Proposal

Code for data:

```
import pandas as pd
chunksize = 100000
filename = 'itineraries.csv'
filename = 'itineraries.csv'
df = pd.DataFrame(), i = 0
for chunk in pd.read_csv(filename, chunksize=chunk_size):
    output_filename = f'output_chunk{i}.csv'
    chunk.to_csv(output_filename, index=False)
    i = i + 1
    print(i)
print(df.head())
```

There are lots of variables in our dataset. I will list some of the important ones here for reference.

- searchDate: The date on which this entry was taken from Expedia.
- flightDate: The date of the flight.
- startingAirport: Three-character IATA airport code for the initial location.
- destinationAirport: Three-character IATA airport code for the arrival location.
- isBasicEconomy: Boolean for whether the ticket is for basic economy.
- isRefundable: Boolean for whether the ticket is refundable.
- isNonStop: Boolean for whether the flight is non-stop.
- totalFare: The price of the ticket (in USD) including taxes and other fees.

We want to using this dataset to build a model for users to predict their budget for their trips. The initial plan is implementing different regression method to build the model, including: simple linear regression, K-Nearest Neighbors Regression, Gradient Boosting Regression, and Random Forest regression.

The computational steps are firstly dividing the dataset with the feature isBasicEconomy. After that, we will divide the dataset with the features isNonStop. We will use the feature isRefundable to do final data split. These three features can cover most of our users' trip preference. Splitting data in this way is not only a good practice of parallel computing, but also gives the model an efficient training process.

Github link: willwzidi/DSCP