Ecole Nationale Supérieure de Statistiques et d'Economie Appliquée ENSSEA

Examen Python S1 (2021/2022)

Présentation du problème:

Dans ce problème on essaye d'analyser les vols pour essayer de déterminer les facteurs influants sur les retard de vols de départ et celles d'arrivée.

Pour cela on va analysé une base de données en provenance de US Departement Of Transportation (DOT) et sont collectées par le Bureau Of Transportation Statistics (BTS). . La base des données contient des données sur tous les vols quittant les grands aéroports des Etats-Unis en une certaine année.

La base de données originale contient plus de 5 millions vols. Un échantillon seulement sera analysé dans ce problème.

Les variables de cette base sont :

YEAR : représente l'année du vol.

MONTH : valeur numérique de 1 à 12 représentant le mois du vol.

DAY : valeur numérique de 1 à 31 représentant le jour du vol.

DAY_OF_WEEK : valuer numérique de 1 à 7 représenant le jour du semain du vol. (la valeur 1 représente le Lundi).

AIRLINE : la compagnie aérienne responsable du vol.

FLIGHT_NUMBER : identifiant qui représente le numéro du vol.

TAIL_NUMBER : Identifiant qui représente l'avion.

ORIGIN_AIRPORT : L'aéroport de départ.

DESTINATION_AIRPORT : L'aéroport d'arrivée.

SCHEDULED_DEPARTURE : Le temps de départ anticipé. Le temps est formaté en 4 chiffres (par exemple le temps est est 10h45 alors cette variable prend la valeur 1045).

DEPARTURE_TIME : Le temps de départ réel (même format que le temps de départ anticipé).

DEPARTURE_DELAY : Le retard du départ (le retard peut prends des valeurs négatives si le départ réel est avant le départ anticipé).

TAXI_OUT : Le roulage de départ, représente le temps entre la sortie de l'avion et le décollage réel de l'avion.

WHEELS_OFF : Le temps de décollage réel de l'avion (même format que les variables représentant le temps).

SCHEDULED_TIME : La durée de vol anticipé (en minutes).

ELAPSED_TIME : La durée réelle du vol.

AIR_TIME : Le temps en air qui est La durée entre le roulage de départ et le roulage d'arrivée.

DISTANCE : La distance parcourue (en Miles).

WHEELS_ON : Le temps de descente réel de l'avion.

TAXI_IN : Le roulage d'arrivée; représente le temps entre le temps de descente réel et l'entrée de l'avion.

SCHEDULED_ARRIVAL : Le temps d'arrivée anticipé.

ARRIVAL_TIME: Le temps d'arrivée réel de l'avion.

ARRIVAL_DELAY : Le retard d'arrivée.

DIVERTED : Variable qui représente si le vol a été dévié vers un autre aéroport. (1 si dévié et 0 sinon).

CANCELLED : variable qui représente si le vol a été annulé. (1 si annulé et 0 sinon).

CANCELLATION_REASON : La raison d'annulaion du vol, prends les valeurs suivantes : **A** : problème de la compagnie aérienne ou de l'avion, **B** : Problèmes relatives au météo, **C** : Problème dans le système et **D** : Porblèmes de sécurité.

AIR_SYSTEM_DELAY : variable indiquant si le vol a un retard dù au sytème d'aviation.

SECURITY_DELAY : variable indiquant si le vol a un retard dù à un problème de sécurité.

AIRLINE_DELAY : variable indiquant si le vol a un retard dù à un problème de la compagnie aérienne.

LATE_AIRCRAFT_DELAY : variable indiquant si le vol a un retard dù à l'avion (retard de vol précédent).

WEATHER_DELAY : variable indiquant si le vol a un retard dù au condition de météo.

Travail à faire

L'étudiant doit lire le fichier **"flights.csv"** et répondre aux questions suivantes. La réponse doit contenir le code nécessaire pour répondre, l'éxécution du code et un commentaire. Le commentaire doit être inclus pour chaque question et non pas seulement pour les questions qui nécessistent une analyse des résultats.

Question 01: Connaitre la base de données

1. Quel est le nombre de vols dans cette base de données ?

2. Quel est le nombre de variables qui ne contiennent aucune valeur manquante ?

Question 02: Variables utiles !!

1. Donner le sommaire statistique pour la variable représentant l'année du vol.

2. Quel est l'écart type de cette variable ? Est ce que cette valeur est logique ? Expliquer.

3. Donner le sommaire statistique pour la variable représentant le numéro du vol. Que représente la moyenne, le min, max et écart-type de cette variable ?

4. Est ce que ces 2 variable sont nécessaire dans l'analyse de cette base de donnée ? Peut-on les supprimer ? Commenter.

Question 03: Analayse temporelle

1. Quels sont les deux mois avec le nombre de vols le plus élevé.

2. Est ce qu'il y a un jour de semaine avec particulièrement plus de vols que les autres ? si oui lequel ? Expliquer.

3. Tracer un graphique linéaire pour le nombre de vols de chaque mois. Est ce que y a t'il des tendances à remarquer ?

4. Tracer un graphique linéaire pour le nombre de vols de chaque jour de semaine ? peut-on observer une certaine tenandance conernant le début de semaine ou les week-end ?

