45-Minute Data Analysis Sprint: Insurance Cost Analysis

Objective: To conduct a rapid, focused analysis of an insurance dataset to answer key business questions under a tight deadline.

Total Time: 45 Minutes

Scenario: The Urgent Analyst Request

Your Role: You are a Data Analyst at "HealthStat Insurance." Your manager has just come to your desk with an urgent request.

The Situation:
"Hi, I'm heading into a last-minute strategy meeting in one hour, and I need some quick data points to support our discussion on future planning. I need you to perform a rapid analysis of our customer insurance data.

Don't worry about a full, formal report. Just create a clean Jupyter Notebook that answers the following specific questions. I need clear visualizations and a one-sentence conclusion for each question so I can quickly understand the key takeaways.

You have 45 minutes. Thanks for your help!"

Instructions & Question List

Instructions for the Student:
Create a new Jupyter Notebook. For each question below, you must:

1. Write the Python code to perform the analysis.
2. Generate one clear and properly labeled visualization.
3. Write a single, concise "Key Takeaway" sentence in a markdown cell immediately after the plot.

Part 1: Data Setup (5 minutes)

- **Task 1:** Import the pandas, matplotlib.pyplot, and seaborn libraries.
- **Task 2:** Load the dataset from this URL: https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/insurance.csv
- **Task 3:** Briefly inspect the data using .info() to ensure it's loaded correctly and check for missing values.

Part 2: Core Analysis Questions (35 minutes)

Question 1: The Smoker Impact

- **Question:** How significant is the difference in medical charges between smokers and non-smokers?
- **Required Visualization:** A **boxplot** comparing the charges for smokers and non-smokers.

Question 2: The Age Factor

- **Question:** What is the general relationship between a customer's age and their charges?
- **Required Visualization:** A **scatterplot** showing age on the x-axis and charges on the y-axis.

Question 3: The BMI Threshold

- **Question:** Our wellness team considers a BMI over 30 to be in the "obese" category. Are medical charges notably higher for this group compared to those with a BMI under 30?
- **Required Visualization:** A **boxplot** comparing charges for customers with bmi < 30 vs. bmi >= 30. (Hint: You may need to create a new column to categorize customers first).

Question 4: The Regional Breakdown

- **Question:** Which geographic region has the highest average insurance charges?
- **Required Visualization:** A **bar chart** showing the average charges for each region.

Question 5: The Combined Effect (The "Money" Chart)

- **Question:** How does the relationship between BMI and charges differ for smokers versus non-smokers?
- **Required Visualization:** A **scatterplot** with bmi on the x-axis, charges on the y-axis, and the points colored by smoker status (using the hue parameter).

Part 3: Final Conclusion (5 minutes)

Question 6: The Final Verdict

- **Question:** In one or two sentences, summarize the most important findings for your manager. Based on your 45-minute analysis, what is the single most critical factor that determines the cost of a customer's health insurance?
- **Required Action:** Write your answer in a final markdown cell at the bottom of the notebook.