**Data Analytics Capstone Report**

*Problem Statement*

Property prices have been rising in the last few years. The objective of this project aims to build a predictive model to estimate the intrinsic value of a property unit based on its characteristics and features, with a focus on private residential sector in the core central region (CCR) in Singapore.

*Data Wrangling*

Dataset:

Private Residential Property Transactions; Private Residental Rental Contracts

Data Period: Jan-2018 to Mar-2023

Source: URA site (www.ura.gov.sg)

*Data Cleaning:*

*Table “Price” -*

1. There are 3 transaction entries marked “Land” in the “Type of area” column. Removed them as they are not considered in the analysis.
2. Removed “Type of area” column as all entries are now “Strata”.
3. Removed “Area (SQM)” column as there is already another column “Area (SQFT)” which is more relevant.
4. Removed “Unit Price ($ PSM)” column as there is already another column “Unit Price ($PSF) which is more relevant.
5. Removed “Nett Price($)” column. There is 14,696 (99.2%) of null values out of 14,813 rows and there is already another “Transacted Price ($)” column.
6. Removed “Number of units” column. There is 14,785 (99.8%) rows with value 1 out of 14,813 rows making this column not relevant to analysis.
7. Removed “market segment” column. All values in this column are marked “core central region” which is also the scope of this project.
8. Format the data in “Sale Date” column to date data type.
9. Add a column “Proj\_ID” with unique identifier for each project.

Table “Rental” -

1. Removed “Property Type” column. All values is this column are “Non-Landed Properties” and column is irrelevant to analysis.
2. Replaced all “NA” values in the “No of Bedrooms” column to “0” to make all data in this column to be integers.
3. Removed “Floor Area (SQM)” column as it is relevant and there is already another “Floor Area (SQFT)” column.
4. Changed “Lease Commencement Date” data type to date data type.
5. Add a column “Proj\_ID” with unique identifier for each project.
6. Added a column “Rental\_psf” to calculate montly rental divided by unit floor area to have a standardized rental matric on same area unit.

Table “Project” -

1. Create a new table with “Proj\_ID” and “Project Name” for all those projects identified in the “Prices” and “Rental” tables.
2. Collect information on project details of selected projects for further analysis.

*Data Dictionary:*

Table: Prices

|  |  |  |
| --- | --- | --- |
| **Columns** | **Data Type** | **Description** |
| Proj \_ID | *Int* | *Unique Identifer for each project* |
| Project\_Name | String | Project Name for each project |
| Transacted\_Price | Int | Selling Price for each transaction in S$ |
| Area | Float | Unit area for unit transacted in sqft |
| Unit\_Price | Int | Transacted Price / Area in psf |
| Sale\_Date | Date | Date of transaction |
| Street\_Name | String | Address of unit |
| Type\_of\_Sale | String | New Sale, Sub sale, Resale |
| Property\_Type | String | Apartment, Condominium |
| Tenure | String | 99 years, Freehold |
| Postal\_District | Int | District 9, 10, 11 |
| Floor\_Level | Category | Categories of transacted unit |

Table: Rental

|  |  |  |
| --- | --- | --- |
| **Columns** | **Data Type** | **Description** |
| Proj\_ID | *Int* | *Unique Identifer for each project* |
| Project\_Name | String | Project Name for each project |
| Street\_Name | String | Address of Project |
| Postal\_District | Int | District 9, 10, 11 |
| No\_of\_Bedroom | Int | Number of bedrooms in unit |
| Monthly\_Rent | Float | Monthly rental in S$ |
| Floor\_Area | Float | Area of unit in sqft |
| Lease\_Commence\_Date | Date | Start date of unit rental |
| Rental\_psf | Float | Monthly rental / floor area |

