Analysis of Water Source Preference for People in Maryland

Aarohi Mehta
amehta17@terpmail.umd.edu

Ankit Dhall
ankitd@terpmail.umd.edu

Nisha Dayananda
dnisha@terpmail.umd.edu

ABSTRACT

Everyday choices and decisions are made subconsciously without realizing the same decision as a relevant one. There can be factors influencing the choice or there might not be but generally, people don‘t think about it. This led us to question whether the choice of bottled water or tap water is independent or dependent on some external factors. Does the common man in Maryland think about the factors when he makes the decision of choosing one over the other? Is there any factor that does influence this decision? This analysis works on the influence of a factor over the decision and choice of the source of drinking water.

AUTHOR KEYWORDS

Hypothesis, Logistic regression, water, categorical variables

INTRODUCTION

Drinking water is a necessity that people care about and hence opens up the opportunity of the decision of choosing the source of drinking water open to influence from other factors such as location, age group, taste, etc. To understand this, we created a study where people‘s choice and factors would be taken into account to try and understand whether there is an association and if possible, a correlation, as well or if the choice is independent and purely based upon the user.

STUDY DESIGN

For this analysis, we wanted to see the preference of the source of drinking water of the people of Maryland. Also, we considered variables that might have an impact on the choice of choosing the water source for drinking. These variables were created with a purpose to define association with the water source so that it could help us reject or fail to reject the null hypothesis we created around the same question that we considered. The null hypothesis, as well as the alternate hypothesis, were created which are as follows:

Null Hypothesis or Ho: µ : The choice of source of drinking water considering tap water and bottled water is independent of external factors.

Alternate Hypothesis or Ha: µ: The choice of source of drinking water considering tap water and bottled water is not independent of external factors.

Since this study took into account the user‘s preference and their decisions with the context of other variable information supplied by the users, this was an observational study.

DATA COLLECTION

To reach the population of Maryland and get information from them regarding the topic, we chose to build a Google Form survey form which allowed us to create a well-structured survey questionnaire. The survey was distributed over social media networks like Facebook and Whatsapp to reach the known people. To expand our reach, we also chose to post the survey form on specific subreddits on Reddit. These subreddits were specifically r/UMD, r/maryland and r/baltimore. This survey form generated a total of 241 responses. After analyzing the collected data and cleaning the unnecessary data, we came down to 198 total responses based on which this study is based.

VARIABLES OF THE EXPERIMENT

The population is all the people living in Maryland and have been living in Maryland for at least 3 months and the sample is the response we got to from our survey. This is an observational study and our dataset has 19 variables 198 observations post cleaning. We dropped all the observations which had come from people not living in Maryland since last three months and people who chose I do not have a preference for drinking water source preference. The table 1 shows the list of variables we used in the study.

The descriptive statistics of the variables are shown in the table 2.

When we did the exploratory analysis for the variable Preferred method of drinking water, we see that there are 157 respondents who have selected Tap water and 41 respondents who have selected Bottled water.

During the survey, when people were asked if they have ever fallen sick because of drinking tap water, 5 people selected the option ‘Yes‘, 181 people selected the option ‘No‘ and 12 participants selected the option ’I don’t know’.

ANALYSIS

Most of the variables in the dataset are categorical, our dependent variable and independent variables are also categorical, so to come with the best analysis, Logistic Regression approach has been used. We used the forward approach for logistic regression, i.e. starting with using one variable to predict the dependent variable and then adding more variables to find the best fit model.

Deviance is a measure of goodness of fit of a model. Higher numbers always indicate a bad fit. The null deviance shows

Variables	*Values*
Living in the state of Maryland for more than 3 months	Yes, No
County	19 Counties of MD
Age group	Under 18, 18-24, 24-34, 35-44, 45-54, 55-65, Above 65+
Preferred method of drinking water	Bottled Water, Tap Water
The average number of glasses of water per day	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10+
The average number of glasses of non-water beverages per day	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10+
Reusable bottle for water	Yes, No, Not applicable
The charge associated with the water supply in the house	Yes, No, Maybe
Rate of buying water bottled	Daily, weekly, monthly, yearly
Average cost spend on water bottles per week	>5, 5-10, 10-15, 15-20, 20-25, 25-30, 30+
Switch to bottled water	Yes, No, Maybe
Trust in bottled water	5 to 1, 5 being high trust
Trust in tap water	5 to 1, 5 being high trust
Device to filter tap water at home	Yes, No, I do not know
Taste of bottled water	5 to 1, 5 being good taste
Fall sick because of drinking tap water	Yes, No, I do not know

