MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE CAPSTONE PROJECT:

PLEASANTON MALT WHISKY CIRCLE STATISTICS

David Melaugh

CAPSTONE OUTLINE

Background, Motivation, Objective

Data Ingestion, Cleanup, Transfer to Python

Initial Visualizations

Output 1: Member Profiles

Output 2: “Easy Grader” Analysis

Output 3: Similarity Analysis

Output 4: Score Prediction

Conclusion

BACKGROUND

Started in November 2009, the Pleasanton Malt Whisky Circle is composed a group of friends that have met roughly every six weeks to connect over a shared love of whisky – primarily Scotch whisky.

We typically taste six whiskies per meeting, and scores are recorded for each attendee. Meetings have 10-20 attendees.

To date, we have held 125 meetings, scoring over 750 whiskies. Once collected and organized for this Capstone Project, this yielded a database of almost 12,000 whisky scores.

MOTIVATION

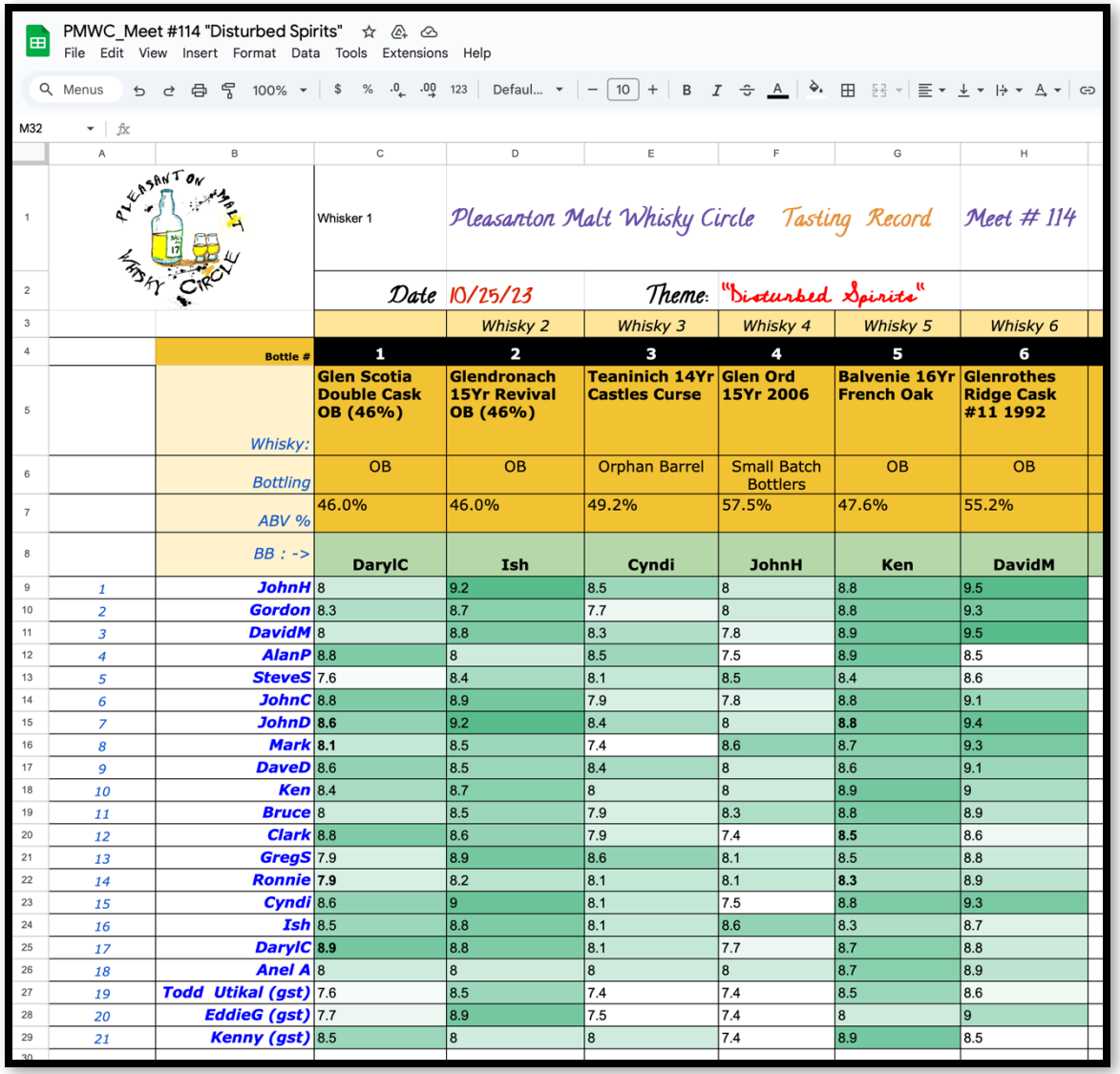
Though I will acknowledge I am not curing cancer here, PMWC meetings have been an enjoyable part of my life for over ten years. This Capstone Project offers a great opportunity to give back to the group to help our members better understand our own shared tastes.

The PMWC whisky scores also happen to offer a nicely sized dataset that is amenable to many of the learnings and techniques imparted by the Machine Learning and Artificial Intelligence Certificate Course.

OBJECTIVE

I will tackle four objectives in this Capstone project:

- Ingest and organize the scores and whisky information
- Prepare individual profiles for each Circle member
- Identify and analyze some interesting questions about the score data
- Generate a predictive analysis, using whisky features to predict PMWC scoring



PMWC_Meet #114 "Disturbed Spirits"

File Edit View Insert Format Data Tools Extensions Help

1		Whisker 1	Pleasanton Malt Whisky Circle	Tasting Record			Meet # 114
2			Date 10/25/23	Theme: "Disturbed Spirits"			
3			Whisky 2	Whisky 3	Whisky 4	Whisky 5	Whisky 6
4	Bottle #	1	2	3	4	5	6
5	Whisky:	Glen Scotia Double Cask OB (46%)	Glendronach 15Yr Revival OB (46%)	Teaninich 14Yr Castles Curse	Glen Ord 15Yr 2006	Balvenie 16Yr French Oak	Glenrothes Ridge Cask #11 1992
6	Bottling	OB	OB	Orphan Barrel	Small Batch Bottlers	OB	OB
7	ABV %	46.0%	46.0%	49.2%	57.5%	47.6%	55.2%
8	BB : ->	DarylC	Ish	Cyndi	JohnH	Ken	DavidM
9	1	JohnH	8	9.2	8.5	8	8.8	9.5
10	2	Gordon	8.3	8.7	7.7	8	8.8	9.3
11	3	DavidM	8	8.8	8.3	7.8	8.9	9.5
12	4	AlanP	8.8	8	8.5	7.5	8.9	8.5
13	5	SteveS	7.6	8.4	8.1	8.5	8.4	8.6
14	6	JohnC	8.8	8.9	7.9	7.8	8.8	9.1
15	7	JohnD	**8.6**	9.2	8.4	8	**8.8**	9.4
16	8	Mark	**8.1**	8.5	7.4	8.6	8.7	9.3
17	9	DaveD	8.6	8.5	8.4	8	8.6	9.1
18	10	Ken	8.4	8.7	8	8	8.9	9
19	11	Bruce	8	8.5	7.9	8.3	8.8	8.9
20	12	Clark	8.8	8.6	7.9	7.4	**8.5**	8.6
21	13	GregS	7.9	8.9	8.6	8.1	8.5	8.8
22	14	Ronnie	**7.9**	8.2	8.1	8.1	**8.3**	8.9
23	15	Cyndi	8.6	9	8.1	7.5	8.8	9.3
24	16	Ish	8.5	8.8	8.1	8.6	8.3	8.7
25	17	DarylC	**8.9**	8.8	8.1	7.7	8.7	8.8
26	18	Anel A	8	8	8	8	8.7	8.9
27	19	Todd Utikal (gst)	7.6	8.5	7.4	7.4	8.5	8.6
28	20	EddieG (gst)	7.7	8.9	7.5	7.4	8	9
29	21	Kenny (gst)	8.5	8	8	7.4	8.9	8.5

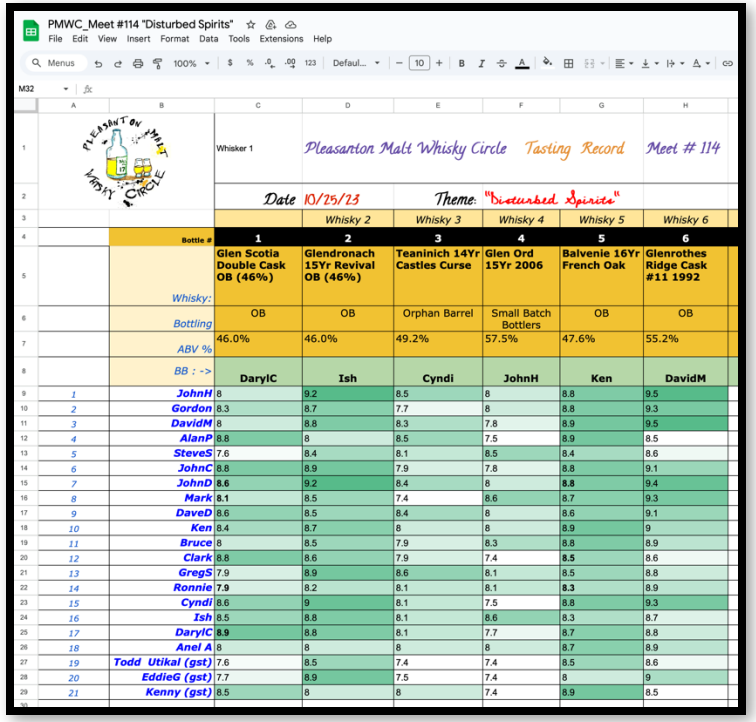
DATA INGESTION

SCORE INGESTION

Initial data source: Member scores are recorded at each meeting in an online Google Sheets file

Capstone ingestion steps:

- Individually export from Google Sheets to Excel
- Standardize data format by copying data in Excel to individual .csv files using Keyboard Maestro macro
- Combine .csv files into master score file
- End output: Meeting_Number, Attendee, Whisky_ID, Whisky_Score



PMWC_Meet #114 "Disturbed Spirits"

