

Rapport

Projet Teisseire

Groupe GREnadine



Groupe GREnadine :

Chef de projet : Morgane MAULET

Équipe : Claire LOUBOUTIN, Lou MARSAIS, Amandine PASCAL, Hélène TE,
Abdessalam ESSABBEUR

Référente technique dans l'entreprise : Madame Sophie GUISNEL

Enseignants référents : Messieurs Pierre KARPMAN et Régis PERRIER

REMERCIEMENTS

Avant de commencer, nous souhaitons remercier l'ensemble des personnes ayant contribué à la réussite de ce projet, en particulier Madame GUISNEL et Messieurs GRIMAUD, PERRIER et KARPMAN.

Nous sommes reconnaissants pour le temps qu'ils nous ont consacré, les suggestions dont ils nous ont fait part lorsque nous avons rencontré des difficultés et pour le temps qu'ils ont pris pour répondre à toutes nos questions.

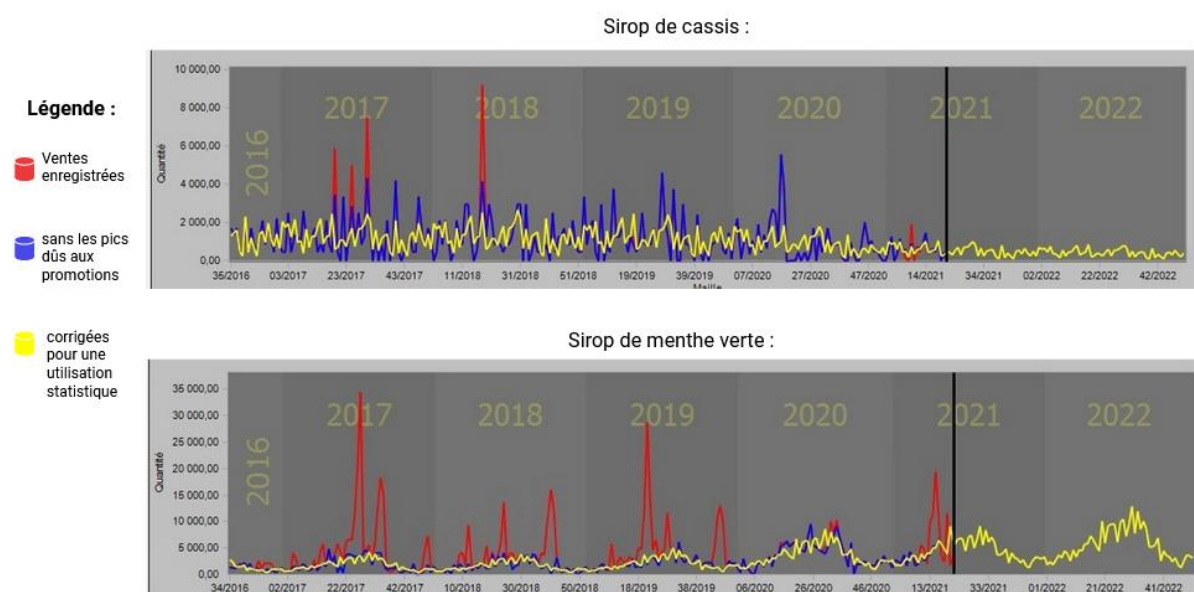
TABLE DES MATIÈRES

INTRODUCTION.....	4
DECOUPAGE DES TACHES ET STRUCTURATION DE L'EQUIPE	6
2.1 DIAGRAMME DE GANTT PREVISIONNEL.....	7
2.2 DIAGRAMME DE GANTT OBSERVE.....	8
ÉLEMENT FOURNI PAR LE CLIENT.....	9
3.1 TRI DE LA BASE DE DONNEES.....	9
3.2 SEPARATION DE LA NOUVELLE BASE DE DONNEES POUR LES STATISTIQUES.....	12
WEB SCRAPING	14
ANALYSE STATISTIQUE	17
5.1 CALCUL DE LA CORRELATION ENTRE LES VENTES DE SIROPS ET LA TEMPERATURE.....	17
5.2 ESTIMATION DE L'IMPACT D'UN ETE CHAUD SUR LES VENTES DE SIROPS	21
CONCLUSION.....	23
TABLE DES FIGURES	24
GLOSSAIRE	25
WEBOGRAPHIE.....	27

INTRODUCTION

Au cours d'une semaine en décembre puis de trois en avril, au sein du campus, nous avons travaillé sur un projet qui nous a été proposé par le groupe Britvic. Britvic est un groupe international spécialisé dans les boissons et rafraîchissements sans alcool, basé en Angleterre. En 2010, Britvic a racheté l'entreprise Teisseire qui elle-même était propriétaire de plusieurs autres marques telles que : Pressade, Moulin de Valdonne et Fruit Shoot. L'usine avec laquelle nous avons travaillé, située à Crolles, produit uniquement des sirops. Tout au long du projet, nous avons été amenés à communiquer avec Madame Sophie GUISNEL, notre contact client chez Britvic. Elle y est la prévisionniste des ventes et pour ce projet, elle était chargée de représenter l'entreprise.

Les ventes de sirops sont impactées par de nombreux facteurs et notamment par les variations climatiques : tous les produits ne réagissent pas de la même façon. Les ventes de certains parfums, comme le cassis, ne sont pas vraiment impactées par la chaleur. D'autres, en revanche, comme celles de la menthe ou encore du citron, sont beaucoup plus "météo-sensibles", c'est-à-dire que leur consommation sera d'autant plus importante en été ou en cas de pic de chaleur. Ce phénomène est identifiable par les courbes jaunes des graphiques suivants, fournis par l'entreprise.



1. Impact de la chaleur sur les ventes de sirops

Notre premier objectif a donc été de mesurer la corrélation* entre les ventes des différents parfums de sirops et la météo afin de remplir le second objectif, à savoir l'estimation de l'impact d'un été chaud sur celles-ci.

Découpage des tâches et structuration de l'équipe

Nous avons d'abord séparé le projet en trois parties distinctes, que nous avons ensuite réparties entre nous en fonction de nos compétences et de nos préférences. Un premier groupe, composé de Claire LOUBOUTIN, Hélène TE et Abdessalam ESSABEUR, s'est chargé de récupérer des données météorologiques à l'aide du web scraping* afin de créer une base de données* exploitable des relevés de températures. Un deuxième groupe, composé de Lou MARSAIS et Morgane MAULET, a travaillé sur l'étude de la corrélation entre les ventes de sirops et les températures observées puis sur une prévision pour l'été prochain. Enfin, Amandine PASCAL, rejointe ensuite par Abdessalam, a travaillé sur une automatisation de l'adaptation de la base de données fournie par notre contact client pour son exploitation.

Pour faciliter l'organisation et le partage des ressources et des informations, nous avons créé un dossier partagé sur Google Drive, un espace de travail sur Notion ainsi qu'un serveur Discord. Nous les avons organisés en fonction des tâches sur lesquelles nous travaillions. Nous tenions également des comptes rendus journaliers de notre avancée.

