ENSSAT LANNION
6 rue Kerampont
22300 LANNION
FRANCE

Rapport de Projet de Bases de Données Avancées

Analyse de résumés de données

Mattias Kockum
Informatique - ENSSAT Promotion 2023

Enseignant ENSSAT Olivier Pivert

February 21, 2024

Table des matières

1 Introduction

Notre travail consiste à étudier des résumés de données. Ces résumés de données sont obtenus grâce à une réécriture d'une base de données classique concernant des millions de vols commerciaux américains. Le code utilisé est une adaptation du code proposé en cours. Pour notre analyse, nous avons choisi d'utiliser un sous ensemble des données brutes, bien que l'utilisateur final ait la possibilité d'utiliser toutes ses données en entrée du programme. Nous avons produit un programme agencé autour d'un 'main.py' qui orchestre les trois parties du programme : l'extraction de sous ensemble, la réécriture en données floues, et l'analyse de ces données.

2 Réécriture

Dans cette première partie, nous avons écrit le code permettant de calculer les moyennes des données et de les sauvegarder au format JSON. Nous avons aussi écrit les fonctions permettant de visualiser l'équilibre des termes. Pour cela, nous regardons la satisfaction moyenne de chaque modalité d'un même terme sur l'ensemble des données. Cela permet de savoir si une des modalités prend le pas sur les autres. [fig:1]

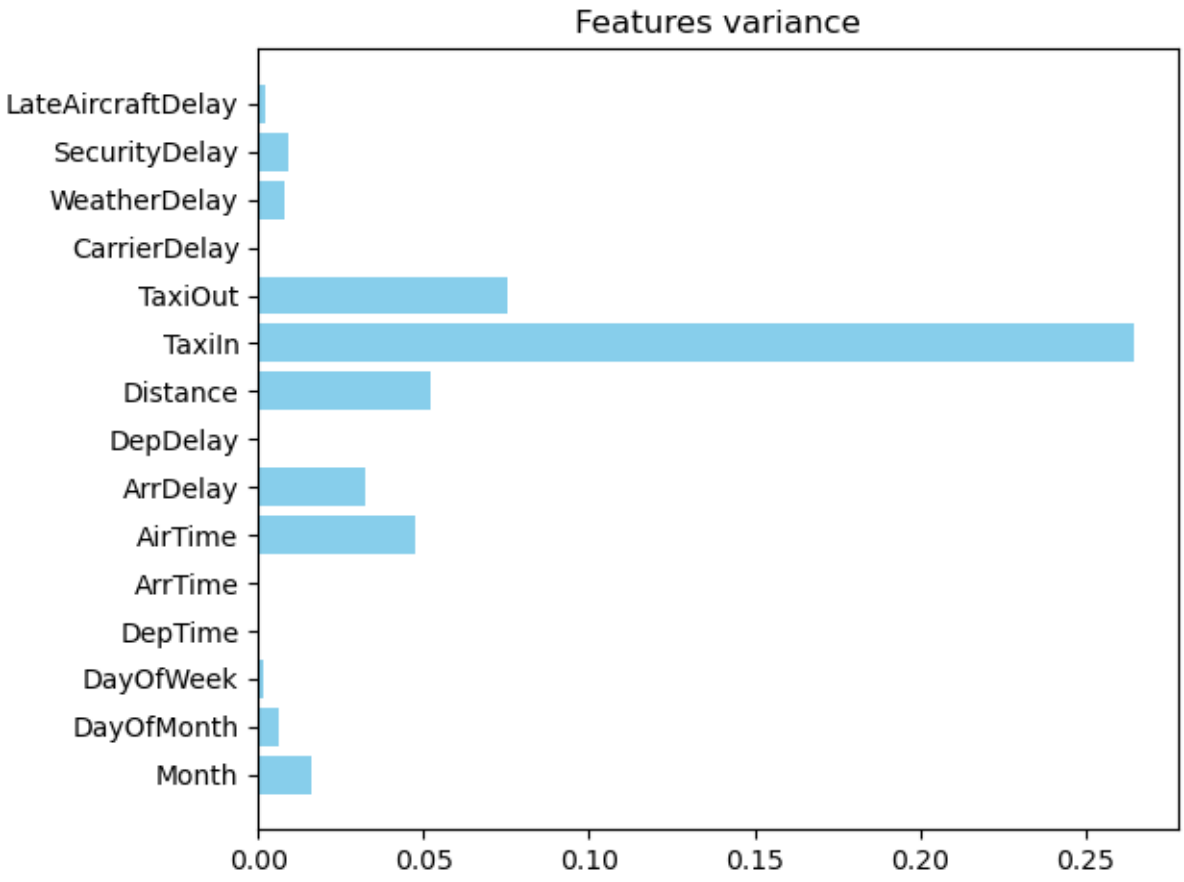

Figure 1

Dans certains cas, cela semble cohérent avec des variations auxquelles on peut s'attendre Par exemple : Plus d'avions lors des vacances d'été et d'hiver [fig:2] ou un équilibre entre les jours du mois. [fig:3]

