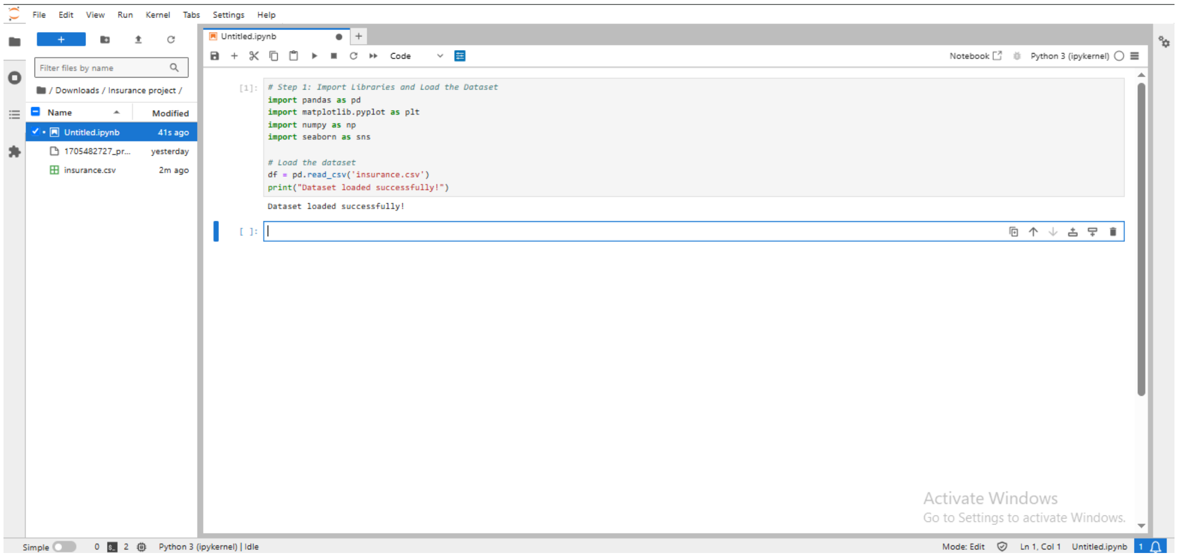
Insurance Data Analysis: Key Observations

Step 1: Import Libraries and Load the Dataset

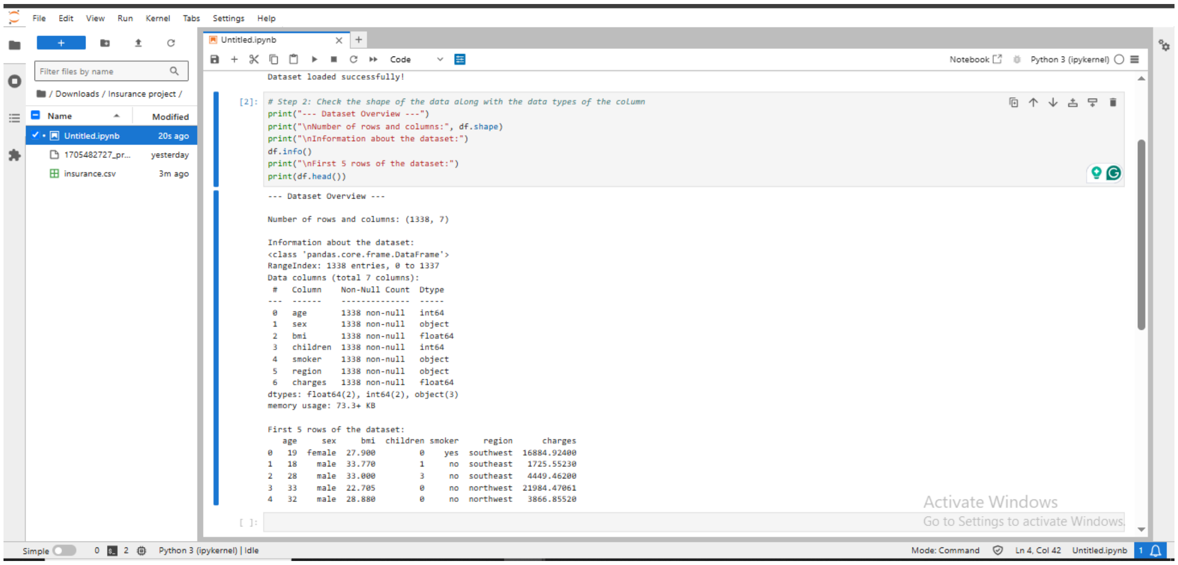
- Successfully imported essential libraries (Pandas, Matplotlib, NumPy, Seaborn).
- The insurance.csv dataset was loaded into a Pandas DataFrame without errors, confirming successful initial setup.



```
[1]: # Step 1: Import Libraries and Load the Dataset
     import pandas as pd
     import matplotlib.pyplot as plt
     import numpy as np
     import seaborn as sns

     # Load the dataset
     df = pd.read_csv('insurance.csv')
     print("Dataset loaded successfully!")

     Dataset loaded successfully!

[ ]:
```

