A comprehensive report on:

“Will the Customer Accept the Coupon”

The dataset

Source of the dataset

The dataset for this study is obtained from UCI’s repository:
https://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation

Description of the columns in the dataset:

```
destination: No Urgent Place, Home, Work
passanger: Alone, Friend(s), Kid(s), Partner (who are the passengers
in the car)
weather: Sunny, Rainy, Snowy
temperature:55, 80, 30
time: 2PM, 10AM, 6PM, 7AM, 10PM
coupon: Restaurant(<$20), Coffee House, Carry out & Take away, Bar,
Restaurant($20-$50)
expiration: 1d, 2h (the coupon expires in 1 day or in 2 hours)
gender: Female, Male
age: 21, 46, 26, 31, 41, 50plus, 36, below21
maritalStatus: Unmarried partner, Single, Married partner, Divorced,
Widowed
has_Children:1, 0
education: Some college - no degree, Bachelors degree, Associates
degree, High School Graduate, Graduate degree (Masters or Doctorate),
Some High School
occupation: Unemployed, Architecture & Engineering, Student,
Education&Training&Library, Healthcare Support,
Healthcare Practitioners & Technical, Sales & Related, Management,
Arts Design Entertainment Sports & Media, Computer & Mathematical,
Life Physical Social Science, Personal Care & Service,
Community & Social Services, Office & Administrative Support,
Construction & Extraction, Legal, Retired,
Installation Maintenance & Repair, Transportation & Material Moving,
Business & Financial, Protective Service,
```

```
Food Preparation & Serving Related, Production Occupations,
Building & Grounds Cleaning & Maintenance, Farming Fishing & Forestry
income: $37500 - $49999, $62500 - $74999, $12500 - $24999, $75000 -
$87499,
$50000 - $62499, $25000 - $37499, $100000 or More, $87500 - $99999,
Less than $12500
Bar: never, less1, 1~3, gt8, nan4~8 (feature meaning: how many times
do you go to a bar every month?)
CoffeeHouse: never, less1, 4~8, 1~3, gt8, nan (feature meaning: how
many times do you go to a coffeehouse every month?)
CarryAway:n4~8, 1~3, gt8, less1, never (feature meaning: how many
times do you get take-away food every month?)
RestaurantLessThan20: 4~8, 1~3, less1, gt8, never (feature meaning:
how many times do you go to a restaurant with an average expense per
person of less than $20 every month?)
Restaurant20To50: 1~3, less1, never, gt8, 4~8, nan (feature meaning:
how many times do you go to a restaurant with average expense per
person of $20 - $50 every month?)
toCoupon_GEQ15min:0,1 (feature meaning: driving distance to the
restaurant/bar for using the coupon is greater than 15 minutes)
toCoupon_GEQ25min:0, 1 (feature meaning: driving distance to the
restaurant/bar for using the coupon is greater than 25 minutes)
direction_same:0, 1 (feature meaning: whether the restaurant/bar is
in the same direction as your current destination)
direction_opp:1, 0 (feature meaning: whether the restaurant/bar is in
the same direction as your current destination)
Y:1, 0 (whether the coupon is accepted)
```

Preliminary insights into the dataset:

Most of the columns are categorical in nature. Although, some columns have int64 as the data-type, they are still categorical, as the values in them are not continuous (ex: `toCoupon_GEQ15min` is defined as int64, but it contains only 0 or 1)

There are 12684 records in the dataset.

Data cleaning

The first and foremost step in an ML exercise is to ‘clean’ the data, otherwise it will have undesirable results in our analysis.

Handling the null values

As mentioned, there are 12684 records In this dataset. Most of the columns have all the records populated. However, the below columns have some null values. The below table shows how many rows are populated with non-null values for these subset of columns. The table also shows the method chosen by this report to handle the null values in the respective columns.