File Edit View Insert Format Data Tools Extensions Help

Whisker 1 — Pleasanton Malt Whisky Circle — Tasting Record — Meet # 114

Date 10/25/23 — Theme: "Disturbed Spirits"

	Bottle #	1	2	3	4	5	6
			Whisky 2	Whisky 3	Whisky 4	Whisky 5	Whisky 6
	Whisky:	Glen Scotia Double Cask OB (46%)	Glendronach 15Yr Revival OB (46%)	Teaninich 14Yr Castles Curse	Glen Ord 15Yr 2006	Balvenie 16Yr French Oak	Glenrothes Ridge Cask #11 1992
	Bottling	OB	OB	Orphan Barrel	Small Batch Bottlers	OB	OB
	ABV %	46.0%	46.0%	49.2%	57.5%	47.6%	55.2%
	BB : ->	DaryIC	Ish	Cyndi	JohnH	Ken	DavidM
1	JohnH	8	9.2	8.5	8	8.8	9.5
2	Gordon	8.3	8.7	7.7	8	8.8	9.3
3	DavidM	8	8.8	8.3	7.8	8.9	9.5
4	AlanP	8.8	8	8.5	7.5	8.9	8.5
5	SteveS	7.6	8.4	8.1	8.5	8.4	8.6
6	JohnC	8.8	8.9	7.9	7.8	8.8	9.1
7	JohnD	8.6	9.2	8.4	8	8.8	9.4
8	Mark	8.1	8.5	7.4	8.6	8.7	9.3
9	DaveD	8.6	8.5	8.4	8	8.6	9.1
10	Ken	8.4	8.7	8	8	8.9	9
11	Bruce	8	8.5	7.9	8.3	8.8	8.9
12	Clark	8.8	8.6	7.9	7.4	8.5	8.6
13	GregS	7.9	8.9	8.6	8.1	8.5	8.8
14	Ronnie	7.9	8.2	8.1	8.1	8.3	8.9
15	Cyndi	8.6	9	8.1	7.5	8.8	9.3
16	Ish	8.5	8.8	8.1	8.6	8.3	8.7
17	DaryIC	8.9	8.8	8.1	7.7	8.7	8.8
18	Anel A	8	8	8	8	8.7	8.9
19	Todd Utikal (gst)	7.6	8.5	7.4	7.4	8.5	8.6
20	EddieG (gst)	7.7	8.9	7.5	7.4	8	9
21	Kenny (gst)	8.5	8	8	7.4	8.9	8.5

Meeting_Number	Attendee	1	2	3	4	5
1	John Houston	7	6	8	6.5	6
1	Jayson	8	5	4	7	7
1	Pat	5.5	4	6.5	5.5	5
1	Nick	6	6	8	8	7
1	Gary	8	5	3	7	8
1	JohnG	7	6	8	8	6

Meeting_Number	Attendee	Whisky_ID	Whisky_Score
1	Gary Brown	1	8
1	Gary Nicolson	1	6
1	Jayson Samuli	1	8
1	John Gowey	1	7
1	John Houston	1	7
1	Pat Henry	1	5.5
1	Gary Brown	2	5
1	Gary Nicolson	2	6
1	Jayson Samuli	2	5
1	John Gowey	2	6
1	John Houston	2	6
1	Pat Henry	2	4

SCORE DATA CLEANUP

Manually clean up score data, including:

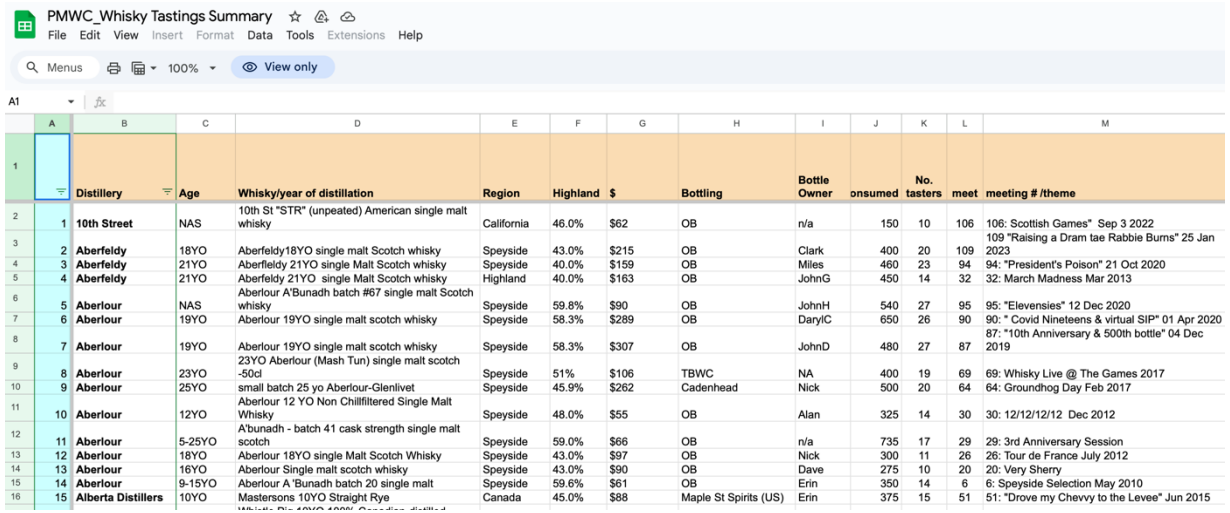
- Sort score data by name to manually standardize names (“GaryB” ”Gary B” “Brown, Gary” = “Gary Brown”)
- Sort by Meeting_Number to identify missing scores, track down missing score sheets with group leadership (end result: six meetings with no score sheet available)
- Sort by score to confirm no obvious outliers (below 0, above 10)
- Spot check master score list against original Google Sheets records to verify that ingestion steps did not introduce copy/paste errors
- Add new column, “Guest,” to distinguish between regular member attendees and one-off occasional attendees by cross-referencing a separate PMWC membership list, for later scoring analysis purposes

WHISKY INGESTION

Initial data source: PMWC master list of whiskies

Capstone ingestion steps:

- Export from Google Sheets to Excel



PMWC_Whisky Tastings Summary

File Edit View Insert Format Data Tools Extensions Help

Menus 100% View only

A1

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Distillery	Age	Whisky/year of distillation	Region	Highland	$	Bottling	Bottle Owner	onsumed	No. tasters	meet	meeting # /theme
2	1	10th Street	NAS	10th St "STR" (unpeated) American single malt whisky	California	46.0%	$62	OB	n/a	150	10	106	106: Scottish Games" Sep 3 2022
3	2	Aberfeldy	18YO	Aberfeldy18YO single malt Scotch whisky	Speyside	43.0%	$215	OB	Clark	400	20	109	109 "Raising a Dram tae Rabbie Burns" 25 Jan 2023
4	3	Aberfeldy	21YO	Aberfleldy 21YO single Malt Scotch whisky	Speyside	40.0%	$159	OB	Miles	460	23	94	94: "President's Poison" 21 Oct 2020
5	4	Aberfeldy	21YO	Aberfeldy 21YO single Malt Scotch whisky	Highland	40.0%	$163	OB	JohnG	450	14	32	32: March Madness Mar 2013
6	5	Aberlour	NAS	Aberlour A'Bunadh batch #67 single malt Scotch whisky	Speyside	59.8%	$90	OB	JohnH	540	27	95	95: "Elevensies" 12 Dec 2020
7	6	Aberlour	19YO	Aberlour 19YO single malt scotch whisky	Speyside	58.3%	$289	OB	DaryIC	650	26	90	90: " Covid Nineteens & virtual SIP" 01 Apr 2020
8	7	Aberlour	19YO	Aberlour 19YO single malt scotch whisky	Speyside	58.3%	$307	OB	JohnD	480	27	87	87: "10th Anniversary & 500th bottle" 04 Dec 2019
9	8	Aberlour	23YO	23YO Aberlour (Mash Tun) single malt scotch -50cl	Speyside	51%	$106	TBWC	NA	400	19	69	69: Whisky Live @ The Games 2017
10	9	Aberlour	25YO	small batch 25 yo Aberlour-Glenlivet	Speyside	45.9%	$262	Cadenhead	Nick	500	20	64	64: Groundhog Day Feb 2017
11	10	Aberlour	12YO	Aberlour 12 YO Non Chillfiltered Single Malt Whisky	Speyside	48.0%	$55	OB	Alan	325	14	30	30: 12/12/12/12 Dec 2012
12	11	Aberlour	5-25YO	A'bunadh - batch 41 cask strength single malt scotch	Speyside	59.0%	$66	OB	n/a	735	17	29	29: 3rd Anniversary Session
13	12	Aberlour	18YO	Aberlour 18YO single Malt Scotch Whisky	Speyside	43.0%	$97	OB	Nick	300	11	26	26: Tour de France July 2012
14	13	Aberlour	16YO	Aberlour Single malt scotch whisky	Speyside	43.0%	$90	OB	Dave	275	10	20	20: Very Sherry
15	14	Aberlour	9-15YO	Aberlour A 'Bunadh batch 20 single malt	Speyside	59.6%	$61	OB	Erin	350	14	6	6: Speyside Selection May 2010
16	15	Alberta Distillers	10YO	Mastersons 10YO Straight Rye	Canada	45.0%	$88	Maple St Spirits (US)	Erin	375	15	51	51: "Drove my Chevvy to the Levee" Jun 2015

