**DATA ANALYTICS ON INSURANCE DATA**

**OVERVIEW:**

This project mainly focuses on analyzing the dataset containing insurance-related data using Python Pandas Library. The primary objectives include the following:

* Cleaning the dataset to handle missing and inconsistent values if exists.
* Preprocessing the data for meaningful insights.
* Performing exploratory data analysis (EDA) to uncover patterns and trends.

The analysis was conducted in Google Colab, leveraging Python’s powerful data analysis libraries.

**DATASET DESCRIPTION:**

**Source:** Provided by the tutor.

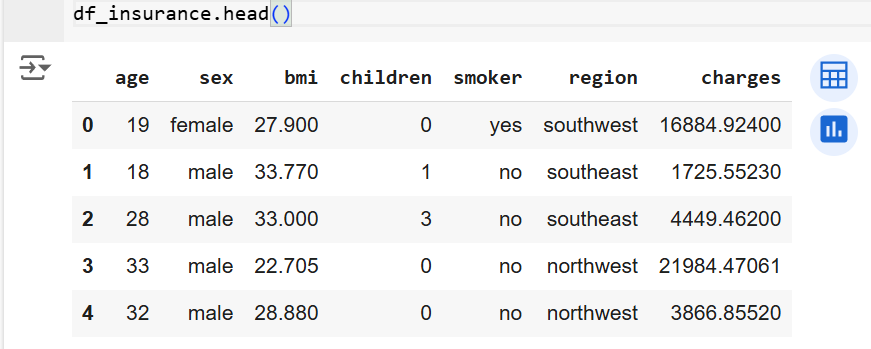
**Size**: Number of rows and columns.

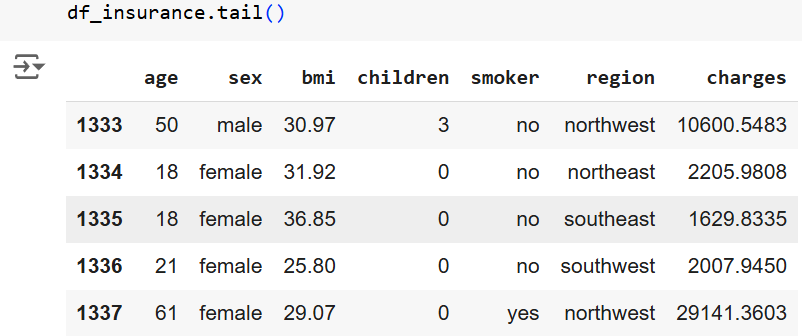
**Features**: Age, Gender, BMI, Children, Smoker, Region, Charges

**Objective:** The goal of this analysis is to uncover trends and correlations within the dataset to gain insights into factors affecting insurance charges.

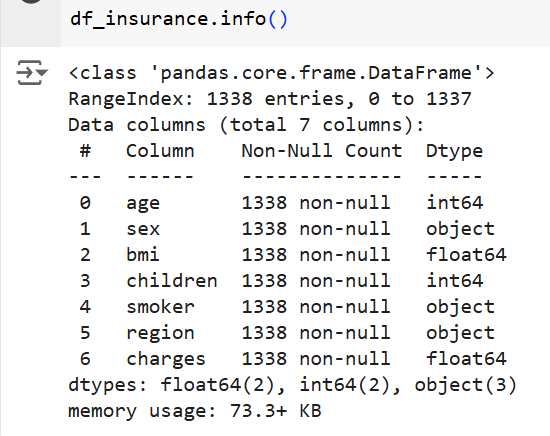
**DATA OVERVIEW:**

1. The first and last five observations of the dataset are listed below respectively.





1. Shape of the dataset: (1338,7). There are 1338 rows and 7 columns in the data set.
2. Information about the data set.



There are four numerical columns (age, bmi, children, charges) and three categorical columns (sex, smoker, region). Surprisingly we have no null values in the dataset which made the analysis process easy. We are using a storage about 73.3kb for this data.

Description about columns of the data is described below.

