BUILDING AN INTERPRETABLE PERSONALIZED RENTAL SCORING MODEL BASED ON PRICE, SIZE, COMMUTE EFFICIENCY, SAFETY, AND CONVENIENCE

Zhu Suying (Sonia)

BEng (CompSc)

3035845710

Today's Agenda

01 Problem Definition

02 Methodologies

03 Key Findings

04 Conclusion

1. PROBLEM DEFINITION

- Background
- Data Science Questions
- Problem Importance
- Difficulties
- Project Goals
- Related Works
- Data Available

1. PROBLEM DEFINITION

From Personal Struggle to Data Solution

- Kicked off by 4 different halls
- Lived in 2 private rental housing
- Rented in Shen Zhen for 2 times
- At least 3 times of movement each year
- Eager for stable living experience

Data Science Questions to Answer

- **How to build an interpretable model to quantify the weights of various factors such as commute efficiency, safety, and convenience for personalized rental recommendations based on my own preference?**

- **How can fragmented data be integrated into a dynamic scoring system to find the optimal rental properties for me?**

1. PROBLEM DEFINITION

Importance: Why Hong Kong's Rental Market Needs a Data-Driven Solution

- Market Fragmentation & Information Overload
- Cost of Inefficient Decisions

Difficulties: Challenges in Solving a Fragmented, Biased Market

- Challenge 1 : Subjectivity in Traditional Advice

  Agents' Advice—Biases Due to vested interests and limited coverage

  Local People – Subjective and restricted to their own neighborhood
- Challenge 2: Limitations of Online Platforms

  Lack personalized recommendations

  Lack detailed information about commute and lifestyle factors

1. PROBLEM DEFINITION

Project Goals

- Short-term: Build a personalized scoring model to find my ideal home (rent, commute, safety, amenities).
- Long-term: Empower renters to quantify trade-offs and make data-driven decisions.

Related Works

- Prior studies: Multi-criteria models eg. MCDM to evaluate rentals based on price, location, and size.
- Gap: Focus on generic criteria, lack individual preferences

Data Availability

- Rental listings: Scraped 5k listings from 28hse.com (price, size, location).
- Commute data: Google Maps API .
- Lifestyle metrics: Government open data (crime stats, amenities per district).

2. METHODOLOGIES

- Data Collection
- Data Cleaning & Preprocessing
- Model Development
- Scoring System Integration

ENVIRONMENT

Platforms:

– R 4.4.3

– RStudio

– Python 3.9

– Visual Studio Code

Libraries:

– rvest 1.0.4

– dplyr 1.1.4

– ggplot 3.5.1

– MASS 7.3–64

– ……

Data Storage:

–Comma–separated values (CSV)

DATA COLLECTION–PROPERTY DATA

Key Codes Snippets

```r
# Location components
district <- node %>%
  html_node("div.district_area a:nth-of-type(1)") %>%
  html_text(trim = TRUE) %||% NA_character_

estate <- node %>%
  html_node("div.district_area a:nth-of-type(2)") %>%  # Capture second link
  html_text(trim = TRUE) %||% NA_character_

unit <- node %>%
  html_node("span.unit_desc") %>%
  html_text(trim = TRUE) %||% NA_character_

# Size extraction for both types
size_lines <- node %>%
  html_nodes("div.areaUnitPrice div") %>%
  html_text() %||% NA_character_

# Initialize both size types as NA
size_gross <- NA
size_saleable <- NA

# Process each size line
for(line in size_lines) {
  if(str_detect(line, "建築面積")) {
    size_gross <- line %>%
      str_extract("\\d{1,3}(?:,\\d{3})+|\\d+") %>%  # Match both comma-separated and plain numbers
      str_remove_all(",") %>%                       # Remove commas
      as.numeric()
  }
  if(str_detect(line, "實用面積")) {
```

```r
# Scraping loop
for(page_num in 1:total_pages) {
  # Respectful crawling
  if(page_num > 1) {
    Sys.sleep(runif(1, 1.5, 3.5) + ifelse(page_num %% 50 == 0, 5, 0))
  }

  # Generate URL
  current_url <- sprintf("%s&page=%d", base_url, page_num)
  print(current_url)
  # Attempt page read
  page_content <- safely_read_html(current_url)

  # Process page
  if(!is.null(page_content)) {
    page_data <- process_page(page_content)
    all_data[[page_num]] <- page_data
    message(sprintf("Page %04d/%04d: %d properties",
                    page_num, total_pages, nrow(page_data)))
  } else {
    failed_pages <- c(failed_pages, page_num)
    message("Failed to scrape page: ", page_num)
  }

  # Intermediate saving
  if(page_num %% 20 == 0) {
    save_scraped_data(all_data, failed_pages)
  }
}
```

