



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

Component 2: Practical: Essay

Name: Kristian Nash

Student ID: 2130196

Module: Data Handling and Business Intelligence

Module Code: LCBB5000

Lecturer: Dr Sadeque Shaikh

Contents

Section 1	3
Introduction	3
Data Overview.....	3
Data Analysis.....	4
Correlations/Graphs.....	8
Section 2	10
Advantages of Using Excel	10
Disadvantages of Using Excel.....	10
Data Mining Cycle	11
Conclusion.....	12
References.....	14

Section 1

Introduction

This report examines housing data to uncover trends and characteristics of homes in the USA from May 9, 2014, to July 10, 2014. I aim to provide insights into different aspects of the housing market like house prices, the number of bedrooms and bathrooms and the size of these homes. This analysis will help us understand the dynamics of the housing market during this specific period.

Our study focuses on several key statistical measurements to offer a view of the housing market:

- **Average Price:** We calculate the average (mean) price of houses to give us an idea of what a typical house might cost during this period.
- **Median Price:** This is the middle price point of all the homes listed, showing us that half the homes cost more and half cost less than this amount. It helps provide a more balanced view by reducing the skew from very high or very low prices.
- **Mode of House Prices:** The mode tells us which price appears most frequently. This information is useful for understanding the most common price point in the market.
- **Variation in Bedrooms:** We will look at how the number of bedrooms varies across different homes to get the diversity in house sizes.
- **Comparison of Variability:** By comparing the variance and standard deviation of house prices and the number of bedrooms, we can see which aspect of housing varies more from one property to another.

To perform this analysis, we are using Microsoft Excel, a tool widely used for its utility in managing and analysing data. Excel is especially suited for this kind of work because it simplifies the process of calculating statistical measures and visualising data trends through charts and graphs.

This report not only aims to provide a snapshot of the housing market but also to show how this data can be effectively used to make informed decisions. For example, understanding price trends and housing features can help potential buyers or real estate professionals in making better investment choices.

Also, our findings will shed light on the broader impacts of housing market trends for economic planning and policymaking. By understanding the factors that influence housing prices and preferences, policymakers can better address issues related to urban development and housing affordability. (Zhao, C. and Liu, F. 2023)

Data Overview

The dataset we're analysing, titled "USA Housing Dataset," includes detailed information on various properties sold between May 9, 2014, and July 10, 2014. It consists of several key columns such as price, bedrooms, bathrooms and square footage of living space. Each entry in the dataset represents

a single property, providing us with insights into the physical characteristics and pricing details of homes across different regions in the USA.

In total the dataset contains records for 4,140 properties, but there are some missing values in categories like price and number of bedrooms, which we will need to address during our analysis. The dataset includes various types of properties, ranging from single-bedroom apartments to large family homes showing the diverse housing options available in the market.

Data Analysis

The **mean**, or average, is a simple way to understand what typical values look like in a set of data. It's like finding out what the common amount is. In our study of house prices, the mean price turned out to be about \$553,063. This average gives us an idea of how much money you might expect to pay for a house. It can be found in figure 1.

However, it's important to take a closer look at how we got this number. Our calculation of the mean includes some houses listed at \$0. These \$0 listings are likely errors or maybe houses that were not actually for sale but got included in the data anyway. Including these in our average makes the mean less accurate because it pulls the average price down lower than it should be.

If we remove the \$0 house prices from our data, the average cost of the houses would likely be higher, giving a truer picture of the market. Removing the \$0 prices would mean only looking at houses that have a real selling price. This would help us avoid wrong conclusions about the average house price. It's a important step because it helps that the information we use is correct and reliable. By adjusting our calculations to skip the \$0 prices, we get a more accurate average that better reflects what people actually pay for houses.

This example shows why it's very important to check all the details in the data we use. Just taking numbers at face value without understanding what they truly represent can lead to mistaken ideas. For someone looking to buy a house, knowing the real average price is very useful because it helps them plan their budget better. It tells them what price range is typical and what they might expect to spend. Therefore, ensuring our data is clean and accurate helps make our analysis useful and trustworthy. (Team, L. 2019)

