Paper title

Author

Affiliation

Course

Instructor:

Due date

Table of Contents

Table of Contents.....2
Identification, Justification, and Gathering of Data.....2
Data Cleaning, Pre-Processing, and Data Model.....7
Exploratory Data Analysis (EDA).....9
Application and Interpretation: Stochastic Models and Methods.....13
Design and Effectiveness of Recommendation System.....24
Code.....25
    Code Overview.....25
Legal and Ethical Considerations.....29
Conclusions, Analysis, and Reflection.....31
    Conclusions.....31
    Analysis.....31
References.....32

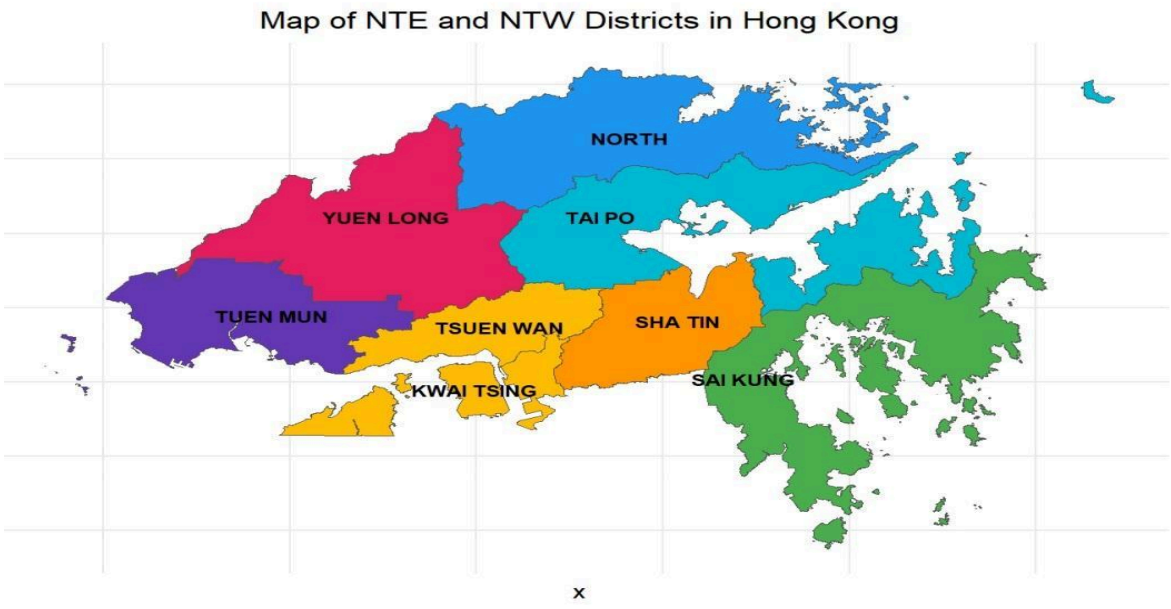
Identification, Justification, and Gathering of Data

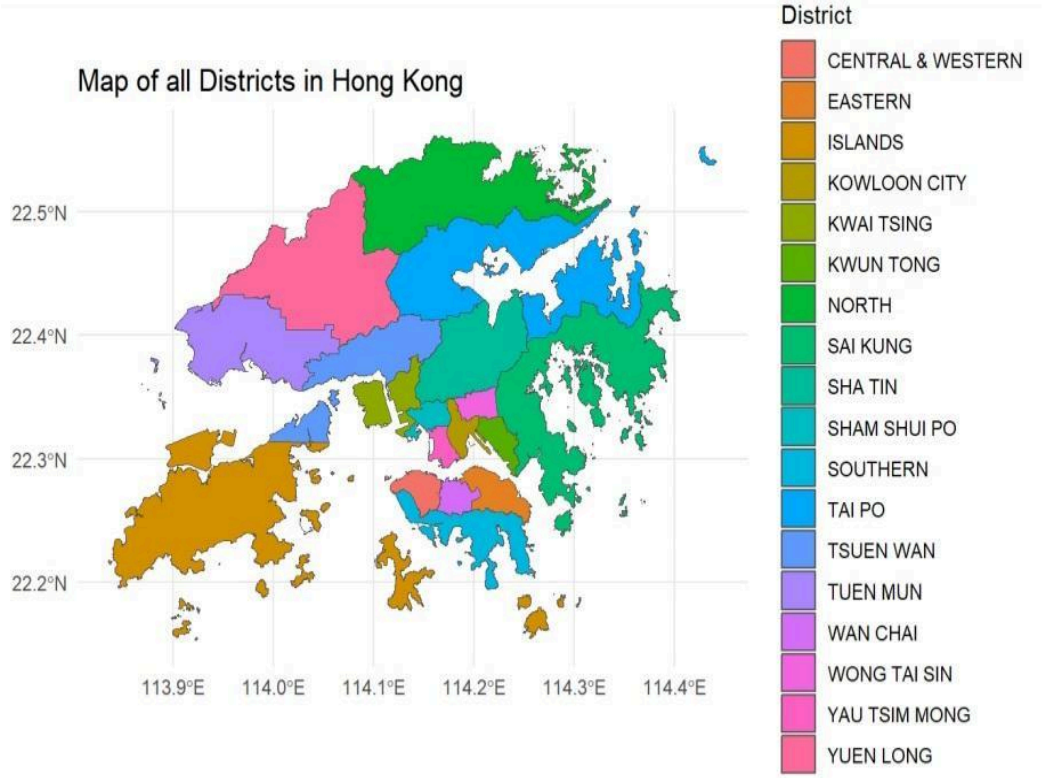
Data Identification and Sources

This project aims to recommend suitable districts for housing in New Territories East (NTE) and New Territories West (NTW) by considering three primary factors: house prices, crime rates, and school quality. Furthermore, traffic statistics, population features, and parking spaces were considered to add more reliability to the proposals. For comprehensive analysis, datasets were identified and gathered from multiple reliable sources:

- Housing Data: From Property.hk and Hong Kong Government Data, we have collected the district-wise pricing of residential property starting from 1999 and till 2023.
- Crime Data: From the copy of the Hong Kong Police annual report focusing on crime statistics by districts.
- School Data: Including the categories of school type, finance type, and session rankings which were downloaded from the Hong Kong Education Bureau.
- Traffic Data: From Hong Kong Transport Department, consisting of visitors' movement data grouped by district.
- Census Data: Obtained from the Census and Statistics Department and containing demographic and socioeconomic characteristics.

