Group_scotch_Project

Levi Quintero

2025-09-15

Contents

1 Introduction

2 Methods

2.1 Exploratory Data Analysis (EDA)

All data analysis were conducted in R version 4.4.3 (R Core Team 2025). We initially assessed if the data derived from Shand et al. (2017) (Table 1) followed a multivariate normal distribution ($X \sim N_{11}(\mu, \Sigma)$) via visual comparisons of observation Mahalanobis distances ($d_M^2(X, \mu)$) to their expected quantiles and a QQ plot line (Fig. x), observation density plots by factor (Fig. x), as well as by conducting a Henze-Zirkler Test of Multivariate normality ($HZ = 1.325, p < 0.001$; Table 1). As the data clearly did not conform to $X \sim N_{11}(\mu, \Sigma)$, we log transformed the observations and re-aplied the same analysis (Fig X2; Table 2) with results now conforming to $X \sim N_{11}(\mu, \Sigma)$ with no outliers ($d_M > 0.99$). This log-transformed data was henceforth used in this analysis.

```
## Warning in attr(x, "align"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

```
## Warning in attr(x, "format"): 'xfun::attr()' is deprecated.
## Use 'xfun::attr2()' instead.
## See help("Deprecated")
```

Table 1: Whisky Origin and Chemical Data

Sample_no	Descriptor	Distillery	P	S	Cl	K	Ca	Mn	Fe	Cu	Zn	Br	Rb
1	Blend	Baile Nicol Jarvie	0.152	1.100	0.173	7.860	1.450	0.032	0.027	0.186	0.015	0.002	0.006
2a	Blend	Bells	0.653	1.580	0.238	4.930	1.400	0.019	0.110	0.242	0.021	0.005	0.003
3a	Blend	Chivas	0.375	0.809	0.193	4.310	1.220	0.019	0.044	0.196	0.007	0.003	0.002
4a	Blend	Dewars	0.121	1.160	0.157	3.200	1.140	0.011	0.050	0.189	0.018	0.003	0.003
5a	Blend	Johnnie Walker	0.326	1.090	0.180	5.480	0.526	0.018	0.103	0.286	0.020	0.002	0.002
6a	Blend	The Famous Grouse	0.145	0.615	0.097	2.740	0.416	0.009	0.050	0.208	0.007	0.002	0.001
7a	Blend	Whyte and Mackay	0.067	0.576	0.151	2.360	0.745	0.012	0.047	0.159	0.019	0.003	0.002
8a	Blend	William Grant	0.239	0.748	0.147	2.840	0.976	0.010	0.021	0.137	0.020	0.003	0.002
9a	Counterfeit	Unknown 1	0.089	4.060	0.066	0.336	1.240	0.007	0.154	0.085	0.038	0.005	0.001
10a	Counterfeit	Unknown 2	0.088	14.700	0.072	1.230	1.400	0.006	0.025	0.052	0.018	0.004	0.001
11a	Counterfeit	Unknown 3	0.279	15.900	0.083	0.811	1.360	0.006	0.057	0.038	0.016	0.002	0.002
12a	Counterfeit	Unknown 4	0.320	22.100	0.596	2.320	1.780	0.008	0.019	0.038	0.015	0.068	0.001
13a	Counterfeit	Unknown 5	0.120	26.100	0.071	2.370	1.630	0.010	0.082	0.187	0.194	0.012	0.005
14a	Grain	Grain matured	0.034	2.230	0.252	6.440	1.040	0.013	0.115	0.174	0.019	0.004	0.006
15a	Grain	Grain unmatured	0.084	5.530	0.113	3.250	1.350	0.012	0.076	0.164	0.046	0.010	0.003
16	Highland	Glengoyne	1.040	5.570	0.343	24.200	0.857	0.023	0.197	1.251	0.041	0.004	0.016
17	Highland	Glenmorangie	0.126	0.796	0.245	6.950	0.859	0.035	0.025	0.523	0.011	0.003	0.006
18a	Island	Bowmore	0.914	6.670	0.316	21.100	0.868	0.037	0.148	0.548	0.032	0.007	0.018
19	Island	Bruichladdie	1.630	5.480	0.697	36.500	4.130	0.038	0.288	0.587	0.066	0.034	0.039
20a	Island	Bunnahabhain	2.240	7.540	1.350	36.200	2.120	0.051	0.184	0.580	0.057	0.014	0.037
21	Island	Talisker	0.034	4.850	0.362	5.670	0.607	0.018	0.070	0.277	0.033	0.003	0.006
22a	Lowland	Auchentoshan	0.169	1.460	0.417	11.700	0.681	0.042	0.128	1.320	0.037	0.006	0.012
23a	Lowland	Glenkinchie	0.108	2.450	0.176	7.760	0.738	0.031	0.106	0.434	0.022	0.002	0.007
24	Speyside	Balvenie	0.695	3.850	0.120	20.300	0.765	0.031	0.121	0.380	0.035	0.005	0.024
25	Speyside	Craigellachie	0.096	0.819	0.177	6.110	0.633	0.024	0.094	0.239	0.025	0.005	0.006
26	Speyside	Dufftown	0.883	4.640	0.130	14.000	1.050	0.030	0.078	0.533	0.024	0.002	0.014
27	Speyside	Glen Elgin	0.115	1.350	0.404	9.270	1.400	0.031	0.046	0.195	0.029	0.006	0.009
28	Speyside	Glenburgie	2.000	7.910	0.185	37.700	1.650	0.053	0.134	0.198	0.043	0.008	0.026
