**Goal is to study customers’ buying patterns so we can answer the following questions:**

1. Do customers in different regions spend more per transaction? Which regions spend the most/least?
2. Are there differences in the age of customers between regions? If so, can we predict the age of a customer in a region based on other demographic data?
3. Can we predict the amount a customer will spend per transaction based on other data we have collected about that customer?
4. Is there any correlation between age of a customer and if the transaction was made online or in the store?
5. Do any other factors predict if a customer will buy online or in our stores?

**Source Data:**

Macintosh HD:Users:thuy:UTA:DA1115.1:AgeRegion.epsBlackwell Electronics demographic data (Demographic\_Data.csv):

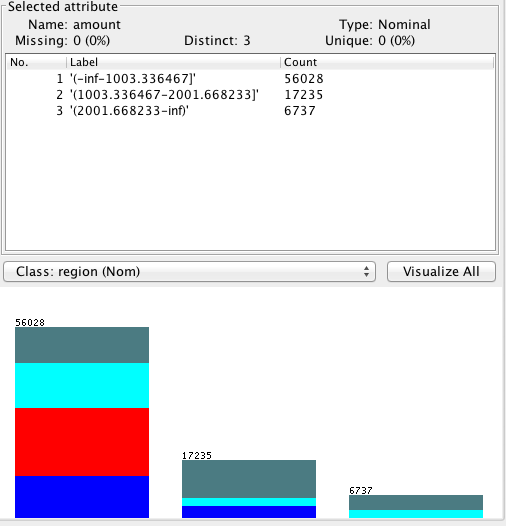
* In-store purchase:
  + in-store = 0; purchased in-store
  + in-store = 1; purchased online
* Customer’s age
  + Range from 18 to 85
* Number of items purchased
* Amount per purchase
  + $5 - $3000
* Region of purchase
  + East (region = 1)

Figure : In-store purchase by Region. Left column shows purchases made in store (in-store = 0). Right column shows purchases made online (in-store = 1)

* + West (region = 2)
  + South (region = 3)
  + Central (region = 4)

**Observations:**

**For amount spent by region**:

****

**Legend:**

Blue: Region 1 (East);

Red: Region 2 (West);

Cyan: Region 3 (South);

Teal: Region4 (Central)

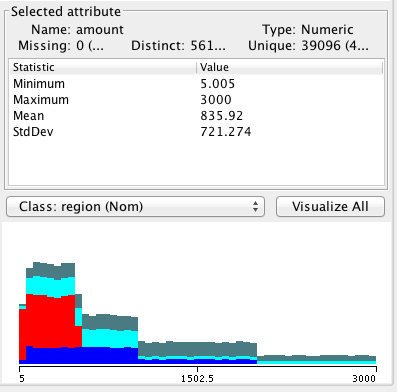
**Note** that entire west region (2) spent less than $1003.34 per transaction.

South and Central regions (3 & 4) spent the most (higher amount per transaction … up to $3000)

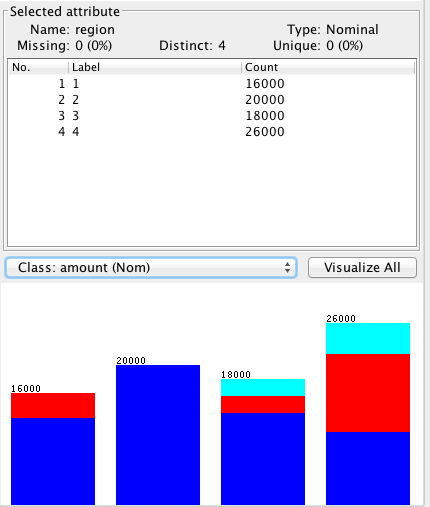
Using XCEL to calculate the average amount spent per transaction for region 2; we get $252.12 /per transaction

This graph shows the amount spent by region before discretizing the amount feature.