Figure 1

B4142		=AVERAGEIF(B2:B4141,"<>0")	
A	B	C	D
4115	02/07/2014	584000	3
4116	02/07/2014	0	3
4117	02/07/2014	0	5
4118	02/07/2014	245000	3
4119	03/07/2014	287919.7826	2
4120	05/07/2014	672500	4
4121	06/07/2014	454790	2
4122	07/07/2014	282508.8889	3
4123	07/07/2014	473200	3
4124	07/07/2014	406062.5	2
4125	07/07/2014	282766.6667	4
4126	07/07/2014	486445.8333	3
4127	07/07/2014	486895	3
4128	07/07/2014	430277.7778	3
4129	08/07/2014	229629.5	2
4130	08/07/2014	0	4
4131	08/07/2014	182805	3
4132	08/07/2014	380680.5556	4
4133	08/07/2014	396166.6667	3
4134	08/07/2014	252980	4
4135	08/07/2014	289373.3077	3
4136	09/07/2014	210614.2857	3
4137	09/07/2014	308166.6667	3
4138	09/07/2014	534333.3333	3
4139	09/07/2014	416904.1667	3
4140	10/07/2014	203400	4
4141	10/07/2014	220600	3
4142	MEANIF	559687.1933	
4143	1.mean	553064.477	

The **mode** is the price that appears most often in our list of house prices. For our dataset, the initial mode we found was \$0. This isn't very helpful because a house price of \$0 doesn't make sense.

To get a better understanding of what a common house price might be, we need to ignore these \$0 listings. When we do that, we find that the most frequently occurring house price is \$300,000. This tells us that \$300,000 is a common price for houses in this data.

Knowing the mode is useful because it shows us a price that many homes are sold at. This helps buyers and sellers see a popular price point. If a lot of homes are sold for \$300,000, buyers might see many options at this price, and sellers might consider this a typical price to aim for if their home is similar to others in the area.

By focusing on this corrected mode, we get a better picture of the housing market. It avoids the confusion that zero prices cause and provides more accurate information for anyone interested in the real estate market. This kind of detail is crucial for making informed decisions about buying or selling property.

Figure 2

B4147					
=MODE(IF(B2:B4141<>0,B2:B4141))					
	A	B	C	D	E
4119	03/07/2014	287919.7826	2	1	870
4120	05/07/2014	672500	4	2.25	2110
4121	06/07/2014	454790	2	1.75	1990
4122	07/07/2014	282508.8889	3	1	1560
4123	07/07/2014	473200	3	1.5	1740
4124	07/07/2014	406062.5	2	1	1290
4125	07/07/2014	282766.6667	4	2	1700
4126	07/07/2014	486445.8333	3	1.75	1880
4127	07/07/2014	486895	3	1	1890
4128	07/07/2014	430277.7778	3	2.25	1620
4129	08/07/2014	229629.5	2	1	770
4130	08/07/2014	0	4	2.25	2890
4131	08/07/2014	182805	3	1	1040
4132	08/07/2014	380680.5556	4	2.5	2620
4133	08/07/2014	396166.6667	3	1.75	1880
4134	08/07/2014	252980	4	2.5	2530
4135	08/07/2014	289373.3077	3	2.5	2538
4136	09/07/2014	210614.2857	3	2.5	1610
4137	09/07/2014	308166.6667	3	1.75	1510
4138	09/07/2014	534333.3333	3	2.5	1460
4139	09/07/2014	416904.1667	3	2.5	3010
4140	10/07/2014	203400	4	2	2090
4141	10/07/2014	220600	3	2.5	1490
4142	MEANIF	559687.1943			
4143	1.mean	553064.477			
4144	2.median	460000			
4145	2.Median Excl 0	464600			
4146	3.Mode	0			
4147	3 Mode Excl 0	300000			

The **median** is a straightforward statistic that helps us understand the central point of a data set. In our analysis of housing prices, the median gives us the price at the exact middle of all the listed prices. This means that half of the homes are priced below this point, and the other half are priced above it. For the houses in our dataset, the median price calculated was \$460,000.

However, this figure includes homes listed at \$0, which are likely to be data errors or non-sale listings. These \$0 values can skew the median downwards because they are treated as actual sales in the calculation, which they are not. To get a more accurate reflection of the housing market, we need to adjust our approach by ignoring these \$0 listings.

When we exclude the \$0 prices and recalculate, we find that the median price of homes adjusts slightly upward to \$464,600. This adjustment is significant. It provides a more accurate representation of the market by only considering actual sales. This adjusted median price tells us that more than half of the homes cost more than the initially calculated \$460,000, reflecting a higher market valuation.