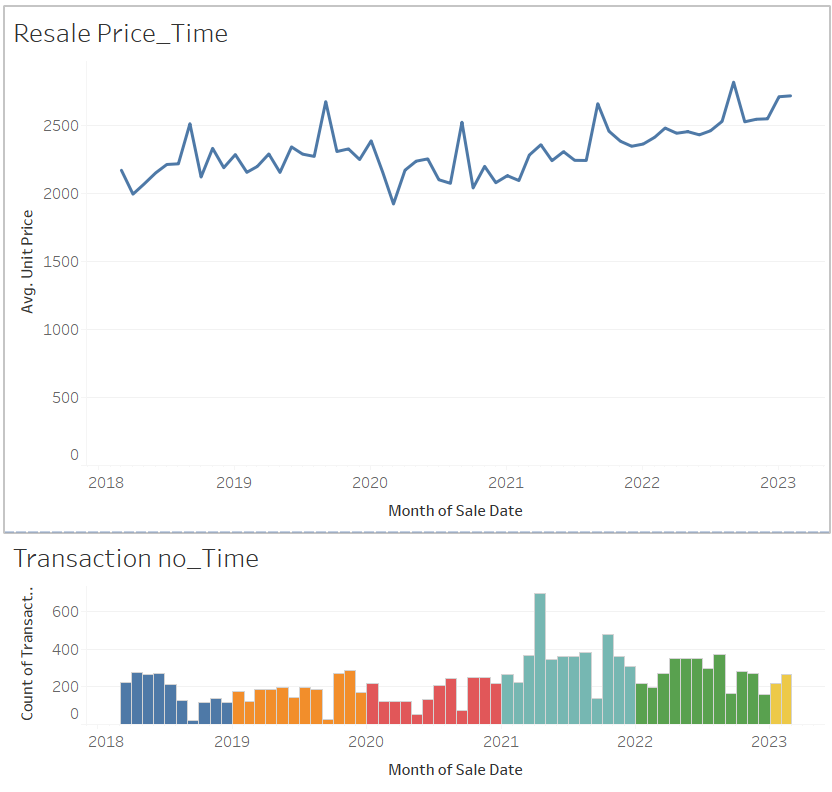
Table: Projects

|  |  |  |
| --- | --- | --- |
| **Columns** | **Data Type** | **Description** |
| Proj\_ID | *Int* | *Unique Identifer for each project* |
| Project\_Name | String | Project Name for each project |
| District | Int | District 9, 10, 11 |
| Developer | String | Developer of Project |
| Site Area (sqm) | Float | Size of project development |
| No\_of\_units | Int | Number of units in project |
| Density | Float | Proj\_size / No\_of\_units |
| Leasehold | String | 99 leasehold, Freehold |
| TOP\_Date | Date | Year of project TOP |
| Dist\_to\_mrt (m) | Float | Distance to nearest mrt station |
| MRT | String | Nearest MRT station |
| Prop\_Type | String | Apartment, Condominium, Mixed Development |

*Exploratory Data Analysis (EDA)*

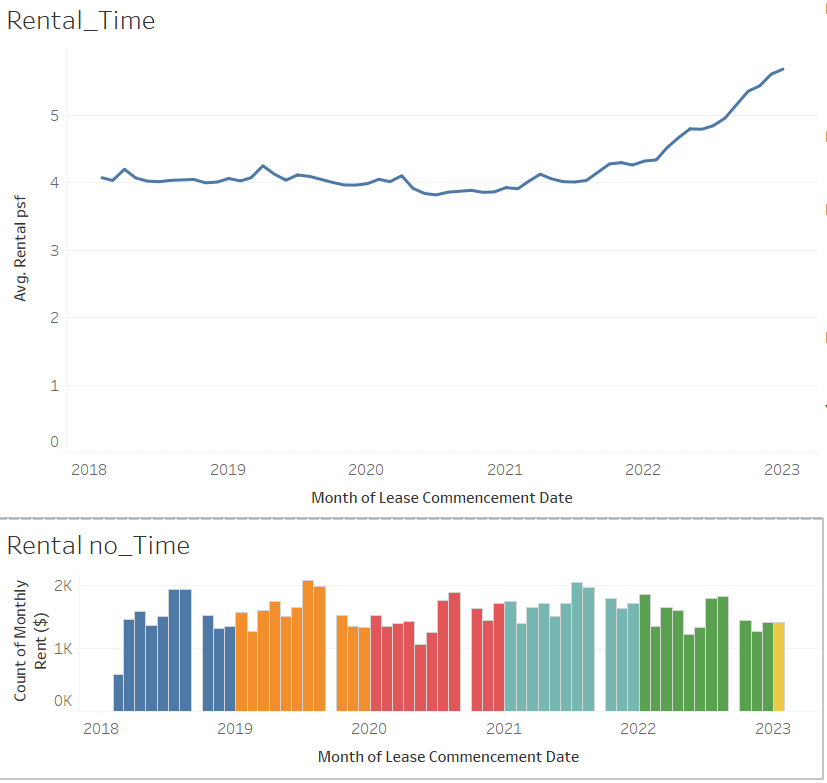
What is the trend of private properties resale prices over the years?

