Project Report: Premium Pricing Optimization

1. Introduction

The goal of this project is to analyze the **insurance dataset** to understand the factors influencing insurance charges and optimize premium pricing. The dataset contains information about policyholders, including their age, sex, BMI, number of children, smoking status, region, and insurance charges. The analysis aims to identify key drivers of insurance costs and provide insights for dynamic pricing models.

2. Dataset Overview

The dataset contains **1338 rows** and **7 columns**:

- **age**: Age of the policyholder.
- **sex**: Gender of the policyholder (male/female).
- **bmi**: Body Mass Index (BMI) of the policyholder.
- **children**: Number of children/dependents covered by the insurance.
- **smoker**: Smoking status of the policyholder (yes/no).
- **region**: Region where the policyholder resides (southeast, southwest, northeast, northwest).
- **charges**: Insurance charges billed to the policyholder.

3. Exploratory Data Analysis (EDA)

3.1. Summary Statistics

- **Age**: Ranges from 18 to 64 years, with an average age of 39.2.
- **BMI**: Ranges from 15.96 to 53.13, with an average of 30.66.
- **Children**: Ranges from 0 to 5, with an average of 1.09.
- **Charges**: Ranges from 1,121.87to1,121.87*to*63,770.43, with an average of $13,270.42.

3.2. Missing Values

- There are **no missing values** in the dataset.

3.3. Distribution of Charges

- The distribution of insurance charges is **right-skewed**, indicating that most policyholders have lower charges, while a few have significantly higher charges.
- The histogram shows a peak around 5,000,withalongtailextendingbeyond5,000,*withalongtailextendingbeyond*50,000.

3.4. Relationship Between Age and Charges

- There is a **positive correlation** between age and charges, indicating that older policyholders tend to have higher insurance costs.
- Smokers generally have higher charges compared to non-smokers across all age groups.

3.5. Relationship Between BMI and Charges

- Policyholders with higher BMI tend to have higher charges, especially if they are smokers.
- The scatter plot shows a cluster of high charges for smokers with BMI above 30.

3.6. Charges by Region

- The **southeast** region has the highest average charges, while the **southwest** region has the lowest.
- The box plot shows that the distribution of charges varies across regions, with the southeast region having more outliers.

3.7. Charges by Smoking Status

- Smokers have significantly higher charges compared to non-smokers.
- The box plot shows a clear distinction between the two groups, with smokers having a much wider range of charges.

3.8. Correlation Matrix

- The correlation matrix reveals that **age** and **BMI** have a moderate positive correlation with charges.
- **Smoking status** is likely a strong predictor of charges, but it is not captured in the correlation matrix due to its categorical nature.

3.9. Pairplot

- The pairplot visualizes pairwise relationships between numerical variables (age, BMI, children, charges) and highlights the impact of smoking status on charges.

4. Grouped Analysis

4.1. Mean Charges by Smoking Status

- **Smokers**: $32,050.23
- **Non-smokers**: $8,434.27
- Smokers pay **almost 4 times more** than non-smokers on average.

4.2. Mean Charges by Region

- **southeast**: $14,735.41
- **southwest**: $12,346.94
- **northeast**: $13,406.38
- **northwest**: $12,417.58
- The **southeast** region has the highest average charges.

4.3. Mean Charges by Number of Children

- **0 children**: $12,365.98
- **1 child**: $12,715.56
- **2 children**: $15,073.32
- **3 children**: $15,325.28
- **4 children**: $13,858.94
- **5 children**: $8,786.00
- Policyholders with **2 or 3 children** tend to have higher charges.

4.4. Mean Charges by Sex

- **Male**: $13,956.75
- **Female**: $12,569.58
- Males have slightly higher charges on average.

4.5. Mean Charges by Age

- Charges increase with age, with the highest charges for policyholders in their 60s.

4.6. Mean Charges by BMI

- Charges increase with BMI, especially for policyholders with a BMI above 30.