Understanding the median price is very useful for potential homebuyers or investors. It helps them gauge the market better a house priced below the median might be considered a bargain, while one priced above might be seen as relatively expensive. Therefore, knowing the accurate median price helps in making informed decisions. (Cottage & Castle Advocates. 2024)

Including \$0 listings can lead to underestimations of both average and median prices, leading potential buyers to expect lower prices than the market truly offers. This could impact their budgeting and market entry strategies.

Figure 3

B4145 {=MEDIAN(IF(B2:B4141<>0, B2:B4141))}					
	A	B	C	D	E
4125	07/07/2014	282766.6667	4	2	1700
4126	07/07/2014	486445.8333	3	1.75	1880
4127	07/07/2014	486895	3	1	1890
4128	07/07/2014	430277.7778	3	2.25	1620
4129	08/07/2014	229629.5	2	1	770
4130	08/07/2014	0	4	2.25	2890
4131	08/07/2014	182805	3	1	1040
4132	08/07/2014	380680.5556	4	2.5	2620
4133	08/07/2014	396166.6667	3	1.75	1880
4134	08/07/2014	252980	4	2.5	2530
4135	08/07/2014	289373.3077	3	2.5	2538
4136	09/07/2014	210614.2857	3	2.5	1610
4137	09/07/2014	308166.6667	3	1.75	1510
4138	09/07/2014	534333.3333	3	2.5	1460
4139	09/07/2014	416904.1667	3	2.5	3010
4140	10/07/2014	203400	4	2	2090
4141	10/07/2014	220600	3	2.5	1490
4142	MEANIF	559687.1943			
4143	1.mean	553064.477			
4144	2.median	460000			
4145	2.Median Excl 0	464600			

The **standard deviation** tells us how much the numbers in a list are spread out from the average. For the bedrooms in the houses we looked at, the standard deviation is 0.9. This is a small number, which means that the number of bedrooms in most houses doesn't change much from the average. (Zach 2021)

A low standard deviation, like the 0.9 we found, shows that most houses have about the same number of bedrooms. This helps people understand that most homes in this dataset are quite similar in terms of how many bedrooms they have. They aren't very different from each other.

For example, if the average (mean) number of bedrooms in the houses is 3, a standard deviation of 0.9 tells us that most of the houses have bedrooms close to this number either a little more or a little less. Someone looking for a house can expect to find many homes with 2, 3, or 4 bedrooms, but not much more or less than that. It is shown in figure 4.

This number helps anyone interested in the housing market to get a quick idea of what most homes are like. It shows that there isn't a big difference in the number of bedrooms from one house to another.

Figure 4

C4148			=STDEV(C2:C4141)	
	A	B	C	D
4132	08/07/2014	380680.5556	4	2.5
4133	08/07/2014	396166.6667	3	1.75
4134	08/07/2014	252980	4	2.5
4135	08/07/2014	289373.3077	3	2.5
4136	09/07/2014	210614.2857	3	2.5
4137	09/07/2014	308166.6667	3	1.75
4138	09/07/2014	534333.3333	3	2.5
4139	09/07/2014	416904.1667	3	2.5
4140	10/07/2014	203400	4	2
4141	10/07/2014	220600	3	2.5
4142	MEANIF	559687.1943		
4143	1.mean	553064.477		
4144	2.median	460000		
4145	2.Median Excl 0	464600		
4146	3.Mode	0		
4147	3 Mode Excl 0	300000		
4148	4.Standard deviation		0.90	

Variance is a way to see how much the numbers in a group differ from the average (mean) of that group. It does this by looking at how far each number in the group is from the average and then finding the average of these differences squared.

In our data on house prices and number of bedrooms, variance helps us see how much these numbers change. For example, the variance of house prices in our data is very large, at 340,689,874,534. This means house prices are spread out a lot from the average house price, which is about \$553,063. Some houses cost a lot more, and some cost a lot less than this average.

On the other hand, the variance for the number of bedrooms is much smaller, just 0.817. Most houses have around these many bedrooms, with not much difference. (Master's in Data Science. n.d.)

So, while house prices can be very different from one house to another, the number of bedrooms is more consistent. This helps people understand that while they might see big differences in prices when they look for a house, they won't see as big a difference in the number of bedrooms. This is shown in figure 5. This makes it easier to predict what kind of house they might get in terms of bedrooms, but harder to guess the price.

