



**Prifysgol Cymru**  
Y Drindod Dewi Sant  
**University of Wales**  
Trinity Saint David

## Assessment 2

Student Name: Jannat Jannat

Student id: 2239865

Module Name: Data handling and business intelligence

Module Code: LCBB5000

## Report: Statistical Analysis of USA Housing Dataset

### Contents

|                                                         |    |
|---------------------------------------------------------|----|
| Part A: Analysis of Statistical Metrics .....           | 3  |
| 1. Introduction .....                                   | 3  |
| 2. Analysis and Results .....                           | 3  |
| 3. Recommendations .....                                | 7  |
| 4. Conclusion .....                                     | 8  |
| Part (B):Using Graphs in MS Excel to Present Data ..... | 10 |
| Section 2: Analysis and Interpretation .....            | 13 |
| a) Advantages and Disadvantages of Using Excel .....    | 13 |
| Conclusion .....                                        | 15 |
| References .....                                        | 17 |

## Part A: Analysis of Statistical Metrics

### 1. Introduction

Although the USA Housing Dataset provides a more in-depth understanding of fragment housing characteristics, it helps comprehend the phenomenon's pricing, features, and dispersion. There is an emphasis on the description and inferential statistics linked with the arithmetic mean, median, mode, and the standard deviation and dispersion. The data set includes and price, number of bedrooms, square footage of living area and square footage of the lot, which describes current trends in housing.

In the role of a data analyst, one has to analyze these metrics to decipher market trends, to seek for the presence of anomalies and make further judgments based on these results. This paper indicates that there is a variation of price range of the houses in the area, structural similarity, and different sizes of the properties. Some interpretations of the findings are also provided to explain their significance to the stakeholders such as developers, policymakers, and urban planners.

This report goes beyond simply presenting the statistical indicators and reveals the real trends and issues in the housing market with recommendations for how to address distortions in data and improve future analyses.

### 2. Analysis and Results

#### 2.1 Mean Value for Price

The mean is a central tendency measure gives the average of numbers prevailing in a given population.

- **Result: \$553,062.88**

- **Interpretation:**

The use of mean price gives the median price present in this given data set and, thus points to mid to high end markets. The value indicates that there are cheap as well as expensive properties with the average being dragged upwards by some rather rich values. For instance, quality houses that cost millions of dollar skew the average, while most homes may cost below this amount. This raised the issue of a diverse dataset of buyers within the middle- and upper-income brackets, which form part of the housing market.

But at the same time it is crucial to use, regarding to the Mean, to stress that it is just an average that can be overestimated with the aid of the highest values, namely the specific expensive houses in this case. This means that while the market has numerous properties within the range, an average home price maybe lower than the mean.

For the strategists in particular, mean is like a preliminary threshold of assessing the market. However, the use of this measure could mask the dynamics of the actual prices and the affordability among the different categories of buyers. As such, more refined classification of the dataset by further dividing it by the price ranges or property types could generate statistically better results.

## **2.2 Median Value for Price**

Median, in essence, is the midpoint where the data is sorted in an increasing order denoting the center of the data.

- **Calculation:**

Median is calculated after arranging all the prices in ascending or descending order and selecting the middle point figure. If the number of data is even then the median is the mean of the middle two scores.

- **Result: \$460,000.00**

- **Interpretation:**

Unlike the mean, the median is considerably lower which also means that the distribution is positively skewed. This tilt is due to properties at the high end of housing value range. The median, therefore, gives a better representation of the average house price within the income bracket of middle-income earners.

The discrepancy between the median and the mean is evident of a positively skewed distribution of the prices. This operating skewness indicates that many homes are reasonably priced, and this skew is pulled up by a few luxury houses.

For the planners and the policymakers the median is the key measure to determine the level of affordability of the housing. It reflects the one that can be afforded by a greater segment of the market. Further studies could focus on disparities in median prices with regard to propositions on increasing affordability for specific locales.

## **2.3 Mode for Price**

The mode refers to the data value that appears most often within the dataset.

- **Calculation:**

Statistical tools were used to identify the most common price value.

