INTERNSHIP PROGRAM

Name: Muhammad Ahmed Almursii

Presentation TOPICS

1st Topic

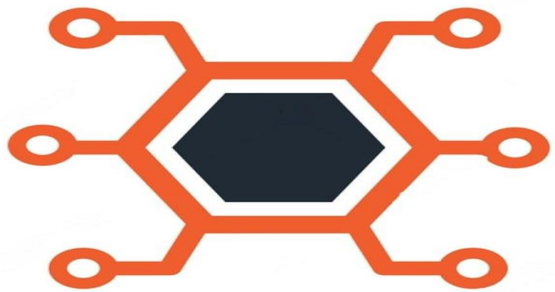
LLM and its application

Introduction to Open Source LLMs

Open source LLMs (large language models) are a type of artificial intelligence (AI) that can be used for a variety of tasks in data science. They are trained on massive datasets of text and code, and can be used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way

2nd
Topic

UNSUPERVISED LEARNING



Unsupervised Learning

- A cluster is represented by a single point, known as centroid (or cluster center) of the cluster.

Types of clustering:
- K-means, Fuzzy C, hierarchical

Similarity functions:
- Euclidean distance, Manhattan distance

Stopping criteria:
- SSD

1st Project

World Population Analysis

Business Understanding

Business Understanding

This project analyzed population and land data for various countries and continents over multiple years. It aimed to understand population trends, population growth, and land availability. The findings provide valuable insights into global population dynamics and the distribution of land resources.

- By examining key indicators such as population counts, growth rates, and land area
- we can better understand the dynamics of global population distribution and resource utilization.

Data Understanding

Data Understanding

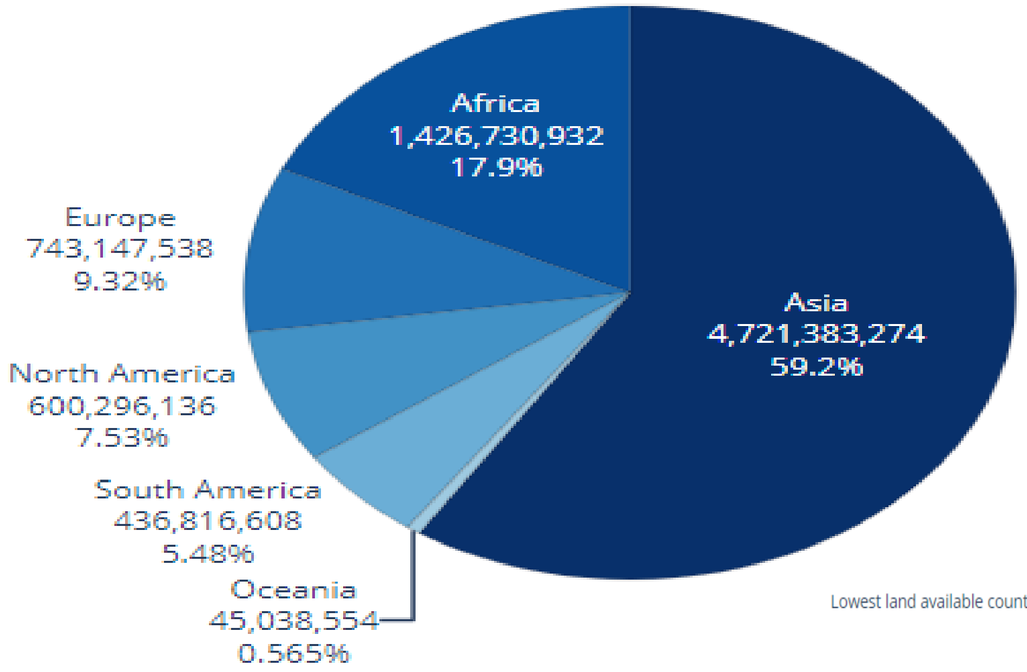
- We have 15 features.
- We have data of population of Years from 1970 till 2022 as following (19970, 1980, 1990, 2000, 2010, 2015, 2020, 2022)
- Our data is 234 records and data are clear having no null values .
- Density (per km²): The population density, calculated as the population per square kilometer.
- Growth Rate: The rate of population growth.
- World Population Percentage: The percentage of the world's population represented by the country/territory