There are wide fluctuations of average resale price with a general slight uptrend from year 2018 to 2023. There seems to be higher transaction volume in the year 2021. Prices ranged between a low at around $2,000 psf around early 2020 to $2,800 near end-2021.



What is the trend of rental rates over the years?

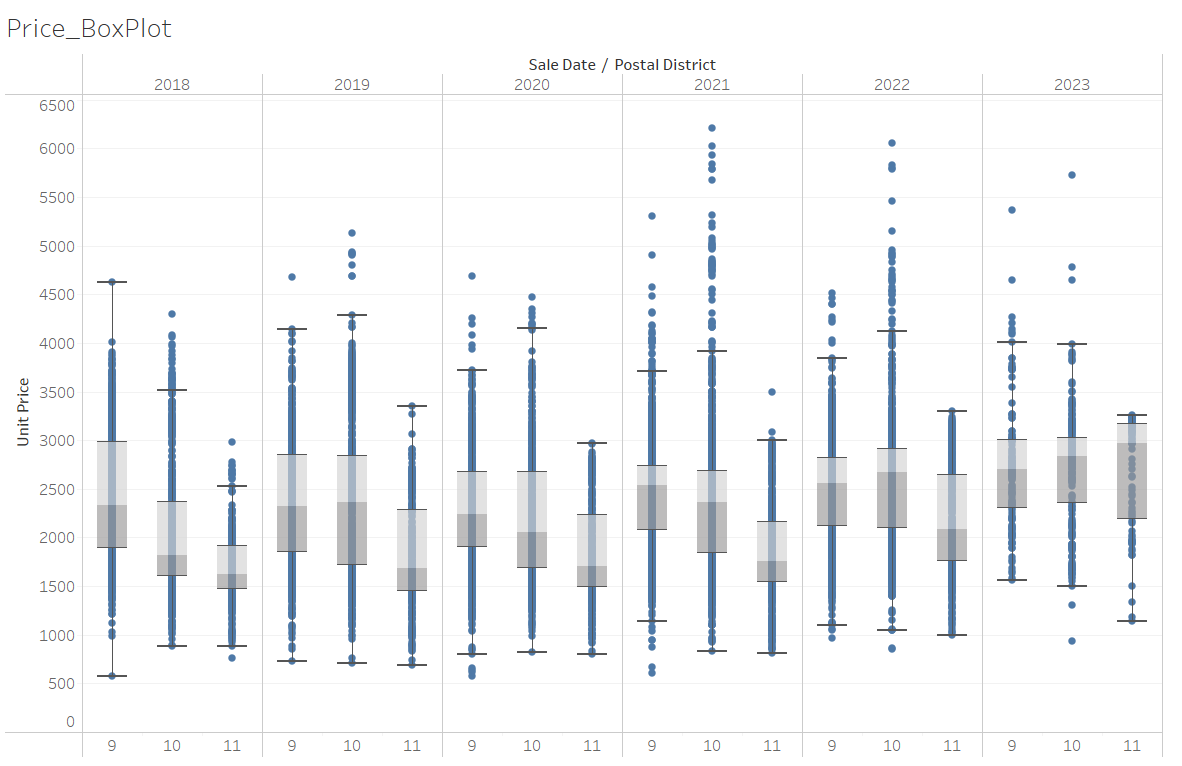
The average rental was generally flat from 2018 to 2021 followed by a steady climb since mid-2021 till current early-2023. Rental volume was generally steady over the years. There don’t seem to be higher rental demand since 2022 to support the rental increase.



