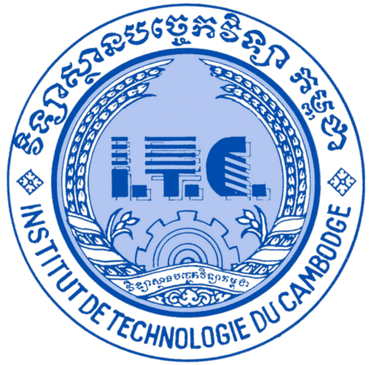


Institute of technology of Cambodia

PREDICTION OF BOOKS PRICE

PROGRAMMING FOR DATA SCIENCE

Lecterer: Chan Sophal

Start Slide

Group Members

Name	ID
HONG Kimmeng	e20200559
KHON Yin Sakal	e20200425
KEO Vonmonyroth	e20200759
KHEANG Tongheang	e20200472
EAB Pisey	e20200994
HONG Kimleng	e20200766

Group : 2

CONTENT

Data Cleaning

	Unnamed: 0	Name	Prices	Page	Format	Language	Pub Date	Publisher	Author	Dimensions	Categories
0	0	Pageboy: A Memoir	27.89	288.0	Hardcover	English	June 06, 2023	Flatiron Books	Elliot Page	5.6 X 8.6 X 1.2 inches | 0.8 pounds	Entertainment & Performing Arts Personal Mem...
1	1	Safe and Sound: A Renter-Friendly Guide to Hom...	23.24	224.0	Hardcover	English	August 22, 2023	DK Publishing (Dorling Kindersley)	Mercury Stardust	0.0 X 0.0 X 0.0 inches | 1.25 pounds	LGBT Maintenance & Repair Reference
2	2	Let This Radicalize You: Organizing and the Re...	16.69	296.0	Paperback	English	May 16, 2023	Haymarket Books	Kelly Hayes	5.5 X 8.4 X 0.8 inches | 0.85 pounds	Political Ideologies - Radicalism Feminism &...
3	3	Quietly Hostile: Essays	15.81	304.0	Paperback	English	May 16, 2023	Vintage	Samantha Irby	5.1 X 7.9 X 0.8 inches | 0.5 pounds	Personal Memoirs Form - Essays Essays
4	4	Demon Copperhead: A Pulitzer Prize Winner	30.23	560.0	Hardcover	English	October 18, 2022	Harper	Barbara Kingsolver	6.4 X 8.9 X 1.7 inches | 1.94 pounds	Literary Coming of Age Small Town & Rural ...

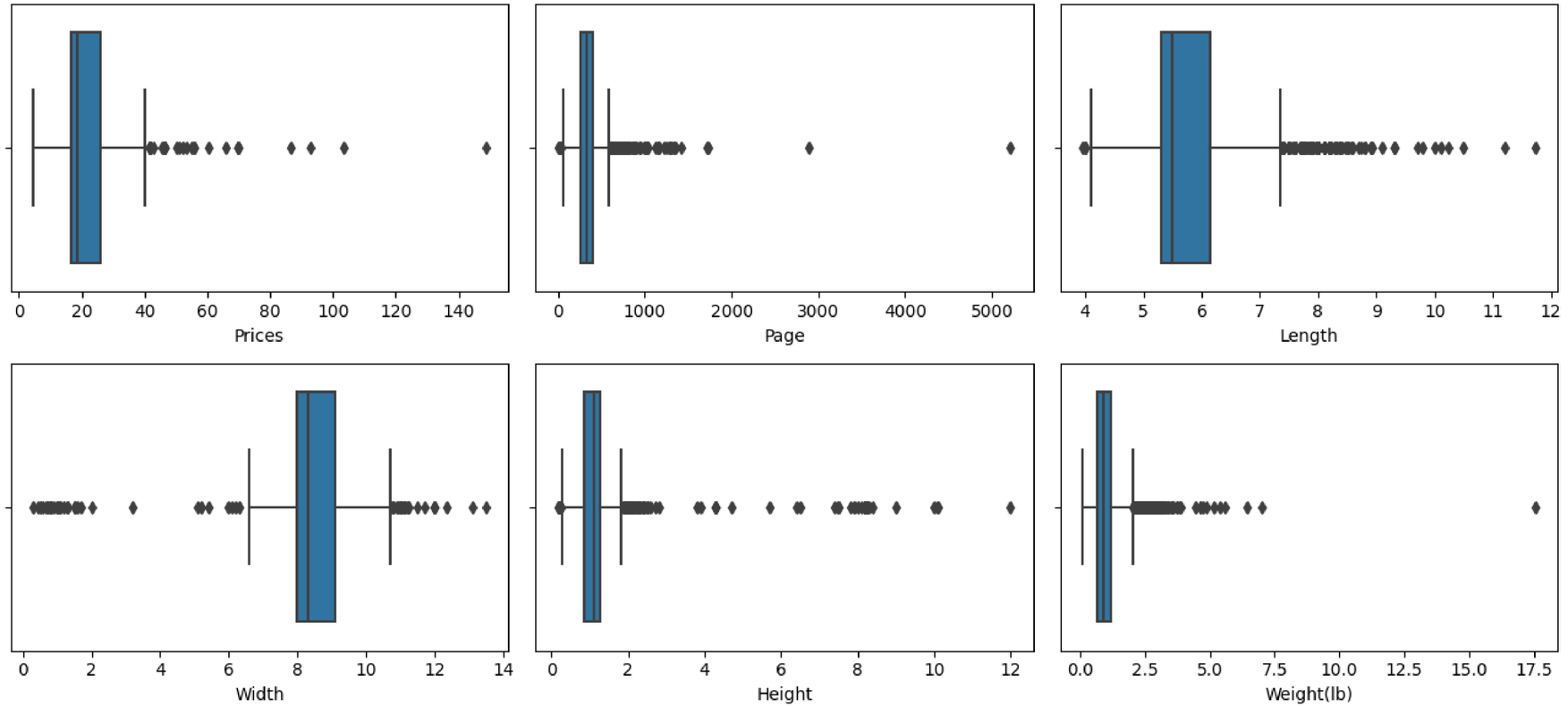
check of NaN

```
Name            0
Prices         14
Page            7
Format          0
Language        0
Pub Date        0
Publisher       0
Author          0
Length          2
Width           2
Height          2
Weight(lb)      2
Categories      0
dtype: int64
```

check if any of these features contain 0

```
Name            0
Prices          0
Page            0
Format          0
Language        0
Pub Date        0
Publisher       0
Author          0
Length        134
Width         134
Height        135
Weight(lb)     94
Categories      0
dtype: int64
```

Handling Outlier



Feature engineering

We start our feature engineering by converting the format of publish date to day count since it first published

