**TEAM 10**

Contents

1. [INTRODUCTION 3](#_bookmark0)
2. [DATA OVERVIEW 3](#_bookmark1)

[Screener Questionnaire 3](#_bookmark2)

[Main Questionnaire 3](#_bookmark3)

1. [SUMMARY OF THE GIVEN RAW DATA 3](#_bookmark4)
2. [DATA PRE-PROCESSING 4](#_bookmark5)
3. [MODEL 6](#_bookmark6)
4. [TABLEAU DASHBOARD 7](#_bookmark7)

[Key performance measures 7](#_bookmark8)

[Fragrance attributes 11](#_bookmark9)

[Addiction, Associations & Appropriateness 12](#_bookmark10)

[Respondent demographics 16](#_bookmark11)

[ANNEXURE I 20](#_bookmark12)

[Logistic Regression 20](#_bookmark13)

[Penalized Logistic Regression 21](#_bookmark14)

[ANNEXURE II 22](#_bookmark15)

[Wrapper Method 22](#_bookmark16)

## INTRODUCTION

## DATA OVERVIEW

### Screener Questionnaire

### Main Questionnaire

## SUMMARY OF THE GIVEN RAW DATA

* + Total variables (columns) = 217
  + Total number of responses = 15217

1. Most of the features given in the data sheet were subparts of the question asked in the Main questionnaire. For example, the selection of the subpart of the question and its type (positive or negative) with respect to its nature.
2. The data sheet file consisted of two sheets.
   * Data
   * Data dictionary
3. Data dictionary- this part of the sheet consists the explanation of every variable/feature in the **main questionnaire**.

## DATA PRE-PROCESSING

* + The “instant like” was discretized as following:

1. like - 1
2. not like - 0
   * The feature for Q3 “strength of the deodorant” is given on the scale of 1 to 5. We rescaled it as following:
3. 3 (“just about right”) -> 1 as it is a positive attribute
4. All other ratings (1, 2, 4, 5) -> 0
   * All the features having the string data types are removed before using the data for final processing because every string value is classified as positive or negative in the subsequent question.
   * The feature ‘*q1\_1 personal opinion of this Deodorant*’ was perfectly correlated with the “instant like” as for rating 5, 6 and 7, Instant Liking was set as *Like* and for ratings 1, 2, 3 and 4, Instant Liking was set as *Not Like* so we trained our model after removing it, otherwise the competition would have been irrelevant.
   * We created a new feature by combining the answers of question 2 in the main questionnaire where 5 words which describe the scent of deodorant were classified as positive or negative. In the data set, positive words were assigned as 1 and negative words were assigned as 2. In our new feature we took the number of positive features and made a new feature ‘q2positive’ which ranged from 0 to 5.
   * In question 4, various attributes of scent of deodorant are rated from 1 to 5 where 1 implies disagree completely and 5 implies agree completely. It consists of both positive and negative attributes so we need to separate all positive and negative attributes. For doing so we calculated the correlation of one attribute ‘attractive’ with all the other attributes and the positive correlation implies that the attribute is positive as ‘attractive’ is a positive attribute and negative correlation implies that the attribute is negative. We were able to separate positive and negative attributes very effectively.

Positive attributes include

* + - 'q4\_2 attractive’,

● 'q4\_3 bold ',

* + - 'q4\_5 casual ',
    - 'q4\_7 clean',
    - 'q4\_8 easy to wear',
    - 'q4\_9 elegant',
    - 'q4\_10 feminine',
    - 'q4\_11 for someone like me',
    - 'q4\_13 high quality',
    - 'q4\_14 long lasting',
    - 'q4\_16 memorable ',
    - 'q4\_17 natural ',

● 'q4\_21 sharp',

* + - 'q4\_22 sophisticated',
    - 'q4\_23 upscale',
    - 'q4\_24 well rounded '

Negative attributes include

* + - 'q4\_1 artificial/chemical',
    - 'q4\_4 boring',
    - 'q4\_6 cheap ',

● 'q4\_12 heavy ',

* + - 'q4\_15 masculine',
    - 'q4\_18 old-fashioned',
    - 'q4\_19 ordinary',
    - 'q4\_20 overpowering’
* Two new features were made corresponding to positive and negative attributes i.e.

‘q4\_positive’ and ‘q4\_negative’

In ‘q4\_positive’ the average rating of all the positive features were taken and in ‘q4\_negative’ the average rating of all the negative features were taken.

* For combining Q6 of the main questionnaire we wanted to separate positive and negative characteristics of the scent of the deodorant. For doing so we saw the corresponding column of Q7 where each characteristic was described as positive or negative. If the number of positives are more than negative, then we classified that characteristic to be positive otherwise it was classified as negative. By this we got all the positive and

negative characteristics of the scent of the deodorant and we made two new features i.e.

‘q6\_positive’ and ‘q6\_negative’.

In ‘q6\_positive’ we added all the positive features and in ‘q6\_negative’ we added all the

negative features.

* + Features of Q7 were combined as the percentage of positive characteristics of all the characteristics selected by the respondent. The missing values were taken as 0 as some characteristics were not selected by the respondent giving survey.
  + For age we have considered “*s2b.coded.age”*.
  + The screener questions 13 was combined by adding the number of deodorants worn regularly be user in past 6 months out of 50 deodorants listed in the survey.
  + Question 13(a) of the screener questionnaire gives us the most often used deodorant of the user. A new feature was made which was set 1 if the most often used deodorant of the person is same for which the person is giving survey. Otherwise it is set to be 0.
  + All the redundant features with 0 variance in the data were removed.