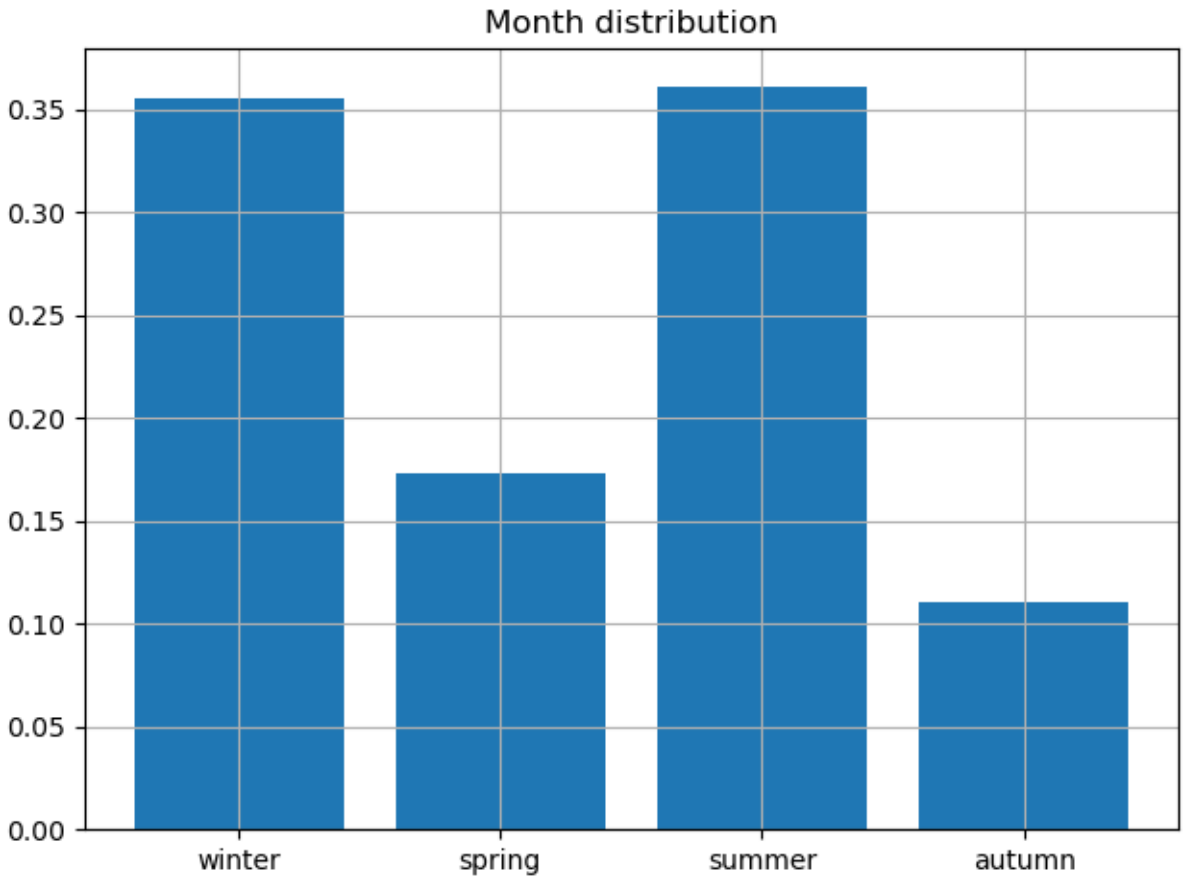
Month distribution
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
winter
spring
summer
autumn

