W241 Final Report

Vous voulez du vin? - Does the use of foreign language advertising in wine increase customer's purchase likelihood?

Rory Liu, Charlotte Swavola, Sharad Varadarajan

Introduction

When walking through the grocery store, one can easily spot products with foreign-language labels - "Vins de France" on a wine bottle, "Wirklich gut!" on a sausage pack, etc. Advertisers often confront consumers with foreign languages, such as German or French [1]. In fact, the use of foreign language in advertising is a well-studied topic. According to the study done by France Leclerc, Bernd Schmitt and Laurette Dubé,showing French pronunciation of a brand name affects the perceived hedonism of the products, attitudes toward the brand, and attitudes toward the brand name[2]. Furthermore, Jos Hornikx and Frank van meurs in their study that the use of foreign languages serve as a strong cue for a product's country of origin, and the associations that the foreign language evoke and those that the country-of-origin evoke are similar[3].

While scholars seem to agree that foreign language advertising shapes product perception by implicitly giving consumers cues on product's country of origin, few of the existing studies dived deep into the causal link between foreign language and consumer's purchase likelihood. This is an important causal link to establish. At the end of the day, the success of an advertising campaign is, or should be measured by the additional sales the campaign generates. Without understanding of whether having foreign language in product communications increases sales, advertiser will be ill-guided in their decision to apply such advertising techniques.

In our study, we seek to understand the relationship between foreign language advertising and consumer willingness to pay. To narrow down our scope for a manageable experiement given our time and resource constaints, we decided to focus on one particular product: wine. The high variance of wine prices, and the strong association between place of origin and the signal on quality makes it a great subject for our study.

Experimental Design

To fully understand the relationship between foreign language advertising and will purchases, we drafted a multi-factorial design for our experiement. There are three main factors that we want to investigate: advertising language, wine's country of origin, and length of flavor profile description.

For advertising language, we will test 3 languages: English (as baseline), French and German. We selected French because France is typically perceived as a premium wine production country, and

[1]Jos Hornikx, Frank van Meurs & Robert-Jan Hof (2013) The Effectiveness of Foreign-Language Display in Advertising for Congruent versus Incongruent Products, Journal of International Consumer Marketing, 25:3, 152-165, DOI: 10.1080/08961530.2013.780451

[2]Leclerc, F., B. H. Schmitt, and L. Dubé. 1994. Foreign branding and its effects on product perceptions and attitudes. Journal of Marketing Research 31 (2):263–270

[3]Jos Hornikx & Frank van Meurs (2017) Foreign Languages in Advertising as Implicit Country-of-Origin Cues: Mechanism, Associations, and Effectiveness, Journal of International Consumer Marketing, 29:2, 60-73, DOI: 10.1080/08961530.2016.1243996

		Page Language					
		English		French		German	
		Flavor Profile Description					
		Long			Short		
Country of Origin	US	1	2	3	4	5	6
	France	7	8	9	10	11	12

Figure 1: Groups Design

we hypothesize that advertsizing in French will increase consumer's willingness to pay. We also selected German as a comparison language. Germany is not typically associated with wine, and therefore, by including German, we will be able to see whether the effect of having foreign language is limited to the language of the country associated with a product, or whether it expands to other languages as well.

For wine's country of origin, we included US and France. Here we are particularly interested in the interaction between country of origin and advertsing language. Prior studies have found that foreign-language advertising serve as country of origin cues. By including country of origin as a variable, we are able to see whether foreign language has any additional impact when country of origin is also given.

For flavor profile description, we have two versions - long and short. We believe that this could be a proxy for foreign language "dosage". A long flavor profile in French might be more noticeable than a short French tagline, and may amplify the effect that we find.

To summarize, we have a 3 x 2 x 2 design, which results in 12 groups of participants. (See summary in figure 1).

Because of our time and resource constraint, we could not conduct a real-life experiement with actual wines in actual stores. Instead, we decided to conduct the study through a Qualtrics survey and gather responses through Mechanical Turk. The survey is structured into the following sections:

We start with some demographics question about participants primary language, income and gender. These will serve as covariates in our later model. Then, we ask participants about their wine-related behavior, including how often they drink wine, whether they like Cabernet Sauvignon, and how often they purchase bottles of wine. We think that these wine-related behavior will explain some of the variation in willingness-to-pay, and therefore will be good covariates that help us reduce our standard error. The third section is the main part of our experiement where we introduce the treament. Here, we randomly assign the participants into one of twelve groups mentioned above, and show them a simulated purchasing page (example seen in figure 3).

In the simulated wine page, we translate all texts into the groups assigned language, and provide english translation for key elements (e.g. flavor profile, 'top wine of the year' tag) to make sure that

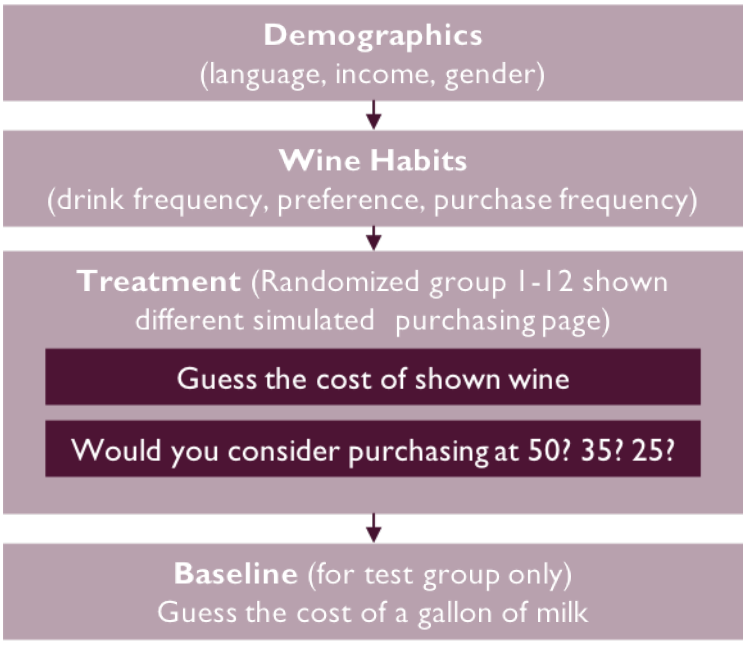


Figure 2: Survey Flow

Figure 3: Page Sample

the respondent understand the information. All groups are shown the exact same wine (which was a hypothetical wine fabricated by us), and the same descriptive information. We made the decision to present the page entirely in French or German for related tracks because we wanted to make sure participants notice our language treatment.

