AI-DRIVEN EXPLORATION AND PREDICTION OF COMPANY REGISTRATION TRENDS WITH REGISTRAR OF COMPANIES(RoC)

PREPARED BY

M.MAHALAKSHMI,
510521205022,
BHARATHIDASAN
ENGINEERING COLLEGE,
PHASE3 PROJECT.

INTRODUCTION

The objective of this project is to harness the power of artificial intelligence (AI) to analyze and predict trends in company registrations using data obtained from the Register of Companies. Company registration trends can provide valuable insights into economic growth, industry dynamics, and business development, making this project highly relevant for policy-makers, investors, and business stakeholders.

Remember that this process can be iterative, and the specific techniques and models used may vary depending on the dataset and the objectives of your analysis. Effective data preprocessing is crucial for building accurate and reliable AI-driven models for exploring and predicting company registration trends

GIVEN DATASET

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	CORPORAT	COMPANY_	COMPANY_	COMPANY_	COMPANY_	COMPANY_	DATE_OF_R	REGISTERED	AUTHORIZE	PAIDUP_CA	INDUSTRIAI	PRINCIPAL_	REGISTERED	REGISTRAR_	EMAIL_ADD	LATEST_YEA	LATEST_YEAR_FINANCIAL_	
2	F00643	HOCHTIEFF	NAEF	NA	NA	NA	1/12/1961	Tamil Nadu	0	0	NA	Agriculture	AMBLE SIDE	ROC微ELHI	NA	NA	NA	
3	F00721	SUMITOMC	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	FLAT NO. 6,	ROC微ELHI	shuchi.chu	NA	NA	
4	F00892	SRILANKAN	ACTV	NA	NA	NA	1/3/1982	Tamil Nadu	0	0	NA	Agriculture	SRILANKAN	ROC微ELHI	shree16us@	NA	NA	
5	F01208	CALTEX IND	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	GOLD CRES	ROC微ELHI	NA	NA	NA	
6	F01218	GE HEALTH(	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	FF-3 Palani	ROC微ELHI	karthick999	NA	NA	
7	F01265	CAIRN ENEI	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	WELLINGT(	ROC微ELHI	neerja.shar	NA	NA	
8	F01269	TORIELLI S.I	ACTV	NA	NA	NA	5/9/1995	Tamil Nadu	0	0	NA	Agriculture	6, Mangaya	ROC微ELHI	chennai@t	NA	NA	
9	F01311	HARDY EXP	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	5TH FLOOR,	ROC微ELHI	venkatesh.	NA	NA	
10	F01314	HOCHTIOF	ACTV	NA	NA	NA	11/4/1996	Tamil Nadu	0	0	NA	Agriculture	NEW NO.8(	ROC微ELHI	kumar@int	NA	NA	
11	F01412	EPSON SIN(	ACTV	NA	NA	NA	25-04-1997	Tamil Nadu	0	0	NA	Agriculture	7C CEATUR	ROC微ELHI	NA	NA	NA	
12	F01426	CARGOLUX	ACTV	NA	NA	NA	11/6/1997	Tamil Nadu	0	0	NA	Agriculture	OFFICE NO	ROC微ELHI	NA	NA	NA	
13	F01468	CHO HEUN(	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	129, MANP	ROC微ELHI	chowelacc	NA	NA	
14	F01543	NYCOMED	ACTV	NA	NA	NA	27-10-1998	Tamil Nadu	0	0	NA	Agriculture	A D 46 1ST	ROC微ELHI	NA	NA	NA	
15	F01544	CHERRINGT	ACTV	NA	NA	NA	1/5/2000	Tamil Nadu	0	0	NA	Agriculture	10HADDOW	ROC微ELHI	NA	NA	NA	
16	F01563	SHIMADZU	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	FIRST FLOO	ROC微ELHI	kousik@vsr	NA	NA	
17	F01565	CORK INTEI	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	ARJAY APE)	ROC微ELHI	NA	NA	NA	
18	F01566	ERBIS ENG(	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	39,2nd Mai	ROC微ELHI	NA	NA	NA	
19	F01589	RALF SCHNI	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	FLAT C, 'SAI	ROC微ELHI	NA	NA	NA	
20	F01593	MITRAJAYA	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	OLD NO 14(	ROC微ELHI	NA	NA	NA	
21	F01618	HEAT AND (	ACTV	NA	NA	NA	13-07-1999	Tamil Nadu	0	0	NA	Agriculture	A40 OLD N(	ROC微ELHI	ncrajagopal	NA	NA	
22	F01628	DIREX SYST	ACTV	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	F-1, FIRST F	ROC微ELHI	direx@vsnl	NA	NA	
23	F01641	NMB-MINE	NAEF	NA	NA	NA	NA	Tamil Nadu	0	0	NA	Agriculture	Level - 2 Re	ROC微ELHI	stsogawa@	NA	NA	

