Group Details: CRN:

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **Name:**  **Name:**  **Name:** |  | **ID:**  **ID:**  **ID:** |
|  |  |  |

**Instructions:**

1. You must submit two separate copies **(one Word file and one PDF file)** using the Assignment Template on Blackboard via the allocated folder. These files **must not be in compressed format**.
2. It is your responsibility to check and make sure that you have uploaded both the correct files.
3. Zero mark will be given if you try to bypass the SafeAssign (e.g., misspell words, remove spaces between words, hide characters, use different character sets, convert text into image or languages other than English or any kind of manipulation).
4. Email submission will not be accepted.
5. You are advised to make your work clear and well-presented. This includes filling your information on the cover page.
6. You must use this template, failing which will result in zero mark.
7. You MUST show all your work, and text must not be converted into an image, unless specified otherwise by the question.
8. Late submission will result in ZERO mark.
9. The work should be your own, copying from students or other resources will result in ZERO mark.
10. Use **Times New Roman** font for all your answers.

Project

Deadline: Tuesday 22/04/2025 @ 23:59

**[Total Mark for this Project is 14]**

***IT475***

***Decision Support Systems***

# Project

Students can form groups consisting of three students and send their names to their instructors before 20th February 2025. Otherwise, the instructors will form the groups randomly and assign any datasets to the groups.

Select one dataset from the datasets provided in the bellow link.

**20 Data Analysis Projects to Boost Your Skills in 2025:**

<https://www.springboard.com/blog/data-analytics/data-analysis-projects/>

For more free public datasets for EDA:

<https://www.tableau.com/learn/articles/free-public-data-sets>

* After the dataset is selected (or assigned), analyze the data using Microsoft Excel to discover the structure of data, trends, patterns, or any anomalies in the data based on your hypothesis.
* Perform the following six tasks.
* You should use visualization to aid your answers.

Your project will include two main parts:

1. The final project report must incorporate all the following 6 tasks and be written using the provided template. (10 marks distributed among the below tasks).
2. A presentation that illustrates your 6 tasks completed in the project. (4 marks)

==========================================================

|  |
| --- |
| **T Task 1:** Understand and describe the nature and structure of the selected dataset. (2 marks)  Describe the dataset. Your description should answer the following questions: is it reliable? how was it collected? What is its size? |

Airbnb, Inc. is an American company, which was founded in 2008. Airbnb operates an online marketplace for short- and long-term home stays and experiences. This dataset describes the latest listing activity in New York City, New York as of January 5th, 2024. It is published originally by the company itself so it is authentic and reliable.

This dataset contains all the information needed about the New York Airbnb listings. This dataset is taken from [Kaggle](https://www.kaggle.com/datasets/vrindakallu/new-york-dataset) which is modified version of original published on website by removing null values. Its size is 4.44 MB and contains 20759 records.

|  |
| --- |
| * Identify the features of the dataset. |

|  |
| --- |
| The dataset consists of following fields and datatypes of columns: |

|  |
| --- |
| * Propose hypothesis/assumptions (between 2 numerical variables) to validate. |

**Pricing-Related Hypotheses:**

**H1:** The number of bedrooms, beds, and baths positively correlates with the price.

**Demand & Availability Hypotheses**

**H2:** Listings with more reviews have higher availability.

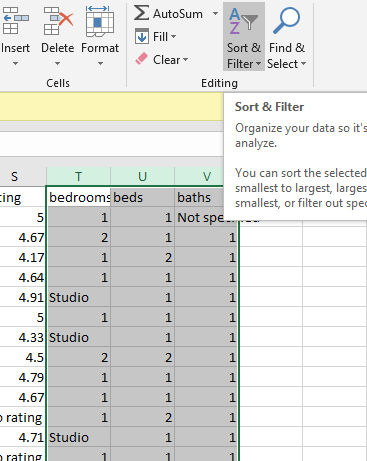
|  |
| --- |
| **Task 2:** Check if your selected features have any of the following issues. Describe how you conducted the tests and how you addressed the issues. Support your answers with screenshots of the issues before and after the fixes. (1 mark)   * Missing Values * *Duplicate values* * Data outliers * Any Noise |

