SEAS 6414 - Python Applications in Data Analytics
Homework 5

Due Date: February 10, 2024 (10:00am EST)

Instructions: To complete the following task using Python, please download an Integrated Development Environment (IDE) of your choice. Ensure that your solution includes both the written code (input) and its corresponding output. Once completed, upload your solution in PDF format or any other format you prefer. **The questions are worth 10 points each**.

Homework Questions

1. Data Cleaning and Exploration

- Load the `zillow_feature_sample.csv` dataset using Pandas and report any missing values per column. Create a strategy to handle these missing values, justifying your approach.
- Generate a summary table that shows the mean, median, and standard deviation of `taxvaluedollarcnt`, `structuretaxvaluedollarcnt`, and `landtaxvaluedollarcnt` for properties built in each decade (1960s, 1970s, etc.).

2. Feature Engineering

- Create a new feature `Age` that represents the age of each property from the `yearbuilt` column, considering the dataset's latest `assessmentyear`.
- Develop a binary feature `HasPool` based on the `poolcnt` column, where 1 indicates the presence of a pool and 0 or NaN indicates no pool.
- Calculate and return the descriptive statistics for the age of the properties. Specifically, report the median age of the properties based on the `yearbuilt` and the latest `assessmentyear`.
- Generate and plot a bar chart of the counts of the binary feature "HasPool" created earlier. Set the y-axis to a logarithmic scale to better visualize the distribution of properties with and without pools.

3. Correlation Analysis

- Using NumPy, calculate the Pearson correlation coefficient between `bedroomcnt` and `bathroomcnt`. Visualize the correlation matrix of the numerical features of the dataset using a heatmap in matplotlib.

4. Geospatial Analysis

- Plot a scatter plot of `latitude` and `longitude` to visualize the geographical distribution of properties. Overlay this plot with a density estimate to highlight property clusters.

5. Market Value Analysis

- Visualize the trend of average `taxvaluedollarcnt` over the years using a line chart. Add a shaded area representing the 95% confidence interval for the average values.
- Create a boxplot to compare the distribution of `taxvaluedollarcnt` across different `buildingqualitytypeid`.

6. Tax Analysis

- Analyze the relationship between `taxamount` and `taxvaluedollarcnt` using a scatter plot and fit a linear regression line to it. Calculate the R-squared value for this fit.

7. Comparative Analysis

- For properties with `numberofstories` more than 1, compare the average `calculatedfinishedsquarefeet` against those with only 1 story using a bar chart.
- Compare the `taxvaluedollarcnt` for properties with and without a fireplace (`fireplaceflag`) using a violin plot.

8. Time-Series Forecasting (Advanced)

- Group the data by `yearbuilt` and calculate the annual mean of `landtaxvaluedollarcnt`. Using this time series data, create a forecast plot for the next 10 years with a rolling mean and standard deviation.

9. Amenities Impact Analysis

- Determine how the presence of a hot tub or spa (`hashottuborspa`) and air conditioning (`airconditioningtypeid`) impacts the `taxvaluedollarcnt`. Use a grouped bar chart to represent the average `taxvaluedollarcnt` for properties with and without these amenities.
- Investigate if there is a significant difference in the `calculatedfinishedsquarefeet` for properties with a basement (`basementsqft`) versus those without. Perform a hypothesis test and visualize the results using a histogram overlaid with the probability density function.

10. Neighborhood and Regional Analysis

- Group the properties by `regionidneighborhood` and plot a horizontal bar chart showing the top 10 neighborhoods with the highest average `taxvaluedollarcnt`.
- Using `regionidzip`, create a pie chart to display the proportion of total `taxamount` contributed by the top 5 zip codes. Include a separate 'other' slice for the remaining zip codes.