output

May 25, 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

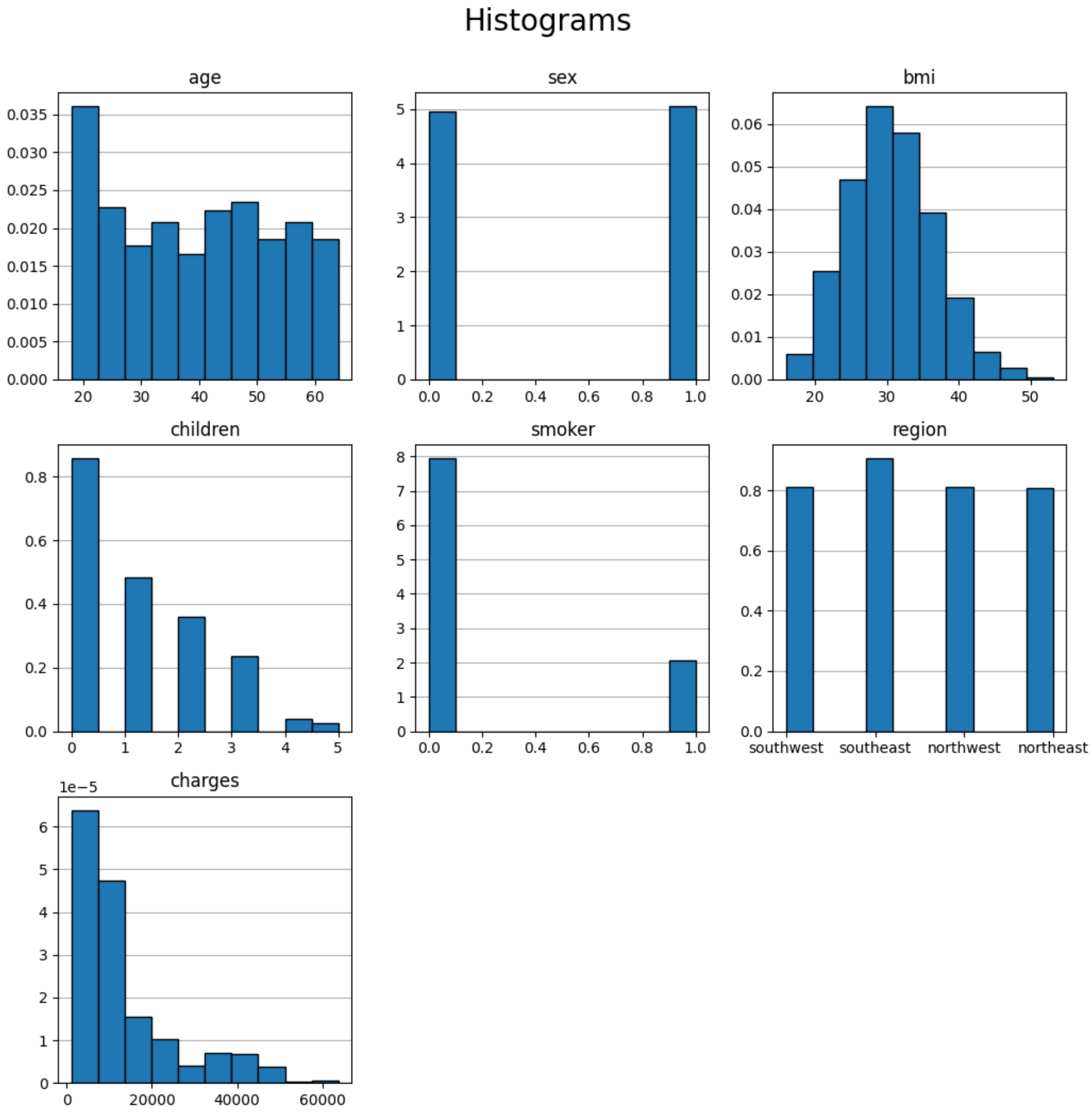
```
#### Initial Variables:

<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
```

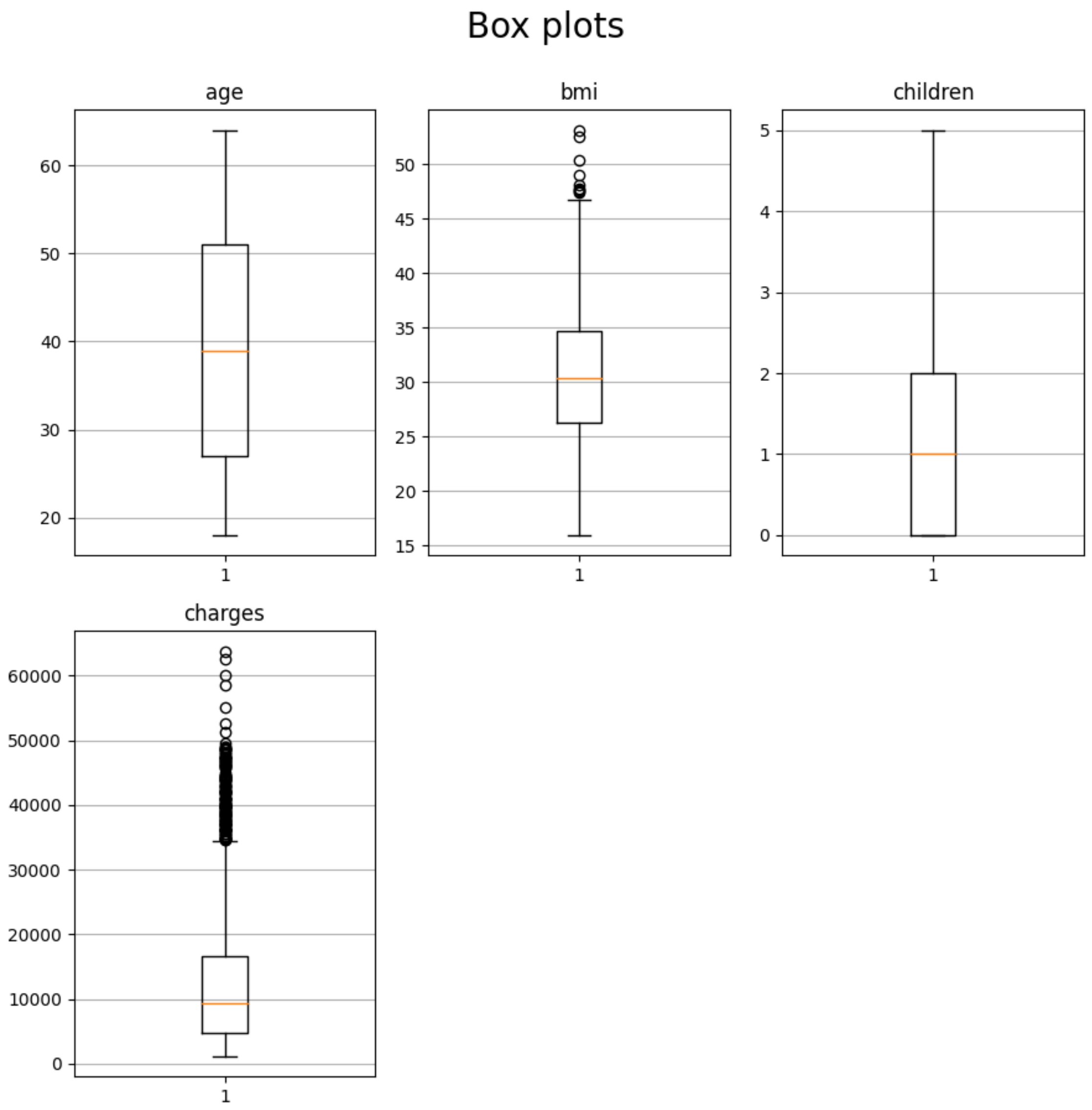
```
First rows:

   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
```

