Project Documentation: Exploratory Data Analysis using Python:

Housing dataset

* Name: Neil Bestin
* DA/DS: DS
* Batch Number:
* Online/Offline: Online
* Roll Number: 7

Introduction: 3

Aim: 3

Business Problem/ Problem statement: 3

Project Workflow: 4

Data Understanding: 4

Data Cleaning : 5

Missing values Imputation: 5

Outliers: 5

Obtaining Derived Metrics: 5

Filtering Data for Analysis: 6

Statistical Analysis: 6

EDA: 7

Univariate Analysis: 7

Bivariate Analysis: 8

Multivariate Analysis: 9

Overall Insights from Analysis 9

Conclusion: 10

# Introduction:

This reinforcement project revolves around the use of python libraries such as Pandas, Numpy, Matplotlib, and more to analysis and provide insights into datasets in order to understand the contents of the dataset and any possible features which may influence other other features in said dataset. The housing dataset contains information regarding households in the Unites States, where data such as the price, area, and amenities available at said household are available. The main axis of analysis is to identify the most important features such that, through analysis, would provide valuable insights for stakeholders in the real estate industry.

# Aim:

The Primary Aim for this project is to find the most influential features in the Housing dataset that, through extensive analysis, would bring insights for all of the parties associated in the Real Estate industry. The goal is to find the most important features of the Housing dataset and use plots and other tools to analyse the relationships between each feature and derive insights valuable for the stakeholders in real estate.

# Business Problem/ Problem statement:

The problem statement for this project is to conduct an in-depth analysis of the dataset to derive valuable insights for stakeholders in the real estate industry. This problem is to address the trend seen in the different regions, as the price of a house in a region might differ from one in another region, hence valuable insights are required for the stakeholders in order to identify important amenities for them, as well as determine a price range for their new home. Essentially, the analysis involved in this project will help refine the data in the Housing dataset to help any interested in either buying or selling a house by providing simple insights.

# Project Workflow:

For this project, the Housing dataset is the main dataset used, where the first process is understanding the contents within the dataset, where I find the number of features in the entire dataset, the number of rows in the dataset, and the different categories of data present. Then, I have to clean the dataset in order to remove any missing or null values in order to maintain a continuous set of data, which will make the analysis smoother and provide less error. Then, I have to use the IQR, or Inter-Quartile Method in order to find outliers within the data, and remove them for finding accurate insights. Then, I filter the less important features, and perform 3 levels of data analysis using plots and statistics to gather insights:

Univariate, Bivariate, and Multivariate.

# Data Understanding:

The Housing dataset, initially, had 18 columns, which include: date, price, bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, waterfront, view, condition, sqft\_above, sqft\_basement, yr\_built, yr\_renovated, street, city, statezip, and country. The dataset had about 4600 entries per column, including null values and outliers. The dataset had 13 numerical columns, and 5 categorical columns. Initially, the data in the price column would shift based on the number of rooms available, whether it be bedrooms or bathrooms, where the mean of the price column was in power of 10, and the bedrooms and bathrooms mean was less than 4, as its seemed that the average house has 3 bathrooms and 3 bedrooms, leading the price to be about 500,000 dollars as the mean price of the 3 bedroom house.

This leads to an insight, where houses are priced based on the number of amenities required in the average family, which seems to be 3 in the United States, hence the price is higher.

# Data Cleaning :

## Missing values Imputation:

First, I used the function is\_null() to check for any empty data for each column, then I used a method called boolean indexing in order to remove the outliers In the sqft\_lot, sqft\_living, and city columns. In the yr\_renovated column, there were still null values, so I used the fillna() function to replace them with the value 0, as the empty value suggests it was not renovated

## Outliers:

For finding outliers, I choose to use a box and whisker plot to find and remove and visible outliers, using the quantile() method in order to find the exact value and use boolean indexing to limit the data unto a lower value, hence removing the outlier. This was done for every column

As for any inconsistent data, the yr\_built, and yr\_renovated columns are in numerical data types, however this was not changed, due to the fact it only contains the year a house was built or renovated, hence it does not need to be converted into the datetime data type, as int64 is a better data type for this context, as it allows me to make plots easier, as well as remove any outliers or simply change the data

# Obtaining Derived Metrics:

In order to find a suitable metric to compare the price columns with the sqft\_living column, I came up with the price\_per\_sqft column, which is found by dividing the price by the sqft\_living. I created a function called calculate\_price\_per\_sqft to the calculation and add it to the final list of features used for analysis

# Filtering Data for Analysis:

As the data had 18 columns, not every column can be used, as it would use up too much memory, hence I reduced it down to 8 columns by creating a array called housing\_df to hold the 8 important columns: price, bedrooms, bathrooms, sqft\_lot, sqft\_living, condition, and yr\_renovated. The other columns such as yr\_built and view were not required, as they are not the first priority for the average person. There were also some features such as waterfront, which had values closer to zero, which was not important to have, and country, as every house in the dataset is the same country.

