Rapport de projet Web Scraping

Roland DENIZOT, Zachary CHENOT- DIA2

Voici le lien vers le GitHub

1 Sujet, Contexte et Hypothèses

1.1 Contexte, idée et problématique

Dans le cadre su projet final de la matière Web Scraping en dernière année à l'ESILV, nous nous plaçons en tant qu'une entreprise proposant du matériel informatique. Nous simulons que nous voulons renouveler et mettre à jour les cartes graphiques que nous proposons sur notre site ainsi qu'établir des prix cohérents par rapport au marché. Pour cela nous allons étudier les produits présents sur plusieurs sites internet et en déduire, par des visualisations et tableaux, les types de cartes graphiques qu'il serait interessant de mettre en vente sur notre site pour compléter ce que la concurrence propose en faible quantité.

En plus de cela, une application cloud va être mise en place afin de pouvoir afficher dynamiquement toutes les visualisations pour n'importe quel dataset concernant notre sujet, tant que celui-ci est formaté comme notre dataset initial.

La problématique pourrait donc se formuler comme suit: Comment pouvons nous placer une offre de cartes graphiques par rapport au marché de la concurrence ?

1.2 Sites à scraper

Nous avons scrapé les deux leaders de la vente de produits informatiques en ligne en France: TopAchat et LDLC. Pour ce faire, une étude approfondie du code source de ces deux pages a été réalisée puis un scraping en python a été fait en fonction. Nous avons pu constater que la manière de concevoir les sites internet pouvait être très différente, allant d'une façon très organisée jusqu'aux manières chaotiques. Les langages de programmation de sites pouvaient aussi être différents et bloquer par exemple le scraping si nous ne simulions pas que nous sommes un vrai utilisateur et donc "faire croire" au site que nous ne sommes pas un programme.

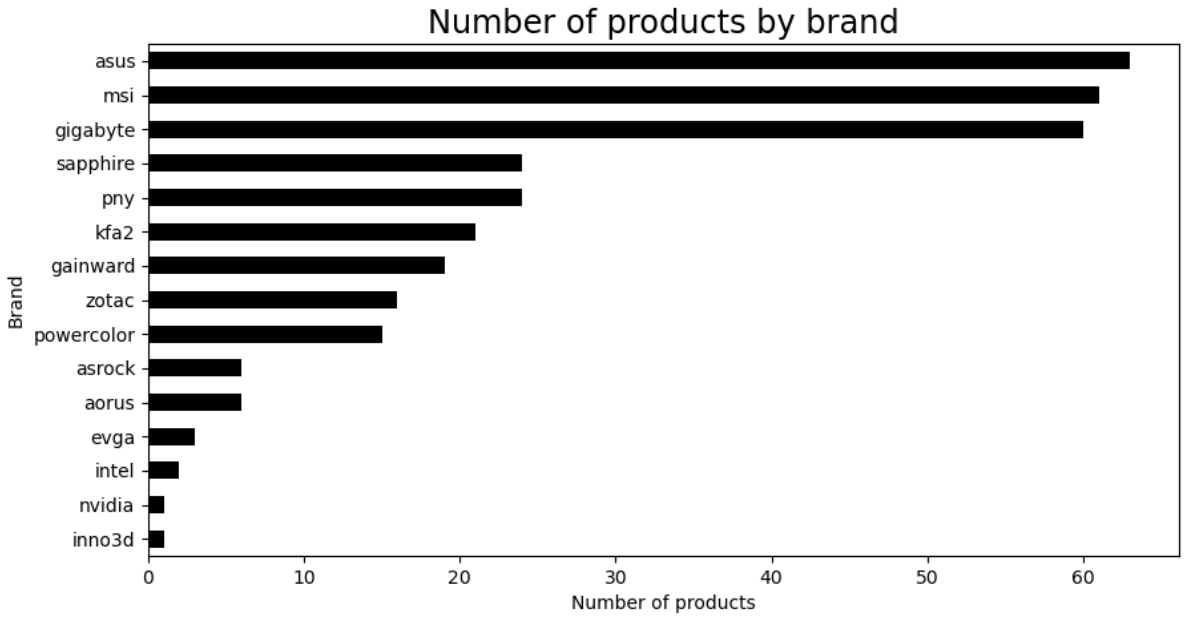
Nous souhaitons ajouter qu'avec le temps, les articles proposés sur les sites ou la manière de mettre en page les sites a pu changer, c'est pourquoi nous avons scrapé puis sauvegardé en csv les tableaux des deux sites, afin de travailler sur des données fixes et stables.