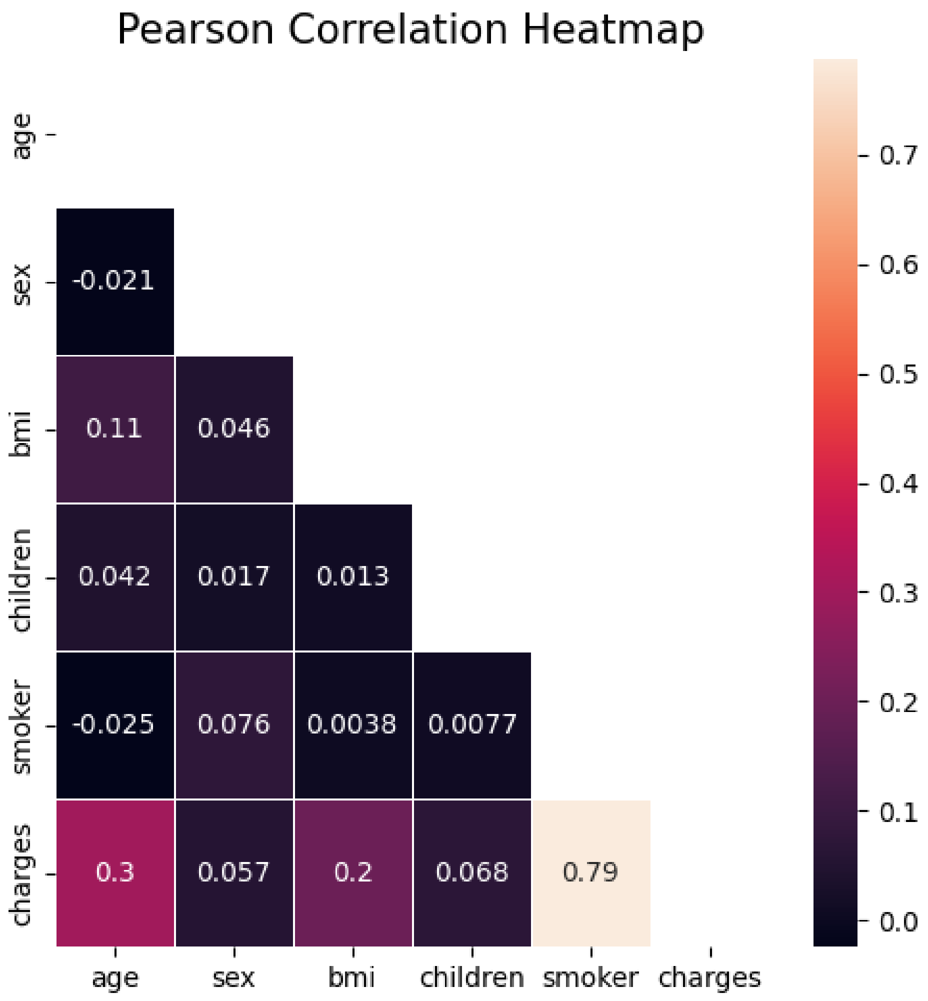
```
Variable Description Before Data Processing:

               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
```

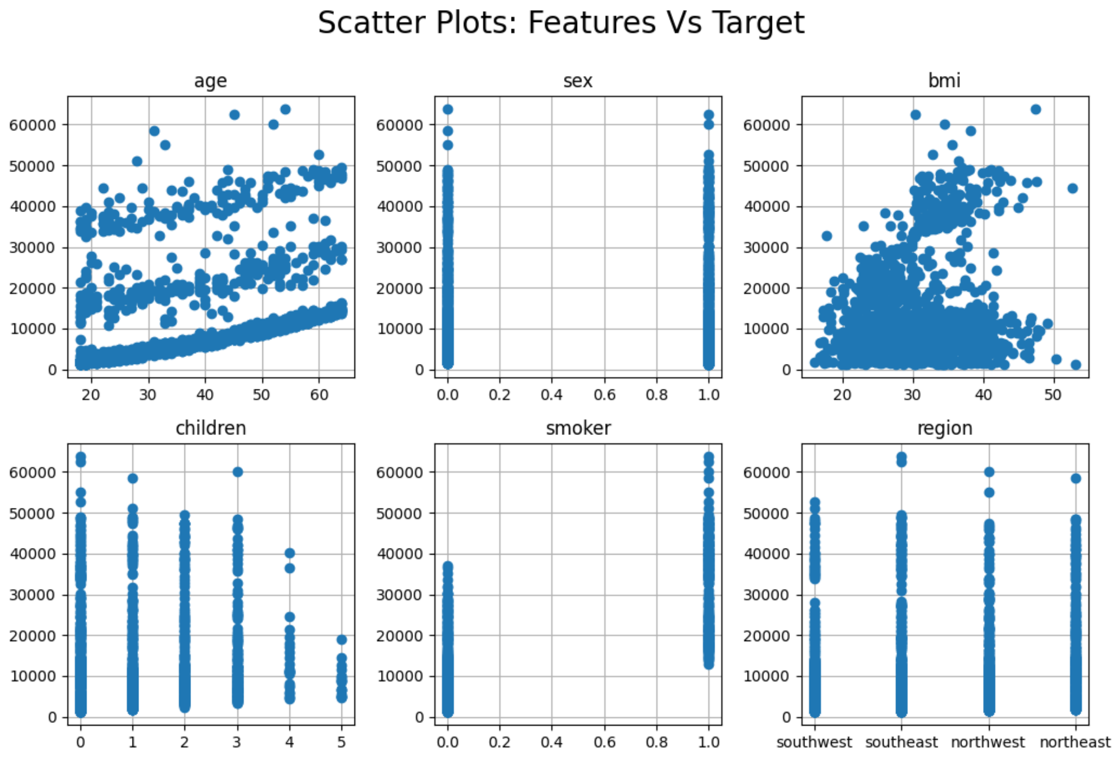
```
#### Variables after transformation:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   age          1338 non-null   int64
 1   sex          1338 non-null   int64
 2   bmi          1338 non-null   float64
 3   children     1338 non-null   int64
 4   smoker       1338 non-null   int64
 5   region       1338 non-null   category
 6   charges      1338 non-null   float64
 7   log_charges  1338 non-null   float64
dtypes: category(1), float64(3), int64(4)
memory usage: 74.8 KB
None

First rows:

   age  sex     bmi  children  smoker     region      charges  log_charges
0   19    0  27.900         0       1  southwest  16884.92400     9.734176
1   18    1  33.770         1       0  southeast   1725.55230     7.453302
```

