

```
In [1]: from analysis import *

drop_cols = [
    "TWD", "COG", "TWA", "TimeUTC", "SecondsSince1970", "ISODateTimeUTC", "Latitude", "Longitude", "Euler_X (deg)", "Euler_Y (deg)",
    "Lat", "LatBow", "LatCenter", "LatStern", "Lon", "LonBow", "LonCenter", "LonStern",
    "Leg", "Log", "LogAlongCourse", "MagneticVariation", "Rank", "TimeLocal",
    "DistanceToLeader", "interval_id", "boat_name", "interval_duration",
    "Heel", "Heel_Lwd", "Line_R", "Line_L", "BelowLineCalc", "VMC", "XTE", "VMG", "gain_forward", "gain_lateral", "gain_vmg", "Total_lines", "LoadCell_1", "LoadCell_2", "LoadCell_3", "LoadCell_4"
]
df = pd.read_csv("all_data.csv")
MAX_NULL_RATIO = 0.2
```

I. All together

```
In [2]: df_numeric = df.select_dtypes(include=["float64", "int64"]).copy()
df_numeric.drop(columns=[c for c in drop_cols if c in df_numeric.columns], inplace=True)
df_numeric.dropna(subset=["SOG"], inplace=True)

print(f"Variables utilisées: {df_numeric.columns.tolist()}")
print(f"Number of rows after filtering: {len(df_numeric)}")
```

Variables utilisées: ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'SOG', 'boat_weight', 'side_line2']
Number of rows after filtering: 80131

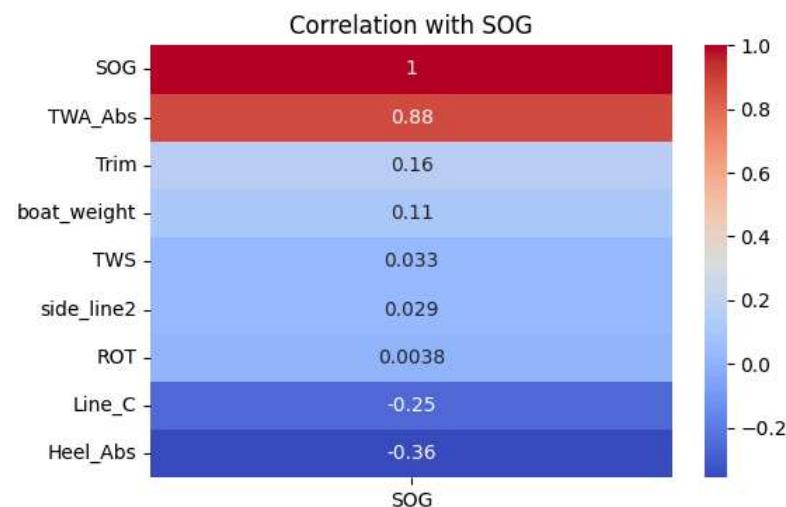
```
In [3]: full_analysis(df_numeric, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Data cleaning:
- Initial samples: 80131
- Removed samples with NaNs: 2598
- Final samples: 77533

Correlation with SOG:
```



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
TWA_Abs	254200.988723	1.0	221120.429266	0.000000e+00	0.740414
TWS	3828.508929	1.0	3330.284206	0.000000e+00	0.041189
boat_weight	3793.260587	1.0	3299.622923	0.000000e+00	0.040825
side_line2	3339.027728	1.0	2904.501860	0.000000e+00	0.036113
Trim	1885.045734	1.0	1639.734464	0.000000e+00	0.020713
Line_C	1724.519306	1.0	1500.098214	0.000000e+00	0.018983
Heel_Abs	808.494576	1.0	703.280773	2.836483e-154	0.008990
ROT	68.871624	1.0	59.908985	1.005414e-14	0.000772
Residual	89121.921096	77524.0	NaN	NaN	0.500000

Polynomial fit:

Model performance:
 $R^2: 0.809$
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 2.412 * TWA_Abs
 0.614 * side_line2
 0.448 * Line_C
 0.237 * boat_weight
 0.227 * TWS
 -0.169 * Trim
 0.127 * Heel_Abs
 0.030 * ROT

	feature	coefficient
4	TWA_Abs	2.412122
7	side_line2	0.613774
1	Line_C	0.447513
6	boat_weight	0.237152
5	TWS	0.227346
3	Trim	-0.169343
0	Heel_Abs	0.127094
2	ROT	0.029847

II Upwind:

II.1. All upwind data

```
In [4]: upwind_data = df[df['TWA'] >= 0]
df_numeric_upwind = upwind_data.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_upwind.drop(columns=[c for c in drop_cols if c in df_numeric_upwind.columns], inplace=True)
df_numeric_upwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_upwind)}")
```

