Exercise: 6.1

**Hypothesis**

**‘**Chocolate Bars whose cocoa bean content comes from South America are superior’

**Data source –** This Data was collected by chocolate tasting experts known as the ‘Manhattan Chocolate Society’. They are industry experts who taste the chocolates and interview producers and manufacturers about how the chocolate is made. The data was made available through Kaggle.

**Data Set** - flavors\_of\_cacao.csv

**Data Collection –** This is a form of administrative data; the data is collected by tasting experts and then gathered into a central system. It is updated regularly, with multiple entries every year and ongoing. The data is collected by members of the ‘Manhattan Chocolate Society’ via, personal examination of product and interviews with manufacturers.

**Data Overview –** This data shows Company name, Origin of the bean, Type of bean, Name of bar, Review date, and Country manufactured in. Data ranges from 2006-2017.

**Limitations –** This data is collected Manually thus there is room for human error, then this data needs to not only be collected but needs to be transferred to a digital storage tidied and organized so it may be used. This could create a timelines issue as we don’t know how long it takes after testing to be uploaded into the system. However, we do know that they have multiple updates a year according to the data. This is also tested by humans so undoubtedly there will be some level of personal bias due to everyone’s tastebuds being different flavours appealing to different people.

Only 1 chocolate is tasted out of a batch and not tested again after thus it may unrepresentative as some chocolates after they have been tested might change the recipe for that specific bar and if it is not tested again its previous result is unrepresentative.

**Relevance –** This data is very relevant to my hypothesis, the contents of cocoa, Origin of bean, manufacturing country and name of manufacturer create a strong platform to find an answer to my hypothesis through hypothesis testing.

**Ethical Considerations –** This data contains no PII issue as it contains no consumer data it is specifically data about chocolate bars and their Manufacturers. There could however be an element of bias as its humans judging what chocolate is better. So, a personal Bias may have affected unintentionally the ranking system.

**Data Profile**

**Data Cleaning and Consistency Checks**

1. **Columns Removed**

* **Bean Origin (City) –** Contained 3 different variable in one column that don’t help my EDA process thus decided to remove.
* **Type of Bean –** over 50% of column was null.

1. **Missing Values**

* Manufacturer - 0
* Country of Manufacturer - 0
* Cocoa (%) - 0
* Type of Bean – 887 missing, 1 NaN (Removed Column)
* Bean Origin (Country) – 73 Missing, 1 NaN (Replaced All With NaN)
* Rating – 0
* Review Year – 0
* REF – 0

1. **Duplicate Values**

20 Duplicate values found and Removed

1. **Data Description**

|  |  |  |
| --- | --- | --- |
| **Column** | **Description** | **Data Type** |
| Manufacturer | Manufacturer of the chocolate bar | Object |
| Country of Manufacturer | Country the manufacturer is from | Object |
| Cocoa (%) | Cocoa % in the chocolate bar | Float64 |
| Bean Origin (Country) | Country cocoa bean is form | Object |
| Review Year | Year chocolate was revied by Manhattan chocolate society | Int64 |
| REF | The higher the number the more recently it was reviewed | Object |

1. **Descriptive Analysis**

**A table with numbers and a number of ones

Description automatically generated with medium confidence**