ANALYSIS OF US HEALTH INSURANCE DATA

BY BJÖRN MÜLLER

CONTENT

- Basic Data Cleaning
- Analysis of Data (mainly Distribution)
- Check of Charges
  - Check of Influence of the single Factors
  - Enhanced Analysis of the Influence of Weight (BMI) and Smoking
- Linear Regression
  - Additional Data Cleaning
  - Comparison of original and predicted Data

DATA CLEANSING

PART I

STEPS OF DATA CLEANSING

- ( ) Missing Values
- ( ) Duplicate Entries
- ( ) Outliers
- ( ) Inconsistent Formatting
- ( ) Incorrect Data Types
- ✓ String/Whitespace Issues
- ✓ Mismatched Data – short Excel check
- ( ) Scaling and Normalization Issues

Check and Preperation of Data

```
# Read in data
insurance_df = pd.read_csv("insurance.csv", sep =",")
# Check the first few rows of the dataset
print(insurance_df.head())
```

[107] ✓ 0.0s

```
...    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
```

STEPS OF DATA CLEANSING

- ✓ Missing Values
- ( ) Duplicate Entries
- ( ) Outliers
- ✓ Inconsistent Formatting
- ✓ Incorrect Data Types
- ✓ String/Whitespace Issues
- ✓ Mismatched Data – short Excel check
- ( ) Scaling and Normalization Issues

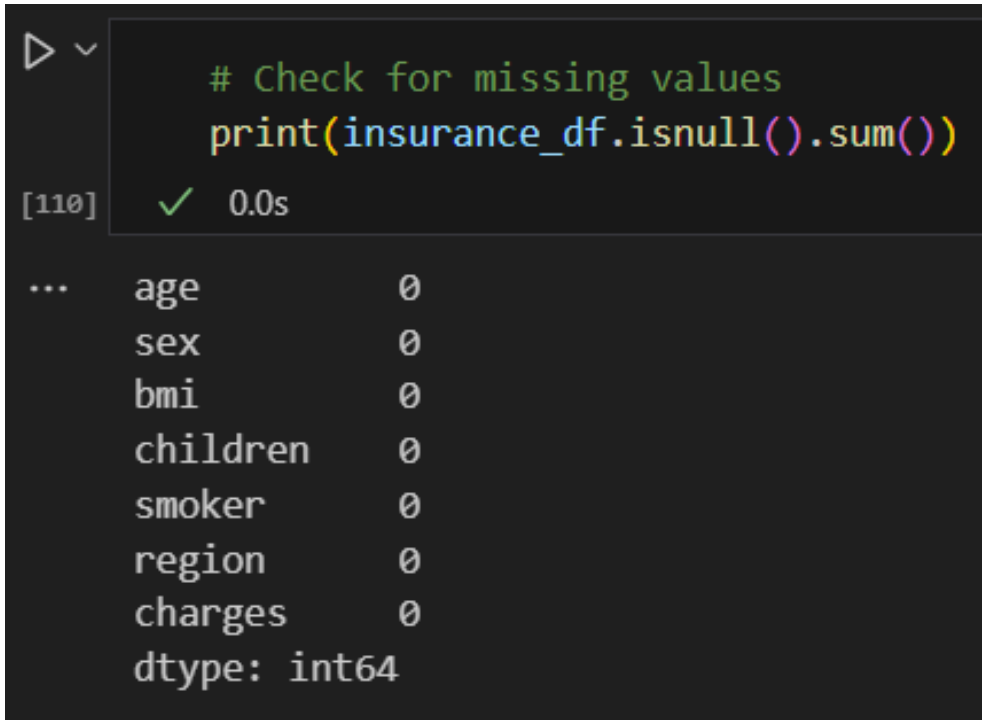
```
# Check dataset info
print(insurance_df.info())
```

[108] ✓ 0.0s

```
<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
```

STEPS OF DATA CLEANSING

- ✓ Missing Values
- ( ) Duplicate Entries
- ( ) Outliers
- ✓ Inconsistent Formatting
- ✓ Incorrect Data Types
- ✓ String/Whitespace Issues
- ✓ Mismatched Data – short Excel check
- ( ) Scaling and Normalization Issues



STEPS OF DATA CLEANSING

- ✓ Missing Values
- ✓ Duplicate Entries
- ( ) Outliers
- ✓ Inconsistent Formatting
- ✓ Incorrect Data Types
- ✓ String/Whitespace Issues
- ✓ Mismatched Data – short Excel check
- ( ) Scaling and Normalization Issues

```
# Check for duplicate rows
duplication_check = insurance_df.duplicated()
print("Number of Duplicates:",duplication_check.sum())

print("\nDuplicates:")
print(insurance_df.iloc[np.where(duplication_check==True)[0]])
```

[111] ✓ 0.0s

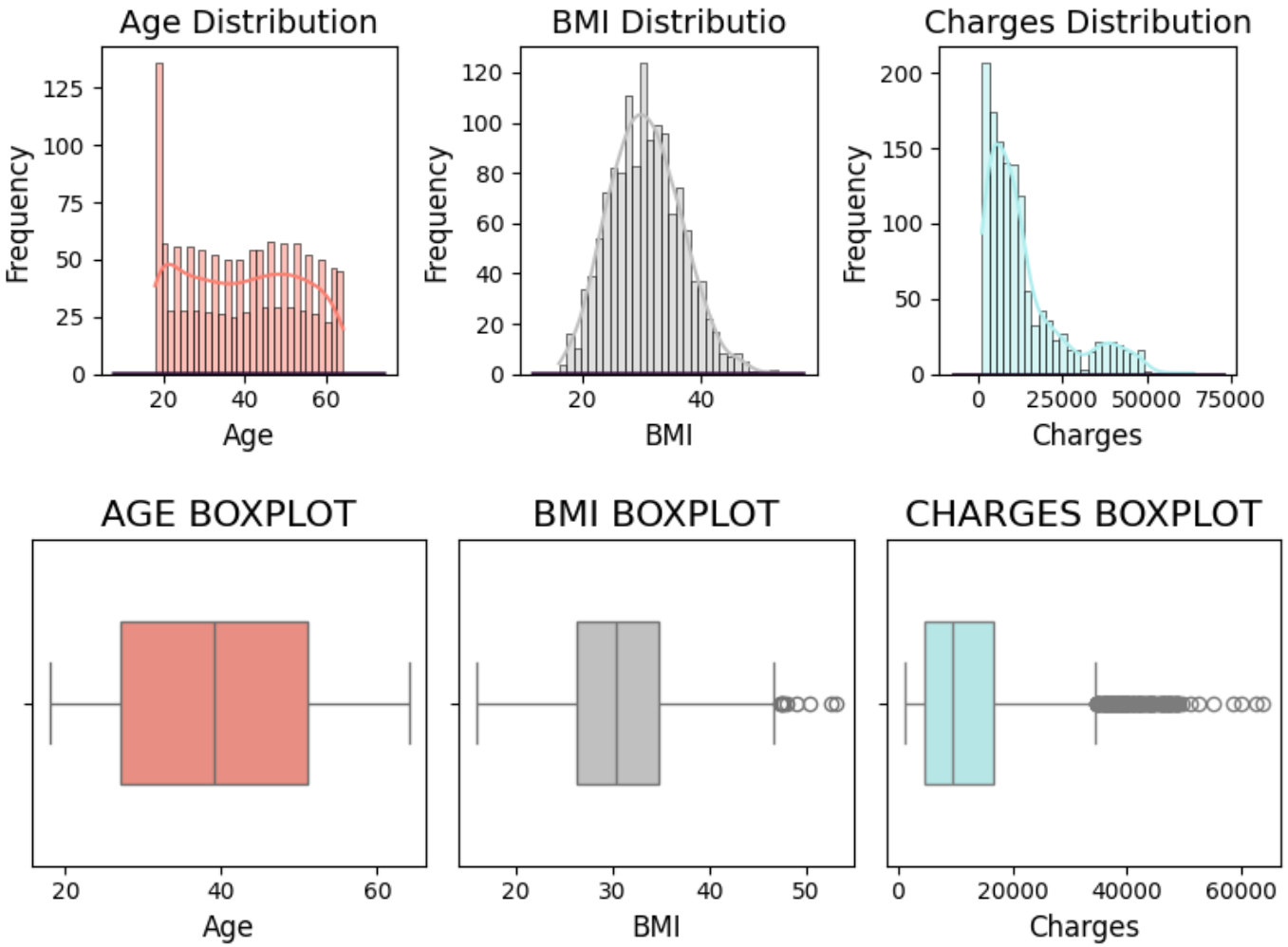
```
Number of Duplicates: 1

Duplicates:
     age   sex    bmi  children smoker     region    charges
581   19  male  30.59         0     no  northwest  1639.5631
```