We then ask the respondents to guess the cost of the wine. This is an un-anchored direct pricing question, aimed to understand respondents' first perception of the wine shown. Afterwards, we use a simplified version of the Gabor Granger method developed by André Gabor and C. W. J. Granger in the 1960s [4]. In this method, we ask respondents whether they would consider purchase the shown wine at $50. If respondents say now, we ask the same question at a lower price point of $35. If they still answers no, we lower the price further to $20. Through this series of 3 questions, we seek to understand where their true willingness-to-pay lies. We selected the 3 price points to cover a wide range of potential willingness-to-pay, and validated our choice through a pilot study of 100 respondents. From the pilot, we saw that most respondents' willingness to pay fall into the 20-50 price range.

Our survey is concluded after the treatment section for our main sample of 1200 respondents. We used the data from this main sample to run various regressions and randomized inferences and seek to find results with statistical significance.

However, to validate the findings from this main sample, we collected another (smaller) sample of 600 respondents. The data from this second batch of respondents served as our validation test. By running the key models we deemed important from main sample with this validation data set, we can confirm that the relationship we found is real.

In our second launch, we maintained the same survey experience by asking exactly the same questions in the same sequence. However, we did introduce a baseline question at the very end of the second survey, asking respondents to estimate the price of a gallon of milk. Because this question came at the very end of the survey, it should not interfere with our treatment, and would not create any difference between the data from first and second launch. However, having this baseline questions helps us establish a reference price points for respondents price guesses, and adjust for any overall inflated / deflated guesses.

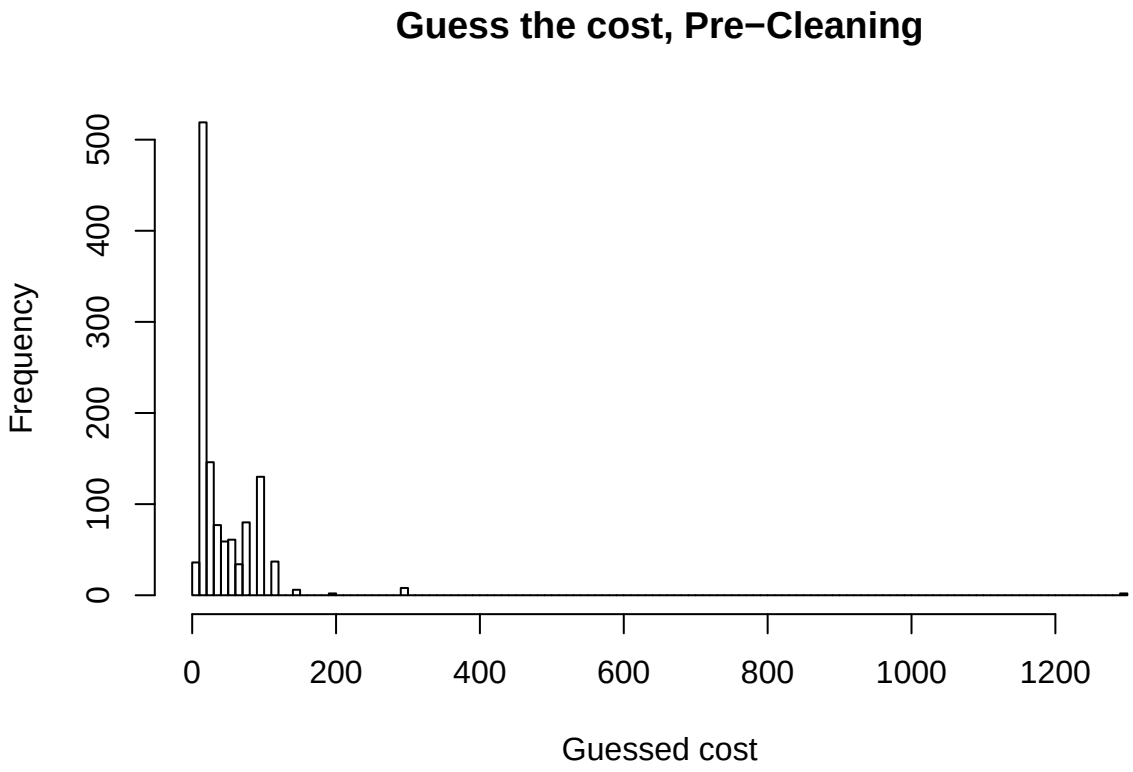
EDA and Data Cleansing

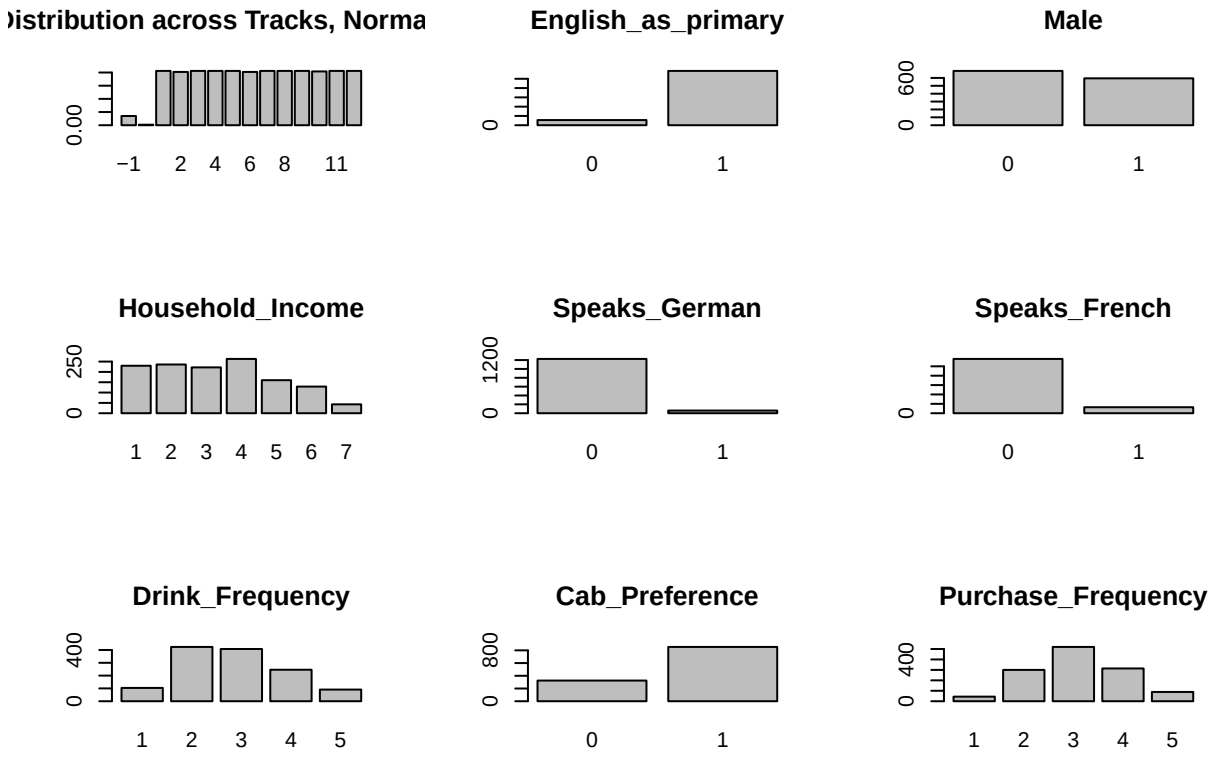
After loading the data gathered from our survey, we first went through some general data processing and examined descriptive statistics to better understand our general results. The three figures below helped support the team in getting a high-level summary of the data. With regards to the distribution for our outcome variables, the team noticed considerable variance in respondent's price perception but that the large majority of guesses fell between 1-100 dollars. This finding greatly influenced our model building decisions. With regards to the covariate barplots below, the team was pleased to see a rather normal distribution for both purchase frequency and drink frequency. Additionally, we noticed that English was the primary language for almost all of our respondents, and that a menial number respondents were fluent in either German or French; these are encouraging statistics considering that we want to evaluate the effect of languages that are "foreign" to our users.

