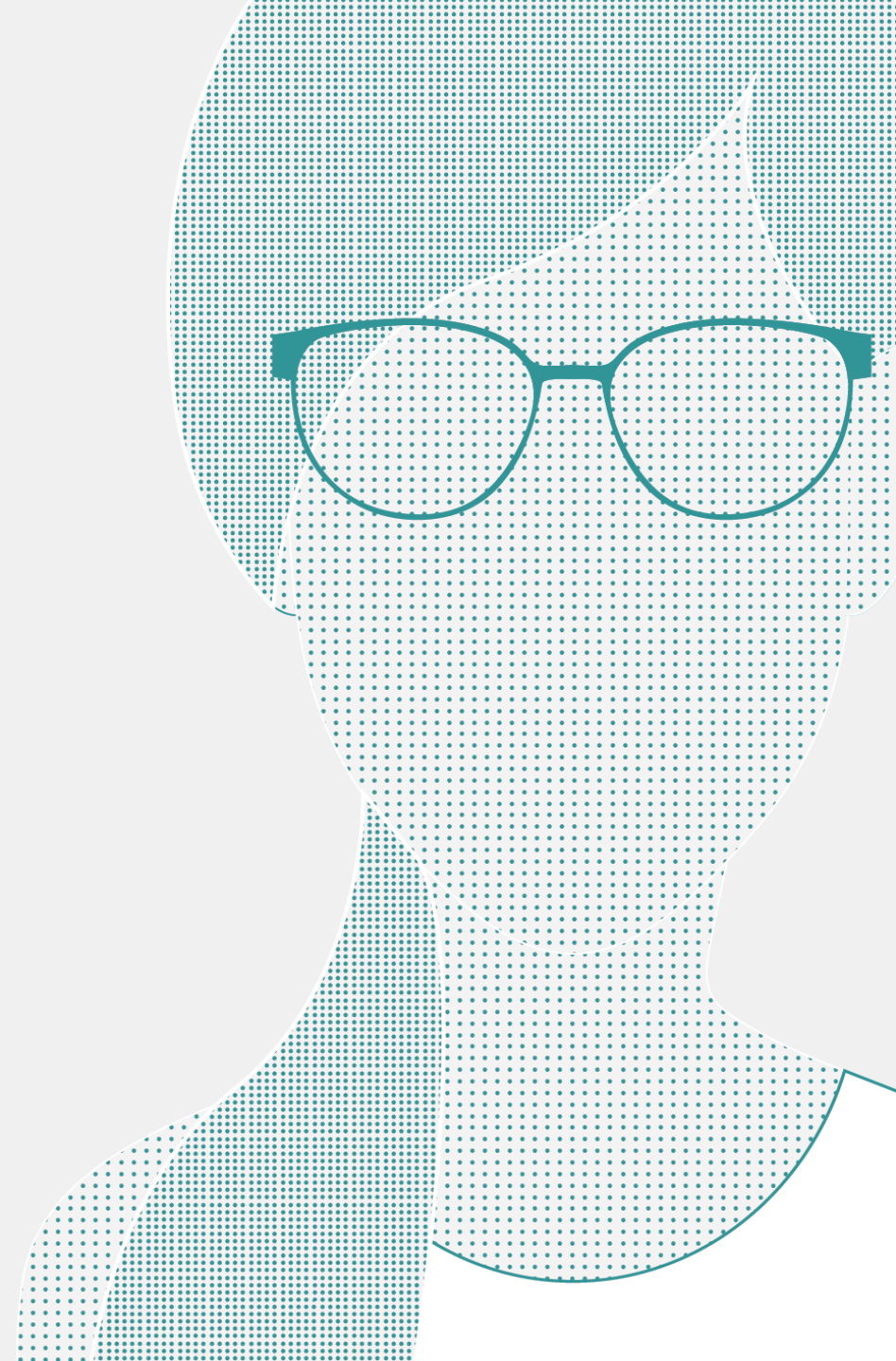# ESTIMATING ANNUAL EARNINGS IN SYDNEYS' HOUSING RENTAL MARKET

WENDY MARIA D'SA
DATA SCIENTIST

BIO

- Data Scientist from IoD

- Expertise in Data Visualisation & Communication

- Master of Business Management

- Master of Architecture & Urban Development

In 2021 residential real estate market value surged to a record

$9 trillion plus

2021 recorded the largest rental increase in a decade of

8.9%

Regional NSW weekly rents grew by

> 10%

Rental rates in Sydney grew by

7.2%

SHORT TERM RENTALS – STR

LONG TERM RENTALS – LTR

$1 billion
to the economy in 2016

18-19%
of GDP

2.7%
of dwellings

32%
of dwellings

30%-300%
higher earnings

consistent
monthly income

short
relationship with guests

contract
with tenant (good or bad)

high
investment (high maintenance, marketing, and operating expenses)

low
investment (low maintenance, marketing, and operating expenses)

up-front cost
to fully furnish the property
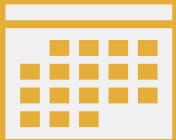
furnishings
not provided

from Nov 2021

New laws for STR in NSW

REGISTRATION
- both Hosted & Non-hosted STR

180 DAY CAP
- Non-hosted STR

FIRE & SAFETY STANDARDS
- Non-hosted STR

LOWER EARNINGS

HIGHER COSTS

MORE LTR SUPPLY & AFFORDABILITY

LESSER TAX EVATIONS

6

POTENTIAL STAKEHOLDERS

JOHN

- Multiple investment property owner.
- Ex non-hosted STR host
- Should he go towards LTR or STR?

MAYA

- Potential STR investment property owner
- Where should she buy to get maximum ROI?