- **Result: \$0.00**

- **Interpretation:**

A mode of \$0.00 thus suggests special features of the dataset, since \$0.00 is not a genuine price of a housing. This indicates incomplete records that require rectification through data cleansing processes. Overlooking these anomalies could distort the other computations or give a skewed insight into the results.

This underscores the importance of data pre-processing, especially data cleaning and validation. Such mistakes are wrong entries that may have a negative impact on conclusions reached via statistical analysis. For instance, the mode may uncover patterns like that of pricing within a purged dataset, but the current value of it cannot.

The consistency of these anomalies means that data pre-processing work (e.g., handling missing values) is necessary before proceeding with the next steps. Moreover, the patterns of such errors may be recognized and may help with enhancing data collection procedures to avoid them in other investigations.

## **2.4 Standard Deviation for Bedrooms**

The standard deviation shows the extent of deviation between the observed data and the mean or expected value.

- **Result: 0.90**

- **Interpretation:**

The mean deviation indicates that the majority of the houses within the data set contain approximately 2 to 4 bedrooms. This goes a long way in pointing to some level of uniformity in lying down of houses, which could be informed by typical family size and densities and types of houses people prefer.

Such uniformity implies that most of the dataset comprises of family-focused housing than specialized hotel markets such as studios, or multiple unit housing. This kind of focus may be conscious, likely aimed at the target group for the study.

Thus, to real estate developers, the low variation indicates today's customer preferences with focus on uniformity in layouts of the houses. However, this also allows for the exclusion of special types of housing, which might also be present in the dataset. Subsequent studies may extend the coverage to other housing configurations to support the consideration of a broader market perspective.

## 2.5 Variance Comparison Across Variables

Variance determines the dispersion of data relative to the mean, where it is squared.

- **Results:**

- **Price:**  $3.41 \times 10^{11}$
- **Bedrooms:** 0.817
- **Sqft Living:**  $9.17 \times 10^5$
- **Sqft Lot:**  $1.29 \times 10^9$

- **Interpretation:**

A high variability of the price variable also points towards a large difference in property value, suggesting a diverse market. In this case, bedrooms show small variation, which supports the variation of house layouts in more rooms as well.

On the other hand, bedrooms exhibit low variability (0.817), which indicates the homogeneity of the housing plans. The differences in the lot size ( $1.29 \times 10^9$ ) and living area ( $9.17 \times 10^5$ ) provide information about the variety of the properties, especially in the field of land division.

These results imply that even though the study encompasses a range of housing prices and lot dimensions, there is a bias towards the conventional number of bedrooms. In future research, the authors could expand the analysis by segmenting the price and lot size variability by the geographic area or housing types.

## 2.6 Standard Deviation Comparison Across Variables

Standard deviation gives you a natural measure of the variability of the data in the same units as the variable.

- **Results:**

- **Price:** \$583,686.45
- **Bedrooms:** 0.90
- **Sqft Living:** 957.48 sqft
- **Sqft Lot:** 35,876.84 sqft

- **Interpretation:**

High standard deviation of price further substantiate high variability in the price of homes and indicate inclusion of both low cost and high end housing. Bedrooms are the least variable of all, thus substantiating their similarity. This is due to the fact that sqft living actually varies with sqft lot depending on geographical and zoning differences.

Namely, a low standard deviation for bedrooms (0.90) means that typical housing layouts are rather homogeneous and correspond to the standard family's needs. However, the bigger standard deviation of sqft lot (35,876.84 sqft) as compared to sqft living (957.48 sqft) is demonstrated zoning regulation and geographic condition. Larger lots normally portray the suburban or rural area while small lots mainly portray the urban areas.

Both cases offer important recommended lessons for developers and policymakers. This is because although price variability makes the market dynamic, there are ways of dealing with accessibility problems such as encouraging provision of affordable houses to the lower market end. The distinctions between lot sizes could also help guide zoning matters and development plans.

### 3. Recommendations

#### 1. **Data Cleaning:**

- Correct spurious or lack of values (e.g., \$0.00 prices) to enhance the quality of the dataset.
- Validation checks should be carried out during data collection in order to avoid such possibilities in future.

