WORLD HAPPINESS INDEX – DATA ANALYTICS PROJECT

IBTIDAA INTERNSHIP TASK – DATA ANALYTICS

SHUBHASHREE BHAT | 13-09-2025

PROBLEM STATEMENT

“Why are some countries happier than others, and can we predict future happiness trends?”

- **Objectives:**
  - Analyze global happiness trends
  - Identify key factors influencing happiness
  - Forecast future scores for selected countries

1. DATASET OVERVIEW

- Source: Custom-built dataset (500+ rows)
- Columns: Country, Year, Happiness Score, GDP, Social Support, Life Expectancy, Freedom, Generosity, Corruption
- Issues: Missing values, duplicates, inconsistent entries
- Cleaned dataset prepared

1. Introduction - **Problem Statement:** Analyze global happiness trends using the World Happiness Index dataset to understand factors affecting well-being and forecast future scores. ¶

- **Dataset:** Uncleaned dataset with ~500 rows, containing missing values and duplicates.
- **Skills Demonstrated:**
  1. Data Cleaning
  2. Exploratory Data Analysis (EDA)
  3. Predictive Modeling
  4. Time-Series Forecasting
  5. (Optional) Geospatial Visualization

[5]:
```
# 2. Import Libraries & Load Data
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("world_happiness_uncleaned_500.csv")
df.head()
```

[5]:

	Country	Year	Happiness_Score	GDP_per_Capita	Social_Support	Life_Expectancy	Freedom	Generosity	Corruption
0	Switzerland	2021.0	4.56	1.23	0.75	0.91	0.31	0.20	0.23
1	South Africa	2017.0	7.73	0.97	0.53	0.45	0.43	-0.07	0.10
2	Mexico	NaN	6.53	0.81	NaN	0.98	0.72	0.42	0.18
3	Germany	2020.0	5.79	1.39	0.72	0.54	0.53	0.54	0.18
4	Switzerland	2016.0	NaN	NaN	NaN	NaN	0.46	0.10	0.03

2. DATA CLEANING

- Missing values filled with column means
- Duplicates removed
- Standardized country names (Unknown, None → NaN)
- Clean dataset saved as world_happiness_cleaned.csv

2. Data Cleaning (Skill 1)

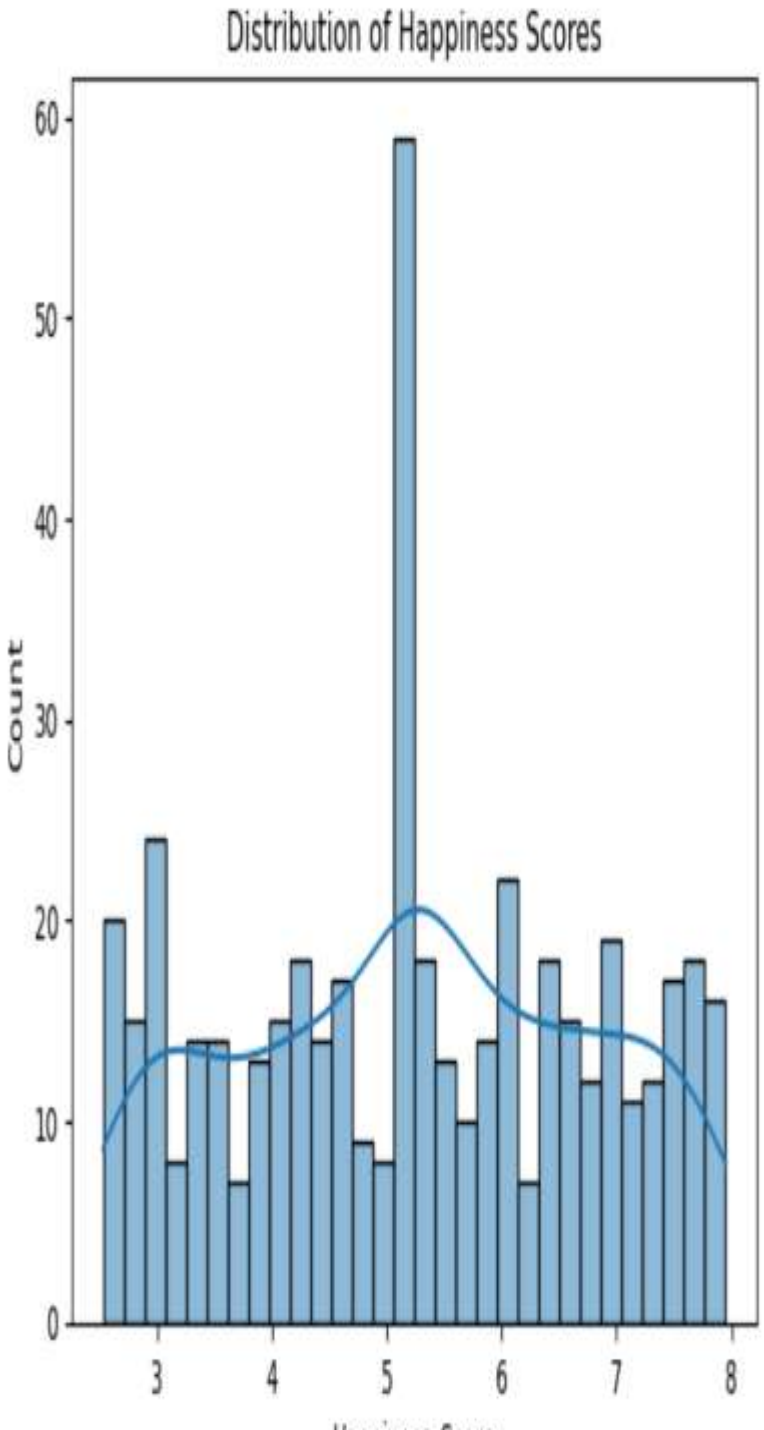
- Drop duplicates
- Handle missing values
- Fix inconsistent country names
- Save cleaned dataset

