Lab 01: Data Cleaning
CS3300 Data Science

Learning Outcomes

1. Identify, access, load, and prepare (clean) a data set for a given problem.
2. Select, apply, and interpret appropriate visual and statistical methods to analyze distributions of individual variables.
3. Clean and transform data for analysis.
4. Communicate findings through generated data visualizations and reports.

Overview

In this lab, you are going to inspect and clean a data set of real estate transactions from California. 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, in a dedicated markdown cell in your notebook - describe the data.

Questions:

- What are the variables?
- What are their inferred types?
- Do any of the columns have null values?

2. Representing Categorical Variables

a. Sometimes a variable can either be represented as an integer or categorical variable. Count the number of unique values for the streets, zip codes, and beds.

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

Questions:

- Do you think it is more appropriate to represent these three variables as categorical or integer variables? Why or why not?

3. Cleaning Continuous Variables

a. Plot histograms of the square footage, latitudes, and longitudes.

Questions:

- Do you notice any “odd” patterns in any of the plots? Do you think they real or artifacts?

4. Cleaning Categorical Variables

a. Plot the beds, baths, type, state, city, and zip codes as count (bar) plots.

Questions:

- Do you notice any “odd” patterns in any of the plots? Do you think they real or artifacts?

5. Engineering New Variables – Part I

Entries with 0 square footage are empty lots. Encoding these cases with values of 0 lead to two different interpretations of the square footage variable. This is a good candidate for creating a new boolean variable.

a. Create a new boolean variable called "empty_lot". This variable should have a value of true if the square footage of a record is 0. Otherwise, it should have a value of false.

b. Add this feature to the dataframe.

c. Create a count (bar) plot for the empty_lot variable.

6. Engineering New Variables – Part II

a. Count of the number of unique values for the addresses (streets variable).

Question:

- Do you think this variable is useful for analysis or as a feature for a ML model in its current form?

b. Street types (e.g., avenue, street, way) can indicate whether a road will be quiet or busy, is in a commercial or suburban area, etc. The street types 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_type(address) that will return the street type (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. Repeat until you parse everything correctly.

e. Add this feature to the dataframe.

f. Plot the street types as a count (bar) plot.

7. Identifying Potential Dependent Variables

This data set can be used for both regression and classification problems.

a. Identify a variable which would make a good dependent (output) variable for a regression problem.

b. Identify a variable which would make a good dependent (output) variable for a classification problem.

Questions:

- What types of variables are appropriate for regression?
- What types of variables are appropriate for classification?

8. Identify and remove 2 outlier records from the dataset.

There are at least two data points where some information is “unknown”. Identify which data points these are and remove them from your data frame.

9. Save the Cleaned Data Set

Save your cleaned data set as a CSV file using to_csv with the index=False option.

Submission Instructions

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

Rubric

Followed submission instructions	5%
Formatting: 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.	5%
Loading and Describing Data	5%
Representing Categorical Variables	10%
Cleaning Continuous Variables	10%
Cleaning Categorical Variables	10%
Engineering New Variables – Part I	10%
Engineering New Variables – Part II (including correct parsing of all street types)	20%
Identification of Dependent Variables	10%
Removal of Unknown Observations	5%
Reasonable effort is made to answers all questions (e.g., 3-4 sentences each)	10%