

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Carthage
Ecole Nationale d'Ingénieurs de Carthage

Rapport de Stage

Spécialité : Informatique

Elaboré par

Farah OUESLETI

Sales & products Forecasting Customer Segmentation

Réalisé au sein de

Medianet

MEDIANET



Encadré par

Hassen BENZARTI

Année universitaire
2023 - 2024

Remerciements

Je tiens à remercier Mr Hassen BENZARTI pour sa présence, et son soutien continu pour la réalisation du projet.

J'aimerais remercier l'Ecole Nationale D'ingénieurs de Carthage d'avoir inclu l'opportunité de réaliser des stages techniques pour ses étudiants de 2ème année, ce qui aboutit à une grande valeur ajoutée pour l'étudiant stagiaire.

Je tiens ainsi à remercier le mouvement des Juniors Entreprises pour m'avoir appris la vie en entreprise, en un milieu professionnel, à la prise de contact avec diverses parties prenantes entre autres clients et partenaires, ce qui a rendu mon intégration dans l'entreprise et ma communication avec l'équipe ainsi que mon superviseur très fluide et professionnelle en même temps.

Table des matières

Remerciements	1
Tables des matieres	2
Liste des figures	5
Introduction Générale	7
Présentation de l'entreprise	8
State Of the Art.....	9
Contexte du projet	11
Partie A : Prédiction des variations temporelles.....	15
Chapitre I : Data Understanding.....	15
A.I.1. Collecte des données	15
A.I.2. Description des données : Structure de données & MindMap	15
A.I.3. Exploration des données.....	20
A.I.3.1. Data DeNesting	20
A.I.3.2. Collecte d'informations à propos les données	21
A.I.3.3. Vérification de la qualité des données	22
Chapitre II : Data Preparation.....	23
A.II.1. Data Selection	23
A.II.2. Data Cleaning	23
A.II.3. Feature Engineering	23
A.II.4. Time Series Visualization.....	24
Chapitre III : Data Modeling	25
A.III.1. Choix préliminaire du modèle et raisons	25
A.III.2. Model Training	25
A.III.3. Decomposition of The Time Series	26
A.III.4. Differencing of The Time Series	27
A.III.4.1. Differencing	27
A.III.4.1.1. Produits du café	27
A.III.4.1.2. Produits du restaurant	28
A.III.4.2. Test AdFuller	28
A.III.5. Évaluation des données pour le choix des paramètres du modèle	29
A.III.5.1. Cas du café	29
A.III.5.2. Cas du restaurant	29
A.III.6. Model Training : Résultat du Fitting et résidus	30
A.III.6.1. Résumé du Training	30
A.III.6.2. Graphes des résidus et densité des résidus	30
A.III.6.2.1. Cas du café	30
A.III.6.2.2. Cas du restaurant	30
A.III.6.2.3. Interprétation des graphes des résidus	31
A.III.6.3. Fitted Model	32

A.III.6.3.1.	Fitted Model du cafétéria (paramètres initiaux)	32
A.III.6.3.2.	Fitted Model du restaurant.....	32
A.III.6.3.3.	Fitted Model du cafétéria (Tuned parameters)	32
Chapitre IV : Évaluation et Interprétation	34	
A.IV.1.	Evaluation par rapport au Fitting	34
A.IV.1.1.	Evaluation relative aux ventes du produit du café	34
A.IV.1.2.	Evaluation relative aux ventes du produit du restaurant	34
A.IV.2.	Evaluation par rapport à des prédictions préliminaires	35
Chapitre V : Solution proposée.....	36	
A.V.1.	Principe de la solution	36
A.V.2.	Data Preparation	37
A.V.2.1.	Data splitting	37
A.V.2.2.	Feature Engineering	37
A.V.3.	Time Series Visualization	38
A.V.4.	Conclusion sur la solution proposée	38
Chapitre VI : Par analogie	39	
A.VI.1.	Prédiction des quantités de produits	39
A.VI.1.1.	Prédictions des ventes du café	39
A.VI.1.1.1.	Modeling	39
A.VI.1.1.1.1.	Differencing the time Series	39
A.VI.1.1.1.2.	Evaluation des données et choix des paramètres	40
A.VI.1.1.1.3.	Graphes des résidus	40
A.VI.1.1.1.4.	Allure du Fitted Values graph	41
A.VI.1.1.2.	Interprétation des résultats	41
A.VI.1.2.	Prédictions des ventes du restaurant	41
A.VI.1.2.1.	Modeling	41
A.VI.1.2.1.1.	Differencing the time Series	41
A.VI.1.2.1.2.	Evaluation des données et choix des paramètres	42
A.VI.1.2.1.3.	Graphes des résidus	42
A.VI.1.2.1.4.	Allure du Fitted Values graph	43
A.VI.1.2.2.	Interprétation des résultats	43
A.VI.1.3.	Prédictions des livraisons et des à emporter	43
A.VI.1.3.1.	Visualisation des Time Series	43
A.VI.1.3.2.	Visualisation des graphes ACF et PACF	45
A.VI.1.3.3.	Visualisation des graphes des résidus	45
A.VI.1.3.4.	Evaluation	46
Conclusion Partie A	47	
Partie B :Segmentation de la base de clients de la start up	48	
Chapitre I : Business Understanding.....	49	
Chapitre II : Data Understanding.....	49	
B.II.1.	Exploration des données	49

B.II.2. Vérification de la qualité des données	50
Chapitre III : Data Preparation.....	51
B.III.1. Data DeNesting	51
B.III.2. Data Selection	52
B.III.3. Data Cleaning	52
Chapitre IV : Modeling	54
A.III.1. Choix préliminaire du modèle	54
A.III.2. Training	54
A.III.3. Résultat du Clustering.....	55
Chapitre IV : Évaluation et Interprétation	56
Conclusion Partie B	57
Conclusion générale et perspectives.....	58

Liste des figures

- Figure 1 : Méthode de CRISP-DM, Wikipédia
Figure 2 : Exemple de MindMap
Figure 3 : MindMap des données du projet
Figure 4 : Allure du symbole Nuage
Figure 5 : Carré aux coins ronds
Figure 6 : Carré aux coins rectangulaires
Figure 7 : MindMap agrandi
Figure 8 : MindMap de la colonne Orders, sujet de la partie A du projet
Figure 9 : Etat des commandes lors de l'importation
Figure 10 : Base de Données après traitement
Figure 11 : Types des variables et nombre
Figure 12 : Noms des colonnes
Figure 13 : Description générale des données
Figure 14 : Visualisation de la structure interne de la colonne
Figure 15 : Nombre de valeurs nulles par colonne “ingredient_used”
Figure 16 : Variation des ventes des produits du café
Figure 17 : Variation des ventes des produits du restaurant
Figure 18 : Décomposition additive des Time Series de la BD du café
Figure 19 : Décomposition additive des Time Series de la BD du café
Figure 20 : Allure des ventes du café avant Differencing
Figure 21 : Allure des ventes du café après Differencing
Figure 22 : Allure des ventes du restaurant avant Differencing
Figure 23 : allure des ventes du restaurant après Differencing
Figure 24 : Résultat du test adfuller - café
Figure 25 : Résultat du test adfuller - restaurant
Figure 26 : Allure des courbes des graphes ACF et PACF pour la cafétéria
Figure 27 : Allure des courbes des graphes ACF et PACF pour le restaurant
Figure 28 : Résultat des résumés du Training des modèles de café et restaurant respectivement
Figure 29 : Allure des graphes des résidus et densité des résidus
Figure 30 : Allure des graphes des résidus et densité des résidus
Figure 31 : Moyennes de la densité des résidus respectivement pour le café, et le restaurant
Figure 32 : Allure des graphes des résidus et densité des résidus du café
Figure 33 : Moyennes de la densité des résidus du café
Figure 34 : Allure du Fitted model par rapport aux valeurs réelles
Figure 35 : Allure du Fitted model par rapport aux valeurs réelles
Figure 36 : Allure du Fitted model par rapport aux valeurs réelles avec des Tuned parameters
Figure 37 : Résumé du graphe des résidus du modèle du restaurant
Figure 38 : Table créé avec les nouveaux Features
Figure 39 : Time Series des ventes du café avec des dates régulières
Figure 40 : allure des quantités vendues du café avant Differencing

- Figure 41 : Allure des quantités vendues du café après Differencing
Figure 42 : Allure des courbes des graphes ACF et PACF pour la cafétéria
Figure 43 : Allure des graphes des résidus et densité des résidus
Figure 44 : Résumé de l'analyse des densités des résidus
Figure 45 : Allure du Fitted model par rapport aux valeurs réelles
Figure 46 : Allure des quantités vendues du restaurant avant Differencing
Figure 47 : Allure des quantités vendues du restaurant après Differencing
Figure 48 : Allure des graphes des résidus et densité des résidus
Figure 49 : Résumé de l'analyse des densités des résidus
Figure 50 : Allure du Fitted model par rapport aux valeurs réelles
Figure 51 : Allure des variations du nombre de livraisons du restaurant
Figure 52 : Allure des variations du nombre des à emporter du café
Figure 53 : Allure des variations du nombre de livraisons du café différencié
Figure 54 : Allure des variations du nombre de livraisons du restaurant différencié
Figure 55 : Allure des graphes ACF et PACF pour les produits du café
Figure 56 : Allure des graphes ACF et PACF pour les produits du restaurant
Figure 57 : Allure des résidus et des densités des résidus pour les produits du café
Figure 58 : Allure des résidus et des densités des résidus pour les produits du restaurants
Figure 59 : Catégories des données et occurrences par colonne
Figure 60 : Description de la base de données des clients
Figure 61 : Informations concernant la base de données des clients
Figure 62 : Colonne Info avant DeNesting
Figure 63 : DataFrame après DeNesting
Figure 64 : DataFrame après sélection préliminaire des Features
Figure 65 : DataFrame après DeNesting des features sélectionnés
Figure 66 : DataFrame partiellement nettoyé
Figure 67 : DataFrame nettoyé
Figure 68 : Elbow Graph
Figure 69 : Plot des Clusters selon les critères choisis

Introduction générale

L'objectif du stage était principalement la montée en compétences techniques en matière de Sciences des données de manière complètement autonome en la mise en pratique de connaissances théoriques apprises de diverses sources en un projet ayant une finalité, une valeur ajoutée sur les parties prenantes et un impact.

Le projet avait eu comme entrée une Base De Données d'une startup partenaire à l'Entreprise avec laquelle le stage a été réalisé.

Vu la confidentialité des données, la base de données était basée sur une structure assez complexe et élaborée, avec peu de données pertinentes et beaucoup de données manquantes.

Présentation de l'entreprise

Medianet est une startup à présence internationale sur quatre continents. Ses domaines d'expertise incluent :

- Transformation digitale
- Communication 360°
- Growth Marketing
- E-réputation
- Formation

Parmi ses prestations de service principales :

- Maintenance préventive
- Maintenance Corrective
- Maintenance évolutive
- Infogérance et Hosting
- Webmastering
- Audit et Test Factory

Elle a un département IT, Qualité, Etude de marché et de projets, Marketing, Événementiel, etc.

C'est une startup, fondée par 3 CEOs, dont la vision est de créer un village de startups en collaboration, en effet, elle est l'une parmi trois autres : Startup Village, Express fm et Qualipro.

Elle a une culture très intéressante et un milieu de travail excellent.

Elle est aussi certifiée ISO 9001, qui certifie de la conformité et performance de son système de management.

State Of the Art

Avant d'entamer, tâchons d'introduire quelques concepts clés à la compréhension du travail présenté.

Nous aborderons l'Intelligence Artificielle(AI), le Machine Learning(ML), le Deep Learning(DL), et, ce qui représente la partie majoritaire de notre travail, Time Series Analysis.

L'intelligence artificielle est le champ de l'informatique qui vise à créer des systèmes capables d'imiter certaines fonctions cognitives humaines, telles que la perception, la résolution de problèmes et l'apprentissage.

L'apprentissage automatique, une branche de l'IA, se concentre sur le développement d'algorithmes et de modèles capables d'apprendre à partir de données et de prendre des décisions basées sur ces apprentissages. Ces deux domaines interagissent étroitement pour créer des systèmes intelligents et autonomes.

Le deep learning, ou apprentissage profond en français, est une sous-catégorie de l'intelligence artificielle, notamment, du ML. Il s'agit d'une approche de l'apprentissage automatique (machine learning) qui repose sur des réseaux de neurones artificiels profonds pour résoudre des tâches complexes.

Les séries temporelles, quant à elles, font référence à des données collectées ou enregistrées dans un ordre chronologique, souvent à intervalles réguliers. Les séries temporelles se retrouvent dans de nombreux domaines, de la finance à la météorologie en passant par l'économie, et elles présentent des caractéristiques uniques, telles que les tendances, les saisons et les cycles, qui les rendent complexes à analyser. L'analyse de séries temporelles consiste à extraire des informations significatives, à effectuer des prévisions et à prendre des décisions éclairées à partir de ces données.

Il serait pertinent d'introduire quelques termes que nous utiliserons plus tard, tel, UnderFitting et OverFitting.

Underfitting, est le fait que notre modèle donne de mauvais résultats sur Training Data et sur les Validation ou Test Data. Cela se produit lorsque notre modèle est très simple, lorsque les inputs features sont peu en valeur ou en pertinence, pour un modèle de Deep Learning, lorsqu'on introduit quelques couches à faibles nombre d'unités, ou forte régularisation.

Overfitting, est le fait que notre modèle donne un très bon résultat sur le Train Data, mais de mauvais résultats sur les Validation ou Test Data. Cela se produit lorsque le modèle est très complexe, et est fait pour convenir parfaitement au Training Data, comme l'utilisation de fonction de régression linéaire à polynômes à grand facteurs.

On parle de High bias et low variance pour l'Underfitting.
et de low bias et High variance pour l'Overfitting.

Parlons un peu plus en détail du ARIMA model, l'un des modèles classiques des Time Series.

ARIMA : Auto Regressive Integrated Moving Average

Il se divise en deux parties principalement :

AR : Représente une dépendance de valeurs entre un Lag à l'instant t, et tous ses précédents aux instants $t-i$ (ou i un entier)

MA : Représente une dépendance d'erreur entre un Lag à l'instant t, et toutes ses erreurs précédentes aux instants $t-i$ (ou i un entier)

Le terme d'erreur a tendance à être aléatoire, tandis que les variations de valeurs de AR sont déterminées par les anciennes valeurs, d'où ils forment un polynôme.

Le nombre de Lags de AR est noté p, alors que celui de MA est noté q

Tout autre terme cité et non expliqué, sera en note de bas de page.

Nous espérons que ce rapport soit facile de compréhension et qu'il atteigne les objectifs de sa réalisation.

Contexte du projet

1. Présentation de l'ensemble du projet

1.1. Contexte du projet

Ce projet vise à utiliser l’Intelligence Artificielle et le Machine Learning, notamment : Time Series pour aboutir à des prédictions des ventes et à tirer des informations pertinentes relatives aux données de différents restaurants et cafés.

1.2. Buts du projet

- Application de l’Intelligence Artificielle dans le domaine commercial pour améliorer la gestion de l’inventaire.
- Aboutir à des recommandations à la direction tirées des analyses de données relatives aux ventes par produit et par période.
- Prédiction des ventes et des revenus pour un prestataire particulier (dont le nom est omis en raison de confidentialité).
- Montée en compétences du stagiaire réalisant le projet et travaillant sur une base de données réelle.

1.3. Éléments d’entrée

Base de données fournie de la part d’une startup en convention avec l’entreprise d’accueil.

Cette base contient les données de caisse datées de plusieurs restaurants et cafés.

2. Description du projet

Le projet se divise en deux axes :

- En première instance, Nous traiterons un cas particulier de l’un des clients de la startup : un restaurant-café et son activité.
- En deuxième lieu, nous nous intéresserons à l’ensemble des clients (étant des restaurants-café) de la startup, en vue de les segmenter.

2.1. Phase d’analyse de données

Nous avons effectué une exploration globale des données pour comprendre leur types, leur contenu, évaluer leur qualité, ensuite, une exploration et une étude plus détaillée avait été réalisée pour chaque phase du projet. (la Base de Données globale et la Base de Données relative à l’étude de cas du restaurant).

Les grandes étapes suivent l’enchaînement du CRISP-DM dont nous expliquerons dans la suite.

- Compréhension et Description des données
- Exploration des données
- Préparation des données

- Nettoyage des données
- Visualisation

2.2. Phase de réalisation du modèle et prédictions

2.2.1. Par rapport au restaurant spécifique

Il existe un seul restaurant-café ayant les ordres passés, c'est celui d'indice 107 dans la base de données nettoyée.

Ainsi, seules les données de celui-ci seront prédites en termes de produit par période, revenues etc.

Les Objectifs à réaliser seront :

- Prédiction des ventes de café à travers le temps
- Prédiction des ventes des éléments du menu global du restaurant à travers le temps
- Prédiction des ventes des éléments du menu du restaurant et du café par produit à travers le temps.
- Prédiction des revenus provenant du café.
- Prédiction des revenus provenant du restaurant.
- Prédiction du nombre de livraisons par rapport au temps (pour allouer les ressources nécessaires au préalable)
- Prédiction du nombre de produits achetés à emporter par rapport au temps (pour allouer les ressources nécessaires au préalable en termes d'espace, de RH etc.)
- Etude de Corrélation entre produits vendus (qui pourra donner lieu à un système de recommandations en une version avancée du projet)

2.2.2. Par rapport à l'ensemble des restaurants

- Segmentation de la base de clients de Poslik
- Forecasting des revenues de différents restaurants et café selon des critères (zone, prix,...) ou Revenue annuel. [Annulé vu l'absence de données sur le chiffre d'affaires ou Revenu Journalier]
- Étude de marché : Prédiction du taux de réussite d'un restaurant / café selon ses features. (Regression model) [Annulé vu l'absence de données sur le chiffre d'affaires ou Revenu Journalier]

3. Outils utilisés

3.1. Phase d'analyse de données

- Python :
 - Nettoyage et Exploration : Pandas

- Visualisation : Matplotlib, Seaborn
- Colab Notebooks
- Google Sheets

3.2. Phase de réalisation du modèle

- Colab
- Python

3.3. Outils de gestion du projet

- Slack
- Réunions d'avancement
- Mail

4. Méthodologie employée

La méthode employée est CRISP - DM : Cross-Industry Standard Process for Data Mining

C'est une méthodologie élaborée par IBM, l'un des leaders de la technologie.