Figure 2

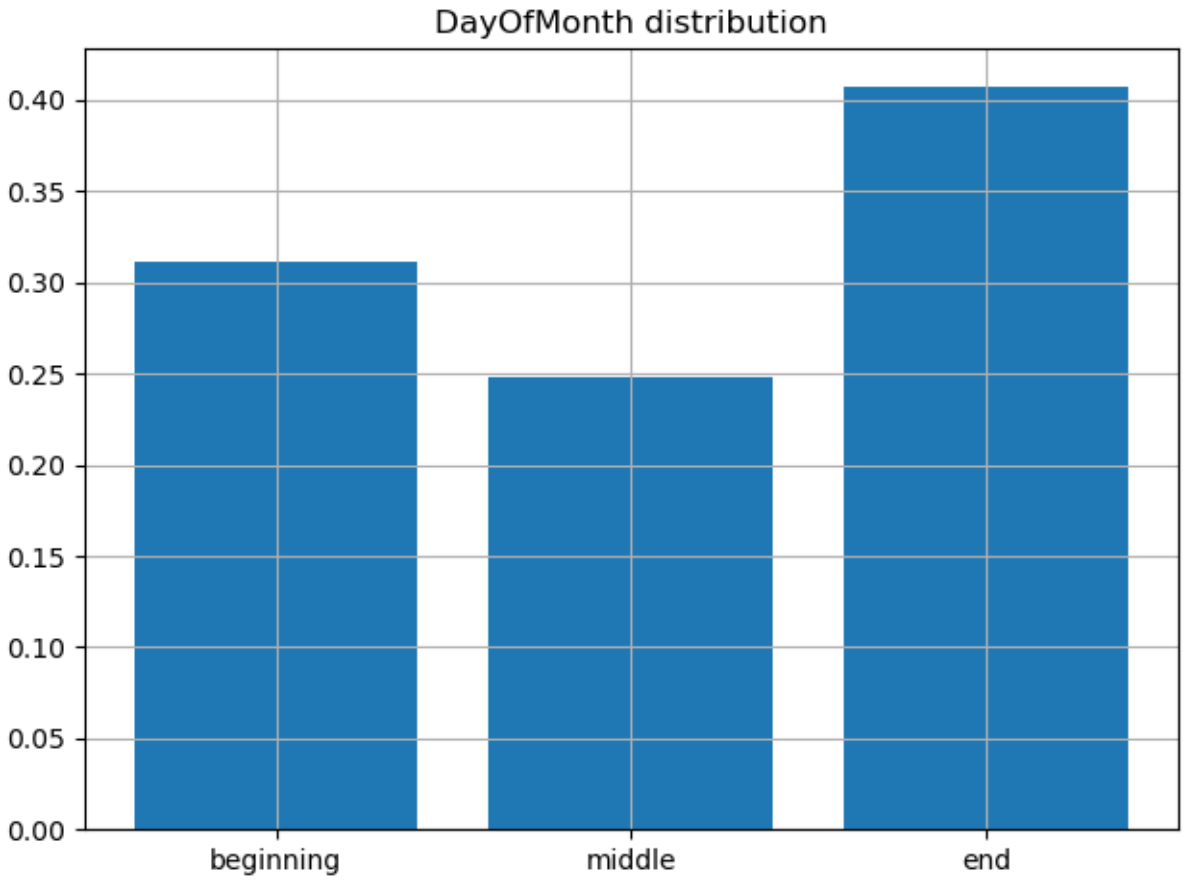

Figure 3

Dans d'autres cas, cela peut être plus étonnant. [fig:4] Ici, dans le cas des taxis cela peut indiquer, soit que les taxis déposent très rapidement les voyageurs, ce qui est possible, soit que les classes sont mal définies. En effet, la modalité 'short' de 'TaxiIn' va jusqu'à 20 minutes d'attente, il serait donc potentiellement judicieux de rajouter une modalité intermédiaire à 'short' et 'medium' pour augmenter la granularité de la donnée, augmenter sa précision et son équilibre.

