INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Addressing Student Challenges in Finding PGs: An EDA Project with Web Scraping MagicBricks Data

About my Group

My name is MD Tahseen Equbal

I hold a graduate degree
in the stream of Computer
Science and Engineering

My name is C.Y. Santosh Kumar Rao

I hold a graduate degree
in the stream of Computer
Science and Engineering

INNOMATICS
RESEARCH LABS

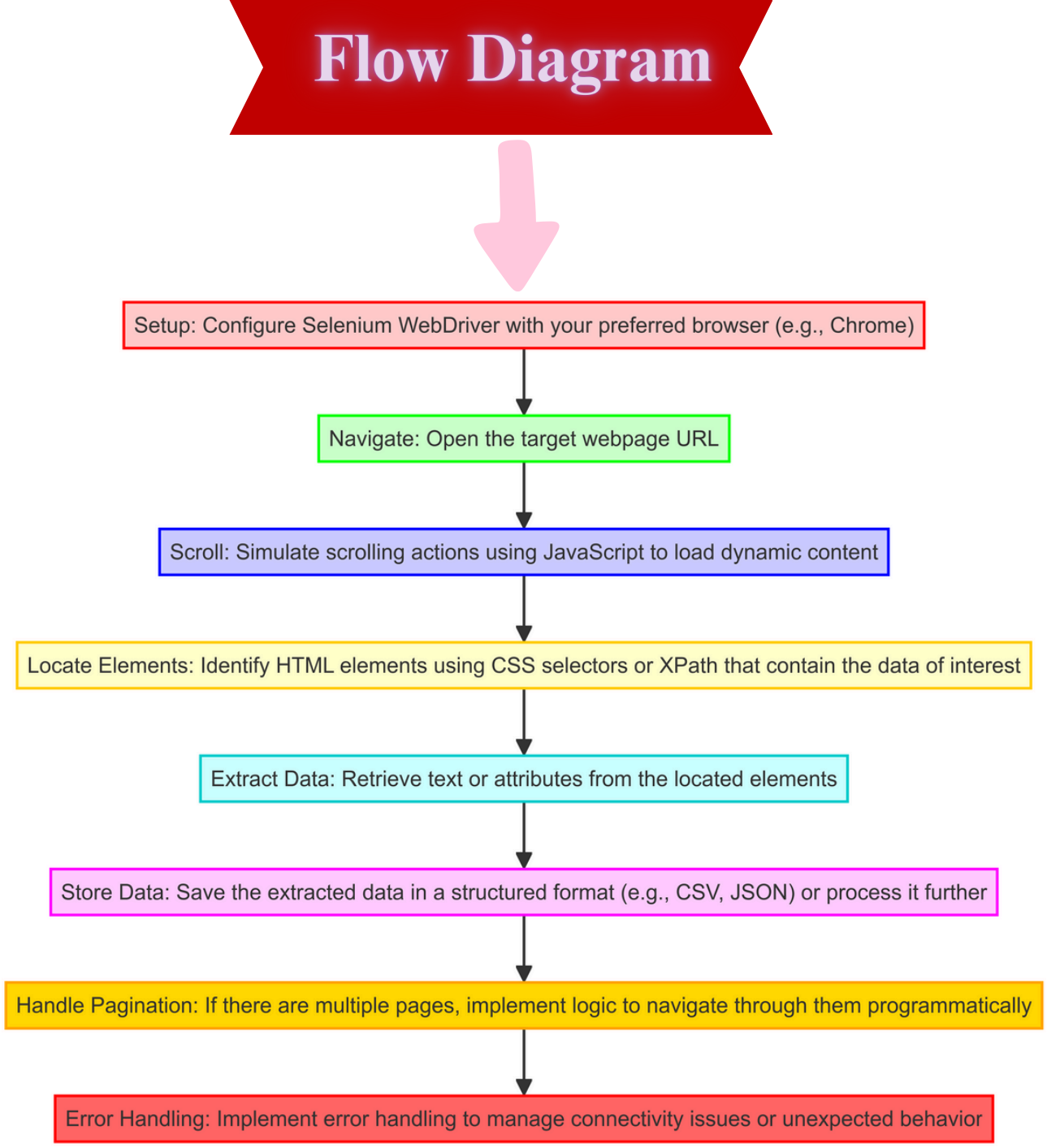
Problem Statement:

Students face numerous challenges when searching for paying guest (PG) accommodations, including lack of transparency in listings, inconsistent rental prices, and inadequate information on amenities. These factors often lead to a time-consuming and frustrating search process.

Objective:

- **Data Acquisition:** Gather comprehensive data on PG accommodations from **MagicBricks** through web scraping.
- **Insight Generation:** Extract actionable insights to understand the distribution of PG accommodations across different areas of Hyderabad, typical rental price ranges, and Nearest distance to Metro
- **Problem Identification and Solution Proposal:** Identify key challenges faced by students in finding PG accommodations and propose data-driven solutions to address these challenges.
- **Visualization and Presentation:** Create compelling visualizations and prepare a comprehensive presentation to communicate findings, insights, and proposed solutions effectively.

