Readme

Dmitri Kazanski

Table of Content:

Data Cleanup and Transformation

I have conducted the following clean-up and transformation of the data:

1) Deleted the “Car” column since it only contained 108 non-null values and was otherwise empty
2) The following columns contained some (not a lot) null values:
    - Bar 107
    - RestaurantLessThan20 130
    - CarryAway 151
    - Restaurant20To50 189
    - CoffeeHouse 217

    Since they are categorical variables, I imputed them with the Mode
3) I noticed some odd/non-standard values such as “never” and “less1”. Sometimes the same column would contain both “never” and “less1” (less than 1). Since fractional values are impossible, “less than 1” must mean 0. So, I imputed both “never” and “less1” with “0”.
4) I noticed that the “toCoupon_GEQ5min” variable was always equal to “1”. Thus, it is a useless variable. I dropped that column.
5) Finally, I had a suspicion that “direction_opp” is the exact opposite of “direction_same” (when one is equal to 1, the other is equal to 0 and vice-versa). This one of these variables is redundant and will be perfectly auto-correlated. So, I dropped “direction_opp”

Having completed the cleanup and transformation, I am now ready for data visualization and Exploratory Data Analysis (EDA).

Data Visualization and Analysis

The suggested prompt leads us through an exploration/visualization of a few values of a few individual variables and their impact on the desired outcome (acceptance of coupons).

Then the prompt suggested creating a few combinations of those variables.

I believe the above is a good exercise, but I do not want to use the same framework to complete my analysis because of the following reasons:

1. For the future modeling exercise, it would be good to know what variables are useful (predictive of the outcome) and what are useless. Thus, I wanted to look individually at the impact of ALL values and ALL variables.
2. I calculated that the total number of combinations of values from each variable is around 19 trillion (19,110,297,600,000). Thus manually combining some values from some variables in the hope of finding something interesting (a strong relationship with coupon acceptance) is futile. Our universe will disintegrate before we will be done.

Instead, I adopted the strategy listed below:

Data Visualization Analysis: Goals

1) As suggested by the prompt, I selected one type of coupon (Cofee House coupon).
2) I looked at ALL values of ALL variables on the desired outcome (coupon acceptance). This analysis will be useful in the future modeling effort.

Data Visualization Analysis: Methods

As I describe in the “Statistical Analysis” section below, all of the data variables in this data frame are actually categorical. None are numerical. Therefore, charts that apply to numerical data (scatter plots, box plots, etc.,) will not be useful here. Count plots seem most useful.

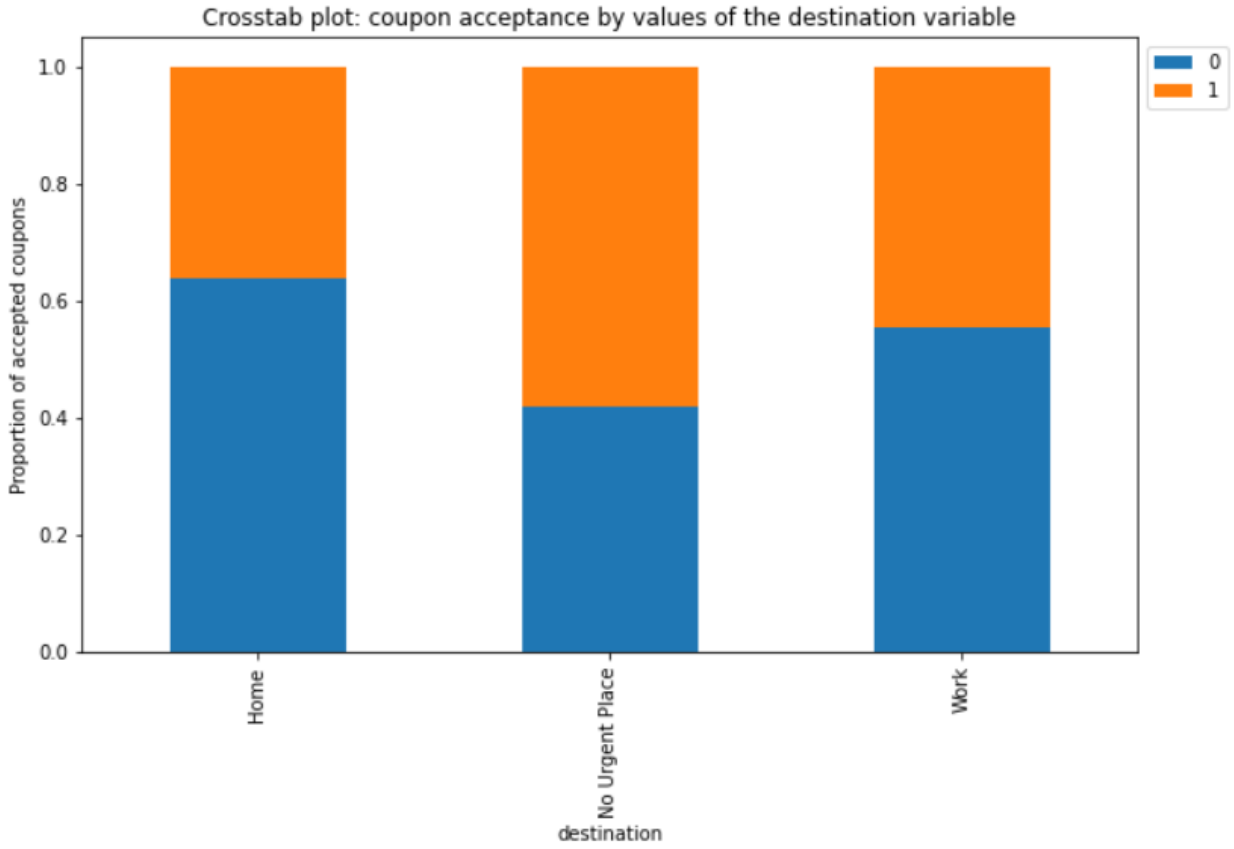
Using a stacked cross-tabulation bar plot, I then plotted and analyzed the impact of all values of all variables on coupon acceptance. To do that, I first created a function that generated the plot. Then, All I needed to do was to drop the name of the column.

The same function that generated the above data in the form of a table so that I can see the exact percentage. I am not showing the percentages here for the sake of brevity. Showing them in the notebook.

Data Visualization Analysis: Results

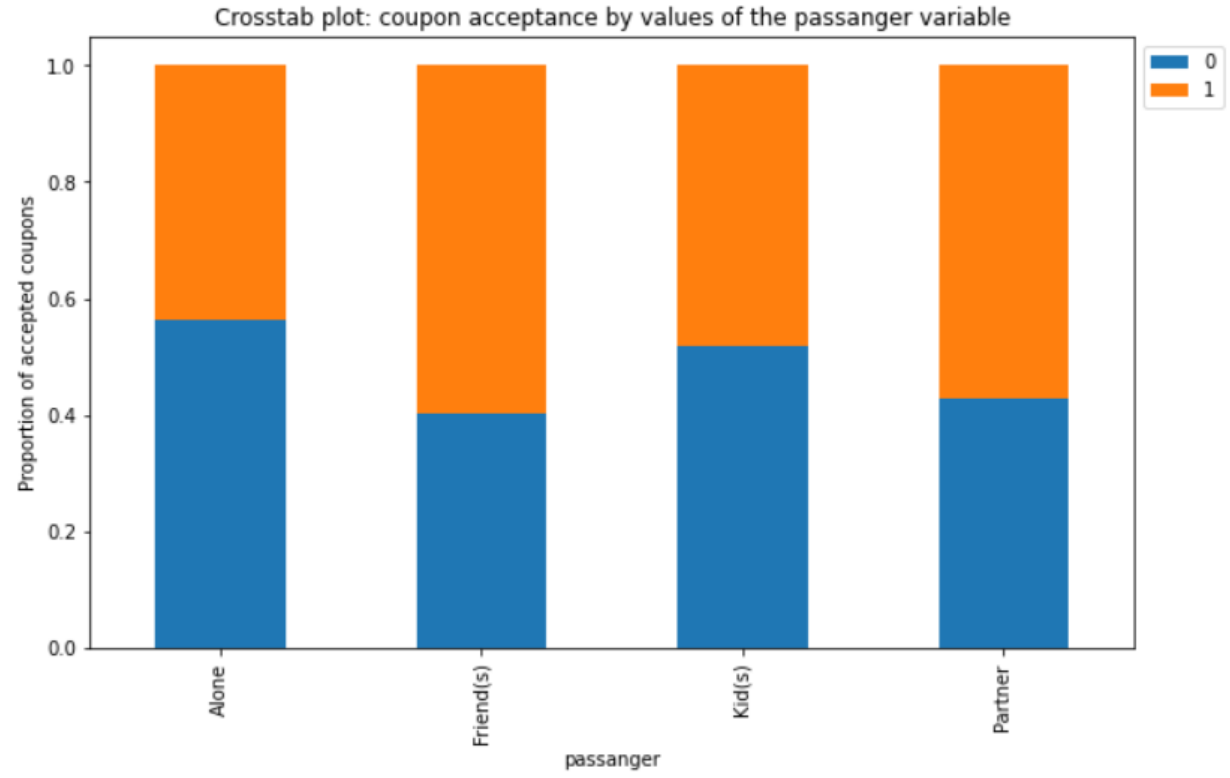
The following is the summary of the analysis of the impact of particular variables (and their values) on the Coffee House coupon acceptance.

Destination:



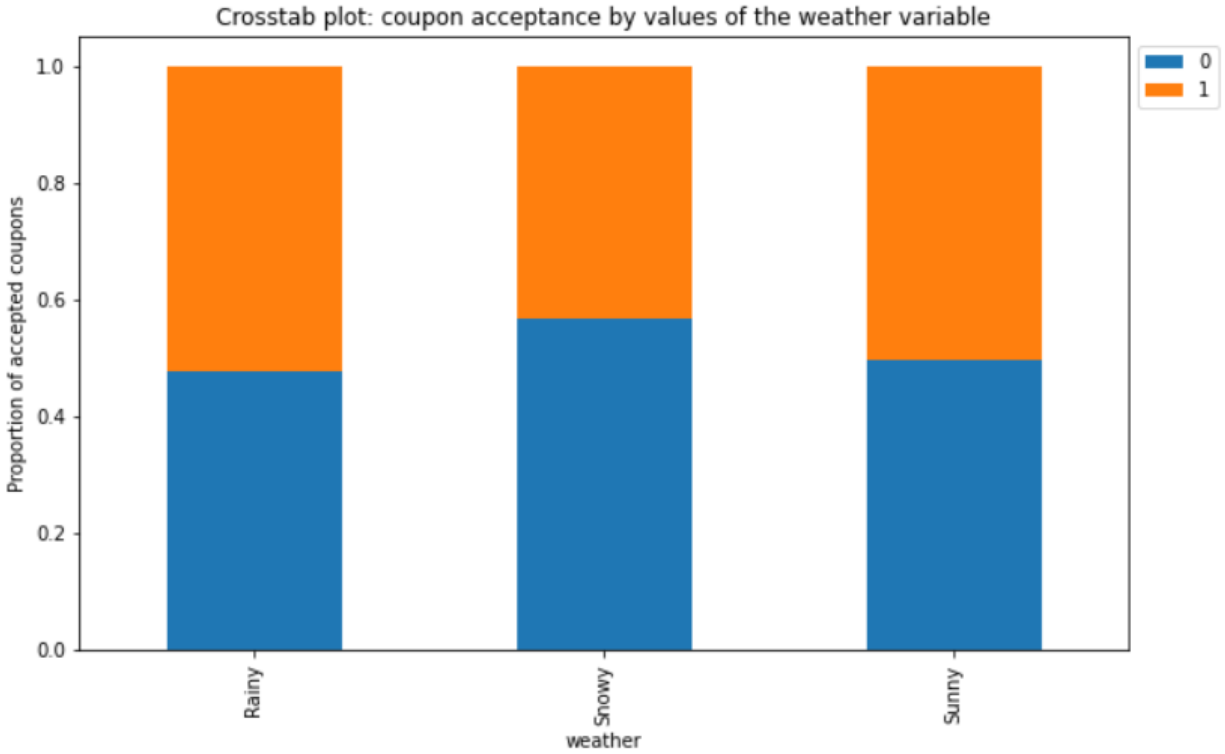
- “No urgent place” destination had the highest Coffee House coupon acceptance (58%)
- ”Home had the lowest (36%)