##

[4] Gabor, A. and Granger, C. (1966). Price as an Indicator of Quality: Report on an Enquiry. Economica, 33(129), p.43.

```
## Output variables, Pre-Cleaning
## ===================================================
## Statistic        N    Mean  St. Dev.  Min     Max
## ---------------------------------------------------
## Guess_the_Cost 1,207 75.243 339.562  1.000 8,600.000
## Purchase_50    1,206 0.363   0.481     0       1
## Purchase_35    1,206 0.532   0.499     0       1
## Purchase_20    1,206 0.824   0.381     0       1
## ---------------------------------------------------
```





After evaluating these high-level statistics, we proceeded to analyze more granular details regarding each submission and discovered the following areas that require attention:

1). There were ~5% attrition (i.e. 5% of the respondents did not finish their survey). The attrition we see falls into 2 types: pre-treatmenet attrition and post-treatment attrition. For pre-treatment attrition, we proceeded to clean them out completely, because these respondents were not exposed to treatment yet, and their dropping off can't have anything to do with treatment we introduce. Dropping them will not create bias in our data. For post-treatment attrition, we wanted to be more careful. We will examine them later on to understand whether we have differential attrition based on treatment assigned.

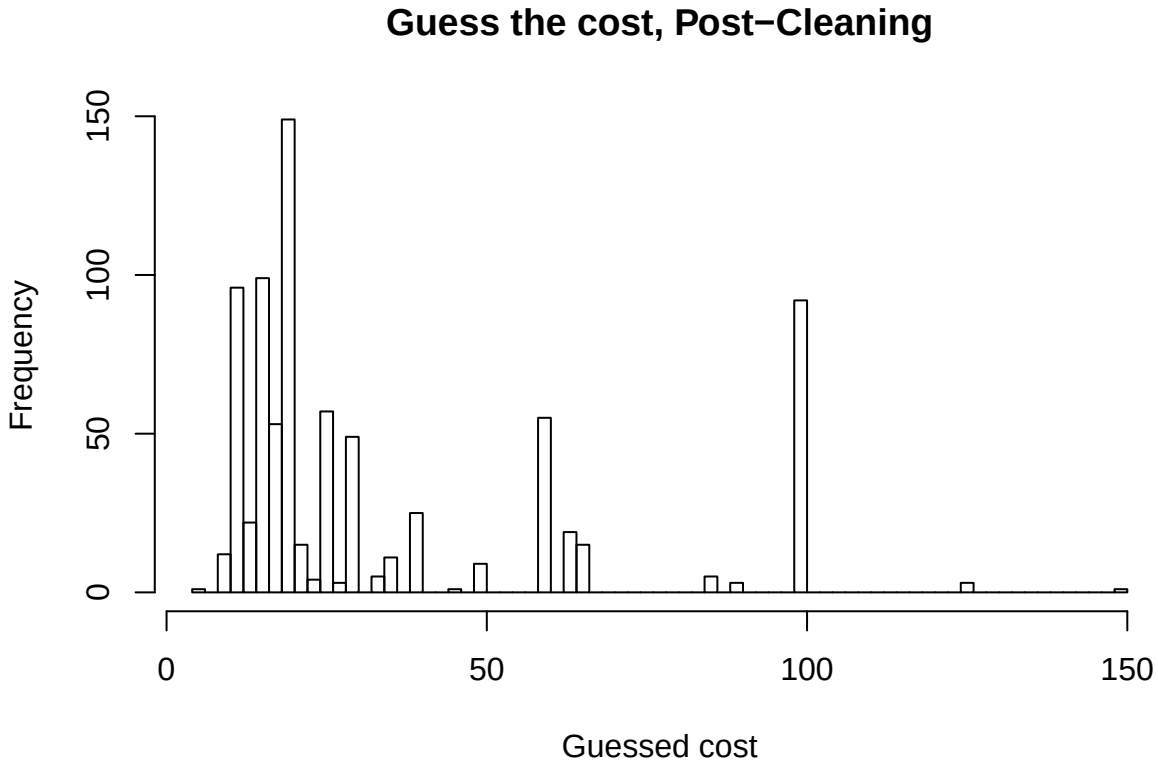
2). There are ~1/3 of respondents that have duplicated IP address and/or geographical coordinates (latitude and longitude). We are worried about these responses because they might be repeat survey takers (that potentially got exposure to more than one treatment), or they might be bots or click farms that do not answer surveys seriously. We proceed to clean out all responses that does not have unique ip addresses and/or geographical coordinates.

