SMU Data Science Review

Volume 1 | Number 1 Article 14

2018

Consumer Welfare and Price Discrimination: A Fine Line

Marie Wallmark
Southern Methodist University, mwallmark@smu.edu

Eyal Greenberg
Southern Methodist University, eyalg@smu.edu

Dan Engels
Southern Methodist University, dwe@smu.edu

Consumer Welfare and Price Discrimination: A Fine Line

Marie Wallmark[1], Eyal Greenberg[1], Daniel Engels[1]

[1] Southern Methodist University
Dallas, TX
mwallmark@smu.edu, eyalg@smu.edu, dwe@smu.edu

Abstract. Traditionally, it was not feasible for businesses to determine the maximum price the buyer was willing to pay, but with the availability of big data and the deployment of sophisticated algorithms, with a great degree of precision businesses can ascertain the maximum willingness price. Some forms of price discrimination are prohibited under the Robinson-Patman Act of Antitrust (1890), provided demographic characteristics such as race and gender are the determining factors. The problem with this interpretation is that sellers are not transparent about what factors are taken into consideration when determining price. Current laws are either limited in their interpretation or inadequate to properly respond to the potential for sellers to exploit the consumer through discrimination. In this paper, we present a common pricing strategy, *behavioral-based price discrimination,* broadly practiced in business, particularly retailers. In general, price discrimination occurs when instead of a set price, pricing for a product is determined by what the seller knows about the customer. This includes historical data indicating what the customer is willing to pay, combined with certain personal attributes. In this scenario, the same product may be offered at different prices to different individuals or market segments. What data points are considered when designing these sophisticated pricing schemes remains a mystery. Using a dataset containing transactions collected from 2500 households, we demonstrate price discrimination empirically by linking consumer spending to certain demographic characteristics. Additionally, we address the implication of price discrimination to the economic welfare of the consumer, to market competition, and to privacy.

1 Introduction

In Economics, price discrimination is a pricing scheme whereby customers are charged different prices for the same product or service. The basis for which is the idea that some consumers are willing to pay more for a product or service, and under a dynamic pricing scheme the seller is able to extract the extra consumer surplus. There are three degrees [8]: in the first degree, the seller charges the buyer the maximum price the buyer is willing to pay. In second-degree price discrimination, price is determined by the quantity purchased. For example, discount offered for bulk purchases. Finally, in third-degree price discrimination the market is divided into

segments, and price is based on membership into a particular group. Before the data gold rush, businesses were only able to leverage the second and third forms of price discrimination to increase profits. Economist viewed first-degree price discrimination only as a theoretical frame for optimal market efficiency. Today however, with the amount of information about consumers that business can now access, first-degree price discrimination need no longer be relegated to theory: it is now possible for the seller to know enough about individual buyers to determine the maximum amount each buyer is willing to pay [1],
[5], [6], [11].

1.1 Big Data

The past few decades have seen a trend toward increasing connectedness and engagement with technology. The trails about ourselves that we leave behind through our interactions with businesses and products are beneficial in that they allow for a more customized experience. For example, every visit or purchase made at a major local retailer is recorded and documented: these data are collected and stored, then used to construct a unique profile that is regularly updated to best reflect and predict individual consumer's purchasing desires [1,2,3]. Perhaps based on purchasing history, moreover, certain products will be recommended or one will be sent marketing materials tailored to one's unique profile based on previous purchases [2].

A simple trip to your local brick-and-mortar retailer creates the opportunity for collecting a large number of data points about your purchasing habits. Your data may be collected from a variety of channels. Most commonly, this occurs when personal information (age, income level, and other personal identifiers) is provided to the retailer as part of the sign-up process for a loyalty or frequent-shopper card in the hope of saving at checkout. This information, combined with data purchased through third-party data brokers (Acxiom, Corelogic, Datalogix, etc.), can be used for personal profiling, allowing retailers to tailor marketing campaigns specifically to each user. Regardless the path taken, many retailers use transaction data as the basis for the types of coupon deals customers are offered. For example, past transactions (historical data) may indicate a preference for a particular brand of cereal, and what one is willing to pay. Rather than send random coupons for products that a given customer may or may not be interested in, the types of deals/discounts one is offered are customer-specific, tailored to perceived individual needs and consistent with previous habits. A company called Catalina, for example, has built a marketing empire providing personalized digital media solutions to the retail industry.[1] When you are at the checkout line and are given your receipt from Catalina's point-of-sales printers, a list of coupons will be printed on the back of your receipt. The product deals you are offered are based on real-time data points collected about you over the lifespan of your relationship with this particular retailer. Sending certain types of coupons to a market segment is an established marketing practice, and is an example of *third-degree price discrimination* (PD). In third-degree price discrimination, the

[1] Catalina Marketing's point-of-sale printers reach 90 million households per year. https://www.informationweek.com/big-data/big-data-analytics/catalina-marketing-aims-for-the-cutting-edge-of-big-data/d/d-id/1099971?page_number=1

market is divided into segments (i.e., students, senior citizens) and the same product may be sold to each segment at different prices. This is a common and accepted practice, but what is different now is the increased leveraging of technological tools to collect and manipulate data for personalization of products and services.

