Coupon Survey EDA Report

By

Stephen Woods

MSDS 430: Python for Data Science

Introduction

This project engages in exploratory data analysis of survey data where respondents are asked if they would accept a coupon under various driving conditions. The goal of this survey was to capture human behaviors related to an in car mobile recommendation system. The survey gives each respondent hypothetical driving scenarios with changing variables like destination, weather, time of day, and type of coupon. The respondent provides their socio demographic information (age, marital status, income), how often they frequent the coupon destination types (restaurants, bars, coffeeshops, and take out) and states whether they would accept the coupon for each scenario. This survey data was sourced from the UCI Machine Learning Repository and was collected by Amazon Mechanical Turk. Amazon Mechanical Turk is a service where 'turkers' are paid to answer survey data. The survey data analyzed was not weighted and contains all responses from turkers who chose to take the survey.

The goal of this project is to identify the traits and behaviors most associated with the decision to accept the coupon. With twelve scenario variables and thirteen demographic and behavioral data columns, we should be able to isolate the traits and behaviors most associated with the coupon decision. My hypothesis is that intuitive logic will hold true in analyzing this data. If the driver has no urgent destination, that driver might be more likely to accept and redeem a coupon. If the coupon is in the same direction as the destination, that driver might be more likely to accept the coupon. One concern with testing my hypothesis was that this survey is not offering the same type of coupon (IE bar, restaurant, takeout) for each scenario. I break down the survey data to account for this so my result is not biased. I am also interested in deriving social insights from this data. If the driver commonly goes out to get food and beverages, is that driver more likely to accept the coupon? With five columns on how frequently a respondent visits bars, restaurants, and so on, can these columns be grouped together and compared with the coupon decision?

Data Preparation and Analysis

This section covers basic steps taken to clean, modify, and interpret this coupon survey data. This data originally contained 12,684 rows with 26 columns. There were only 605 rows with null values in my data set, which I dropped as they did not represent a significant chunk of the data. I also dropped 22 rows where a respondent indicated they do not drive. This survey aims to collect data on when a driver would accept a coupon from an in-vehicle recommendation system, so results from a respondent that does not drive do not belong. I ultimately dropped 10 columns and added one calculated column through my EDA process. I did not identify any outliers within the data set as respondents filled in the survey by selecting responses from a list of available responses. For example, the "Bar" column questions the respondent how many times they go to a bar per month. If they go to a bar more than eight times, they select "gt8" (greater than eight). Table 1 below is a data dictionary, providing descriptions for each field I determined relevant to the coupon decision.