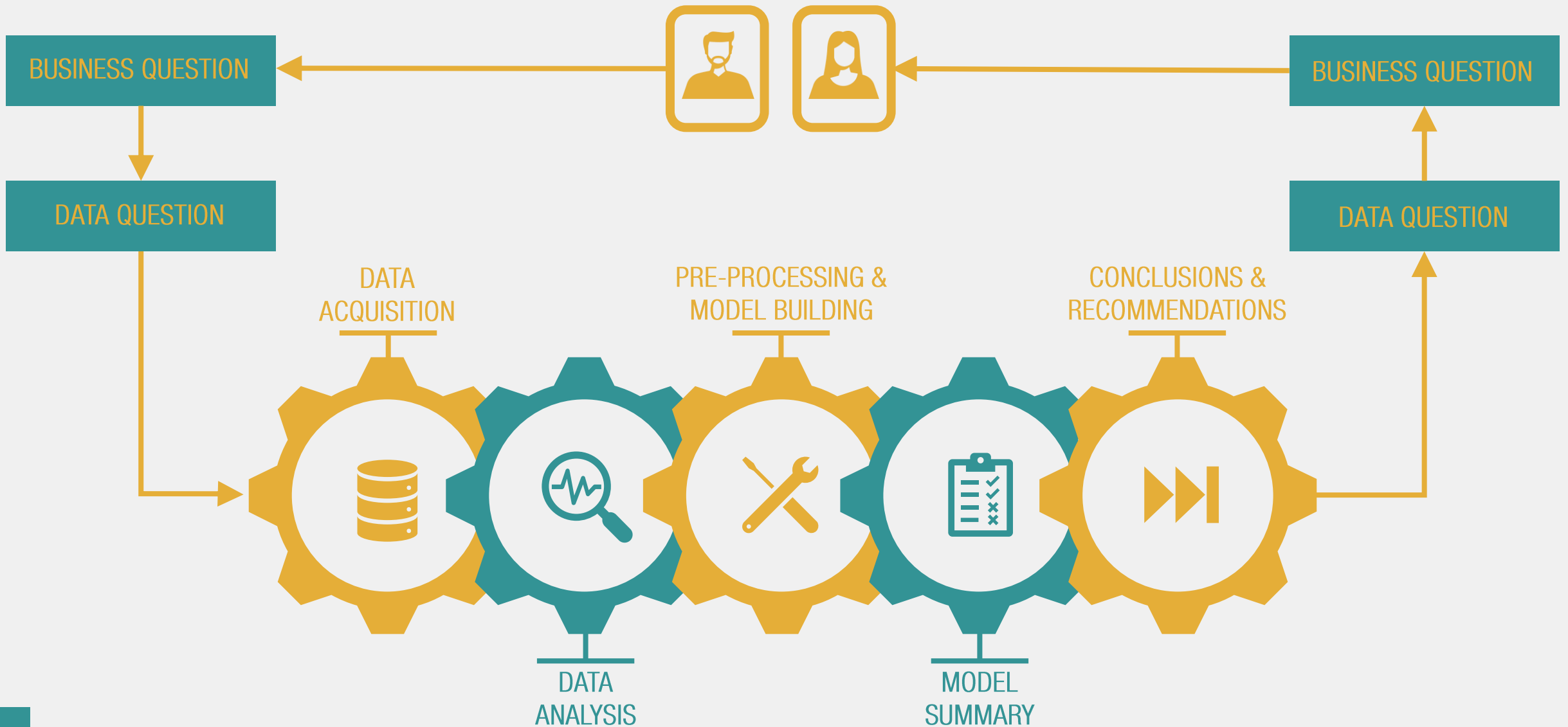# BUSINESS QUESTION

Short Term vs Long Term

WHICH REAL ESTATE INVESTMENT IS RIGHT FOR YOU ?

DATA QUESTION

? Based on AirBnB data and neighbourhood data how accurately can we predict the price of a short-term rental?

? What features can help predict the price?

? What can we do with the predictions?

PROJECT PIPELINE

BUSINESS QUESTION

BUSINESS QUESTION

DATA QUESTION

DATA QUESTION

DATA
ACQUISITION

PRE-PROCESSING &
MODEL BUILDING

CONCLUSIONS &
RECOMMENDATIONS

DATA
ANALYSIS

MODEL
SUMMARY

8

DATA ACQUISITION

ORIGINAL DATA SOURCE (archived): http://insideairbnb.com/get-the-data.html

WORKING DATA SOURCE: https://www.kaggle.com/tylerx/sydney-airbnb-open-data

36,662 observations with 96 features ▶ NLP | categorical | numerical | boolean

Listing & Host

- Id
- host_location,
- response_time
- response_rate
- superhost
- host location

Location

- neighbourhood
- city
- zipcode
- market
- smart_location
- country_code
- latitude
- longitude

Features engineered

- lga
- distance to cbd
- café_density
- transport rating
- beach access
- cultural access
- national park access

Property

- property_type, room_type
- accommodates, beds
- bathrooms, bedrooms
- bed_type
- amenities
- price, security_deposit
- cleaning_fee
- guests_included
- extra_people
- availability_90

Reviews

- number of reviews
- review scores rating
- review_accuracy
- review_cleanliness
- review_checkin
- review_communication
- review_location
- review_value
- instant_bookable
- cancellation_policy