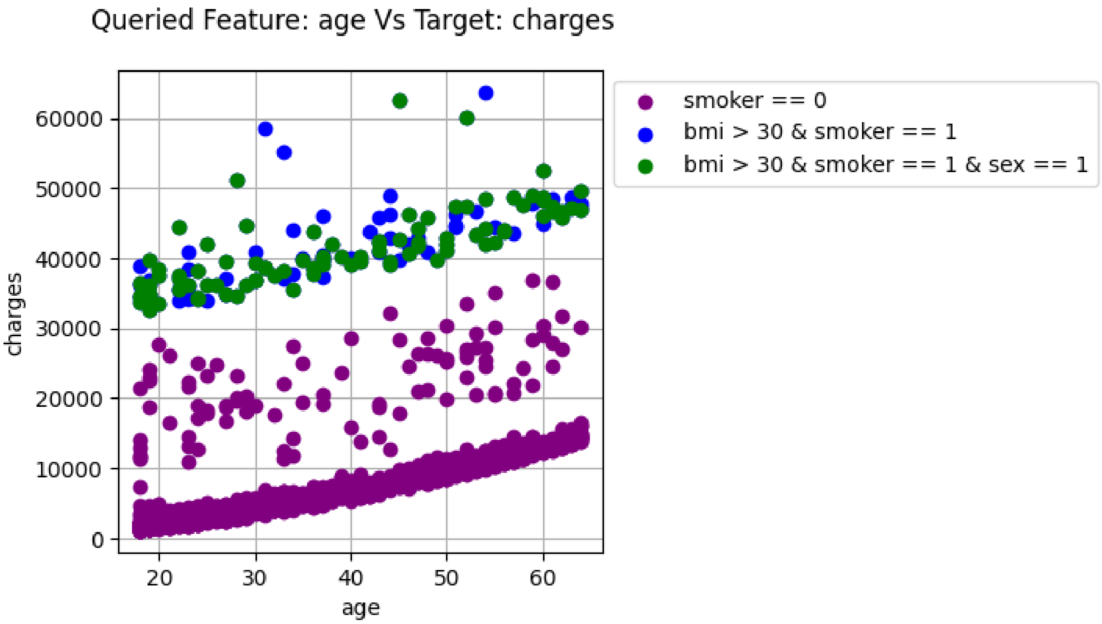
```
2   28    1  33.000         3       0  southeast   4449.46200     8.400538
3   33    1  22.705         0       0  northwest  21984.47061     9.998092
4   32    1  28.880         0       0  northwest   3866.85520     8.260197

Variable Description After Data Processing:

               age          sex          bmi     children       smoker  \
count  1338.000000  1338.000000  1338.000000  1338.000000  1338.000000
mean     39.207025     0.505232    30.663397     1.094918     0.204783
std      14.049960     0.500160     6.098187     1.205493     0.403694
min      18.000000     0.000000    15.960000     0.000000     0.000000
25%      27.000000     0.000000    26.296250     0.000000     0.000000
50%      39.000000     1.000000    30.400000     1.000000     0.000000
75%      51.000000     1.000000    34.693750     2.000000     0.000000
max      64.000000     1.000000    53.130000     5.000000     1.000000

            charges  log_charges
count   1338.000000  1338.000000
mean   13270.422265     9.098659
std    12110.011237     0.919527
min     1121.873900     7.022756
25%     4740.287150     8.463853
50%     9382.033000     9.146552
75%    16639.912515     9.719558
max    63770.428010    11.063045
```


Histograms
age
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
20
30
40
50
60
sex
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
bmi
0.00
0.01
0.02
0.03
0.04
0.05
0.06
20
30
40
50
children
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
smoker
0
1
2
3
4
5
6
7
8
0.0
0.2
0.4
0.6
0.8
1.0
region
0.0
0.2
0.4
0.6
0.8
southwest
southeast
northwest
northeast
charges
1e−5
0
1
2
3
4
5
6
0
20000
40000
60000


Box plots
age
60
50
40
30
20
1
bmi
50
45
40
35
30
25
20
15
1
children
5
4
3
2
1
0
1
charges
60000
50000
40000
30000
20000
10000
0
1