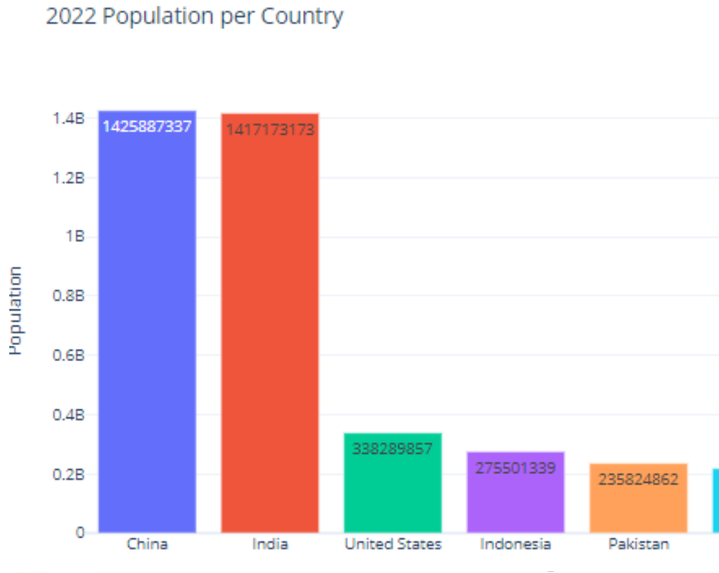
Countries per Continent ->>

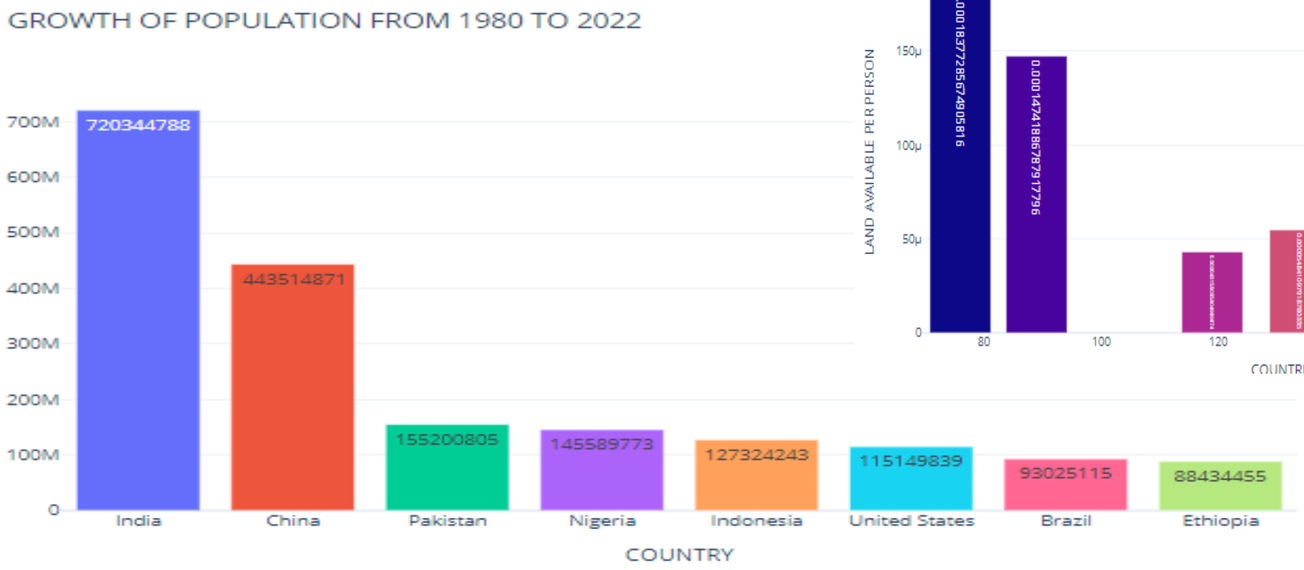
- Africa 57
- Asia 50
- Europe 50
- North America 40
- Oceania 23
- South America 14

