ISyE 6740 – Summer 2025
Final Project Report

Team Member Names: Team 028, Li Yat Hei Jack

Project Title

Finding possible expansion region(s) and their medical charge prediction for a physiotherapy clinic using demographic data at Hong Kong

Problem Statement

The physiotherapy industry in Hong Kong is facing significant challenges in adapting to the demographic shifts driven by an ageing population. Currently, one out of every eight residents is aged 65 or above, with an average life expectancy of 80 years for men and 86 years for women as of 2010. Projections indicate that in 20 years, one in every four Hong Kong residents will be 65 or older. By 2050, the World Health Organization forecasts that Hong Kong will rank fifth globally for cities with the highest percentage of older adults, with an estimated 40% of the population being 65 or older *(Hong Kong population projections for 2022-2046, 2025)*.
This trend of population ageing is expected to persist, particularly as post-war baby boomers enter old age and life expectancy continues to rise. The number of elderly individuals aged 65 and over is projected to nearly double from 1.45 million in 2021 to 2.74 million by 2046. Consequently, the proportion of elderly in the population is anticipated to increase from 20.5% to 36.0%, meaning more than one in every three Hong Kong residents will be elderly.*(Excel@PolyU, 2025)*
Given this context, existing physiotherapy clinics face challenges in identifying optimal locations for expansion that would effectively meet the needs of this growing demographic. Current expansion strategies often lack a data-driven approach, leading to potential missed opportunities in underserved areas. This project aims to analyze demographic data in Hong Kong to identify potential expansion regions for a physiotherapy clinic. By leveraging geographic and demographic insights, we seek to determine areas with a high concentration of target populations—such as seniors, individuals with chronic conditions, and young adults engaged in sports—who would benefit from physiotherapy services. The findings will guide strategic decisions on clinic locations, ensuring that services are accessible to those who need them most.

The objective of this project is to assist a physiotherapy clinic in identifying potential regions for expansion in Hong Kong, as well as predicting medical charges based on demographic data. Given that insurance coverage is a significant factor influencing medical charges, the project aims to:

1. Identify regions in Hong Kong where a physiotherapy clinic could expand its services.
2. Visualize these potential expansion regions on a geospatial map.
3. Develop a predictive model for medical charges, emphasizing the impact of insurance coverage.

Data Source

Due to data security reasons, I cannot obtain demographic data from the Hong Kong government, unless I’m a government officer. Therefore, I use simulated datasets from a python library called Faker. It will create 5000 demographic data points. The features include Age, Ethnicity,Gender,Income,Residential Region,Insurance Coverage, Insurance Amount, Sports,Height and Weight.

Methodology

This project employs a data-driven approach to identify potential expansion regions for a physiotherapy clinic in Hong Kong using simulated demographic data. The methodology consists of several key steps:

Data Generation

To simulate the demographic data for the project, I use the ***Faker*** library in Python to create a dataset containing 5,000 data points with the following features:

Feature Variable	Description	Value
Age	Age of a patient	Integer between 18 and 70
Ethnicity	Ethnicity of a patient	One of "Chinese", "Filipinos", "South Asians", "Whites" and "Other"
Gender	Gender of a patient	'Male' or 'Female'.
Income	Income of a patient	Integer between 20,000 and 80000 in HKD
District	District where the patient resides	One of "Wan Chai", "Central and Western", "Eastern", "Southern", "Kowloon City", "Yau Tsim Mong", "Wong Tai Sin", "Sham Shui Po", "Kwun Tong", "Tsuen Wan", "North", "Sha Tin", "Islands", "Yuen Long", "Tai Po", "Tuen Mun", "Kwai Tsing", "Sai Kung
Residential Region	A region a district belongs to	One of "Hong Kong Island", "Kowloon" and "New Territories"
Insurance Coverage	Indicate whether the patient has insurance coverage	Either "Yes" or "No"
Insurance Amount	Insurance amount	If Insurance Coverage = "Yes", the insurance amount is Integer range from 300 – 3000. Insurance amount is 0 if Insurance Coverage = "No"
Sports	Patient's regular sport	One of "Football", "Basketball", "Tennis", "Badminton", "Swimming", "Jogging","Gymnastics", "Fencing", "Cycling","Dragon Boat", "Rugby" and "Martial Arts"
Height	Height of a patient	Integer between 140 and 200 in cm
Weight	Weight of a patient	Integer between 50 and 150 in kg
Pricepoint	Medical charge on the patient per treatment	Integer range from 600 to 3000 in HKD