Step 2: Check the Shape of the Data and Data Types

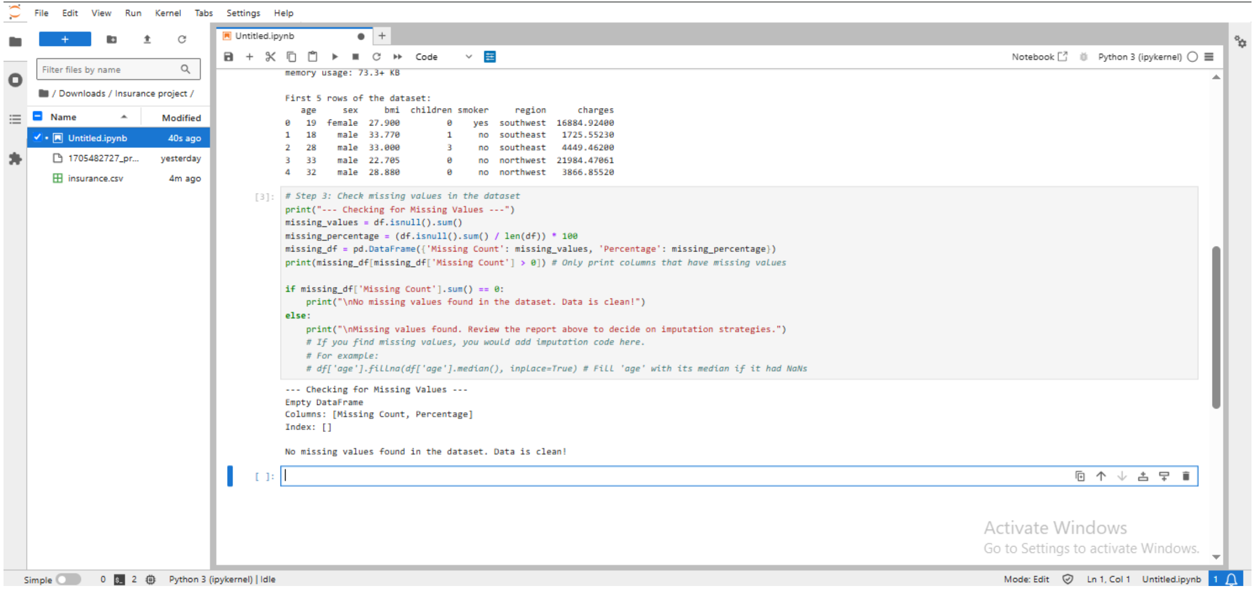
- The dataset contains **1338** rows (representing policyholders) and **7** columns (representing features).
- Data types for each column are appropriate:
  - age (int)
  - sex (object)
  - bmi (float)
  - children (int)
  - smoker (object)
  - region (object)
  - charges (float)
- All columns have an equal number of non-null entries, indicating no immediate missing values.



```
     Dataset loaded successfully!

[2]: # Step 2: Check the shape of the data along with the data types of the column
     print("--- Dataset Overview ---")
     print("\nNumber of rows and columns:", df.shape)
     print("\nInformation about the dataset:")
     df.info()
     print("\nFirst 5 rows of the dataset:")
     print(df.head())

     --- Dataset Overview ---

     Number of rows and columns: (1338, 7)

     Information about the dataset:
     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 1338 entries, 0 to 1337
     Data columns (total 7 columns):
      #   Column    Non-Null Count  Dtype  
     ---  ------    --------------  -----  
      0   age       1338 non-null   int64  
      1   sex       1338 non-null   object 
      2   bmi       1338 non-null   float64
      3   children  1338 non-null   int64  
      4   smoker    1338 non-null   object 
      5   region    1338 non-null   object 
      6   charges   1338 non-null   float64
     dtypes: float64(2), int64(2), object(3)
     memory usage: 73.3+ KB

     First 5 rows of the dataset:
        age     sex     bmi  children smoker     region      charges
     0   19  female  27.900         0    yes  southwest  16884.92400
     1   18    male  33.770         1     no  southeast   1725.55230
     2   28    male  33.000         3     no  southeast   4449.46200
     3   33    male  22.705         0     no  northwest  21984.47061
     4   32    male  28.880         0     no  northwest   3866.85520

[ ]:
```

Step 3: Check Missing Values in the Dataset

- No missing values were found in any of the columns (df.isnull().sum() showed all zeros). The dataset is clean in terms of missing data, and no imputation steps are required.



```
memory usage: 73.3+ KB

First 5 rows of the dataset:
   age     sex     bmi  children smoker     region      charges
0   19  female  27.900         0    yes  southwest  16884.92400
1   18    male  33.770         1     no  southeast   1725.55230
2   28    male  33.000         3     no  southeast   4449.46200
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
```

```
[3]: # Step 3: Check missing values in the dataset
print("--- Checking for Missing Values ---")
missing_values = df.isnull().sum()
missing_percentage = (df.isnull().sum() / len(df)) * 100
missing_df = pd.DataFrame({'Missing Count': missing_values, 'Percentage': missing_percentage})
print(missing_df[missing_df['Missing Count'] > 0]) # Only print columns that have missing values

if missing_df['Missing Count'].sum() == 0:
    print("\nNo missing values found in the dataset. Data is clean!")
else:
    print("\nMissing values found. Review the report above to decide on imputation strategies.")
    # If you find missing values, you would add imputation code here.
    # For example:
    # df['age'].fillna(df['age'].median(), inplace=True) # Fill 'age' with its median if it had NaNs
```

