# 

**ANL252**

**Python for Data Analytics**

**Group-Based Assignment**

**July 2023 Presentation**

**Submitted By:**

|  |  |  |
| --- | --- | --- |
| **S.N.** | **Name** | **PI. No** |
| 1 | Ng Ling Ying | E2181019 |
| 2 | Tan Zekai Marcus | H2011558 |
| 3 | Pereira Kathlyn Therese | B2272579 |
| 4 | Goh Qing Feng | Q2181339 |

**Tutorial Group:** TG03

**Submission Date:**  12/10/2023

**Question 1(a)**

To read the dataset in Python, we would have to import the Pandas library and convert the .csv dataset to pandas DataFrame using the read\_csv() function. Pandas is a Python library that allows us to read and analyze datasets. The following code allows us to read the dataset in Python.

import pandas as pd

hdb\_data = pd.read\_csv("GBA\_HDB.csv")

Once the dataset is converted into a DataFrame, which is a 2-dimensional data structure, we can identify its dimensions using the ‘shape’ attribute of the DataFrame. This attribute would return a tuple representing the dimensionality of a particular data frame, which refers to the total number of rows and columns of the dataset. We are able to identify the dimensions by running the following code.

hdb\_data.shape

Output:

(1250, 11)

The output indicates that there are 1,250 rows and 11 columns in the dataset.

[130 Words]

**Question 1(b)**

We can identify variables with missing values using the code below.

import pandas as pd

hdb\_data = pd.read\_csv("GBA\_HDB.csv")

missing\_var = hdb\_data.isnull().any()

print(missing\_var)

This will produce an output where if the variable is indicated ‘True’, it contains missing values:

month False

town False

flat\_type True

block False

street\_name True

storey\_range False

floor\_area\_sqm False

flat\_model False

lease\_commence\_date False

remaining\_lease False

resale\_price True

dtype: bool

If we only want to show the ‘True’ Values, we can use the following code below to filter and retrieve the index.

missing\_var[missing\_var == True].index

Output:

Index(['flat\_type', 'street\_name', 'resale\_price'], dtype='object')

The output indicates that variables ‘flat\_type’, ‘street\_name’, and ‘resale\_price’ contain missing values.

Identifying missing data is important because large amounts can reduce the statistical estimates of an analysis such as the average or median, which leads to biased and skewed conclusions that may have business consequences. It also impacts the quality of data visualization charts, missing values in the dataset can lead to a drop to zero break in a bar/line chart. Also, if machine learning is incorporated, it can’t handle missing values and might throw errors if there are too many. Even if they do somehow manage to interpret the data, it might lead to misleading results. Therefore, the best way to handle missing data is to either eliminate or replace them.

[173 Words]

**Question 1(c)**

Using the fillna() function, missing data can be replaced by mean, which is the measure of central tendency. This method is suitable when the data is a numerical variable. Using the mean will ensure that the imputed values align with the data value.

Another method is using the dropna() function. This method is suitable when the amount of missing data is small compared to the dataset, and removing it will not impact the analysis.

In this case, the missing data under the columns ‘flat\_type’ and ‘street\_name’ are non-numerical data, which suggests that the dropna() method is appropriate.

hdb\_data = hdb\_data.dropna(subset=['flat\_type', 'street\_name'], how = 'any')

On the other hand, the missing data under the column ‘resale\_price’ are numerical data and the number of ‘null’ values is more than 10% of the dataset. The following code is used to generate the count of ‘null’ values.

hdb\_data.isnull().sum()

Output:

month 0

town 0

flat\_type 0

block 0

street\_name 0

storey\_range 0

floor\_area\_sqm 0

flat\_model 0

lease\_commence\_date 0

remaining\_lease 0

resale\_price 133

dtype: int64

The output indicates that there are 133 ‘null’ values in ‘resale\_price’.

hdb\_data.shape

Output:

(1209, 11)