Analysis and Prediction

In the realm of strategic expansion, a comprehensive analysis of demographic data is crucial. By examining factors such as age, income levels, and sports participation, we can identify regions with heightened demand for services. This analysis will help pinpoint areas where our offerings could resonate most effectively with potential customers, thereby maximizing our outreach and impact.

To visualize these potential expansion regions, we will employ geospatial mapping techniques using the Folium library in Python. This powerful tool allows us to create interactive maps that clearly illustrate the identified areas of opportunity.

(*Moos, N., Juergens, C., & Redecker, A. P. , 2021*) By overlaying demographic data onto these maps, stakeholders can easily grasp the geographical landscape and make informed decisions regarding resource allocation and marketing strategies.

Furthermore, understanding the financial implications of expansion is essential. We will implement a Linear Regression model using the scikit-learn library to predict medical charges in these targeted regions. By giving particular weight to insurance coverage in our model, we can achieve more accurate predictions that reflect real-world scenarios. This predictive analytics approach will not only guide pricing strategies but also ensure that our services remain accessible to a broader audience. (*Madhuri, C. R., Anuradha, G., & Pujitha, M. V. , 2019*)

Evaluation and Final Results

The evaluation of our predictive model is a critical step in ensuring its effectiveness and reliability. To assess the model's performance, we will employ several key metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. These metrics will provide a comprehensive understanding of how well our model predicts medical charges, allowing us to quantify its accuracy and make necessary adjustments. By analyzing these performance indicators, we can ensure that our model meets the desired standards of precision and robustness.

In addition to performance metrics, examining feature importance is vital for understanding the factors that influence our predictions. By focusing on the role of insurance coverage, we can confirm its significant impact on medical charge estimates. This analysis not only validates our model's structure but also offers insights into how various factors interact to shape healthcare costs, thus guiding future enhancements to our predictive approach.

Moreover, visualizing potential expansion regions based on demographic data is essential for strategic planning. Using the Folium library, we will create interactive maps that highlight these regions, making the data easily accessible and interpretable. To further enhance our analysis, we will develop choropleth maps that represent medical charge predictions across different areas. This visual representation will allow stakeholders to identify trends and disparities in medical costs, supporting informed decision-making regarding resource allocation and market entry strategies.

Model selection

I started off with Linear Regression and Random Forest for price prediction because Linear Regression is a straightforward and interpretable model that assumes a linear relationship between the independent variables—such as Age, Income, and Weight—and the dependent variable, which is the Pricepoint. This model is beneficial when the relationships in the data are linear and when the primary goal is to understand the influence of individual demographic factors on pricing. However, it can struggle with non-linear relationships and interactions among variables, which are common in complex datasets.

On the other hand, Random Forest is a more robust and flexible ensemble learning method that can capture complex non-linear relationships and interactions between demographic features like Ethnicity, Gender, and Insurance Coverage. It operates by constructing multiple decision trees and averaging their predictions, which helps in reducing overfitting and improving accuracy. Given the diverse nature of demographic data, including categorical variables such as District and Region, Random Forest can effectively handle these complexities, making it a strong candidate for capturing the nuances in price prediction for physiotherapy.

