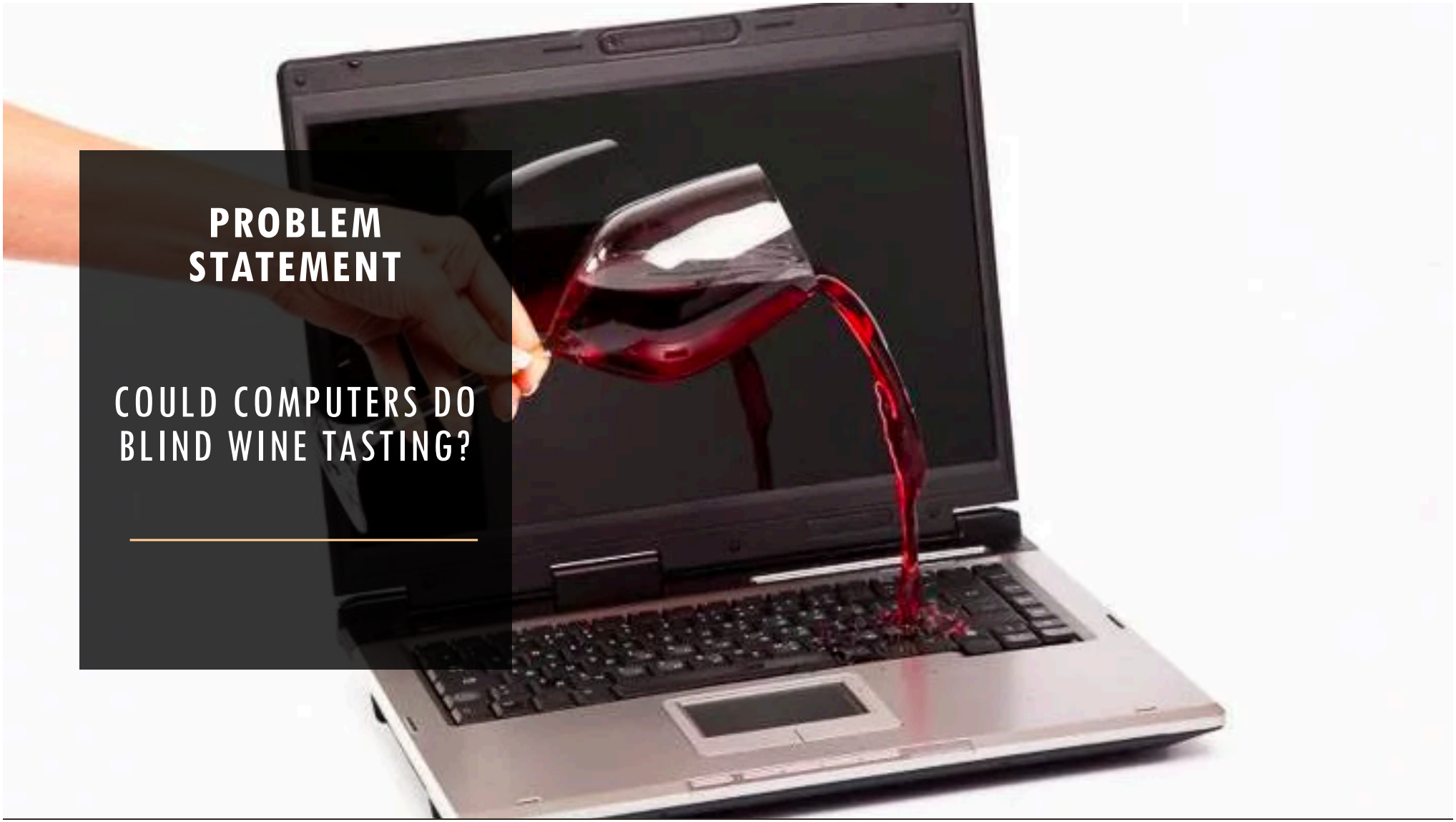

BLIND WINE TASTING
MARIA GIMENO

AGENDA

1. Problem statement
2. Goal
3. Datasets
4. Clean the data
5. Exploration Data Analysis
6. Predictor
7. Targets
8. Approach
9. Model Selection
10. Summary
11. Conclusion
12. Lessons learnt

