Projet Tableau

M2 IMSD 2018/2019

Hyesu KIM & Aboudalla DIABATE

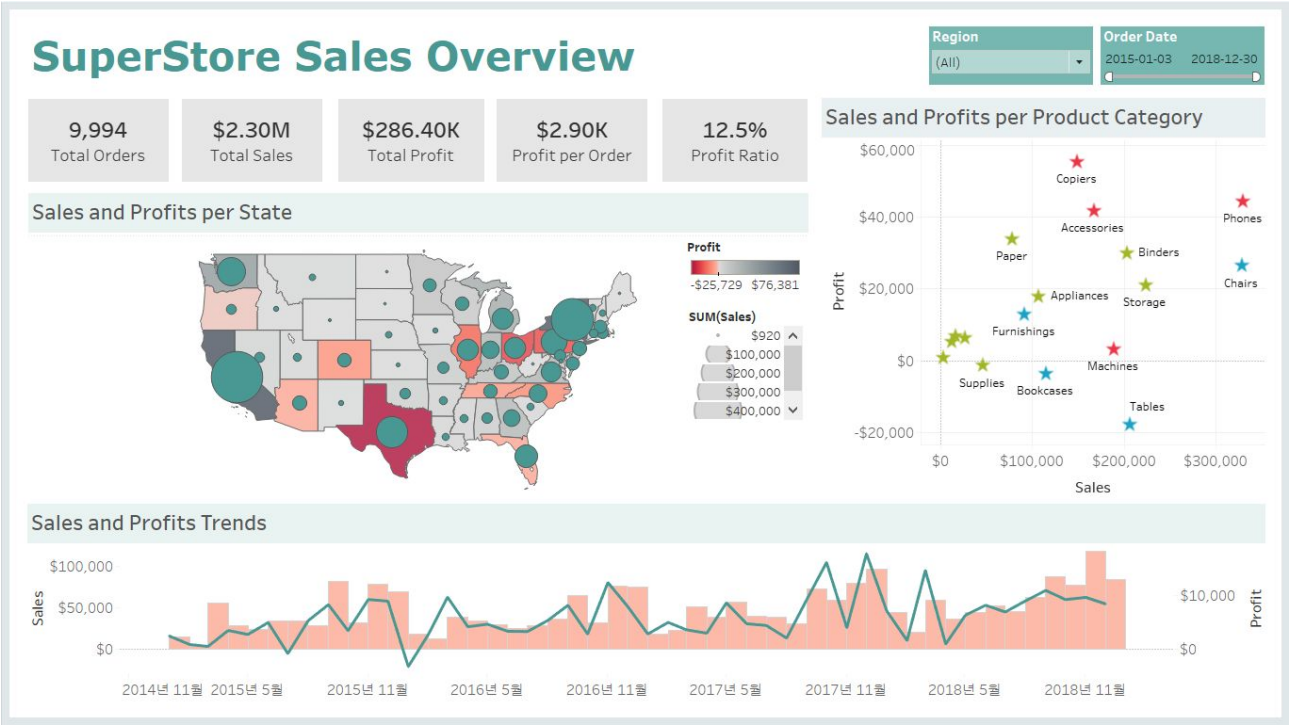
Notre projet Tableau est structuré en deux parties. La première est plus simple et basique à réaliser la visualisation à partir de la base des données sur Tableau “SuperStore”. La deuxième est plus complète et plus créative c’est-à-dire nous avons été amené à trouver nous même une base de données, à la traiter pour l’étape visualisation et à l’analyser.

1. Première partie

Nous avons créé deux feuilles de tableaux de bord : une vue d’ensemble et une analyse produits. Dans la première feuille, nous présenterons une synthèse de la base des données, particulièrement sur la performance de ventes et profits. La deuxième est dédiée à l’analyse sur les produits en analysant la précision de livraison et la performance vente de chaque produit par segment et catégorie.