CHECK OF DISTRIBUTION

CHECK OF DISTRIBUTIONS (1) + OUTLIERS

Observation

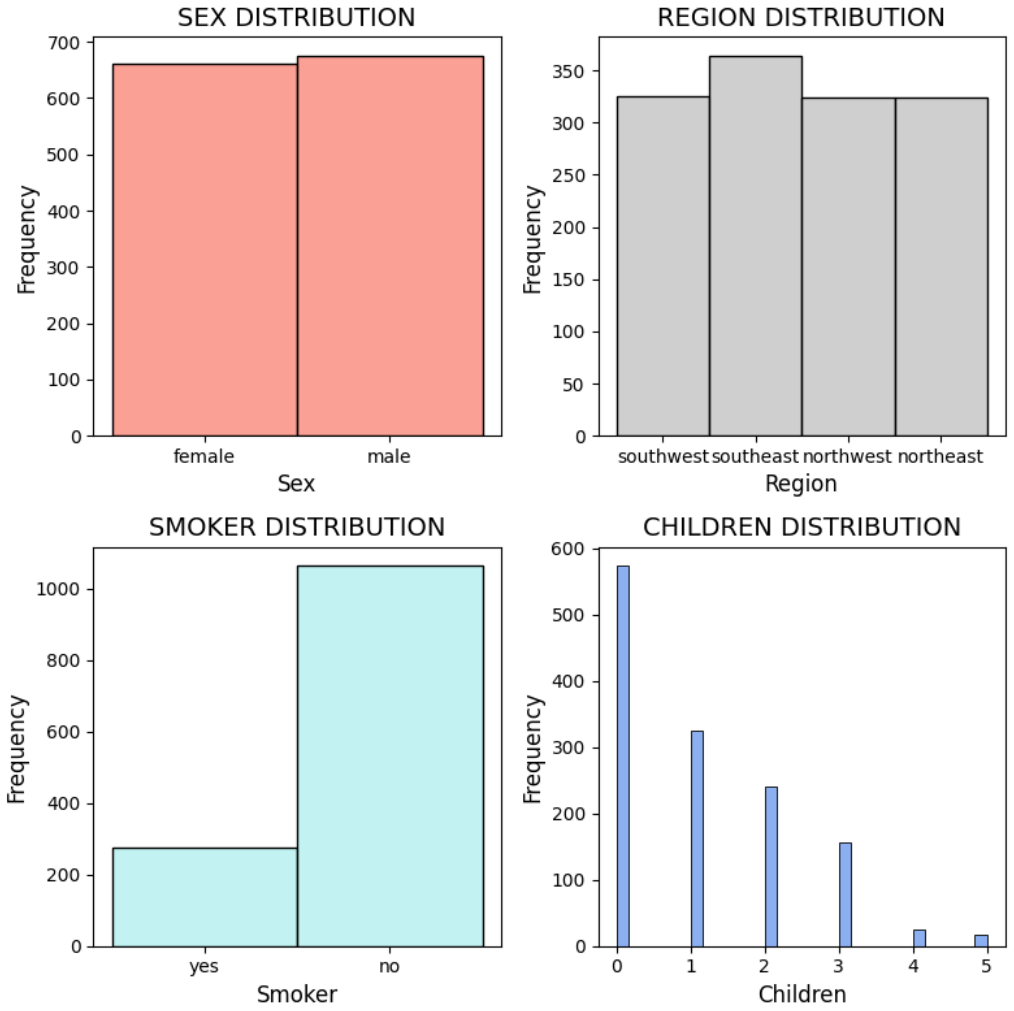
- Distribution
  - Even distribution in age
  - BMI: Gauss curve
  - Charges: great left skew
- Outliers
  - 9 Outliers at BMI
  - > 100 Outliers at Charges



CHECK OF DISTRIBUTIONS (2) + STRING CHECK

Observation

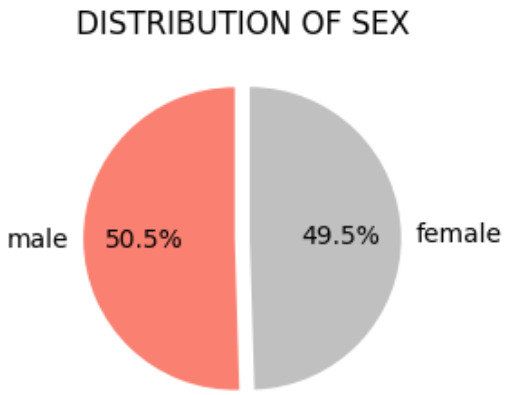
- Relatively even distribution between sex and region
- Uneven distribution between
  - Smokers
  - Children (w/ has a far higher concentration)

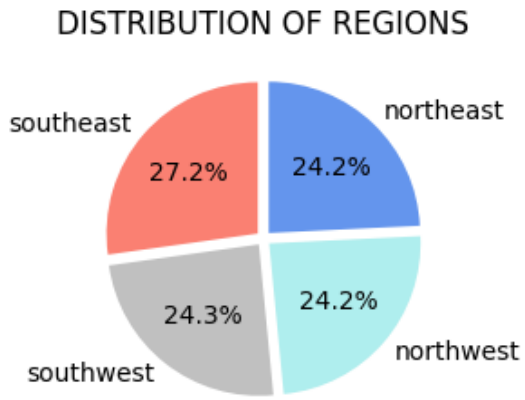


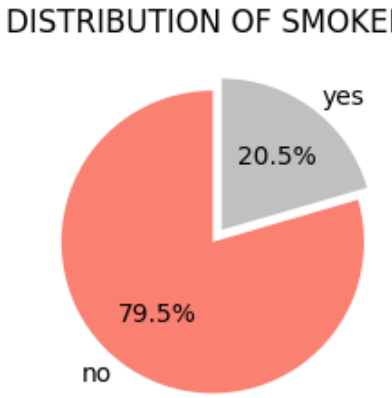
CHECK OF DISTRIBUTIONS (3)
SEX, REGION, SMOKER CHILDREN

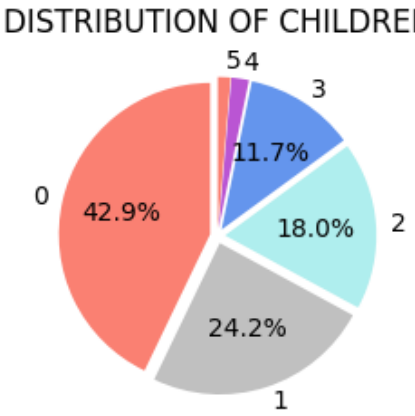
Observation

- Relatively even distribution between sex and region
- Uneven distribution between
  - Smokers
  - Children (w/ has a far higher concentration)