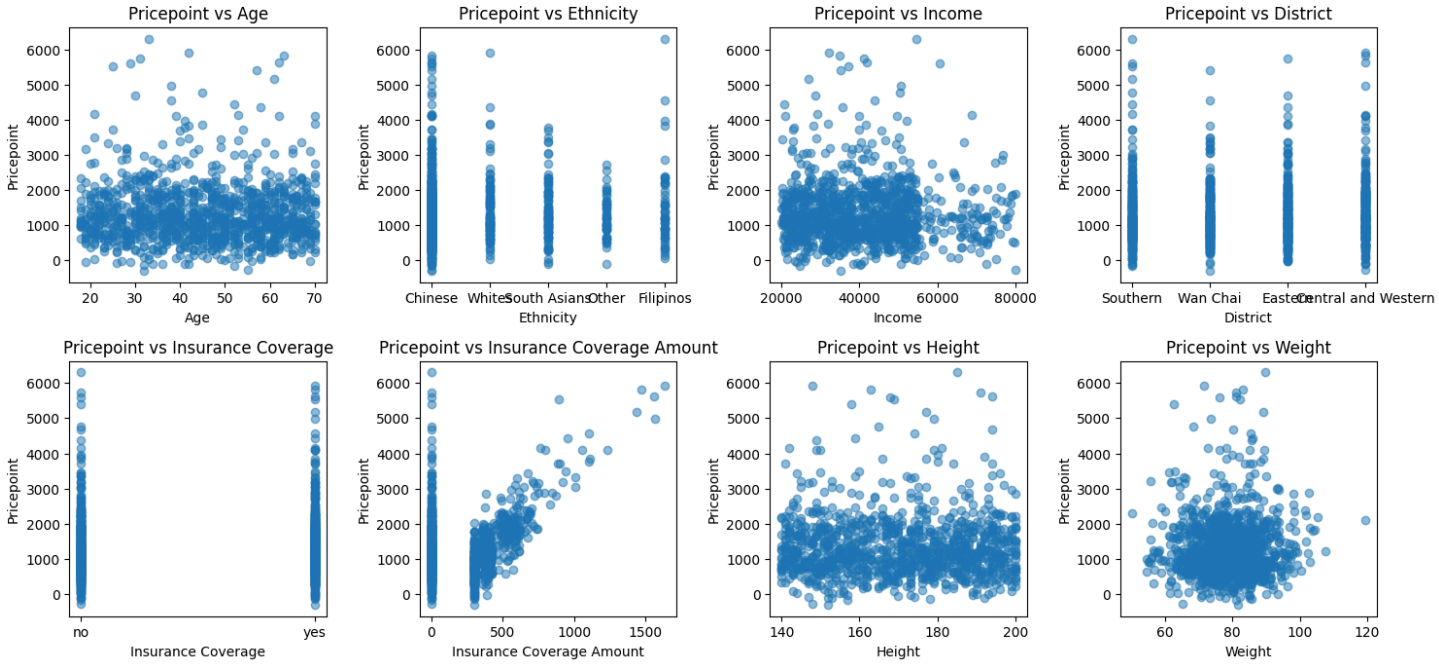
While both models have their strengths and weaknesses, I select the model based on the accuracy of the pricepoint of the physiotherapist's advice, the R-squared value and Mean Absolute Error and Mean Squared Error. Here are the results of price prediction of both models

Model	**Random Forest**	**Linear Regression**
Mean	1357.3	1360.0
Standard Deviation	625.7	600.0
Quantiles (25th, 50th, 75th)	1025, 1245, 1469	1090, 1280, 1390
Mean Absolute Error	524	513
Mean Squared Error	547959	526774
R-squared (R^2)	0.3897	0.4133

Linear regression performs slightly better than Random Forest in all metrics. The physiotherapists said that the mean, standard deviation and quartiles align closer to their current price points. In addition, Linear Regression allows me to investigate which factors are essential to affect price points. Hence, I choose the Linear Regression model for price prediction.

