**TECHNICAL REPORT**

Data Analysis and Hypothesis Testing for Melbourne Letting Agency’s Real Estate Study

Assignment 001- CT4031

Contents

[Introduction 3](#_Toc136518716)

[Dataset analysis and Pre-processing 3](#_Toc136518717)

[Dataset visualization and proposed hypotheses 4](#_Toc136518718)

[Hypothesis testing 5](#_Toc136518719)

[Conclusion 5](#_Toc136518720)

[References 6](#_Toc136518721)

[Appendix with python code 7](#_Toc136518722)

# Introduction

This technical report consists of real estate listings in Melbourne, Australia, which is the second-most populous city in both Australia and Oceania. The investigation is based upon predicting the highly competitive real estate market making it crucial for industry players like the Melbourne Letting Agency (MLA) to gain insights and identify trends to stay ahead. In order to achieve this, MLA has collected and compiled a comprehensive dataset to support their studies and inform their decision-making process. The dataset contains valuable information about real estate listings in Melbourne, including property type, number of bedrooms, car spots, rental prices, and listing years.

Our objective is to conduct a thorough analysis of the dataset provided MLA and address two hypotheses proposed by the agency. The first hypothesis suggests that the probability of having two car spots, given that the house has two bedrooms, is below 7%. The second hypothesis proposes a significant price difference between renting units in the year 2016 compared to 2017.

To accomplish this, on follow a structured approach that includes dataset analysis and pre-processing, dataset visualization, hypothesis testing, referencing, and a Python code implementation. By fulfilling these requirements, the aim is to provide a comprehensive technical report that showcases our ability to analyse and interpret real estate data, apply statistical tests to evaluate hypotheses, and present the result in a clear and concise manner. The mathematical applications have also been used such as Bayes and KNN algorithms for the clarification.

# Dataset analysis and Pre-processing

During the analysis, several characteristics of the dataset were examined. The dataset contains 58,037 records. In the provided dataset, there are 21 columns and 2867 rows. The columns are Suburb, Address, Rooms, Type, Price, Method, SellerG, Date, Distance, Postcode, Bedroom2, Bathroom, Car, Land size, Building Area, Year Built, Council Area, Latitude, Longitude, Region name, Property count.

Data manipulation involves transforming, restricting, or modifying the dataset to make it suitable for analysis or to derive new insights. Aggregating and summarizing data, creating derived variables, restricting data

The problems encountered when cleaning the data that required attention and resolution were data inconsistencies, missing values, dropping columns, changing indexes, tidying up fields in the data, renaming columns and skipping rows. In data transformation, we perform many methods such as converting into uniform value.

|  |
| --- |
| Data Inconsistencies |
| Back slashes (//) mentioned in the Date’s column |
| Hashes (##) mentioned in the Room’s column |
| Question marks (??) mentioned in the Suburb and Building Area columns |
| The Euro symbol (£) mentioned in the Price column |

Table 1.1

The missing values have been added using mean, median and mode in the dataset, and some have been valued as 0.

# Dataset visualization and proposed hypotheses

Data visualization is the graphical presentation of data using visual elements like charts, graphs, and maps. Data visualization tools provide a comprehensible way to view and acknowledge trends, outliers, and patterns in data (Mohammad Farma et al., 2021). In the figure.1, it is known that the probability of having two car spots given that the house is comprised of two bedrooms is below 7%.

In the context of this assignment, dataset visualization played a crucial role in understanding the rental unit trends and supporting the hypotheses proposed by the Melbourne Letting Agency (MLA). Using various types of graphs and visualizations, valuable insights were gained, facilitating data analysis and interpretation.

One of the primary objectives of dataset visualization was to examine the price differences between renting units in the years 2016 and 2017. This involved creating visual representations such as box plots or bar charts to compare the rental prices across the two years. These visualizations allowed for a clear comparison, highlighting any significant differences in prices, and providing a basis for further analysis.

Additionally, scatter plots were employed to explore relationships between different variables, such as the number of bedrooms and car spots. By plotting these variables against each other, patterns or correlations could be identified, supporting, or refuting the hypotheses put forth by the agency. For example, scatter plots could reveal whether there was a relationship between the number of bedrooms and the availability of car spots.

Furthermore, histogram visualizations were utilized to examine the distribution of rental prices or other relevant variables. By binning the data and creating a histogram, the frequency and distribution of prices could be observed, aiding in identifying any notable trends or patterns.