Column	Non-null count	Method chosen to clean the null entries
car	108	Out of 12684 entries, only 108 entries are populated. This column is almost useless for our analysis. I chose to populate replace nulls with string 'unknown' and move on.
Bar	12577	Most rows are populated. Only 107 (ie, <1%) are not populated. I chose to replace them with most frequent # answer in the column (which is 'never')
CoffeeHouse	12467	Again, most rows are populated. However, unlike 'Bar' column, there's no clear dominant answer (ie, value_counts are close for many choices). I chose to use bfill method which picks up the next valid observation
CarryAway	12533	Same method as 'CoffeeHouse' column
RestaurantLessThan20	12554	Same method as 'CoffeeHouse' column
Restaurant20To50	12495	Same method as 'CoffeeHouse' column

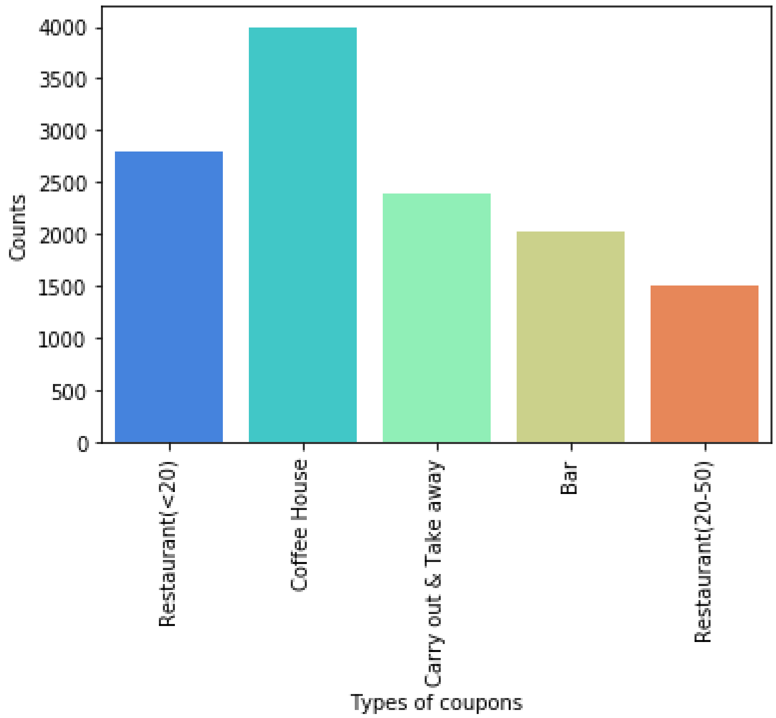
Data analysis

This section contains a series of questions and answers to them in the form of visualizations and descriptions. The study's conclusions will be drawn from these questions and answers.

Preliminary questions

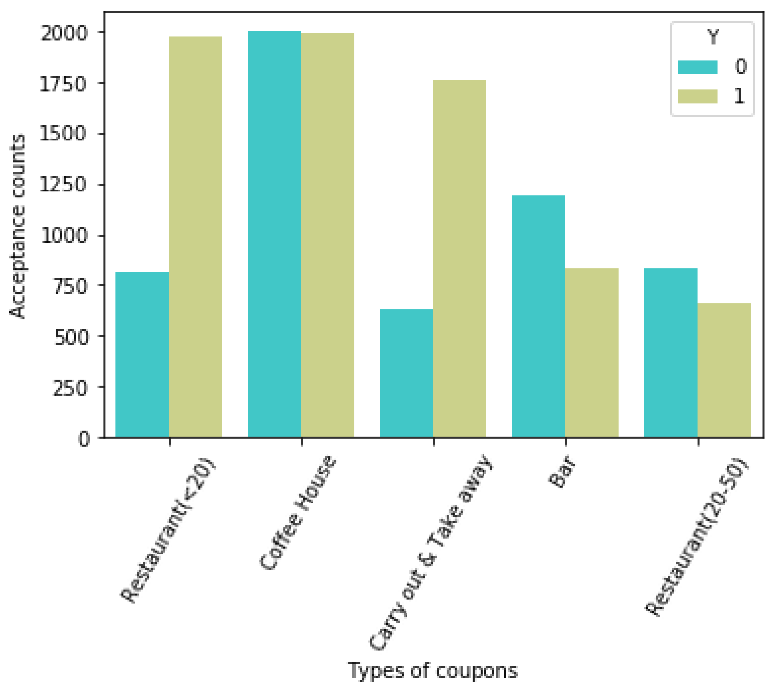
```
Question:
How many types of coupons are there and what are the counts of each of those
coupon types?
```

A countplot is best suited to answer this question



`Question: How many coupons were accepted by the drivers in each of these coupon types`

Again, a countplot would be the best way to see this with a hue on column ‘Y’.



