Appendix V: Forward Selection

In [7]:
```
"""Imports necessary packages"""

import itertools
import math
from typing import Dict, Iterable, List, Union

import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import pylab
import scipy
import scipy.stats as stats
import seaborn as sns
import statsmodels.api as sm
from sklearn.model_selection import train_test_split

sns.set_style("whitegrid")
```

In [8]:
```
def forward_p_vals(input_model_str: str, vars: List[str], data: Iterable) -> Dict[st
    """Creates a new model for every variable in `vars` such that it contains all va

    Args:
        input_model_str (str): a model string as required by statsmodels.api.formula
        vars (List[str]): the list of explanatory variables that can be considered i
        data (Iterable): the two dimensional data for fitting the new models.

    Returns:
        Dict[str, float]: the dictionary of new model strings (key) and the associat
    """
    result = {}
    for var in vars:
        if input_model_str.endswith("~"):
            model_str = "%s %s" % (input_model_str, var)
        else:
            model_str = "%s + %s" % (input_model_str, var)
        model = sm.formula.ols(formula=model_str, data=data)
        model_fitted = model.fit()
        p_vals = model_fitted.pvalues.to_dict()
        var = var.replace("*", ":")
        for k in p_vals.keys():
            if var+"[" in k:
                var = k
        result[model_str] = p_vals.get(var, 1)
    return result
```

In [9]:
```
def print_p_vals_from_models(models: Dict[str, float]) -> None:
    """Prints p values from a dictionary.

    Args:
        models (Dict[str, float]): the dictionary of variables newly added to a mode
    """
    for model, value in models.items():
        print("p-value of %s: %.8f" % (model.split(" ")[-1], value))
```

In [10]:
```
def forward_selection(response_var: str, explanatory_vars: List[str], data: Iterable
```

```
    """Performs the forward selection.

    Args:
        response_var (str): the variable to predict.
        explanatory_vars (List[str]): the list of explanatory vriables that may be u
        data (Iterable): the two-dimensional data to use for model fitting.

    Returns:
        str: the resulting model string in the format required by statsmodels.api.fo
    """
    i = 1
    previous_model = "%s ~" % response_var
    while len(explanatory_vars) > 0:
        print("--- STEP %i ---" % i)
        print("current model: %s" % previous_model)

        models = forward_p_vals(previous_model, explanatory_vars, data)
        print("possible variables:")
        print_p_vals_from_models(models)

        best_next_model = min(models, key=models.get)
        if models[best_next_model] > 0.05:
            print("The minimal p-value is higher than 0.05, returning the previous m
            return previous_model

        previous_model = best_next_model
        explanatory_vars = [var for var in explanatory_vars if var != best_next_mode
        i += 1
```

In [11]:

```
data = pd.read_csv("D:/School/frequentist-statistics/ITM-song-popularity/database/it
data = data.drop("Unnamed: 0", axis=1)
```

In [12]:

```
explanatory_vars = ["name_len", "track_number", "duration", "acousticness", "danceab
```

