**สามารถนำข้อมูลต่อไปนี้มาทำนายอะไรได้บ้างโดยใช้ data science methodology**

ขั้นตอน

วิเคราะห์ข้อมูลการท่องเที่ยในประเทศไทยเพื่อหาความสัมพนธ์ระหว่างจำนวนผู้เข้าชมและราได้

|  |
| --- |
| import pandas as pd  import matplotlib.pyplot as plt  import seaborn as sns  *# Load the data*  data = pd.read\_csv('thailand\_domestic\_tourism.csv')  *# Check the data structure*  data.head() |

|  |
| --- |
| # Analyze the relationship between the number of tourists and revenue  # Scatter plot for total tourists vs total revenue  plt.figure(figsize=(10, 6))  sns.scatterplot(x=data['no\_tourist\_all'], y=data['revenue\_all'])  plt.title('Relationship between Total Tourists and Total Revenue')  plt.xlabel('Number of Tourists (All)')  plt.ylabel('Total Revenue')  plt.show()  # Correlation analysis  correlation = data[['no\_tourist\_all', 'revenue\_all']].corr()  correlation |

สร้าง model การทำนายรายได้จากการท่องเที่ยว โดยใช้ข้อมูลที่มีอยู่

|  |
| --- |
| from sklearn.model\_selection import train\_test\_split  from sklearn.linear\_model import LinearRegression  from sklearn.metrics import mean\_squared\_error, r2\_score  # Prepare the data for modeling  X = data[['no\_tourist\_all']]  y = data['revenue\_all']  # Split the 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)  # Create and train the model  model = LinearRegression()  model.fit(X\_train, y\_train)  # Make predictions  y\_pred = model.predict(X\_test)  # Evaluate the model  mse = mean\_squared\_error(y\_test, y\_pred)  r2 = r2\_score(y\_test, y\_pred)  mse, r2, model.coef\_, model.intercept\_ |

Output

|  |
| --- |
| (7.823276297907415e+18,  0.7353118718290527,  array([13075.51343219]),  -1362064813.0760434) |

สรุปผลการวิเคราะห์และการทำนายที่ได้จากข้อมูล

**Answer**

A graph showing the number of tourists

Description automatically generated

ข้อมูลที่แสดงถึงความสัมพันธ์ระหว่างจำนวนผู้เยี่ยมชม (นักท่องเที่ยว) และรายได้จากการท่องเที่ยวในประเทศไทยสามารถนำไปใช้ในการวิเคราะห์และทำนายผลลัพธ์ต่าง ๆ ได้โดยใช้วิธีการทาง Data Science ดังนี้:

**สิ่งที่สามารถวิเคราะห์และทำนายได้**

**1. การทำนายรายได้จากการท่องเที่ยว**

* จากความสัมพันธ์ที่ชัดเจนระหว่างจำนวนผู้เยี่ยมชมและรายได้ (Correlation = 0.89), สามารถสร้างโมเดลการถดถอยเชิงเส้น (Linear Regression) เพื่อทำนายรายได้จากการท่องเที่ยวบนพื้นฐานของจำนวนผู้เยี่ยมชม.
* ตัวอย่างผลลัพธ์:
  + ค่าสัมประสิทธิ์ (β1) = 13075.51 หมายความว่า ทุก ๆ นักท่องเที่ยวเพิ่มขึ้น 1 คน จะเพิ่มรายได้ประมาณ 13,075 บาท
  + ค่า R^2 = 0.735 แสดงว่าโมเดลสามารถอธิบายความแปรปรวนของข้อมูลได้ประมาณ 73.5%.

**2~~. การวิเคราะห์แนวโน้มและฤดูกาล~~**

* ~~ข้อมูลที่มีวันที่ (travel\_date) สามารถใช้วิเคราะห์แนวโน้ม (Trend) และฤดูกาล (Seasonality) เช่น:~~
  + ~~ช่วงเวลาที่มีนักท่องเที่ยวสูงสุด/ต่ำสุด~~
  + ~~ผลกระทบของเทศกาลหรือวันหยุดยาวต่อจำนวนนักท่องเที่ยวและรายได้~~

**3. การแบ่งกลุ่มจังหวัดตามศักยภาพด้านการท่องเที่ยว**

* สามารถใช้เทคนิคการจัดกลุ่ม (Clustering) เพื่อจัดกลุ่มจังหวัดตามจำนวนผู้เยี่ยมชมและรายได้ เช่น:
  + กลุ่มจังหวัดที่มีนักท่องเที่ยวต่างชาติสูง
  + กลุ่มจังหวัดที่พึ่งพานักท่องเที่ยวในประเทศ