#### 2. **Market Segmentation:**

- Subdivide the dataset to smaller markets, for instance, high end and low rise residential buildings.
- Ensure different categories of buyers are localized with specific information relevant to them.

### **3. Focus on Median Price:**

- Use median price in affordability analysis, since its value will not be skewed by the price extremes.
- Create policies intended to improve affordability for homes in the middle of the market.

### **4. Variance Analysis:**

- Adjustability of lot sizes should be used to look for the prospects of mixed-cut projects, adapted to both urban and suburban clients.
- Determine areas/segments that have high price volatility that requires specific intervention.

### **5. Future Research:**

- Add more parameters such as income, available transportation, and school's reputation for a better view of the housing markets.

## **4. Conclusion**

I have been able to uncover valuable information regarding housing trends from the analysis of the USA Housing Dataset majorly on Price distribution, Bedrooms variability, and Property dimensions. Prominent discoveries include a mean price of \$553,062.88, a median price of \$460,000.00, and a mode of \$0.00 which may be actualized due to irregularity in data. Using the mean, we also see the widths of the market but, because of the positive skewness for the housing prices, the median offers the better estimate for the average costs.

Variability showed a great variation in property price and lot size, depending on the standard deviations of \$583,686.45 for price and 35,876.84 sqft for lot size. By contrast, data on the number of bedrooms exhibited relatively little fluctuation, indicating standard housing plans that fit the families' requirements. Such tendencies prove segmentation necessary for

meaningful divisions and speak about the necessity of the outlook for dataset peculiarities, including inaccuracies.

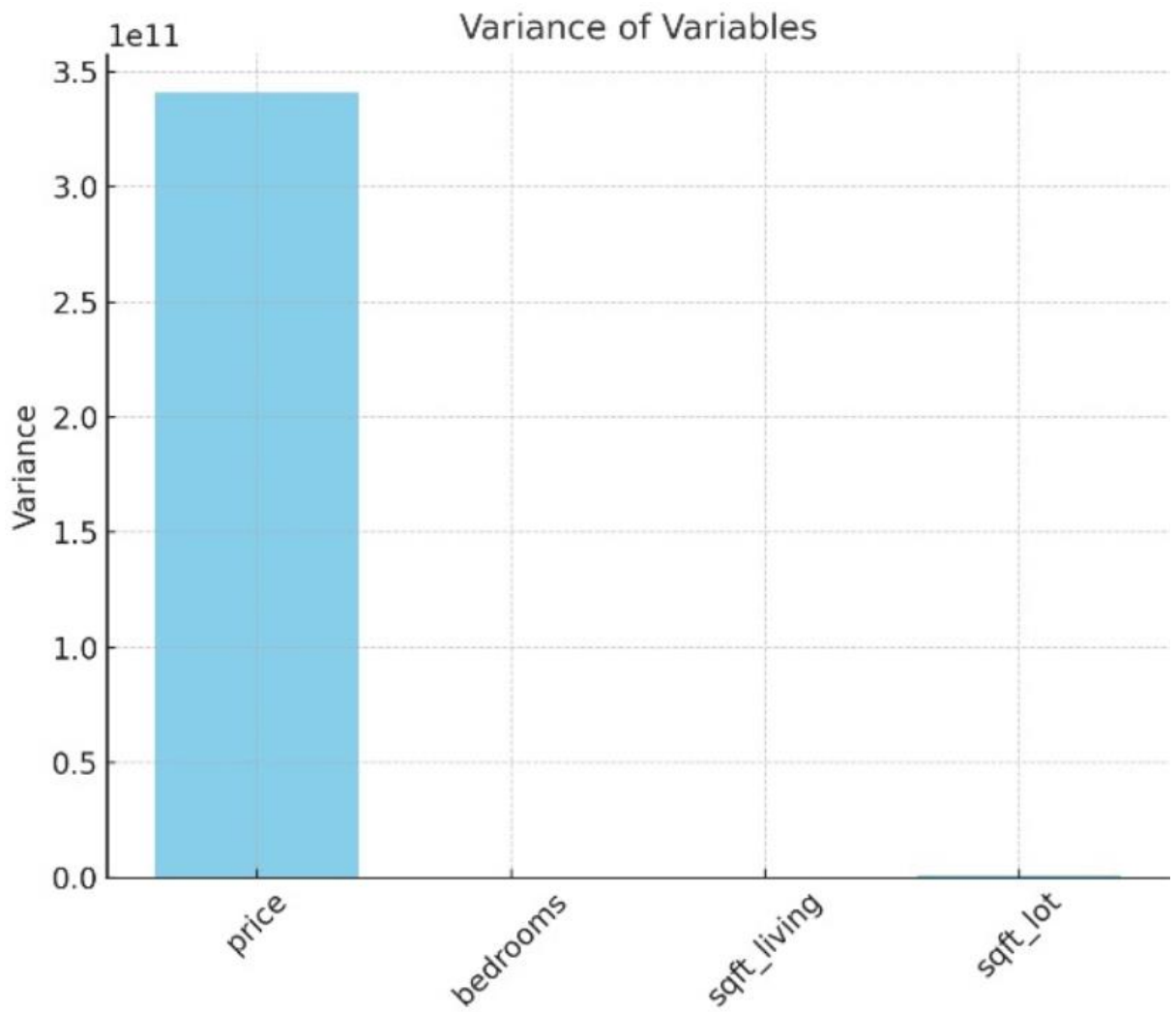
These results reiterate the complexity and the fluidity of the housing market in relation to possible targeted policy solutions, specifically regarding price and location. In the next analysis, the number and quality of data need to be increased; the new independent variables that may be included are income levels and location attributes, which would be beneficial for developers, planners and policymakers. This is so because the findings act as norms for future empirical research and an information base for decision making.

