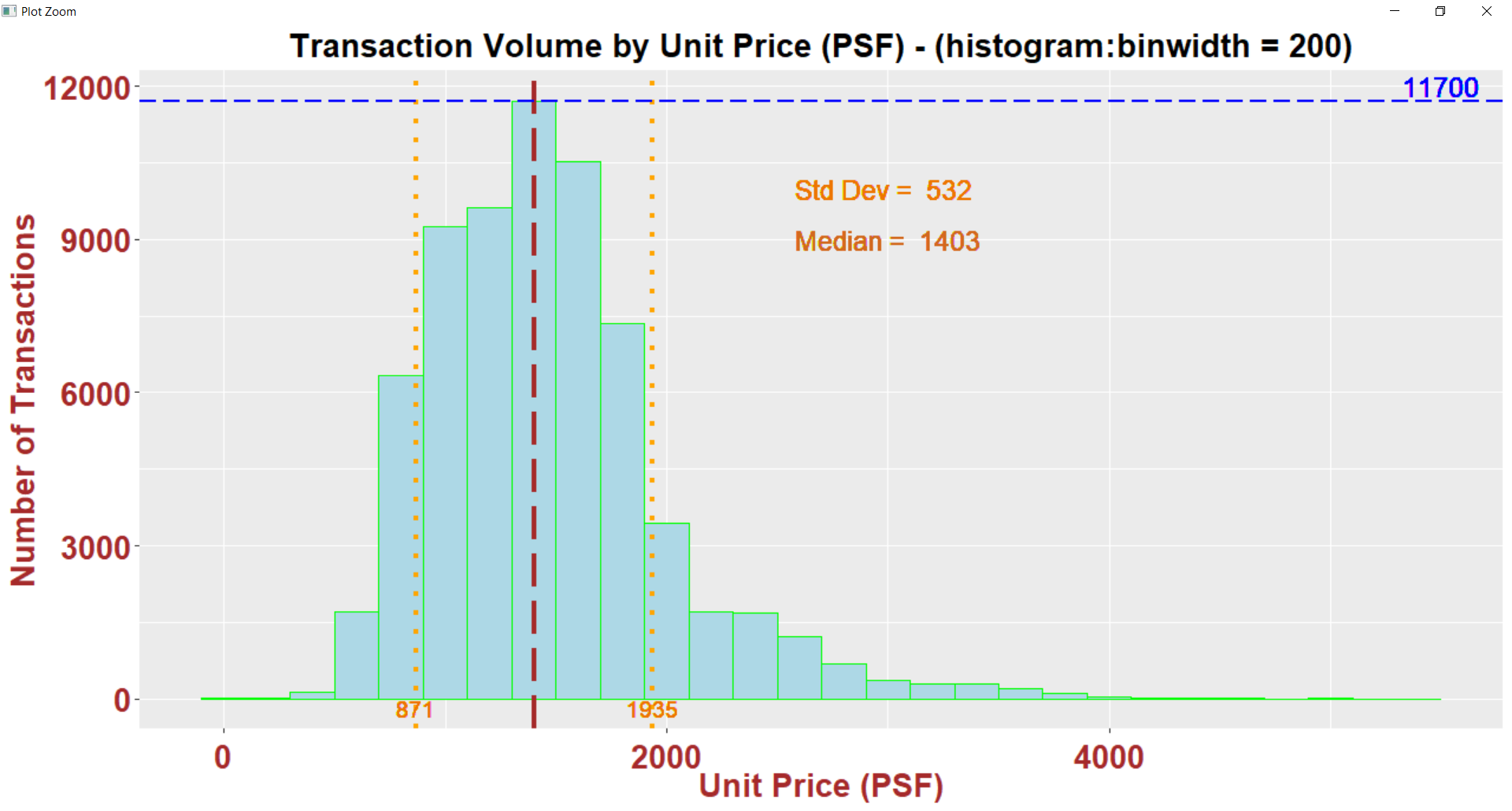
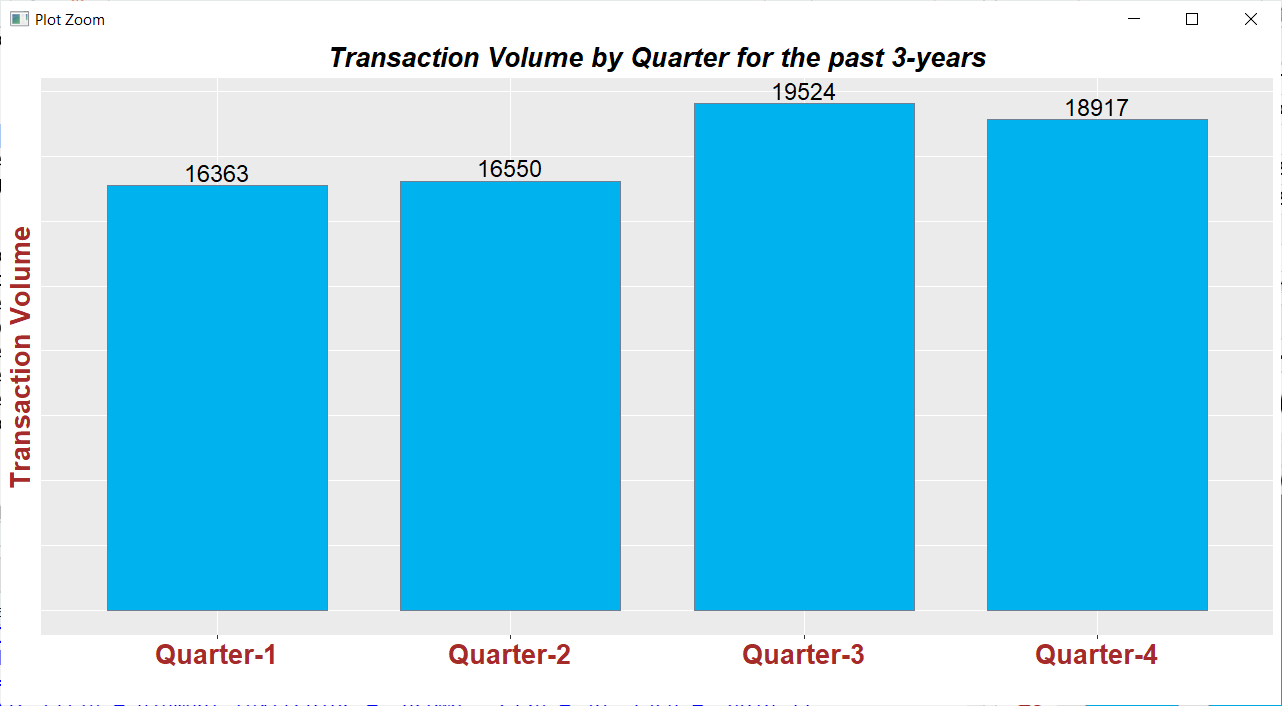
**Data Understanding using EDA (16 Marks)**