1.3 Hypothèses

Avant la mise en oeuvre de ce projet, nous avons fait l'hypothèse que les sites à scraper forment une bonne représentation de la vente des cartes graphiques en France en raison du fait que ce sont les deux leaders de ce type de commerce dans notre pays. De plus, nous pouvions prévoir le fait que ces sites, visant majoritairement les joueuers de jeux-vidéo et les professionnels, proposent des cartes majoritairement haut de gamme.

2 Résultats

2.1 Nombre de produits par marque

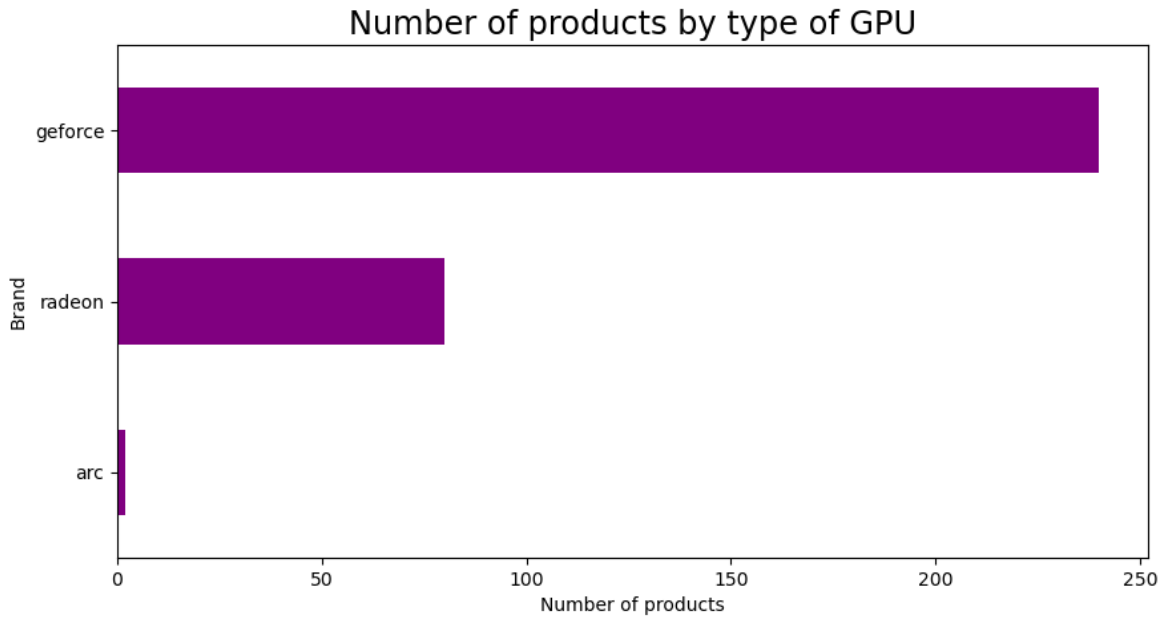


A l'obeservation de ce graphique, il faudrait limiter la quantité de produits de la marque Asus, MSI et Gigabyte car ce sont les plus représentés chez les concurrents. Il faurdrait plutot favoriser les marques réputées comme étant de qualité mais moins présentes sur les autres sites tels que Sapphire, KFA2, Gainward, Zotac, Aorus.