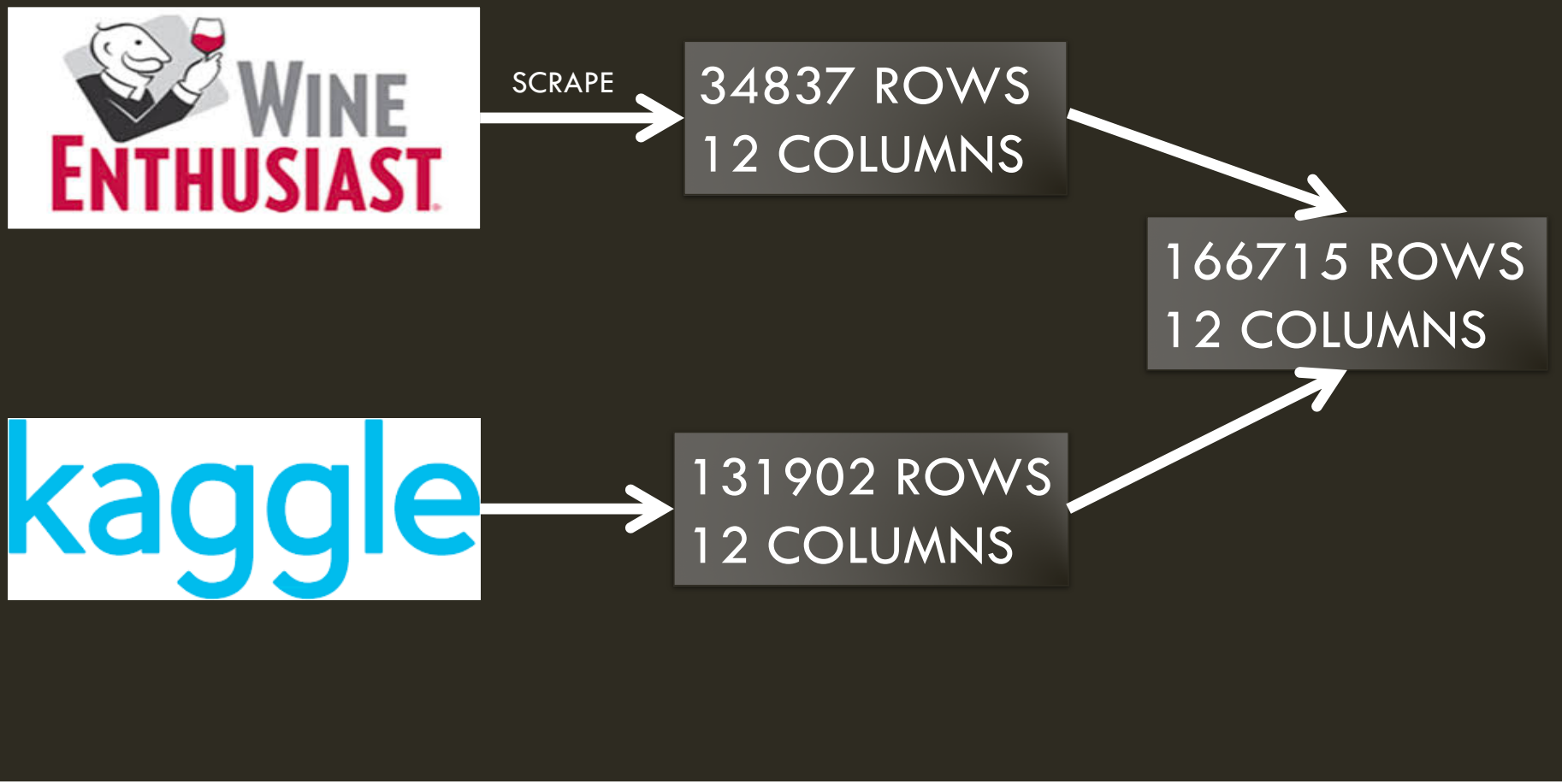


PROBLEM STATEMENT

COULD COMPUTERS DO BLIND WINE TASTING?


GOAL
To build a classification model that is able to tell what country, province and variety are the wines from, based on a wine description given by an expert