Is there any variance in resale prices across different districts?

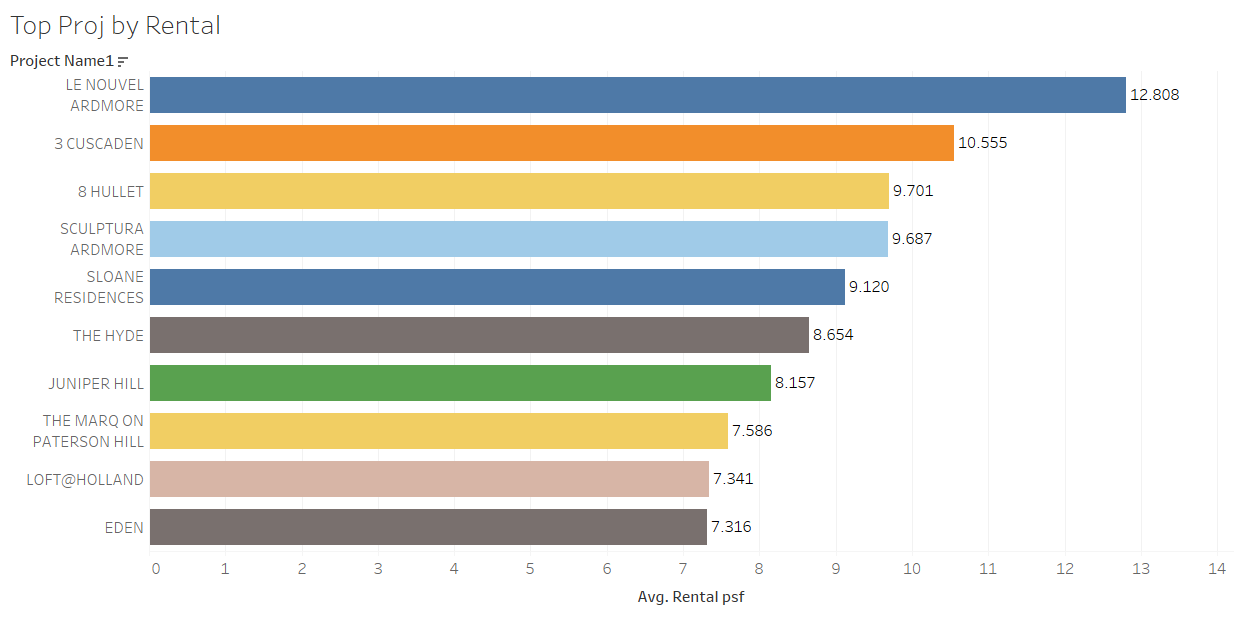
District 9 seems to have a generally higher median transacted psf compared to Districts 10 and 11.

District 10 has a greater range of transacted prices observed in year 2021 and 2022.