A l'inverse, certaines marques telles que Intel ou Inno3D ne sont pas vraimpent des manufactureurs de cartes graphiques, c'est pourquoi il n'y a pas forcément nécessité de proposer une grande quantité de ces produits.

Enfin, Nvidia correspond à l'entrée de gamme, les cartes sont basiques sans modifications hardware particulière, cela représente plutôt la base de ce qui est possible pour chaque chipset graphique qu'elle produit. Nous pourrions peut-être augmenter la quantité de produits de ce type sur notre site dans le but d'atteindre une clientèle qui ne joue pas fréquemment aux jeux videos ou n'a tout simplement pas besoin de grandes performances graphiques.

2.2 Nombre de produits par GPU

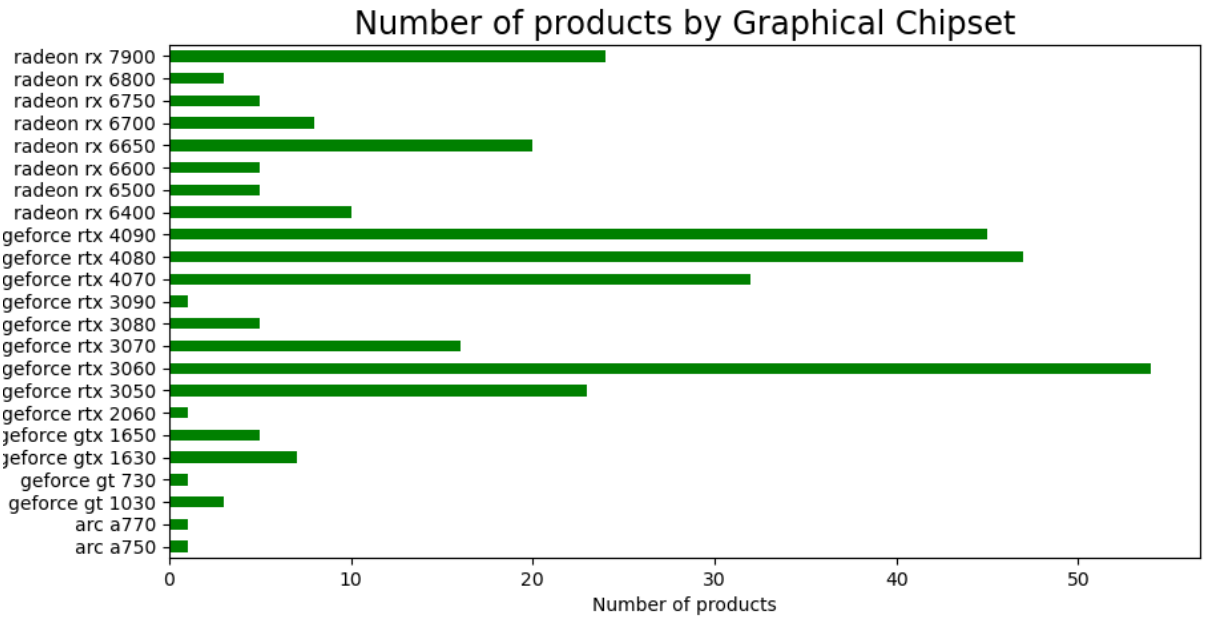


En regardant le graphique précédent, nous constatons très distinctement une large différence dans le nombre de produits en fonction de leur type de GPU. Il serait pertinent de prioriser les cartes de type Radeon et Arc (présentes uniquement sur TopAchat et non LDLC).

Il faudrait tout de même garder un certain nombre de cartes graphiques de type GeForce car ces dernières sont très réputées auprès de la communauté. Radeon possède une qualité comparable aux GeForce mais sont moins connues, il serait intéressant de proposer un nombre significatif de ces cartes pour récupérer cette communauté de connaisseurs.

Enfin, les cartes de type Arc sont considérées comme de l'entrée de gamme, il est important de les mettre en vente sur notre site mais peut-être dans des quantités plus appropriées.