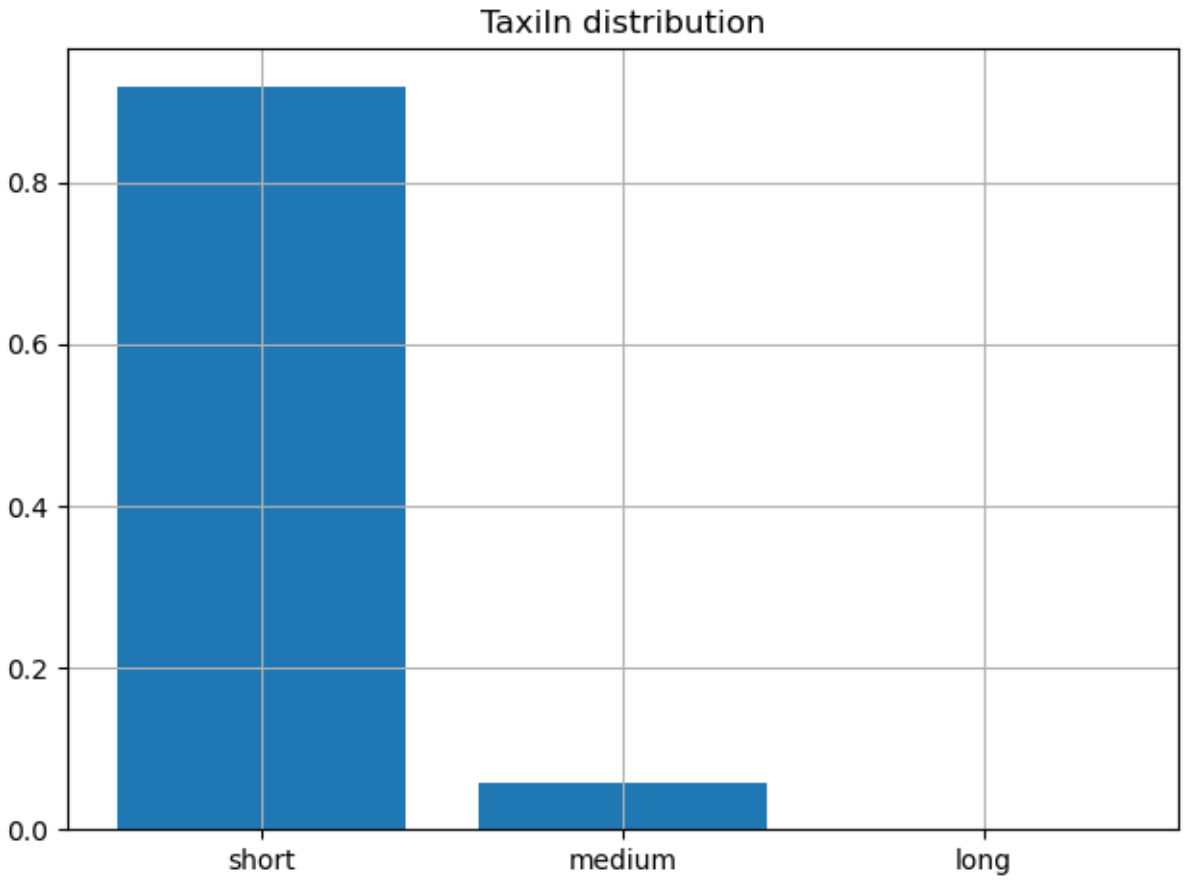

Figure 4

3 Exploration des données

Pour extraire les données correspondant à un certain seuil, nous avons utilisé la bibliothèque pandas permettant d'extraire des sous-ensembles de n-uplets dans un tableau en fonction de certains critères. Pour cela, nous avons écrit un code très simple [fig:5] où R est le tableau de n-uplets et v est un dictionnaire contenant en clefs les modalités et en valeurs leur valeur seuil au-dessous de laquelle les n-uplets sont écartés.

```
def cut(R, v):
    for feature, threshold in v.items():
        R = R[R[feature] > threshold]
    return R
```

Figure 5

Pour enregistrer le JSON associé, nous avons choisi une nomenclature explicite qui correspond au dictionnaire de critères v, à savoir modalité1_seuil1-modalité2_seuil2-etc.json .

4 Termes corrélés

Pour identifier les termes corrélés, nous avons opté pour une matrice de corrélation. Ici, la bibliothèque pandas permet d'effectuer le calcul de la corrélation sur tout le tableau.