Data_Gov_Tamil_Nadu

LOADING AND PREPROCESSING DATASET STEPS

Loading and preprocessing a dataset for AI-driven exploration and production of company registration trends involving the Register of Companies typically involves the following steps:

1. Data Collection:Gather data from the Register of Companies, which includes information about newly registered companies, their locations, industries, registration dates, and other relevant details.
2. Data Storage:Store the collected data in a secure and scalable data storage solution, such as a database or cloud-based storage (e.g., AWS S3, Google Cloud Storage).
3. Data Cleaning: Clean the dataset to handle missing values, data inconsistencies, and errors. This ensures the dataset's accuracy and reliability.
4. Data Integration: If you have data from multiple sources or historical data, integrate and harmonize it for a comprehensive view of company registration trends.

5. Feature Extraction: Extract relevant features from the data, such as registration dates, company types, and geographic regions, which are essential for trend analysis.

6. Data Normalization: Normalize the data, especially numerical values, to ensure consistent scales and facilitate AI model training.

7. Data Splitting: Divide the dataset into training, validation, and test sets to develop and evaluate AI models effectively.

8. Data Preprocessing: Further preprocess the data, such as converting date formats, encoding categorical variables, and handling text data if necessary.

9. Labeling:Define the target variable or labels for your AI model, such as predicting the number of new registrations within a specific time frame.

10. Model Training:Train AI models, such as machine learning or deep learning models, using the preprocessed data to predict company registration trends.

11. Model Evaluation: Assess the performance of the trained AI models using appropriate evaluation metrics to ensure their effectiveness in trend prediction.

12. Deployment: Deploy the AI-driven models in a production environment to monitor and analyze company registration trends in real time. This deployment may involve setting up APIs, dashboards, or automated reports.

The specific tools and techniques used at each step will depend on your project's requirements and the technologies you choose. Effective data loading and preprocessing are crucial for developing AI models that can provide valuable insights into company registration trends, potentially aiding in business strategy and decision-making.

LOADING AND PREPROCESSING DATASET PROGRAM

Pandas and Numpy have been used for Data Manipulation and numerical Calculations
Matplotlib and Seaborn have been used for Data visualizations.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#to ignore warnings
import warnings
warnings.filterwarnings('ignore')
data = pd.read_csv("used_cars.csv")
data.head()
```

OUTPUT:

	S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.60	998.0	58.16	5.0	NaN	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	NaN	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.20	1199.0	88.70	5.0	8.61	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	NaN	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968.0	140.80	5.0	NaN	17.74

data.info

OUTPUT:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7253 entries, 0 to 7252
Data columns (total 14 columns):
 #    Column                Non-Null Count  Dtype
---   ------                --------------  -----
 0    S.No.                 7253 non-null   int64
 1    Name                  7253 non-null   object
 2    Location              7253 non-null   object
 3    Year                  7253 non-null   int64
 4    Kilometers_Driven     7253 non-null   int64
 5    Fuel_Type             7253 non-null   object
 6    Transmission          7253 non-null   object
 7    Owner_Type            7253 non-null   object
 8    Mileage               7251 non-null   float64
 9    Engine                7207 non-null   float64
 10   Power                 7078 non-null   float64
 11   Seats                 7200 non-null   float64
 12   New_price             1006 non-null   float64
 13   Price                 6019 non-null   float64
dtypes: float64(6), int64(3), object(5)
memory usage: 793.4+ KB
```

```
cat_cols=data.select_dtypes(include=['object']).columns
num_cols = data.select_dtypes(include=np.number).columns.tolist()
print("Categorical Variables:")
print(cat_cols)
print("Numerical Variables:")
print(num_cols)
```

OUTPUT:

```
Categorical Variables:
Index(['Name', 'Location', 'Fuel_Type', 'Transmission', 'Owner_Type', 'Brand',
       'Model'],
      dtype='object')
Numerical Variables:
['Year', 'Kilometers_Driven', 'Mileage', 'Engine', 'Power', 'Seats', 'New_price', 'Price', 'Car_Age']
```

```
from datetime import date
date.today().year
data['Car_Age']=date.today().year-data['Year']
data.head()
```