INNOMATICS RESEARCH LABS

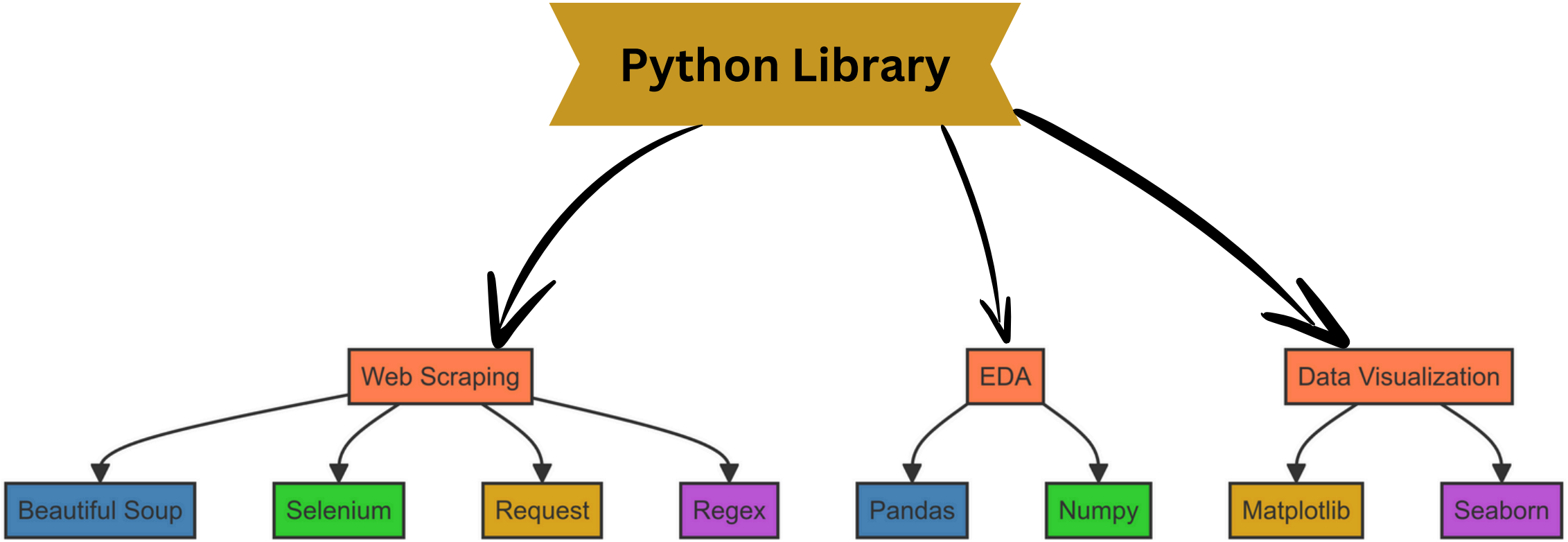


WEB SCRAPPING

Web scrapping is a technique for extracting data and content from websites. Web scraping can be done manually.

INNOMATICS RESEARCH LABS

Libraries Used:



INNOMATICS RESEARCH LABS

DATA EXTRACTION

	Unnamed: 0	Title	Location	Price	Room Info	Type	Places
0	0	N GRAND MEN'S PG/Paying Guest	in Kondapur	#10,000 Onwards	Single Room #22,000 Twin Sharing #12,000 Tri...	Boys#10,000 OnwardsFood IncludedBeds Available...	0.2 Km from Chirec International School 1.4 Km...
1	1	Sugathi Hometel PG/Paying Guest For Women	in Nanakaramguda	#7,000 Onwards	Single Room #16,000 Twin Sharing #9,000 Trip...	Girls#7,000 OnwardsFood IncludedBeds Available...	0.9 Km from Isb (Indian School Of Business) 1....
2	2	MS Comforts Women's PG/Paying Guest	in Gachibowli	#8,000 Onwards	Single Room #18,000 Twin Sharing #10,000 Tri...	Girls#8,000 OnwardsFood IncludedBeds Available...	0.3 Km from Asha Kiran 0.6 Km from Dlf Cyberci...
3	3	SKY LINE CO LIVING PG/Paying Guest	in Gachibowli	#18,000 Onwards	Single Room With AC #30,000 Twin Sharing With...	Coed#18,000 OnwardsFood IncludedBeds Available...	0.9 Km from Isb (Indian School Of Business) 1....
4	4	Manjula Luxury PG/Paying Guest For MEN'S	in Gowlidoddy	#8,000 Onwards	Single Room #18,000 Twin Sharing #10,000 Tri...	Boys#8,000 OnwardsFood IncludedBeds Available ...	1.6 Km from Isb (Indian School Of Business) 1....

Data Overview

Data Source - Magic Bricks

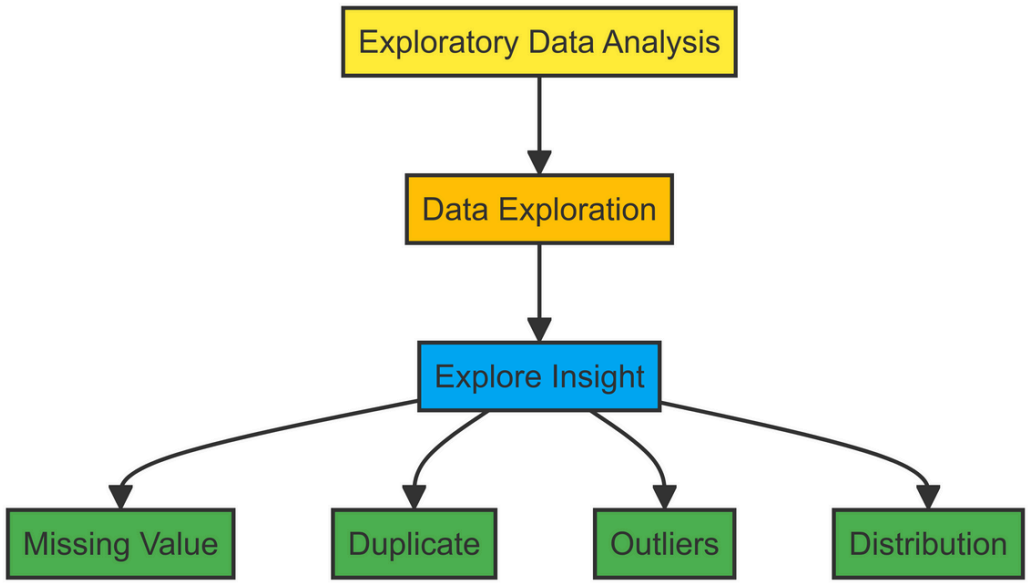
Data Info- I Extract 980 rows and 6 Features

INNOMATICS RESEARCH LABS

EXPLORATORY DATA ANALYSIS

Missing Value

1 Data Exploration




```
df.isnull().sum()
```

```
Title                0
Location             0
Single Room         35
Twin Sharing         0
Triple Sharing     105
Four Sharing       525
Has AC               0
Nearest Metro        0
Metro Distance       0
Gender               0
dtype: int64
```

3 Features have Missing Value:

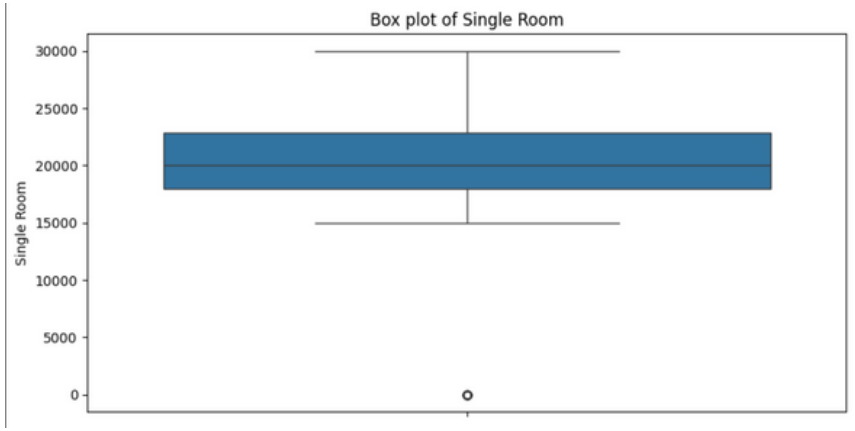
- Single Room
- Triple Sharing
- Four Sharing

