

PLEASE USE THIS CSV FILE TO ANALYZE INSIGHTS AND
FORECAST FORESIGHTS ACCORDING TO DATA SCIENCE LIFE
CYCLE PROCESS.

CONTENTS	
Analyze Step	2
Perplexity Answer.....	5
Data Understanding and Preparation.....	5
Exploratory Data Analysis	6
Insights.....	6
Forecasting Foresights	7

ANALYZE STEP

Analyze the provided CSV file for insights based on the data science life cycle process.

```
import pandas as pd

# Load the CSV data
data = pd.read_csv('thailand_domestic_tourism.csv')

# Display basic information about the dataset
data_info = data.info()

# Display summary statistics
summary_stats = data.describe()

# Check for missing values
missing_values = data.isnull().sum()

# Group by region and calculate total tourists and revenue
region_analysis = data.groupby('region_eng').agg({
    'no_tourist_all': 'sum',
    'revenue_all': 'sum'
}).reset_index()

# Analyze the ratio of tourists staying overnight by region
region_analysis['avg_ratio_tourist_stay'] =
data.groupby('region_eng')['ratio_tourist_stay'].mean().values

# Analyze the top 5 provinces by total revenue
top_revenue_provinces = data[['province_eng',
'revenue_all']].sort_values(by='revenue_all', ascending=False).head(5)

# Analyze the top 5 provinces by number of tourists
top_tourist_provinces = data[['province_eng',
'no_tourist_all']].sort_values(by='no_tourist_all', ascending=False).head(5)

(data_info, summary_stats, missing_values, region_analysis,
top_revenue_provinces, top_tourist_provinces)
```

Forecast future trends based on the analysis of the CSV data

```
from statsmodels.tsa.holtwinters import ExponentialSmoothing
import numpy as np
import matplotlib.pyplot as plt

# Convert travel_date to datetime and set as index
data['travel_date'] = pd.to_datetime(data['travel_date'])
data.set_index('travel_date', inplace=True)

# Aggregate data by month for forecasting
data_monthly = data.resample('M').sum()

# Select the column for forecasting (e.g., total revenue)
revenue_series = data_monthly['revenue_all']

# Apply Exponential Smoothing for forecasting
model = ExponentialSmoothing(revenue_series, seasonal='add',
                              seasonal_periods=12)
fit_model = model.fit()

# Forecast for the next 12 months
forecast = fit_model.forecast(12)

# Plot the original data and forecast
plt.figure(figsize=(10, 6))
plt.plot(revenue_series, label='Original Data')
plt.plot(forecast, label='Forecast', color='red')
plt.legend()
plt.title('Revenue Forecast')
plt.xlabel('Date')
plt.ylabel('Revenue')
plt.grid()
plt.show()

forecast
```

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split

# Prepare data for linear regression forecasting
revenue_series = data_monthly['revenue_all'].reset_index()
```

```

revenue_series['time_index'] = np.arange(len(revenue_series))

# Define features and target
X = revenue_series[['time_index']]
y = revenue_series['revenue_all']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42, shuffle=False)

# Train the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict future values
future_time_index = np.arange(len(revenue_series), len(revenue_series) +
12).reshape(-1, 1)
forecast = model.predict(future_time_index)

# Plot the original data and forecast
plt.figure(figsize=(10, 6))
plt.plot(revenue_series['time_index'], revenue_series['revenue_all'],
label='Original Data')
plt.plot(future_time_index, forecast, label='Forecast', color='red')
plt.legend()
plt.title('Revenue Forecast')
plt.xlabel('Time Index')
plt.ylabel('Revenue')
plt.grid()
plt.show()

forecast

```

```

# Re-aggregate data by month for forecasting
monthly_data = data.resample('M').sum()

# Prepare data for linear regression forecasting
revenue_series = monthly_data['revenue_all'].reset_index()
revenue_series['time_index'] = np.arange(len(revenue_series))

# Define features and target
X = revenue_series[['time_index']]
y = revenue_series['revenue_all']

```

```

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42, shuffle=False)