**Takeaway 1:** What is the mean/median of the Unit.Price.PSF for the overall property transaction ?



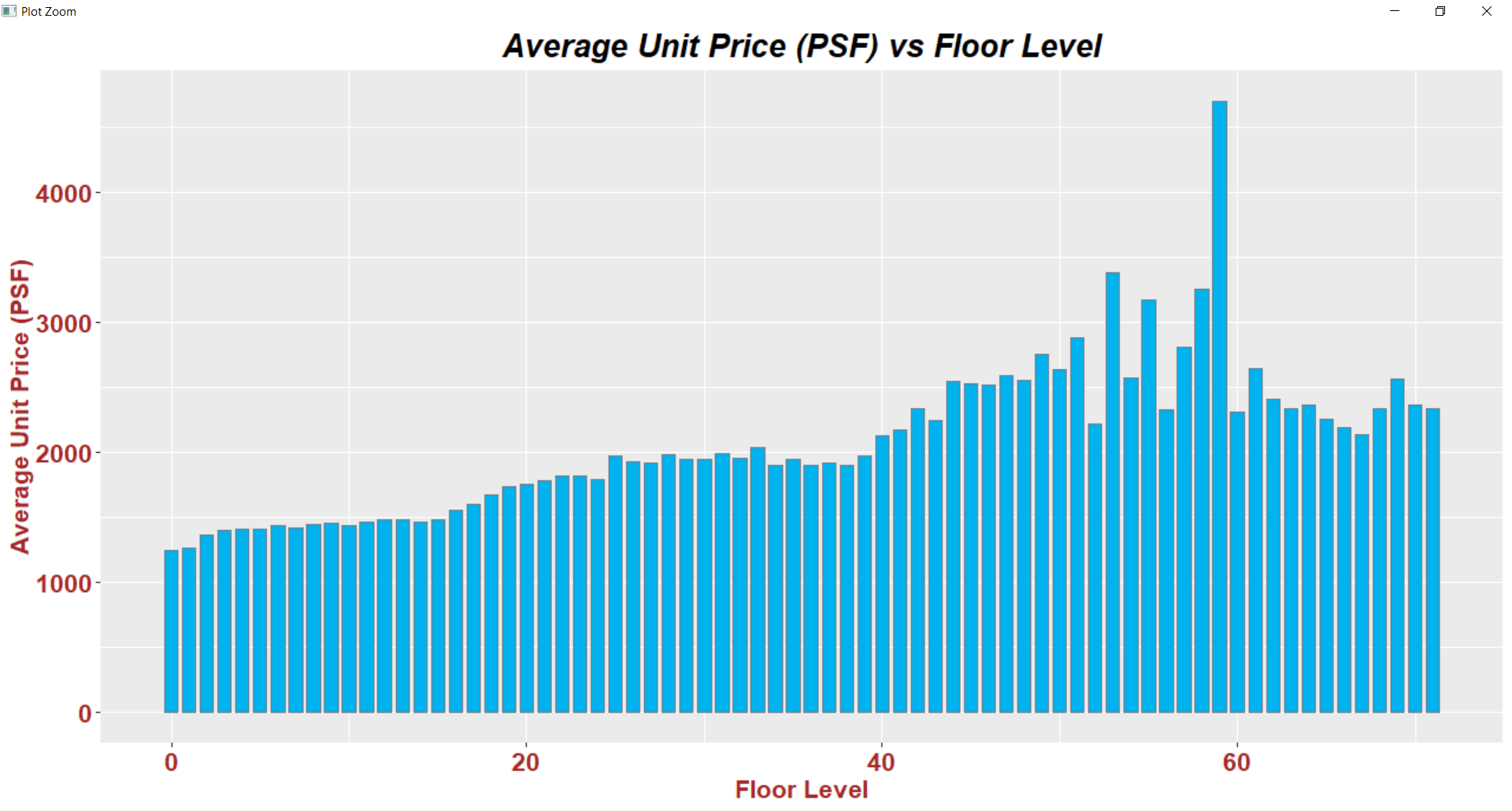
Above chart shows the histogram of the Unit.Price.PSF for the overall transaction. The median unit price (psf) transacted is at $1403. About 68% (1-σ deviation) of the properties transacted between the unit price of $871 to $1935 psf