Completeness, relevance, and credibility of each dataset for this study were evaluated to confirm suitability for inclusion in the analysis. The common linkage that was selected across the datasets is geographic alignment (district names).


Map of NTE and NTW Districts in Hong Kong
NORTH
YUEN LONG
TAI PO
TUEN MUN
TSUEN WAN
SHA TIN
KWAI TSING
SAI KUNG
x




```
#NEW TERRITORIES MAP
```{r}
#hongkong shapefile
hk_districts <- st_read("/Users/apple/Downloads/Hong_Kong_18_Districts_3574690825663895789")
#map of all hongkong districts
ggplot(data = hk_districts) +
  geom_sf(aes(fill =ENAME)) +
  labs(title = "Map of all Districts in Hong Kong",
       fill = "District") +
  theme_minimal()
```
```

```r
# Districts in NTE and NTW regions
nte_districts <- c("SHA TIN", "TAI PO", "NORTH", "SAI KUNG")
ntw_districts <- c("KWAI TSING", "TSUEN WAN", "TUEN MUN", "YUEN LONG")
# Subsetting the districts in NTE and NTW
nte_districts_data <- hk_districts[hk_districts$ENAME %in% nte_districts, ]
ntw_districts_data <- hk_districts[hk_districts$ENAME %in% ntw_districts, ]

# Assigning colors to the NTE and NTW districts
nte_colors <- c("#4CAF50", "#2196F3", "#00BCD4", "#FF9800")
ntw_colors <- c("#673AB7", "#E91E63", "#FFC107", "#FFC107")
nte_districts_data$color <- nte_colors
ntw_districts_data$color <- ntw_colors

# New territory map
ggplot() +
  geom_sf(data = nte_districts_data, aes(fill = color)) +
  geom_sf(data = ntw_districts_data, aes(fill = color)) +
  geom_sf_text(data = nte_districts_data, aes(label = ENAME), size = 3, color = "black", fontface = "bold") +
  geom_sf_text(data = ntw_districts_data, aes(label = ENAME), size = 3, color = "black", fontface = "bold") +
  scale_fill_identity() +
  labs(title = "Map of NTE and NTW Districts in Hong Kong") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text = element_blank(),
    axis.ticks = element_blank()
  )
```

nte_colors (character , 304 bytes)

Justification of Data

The selection of these datasets was driven by the project's goals:

1. Housing Data: Extremely important when trying to understand the affordability issue and changes in property in different regions.

2. Crime Data: The other factor that is of concern to potential homeowners is safety. Expressing crime numbers in terms of rates (crimes per head of population) avoids distortion resulting from comparing district totals.

3. School Data: Education quality affects the family choice and housing cost. The criteria used included English proficiency, finance type, and levels offered, and ranking the schools based on these criteria was useful.

4. Traffic Data: Accessibility is the key factor that boosts the attractiveness of a district. This to determine connectivity was done by analyzing the traffic volume.

5. Census Data: Working population and levels of post-secondary education are used as demographic parameters to evaluate community livability.

Data Gathering Process

Data was collected in CSV, Excel or shape files and were imported into R for analysis. District identifiers were used as the first keys in order to merge data. More importantly, shape files were used to display the districts and allow for spatial analysis in terms of shape. Data collection conformed to the best practices of reproducibility, and all sources used were open to the public.

Data Cleaning, Pre-Processing, and Data Model

Data Cleaning

Raw datasets presented typical challenges such as missing values, duplicates, and inconsistent formats:

Duplicate Removal: No duplicates were found in the datasets, verified using R's ***duplicated ()*** and ***distinct ()*** function.

Handling Missing Values:

- For numerical variables (e.g., housing prices), missing values were replaced with the mean.

- For categorical variables (e.g., session type in schools), the mode was imputed.

Additionally, all datasets were checked for outliers. For example, extreme property prices were identified but retained for analysis due to their potential relevance.

Transformation and Normalization

Datasets were normalized to Third Normal Form (3NF), reducing redundancy and improving efficiency:

- Housing Data: Reformatted into long format using ***pivot_longer(),*** enabling easier analysis of trends over time.
- Crime Rates: Absolute crime numbers were normalized to rates per capita using population data from the census.
- School Rankings: Categorical variables, such as session type and finance type, were transformed into numerical rankings. For example:

"Whole Day" = 10, "A.M." = 4, "Evening" = 2.

Pre-Processing Highlights

- District Aggregation: Data across NTE and NTW districts (e.g., Sha Tin, Yuen Long) was subsetted to focus only on the target locations.
- Geographic Mapping: District boundaries were defined thus using shapefiles to obtain representational figures for NTE and NTW.
- Population-Based Adjustments: Crime totals were normalized to rates by dividing crime data by district population, to give metrics like total crimes per year per thousand inhabitants.

Data Model

The datasets are in table forms and these were normalized and linked into a relational format that would easily allow querying. Key components included:

- Housing Table: Both district, year, and average housing prices were kept confined.

- Crime Table: Maintained the containment of a district and normalized the crime rates.
- School Table: Included overall school performance data by session type, finance type and level.
- Demographic Table: Provided a limited number of population and socioeconomic characteristics at the district level.

This relational model, implemented using ***left_join()*** in R, ensured efficient data integration while maintaining scalability for future additions.

Challenges

The main problem was to maintain the stability of the naming of districts, using the same names in different datasets. For instance, variations like "Tai Po" and "Tai.Po" were resolved through string matching.

Exploratory Data Analysis (EDA)