In general, the coupon acceptance rate is good for less-expensive restaurants and ‘Carry out & Take away’ places.
The acceptance and denial rates for Coffee Houses is about the same; may-be, it is too much work to go use a coupon for just a coffee.
The acceptance rate for Bar and the more expensive restaurants is poor, suggesting that these customers are probably not in the coupon-using-state-of-mind when going out.

Again, this is a high level analysis. We will ask more questions and find answers for some of these coupon types below and see if the pattern changes.

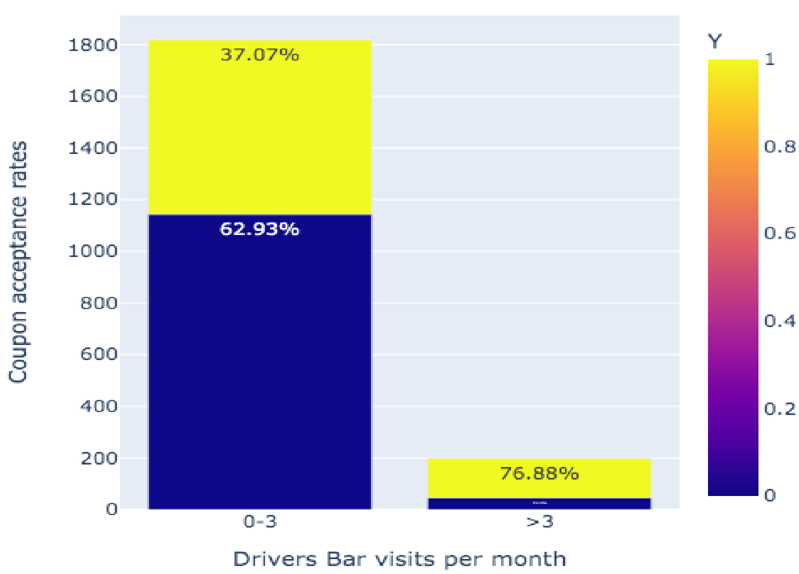
Investigating the Bar Coupons

In the previous section, we noticed that the Bar coupons have relatively poor acceptance rate. Let’s ask some questions to see what type of Bar customers had accepted the Bar coupons better than the others. This will help to focus business of these type of customers than everyone.

```
Question: Compare the acceptance rate between those who went to a bar 3 or fewer times a month to those who went more.
```

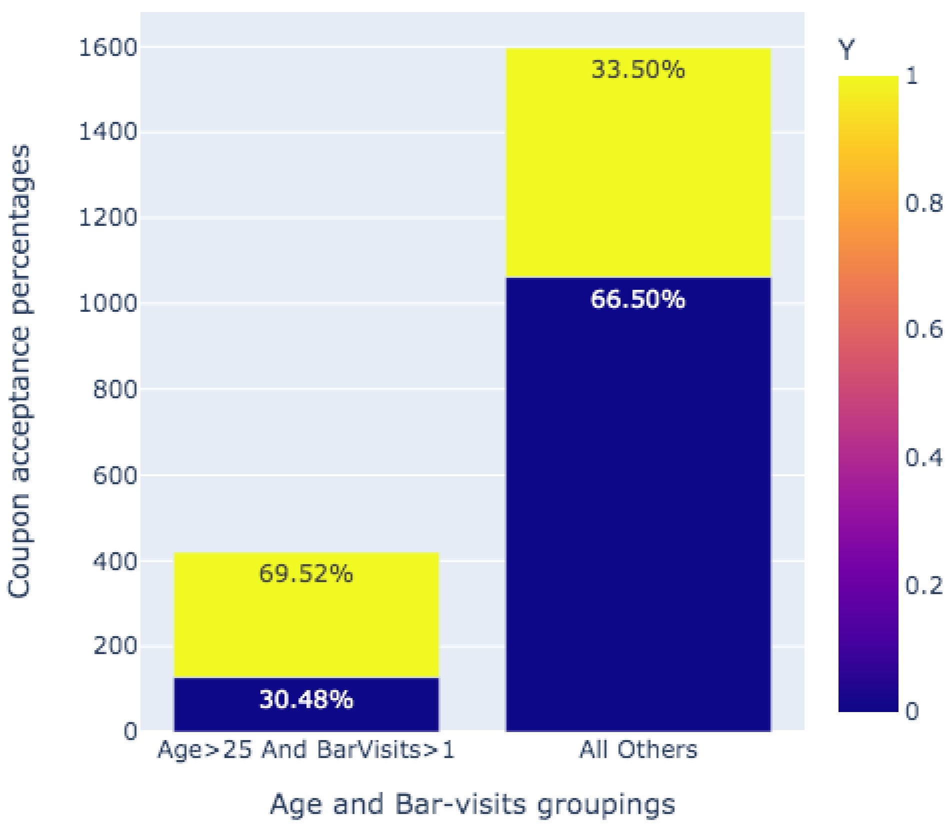
To answer this, we will again use a countplot but this time, we will also spit out the percentages to make it easier.