## MODEL

In our model we have used L2 penalized Logistic regression method.

1. We reduced the dimensionality of the data while preprocessing. Low dimensionality prevent overfitting and gives better estimates. Also it is easier to visualize the data with low dimensions and the time required to train the data.
2. We created five smaller datasets for the relevant deodorants (B, F, G, H, J).
3. We fit a fused lasso regularized logistic regression model for each of the dataset on the dataset. We fit the model with five different values of **regularization parameters** and we got the best fit for regularization value 0.01. varied five times. We got 5 accuracy scores,

a) B=0.763 b) F=0.755 c) G=0.761 d) H=0.756 e) J=0.758

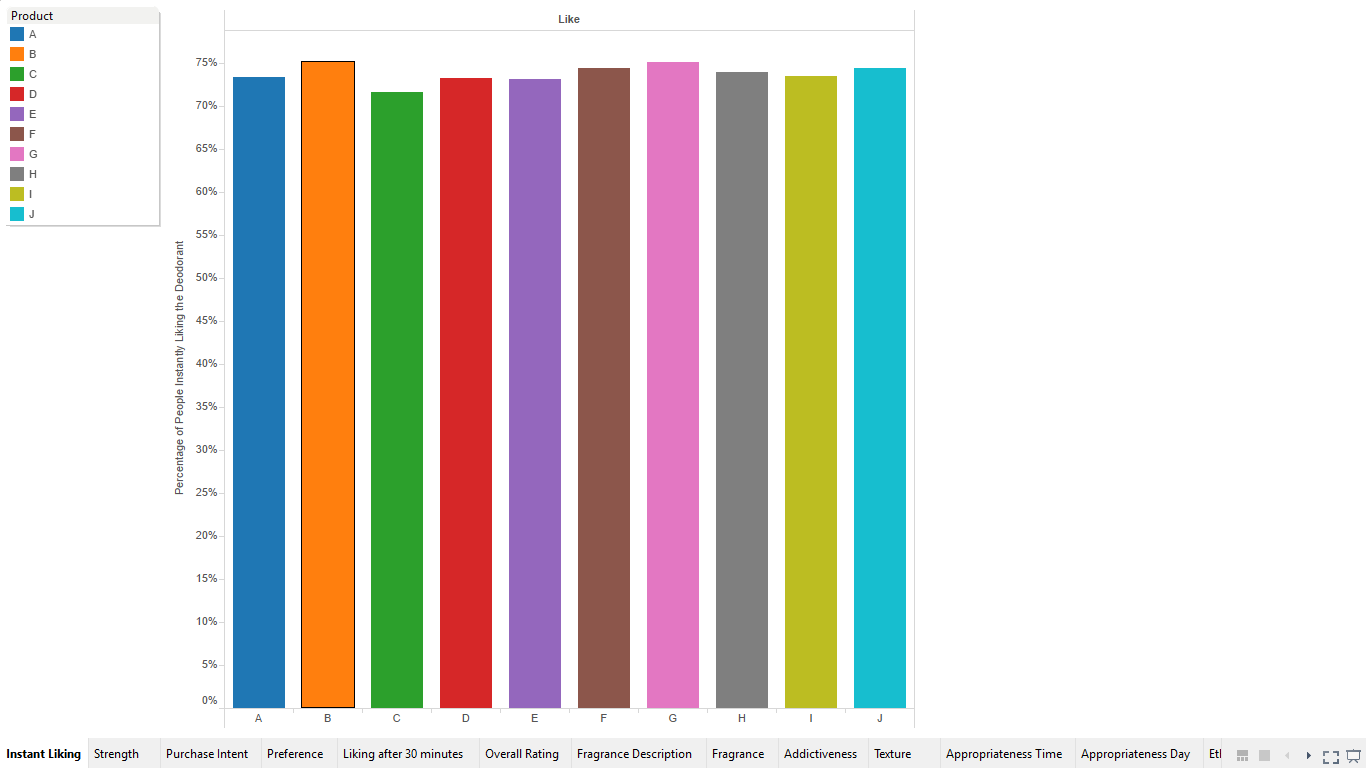
1. So, we applied wrapper method and then tried to combine the variables.