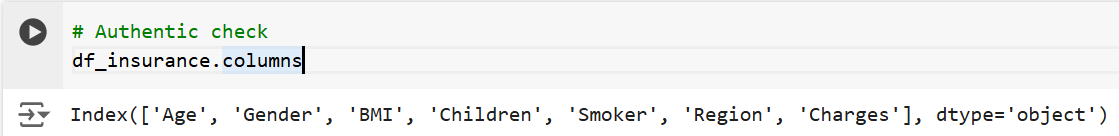
* Age: Age of the insurance owner.
* Sex: Gender of the insurance owner.
* BMI: BMI of insurance owner.
* Children: Number of children an insurance owner have.
* Smoker: Whether insurance owner is smoker or not.
* Region: Which part of USA does the owner belong to.
* Charges: Charges for the insurance.

**DATA PREPROCESSING:** Data cleaning and null value handling are two major parts of data preprocessing.

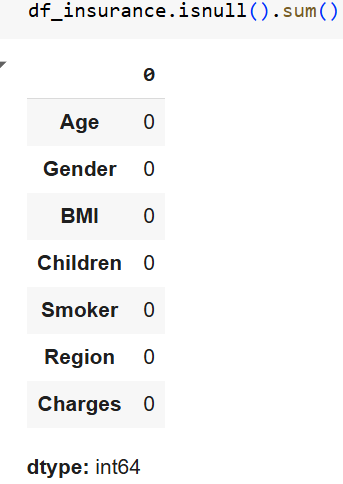
**Data Cleaning:**

1. Standardized/renamed the column names for consistency.



1. Identified that there are no null values in the data set.



**Observations:** Preprocessing confirmed that there are no null values in the data set and ensured uniformity in data representation, which is crucial for data analysis.

**EXPLORATORY DATA ANALYSIS (EDA):**

**Techniques used:**

Univariate Analysis: Performed analysis on each column.

Bivariate Analysis: Relationships between pairs of features using bar graph.

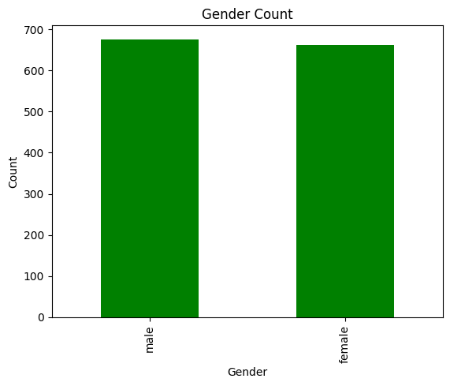
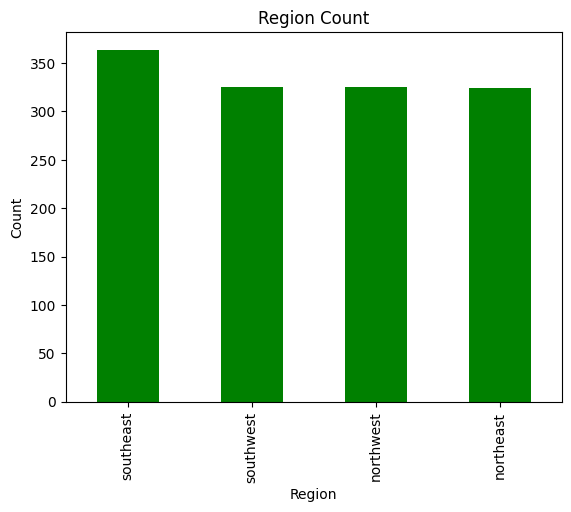
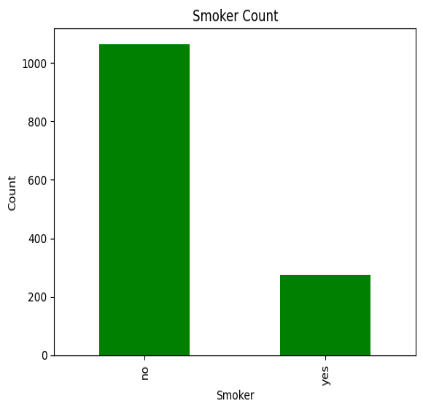
Multivariate Analysis: Investigating combined trends using bar graph.

