Data quality

Sabin Basnet

MSDS-631-A01 R Programming

Dr. Rajan Thapaliya

September 8th, 2024

1. **There is a surprisingly cheap 5 carat diamond, and some cheap 3 carat diamonds. How can we identify those points?**

We identify those points by filtering out the diamonds based on their carat weight and price and then plotting them to observe outliers.

A screenshot of a computer program

Description automatically generated

A graph showing a diagram of diamonds

Description automatically generated

1. **Use an interactive scatterplot to identify outliers in these variables. Check prices, carat and other information and think about if any of the outliers can be due to data errors. A blue screen with white text

   Description automatically generated**

A chart with yellow and green dots

Description automatically generated

1. **Discuss the following:**
   1. **How can you tell if the data is an outlier or if it is something important?**

As we know outliers are the points that deviate significantly from other data. To find the price outliers we need to compare diamonds of similar carat, color, and clarity. Also, we can use boxplots or calculate z-scores to detect extreme values.

For Carat vs Price outliers, if a diamond with a high carat weight is priced significantly lower than expected, it could either be a low-quality diamond or some error in data that is incorrectly entered or missing features like poor cut or clarity.

* 1. **Which data is the noise and how is the noise different from outliers?**

Noise means random error or irrelevant variations in the data, which can distort the analysis but doesn’t provide useful information. And outliers are the extreme values that may hold significant meaning for example unique cheap diamonds.

Outliers often follow patterns or relationships with other variables, while noise typically does not. For example, a 5-carat diamond priced at an unusually low value could be an outlier if other variables cut or clarity justify the lower price. Noise, on the other hand, might appear as random, inexplicable anomalies without correlation to other variables.

1. **When there are missing values, explain the pros and cons of the following strategies:**
   1. **Elimination of Data Objects**

The pros of the elimination of data objects are that it simplifies the analysis by removing incomplete data which can improve the quality of results, and the cons are that it reduces data size and potentially discards valuable information, especially if a large portion of the data has missing values.

* 1. **Estimation of Missing Values**

The pros of the estimation of missing values it that it retains dataset size by allowing for more comprehensive analysis and the cons are that imputation methods like mean median may introduce biases which may lead to inaccurate insights.

1. **What are the limitations of analyzing real data with missing values and why is it impossible to really know such data?**

Real-world data often suffers from missing or inaccurate information, which can introduce limitations to the analysis. If missing values are not random, removing or imputing them can introduce bias into the results. Imputed values are estimates and may not accurately represent the actual values, leading to reduced confidence in the findings.

Additionally, there may be hidden factors causing missing data that are not captured in the analysis. It's important to remember that even the cleanest datasets may contain undetected errors or omissions due to limitations in data collection methods, making it impossible to fully know the true nature of real data.

**Conclusion**

In this project, we used the diamonds dataset from the ggplot2 package to identify outliers and noise in diamond prices, focusing on unusually cheap 3-carat and 5-carat diamonds. Outliers were visualized through interactive scatterplots using plotly, and strategies for handling missing data were discussed. While real-world data analysis is often imperfect, careful handling of outliers and missing values can lead to more accurate insights.



University of the Cumberlands

School of Computing and Information Sciences

# Certification of Authorship

**Submitted to**: Dr. Rajan Thapaliya

**Course**: MSDS 631- A01 R programming

**Student’s Name**: Sabin Basnet

**Date of Submission**: 09/08/2024

**Purpose and Title of Submission**: Data quality

**Certification of Authorship**: I hereby certify that I am the author of this document and that any assistance I received in its preparation is fully acknowledged and disclosed in the document. I have also cited all sources from which I obtained data, ideas, or words that are copied directly or paraphrased in the document. Sources are properly credited according to accepted standards for professional publications. I also certify that this paper was prepared by me for this purpose.

**Student’s Signature**: Sabin Basnet