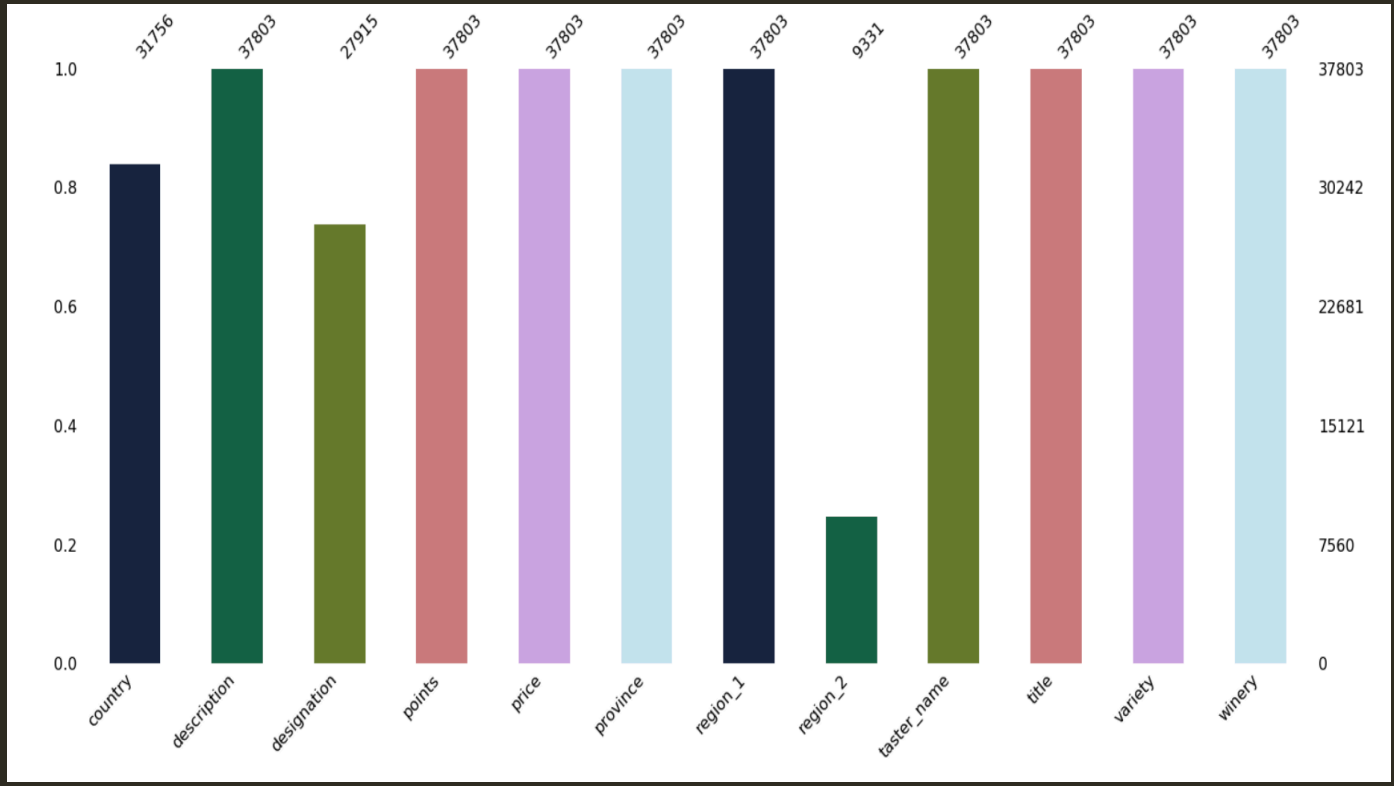
DATASETS


WINE ENTHUSIAST
SCRAPE
34837 ROWS
12 COLUMNS
kaggle
131902 ROWS
12 COLUMNS
166715 ROWS
12 COLUMNS

CLEAN THE DATA

NULL VALUES

country	6047
description	0
designation	9888
points	0
price	0
province	0
region_1	0
region_2	28472
taster_name	0
title	0
variety	0
winery	0



CLEAN THE DATA

Out[540]:

d:.1	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
0	France	Voluptuous and penetrating, this pure Grenache...	Cuvée de Mon Aïeul	99points	N/A,	Châteauneuf-du-Pape	Rhône Valley	NaN	Anna Lee C. Iijima	NaN	99pointsDomaine Pierre Usseglio et Fils 2016 C...	Grenache	Domaine Pierre Usseglio et Fils
1	France	Buxom and heady, this is a delightfully rich, ...	Vieilles Vignes	99points	$114,	Châteauneuf-du-Pape	Rhône Valley	NaN	Anna Lee C. Iijima	NaN	99pointsDomaine de la Janasse 2016 Vieilles Vi...	Rhône-style Red Blend	Domaine de la Janasse
2	France	Concentrated, unctuous black-cherry and plum f...	La Réserve des Cieux	98points	N/A,	Châteauneuf-du-Pape	Rhône Valley	NaN	Anna Lee C. Iijima	NaN	98pointsDomaine l'Abbé Dîne 2016 La Réserve de...	Grenache	Domaine l'Abbé Dîne
3	France	Sultry and silken on the palate, this wine sta...	La Réserve	98points	$175,	Châteauneuf-du-Pape	Rhône Valley	NaN	Anna Lee C. Iijima	NaN	98pointsDomaine le Clos du Caillou 2016 La Rés...	Rhône-style Red Blend	Domaine le Clos du Caillou
4	NaN	The wine's fine perfumed black plum fruits giv...	NaN	98points	$120,	Port	Port Blend	Portugal	Roger Voss	NaN	98pointsFonseca 2017 Port	Port	Fonseca

CLEAN THE DATA

Out[540]:

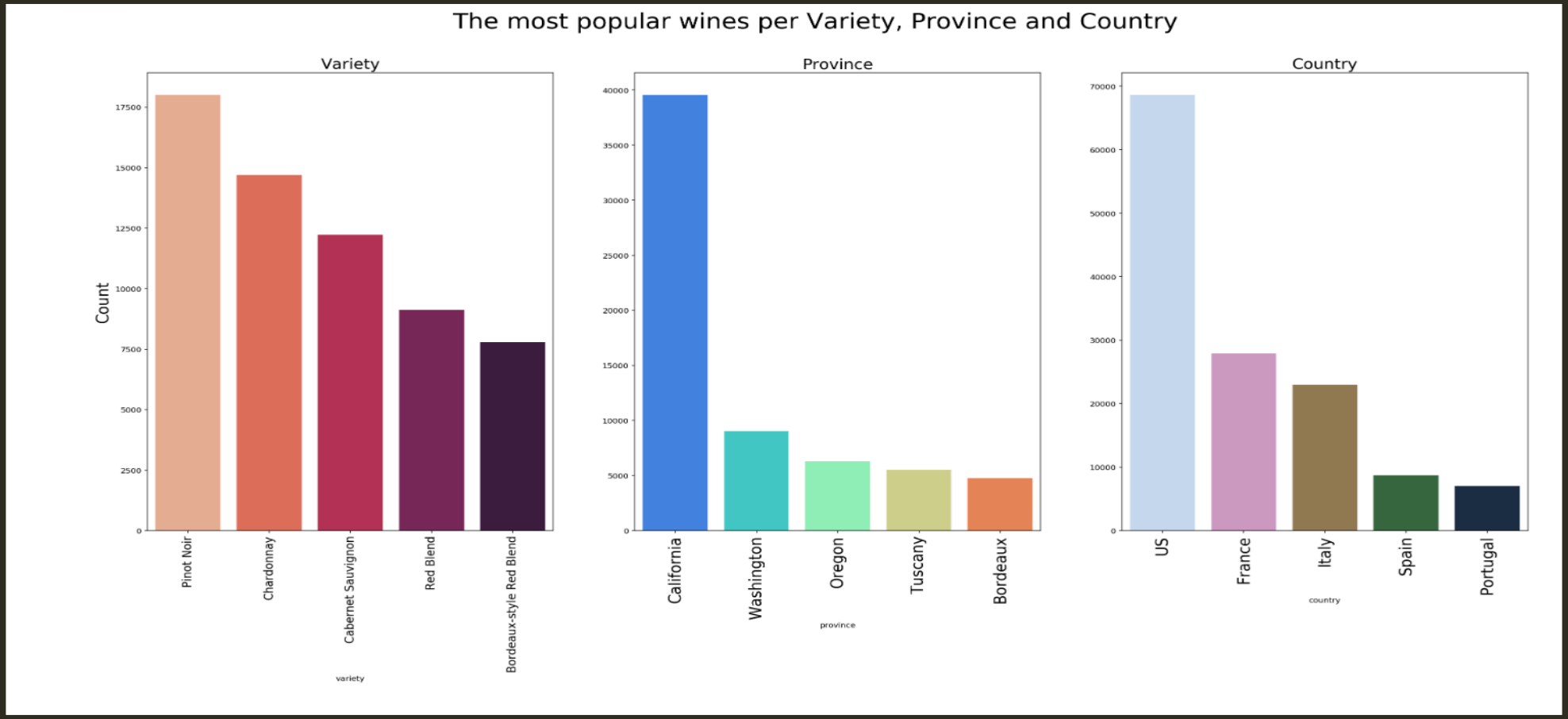
d: .1	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle		title	variety	winery
0	France	Voluptuous and penetrating, this pure Grenache...	Cuvée de Mon Aïeul	99points	N/A,	Châteauneuf-du-Pape	Rhône Valley	NaN	Anna Lee C. Iijima		NaN	99pointsDomaine Pierre Usseglio et Fils 2016 C...	Grenache	Domaine Pierre Usseglio et Fils
1	France	Buxom and heady, this is a delightfully rich, ...	Vieilles Vignes	99points	$114,	Châteauneuf-du-Pape	Rhône Valley		Anna Lee C. Iijima		NaN	99pointsDomaine de la Janasse 2016 Vieilles Vi...	Rhône-style Red Blend	Domaine de la Janasse
2	France	Concentrated, unctuous black-cherry and plum f...	La Réserve des Cieux	98points	N/A,	Châteauneuf-du-Pape	Rhône Valley	N	Anna Lee C. Iijima		NaN	98pointsDomaine l'Abbé Dîne 2016 La Réserve de...	Grenache	Domaine l'Abbé Dîne
3	France	Sultry and silken on the palate, this wine sta...	La Réserve	98points	$175,	Châteauneuf-du-Pape	Rhône Valley	N	Anna Lee C. Iijima		NaN	98pointsDomaine le Clos du Caillou 2016 La Rés...	Rhône-style Red Blend	Domaine le Clos du Caillou
4	NaN	The wine's fine perfumed black plum fruits giv...	NaN	98points	$120,	Port	Port Blend	Portuga	Roger Voss		NaN	98pointsFonseca 2017 Port	Port	Fonseca