Number of rows after filtering: 48787

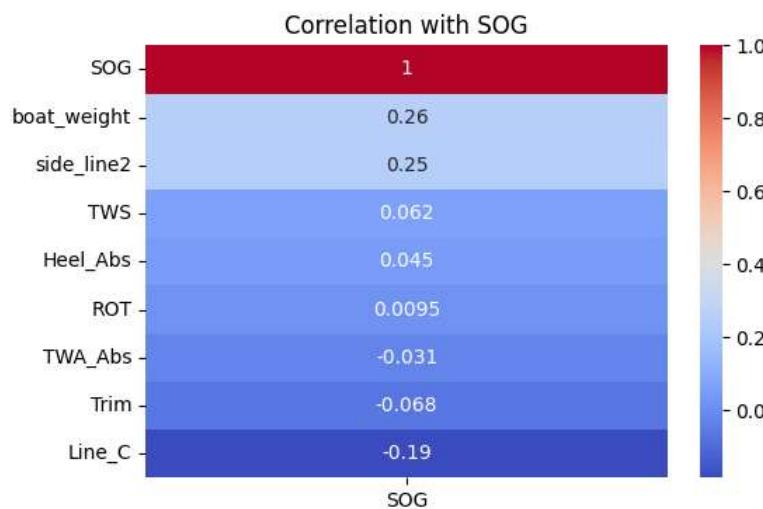
```
In [5]: full_analysis(df_numeric_upwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Data cleaning:
- Initial samples: 48787
- Removed samples with NaNs: 1096
- Final samples: 47691
```

Correlation with SOG:



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
boat_weight	2609.591463	1.0	2429.310621	0.000000e+00	0.048478
side_line2	1922.353046	1.0	1789.549336	0.000000e+00	0.036173
Line_C	896.899769	1.0	834.938405	5.115309e-182	0.017209
Trim	769.317790	1.0	716.170291	1.315186e-156	0.014797
TWS	318.629387	1.0	296.617216	2.855190e-66	0.006182
TWA_Abs	169.325415	1.0	157.627749	4.257421e-36	0.003295
Heel_Abs	97.892644	1.0	91.129836	1.406612e-21	0.001908
ROT	26.345069	1.0	24.525048	7.360030e-07	0.000514
Residual	51220.514612	47682.0	NaN	NaN	0.500000

Polynomial fit:

Model performance:
 $R^2: 0.141$
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 $0.631 * \text{side_line2}$
 $0.428 * \text{Line_C}$
 $0.247 * \text{boat_weight}$
 $-0.131 * \text{Trim}$
 $0.089 * \text{TWS}$
 $0.066 * \text{TWA_Abs}$
 $0.048 * \text{Heel_Abs}$
 $0.024 * \text{ROT}$

	feature	coefficient
7	side_line2	0.630941
1	Line_C	0.428131
6	boat_weight	0.247453
3	Trim	-0.130755
5	TWS	0.089033
4	TWA_Abs	0.066272
0	Heel_Abs	0.048433
2	ROT	0.023993

II.2. Upwind: Gian vs Karl

II.2.1. Upwind: Gian

```
In [6]: gian_data_upwind = upwind_data[
    (upwind_data['boat_name'] == "Gian Stragiotti") |
    ((upwind_data['boat_name'] == "SenseBoard") & (upwind_data['opponent_name'] == "Karl Maeder"))
]
df_numeric_gian_upwind = gian_data_upwind.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_gian_upwind.drop(columns=[c for c in drop_cols if c in df_numeric_gian_upwind.columns], inplace=True)
df_numeric_gian_upwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_gian_upwind)})")
```

Number of rows after filtering: 24390

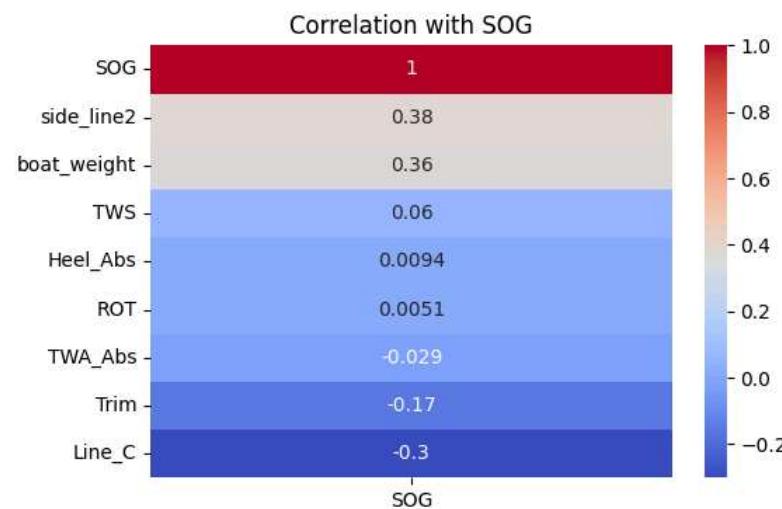
```
In [7]: full_analysis(df_numeric_gian_upwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Data cleaning:
- Initial samples: 24390
- Removed samples with NaNs: 187
- Final samples: 24203

Correlation with SOG:
```