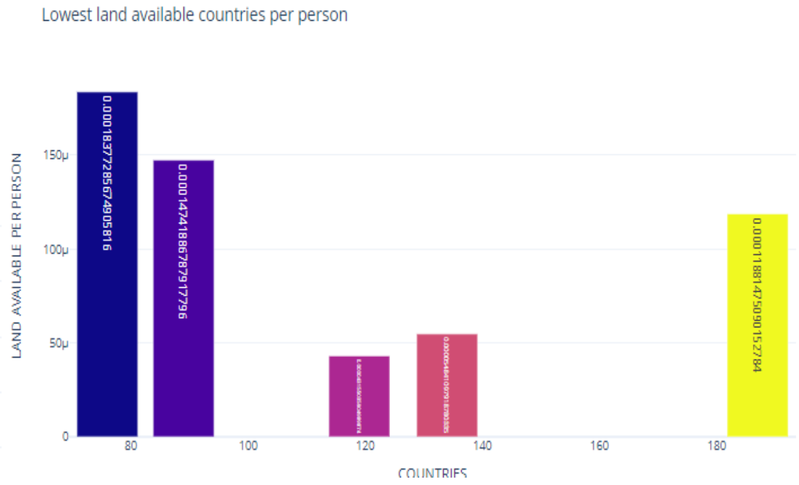
EDA Charts

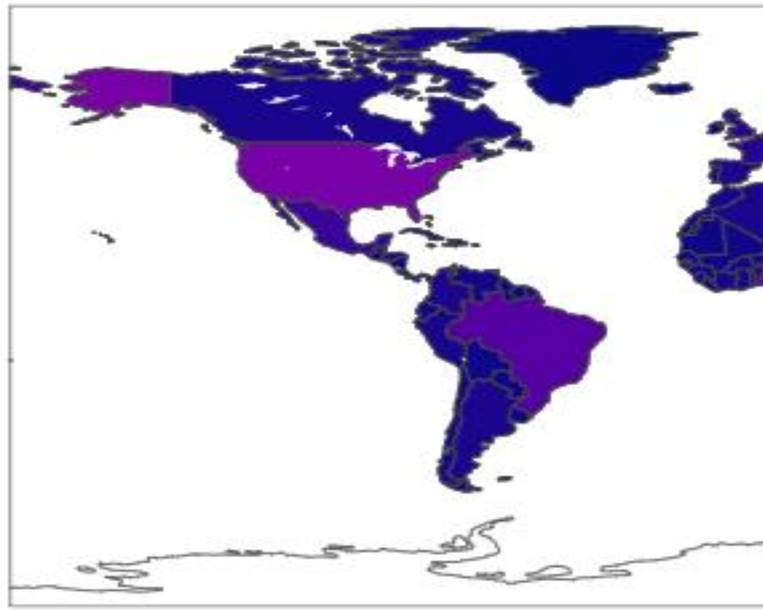
EDA Charts


Africa
1,426,730,932
17.9%
Europe
743,147,538
9.32%
North America
600,296,136
7.53%
South America
436,816,608
5.48%
Oceania
45,038,554
0.565%
Asia
4,721,383,274
59.2%


2022 Population per Country
Population
1.4B
1.2B
1B
0.8B
0.6B
0.4B
0.2B
0
1425887337
1417173173
338289857
275501339
235824862
China
India
United States
Indonesia
Pakistan


GROWTH OF POPULATION FROM 1980 TO 2022
700M
600M
500M
400M
300M
200M
100M
0
720344788
443514871
155200805
145589773
127324243
115149839
93025115
88434455
India
China
Pakistan
Nigeria
Indonesia
United States
Brazil
Ethiopia
COUNTRY


Lowest land available countries per person
LAND AVAILABLE PER PERSON
150µ
100µ
50µ
0
0.00018377285749581 6
0.00014741886787917796
0.00011881475090152784
80
100
120
140
160
180
COUNTRIES


2022 Population

Conclusion & Recommendation

Conclusion & Recommendation

- Population Growth: The population of several countries and continents has experienced significant growth over the years, with Asia and Africa showing the highest population increases. This growth poses challenges and opportunities in areas such as infrastructure, healthcare, and resource management.

- Regional Variations: Population distribution varies across continents, with Asia having the largest population overall. The growth rates and patterns differ among continents, influenced by factors such as fertility rates, migration, and economic development.

- Population by Country: Countries like China and India have the highest populations, which impact various aspects of their economies and societies. Understanding population dynamics in these countries is crucial for effective policy-making and resource allocation

Recommendation

- Implement comprehensive family planning programs to ensure access to reproductive healthcare, education, and contraceptive methods.

2nd Project

Loan Approval Prediction

Business Understanding

Business Understanding

The Loan Approval Prediction project aims to develop a predictive model to assist in the loan approval decision-making process. By analyzing various features of loan applicants, such as credit history, income, education, and property area,

Other Influential Factors: Features such as property area, marital status, education level, and income also play a significant role in loan approval decisions.

Loan Approval Objective:

- The main objective is to accurately predict whether a loan application will be approved or not based on the given features. This prediction model can help streamline the loan approval process and make informed decisions.

Data Understanding

Data Understanding

- Column Identification: The dataset consists of the following columns: Loan_ID, Gender, Married, Dependents, Education, Self_Employed, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area, and Loan_Status.