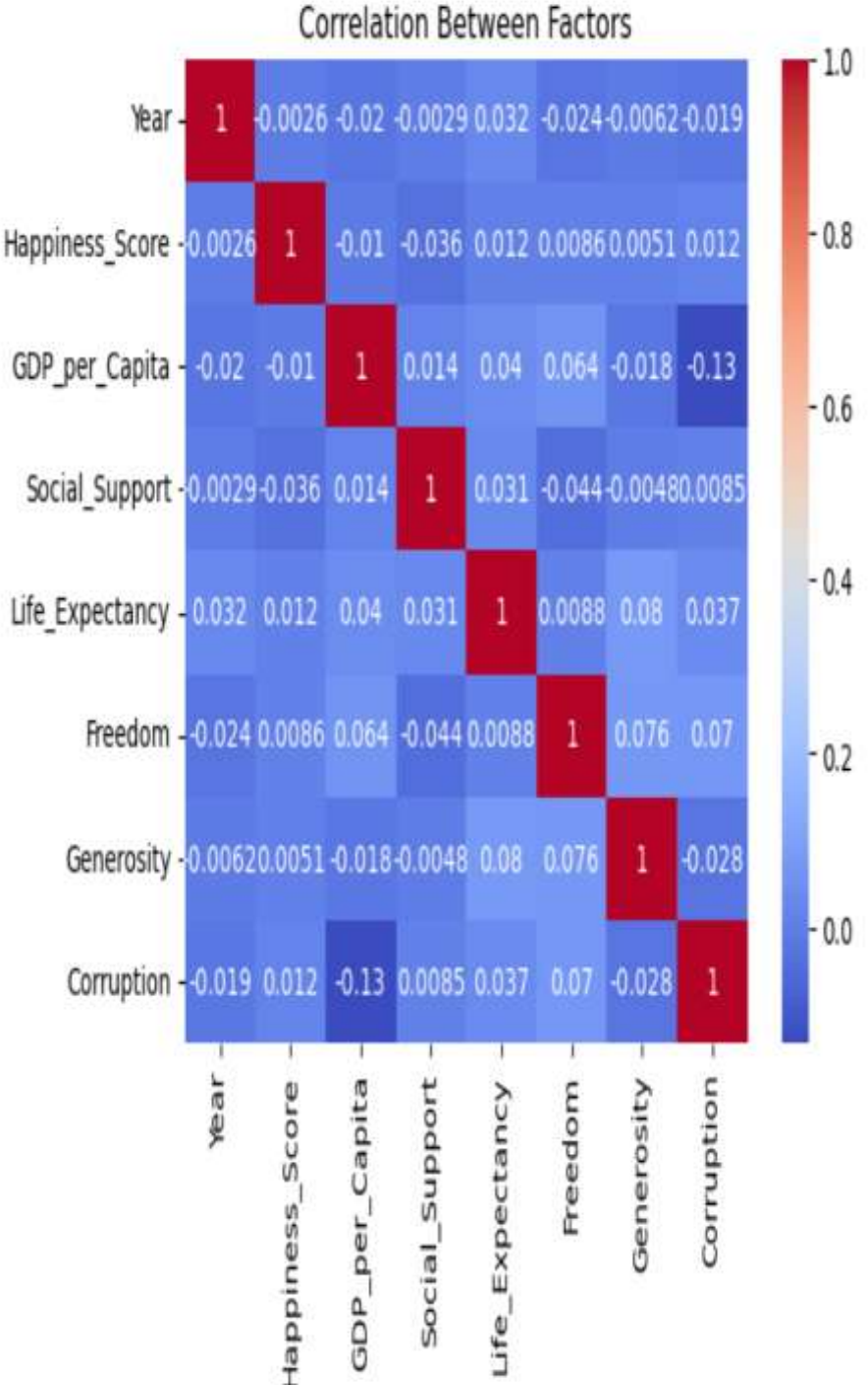
[6]:
```python
# Drop duplicates
df = df.drop_duplicates()

# Handle missing values
df["Happiness_Score"] = df["Happiness_Score"].fillna(df["Happiness_Score"].mean())
df = df.dropna(subset=["Country"])

# Clean country names
df["Country"] = df["Country"].replace(["Unknown", None], np.nan)

# Save cleaned dataset
df.to_csv("world_happiness_cleaned.csv", index=False)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 477 entries, 0 to 499
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Country          463 non-null    object
 1   Year             418 non-null    float64
 2   Happiness_Score  477 non-null    float64
 3   GDP_per_Capita   428 non-null    float64
 4   Social_Support   402 non-null    float64
```