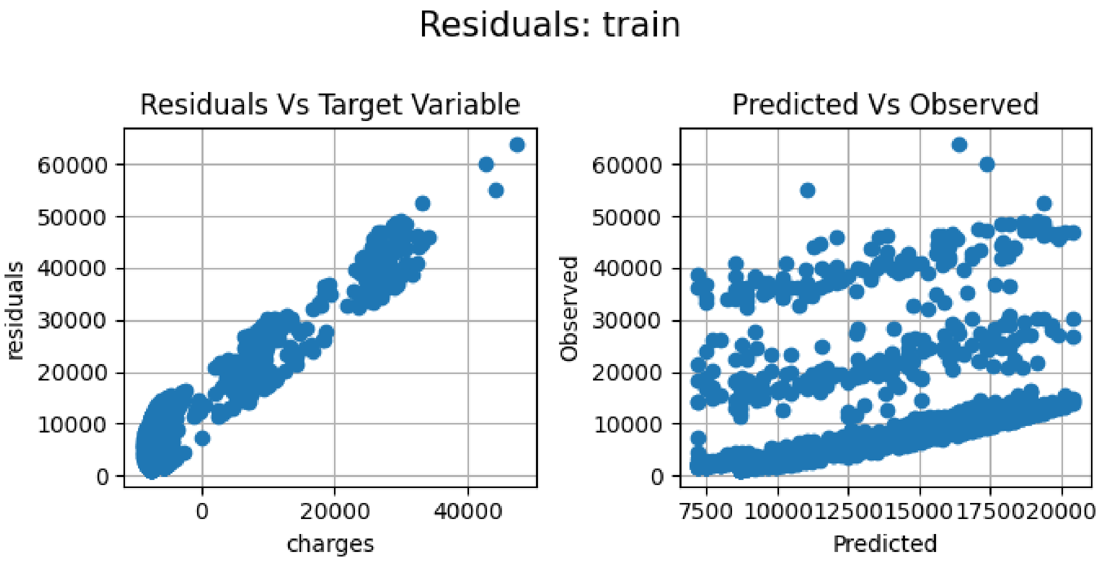

Pearson Correlation Heatmap
age
sex
bmi
children
smoker
charges
-0.021
0.11
0.046
0.042
0.017
0.013
-0.025
0.076
0.0038
0.0077
0.3
0.057
0.2
0.068
0.79
age
sex
bmi
children
smoker
charges
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

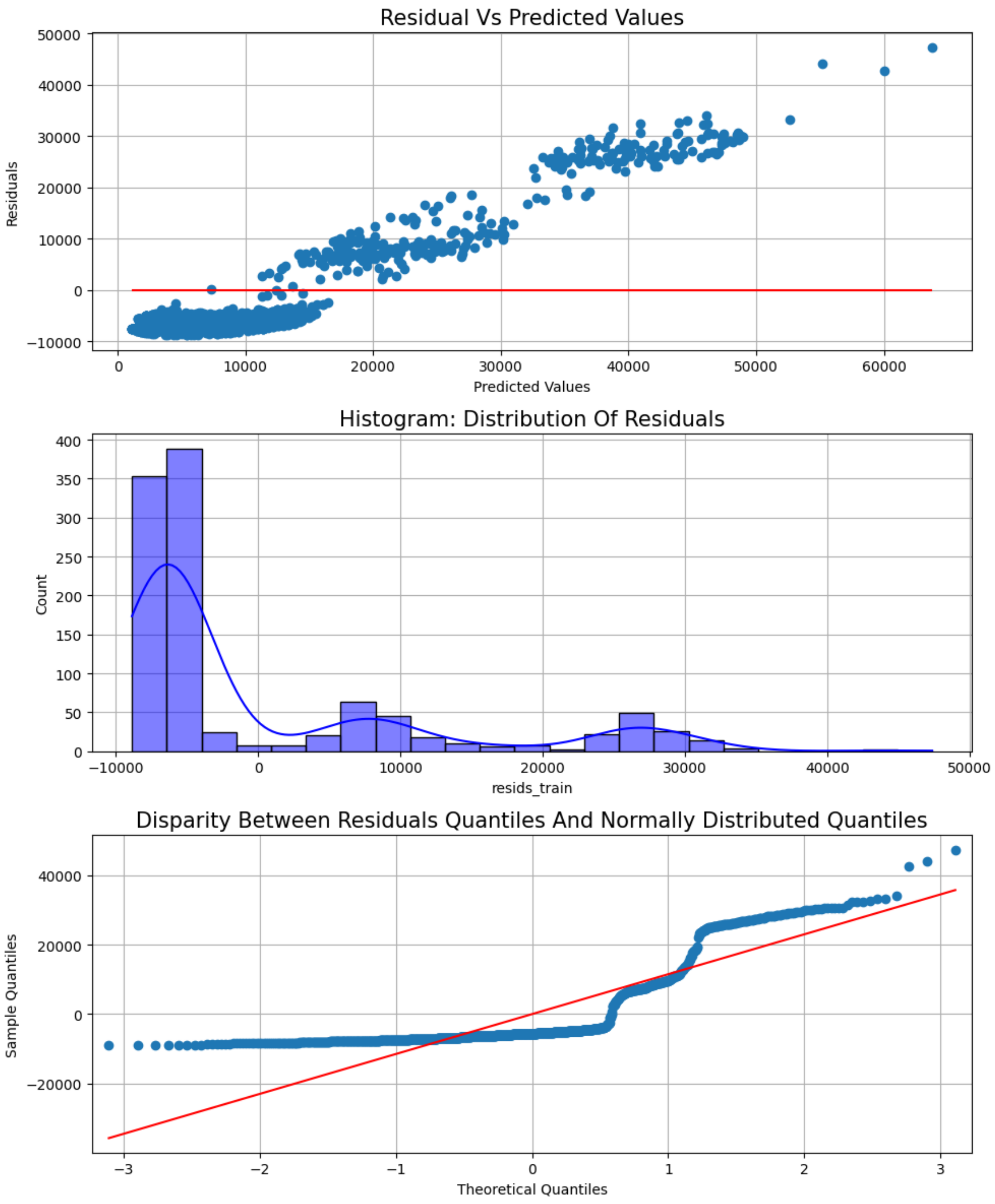




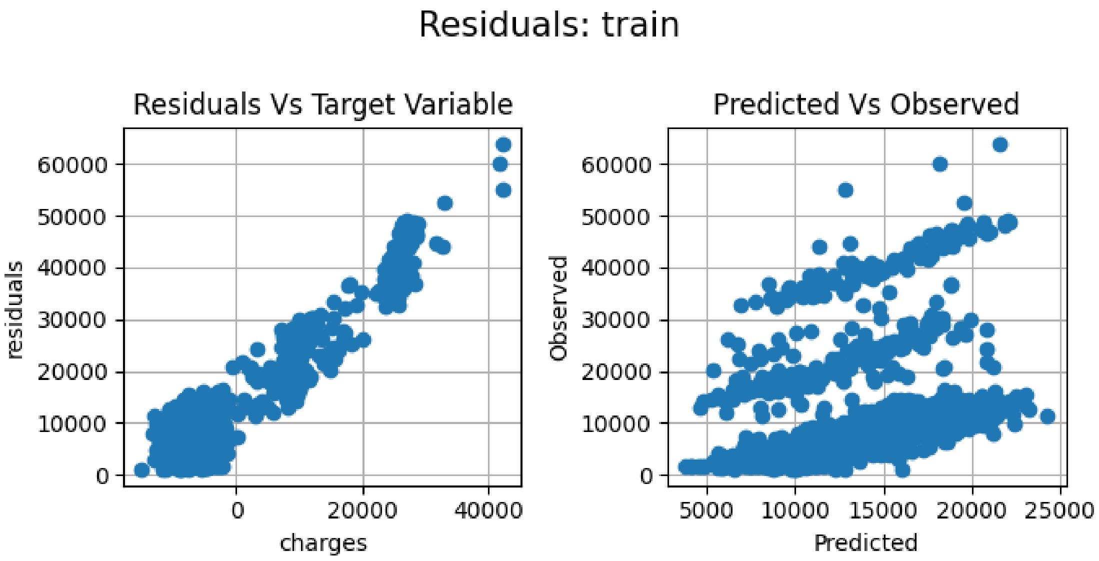
## Analysis of Variance Inflation Factor:

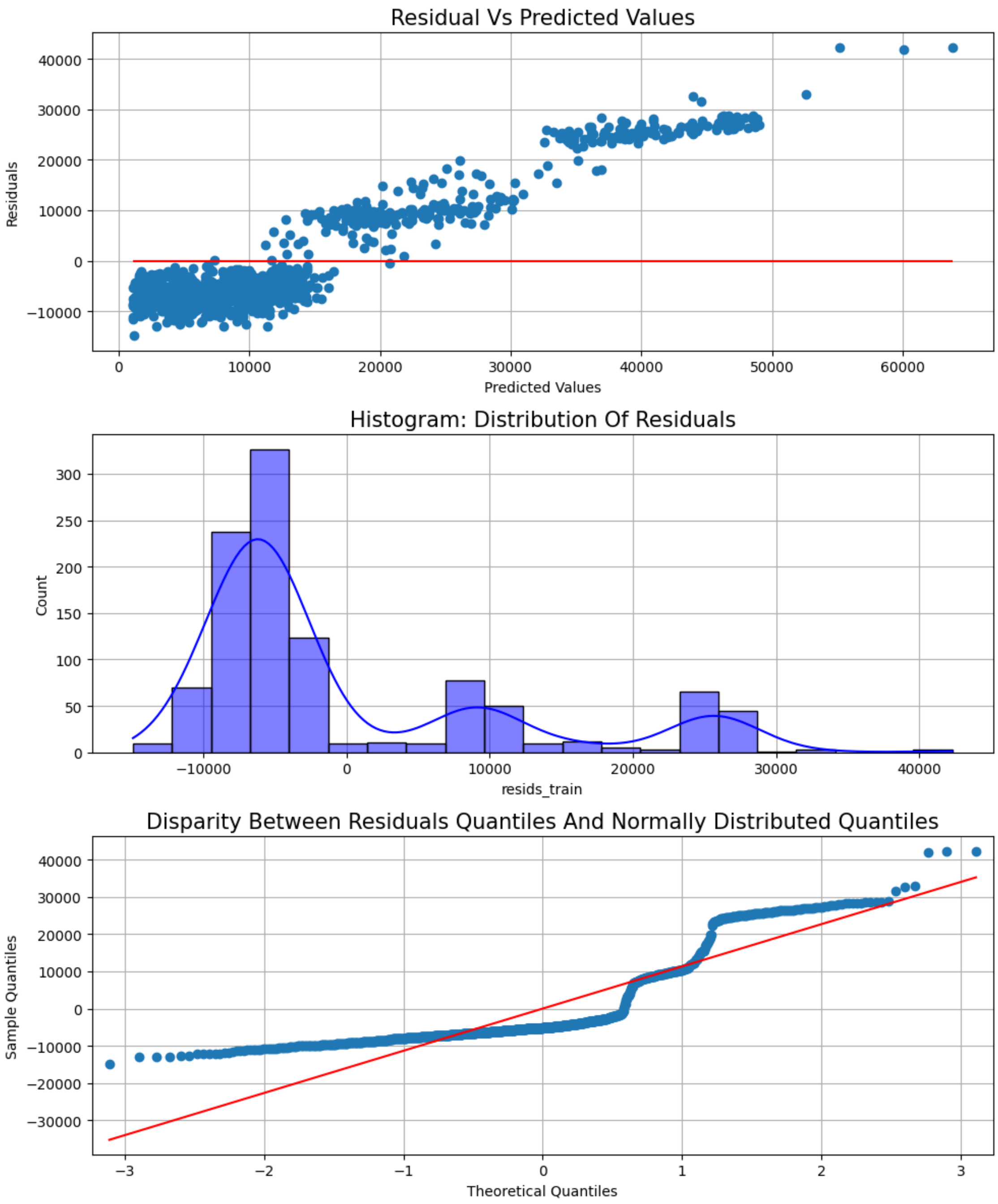
```
Features group: 1

sex    1.000435
age    1.000435
Name: VIF, dtype: float64

------------------------------------

Features group: 2

sex    1.002838
age    1.012775
bmi    1.014516
Name: VIF, dtype: float64

------------------------------------

Features group: 3

age       1.000988
sex       1.006202
smoker    1.006394
Name: VIF, dtype: float64

------------------------------------

Features group: 4

children    1.002242
smoker      1.006457
sex         1.008878
bmi         1.014578
age         1.015129
Name: VIF, dtype: float64

------------------------------------


#### Regression Results


## Regression number: 1

Target variable (Y): charges

Explanatory Variables:
    - x1: age
    - x2: sex

                   Results: Ordinary least squares
===================================================================
Model:              OLS              Adj. R-squared:     0.088
Dependent Variable: y                AIC:                23049.7374
Date:               2023-05-25 18:56 BIC:                23064.6636
No. Observations:   1070             Log-Likelihood:     -11522.
Df Model:           2                F-statistic:        52.66
Df Residuals:       1067             Prob (F-statistic): 1.55e-22
R-squared:          0.090            Scale:              1.3237e+08
--------------------------------------------------------------------
            Coef.     Std.Err.     t     P>|t|    [0.025     0.975]
--------------------------------------------------------------------
const     2605.5835  1128.9424   2.3080  0.0212  390.3843  4820.7826
x1         255.4248    25.2778  10.1047  0.0000  205.8251   305.0246
x2        1462.8520   703.7863   2.0785  0.0379   81.8898  2843.8142
-------------------------------------------------------------------
Omnibus:              292.892       Durbin-Watson:          1.933
Prob(Omnibus):        0.000         Jarque-Bera (JB):       581.679
Skew:                 1.642         Prob(JB):               0.000
Kurtosis:             4.506         Condition No.:          139
===================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.

## Error measurement:

MSE: 136290734.7
RMSE: 11674.36

## Residuals Analysis for the train set.

Test: Shapiro-Wilk
    - Statistic: 0.6924, p-value: 0.0
Test: D'Agostino's
    - Statistic: 292.8918, p-value: 0.0
Test: Kolmogorov-Smirnov
    - Statistic: 0.7206, p-value: 0.0
Test: Jarque-Bera
    - Statistic: 581.6791, p-value: 0.0
```