Whisky_ID	Whisky_Distillery	Whisky_Age_Corrected	Whisky_Age	Whisky_Description	Whisky_Region	Whisky_ABV	Whisky_Price	Whisky_Bottling	Bottle_Owner	Meeting_Number	Meeting_Theme
1	Bladnoch	10	10YO	Utd. distillers.Flora & Fauna. Single malt	Lowland	0.4300	83	OB	John Houston	1	01: what's in your cupboard Nov 2009
2	Glenlivet	12	12YO	Glenlivet Single Malt.	Speyside	0.4000	27	OB	John Houston	1	01: what's in your cupboard Nov 2009
3	Dufftown	12	12YO	The Singleton of Dufftown	Speyside	0.4000	53	OB	Gary Brown	1	01: what's in your cupboard Nov 2009
4	Balvenie	12	12YO	Balvenie Signature batch #2. Single malt	Speyside	0.4000	58	OB	Jayson Samuli	1	01: what's in your cupboard Nov 2009
5	Bowmore	18	18YO	Bowmore single malt	Islay	0.4000	110	OB	John Gowey	1	01: what's in your cupboard Nov 2009
6	Arran	10	10YO	The Arran Single malt	Island	0.4600	48	OB	Daryl Coon	2	2: Around the Islands Dec 2009
7	Highland Park	18	18YO	Highland Park single malt (Orkney)	Island	0.4300	100	OB	Pat Henry	2	2: Around the Islands Dec 2009
8	Talisker	18	18YO	Talisker single malt (Skye)	Island	0.4580	81	OB	Gary Nicolson	2	2: Around the Islands Dec 2009
9	Bruichladdich		NAS	Bruichladdich Waves Cuvee single malt	Islay	0.4600	57	OB	John Houston	2	2: Around the Islands Dec 2009
10	Jura		NAS	Jura Superstition single malt	Island	0.4300	45	OB	Chris San Marchi	2	2: Around the Islands Dec 2009

WHISKY DATA CLEANUP

Manually clean up whisky data, including:

- Sort by distillery, region, bottling to manually standardize (abbreviations, misspellings, synonyms, etc.)
- Convert Whisky_ABV to decimal; round Whisky_Price to nearest dollar
- Add new column, “Whisky_Age_Corrected”, manually:
  - Convert from string to integer
  - “No age statement” whiskies now blank value
  - Whiskies with age range (“12-18 year”) given lowest value (12)
  - Original data preserved as “Whisky_Age” field in case needed

END OUTPUT

Master Data File.xlsx:

- 82 distinct attendees
- 11,742 scores
- 375 whiskies
- 194 distilleries, across 20 whisky regions
- Bottle prices ranging from $21 to $1,750
- Ages ranges from No Age Statement to 50 years old

UPTAKE INTO PYTHON

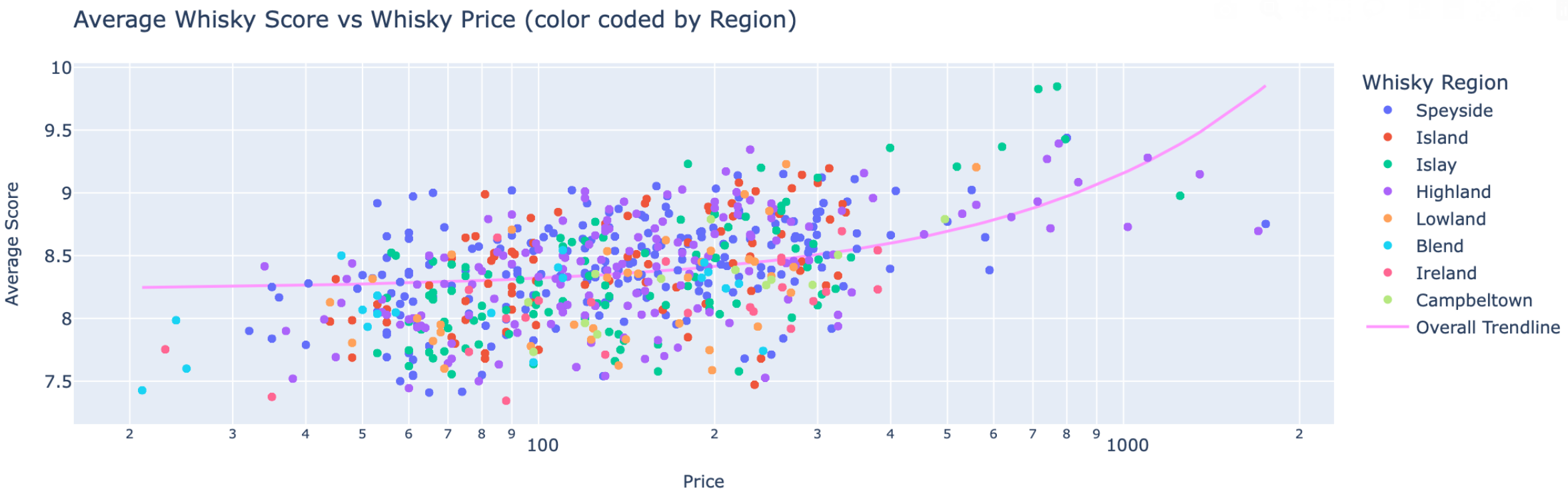
Data_Loading.py:

- Reads the scores and whiskies tabs from the Excel master data file, outputs a joined dataframe
- Passes: Whisky_ID, Whisky_Distillery, Whisky_Age_Corrected, Whisky_Description, Whisky_Region, Whisky_ABV, Whisky_Price, Meeting_Number, Whisky_Bottling

Over course of project, I built in variables to customize data uptake:

- remove_guests: If True, removes guest scores (default true)
- remove_USwhiskies: If True, removes US whiskies (default false)
- remove_thresh: If above 0, removes outlier values below X (default 0)
- pointscale: Used in combination with remove_thresh to rescale scores (e.g., if scores below 7 removed, rescales 7-10 as 0-30; default false)
- fill_missing_age: If True, sets a placeholder for missing whisky ages and adds a field noting this (default true)
- min_whiskies_per_region: If above 0, removes any region that has lower than X unique whiskies (default 0)

INITIAL VISUALIZATIONS



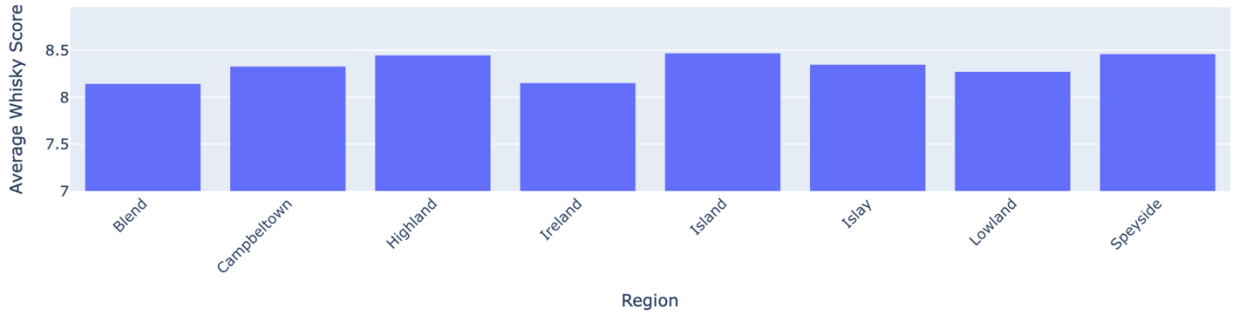
INITIAL VISUALIZATIONS

Data Analysis.Visualization.ipynb:

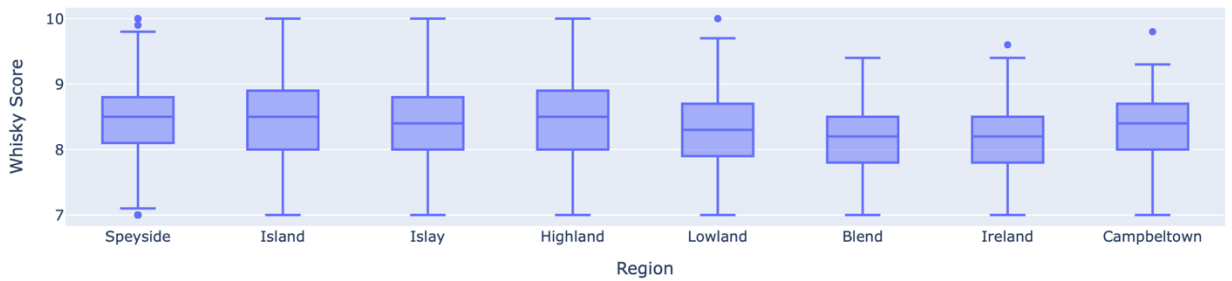
- Dropped outliers below 7 for cleaner visualization (very few scores are below this value)
- Selected visualizations of whisky score distribution, regional averages
- Experimented with simpler and more complex visualizations
- (examples follow)

