**Lab 03: Exploratory Data Analysis (EDA)**

**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.

**Overview**

In this lab, you are going to perform exploratory data analysis (EDA) on a data set of real estate transactions from California. Your goal is to identify and explain the relationships between the sale prices of the properties and other 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**

**1. Loading the Data and Initial Assessment**

a. Load the Sacramentorealestatetransactions.csv file as a DataFrame.

b. Using the output of the initial head() and info() commands, describe the data. What are the variables? What are their inferred types? Do any of the columns have null values?

**2. Cleaning Categorical Variables**

a. Sometimes a variable can be either an integer or categorical variable. Count the number of unique values for the streets, zip codes, and beds. Do you think it is more appropriate to represent these three variables as categorical variables? Why or why not?

b. Convert the following variables to categorical variables: city, state, zip, beds, baths, type

**3. Guess which variable should be the independent variable.**

**4. EDA of Continuous Variables**

a. Plot the square footage, latitude, and longitude against the price. What type of plot is appropriate?

b. Describe whether the variables are strongly correlated with the price. Hypothesize why these variables may or may not be correlated.

c. Do you notice “odd” patterns in any of the plots? Do you think the odd pattern(s) is/are real or artifact(s)?

**5. EDA of Categorical Variables**

a. Plot the beds, baths, type, state, city, and zip codes against the price. What type of plot is appropriate?

b. Do the prices vary with the number of beds, baths, and type? Hypothesize why these variables may or may not be correlated.

c. Is there anything “odd” about the numbers of beds and baths? Can you hypothesize what this odd value might mean?

**6. Engineering New Variables**

a. Count of the number of unique values for the addresses (streets variable). Do you think this variable is useful to analyze in its current form?

b. Streets (e.g., avenue, street, way) can indicate whether a road will be quiet or busy, is in a commercial or suburban area, etc. The streets can be extracted from the address. Using the head() command, look at the first 20 records to identify address patterns.

c. Write a function get\_street(address) that will return the street (as a String) of an address.

d. Use that function to create a new categorical variable of street types. Count the unique elements and print them out. If it looks like any addresses were parsed incorrectly (e.g., a number is returned instead of a street type), update your function in part c.

e. Plot the street type against the price. Do the prices vary with the street type?

**Submission Instructions**

Save the notebooks as HTML files by selecting Download as -> HTML (.html) from the File menu. Add the two HTML files to a zip (not 7zip!) file and submit through Blackboard. Do not include .ipynb files.

**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. | 20% |
| Appropriate plot types are selected for continuous vs continuous and categorical vs continuous comparisons. | 10% |
| Reasonable effort is made to answers all questions (e.g., 1-2 sentences each) | 45% |
| Street types are correctly extracted from the addresses. | 20% |