<Title of the Project>

# Introduction

In this paper and the follow documents presented along the way of this project, we will attempt to demonstrate if the online browsing history is correlated to final purchase.

In that case, we will be using four data files to learn from. The first file is composed with the list of 22.874 customers and some personal information that will be described in the dataset section. The second file to be used is a file that contains information related to users buying voucher. And the third file to be used contains the data related to users browsing in the site. The last file to be used is the coupon detail information that contains all the data specifically about the coupon.

# Literature Review

There are several researches and projects done related to online customer behavior and coupon usage.

• Blattberg et al. (1978) suggested that the coupon usage would be related to demographics characteristics where the consumers are assumed to minimize the sum of transaction costs, storage costs and the price of the item. He basically suggested that the upper income households, the more likely to redeem the coupon.

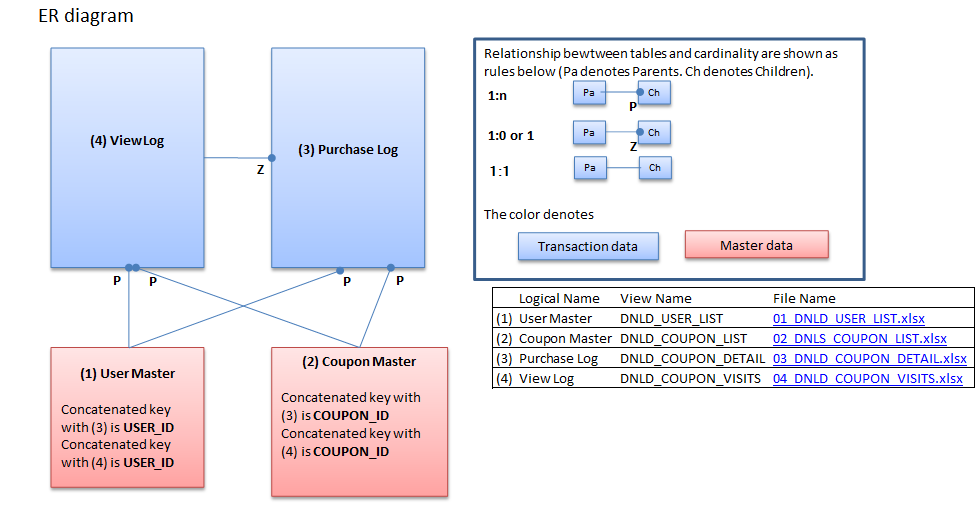
• Narasimhan (1984) proposed that intensity of coupon usage is related inversely to a household’s opportunity cost of time. Therefore it would be expected that in households that are more educated, have children under six and husband and wife are employed would have a lower prone to use coupons.

• Bawa and W. Shoemaker (1987) suggested that the intention of using the coupon (which in their project is called CPI - coupon proneness index) is a function of household characteristics and customer behavior.

-Kwon Jung (2010) suggested that the online usage of coupon is a funcion of the percentage of discount offered and demographics.

# Dataset

As mentioned before, the dataset used is composed by 4 files and they all have an entity in common.



Above you can see the ER diagram for the data.

• user\_list.csv: contains 6 features and 22,873 users. The features are related to (registered day, gender, age, date that unregistered, preferable name and user id).

* coupon\_detail.csv: contains 6 features and 168,997 entries. The columns consist in information about quantity bought, purchase date, geographic area that was bought, purchase identifier, user id and coupon id.
* coupon\_visits.csv: contains 8 features that refer manly to the browsing logs. The columns are purchase flag, purchase id in case it happened, log date, page serial, refer, coupon id, user id, session id.

coupon\_list.csv contains 24 features related to the coupon like category, expire date, what week days it’s available, discount value, and so on.

# Approach

In our process to model the data we will have 4 main steps.

• Data preparation: where will merge/join the tables creating one single table to be used. In addition we will check if there is any missing information or data that should be transformed.

* Descriptive Analysis: where we will calculate basic statistics to understand how the data is distributed.
* Modeling: where the methodology will be tested in order to predict the coupon

Once this is done, explain each of the steps in detail. What are you planning to do in each step or have already done. For example, in the above case you would create subheadings for each of the steps.

## Step 1: Data

The first step to be taken in this project is to transform the data into a flattened table with the features of the users and the coupons considering if the interaction was a purchase or not.

The code to transform the data can be found in this github [repository](https://github.com/dresenhista/StatisticsProjects/blob/master/Project/assignment%202/1-clean.R).

This final table has around 2 million of interactions, therefore we selected randomly around 8 thousand interactions in order to run the model.

With the 8 thousand interactions we divided 80 % of the data to train and the rest to test. The training data was then used to work with the machine learning model using cross validation for 10 folds.

After the model selection the plan was to test in the test partition.

## Step 2: Cross Validation with Logistic Regression

The cross validation with model selection consists in splitting the data randomly in 10 folds and then applying the logistic regression to each fold. After the model is adjusted, compare what the prediction would be to the rest of the training data, aka 9 folds left. You can find the code to create the data [here](https://github.com/dresenhista/StatisticsProjects/blob/master/Project/assignment%202/3-rcode_final.R)

The model in this first stage consists in having the follow function: purchase = function (features). And because majority of the times we have more people not buying then buying, if the model predicts that everyone don’t buy we have a smaller error than the opposite. Therefore any model is biased to predict a unsuccessful purchase. The problem is basically minimize the picture below:

[]insert here the teacher’s drawing]



Running the cross validation the model that minimizes the false positive and maximizes the true positive is the 5th interaction. See all the detailed data in this link

|  |
| --- |
|  |

## Step N: <Name of the step>

Write details of the step N. If there is any source code that you’d like to share then provide the link of the Github.

# Results

Explain your results here. Consider that you need to communicate your results to executives in an organization. For example:

1. Insert tables and/or charts showing the results
2. Write description of the tables and charts, such that they show the usefulness for an organization
3. Identify the evaluation measures, such as accuracy, precision, recall, etc.

# Conclusions

Give a short summary (one to two paragraphs) of your analysis and conclude the discussion by defining the usefulness of your analysis.