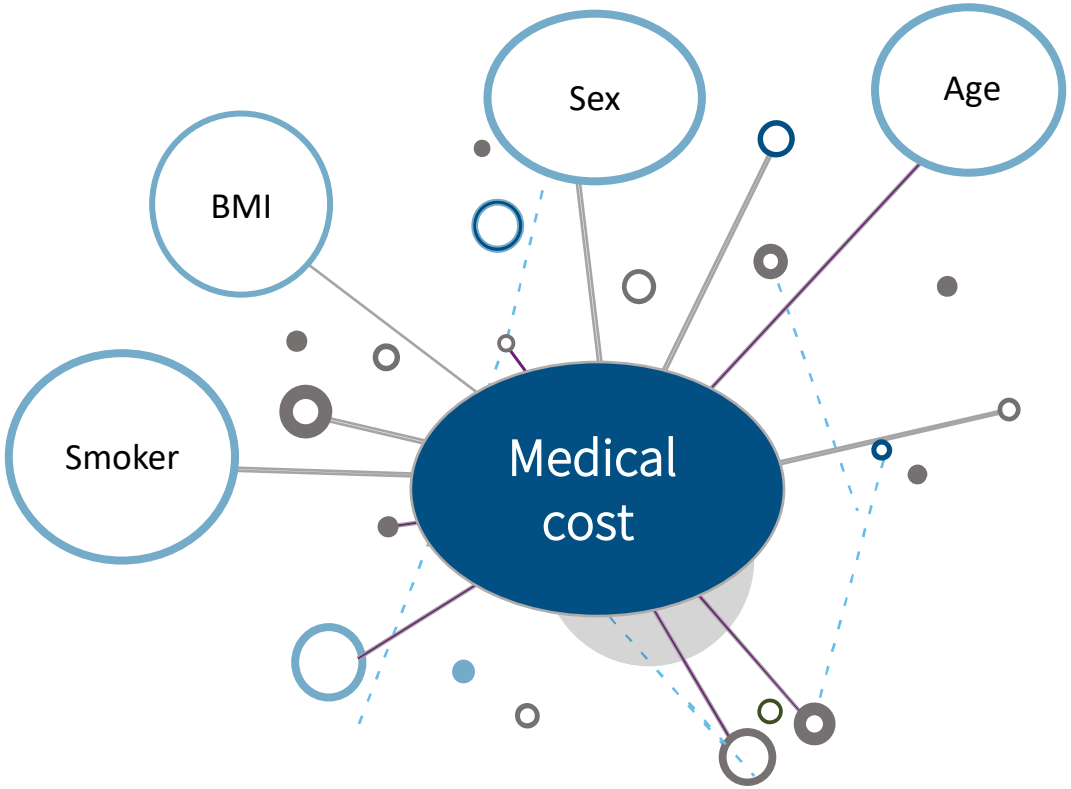
Medical cost

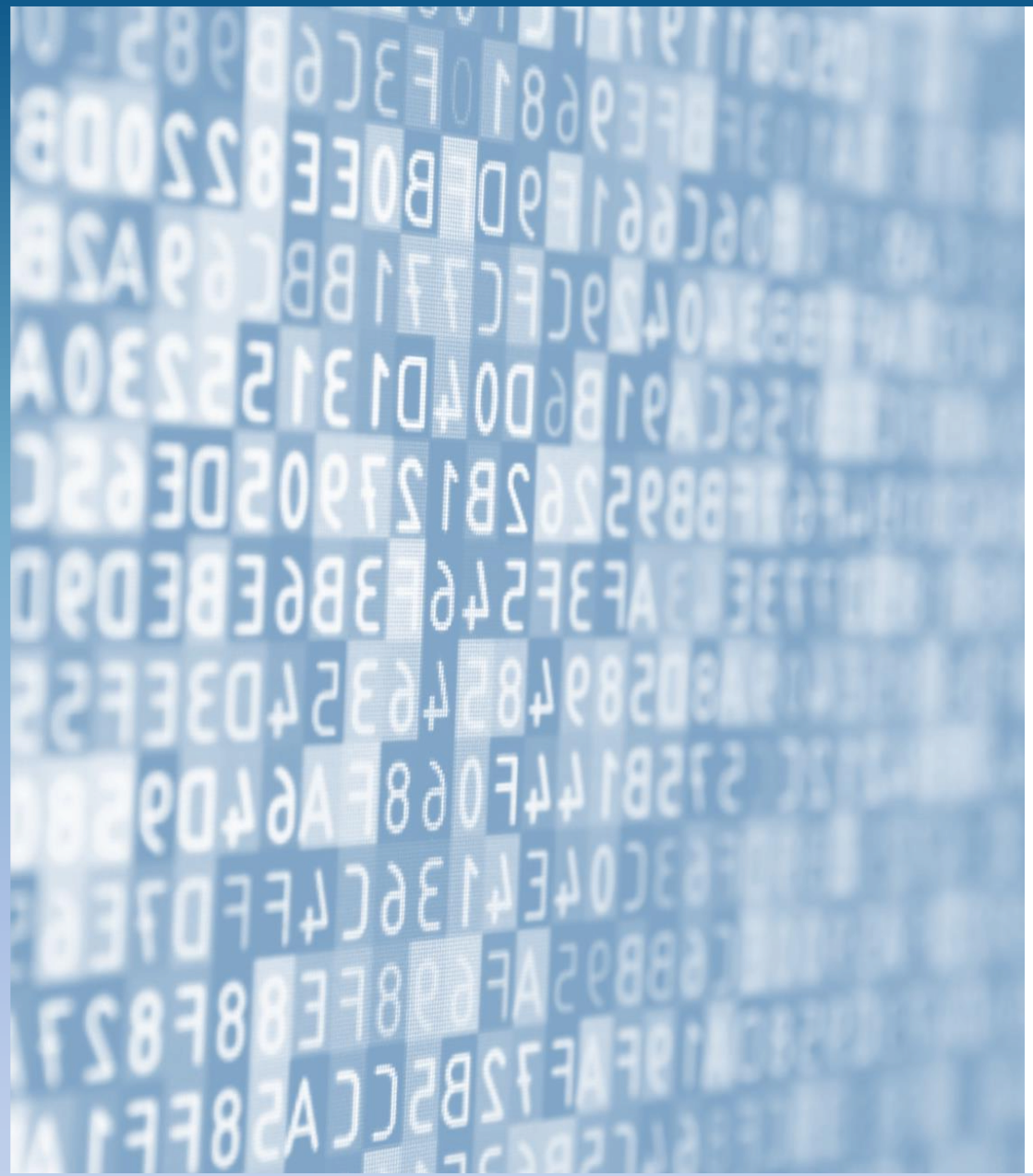
Regresion Project

Tomer badug. Shirli miller. Judi Eliya

project goal



prediction the medical cost



Data content

the input variables are:

1. Age
2. Sex
3. BMI
4. Children
5. Smoker

Output variable – Charges (Medical cost)

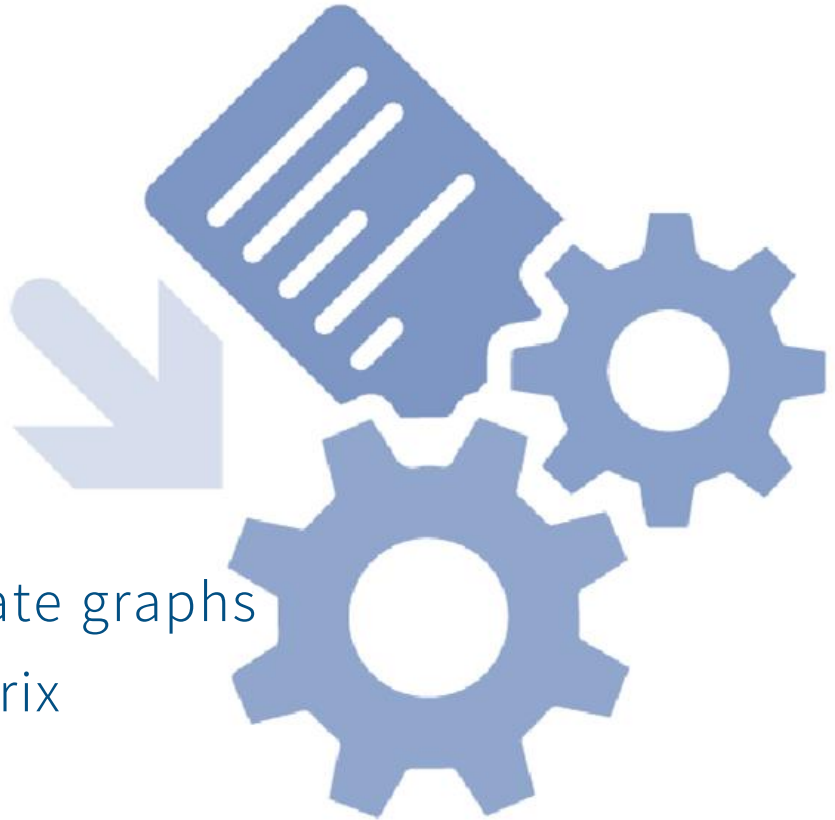
The Dataset

1338 row and 7 columns

- **Age** - age of primary beneficiary (18-64)
- **Sex** - insurance contractor gender: female / male
- **BMI** - Body mass index (kg / m ^ 2) using the ratio of height to weight (ideally 18.5 to 24.9)
- **Children** - Number of dependents
- **Smoker** – Yes / No
- **Region** – divided to: northeast, southeast, southwest, northwest of US
- **Charges** - Individual medical costs billed (currency amount in thousands)

