***Machine learning-based analysis to predict property sales transactions in Dubai, UAE Master of Science in Professional Studies:***

***Data Analytics***

1. Data Preparation

**Data preparation** is the act of manipulating (or pre-processing) [raw data](https://en.wikipedia.org/wiki/Raw_data) (which may come from disparate data sources) into a form that can readily and accurately be analyzed, e.g. for business purposes.

Data preparation is the first step in data analytics projects and can include many discrete tasks such as loading data or data ingestion, [data fusion](https://en.wikipedia.org/wiki/Data_fusion), [data cleaning](https://en.wikipedia.org/wiki/Data_cleaning), [data augmentation](https://en.wikipedia.org/wiki/Data_augmentation), and data delivery.

The issues to be dealt with fall into two main categories:

* systematic errors involving large numbers of data records, probably because they have come from different sources;
* individual errors affecting small numbers of data records, probably due to errors in the original data entry.

2.1. We will need first to identify some necessary technologies and libraries that we are going to need throughout this research:

- Python: is a [high-level](https://en.wikipedia.org/wiki/High-level_programming_language), [general-purpose programming language](https://en.wikipedia.org/wiki/General-purpose_programming_language). Its design philosophy emphasizes [code readability](https://en.wikipedia.org/wiki/Code_readability) with the use of [significant indentation](https://en.wikipedia.org/wiki/Off-side_rule).[

Python is [dynamically-typed](https://en.wikipedia.org/wiki/Type_system#DYNAMIC) and [garbage-collected](https://en.wikipedia.org/wiki/Garbage_collection_(computer_science)). It supports multiple [programming paradigms](https://en.wikipedia.org/wiki/Programming_paradigm), including [structured](https://en.wikipedia.org/wiki/Structured_programming) (particularly [procedural](https://en.wikipedia.org/wiki/Procedural_programming)), [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) and [functional programming](https://en.wikipedia.org/wiki/Functional_programming). It is often described as a "batteries included" language due to its comprehensive [standard library](https://en.wikipedia.org/wiki/Standard_library).

[Guido van Rossum](https://en.wikipedia.org/wiki/Guido_van_Rossum) began working on Python in the late 1980s as a successor to the [ABC programming language](https://en.wikipedia.org/wiki/ABC_(programming_language)) and first released it in 1991 as Python 0.9.0.[[35]](https://en.wikipedia.org/wiki/Python_(programming_language)#cite_note-35) Python 2.0 was released in 2000 and introduced new features such as [list comprehensions](https://en.wikipedia.org/wiki/List_comprehension), [cycle-detecting](https://en.wikipedia.org/wiki/Cycle_detection) garbage collection, [reference counting](https://en.wikipedia.org/wiki/Reference_counting), and [Unicode](https://en.wikipedia.org/wiki/Unicode) support. Python 3.0, released in 2008, was a major revision that is not completely [backward-compatible](https://en.wikipedia.org/wiki/Backward_compatibility) with earlier versions. Python 2 was discontinued with version 2.7.18 in 2020.

- Project Jupyter: is a project with goals to develop [open-source software](https://en.wikipedia.org/wiki/Open-source_software), [open standards](https://en.wikipedia.org/wiki/Open_standard), and services for [interactive computing](https://en.wikipedia.org/wiki/Interactive_computing) across multiple [programming languages](https://en.wikipedia.org/wiki/Programming_language). It was spun off from [IPython](https://en.wikipedia.org/wiki/IPython" \o "IPython) in 2014 by [Fernando Pérez](https://en.wikipedia.org/wiki/Fernando_P%C3%A9rez_(software_developer)) and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are [Julia](https://en.wikipedia.org/wiki/Julia_(programming_language)), [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) and [R](https://en.wikipedia.org/wiki/R_(programming_language)). Its name and logo are an [homage](https://en.wikipedia.org/wiki/Homage_(arts)) to [Galileo](https://en.wikipedia.org/wiki/Galileo_Galilei)'s discovery of the [moons of Jupiter](https://en.wikipedia.org/wiki/Moons_of_Jupiter), as documented in notebooks attributed to Galileo. Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is financially sponsored by NumFOCUS.

- Numpy: a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Higuaín with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

- Pandas: a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals. Its name is a play on the phrase "Python data analysis" itself. Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007 to 2010.

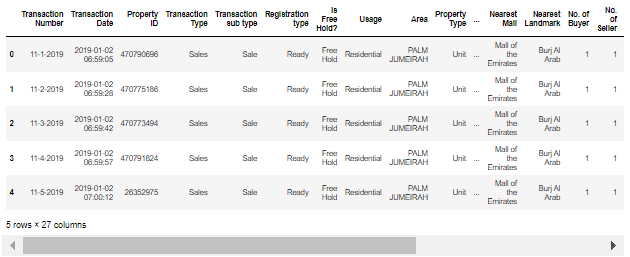
- matplotlib: a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

- Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

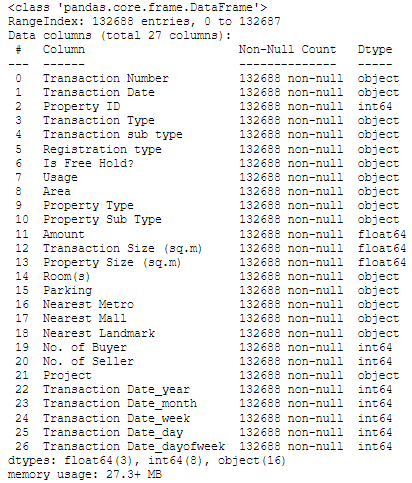
- Plotly: a technical computing company headquartered in Montreal, Quebec, that develops online data analytics and visualization tools. Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python, R, MATLAB, Perl, Julia, Arduino, and REST.

**2.2. What is our data looks like?**

2.2.1 Let's take a look on a sample of the data:

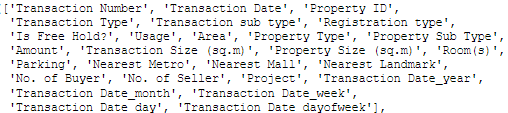


2.2.2 showing information about the dataset:



We have 132688 records about transactions and 26 columns.

