```python
import pandas as pd
import numpy as np
from lifelines import CoxPHFitter
from sklearn.preprocessing import StandardScaler
import matplotlib.pyplot as plt
from lifelines.utils import concordance_index

# Load the data from the .xlsx file
data = pd.read_excel('data1.xlsx')

# Define categorical variables
categorical_cols = ['SEX', 'CompositeStage', 'LNInvolment', 'Comorbidity', 'FamiliyHistoryOfCancer']
data[categorical_cols] = data[categorical_cols].astype('category')

# One-hot encode categorical variables
data_encoded = pd.get_dummies(data, columns=categorical_cols, drop_first=True)

# Standardize the covariates
scaler = StandardScaler()
data_encoded[['DEATH', 'AGE']] = scaler.fit_transform(data_encoded[['DEATH', 'AGE']])
buckley_james_data = data_encoded[['Months', 'DEATH', 'AGE'] + [col for col in data_encoded.columns if col.startswith('SEX_')
cph = CoxPHFitter(penalizer=0.1)
cph.fit(buckley_james_data, 'Months', 'DEATH', show_progress=True)
print(cph.summary)
```

```
Iteration 1: norm_delta = 0.68300, step_size = 0.9500, log_lik = -1663.17959, newton_decrement = 54.49713, seconds_since_start
= 0.0
Iteration 2: norm_delta = 0.08068, step_size = 0.9500, log_lik = -1616.75385, newton_decrement = 1.16193, seconds_since_start =
0.0
Iteration 3: norm_delta = 0.00455, step_size = 0.9500, log_lik = -1615.59252, newton_decrement = 0.00340, seconds_since_start =
0.0
Iteration 4: norm_delta = 0.00000, step_size = 1.0000, log_lik = -1615.58912, newton_decrement = 0.00000, seconds_since_start =
0.0
Convergence success after 4 iterations.
```

	coef	exp(coef)	se(coef)	coef lower 95%
covariate				
AGE	0.026478	1.026831	0.056826	-0.084899
SEX_2	-0.027535	0.972841	0.107577	-0.238382
CompositeStage_2	-0.151530	0.859392	0.173728	-0.492030
CompositeStage_3	0.026319	1.026669	0.188558	-0.343248
CompositeStage_4	1.135923	3.114047	0.180333	0.782478
LNInvolment_1	-0.343457	0.709314	0.138901	-0.615698
Comorbidity_1	-0.053072	0.948312	0.111065	-0.270755
FamiliyHistoryOfCancer_1	-0.055868	0.945664	0.157606	-0.364770

	coef upper 95%	exp(coef) lower 95%
covariate		
AGE	0.137854	0.918605
SEX_2	0.183313	0.787902
CompositeStage_2	0.188969	0.611384
CompositeStage_3	0.395887	0.709462
CompositeStage_4	1.489369	2.186884
LNInvolment_1	-0.071217	0.540264
Comorbidity_1	0.164612	0.762803
FamiliyHistoryOfCancer_1	0.253035	0.694356

	exp(coef) upper 95%	cmp to	z	p
covariate				
AGE	1.147808	0.0	0.465942	6.412567e-01
SEX_2	1.201190	0.0	-0.255952	7.979879e-01
CompositeStage_2	1.208004	0.0	-0.872229	3.830834e-01
CompositeStage_3	1.485701	0.0	0.139582	8.889905e-01
CompositeStage_4	4.434295	0.0	6.299046	2.994829e-10
LNInvolment_1	0.931260	0.0	-2.472683	1.341029e-02
Comorbidity_1	1.178936	0.0	-0.477843	6.327621e-01

```
FamiliyHistoryOfCancer_1               1.287929      0.0 -0.354475  7.229829e-01

                           -log2(p)
covariate
AGE                        0.641026
SEX_2                      0.325561
CompositeStage_2           1.384270
CompositeStage_3           0.169760
CompositeStage_4          31.636807
LNInvolment_1              6.220515
Comorbidity_1              0.660265
FamiliyHistoryOfCancer_1   0.467966
```