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
side_line2	2473.857226	1.0	2510.891887	0.000000e+00	0.094024
Line_C	1386.804632	1.0	1407.565668	1.783505e-299	0.054980
TWS	651.785654	1.0	661.543154	5.935661e-144	0.026616
Trim	630.165221	1.0	639.599054	2.635748e-139	0.025755
boat_weight	168.776892	1.0	171.303551	5.212697e-39	0.007031
TWA_Abs	87.892768	1.0	89.208559	3.863905e-21	0.003674
ROT	4.681050	1.0	4.751127	2.928872e-02	0.000196
Heel_Abs	0.666845	1.0	0.676828	4.106892e-01	0.000028
Residual	23837.148080	24194.0	Nan	Nan	0.500000

Polynomial fit:

Model performance:
 $R^2: 0.247$
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 1.134 * side_line2
 0.789 * Line_C
 0.204 * TWS
 -0.164 * Trim
 0.129 * boat_weight
 0.067 * TWA_Abs
 0.014 * ROT
 -0.006 * Heel_Abs

	feature	coefficient
7	side_line2	1.134301
1	Line_C	0.789013
5	TWS	0.204137
3	Trim	-0.163660
6	boat_weight	0.129031
4	TWA_Abs	0.066695
2	ROT	0.014175
0	Heel_Abs	-0.005710

II.2.2. Upwind: Karl

```
In [8]: karl_data_upwind = upwind_data[
    (upwind_data['boat_name'] == "Karl Maeder") |
    ((upwind_data['boat_name'] == "SenseBoard") & (upwind_data['opponent_name'] == "Gian Stragiotti"))
]
df_numeric_karl_upwind = karl_data_upwind.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_karl_upwind.drop(columns=[c for c in drop_cols if c in df_numeric_karl_upwind.columns], inplace=True)
df_numeric_karl_upwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_karl_upwind)}")
```

Number of rows after filtering: 24397

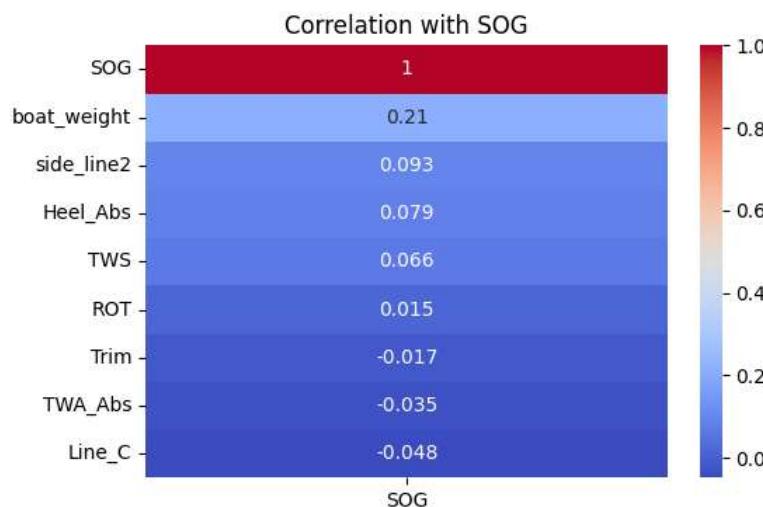
```
In [9]: full_analysis(df_numeric_karl_upwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Data cleaning:
- Initial samples: 24397
- Removed samples with NaNs: 909
- Final samples: 23488
```

Correlation with SOG:



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
boat_weight	1622.763434	1.0	1534.448387	0.000000e+00	0.061345
TWS	312.057877	1.0	295.074868	9.836020e-66	0.012412
Heel_Abs	284.674538	1.0	269.181802	3.700284e-60	0.011335
side_line2	181.121364	1.0	171.264264	5.365365e-39	0.007242
Line_C	102.292592	1.0	96.725561	8.813236e-23	0.004103
Trim	96.544845	1.0	91.290621	1.357890e-21	0.003873
TWA_Abs	73.304792	1.0	69.315353	8.843901e-17	0.002944
ROT	28.497316	1.0	26.946417	2.109160e-07	0.001146
Residual	24830.331859	23479.0	Nan	Nan	0.500000

Polynomial fit:

Model performance:
 $R^2: 0.087$
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 0.282 * boat_weight
 0.270 * side_line2
 0.201 * Line_C
 0.138 * TWS
 0.118 * Heel_Abs
 -0.066 * Trim
 0.065 * TWA_Abs
 0.036 * ROT

feature	coefficient
6 boat_weight	0.282362
7 side_line2	0.270422
1 Line_C	0.200587
5 TWS	0.137897
0 Heel_Abs	0.117738
3 Trim	-0.066073
4 TWA_Abs	0.064662
2 ROT	0.035691

II.2.3. Upwind: Karl vs Gian t-test

```
In [10]: t_test(df_numeric_gian_upwind, df_numeric_karl_upwind)

T-statistic: 31.779, p-value: 0.000000000000000
The difference is statistically significant, keeping data split.
```