Ces étapes sont :

- A. Business Understanding
- B. Data Understanding
 - a. Collecte des données
 - b. Description des données
 - c. Exploration des données
 - d. Vérification de la qualité des données
- C. Data Preparation
 - a. Data Selection (Features and subjects)
 - b. Data cleaning
 - c. Feature and data engineering
 - d. Data Integration
 - e. Data Formatting
- D. Modeling
 - a. Model selection
 - b. Model Creation
 - c. Model Evaluation
- E. Evaluation
- F. Deployment

Ci-dessous un schéma explicatif.

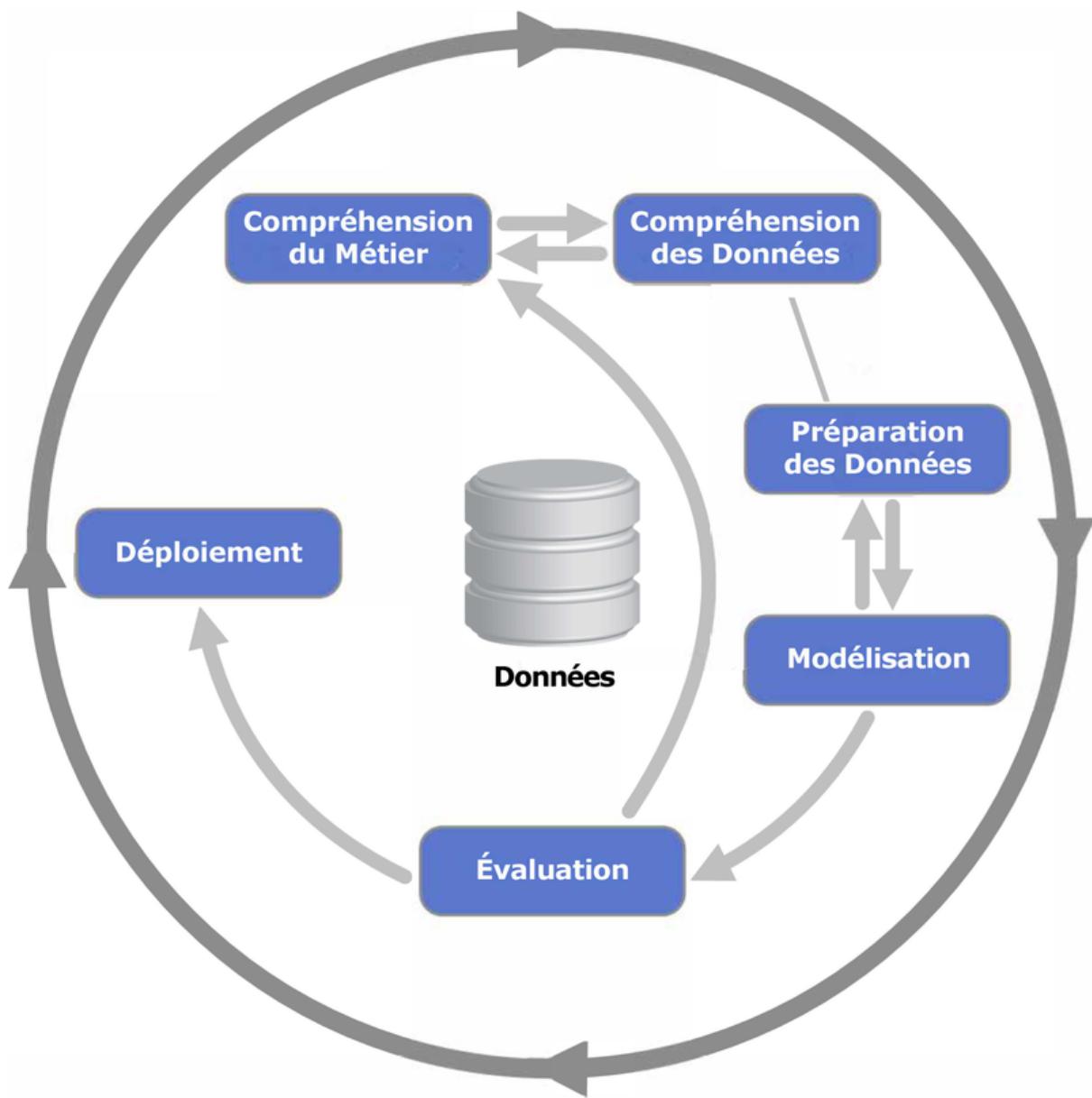


Figure 1 : Méthode de CRISP-DM, Wikipédia

La phase de déploiement n'a pas été réalisée vu la durée courte du stage.

5. Durée du projet

Le projet débute le **01/08/2023** et s'achève le **15/09/2023**.

Partie A : Prédiction des variations temporelles

Chapitre I : Data Understanding

A.I.1. Collecte des données

La base de données a été fournie par l'entreprise d'une partie externe à l'entreprise, des données de tierce partie.

A.I.2. Description des données : Structure des données & MindMap

La base de données délivrée par l'entreprise, provenant d'une Startup Partenaire, est extrêmement Nested, sa structure contient plusieurs Dictionnaires à l'intérieur de listes etc. Ainsi, Nous avons jugé pertinent de tracer l'arborescence de la base de données pour une meilleure visibilité et compréhension de sa structure globale et en détail.

Pour cela, nous avons recouru à un outil en ligne “MindMeister” qui permet de tracer des MindMap.

En effet, un MindMap est un outil utilisée dans plusieurs contextes entre autre, la gestion de projet, il part d'une idée centrale, et explore des aspects de 1er degré provenant de cette idée principale, ensuite, il explore des idées de 2ème degré provenant des idées de 1er degré et ainsi de suite autant que nécessaire.

Ci-joint un exemple simple explicatif :

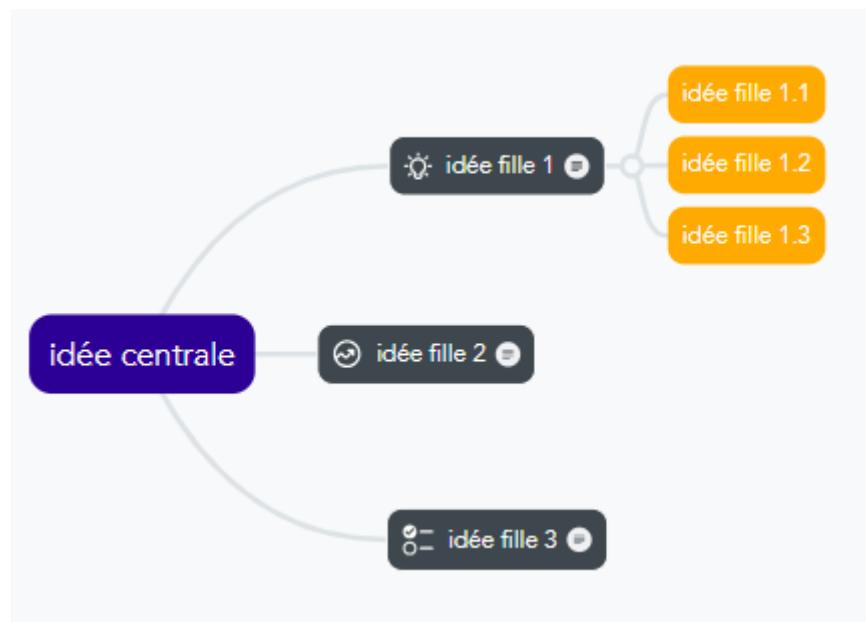


Figure 2 : Exemple de MindMap

Ci-dessous Une figure du MindMap de la structure d'une partie de la base de données sur laquelle nous avons travaillé :

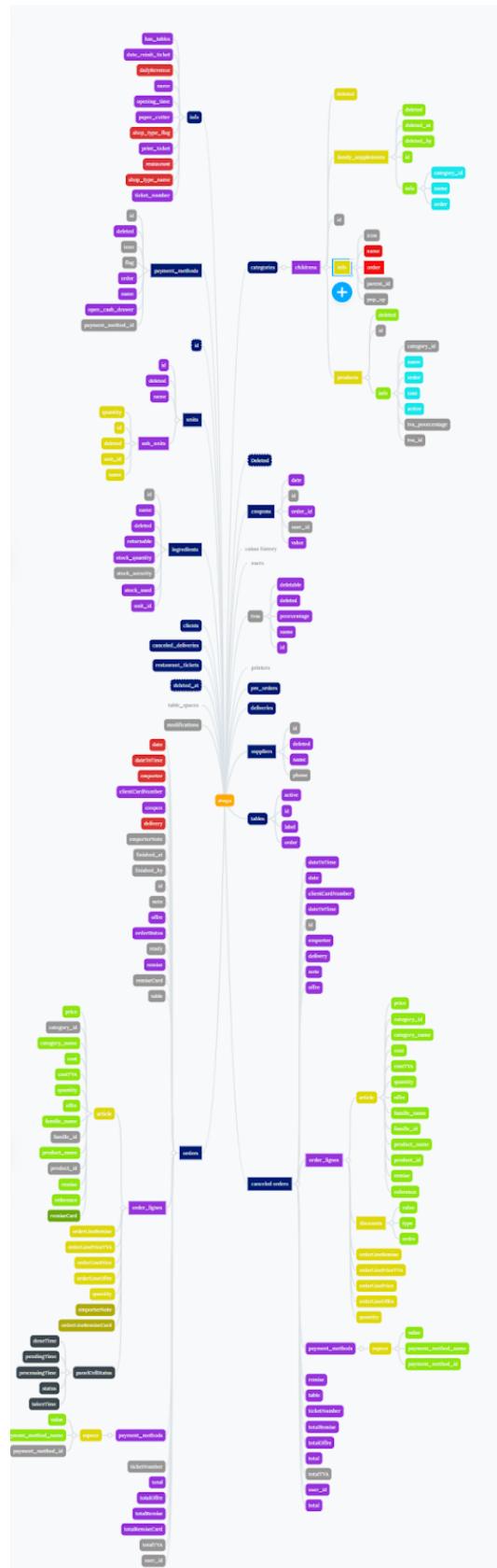


Figure 3 : MindMap des données du projet

Nous expliquons la clé du MindMap dont nous avons défini en raison de clarté et de fournir des schémas significatives à première vue :

Les couleurs utilisées pour indiquer le niveau de profondeur du Mindmap suivent l'ordre suivant (dans un sens croissant) :

Orange > Bleu Marine > Violet > Jaune Motard > Vert > Turquoise

Nomenclature des autres couleurs :

- **Rouge** : Colonne éliminée après visualisation des données car elle a plusieurs valeurs nulles ou car elle est non pertinente.
- **Gris** : Colonne éliminée dès le début avant visualisation extensive de la Base de Données.

Nomenclature des formes utilisées :

- **Nuage** : Noeud final (feuille)
- **Carré aux coins ronds** : Le contenu de cette colonne est un seul exemplaire de l'élément décrit (qui peut avoir des noeuds enfants sous forme de dictionnaire ou liste)
- **Carré aux coins rectangulaires** : Le contenu de cette colonne est présent en plusieurs exemplaires (dictionnaire de clé étant l'identifiant, et valeurs étant la suite de l'arborescence)

Ci-dessous les exemples des formes employées :



Figure 4 : Allure du symbole Nuage



Figure 5 : Carré aux coins ronds

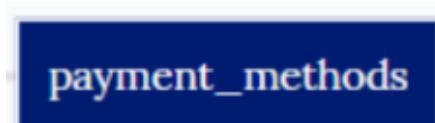


Figure 6 : Carré aux coins rectangulaires

Ci-dessous une coupe en taille lisible du MindMap.

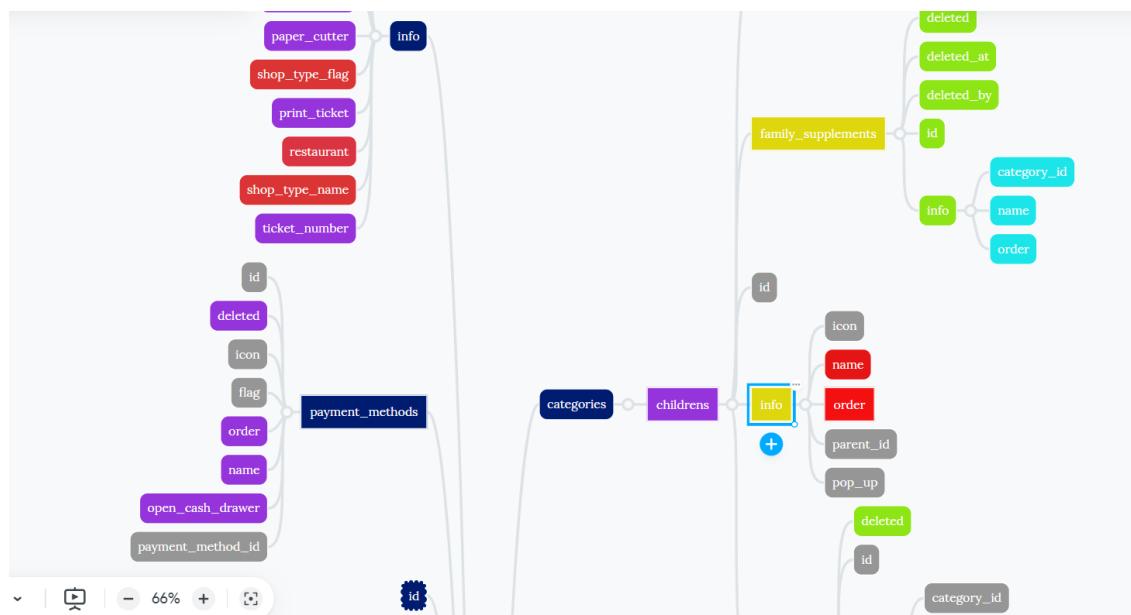


Figure 7 : MindMap agrandi

Cette phase a été réalisée de manière préliminaire pour la Base de Données entière, pour une compréhension globale, ensuite plus particulièrement pour l'attribut “Orders” qui sera traité rigoureusement dans la partie A.

Ci-dessous le mindmap de “Orders”.



Figure 8 : MindMap de la colonne Orders, sujet de la partie A du projet

A.I.3. Exploration des données

Une phase d'exploration des données au moyen de pandas et google sheets a permis une compréhension initiale de la signification des features employés, des types de variables obtenues, des catégories des valeurs des variables, du nombre de chaque catégorie, du nombre de valeurs nulles pour chaque attribut.

A.I.3.1. Data DeNesting

Nous avons transformé une Base de Données ayant une structure de plusieurs commandes passées initialement contenues dans un dictionnaire dont les clés sont les id des commandes, et les valeurs sont un autre sous dictionnaire contenant la date, les méthodes de paiement, (lui-même contenant un sous dictionnaire à élaborer...), etc.

Nous avons transformé une Base de Données initialement importée sous forme de fichier json, ayant cet état :

```
{'-N1s7krGSiw3iirC7SO1':  
  {'clientCardNumber': ...}  
  
{'-  
MwzWeQmvpJKjT4VuFb_':  
  {'clientCardNumber': ...}  
  
{'-  
N1UgjHZLmAy8iYuQ_Q5':  
  {'clientCardNumber': ...}  
  
{'-MySi4uCuuSdtQiwBZ1n':  
  {'clientCardNumber': ...}  
  
{'-  
N2SEdRz55WtMqaCWAIO':  
  {'clientCardNumber': ...}
```

Figure 9 : Etat des commandes lors de l'importation

En une Base de Données lisible dont les features sont explicités :

	index	coupon	date	dateToTime	delivery	emporter	delivery_infos.heure_recup	remise	ticketNumber	total	...	order_lignes.article.ingredient_useds	order_lignes
0	0	0.0	29/03/2023	08:51	1680079866815	0	0	None	0	1	5.4	...	NaN
1	0	0.0	29/03/2023	08:51	1680079866815	0	0	None	0	1	5.4	...	NaN
2	0	0.0	29/03/2023	08:51	1680079866815	0	0	None	0	1	5.4	...	NaN
3	0	0.0	29/03/2023	08:51	1680079866815	0	0	None	0	1	5.4	...	NaN
4	0	0.0	29/03/2023	08:51	1680079866815	0	0	None	0	1	5.4	...	NaN
...
1020	535	0.0	08/08/2023	12:58	1691495924308	0	0	None	0	3	12.0	...	NaN
1021	535	0.0	08/08/2023	12:58	1691495924308	0	0	None	0	3	12.0	...	NaN
1022	536	0.0	08/08/2023	13:00	1691496033435	0	0	None	0	4	25.0	...	NaN
1023	537	0.0	08/08/2023	13:04	1691496270995	0	0	None	0	5	11.0	...	NaN
1024	537	0.0	08/08/2023	13:04	1691496270995	0	0	None	0	5	11.0	...	NaN

Figure 10 : Base de Données après traitement

A.I.3.2. Collecte d'informations à propos la Data en utilisant Pandas et sheets

```
f.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 53 columns):
 #   Column           Non-Null Count  Dtype  
 --- 
 0   index            1025 non-null   int64  
 1   coupon            1025 non-null   float64 
 2   date              1025 non-null   object  
 3   dateToTime        1025 non-null   int64  
 4   delivery          1025 non-null   int64  
 5   emporter          1025 non-null   int64  
 6   delivery_infos.heure_recup    1025 non-null   object  
 7   remise             1025 non-null   int64  
 8   ticketNumber      1025 non-null   int64  
 9   total              1025 non-null   float64 
 10  totalOffre         1025 non-null   float64 
 11  totalRemise        1025 non-null   float64 
 12  data.ingredient_useds 817 non-null   object  
 13  payment_methods_value 876 non-null   float64 
 14  unpaid_values      1025 non-null   float64 
 15  espece             1025 non-null   int64  
 16  cheque              1025 non-null   int64  
 17  carte_bancaire     1025 non-null   int64  
 18  livraison           1025 non-null   int64  
 19  emporter.1          1025 non-null   int64  
 20  offer               1025 non-null   int64  
 21  placement           1025 non-null   object  
 22  order_lignes.article.category_id 1025 non-null   object  
 23  order_lignes.article.category_name 1025 non-null   object  
 24  order_lignes.article.cost          1025 non-null   float64 
 25  order_lignes.article.costTVA       1025 non-null   int64 
```

