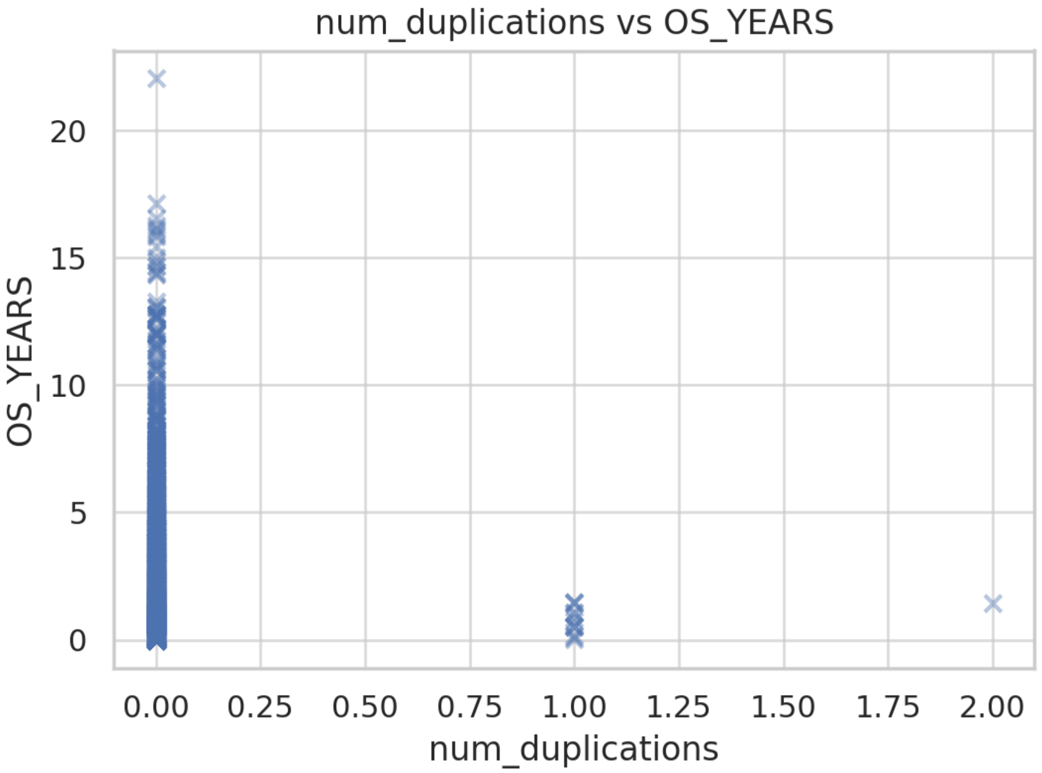
The scatter plot above shows the relationship between frac_deletions and overall survival (OS_YEARS). This helps assess whether frac_deletions might influence survival duration.

num_deletions vs OS_YEARS



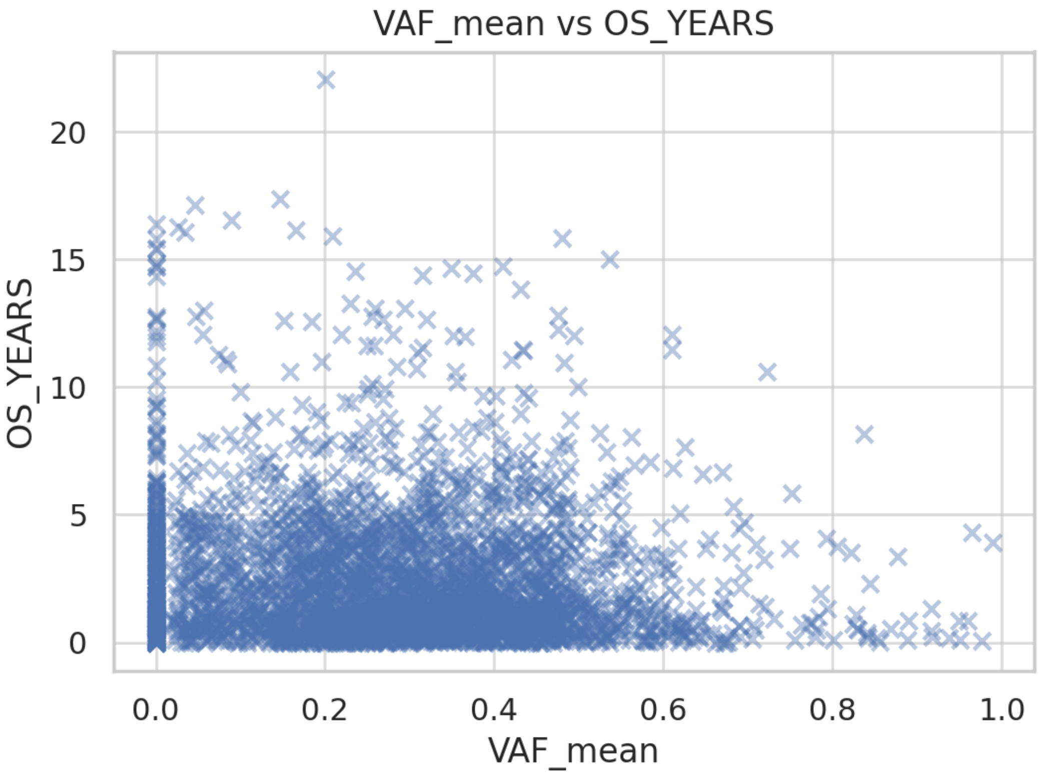
The scatter plot above shows the relationship between num_deletions and overall survival (OS_YEARS). This helps assess whether num_deletions might influence survival duration.

