Prepare Data

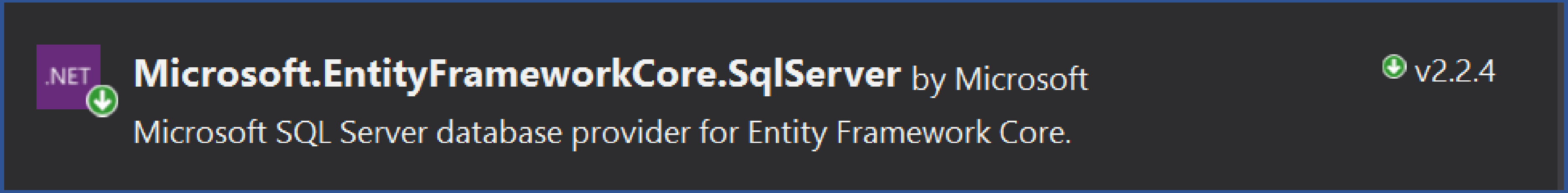
What’s in this session?

- Create data model via Classes
- Filter data
- Replace missing data
- Use normalizer
- Use Binning
- Work with categorical data
- Work with Text
- Work with text pipeline

Create New Project

Create new .NET CORE console app project name = “Prepare”

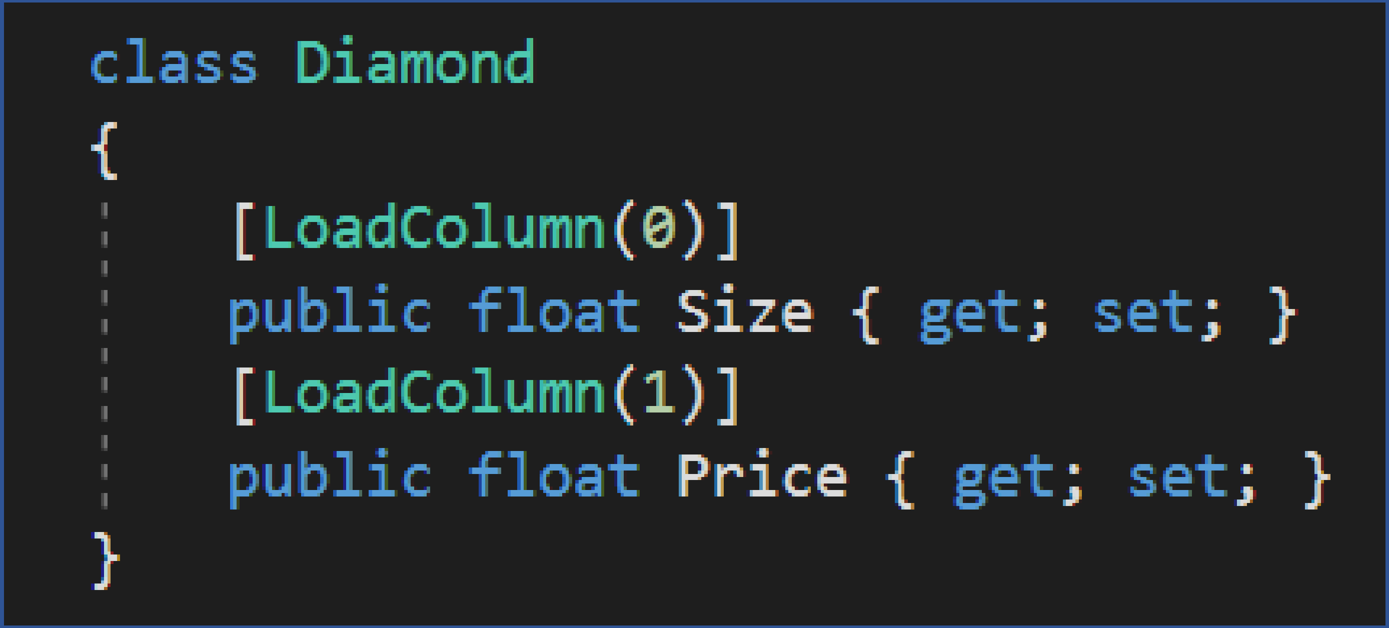
Add NuGet Packages