vintage

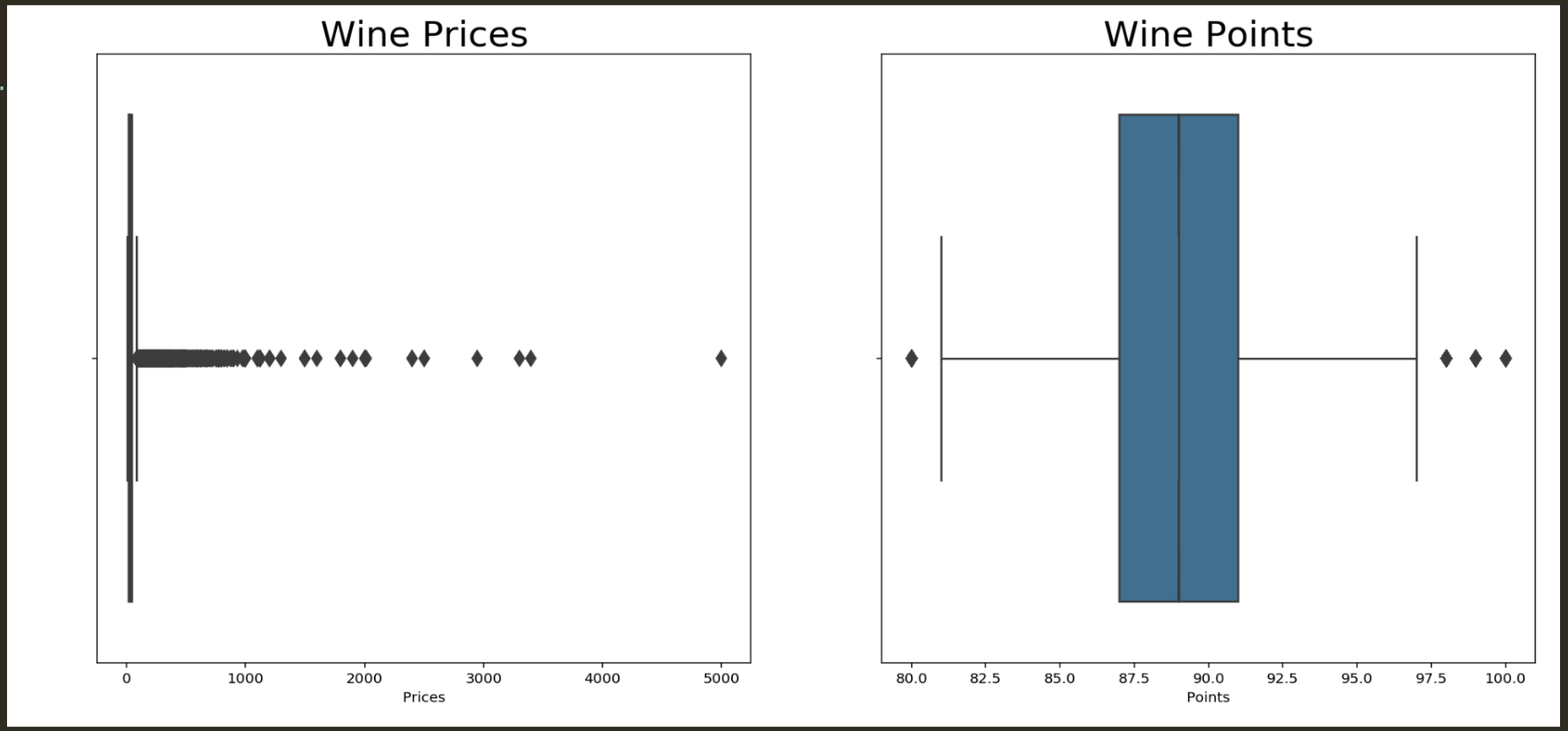
CLEAN THE DATA

country	description	designation	points	price	province	region_1	taster_name	title	variety	winery	vintage
Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	Roger Voss	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos	2011.0
US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Paul Gregutt	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm	2013.0
US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	Alexander Peartree	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian	2013.0
US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Paul Gregutt	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks	2012.0
Spain	Blackberry and raspberry aromas show a typical...	Ars In Vitro	87	15.0	Northern Spain	Navarra	Michael Schachner	Tandem 2011 Ars In Vitro Tempranillo-Merlot (N...	Tempranillo-Merlot	Tandem	2011.0