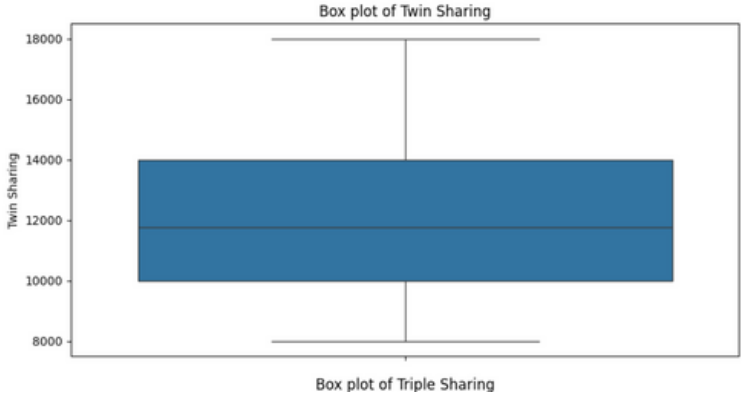
INNOMATICS RESEARCH LABS

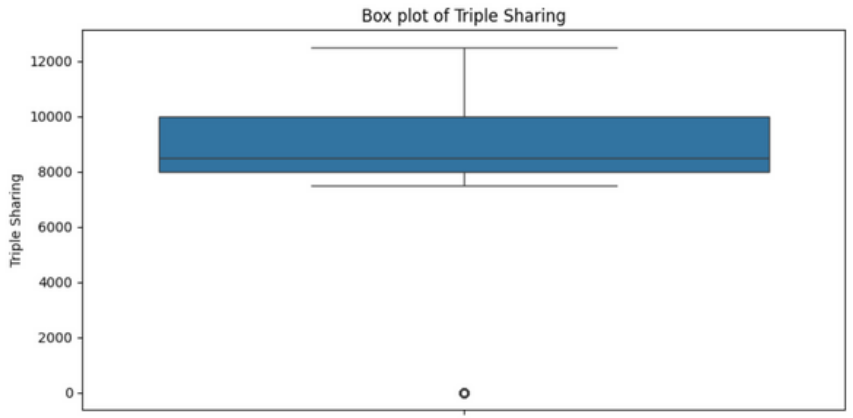
Outlier

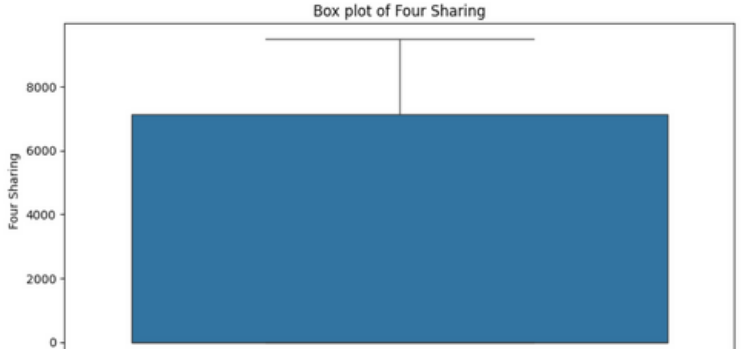
3 Features have Outliers Value:

- Single Room
- Triple Sharing
- Four Sharing









INNOMATICS RESEARCH LABS

Structural Error

```
[70] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 980 entries, 0 to 979
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Title           980 non-null    object 
 1   Location        980 non-null    object 
 2   Single Room     945 non-null    float64
 3   Twin Sharing    980 non-null    int64  
 4   Triple Sharing  875 non-null    float64
 5   Four Sharing    455 non-null    float64
 6   Has AC          980 non-null    object 
 7   Nearest Metro   980 non-null    object 
 8   Metro Distance  980 non-null    object 
 9   Gender          980 non-null    object 
dtypes: float64(3), int64(1), object(6)
memory usage: 76.7+ KB
```

1 Features have Dtype Error Value:

- Metro Distance

INNOMATICS RESEARCH LABS

Duplicated

Explanation

1. Multiple Listings of the Same PG:
   - Reason: A single PG accommodation might be listed multiple times across various platforms or by different users. Each listing could provide unique attributes such as price variations, room availability, or promotional offers.
   - Example: A PG named "Comfort Stay" might be listed with different prices or room types (single room, double sharing, etc.).
   - Usage: Retaining these duplicates helps in analyzing price trends, room preferences, and availability.
2. Proximity to Multiple Metro Stations:
   - Reason: A PG accommodation might be located near more than one metro station, making it accessible from different routes.
   - Example: A PG near both "Central Metro" and "Park Street Metro" could have listings showing proximity to either station.
   - Usage: Keeping these duplicates is crucial for users who prioritize accessibility and wish to choose accommodations based on their daily commute routes.