The Bar coupon acceptance rate is ~77% in drivers who went to a bar more than 3 times a month while it is only ~37% in drivers who went 3 or fewer times to a bar a month. This is a great insight for the business to focus on this subset.



`Question: Compare the acceptance rate between drivers who go to a bar more than once a month and are over the age of 25 to the all others. Is there a difference?`

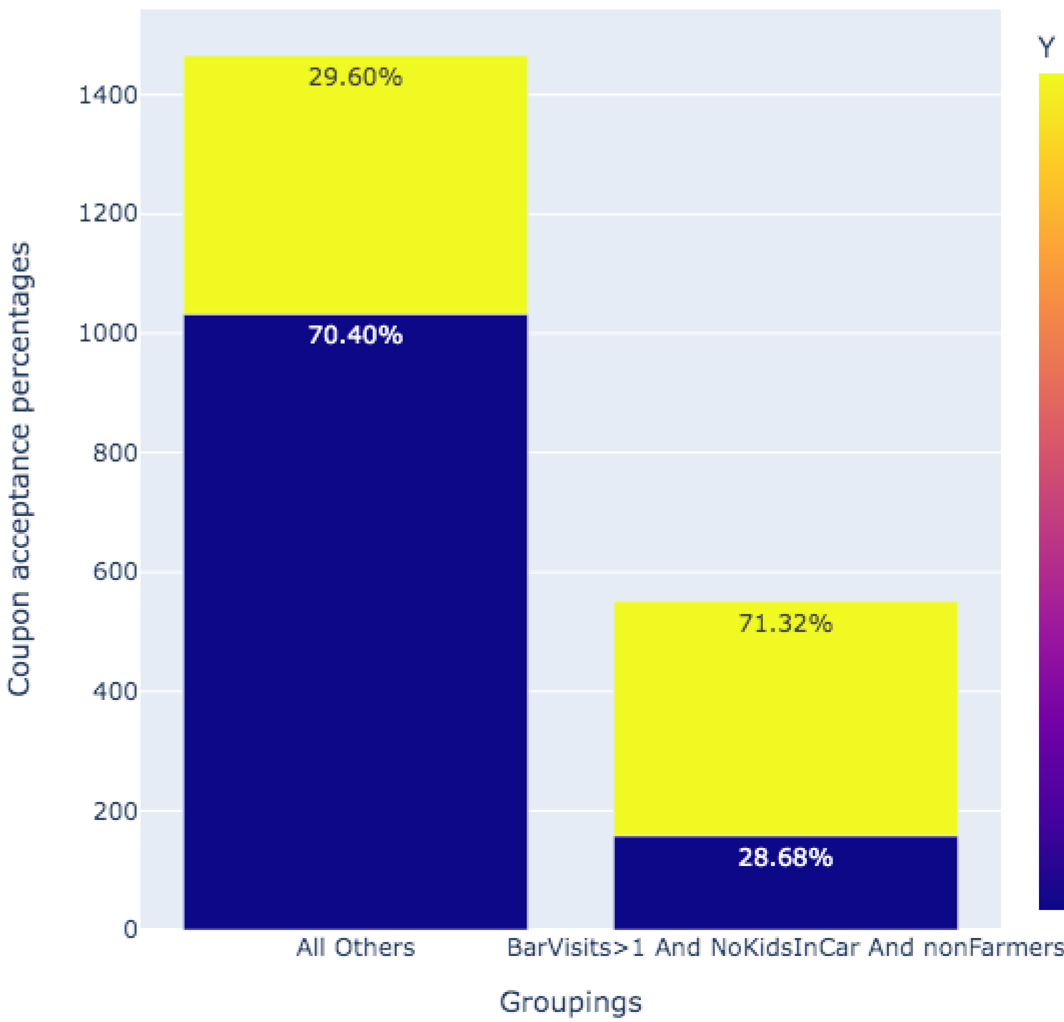
Sure, we will answer the question using a similar visualization.



