# **NYC Airbnb Price Prediction**

**DATASET**

Dataset link: <https://fred.stlouisfed.org/>

# **DATA PIPELINE**

## STEP - 1 : Data Scraping

* Used Fred api to
* Beautiful Soup is used to scrape data into S3 with AWS LAMBDA as a compute service
* The zipped files are then extracted and pushed to a **Landing S3 Bucket** which serves as the source for the processing aspect of the pipeline

## STEP - 2 : Data Pre-Processing

* Used AWS Glue for Cleaning and Transformation of the data
* The raw data obtained was first resized from the original 147 columns down to 38 columns, since all other features had a lot of empty columns and were not useful for our predictions
* Transformed and cleansed data is written to S3 bucket using ETL job into Parquet format
* AWS Crawler is being used to read data from S3 to Athena

## STEP - 3 : Machine Learning Model

* Used BorutaPy for Feature Selection and GridSearchCV to tune the Random Forest Model
* Also, we have used H2O as an autoML service to predict the prices

## STEP - 4 : Visualizations using AWS QuickSight

* We have used QuickSight to visualize some important features for the end users
* Users get detailed insights from the Dashboards

## STEP - 5 : Flask application

* Contextualize the prediction with respect to the average prices matching the features

# 

# 

# AMUNDSEN

* Amundsen is a meta-data store for extracting information out of the our data
* We have integrated Amundsen to our project to poll data from the AWS Glue Data Catalog at every 10 mins so as to keep the meta-data store updated

# 