Youan Bi D Rubinel

Elena West

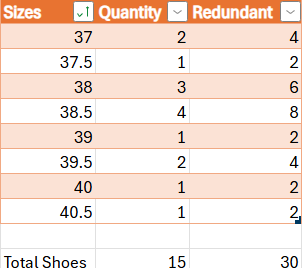
Ezekial Curran

Manager's question? "What sizes of shoes do you need and how many?

Answer:

This table shows how many shoes of each size to buy. The second column (Quantity) is the quantity of shoes you can buy if you want to only buy 15. The second column (labelled “Redundant”), shows how many shoes of each kind you should buy if you are willing to buy 30 instead.

Part 2



**Thought questions:**

* What problems might you run into using this approach?
* What assumptions about the data might cause you problems? The roster? The players?
* What biases may be in the data?
* What other thoughts came up in your discussion?

Problems that may occur from the data established is the existence of variation in the data which may result in inaccurate estimations of shoes. To compensate for this we have doubled the amount of each shoe size from the range of sizes while still staying within 30 shoes.

Assumptions about the data were made relating to accurate information for the volleyball roster players. Other assumptions are made on the data collected that it was collected correctly. There were also predictions made on other data collected relating to women’s height and shoe sizes.

One of the biases could be using non-athlete people data as opposed to using athlete data.

We noticed that the data was in European shoe sizes. We also noticed some strange outliers and a null row. We also noticed that there were more women than men.

For our prediction, we made the assumption that the data was collected correctly. This is never guaranteed and can always skew results.

We used a linear regression model, and the r-squared was under .5 and not very good.

A short summary of the insights you gained from both parts of the activity and how it relates to machine learning (Minimum of 200 words)

The first helped us navigate through a row data and get some quick insights about the dataset composition.

We learned about interpreting statistical results into answers to questions posed by a manager. It takes extra communication skills to translate technical explanations into interpretable results. That relates to machine learning in that many people are very excited about AI right now. Many clients will want to get predictions, and many of them will have no idea what AI is. We will need to know how to interpret our results to people who don’t know what exactly we do.

Some insights gained about the data is the value of clean data. Models and predictions can quickly deviate or mislead when data has significant variation or a substantial amount of outliers. The ways data can be interpreted is generally fairly open and it is best to adjust for the desired outcome. For predicting whether a shoe and height are a woman or a man, methods of classification were used on the datasets. Methods of regression was then used for anticipating shoe sizes for the players on the roster and how many shoes of each size to get.

We learned how to use linear regression in R. We also learned the basics of navigating RStudio.