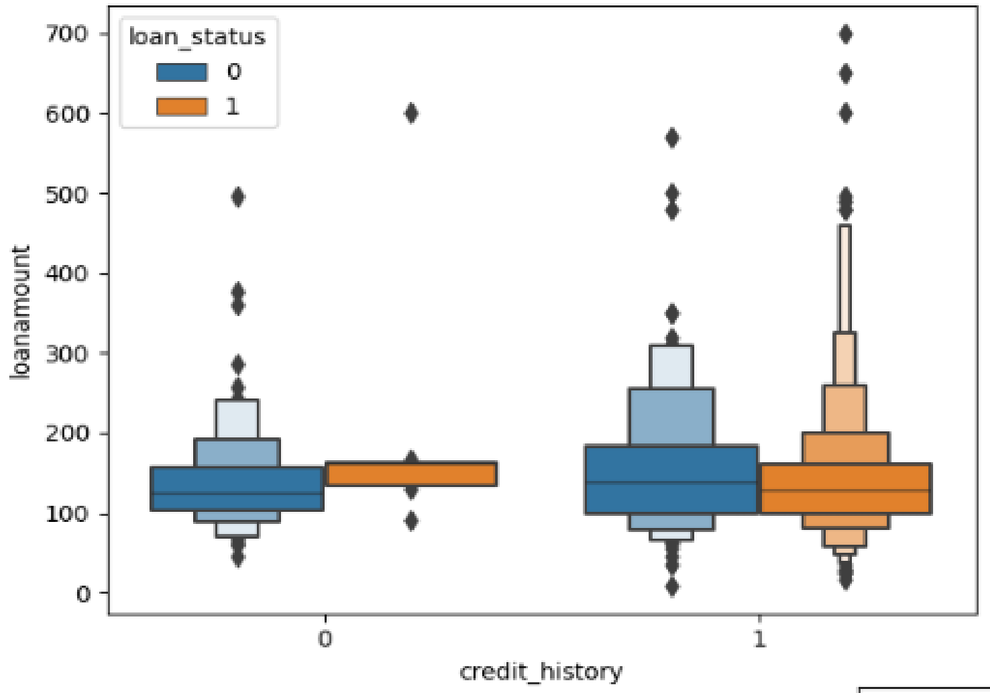
- Data Exploration: The dataset is explored by analyzing the structure, content, and distribution of each column. This includes checking the data types, unique values, and missing values in the dataset.

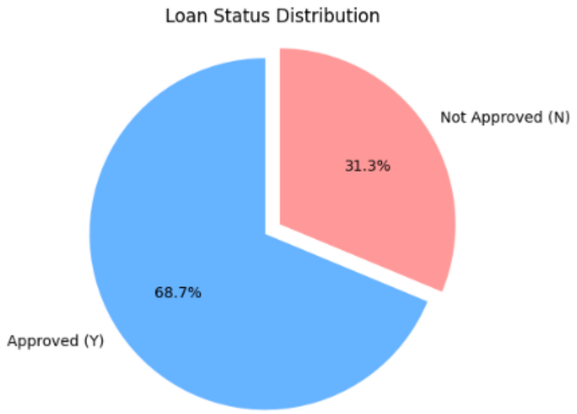
- Loan Status Distribution: The distribution of the Loan_Status column is examined to understand the proportion of approved and not approved loans. This helps in understanding the class imbalance and the overall loan approval rate in the dataset.

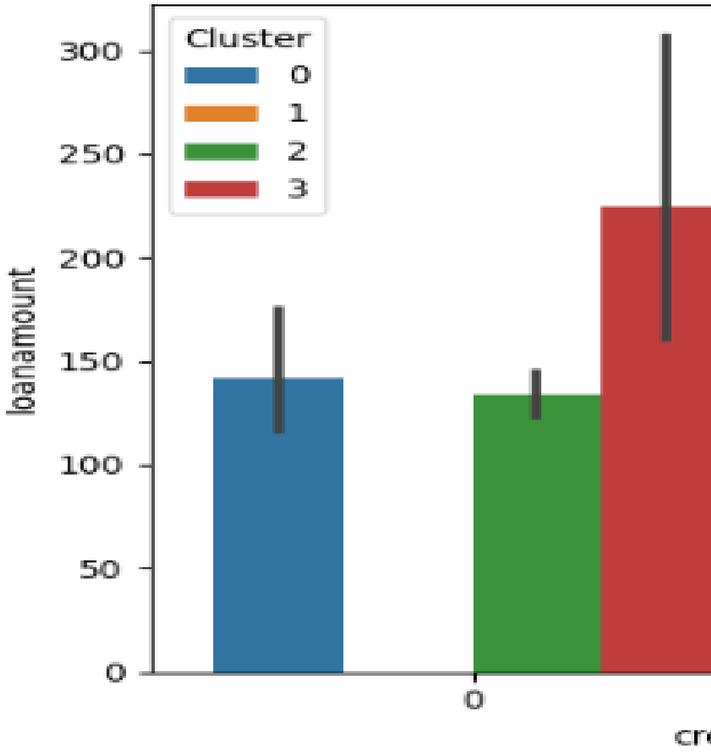
- Missing values are handled by imputing them with appropriate values, and categorical variables are encoded for further analysis.