In [240…
```python
concordance_values = {}
for column in cph.summary.index:
    if column != 'ID':
        concordance_values[column] = concordance_index(buckley_james_data[column], cph.predict_partial_hazard(buckley_james_da
print("Concordance values of Univariate Variables:\n")
print(concordance_values)
```

Concordance values of Univariate Variables:

{'AGE': 0.5306817586333386, 'SEX_2': 0.45524296675191817, 'CompositeStage_2': 0.2839379059084429, 'CompositeStage_3': 0.1588888
8888888889, 'CompositeStage_4': 1.0, 'LNInvolment_1': 0.1500506072874494, 'Comorbidity_1': 0.3904086109968463, 'FamiliyHistoryO
fCancer_1': 0.45142900577683187}

In [228…
```python
univariate_results = []
for col in data.columns:
    if col not in ['Months', 'ID','DEATH']:
        cph_univariate = CoxPHFitter(penalizer=0.1)
        cph_univariate.fit(data[[col, 'Months', 'DEATH']], 'Months', 'DEATH', show_progress=True)
        univariate_results.append((col, cph_univariate.summary))
        n = len(data)
        llf = cph_univariate.log_likelihood_
        k = cph_univariate.params_.shape[0]
        aic = -2 * llf + 2 * k
        bic = -2 * llf + k * np.log(n)
        univariate_aic_bic.append((col, aic, bic))
        print(f"\nAIC value of {col}:", aic)
        print(f"BIC value of {col}:", bic)
```

```python
# Print the summaries of the univariate analysis
for col, summary in univariate_results:
    print(f"Univariate analysis of: {col}")
    print(summary)
    print("\n")
```

```
Iteration 1: norm_delta = 0.09094, step_size = 0.9500, log_lik = -943.70062, newton_decrement = 0.84604, seconds_since_start =
0.0
Iteration 2: norm_delta = 0.00646, step_size = 0.9500, log_lik = -942.84679, newton_decrement = 0.00411, seconds_since_start =
0.0
Iteration 3: norm_delta = 0.00033, step_size = 0.9500, log_lik = -942.84269, newton_decrement = 0.00001, seconds_since_start =
0.1
Iteration 4: norm_delta = 0.00000, step_size = 1.0000, log_lik = -942.84268, newton_decrement = 0.00000, seconds_since_start =
0.1
Convergence success after 4 iterations.

AIC value of AGE: 1887.6853655755
BIC value of AGE: 1891.523096022666
Iteration 1: norm_delta = 0.02152, step_size = 0.9500, log_lik = -943.70062, newton_decrement = 0.04824, seconds_since_start =
0.0
Iteration 2: norm_delta = 0.00117, step_size = 0.9500, log_lik = -943.65238, newton_decrement = 0.00014, seconds_since_start =
0.0
Iteration 3: norm_delta = 0.00006, step_size = 0.9500, log_lik = -943.65224, newton_decrement = 0.00000, seconds_since_start =
0.0
Iteration 4: norm_delta = 0.00000, step_size = 1.0000, log_lik = -943.65224, newton_decrement = 0.00000, seconds_since_start =
0.0
Convergence success after 4 iterations.

AIC value of SEX: 1889.3044715567958
BIC value of SEX: 1893.1422020039618
Iteration 1: norm_delta = 0.60207, step_size = 0.9500, log_lik = -943.70062, newton_decrement = 33.86694, seconds_since_start =
0.0
Iteration 2: norm_delta = 0.06882, step_size = 0.9500, log_lik = -909.00329, newton_decrement = 0.38483, seconds_since_start =
0.0
Iteration 3: norm_delta = 0.00442, step_size = 0.9500, log_lik = -908.61639, newton_decrement = 0.00154, seconds_since_start =
0.0
Iteration 4: norm_delta = 0.00000, step_size = 1.0000, log_lik = -908.61485, newton_decrement = 0.00000, seconds_since_start =
0.0
Convergence success after 4 iterations.

AIC value of CompositeStage: 1819.2296914692981