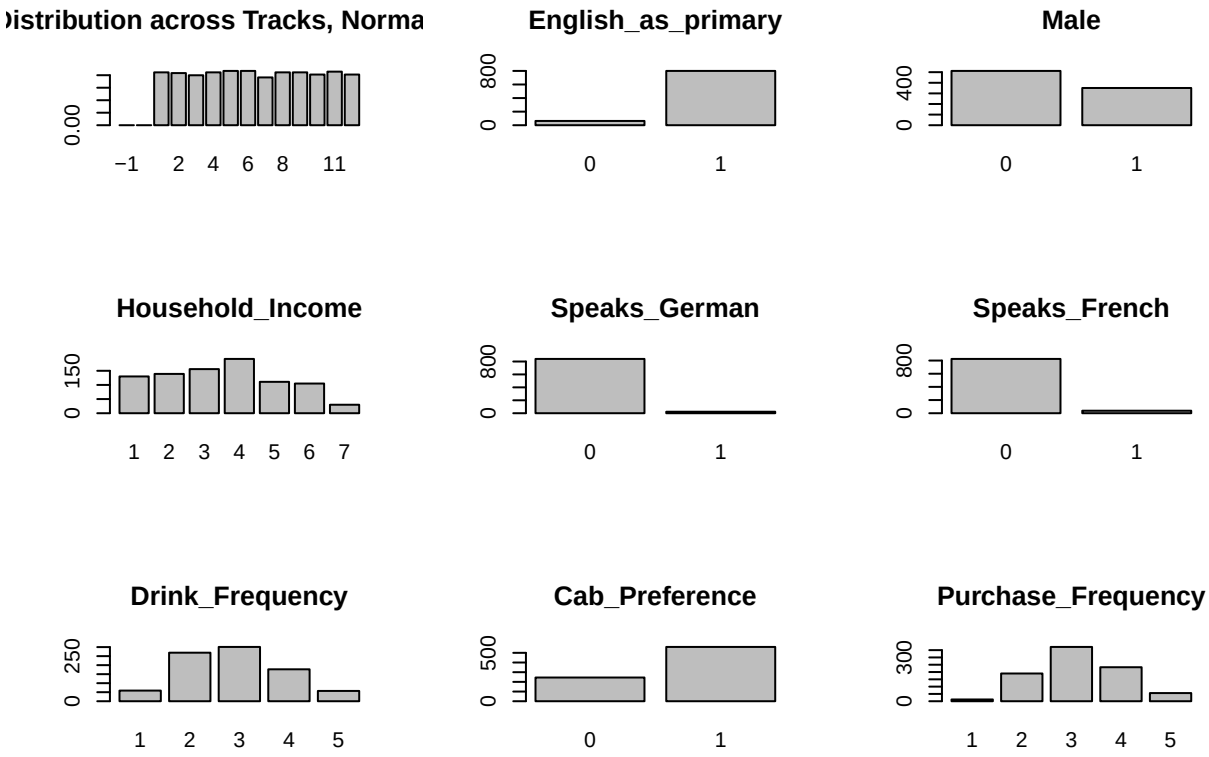
3). As seen from the histogram above, we reiterate that we noticed many large outliers for guess-the-cost. The large variation in the answers will hurt our ability to obtain a reasonable standard error and find significant findings. We need to pay special attenion in our modeling phase later, to try to minize the effect of outliers.

After cleaning the data, we analyzed the previous descriptive statistics once more to see the effects. Below we see that cleaning reduced our outlier problem significantly for guess-the-cost, although considerable variation still exists.