Overall, dataset visualization was instrumental in providing a visual representation of the rental unit data, allowing for a comprehensive understanding of the trends, patterns, and relationships within the dataset. These visualizations served as a foundation for hypothesis evaluation and supported the analysis and conclusions drawn from the dataset.

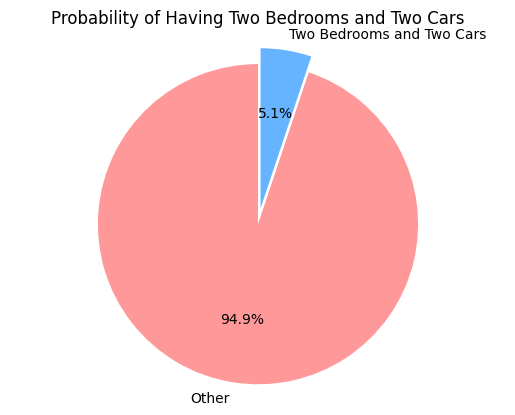


Figure.1

Less than 7% = 5.129099790648988

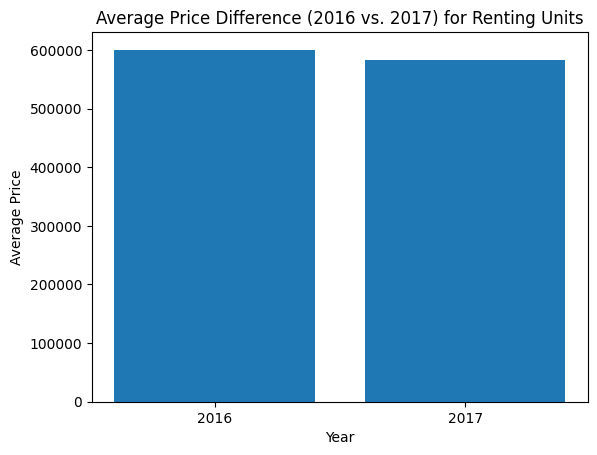


Figure.2

# Hypothesis testing

In hypothesis testing, we typically perform statistical tests to evaluate the hypotheses proposed by the agency and provide evidence to support your arguments. These are the steps to follow for the first hypothesis:

1. Null hypothesis (H0): The probability of having two car spots given that the house is comprised of two bedrooms is 7% or higher.
2. Alternative hypothesis (H1): The probability of having two car spots given that the house is comprised of two bedrooms is below 7%.

The significance level (denoted as α) is the probability of rejecting the null hypothesis when it is true. Commonly used significance levels are 0.05 (5%) and 0.01 (1%). These are the statistical tests chosen-

1. Chi-squared test: This test is used to determine if there is a significant association between two categorical variables, such as the presence of two car spots and the number of bedrooms (Kim, 2017) (Studio, 2022).
2. T-test: A t-test can be used to compare the means of two groups, such as comparing the rental prices in 2016 and 2017 to evaluate the hypothesis of a massive price difference.
3. Analysis of Variance (ANOVA): ANOVA is useful when comparing means across more than two groups. If your dataset includes multiple years or other categorical variables, ANOVA can help evaluate the price difference hypothesis (Liu & Wang, 2020).
4. Correlation analysis: You can perform correlation analysis to assess the strength and direction of the relationship between variables, such as the correlation between the number of bedrooms and the number of car spots.
5. Regression analysis: Regression analysis can be used to model and examine the relationship between variables, allowing you to assess how well the variables predict each other (Tyagi et al., 2022).

On attempting the tests on the data given the null hypothesis has been successful, suggesting evidence in support of the alternative hypothesis whereas there is insufficient evidence to support the alternative hypothesis. (Refer to Figure.7 in the appendix)

For the second hypothesis:

1. Null Hypothesis (H0): There is no significant price difference between renting units in year 2016 and year 2017.
2. Alternative Hypothesis (Ha): There is a significant price difference between renting units in year 2016 and year 2017.

The tests which can be done for this hypothesis is:

1. Independent t-test: If you have two independent samples of rental unit prices for the years 2016 and 2017, you can perform an independent t-test. This test compares the means of the two groups (prices in 2016 and prices in 2017) to determine if there is a significant difference between them.
2. Mann-Whitney U test: If the assumptions of the independent t-test are not met (e.g., non-normal distribution), you can use the Mann-Whitney U test. This test is a non-parametric alternative to the t-test and compares the medians of the two groups.
3. Paired t-test: If you have paired data where rental unit prices for the same units are collected in both 2016 and 2017, you can perform a paired t-test. This test compares the means of the paired samples to determine if there is a significant difference.
4. Wilcoxon signed-rank test: If the assumptions of the paired t-test are not met, you can use the Wilcoxon signed-rank test. This non-parametric test compares the medians of the paired samples.

