**Data Science**

**Problem 1: Predictive shopping list for Monoprix**

Monoprix is a french retailer whose activity is primarily focused on cities. In order to reward its loyal customers, Monoprix started to deploy new innovative services. The goal is to provide new ways for the customer to interact with Monoprix, and help him in his daily life.

In this situation has Monoprix decided to launch a new vocal based experience, using smart speakers like Google Home. The purpose of the service is to build shopping lists by simply talking to the smart speaker. For example, you can ask « Remind me to buy egg » and the speaker will add eggs to the shopping list.

Monoprix wants this service to be intelligent, and asked us to create an algorithm that could learn from the customers habits and suggest them the products they might have forgotten to add on the list.

Therefore, Monoprix provides you with 10 years of purchase history, for the customers of their loyalty program. Also, you have at your disposal the full product catalog.

1. **First step**: Recommendation engineQuestion: How can you use Monoprix's data to build the recommendation algorithm ?

In view of the problem of building the recommendation algorithm for Monoprix, I would follow the pipeline steps to develop the recommendation algorithm:

1. I would clean and process the data (cleaning sales history, product reviews, etc.).

2. To gain insights, I would conduct a thorough investigative analysis exploratory and explanatory of the information to discover underlying patterns and relationships.

3. The data would best be partitioned with 70% dedicated to model training, 20% for testing purposes, and the final ten percent allocated towards evaluation of results.

4. **I would use a hybrid recommendation model approach based on user-based collaborative filtering and content-based filtering, due this approach can better capture user behavior, resulting in more accurate and personalized recommendations for users and also can generalize better than only one type of model recommendation (model approach based on user-based collaborative filtering). This hybrid model use cosine similarity and k-nearest neighbors.**

5. Would train the recommendation hydro model

6. After rigorously evaluating the model's performance in training and testing, I would consider implementing it in production more broadly if it proves sufficiently effective and generalist on the evaluation set.

1. **Second step**: Natural LanguageQuestion: During the project, we realize that the product catalog is really dirty because products are wrongly named. Which solutions can you propound to correct products' names?

Given that the product catalog is "very dirty", a manual approach may not be viable, in this case I would follow a pipeline to clean the catalog data using NLP.

1. Text Preprocessing (Removing punctuations, special characters)

2. Tokenization (Division of product names into words)

3. Removal of Stopwords (Elimination of common words, "is", "the", "of", etc.)

4. Lemmatization and Stemming (Reduction of inflected or derived words to their base form)

5.

6. Correction of Specific Product Names in the catalog (For example: "Playstation XYZ Pro" which was actually "Playstation XYZ")

**Problem 2: Measuring the effect of a marketing campaign**

A Pharmaceutical company is trying to measure how much their last marketing campaign on a specific product has helped increase its sales and asks for your advice on how to do it. Previewing this, the Brand/Product manager already built a marketing campaign set up in which he separated one control region ( specific region where he didn't roll out the campaign ) from the others.

1. Assume the product has been on market in the last 2 years with a stable demand**. Explain a model you would advise the company to use and its main assumptions.**
2. Assume now that the product is new, so that the campaign was a launching one.

**In this scenario is it possible to measure the effect of the campaign on sales? If yes, what model would you suggest and why?**

**Problem 3: Regression Analysis**

A supermarket company has a new internal policy to not discriminate **significantly** salary according to the location of their employees. They gathered the data from all of their employees and want you to verify if they are already following the new policy.

Before answering the questions below take a look at the annexed dataset: (**1stPhase-SelectiveProcess-Data Science-Data Base.csv**)

1. Question: Describe how can you use the supermarket data to verify if employees from different locations have significantly different salaries ? (Include here how you are going to treat the variables before feeding into the model)

Para este problema, se somente queremos saber se a variavel “locations” existe uma diferença estatisticamente significativa nos salários entre os grupos e a empresa segue já a nova politica, podemos ter uma abordagem mais simples, sem precisar criar um modelo de regressão, podemos usar o somente o teste t de Student para verificar se a variável “locations” tem um diferença estatisca nos salários dos funcionários (este método compara a distribuição entre a média valores nos grupos). Caso exista uma diferença estatistica significante, podemos criar um modelo de regressão para entender melhor a contribuição da variavel “locations” para ps sálarios dos funcionarios.

1) Primeiramente devemos saber o tipo de “localizações” da variável, usarei métodos para inferir o tipo de variável.

Podemos ver que a variável “locations” é uma variável categórica nominal binária. Agora que sabemos o tipo de variável e o alvo é uma variável numérica contínua.

1. Question: Implement the approach you described in python or r