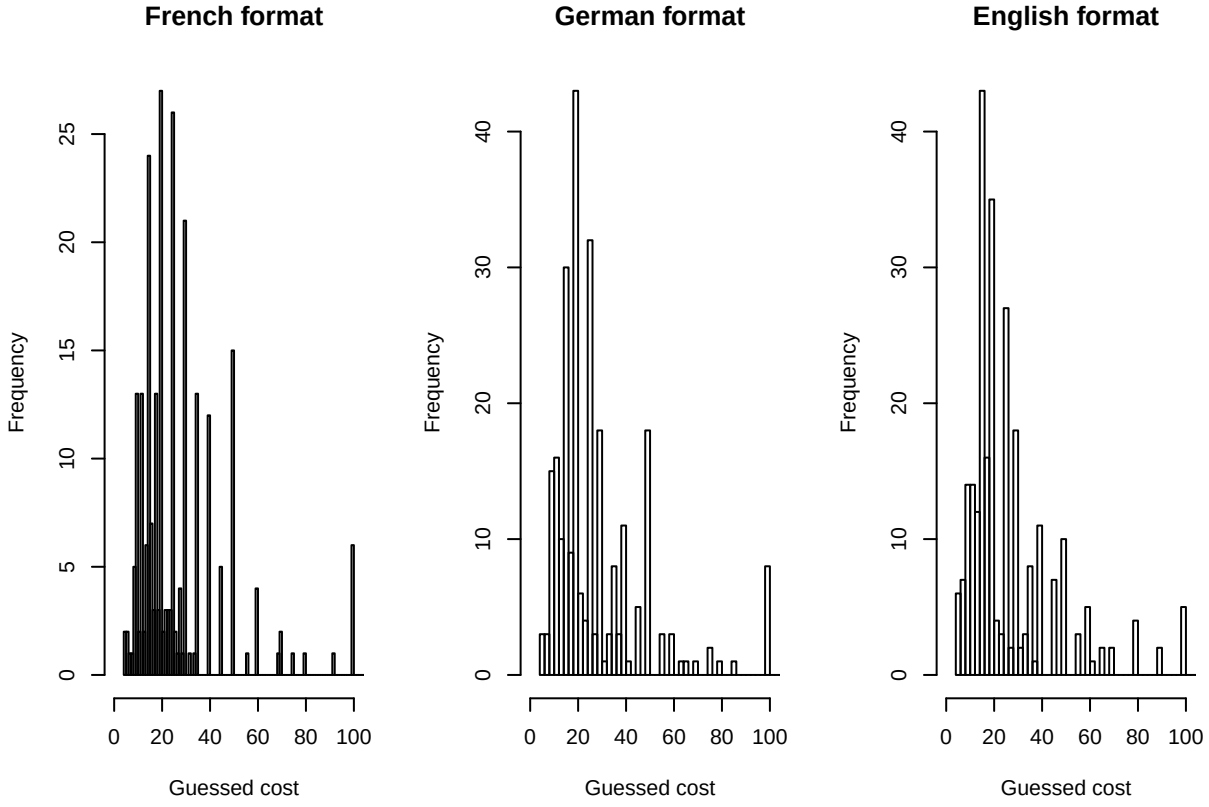
Post cleaning

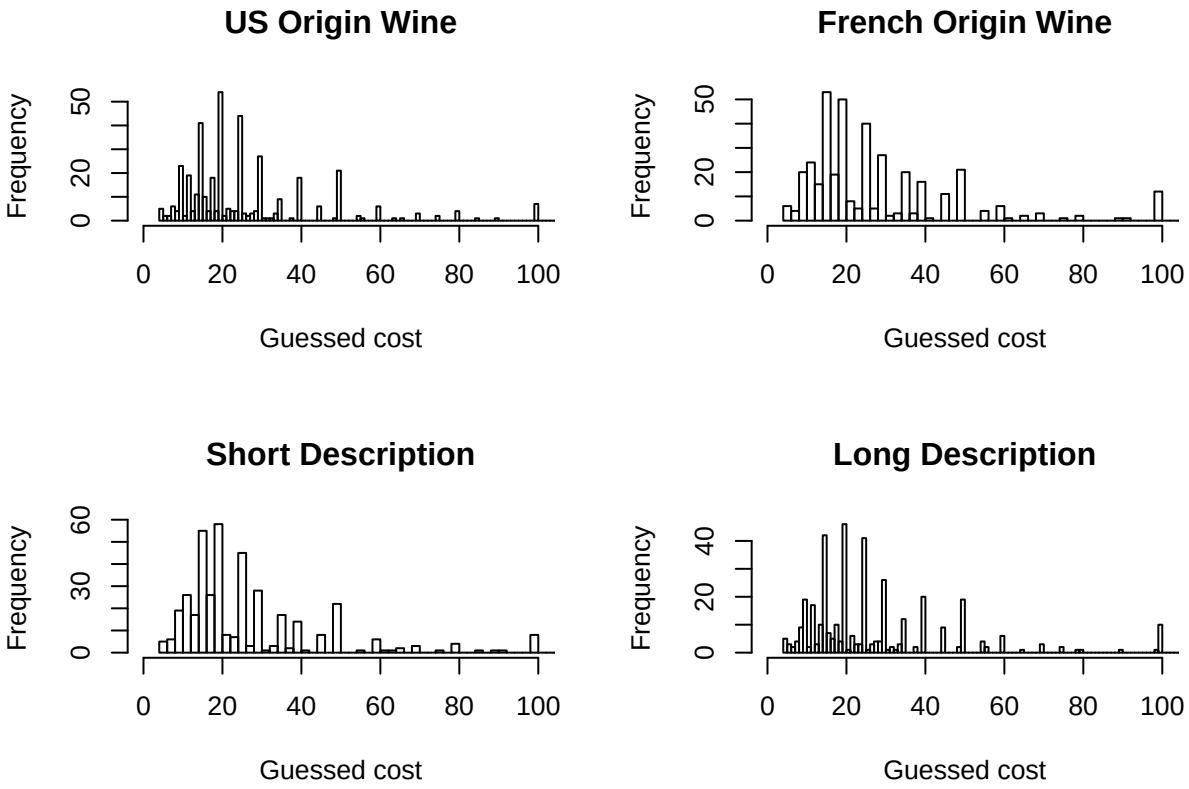
```
##
## Output variables, Pre-Cleaning
## ================================================
## Statistic       N   Mean  St. Dev.  Min     Max
## ------------------------------------------------
## Guess_the_Cost 807 39.715 115.716  4.000 1,999.000
## Purchase_50    806 0.261   0.439     0       1
## Purchase_35    806 0.450   0.498     0       1
## Purchase_20    806 0.793   0.406     0       1
## ------------------------------------------------
```





Next, the team decided to dive a bit deeper into the guess-the-cost results, to see whether a treatment effect was obvious through histograms. We evaluated histograms that were summarized at each level of our three treatment factors:





From these historgrams of the answers, we don't observe a very apparent difference based on page lanugae, country of origin, or length of description. This is not to say that we don't have any treatment effect, it's that the effect that our treament have may be too small to appear in these visualization.

Experimental Assumptions

Before beginning our modeling, the team proceeded to first validate our experimental assumptions, including conducting a covariance balance check, a check for potential differential attrition, and a check for potential heterogeneous treatment effects.

Demographic/Covariate Balance Check

For a covaraite balance check, the team regressed each treatment factor against our covariates. The goal was to use the F-statistic as an indicator of whether our covariates had any explanatory power on which treatment track a respondent was assigned to. For 3 treatment factors (Origin, Language Format, Description length) with 2, 3 and 2 levels respectively, the team ended up constructing four regressions that are listed below:

1)$FrenchPage=\beta_0+\beta_1 Male+\beta_2 Income+\beta_3 EnglishPrimary+\beta_4 DrinkFrequency+\beta_5 CabPreference+\beta_6 PurchaseFrequency+u$

2)$GermanPage=\beta_0+\beta_1 Male+\beta_2 Income+\beta_3 EnglishPrimary+\beta_4 DrinkFrequency+\beta_5 CabPreference+\beta_6 PurchaseFrequency+u$

3)$USOrigin=\beta_0+\beta_1 Male+\beta_2 Income+\beta_3 EnglishPrimary+\beta_4 DrinkFrequency+\beta_5 CabPreference+\beta_6 PurchaseFrequency+u$

4)$LongDescription=\beta_0+\beta_1 Male+\beta_2 Income+\beta_3 EnglishPrimary+\beta_4 DrinkFrequency+\beta_5 CabPreference+\beta_6 PurchaseFrequency+u$

In these regressions, Male, EnglighPrimary and CabPreference are binary variables. Income, Drink Frequency and Purchase Frequency represent factor vectors. Equations 1 and 2 were evaluated on subsets of our original data frame. For equation 1, the data frame only contained observations where the purchasing page was in English or French. For equation 2, the data frame only contained observations where the purchasing page was in English or German. After running all four of these regressions, the p-values for each of the F-statistics were much greater than 0.05. The team concludes that our covariate balance check passes, seeing that we cannot reject the null hypothesis that all beta values are zero for these regressions.

Attrition check

As mentioned earlier, our attrition is rather low for the entire study (~5%), so we are not very concerned about potential bias caused by potential differential attrition. However, we decided to do a check for completeness. Here, we used logistic regression (since whether or not finishes the survey is binary) to check whether being assigned to certain treamtent leads to more or less attrition than others.