num_duplications vs OS_YEARS



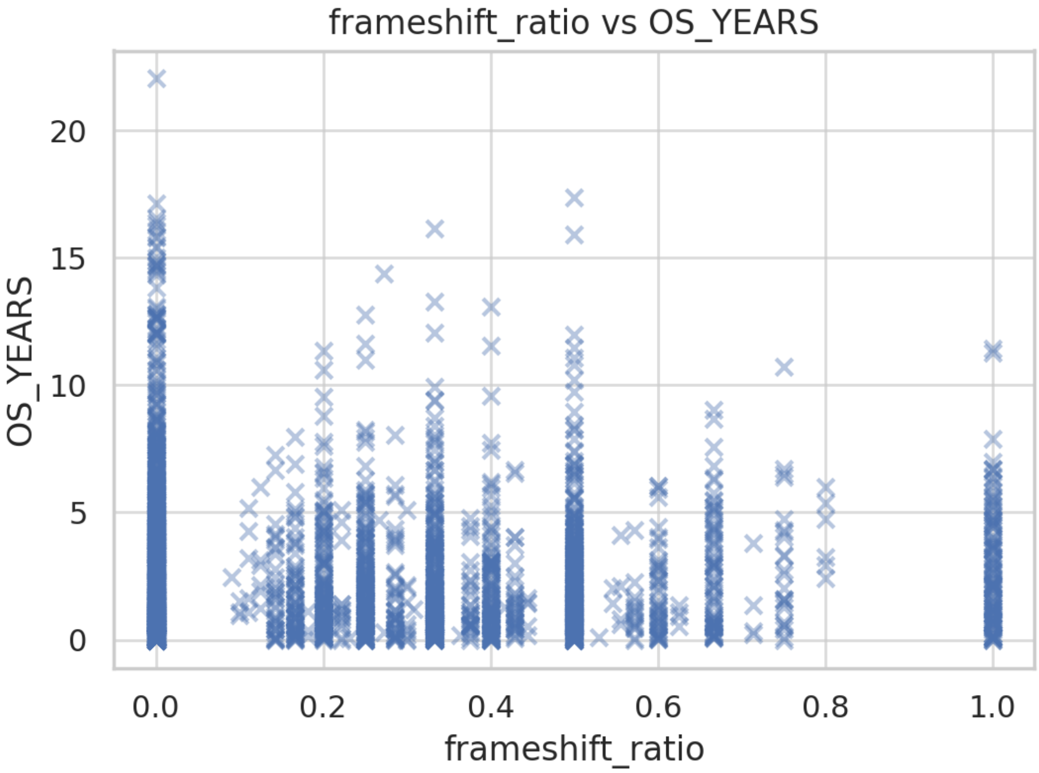
The scatter plot above shows the relationship between num_duplications and overall survival (OS_YEARS). This helps assess whether num_duplications might influence survival duration.