2 Pricing Models and Transparency

While this may seem a win-win for both retailers interested in increasing their bottom-line and for consumers in getting tailored discount offers, the practice of collecting data on consumers and attempting to influence purchasing habits without explicit consent does raise certain questions. Often consumers are unaware that data is being collected about them and they do not know how it is being used [3]. Though the collection of data has significantly improved the way businesses engage with customers and thereby the potential for improved profit, the question of improved overall consumer welfare remains.

Most retailers who use this practice maintain that only historical data (past transactions) are employed when designing the models that underpin the targeted marketing campaigns which incorporate the various coupon offers, there remains no way for us as the consumers to corroborate this claim. Further, if demographic variables are used in constructing these models, it is likely concealed.

2.2 Consumer Welfare and Market Competition

Individualized pricing schemes can be implemented in a variety of ways. If viewed from the perspective of market efficiency, it may seem like a good deal for consumers to minimize transaction time spent searching for a desired product, or to conveniently receive customized coupons at an agreeable price. From the retailer's perspective, they can use what they know about you to tailor discount offers to you at the most optimal price point, which may vary from customer to customer based on a variety of personal factors. With such competing interests, one wonders if consumers, particularly those enrolled in loyalty programs, are getting the savings expected. It seems counterintuitive to think that businesses would forgo the ability to increase profits by utilizing all the tools available to them, especially big data, particularly when the specifics do not need to be disclosed or can be easily concealed. It is fair to say that those subscribed to loyalty programs may not necessarily get the lowest prices [1]. In this respect, the argument could be made that the consumer stands to see a reduction in overall financial well-being [4]. If market efficiency is contingent upon the empowered consumer, it is fair to say that with the ubiquitous deployment of big data and pricing algorithms, the power dynamic has shifted in the favor of business, thereby challenging the notion of market efficiency.

Further, the convergence of the real and online environments has created for businesses the opportunity to merge data from various points of contact allowing for even more granularity in their ability to profile the consumer. Collecting more disparate information allows business to form a more nuanced and specific profile, thereby improving product and service offering. This creates a data feedback loop (illustrated in Figure 1) where the more data a business has, the better the products and services it can offer to consumers, the more consumers they are able to attract, and the more data they are able to collect. This is purported to benefit the consumer through improved product and service quality, and allow for more choices and market competition.

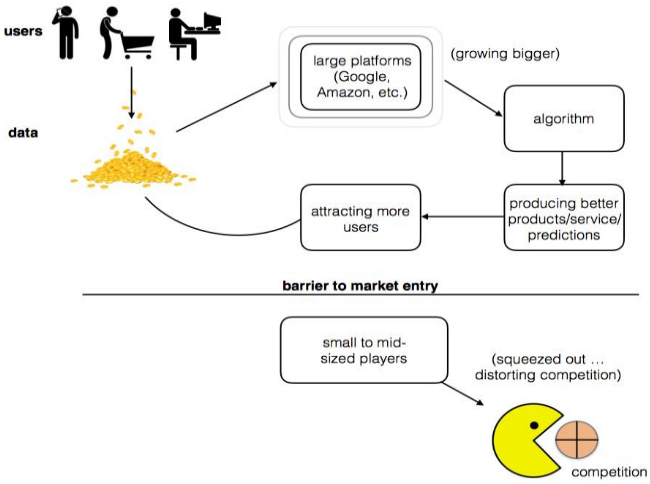

Fig. 1. Firms collect data, use algorithms to improve products and services, overall production improves, attract more customer and collect more data, creating a loop, and eventually, smaller players can't compete and new players cannot gain entry to markets [11].

