CX

Christopher Eshleman

10/20/2019

CX Correlation

I'm looking at two variables in customer experience data, overall experience and one potential predictor (security screening time). Both are categorical and ordinal. First, check if they're even related, globally. Spearman rho correlation should do the trick. Second, visualize that relationship. A cross tab and, if possible, an accompanying heat map. Or faceted bar plots. Third, use an ordinal logistic regression or linear discriminant analysis to look closer. Fourth, if needed, consider controls via mutivariate analysis. This may require some automated variable selection given the nature of the data set (thousands of potential covariates).

Note: While trying to read a SPSS file with the read.spss function I found an error. It was due to label (not variable name) repetition in the raw data. A custom (not from me) function provides a work-around:

Now call read.spss with the argument use.value.labels set to FALSE and, later, convert the SPSS categorical variables into R factors with the function:

```
## re-encoding from CP1252
```

```
## Duplicated labels: Gate 8 Other Gate 6 Gate 5 Gate 4 Gate 3 Gate 2 Gate 1 Other Gate 8 Gate 7 Gate 6
## Duplicated labels: WestJet Air Canada Jazz Other US Airways (also Express, Piedmont. Republic Air, Cl
## Duplicated labels: Wyoming Washington Indiana Delaware Warren Union Somerset Middlesex Mercer Cumberl
## Duplicated labels: Rock Tavern Factoryville Thomaston Thomaston Watchung Central Islip Thomaston Moni
## Duplicated labels: Wyoming Washington Indiana Delaware Warren Union Somerset Middlesex Mercer Cumberl
## Duplicated labels: Wyoming Washington Indiana Delaware Warren Union Somerset Middlesex Mercer Cumberl
## Duplicated labels: Wyoming Washington Indiana Delaware Warren Union Somerset Middlesex Mercer Cumberl
## Duplicated labels: Wyoming Washington Indiana Delaware Warren Union Somerset Middlesex Mercer Cumberl
## Duplicated labels: Rock Tavern Factoryville Thomaston Thomaston Watchung Central Islip Thomaston Moni
## Duplicated labels: Dublin Ireland Atlanta(Intl) GA USA Wilmington NC USA Sarasota/Bradenton FL USA Ro
```

Any correlation analysis needs to grapple with the wide (big) nature of the data — nearly 2,600 variables. A few obvious candidates for focus pop out. First, and most importantly, OVERALLEXPERIENCEFS looks like a catch-all for — well, for travelers' individual experiences using the facility:

```
##  Factor w/ 10 levels "10-Outstanding",..: 5 2 1 1 1 2 4 4 4 2 ...
```

Overall experience is ranked 1-10. But these numbers represent rankings along a qualitative experience scale. So technically it's a qualitative variable (not quantitative) despite the deceptive numeric label. It's an ordinal data point — ordered categories. This is in the weeds but it has implications for analytical purposes later.

```
##  Factor w/ 10 levels "10-Outstanding",..: 5 2 1 1 1 2 4 4 4 2 ...
```

```
## 
##      10-Outstanding    9    8    7    6    5    4    3    2 1-Unacceptable
##   1               0    0    0    0    0    0    0    0    0              0
##   2               0    0    0    0    0    0    0    0   45              0
##   3               0    0    0    0    0    0    0   84    0              0
##   4               0    0    0    0    0    0  158    0    0              0
##   5               0    0    0    0    0  378    0    0    0              0
##   6               0    0    0    0  596    0    0    0    0              0
##   7               0    0    0 1556    0    0    0    0    0              0
##   8               0    0 1642    0    0    0    0    0    0              0
##   9               0 1155    0    0    0    0    0    0    0              0
##   10              0    0    0    0    0    0    0    0    0              0
```

```
##  Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 6 9 NA NA NA 9 7 7 7 9 ...
```

Beyond that, analysts' speculation around JD Power airport satisfaction rankings identified security screening as one predictor. The data includes SecurityScreeningTime as a variable:

```
str(cx$SecurityScreeningTime)
```

```
##  Factor w/ 8 levels "More than 45 mins.",..: 2 NA 6 6 2 NA 4 5 NA 7 ...
```

This is also reported qualitatively, despite the fact that the underlying variable being measured is quantitative and continuous. We can set up a continuous (imputed) representation even though it'll lean on certain assumptions regarding actual distribution. For now let's stick to using the data as is (qualitative).

