General Information

The goal of this assignment is to familiarize yourself with quantitative methods and statistics in practice. In the course of this assignment, you will parse and analyze log files of an online computer parts retailer. You will also learn to interpret these values to draw conclusions based on quantitative insights.

Introduction

You work for a regional online computer parts retailer 'ComVille' in Fresno, California and are responsible for the technical infrastructure. From January 1st, 2023 the California Consumer Privacy Act (CCPA) has become effective and you now realize that it requires a "Do not sell" opt-out option for all customers.

Since you have a background in privacy, you want to go above and beyond this requirement by showing customers a 90 second long video that explains which data your company collects, how it is used, and under which circumstances you transmit them to third parties.

You go to the CEO of the company and pitch what you want to do. However, the CEO isn't convinced about your plan and tells you that (a) she would rather only implement the new law as minimally as required, and (b) you may only change the checkout procedure if you can demonstrate that it does not affect the sales. Since the CEO attended a class in statistics in college, she asks you for a detailed statistical report until March 22nd, 2023, 09:59 CEST, to check if your proposed strategy affect sales.

In order to eliminate the concerns of the CEO, you conduct a limited comparative test for one week with three conditions:

(A) current checkout-procedure without any privacy related options

(B) "Do not sell" opt-out possibility as required by CCPA

(C) your 90 seconds video explanation with an opt-in at the end

You can download the three resulting Apache logfiles, and the inventory list from the 'Materials' section on the CMS. Your job is to find out if either of the two new options affect sales, i.e. the number of aborted checkout procedures, or the amount of money that shoppers spend on computer parts. You may solve this exercise either in R or Python. We expect your submission in a Jupyter Notebook that includes all necessary code and your written answers. You can download a template notebook from the 'Materials' section on CMS.

1. Parse and Prepare the log Files (3P)

Extract the following information from the three log files and the inventory list:

1. The amount of aborted checkout procedures
2. The amount of successful checkout procedures
3. The amount of money spent in each successful checkout procedure

Review the initial data and the resulting datapoints for factors that could impact the validity of your results. Exclude these datapoints from your dataset and provide an argument for their exclusion in your submission.

Possibly helpful information:

- Each line of the log file is structured as follows:
  `<Remote IP>␣<Remote User>␣<Remote logname>␣<Receive Time>␣<Request Line>␣<User Agend>␣<Status>␣<Size of Response>`
- Customers visit `/item/details` to view the details of a certain computer parts. Customers visit `/cart/add` or `/cart/remove` to add respectively remove computer parts from their shopping cart.
- Customers start the checkout procedure by visiting `/cart/checkout` and they are forwarded to `/thank_you_for_your_order` if it was successful.
- You may assume that each customer uses a unique IP address.
- If you choose to use Python for this exercise then we recommend using the apache-log-parser library (https://github.com/rory/apache-log-parser) which you can install using pip (`pip3 install apache-log-parser`).

2. Analyze the Extracted Information (22P)

You may use Python or R to solve the tasks in this exercise were applicable.

For each statistical test you run you should (a) explicitly state H_0, and (b) list all of the results (including the non-significant ones).

1. Effect on checkout completion (6P)
   (a) Assess the global effect of authentication methods on the checkout success/abortion using a non-parametric test.
   (b) In case of a global effect, run a post-hoc analysis to find out which condition shows an effect.

(c) Write a short explanation for the report, that describes why you choose these tests and what the results mean for the online shop.

2. Describe the distributions of money spent by the customers (6P)

   (a) Calculate the mean and the standard deviation of the money spent in successful checkout procedures for all three log files.

   (b) Check each of the three distributions for normality and homogeneity of variances.

   (c) Explain what each of your results means in simple terms for the report.

3. Effect on the amount of money spent in each successful checkout procedure (6P)

   (a) Use a parametric test to find out if there is a global effect on the money spent in each checkout procedure.

   (b) In case of a global effect, run a post-hoc analysis to find out which conditions show an effect.

   (c) Write a short explanation for the CEO that describes why you chose these tests and what the results mean for the future privacy options on the online shop.

4. Argue which kind of additional data and corresponding statistical tests could provide evidence that your preferred option (C - *video explanation with opt-in*) trumps the other two. (4P)

Possibly helpful information:

- You may use the python libraries numpy, scipy, pandas, and statsmodels for the required calculations in this exercise. Install them using pip (`pip3 install numpy scipy pandas statsmodels`).
- If you are unsure about which statistical test to use, read up on them in the lecture slides or Chapter 4 of Lazar et al.[1].

Resources

You can find the following resources in the 'Material' section on the CMS:

- Jupyter Notebook Template
- ComVille Prices in a `.csv` file
- `.log` files for conditions (A), (B) and (C) respectively
- Chapter 4 of Lazar et al.[1]

Office Hours

We offer office hours on the off days (each Tuesday and Thursday) for students to work on assignments and ask questions to tutors who are present.

Submission

Your submission should include the following:

1. all necessary code
2. your written answers

Submit the Jupyter notebook containing these parts via the CMS until **Wednesday, March 22nd, 2023, 09:59 CEST**.

References

[1] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research Methods in Human-Computer Interaction*. Wiley Publishing, 2010. ISBN: 9780128093436.