Figure 5

C4149			=VARP(C2:C4141)	
	A	B	C	
4140	10/07/2014	203400	4	
4141	10/07/2014	220600	3	
4142	MEANIF	559687.1943		
4143	1.mean	553064.477	3.400483	
4144	2.median	460000		
4145	2.Median Excl 0	464600		
4146	3.Mode	0		
4147	3 Mode Excl 0	300000		
4148	4.Standard devia	583686.45	0.90	
4149	5. Variances	3.40608E+11	0.816908	

Correlations/Graphs

To help us understand the data better, we've made some graphs using Excel. These graphs show how much the houses cost and how many bedrooms they have.

Figure 6 shows all the house prices we looked at. It shows each house's price on a vertical line, with the lowest prices starting at \$7,800 and the highest reaching up to \$26,590,000. However, most of the house prices fall between \$7,800 and \$500,000. This range is where you'll find many of the houses being sold.

Figure 6

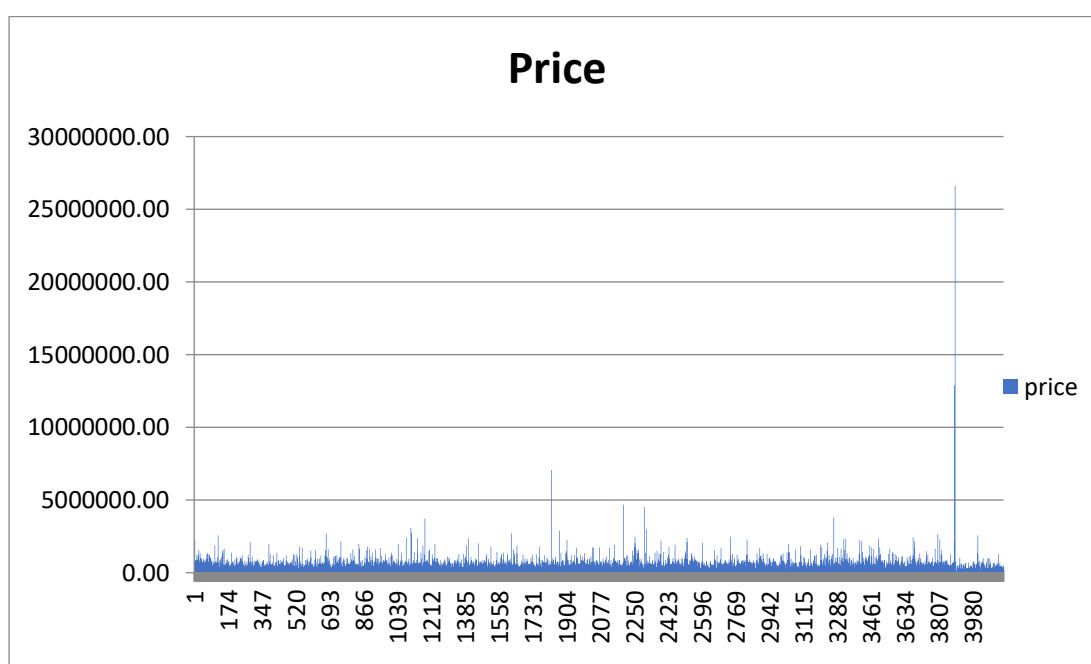
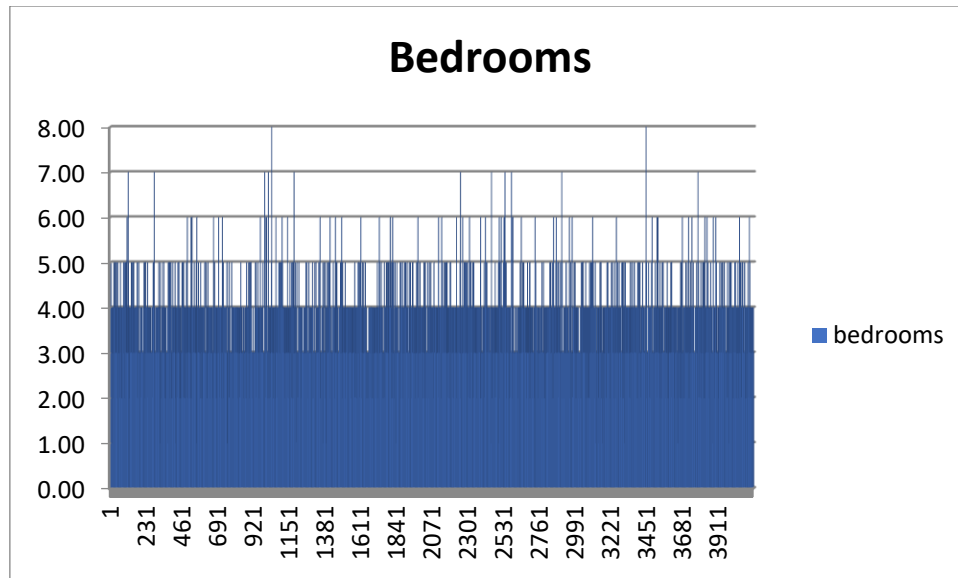