Le diagramme de Gantt que nous avons composé lors de la première semaine, en décembre, n'a pas été complètement respecté, bien qu'il nous ait été très utile pour commencer et nous organiser. En effet, quelques tâches ont nécessité moins de temps que prévu, des problèmes ont été rencontrés et des missions à compléter ont été ajoutées ou fusionnées. Tout ceci nous a contraint de changer les effectifs et les affectations aux tâches, ou d'accorder plus de temps à certaines qu'à d'autres. Cette adaptabilité du planning nous a permis une plus grande efficacité et une perte de temps bien moindre que si nous l'avions suivi à la lettre.

2.1 Diagramme de Gantt prévisionnel

	Morgane
	Claire
	Hélène
	Amandine
	Lou
	Abdessalam

2. Légende du diagramme de Gantt

Tâches	S1					S2					S3					S4				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Pre conception																				
Rédiger le cahier des charges																				
Réaliser le diaporama de présentation																				
Réaliser le diagramme de Gantt																				
Rechercher un site regroupant les données météorologiques de France																				
Se documenter sur le web scraping																				
Se documenter sur les méthodes utilisées dans les statistiques																				

3. Diagramme de Gantt de la préconception

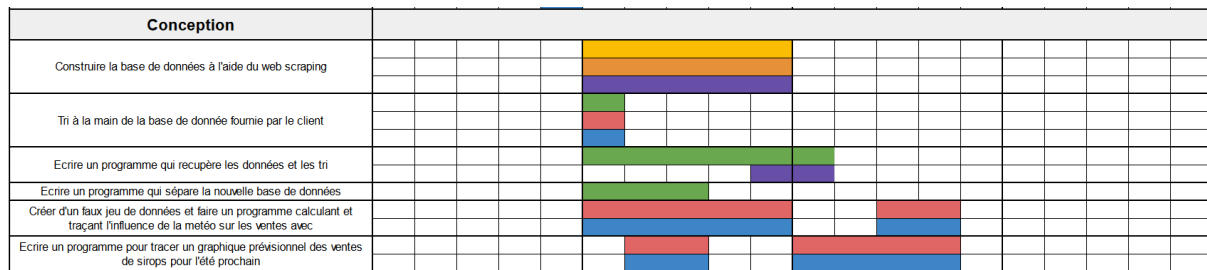
Conception																				
Construire la base de données à l'aide du web scraping																				
Ecrire un programme qui récupère les données pour mieux les traiter																				
Faire un programme d'analyse statistique avec un jeu de données aléatoires																				
Calculer et analyser l'influence de la météo sur les ventes																				
Ecrire un programme permettant de tracer le graphique de l'influence de la météo sur les ventes																				
Faire un bilan prévisionnel des ventes de sirops pour l'été prochain																				
Ecrire un programme pour tracer un graphique prévisionnel																				

4. Diagramme de Gantt de la conception

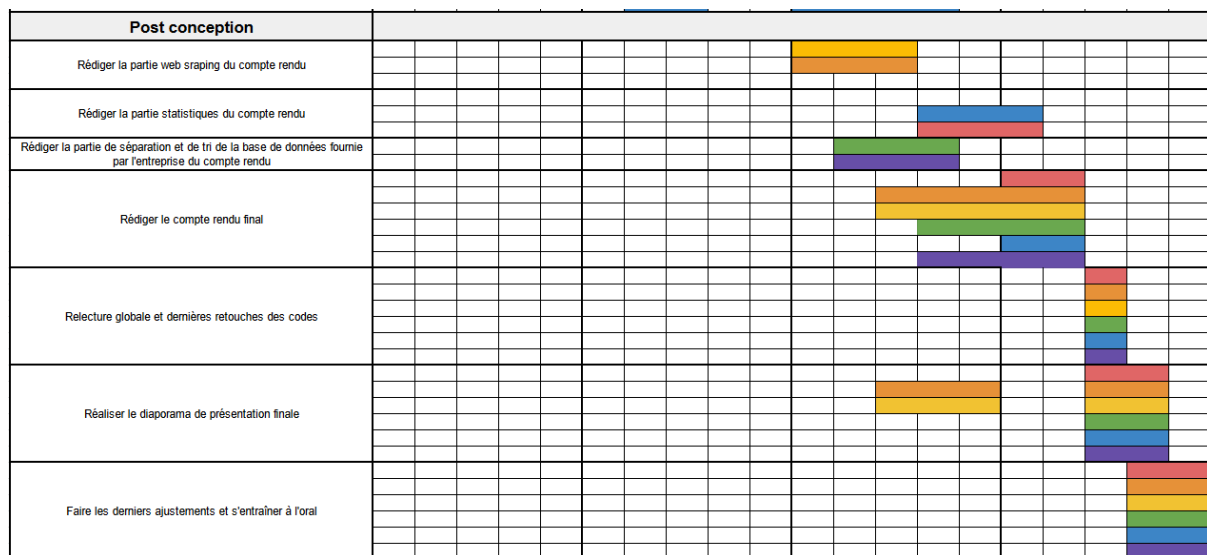
Post conception																				
Rédiger le compte rendu du web scraping																				
Rédiger le compte rendu des statistiques																				
Rédiger le compte rendu final																				
Réaliser le diaporama de présentation finale																				
Faire les derniers ajustements et s'entraîner à l'oral																				

5. Diagramme de Gantt de la post conception

2.2 Diagramme de Gantt observé



6. Diagramme de Gantt observé de la conception



7. Diagramme de Gantt observé de la post conception

Élément fourni par le client

3.1 Tri de la base de données

Nous avons en premier lieu récupéré la base de données fournie par l'entreprise, qui contient les données de ventes par mois des sirops commercialisés de janvier 2018 à décembre 2021. Cette base de données se présente sous la forme d'un fichier Excel*. Il est composé de trois feuilles de calcul* : la première contenant les données de ventes de T, la deuxième celles de MDV et la troisième celles des MDD.

Nous avons ensuite installé la bibliothèque* Pandas qui permet la manipulation de données. Elle nous a permis de rassembler les trois feuilles de calcul de la base de données en une seule afin de pouvoir effectuer un unique tri. Les trois feuilles de calcul possédant le même header, c'est-à-dire le nom des colonnes, composé des marques ou enseignes, des gammes, des parfums, des ventes en litres et des mois et années la concaténation aurait dû être assez simple.

Nous avons cependant rencontré un problème lors de la récupération des dates. En effet, celles-ci sont dans un format *jour/mois/année heures:minutes:secondes*. Après récupération et remise en place dans la nouvelle base de données, l'affichage des dates n'était pas correct puisque nous avions des dièses à la place de celles-ci. Cela vient du fait que les dates étaient trop longues. Nous les avons donc réduites en ne gardant que les mois et années, qui ne sont au final que les parties qui nous intéressent.

D	E	F	G	H	I
Parfum	Ventes en litres	#####	#####	#####	#####
CITRON	Fond de rayon	0	0	0	0
GRENADINE	Fond de rayon	0	0	0	0
MENTHE VERTE	Fond de rayon	0	0	0	0
FRAISE	Fond de rayon	0	0	0	0
PECHE	Fond de rayon	0	0	0	0
MOJITO	Fond de rayon	33250	33336	42433	55638
PINA COLADA	Fond de rayon	5947	5814	6570	9713
THE PÊCHE	Fond de rayon	29142	32785	36236	39254
TROPICAL	Fond de rayon	12359	9410	11761	11074
GRENADINE	Fond de rayon	0	0	0	0

8. Mauvaise mise en forme des dates

Nous avons aussi traité comme un cas particulier la première case de la première colonne. Comme nous avons une entête intitulée ‘Marque’ sur les deux premières feuilles et une autre intitulée ‘enseigne’ sur la dernière, nous n’avions rien d’affiché lors de la concaténation automatique.