VAF_mean vs OS_YEARS



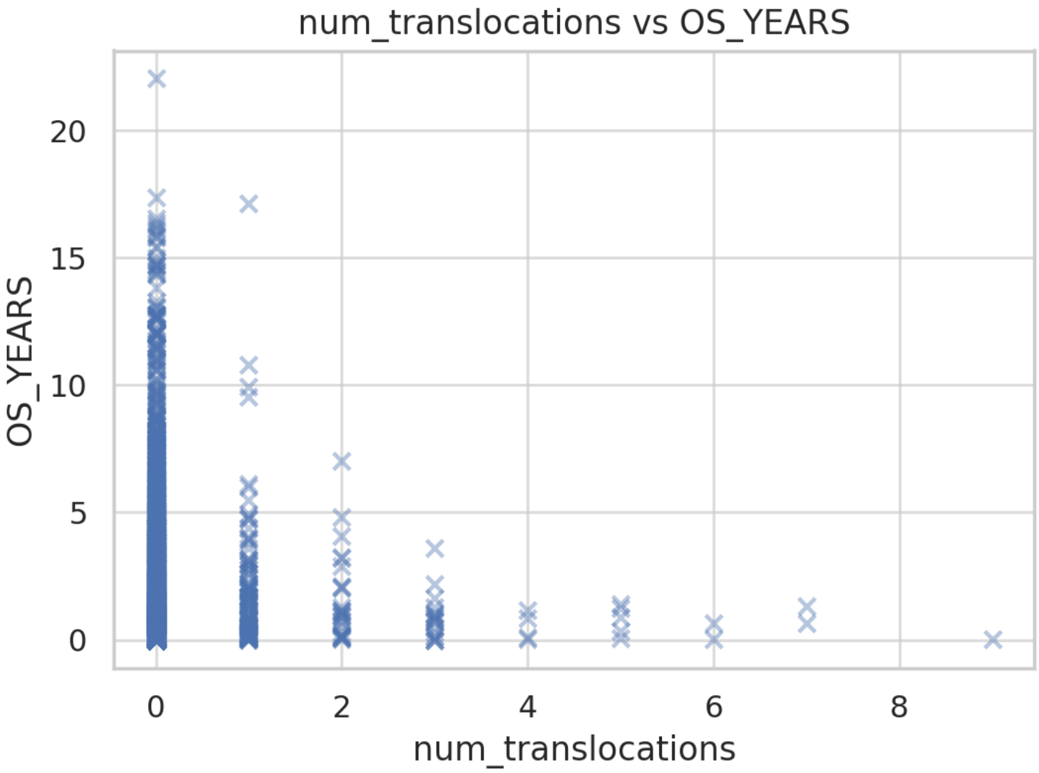
The scatter plot above shows the relationship between VAF_mean and overall survival (OS_YEARS). This helps assess whether VAF_mean might influence survival duration.

frameshift_ratio vs OS_YEARS



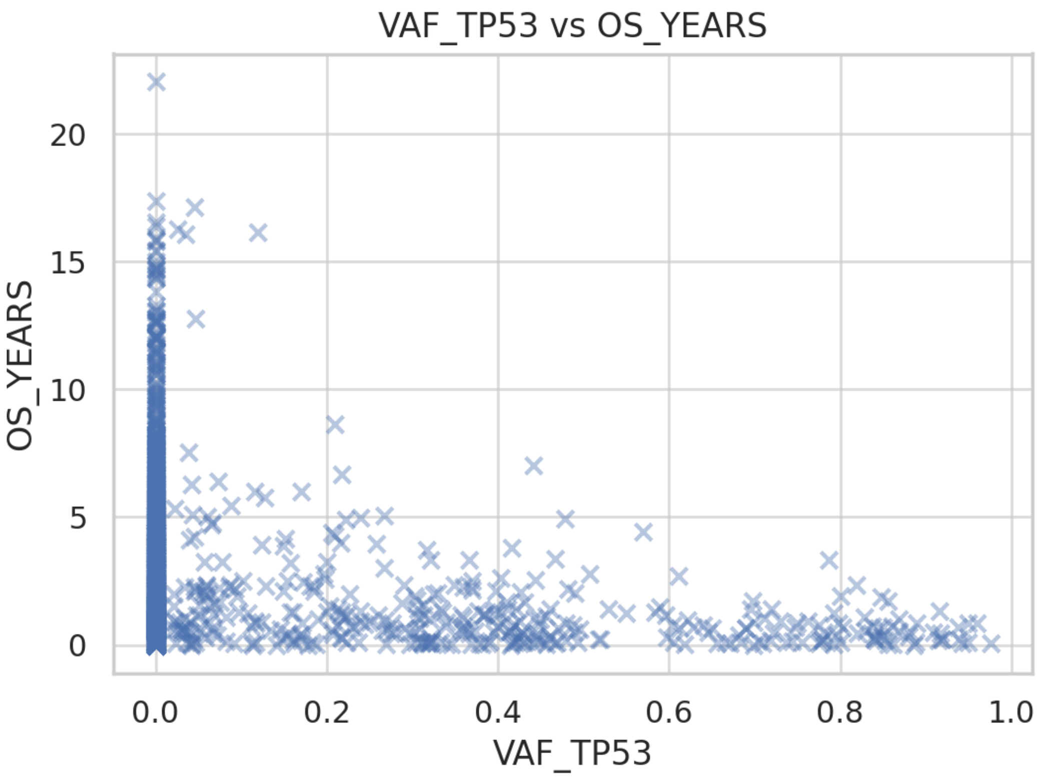
The scatter plot above shows the relationship between frameshift_ratio and overall survival (OS_YEARS). This helps assess whether frameshift_ratio might influence survival duration.