pre processing - EDA

- Handling null
- Remove duplicate
- Label Encoder
- describe
- pairwise relationships in a dataset
- Explore two variables with bivariate and univariate graphs
- Compute pairwise correlation of columns – matrix



pre processing - EDA

- Handling null

```
[ ]   1 df.isnull().sum()
```

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

pre processing - EDA

- Remove duplicate

```
[ ]  1 df[df.duplicated()]
```

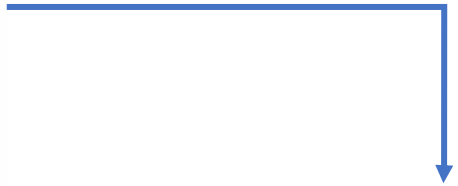
	age	sex	bmi	children	smoker	region	charges
581	19	male	30.59	0	no	northwest	1.64

pre processing - EDA

- Label Encoder

```
1 df.head()
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16.88
1	18	male	33.770	1	no	southeast	1.73
2	28	male	33.000	3	no	southeast	4.45
3	33	male	22.705	0	no	northwest	21.98
4	32	male	28.880	0	no	northwest	3.87



```
['female' 'male']
['no' 'yes']
['northeast' 'northwest' 'southeast' 'southwest']
```

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16.88
1	18	1	33.770	1	0	2	1.73
2	28	1	33.000	3	0	2	4.45
3	33	1	22.705	0	0	1	21.98
4	32	1	28.880	0	0	1	3.87

pre processing - EDA

- describe

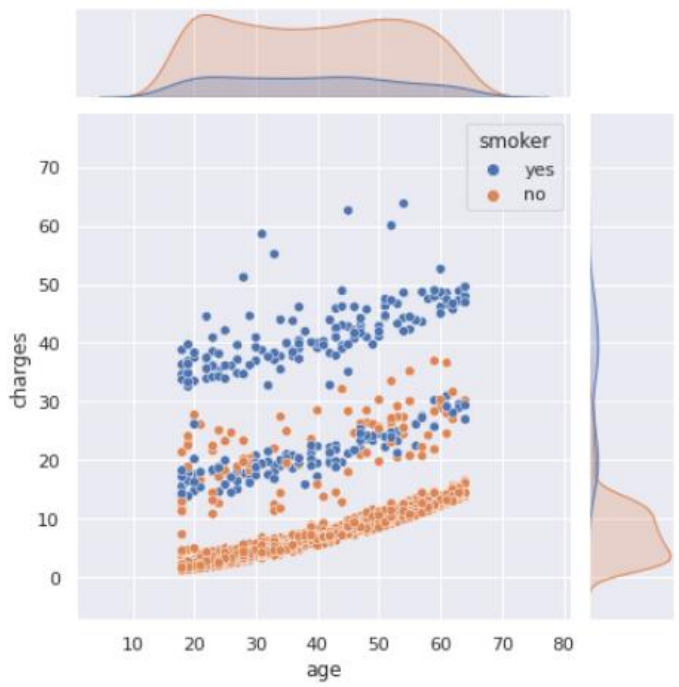
	age	sex	bmi	children	smoker	region	charges
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	0.505232	30.663397	1.094918	0.204783	1.515695	13.270433
std	14.049960	0.500160	6.098187	1.205493	0.403694	1.104885	12.109948
min	18.000000	0.000000	15.960000	0.000000	0.000000	0.000000	1.120000
25%	27.000000	0.000000	26.296250	0.000000	0.000000	1.000000	4.742500