Before converting

	Name	Prices	Page	Format	Pub Date	Publisher	Author
0	Pageboy: A Memoir	27.89	288.0	Hardcover	June 06, 2023	Flatiron Books	Elliot Page
1	Safe and Sound: A Renter-Friendly Guide to Hom...	23.24	224.0	Hardcover	August 22, 2023	DK Publishing (Dorling Kindersley)	Mercury Stardust
2	Let This Radicalize You: Organizing and the Re...	16.69	296.0	Paperback	May 16, 2023	Haymarket Books	Kelly Hayes
3	Quietly Hostile: Essays	15.81	304.0	Paperback	May 16, 2023	Vintage	Samantha Irby
4	Demon Copperhead: A Pulitzer Prize Winner	30.23	560.0	Hardcover	October 18, 2022	Harper	Barbara Kingsolver

After converting

	Name	Prices	Page	Format	Pub Date	Publisher	Author
0	Pageboy: A Memoir	27.89	288.0	Hardcover	39	Flatiron Books	Elliot Page
1	Safe and Sound: A Renter-Friendly Guide to Hom...	23.24	224.0	Hardcover	0	DK Publishing (Dorling Kindersley)	Mercury Stardust
2	Let This Radicalize You: Organizing and the Re...	16.69	296.0	Paperback	60	Haymarket Books	Kelly Hayes
3	Quietly Hostile: Essays	15.81	304.0	Paperback	60	Vintage	Samantha Irby
4	Demon Copperhead: A Pulitzer Prize Winner	30.23	560.0	Hardcover	270	Harper	Barbara Kingsolver

Feature engineering

Target encoding

```
import category_encoders as ce
target = ce.TargetEncoder(cols=['Format','Publisher','Author'])
# Fit the TargetEncoder on the training data
df_label[['Format','Publisher','Author']] = target.fit_transform(df_clean[['Format','Publisher','Author']],df_clean['Prices']
```

	Name	Prices	Page	Format	Pub Date	Publisher	Author
	Pageboy: A Memoir	27.89	288.0	27.229933	39	22.866097	21.964441
	Safe and Sound: A Renter-Friendly Guide to Hom...	23.24	224.0	27.229933	0	22.866245	21.749272
	Let This Radicalize You: Organizing and the Re...	16.69	296.0	17.386380	60	19.906852	20.897062
	Quietly Hostile: Essays	15.81	304.0	17.386380	60	15.814177	20.592877
	Demon Copperhead: A Pulitzer Prize Winner	30.23	560.0	27.229933	270	22.078555	21.249757

Feature engineering

Ordinal Encoding

```
oe = OrdinalEncoder(categories = [['Small', 'Medium', 'Large'], ['Thin', 'Medium', 'Thick']])
```

```
oe.fit(df_clean[['Cover Size', 'Thickness']])
```

```
OrdinalEncoder(categories=[['Small', 'Medium', 'Large'],
                           ['Thin', 'Medium', 'Thick']])
```

```
df_label[['Cover Size', 'Thickness']] = oe.transform(df_clean[['Cover Size', 'Thickness']])
```

	Name	Prices	Page	Format	Pub Date	Publisher	Author
	Pageboy: A Memoir	27.89	288.0	27.229933	39	22.866097	21.964441
	Safe and Sound: A Renter-Friendly Guide to Hom...	23.24	224.0	27.229933	0	22.866245	21.749272
	Let This Radicalize You: Organizing and the Re...	16.69	296.0	17.386380	60	19.906852	20.897062
	Quietly Hostile: Essays	15.81	304.0	17.386380	60	15.814177	20.592877
	Demon Copperhead: A Pulitzer Prize Winner	30.23	560.0	27.229933	270	22.078555	21.249757

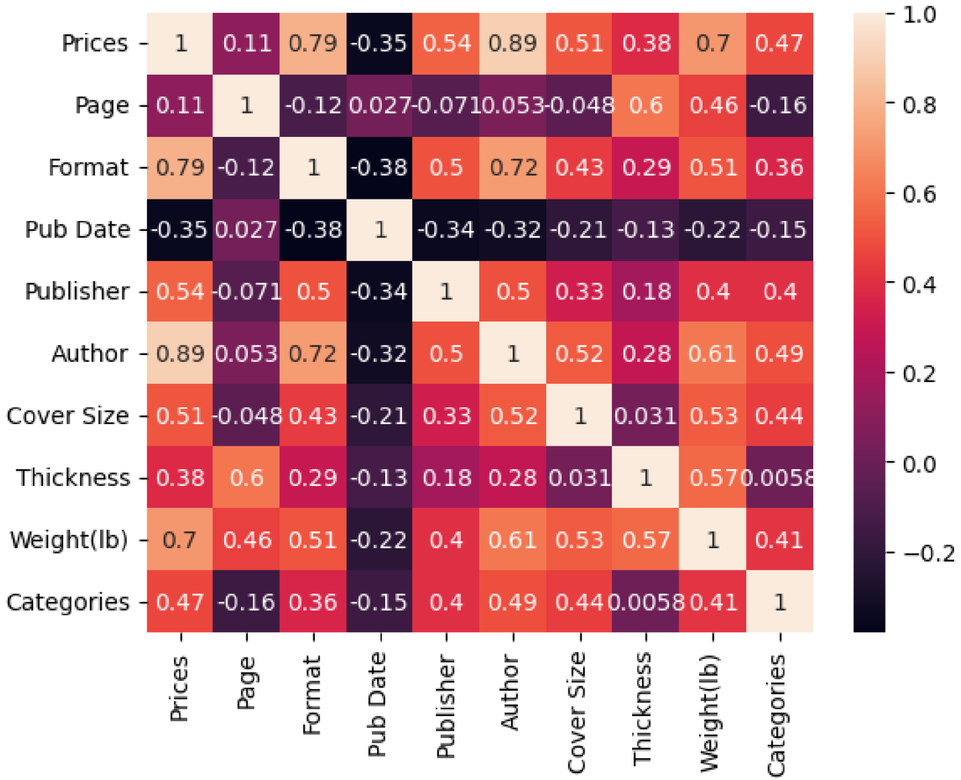
Feature engineering

Target Encoding on Book Categories

Pub Date	Publisher	Author	Cover Size	Thickness	Weight(lb)	Categories
39	22.866097	21.964441	1.0	1.0	0.80	22.494100
0	22.866245	21.749272	1.0	1.0	1.25	22.229108
60	19.906852	20.897062	1.0	0.0	0.85	21.131040
60	15.814177	20.592877	1.0	0.0	0.50	21.199378
270	22.078555	21.249757	1.0	2.0	1.94	20.230740