Passenger:



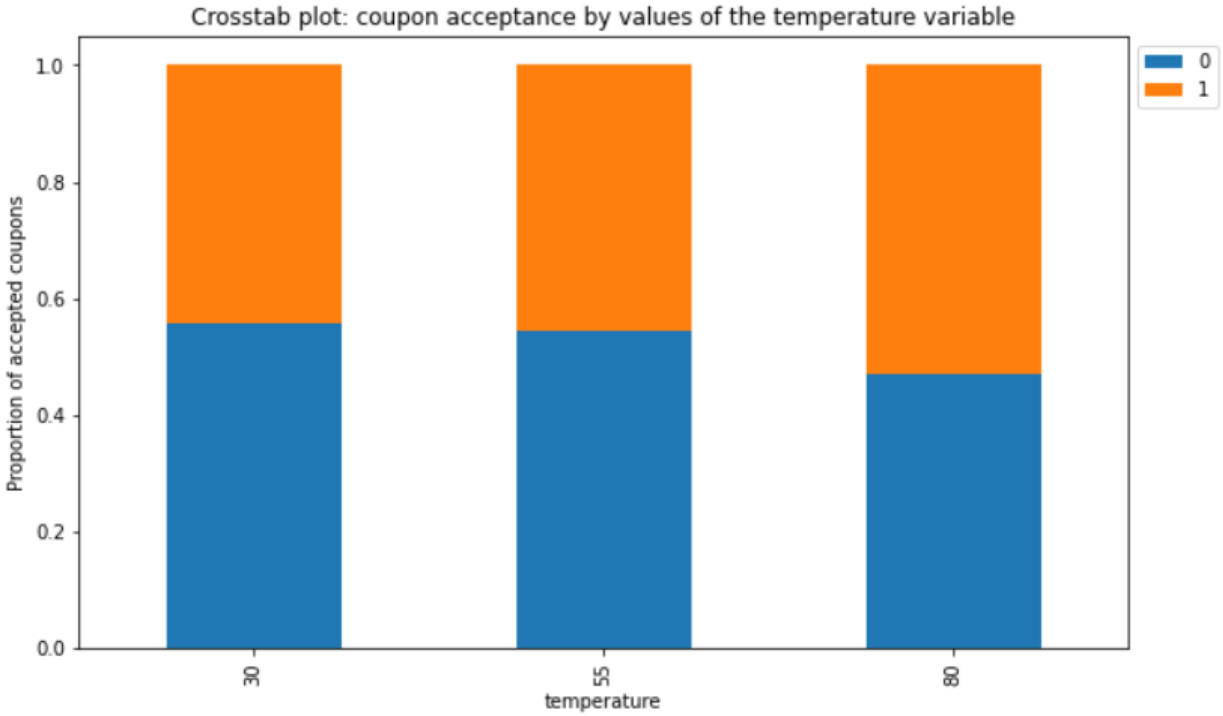
- “Friend” and ”Partner” passengers had the highest Coffee House coupon acceptance at 60% and 57%, respectively.

Weather:



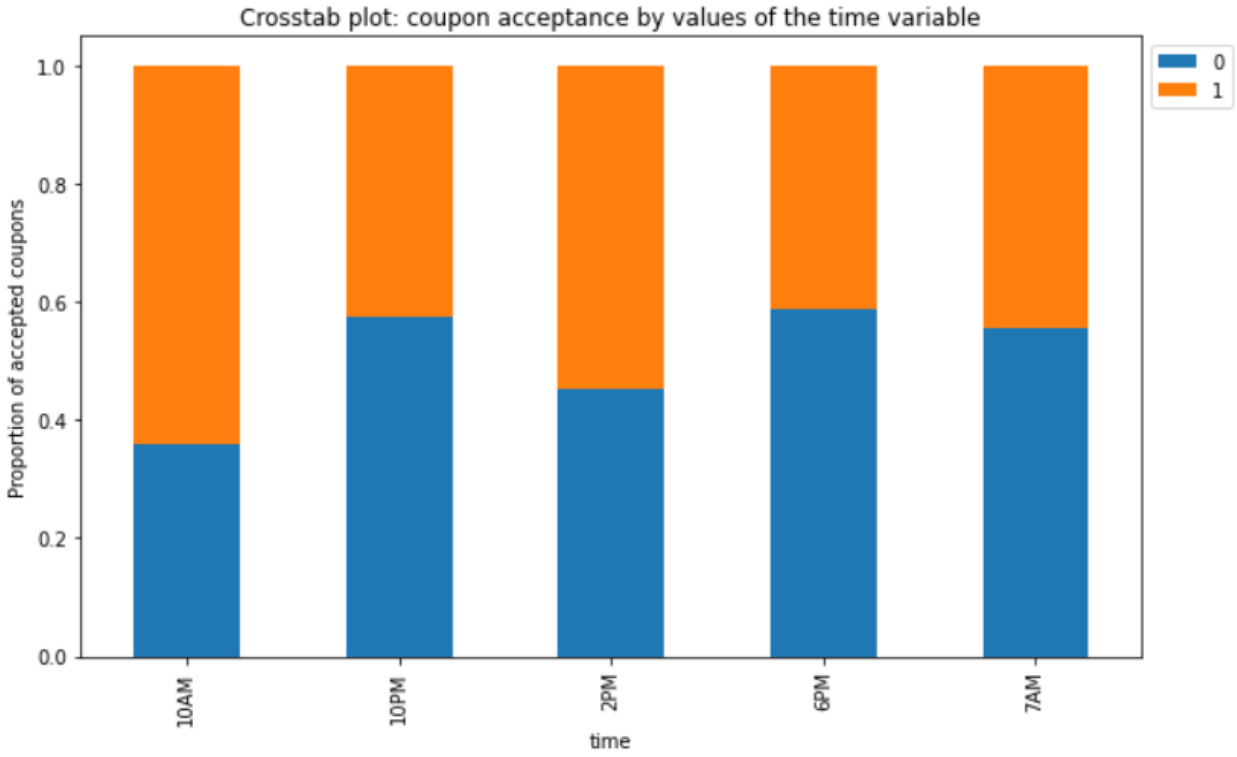
- Coffee House coupon acceptance was the lowest (43%) during Snowy weather, which is not surprising because driving in the snow is risky and slow.
- Rainy weather was the best (52% acceptance). It is known that people who live in rainy climates (ex.g., Seattle, consume more coffee).

Temperature:



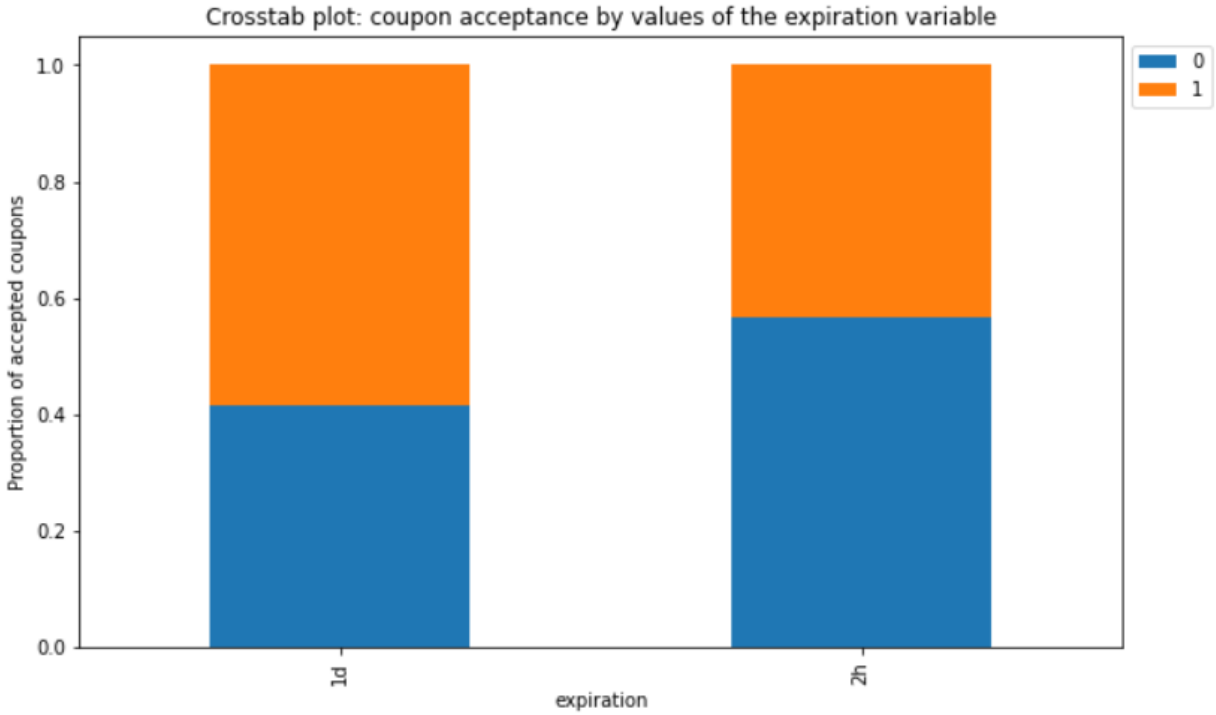
- Warm weather (80F) had the highest acceptance rate (53%)

Time of day:



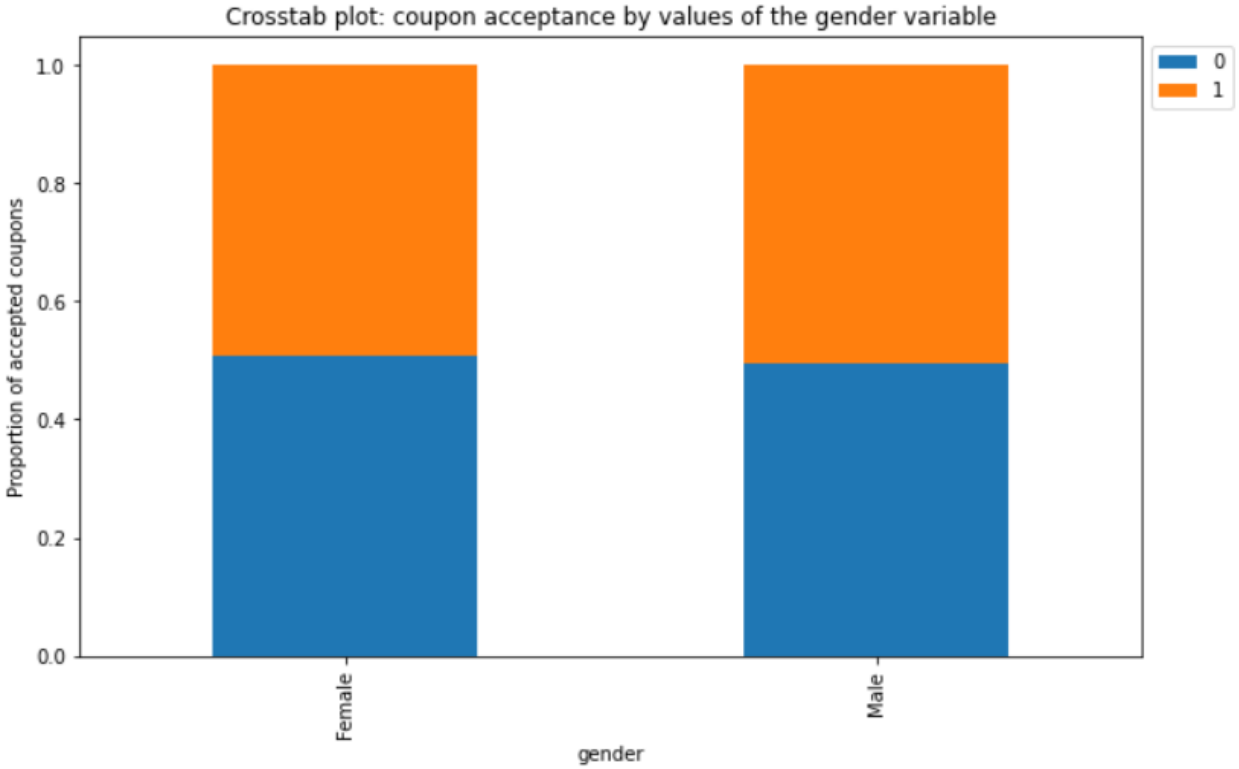
- Late morning (10 am) time had the highest acceptance of Coffee Shop coupons (64%).
- 2 pm was the 2nd highest (54%). This is when people often feel sleepy and need another coffee.