<u>Relationship investigation between Pricepoint and all dependent variables</u>

I run a set of plots on Pricepoint with each independent variable to find any relationship between them so I can create a better regression model. For instance, if Pricepoint has a quadratic relationship with Income, then I will include this relationship in the regression equation.



Based on the plots, the price point shows no significant relationships with most variables, except for the Insurance Coverage Amount, which exhibits a moderate linear relationship. This finding aligns with common intuition: as the insurance coverage amount increases, so does the price point that patients are willing to pay for

treatment. Patients often prefer clinics that maximize their insurance benefits, leading to higher expenditures when coverage is substantial.

It's important to note that while higher insurance coverage can encourage patients to seek treatment by reducing out-of-pocket costs, it does not guarantee that patients will pursue care solely because their coverage is extensive. Additionally, we observe that when insurance coverage is zero, there is no discernible pattern in the price point. This scenario reflects the behavior of uninsured patients, who often seek treatment due to the urgency of their injuries, prioritizing immediate care over cost considerations.

Feature Analysis

Generalized Linear Model Regression Results

Dep. Variable:	Pricepoint	No. Observations:	893
Model:	GLM	Df Residuals:	870
Model Family:	Gaussian	Df Model:	22
Link Function:	Identity	Scale:	4.5276e+05
Method:	IRLS	Log-Likelihood:	-7070.3
Date:	Mon, 21 Jul 2025	Deviance:	3.9390e+08
Time:	14:35:59	Pearson chi2:	3.94e+08
No. Iterations:	3	Pseudo R-squ. (CS):	0.4479
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1168.2388	285.913	4.086	0.000	607.859	1728.618
Age	-1.3025	1.546	-0.842	0.400	-4.333	1.728
Income	-0.0004	0.002	-0.241	0.810	-0.004	0.003
Insurance Coverage Amount	3.9950	0.178	22.461	0.000	3.646	4.344
Height	-0.4179	1.386	-0.302	0.763	-3.134	2.298
Weight	3.1639	2.670	1.185	0.236	-2.070	8.397
Ethnicity_Filipinos	84.0489	120.217	0.699	0.484	-151.573	319.671
Ethnicity_Other	-89.9111	104.257	-0.862	0.388	-294.251	114.429
Ethnicity_South Asians	2.5629	80.727	0.032	0.975	-155.658	160.784
Ethnicity_Whites	120.6867	79.335	1.521	0.128	-34.806	276.180
Gender_Male	33.7667	45.821	0.737	0.461	-56.041	123.574
Insurance Coverage_yes	-1706.2124	88.607	-19.256	0.000	-1879.879	-1532.545
Sports_Basketball	-41.1242	121.543	-0.338	0.735	-279.343	197.095
Sports_Cycling	-42.9539	125.541	-0.342	0.732	-289.010	203.102
Sports_Dragon Boat	47.5323	118.892	0.400	0.689	-185.492	280.557
Sports_Fencing	-47.6199	121.336	-0.392	0.695	-285.435	190.195
Sports_Football	11.5427	118.035	0.098	0.922	-219.801	242.886
Sports_Gymnastics	59.0713	119.981	0.492	0.622	-176.086	294.229
Sports_Jogging	139.8693	122.834	1.139	0.255	-100.880	380.619
Sports_Martial Arts	48.8629	122.900	0.398	0.691	-192.016	289.742
Sports_Rugby	68.7152	123.726	0.555	0.579	-173.784	311.215
Sports_Swimming	76.3000	116.244	0.656	0.512	-151.534	304.134
Sports_Tennis	-40.4919	120.589	-0.336	0.737	-276.843	195.859

According to the regression model, both Insurance Coverage and its amount are highly significant predictors of the price point, while the variables weights, Sports_Jogging, and Ethnicity_Whites show moderate significance.