CORRELATION OVERVIEW

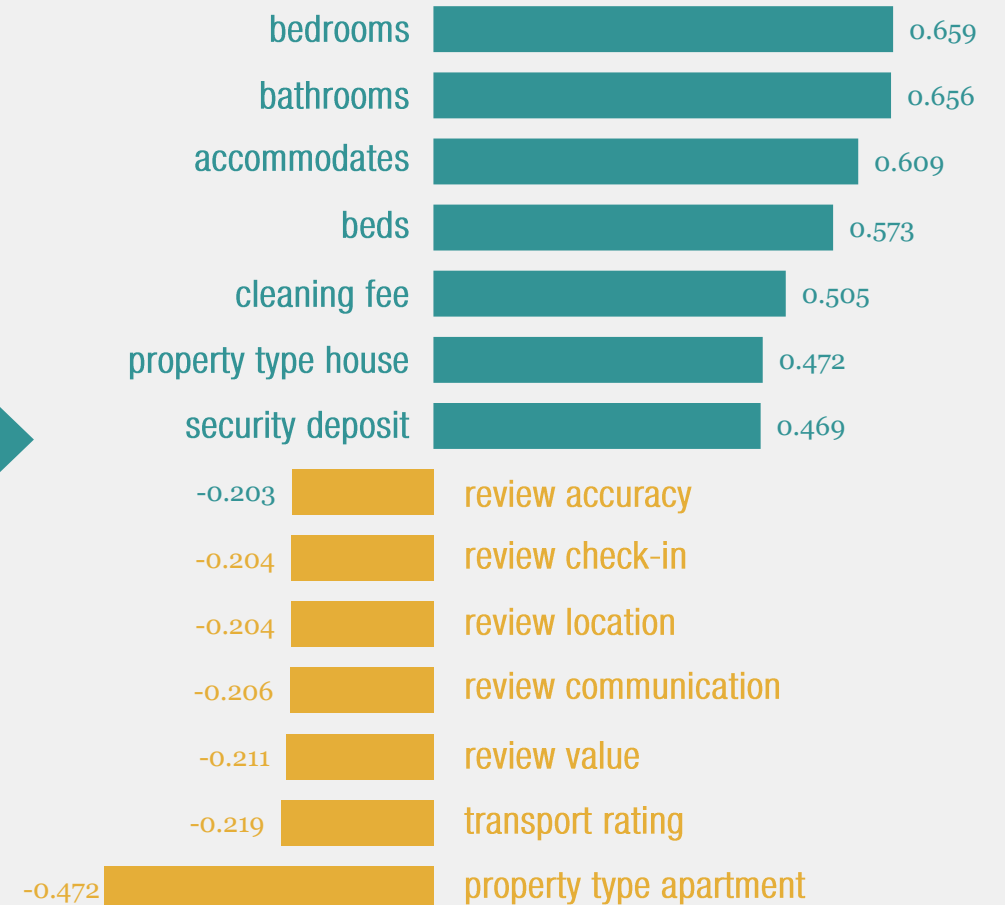## HIGH CORRELATION WITH PRICE

- Bedrooms, Bathrooms, Accommodates, and Beds
- Property type House
- Cleaning fee & Security deposit
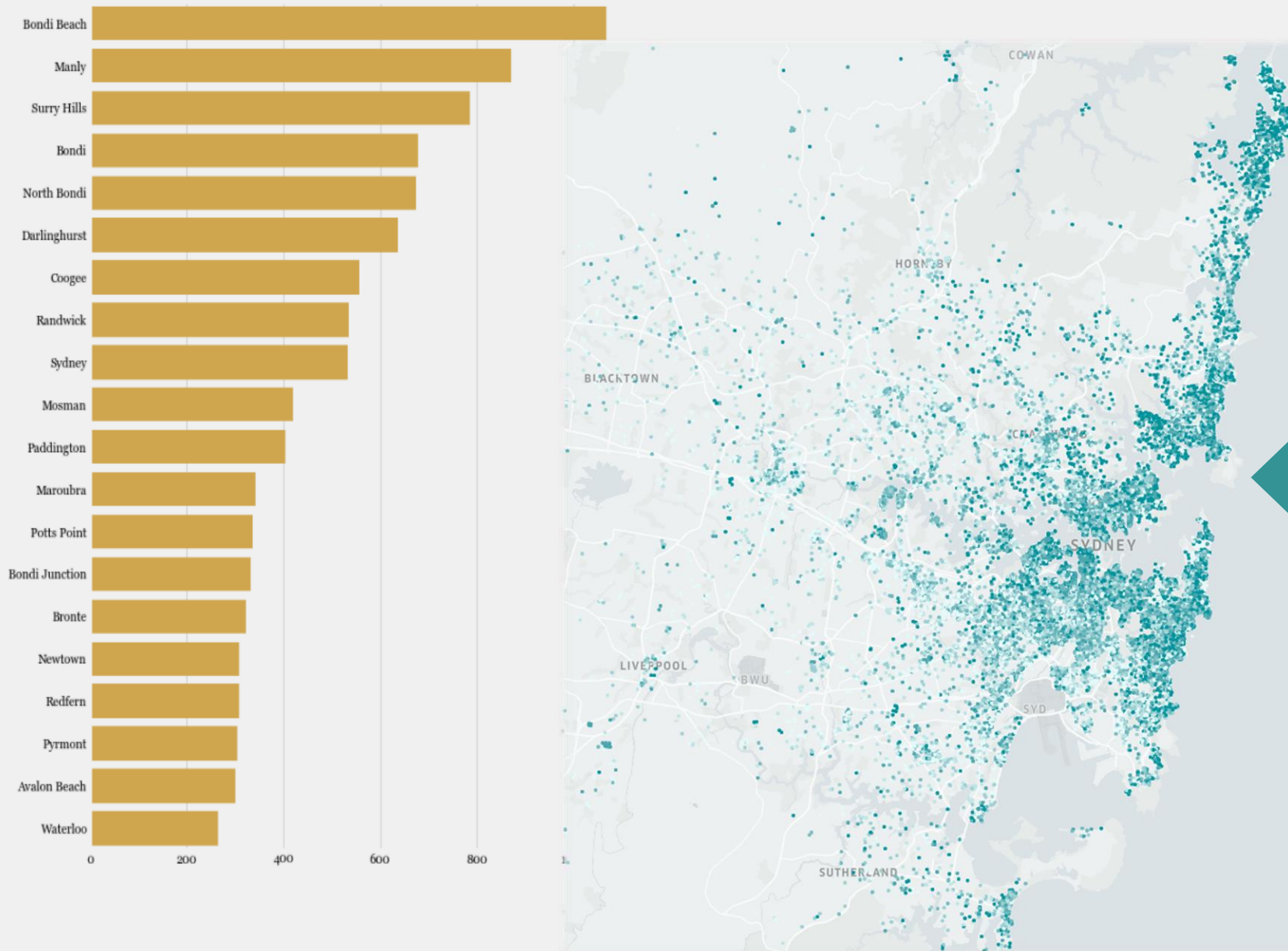
LOW CORRELATION WITH PRICE