2.2.3 dataset column names:

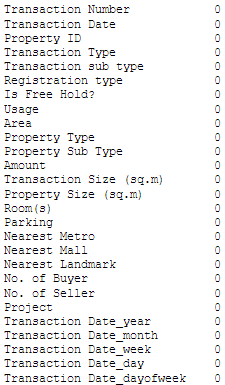


**2.3. Checking the validity of our data**

#### Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance,[[1]](https://en.wikipedia.org/wiki/Data_pre-processing#cite_note-1) and is an important step in the [data mining](https://en.wikipedia.org/wiki/Data_mining) process. The phrase ["garbage in, garbage out"](https://en.wikipedia.org/wiki/GIGO) is particularly applicable to [data mining](https://en.wikipedia.org/wiki/Data_mining) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning) projects. [Data-gathering](https://en.wikipedia.org/wiki/Data_collection) methods are often loosely controlled, resulting in [out-of-range](https://en.wikipedia.org/w/index.php?title=Range_error&action=edit&redlink=1) values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), and [missing values](https://en.wikipedia.org/wiki/Missing_values), etc.

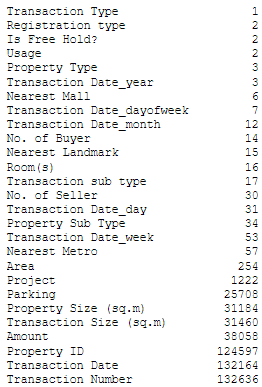
#### 2.3.1 Check if there're any null values in the data:

A null value in a relational database is used when the value in a column is unknown or missing. A null is neither an empty string (for character or datetime data types) nor a zero value (for numeric data types).



#### Fortunately, there're not.

#### 2.3.2 How much unique values in each column of the dataset?



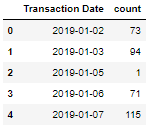
#### Since transaction Type column has only 1 unique value that is Sales, I don't think that this column would benefit any of our processes, so I'll drop it down from the dataset.

3. Analysis:

##### *Data analysis: is a process of inspecting,*[*cleansing*](https://en.wikipedia.org/wiki/Data_cleansing)*,*[*transforming*](https://en.wikipedia.org/wiki/Data_transformation)*, and*[*modeling*](https://en.wikipedia.org/wiki/Data_modeling)[*data*](https://en.wikipedia.org/wiki/Data)*with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.*

## 3.1 How did transactions go through time?

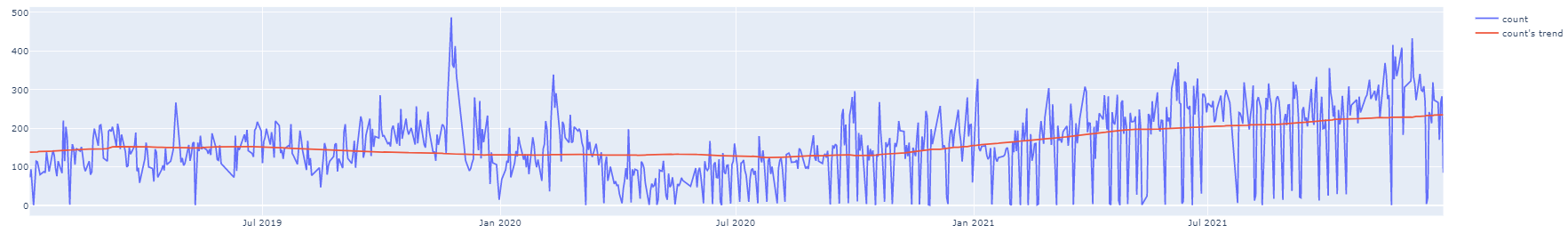
##### *First we need to create a dated dataset that has just the date and the number of transactions that occurred in that date:*[*¶*](http://localhost:8888/notebooks/project.ipynb#First-we-need-to-create-a-dated-dataset-that-has-just-the-date-and-the-number-of-transactions-that-occured-in-that-date:)



##### *Before jumping to the charts, let's first know What is a line chart:*

A line chart is a graphical representation of an asset's historical price action that connects a series of data points with a continuous line. This is the most basic type of chart used in finance, and it typically only depicts a security's closing prices over time. Line charts can be used for any timeframe, but they most often use day-to-day price changes.

#### Now we can use this dated dataset in showing how the transactions goes through years and how its trend acts:

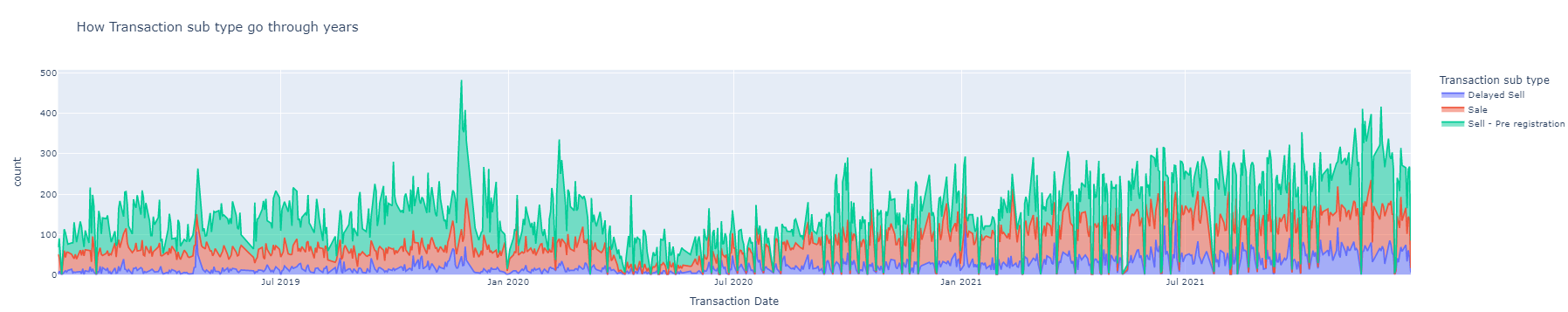


## **The number of daily transactions has been increasing since 2021.**

## 3.2 What is the most frequent Transaction sub type?

here we'll be showing a bar chart, but let's first know what a bar char is: A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

### Here we can find that Sell - Pre registration is the most frequent Transaction sub type, after it comes the Sale and delayed Sell, with that said, I wonder how have these three top sub types been going through the years? well, to figure out the answer of this question we need to create an area chart that looks like this:



To understand this chart you need to know first what is an area chart: An area chart is a graph that combines a line chart and a bar chart to show changes in quantities over time. It’s similar to a line graph in that data points are plotted and connected by line segments. However, the area below the line is colored in or shaded. Then, other values are plotted below the lines and shaded in a different color, resulting in a chart with layers.