# Statistical Analysis:

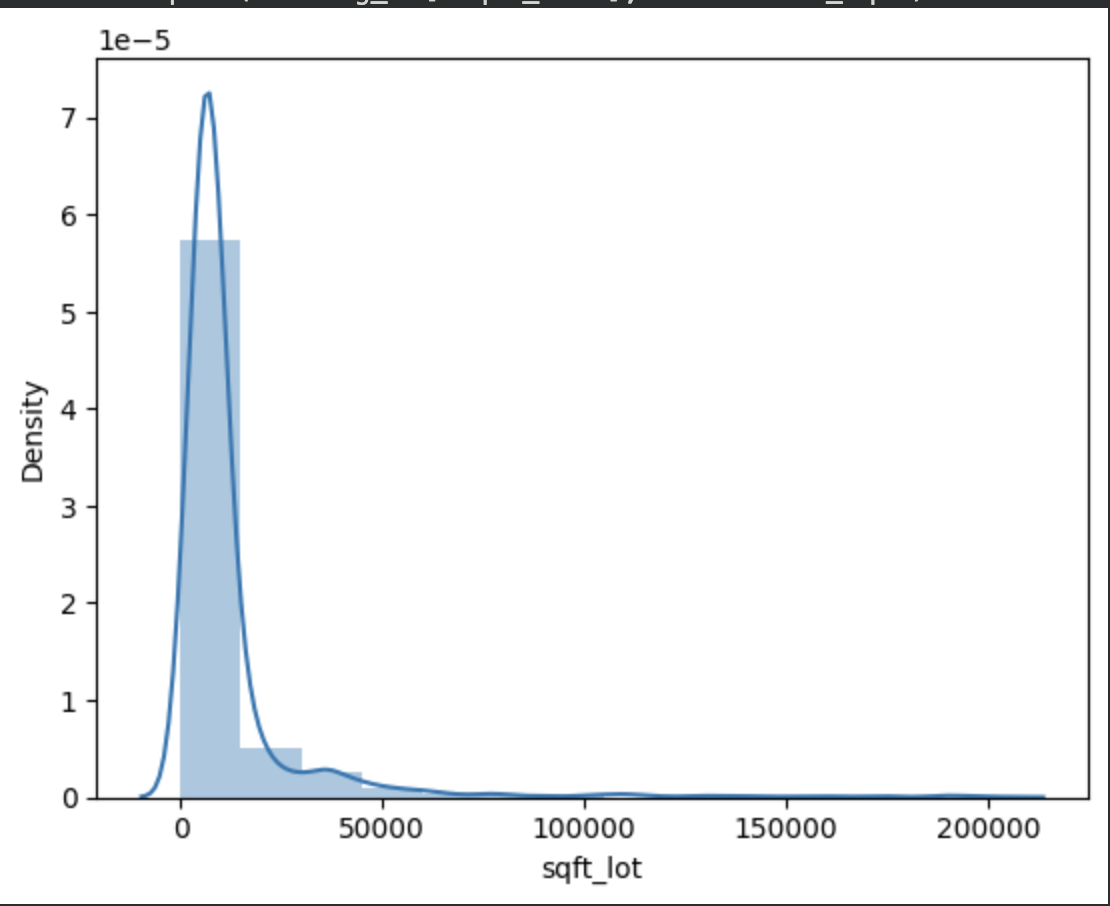
After filtering the dataset, the mean values for the housing\_df has decreased for all columns, as the number of values, including outliers has decreased, or removed due to null values in the other respective columns during data cleaning. The maximum values has also decreased due to the outliers being removed, also affecting the quantiles from 25% to 75%. However, the yr\_renovated feature is an exception, as for the 25% and 50% quantile, it was 0 due to null values, however, the 75% quantile and the maximum value has stayed the same 1999 and 2014 respectively.

