First Assigment

Due to: 18 oct 2023

Statistical Analysis

We choose 4 datasets for your assignment and you are supposed to do only 2/4 problems(US Accidents and Microsoft Malware involves bonus points. obviously are more challenging datasets too). Use descriptive statistics and inferential statistics to write report. For inspiration, we write some questions about datasets. Use various types of statistical tests and questions for your report.

Problem 1.

House Prices

The house price dataset provides comprehensive information on residential properties, including key features such as square footage, number of bedrooms and bathrooms, location, and sale prices. This dataset is a valuable resource for real estate market analysis, helping to understand housing market trends and factors influencing property values. (use only the train part of the data for your report.)

- What are The building class and why it is important?
- How does the overall quality (OverallQual) of a house relate to its sale price?
- How do the different types of heating (Heating) affect the sale prices?
- How do the different types of utilities (Utilities) available in a property relate to sale prices?

Download dataset

- House Prices

Problem 2.

Top songs on spotify

A comprehensive collection of 10,000 of the most popular songs that have dominated the music scene from 1960 to the present day. It's a musical journey that resonates with listeners of all ages, transcending genres and striking a chord with the soul.

Download dataset

- Top 10000 Songs on Spotify 1960-Now

Problem 3.

US Accidents

This is a countrywide car accident dataset that covers 49 states of the USA. The accident data were collected from February 2016 to March 2023. The dataset currently contains approximately 7.7 million accident records.

Download dataset

- US Accidents (2016 - 2023)

Problem 4.

Microsoft Malware

Once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways. With more than one billion enterprise and consumer customers, Microsoft takes this problem very seriously and is deeply invested in improving security.As one part of their overall strategy for doing so, Microsoft is challenging the data science community to develop techniques to predict if a machine will soon be hit with malware(you are supposed to only analyse dataset. No prediction:). Only use the train part of the dataset for your report.

Download dataset

- Microsoft Malware Prediction

Notes

Note:You should consider new additional creative questions for both tasks and answer them with proper statistical tests. Each task should have its report and IPython Notebook. Once again, we emphasize the report; it should contain all your questions and your proper statistical answers. Use figures, pictures, and tables. DO NOT PUT ANY CODE IN THE REPORT.