Insurance Data Analysis

Problem Statement:
An insurance agency, ABC Insurance, has a large dataset containing information about their policyholders and claims. They want to perform exploratory data analysis (EDA) on this dataset to gain insights that can help them make better business decisions and improve their operations.

The agency wants to analyze the different body types and the environment that affect the premium. The disease's effect or the cost of treatment differs depending on the circumstances. For example, a smoker's medical insurance premium may be higher than that of a healthy person, because smokers are more likely to develop chronic diseases. The agency wants to analyze the data to research healthcare premium costs.

Objective: To analyze the dataset that will help to create a model that will predict the cost of medical insurance based on various input features

Domain: Healthcare

Dataset: insurance dataset (insurance.csv)

Dataset Description:

age	Age of the person
sex	Female or Male
BMI	BMI value to estimate an individual's health and fitness condition
children	number of children (1,2,3,4, or 5)
smoker	The person is a smoker or not
region	Specifies the region (northeast, northwest, southeast, southwest)
charges	the amount of insurance

Steps to Be Followed:

1. Import libraries such as Pandas, matplotlib, NumPy, and seaborn and load the insurance dataset

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv("insurance.csv")
```

2. Check the shape of the data along with the data types of the column

```
df.shape
```

```
(1338, 7)
```

```
df.info()
```

```
<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
```

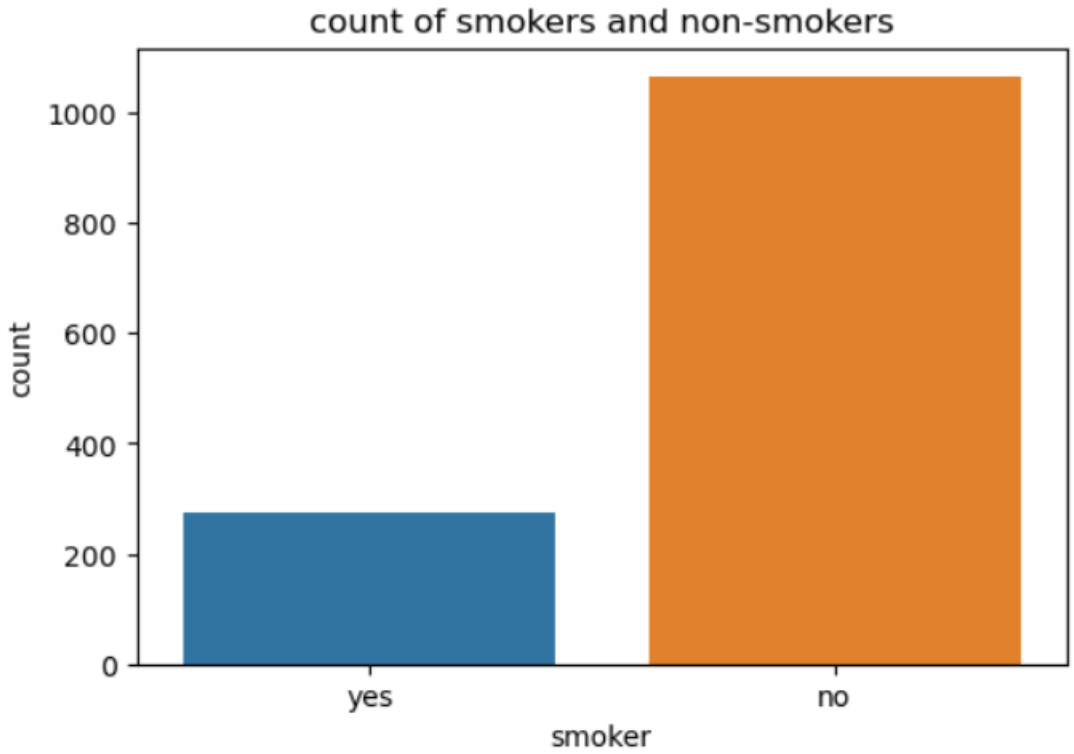
3. Check missing values in the dataset and find the appropriate measures to fill in the missing values