29	Speyside	Glennfiddich	0.317	2.720	0.344	12.400	0.660	0.029	0.132	0.519	0.193	0.004	0.013
30	Speyside	Glenrothes	0.953	4.110	0.399	16.700	1.830	0.041	0.137	1.030	0.029	0.007	0.014
31	Speyside	Knockando	0.051	1.030	0.191	5.140	0.605	0.017	0.094	0.432	0.020	0.008	0.005
32	Speyside	Linkwood	0.276	1.050	0.207	6.220	1.010	0.020	0.064	0.769	0.019	0.004	0.006

Note: Chemical concentrations reported in mg per L. All samples analyzed using total reflection X-ray fluorescence. Derived from Shand et al. 2017.

Due to Shand et al. (2017)'s inabilty to draw cohesive classification results at the regional level, we sought to allow the data as well as whisky type composition knowledge to intuitively guide our selection of the number of assigned clusters (k^*) in this analysis. We first viewed box-plots by variable class and the overall correlation structure of the chemical variables. Drawing upon observations here (Fig. X3), we then isolated the chemical variable correlation structures of the largest classes available to view and assessed mean vector differences between these. We then viewed the observations within three-dimensional principal component (PC) space to reduce dimensionality for intuitive viewing of possible clusters (Fig x4).

Counterfeits, Blends and Speyside whisky classes all displayed mean significant vector differences while Island and Speyside (Both single-region origin malted-barley whiskies) did not. Counterfeits were visually clearly differentiated with the PC space, as well as tentative groups emerging between blended/grain whiskies and regional whiskies (Speyside, Highland, Lowland, and Island). Further, blended whiskies are composed primarily of grains (wheat or maize, 60-80%) with little malted-barley, while single origin whiskies are of a single-malt or multi-malt character (Storrie 1962; Bower 2016; Kew et al. 2016; Scotch Whisky Association Cereals Working Group 2021).

Thus we re-aggregated the sampled whiskies into three logical overarching predictive whisky classes: "Counterfeits", "Grains & Blends" (a combination of gran and blended classes), and "Provenance" (all whiskies of a single-region origin). This reclassification is further supported as both grain type (barley vs. other) and

regionality being shown to influence chemical composition of whiskies under other analytical methods (Kew et al. 2016; Roullier-Gall et al. 2020).

We then reassessed correlation plots of chemical observations for each of these three classes, as well as mean vector differences between them. We found striking correlation structure differences between the classes (particularly between counterfeit and provenance whiskies) as well as significant mean vector differences (Fig x6, Table 3).

This led us to conduct Principle Component Analysis (PCA) to ascertain the magnitude and directionality of each chemical driver within a reduced space, as to infer and categorize chemical differences between groups later drawn from cluster analysis. Principal components analysis was conducted via using the prcomp() function on our scaled, log-transformed data.

As we wished to utilize all of the available data within our $n\ x\ p$ dimensional space, we implemented k-means, partitioning around medoids (PAM), and agglomerative hierarchical clustering to draw data-supported groups to test our predictive ones. In doing so, we scaled our log-transformed observations, except in the case of correlation distance hierarchical clustering which was conducted on the log-transformed data. Further, we wished to see if we could draw general consensus between these methods in support of using XTRF chemical composition sampling as a widely applicable method to produce suitable data for general multivariate analysis to categorize whiskies by counterfeit, blend, or single origin status.