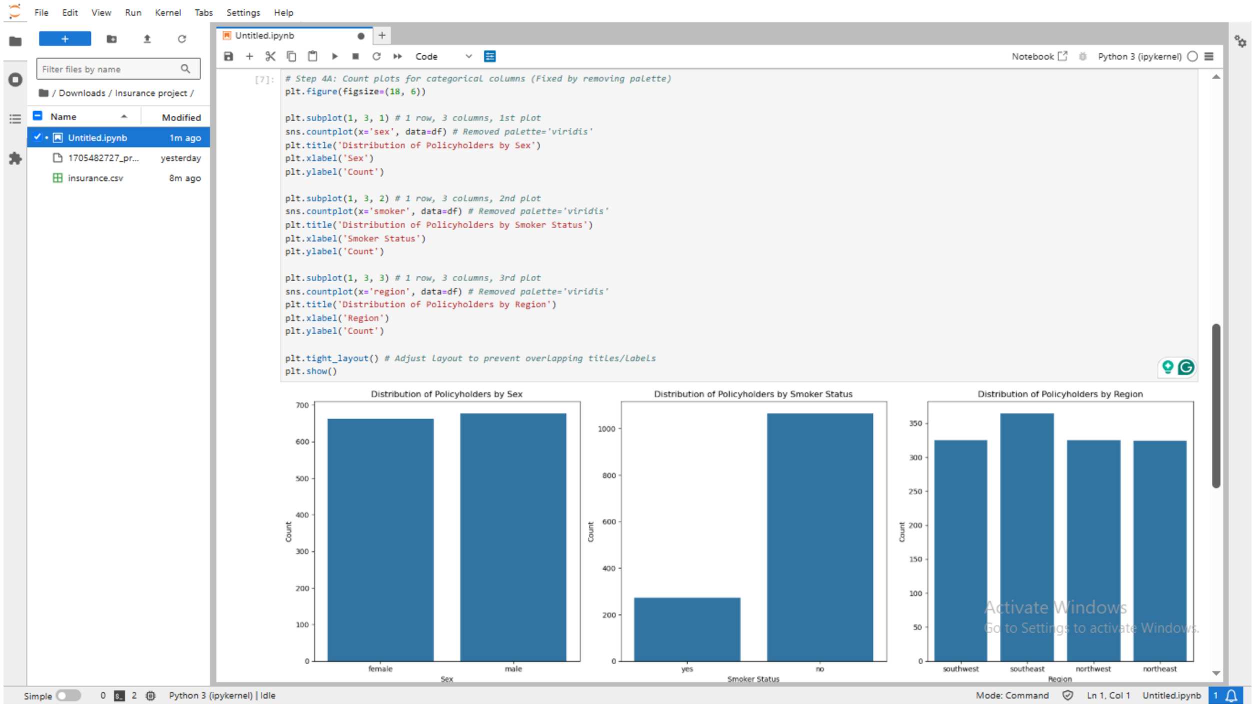
```
--- Checking for Missing Values ---
Empty DataFrame
Columns: [Missing Count, Percentage]
Index: []

No missing values found in the dataset. Data is clean!
```

```
[ ]:
```

Step 4: Explore the Relationship Between Features and Target Column (charges)

A. Count Plots for Categorical Features (Distribution)

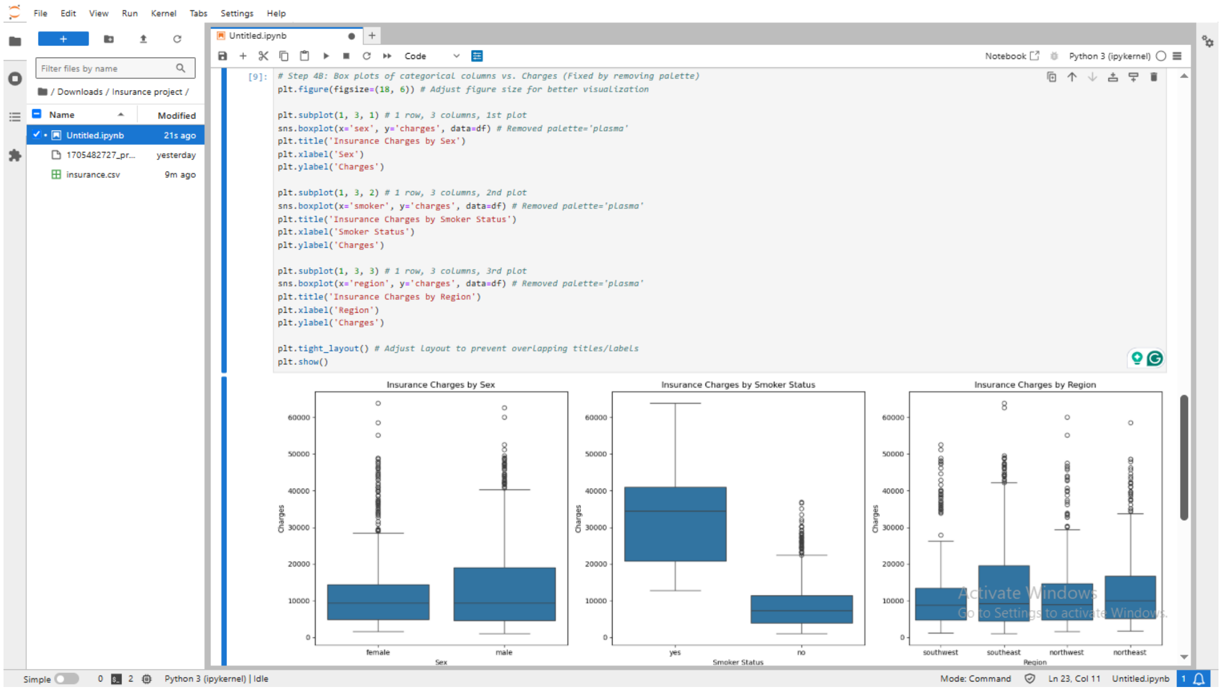
- **Sex:** The distribution between male and female policyholders is relatively balanced.
- **Smoker Status:** A significant majority of policyholders are non-smokers, with a smaller proportion being smokers.
- **Region:** Policyholders are quite evenly distributed across the four geographical regions (northeast, northwest, southeast, southwest).



```
[7]: # Step 4A: Count plots for categorical columns (Fixed by removing palette)
plt.figure(figsize=(18, 6))

plt.subplot(1, 3, 1) # 1 row, 3 columns, 1st plot
sns.countplot(x='sex', data=df) # Removed palette='viridis'
plt.title('Distribution of Policyholders by Sex')
plt.xlabel('Sex')
plt.ylabel('Count')

plt.subplot(1, 3, 2) # 1 row, 3 columns, 2nd plot
sns.countplot(x='smoker', data=df) # Removed palette='viridis'
plt.title('Distribution of Policyholders by Smoker Status')
plt.xlabel('Smoker Status')
plt.ylabel('Count')

plt.subplot(1, 3, 3) # 1 row, 3 columns, 3rd plot
sns.countplot(x='region', data=df) # Removed palette='viridis'
plt.title('Distribution of Policyholders by Region')
plt.xlabel('Region')
plt.ylabel('Count')

plt.tight_layout() # Adjust layout to prevent overlapping titles/labels
plt.show()
```