In [13]:

```
best_fs_abs = forward_selection("popularity_abs", explanatory_vars, data)
print("The best model for absolute popularity excluding correlations obtained via fo
```

```
--- STEP 1 ---
current model: popularity_abs ~
possible variables:
p-value of name_len: 0.46054799
p-value of track_number: 0.21061084
p-value of duration: 0.00087687
p-value of acousticness: 0.00576701
p-value of danceability: 0.00328833
p-value of energy: 0.03409356
p-value of loudness: 0.00488552
p-value of speechiness: 0.22617057
p-value of valence: 0.03859374
p-value of tempo: 0.54998284
p-value of complexity: 0.00001846
p-value of age_days: 0.00000017
p-value of mode: 0.91191181
--- STEP 2 ---
current model: popularity_abs ~ age_days
possible variables:
p-value of name_len: 0.91292203
p-value of track_number: 0.04107383
p-value of duration: 0.00014822
p-value of acousticness: 0.00073659
p-value of danceability: 0.00018752
p-value of energy: 0.00038229
```

```
p-value of loudness: 0.00000553
p-value of speechiness: 0.23901821
p-value of valence: 0.02374235
p-value of tempo: 0.08657363
p-value of complexity: 0.00000317
p-value of mode: 0.75308241
--- STEP 3 ---
current model: popularity_abs ~ age_days + complexity
possible variables:
p-value of name_len: 0.44094297
p-value of track_number: 0.00001266
p-value of duration: 0.53709664
p-value of acousticness: 0.02029146
p-value of danceability: 0.05820237
p-value of energy: 0.00832388
p-value of loudness: 0.01648438
p-value of speechiness: 0.70579816
p-value of valence: 0.47059392
p-value of tempo: 0.23158259
p-value of mode: 0.47487225
--- STEP 4 ---
current model: popularity_abs ~ age_days + complexity + track_number
possible variables:
p-value of name_len: 0.62457759
p-value of duration: 0.76106354
p-value of acousticness: 0.21122935
p-value of danceability: 0.09193785
p-value of energy: 0.07424965
p-value of loudness: 0.13148060
p-value of speechiness: 0.30669822
p-value of valence: 0.79107073
p-value of tempo: 0.49622858
p-value of mode: 0.32506313
The minimal p-value is higher than 0.05, returning the previous model
The best model for absolute popularity excluding correlations obtained via forward s
election is `popularity_abs ~ age_days + complexity + track_number`.
```

In [14]:

```
best_fs_rel = forward_selection("popularity_norm", explanatory_vars, data)
print("The best model for relative popularity excluding correlations obtained via fo
```

```
--- STEP 1 ---
current model: popularity_norm ~
possible variables:
p-value of name_len: 0.46054799
p-value of track_number: 0.21061084
p-value of duration: 0.00087687
p-value of acousticness: 0.00576701
p-value of danceability: 0.00328833
p-value of energy: 0.03409356
p-value of loudness: 0.00488552
p-value of speechiness: 0.22617057
p-value of valence: 0.03859374
p-value of tempo: 0.54998284
p-value of complexity: 0.00001846
p-value of age_days: 0.00000017
p-value of mode: 0.91191181
--- STEP 2 ---
current model: popularity_norm ~ age_days
possible variables:
p-value of name_len: 0.91292203
p-value of track_number: 0.04107383
p-value of duration: 0.00014822
p-value of acousticness: 0.00073659
p-value of danceability: 0.00018752
p-value of energy: 0.00038229
p-value of loudness: 0.00000553
p-value of speechiness: 0.23901821
p-value of valence: 0.02374235
```

```
p-value of tempo: 0.08657363
p-value of complexity: 0.00000317
p-value of mode: 0.75308241
--- STEP 3 ---
current model: popularity_norm ~ age_days + complexity
possible variables:
p-value of name_len: 0.44094297
p-value of track_number: 0.00001266
p-value of duration: 0.53709664
p-value of acousticness: 0.02029146
p-value of danceability: 0.05820237
p-value of energy: 0.00832388
p-value of loudness: 0.01648438
p-value of speechiness: 0.70579816
p-value of valence: 0.47059392
p-value of tempo: 0.23158259
p-value of mode: 0.47487225
--- STEP 4 ---
current model: popularity_norm ~ age_days + complexity + track_number
possible variables:
p-value of name_len: 0.62457759
p-value of duration: 0.76106354
p-value of acousticness: 0.21122935
p-value of danceability: 0.09193785
p-value of energy: 0.07424965
p-value of loudness: 0.13148060
p-value of speechiness: 0.30669822
p-value of valence: 0.79107073
p-value of tempo: 0.49622858
p-value of mode: 0.32506313
The minimal p-value is higher than 0.05, returning the previous model
The best model for relative popularity excluding correlations obtained via forward s
election is `popularity_norm ~ age_days + complexity + track_number`.
```