This indicates that the clinic should prioritize understanding various insurance coverage schemes, particularly regarding the types of injuries they encompass. By partnering with insurance companies, the clinic can streamline the process for patients, enabling them to easily identify which injuries are covered without having to navigate the complexities on their own. This collaboration not only enhances patient satisfaction but also benefits both the clinic and the insurance providers by potentially increasing sales, as patients are more likely to seek treatment when they have clarity about their coverage.

Conversely, the moderately significant variables—weights, Sports_Jogging, and Ethnicity_Whites—suggest that the clinic should target specific demographics. Focusing on white individuals, jogging enthusiasts, and those with higher body weights may yield better results, as these groups are statistically more likely to seek physiotherapy services compared to others. Tailoring marketing strategies to engage these audiences could enhance the clinic's outreach and patient acquisition efforts.

Price Prediction

	District	Mean	Standard Deviation	Quantiles (25th, 50th, 75th)
0	Tsuen Wan	800.0	530.0	[270.0, 820.0, 1310.0]
1	Islands	820.0	540.0	[300.0, 770.0, 1330.0]
2	North	820.0	530.0	[310.0, 930.0, 1310.0]
3	Tuen Mun	820.0	530.0	[260.0, 880.0, 1310.0]
4	Tai Po	830.0	550.0	[270.0, 920.0, 1320.0]
5	Yuen Long	850.0	520.0	[290.0, 1200.0, 1300.0]
6	Kwai Tsing	860.0	520.0	[280.0, 1170.0, 1320.0]
7	Sha Tin	870.0	520.0	[340.0, 1200.0, 1320.0]
8	Sai Kung	880.0	500.0	[330.0, 1190.0, 1330.0]
9	Sham Shui Po	1010.0	450.0	[580.0, 1190.0, 1340.0]
10	Wong Tai Sin	1030.0	420.0	[630.0, 1210.0, 1340.0]
11	Kwun Tong	1030.0	410.0	[620.0, 1230.0, 1340.0]
12	Kowloon City	1040.0	470.0	[600.0, 1230.0, 1350.0]
13	Yau Tsim Mong	1050.0	460.0	[630.0, 1240.0, 1350.0]
14	Southern	1290.0	480.0	[1040.0, 1270.0, 1380.0]
15	Eastern	1300.0	440.0	[1110.0, 1280.0, 1360.0]
16	Wan Chai	1320.0	490.0	[1050.0, 1290.0, 1390.0]
17	Central and Western	1410.0	690.0	[1200.0, 1290.0, 1410.0]