香港屋網 28Hse.com

Sample Property Raw Data

	Price <dbl>	Location <chr>	Size_Gross_sqft <dbl>	Size_Saleable_sqft <dbl>	Features <chr>
1	19500	旺角SKYPARK	NA	404	1 房 , 1 浴室, 向東, 私人屋苑, 望開揚景, 雅緻裝修, 包全屋家電, 冷氣機, 洗衣機, 熱水爐, 雪櫃
2	15500	紅磡黃埔新村	NA	321	2 房 , 1 浴室, 向西南, 私人屋苑, 豪華裝修
3	12500	紅磡環海東岸	NA	205	開放式間隔 , 1 浴室, 向西北, 私人屋苑, 村屋, 望園景, 望開揚景, 望樓景, 包全屋家電, 微波爐, 洗衣機
4	18500	東涌映灣園	NA	563	2 房 , 1 浴室, 私人屋苑, 望山景, 望開揚景, 包全屋傢俬, 可養貓狗
5	17300	東涌東環	NA	452	2 房 , 1 浴室, 私人屋苑
6	16500	紅磡BakerCircleGreenwich	NA	264	1 房 , 1 浴室, 私人屋苑, 雅緻裝修

DATA COLLECTION-LIFE DATA

各區域及地區
AREAS AND DISTRICTS

區域 Area	地區 District	地區內的分區名稱	Names of Sub-districts within District Boundaries	規劃統計小區 Tertiary Planning Units
九龍 KOWLOON	深水埗 Sham Shui Po	美孚、荔枝角、長沙灣、深水埗、石硤尾、又一村、大窩坪、昂船洲	Mei Foo, Lai Chi Kok, Cheung Sha Wan, Sham Shui Po, Shek Kip Mei, Yau Yat Tsuen,Tai Wo Ping, Stonecutters Island	255, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269
	九龍城 Kowloon City	紅磡、土瓜灣、馬頭角、馬頭圍、啟德、九龍城、何文田、九龍塘、筆架山	Hung Hom, To Kwa Wan, Ma Tau Kok, Ma Tau Wai, Kai Tak, Kowloon City, Ho Man Tin, Kowloon Tong, Beacon Hill	213, 231, 232, 233, 234, 235, 236, 237, 241, 242, 243, 244, 245, 246, 247, 271, 272, 285, 286
	黃大仙 Wong Tai Sin	新蒲崗、黃大仙、東頭、橫頭磡、樂富、鑽石山、慈雲山、牛池灣	San Po Kong, Wong Tai Sin, Tung Tau, Wang Tau Hom, Lok Fu, Diamond Hill, Tsz Wan Shan, Ngau Chi Wan	281, 282, 283, 284, 287, 288, 289

District	Reported Crimes	Bars & Restaurants	Malls & Supermarkets	Parks
Yau Tsim Mong	8883	6243	324	95
Kwun Tong	7608	3160	199	128
Sha Tin	6519	2520	193	161
Yuen Long	6327	3256	193	188
Tai Po	5887	1403	78	93
Sham Shui Po	5357	2991	174	86
Tuen Mun	4923	1699	135	74
Kwai Tsing	4832	1235	89	54
Central and Western	4609	4462	154	99
Eastern	4347	4179	226	115
Sai Kung	3904	1204	88	71
Wong Tai Sin	3737	1170	101	62
Kowloon City	3531	2667	165	108
Southern	3040	777	70	60
Tsuen Wan	2711	3401	197	87
Wan Chai	2523	2322	79	59
Islands	2208	1232	46	61
North	586	1755	104	131