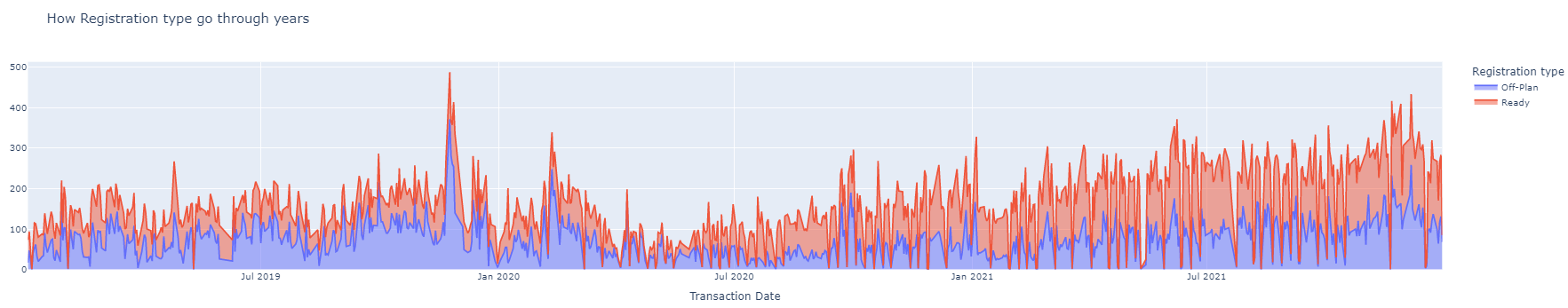
**You can see that Sale and Delayed Sell has been increasing in last year and half when Sell - Pre-registration has been the same.**

3.3. What is the most frequent Registration type?



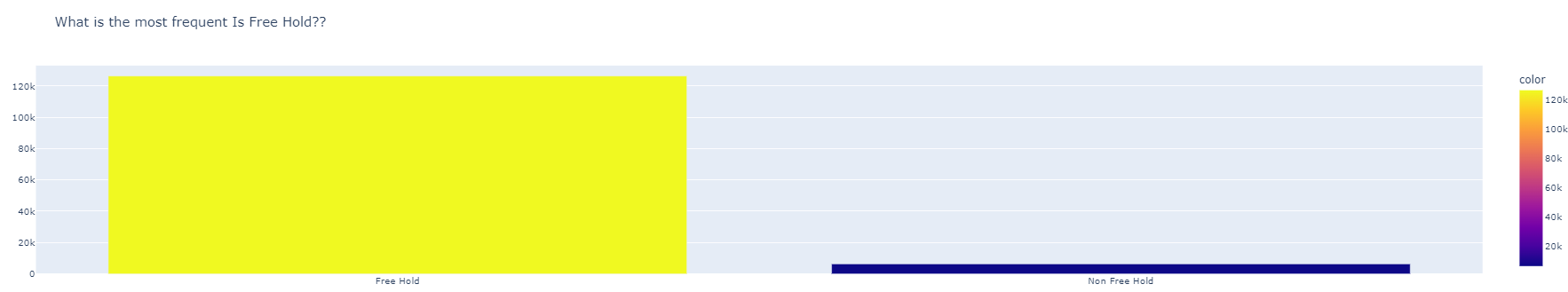
### Ready Registration type has appeared in the data for 54% of all the transactions.

### 3.4. How does Registration type go through the years?

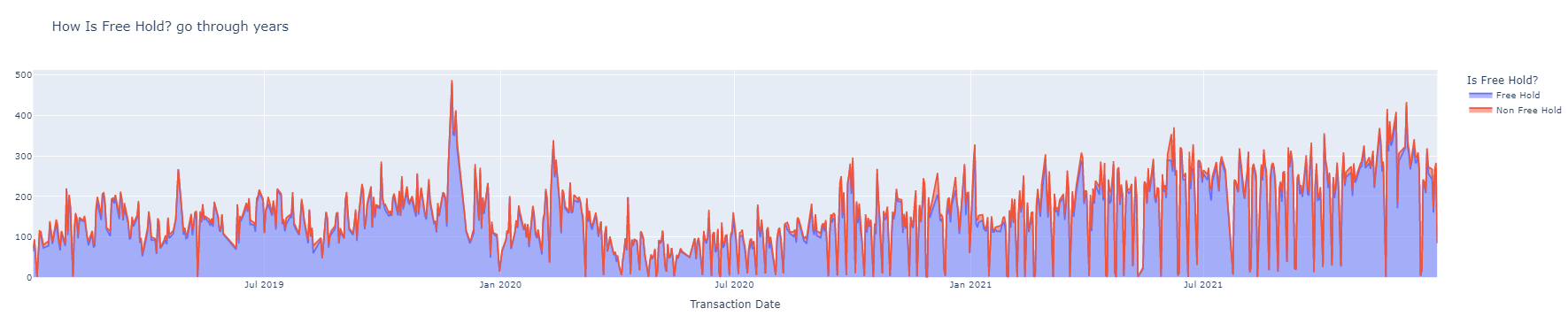


## **Ready Registration type has been also increasing in the last year and half.**

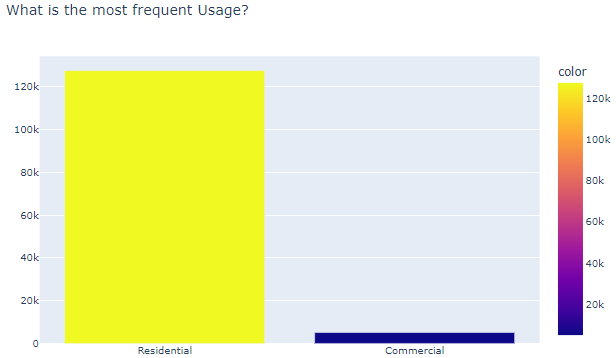
### 3.5. Were most of the transactions Free Hold or Non-Free Hold?