III. Upwind: Master vs Slave

III.1. Master

```
In [11]: master_data_upwind = upwind_data[upwind_data['boat_role'] == "master"]
df_numeric_master_upwind = master_data_upwind.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_master_upwind.drop(columns=[c for c in drop_cols if c in df_numeric_master_upwind.columns], inplace=True)
df_numeric_master_upwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_master_upwind)}")
```

Number of rows after filtering: 24391

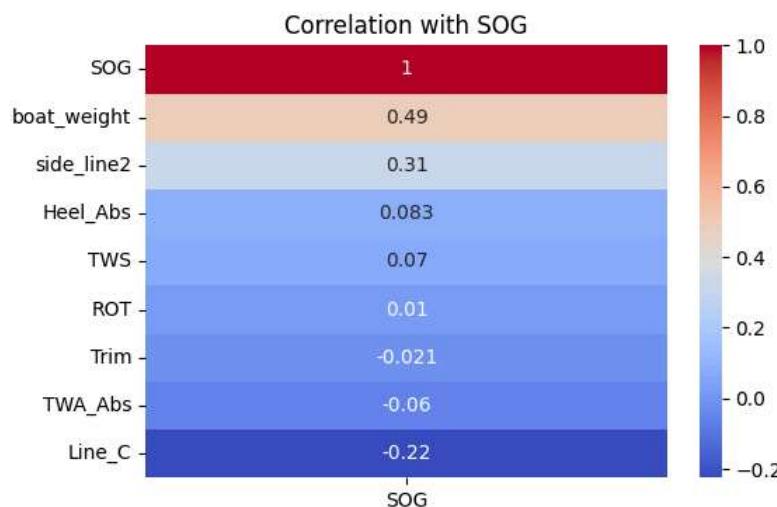
```
In [12]: full_analysis(df_numeric_master_upwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Data cleaning:
- Initial samples: 24391
- Removed samples with NaNs: 145
- Final samples: 24246

Correlation with SOG:
```



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
boat_weight	5884.476743	1.0	6111.225350	0.000000e+00	0.201370
side_line2	938.473825	1.0	974.636366	8.329006e-210	0.038658
Line_C	461.900622	1.0	479.699200	2.611363e-105	0.019408
Trim	285.468582	1.0	296.468643	4.790029e-66	0.012084
TWA_Abs	114.786159	1.0	119.209254	1.093372e-27	0.004894
Heel_Abs	92.135316	1.0	95.685599	1.482469e-22	0.003932
TWS	7.952304	1.0	8.258733	4.059167e-03	0.000341
ROT	0.521380	1.0	0.541470	4.618312e-01	0.000022
Residual	23337.719467	24237.0	Nan	Nan	0.500000

Polynomial fit:

Model performance:
 $R^2: 0.314$
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 0.599 * side_line2
 0.536 * boat_weight
 0.412 * Line_C
 -0.112 * Trim
 -0.079 * TWA_Abs
 0.066 * Heel_Abs
 0.020 * TWS
 0.005 * ROT

	feature	coefficient
7	side_line2	0.598989
6	boat_weight	0.536384
1	Line_C	0.411749
3	Trim	-0.111803
4	TWA_Abs	-0.079015
0	Heel_Abs	0.065715
5	TWS	0.019742
2	ROT	0.004735

II.3.2 Slave

```
In [13]: slave_data_upwind = upwind_data[upwind_data['boat_role'] == "slave"]
df_numeric_slave_upwind = slave_data_upwind.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_slave_upwind.drop(columns=[c for c in drop_cols if c in df_numeric_slave_upwind.columns], inplace=True)
df_numeric_slave_upwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_slave_upwind)})")
```

Number of rows after filtering: 24396

```
In [14]: full_analysis(df_numeric_slave_upwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

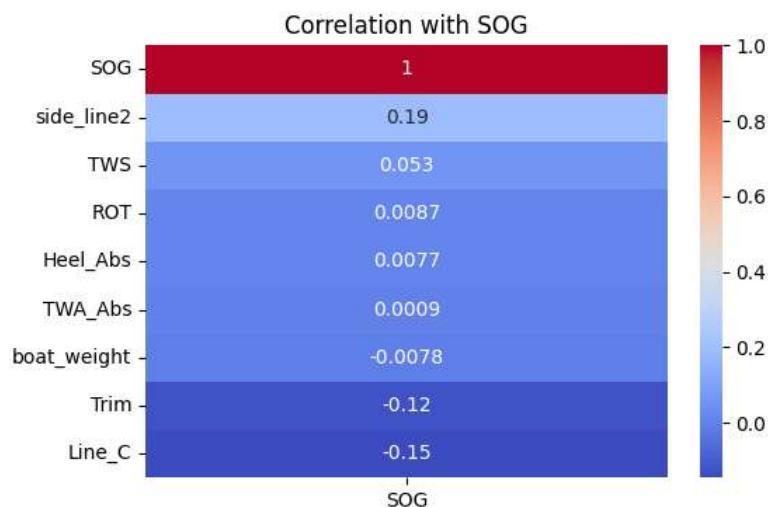
Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']
```