Here, we can see that there is not a massive price difference in the year 2016 to the year 2017. (Please refer to figure.5 in the appendix)

# Conclusion

In conclusion, this assignment involved analyzing and evaluating hypotheses proposed by the Melbourne Letting Agency (MLA) regarding rental units. The assignment consisted of three main sections: Dataset Analysis and Pre-Processing, Dataset Visualization, and Hypothesis Testing.

In the Dataset Analysis and Pre-Processing section, the provided dataset was examined, and various problems encountered during data manipulation were addressed. Steps were taken to clean the dataset, ensuring the data was valid and suitable for analysis.

The Dataset Visualization section involved visualizing the dataset using different types of graphs to explore and support the proposed hypotheses. Graphs such as box plots, scatter plots, or histograms were used to analyze and present relevant information related to the hypotheses.

Finally, in the Hypothesis Testing section, statistical tests were applied to evaluate the hypotheses proposed by the agency. Specific tests such as the independent t-test, Mann-Whitney U test, paired t-test, or Wilcoxon signed-rank test were chosen based on the nature of the data and the research questions. These tests provided insights into the hypotheses and allowed for a statistically supported analysis of the rental unit trends.

In this document, the two hypotheses have been thoroughly proved through the Python code using comments in the appendix. This gave us insight on big data and mathematical conclusions in real-life experience.

Overall, this assignment aimed to demonstrate the ability to analyze and evaluate hypotheses using statistical tests and provide meaningful conclusions based on the results. By conducting a thorough dataset analysis, performing appropriate statistical tests, and interpreting the findings, the assignment facilitated a comprehensive understanding of the rental unit trends and provided insights for decision-making in the real estate industry.

# References

F. Mohammad Farma, K. Barevan, I. Tarigan and R. Fitri Sari, "Sentiment Analysis of Video Game Console Pre-launching Tweets Using Python," 2021 IEEE 7th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia, 2021, pp. 1-6, doi: 10.1109/ICCED53389.2021.9664856.

*Indexing and selecting data#* (no date) *Indexing and selecting data - pandas 2.0.2 documentation*. Available at: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy (Accessed: 01 June 2023).

*IEEE Xplore* (no date). Available at: https://ieeexplore.ieee.org/Xplore/guesthome.jsp (Accessed: April 22, 2023).

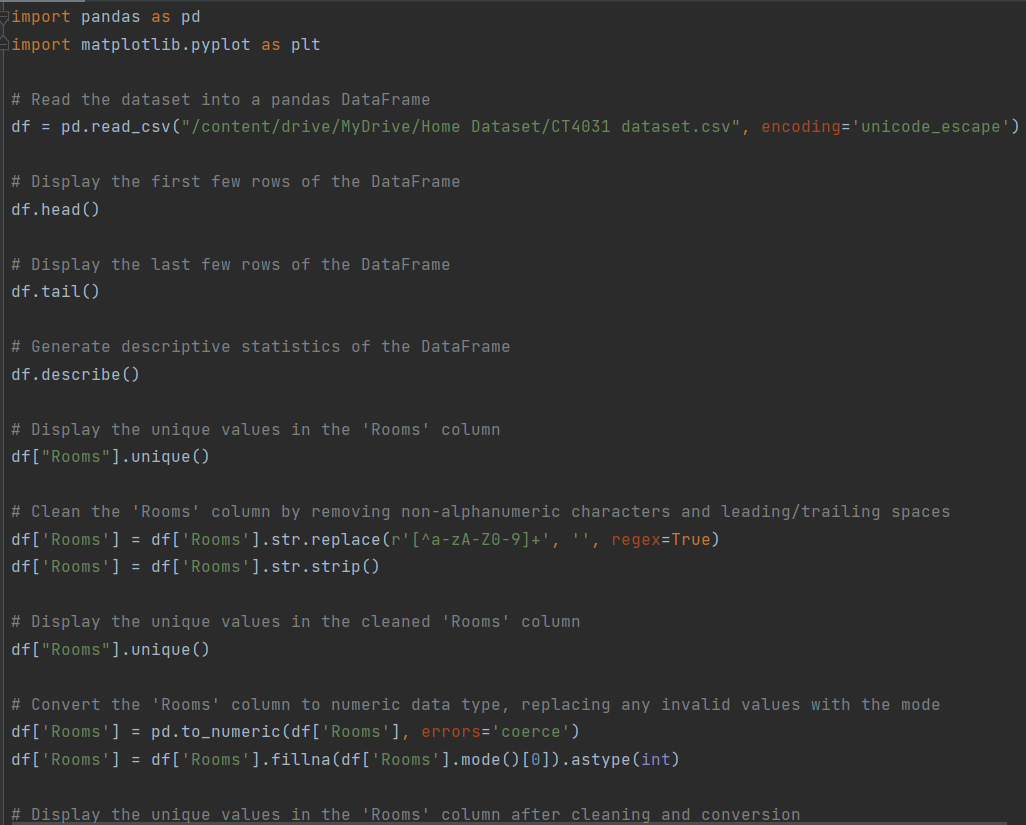
Kim, H.-Y. (2017) ‘Statistical notes for clinical researchers: Chi-Squared Test and Fisher’s exact test’, *Restorative Dentistry & Endodontics*, 42(2), p. 152. doi:10.5395/rde.2017.42.2.152.