2.3 Nombre de produits par chipset graphique

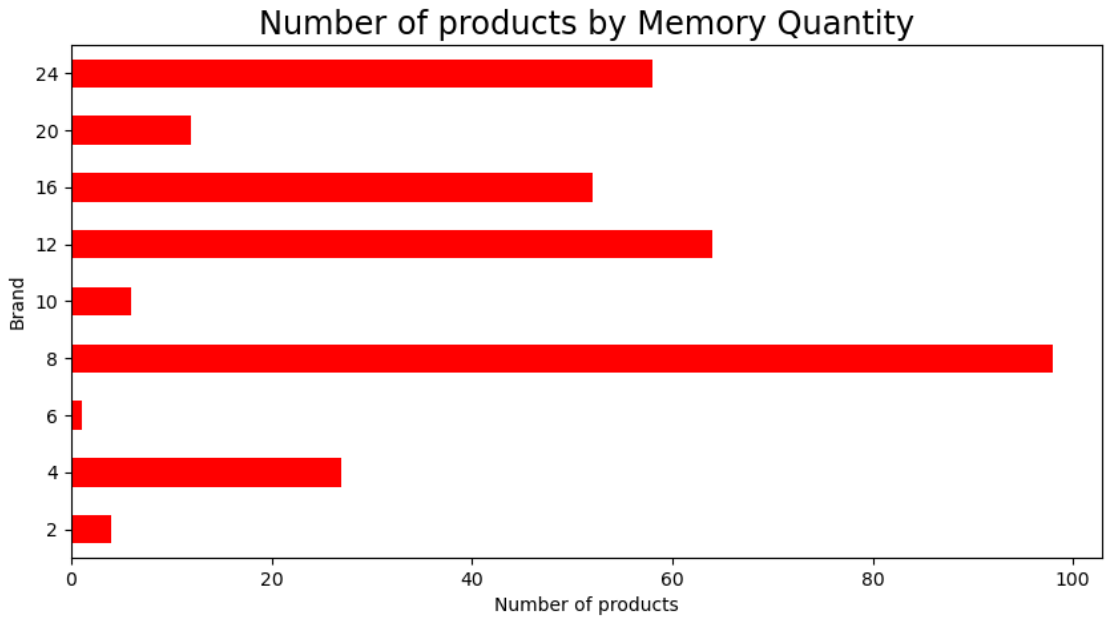


Ce graphique entre un peu plus dans le détail par rapport au précédent. Nous pouvons constater qu'une opportunité s'offre à nous pour les cartes de chipset graphique geForce RTX 3080, 3090 et 3070. Ces produits ne sont pas de la dernière technologie en date mais proposent toujours des performances très bonnes, qui pourraient être tout à fait suffisante pour la majorité des clients (cf. prix). Il en est de même pour les Radeon 6800, 6750 et 6700 qui sont des produits moins connus par le public mais sont tout aussi puissants.

Il serait de plus pertinent de proposer plus de cartes d'entrée de gamme type Arc A770, A750 ou GeForce 2060, pour un public moins informé ou nécessitant moins de performances.

Enfin, il semble important de limiter la quantité de GeForce RTX 3060, 4080 et 4090.