2.3 Sherman Antitrust Act

Many consumers object to the practice of price discrimination on the basis that it is a violation of fairness, and on the false belief that price discrimination is illegal [2]. Price discrimination is in fact perfectly legal: it is not generally assumed to run afoul of the Sherman Antitrust Law, which protects consumers from discrimination based on race, gender, or other sensitive demographic information. Big data has made it possible for sellers to cobble together enough data points about a buyer to a great degree of accuracy to predict the maximum they are willing to pay. If, however, the data points taken into consideration were among the prohibited ones listed above, then price discrimination would be in violation of the Robertson-Patman Act of the Sherman Antitrust law. According to the Robertson-Patman Act, "price discrimination is lawful if the prices reflect actions taken to promote loss of doing business or an attempt to meet competitors offering [13]." The intent of the Robertson-Patman Act was to protect retailers against unfavorable prices that could adversely impact their competitive ability. Further, the Act only applies specifically to commodities—goods of like grade and quality—and stipulates that sales must be made through interstate commerce. This interpretation is limited, and offers consumers very little protection against exploitative price strategies. In general, the Antitrust Act seeks to stifle the formation of pricing power through anticompetitive conduct, rather than reducing existing power or regulating the manner of its exercise [1], [11]. In other words, the focus has been in protecting consumer welfare indirectly through promotion of market competition, but big data has enable those firms with data advantage in some ways to circumvent the traditional consolidation practices usually viewed as anti-competitive. As it is currently interpreted, the ubiquity of big data renders obsolete the ability of antitrust to effectively regulate market competition and protect consumer welfare. The only recourse left to consumers is the hope that companies, as they thread that fine line between profit optimization and consumer welfare, will forego their own economic interest and put the consumer's first by using the data they collect responsibly. This is an unrealistic prospect. The current systems in place cannot adequately address this new era of big data. Further, there are no mechanisms currently in place that could make clear to consumers and the public at large the exact composition of the models that undergird the pricing strategies used to price discriminate.

3 Our Approach

To detect price discrimination, we look at what variables most explain or account for overall household spending after all discounts are applied. Without access to the pricing models and algorithms used to price discriminate on the household level, or in this case, coupon offers sent, it is very difficult to establish a causal link. To that end, based on the premise of consumer welfare (that is, being better off as a result of tailored product offerings at the best possible values),[2] we take the approach that overall household spending is a good proxy measure for price discrimination if significantly influenced by demographic attributes. In this case, we define "best possible value" simply as a proportion of discounts received to total amount spent. If retailers are using the data collected for better product customization rather than to exploit personal markers to induce spending, we should expect to see a random distribution in total spending. If instead patterns begin to emerge based on certain demographic attributes, then we may be able to point to a systematic mechanism that cannot be attributed to chance alone.

3.1 Data

The dataset used for this analysis was obtained from Dunnhumby[3], a consultancy company for retail and consumer data science projects. This dataset contains

[2] Consumer welfare refers to the individual benefits derived from the consumption of goods and services https://stats.oecd.org/glossary/detail.asp?ID=3177.

[3] Dunnhumby Source Files, The complete journey [Online]. Available: https://www.dunnhumby.com/sourcefiles

household-level transactions collected over a two-year timespan from a group of 2,500 households who are frequent retail shoppers. The data are dispersed amongst eight tables. For this analysis, we use only two of the eight tables, the demographic table containing 801 observations and 8 variables, and the transaction table containing 2,500 households, 1048575 observations and 12 variables. We know that the data was collected over a two year time frame, but the exact time frame is not specified. Therefore, we cannot take into account seasonal trends and potential impact to spending. We instead assume that any fluctuation in spending would be observed across all categories. Further, demographic variables are factorized, such that comparisons to the general population is not possible. For example, households are represented with a unique ID. The composition of the household is unknown. The literature provided with the data indicate that the demographic information used in this project was captured through the retailer's loyalty card program. Using the loyalty card price, we could see exactly what the customer paid at checkout once the discounts are applied. We used the non-loyalty card price as a baseline measure. Presumably the non-loyalty price is what regular customers would pay if no retailer discount was present, and if nothing was known about the shopper's buying habits or personal information.

Using these two datasets, we created two distinct tables: one containing demographic and transaction data for a group identified as loyalty card members and another table containing just transaction data for the remaining households. This group is identified as our control group (non-loyalty card members). The original demographic and transaction dataset contained individual observations for every household transaction. Since we are only concerned with total spending over this time period, transactions were summed per household without regard to product type. From there we used the household_key from the demographic table to join together all related transactions, establishing this as our loyalty card group. Then, we filtered the remaining observations as our control group, reducing the original transaction dataset from 1048575 observations to 1695.