## TABLEAU DASHBOARD

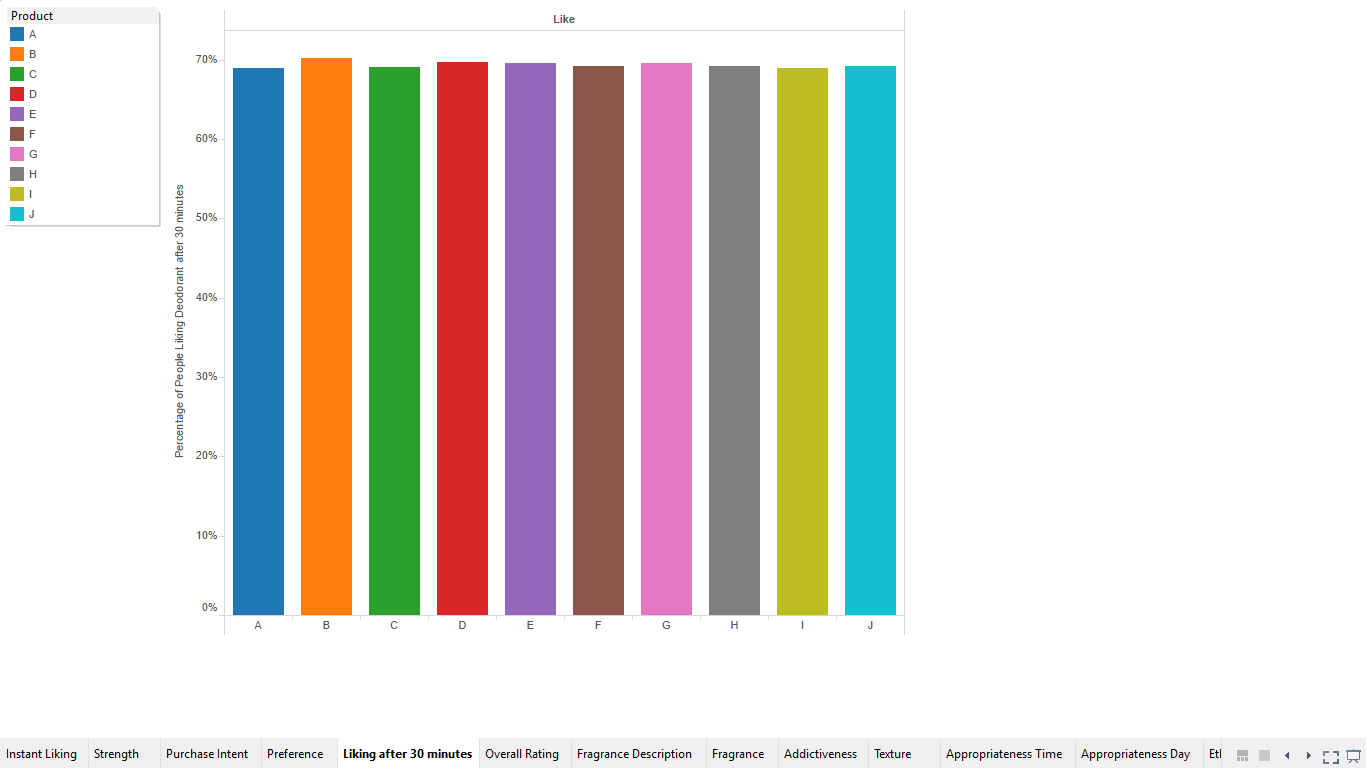
We need to present the data from the survey in form of an interactive dashboard. The aim of the dashboard is to present factors of all the ten deodorants to a business client so that she/he can improve the reception and sales of these deodorants. **Please use the cleaned data present inside the tableau folder to use the visualizations.**

We display four dashboards in the tableau file as follows:

### Key performance measures



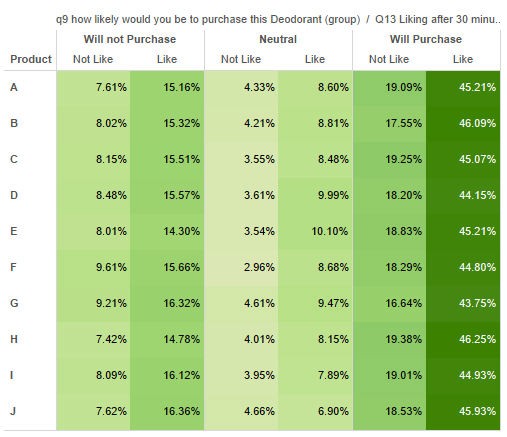
In this bar graph we show, percentage of people instantly liking the deodorants and we find that Deodorant B is the most liked deodorant. Note that the percentage here means that if you take a sample of 100 people, the percentage value denotes the number of people instantly liking the deodorant.



In this bar graph we show, percentage of people liking the deodorant after 30 minutes and we find that Deodorant B is the most liked deodorant. Note that the percentage here means that if you take a sample of 100 people, the percentage value denotes the number of people liking the deodorant after 30 minutes.

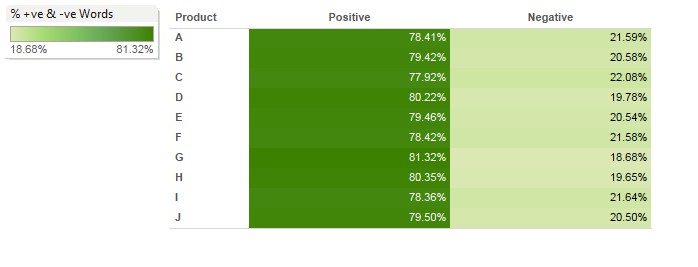


The above heat map represents the change of liking status after 30 minutes. We note that almost 50% like the deodorant both instantly and after 50 minutes, but around 20% of the people disliked the deodorant 30 minutes of use even though they liked they had “instant liked” the deodorant. We need to target this set of people.