BIC value of CompositeStage: 1823.0674219164641
Iteration 1: norm_delta = 0.20088, step_size = 0.9500, log_lik = -943.70062, newton_decrement = 4.61862, seconds_since_start =
0.0
Iteration 2: norm_delta = 0.02598, step_size = 0.9500, log_lik = -938.91722, newton_decrement = 0.06679, seconds_since_start =
0.0
```

```
Iteration 3: norm_delta = 0.00161, step_size = 0.9500, log_lik = -938.85015, newton_decrement = 0.00025, seconds_since_start = 0.0
Iteration 4: norm_delta = 0.00000, step_size = 1.0000, log_lik = -938.84989, newton_decrement = 0.00000, seconds_since_start = 0.0
Convergence success after 4 iterations.

AIC value of LNInvolment: 1879.6997872387765
BIC value of LNInvolment: 1883.5375176859425
Iteration 1: norm_delta = 0.07263, step_size = 0.9500, log_lik = -943.70062, newton_decrement = 0.54710, seconds_since_start = 0.0
Iteration 2: norm_delta = 0.00311, step_size = 0.9500, log_lik = -943.15727, newton_decrement = 0.00102, seconds_since_start = 0.0
Iteration 3: norm_delta = 0.00016, step_size = 0.9500, log_lik = -943.15625, newton_decrement = 0.00000, seconds_since_start = 0.0
Iteration 4: norm_delta = 0.00000, step_size = 1.0000, log_lik = -943.15625, newton_decrement = 0.00000, seconds_since_start = 0.0
Convergence success after 4 iterations.

AIC value of Comorbidity: 1888.3125005206216
BIC value of Comorbidity: 1892.1502309677876
Iteration 1: norm_delta = 0.02707, step_size = 0.9500, log_lik = -943.70062, newton_decrement = 0.07765, seconds_since_start = 0.0
Iteration 2: norm_delta = 0.00205, step_size = 0.9500, log_lik = -943.62205, newton_decrement = 0.00043, seconds_since_start = 0.0
Iteration 3: norm_delta = 0.00011, step_size = 0.9500, log_lik = -943.62163, newton_decrement = 0.00000, seconds_since_start = 0.0
Iteration 4: norm_delta = 0.00000, step_size = 1.0000, log_lik = -943.62163, newton_decrement = 0.00000, seconds_since_start = 0.0
Convergence success after 4 iterations.

AIC value of FamiliyHistoryOfCancer: 1889.243255971226
BIC value of FamiliyHistoryOfCancer: 1893.080986418392
Univariate analysis of: AGE
              coef  exp(coef)  se(coef)  coef lower 95%  coef upper 95%  \
covariate
AGE       0.007719   1.007749  0.005933        -0.00391        0.019349


          exp(coef) lower 95%  exp(coef) upper 95%  cmp to        z  \
covariate
AGE                  0.996098             1.019537     0.0  1.30097
```

```
                  p  -log2(p)
covariate
AGE        0.193269   2.37132


Univariate analysis of: SEX
             coef  exp(coef)  se(coef)  coef lower 95%  coef upper 95%  \
covariate
SEX        0.044123    1.04511  0.142053       -0.234297        0.322542

          exp(coef) lower 95%  exp(coef) upper 95%  cmp to         z       p  \
covariate
SEX                  0.791127             1.380633     0.0  0.310606  0.7561

          -log2(p)
covariate
SEX        0.40335


Univariate analysis of: CompositeStage
                    coef  exp(coef)  se(coef)  coef lower 95%  coef upper 95%  \
covariate
CompositeStage  0.707114    2.028129  0.087683        0.535259        0.878968

                exp(coef) lower 95%  exp(coef) upper 95%  cmp to         z  \
covariate
CompositeStage              1.70789             2.408414     0.0  8.064462

                           p    -log2(p)
covariate
CompositeStage  7.355908e-16   50.271946


Univariate analysis of: LNInvolment
                coef  exp(coef)  se(coef)  coef lower 95%  coef upper 95%  \