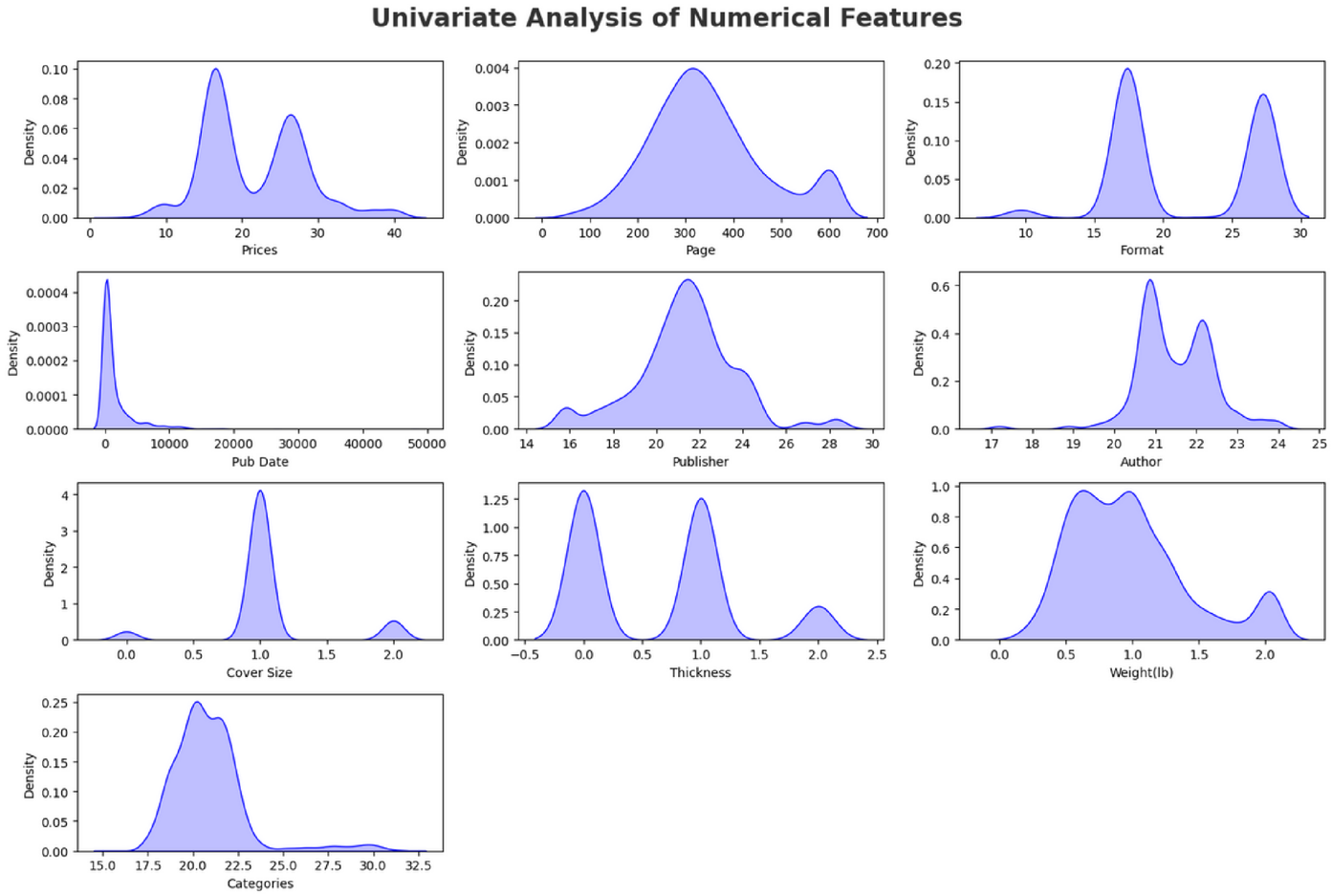
Features Visualization

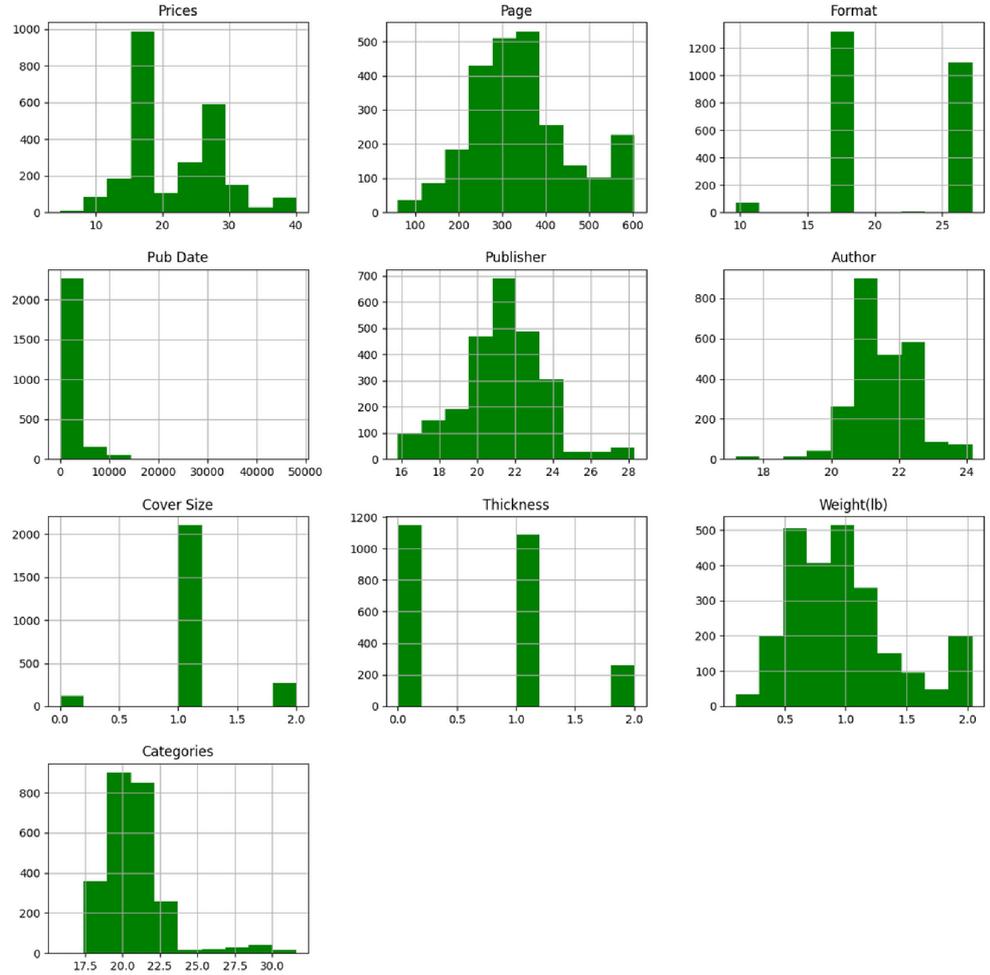
we Create a scatter plot matrix to visualize the relationships between variables.



Compute the correlation matrix to quantify the relationships between variables.