**Results of preforming univariate analysis on numerical data:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Min** | **Max** | **Avg** |
| **Age** | 18 | 64 | 39.207 |
| **BMI** | 15.96 | 53.13 | 30.663 |
| **Children** | 0 | 5 | 1.09 |
| **Charges** | 1121.873 | 63770.428 | 13270.4222 |

**Results of performing univariate analysis on categorical columns:**

Categorical columns are Gender, Smoke, Region

|  |  |  |
| --- | --- | --- |
| **Columns** | | **Count** |
| **Gender** | Male | 676 |
| Female | 662 |
| **Smoker** | Yes  No | 274  1064 |
| **Region** | Northeast  Northwest  Southeast  Southwest | 324  325  364  325 |

**Bivariate Analysis:**

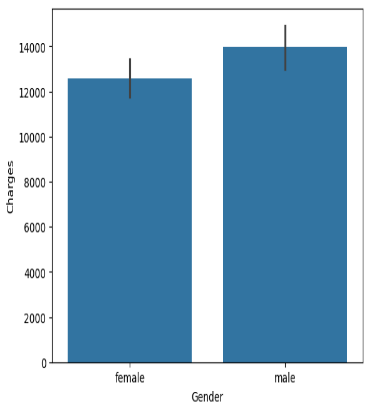
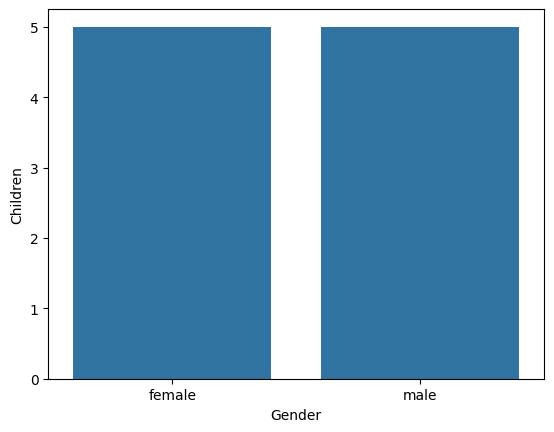
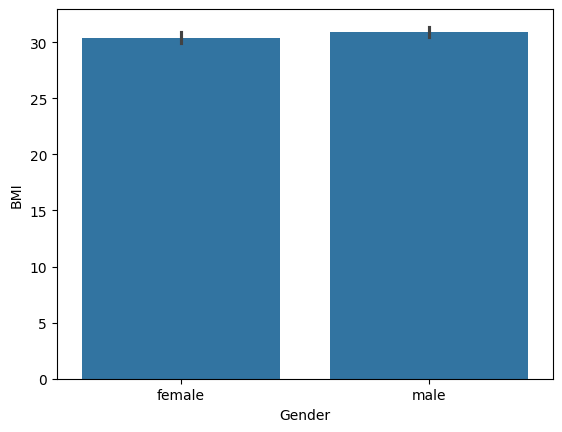
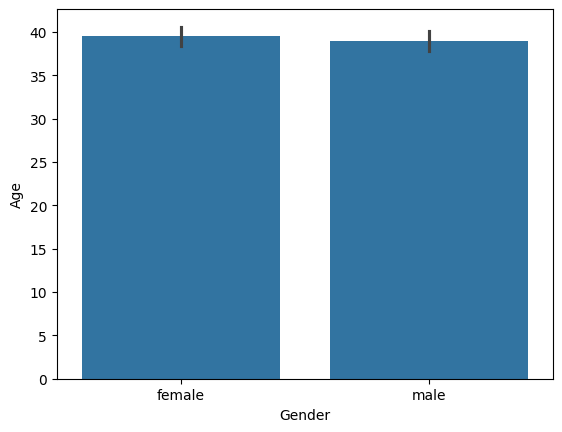
Performed bivariate analysis on one numerical and one categorical column:

Categorical Columns: Gender, Smoker, Region

Numerical Columns: Age, BMI, Children, Charges

(Gender, Age), (Gender, BMI), (Gender, Children), (Gender, Charges), (Smoker, Age), (Smoker, BMI), (Smoker, Children), (Smoker, Charges), (Region, Age), (Region, BMI), (Region, Children), (Region, Charges)