Residuals: train
Residuals Vs Target Variable
residuals
charges
0
10000
20000
30000
40000
50000
60000
0
20000
40000
Predicted Vs Observed
Observed
Predicted
0
10000
20000
30000
40000
50000
60000
7500
10000
12500
15000
17500
20000

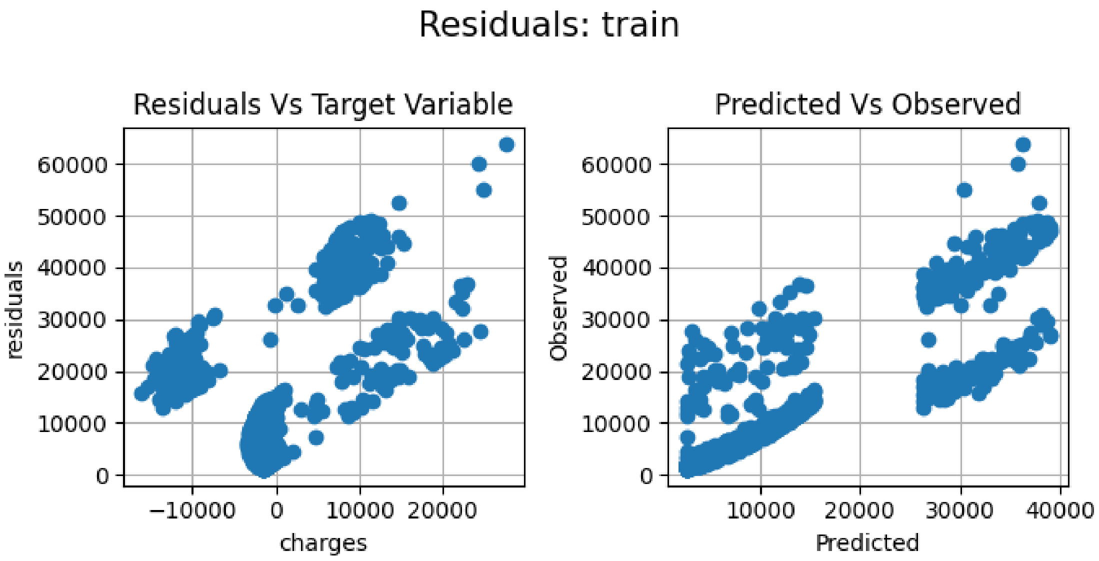


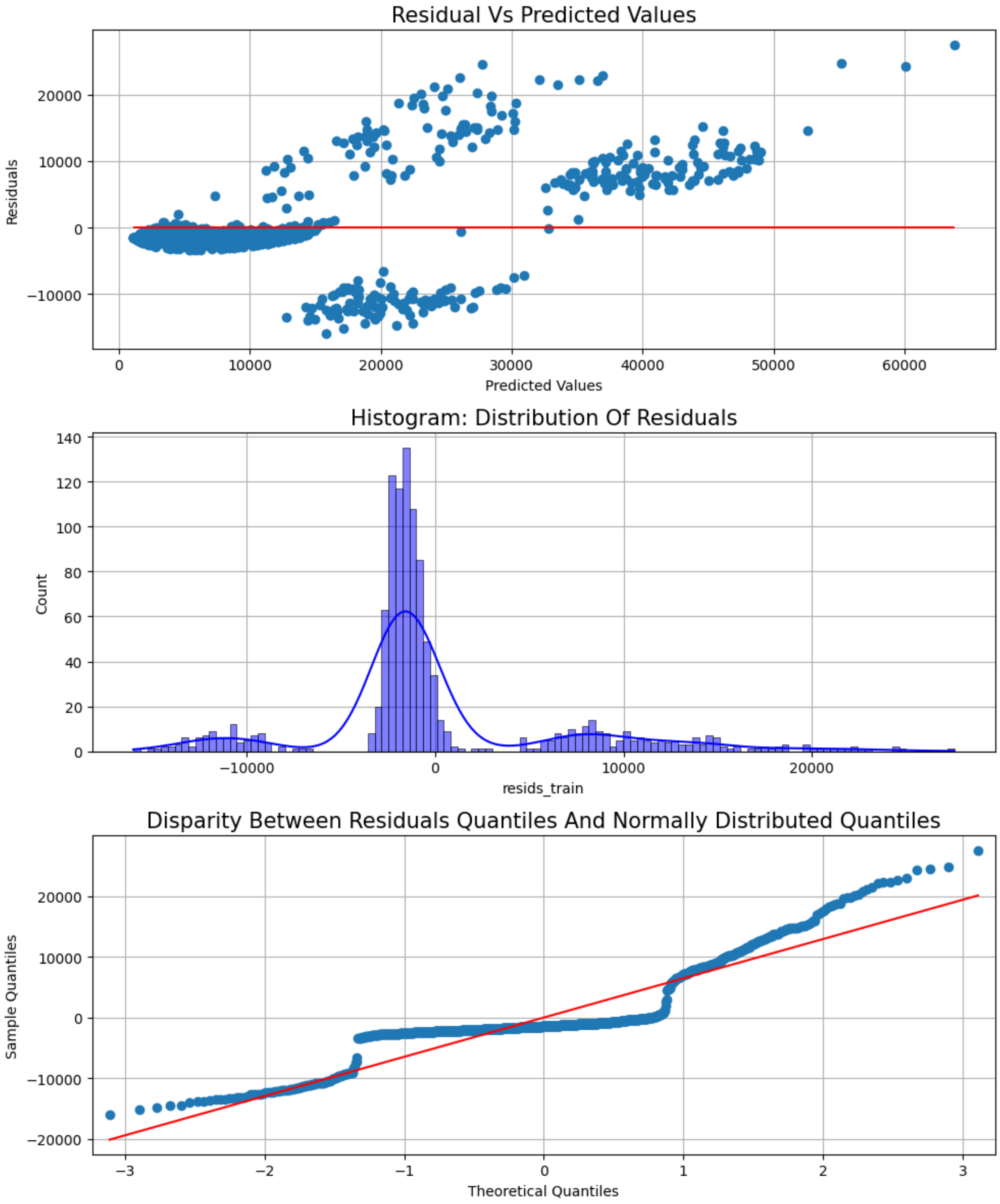
```
## Regression number: 2

Target variable (Y): charges

Explanatory Variables:
    - x1: age
    - x2: sex
    - x3: bmi

                   Results: Ordinary least squares
===================================================================
Model:              OLS              Adj. R-squared:     0.113
Dependent Variable: y                AIC:                23021.0051
Date:               2023-05-25 18:56 BIC:                23040.9068
No. Observations:   1070             Log-Likelihood:     -11507.
Df Model:           3                F-statistic:        46.45
Df Residuals:       1066             Prob (F-statistic): 3.26e-28
R-squared:          0.116            Scale:              1.2875e+08
-------------------------------------------------------------------
           Coef.     Std.Err.    t     P>|t|     [0.025     0.975]
-------------------------------------------------------------------
const    -6373.0901 1958.3127 -3.2544 0.0012 -10215.6752 -2530.5049
x1         238.2345   25.1191  9.4842 0.0000    188.9459   287.5230
x2        1199.9390  695.6781  1.7248 0.0848   -165.1149  2564.9929
x3         318.9589   57.2301  5.5733 0.0000    206.6624   431.2554
-------------------------------------------------------------------
Omnibus:              233.907       Durbin-Watson:          1.933
Prob(Omnibus):        0.000         Jarque-Bera (JB):       403.213
Skew:                 1.434         Prob(JB):               0.000
Kurtosis:             3.907         Condition No.:          292
===================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.

## Error measurement:

MSE: 131639993.92
RMSE: 11473.45

## Residuals Analysis for the train set.

Test: Shapiro-Wilk
    - Statistic: 0.7686, p-value: 0.0
Test: D'Agostino's
    - Statistic: 233.9065, p-value: 0.0
Test: Kolmogorov-Smirnov
    - Statistic: 0.7196, p-value: 0.0
Test: Jarque-Bera
    - Statistic: 403.2125, p-value: 0.0
```