3. EXPLORATORY DATA ANALYSIS (EDA)

- The goal of EDA is to gain a deep understanding of the data's features before moving on to more complex modelling or analysis.
- Histogram: Distribution of Happiness Scores
- Heatmap: Correlation between factors


Distribution of Happiness Scores
Count
0
10
20
30
40
50
60
3
4
5
6
7
8


Correlation Between Factors
Year 1 -0.0026 -0.02 -0.0029 0.032 -0.024 -0.0062 -0.019
Happiness_Score -0.0026 1 -0.01 -0.036 0.012 0.0086 0.0051 0.012
GDP_per_Capita -0.02 -0.01 1 0.014 0.04 0.064 -0.018 -0.13
Social_Support -0.0029 -0.036 0.014 1 0.031 -0.044 -0.0048 0.0085
Life_Expectancy 0.032 0.012 0.04 0.031 1 0.0088 0.08 0.037
Freedom -0.024 0.0086 0.064 -0.044 0.0088 1 0.076 0.07
Generosity -0.0062 0.0051 -0.018 -0.0048 0.08 0.076 1 -0.028
Corruption -0.019 0.012 -0.13 0.0085 0.037 0.07 -0.028 1
Year
Happiness_Score
GDP_per_Capita
Social_Support
Life_Expectancy
Freedom
Generosity
Corruption
1.0
0.8
0.6
0.4
0.2
0.0