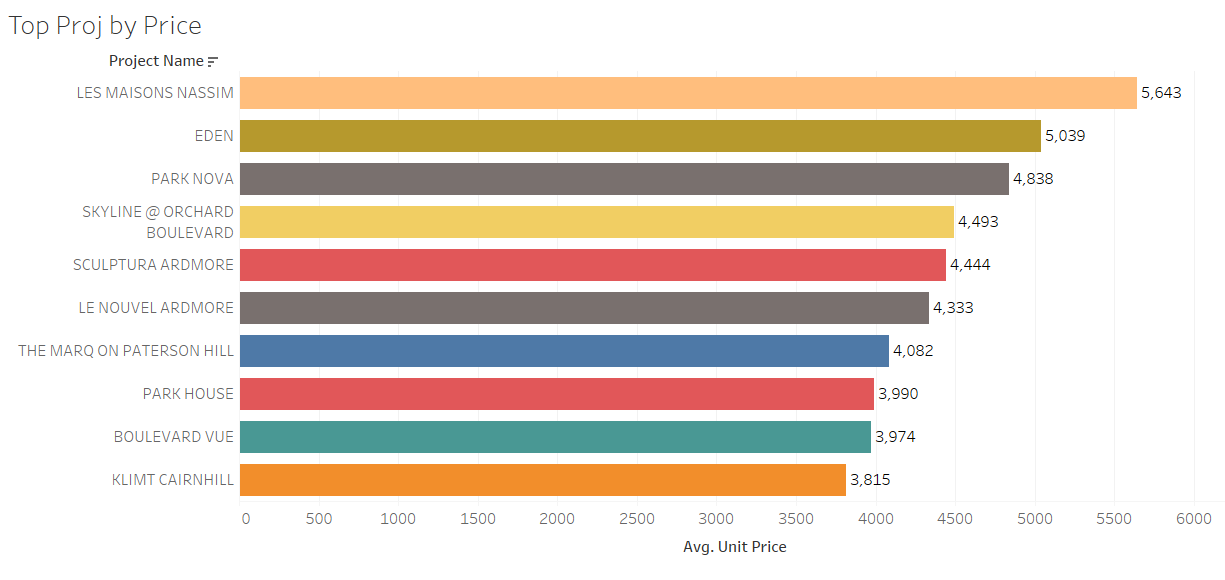
Which are the top 5 projects that have the highest rental.

They are Le Nouvel Ardmore, 3 Cuscaden, B Hullet, Sculptura Admore and Sloane Residences.



Which are the top 5 projects that have the highest resale prices.

They are Les Maisons Nassim, Eden, Park Nova, Skyline @ Orchard Boulevard, and Sculptura Admore.