**Takeaway 2:** Is there certain quarter prefer by property owner/owner-to-be to have transaction done?



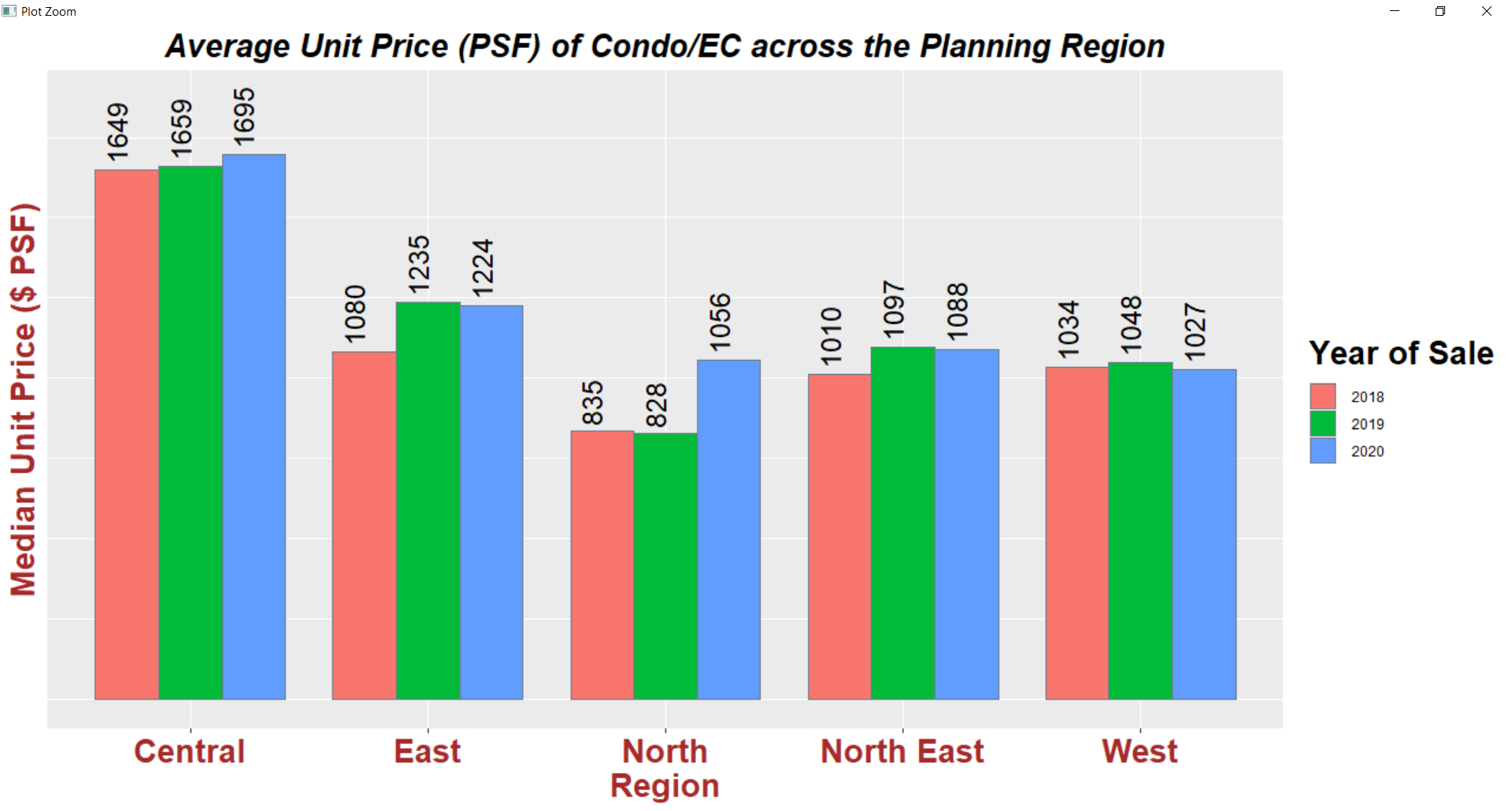
The bar chart above shows the sum of each quarter transactions done over the past 3-years (from Oct-2017 to Oct-2020). Third quarter seem to be the preferred season by the property owner/potential owner to have the transaction done. This is followed by the 4th quarter. 1st and 2nd quarters the transaction activities are about the same and they are the least active quarters.