This confirms that the West region (2) spent the least and that the Central region (4) also made the most purchase (transactions)



Both South and Central regions have customers spending the most (up to $3000 per purchase)

****

**Amount Spent per Region:**

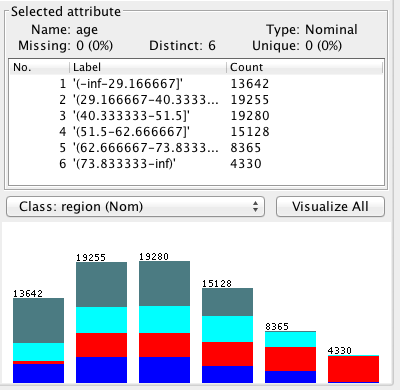
East: 16000 transactions

West: 20000 transactions (100% under $1003)

South: 18000 transactions

Central: 26000 transactions

**Age by Region:**

****For age group 73.8 (74) to 85, most are in West region (2); this region also has the fewest customers in age group 18-28.

**Spending by transaction**

=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0

Relation: Demographic\_Data-weka.filters.unsupervised.attribute.NumericToNominal-R1,5-weka.filters.unsupervised.attribute.Discretize-B6-M-1.0-R2

Instances: 80000

Attributes: 5

in-store

age

items

amount

region

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

M5 pruned model tree:

(using smoothed linear models)

region=1,3,4 <= 0.5 : LM1 (20000/19.778%)

region=1,3,4 > 0.5 :

| in-store=0 <= 0.5 :

| | region=4 <= 0.5 :

| | | region=3,4 <= 0.5 : LM2 (16000/66.728%)

| | | region=3,4 > 0.5 : LM3 (10999/37.952%)

| | region=4 > 0.5 : LM4 (13001/77.69%)

| in-store=0 > 0.5 : LM5 (20000/115.628%)

LM num: 1

amount =

0.5342 \* in-store=0

- 0.051 \* age='(62.666667-73.833333]','(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.0521 \* age='(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

- 3.1614 \* age='(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.0115 \* age='(29.166667-40.333333]','(-inf-29.166667]'

+ 0.0238 \* age='(-inf-29.166667]'

+ 0.8855 \* region=1,3,4

- 0.0745 \* region=3,4

+ 0.1986 \* region=4

+ **252.8705**

LM num: 2

amount =

0.4433 \* in-store=0

- 0.017 \* age='(62.666667-73.833333]','(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 119.2247 \* age='(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 130.0925 \* age='(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 25.9971 \* age='(29.166667-40.333333]','(-inf-29.166667]'

+ 61.3221 \* age='(-inf-29.166667]'

+ 0.2953 \* region=1,3,4

- 0.3638 \* region=3,4

+ 0.4164 \* region=4

**+ 526.2351**

LM num: 3

amount =

0.4433 \* in-store=0

- 0.017 \* age='(62.666667-73.833333]','(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.1883 \* age='(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.1466 \* age='(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.0465 \* age='(29.166667-40.333333]','(-inf-29.166667]'

+ 0.1022 \* age='(-inf-29.166667]'

+ 0.2953 \* region=1,3,4

- 0.4497 \* region=3,4

+ 0.4164 \* region=4

**+ 520.9323**

LM num: 4

amount =

0.4433 \* in-store=0

- 0.017 \* age='(62.666667-73.833333]','(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.1553 \* age='(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.0872 \* age='(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.0294 \* age='(29.166667-40.333333]','(-inf-29.166667]'

+ 0.0439 \* age='(-inf-29.166667]'

+ 0.2953 \* region=1,3,4

- 0.2973 \* region=3,4

+ 0.6908 \* region=4

**+ 1023.4124**

LM num: 5

amount =

0.7083 \* in-store=0

- 0.017 \* age='(62.666667-73.833333]','(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.121 \* age='(51.5-62.666667]','(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 0.0436 \* age='(40.333333-51.5]','(29.166667-40.333333]','(-inf-29.166667]'

+ 19.3898 \* age='(29.166667-40.333333]','(-inf-29.166667]'

+ 0.0268 \* age='(-inf-29.166667]'