DATA CLEANING & PREPROCESSING

district	Unnamed..0	Price	Location	Size_Gross_sqft	Size_Saleable_sqft	
Central and Western	1735	20000	上環嘉寶大廈	NA	343	2 房 , 1 浴室, 洋樓, 連天台, 望開揚景
Central and Western	1530	16800	上環世瑛大廈	720	NA	開放式間隔, 寫字樓, 望開揚景
Central and Western	2824	16000	上環家樂花園	512	298	2 房 , 1 浴室, 私人屋苑
Central and Western	3013	13500	上環美輪樓	NA	288	開放式間隔 , 1 浴室, 唐樓, 單幢式大廈, 雅緻裝修,
Central and Western	641	18800	堅尼地城采逸軒	NA	327	2 房 , 1 浴室, 單幢式大廈, 包全屋家電, 有露台, 獨

Total_Time	Transit_Details	Crime_Per_Person	Bars_Per_SqKm
14	3 min WALKING, 8 min TRAM (石塘咀 - 銅鑼灣), 3 min WALKING	19.8492678725237	355.537848605578
11	2 min WALKING, 5 min TRAM (上環(西港城) - 筲箕灣), 3 min WALKING	19.8492678725237	355.537848605578
22	5 min WALKING, 13 min TRAM (堅尼地城 - 筲箕灣), 3 min WALKING	19.8492678725237	355.537848605578
15	3 min WALKING, 8 min TRAM (堅尼地城 - 跑馬地), 3 min WALKING	19.8492678725237	355.537848605578
19	4 min WALKING, 7 min SUBWAY (港島綫), 8 min WALKING	19.8492678725237	355.537848605578

Malls_Per_SqKm	Parks_Per_SqKm	Population_Density	Walking_Time	Bus_Time	Subway_Time	Tram_Time	Total_Transit_Time	Walking_Percent
12.2709163346614	7.88844621513944	18501.9920318725	6	0	0	8	14	42.8571428571429
12.2709163346614	7.88844621513944	18501.9920318725	5	0	0	5	10	50
12.2709163346614	7.88844621513944	18501.9920318725	8	0	0	13	21	38.0952380952381
12.2709163346614	7.88844621513944	18501.9920318725	6	0	0	8	14	42.8571428571429
12.2709163346614	7.88844621513944	18501.9920318725	12	0	7	0	19	63.1578947368421