# Train the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict future values
future_time_index = np.arange(len(revenue_series), len(revenue_series) +
12).reshape(-1, 1)
forecast = model.predict(future_time_index)

# Plot the original data and forecast
plt.figure(figsize=(10, 6))
plt.plot(revenue_series['time_index'], revenue_series['revenue_all'],
label='Original Data')
plt.plot(future_time_index, forecast, label='Forecast', color='red')
plt.legend()
plt.title('Revenue Forecast')
plt.xlabel('Time Index')
plt.ylabel('Revenue')
plt.grid()
plt.show()

forecast

```

PERPLEXITY ANSWER

Based on the provided CSV data on Thailand's domestic tourism, I'll analyze insights and forecast foresights following the data science lifecycle process:

DATA UNDERSTANDING AND PREPARATION

The dataset contains information on domestic tourism in Thailand, including:

- Tourist numbers (total, foreign, Thai, staying overnight)
- Revenue (total, from foreign tourists, from Thai tourists)
- Geographic information (province, region)
- Time information (travel date)

Key observations:

- The data covers 77 provinces across 5 regions of Thailand
- Time period spans from 2019 to 2023
- No missing values were detected

EXPLORATORY DATA ANALYSIS

Regional Analysis

1. Tourist Distribution:

- Central region attracts the most tourists (354.4 million)
- East Northeast region has the lowest tourist numbers (115.8 million)

2. Revenue Generation:

- Central region generates the highest revenue (2.29 trillion baht)
- East Northeast region generates the lowest revenue (231.4 billion baht)

3. Overnight Stay Ratio:

- North region has the highest average ratio of tourists staying overnight (41.02%)
- South region has the lowest average ratio (36.51%)

Top Performing Provinces

1. By Revenue:

- Bangkok consistently tops the list, generating over 100 billion baht in revenue
- Other top provinces likely include popular tourist destinations like Phuket and Chiang Mai

2. By Tourist Numbers:

- Bangkok leads with over 6 million tourists
- Other provinces in the top 5 receive around 5.9-6.0 million tourists

Seasonal Trends

While not explicitly shown in the provided output, analyzing the 'travel_date' column would reveal seasonal patterns in tourist numbers and revenue.

INSIGHTS

1. *Bangkok Dominance:* The capital city is the primary driver of Thailand's domestic tourism, leading in both tourist numbers and revenue generation.
2. *Regional Disparities:* There's a significant gap between the central region and others in terms of tourist attraction and revenue generation, indicating potential for development in other regions.
3. *Overnight Stays:* The north region's high ratio of overnight stays suggests it's more effective at retaining tourists, possibly due to its natural attractions and cultural sites.

4. *Foreign vs. Domestic Tourists*: The presence of separate columns for foreign and Thai tourists indicates the importance of both markets to Thailand's tourism industry.

FORECASTING FORESIGHTS

1. *Post-Pandemic Recovery*: Analyzing trends from 2019 to 2023 could reveal the impact of the COVID-19 pandemic and the pace of recovery in different regions and provinces.
2. *Emerging Destinations*: Identifying provinces with rapidly growing tourist numbers or revenue could highlight emerging tourism hotspots.
3. *Seasonal Strategies*: Understanding seasonal patterns could help in developing targeted marketing strategies and resource allocation for peak and off-peak seasons.
4. *Investment Opportunities*: Regions with high tourist numbers but lower revenue might present opportunities for developing higher-value tourism offerings.
5. *Sustainability Concerns*: High tourist numbers in certain areas might raise concerns about overtourism and environmental sustainability, necessitating careful management strategies.

To further refine these insights and develop more accurate forecasts, additional steps would include:

- Time series analysis to identify trends and seasonality
- Predictive modeling to forecast future tourist numbers and revenue
- Segmentation analysis to understand different tourist profiles
- Correlation analysis to identify factors influencing tourism performance

These analyses would provide a more comprehensive understanding of Thailand's domestic tourism landscape and enable more accurate predictions of future trends.