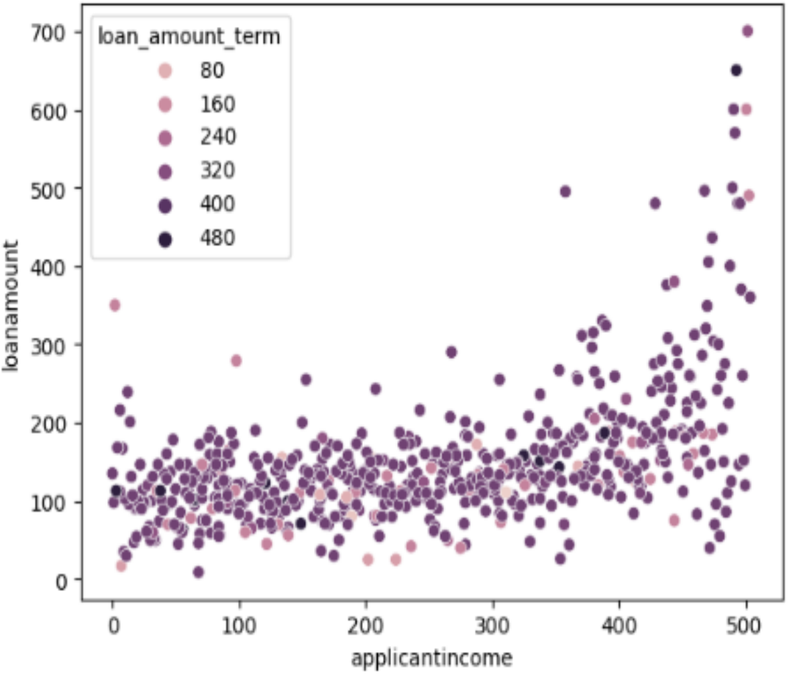
EDA Charts

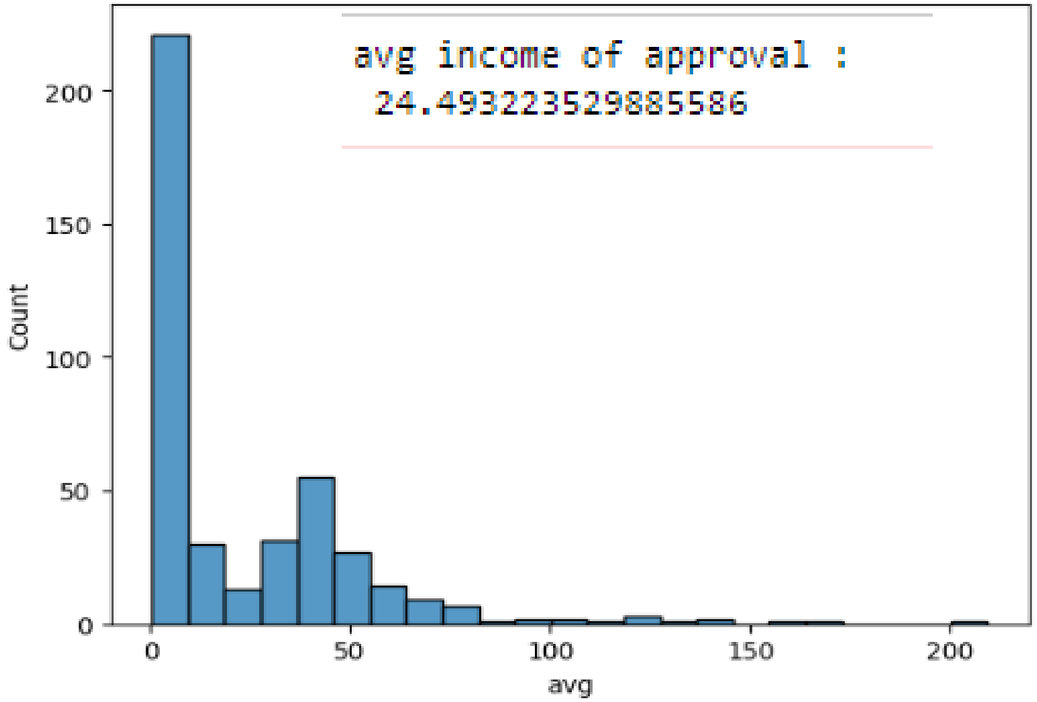
EDA Charts











	applicantincome	coappli
applicantincome	1.000000	
coapplicantincome	-0.205261	
loanamount	0.497118	
loan_amount_term	-0.026696	

ML Models (If Available)

ML Models (If Available)

Adjusted xgboost

```
from xgboost import XGBClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score


XGBclassifier = XGBClassifier()

# Define the parameter grid to search
param_grid = {
    'n_estimators': [100, 200, 300],  # Number of trees
    'max_depth': [3, 4, 5],  # Maximum depth of each tree
    'learning_rate': [0.1, 0.01, 0.001]  # Learning rate
}
```

```
# Get the best parameters and best score
best_params = grid_search.best_params_
best_score = grid_search.best_score_

# Create a new XGBoost classifier with the best parameters
XGBclassifier_best = XGBClassifier(**best_params)

# Fit the classifier to the training data
XGBclassifier_best.fit(X_train, y_train)

# Make predictions on the test data
y_pred = XGBclassifier_best.predict(X_test)
```

```
accuracy
```

```
0.827027027027027
```

Conclusion & Recommendation

Conclusion & Recommendation

- After preprocessing the data and conducting exploratory data analysis, key insights were derived. Credit history was found to be the most crucial factor in loan approval, followed by property area, marital status, and education. Other factors like income, loan amount, and employment type also played significant roles in loan approval decisions.
- Applicants in semiurban and urban areas had higher loan approval rates compared to those in rural areas.
- Married individuals and graduates had higher loan approval rates.
- Gender and self-employment status had minimal impact on loan approval.
- The models achieved varying levels of accuracy, with Logistic Regression and XGBoost Classifier performing relatively well.
- Adjustments and hyperparameter tuning were applied to improve model performance.

