Option #1: Hot vs. Cold Cereals

Scott Miner

Colorado State University – Global Campus

**Option #1: Hot vs. Cold Cereals**

The assignment asks the student to explore the “*Cereals.csv*” file, using the code in chapters 3 and 4 of *Data Mining for Business Analytics* by Shmueli et al. (n.d.) as a template for exploratory data analysis (EDA). The student is to provide summary statistics for the data frame, including the mean, standard deviation, minimum, maximum, median, length, and the number of missing values for all quantitative variables. Additionally, the student will plot histograms for all quantitative variables, using both the “hist()” and “ggplot()” functions. Lastly, the student will provide a side-by-side boxplot comparing calories in “cold” versus “hot” cereals.

Statistician Karl Pearson coined the term “histogram” in 1892 to refer to a graphical representation of data that displays frequency counts, corresponding to bin ranges (Ioannidis, 2003). We can learn three things from histograms: (a) the distribution shape, (b) span, and (c) whether the distribution contains outliers. The shape tells us whether data are symmetric (e.g., left- or right-skewed). The span, which we convey via bins, tells us the range, along with the minimum and maximum approximations. Lastly, outliers, represented by extreme high and low values, inform us of data anomalies, including unusual cases and entry errors (Editor, n.d.). Tukey introduced the boxplot in 1977 to graph univariate data (Ospina et al., 2014). Boxplots comprise five key statistics: (a) “minimum,” (b) first quartile, (c) median, (d) third quartile, and (e) “maximum” (Galarnyk, 2019). From a boxplot, we again learn: (a) the distribution shape, (b) span, and (c) information on outliers. Additionally, we learn (d) the above-stated summary statistics, and (e) how tightly coupled the data are. The boxplot in Figure 6, for instance, tells us very few cereals are of type “hot,” and the median and average for this subset are the same. The remaining screenshots in this paper provide the code and output to generate the summary statistics and data visualizations, including histograms and boxplots, for the *cereals.csv* data set.

A screenshot of a social media post

Description automatically generated

Figure . Summary Statistics for the cereals.csv data set, including the mean, standard deviation, min, max, median, length, and missing values for all numeric columns.

A screenshot of a computer

Description automatically generated

Figure . A histogram with labels displaying the frequency counts of sodium in the cereals.csv data set and the corresponding bins and code.

A screenshot of a cell phone

Description automatically generated

Figure . The code to provide histograms for quantitative variables # 1 - 6 in the cereals.csv data set (Part 1 of 3).