5. Key Insights

1. **Smoking Status**: Smoking is the most significant factor influencing insurance charges. Smokers pay significantly higher premiums than non-smokers.
2. **Age**: Older policyholders tend to have higher charges, likely due to increased health risks.
3. **BMI**: Higher BMI is associated with higher charges, particularly for smokers.
4. **Region**: The **southeast** region has the highest average charges, possibly due to higher healthcare costs or risk factors.
5. **Children**: Policyholders with **2 or 3 children** have higher charges, possibly due to additional dependents.

6. Recommendations

1. **Dynamic Pricing Model**:
    - Incorporate **smoking status**, **age**, and **BMI** as key variables in the pricing model.
    - Offer discounts or incentives for non-smokers and policyholders with lower BMI.
2. **Regional Adjustments**:
    - Adjust premiums based on regional healthcare costs and risk factors.
    - Consider offering lower premiums in regions with lower average charges (e.g., southwest).
3. **Family Plans**:
    - Develop family plans with adjusted premiums for policyholders with children.
4. **Wellness Programs**:
    - Introduce wellness programs to encourage healthy lifestyles (e.g., smoking cessation, weight management).

7. Conclusion

The analysis highlights the key factors influencing insurance charges and provides actionable insights for optimizing premium pricing. By leveraging these insights, insurers can develop

dynamic pricing models that balance risk and profitability while offering competitive premiums to policyholders.

8. Next Steps

- Build predictive models (e.g., regression, machine learning) to estimate charges based on policyholder characteristics.
- Conduct further analysis to explore interactions between variables (e.g., age and smoking status).
- Validate the pricing model with real-world data and refine it based on feedback.

9. Visualizations

- **Distribution of Charges**: Right-skewed distribution with a peak around $5,000.
- **Age vs Charges**: Positive correlation, with smokers having higher charges.
- **BMI vs Charges**: Higher BMI leads to higher charges, especially for smokers.
- **Charges by Region**: Southeast region has the highest average charges.
- **Charges by Smoking Status**: Smokers have significantly higher charges.

10. Appendix

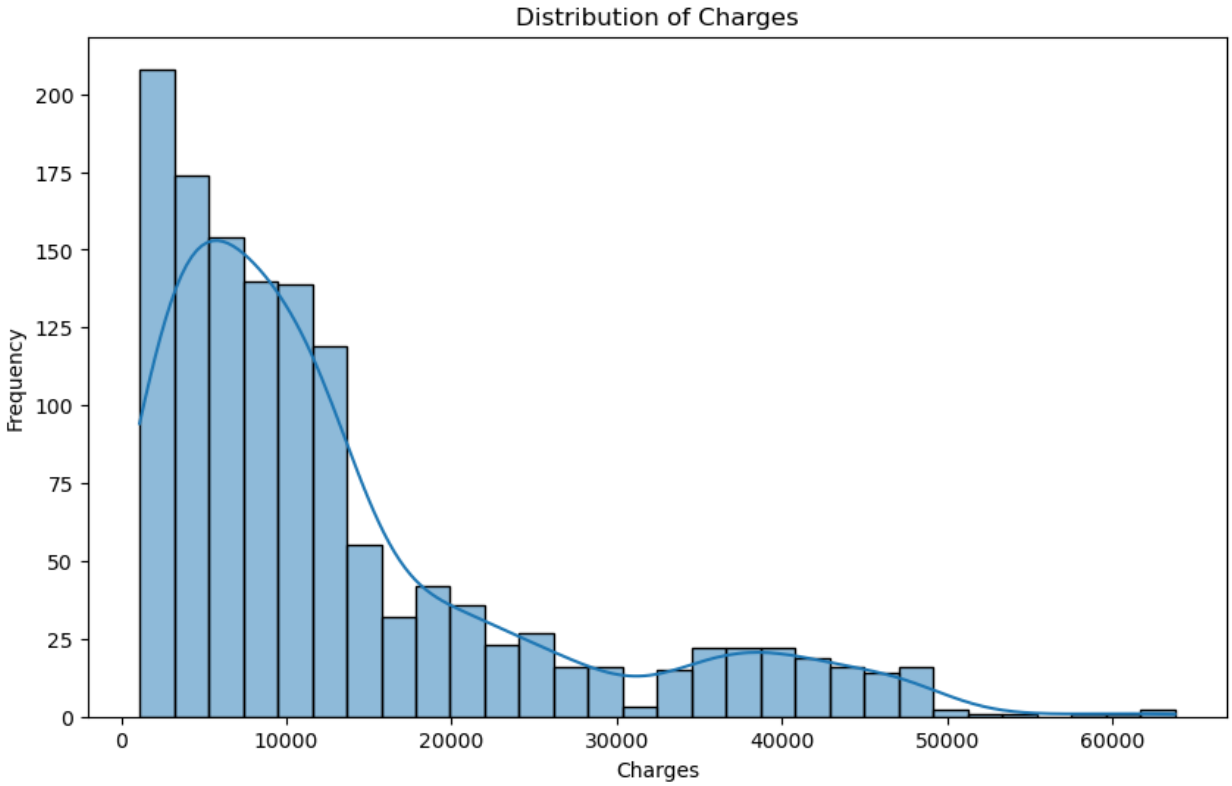
- **Dataset**: insurance.csv
- **Tools Used**: Python, Pandas, NumPy, Matplotlib, Seaborn
- **Code**: Provided in the project file.