After removing the missing data in ‘flat\_type’ and ‘street\_name’, there are 1,209 rows remaining. As the amount of missing data in ‘resale\_price’ is relatively significant to the dataset, removing them with the dropna() method may be inappropriate. Hence, the missing data will be replaced with the mean using the fillna() method.

hdb\_data['resale\_price'] = hdb\_data['resale\_price'].fillna(hdb\_data['resale\_price'].mean())

(199 Words)

**Question 1(d)**

##Import HDB\_Data Excel to jupyter. DataFrame\_name=pd read\_csv csv\_file\_name.csv

import pandas as pd

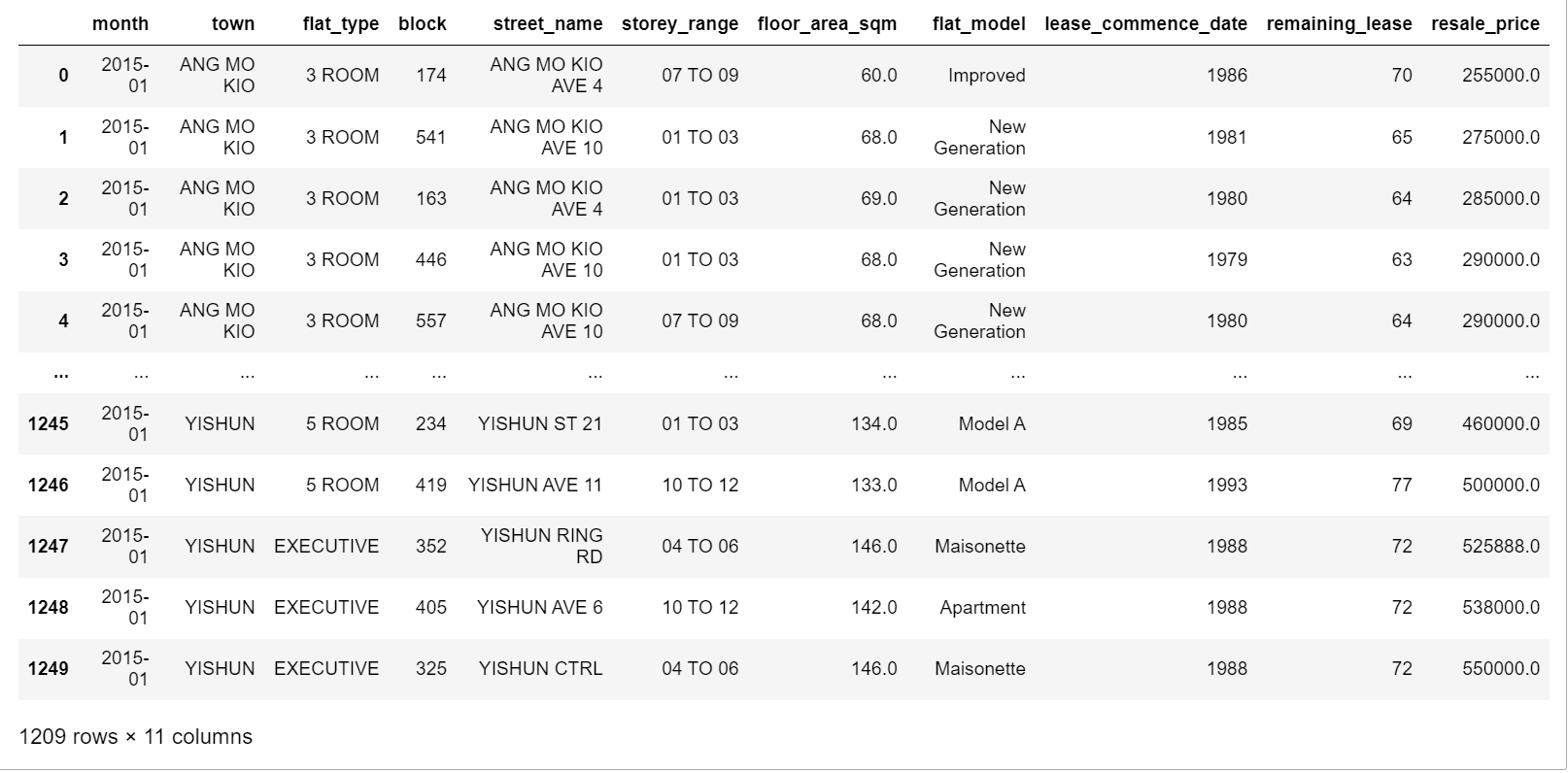
import matplotlib.pyplot as plt

hdb\_data = pd.read\_csv(r"D:\for school\ANL252-Python\HDB\_Data.csv")