In [15]:

```
correlations = ["duration*complexity", "acousticness*energy", "energy*loudness", "tr
explanatory_vars.extend(correlations)
```

In [16]:

```
best_corr_fs_abs = forward_selection("popularity_abs", explanatory_vars, data)
print("The best model for absolute popularity including correlations obtained via fo
```

```
--- STEP 1 ---
current model: popularity_abs ~
possible variables:
p-value of name_len: 0.46054799
p-value of track_number: 0.21061084
p-value of duration: 0.00087687
p-value of acousticness: 0.00576701
p-value of danceability: 0.00328833
p-value of energy: 0.03409356
p-value of loudness: 0.00488552
p-value of speechiness: 0.22617057
p-value of valence: 0.03859374
p-value of tempo: 0.54998284
p-value of complexity: 0.00001846
p-value of age_days: 0.00000017
p-value of mode: 0.91191181
p-value of duration*complexity: 0.73151876
p-value of acousticness*energy: 0.00982869
p-value of energy*loudness: 0.25318394
p-value of track_number*complexity: 0.31355216
p-value of track_number*duration: 0.15667675
p-value of duration*loudness: 0.29716809
p-value of duration*speechiness: 0.02035317
p-value of acousticness*loudness: 0.46479656
p-value of danceability*valence: 0.10095925
p-value of danceability*complexity: 0.20162847
p-value of loudness*complexity: 0.19865457
```

```
p-value of valence*complexity: 0.88515857
--- STEP 2 ---
current model: popularity_abs ~ age_days
possible variables:
p-value of name_len: 0.91292203
p-value of track_number: 0.04107383
p-value of duration: 0.00014822
p-value of acousticness: 0.00073659
p-value of danceability: 0.00018752
p-value of energy: 0.00038229
p-value of loudness: 0.00000553
p-value of speechiness: 0.23901821
p-value of valence: 0.02374235
p-value of tempo: 0.08657363
p-value of complexity: 0.00000317
p-value of mode: 0.75308241
p-value of duration*complexity: 0.24991864
p-value of acousticness*energy: 0.02824305
p-value of energy*loudness: 0.44533786
p-value of track_number*complexity: 0.09104340
p-value of track_number*duration: 0.00308632
p-value of duration*loudness: 0.06517086
p-value of duration*speechiness: 0.05489606
p-value of acousticness*loudness: 0.60675094
p-value of danceability*valence: 0.13839380
p-value of danceability*complexity: 0.25585069
p-value of loudness*complexity: 0.11467147
p-value of valence*complexity: 0.69210417
--- STEP 3 ---
current model: popularity_abs ~ age_days + complexity
possible variables:
p-value of name_len: 0.44094297
p-value of track_number: 0.00001266
p-value of duration: 0.53709664
p-value of acousticness: 0.02029146
p-value of danceability: 0.05820237
p-value of energy: 0.00832388
p-value of loudness: 0.01648438
p-value of speechiness: 0.70579816
p-value of valence: 0.47059392
p-value of tempo: 0.23158259
p-value of mode: 0.47487225
p-value of duration*complexity: 0.24991864
p-value of acousticness*energy: 0.22874350
p-value of energy*loudness: 0.34073603
p-value of track_number*complexity: 0.09104340
p-value of track_number*duration: 0.01941142
p-value of duration*loudness: 0.09515501
p-value of duration*speechiness: 0.16117078
p-value of acousticness*loudness: 0.48848629
p-value of danceability*valence: 0.39121081
p-value of danceability*complexity: 0.25585069
p-value of loudness*complexity: 0.11467147
p-value of valence*complexity: 0.69210417
--- STEP 4 ---
current model: popularity_abs ~ age_days + complexity + track_number
possible variables:
p-value of name_len: 0.62457759
p-value of duration: 0.76106354
p-value of acousticness: 0.21122935
p-value of danceability: 0.09193785
p-value of energy: 0.07424965
p-value of loudness: 0.13148060
p-value of speechiness: 0.30669822