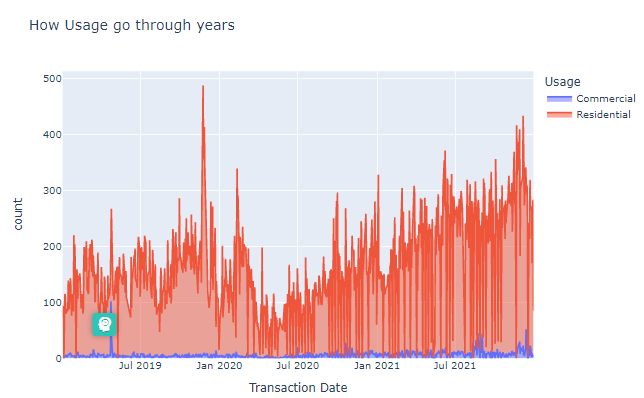
#### About 95% of the transaction were Free Hold transactions.



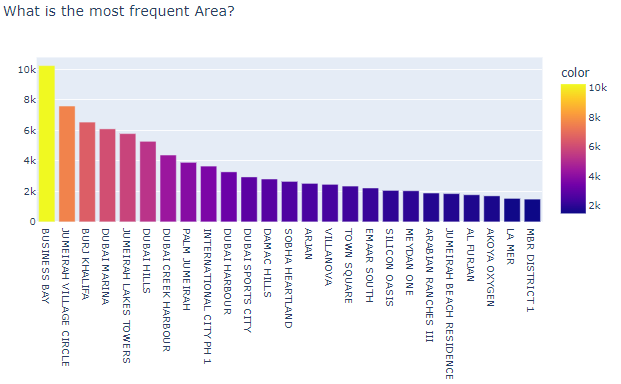
### 3.6. What is the most frequent Usage?



#### About 96% of Usages were Residential.

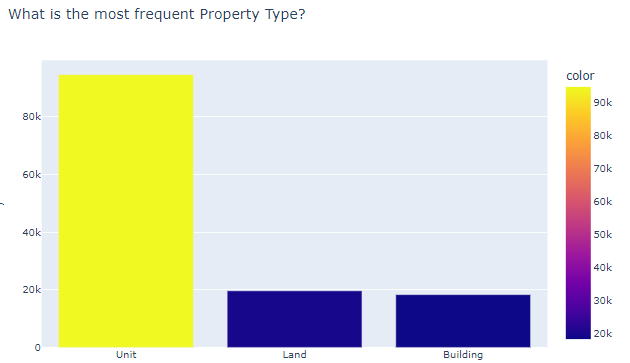


### 3.7. What is the most frequent Area?

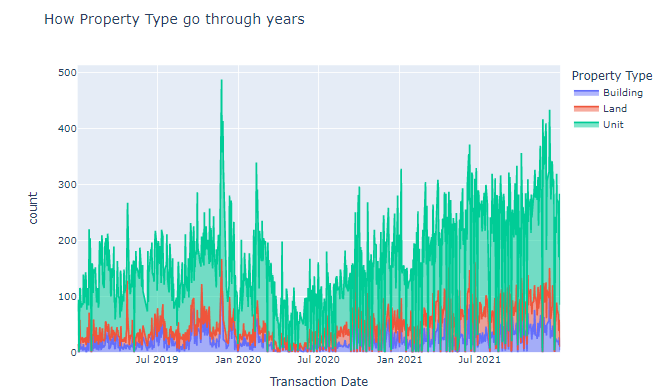


#### Business Bay was the most frequent area of all the transaction, 7.7% of the transactions were in that area.

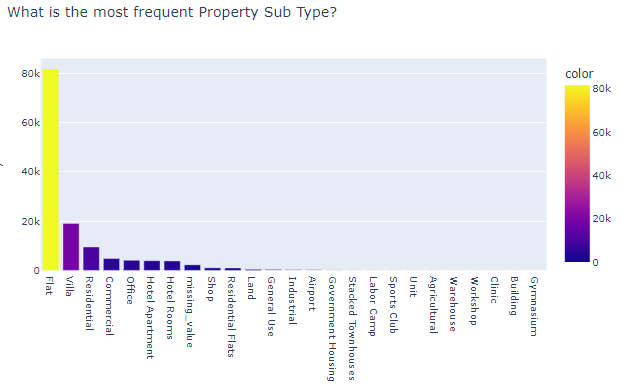
### 3.8. What is the most frequent Property Type?



#### Unit is the most frequent Property Type of all the transactions as 71% of the transactions were Unit Property Types.

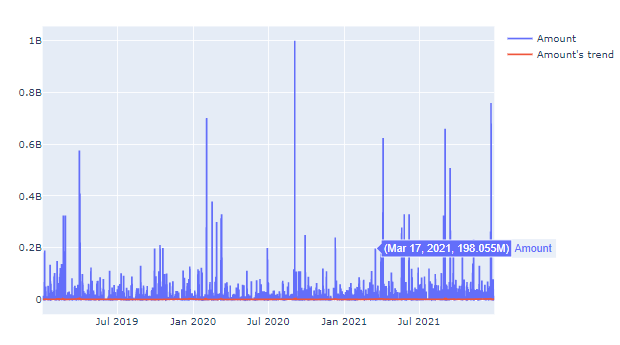


### 3.9. What is the most frequent Property Sub Type?



### Flat is the most frequent Property Sub Type as 64% of all the transactions were Flat Property Sub Type.

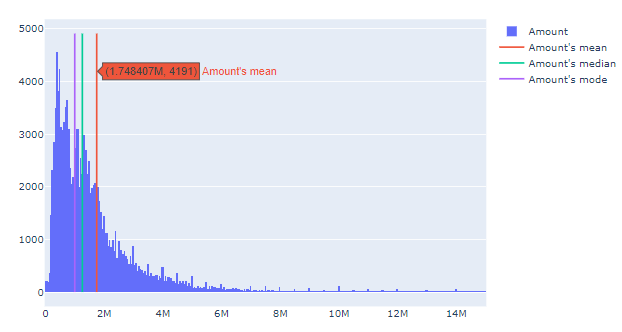
## 3.10. Amount through time:



What is distribution: A statistical distribution, or probability distribution, describes how values are distributed for a field. In other words, the statistical distribution shows which values are common and uncommon. There are many kinds of statistical distributions, including the bell-shaped normal distribution.