hdb\_data = hdb\_data.dropna(subset=['flat\_type', 'street\_name'], how = 'any')

hdb\_data['resale\_price'] = hdb\_data['resale\_price'].fillna(hdb\_data['resale\_price'].mean())

display (hdb\_data)



By displaying the dataset as a table like the above, it provides us with a clear visualization of the dataset and from the information imported, we can also come up with three different graphs that would represent different meanings with different insights as shown below.

##1st Graph

##Histogram Graph-To show the distribution of a variable.

df = pd.DataFrame(hdb\_data)

price = df['resale\_price']

plt.hist(price,bins=10)

plt.ylabel('Frequency')

plt.xlabel('Resale prices')

plt.title('Histogram of Resale Prices')

##Adding the annotations

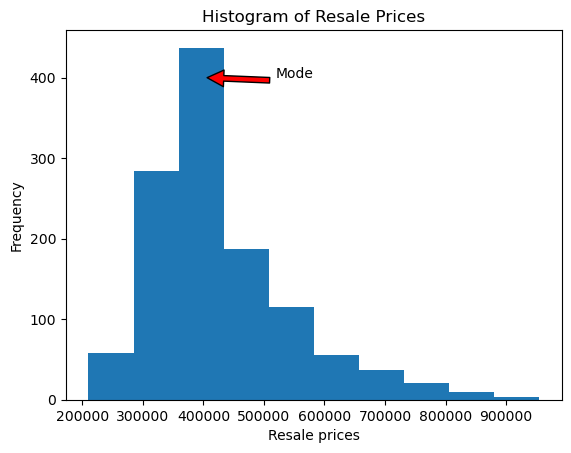
plt.annotate("Mode", xy = (400000, 400),

xytext = (520000, 400),

arrowprops = dict (facecolor = 'red',

shrink = 0.05))

plt.show()



The right-skewed histogram above shows the distribution of the resale prices where we can identify that the mode of the resale prices is around 400 thousand dollars. This shows that most of the houses are resaleable around that price range. The right-skewed histogram indicates that the mean of resale prices is greater than the median due to the houses with significantly higher resale prices at around 800 to 900 thousand dollars. Thus, the median, being more resistant to outliers than the mean, would be the preferred measure of central tendency for resale price.

##2nd Graph

##Bar Chart

##Bar Graph-Comparison

avg\_resale = df.groupby('town')['resale\_price'].mean().reset\_index()

plt.bar(avg\_resale['town'],avg\_resale['resale\_price'])

plt.xticks(rotation=90)

plt.xlabel("Towns")

plt.ylabel("Average Resale Pricing($)")

##Adding the annotations

plt.annotate("$687667", xy = ("MARINE PARADE", 600000),

xytext = ("PUNGGOL", 650000),

arrowprops = dict (facecolor = 'red',

shrink = 0.05))