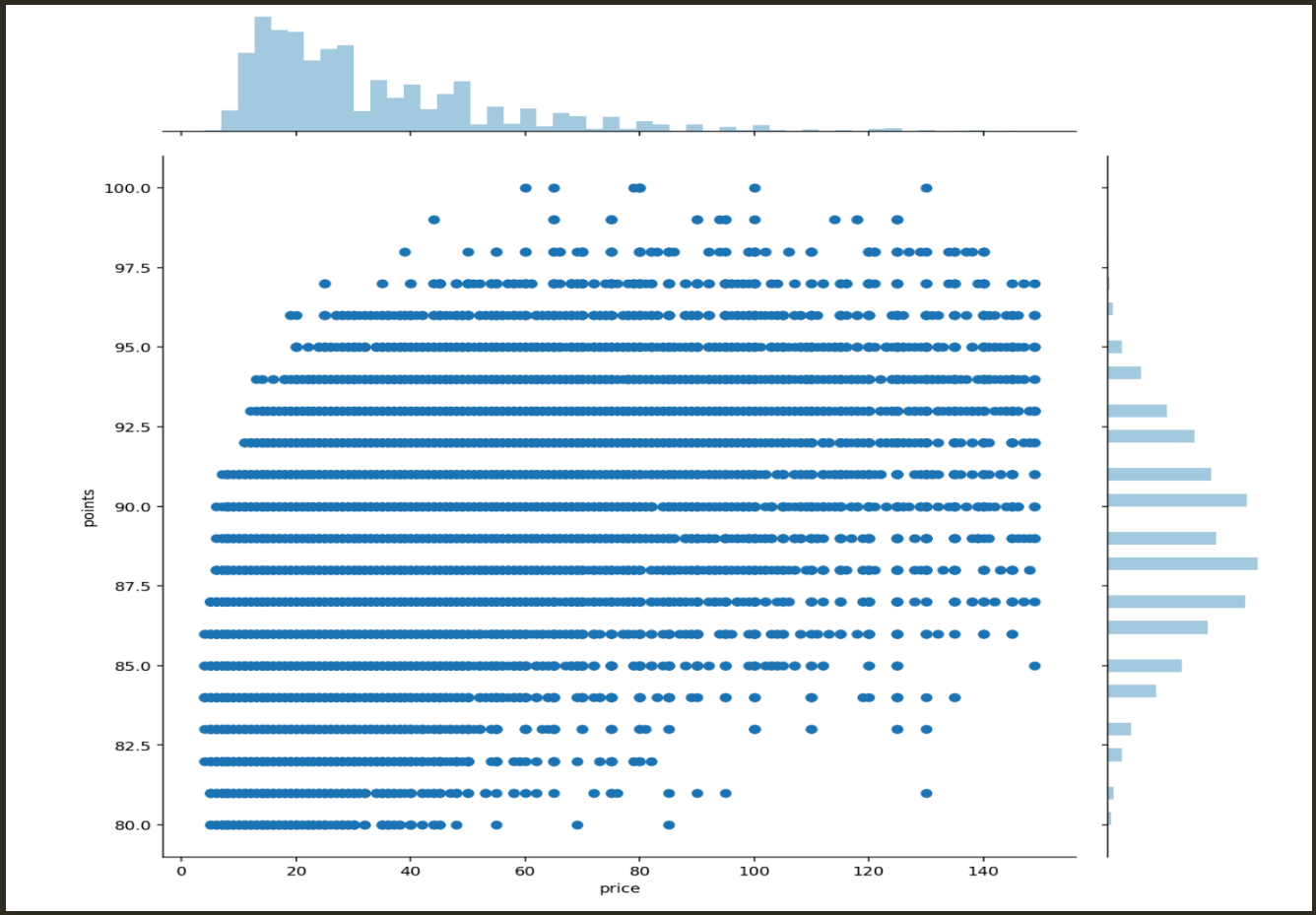
EXPLORATION DATA ANALISYS



EXPLORATION DATA ANALISYS



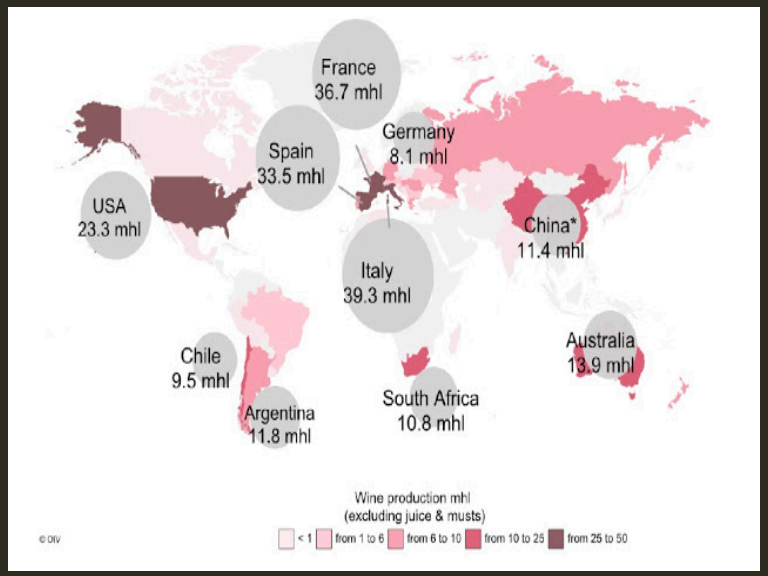
EXPLORATION DATA ANALYSIS



PREDICTOR
WINE DESCRIPTION

TARGETS

COUNTRY



VARIETY

PROVINCE

APPROACH

❑ SPLIT THE DATA: 80 % - 20%

		COUNTRY			VARIETY			PROVINCE	
DATASET	NUMBER	NORMAL	REDUCTION	STRATIFY	NORMAL	REDUCTION	STRATIFY	REDUCTION	STRATIFY
WINE ENTHUSIAST	42	✓							
	39		✓	✓					
	541				✓				
	57					✓	✓		
	77							✓	✓
kaggle	43	✓							
	39		✓	✓					
	87					✓	✓		
	84							✓	✓
WINE ENTHUSIAST kaggle	46	✓							
	42		✓	✓					
	89					✓	✓		
	149							✓	✓

MODEL SELECTION

1. **Natural Language Processing** → CountVectorizer / TfidfVectorizer

2. **Logistic Regression** → Lasso / Ridge

3. **Random Forest Classification** → 200 trees in the forest

4. **Naive Bayes** → MultinomialNB / BernoulliNB

COUNTRY

	Wine Enthusiast		kaggle		Wine Enthusiast + kaggle	
	COUNTRY	COUNTRY REDUCTIONS	COUNTRY	COUNTRY REDUCTIONS	COUNTRY	COUNTRY REDUCTIONS
Baseline	0.35931	0.36000	0.44360	0.10774	0.42600	0.10807
CountVectorizer and Logistic Regression with Ridge regularization	0.91339	0.91339	0.85178	0.85407	0.87310	0.87348
CountVectorizer and Logistic Regression with Lasso regularization	0.91124	0.91381	0.84974	0.84963	0.87037	0.87255
CountVectorizer and Logistic Regression with Lasso regularization and multinomial multiclass	0.91239	-	0.85504	0.85288	-	-
TfidfVectorizer and Logistic Regression with Ridge regularization	0.90033	-	0.81240	-	-	0.86145
TfidfVectorizer and Logistic Regression with Lasso regularization	0.90650	-	0.85095	-	0.86123	0.86397
TfidfVectorizer and Logistic Regression with Lasso regularization and multinomial class	0.90636	0.91079	-	-	-	-
Random Forest with CountVectorizer	0.88209	0.88475	0.74781	-	0.78265	-
Random Forest with TfidfVectorizer	0.88226	-	0.75110	0.75328	0.79114	0.79114
Random Forest with TfidfVectorizer and max_depth = 10	0.68820	-	0.75114	-	-	-
MultinomialNB with CountVectorizer	0.87103	0.87167	0.81952	0.72840	0.81799	0.81799
BernoulliNB with CountVectorizer	0.85939	0.85773	0.81437	-	0.82206	-
MultinomialNB with TfidfVectorizer	0.78371	-	0.72995	-	0.73565	-
MultinomialNB with CountVectorizer and TfidfTransformer	0.78342	-	0.72961	-	0.73511	-
Cross validation	0.90965	0.91282	0.85316	0.84172	0.86965	0.86982