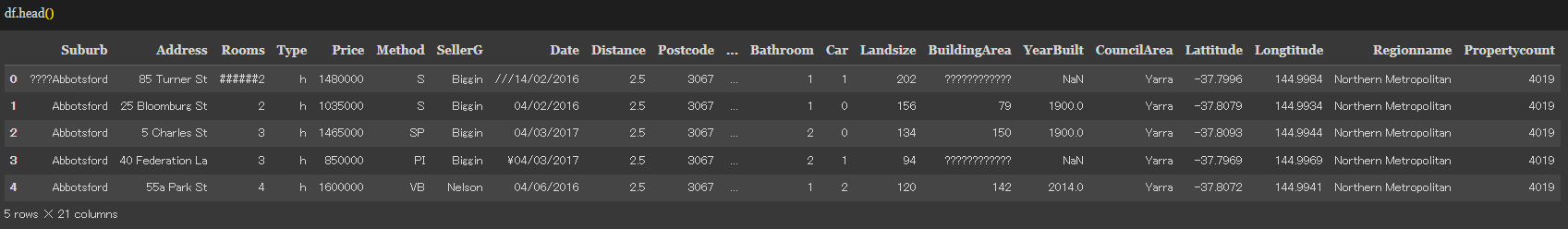
Liu, Q. and Wang, L. (2020) ‘T-test and ANOVA for data with ceiling and/or floor effects’, *Behavior Research Methods*, 53(1), pp. 264–277. doi:10.3758/s13428-020-01407-2.

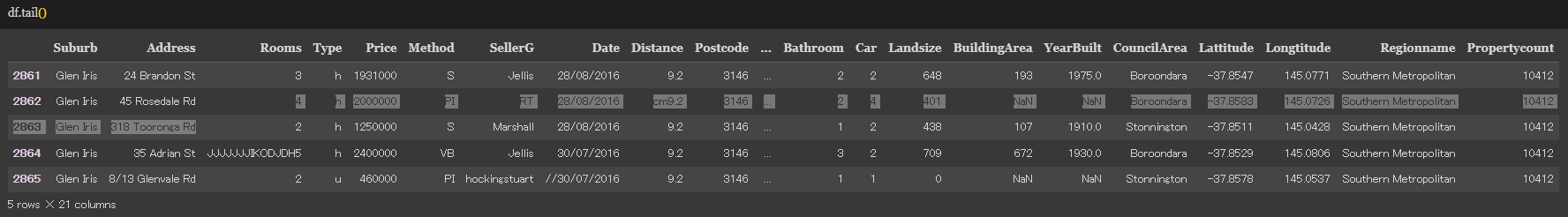
Studio, V.D. (2022) *An interactive guide to hypothesis testing in Python*, *Visual Design*. Available at: https://www.visual-design.net/post/an-interactive-guide-to-hypothesis-testing-in-python (Accessed: 01 June 2023).