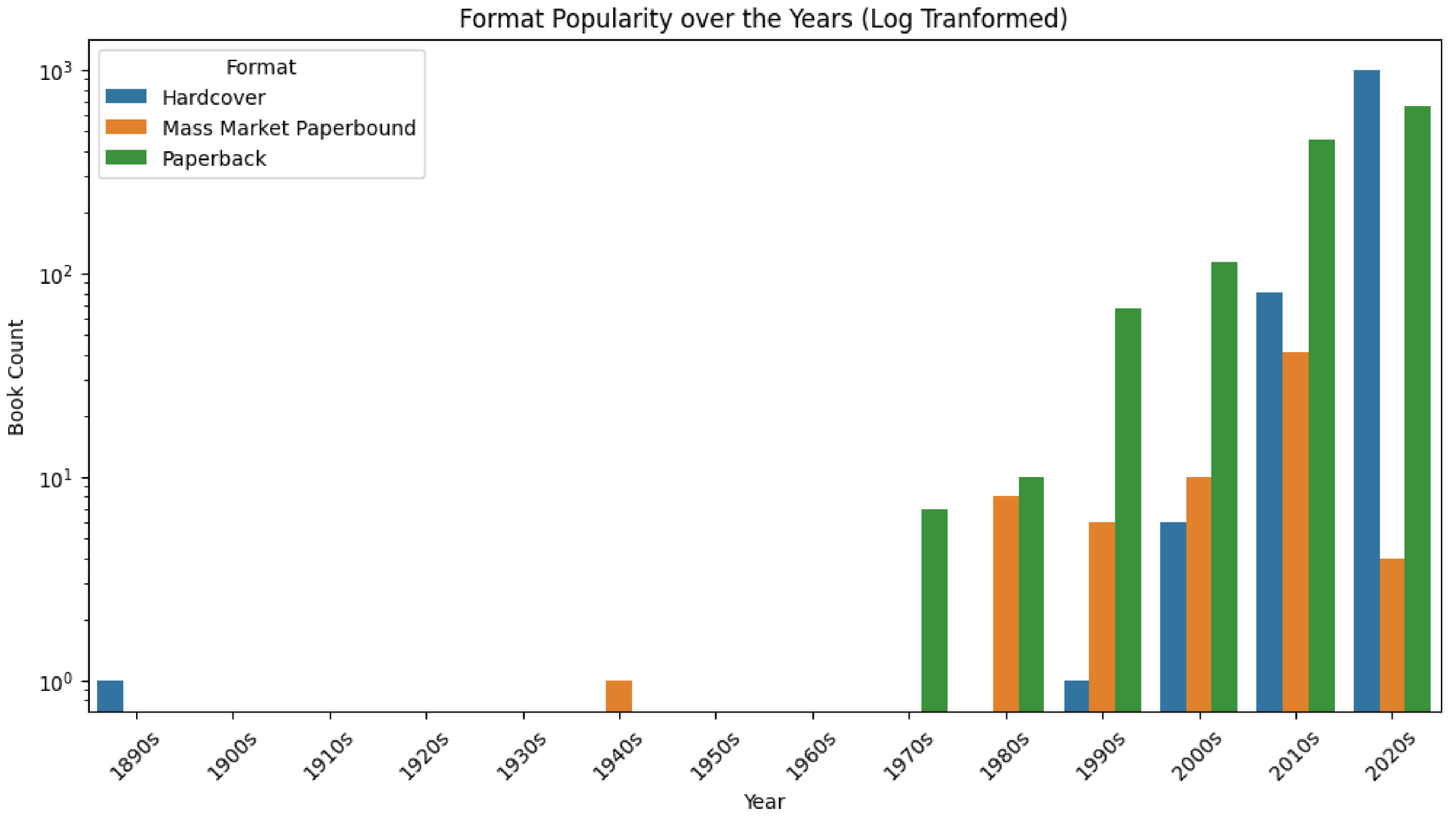
Features Visualization

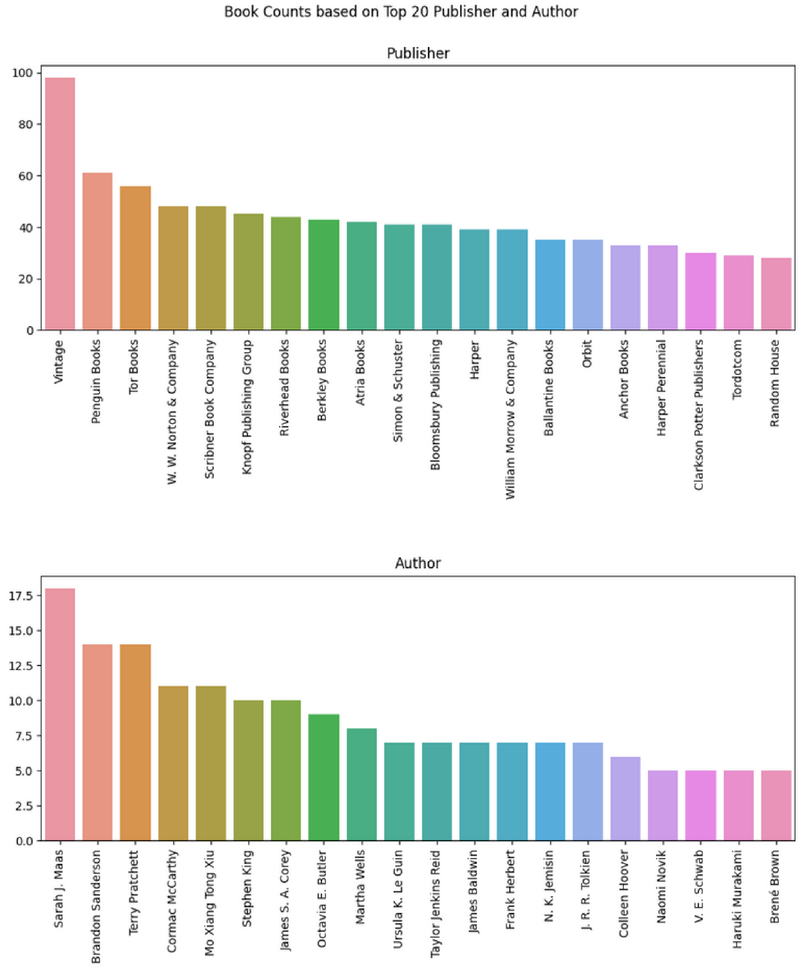




For numerical variables we histogram of density and define skewness value of each numerical feature