50%	39.000000	1.000000	30.400000	1.000000	0.000000	2.000000	9.385000
75%	51.000000	1.000000	34.693750	2.000000	0.000000	2.000000	16.642500
max	64.000000	1.000000	53.130000	5.000000	1.000000	3.000000	63.770000

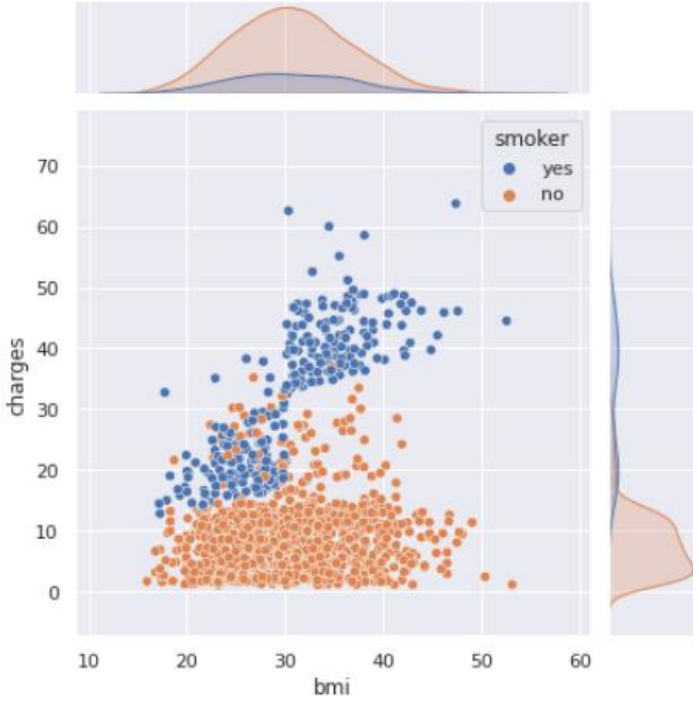
pre processing - EDA

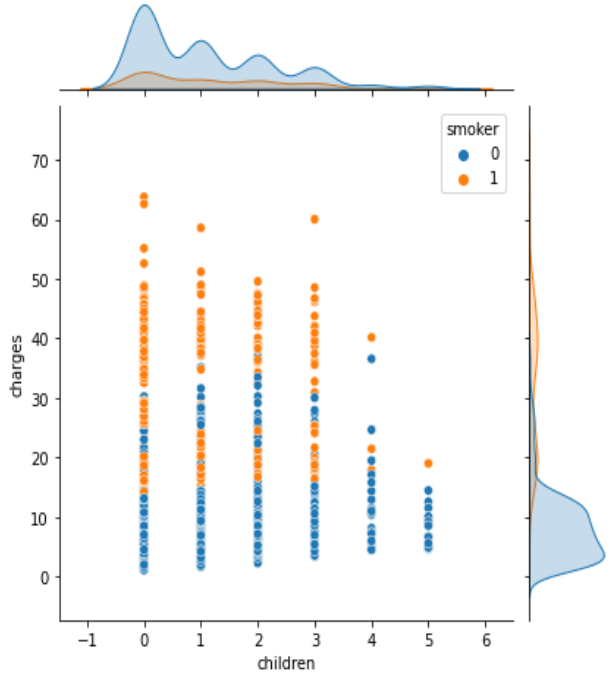
- pairwise relationships in the dataset

pre processing - EDA

- Explor two variables with bivariate and univariate graphs



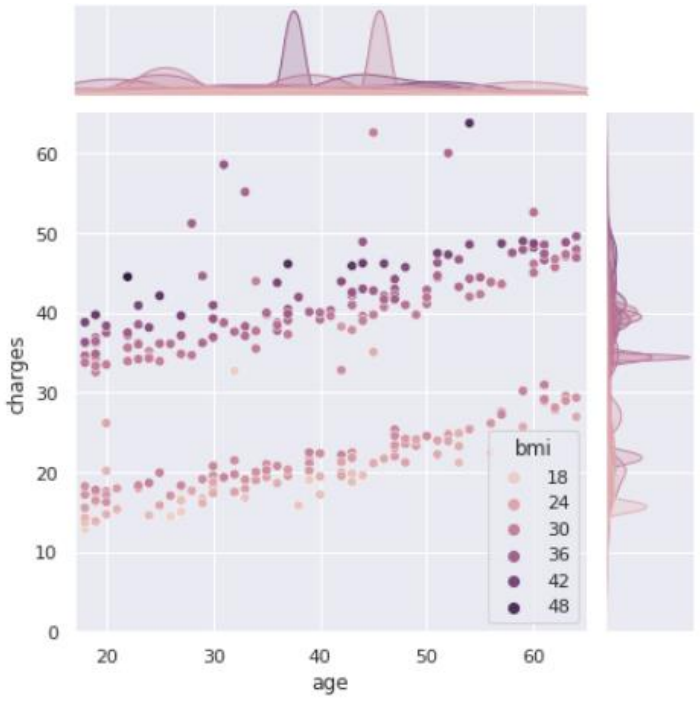




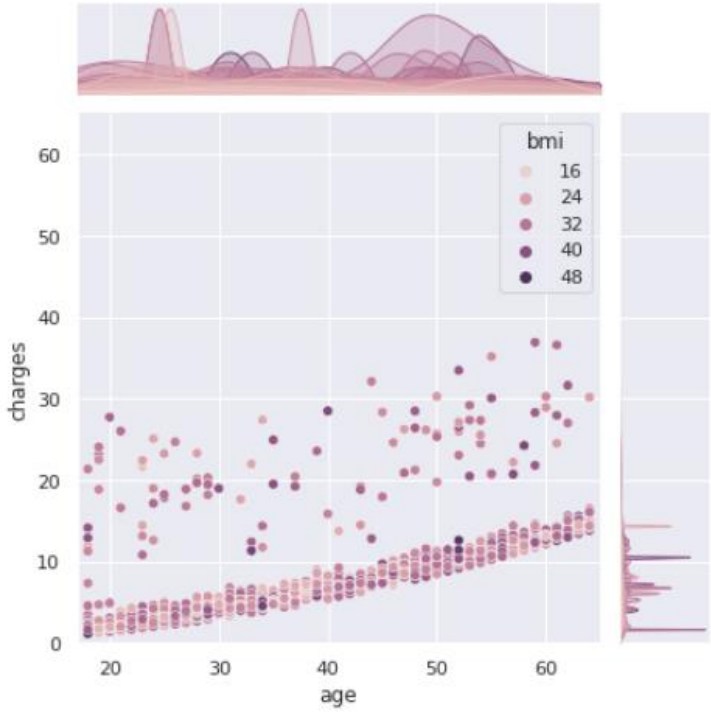
pre processing - EDA

- Explor two variables with bivariate and univariate graphs

smoker



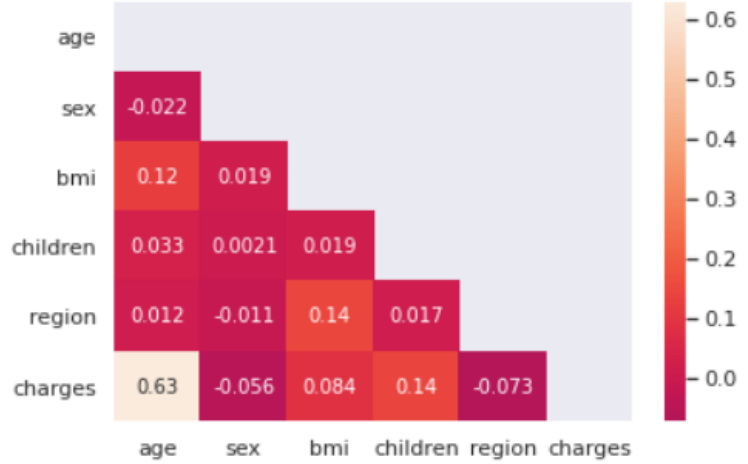
Non-smoker



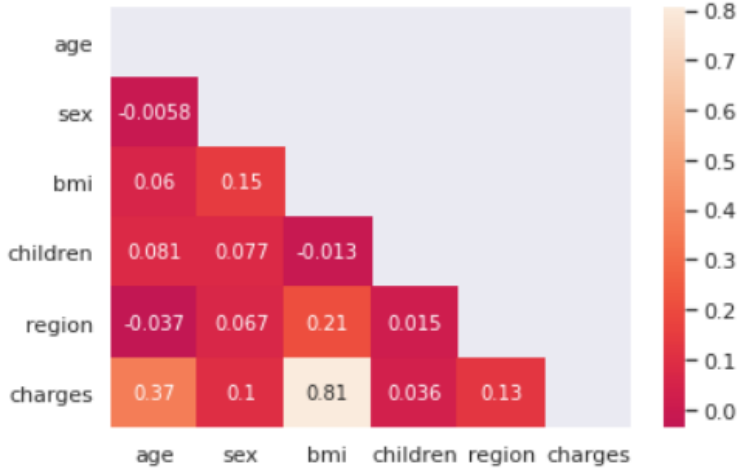
pre processing - EDA

➢Compute pairwise correlation of columns - matrix

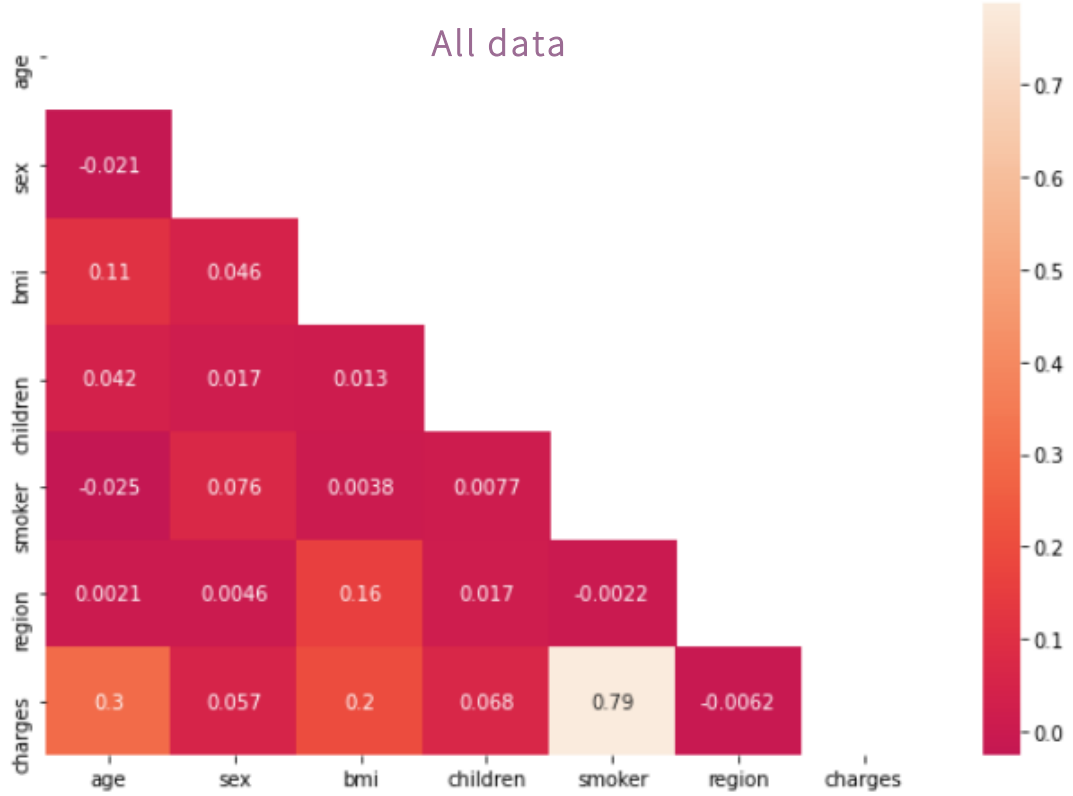
Non-smoker



smoker



All data



Conclusion and the next steps

- Smoking has the highest impact on medical costs, even though the costs are growing with age, bmi and children.
- Also people who have children generally smoke less

We would like to separate the data

to smoker and non-smoker

Examining statistical models

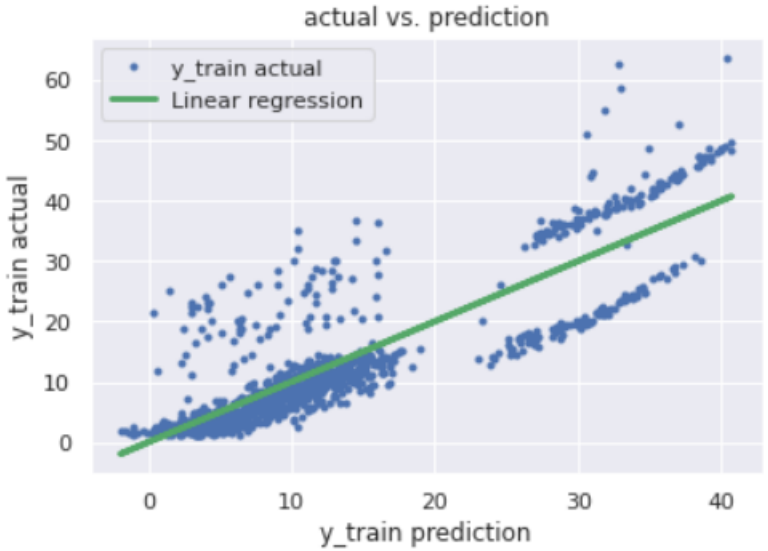
- Linear regression
- Tree regression
- KNN

measure of success

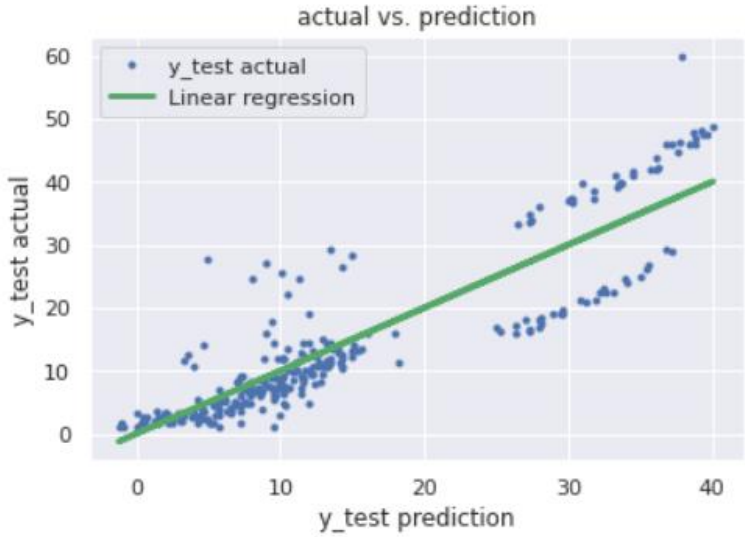
RMSE

Examining statistical models

- Linear regression – All Data



LR ALL Data RMSE (train) = 6.11

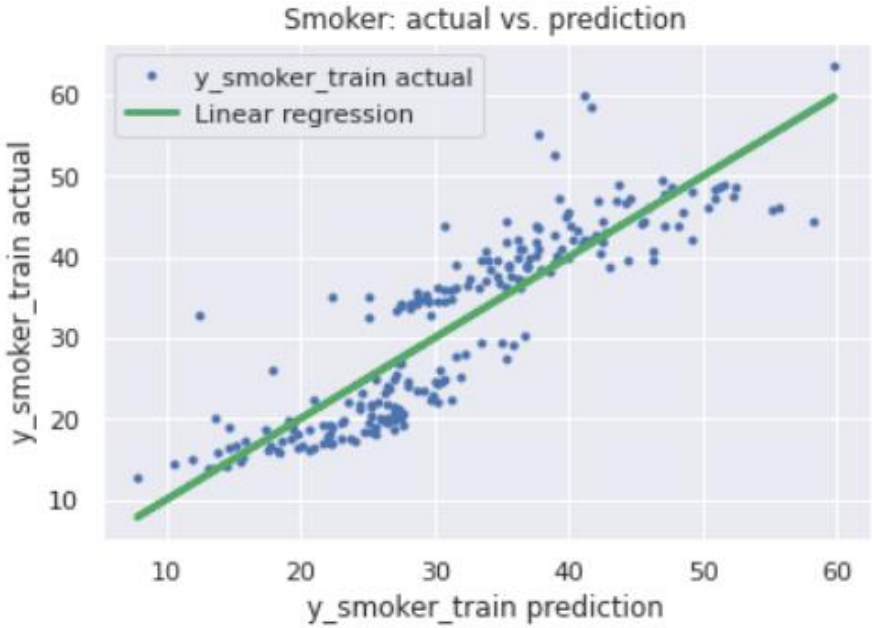


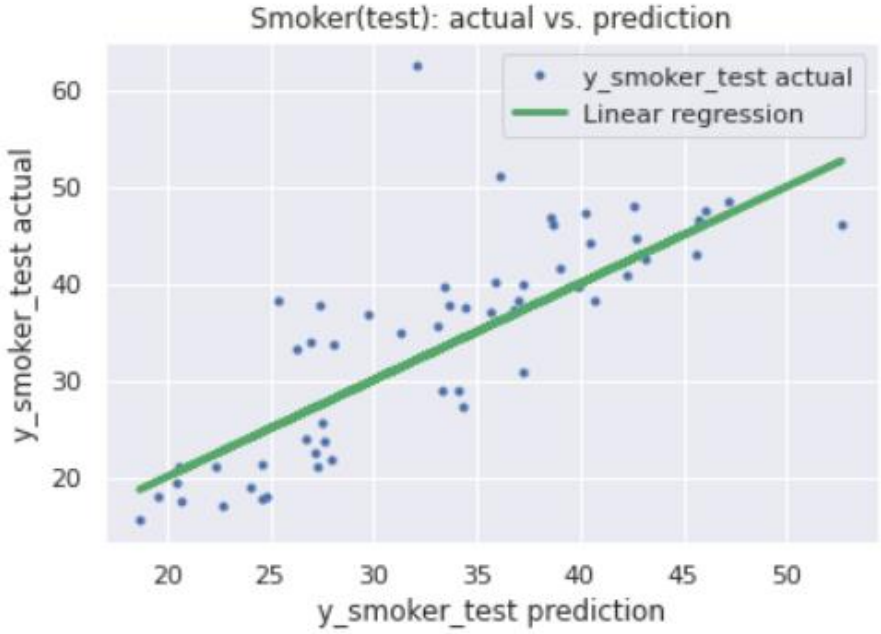
LR ALL Data RMSE (test) = 5.81