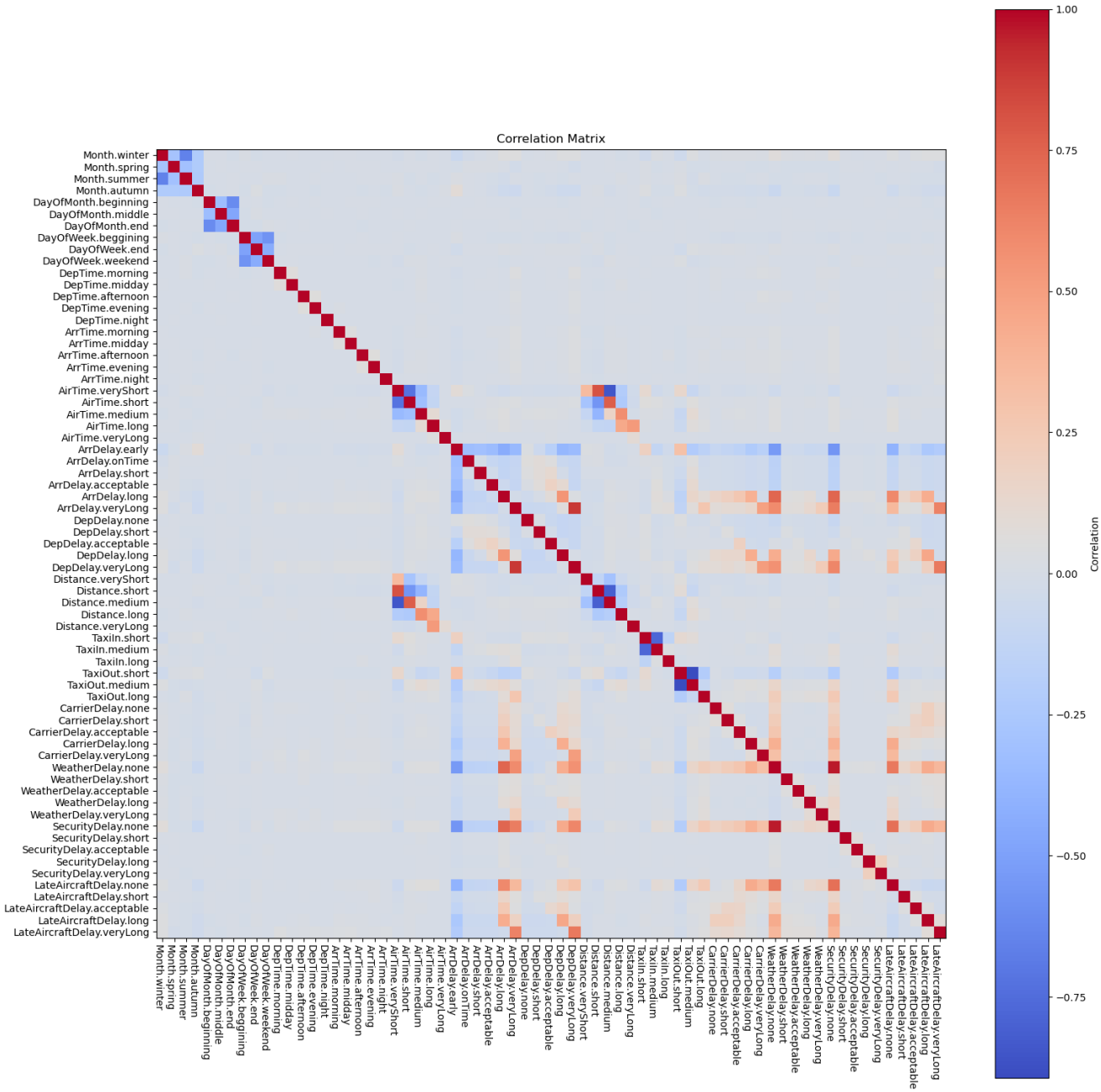
Correlation Matrix
Month.winter
Month.spring
Month.summer
Month.autumn
DayOfMonth.beginning
DayOfMonth.middle
DayOfMonth.end
DayOfWeek.beggining
DayOfWeek.end
DayOfWeek.weekend
DepTime.morning
DepTime.midday
DepTime.afternoon
DepTime.evening
DepTime.night
ArrTime.morning
ArrTime.midday
ArrTime.afternoon
ArrTime.evening
ArrTime.night
AirTime.veryShort
AirTime.short
AirTime.medium
AirTime.long
AirTime.veryLong
ArrDelay.early
ArrDelay.onTime
ArrDelay.short
ArrDelay.acceptable
ArrDelay.long
ArrDelay.veryLong
DepDelay.none
DepDelay.short
DepDelay.acceptable
DepDelay.long
DepDelay.veryLong
Distance.veryShort
Distance.short
Distance.medium
Distance.long
Distance.veryLong
TaxiIn.short
TaxiIn.medium
TaxiIn.long
TaxiOut.short
TaxiOut.medium
TaxiOut.long
CarrierDelay.none
CarrierDelay.short
CarrierDelay.acceptable
CarrierDelay.long
CarrierDelay.veryLong
WeatherDelay.none
WeatherDelay.short
WeatherDelay.acceptable
WeatherDelay.long
WeatherDelay.veryLong
SecurityDelay.none
SecurityDelay.short
SecurityDelay.acceptable
SecurityDelay.long
SecurityDelay.veryLong
LateAircraftDelay.none
LateAircraftDelay.short
LateAircraftDelay.acceptable
LateAircraftDelay.long
LateAircraftDelay.veryLong
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Correlation

Figure 6

Sur la matrice de corrélation obtenue [fig:6], nous pouvons encore une fois observée des choses qui étaient attendues, comme le fait que 'AirTime.veryShort' soit positivement corrélé à 'Distance.short' et négativement à 'Distance.medium'.

