Capstone Project Option 1, Module 8: Final Report

Sean Senko

MIS581 – Capstone – Business Intelligence and Data Analytics

Colorado State University – Global Campus

Dr. Osama Morad

October 4, 2020

Capstone Project Option 1, Module 8: Final Report

**Abstract**

Peterbrooke Chocolatier is a chocolatier founded in the San Marco district of Jacksonville, Florida, and which has expanded to various other locations. The Peterbrooke Chocolatier company is attempting to expand its selection of chocolates by looking for a potentially superior recipe of chocolate. It is doing this through analysis of the Kaggle Chocolate Bars 2020 data set as general guidelines. Analysis attempted to determine whether cocoa beans from a specific country of origin, or chocolates bars with a specific cocoa percentage in their composition, were superior to others.

**Research Hypothesis**

Given the data in the Kaggle Chocolate Bars 2020 data set, there exists a potential formulation for chocolate in terms of either the origin of the cocoa beans or the cocoa percentage within the chocolate bars which is of superior quality (determined by rating). The goal is to determine if one or both of these statements holds true; if both statements prove false, then it simply means there is inconclusive evidence from the data set to support the idea that a superior formulation exists from what has been determined with the data.

**Introduction**

Peterbrooke Chocolatier is a chocolatier founded in the San Marco district of Jacksonville, Florida, and which has expanded to various other locations. The Peterbrooke Chocolatier company offers a variety of chocolate-based products, which include hand-dipped chocolate Oreos, chocolate potato chips, chocolate graham crackers, pretzel rods, pretzel twists, Nutter Butter biscuits, and even popcorn and caramel corn. The Peterbrooke Chocolatier also currently offers four types of chocolate bars for sale: white chocolate bars, milk chocolate bars, dark chocolate bars, and 72% cocoa dark chocolate bars. Peterbrooke Chocolatier also offers a variety of traditional mixed chocolate boxes. As vast quantities of chocolate in the of billions of kilograms per year are consumed, with China alone consuming roughly 6.9 billion kilograms in the year 2012 according to Llanos-Contreras et al. (2020), there is undoubtedly a market for chocolate consumption. The question is how Peterbrooke Chocolatier might take advantage of this fact given their current product range.

Given the wide range of chocolate products offered by Peterbrooke Chocolatier, it could be worthwhile attempting to find another chocolate recipe to use in their products. However, it is important to remember that changing products or introducing new products does not always help businesses; as Grumet (2009) recalls, in April of 1985 the Coca Cola Company launched a new variation of its signature soft drink which they called “New Coke”, and advertised it as a better product. Consumers disagreed however, and the Coca Cola Company chose to withdraw their new product from the market three months after its debut.

**Overview of study**

The purpose of this capstone study is to determine if a potentially superior chocolate product could be created by using the data from the Kaggle Chocolate Bars 2020 data set as a guideline. This data set was chosen because it is more up to date and contains more entries than an older, similar data set on Kaggle titled “Chocolate Bar Ratings”. Also, the fact that the data is available freely to the public made the data set an attractive option. It may prove possible, given the nature of the data, to determine that having a certain percentage of cocoa in the chocolate has a better chance of creating superior chocolate. It might also be shown that cocoa beans from specific country origins produce superior chocolate. The hope of this study is to find potentially superior cocoa beans or a potentially superior percentage of cocoa in chocolate so far as quality is concerned in the rating of the bars in the data set.

Therefore, an overall research question was developed which said, “Do either the cocoa percentage in the chocolate or the country of origin for the cocoa beans make a significant impact on the quality of the chocolate?”. This was broken into two smaller questions, namely, “Does the cocoa percentage in the chocolate make a significant difference in the overall quality of a batch of chocolate?” and “Does any particular country produce superior cocoa beans compared with others?”. The null hypothesis of the first question states that the cocoa percentage in a batch of chocolate has no significant impact on the overall quality of the chocolate. The alternate hypothesis states the opposite, namely that the cocoa percentage in a batch of chocolate does have a significant impact on the overall quality of the chocolate. The null hypothesis of the second question is as follows: there is no statistically significant difference in the overall quality of batches of chocolate made from cocoa beans in different countries. The alternate hypothesis of the second question states that there is a statistically significant difference in the overall quality of batches of chocolate made from cocoa beans in different countries such that the batches of chocolate made from beans from one or more countries of origin are significantly better quality than others.