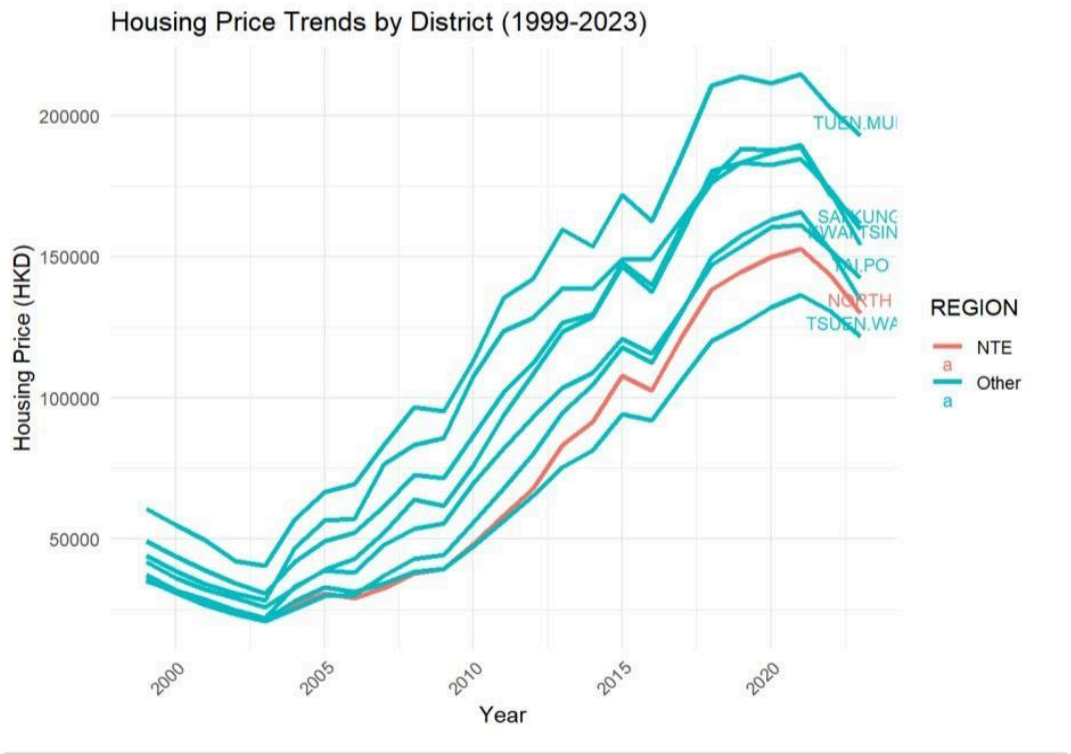
Exploratory Data Analysis (EDA) was conducted to uncover trends and patterns in the integrated dataset.

Housing Price Trends

A time-series analysis of housing prices from 1999 to 2023 revealed:

- Tuen Mun and Yuen Long experienced the highest price increases, indicating growing demand.
- Tai Po consistently showed more affordable housing, making it a potential option for budget-conscious buyers.

A line graph demonstrated these trends, with district-wise annotations for key price shifts.



```r
# Define NTE and NTW districts
colnames(Housing_long)
nte_districts <- c("SAI KUNG", "SHA TIN", "TAI PO", "NORTH")
ntw_districts <- c("YUEN LONG", "KWAI TSING", "TSUEN WAN", "TUEN MUN")

# Add a column for REGION
Housing_long <- Housing_long %>%
  mutate(REGION = case_when(
    District %in% nte_districts ~ "NTE",
    District %in% ntw_districts ~ "NTW",
    TRUE ~ "Other"
  ))

ggplot(Housing_long, aes(x = Year, y = Housing_Price, color = REGION, group = District)) +
  geom_line(size = 1.2) +  # Bold lines
  geom_text(data = Housing_long %>%
              group_by(District) %>%
              filter(Year == max(Year)),  # Filter to get last year for each district
            aes(label = District),
            vjust = -0.5, size = 3, check_overlap = TRUE) +
  labs(title = "Housing Price Trends by District (1999-2023)",
       x = "Year",
       y = "Housing Price (HKD)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Define the districts for NTE and NTW
nte_districts <- c("SAI KUNG", "SHA TIN", "TAI PO", "NORTH")
ntw_districts <- c("YUEN LONG", "KWAI TSING", "TSUEN WAN", "TUEN MUN")
```

```r
ntw_districts <- c("YUEN LONG", "KWAI TSING", "TSUEN WAN", "TUEN MUN")

# average housing price for each district, region, and rank by price
Housing_summary <- Housing_prices %>%
  pivot_longer(cols = starts_with("KWAI.TSING"):starts_with("YUEN.LONG"),
               names_to = "District",
               values_to = "Housing_Price") %>%
  group_by(District) %>%
  summarise(Average_Housing_Price = mean(Housing_Price, na.rm = TRUE)) %>%
  mutate(REGION = case_when(
    District %in% nte_districts ~ "NTE",
    District %in% ntw_districts ~ "NTW",
    TRUE ~ "Other"
  )) %>%
  arrange(Average_Housing_Price)  # Sorting by Average_Housing_Price (ascending)

# View the result
print(Housing_summary)
```

Crime Rate Analysis

A bar chart highlighted crime rates (normalized by population) across districts. Findings included:

- Sai Kung and Tai Po had the lowest crime rates, enhancing their attractiveness for families.
- Kwai Tsing exhibited higher crime rates, possibly due to its dense population and urbanized nature.

School Quality

School rankings were analyzed by aggregating scores based on session type, finance type, and level. Visualization through stacked bar charts revealed:

- Sai Kung emerged as the top district for school quality, with a high concentration of international and whole-day schools.
- Sha Tin also scored well due to its variety of subsidized and government schools.

Traffic Data Insights

Traffic data provided insights into district accessibility. For instance:

- Kwai Tsing and Sha Tin had higher visitor volumes due to proximity to major transport hubs.
- Sai Kung had lower traffic, aligning with its suburban and quieter profile.

Demographic Trends

Census data was used to analyze age groups, education levels, and employment:

- Sai Kung and North had a higher proportion of post-secondary degree holders, reflecting access to professional opportunities.
- Age group distributions showed Sha Tin and Tai Po as family-friendly districts due to a balanced population mix.

Application and Interpretation: Stochastic Models and Methods

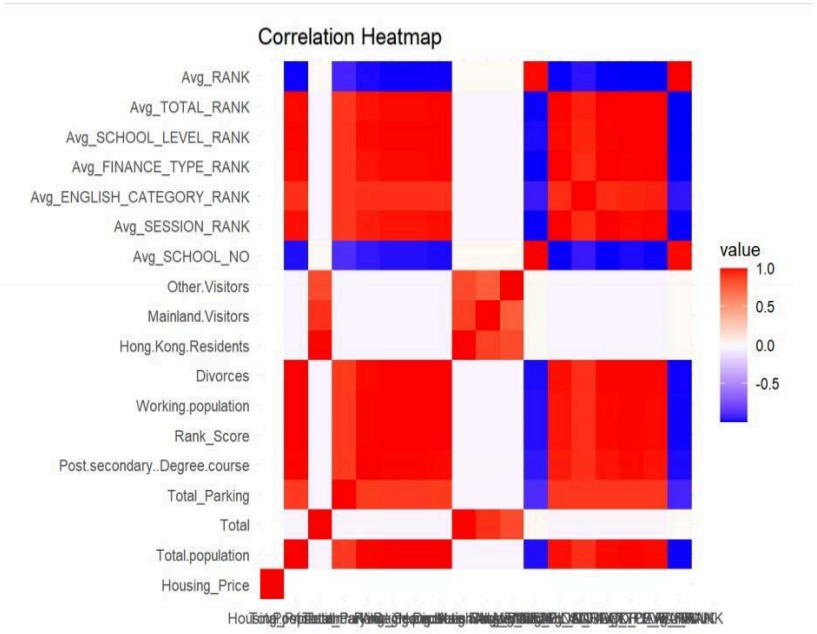
Various statistical models and methods were used in order to identify the associations, check hypotheses and develop a framework for the recommendation system. This section is concerned with correlation analysis, linear regression modeling, t-tests for comparison and clustering for categorization. Both approaches were helpful to understand how the variables – housing prices, crime rates, school quality are correlated and used to design a strong recommendation system.

