DATA Initiative Research Assistant position: Data Analysis task (Project Gutenberg)

In [1]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.linear_model import LassoCV
from sklearn.preprocessing import StandardScaler

# Function to load and merge data from CSV files
def load_data():
    metadata = pd.read_csv('D:/NEU/Oncampus/DATA Initiative Fall/data/SPGC-metadata-2018
    kld_scores = pd.read_csv('D:/NEU/Oncampus/DATA Initiative Fall/data/KLDscores.csv')
    extra_controls = pd.read_csv('D:/NEU/Oncampus/DATA Initiative Fall/data/extra_contro

    # Rename columns for merging
    metadata = metadata.rename(columns={'id': 'book_id'})
    kld_scores = kld_scores.rename(columns={'filename': 'book_id'})
    extra_controls = extra_controls.rename(columns={'id': 'book_id'})

    # Merge datasets on 'book_id'
    data = metadata.merge(kld_scores, on='book_id').merge(extra_controls, on='book_id')
    return data

# Function to handle missing values and infinite values in 'log_downloads'
def handle_missing_values(data):
    # Fill missing numeric values with column mean
    data.fillna(data.mean(numeric_only=True), inplace=True)

    # Calculate log of downloads and handle infinite values
    data['log_downloads'] = np.log(data['downloads'].replace(0, np.nan))
    data['log_downloads'].replace([np.inf, -np.inf], np.nan, inplace=True)
    data.dropna(subset=['log_downloads'], inplace=True)
    return data

# Function to calculate book-level measures of KLD
def calculate_kld_measures(data):
    data['kld_values'] = data['kld_values'].apply(eval)
    data['kld_mean'] = data['kld_values'].apply(np.mean)
    data['kld_variance'] = data['kld_values'].apply(np.var)
    data['kld_slope'] = data['kld_values'].apply(lambda x: np.polyfit(range(len(x)), x,
    return data

# Function to check multicollinearity using VIF
def check_multicollinearity(X):
    vif_data = pd.DataFrame()
    vif_data["feature"] = X.columns
    vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(len(X.colum
    return vif_data

# Function to fit OLS regression model
def fit_ols_model(X, y):
    X = sm.add_constant(X)
    model = sm.OLS(y, X).fit()
    return model

# Function to perform LASSO regression and identify most predictive variables
def perform_lasso_regression(X, y):
    scaler = StandardScaler()
    X_scaled = scaler.fit_transform(X)
    lasso = LassoCV(cv=5).fit(X_scaled, y)
```

```
    predictive_vars = X.columns[lasso.coef_ != 0]
    return predictive_vars

# Main function to orchestrate the entire process
def main():
    # Load and merge data
    data = load_data()

    # Handle missing values and calculate log downloads
    data = handle_missing_values(data)

    # Calculate KLD measures
    data = calculate_kld_measures(data)

    # Prepare the regression model using available columns
    X = data[['kld_mean', 'kld_variance', 'kld_slope', 'subj2_war', 'subj2_adventure', '
    X = pd.get_dummies(X, drop_first=True)  # Convert categorical variables to dummy var
    y = data['log_downloads']

    # Check for multicollinearity
    vif_data = check_multicollinearity(X)
    print(vif_data)

    # Drop variables with high VIF (example)
    X = X.drop(columns=['kld_mean', 'speed', 'sentiment_avg', 'sentiment_vol'])

    # Fit OLS regression model
    model = fit_ols_model(X, y)
    print(model.summary())

    # Perform LASSO regression
    predictive_vars = perform_lasso_regression(X, y)
    print("Most predictive variables:", predictive_vars)

    # Summary of Analysis and Variable Descriptions
    summary = """
    The analysis reveals that certain characteristics of the Kullback-Leibler divergence

    LASSO regression further identifies that genre-specific variables like romance, fant
    """
    print(summary)

