Appendix IV: The relation of explanatory variables and the response variable

In [6]:
```
"""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 [7]:
```
def make_scatterplot(x_data: Iterable, y_data: Iterable, xlabel: str = "x", ylabel:
    """Prints a scatterplot.

    Args:
        x_data (Iterable): the one dimensional data to plot on the x axis.
        y_data (Iterable): the one dimensional data to plot on the y axis.
        xlabel (str, optional): the label for the x axis. Defaults to "x".
        ylabel (str, optional): the label for the y axis. Defaults to "y".
        title (str, optional): the title of the plot. Defaults to "A scatterplot".
    """
    plt.figure()
    plt.scatter(x_data, y_data)
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.title(title)
    plt.show()
```

In [8]:
```
def make_stripplot(cat_var: str, num_var: str, data: Iterable, title:str="The distri
    """Prints a scatter plot of a categorical variable's values [x axis] against a n

    Args:
        cat_var (str): the name of the categorical variable.
        num_var (str): the name of the numerical variable.
        data (Iterable): the two dimensional data which includes the categorical and
        title (str, optional): the title of the plot. Defaults to "The distribution"
    """
    plt.figure()
    sns.stripplot(x=cat_var, y=num_var, data=data)
    plt.title(title)
    plt.show()
```

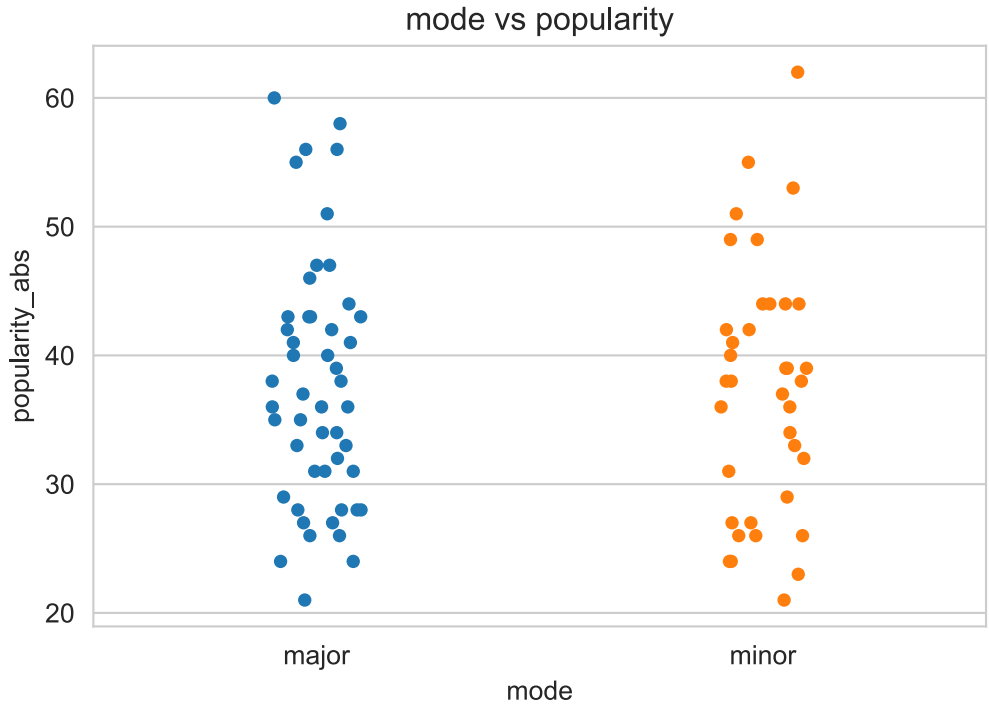
In [9]:
```
data = pd.read_csv("D:/School/frequentist-statistics/ITM-song-popularity/database/it
data = data.drop("Unnamed: 0", axis=1)
```

In [10]:
```
```

```
numerical_variables = ["name_len", "track_number", "duration", "acousticness", "danc
categorical_variables = ["mode"]
```

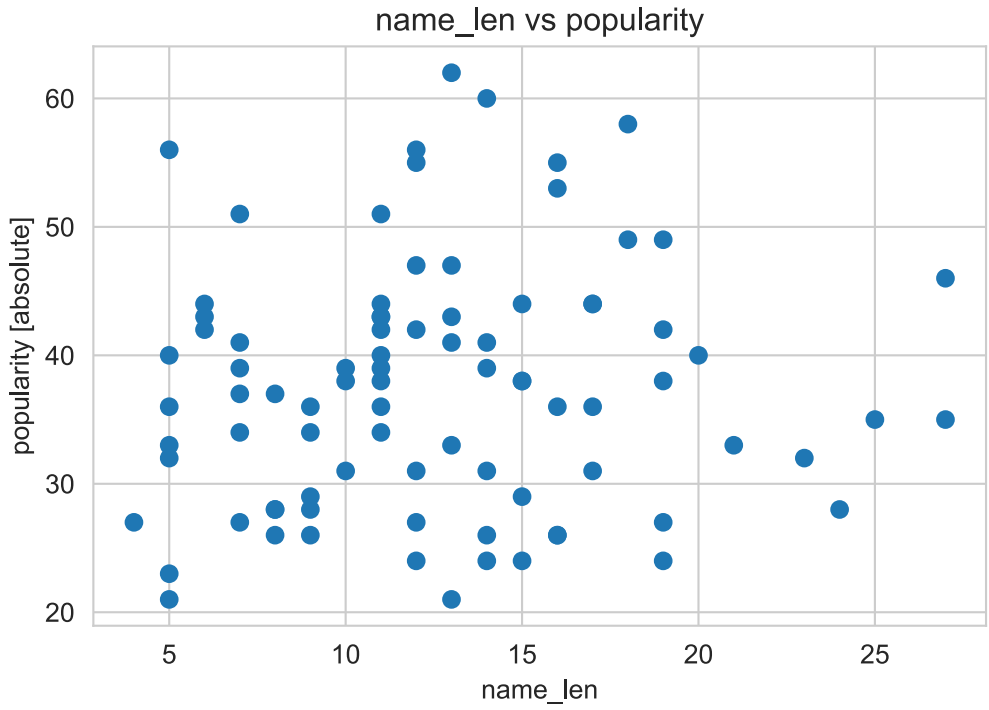
In [11]:

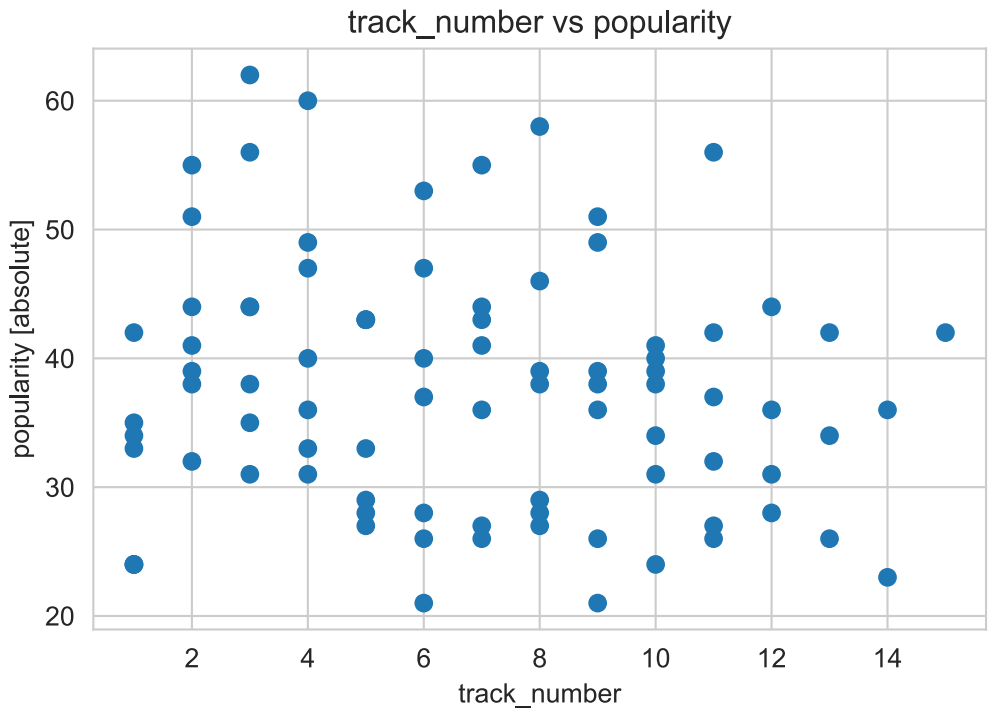
```
for x in categorical_variables:
    make_stripplot(x, "popularity_abs", data, title="%s vs popularity" % x)
```

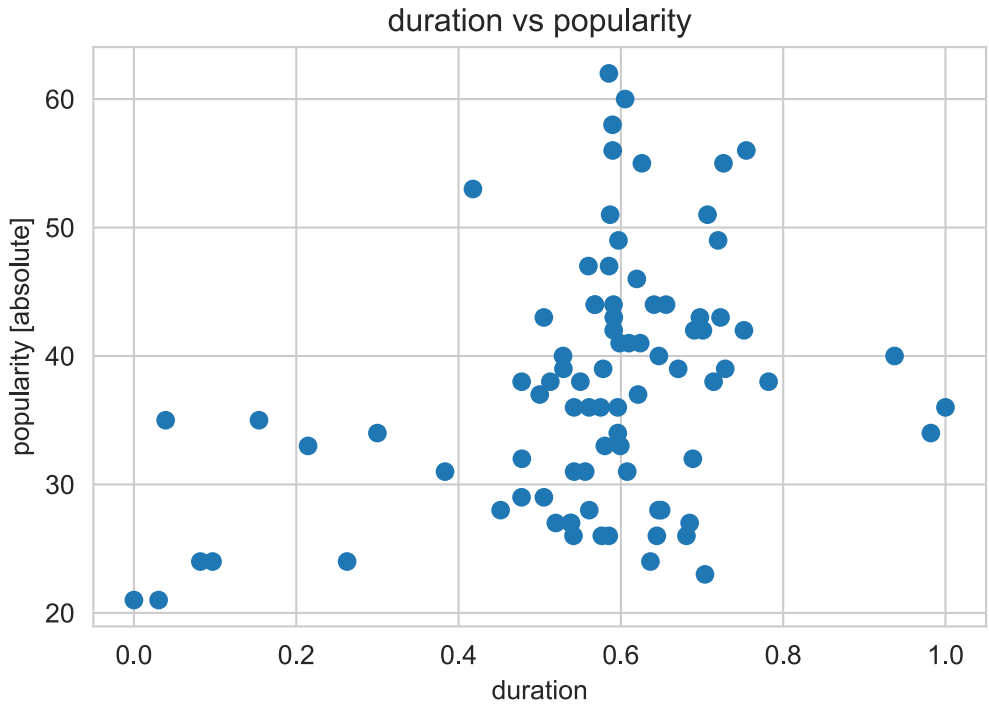


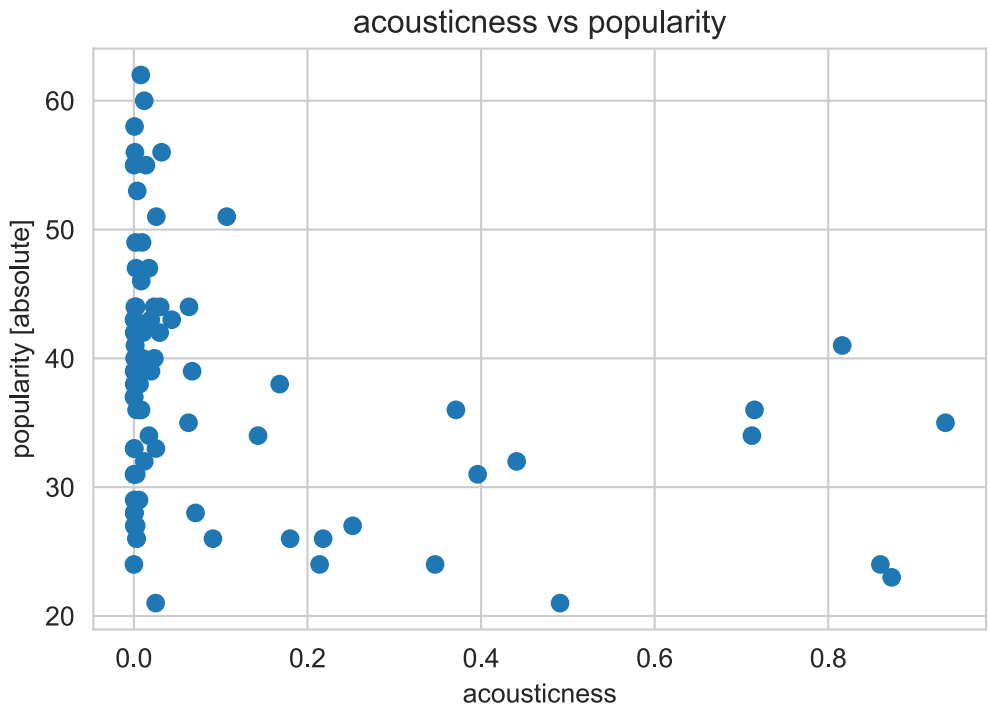
In [12]:

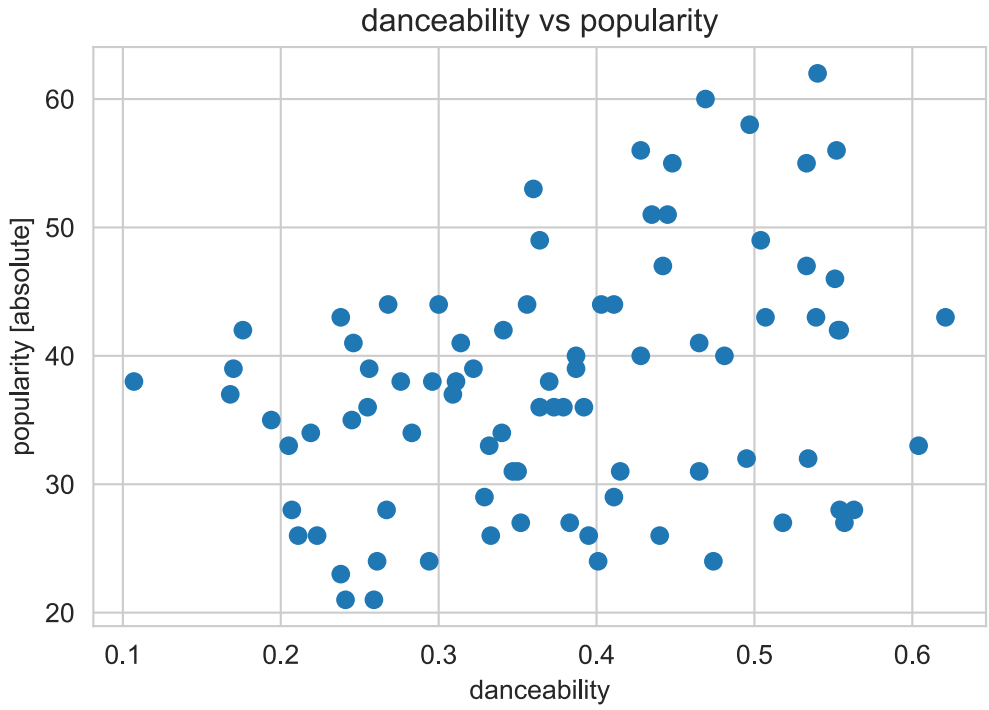
```
for x in numerical_variables:
    make_scatterplot(data[x], data["popularity_abs"], xlabel=x, ylabel="popularity [
```

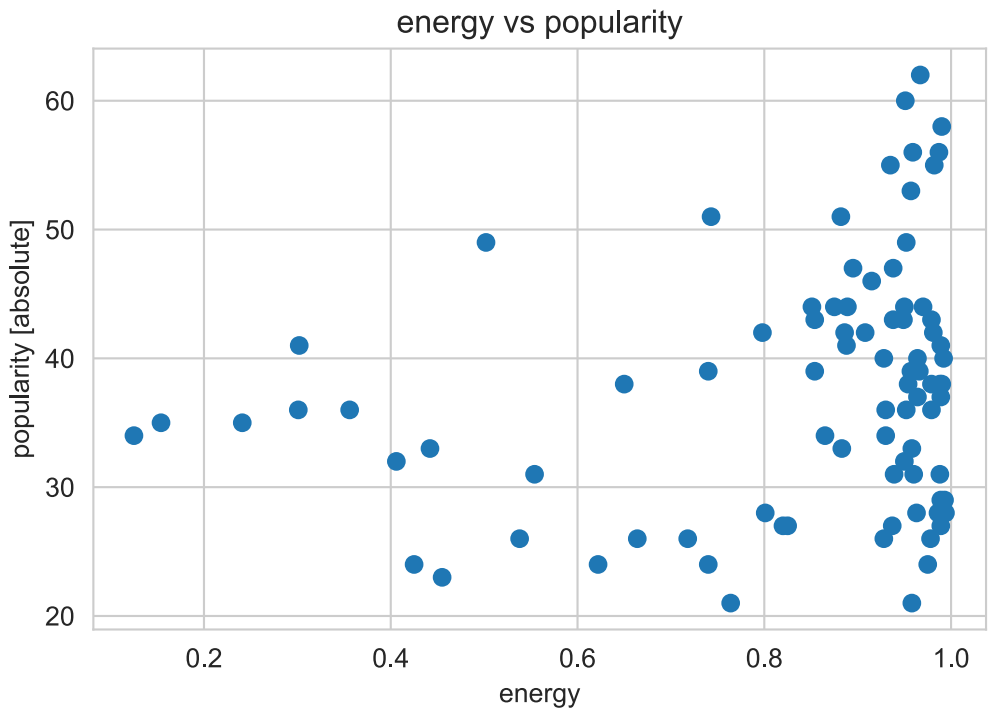