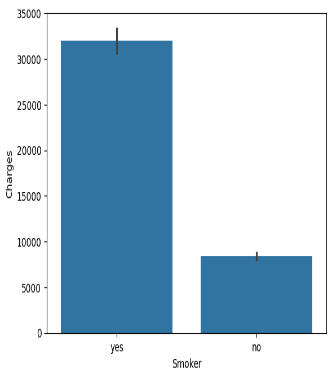
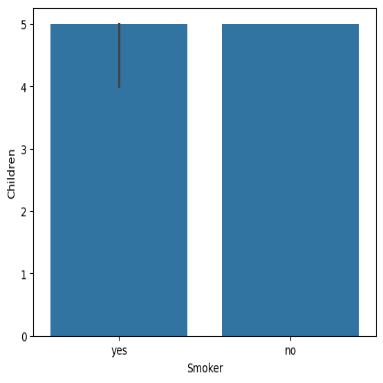
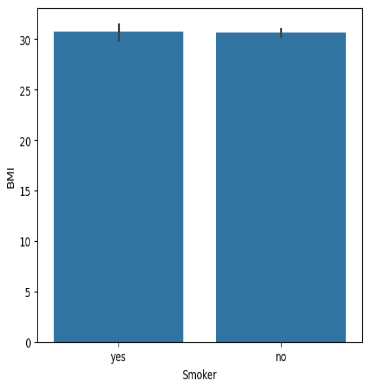
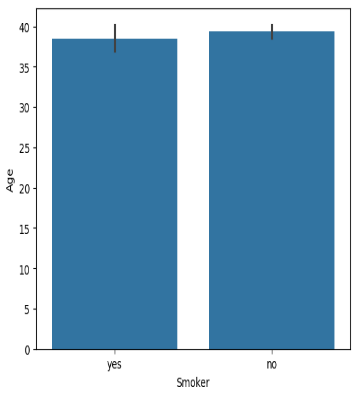
1. **(Gender, Age), (Gender, BMI), (Gender, Children), (Gender, Charges)**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gender** | **Age** | **BMI** | **Children** | **Charges** |
| Male | 38 | 30+ | 5 | 13000-14000 |
| Female | 39 | 30 | 5 | 12000-13000 |

From this I can conclude that charges paid by males are a bit more than females due to higher BMI.

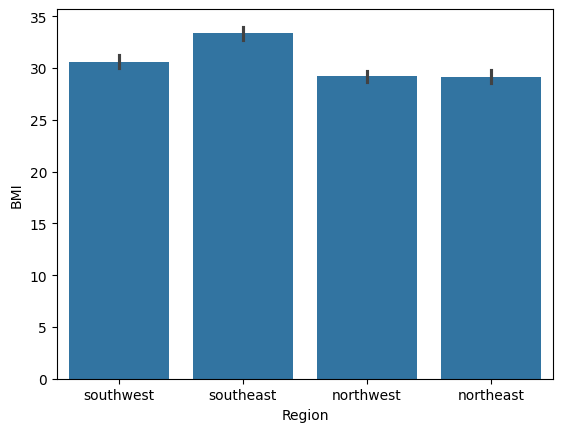
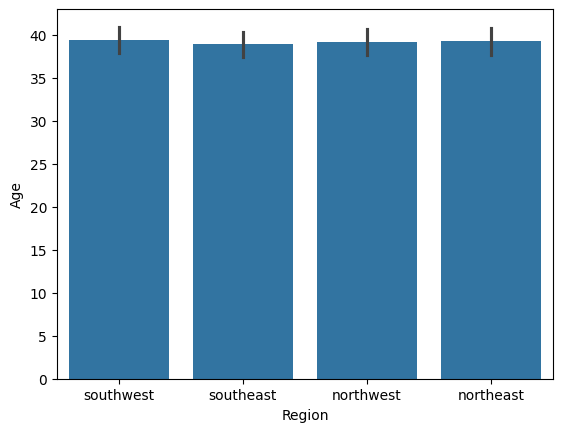
1. **(Smoker, Age), (Smoker, BMI), (Smoker, Children), (Smoker, Charges)**

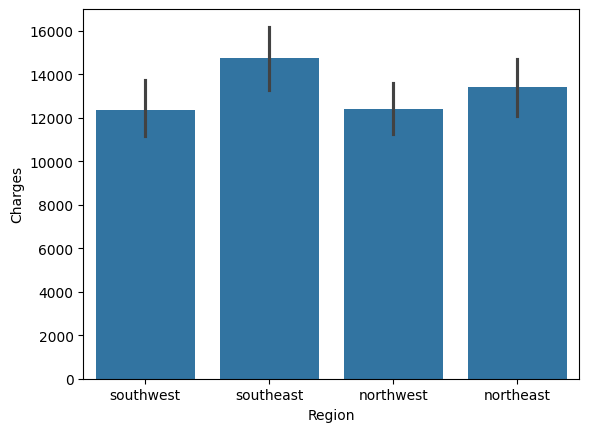
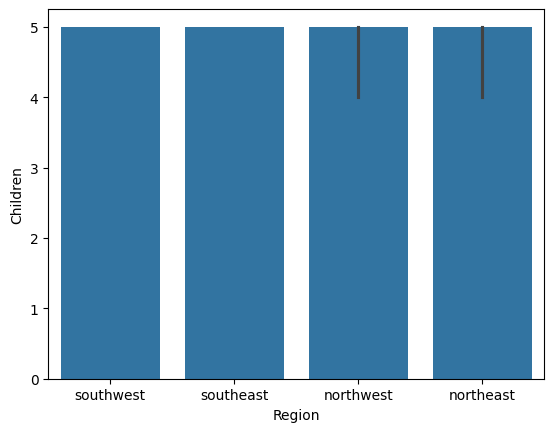