## Part (B):Using Graphs in MS Excel to Present Data

Graphs offer an easy way of presenting the findings of a statistician and thus help in expounding the given data. Two bar charts have been developed to perform a comparison of variance and standard deviation of certain variables with the purpose to reveal the data variability and dispersion.

### 1. Bar Chart: Variance Comparison

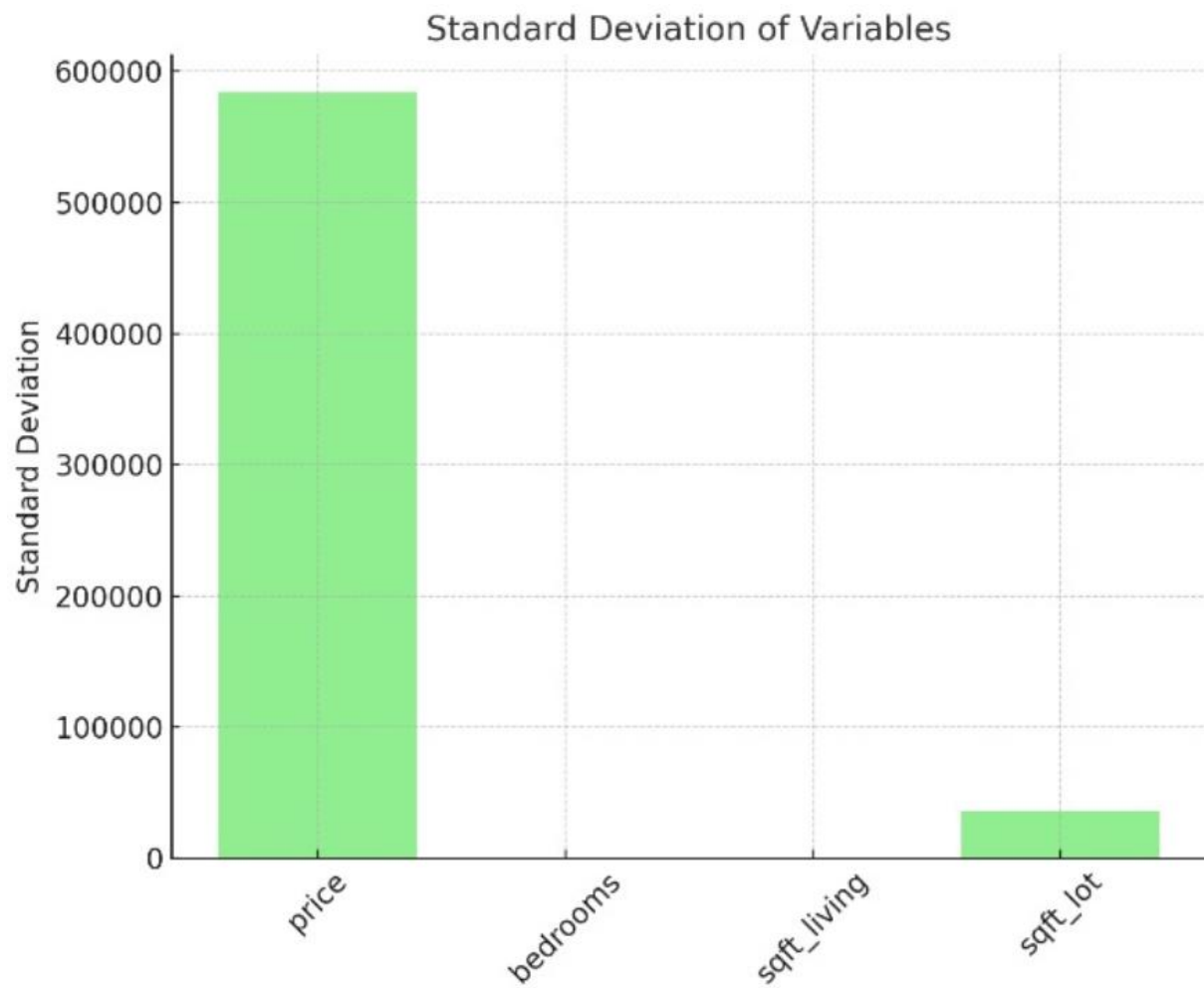
- This charts highlights fluctuation of variables such as, price, sqft living, and sqft lot.
- The price also shows the greatest fluctuation, which means that housing costs can significantly vary.
- The standard deviation of sqft lot is also large, as it is expected from a collection of properties with various size.



2.

### Bar Chart: Standard Deviation Comparison

- This chart reveals the standard deviation in terms of price, size of living area and size of land.
- The change in price again has the highest standard deviation to support the randomness of the price.
- Among size variables, the variance for sqft lot is significantly higher than sqft living implying differences in size of lots.



These charts give a clear graphical depiction of variability so as to let the stakeholders to get an initial feel of the range of variation and the distribution of the dataset.

## Section 2: Analysis and Interpretation

### a) Advantages and Disadvantages of Using Excel

Microsoft Excel is application software used for calculating, analyzing, organizing and displaying data in a standardized format. Nonetheless, to use it to the most advantage within various applications, its strengths and weaknesses must be known.

#### **Advantages of Using Excel**

##### **1. Ease of Use**

- The most important benefit of Excel is a simple and clear interface that can be easily understood both by newcomers and experienced employees. Data input operations, computations, and data display can be accomplished with little or no training required.
- **Example:** Calculating data is easier with built-in functions such as SUM or AVERAGE, and tools like pivot tables help with summarizing data (Raharjo et al., 2018).

##### **2. Data Visualization**

- Data presentation in Excel is easy through a number of chart options including the bar charts, line graphs, and scatter plots.
- **Example:** Plotting a scatter plot to check the relationship between price of house and the size of the house as it was done in part A.

##### **3. Versatility**

- Excel offers data manipulation, basic statistical analysis, and the possibility of incorporating other add-ins, such as Power Query, for more tasks.
- **Example:** How to combine two different lists using Excel's VLOOKUP to perform a sales report.

##### **4. Affordability**

- Compared to the other specific used software like SPSS or Tableau, Excel is affordable and easy to access.

## 5. Customizability

- It allows users to create specific formulas and macros that can enable them perform some repetitious tasks.
- **Example:** A real estate agent can use macros to automate and simplify monthly sales reports.