B. Box Plots for Categorical Features vs. Charges

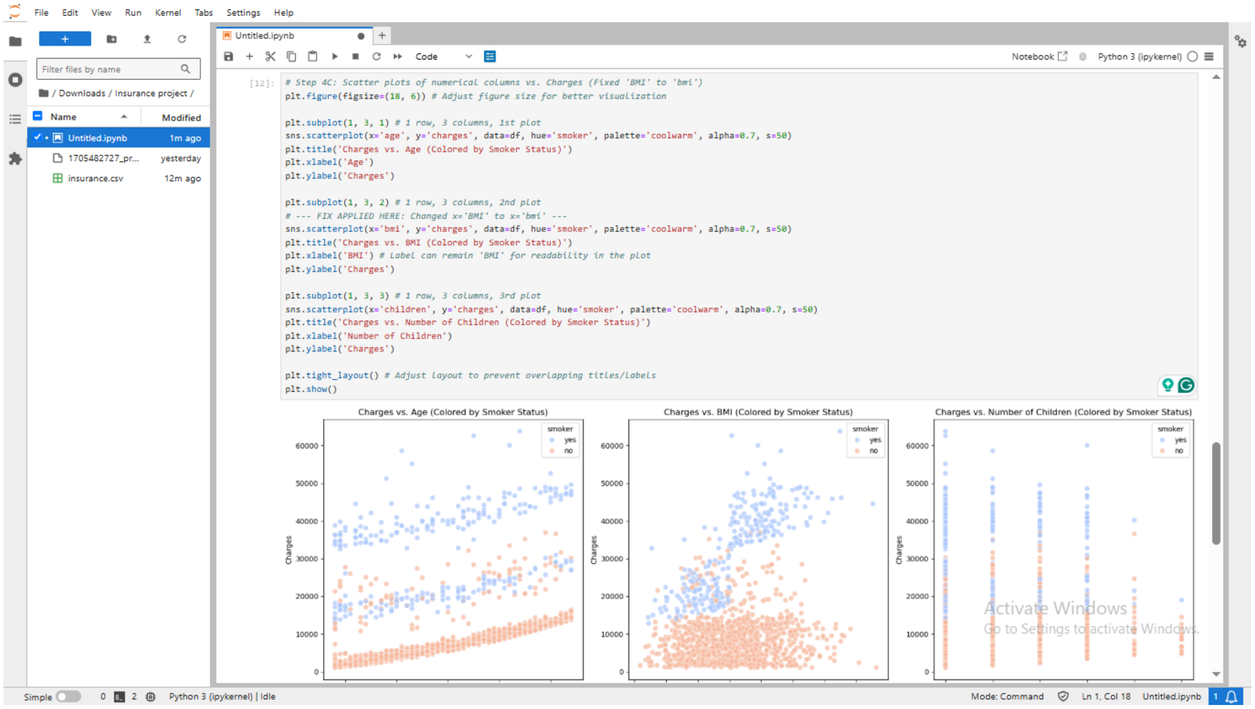
- **Smoker:** There is a stark and significant difference in insurance charges between smokers and non-smokers. Smokers consistently face much higher premiums, indicating that smoking status is a primary determinant of insurance cost.
- **Sex:** The median charges for males and females are relatively similar, with no dramatic difference based solely on gender.
- **Region:** While minor variations exist, the median charges across different regions are quite similar, with no single region showing an extremely high or low premium compared to others.