Une autre complication est survenue lors de la compilation des feuilles de calcul. Nous avons pu voir dans le dataframe* qu’il y avait des colonnes en plus, qui cependant n’apparaissaient pas lors de la conversion du dataframe en un fichier Excel. Cela nous empêchait de rajouter par la suite les entêtes des colonnes dans le dataframe. Il a donc fallu supprimer ces colonnes en trop.

La dernière difficulté rencontrée lors de la jointure des feuilles de calcul, fut que la colonne des ‘enseigne’ de la troisième feuille de calcul n’était pas recopiée et affichait des valeurs “NaN” dans le dataframe. Cette erreur vient probablement du fait que la colonne récupérée se nomme ‘Marque’ puis ‘enseigne’. Pour résoudre cela, nous avons récupéré les colonnes des marques et enseignes à part dans une liste puis nous l’avons concaténé avec le dataframe contenant le reste des données.

Élément fourni par le client

106	35	MDV	MDV tradi	POMME V.VAI	fond de rayon	7060	5313
107	36	MDV	MDV tradi	RECETTE PRO	fond de rayon	19144	16657
108	37	MDV	MDV tradi	PECHE ABRICO	fond de rayon	44797	37212
109	0		BIO (btl 50	THE PÊCHE	fond de rayon	0	0
110	1		BIO (btl 50	GRENADINE	fond de rayon	0	0
111	2		BIO (btl 50	MENTHE VER	fond de rayon	0	0
112	3		BIO (BID P	CASSIS	fond de rayon	0	0
113	4		BIO (BID P	CITRON	fond de rayon	0	0
114	5		BIO (BID P	FRAISE	fond de rayon	0	0
115	6		BIO (BID P	GRENADINE	fond de rayon	0	0
116	7		BIO (BID P	MENTHE VER	fond de rayon	0	0
117	8		BIO (BID P	PECHE	fond de rayon	0	0
118	9		BIO BIDON	CASSIS	fond de rayon	0	0
119	10		BIO BIDON	CITRON	fond de rayon	0	0
120	11		BIO BIDON	FRAISE	fond de rayon	0	0
121	12		BIO BIDON	GRENADINE	fond de rayon	0	0
122	13		BIO BIDON	MENTHE VER	fond de rayon	0	0
123	14		BIO BIDON	PECHE	fond de rayon	0	0
124	15		BIO	CASSIS	fond de rayon	5070	3510

9. Mauvaise récupération de la colonne des enseignes

Une fois la base correctement récupérée, nous l'avons triée pour ne garder que les valeurs significatives. Nous avons tout d'abord simplifié certains noms de gammes. En effet, des produits tels que le 'T Classique 60cl Grenadine' et le 'T Classique 75cl Grenadine' par exemple, sont devenus le même produit 'T Classique Grenadine' car il ne s'agit en fait que d'un changement de volume de la part de la marque. Cela nous a permis par la suite de regrouper les lignes en sommant leurs valeurs.

Enfin, nous n'avons gardé que les produits qui ont été vendus pour au moins 15 mois, pas forcément consécutifs. Pour cela, nous avons importé la bibliothèque Numpy, dont nous nous sommes servis dans la fonction *replace()* pour transformer tous les 0 en "NaN" (valeur nulle).

	A	B	C	D	E	F	G	H	I	J
1	Marque_Enseigne	Gamme	Parfum	Ventes en litres	1 / 2018	2 / 2018	3 / 2018	4 / 2018	5 / 2018	6 / 2018
2	Auchan	0% 60CL	AGRUMES	fond de rayon	1670	1253	4594	4594	6264	6264
3	Auchan	0% 60CL	GRENADIN	fond de rayon	11797	10858	13781	12946	17957	18792
4	Auchan	0% 60CL	MENTHE V	fond de rayon	8352	6682	10022	10440	15451	16704
5	Auchan	CLA 75 CL	GRENADIN	fond de rayon	24750	9000	29250	46688	46688	36000
6	Auchan	CLA 75 CL	MENTHE V	fond de rayon	16688	8438	21938	20813	25875	12375
7	Auchan	Plaisir 60C	ANIS	fond de rayon	835	4594	3341	4594	6682	7517
8	Auchan	Plaisir 60C	BANANE/F	fond de rayon	9605	5011	7934	10022	16286	15451
9	Auchan	Plaisir 60C	CASSIS	fond de rayon	15869	13781	15034	18374	19210	26726
10	Auchan	Plaisir 60C	CITRON V	fond de rayon	7517	6682	7099	10440	15451	19627
11	Auchan	Plaisir 60C	FRAISE	fond de rayon	29023	20880	30902	37166	33826	56376
12	Auchan	Plaisir 60C	FRAMBOIS	fond de rayon	8770	9187	10022	14198	15869	17122
13	Auchan	Plaisir 60C	FRUITS PA	fond de rayon	6264	6264	10022	12528	12946	19209
14	Auchan	Plaisir 60C	MOJITO	fond de rayon	0	0	0	0	0	0
15	Auchan	Plaisir 60C	ORANGE	fond de rayon	11275	8352	10858	12110	13781	13363
16	Auchan	Plaisir 60C	PAMPLEM	fond de rayon	835	3341	4594	4176	8770	7934
17	Auchan	Plaisir 60C	PECHE	fond de rayon	21298	16704	22968	25891	33826	32573
18	Auchan	Plaisir 60C	POMME F	fond de rayon	12528	11275	13781	13781	18792	18374
19	Francap	100CL	CITRON	fond de rayon	1368	3648	2736	4104	5016	4104

10. Nouvelle base de données

3.2 Séparation de la nouvelle base de données pour les statistiques

Une fois la base de données triée, il a fallu la séparer en deux parties jointes par des identifiants afin de faciliter son utilisation par l'équipe de réalisation statistique. En effet, dans la mesure où il y aurait des calculs réalisés sur la table des données, il fallait qu'elle soit la plus simple possible, avec le moins de données parasites. Notre première idée était de créer des identifiants à partir de la marque, la gamme et le parfum de chaque produit, permettant ainsi de rapidement avoir toutes les informations sur le sirop étudié. Ainsi, on avait des identifiants commençant par la marque en une lettre, puis composés de trois lettres représentant la gamme et enfin de trois lettres pour le parfum.

Cependant, cette méthode n'est pas réalisable car il y a trop de cas particuliers. Par exemple, si l'on prend les trois premières lettres des parfums, on se retrouve avec les mêmes identifiants pour les parfums "Fraise" et "Framboise". De plus, si le client ajoute des marques, gammes ou parfums, il faudrait gérer les nouveaux cas particuliers. Pour contourner ce problème, les identifiants sont devenus des nombres entiers commençant à 1.