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Smoker** | **Age** | **BMI** | **Children** | **Charges** |
| Yes | 38 | 30+ | 5 | 33000 |
| No | 39 | 30 | 5 | 8000 |

From this I can say that smokers are paying high amount compared to non-smokers.

1. **(Region, Age), (Region, BMI), (Region, Children), (Region, Charges)**





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Region** | **Age** | **BMI** | **Children** | **Charges** |
| Southwest | 39 | 30 | 5 | ~12000 |
| Southeast | 38 | 33 | 5 | ~14000 |
| Northwest | 39 | 29 | 5 | ~12000 |
| Northeast | 40 | 29 | 5 | ~12000 |

Highest charges are imposed on Southeast region people as their BMI is comparatively high.

**Multivariate Analysis:**

# Gender, smoker, age

# gender, region, age

# smoker, region, age

# Gender, smoker, BMI

# gender, region, BMI

# smoker, region, BMI

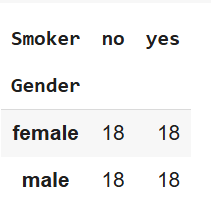
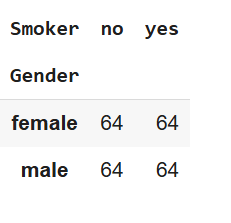
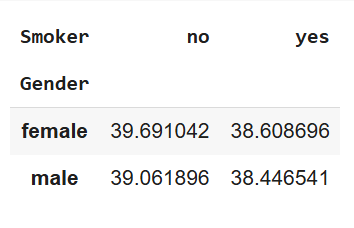
# Gender, smoker, charges

# gender, region, charges

# smoker, region, charges

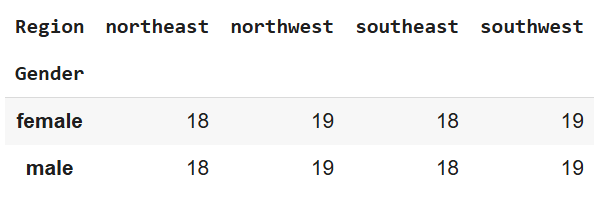
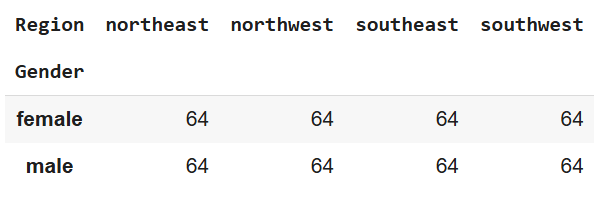
# Gender, smoker, age

**MIN MAX AVG**

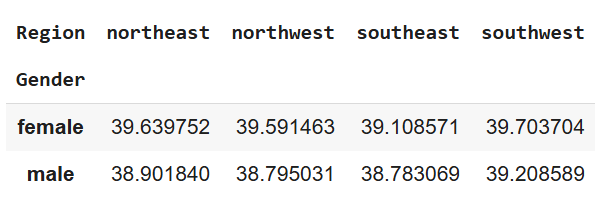
  

# gender, region, age

**MIN MAX**

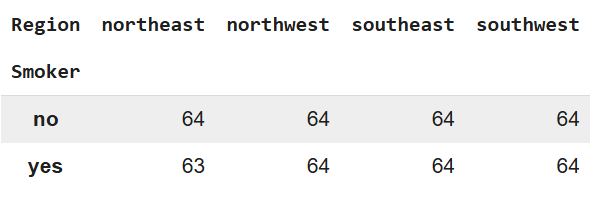
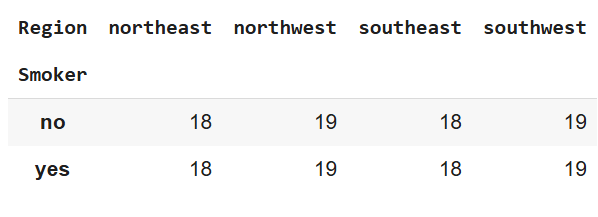
 