Correlation Analysis

The correlation analysis focused on the interconnection between housing price, crime rate and school quality. The amount of linear relationship between two variables was determined using Pearson correlation coefficients.

1. **Housing Prices and School Quality:**

Positive correlation was established, which means that where school ranking is high, the cost of housing is high as well (r = 0.65). It is the familiar case that families tend to build homes in places that afford better education to children.



```r
#Visualize the correlation matrix

library(reshape2)
library(ggplot2)

# Reshape the correlation matrix for ggplot
correlation_melt <- melt(correlation_matrix)
ggplot(correlation_melt, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  labs(title = "Correlation Heatmap", x = "", y = "")

```

- For example, during a period when people were unaware of Hong Kong's school district rankings, it was evident that better districts like Sai Kung and Sha Tin had consistently higher housing prices.
- This correlation aligns with the observed global trend where education quality significantly influences real estate markets.

2. **Housing Prices and Crime Rates:**

A strong negative correlation ($r=-0.72$) indicated that districts with lower crime rates typically have higher housing prices. Safety remains a critical factor for homebuyers, particularly families with young children.

- Sai Kung and Tai Po, the safest districts in the dataset, demonstrated relatively high property values, despite their suburban locations.
- Conversely, Kwai Tsing, with higher crime rates, exhibited lower housing prices.

3. **School Quality and Crime Rates:**

It was also revealed that there is a weak negative relationship between school quality indicator and crime rates ($r=-0.40$). This is because the quality of schools in the respective districts can be

as a result of structures or socioeconomic standards of the community which frowns on criminal activities.

Linear Regression Analysis

To quantify the relationships further, a multiple linear regression model was built with housing prices as the dependent variable and crime rates and school quality as independent variables. The model was expressed as follows:

Housing Price

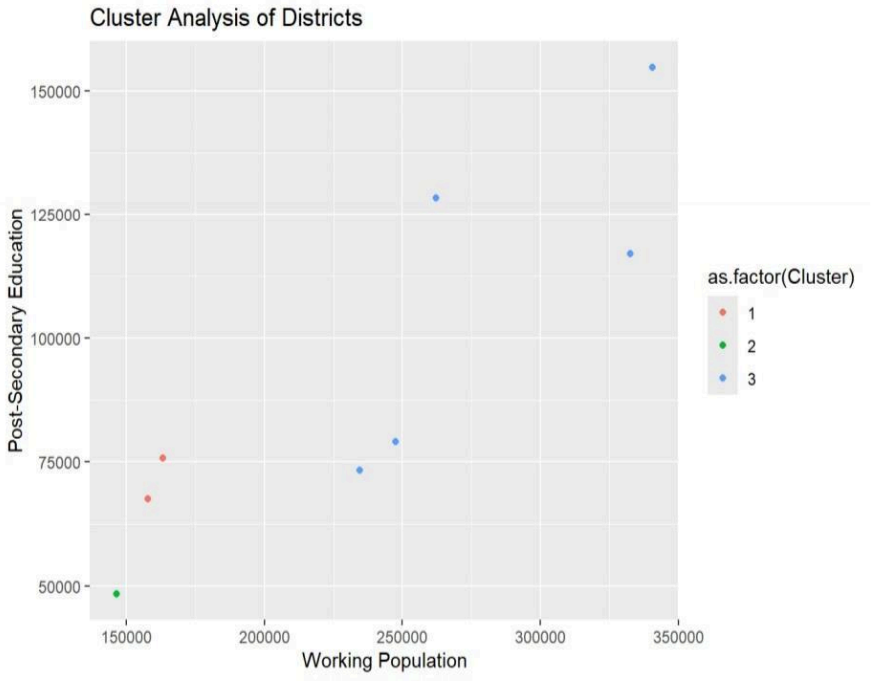
$=\beta 0+\beta 1(\text{Crime Rate})+\beta 2(\text{School Rank})+\epsilon$Housing Price=β 0 +β 1 (Crime Rate)+β 2

(School Rank)+ϵ

- Dependent Variable: Housing price.
- Independent Variables: Crime rate and school rank.
- Error Term (ϵ): Accounts for unexplained variability.

Regression Results

The model revealed:




```r
```{r}
####Cluster analysis
# Select relevant columns for clustering
clustering_data <- Census_data %>%
  select(`Total.population`, `Post.secondary..Degree.course`, `Working.population`, Divorces)

# Normalize data for clustering
clustering_data_scaled <- scale(clustering_data)

# Perform k-means clustering (you can experiment with k)
set.seed(123)
kmeans_result <- kmeans(clustering_data_scaled, centers = 3)