+ 0.2953 \* region=1,3,4

- 0.0964 \* region=3,4

+ 0.2564 \* region=4

**+ 1532.0856**

Number of Rules : 5

Time taken to build model: 232.23 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient 0.6699

Mean absolute error 399.9915

Root mean squared error 535.5058

Relative absolute error 68.873 %

Root relative squared error 74.2446 %

Total Number of Instances 80000

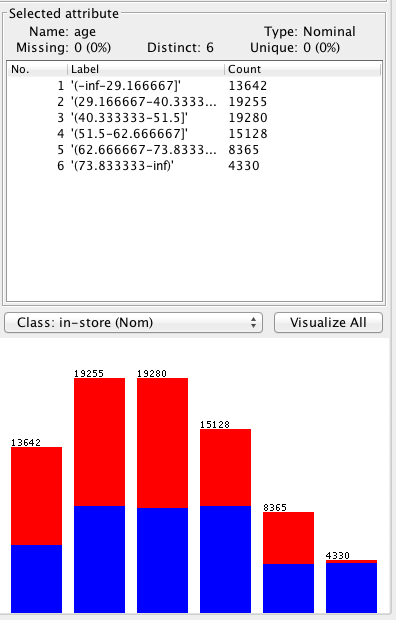
**Online vs. In-store (by Age)**

**Blue:** Online purchase

**Red:** In-store purchase

For age group 74-85, can see clearly that most purchases are done online.

Need to come up with stats for each.

****

So, run J48 on “In-store”:

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Demographic\_Data-weka.filters.unsupervised.attribute.NumericToNominal-R1,5-weka.filters.unsupervised.attribute.Discretize-B6-M-1.0-R2

Instances: 80000

Attributes: 5

in-store

age

items

amount

region

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

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

region = 1: 1 (16000.0)

region = 2: 0 (20000.0)

region = 3

| amount <= 1000.3: 1 (13186.0/2187.0)

| amount > 1000.3: 0 (4814.0)

region = 4

| amount <= 1999.9

| | age = '(-inf-29.166667]': 1 (6020.0/2168.0)

| | age = '(29.166667-40.333333]': 1 (6036.0/2118.0)

| | age = '(40.333333-51.5]': 1 (6090.0/2112.0)

| | age = '(51.5-62.666667]'

| | | amount <= 99.995: 1 (34.0)

| | | amount > 99.995: 0 (3344.0/1219.0)

| | age = '(62.666667-73.833333]': 0 (89.0)

| | age = '(73.833333-inf)': 1 (0.0)

| amount > 1999.9: 0 (4387.0)

Number of Leaves : 12

Size of the tree : 17

Time taken to build model: 0.57 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 70196 87.745 % 🡨 This is very **good**

Incorrectly Classified Instances 9804 12.255 %

Kappa statistic 0.7549

Mean absolute error 0.1685

Root mean squared error 0.2903

Relative absolute error 33.7073 %

Root relative squared error 58.0626 %

Total Number of Instances 80000

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.785 0.03 0.963 0.785 0.865 0.945 0

0.97 0.215 0.819 0.97 0.888 0.945 1

Weighted Avg. 0.877 0.123 0.891 0.877 0.876 0.945

=== Confusion Matrix ===

a b <-- classified as

31415 8585 | a = 0

1219 38781 | b = 1

Based on J48 tree, we can write the rules as:

* East region: shop in-store
* West region: shop online
* South region:
  + If amount <= $1000, shop in-store
  + If amount > $1000, shop online
* Central region:
  + If amount <= $2000, and
    - If age is 18 – 51, shop in-store
    - If age is 73 – 85, shop in-store
    - If age is 52 – 63, and
      * If amount <= $100, shop in-store
      * If amount > $100, shop online
  + If amount > $2000, shop online

**Online vs In-store**

**(Classify by regions)**

Use “Classify” tab and use tree.J48

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Demographic\_Data-weka.filters.unsupervised.attribute.NumericToNominal-R1,5-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R4-weka.filters.unsupervised.attribute.Discretize-B6-M-1.0-R2