Features Visualization

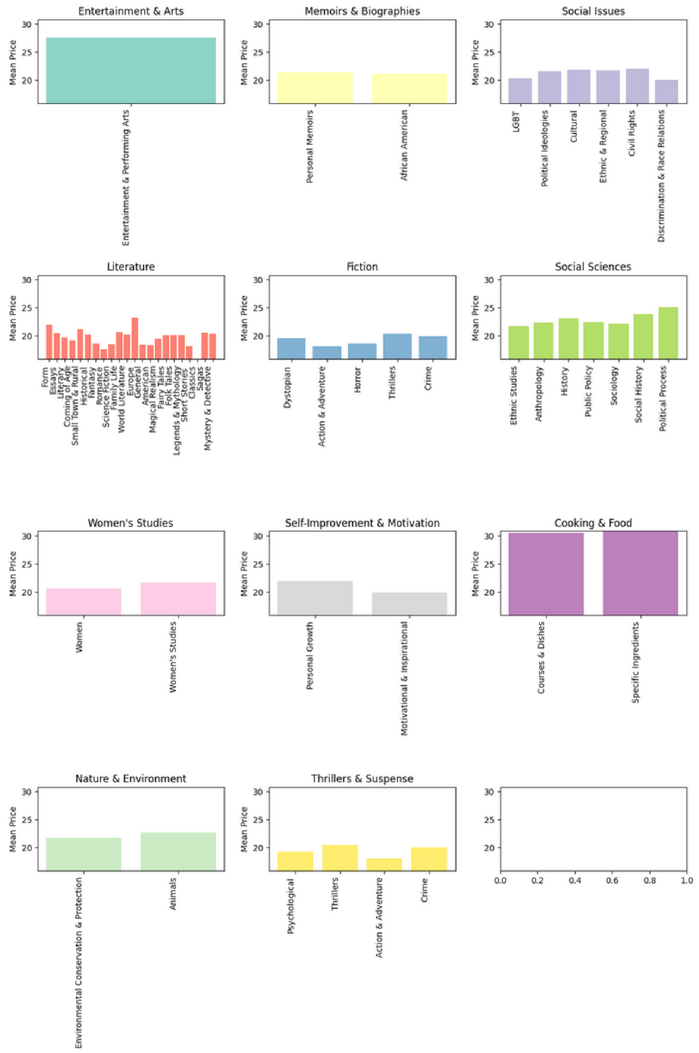


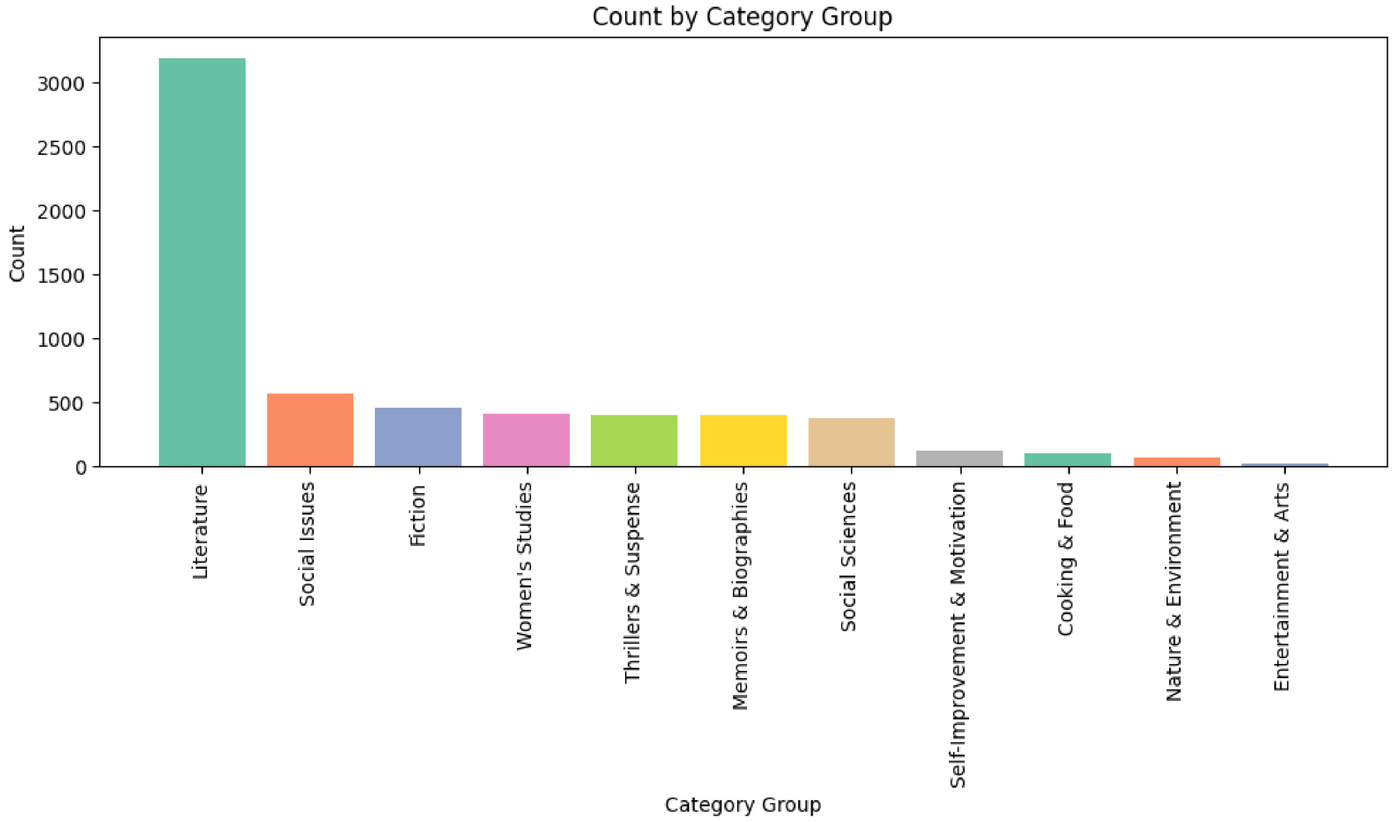


Compare to GPT Finding:

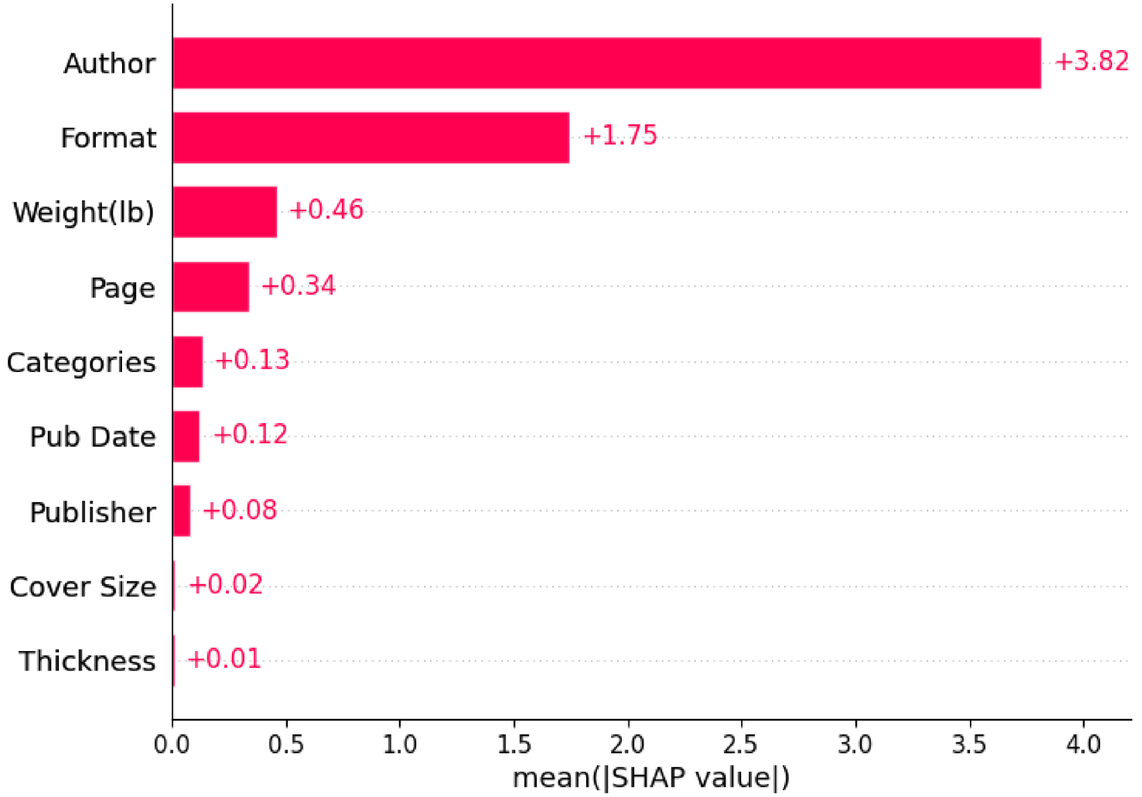
- **HardCover : Have the longest history, became more prevalent during the 19th century.**
- **MMPB : Gained popularity in the 1940s and 1950s, more cost effective and portable.**
- **PaperBack : Have more modern sense, popularity start in the 20th century.**