	A	B	C	D	E	F	G	H	I	J
1	Années\Id	1	2	3	4	5	6	7	8	9
2	1 / 2018	1670	11797	8352	24750	16688	835	9605	15869	7517
3	2 / 2018	1253	10858	6682	9000	8438	4594	5011	13781	6682
4	3 / 2018	4594	13781	10022	29250	21938	3341	7934	15034	7099
5	4 / 2018	4594	12946	10440	46688	20813	4594	10022	18374	10440
6	5 / 2018	6264	17957	15451	46688	25875	6682	16286	19210	15451
7	6 / 2018	6264	18792	16704	36000	12375	7517	15451	26726	19627
8	7 / 2018	6264	16704	18792	56250	26662	5429	10440	10858	13363
9	8 / 2018	5846	16704	21715	43313	29025	8770	19210	23386	21715
10	9 / 2018	1253	9605	2923	22500	10688	3341	10440	15869	4176
11	10 / 2018	4594	13781	13363	40500	18000	8769	17121	25890	21297
12	11 / 2018	2923	11693	8352	31500	15750	1253	6682	10440	4176
13	12 / 2018	2088	10858	9187	38813	16313	2923	9605	13363	4176
14	1 / 2019	3758	12110	7934	30375	12938	2923	10440	17539	8352
15	2 / 2019	2923	15869	11275	38813	16875	5429	11693	19627	11275
16	3 / 2019	4176	11275	10858	35438	15750	3640	12204	14170	7826

11. Base de données composée des valeurs concernant les sirops associés à leur identifiant

Nous avons donc isolé d'un côté les valeurs de ventes des produits associés à leur identifiant et de l'autre les marques, gammes, parfums et ventes en litres, également associés à leur identifiant.

	A	B	C	D	E	F
1	Identifiant	Marque-En	Gamme	Parfum	Ventes en litres	
2	1	Auchan	0% 60CL	AGRUMES	fond de rayon	
3	2	Auchan	0% 60CL	GRENADINE	fond de rayon	
4	3	Auchan	0% 60CL	MENTHE VERTE	fond de rayon	
5	4	Auchan	CLA 75 CL	GRENADINE	fond de rayon	
6	5	Auchan	CLA 75 CL	MENTHE VERTE	fond de rayon	
7	6	Auchan	Plaisir 60CL	ANIS	fond de rayon	
8	7	Auchan	Plaisir 60CL	BANANE/FRAISE	fond de rayon	
9	8	Auchan	Plaisir 60CL	CASSIS	fond de rayon	
10	9	Auchan	Plaisir 60CL	CITRON VERT	fond de rayon	
11	10	Auchan	Plaisir 60CL	FRAISE	fond de rayon	
12	11	Auchan	Plaisir 60CL	FRAMBOISE	fond de rayon	
13	12	Auchan	Plaisir 60CL	FRUITS PASSION	fond de rayon	
14	13	Auchan	Plaisir 60CL	MOJITO	fond de rayon	
15	14	Auchan	Plaisir 60CL	ORANGE	fond de rayon	
16	15	Auchan	Plaisir 60CL	PAMPLEM. ROSE	fond de rayon	

12. Base de données composée des informations concernant les sirops associées à leur identifiant

Le logiciel RStudio* ne lisant pas les fichiers .xlsx*, nous avons ensuite converti les deux nouvelles bases de données obtenues en fichiers .csv*. Cela permet d'avoir toutes les données dans une seule case, séparées par des virgules.

	A	B
1	Identifiants,Marque-Enseigne,Gamme,Parfum,Ventes en litres	
2	1,Auchan,0% 60CL,AGRUMES,fond de rayon	
3	2,Auchan,0% 60CL,GRENADINE,fond de rayon	
4	3,Auchan,0% 60CL,MENTHE VERTE,fond de rayon	
5	4,Auchan,CLA 75 CL,GRENADINE,fond de rayon	
6	5,Auchan,CLA 75 CL,MENTHE VERTE,fond de rayon	
7	6,Auchan,Plaisir 60CL,ANIS,fond de rayon	
8	7,Auchan,Plaisir 60CL,BANANE/FRAISE,fond de rayon	
9	8,Auchan,Plaisir 60CL,CASSIS,fond de rayon	
10	9,Auchan,Plaisir 60CL,CITRON VERT,fond de rayon	
11	10,Auchan,Plaisir 60CL,FRAISE,fond de rayon	
12	11,Auchan,Plaisir 60CL,FRAMBOISE,fond de rayon	
13	12,Auchan,Plaisir 60CL,FRUITS PASSION,fond de rayon	
14	13,Auchan,Plaisir 60CL,MOJITO,fond de rayon	
15	14,Auchan,Plaisir 60CL,ORANGE,fond de rayon	
16	15,Auchan,Plaisir 60CL,PAMPLEM. ROSE,fond de rayon	

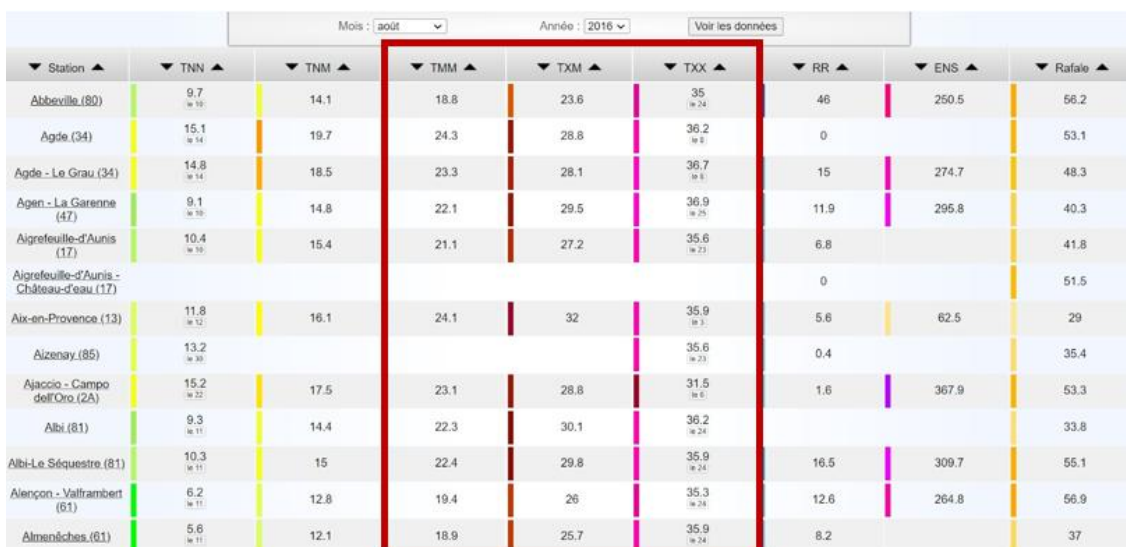
13. Version en extension .csv du fichier des identifiants

Web scraping

Nous avons choisi le site [infoclimat.fr](https://www.infoclimat.fr) pour créer la base de données des archives météo des quatre dernières années à l'aide du web scraping. Le site regroupe toutes les informations dont nous avons besoin et il ne comporte pas d'animations qui pourraient rendre l'extraction des données difficile. En effet, nous pouvons naviguer assez facilement entre les pages en modifiant les mois et les années directement dans l'url <https://www.infoclimat.fr/stations-meteo/analyses-mensuelles.php?mois=01&annee=2018> et récupérer précisément des données ciblées.

Pour ce faire, nous avons choisi de coder en Python* car il s'agit d'un langage qui nous est familier. De plus, il comporte des bibliothèques très utilisées dans ce domaine telles que BeautifulSoup, Request ou encore Pandas.

