Test de recrutement - Data Scientist

Air France - KLM

Commercial Data Analytics - Mai 2022

Exercice 1 - 📊 Exploration et visualisation de données

L'un des objectifs de l'équipe RMP (Revenue Management and Pricing) est **maximiser la recette des vols d'Air France**. Elle vous sollicite afin d'analyser l'impact de la pandémie sur le comportement de voyage et de réservation des clients.

Un jeu de données a été extrait (*flight_data_extract.csv*) des bases d'Air France et vous trouverez en annexe une description des champs qui ont été sélectionnés pour cette étude. Il contient trois semaines de ventes (première semaine d'Avril 2019, 2020 et 2021). Une deuxième table (*area.csv*) permet d'établir une correspondance entre l'aéroport de départ (aussi appelé aéroport d'origine) et différents niveaux de regroupement géographique (zone, sous-zone, pays, etc.) utilisés pour analyser les tendances. Nous ferons l'approximation, dans le cadre de cet exercice, que la zone géographique de l'aéroport d'origine correspond au marché de vente du billet.

Le rendu de cette analyse est une **présentation Powerpoint en quelques slides** (*5-6 slides pour 10 minutes de présentation*) contenant vos graphiques et vos conclusions, et destinée au responsable du RMP. Il vous sera aussi demandé de nous transmettre le **script qui a été utilisé pour analyser ces données** (dans le langage de programmation de votre choix).

Dans un premier temps, analysez les tendances constatées sur les 3 périodes extraites, en mettant en avant les **évolutions significatives et pertinentes, selon vous, du comportement de voyage et de réservation depuis de début de la pandémie**. On s'intéressera notamment à l'évolution du nombre de billets vendus, du chiffre d'affaires généré et de leur répartition en fonction de différentes dimensions telles que le canal de vente, le profil du client (niveau de séniorité dans le programme de fidélité), le type de courrier, le type de voyage (aller retour/aller simple), etc. D'autres éléments de comparaison seront également bienvenus.

Quelles constatations et recommandations tirez-vous de Cette comparaison ? **Synthétisez les messages clés et choisissez une représentation graphique adéquate pour les illustrer.**

L'équipe du RMP vous demande d'analyser les tendances par marché de vente, en utilisant un découpage spécifique en 7 groupes : France, reste de l'Europe, Afrique, Amérique Nord, Amérique Sud, Caraibes & Océan Indien et Asie, et par niveau de contribution du client. Commentez votre analyse sur ces 2 dimensions. On pourra catégoriser les clients en trois niveaux de contribution (basse, moyenne et haute contribution) en fonction du chiffre d'affaire généré pendant la période, en donnant le label "basse contribution" aux clients dont le chiffre d'affaire total sur la période est dans le 1er quartile de la distribution du chiffre d'affaire par client, et "haute contribution" à ceux dans le 3ème quartile.

Dans un second temps, l'équipe Digital Marketing d'Air France affirme que le "lead time moyen" (différence entre la date de réservation et la date du vol) a chuté de manière significative pour les ventes online de nos clients "haute contribution" pendant la pandémie. Pouvez-vous vérifier cette affirmation en illustrant graphiquement ?

Exercice 2 - 🎰 Expérimentation statistique

Les deux parties suivantes sont indépendantes. Le format de restitution pour cet exercice est libre.

Première partie

On étudie la recette générée par l'achat d'un billet sur le vol du 3 janvier 2022 de la ligne Paris – Bangkok, et on compare 2 sous-populations d'acheteurs : les étudiants vs les actifs.

Pour cela, on extrait un échantillon de 15 billets achetés sur ce vol par des clients ayant déclaré être « actifs » et de 13 billets achetés sur le même vol par des clients ayant déclarés être « étudiants ». Le tableau ci-dessous donne la recette (en euros) de chaque billet vendu, trié par type de population. On suppose que la recette d'un billet suit une loi normale, d'espérance μ_e et de variance σ_e^2 pour les étudiants, et d'espérance μ_a et de variance σ_a^2 pour les actifs.