    # Display regression table
    regression_table = model.summary().tables[1]
    print(regression_table)

    # Variable descriptions
    variable_descriptions = """
    - kld_mean: Average Kullback-Leibler divergence across the narrative.
    - kld_variance: Variance of Kullback-Leibler divergence across the narrative.
    - kld_slope: Slope of a linear regression fitted to the Kullback-Leibler divergence
    - log_downloads: Natural logarithm of the download counts.
    - subj2_*: Binary indicators for various subjects (e.g., war, adventure, comedy, bio
    - speed: Reading speed.
    - sentiment_avg: Average sentiment score.
    - sentiment_vol: Sentiment volatility.
    - wordcount: Total word count.
    """
    print(variable_descriptions)

if __name__ == "__main__":
    main()
```

```
C:\Users\pramo\anaconda3\lib\site-packages\statsmodels\regression\linear_model.py:1738:
RuntimeWarning: invalid value encountered in scalar divide
```

```
  return 1 - self.ssr/self.uncentered_tss
```

	feature	VIF
0	kld_mean	320.290673
1	kld_variance	3.352122
2	kld_slope	1.854000
3	subj2_war	1.591647
4	subj2_adventure	1.628347
5	subj2_comedy	NaN
6	subj2_biography	1.066063
7	subj2_romance	1.589978
8	subj2_drama	1.021428
9	subj2_fantasy	1.187940
10	subj2_family	1.248184
11	subj2_sciencefiction	1.372743
12	subj2_action	1.003853
13	subj2_thriller	NaN
14	subj2_western	1.437641
15	subj2_horror	1.062477
16	subj2_mystery	1.804978
17	subj2_crime	1.042326
18	subj2_history	2.837638
19	subj2_periodicals	1.468506
20	subj2_others	13.224521
21	speed	192.039259
22	sentiment_avg	45.749528
23	sentiment_vol	24.265637
24	wordcount	6.530323

OLS Regression Results

Dep. Variable:	log_downloads	R-squared:	0.118
Model:	OLS	Adj. R-squared:	0.116
Method:	Least Squares	F-statistic:	60.21
Date:	Tue, 02 Jul 2024	Prob (F-statistic):	6.74e-216
Time:	16:27:33	Log-Likelihood:	-13009.
No. Observations:	8541	AIC:	2.606e+04
Df Residuals:	8521	BIC:	2.620e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.8563	0.060	47.531	0.000	2.738	2.974
kld_variance	-2.8118	13.279	-0.212	0.832	-28.841	23.218
kld_slope	-215.6133	29.156	-7.395	0.000	-272.766	-158.461
subj2_war	-0.0596	0.057	-1.050	0.294	-0.171	0.052
subj2_adventure	-0.0826	0.059	-1.405	0.160	-0.198	0.033
subj2_comedy	-4.56e-12	6.42e-13	-7.099	0.000	-5.82e-12	-3.3e-12
subj2_biography	-0.0376	0.170	-0.221	0.825	-0.371	0.296
subj2_romance	0.1346	0.068	1.979	0.048	0.001	0.268
subj2_drama	0.1836	0.311	0.590	0.555	-0.427	0.794
subj2_fantasy	0.9772	0.110	8.912	0.000	0.762	1.192
subj2_family	0.0864	0.109	0.793	0.428	-0.127	0.300
subj2_sciencefiction	1.1693	0.089	13.184	0.000	0.995	1.343
subj2_action	-0.9332	1.113	-0.839	0.402	-3.114	1.248
subj2_thriller	2.891e-12	4.34e-13	6.655	0.000	2.04e-12	3.74e-12
subj2_western	0.0987	0.086	1.150	0.250	-0.070	0.267
subj2_horror	1.7572	0.175	10.053	0.000	1.415	2.100
subj2_mystery	0.2809	0.073	3.846	0.000	0.138	0.424
subj2_crime	0.3152	0.227	1.389	0.165	-0.130	0.760
subj2_history	0.0123	0.055	0.224	0.822	-0.095	0.119
subj2_periodicals	-0.5853	0.090	-6.513	0.000	-0.761	-0.409
subj2_others	-0.2135	0.058	-3.680	0.000	-0.327	-0.100
wordcount	3.174e-06	1.75e-07	18.118	0.000	2.83e-06	3.52e-06

Omnibus:	1346.336	Durbin-Watson:	1.527

Prob(Omnibus):	0.000	Jarque-Bera (JB):	3495.680