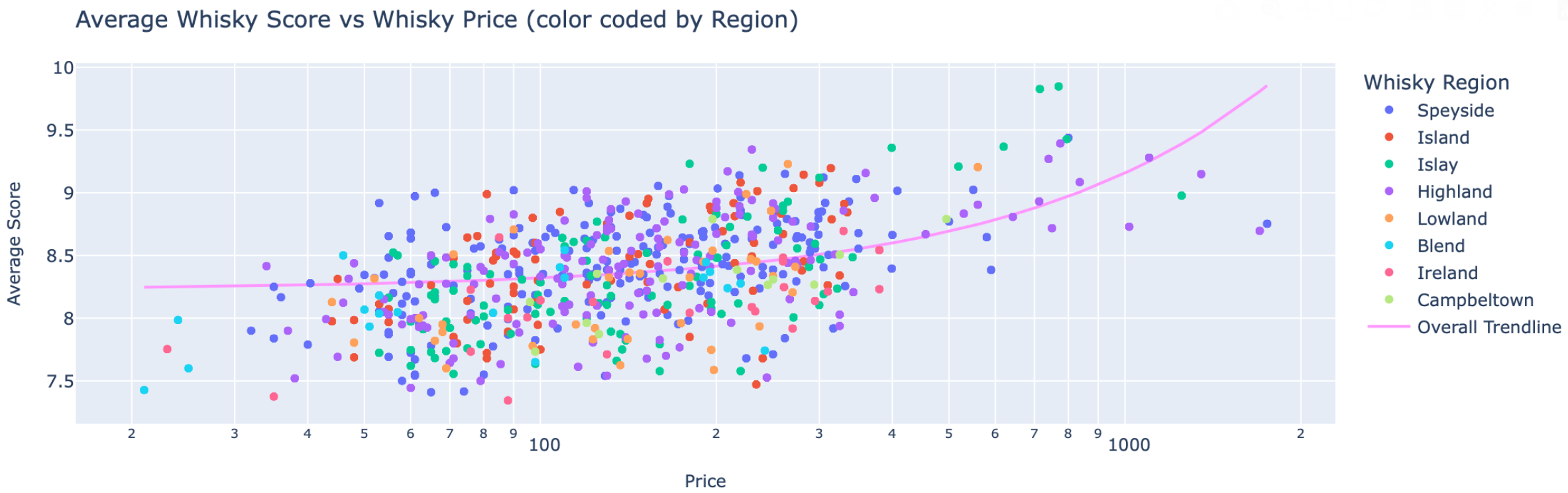
Distribution of Whisky Scores

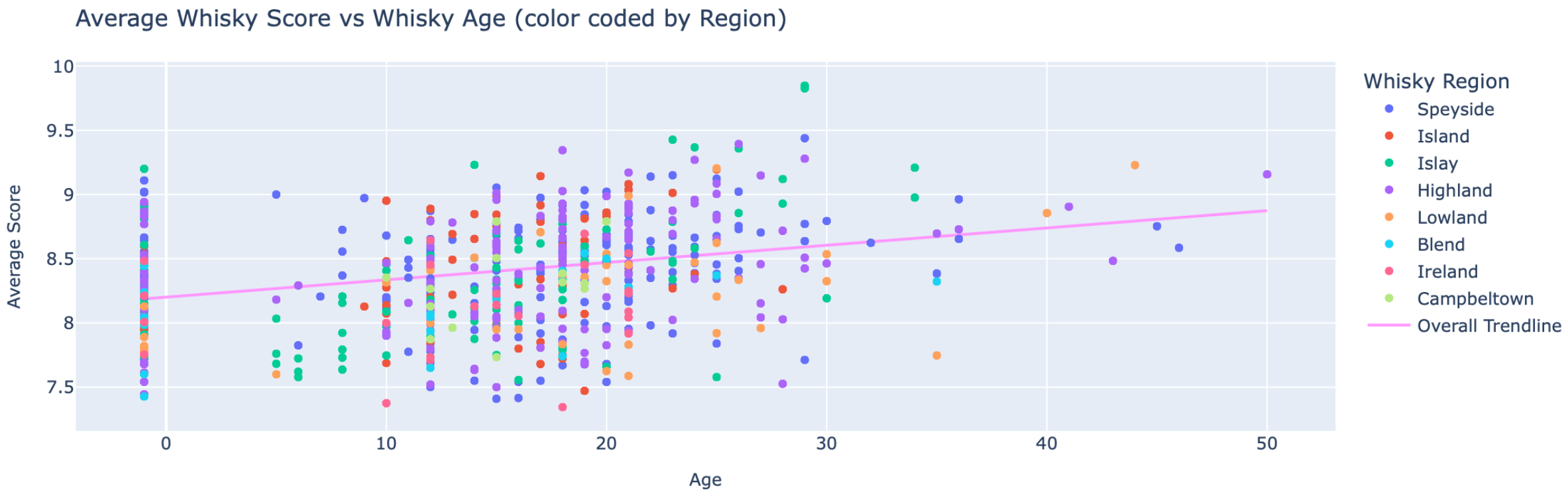
Average Whisky Score by Region

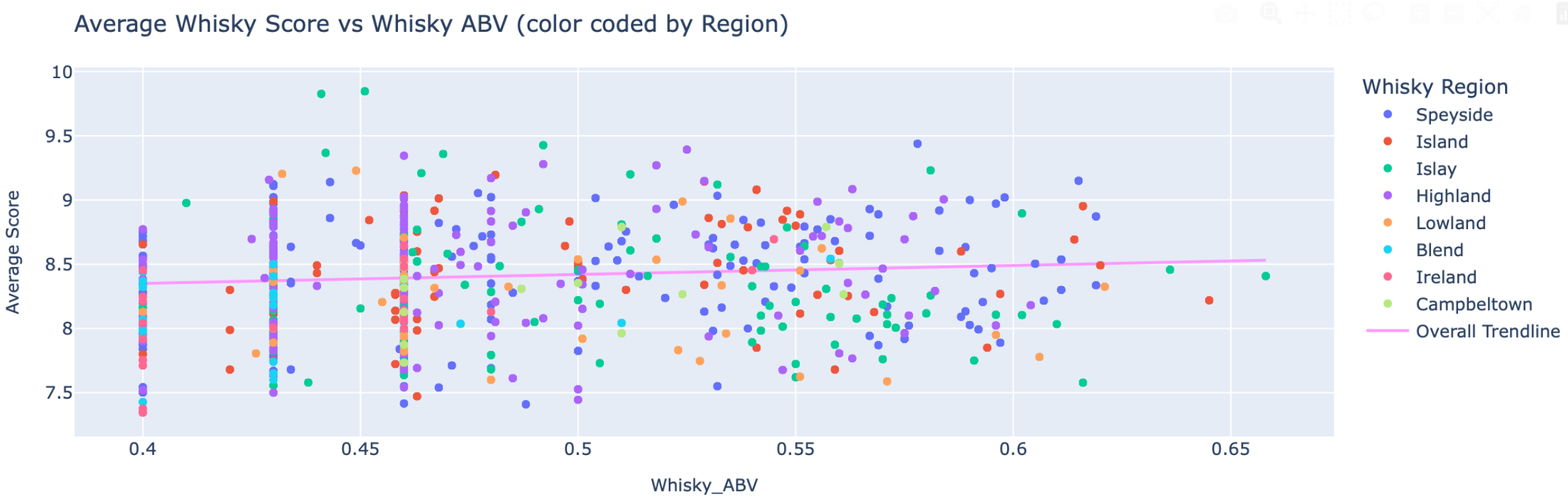


Whisky Score by Region




Average Whisky Score vs Whisky Price (color coded by Region)
Average Score
10
9.5
9
8.5
8
7.5
2
3
4
5
6
7
8
9
100
2
3
4
5
6
7
8
9
1000
2
Price
Whisky Region
Speyside
Island
Islay
Highland
Lowland
Blend
Ireland
Campbeltown
Overall Trendline


Average Whisky Score vs Whisky Age (color coded by Region)
Average Score
10
9.5
9
8.5
8
7.5
0
10
20
30
40
50
Age
Whisky Region
Speyside
Island
Islay
Highland
Lowland
Blend
Ireland
Campbeltown
Overall Trendline


Average Whisky Score vs Whisky ABV (color coded by Region)
Average Score
10
9.5
9
8.5
8
7.5
0.4
0.45
0.5
0.55
0.6
0.65
Whisky_ABV
Whisky Region
Speyside
Island
Islay
Highland
Lowland
Blend
Ireland
Campbeltown
Overall Trendline

PMWC Report for David Melaugh

Meetings Attended: 73

Whiskies Scored: 443

Total Price of Whiskies Scored: $89,826.50

Average Score: 8.14

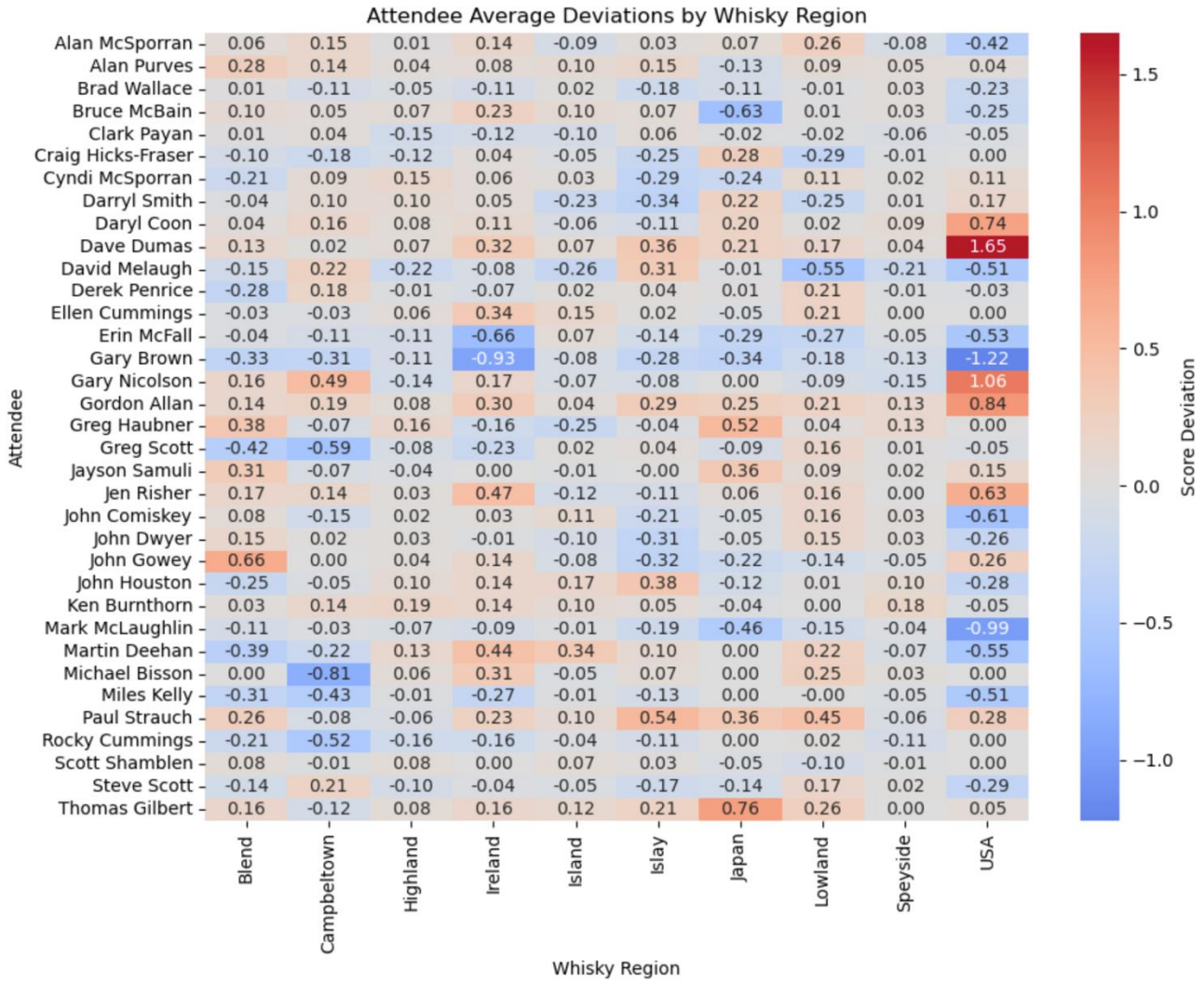
CAPSTONE OUTPUT 1:

MEMBER PROFILES

MEMBER PROFILES

whisky_report_generator.py:

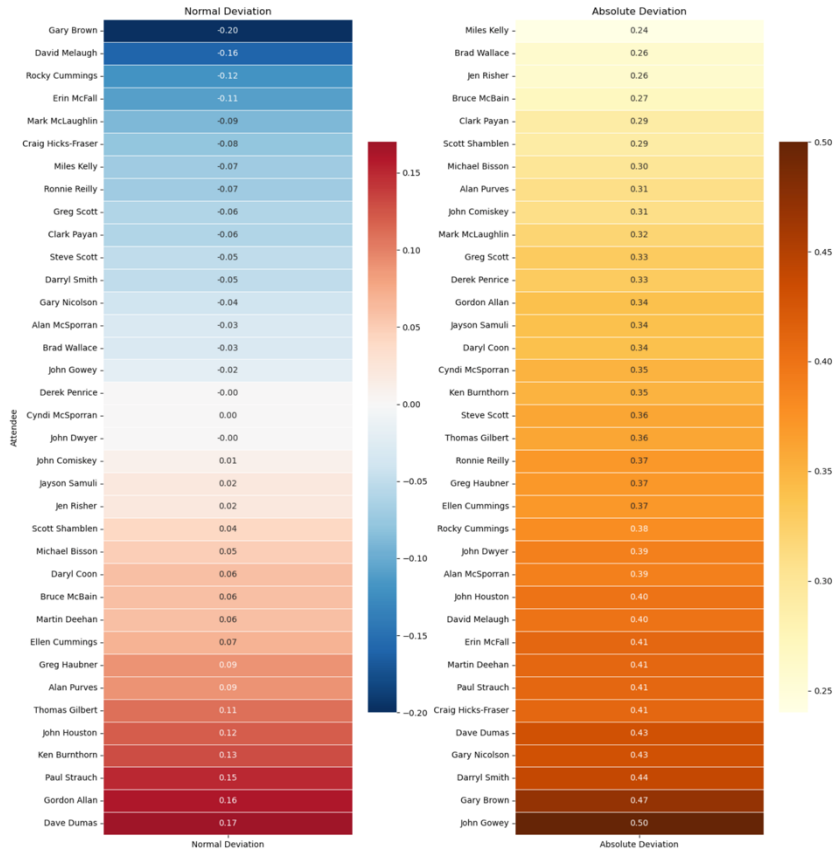
- Presents a variety of information about a member's participation in the PMWC
- Report begins with meeting and whisky count, average score, and total price of whiskies tasted, along with a graph of moving average of scores
- Distillery and Region analysis: average score, top and bottom 5
- Scoring patterns: Most and least similar members (this analysis is discussed further in Output 2), largest "disagreement" with another member
- Report concludes with complete scoring history
- Example output included in Git: whisky_report_david_melaugh.pdf



CAPSTONE OUTPUT 2:

“EASY GRADER” ANALYSIS

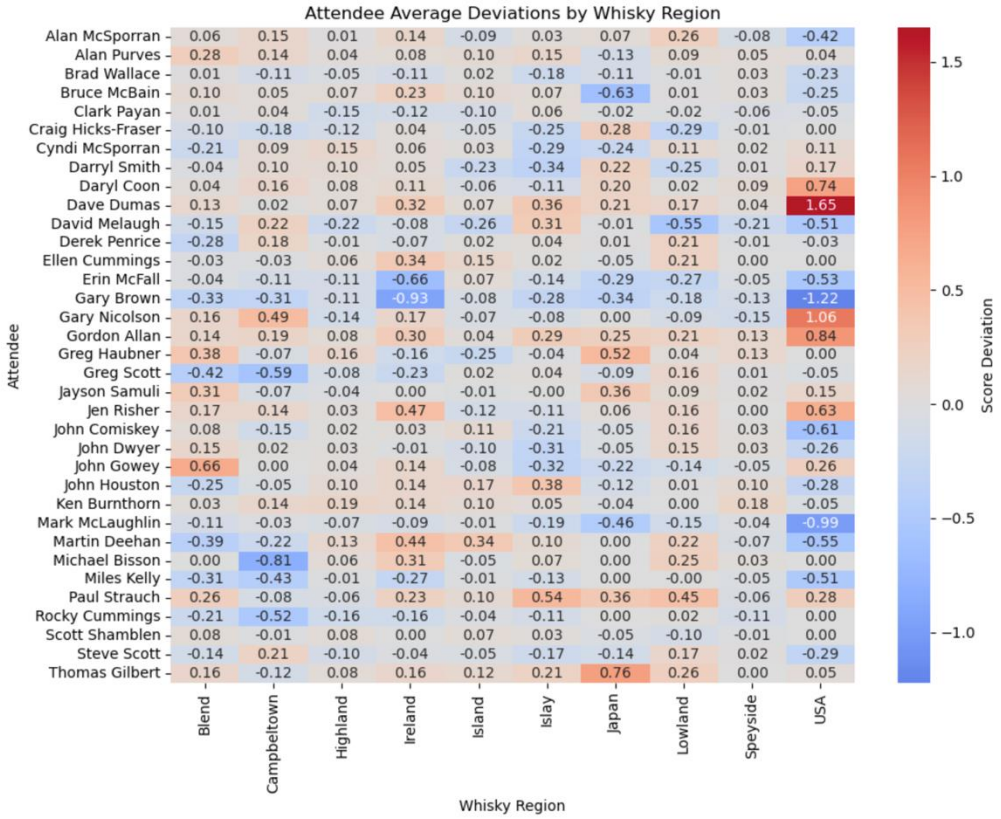
SCORING TREND ANALYSIS



[Data Analysis.Easy Graders.ipynb]:

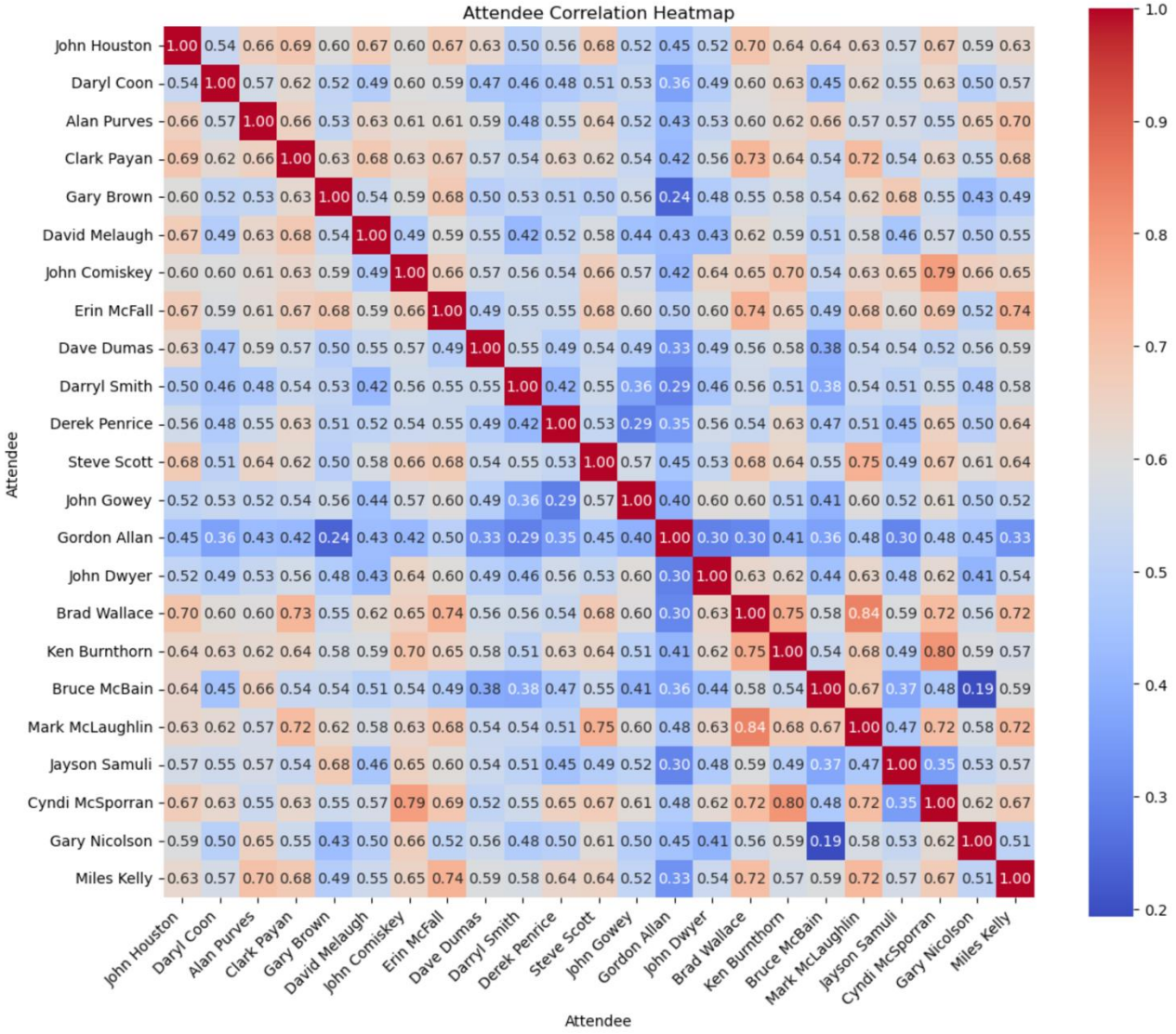
- Seeks to answer the question: are some members easy graders, habitually scoring whiskies higher than the group? Are some curmudgeons?
- Calculates on a whisky-by-whisky basis the group average score, the member's deviation, then totals the deviation and absolute deviation
- How about folks who march to their own drummer? The absolute value of the deviation should tell us that
- Result: Gary Brown and I are curmudgeons, regularly scoring lower than the group; Dave Dumas is the easy grade, regularly scoring higher. John Gowey marches to his own beat, with the highest absolute deviation from the group average score.
- [Drops guest scores and has a filter for number of meeting attendees to limit output for legibility]

SCORING TREND ANALYSIS



Data Analysis.Easy Graders.ipynb:

- Drill down further: could score deviation be better explained by individual preferences that deviate from group, as opposed to a blanket easy/hard grading curve?
- Test this by segregating deviation by region, age, and price
- Answer: yes, attitude in particular on US whiskey (i.e., bourbon) drives a chunk of the deviation – Dave Dumas very much likes bourbon, while Gary Brown does not (but both still deviate to some degree across all categories)



CAPSTONE OUTPUT 3:

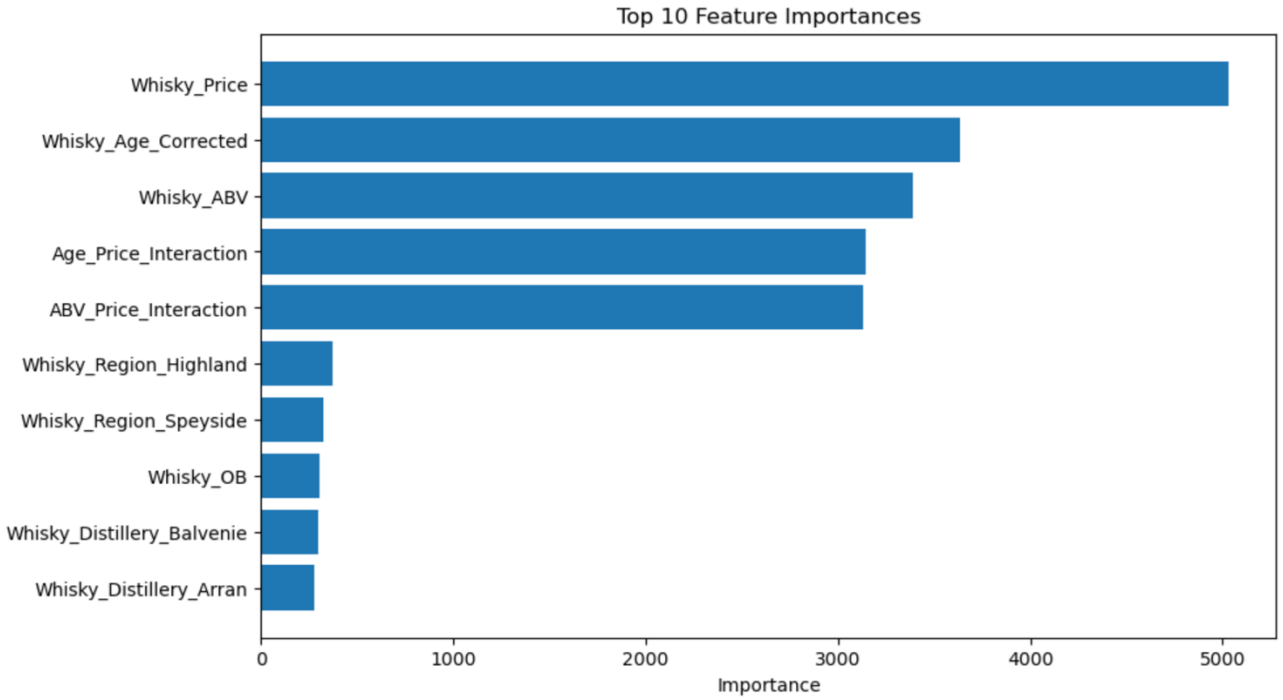
SIMILARITY ANALYSIS

SIMILARITY ANALYSIS

```
Most Correlated Attendee for Each Attendee:
John Houston is most correlated with Brad Wallace (Correlation: 0.70)
Daryl Coon is most correlated with Ken Burnthorn (Correlation: 0.63)
Alan Purves is most correlated with Miles Kelly (Correlation: 0.70)
Clark Payan is most correlated with Brad Wallace (Correlation: 0.73)
Gary Brown is most correlated with Erin McFall (Correlation: 0.68)
David Melaugh is most correlated with Clark Payan (Correlation: 0.68)
John Comiskey is most correlated with Cyndi McSporran (Correlation: 0.79)
Erin McFall is most correlated with Miles Kelly (Correlation: 0.74)
Dave Dumas is most correlated with John Houston (Correlation: 0.63)
Darryl Smith is most correlated with Miles Kelly (Correlation: 0.58)
Derek Penrice is most correlated with Cyndi McSporran (Correlation: 0.65)
Steve Scott is most correlated with Mark McLaughlin (Correlation: 0.75)
John Gowey is most correlated with Cyndi McSporran (Correlation: 0.61)
Gordon Allan is most correlated with Erin McFall (Correlation: 0.50)
John Dwyer is most correlated with John Comiskey (Correlation: 0.64)
Brad Wallace is most correlated with Mark McLaughlin (Correlation: 0.84)
Ken Burnthorn is most correlated with Cyndi McSporran (Correlation: 0.80)
Bruce McBain is most correlated with Mark McLaughlin (Correlation: 0.67)
Mark McLaughlin is most correlated with Brad Wallace (Correlation: 0.84)
Jayson Samuli is most correlated with Gary Brown (Correlation: 0.68)
Cyndi McSporran is most correlated with Ken Burnthorn (Correlation: 0.80)
Gary Nicolson is most correlated with John Comiskey (Correlation: 0.66)
Miles Kelly is most correlated with Erin McFall (Correlation: 0.74)
```

Data Analysis.Most Similar.ipynb:

- Seeks to answer the question: for a given member, what other members have similar tastes? What members have different tastes?
- First observation: no negative correlations. We're all swimming in generally the same direction, meaning that if any given member scores a whisky higher, every other member is at least somewhat likely to score that whisky higher.
- Output: heatmap plus list of most-correlated for each member
- Limit analysis to members who have attended over 200 meetings, to ensure strong basis for analysis and to make heatmap legible



CAPSTONE OUTPUT 4:

SCORE PREDICTION

SCORE PREDICTION

Data Analysis.Regression.ipynb:

- Most complex of the Capstone Outputs
- Objective: predict Whisky_Score using various whisky data
- Evaluation metrics: R^2 (measuring variance explained by model) and RMSE (measuring average magnitude of error, in same unit as target)
- Process overview:
  - Test various models
  - Select LightGBM and CatBoost as a good balance between speed and accuracy, hyperparameter tune both and compare
  - Select LightGBM as preferred model, improve via feature engineering
  - Step back, bucketize scores into low, medium, high in attempt to gain more predictive power at expense of precision
  - [Repeat model selection and hyperparameter tuning]
  - Test combination of models via ensemble with voting

MODEL TESTING

Model Comparison: Regression on Whisky Scores
We're testing 7 models. Higher **R²** and lower **NMSE** indicate better performance.

Linear Regression
- R² (Train Set): **0.27**
- NMSE (Train Set): **0.73**
- RMSE (Train Set): **4.92**
- MAE (Train Set): **3.93**
- Training Time: **0.0 seconds**

XGBoost
- Best Parameters: {'model__n_estimators': 200, 'model__max_depth': 5, 'model__learning_rate': 0.1}
- R² (Train Set): **0.40**
- NMSE (Train Set): **0.60**
- RMSE (Train Set): **4.46**
- MAE (Train Set): **3.54**
- Training Time: **1.3 seconds**

LightGBM
- Best Parameters: {'model__n_estimators': 200, 'model__max_depth': -1, 'model__learning_rate': 0.1}
- R² (Train Set): **0.48**
- NMSE (Train Set): **0.52**
- RMSE (Train Set): **4.16**
- MAE (Train Set): **3.27**
- Training Time: **3.7 seconds**

Gradient Boosting
- Best Parameters: {'model__n_estimators': 200, 'model__max_depth': 5, 'model__learning_rate': 0.1}
- R² (Train Set): **0.41**
- NMSE (Train Set): **0.59**
- RMSE (Train Set): **4.44**
- MAE (Train Set): **3.52**
- Training Time: **8.6 seconds**

Support Vector Regression
- Best Parameters: {'model__gamma': 'scale', 'model__C': 10}
- R² (Train Set): **0.40**
- NMSE (Train Set): **0.60**
- RMSE (Train Set): **4.48**
- MAE (Train Set): **3.43**
- Training Time: **13.5 seconds**

CatBoost
- Best Parameters: {'model__learning_rate': 0.05, 'model__iterations': 1000, 'model__depth': 8}
- R² (Train Set): **0.48**
- NMSE (Train Set): **0.52**
- RMSE (Train Set): **4.16**
- MAE (Train Set): **3.28**
- Training Time: **14.2 seconds**

Random Forest
- Best Parameters: {'model__n_estimators': 300, 'model__max_depth': 20}
- R² (Train Set): **0.47**
- NMSE (Train Set): **0.53**
- RMSE (Train Set): **4.19**
- MAE (Train Set): **3.31**
- Training Time: **55.2 seconds**

- In preprocessing, dropped guest scores, scores below 7, and whiskies from uncommon regions (e.g., New Zealand), to reduce effects of outliers. Rescaled scores from 0-30 (formerly 7-10 with decimal scores).
- Tested linear regression, XGBoost, Gradient Boosting, SVR, Random Forest, plus two additional models suggested by research, LightGBM and CatBoost.
- **Result:** LightGBM and CatBoost present a good balance of accuracy and speed, achieving near or better performance with Random Forest with much less processing time (LightGBM = 14x faster than Random Forest)

LIGHTGBM VS. CATBOOST

- Hyperparameter tune LightGBM and CatBoost
- Result: a tie. LightGBM takes longer to train, but that is caused by greater diversity of parameters to test. LightGBM chosen for further optimization given its speed in prior round (3.7 sec. vs. 14.2 for CatBoost)

Model Comparison: LightGBM Regression on Whisky Scores
Tuning LightGBM deeply. Higher **R²** and lower **NMSE** indicate better performance.

```
================================================================
  ▸ Best Parameters:       {'model__subsample': 1.0, 'model__reg_lambda': 0, 'model__reg_alpha': 0, 'model__n_estimators': 100, 'model__min_c
hild_samples': 10, 'model__max_depth': -1, 'model__learning_rate': 0.2, 'model__colsample_bytree': 1.0}
  ▸ R² (Train Set):         0.48
  ▸ NMSE (Train Set):       0.52
  ▸ RMSE (Train Set):       4.14
  ▸ MAE (Train Set):        3.25
  ▸ Training Time:          185.1 seconds
----------------------------------------------------------------
```

Model Comparison: CatBoost Regression on Whisky Scores
Tuning CatBoost deeply. Higher **R²** and lower **NMSE** indicate better performance.

```
================================================================
  ▸ Best Parameters:       {'model__learning_rate': 0.1, 'model__l2_leaf_reg': 1, 'model__iterations': 500, 'model__depth': 8}
  ▸ R² (Train Set):         0.48
  ▸ NMSE (Train Set):       0.52
  ▸ RMSE (Train Set):       4.15
  ▸ MAE (Train Set):        3.26
  ▸ Training Time:          20.4 seconds
----------------------------------------------------------------
```

Summary of Model Performance

	Model	R2	NMSE	RMSE	MAE	Training Time (s)
0	LightGBM	0.484415	0.515585	4.140438	3.253919	185.104894
1	CatBoost	0.482295	0.517705	4.148941	3.264770	20.401227

LIGHTGBM FEATURE ENGINEERING

- Hardcode the best parameters for LightGBM, with some tweaking
- Test addition of feature combinations:
  - ABV_Price_Interaction = Whisky_ABV * Whisky_Price
  - ABV_Squared = Whisky_ABV^2
  - Log_Whisky_Price = log(Whisky_Price)
  - Age_Price_Interaction = 'Whisky_Age_Corrected * Whisky_Price
- Result: further (small) increase in accuracy

Further Improvement: LightGBM Regression on Whisky Scores
Hard-coded parameters applied, now adding feature engineering. Higher **R²** and lower **NMSE** indicate better performance.