Instances: 80000

Attributes: 5

in-store

age

items

amount

region

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

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

in-store = 0

| amount = '(-inf-1003.336467]'

| | age = '(-inf-29.166667]': 4 (2067.0/1106.0)

| | age = '(29.166667-40.333333]': 2 (5334.0/1524.0)

| | age = '(40.333333-51.5]': 2 (5350.0/1554.0)

| | age = '(51.5-62.666667]': 2 (5444.0/1559.0)

| | age = '(62.666667-73.833333]': 2 (3923.0/75.0)

| | age = '(73.833333-inf)': 2 (4109.0)

| amount = '(1003.336467-2001.668233]': 4 (7036.0/2446.0)

| amount = '(2001.668233-inf)': 4 (6737.0/2360.0)

in-store = 1

| amount = '(-inf-1003.336467]'

| | age = '(-inf-29.166667]': 1 (5017.0/3081.0)

| | age = '(29.166667-40.333333]': 1 (7317.0/4301.0)

| | age = '(40.333333-51.5]': 1 (7587.0/4543.0)

| | age = '(51.5-62.666667]': 3 (5353.0/2930.0)

| | age = '(62.666667-73.833333]'

| | | items <= 2: 3 (895.0/374.0)

| | | items > 2

| | | | items <= 3: 1 (608.0/291.0)

| | | | items > 3: 3 (2803.0/1290.0)

| | age = '(73.833333-inf)': 3 (221.0/86.0)

| amount = '(1003.336467-2001.668233]': 4 (10199.0/3622.0)

| amount = '(2001.668233-inf)': 1 (0.0)

Number of Leaves : 18

Size of the tree : 25

Time taken to build model: 0.48 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 48838 61.0475 %

Incorrectly Classified Instances 31162 38.9525 %

Kappa statistic 0.4737

Mean absolute error 0.2381

Root mean squared error 0.3451

Relative absolute error 64.2374 %

Root relative squared error 80.1573 %

Total Number of Instances 80000

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.501 0.184 0.405 0.501 0.448 0.826 1

0.972 0.079 0.805 0.972 0.881 0.974 2

0.255 0.075 0.495 0.255 0.337 0.769 3

0.645 0.185 0.627 0.645 0.636 0.796 4

Weighted Avg. 0.61 0.134 0.597 0.61 0.592 0.84

=== Confusion Matrix ===

a b c d <-- classified as

8019 0 4065 3916 | a = 1

0 19448 0 552 | b = 2

6235 1641 4592 5532 | c = 3

5535 3071 615 16779 | d = 4

**Select “(Nom) in-store” then “Start”**

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Demographic\_Data-weka.filters.unsupervised.attribute.NumericToNominal-R1,5-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R4-weka.filters.unsupervised.attribute.Discretize-B6-M-1.0-R2

Instances: 80000

Attributes: 5

in-store

age

items

amount

region

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

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

region = 1: 1 (16000.0)

region = 2: 0 (20000.0)

region = 3

| amount = '(-inf-1003.336467]': 1 (13194.0/2195.0)

| amount = '(1003.336467-2001.668233]': 0 (2446.0)

| amount = '(2001.668233-inf)': 0 (2360.0)

region = 4

| amount = '(-inf-1003.336467]'

| | age = '(-inf-29.166667]': 1 (2864.0/961.0)

| | age = '(29.166667-40.333333]': 1 (2913.0/997.0)

| | age = '(40.333333-51.5]': 1 (2991.0/1001.0)

| | age = '(51.5-62.666667]': 0 (1643.0/615.0)

| | age = '(62.666667-73.833333]': 0 (45.0)

| | age = '(73.833333-inf)': 1 (0.0)

| amount = '(1003.336467-2001.668233]'

| | age = '(-inf-29.166667]': 1 (3161.0/1212.0)

| | age = '(29.166667-40.333333]': 1 (3123.0/1121.0)

| | age = '(40.333333-51.5]': 1 (3101.0/1113.0)

| | age = '(51.5-62.666667]': 0 (1738.0/638.0)

| | age = '(62.666667-73.833333]': 0 (44.0)

| | age = '(73.833333-inf)': 1 (0.0)

| amount = '(2001.668233-inf)': 0 (4377.0)