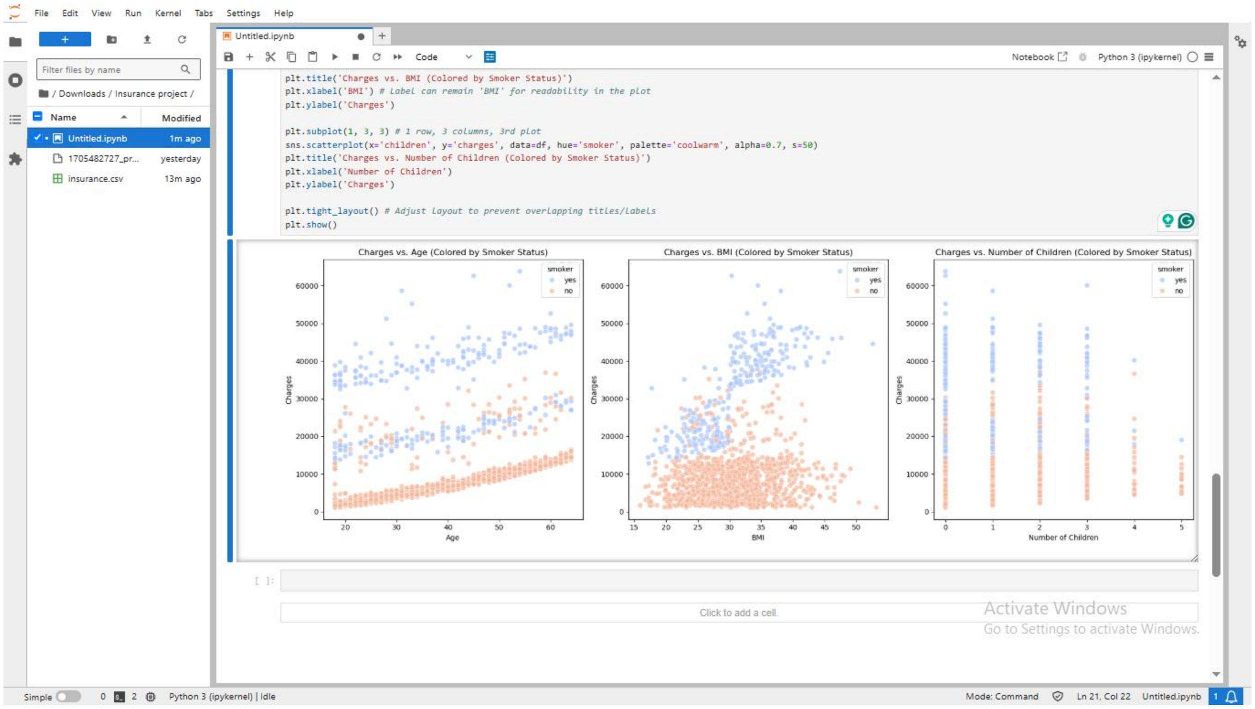
```
[9]: # Step 4B: Box plots of categorical columns vs. Charges (Fixed by removing palette)
     plt.figure(figsize=(18, 6)) # Adjust figure size for better visualization

     plt.subplot(1, 3, 1) # 1 row, 3 columns, 1st plot
     sns.boxplot(x='sex', y='charges', data=df) # Removed palette='plasma'
     plt.title('Insurance Charges by Sex')
     plt.xlabel('Sex')
     plt.ylabel('Charges')

     plt.subplot(1, 3, 2) # 1 row, 3 columns, 2nd plot
     sns.boxplot(x='smoker', y='charges', data=df) # Removed palette='plasma'
     plt.title('Insurance Charges by Smoker Status')
     plt.xlabel('Smoker Status')
     plt.ylabel('Charges')

     plt.subplot(1, 3, 3) # 1 row, 3 columns, 3rd plot
     sns.boxplot(x='region', y='charges', data=df) # Removed palette='plasma'
     plt.title('Insurance Charges by Region')
     plt.xlabel('Region')
     plt.ylabel('Charges')

     plt.tight_layout() # Adjust layout to prevent overlapping titles/labels
     plt.show()
```

C. Scatter Plots for Numerical Features vs. Charges (Colored by Smoker Status)

- **Age vs. Charges:** There's a clear positive linear relationship. As age increases, insurance premiums generally increase for both smokers and non-smokers. Two distinct bands of data points are visible, separating smokers (higher charges) from non-smokers (lower charges).
- **BMI vs. Charges:** A general positive trend is observed where higher BMI tends to be associated with higher charges, especially noticeable within both the smoker and non-smoker groups. However, the relationship is more scattered than with age.
- **Children vs. Charges:** The number of children does not show a strong direct linear correlation with charges. The data points are widely dispersed, suggesting that this feature alone is not a major predictor of premium costs.



```
[12]: # Step 4C: Scatter plots of numerical columns vs. Charges (Fixed 'BMI' to 'bmi')
plt.figure(figsize=(18, 6)) # Adjust figure size for better visualization

plt.subplot(1, 3, 1) # 1 row, 3 columns, 1st plot
sns.scatterplot(x='age', y='charges', data=df, hue='smoker', palette='coolwarm', alpha=0.7, s=50)
plt.title('Charges vs. Age (Colored by Smoker Status)')
plt.xlabel('Age')
plt.ylabel('Charges')

plt.subplot(1, 3, 2) # 1 row, 3 columns, 2nd plot
# --- FIX APPLIED HERE: Changed x='BMI' to x='bmi' ---
sns.scatterplot(x='bmi', y='charges', data=df, hue='smoker', palette='coolwarm', alpha=0.7, s=50)
plt.title('Charges vs. BMI (Colored by Smoker Status)')
plt.xlabel('BMI') # Label can remain 'BMI' for readability in the plot
plt.ylabel('Charges')

plt.subplot(1, 3, 3) # 1 row, 3 columns, 3rd plot
sns.scatterplot(x='children', y='charges', data=df, hue='smoker', palette='coolwarm', alpha=0.7, s=50)
plt.title('Charges vs. Number of Children (Colored by Smoker Status)')
plt.xlabel('Number of Children')
plt.ylabel('Charges')

plt.tight_layout() # Adjust layout to prevent overlapping titles/labels
plt.show()
```

