Implement filter method on training table

Author: Xin Wei a1912958
Team: RAIL-PG-2

In this Sprint, my main contributions were focused on the feature selection task. First, I excluded all non-numerical features and, by analyzing the distribution of the target variable, identified the issue of sample ratio imbalance. In subsequent work, the impact of this problem on the model needs to be given due attention.

```
Number of numerical features: 36
Total sample size: 1065013
Sample distribution:
0    944261
1    120752
Name: Tc_target, dtype: int64
```

By generating descriptive statistical tables and distribution histograms of these numerical features, I further excluded two numerical features that had a relatively minor impact on the model.

```python
desc = X_num.describe().T
desc["missing_ratio"] = X_num.isna().mean()
desc["unique_count"] = X_num.nunique()
desc.reset_index(inplace=True)
desc.rename(columns={"index": "Feature"}, inplace=True)
print(f"Descriptive Statistics Table:")
display(desc)
```

desc: pandas.core.frame.DataFrame = [Feature: object, count: float64 ... 9 more fields]

Descriptive Statistics Table:

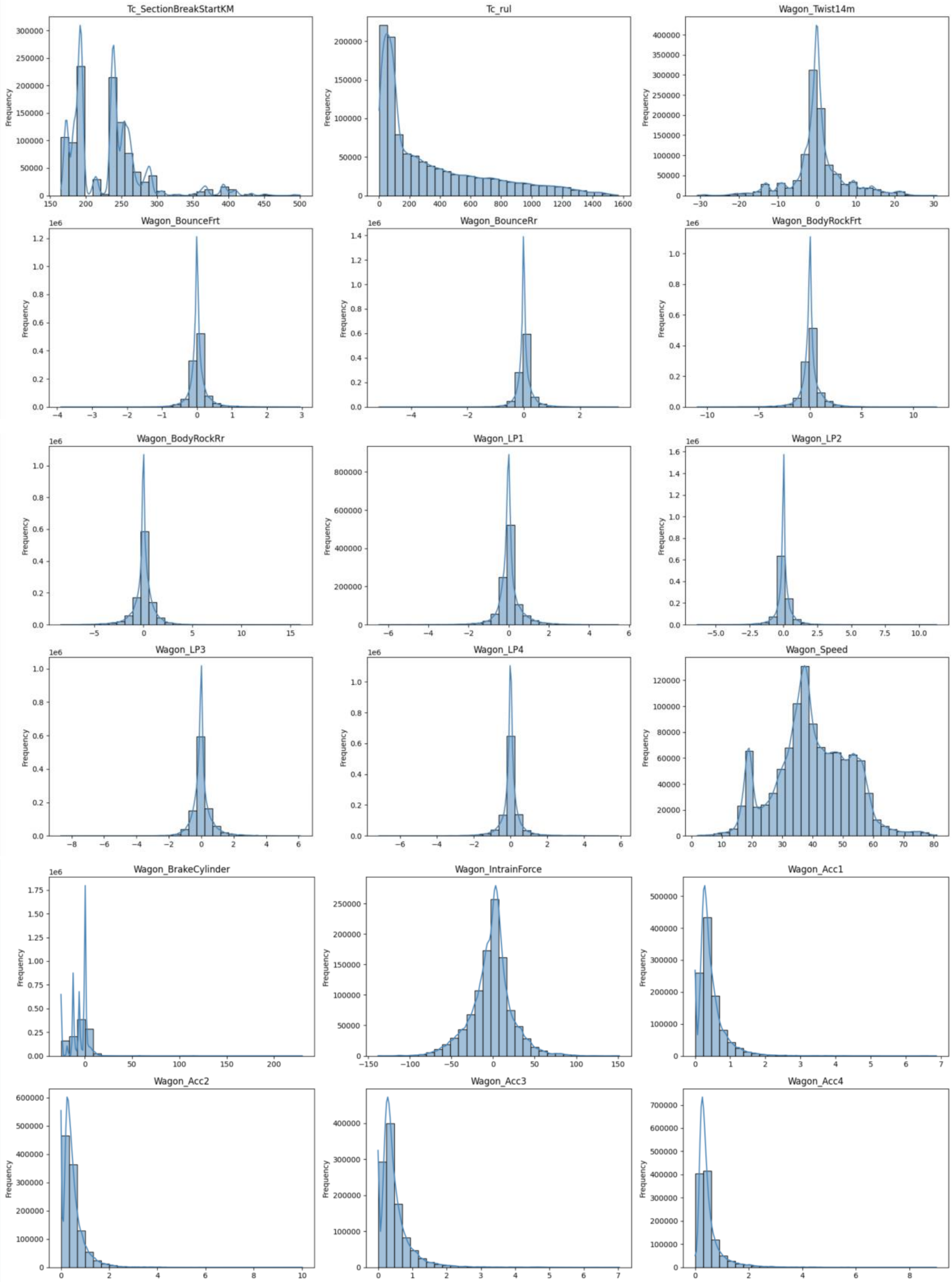
Table

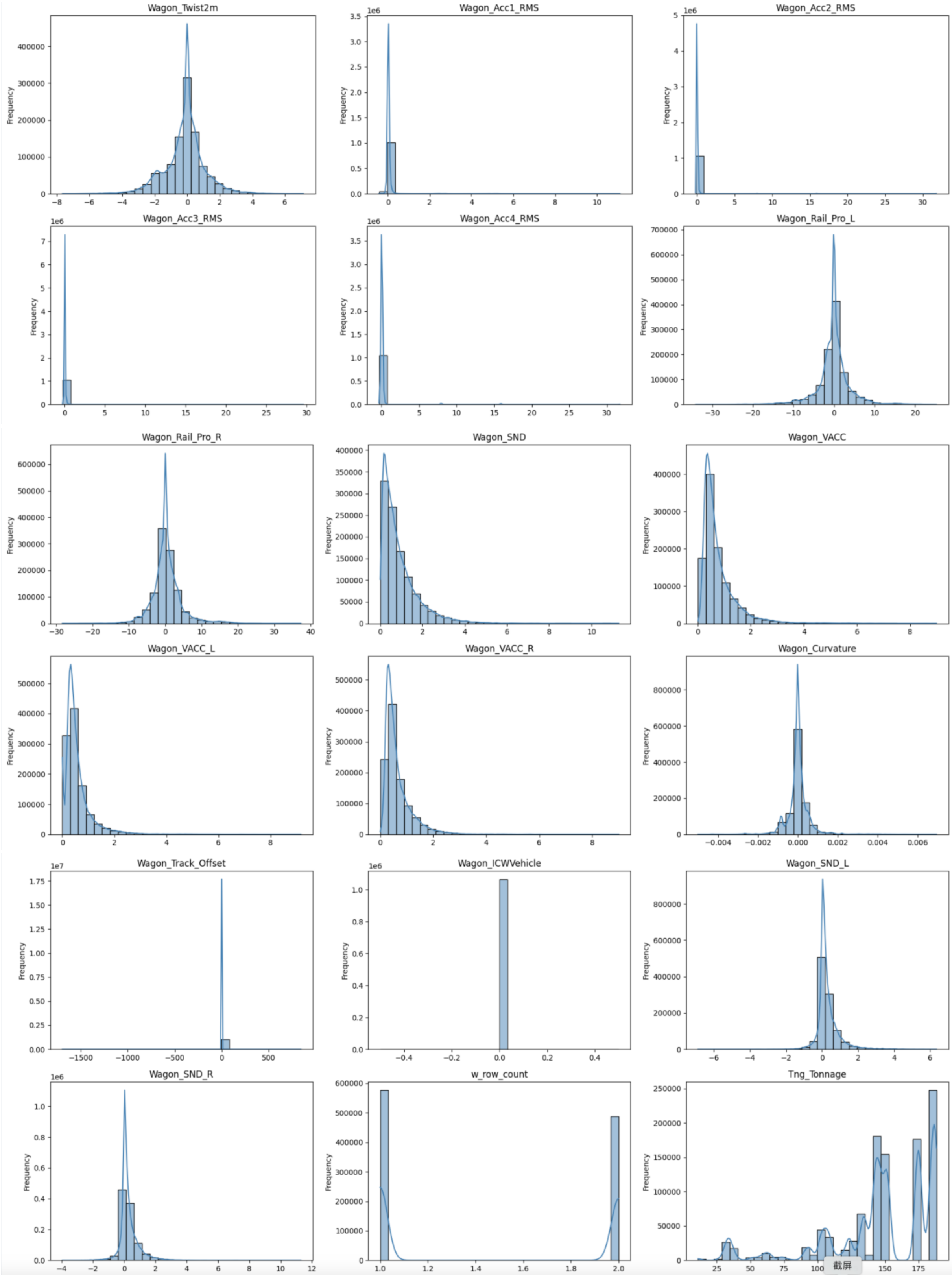
	Feature	count	mean	std	min	25%	50%	75%	max
1	Tc_SectionBreakStartKM	1065013	231.7947490781802	52.17856043452325	165.01	192.02	238.3	253.73	500.27
2	Tc_rul	1065013	342.155538946473	362.8119755769241	1	65	182	543	1568
3	Wagon_Twist14m	1065013	0.3081851591546329	6.870448905456213	-30.683843975000002	-1.941045210375	0	2.2609647295	30.72042675
4	Wagon_BounceFrt	1065013	0.020407908145347462	0.28147693557369263	-3.8921581625	-0.06550402731125	0	0.08442983249999998	2.9464458575
5	Wagon_BounceRr	1065013	0.015614604084433215	0.3013044744165423	-5.136549815	-0.0747475305	0	0.08442129055	3.35746153
6	Wagon_BodyRockFrt	1065013	0.012436152325886152	1.0454949105850977	-10.934176035	-0.28629674	0	0.2817801605	12.26408085
7	Wagon_BodyRockRr	1065013	-0.005714067719005067	1.0854381663954402	-8.414818433466667	-0.28793692975	0	0.37246725112500007	16.000335999999997