Figure 11 : Types des variables et nombre

```
f.columns
```

```
Index(['index', 'coupon', 'date', 'dateToTime', 'delivery', 'emporter',
       'delivery_infos.heure_recup', 'remise', 'ticketNumber', 'total',
       'totalOffre', 'totalRemise', '...', 'order_lignes.orderLineOffre'],
      dtype='object')
```

Figure 12 : Noms des colonnes

	index	coupon	dateToTime	delivery	emporter	remise	ticketNumber	total	totalOffre	totalRemise	...	order_lignes.orderLineOffre
count	1025.000000	1025.000000	1.025000e+03	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	...	1025.000000
mean	257.169756	0.017171	1.687555e+12	0.013659	0.013659	0.887805	10.400000	38.763253	0.316195	0.767346	...	0.123512
std	162.176320	0.231499	3.410636e+09	0.116126	0.116126	5.641650	9.449702	88.767121	2.040539	3.945956	...	1.160490
min	0.000000	0.000000	1.680080e+12	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000
25%	112.000000	0.000000	1.685972e+12	0.000000	0.000000	0.000000	3.000000	10.000000	0.000000	0.000000	...	0.000000
50%	243.000000	0.000000	1.688470e+12	0.000000	0.000000	0.000000	7.000000	22.500000	0.000000	0.000000	...	0.000000
75%	404.000000	0.000000	1.690463e+12	0.000000	0.000000	0.000000	15.000000	44.720000	0.000000	0.000000	...	0.000000
max	537.000000	4.300000	1.691496e+12	1.000000	1.000000	50.000000	50.000000	1800.000000	27.000000	27.540000	...	27.000000

Figure 13 : Description générale des données

```
index                      0
coupon                     0
date                       0
dateToTime                 0
delivery                   0
emporter                   0
delivery_infos.heure_recup 1007
remise                      0
ticketNumber                0
total                      0
totalOffre                  0
totalRemise                 0
data.ingredient_useds      208
payment_methods_value      149
unpaid_values                0
espece                      0
cheque                      0
carte_bancaire                0
livraison                   0
emporter.1                  0
offer                        0
placement                   0
order_lignes.article.category_id 0
order_lignes.article.category_name 0
order_lignes.article.cost      0
order_lignes.article.costTVA     0
```

Figure 14 : Nombre de valeurs nulles par colonne “ingredient_used”

```
[{'ingredient_id': '-NMcv4baLqfktnvkvsw',
 'ingredient_name': 'Thon',
 'prix_moyen': 0.079,
 'quantity': 0.05,
 'returnable': False},
 {'ingredient_id': '-NMcv4p6j4eadnjf0ydk',
 'ingredient_name': 'Escalope',
 'prix_moyen': 0,
 'quantity': 0.75,
 'returnable': True},
 {'ingredient_id': '-NMcv4xzduw3fgkyc_0f',
 'ingredient_name': 'Fromage',
 'prix_moyen': 0.9545,
 'quantity': 0.1,
 'returnable': True}]
```

Figure 15 : Visualisation de la structure interne de la colonne

A.I.4. Vérification de la qualité des données

Les données contiennent plusieurs colonnes à valeurs presque entièrement nulles.

Certaines valeurs ne sont pas très logiques.

On trouve quelques Outliers

L'une des données pertinentes : Revenus Journaliers n'est pas présente.

Les points les plus importants à signaler concernant les données demeurent :

- Les données portent sur la période entre le 29/03/2023 et le 08/08/2023, qui est une durée très courte, sachant qu'elle inclut le mois de Ramadan, qui est un mois durant lequel la consommation de café et de nourriture devient très irrégulière et axée sur une période non conventionnelle de la journée, ce qui fait que, inclure cette partie dans la phase de Training du Modèle va fausser les prédictions; mais l'éliminer fera que nous aurons des données d'approximativement 3 mois et demi.
- Les données sont datées, mais il y a trop d'irrégularité dans les dates, parfois nous trouvons des données relatives à uniquement deux jours durant dix jours, d'autres fois on trouve des données séparées de plus de 12 jours pour un café ou un restaurant (Donnée manquante puisqu'il est impossible de n'avoir aucune commande pour des jours de suite pour un restaurant ou un café, et de manière récurrente).
- Des fois, nous trouvons une journée à une seule commande.

Les irrégularités des dates couplées avec la taille restreinte de la Base de Données, rendent les prédictions impossibles dans ce cas, seule la phase de training est applicable, vu que la fonction .predict() n'est pas applicable à un modèle de Time Series à des intervalles de temps irrégulières.

Une solution a été proposée par l'étudiant pour essayer de recourir à ce problème.

Pour la suite, nous expliquerons la démarche et le training avec la Base de Données présentant des irrégularités, sans prédictions vu l'impossibilité de cela ; ensuite, les solutions alternatives proposées seront exposées ainsi que leurs limites.

Chapitre II : Data Preparation

Cette Phase, couplée avec la phase de Data Exploration, occupe 80% de la durée du projet.

A.II.1. Data Selection

Les colonnes à features non pertinentes ont été supprimées.

Les colonnes à valeurs majoritairement nulles ont été supprimées.

Les lignes à valeurs surpassant le 3ème et 1er quartile ont été supprimées.

Certains de ces traitements ont été réalisés en utilisant la librairie Pandas de Python, d'autres ont été réalisés sur google Sheets.

A.II.2. Data Cleaning

Les données à noms mal écrits (avec des symboles tel “%”) ont été réécrites correctement.

Certains Outliers ont vu leur valeurs extrémales remplacées par la moyenne des valeurs pour le feature concerné.

A.II.3. Feature Engineering

Cette phase avait été adoptée dans le cadre des solutions alternatives à l'utilisation de la Base de Données irrégulière. Elle sera donc explicitée le moment opportun.

A.II.4. Time Series Visualization

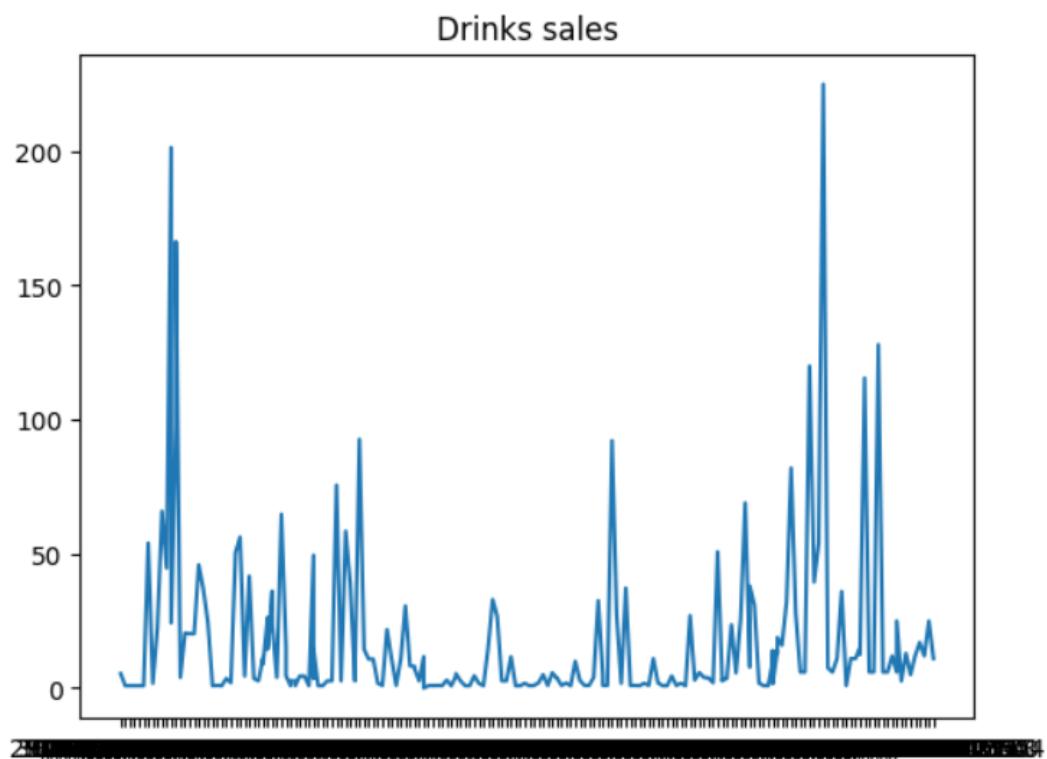


Figure 16 : Variation des ventes des produits du café

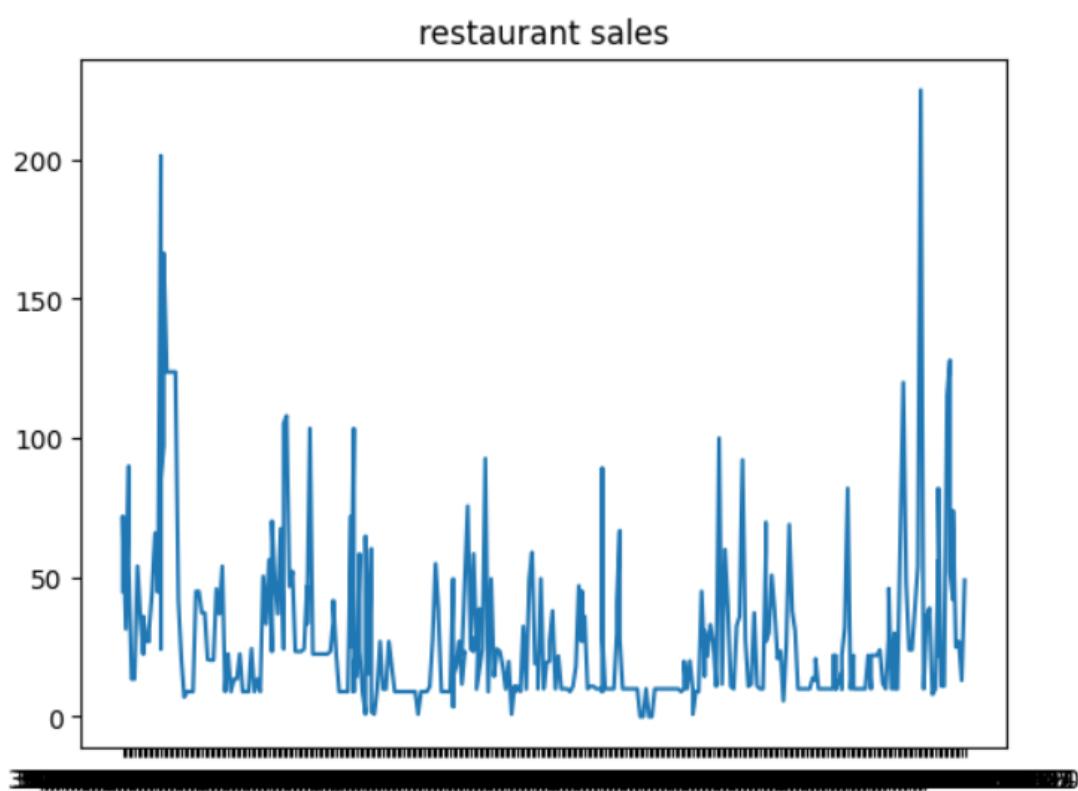


Figure 17 : Variation des ventes des produits du restaurant

Chapitre III : Data Modeling

A.III.1. Choix préliminaire du modèle et raisons

Il est possible de recourir à des modèles classiques du Time Series Analysis and Predictions, notamment the SARIMAX package, contenant plusieurs modèles ARIMA, SARIMA, etc. Ces modèles ont des paramètres, trois pour ARIMA, sept pour SARIMA (Seasonal ARIMA), le tuning se fait suite à plusieurs méthodes d'analyse, parmi lesquelles les diagrammes ACF¹ et PACF².

Il est aussi possible de recourir à des modèles de Deep Learning, notamment les LSTM (Long Short Term Memory). Les recherches ont montré que ces modèles donnent des résultats en général meilleurs que les modèles classiques, sauf dans le cas de Seasonality, dans ce dernier, ce sont les modèles de la famille SARIMAX qui ont le dessus.

Nous avons choisi d'utiliser les modèles classiques, en particulier the ARIMA Model, vu que nous voulons principalement appliquer les méthodes classiques et les maîtriser comme objectif de ce stage d'une part, d'autre part, la décomposition de la série montre qu'il y a présence de Seasonality, ce qui supporte l'utilisation de ce modèle.

De plus, un modèle de Deep Learning non LSTM avait été implémenté à titre d'essai.

A.III.2. Model Training

Nous subdivisons la Base de Données en Training_Data et Testing_Data.

Le projet traite 4 parties principales en relation avec l'étude de cas d'un restaurant-café spécifique, lesquels :

- Prédiction des Quantités vendues de produits en total en fonction du temps :
 - Point de vue restaurant
 - Point de vue café
- Prédiction des ventes de produits en fonction du temps :
 - Point de vue restaurant
 - Point de vue café
- Prédiction des Produits à livrer et à emporter en fonction du temps :
 - Point de vue restaurant
- Prédiction des Quantités vendues par produit et recherche de corrélation en fonction du temps.

Chacun des quatre points principaux a été traité dans un Colab Notebook à part pour la clarté et pour le loading de différentes Base de Données adaptées au besoin pour chaque Notebook.

¹ ACF : AutoCorrelation Function : Indique le degré de dépendance de la valeur d'un Lag donné avec tous ses précédents.

² PACF : Partial AutoCorrelation Function : Indique le degré de dépendance de la valeur d'un Lag donné avec son précédent direct uniquement.

A.III.3. Decomposition of The Time series

Nous avons recouru à une décomposition additive.

Ci-dessous le résultat de la décomposition.

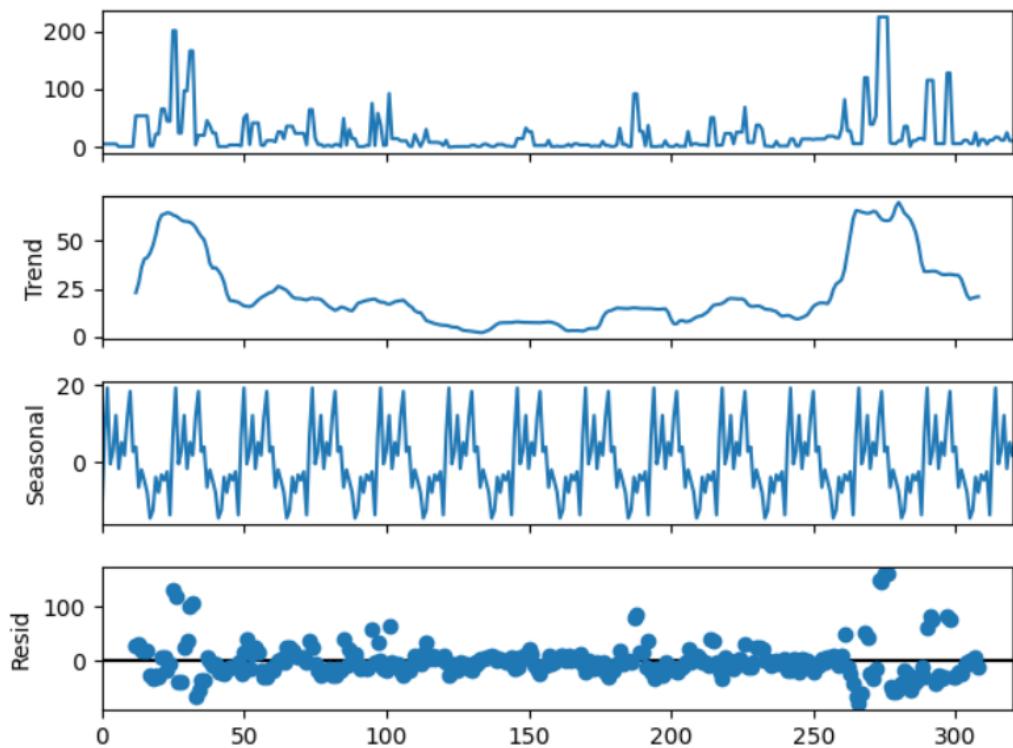


Figure 18 : Décomposition additive des Time Series de la BD du café

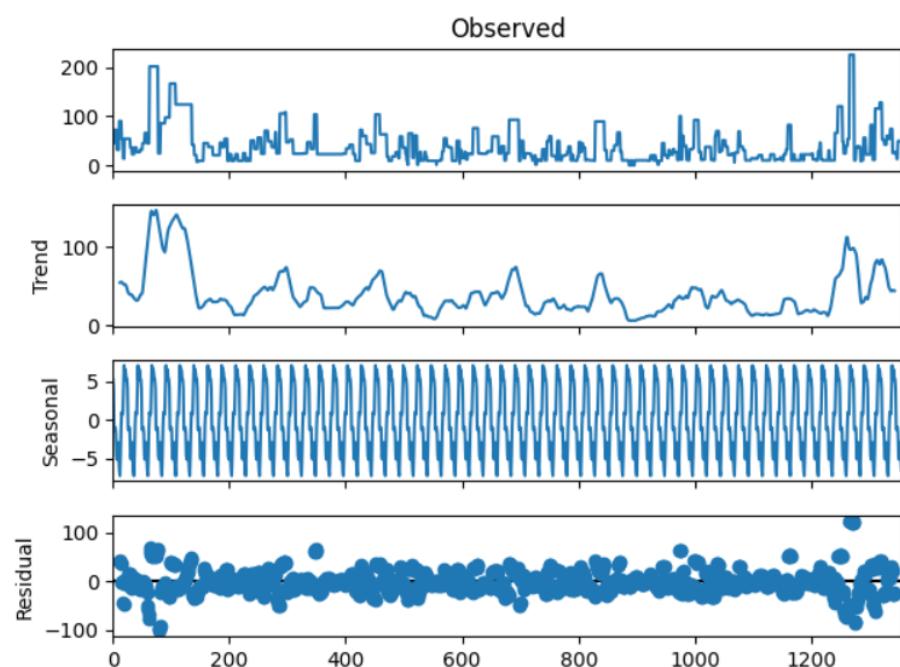


Figure 19 : Décomposition additive des Time Series de la BD du café

A.III.4. Differencing of the Time Series

4.1. Differencing

Nous procémons au Differencing pour éliminer le

- Trend
- Seasonality

et aboutir à des séries stationnaires^{*1}, ainsi que pour pouvoir utiliser les AdFuller tests qui supposent que les séries sont stationnaires.