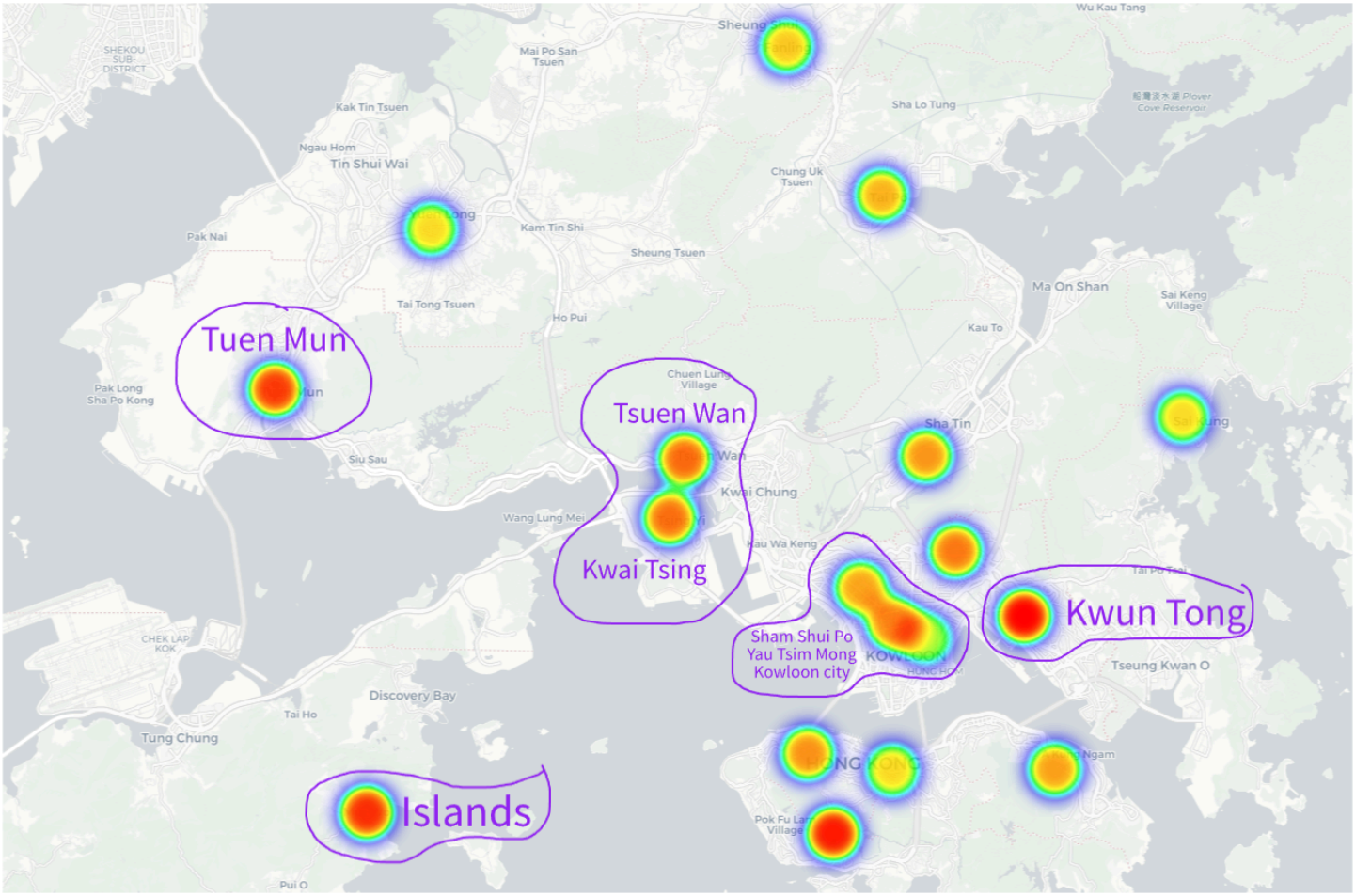
New Territories

Kowloon

Hong Kong Island

Based on the table, the price points for Kowloon and New Territories are lower than Hong Kong Island in terms of mean. Kowloon has the lowest standard deviation while New Territories is highest.

Geospatial Analysis and Age by district



The heat map illustrates the population distribution across various districts, with the intensity of red indicating higher population density. Districts such as Yau Tsim Mong, Sham Shui Po, Kowloon City, Kwun Tong, Tuen Mun, Islands, Tsuen Wan, and Kwai Tsing all exhibit significant population concentrations. Notably, these areas are characterized by a high number of apartments and office towers, leading to a daily influx of workers who prefer to visit clinics conveniently located near their workplaces for efficiency.