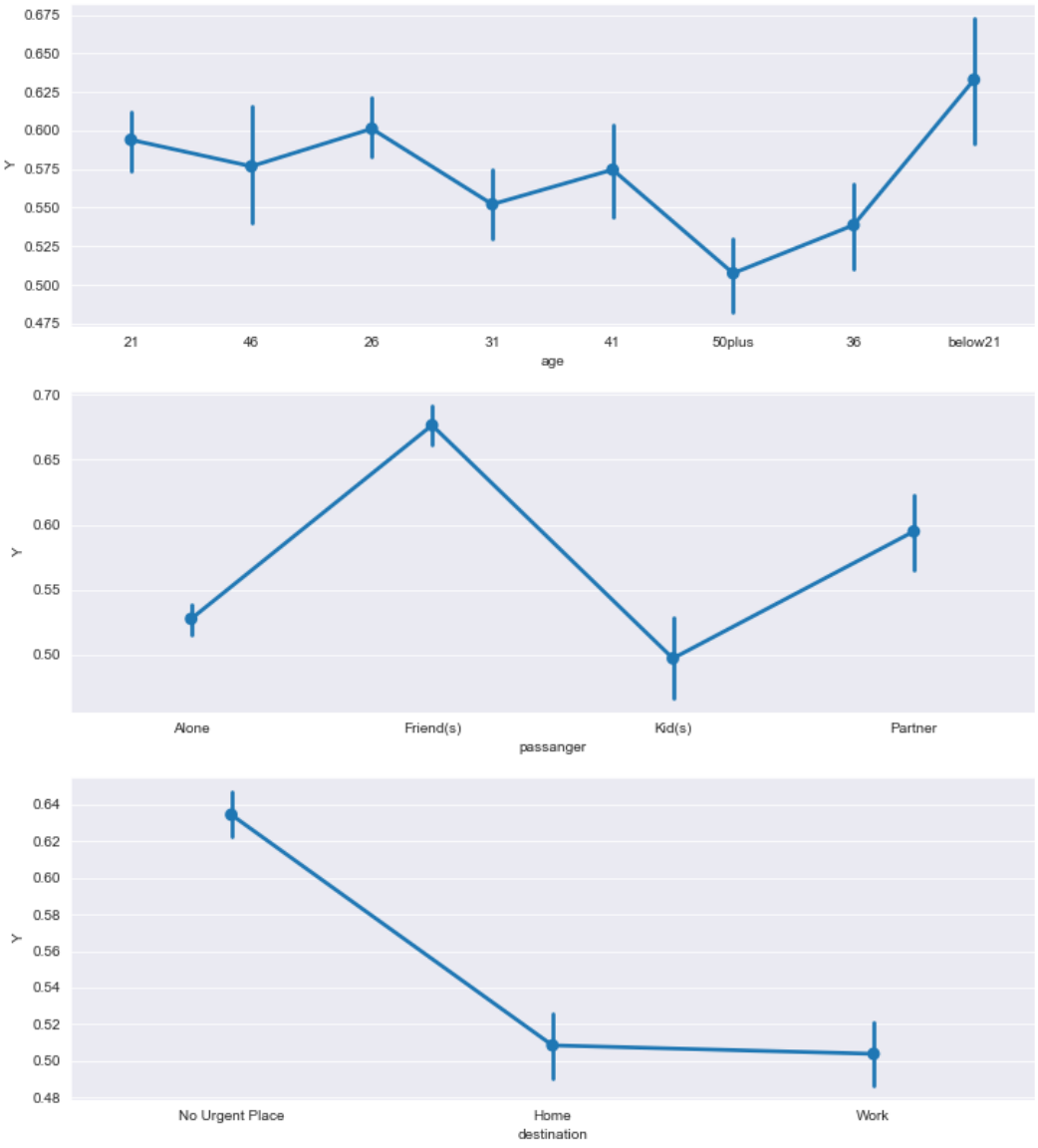
Table 1: Data Dictionary from Coupon Survey

Column	Description	Sample Data
destination	destination type	No Urgent Place
passanger	passenger type	Alone
weather	weather conditions	Sunny
time	time	2PM
coupon	coupon type	Bar
expiration	time before coupon expires	1d
age	age	21
maritalStatus	marrital status	Single
education	education level	Bachelors degree
Bar	how many times do you go to a bar every month	never
CoffeeHouse	how many times do you go to a cofeehouse every month	less1
CarryAway	how many times do you get takeout every month	4~8
RestaurantLessThan20	... restaurant with an average expense per person of less than $20 every month?	4~8
Restaurant20To50	... restaurant with average expense per person of $20 - $50 every month?	less1
toCoupon_GEQ25min	driving distance to redeem the coupon is greater than 25 minutes	0
Y	whether the coupon is accepted	1
coupon_type_match	Added column: FALSE if the respondant never goes to the coupon destination type.	TRUE

With 26 original columns, I only kept columns with responses that demonstrated at least a 10% difference in acceptance rate. For example, the gender column showed that males accepted the coupon

59% of the time and Females accepted 55% of the time, so that column was dropped. Figure 1 below features a point plot showing the acceptance rates from the age, passenger, and destination columns.

Figure 1: Acceptance Rate Point Plot



Respondents below 21 were most likely to accept the coupon and respondents above 50 were least likely to accept the coupon. Similarly, a respondent would accept the coupon 68% of the time if driving with friends in the car, and 64% of the time if there was no urgent destination.

After identifying which columns showed the strongest relationships to the coupon decision, I looked to add columns to break down the coupon decision drivers even more. There are five different coupons being offered in this data set, with the acceptance rate and count of how many times each coupon was offered shown below in Table 2.