8	Wagon_LP1	1065013	0.026463526813276086	0.5987222593105069	-6.460575815	-0.170727342	-0.005581715099999995	0.17214287755000002	5.4607134975000005
9	Wagon_LP2	1065013	0.009583567658789637	0.6221379361323564	-6.360016575	-0.14046705725000003	0	0.16589957116666665	11.29129695
10	Wagon_LP3	1065013	0.020668639565442817	0.626828607965538	-8.6637123575	-0.215467548	-0.008212978003750003	0.18195207199374996	6.089845609999999
11	Wagon_LP4	1065013	0.014151778622986057	0.6083128758370759	-7.145313145	-0.14215325325	0	0.14847469593749998	5.8514997275
12	Wagon_Speed	1065013	39.707285916463	12.267110957314044	1.9263749999999997	32.43	38.81675	48.69125	80.904
13	Wagon_BrakeCylinder	1065013	-5.822940472043602	10.457170426062596	-25.147091924999998	-12.2664914671	-2.8220904807999996	0.45656554699999996	229.75761
14	Wagon_IntrainForce	1065013	-1.7744660419866962	25.181440233878217	-138.39074575	-14.180494599874999	0.113047638	10.4822166075	151.82294325
15	Wagon_Acc1	1065013	0.45744584055320964	0.4152906142005321	0	0.23177733475	0.3501972013333333	0.561348819625	6.871273239999999