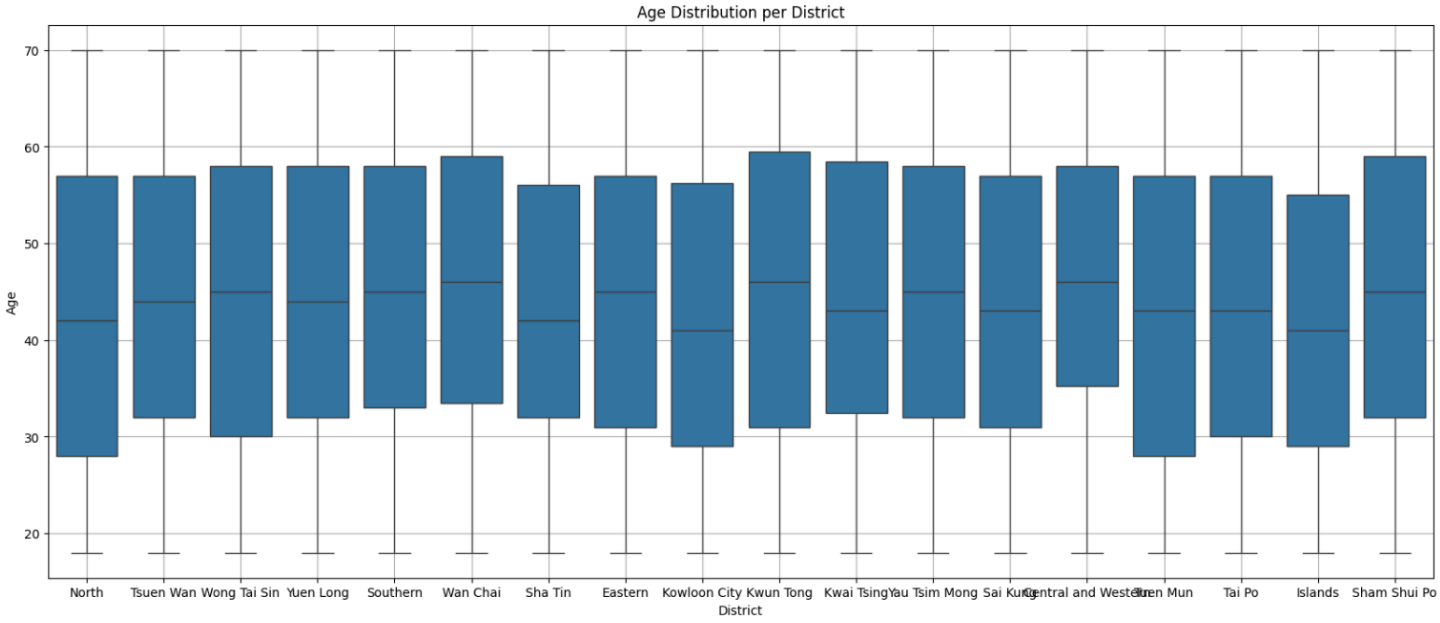
While all these districts appear promising for expansion, the ideal locations should be focused on: i) Yau Tsim Mong, Sham Shui Po, and Kowloon City; ii) Tsuen Wan and Kwai Tsing; or iii) Kwun Tong. These districts are home to stadiums and sports arenas, making them particularly advantageous for establishing a physiotherapy clinic. Proximity to sports facilities can significantly enhance accessibility for individuals seeking treatment, ultimately driving higher patient volume and increasing clinic revenue.

In contrast, although Tuen Mun and Islands districts showcase high population figures, their rural nature poses accessibility challenges, which may deter potential patients from seeking physiotherapy services.

Age by District

It is crucial to analyze the age distribution within each district, as individuals aged 1 to 25 and those over 75 are generally less likely to seek physiotherapy. Young children, for instance, often lack the financial means to afford treatment. They may also underestimate their injuries or choose self-care over professional assistance, as their healthcare interventions typically focus on developmental milestones rather than physiotherapy.

On the other hand, older adults frequently encounter mobility challenges that hinder their ability to visit physiotherapists. Additionally, many prioritize treatments for chronic conditions over physiotherapy, which can further limit their access to these services. Understanding these age-related dynamics is essential for effectively targeting physiotherapy services in each district.



The graph suggests that age groups across all districts are similar. In other words, the clinic can choose any districts to expand.

Conclusion

Kowloon and the New Territories exhibit a lower price point compared to the Hong Kong Island region. Potential expansion areas include i) Yau Tsim Mong, Sham Shui Po, and Kowloon City; ii) Tsuen Wan and Kwai Tsing; and iii) Kwun Tong, all of which boast high population densities. The average price points and ranges for these three groups are as follows: Yau Tsim Mong (average 1,030; range 630-1,350), Sham Shui Po (average 830; range 500-1,200), and Kowloon City (average 1,030; range 620-1,340).