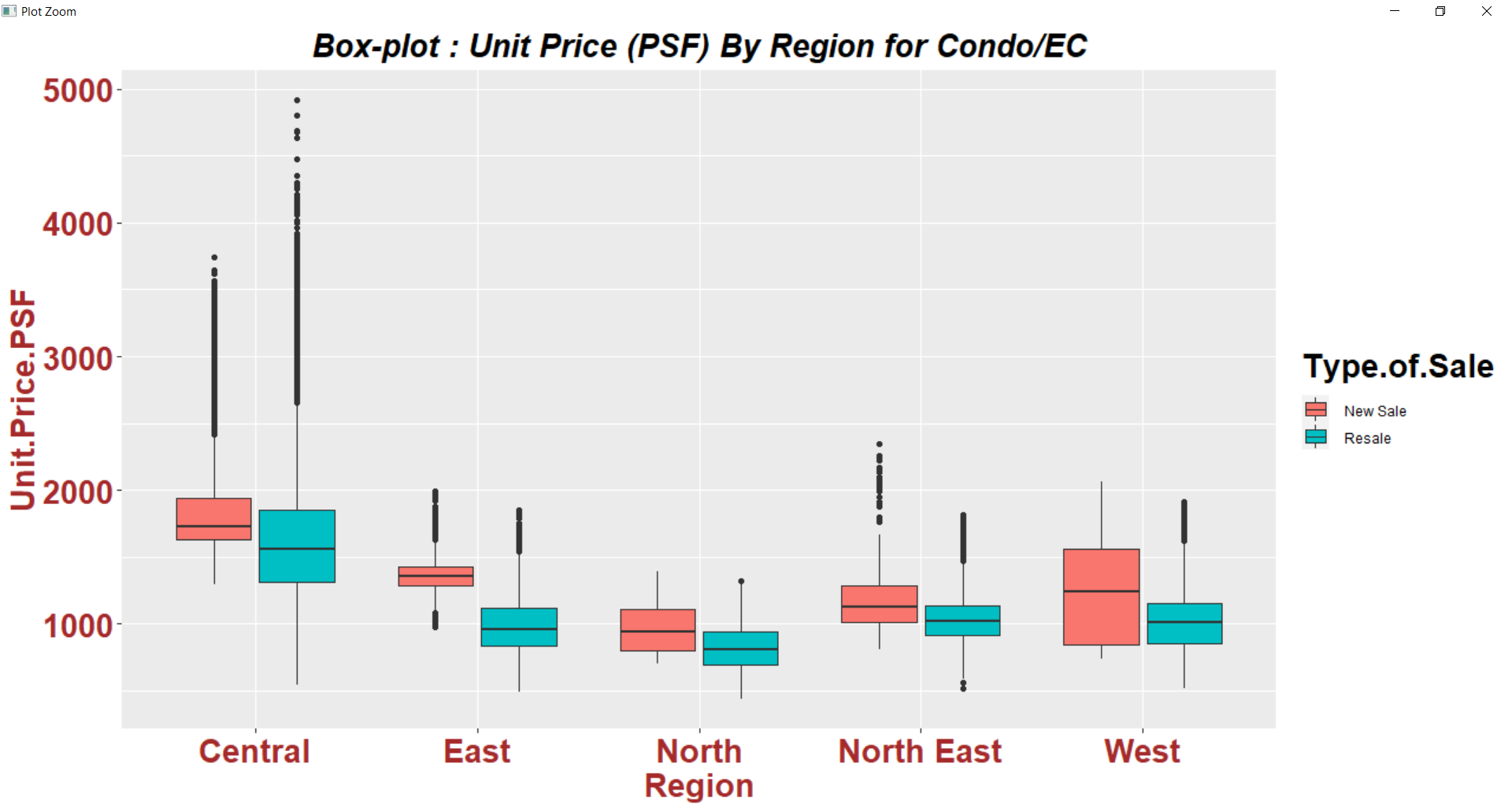
**Takeaway 3:** Is there any impact on Unit Price(PSF) for higher floor properties?



The plot above shows there is a positive correlation between the unit price (psf) and the property floor-level. The higher the property floor is located the higher the unit price it can command.

**Takeaway 4:** What is the impact of average unit price (psf) of Condo/EC for past 3-years across the region?





From the above bar chart and boxplot, it is observed that Condo/EC properties in the central region can command the highest unit price. From the boxplot it also shows central region has lot more upper outlier than other region, indicating the property in central region can command higher price. This is because central region is always perceived as the prime location. For the other region, the unit price (psf), is lower than the national median price.

For the past 3 years the unit price (psf) is relatively about the same across the regions, except for North region which goes up by about 27% in 2020.

**Data Preparation (8 Marks)**

The dataset given by the file “REALIS\_Oct17toOct20.csv”, cannot be use directly for analysis. The following actions are needed to get the dataset ready for analysis.

Below are the actions were performed for data preparation:-

* Remove the comma and change the following column to numeric.
  + **Transacted Price, Area.SQFT, Area.SQM, Unit.Price.PSF, Unit.Price.PSM, Number.of.Units**

*realEstate$Transacted.Price<-as.numeric(gsub(",", "", realEstate$Transacted.Price))*

*realEstate$Area.SQFT<-as.numeric(gsub(",", "", realEstate$Area.SQFT))*

*realEstate$Area.SQM<-as.numeric(gsub(",", "", realEstate$Area.SQM))*

*realEstate$Unit.Price.PSF<-as.numeric(gsub(",", "", realEstate$Unit.Price.PSF))*

*realEstate$Unit.Price.PSM<-as.numeric(gsub(",", "", realEstate$Unit.Price.PSM))*

*realEstate$Number.of.Units<-as.numeric(realEstate$Number.of.Units)*

* Factor following categorical column.
  + **Planning Region, Type.of.Sale, Type.of.Area, Property.Type, Postal.District, Postal.Sector**

*realEstate$Planning.Region <- as.factor(gsub(" Region", "", realEstate$Planning.Region))*

*realEstate$Type.of.Sale <- factor(realEstate$Type.of.Sale)*

*realEstate$Type.of.Area <- factor(realEstate$Type.of.Area)*

*realEstate$Property.Type <- factor(realEstate$Property.Type)*

*realEstate$Postal.District <- factor(realEstate$Postal.District)*

*realEstate$Postal.Sector <- factor(realEstate$Postal.Sector)*

* Change the **“Sale.Date**” column from character type to Date type.

*realEstate$Sale.Date <- as.Date(realEstate$Sale.Date, "%d-%b-%y")*

* Create a new column call “**Transacted.Price.Per.Unit**”, and this column is derived from “Transacted Price” and “Number.of.Units” columns.

The “Transacted.Price.Per.Unit column is needed, because in the “Transacted Price” column it has quite a few en-bloc (group purchase) transactions where the amount is very huge (close to a billion) but are share across few hundred units. This will skew the mean transacted price.

*realEstate$Transacted.Price.Per.Unit <-(realEstate$Transacted.Price / realEstate$Number.of.Units)*

* Likewise, for the property area, create a new column call **Area.SQFT.Per.Unit**, and is derived from “Area.SQFT” and “Number.of.Units” columns.

*realEstate$Area.SQFT.Per.Unit <-(realEstate$Area.SQFT / realEstate$Number.of.Units)*

* Create a new column call **“Storey”** (floor). The floor level of each property can be extracted from the “Address” column. For Landed property, the floor level is set to 0. The floor level of each property also contributes to the property price analysis.

#Extract Storey from Address column

*realEstate$Storey <- str\_extract(realEstate$Address,regex("[#][:digit:][:digit:]"))*

*realEstate$Storey[is.na(realEstate$Storey)] = "#0"*

*realEstate$Storey = as.numeric(gsub("#", "", realEstate$Storey))*

* Create a new column call **“Year.of.Sale”.** The year is extract from the “Sale.Date” column.