5 Termes atypiques

Pour identifier les termes atypiques, nous avons réutilisé le code vu précédemment pour obtenir l'équilibre des modalités. De cette manière, nous pouvons par exemple identifier que les départs de nuit sont peu fréquents. [fig:7]

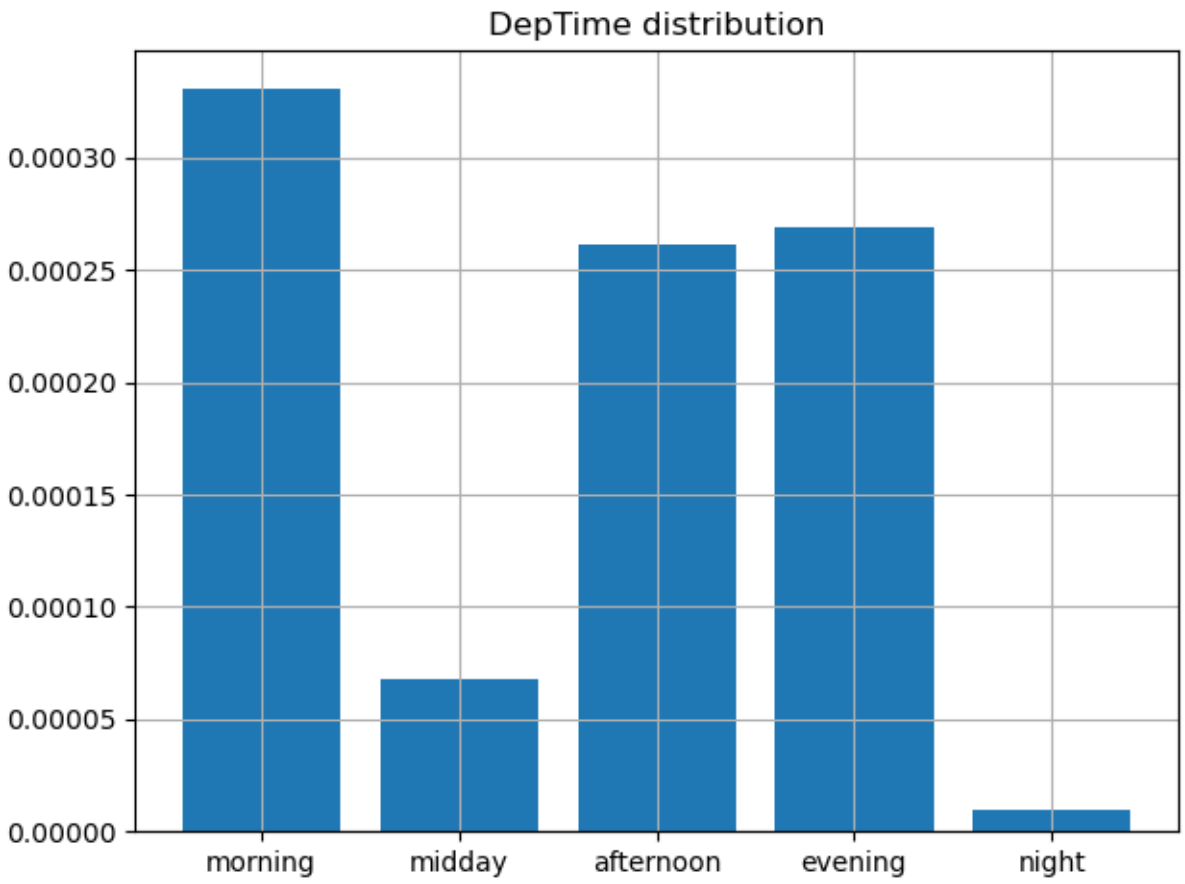
DepTime distribution
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
morning
midday
afternoon
evening
night

Figure 7