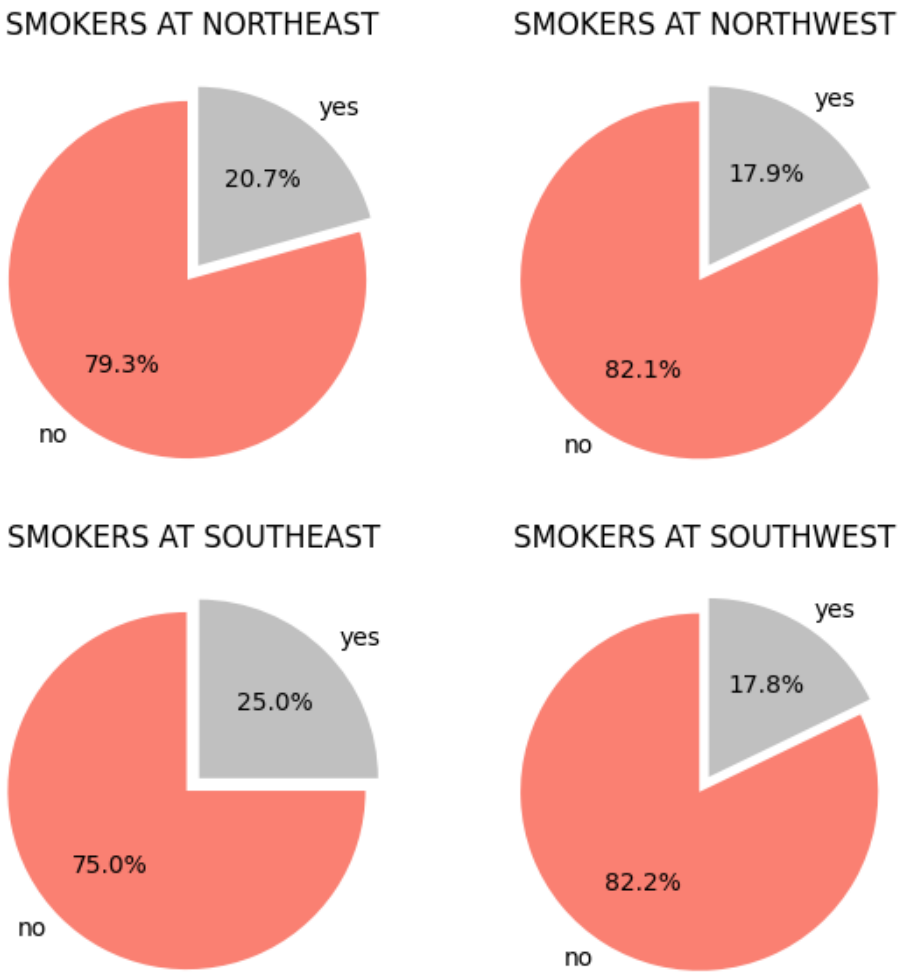




CHECK OF DISTRIBUTIONS (4)
(NON) SMOKERS PER SEX & REGION

Observation

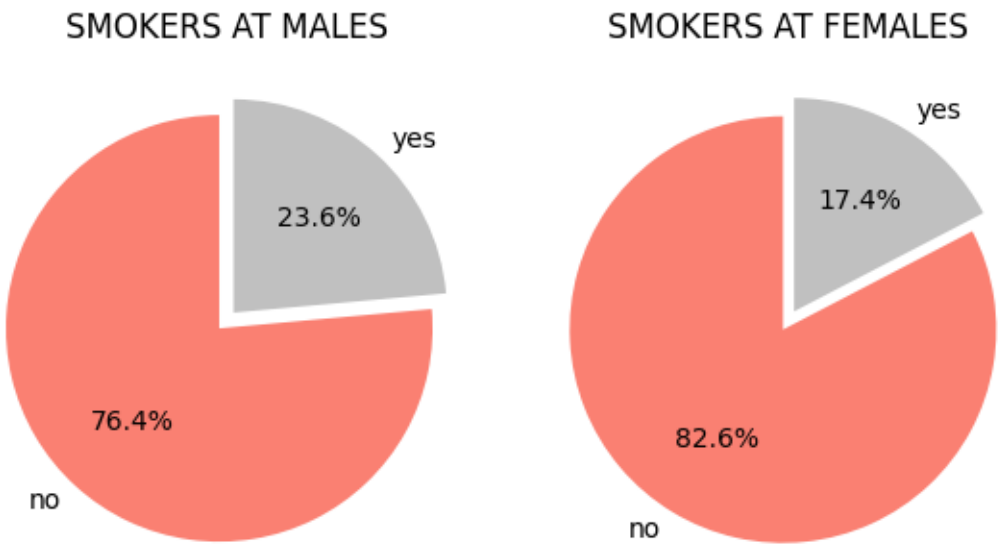
- No great variation of smokers by region
- Small outlier: Southeast

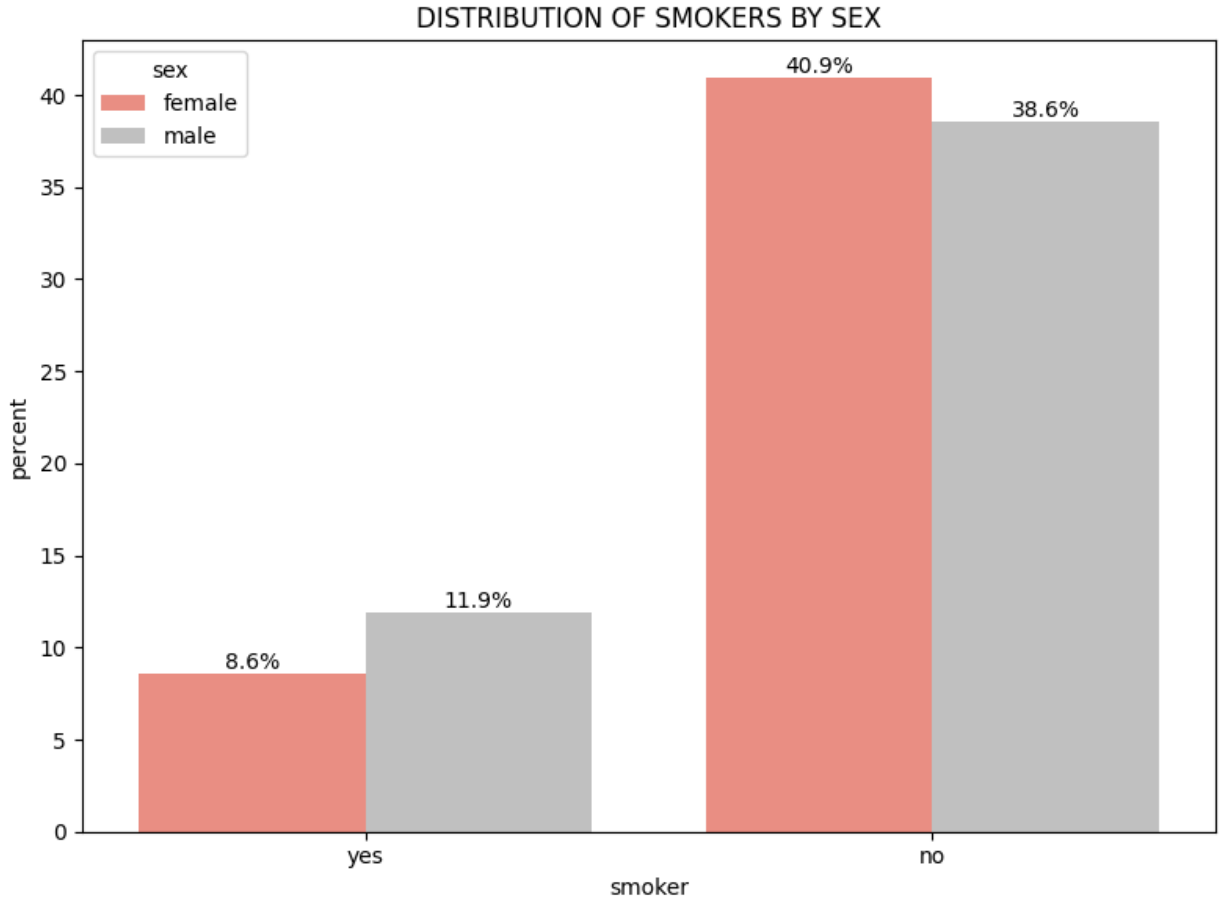


CHECK OF DISTRIBUTIONS (5)
(NON) SMOKERS PER SEX & REGION

Observation

- Relatively even distribution of smokers at a certain sex
- Males slightly higher than females