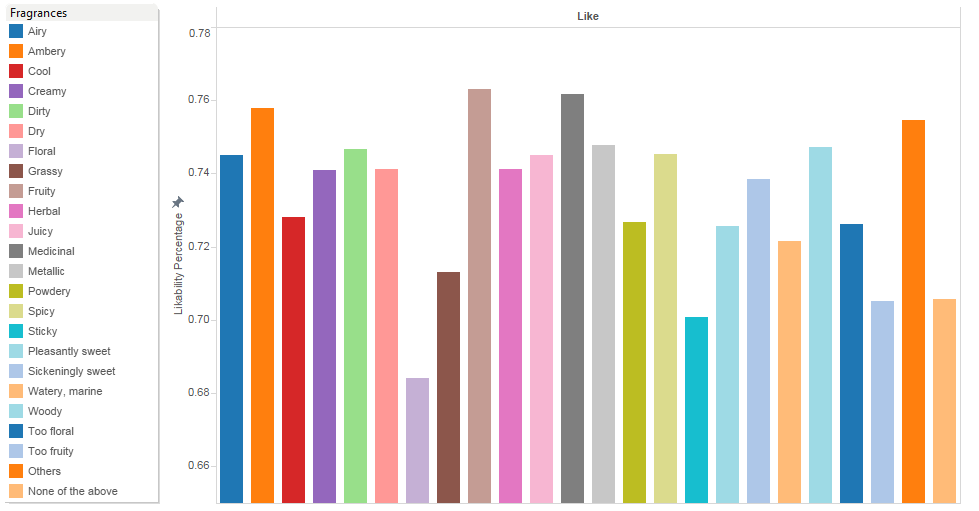


The above heat map represents the purchase intent for the deodorant in relation to their instant like status. We note that they are around 15% of the respondents who said that they won’t purchase the deodorant even though they liked it.

### Fragrance attributes

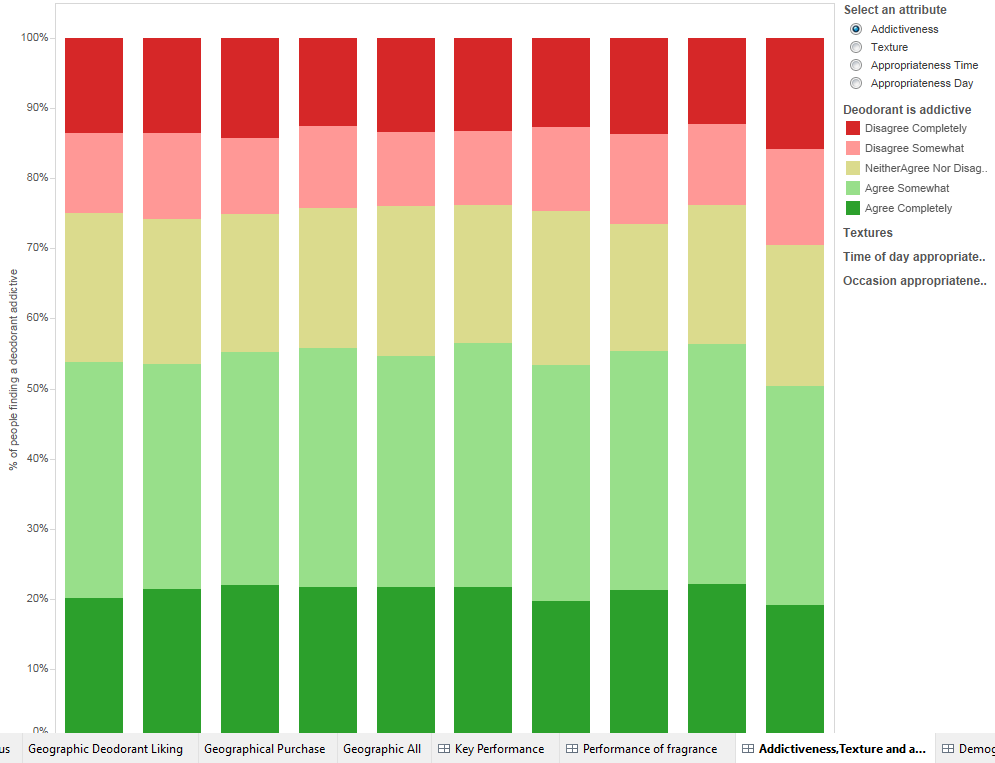


We present here the percentage of positive and negative reviews given by the respondents based on Question 4 in the questionnaire for each of the deodorants.