Expiration:



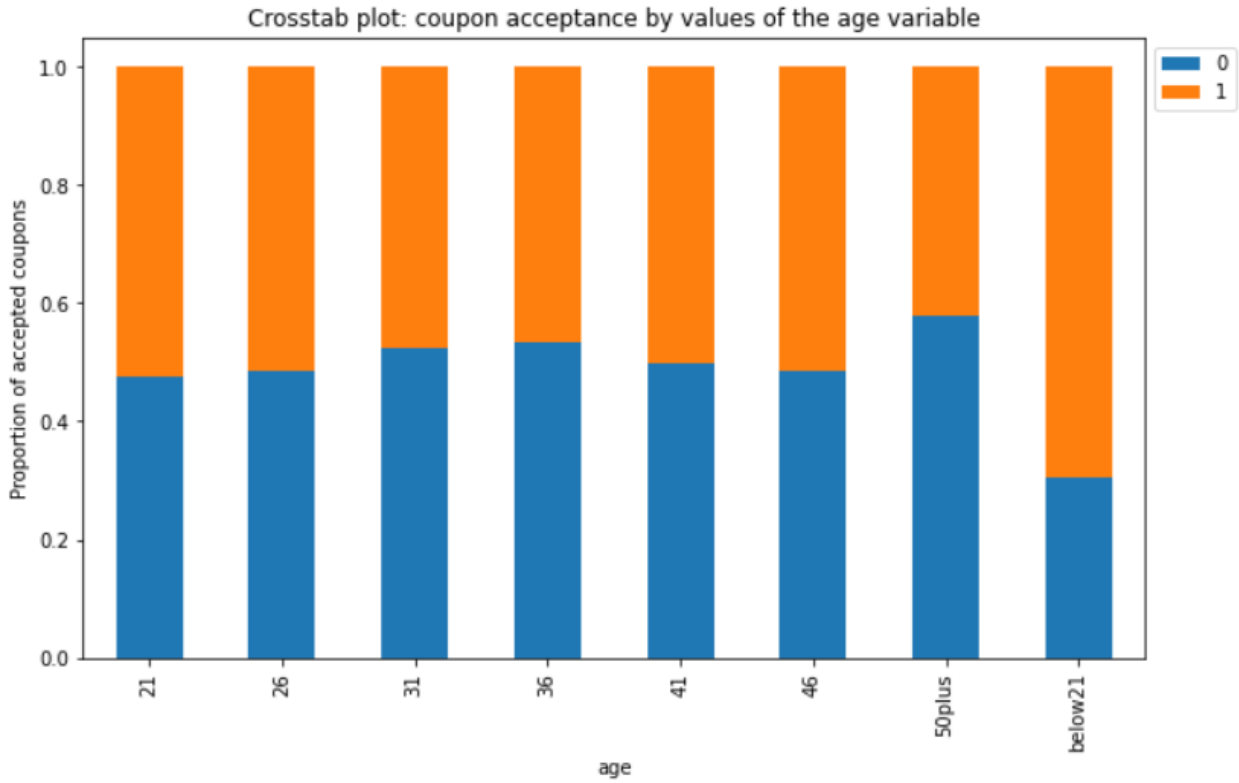
- Coupons that were set to expire in one day had a greater acceptance rate (58%) than the coupons that were set to expire in two hours (43%). Perhaps the users found the former to be more useful.

Gender:



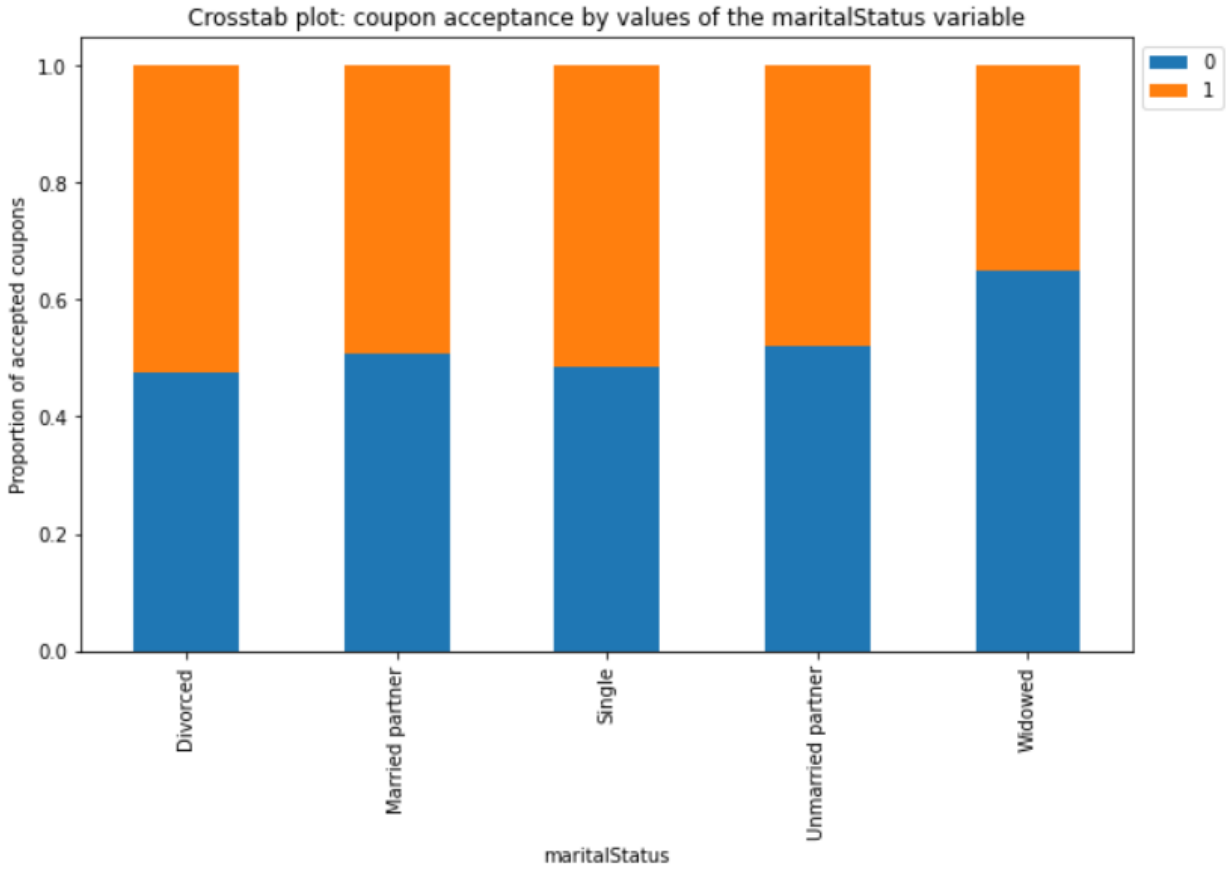
- Gender has no impact. The difference in acceptance rate is negligible.

Age:



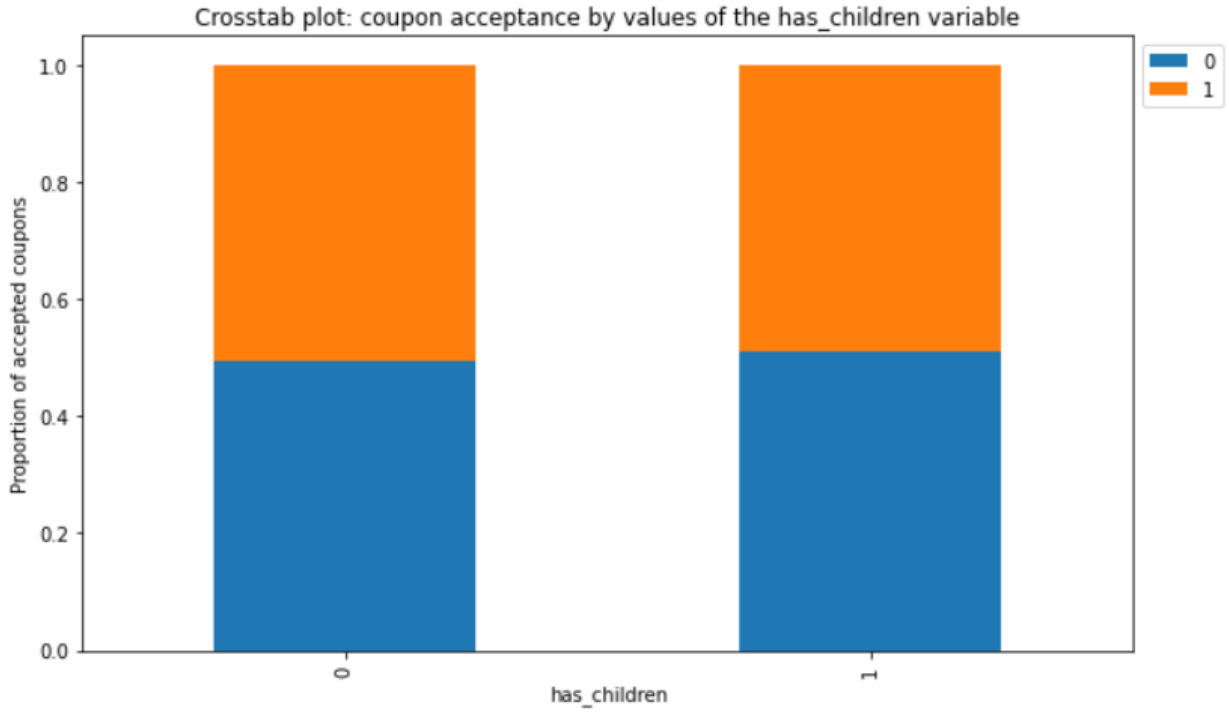
- Younger folks had a better Coffee House Coupon acceptance rate.
- Folks under 21 years of age had it as high as 52%
- Folks over 50 had the lowest acceptance rate (42%)

Marital Status:



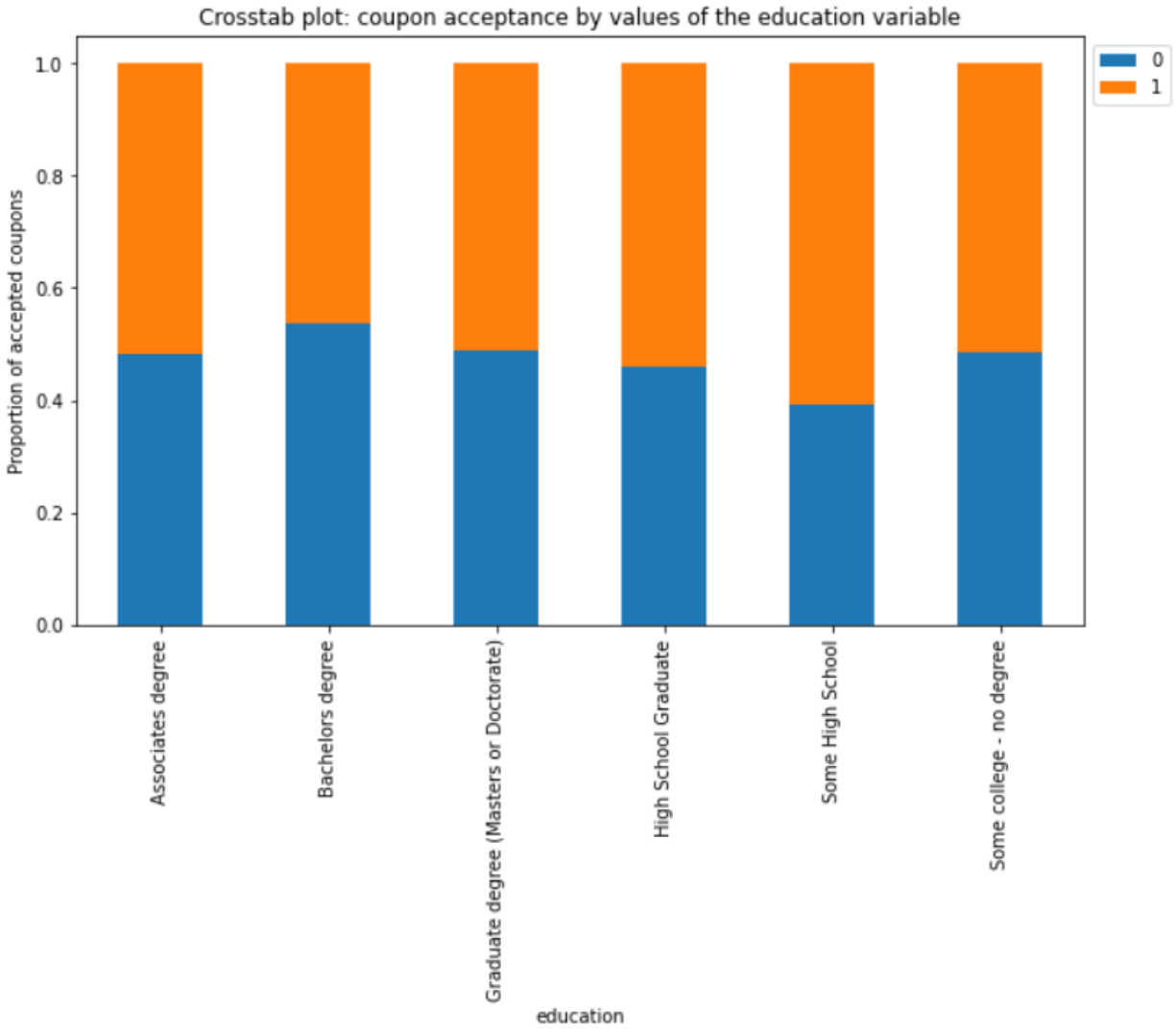
- Divorced and single people had the highest acceptance rate (52% and 51.6%, respectively)
- Widowed people had the lowest rate (35%)