Nous nous sommes concertés avec Madame Sophie GUISNEL, au sujet de la récolte des informations et avons décidé de prendre les données entre janvier 2018 et décembre 2021 pour correspondre à la base de données des ventes de sirops fournie. Avec le binôme qui s'occupe des statistiques, nous avons convenu de récupérer les températures moyennes, températures moyennes maximales et températures maximales extrêmes du mois, pour toutes les stations.



Station	TNN	TNM	TMM	TXM	TXX	RR	ENS	Rafale
Abbeville (80)	9,7 (n 19)	14,1	18,8	23,6	35 (n 24)	46	250,5	56,2
Agde (34)	15,1 (n 14)	19,7	24,3	28,8	36,2 (n 8)	0		53,1
Agde - Le Grau (34)	14,8 (n 14)	18,5	23,3	28,1	36,7 (n 8)	15	274,7	48,3
Agen - La Garenne (47)	9,1 (n 10)	14,8	22,1	29,5	36,9 (n 25)	11,9	295,8	40,3
Aigrefeuille-d'Aunis (17)	10,4 (n 10)	15,4	21,1	27,2	35,6 (n 23)	6,8		41,8
Aigrefeuille-d'Aunis - Château-d'eau (17)						0		51,5
Aix-en-Provence (13)	11,8 (n 12)	16,1	24,1	32	35,9 (n 3)	5,6	62,5	29
Aizenay (85)	13,2 (n 28)				35,6 (n 23)	0,4		35,4
Ajaccio - Campo dell'Oro (2A)	15,2 (n 22)	17,5	23,1	28,8	31,5 (n 6)	1,6	367,9	53,3
Albi (81)	9,3 (n 11)	14,4	22,3	30,1	36,2 (n 24)			33,8
Albi-Le Séquestre (81)	10,3 (n 11)	15	22,4	29,8	35,9 (n 24)	16,5	309,7	55,1
Alençon - Valfrembert (61)	6,2 (n 11)	12,8	19,4	26	35,3 (n 24)	12,6	264,8	56,9
Almenèches (61)	5,6 (n 11)	12,1	18,9	25,7	35,9 (n 24)	8,2		37

14. Image du site [infoclimat.fr](https://www.infoclimat.fr), utilisé pour la récolte des données météorologiques
Encadré : les températures que l'on souhaite récupérer

Afin de récupérer ces informations, nous avons utilisé la bibliothèque Request pour accéder au site internet et la bibliothèque BeautifulSoup pour analyser la page HTML et récupérer les données.

Une fois la première version du code terminée, nous avons réalisé que sur le site utilisé, d'un mois à l'autre, les stations d'observation changeaient. Certaines étaient supprimées et d'autres ajoutées, ce qui a rendu notre code obsolète, puisque valable uniquement mois par mois. Nous avons dû trouver une solution pour associer une température à un mois et une station donnés. Nous avons effectivement deux listes indépendantes pour stocker ces informations. La solution a été les dictionnaires, dans lesquels nous avons pu assigner à chaque station une date et une mesure. Ensuite, nous avons utilisé la bibliothèque Pandas, pour transformer ces dictionnaires en dataframe et les écrire dans un fichier csv.

Il a également fallu adapter l'écriture des données dans le fichier csv pour une utilisation plus facile par le groupe statistique. Nous leur avons transmis trois fichiers différents pour chaque type de mesure qui nous intéressent avec pour chacun, les dates en colonne et les départements en ligne.

	A	B	C	D	E
1	Date	departement 01	departement 02	departement 03	departement 04
2	janv-18	6.79	6.74	7.16	4.808333333
3	févr-18	1.26	1.22	1.14	0.991666667
4	mars-18	6.82	6.1	7.616666667	4.508333333
5	avr-18	13.74	12.62	13.11666667	10.72727273
6	mai-18	16.29090909	15.66	14.9	13.35
7	juin-18	19.42	18.02	18.61666667	17.09166667
8	juil-18	22.39	21.82	21.86666667	20.33333333
9	août-18	22.01	19.46	21.53333333	18.62
10	sept-18	18.65	15.92	18.46666667	16.47
11	oct-18	12.67	12.52	13.13333333	11.8
12	nov-18	7.69	7.44	8.666666667	5.48
13	déc-18	5.09	5.98	6.133333333	3.154545455
14	janv-19	1.8	3.38	2.4	-0.21818182
15	févr-19	5.81	6.56	7.485714286	4.736363636
16	mars-19	8.45	8.66	8.742857143	5.909090909
17	avr-19	10.61	10.66	10.41428571	6.890909091
18	mai-19	12.55	12.12	12.28571429	9.663636364
19	juin-19	20.1	18.64	19.61428571	17.78181818

15. Base de données formée à partir du site infoclimat.fr

Le site contient environ 850 stations par mois. Une telle quantité de données n'était pas nécessaire mais nous ne voulions pas en perdre par souci de précision, c'est pourquoi nous avons opté pour une moyenne départementale des observations de température par mois. Nous avons également supprimé les stations d'outre-mer que la base de données fournie par notre client ne couvrait pas.

Pour obtenir une moyenne des températures par département, nous avons récupéré, à l'aide de la bibliothèque Re, le numéro de département. Celui-ci se situe dans le nom de chaque station et correspond au dernier texte entre parenthèses.

Aouste-sur-Sye (Saint-
Pierre) (26)

16. Une station dont une partie du nom est aussi entre parenthèses

Une fois cette base de données complétée, elle a été transmise au groupe d'études statistiques.

Analyse statistique

Nous avons choisi de travailler sur RStudio car le langage R* est idéal pour faire de l'analyse statistique et notre contact client y est familier.

5.1 Calcul de la corrélation entre les ventes de sirops et la température

La première partie de notre travail consistait à calculer la corrélation entre les ventes de sirops de l'entreprise Britvic et la température. La première chose que nous avons faite a été de regrouper toutes les informations que nous avions sur ce sujet. Une fois cette tâche accomplie et n'ayant pas encore accès aux bases de données, nous avons décidé d'en créer des fausses afin de pouvoir tester notre code. Cependant, nous voulions tout de même que nos bases de données soient les plus réalistes possible. Par conséquent, nous avons décidé, pour les ventes, de reprendre le fichier Excel fourni par l'entreprise puis de le trier rapidement à la main. En ce qui concerne les données des températures, nous avons décidé de prendre aléatoirement les informations de deux stations, Albi et Aix-en-Provence, du site de météorologie et de les rentrer à la main dans un fichier Excel. Enfin, nous avons exporté ces fichiers au format csv pour qu'ils soient lisibles avec RStudio.

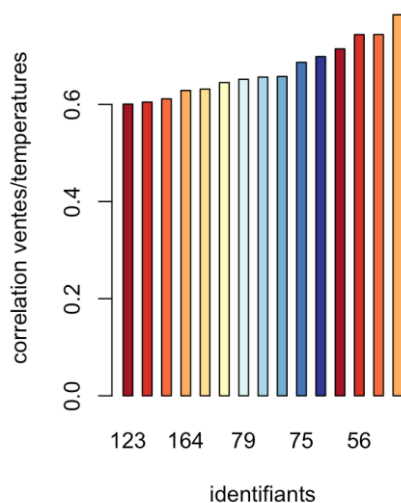
Nous avons alors pu commencer à travailler sur une fonction que nous avons nommée *correlation_tri()* et qui calcule la corrélation entre les ventes de sirops et les températures moyennes sur la France.