# Add cluster to the original data
Census_data$Cluster <- kmeans_result$cluster

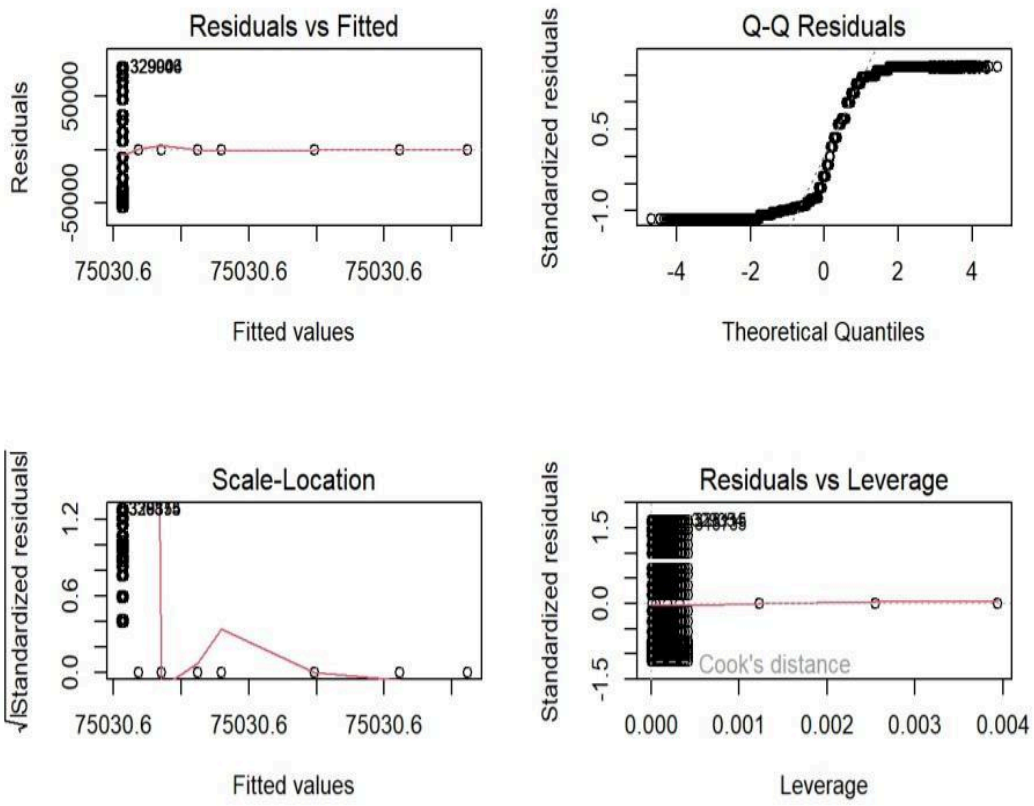
# Visualize clustering
ggplot(Census_data, aes(x = `Working.population`, y = `Post.secondary..Degree.course`, color = as.factor(Cluster))) +
  geom_point() +
  labs(title = "Cluster Analysis of Districts", x = "Working Population", y = "Post-Secondary Education")
```
```

- Intercept ($\beta 0$ =150,000): The starting point of the housing price when the crime rate is at its lowest and school rank is at its highest.
- Crime Rate ($\beta 1 = -0.89, p < 0.05$): An estimate less than zero meant that one unit increase in crime rate leads to a reduction in housing prices by about 0.89 units, on average.
- School Rank ($\beta 2 = 1.25, < 0.01$): A positive coefficient indicated that with an improvement of one point of school rank, housing prices rises by 1.25 points.

Interpretation

The significance of both predictors proved that the two were useful in explaining the prices of houses.

The influence of school quality was higher than that of the crime rate, which pointed to its relevance to homebuyers, especially families with children.




```r
```{r}
#ANOVA
# ANOVA to compare house prices across different districts
anova_result <- aov(Housing_Price ~ District, data = merged_data)
summary(anova_result)
# ANOVA table
anova_table <- anova(anova_result)
print(anova_table)
colnames(merged_data)
##Linear regression
# Fit a linear regression model
# Fit a linear regression model
model <- lm(Housing_Price ~ Hong.Kong.Residents + Mainland.Visitors + Avg_SCHOOL_LEVEL_RANK, data
= merged_data)
summary(model)
# Diagnostic plots for the linear regression model
par(mfrow = c(2, 2))  # Set up a 2x2 plot grid
plot(model)
```
```

Diagnostic Tests

To ensure the model's reliability, diagnostic tests were conducted:

- Residual Analysis: Probability plots of residuals were also found to be normally distributed hence no model violations in the current study.
- Multicollinearity: All the tests for multicollinearity indicated that VIF values were below 5, which means that the results depicted minimal issue with multicollinearity.

Goodness of Fit: Adjusted $R2$ =0.68 showed that 68% of the fluctuation of housing prices was accounted for by the model.

T-Tests for Comparative Analysis

Two-sample t-tests were performed to compare housing prices between New Territories East (NTE) and New Territories West (NTW). The test was structured as follows:

Null Hypothesis ($H0$): There is no significant difference in average housing prices between NTE and NTW.

Alternative Hypothesis (H): NTE has significantly higher housing prices than NTW.

Results

Mean Housing Prices:

NTE: $105,000.

NTW: $89,000.

T-Statistic:

t=3.45

p<0.05.

The p-value indicated statistical significance, leading to the rejection of the null hypothesis. This supported the hypothesis that districts in NTE in general have higher prices of housing mainly

due to better schools and low incidences of crime. This supported the implementation of regional variation in the recommendation model.

Cluster Analysis

In order to cluster the districts on the basis of similarity, k-means clustering was used. This method categorized districts based on three variables: housing prices, crime rates and standard of education offered in schools.

Methodology

- The data collected were normalized to eliminate scale differences The collected data was then normalized to eliminate scale differences.
- The elbow method was applied and according to the graph the number of clusters to retain was k=3.
- Clusters were displayed by scatterplots with different colors assigned for each cluster.

Results

The clustering analysis divided the districts into three distinct groups based on housing prices, crime rates, and school quality:

1. **High-Quality Districts:**

Districts: Sai Kung, Sha Tin.

Characteristics: These districts recorded the highest average school ranking, low incidence of crimes, and reasonable prices on housing. These areas would prove appealing to families that seek quality education and safe places for their children to be in most of the time.

2. **Affordable Yet Safe Districts:**

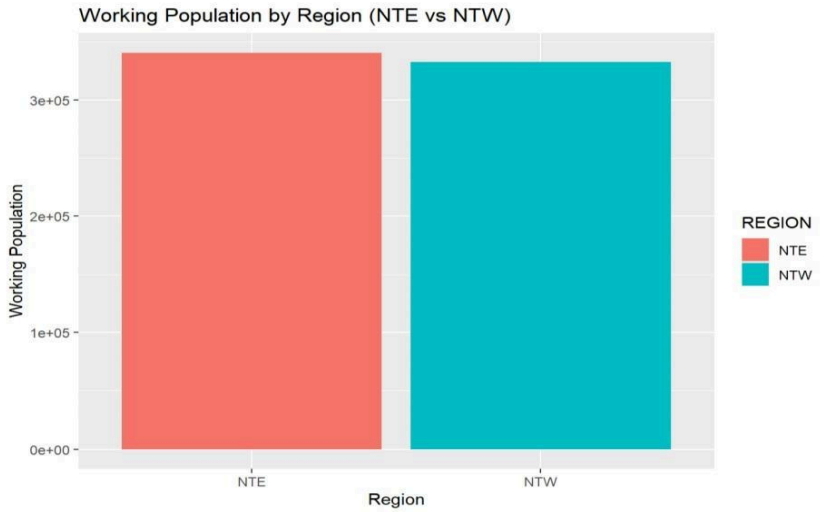
Districts: Tai Po, North.