- Property type Apartment
- Transport rating
- Review scores – accuracy, location, check-in, communications, and value

HIGH CORRELATION WITH REVIEW SCORES RATING

- Review scores – accuracy, location, check-in, cleanliness, communications, and value

| Feature	Value
bedrooms	0.659
bathrooms	0.656
accommodates	0.609
beds	0.573
cleaning fee	0.505
property type house	0.472
security deposit	0.469
review accuracy	-0.203
review check-in	-0.204
review location	-0.204
review communication	-0.206
review value	-0.211
transport rating	-0.219
property type apartment	-0.472

Top 20 Suburbs

LISTINGS OVERVIEW

- Top 20 Suburbs topped by the beach suburbs of Bondi and Manly and the centrally located suburb of Surry Hills

- Most suburbs are located along the eastern and northern coasts of Sydney

PRICING OVERVIEW

Price ($)

House vs Apartment Price ($)

- Notable drop in pricing after the $200 mark.

- Only 0.5% observations above the $1000 mark with an extreme of $14,999.
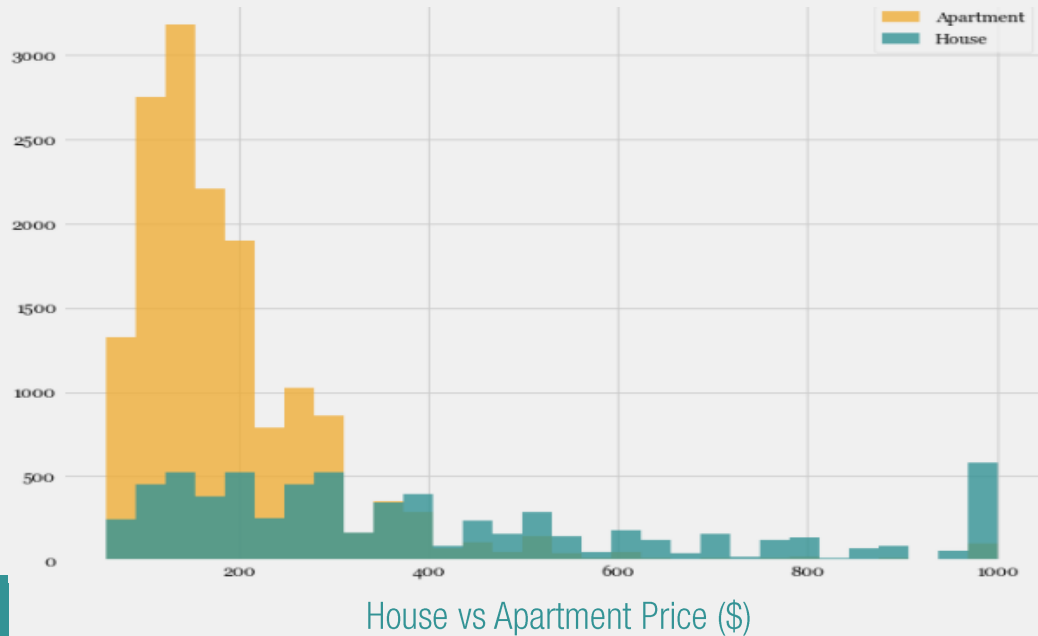
- Apartments dominate the listings and mostly range from $60 - $300.

- Houses cover the entire price range but also dominate the higher price range.

12

LGA OVERVIEW

TOP 5 PRICIEST LGA'S

1. Mosman
2. Northern Beaches
3. Hunters Hill
4. Lane Cove
5. Woollahra

- Mosman has much lower listings than the eastern suburbs but commands higher price - usually rent out houses at the top end of the price range.

- Suburbs in the Northern Beaches barely touch the Top 20 listings numbers – Beach suburbs of Northern Beaches are much pricier than the eastern beach suburbs.