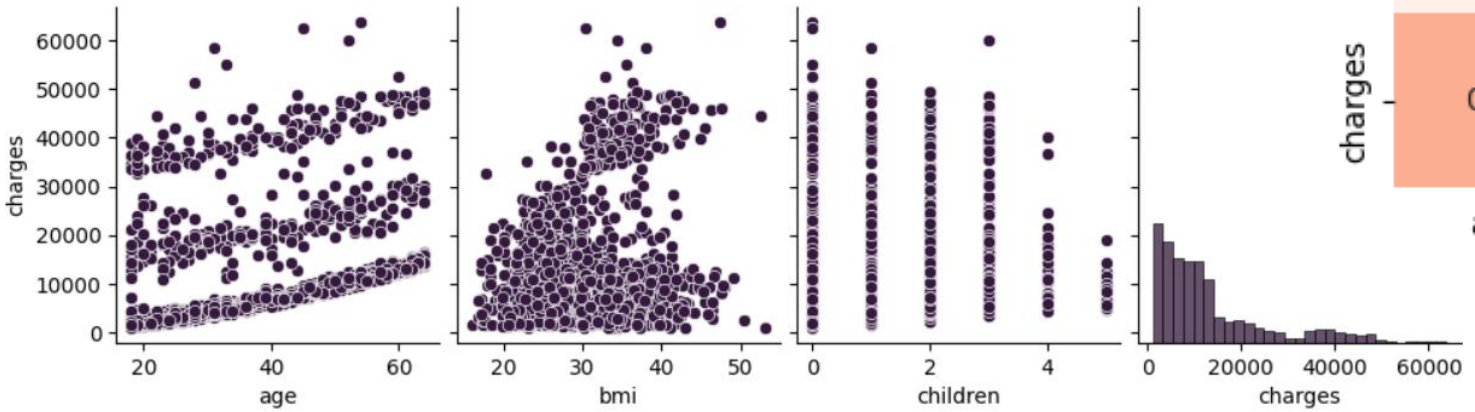
CHECK OF CHARGES

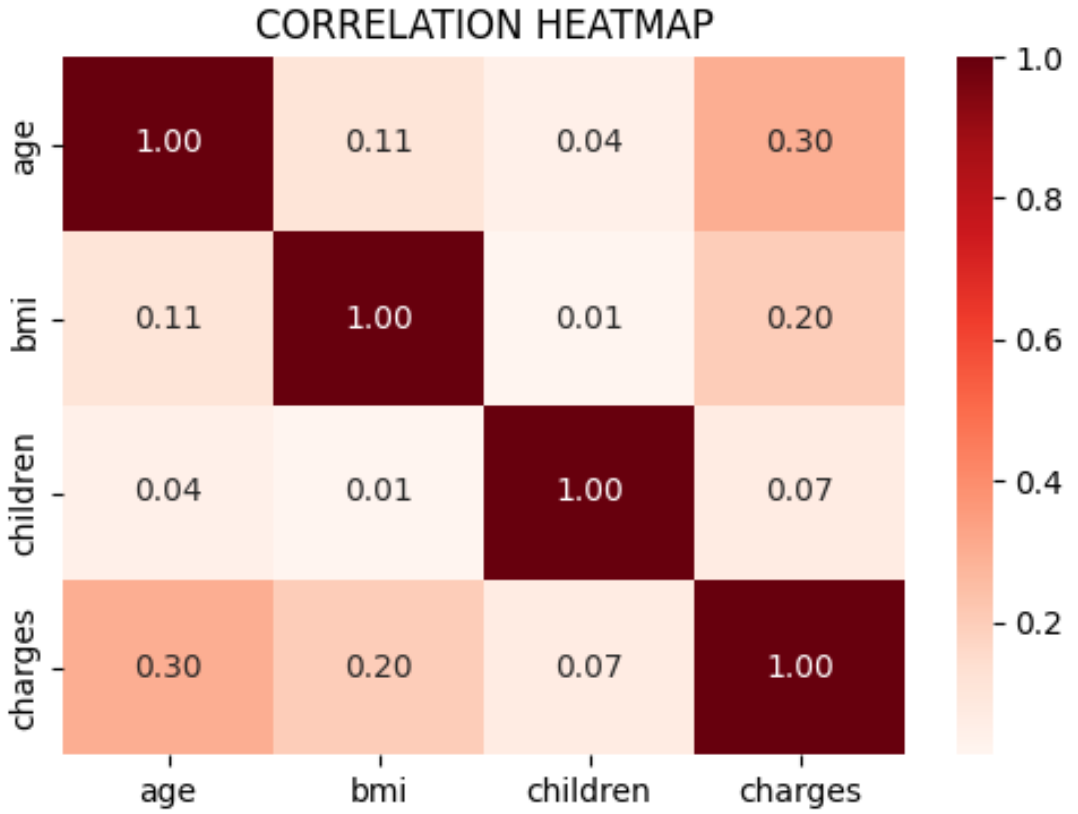
```
sns.pairplot(df_cleaned)  plt.show()
```

CHECK OF CHARGES (1) CORRELATIONS

Observation

- Look on heatmap → lowest horizontal line
  - Greatest impact by age and BMI
  - Low impact by number of children