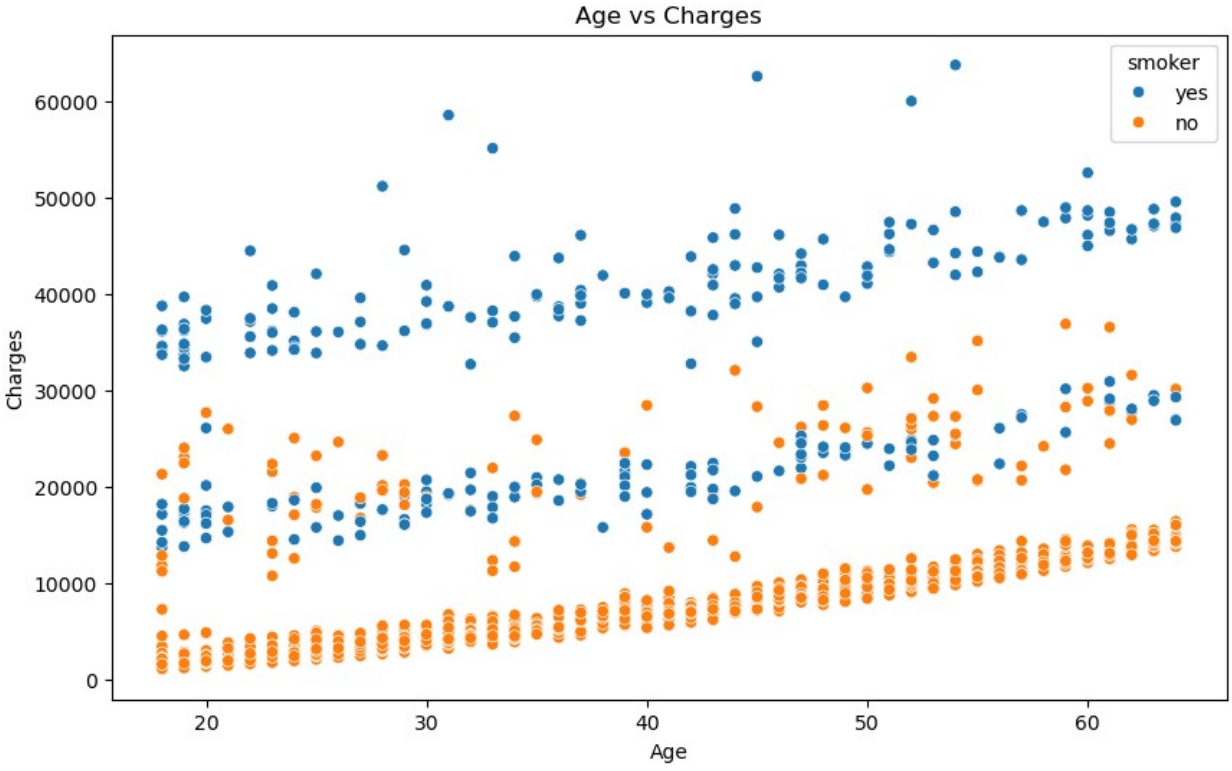
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv('insurance.csv')