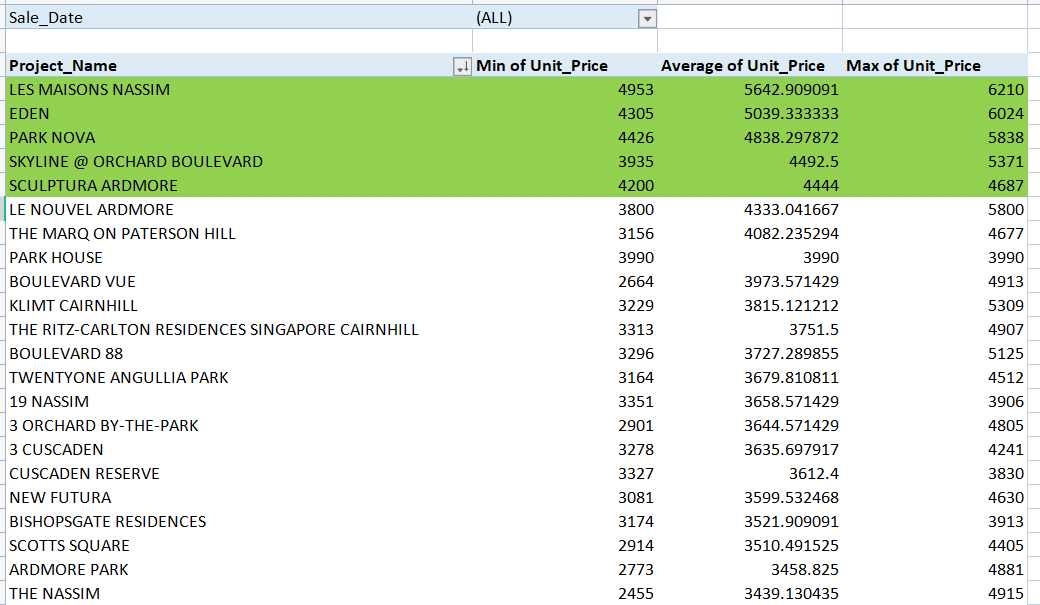


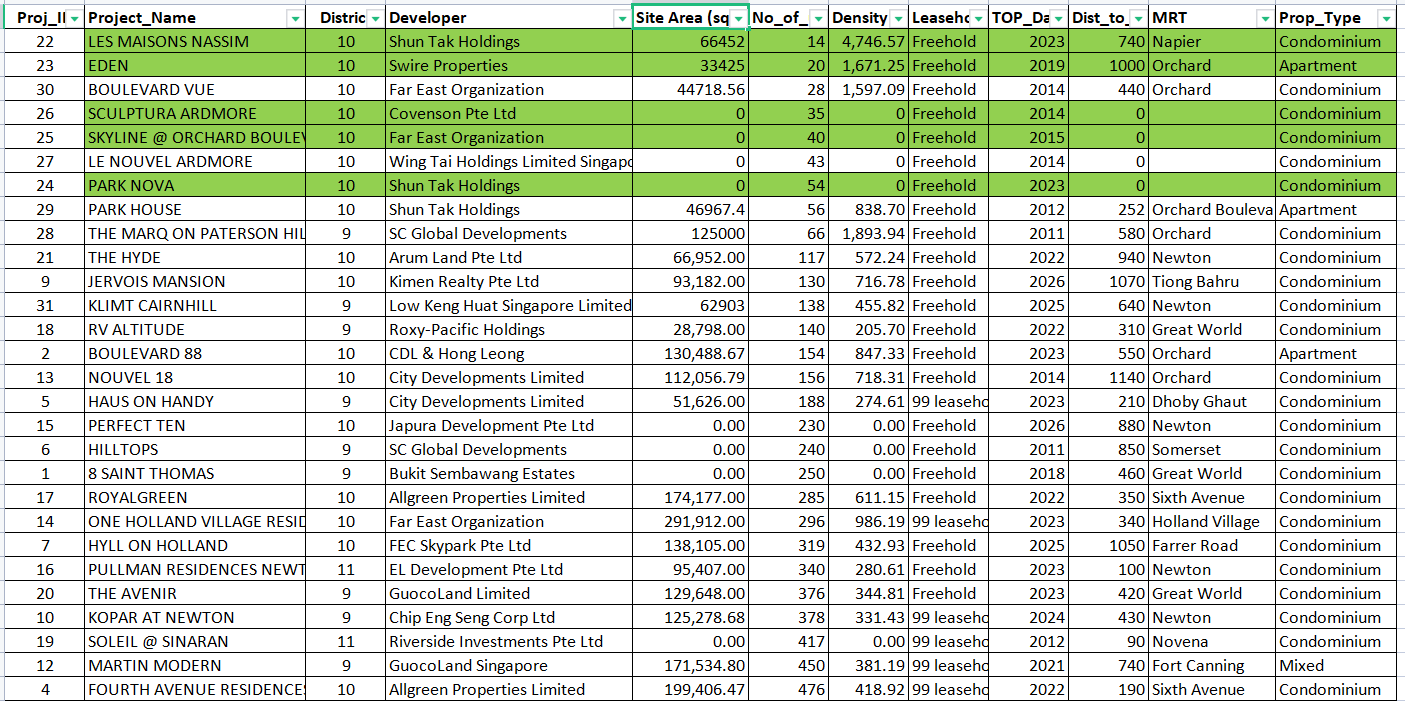
Is there any common features of projects in the Top 5 by highest mean resale prices?

Below is a pivot table of transactions grouped by projects with the Top 5 projects with highest mean resale prices highlighted green.

These projects have minimal number of units in each project giving them a higher site area per unit.

This make each unit appear more prestigious and buyers are willing to pay a higher price for them due to the limited supply.





*Modelling*

*Preparing Data:*

1. Narrow the scope by including only those projects within the top 20 by transaction volume and top 10 in term of resale prices that have complete data. This filtered down the data for analysis down to total 23 projects and 4,718 transactions.
2. Drop outliners
3. Add new columns to dummify the features as below so as to quantify the categorical features:

Column “D\_TypeOfSale” -

New Sale = 1; Sub Sale = 2; Resale = 3

Column “D\_PropertyType” -

Apartment = 1; Condominium = 2

Column “D\_Tenure” -

99-years leasehold = 1; Freehold/999-years leasehold = 2

Column “D\_FloorLevel” -

0 = data not available

1 = floor B1 to B5

2 = floor 01 to 05

3 = floor 06 to 10

4 = floor 11 to 15

5 = floor 16 to 20

6 = floor 21 to 25

7 = floor 26 to 30

8 = floor 31 to 35

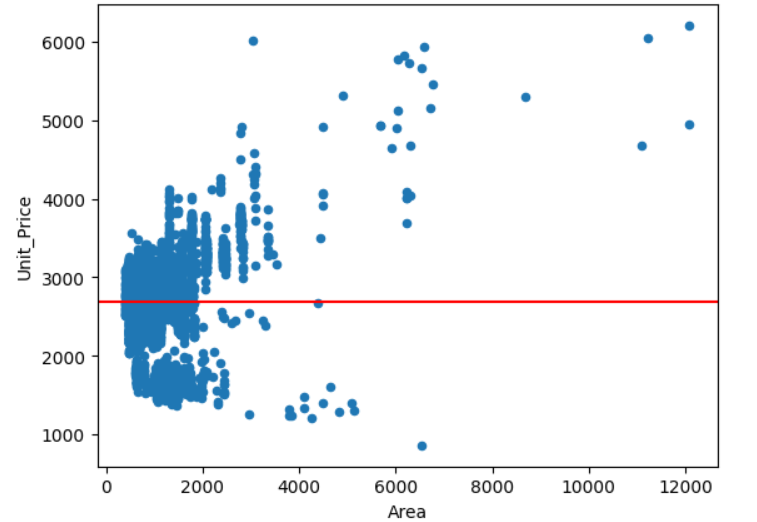
9 = floor 36 to 40

10 = floor 41 to 45

11 = floor 46 to 50

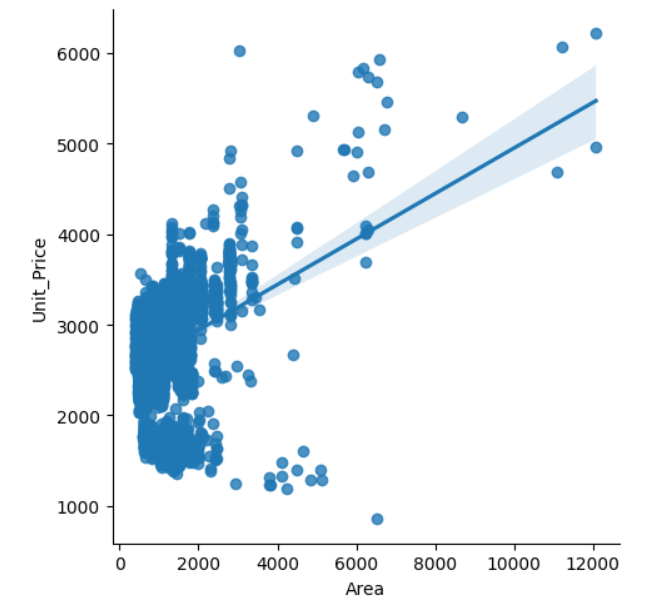
|  |  |  |
| --- | --- | --- |
| **Model** | **Description** | **Remarks** |
| 1. Benchmark model | Y = y-mean | Taking mean of y as predicted y value |
| 1. Simple Linear Regression Model | Y = mX + C | Where  Y is dependent variable “Unit\_Price”,  X is independent variable “Area”,  m is coefficient,  C is y-intercept |
| 1. Multi-Factor Linear Regression Model | Y = m1(X1) + m2(X2) + … m5X5 + C | Where  Features X are [Area, PropertyType, Tenure, Unit\_Age, Dist\_to\_mrt] |