*realEstate$Year.of.Sale <- as.integer(format(realEstate$Sale.Date,"%Y"))*

* Create a new column call **“Quarters”**. This is derived and transformed from the “Sale.Date” column. Quarters column can be used to analyse to monitor which quarters or period has more transactions or activities.

*realEstate$Sale.Mth <- format(realEstate$Sale.Date,"%b")*

*realEstate$Sale.Mth = factor(realEstate$Sale.Mth, levels = c("Jan", "Feb", "Mar", "Apr",* "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))

# Change month to quarter

*realEstate$Sale.Quarter <- gsub("Jan", "Quarter-1", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Feb", "Quarter-1", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Mar", "Quarter-1 realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Apr", "Quarter-2", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("May", "Quarter-2", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Jun", "Quarter-2", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Jul", "Quarter-3", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Aug", "Quarter-3", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Sep", "Quarter-3", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Oct", "Quarter-4", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Nov", "Quarter-4", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- gsub("Dec", "Quarter-4", realEstate$Sale.Mth)*

*realEstate$Sale.Quarter <- factor(realEstate$Sale.Quarter)*

* Create a new column call **“Lease.Length”**. This is a categorical data that has five level (“Very Short, Short, Long, Very Long, Freehold). It is derived from the Tenure column. Can be used for regression modelling.

#Convert those 999999 and 9999 to 999

*realEstate$Tenure <- gsub("999999 yrs", "999 yrs", realEstate$Tenure)*

*realEstate$Tenure <- gsub("9999 yrs", "999 yrs", realEstate$Tenure)*

*tenure.start <- str\_extract(realEstate$Tenure,regex("[:digit:][:digit:][/][:digit:][:digit:][/][:digit:][:digit:][:digit:][:digit:]"))*

*tenure.start <- gsub("/","-", tenure.start)*

*tenure.start <- as.Date(tenure.start, "%d-%m-%Y")*

*tenure.start.year <- as.integer(format(tenure.start,"%Y"))*

*tenure.duration <- as.integer(str\_extract(realEstate$Tenure,regex("[:digit:][:digit:]+")))*

*tenure.end <-ymd(tenure.start)+years(tenure.duration)*

*tenure.end <- as.Date(tenure.end, "%d-%m-%Y")*

*tenure.end.year <- as.integer(format(tenure.end,"%Y"))*

*remaining.lease.year <- as.integer(tenure.end.year - (as.integer(format(today(),"%Y"))))*

*#Convert those NA result from Freehold status to 9999*

*remaining.lease.year[is.na(remaining.lease.year)] = 9999*

*#Create a column call Lease.Length*

*realEstate$Lease.Length <- remaining.lease.year*

*realEstate <- realEstate %>% mutate(Lease.Length= ifelse(Lease.Length < 30, "Very Short",*

*ifelse(Lease.Length >= 30 & Lease.Length < 60, "Short",*

*ifelse(Lease.Length >= 60 & Lease.Length < 110, "Long",*

*ifelse(Lease.Length >= 110 & Lease.Length < 1000, "Very Long", "Freehold")))))*

*realEstate$Lease.Length <- factor(realEstate$Lease.Length)*

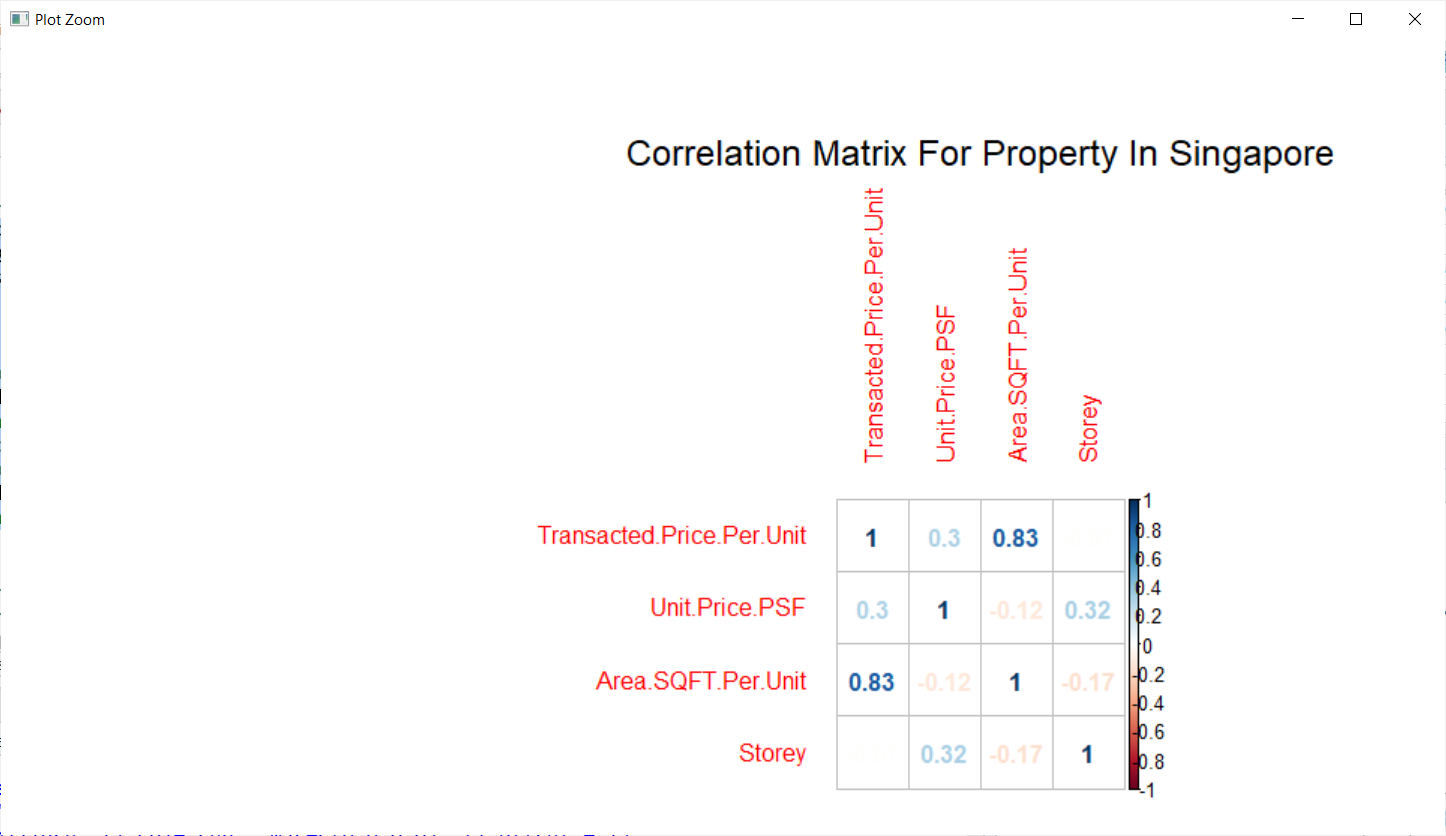
* In the **“Purchaser Address Indicator”** column, the data is either “HDB” or “Private” and there are quite several “N.A”. Substitute the “N.A” to “Others” as it could be first-time buyer or foreigner whose status are neither “HDB” or “Private” owner.