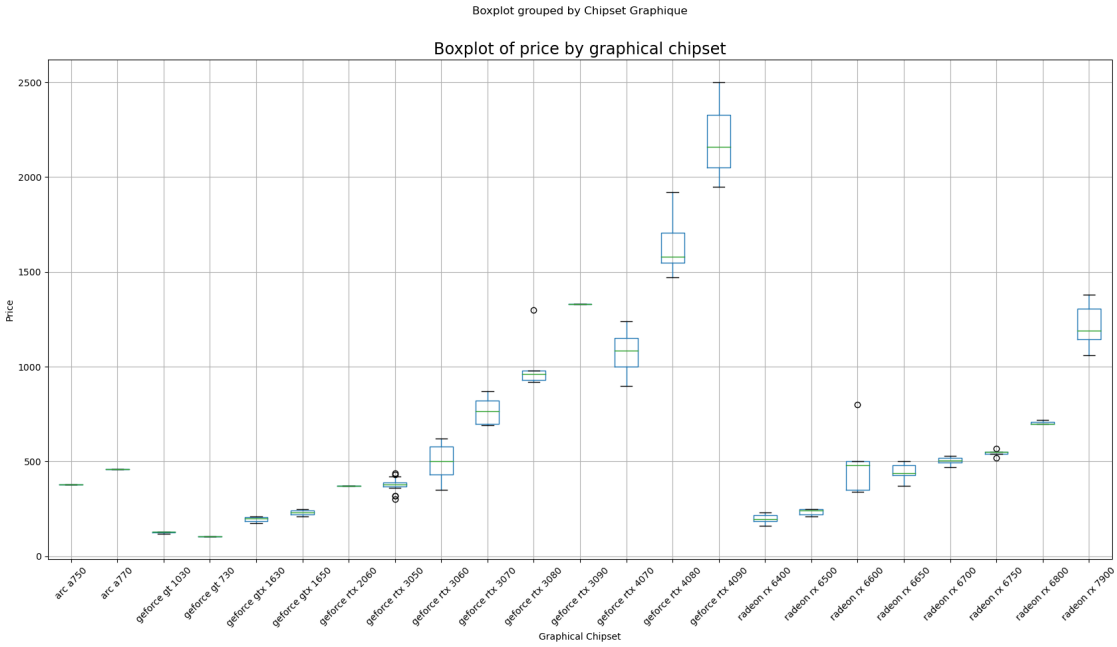
2.4 Nombre de produits par quantité de mémoire



A la vue du graphique précédent, il serait intéressant pour notre entreprise de compléter l'offre sur les quantités de mémoire 2, 6, 10 et 20GB. la quantité de cartes possédant ces tailles de mémoire sont très limitées sur les sites de la concurrence. Nous constatons de plus qu'aucune carte avec 14GB de mémoire n'est proposée.

Nous pouvons toujours continuer d'offir à nos clients la possibilité d'acheter des cartes graphiques possédant 12, 16, 24GB de mémoire mais en adaptant en quantité égales pour chacunes d'elles.

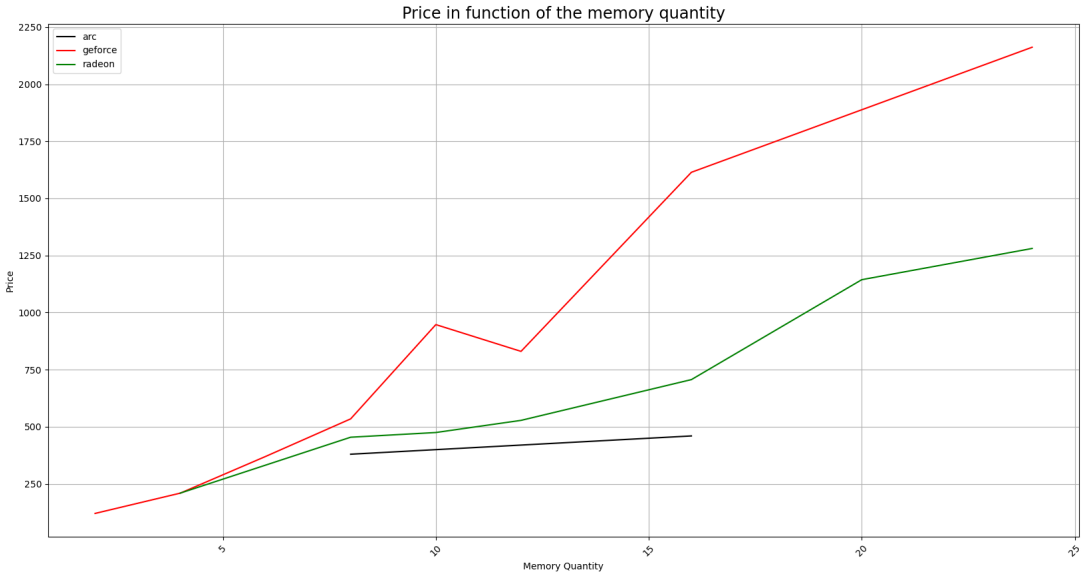
2.5 Boxplot du prix par chipset graphique