**~~4. การวางแผนกลยุทธ์การตลาด~~**

* ~~วิเคราะห์ข้อมูลเพื่อระบุพื้นที่ที่ควรมุ่งเน้นการส่งเสริมการตลาด เช่น จังหวัดที่มีศักยภาพสูงแต่ยังไม่ได้รับความนิยมมากนัก~~

**5. การประเมินผลกระทบทางเศรษฐกิจ**

* ใช้ข้อมูลในการประเมินผลกระทบทางเศรษฐกิจจากการเปลี่ยนแปลงจำนวนนักท่องเที่ยว เช่น:
  + ผลกระทบจาก COVID-19 ต่อรายได้ในแต่ละภูมิภาค
  + การฟื้นตัวของอุตสาหกรรมหลังวิกฤต

**~~6. การจัดสรรทรัพยากร~~**

* ~~ใช้ข้อมูลเพื่อช่วยในการจัดสรรทรัพยากร เช่น งบประมาณสำหรับโครงสร้างพื้นฐานด้านการท่องเที่ยวในพื้นที่ที่มีศักยภาพสูง~~

**7. การพัฒนาโมเดลพยากรณ์ขั้นสูง**

* ใช้ Machine Learning เช่น Random Forest หรือ Gradient Boosting เพื่อสร้างโมเดลพยากรณ์ที่แม่นยำมากขึ้น โดยรวมปัจจัยอื่น ๆ เช่น:
  + สัดส่วนของนักท่องเที่ยวต่างชาติ/ในประเทศ
  + ระยะเวลาการพักเฉลี่ย
  + อัตราการเข้าพักโรงแรม

**Please use this csv file to analyze insights and forecast foresights according to data science life cycle process.**

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**

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.

How can I visualize the regional distribution of tourists and revenue

Visualize the regional distribution of tourists and revenue from the provided

|  |
| --- |
| import pandas as pd  import matplotlib.pyplot as plt  # Load the dataset  data = pd.read\_csv('thailand\_domestic\_tourism.csv')  # Group data by region and calculate total tourists and revenue  regional\_data = data.groupby('region\_eng').agg({  'no\_tourist\_all': 'sum',  'revenue\_all': 'sum'  }).reset\_index()  # Plot regional distribution of tourists  plt.figure(figsize=(12, 6))  plt.subplot(1, 2, 1)  plt.bar(regional\_data['region\_eng'], regional\_data['no\_tourist\_all'], color='skyblue')  plt.title('Regional Distribution of Tourists')  plt.xlabel('Region')  plt.ylabel('Total Tourists')  plt.xticks(rotation=45)  # Plot regional distribution of revenue  plt.subplot(1, 2, 2)  plt.bar(regional\_data['region\_eng'], regional\_data['revenue\_all'], color='lightgreen')  plt.title('Regional Distribution of Revenue')  plt.xlabel('Region')  plt.ylabel('Total Revenue (Baht)')  plt.xticks(rotation=45)  plt.tight\_layout()  plt.show() |

Finalize the visualization and prepare to review

**Answer**

A comparison of a graph

Description automatically generated with medium confidence

To visualize the regional distribution of tourists and revenue in Thailand, the dataset was analyzed and grouped by regions. Two bar charts were created:

1. **Regional Distribution of Tourists**:
   * The total number of tourists for each region was summed up and displayed.
   * The Central region had the highest number of tourists, significantly surpassing other regions.
   * Other regions (East, East Northeast, North, and South) showed relatively similar tourist numbers but were much lower compared to the Central region.
2. **Regional Distribution of Revenue**:
   * The total revenue generated by tourists in each region was calculated.
   * Again, the Central region dominated in terms of revenue generation, with a noticeable gap compared to other regions.
   * The South region followed as the second-highest revenue generator, likely due to its popular beach destinations like Phuket and Krabi.

Visual Insights:

* The bar charts clearly highlight regional disparities, with the Central region being the focal point for both tourist numbers and revenue.
* These visualizations can be used to identify areas for tourism development or marketing efforts to balance regional tourism contributions.

If you want to replicate this analysis or modify it further, the code used is available in the provided execution results.