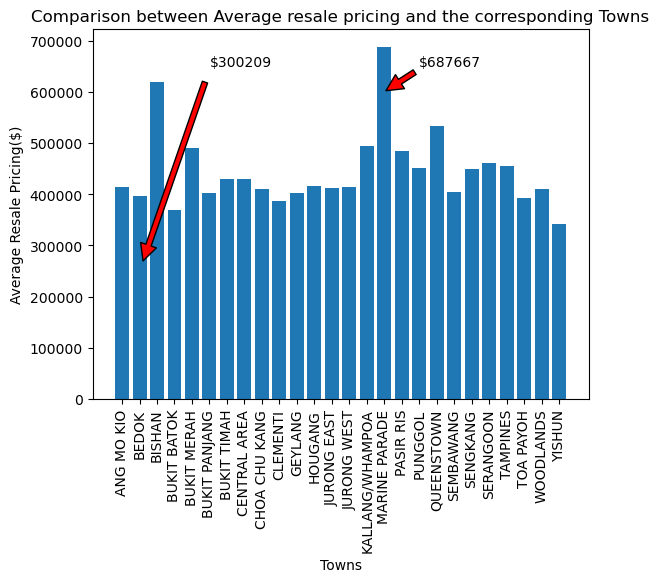
plt.annotate("$300209", xy = ("BEDOK", 250000),

xytext = ("BUKIT PANJANG", 650000),

arrowprops = dict (facecolor = 'red',

shrink = 0.05))

plt.show ()



The bar chart above shows a comparison between the average resale pricing and the different towns. The chart shows that the highest average resale pricing of houses lies in areas such as Marine Parade and Bishan while the lowest average resale pricing belongs to houses in Bedok. Furthermore, most of the prices are around 400 thousand dollars, which is coherent with the histogram above.

##3rd Graph

##Scatter Plot. To study the correlation between two variables.

x = df["resale\_price"]

y = df["floor\_area\_sqm"]

plt.scatter (x,y)

plt.xlabel("Resale Price according to flat area")

plt.ylabel("Floor area (sqm)")

## Adding Annotations

plt.annotate("Around 115 sqm", xy = (370000,115),

xytext = (400000,160),

arrowprops = dict(facecolor = 'red',

shrink = 0.05))

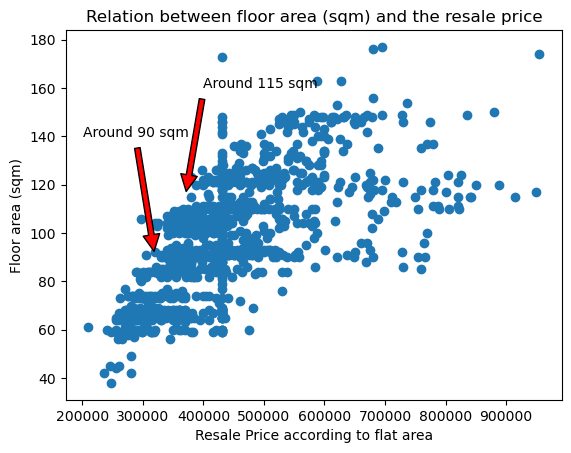
plt.annotate("Around 90 sqm", xy = (320000,90),

xytext = (201000,140),

arrowprops = dict(facecolor = 'red',

shrink = 0.05))

plt.show()



This is a scatter plot which shows the correlation between the floor area per square meter of a house and the resale price (Kumar, 2023). From this scatter plot, we can infer that the larger the floor area per square meter, the higher the resale pricing and we can also infer that there seem to be more resale houses with 90 - 115 square meters of space. To add on, we can infer that while generally the larger the area, the higher the resale price, there are some instances where the houses are priced lower despite the bigger floor area and this could be due to the town they are in. Referring to the bar graph above, these houses may be in towns with lower average resale prices, such as Bedok. Similarly, houses with smaller floor areas but relatively high resale prices may be in towns with higher average resale prices, such as Marine Parade.

[357 Words]

**References:**

Kumar, M. (2023). *Study Unit 3: Arrays and Plots*  [PowerPoint Slides]. Singapore University of Social Sciences.

**Declaration Page**

We, members of **group 8** (*kathlyn001, qfgoh001, lyng007, marcustan015)* do hereby declare that we each contributed to this assignment and that we collectively agree to a shared grade.

|  |  |  |
| --- | --- | --- |
| Name | Contribution | Signature |
| Pereira Kathlyn Therese (Team Lead) | I did D |  |
| Goh Qing Feng | I did A and D |  |
| Tan Zekai, Marcus | I did B |  |
| Ng Ling Ying | I did C |  |