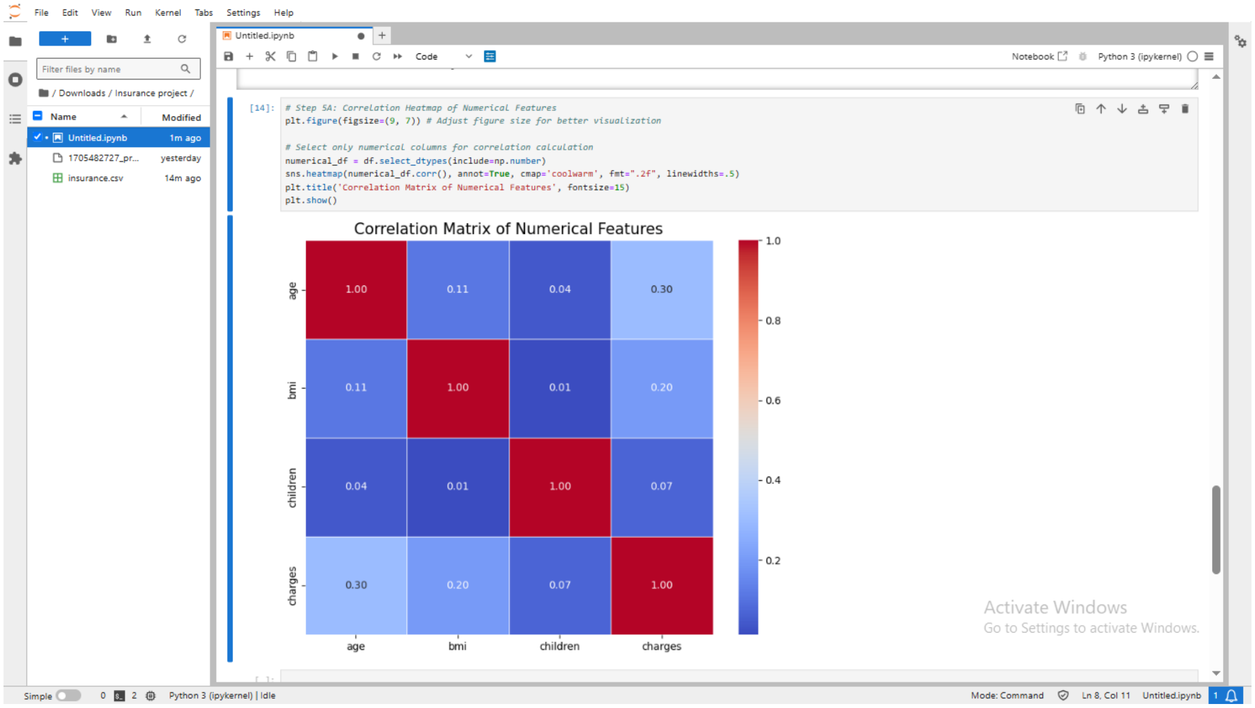


```
plt.title('Charges vs. BMI (Colored by Smoker Status)')
plt.xlabel('BMI') # Label can remain 'BMI' for readability in the plot
plt.ylabel('Charges')

plt.subplot(1, 3, 3) # 1 row, 3 columns, 3rd plot
sns.scatterplot(x='children', y='charges', data=df, hue='smoker', palette='coolwarm', alpha=0.7, s=50)
plt.title('Charges vs. Number of Children (Colored by Smoker Status)')
plt.xlabel('Number of Children')
plt.ylabel('Charges')

plt.tight_layout() # Adjust layout to prevent overlapping titles/labels
plt.show()
```

Step 5: Perform Data Visualization Using Plots of Feature vs. Feature

A. Correlation Matrix of Numerical Features

- The correlation heatmap shows moderate positive correlations between charges and age (~0.30) and bmi (~0.20), confirming observations from scatter plots.
- age, bmi, and children show very low linear correlation with each other, indicating no significant multicollinearity among these independent numerical features.



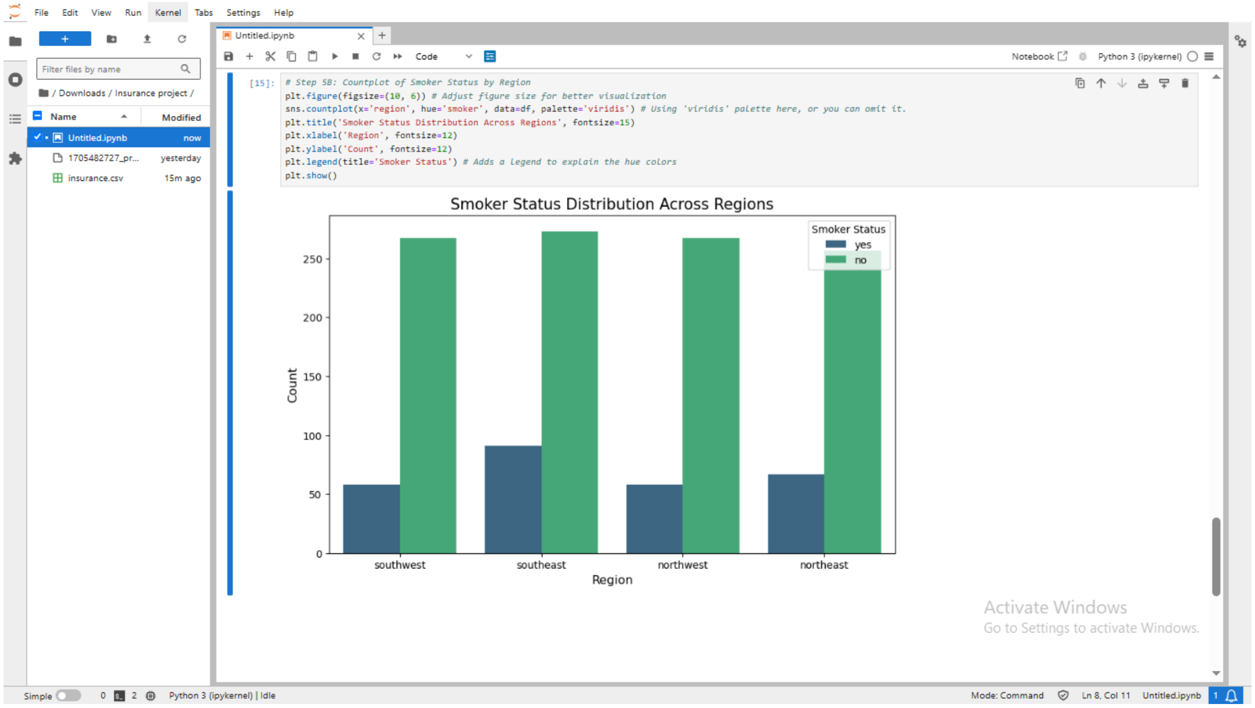
```
[14]: # Step 5A: Correlation Heatmap of Numerical Features
plt.figure(figsize=(9, 7)) # Adjust figure size for better visualization

# Select only numerical columns for correlation calculation
numerical_df = df.select_dtypes(include=np.number)
sns.heatmap(numerical_df.corr(), annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Correlation Matrix of Numerical Features', fontsize=15)
plt.show()
```