*realEstate$Purchaser.Address.Indicator <- gsub("N.A", "Others", realEstate$Purchaser.Address.Indicator)*

*realEstate$Purchaser.Address.Indicator <- factor(realEstate$Purchaser.Address.Indicator)*

* There is quite a fair bit of missing data in Nett.Price, “Completion Date” and “Project Name” columns. Decided not to do anything since they are not meaningful in the analysis.

**Correlation Analysis**

****

From the correlation matrix, it shows there is a **positive** correlation between Transacted.Price.Per.Unit and Area.SQFT.Per.Unit.

It is also observed that there is a weak positive correlation between property floor level and Unit.Price.PSF. That is the higher the property it can command better pricing.

There is a weak negative correlation between the property size and unit price psf. That is, the bigger the property area the smaller unit price psf will be.

There is a weak negative correlation between the property floor level and property area. That is, the lower the floor the property is located the bigger the size of the property. This is the case of all landed property which is usually bigger and is located at ground floor (0 level).

The dataset does not have many strong correlations other than the correlation between transacted.price.per.unit and the area.sqft.per.unit. Simple linear regression with just only one independent variable is not fitting enough.

Multiple regression is needed. This will involve the use of categorical data such as planning region, property type, type of sale, type of area, purchaser address indicator, lease length and property floor level (storey).

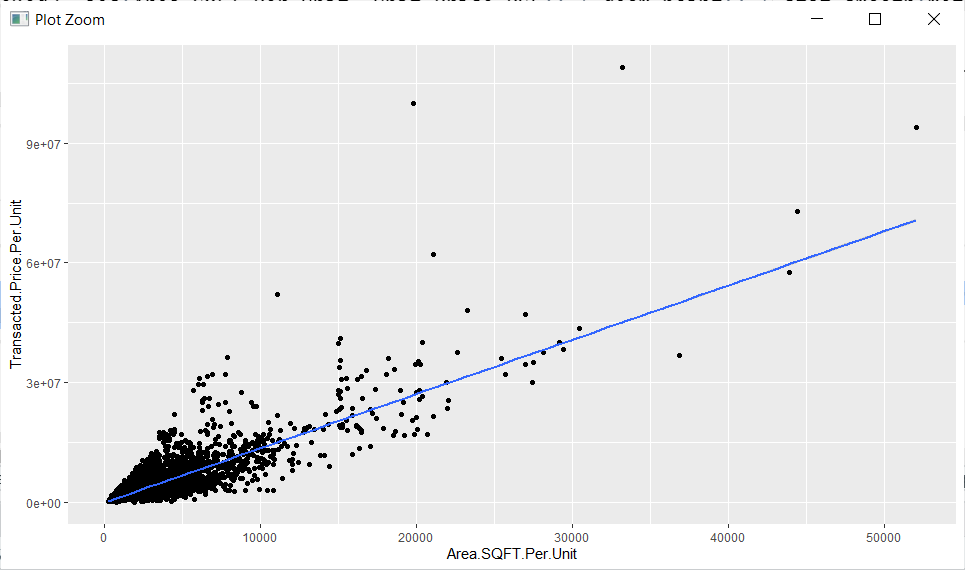
**Data Modelling (18 Marks)**

1. **Multiple Regression model:**

**model1 :**

Dependent variable :: **Transacted.Price.Per.Unit**

Independent variable ::Area.SQFT.Per.Unit + Property.Type + Lease.Length + Storey

****

From **model1** it is observed from the scatter plot, that the data are quite far apart for large size property. This model is good for small size property. A separate model can be generated to focus on just the larger property (see below next page for regression model above 8000sqft)

It was also found that floor level (Storey), property type and lease length do contribution to the prediction of the transacted price. In the earlier section it was presented that the floor level has an impact on the property price (see Takeaway 3). The regression line for this model is **72.98%**  fitting as shown in the adjusted R-squared value.