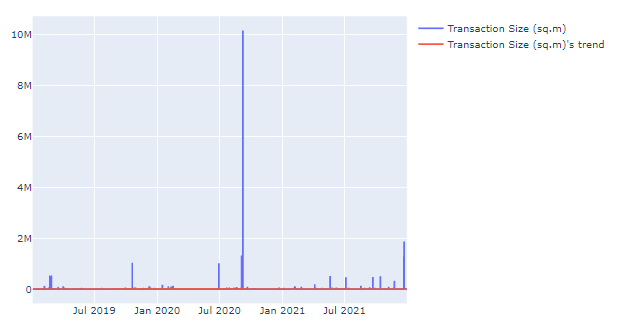
What Is a Histogram? A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

## The distribution of amounts:

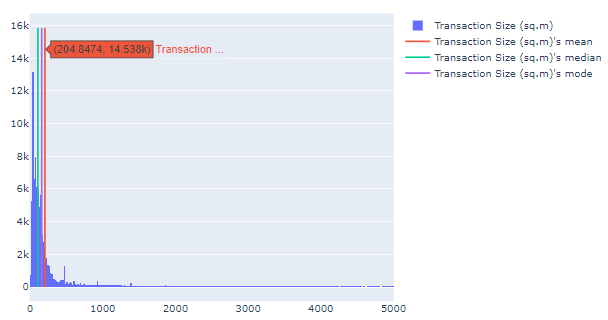


#### Although the average amount value is 1.75M, but the most frequent amount value was 1M and half of amounts values are below 1.26M.

## 3.11. Transaction Size through time:

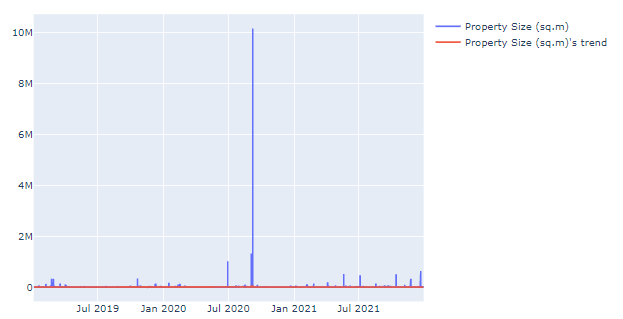


## The distribution of Transaction Size:

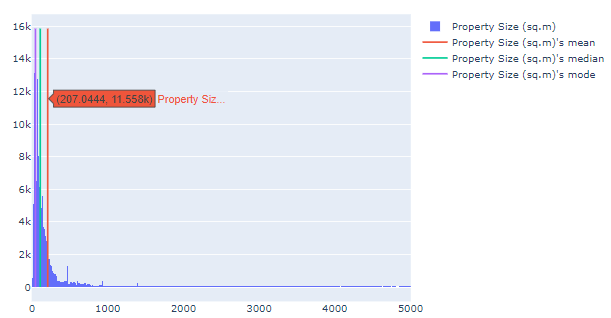


#### Although the average Transaction Size value is 205, but the most frequent Transaction Size value was 161 and half of Transaction Sizes values are below 108.

## 3.12. Property Size through time:

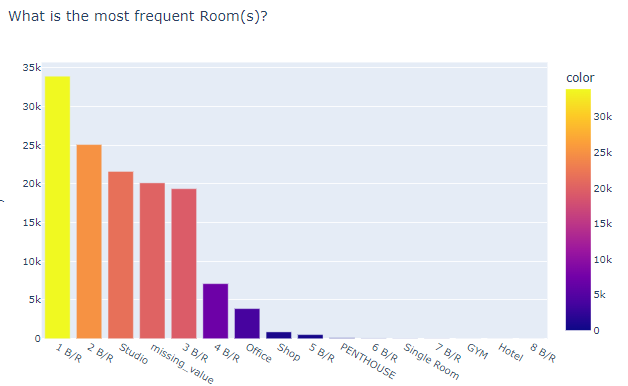


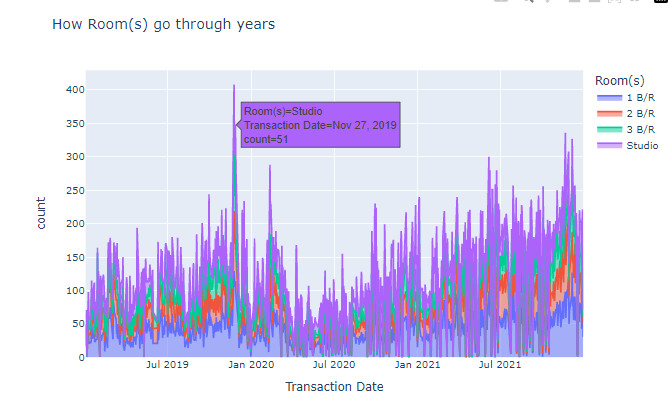
## The distribution of Property Size:



#### Although the average Transaction Size value is 207, but the most frequent Transaction Size value was 45 and half of Transaction Sizes values are below 108.

### 3.13. What is the most frequent Rooms Type?





# 4. Time series analysis and forecasting:

**4.1. INTRO:**

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

Time series forecasting occurs when you make scientific predictions based on historical time stamped data. It involves building models through historical analysis and using them to make observations and drive future strategic decision-making. An important distinction in forecasting is that at the time of the work, the future outcome is completely unavailable and can only be estimated through careful analysis and evidence-based priors.

##### *Libraries that we'll be using:*

* Statsmodels is a Python package that allows users to explore data, estimate statistical models, and perform statistical tests. An extensive list of descriptive statistics, statistical tests, plotting functions, and result statistics are available for different types of data and each estimator. It complements SciPy's stats module.
* Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS fiscally sponsored project.