Microsoft.EntityFrameworkCore.SqlServer by Microsoft

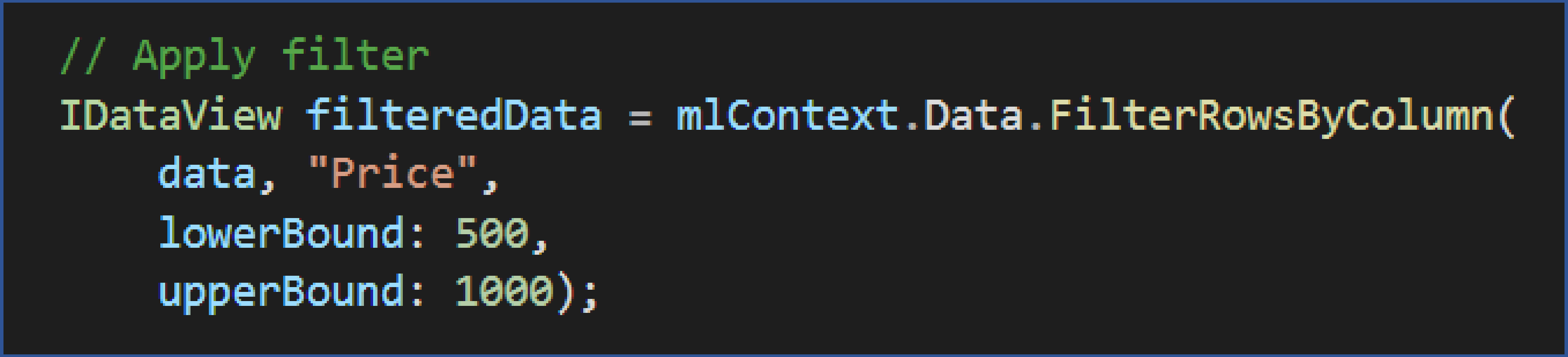
Microsoft SQL Server database provider for Entity Framework Core.

Create data model via Classes

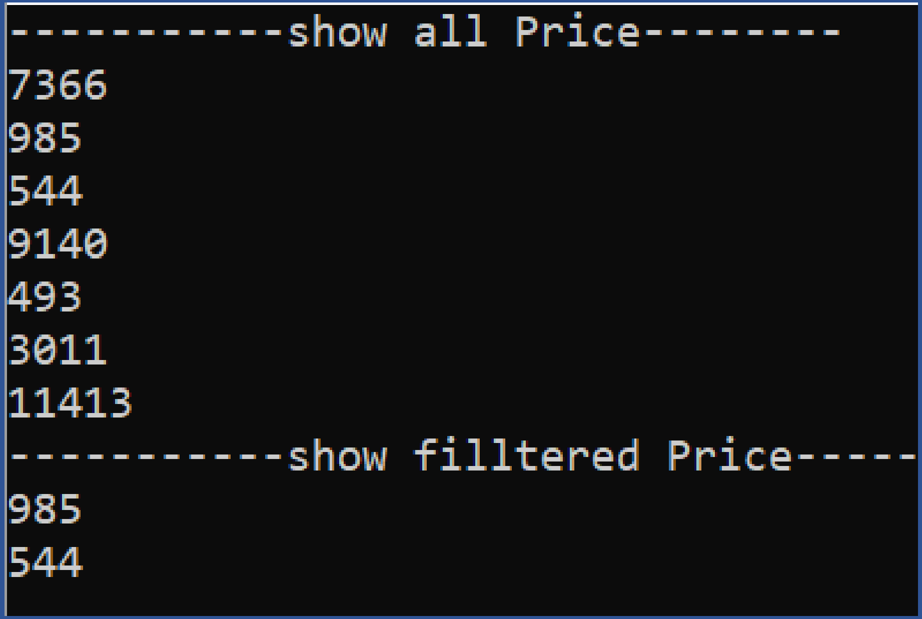


```
class Diamond
{
    [LoadColumn(0)]
    public float Size { get; set; }
    [LoadColumn(1)]
    public float Price { get; set; }
}
```