num_translocations vs OS_YEARS



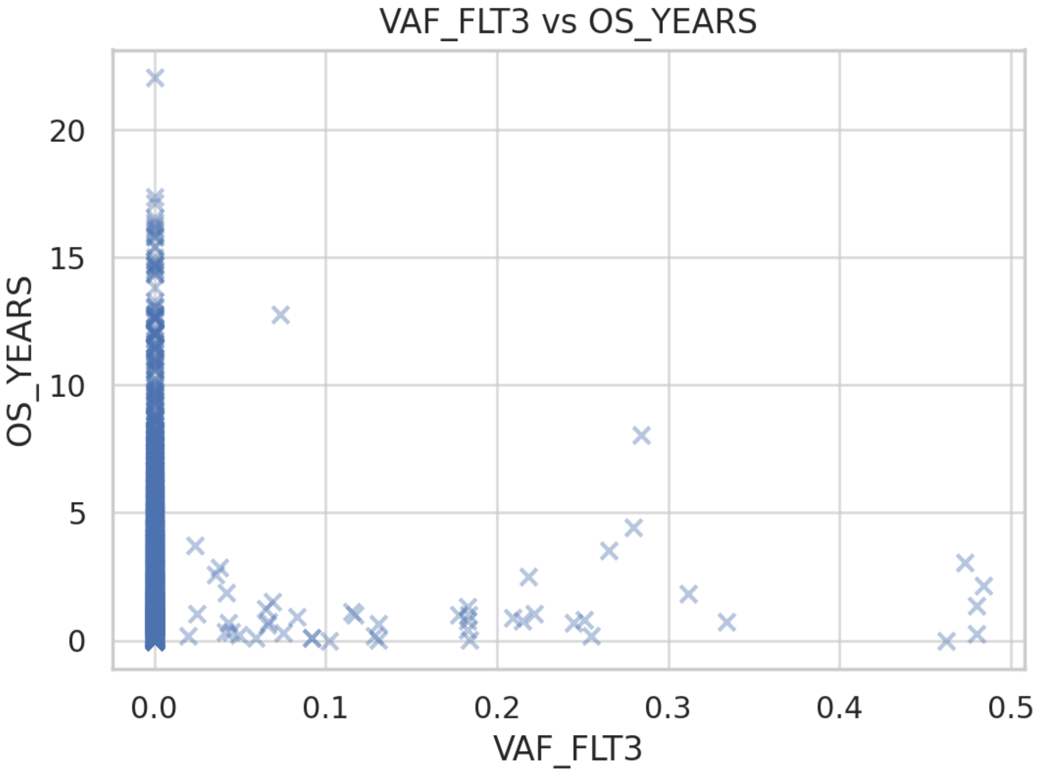
The scatter plot above shows the relationship between num_translocations and overall survival (OS_YEARS). This helps assess whether num_translocations might influence survival duration.