Recommendation

- The project's recommendations include giving priority to applicants with a good credit history, focusing on semiurban and urban areas, considering marital status and education

3nd Project

Sports Wear

Business Understanding

Business Understanding

The analysis of the dataset revealed valuable insights into the factors influencing the purchase decision and sales performance of articles. Several key features were examined, including country, article, price, discount, and promotional activities. By understanding the relationship between these features and the purchase decision, businesses can optimize their marketing strategies, pricing, and promotional efforts to drive sales and improve customer satisfaction.

Sports Wear Objective:

- Improve sales and customer satisfaction in the sports wear category. Through the analysis of various factors such as sales performance, customer preferences, pricing, promotions, and market trends, several insights and recommendations were generated to achieve the objective.

Data Understanding

(100000, 24)

Data Understanding

- Gather the dataset related to sports wear, including information about articles, prices, promotions, sales, customer demographics, and other relevant variables.

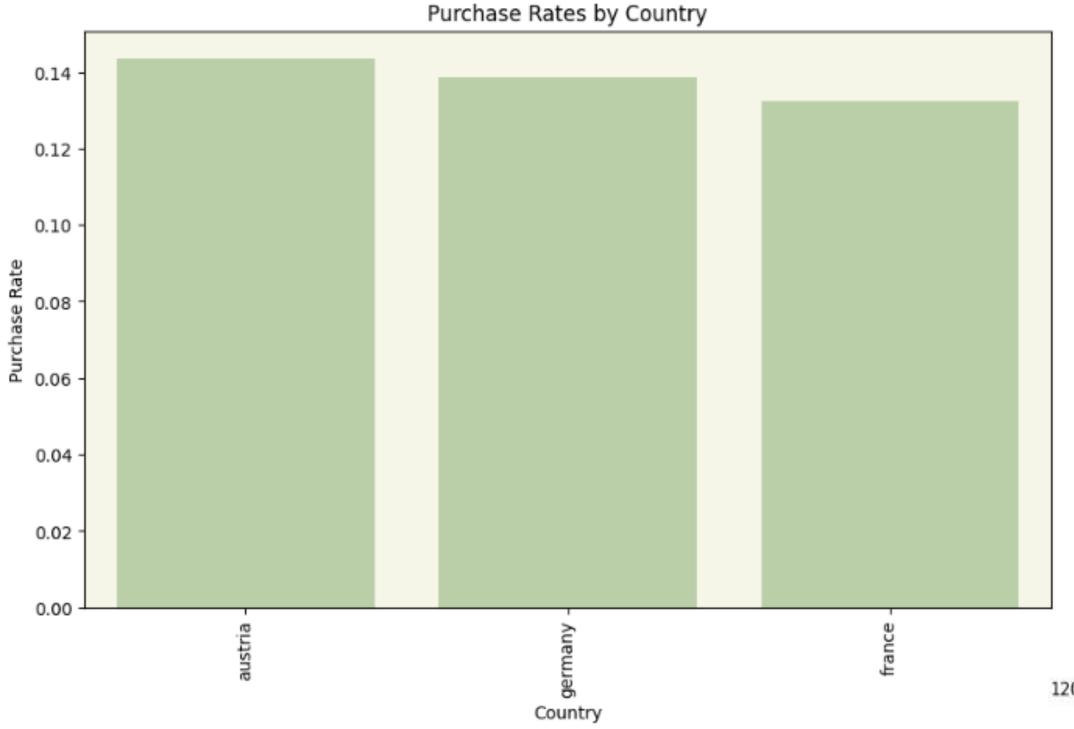
- Our dataset contains 10000 records and 24 Features .

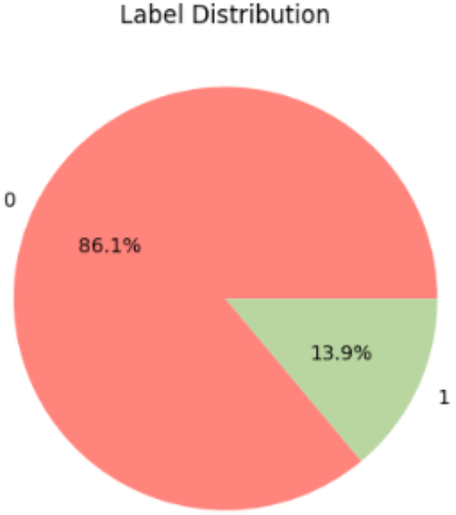
- label Distribution: The label distribution is imbalanced, with Label 0 representing approximately 86.1% of the dataset, while Label 1 represents around 13.9% of the dataset.