4.1.1. Produits du café

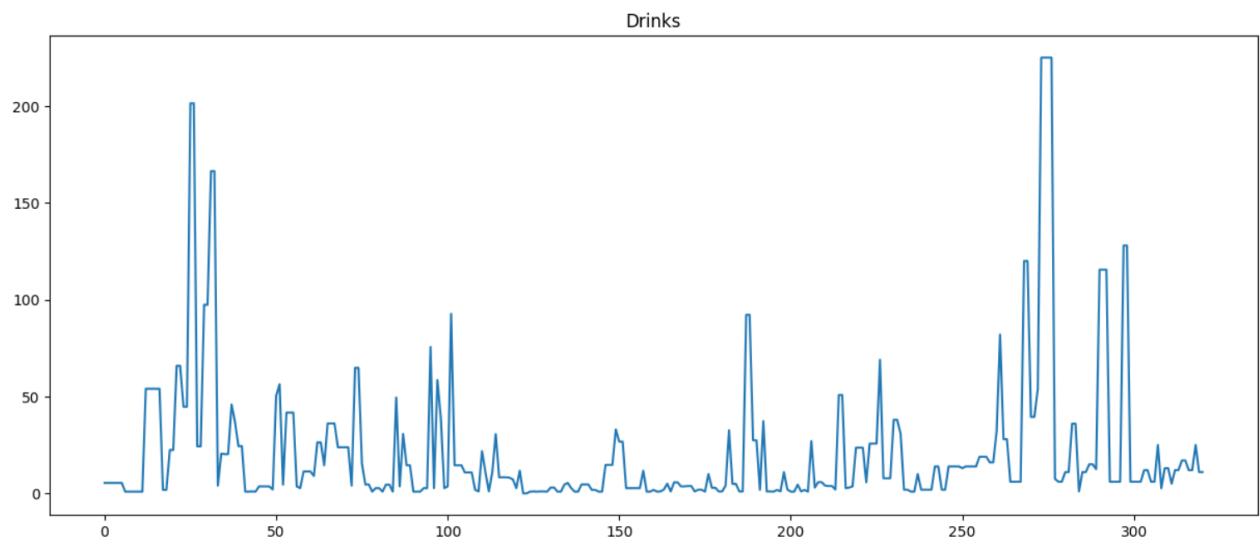


Figure 20 : Allure des ventes du café avant Differencing

3

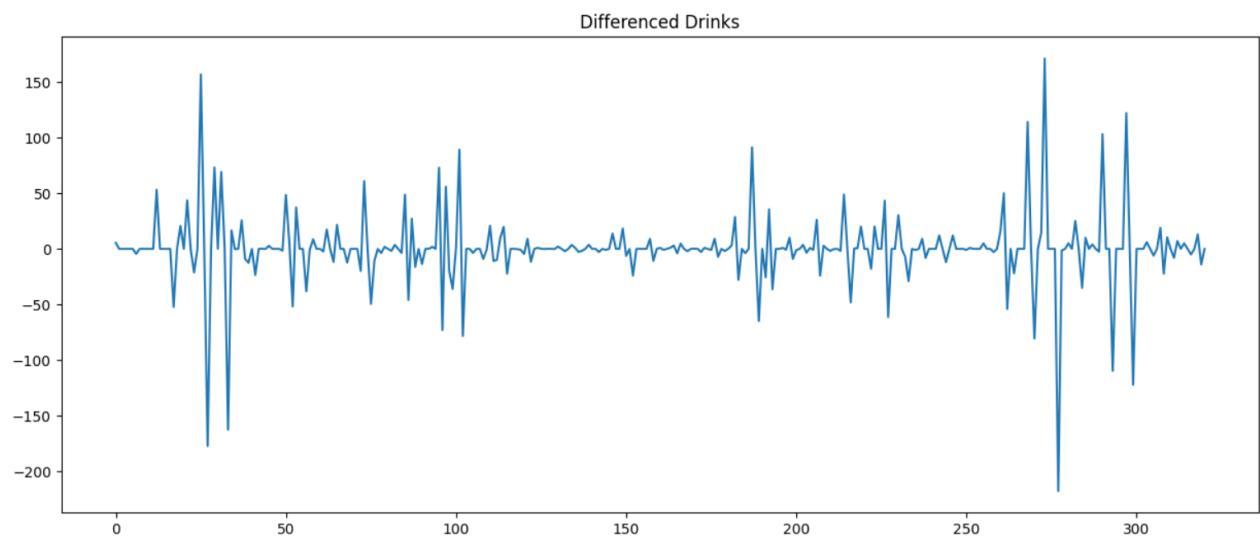


Figure 21 : allure des ventes du café après Differencing

³ stationnaires : Ayant une moyenne et un écart-type nuls (Weak Stationarity)

4.1.2. Produits du restaurant

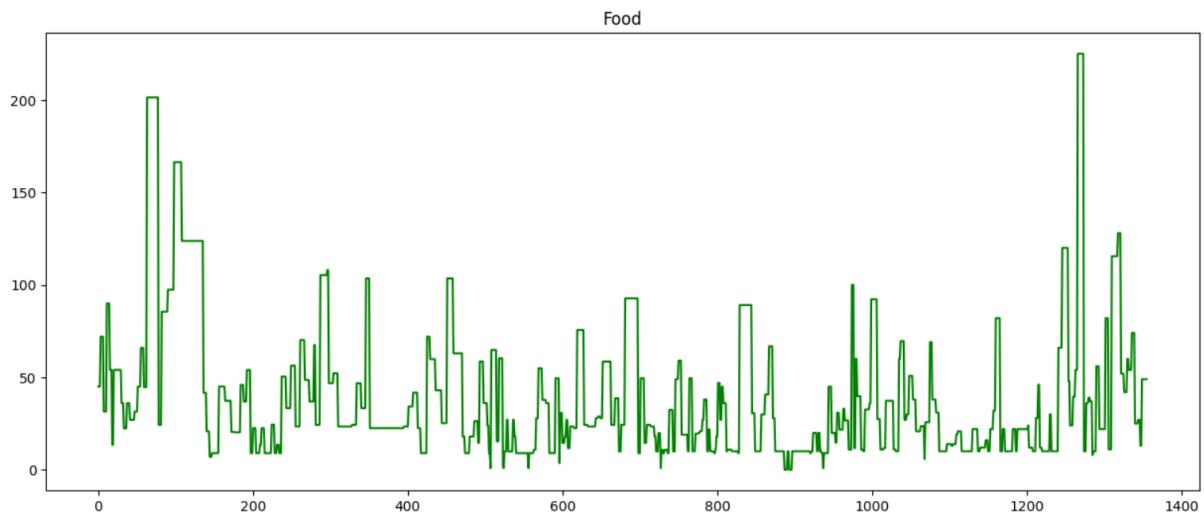


Figure 22 : Allure des ventes du restaurant avant Differencing

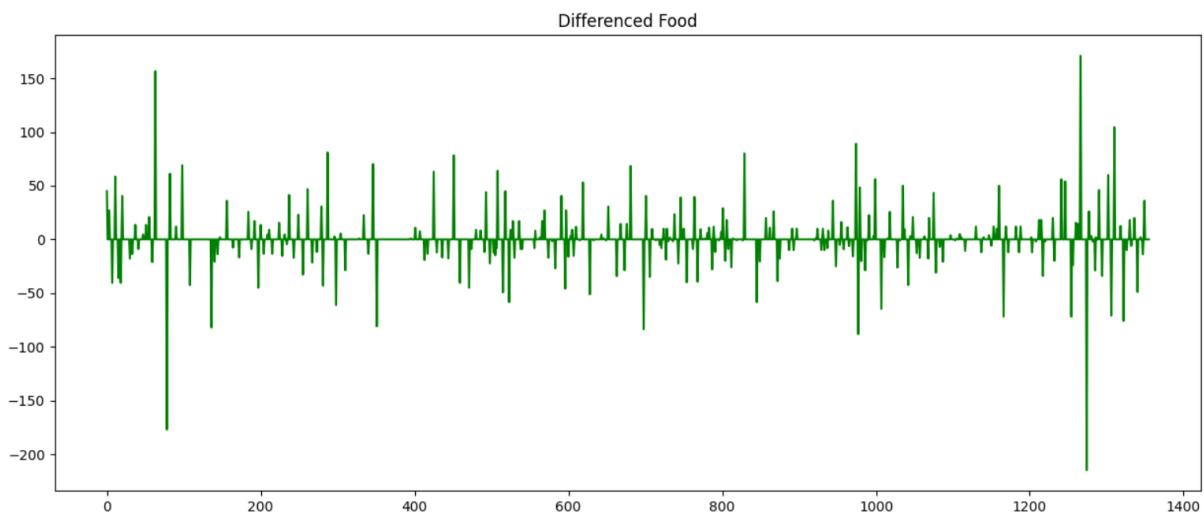


Figure 23 : allure des ventes du restaurant après Differencing

4.2. Test ADFuller

ADF Statistic: -7.712705
p-value: 0.000000

ADF Statistic: -11.786136
p-value: 0.000000

Figure 24 : Résultat du test adfuller - café

Figure 25 : Résultat du test adfuller - restaurant

Les résultats des deux tests assurent l'absence de Seasonality pour les deux cas.

A.III.5. Évaluation des données pour le choix des paramètres du modèle

Nous avons recouru aux diagrammes ACF et PACF des données pour déterminer les paramètres du modèle ARIMA.

Ci-dessous les graphes relatifs aux données des ventes du restaurant et du café.

5.1. Cas du café

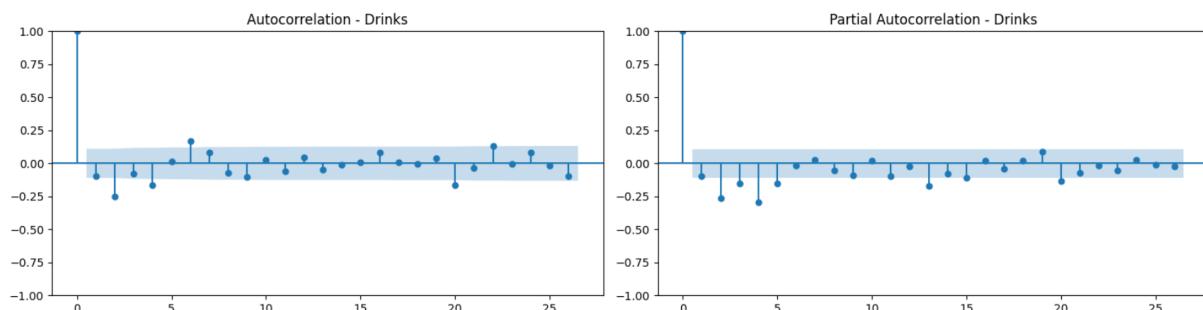


Figure 26 : Allure des courbes des graphes ACF et PACF pour la cafétéria

Ainsi, les paramètres choisis pour le modèle ARIMA sont :

AR : $p = 5$

MA : $q= 6$

I : 0 (vu qu'on avait déjà effectué le Differencing et le test ADF donne un résultat satisfaisant)

5.2. Cas du restaurant

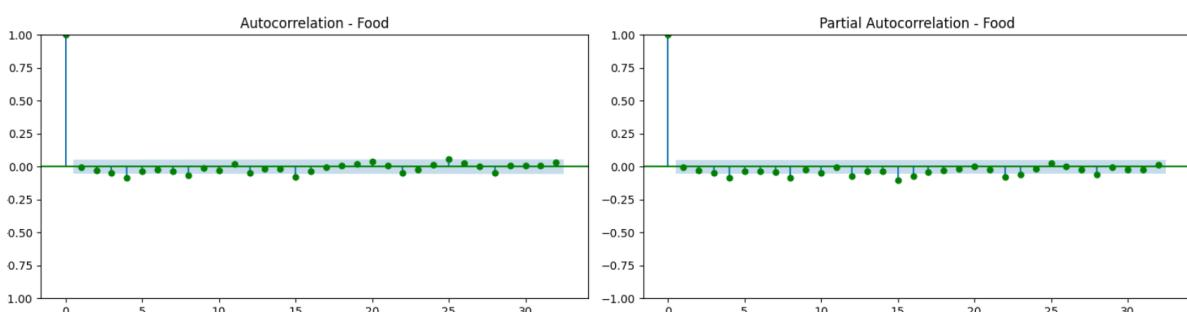


Figure 27 : Allure des courbes des graphes ACF et PACF pour le restaurant

Ainsi, les paramètres choisis pour le modèle ARIMA sont :

- AR : $p=8$: Le dernier Lag après lequel les valeurs sont presque nulles dans PACF (périodiquement)
- MA : $q=8$: Le dernier Lag après lequel les valeurs sont presque nulles dans ACF (périodiquement)

- I : 0 (vu qu'on avait déjà effectué le Differencing et le test ADF donne un résultat satisfaisant)

A.III.6. Model Training : Résultat du Fitting et résidus

6.1. Résumé du Training

SARIMAX Results						
Dep. Variable:	0	No. Observations:	321			
Model:	ARIMA(6, 0, 8)	Log Likelihood:	-1533.095			
Date:	Wed, 20 Sep 2023	AIC:	3092.191			
Time:	12:01:42	BIC:	3141.219			
Sample:	0	HQIC:	3111.767			
Covariance Type:	opg					
coef	std err	z	P> z	[0.025	0.975]	
const	0.0097	0.183	0.095	0.925	-0.191	0.211
ar.L1	0.0624	0.240	0.259	0.795	-0.499	0.533
ar.L2	0.2372	0.261	1.178	0.239	-0.158	0.322
ar.L3	-0.1396	0.268	1.240	0.149	-0.025	0.196
ar.L4	0.1685	0.095	0.866	0.386	-0.213	0.550
ar.L5	0.0366	0.124	0.309	0.842	0.226	0.389
ar.L6	0.2580	0.052	4.830	0.000	0.149	0.351
ma.L1	-0.3648	0.245	-1.489	0.136	-0.845	0.115
ma.L2	-0.5929	0.245	-2.422	0.015	-1.073	-0.113
ma.L3	0.3299	0.296	1.116	0.265	-0.250	0.999
ma.L4	-0.3276	0.223	-1.468	0.142	-0.765	0.110
ma.L5	-0.0321	0.222	-0.144	0.885	-0.468	0.404
sigma2	828.0865	38.912	21.281	0.000	751.820	904.353
Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	1672.70			
Prob(Q):	0.99	Prob(JB):	0.00			
Heteroskedasticity (H):	1.27	Skew:	1.56			
Prob(H) (two-sided):	0.21	Kurtosis:	13.74			

SARIMAX Results						
Dep. Variable:	0	No. Observations:	321			
Model:	ARIMA(6, 0, 8)	Log Likelihood:	-5720.859			
Date:	Wed, 20 Sep 2023	AIC:	11477.717			
Time:	12:01:56	BIC:	11571.539			
Sample:	0	HQIC:	11512.846			
Covariance Type:	opg					
coef	std err	z	P> z	[0.025	0.975]	
const	-0.0171	0.020	-0.841	0.400	-0.057	0.023
ar.L1	-0.9958	0.275	-3.622	0.000	-1.535	-0.457
ar.L2	-0.6180	0.128	-5.158	0.000	-0.853	-0.383
ar.L3	-0.8353	0.121	-6.981	0.000	-1.073	-0.598
ar.L4	-0.3069	0.138	-2.222	0.022	-0.567	0.154
ar.L5	0.3126	0.134	2.327	0.020	0.049	0.576
ar.L6	0.2777	0.081	3.445	0.001	0.120	0.436
ar.L7	0.7865	0.109	7.231	0.000	0.573	1.000
ar.L8	0.6400	0.106	3.772	0.001	0.277	1.025
ma.L1	-0.212	0.247	-0.854	0.430	-0.405	0.005
ma.L2	0.4729	0.148	3.295	0.001	0.184	0.762
ma.L3	0.6376	0.201	3.176	0.001	0.244	1.031
ma.L4	-0.1008	0.255	-0.396	0.692	-0.600	0.399
ma.L5	-0.5230	0.276	-2.279	0.023	-1.100	-0.688
ma.L6	-0.5313	0.246	-2.146	0.025	-0.925	0.138
ma.L7	-0.9974	0.182	-5.493	0.000	-1.353	-0.642
ma.L8	-0.7724	0.213	-3.634	0.000	-1.189	-0.356
sigma2	272.3222	19.488	13.974	0.000	234.126	310.519
Ljung-Box (L1) (Q):	0.35	Jarque-Bera (JB):	86731.13			
Prob(Q):	0.56	Prob(JB):	0.00			
Heteroskedasticity (H):	1.20	Skew:	-0.03			
Prob(H) (two-sided):	0.05	Kurtosis:	42.18			

Figure 28 : Résultat des résumés du Training des modèles de café et restaurant respectivement

6.2. Graphes des résidus et densité des résidus

6.2.1. Cas du café

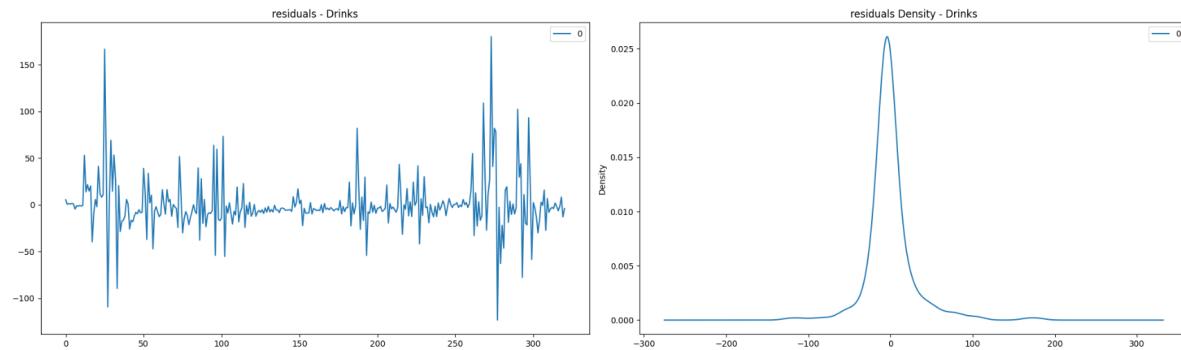


Figure 29 : Allure des graphes des résidus et densité des résidus

6.2.2. Cas du restaurant

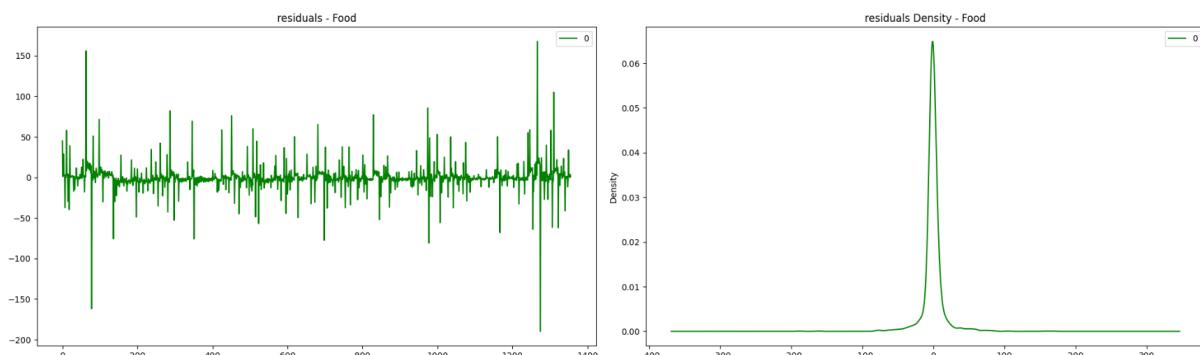


Figure 30 : Allure des graphes des résidus et densité des résidus

6.2.3. Interprétation des graphes des résidus

mean -0.076729

mean -0.202200

Figure 31 : Moyennes de la densité des résidus respectivement pour le café, et le restaurant

Les valeurs moyennes des graphes des résidus sont centrées autour de zéro avec un faible écart-type, en particulier pour les données du restaurant.