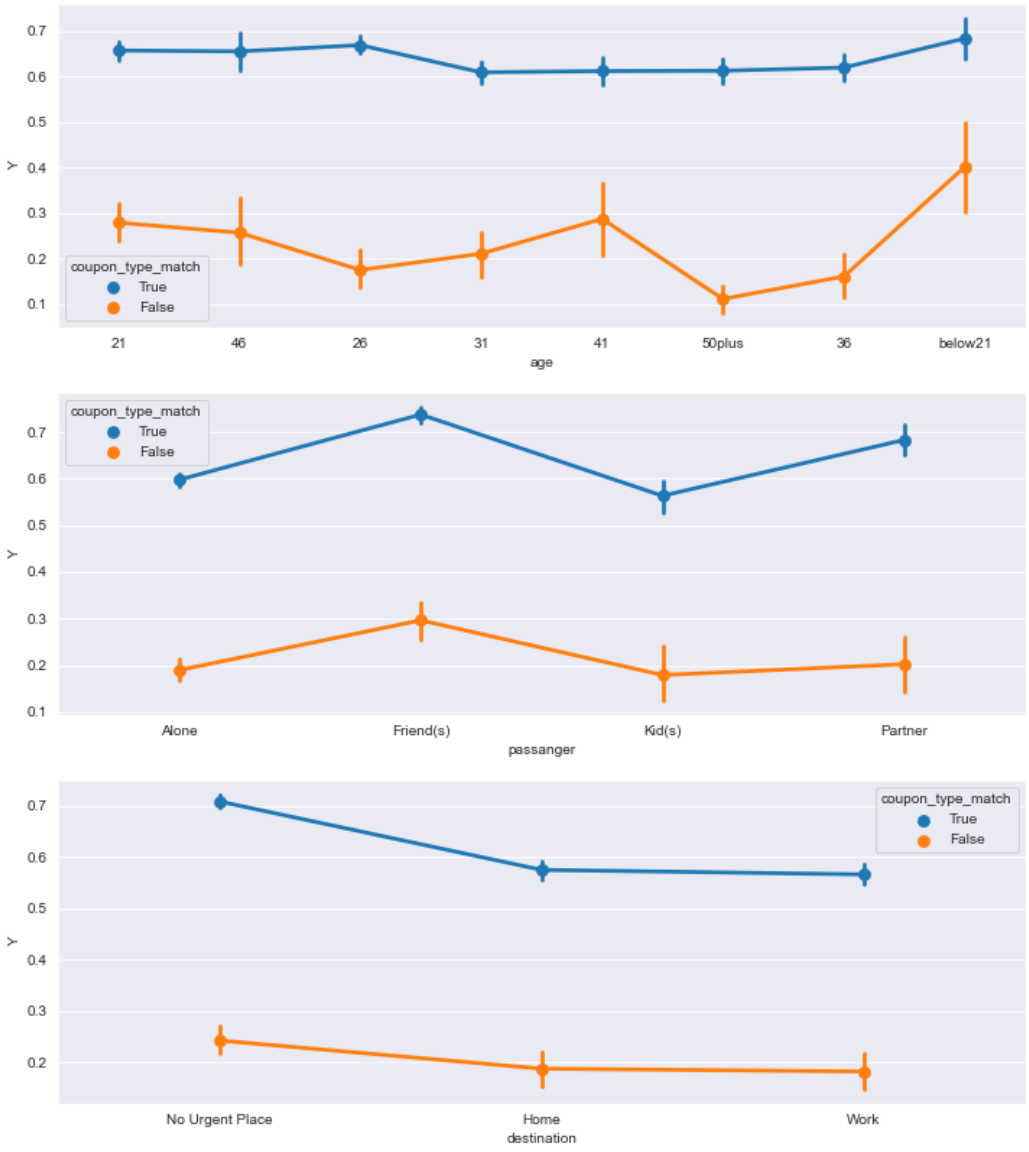
Table 2: Acceptance Rate and Count by Coupon Type

Coupon Type	Acceptance Rate	Count
Carry out & Take away	74%	2276
Restaurant(<20)	71%	2648
Coffee House	50%	3809
Restaurant(20-50)	45%	1413
Bar	41%	1911

Table 2 shows that some coupons have higher acceptance rates than others and some are offered more frequently than others, potentially biasing calculated summaries of acceptance rates for other columns. For example, if a coupon for a bar is being offered and we are looking at the acceptance rates of another column, we can filter out those who indicated they never go to bars to increase the acceptance rate. I added a true/false column that checks whether the respondent goes to the destination that the coupon is being offered for at all to account for this aspect of my data. For example, if the coupon type in is *Bar* and the respondent selected that they *never* go to bars in the Bar column, my calculated column will