```
charges = -11.067+0.251*age+0.104*sex+0.316*bmi+0.507*children+23.783*smoker-0.418*region
```

Examining statistical models

- Linear regression – Smoker





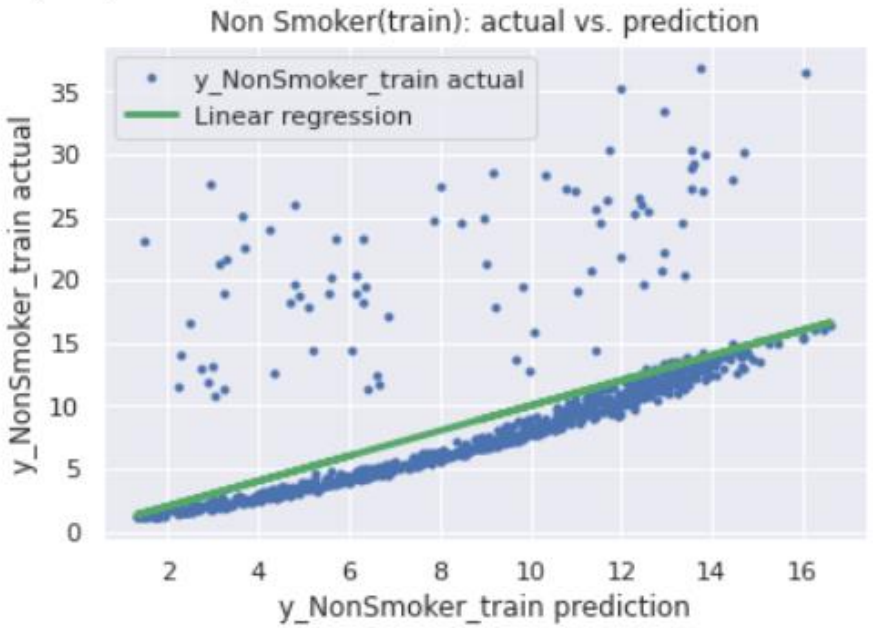
Smokers only RMSE (train) = 5.46

Smokers only RMSE (test) = 6.69

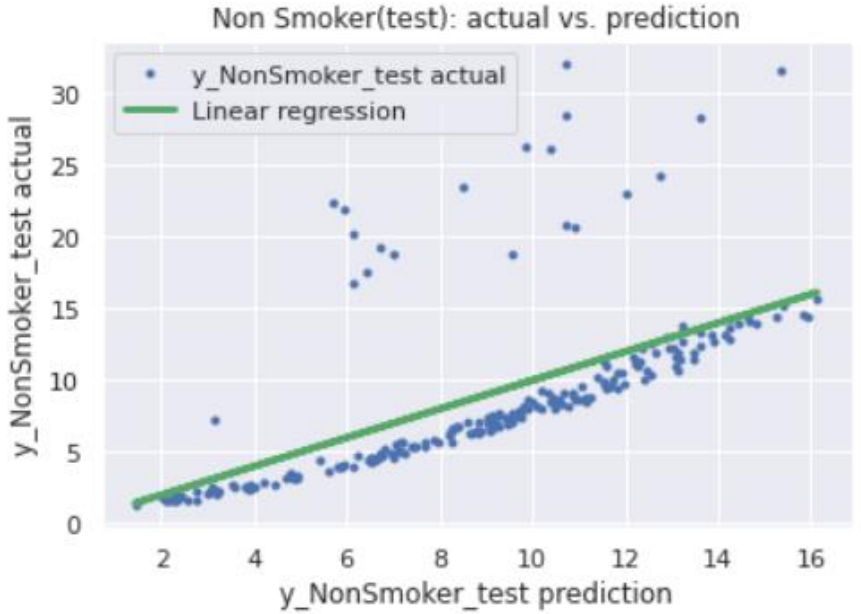
charges = -21.397+0.261*age-0.852*sex+1.439*bmi+0.243*children-0.534*region

Examining statistical models

- Linear regression – Non-smoker



Non Smokers only RMSE (train) = 4.59

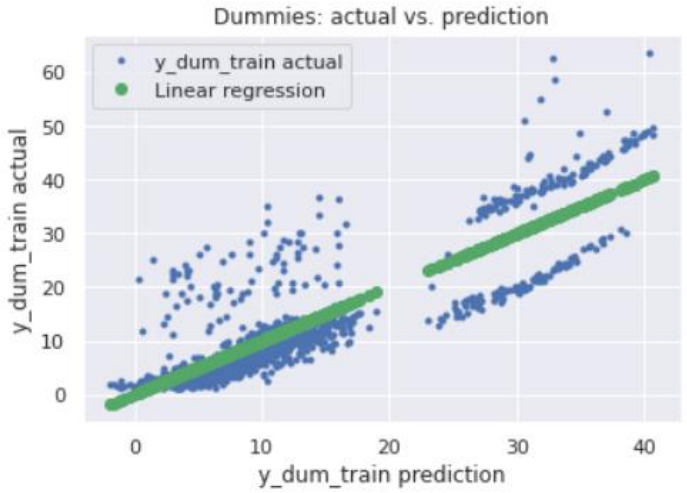


Non Smokers only RMSE (test) = 4.53

charges = -1.921+0.260*age-0.513*sex+0.012*bmi+0.660*children-0.477*region

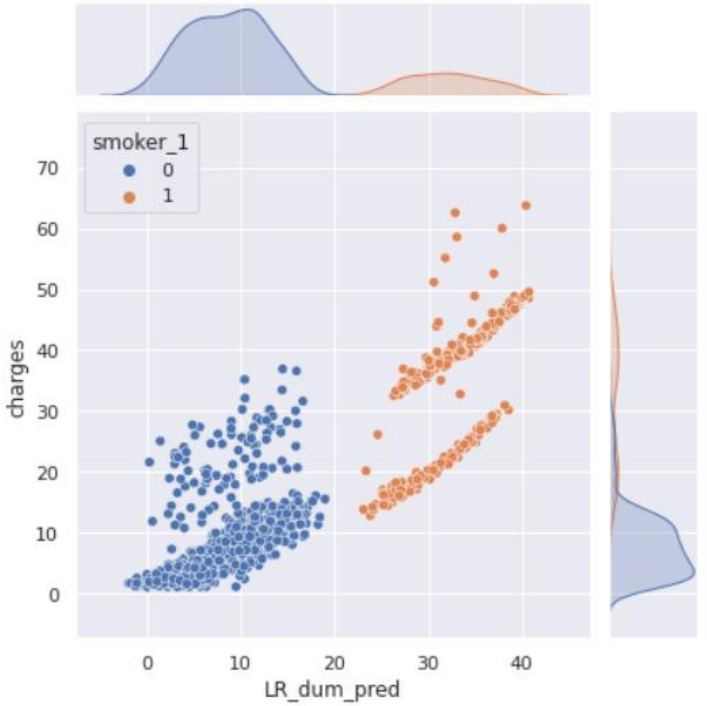
Examining statistical models

- Linear regression – dummies to smoker and Non-smoker



LR Dummies RMSE (train) = 6.11

LR Dummies RMSE (test) = 5.81

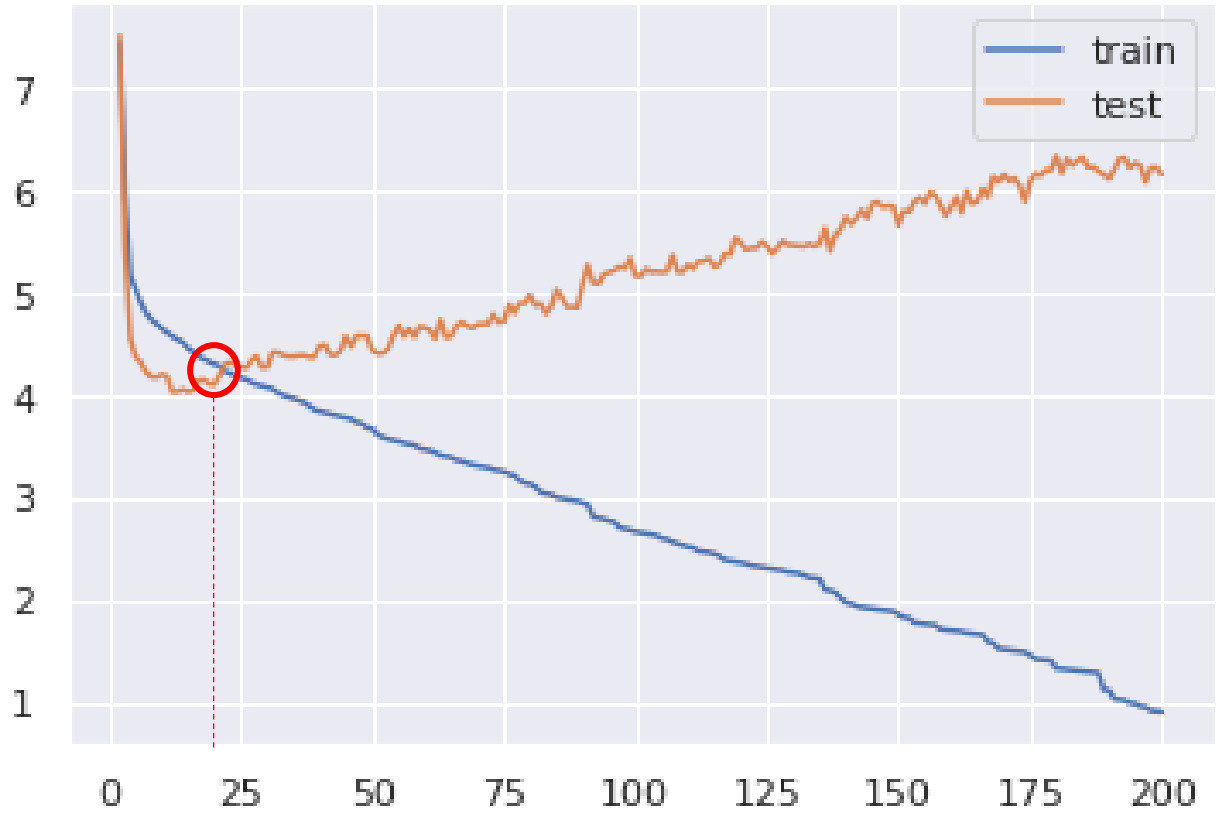


charges = 0.825+0.251*age+0.104*sex+0.316*bmi+0.507*children-0.418*region-11.891*smoker_0+11.891*smoker_1

Examining statistical models

- Tree regression:

22 leaf



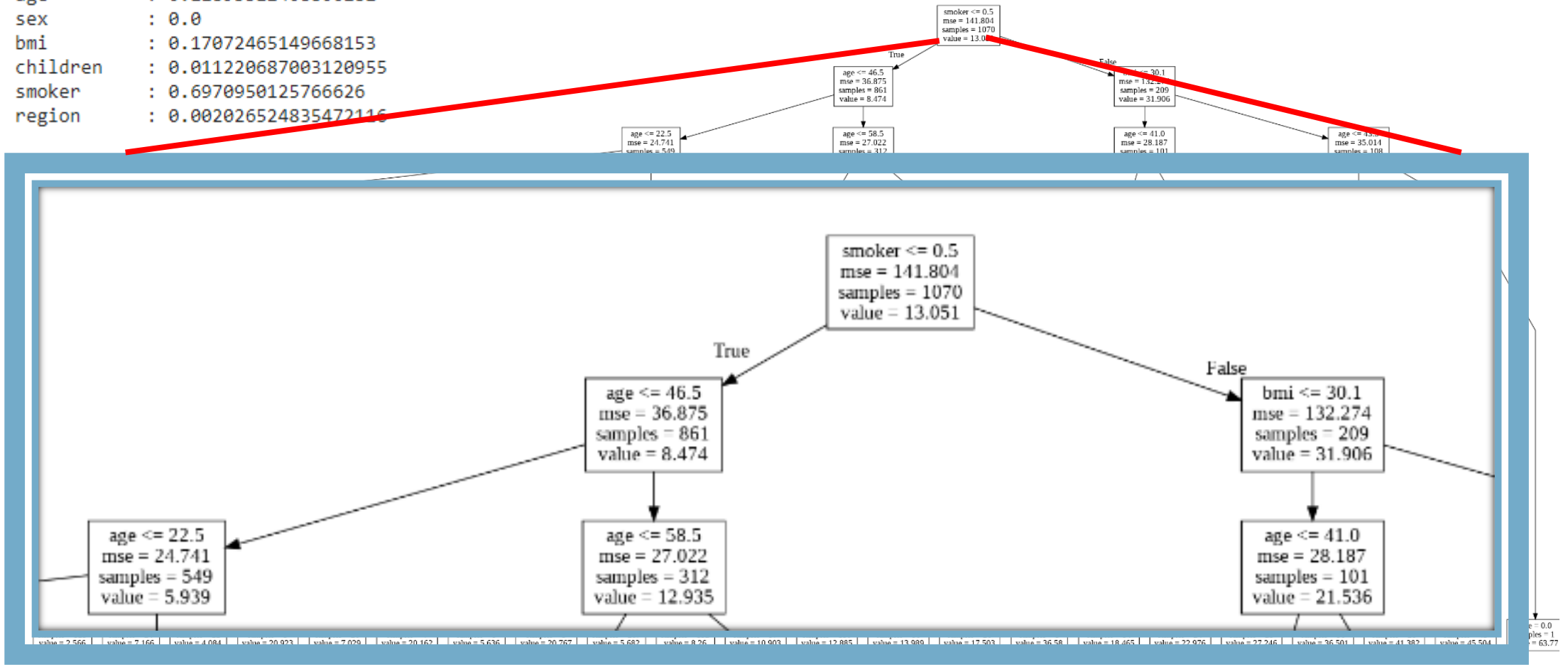
Examining statistical models

- Tree regression