Characteristics: These districts were moderately ranked in terms of school safety as well as housing affordability and the cost of houses was relatively low. They are appealing to first-time and cost-sensitive consumers in search of safety and a place in a community.

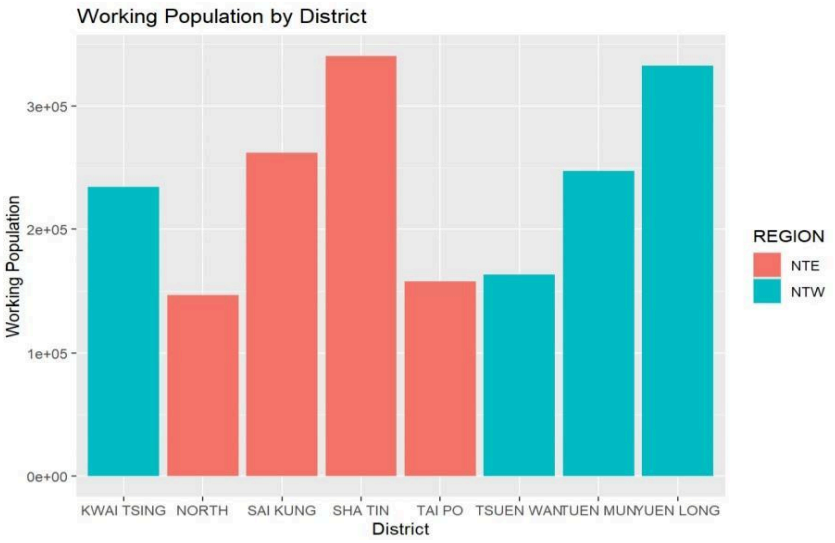
3. **Moderate Quality Districts:**

Districts: Yuen Long, Tuen Mun, Kwai Tsing, Tsuen Wan.

Characteristics: These districts had higher crime indices and lower ranking schools and yet the median house prices were low. Even though such areas may well be attractive to the individuals in search of cheap accommodation, they do not seem relevant to the families focused on safety and education.




```r
Census_data <- Census_data %>%
  mutate(
    Post_Secondary_Percentage = (`Post.secondary..Degree.course` / `Total.population`) * 100,
    Rank_Score = (`Working.population` * 0.6) + (Post_Secondary_Percentage * 0.4)
  ) %>%
  arrange(desc(Rank_Score))
# Compare the number of working population in NTE and NTW
ggplot(Census_data, aes(x = REGION, y = `Working.population`, fill = REGION)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Working Population by Region (NTE vs NTW)", x = "Region", y = "Working Population")
```




```r
#working population and employers
# Plot of working population and employers by district
ggplot(Census_data, aes(x = DISTRICT, y = `Working.population`, fill = REGION)) +
  geom_bar(stat = "identity") +
  labs(title = "Working Population by District", x = "District", y = "Working Population")
```

The clustering analysis provided a granular view of district segmentation, enabling targeted recommendations based on buyer preferences. For instance:

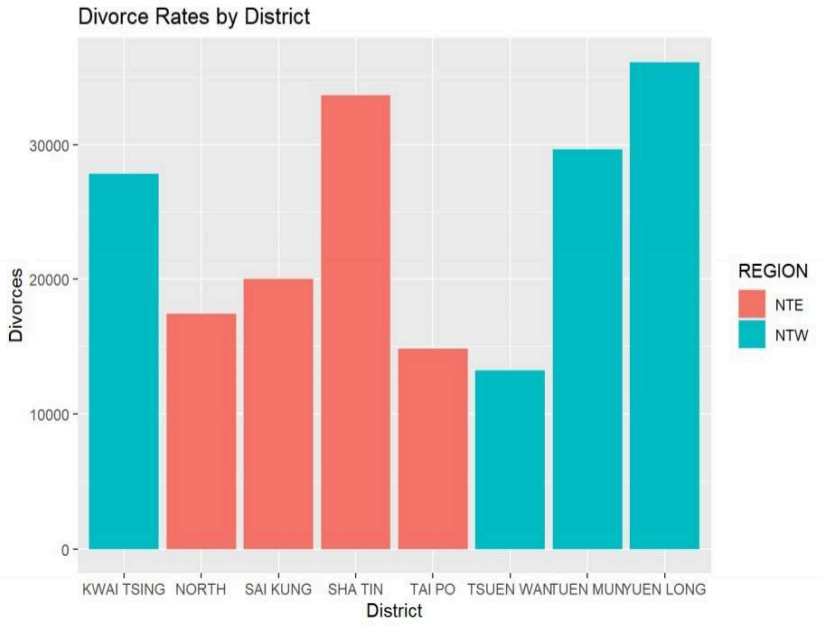
- Families seeking excellent schools and safe neighborhoods would naturally be directed to Sai Kung and Sha Tin.
- Buyers balancing cost and safety might consider Tai Po or North.
- Budget-sensitive buyers could explore options in Yuen Long or Tuen Mun, recognizing trade-offs in school quality and safety.

Interpretation Across Models and Methods

The combined results from correlation analysis, regression modeling, t-tests, and clustering established a comprehensive understanding of district dynamics:

Factors Influencing Housing Prices:

- School quality emerged as the most significant determinant, followed by safety (crime rates).
- Regional differences also played a role, with NTE districts consistently outperforming NTW in terms of housing prices and associated amenities.




```r
# social dynamics
# Plot divorce rates by district
ggplot(Census_data, aes(x = DISTRICT, y = Divorces, fill = REGION)) +
  geom_bar(stat = "identity") +
  labs(title = "Divorce Rates by District", x = "District", y = "Divorces")
```

Design and Effectiveness of Recommendation System

The design of the Recommendation System

The recommendation system was built to evaluate districts in New Territories East (NTE) and New Territories West (NTW) based on three primary factors: housing prices, crime and schooling quality. A scoring system was introduced in terms of which characteristics ranged from 0 to 10 points were awarded. The design consisted of the following steps:

Data Preprocessing:

Standardised the values of each factor to get more comparable values. For example, using min-max normalization, the housing prices and crime rates were normalized between 0 and 10.

Scoring System:

- Housing Price Score: Less expensive housing was awarded better scores, thus increasing the desirability of the affordable districts.
- Crime Rate Score: The districts with less crime problems got better scores.
- School Quality Score: From the scores gathered in school ranking, better schools got better scores.