**Literature Review**

The overall quality of a batch of chocolate can come from a variety of factors, such as the presence of antioxidants (Meng et al., 2009), the addition of milk powders (Liang and Hartel, 2004), and even the process of manufacturing the chocolate itself (Barišić et al., 2019). It is also possible to partially replace the wheat flour in a chocolate cookie recipe with a formulation of overripe banana pulp powder to add dietary fiber to the cookies (Ng et al., 2020), or make other alterations to the recipes to add or remove ingredients. Therefore, it is quite possible that superior quality chocolate products could be created with the right ingredients and/or the use of specific procedures.

**Research Design**

This capstone study was conducted using Python and Rstudio, using a quantitative approach to the research (O’Leary, 2017). The quantitative approach to research involves the assumptions of positivism and empiricism, uses the scientific method and more specifically a methodology which is hypothesis driven, deductive, reliable, valid, reproducible, objective, and generalizable. The methods involved in the quantitative approach to research are large in scale, and done either through surveys or a randomized controlled trial. Analysis is through statistics. Such is the case in this study.

**Methods**

The specific method of research conducted within this capstone study was experimentation. Two experiments were conducted using Rstudio: an experiment to determine if there was any percentage of cocoa in the chocolate bars in the data set which produced a chocolate bar on average of statistically significant high quality (as formulated through the mean of the ratings of each percentage of cocoa in the bars, adjusted by the standard deviation), and an experiment to determine if there were cocoa beans from a specific country of origin which produced a chocolate bar on average of statistically significant high quality (as formulated through the mean of the ratings of each country of origin adjusted by the standard deviation). Additional analysis was conducted using Python.

**Limitations**

The Kaggle Chocolate Bars 2020 data set may be incomplete. Some potential conclusions are impossible to make given the size and entries of the data set. Future research might be possible regarding 50 percent cocoa chocolate bars, but the data set only contains one entry so the data is inconclusive about the statistical significance of the quality of those bars of chocolate.

**Ethical considerations**

The data set is freely available to the public. The potential dilemma of informed consent is not present because the data does not involve humans as subjects. However, Peterbrooke Chocolatier is still obligated to maintain the usual security, privacy, and ethical standards as would be required in other studies of this nature. For example, even though the initially processed data is public, the analyzed data needs to be protected so as to not lose work which has already taken place. Furthermore, there is the potential consideration of four common elements of a framework for big data ethics based on the work of Davis (2012). These elements include identity (what relationship exists between somebody’s offline identity and online identity?), privacy (Who should control access to the data?), ownership (Who owns data, can rights to it be transferred, and what are the obligations of people who generate and use that data?), and reputation (How can it be determined that the data is trustworthy?). One caveat exists: identity and privacy should not be confused with one another. Identity has to do with anonymity or otherwise being identifiable, whereas privacy covers the use of and access to data.

Clearly, identity is not a factor in this study. Privacy is a concern so far as keeping unwanted users from accessing, modifying, and deleting the edited data used in the project. Ownership will involve Peterbrooke Chocolatier owning the edited data; presumably, rights to it are nontransferrable. The obligations of anybody generating and using the data are to keep it safe from unwanted access and to handle it appropriately. Finally, the reputation of the data set is unknown.

Additionally, one way that security may be maintained within visual data analytics is to use disconnected data sources so that a data breach in one source will not compromise the security of another data source. If a presentation is being delivered, then perhaps the data might be presented from a local source instead of using the cloud. For example, a presenter might choose to use an external USB stick drive which contains the data utilized within the presentation instead of relying on downloading the presentation and its data from an external source like Dropbox or Google Drive.