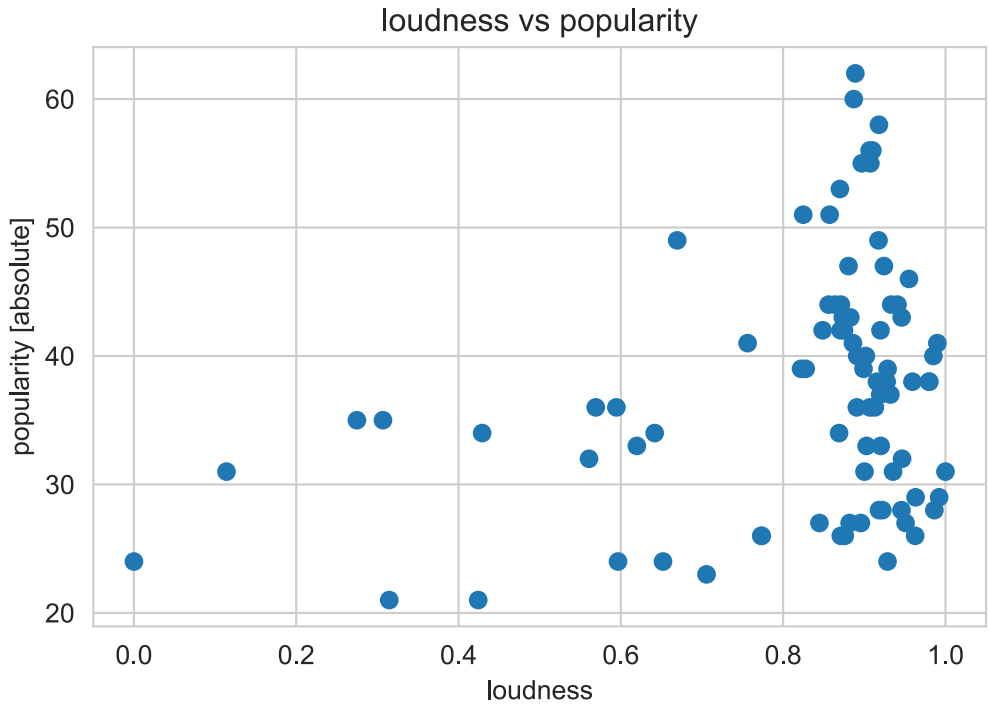

track_number vs popularity
popularity [absolute]
track_number
20
30
40
50
60
2
4
6
8
10
12
14

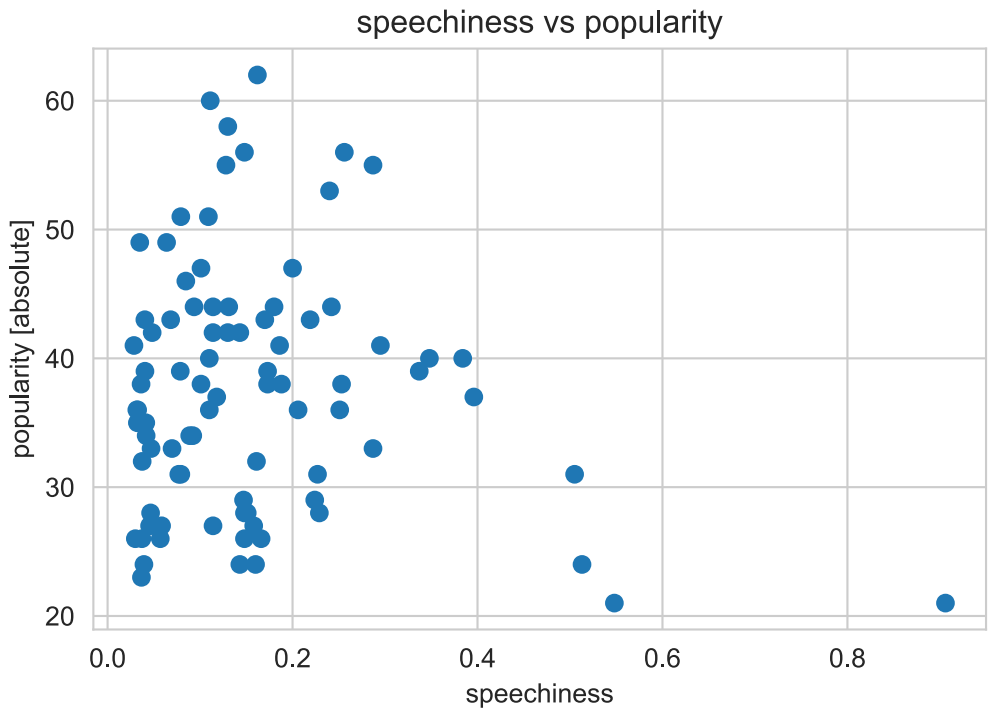

duration vs popularity
popularity [absolute]
duration
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0

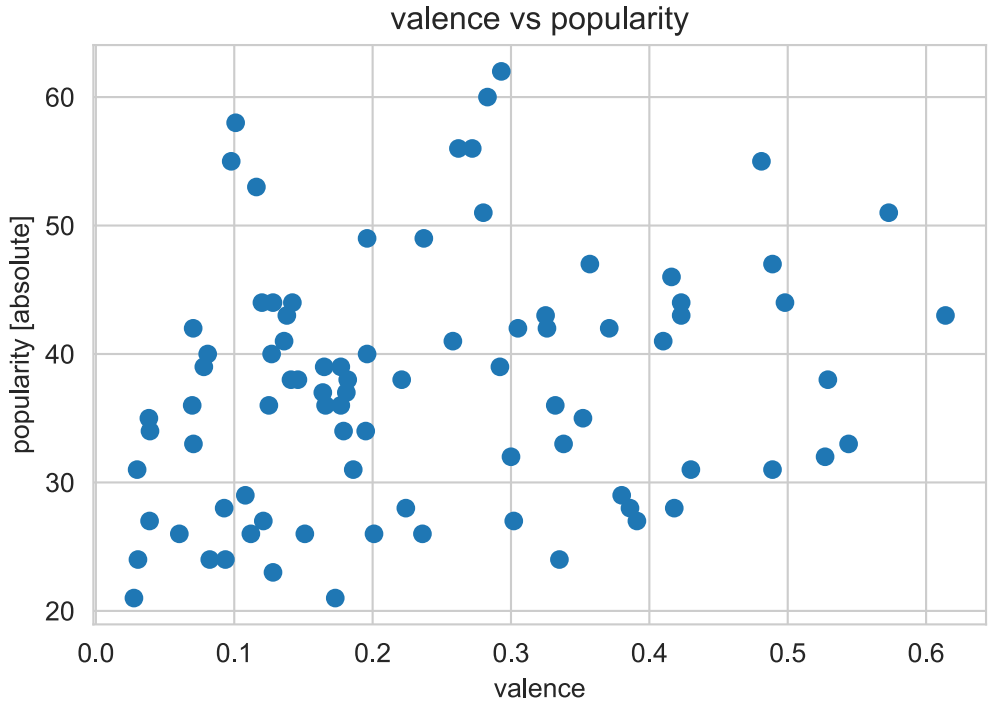

acousticness vs popularity
popularity [absolute]
60
50
40
30
20
0.0
0.2
0.4
0.6
0.8
acousticness