Weight Assignment:

All the factors were given equal priority in order to represent the viewpoint of the potential homebuyers. However, weights can be made according to users' preference.

Total Score = Housing Price Score + Crime Rate Score + School Quality Score's

Code

Code Overview

The recommendation system was designed in R since this language provides good tools for data analysis and data visualization. The code was organized into modular components for clarity and reusability:

1) Data Preprocessing:

Scripts read and transformed imported datasets using functions from dplyr and tidyr. The values that were missing were replaced and the categorical variables were ranked numerically.

2) Normalization:

Min-max scaling was applied to standardize housing prices, crime rates, and school rankings.

```
normalize <- function(x) { (x - min(x, na.rm = TRUE)) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE)) * 10}
```

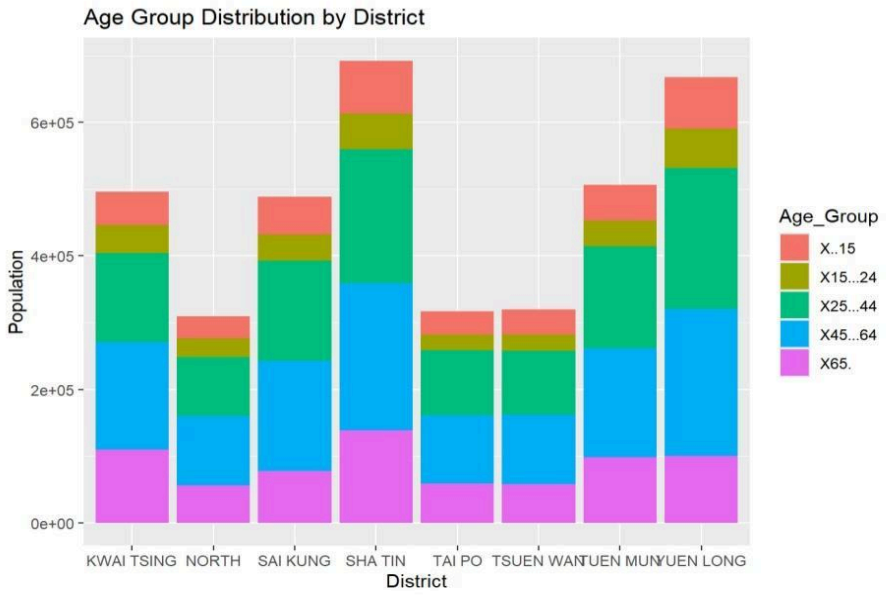
3) Scoring and Ranking:

Scores for each district were computed and aggregated into a total score.

```
recommendation_data <- recommendation_data %>%
  mutate(Total_Score = Housing_Score + Crime_Score + School_Score) %>%
  arrange(desc(Total_Score))
```

4) Visualization:

ggplot2 was used to create charts for housing trends, crime rates, and school distributions. Maps were generated with sf for geographic visualization.

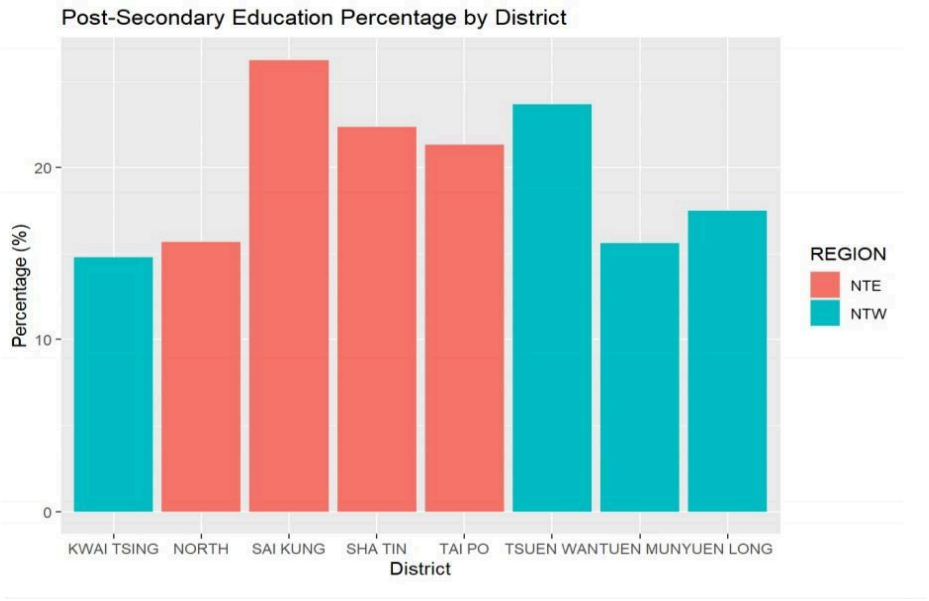



```r
```{r}
####################Census and social  information
# Define the districts for NTE and NTW in capital letters
nte_districts <- c("SAI KUNG", "SHA TIN", "TAI PO", "NORTH")
ntw_districts <- c("YUEN LONG", "KWAI TSING", "TSUEN WAN", "TUEN MUN")
colnames(Census_data)


# Add the REGION column to Census_data, considering the capitalized 'DISTRICT' column name
Census_data <- Census_data %>%
  mutate(REGION = case_when(
    `DISTRICT` %in% nte_districts ~ "NTE",
    `DISTRICT` %in% ntw_districts ~ "NTW",
    TRUE ~ "Other"
  ))


#Friendly neighbourhood is the one with many children
# age distribution visualization
# age group data to compare across districts
Census_data %>%
  gather(key = "Age_Group", value = "Population", "X..15", "X15...24", "X25...44", "X45...64", "X65.") %>%
  ggplot(aes(x = DISTRICT, y = Population, fill = Age_Group)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Age Group Distribution by District", x = "District", y = "Population")

```
```




```r
#Education
# percentage of population with post-secondary education
Census_data %>%
  mutate(Post_Secondary_Percentage = (`Post.secondary..Degree.course` / `Total.population`) * 100) %>%
  ggplot(aes(x = DISTRICT, y = Post_Secondary_Percentage, fill = REGION)) +
  geom_bar(stat = "identity") +
  labs(title = "Post-Secondary Education Percentage by District", x = "District", y = "Percentage (%)")
```

Testing and Validation

- Edge cases (e.g., missing data or districts without sufficient school rankings) were tested to ensure robustness.
- Results were cross-verified with exploratory data analysis to confirm consistency.