* ***Missing values***

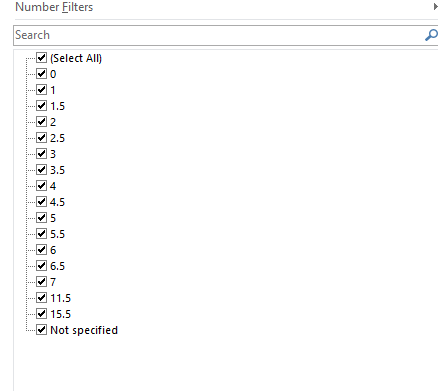
For missing values, we checked the count of blank in each column of excel and found none as is specified by the details given in dataset. The dataset was already processed, cleaned of null values and blanks. Our analysis is also reflective of this fact.

|  |
| --- |
|  |

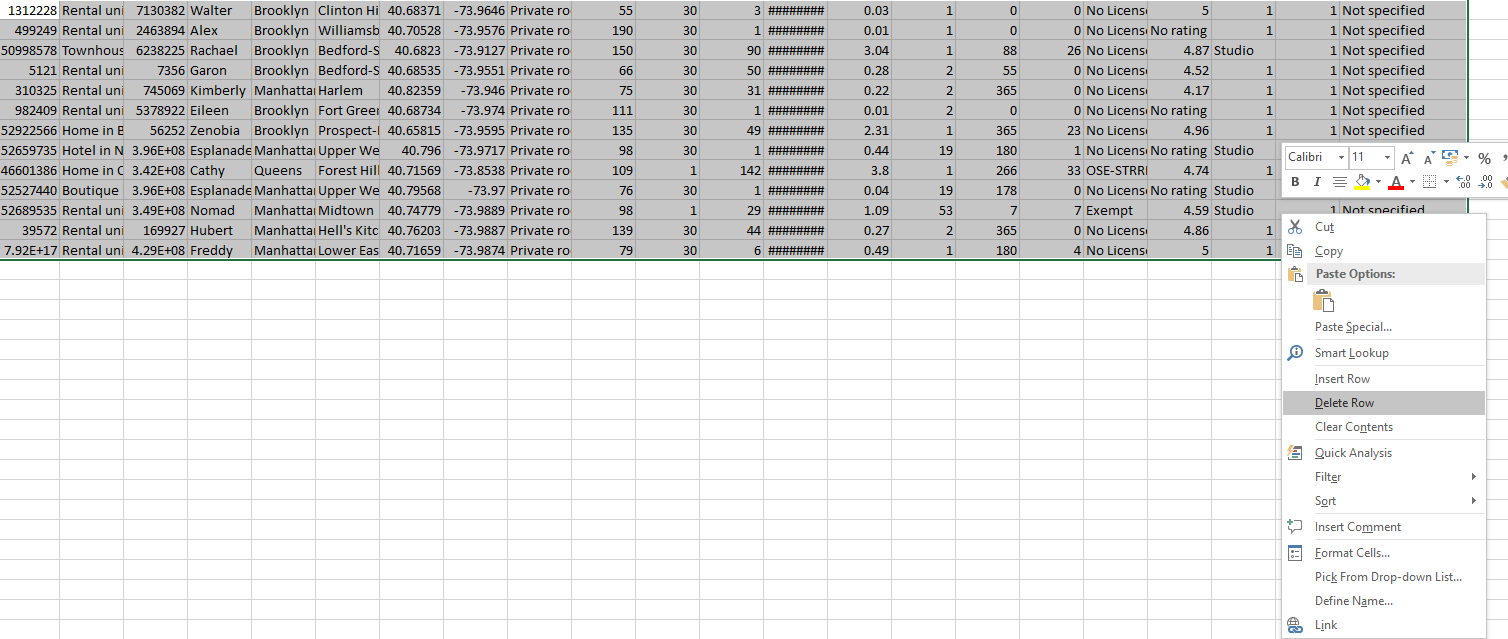
To handle other types of missing data, we applied filters on our relevant columns including baths, beds, number of reviews and availability.



We only have some missing values in baths column that are mentioned as not specified. These are just 13 records. So, we can remove these records.



We filtered and selected all the values that are not specified and then delete all these rows from the excel dataset.



***Duplicate values***

To find the duplicate records in the dataset, we use function of remove duplicates in datasheet.

|  |
| --- |
|  |
|  |
|  |

Here, 12 exact duplicates record were found and removed.

***Data outliers***

Data values deviating from majority of the samples are known as data outliers. Before we fix data outliers we need to identify them. As per our set of hypotheses, number of reviews, rating, prices, listing, no\_of\_bedrooms, beds, baths and availability are the variables that we need to validate.