Data cleaning:

- Initial samples: 24396
- Removed samples with NaNs: 951
- Final samples: 23445

Correlation with SOG:



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
side_line2	734.530489	1.0	725.655649	1.966818e-157	0.030033
Trim	389.654239	1.0	384.946308	5.019941e-85	0.016160
Line_C	329.740541	1.0	325.756507	2.495429e-72	0.013709
TWS	85.738577	1.0	84.702656	3.749702e-20	0.003601
TWA_Abs	82.705752	1.0	81.706474	1.697941e-19	0.003474
ROT	9.622841	1.0	9.506575	2.049748e-03	0.000405
boat_weight	4.283971	1.0	4.232211	3.967479e-02	0.000181
Heel_Abs	0.787381	1.0	0.777868	3.778031e-01	0.000033
Residual	23722.624582	23436.0	Nan	Nan	0.500000

Polynomial fit:

Model performance:
 $R^2: 0.071$
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 0.580 * side_line2
 0.390 * Line_C
 -0.133 * Trim
 0.068 * TWA_Abs
 0.068 * TWS
 0.021 * ROT
 -0.015 * boat_weight
 0.006 * Heel_Abs

	feature	coefficient
7	side_line2	0.580206
1	Line_C	0.390425
3	Trim	-0.132962
4	TWA_Abs	0.068202
5	TWS	0.067637
2	ROT	0.020735
6	boat_weight	-0.014864
0	Heel_Abs	0.006240

III.3.3. Upwind: Master vs Slave t-test

```
In [15]: t_test(df_numeric_master_upwind,df_numeric_slave_upwind)

T-statistic: 16.509, p-value: 0.000000000000000
The difference is statistically significant, keeping data split.
```

III Downwind

III.1. All downwind data

```
In [16]: downwind_data = df[df['TWA'] < 0]
df_numeric_downwind = downwind_data.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_downwind.drop(columns=[c for c in drop_cols if c in df_numeric_downwind.columns], inplace=True)
df_numeric_downwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_downwind)}")

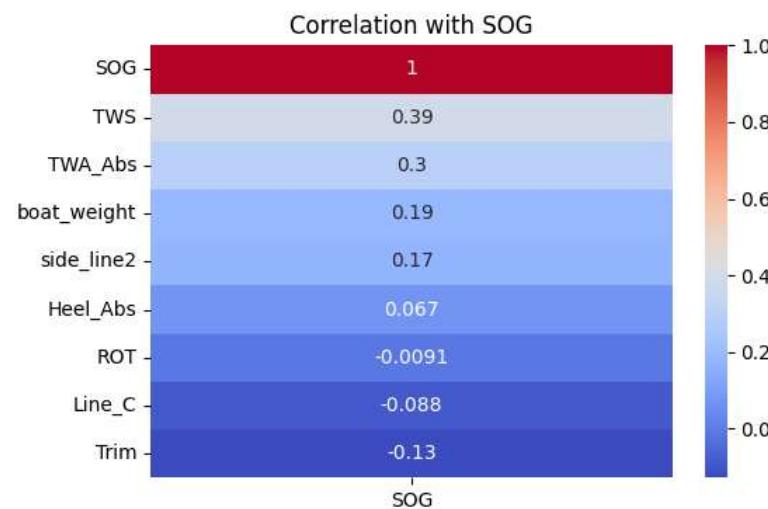
Number of rows after filtering: 31344
```

```
In [17]: full_analysis(df_numeric_downwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']
Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Data cleaning:
- Initial samples: 31344
- Removed samples with NaNs: 1502
- Final samples: 29842
```

Correlation with SOG:



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
TWS	3896.182595	1.0	3524.962025	0.000000e+00	0.105671
boat_weight	1434.485818	1.0	1297.810847	3.132441e-278	0.041689
TWA_Abs	1427.393806	1.0	1291.394548	6.795910e-277	0.041491
side_line2	774.975803	1.0	701.137642	9.883441e-153	0.022962
Heel_Abs	709.498665	1.0	641.899036	3.925800e-140	0.021063
Trim	656.232447	1.0	593.707918	7.290146e-130	0.019513
Line_C	593.135497	1.0	536.622720	1.121074e-117	0.017670
ROT	54.624242	1.0	49.419752	2.110959e-12	0.001654
Residual	32974.770944	29833.0	NaN	NaN	0.500000

Polynomial fit:

Model performance:
 $R^2: 0.296$
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 $0.448 * \text{side_line2}$
 $0.429 * \text{TWS}$
 $0.396 * \text{Line_C}$
 $0.267 * \text{TWA_Abs}$
 $0.245 * \text{boat_weight}$
 $0.203 * \text{Heel_Abs}$
 $-0.178 * \text{Trim}$
 $-0.044 * \text{ROT}$

	feature	coefficient
7	side_line2	0.447631
5	TWS	0.428931
1	Line_C	0.396126
4	TWA_Abs	0.266872
6	boat_weight	0.244961
0	Heel_Abs	0.202721
3	Trim	-0.178111
2	ROT	-0.044311