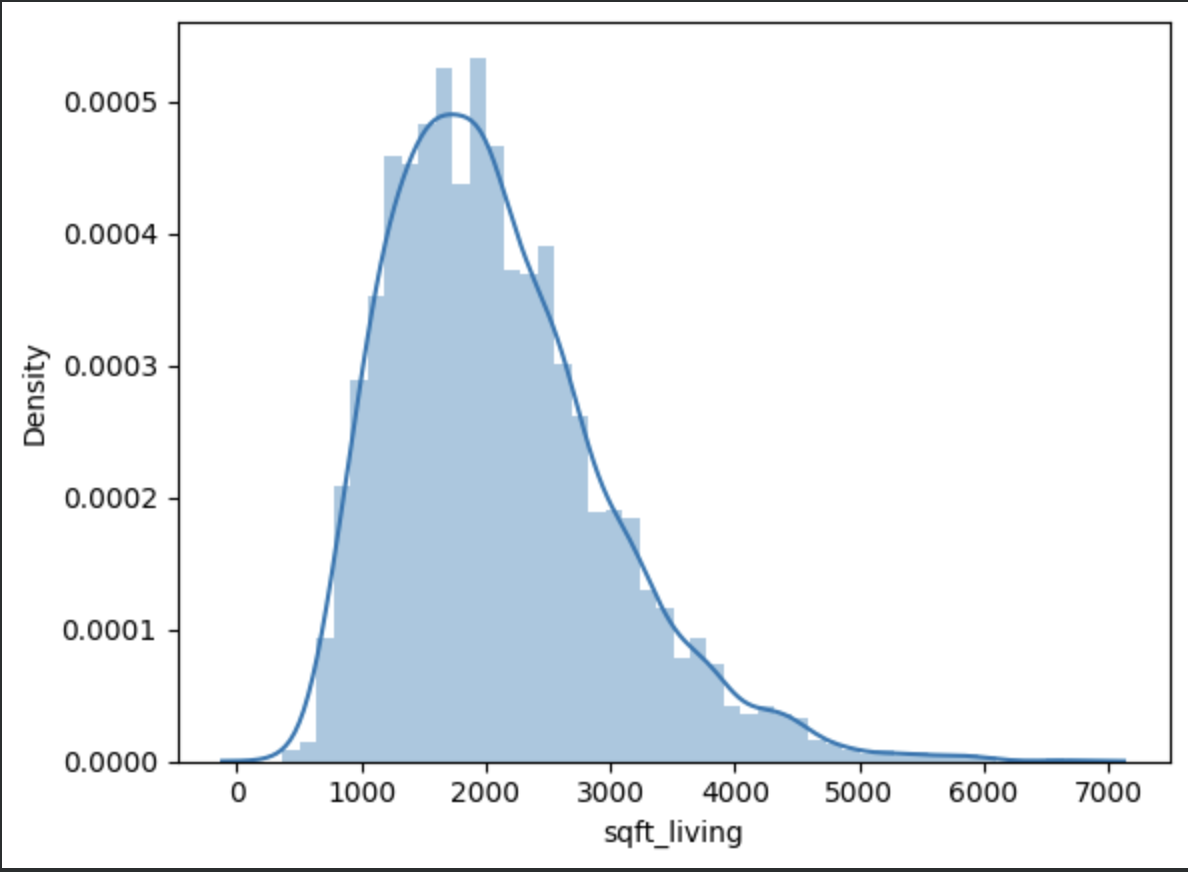
As for the statistical tests, as there are no categorical data within the filtered dataset, I decided to test for normality using the Shapiro-Wilk test, before and after removing outliers in order to understand whether removing outliers caused any effect on the features to achieve normal distribution. Before removing outliers, all 8 features had rejected the initial hypothesis that they follow a normal distribution. After removing outliers, the same is seen, with the test statistic for features like price and sqft\_lot increasing due to the removal of outliers, but however the p-value of both tests remain 0 for 9 features, including the derived price\_per\_sqft feature