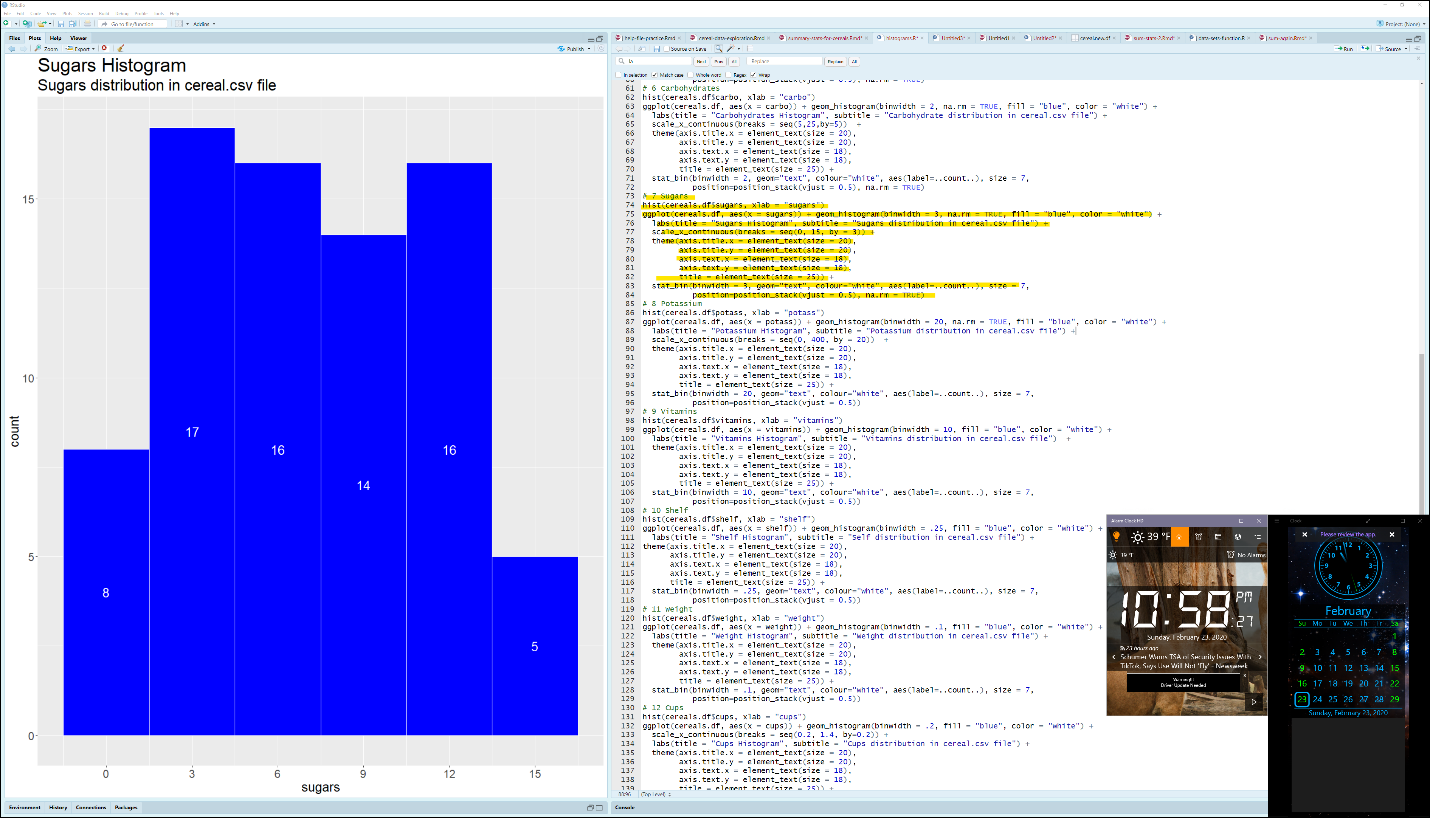


Figure . The code to provide histograms for quantitative variables # 6 - 11 in the cereals.csv data set (Part 2 of 3) with an additional histogram showing the sugars variable.

A screenshot of a computer

Description automatically generated

Figure . The code to provide histograms for quantitative variables # 11 - 13 in the cereals.csv data set (Part 3 of 3) with an additional histogram showing the rating variable.

A screenshot of a computer

Description automatically generated

Figure . Boxplots displaying calories by cereal type for two types of cereals, hot and cold, with corresponding labels, including the mean, median, and axes. Notice the median and the average are the same for the "hot" type cereal. There were only three records with cereal type “H.”

A screenshot of a social media post

Description automatically generated

Figure . Zooming in on a portion of code that provides the sodium histogram plot using both the hist() and geom\_histogram() functions.

References

Editor, M. B. (n.d.). 3 Things a Histogram Can Tell You. Retrieved February 23, 2020, from <https://blog.minitab.com/blog/3-things-a-histogram-can-tell-you>

Galarnyk, M. (2019, November 7). Understanding Boxplots. Medium. https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Ioannidis, Y. (2003). - The History of Histograms (abridged). In J.-C. Freytag, P. Lockemann, S. Abiteboul, M. Carey, P. Selinger, & A. Heuer (Eds.), Proceedings 2003 VLDB Conference (pp. 19–30). Morgan Kaufmann. <https://doi.org/10.1016/B978-012722442-8/50011-2>

Ospina, R., Larangeiras, A. M., & Frery, A. C. (2014). Visualization of Skewed Data: A Tool in R. Visualización de Datos Sesgados: Una Herramienta En R., 37(2), 399–417. https://doi.org/10.15446/rce.v37n2spe.47945

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., Lichtendahl, K. C., & Jr. (n.d.). Data Mining for Business Analytics. Retrieved from <https://platform.virdocs.com/r/s/0/doc/503437/sp/21743563/mi/74416504?cfi=%2F4%2F2%2F10%2F20%2F28%2C%2F1%3A386%2C%2F1%3A386>