Table 1: Attrition Regressions-Logistic with Odds Ratio

	Dependent variable:		
	Finished		
	(1)	(2)	(3)
French_Purchasing_Page1	0.326***		
	(0.396)		
German_Purchasing_Page1	0.405**		
	(0.405)		
US_Origin1		0.990	
		(0.274)	
Long_Description1			0.653
			(0.279)
Constant	30.556***	14.214***	17.826***
	(0.339)	(0.196)	(0.214)
Observations	863	863	863
Log Likelihood	−205.143	−209.964	−208.780
Akaike Inf. Crit.	416.287	423.928	421.560
Note:		*p<0.1; **p<0.05; ***p<0.01	

After exponentiating the output coefficients, we saw that logistic regression suggests respondents assigned the French or German purchasing page were around 3 times more likely to drop off before finishing (compared to those assigned the English purchasing page). In other words, we do find

evidence for differential attrition (this statistically significant finding was replicated with linear regression as well). However, since our attrition rate is very low, even if the 5% that dropped out have very extreme outcomes, they are unlikely to bias our estimate too significantly. Therefore, we are not very concerned about this effect. Additionally, when conducting a differential attrition check for our validation study, the team was not able to replicate the finding, and found no statistically significant effects.

Heterogeneous treatment effects

We tested for heterogeneous treatment effects using regression for those who answered "Never" for their wine purchasing behavior. We hypothesized that people that never purchase wine may have a rather un-firm value perception of wine, and may be influenced by our treatment more easily. In our test, we found a significance in the interaction term between never-pruchase and German language, which does suggest the presence of HTE.

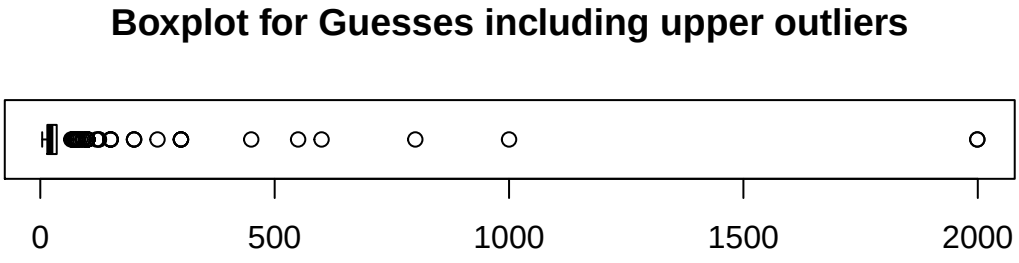
Table 2: HTE

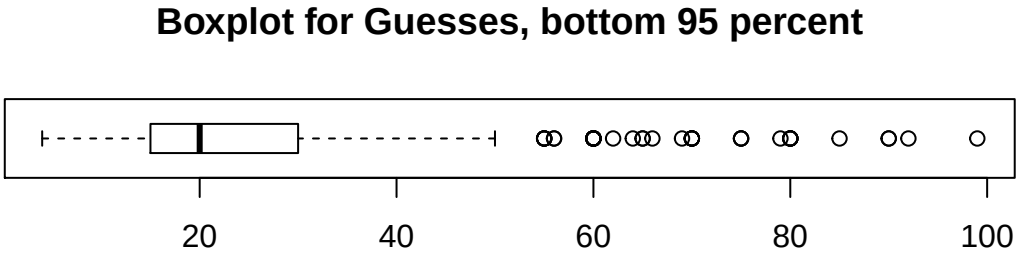
	Dependent variable:		
	Guess_the_Cost		
	(1)	(2)	(3)
French_Purchasing_Page	0.842		
	(1.290)		
German_Purchasing_Page	1.301		
	(1.275)		
US_Origin		−1.262	
		(1.057)	
Long_Description			1.055
			(1.072)
Never_purchased	9.218**	13.656***	14.007***
	(3.730)	(3.070)	(3.240)
French_Purchasing_Page:Never_purchased	−0.573		
	(5.361)		
German_Purchasing_Page:Never_purchased	12.145**		
	(5.357)		
US_Origin:Never_purchased		−3.234	
		(4.431)	
Long_Description:Never_purchased			−3.675
			(4.493)
Constant	23.786***	25.209***	24.078***
	(0.900)	(0.753)	(0.751)
Observations	807	807	807
Residual Std. Error	14.281 (df = 801)	14.516 (df = 803)	14.630 (df = 803)
Note:			*p<0.1; **p<0.05; ***p<0.01

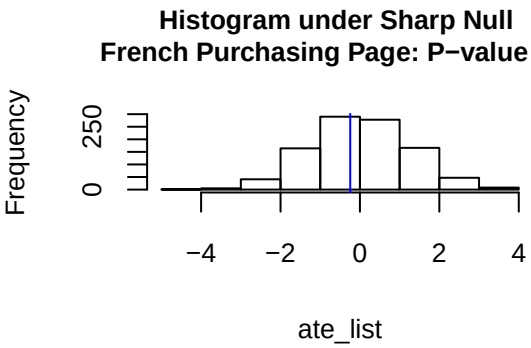
Modeling

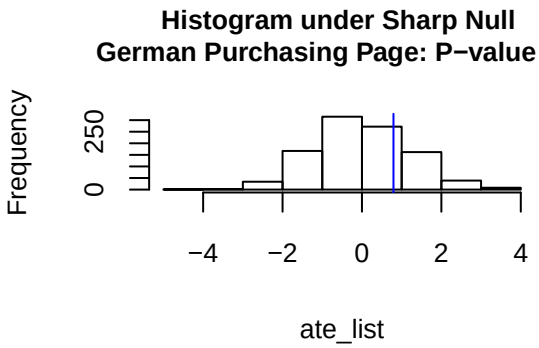
1 RI with limit on guesses

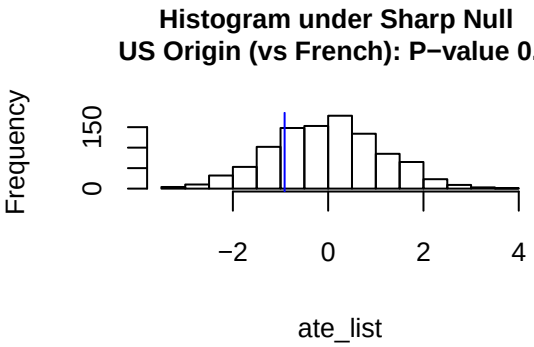
Our first attempt to isolate a treatment effect using random inference and the sharp null hypothesis. By nature of RI, in which a high outlier may swing the entire treatment- or control-assigned group, we limited the guesses to the bottom 95%. Using the reduced outcome set, we tested the simple treatments and our specified interaction (same format*origin) hypotheses.