VAF_TP53 vs OS_YEARS



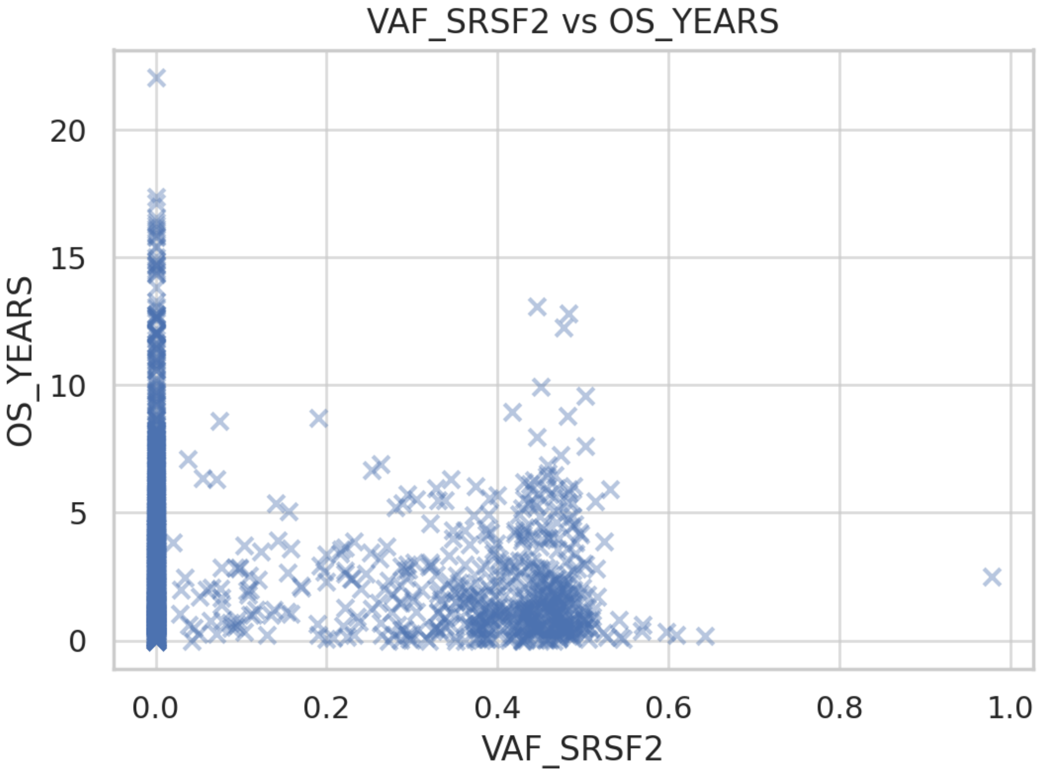
The scatter plot above shows the relationship between VAF_TP53 and overall survival (OS_YEARS). This helps assess whether VAF_TP53 might influence survival duration.

VAF_FLT3 vs OS_YEARS



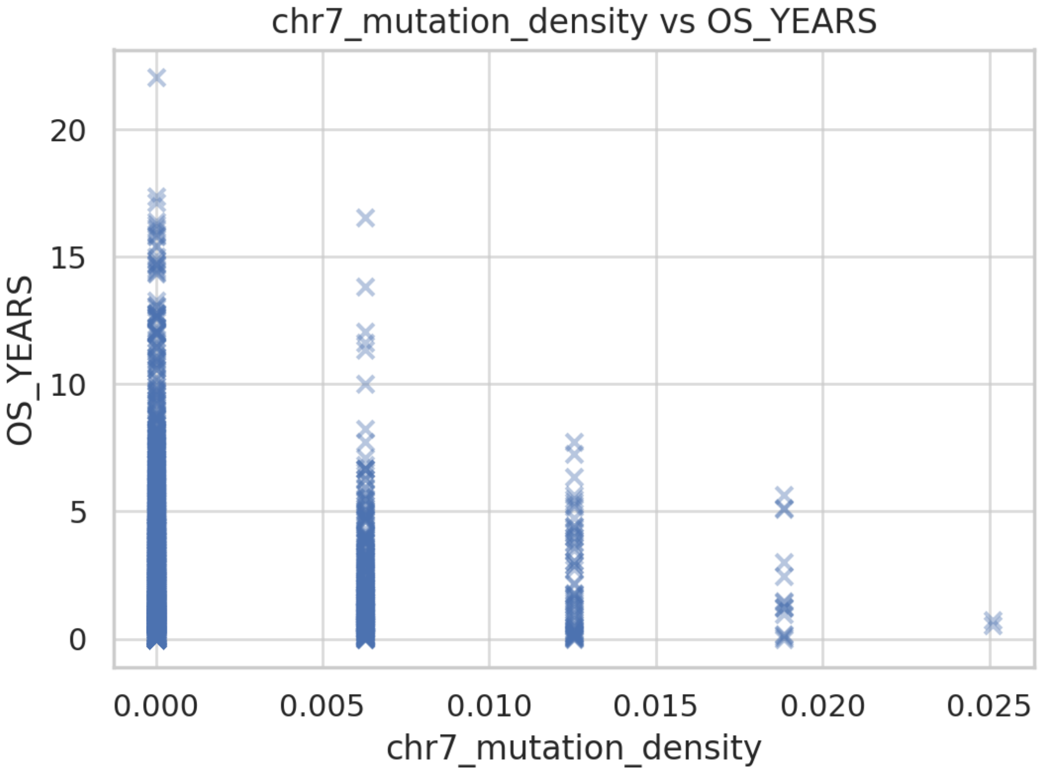
The scatter plot above shows the relationship between VAF_FLT3 and overall survival (OS_YEARS). This helps assess whether VAF_FLT3 might influence survival duration.