# EDA:

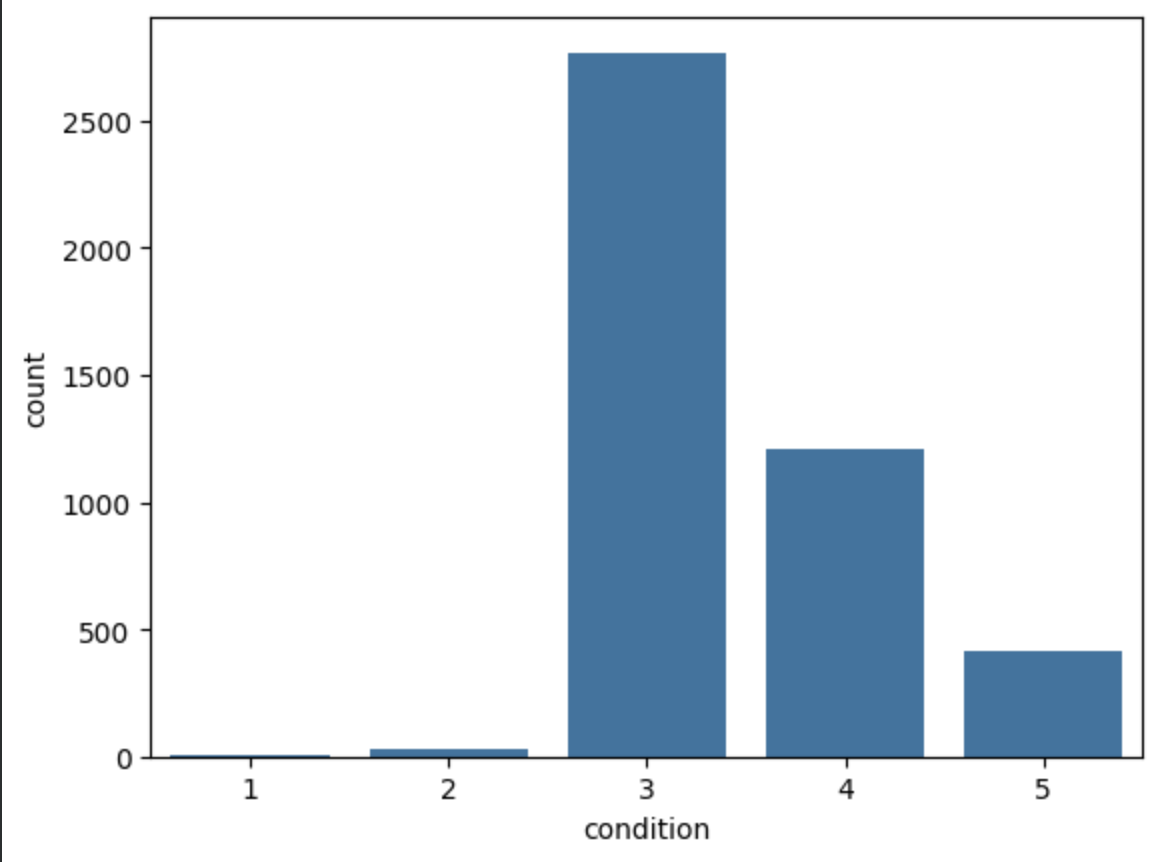
After reviewing the statistical tests, I started the EDA, or Exploratory Data Analysis on all features

## Univariate Analysis:

First, for the sqft\_lot feature, as seen in the plot, peaks as about 15000 square feet, showing that in the housing market, smaller plots of land are more preferred than higher plots, as we can see that the kernel density function curve dips.