Number of Leaves : 18

Size of the tree : 23

Time taken to build model: 0.23 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 70147 87.6838 % 🡸This is a **good** model …

Incorrectly Classified Instances 9853 12.3163 % Tree shows: Region 🡪 Amount 🡪 Age

Kappa statistic 0.7537

Mean absolute error 0.169

Root mean squared error 0.2907

Relative absolute error 33.7989 %

Root relative squared error 58.1445 %

Total Number of Instances 80000

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.785 0.031 0.962 0.785 0.864 0.946 0

0.969 0.215 0.818 0.969 0.887 0.946 1

Weighted Avg. 0.877 0.123 0.89 0.877 0.876 0.946

=== Confusion Matrix ===

a b <-- classified as

31400 8600 | a = 0

1253 38747 | b = 1

**Select (Nom) amount and “Start”**

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Demographic\_Data-weka.filters.unsupervised.attribute.NumericToNominal-R1,5-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R4-weka.filters.unsupervised.attribute.Discretize-B6-M-1.0-R2

Instances: 80000

Attributes: 5

in-store

age

items

amount

region

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

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

region = 1: '(-inf-1003.336467]' (16000.0/3622.0)

region = 2: '(-inf-1003.336467]' (20000.0)

region = 3

| in-store = 0: '(1003.336467-2001.668233]' (7001.0/4555.0)

| in-store = 1: '(-inf-1003.336467]' (10999.0)

region = 4

| in-store = 0

| | age = '(-inf-29.166667]': '(1003.336467-2001.668233]' (3310.0/2098.0)

| | age = '(29.166667-40.333333]'

| | | items <= 7: '(1003.336467-2001.668233]' (2948.0/1913.0)

| | | items > 7: '(2001.668233-inf)' (259.0/156.0)

| | age = '(40.333333-51.5]': '(1003.336467-2001.668233]' (3150.0/2037.0)

| | age = '(51.5-62.666667]'

| | | items <= 4

| | | | items <= 3

| | | | | items <= 2: '(-inf-1003.336467]' (707.0/451.0)

| | | | | items > 2: '(1003.336467-2001.668233]' (453.0/293.0)

| | | | items > 3: '(2001.668233-inf)' (440.0/278.0)

| | | items > 4: '(1003.336467-2001.668233]' (1604.0/1017.0)

| | age = '(62.666667-73.833333]'

| | | items <= 7

| | | | items <= 2: '(2001.668233-inf)' (24.0/12.0)

| | | | items > 2

| | | | | items <= 6: '(-inf-1003.336467]' (72.0/40.0)

| | | | | items > 6: '(2001.668233-inf)' (24.0/13.0)

| | | items > 7: '(1003.336467-2001.668233]' (8.0/2.0)

| | age = '(73.833333-inf)': '(1003.336467-2001.668233]' (0.0)

| in-store = 1: '(1003.336467-2001.668233]' (13001.0/6424.0)

Number of Leaves : 18

Size of the tree : 29

Time taken to build model: 0.27 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 56837 71.0463 %

Incorrectly Classified Instances 23163 28.9538 %

Kappa statistic 0.4098

Mean absolute error 0.2155

Root mean squared error 0.3284

Relative absolute error 70.8899 %

Root relative squared error 84.2286 %

Total Number of Instances 80000

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.789 0.201 0.902 0.789 0.841 0.869 '(-inf-1003.336467]'

0.697 0.273 0.412 0.697 0.518 0.807 '(1003.336467-2001.668233]'

0.094 0.017 0.339 0.094 0.147 0.9 '(2001.668233-inf)'

Weighted Avg. 0.71 0.201 0.749 0.71 0.713 0.858

=== Confusion Matrix ===

a b c <-- classified as

44194 11266 568 | a = '(-inf-1003.336467]'

4557 12010 668 | b = '(1003.336467-2001.668233]'

263 5841 633 | c = '(2001.668233-inf)'