L'obtention d'un résultat meilleur pour les données du restaurant peut être dû au nombre de données des commandes du restaurant grandement supérieur à celui des données du café.

Nous essayerons d'améliorer le résultat du fitting pour la cafétéria, pour cela, nous choisirons des valeurs correspondantes pour les paramètres q et p, plus grandes que celles choisies précédemment.

Les graphes ACF et PACF donnent les meilleures choix de résultats suivant :

- p= 20
- q= 13

Le graphe des densités des résidus donne ci dessous :

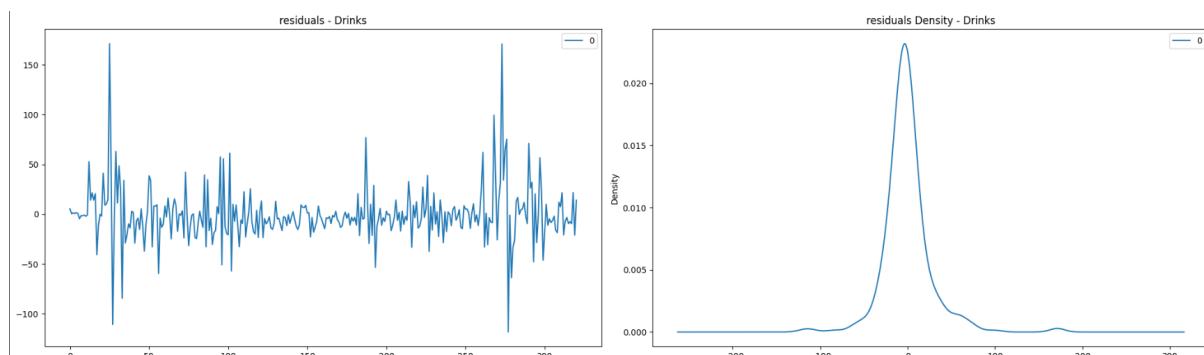


Figure 32 : Allure des graphes des résidus et densité des résidus du café

mean -0.362033
std 27.596148

Figure 33 : Moyennes de la densité des résidus du café

La nouvelle valeur de la moyenne n'a pas beaucoup changé, l'écart-type (Std) non plus, il a varié de 28. à 27.6

L'amélioration de l'erreur est négligeable malgré une multiplication des paramètres d'un facteur de 10, ce qui met notre modèle dans le risque de l'Overfitting.

La validation du choix des paramètres pour le modèle des ventes de la cafétéria dépendra du résultat des prédictions :

- En présence d'Overfitting, les paramètres initiaux correspondent déjà bien au modèle.
- En absence d'Overfitting, on pourra garder les paramètres post Tuning malgré les pertes temporelles du Training et comparer les résultats (Accuracy et perte temporelle) à un modèle LSTM éventuellement. (*Cette étude est hors du contexte de ce projet mais c'est une perspective à considérer*)

6.3. Fitted Model

6.3.1. Fitted Model du cafétéria (paramètres initiaux)

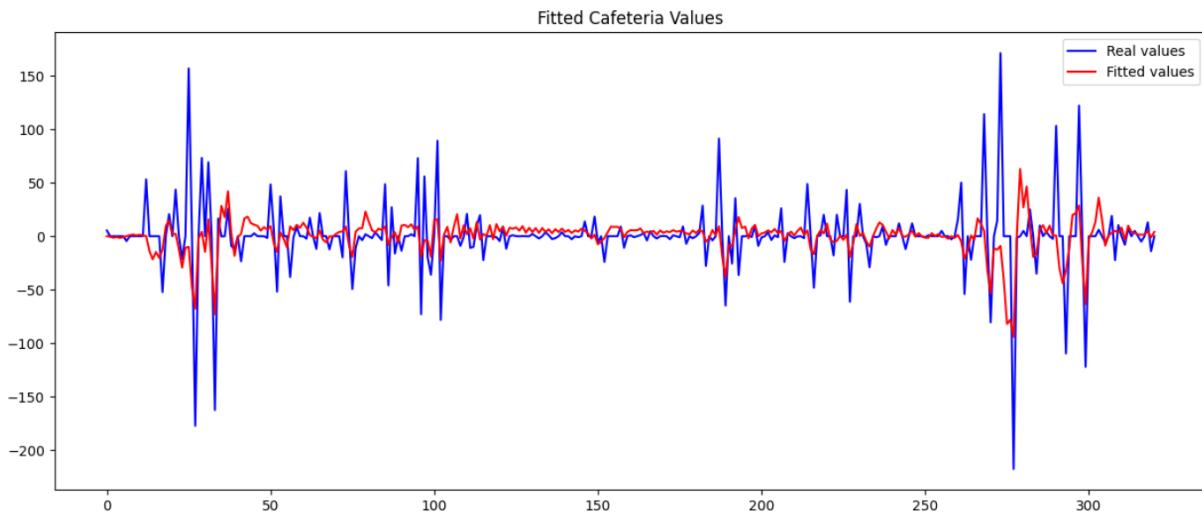


Figure 34 : Allure du Fitted model par rapport aux valeurs réelles

6.3.2. Fitted Model du restaurant



Figure 35 : Allure du Fitted model par rapport aux valeurs réelles

6.3.3. Fitted Model du cafétéria (Tuned Parameters)

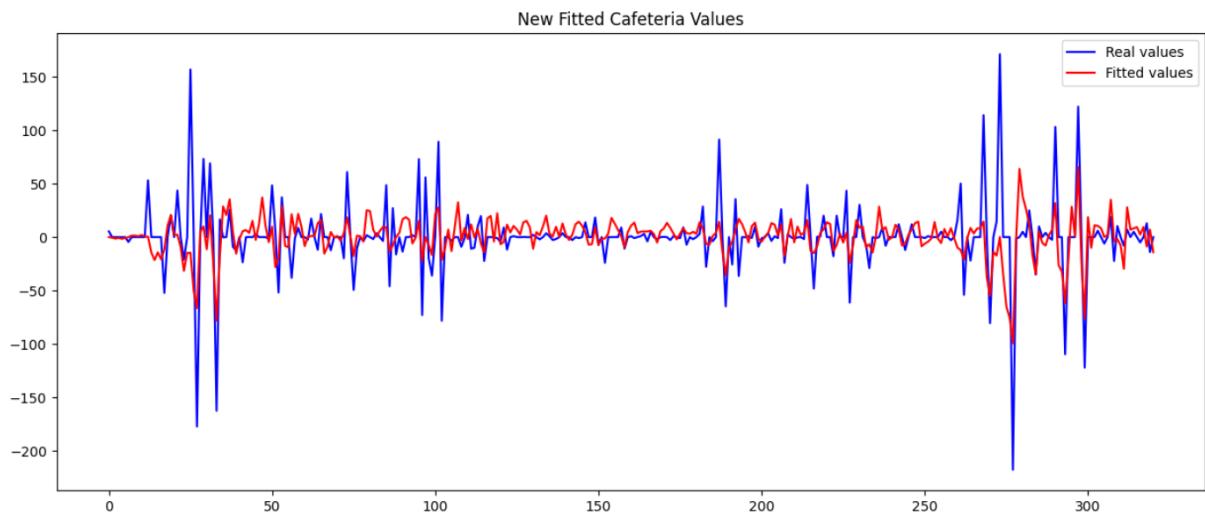


Figure 36 : Allure du Fitted model par rapport aux valeurs réelles avec des Tuned parameters

Chapitre IV : Évaluation et Interprétation

Vu notre incapacité à fournir des prédictions avec des données irrégulières, nous évaluerons le résultat du fitting et le graphe des résidus.

A.IV.1. Evaluation par rapport au Fitting

A.IV.1.1. Evaluation relative aux ventes de produits du café

Nous avons déjà évalué le modèle de point de vue fitting, et nous avons dégagé des paramètres post-Tuning.

Nous analyserons ainsi, le résultat des deux Fitted models.

On a remarqué que la multiplication des paramètres d'un facteur de 10 n'a pas eu une valeur ajoutée remarquable par rapport à l'erreur et la densité des résidus.

La visualisation du fitted Model pour les paramètres initiaux, montre une concordance des points prédits par rapport aux valeurs réelles, celles-ci sont légèrement au-dessous des valeurs de la prédiction, mais l'erreur reste minimale (par rapport à la quantité et qualité des données fournies).

La visualisation du Fitted Model avec les paramètres post-Tuning, montre par contre une grande variabilité par rapport aux valeurs réelles. Bien que la nouvelle densité d'erreur soit inférieure à l'ancienne valeur, les prédictions locales des valeurs des prix présentent un écart bien plus remarquable que celui de l'écart de l'ancien modèle.

On peut conclure que le premier cas présente de l'Underfitting, alors que le deuxième, présente non seulement de l'Underfitting, mais aussi de l'Overfitting, ce qui s'explique d'une part par le manque de données, d'autre part, par l'augmentation de la complexité de la fonction polynomiale à des facteurs de puissance 20 ce qui résulte inévitablement en un modèle adapté seulement au Training Data.

Des prédictions donneront probablement un résultat qui soutient ce propos.

A.IV.1.2. Evaluation relative aux ventes de produits du restaurant

Nous remarquons que comme déjà mentionné dans la partie *6.2 Graphes des résidus et densité des résidus*, la moyenne est centrée autour de 0, et l'écart-type est assez faible : il présente 16.5 or que la valeur maximale est de 167.8, ce qui donne 9.8% d'erreur sur le résultat du fitting.

Ce résultat semble satisfaisant par rapport à la quantité des données introduites, son irrégularité, et la présence de comportements non conventionnels des variations, lié au mois de Ramadan.

count	1356.000000
mean	-0.202200
std	16.460413
min	-189.831857
25%	-3.139878
50%	-0.965669
75%	2.311482
max	167.768327

Figure 37 : Résumé du graphe des résidus du modèle du restaurant

A.IV.2. Evaluation par rapport à des prédictions préliminaires

Une étude préliminaire des prix d'un produit du café et du restaurant, indique que celui-là n'est pas particulièrement l'un des prestataires de haute gamme, nous nous attendons ainsi à une consommation qui suit cette variation :

- Pour le café :
 - Hausse : Le matin entre 6:30 et 9:30 (horaires de travail des employés, et étudiants qui auront tendance à se procurer leur petit-déjeuner quotidien chez un tel prestataire) le taux de hausse peut varier vu que les données présentes sont durant les mois d'été, donc on ne tiendra compte que des employés.
 - Hausse les après-midi après les horaires de travail (entre 19:00 et 22:00)
- Pour les restaurants :
 - Hausse : Durant l'horaire du déjeuner (12:00 - 14:30) mais probablement à un nombre qui représente 70% du nombre des personnes qui viennent à la cafétéria le matin.
 - Hausse : Durant les horaires de dîner (19:00 - 21:30)
 - Des pics moyens et peu réguliers entre 15:00 et 18:00 pour des groupes d'étudiants qui ont sauté des repas.

Nous évaluerons si nos prévisions de variations sont conformes aux valeurs réelles en première instance.

Une vérification de la Base de données couplée avec les plots, montre que les majorité des pics ont lieu soit aux alentours de 10H ou aux alentours de 18H, ce qui concorde avec les attentes partiellement.

La non conformité totale peut revenir à:

- Absence des données de consommation aux alentours de 8H (Donnée manquantes)
- Tendance vers la consommation des produits l'après midi vu que l'activité hormonale des individus fait qu'ils n'ont pas besoin de consommation élevée en nourriture après le mois de Ramadan.

Chapitre V : Solution proposée

A.V.1. Principe de la solution

Pour recourir au problème des données irrégulières, nous tentons de les discréteriser.

Nous avons considéré l'idée d'ajouter des lignes manquantes, mais il y a tellement d'écart et tellement de valeurs à ajouter, que la data perdra son intégrité et qu'ainsi, les résultats livrées ne seront pas fiables.

Ainsi, nous avons décidé de la rendre régulière en suivant une subdivision logique, pertinente, et porteuse d'informations.

Telle est :

Nous opterons pour une subdivision sur deux étapes :

- Par rapport aux jours
- Par aux heures de la journée
- Par rapport aux jours : Nous l'avons subdivisé en des parties du mois :
 - jour 1 → 5
 - jour 5 → 11
 - jour 11 → 16
 - jour 16 → 21
 - jour 21 → 26
 - jour 26 → 31

Cette division relève de l'état financier des individus et de sa variation selon les différentes phases du mois (début du mois : entre jour 1 et jour 5 nous nous attendons à hausse des taux de consommation des individus, tandis qu'à la fin du mois nous nous attendons à une baisse.)

- Par aux heures de la journée : Nous l'avons subdivisé ainsi :
 - Heure : 4 → 10
 - Heure : 10 → 12
 - Heure : 12 → 15
 - Heure : 15 → 18
 - Heure : 18 → 00
 - Heure : 0 → 4

De même, cette subdivision tient compte des horaires de pointe probable selon les besoins des individus

Nous obtiendrons ainsi, des intervalles réguliers sans perte de pertinence des informations.

Nous symboliserons une date selon cette nouvelle distribution par : $G(x,y)$ ou:

- x représente la phase du jour dans le mois [1..6],
- y représente la phase de l'heure de la journée [1..6]

A.V.2. Data Preparation :

A.V.2.1. Data splitting :

Il y a eu séparation de la Base de Données des commandes en deux bases de données des produits du cafétéria, et du produit du restaurant.

A.V.2.2. Feature Engineering :

Nous décidons de créer un nouveau feature : **Converted_Date**, qui représente la nouvelle date et heure formatée selon le principe décrit précédemment.

D'autres Features avaient été créés :

- **Sum_Revenues** : Présente la somme des revenus pour un G(x,y)
- **Total_Quantity** : Présente la quantité totale vendue pour un G(x,y)
- **NB_Delivery** : Présente le Nombre de livraisons pour un G(x,y)
- **NB_Total_Commandes** : Présente le Nombre total de commandes pour un G(x,y)
- **Sup_caramel**
- **Sup_nutella**
- **Jus_Orange**
- **Citronnade**
- **Express**
- **Direct**
- **Turc**

: Présentent la quantité totale commandée d'un élément Z pour un G(x,y), ou Z appartient à {Sup_caramel, Sup_nutella, Jus_Orange, Citronnade, Express , Direct, Turc}

Nous créons un nouveau DataFrame avec les nouveaux Features, et obtenons ainsi, le tableau suivant.

Converted_Date	Sum_Revenues	Total_Quantity	NB_Delivery	NB_Total_Commandes	Sup_caramel	Sup_nutella	Jus_Orange	Citronnade	Express	Direct	Turc
2023-03-29 08:00:00	2023-03-29 08:00:00	32.40	6	0	6	6	0	0	0	0	0
2023-03-29 00:00:00	2023-03-29 00:00:00	3.60	4	0	4	3	0	0	0	0	0
2023-03-29 22:00:00	2023-03-29 22:00:00	1.80	2	0	2	1	0	0	0	0	0
2023-03-29 23:00:00	2023-03-29 23:00:00	270.00	5	0	5	4	0	0	0	0	0
2023-04-03 22:00:00	2023-04-03 22:00:00	3.60	2	0	2	1	0	0	0	0	0
...
2023-04-07 09:00:00	2023-04-07 09:00:00	39.84	4	0	4	0	0	0	1	0	0
2023-04-07 10:00:00	2023-04-07 10:00:00	26.00	2	0	2	0	0	0	0	1	0
2023-08-07 12:00:00	2023-08-07 12:00:00	12.00	2	1	1	0	0	0	0	0	0
2023-08-07 12:00:00	2023-08-07 12:00:00	82.00	13	2	6	0	0	3	2	0	0
2023-08-07 13:00:00	2023-08-07 13:00:00	33.00	5	0	3	0	0	1	1	0	0

Figure 38 : Table créé avec les nouveaux Features

Bien que nous ayons obtenu des données régulières, le nombre de sujets a diminué de 321 à 69 sujets.

A.V.3. Time Series Visualization :

Le plot de la nouvelle Base de Données discrète, donne le schéma suivant.