```
df.isna().sum()
```

```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

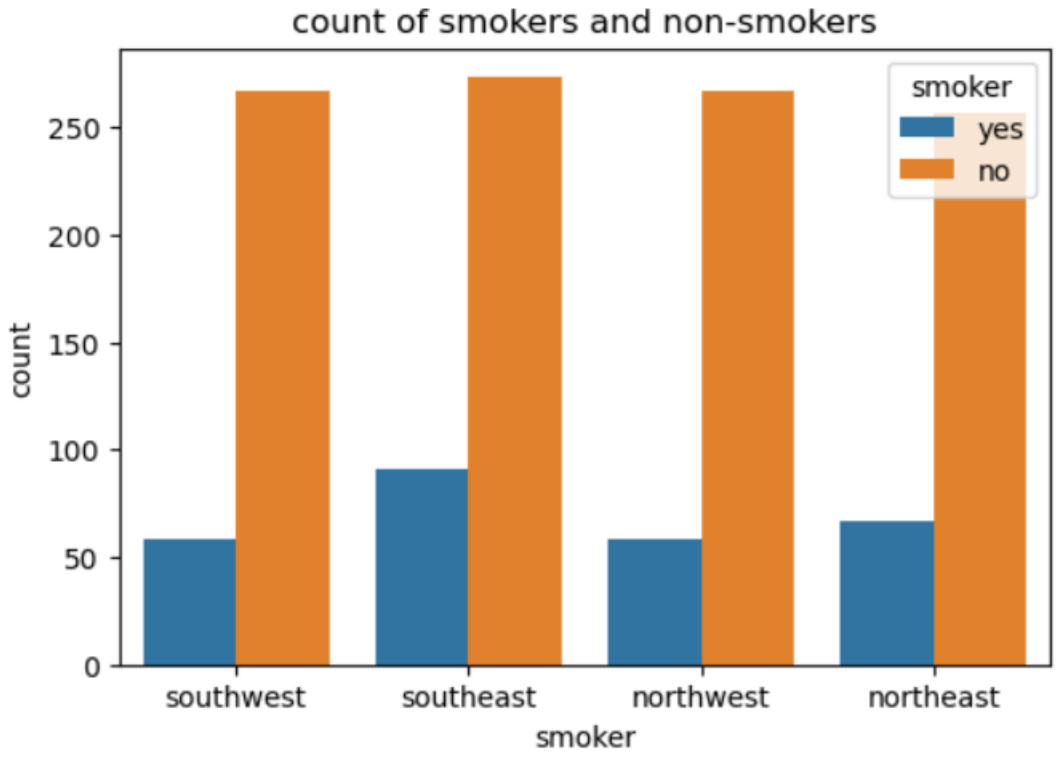
4. Explore the relationship between the feature and target column using a count plot of categorical columns and a scatter plot of numerical columns

```
#count plot for categorical columns : sex--- charges
plt.figure(figsize=(6,4))
sns.countplot(x="smoker",data=df)
plt.title("count of smokers and non-smokers")
plt.xlabel("smoker")
plt.ylabel("count")
plt.show()
```



Observation: This count plot displays the number of smokers versus non-smokers in the dataset. It's immediately evident that the number of non-smokers far exceeds the number of smokers. This imbalance suggests that smoking is relatively less prevalent in this dataset's population, which might mirror real-world insurance data trends where most policyholders are non-smokers. However, it's crucial to remember that even with fewer smokers, their health insurance charges and risks might be significantly different, as we'll see in later plots.

