Singapore's Airbnb Price Predictor

Yosi Aditya Nugroho
Job Connector Data Science Purwadhika

Table of Contents

01 Problem Statement

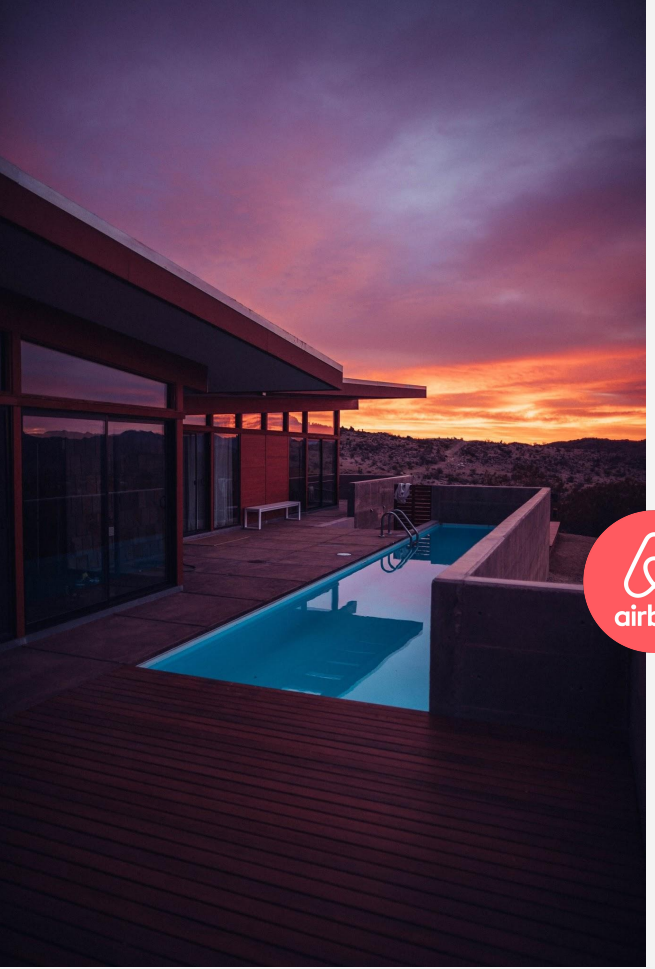
COVID-19 Pandemic

One year has passed since COVID-19 pandemic spread throughout the world, slowing down the economic growth in all parts of the world. Many countries applied lockdown as their measures to contain the spread of the virus, resulting in slump demand for tourism all over the world. But, recently some countries has lift the lockdown measures as their effort to suppressed the COVID-19 cases. Resulting in thriving economic post-pandemic and increasing demand for tourism

Why Singapore

Singapore is one of those successful countries to fight against pandemic. They start to open their border for travellers and hopefully can revive its tourism industry as soon as possible. Singapore's tourism industry itself is vital for Singapore as it makes **4.1%** of their national GDP (2017)

As the tourism industry started to grow, one of the important thing the traveler will need is **accommodation** as a place to spend the night



About Airbnb

Aside from hotel, Airbnb is another choice for traveller to look for accomodation in their travel destination. Airbnb is an American vacation rental online marketplace company. Through the service, users can arrange lodging, primarily homestays, and tourism experiences or list their properties for rental. Airbnb does not own any of the listed properties; instead, it profits by receiving commission from each booking

Problem

As the lockdown measure has been lifted and tourism demand increases, Singapore resident can rent their properties in Airbnb to gain extra income sources and help to boost the economic growth in their countries. But, choosing the right competitive price is difficult since there are so many factors to be considered and calculated

Solution

Machine learning can help to do those calculations and predict the correct listing price for new host as accurate as possible. This model can also be helpful for former host to revalue their listing into new competitive price and also for traveller to plan their budget for accommodations suiting their needs

Dataset Information

02

The dataset used for this project is downloaded from Inside Airbnb.

The data behind the Inside Airbnb site is sourced from publicly available information from the Airbnb site. The data has been analyzed, cleansed and aggregated where appropriate to facilitate public discussion

The data that I will be using is Singapore's Airbnb listing that is last scraped on **December 29, 2020** and contains **4387** listing information

The Four Datasets

Calendar

Listings availability information in time series format

1.6M rows and 7 columns

Reviews

Detailed Review Data for listings in Singapore

55 rows and 2 columns

Listings

Detailed Listings data for Singapore

4387 rows and 74 columns

Neighbourhood

Neighbourhood list for geofilter. Sourced from city or open source GIS files.

53984 rows and 6 columns

03 Data Cleaning Process

Data Cleaning Process

05

Missing Values and Encoding Strategy

Decide strategy for imputing missing values and encoding categorical variable

04

Dropping Outlier Values

Drop some outlier records to prevent worse model performance

03

Feature Engineering

Create new features derived from existing column on dataset

02

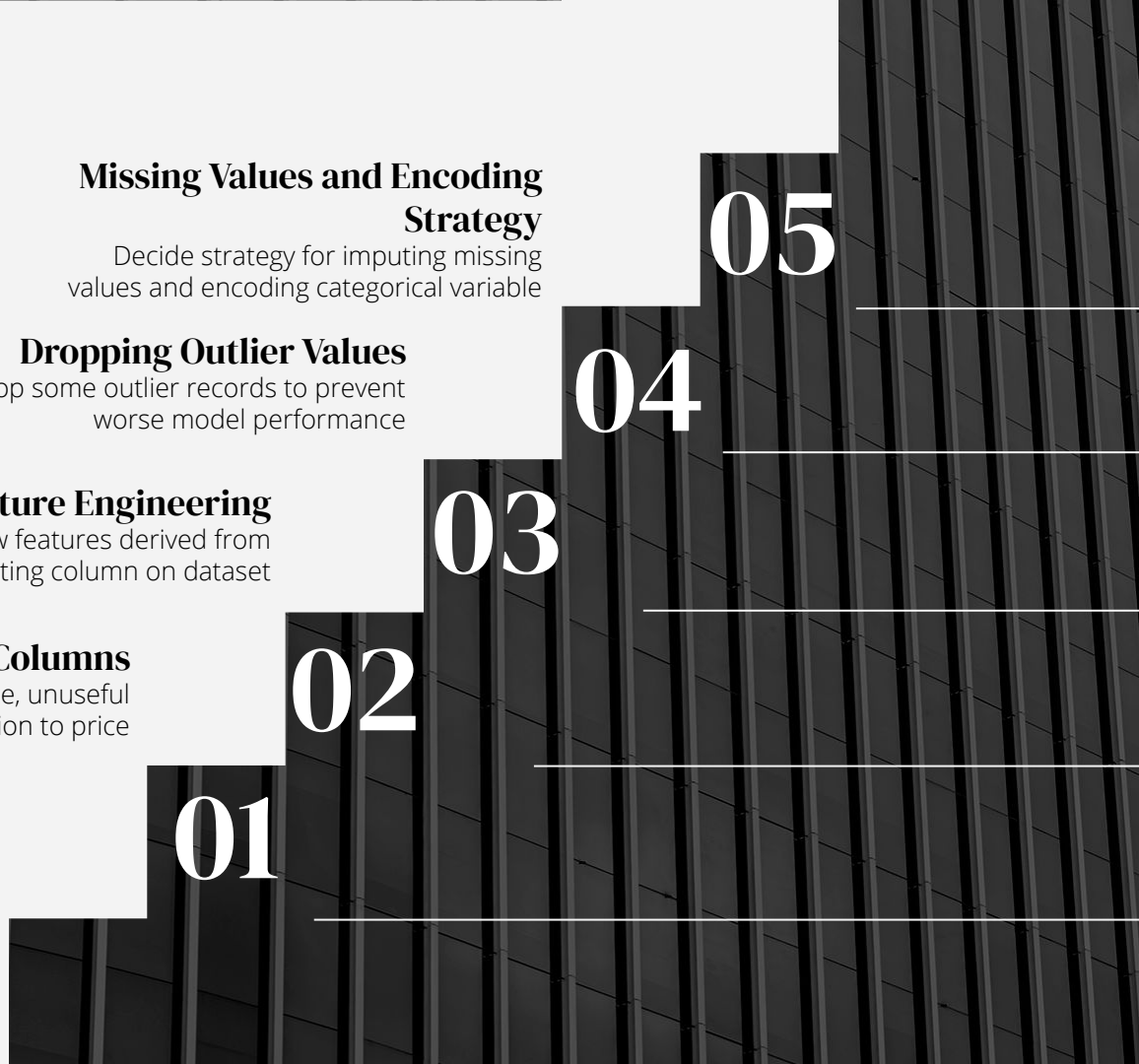
Drop Unnecessary Columns

Drop columns with zero/little variance, unuseful columns and columns with small correlation to price

01

Fix Incorrect Data Types

Removing unwanted characters and recast as appropriate data type

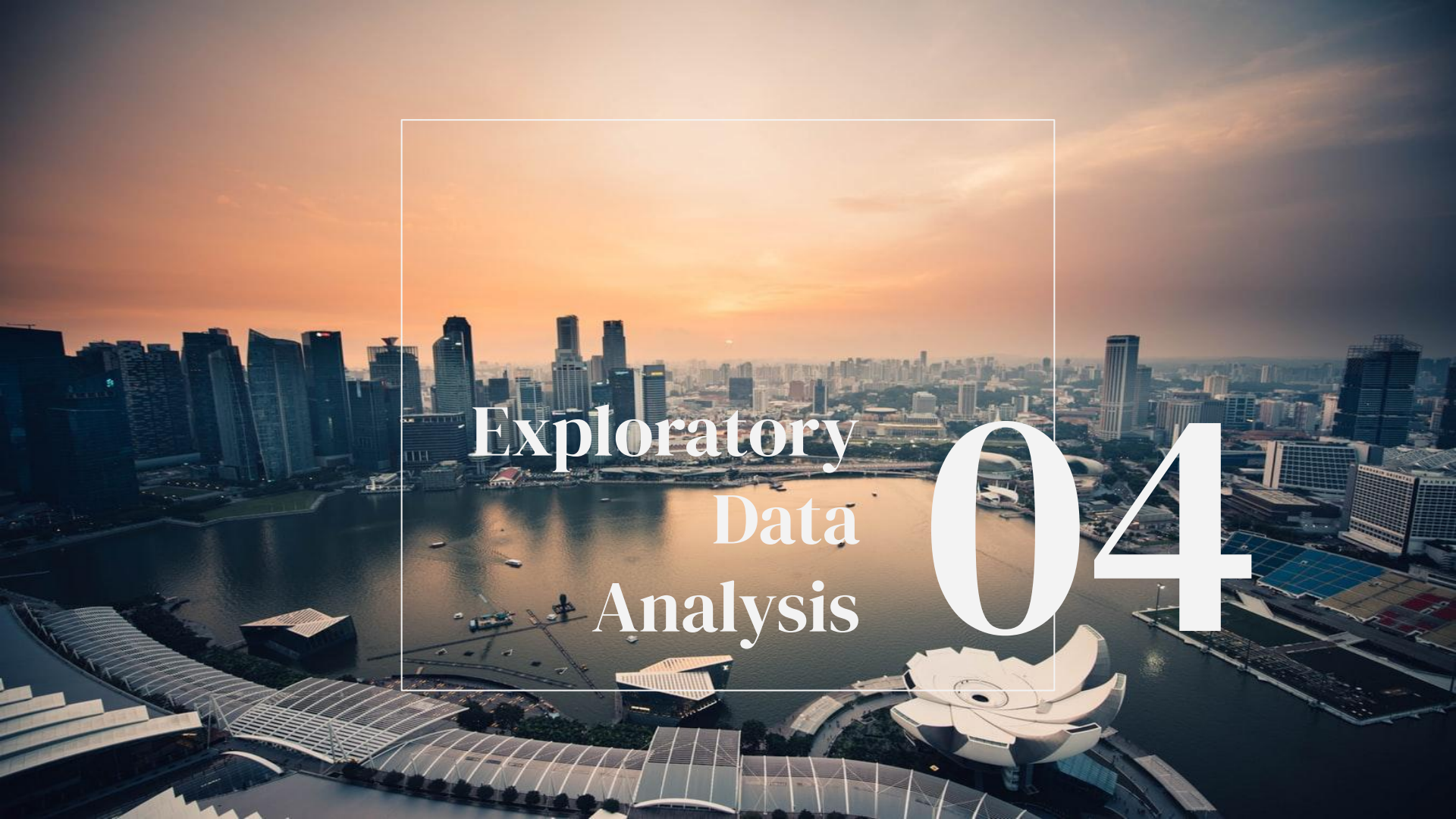


Before Cleaning

4387 Rows
74 Columns

After Cleaning

4372 Rows
18 Feature Columns
50 Amenities Columns



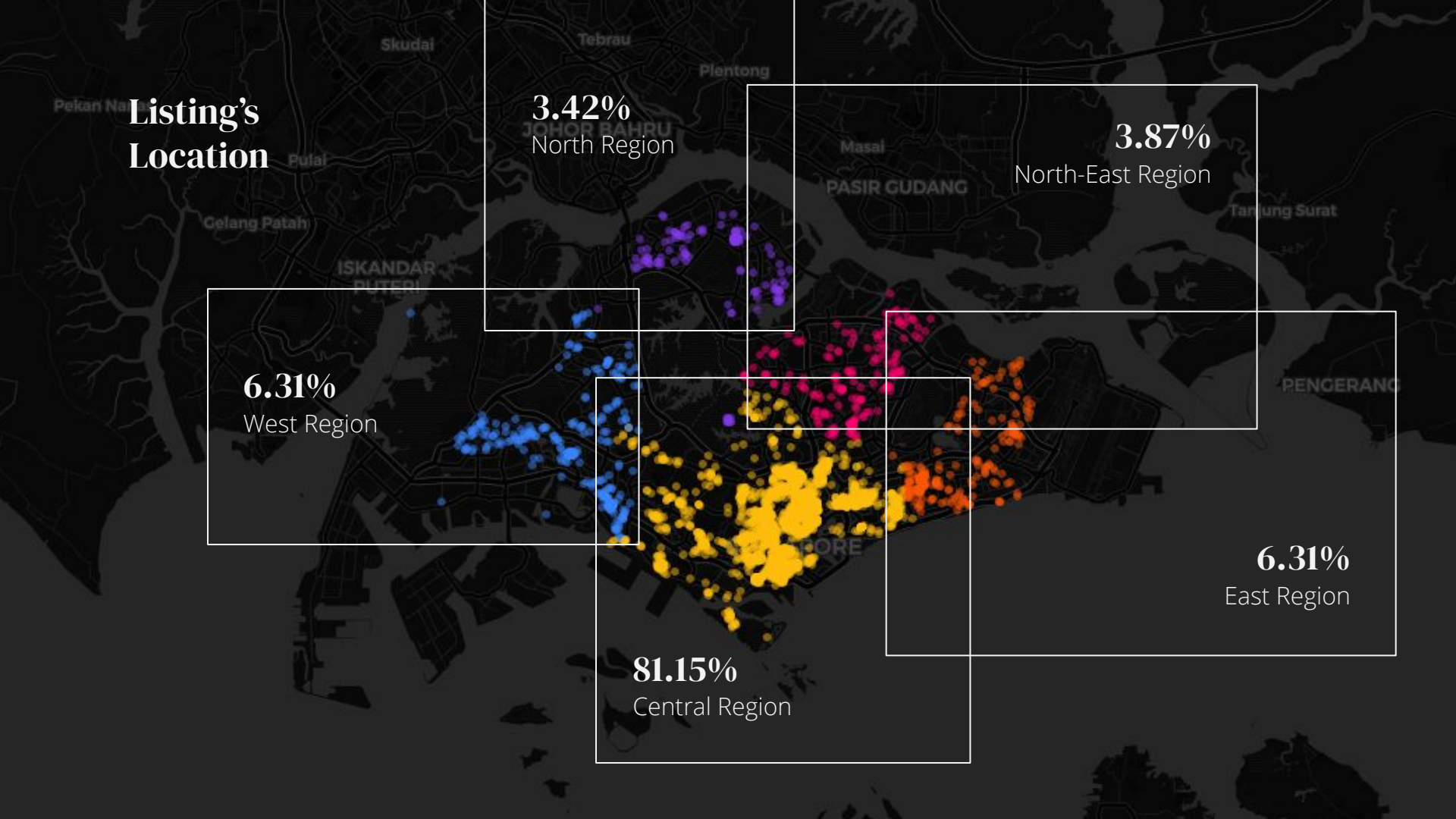
Exploratory Data Analysis

04

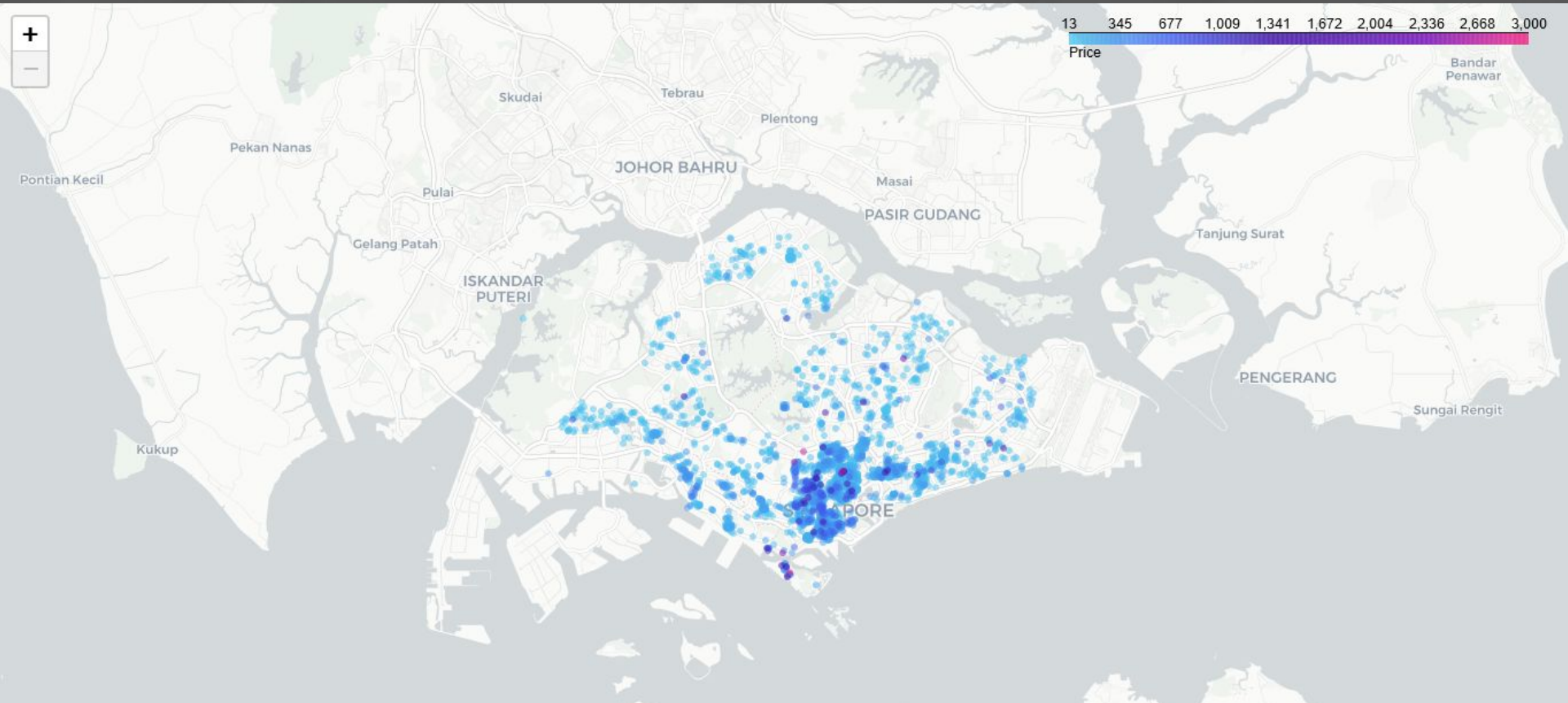
Price Distribution

The price distribution for 4372 Airbnb listing in Singapore is heavily skewed as there are some listing with much higher price than the other. The median price is at **SGD114**, minimum price at **SGD13** and maximum price at **SGD3000**

Price Percentiles


Listing's Location
3.42%
North Region
3.87%
North-East Region
6.31%
West Region
6.31%
East Region
81.15%
Central Region
Skudai
Tebrau
Plentong
Pekan Nanas
JOHOR BAHRU
Pulai
Masai
PASIR GUDANG
Tanjung Surat
Gelang Patah
ISKANDAR PUTERI
PENGERANG

Price and Location



Price and Neighbourhood

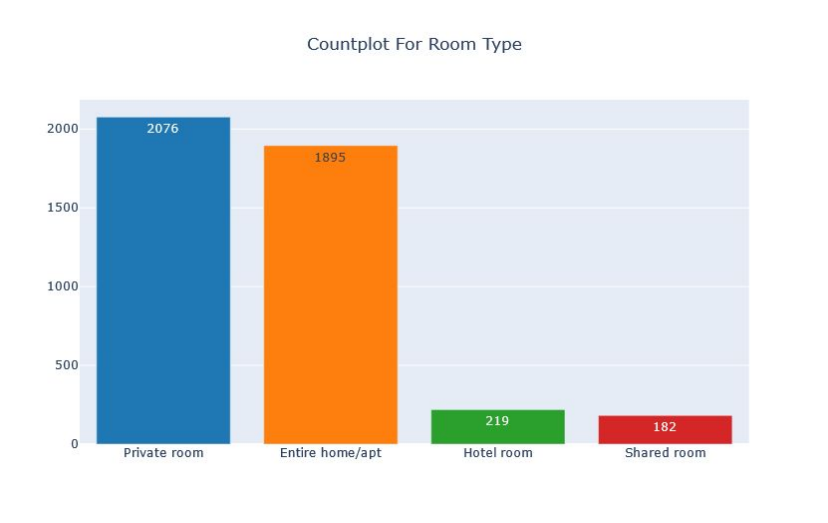
KDE Plot

Boxplot

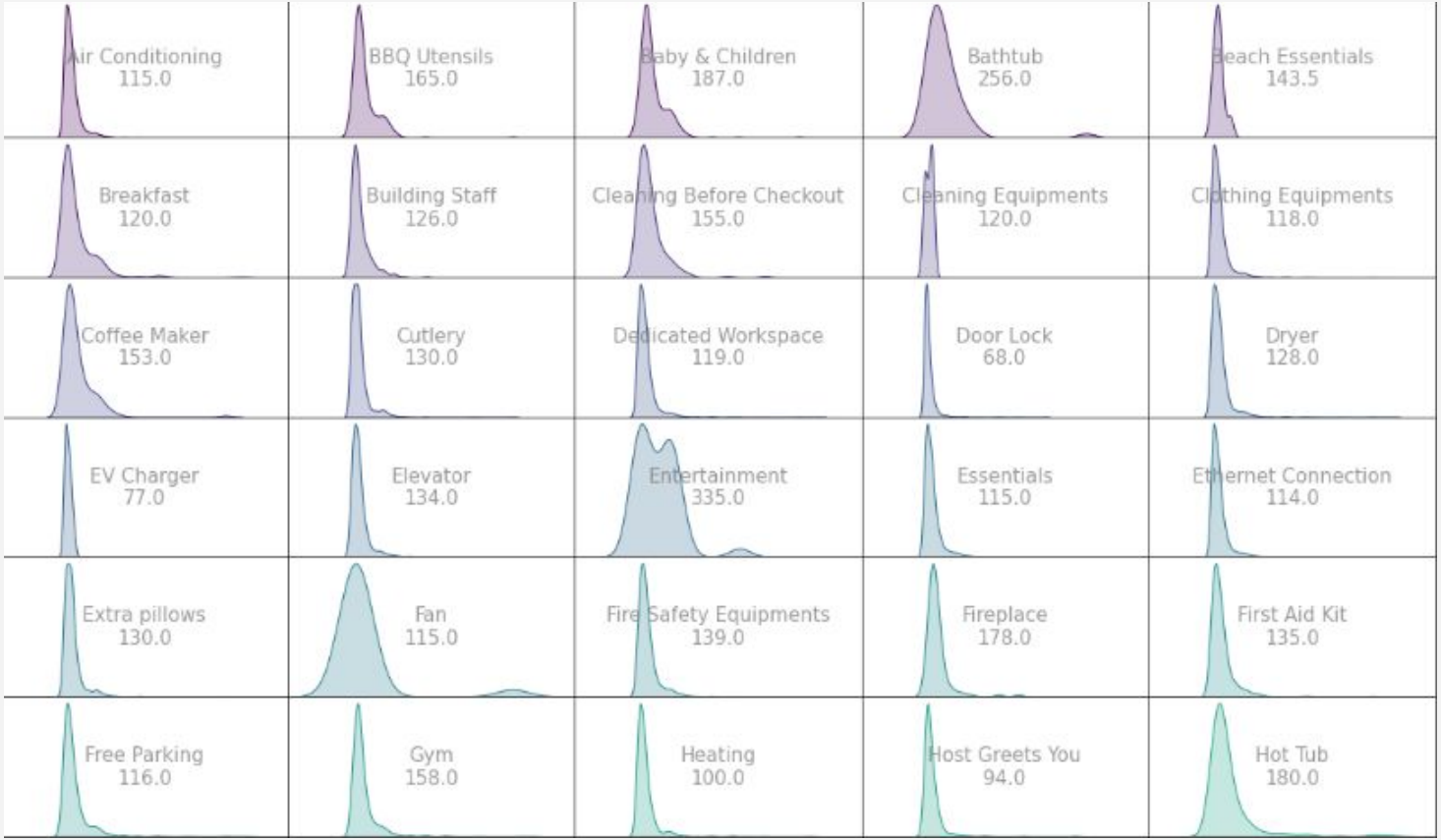
Price and Room Type

KDE Plot

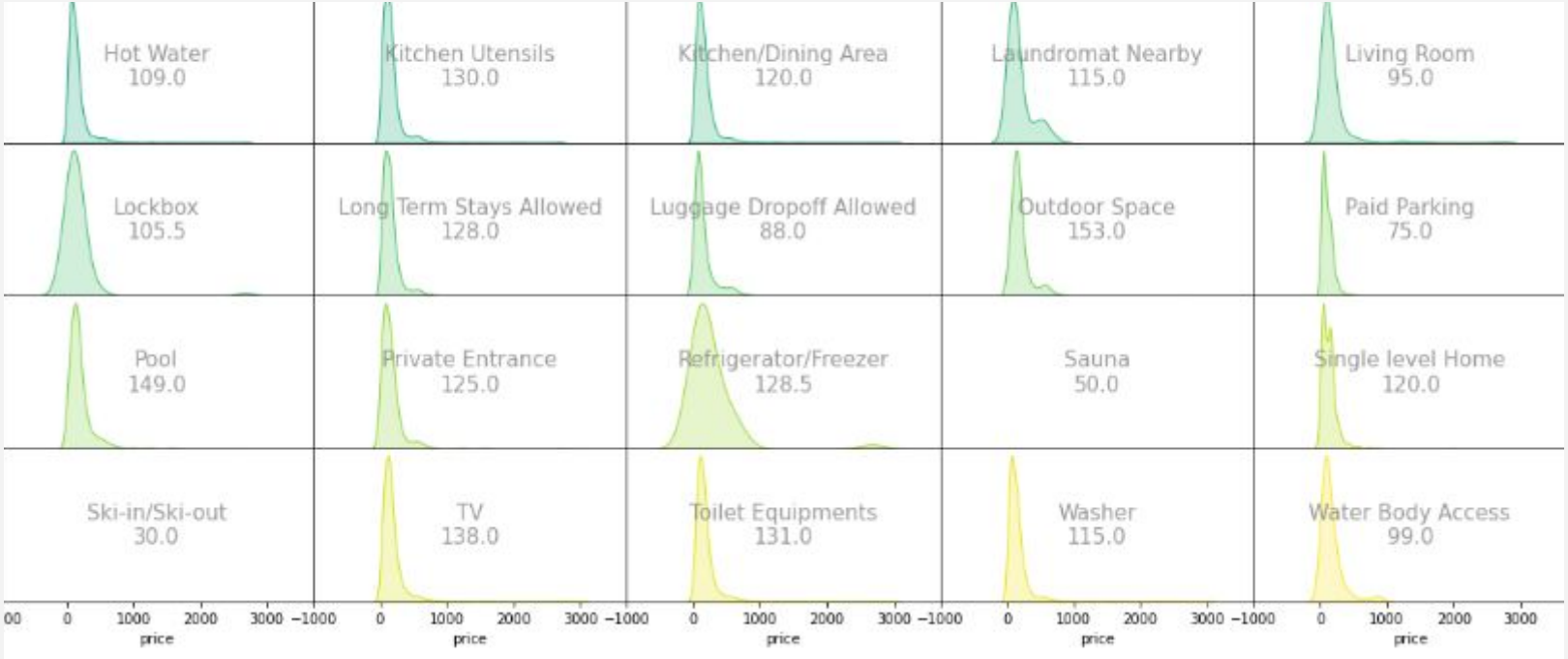
Count Plot

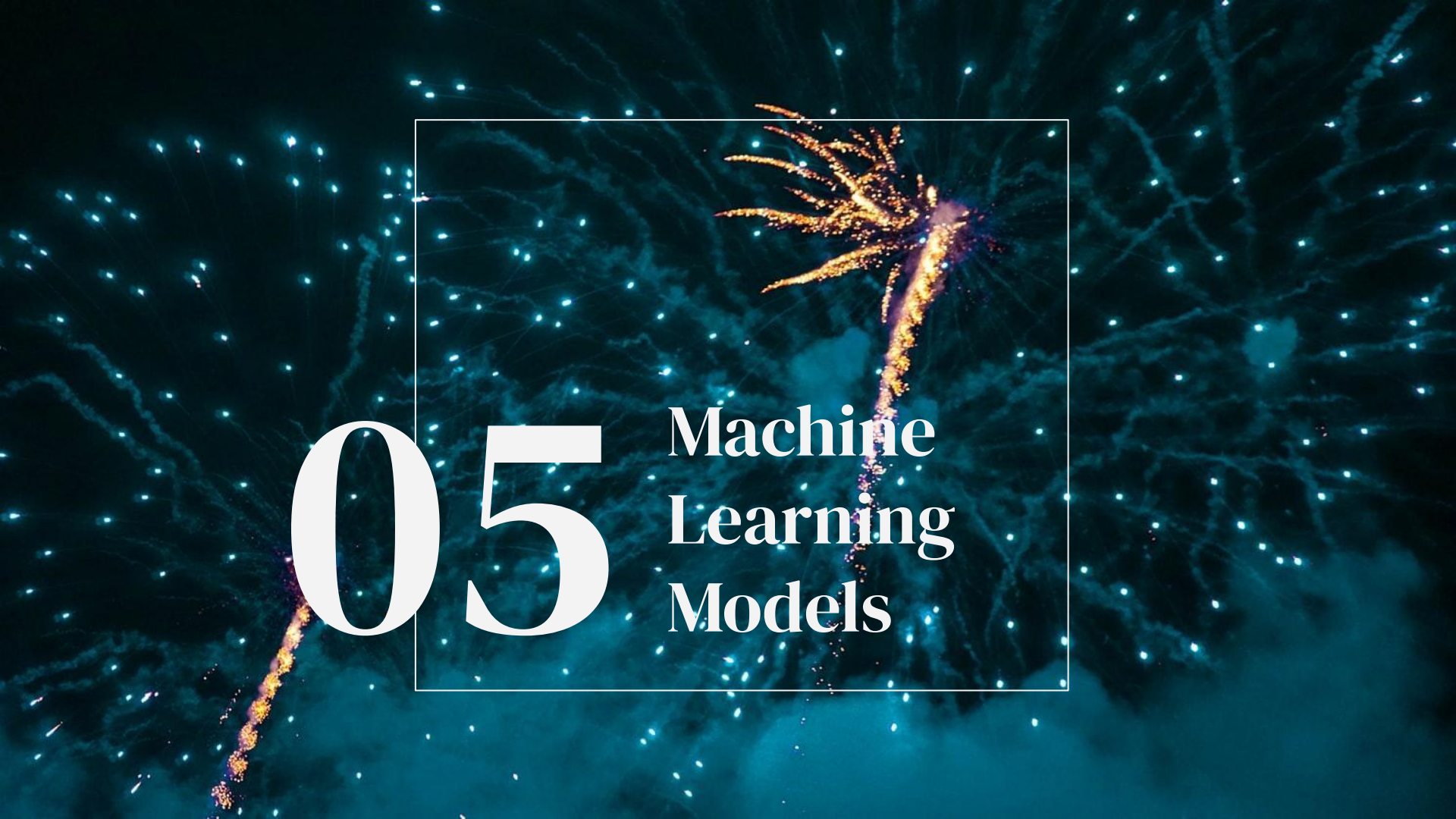


Price and Amenities

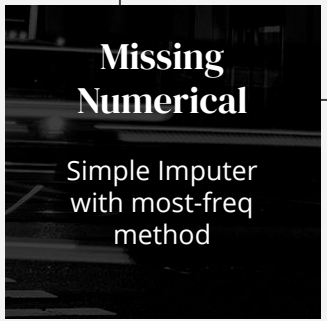


Price and Amenities



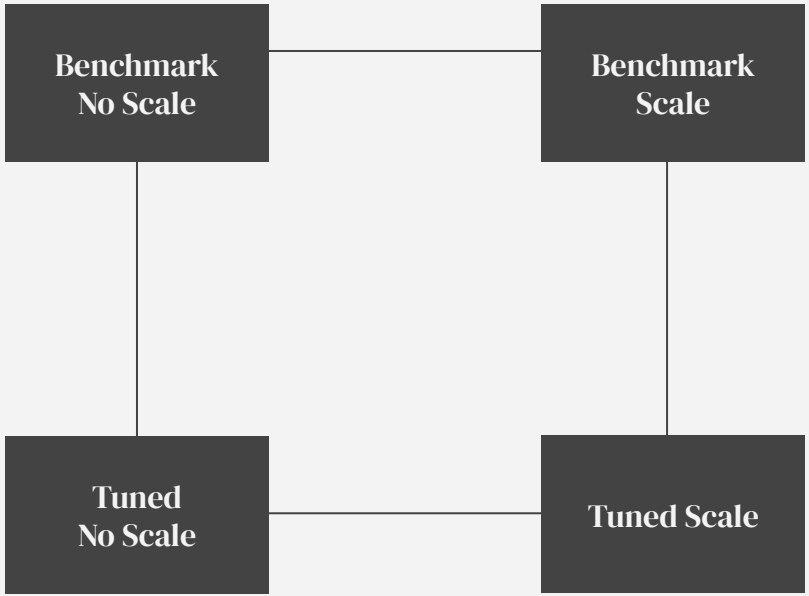

05
Machine
Learning
Models

Column Transformer



Model Pipeline

Column Transformer
Regression Model
Transformed Target Regressor
cross_val_score using RMSE



Column Transformer
Regression Model
Robust Scaler
Transformed Target Regressor
cross_val_score using RMSE

Column Transformer
Regression Model
Transformed Target Regressor
RandomizedSearchCV 50 iterations

Column Transformer
Regression Model
Transformed Target Regressor
RandomizedSearchCV 50 iterations

Machine Learning Regression Models

Cross Validation Result

	Ridge	Lasso	Elastic Net	Decision Tree	KNN	Random Forest	XGB
Benchmark_NoScale	153.949783	186.734405	184.544254	177.997795	159.456351	145.183703	143.462260
Benchmark_Scale	152.794429	193.252952	190.019028	177.999686	150.664259	145.142675	143.466311
Tuned_NoScale	152.593847	154.860982	154.621594	155.512006	153.575697	145.960197	141.993254
Tuned_Scale	152.591324	153.611014	153.012793	155.510940	146.111794	145.957072	141.988715

Train Dataset Prediction Result

Model Variation	Ridge	Lasso	Elastic Net	Decision Tree	KNN	Random Forest	XGB
Benchmark_NoScale	150.702283	187.389928	185.197649	0.058581	144.233912	92.267116	40.443399
Benchmark_Scale	149.548602	193.961983	190.878637	0.058581	137.875329	92.266141	40.443399
Tuned_NoScale	149.309731	152.313516	151.661226	131.024848	0.058581	104.137009	24.264043
Tuned_Scale	149.307406	151.004787	149.981613	131.024848	0.058581	104.143985	24.264043

Train Dataset Prediction Result

Model Variation	Ridge	Lasso	Elastic Net	Decision Tree	KNN	Random Forest	XGB
Benchmark_NoScale	150.702283	187.389928	185.197649	0.058581	144.233912	92.267116	40.443399
Benchmark_Scale	149.548602	193.961983	190.878637	0.058581	137.875329	92.266141	40.443399
Tuned_NoScale	149.309731	152.313516	151.661226	131.024848	0.058581	104.137009	24.264043
Tuned_Scale	149.307406	151.004787	149.981613	131.024848	0.058581	104.143985	24.264043

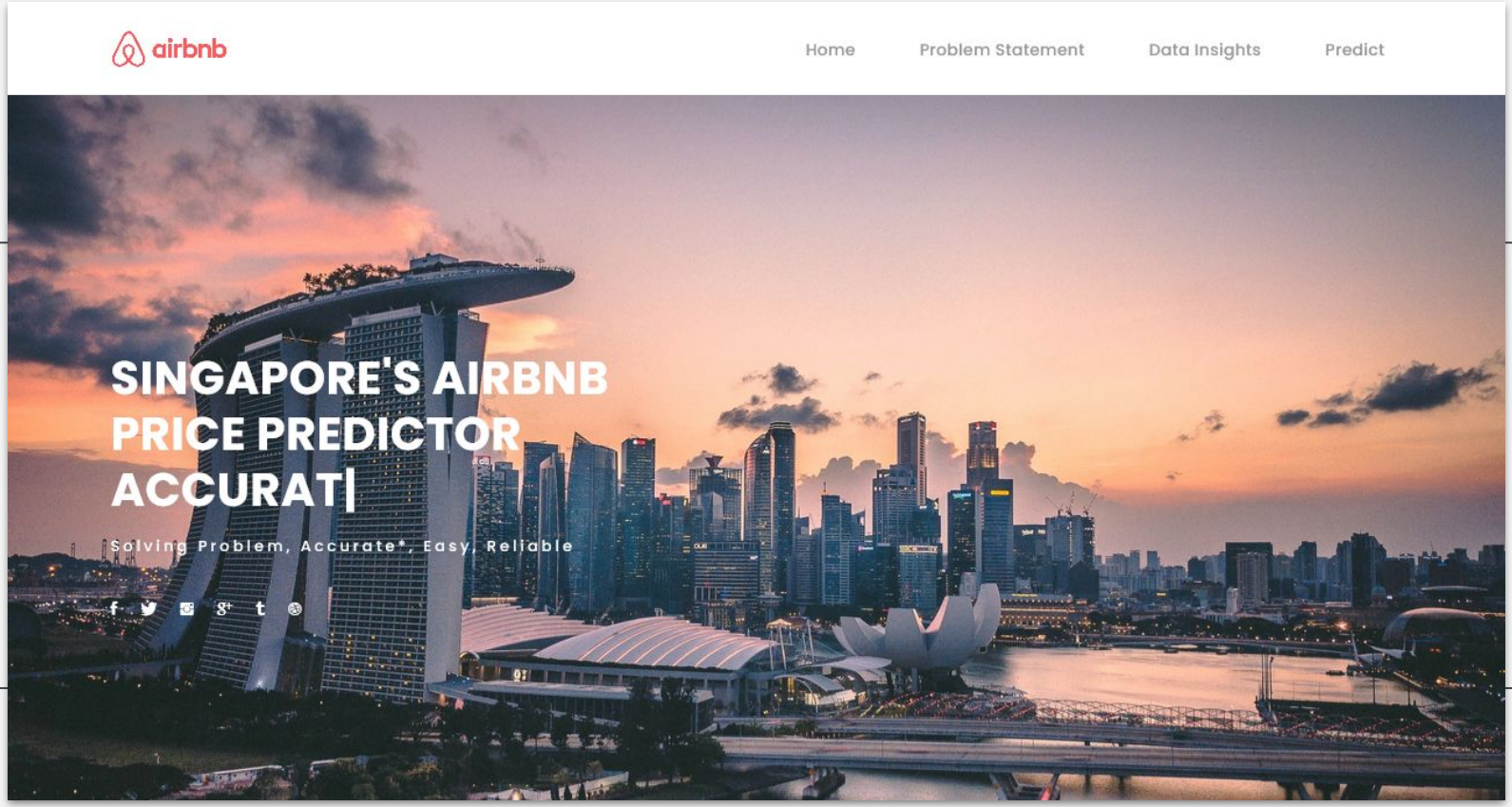
Test Dataset Prediction Result

Model Variation	Ridge	Lasso	Elastic Net	Decision Tree	KNN	Random Forest	XGB
Benchmark_NoScale	159.498346	201.180518	199.812053	188.900339	167.376372	144.288902	134.651870
Benchmark_Scale	160.560307	205.929145	203.527709	188.900339	156.939328	144.288721	134.634701
Tuned_NoScale	159.693977	161.462637	160.035033	164.104908	143.968356	148.335935	133.756034
Tuned_Scale	159.697172	161.877570	161.313851	164.104908	145.625439	148.327971	133.753524

Choosing Best Model

After fitting, training and scoring each models we get each RMSE score for cross validation, train dataset and train dataset scenario. Decision Tree and KNN model had exquisitely good performance with RMSE score less than 1 in train dataset scenario but those did worse in test dataset scenario which may indicate overfitting. Hence, model that has best result in all scenario would be **XGB Model** and we will use this model as our prediction model