p-value of valence: 0.79107073
p-value of tempo: 0.49622858
p-value of mode: 0.32506313
p-value of duration*complexity: 0.32524097
p-value of acousticness*energy: 0.46456826
```

```
p-value of energy*loudness: 0.75991528
p-value of track_number*complexity: 0.09104340
p-value of track_number*duration: 0.01941142
p-value of duration*loudness: 0.30268858
p-value of duration*speechiness: 0.43587191
p-value of acousticness*loudness: 0.47629115
p-value of danceability*valence: 0.18085494
p-value of danceability*complexity: 0.63639693
p-value of loudness*complexity: 0.51885671
p-value of valence*complexity: 0.97031593
--- STEP 5 ---
current model: popularity_abs ~ age_days + complexity + track_number + track_number*
duration
possible variables:
p-value of name_len: 0.40263598
p-value of duration: 0.07554685
p-value of acousticness: 0.75605238
p-value of danceability: 0.04040702
p-value of energy: 0.49257598
p-value of loudness: 0.57376604
p-value of speechiness: 0.62767628
p-value of valence: 0.72289917
p-value of tempo: 0.98619012
p-value of mode: 0.47753492
p-value of duration*complexity: 0.99458273
p-value of acousticness*energy: 0.32890771
p-value of energy*loudness: 0.62738498
p-value of track_number*complexity: 0.45650402
p-value of duration*loudness: 0.54217444
p-value of duration*speechiness: 0.56114382
p-value of acousticness*loudness: 0.63978009
p-value of danceability*valence: 0.22002064
p-value of danceability*complexity: 0.09333917
p-value of loudness*complexity: 0.60612670
p-value of valence*complexity: 0.79715605
--- STEP 6 ---
current model: popularity_abs ~ age_days + complexity + track_number + track_number*
duration + danceability
possible variables:
p-value of name_len: 0.42546278
p-value of duration: 0.01540413
p-value of acousticness: 0.58426140
p-value of energy: 0.34952138
p-value of loudness: 0.42081623
p-value of speechiness: 0.32402119
p-value of valence: 0.35214988
p-value of tempo: 0.36070424
p-value of mode: 0.43042383
p-value of duration*complexity: 0.91346388
p-value of acousticness*energy: 0.29808733
p-value of energy*loudness: 0.65180163
p-value of track_number*complexity: 0.38652646
p-value of duration*loudness: 0.34584063
p-value of duration*speechiness: 0.29933269
p-value of acousticness*loudness: 0.56688441
p-value of danceability*valence: 0.22002064
p-value of danceability*complexity: 0.09333917
p-value of loudness*complexity: 0.45009756
p-value of valence*complexity: 0.86123023
--- STEP 7 ---
current model: popularity_abs ~ age_days + complexity + track_number + track_number*
duration + danceability + duration
possible variables:
p-value of name_len: 0.42546278
p-value of acousticness: 0.58426140
p-value of energy: 0.34952138
p-value of loudness: 0.42081623
p-value of speechiness: 0.32402119
p-value of valence: 0.35214988
```

```
p-value of tempo: 0.36070424
p-value of mode: 0.43042383
p-value of duration*complexity: 0.91346388
p-value of acousticness*energy: 0.29808733
p-value of energy*loudness: 0.65180163
p-value of track_number*complexity: 0.38652646
p-value of duration*loudness: 0.34584063
p-value of duration*speechiness: 0.29933269
p-value of acousticness*loudness: 0.56688441
p-value of danceability*valence: 0.22002064
p-value of danceability*complexity: 0.09333917
p-value of loudness*complexity: 0.45009756
p-value of valence*complexity: 0.86123023
The minimal p-value is higher than 0.05, returning the previous model
The best model for absolute popularity including correlations obtained via forward s
election is `popularity_abs ~ age_days + complexity + track_number + track_number*du
ration + danceability + duration`.
```

