Name: Zaid Saleh

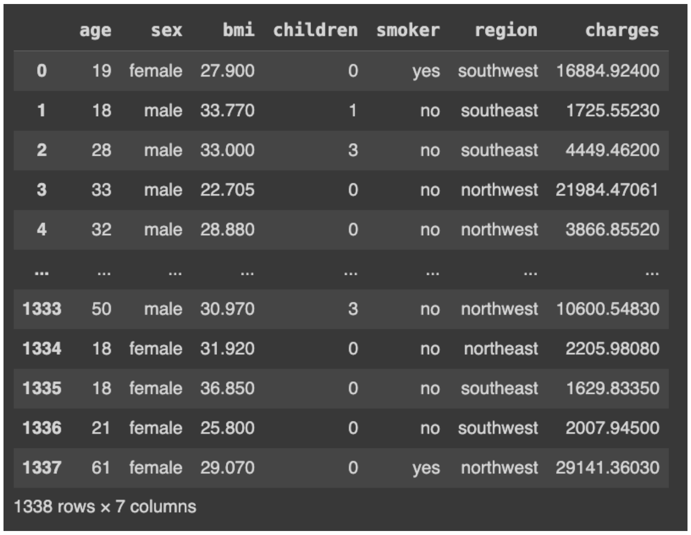
Federal Reserve Bank Economic Data

https://www.kaggle.com/datasets/mirichoi0218/insurance

Medical Cost Personal Datasets. Downloadable Excel, CSV formats. Also available on GitHub https://github.com/stedy/Machine-Learning-with-R-datasets

Description of the data:

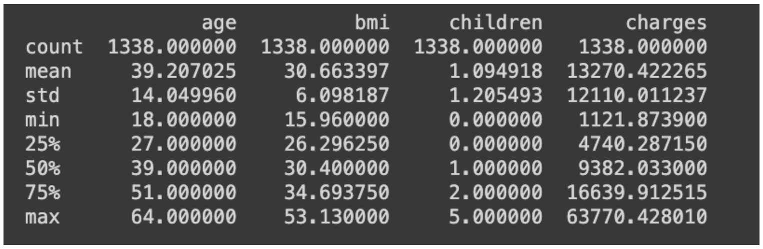
General Overview: The dataset is a collection of medical cost personal datasets aimed at insurance forecast. It is part of the book "Machine Learning with R" by Brett Lantz.



	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
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	northwest	10600.54830
1334	18	female	31.920	0	no	northeast	2205.98080
1335	18	female	36.850	0	no	southeast	1629.83350
1336	21	female	25.800	0	no	southwest	2007.94500
1337	61	female	29.070	0	yes	northwest	29141.36030

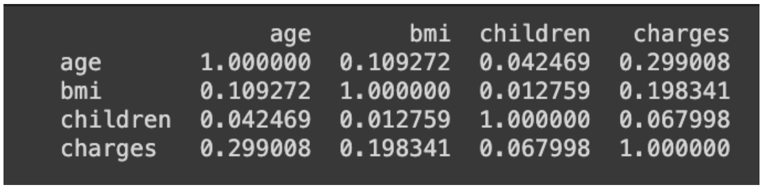
1338 rows × 7 columns

Size and Scope: The dataset includes 1338 records, providing a comprehensive view of individual insurance charges and related factors.



```
              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
```

Summary of how much correlation there is between each columns dataset.



