output

June 30, 2023

1 Data Science Project - Predicting Insurance Via Linear Regression

1.1 ## Introduction

From a data set that compiles information on peoples' medical history we implement a linear regression model that attempts to predict the insurance costs of patients.

Data Set Description (source)

- `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
- `region`: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- `charges`: Individual medical costs billed by health insurance

```
Set-up script: parameters_template.py


#### Initial Variables:

<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
```

Data Viewer 1:

	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
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692
10	25	male	26.220	0	no	northeast	2721.32080

Variable Description Before Data Processing:

	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

Missing values (NAs) per column before removal:

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
```

```
Data columns (total 9 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   charges    1338 non-null   float64
 6   northwest  1338 non-null   uint8
 7   southeast  1338 non-null   uint8
 8   southwest  1338 non-null   uint8
dtypes: float64(2), int64(2), object(2), uint8(3)
memory usage: 66.8+ KB
None

#### Variables after transformation:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 9 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   charges    1338 non-null   float64
 6   northwest  1338 non-null   uint8
 7   southeast  1338 non-null   uint8
 8   southwest  1338 non-null   uint8
dtypes: float64(2), int64(2), object(2), uint8(3)
memory usage: 66.8+ KB
None


Data Viewer 2:

    age     sex     bmi  children smoker      charges  northwest  southeast
southwest
0    19  female  27.900         0    yes  16884.92400          0          0
1
1    18    male  33.770         1     no   1725.55230          0          1
0
2    28    male  33.000         3     no   4449.46200          0          1
0
3    33    male  22.705         0     no  21984.47061          1          0
```

```
0
4    32    male  28.880         0     no   3866.85520          1          0
0
5    31  female  25.740         0     no   3756.62160          0          1
0
6    46  female  33.440         1     no   8240.58960          0          1
0
7    37  female  27.740         3     no   7281.50560          1          0
0
8    37    male  29.830         2     no   6406.41070          0          0
0
9    60  female  25.840         0     no  28923.13692          1          0
0
10   25    male  26.220         0     no   2721.32080          0          0
0
```

Variable Description After Data Processing:

	age	bmi	children	charges	northwest	southeast	southwest
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265	0.242900	0.272048	0.242900
std	14.049960	6.098187	1.205493	12110.011237	0.428995	0.445181	0.428995
min	18.000000	15.960000	0.000000	1121.873900	0.000000	0.000000	0.000000
25%	27.000000	26.296250	0.000000	4740.287150	0.000000	0.000000	0.000000
50%	39.000000	30.400000	1.000000	9382.033000	0.000000	0.000000	0.000000
75%	51.000000	34.693750	2.000000	16639.912515	0.000000	1.000000	0.000000
max	64.000000	53.130000	5.000000	63770.428010	1.000000	1.000000	1.000000

- Non-numeric variables in the main data frame:
  - sex
  - smoker

[End Of Report]