Report - Week 1

Created	@July 5, 2024 9:23 AM
Reviewed	☐

Internship Week 1 Report: Advanced Data Preparation and EDA

Dataset: NYC Property Sales

1. Data Cleaning

- **Handling Missing Values:**
  - Utilized advanced techniques to handle missing values, replacing empty strings and specific placeholders (`NA`, ` `) with NaN across selected columns (`GROSS SQUARE FEET`, `LAND SQUARE FEET`, `SALE PRICE`).
  - Imputed missing values for `GROSS SQUARE FEET` and `LAND SQUARE FEET` using their respective means after cleaning.

- **Outliers Detection and Handling:**
  - Employed Interquartile Range (IQR) method to identify and replace outliers in columns such as `SALE PRICE`, `GROSS SQUARE FEET`, `LAND SQUARE FEET`, `TAX CLASS AT TIME OF SALE`, `YEAR`, `DAY`, `MONTH`, and `YEAR BUILT`.
  - Outliers were replaced with the mean values of respective columns to ensure data integrity and mitigate the impact of extreme values on analysis.

2. Data Transformation

- **Feature Engineering:**
  - Engineered new features from `SALE DATE` to extract `YEAR`, `MONTH`, and `DAY` attributes, facilitating time-based analysis.

- **Encoding:**
  - Utilized Frequency Encoding for columns `TAX CLASS AT PRESENT` and `ADDRESS` to transform categorical variables into numerical representations based on their frequency in the dataset.
  - Implemented Target Encoding for columns `NEIGHBORHOOD`, `BUILDING CLASS CATEGORY`, `BUILDING CLASS AT PRESENT`, and `BUILDING CLASS AT TIME OF SALE` using K-Fold cross-validation, replacing categorical values with the mean of `SALE PRICE` for each category.

3. Advanced Data Visualization

- **Interactive Plots with Plotly:**
  - Created interactive scatter plots to visualize relationships such as `Year Built` vs. `Sale`, `Land Square Feet` vs. `Gross Square Feet`, `Gross Square Feet` vs. `Sale Price`, and `Land Square Feet` vs. `Sale Price`.
  - Developed histograms and box plots to analyze distributions and variations in `Sale Price`.

Key Insights and Actions:

- **Data Integrity:** Ensured data cleanliness and consistency by addressing missing values and outliers early in the data preparation phase.
- **Feature Engineering:** Enhanced dataset capabilities by deriving new features from existing data, improving model performance and interpretability.
- **Visualization:** Leveraged interactive plots to effectively communicate relationships and distributions, aiding in exploratory analysis and insight generation.

Conclusion:

Week 1 of the internship focused on foundational tasks of data cleaning, transformation, and visualization, setting a solid groundwork for subsequent analytical and modeling tasks. The approach adopted ensures robustness and reliability in preparing the NYC Property Sales dataset for advanced data science tasks

---

Issues Faced and Questions

- **Dealing with Outliers**: During data cleaning, I encountered challenges in deciding whether to remove outliers outright or replace them with mean values. I was unsure about the impact of these decisions on the accuracy of my predictive models. For example, in the case of **GROSS SQUARE FEET** and **LAND SQUARE FEET**, should outliers be replaced with mean values, or should they be removed altogether?

- **Data Quantity vs. Data Quality**: While cleaning the dataset, I noticed a significant reduction in data volume due to dropping rows with missing values and outliers. This raised the question of whether it is better to prioritize data quality by ensuring cleaner data, even if it means reducing the dataset size.

- **Encoding Query**: The process of encoding categorical variables, especially using methods like Target Encoding, resulted in the creation of numerous new columns. I'm concerned about the practicality and potential impact on model performance of having such a large number of features. Should I employ dimensionality reduction techniques or use alternative encoding methods to manage this issue effectively?

- **Preprocessing on Encoded Columns**: After encoding categorical variables, I'm uncertain about whether I need to address outliers in these newly encoded columns. Should outliers in encoded columns be removed or replaced, and should these columns be scaled similarly to numerical features before model training?

- **Visually Detecting Outliers**: I faced challenges in visually identifying outliers during exploratory data analysis (EDA). What are the recommended techniques or visualizations to effectively detect outliers in different types of data distributions (e.g., skewed distributions)?