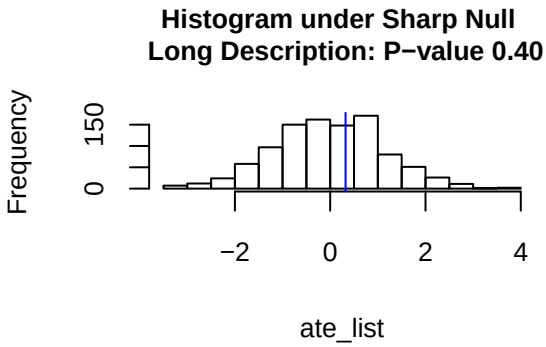












The results from our Randomization Inference attempt are displayed in the 4 charts above. Here we see no statistically significant results for all of the 4 treatments (French purchasing page, German purchasing page, US origin, and Long description. Therefore, we could not reject the sharp null hypotheses that the treatment effect is zero for all respondents. The team moved on to regression models.

2 Model Evaluation Strategy

The first outcome of interest was respondents guesses for the price of a wine. Initially, the team wanted to see how standard OLS regressions would perform with our data. We were not very optimistic based on the considerable variance observed from our earlier descriptive statistics. Preliminary models that we built reinforced this concern. The table below contains simple OLS models where we used only the treatment variables as predictors. As you can see from the table below, the most simplistic of regressions were returning large standard errors.

After running a few more models with covariates, the team realized that these large standard errors were going to be an issue in all OLS regression models we run. Therefore, the team decided to go with a different strategy to evaluate a treatment effect for this outcome, and that was through Robust Linear Models. The idea of this model is to weigh observations differently based on how high their residuals are. We viewed this a compromise between eliminating extreme guesses entirely from the analysis and treating all price guesses equally like we would in OLS regression. When re-running the same models, where treatment variables were the only predictors, but with RLM we saw that the treatment effect dropped considerably; however the standard error was more reasonable and reflective of the distribution among the majority of observations.

For the second outcome of interest, the team decided to evaluate three different outcomes of

Table 3: Standard OLS Regressions

	Dependent variable:			
	Guess_the_Cost			
	(1)	(2)	(3)	(4)
French_Purchasing_Page1	−11.295			−11.426
	(10.030)			(10.034)
German_Purchasing_Page1	−3.094			−2.892
	(9.905)			(9.910)
US_Origin1		−8.508		−8.659
		(8.157)		(8.168)
Long_Description1			−5.034	−4.877
			(8.161)	(8.165)
Constant	44.413***	44.046***	42.213***	51.167***
	(6.985)	(5.804)	(5.720)	(9.015)
Observations	806	806	806	806
R^2	0.002	0.001	0.0005	0.004
Adjusted R^2	−0.001	0.0001	−0.001	−0.001
Residual Std. Error	115.833 (df = 803)	115.780 (df = 804)	115.831 (df = 804)	115.869 (df = 801)
F Statistic	0.672 (df = 2; 803)	1.088 (df = 1; 804)	0.381 (df = 1; 804)	0.710 (df = 4; 801)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4: RLM Basic Regressions

	Dependent variable:			
	Guess_the_Cost			
	(1)	(2)	(3)	(4)
French_Purchasing_Page1	0.837			0.842
	(1.274)			(1.248)
German_Purchasing_Page1	1.611			1.640
	(1.258)			(1.232)
US_Origin1		−1.478		−1.561
		(1.023)		(1.016)
Long_Description1			1.009	1.066
			(1.031)	(1.015)
Constant	24.300***	25.656***	24.480***	24.429***
	(0.887)	(0.728)	(0.723)	(1.121)
Observations	806	806	806	806
Residual Std. Error	14.634 (df = 803)	13.840 (df = 804)	14.103 (df = 804)	14.084 (df = 801)
Note:				*p<0.1; **p<0.05; ***p<0.01

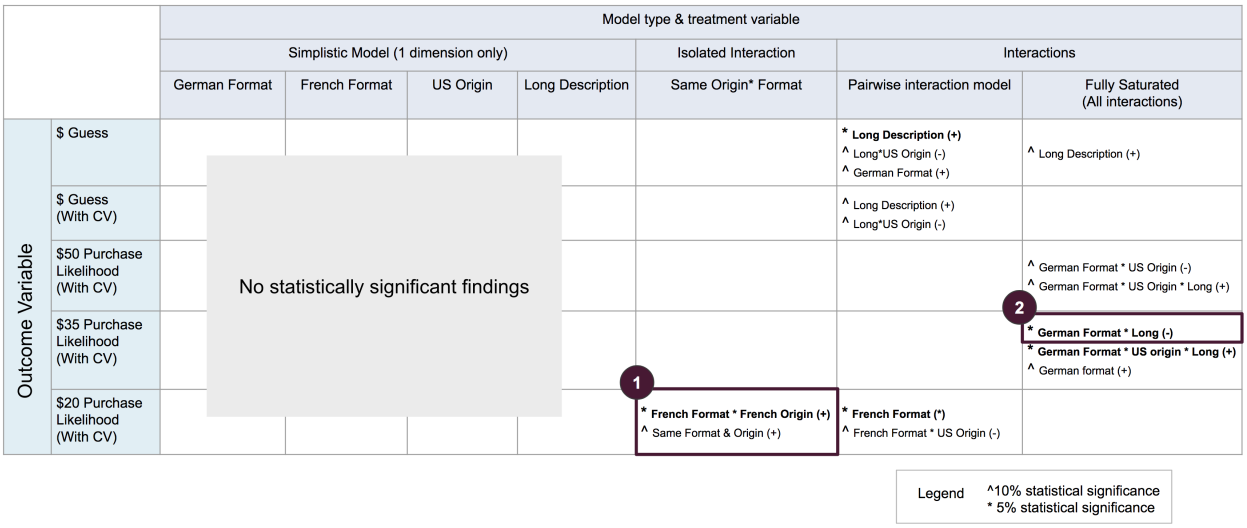


Figure 4: Model summary

willingness to purchase. Since our outcomes for each of these 3 price points are binary, we decided to use binomial logistic regression as our method of analysis. With regards to the output of logistic regression models, the team chose to exponentiate the coefficients, as we could then interpret effects as the odds of purchasing the wine at a certain price point (odds ratio). The team also considered using an ordinal logistic regression instead, and looking at maximum pricepoint at which individuals would purchase; however we did not believe the levels of our willingness to purchase outcome met the conditions of the proportional odds assumptions. Therefore we thought it would be more prudent to use binary logistic regression with three separate outcomes.

Model Matrix

The matrix above describes the different RLM and Logistic regression models we ran on our training set. Each column describes the details of predictors we used in the model, and each row describes the outcome variable for the model and whether or not covariates were included:

1) The left-hand side of the table pertains to our most simple models, where we only used the treatment variables as predictors without any interactions

2) For Isolated Interactions we only examined the effect of treatment where the wine origin and purchasing page format are the same. For example the effect of French origin-French language treatment, or US origin-English language treatment.