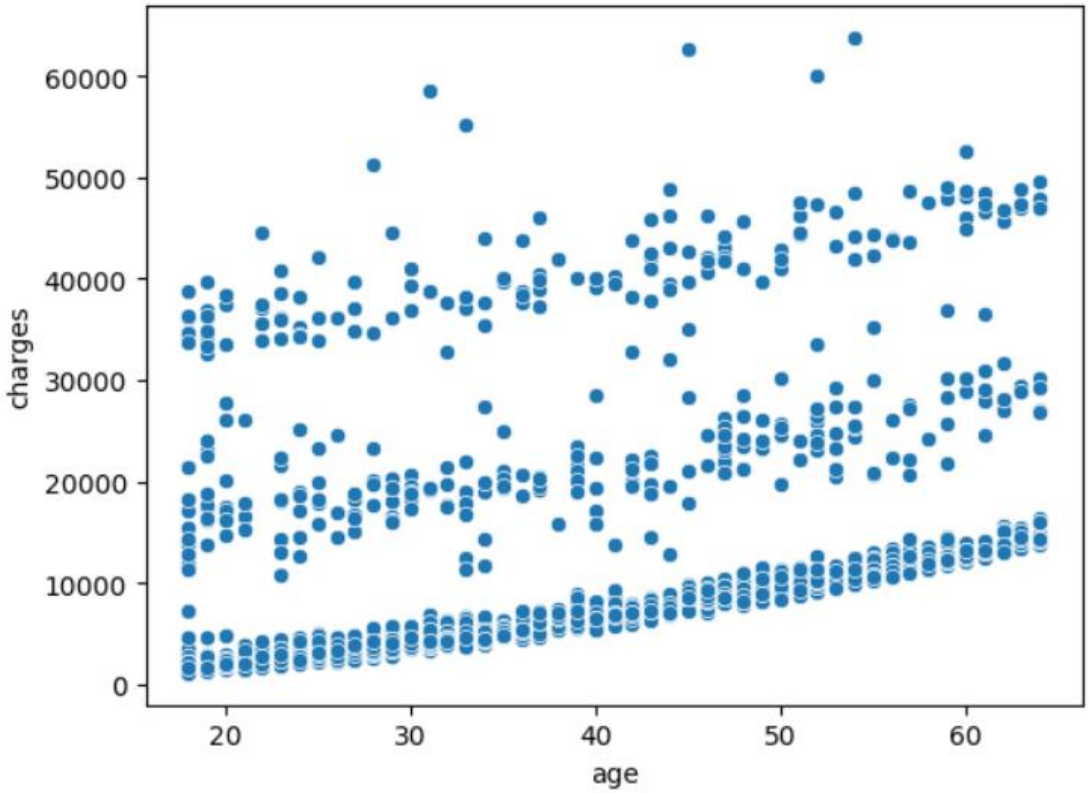
```
plt.figure(figsize=(6,4))
sns.countplot(x="region",data=df,hue="smoker")
plt.title("count of smokers and non-smokers")
plt.xlabel("smoker")
plt.ylabel("count")
plt.show()
```



Observation: This count plot breaks down smokers and non-smokers across different regions. It shows that across all four regions (northeast, northwest, southeast, southwest), non-smokers consistently outnumber smokers. Interestingly, the southeast region appears to have the highest number of smokers compared to the other regions. This could be relevant for insurers or public health policymakers to identify regional health behavior patterns. Despite the regional variation, the overall trend remains clear — smoking is less common than non-smoking in all areas.

```
#scatter plot of numerical columns
sns.scatterplot(x="age", y='charges', data=df)
```

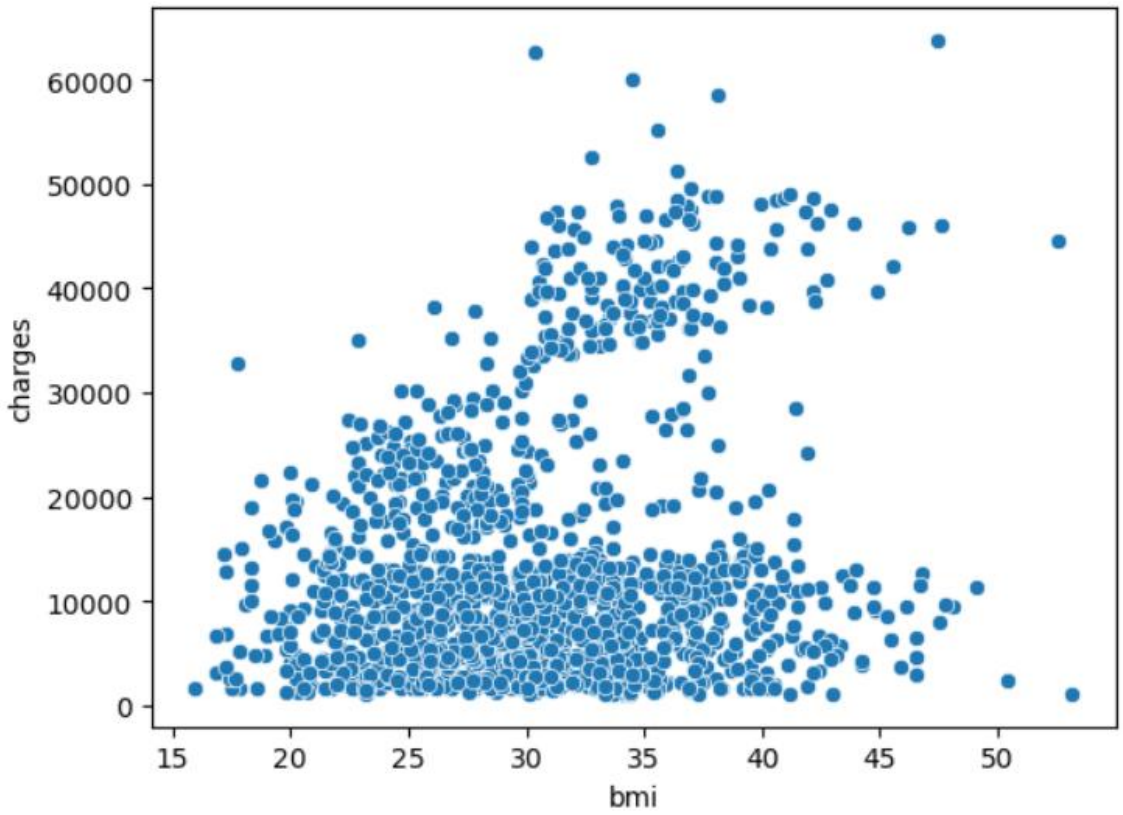
```
<Axes: xlabel='age', ylabel='charges'>
```