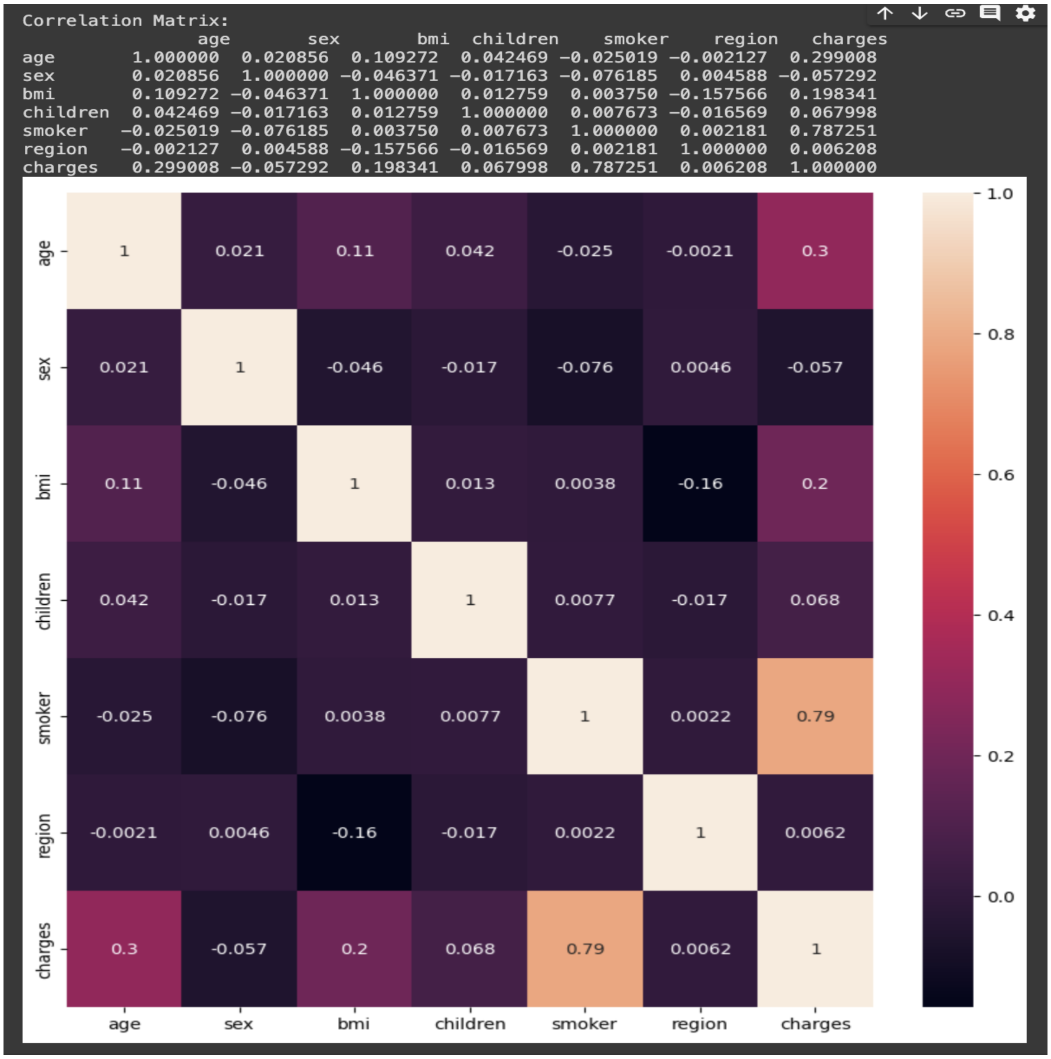
```
               age       bmi  children   charges
age       1.000000  0.109272  0.042469  0.299008
bmi       0.109272  1.000000  0.012759  0.198341
children  0.042469  0.012759  1.000000  0.067998
charges   0.299008  0.198341  0.067998  1.000000
```

Key Features/Variables:

age: age of primary beneficiary
sex: insurance contractor gender, female, male
bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
children: Number of children covered by health insurance / Number of dependents
smoker: Smoking status.
region: Beneficiary's residential area in the US (northeast, southeast, southwest, northwest).
charges: Individual medical costs billed by health insurance.

Data Integrity and Validation
Source Verification: The dataset was sourced from Kaggle, a reliable and reputable data hosting platform.

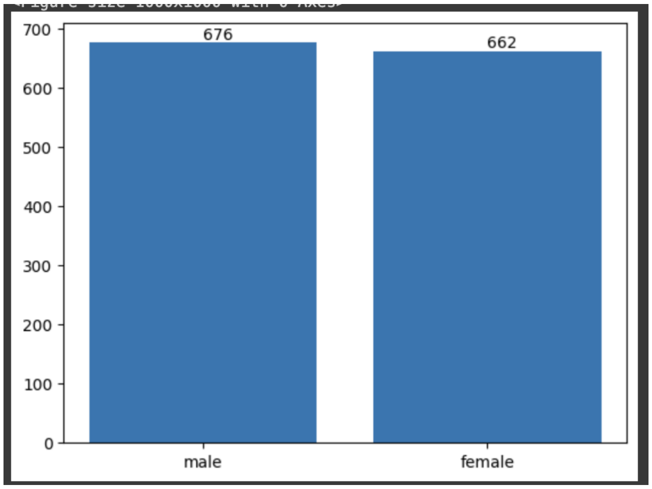
Data visualizations:
Correlation matrix:



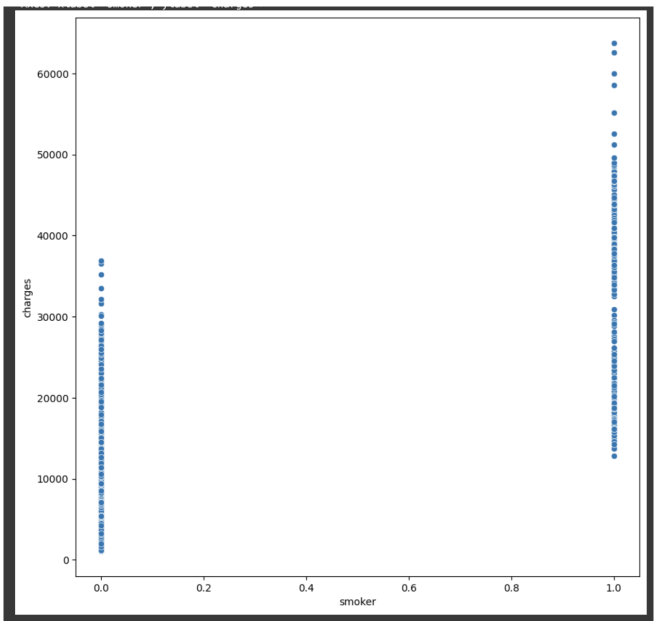
```
Correlation Matrix:
               age       sex       bmi  children    smoker    region   charges
age       1.000000  0.020856  0.109272  0.042469 -0.025019 -0.002127  0.299008
sex       0.020856  1.000000 -0.046371 -0.017163 -0.076185  0.004588 -0.057292
bmi       0.109272 -0.046371  1.000000  0.012759  0.003750 -0.157566  0.198341
children  0.042469 -0.017163  0.012759  1.000000  0.007673 -0.016569  0.067998
smoker   -0.025019 -0.076185  0.003750  0.007673  1.000000  0.002181  0.787251
region   -0.002127  0.004588 -0.157566 -0.016569  0.002181  1.000000  0.006208
charges   0.299008 -0.057292  0.198341  0.067998  0.787251  0.006208  1.000000
```