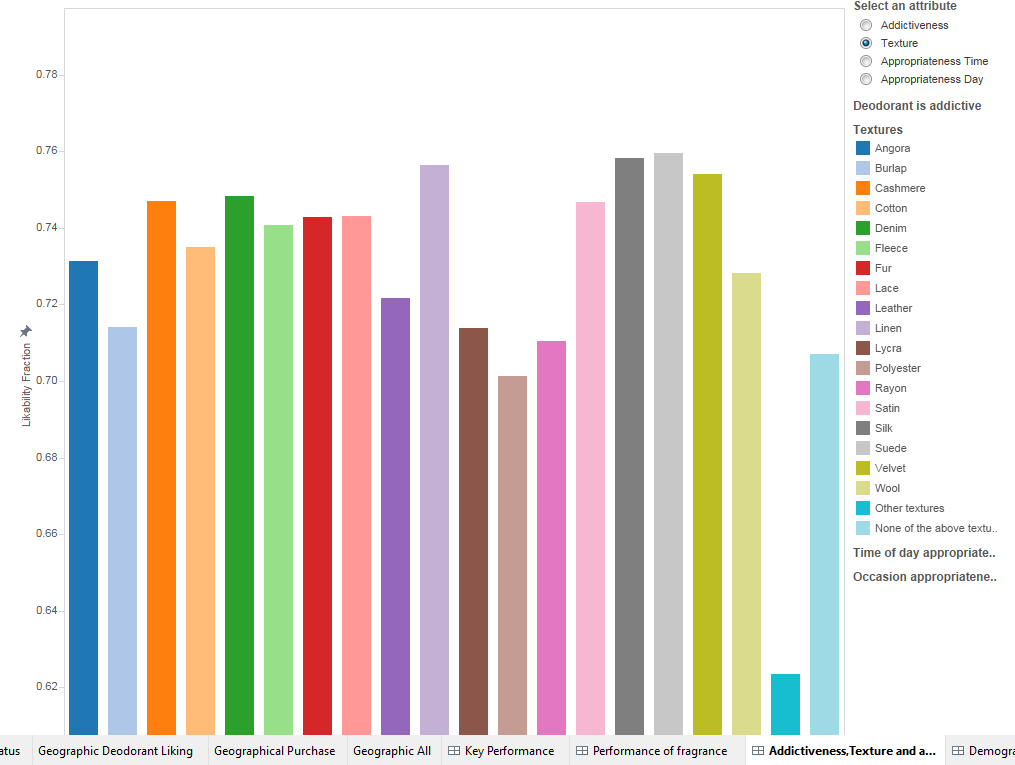


The above bar graph presents the likability percentage of different fragrances. We find that Ambery, Fruity and Medicinal are the most liked fragrances. Floral, Sticky and Too Fruity fragrances are least liked. Note that the percentage here means that if you take a sample of 100 people, the percentage value denotes the number of people liking a particular fragrance.

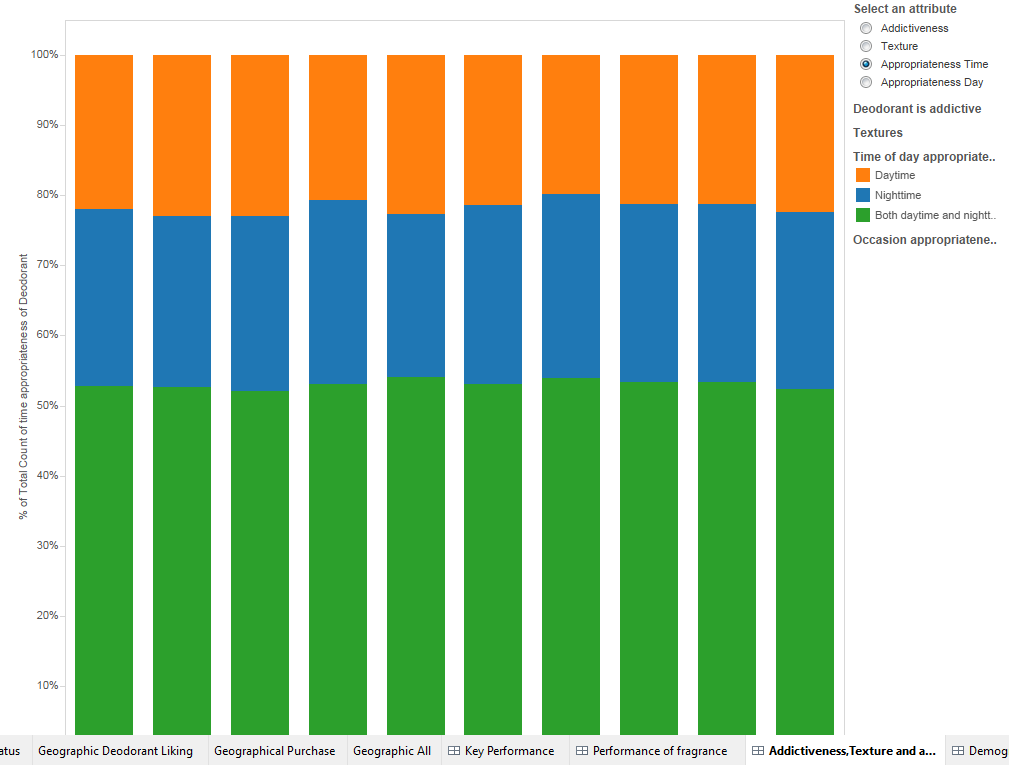
### Addiction, Associations & Appropriateness