Figure 7 shows how many bedrooms each of these houses has. In this graph, the number of bedrooms a house has is shown on a vertical line. This helps us see the common number of bedrooms that most houses have, just like how the first graph showed us about the prices.

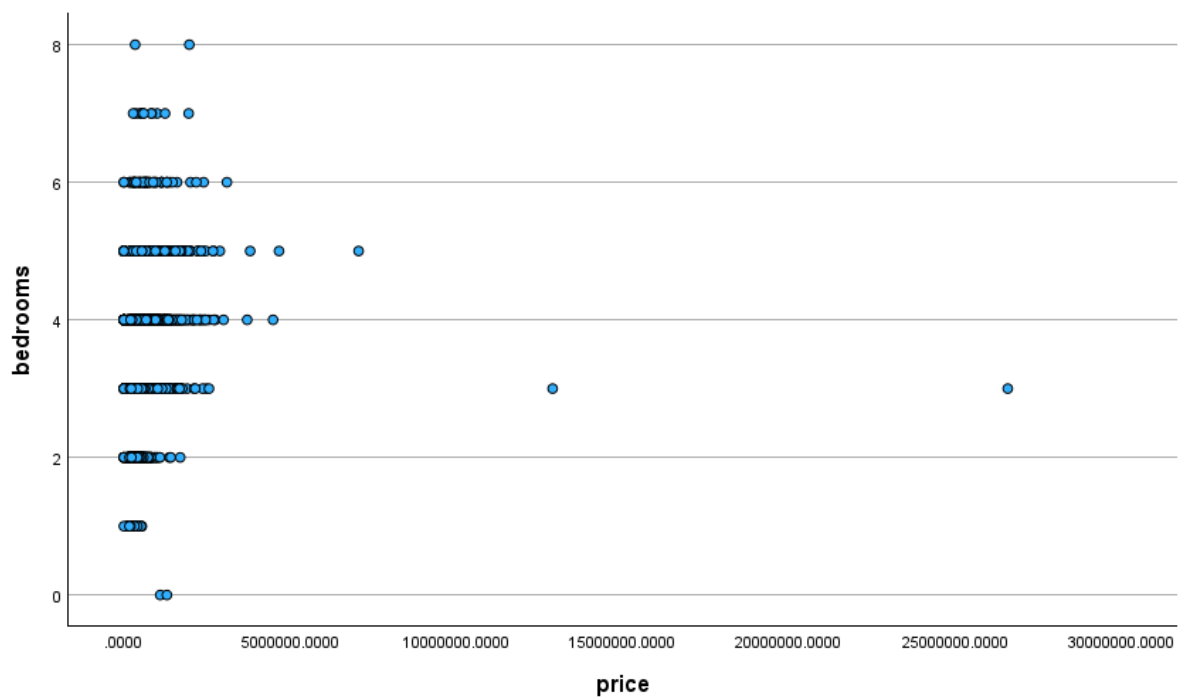
By looking at these graphs, we can also see if there's a link between how much a house costs and how many bedrooms it has. We find out more expensive houses tend to have more bedrooms. This kind of information is useful for people thinking about buying a house because it helps them understand what they can expect to find within their budget.

Figure 7



A scatter plot is a type of chart that we use to show how two variables relate to each other. In this chart, each dot represents different pieces of data. Where the dot is placed on the plot shows us the values of the two things we are comparing. For our example, we are looking at house prices compared to the number of bedrooms they have.

Figure 8



This chart is useful because it lets us see patterns between the price of a house and how many bedrooms it has. For instance, we can see if more expensive houses tend to have more bedrooms. Each dot's position on the horizontal line (left to right) shows us the price, and its position on the vertical line (up and down) shows us the number of bedrooms.