To detect data outliers, we will plot the box and whisker chart to identify the outliers:

|  |
| --- |
|  |

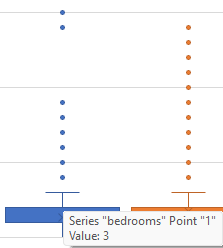
For detection of outliers in other variables, we plotted several box and whiskers plot.

|  |  |
| --- | --- |
|  | |
|  |  |

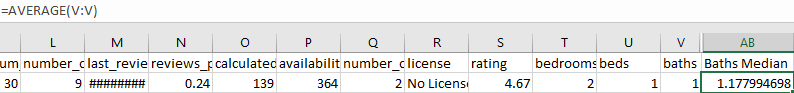
***Fixing Data Outliers:***

Analyzing these graphs, we know that there are some outliers in no of bedrooms, beds and baths, one outlier in prices and several in number of reviews variables.

From graphs we can detect the outlier’s points are above a certain value by analyzing the ending range of normal values, for instance, in case of bedrooms, values above than 3 are marked as outliers.



We can fix the outliers by replacing these values by the mean of the column. So, we calculated the mean using average function in excel.



The using the if statement formula to replace the value with mean if they are above the upper limit of normal values range. For instance, in case of baths, if statement is written as following:



We applied the same technique for all the relevant columns. After removing the outlier’s

box and whisker plots look like:

|  |  |
| --- | --- |
|  |  |
|  |

***Any noise or irregularities***

Further to check if any column contains NA or null values. I have applied filters on each column and check the range of values that can be found in each of these columns:

|  |
| --- |
|  |

Most of the column contains no N/A null or blank values. Some are shown below:

|  |  |
| --- | --- |
|  |  |

Following are the three columns that have some missing values:

|  |  |  |
| --- | --- | --- |
|  |  |  |

For no license there are total 17000+ records, so we neither eliminate them nor replace it with something else.

There were 13 records in bath column which are not specified and removed in earlier steps.

|  |
| --- |
| **Task 3:** Provide descriptive statistics for the selected features using statistical methods to understand the dataset more and answer the following analysis questions:   * Include any of the measures of central tendency such as the mean, median, and mode. * Describe the spread of your data. This may include the measure of variance, standard deviation, skewness, and kurtosis. |

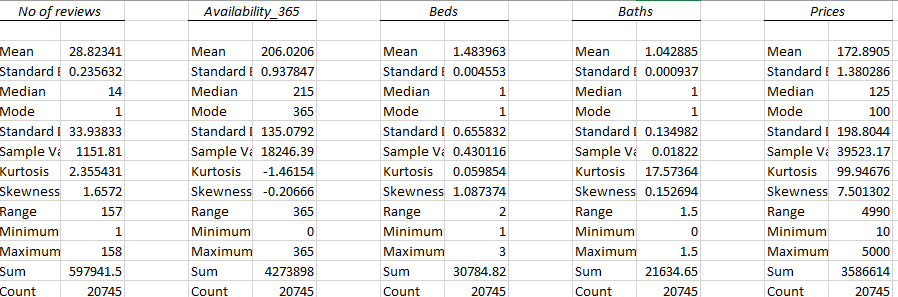
We can analyze the data using descriptive statistics option of excel. We click the data analysis option from data tab.



From data analysis we choose the descriptive statistics options and entered the column range as an input parameter and enabled the option of summary statistics

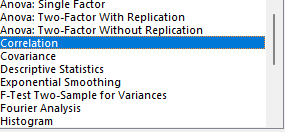
|  |  |
| --- | --- |
|  |  |

After doing summary statistics on each of our relevant column we reached with the following statistical measures. The summary gives us central tendency as well as spread of the data.

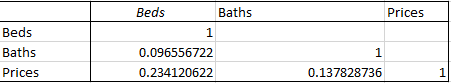


|  |
| --- |
| **Task 4:** Validate the hypothesis in Task 1 by investigating the relationship between two quantitative variables you have chosen using correlation, regression, and R-squared with possible conclusions. (2 marks) |

For hypothesis in Task 1, we calculated the correlation between no\_of\_beds, no\_of baths and prices. To calculate the correlation, we used the data analysis option in excel and this time we choosed the correlation option from data analysis tab



For hypothesis 1, we gave both no\_of\_beds and no\_of baths as x axis variables and price as y\_axis.

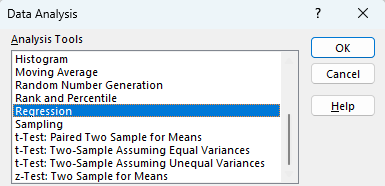


From correlation values we can deduce the no\_of\_beds and no\_of\_baths have positive correlation with each other. Further both have positive correlation with price column as well.