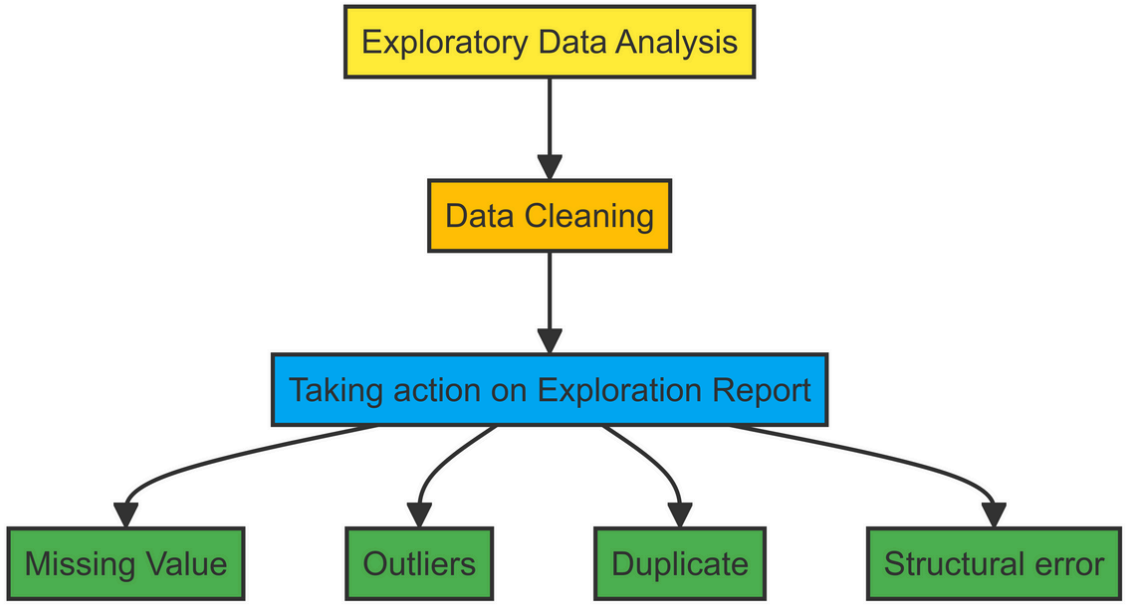
INNOMATICS RESEARCH LABS

Duplicated

3. Temporal Changes in Data:
    - Reason: Prices and availability of PG accommodations can change over time due to demand, season, or special events.
    - Example: The same PG might have different prices in summer compared to winter.
    - Usage: Analyzing these temporal changes requires retaining duplicates to track how prices or availability fluctuate over time.
4. Data from Multiple Sources:
    - Reason: Collecting data from multiple sources can result in duplicates, each source providing additional context or updates.
    - Example: Listings from different rental websites or apps might overlap but offer unique details.
    - Usage: Aggregating data from multiple sources enhances the richness of the dataset, offering a more comprehensive view.

INNOMATICS RESEARCH LABS

2. Data Cleaning



Missing Value

```
# TreatMent Using Domain Knowledge
df.fillna("0",inplace = True)
```

```
Unnamed: 0          0
Name                0
Location            0
Single Room         0
Twin Sharing        0
Triple Sharing      0
Four Sharing        0
Has AC              0
Nearest Metro       0
Metro Distance      0
Type                0
dtype: int64
```

To handle missing values for unavailable room types (e.g., Single Room, Three Sharing, Four Sharing) in PG listings, we used the fillna method to replace them with a constant value of 0.

Structural Error

```python
# Dtpye Treatment
data = {
    'Single Room': ['1000', '2000', 'N/A', '1500'],
    'Triple Sharing': ['3000', '2500', '3500', 'N/A'],
    'Four Sharing': ['2000', 'N/A', '1800', '2100']
}

df = pd.DataFrame(data)

# Convert columns to numeric, coerce errors to NaN, and fill NaN with 0
df['Single Room'] = pd.to_numeric(df['Single Room'], errors='coerce').fillna(0).astype(np.int64)
df['Triple Sharing'] = pd.to_numeric(df['Triple Sharing'], errors='coerce').fillna(0).astype(np.int64)
df['Four Sharing'] = pd.to_numeric(df['Four Sharing'], errors='coerce').fillna(0).astype(np.int64)

print(df.dtypes)
print(df)
```

```
Single Room       int64
Triple Sharing    int64
Four Sharing      int64
dtype: object
   Single Room  Triple Sharing  Four Sharing
0         1000            3000          2000
1         2000            2500             0
2            0            3500          1800
3         1500               0          2100
```

Correct Value

```python
df.info()
```

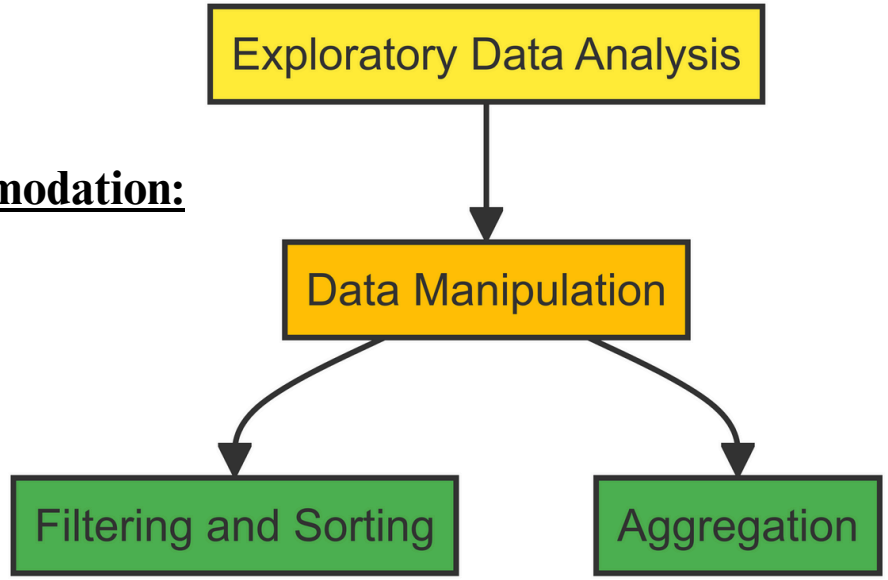
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 980 entries, 0 to 979
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Unnamed: 0      980 non-null    int64
 1   Name            980 non-null    object
 2   Location        980 non-null    object
 3   Single Room     945 non-null    float64
 4   Twin Sharing    980 non-null    int64
 5   Triple Sharing  875 non-null    float64
 6   Four Sharing    455 non-null    float64
 7   Has AC          980 non-null    object
 8   Nearest Metro   980 non-null    object
 9   Metro Distance  980 non-null    float64
 10  Type            980 non-null    object
dtypes: float64(4), int64(2), object(5)
memory usage: 84.3+ KB
```

Converted the data type of 'Metro Distance' from object to float using the astype method for accurate analysis

CLEANED DATA

	Name	Location	Single Room	Twin Sharing	Triple Sharing	Four Sharing	Has AC	Nearest Metro	Metro Distance	Type
	N GRAND MEN'S PG/Paying Guest	Kondapur	22000.0	12000	10000.0	0	No	Raidurg Metro Station	3.0	Boys
	Sugathi Hometel PG/Paying Guest For Women	Nanakaramguda	16000.0	9000	8500.0	7000.0	No	Raidurg Metro Station	4.0	Girls
	MS Comforts Women's PG/Paying Guest	Gachibowli	18000.0	10000	9000.0	8000.0	No	Raidurg Metro Station	3.0	Girls
	SKY LINE CO LIVING PG/Paying Guest	Gachibowli	30000.0	18000	0	0	Yes	Raidurg Metro Station	4.0	Coed
	Manjula Luxury PG/Paying Guest For MEN'S	Gowlidoddy	18000.0	10000	8000.0	0	No	Raidurg Metro Station	5.0	Boys
	SVS Co living Stays PG/Paying Guest	Kondapur	20000.0	12000	10000.0	0	Yes	Hitech Metro Station	3.0	Coed
	TULASI PREMIUM CO LIVING AND GUEST ROOMS PG/Pa...	Kondapur	26000.0	16000	12000.0	0	Yes	Miyapur Metro Station	2.0	Coed
	Unique Women's PG/Paying Guest	Gachibowli	24000.0	14000	10000.0	9000.0	No	Raidurg Metro Station	1.0	Girls
	GOPI MEN'S PG/Paying Guest	Kondapur	20000.0	10000	8000.0	7000.0	No	Miyapur Metro Station	2.0	Boys

3. Data Manipulation



I Take Some Data Manipulation (7) Problem Statements for PG Accommodation:

1.Filter PGs with Twin Sharing Price Less Than ₹10,000