Le premier problème qui s'est imposé à nous concernait les périodes durant lesquelles certains sirops n'étaient pas encore ou plus vendus. En effet, si nous avions pris ces périodes en compte dans nos calculs, cela aurait faussé les résultats. Pour pallier cette complication, nous avons alors créé la fonction *date()* qui permet de récupérer les dates de début et de fin de ventes de chaque produit. Suite à cela, nous avons pu modifier la fonction *correlation_tri()* pour qu'elle calcule la corrélation seulement entre ces deux dates. Une fois la fonction exécutée, nous obtenons le tableau suivant représentant, par ordre croissant, la corrélation de chacun des sirops avec la température.

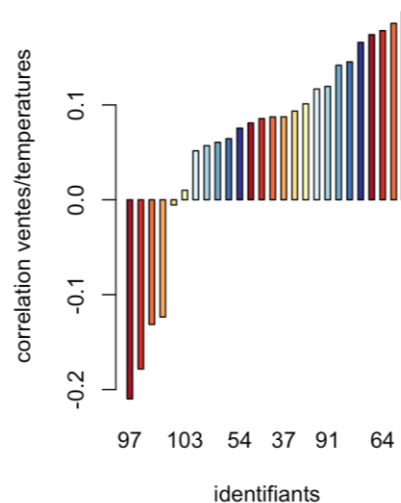
nom_sirops	correlation			
97 SU CLA 75 CL GRENADINE	-0.209896290656696	160 T ZEROS 0% CITRON	0.604839838986434	
102 SU GOU 75CL FRAISE	-0.178391914537658	118 T BIO MENTHE VERTE	0.611524203790365	
100 SU GOU 75CL CASSIS	-0.131444360106321	164 T ZEROS 0% MENTHE VERTE	0.628754272938783	
107 SU GOU 75CL PECHE	-0.123549574064862	46 Galec Max CITRON	0.631506407660095	
136 T MEGA PECHE ABRICOT	-0.00550292564108554	44 Galec CLA 75 CL MENTHE VERTE	0.645140038065074	
103 SU GOU 75CL FRAMBOISE	0.0100250788283846	79 MDV MDV tradition (classique 70cl) MENTHE VERTE	0.65181404309145	
60 MDV MDV petits producteurs CASSIS	0.0516499714274091	3 Auchan 0% 60CL MENTHE VERTE	0.656279431832674	
74 MDV MDV tradition (Plaisir 70cl) MÛRE	0.0570709915896126	83 MDV MDV tradition (recette 70cl AN++) RECETTE P...	0.657550421671509	
137 T MEGA POMME FRAMBOISE CASSIS	0.0605532271038709	75 MDV MDV tradition (Plaisir 70cl) ORGEAT	0.686769264150048	
96 SU CLA 75 CL CITRON	0.0644488222985687	129 T CONC DE GOUT MOJITO	0.698613705628899	
54 Galec Plaisir FRAMBOISE	0.0755417358642309	66 MDV MDV tradition (Plaisir 70cl) ANIS	0.71445630531887	
63 MDV MDV sirops de camille FRAISE DES BOIS	0.0810038388357814	56 Galec Plaisir MOJITO	0.74391525163373	
51 Galec Plaisir CASSIS	0.0855377170803275	150 T PLAISIRS 60CL MENTHE GLACIALE	0.744009434134621	
		72 MDV MDV tradition (Plaisir 70cl) MENTHE GLACIALE	0.784812965484921	

17. Une partie du tableau trié des corrélations par sirop

Nous avons aussi créé une fonction `correlation_graphes()` qui trace, à partir du tableau précédent, deux diagrammes. Le premier représente les sirops ayant une forte corrélation avec la température, c'est-à-dire supérieure à 0.6 et le deuxième les sirops ayant une faible corrélation avec la température, c'est-à-dire inférieure à 0.2. Nous avons fait le choix de ces seuils car nous savons que plus une corrélation est proche de 1, plus elle est forte et plus elle est proche de 0, plus elle est faible. Nous avons dans un premier temps choisis un seuil de 0.8 pour la forte corrélation et de 0.2 pour la faible corrélation. Cependant, nous n'avons pas beaucoup de données pour la forte corrélation. Nous avons donc décidé de définir, en accord avec Madame GUISNEL, le seuil à 0.6.



18. Diagramme des fortes corrélations par sirop



19. Diagramme des faibles corrélations par sirop

Sur les diagrammes précédents, nous pouvons observer que les ventes de certains sirops, comme le sirop MDV Menthe Glaciale ou le sirop T Menthe Glaciale, sont très corrélés à la température. Nous pouvons donc penser que si les températures augmentent, une augmentation des ventes se fera ressentir. À contrario, d'autres sirops comme le sirop SU

Grenadine ne sont pas corrélés. On peut alors penser que la variation des températures n'aura que peu d'effet sur ces ventes-ci.

Par la suite, notre contact client nous a demandé si l'on pouvait calculer cette corrélation, non plus en prenant en compte les sirops indépendamment les uns des autres, mais en les regroupant par parfum puis par gamme et enfin par marque. Nous avons alors créé la fonction *correlation_par_type()* qui permet d'obtenir les tableaux suivants.

nom_sirops	correlation	27	PECHE BLANCHE	0.511796782571123
25 MÛRE	0.0570709915896126	38	PINA COLADA	0.524711344415857
41 CLEMENTINE	0.235395605904719	18	ABRICOT PASSION	0.532398612680446
39 ABRICOT	0.23690390119849	13	PAMPLEM. ROSE	0.554184328192103
15 POMME FRAMBOISE CASSIS	0.23916398469504	40	CERISE	0.554580232776667
6 CASSIS	0.240218737271063	16	CITRON	0.556168107338966
44 POMME	0.247571162638615	21	Pêche blanche	0.560941526416527
9 FRAMBOISE	0.247619067762731	7	CITRON VERT	0.599992819172827
20 FRAISE DES BOIS	0.25145687384416	3	MENTHE VERTE	0.635520415722518
8 FRAISE	0.265738089693894	30	RECETTE PROVENC	0.657550421671509
29 PECHE ABRICOT	0.266620226774665	11	MOJITO	0.67043844435473
35 PAMPLEMOUSSE	0.269561472197754	26	ORGEAT	0.686769264150048
2 GRENADINE	0.29969917246023	4	ANIS	0.725688625927052
12 ORANGE	0.315007484577364	23	MENTHE GLACIALE	0.790000207048745

20. Une partie du tableau trié des corrélations par parfum

Comme précédemment, nous pouvons remarquer que les ventes de certains parfums sont influencées par la température, en particulier la Menthe Glaciale, l'Anis et l'Orgeat, ce qui correspond à nos attentes.