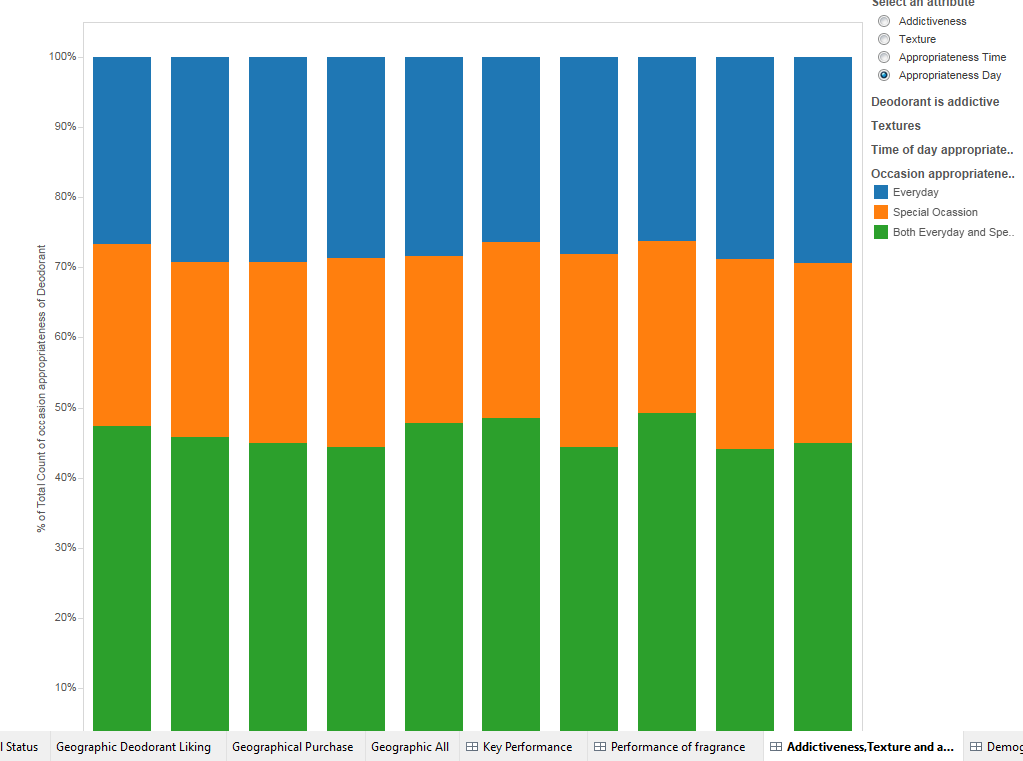
The above stacked graph represents the addictiveness of various deodorants. People found Deodorant F to be the most addictive deodorant and the J to be the least addictive deodorant.



The above bar graph presents the likability percentage of different textures. We find that Silk, Suede and Linen are the most liked textures. Polyester, Rayon, Lycra and Burlap are least liked textures. Note that the percentage here means that if you take a sample of 100 people, the percentage value denotes the number of people liking a particular texture.

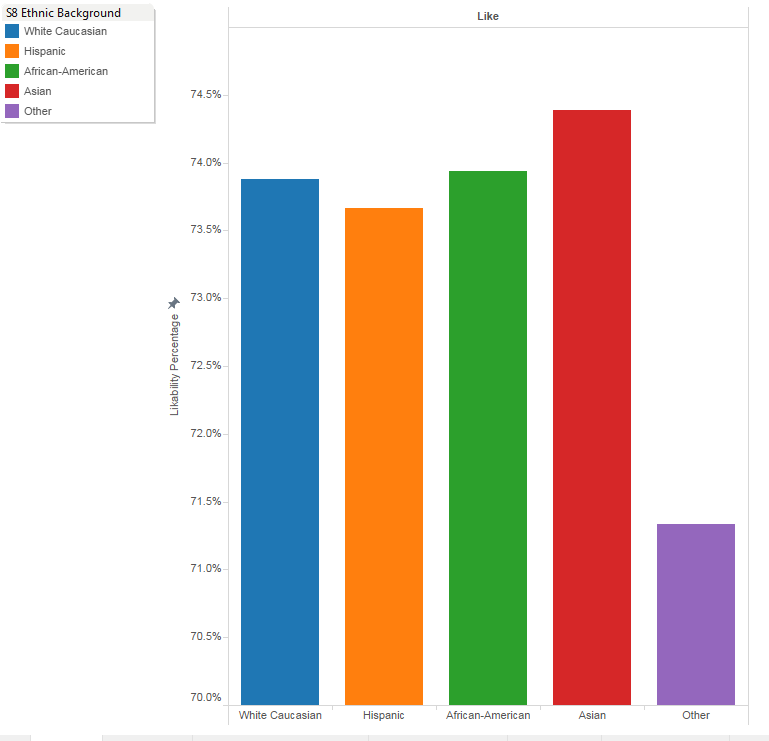


The above stacked graph represents the time at people find it appropriate to use the deodorant. Most people prefer to use the deodorant at both daytime and night. There is no strong dependence of appropriateness on the deodorant used.

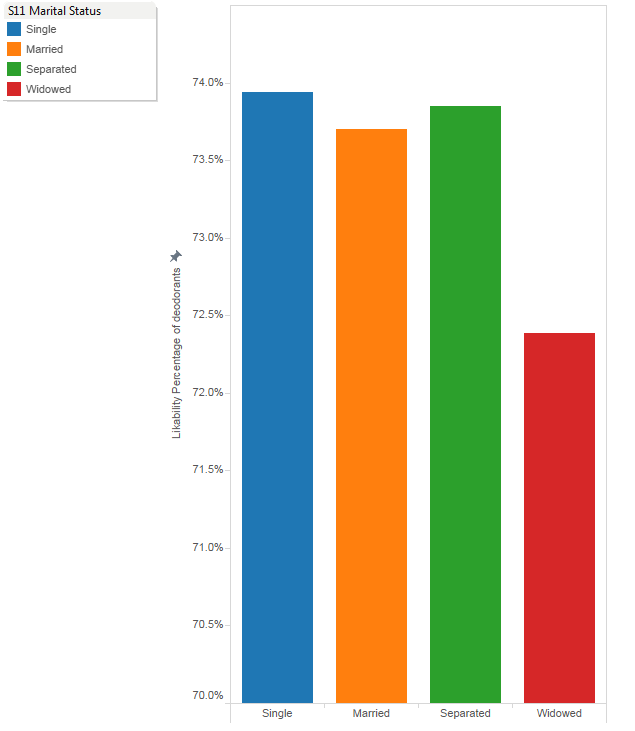


The above stacked graph shows the occasions on which people find it appropriate to use the deodorant. Most people like to use deodorant on both special occasion and on everyday basis.