Filter data

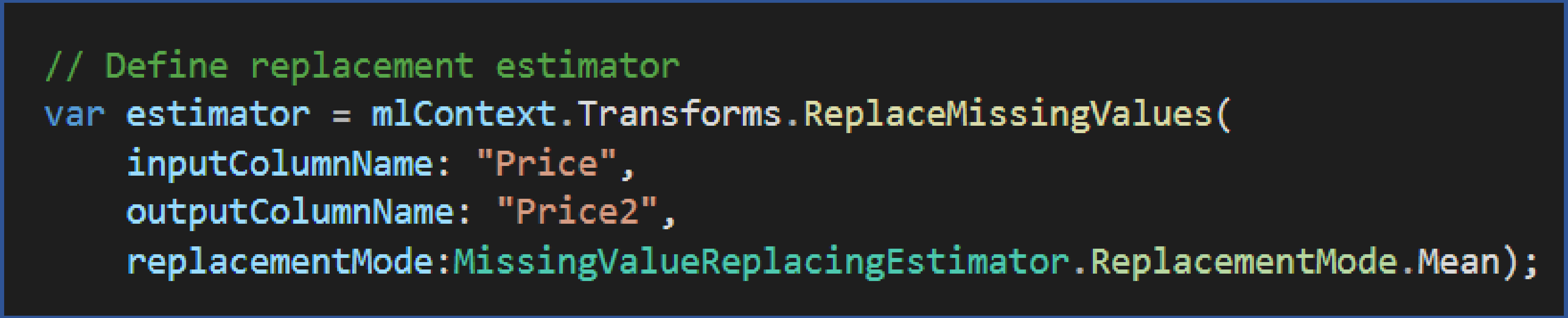


```
// Apply filter
IDataView filteredData = mlContext.Data.FilterRowsByColumn(
    data, "Price",
    lowerBound: 500,
    upperBound: 1000);
```

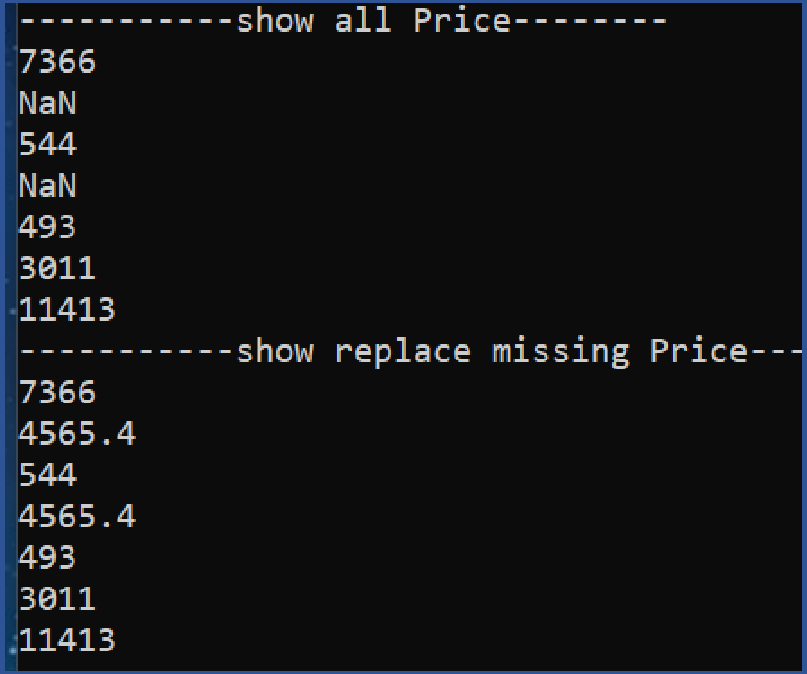


```
-----------show all Price--------
7366
985
544
9140
493
3011
11413
-----------show fillteered Price-----
985
544
```