Children at home:



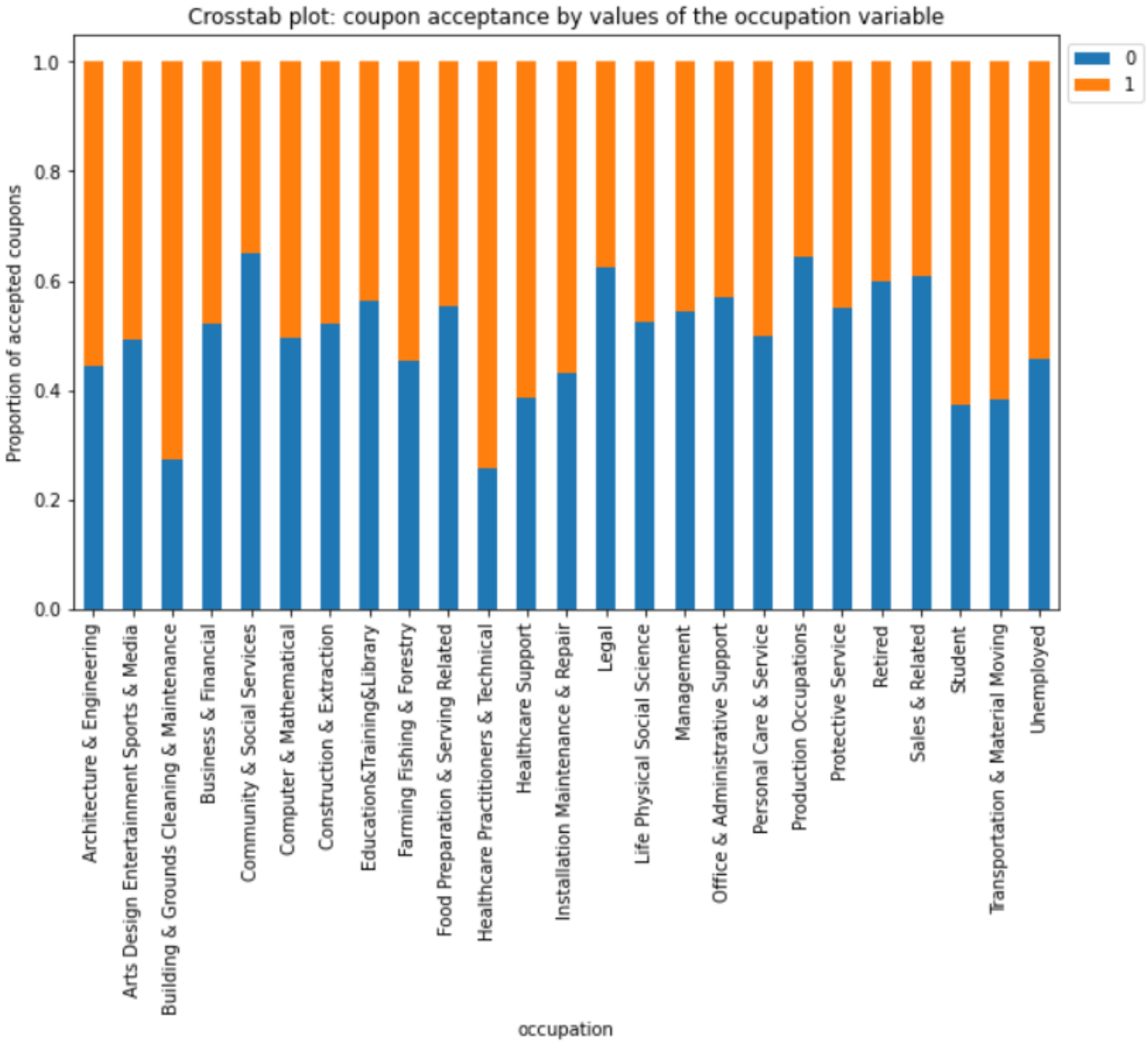
- The impact of having children was negligible, with about 1.1% difference in acceptance rate.

Education:



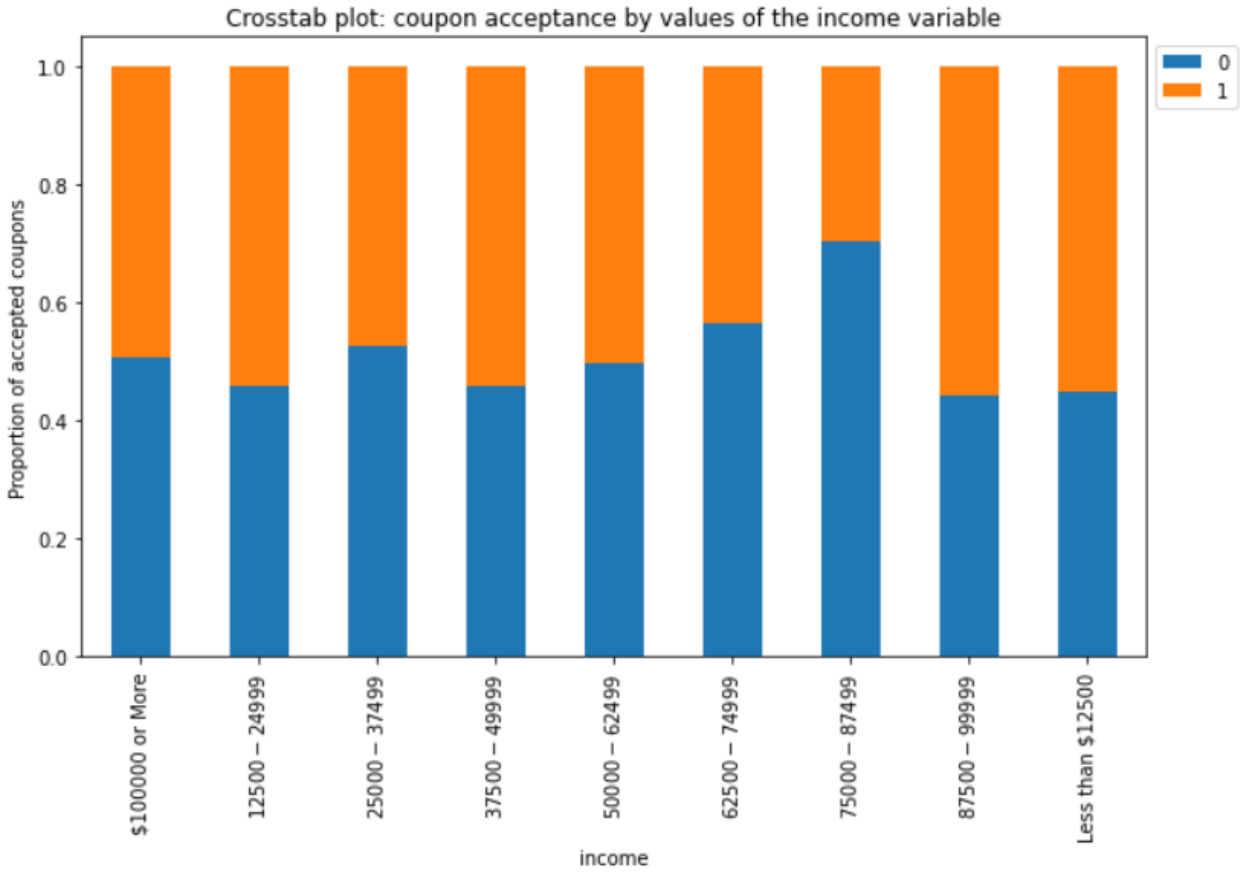
- Less educated people (some high school, high school graduates, and folks with associate degrees) had the highest acceptance rates.

Occupation:



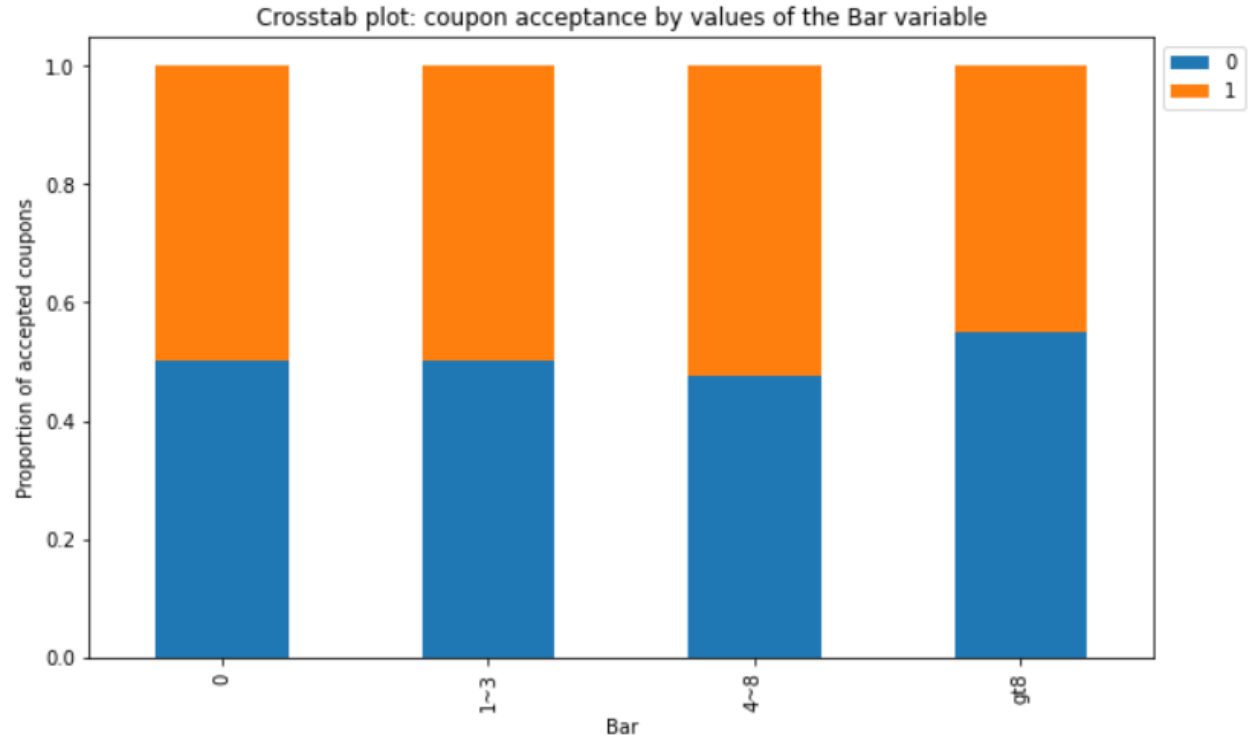
- Healthcare practitioners had the highest acceptance rate for Coffee House coupons (74%). This is not surprising. Often, healthcare professionals, especially medical residents, are overworked and are in strong need of coffee to stay alert.
- Building & Ground Cleaning & Maintenance group was the 2nd highest (73%)
- Community and Social workers had the lowest acceptance rates (35%)

Income:



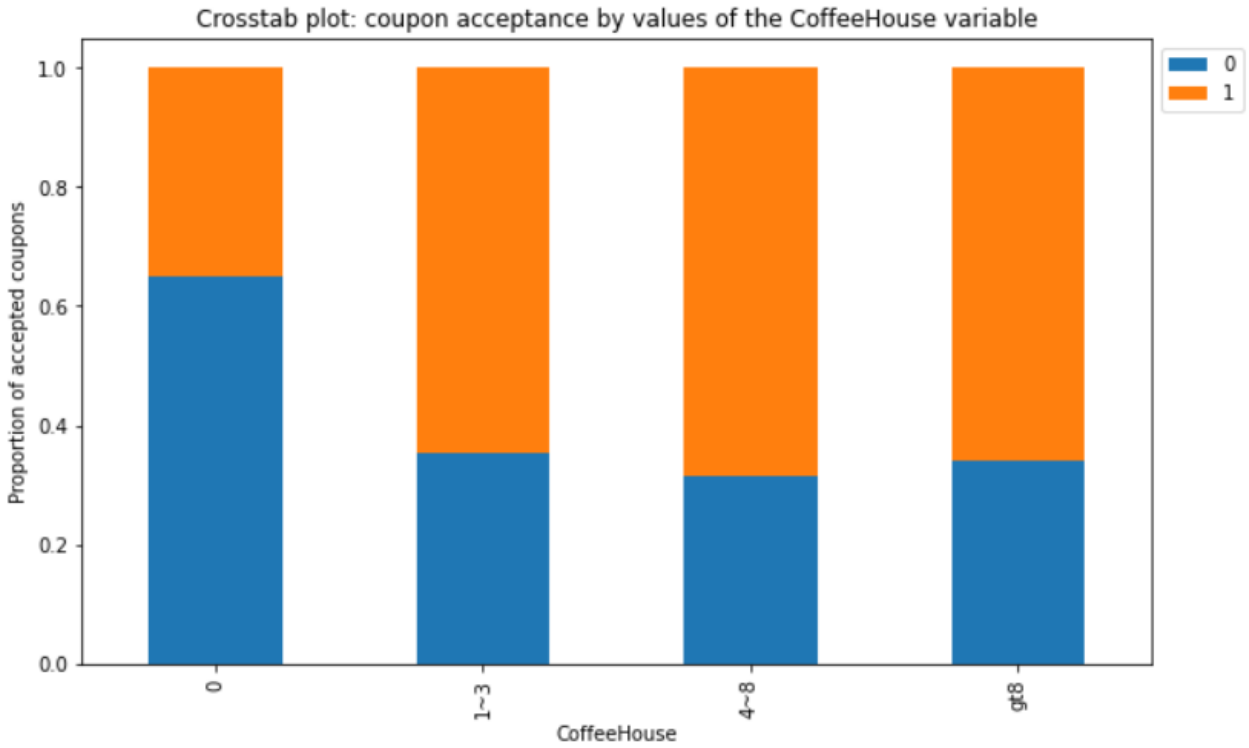
- The acceptance rate by income data looked a bit odd::
    - The lowest income bracket and one of the highest brackets had high acceptance rates (55% and 56% respectively)
    - but the 2nd highest income bracket ($75-$87k) had the lowest acceptance rate (30%)

Bar Visits:



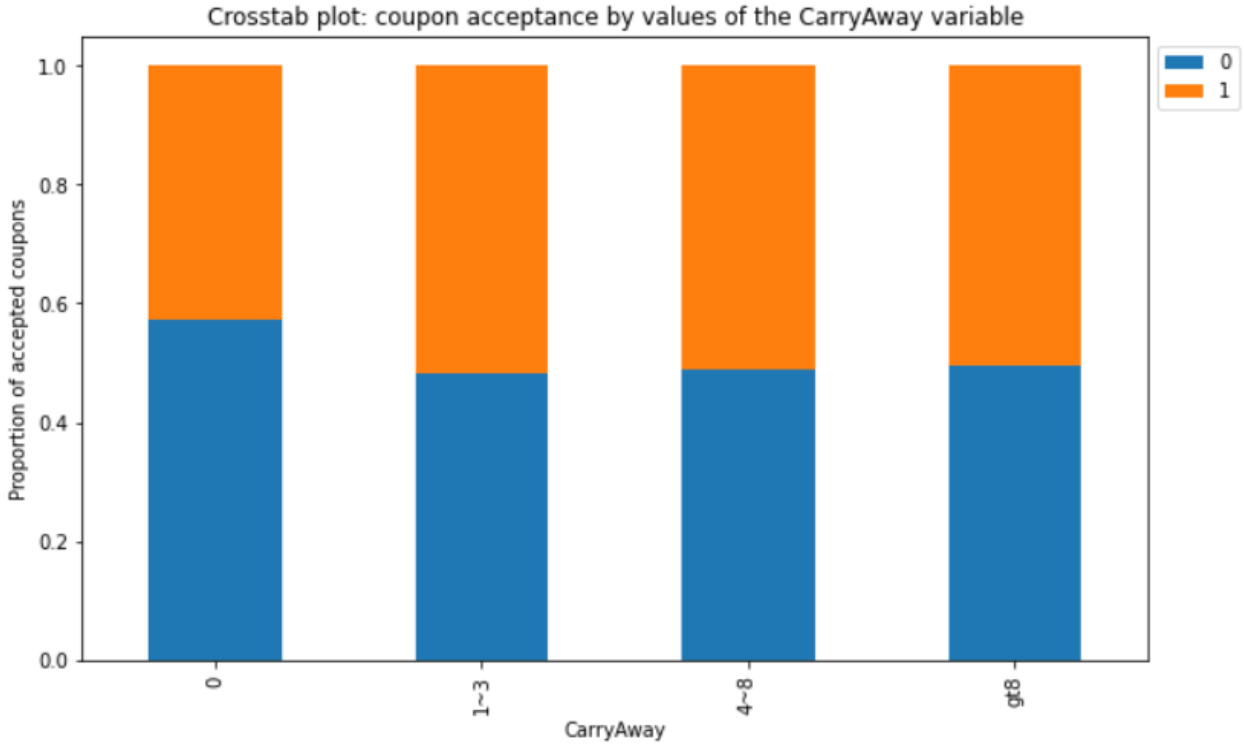
- Bar visits have little impact on the Coffee House Coupon acceptance. The differences are pretty minor.

Coffee House Visits:



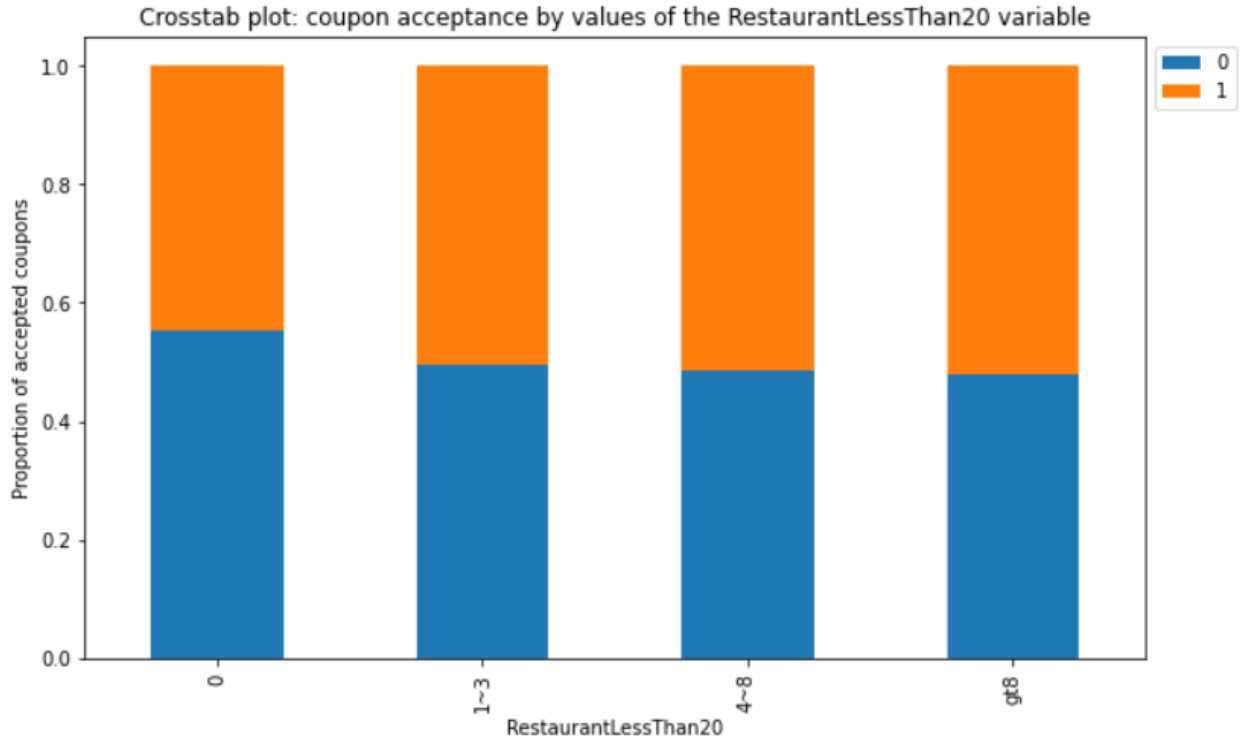
- As one would expect, prior Coffee House visits have a **massive** influence on the Coffee House coupon acceptance:
    - Whereas folks who never visited Coffee House, accepted the coupons only 19% of the time, folks who visited a lot (4+ times) accepted the coupon 66-69% of the time.
    - This will likely be the greatest predictor of the Coffee House coupon acceptance.
    - The same is likely true for other types of coupons (but needs to be verified).

Carry Away:



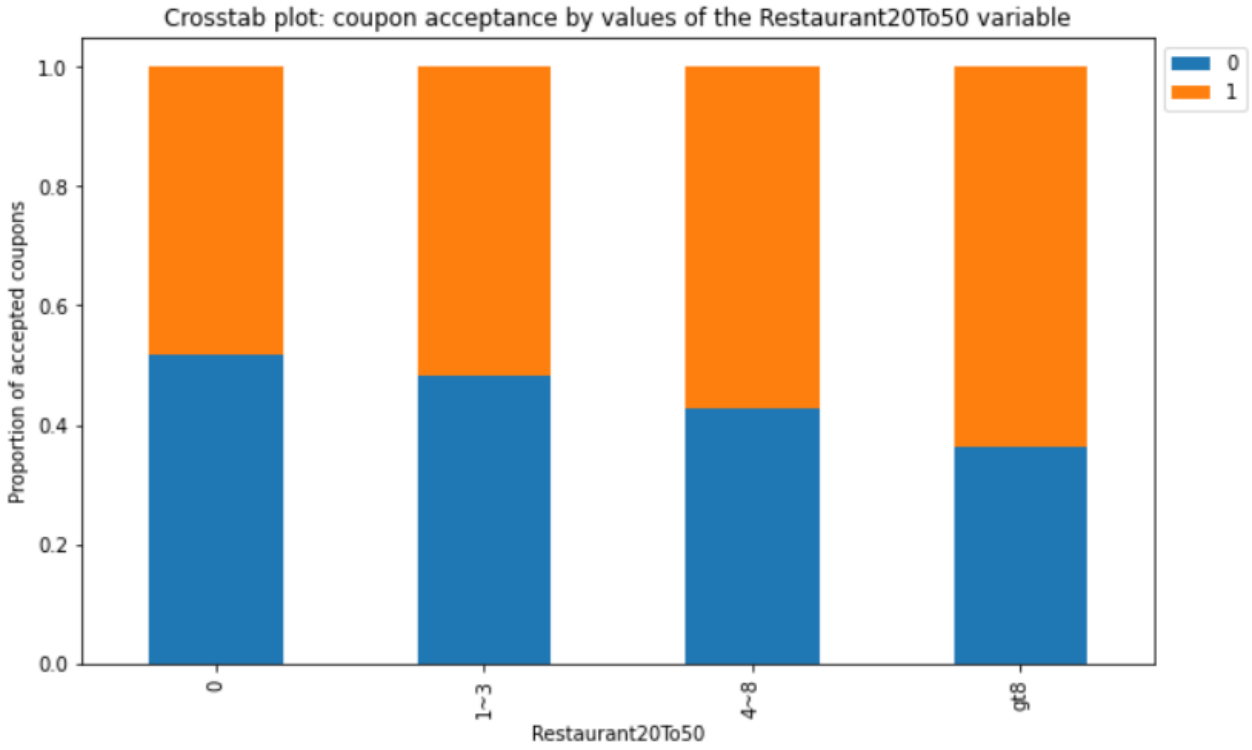
- There might be a small dependency here. People who never did Carry Away are less likely to accept the Coffee House coupons.