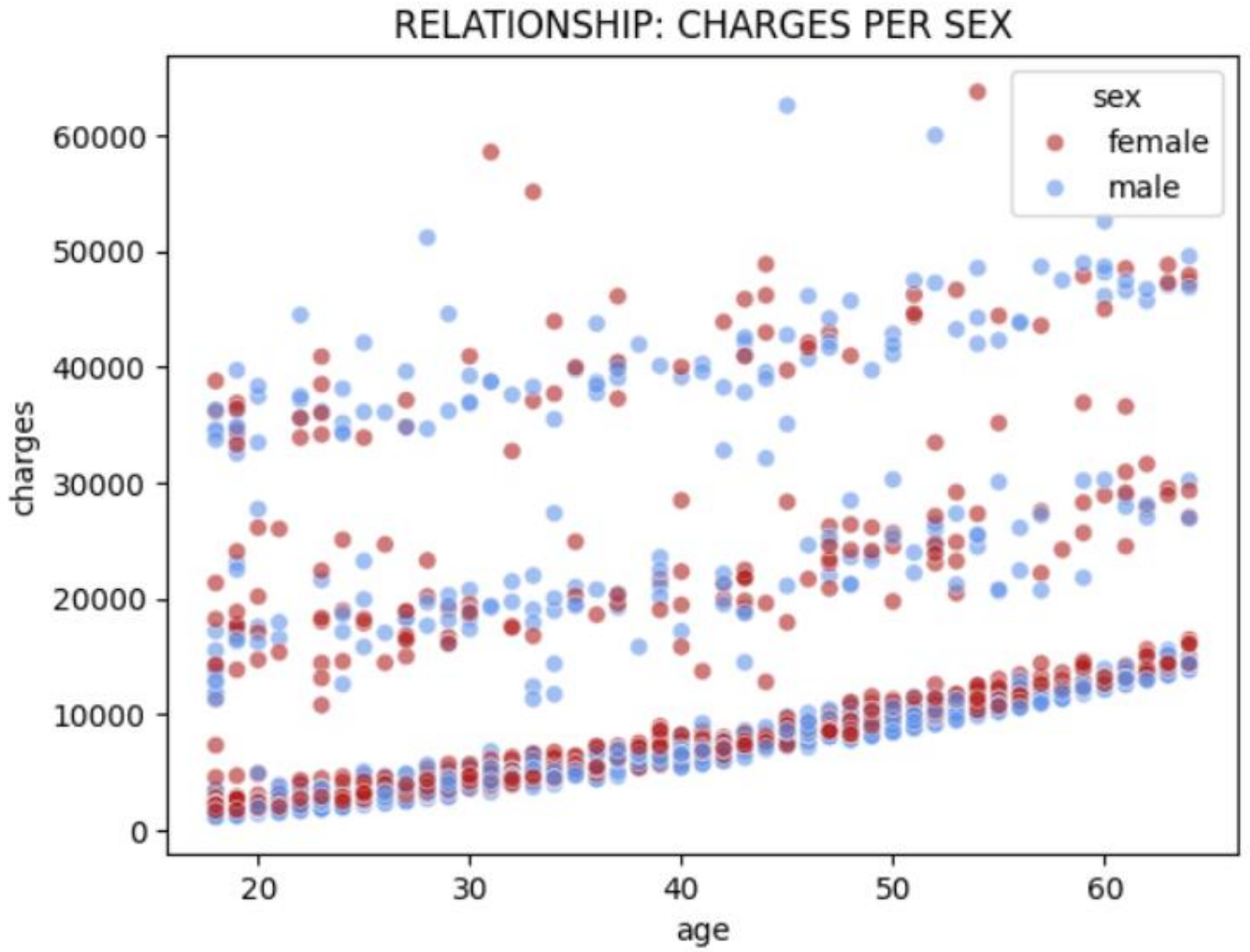




CHECK OF CHARGES (2) CORRELATIONS

Observation

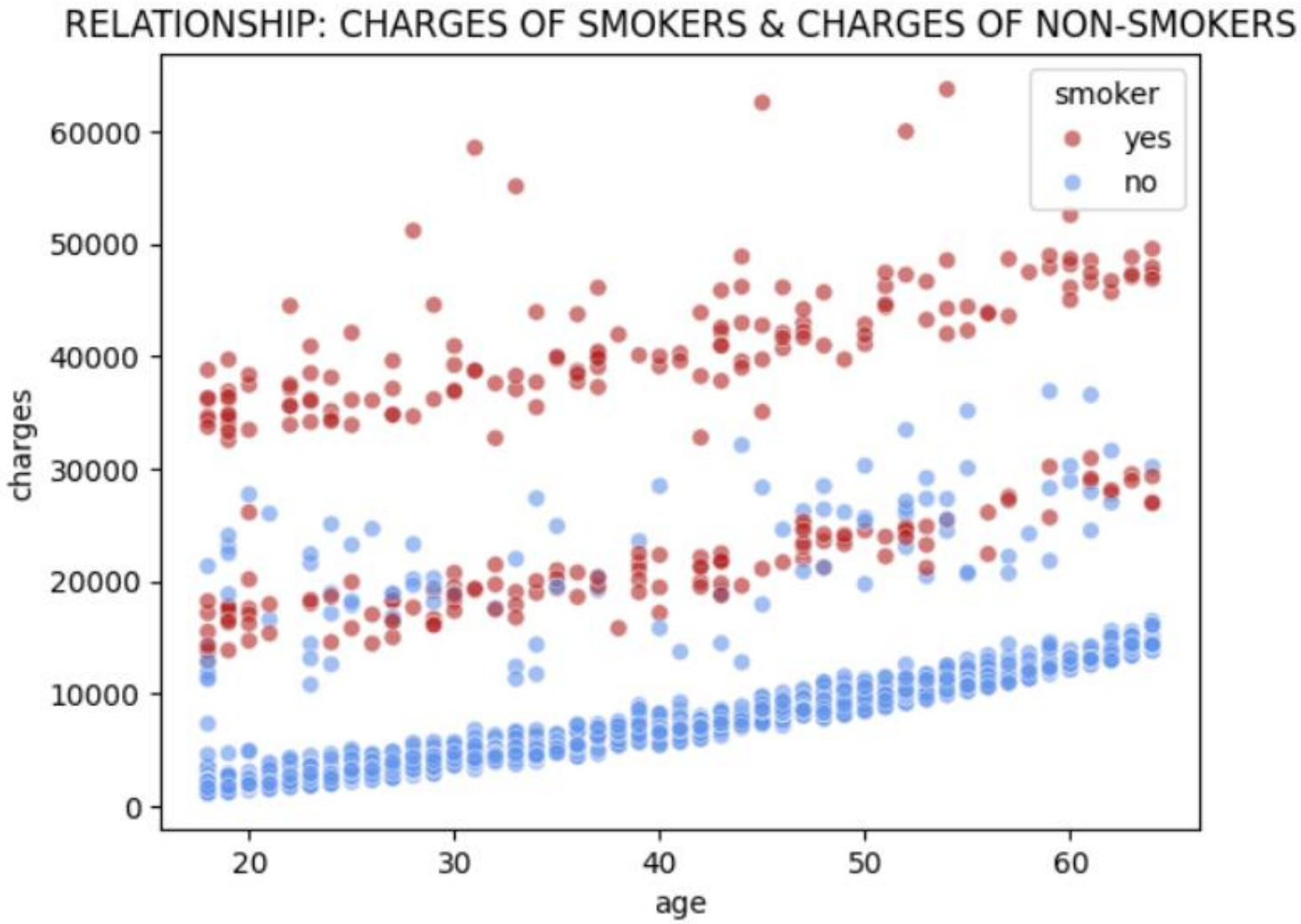
- Impact by sex is evenly distributed by male and female
- Age is a higher impact



CHECK OF CHARGES (3) CORRELATIONS

Observation

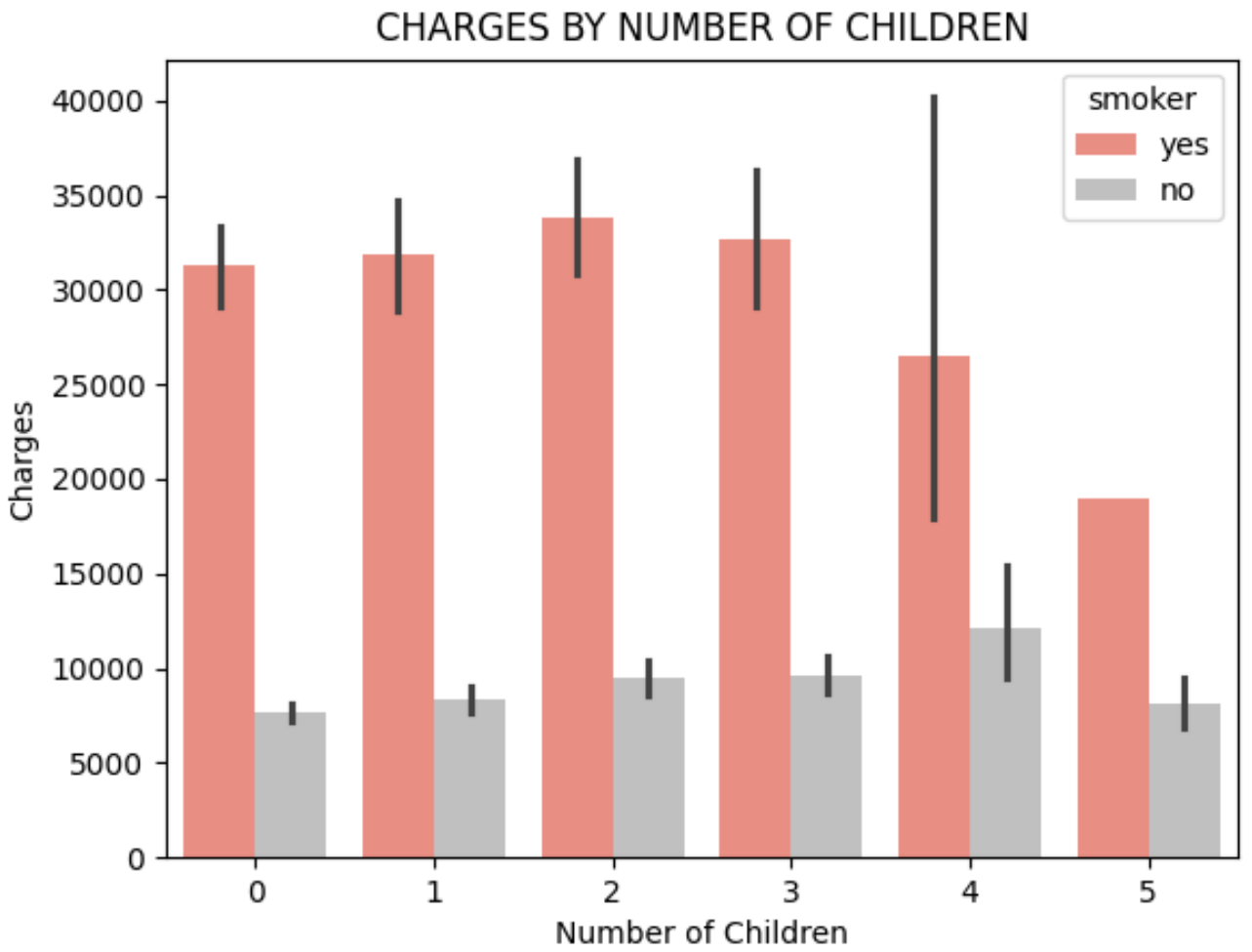
- Critical impact by smoking
- Age has an impact, too