*Call:*

*lm(formula = Transacted.Price.Per.Unit ~ Area.SQFT.Per.Unit +*

*Property.Type + Lease.Length + Storey, data = realEstate)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-17010618 -339476 -9024 254628 71669115*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 1.394e+05 1.111e+04 12.54 <2e-16 \*\*\**

*Area.SQFT.Per.Unit 1.496e+03 5.008e+00 298.78 <2e-16 \*\*\**

*Property.TypeCondominium -1.552e+05 8.838e+03 -17.56 <2e-16 \*\*\**

*Property.TypeDetached House -1.438e+06 5.533e+04 -25.98 <2e-16 \*\*\**

*Property.TypeExecutive Condominium -6.099e+05 1.578e+04 -38.64 <2e-16 \*\*\**

*Property.TypeSemi-Detached House -1.212e+06 2.999e+04 -40.43 <2e-16 \*\*\**

*Property.TypeTerrace House -7.138e+05 1.998e+04 -35.73 <2e-16 \*\*\**

*Lease.LengthLong -4.386e+05 9.554e+03 -45.91 <2e-16 \*\*\**

*Lease.LengthShort -1.393e+06 4.403e+04 -31.63 <2e-16 \*\*\**

*Lease.LengthVery Long -2.987e+05 2.210e+04 -13.52 <2e-16 \*\*\**

*Lease.LengthVery Short -2.985e+06 1.822e+05 -16.38 <2e-16 \*\*\**

*Storey 2.889e+04 5.197e+02 55.59 <2e-16 \*\*\**

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 1011000 on 66775 degrees of freedom*

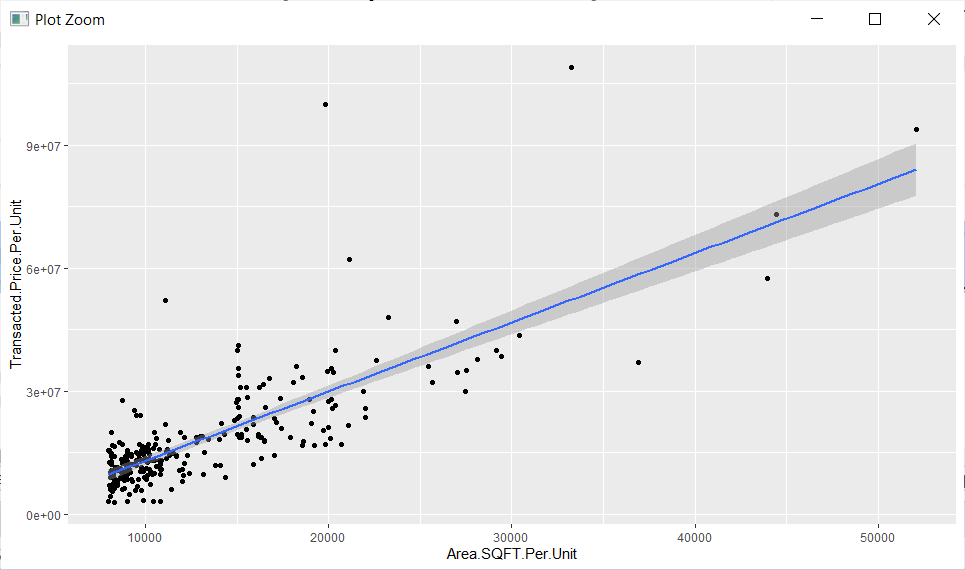
*Multiple R-squared: 0.7299,* ***Adjusted R-squared: 0.7298***

*F-statistic: 1.64e+04 on 11 and 66775 DF,* ***p-value: < 2.2e-16***

**model1\_Above8000sqft :**

Dependent variable :: **Transacted.Price.Per.Unit**

Independent variable ::Area.SQFT.Per.Unit + Property.Type + Lease.Length



**realEstate %>% filter(Area.SQFT.Per.Unit > 8000)**

This **model1\_Above8000sqft** focus only on property that is more than 8000sqft. It filters away those data where the property size is below 8000sqft. It only takes in property that are 8000sqft. With this model the adjusted R-squared value **(66%)** is not as higher as the earlier model, but prediction is better for large property size. The table below shows the comparison of the predicted transacted price against the value given in the dataset.

In this model which is for large properties; removing the “Storey” independent parameter also gives better result. Because most property with huge area are landed property and is at ground level (whereby Storey is set to 0).

***#filter to accept property that is above 8000sqft***

***realEstateReg2 <- realEstate %>% filter(Area.SQFT.Per.Unit > 8000)***

***model1\_Above8000sqft<-lm(Transacted.Price.Per.Unit~ Area.SQFT.Per.Unit + Property.Type + Lease.Length, data=realEstateReg2)***

***summary(model1\_Above8000sqft)***

*Call:*

*lm(formula = Transacted.Price.Per.Unit ~ Area.SQFT.Per.Unit +*

*Property.Type + Lease.Length, data = realEstateReg2)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-21527203 -3404833 -577747 2533619 70437268*

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) 2.364e+07 8.413e+06 2.811 0.005324 \*\**

*Area.SQFT.Per.Unit 1.689e+03 7.948e+01 21.247 < 2e-16 \*\*\**

*Property.TypeCondominium -1.381e+07 9.109e+06 -1.516 0.130809*

*Property.TypeDetached House -2.752e+07 8.218e+06 -3.349 0.000933 \*\*\**

*Property.TypeSemi-Detached House -3.238e+07 8.605e+06 -3.762 0.000208 \*\*\**

*Lease.LengthLong 2.711e+06 1.762e+06 1.538 0.125277*

*Lease.LengthShort -7.376e+06 4.052e+06 -1.820 0.069895 .*

*Lease.LengthVery Long -5.848e+06 2.074e+06 -2.820 0.005178 \*\**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 8001000 on 258 degrees of freedom*