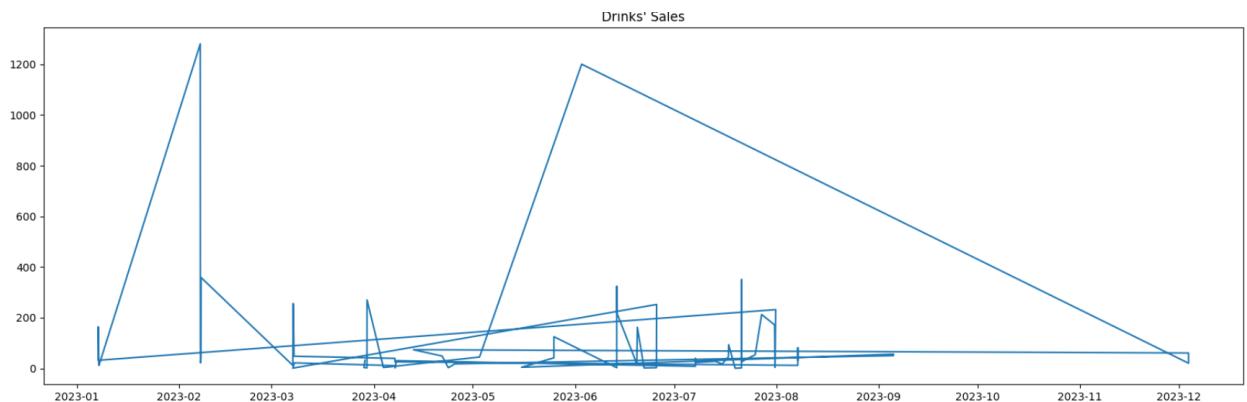


Figure 39 : Time Series des ventes du café avec des dates régulières

A.V.4. Conclusion sur la solution proposée :

Nous pourrons bien réaliser l'impossibilité d'employer cette Base de Données pour les prédictions.

Dans la même approche, il est possible de subdiviser la journée en 24H, et de donner la valeur 0 lors d'absence de data.

Il est aussi possible de remplacer les valeurs ajoutées par la moyenne à la place de 0.

Mais ces solutions mèneront encore une fois à la non intégrité des données, rendant nos prédictions inutiles, non fiables et non descriptives de la réalité.

Ainsi, le plus logique à faire, serait, d'attendre quelques mois, pour avoir une Base de Données descriptive d'une année entière, ainsi, les Seasonalités pourront bien être capturées, ainsi que les variations de la consommation selon les saisons : saison ou les étudiants et écoliers sont présents, saisons ou se rendre à un café ou un restaurant devient une activité de distraction et non pas un besoin quotidien etc.

Chapitre VI : Par analogie

Comme le titre l'indique, les autres points évoqués dans la section 2.2.1. Par rapport au restaurant spécifique , seront traités d'une manière complètement analogue à la précédente, pour cela, nous éviterons les détails exhaustives de description des étapes du CRISP-DM, et nous nous focaliserons sur l'interprétation des densités des erreurs du Fitting, ainsi que les variations des valeurs réelles.

Les phases de Business Understanding, Data Understanding, et Data Preparation sont communes à la partie A du projet.

A.VI.1. Prédition des quantités de produits vendus

A.VI.1.1. Prédictions des ventes du café

A.VI.1.1.1. Modeling

A.VI.1.1.1.1. Differencing the time Series

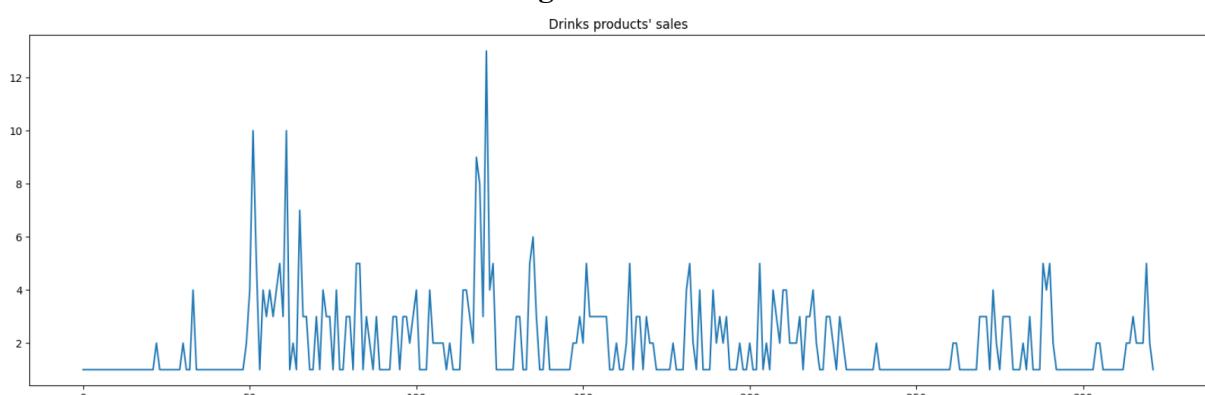


Figure 40 : allure des quantités vendues du café avant Differencing

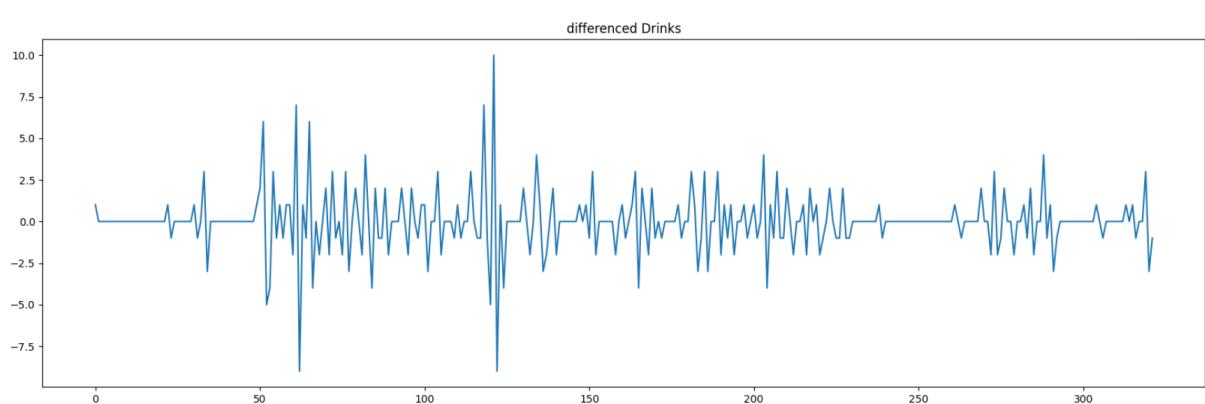


Figure 41 : allure des quantités vendues du café après Differencing

A.VI.1.1.1.2. Evaluation des données et choix des paramètres

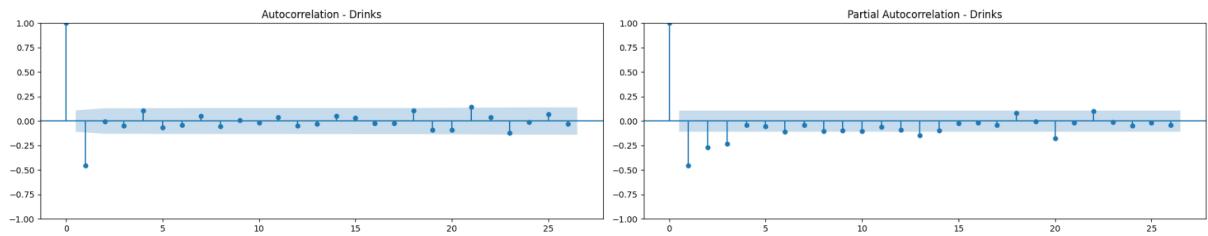


Figure 42 : Allure des courbes des graphes ACF et PACF pour la cafétéria

Ainsi, les paramètres choisis pour le modèle ARIMA sont :

$$p=1, \quad I=0, \quad q=3$$

A.VI.1.1.1.3. Graphes des résidus

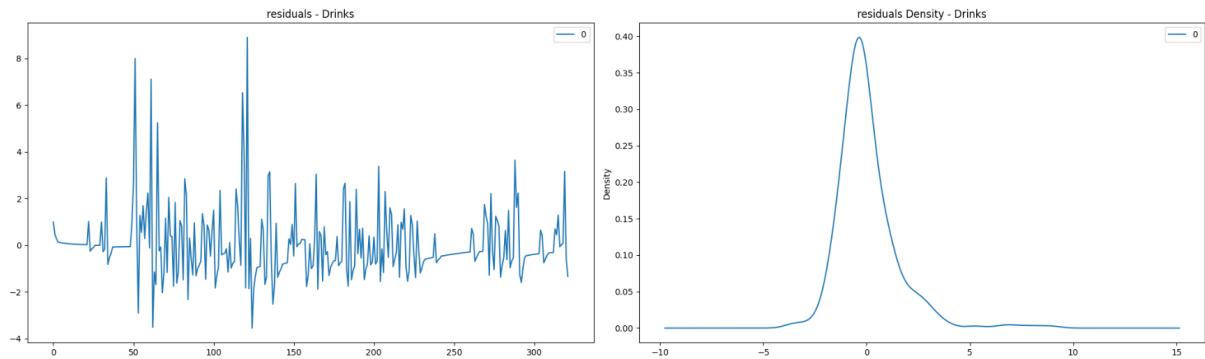


Figure 43 : Allure des graphes des résidus et densité des résidus

0	
count	322.000000
mean	0.032374
std	1.468261
min	-3.544454
25%	-0.766583
50%	-0.275983
75%	0.492650
max	8.911291

Figure 44 : Résumé de l'analyse des densités des résidus

A.VI.1.1.4. Allure du Fitted Values graph

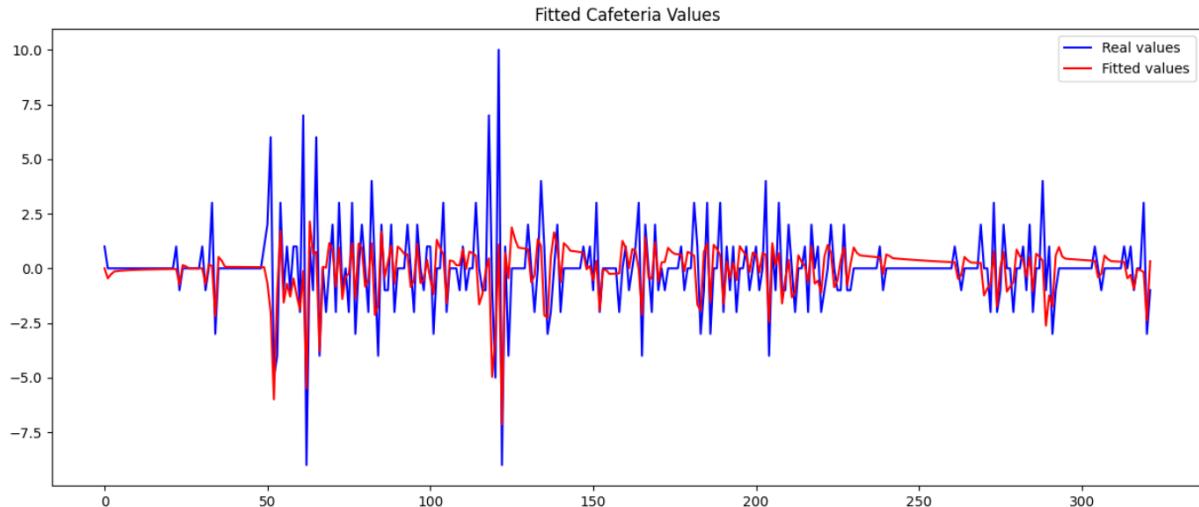


Figure 45 : Allure du Fitted model par rapport aux valeurs réelles

A.VI.1.1.2. Interprétation des résultats

La moyenne est de 0.03, centrée autour de 0, mais l'écart-type est assez remarquable par rapport aux valeurs maximales prises par les quantités de produits : 1.8 alors que la valeur maximale est aux alentours de 9.

On pourrait conclure qu'il faudrait recourir à d'autres paramètres ou à un autre modèle qui tient compte de la Seasonality.

Toutefois, le graphe des fitted values semble bien correspondre aux valeurs réelles, la courbe des Fitted Values est légèrement en dessus de celle des valeurs réelles, mais elle la suit dans toutes ses variations locales.

Cela pourrait relever de l'underFitting vu la quantité réduite des données.

A.VI.1.2. Prédictions des ventes du restaurant

A.VI.1.2.1. Modeling

A.VI.1.2.1.1. Differencing the time Series

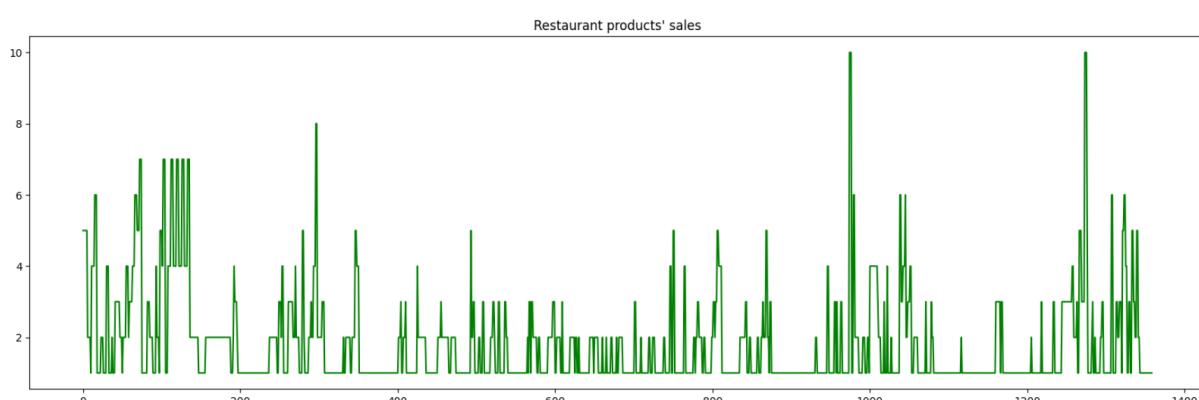


Figure 46 : Allure des quantités vendues du restaurant avant Differencing

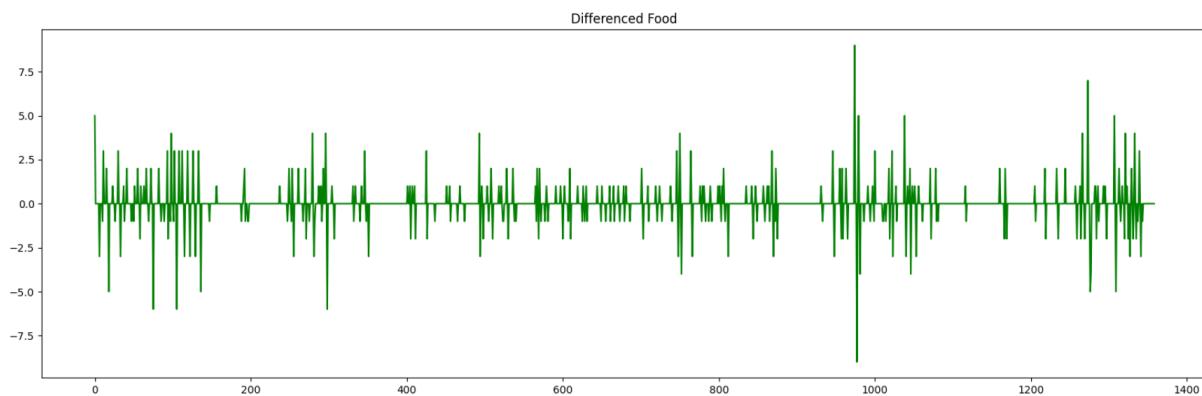


Figure 47 : Allure des quantités vendues du restaurant après Differencing

A.VI.1.2.1.2. Evaluation des données et choix des paramètres

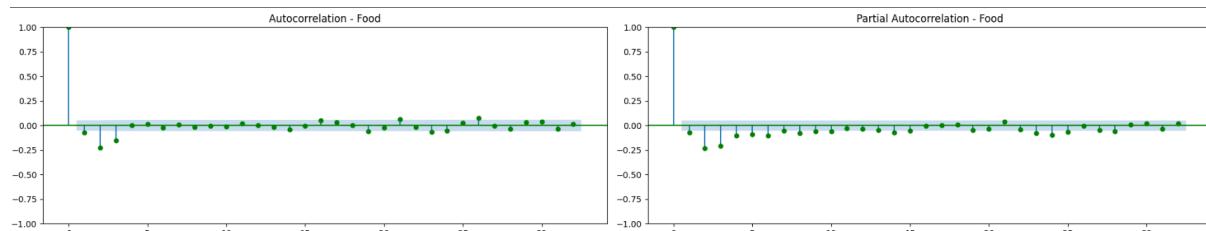


Fig : Allure des courbes des graphes ACF et PACF pour le restaurant

Ainsi, les paramètres choisies pour le modèle ARIMA sont : $p=1$, $I=0$, $q=3$

A.VI.1.2.1.3. Graphes des résidus

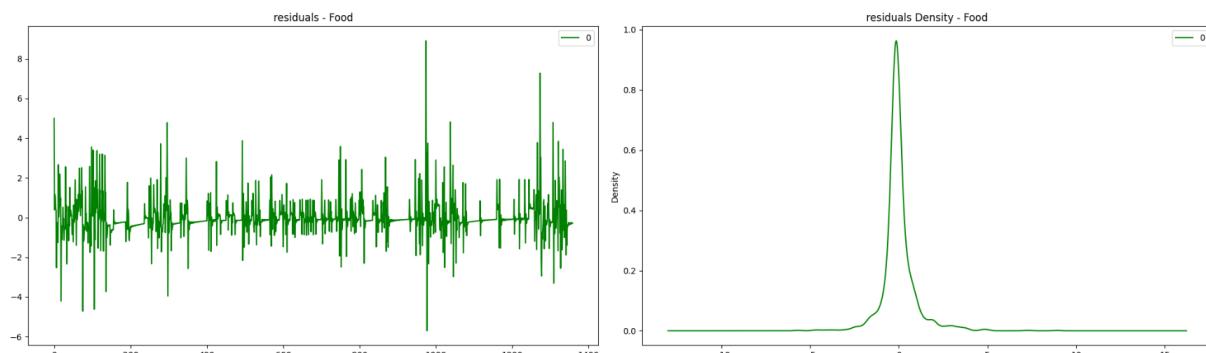


Figure 48 : Allure des graphes des résidus et densité des résidus

```

count    1359.000000
mean      0.003049
std       0.951002
min     -5.705374
25%    -0.269618
50%    -0.120682
75%     0.163442
max      8.909125

```

Figure 49 : Résumé de l'analyse des densités des résidus

A.VI.1.2.1.4. Allure du Fitted Values graph



Figure 50 : Allure du Fitted model par rapport aux valeurs réelles

A.VI.1.2.2. Interprétation des résultats

La moyenne est centré autour de 0, La valeur de l'écart-type semble être relativement réduite par rapport aux valeurs maximales pouvant être atteinte par les quantités de produits. De plus, la courbe de densité est fine, ce qui soutient le jugement de la valeur de l'écart-type en tant que “relativement réduite”.