36 rows | 3.48s runtime

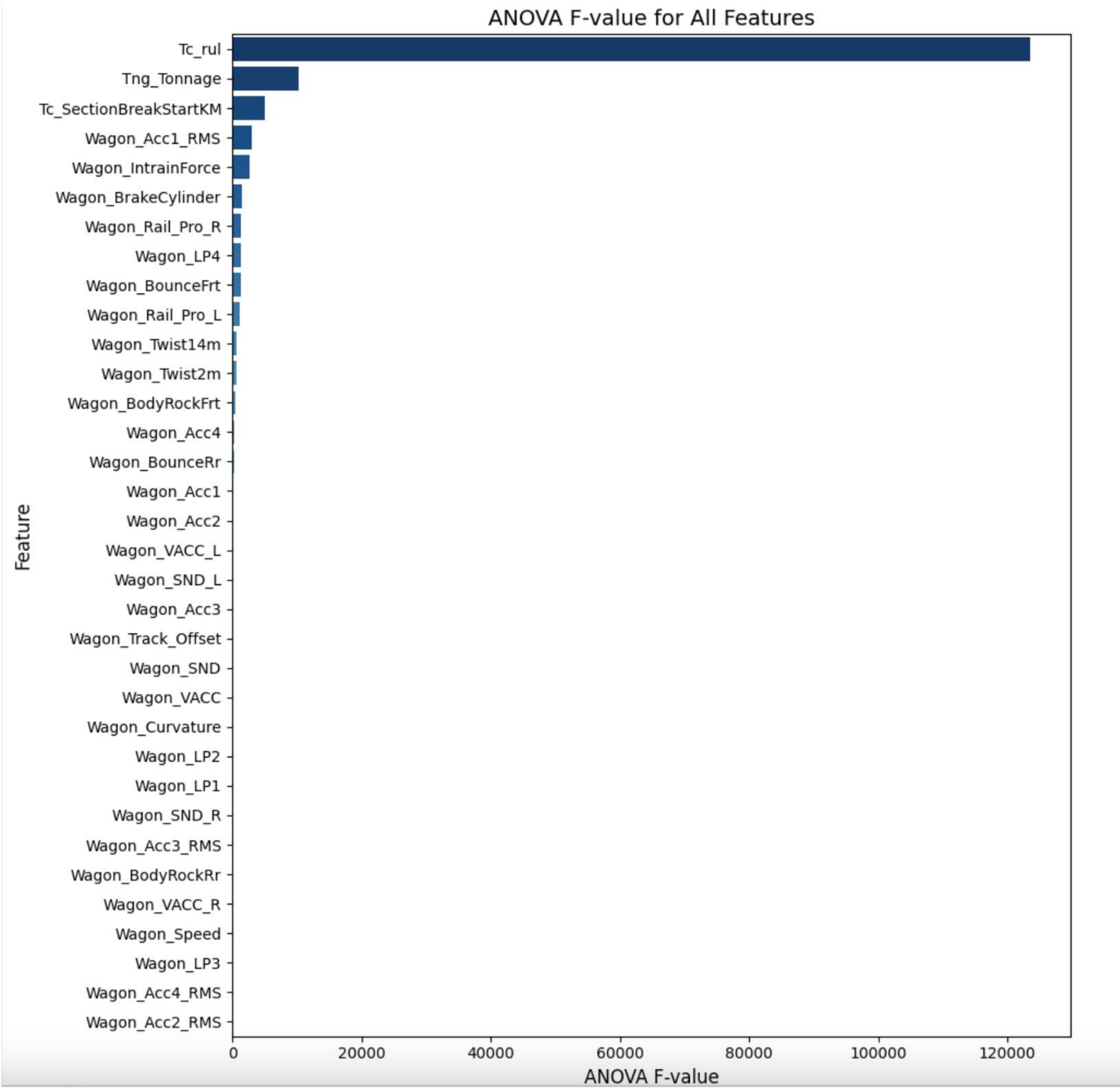
Refreshed 2 days ago

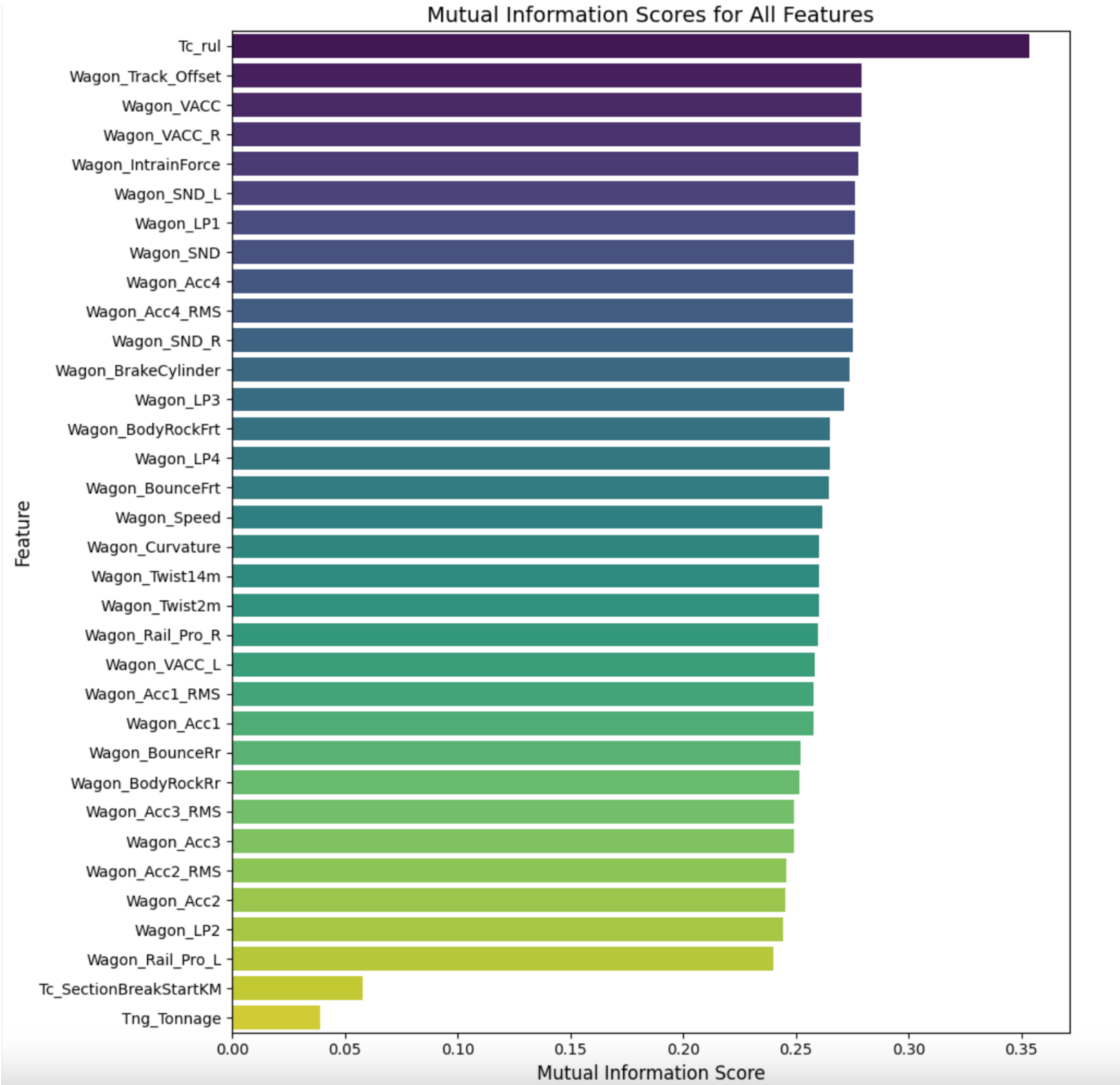
Univariate Distribution Histogram:





After the initial screening of numerical features, it was observed that the majority of features have a unique_count value exceeding 30,000. This indicates that these features possess a rich range of values within the sample space, and their distribution is closer to that of continuous variables rather than finite categorical (discrete) variables. Based on this characteristic, the subsequent analysis will employ methods suitable for continuous features with a discrete target variable, namely ANOVA F-test and Mutual Information.





ANOVA is effective in detecting linear mean differences, while MI is better suited for uncovering nonlinear dependencies. Using both methods in tandem provides a more comprehensive and balanced assessment of feature importance, reducing the risk of bias from relying on a single method.