Replace missing data

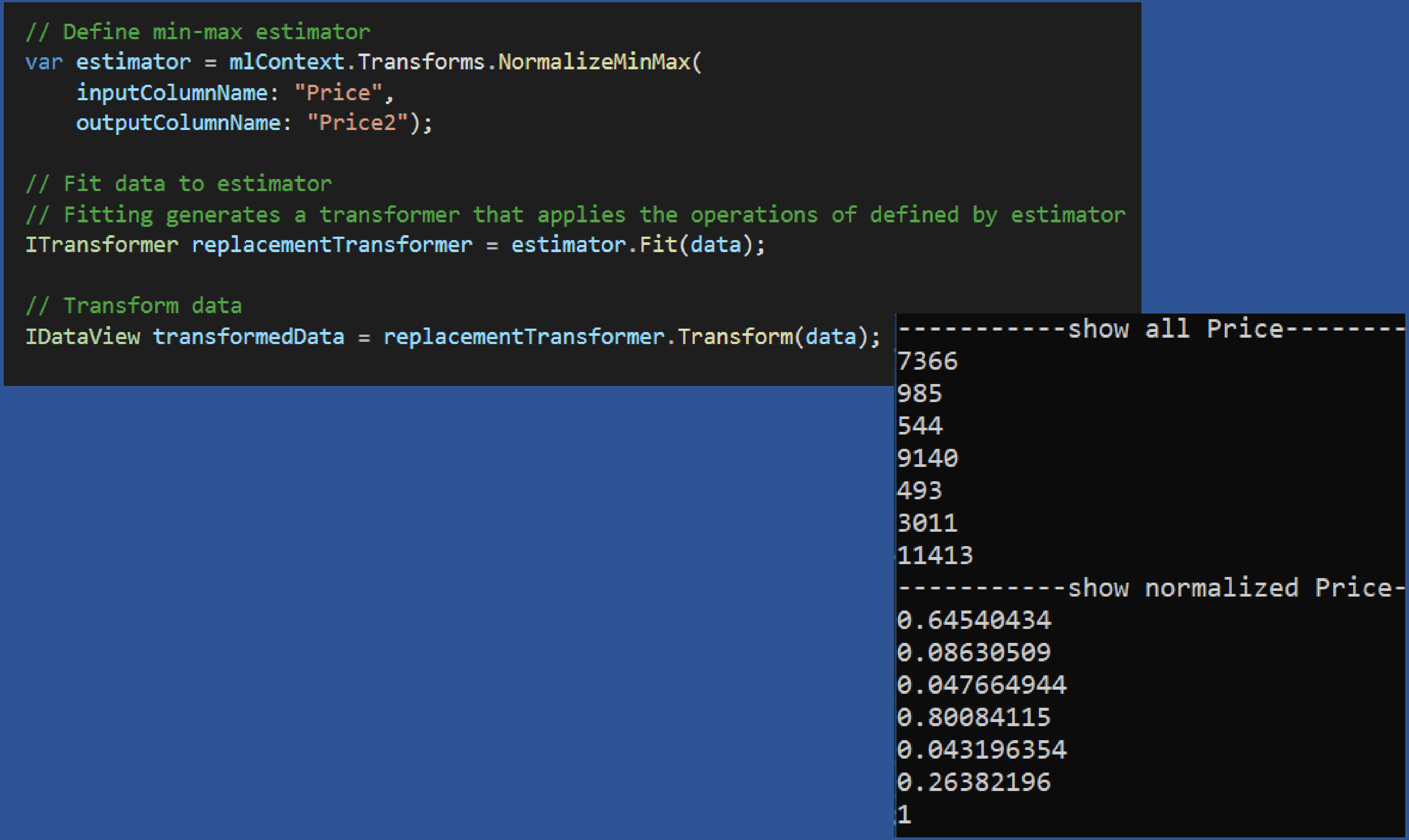


```
// Define replacement estimator
var estimator = mlContext.Transforms.ReplaceMissingValues(
    inputColumnName: "Price",
    outputColumnName: "Price2",
    replacementMode:MissingValueReplacingEstimator.ReplacementMode.Mean);
```



```
-----------show all Price--------
7366
NaN
544
NaN
493
3011
11413
-----------show replace missing Price---
7366
4565.4
544
4565.4
493
3011
11413
```

