Project Checkpoint 2 (Status Report)

**Updates** **to Project Checkpoint 1:**

* Unable to locate specific data for whisky sales over time. Removed time series analysis.
* Solidified team roles, data, and analysis (below)

|  |  |
| --- | --- |
| **Team Member** | **Role** |
| Brooke McCurdy | Exploratory data analysis, visualizations, presentation |
| Kelsey Johnson | Natural Language Processing |
| Lee Scott | Linear regression |
| MJ Schultz | Clustering |
| Sam Hochheimer | Data cleaning, exploratory data analysis, visualizations |

**Data**:

1. Scotch Whisky Dataset
   * Attribute scores (bold, smokey, sweetness, etc.) of 86 malt scotch whiskies
   * 86 rows x 17 columns (16 attributes likely to be used)
   * Thatdataanalyst. (2018, March 22). *Scotch Whiskey Dataset* [data set]. Kaggle. https://www.kaggle.com/koki25ando/scotch-whisky-dataset
2. Meta-Critic Whisky Database
   * Critic ratings, cost, and location data affiliated with specific whisky
   * 1788 rows x 10 columns
     + 5 attributes likely to be used: Whisky, Meta Critic rating, Cost, Country, Type)
   * Selfbuilt. (2021, August 27). Meta-Critic Whisky Database. https://whiskyanalysis.com/index.php/database/
3. 2,2k+ Scotch Whisky Reviews
   * Brand, type, price, and reviews
   * 2245 rows x 7 columns
   * Thatdataanalyst. (2018, June 13). 2,2k+ Scotch Whisky Reviews. https://www.kaggle.com/koki25ando/22000-scotch-whisky-reviews

**Project Checkpoint 2:**

|  |  |
| --- | --- |
| **Specification** (Problems/Questions): | **Analysis** |
| 1. Can cost be predicted based on attribute scores of whiskies? | Linear Regression |
| 1. Can ratings be predicted based on attributes such as alcohol percentage, brand, and/or country (among other things)? | Linear Regression |
| 1. What are customer reviews telling us about whisky preferences? | k-means Clustering |
| 1. What recommendations can be made to consumers (based on specific preferences, attributes, etc.)? | NLP |

**Observation:** Significant data cleansing required, including splitting columns (i.e., column in 2,2k+ dataset has both whisky name and alcohol percent), matching whisky names, and potential averaging of prices across the larger dataset to result in one price per type of whisky. More to share as analysis moves forward.

**Recommendation:** Continuing weekly group meetings to check in on progress; continued analysis; continued review of each other’s work through sharing script through Colab; PowerPoint presentation to be generated