SIT742: Modern Data Science

(Assessment Task 01: Wine Rating Data Exploration)

Student Information: Please fill your information below

- Name: Jose Arturo Gil Alonso
- Student ID: 218659676
- Email: jgilalonso@deakin.edu.au

Part 0. Data Files

0.1 Download Data

In [0]:
```
!pip install wget
```

```
Requirement already satisfied: wget in /usr/local/lib/python3.6/dist-pa
ckages (3.2)
```

In [0]:
```
import wget

link_to_data = 'https://github.com/tulip-lab/sit742/raw/master/Assessme
nt/2019/data/wine.json'
DataSet = wget.download(link_to_data)

link_to_data = 'https://github.com/tulip-lab/sit742/raw/master/Assessme
nt/2019/data/stopwords.txt'
```

```
DataSet = wget.download(link_to_data)
```

In [0]:
```
!ls
```

```
 sample_data            'stopwords (1).txt'   'wine (1).json'
 statisticByState.csv    stopwords.txt         wine.json
```

0.2 Load Data

In [0]:
```
import json
import pandas as pd
import matplotlib.pyplot as plt
```

In [0]:
```
file = 'wine.json'
```

In [0]:
```
import numpy as np
```

In [0]:
```
DAF=pd.read_json('wine.json')
```

In [0]:
```
# write your code here
# read the json file and drop rows with invalid values in the attribute
s of 'points' and 'price'.
```

In [0]:
```
import math
DAFE=pd.read_json(file)
DAFA= DAFE[DAFE['price'].notnull()& DAFE['points'].notnull()]

DAFA.shape[0] - DAFA.count()
DAFA['points'] = DAFA['points'].astype('float')
DAFA.head()
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:6: Setting
WithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
```

```
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-d
ocs/stable/indexing.html#indexing-view-versus-copy
```

Out[0]:

	country	description	designation	points	price	province	region_1	region_2	taster_name
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87.0	15.0	Douro	None	None	Roger Voss
2	US	Tart and snappy, the flavors of lime flesh and...	None	87.0	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt
3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87.0	13.0	Michigan	Lake Michigan Shore	None	Alexander Peartree
4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87.0	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt
5	Spain	Blackberry and raspberry aromas show a typical...	Ars In Vitro	87.0	15.0	Northern Spain	Navarra	None	Michael Schachner

Part 1: numeric anaysis

1.1 Explore the data distribution for each column.

In [0]:
```
# write your code here
# you may use functions such as describe() on each attribute
```

In [0]:
```
DAFA.describe()
```

Out[0]:

	points	price
count	120975.000000	120975.000000
mean	88.421881	35.363389
std	3.044508	41.022218
min	80.000000	4.000000
25%	86.000000	17.000000
50%	88.000000	25.000000
75%	91.000000	42.000000
max	100.000000	3300.000000

1.2 Find the 10 varieties of wine which receives the highest number of reviews

In [0]:
```
# write your code here
# you may use functions such as value_counts()
```

In [0]:
```
DAFA['variety'].value_counts().head(10)
```

Out[0]:
```
Pinot Noir                   12787
Chardonnay                   11080
Cabernet Sauvignon            9386
Red Blend                     8476
Bordeaux-style Red Blend      5340
Riesling                      4972
Sauvignon Blanc               4783
Syrah                         4086
Rosé                          3262
Merlot                        3062
Name: variety, dtype: int64
```

1.3 Find varieties of wine having the average price less than 20, with the average pointsat least 90

In [0]:
```
# write your code here
# you may use functions such as groupby()
```

In [0]:
```
DAFA.dtypes
DAFA1=DAFA.groupby('variety').agg({'points':'mean','price':'mean'})
DAFA1[(DAFA1["points"]>=90)&(DAFA1["price"]<20)]
```

Out[0]:

variety	points	price
Blauburgunder	93.0	19.000000
Caprettone	92.0	19.000000
Kotsifali	92.0	13.000000
Ondenc	90.0	15.000000
Roussanne-Grenache Blanc	91.0	16.000000
Shiraz-Malbec	90.0	18.666667
Tinta Cao	90.0	14.000000

1.4 Build statistic table

In [0]:
```
# write your code here
# you may use functions such as groupby() and round(decimals=2)
```

In [0]:
```
DAFA1.describe()
```

Out[0]:

	points	price
count	697.000000	697.000000
mean	87.854013	28.190650
std	2.014091	24.678005
min	80.500000	7.000000
25%	86.666667	18.000000
50%	88.000000	23.400000
75%	89.000000	32.000402
max	95.000000	495.000000

In [0]:
```
np.round(DAFA1.describe(),2)
```

Out[0]:

	points	price
count	697.00	697.00
mean	87.85	28.19
std	2.01	24.68
min	80.50	7.00
25%	86.67	18.00
50%	88.00	23.40

	points	price
75%	89.00	32.00
max	95.00	495.00

In [0]:
```
# save your table to 'statisticByState.csv'
```

In [0]:
```
np.round(DAFA1.describe(),2).to_csv("statisticByState.csv",sep=',')
```

1.5 Recommendations

Based on the analysis, which country/countries would you recommend *HOTEL TULIP* to source wine from? Pleas state your reasons.