MODEL DEVELOPMENT

- Manually Rated 100 properties
- Linear Model vs. Random Forest

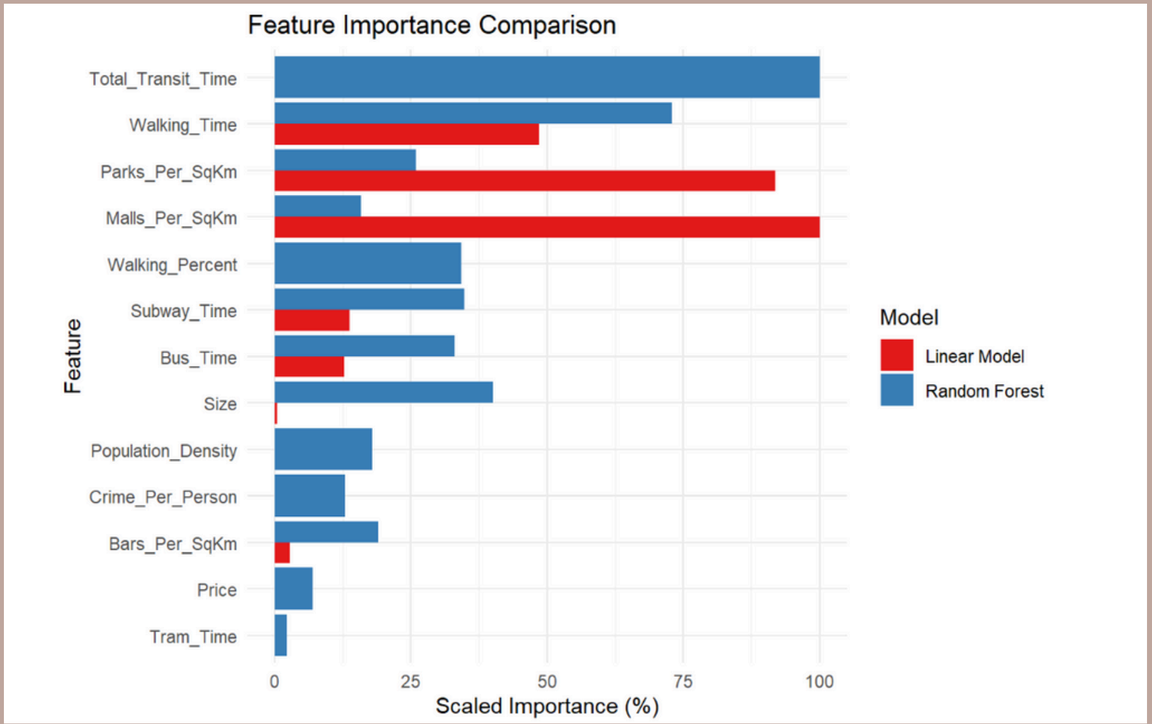
SCORING SYSTEM INTEGRATION

- Rated the Rested Properties
- Rank by scores

3. KEY FINDINGS

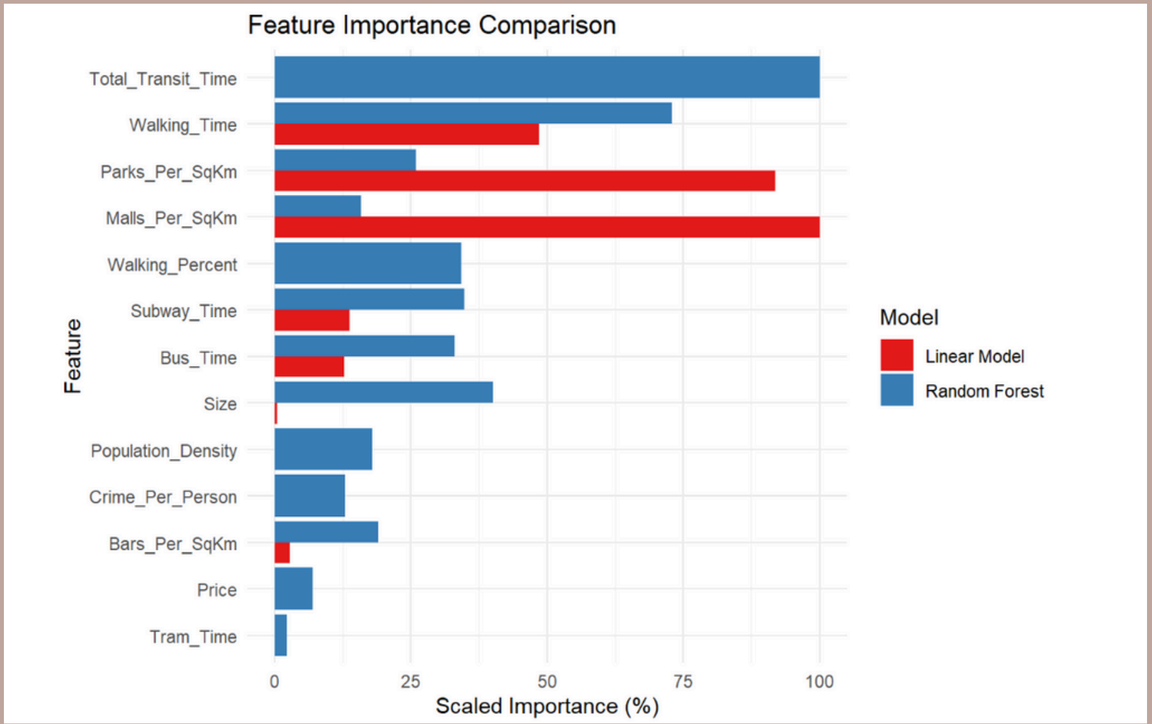
Case Study based on
my personal preferences