Features Visualization





Features Importantn



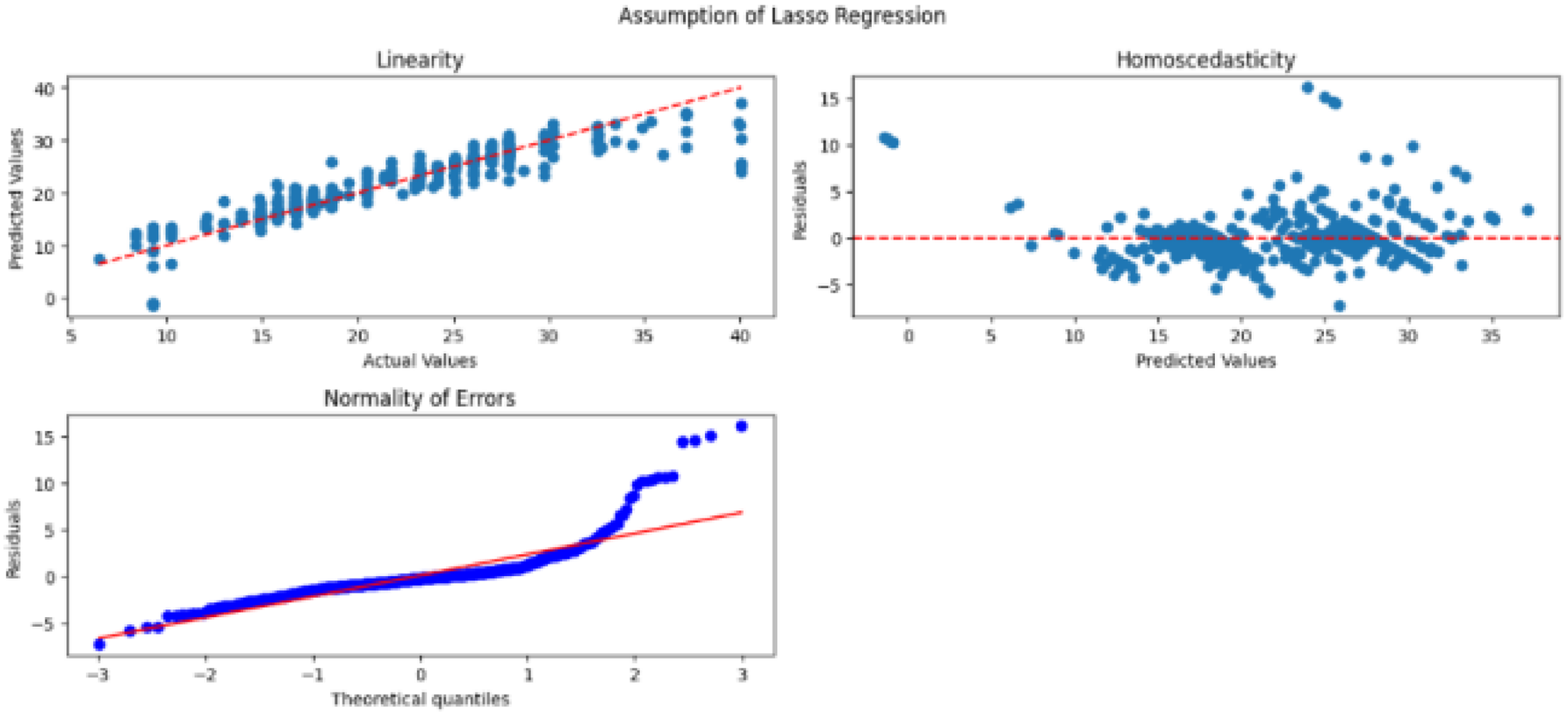
Model building

In the book price prediction project, several regression models were built to predict book prices based on the available features.
The following models were employed are

Model	Train MSE	Train R2	Predict MSE	Predict R2	Parameter
Linear Regression	5.040807	0.883399	6.620211	0.854223	None
Ridge Regression	5.040807	0.883399	6.620117	0.854225	alpha = 0.1
Lasso Regression	5.096687	0.882107	6.637076	0.853852	alpha = 0.1
Random Forest Regression	0.457943	0.989407	3.056361	0.932699	n_estimators = 100, max_depth = 10, min_samples_split = 10, min_samples_leaf = 4, max_features = 'auto', bootstrap = True
Support Vector Regression	1.825713	0.957768	4.191129	0.907711	gamma = 0.1, coef0 = 0.01, C = 10, epsilon = 0.5
MultiLayer Perception	1.986323	0.954053	3.786791	0.916615	activation = 'tanh', solver = 'adam', learning_rate = 'constant', alpha = 0.0001
Gradient Boosting Regressor	0.611615	0.985852	3.445201	0.924136	learning_rate = 0.1, max_depth = 3, n_estimators = 300, alpha = 0.1
Extreme gradient Boosting	1.451656	0.966421	3.387203	0.925421	learning_rate = 0.1, max_depth = 3, n_estimators = 100, alpha = 0.1