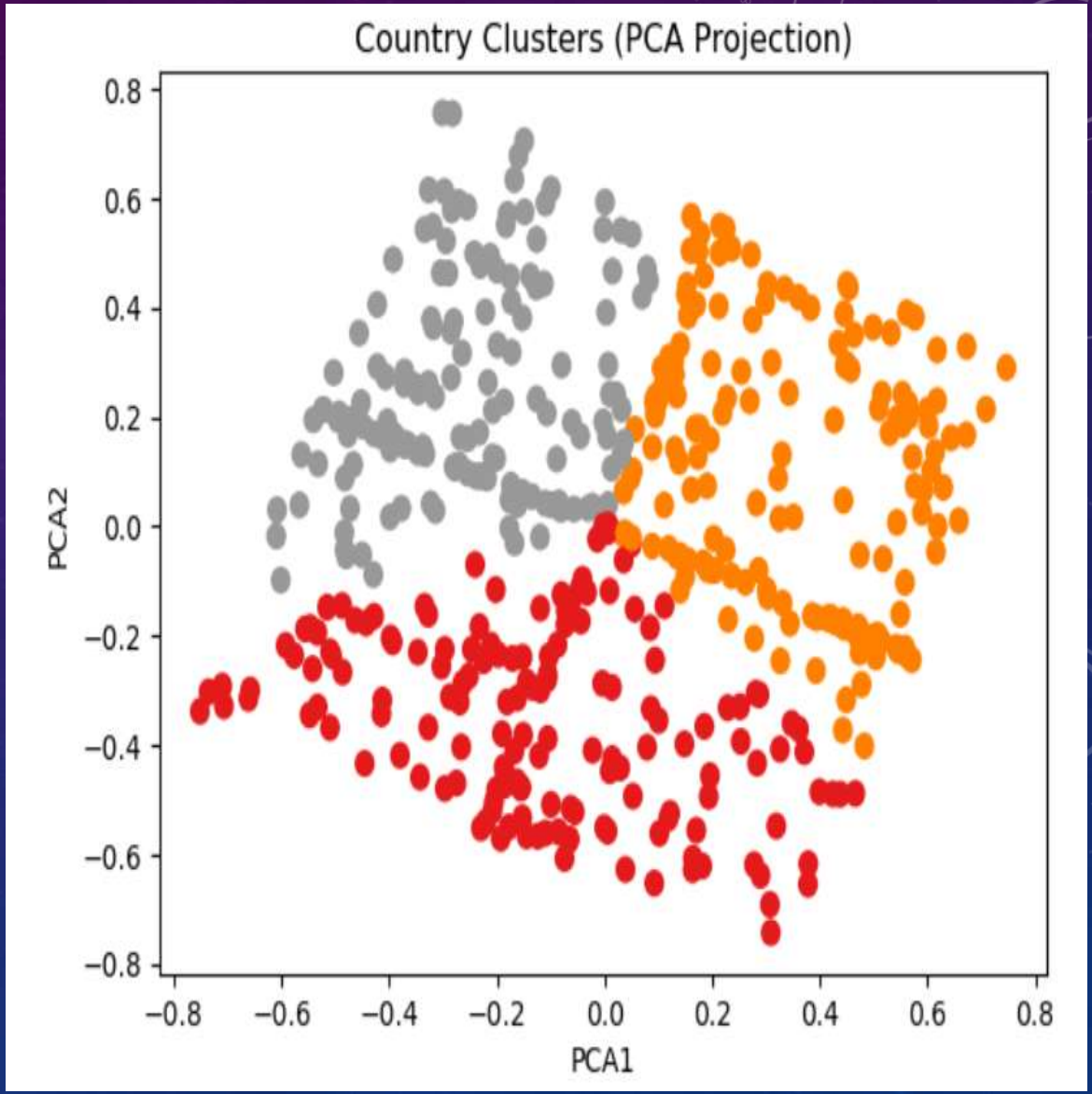
3.1 COUNTRY RANKINGS

- Top 10 Happiest Countries (bar chart)
- Bottom 10 Least Happy Countries (bar chart)

```
Top 10 Happiest:
 Country
Spain           6.043750
Brazil          5.958885
Austria         5.892089
Australia       5.769736
Luxembourg      5.735022
Germany         5.509145
Canada          5.467192
UK              5.449670
India           5.402051
Sweden          5.372216
Name: Happiness_Score, dtype: float64

Bottom 10 Least Happy:
 Country
USA               4.175000
Netherlands       4.475000
Russia            4.637454
Switzerland       4.788275
New Zealand       4.804694
France            4.950816
China             4.960551
Japan             5.064783
South Korea       5.122735
South Africa      5.159776
Name: Happiness_Score, dtype: float64
```

3.2 CLUSTERING ANALYSIS

- Method: MiniBatch KMeans + PCA
- 3 clusters: High Happiness, Medium Happiness, Low Happiness
- PCA scatter plot visualization



4. PREDICTIVE MODELING

- Model: Linear Regression
- Features: GDP, Social Support, Life Expectancy, Freedom, Generosity, Corruption
- Metrics: R^2 Score, RMSE
- Key influencing features identified
- Used regression to predict happiness score

R^2 Score: -0.026444708421588414

RMSE: 1.4369319922451533

5. TIME-SERIES FORECASTING

- Example: India (2015–2021 data)
- Model: ARIMA
- Forecast: 2022–2024
- Visualization: Line chart (actual vs forecast)
- Forecasted Scores (next 3 years):

1. 20 4.704259
2. 21 5.552971
3. 22 5.314909

INSIGHTS & CONCLUSION

- GDP & Social Support strongly correlate with happiness
- Nordic countries consistently rank highest
- Forecast shows stable happiness trends
- Future work: Extend dataset, build Streamlit dashboard