Sur cette figure, il est clair que certains chipsets ont un très grand écart de prix d'une carte à l'autre, nous nous devons donc d'être très attentifs lors du choix du prix à assigner à chacune d'elles dans le but d'être attractifs par rapport à la concurrence, particulièrement pour les plus chères d'entre elles.

Certains autres produits n'ont pas un grand spectre de prix en raison du fait que leur variété n'est pas grande. Il serait pertinent de proposer sur notre site plus de cartes de ce type pour élargir les possibilités.

2.6 Price en fonction de la quantité de mémoire



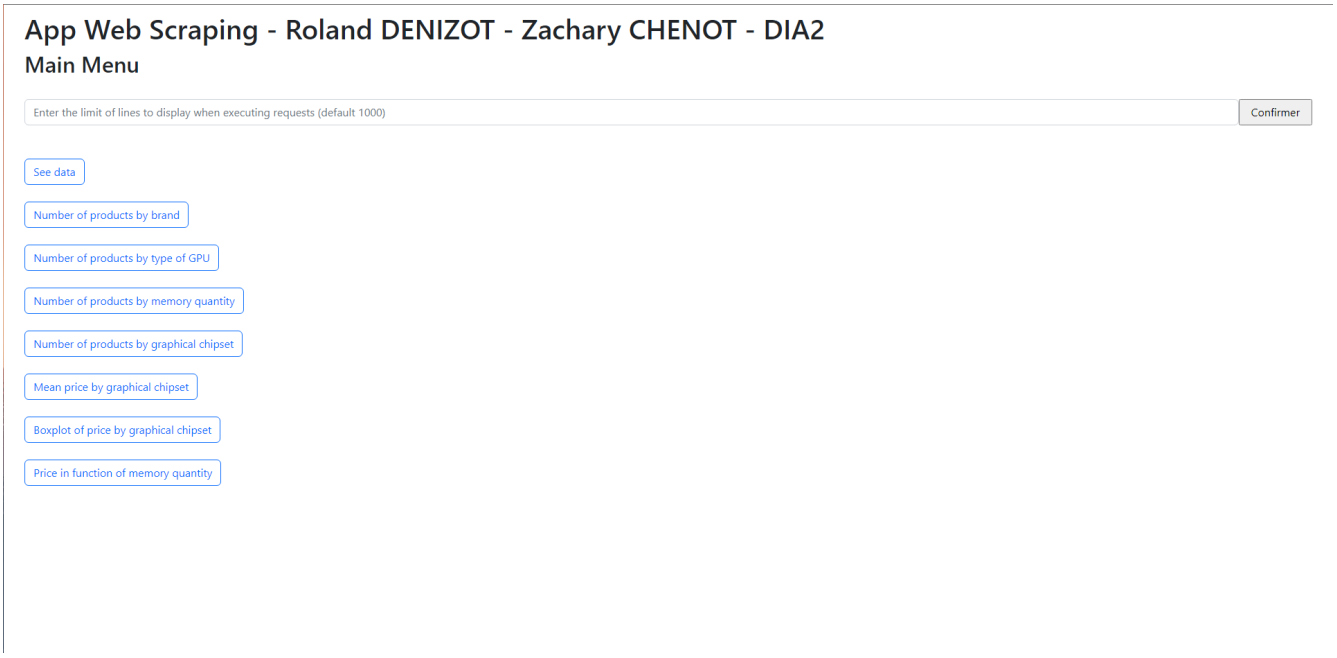
Sur ce dernier graphique, nous pouvons noter que les cartes de type GeForce possèdent des prix bien plus élevé que les autres à quantité de mémoire égale. Une nouvelle fois, proposer plus de Radeon et de Arc semble une bonne idée car les prix sont plus attractifs à performances égales.