CHECK OF CHARGES (4) CORRELATIONS

Observation

- Number of children has almost no impact until a number of 3
- Great decrease at 4 and 5 children
- Great impact by smoking



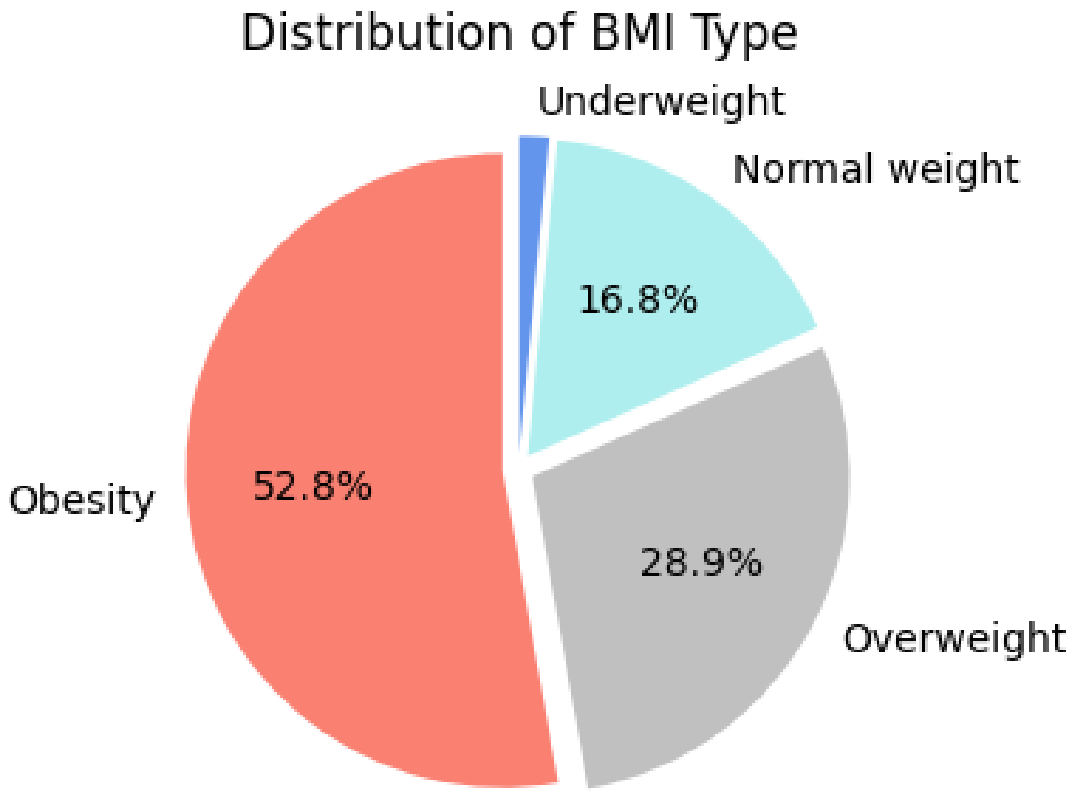
CHECK OF CHARGES (5) CORRELATIONS

Clustering

- Clustering of BMI according to current standards:
  - BMI > 18.5 → Underweight
  - 18.5 < BMI < 25 → Normal Weight
  - 25 =< BMI < 30 → Overweight
  - 30 < BMI → Obesity

Observation

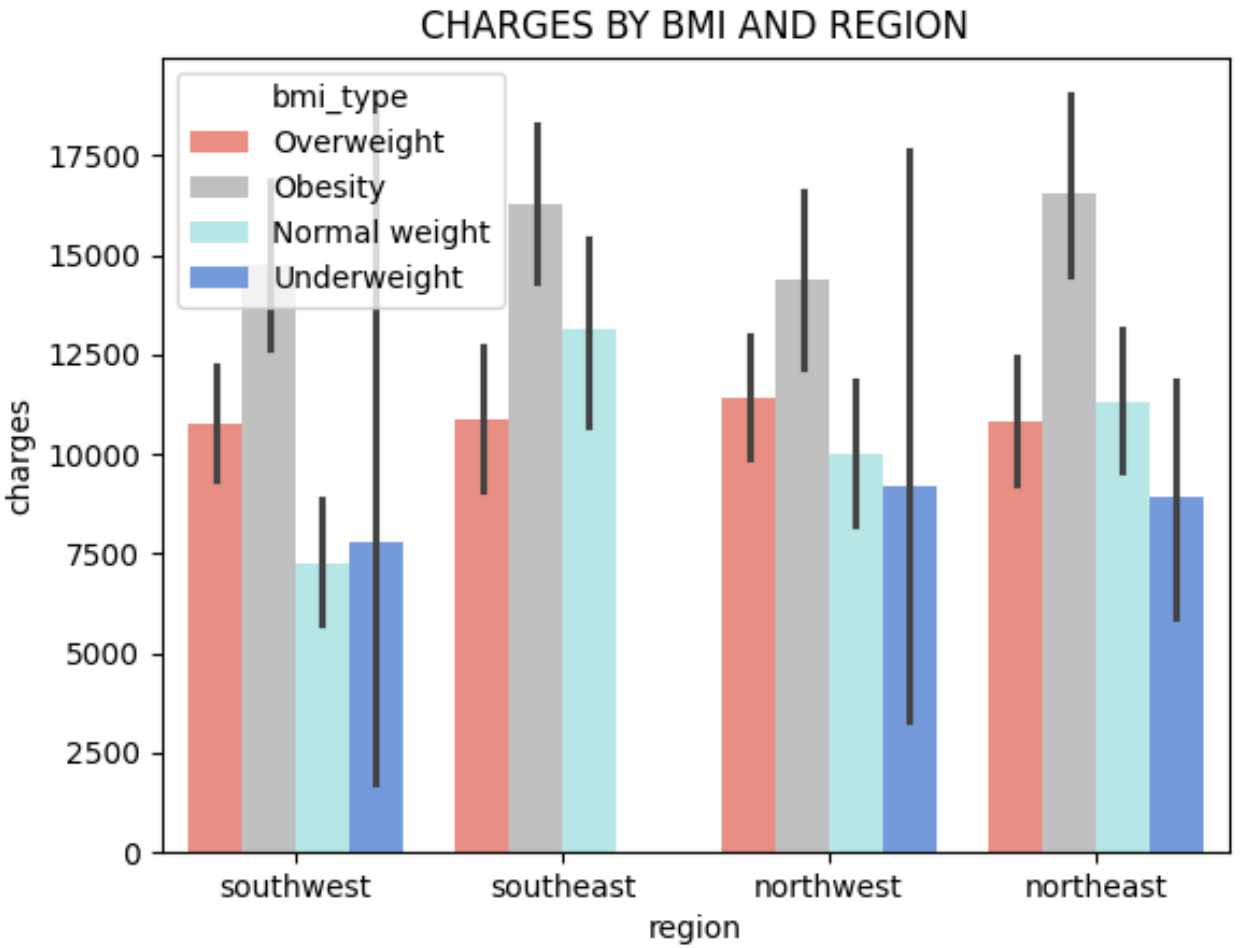
- High percentage of obesity and overweight
- Keep in mind: BMI one of the highest impacts on charges



