**Project 2: Analysis of Real Estate**

**CS6230 AI Tools**

**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 project, you are going to analyze a data set of real estate transactions from California. In the first part, you will perform data analysis in a Jupyter notebook. For the second part of the project, you are going to use what you learned to characterize the properties based on their spatial distributions and prepare a short written report.

Your exploratory data analysis and technical work will be performed in a Jupyter notebook. Your notebook should be divided into sections, labeled and introduced with Markdown cells. Code should be divided into cells in logical groupings (if you have two code cells back-to-back without some sort of output like a plot, think about whether that is necessary.) Each plot should be labeled appropriately (axes, legends, units, etc.) and interpreted in a succeeding Markdown cell. The notebook and its code should be clean and polished – consider it to be a final product that you will present to other professionals.

**Instructions**

**Part I: Data Exploration and Cleaning**

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() methods, in a dedicated markdown cell in your notebook - describe the data. What are the variables? What are their inferred types? Do any of the columns have missing values?

2. Continuous Variables

a. Run the describe() method. Did summary statistics appear for all of the variables you expected to be numeric (floats or ints) from the output of the info() method? (If not, this is a good place to double check your work.)

b. Plot histograms of the prices, square footage, latitudes, and longitudes using Seaborn's displot() method with kde=False. Do you notice “odd” patterns in any of the plots? Do you think the odd pattern(s) is/are real or artifact(s)?

c. Are the latitudes and longitudes symmetric or skewed? Would you expect them to be? (Hint: think about how latitudes and longitudes might relate to neighborhoods)

d. Are there any numerical variables with inappropriate values (e.g., 0 or -1 for the number of bathrooms)? What do you think these indicate?

3. Categorical Variables

a. Pandas does not determine which variables are categorical by default. Strings are left as strings (and called objects in the info() method). Good categorical variables often have a relatively small number of unique values. Using the results of the info() method, make a list of all variables that you think could be categorical variables.

b. Count the number of unique values for each variable you find in (3a). Which variables do you think would be good categorical variables? Convert them to categorical variables. Run the info() method again to verify that these variables have the Dtype "category" next to them.

c. Sometimes it can be useful to represent numerical variables as categorical variables. Count the number of unique values for the zip codes, beds, and baths. Do you think it is more appropriate to represent these variables as categorical or integer variables? Why or why not? Create two new columns (cat\_beds, cat\_baths) that encode the number of bedrooms and bathrooms as categorical variables. Run the info() method to verify that these variables not have the Dtype "category" next to them.

d. Create a new DataFrame of only the categorical columns. Run the describe("category") method on this DataFrame. Based on the number of unique values, do you think the state is a useful variable? Why or why not?

e. Plot the types as a bar (count) plot. Is there anything odd about any of the types?

f. Plot the number of beds and baths as bar (count) plots. Is there anything “odd” about the numbers of beds and baths? If so, can you hypothesize what these outlier values might mean?

g. Plot the zip codes as a bar (count) plot. Even if the number of unique zip codes is large, why do you think it is better to use a categorical variable for zip code instead of a numerical variable?

i. Plot the city as a bar (count) plot. Which cities have the most properties in the data set? Which cities have the least?

4. Engineering New Variables

Assume that 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. Create a count (bar) plot for the empty\_lot variable.

**Part II: Bokeh Dashboard**

After cleaning the data, use the [Bokeh library](https://docs.bokeh.org/en/latest/index.html) to create a dashboard like so:



You will need to use pip to install the bokeh library in your virtual environment.

The OMDB [movie data explorer](https://github.com/bokeh/bokeh/tree/branch-3.0/examples/app/movies) is a good example to follow. It is much easier to run Bokeh in server mode so that you can use Python to write callbacks instead of having to revert to JavaScript:

$ bokeh serve --show real\_estate\_dashboard.py

Your dashboard should have widgets for filtering on the number of beds, number of baths, price, square footage, type of property, and whether the property is on an empty lot or not. You should plot the locations of the properties by their latitudes and longitudes using a scatter plot.

**Part III: Analysis of Geospatial Distribution of Properties**

You are now going to use your dashboard to analyze your data. Think of the city of Milwaukee and its surrounding areas. Think about urban, suburban, and rural areas. What differences might you expect to see in the residence types, square footage, number of beds and baths, and prices? Where would empty lots most likely be located? Where are the low-, middle-, and high-income neighborhoods located relative to the city? What differences in properties might you expect between these three types of neighborhoods (along with whether they are urban, suburban, or rural)?

I want you to use your dashboard to analyze the real estate data to try to identify properties of the real estate that differ across geographic regions. For example, you might expect to find most of the condos are near the city center. Using the groups of properties that you find, attempt to character the regions around Sacramento as urban, suburban, and rural and low, middle, and high income. Support your predictions with evidence from the real estate properties.

Write up your analysis in a separate report with supporting figures (mostly screenshots from your dashboard).

**Submission Instructions**

Submit your Jupyter notebook (PDF or HTML), dashboard source (zipped up .py file), and your report (PDF) through Canvas.

|  |  |
| --- | --- |
| Data Exploration, Cleaning, and Preparation: Columns were converted to the appropriate types. Outliers and "odd" values were identified. | 15% |
| Plots: appropriate types of plots were chosen for each analysis, axes are properly labeled, used correct axes for variables, points were colored as required, lines were coloring as required, used a legend if appropriate, chose appropriate axes limits to make plot readable and do not cause misleading interpretations, font sizes and figure resolution are legible. | 20% |
| Feature Engineering: Successfully created new variable with the appropriate type and values. | 5% |
| Bokeh Dashboard: The dashboard functions properly. Properties can be filtered by the variable values. | 20% |
| Report: Report is written in a professional manner using proper grammar and spelling. Report is a useful standalone document that can be shared with a business partner. | 20% |
| Reflection Questions: Reasonable attempts were made to answer each reflection question. Answers are supported by appropriate plots. | 20% |