```
age        : 0.11893312408806282
sex        : 0.0
bmi        : 0.17072465149668153
children   : 0.011220687003120955
smoker     : 0.6970950125766626
region     : 0.002026524835472116
```

RMSE (Tree-train)= 4.24

RMSE (Tree- test)= 4.31



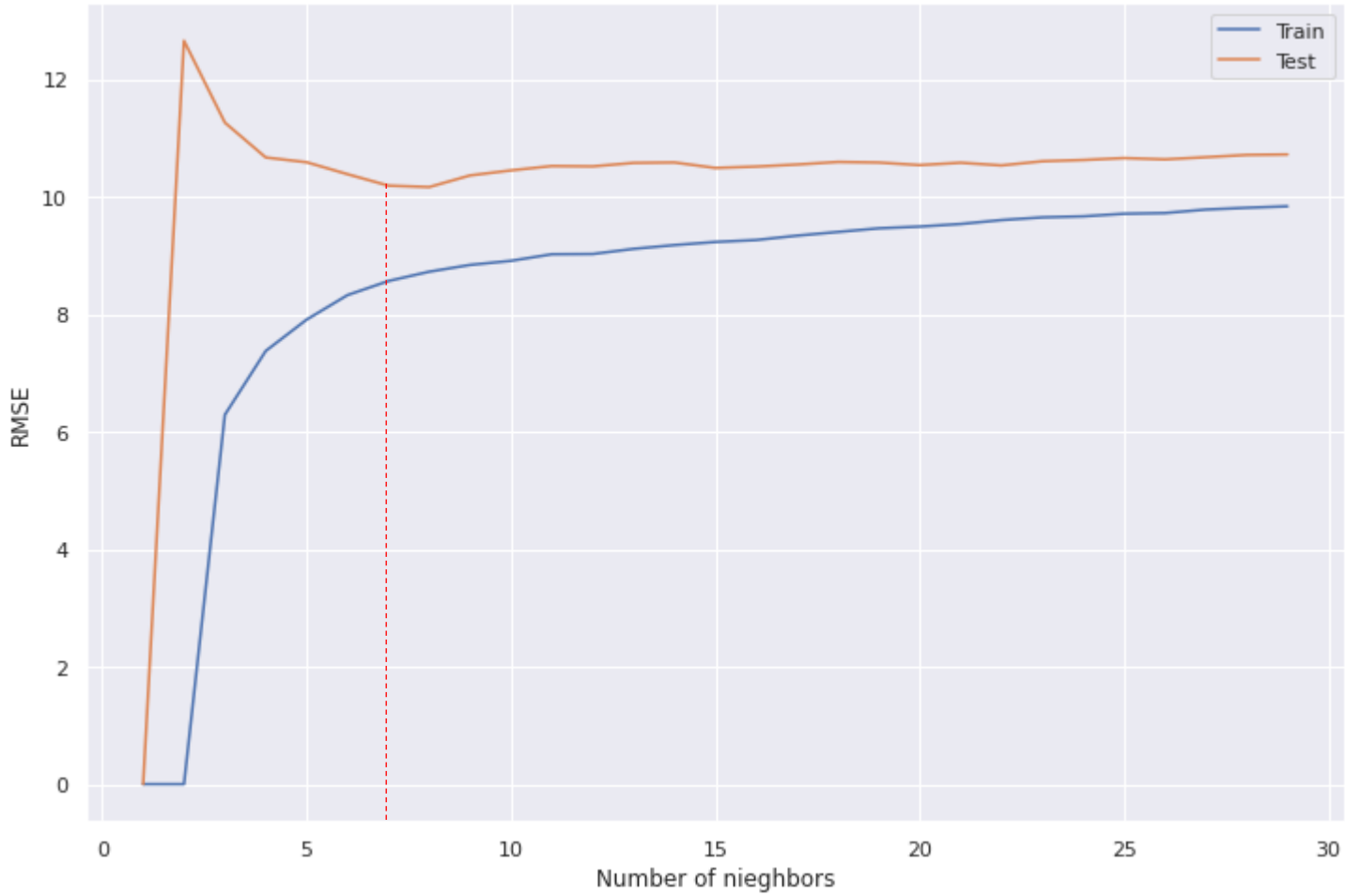
Examining statistical models

- KNN:

7 neighbors

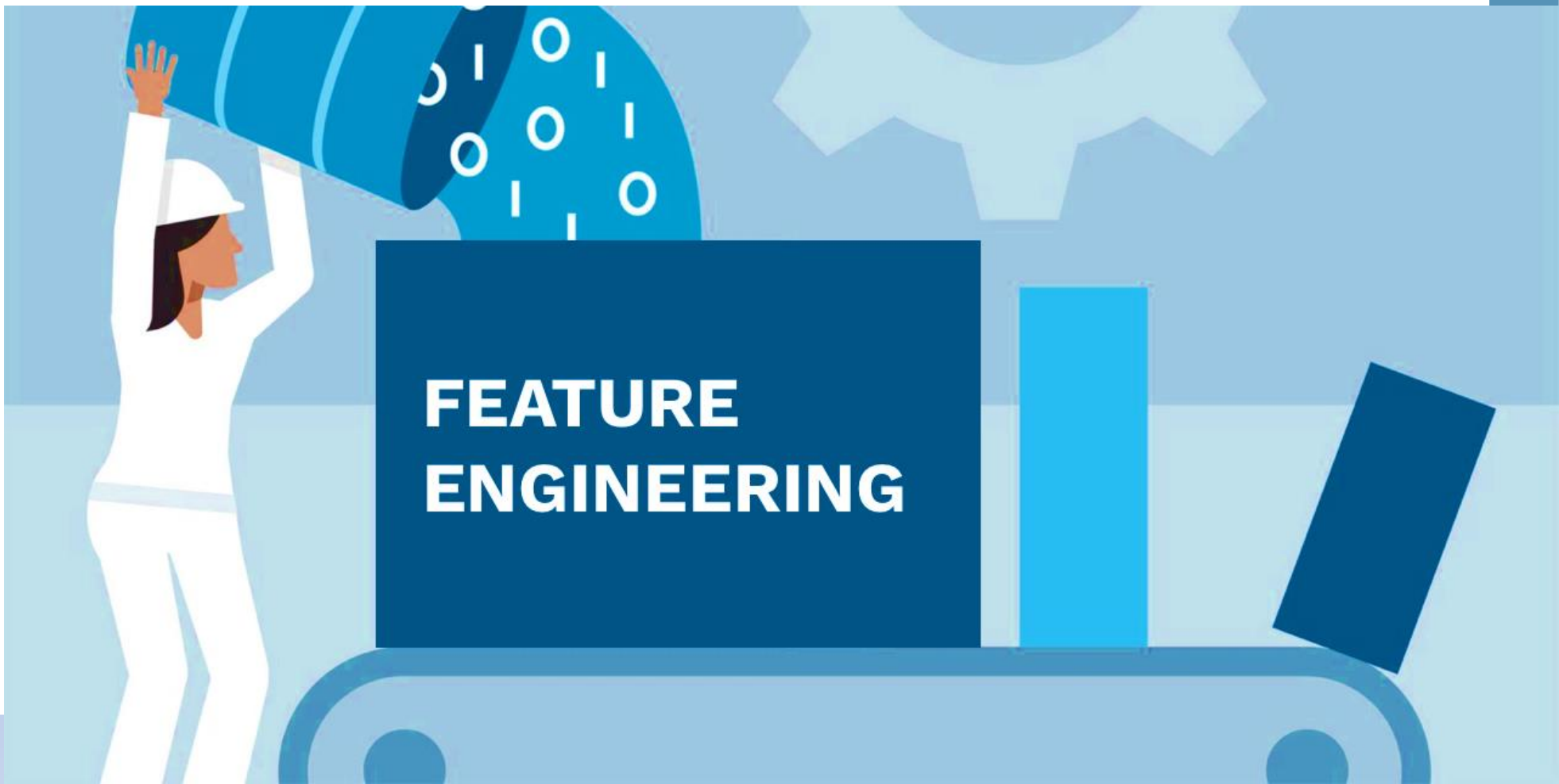
```
KNN RMSE (train)= 8.72