covariate
LNInvolment -0.48295    0.616961  0.160078       -0.796698       -0.169202

              exp(coef) lower 95%  exp(coef) upper 95%  cmp to         z  \
```

```
covariate
LNInvolment                0.450815                  0.844338      0.0 -3.01696

                     p  -log2(p)
covariate
LNInvolment  0.002553   8.61346


Univariate analysis of: Comorbidity
                 coef  exp(coef)  se(coef)  coef lower 95%  coef upper 95%  \
covariate
Comorbidity -0.144597   0.865371  0.138333       -0.415726        0.126531

             exp(coef) lower 95%  exp(coef) upper 95%  cmp to         z  \
covariate
Comorbidity             0.659861             1.134884      0.0 -1.045284

                    p  -log2(p)
covariate
Comorbidity  0.295892  1.756859


Univariate analysis of: FamiliyHistoryOfCancer
                            coef  exp(coef)  se(coef)  coef lower 95%  \
covariate
FamiliyHistoryOfCancer -0.082937   0.920409  0.210487       -0.495484

                        coef upper 95%  exp(coef) lower 95%  \
covariate
FamiliyHistoryOfCancer         0.32961             0.609276

                        exp(coef) upper 95%  cmp to         z         p  \
covariate
FamiliyHistoryOfCancer             1.390425      0.0 -0.394026  0.693562

                        -log2(p)
covariate
FamiliyHistoryOfCancer  0.527903
```

```python
# Print AIC and BIC for univariate models
print("\nAIC and BIC for univariate models:")
for col, aic, bic in univariate_aic_bic:
    print(f"{col}: AIC={aic}, BIC={bic}")
```

```
AIC and BIC for univariate models:
SEX_2: AIC=3330.115165027162, BIC=3337.790625921494
CompositeStage_2: AIC=3321.5101024525675, BIC=3329.1855633468995
CompositeStage_3: AIC=3317.0745678980034, BIC=3324.7500287923353
CompositeStage_4: AIC=3242.6468994706224, BIC=3250.3223603649544
LNInvolment_1: AIC=3322.1186323122874, BIC=3329.7940932066194
Comorbidity_1: AIC=3328.7696858848362, BIC=3336.445146779168
FamiliyHistoryOfCancer_1: AIC=3329.9669227876907, BIC=3337.6423836820227
AGE: AIC=1887.6853655755, BIC=1891.523096022666
SEX: AIC=1889.3044715567958, BIC=1893.1422020039618
CompositeStage: AIC=1819.2296914692981, BIC=1823.0674219164641
LNInvolment: AIC=1879.6997872387765, BIC=1883.5375176859425
Comorbidity: AIC=1888.3125005206216, BIC=1892.1502309677876
FamiliyHistoryOfCancer: AIC=1889.243255971226, BIC=1893.080986418392
AGE: AIC=1887.6853655755, BIC=1891.523096022666
SEX: AIC=1889.3044715567958, BIC=1893.1422020039618
CompositeStage: AIC=1819.2296914692981, BIC=1823.0674219164641
LNInvolment: AIC=1879.6997872387765, BIC=1883.5375176859425
Comorbidity: AIC=1888.3125005206216, BIC=1892.1502309677876
FamiliyHistoryOfCancer: AIC=1889.243255971226, BIC=1893.080986418392
AGE: AIC=1887.6853655755, BIC=1891.523096022666
SEX: AIC=1889.3044715567958, BIC=1893.1422020039618
CompositeStage: AIC=1819.2296914692981, BIC=1823.0674219164641
LNInvolment: AIC=1879.6997872387765, BIC=1883.5375176859425
Comorbidity: AIC=1888.3125005206216, BIC=1892.1502309677876
FamiliyHistoryOfCancer: AIC=1889.243255971226, BIC=1893.080986418392
```

```python
significant_variables_multivariate = [(var, summary) for var, summary in multivariate_results if summary['p'][var] < 0.05]
print("\nSignificant variables from univariate analysis:")
for var, summary in significant_variables_multivariate:
    print(f"\n{var}:")
    print(summary)
```