```
# 2 Filter PGs that offer Triple Sharing and have AC
pg_triple_sharing_ac = df[(df['Triple Sharing'] != 'Not Available') & (df['Has AC'] == 'Yes')]
pg_triple_sharing_ac[['Name', 'Location', 'Triple Sharing', 'Has AC']]
```

	Name	Location	Triple Sharing	Has AC
5	SVS Co living Stays PG/Paying Guest	Kondapur	10000	Yes
6	TULASI PREMIUM CO LIVING AND GUEST ROOMS PG/Pa...	Kondapur	12000	Yes
12	TECHIES ELITE CO-LIVING PG/Paying Guest	Hafeezpet, NH 9	12500	Yes
16	COZY EXECUTIVE WOMEN'S PG/Paying Guest	Gachibowli	8500	Yes
19	Le bestow coliving - Bhuvan PG/Paying Guest	Kondapur	12000	Yes
...	...	...	...	...
971	Le bestow coliving - Bhuvan PG/Paying Guest	Kondapur	12000	Yes
973	RRR COMFORT STAY WOMENS separate building MENS...	Gachibowli	8500	Yes
976	Estay Executive Men's PG/Paying Guest	Hitech City	8500	Yes
977	R3 ATMOS LIVE THE SWAG PG/Paying Guest	Hitex Road	11000	Yes
978	SVS Co living Stays PG/Paying Guest	Kondapur	9000	Yes

315 rows × 4 columns

Insight

- Out of 978 PG 315 PG price is less than 10000 in Two Sharing type

2. Filter PGs Offering Triple Sharing and Have AC

```
# 2 Filter PGs that offer Triple Sharing and have AC
pg_triple_sharing_ac = df[(df['Triple Sharing'] != 'Not Available') & (df['Has AC'] == 'Yes')]
pg_triple_sharing_ac[['Name', 'Location', 'Triple Sharing', 'Has AC']]
```

	Name	Location	Triple Sharing	Has AC
5	SVS Co living Stays PG/Paying Guest	Kondapur	10000	Yes
6	TULASI PREMIUM CO LIVING AND GUEST ROOMS PG/Pa...	Kondapur	12000	Yes
12	TECHIES ELITE CO-LIVING PG/Paying Guest	Hafeezpet, NH 9	12500	Yes
16	COZY EXECUTIVE WOMEN'S PG/Paying Guest	Gachibowli	8500	Yes
19	Le bestow coliving - Bhuvan PG/Paying Guest	Kondapur	12000	Yes
...	...	...	...	...
971	Le bestow coliving - Bhuvan PG/Paying Guest	Kondapur	12000	Yes
973	RRR COMFORT STAY WOMENS separate building MENS...	Gachibowli	8500	Yes
976	Estay Executive Men's PG/Paying Guest	Hitech City	8500	Yes
977	R3 ATMOS LIVE THE SWAG PG/Paying Guest	Hitex Road	11000	Yes
978	SVS Co living Stays PG/Paying Guest	Kondapur	9000	Yes

315 rows × 4 columns

Insight

- Out of 978 PG 315 PG having AC in Triple Sharing type

3. List Names and Locations of PGs with Metro Distance Less Than 4 km