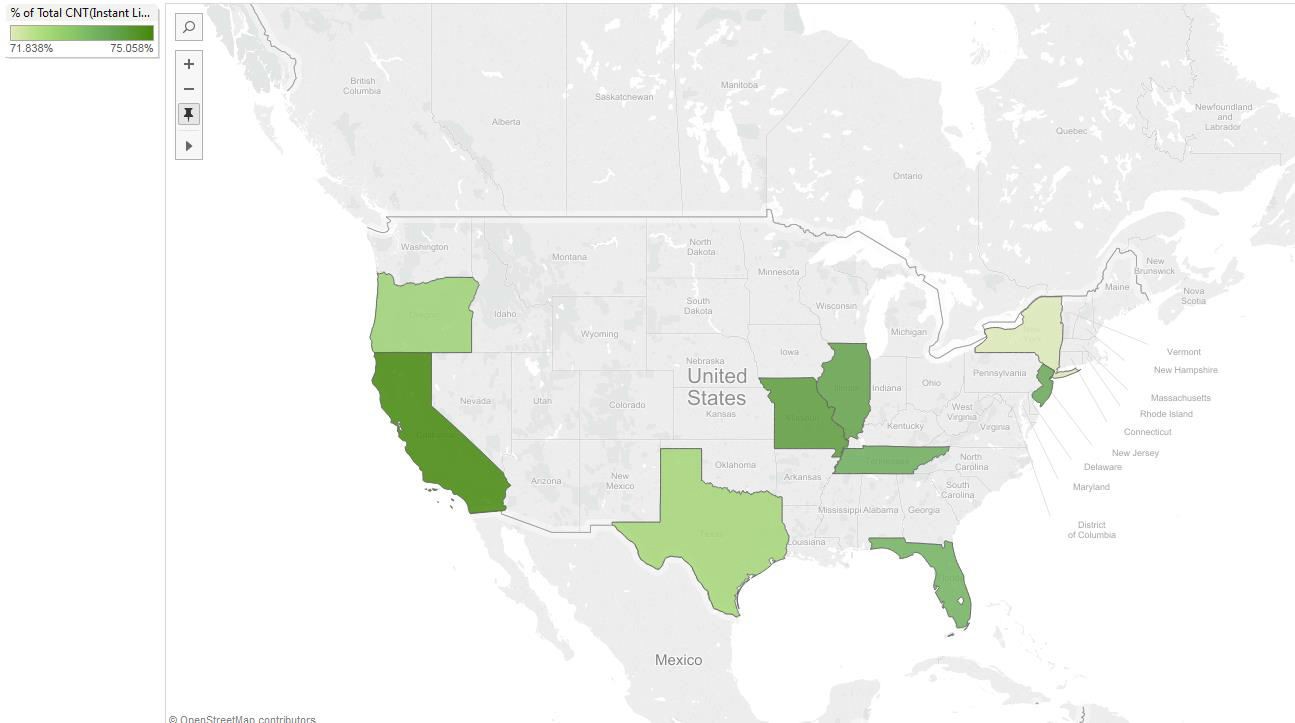
### Respondent demographics



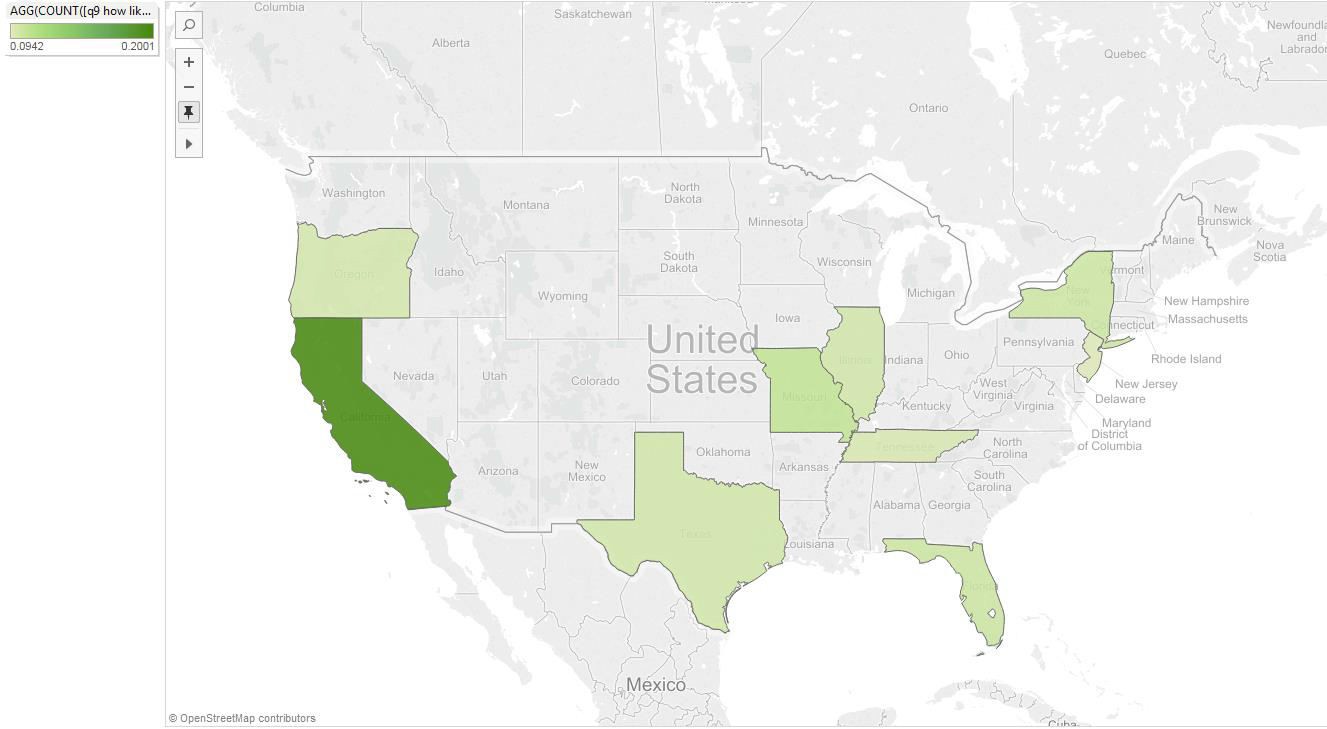
The above bar graph represents the likability percentage by the ethnic group of the respondents. The deodorants are well liked by Asians and we need to target Hispanic and other minority communities.



The above bar graph represents the likability percentage by the marital status of the respondents. The deodorants are well liked by Single and Separated people as expected.



Above graph represent the instant likability of our deodorants by the geographic location of the respondents. Our deodorants are more liked in California and Missouri. We need to spend more energy targeting people in New York and Texas.



Above graph represent the purchase intent of our deodorants by the geographic location of the respondents. Our deodorants are more likely to be purchased in California. Even though people in Missouri and Illinois like the deodorant they don’t show strong purchase intent.

**ANNEXURE**

# ANNEXURE I

### Logistic Regression

Logistic regression provides a good method for classification by modeling the probability of membership of a class with transforms of linear combinations of explanatory variables.

Let Yj indicate the status of array j, j = 1, . . . , n and let xij , i = 1, . . . , m be the normalized and scaled expressions of the m genes on that array. Imagine that one specific gene has been selected as a good candidate for discrimination between the two classes (say y = 0 and y = 1). If X is the expression measured for this gene, let p ( x ) be the probability that an array with measured expression X = x represents a class of type Y = 1.

A simple regression model would be

p ( x) = α + β x

with α and β estimated with a standard linear regression procedure.

Not a good idea!

* + p ( x ) may be estimated with negative values
  + p ( x ) may be larger than 1

Solution

Transform p(x) to η(x):

η(x) = log ( p(x) / (1 − p(x)) ) = α + βx.

The curve that computes p(x) from η(x), p(x) = 1 /( 1 + exp(−η(x)) ) is called the logistic curve. This a special case of the generalized linear model. Fast and stable algorithms to estimate the parameters exist (glm package in R).