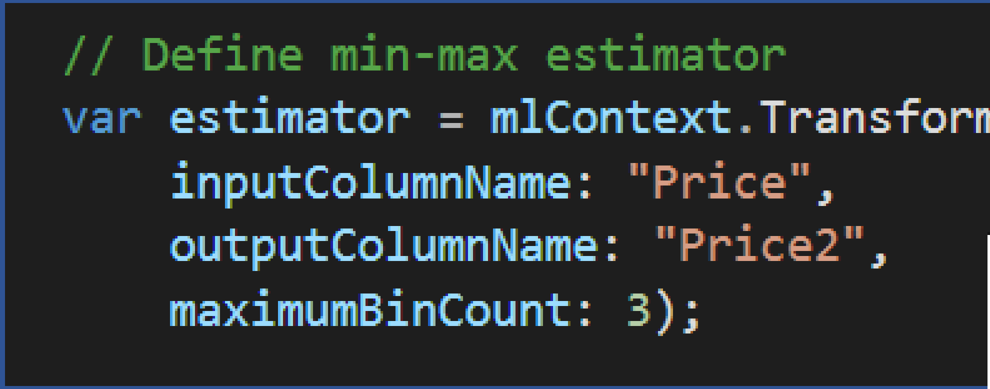
Use normalizer



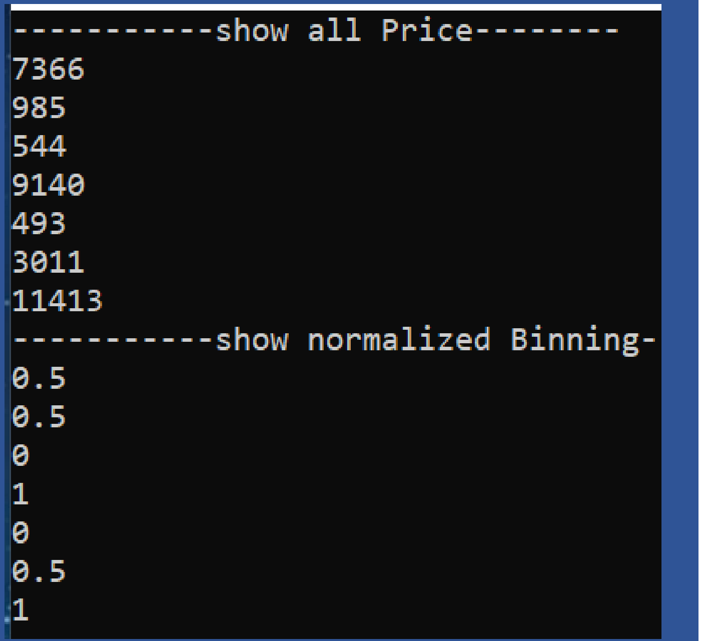
```
// Define min-max estimator
var estimator = mlContext.Transforms.NormalizeMinMax(
    inputColumnName: "Price",
    outputColumnName: "Price2");

// Fit data to estimator
// Fitting generates a transformer that applies the operations of defined by estimator
ITransformer replacementTransformer = estimator.Fit(data);

// Transform data
IDataView transformedData = replacementTransformer.Transform(data);
```

```
-----------show all Price--------
7366
985
544
9140
493
3011
11413
-----------show normalized Price-
0.64540434
0.08630509
0.047664944
0.80084115
0.043196354
0.26382196
1
```

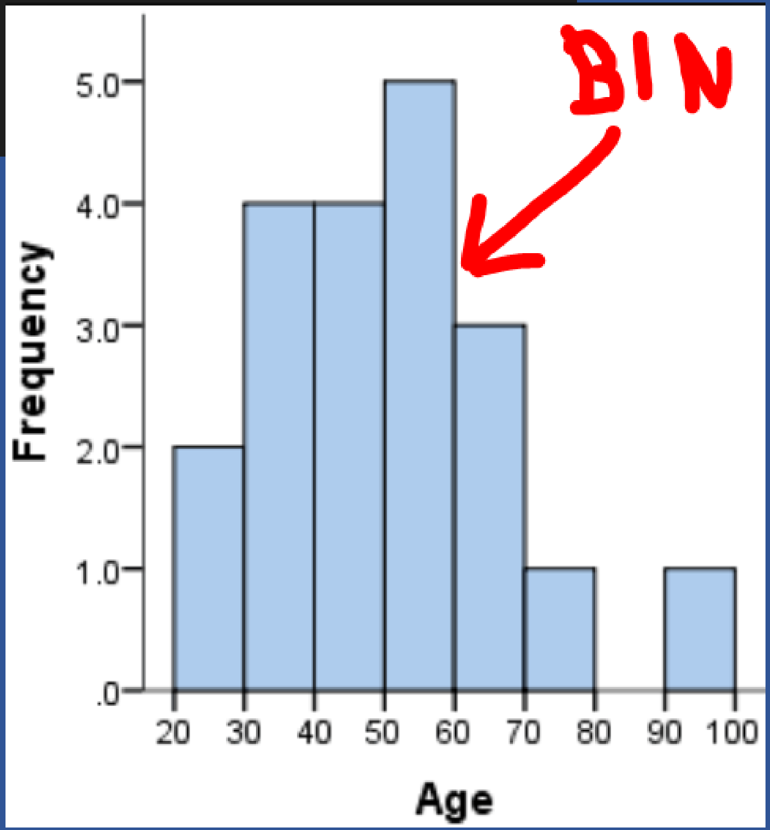
Use Binning



```
// Define min-max estimator
var estimator = mlContext.Transforms.NormalizeBinning(
    inputColumnName: "Price",
    outputColumnName: "Price2",
    maximumBinCount: 3);
```