La courbe des Fitted Values semble bien correspondre aux données réelles, elle la suit dans toutes ses variations locales (housse et baisse) par contre, l'intensité des pics est beaucoup inférieure à celle des valeurs réelles.

A.VI.2. Prédition des produits à livrer et à emporter

A.VI.2.1. Visualisation des Time series

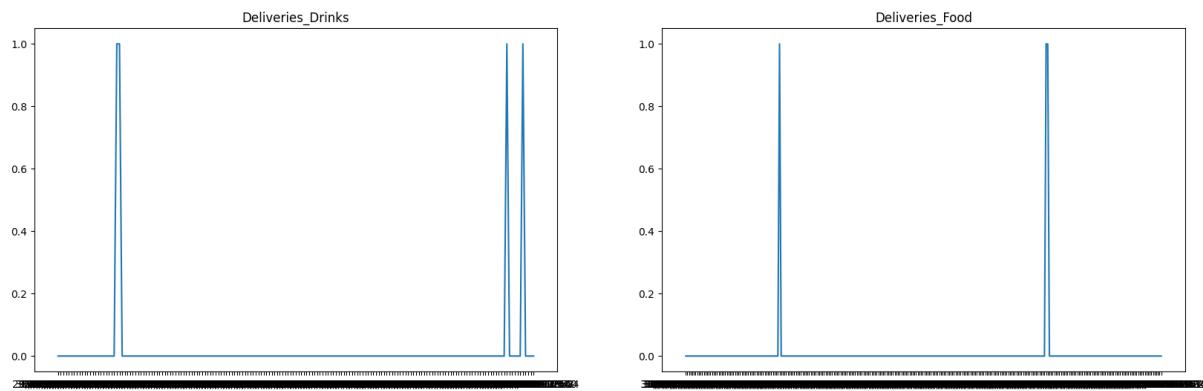


Figure 51 : Allure des variations du nombre de livraisons du restaurant

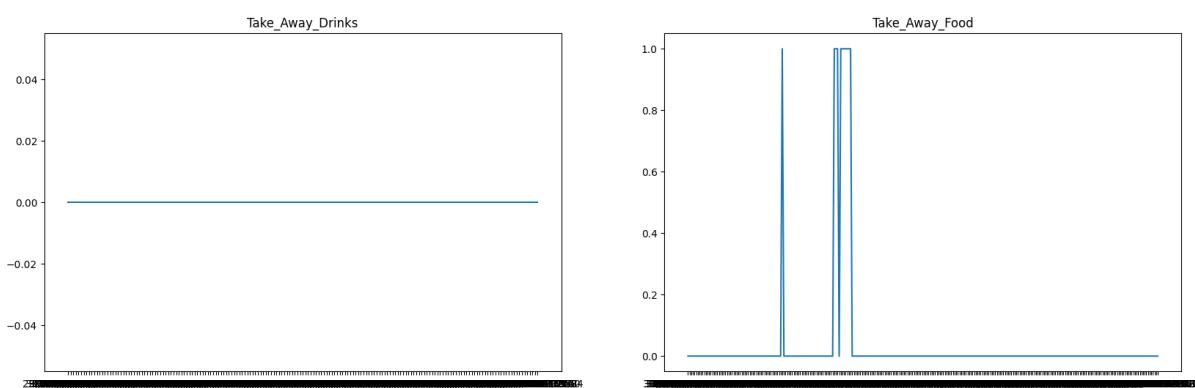


Figure 52 : Allure des variations du nombre des “à emporter” du café

Ci-dessous les graphes différenciés :

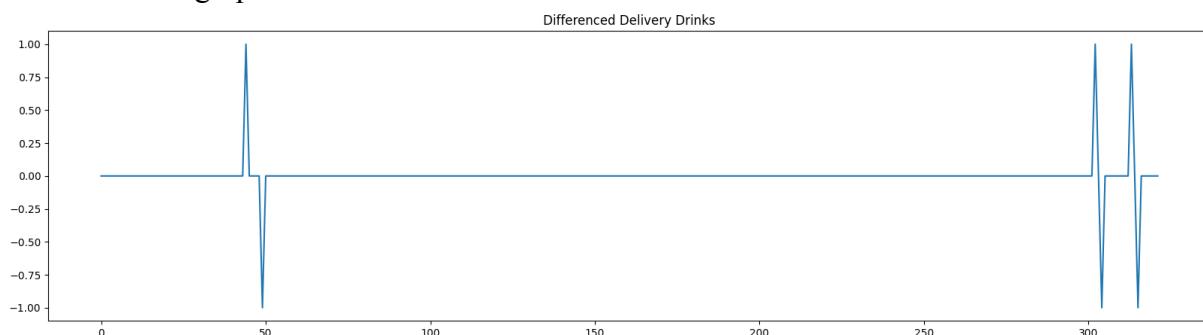


Figure 53 : Allure des variations du nombre de livraisons du café différencié

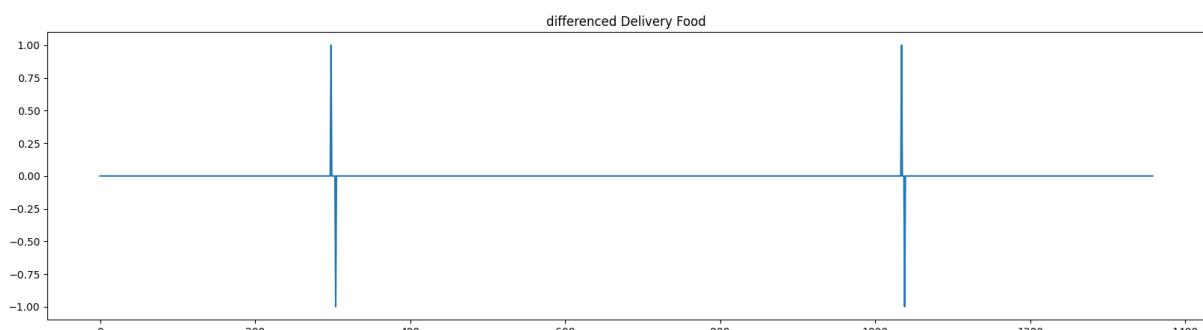


Figure 54 : Allure des variations du nombre de livraisons du restaurant différencié

A.VI.2.2. Visualisation des graphes ACF et PACF

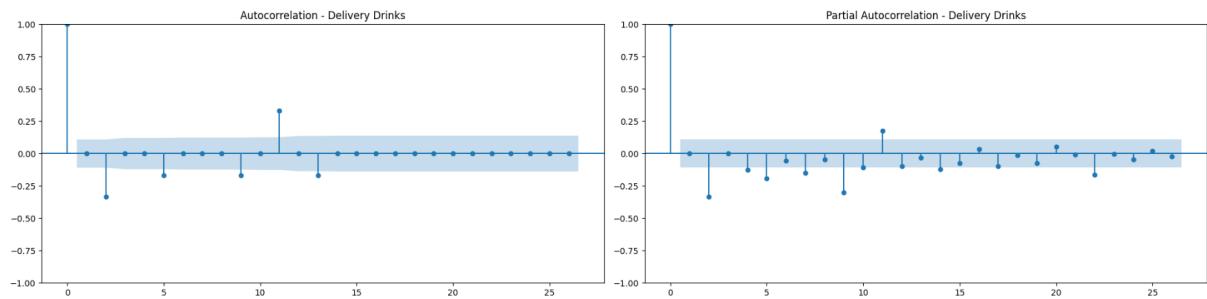


Figure 55 : Allure des graphes ACF et PACF pour les produits du café

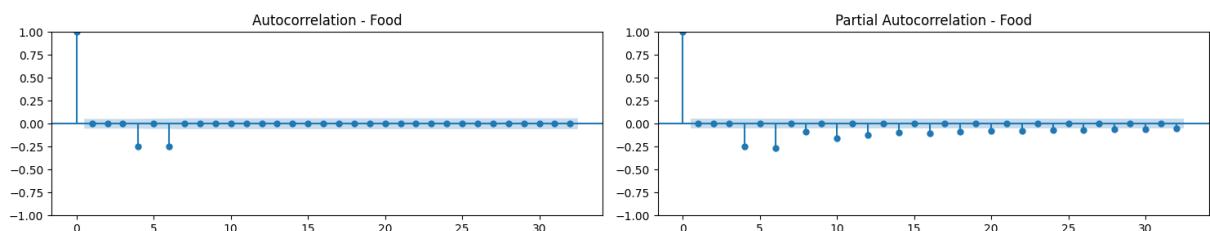


Figure 56 : Allure des graphes ACF et PACF pour les produits du restaurant

A.VI.2.3. Visualisation du graphe des résidus

Nous présentons ci-dessous les graphes des résidus pour la cafétéria et le restaurant

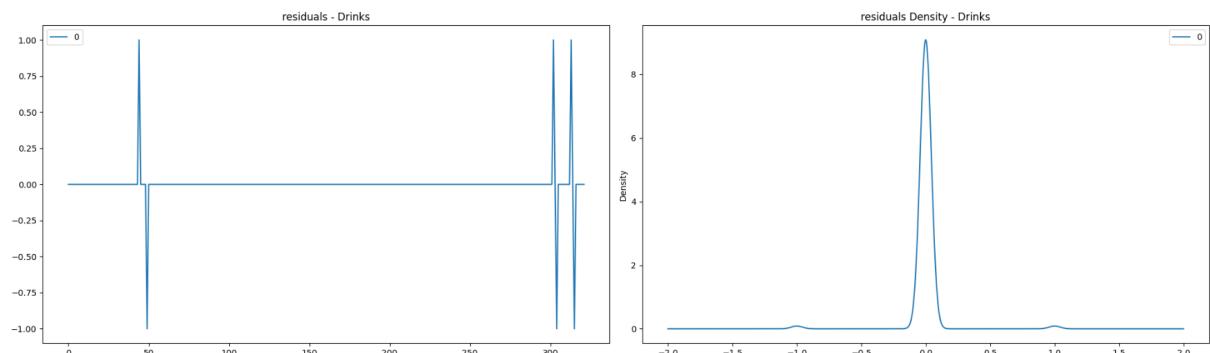


Figure 57 : Allure des résidus et des densités des résidus pour les produits du café

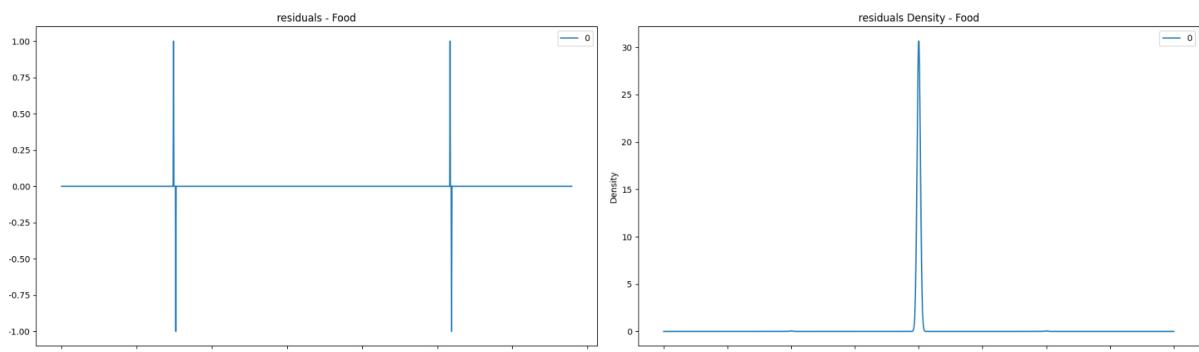


Figure 58 : Allure des résidus et des densités des résidus pour les produits du restaurants

A.VI.2.4. Evaluation

Nous ne trouvons que deux pics pour les livraisons, au plus de trois, sur une durée de 3 à 4 mois.

Le graphe des densités des résidus donne un pic très fin centré sur 0, indicatif d'un faible écart-type.

Donc, nous concluons que le fitting des données donne un bon résultat.

Conclusion Partie A

Nous éviterons de ré-évoquer les interprétations détaillées déjà mentionnées à chaque sous-partie du projet.

La qualité des données très particulière (courte durée, 2 mois d'été et un mois de jeûn en Tunisie, Plusieurs trous de données de longue durée) fait que la souche utilisée n'est pas très rassurante en terme de fiabilité des résultats ou d'interprétation,
Pour cela, nous voulons bien réétudier ce cas, mais avec des données de meilleure qualité pour pouvoir tirer des conclusions constructives et faire des prédictions fiables et significatives.

Parmi les perspectives, nous pourrons recourir à un modèle LSTM, et SARIMA, et comparer les résultats obtenus.

Nous pourrons aussi utiliser un Grid pour essayer les paramètres du modèle SARIMAX et chercher ceux ayant le meilleur résultat (minimum Loss, maximum Accuracy, Marge de temps acceptable selon les ressources disponibles en ligne)

Partie B :Segmentation de la base de clients de la start up

Chapitre I : Business Understanding

La startup permet à des restaurants et cafés d'enregistrer les commandes passées en caisse dans une Base de Données. Elle nécessite ainsi l'enregistrement préalable des menus.

Les clients que nous tâchons de classifier dans cette partie, sont les restaurants et cafés qui ont décidé d'opter pour ce choix d'enregistrer leurs commandes avec la startup.

Ce qui nous intéresse ici, c'est quelles sont les caractéristiques communes qui font de cette start up un prestataire de service de valeur, et pour quelle cible ?

Chapitre II : Data Understanding

La phase de collecte et de description des données avait été déjà expliquée dans la partie A vu le besoin.

B.II.1. Data Exploration

1.1. Exploration Pré-DeNesting

Nous avons 152 clients, parmi leurs features pertinents :

- Info (qui contient la majorité des informations utiles concernant le client),
- Suppliers : qui n'est présente que pour 6 clients
- Clients : une colonne qui indique les clients des restaurants-café, celle-ci est vide pour 126 clients.

Nous cherchons à voir les catégories des valeurs des colonnes de chaque feature.

1.2. Exploration Post-DeNesting

Nous avons exploré les valeurs prises par chaque colonne, et les nombres d'occurrence de chacune de ces valeurs.

has_tables :
True 95
False 18
Name: has_tables, dtype: int64

dailyRevenue :
0.0 119
Name: dailyRevenue, dtype: int64

Figure 59 : Catégories des données et occurrences par colonne

Nous avons visualisé une description globale des données, comme démontré dans le graphique qui suit.

	deleted	info.dailyRevenue	info.ticket_number	info.city.country_id	info.city.flag	info.city.id	info.city.state_id	info.country.flag	info.country.id	info.state.country_id	info.s
count	150.000000	119.0	151.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
mean	0.260000	0.0	19.370861	224.0	1.0	106929.0	2558.0	1.0	224.0	224.0	
std	0.440104	0.0	48.681087	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
min	0.000000	0.0	0.000000	224.0	1.0	106929.0	2558.0	1.0	224.0	224.0	
25%	0.000000	0.0	0.000000	224.0	1.0	106929.0	2558.0	1.0	224.0	224.0	
50%	0.000000	0.0	2.000000	224.0	1.0	106929.0	2558.0	1.0	224.0	224.0	
75%	1.000000	0.0	14.000000	224.0	1.0	106929.0	2558.0	1.0	224.0	224.0	
max	1.000000	0.0	369.000000	224.0	1.0	106929.0	2558.0	1.0	224.0	224.0	

Figure 60 : Description de la base de données des clients

Une Description des types de données a été visualisée ci-dessous :