In [17]:

```
best_corr_fs_rel = forward_selection("popularity_norm", explanatory_vars, data)
print("The best model for relative popularity including correlations obtained via fo
```

```
--- STEP 1 ---
current model: popularity_norm ~
possible variables:
p-value of name_len: 0.46054799
p-value of track_number: 0.21061084
p-value of duration: 0.00087687
p-value of acousticness: 0.00576701
p-value of danceability: 0.00328833
p-value of energy: 0.03409356
p-value of loudness: 0.00488552
p-value of speechiness: 0.22617057
p-value of valence: 0.03859374
p-value of tempo: 0.54998284
p-value of complexity: 0.00001846
p-value of age_days: 0.00000017
p-value of mode: 0.91191181
p-value of duration*complexity: 0.73151876
p-value of acousticness*energy: 0.00982869
p-value of energy*loudness: 0.25318394
p-value of track_number*complexity: 0.31355216
p-value of track_number*duration: 0.15667675
p-value of duration*loudness: 0.29716809
p-value of duration*speechiness: 0.02035317
p-value of acousticness*loudness: 0.46479656
p-value of danceability*valence: 0.10095925
p-value of danceability*complexity: 0.20162847
p-value of loudness*complexity: 0.19865457
p-value of valence*complexity: 0.88515857
--- STEP 2 ---
current model: popularity_norm ~ age_days
possible variables:
p-value of name_len: 0.91292203
p-value of track_number: 0.04107383
p-value of duration: 0.00014822
p-value of acousticness: 0.00073659
p-value of danceability: 0.00018752
p-value of energy: 0.00038229
p-value of loudness: 0.00000553
p-value of speechiness: 0.23901821
p-value of valence: 0.02374235
p-value of tempo: 0.08657363
p-value of complexity: 0.00000317
p-value of mode: 0.75308241
p-value of duration*complexity: 0.24991864
p-value of acousticness*energy: 0.02824305
p-value of energy*loudness: 0.44533786
p-value of track_number*complexity: 0.09104340
```

```
p-value of track_number*duration: 0.00308632
p-value of duration*loudness: 0.06517086
p-value of duration*speechiness: 0.05489606
p-value of acousticness*loudness: 0.60675094
p-value of danceability*valence: 0.13839380
p-value of danceability*complexity: 0.25585069
p-value of loudness*complexity: 0.11467147
p-value of valence*complexity: 0.69210417
--- STEP 3 ---
current model: popularity_norm ~ age_days + complexity
possible variables:
p-value of name_len: 0.44094297
p-value of track_number: 0.00001266
p-value of duration: 0.53709664
p-value of acousticness: 0.02029146
p-value of danceability: 0.05820237
p-value of energy: 0.00832388
p-value of loudness: 0.01648438
p-value of speechiness: 0.70579816
p-value of valence: 0.47059392
p-value of tempo: 0.23158259
p-value of mode: 0.47487225
p-value of duration*complexity: 0.24991864
p-value of acousticness*energy: 0.22874350
p-value of energy*loudness: 0.34073603
p-value of track_number*complexity: 0.09104340
p-value of track_number*duration: 0.01941142
p-value of duration*loudness: 0.09515501
p-value of duration*speechiness: 0.16117078
p-value of acousticness*loudness: 0.48848629
p-value of danceability*valence: 0.39121081
p-value of danceability*complexity: 0.25585069
p-value of loudness*complexity: 0.11467147
p-value of valence*complexity: 0.69210417
--- STEP 4 ---
current model: popularity_norm ~ age_days + complexity + track_number
possible variables:
p-value of name_len: 0.62457759
p-value of duration: 0.76106354
p-value of acousticness: 0.21122935
p-value of danceability: 0.09193785
p-value of energy: 0.07424965
p-value of loudness: 0.13148060
p-value of speechiness: 0.30669822
p-value of valence: 0.79107073
p-value of tempo: 0.49622858