# Display the first few rows of the dataset
print(df.head())

# Get basic information about the dataset
print(df.info())

# Get summary statistics
print(df.describe())

# Check for missing values
print(df.isnull().sum())

# Visualize the distribution of charges
plt.figure(figsize=(10, 6))
sns.histplot(df['charges'], bins=30, kde=True)
plt.title('Distribution of Charges')
plt.xlabel('Charges')
plt.ylabel('Frequency')
plt.show()

# Visualize the relationship between age and charges
plt.figure(figsize=(10, 6))
sns.scatterplot(x='age', y='charges', data=df, hue='smoker')
plt.title('Age vs Charges')
plt.xlabel('Age')
plt.ylabel('Charges')
plt.show()

# Visualize the relationship between BMI and charges
plt.figure(figsize=(10, 6))
sns.scatterplot(x='bmi', y='charges', data=df, hue='smoker')
plt.title('BMI vs Charges')
plt.xlabel('BMI')
plt.ylabel('Charges')
plt.show()

# Visualize the distribution of charges by region
plt.figure(figsize=(10, 6))
sns.boxplot(x='region', y='charges', data=df)
plt.title('Charges by Region')
plt.xlabel('Region')
plt.ylabel('Charges')
```

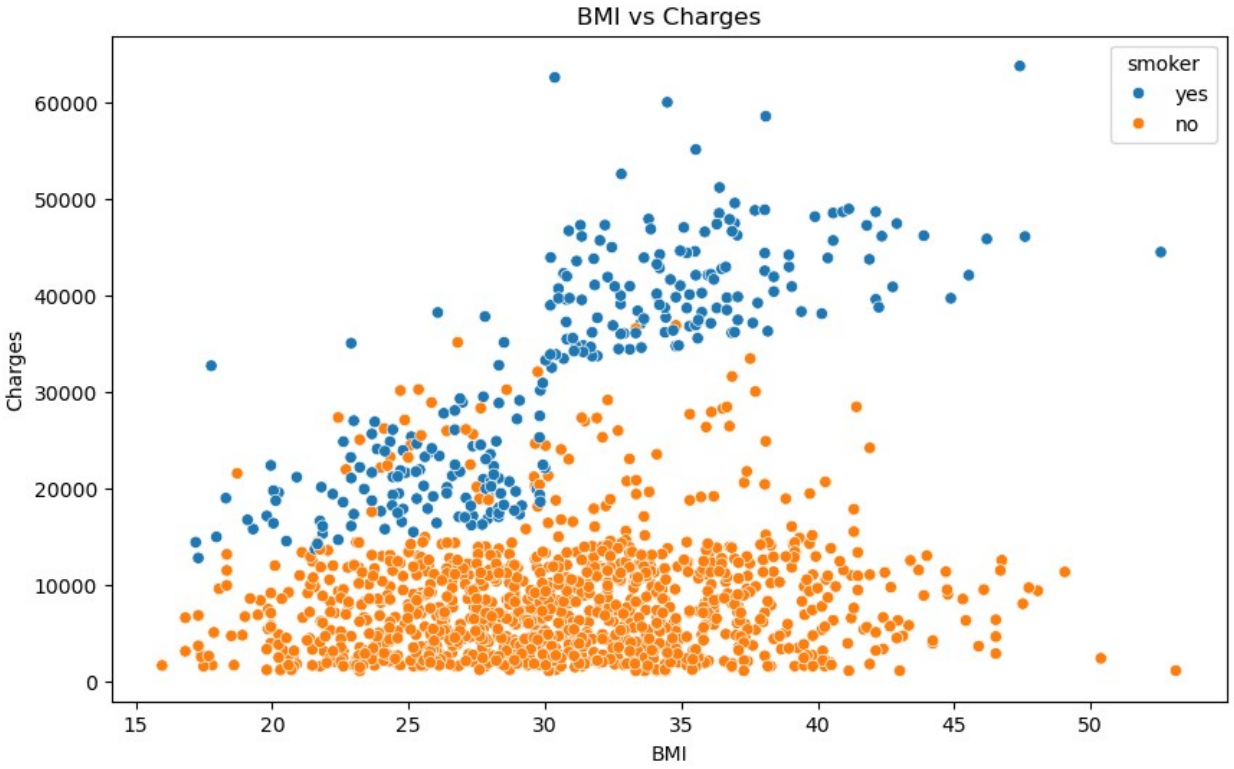
```
plt.show()