Cheap (less than $20) Restaurant visits:



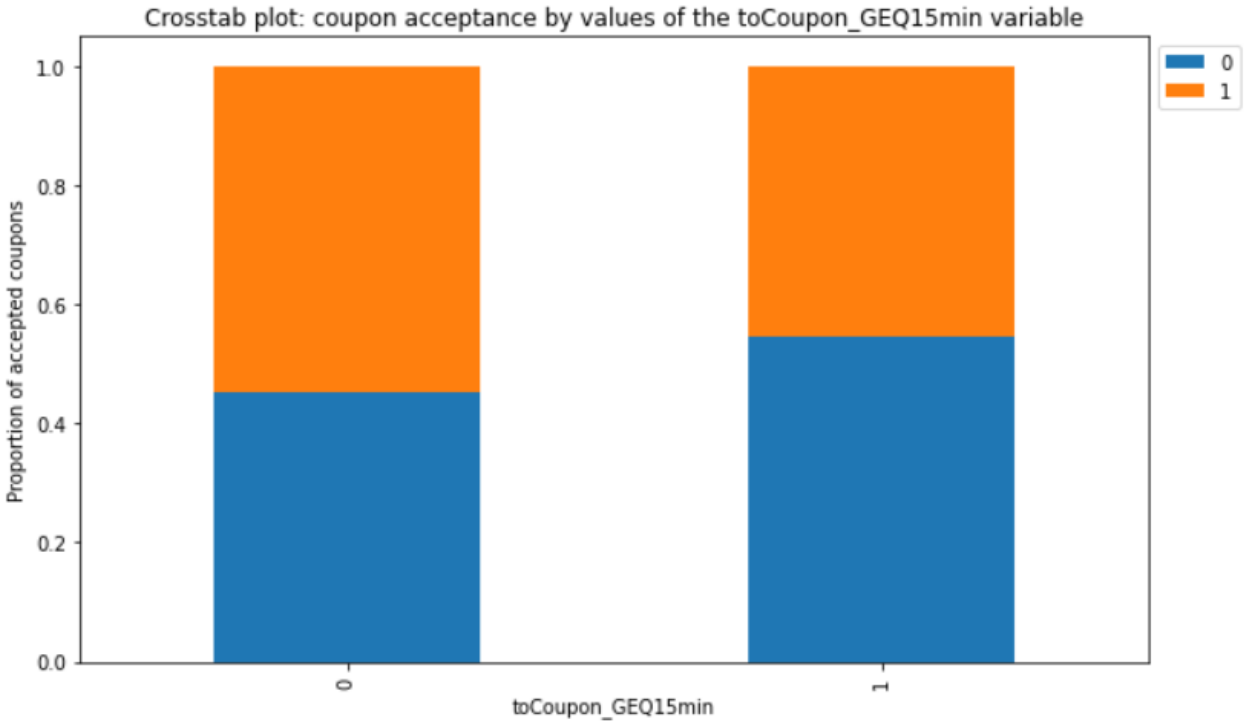
- There is a small dependency here. People who never went to a cheap restaurant coupon are less likely to accept the Coffee House coupon.

More expensive ($20 to $50) restaurant visits:



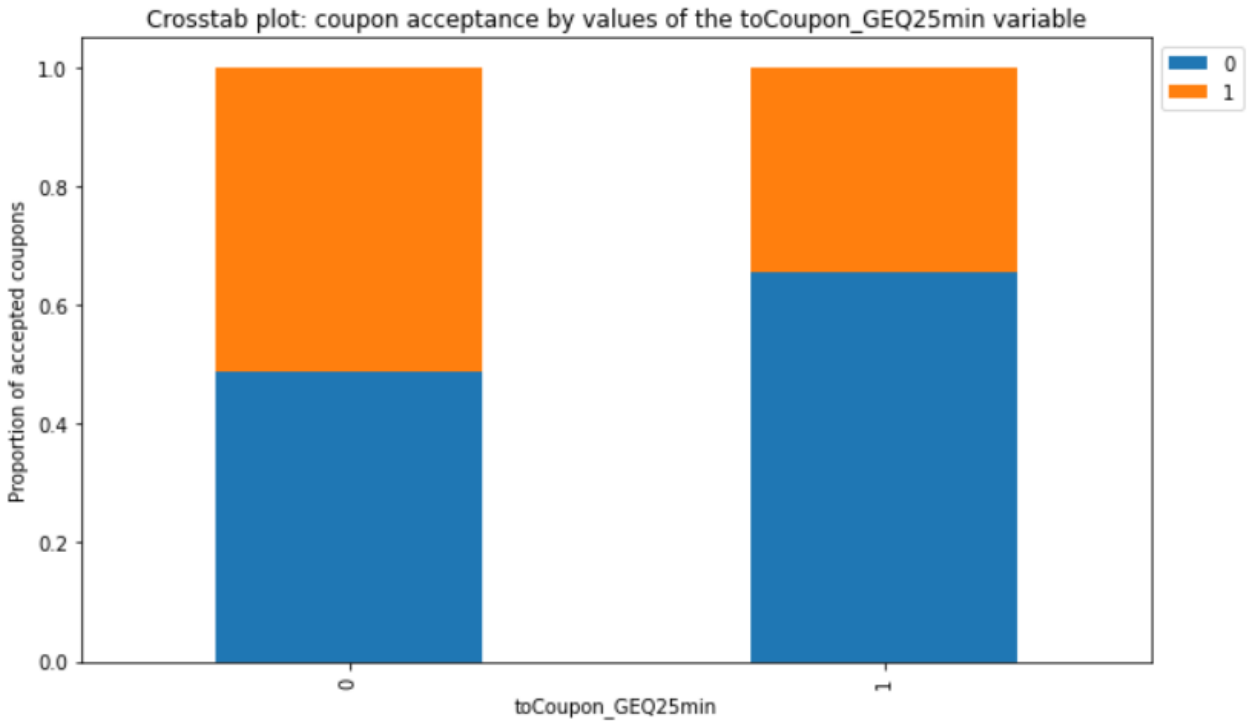
- The Restaurant 20 to 50 variable has an impact. The more people go to these restaurants, the higher the likelihood of them accepting the Coffee House coupon.

GEQ15min



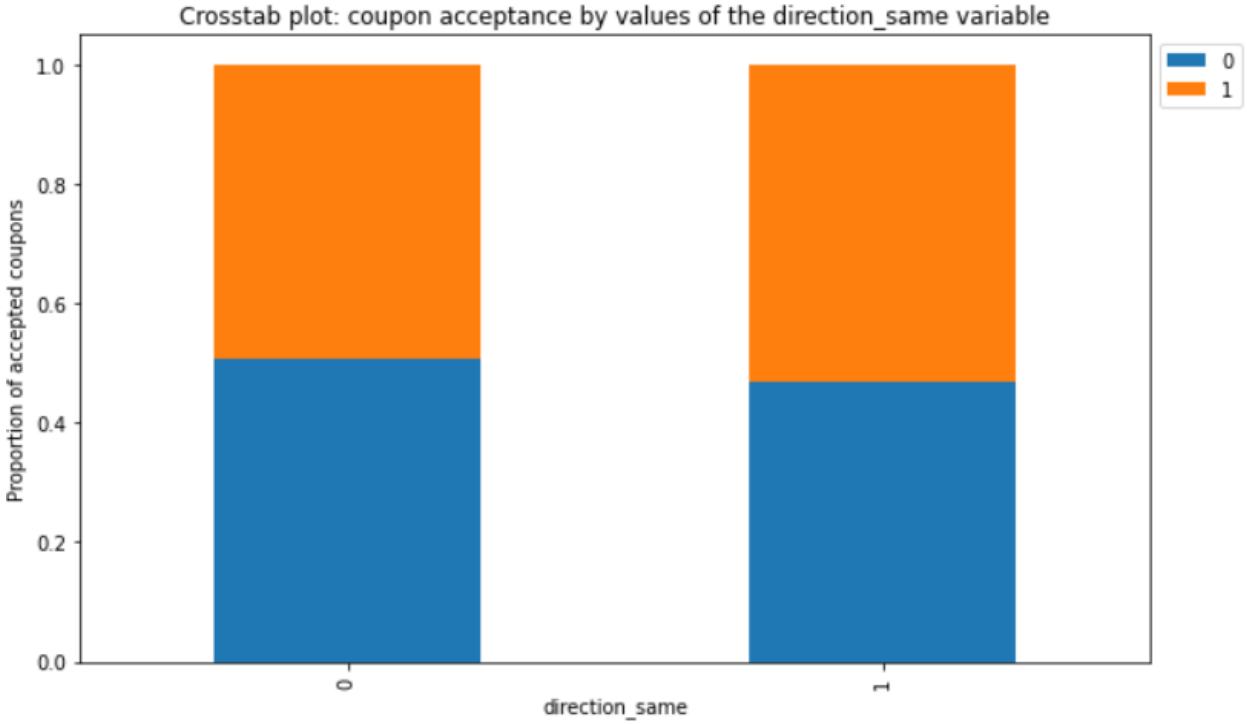
- There appears to be a negative relationship here. toCoupon_GEQ15min = 1 is associated with lower CoffeeHouse coupon acceptance rates.

GEQ25min



- There appears to be an even stronger negative relationship here. toCoupon_GEQ25min = 1 is associated with significantly lower CoffeeHouse coupon acceptance rates.

Direction_Same



- There might be a slight positive relationship here. direction_same = 1 is associated with a somewhat higher CoffeeHouse coupon acceptance rate, which is to be expected.

Statistical Analysis

The Rationale:

Statistical analysis is outside of the scope of this part of the course. Nevertheless, the prompt encouraged us to do a basic Correlation Matrix analysis.

However, when looking at the values of all of our variables, I noticed that all of them. are actually categorical (and not numerical) variables. For example "age" variable might seem numeric, but upon closer look, we see the following possible values for "age":

- '21', '26', '31', '36', '41', '46', '50plus', 'below21'

'50plus' and 'below21' cannot be converted to a number, since we do not know what the median (or mean) age for people in those buckets would be. Even the rest of the values seem like 5-year-increment buckets rather than numbers. So, we should treat these values as categories and not numbers.

"Temperature" is the only variable that contains only numbers, but the list of all possible values for "weather" is:

- 30, 55, 80 (these also seem like bucket IDs)

It seems unlikely that the temperature was equal only to one of these three values during the time the data was collected. Most likely, the temperature was sorted into approximate “cold,” “moderate,” and “warm” buckets.

Since we are dealing with Categorical (and not numerical) variables, we cannot do a correlation matrix analysis. In order to measure how each of our categorical features influences the outcome and in order to see how any pair of variables relate to each other, we will use Cramer's V association measure instead of the correlation.

The theory is described on this Wikipedia page.

- Cramer's v calculation relies on Chi-Squared statistics, so I had to import the SciPy Module.
- Just like the correlation matrix, the values range from 0 (no association) to 1 (perfect association).
- Cramer’s V matrix is also symmetrical with the diagonals equal to 1 (perfect association of each variable to itself)

Preparation steps:
1) Create a Cramer’s V function for each pair of variables (it takes into account the degrees of freedom)
2) Create another function that populates the matrix using the above function.
3) Create a function that draws a n Seaborn heatmap for Cramer’s V.

Application:

Since we can do quite a lot of analysis using just one Cramer’s V association matrix, I went ahead and applied it to ALL possible coupons (not just Coffee House).

In order to do so, I created copies of the data frame that were specific to particular coupons: “Bar,” “Carry out,” etc.

I was interested in learning whether each type of coupon would have different predictors. Indeed, this was the case.

As a result of the analysis, we could see how each of the variables is associated (interacts) with any other variable.

One of the variables happens to be “Y” (coupon acceptance), which is the desired outcome. I paid particular attention to what impacts that variable.

Summary of Statistical Analysis:

Please note that the graphics are easier to read within the Notebook file (especially on larger monitors). I provide them here mostly for illustration.