III.2. Downwind: Gian vs Karl

III.2.1. Downwind: Gian

```
In [18]: gian_data_downwind = downwind_data[
    (downwind_data['boat_name'] == "Gian Stragiotti") |
    ((downwind_data['boat_name'] == "SenseBoard") & (downwind_data['opponent_name'] == "Karl Maeder"))
]
df_numeric_gian_downwind = gian_data_downwind.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_gian_downwind.drop(columns=[c for c in drop_cols if c in df_numeric_gian_downwind.columns], inplace=True)
df_numeric_gian_downwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_gian_downwind)}")
```

Number of rows after filtering: 15667

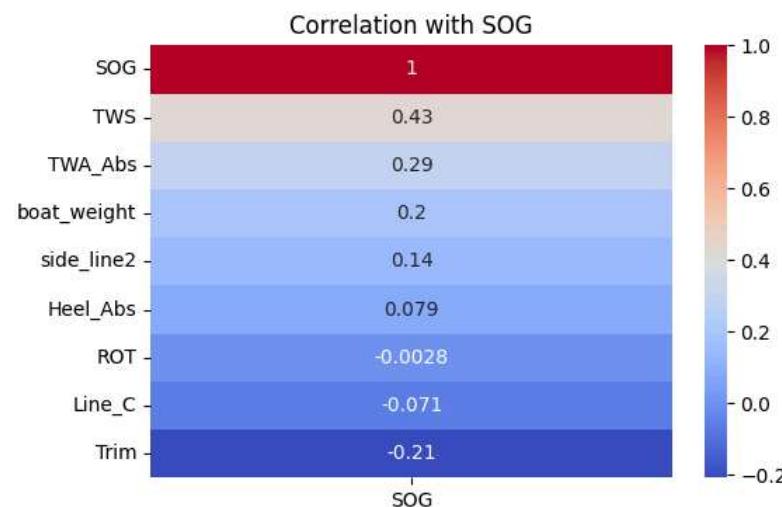
```
In [19]: full_analysis(df_numeric_gian_downwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Data cleaning:
- Initial samples: 15667
- Removed samples with NaNs: 127
- Final samples: 15540

Correlation with SOG:
```



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
TWS	4311.012925	1.0	4066.402641	0.000000e+00	0.207497
side_line2	1185.048459	1.0	1117.807872	1.001755e-236	0.067140
TWA_Abs	1110.379257	1.0	1047.375460	2.049068e-222	0.063177
Trim	568.227512	1.0	535.985834	1.316268e-116	0.033359
Heel_Abs	523.150912	1.0	493.466916	1.183968e-107	0.030795
Line_C	459.844187	1.0	433.752264	4.914218e-95	0.027169
boat_weight	116.883594	1.0	110.251526	1.052409e-25	0.007049
ROT	4.190973	1.0	3.953174	4.680057e-02	0.000254
Residual	16465.251391	15531.0	Nan	Nan	0.500000

Polynomial fit:

Model performance:
 $R^2: 0.380$
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 0.881 * side_line2
 0.651 * TWS
 0.538 * Line_C
 0.331 * TWA_Abs
 0.240 * Heel_Abs
 -0.228 * Trim
 -0.151 * boat_weight
 -0.017 * ROT

	feature	coefficient
7	side_line2	0.880637
5	TWS	0.651012
1	Line_C	0.537877
4	TWA_Abs	0.330501
0	Heel_Abs	0.240474
3	Trim	-0.228298
6	boat_weight	-0.151046
2	ROT	-0.016990

III.2.2. Downwind: Karl

```
In [20]: karl_data_downwind = downwind_data[
    (downwind_data['boat_name'] == "Karl Maeder") |
    ((downwind_data['boat_name'] == "SenseBoard") & (downwind_data['opponent_name'] == "Gian Stragiotti"))
]
df_numeric_karl_downwind = karl_data_downwind.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_karl_downwind.drop(columns=[c for c in drop_cols if c in df_numeric_karl_downwind.columns], inplace=True)
df_numeric_karl_downwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_karl_downwind)}")
```