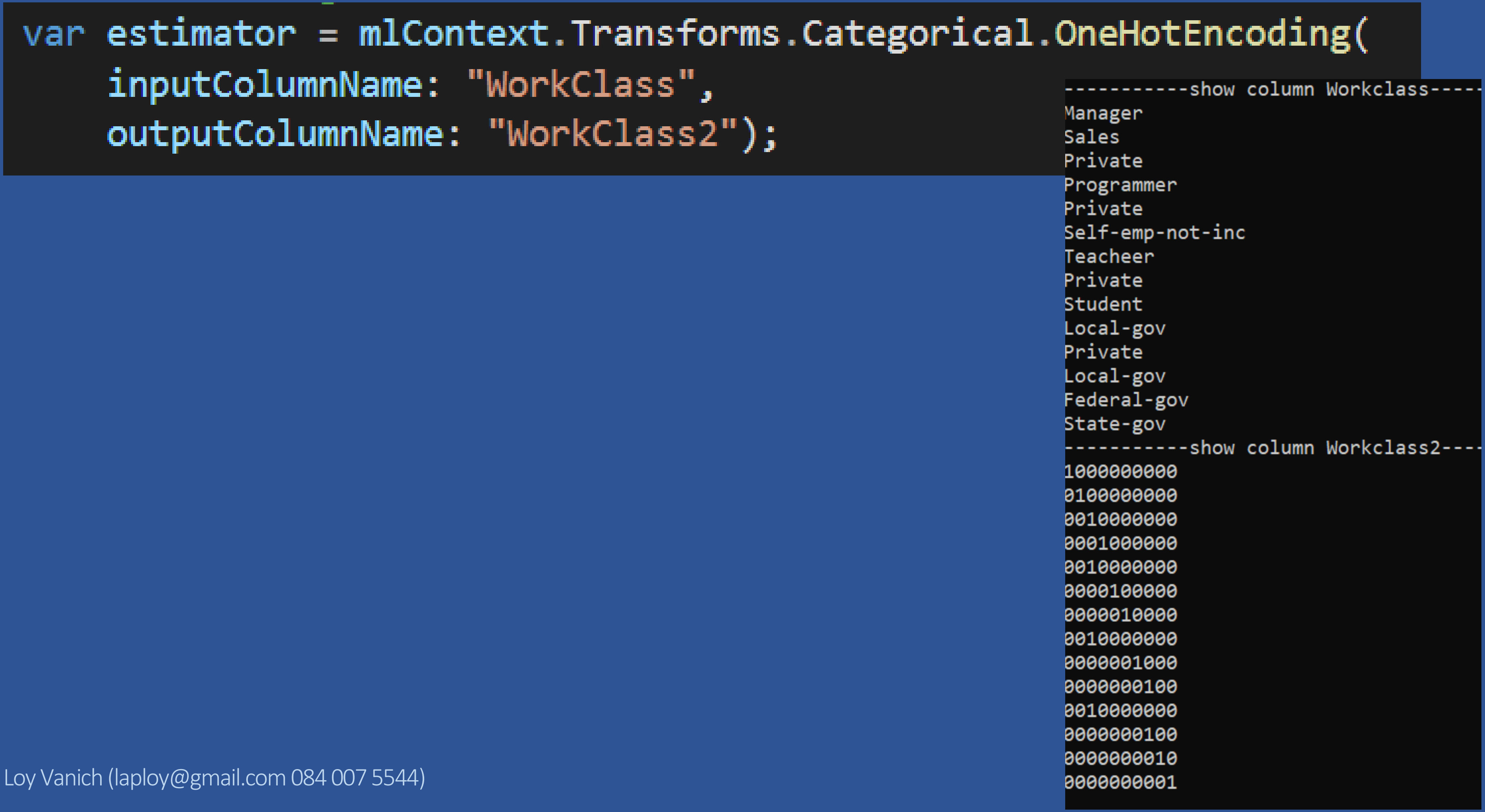


```
-----------show all Price--------
7366
985
544
9140
493
3011
11413
-----------show normalized Binning-
0.5
0.5
0
1
0
0.5
1
```



Work with categorical data

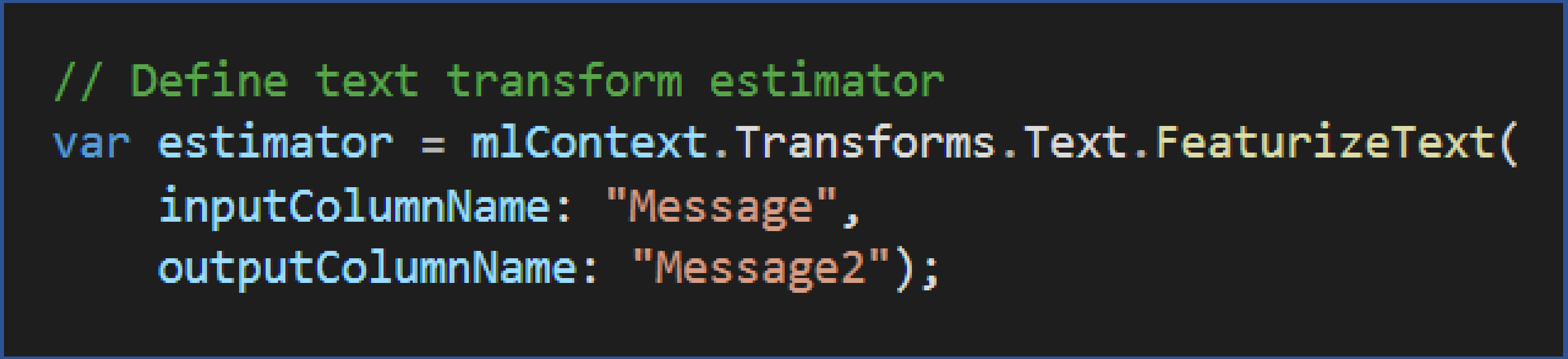
One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.



