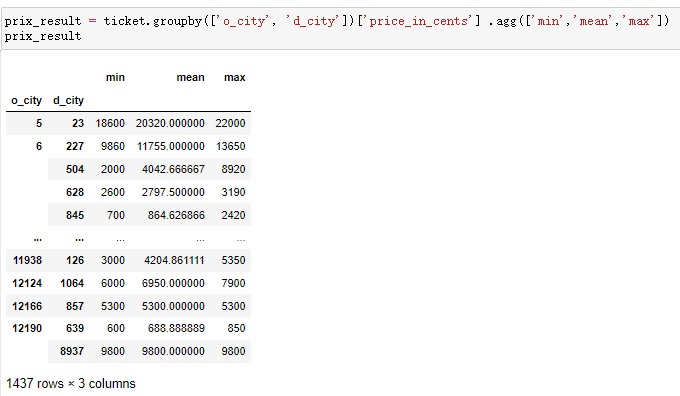
# Results:

## 1. Price & Time

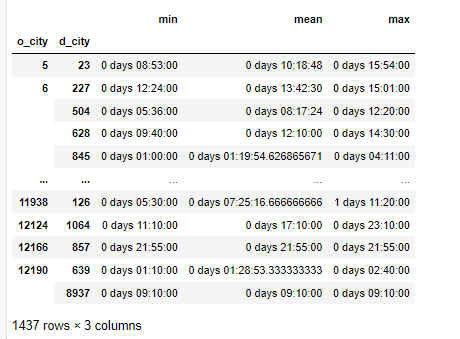
Based on the ticket\_data.csv file, the first objective is to find the lowest, average, and highest prices, as well as the lowest, highest, and average travel time for each trip.

My understanding is that the objective is to group the data by the departure city (o\_city) and arrival city (d\_city), as the trip is defined as the journey from one city to another. The final results will be presented accordingly.

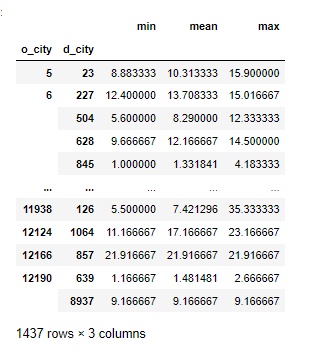
### 1.1 Price



### 1.2 Time



**Time in hours：**



## 2. The difference in price and time is derived from different distances and different modes of travel.

### 2.1 Straight line distance

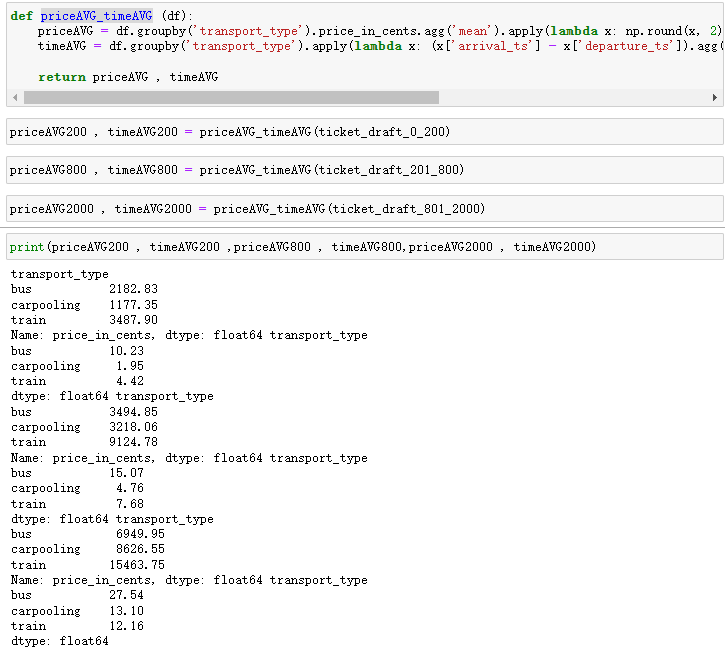
First, we will calculate the straight-line distance between two cities based on the departure city (o\_city) and arrival city (d\_city) from the ticket\_data.csv file, as well as the latitude and longitude information from the cities.csv file.

Since the distances in the table are all within 2000 km, we will divide the data into three groups: 0-200 km, 201-800 km, and 800-2000 km.