- Suburbs in western and south western Sydney command some of the cheapest prices.
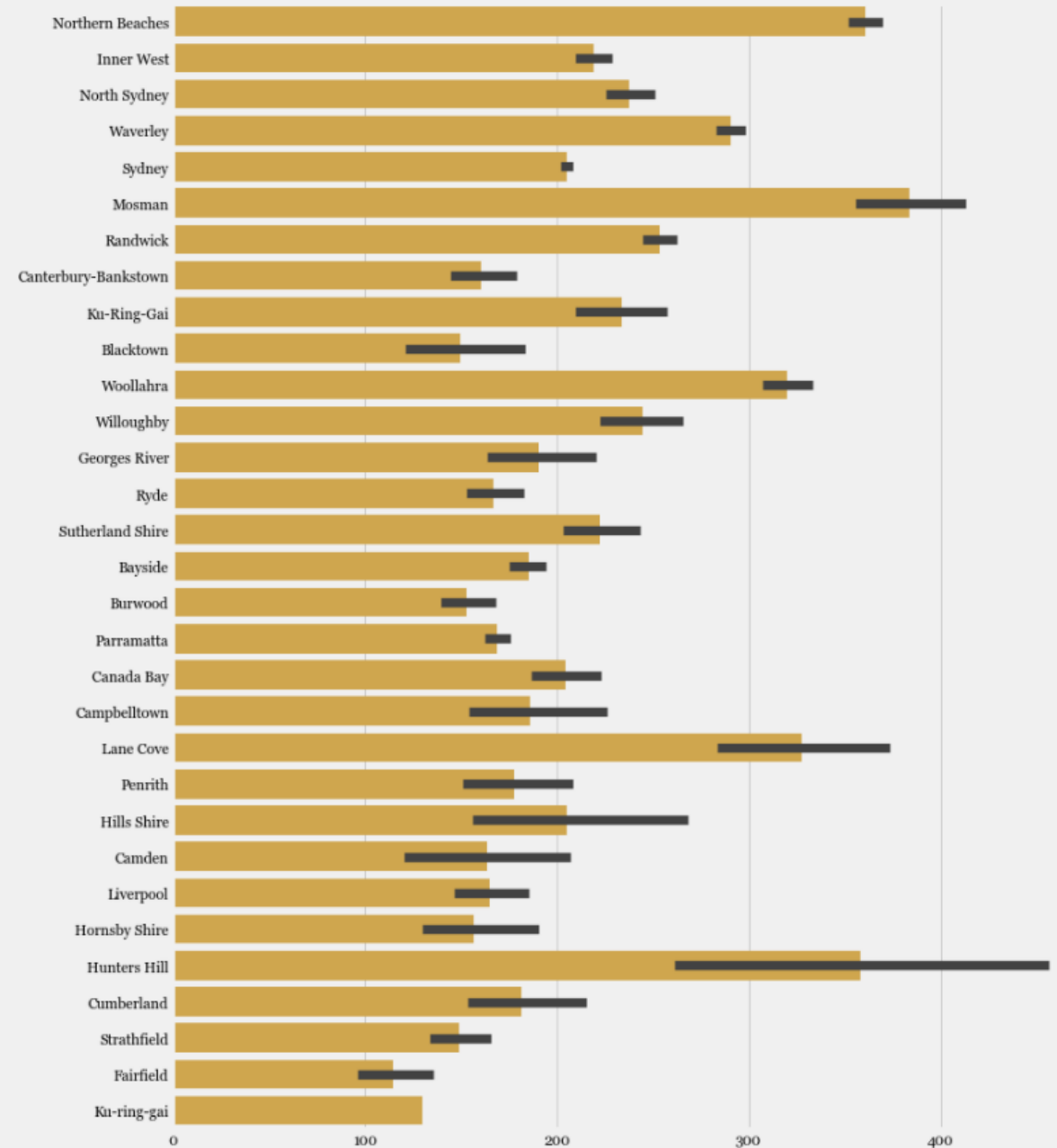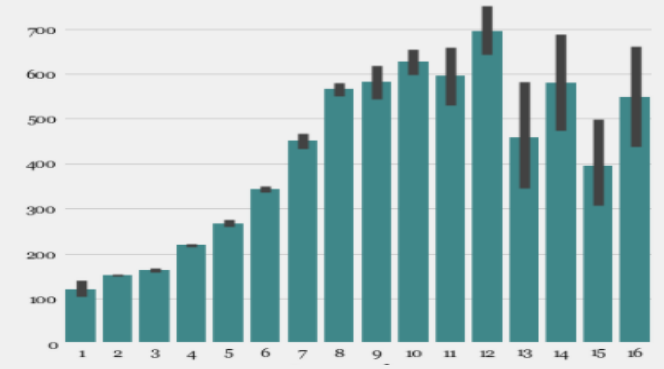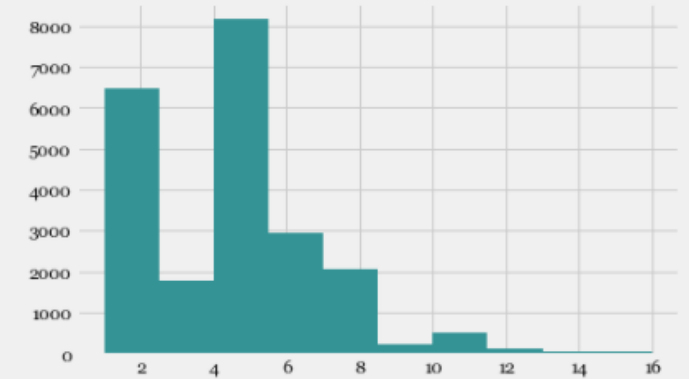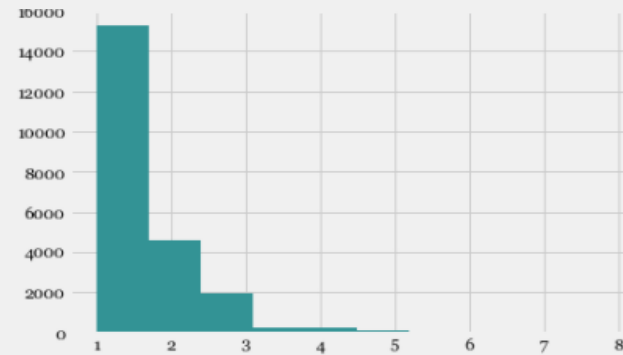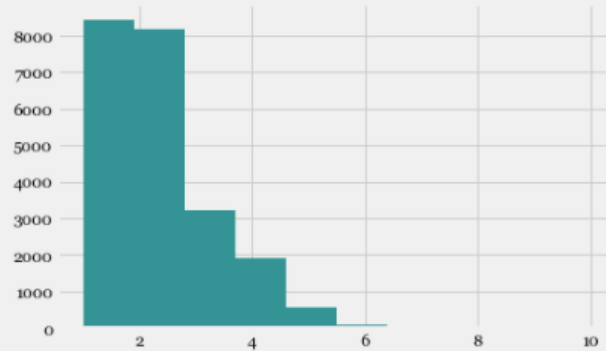
ACCOMMODATION OVERVIEW
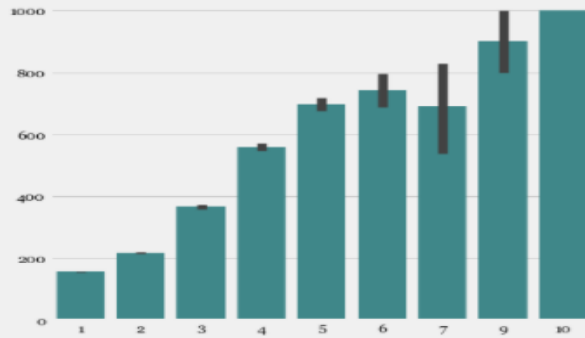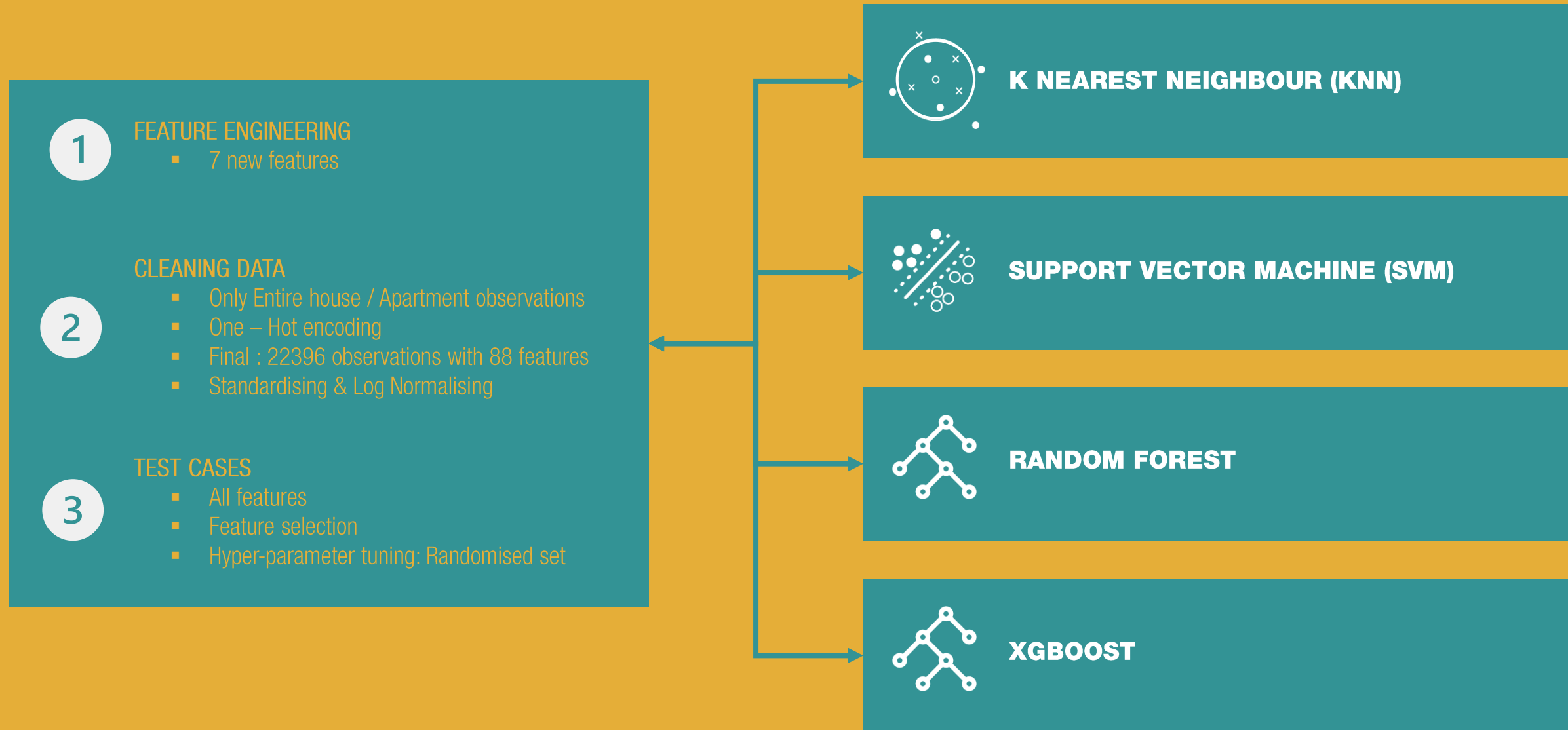