3) For pairwise interaction models, we would include only 2 of 3 treatments per model and observe their dynamic. For example, one model could include language format and description length as predictors, as well as their interaction, regardless of the origin of the wine.
4) For the fully saturated models, we included each of the treatment variables, and all pairwise and three-prong intereactions between treatment.

The exact model equations can be found in Appendix A.

On the left-hand side of the table there is a box stating, “No Statistically Significant Findings” and

that pertains to the most simplistic regressions we ran. For example, when trying to isolate the effect of wines of US Origin, regardless of the length of the flavor profile or purchasing page format, we saw no significance. As we move left to right on the matrix we come across some of our more detailed models. At this point, the team started to uncover some findings that are significant at the 10% level and others at the 5% level. For brevity, we will discuss the two findings we found most interesting. The first was the positive effect of the interaction between French Format and French Origin on the willingness to purchase at $20.

$$Purchase20=\beta_0+\beta_1 FrenchFormatOrigin+\beta_2 EnglishFormatOrigin+\beta_3 Male+\beta_4 Income+$$
$$\beta_5 EnglishPrimary+\beta_6 DrinkFrequency+\beta_7 CabPreference+\beta_8 PurchaseFrequency+u$$

The results indicated that a respondent was 2.075 times more likely (statistically significant at 5%) to purchase a wine of French Origin-French Format for $20 than a respondent who saw a wine of English Origin-French Format. This made a lot of sense to us, considering that adding French language to describe a French wine could possibly attribute more authenticity to the wine.

The second model of interest was the fully saturated model where there was a negative effect involving German Format on the willingness to purchase at $35.

$$Purchase35=\beta_0+\beta_1 FrenchPurchasingPage+\beta_2 GermanPurchasingPage+\beta_3 USOrigin+\beta_4 Long$$
$$Description+\beta_5 FrenchPurchasingPage*USOrigin+\beta_6 GermanPurchasingPage*USOrigin+$$
$$\beta_7 FrenchPurchasingPage*LongDescription+\beta_8 GermanPurchasingPage*LongDescription+$$
$$\beta_9 USOrigin*LongDescription+\beta_{10} FrenchPurchasingPage*USOrigin*LongDescription+$$
$$\beta_{11} GermanPurchasingPage*USOrigin*LongDescription+\beta_{12} Male+\beta_{13} Income+$$
$$\beta_{14} EnglishPrimary+\beta_{15} DrinkFrequency+\beta_{16} CabPreference+\beta_{17} PurchaseFrequency+u$$

The results indicated that a respondent was about 4 times less likely (statistically significant at 5%) to purchase a wine with German Language-Long Description-French Origin for $35 than a respondent who saw a wine of English Language-Long Description-French Origin. Since German is not a language typically associated with wine, the team was curious whether German was the driving factor in respondents choice not to purchase the wine at $35.

Validation Study

Based on the previously mentioned matrix, the team evaluated many regression models. Therefore, making any causal inferences from these model results would not be prudent, considering the chances of retrieving a false positive (type II error) increases considerably when running all of these models. To avoid making false causal claims and verify that our findings were legitimate, the team re-evaluated our two findings of interest on a validation set, to see if we could replicate the direction of the effect with statistical significance.

As mentioned in the introduction, the team added a baseline price-perception question for a gallon of milk post-treatment. By adding this baseline question, the team was able to gauge whether or not a respondent was providing guesses in terms of US currency; we chose to remove all observations where respondents guessed over $10 for a gallon of milk, with the hope of ridding ourselves of harmful outliers for our outcome variables. While this method helped remove many inflated guesses,

there were still a number of outliers remaining in our dataset. Therefore, the team continued with RLM instead of OLS regressesion for our Guess the Price outcome variable.

After data collection and data cleansing for our validation set, the team went through the following tasks prior to model analysis:

1) Conducted a covariate balance check and confirmed there were no baseline differences between the groups (in terms of our measured covariates)

2) Observed that the differential attrition identified in our training set did not exist in our validation set. In fact, respondents who were assigned the French/German purchasing page treatments were less likely to attrit in our validation study compared to those assigned the English purchasing page (though not statistically significant). The direction of this differential attrition effect is opposite to what was observed in our training study.

3) The heterogeneous treatment effect for never-purchasers did not persist in the validation study as well

After re-running our two models of interest we observed the following:

1) Respondents were 2.252 times more likely (statistically significant at 10%) to purchase a wine of French Origin-French Format for $20 than respondents who saw a wine of English Origin-French Format. While the direction and relative magnitude of the effect persisted in the validation study, the statistical significance dropped from 5% to 10%.

2) Respondents were 1.592 times more likely (not statistically significant) to purchase a wine with German Language-Long Description-French Origin for $35 than respondents who saw a wine of English Language-Long Description-French Origin. Compared to what we observed in the training study, the direction of the effect switched directions and the magnitude of the effect dropped considerably. Additionally, the finding was no longer statistically significant.

Conclusion

The one significant finding that we have that persisted in both main and validation study (although only at 10%) is the positive impact of French wine in French purchasing page. We found that this combintion drives up purchase likelihood at $20. Advertsing in French for a entry-level French wine may be a good idea for businesses (further validation needed).

Having gone through the study, we learned a lot about experimental design, and if we were to do the study again, we would improve oru design in the following ways:

1. Introducing a baseline question in main study. The milk-cost baseline question helped us control for some outlier answer, especially those caused by currency perception. Having this in the main study would definitely help in our modeling process.
2. Limiting geographical location of respondents. In mechanical Turk, we did not limit to only US respondents, for fear that we may not achieve our ambitious sample on time. If time permits, this limitation would be useful because it will then eliminate currency bias.

3. Increasing sample size, especially in validation study. The large variation we see in responses also calls for large sample sizes. If we have enough time and resources, gathering a larger sample helps increase statistical power.

4. Changing treatment dosage. For this treatment, we used the "maximum dosage possible" by translating everything into the treatment language. This may have overwhelmed or confused respondents, which lead to some differential attrition. In a future study, it may be wise to reduce the treatment dosage to a more manageable way (e.g. only translating flavor profile and tagline),

5. Redesigning the guess-the-cost question. We liked the question because it is completely unanchored, which allowed us to find psychological price thresholds (like $20). however, for nature of this question is outlier prone. In a new study, we may consider changing this question type, or at least putting a maximum answer threshold on guess-the-cost questions implicitly, without anchoring respondents. Having an implicit max price (the highest price imaginable for a wine, say 300), where only people that entered more than the value would see, may help us reduce the magnitude of outliers.