Lab 02: Exploratory Data Analysis (EDA) -- Visualization
CS3300 Data Science

Learning Outcomes

1. Understand the basic process of data science and exploratory data analysis including modes of inquiry (hypothesis driven, data driven, and methods driven).

2. Identify, access, load, and prepare (clean) a data set for a given problem.

3. Select, apply, and interpret appropriate visual and statistical methods to analyze distributions of individual variables and relationships between pairs of variables.

4. Communicate findings through generated data visualizations and reports.

5. Identify correlated and predictive variables.

Overview

In the previous lab, you loaded and inspected a data set of real estate transactions. In this lab, you are going to perform exploratory data analysis (EDA) to identify and explain the relationships between a target variable and other feature variables. You should prepare your results as a Jupyter notebook. In addition to code and plots, you should have text offering interpretations and explanations. Your notebook should be organized into sections with appropriate headers. The notebook and its code should be clean and polished. Use the Blood Glucose Tutorial as a template and reference.

Instructions

Note: In this lab, focus on the following columns: city, state, zip, beds, baths, type, street type.

Loading the Data

a. Load the CSV file of the cleaned data set you created in Lab 1.

Part I: Regression on Price

In part I, you are going to explore which variables are predictive of the price (serving as dependent (target) variable).

a. For each continuous variable, create a scatter plot of the continuous variable versus price. Make sure to put the independent variable (feature variable) on the horizontal axis and the dependent variable (target variable) on the vertical axis.

b. A predictive continuous independent variable (feature variable) will correlate with the output variable (target variable). For each continuous independent variable (feature variable), explain if you think the variable will be predictive or not. You could describe if one variable seems to be more predictive than the other based on the strength of the observed relationship.

c. For each categorical variable, create a box plot of the categorical variable versus price. Make sure to put the categorical variable on the horizontal axis and the dependent variable (target variable) on the vertical axis.

d. A predictive categorical independent variable will have different distributions of the output variable for each categorical value. For each categorical independent variable (feature variable), explain if you think the variable will be predictive or not. You could describe if one variable seems to be more predictive than the other based on the strength of the observed relationship.

Part II: Classification on Property Type

In part II, you are going to explore which variables are predictive of the property type (serving as dependent or target variable).

a. For each continuous variable, create a box plot of the continuous variable versus property type. Make sure to put the property type on the horizontal axis and the continuous variable on the vertical axis.

b. A predictive continuous independent (feature) variable will have different distributions for each categorical output value. Describe if each continuous feature variable is predictive or not.

c. For each categorical variable, create a cross-tabulation table that shows the counts of each categorical variable value for each property type. For each categorical feature, visualize the ***normalized*** cross-tabulation table using a bar plot for each possible property type.
Note: You may consider using a heatmap instead of a bar plot if the categorical feature contains lots of categories.

d. A categorical variable is predictive if each property type distributes differently across the categories. Describe if each categorical feature variable is predictive or not.

Part III: Compare Predictive Variables

a. How many variables are predictive for both problems?

b. Explain why you think each variable would be predictive of both or only one problem.

Submission Instructions

Save your Jupyter notebook as a PDF and upload that file through Canvas.

Rubric

Followed submission instructions	5%
Report is polished and clean. No unnecessary code. Section headers are used. Plots are described and interpreted using text. The report contains an introduction and conclusion.	10%
Loaded data	5%
Part I: Regression	
Scatter plots	10%
Box plots	10%
Variable Predictiveness	10%
Part II: Classification	
Box plots	10%
Bar plots of cross-tab table	15%
Variable Predictiveness	10%
Part III: Comparison	
Explanations of why variables are predictive of only one or both problems	10%
Exceeded Expectations	5%