# Visualize the distribution of charges by smoking status
plt.figure(figsize=(10, 6))
sns.boxplot(x='smoker', y='charges', data=df)
plt.title('Charges by Smoking Status')
plt.xlabel('Smoker')
plt.ylabel('Charges')
plt.show()

# Correlation matrix to understand relationships between numerical
variables
corr_matrix = df.corr()
plt.figure(figsize=(10, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()

# Pairplot to visualize pairwise relationships in the dataset
sns.pairplot(df, hue='smoker')
plt.show()

# Group by smoking status and calculate mean charges
smoker_charges = df.groupby('smoker')['charges'].mean()
print(smoker_charges)

# Group by region and calculate mean charges
region_charges = df.groupby('region')['charges'].mean()
print(region_charges)

# Group by number of children and calculate mean charges
children_charges = df.groupby('children')['charges'].mean()
print(children_charges)

# Group by sex and calculate mean charges
sex_charges = df.groupby('sex')['charges'].mean()
print(sex_charges)

# Group by age and calculate mean charges
age_charges = df.groupby('age')['charges'].mean()
print(age_charges)

# Group by BMI and calculate mean charges
bmi_charges = df.groupby('bmi')['charges'].mean()
print(bmi_charges)
```

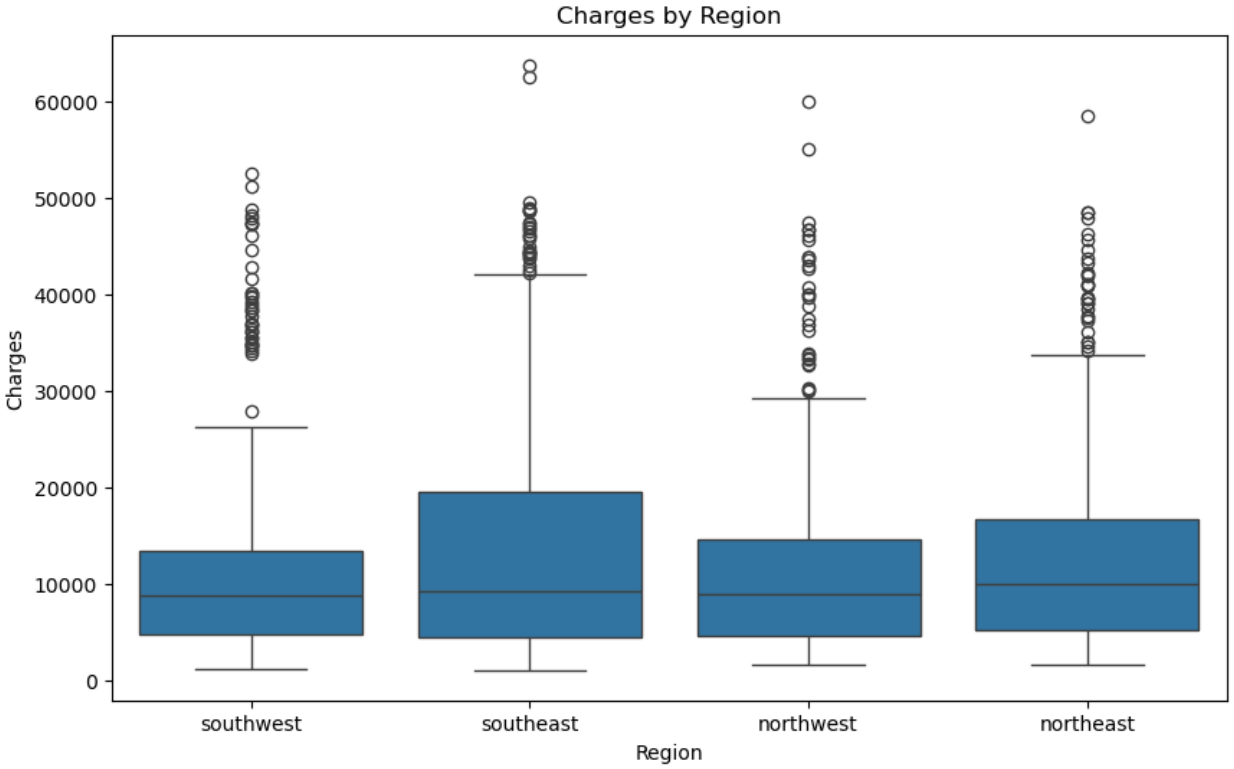
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200

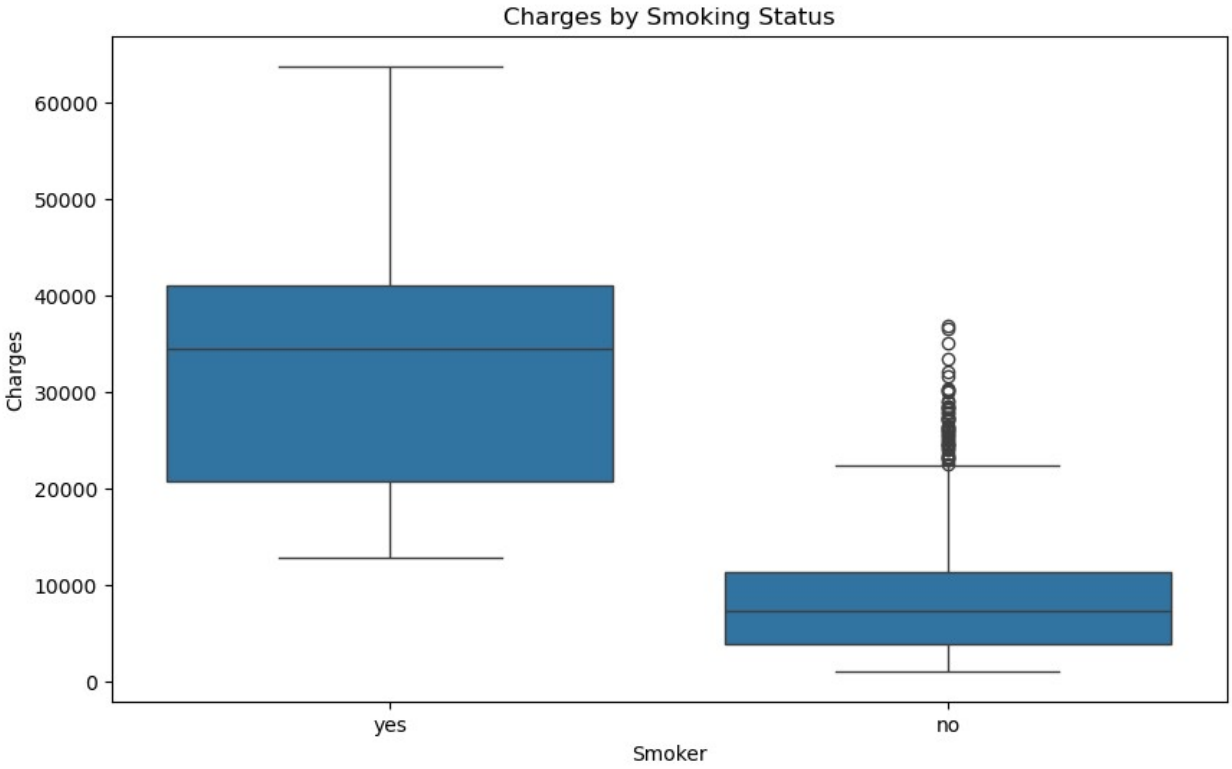
```
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 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
None
                age          bmi     children       charges
count  1338.000000  1338.000000  1338.000000   1338.000000
mean     39.207025    30.663397     1.094918  13270.422265
std      14.049960     6.098187     1.205493  12110.011237
min      18.000000    15.960000     0.000000   1121.873900
25%      27.000000    26.296250     0.000000   4740.287150
50%      39.000000    30.400000     1.000000   9382.033000
75%      51.000000    34.693750     2.000000  16639.912515
max      64.000000    53.130000     5.000000  63770.428010
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```