Correlation Matrix of Numerical Features

	age	bmi	children	charges
age	1.00	0.11	0.04	0.30
bmi	0.11	1.00	0.01	0.20
children	0.04	0.01	1.00	0.07
charges	0.30	0.20	0.07	1.00

B. Categorical Feature Relationships (Smoker Status by Region)

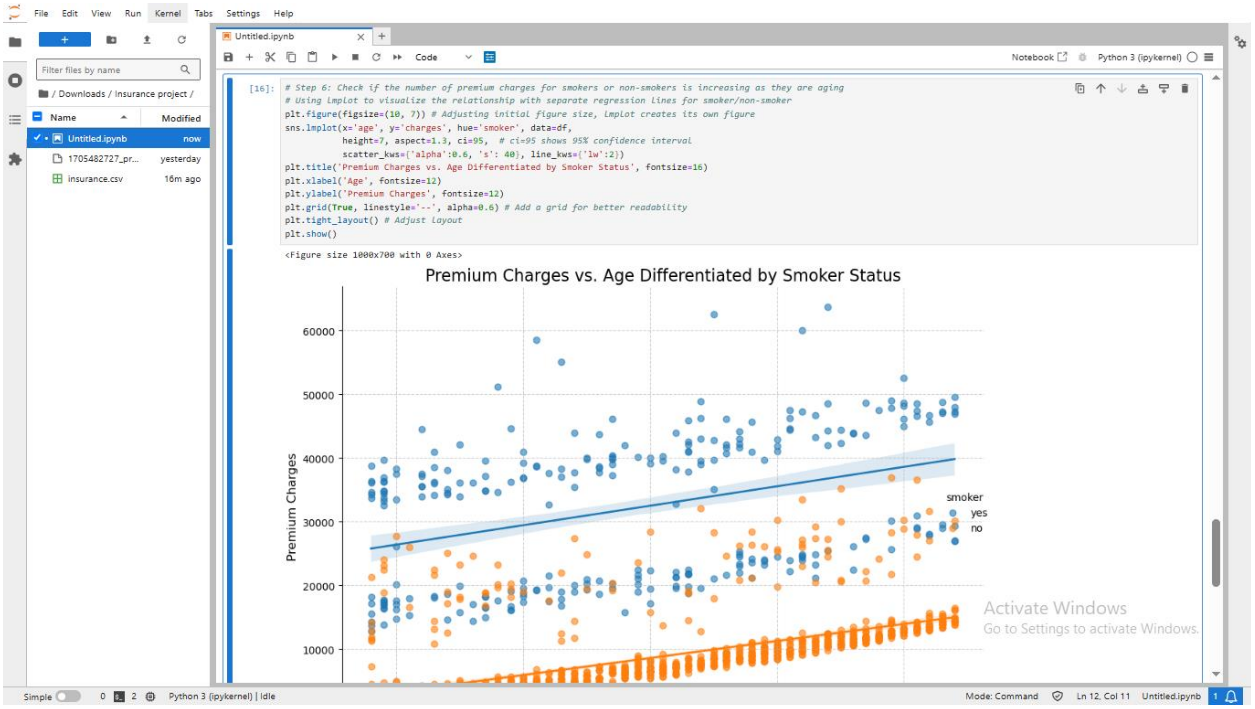
- The proportion of smokers to non-smokers is relatively consistent across all four geographical regions, meaning no particular region has a disproportionately higher or lower number of smokers.



```
[15]: # Step 5B: Countplot of Smoker Status by Region
plt.figure(figsize=(10, 6)) # Adjust figure size for better visualization
sns.countplot(x='region', hue='smoker', data=df, palette='viridis') # Using 'viridis' palette here, or you can omit it.
plt.title('Smoker Status Distribution Across Regions', fontsize=15)
plt.xlabel('Region', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.legend(title='Smoker Status') # Adds a legend to explain the hue colors
plt.show()
```