```
# 4 Filter PGs with Metro Distance less than 4 km
pg_metro_distance_less_4km = df[df['Metro Distance'] < 4]
pg_metro_distance_less_4km[['Name', 'Location', 'Metro Distance']]
```

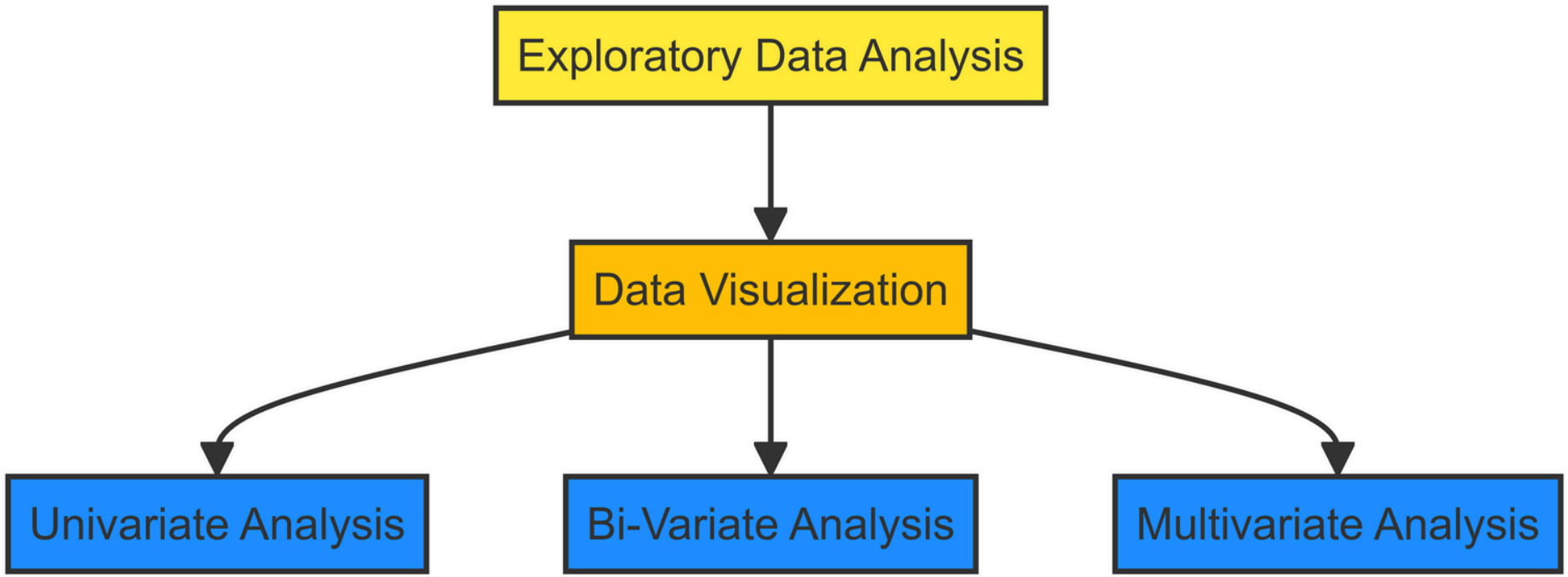
	Name	Location	Metro Distance
0	N GRAND MEN'S PG/Paying Guest	Kondapur	3.3
2	MS Comforts Women's PG/Paying Guest	Gachibowli	3.0
5	SVS Co living Stays PG/Paying Guest	Kondapur	3.3
6	TULASI PREMIUM CO LIVING AND GUEST ROOMS PG/Pa...	Kondapur	2.5
7	Unique Women's PG/Paying Guest	Gachibowli	1.6
...	...	...	...
975	My Spaces Luxury Men's PG/Paying Guest	Kondapur	2.6
976	Estay Executive Men's PG/Paying Guest	Hitech City	0.4
977	R3 ATMOS LIVE THE SWAG PG/Paying Guest	Hitex Road	1.7
978	SVS Co living Stays PG/Paying Guest	Kondapur	2.9
979	Sasya Elite Co-Living PG/Paying Guest	Gachibowli	2.3

805 rows × 3 columns

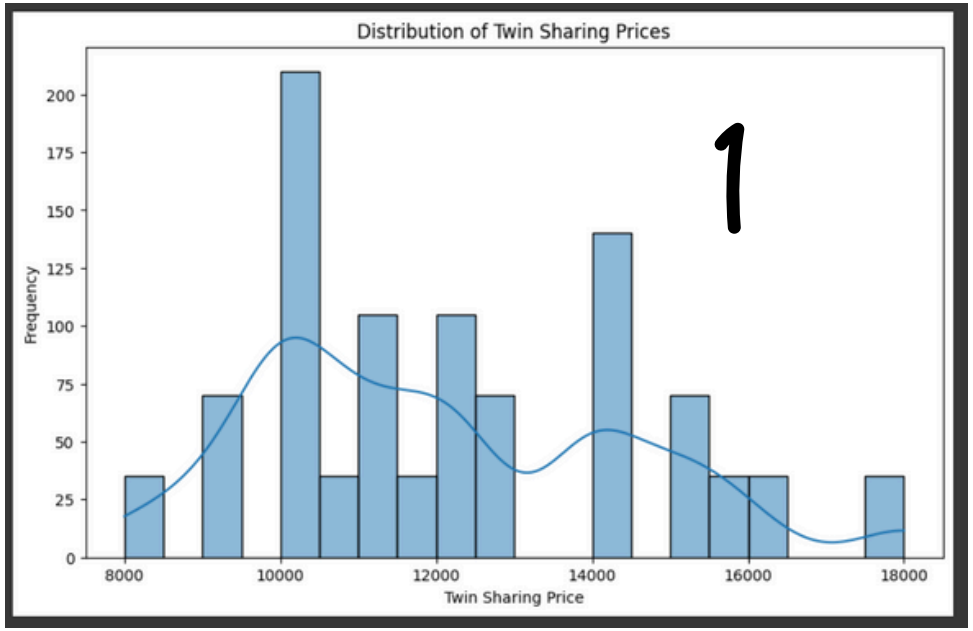
Insight

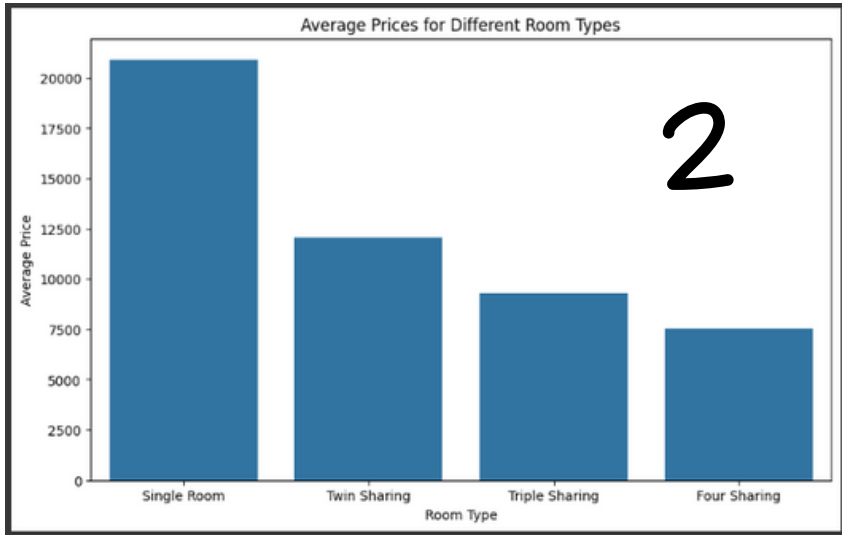
- Out of 978 PG 805 with metro distance less than 4km

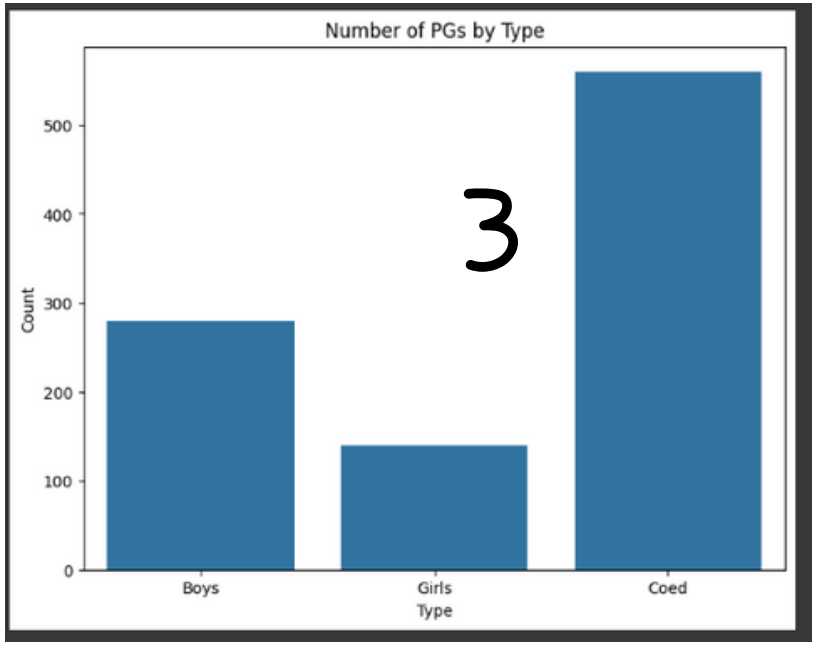
4 . Data Visualization



UNIVARIATE ANALYSIS





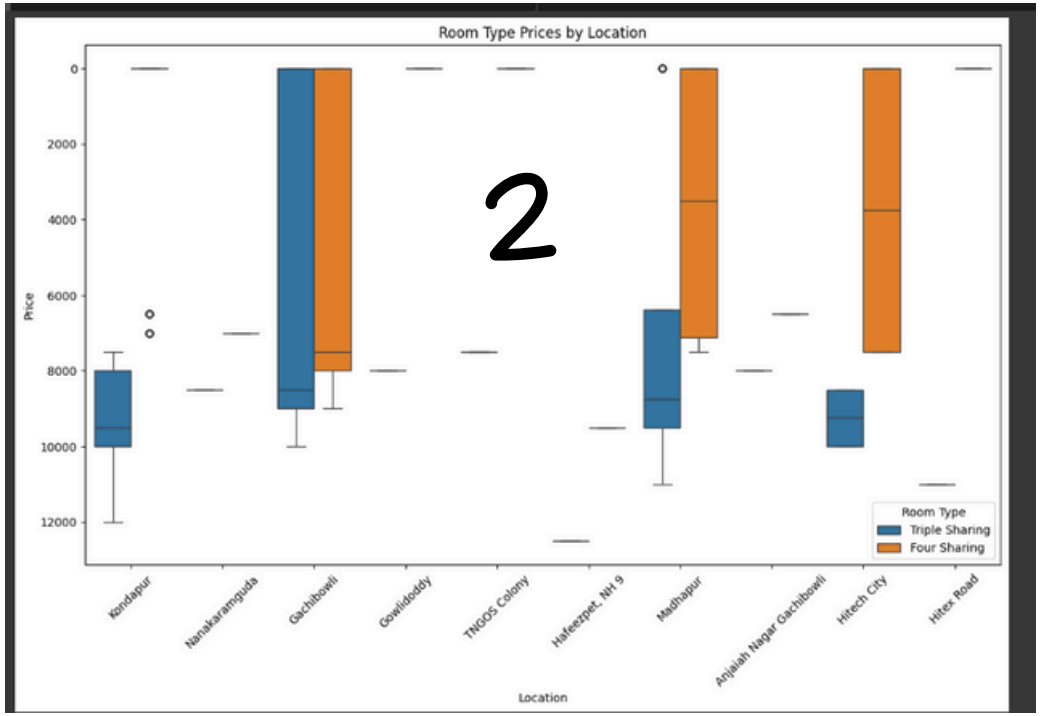


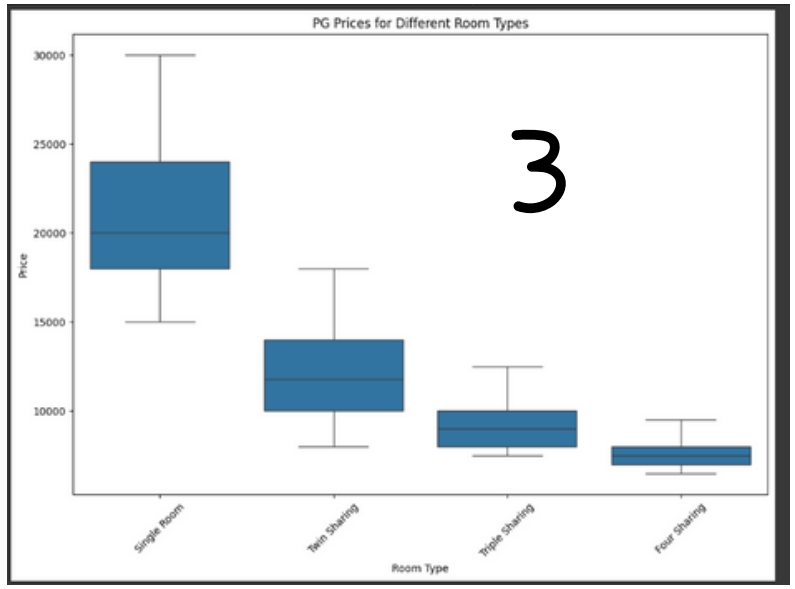
Problem Statement and Insights

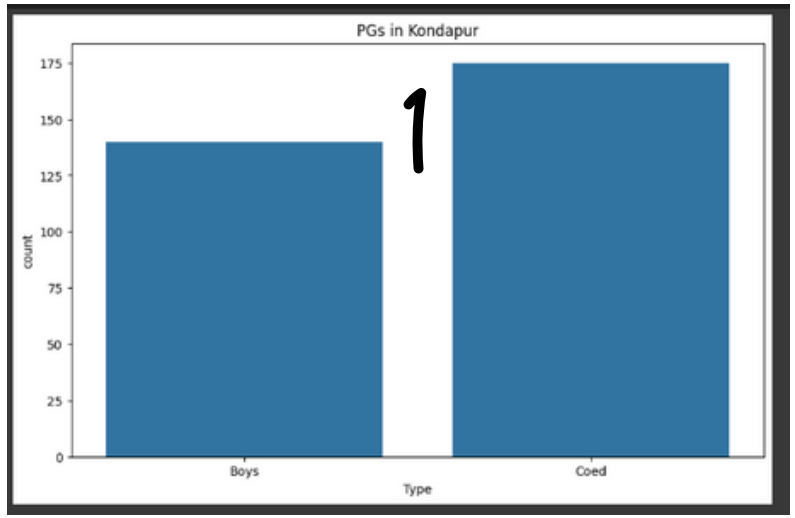
Problem Statement	Insight
1. Find all PGs with a Twin Sharing price less than 10,000.	The distribution plot of Twin Sharing prices shows that there is a significant number of PGs with prices around 10,000. Finding those below 10,000 requires filtering the dataset.
2. Find the average price for each room type (Single Room, Twin Sharing, Triple Sharing, Four Sharing).	The bar plot indicates that Single Rooms have the highest average price, followed by Twin Sharing, Triple Sharing, and Four Sharing.
3. Find the number of PG By Type (Boys, Girls, Coed).	The count plot of PGs by type reveals that Coed PGs are the most common, followed by Boys and Girls PGs. To find PGs without AC, further filtering is needed.

INNOMATICS RESEARCH LABS