Table 1. Variables used in the study

Variable	*Mean*	*Median*	*1st Qrt*	*3rd Qrt*	*SD*
Trust in bottled water	3.7	4	3	5	1.10
Trust in tap water	3.37	3.0	3	4	1.15
Taste in tap water	3.93	4	3	5	1.13
Taste in bottled water	4.1	4	4	5	0.94

Table 2. Descriptive statistics

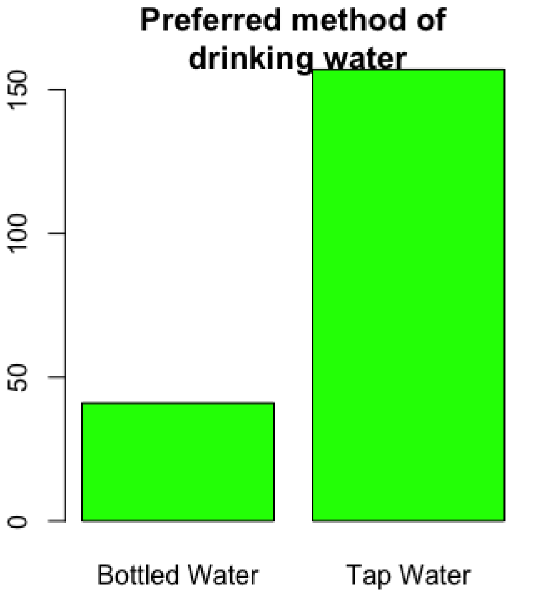

Figure 1. Distribution of drinking water preference

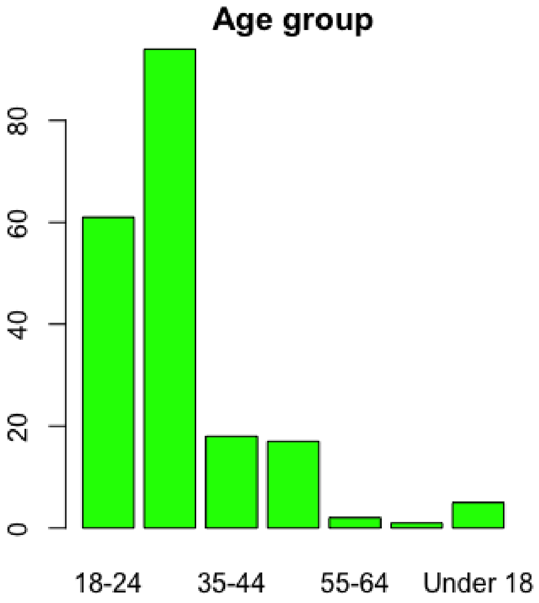

Figure 2. Distribution of Age group

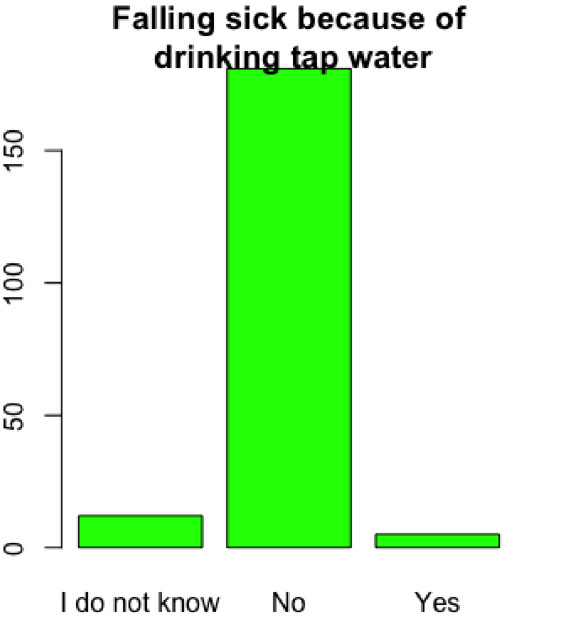

Figure 3. Distribution of people falling sick drinking tap water

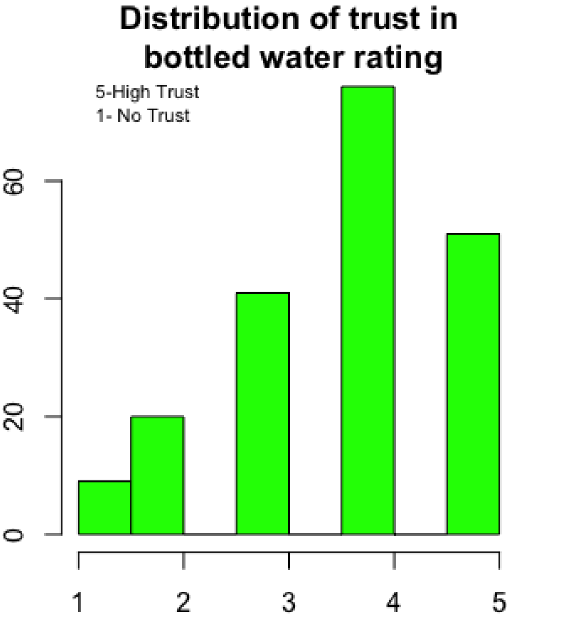

Figure 4. Distribution of trust in bottled water rating

how well the response variable is predicted by a model that includes only the intercept (grand mean) whereas residual with the inclusion of independent variables. If your Null Deviance is really small, it means that the Null Model explains the data pretty well. Likewise with your Residual Deviance. When we compare the null and the residual deviance for all the models, we see that the models Residual Deviance < Null Deviance and thus we are doing good with the predictor variables, but not all models are statistically significant. For all the models, when we compare the residual deviance, for the model which uses counties, Age group, falling sick because of tap water, glasses of water, taste of tap water, trust in tap water, trust in bottled water, taste of bottled water to predict the preference of water, the residual deviance is the least which is 104.82.

We built 13 different models to check which model is statistically significant. In this paper, we are explaining only the first model we tried and the best model we got. We are also interpreting the odds ratio.

The residual deviance for the first model which predicts the preference of water with counties is 191.86 with Baltimore City county as a reference group and their odds ratio are shown in table 4. Thus with the forward approach, we came with the best fit model with least residual deviance.

The odds ratio can be interpreted as follows.

Compared to those living in Baltimore City, the odds of people living in Baltimore county preferring tap water are 71% less likely to prefer tap water versus bottled water. Compared to those living in Baltimore City, the odds of people living in Montgomery county preferring tap water are 73% less likely to prefer tap water versus bottled water. Compared to those living in Baltimore City, the odds of people living in Prince George county preferring tap water are 77% less likely to prefer tap water versus bottled water.