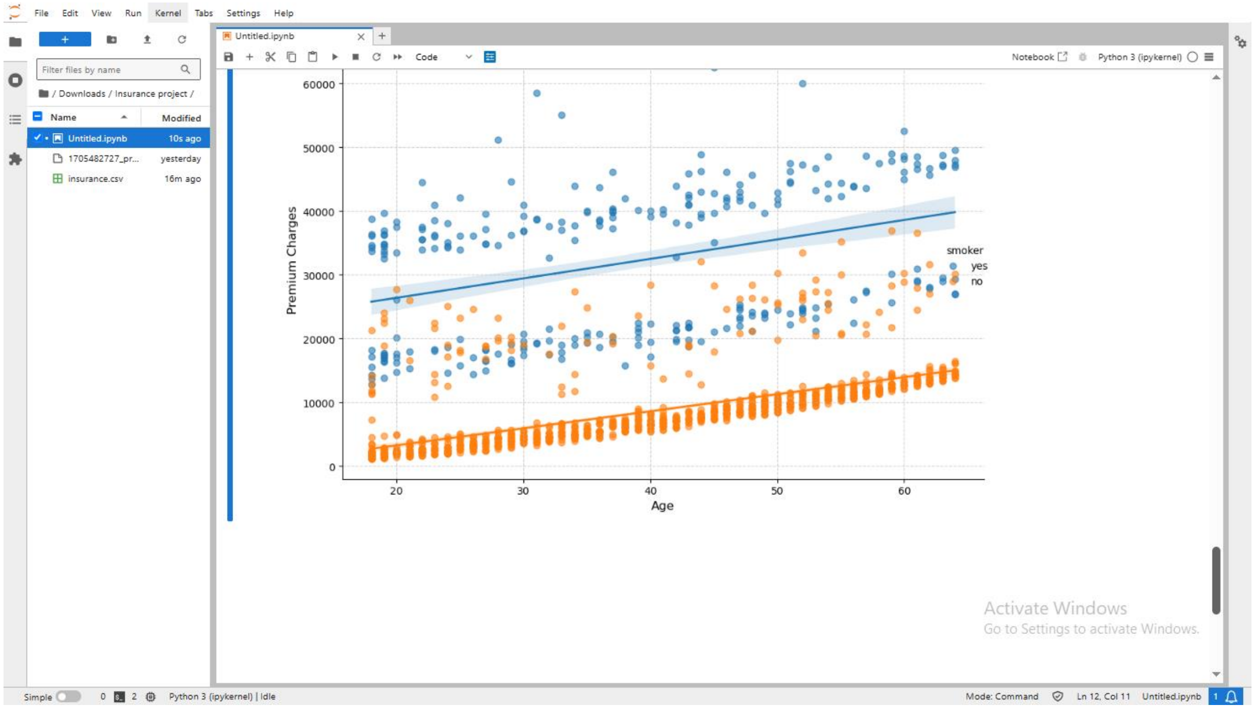
Step 6: Check if the Number of Premium Charges for Smokers or Non-Smokers is Increasing as they are Aging

- The lmplot clearly demonstrates two distinct groups of charges based on smoker status.
- **Smokers:** Individuals who smoke consistently face significantly higher premium charges across all age groups compared to non-smokers.
- **Age Trend:** Within both the smoker and non-smoker groups, there is a clear positive linear trend: as age increases, the premium charges also increase.
- **Slope Comparison:** The regression line for smokers appears to have a slightly steeper slope than that for non-smokers, suggesting that the increase in charges with age might be somewhat more pronounced for smokers.



```
[16]: # Step 6: Check if the number of premium charges for smokers or non-smokers is increasing as they are aging
      # Using lmplot to visualize the relationship with separate regression lines for smoker/non-smoker
      plt.figure(figsize=(10, 7)) # Adjusting initial figure size, lmplot creates its own figure
      sns.lmplot(x='age', y='charges', hue='smoker', data=df,
                 height=7, aspect=1.3, ci=95,  # ci=95 shows 95% confidence interval
                 scatter_kws={'alpha':0.6, 's': 40}, line_kws={'lw':2})
      plt.title('Premium Charges vs. Age Differentiated by Smoker Status', fontsize=16)
      plt.xlabel('Age', fontsize=12)
      plt.ylabel('Premium Charges', fontsize=12)
      plt.grid(True, linestyle='--', alpha=0.6) # Add a grid for better readability
      plt.tight_layout() # Adjust layout
      plt.show()
```

```
<Figure size 1000x700 with 0 Axes>
```



Conclusion (Summary of Key Findings)

This exploratory data analysis of the insurance dataset has provided valuable insights into the factors influencing medical insurance premiums. My/Our analysis revealed several key determinants:

- **Smoker Status** is the most significant predictor of insurance charges. I/We observed that smokers consistently incur substantially higher premiums compared to non-smokers, a difference that holds true across all age groups. This finding strongly supports the agency's initial hypothesis regarding the impact of smoking on chronic disease risk and associated healthcare costs.
- **Age** exhibits a clear positive linear relationship with insurance charges. As individuals age, their premiums tend to increase, a trend I/we found evident in both smoker and non-smoker groups, albeit with potentially a slightly steeper increase for smokers.
- **BMI** shows a positive correlation with charges, indicating that higher body mass index can lead to increased premiums, though this relationship is less linear and more dispersed than age.
- Features such as **sex, number of children**, and **geographical region** showed less direct or significant impact on the magnitude of insurance charges compared to smoker status and age. While minor variations exist, they are not as pronounced.

Overall, the dataset was found to be clean and complete, providing a robust foundation for further analysis. These insights are crucial for ABC Insurance to understand their current premium structure and identify key risk factors among their policyholders.

Next Steps / Future Work

The insights gained from this exploratory data analysis lay a strong foundation for building a predictive model for medical insurance charges. The logical next steps for this project would involve:

1. **Data Preprocessing:**
   - **Encoding Categorical Features:** Convert categorical variables such as 'sex', 'smoker', and 'region' into numerical representations (e.g., using One-Hot Encoding) so they can be processed by machine learning algorithms.
   - **Feature Scaling (Optional but Recommended):** Scale numerical features like 'age' and 'bmi' to ensure that no single feature dominates the model due to its larger numerical range.
2. **Model Building:** Select and train appropriate regression models (e.g., Linear Regression, Decision Trees, Random Forests, Gradient Boosting) using the prepared dataset.
3. **Model Evaluation:** Evaluate the performance of the trained model(s) using relevant metrics (e.g., R-squared, Mean Absolute Error, Root Mean Squared Error) to determine its accuracy and predictive power in forecasting medical insurance costs.

By following these steps, ABC Insurance can develop a robust model to predict insurance premiums, further enhancing their business decision-making and operational efficiency.