Additional trimming was performed on the final tables to eliminate outliers due to errors in data entry. For example, we observed in the transaction table that the mean number of items purchased per trip was 94, the median 1, and the range 0–83055. Buying 83055 items on one shopping trip seemed improbable. Looking at the top 20 and bottom 20 observations, we could see that on average a typical basket of goods contained fewer than 10 items. We set the filter on the quantity variable from the transaction dataset to only include observations > 10. We also removed superfluous variables not pertinent to our analysis, such as Basket_ID, Day, Product_ID, Store_ID, Trans_Time and Week_No, and converted discount units to absolute numbers, leaving us with the following demographic variables:

- Age (ordinal factors: 6 levels)
- Income (ordinal factors: 12 levels)
- Marriage status (factor: 3 levels)
- Homeowner status (factor: 5 levels)
- Household size (ordinal factors: 5 levels)

Dependent continuous variables documenting amount spent and discounted included:

- Retail discount (retail_disc): Total amount discounted
- Coupon discount (coupon_disc): Amount discounted in retail coupons
- Manufacturers coupons (coupon_match): Amount discounted through manufacturer coupons (i.e., paid through manufacturer promotional)

Additionally, we created the following two new variables: *Total*, which consisted of quantity * sales_value to produce a single value of total amount spent per household; and, in order to place all observations on a common ratio scale, *proportion* of retail discount compared to total amount spent (retail_disc / total). The proportion variable indexes the "best possible value" a household can receive from the retailer, with a proportion of 0 indicating no discount, and 1 indicating 100% discount.

4 Analysis

To determine whether membership in the loyalty card program had any effect upon total spent, retail discount received, and proportion of discount/total, we first compared control and loyalty group on three variables using independent samples t-tests. The loyalty group spent significantly more than control (see Table 1), $t(1133) = 19.73$, $p < .0001$; also unsurprisingly, the retail discount received was also significantly higher in the loyalty group, $t(1130) = 19.43$, $p < .0001$. However, proportion of discount/total was greater in the control group, $t(2054) = -3.30$, $p < .001$. Comparisons are shown in Figure 2. The latter result was unexpected, and suggests that membership in the loyalty program does not necessarily amount to best possible value.

Table 1: Descriptive statistics of Control and Loyalty Groups for three variables of interest

	Control	**Loyalty group**
Total $ *M*	1166.69	2818.82
Total $ *SD*	1413.20	2160.02
Discount $ *M*	154.81	360.18
Discount $ *SD*	177.73	272.89
Proportion *M*	.143	.135
Proportion *SD*	.067	.05

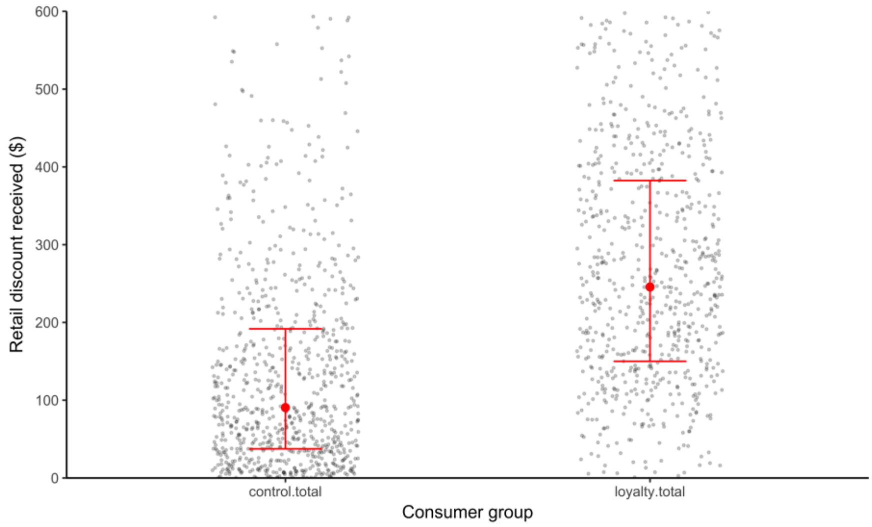

Fig. 2a.Comparison of retail discount received by control and loyalty group (*Note:* Bars show median and middle quartiles)

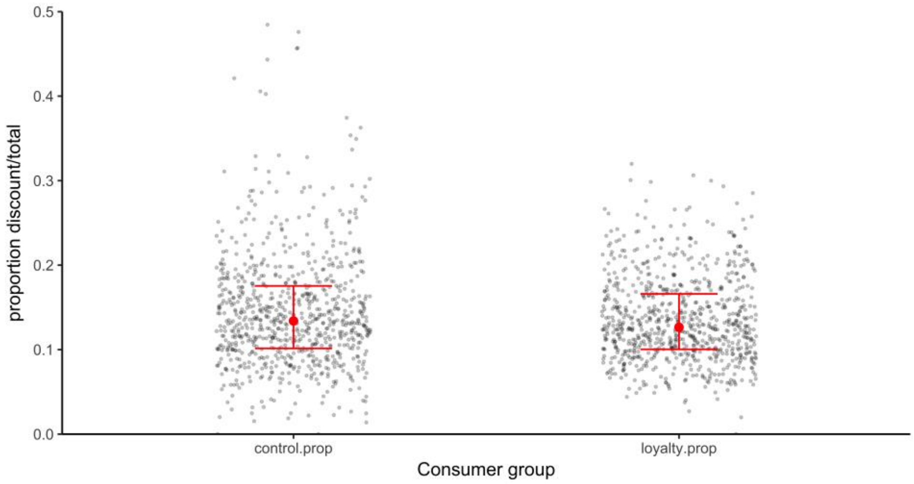

Fig. 2b. Comparison of discount proportion received over total by control and loyalty group (*Note:* Bars show median and middle quartiles)