1. *Benchmark Model*





1. *Simple Linear Regression Model*



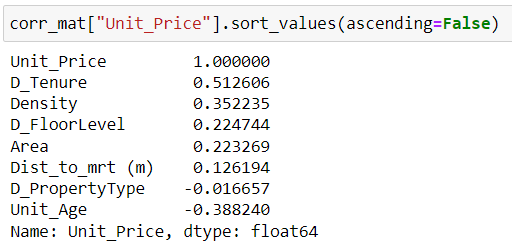


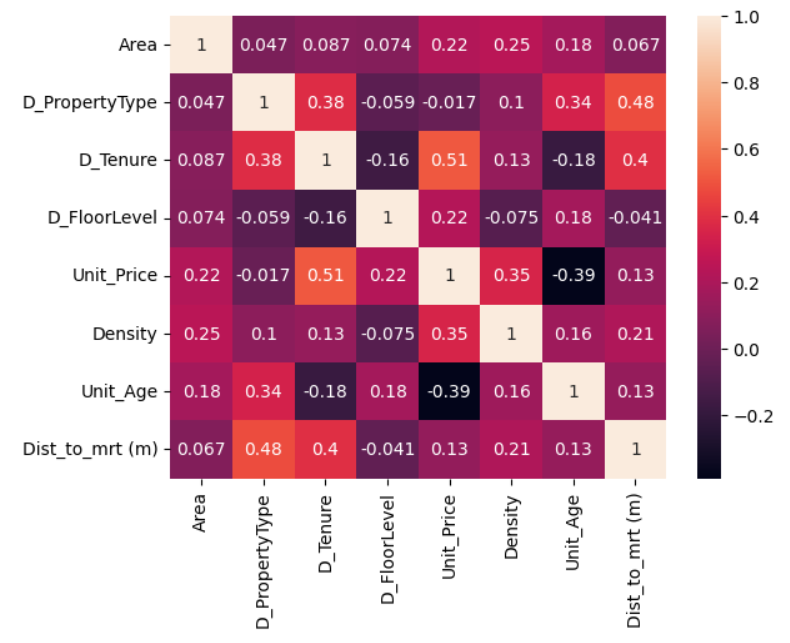


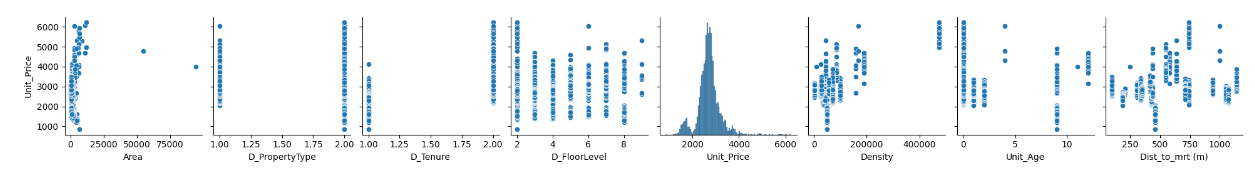
1. *Multi-Factors Linear Regression Model*

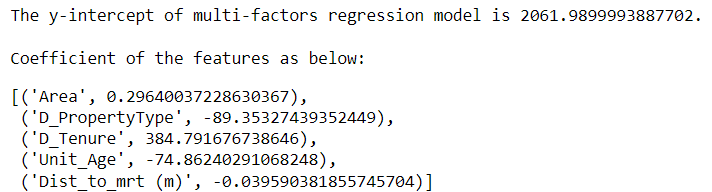
Correlation Matrix













Conclusion

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Description** | **Remarks** | **R-square** | **RMSE as percentage of y-mean** |
| 1. Benchmark model | Y = 2,691.04 | Taking mean of y as predicted y value | NA | 17.59% |
| 1. Simple Linear Regression Model | Y = (0.2367)X + 2449.2244 | Where  Y is “Unit\_Price” in psf,  X is “Area” in sqft | 0.1306 | 15.70% |
| 1. Multi-Factor Linear Regression Model | Y = (0.2964)X1 + (-89.3533)X2 + (384.7917)X3 + (-74.8624)X4 + (-0.0396)X5 + 2061.99 | Where  X1 = Area in sqft,  X2 = Property Type (1=apartment; 2=condo),  X3 = Tenure (1=99-leasehold; 2= Freehold),  X4 = Unit Age in year,  X5 = Distance to mrt in meters | 0.5475 | 11.53% |

*Deployment*

Streamlit app:

<https://ngchekwee-capstone-private-property-prop-app-ph450p.streamlit.app/>

