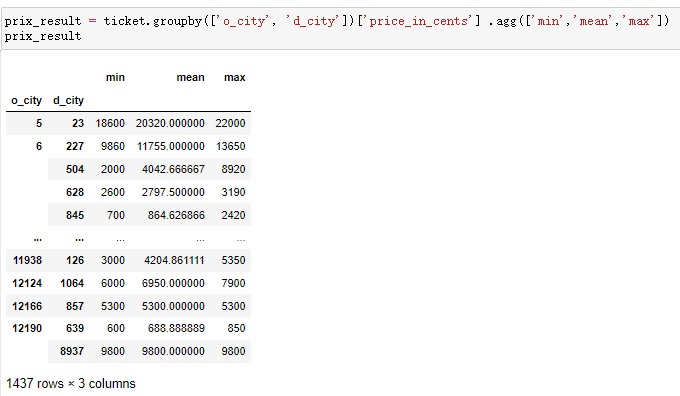
# Results:

## 1. Prix et Durée

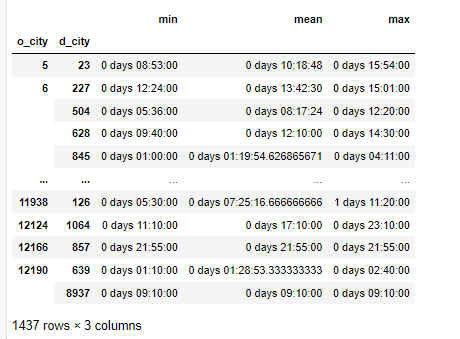
Sur la base du fichier ticket\_data.csv, le premier objectif est de trouver le prix le plus bas, le prix moyen et le prix le plus élevé, ainsi que le temps de trajet le plus bas, le plus élevé et le temps de trajet moyen pour chaque voyage.

Si j'ai bien compris, l'objectif est de regrouper les données par ville de départ (o\_city) et par ville d'arrivée (d\_city), le voyage étant défini comme le déplacement d'une ville à l'autre. Les résultats finaux seront présentés en conséquence.

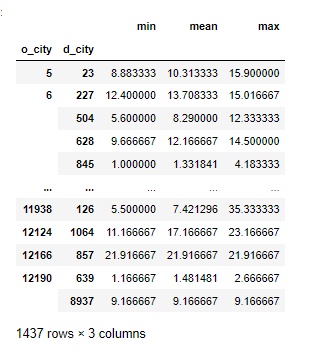
### 1.1 Le prix



### 1.2 Durée



**Durée en heures：**



## 2. La différence de prix et de temps s'explique par des distances et des modes de transport différents.

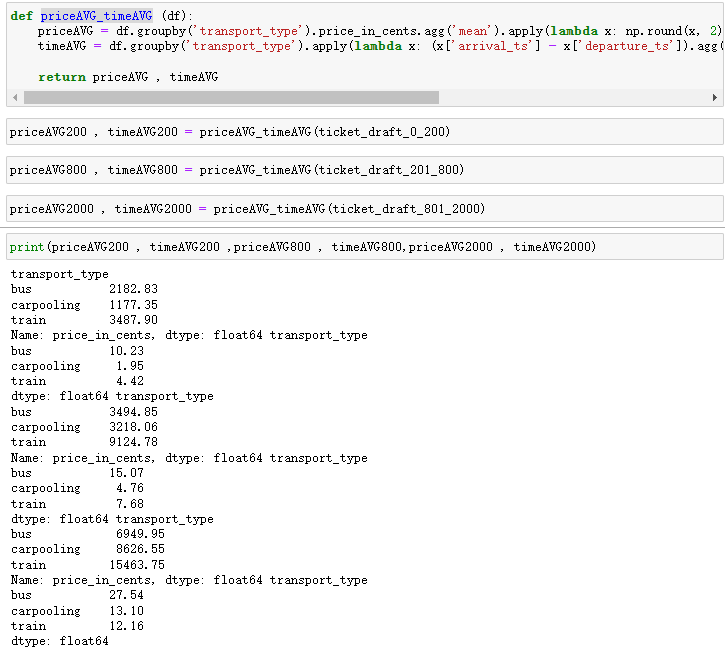
### 2.1 Distance en ligne droite

Tout d'abord, nous calculons la distance en ligne droite entre deux villes en nous basant sur la ville de départ (o\_city) et la ville d'arrivée (d\_city) du fichier ticket\_data.csv, ainsi que sur les informations relatives à la latitude et à la longitude du fichier cities.csv.

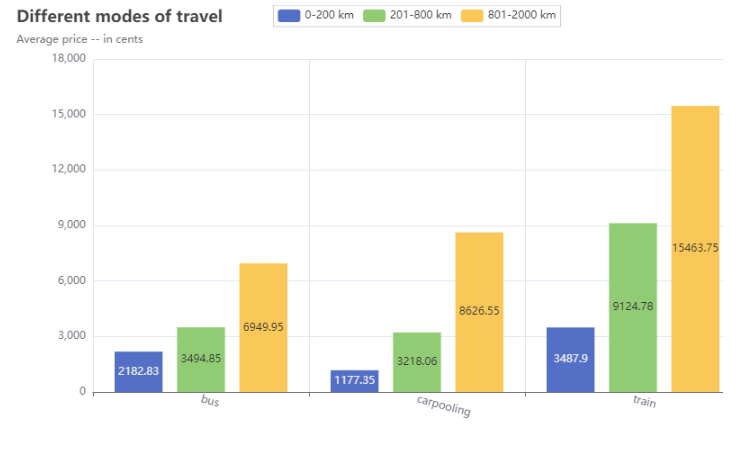
Étant donné que les distances dans le tableau sont toutes inférieures à 2000 km, nous diviserons les données en trois groupes : 0-200 km, 201-800 km et 800-2000 km.