From figure 8 , we can see that most of the houses priced up to \$500,000 have between 3 and 4 bedrooms. This gives us a quick way to see that houses in this price range usually have about that many bedrooms.

However, this kind of chart doesn't tell us about other important things like where the house is located or how big it is. If we wanted to understand the whole picture, we would need to look at these things too. This deeper look could help us understand why some houses cost and how size or location might affect the price.

Section 2

Advantages of Using Excel

Excel is known for its robust functionality in data analysis, making it a great tool for managing and analysing large datasets, such as the "USA Housing Dataset." Its integration with the Microsoft Office suite enhances accessibility, as many users are already familiar with it. This familiarity allows for efficient navigation and manipulation of data, including performing calculations like sums, averages, and more complex statistical functions directly within the software. (Bala_Excel_Insights 2024)

One of Excel's key strengths lies in its powerful data visualisation capabilities. Users can easily create a variety of charts, graphs and tables to visually represent data, making it easier to identify trends and patterns quickly. For example, generating scatter plots to explore the relationship between house prices and the number of bedrooms can be accomplished with just a few clicks, simplifying the process of data interpretation.

Also Excel supports a wide array of functions, such as IF, mean, mode, and median, which are logically organised and easily accessible. This setup streamlines the analytical process and increases productivity by reducing the time spent on data processing. The ability to integrate smoothly with other Microsoft products like Word makes for seamless transfer of data and results, further enhancing its utility in professional settings.

Disadvantages of Using Excel

While Excel is a powerful tool for data analysis, it does have several disadvantages that can impact its efficiency and reliability, particularly in more complex situations. These limitations can be significant depending on the nature of the data and the requirements of the analysis. (www.sigmacomputing.com. n.d.).

Excel is prone to human error, especially when dealing with large datasets or complex calculations. Manual data entry and formula setup can lead to mistakes such as mis-typed figures or incorrect formula references, which can skew analysis results significantly. Errors in applying conditions or

excluding certain data points (like \$0 house prices) could lead to incorrect interpretations of trends in the housing market.

While Excel is capable of handling datasets with up to a million rows, it begins to slow down as it reaches its capacity. This can be annoying for users working with very large datasets, as performance issues may lead to inefficiencies in processing and analysing data. The slowdown can affect productivity, especially when performing operations that require recalculating large portions of data, such as sorting, applying filters or updating formulas.

Another limitation is Excel's basic set of statistical functions, which might not be good for advanced statistical analyses. For deeper data analysis, such as complex models or time series analysis, users often need to rely on additional software or plugins. This requirement can complicate the analysis process, as it might involve transferring data between different platforms, increasing the risk of data corruption or loss during transfers.

Excel's compatibility is limited to Windows and Mac OS X, excluding users on other platforms such as Linux. This can be a barrier in diverse computing environments, requiring alternative solutions or additional virtualisation software to run Excel, further complicating its use. (Hiron-Grimes, B. 2022)

In summary, while Excel offers great utility and ease of use for most data processing and analysis tasks, its disadvantages in performance with large datasets, advanced capabilities, security and cross-platform compatibility must be carefully considered, particularly in professional and data-intensive environments.

Data Mining Cycle

The Data Mining Cycle is a critical process for analysing large volumes of data to uncover patterns and actionable insights. This process is important across various fields such as business, healthcare, and science, where it helps decision-making based on data analysis rather than assumptions. (Twin, A. 2023)

Understanding the Data Mining Cycle

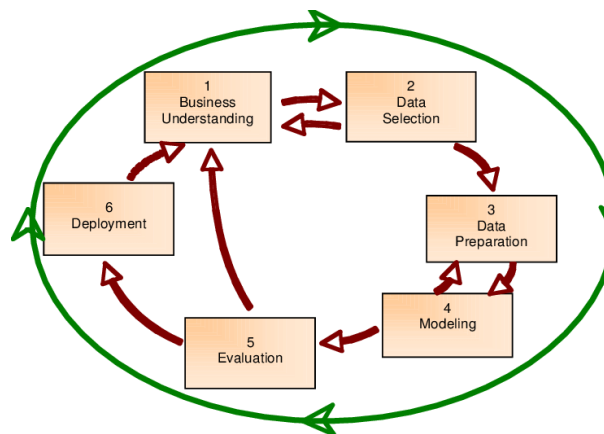
The Data Mining Cycle starts by identifying a specific problem that needs solving. This could be anything from a company wanting to know which products sell best to improving their marketing tactics. Once the issue is clear, the next step is to gather the needed data, like sales records or customer profiles.