Skew:	0.871	Prob(JB):	0.00
Kurtosis:	5.605	Cond. No.	6.50e+22

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
[2] The smallest eigenvalue is 2.81e-32. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
Most predictive variables: Index(['kld_slope', 'subj2_romance', 'subj2_fantasy', 'subj2_
sciencefiction',
       'subj2_horror', 'subj2_mystery', 'subj2_periodicals', 'subj2_others',
       'wordcount'],
      dtype='object')

The analysis reveals that certain characteristics of the Kullback-Leibler divergence (KLD) within a narrative are significant predictors of book popularity, measured as the log of download counts. Specifically, the slope of KLD over the narrative has a signific ant negative impact on log downloads, suggesting that a faster rate of information revel ation tends to decrease popularity. Additionally, various genres, such as fantasy, scien ce fiction, and horror, show strong positive effects on downloads.

LASSO regression further identifies that genre-specific variables like romance, fant asy, science fiction, horror, mystery, periodicals, and others, as well as the word coun t, are among the most independently predictive variables of book downloads. This undersc ores the importance of both the structure of information revelation and genre-specific c haracteristics in understanding book popularity.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.8563	0.060	47.531	0.000	2.738	2.974
kld_variance	-2.8118	13.279	-0.212	0.832	-28.841	23.218
kld_slope	-215.6133	29.156	-7.395	0.000	-272.766	-158.461
subj2_war	-0.0596	0.057	-1.050	0.294	-0.171	0.052
subj2_adventure	-0.0826	0.059	-1.405	0.160	-0.198	0.033
subj2_comedy	-4.56e-12	6.42e-13	-7.099	0.000	-5.82e-12	-3.3e-12
subj2_biography	-0.0376	0.170	-0.221	0.825	-0.371	0.296
subj2_romance	0.1346	0.068	1.979	0.048	0.001	0.268
subj2_drama	0.1836	0.311	0.590	0.555	-0.427	0.794
subj2_fantasy	0.9772	0.110	8.912	0.000	0.762	1.192
subj2_family	0.0864	0.109	0.793	0.428	-0.127	0.300
subj2_sciencefiction	1.1693	0.089	13.184	0.000	0.995	1.343
subj2_action	-0.9332	1.113	-0.839	0.402	-3.114	1.248
subj2_thriller	2.891e-12	4.34e-13	6.655	0.000	2.04e-12	3.74e-12
subj2_western	0.0987	0.086	1.150	0.250	-0.070	0.267
subj2_horror	1.7572	0.175	10.053	0.000	1.415	2.100
subj2_mystery	0.2809	0.073	3.846	0.000	0.138	0.424
subj2_crime	0.3152	0.227	1.389	0.165	-0.130	0.760
subj2_history	0.0123	0.055	0.224	0.822	-0.095	0.119
subj2_periodicals	-0.5853	0.090	-6.513	0.000	-0.761	-0.409
subj2_others	-0.2135	0.058	-3.680	0.000	-0.327	-0.100
wordcount	3.174e-06	1.75e-07	18.118	0.000	2.83e-06	3.52e-06

- kld_mean: Average Kullback-Leibler divergence across the narrative.
- kld_variance: Variance of Kullback-Leibler divergence across the narrative.
- kld_slope: Slope of a linear regression fitted to the Kullback-Leibler divergence values over the narrative.
- log_downloads: Natural logarithm of the download counts.
- subj2_*: Binary indicators for various subjects (e.g., war, adventure, comedy, bio graphy, romance, drama, fantasy, family, science fiction, action, thriller, western, hor ror, mystery, crime, history, periodicals, others).
- speed: Reading speed.
- sentiment_avg: Average sentiment score.

- sentiment_vol: Sentiment volatility.
- wordcount: Total word count.

In [ ]: