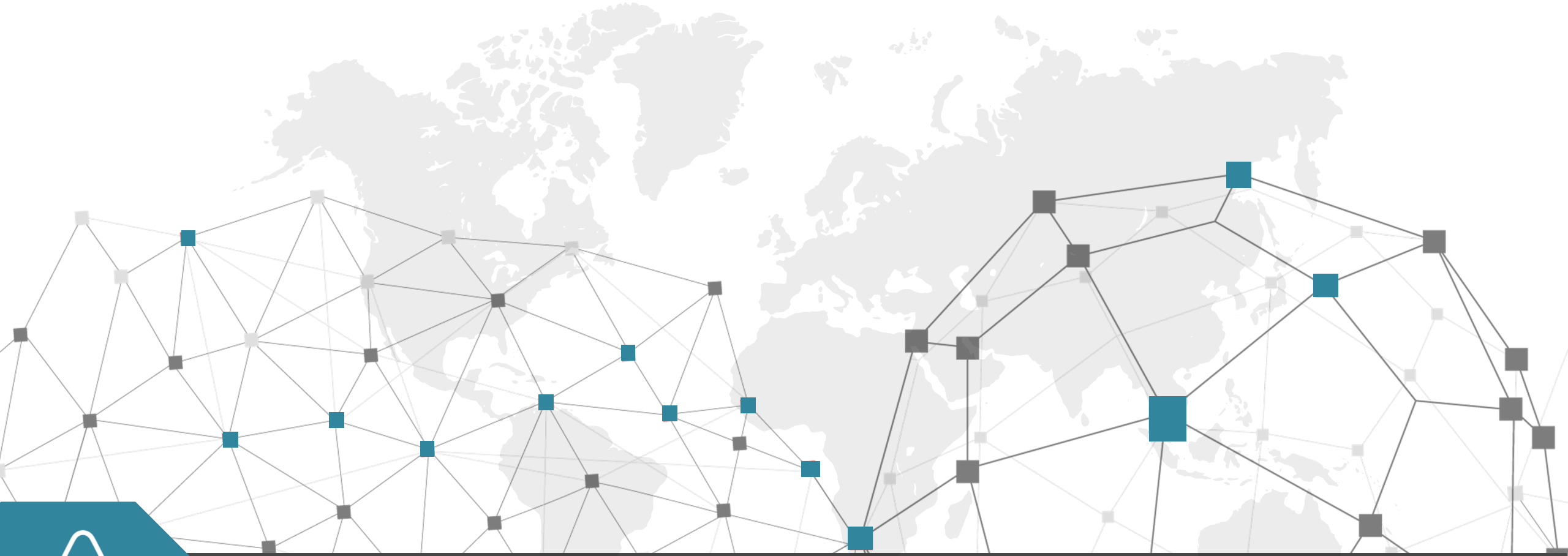
# IS5126 HANDS-ON WITH APPLIED ANALYTICS:
# KEY FACTORS AFFECTING THE PRICING STRATEGY FOR AIRBNB LISTING




Goh Khai Hong
e0503476

Sonakshi Mendiratta
e0444087

Toshal Patel
e0403964

Vignesh Thangaraju
e0503517

# Overview

NUS
National University
of Singapore

# Introduction & Objectives

# Problem Statement and Project Scope

## What we aim to do

Study factors influencing Airbnb pricing using accommodation attributes along with external factors like macro-economic variables, weather data, etc.

## Importance of Study

- Understand the quantitative impact and relationship of various factors with pricing
- Assist Airbnb hosts make more data-driven decisions in setting listing prices

## How we aim to do

- Enrich traditional data with alternative data sources like weather data, tourism data, listing reviews, host statistics, location factors and macro-economic factors
- Perform descriptive, sentiment, panel and causal analytics to effectively analyse the pricing strategy

**Data Collection & Preparation**

Identify and collect potentially useful information from various data sources. Integrate the disparate sources of data

**Exploratory Data Analysis**

Analyse the data set to better understand the various attributes, uncover previously hidden insights and identify the outliers

**Panel and Casual Analysis**

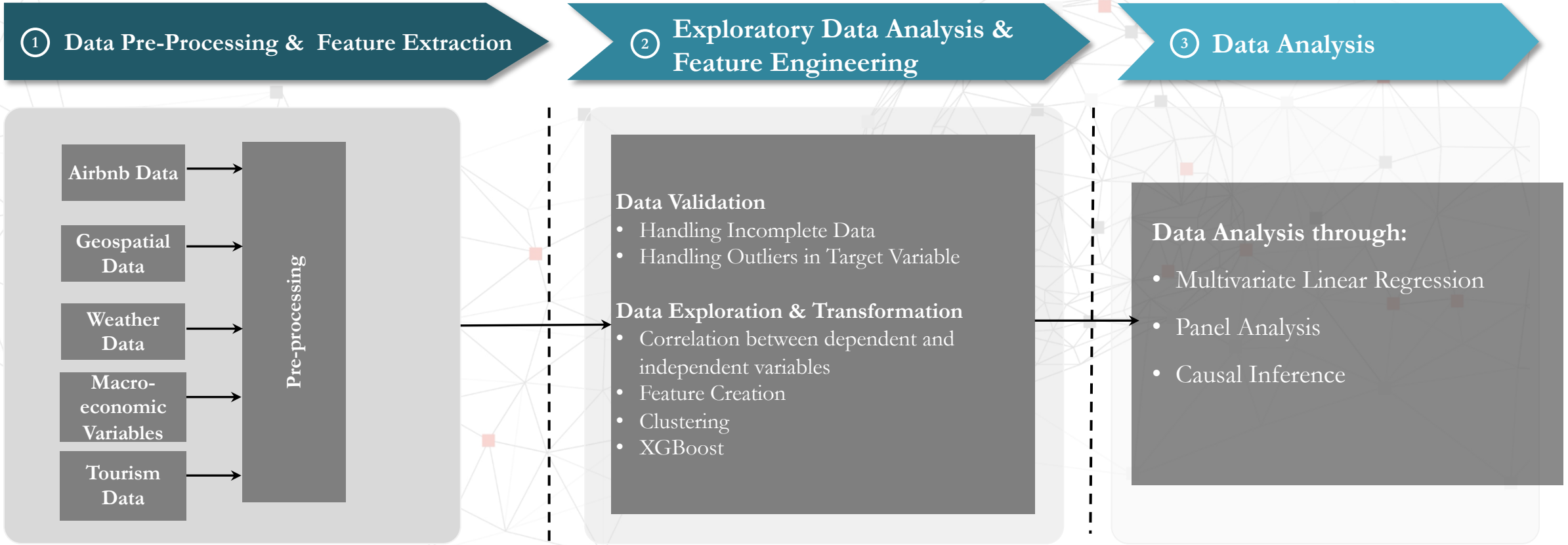
The panel analysis helps to understand the influence of time varying and time invariant variables. It is our first step towards causal analysis/interpretation

NUS National University of Singapore

# Approach

# Approach Implemented

The following framework was used to pre-process data & study the Airbnb pricing



NUS National University of Singapore

# Data & Sources

## Datasets used for Analytics

# Data Sources for Analytics

### Inside Airbnb

Built a crawler and scrape the Airbnb datasets from Inside Airbnb website (in csv)


### Economist Intelligence Unit

Get the monthly/quarterly macro-economic datasets from EIU website (in csv format)


### International Monetary Fund

Obtained the monthly/quarterly macro-economic datasets from IMF website (in csv format)


### Data.gov.sg

Downloaded the monthly public weather datasets from Data.gov.sg (in csv format)


### OpenMap API

Developed a script to search and count the nearby amenities based on the latitude and longitude of the listing(within 200m &1000m)


### Singapore Tourism Analytics Network

Retrieved the monthly hotel statistics datasets from STAN website (in csv format)

NUS
National University
of Singapore

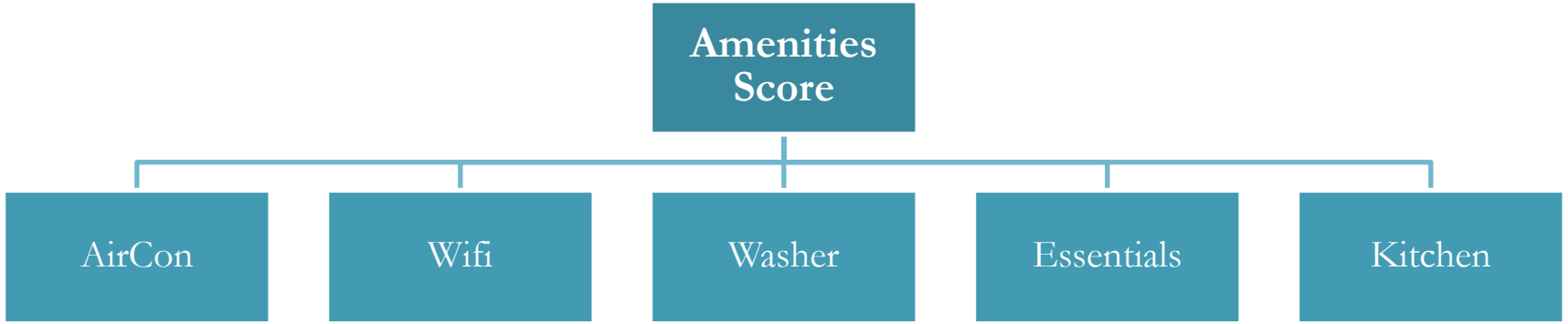
# Data Pre-processing & EDA

# Feature Engineering

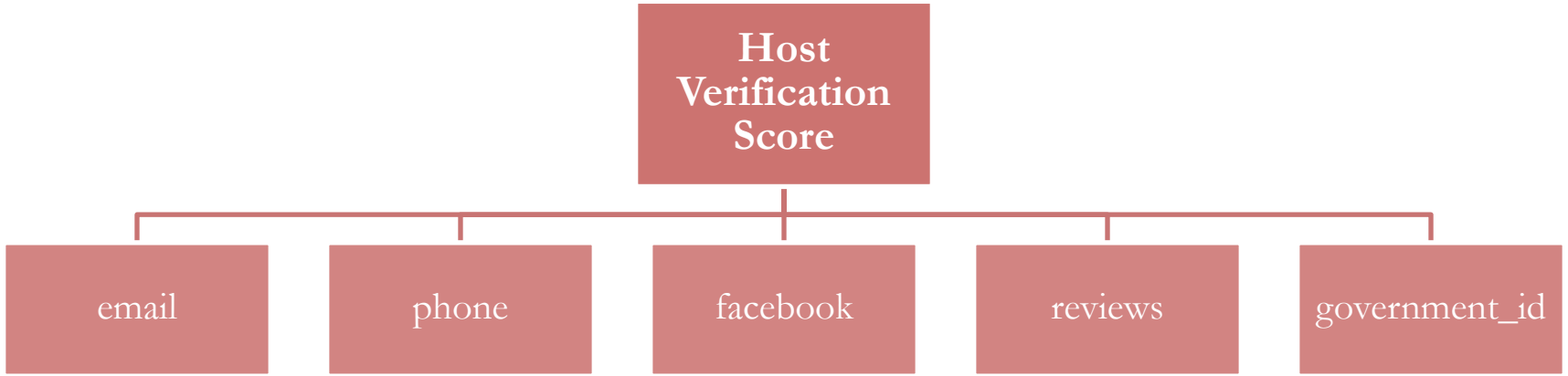
| Feature Category | Features | Data Description |
|---|---|---|
| **Listing features** | Bedrooms, bathrooms, beds, room_type, property_type, amenities, minimum_nights, accomodates, availability_30, reviews, review_scores, instant_bookable, cleaning_fee, security deposit, price | These features describe each Airbnb listing |
| **Host features** | Host_neighbourhood, host_acceptance_rate, host_response_rate, is_host_superhost, host_total_listings_count, is_host_verified | These features describe the host of the Airbnb listing |
| **Geospatial Data** | Subway_count_within_200m, Subway_count_within_1000, … Bus, Shops, Restaurants, Attractions | These features give a count of busstops, MRT stations, shops, restaurants and tourist attractions within 200m and 1000m radius of the Airbnb listing |
| **Weather Data** | total_rainfall, mean_humidity, mean_temp | This gives an overview of the weather variables for the month |
| **Macro-Economic Variable** | Nominal GDP % Change YOY, Unemployment Rate, Consumer Prices Exchange rate S$:US$ (av), SG Straits Times Index, | These features describe Singapore's macro-economic factors |

# Feature Engineering - Scores

**Top 5 amenities found in Singapore Airbnb Listings**




**Top 5 verification modes of Singaporean Hosts of Airbnb**

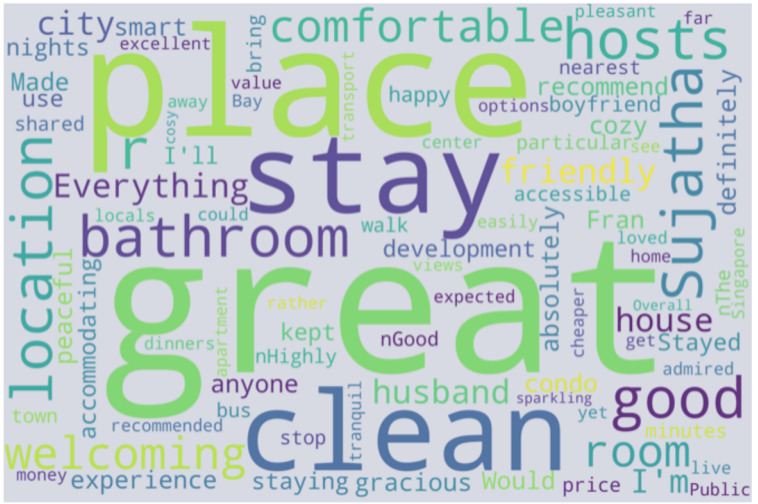


# Feature Engineering - Sentiment Analysis on Reviews

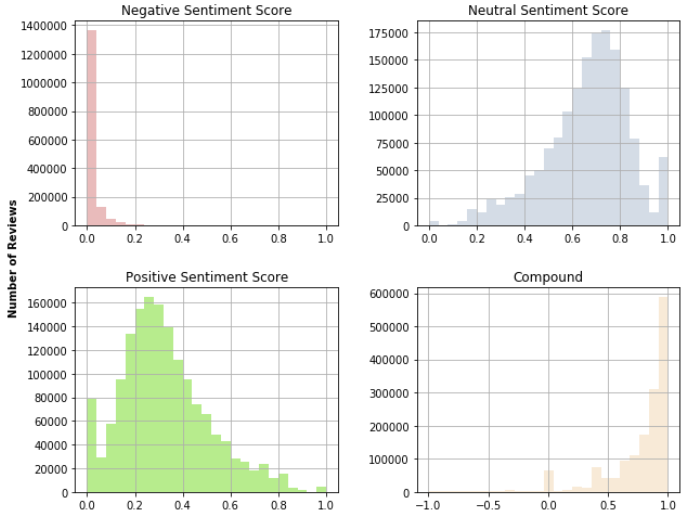
**V**alence **A**ware **D**ictionary for s**E**ntiment **R**easoning (VADER) model is used for sentiment analysis. NLTK package is leveraged in Python to perform sentiment analysis on the Airbnb Reviews. The output of the VADER model is the negative, neutral and positive strength of the review as well as a compounded score calculated based on the various sentiment strengths.

Average Sentiment Scores were computed for each month per listing

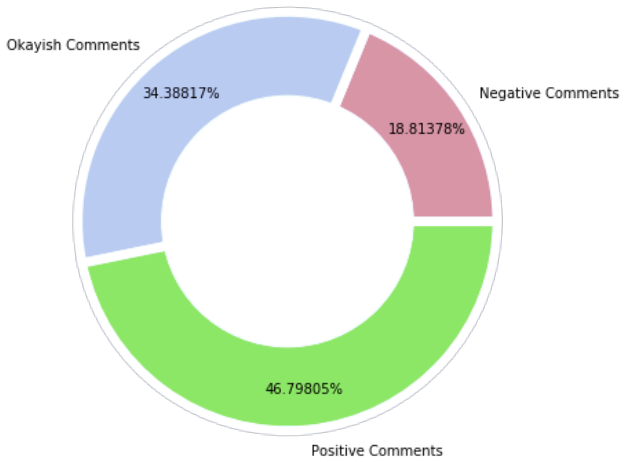
**Word Cloud on Airbnb Reviews**



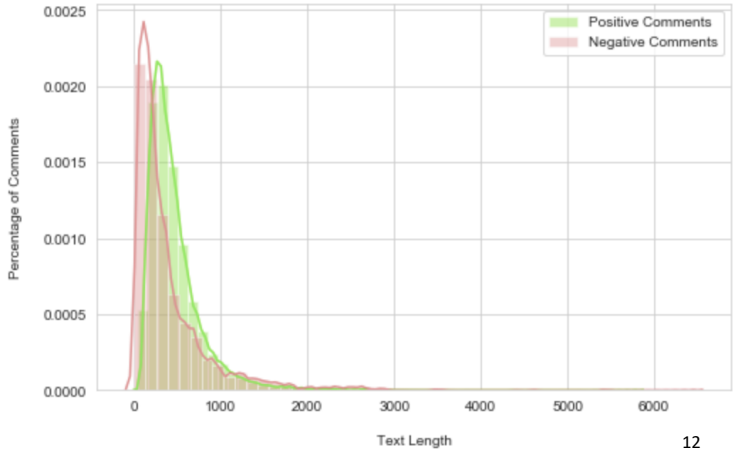

**Sentiment Scores**



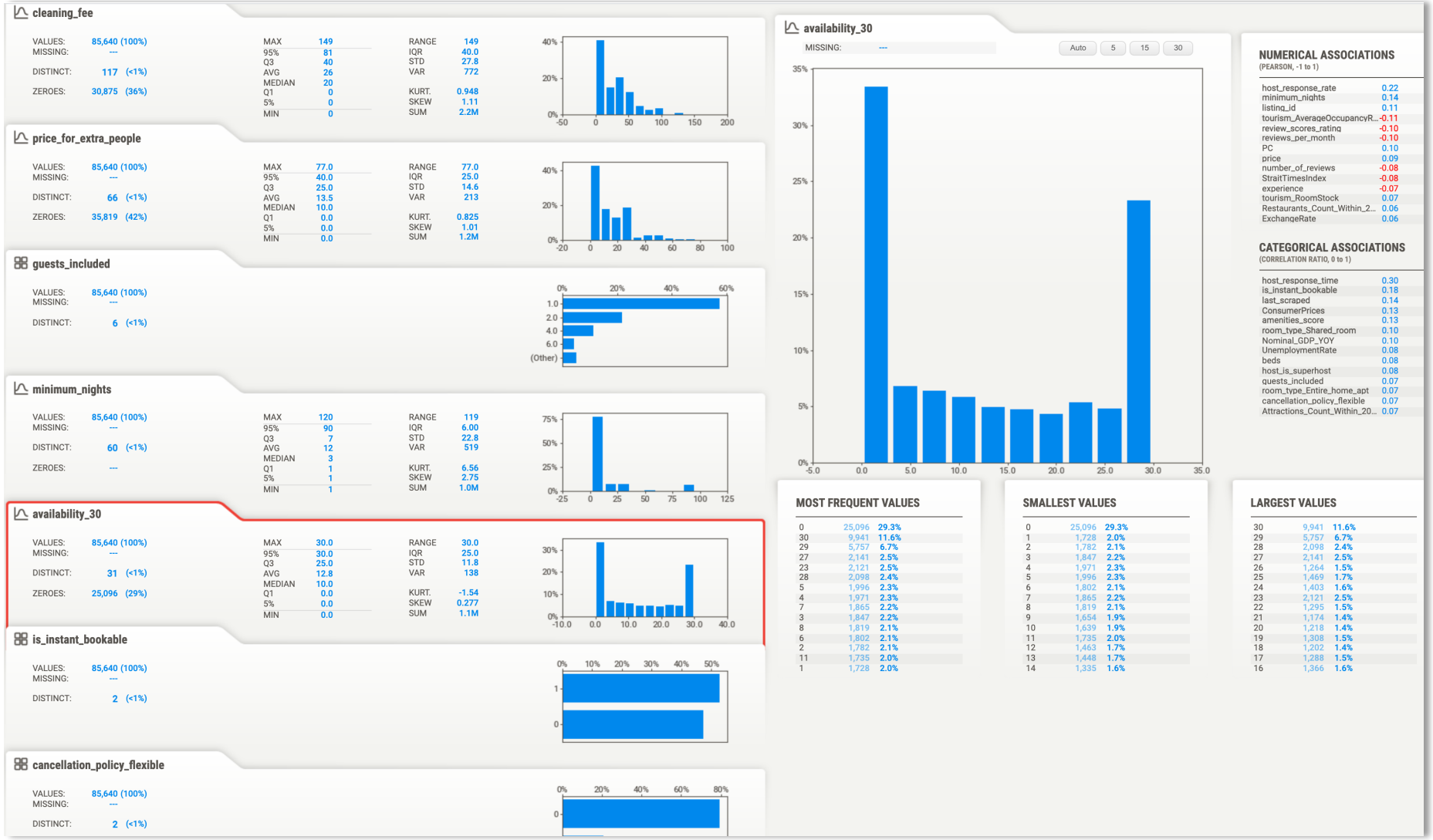

**Distribution of Sentiments**




**Distribution Plot For Length of Comments**



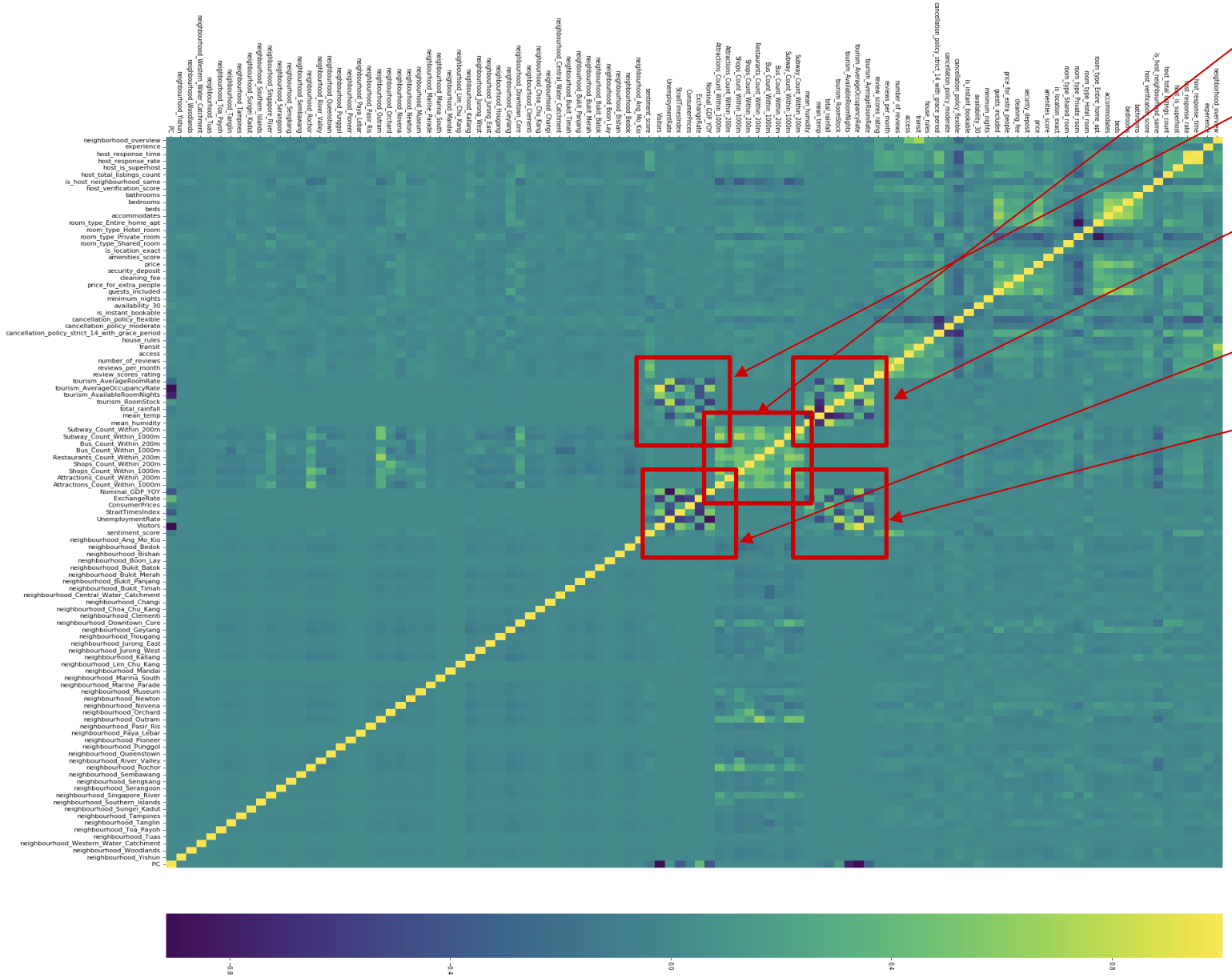
# Exploratory Data Analysis





- Used SweetViz open source Python library to kickstart the EDA
- Visualised and analysed the min, max, average, median and quartiles of each feature
- Studied the range, variance and standard deviation
- Removed outliers for Price variables (price, cleaning fee, etc.), no. of beds, bathrooms, etc.

# Exploratory Data Analysis - Correlation



Geospatial Data

Macro-economic variables VS tourism and reviews

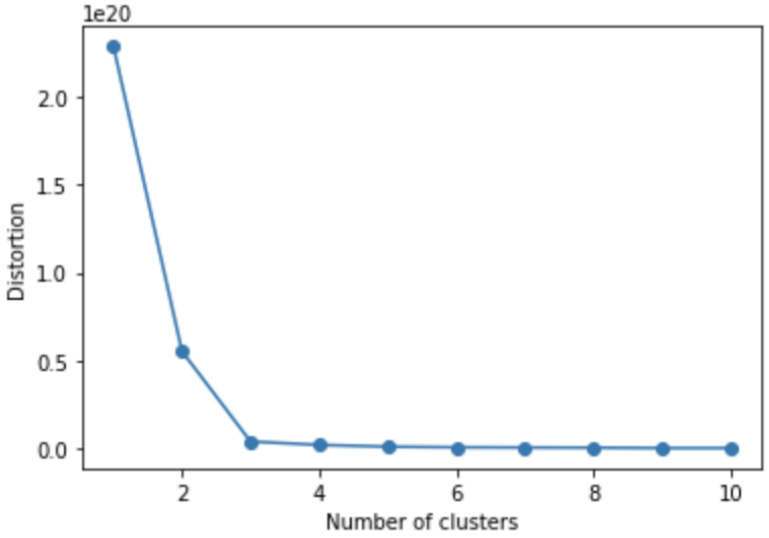
Weather and tourism features VS reviews and tourism

Macro-economic variables

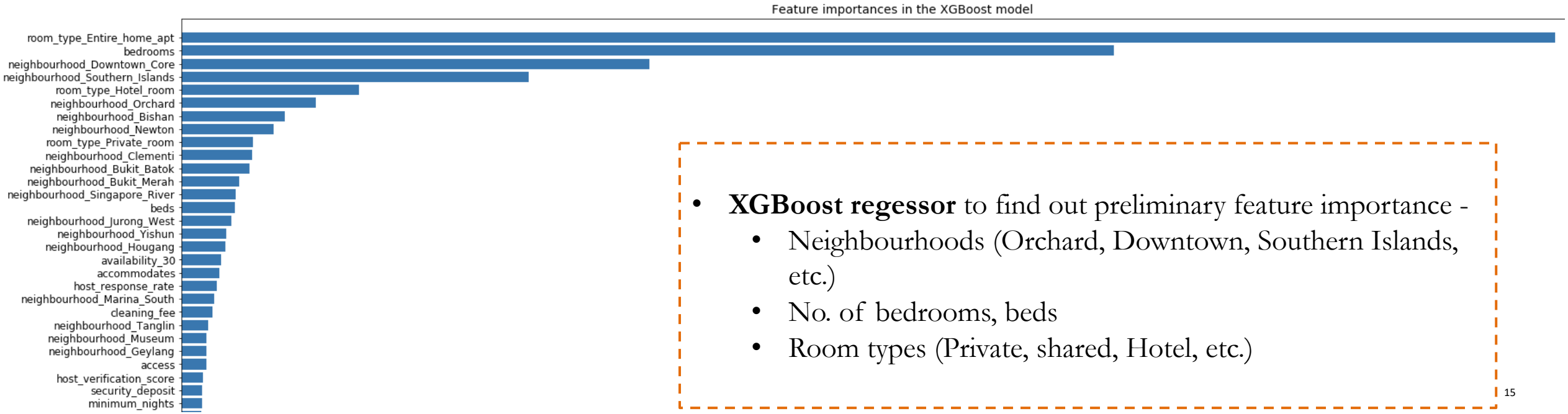
Macro-economic VS weather and tourism

- Plotted correlation matrix to identify the highly correlated variables
- Tourism data pertaining to the hotel occupancy was highly correlated with the visitors in Singapore for that month
- Multicollinearity was resolved using PCA

# Exploratory Data Analysis – Feature Importance




- **KMeans** clustering – found 3 clusters for listings
  - High priced, mid-priced, low-priced



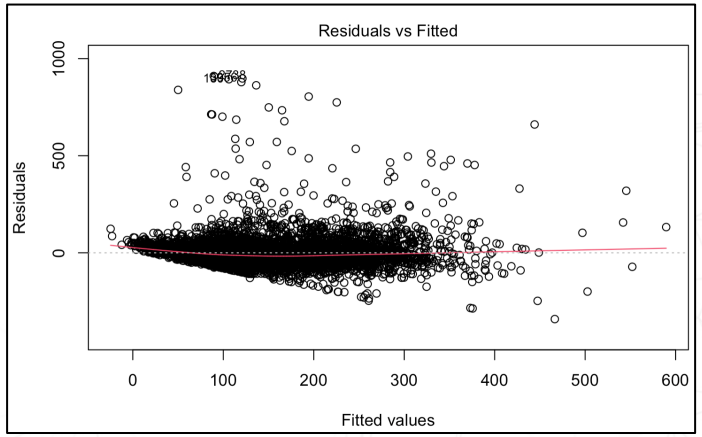

- **XGBoost regessor** to find out preliminary feature importance -
  - Neighbourhoods (Orchard, Downtown, Southern Islands, etc.)
  - No. of bedrooms, beds
  - Room types (Private, shared, Hotel, etc.)

NUS National University of Singapore

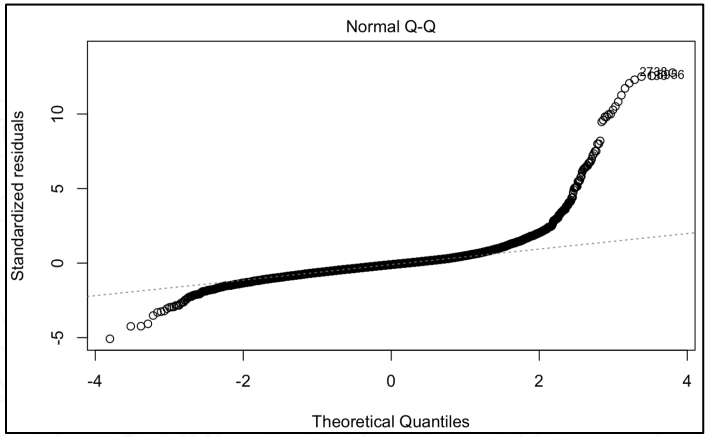
# Multivariate Regression

# Multivariate Regression on cross-sectional data

- Selected cross-sectional Airbnb data for October 2019
- Implemented stepwise linear regression with all the features to obtain model with optimal AIC results
- Tested the OLS Assumptions
- Resulted in a model with most relevant features that showed relatively robust statistical relationship with price of listings




**Residual Chart – Validates mean zero error assumption**




**Normality exists in the middle ranges and deviation is seen at the ends**

## Features Selected

- neighborhood_overview
- experience
- host_response_rate
- host_is_superhost
- bathrooms
- Bedrooms
- Accommodates
- room_type
- amenities_score
- minimum_nights
- availability_30
- cancellation_policy
- reviews_per_month
- review_scores_rating
- Subway_Count_Within_1000m
- Bus_Count_Within_1000m
- Restaurants_Count_Within_200m
- Shops_Count_Within_1000m
- neighbourhood

NUS National University of Singapore

# Panel Analysis with Fixed Effect and Random Effect Models

# Fixed Effect Model for Panel Analysis

For panel analysis, the fixed effect model is used, and the statistics are shown below:

```
Call:
plm(formula = price ~ experience + host_is_superhost + security_deposit +
    cleaning_fee + availability_30 + number_of_reviews + reviews_per_month +
    review_scores_rating + total_rainfall + mean_temp + mean_humidity +
    ConsumerPrices + Nominal_GDP_YOY + ExchangeRate + StraitTimesIndex +
    UnemploymentRate + sentiment_score,
    data = pdata,
    effect = "twoways", model = "within",
    index = c("listing_id", "last_scraped"))

Unbalanced Panel: n = 12471, T = 1-12, N = 85640

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-7.5375e+02 -6.8479e-01 -6.3808e-03  7.0685e-01  3.2662e+02
```

No time-variant variables in FE

```
Coefficients:
                       Estimate Std. Error t-value  Pr(>|t|)
experience           -0.1527116  0.2304338 -0.6627 0.5075162
host_is_superhost    -0.0494391  0.2350932 -0.2103 0.8334375
security_deposit      0.0010259  0.0011480  0.8936 0.3715465
cleaning_fee          0.0571436  0.0077202  7.4018 1.358e-13 ***
availability_30       0.0204300  0.0061402  3.3273 0.0008774 ***
number_of_reviews     0.0517190  0.0103562  4.9940 5.928e-07 ***
reviews_per_month    -0.3226263  0.1368337 -2.3578 0.0183863 *
review_scores_rating -0.0117558  0.0029514 -3.9831 6.810e-05 ***
sentiment_score      -0.4298614  0.1717613 -2.5027 0.0123283 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    8366400
Residual Sum of Squares: 8349200
R-Squared:      0.0020618
Adj. R-Squared: -0.16833
F-statistic: 16.7923 on 9 and 73149 DF, p-value: < 2.22e-16
```

# Random Effect Model for Panel Analysis

The random effect model is also used, and *Hausman Test* is performed to compare the *fixed effect* vs *random effect* model :

```
Call:
plm(formula = price ~ experience + host_is_superhost + security_deposit +
    cleaning_fee + availability_30 + number_of_reviews + reviews_per_month +
    review_scores_rating + total_rainfall + mean_temp + mean_humidity +
    ConsumerPrices + Nominal_GDP_YOY + ExchangeRate + StraitTimesIndex +
    UnemploymentRate + sentiment_score,
    data = pdata,
    model = "random",
    index = c("listing_id", "last_scraped"))

Unbalanced Panel: n = 12471, T = 1-12, N = 85640

Effects:
                 var std.dev share
idiosyncratic  114.15   10.68 0.012
individual    9748.54   98.73 0.988
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.8924  0.9591  0.9688  0.9614  0.9688  0.9688

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-732.55   -3.10   -1.14   -0.04    1.92  329.92
```

Time-variant variables in RE

```
Coefficients:
                      Estimate  Std. Error z-value  Pr(>|z|)
(Intercept)         95.44144426 14.31104218  6.6691 2.574e-11 ***
experience          -1.26294301  0.20284624 -6.2261 4.782e-10 ***
host_is_superhost    0.16663552  0.23768937  0.7011  0.483263
security_deposit     0.00322029  0.00113606  2.8346  0.004588 **
cleaning_fee         0.11964625  0.00763215 15.6766 < 2.2e-16 ***
availability_30      0.02407110  0.00617620  3.8974 9.723e-05 ***
number_of_reviews    0.02869443  0.01008084  2.8464  0.004421 **
reviews_per_month   -0.22961330  0.13777118 -1.6666  0.095588 .
review_scores_rating -0.01763165  0.00296607 -5.9445 2.774e-09 ***
total_rainfall      -0.00231990  0.00230865 -1.0049  0.314958
mean_temp           -0.43642962  0.33164707 -1.3159  0.188192
mean_humidity       -0.02800743  0.02429443 -1.1528  0.248979
ConsumerPrices       1.91647512  0.37847360  5.0637 4.112e-07 ***
Nominal_GDP_YOY      1.63667263  0.24013751  6.8156 9.389e-12 ***
ExchangeRate        -0.52168012  9.15430358 -0.0570  0.954555
StraitTimesIndex     0.00052774  0.00094155  0.5605  0.575141
UnemploymentRate    19.84064507  1.98454304  9.9976 < 2.2e-16 ***
sentiment_score     -0.51910296  0.17363181 -2.9897  0.002793 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    10569000
Residual Sum of Squares: 10062000
R-Squared:      0.048045
Adj. R-Squared: 0.047856
Chisq: 606.617 on 17 DF, p-value: < 2.22e-16
```

# Hausman Test: Fixed Effect vs Random Effect Model

The Hausman Test result is shown below:

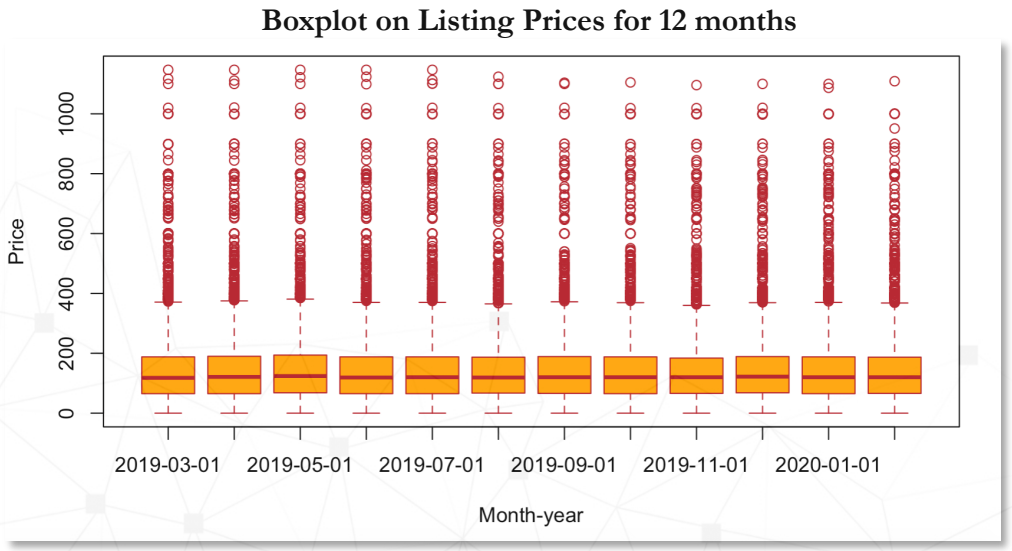
```
Hausman Test

data:  price ~ experience + host_is_superhost + security_deposit + cleaning_fee +  ...
chisq = 2845.6, df = 9, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent
```

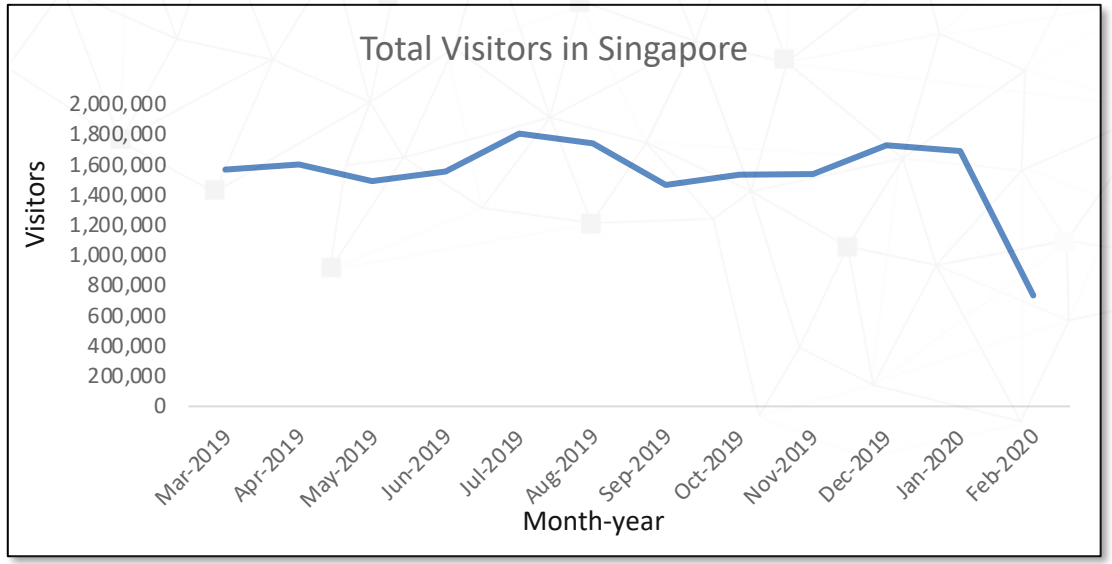
We see that the p-value < 0.05
Hence, reject the null hypothesis, that is,
*"fixed effect $a_i$ is uncorrelated with all covariates for all periods and thus random effect is the proper one".*

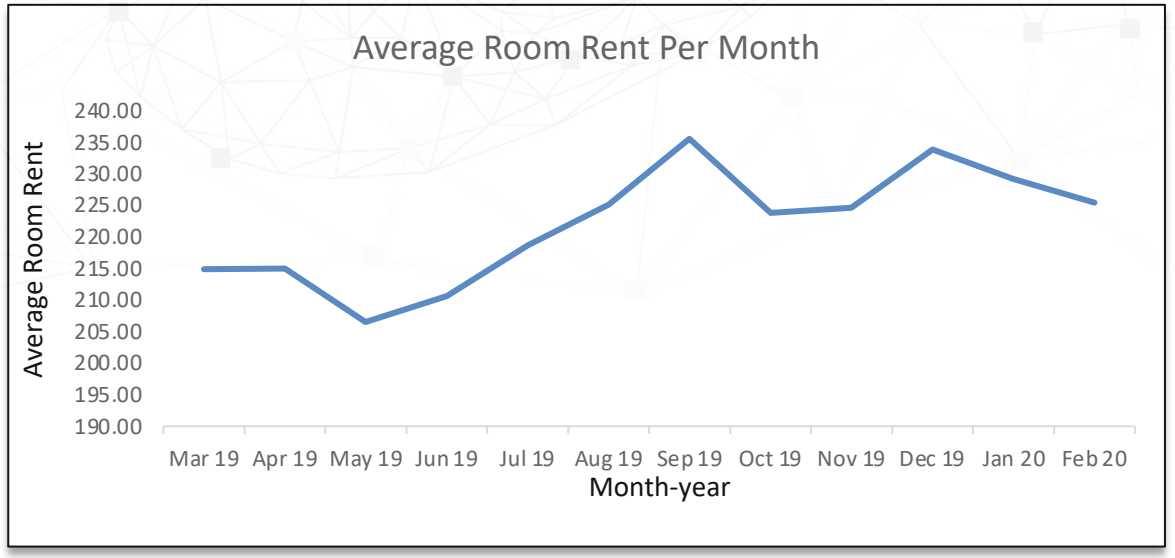
Therefore, we select a **fixed-effect model.**

# Seasonality in Panel Data




- No seasonality reflected in the price for different boxplot values of price.
- Although, we see seasonality in the number of visitors and the pricing for hotel rooms is observed to be seasonal as well



NUS National University of Singapore

# Causal Inference using Instrumental Variables

# Instrumental Variables Method for Causal Inference

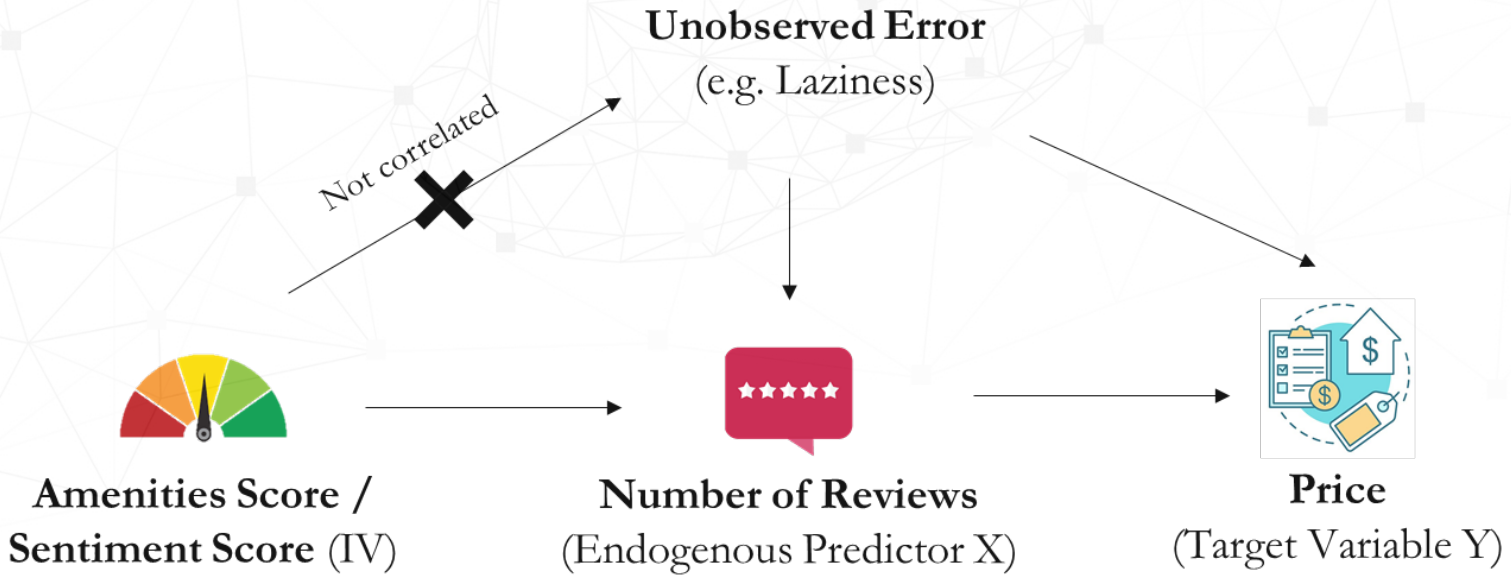
For Causal Inference (ceteris paribus effect), the criteria to pick IV are the following:

- **Relevance** The instrument variable Z must be correlated with X
- **Exclusion** The instrument variable Z cannot be correlated with the unobserved error (omitted confounding variable)

A valid IV breaks endogeneity and establish causal interpretation with its exogenous variation it introduced along X's dimension.