**AVG**

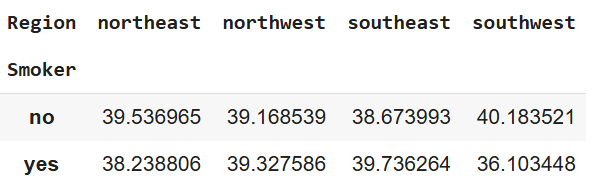


# smoker, region, age

**MAX MIN**

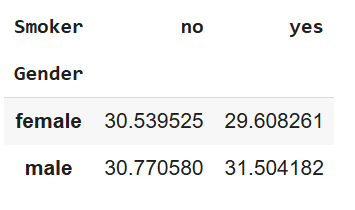
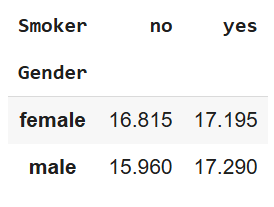
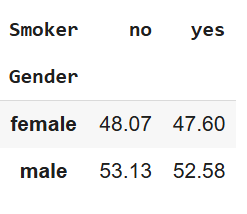
 

**AVG**



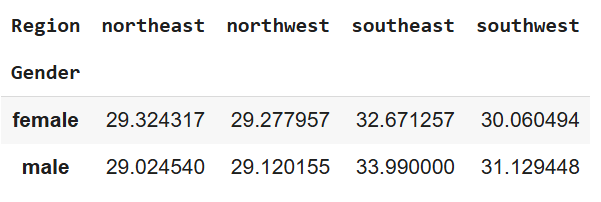
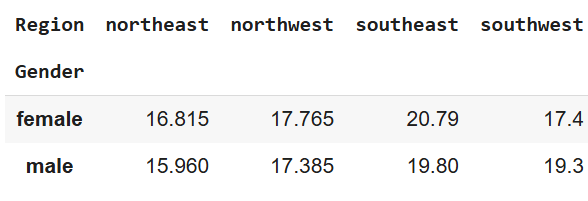
# Gender, smoker, BMI

**AVG MIN MAX**

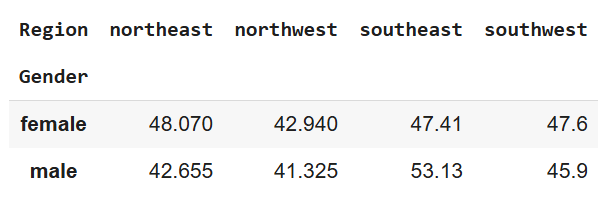
  

# gender, region, BMI

**AVG MIN**

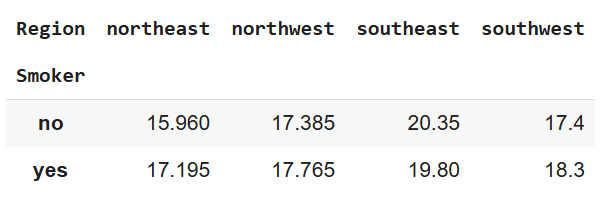
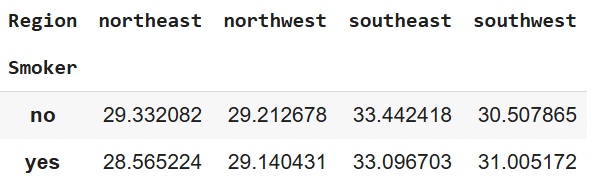
 