1.1 Visualization des données

1.1.1 Vue d’ensemble



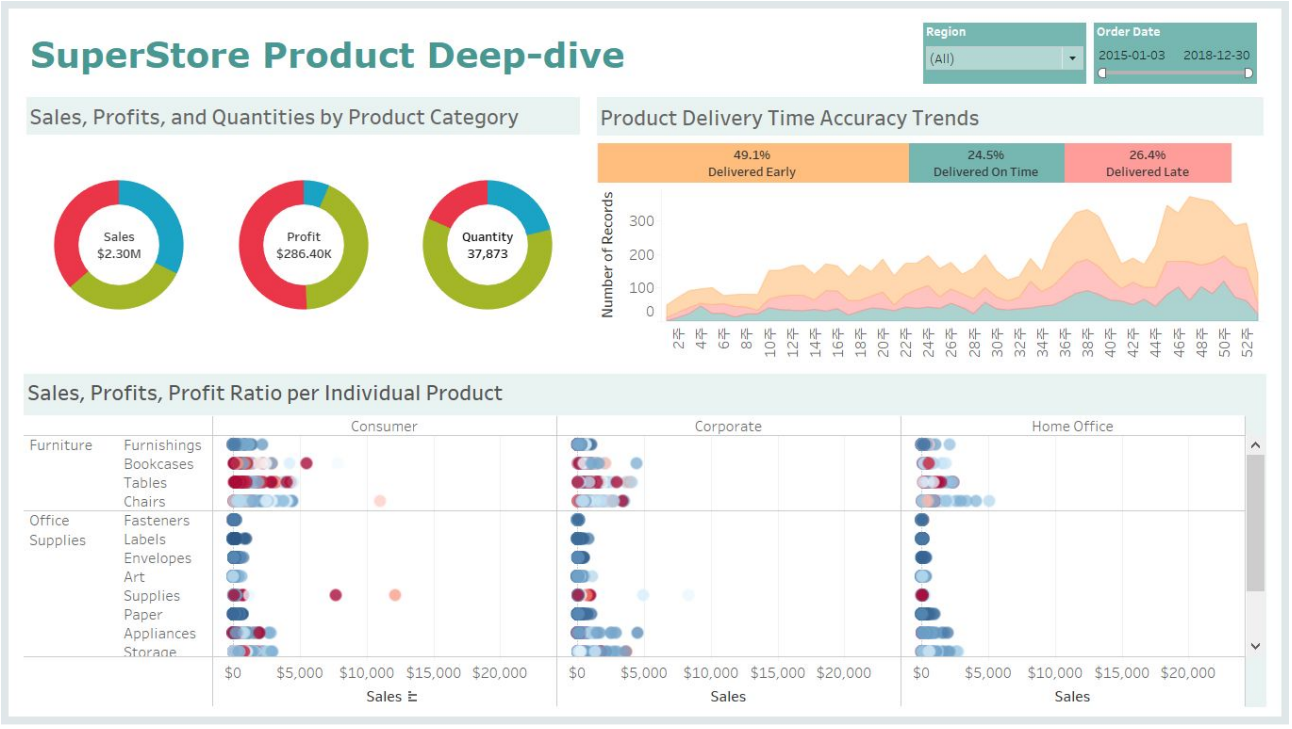
Ce tableau de bord est constituée de 6 parties avec un titre, deux filtres et cinq boîtes de KPIs. En utilisant les filtres, les chiffres et les tableaux vont dynamiquement changer. S’ils ne sont pas précisés, le tableaux de bord sera mis pour tous les régions et toutes les dates par défaut.

Cette carte géographique “Sales and Profits per State” nous montre la somme de ventes et profits par l’Etat. Les parties en couleur indiquent l’importance de gains, la plus rouge signifie le moins de gains. La taille de rond représente l’importance de ventes. L’Etat qui a le meilleur résultat de ventes et de profits est Californie. Au contraire, le Dakota du Nord est l’Etat qui a le plus mauvais résultat de ventes et de profits.

La diagramme de dispersion (scatter plot) “Sales and Profits per Product Category” est une tableau qui explique la distribution de chaque catégorie et sub-catégorie à travers la somme de ventes et de protis. La couleur d’étoiles correspond à chaque catégorie, Furniture (blue), Office Supplies (vert) et Technology (rouge). Globalement, la catégorie Technology est la catégorie qui a vendu le plus et a gagné le plus, spécialement la sub-catégorie Copier. Les catégories Office Supplies et Furniture sont positionnées à l’opposé de la catégorie Technology dans le graphe, vend le moins et gagne moins. La sub-catégorie Table est celle qui a le plus faible profit.

Dans ce graphique barre et ligne combinée “Sales and Profits Trends”, les barres représentent la somme des ventes par mois et celle de ligne est pour la somme des profits par mois. Nous remarquons une augmentation de la tendance sur les ventes et les gains. La performance de ventes dans le quatrième trimestre est plus forte que les autres trimestres.

1.1.2 Analyse de produits



Ce tableau de bord est constitué de 5 parties incluant le titre et les deux filtres. Avec les mêmes filtres que ceux du premier tableau de bord, on peut voir les chiffres et les graphiques qui correspondent à chaque région et à une date particulière.

Dans ce graphique camembert “Sales, Profits, and Quantities by Product Category”, chaque graph camembert signifie le pourcentage de catégorie pour trois KPIs, ventes, gains, et quantités. Les couleurs montrent les catégories : Furniture (blue), Office Supplies (vert) et Technology (rouge). Les trois catégories sont quasiment également vendu par rapport la somme de ventes. La catégorie

Technologies occupe plus de 50 % de la somme de profits alors que la catégorie Furniture ne se représente que 6.4 %. La catégorie Office Supplies constitue la majorité de la somme de quantité. Concernant le prix de chaque produit, il nous semble qu'il est normal de les observer.