COUNTRY

	Wine Enthusiast		kaggle		Wine Enthusiast + kaggle	
	COUNTRY	COUNTRY REDUCTIONS	COUNTRY	COUNTRY REDUCTIONS	COUNTRY	COUNTRY REDUCTIONS
Baseline	0.35931	0.36000	0.44360	0.10774	0.42600	0.10807
CountVectorizer and Logistic Regression with Ridge regularization	0.91339		0.85178	0.85407	0.87310	0.87348
CountVectorizer and Logistic Regression with Lasso regularization		0.91381				
Cross validation	0.90965	0.91282	0.85316	0.84172	0.86965	0.86982

PROVINCE

	Wine Enthusiast	kaggle	Wine Enthusiast + kaggle
Baseline	0.10805	0.30702	0.11543
CountVectorizer and Logistic Regression with Ridge regularization	0.55534	0.70735	0.68846
CountVectorizer and Logistic Regression with Lasso regularization	0.56304	0.71003	0.68636
CountVectorizer and Logistic Regression with Lasso regularization and multinomial multi class	0.54692	-	-
TfidfVectorizer and Logistic Regression with Ridge regularization	0.55341	0.57738	0.65853
TfidfVectorizer and Logistic Regression with Lasso regularization	0.56015	0.58380	0.68636
TfidfVectorizer and Logistic Regression with Lasso regularization and multinomial class	0.55510	0.70273	-
Random Forest with CountVectorizer	0.52382	0.53677	0.52174
Random Forest with TfidfVectorizer	-	0.54297	-
Random Forest with TfidfVectorizer and max depth = 10	0.43695	-	0.47790
# MultinomialNB with CountVectorizer	0.47930	0.63479	0.57831
BernoulliNB with CountVectorizer	0.40495	0.59819	0.53596
MultinomialNB with TfidfVectorizer	0.40255	0.43030	0.39322
MultinomialNB with CountVectorizer and TfidfTransformer	-	0.42951	-
Cross Validation	0.54262	0.70460	0.67990

PROVINCE

	Wine Enthusiast	kaggle	Wine Enthusiast + kaggle
Baseline	0.10805	0.30702	0.11543
CountVectorizer and Logistic Regression with Ridge regularization			0.68846
CountVectorizer and Logistic Regression with Lasso regularization	0.56304	0.71003	
Cross Validation	0.54262	0.70460	0.67990

VARIETY

	Wine Enthusiast		kaggle	Wine Enthusiast + kaggle
Baseline	0.10926	0.11847	0.11419	0.11543
CountVectorizer and Logistic Regression with Ridge regularization	0.57719	0.63640	0.63836	0.6645
CountVectorizer and Logistic Regression with Lasso regularization	0.59385	0.64185	0.64487	0.65616
CountVectorizer and Logistic Regression with Lasso regularization and multinomial mult_class	0.58782	0.63484	-	-
TfidfVectorizer and Logistic Regression with Ridge regularization	0.57274	0.61133	0.57738	0.63600
TfidfVectorizer and Logistic Regression with Lasso regularization	0.58394	0.62488	0.59023	0.64745
TfidfVectorizer and Logistic Regression with Lasso regularization and multinomial class	0.58667	0.63344	-	-
Random Forest with CountVectorizer	0.54875	0.59747	0.56556	0.64431
Random Forest with TfidfVectorizer	0.54559	-	0.56404	0.64494
Random Forest with TfidfVectorizer and max_depth = 10	0.33074	0.35970	-	-
# MultinomialNB with CountVectorizer	0.4707	0.53176	0.52760	0.53763
BernoulliNB with CountVectorizer	0.42294	0.47960	0.50208	0.51979
MultinomialNB with TfidfVectorizer	0.37440	0.41980	0.40012	0.42021
MultinomialNB with CountVectorizer and TfidfTransformer	0.37426	-	0.40020	-
Cross Validation	0.58221	0.63435	0.64016	0.67990

VARIETY

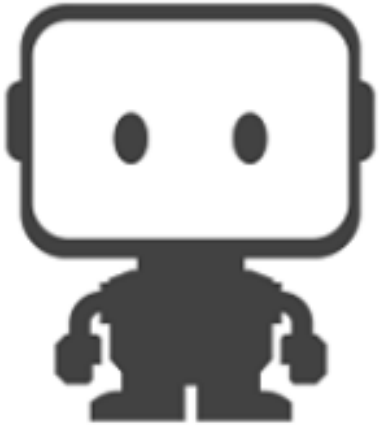
	Wine Enthusiast		kaggle	Wine Enthusiast + kaggle
Baseline	0.10926	0.11847	0.11419	0.11543
CountVectorizer and Logistic Regression with Ridge regularization				0.6645
CountVectorizer and Logistic Regression with Lasso regularization	0.59385	0.64185	0.64487	
Cross Validation	0.58221	0.63435	0.64016	0.67990