This model is our best fit model. When predicting the preference of water source using counties, age group, falling sick because of tap water, glasses of water, taste of tap water, trust in tap water, trust in bottled water, taste of bottled water and having Baltimore City, age 25-34, not having a device for filtering water at home and not falling sick because of tap water as the reference group, we get the following table as the results.

We see that Baltimore County, Montgomery County, other counties, age 18-24, the taste of tap water, trust in tap water and taste of bottled water are statistically significant predictors with alpha = 0.05. When making to odds scale, Compared to people belonging to age group 25-34 living in Baltimore City and not having a device for filtering water at home and not falling sick because of tap water, the odds of people belonging to Baltimore county preferring tap water are 91% less likely to prefer tap water versus bottled water. Compared to people belonging to age group 25-34 living in Baltimore City and not having a device for filtering water at home and not falling sick because of tap water, the odds of people belonging to Montgomery county preferring tap water are 93% less likely to prefer tap water versus bottled water. Compared to people

	Estimate	*Std. Error*	*zvalue*	*Pr(>|z|)*	*OR*
Intercept	2.33	0.427	5.46	<0.001	10.33
countyBaltimore County	-1.23	0.63	-1.94	<0.01	0.29
countyMoCo	-1.27	0.59	-2.15	<0.05	0.27
countyPG	-1.45	0.56	-2.56	<0.05	0.23
countyOthers	-1.33	0.55	-2.40	<0.05	0.26

Table 3. The results of the model including county variable

	Estimate	*Std. Error*	*zvalue*	*Pr(>|z|)*	*OR*
Intercept	1.79	1.78	1.01	0.313	5.991
countyBaltimore County	-2.39	0.99	-2.39	<0.05	0.091
countyMoCo	-2.58	0.98	-2.64	<0.01	0.076
countyPG	-1.57	0.87	-1.80	<0.1	0.207
countyOthers	-2.09	0.95	-2.19	<0.05	0.124
age18-24	-1.47	0.69	-2.13	<0.05	0.229
age35-44	-0.65	1.19	-0.55	0.583	0.521
age45-54	-0.95	0.99	-0.97	0.333	0.383
ageothers	-1.72	1.28	-1.34	0.180	0.179
sickYes and Don't know	-1.48	0.82	-1.82	<0.1	0.226
glasses	-0.04	0.12	-0.29	0.768	0.963
taste	1.24	0.30	4.07	<0.001	3.452
trust	0.60	0.30	1.99	<0.05	1.822
trust$_bottled$	0.53	0.35	1.53	0.126	1.702
taste$_bottled$	-1.37	0.44	3.12	<0.01	0.253

