DSA 8010 - mini-project 2 (10 points)

Please submit your project through Canvas by 11:59 pm on Tuesday, September 24. You are encouraged to work in a group of up to 4 people, but you are not required to. If you work in a group, each person in the group should submit the report in Canvas. Include all group members' names at the top.

The purpose of this mini-project is to assess your ability to explore associations in a data set using descriptive methods.

Objectives

In this project you will demonstrate your ability to:

- Perform exploratory data analysis in R.
- Explore associations between two variables with summary statistics and plots.
- Draw accurate conclusions regarding associations between variables.
- Create a plot or table that explores associations among three variables.
- Think critically about data quality, missing values, and unusual features in a data set.

Data

The data for this project is the Pierce County House Sales data set from Mini Project 1. This data set represents home sales in 2020 in Pierce County, Washington. Please visit this site for variable names and definition:

https://www.openintro.us/data/index.php?data=

Assignment

Write a short report that describes associations between `sale_price` and other variables. You should use methods learned in Weeks 0-2 of the course.

Details

Part 1: Can sales price be predicted? Choose any three variables other than `sale_price`. For each variable, explore its association with sale price using plots, tables, and summary statistics. Include at least one plot or table for each variable. Report appropriate statistics for each variable. Summarize your findings in 2-3 sentences for each variable.

Part 2: Is missingness informative? Create a binary variable that indicates whether the `view_quality` is missing. Missing values in this data file are represented as an empty character string; that is, `''`. Choose three variables. (They may, but do not need to be, the same variables from part 1.) Explore the association between these variables and the binary variable that you created. Follow the same instructions as in Part 1.

Data quality and unusual features. In your analysis, pay attention to outliers, missing values, and excessive zeros. In both questions, include in the text a short description of any of these features and describe the impact that they may have on your analysis.

Other instructions

Your analysis should be presented as a report, with tables and plots embedded in the report and formatted to be appropriately sized and readable. Make sure to write clearly. It is fine to write in a conversational tone. Write the report in R Markdown and submit the document in an html, doc, or pdf file. The text for each question does not need to exceed 1-2 paragraphs.

The focus of this project is descriptive analysis; therefore, it is not encouraged to use statistical methods that we have not covered yet in class (e.g. linear regression and tests of significance). Please be aware that using these methods incorrectly may result in deductions of points. If you use any outside resources, make sure to include an appropriate citation. For websites or code resources, include as many of the following pieces of information that are available: url, name of author, date accessed, name of webpage.

Rubric

Your project will be graded according to this rubric. Make sure to follow the specific guidelines about the required plots, statistics, and discussion elements.

Category	Points
The presentation style. The results are presented clearly in a readable, narrative form. The writing is clear and concise. All quantitative results, including plots, tables, and summary statistics are referenced and interpreted in the text. Captions, graph titles, and axis labels are included where appropriate. Excessive computer output is not included in the report.	1
Appropriate methods are used to assess associations.	3
Appropriate conclusions are drawn from statistics and plots.	3
The required number of plots, tables, and summary statistics are included.	1
The required number of variables are selected for analysis.	1
The text includes an exploration of outliers, excessive zeros, and missing values and, if applicable, a discussion of their potential impact on overall data quality.	1