Recently, the Hong Kong government has made substantial investments in building sports facilities in the Kowloon City district. This area features two running stadiums, one bowling alley, two bouldering arenas, 20 badminton courts, three basketball courts, and six multipurpose indoor arenas. Given its concentration of sports facilities, I believe Kowloon City represents the most promising expansion location. The average price point in Kowloon City is 1,040, with a range of 600-1,350, making it an attractive option for physiotherapy services aimed at active individuals.

Future directions

Increasing Sample Size

To enhance the reliability and validity of our findings, it is essential to increase the sample size for our analysis. A larger sample will provide more comprehensive data, allowing us to draw more accurate conclusions about potential expansion locations for our physiotherapy clinics. By capturing a broader range of demographics and behaviors, we can better understand the needs and preferences of various communities, ultimately leading to more informed decision-making.

Transportation Modes for Geospatial Analysis

When conducting geospatial analysis to identify potential expansion locations, it is crucial to consider the transportation modes individuals use to travel to work and other destinations. Many people prefer to visit clinics that are conveniently located near their workplaces for ease and efficiency. By mapping out transportation patterns and assessing clinic proximity to major employment centers, we can strategically position our clinics to maximize accessibility and attract a steady flow of clients throughout the week.

Competitor Analysis

Incorporating the number of competitors per district into our analysis is vital for assessing market saturation. Areas with a high concentration of physiotherapy clinics may indicate a competitive landscape that could impede our expansion efforts. By identifying districts with fewer competitors, we can target our resources more effectively, ensuring that we establish our clinics in regions where demand is likely to outpace supply, thus increasing our chances for success.

Rental Fees Consideration

Another important factor to consider is the rental fees associated with each district. The cost of leasing commercial space can significantly impact our operational budget and overall profitability. By analyzing rental prices across different districts, we can identify areas where the cost of entry aligns with our financial goals, allowing for sustainable growth while minimizing financial risk associated with high-rent locations.

Seasonality Effects

Lastly, it is essential to take seasonality into account when planning our expansion. The patterns of outdoor and indoor activities vary significantly between seasons, influencing the demand for physiotherapy services. For instance, during the summer months, individuals are more likely to engage in outdoor sports, potentially leading to a higher incidence of related injuries. Conversely, winter may see an uptick in indoor activities. By aligning our marketing strategies and service offerings with seasonal trends, we can optimize our clinic operations and better meet the needs of our clients throughout the year.

Bibliography

1) *Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019). House Price Prediction Using Regression Techniques: A Comparative study. House Price Prediction Using Regression Techniques: A Comparative Study, 1–5. https://doi.org/10.1109/icsss.2019.8882834*

2) *Moos, N., Juergens, C., & Redecker, A. P. (2021). Geo-Spatial analysis of population density and annual income to identify Large-Scale Socio-Demographic disparities. ISPRS International Journal of Geo-Information, 10(7), 432. https://doi.org/10.3390/ijgi10070432*

3) *Hong Kong population projections for 2022-2046 released [15 Aug 2023]*. (n.d.). Census and Statistics Department. Retrieved June 23, 2025, from https://www.censtatd.gov.hk/en/press_release_detail.html?id=5368#:~:text=With%20post%2Dwar%20baby%20boomers,people%20will%20be%20an%20elderly.

4) *Excel@PolyU*. (n.d.). Retrieved June 24, 2025, from https://www.polyu.edu.hk/cpa/Excel@PolyU/2011/11/viewpoint.html#:~:text=The%20ageing%20trend%20is%20becoming,will%20be%2065%20or%20above.

5) *Plotting with Folium — GeoPandas 1.1.1+0.ge9b58ce.dirty documentation*. (n.d.). Retrieved June 27, 2025, from https://geopandas.org/en/stable/gallery/plotting_with_folium.html