Bedrooms vs Price ($)
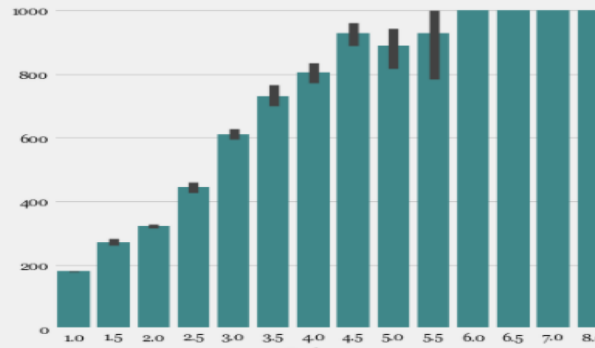
Bathrooms vs Price ($)

Accommodates vs Price ($)

- As beds and bathrooms increase so does the price

- As accommodates increase so does the price, but has diminishing returns after 12 guests

- The median common setup is 2 beds in 2 bedroom, with 1 bathroom that accommodates 4 guests.

PRE-PROCESSING & MODEL SETUP

1 FEATURE ENGINEERING
- ▪ 7 new features

2 CLEANING DATA
- ▪ Only Entire house / Apartment observations
- ▪ One – Hot encoding
- ▪ Final : 22396 observations with 88 features
- ▪ Standardising & Log Normalising

3 TEST CASES
- ▪ All features
- ▪ Feature selection
- ▪ Hyper-parameter tuning: Randomised set

K NEAREST NEIGHBOUR (KNN)

SUPPORT VECTOR MACHINE (SVM)

RANDOM FOREST

XGBOOST

|	size	train rmse	test rmse	train r2	test r2
RANDOM FOREST	All features	0.38	0.39	0.65	0.63
	Imp features	0.38	0.39	0.64	0.63
	Hyper-param	0.34	0.38	0.71	0.66
XGBOOST	All features	0.34	0.36	0.71	0.68
	Imp features	0.37	0.38	0.67	0.65
	Hyper-param	0.34	0.37	0.71	0.66
K NEAREST NEIGHBOUR (KNN)	All features	0.39	0.42	0.63	0.58
	Imp features	0.37	0.39	0.66	0.62
	Hyper-param	0.37	0.40	0.65	0.61
SUPPORT VECTOR MACHINE (SVM)	All features	0.28	0.37	0.8	0.66
	Imp features	0.37	0.38	0.67	0.66
	Hyper-param	0.37	0.38	0.66	0.64

