**MIDTERM PROJECT**

**The data results from an advertising campaign of a large travel agency.**

**You have access to two datasets**

**Adandoned.csv (ABD)**

**Reservation.csv (RES)**

**Introduction**

The first contains demographic information about customers who engaged with the travel agency, but that did not buy a vacation package. You may notice the sparseness/missingness of the data and also the presence of potential duplicates.

These customers were randomly assigned to a retargeting campaign; you can see their status in the “Test\_Control” column.

Sometime later, data about reservations were pulled from the system, so have obtained the “Reservation.csv” File. Note that you will just see those who bought a package, whether in the treatment or control group, or neither.

You are generally tasked with matching the data across the files to test whether the retargeting campaign has worked statistically.

In doing so, you will have to make assumptions about reconciling the records (missing info, matching rules, duplicates, etc.)

There is no solution, but the instructor knows the reasonable range of results we should expect.

Delivery: you can use this doc as a template, you are also expected to upload the Rscript that runs all the analyses top to bottom, and the cleaned data (csv format preferred) that you used to answer the questions.

Note: this project is individual based, you can discuss with your peers, but you are expected to disclose all the people you collaborated with.

**I. The Business Problem**

ABD contains data for all the customers in the dataset that were already pursued (advertised) but did not buy a vacation package.

Business Problem: Should we retarget those customers?

**Q1:** In light of your experience as a businesswoman/man, argue why this is a sensible business question.

An experiment is run where customers in the abandoned dataset are randomly placed in a treatment or a control group (see column L in both files).

Those marked as “test” are retargeted (treated); the others marked as control are part of the control group.

**Q2:** Investigate the test/control variable. Does the experiment seem to be run properly?

**Q3:** compute the same summary statistics for this Test\_variable by stratifying on States (meaning considering only the entries with known “State”), wherever this information is available.

**II. Data Matching**

About three months later, the experiment/retargeting campaign is over.

Customers, presented in the ABD excel file, who bought vacation packages during the time frame, are recorded in the RS excel file.

**Q5:** After observing the data in both files, argue that customers can be matched across some “data keys” (column labels). Correctly identify all these data keys (feel free to add a few clarifying examples if needed)

**Q6**: EXTREMELY CAREFULLY DESCRIBE YOUR DATA MATCHING PROCEDURE to IDENTIFY: (1) Customers in the TREATMENT group who bought (2) Customers in the TREATMENT group who did not buy (3) Customers in the Control group who bought, and (4) Customers in the Control group who did not buy. Be as precise as possible.

**Q7:** Are there problematic cases? i.e. data records not matchable? If so, provide a few examples and toss those cases out of the analysis.

**Q8: Complete the following cross-tabulation:**

|  |  |  |
| --- | --- | --- |
| **Group \ Outcome** | **Buy** | **No Buy** |
| **Treatment** |  |  |
| **Control** |  |  |

**Q9: Repeat Q8 for 5 randomly picked states. Report 5 different tables by specifying the states you “randomly picked”.**

**III. Data Cleaning:**

You have now identified all the relevant customers for the analysis and their outcome, and you also know if they are in a treated or in a control group.

Produce an Excel File with the following columns

Customer ID | Test Variable | Outcome | D\_State | D\_Email |

Where Test Variable indicates the treatment or the control group, the Outcome is a binary variable indicating whether a vacation package was ultimately bought. D\_State and D\_Email identify whether the information is present on file.

(Note that you should have as many rows as customers you were able to match across the two data sets. Be sure to attach this excel file to the submission for proper verification.)

**IV. Statistical Analysis**

We are finally in a condition to try to answer the relevant business question.

**Q10:** Run a Linear regression model for

Outcome = alpha + beta \* Test\_Variable + error

And Report the output.

**Q11:** Argue this is statistically equivalent to an ANOVA/t-test.

**Q12:** Argue whether this is a properly specified linear regression model, if so if we can draw any causal statement about the effectiveness of the retargeting campaign. Is this statistically significant?

**Q13:** Now add the dummies for State and Emails to the regression model. Also consider including interactions with the treatment. Report the outcome and comment on the results. (You can compare with Q9)

**VI: Conclusion**

**Q14: Lesson Learned. What would you have done differently in designing the experiment? Any other directions you could have taken with better data? Are there any prescriptive managerial implications of this study? Please answer briefly**

**Q15: Self-evaluation. Please score your effort on a scale 0-100. Please score your expected performance on the same scale. Add comments if necessary, including whether you collaborate with your peers.**