KNN RMSE (test)= 10.16
```



Feature engineering

- Label Encoder
- Dummies
- Standard scalar
- Box Cox
- Normalization



Examining statistical models

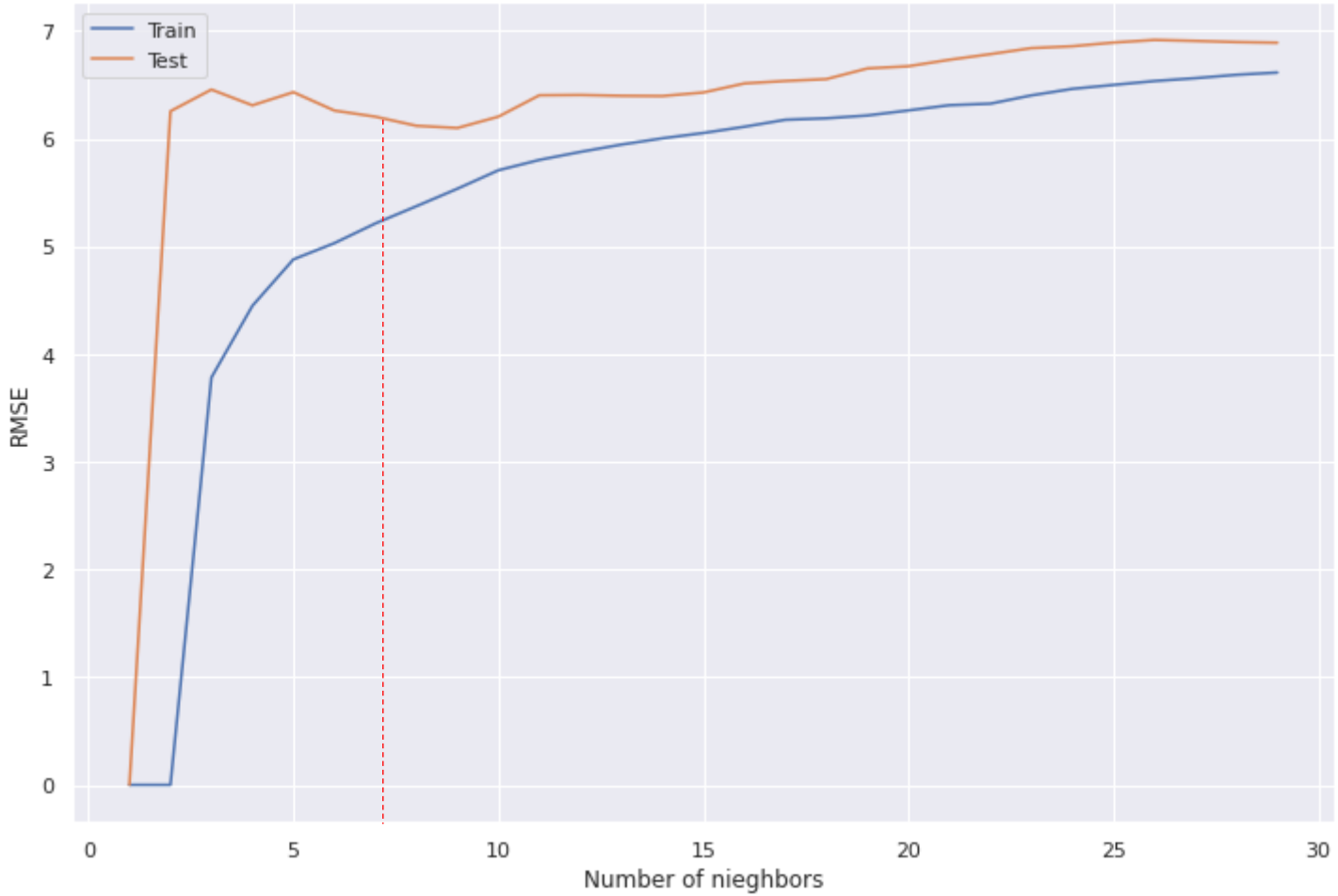
➢KNN:

7 neighbors

Standard Scalar