Pour le graphique Area "Product Delivery Time Accuracy Trends", nous avons créé deux variables supplémentaires, 'Delivery Actual' pour compter les délais de livraison actuelle et 'Delivery Scheduled' pour compter les délais de livraison prévu. Nous avons aussi créé une variable catégorielle, 'Delivery Status'. Cette variable signifie si le produit est bien arrivé comme prévu en comparant les délais de 'Delivery Actual' et 'Delivery Scheduled'. De plus ce graphique est utilisé comme un filtre pour le graphique suivant, "Sales, Profits, Profit Ratio per Individual Product".

Avec ce graphique, on peut voir la tendance des états de livraison de produit pendant un an. Le nombre de commande a augmenté au fil du temps. Plus de 70% des commandes ont été livrées à l'heure ou en avance. La quantité de ces dernières a beaucoup augmenté à la fin d'année mais les livraisons en retard ont aussi augmenté.

Dan ce graphique barre et figure combinées "Sales, Profits, Profit Ratio per Individual Product", nous avons croisé trois variables continues, la somme de ventes, de profits, et de taux de profits, à cinq variables catégorielles, le segment, la catégorie, la sub-catégorie et l'état de livraison et le nom de produit. Grâce à cette tableau, on peut observer toutes les informations utiles comme la somme de ventes, de gains, et de taux de gains et l'état de livraison par chaque produit individu.

2. Deuxième partie

2.1 Base de données

Dans notre projet, nous avons décidé d'analyser les éléments qui expliquent le plus la satisfaction et la non satisfaction des clients à l'égard des compagnies aériennes avec lesquelles ils effectuent leurs voyages. Pour cela, nous avons utilisé une base de données provenant d'une enquête sur le site de IBM WATSON. Cette base de données représente les réponses aux questions de certains clients voyageant à l'intérieur des Etats-Unis avec les principales compagnies aériennes américaines.

Notre base de données brutes est composée de 129.889 lignes qui représentent chacune un client et de 30 variables. Après l'analyse de données de la base brute au cours de laquelle, nous avons supprimé certaines valeurs manquantes, créer des variables et supprimer d'autres, notre base de données corrigées contient finalement 129.549 observations et 31 variables suivantes :

- Satisfaction : le niveau de satisfaction du client. Il est compris en 1 et 5. Le niveau 1 est le niveau de satisfaction le plus faible et 5 le niveau le plus élevé.
- Membership_status : Programme de fidélité des compagnies aériennes. Nous en avons 4 : 'Blue', 'Silver', 'Gold', et 'Platinum'. Le blue est l'état le plus base.
- Age : l'âge du client.
- Age_range : La tranche d'âge à laquelle le client appartient
- Gender : Homme ou Femme
- Price_sensitivity : le niveau auquel le prix affecte l'achat du billet par le client. Il est compris entre 0 et 5.
- no_of_flights_p,a, : Le nombre de vols que le client a pris par an. Ce nombre est compris entre 0 et 100.

- no_of_flights_p,a,_grouped : la tranche du nombre de vols pris par an.
- travel_type : Motif du voyage qui sont 'Business travel' (voyage d'affaire), 'Mileage tickets' (voyage fait avec les mileage accumulé) et 'Personal Travel' (Voyage personnel ou voyage en famille ou vacance).
- shopping_amount_at_airport : Montant des achats du clients à l'aéroport de départ.
- eating_and_drinking_at_airport : Ses dépense en nourritures et boissons à l'aéroport de départ.
- cabin_class : La classe de la cabine. Elle peut être Business, Eco ou Eco Plus.
- flight_date : La date du voyage. Tous les voyages sont effectués en janvier, février ou mars 2014.
- airline_name : Le nom de la compagnie aérienne. Il y a 14 différentes compagnies aériennes.
- departure_city : Ville de départ de l'avion.
- departure_state : Etat de départ de l'avion.
- destination_city : Ville de destination de l'avion.
- destination_state : Etat de destination de l'avion.
- scheduled_departure_hour : Les heures auxquelles les client sont programmés à prendre le vol. Les heures sont comprises entre entre 1h du matin et 24h.
- flight_time_in_mins : cette variable indique le temps du vol en minutes.
- flight_distance_in_miles : La distance en mile entre 2 le lieu de départ du vol et son lieu de destination.
- total_delay_in_mins : Combien de minutes au total le vol fut-il en retard à la fois à son départ et à son arrivée.
- flight_status : Statut du vol. Soit le vol est arrivé à l'heure, soit il a été en retard, soit il a été supprimé.
- country : Le pays de départ et de destination des vols. Comme préciser plus haut, l'enquête a été réalisé auprès des clients qui effectué leur voyage à l'intérieur des Etats-Unis. Mais nous avons ajouté cette variable afin de pouvoir faire un graphe de géolocalisation.