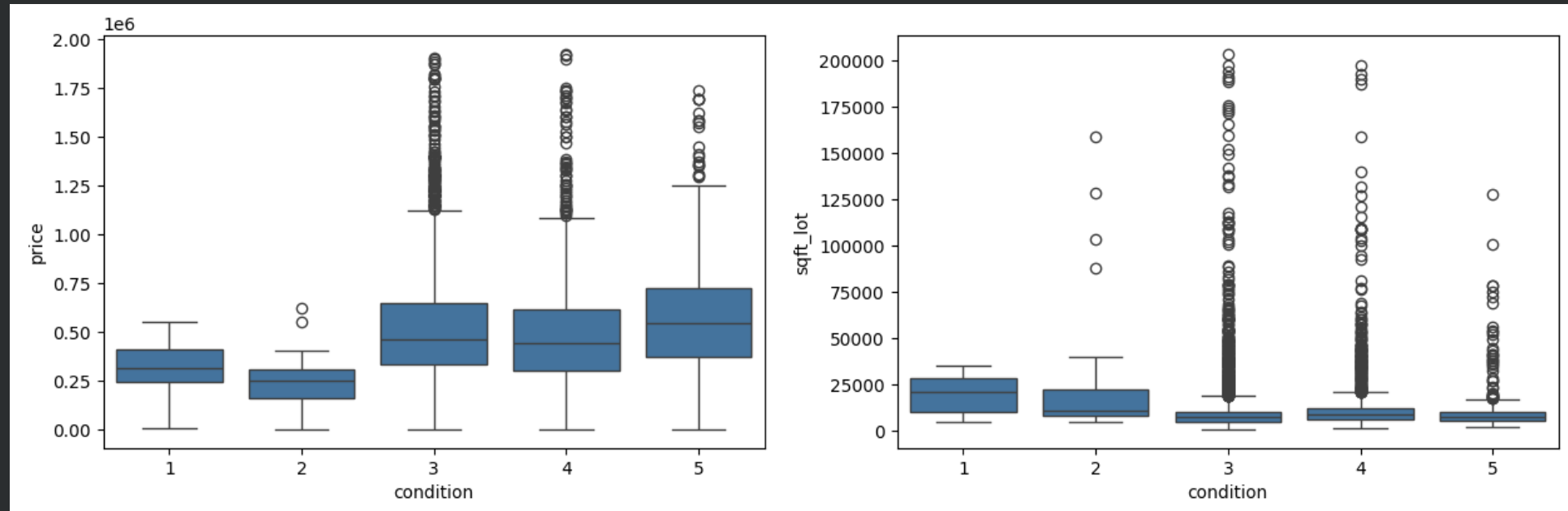


Next, the sqft\_living plot has a curve with larger area than the sqft\_lot, suggesting that majority of houses in the housing market have close to 2000 square feet of living space being used on the property itself, with the rest either being unused space or act as a back or front yard. This shows that stakeholders prefer housing with more space for the choice to extend their property years after purchase.

Next, the displots of the bathrooms and bedrooms features shows that most homes tend to have about 3 bedrooms, with some having either 2 bathrooms and either a hand-washing area or another bathroom, as the bathrooms plot peaks at 2.5, showing that smaller families tend to be the target audience in the housing market, where the price displot, according to the above metrics, tend to cost around 300,000 USD, with the average price\_per\_sqft being close to 250 USD, as seen in the python file

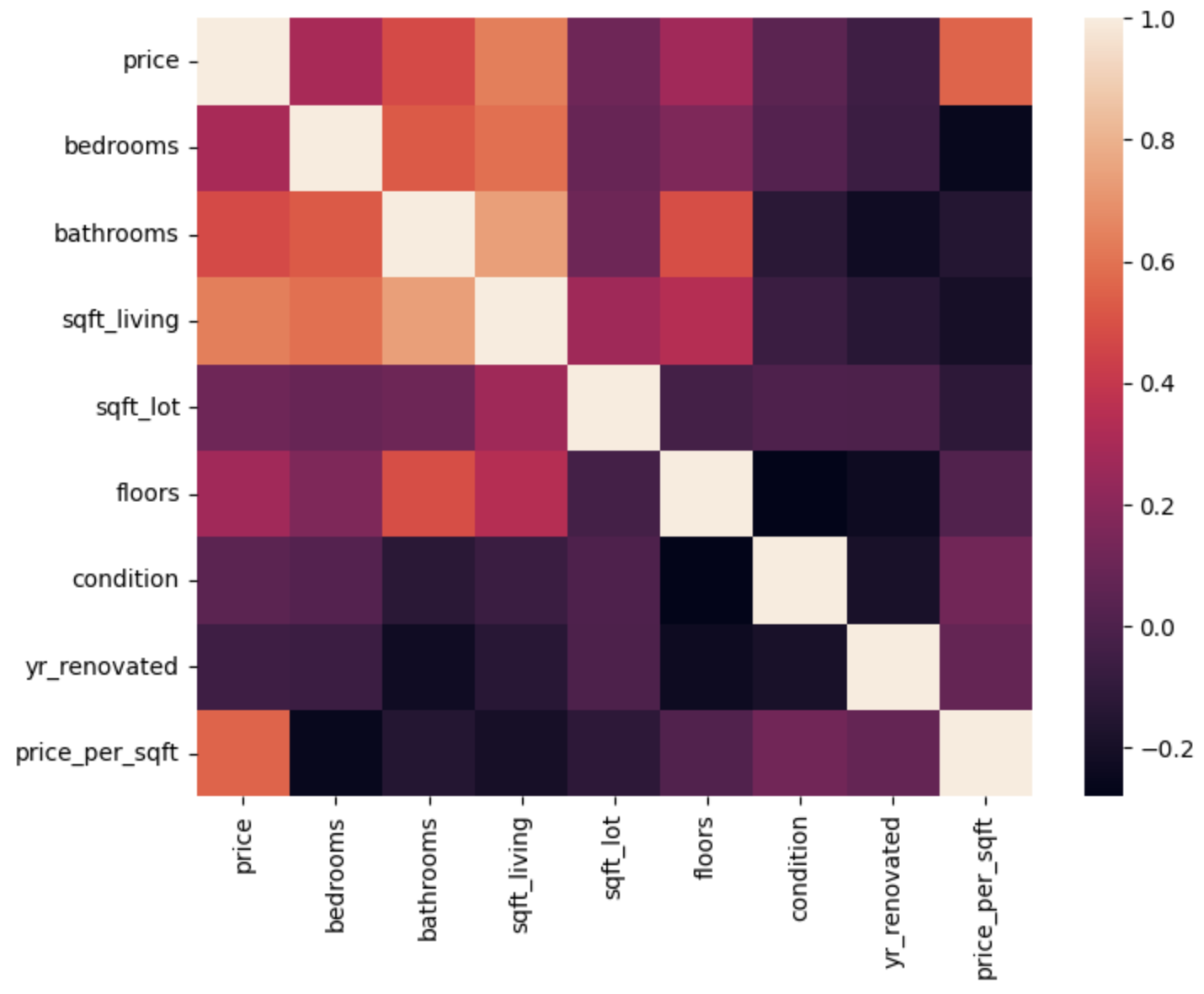
The last plot for the condition feature shows that most houses being sold are average condition, having basic amenities.

