Homework 3

Building a Reporting Engine

All of the previous homework assignments have asked you to perform analyses while writing a report. This time, you will build a dynamic reporting system that can display a wider range of information. Each of the questions below will include an analytical component. Then you will build a section of a reporting engine that can answer a whole class of similar questions.

The analytical questions may be written up in the usual style of a report. We will also ask you to turn in your reporting engine as an RMarkdown file.

About The Data

We will be working with a simulated data set related to market research surveys for mobile phone products.

Main File: Homework 3 Data.csv

Delimiter: Each column of each file is separated with a comma **,** delimiter.

Header The first row of the data set includes the column names, and each subsequent row includes one observation of values. Here is a selection of 20 randomly sampled lines from the data set:

Show 10 entries Search:

id	Age	Gender	Income	Region	Persona	Product	Awareness	BP_User_Friendly_0_1
6351	20	Female	20000	Northeast	Ambivalent Adventurer	MobilitEE	0	
7937	71	Female	62000	South	Consistent Compromiser	Off the Hook	0	
412	63	Male	40000	Northeast	Ambivalent Adventurer	Ring Ring	0	
6090	29	Female	84000	West	Materialistic Meditator	Smartophonic	1	
4323	32	Male	43000	South	Ambivalent Adventurer	Phone Zone	1	
761	37	Male	85000	Midwest	Ambivalent Adventurer	MobilitEE	0	
5591	33	Female	88000	South	Outdoorsy Ombudsman	Speed Dials	0	
4564	65	Female	30000	Midwest	Ambivalent Adventurer	Phone Zone	0	
5363	72	Male	67000	Midwest	Ambivalent Adventurer	MobilitEE	0	

id	Age	Gender	Income	Region	Persona	Product	Awareness	BP_User_Friendly_0_1
5577	58	Female	52000	Northeast	Technological Triumphalist	Phone Zone	0	

Showing 1 to 10 of 20 entries

Previous 1 2 Next

Your organization's market research team created a survey to collect information about the customer base. A large, representative sample of customers was surveyed. Each row of the data set records the information for a single respondent's reactions about a single product. The data are organized in long, melted format. Each person in multiple rows, with one for each product. The Main File includes the following variables:

- **id**: This is a unique identifier for the respondent. The data are structured in a **melted** format. Each person's responses show up in multiple rows, with 1 row for each product.
- **Age**: This is the subject's age in years (rounded down) at the time of survey. For the purpose of this study, all of the respondents should be at least 18 years old. A number of questions will ask you to categorize the respondents into the following groups based on their age:
- **Age Groups**:
  - At least 18 and under 35. (Don't include anyone who is 35.)
  - At least 35 and under 50.
  - At least 50 and under 65.
  - At least 65.
- **Gender**: This identifies the respondent's gender as Male or Female.
- **Income**: This is the respondent's household income – the combined income of all members of the household – rounded to the nearest thousand dollars. A number of questions will ask you to categorize the respondents into the following groups based on their income:
- **Income Group**:
  - Under $50,000.
  - At least $50,000 and under $75,000.
  - At least $75,000 and under $100,000.
  - At least $100,000 and under $150,000.
  - At least $150,000.
- **Region**: This is the geographial region within the U.S.A. in which the respondent lives.
- **Persona**: This is the respondent's marketing profile category. These were created previously by the marketing organization as a method of dividing the respondents into a number of illustrative groups.
- **Product**: This is the name of each brand of mobile phone that was surveyed.
- **Brand Perceptions**: There are a number of variables about the respondent's perceptions of the brands. Each of these variables is labeled with the form **BP_quality_min_max**. The word or phrase used in place of the quality is the perception that was surveyed. The respondents were asked to rate that perception on an integer scale from the minimum to the maximum listed values.
- **Outcomes**: These are the marketing states of engagement that the survey was designed to investigate. The outcomes include Awareness, Consideration, Consumption, Satisfaction, and Advocacy. Satisfaction was assessed on an integer scale from 0 to 10. All of the other outcomes are binary variables. For the purposes of this assignment, it would be reasonable to place all of the outcomes on a percentage scale from 0 to 100.

Note: A dynamic progression of the questions in the survey was utilized. Those not aware of a product were not asked about any further states of engagement. Those who were aware were asked about their perception of the brand and also their consideration. Those who had considered the product were asked about their consumption. Those who had consumed the product were asked about both their satisfaction and advocaccy. Any questions that were not asked should result in missing (NA) values for the record.