Dashboard Preview



Dashboard Preview

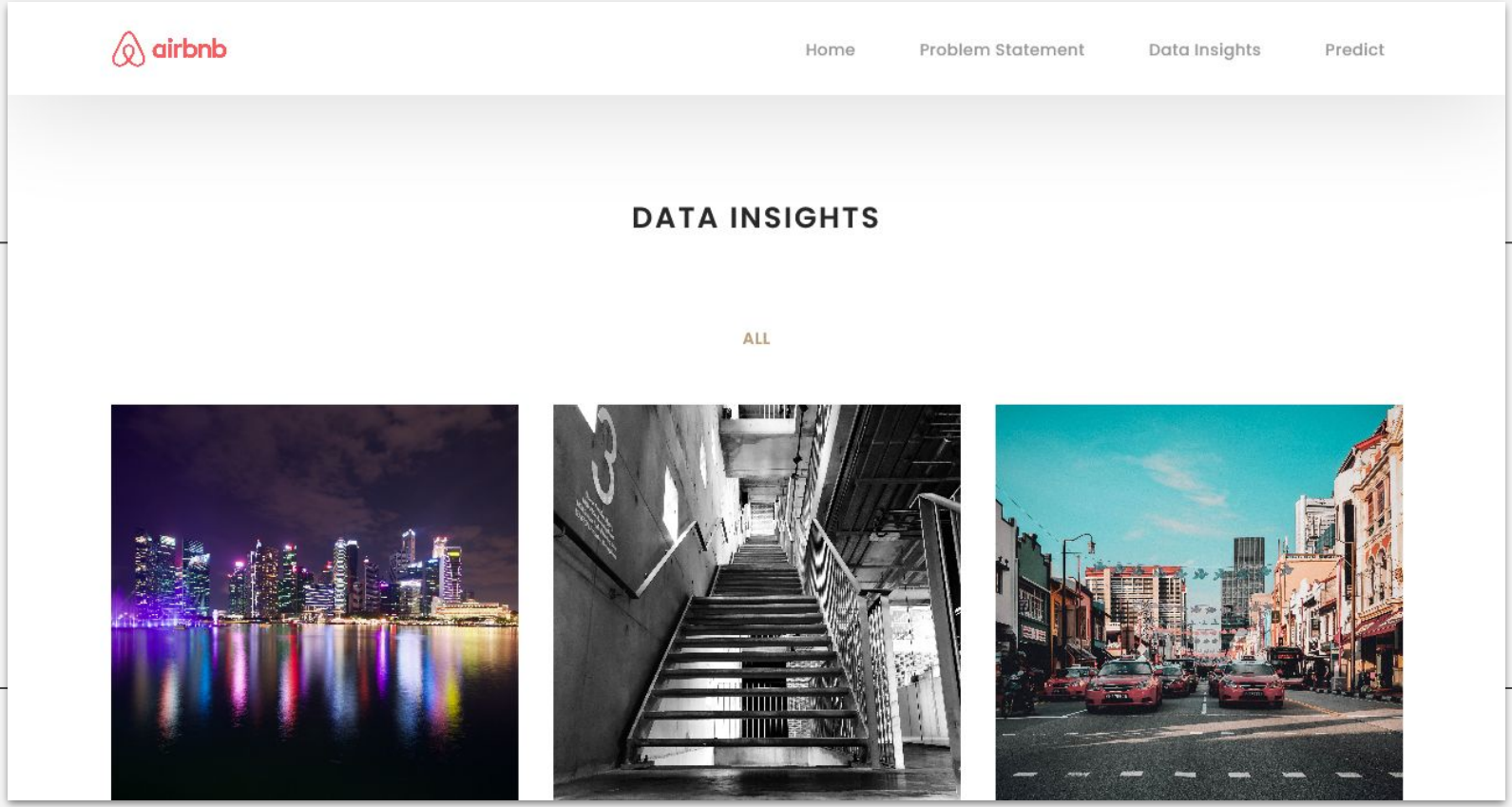
airbnb

Home Problem Statement Data Insights Predict

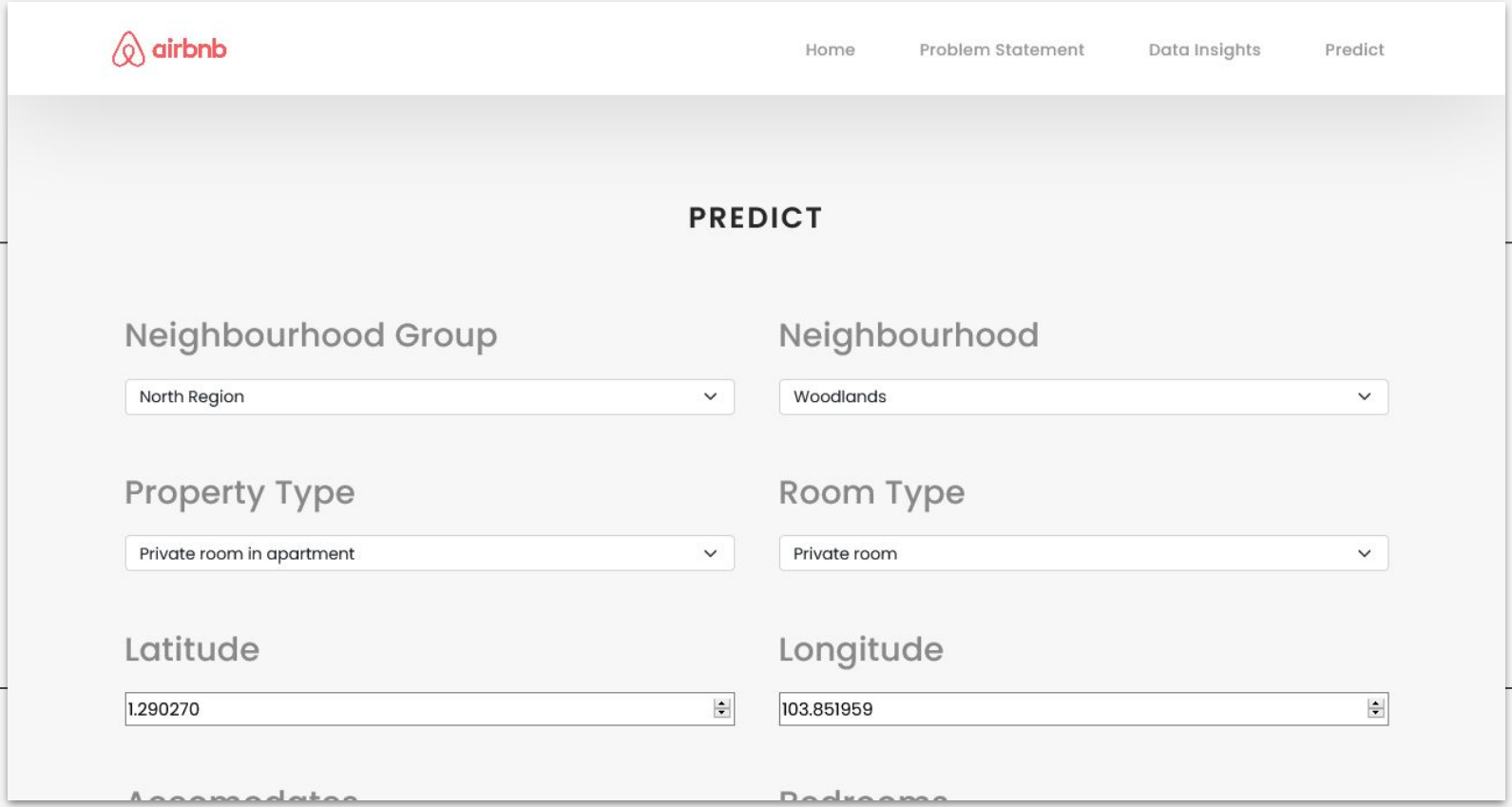
COVID-19 Pandemic

One year has passed since COVID-19 pandemic spread throughout the world, slowing down the economic growth in all parts of the world. Many countries applied lockdown as their measures to contain the spread of the virus, resulting in slump demand for tourism all over the world. But, recently some countries has lift the lockdown measures as their effort to suppressed the COVID-19 cases. Resulting in thriving economic post-pandemic and increasing demand for tourism.