**Amenities Score**
- AirCon
- Wifi
- Washer
- Essentials
- Kitchen

**Sentiment Score:**
- Sentiment Analysis on reviews



# Instrumental Variables Method for Causal Inference

The statistical packages ivreg can be used executing Two Stage Least Square (2SLS):

- The formula of ivreg highlights the two stages: y ~ X + S | X + Z
  a. X are exogenous controls;
  b. S is endogenous variable
  c. Z are plausible IVs

There are 3 tests needed for model specification in 2SLS and IV approach.

- **Weak Instrument Test:** if all plausible instruments are jointly correlated with the endogenous variable
- **Hausman Test for Endogeneity:** if the suspicious variable is endogenous at the first place
- **Over-Identification (Sagan-Hausman) Test:** if all plausible IVs are exclusively exogenous, when we have more IVs than we need and assuming at least one of them is exogenous

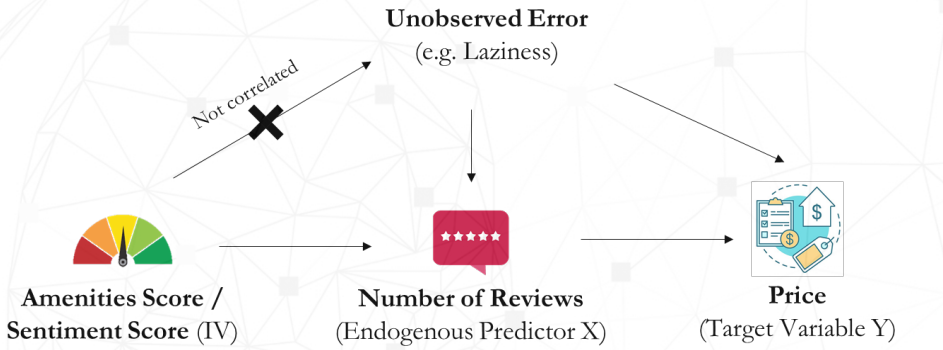
# Instrumental Variables Method for Causal Inference

Specification Test in Two Stage Least Square (2SLS) & Correlation Test:

Causal_Inference.html

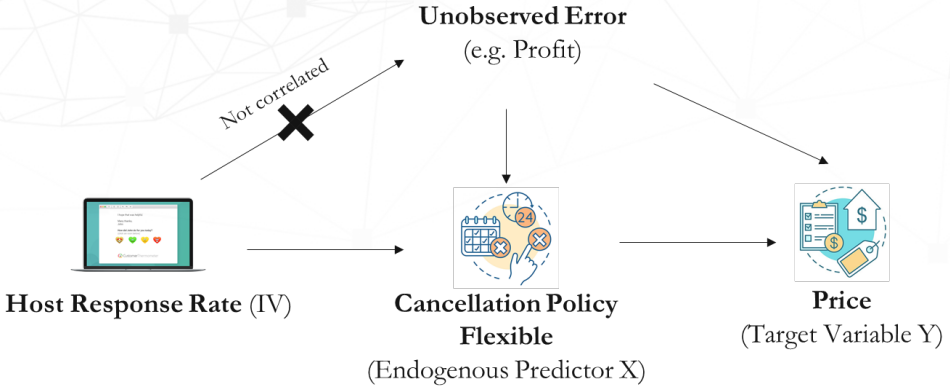
**Case 1:** X = number_of_reviews , $Z_1$ = sentiment_score , $Z_2$ = amenities_score

```
## Diagnostic tests:
##                   df1   df2 statistic p-value
## Weak instruments    2 85540   130.394  <2e-16 ***
## Wu-Hausman          1 85540   189.529  <2e-16 ***
## Sargan              2    NA     0.558   0.756
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.6 on 85541 degrees of freedom
## Multiple R-Squared: -12.51,  Adjusted R-squared: -12.53
## Wald test:  33.9 on 98 and 85541 DF,  p-value: < 2.2e-16
```




**Case 2:** X = cancellation_policy_flexible, Z = host_response_rate,

```
## Diagnostic tests:
##                   df1   df2 statistic  p-value
## Weak instruments    1 85540    167.08  < 2e-16 ***
## Wu-Hausman          1 85539     17.16 3.44e-05 ***
## Sargan              0    NA        NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.69 on 85540 degrees of freedom
## Multiple R-Squared: 0.4761,  Adjusted R-squared: 0.4755
## Wald test: 850.8 on 99 and 85540 DF,  p-value: < 2.2e-16
```




** we also tried Causality for weather and macro-economic attribute

NUS National University of Singapore

# Conclusion & Recommendation

# Conclusion & Recommendations

- Apart from the expected attributes like number of bedrooms, bathrooms etc. features like **amenities provided by the host** such as Wifi, AirCon, Washer, Essentials and Kitchen and the **connectivity of the place** are definitely the **key factors in determining the pricing**. Number of reviews as well as sentiment scores are important factors as well.

- There is **not much fluctuation in listing prices across the 12 months of analysis**. However, fluctuation in pricing can be seen in hotel room prices based on demand. The Airbnb hosts can also adjust prices i.e., increase during the peak periods like December – January and July- August and decrease prices during non-peak times like May-June and October-November.

- We performed multiple causal inference experiments on the entire data and found that **number of reviews**, **cancellation policy flexible had causal relationship with price**. Similarly, we tried causality for weather and macro economic variables against price but we didn't get conclusive evidence.

NUS National University of Singapore

# Thank You!

# References


- Inside Airbnb: http://insideairbnb.com/get-the-data.html
- Data.gov.sg: https://data.gov.sg/dataset?q=weather
- The Economist Intelligence Unit: http://country.eiu.com/singapore
- International Monetary Fund: https://data.imf.org/?sk=388dfa60-1d26-4ade-b505-a05a558d9a42
- Singapore Tourism Analytics Network: https://stan.stb.gov.sg/content/stan/en/tourism-statistics.html
- Holiday Lets, Homes, Experiences & Places. (n.d.). Airbnb. https://www.airbnb.com.sg/
- The best amenities to offer right now – Resource Centre. (n.d.). Airbnb. https://www.airbnb.com.sg/resources/hosting-homes/a/the-best-amenities-to-offer-right-now-203
- Y. Li, Q. Pan, T. Yang, and L. Guo, “Reasonable price recommendation on Airbnb using Multi Scale clustering,” 2016. 35th Chinese Control Conference (CCC) 2016.
- Perez Sanchez, V. Raul, et al. "The what, where, and why of Airbnb price determinants." Sustainability 10.12 (2018) 4596