p-value of mode: 0.32506313
p-value of duration*complexity: 0.32524097
p-value of acousticness*energy: 0.46456826
p-value of energy*loudness: 0.75991528
p-value of track_number*complexity: 0.09104340
p-value of track_number*duration: 0.01941142
p-value of duration*loudness: 0.30268858
p-value of duration*speechiness: 0.43587191
p-value of acousticness*loudness: 0.47629115
p-value of danceability*valence: 0.18085494
p-value of danceability*complexity: 0.63639693
p-value of loudness*complexity: 0.51885671
p-value of valence*complexity: 0.97031593
--- STEP 5 ---
current model: popularity_norm ~ age_days + complexity + track_number + track_number
*duration
possible variables:
p-value of name_len: 0.40263598
p-value of duration: 0.07554685
p-value of acousticness: 0.75605238
p-value of danceability: 0.04040702
p-value of energy: 0.49257598
p-value of loudness: 0.57376604
```

```
p-value of speechiness: 0.62767628
p-value of valence: 0.72289917
p-value of tempo: 0.98619012
p-value of mode: 0.47753492
p-value of duration*complexity: 0.99458273
p-value of acousticness*energy: 0.32890771
p-value of energy*loudness: 0.62738498
p-value of track_number*complexity: 0.45650402
p-value of duration*loudness: 0.54217444
p-value of duration*speechiness: 0.56114382
p-value of acousticness*loudness: 0.63978009
p-value of danceability*valence: 0.22002064
p-value of danceability*complexity: 0.09333917
p-value of loudness*complexity: 0.60612670
p-value of valence*complexity: 0.79715605
--- STEP 6 ---
current model: popularity_norm ~ age_days + complexity + track_number + track_number
*duration + danceability
possible variables:
p-value of name_len: 0.42546278
p-value of duration: 0.01540413
p-value of acousticness: 0.58426140
p-value of energy: 0.34952138
p-value of loudness: 0.42081623
p-value of speechiness: 0.32402119
p-value of valence: 0.35214988
p-value of tempo: 0.36070424
p-value of mode: 0.43042383
p-value of duration*complexity: 0.91346388
p-value of acousticness*energy: 0.29808733
p-value of energy*loudness: 0.65180163
p-value of track_number*complexity: 0.38652646
p-value of duration*loudness: 0.34584063
p-value of duration*speechiness: 0.29933269
p-value of acousticness*loudness: 0.56688441
p-value of danceability*valence: 0.22002064
p-value of danceability*complexity: 0.09333917
p-value of loudness*complexity: 0.45009756
p-value of valence*complexity: 0.86123023
--- STEP 7 ---
current model: popularity_norm ~ age_days + complexity + track_number + track_number
*duration + danceability + duration
possible variables:
p-value of name_len: 0.42546278
p-value of acousticness: 0.58426140
p-value of energy: 0.34952138
p-value of loudness: 0.42081623
p-value of speechiness: 0.32402119
p-value of valence: 0.35214988
p-value of tempo: 0.36070424
p-value of mode: 0.43042383
p-value of duration*complexity: 0.91346388
p-value of acousticness*energy: 0.29808733
p-value of energy*loudness: 0.65180163
p-value of track_number*complexity: 0.38652646
p-value of duration*loudness: 0.34584063
p-value of duration*speechiness: 0.29933269
p-value of acousticness*loudness: 0.56688441
p-value of danceability*valence: 0.22002064
p-value of danceability*complexity: 0.09333917
p-value of loudness*complexity: 0.45009756
p-value of valence*complexity: 0.86123023
The minimal p-value is higher than 0.05, returning the previous model
The best model for relative popularity including correlations obtained via forward s
election is `popularity_norm ~ age_days + complexity + track_number + track_number*d
uration + danceability + duration`.
```