	age	sex	bmi	children	smoker	region	charges
age	1	0.021	0.11	0.042	-0.025	-0.0021	0.3
sex	0.021	1	-0.046	-0.017	-0.076	0.0046	-0.057
bmi	0.11	-0.046	1	0.013	0.0038	-0.16	0.2
children	0.042	-0.017	0.013	1	0.0077	-0.017	0.068
smoker	-0.025	-0.076	0.0038	0.0077	1	0.0022	0.79
region	-0.0021	0.0046	-0.16	-0.017	0.0022	1	0.0062
charges	0.3	-0.057	0.2	0.068	0.79	0.0062	1

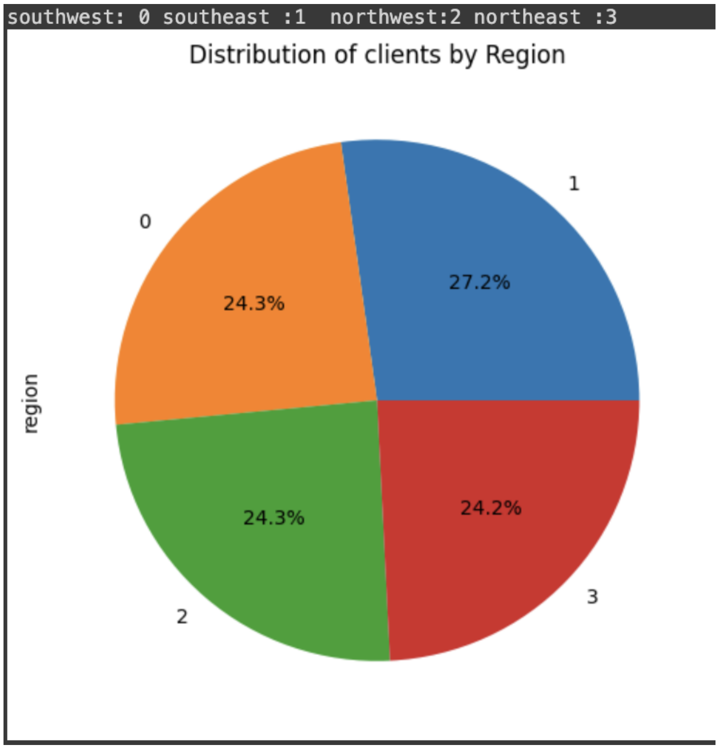
2. The barplot representing how many male and female in the dataset



3. The graph is comparing what is the medical charge between the smoker and non smoker groups.

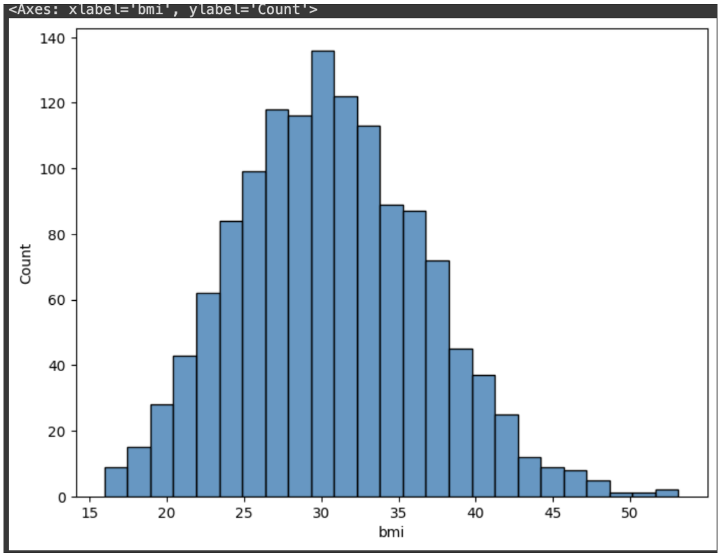


4. SHowing the distribution of the clients based on the region



```
southwest: 0 southeast :1  northwest:2 northeast :3
```

5. Histogram of the BMI of the people in the dataset



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

Special instructions: Download Instructions: The dataset is available for download on Kaggle upon creating an account. Preprocessing Requirements: Some preprocessing might be required, such as encoding categorical variables for analysis, normalizing the value, removing the columns that is not correlated to the output and replacing the non value with the mean of the column. Usage Constraints: The dataset is in the public domain, with no specific usage constraints mentioned