Note: The description above tells you *the intended structure* of the data set. However, it's possible that there could be problems lurking in the records. In the course of doing this assignment, you may uncover some issues. For instance, you may find an erroneous value. In this circumstance, it will be necessary to resolve the situation. Here are some guidelines for doing so:

- If the issue has an obvious solution, then you may recode the data. For instance, if you see a value of **"True"** for a binary variable, then you may safely assume that this value should have been coded as a 1.
- If the issue does not have an obvious solution, then you can replace the erroneous value with **NA** to denote a missing value.

In either circumstance, note the problem in your solution and briefly describe the work you did to clean the data.

Then, use the data to answer the following questions and to build a reporting engine according to the specifications described.

Question 1: Respondent Variables

a. In percentage terms, how were the survey's respondents divided into categories for the following variables? Answer separately for each variable. Round all percentages to 1 decimal place (e.g. 84.2%).

Hint: Keep in mind that each respondent may appear multiple times in the data set.

- **Age Group**
- **Gender**
- **Income Group**
- **Region**:
- **Persona**

b. Now create a visual display of this information. Allow the user to select which variable to explore. Then create a graph that depicts the percentages of respondents in each category for that variable.

Question 2: Segmented Outcomes

a. What are the top 5 products by Awareness rates in the Northeast? Round the percentages to 1 decimal place, e.g. 84.2%.

b. What are the top 5 products by Advocacy rates among females who earn at least $100,000? Round the percentages to 1 decimal place, e.g. 84.2%.

c. Now create a dynamic, visual display ranking the products by their outcomes. The user will make the following selections:

State of engagement: Only a single state may be selected at once.

Other variables: Age Group, Gender, Income Group, Region, Persona

Then, for all of the other variables, any combination of categories may be selected, so long as at least one category from each variable is chosen. For instance, for Gender, the user may select Male only, Female only, or both Male and Female.

Then, the user should be able to select how many products to display. Once a number is selected, the outcome rates should be graphically displayed in sorted decreasing order for the top products in the selected subgroups. If 5 is selected for Awareness, then the 5 products with the highest rates of Awareness for the specified subgroup will be depicted. Make sure to include the percentages in the graph, each rounded to 1 decimal place (e.g. 84.2%).

Question 3: Overall Brand Perceptions

a. What are the top 5 brands by the overall average perception?

Evaluating this question can be tricky. Some of the perceptions are for positive traits, and others are for negative traits. The brand with the best overall perception would have the highest scores for the positive traits and the lowest scores for the negative traits. To aggregate these scores, we will follow a number of steps:

1. For each brand, compute the average score of each brand perception variable. In computing these averages, remove any missing values from the calculations.
2. Then, for the negative perceptions, invert the scores to place them on a comparable scale with the positive traits. To do this, use the conversion formula:

Inverted Score = min possible score + max possible score - recorded score = 10 - recorded score.

The minimum and maximum possible scores here are 0 and 10. Therefore, the inverted average score is:

Inverted Average Score = 10 - Average Score.

3. With all of the average scores of each perception now recorded on the same scale, we can aggregate them into one measure, the Overall Average Perception. For each brand, compute the mean of these variable averages. (To be clear: within a single product, you can add up the average scores for each perception and then divide by the number of perceptions.)
4. Now rank the brands in decreasing order of their Overall Average Perception scores.
5. Show the results for the top 5 brands.

b. Now create a dynamic, graphical display that allows the user to perform this calculation in selected subgroups. Much like the previous question, the user may make any combination of selections in the following variables, provided that at least one category of each variable is selected: Age Group, Gender, Income Group, Region, Persona.

Also allow the user to select how many brands should be displayed, with the top k brands depicted in decreasing sorted order. All results should display the overall average perception for the brand, rounded to 1 decimal place (e.g. 6.1).

Question 4: Outcomes Gaps

The marketing department wants to identify products with engagement that is underperforming in some ways. The best products should have high rates of engagement across all of the outomes, but that is not always the case.

For the purposes of this question, we will work with the average rate of each state of engagement. To ensure a fair comparison, we will place all of the outcomes on a percentage scale from 0 to 100. For binary outcomes (awareness, consideration, consumption, and advocacy), the average will be the percentage of the respondents who answered yes to the question among those who were asked. For outcomes on an integer scale (e.g. Satisfaction), the average will be percentage of the maximum score. So, for instance, if the average satisfaction for a product is 7, then its percentage rating would be 70%.

a. Which 5 products have the largest gap between the rate of consumption and the rate of awareness? This would correspond to a formula of Difference = Rate of Consumption - Rate of Awareness. Products with higher rates of awareness than the corresponding rates of consumption will have negative differences. Display a bar graph showing the 5 largest differences in decreasing sorted order. Include the differences as percentages rounded to 1 decimal place (e.g. 84.2%).

b. Which 5 products have the largest gap between the rate of awareness and the average satisfaction (in percentage terms)? Here the formula would be Difference = Rate of Awareness - Percentage Average Satisfaction. Display a bar graph showing the 5 largest differences in decreasing sorted order. Include the differences as percentages rounded to 1 decimal place (e.g. 84.2%).

c. Now create a dynamic, graphical display that ranks the products in terms of the difference in averages between any two selected outcomes. The user will be allowed to make the following selections:

First Outcome: One of the outcome variables.