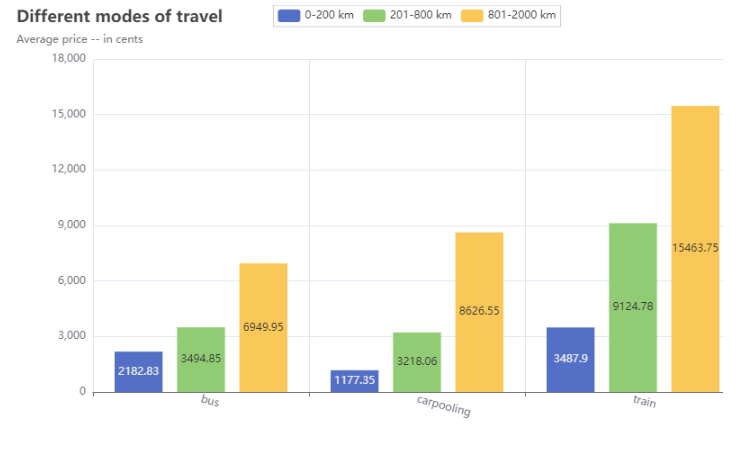


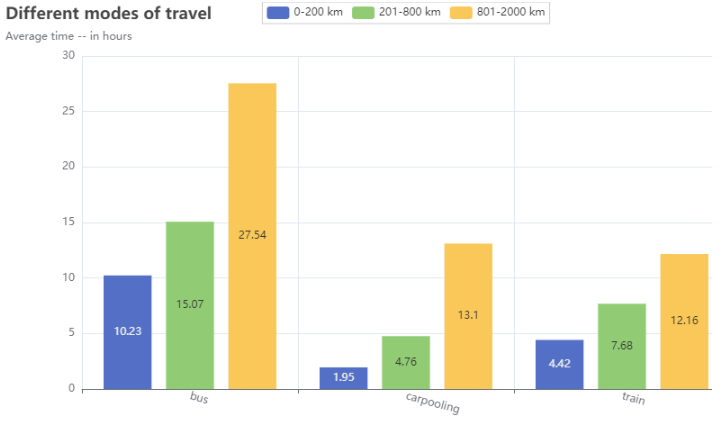
### 2.2 Defining functions

Define two functions (priceAVG) and (timeAVG) to derive the average of price and time for different categories based on different distances



## 3. Introducing ‘pyecharts’ for data visualisation





## 4. Making price forecasts

### 4.1 Using dummie

First, we will use "dummy" variables to categorize the transport\_type. This is necessary for predicting prices in the future. Although there are only three types of data in this case, replacing them would be possible, but using "dummy" variables would be more appropriate in case there are multiple types of data in the future.

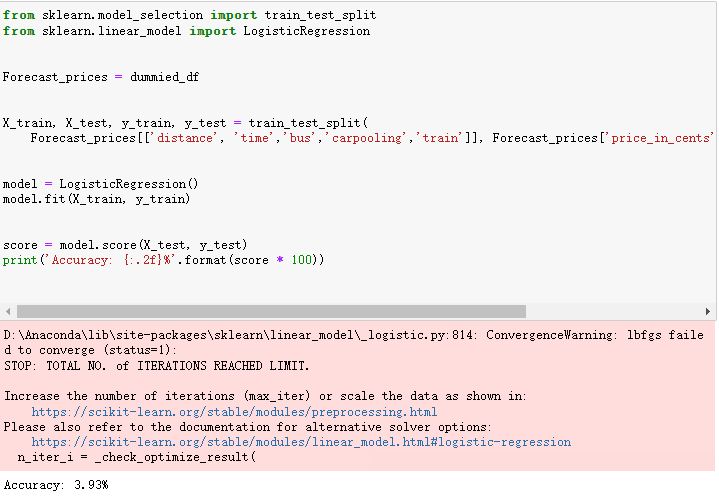
### 4.2 First a simple prediction

Changing 'new\_distance' or 'new\_time' will give different results.



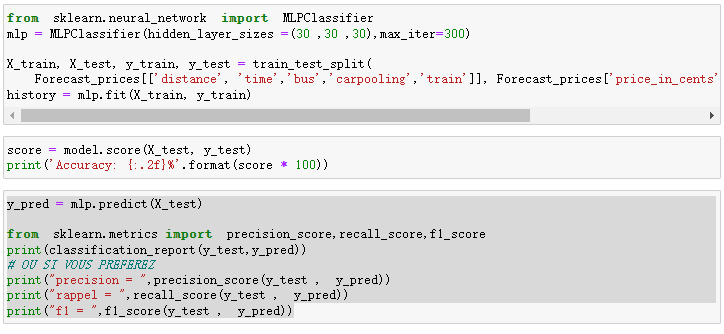
### 4.3 Linear regression

Next, a simple machine learning model is used, the model is trained and the results are obtained



### 4.4 Deep Learning--MLP

I chose to use deep learning as the results from the machine learning model were not as good as I would have liked. However, due to the sheer volume of data, my computer never ran the results, but I am sure there is nothing wrong with my ideas and code, so please try to come up with results if you can.



# Questions:

1. I was going to use databricks for price prediction, but for some reason the site keeps getting buggy, hopefully I can try it in the future.

2. The 'id' column in the providers.csv table is equal to the 'company' column in ticket\_data.csv, so what does the What does 'company\_id' mean?