nom_sirops	correlation	6	Plaisir 75CL	0.388359584323203
19 GOU 75CL	0.107803162848217	21	CLASSIQUES	0.39839195051083
26 PLAISIRS 60CL AROMES NAT	0.160433459587301	2	CLA 75 CL	0.430640147089222
11 MDV petits producteurs	0.172612699656036	25	PLAISIRS 60CL	0.45729811507136
12 MDV sirops de camille	0.200799313802186	16	MDV tradition (recette 70cl AN++)	0.49679526410466
15 MDV tradition (classique 70cl AN++)	0.290696240244287	9	Max	0.511903205929783
10 Plaisir	0.298369384277401	8	Bidon+ fruité	0.537615110110343
4 100CL	0.305358506011837	1	0% 60CL	0.56069017217964
20 SS SUCRE	0.355401962857154	27	ZEROS 0%	0.573719530275515
5 BIO (bti 50cl)	0.355710311000385	17	MDV tradition (recette 70cl)	0.574310309223651
18 BIO 50cl	0.376180256062445	22	CLASSIQUES AROMES NAT	0.583377852973672
24 MEGA	0.378379167239604	14	MDV tradition (classique 70cl)	0.62941417131096
7 BIO	0.383891789936932	23	CONC DE GOUT	0.632692247881096
3 Plaisir 60CL	0.384026206884611	13	MDV tradition (Plaisir 70cl)	0.63663520344041

21. Le tableau trié des corrélations par gamme

	nom_sirops	correlation
1	MDD	0.434834872979518
3	T	0.496542766076634
2	MDV	0.634435444610627

22. Tableau trié des corrélations par marque

Le tableau de corrélations par marque nous indique que les MDD sont moins “météo-sensibles” que MDV et T. Cette observation conforte les hypothèses de l’entreprise à savoir que les consommateurs occasionnels de sirops, c’est-à-dire ceux qui n’en boivent qu’en été, ont plutôt tendance à acheter des sirops de marques. À l’inverse, les consommateurs réguliers, autrement dit ceux qui en boivent toute l’année, ont tendance à acheter des sirops de MDD.

Les tableaux de corrélations par parfum, par gamme et par marque ont un réel intérêt pour Britvic car ils peuvent donner des informations sur le comportement de leurs clients. Toutefois, pour la suite de nos calculs, nous avons décidé de ne pas approfondir l’analyse de ces données.

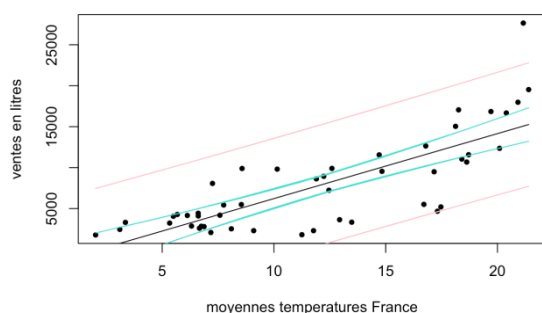
Pour finir, Madame GUISNEL nous a demandé s’il était possible de mettre ces quatre tableaux dans des fichiers Excel. RStudio ne travaillant que sur des fichiers .csv, nous en avons donc créé un par tableau.

5.2 Estimation de l'impact d'un été chaud sur les ventes de sirops

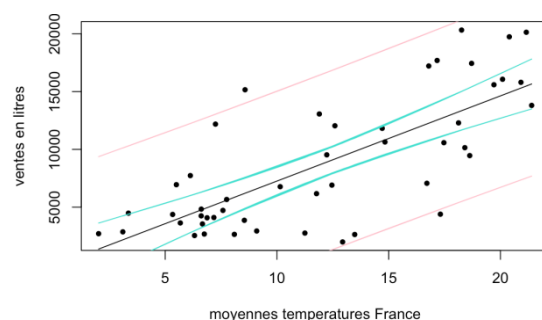
La deuxième partie de notre travail consistait à estimer l'impact d'un été chaud sur les ventes de sirops de l'entreprise. Pour cela, nous avons décidé, grâce à une régression linéaire*, de linéariser le modèle des ventes de sirops en fonction de la température. Nous avons donc défini une fonction qui trace, dans un premier temps, les données de ventes d'un sirop choisi en fonction des températures. Par la suite, celle-ci calcule une régression linéaire, grâce à la fonction $lm()$ de R, sur ce sirop et la trace sur le même graphique.

Ensuite, pour estimer au mieux l'impact d'un été chaud sur les ventes de sirops de l'été prochain, nous avons décidé, à partir de ce modèle linéarisé, de calculer un intervalle de confiance et un intervalle de prédiction* à un niveau de confiance de 95%. Nous avons choisi ce niveau de confiance pour que nos intervalles soient le plus précis possible. Notre fonction se charge donc aussi de ces calculs dont elle trace les intervalles. L'intervalle de confiance permet d'appuyer la précision de notre régression linéaire. L'intervalle de prédiction, quant à lui, permet d'estimer assez précisément que, pour une température donnée, le total des ventes se situera dans cet intervalle.

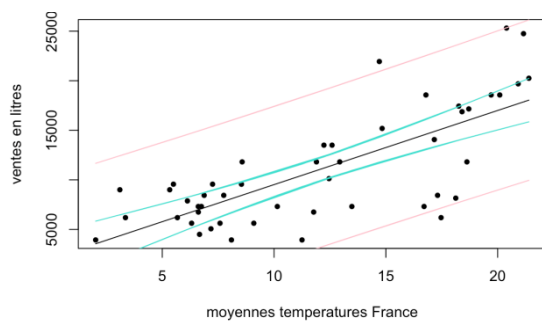
Ci-dessous, se trouvent les graphiques obtenus pour les trois sirops les plus corrélés, à savoir le sirop MDV Menthe Glaciale, le sirop T Menthe Glaciale et le sirop Galec Mojito. Nous avons choisi ces sirops car le schéma de régression linéaire n'est intéressant que s'il y a corrélation. En effet, si nous avions pris des sirops non corrélés avec la météo, les points du graphe n'auraient pas suivi une droite et la régression linéaire aurait été impossible.



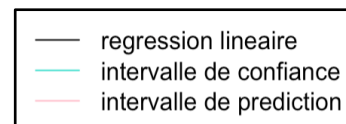
23. Graphe de la régression linéaire des ventes du sirop MDV tradition (Plaisir 70cl) Menthe Glaciale



24. Graphe de la régression linéaire des ventes du sirop T Plaisirs 60cl Menthe Glaciale



25. Graphe de la régression linéaire des ventes du sirop Galec Plaisir Mojito



26. Légende des 3 graphes de régression linéaire

La régression linéaire obtenue n'est pas très précise car beaucoup de points sont éloignés de la droite. Ceci rend notre intervalle de prédiction très grand, plus ou moins 7 500 L/mois par rapport à la régression linéaire, pour le sirop MDV Menthe glaciale. Ce phénomène est dû aux nombreux facteurs extérieurs qui influent également sur les ventes de sirops et qui rendent, si l'on ne les connaît pas, la prédiction très difficile.

Au vu des résultats obtenus, nous avons décidé d'exécuter notre code avec les températures moyennes maximales en France puis avec les températures maximales extrêmes. Nous pensions obtenir des résultats plus précis, ce qui nous aurait permis d'estimer plus rigoureusement les ventes pour l'été prochain. Seulement, il s'est avéré que la précision gagnée n'a pas été suffisante pour apporter de nouvelles informations.