But classify it as ordinal with the rank we want.

```
## 
## More than 45 mins.        30-45 mins.        20-29 mins. 
##                105                241                460 
##        15-19 mins.        10-14 mins.          5-9 mins. 
##                711               1207               1832 
##          1-4 mins.         < 1 minute 
##                697                147
```

```
##  Ord.factor w/ 8 levels "< 1 minute"<"1-4 mins."<..: 7 NA 3 3 7 NA 5 4 NA 2 ...
```

What's a raw correlation of these two variables look like? Use a Spearman correlation coefficient for categorical (rank) variables such as these.

```
##                            factor_SecurityScreeningTime
## factor_OVERALLEXPERIENCEFS < 1 minute 1-4 mins. 5-9 mins. 10-14 mins.
##                          1            0         0         0           0
##                          2            0         5         8           7
##                          3            1         2        11          15
##                          4            1        11        26          33
##                          5            7        23        74          77
##                          6           11        50       142         121
##                          7           23       168       531         294
##                          8           29       165       478         330
##                          9           31       153       354         210
##                          10           0         0         0           0
##                            factor_SecurityScreeningTime
## factor_OVERALLEXPERIENCEFS 15-19 mins. 20-29 mins. 30-45 mins.
##                          1             0           0           0
##                          2            10           2           2
##                          3            23          10           5
##                          4            23          13           8
##                          5            63          49          21
##                          6            81          60          31
##                          7           156         115          62
##                          8           170         105          60
##                          9           119          75          30
##                          10            0           0           0
##                            factor_SecurityScreeningTime
## factor_OVERALLEXPERIENCEFS More than 45 mins.
##                           1                   0
##                           2                   0
##                           3                   5
##                           4                  10
##                           5                  15
##                           6                  14
##                           7                  17
##                           8                  21
##                           9                   8
##                           10                  0
```

cor(cx$factor_OVERALLEXPERIENCEFS, cx$factor_SecurityScreeningTime, method=“spearman”, use=“pairwise.complete.obs”)

cor.test(cx$factor_OVERALLEXPERIENCEFS, cx$factor_SecurityScreeningTim method=“spearman”, use=“pairwise.complete.obs”)

Good. Figure out how to plot this as bar charts, a heat map, or both. For now, let’s just run things. We could use an ordinal logistic regression or a linear discriminant analysis. There’s a fairly accessible MASS package that includes proportional odds logistic regression functionality:

```
##
##                          1   2   3   4   5   6   7   8   9  10
##   < 1 minute             0   0   1   1   7  11  23  29  31   0
##   1-4 mins.              0   5   2  11  23  50 168 165 153   0
##   5-9 mins.              0   8  11  26  74 142 531 478 354   0
##   10-14 mins.            0   7  15  33  77 121 294 330 210   0
##   15-19 mins.            0  10  23  23  63  81 156 170 119   0
##   20-29 mins.            0   2  10  13  49  60 115 105  75   0
##   30-45 mins.            0   2   5   8  21  31  62  60  30   0
##   More than 45 mins.     0   0   5  10  15  14  17  21   8   0
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = factor_OVERALLEXPERIENCEFS ~ SecurityScreeningTime,
##     data = cx)
##
## Coefficients:
##                                   Value Std. Error t value
## SecurityScreeningTime30-45 mins. 0.7042     0.2294   3.070
## SecurityScreeningTime20-29 mins. 0.7581     0.2140   3.543
## SecurityScreeningTime15-19 mins. 0.8028     0.2084   3.853
## SecurityScreeningTime10-14 mins. 1.0733     0.2030   5.288
## SecurityScreeningTime5-9 mins.   1.2597     0.2005   6.282
## SecurityScreeningTime1-4 mins.   1.4136     0.2095   6.746
## SecurityScreeningTime< 1 minute  1.4939     0.2664   5.609
##
## Intercepts:
##      Value     Std. Error t value
## 1|2   -10.1215    3.6295    -2.7887
## 2|3    -3.8938    0.2553   -15.2506
## 3|4    -2.7391    0.2130   -12.8607
## 4|5    -1.9276    0.2014    -9.5718
## 5|6    -0.9549    0.1965    -4.8589
## 6|7    -0.1639    0.1958    -0.8368
## 7|8     1.1388    0.1965     5.7963
## 8|9     2.4638    0.1983    12.4229
## 9|10  213.5455    0.1983  1076.7244
##
## Residual Deviance: 15719.42
## AIC: 15751.42
## (1603 observations deleted due to missingness)
```

For a one-step increase in screening time (around 5 or 10 minutes), expect a x increase in the expected value of Y. (In the log odds scale.)

LDA . . . .

So does security wait time carry a strong association with overall experience?

These numbers represent decreasing odds of reporting lower satisfaction categories as you experience higher security screening times. They differ significantly across wait times. A next step might be to try and isolate which pairs of wait time-satisfaction differed, but that'd take a little control work.