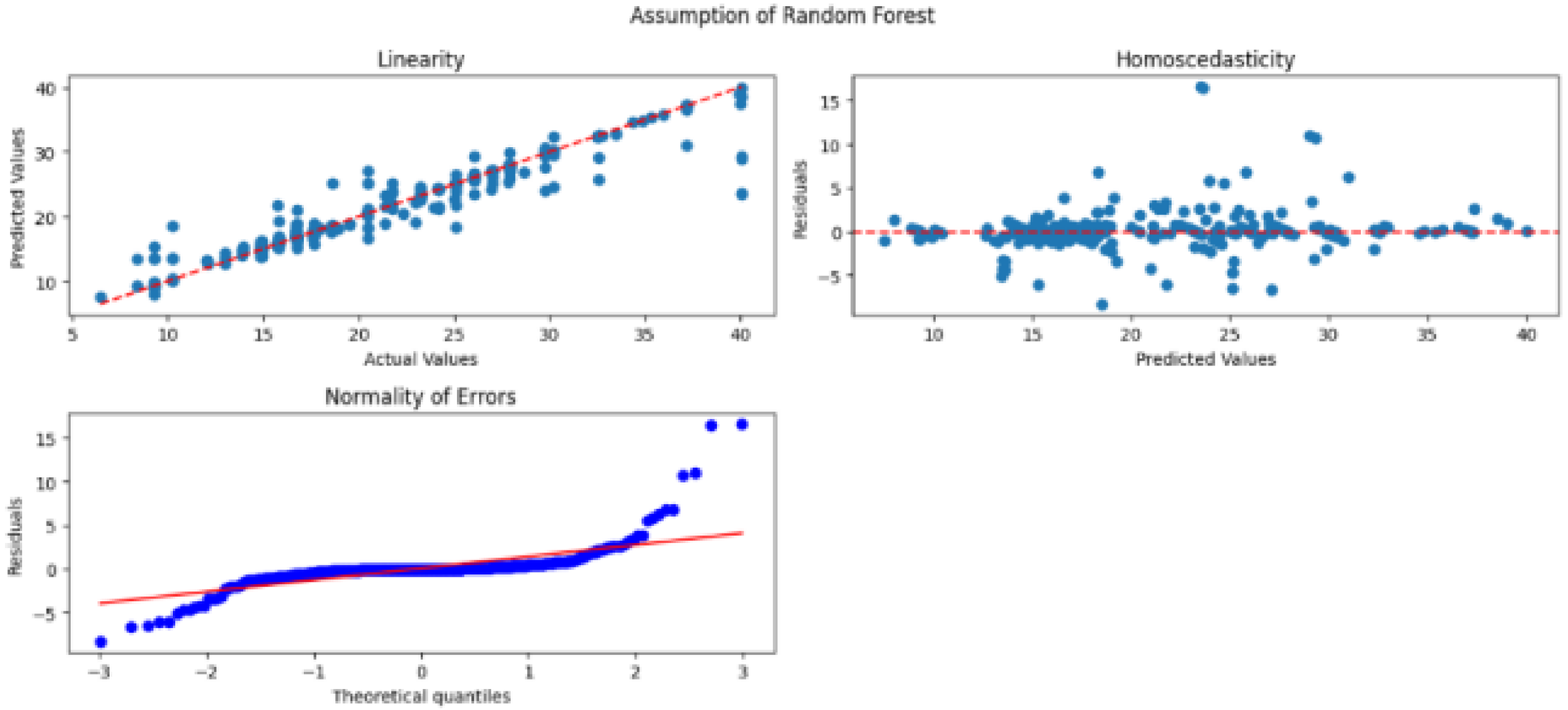
Model Diagnostic

Linear Regression



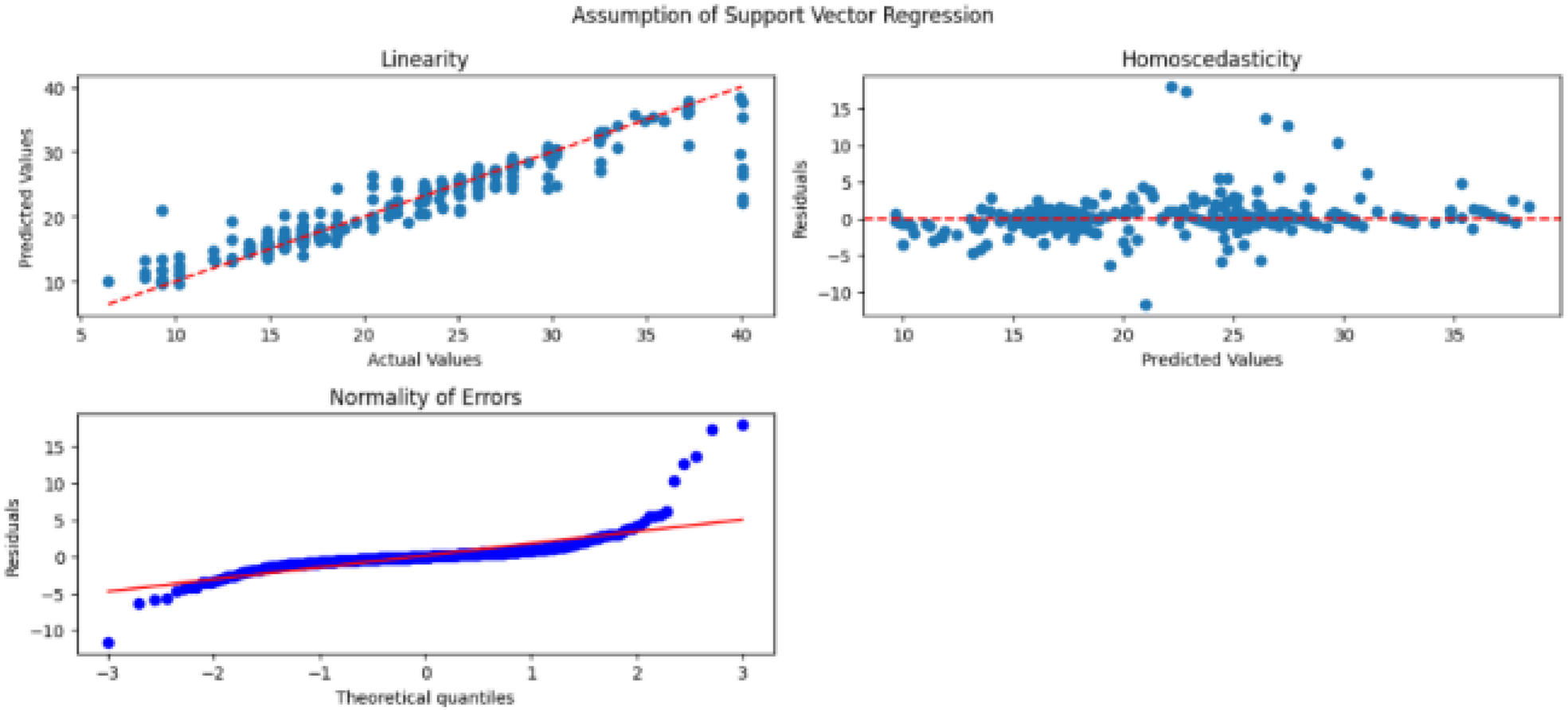
Model Diagnostic

Random Forest Regression



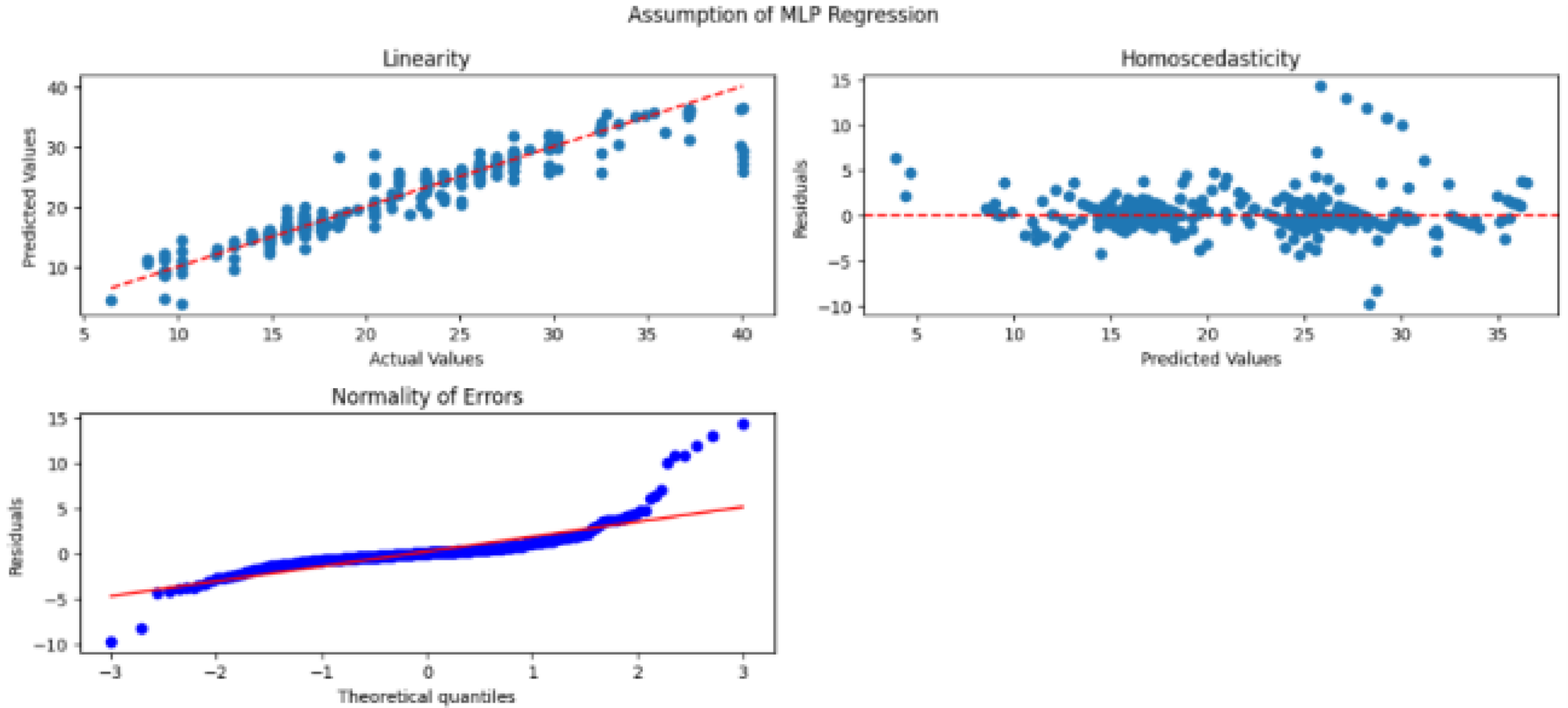