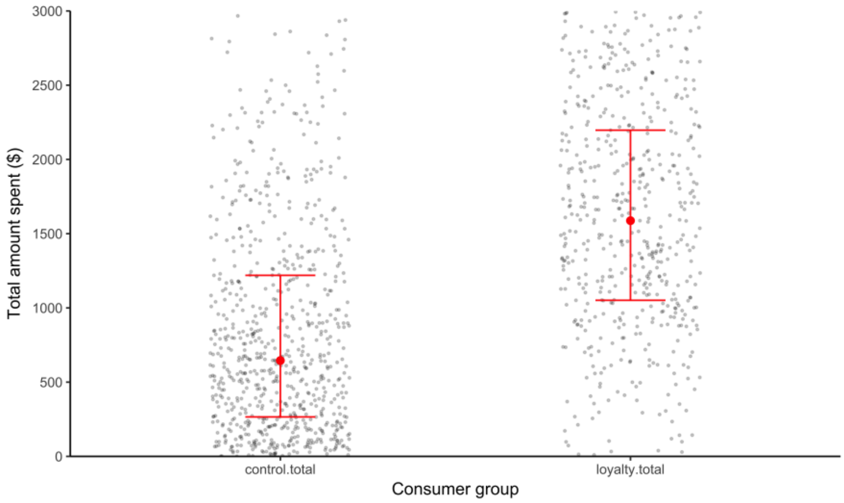

Fig. 2c. Comparison of overall total spent by control and loyalty group (*Note:* Bars show median and middle quartiles*)*

Having determined that control and loyalty groups differed on the three dependent variables of interest, we next explored their association with demographic factors within the loyalty group. Since the income variable was considered a priori to be most significantly associated with consumer spending and discount patterns, we examined the frequency distribution among the 12 original income levels on our three dependent variables of interest. As shown in Figure 3, the distribution of incomes was highly skewed towards the lower brackets. In order to simplify this factor and control for large differences in *N* between income groups in later analyses, therefore, we reduced the factorization to 4 levels: low (under $15 – 34K), low.mid ($35 – 49K), mid-high ($50 – 74K), and high ($75K +). As can be seen, the reduction led to roughly equivalent numbers in each group.