It is straightforward to extend the model with more variables (genes expressions), introducing explanatory variables x 1 , x 2, . . . , x p:

η ( x 1 , x 2, . . . , x p) = log p ( x ) / (1 − p ( x )) = α + p ∑ βixi

The maximum likelihood algorithm for glm’s can be extended to this case very

easily.

### Penalized Logistic Regression

L1 and L2 penalized estimation methods shrink the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates. The purpose of this shrinkage is to prevent overfit arising due to either collinearity of the covariates or high-dimensionality. Although both methods are shrinkage methods, the effects of L1 and L2 penalization are quite different in practice.

Applying an L2 penalty tends to result in all small but non-zero regression coefficients, whereas applying an L1 penalty tends to result in many regression coefficients shrunk exactly to zero and a few other regression coefficients with comparatively little shrinkage. Combining L1 and L2 penalties tends to give a result in between, with fewer regression coefficients set to zero than in a pure L1 setting, and more shrinkage of the other coefficients. The fused lasso penalty, an extension of the lasso penalty, encourages sparsity of the coefficients and their differences by penalizing the L1-norm for both of them at the same time, thus producing sparse and piecewise constant stretches of non-zero coefficients. The amount of shrinkage is determined by tuning parameters λ1 and λ2. A value of

zero always means no shrinkage (= maximum likelihood estimation) and a value of infinity means infinite shrinkage (= setting all regression coefficients to zero)

# ANNEXURE II

### Wrapper Method

### In wrapper method, we use hit and trial method of including the variable/feature in the logistic regression model. If the variable’s contribution to the accuracy is positive the variable is kept in the model. Similarly, all other combinations of variables are applied until only the useful contributing variables are left in the model.

