**‘HOUSING’ Project**

**DSA 8401 Applied Machine Learning Assignment**

**August 1, 2022**

This project contributes **15%** of your total points.

Due date: **August 26th 2022** (No assignments will be accepted after this date)

**Dataset Description:** Dataset contains information from an Assessor's Office used in computing value of home (SalePrice) for residential properties sold in Iowa from 2006 to 2010. The dataset is used to predict SalePrice based on a number of features.

Please refer to the companion document (**ahds.doc**) to view the data dictionary.

**Instructions:**

The purpose of this project is to expose the student to hands-on EDA and some Feature Engineering exercises.

You are required to use the dataset **'ahds.csv'** for this exercise.

Work through all sections and copy the relevant output (charts, comments) into a PowerPoint deck. You are only required to show the most relevant plots/charts/outputs related to the tasks below. Also provide short summary answers to any questions asked below.

Please keep to a **maximum of 6 slides (including a title slide)**.

**Title slide 1:** Full name/ date of submission

**SLIDE 2 (15pts):**

1. Perform a high-level look at the dataset and record your initial overall observations of the dataset showing the shape of the dataset, number of categorical features, number of numeric features, nature of the target feature (SalePrice).
2. Examine columns 'PID' and 'Order'. Would you use them for modeling? If not, why?
3. Focusing only on numerical features, identify 3 features which you suspect have outliers (show plots to indicate this fact).
4. Perform a fix for these outliers and show plots after the fix. Describe your approach(s) and reason you chose the approach(s).

**SLIDE 3 (15 pts):**

1. Find the top 3 features with the highest number of missing values (look into both numerical and categorical). Show the number of missing values in each.
2. How would you handle these 3 features identified in task 5 above? Perform an appropriate imputation for the missing values and show summary statistics after the fix. Briefly describe your approach(s) and reason(s) you chose the approach(s).

**SLIDE 4 (30 pts):**

1. Identify 2 highly skewed numerical features (hint: can use skew\_vals function). Show histograms of these features.
2. Try out a few transformations on the features identified in (7) to remove skewness. Describe the method(s) used and show histograms resulting from your best transformation.
3. Identify 3 categorical features with the highest cardinality. For each of these features, describe how you would convert them into numeric.
4. Perform one-hot encoding on the feature 'Garage Type'. Show the distribution (Min, Mean, Max) of the newly generated columns.

**SLIDE 5 (15 pts):**

12. Perform a bivariate analysis (box plots) of these features ('MS Zoning', 'Foundation','Garage Type') versus the 'SalePrice'. Which of these features do you think is more important for predicting sale price?

**SLIDE 6 (25pts):**

13. Isolate all numeric features (except SalePrice) and perform a PCA , reducing the dimensionality to 2 Principal Components. What is the proportion of variance explained? Plot 2 scatter plots (PCA1 vs SalePrice and PCA2 vs SalePrice). Which Principal component has a higher correlation with SalePrice?

14. How many Principal Components would you need to capture over 80% of the variation?