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 152 entries, 0 to 151
Data columns (total 27 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   id               152 non-null    object  
 1   deleted          150 non-null    float64 
 2   deleted_at       40 non-null    object  
 3   info.dailyRevenue 119 non-null    float64 
 4   info.date_reinit_ticket 151 non-null    object  
 5   info.has_tables  113 non-null    object  
 6   info.name         152 non-null    object  
 7   info.opening_time 152 non-null    object  
 8   info.paper_cutter 152 non-null    bool    
 9   info.print_ticket 111 non-null    object  
 10  info.shop_type_flag 152 non-null    object  
 11  info.shop_type_name 152 non-null    object 

```

Figure 61 : Informations concernant la base de données des clients

B.II.2. Vérification de la qualité des données

Nous trouvons que la sous colonne “Daily Revenu” (de “Info”) est nulle pour tous les restaurants quelqusoit le nombre de commandes passées.

La seule colonne entièrement remplie pour tous les clients est la colonne “Info”.

Il ya plusieurs données nulles qui auront pu être utiles dans la segmentation, tel “Suppliers”, “DailyRevenu”, et “emplacement”.

Chapitre III : Data Preparation

B.III.1. Data DeNesting

Nous recourons à une phase de DeNesting des données de la colonne “Info” jugées pertinentes.

Ci-dessous la colonne “Info” avant le processus de Denesting :

	id	info
	1LeuTEU1OaTbPIDjbkfa2LdHJ2m2	{"dailyRevenue": 0, "date_reinit_ticket": "13/07/2022", "has_tables": true}
	1YUuyZj3vaguDV6kniXkjZHGsVJ3	{"dailyRevenue": 0, "date_reinit_ticket": "13/07/2022", "has_tables": false}
	1fdugKSBzPdBloqLvtlzHhfKZ22	{"dailyRevenue": 0, "date_reinit_ticket": "14/07/2022", "has_tables": true}
	1HIHQ4SKtgdB097iGGt2cSiNx92	{"dailyRevenue": 0, "date_reinit_ticket": "24/03/2022", "has_tables": true}
	1wkiFE5YD3fxqTICRHJyGRoCs3k2	{"dailyRevenue": 0, "date_reinit_ticket": "14/07/2022", "has_tables": true}

	yPFnCPwL8dNnlvz0PzbKRwUOAfF2	{"date_reinit_ticket": "03/08/2021", "has_tables": true}

Figure 62 : Colonne Info avant DeNesting

Ci-dessous le DataFrame après Denesting :

id	deleted	deleted_at	info.dailyRevenue	info.date_reinit_ticket	info.has_tables	info.name	info.opening_time	info.paper_cutter	info.print_ticket	...	info
dHJ2m2	0.0	None	0.0	13/07/2022	True	ILIADE BAR 1	05:00	True	True	...	
HGsVJ3	0.0	None	0.0	13/07/2022	False	THE GRILL	06:00	False	True	...	
HhfKZ22	0.0	None	0.0	14/07/2022	True	MISK	05:00	False	True	...	
cSiNx92	1.0	24-03-2022	0.0	24/03/2022	True	STOCK TEST	06:00	True	True	...	
RoCs3k2	0.0	None	0.0	14/07/2022	True	365 CLUB CAFE	07:00	True	True	...	
...
UOAfF2	1.0	27-12-2021	NaN	03/08/2021	True	BAIA TEST	05:00	True	NaN	...	
RMUg73	0.0	None	0.0	13/07/2022	True	BAR CENTRAL CWM	05:00	True	True	...	
UePUH3	1.0	27-12-2021	0.0	29/05/2021	NaN	CAFE'IN	05:00	False	NaN	...	
JOeoty2	0.0	None	NaN	07/07/2023	True	590	09:09	True	NaN	...	
WVmml1	0.0	None	0.0	08/01/2022	True	Chocolat test	06:00	True	NaN	...	

Figure 63 : DataFrame après DeNesting

B.III.2. Data Selection

Il y a eu une sélection avant le Denesting, qui a abouti à une suppression de toutes les colonnes sauf : id, info, deleted, deleted_at.

Ci-dessous la table conservée :

		id	deleted	deleted_at	info
0		1LeuTEU1OaTbPIDjbkfa2LdHJ2m2	0.0	NaN	{'dailyRevenue': 0, 'date_reinit_ticket': '13/...
1		1YUuyZj3vaguDV6kniXkjZHGsVJ3	0.0	NaN	{'dailyRevenue': 0, 'date_reinit_ticket': '13/...
2		1fdugKSBzPdBloqLvtlzHhfKZ22	0.0	NaN	{'dailyRevenue': 0, 'date_reinit_ticket': '14/...
3		1IHIQ4SKtgdB097iGGt2cSiNxr92	1.0	24-03-2022	{'dailyRevenue': 0, 'date_reinit_ticket': '24/...
4		1wkiFE5YD3fxqTICRHJyGRoCs3k2	0.0	NaN	{'dailyRevenue': 0, 'date_reinit_ticket': '14/...
...	
147		yPFnCPwL8dNnlvz0PzbKRwUOAfF2	1.0	27-12-2021	{'date_reinit_ticket': '03/08/2021', 'has_table...
148		yrtyvXC2bkVtAW6REaBy1MRMUg73	0.0	NaN	{'dailyRevenue': 0, 'date_reinit_ticket': '13/...
149		ywAv0kkVVifXpe108MhktRUePUH3	1.0	27-12-2021	{'dailyRevenue': 0, 'date_reinit_ticket': '29/...
150		z6g3Pjyf6PPLhQb9XhUg7JOeoty2	0.0	NaN	{'currency': 'TND', 'date_reinit_ticket': '07/...
151		zNRmMJOGbretRoMLaE5vDjvWVmm1	0.0	NaN	{'dailyRevenue': 0, 'date_reinit_ticket': '08/...

152 rows x 4 columns

Figure 64 : DataFrame après sélection préliminaire des Features

Une deuxième sélection a eu lieu en se basant sur la pertinence de l'apport du Feature, et du nombre de valeurs nulles de celui-ci.

Cette sélection a abouti à la table suivante.

id	deleted	deleted_at	info.dailyRevenue	info.date_reinit_ticket	info.has_tables	info.name	info.opening_time	info.paper_cutter	info.print_ticket	info.shop_type_flag	info.shop_type_name
1LeuTEU1OaTbPIDjbkfa2LdHJ2m2	0.0	None	0.0	13/07/2022	True		05:00	True	True	restaurant	Restaurant
1YUuyZj3vaguDV6kniXkjZHGsVJ3	0.0	None	0.0	13/07/2022	False		06:00	False	True	restaurant	Restaurant
1fdugKSBzPdBloqLvtlzHhfKZ22	0.0	None	0.0	14/07/2022	True		05:00	False	True	restaurant	Restaurant
1IHIQ4SKtgdB097iGGt2cSiNxr92	1.0	24-03-2022	0.0	24/03/2022	True		06:00	True	True	restaurant	Restaurant
1wkiFE5YD3fxqTICRHJyGRoCs3k2	0.0	None	0.0	14/07/2022	True		07:00	True	True	restaurant	Restaurant
...
yPFnCPwL8dNnlvz0PzbKRwUOAfF2	1.0	27-12-2021	NaN	03/08/2021	True		05:00	True	NaN	restaurant	Restaurant
yrtyvXC2bkVtAW6REaBy1MRMUg73	0.0	None	0.0	13/07/2022	True		05:00	True	True	restaurant	Restaurant
ywAv0kkVVifXpe108MhktRUePUH3	1.0	27-12-2021	0.0	29/05/2021	NaN		05:00	False	NaN	restaurant	Restaurant
z6g3Pjyf6PPLhQb9XhUg7JOeoty2	0.0	None	NaN	07/07/2023	True		09:09	True	NaN	restaurant	Restaurant
zNRmMJOGbretRoMLaE5vDjvWVmm1	0.0	None	0.0	08/01/2022	True		06:00	True	NaN	restaurant	Restaurant

Figure 65 : DataFrame après DeNesting des features sélectionnés

B.III.3. Data Cleaning

Les noms des colonnes avaient été modifiés, et les colonnes à valeurs nulles avaient été éliminées, tel est l'état du nouveau DataFrame ci-dessous.

index		id	deleted	has_tables	name	opening_time	shop_type_flag	Number_Clients
0	0	1LeuTEU1OaTbPIDjbkfa2LdHJ2m2	0.0	1.0		05:00	restaurant	8.0
1	1	1YUuyZj3vaguDV6kniXkjZHGsVJ3	0.0	0.0		06:00	restaurant	1.0
2	2	1fdugKSBzPdBilloqLvtlzHhfKZ22	0.0	1.0		05:00	restaurant	21.0
3	3	1IHQ4SKtgdB097iGGt2cSiNxr92	1.0	1.0		06:00	restaurant	8.0
4	4	1wkiFE5YD3fxqTICRHJyGRoCs3k2	0.0	1.0		07:00	restaurant	151.0
...
147	147	yPFnCPwL8dNnlvz0PzbKRwUOAfF2	1.0	1.0		05:00	restaurant	0.0
148	148	yrtyvXC2bkVtAW6REaBy1MRMUg73	0.0	1.0	BAF	05:00	restaurant	0.0
149	149	ywAv0kkVViXpe108MhktRUePUH3	1.0	NaN		05:00	restaurant	0.0
150	150	z6g3Pjf6PPLhQb9XhUg7JOeoty2	0.0	1.0		09:09	restaurant	0.0
151	151	zNRmMJOGbretRoMLaE5vDjvWVmm1	0.0	1.0		06:00	restaurant	2.0

Figure 66 : DataFrame partiellement Nettoyé

Il y avait eu modification des valeurs des colonnes pour les adapter au type de traitement durant la phase de Modeling.

Les “Opening Time” étaient transformés en float values.

Les shop_type etait codée de la sorte :

1 : restaurant, 2: fast food, 3:pizzeria, 4:beach bar

Quelques sujets de test avaient été supprimés, ce qui donne finalement 148 sujets.

Ci-dessous le DataFrame final.

index		id	deleted	has_tables	name	opening_time	shop_type_flag	Number_Clients
0		1LeuTEU1OaTbPIDjbkfa2LdHJ2m2	0	1	ILIADE BAR 1	5.0	1	8
1		1YUuyZj3vaguDV6kniXkjZHGsVJ3	0	0	THE GRILL	6.0	1	1
2		1fdugKSBzPdBilloqLvtlzHhfKZ22	0	1	MISK	5.0	1	21
3		1IHQ4SKtgdB097iGGt2cSiNxr92	1	1	STOCK TEST	6.0	1	8
4		1wkiFE5YD3fxqTICRHJyGRoCs3k2	0	1	365 CLUB CAFE	7.0	1	151
...	
144		yPFnCPwL8dNnlvz0PzbKRwUOAfF2	1	1	BAIA TEST	5.0	1	0
145		yrtyvXC2bkVtAW6REaBy1MRMUg73	0	1	BAR CENTRAL CWM	5.0	1	0
146		ywAv0kkVViXpe108MhktRUePUH3	1	0	CAFE'IN	5.0	1	0
147		z6g3Pjf6PPLhQb9XhUg7JOeoty2	0	1	590	9.0	1	0
148		zNRmMJOGbretRoMLaE5vDjvWVmm1	0	1	Chocolat Test	6.0	1	2

Figure 67 : DataFrame nettoyé

Chapitre IV : Modeling

B.IV.1. Choix préliminaire du modèle

Nous avons choisi KMeans clustering algorithm.

C'est un choix personnel qui relève de notre volonté de sa mise en œuvre sur un cas réel.

B.IV.2. Détermination du nombre de Clusters

Après sélection des critères du Clustering, dont nous avons varié, et sommes arrivés au choix des 2 critères suivants qui donnent des résultats pertinents : opening time, Number Clients. Le Elbow Graph ci-dessous montre le nombre de clusters optimal :

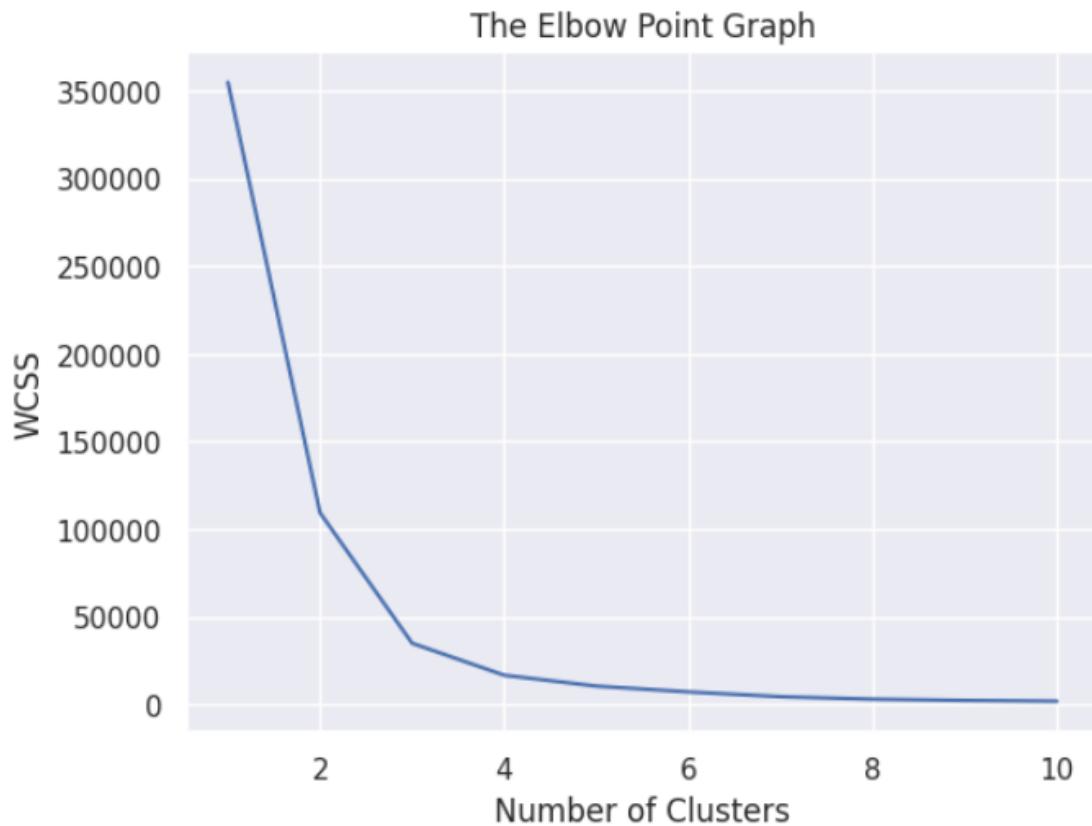


Figure 68 : Elbow Graph

B.IV.3. Résultat du clustering :

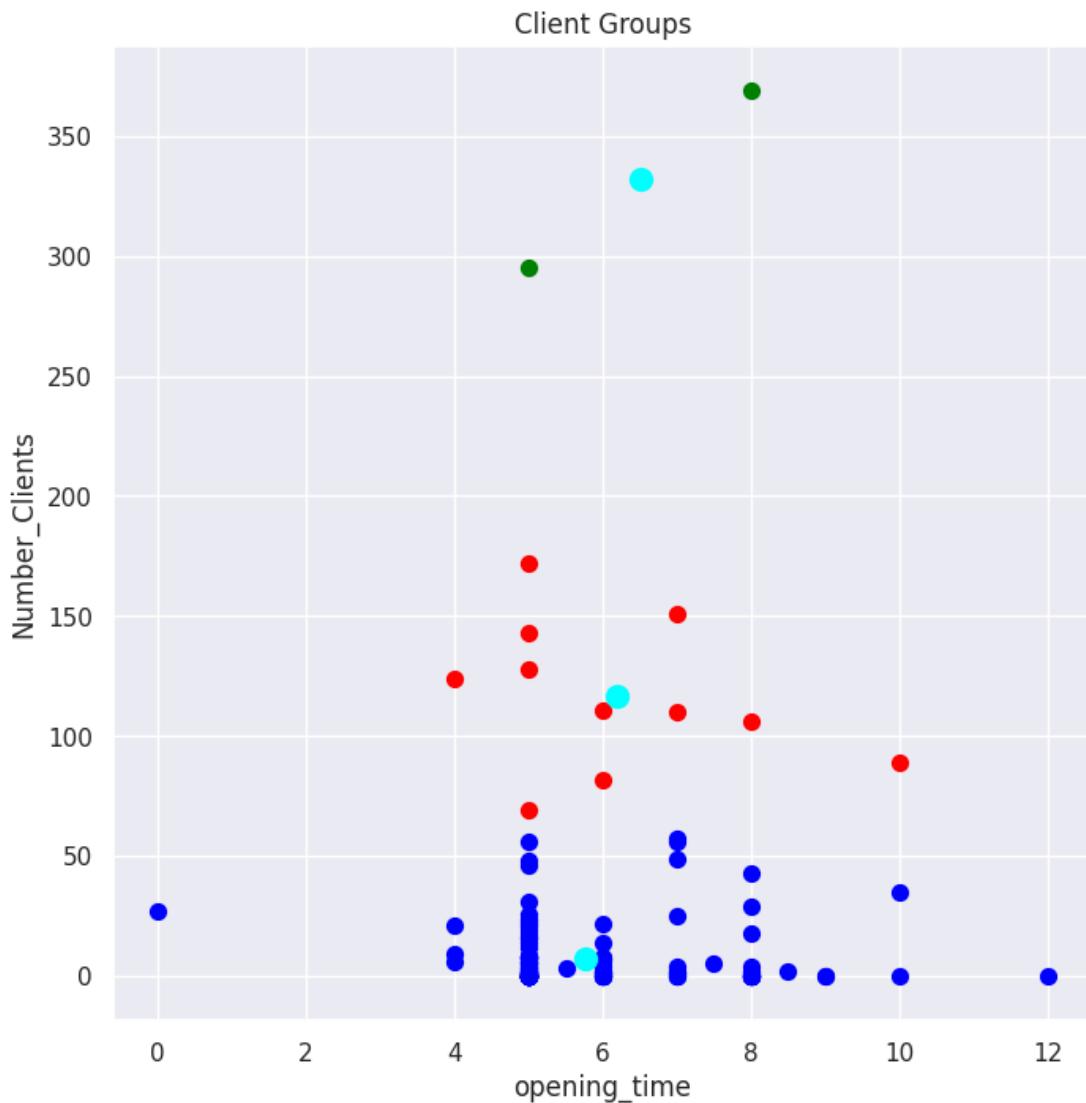


Figure 69 : Plot des Clusters selon les critères choisis

Chapitre V : Évaluation et Interprétation

Nous obtenons 3 clusters :

- 1er : Horaires d'ouverture entre 5h et 8H, et ayant un nombre très élevé de clients.
- 2ème : Horaires d'ouverture entre 4h et 10H, et ayant un nombre relativement faible de clients par rapport à d'autres.
- 3ème : Horaires d'ouverture entre 4h et 10H, et ayant un nombre moyennement élevé de clients.

→ Le 1er cluster représente probablement des cafétéria proche directement à une zone industrielle et avec une fourchette de prix moyenne, donc une consommation quotidienne de la classe des employées, ou étudiants.

→ Le 2ème Cluster représente probablement des cafétéria dans des zones relativement éloignées des entreprises et à fourchette de prix relativement en dessous de la moyenne, en effet, la majorité de leurs commandes commencent probablement aux alentours de 9H ou plus, vu qu'ils ciblent des individus visitant un café pour le simple plaisir.

→ Le 3ème Cluster représente probablement des cafétéria dans des zones proches d'employeurs et entreprises mais à fourchette de prix relativement au dessus de la moyenne.

Nous pourrons effectuer une subdivision selon d'autres critères pour distinguer entre les 2 derniers clusters.

Des données concernant l'emplacement des cafétérias auraient pu être un très bon critère de distinction entre les clusters et fournira une validation des hypothèses émises.
Une telle donnée n'est pas présente.

Conclusion Partie B

L'étude pourra toucher plusieurs autres critères combinés, tel : Emplacement, Alentours de l'emplacement (zone peu ou beaucoup fréquentée et par quelles type d'individus), fourchette de prix moyenne pour les produits les plus consommées, et les moins consommées, type de café et clients qu'elle cible, Chiffre d'affaires, type d'activité axée sur les ventes de café, croissants et jus, ou sur les autres types de produits(gateau,pancakes,milkshakes etc.)

Même une seule donnée de cette liste aurait pu donner un résultat plus pertinent que celui présenté.

Cette analyse permettra à la startup de comprendre les clients qu'elle obtient par rapport à sa cible, et ainsi, de mieux adapter sa stratégie de marketing, sa prospection, et ses prestations aux prospects.

Conclusion générale et perspectives

Une meilleure qualité de données permettra des points d'améliorations majeurs et une valeur ajoutée remarquable sur chaque cafétéria ou restaurant de la base de données d'une part, et sur la startup d'autre part.

L'objectif était principalement de mettre en pratique des connaissances théoriques apprises de façon autonome sur des cas réels, Nous jugeons cet objectif atteint.

De plus, ce projet, en ses deux parties, assure une valeur ajoutée remarquable sur les parties prenantes, malgré la qualité des données, ainsi, une proposition de collaboration avec la startup pour un travail effectué sur une base de données complète pourra être envisagée.