Actifs	**1150 1500 1700 1800 1800 1850 2200 2700 2900 3000 3100 3500 3900 4000 5400**
Etudiants	**300 700 850 900 1000 1420 1500 1500 1800 1800 1900 2200 2400**

1. Donnez un intervalle de confiance de μ_e et μ_a au seuil de 5%.
2. Votre manager vous demande de vérifier qu'en moyenne la recette d'un billet vendu sur ce vol ne dépend pas du statut étudiant vs actif. Construisez le test adéquat pour vérifier cette hypothèse, en supposant que les variances des 2 sous-populations sont égales (au niveau de confiance 95%). Détaillez les hypothèses et le test statistique avec les formules et les valeurs calculées sur cet échantillon. Que pouvez-vous conclure et répondre à votre manager ?
3. Vérifiez ces conclusions en utilisant la fonction adéquate sous `R/python` : reportez le résultat et commentez.
4. Construisez la même démarche sans supposer l'égalité des variances (au seuil de 5%). A nouveau, expliquez le test statistique réalisé, calculez la variance commune et le nombre de degrés de liberté associé à cette variance commune (détaillez les formules et valeurs sur cet échantillon). Confirmez ces résultats en utilisant la fonction adéquate sous R/Python et en commentant les résultats obtenus.
5. La recette générée par un billet Paris-Bangkok dépend-elle de manière significative du profil client : étudiant versus actif ? Comparez, sur cet exemple, les résultats des 2 tests.
6. `[Question facultative]` On souhaite maintenant comparer les variances des 2 sous-populations (σ_a^2 et σ_e^2) : donnez une estimation de chaque variance. Quelle loi appliquez-vous pour calculer un intervalle de confiance à 95% (détaillez les hypothèses qui vous permettent de l'appliquer).
7. `[Question facultative]` Peut-on considérer que les étudiants et les actifs ont la même variance (avec un niveau de confiance de 95%) – détaillez le test statistique associé ? Ainsi, parmi les 2 tests construits en question 2 à 4, lequel est adéquat ? Que peut-on dire de l'homogénéité de ces 2 sous-populations ? Comment l'expliqueriez-vous, intuitivement ?

Deuxième partie

Le département Marketing d'Air France souhaite choisir le vin qui sera proposé en cabine *Economy* à bord de nos vols, et organise pour cela une dégustation de 6 vins auprès de 12 clients voyageant sur un trajet Toulouse – New-York. A chaque dégustation, le client note sa préférence entre 0 (peu apprécié) et 10 (très apprécié). La dégustation est répétée dans 3 contextes différents : **(1)** dans la salle d'embarquement du vol Toulouse - Paris, **(2)** à bord du vol Toulouse – Paris et **(3)** à bord du vol Paris-New York. Ainsi, chaque client a évalué trois fois chacun des 6 vins (cf. tableau illustratif ci-dessous).

	Contexte (1)			Contexte (2)			Contexte (3)		
	vin 1	…	vin 6	vin 1	…	vin 6	vin 1	…	vin 6
Client 1	4.5		6	5.5		7	8		4.5
Client 2	7.5		8.5			8.5	6.5		5.5
…	…		…	…		…	…		…
Client 12	3.5	…	4	4.5	…	5	6	…	3.5

Dans un premier temps, on cherche les facteurs influents sur l'évaluation des vins.

1. Proposer un modèle pour analyser ces données
2. Que signifie concrètement l'interaction entre les facteurs type de vin et client ?

L'analyse décide de négliger a priori les interactions type de $vin.contexte$ et $client.contexte$. Le tableau ci-dessous donne les résultats de l'analyse de la variance pour le modèle utilisant tous les effets principaux et l'interaction type de $vin.client$.

3. On s'intéresse à l'interaction type de $vin.client$. Décrire le test de significativité de cette interaction (hypothèses, statistique du test, loi de cette statistique de test sous H_0). A l'aide du listage, conclure.
4. A l'aide du listage, quels effets choisissez-vous de retenir dans le modèle (ne pas décrire les tests).
5. Quel vin est le plus apprécié ? A bord de nos avions, utiliseriez-vous seulement ce vin ?
6. Interprétez, concrètement le coefficient relatif au client n°2 et le coefficient relatif au client n°5

```
>AovSum(0.Appreciation~Vin+Client+Contexte+Client:Vin, data=Vins)
```

$Ftest

	SS	df	MS	F value	Pr(>F)	
Vin	604.09	5	120.819	115.1583	< 2.2e-16	***
Client	66.89	11	6.081	5.7957	1.056e-07	***
Contexte	2.64	2	1.320	1.2582	0.287313	
Client:Vin	99.26	55	1.805	1.7201	0.005743	**
Residuals	148.98	142	1.049			