All (any kind of) coupons Variable Association Analysis:

Observations:

When it comes to the data on acceptance of all coupons (our main data frame), the following pairs of variables have the strongest association: (association of over 0.6 is considered "strong")

- Time and distance: 0.84
- Weather and Temperature (not surprisingly): 0.63

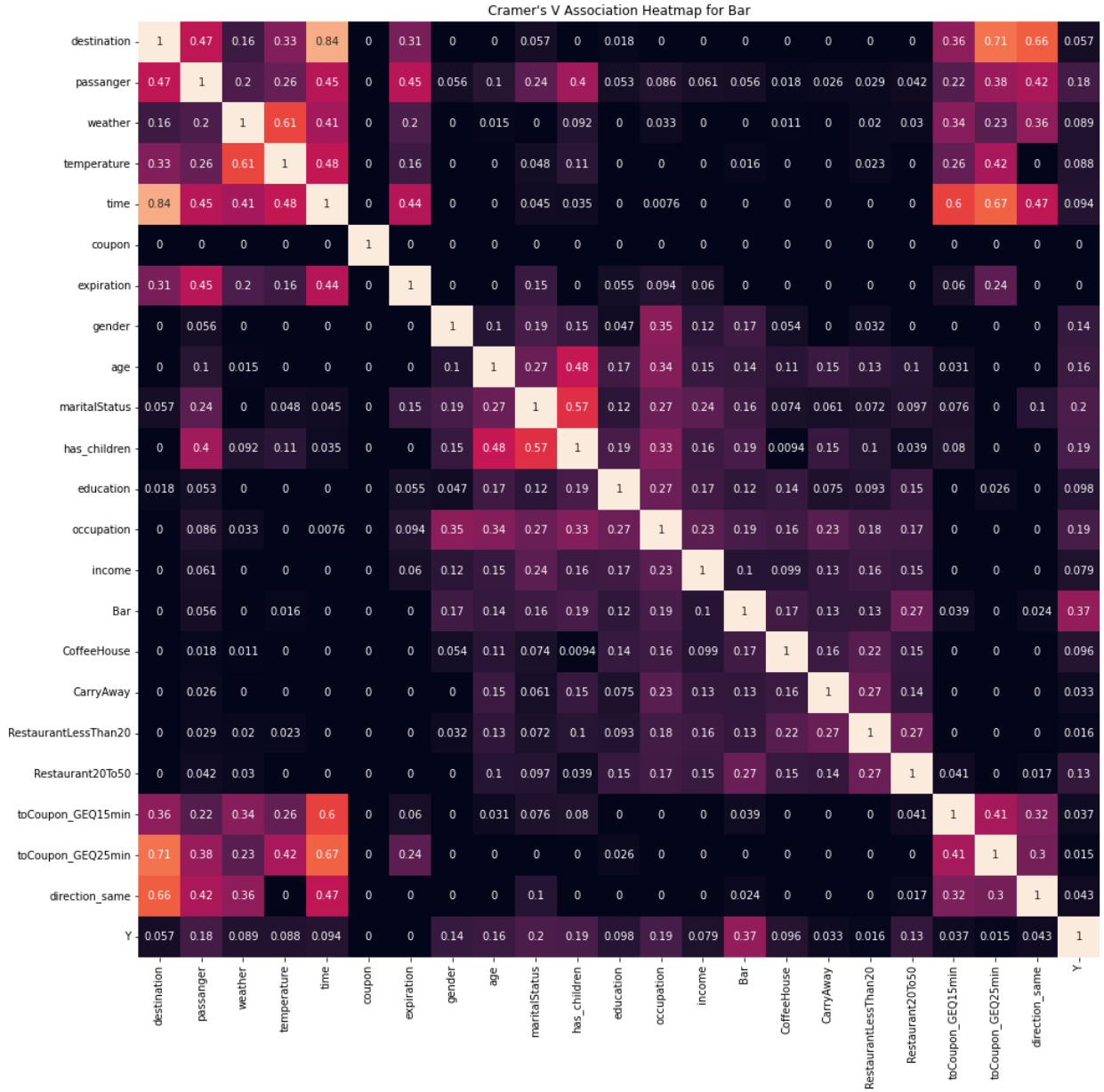
The following pairs of variables have "Decent" association (between 0.4 and 0.6)

- Marital Status and Has kids (not surprisingly, since Married people are expected to have more kids than singles): 0.57
- Passenger and Destination
- Direction Same and Destination
- Direction Opposite and Destination
- Age and has Children (not surprisingly since people who are two young are less likely to have children)
- to Coupon GEQ25 min and Destination

The strongest predictors (association) to accepting any coupons (the Y variable) are the following:

- Coupon
- Destination
- Passenger
- Coffee House

Bar Coupon variable association analysis:



Observations:

When it comes to the data on acceptance of Bar coupons, the following features have a strong association (over 0.6):

- Time and destination: 0.84
- ToCoupon_GEQ25 min and Destination: 0.71
- Weather and Temperature (not surprisingly): 0.61

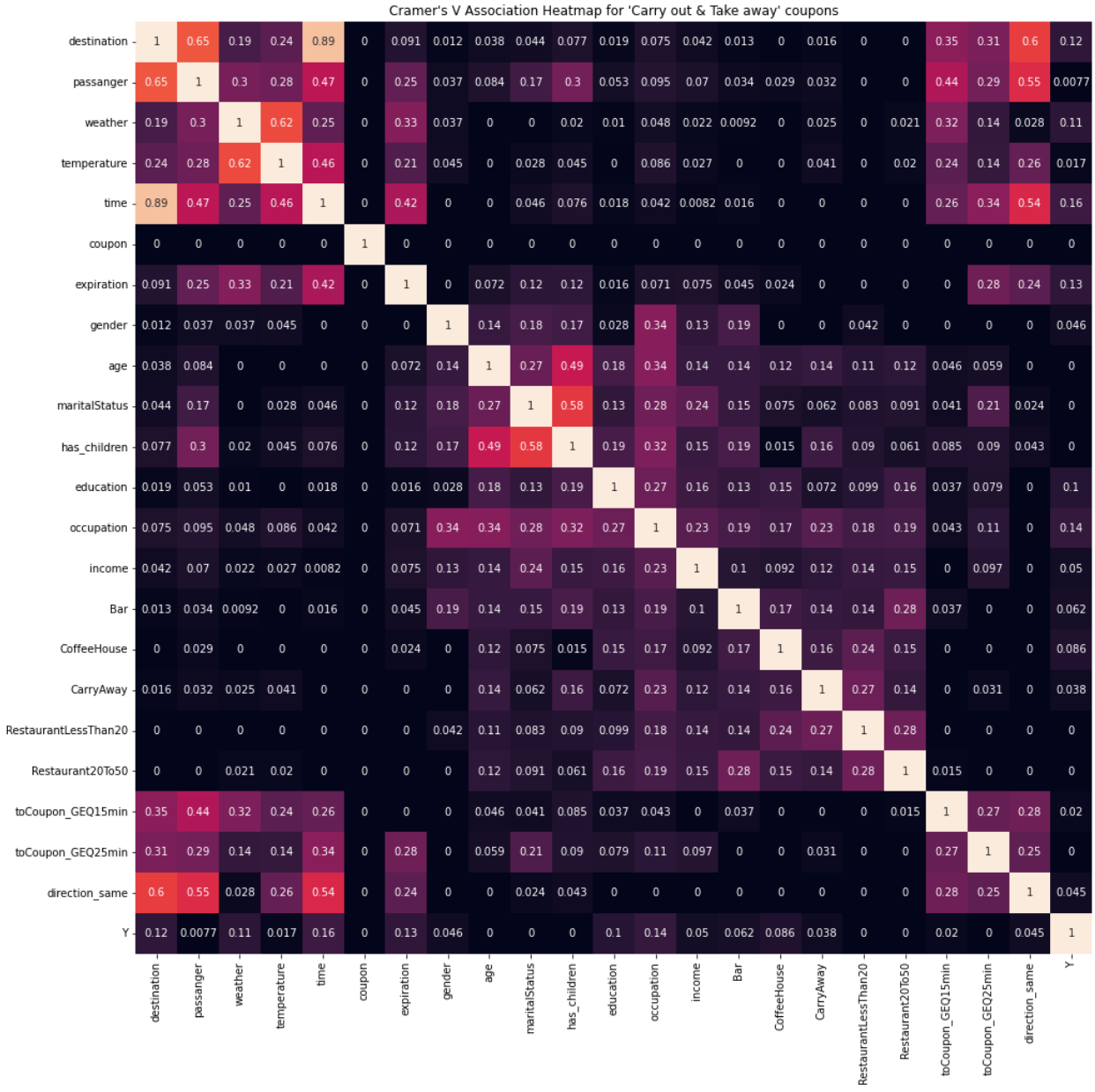
The following pairs of variables have "Decent" association (between 0.4 and 0.6)

- Marital Status and Has kids (not surprisingly since Married people are expected to have more kids than singles): 0.57
- Has Kids and Age
- Time and Temperature
- Time and Passengers

The strongest predictors of the acceptance of the Bar coupon (the Y variable) are:

- Bar (not surprisingly)
- Marital Status
- Has Children
- Occupation
- Passenger
- Age

'Carry out & Take away' coupon Association Analysis



Observations:

When it comes to the data on acceptance of Bar coupons, the following features have Strong association (over 0.6):

- Time and destination: 0.89

- Passenger and destination: 0.65
- Weather and Temperature (not surprisingly): 0.62

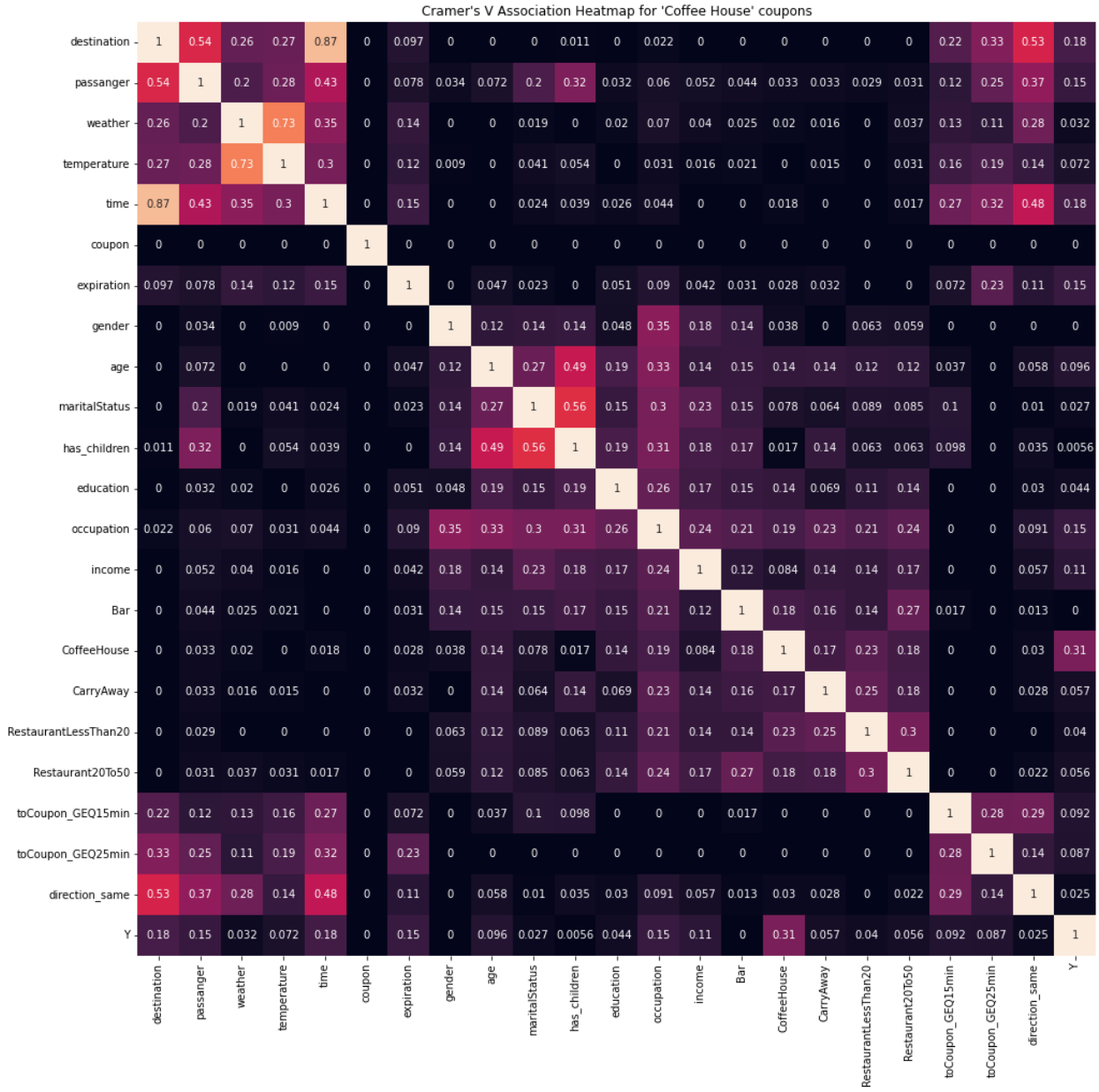
The following pairs of variables have "Decent" association (between 0.4 and 0.6)