Model Diagnostic

Support Vector Regression



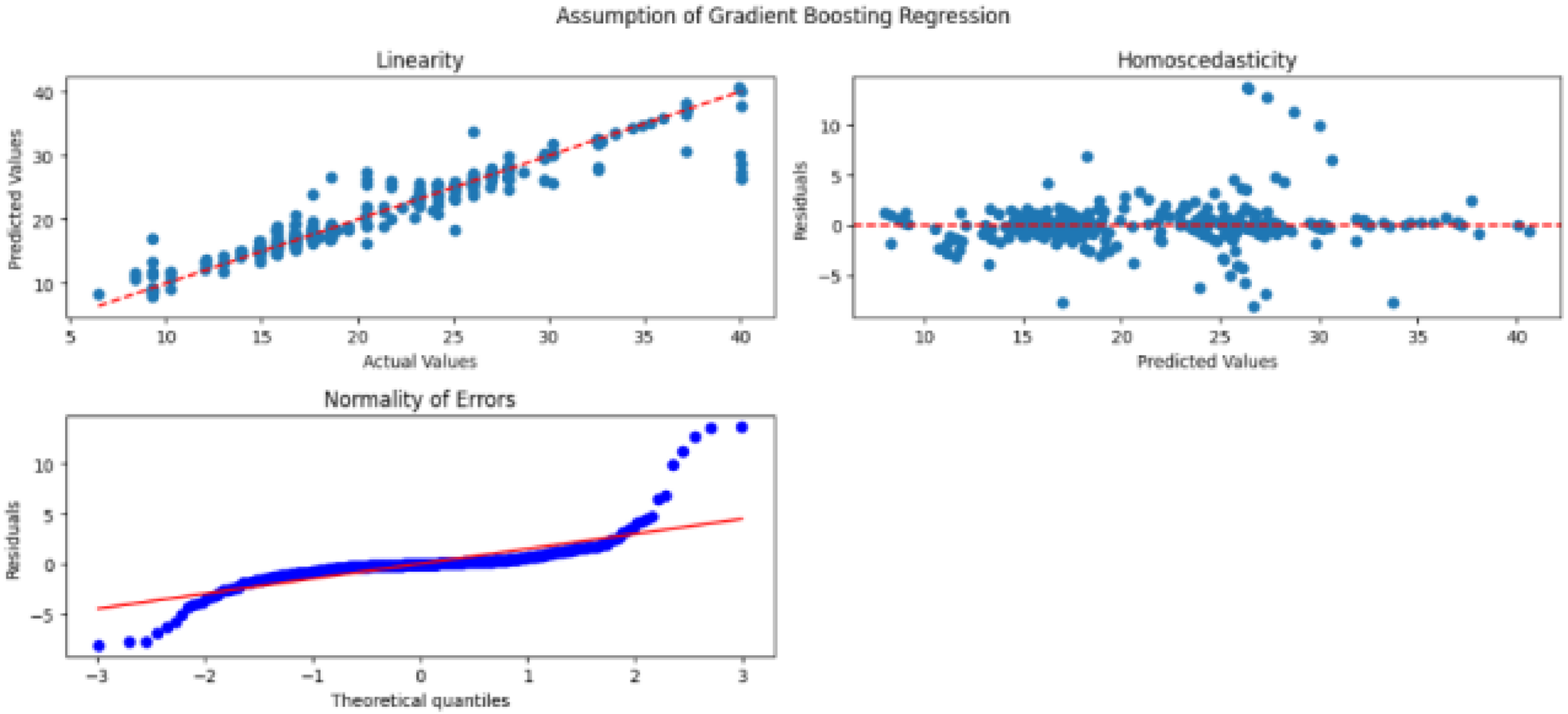
Model Diagnostic

Mulitlayer Perception Regression



Model Diagnostic

Gradient Boosting Regression



THANK YOU

End Slide