```
Knn Standard Scalar RMSE (train)= 5.37

Knn Standard Scalar RMSE (test)= 6.12
```



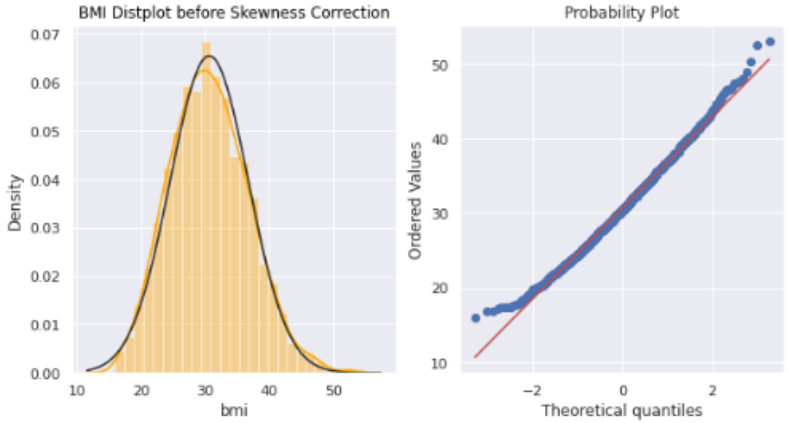
Examining statistical models

- KNN:

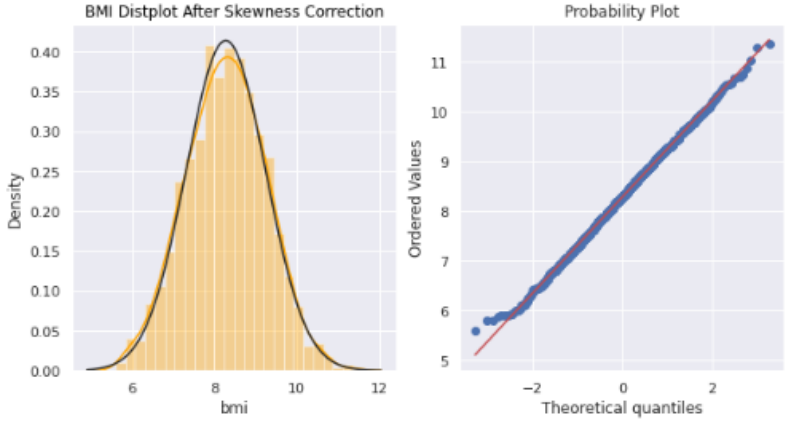
7 neighbors

Box Cox

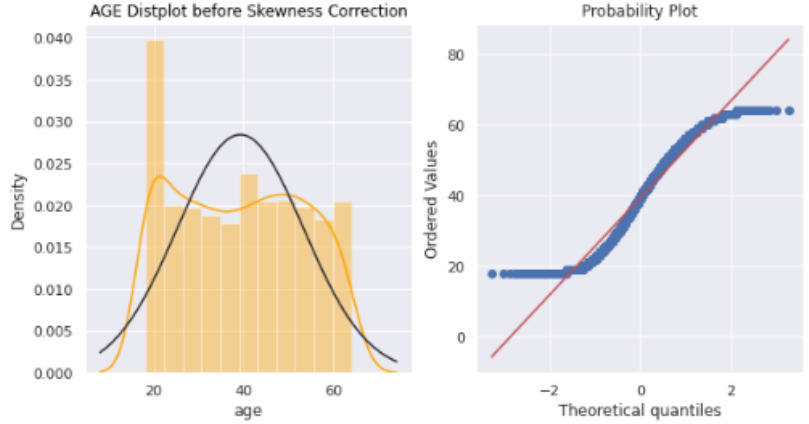
```
Before Correcting
Mu before correcting BMI : 30.66339686098655, Sigma before correcting BMI : 6.0959076415894256
```