Yes, there is a glaring difference. **There's a ~70% Bar coupon acceptance rate in drivers who are more than 25 years old and typically visit a bar more than once a month while it is only ~33.5% in rest of the drivers who received Bar coupons**

Question: Compare the acceptance rate between drivers who go to bars more than once a month and had passengers that were not a kid and had occupations other than farming, fishing, or forestry

Here's how these acceptance rates look.

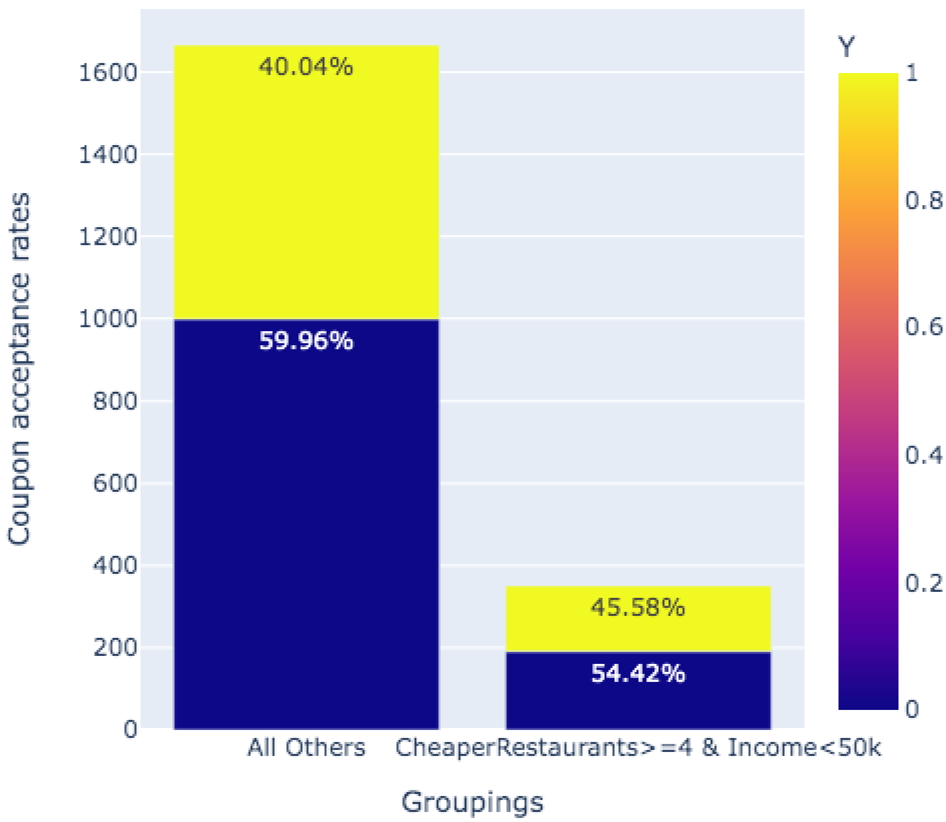


The Bar coupon acceptance rate in the subset of drivers requested is ~71% while it is < 30% in the rest of the drivers.