BIVARIATE ANALYSIS



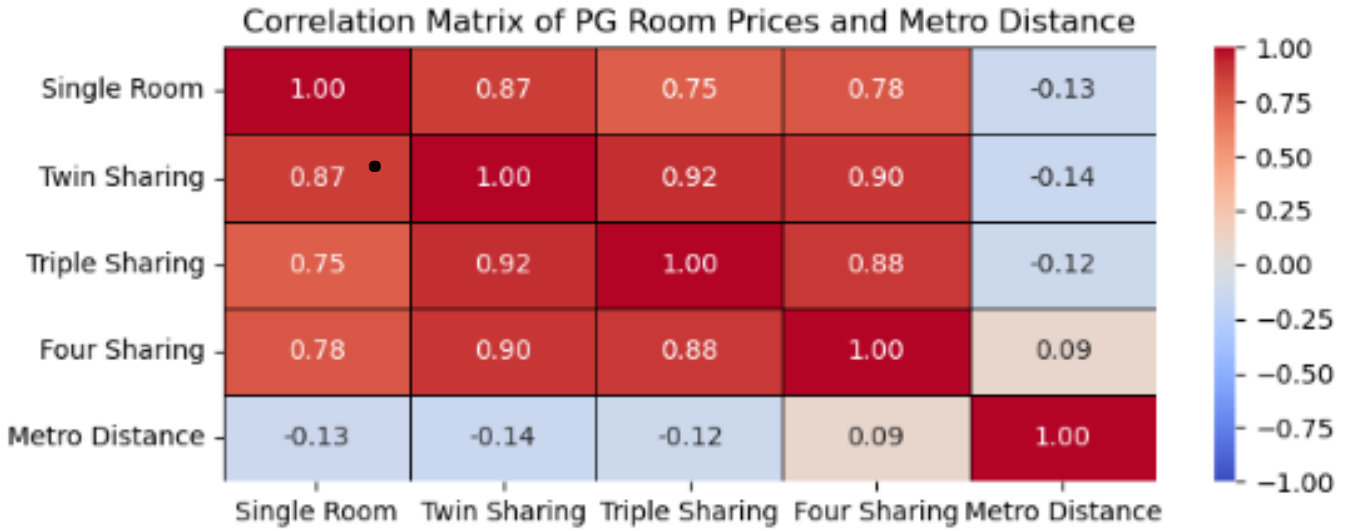




Visualization	Problem Statement	Insight
Visualization 1	Filter PGs in Kondapur and visualize the distribution of PG types (Boys, Coed).	The number of Coed PGs in Kondapur is significantly higher than Boys PGs.
Visualization 2	Identify PGs that offer Triple Sharing and Four Sharing with AC and visualize their prices by location.	Triple Sharing prices vary significantly across locations, with Kondapur and Miyapur having the highest prices.
Visualization 3	Find PGs for girls that offer Twin Sharing at a price of 9,000 or less and visualize prices for different room types.	Twin Sharing prices are generally lower than other room types, and many locations offer Twin Sharing below 9,000.

INNOMATICS RESEARCH LABS

MULTIVARIATE ANALYSIS: CORRELATION FOR NUMERICAL DATA



Insight

Insight	Details
High Positive Correlations	
Single Room and Twin Sharing	Correlation: 0.87 - Higher single room prices correlate with higher twin sharing prices
Twin Sharing and Triple Sharing	Correlation: 0.92 - Higher twin sharing prices correlate with higher triple sharing prices
Triple Sharing and Four Sharing	Correlation: 0.88 - Higher triple sharing prices correlate with higher four sharing prices
Weak Negative Correlations with Metro Distance	
Single Room and Metro Distance	Correlation: -0.13 - Slight tendency for higher prices closer to metro stations
Twin Sharing and Metro Distance	Correlation: -0.14 - Slight tendency for higher prices closer to metro stations
Triple Sharing and Metro Distance	Correlation: -0.12 - Slight tendency for higher prices closer to metro stations
Negligible Correlation	
Four Sharing and Metro Distance	Correlation: 0.09 - Almost no relationship between four sharing prices and metro proximity

INNOMATICS RESEARCH LABS

CONCLUSION

In conclusion, our analysis of PG accommodations in Kondapur and surrounding areas has yielded several key insights that can guide both prospective tenants and property managers.