Next we prepare the data for analysis. Often, the data we collect is raw and unorganised, so we need to clean and organize it. This includes removing errors, filling in missing information and making sure the data is in a format we can use.

Once the data is ready it is analysed. We use math and statistics to look through the data and find patterns or connections. Tools like Excel or more advanced software help us do calculations, build models and test our ideas.

After analysing the data, we need to understand what the results mean. This step turns our findings into clear insights that can guide decision-making. For example, if we find out that lowering prices increases sales, a company might start offering discounts more often.

The last step of the Data Mining Cycle is to use these insights. This means applying what we've learned to make smart decisions or solve the problem we started with. This could involve changing how things are done, improving products, or updating policies. The goal here is to act based on what the data tells us.



Importance of the Data Mining Cycle

The Data Mining Cycle is important because it transforms raw data into valuable insights that help make smarter, fact-based decisions. Without this organised way of analysing data, organisations could make decisions based on wrong assumptions, leading to poor practices and missed chances. (Chartexpo.com. 2024)

For example, a real estate company might use the Data Mining Cycle to find out what affects house prices. First, they define what they need to solve figuring out what influence's prices. Next, they gather and prepare important data like sales records, house features and information about the area. After analysing this data, they find out which factors are most important for setting prices. These insights help sellers price their homes competitively and help buyers make smart choices when they purchase homes.

Conclusion

In conclusion, our analysis of the "USA Housing Dataset" has highlighted key characteristics and trends in the US housing market from May 9, 2014, to July 10, 2014. By looking at statistical measures such as the average, median, mode and variance of house prices and bedroom numbers, we have gained insights into the housing dynamics during this period.

Our findings show that the corrected average house price was approximately \$553,063, and the median, adjusted to remove \$0 listings, was about \$464,600, offering a more accurate view of the

market's pricing landscape. The most common house price, once adjusted, was around \$300,000, indicating a significant price point for potential buyers and sellers.

Overall, this analysis helps us understand market behaviours and supports better decision-making in the real estate sector. As data and analysis methods improve, the value of such insights will continue to grow, influencing future housing market strategies and policies.

References

- Zhao, C. and Liu, F. (2023). Impact of housing policies on the real estate market - systematic literature review. *Heliyon*, [online] 9(10). doi:<https://doi.org/10.1016/j.heliyon.2023.e20704>
- Team, L. (2020). *6 Ways to Make Your Data Analysis More Reliable*. [online] Leadspace. Available at: <https://www.leadspace.com/blog/ways-to-make-your-data-analysis-more-reliable/>.
- Cottage & Castle Advocates. (2024). *Understanding The Median House Price*. [online] Available at: <https://www.cottageandcastle.com.au/home-buyers/median-house-price/>
- Zach (2021). *What is Considered a Low Standard Deviation?* [online] Statology. Available at: <https://www.statology.org/what-is-a-low-standard-deviation/>.
- Master's in Data Science. (n.d.). *What Is the Difference Between Bias and Variance?* [online] Available at: <https://www.mastersindatascience.org/learning/difference-between-bias-and-variance/>.
- Bala_Excel_Insights (2024). *Excel in Research: Data Management and Analysis - Bala_Excel_Insights - Medium*. [online] Medium. Available at: <https://medium.com/@kbala7092/excel-in-research-data-management-and-analysis-2a50d43b3263>.
- www.sigmacomputing.com. (n.d.). *5 Disadvantages of Excel for Data Visualization | Sigma Computing*. [online] Available at: <https://www.sigmacomputing.com/blog/5-excel-disadvantages-for-data-visualization>.
- Hiron-Grimes, B. (2022). *Excel for Mac vs Excel for Windows*. [online] The Excel Experts. Available at: <https://www.theexcelexperts.com/excel-for-mac-vs-excel-for-windows/>.
- Twin, A. (2023). *Data Mining: How Companies Use Data to Find Useful Patterns and Trends*. [online] Investopedia. Available at: <https://www.investopedia.com/terms/d/datamining.asp>.
- Chartexpo.com. (2024). *Data Mining: How to Extract Valuable Insights from Your Data?* [online] Available at: <https://chartexpo.com/blog/data-mining>