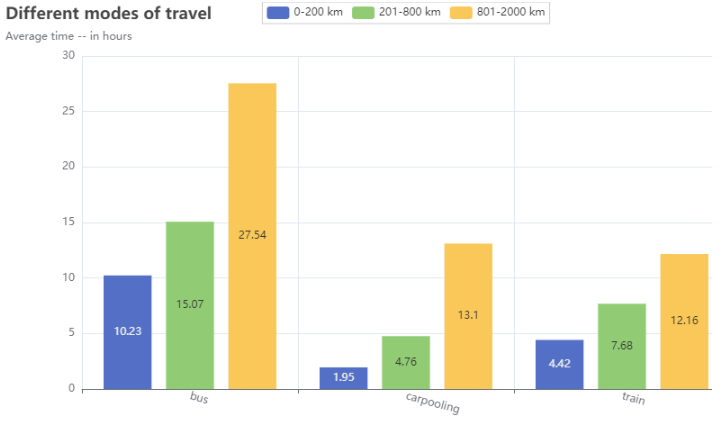
### 2.2 Définition des fonctions

Définir deux fonctions (prixAVG) et (tempsAVG) pour calculer la moyenne du prix et du temps pour différentes catégories sur la base de différentes distances.



## 3. Présentation de "pyecharts" pour la visualisation de données





## 4. Faire des prévisions de prix

### 4.1 Utiliser dummie

Tout d'abord, nous utiliserons des variables "dummy" pour catégoriser le transport\_type. Cela est nécessaire pour prédire les prix à l'avenir. Bien qu'il n'y ait que trois types de données dans ce cas, les remplacer serait possible, mais l'utilisation de variables "dummy" serait plus appropriée au cas où il y aurait plusieurs types de données à l'avenir.

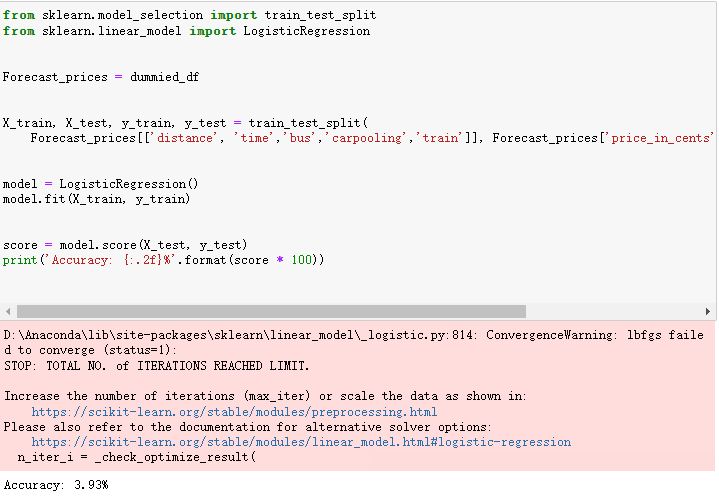
### 4.2 Tout d'abord, une simple prédiction

La modification de "new\_distance" ou "new\_time" donnera des résultats différents.



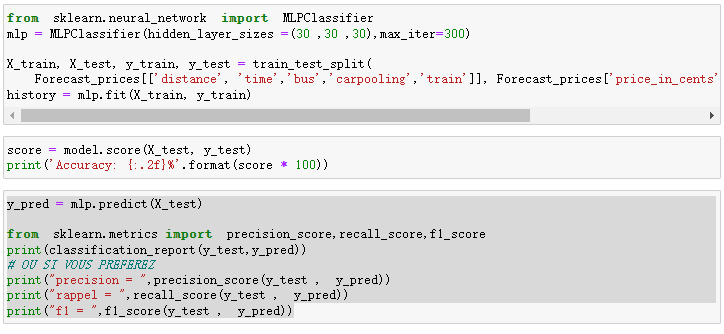
### 4.3 Régression linéaire

Ensuite, un modèle simple d'apprentissage automatique est utilisé, le modèle est entraîné et les résultats sont obtenus :



### 4.4 Apprentissage profond--MLP

J'ai choisi d'utiliser l'apprentissage profond car les résultats du modèle d'apprentissage automatique n'étaient pas aussi bons que je l'aurais souhaité. Cependant, en raison du volume de données, mon ordinateur n'a jamais exécuté les résultats, mais je suis sûr qu'il n'y a rien de mal dans mes idées et mon code, alors s'il vous plaît, essayez de trouver des résultats si vous le pouvez.



# Questions:

1. J'avais l'intention d'utiliser databricks pour la prédiction des prix, mais pour une raison ou une autre le site continue à être buggé, j'espère pouvoir l'essayer à l'avenir.

2. La colonne 'id' dans le tableau providers.csv est égale à la colonne 'company' dans ticket\_data.csv, donc que signifie 'company\_id' ?