Distribution of Charges
Frequency
Charges
0
25
50
75
100
125
150
175
200
0
10000
20000
30000
40000
50000
60000


Age vs Charges
smoker
yes
no
Charges
Age
0
10000
20000
30000
40000
50000
60000
20
30
40
50
60


BMI vs Charges
smoker
yes
no
Charges
BMI
0
10000
20000
30000
40000
50000
60000
15
20
25
30
35
40
45
50


Charges by Region
Charges
Region
0
10000
20000
30000
40000
50000
60000
southwest
southeast
northwest
northeast




```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
Cell In[1], line 62
     59 plt.show()
     61 # Correlation matrix to understand relationships between numerical variables
---> 62 corr_matrix = df.corr()
     63 plt.figure(figsize=(10, 6))
     64 sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')

File ~\anaconda3\Lib\site-packages\pandas\core\frame.py:11049, in DataFrame.corr(self, method, min_periods, numeric_only)
  11047 cols = data.columns
  11048 idx = cols.copy()
> 11049 mat = data.to_numpy(dtype=float, na_value=np.nan, copy=False)
  11051 if method == "pearson":
  11052     correl = libalgos.nancorr(mat, minp=min_periods)

File ~\anaconda3\Lib\site-packages\pandas\core\frame.py:1993, in DataFrame.to_numpy(self, dtype, copy, na_value)
   1991 if dtype is not None:
   1992     dtype = np.dtype(dtype)
-> 1993 result = self._mgr.as_array(dtype=dtype, copy=copy, na_value=na_value)
```

```
   1994 if result.dtype is not dtype:
   1995     result = np.asarray(result, dtype=dtype)

File ~\anaconda3\Lib\site-packages\pandas\core\internals\
managers.py:1694, in BlockManager.as_array(self, dtype, copy,
na_value)
   1692         arr.flags.writeable = False
   1693 else:
-> 1694     arr = self._interleave(dtype=dtype, na_value=na_value)
   1695     # The underlying data was copied within _interleave, so no
need
   1696     # to further copy if copy=True or setting na_value
   1698 if na_value is lib.no_default:

File ~\anaconda3\Lib\site-packages\pandas\core\internals\
managers.py:1753, in BlockManager._interleave(self, dtype, na_value)
   1751     else:
   1752         arr = blk.get_values(dtype)
-> 1753     result[rl.indexer] = arr
   1754     itemmask[rl.indexer] = 1
   1756 if not itemmask.all():

ValueError: could not convert string to float: 'female'

pip install nbconvert

Requirement already satisfied: nbconvert in c:\users\rishi\anaconda3\
lib\site-packages (7.16.4)
Requirement already satisfied: beautifulsoup4 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (4.12.3)
Requirement already satisfied: bleach!=5.0.0 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (4.1.0)
Requirement already satisfied: defusedxml in c:\users\rishi\anaconda3\
lib\site-packages (from nbconvert) (0.7.1)
Requirement already satisfied: jinja2>=3.0 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (3.1.4)
Requirement already satisfied: jupyter-core>=4.7 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (5.7.2)
Requirement already satisfied: jupyterlab-pygments in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (0.1.2)
Requirement already satisfied: markupsafe>=2.0 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (2.1.3)