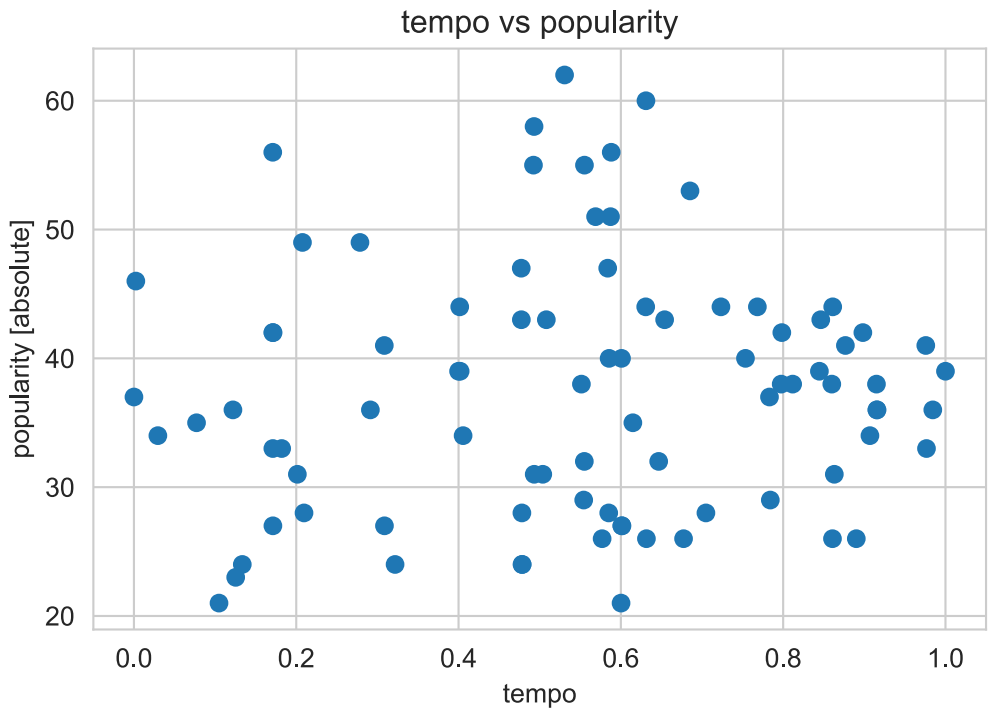

danceability vs popularity
popularity [absolute]
60
50
40
30
20
0.1
0.2
0.3
0.4
0.5
0.6
danceability

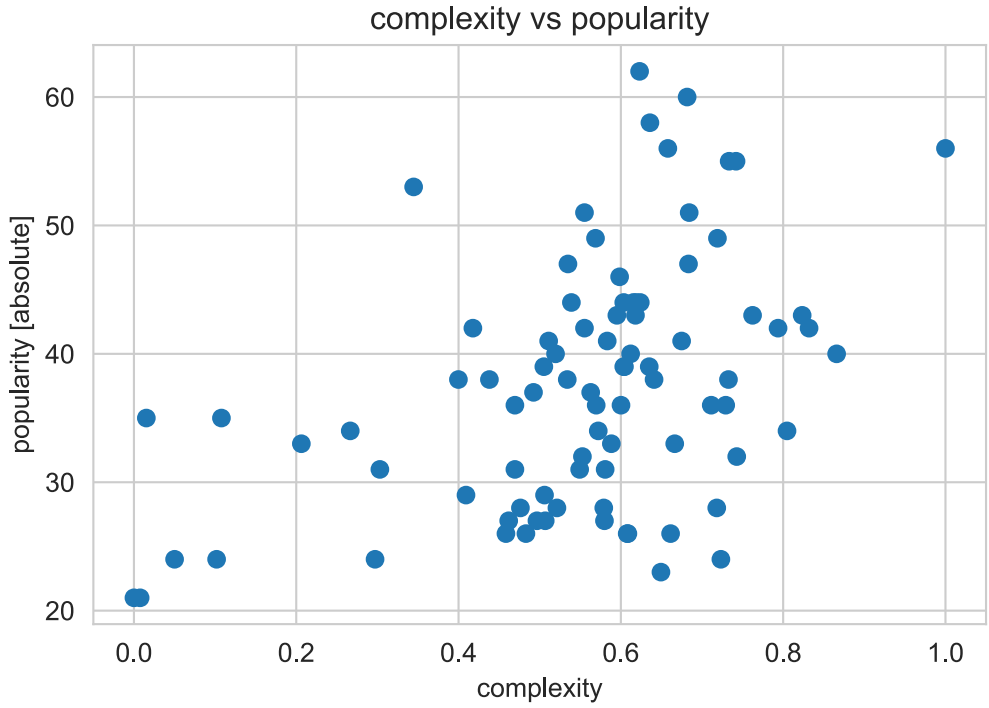

energy vs popularity
popularity [absolute]
energy
60
50
40
30
20
0.2
0.4
0.6
0.8
1.0

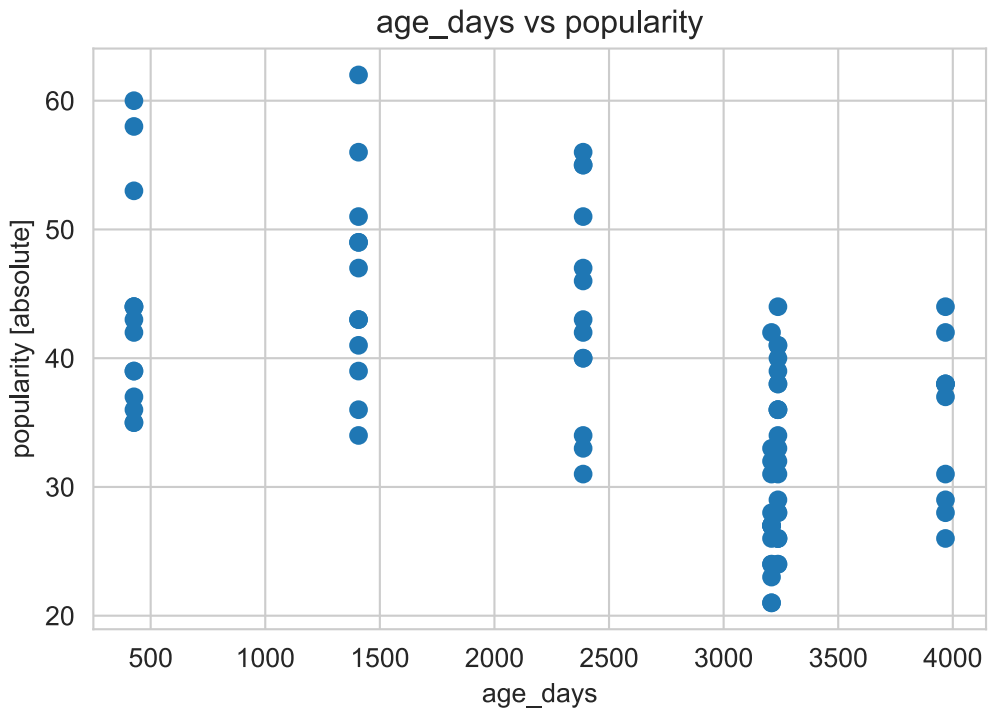

loudness vs popularity
popularity [absolute]
loudness
60
50
40
30
20
0.0
0.2
0.4
0.6
0.8
1.0


speechiness vs popularity
popularity [absolute]
60
50
40
30
20
0.0
0.2
0.4
0.6
0.8
speechiness


valence vs popularity
popularity [absolute]
60
50
40
30
20
0.0
0.1
0.2
0.3
0.4
0.5
0.6
valence


tempo vs popularity
popularity [absolute]
60
50
40
30
20
0.0
0.2
0.4
0.6
0.8
1.0
tempo


complexity vs popularity
popularity [absolute]
60
50
40
30
20
0.0
0.2
0.4
0.6
0.8
1.0
complexity


age_days vs popularity
popularity [absolute]
60
50
40
30
20
500
1000
1500
2000
2500
3000
3500
4000
age_days