**MAX**

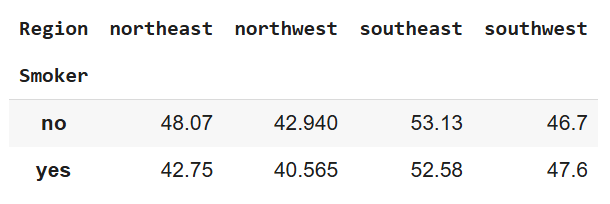


# smoker, region, BMI

**MEAN MIN**

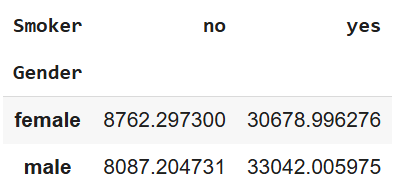
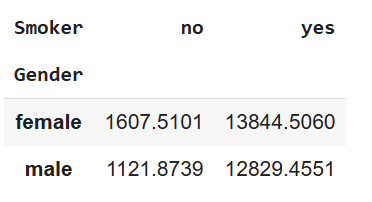
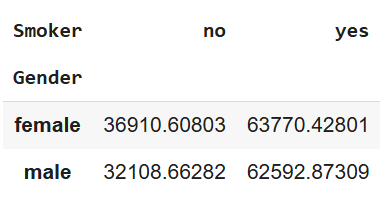


**MAX**



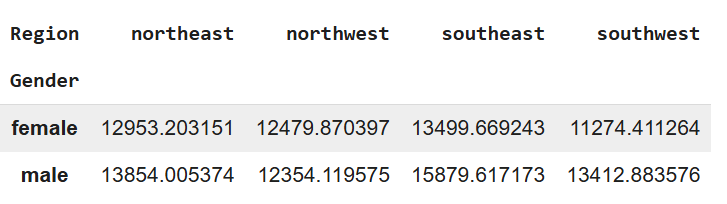
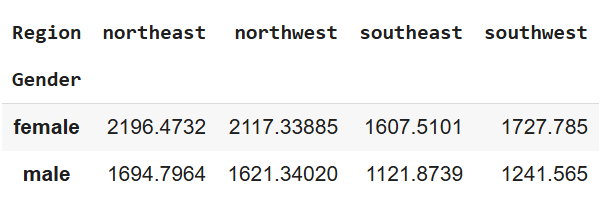
# Gender, smoker, charges

**MEAN MIN MAX**

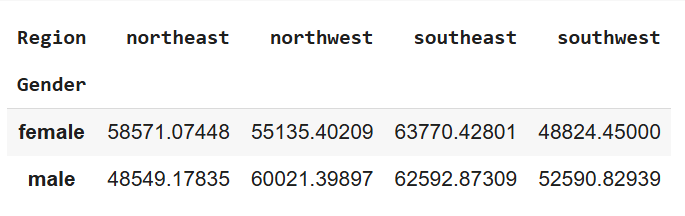
  

# gender, region, charges

**MEAN MIN**

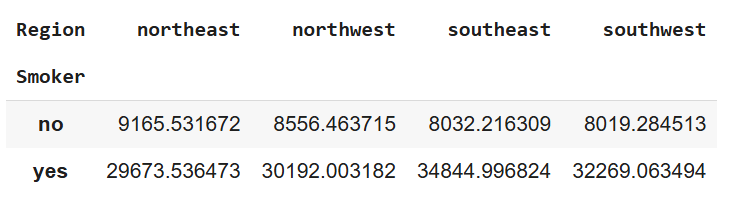
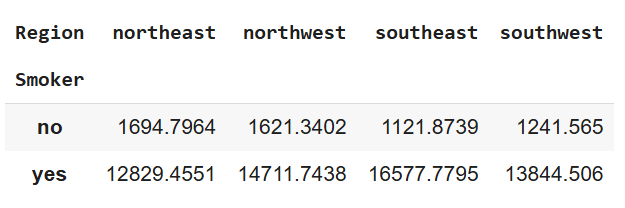
 

**MAX**

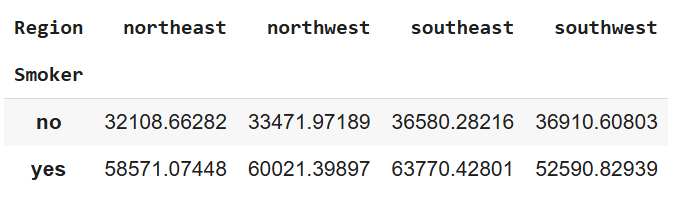


#smoker, region, charges

**MEAN MIN**

**MAX**



CONCLUSION:

So I can conclude that on an average Smoker’s from various regions are paying high compared to Non-Smoker’s that too Males are paying more compared to females as they are having high BMI and more smokers are Males