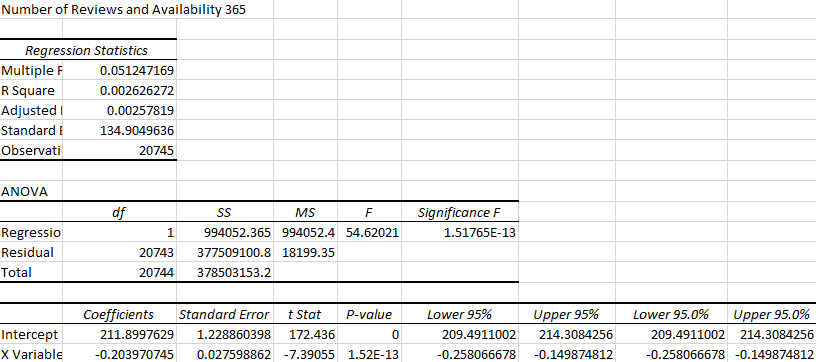
For hypothesis 2, we calculated the correlation between no\_of\_reviews and availability\_365. From correlation value, it is deduced that availability\_365 is negatively correlated with number of reviews.

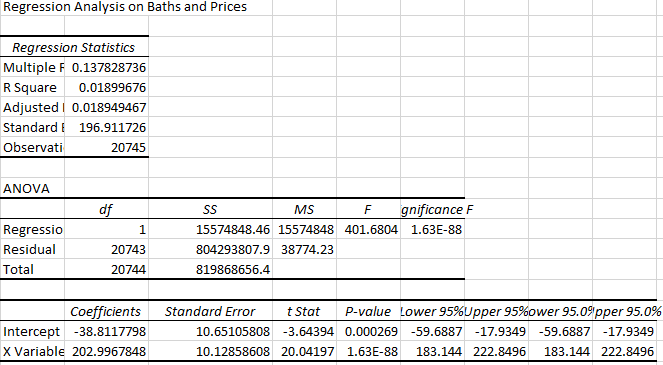
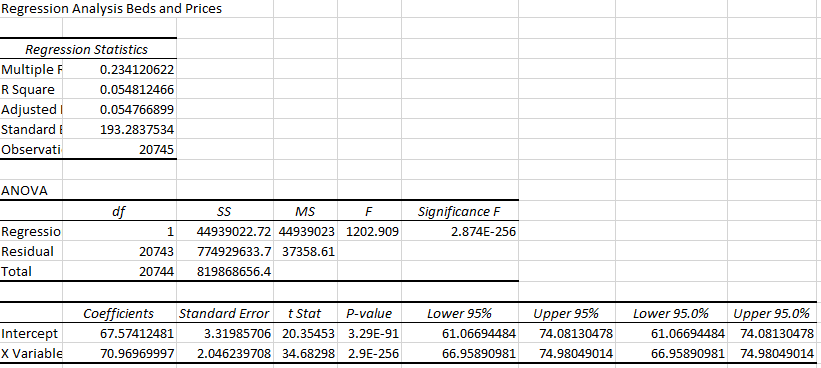


To conduct further analysis, we used the regression analysis option of data analysis tab using excel.



Performing regression of beds and prices, baths and prices and availability and number of reviews, following are the results





**Conclusion:**

* + From regression analysis, more reviews slightly reduce availability, possibly indicating higher competition. The coefficient indicates that there is a negative relationship between these two variables.
  + However, there’s a positive and significant relationship between the number of beds and prices. From R square value we deduced that about 5.5% of the variation in prices is explained by the number of beds and coefficient value 70.97 reflects each additional bed increases the price by approximately $71.
  + For baths and price relationship R Square value is 0.01899 showing only 1.9% of price variation is explained by number of baths. The coefficient value 202.99 indicates that each additional bath increases the price by approximately $20

|  |
| --- |
| **Task 5:** Show a visual representation of your analysis (hint: use the right chart/graph for your data analysis). (1 mark) |

We can visualize the analysis using two charts. Scatter chart and Linear chart can show association between Census and Time in the selected dataset.