## Bivariate Analysis:

As for the Bivariate Analysis, I started by comparing each feature to the condition feature, understanding which condition of a house is more sought after in the housing market

As seen in the above box and whisker plots(the rest of the plots are in the python file), as well as the single plot relating price and condition in displot, most features tend to show that most houses available in the housing market tend to have an average condition, where the house has basic amenities, and has been renovated a few years prior being listed in the market. This shows that buyers tend to invest in an average home than a better quality household, as it has enough amenities and space to accommodate them.

## Multivariate Analysis:

As for the multivariate analysis, I used a heat map for all of the features and found out that statistically, the only features which have a relevant coefficient is price and price\_per\_sqft, where the sqft\_living feature has no correlation with each other, even though the price and soft living features have a correlation of around 0.7. this leads to a insight that the relationship between the price of a house, and the space used as living space does not have a linear relationship. All other features have low correlation with each other except for floors and sqft\_living, and floors with price, as both correlations have a value of around 0.3

# Overall Insights from Analysis

As for the overall analysis, for the three levels of analysis conducted, it is seen that only few features are statistically related with other features, leaving to the idea that only the interpretation of those data in the real world is what correlates of the the 9 features. Mainly, we have seen that the price feature can directly correlate to the sqft\_living and floors features, but not to the bedrooms and bathrooms features, even though in the real world, they have a correlation. However, most of the statistical test done assume that all of the above features have a linear relationship with one another. This is true for the relationships mentioned above, as more living space would allow for Moree bedrooms and bathrooms, as it costs more for more living space, hence the price, bedrooms, and bathrooms features are related to each other, but not in a linear fashion. This is further enhanced with the addition of the derived column price\_per\_sqft, as it directly correlates with price, with a little to no correlation with other features, showing that the derived feature has a different type of relationship with the other features.

This analysis can provide answers to key stakeholders in the housing industry, as it identifies that the price of a house is directly related to not only the area of the lot being sold, but the actual space being used as a house. We have already seen in the bivariate analysis that all features point to popular houses having an average condition, meaning that most houses do not tend to be very large, hence it is most suitable for small families. This is further insinuated by the fact that the plot shown below, as most houses fall under average condition, and are priced accordingly. This plot shows a key relationship between all the other features, as the darkest hue, representing the best houses, are rarely visible due to their high price, but contain the most amenities and space.

# Conclusion:

To conclude, for the analysis conducted in this project, there are only 4 features with direct relationships: price, price\_per\_sqft, floors, and bathrooms. All of the other features chosen through filtration do not have a direct relationship, as seen from the heatmaps, plots, and correlation matrices from Pearsons and Spearmans correlation coefficient, shows that there is no direct correlation between the remaining features, however, due to the scatter plot below, one is able to identify that, there may be higher degrees of relations between these features, which cannot be seen through the analysis conducted in this project. In the future, regression analysis will be required in order to further determine the relationship between these features, where polynomial or non-linear regression is recommended, due to the lack of direct relations between the 5 remaining features.