**Findings**

The initial analysis in Python showed that there were thirteen distinct ratings, all separated in quarterly segments save for 2.6, on a scale of 1 to 4. Subsequent analysis in Rstudio revealed that there were chocolate bars with 46 distinct percentages of cocoa powder within the recipes, ranging from 42.0 to 100.0. Furthermore, there were 62 distinct countries of origin for the cocoa beans. The mean and standard deviation of the rating column were calculated; these were 3.198561 and 0.434329 respectively. Analysis of the chocolate by origin of the cocoa beans found that chocolate bars with cocoa beans having an origin in Puerto Rico have a statistically significant below average quality as determined by rating; cocoa beans from all other origins either have insufficient data points, or fall within the standard deviation. Additional analysis of the chocolate by the percentage of cocoa in the formula was largely inconclusive; the overall data was inconclusive about the quality of chocolate which was made of 50% cocoa, which had one data point both above the mean and a standard deviation. Chocolate bars of different percentages of cocoa either had insufficient data or fell within one standard deviation of the mean.

**Conclusion**

The null hypothesis of the question, “Does any particular country produce superior cocoa beans compared with others?” was proven due to a lack of support for the alternate hypothesis of the question. In other words, there was not found to be any country or set of countries with significantly superior quality cocoa beans compared the rest. Additionally, the null hypothesis of the question “Does the cocoa percentage in the chocolate make a significant difference in the overall quality of a batch of chocolate?” was proven from a lack of support for the alternate hypothesis in the analysis conducted. The alternate hypothesis stated that the cocoa percentage makes a significant difference in the overall quality of a batch of chocolate, and there was insufficient evidence to support this hypothesis. As the null hypothesis of both of these questions was shown from lack of support for either alternate hypothesis, the answer to the overall research question is that there is insufficient evidence from the Chocolate Bars 2020 dataset from Kaggle to produce a superior quality chocolate recipe. This does not mean that a superior formulation does not exist; however, the data did not show that a superior formulation existed within the data itself.

**Recommendations**

If Peterbrooke Chocolatier wishes to find a superior chocolate recipe through data, it will most likely have to look elsewhere. Although there are other columns within the Kaggle Chocolate Bars 2020 Data Set, they are less likely to be of significance. Therefore, the search would most likely continue elsewhere.

References

Barišić, V., Kopjar, M., Jozinović, A., Flanjak, I., Ačkar, Đ., Miličević, B., . . . Babić, J. (2019).

The chemistry behind chocolate production. *Molecules, 24*(17) doi:http://dx.doi.org.csuglobal.idm.oclc.org/10.3390/molecules24173163

Davis, K. (2012). *Ethics of Big Data*. Sebastopol, CA: O’Reilly Media, Inc.

Grumet, L. (2009). Stretch the brand, don't lose the equity: Certified public accountant. *The CPA Journal, 79*(1), 7. Retrieved from https://csuglobal.idm.oclc.org/login?url=https://www-proquest-com.csuglobal.idm.oclc.org/docview/212271514?accountid=38569

Liang, B., & Hartel, R. W. (2004). Effects of milk powders in milk chocolate. *Journal of Dairy*

*Science, 87*(1), 20-31. Retrieved from https://csuglobal.idm.oclc.org/login?url=https://www-proquest-com.csuglobal.idm.oclc.org/docview/195825653?accountid=38569

Llanos-Contreras, O., Baier-Fuentes, H., & Valdés, C. Y. (2020). ROGGENDORF: THE CHALLENGE OF INTERNATIONALIZATION OF A FAMILY ENTERPRISE. *Academy of Entrepreneurship Journal, 26*(2), 1-15. Retrieved from https://csuglobal.idm.oclc.org/login?url=https://search-proquest-com.csuglobal.idm.oclc.org/docview/2424655909?accountid=38569

Meng, C. C., Jalil, A. M. M., & Ismail, A. (2009). Phenolic and theobromine contents of commercial dark, milk and white chocolates on the Malaysian market. *Molecules, 14*(1), 200-209. doi:http://dx.doi.org.csuglobal.idm.oclc.org/10.3390/molecules14010200

Ng, Y. V., Alina, T. I. T., & Rosli, W. I. W. (2020). Effect of overripe banana pulp incorporation on nutritional composition, physical properties, and sensory acceptability of chocolate cookies. *International Food Research Journal, 27*(2), 252-260. Retrieved from https://csuglobal.idm.oclc.org/login?url=https://search-proquest-com.csuglobal.idm.oclc.org/docview/2419451369?accountid=38569

O’Leary, Z. (2017). The essential guide to doing your research project (3rd ed.). Thousand Oaks,

CA: Sage Publishing.