Let’s now see if the income limit has any insight in bar coupon acceptance:

`Question: Compare the acceptance rates between those drivers who go to cheaper restaurants 4 or more times a month and income is less than 50K.`

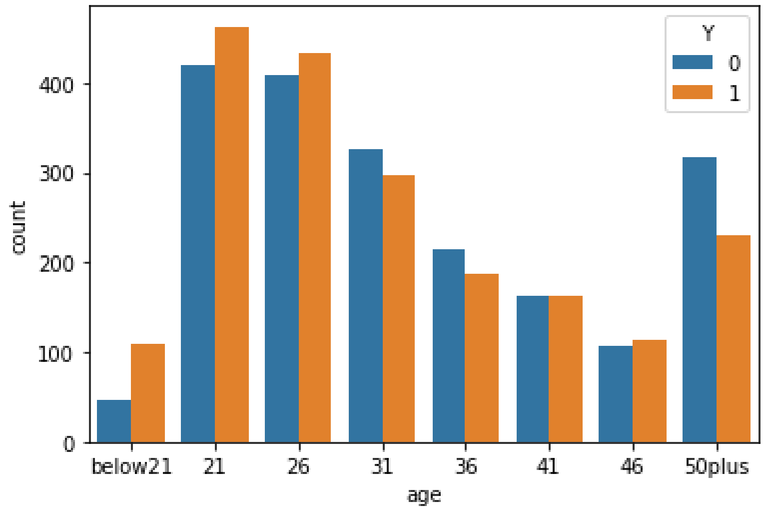
Let’s see the visual for this:



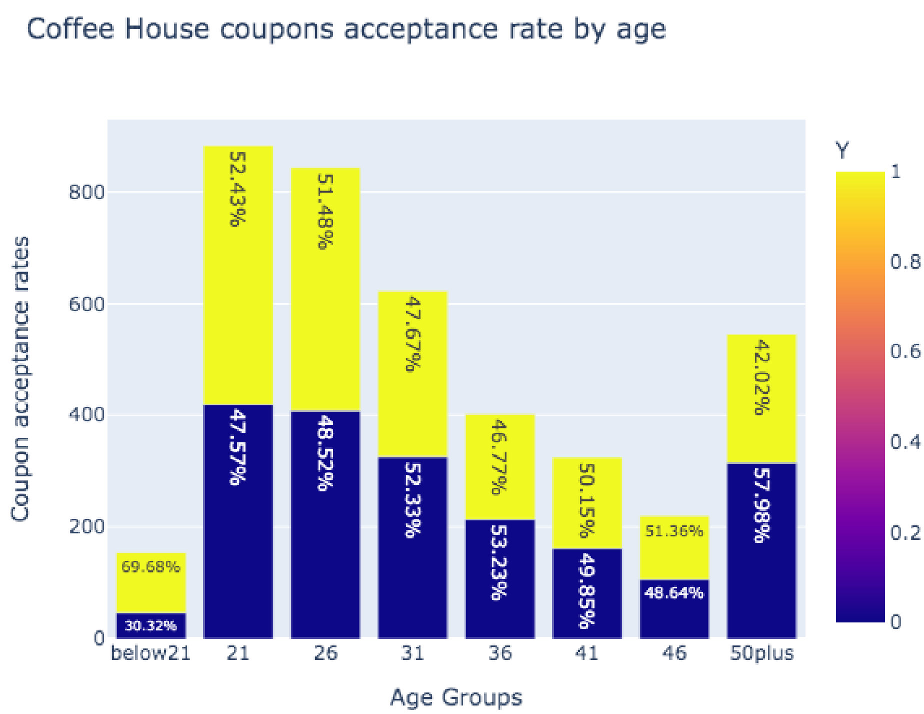
The data shows that the acceptance rate for the subset of drivers in question is not that different from the rest of the drivers. Little surprising insight but it is what it is!!!

Now, let’s try to peek into CoffeeHouse data and try to analyze the almost-equal acceptance-and-denial rate of this coupon type.

`Question: What are the CoffeeHouse coupons acceptance rate by age`
Let’s see a bar chart by age with hue of acceptance