2.2 K-means Unsupervised Clustering

K-means analysis was conducted on the scaled-log transformed data using the kmeans() function with the default recommended Hartigan-Wong algorithm, iterations set to 100 (iter. max = 100), and 50 random starts (nstart = 50) for all values of k. Setting random initializations to 50 reduces the chance of poor initial centroid allocation, while allowing 100 maximum iterations ensures convergence during the process of iterative group allocation of data points (ref).

This process was generated for $k = 1 - 10$, allowing k^* to be chosen via visual assessment of total within sum of squares (WSS) reductions with concurrent silhouette plot analysis.

2.3 Partitioning Around Medoids (PAM) Unsupervised Clustering

PAM clustering was conducted using the pam() function from the cluster package with Euclidean distance (Kaufman and Rousseeuw, n.d.). k^* assessment was conducted via assessing silhouette plots and widths as well as via comparing clustering results to k-means clustering for respective k^*.

2.4 Agglomerative Hierarchical Clustering

All agglomerative hierarchical clustering was performed using hclust(), with the original clustering results of (ref) reproduced via using euclidian distance with complete linkage (method = "complete"). We attempted to find consistent clustering results to (ref) using other distances (Minkowski, Manhattan and Chebyshev) and associated linkages (complete and Ward) as well as develop our own agglomerative hierarchical clustering models with better performance. Of those trialed, Euclidian (Ward linkage), Manhattan (complete and Ward linkage), and Correlation ($1 - r$, Ward linkage) distances were retained for comparison and analysis.

2.5 Quality Metrics

After k^* was established confusion matrices were produced for all clustering results and compared to our proposed groups (Provenance, blend/grain, and counterfeit whiskies) with global quality statistics (Acc, $F1\text{-score}_M$, TNR_M, $F1\text{-score}_\mu$, PPV_μ, TPR_μ) calculated. The best scoring clustering results were then taken and class-wise quality statistics calculated (Acc_i, MR_i, PPV_i, TPR_i, TNR_i, $F1\text{-score}_i$).

3 Results

4 Discussion

Reference

Bower, Julie. 2016. "Scotch Whisky: History, Heritage and the Stock Cycle." *Beverages* 2 (2): 11.

Kaufman, L., and Peter J. Rousseeuw. n.d. "PAM Clustering Algorithm."

Kew, Will, Ian Goodall, David Clarke, and Dušan Uhrín. 2016. "Chemical Diversity and Complexity of Scotch Whisky as Revealed by High-Resolution Mass Spectrometry." *Journal of the American Society for Mass Spectrometry* 28 (1): 200–213.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Roullier-Gall, Chloé, Julie Signoret, Christian Coelho, Daniel Hemmler, Mathieu Kajdan, Marianna Lucio, Bernhard Schäfer, Régis D Gougeon, and Philippe Schmitt-Kopplin. 2020. "Influence of Regionality and Maturation Time on the Chemical Fingerprint of Whisky." *Food Chemistry* 323: 126748.

Scotch Whisky Association Cereals Working Group. 2021. "Scotch Whisky Cereals Technical Note." The Scotch Whisky Association. https://www.scotch-whisky.org.uk/media/1900/cereals-technical-note-6th-edition-240821.pdf.

Shand, Charles A, Renate Wendler, Lorna Dawson, Kyari Yates, and Hayleigh Stephenson. 2017. "Multivariate Analysis of Scotch Whisky by Total Reflection x-Ray Fluorescence and Chemometric Methods: A Potential Tool in the Identification of Counterfeits." *Analytica Chimica Acta* 976: 14–24.

Storrie, Margaret C. 1962. "The Scotch Whisky Industry." *Transactions and Papers (Institute of British Geographers)*, no. 31: 97–114.