==========

Training Performance:
- ▸ R² (Train Set): **0.49**
- ▸ NMSE (Train Set): **0.51**
- ▸ RMSE (Train Set): **4.11**
- ▸ MAE (Train Set): **3.22**

Test Performance:
- ▸ R² (Test Set): **0.39**
- ▸ NMSE (Test Set): **0.61**
- ▸ RMSE (Test Set): **4.52**
- ▸ MAE (Test Set): **3.53**

- ▸ Training Time: **2.0 seconds**

Top 10 Important Features:

	Feature	Importance
2	Whisky_Price	5028
0	Whisky_Age_Corrected	3637
1	Whisky_ABV	3391
8	Age_Price_Interaction	3143
5	ABV_Price_Interaction	3130
168	Whisky_Region_Highland	373
173	Whisky_Region_Speyside	325
4	Whisky_OB	304
23	Whisky_Distillery_Balvenie	295
16	Whisky_Distillery_Arran	279

SHIFT TO SCORE BUCKET MODEL

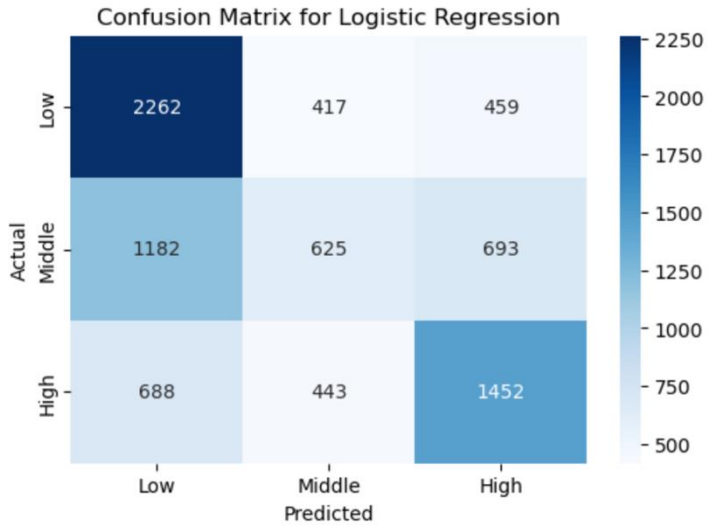
- Instead of trying to predict specific whisky scores, let's bucket the scores into low, medium, and high, and try to predict the bucket
- Sacrifices score prediction precision, but potentially gains accuracy
- In practice, difference between, e.g., a "7.8" whisky an "8.2" whisky is negligible
- New evaluation metric: accuracy (all errors are equally bad, so we don't need to weight false negatives or positives more heavily)
- Results on next page
  - Predictive accuracy now 62%
  - LightGBM was again strongest performer

```
Class Distribution (0=Low, 1=Middle, 2=High):
Whisky_Score
0    3941
1    3082
2    3254
Name: count, dtype: int64
Score Cutoffs:
Low: <= 12.0, Middle: > 12.0 and <= 17.0, High: > 17.0
Converted Score Cutoffs (x / 10 + 7):
Low: <= 8.2, Middle: > 8.2 and <= 8.7, High: > 8.7
============================================================
```

==========

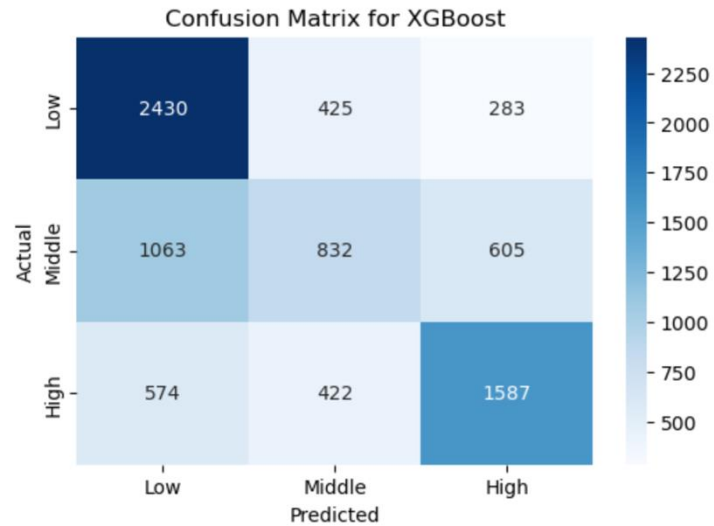
Logistic Regression

- Accuracy (Train Set): **0.53**
- Training Time: **0.1 seconds**



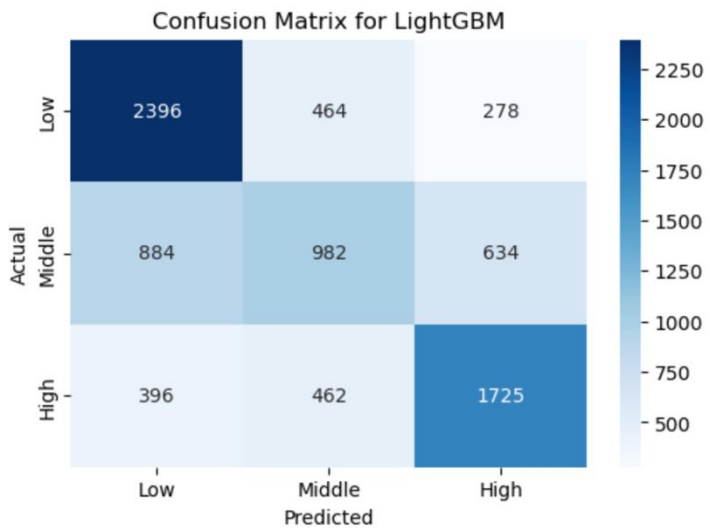
XGBoost

- Best Parameters: {'model__n_estimators': 200, 'model__max_depth': 5, 'model__learning_rate': 0.1}
- Accuracy (Train Set): **0.59**
- Training Time: **3.4 seconds**



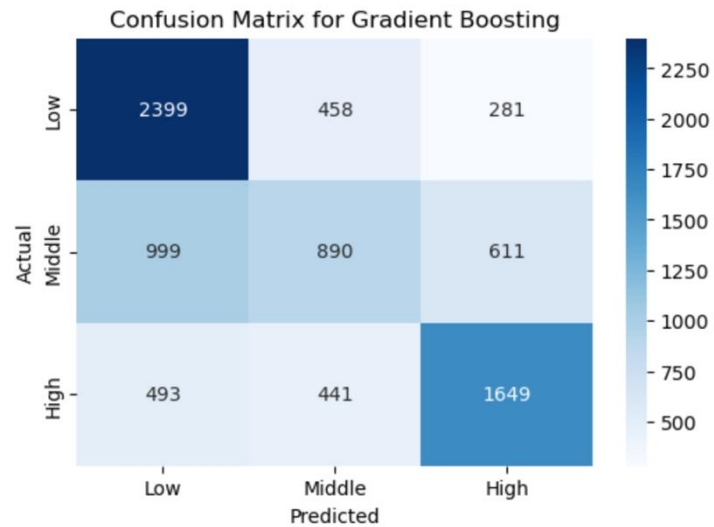
LightGBM

- Best Parameters: {'model__n_estimators': 200, 'model__max_depth': 10, 'model__learning_rate': 0.1}
- Accuracy (Train Set): **0.62**
- Training Time: **9.6 seconds**

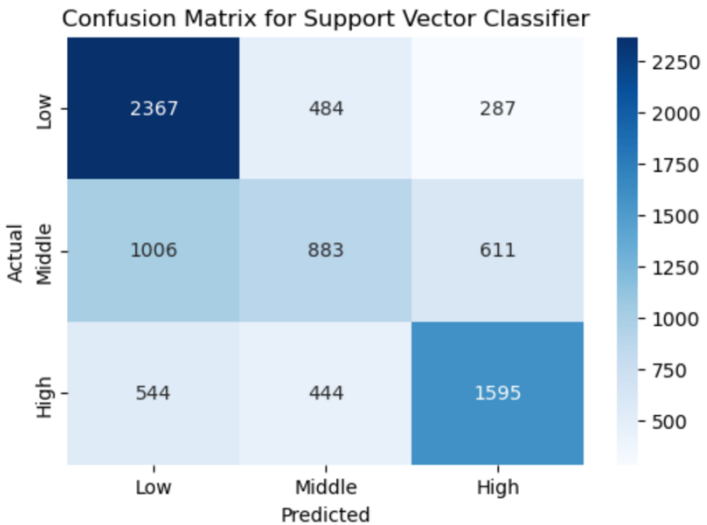


Gradient Boosting

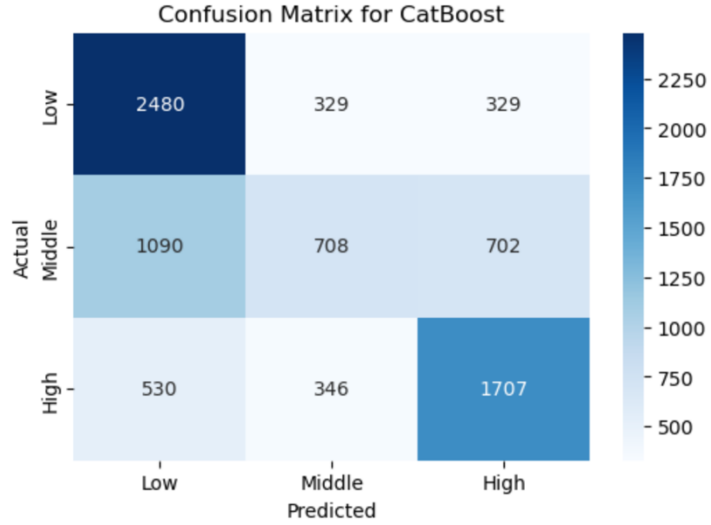
- Best Parameters: {'model__n_estimators': 200, 'model__max_depth': 5, 'model__learning_rate': 0.1}
- Accuracy (Train Set): **0.60**
- Training Time: **28.1 seconds**