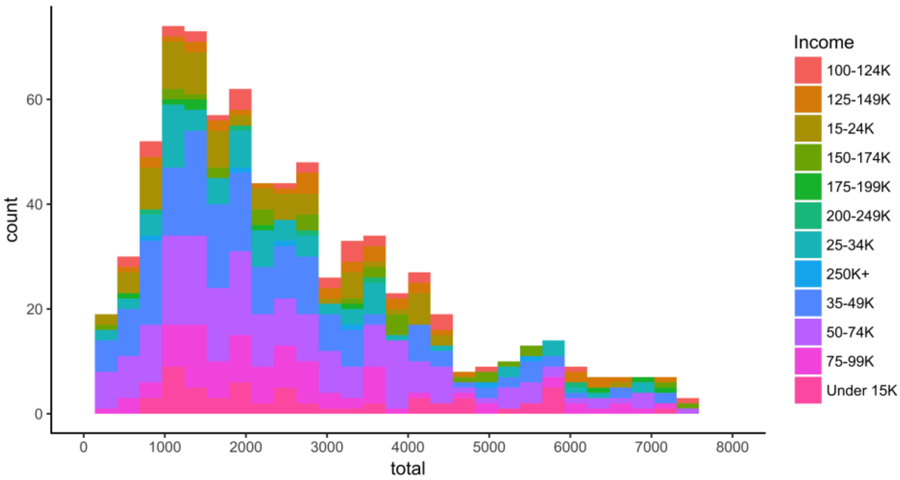

Fig. 3a. Original 12 income levels across total spending

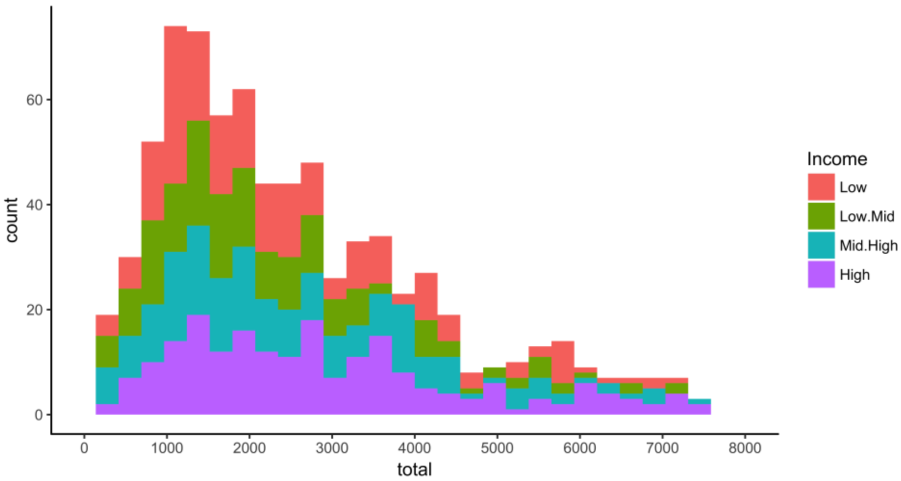

Fig. 3b. Income variable reduced to 4 levels across total spending

In order to test for collinearity of variables before conducting an exploratory linear regression analysis, we first computed a correlation matrix of all pairs of continuous variables. Since the proportion variable was not independent (i.e., it is a composite index of two other variables), it was left out of the analysis. Total and retail_disc were strongly correlated, $r(798) = .82$, $p < .00001$. Coupon variables were correlated with both total and retail_disc, though at weak levels ($r < .30$). In initial regression models, then, these two variables were kept independent in order to avoid distortions due to collinearity. All demographic variables were entered as predictors into three regression models for total, retail_disc, and proportion variables. The explanatory power for all three models was low (less than 11% variance explained in all models); however, all income categories showed a statistically significant association with total, retail_disc, and proportion. Thus, while the magnitude of the effect was small in absolute terms, income appeared to systematically influence the three dependent variables.

To determine whether differences in spending/discounts between income groups were statistically significant, we next computed three one-way ANOVAs (Figure 4). Total amount spent varied by income, $F(3, 796) = 8.13$, $p < .0001$, $\eta^2 = .03$. Tukey HSD post-hoc testing revealed that the high income bracket was significantly different from the other three. This is logical given that many higher-income people have more money to spend, which may enable more expensive purchases relative to households with lower incomes.

In contrast, there was no main effect of income on mean retail discount, $F(3, 796) = 1.53$, $p = .21$. No income group received significantly more or less discounts than the others: thus, while total trended upward in the high-income group, the discount remained flat.

This relationship is reflected in the significant difference in proportion related to income, $F(3, 796) = 10.7$, $p < .0001$, $\eta^2 = .04$, with a similar post-hoc result indicating that high income people had significantly *lower* proportion of discount/total ($M = .11$) than the other brackets ($M = .14$). Though income explains only a small portion of the total variance (4%), people in the lower 75% of the income distribution received a greater discount benefit relative to total amount spent. Although no causation can be inferred from this relationship, it is nonetheless interesting that high-income people got a "worse deal" than the others: unlike the other groups, for whom discount given appeared to constrain total spent, high-income people were more likely to have a higher total despite hitting a ceiling on discount. This could be the result of spending habits among this group—they are more likely not to need the extra discount benefit than the other groups, and may be more inclined to spend on more expensive specialty items that are not offered at discounted prices. It could also be the result of steering, targeted discounting, and other measures aimed to nudge those of means to spend at full price. Further research will be required to explore this connection.

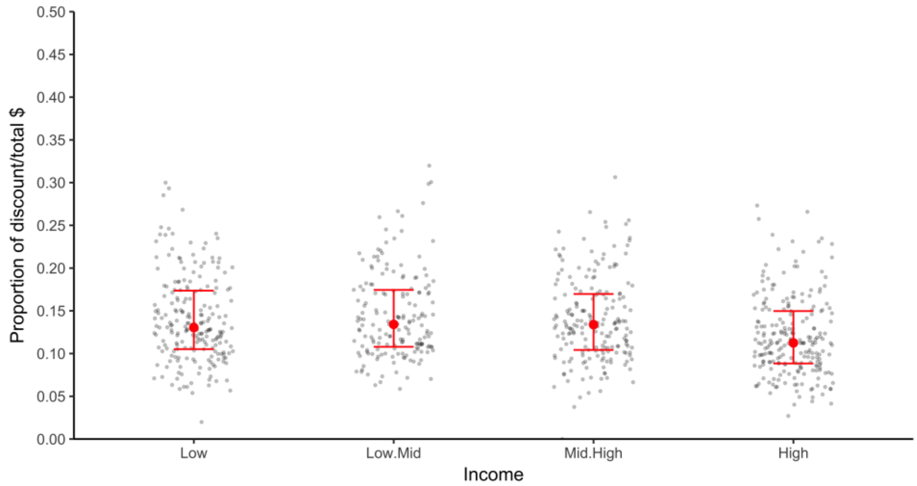

Fig. 4a. ANOVA of 4 level income variable showing the proportion of discount over total spending