Residuals: train
Residuals Vs Target Variable
residuals
charges
0
10000
20000
30000
40000
50000
60000
0
20000
40000
Predicted Vs Observed
Observed
Predicted
0
10000
20000
30000
40000
50000
60000
5000
10000
15000
20000
25000



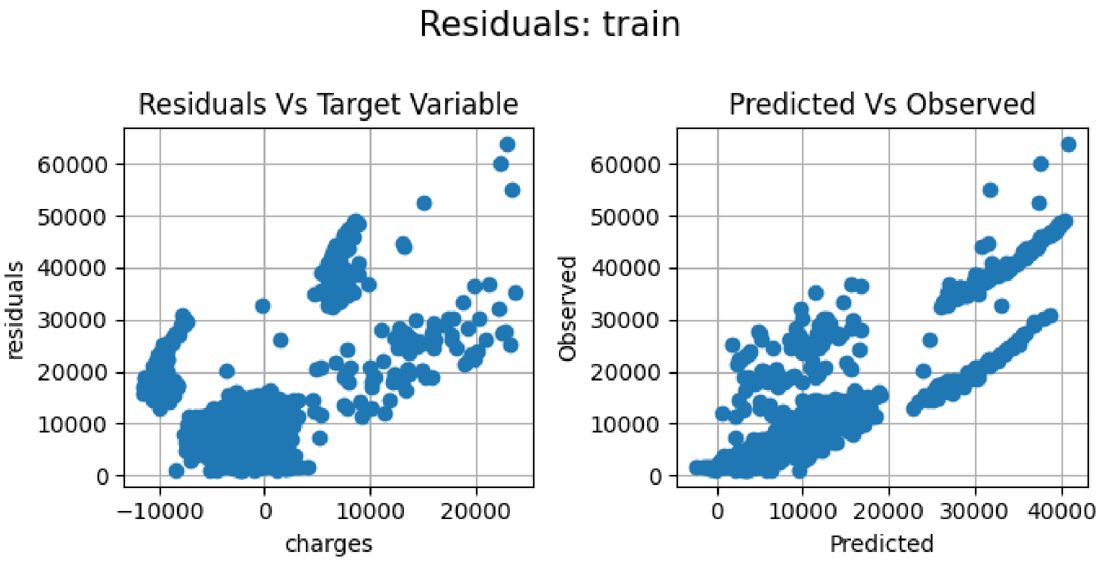
```
## Regression number: 3

Target variable (Y): charges

Explanatory Variables:
    - x1: age
    - x2: sex
    - x3: smoker

                   Results: Ordinary least squares
===================================================================
Model:              OLS              Adj. R-squared:     0.712
Dependent Variable: y                AIC:                21819.0882
Date:               2023-05-25 18:56 BIC:                21838.9899
No. Observations:   1070             Log-Likelihood:     -10906.
Df Model:           3                F-statistic:        880.1
Df Residuals:       1066             Prob (F-statistic): 7.72e-288
R-squared:          0.712            Scale:              4.1869e+07
--------------------------------------------------------------------
         Coef.     Std.Err.     t     P>|t|     [0.025      0.975]
--------------------------------------------------------------------
const  -2374.9705  643.3317  -3.6917  0.0002  -3637.3107  -1112.6303
x1       277.1770   14.2235  19.4872  0.0000    249.2677    305.0862
x2        42.4621  396.9149   0.1070  0.9148   -736.3612    821.2854
x3     23630.4702  491.9384  48.0354  0.0000  22665.1927  24595.7476
-------------------------------------------------------------------
Omnibus:              198.150       Durbin-Watson:          2.072
Prob(Omnibus):        0.000         Jarque-Bera (JB):       413.935
Skew:                 1.053         Prob(JB):               0.000
Kurtosis:             5.202         Condition No.:          142
===================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.

## Error measurement:

MSE: 37362100.73
RMSE: 6112.45

## Residuals Analysis for the train set.

Test: Shapiro-Wilk
    - Statistic: 0.8169, p-value: 0.0
Test: D'Agostino's
    - Statistic: 198.1497, p-value: 0.0
Test: Kolmogorov-Smirnov
    - Statistic: 0.7776, p-value: 0.0
Test: Jarque-Bera
    - Statistic: 413.9351, p-value: 0.0
```