Table 4. The results of the model including county, age, taste, trust, falling sick, number glasses variables

belonging to age group 25-34 living in Baltimore City and not having a device for filtering water at home and not falling sick because of tap water, the odds of people belonging to other countries (except Prince George's, Baltimore County, Montgomery County) county preferring tap water are 88% less likely to prefer tap water versus bottled water. Compared to people belonging to age group 25-34 living in Baltimore City and not having a device for filtering water at home and not falling sick because of tap water, the odds of people belonging to age group 18-24 preferring tap water are 78% less likely to prefer tap water versus bottled water. Compared to people belonging to age group 25-34 living in Baltimore City and not having a device for filtering water at home and not falling sick because of tap water, the odds of people falling sick and not knowing if they fall sick because of tap water preferring tap water is 78% less likely to prefer tap water versus bottled water. We see that for every rank increase in the taste of tap water, the chances of people preferring tap water increases by 1.24 on the log odds scale. We see that for every rank increase in the trust in tap water, the chances of people preferring tap water increases by 0.60 on the log odds scale. We see that for every rank increase in the taste of bottled water, the chances of people preferring tap water decreases by 1.37 on the log odds scale.

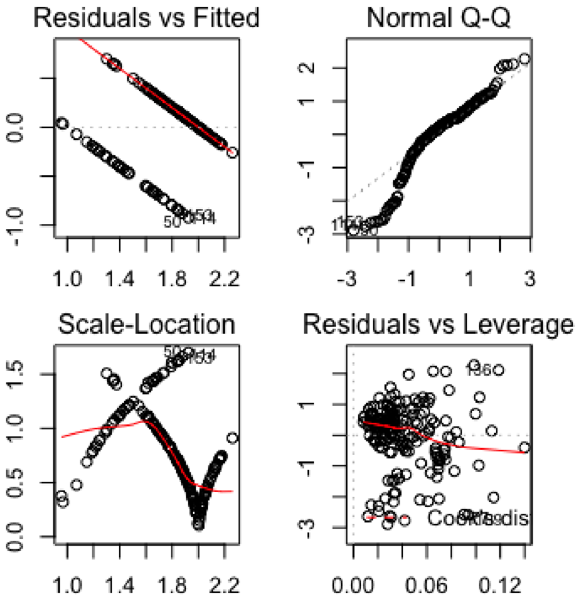

Figure 5. Regression Diagnostic Plot

When we test the assumptions in the model, we see the normally distributed errors. To assess normality, we see normal distribution when plotted points form a straight, 45-degree diagonal line. Here we can see that we are deviating on the sides in the Q-Q plot. Secondly, we see if there are no influential outliers by seeing the Residual vs Leverage Plot in Figure 2. Cook's distance tells about the impact of outliers and influential cases. We know there is a problem when for any observation the Cook's distance is close to 1 or more. For our model, we don‘t see that to be very problematic. Thirdly, we access homoscedasticity by seeing the Residual vs Fitted plot. For our model, we do not see an apparent cloud formation and that can be a problem. Also, we see the Scale-Location plot and see that the observations are spread relatively equally above and below the horizontal line, which is not true for our model. To access multicollinearity, we run a cor.test using various combinations for the independent variables and see that we see a small correlation effect size for all. To see that errors are independent, we see the Residual vs Fitted plot and we do not see that the red fitted line is relatively flat and straight and thus limiting the predictions.

CONCLUSION

When testing if the entire model is statistically significant or not, we see that the model which predicts the preference of water using County, Age group, Falling sick because of tap water, Glasses of water, Taste of tap water, Trust in Tap water, Trust in bottled water, Taste of bottled water is statistically significant at alpha = 0.10. Thus, we conclude that the choice of source for drinking water is definitely affected by external factors and we have busted the myth to reject the null hypothesis and accept the alternate hypothesis, but there are assumptions like homoscedasticity, errors being independent, normality which are affecting our model.

LIMITATIONS

The survey has major limitations. Firstly, the survey is not a representative of the entire population of Maryland as it has a lot of data from the 4 main counties namely; Baltimore City, Baltimore County, Prince George’s County, and Montgomery County. Secondly, the maximum data is from the age group 25-34 and thus limiting our conclusions about the different age groups. Thirdly, there is a feeling of overfitting the model highly, meaning when the model will be applied to a lot of data from the real world, the model not give accurate results.

REFERENCES

Morgan, J. N., Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. Journal of the American statistical association, 58(302), 415-434.

Doria, M. F. (2006). Bottled water versus tap water: understanding consumers’ preferences. Journal of water and health, 4(2), 271-276.