**Scatter chart**

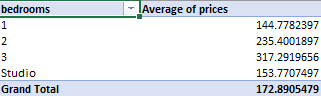
A scatter chart is used to reveal trends between two variables. We can draw a scatter chart by:

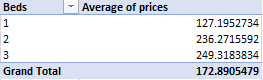
Insert >> Scatter chart >> OK

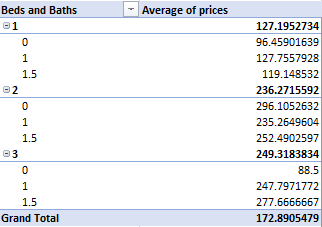
|  |
| --- |
| There exists a directly proportional relationship showing that when no of beds increases the price of listing also increased. |
| There exists a directly proportional relationship between baths and prices as well showing that when no of bath increases the price of listing also increased. |
| Because of 20000+ records, it is difficult to visualize however the trend line indicates that when the number of reviews increases for a listing its availability decreases |

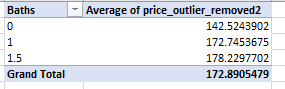
|  |
| --- |
| **Task 6:** Build an active Dashboard that summarizes the most crucial factors (variables) that will help in the decision-making process, and then demonstrate the effectiveness of your selection of those factors in the decision-making process. (2 marks) |

Forbuilding a dashboard, we inserted a pivot chart on selected data sheet. To see the impact of no of bedrooms, no of beds, no of baths and average prices, we plotted several pivot charts.  
Some of those can be seen in following screenshots:



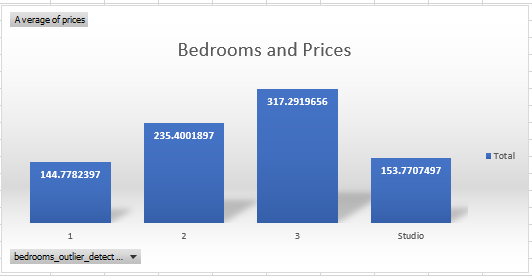




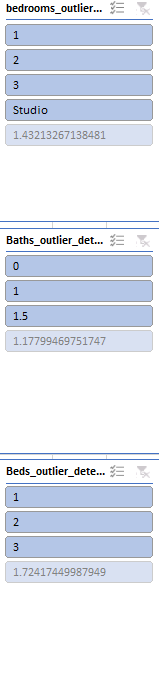


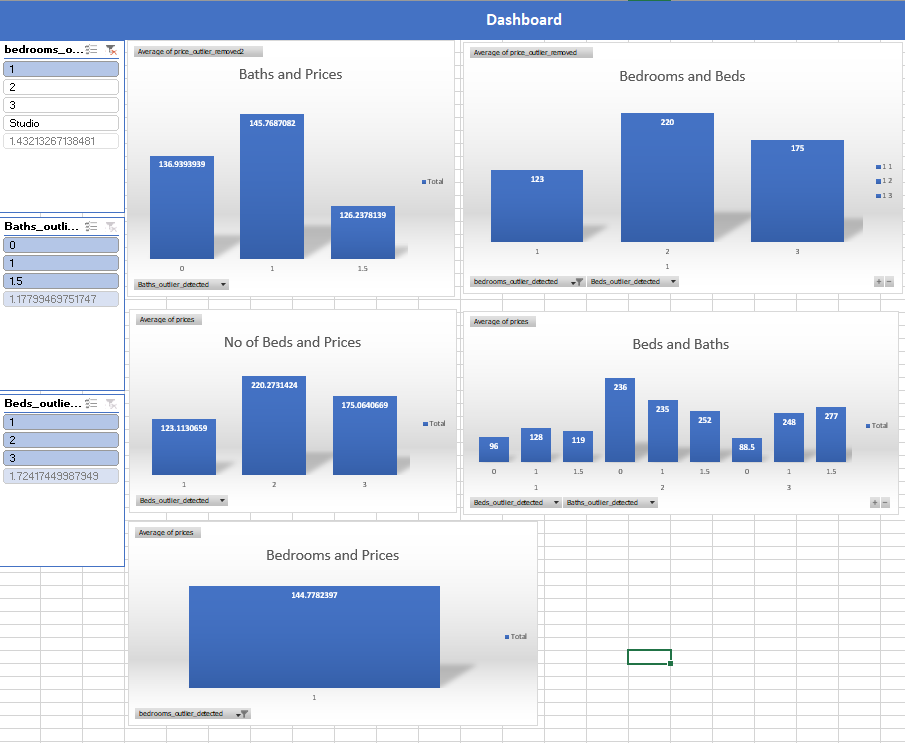
To make the visualization more effected we created pivot chart using bar charts:





To make our dashboard active, we inserted slicers over no of beds, bedrooms and no of baths. Linking these slicers to the relevant charts make charts more dynamic.



To filter out values, we applied filter of no of bedrooms = 1 and all the charts are updated accordingly.  


For availability\_365 and no of reviews we inserted a line chart to see the trend.