2 → Performs the best with least features

1 → Performs the best overall

- Models did not improve with further features
- Can be improved with K-means clustering of LGA's and suburbs.
- Test with Transformed target regressor, and Grid Search CV
- Use ensemble LGBM & Gradient Boosting

16
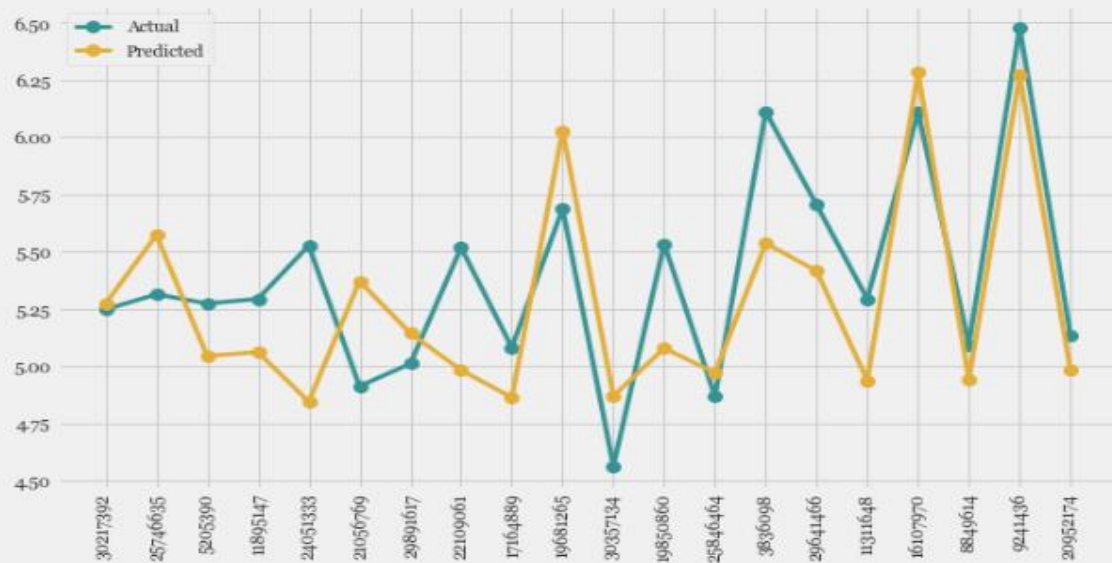
Actual vs. Predicted (log price)
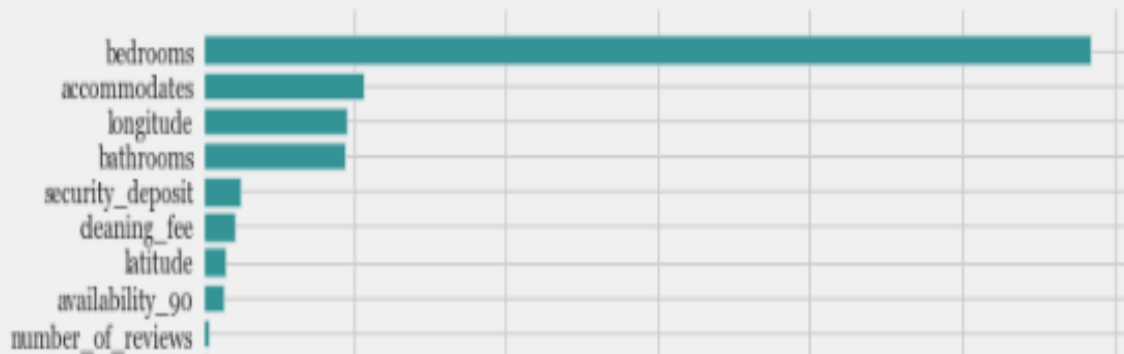
Random Forest with top 9 features

Overall Best
RANDOM FOREST WITH 9 FEATURES
explains 66% variance in price

TOP 9 FEATURES FOR DEPLOYMENT

Bedrooms

Accommodates

Longitude

Bathrooms

Security deposit

Cleaning fee

Latitude

Availability in 90 days

Number of reviews

17

RECOMMENDATIONS

WHAT PEOPLE LOOK FOR

- Bedrooms, bathrooms and number accommodated
- Cleaning fees & security deposit also has an effect on bookings.