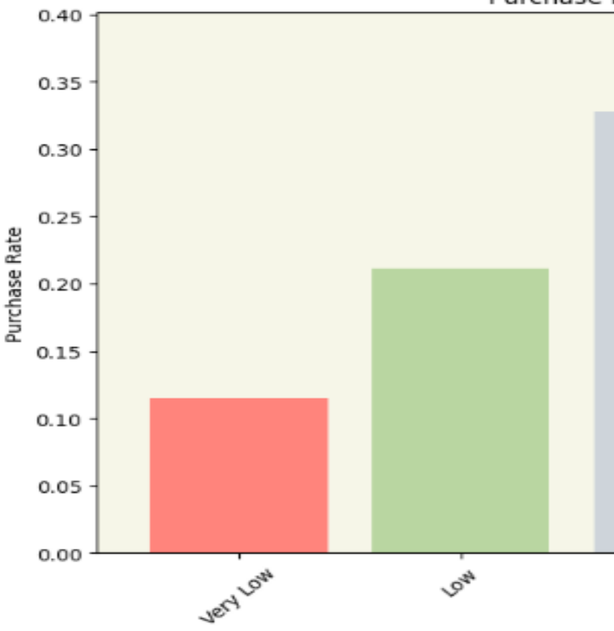
- This indicates a significant class imbalance, with Label 0 being the majority class and Label 1 being the minority class. Imbalanced data can impact the performance of machine learning models, as they may have a tendency to favor the majority class, potentially leading to biased predictions and lower accuracy for the minority class.