Question 04: Aéroport de départ et d'arrivée

1. Combien y-a-t'il d'aéroport de départ et d'arrivée dans la base ?

2. Donner les 5 aéroports de départ avec le plus nombre de vols.

3. Donner les 5 aéroports d'arrivée avec le plus nombre de vols.

4. Combien y-a-t'il de vols incluant un des aéroports cités dans les 2 questions précédentes ? Que représente ce nombre par rapport au nombre total de vols ?

Question 05: Compagnies aériennes

1. Combien de compagnie aérienne dans cette base ?

2. Quels sont les 05 compagnie aérienne avec le plus nombre de vols ?

3. Tracer un diagramme par bar pour la fréquence de vols pour ces 05 compagnies aériennes.

4. Tracer un diagramme circulaire pour la fréquence relative des vols pour ces 05 compagnies aériennes.

Question 06: Statistiques descreptives

1. Calculer la distance moyenne parcourues par les vols originaires des 05 aéroports de départ cités en question 4.2.

2. Calculer l'écart-type de la durée du vol à destination des 05 aéroports d'arrivée cités en question 4.3.

3. Calculer l'erreur quadratique moyenne pour le roulage de départ.

4. Calculer l'écart interquartile pour le roulage d'arrivée.

5. Calculer la variance de la durée réelle du vol pour les 3 premiers mois de l'année.

6. Calculer la médiane de la durée de vol anticipée durant les week-ends.

7. Calculer les quantiles à 35%, 65% et 85% de la durée de vol anticipée.

8. Calculer l'étendue du temps en air pour les vols qui ont parcourus une distance supérieure à la distance moyenne.

Question 07: Retard de vol

1. Donner les sommaire statistique des retards de vols de départ et d'arrivée.

2. Donner le nombre de vols qui ont fair un retard de départ (supérieur à 0) et le nombre de vols qui ont démarrés avant le temps anticicpé (inférieur à 0).

3. Donner la fréquence relative des vols qui ont arrivées avant le temps anticipé et après le temps anticipé.

Selon BTS un vol avec un retard est considéré si le vol fait plus de 15 minutes de retard.

4. Créer une nouvelle variable dans la base **DELAYED_D** qui prends la valeur 1 si le vol a un retard de départ selon la définition du BTS et 0 sinon.

5. Créer une nouvelle variable dans la base **DELAYED_A** qui prends la valeur 1 si le vol a un retard d'arrivée selon la définition du BTSet 0 sinon.

6. Quelle est la fréquence relative des vols avec un retard de départ ? et retard d'arrivée ? selon la définition du BTS.

Question 08 : Les raisons de retard

1. Quel est le nombre de valeurs manquantes dans les 5 variables qui contiennent **.___DELAY** dans leurs noms (les 5 dernières variables de la base). Commenter.

2. Que représente ce nombre ?.

3. Comparer ce nombre avec les résultats obtenues dans la question 07.

4. Tracer un diagramme circulaire pour chacune des 5 variables.

Question 09 : dévié ou annulé !

1. Préciser le pourcentage de vols annulés par rapport aux autres vols.

2. Classer les raisons d'annulations des vols (du pourcentage le plus faible au plus élevé).

3. Tracer un diagramme à barres pour les raisons d'annulations (le diagramme doit contenir les raisons classées de la même manière que la question précédente).

4. Quel est le pourcentage des vols dévié ?

5. Calculer la distance moyenne des vols dévié ?

6. Calculer la médiane de la durée réelle des vols déviés et la médiane de la durée anticipé des vols déviés. Comparer et Commenter.

Question 10 : Le roulage de départ et le roulage d'arrivée.

1. Calculer les quartiles du roulage de départ et du roulage d'arrivée.

2. Tracer l'histogramme du roulage de départ et du roulage d'arrivée (chaque variable dans un graphique).

3. Tracer la boite à moustaches du roulage de départ et du roulage d'arrivée (chaque variable dans un graphique).

4. Tracer l'histogramme et la boite à moustaches du roulage de départ en distinguant entre les vols avec un retard de départ et les autres.

5. Tracer l'histogramme et la boite à moustaches du roulage d'arrivée en distinguant entre les vols avec un retard d'arrivée et les autres.

Question 11 : Relation entre le retard de départ et le retard d'arrivée.

1. Calculer la corrélation entre le retard de départ et le retard d'arrivée. Commenter.

2. Tracer le nuage de points pour le retard de départ vs. le retard d'arrivée.

3. Calculer la matrice de corrélation entre les variables numériques suivantes : retard de départ, retard d'arrivée, durée réelle, durée anticipée, durée de vol et roulage de départ et roulage d'arrivée.

4. Analayser la matrice de corrélation.

Question 12 : Durée de vol vs Distance

1. Calculer la corrélation entre la durée de vol réelle et la distance parcourue.

2. Calculer la corrélation entre la distance parcourue et la durée de vol anticipée.

3. Calculer la corrélation entre le temps en air et la distance.

4. Tracer les nuages de points des variables des trois questions précédentes.

5. Tracer la boite à moustaches de la distance parcourue.

6. Tracer la boite à moustaches de la distance parcourue en distinguant entre les vols avec retard et les autres.

7. Tracer la boite à moustaches de la durée réelle de vol. Tracer la boite à moustaches de la durée de vol réelle en distinguant entre les vols avec retard et les autres.

8. Tracer la boite à moustaches de la durée anticipée de vol. Tracer la boite à moustaches de la durée de vol anticipée en distinguant entre les vols avec retard et les autres.

9. Concluez si l'une de ses variables peut influencer la surveance du retard de vol.