OUTPUT:

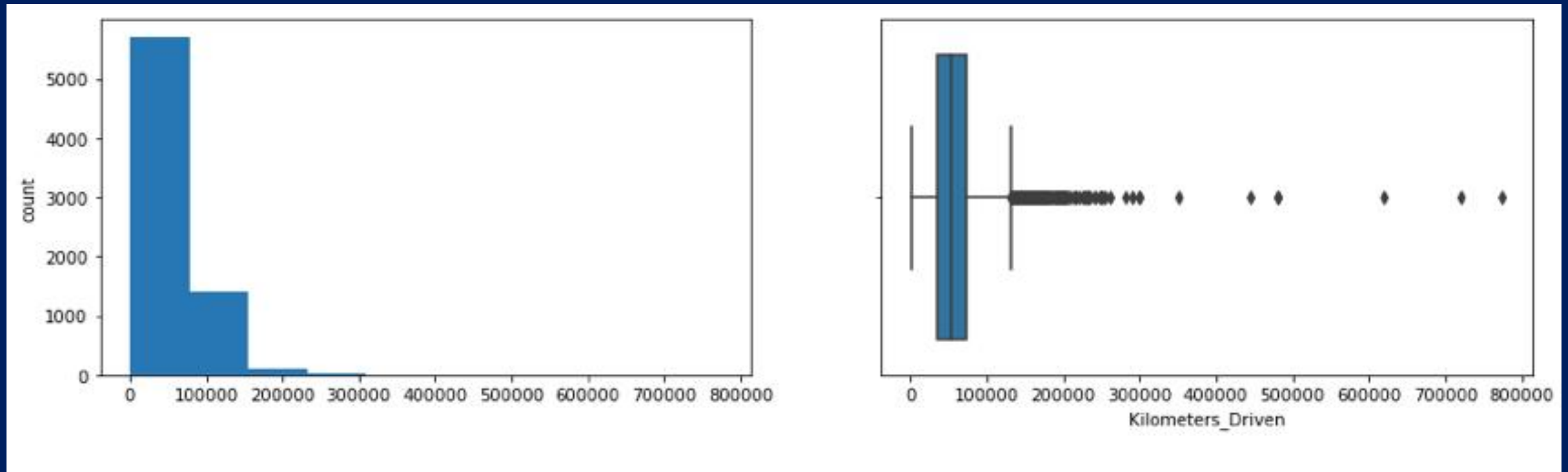
Out[32]:

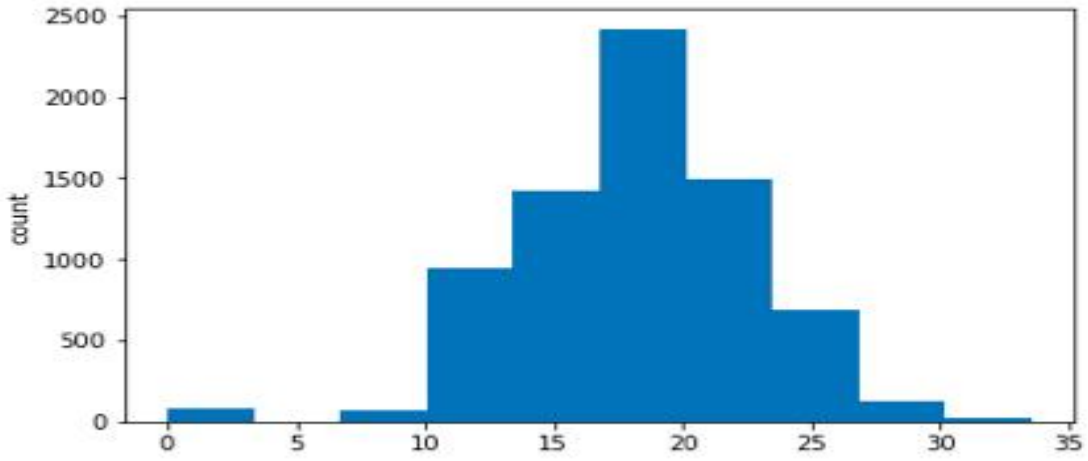
	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_price	Price	Car_Age
0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.60	998.0	58.16	5.0	NaN	1.75	12
1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	NaN	12.50	7
2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.20	1199.0	88.70	5.0	8.61	4.50	11
3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	NaN	6.00	10
4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968.0	140.80	5.0	NaN	17.74	9