Second Outcome: Another outcome variable. In practice, it would be nice to exclude the outcome that was selected first. In practice, that requires some additional programming tools. So it's OK to select the same variable twice. In that case, all of the products should necessarily show a difference of zero.

The difference in rates will be Difference = Average First Outcome - Average Second Outcome per product.

Number of Top Products: The user will select how many products to display.

Display Percentages: If checked, the bargraph will display the percentages for each product.

Digits: How many digits should the percentages be rounded to? 1 digit would be a number like 84.2%.

Question 5: Cross-Product Measures

How much does a respondent's engagement depend on the product, and how much depends on the respondent? One way we might investigate this further is to see whether the respondent's outcomes in other products has an impact on this one. We will investigate this by the following steps:

a. How much impact does respondent's overall trends in awareness have for that person's awareness with Buzzdial phones? To answer this question, we want to create a logistic regression model. The outcome will be the respondents' Awareness of Buzzdial. The variables in the model will include age group, gender, income group, region, persona, and the **aggregated awareness**. The aggregated awareness will be the average of the respondent's awareness scores for all of the products *except for Buzzdial*. Each respondent will have a different value of aggregated awareness. Any missing scores should be removed from the calculation of the aggregated awareness. Then, fit the logistic regression model. Display a table including the model's Odds Ratios, 95% confidence intervals for the Odds Ratios, and the p-values. In particular, show these values for the aggregated awareness variable and comment on the results. Round all of the results to 3 decimal places.

b. How much impact does respondent's overall trends in satisfaction have for that person's satisfaction with Buzzdial phones? To answer this question, we want to create a linear regression model. The outcome will be the respondents' Satisfaction with Buzzdial. The variables in the model will include age group, gender, income group, region, persona, and the **aggregated satisfaction**. The aggregated satisfaction will be the average of the respondent's satisfaction scores for all of the products *except for Buzzdial*. Each respondent will have a

different value of aggregated satisfaction. Any missing scores should be removed from consideration. Then, fit the linear regression model. Display a table including the model's coefficients, 95% confidence intervals for the coefficients, and the p-values. In particular, show these values for the aggregated satisfaction variable and comment on the results. Round all of the results to 3 decimal places.

c. Now we will create a dynamic model that allows the user to build a model including an aggregated outcome for a specific product. The site should include the following features:

- The user can select the product.
- The user can select the state of engagement as the outcome.
- The user can select the other variables to include in the model. The list of choices should include the age group, gender, income group, region, persona, brand perceptions, and the Aggregated Engagement. Each person's aggregated engagement will be calculated as the average score of the selected state of engagement across the measured values of the other products . You can give this variable a name like "Aggregated.Engagement".

The user's selections will then be incorporated into a model. For Satisfaction outcomes, use a linear regression. For all of the other outcomes, use a logistic regression. Then create a dynamic table showing the model's results. For logistic regressions, this must include the Odds Ratios, 95% confidence intervals for the Odds ratios, and the p-values. For linear regressions, this must include the coeffiients, 95% confidence intervals for the coefficients, and the p-values. Other factors may be included but are not necessary. Round all of the results to 3 decimal places.

Completing the Assignment

Each of the questions include two different components:

- Building a **reporting engine** that will display a range of information based on the user's selections.
- Answering **specific questions** based on a single set of selections to the reporting engine.

To best facilitate this work, we are asking you to generate output in two different files:

- An Rmarkdown application (using shiny or flexdashboard) that can be run in real time.
- A static report built in RMarkdown (e.g. an HTML file) that answers the specific questions.

Template files for each of these components will be provided. This is also a good opportunity to design your code to be reusable in both applications. Given the code that can generate the reporting engine, each of the specific questions can subsequently be answered in a single line of code (e.g. by calling a function of your design).

The submission will require the following files:

- Reporting Engine (in RMarkdown format);
- Source Code for the Static Report (in RMarkdown format);
- Output of the Static Report (HTML preferred).