Significant variables from univariate analysis:

CompositeStage:

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	\
covariate						
CompositeStage	0.451465	1.570611	0.061942	0.330061	0.572868	
AGE	0.013370	1.013460	0.053478	-0.091446	0.118185	

	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	\
covariate					
CompositeStage	1.391053	1.773346	0.0	7.288543	
AGE	0.912611	1.125453	0.0	0.250006	

	p	-log2(p)
covariate		
CompositeStage	3.133250e-13	41.537405
AGE	8.025824e-01	0.317279

LNInvolment:

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	\
covariate						
LNInvolment	-0.143911	0.865965	0.051475	-0.244800	-0.043022	
AGE	-0.027609	0.972769	0.052098	-0.129719	0.074502	

	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	\
covariate					
LNInvolment	0.782861	0.957891	0.0	-2.795740	
AGE	0.878342	1.077347	0.0	-0.529935	

	p	-log2(p)
covariate		
LNInvolment	0.005178	7.593362
AGE	0.596157	0.746236

In [231...
```python
#significant_variables = [(var, p_value) for var, p_value in multivariate_results if p_value < 0.05]
print("Updated data with significant variables as categorical data:")
print(data)

#data_encoded = pd.get_dummies(data, columns=[var for var, _ in significant_variables], drop_first=True)

# Update the Buckley-James data with the new categorical variables
```

```
categorical_columns = ['SEX_', 'CompositeStage_', 'LNInvolment_', 'Comorbidity_', 'FamiliyHistoryOfCancer_']
buckley_james_data = data_encoded[['Months', 'DEATH', 'AGE'] + [col for col in data_encoded.columns if any(col.startswith(cat_
```

Updated data with significant variables as categorical data:

```
      ID  Months  DEATH  AGE SEX CompositeStage LNInvolment Comorbidity  \
0      1      70      0   50   1              3           1           1
1      2      68      0   50   2              1           0           1
2      3      69      0   52   1              2           0           1
3      4      43      1   55   2              2           0           0
4      5      71      0   69   2              3           1           1
..   ...     ...    ...  ...  ..            ...         ...         ...
338  339      65      0   41   1              3           1           1
339  340      61      0   52   1              2           0           1
340  341      65      0   61   2              2           0           1
341  342      16      1   71   2              4           0           0
342  343      31      1   60   2              4           1           0

     FamiliyHistoryOfCancer
0                         0
1                         0
2                         0
3                         0
4                         0
..                      ...
338                       0
339                       0
340                       0
341                       0
342                       0

[343 rows x 9 columns]
```