## 4.2. Moving Average - Smoothing:

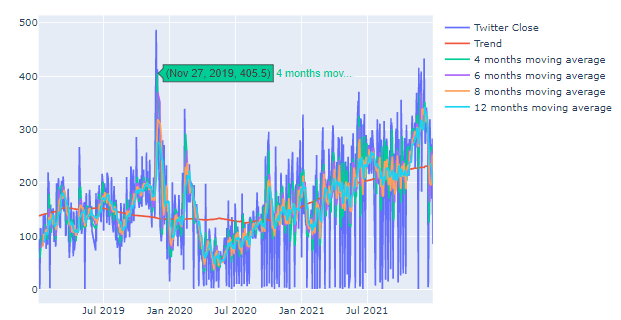
#### According to [Machine Learning Mastery](https://machinelearningmastery.com/moving-average-smoothing-for-time-series-forecasting-python/) -> Smoothing is a technique applied to time series to remove the fine-grained variation between time steps.

#### The hope of smoothing is to remove noise and better expose the signal of the underlying causal processes. Moving averages are a simple and common type of smoothing used in time series analysis and time series forecasting.

#### Calculating a moving average involves creating a new series where the values are comprised of the average of raw observations in the original time series.

#### A moving average requires that you specify a window size called the window width. This defines the number of raw observations used to calculate the moving average value.

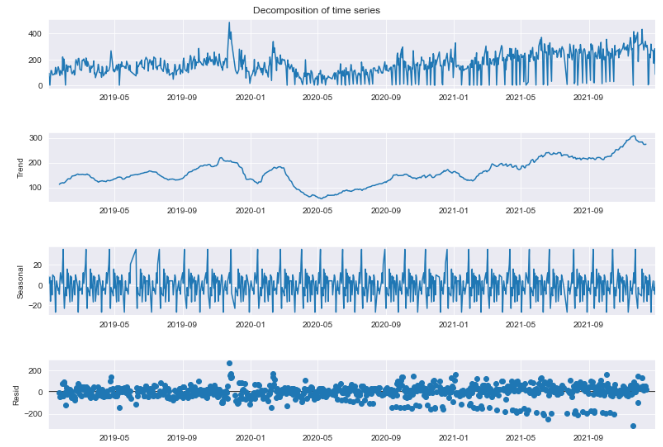
#### The “moving” part in the moving average refers to the fact that the window defined by the window width is slid along the time series to calculate the average values in the new series.[¶](http://localhost:8888/notebooks/project.ipynb#The-%E2%80%9Cmoving%E2%80%9D-part-in-the-moving-average-refers-to-the-fact-that-the-window-defined-by-the-window-width-is-slid-along-the-time-series-to-calculate-the-average-values-in-the-new-series.)



Note: Since this is a Plotly plot, you can hide and show moving averages by a double click if it's so noisy for you to get a clearer look at every moving average.

#### To get a deeper analysis of how the data goes let's use statsmodels.tsa.seasonal.seasonal\_decompose which gives us a seasonal decomposition using moving averages.

The seasonal\_decompose() function returns a result object. The result object contains arrays to access four pieces of data from the decomposition. For example, the snippet below shows how to decompose a series into trend, seasonal, and residual components assuming an additive model.

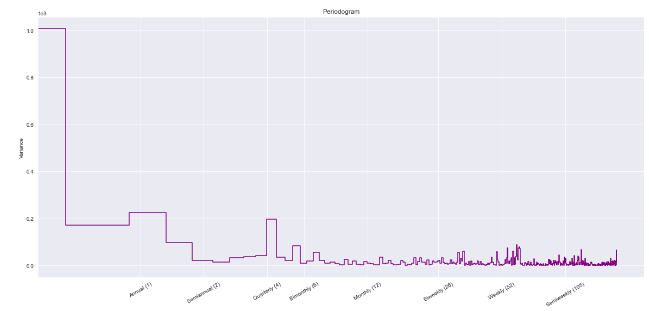


## 4.3. Seasonality:

#### according to [Kaggle](https://www.kaggle.com/code/ryanholbrook/seasonality) -> We say that a time series exhibits seasonality whenever there is a regular, periodic change in the mean of the series. Seasonal changes generally follow the clock and calendar -- repetitions over a day, a week, or a year are common. Seasonality is often driven by the cycles of the natural world over days and years or by conventions of social behavior surrounding dates and times.

#### Now let's use the following very helpful functions that was provided by Kaggle's experts to check whether our data is seasonal or not.[¶](http://localhost:8888/notebooks/project.ipynb#Now-let's-use-the-following-very-helpful-functions-that-was-provided-by-Kaggle's-experts-to-check-whether-our-data-is-seasonal-or-not.)

#### 



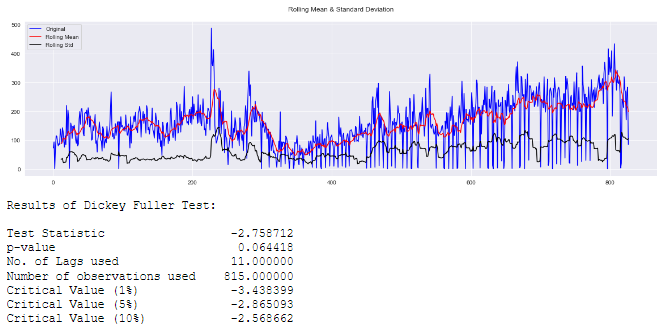
#### Looking at the previous plots (especially the last one) we can find that our time series here do has a double-annual seasonality.

## 4.4. Stationary:

#### According to [otexts.com](https://otexts.com/fpp2/stationarity.html) ->

#### A stationary time series is one whose properties do not depend on the time at which the series is observed.14 Thus, time series with trends, or with seasonality, are not stationary — the trend and seasonality will affect the value of the time series at different times. On the other hand, a white noise series is stationary — it does not matter when you observe it, it should look much the same at any point in time.

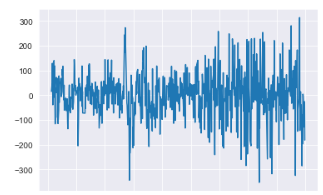
#### Now thanks to [ABHISHEKMAMIDI](https://www.kaggle.com/code/abhishekmamidi/time-series-preprocessing-to-modelling/notebook) who has made the following function that can check time series stationarity we can check whether our time series is stationary or not.