Nous pouvons constater une légère baisse du prix au passage de 10 à 12GB de mémoire, cela est très certainement dû au passage à une nouvelle génération qui prend comme entrée de gamme, et donc comme prix de référence, les cartes avec 12GB de mémoire.

Enfin, pour les meilleures cartes (celles avec le plus de mémoire), l'écart de prix est immense entre Radeon et GeForce, nous pouvons donc essayer de baisser les prix des GeForce car il semblerait que la marge des concurrents soit assez forte sur ces produits.

2.7 Application Flask

L'application flask possède 3 types de pages: le menu principal, l'affichage de plots et l'affichage des données sous forme de tableau. Elle peut générer les graphiques vus précédemment pour nimporte quel dataset, tant que celui-ci est formaté comme le dataset initial.



App Web Scraping - Roland DENIZOT - Zachary CHENOT - DIA2

See data

Main Menu

Nom_produit	Chipset Graphique	Marque	Quantité Mémoire (Go)	Type Mémoire	Bus	Prix (€)	Site Internet	Longueur	Largeur	Epaisseur	Type du GPU
kfa2 geforce rtx 3060 ti (1-click oc) (lhr)	geforce rtx 3060	kfa2	8	gddr6	pci express 4.0	449.99	topachat.com	256.50	131.50	41.50	geforce
sapphire radeon rx 6700 pulse	radeon rx 6700	sapphire	10	gddr6	pci express 4.0	469.99	topachat.com	260.00	119.85	49.00	radeon
msi radeon rx 6750 xt mech 2x oc	radeon rx 6750	msi	12	gddr6	pci express 4.0	519.99	topachat.com	249.00	132.00	52.00	radeon
sapphire radeon rx 6600 pulse	radeon rx 6600	sapphire	8	gddr6	pci express 4.0	339.99	topachat.com	193.00	120.05	40.05	radeon
gigabyte radeon rx 6650 xt gaming oc	radeon rx 6650	gigabyte	8	gddr6	pci express 4.0	369.99	topachat.com	282.00	115.00	50.00	radeon
kfa2 geforce rtx 3050 ex (lhr)	geforce rtx 3050	kfa2	8	gddr6	pci express 4.0	299.99	topachat.com	224.00	133.00	44.00	geforce
kfa2 geforce gtx 1630 ex (1-click oc)	geforce gtx 1630	kfa2	4	gddr6	pci express 3.0	174.99	topachat.com	196.00	126.00	39.00	geforce
kfa2 geforce rtx 3060 (12 go) (1-click oc) (lhr)	geforce rtx 3060	kfa2	12	gddr6	pci express 4.0	399.99	topachat.com	258.00	126.00	41.50	geforce
kfa2 geforce rtx 3060 (8 go) (1-click oc) (lhr)	geforce rtx 3060	kfa2	8	gddr6	pci express 4.0	349.99	topachat.com	258.00	126.00	41.50	geforce
gainward geforce rtx 4070 ti phantom reunion (12 go)	geforce rtx 4070	gainward	12	gddr6x	pci express 4.0	999.99	topachat.com	313.90	137.00	59.70	geforce
gigabyte geforce rtx 3060 eagle oc rev 2.0 (12 go) (lhr)	geforce rtx 3060	gigabyte	12	gddr6	pci express 4.0	439.99	topachat.com	242.00	124.00	41.00	geforce
gigabyte geforce rtx 4080 gaming oc (16 go)	geforce rtx 4080	gigabyte	16	gddr6x	pci express 4.0	1499.99	topachat.com	342.00	150.00	75.00	geforce
kfa2 geforce rtx 3060 ti plus v2 gddr6x (1-click oc) (lhr)	geforce rtx 3060	kfa2	8	gddr6x	pci express 4.0	499.99	topachat.com	254.00	142.00	54.00	geforce
sapphire radeon rx 7900 xt	radeon rx 7900	sapphire	20	gddr6	pci express 4.0	1099.99	topachat.com	276.40	112.70	51.25	radeon
asus geforce rtx 3050 dual o8g (lhr)	geforce rtx 3050	asus	8	gddr6	pci express 4.0	359.99	topachat.com	200.00	123.00	38.00	geforce
msi geforce rtx 3060 ventus 2x oc (12 go) (lhr)	geforce rtx 3060	msi	12	gddr6	pci express 4.0	459.99	topachat.com	235.00	124.00	42.00	geforce

