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 dependent (output) and other independent 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

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 variable).

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

b. A predictive continuous independent variable will correlate with the output variable. Create a table in which you list describe each continuous independent variable as predictive or not.

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

d. A predictive categorical independent variable will have different distributions of the output variable for each categorical value. Create a table in which you describe each categorical independent variable as predictive or not.

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 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 variable will have different distributions of the for each categorical output value. Create a table in which you list describe each continuous independent variable as predictive or not.

c. For each categorical variable, create a heat map of the counts of each categorical variable value for each property type.

```
combination_counts = df.value_counts(subset=["type",
                                              "street_type"])
                      .unstack(level=0)
                     .fillna(0)
sns.heatmap(combination_counts)
```

d. A categorical variable is predictive if each value occurs frequently with one value of the output variable. Create a table in which you describe each categorical independent variable as predictive or not.

Part III: Compare Predictive Variables

a. Make a Venn diagram of the variables you described as predictive.

b. How many variables are predictive for both problems?

c. Create a table with an explanation as to 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%
Tables of Variable Predictiveness	10%
Part II: Classification	
Box plots	10%
Heat maps	10%
Tables of Variable Predictiveness	10%
Part III: Comparison	
Venn Diagram	10%
Explanations of why variables are predictive of only one or both problems	10%
Exceeded Expectations	5%