In [232...
```
cph_multivariate = CoxPHFitter(penalizer=0.1)
cph_multivariate.fit(buckley_james_data[['Months', 'DEATH', 'AGE'] + [var for var, _ in significant_variables]], 'Months', 'DE
print(cph_multivariate.summary)
```

```
Iteration 1: norm_delta = 0.68301, step_size = 0.9500, log_lik = -1663.17959, newton_decrement = 54.30464, seconds_since_start = 0.0
Iteration 2: norm_delta = 0.08057, step_size = 0.9500, log_lik = -1616.92886, newton_decrement = 1.14381, seconds_since_start = 0.1
Iteration 3: norm_delta = 0.00450, step_size = 0.9500, log_lik = -1615.78591, newton_decrement = 0.00329, seconds_since_start = 0.1
Iteration 4: norm_delta = 0.00000, step_size = 1.0000, log_lik = -1615.78262, newton_decrement = 0.00000, seconds_since_start = 0.1
Convergence success after 4 iterations.
                     coef   exp(coef)   se(coef)   coef lower 95%  \
covariate
AGE              0.018283    1.018451   0.053798        -0.087159
CompositeStage_2 -0.149303    0.861308   0.173490        -0.489338
CompositeStage_3  0.032246    1.032772   0.187625        -0.335493
CompositeStage_4  1.137077    3.117641   0.179401         0.785458
LNInvolment_1    -0.345846    0.707622   0.138427        -0.617158

                  coef upper 95%   exp(coef) lower 95%   exp(coef) upper 95%  \
covariate
AGE                    0.123725             0.916532              1.131705
CompositeStage_2       0.190732             0.613032              1.210135
CompositeStage_3       0.399985             0.714986              1.491802
CompositeStage_4       1.488695             2.193411              4.431311
LNInvolment_1         -0.074533             0.539475              0.928177

                  cmp to          z              p   -log2(p)
covariate
AGE                  0.0   0.339851   7.339686e-01   0.446210
CompositeStage_2     0.0  -0.860585   3.894667e-01   1.360428
CompositeStage_3     0.0   0.171864   8.635445e-01   0.211658
CompositeStage_4     0.0   6.338196   2.324704e-10  32.002234
LNInvolment_1        0.0  -2.498390   1.247587e-02   6.324716
```

```python
concordance_dict = {}

for var in cph_multivariate.params_.index:
    concordance = cph_multivariate.concordance_index_
    concordance_dict[var] = concordance
print("Concordance values of Multivariate Variables:\n")
print(concordance_dict)
```

Concordance values of Multivariate Variables:

{'AGE': 0.6582635491564964, 'CompositeStage_2': 0.6582635491564964, 'CompositeStage_3': 0.6582635491564964, 'CompositeStage_4': 0.6582635491564964, 'LNInvolment_1': 0.6582635491564964}

In [233...
```python
n = len(buckley_james_data)
llf = cph_multivariate.log_likelihood_
k = cph_multivariate.params_.shape[0]
multivariate_aic = -2 * llf + 2 * k
multivariate_bic = -2 * llf + k * np.log(n)
print(cph_multivariate.summary)
```

covariate	coef	exp(coef)	se(coef)	coef lower 95% \
AGE	0.018283	1.018451	0.053798	-0.087159
CompositeStage_2	-0.149303	0.861308	0.173490	-0.489338
CompositeStage_3	0.032246	1.032772	0.187625	-0.335493
CompositeStage_4	1.137077	3.117641	0.179401	0.785458
LNInvolment_1	-0.345846	0.707622	0.138427	-0.617158

covariate	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95% \
AGE	0.123725	0.916532	1.131705
CompositeStage_2	0.190732	0.613032	1.210135
CompositeStage_3	0.399985	0.714986	1.491802
CompositeStage_4	1.488695	2.193411	4.431311
LNInvolment_1	-0.074533	0.539475	0.928177

covariate	cmp to	z	p	-log2(p)
AGE	0.0	0.339851	7.339686e-01	0.446210
CompositeStage_2	0.0	-0.860585	3.894667e-01	1.360428
CompositeStage_3	0.0	0.171864	8.635445e-01	0.211658
CompositeStage_4	0.0	6.338196	2.324704e-10	32.002234
LNInvolment_1	0.0	-2.498390	1.247587e-02	6.324716

In [235...
```python
# Print AIC and BIC for multivariate model
print("\nAIC value of the multivariate model:", multivariate_aic)
print("BIC value of the multivariate model:", multivariate_bic)
```

```
AIC value of the multivariate model: 3241.5652333399553
BIC value of the multivariate model: 3260.753885575785
```