CHECK OF CHARGES (6) CORRELATIONS

Observation

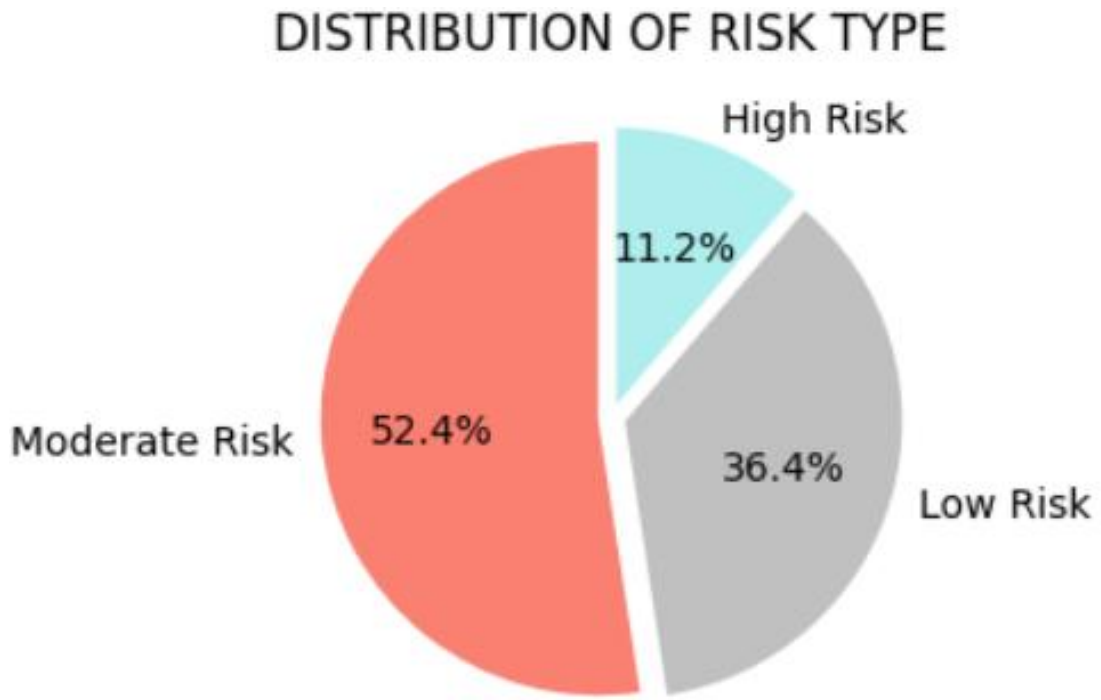
- No underweight people in southeast
- Highest charges for obesity
- Tie 2nd highest charges for normal and overweight
  - Assumption: affect by smokers



CHECK OF CHARGES (7) CORRELATIONS

Clustering

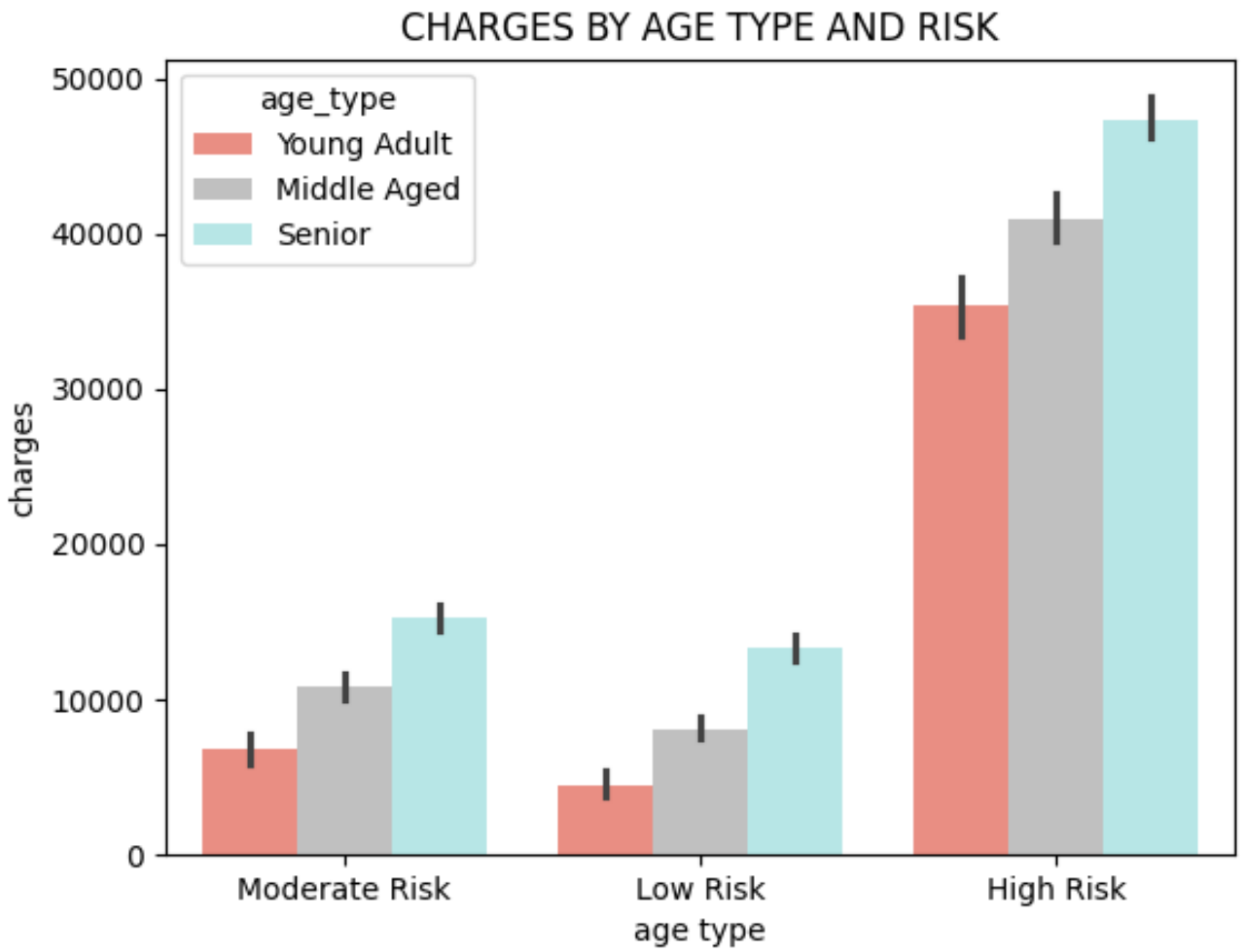
- Clustering to risk level:
  - High risk: Obesity + Smoker
  - Moderate risk: normal/overweight + smoker
  - Else: low risk
- Clustering by age:
  - Age < 30 years
  - 30 years < Age < 50 years
  - Age > 50 years



