# **NYC Airbnb Price Prediction**

**DATASET**

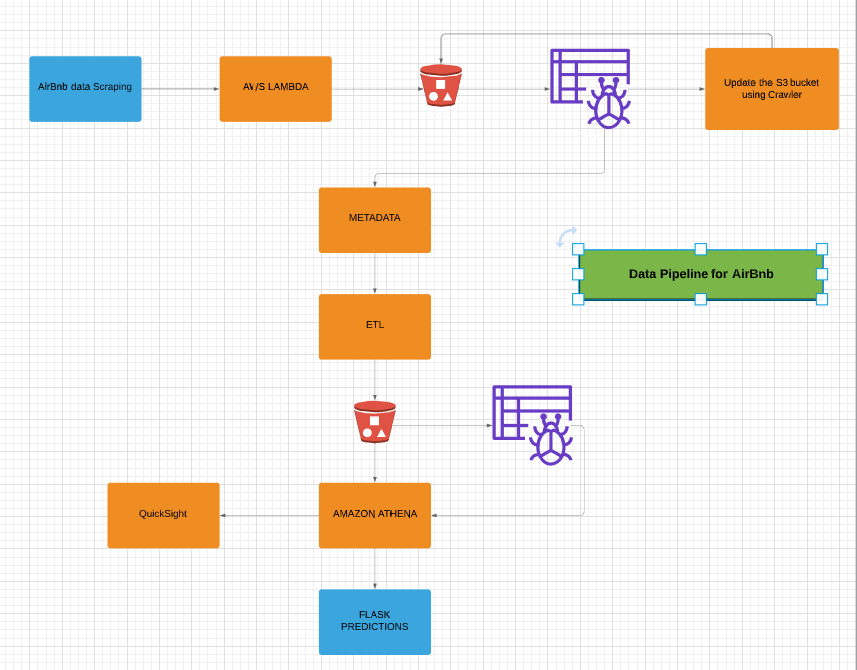
Dataset link: <http://insideairbnb.com/new-york-city/>

New York City is a city of renters, vacancy rates are at crisis levels, and rents continue to rise. Income levels for the average New Yorker haven't kept pace, and affordability is at record lows. Housing is scarce; homelessness levels are increasing; food insecurity is growing; and economic and racial inequality rates in New York City are near the highest in the United States. It's at this time that short term rental platforms, dominated by Airbnb, have entered the market, and grown to have listings of tens of thousands of rooms and entire apartments. Airbnb addresses the demand for tourist accommodation and creates an income stream for "hosts," and ignores both the need for and loss of housing. Behind buzzwords like "the sharing economy" and "disruptive," Airbnb as a platform simply allows many of their hosts to operate unlicensed, unregulated and untaxed hotels in residential neighborhoods. The taxes that Airbnb refuses to collect for the city are being used as leverage in Airbnb's corporate lobbying and marketing campaigns to force the city and state to legitimize their business model and legalize Airbnb hosts' activities. This "bribe" hides the fact that Airbnb is enabling a massive abuse of housing in New York City, and, in rapidly changing and gentrifying neighborhoods.

# **PROJECT GOALS**

* Predict the prices based on the features for the NY Airbnb listings
* Create a pipeline for Data Extraction,Ingestion and Inference
* Visualize the insights from the Data using QuickSight
* Compare predicted and average prices to give a fair idea to the users

# **DATA PIPELINE**



## STEP - 1 : Data Scraping

* The data set was programmatically downloaded from insideairbnb.com
* 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

# 