**Exploratory Data Analysis on Real Estate Data Set**



**Introduction**

The real estate data set contains 20 columns and 4734 rows of information. I imported varies libraries such as pandas, Numpy, Seaborn and Matplotlib to assist my analysis and loaded the excel file into Jupyter notebook.

**Objectives**

1. To glean any patterns or relationships from the data
2. To make recommendations to stakeholders for future areas of growth and potential savings.

**DATA CLEANING and MISSING DATA**

In summary I carried out the following:

* I identified missing values were present in nine columns. I removed ‘ApartmentUnitNumber’ and ‘OwnerFirstName’ as they were not relevant and had the largest missing values.
* For the other 7 remaining columns I used Scikit- learn mean strategy to fill in the missing values of ‘LandSF’, ‘exBuildingTypeID’, ‘LocationStartNumber’, ‘TotalFinishedArea’, ‘LivingUnits’. I left the 2 remaining columns ‘OwnerLastName’ with 1 result and ‘LegalReference’ with 7 missing values.
* To assist me with my analysis I extracted the month from the Sale Date column and created a new column called Sale Month in the dataframe
* Then I extracted the month from the Sale Date column and created a new column called Sale Quarter in the dataframe.

**DATA STORIES AND VISUALISATIONS**

In summary I carried out the following:

### **Vertical Bar Chart of Total Sales Per Quarter**

A picture containing text, screenshot, diagram, plot

Description automatically generated

Based on the vertical bar chart visualisation it appears that the 4th Quarter has the most total sales across the year. This is quite surprising given that they are generally the colder months, in October, November and December time. However I am not aware of the country/region of where this data is taken from so I cannot look into other factors at play such as local market conditions, economic factors or seasonal patterns. This could suggest the climate of the country this data was taken from is opposite to the UK and October – December is the spring and summer months.

### **Horizontal Bar Chart of Average Sales price per Area**

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

Based on the horizontal bar chart visualisation it appears that the average sale price per Total Finished area is just over 800. However I am not aware of the country/region of where this data is taken from so I cannot look into other factors at play such as local market conditions, economic factors, interest rates, demographics or seasonal patterns.

.

### **Bar Chart of which Month has the biggest sale**[**¶**](http://localhost:8888/notebooks/Dropbox/HI%20TECH%20PLUS/Week%2011/Real%20Estate%20Analysis.ipynb#Bar-Chart-of-which-Month-has-the-biggest-sale)

A picture containing text, screenshot, diagram, plot

Description automatically generated

Based on the bar chart visualisation it appears that the month with the highest total sales by a huge margin is December. This is quite surprising given that they are generally the colder months, in October, November and December time. However I am not aware of the country/region of where this data is taken from so I cannot look into other factors at play such as local market conditions, economic factors or seasonal patterns. This could suggest the climate of the country this data was taken from is opposite to the UK (maybe southern hemisphere) and October – December is the spring and summer months.

.

.

### **Stacked Bar Chart of which Month has the biggest sale by Deed ID**

A picture containing text, screenshot, plot, diagram

Description automatically generated

In order to create a stacked bar chart I needed another categorical variable to make it meaningful so I choose the Deed ID column. Based on the stacked bar chart visualisation it appears that the month with the highest total sales by a huge margin is December and the Deed ID 19 has the highest sale price. This is quite surprising given that they are generally the colder months, in October, November and December time. However I am not aware of the country/region of where this data is taken from so I cannot look into other factors at play such as local market conditions, economic factors or seasonal patterns. This could suggest the climate of the country this data was taken from is opposite to the UK (maybe southern hemisphere) and October – December is the spring and summer months.

.

**My findings were as follows:**

The quarter with the highest sales is quarter 4 in this dataframe.

The month with the highest sales is December in this dataframe.

The average sale price per Total Finished area is just over 800 in this dataframe.

**My conclusions are as follows:**

This country where the data is captured from is the Southern Hemisphere of the world due to the highest sales month as December.

I would like to explore the data further to try and evaluate properties. I could compare sales between different countries and regions based on sales price and the same area in square foot. Also if I knew the age, condition and number of bedrooms of the property I could analysis and present these findings.

Finally if I knew external factors such as the region I could compare schools nearby, transport links, train station nearby, and local crime rate.

**THIS REPORT WAS WRITTEN BY: LOUISE MARIE RANDALL**