CONCLUSION

Lors de ce stage applicatif, nous avons étudié la corrélation entre les ventes des différents sirops et la météo afin d'estimer l'impact d'un été chaud sur celles-ci. Nous n'avons eu aucun mal à montrer la corrélation entre certains sirops et la température. Cependant, lorsqu'il s'agit d'estimer une éventuelle augmentation des ventes due à un été chaud en tenant compte de la corrélation obtenue, la tâche n'est pas si aisée. En effet, de nombreux autres facteurs, dont nous n'avons pas connaissance, entrent en jeu.

Ce projet que nous avons réalisé pendant trois semaines, nous a permis de mettre en pratique des connaissances théoriques vues en cours de Probabilités et Statistiques, de découvrir l'implémentation du web scraping, la manipulation de feuilles de calcul via Python et la notion d'intervalle de prédiction. Cela nous a également permis de devenir plus rigoureux, d'améliorer notre professionnalisme et notre capacité à mener à bien un projet.

Durant la phase active du projet, nous avons dû redéfinir, au fur et à mesure, les tâches à exécuter ainsi que les effectifs mis en place sur chaque tâche. Malgré cela, nous avons pu réaliser l'ensemble de ces tâches dans les temps.

En plus d'avoir été enrichissant d'un point de vue technique, ce projet a été instructif d'un point de vue humain. Grâce au projet, nous avons amélioré notre capacité à travailler en groupe, à nous organiser, à répartir des tâches et à nous adapter.

TABLE DES FIGURES

1. IMPACT DE LA CHALEUR SUR LES VENTES DE SIROPS	4
2. LEGENDE DU DIAGRAMME DE GANTT	7
3. DIAGRAMME DE GANTT DE LA PRECONCEPTION.....	7
4. DIAGRAMME DE GANTT DE LA CONCEPTION.....	7
5. DIAGRAMME DE GANTT DE LA POST CONCEPTION.....	7
6. DIAGRAMME DE GANTT OBSERVE DE LA CONCEPTION	8
7. DIAGRAMME DE GANTT OBSERVE DE LA POST CONCEPTION	8
8. MAUVAISE MISE EN FORME DES DATES.....	10
9. MAUVAISE RECUPERATION DE LA COLONNE DES ENSEIGNES.....	11
10. NOUVELLE BASE DE DONNEES	11
11. BASE DE DONNEES COMPOSEE DES VALEURS CONCERNANT LES SIROPS ASSOCIES A LEUR IDENTIFIANT	12
12. BASE DE DONNEES COMPOSEE DES INFORMATIONS CONCERNANT LES SIROPS ASSOCIEES A LEUR IDENTIFIANT	13
13. VERSION EN EXTENSION .CSV DU FICHIER DES IDENTIFIANTS	13
14. IMAGE DU SITE INFOCLIMAT.FR, UTILISE POUR LA RECOLTE DES DONNEES METEOROLOGIQUES ENCADRE : LES TEMPERATURES QUE L'ON SOUHAITE RECUPERER.....	14
15. BASE DE DONNEES FORMEE A PARTIR DU SITE INFOCLIMAT.FR.....	15
16. UNE STATION DONT UNE PARTIE DU NOM EST AUSSI ENTRE PARENTHESES	16
17. UNE PARTIE DU TABLEAU TRIE DES CORRELATIONS PAR SIROP	18
18. DIAGRAMME DES FORTES CORRELATIONS PAR SIROP.....	18
19. DIAGRAMME DES FAIBLES CORRELATIONS PAR SIROP.....	18
20. UNE PARTIE DU TABLEAU TRIE DES CORRELATIONS PAR PARFUM.....	19
21. LE TABLEAU TRIE DES CORRELATIONS PAR GAMME	19
22. TABLEAU TRIE DES CORRELATIONS PAR MARQUE.....	20
23. GRAPHE DE LA REGRESSION LINEAIRE DES VENTES DU SIROP MDV TRADITION (PLAISIR 70CL) MENTHE GLACIALE ...	21
24. GRAPHE DE LA REGRESSION LINEAIRE DES VENTES DU SIROP T PLAISIRS 60CL MENTHE GLACIALE	21
25. GRAPHE DE LA REGRESSION LINEAIRE DES VENTES DU SIROP GALEC PLAISIR MOJITO.....	22
26. LEGENDE DES 3 GRAPHS DE REGRESSION LINEAIRE	22

GLOSSAIRE

Base de données : Un ensemble d'informations qui est organisé de manière à être facilement accessible, géré et mis à jour. Elle est utilisée par les organisations comme méthode de stockage, de gestion et de récupération de l'information.

Bibliothèque : Un ensemble de fonctions utilitaires, regroupées et mises à disposition afin de pouvoir être utilisées sans avoir à les réécrire.

Corrélation : Un lien entre deux caractères quantitatifs d'une distribution qui décrit le type, le sens et la force de ce lien. Lorsque la corrélation est forte on parlera de corrélation par abus de langage et lorsque la corrélation est faible on parlera de “non corrélation”.

Csv : Un fichier CSV (en anglais, Comma Separated Values) est le fichier de base des données recueillies, sans formatage particulier et dont le champ est séparé par une virgule.

Dataframe : Une matrice pouvant avoir des colonnes de types différents (numérique, texte, facteur, ...).

Excel : Un logiciel tableur de la suite bureautique Microsoft Office développé et distribué par l'éditeur Microsoft.

Feuille de calcul : Un document de base des tableurs, constitué d'un tableau de grande taille destiné à contenir des données et / ou des formules calculées dynamiquement.

Intervalle de prédiction : Une estimation d'un intervalle dans lequel une observation future tombera, avec une certaine probabilité, compte tenu de ce qui a déjà été observé.

Python : Un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet.

R : Un langage de programmation destiné aux statistiques et à la science des données.

Régression linéaire : Une modélisation linéaire qui permet d'établir des estimations dans le futur à partir d'informations provenant du passé.

RStudio : Un environnement de développement gratuit, libre et multiplateforme pour R.

Web scraping : Une technique d'extraction du contenu de sites Web, via un script ou un programme.

Xlsx : Le format ordinaire des fichiers enregistrés sans macros avec Microsoft Excel.

WEBOGRAPHIE

Excel :

- Documentation pour les dates : <https://medium.com/france-school-of-ai/travailler-facilement-avec-les-dates-sur-pandas-14b14b2ea51>
- Documentation Xlsxwriter :
<https://xlsxwriter.readthedocs.io/>

Web scraping + Excel :

- Documentation Pandas :
<https://pandas.pydata.org/docs/index.html>
<https://www.delftstack.com/fr/howto/python-pandas/>

Web scraping :

- Vidéo introductive sur le web scraping :
https://www.youtube.com/watch?v=XQgXKtPSzUI&list=PL8eNk_zTBST-SaABhXwBFbKvvA0tIRSRV
- Documentation Beautiful Soup :
<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- Site de récupération des données météorologiques :
<https://www.infoclimat.fr/stations-meteo/analyses-mensuelles.php?>

Statistiques :

- Cours sur les vecteurs aléatoires de Monsieur Jean-François COEURJOLLY
- Calcul de la régression linéaire :
<https://fermin.perso.math.cnrs.fr/Files/Chap3.pdf>
- Calculs des intervalles de confiance et de prédiction:
<https://delladata.fr/intervalle-confiance-intervalle-prediction/#:~:text=L'intervalle%20de%20confiance%20%C3%A0,de%20r%C3%A9gression%20de%20la%20population>
- Documentation R :
<https://www.rdocumentation.org>