VAF_SRSF2 vs OS_YEARS



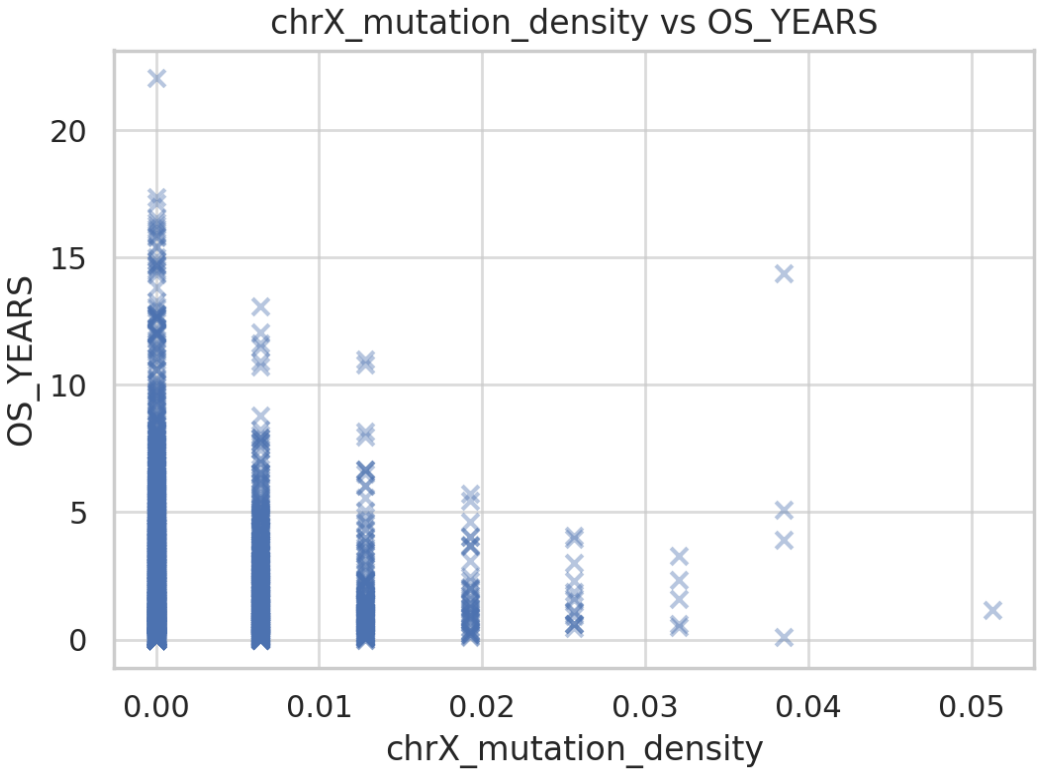
The scatter plot above shows the relationship between VAF_SRSF2 and overall survival (OS_YEARS). This helps assess whether VAF_SRSF2 might influence survival duration.

chr7_mutation_density vs OS_YEARS



The scatter plot above shows the relationship between chr7_mutation_density and overall survival (OS_YEARS). This helps assess whether chr7_mutation_density might influence survival duration.

chrX_mutation_density vs OS_YEARS



The scatter plot above shows the relationship between chrX_mutation_density and overall survival (OS_YEARS). This helps assess whether chrX_mutation_density might influence survival duration.

Conclusion

In conclusion, this exploratory analysis has highlighted several engineered features with strong clinical relevance and potential predictive value for survival in adult myeloid leukemia. The inclusion of gene-specific VAFs and chromosome-specific mutation densities provides a genomic dimension that complements clinical observations. The derived features have demonstrated acceptable multicollinearity and largely low inter-feature correlation, making them well-suited for integration into survival models such as Extra Survival Trees or Cox Proportional Hazards. These findings will guide the subsequent model development phase to improve personalized prognosis and treatment planning in hematologic oncology.