KEY FINDINGS 1
TRANSPORTATION DOMINATES



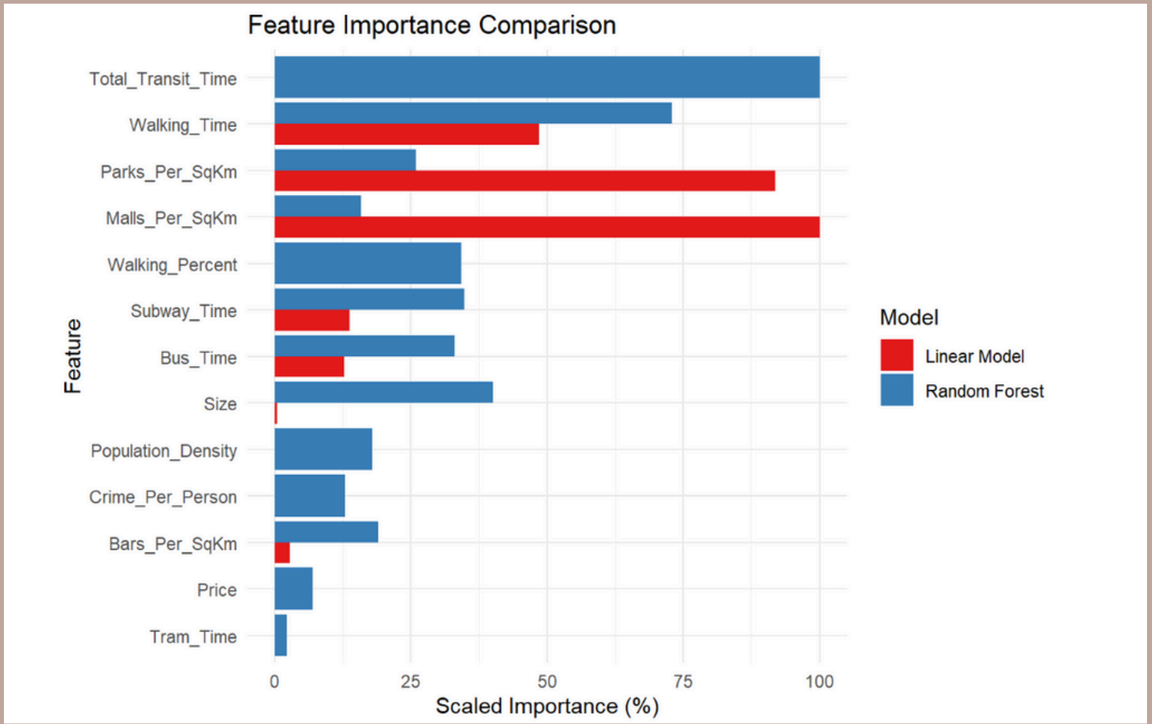
- Total transit time is the top factor in Random Forest model
- Walking time ranks high in both models
- Transportation factors account for 4 of the top 6 features