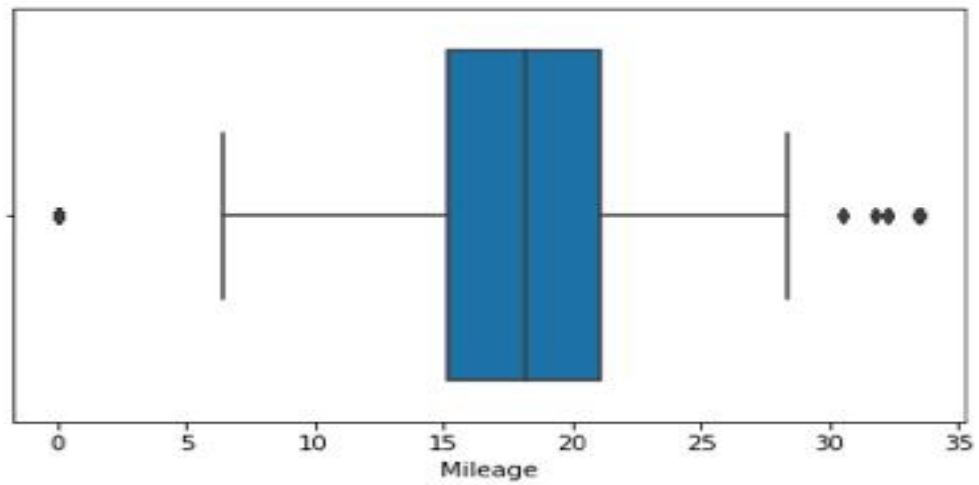
DATA VISUALIZATION

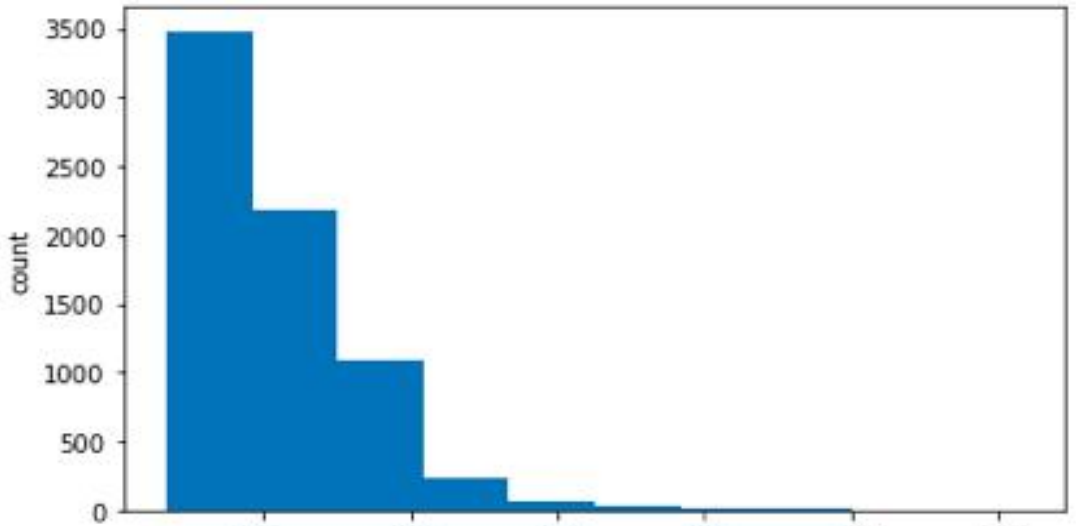
```
for col in num_cols:
    print(col)
    print('Skew :', round(data[col].skew(), 2))
    plt.figure(figsize = (15, 4))
    plt.subplot(1, 2, 1)
    data[col].hist(grid=False)
    plt.ylabel('count')
    plt.subplot(1, 2, 2)
    sns.boxplot(x=data[col])
    plt.show()
```