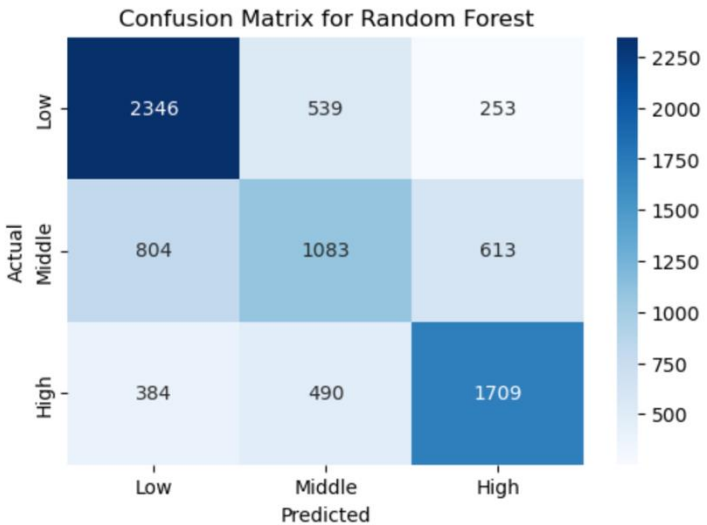
Support Vector Classifier
- ▸ Best Parameters: {'model__gamma': 'scale', 'model__C': 10}
- ▸ Accuracy (Train Set): **0.59**
- ▸ Training Time: **13.2 seconds**



CatBoost
- ▸ Best Parameters: {'model__learning_rate': 0.03, 'model__iterations': 1000, 'model__depth': 6}
- ▸ Accuracy (Train Set): **0.60**
- ▸ Training Time: **22.8 seconds**

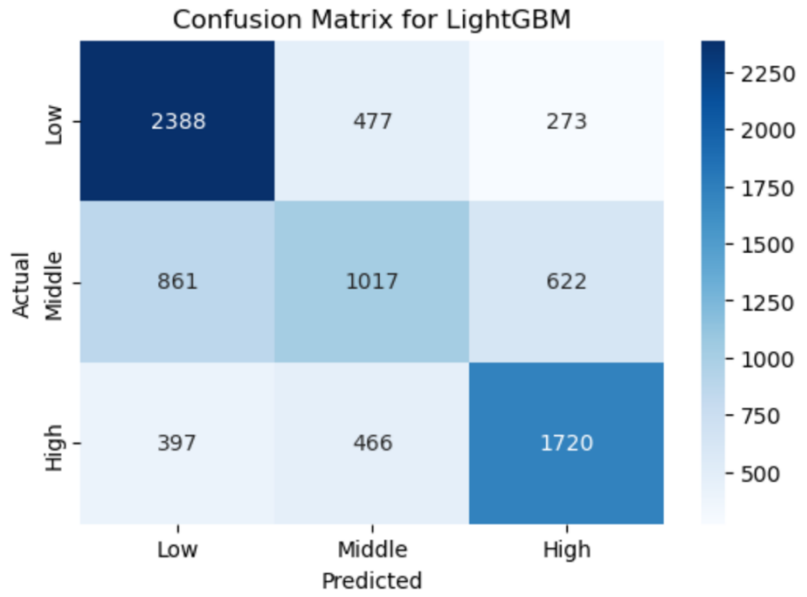


Random Forest
- ▸ Best Parameters: {'model__n_estimators': 200, 'model__max_depth': None}
- ▸ Accuracy (Train Set): **0.62**
- ▸ Training Time: **20.7 seconds**



SCORE BUCKET MODEL : LIGHTGBM TUNING

- Selected LightGBM, engaged hyperparameter tuning
- **Result:** increase in accuracy, especially in middle bucket (986 to 1017)

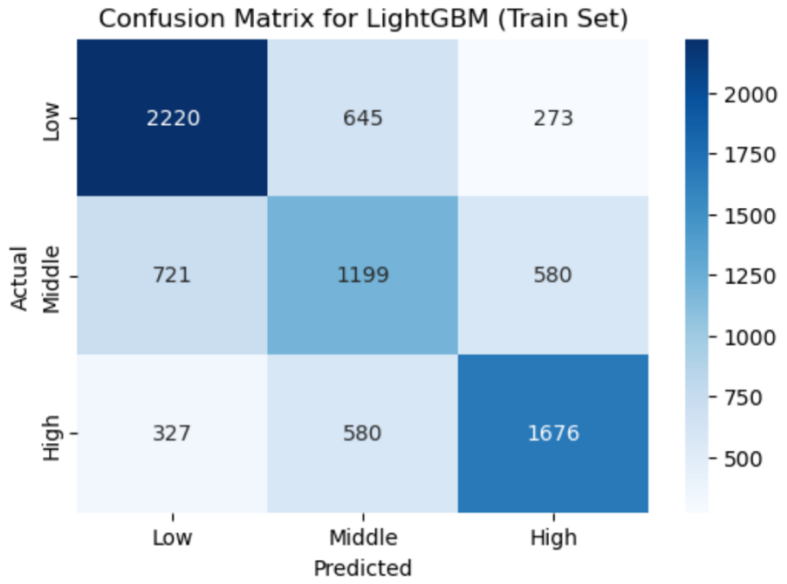


Class Distribution (0=Low, 1=Middle, 2=High):
Whisky_Score
0 3941
1 3082
2 3254
Name: count, dtype: int64
Score Cutoffs:
Low: <= 12.0, Middle: > 12.0 and <= 17.0, High: > 17.0
Converted Score Cutoffs (x / 10 + 7):
Low: <= 8.2, Middle: > 8.2 and <= 8.7, High: > 8.7

Model Comparison: LightGBM Classification on Whisky Score Buckets
Tuning LightGBM deeply. Higher **Accuracy** indicates better performance.

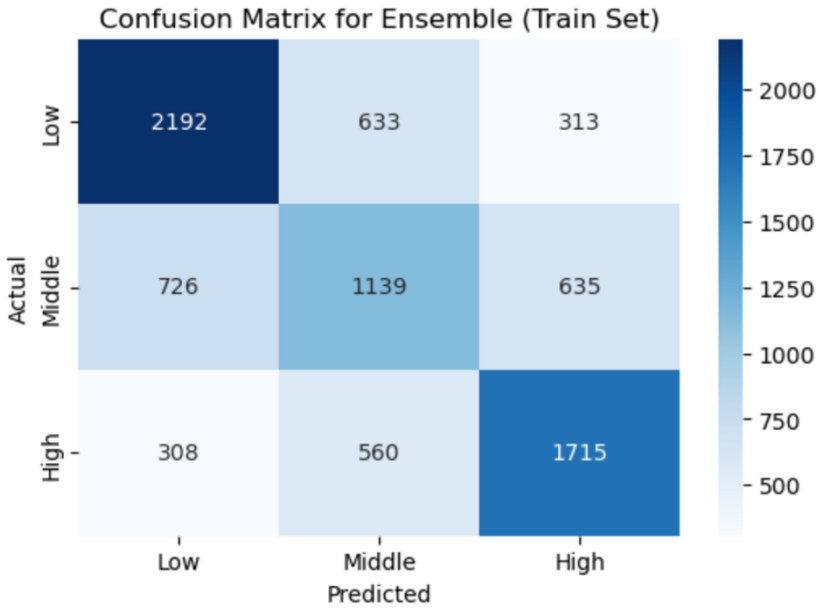
▸ Best Parameters: {'model__subsample': 0.6, 'model__reg_lambda': 0.01, 'model__reg_alpha': 0, 'model__n_estimators': 500, 'model__min_child_samples': 10, 'model__max_depth': 15, 'model__learning_rate': 0.05, 'model__colsample_bytree': 0.6}
▸ Accuracy (Train Set): **0.62**
▸ Training Time: **828.5 seconds**

SCORE BUCKET MODEL: LIGHTGBM FEATURE ENGINEERING



- Hardcode the best LightGBM parameters, with tweaking
- Implement feature engineering, with some new additions:
  - ABV_Price_Interaction = Whisky_ABV * Whisky_Price
  - ABV_Squared = $Whisky\_ABV^2$
  - Log_Whisky_Price = log(Whisky_Price)
  - Age_Price_Interaction = 'Whisky_Age_Corrected * Whisky_Price
  - Price_vs_Region_Avg = Whisky_Price - region_price_avg
  - ABV_vs_Region_Avg = Whisky_ABV - Region_ABV_Avg
- Add Synthetic Minority Over-sampling Technique (SMOTE) to help offset bucket imbalance
- **Result:** strong increase in accuracy for middle (1017 to 1199), decrease at high (1720 to 1676)

SCORE BUCKET MODE: ENSEMBLE



- Let's throw everything at this dataset, as a last try to improve accuracy
- Ensemble LightGBM + XGBoost + CatBoost + RF + SVC + Logistic Regression, with LightGBM weighted most heavily
- **Result:** no meaningful improvement over tuned LightGBM model
- Suggests not much room for further enhancement – we are hitting limits of underlying noise in data

THANK YOU

Presented by David Melaugh