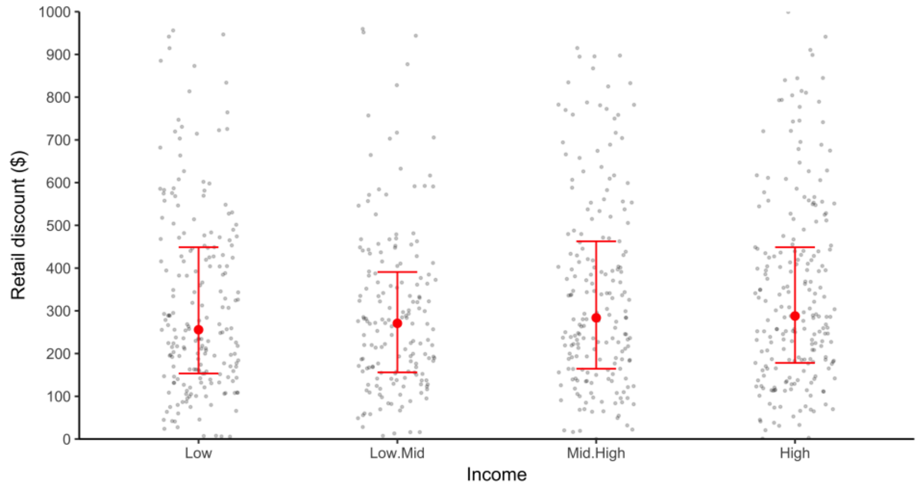

Fig. 4b. ANOVA of 4 level income variable showing retail discount across total spending

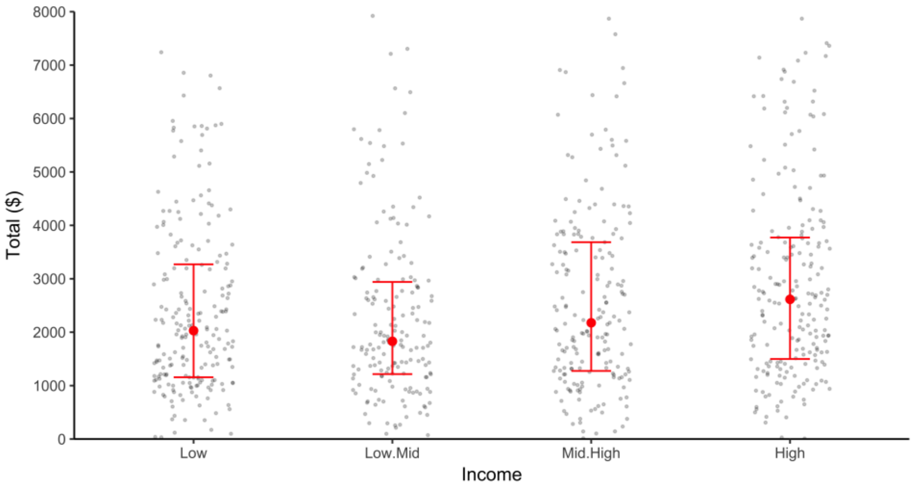

Fig. 4a. ANOVA of 4 level income variable across total spending

5 Previous work

Much of the work done on price discrimination and the issues related (privacy, ethics and regulation) to the practice focused on e-commerce. Though the focus of this paper is on the retail environment, there are still many parallels between the more static brick-and-mortar environment and the dynamic world of online retailers. Like Facebook and Google, whose main source of revenue is derived from targeted ads, with the use of personal information captured from loyalty card program memberships coupled with sophisticated point-of-sales printers[1], retailers are able to collect data in real time, and respond to consumer demands quickly. The reality of which is that real and online environments are converging.

