```
import pandas as pd
from google.colab import drive
import matplotlib.pyplot as plt
import sklearn as sk
from sklearn.linear_model import LinearRegression
import numpy as np
```

```
drive.mount('/content/gdrive/', force_remount = True)
```

```
Mounted at /content/gdrive/
```

```
df = pd.read_csv('/content/gdrive/MyDrive/442B//insurance.csv')
df
```

	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

```
df.loc[df["sex"]=="female", "sex"] = 1
df.loc[df["sex"]=="male", "sex"] = 0
df.loc[df["smoker"]=="yes", "smoker"] = 1
df.loc[df["smoker"]=="no", "smoker"] = 0
df.loc[df["region"]=="southwest", "region"] = 0
df.loc[df["region"]=="southeast", "region"] = 1
df.loc[df["region"]=="northwest", "region"] = 2
df.loc[df["region"]=="northeast", "region"] = 3
df.head(5)
```

	age	sex	bmi	children	smoker	region	charges
0	19	1	27.900	0	1	0	16884.92400
1	18	0	33.770	1	0	1	1725.55230
2	28	0	33.000	3	0	1	4449.46200
3	33	0	22.705	0	0	2	21984.47061
4	32	0	28.880	0	0	2	3866.85520

```
#Split the data
```

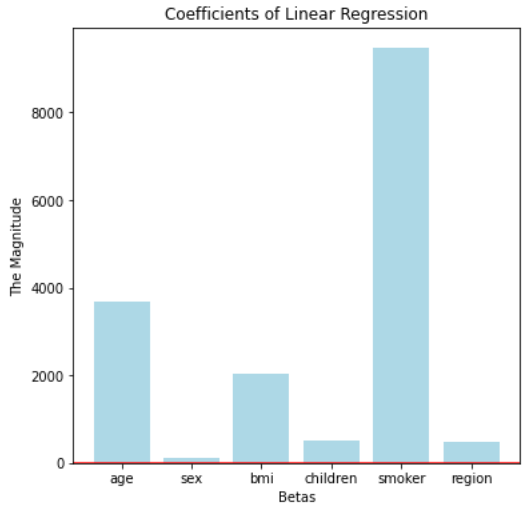
```
data = np.array(df.iloc[:,:-1])
target=np.array(df.iloc[:,-1])
```

```
cut = int((len(data) *.8) // 1)
```

```
in_data=data[:cut]
out_data=data[cut:]

in_target=target[:cut]
out_target=target[cut:]
```

```
#Normalize the data


from sklearn import preprocessing
scaler = preprocessing.StandardScaler().fit(in_data)
in_data_scale = scaler.transform(in_data)

out_data_scale = scaler.transform(out_data)


#Get lambda from lasso cross validation


from sklearn.linear_model import LassoCV
modCV = LassoCV().fit(in_data_scale,in_target)
a=modCV.alpha_


#run a regression with lambda


from sklearn.linear_model import Lasso
mod1 = Lasso(alpha = 1.0).fit(in_data_scale,in_target)


#Visualize


import matplotlib.pyplot as plt
df1 = pd.DataFrame(zip(df.columns[:-1], mod1.coef_))

plt.figure(figsize=(6,6))
plt.bar(df1[0],df1[1],color = 'lightblue')
plt.axhline(0,color='red')
plt.xlabel('Betas')
plt.ylabel('The Magnitude')
plt.title('Coefficients of Linear Regression')
```

```
Text(0.5, 1.0, 'Coefficients of Linear Regression')
```




```
#Compare MSE


from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression

pred_train = mod1.predict(in_data)
print((mean_squared_error(in_target,pred_train)))
```

```
47196178631.99201
```

```
pred_test = mod1.predict(out_data)
print((mean_squared_error(out_target,pred_test)))
```

```
43880001331.24989
```

```
#from here, MSE for out sampling is lower and MSE for in sampling is higher
```

✓ 0s completed at 1:12 AM