KEY FINDINGS 2
AMENITIES MATTER



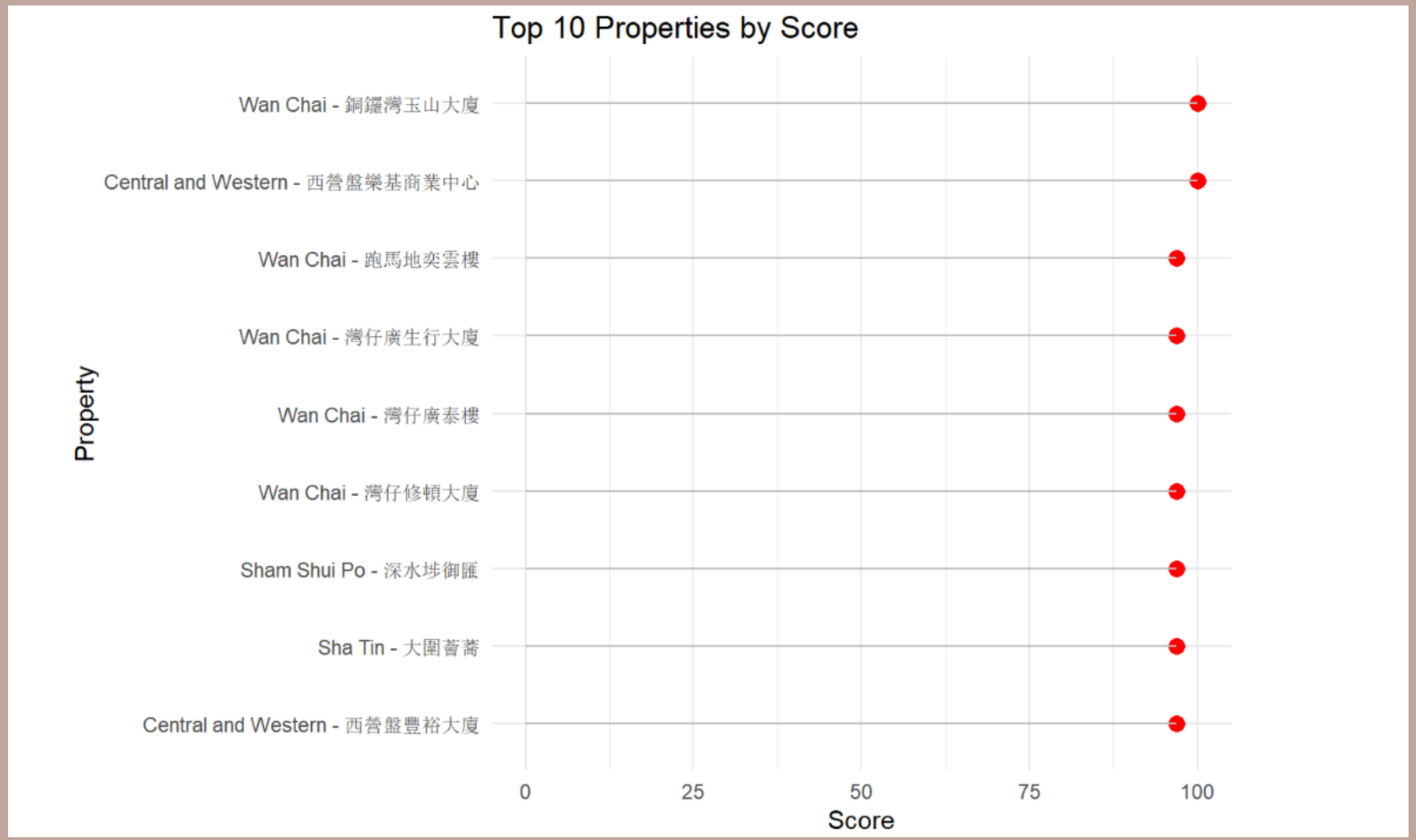
- Malls_Per_SqKm is the top factor in the linear model
- Parks_Per_SqKm is the second most important
- Both are statistically significant predictors

KEY FINDINGS 3
SIZE VS. PRICE



- Size ranks #3 in Random Forest model
- Price ranks surprisingly low (#12)
- Property characteristics less important than location factors

KEY FINDINGS 4
Distribution of Properties by My Preference



4. CONCLUSION

DATA-DRIVEN PROPERTY DECISIONS

- Model successfully quantifies my preference weights by scoring system
- Transportation accessibility emerged as dominant factor
- Amenities proximity and size outweigh price
- Dynamic scoring system identifies optimal properties matching my personal preferences
- New Rental Recommendations:
  Sha Tin & Happy Valley

FUTURE WORKS
EMPOWERING DATA-DRIVEN PROPERTY DECISIONS

Model Enhancement

- Real-time transit data integration
- Preference drift detection
- Multi-user preference balancing

Long-Term Vision

- Democratize data-driven property decisions
- Quantifiable trade-off analysis for all renters
- Transparent rental market with informed consumers

THANK YOU

References

Cost of living in Hong Kong – single, family & student. (2025, March 10). Instarem | International Money Transfers. https://www.instarem.com/blog/cost-of-living-in-hong-kong/

Hong Kong fun in 18 districts - home. (n.d.). https://www.gohk.gov.hk/en/

Hse. (2025, May 7). 28HsE 租屋網站, 租屋,最新租盤資料. 28Hse 香港屋網. https://www.28hse.com/rent

Miguelez, C. (2025, April 1). The most liveable districts in Hong Kong - Preply. Language learning with Preply Blog. https://preply.com/en/blog/most-liveable-districts-in-hong-kong/

Mulliner, E., Malys, N., & Maliene, V. (2015). Comparative analysis of MCDM methods for the assessment of sustainable housing affordability. Omega, 59, 146–156. https://doi.org/10.1016/j.omega.2015.05.013