Variance Inflation Factor (VIF)

Variance inflation factor is used in linear regression to detect the multi-collinearity in features. Generally if VIF is larger than 5 or 10 we exclude the feature from the regression.

Calculate VIF using numpy

Numpy uses matrix inversion, which is expensive. So we can use this method only for small datasets

```
import numpy as np

cc = np.corrcoef(X, rowvar=False)
arr_vif = np.linalg.inv(cc).diagonal()
```

Using statsmodels needs adding a constant

Statsmodels uses OLS method to fit the data.

```
import statsmodels.api as sm

X1 = sm.add_constant(X)
lst_vif = [variance_inflation_factor(X1, i) for i in range(X1.shape[1])]
```

Using scikit learn

```
def vif_sklearn(exogs, data):
    import pandas as pd
    from sklearn.linear_model import LinearRegression

    # initialize dictionaries
    vif_dict, tolerance_dict = {}, {}

    # form input data for each exogenous variable
    for exog in exogs:
        not_exog = [i for i in exogs if i != exog]
        X, y = data[not_exog], data[exog]

        # extract r-squared from the fit
        r_squared = LinearRegression().fit(X, y).score(X, y)

        # calculate VIF
        vif = 1/(1 - r_squared)
        vif_dict[exog] = vif

        # calculate tolerance
        tolerance = 1 - r_squared
        tolerance_dict[exog] = tolerance

    # return VIF DataFrame
    df_vif = pd.DataFrame({'VIF': vif_dict, 'Tolerance': tolerance_dict})

    return df_vif

# Usage
import seaborn as sns

df = sns.load_dataset('car_crashes')
exogs = ['alcohol', 'speeding', 'no_previous', 'not_distracted']
vif_sklearn(exogs=exogs, data=df)
```