CHECK OF CHARGES (8) CORRELATIONS

Observations

- Charges increase with risk level
- Age as greater impact



KEY FINDINGS

- Findings (Data Distribution):
  - Outliers:
    - BMI: 9
    - Charges: > 100 + Skew to the right
  - More or less balanced distribution per
    - Age
    - Sex
    - Smokers
    - Region

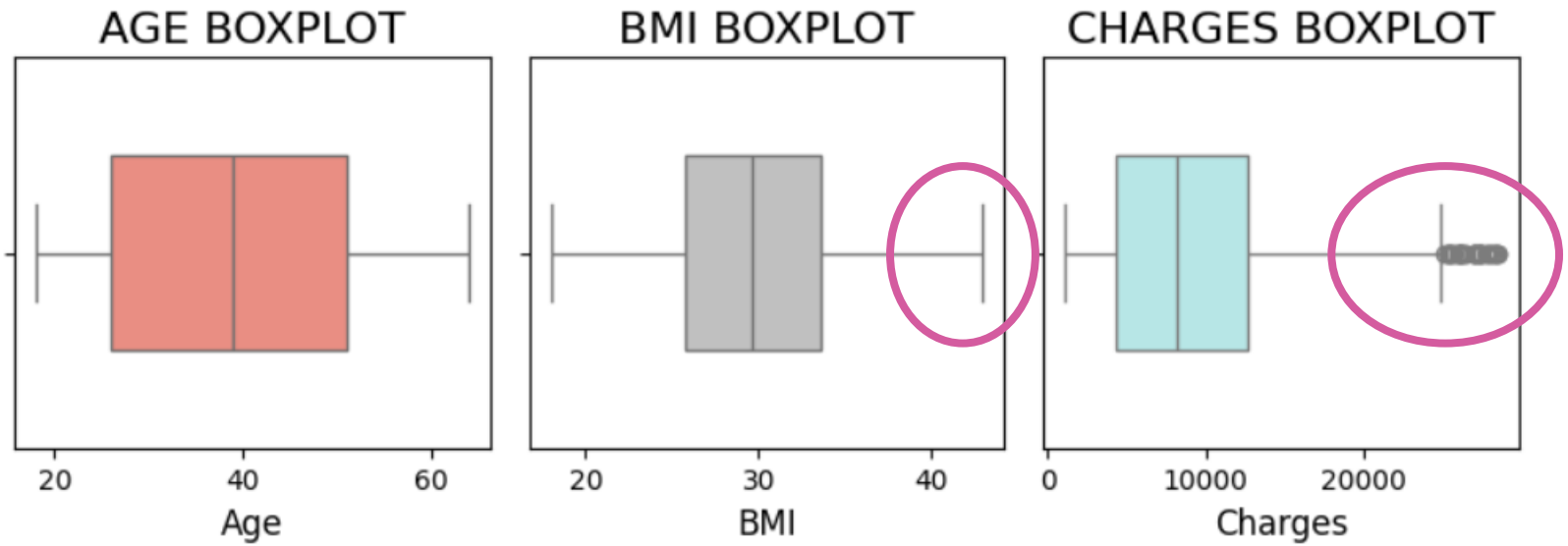
- Insights
  - Smoking has a high impact on charges
  - Also, obesity affects charges heavily
  - Charges rise continuously per age
  - Regional distribution of the factors including the sex mentioned above is relatively evenly done
  - Low impact on charges by the number of children

DATA CLEANSING

PART 2

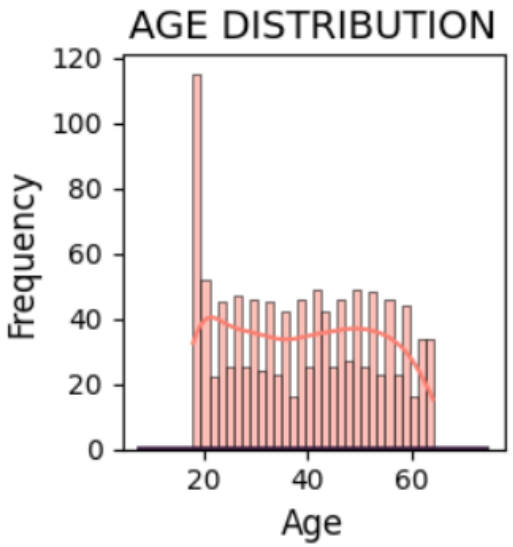
STEPS OF DATA CLEANSING

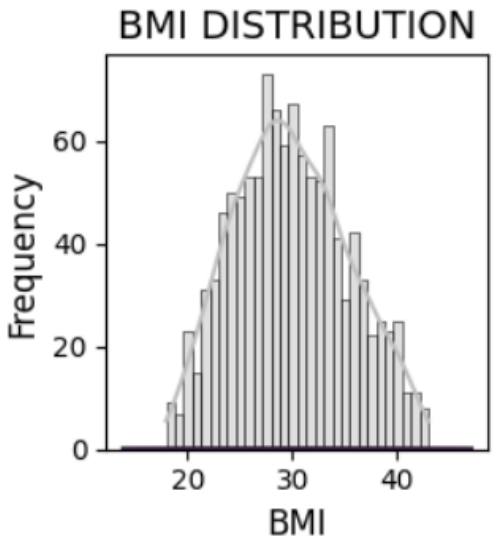
- ✓ Missing Values
- ✓ Duplicate Entries
- ✓ Outliers
- ✓ Inconsistent Formatting
- ✓ Incorrect Data Types
- ✓ String/Whitespace Issues
- ✓ Mismatched Data
- ( ) Scaling and Normalization Issues

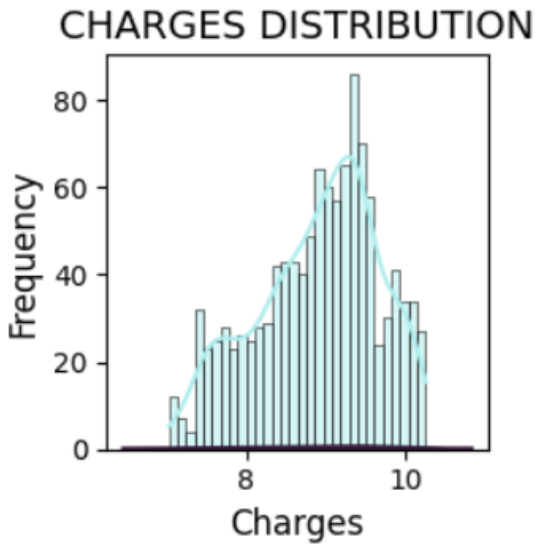


STEPS OF DATA CLEANSING

- ✓ Missing Values
- ✓ Duplicate Entries
- ✓ Outliers
- ✓ Inconsistent Formatting
- ✓ Incorrect Data Types
- ✓ String/Whitespace Issues
- ✓ Mismatched Data
- ✓ Scaling and Normalization Issues







	age	sex	bmi	children	smoker	charges	southwest	southeast	northwest	northeast
0	-1.436011	1	-0.361123	-0.888643	1	1.089896	1	0	0	0
1	-1.507805	0	0.709290	-0.068301	0	-1.863695	0	1	0	0
2	-0.789863	0	0.568878	1.572383	0	-0.637262	0	1	0	0
3	-0.430892	0	-1.308447	-0.888643	0	1.431710	0	0	1	0
4	-0.502686	0	-0.182417	-0.888643	0	-0.818993	0	0	1	0

COMPARISON: PREDICTED VS. ACTUAL VALUES