Observation: In this scatter plot of age versus insurance charges, a positive correlation can be observed: as age increases, the insurance charges tend to rise. However, there's noticeable dispersion, with some younger individuals facing high charges as well. This likely hints at other influencing factors, such as smoking status, BMI, or health conditions. The charges spike especially for older individuals, suggesting age is a significant risk factor in premium calculations.

```
sns.scatterplot(x="bmi", y='charges', data=df)
```

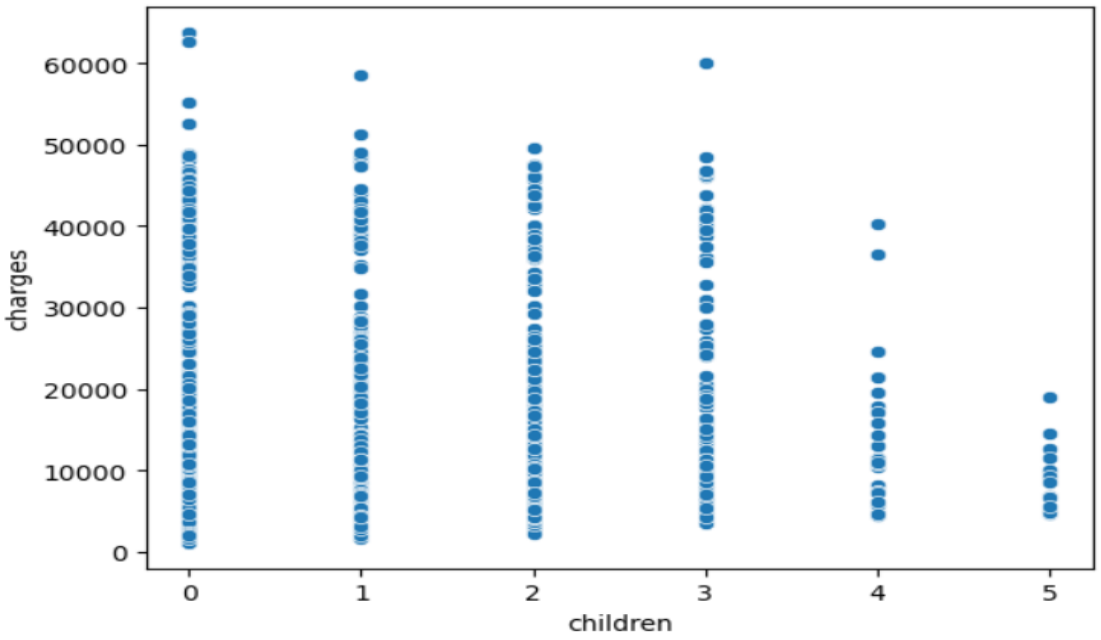
```
<Axes: xlabel='bmi', ylabel='charges'>
```