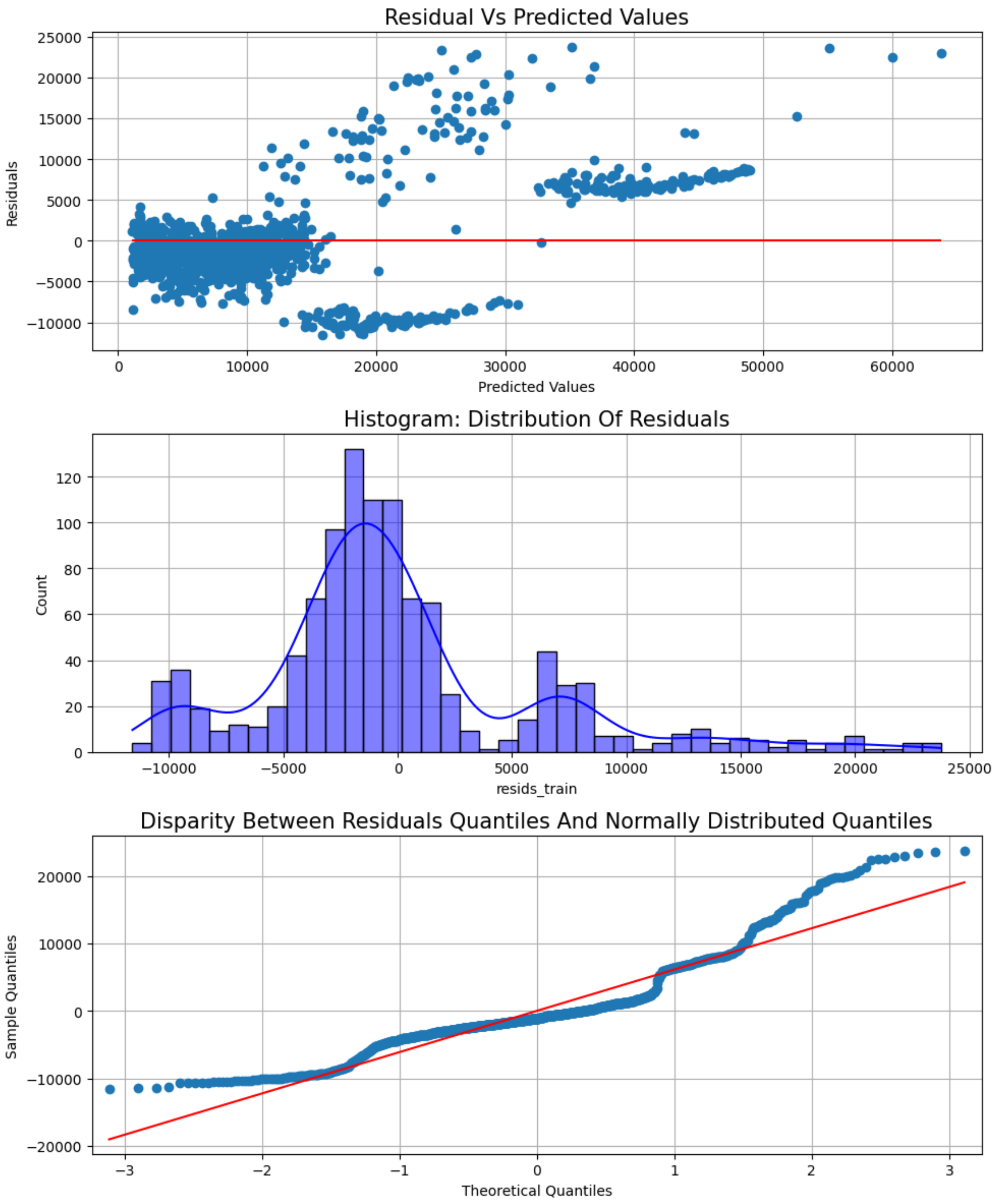

Residuals: train
Residuals Vs Target Variable
residuals
charges
60000
50000
40000
30000
20000
10000
0
−10000
0
10000
20000
Predicted Vs Observed
Observed
Predicted
60000
50000
40000
30000
20000
10000
0
10000
20000
30000
40000



```
## Regression number: 4

Target variable (Y): charges

Explanatory Variables:
    - x1: age
    - x2: sex
    - x3: bmi
    - x4: children
    - x5: smoker

                   Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.741
Dependent Variable: y                AIC:                21707.6711
Date:               2023-05-25 18:56 BIC:                21737.5236
No. Observations:   1070             Log-Likelihood:     -10848.
Df Model:           5                F-statistic:        611.4
Df Residuals:       1064             Prob (F-statistic): 8.65e-310
R-squared:          0.742            Scale:              3.7659e+07
------------------------------------------------------------------
          Coef.     Std.Err.    t     P>|t|     [0.025     0.975]
------------------------------------------------------------------
const  -11922.8934 1071.8911 -11.1232 0.0000 -14026.1540 -9819.6328
x1        257.2126   13.6116  18.8965 0.0000    230.5039   283.9212
x2       -266.7664  377.4882  -0.7067 0.4799  -1007.4724   473.9395
x3        321.6202   30.9536  10.3904 0.0000    260.8832   382.3572
x4        559.8364  158.1266   3.5404 0.0004    249.5610   870.1119
x5      23622.1141  466.5819  50.6280 0.0000  22706.5890 24537.6392
------------------------------------------------------------------
Omnibus:              220.123       Durbin-Watson:         2.075
Prob(Omnibus):        0.000         Jarque-Bera (JB):      451.903
Skew:                 1.172         Prob(JB):              0.000
Kurtosis:             5.156         Condition No.:         296
==================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors
is correctly specified.

## Error measurement:

MSE: 33733072.88
RMSE: 5808.02

## Residuals Analysis for the train set.

Test: Shapiro-Wilk
    - Statistic: 0.9019, p-value: 0.0
Test: D'Agostino's
    - Statistic: 220.1229, p-value: 0.0
Test: Kolmogorov-Smirnov
    - Statistic: 0.6355, p-value: 0.0
```

```
Test: Jarque-Bera
    - Statistic: 451.9032, p-value: 0.0
```





** [No more experiments] **

## Error Measurement Comparison

```
                                          mse      rmse
age, sex, bmi, children, smoker   33733072.88   5808.02
age, sex, smoker                  37362100.73   6112.45
age, sex, bmi                    131639993.92  11473.45
age, sex                          136290734.7  11674.36


[End Of Report]
```