Number of rows after filtering: 15677

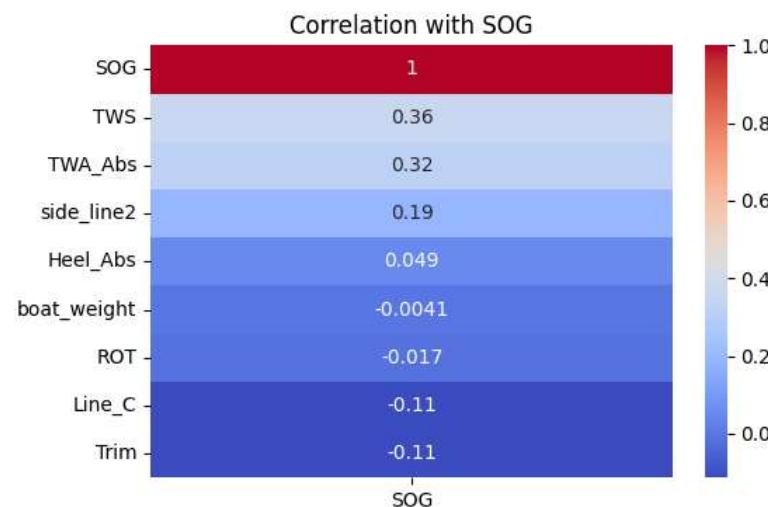
```
In [21]: full_analysis(df_numeric_karl_downwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Data cleaning:
- Initial samples: 15677
- Removed samples with NaNs: 1375
- Final samples: 14302
```

Correlation with SOG:



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
TWA_Abs	743.176205	1.0	702.927062	3.049950e-151	0.046875
TWS	718.142331	1.0	679.248980	2.488059e-146	0.045367
boat_weight	269.051547	1.0	254.480179	8.467459e-57	0.017493
Heel_Abs	182.538293	1.0	172.652333	3.289464e-39	0.011935
Trim	156.310689	1.0	147.845172	7.535432e-34	0.010238
side_line2	132.993940	1.0	125.791218	4.517590e-29	0.008724
Line_C	86.710621	1.0	82.014523	1.523214e-19	0.005705
ROT	56.732255	1.0	53.659733	2.511485e-13	0.003740
Residual	15111.407819	14293.0	Nan	Nan	0.500000

Polynomial fit:

Model performance:
 R²: 0.216
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 0.304 * TWS
 0.286 * TWA_Abs
 0.249 * side_line2
 0.209 * Line_C
 0.150 * Heel_Abs
 0.150 * boat_weight
 -0.127 * Trim
 -0.065 * ROT

	feature	coefficient
5	TWS	0.303895
4	TWA_Abs	0.286355
7	side_line2	0.249054
1	Line_C	0.209204
0	Heel_Abs	0.150211
6	boat_weight	0.150168
3	Trim	-0.127286
2	ROT	-0.065463

III.2.3. Downwind: Karl vs Gian t-test

```
In [22]: t_test(df_numeric_karl_downwind, df_numeric_gian_downwind)

T-statistic: -38.976, p-value: 0.000000000000000
The difference is statistically significant, keeping data split.
```

III.3. Downwind: Master vs Slave

III.3.1 Downwind Master

```
In [23]: master_data_downwind = downwind_data[downwind_data['boat_role'] == "master"]
df_numeric_master_downwind = master_data_downwind.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_master_downwind.drop(columns=[c for c in drop_cols if c in df_numeric_master_downwind.columns], inplace=True)
df_numeric_master_downwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_master_downwind)}")
```

Number of rows after filtering: 15670

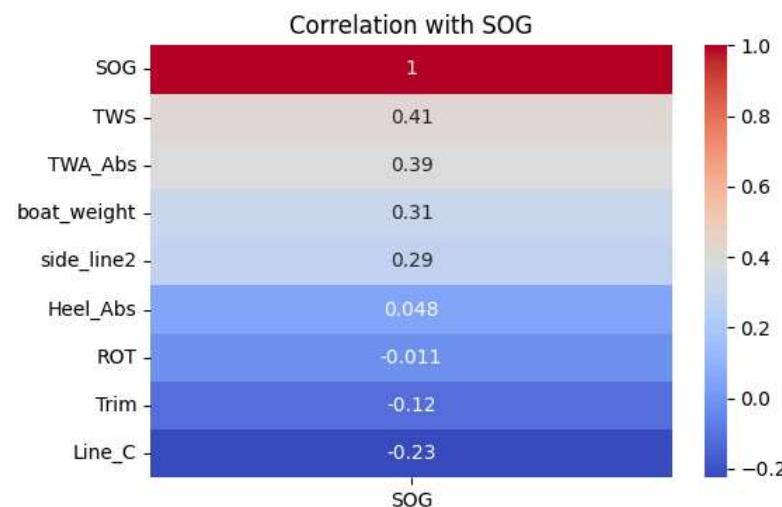
```
In [24]: full_analysis(df_numeric_master_downwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Data cleaning:
- Initial samples: 15670
- Removed samples with NaNs: 90
- Final samples: 15580

Correlation with SOG:
```



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
TWA_Abs	1786.999377	1.0	1682.555326	0.000000e+00	0.097519
TWS	1718.820938	1.0	1618.361686	0.000000e+00	0.094149
boat_weight	1562.558166	1.0	1471.231943	1.170051e-307	0.086329
Trim	634.462512	1.0	597.380330	1.691381e-129	0.036947
Heel_Abs	623.315846	1.0	586.885149	2.675749e-127	0.036322
side_line2	153.359460	1.0	144.396120	4.078270e-33	0.009188
ROT	64.685129	1.0	60.904503	6.370158e-15	0.003896
Line_C	20.091797	1.0	18.917500	1.373625e-05	0.001213
Residual	16537.564543	15571.0	Nan	Nan	0.500000