2.2 Traitement des données

Avant de commencer l'étape data visualisation, nous avons traité la base de données sur Python. Cette étape constitue 4 parties, traitement de valeurs manquantes, traitement de valeurs aberrantes, conversion de variables, ainsi que la création de certaines variables et la suppression d'autres dont on a pas besoin dans notre analyse. Les détails sont trouvés dans ci-joint le fichier de Jupyter Notebook, '[Tableau project] Part 2.1 Data Processing.ipynb'.

2.2.1 Traitement de valeurs manquantes

Pendant l'exploration de la base données originales, nous avons observé les valeurs manquantes dans les variables suivantes :

- 'departure_delay_in_mins' : 1.8% (2345/129889)
- 'arrival_delay_in_mins' : 2.1% (2738/129889)
- 'flight_time_in_mins' : 2.1% (2738/129889)

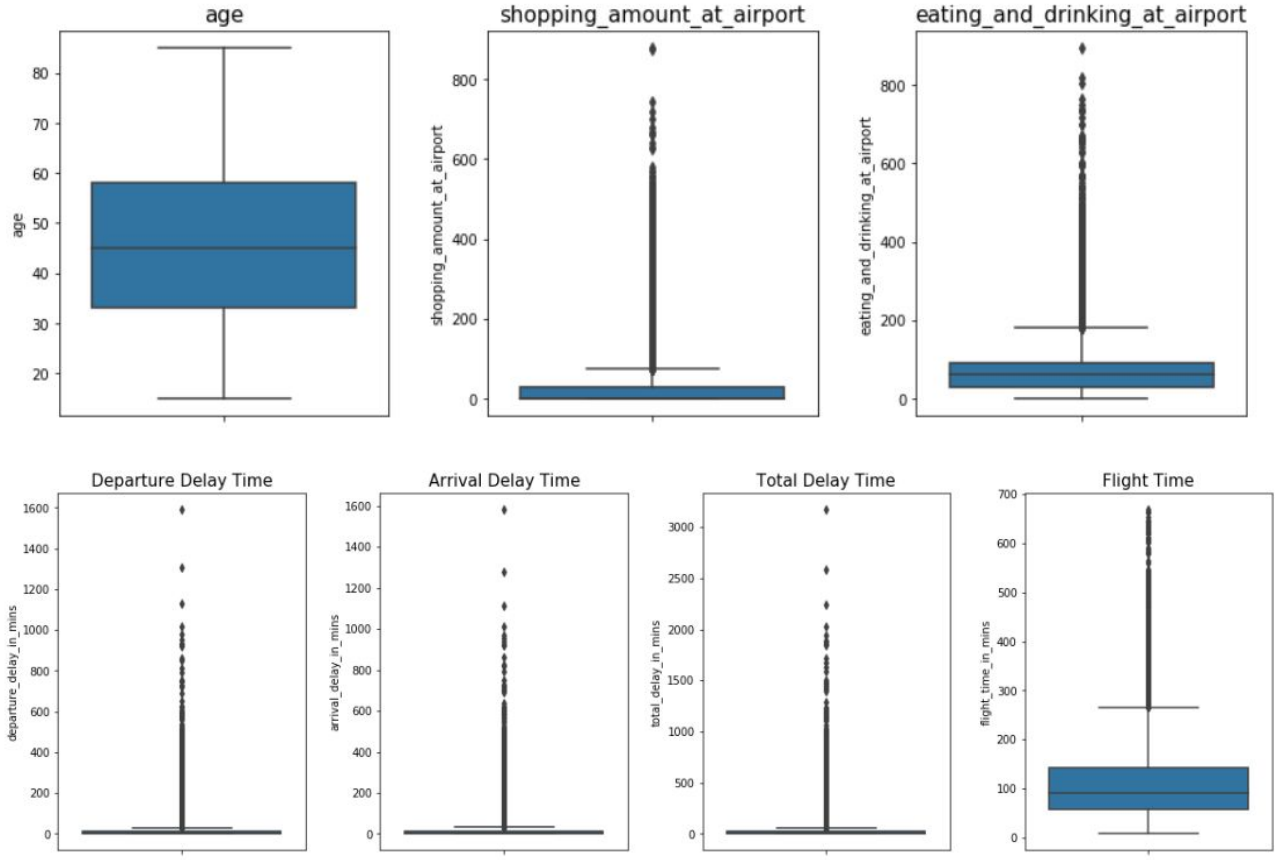
Après avoir étudié les autres variables, nous nous sommes rendu compte qu'il s'agit des vols qui ont été annulés (cela représente 2401 observations). Ainsi, nous avons trouvé légitime de ne pas trouver de valeurs pour ces lignes correspondant à ces variables. De plus, nous avons constaté que les 2345 observations dont les valeurs manquantes dans tous ces trois variables sont des vols qui ont été annulés avant l'heure départ. Les valeurs manquantes dans les dernières variables (56 observations)

ont été annulés après l'heure départ. C'est la raison pour laquelle les valeurs dans la variable 'departure_delay_in_mins' sont présentes car les clients ont attendu.
Comme la présence de ces valeurs manquantes est normal et il n'y a pas une autre valeur applicable pour les remplacer, nous avons donc décidé de les laisser vides. Dans le tableau, ces valeurs ne vont pas être comptées pour le calcul.