App Web Scraping - Roland DENIZOT - Zachary CHENOT - DIA2

Display of Price in function of the memory quantity

Main Menu

2.8 Produits communs

Afin de pouvoir identifier la stratégie de pricing de chacun des deux sites et se placer par rapport à eux, nous avons dû dans un premier temps trouver les produits commun aux deux sites. Pour ce faire nous avons utiliser deux filtre. Le premier, consistait à réunir les produits de ldlc et de topachat dans un même dataframe. Ensuite, nous avons demandé de trouver quel produit avait au moins un duplicat dans ce dataframe. Or ce duplicat doit être trouvé en fonction de certaines caractéristiques, c'est-à-dire que nous avons indiqué que si certains colonnes était exactement pareil qu'un certains produit, alors il s'agit duplicat. Concrètement, pour un produit donné, on regarde son type de carte graphique, le quantité de mémoire (en Go), et ses dimensions par exemple, et s'il existe au moins un autre produit qui a exactement les mêmes valeurs pour ces colonnes alors on les garde tous.

Ensuite un deuxième filtre a été appliqué sur les noms des produits. Nous avons de nouveau séparé le dataframe en deux : un pour ldlc, et un pour topachat. Puis pour chaque produit dans le dataframe avec le moins de produits (ici topachat) nous calculons le score de correspondance entre les string du nom de produit topachat et ldlc. On garde alors celui avec le plus haut score. Néanmoins malgré ces deux filtres nous avons obtenu environ 120 produits similaire dont une quinzaine avec des anomalies.

3 Conclusion

En conclusion de l'étude menée, notre entreprise fictive pourrait donc s'orienter vers une gamme de produits entrée et milieu de gamme pour satisfaire les clients qui ont tout autre utilisation des cartes graphiques que pour les jeux vidéos: par exemple les mineurs de crypto-monnaies qui préfèrent multiplier les cartes milieu de gamme moins chères et plus optimisées pour leur activité ou les personnes ayant uniquement besoin d'un affichage multiple nécessitant une carte graphique. Afin de pouvoir tout de même atteindre les clients nécessitant de grandes performances graphiques, nous continuerons de proposer des cartes haut de gamme.

Tout cela nous emmène vers la conclusion qu'une meilleure répartition des gammes serait appropriée pour notre objectif de ventes, c'est-à-dire proposer en nombre similaire des cartes de tout type, en baissant le nombre de cartes largement représentées chez la concurrence et en augmentant le nombre de cartes peu présentes.

Concernant les prix, nous serions enclin à adopter la stratégie de Topachat par rapport à ldlc, à savoir qu'ils vendent systématiquement les mêmes produits 10€ moins cher(voire 20€-30€). Nous vendrions donc moins cher les produits présents en grande quantité chez nos concurrent (Asus, Gigabyte...) mais potentiellement plus cher les produits dont eux manque (Nvidia par exemple).

Une potentielle amélioration serait d'essayer d'obtenir un nombre de vente de chaque produits sur les sites ldlc et topachat qui rendrait notre business plan plus précis et solide.