Observation: This plot illustrates the relationship between BMI and insurance charges. Unlike the age plot, there isn't a strong linear relationship here. Most individuals with a BMI below 30 cluster around lower charges. However, a group of outliers with high charges exists across various BMI levels, suggesting factors like smoking or age play a more pivotal role than BMI alone in determining premiums. Some individuals with higher BMI also incur higher charges, but the correlation isn’t consistent.

```
: sns.scatterplot(x="children", y='charges', data=df)
```

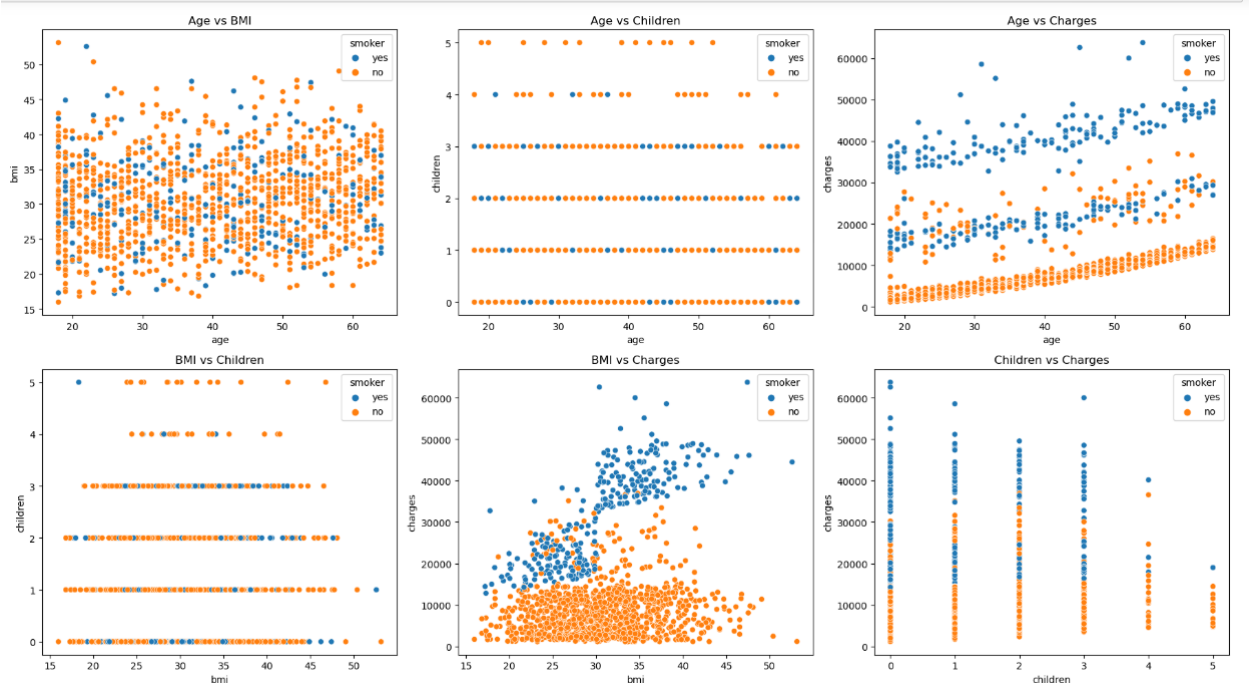
```
: <Axes: xlabel='children', ylabel='charges'>
```



Observation: The scatter plot of the number of children versus insurance charges shows no clear pattern or correlation. Charges seem widely dispersed regardless of how many children an individual has. Most data points cluster below a certain charge value, irrespective of the number of children, indicating that this variable might have a minimal direct effect on insurance charges.

5. Perform data visualization using plots of feature vs feature

```
: fig, axs = plt.subplots(2, 3, figsize=(18, 10))

sns.scatterplot(x="age", y="bmi", hue="smoker", data=df, ax=axs[0, 0])
axs[0, 0].set_title("Age vs BMI")

sns.scatterplot(x="age", y="children", hue="smoker", data=df, ax=axs[0, 1])
axs[0, 1].set_title("Age vs Children")

sns.scatterplot(x="age", y="charges", hue="smoker", data=df, ax=axs[0, 2])
axs[0, 2].set_title("Age vs Charges")

sns.scatterplot(x="bmi", y="children", hue="smoker", data=df, ax=axs[1, 0])
axs[1, 0].set_title("BMI vs Children")

sns.scatterplot(x="bmi", y="charges", hue="smoker", data=df, ax=axs[1, 1])
axs[1, 1].set_title("BMI vs Charges")

sns.scatterplot(x="children", y="charges", hue="smoker", data=df, ax=axs[1, 2])
axs[1, 2].set_title("Children vs Charges")

plt.tight_layout()
```



Observation: Age vs BMI: No obvious trend between age and BMI for both smokers and non-smokers. Both categories appear evenly spread across the age and BMI ranges, indicating independence between these two variables.