Il reste encore 337 valeurs manquantes des variables 'arrival_delay_in_mins' et 'flight_time_in_mins' mais ils ne sont pas des vols annulés et leurs valeurs dans la variable 'departure_delay_mins' sont existantes. Nous avons par conséquent décidé de les supprimer (car elles ne représente que 0.26% des observations totale).

2.2.2 Traitement de valeurs aberrantes

Comme nous observons dans le boxplot des variables continues, la plupart ont certaines valeurs aberrantes.



On pourrait les traiter mais on a décidé de les conserver, car il semble qu'ils se soient produites réellement (et non introduites par erreur). Par exemple, la variable 'departure_delay_in_mins' est exprimée en minutes, le temps le plus long (1592 minutes) correspond donc à environ un jour. Nous pensons que ce genre de retard des vols bien qu'il soit rare peut exister dans une situation exceptionnelle telle que un signal d'alertes ou une menace terroriste par exemple.

2.2.3 Conversion variables

Comme il y avait deux variables qui ont eu des types inappropriés, nous les avons transformé en type appropriés :

- ‘satisfaction’ : type ‘string’ → type ‘float64’
- ‘flight_date’ : type ‘string’ → type ‘date’

2.2.4 Création de variables

Afin de rendre notre analyse plus simple et la visualisation de nos graphes, nous avons créé 3 variables, ‘total_delay_in_mins’, ‘flight_status’, et ‘country’.

En sommant les variables ‘departure_delay_in_mins’ et ‘arrival_delay_in_mins’, nous avons créé une variable continue, ‘total_delay_in_mins’ qui peut être une variable expliquant le niveau de ‘satisfaction’ des clients.

Avec cette nouvelle variable, nous avons créé une variable de catégorie, qui indique l’état du vol : ‘on-time’, ‘delayed’ ou ‘cancelled’. Puis on a supprimé la variable ‘flight_cancelled’ qui contient le même contenue (Yes/No).

En faisant la visualisation, nous avons remarqué que sans l’information de pays, on ne peut pas avoir la carte géographique. Par conséquent nous avons créé une variable de catégorie qui prend une seule modalité, ‘United States’.

2.3 Visualization des données

Nous avons à présent une base de données corrigées nous permettant de procéder à leur visualisation et de faire une analyse sur la satisfaction des clients des compagnies aériennes présentes dans notre donnée.

Nous ferons notre analyse en quatre parties qui correspondront chacune à un tableau de bord. Dans la première partie, nous présenterons une vue d’ensemble de nos données. Ensuite, dans la seconde partie nous croiserons les variables liées aux retards des vols à d’autres variables telles que la destination des vols, la variable du temps afin de voir si ces retards diffèrent d’une destination à une autre ou d’une période à une autre. Dans la troisième partie, nous ferons une analyse client en déterminant les typologies de clientèles et les programmes de fidélité qu’ils détiennent dans les compagnies aériennes. Enfin, nous verrons les variables qui influencent le plus la satisfaction des clients.

2.3.1 Vue d'ensemble

Ce tableau de bord est constituée de 6 parties avec un titre, deux filtres et six boîtes de KPIs. Comme la première partie, les chiffres et les tableaux vont dynamiquement changer par les deux filtre. S'ils ne sont pas précisés, le tableaux de bord sera mis pour tous les compagnies aériennes et toutes les dates par défaut.

