Assignment 4

Problem Statement:

Data Wrangling

Data Wrangling on Real Estate Market

Dataset: "RealEstate_Prices.csv"

Description: The dataset contains information about housing prices in a specific real estate market. It includes various attributes such as property characteristics, location, sale prices, and other relevant features. The goal is to perform data wrangling to gain insights into the factors influencing housing prices and prepare the dataset for further analysis or modeling.

Tasks to Perform:

1. Import the "RealEstate_Prices.csv" dataset. Clean column names by removing spaces, special characters, or renaming them for clarity.
2. Handle missing values in the dataset, deciding on an appropriate strategy (e.g., imputation or removal).
3. Perform data merging if additional datasets with relevant information are available (e.g., neighborhood demographics or nearby amenities).
4. Filter and subset the data based on specific criteria, such as a particular time period, property type, or location.
5. Handle categorical variables by encoding them appropriately (e.g., one-hot encoding or label encoding) for further analysis.
6. Aggregate the data to calculate summary statistics or derived metrics such as average sale prices by neighborhood or property type.
7. Identify and handle outliers or extreme values in the data that may affect the analysis or modeling process.

Objective:

The objective of data wrangling in the real estate market is to collect, clean, transform, and prepare raw real estate data to make it suitable for analysis and decision-making. Data wrangling, also known as data munging or data preprocessing, is a crucial step in the data analysis process. By performing data wrangling, analysts and stakeholders can gain valuable insights and make informed decisions in the real estate industry.

Here are some specific objectives of data wrangling in the real estate market:

1) **Data Collection:** Gather raw data from various sources, such as real estate listings, property databases, government agencies, real estate agents, and online platforms.
2) Data Cleaning: Identify and handle data quality issues, such as missing values, outliers, and

inconsistencies, to ensure the data is accurate and reliable.

3) **Data Transformation**: Convert data into a consistent format and structure to facilitate analysis. This may involve converting data types, standardizing units of measurement, and handling categorical variables.

4) **Data Integration:** Combine data from different sources and merge relevant information to create a comprehensive dataset for analysis.

5) **Feature Engineering:** Create new variables or features from existing data that might provide valuable insights. For example, calculating price per square foot, creating location-based features, or deriving property age from the construction year.

6) **Data Enrichment:** Augment the dataset with additional relevant information, such as demographic data, economic indicators, or market trends, to provide a broader context for analysis.

7) **Data Reduction:** If the dataset is too large or contains redundant information, data wrangling can involve reducing the data size while retaining its essential characteristics.

8) **Handling Missing Data:** Develop strategies to handle missing data points, such as imputation techniques or excluding records with missing values, based on the impact on the analysis.

9) **Data Visualization:** Generate visual representations of the data during the wrangling process to explore patterns, identify anomalies, and verify the effectiveness of cleaning and transformation steps.

10) **Data Documentation:** Maintain detailed documentation of the data wrangling process to ensure transparency, reproducibility, and collaboration among analysts and stakeholders.

11) By accomplishing these objectives, data wrangling empowers real estate professionals, investors, and policymakers to make better-informed decisions, identify market trends, understand property valuation, assess risk, and discover opportunities in the real estate market.

Theory:

What is Data Wrangling?

Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data wrangling is the practice of converting and then plotting data from one “raw” form into another.

Data Wrangling is also known as Data Munging, data cleansing, data scrubbing, data cleaning, or data remediation.

Data Wrangling in Python

Data Wrangling is a crucial topic for Data Science and Data Analysis. Pandas Framework of Python is used for Data Wrangling. Pandas is an open-source library in Python specifically developed for Data Analysis and Data Science. It is used for processes like data sorting or filtration, Data grouping, etc. Data wrangling in Python deals with the below functionalities:

1. **Data exploration:** In this process, the data is studied, analyzed, and understood by visualizing representations of data.
2. **Dealing with missing values:** Most of the datasets having a vast amount of data contain missing values of *NaN, they are needed to be taken* care of by replacing them with mean, mode, the most frequent value of the column, or simply by dropping the row having a *NaN* value.
3. **Reshaping data:** In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.
4. **Filtering data:** Some times datasets are comprised of unwanted rows or columns which are required to be removed or filtered

Fig: steps in data wrangling

There are 6 steps as follows:

1. **Data Discovery**: This is an all-encompassing term that describes understanding what your data is all about. In this first step, you get familiar with your data

2. **Data Structuring**: When you **collect** raw data, it initially is in all shapes and sizes, and has no definite structure. Such data needs to be restructured to suit the analytical model that your enterprise plans to deploy

3. **Data Cleaning**: Raw data comes with some errors that need to be fixed before data is passed on to the next stage. **Cleaning** involves the tackling of outliers, making corrections, or deleting bad data completely
4. **Data Enriching**: By this stage, you have kind of become familiar with the data in hand. Now is the time to ask yourself this question – do you need to embellish the raw data? Do you want to augment it with other data?

5. **Data Validating**: This activity surfaces data quality issues, and they have to be addressed with the necessary transformations. The rules of **validation** rules require repetitive programming steps to check the authenticity and the quality of your data

6. **Data Publishing**: Once all the above steps are completed, the final output of your data **wrangling efforts** is pushed downstream for your analytics needs

Algorithm:

Data wrangling is the process of gathering, cleaning, transforming, and organizing raw data into a format suitable for analysis. When it comes to the real estate market, data wrangling plays a crucial role in making the data usable and extracting valuable insights. Here's a step-by-step guide to data wrangling for the real estate market:

1) **Data Collection:**
   Identify the sources of real estate data you want to analyze. These could include public databases, real estate websites, APIs, or data provided by real estate agencies. Decide on the specific variables you need, such as property prices, location, property size, number of bedrooms, etc.
2) **Data Cleaning**:
   Remove any duplicate records from the dataset to ensure data accuracy. Handle missing values. You can either remove rows with missing values or use imputation techniques (e.g., mean, median, regression) to fill in missing data where appropriate. Check for and correct data entry errors or inconsistencies. Standardize data formats (e.g., converting dates to a uniform format) to facilitate analysis.
3) **Data Transformation:**
   Convert categorical variables to numerical representations using techniques like one-hot encoding or label encoding. Extract relevant features from the existing data. For example, you might extract the year from a date variable to create a separate "Year" feature. If you have unstructured data (e.g., property descriptions), consider using natural language processing (NLP) techniques to extract meaningful information.
4) **Data Integration:**
   If you have data from multiple sources, integrate them into a single dataset. Ensure that the

data formats are consistent. Merge datasets using common identifiers (e.g., property IDs) if you need to combine information from different sources.

5) **Data Exploration:**
   Perform exploratory data analysis (EDA) to understand the distribution and relationships of variables. Visualize the data through plots, histograms, scatter plots, and other relevant charts to identify patterns and outliers.
6) **Data Analysis:**
   Apply statistical methods and machine learning algorithms to extract insights from the data.
   Perform regression analysis to understand the relationship between property features and prices.
   Use clustering techniques to group similar properties together.
7) **Data Presentation:**
   Summarize your findings in a clear and understandable manner.Create visualizations and reports to present the results effectively.

Conclusion

In this way, we perform Data Wrangling on Real Estate Market .

Oral Questions:

1) What is Data Wrangling?
2) What are the different steps in data wrangling?
3) What is data cleaning?
4) How to perform Data Wrangling on Real Estate Market
5) What is data visualization?