Age vs Children: Similar lack of relationship here. The number of children doesn't show a trend with age for either group. Most individuals, irrespective of smoking status, have 0–2 children.

Age vs Charges: A striking difference emerges. Smokers consistently incur significantly higher charges than non-smokers, especially as age increases. Charges for smokers escalate rapidly with age, creating a clear upward trend, while non-smokers' charges increase more gradually.

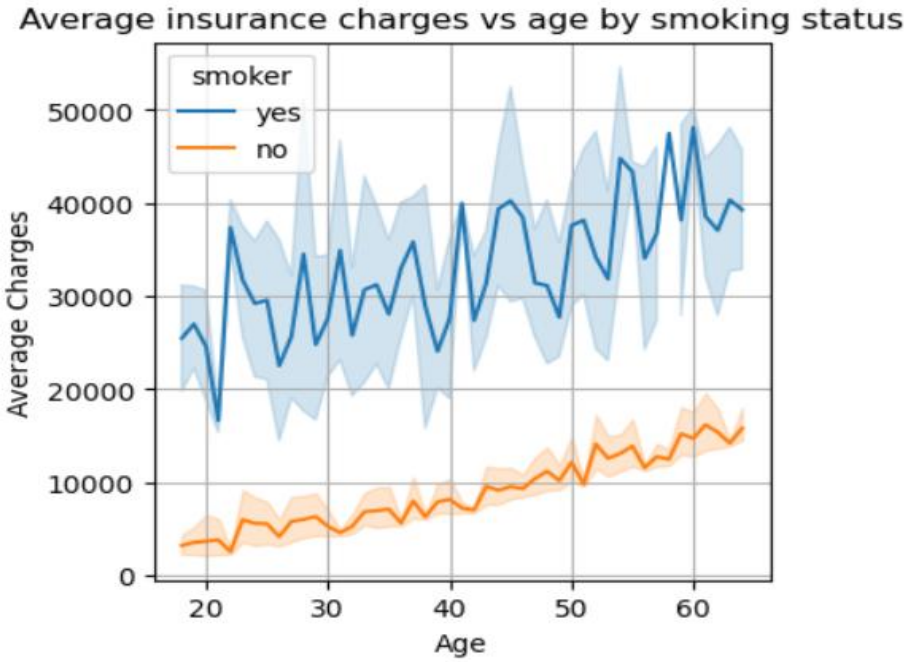
BMI vs Children: Again, no clear relationship is visible for either group, suggesting these two variables are independent.

BMI vs Charges: For smokers, insurance charges remain higher at nearly all BMI levels, especially beyond a BMI of 30. Non-smokers' charges increase modestly with BMI but never reach the high levels seen for smokers, reinforcing the heightened risk insurers associate with smoking.

Children vs Charges: Number of children has a minimal impact on charges for both groups, though smokers generally face higher charges regardless of how many children they have.

6. Check if the number of premium charges for smokers or non-smokers is increasing as they are aging.

```
plt.figure(figsize=(4,4))
sns.lineplot(data=data,x="age", y ="charges" , hue="smoker",estimator="mean")
plt.title("Average insurance charges vs age by smoking status")
plt.xlabel("Age")
plt.ylabel("Average Charges")
plt.grid(True)
plt.show()
```



Observation: The line plot shows that average insurance charges increase with age for both smokers and non-smokers, but the rise is much steeper for smokers. Smokers consistently face higher premium charges starting from age 18, with a significant upward trend as they age. By age 60+, their charges often exceed 40,000 to 50,000 units, over three times higher than non-smokers in the same age group.

Non-smokers experience a gradual and steady increase in charges, reaching around 15,000–20,000 units by age 60. The variability in charges is also greater among smokers, likely due to additional health risk factors, while non-smokers show more consistent premium patterns.

Overall, the plot highlights how smoking dramatically amplifies the impact of aging on insurance costs, reflecting the compounded health risks and financial implications for insurers.