Here we will nlearen to find outliers using z\_score & INNER QUARTILE RANGE(iqr)

Def cal\_z(df, col):

Mean = df[col].mean()

Std = df[col].std()

Return (df[col]-mean)/std

This is the function to calculate the z value of a column, now we can just call this cal\_z with any column as shown below,

(housing

.pipe(cal\_z, col=’SalePrice’)

)

So now we will see how to also filterout the Z values >=3. 3 is used in normal distribution, anything above 3 is outlayer in normal distribution

(housing

.assign(Z\_score = cal\_z(housing, col=’SalesPrice’))

.query(‘Z\_score.abs() >=3’)

)

Here to check for -3, we will njot use abs

.query(‘Z\_score<= -3’)

Another way to calculate outlayers is inner quartile range, and values outside are calsified as outliners

Def cal\_iqr\_outlier(df, col)

Ser = df[col]

Iqr = ser.quantile(.75) – ser.quantile(.25)

Med = ser.median()

Small\_mask = ser < med-iqr\*3

Large\_mask = ser > med+iqr\*3

Return small\_mask | large\_mask

Here ser.quantile(.75) is 75 percentile of the series.

Housing[

Cal\_iqr\_outliner(housing, ‘SalesPrice’)

So this gives out the Boolean results, true for anything satisfyting small or large, false otherwise.

Now we will throw htrese results to the column

Housing(

.assign(Iqr\_outlier = cal\_iqr\_outlier(housing, col= ‘SalesPrice’))

.query(‘iqr\_outlier’) //this filters for only true

)