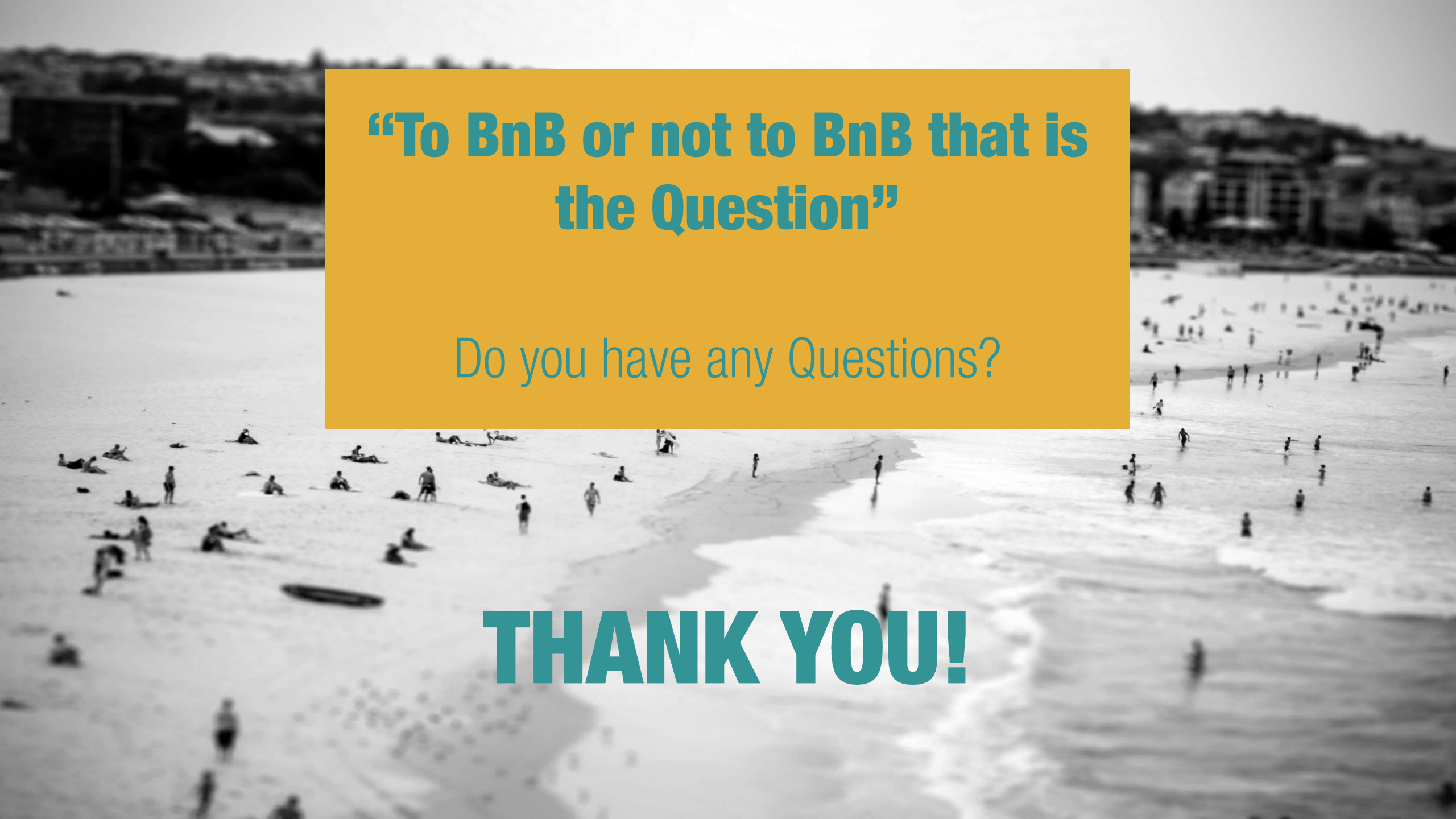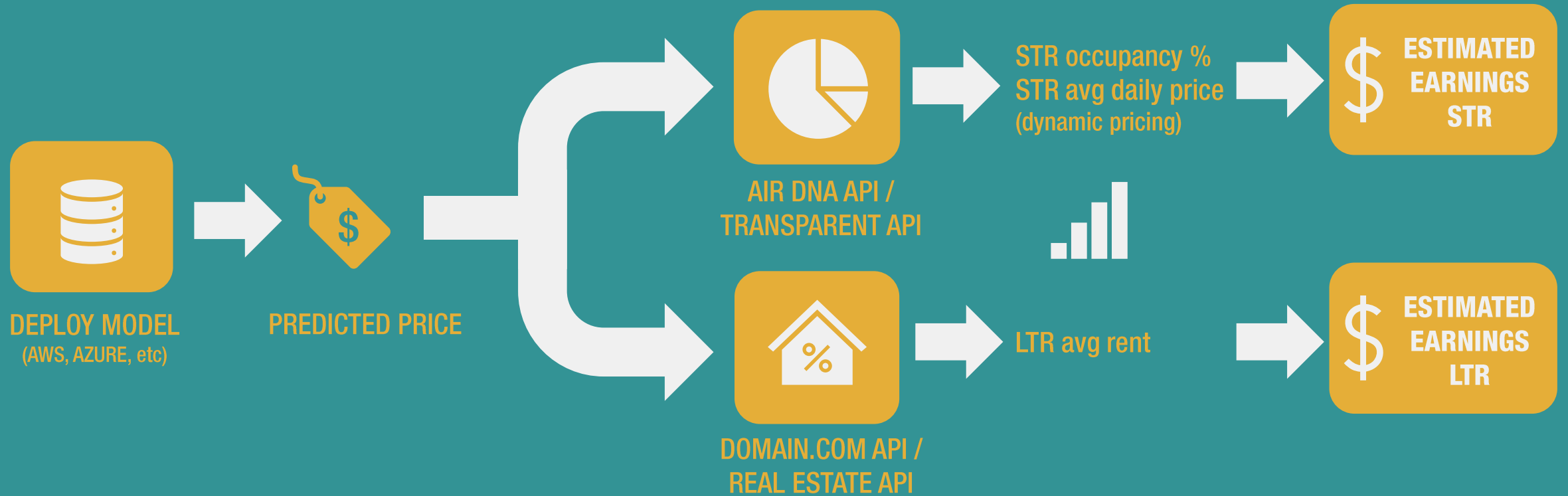## WHERE TO INVEST FOR STRs

- Beach suburbs
- City centre suburbs

NEXT STEPS

- Image analysis using CNN to assess image quality as an input into the pricing model
- Sentiment analysis on the reviews
- Use better quality/more accurate data which includes the actual average prices paid per night

THE APPLICATION PIPELINE

DEPLOY MODEL
(AWS, AZURE, etc)

→

PREDICTED PRICE

**AIR DNA API /
TRANSPARENT API**

→ STR occupancy %
STR avg daily price
(dynamic pricing)

→ **ESTIMATED EARNINGS STR**

**DOMAIN.COM API /
REAL ESTATE API**

→ LTR avg rent

→ **ESTIMATED EARNINGS LTR**

"To BnB or not to BnB that is the Question"

Do you have any Questions?

THANK YOU!