```
After Correcting
Mu after correcting BMI : 30.66339686098655, Sigma after correcting BMI : 6.0959076415894256
```




```
Before Correcting
Mu before correcting AGE : 39.20702541106129, Sigma before correcting AGE : 14.044709038954522
```




```
After Correcting
Mu after correcting AGE : 39.20702541106129, Sigma after correcting AGE : 14.044709038954522
```

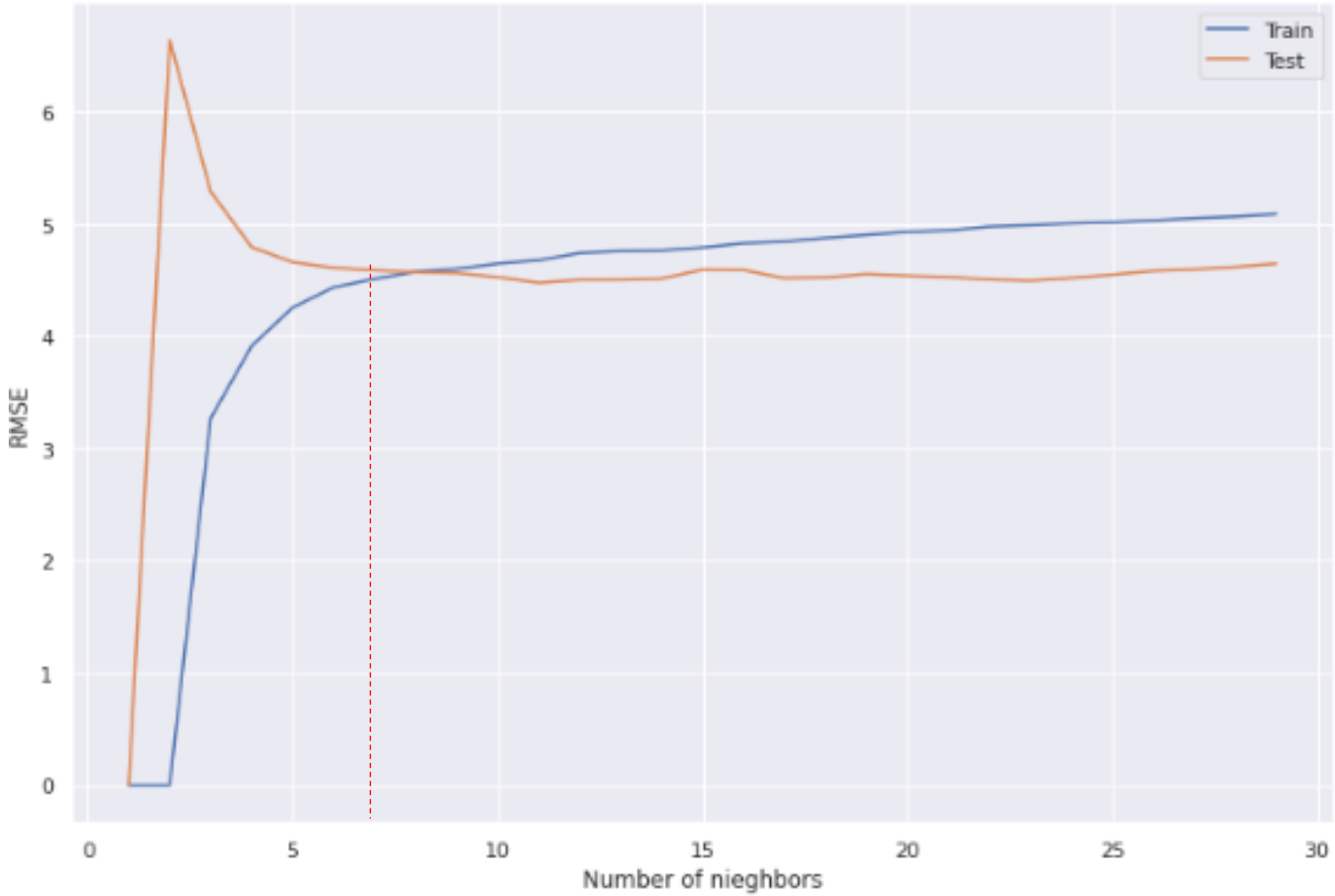
Examining statistical models

- KNN:

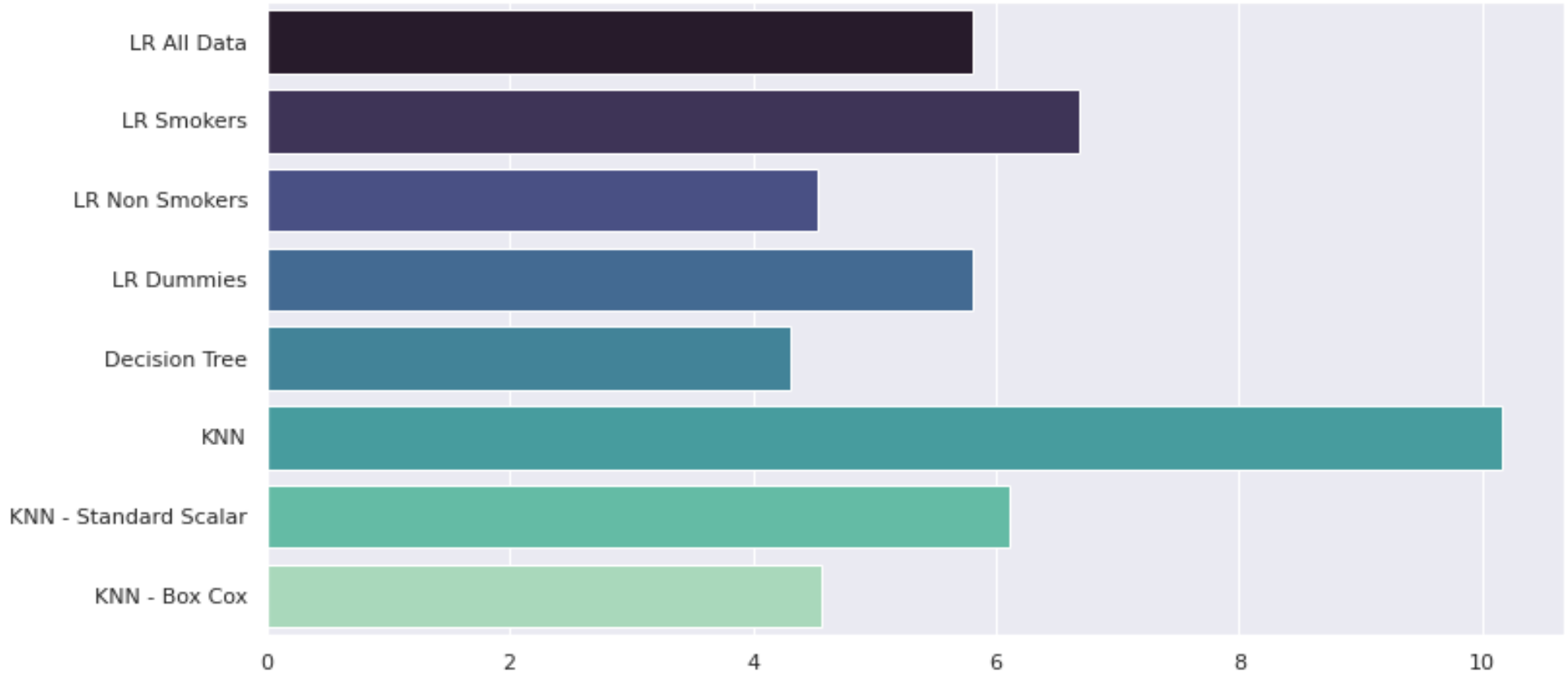
7 neighbors

Box Cox

```
RMSE (knn-train)= 4.58
RMSE (Tree- test)= 4.57
```



Comparing Methods



Thanks

Tomer badug

Shirli miller

Judi Eliya