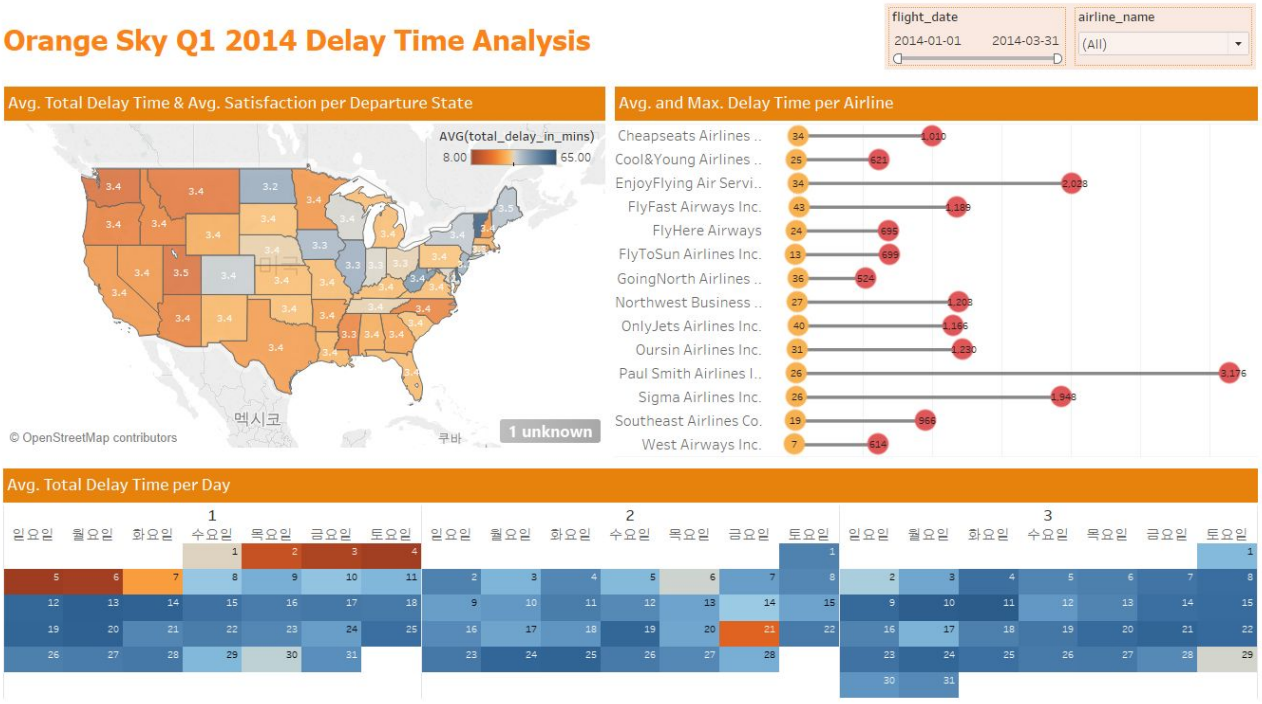
Ce tableau de bord nous présente, en moyenne par jour, 1.439 passagers ont pris leurs vols avec les 14 compagnies aériennes. Les réponses données par les clients qui ont été interrogés ont révélé que 44% des vols ont respecté leurs horaires de départ et d'arrivée contre 54% qui ont été retardés et 2% de ces vols ont été annulés. Donc plus de la moitié des vols ont été en retard d'environ 30 minutes en moyenne, ce qui pourrait peut être avoir une conséquence sur le niveau de satisfaction des clients qui est en moyenne de 3.38 sur 5.

Le graphique "Avg. Departure and Arrivals Delay Time Series" nous montre les retards moyens au départ et à l'arrivée des vols au cours du temps. On remarque que les deux courbes ont la même allure, ce qui nous paraît logique car on penserait, de façon générale, qu'un vol qui prend du retard à son départ arrivera forcément à sa destination plus ou moins proportionnellement au temps de retard pris au départ.

Au début du mois de janvier, les deux courbes observent un pic de retard des vols d'environ de 42 minutes en moyenne avant de baisser jusqu'à mi-janvier à environ 8 min en moyenne. Ensuite, ces temps de retard augmentent vers fin janvier et reste à environ une vingtaine de minutes jusqu'à mis février. Les vols observent encore un autre pic de retard de 36 minutes après mi-février et reste à une vingtaine de minutes jusqu'à fin mars. Mais globalement au cours de la période janvier-mars les retards combinés aux départs et arrivées subis par les vols étaient d'environ une trentaine de minutes.

Le graphique "No. of flights ans Avg. satisfaction rate per airline" est le niveau moyen de satisfaction des clients et le nombre de vols pour chaque compagnie aérienne. Il tourne autour de 3.4 sur 5. Cheapseats Airlines Inc. est la compagnie qui a plus de voyageurs dans notre base de données soit environ 25.000 voyageurs.