- Marital Status and Has kids (not surprisingly since Married people are expected to have more kids than singles)
- Has Kids and Age
- Time and Passengers
- Time and Temperature

The strongest predictors of the acceptance of the Bar coupon (the Y variable) are:

- Time
- Occupation
- Expiration
- Destination
- WeatherMarital Status
- Education

Coffee House Coupon association analysis



Observations:

When it comes to the data on acceptance of Coffee House coupons, the following features have a strong association (over 0.6):

- Time and destination: 0.87
- Temperature and Weather: 0.73

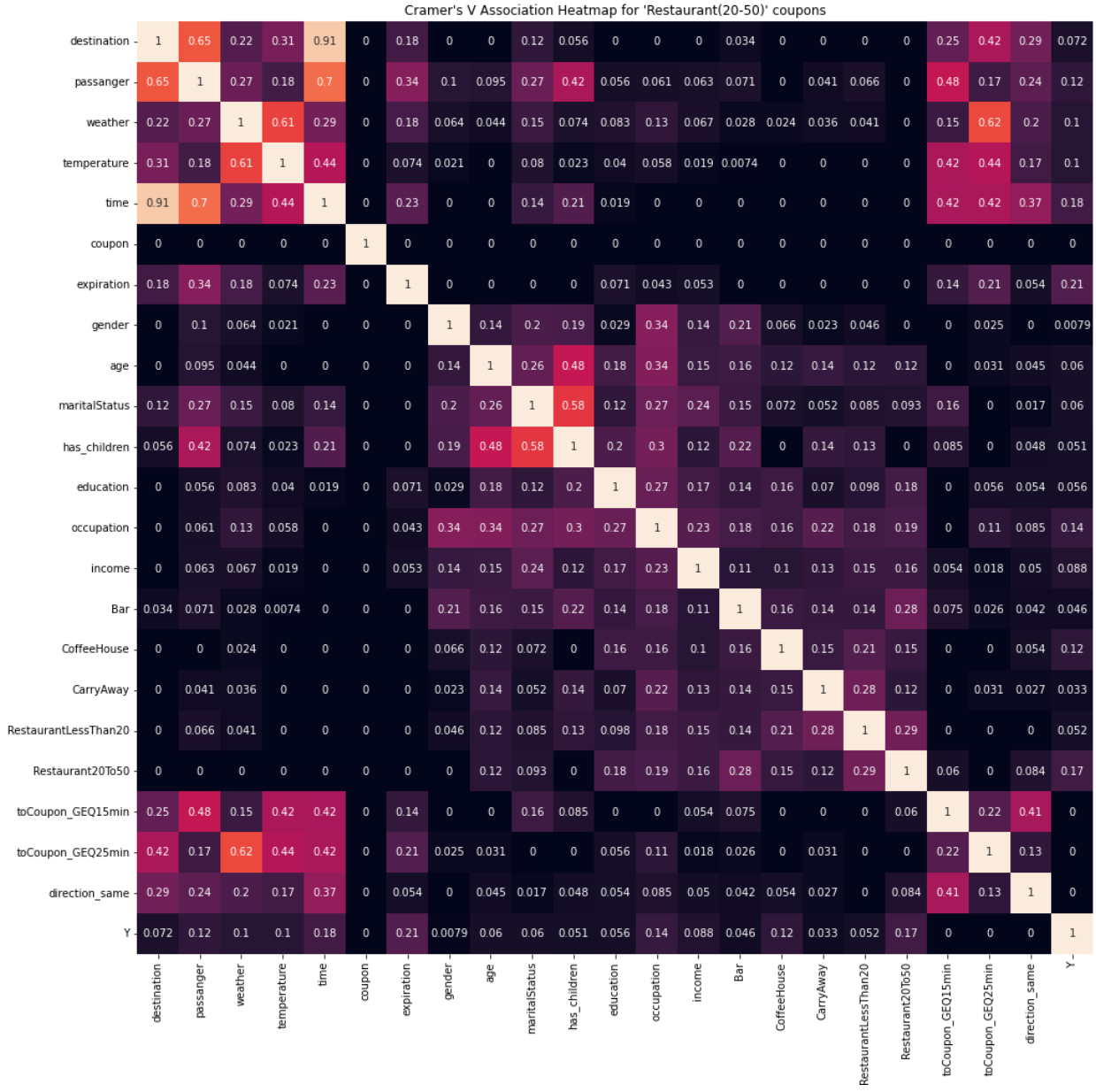
The following pairs of variables have a "Decent" association (between 0.4 and 0.6)

- Has Children and Marital Status
- Passenger and Destination
- Direction Same and Destination
- Has Children and Age
- Direction Opposite and Destination
- Time and Direction Same
- Time and Direction Opposite
- Time and passenger

The strongest predictors of the acceptance of the Bar coupon (the Y variable) are:

- Coffee House (was by far, the strongest variable, which is to be expected
- Destination
- Time
- Passenger
- Occupation
- Expiration These will be the variables that we should focus on when creating our Coffee house models.

'Restaurant(20-50)' Coupon Variable Association Analysis



Observations:

When it comes to the data on acceptance of Restaurants 20-50 coupons, the following features have a strong association (over 0.6):

- Time and Destination: 0.91
- Passenger and Time: 0.7
- Passenger and Destination: 0.65
- toCoupon_GEQ25min and Weather: 0.62
- Temperature and Weather: 0.61

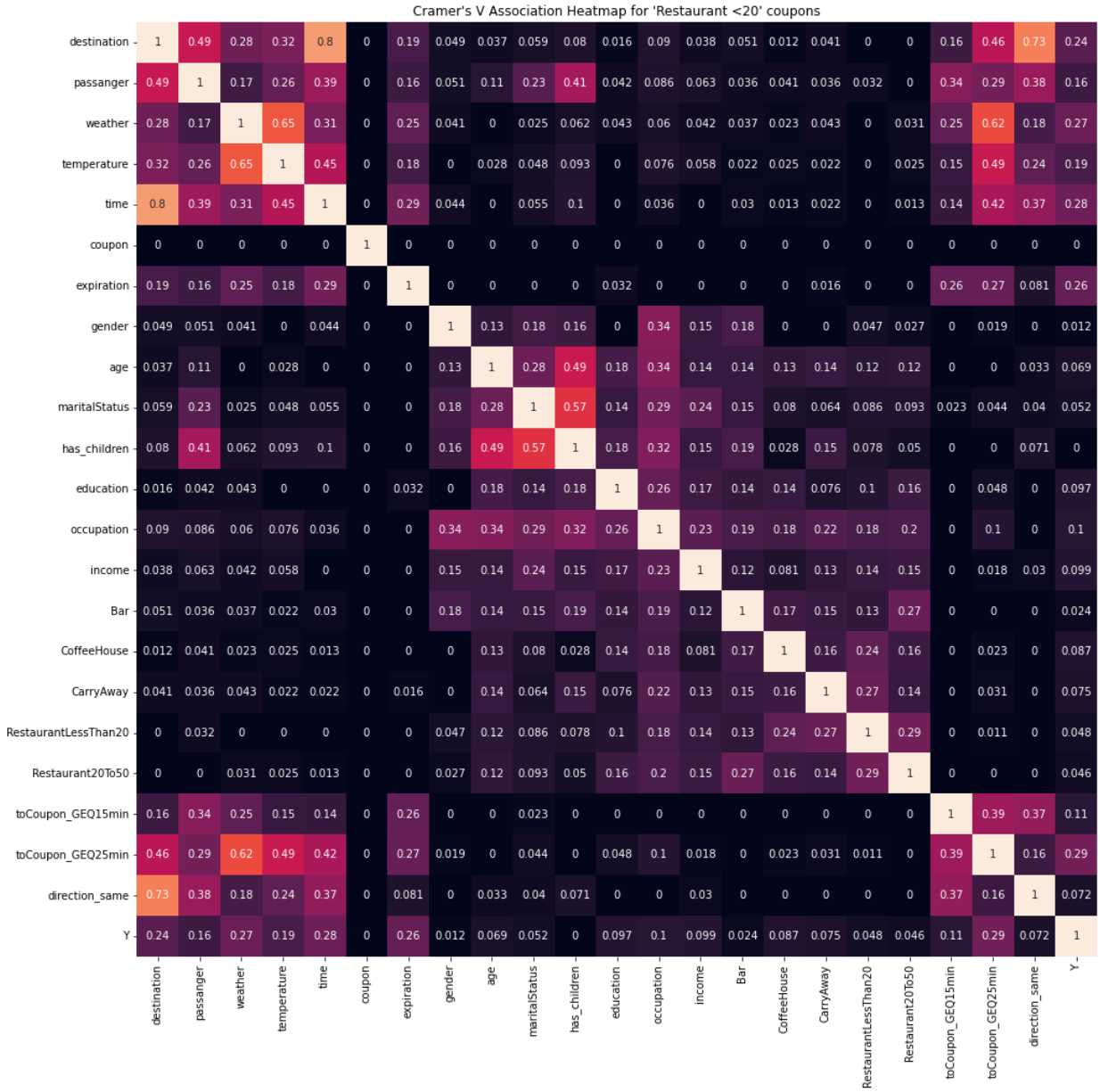
The following pairs of variables have "Decent" association (between 0.4 and 0.6)

- Has Children and Marital Status
- ToCoupon_GEQ15 min and Passenger
- ToCoupon_GEQ15 min and Temperature
- ToCoupon_GEQ15 min and Time
- ToCoupon_GEQ25 min and Temperature
- ToCoupon_GEQ25 min and Time
- ToCoupon_GEQ25 min and Destination
- Direction Same and ToCoupon_GEQ15
- Direction Opposite and ToCoupon_GEQ15
- Has Children and Passenger
- Passenger and Destination

The strongest predictors of the acceptance of the Bar coupon (the Y variable) are:

- Expiration
- Time
- Restaurant 20 to 50 attendance (not surprisingly)
- Occupation
- Coffee House attendance
- Passenger These will be the variables that will likely have significant contribution to the future model that predict the likelihood of accepting the Restaurants 20-50 coupons

'Restaurant(<20)' Coupon Variable Association Study



Observations:

When it comes to the data on acceptance of Restaurants 20-50 coupons, the following features have Strong association (over 0.6):

- Time and Destination: 0.8
- Direction Same and Destination: 0.73
- Direction Opposite and Destination: 0.73
- Temperature and Weather: 0.65 (as one would expect)
- ToCoupon_GEQ25 min and Weather: 0.62

The following pairs of variables have "Decent" association (between 0.4 and 0.6):

- Has Children and Marital Status
- Passenger and Destination
- Time and Temperature
- ToCoupon_GEQ25 min and Destination
- ToCoupon_GEQ25 min and Temperature
- ToCoupon_GEQ25 min and Time

- Has Childre and Passenger

The strongest predictors of the acceptance of the Bar coupon (the Y variable) are:

- ToCoupon_GEQ25
- Time
- Weather
- Expiration
- Destination
- Temperature
- Passenger
- ToCoupon_GEQ15

Those will likely be most useful (impactful) variables in the future models that predict acceptance of the cheap restaurant coupons.

General Observations from Statistical Analysis:

1. When it comes to the association to the "Y" variable (the coupon acceptance), the specialized data frames (those filtered to specific coupons), produce greater variable association as compared to the data frame that contains all coupons. This is not surprising, because both the users and the contextual variables that influence each type of coupon acceptance are different. For example, one is more likely to accept a Coffee shop coupon in the morning and a Bar coupon in the evening. So, when we create models for this data, it would be a good idea to model the acceptance of each type of coupon separately--as opposed to creating a generic model that predicts acceptance of any type of coupon.
2. I expect that there will be some interactions between variables. I.e., The effect of two or more variables could be different than the combined effect of each of them. Therefore, the best way to see what variables are most impactful is to go ahead and build the model.