SUMMARY

DATASET	TARGET	BASELINE	ACCURACY	CROSS_VAL	LOGISTIC REGRESSION
	COUNTRY	0.35931	0.91339	0.90984	RIDGE
	COUNTRY REDUCE THE NUMBER AND STRATIFY	0.36000	0.91381	0.91282	LASSO
	VARIETY	0.10926	0.59385	0.58221	LASSO
	VARIETY REDUCING NUMBER AND STRATIFY:	0.11847	0.64185	0.63435	LASSO
	PROVINCE REDUCING NUMBER AND STRATIFY	0.10805	0.56256	0.54262	LASSO
	COUNTRY	0.44360	0.85679	0.85316	RIDGE
	COUNTRY REDUCE THE NUMBER AND STRATIFY	0.10774	0.85407	0.85012	RIDGE
	VARIETY REDUCING NUMBER AND STRATIFY:	0.11419	0.64487	0.64016	LASSO
	PROVINCE REDUCING NUMBER AND STRATIFY	0.30702	0.71003	0.70460	LASSO
	COUNTRY	0.42600	0.87310	0.86965	RIDGE
	COUNTRY REDUCE THE NUMBER AND STRATIFY	0.10807	0.87348	0.86982	RIDGE
	VARIETY REDUCING NUMBER AND STRATIFY:	0.11543	0.66456	0.67990	RIDGE
	PROVINCE REDUCING NUMBER AND STRATIFY	0.26774	0.68843	0.67990	RIDGE

CONCLUSION

- The simplest model has to be selected
- When the number of classes in the targets were reduced the model performed better
- The results don't show a huge difference between the datasets
- The data supports the initial hypothesis



LESSONS LEARNT

Improvements:

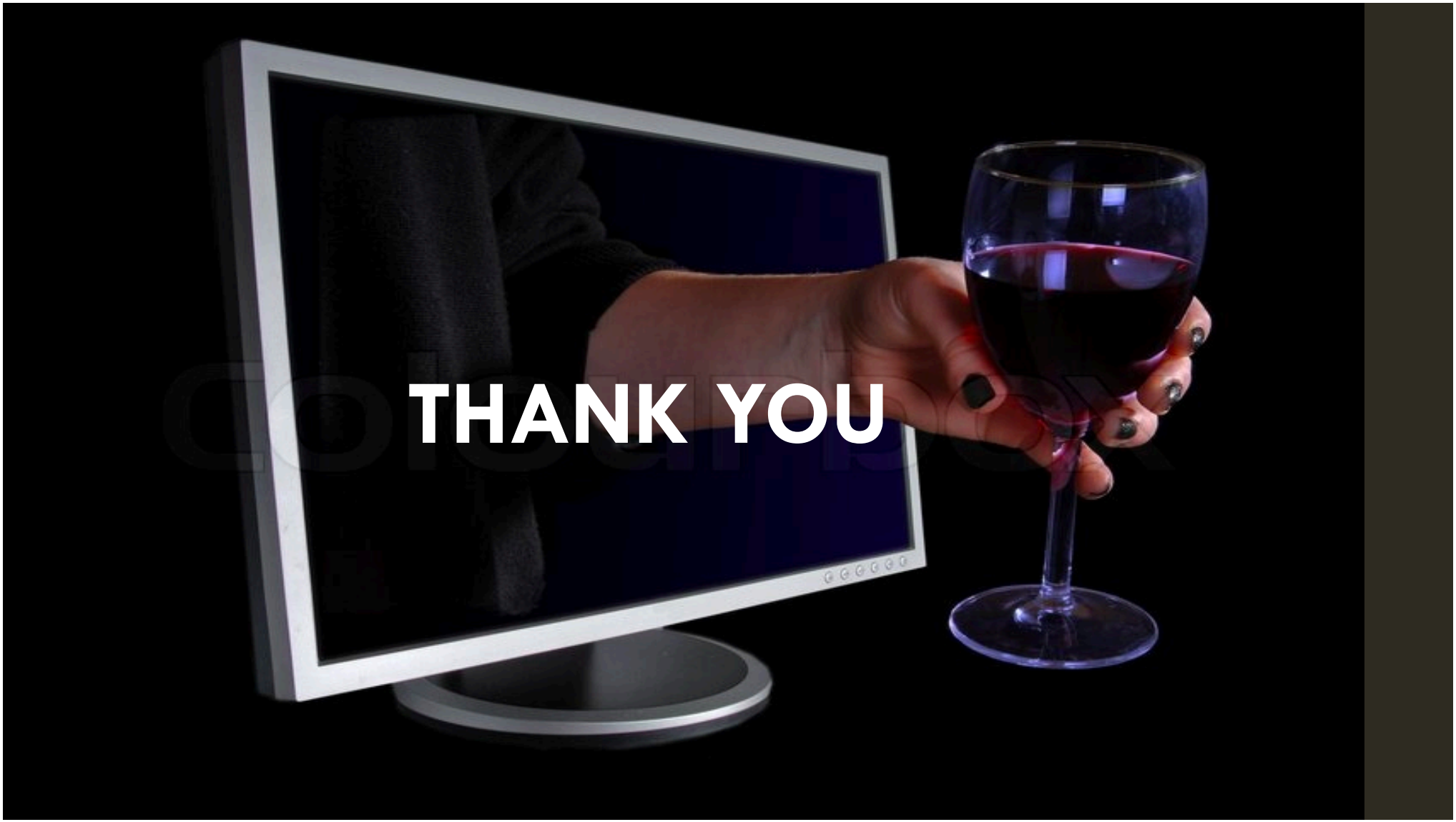
- Principal component analysis (PCA)
- Define other models
- Worked with other features
- GridSearch

Learn:

- Organize the ideas of what data you want to use before you start running models.
- Be careful to define the target
- Run a model cost a lot of time and it's essential to define the model carefully before start running it.

“All things being equal, the simplest solution tends to be the best one.”

William of Ockham


THANK YOU