2.3.2 Analyse des retards des vols



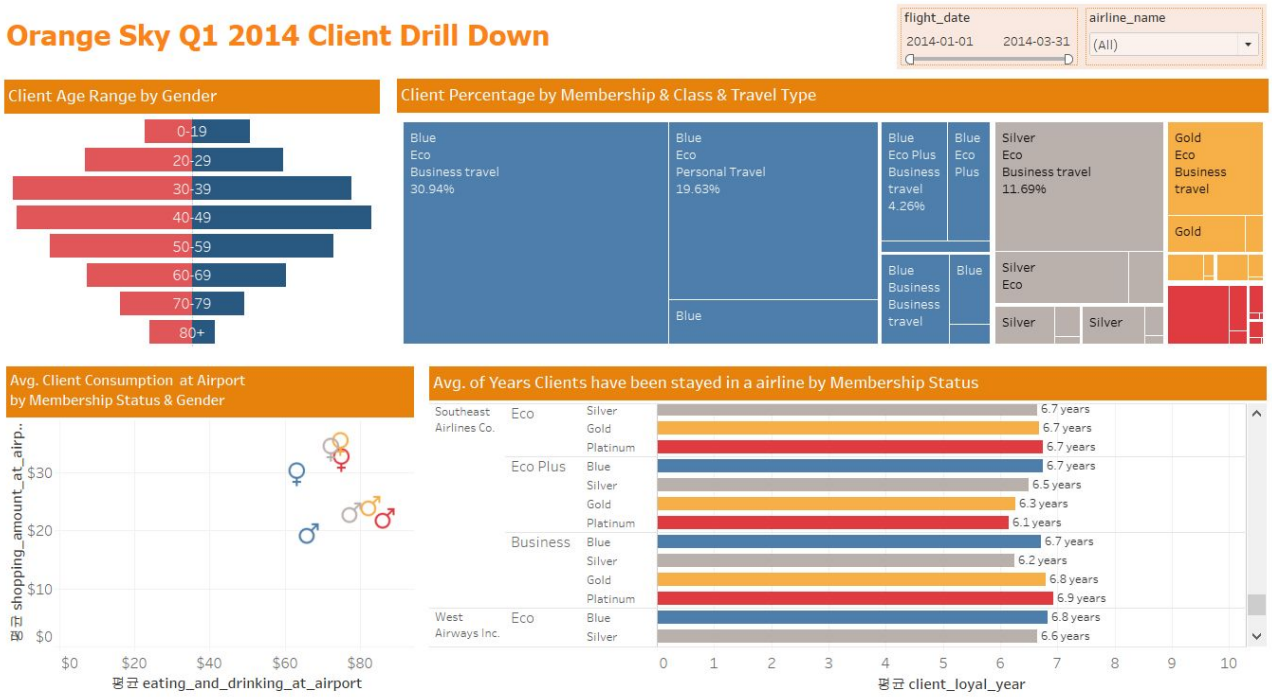
Ce tableau de bord est constitué de 5 parties incluant le titre et les deux filtres. Avec les mêmes filtres que ceux du premier tableau de bord, on peut voir les chiffres et les graphiques qui correspondent à chaque compagnie aérienne et à une date particulière.

Cette carte géographique "Avg. Total Delay Time & Avg. Satisfaction per Departure State" nous montre la durée moyenne de retard et le niveau moyen de satisfaction par l'Etat. Les vols au départ des Etats de Vermont, Delaware, West Virginia observent un retard moyen plus élevé au départ et à l'arrivée soit plus de 50 minutes.

Le graphique "Avg and max Delay Time per airline" nous montre la durée moyenne de retard et celle de retard maximum de chaque compagnie aérienne. En visualisant l'écart de ces deux nombres, on peut facilement remarquer laquelle des compagnies aériennes a la durée de retard la plus longue. Paul Smith Airlines Inc. note le plus grand écart entre deux nombres, 26 min en moyen et 3176 min au maximum. A l'inverse, West Airways Inc observe le plus petit écart entre la moyenne et la maximum de la durée de retard, 7 min en moyen et 614 min au maximum.

Avec la figure heat map calendar "Avg. Total Delay Time per Day", on souhaitait savoir si la durée moyenne de retard changé d'une période à une autre. Si on observe les retards de vols pour toutes les dates de voyage dans la figure, la période du jeudi 2 janvier au lundi 6 janvier sont les dates qui ont connues les pics de retard de vols avec plus d'une heure de retard en moyenne. On peut voir cela comme la cause de long weekend pris après les vacances de noël et du nouvel an.

2.3.3 Analyse client



Ce tableau de bord est constitué de 6 parties incluant le titre et les deux filtres.