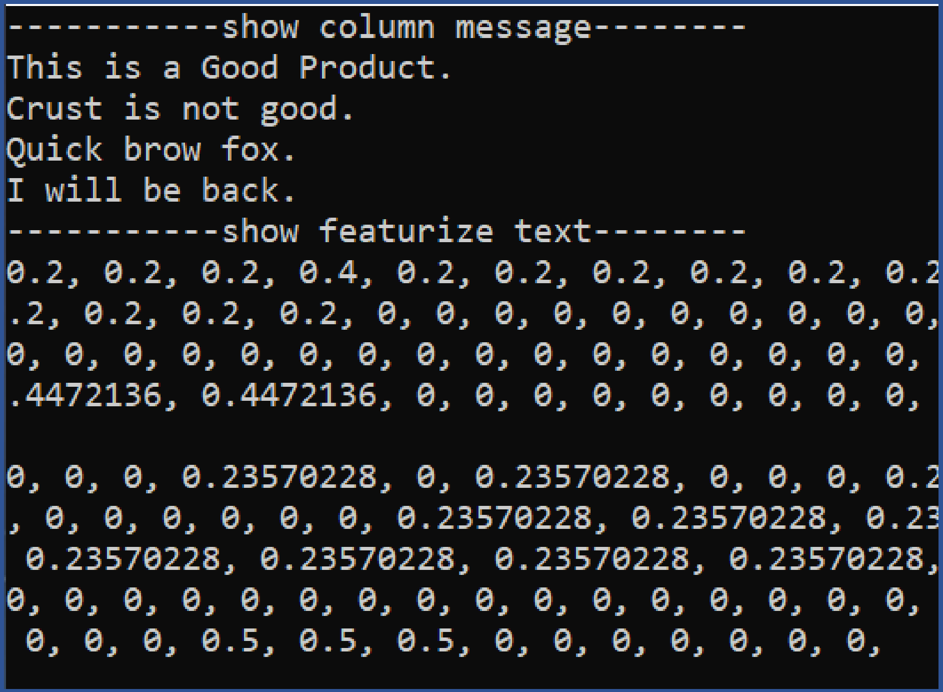
```
var estimator = mlContext.Transforms.Categorical.OneHotEncoding(
    inputColumnName: "WorkClass",
    outputColumnName: "WorkClass2");
```

```
----------show column Workclass----
Manager
Sales
Private
Programmer
Private
Self-emp-not-inc
Teacheer
Private
Student
Local-gov
Private
Local-gov
Federal-gov
State-gov
----------show column Workclass2---
1000000000
0100000000
0010000000
0001000000
0010000000
0000100000
0000010000
0010000000
0000001000
0000000100
0010000000
0000000100
0000000010
0000000001
```