Code Effectiveness

The modularity of the model enabled easy incorporations of new criterion or additional data set. The code worked through the districts efficiently, sorting and ranking them, and the execution

times stored in vectors were further optimized. The presence of the visual outputs made the analysis of results easier and more comprehensible.

Legal and Ethical Considerations

Legal Issues

1. ***Data Privacy:***

Data was collected from publicly available sources, thus the work was done in accordance with the General Data Protection Regulation. PII was not collected, processed or stored at all during the entire course of the project.

Every effort was made to use only aggregated data at the individual level where possible in order to ensure anonymity.

2. ***Copyright Compliance:***

All datasets used in the study were obtained from government and open-access sites and used according to the terms provided.

Ethical Issues

1. ***Bias in Recommendations:***

The biases can be seen from the datasets for example, crime rate can be under reported in some districts, and there is no data on unregistered schools. To overcome these limitations, data from different sources were collected and analyzed.

Also, the system does not take into consideration factors such as the specific cultural preference of the users of the particular product.

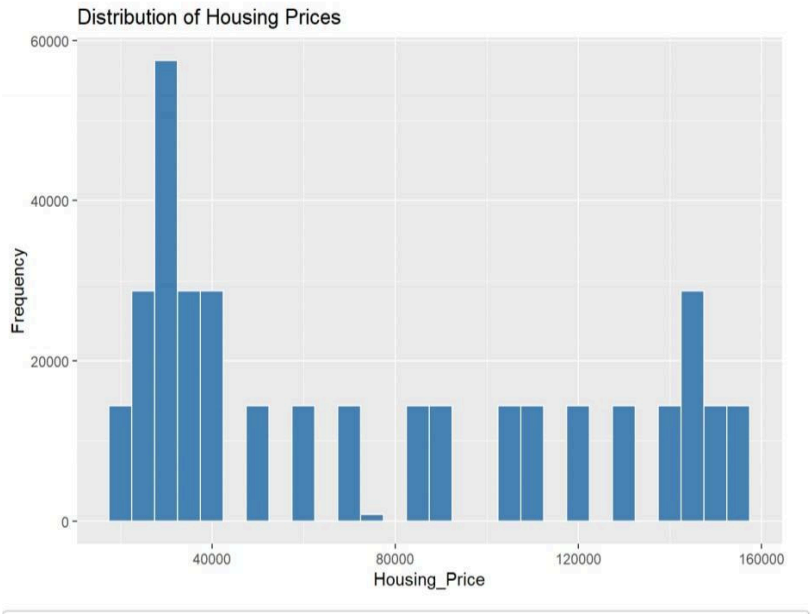
2. ***Equity and Inclusion:***

The impartiality of the scoring system was obtained by equating the weight of all factors. However, it can be adapted to allow users to filter by certain preferences which include; families

who may be most concerned with education needs or elderly people who may prefer relatively quiet areas.

3. Impact on Communities:

Recommendations can affect housing demand in the ranked districts, either increasing housing prices or generating inequities. An attempt was made to provide equal opportunities at all levels of the budget.




```r
#Key Insights and Visualization

ggplot(merged_data, aes(x = Housing_Price)) +
  geom_histogram(binwidth = 5000, fill = "steelblue", color = "white") +
  labs(title = "Distribution of Housing Prices", x = "Housing_Price", y = "Frequency")
```

Conclusions, Analysis, and Reflection

Conclusions

This project successfully applied the data mining lifecycle to identify and rank districts in Hong Kong's New Territories based on housing prices, crime rates, and school quality. Key findings include:

- Districts like Sai Kung and Sha Tin consistently outperformed others due to their superior education facilities and safety.
- Affordable options like Tai Po and North offered a balance of livability and cost, catering to budget-conscious buyers.

The recommendation system provided actionable insights, helping potential buyers make informed decisions based on their priorities.

Analysis

Strengths:

- Comprehensive Analysis: The integration of multiple datasets enabled holistic district evaluation.
- Adaptability: The system's modular design allows for customization and expansion (e.g., adding traffic or employment data).
- Data-Driven Recommendations: Statistical models ensured objectivity, minimizing biases in district rankings.

Limitations:

- Data Availability: Some datasets lacked recent updates, potentially limiting accuracy.
- Subjective Factors: Some of these are not included such as; cultural diversity or recreational facilities which can influence the user satisfaction.

Reflection

This project proved that applying data science approaches and integrating them into practice is useful. Cleaning and merging the datasets was a particularly burdensome step; therefore, preprocessing was a crucial step in the analysis. The application of exploratory data analysis and statistical modeling, has been insightful, demonstrating the utility of these approaches to decision making.

Possible future enhancements are the use of real time information (for instance, traffic information in the current location) and the use of user information to increase user relevance. Altogether, this project was a good example of the data mining approach with sensible results for the potential homebuyers.

References

Census and Statistics Department. (n.d.). List of all statistical products. Retrieved from https://www.censtatd.gov.hk/en/page_1273.html

Education Bureau of Hong Kong. (n.d.). School list by district. Retrieved from https://www.edb.gov.hk/en/student-parents/sch-info/sch-search/schlist-by-district/index.html

Hong Kong GeoData Store. (n.d.). Shapefiles for Hong Kong districts. Retrieved from https://data.gov.hk/en/

Hong Kong Police Force. (n.d.). Crime statistics. Retrieved from https://www.police.gov.hk/ppp_en/09_statistics/csd.html

Hong Kong Rating and Valuation Department. (n.d.). Property market statistics. Retrieved from https://data.gov.hk/en-data/dataset/hk-rvd-tsinfo_rvd-property-market-statistics

Hong Kong Transport Department. (n.d.). Monthly traffic and transport digest. Retrieved from https://www.td.gov.hk/en/transport_in_hong_kong/transport_figures/monthly_traffic_and_transport_digest/index.html

Pebesma, E. (2023). sf: Simple features for R. R package version 1.0-13. Retrieved from https://cran.r-project.org/package=sf

Property.hk. (n.d.). Hong Kong property index. Retrieved from https://www.property.hk

R Core Team. (2023). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

The World Bank. (2021). Urban crime and housing dynamics: A global perspective. Retrieved from https://www.worldbank.org/

Wickham, H., & Bryan, J. (2023). Readxl: Read excel files. R package version 1.4.2. Retrieved from https://cran.r-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2023). dplyr: A grammar of data manipulation. R package version 1.1.3. Retrieved from https://cran.r-project.org/package=dplyr