STAT394 Group Project Report2

Ken MacIver, Tom Tribe, Jundi Yang, Mei Huang

2022-08-11

Contents

1 Second Project Report (Milestone 2)

1.1 Introduction

We have selected the 'diamonds' dataset for our Group Project. It contains 53940 rows of data on different diamonds. There are three categorical variables: cut, clarity and color, and seven numerical variables: carat, depth, table, price, and measurements of the diamonds x, y, z.

1.2 Platform and Data

The dataset is freely available on the Kaggle website. It can be located at "Diamonds Dataset, Kaggle.com" (2016). There is also a .csv version available in the 'Data' folder in the GitHub repository for this project.

Categorical data:

- cut: levels of diamonds proportion, symmetry and polish
- clarity: levels of clarity
- color: color of the diamonds

Numeric data:

- carat = weight in carat
- depth = the distance between the culet and the table in millimeters
- table = width of the table facet divided by the width of the diamond in percentage
- price = price of the diamonds
- x = length in mm

- y = width in mm
- z = depth in mm

1.2.1 Initial Goals:

As we have no expertise in subject area of the dataset we will be relying on our exploratory data analysis (EDA) to determine the exact direction of our investigation. A possible initial guiding question could be to investigate which of the variables is most predictive of diamond price, which seems like the natural response variable.

Reference list

"Diamonds Dataset, Kaggle.com." 2016. https://www.kaggle.com/datasets/shivam2503/diamonds.