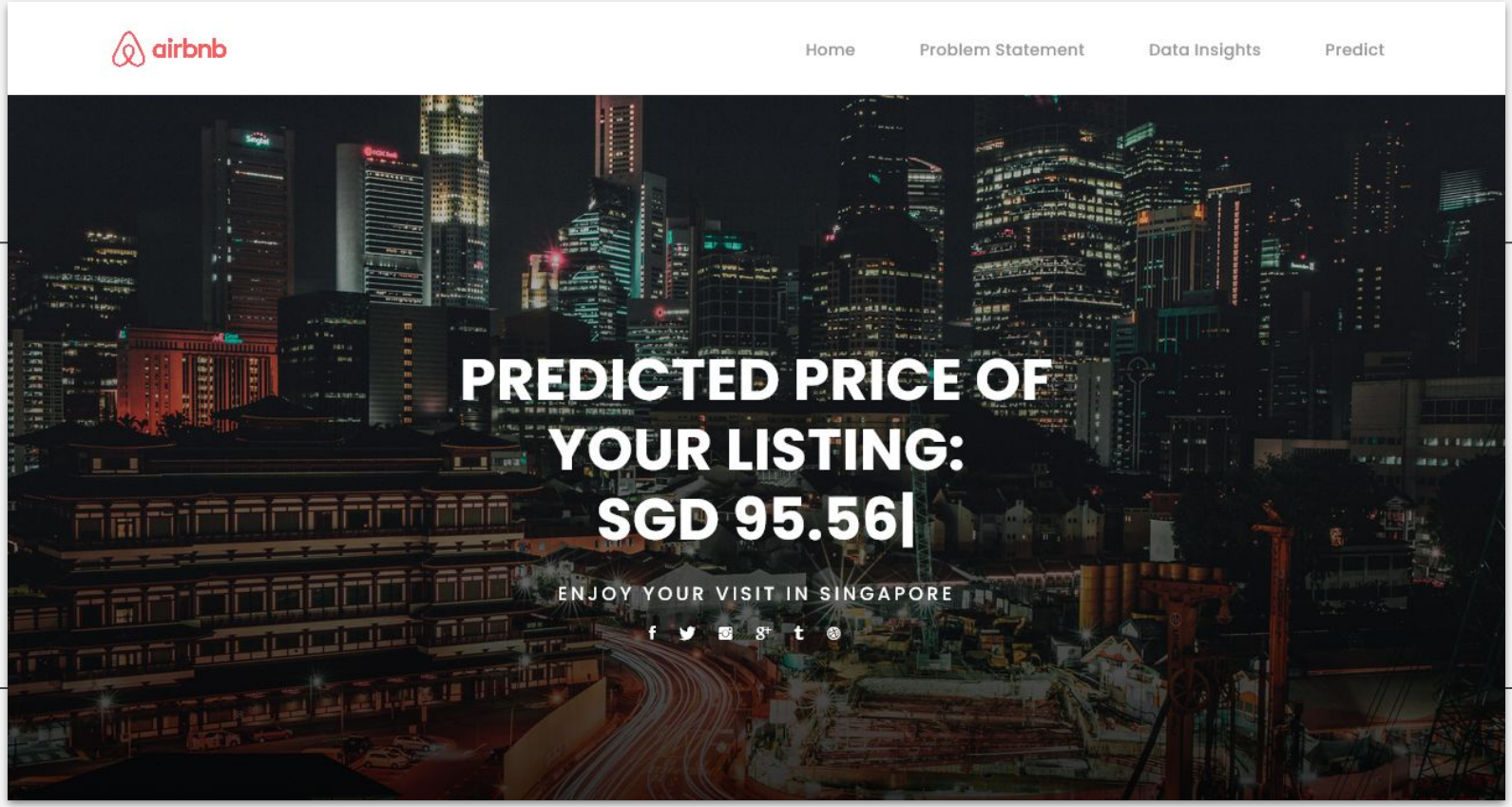
Dashboard Preview



Dashboard Preview



Dashboard Preview



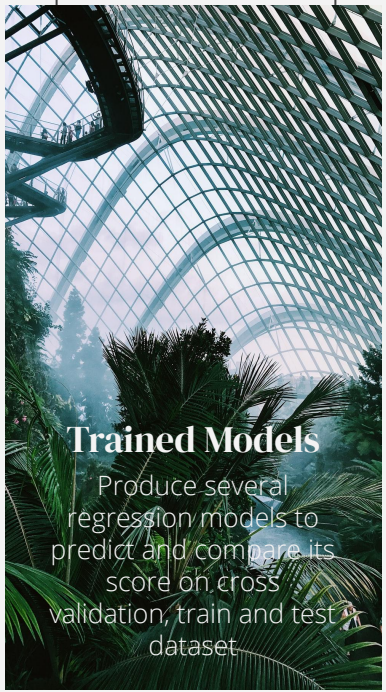
Conclusions And Improvement

06

Conclusions

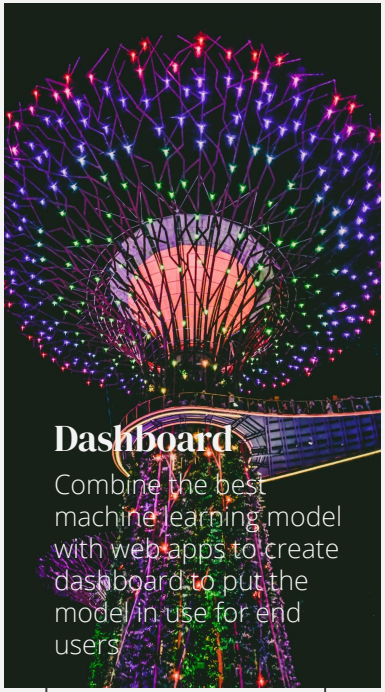
Insights

Get some insights about Singapore's Airbnb listing such as price distribution, listings location and some correlation between features and price.



Trained Models

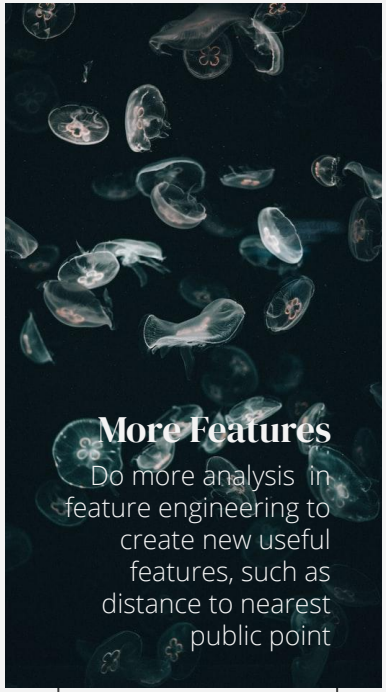
Produce several regression models to predict and compare its score on cross validation, train and test dataset



Dashboard

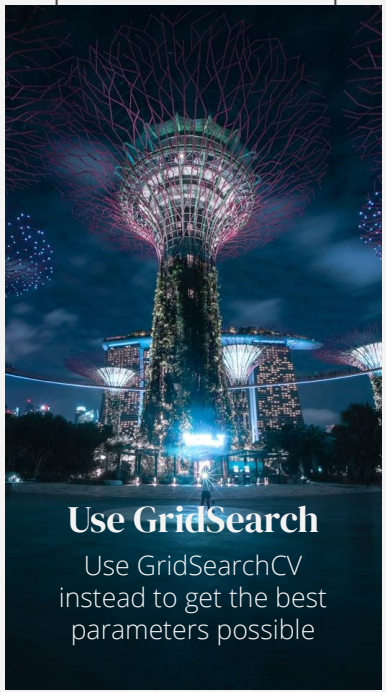
Combine the best machine learning model with web apps to create dashboard to put the model in use for end users

Improvements



More Features

Do more analysis in feature engineering to create new useful features, such as distance to nearest public point



Use GridSearch

Use GridSearchCV instead to get the best parameters possible

More Analysis

Use other datasets to get more insights about Airbnb listing in Singapore

Thanks

Does anyone have any questions?

yosiadityan@gmail.com
linkedin.com/in/@yosiadityan
github.com/yosiadityan/sg-airbnb

Credits

This is where you give credit to the ones who are part of this project.

- Presentation template by Slidesgo
- Icons by Flaticon
- Infographics by Freepik
- Images created by Freepik
- Author introduction slide photo created by Freepik
- Text & Image slide photo created by Freepik