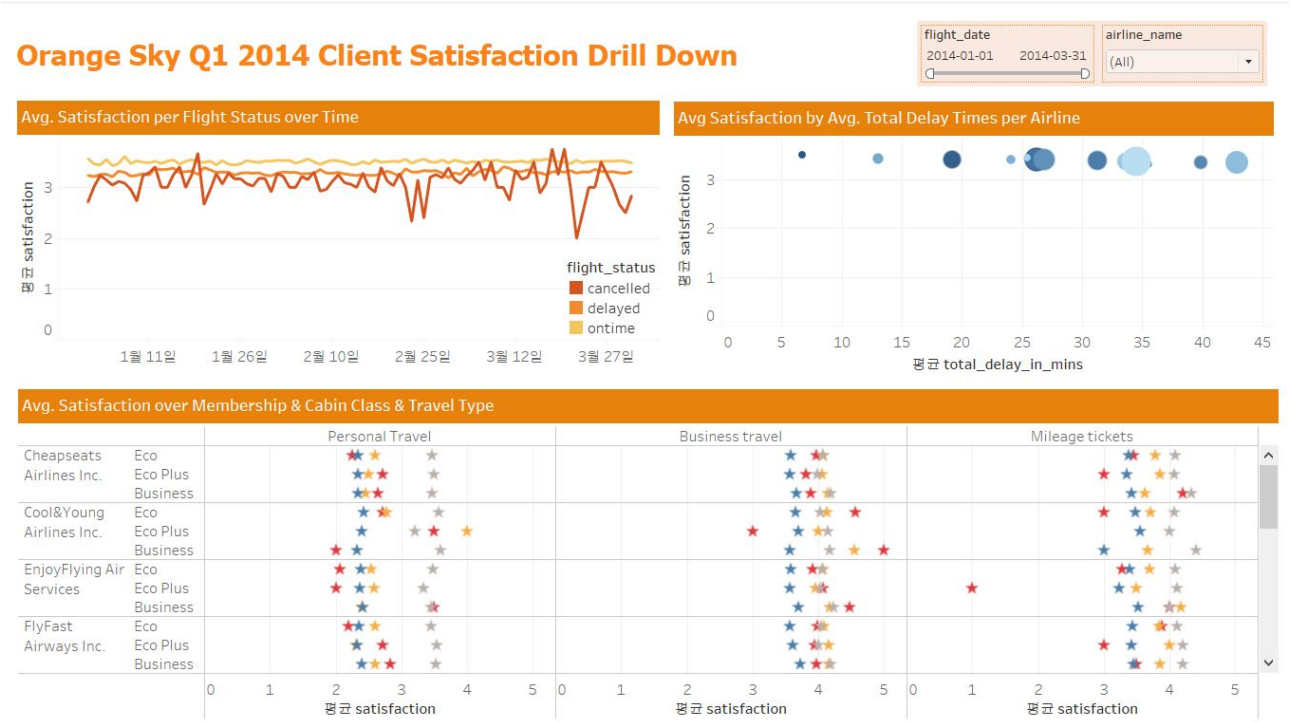
La première figure butterfly chart, “Client Age Range by Gender” nous montre la distribution de la tranche d'âge des client par sexe ainsi que leur niveau moyen de satisfaction. La moyenne d’âge des voyageurs est d’environ 46 ans. Ils sont composés majoritairement des personnes âgées de 30 à 39 et 40 à 49 ans dans les deux sexes (plus de 40%). De plus, ces groupes ont le niveau moyen de satisfaction le plus élevé (3,7/5). Or les seniors ont un niveau moyen de satisfaction le plus faible (l’intérieur de 3) et ils ne sont pas nombreux. On observe donc la relation négative entre la tranche d’âge et la note de satisfaction.

Ensuite, dans le graphique treemap “Client Percentage by Membership & Cabin Class & Travel Type”, nous pouvons voir la proportion de tous les groupes de clients à travers la fidélité, la cabine, et le type de voyage. Environ 31% des clients ont des cartes de fidélité Blue et font des voyages d’affaires en cabine “Eco” et seulement 4,26% des clients occupent des places “Eco Plus”.

Avec le diagramme de dispersion (scatter plot) “Avg. Client Consumption at Airport by Membership Status & Gender”, on souhaite étudier le comportement du consommateur. Les couleurs montrent le programme de fidélité : Blue (blue), Sliver (gris), Gold (jaune), et Platinum (rouge). Selon le graphe, les homme dépensent moins en shopping que les femmes mais le montant de dépenses en nourritures de deux sexes sont similaires.

Le graphique “Avg. of years clients have been stayed in airline by membership status” nous montre le nombre d’année moyen de fidélité des clients à leur compagnie aérienne. Globalement, les clients ont des cartes de fidélité en moyenne de 6 à 7 ans.

2.3.4 Analyse de la satisfaction client



Ce tableau de bord est constitué de 5 parties incluant le titre et les deux filtres.

Le graphique “Avg. satisfaction per flight status over time” nous montre bien que l’état du vol a des conséquence sur le niveau de satisfaction des clients. Les clients qui ont vu leurs vols annulés sont moins satisfaits que ceux ayant vu leur vols en retard ou respecter les horaires.

Dan ce graphique barre et figure combinées “Avg. satisfaction over membership and cabin class and travel type”, nous avons croisé trois variables catégorielles, la fidélité, la cabine, et le type de voyage à la variable satisfaction. Globalement, pour tous les motifs de voyages, les clients ayant les types de fidélité “Silver” montrent plus de satisfaction que les clients ayant les autres types.

3. Conclusion

En conclusion, nous pouvons constater que bien que les toutes les compagnies aériennes dans nos données aient à peu près le même niveau moyen de satisfaction des clients, nous retenons par ailleurs que le status d’un vol c’est-à-dire qu’un vol soit à l’heure, en retard ou annulé ainsi que le programme de fidélité influencent ce niveau de satisfaction.