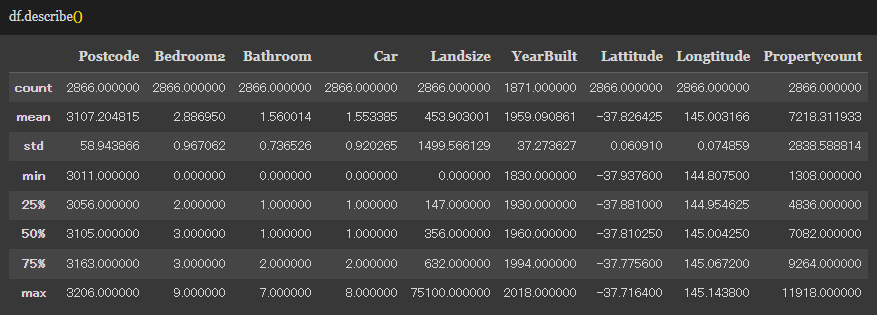
Tyagi, K. *et al.* (2022) ‘Regression analysis’, *Artificial Intelligence and Machine Learning for EDGE Computing*, pp. 53–63. doi:10.1016/b978-0-12-824054-0.00007-1.

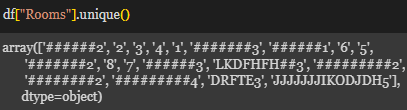
# Appendix with the Python code

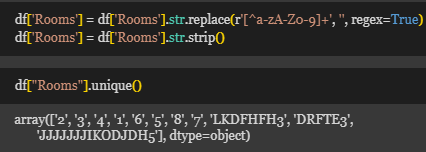


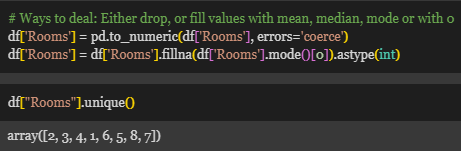


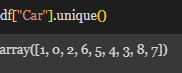


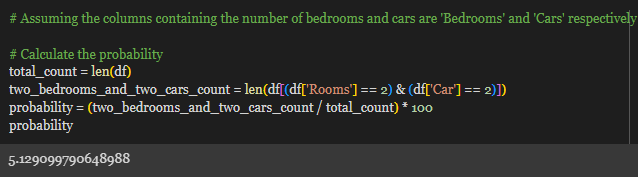




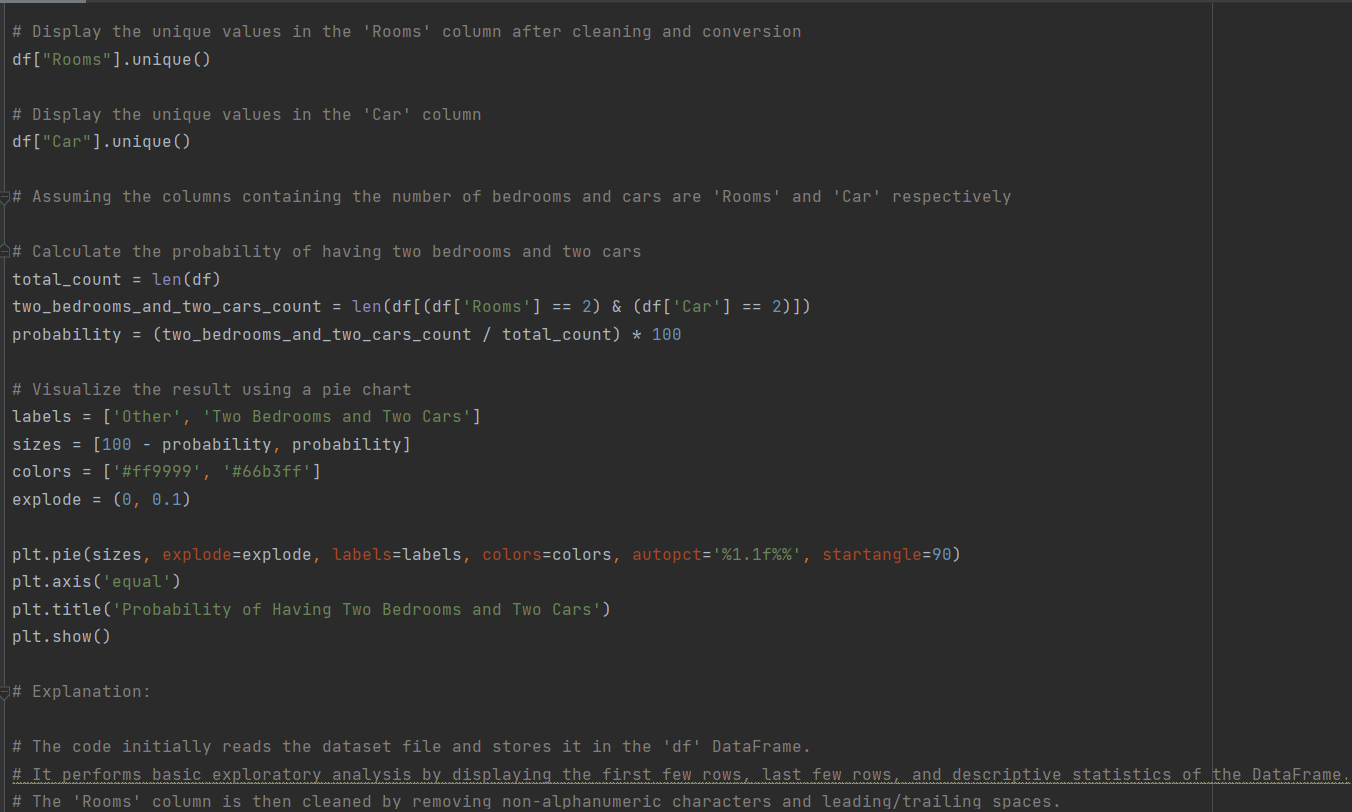








*Figure.7*



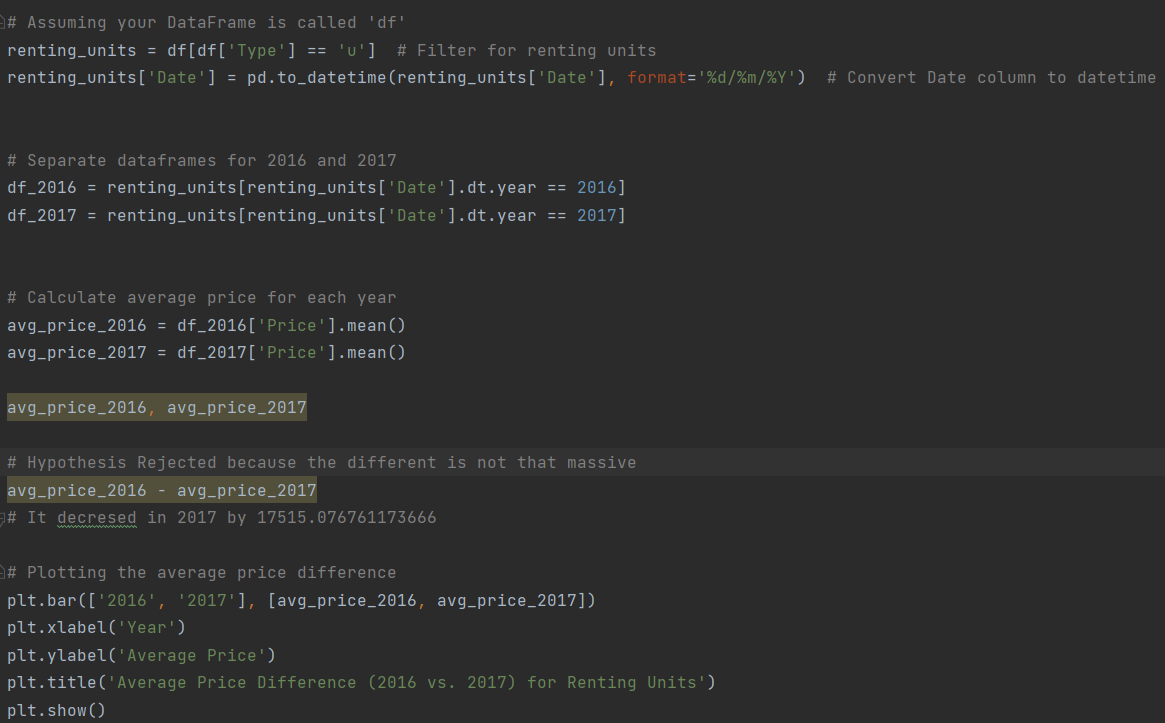


Figure.5

To test the hypothesis (see figure.5 for reference) of a massive price difference between 2016 and 2017 for renting units, you can follow these steps:

Filter the dataset to include only the "Type" column values indicating renting units. Extract the relevant columns, including "Price" and "Date". Convert the "Date" column to the appropriate datetime format. Create separate data frames for 2016 and 2017. Calculate the average price for each year. Plot the average price for 2016 and 2017 to visualize the price difference.

renting units['Date'] = pd.to\_datetime(renting units['Date'], format='%d/%m/%Y') this code Converts Date column to datetime