*Multiple R-squared: 0.669,* ***Adjusted R-squared: 0.66***

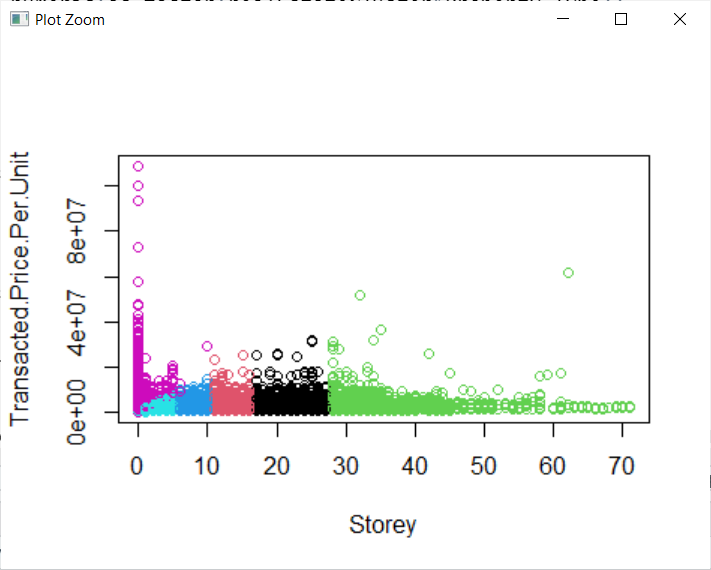
*F-statistic: 74.49 on 7 and 258 DF,* ***p-value: < 2.2e-16***

Below table shows the prediction for the transacted price and evaluate against the given dataset for 3 records (2 small properties size and 1 big property size).

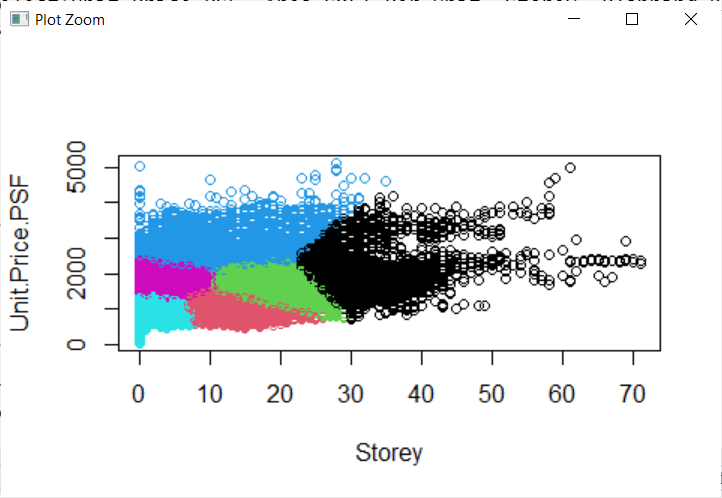
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Use** | **Area (SQFT)** | **Value from dataset (D)**  **Transacted Price** | **Predicted Transacted Price from model (M)** | **Error :**  **(D-M)/D** | **Other parameters**  **(Property Type, Lease Length, Storey)** |
| model1 | 968.76  (small size property) | $1.56M | $1.60M | **-2.56%** | See record 19367  “Condominium”, “Long”  Storey = 21 |
| 1033.34  (small size property) | $1.37M | $1.236M | **9.78%** | See record 33756  “Condominium”. “Long”  Storey = 5 |
| 52059.01  (big size property) | $93.9M | $76.60M | **18.42%** | See record 33756  “Detach House”. “Freehold”  Storey = 0 |
| model1\_Above8000sqft | 52059.01  (big size property) | $93.9M | $84.037M | **10.50%**  **(Improved)** | See record 33756  “Detach House”. “Freehold”  Storey = 0 |

**Table1:** Comparison between the predicted transacted price and the given value in the dataset

1. **Clustering scatter plot:**

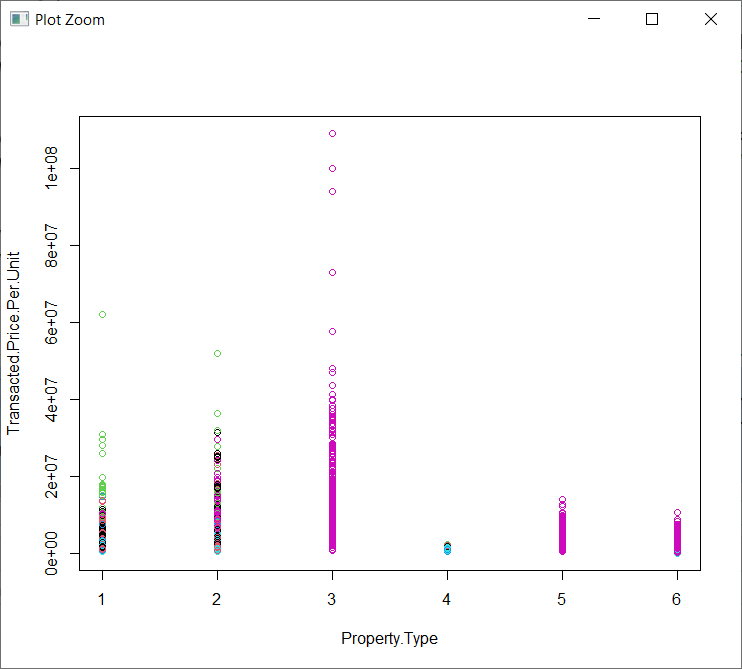


6 clusters



6 clusters

From the above 2 scatter plots it can infer that there is some distinct cluster for different floor level when plot against transacted price or unit price. Thus, they are classified neatly.



From the scatter plot above it can infer that there is some cluster between the property type and the transacted price except for property type 1 and 2. Property 1 and 2 type has a mixture of clusters. Thus, they are only classified moderately.

**Recommendation**

For prospective property buyer planning to get a property in the central region, be prepared to pay between $1650 to $1700 psf unit price (higher than national average of $1403).

For those with tight budget, north region will be a better consideration. It has a much lower asking price between $828 to $1056 (lower than national average by 32%). However, the unit price (psf) in North region is picking up pace and had increased by a 27% in 2020 compare to 2019 as shown in takeaway 4 chart.