Polynomial fit:

Model performance:
 R²: 0.398
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 0.401 * TWA_Abs
 0.387 * TWS
 0.359 * boat_weight
 0.295 * side_line2
 0.260 * Heel_Abs
 -0.244 * Trim
 0.106 * Line_C
 -0.067 * ROT

	feature	coefficient
4	TWA_Abs	0.400926
5	TWS	0.387390
6	boat_weight	0.359274
7	side_line2	0.294504
0	Heel_Abs	0.260317
3	Trim	-0.243639
1	Line_C	0.105793
2	ROT	-0.066867

III.3.2 Downwind Slave

```
In [25]: slave_data_downwind = downwind_data[downwind_data['boat_role'] == "slave"]
df_numeric_slave_downwind = slave_data_downwind.select_dtypes(include=["float64", "int64"]).copy()
df_numeric_slave_downwind.drop(columns=[c for c in drop_cols if c in df_numeric_slave_downwind.columns], inplace=True)
df_numeric_slave_downwind.dropna(subset=["SOG"], inplace=True)
print(f"Number of rows after filtering: {len(df_numeric_slave_downwind)}")
```

Number of rows after filtering: 15674

```
In [26]: full_analysis(df_numeric_slave_downwind, target_variable="SOG", max_null_ratio=MAX_NULL_RATIO)

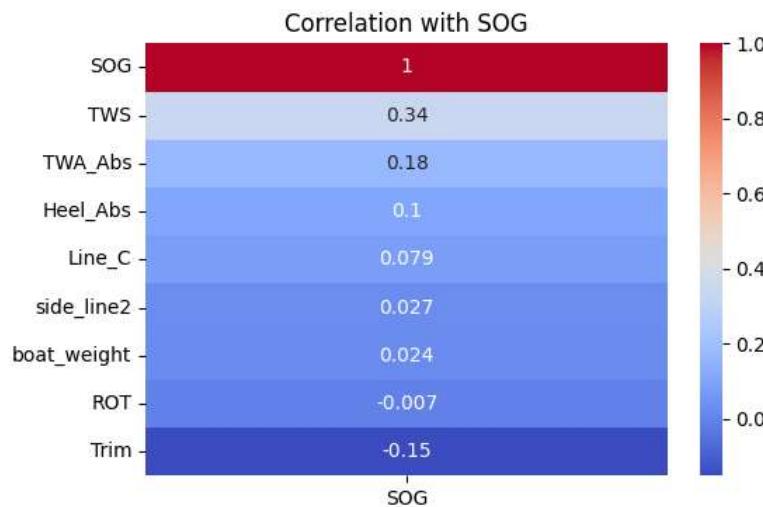
Initial features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']

Selected features (8): ['Heel_Abs', 'Line_C', 'ROT', 'Trim', 'TWA_Abs', 'TWS', 'boat_weight', 'side_line2']
```

Data cleaning:

- Initial samples: 15674
- Removed samples with NaNs: 1412
- Final samples: 14262

Correlation with SOG:



ANOVA:

	sum_sq	df	F	PR(>F)	partial_eta_sq
TWS	1918.077106	1.0	1869.326544	0.000000e+00	0.115946
Line_C	567.728972	1.0	553.299361	4.612276e-120	0.037369
side_line2	433.683402	1.0	422.660743	1.409782e-92	0.028800
Trim	195.156282	1.0	190.196118	5.439139e-43	0.013169
Heel_Abs	183.229397	1.0	178.572370	1.740739e-40	0.012374
boat_weight	143.792360	1.0	140.137680	3.514383e-32	0.009736
TWA_Abs	88.559083	1.0	86.308233	1.758919e-20	0.006019
ROT	3.786772	1.0	3.690526	5.474229e-02	0.000259
Residual	14624.707003	14253.0	Nan	Nan	0.500000

Polynomial fit:

Model performance:
 $R^2: 0.231$
 Used features: 8
 Polynomial terms generated: 8

Top terms in formula:
 $0.537 * \text{Line_C}$
 $0.451 * \text{TWS}$
 $0.450 * \text{side_line2}$
 $0.151 * \text{Heel_Abs}$
 $-0.140 * \text{Trim}$
 $0.111 * \text{boat_weight}$
 $0.101 * \text{TWA_Abs}$
 $-0.017 * \text{ROT}$

	feature	coefficient
1	Line_C	0.537020
5	TWS	0.451362
7	side_line2	0.449649
0	Heel_Abs	0.151213
3	Trim	-0.140370
6	boat_weight	0.111411
4	TWA_Abs	0.101155
2	ROT	-0.016852

III.3.3. Downwind: Master vs Slave t-test

```
In [27]: t_test(df_numeric_master_downwind,df_numeric_slave_downwind)
```

T-statistic: 0.953, p-value: 0.340575073568009
The difference is not statistically significant, keeping data combined.