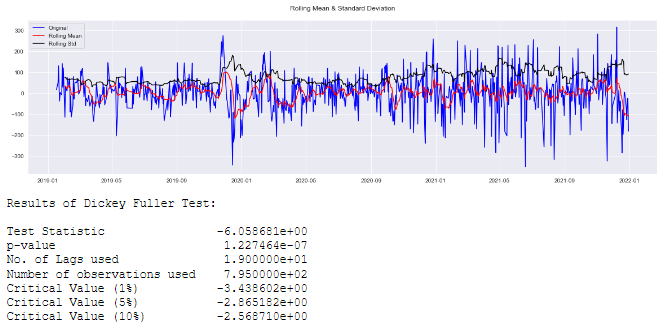


#### Since the [p-value](https://machinelearningmastery.com/time-series-data-stationary-python/) is more than 0.05, We can say that our time series here is not stationary, but what can we do about that?

#### Well one way to make a non-stationary time series stationary is by [Differencing](https://machinelearningmastery.com/remove-trends-seasonality-difference-transform-python/).

Differencing is a popular and widely used data transform for making time series data stationary.





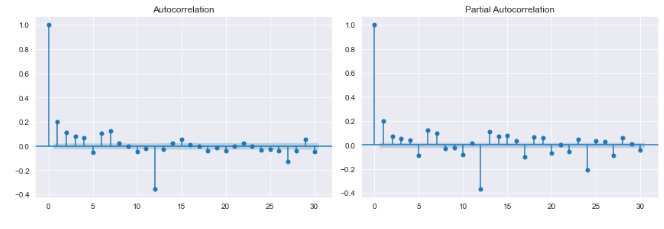
#### Again, after looking at the p-value here which is very close to 0 (less than 0.05), our time series has become stationary and ready to be forecasted.

## 4.5. Autocorrelation and partial autocorrelation:

#### according to [IBM](https://www.ibm.com/docs/en/spss-modeler/saas?topic=data-autocorrelation-partial-autocorrelation-functions) -> Autocorrelation and partial autocorrelation are measures of association between current and past series values and indicate which past series values are most useful in predicting future values. With this knowledge, you can determine the order of processes in an ARIMA model. More specifically,

#### \* Autocorrelation function (ACF). At lag k, this is the correlation between series values that are k intervals apart.

#### \* Partial autocorrelation function (PACF). At lag k, this is the correlation between series values that are k intervals apart, accounting for the values of the intervals between.

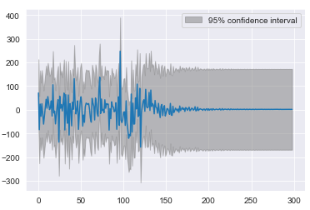


#### From that previous plots we can find that all lags under 30 can be helpful in predicting the future of this time series, let's use them in our model as the p value, the d value will be 0 as we've already differentiate the time series.

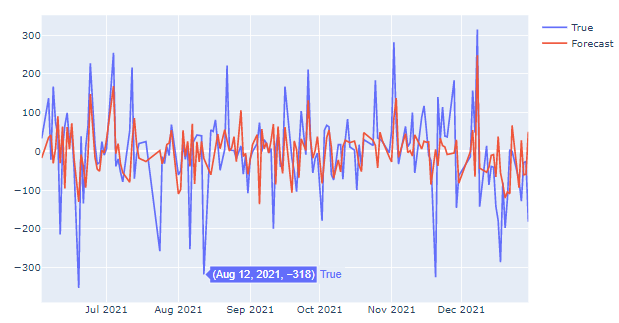
For more information you can click [here](https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average).

## 4.6. ARIMA model:

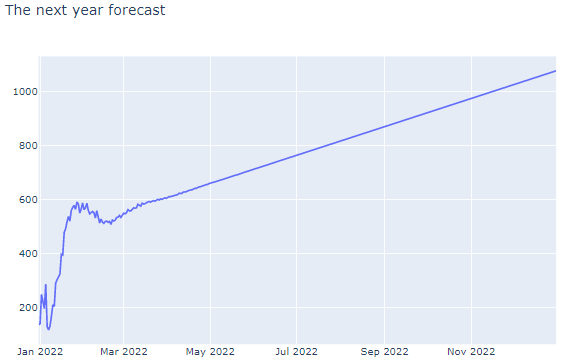
In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity in the sense of mean (but not variance/autocovariance), where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity of the mean function (i.e., the trend. When the seasonality shows in a time series, the seasonal-differencing could be applied to eliminate the seasonal component. Since the ARMA model, according to the World's decomposition theorem is theoretically sufficient to describe a regular (a.k.a. purely nondeterministic) wide-sense stationary time series, we are motivated to make stationary a non-stationary time series, e.g., by using differencing, before we can use the ARMA model. Note that if the time series contains a predictable sub-process (a.k.a. pure sine or complex-valued exponential process), the predictable component is treated as a non-zero-mean but periodic (i.e., seasonal) component in the ARIMA framework so that it is eliminated by the seasonal differencing.



The root mean squared error is 95.02202186670867.



#### Now let's create a function that give us a real forecast of property sales transactions in Dubai for the next year!



### From this graph of forecasting the property sales transactions in Dubai, we can find that the sales are going up, although it's not a good approach to count on more than 2 weeks' forecast.

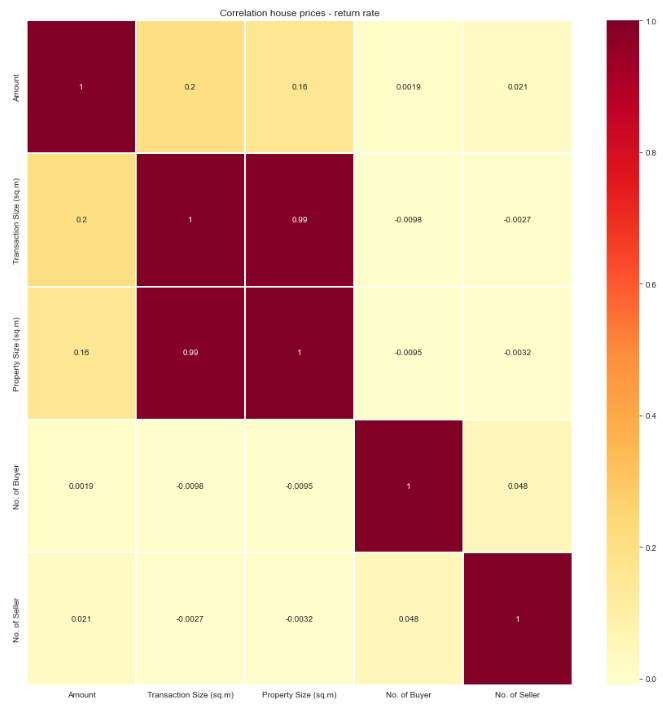
# 5. Clustering:

**5.1. INTRO:**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data analysis, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

#### 5.2. Getting the correlation between each numerical column:

Correlation is used to summarize the strength and direction of the linear association between two quantitative variables. It is denoted by r and values between -1 and +1. A positive value for r indicates a positive association, and a negative value for r indicates a negative association.