6 Conclusion

To be clear, this study is correlational only: it does not reveal a causative link between the demographic variable of income and retail prices paid. Particularly, considering the limited nature of the dataset used for this analysis, as delineated in section 3.1. We cannot therefore definitely say whether price discrimination was a factor in the results we observed. Moreover, the magnitude of the income-based difference in total spend, retail discount received, and proportion of discount/total was quite weak, as indicated in small effect sizes in the ANOVA tests and low R^2 values in our exploratory linear regression models. However, interpreted through the lens of business analytics, if targeting pricing strategies were employed, such effects would need to be quite small and unnoticeable: indeed, it is hard to imagine that total expenditures and discounts could be driven by income to a notable extent without consumer backlash. This is consistent with the literature on this issue: for example, Benjamin Shiller [10] in their study of the characteristics of the Netflix subscription base produced a statistically significant model relating subscriptions to web use that only accounted for 8 to 12% of variance explained. While our results cannot support any claims regarding demographic-based price discrimination, differences in spending as a function of income is nevertheless consistent with this hypothesis. It is expected that high-income people would spend more; however, coupled with a non-significant difference in discounts between the income levels, it is clear that high-income consumers received a lower proportional discount than others. If income was unrelated to the proportion of total spending that was discounted, we would expect to observe no differences on this variable between income groups. As stated previously, there are numerous plausible explanations for this result. However, price discrimination should not be ruled out as a possible variable. Further research is necessary to test this claim, though with the proprietary nature of most commercial pricing algorithms, "smoking gun" evidence will likely be difficult if not impossible to gather.

A. **Appendix:** Dataset Details

Table 1. HH_DEMOGRAPHIC (801 HOUSEHOLDS)

Variable	Description
Household_Key	Uniquely identifies each household
Age_Desc	Estimated age range
Maritial_Status_Code	Marital Status (A-Married, B-Single, U-Unknown
Income_Desc	Household income
Homeowner_Desc	Homeowner, renter, etc.
HH_Comp_desc	Household composition
Household_Size_Desc	Size of Household up to 5+
Kid_Category_Desc	Number of Children present up to 3+

Table 2. TRANSACTION_DATA (2500 HOUSEHOLDS)

Variable	Description
Household_Key	Uniquely identifies each household
Basket_ID	Uniquely identifies a purchase occassion
Day	Day when transaction occured
Product_ID	Uniquely identifies each product
Quantity	Number of the products purchased during the trip
Sales_Value	Amount of dollars retailer receives from sale
Store_ID	Identifies unique stores
Coupon_Match_Disc	Discount applied due to retailer's match of manufacturer coupon
Coupon_Disc	Discount applied due to manufacturer coupon
Retail_Disc	Discount applied due to retailer's loyalty card program
Trans_time	Time of day when the transaction occurred
Week_no	Week of the transaction. Ranges 1-102

B Appendix: Program Code

Source code for data wrangling, visualization and analysis is available at the link provided below:

goo.gl/MEhraL

References

1. R. Woodcock, (2016, August 02). Big Data, Price Discrimination, and Antitrust. Hastings Law Journal, Vol. 68, 2017.
2. J. Turow, L. Feldman, and K. Meltzer, (2005, June 01). Open to exploitation: American shoppers online and offline. *UPenn Repository* [Online]. Available: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1035&context=asc_papers
3. G. Pingitore *et al,* (2013, October 29). Consumer concern about data privacy: What can businesses do? *JD Power and Associates* [Online]. Available: http://www.jdpower.com/sites/default/files/Consumer_Concerns_Data_Privacy.pdf
4. M. E. Stucke, A.P. Grunes. *Big Data and Competition Policy*. Oxford, UK: Oxford University Press, 2016.A.
5. Hannak *et al*, "Measuring Price Discrimination and Steering on E-commerce", *Internet Measurement Conference '14*, Vancouver, BC, Canada, Nov. 5-7, 2014.
6. The Executive Office of the President of the United States (EOP), "Big Data and Differential Pricing", EOP, Washington, DC, Feb. 2015.
7. J. Mikians *et al*, "Detecting price and search discrimination on the internet", *Hotnets*, Seattle, WA, Oct. 2012.
8. M. Armstrong, M. (2006, October). Price discrimination [Online]. Available: http://else.econ.ucl.ac.uk/papers/uploaded/222.pdf
9. J. Mikians *et al*, "Crowd-assisted Search for Price Discrimination in E-Commerce: First results", *CoNEXT*, Santa Barbara, CA, Dec. 2013.
10. B.R. Shiller. (2014, January 30). First Degree Price Discrimination Using Big Data. [Online]. Available at SSRN: https://ssrn.com/abstract=2314997
11. A. Ezrachi, and M. Stucke. *Virtual Competition*. Cambridge, MA: Harvard University Press, 2016.
12. L. Chen, A. Mislove and C. Wilson, "An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace", *International World Wide Web Conference Committee (IW3C2)*, Montreal, Quebec, CA, Oct. 2012.
13. Federal Trade Commission [Online]. Available: https://www.ftc.gov/about-ftc/bureaus-offices/bureau-consumer-protection/office-technology-research-investigation
14. *The Economist* (2016, January 28). Schumpeter: Flexible figures [Online]. Available: http://www.economist.com/news/business/21689541-growing-number-companies-are-using-dynamic-pricing-flexible-figures