$Ttest

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.3921	0.0697	77.3691	0.0000
Vin - 1	1.5329	0.1558	9.8362	0.0000
Vin - 2	0.7301	0.1558	4.6849	0.0000
Vin - 3	-3.6255	0.1558	-23.2641	0.0000
Vin - 4	0.7329	0.1558	4.7027	0.0000
Vin - 5	0.4106	0.1558	2.6351	0.0093
Vin - 6	0.2190	0.1558	1.4052	0.1622
Client - 1	-0.4255	0.2311	-1.8407	0.0678
Client - 2	0.0690	0.2311	0.2984	0.7658
Client - 3	-0.6644	0.2311	-2.8741	0.0047
Client - 4	-0.2810	0.2311	-1.2158	0.2261
Client - 5	-0.7255	0.2311	-3.1385	0.0021
Client - 6	1.2968	0.2311	5.6101	0.0000
Client - 7	0.4801	0.2311	2.0770	0.0396
Client - 8	-0.0644	0.2311	-0.2784	0.7811
Client - 9	0.4801	0.2311	2.0770	0.0396
Client - 10	-0.4810	0.2311	-2.0810	0.0392
Client - 11	-0.0088	0.2311	-0.0381	0.9697
Client - 12	0.3245	0.2311	1.4040	0.1625
Contexte - 1	0.0384	0.0986	0.3899	0.6972
Contexte - 2	-0.1505	0.0986	-1.5266	0.1291
Contexte - 3	0.1120	0.0986	1.1367	0.2576
Client - 1 : Vin - 1	-0.3995	0.5169	-0.7730	0.4408
Client - 2 : Vin - 1	-0.3273	0.5169	-0.6333	0.5276
...				
Client - 12 : Vin - 6	-0.3023	0.5169	-2.5197	0.0129

Exercice 3- 💻 Modélisation et apprentissage supervisé

Le **programme Flying Blue d'Air France** est un programme de fidélité qui propose des avantages (par ex. l'embarquement prioritaire) aux clients fidélisés. En voyageant sur Air France, les clients ayant souscrits à ce programme cumulent des points de fidélité (aussi appelés *miles*) et peuvent dépenser leurs points de fidélité sur un achat de billet ou sur l'ensemble des partenaires du programme. Chaque client est associé à un **statut** dans le programme, ou "tier level" en anglais : (par ordre de fidélité croissante) *Explorer, Silver, Gold et Platinum*. A chaque statut est associé un ensemble d'avantages, qui est d'autant plus attractif que le statut est élevé. Au fur et à mesure de ses voyages sur Air France et sur l'ensemble des compagnies aériennes du programme Flying Blue, le client cumule des points, et son statut évolue (à la hausse ou à la baisse).

L'équipe Marketing & Communication du programme de fidélité Flying Blue vous sollicite en préparation d'une campagne. Ils souhaiteraient tester une offre incitative pour **accélérer un changement de statut de leurs clients vers des niveaux plus élevés**. Pour cela, ils vous demandent de les aider à cibler les clients dont on prédit un changement de statut prochainement vers les statuts *Gold* ou *Platinum*, à partir des informations de profil et de l'historique d'achat du client.

Le rendu de cet exercice est un **notebook** que vous présenterez le jour de votre entretien.

Le fichier **classification_data.csv** contient au grain client différentes données agrégées qui vous sont décrites en annexe. Elles contiennent l'historique d'achat et de vol des clients ayant un statut Explorer & Silver sur la période de 2016 & 2017, ainsi qu'une colonne qui indique si ces clients ont changé de statut vers "Gold" ou "Platinum" en 2018 & 2019.

Précisez quelle méthode vous choisissez pour sélectionner les features les plus pertinentes. Quelles autres features auriez-vous pu intégrer au modèle si vous aviez accès aux données ? Après avoir sélectionné les features qui vous paraissent les plus pertinentes, construisez un modèle de classification qui permet de prédire le champ `tierStatus2Cl`. Il vous est demandé d'évaluer les performances de votre modèle, de justifier votre choix et de le comparer à un second modèle de classification.

Quel est le profil type des clients qui vont bientôt passer au statut *Gold* ?