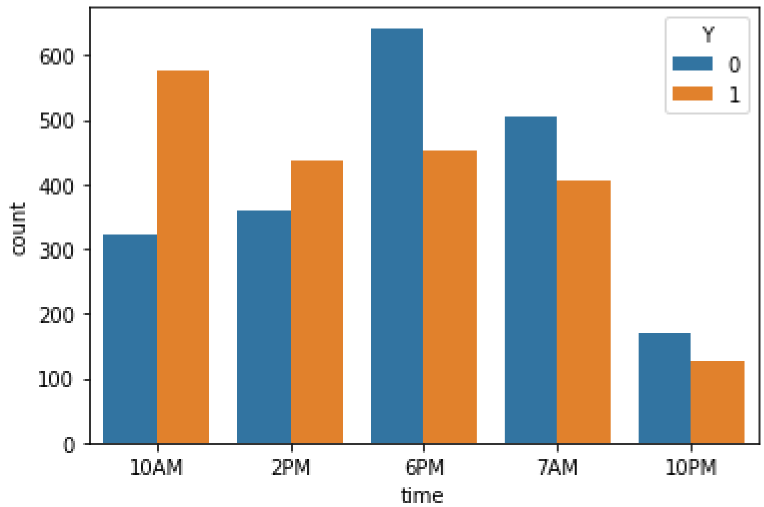


Also, we can see by acceptance percentages:

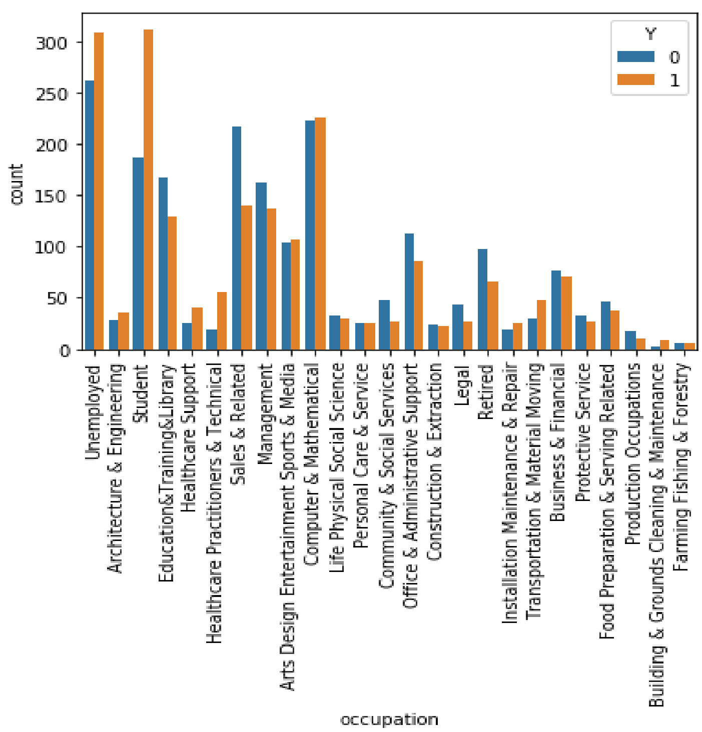


Seems like the acceptance rate is high with drivers below 21 years. Also, the acceptance rate seems to be going down with age, with a slight anomaly from 41 to 46 (very minor, that we can ignore the anomaly). So, in general, we can say that the Coffee House coupons are popular in younger age drivers.

`Question: Does the time of day affect Coffee house coupons?`
Seems like it does. They are more popular in the morning around 10AM. The acceptance rate goes down as the sun hits the sky. See below:



`Question: Which professions accept more Coffee House coupons`



Looks like the acceptance rate of this coupon type is high in Unemployed and Students.

Summary

In general, the coupon acceptance rate is good for less-expensive restaurants and 'Carry out & Take away' places
There is a decent probability (77%) of a driver who goes to a bar 3 or more times in a month, to accept a Bar coupon
There's a ~70% Bar coupon acceptance rate in drivers who are more than 25 years old and typically visit a bar more than once a month while it is only ~33.5% in rest of the drivers who received Bar coupons
The acceptance rate for the subset of drivers who go to cheaper restaurants 4 or more times a month and income is less than 50K is not that different from the rest of the drivers. Little surprising insight but it is what it is!!!
Coffee house coupons are popular with younger age drivers and the acceptance rate is going down as the age increases
Coffee house coupons are more popular in the morning around 10AM (probably because of the rush hour traffic)
Coffee house coupons acceptance rate is high among Students and Unemployed

There are many more insights that can be extracted from different coupon types.

One short-coming I found in this dataset is, all the columns have categorical data. There's only so much we can do with categorical data. With continuous data, we could have plotted distributions, scatter plots, etc that would give even more insights.