1. **Distribution of PG Types:** The data reveals a significant prevalence of Coed PGs in Kondapur compared to Boys PGs. This insight can help property managers understand market demand and adjust their offerings accordingly.

2. **Price Variations by Room Type:** Prices for Triple Sharing rooms vary significantly across different locations, with Kondapur and Miyapur showing the highest prices. This indicates a potential premium for these areas, which could be due to better amenities or proximity to key locations.

3. **Affordable Twin Sharing Options:** Many locations offer Twin Sharing rooms at prices below 9,000, making them a viable option for budget-conscious tenants. This insight is crucial for students and young professionals looking for affordable housing options.

CONCLUSION

4. **Correlation Insights:** - There is a high positive correlation between the prices of different room types, indicating that as the price of one room type increases, the prices of other room types tend to increase as well. - Weak negative correlations with metro distance suggest a slight tendency for higher prices closer to metro stations, although this is not a strong determinant. - Four Sharing rooms show negligible correlation with metro proximity, indicating that their prices are less influenced by distance to metro stations.

5. **Filtering for Specific Needs:** The ability to filter PGs based on specific criteria such as price, room type, and amenities (like AC) allows for a more tailored search, making it easier for tenants to find accommodations that meet their specific needs. Overall, these insights provide a comprehensive understanding of the PG market in Kondapur and can help in making informed decisions regarding accommodation choices. Future analyses could further explore other factors influencing PG prices and availability, such as seasonal trends and additional amenities.

INNOMATICS RESEARCH LABS

THANK YOU

INNOMATICS RESEARCH LABS