Requirement already satisfied: mistune<4,>=2.0.3 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (2.0.4)
Requirement already satisfied: nbclient>=0.5.0 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (0.8.0)
Requirement already satisfied: nbformat>=5.7 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (5.10.4)
Requirement already satisfied: packaging in c:\users\rishi\anaconda3\
lib\site-packages (from nbconvert) (24.1)
```

```
Requirement already satisfied: pandocfilters>=1.4.1 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (1.5.0)
Requirement already satisfied: pygments>=2.4.1 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (2.15.1)
Requirement already satisfied: tinycss2 in c:\users\rishi\anaconda3\
lib\site-packages (from nbconvert) (1.2.1)
Requirement already satisfied: traitlets>=5.1 in c:\users\rishi\
anaconda3\lib\site-packages (from nbconvert) (5.14.3)
Requirement already satisfied: six>=1.9.0 in c:\users\rishi\anaconda3\
lib\site-packages (from bleach!=5.0.0->nbconvert) (1.16.0)
Requirement already satisfied: webencodings in c:\users\rishi\
anaconda3\lib\site-packages (from bleach!=5.0.0->nbconvert) (0.5.1)
Requirement already satisfied: platformdirs>=2.5 in c:\users\rishi\
anaconda3\lib\site-packages (from jupyter-core>=4.7->nbconvert)
(3.10.0)
Requirement already satisfied: pywin32>=300 in c:\users\rishi\
anaconda3\lib\site-packages (from jupyter-core>=4.7->nbconvert)
(305.1)
Requirement already satisfied: jupyter-client>=6.1.12 in c:\users\
rishi\anaconda3\lib\site-packages (from nbclient>=0.5.0->nbconvert)
(8.6.0)
Requirement already satisfied: fastjsonschema>=2.15 in c:\users\rishi\
anaconda3\lib\site-packages (from nbformat>=5.7->nbconvert) (2.16.2)
Requirement already satisfied: jsonschema>=2.6 in c:\users\rishi\
anaconda3\lib\site-packages (from nbformat>=5.7->nbconvert) (4.23.0)
Requirement already satisfied: soupsieve>1.2 in c:\users\rishi\
anaconda3\lib\site-packages (from beautifulsoup4->nbconvert) (2.5)
Requirement already satisfied: attrs>=22.2.0 in c:\users\rishi\
anaconda3\lib\site-packages (from jsonschema>=2.6->nbformat>=5.7-
>nbconvert) (23.1.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
c:\users\rishi\anaconda3\lib\site-packages (from jsonschema>=2.6-
>nbformat>=5.7->nbconvert) (2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in c:\users\rishi\
anaconda3\lib\site-packages (from jsonschema>=2.6->nbformat>=5.7-
>nbconvert) (0.30.2)
Requirement already satisfied: rpds-py>=0.7.1 in c:\users\rishi\
anaconda3\lib\site-packages (from jsonschema>=2.6->nbformat>=5.7-
>nbconvert) (0.10.6)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\
rishi\anaconda3\lib\site-packages (from jupyter-client>=6.1.12-
>nbclient>=0.5.0->nbconvert) (2.9.0.post0)
Requirement already satisfied: pyzmq>=23.0 in c:\users\rishi\
anaconda3\lib\site-packages (from jupyter-client>=6.1.12-
>nbclient>=0.5.0->nbconvert) (25.1.2)
Requirement already satisfied: tornado>=6.2 in c:\users\rishi\
anaconda3\lib\site-packages (from jupyter-client>=6.1.12-
>nbclient>=0.5.0->nbconvert) (6.4.1)
Note: you may need to restart the kernel to use updated packages.
```