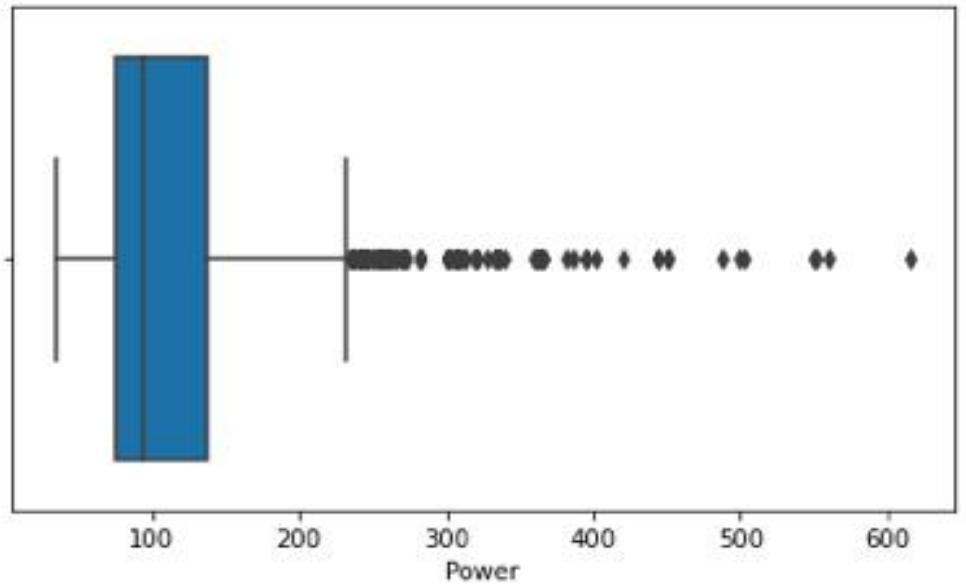
OUTPUT:




count
2500
2000
1500
1000
500
0
0
5
10
15
20
25
30
35

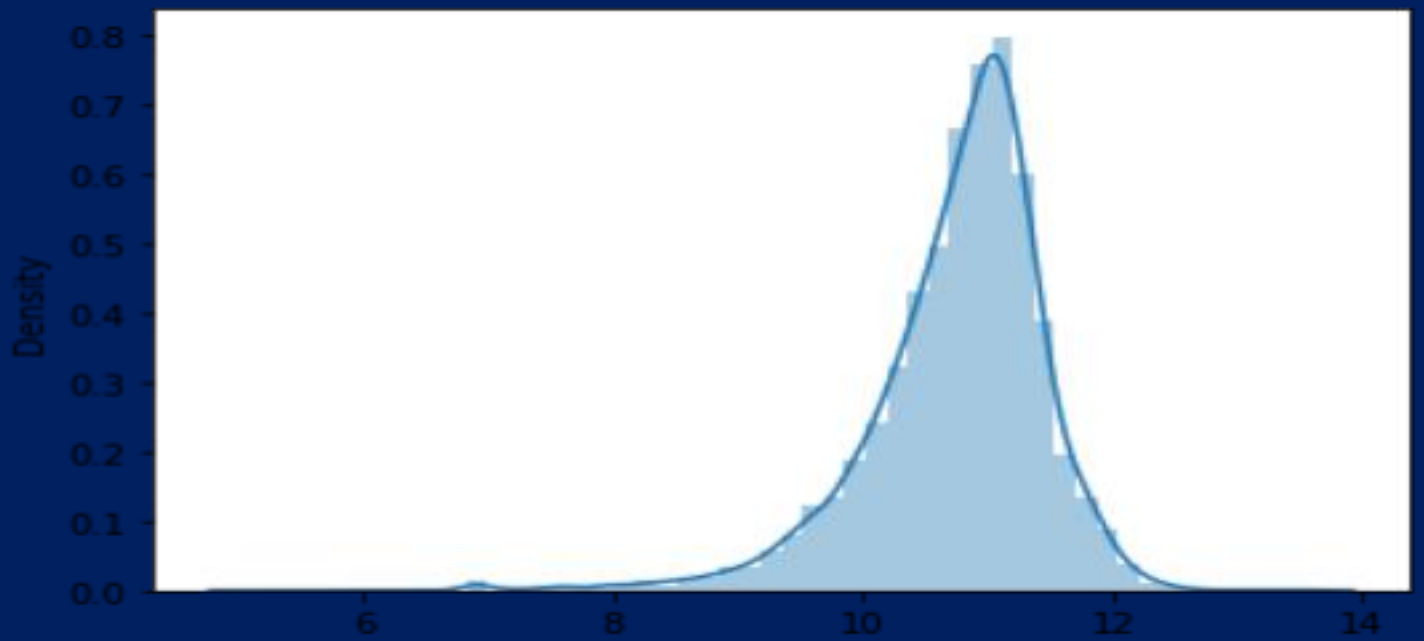

0
5
10
15
20
25
30
35
Mileage


count
3500
3000
2500
2000
1500
1000
500
0
100
200
300
400
500
600


100
200
300
400
500
600
Power

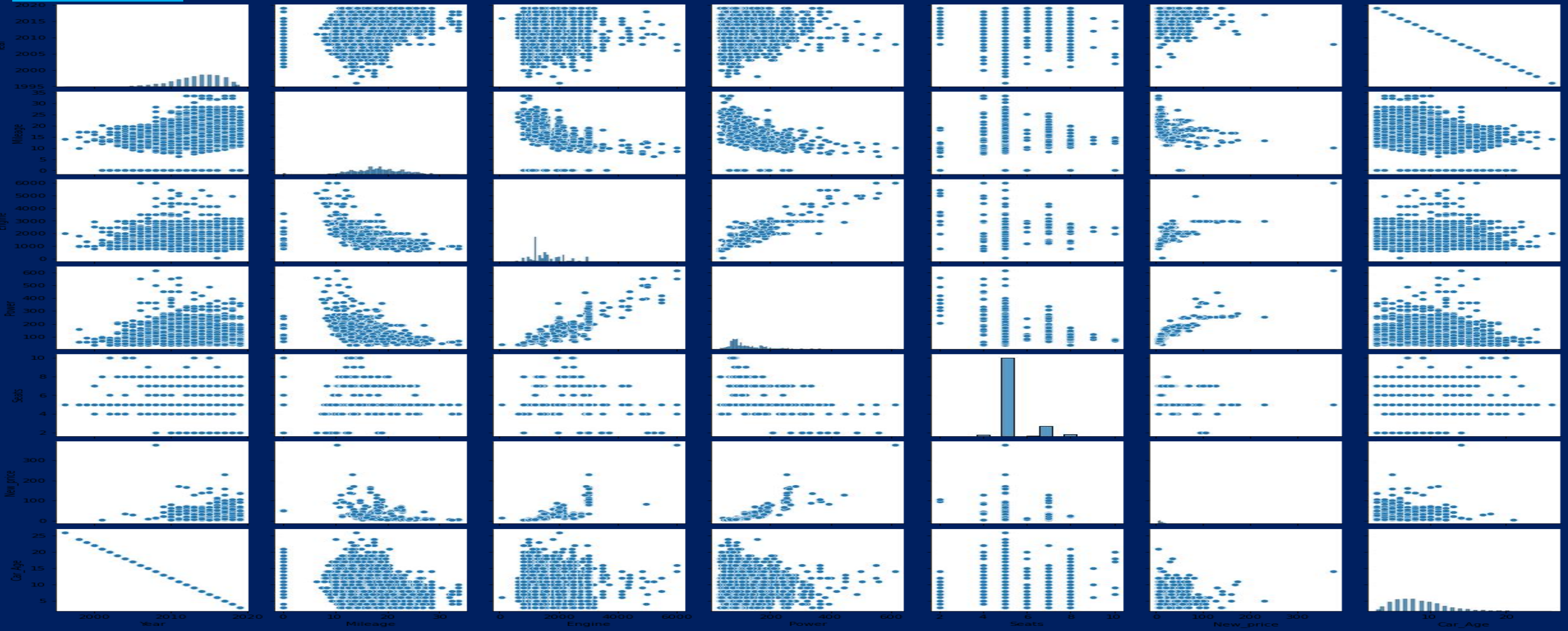
```
# Function for log transformation of the column
def log_transform(data,col):
    for colname in col:
        if (data[colname] == 1.0).all():
            data[colname + '_log'] = np.log(data[colname]+1)
        else:
            data[colname + '_log'] = np.log(data[colname])
    data.info()
log_transform(data,['Kilometers_Driven','Price'])
#Log transformation of the feature 'Kilometers_Driven'
sns.distplot(data["Kilometers_Driven_log"], axlabel="Kilometers_Driven_log");
```

OUTPUT:



```
plt.figure(figsize=(13,17))
sns.pairplot(data=data.drop(['Kilometers_Driven','Price'],axis=1))
plt.show()
```

OUTPUT:



CONCLUSION

Overall, the loading and preprocessing of data is the crucial foundation upon which the success of AI-driven analysis of company registration trends rests. Ensuring data accuracy and suitability for analysis is essential for producing actionable insights and informed decision-making in the field of business registration trend analysis.

This is a simplified example, and you may need to adapt it to your specific dataset and analysis requirements. Additionally, you can explore more advanced machine learning models and techniques to enhance your trend prediction.

THANKS