Comment proposez-vous d'évaluer la recette incrémentale de la campagne Flying Blue et plus précisément de son ciblage ?

Annexes

Table flight_data_extract.csv

Champ	Description
FLIGHT_DATE	Date de vol
TKT_NUM	Numéro de billet
ACTIVITY_TIER	Tier (niveau de fidélité au programme FB, A : Explorer, B : Silver, R : Gold, M : Platinum)
OPE_CRR	Compagnie aérienne opératrice
ONLINE_FLAG	True si billet acheté sur un canal de vente online
OW_FLAG	True si le billet correspond à un One-Way (Aller simple), False si aller-retour ou autre parcours plus complexe (Open Jaw, ...)
OND_ORI	Origine
NET_REV	Revenu
HAUL	Courrier (SH pour Short Haul, MH Medium Haul, LH Long Haul)
DIS_CHA	Canal de vente
BPA	Cabine (EC: Economy, BU: Business)
PNR_DAT_AMD	Date de booking

Table area.csv

Champ	Description 1	Description 2
COD_MCR_ZON	Departure IATA Macro Zone (Code)	Scheduled departure macro-zone code. The world is divided into 4 macro-zones. Ex: FCE for France.
LIB_MCR_ZON_ANG	Departure IATA Macro Zone (Label)	Scheduled departure macro-zone label.
COD_ZON_IAT	Departure IATA Zone (Code)	Scheduled departure IATA zone code. The world is divided into 8 IATA zones. Ex : NOAM for north America.
LIB_ZON_IAT_ANG	Departure IATA Zone (Label)	Scheduled departure IATA zone label.
COD_RGN_IATA	Departure IATA Region (Code)	Scheduled departure IATA region code. The world is divided into 14 IATA regions. Ex: ASE for the south-east of Asia
LIB_RGN_IAT_ANG	Departure IATA Region (Label)	Scheduled departure IATA area label.
COD_SUB_RGN_IAT	Departure IATA Sub-Region (Code)	Scheduled departure IATA sub-region code. The world is divided into 18 sub-regions. Ex: AF3 for Western Africa
LIB_SUB_RGN_ANG	Departure IATA Sub-Region (Label)	Scheduled departure IATA sub-area label.
COD_PAY	Departure Country (Code)	Scheduled departure country code. Ex : SE for Sweden.
LIB_PAY_ANG	Departure Country (Label)	Scheduled departure country label. Ex: Sweden.
CPN_ORI_PLN_CIT	Departure City (Code)	Scheduled departure city code. Ex : PAR for Paris
LIB_VIL_IAT_ANG	Departure City (Label)	Scheduled departure city label. Ex: Paris
CPN_ORI_PLN_ARP	Departure Airport (Code)	Planned departure airport code Ex : CDG (Charles de Gaulle)

Table classification_data.csv

Champ	Description
age	Age du client
seniority	Nombre d'années depuis l'adhésion au programme Flying Blue
country	Regroupement géographique selon pays de résidence (IDF si client habitant en Ile de France, FR reste de la France)
isCorporate	1 si le client a un contrat corporate avec Air France, 0 sinon
diffLastActivity	Nb de jours écoulés entre le dernier achat du client et le 31 Decembre 2017
nbTickets	Nb de billets achetés en 2016 et 2017
revenueLH	Chiffre d'affaire généré sur le long courrier en 2016 et 2017
revenue2016	Chiffre d'affaire en 2016
revenue2017	Chiffre d'affaire en 2017
nbOptions	Nombre d'options payantes achetées en 2016 et 2017 (bagage supplémentaire, repas à bord, siège payant)
shareBag	Ratio du nombre d'option payante de type bagage sur le nombre total d'options payantes achetées en 2016 et 2017
shareMeal	Ratio du nombre d'option payante de type repas sur le nombre total d'options payantes achetées en 2016 et 2017
shareSeat	Ratio du nombre d'option payante de type siège payant sur le nombre total d'options payantes achetées en 2016 et 2017
timeToGold	Si le client est passé Gold après le 31/12/2017, nombre de jours entre le 31/12/2017 et son passage au statut Gold
dateToGold	Si le client est passé Gold après le 31/12/2017, date de passage au statut Gold
tierStatus2Cl	(Variable à prédire) Statut du client après le 31/12/2017 (est-il passé au statut gold ou est-il resté explorer ?)