Since 'No. of Buyer', 'No. of Seller' are not correlated with the other column I'll remove them.

#### 5.3. Scaling the data:

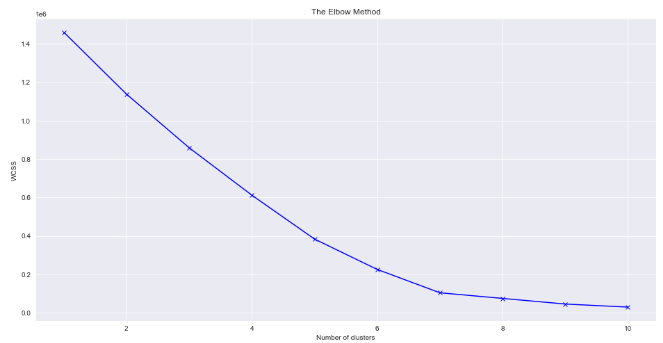
scaling: This means that you're transforming your data so that it fits within a specific scale, like 0-100 or 0-1. You want to scale data when you're using methods based on measures of how far apart data points are, like support vector machines (SVM) or k-nearest neighbors (KNN).

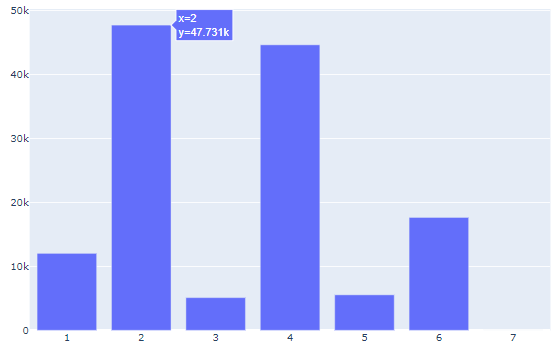
k-means clustering:

is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.

#### 5.4. elbow method:

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.





there's only one row that has clustered in a single cluster, for sure it's an outlier, I'll remove it from the data so that remains only 6 clusters (groups of the day).

#### 5.5. Identifying the criteria for decision-making by the UAE consumers and behavior when buying a property in Dubai:

for cluster 1:

* there were 12K transactions
* the average Amount was 5.25M
* the average Transaction Size was 1.5K
* the average Property Size was 1.15K
* all was a Free Hold
* all was Residential Usage
* all was Land Property Type
* all was Ready Registration type exept 1 transaction was Off-Plan

for cluster 2:

* there were 48K transactions
* the average Amount was 1.6M
* the average Transaction Size was 84
* the average Property Size was 83
* all was a Free Hold
* all was Residential Usage
* all was Unit Property Type
* all was Off-Plan Registration type

for cluster 3:

* there were 5K transactions
* the average Amount was 8.8M
* the average Transaction Size was 1935
* the average Property Size was 1986
* there were 4321 transactions that were a Free Hold, and other 783 Non Free Hold
* all was Commercial Usage
* all was Land Property Type
* all was Ready Registration type

for cluster 4:

* there were 45K transactions
* the average Amount was 1.4M
* the average Transaction Size was 120
* the average Property Size was 121
* all was a Free Hold
* all was Residential Usage
* all was Unit Property Type
* all was Ready Registration type

for cluster 5:

* there were 5.5K transactions
* the average Amount was 3.25M
* the average Transaction Size was 1.2K
* the average Property Size was 1.3K
* all was Non-Free Hold
* all was Residential Usage
* there were 2533 Land Property Type, 2319 Unit and 688 Building.
* there were 4944 Ready Registration type, and other 596 Off-Plan

for cluster 6:

* there were 17.6K transactions
* the average Amount was 1.9M
* the average Transaction Size was 219
* the average Property Size was 220
* all was a Free Hold
* all was Residential Usage
* all was Building Property Type
* There were 12491 Off-Plan Registration type and 5153 Ready

***Conclusion:***

In this research, we have discussed and calculated many important topics and results to:

1- use machine learning-based analysis to predict property sales transactions in Dubai for 2023 based on 2019, 2020, and 2021 data

2- identify the criteria for decision-making by the UAE consumers and behavior when buying a property in Dubai

Some of those discussions and accounts were about:

1- We have read and prepared the data in a programming environment using both Python and Jupiter and many other libraries and techniques to prepare and clean the data to be ready for all subsequent operations.

2- We analyzed the data exploratory analysis and drew more than 20 statistic graphs that describe different types of data that were carried by our data set. We answered questions such as:

- How did transactions go through time?

- What is the most frequent transaction sub type?

-What is the most frequent Property Type?

-What is the distribution of amounts

and more...

3- Then we go through time series analysis and forecasting using the ARIMA model process that ended up with the knowledge that the payments are likely to increase in the next year.

4- Finally, we divided the types of operations into six that identify the criteria for decision-making by the UAE consumers and behavior when buying a property in Dubai, by presenting the differences between the six groups inherent in:

- The number of processes in which it is possible to describe the percentage of chance of occurrence of a similar process to the processes belonging to that group

- Average quantity for each group

- Average Transaction Size for each group

- the average Property Size for each group

- Property Type for each group

- Registration type for each group

and more using one of the machine learning type which is the unsupervised machine learning that is a machine learning paradigm for problems where the available data consists of unlabeled examples, meaning that each data point contains features (covariates) only, without an associated label. The goal of unsupervised learning algorithms is learning useful patterns or structural properties of the data. Examples of unsupervised learning tasks are clustering, dimension reduction, and density estimation.