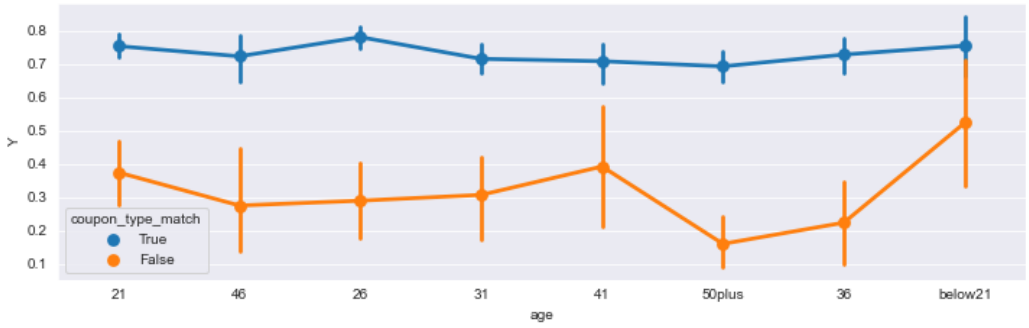
show false for that row. Figure 2 below features a point plot on the same columns as Figure 1 but broken out by the calculated coupon_type_match column.

Figure 2: Acceptance Rate Point Plot by coupon_type_match



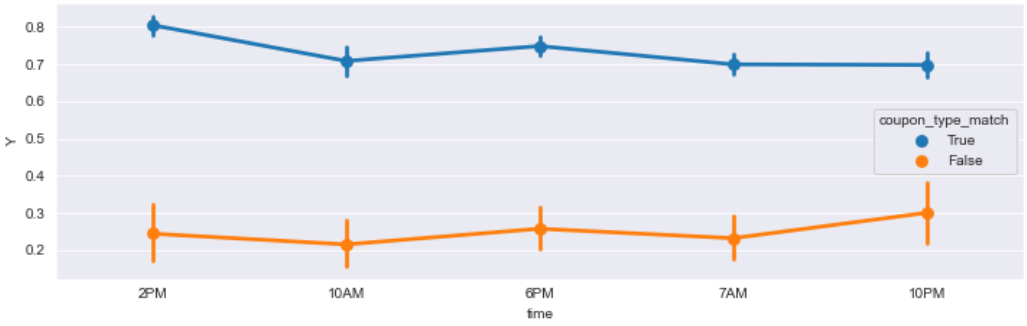
With this additional filter, we see respondents are much less likely to accept the coupon if they never go to the coupon destination across all three plots. For the age plot, we see lower differences in acceptance rates between age groups where the coupon type match is true. We see much larger differences between age groups where respondents never go to the coupon destination (coupon type match of false), potentially reflecting that younger respondents are more likely to try something new or different. Differences between acceptance rates in the passenger and destination columns were more uniform when broken out by the coupon type match field. Those traveling with friends in the car and those with no urgent destination who go to the coupon destination at least once per month accepted the coupon 73% and 71% of the time. Taking this one step further, Figure three below shows that if we filter for those with no urgent destination and friends in the car, all age groups that go to the coupon destination at least once per month have at least a 70% acceptance rate.

Figure 3: Acceptance Rate Point Plot: Driving with Friends + No Urgent Destination by Age



Note the error bars for the group who never frequent the destination type are significantly larger as fewer respondents fell into this category. Again, differences in acceptance rates are strongest between those 50 plus and below 21 in this sub group that do not go to the coupon destination being offered. Figure four below filters for environmental factors associated with the coupon being accepted.

Figure 4: Acceptance Rate Point Plot: Sunny Weather + Coupon Expires in One Day by Time



If it is sunny outside and the coupon expires in one day (vs two hours), respondents were over 80% likely to accept the coupon at 2PM.

Conclusion

Overall, the coupon was accepted 57% of the time. This analysis identified individual demographic traits and social/environmental factors that had higher rates of accepting the coupon. These individual factors were then combined to identify groups of respondents that accepted the coupon between 70% and 85% of the time. My hunch that the scenarios and traits that seemed intuitively inducive to accepting a coupon held true. Those driving with friends to no urgent destination that frequent the coupon destination at least once per month were very likely to accept the coupon. Similarly, high acceptance rates were shown if it was sunny outside and the respondent had more time to redeem the coupon. Further, younger respondents were more likely than older respondents to accept the coupon, significantly so among the group that never frequents the coupon destination. More filters could be applied to the other columns identified as relevant to the coupon decision to create more subgroups. Another goal I had was to identify social insights from this data set but was unable to do so. I created another calculated column that added up a total social score from the five columns that showed how frequently respondents visited a coupon destination. This numeric social score column had no significant relationship with the coupon decision. One next step for this project is to build a machine learning model around the data deemed highly relevant to predict when the coupon will be accepted.

Citations

Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. 'A bayesian framework for learning rule sets for interpretable classification.' The Journal of Machine Learning Research 18, no. 1 (2017): 2357-2393.