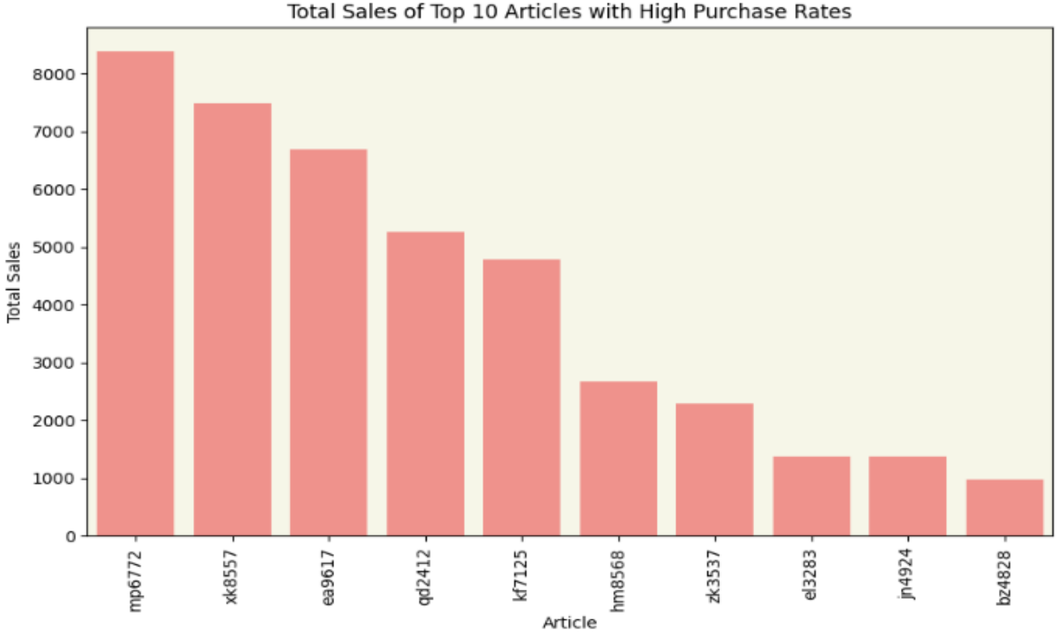
EDA Charts

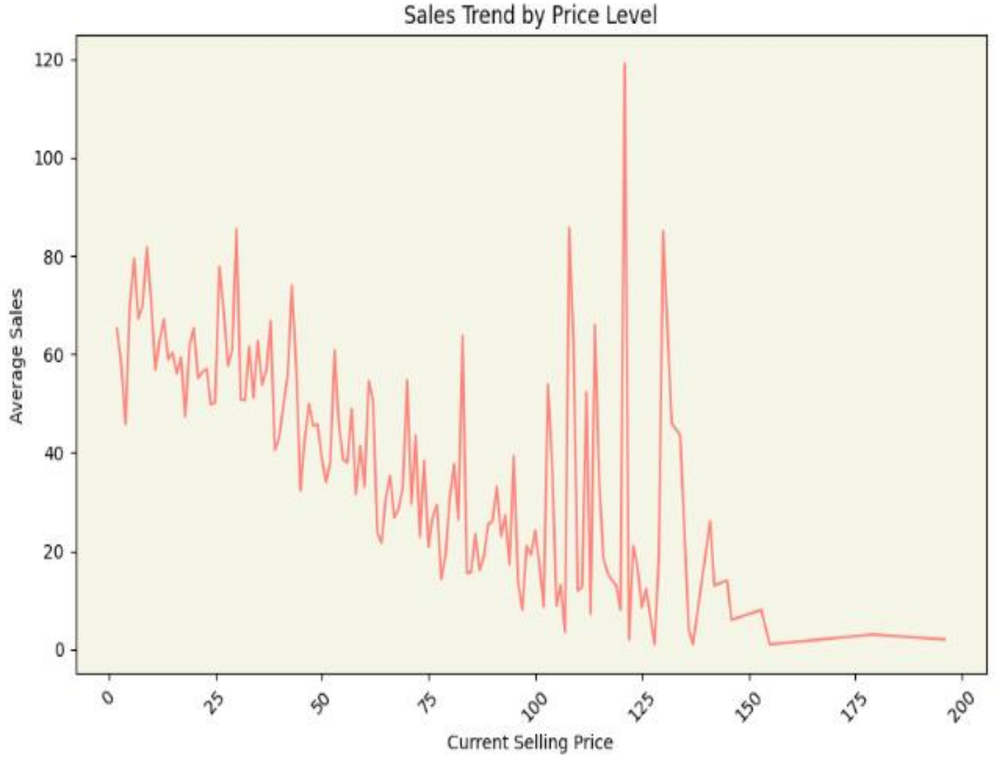
EDA Charts

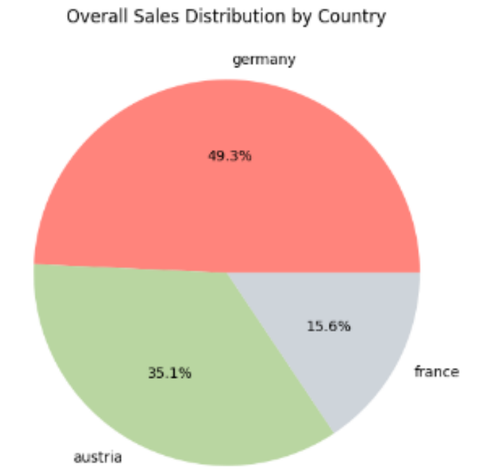

Purchase Rates by Country
Purchase Rate
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
austria
germany
france
Country


Label Distribution
0
86.1%
13.9%
1


Purchase
Purchase Rate
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Very Low
Low


Total Sales of Top 10 Articles with High Purchase Rates
Total Sales
8000
7000
6000
5000
4000
3000
2000
1000
0
mp6772
xk8557
ea9617
qd2412
kf7125
hm8568
zk3537
el3283
jn4924
bz4828
Article


Sales Trend by Price Level
Average Sales
120
100
80
60
40
20
0
0
25
50
75
100
125
150
175
200
Current Selling Price


Overall Sales Distribution by Country
germany
49.3%
15.6%
35.1%
france
austria

ML Models (If Available)

Conclusion & Recommendation

Conclusion & Recommendation

1. Country Influence: Austria, Germany, and France showed higher purchase rates and sales performance. Businesses should focus on these countries for targeted marketing and expansion opportunities.
2. Article Popularity: Certain articles, such as "XK8557" and "MP6772," showed higher purchase rates and total sales. Promoting these top-performing articles can maximize profitability.
3. Price Impact: The regular price of articles did not show a significant difference between purchased and non-purchased items. However, analyzing the current selling price revealed that lower prices and moderate discounts had a positive correlation with higher purchase rates.
4. Discount Effect: High discounts were not associated with higher purchase rates. Instead, articles with low to medium discounts showed better sales performance. Businesses should focus on providing moderate discounts that strike a balance between attractiveness and maintaining profit margins.
5. Promotional Activities: Media advertisements (promo1) and store events (promo2) had a positive impact on the purchase decision and sales. Coordinating different promotions and aligning them with price ranges can maximize their effectiveness in driving sales.

Recommendation

- Targeted Marketing: Focus on countries with higher purchase rates, such as Austria, Germany, and France, to maximize marketing efforts and expand customer base.

I would like to express my deep appreciation to Mr..Muhammad Abushamma and Mr.Yousif Yassin for their exceptional mentorship during my internship. Their guidance and support have been invaluable to my professional growth. I would also like to thank my colleagues, Essam Abdelhamid and Yousif Thabet, for their collaboration and support throughout this experience.

Fixed Solutions

I am grateful for the time you invested in us, constructive feedback on my projects, or simply taking the time to answer my questions. Your willingness to share your experiences and provide guidance, even amidst your busy schedule, has made a lasting impression on me.