Work with Text

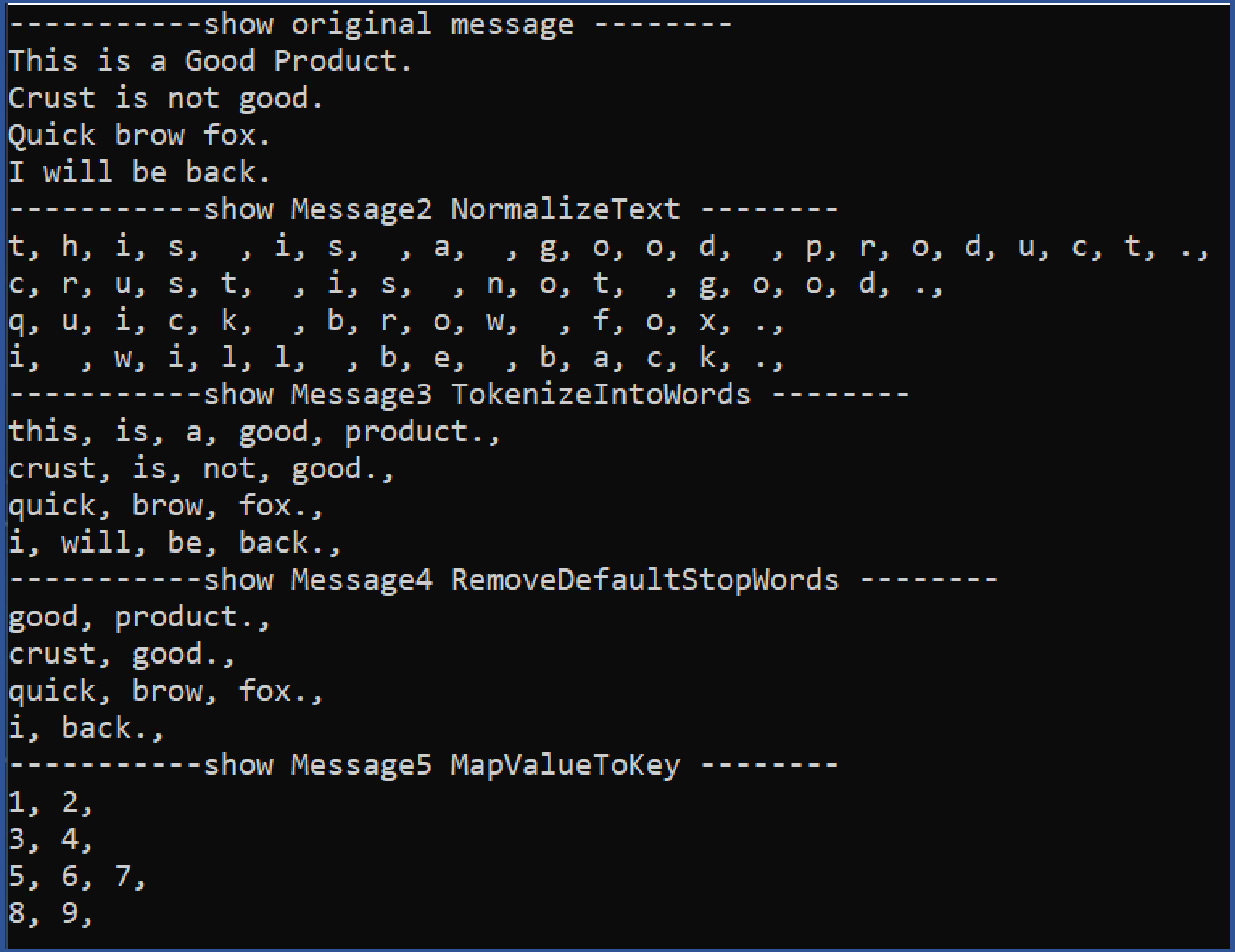


```
// Define text transform estimator
var estimator = mlContext.Transforms.Text.FeaturizeText(
    inputColumnName: "Message",
    outputColumnName: "Message2");
```



```
-----------show column message--------
This is a Good Product.
Crust is not good.
Quick brow fox.
I will be back.
-----------show featurize text--------
0.2, 0.2, 0.2, 0.4, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2
.2, 0.2, 0.2, 0.2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
.4472136, 0.4472136, 0, 0, 0, 0, 0, 0, 0, 0, 0,

0, 0, 0, 0.23570228, 0, 0.23570228, 0, 0, 0, 0.2
, 0, 0, 0, 0, 0, 0, 0.23570228, 0.23570228, 0.23
 0.23570228, 0.23570228, 0.23570228, 0.23570228,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0.5, 0.5, 0.5, 0, 0, 0, 0, 0, 0, 0,
```

Work with text pipeline



```
-----------show original message --------
This is a Good Product.
Crust is not good.
Quick brow fox.
I will be back.
-----------show Message2 NormalizeText --------
t, h, i, s,  , i, s,  , a,  , g, o, o, d,  , p, r, o, d, u, c, t, .,
c, r, u, s, t,  , i, s,  , n, o, t,  , g, o, o, d, .,
q, u, i, c, k,  , b, r, o, w,  , f, o, x, .,
i,  , w, i, l, l,  , b, e,  , b, a, c, k, .,
-----------show Message3 TokenizeIntoWords --------
this, is, a, good, product.,
crust, is, not, good.,
quick, brow, fox.,
i, will, be, back.,
-----------show Message4 RemoveDefaultStopWords --------
good, product.,
crust, good.,
quick, brow, fox.,
i, back.,
-----------show Message5 MapValueToKey --------
1, 2,
3, 4,
5, 6, 7,
8, 9,
```

What’s next?