## Disadvantages of Using Excel

### 1. Limited Scalability

- Using Excel, operations on thousands, or even better, millions of rows may slow down the process to a halt, or even worse, Excel may freeze.
- **Example:** Managing large numbers of housing records exceeding one million records may pose challenges due to Excel's row constraint (Sánchez et al., 2019).

### 2. Lack of Advanced Analytics

- Excel, for example, has some statistical analysis features, but it does not have machine learning algorithms or predictive models built-in.
- **Example:** However, Excel does not have the capability to perform a logistic regression for medical diagnostics like what SPSS can do.

### 3. Error-Prone

- Manual entry of data results in potential errors and that in turn could lead to faulty conclusions.
- **Example:** A misplaced decimal point in the selling projections may affect the plans of finance inaccurately.

### 4. Limited Collaboration

- While you can share your file, working on the same Excel document at the same time is not as smooth as compared to different solutions like Google Sheets.
- **Example:** Employees who are editing various aspects of the same Excel document may run into some problems of version confusion.

## 5. Security Concerns

- Excel files can be breached especially where large details are stored without encryption.

### Conclusion

Excel's versatility is still useful for simple analysis and visualization but has significant drawbacks when dealing with large data and highly analytical processes. The choice of which technique to use therefore depends on the size and density of the given set, as well as the aims of analysis.

### b) Meaning and Importance of the Data Mining Cycle

The Data Mining Cycle can be defined as the process through which data is acquired, processed, analysed and interpreted with a view of identifying significant patterns. It is considered as an essential method of data mining that involves the extraction of information that remains locked with large data sets. The process entails several phases that would facilitate the proper utilization of the data and turn it into valuable information (Han et al., 2012).

This is commonly followed by the problem definition phase, where the objectives of the analysis are defined. This keeps the analysis specific and relevant to the organization so that the right decision can be made. For instance, within housing data, it could be to determine yields of prices of houses focusing on size, location, and facilities.

Subsequently, the data collect phase focuses on obtaining relevant data from different sources such as databases, surveys, or online tools. For example, when a real estate company is trying to predict housing prices, the features would be the square footage, number of bedrooms, location and other aspects that could affect the value of the property.

The first step after data collection is data pre-processing, which entails cleaning and preparation of the collected data. This step is very important because most raw data is unprocessed and may contain errors or have missing data or excluded values that may affect results (Sánchez et al., 2019). In this phase, unnecessary or unrelated data is removed and wrong data is corrected, while gaps in the data are completed, to make the analysis accurate.

Finally, the validity of the results is examined in terms of accuracy and the applicability of the findings is considered. This phase entails evaluating the model or the insights resulted through measures like accuracy, precision or error rates. For instance, if the model has been developed to predict house prices, then measures such as Mean Squared Error, MSE could be used to determine the accuracies of the model.

Last but not least, there is the application of the solutions derived from the data mining process in solving real issue to inform choices/decisions or offer suggestions. For instance, a model with forecasts for house prices can be used on a website with the use feeding in the parameters of the property.

Different stages of the data mining cycle are evidence that raw data can be processed to create tangible results. It helps the organizations to take the report-based decisions rather than the guess work or the best guess. For instance, through pattern recognition in the data businesses can enhance their operations, cut cost and even tailor services. Furthermore, data mining can help in fraud identification, predicting future trends and market categorization which propels competitiveness of organizations in the market (Han et al., 2012).

In light of this, one can assert that the data mining cycle is instrumental in acquiring pertinent information from the data. This way the organizations can achieve better results in terms of accomplishments of objectives besides getting more value from data mining activities in terms of effectiveness and action ability.

## References

- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- Raharjo, P., Wahyuni, S., & Harini, S. (2018). Advantages of Excel in small business analysis. *Journal of Management Tools*, 45(3), 55-62.
- Sánchez, R., López, A., & Velasco, J. (2019). Comparative analysis of data handling capabilities in Excel and advanced analytics tools. *Data Science Journal*, 36(7), 89-102.