Your Answer:

- Based on the analysis, 5 different types of wines were found, which reached a score of 87.00. Those are from countries such as Portugal, United States and Spain. It is the description:
- The first wine, which is from Douro, Portugal, is called Quinta Dos Avidagos. It Costs $15.00
- The second wine, which is from Oregon, US, with a designation of Rainstorm 2013, Pinot Gris. It costs $14.00
- The third wine, which is from Michigan, US, is St. Julian 2013 Reserve Late Harvest Riesling. It costs $13.00
- The four wine, which is from Oregon, US, is identified as Sweet Cheeks 2012. It is a Pinot Noir, with a cost of $65.00
- The last wine highly rated is from Northern Spain, which is recognised as Tandem 2011 Ars In Vitro. It costs $15.00

Furthermore, after analysing the Statistic Table, it was identified that Pinot Noir was the variety of wine that received the most amount of reviews and also was on the top after having filtered the

wines with ranking points greater than 90. It received 93 points, being recognised in Germany as Blauburgunder, which continued on the top in the sample of 697 reviews after the last filter.

For instance, the wines located in the US are a good option and from my point of view, after analysing the statistics, I would recommend to Hotel Tulip, the Pinot Noir variety of wine.

Part 2. Text Analysis

2.1 extract high requency words in description

In [0]:
```
import re
import nltk
from nltk.tokenize import RegexpTokenizer
from nltk.probability import *
from itertools import chain
#from tqdm import tqdm
import codecs
```

In [0]:
```
with open('stopwords.txt') as f:
    stop_words = f.read().splitlines()
stop_words = set(stop_words)
```

In [0]:
```
# write your code here
# define your tokenize

tokenizer = RegexpTokenizer(r"\w+(?:[-']\w+)?")
tkn=[]

for wine in DAFA['description']:
  tkn1=tokenizer.tokenize(wine)
  for text in tkn1:
    words=text.lower()
    if words not in stop_words:
      tkn.append(words)
```

In [0]:
```
# find top common words with document frequencies > 5000
# you may use function FreqDist() and sort()
```

In [0]:
```
from nltk.probability import *
cmw=FreqDist(tkn)
cmw.most_common(5000)
```

Out[0]:
```
[('wine', 70325),
 ('flavors', 60179),
 ('fruit', 41726),
 ('aromas', 37503),
 ('palate', 36399),
 ('finish', 33696),
 ('acidity', 31530),
 ('tannins', 28123),
 ('drink', 27873),
 ('cherry', 26063),
 ('ripe', 24304),
 ('black', 23824),
 ('notes', 18111),
 ('red', 17570),
 ('spice', 17209),
 ('oak', 16326),
 ('nose', 16319),
 ('rich', 15682),
 ('fresh', 15347),
 ('dry', 14572),
 ('berry', 14438),
 ('plum', 13510),
 ('blend', 12533),
 ('soft', 12392),
 ('blackberry', 12155),
 ('apple', 12122),
 ('fruits', 11785),
 ('offers', 11737),
 ('crisp', 11710),
 ('sweet', 11641),
 ('white', 11454),
 ('texture', 11284),
```

```
('citrus', 10705),
('shows', 10645),
('light', 10634),
('dark', 10511),
('vanilla', 10292),
('cabernet', 10232),
('bright', 9912),
('pepper', 9514),
('full', 9210),
('raspberry', 9021),
('good', 8995),
('juicy', 8907),
('green', 8570),
('fruity', 8411),
('firm', 8156),
('lemon', 8092),
('peach', 8039),
('chocolate', 7905),
('touch', 7878),
('balanced', 7325),
('dried', 7313),
('pear', 7295),
('sauvignon', 7183),
('character', 7072),
('years', 6953),
('spicy', 6795),
('structure', 6609),
('pinot', 6501),
('smooth', 6426),
('made', 6219),
('herb', 6060),
('tart', 6059),
('herbal', 6019),
('tannic', 6003),
('concentrated', 5826),
('note', 5738),
('bit', 5729),
('licorice', 5625),
('flavor', 5620),
```

```
('long', 5584),
('hint', 5577),
('merlot', 5513),
('fine', 5401),
('mineral', 5349),
('mouth', 5342),
('clean', 5334),
('balance', 5283),
('give', 5239),
('currant', 5237),
('creamy', 5234),
('toast', 5188),
('orange', 5154),
('syrah', 5137),
('wood', 5110),
('opens', 5042),
('style', 5009),
('savory', 5008),
('earthy', 5007),
('full-bodied', 5000),
('lead', 4999),
('age', 4998),
('alongside', 4980),
('leather', 4927),
('slightly', 4918),
('vineyard', 4874),
('tobacco', 4867),
('hints', 4800),
('lime', 4746),
('dense', 4735),
('elegant', 4687),
('delicious', 4655),
('structured', 4577),
('tight', 4537),
('chardonnay', 4528),
('complex', 4495),
('aging', 4440),
('ready', 4421),
('great', 4414),
```

```
('mouthfeel', 4408),
('cassis', 4395),
('herbs', 4297),
('cherries', 4159),
('feels', 4124),
('smoky', 4123),
('time', 4086),
('melon', 4026),
('lively', 3998),
('richness', 3990),
('young', 3980),
('strawberry', 3970),
('cola', 3951),
('coffee', 3933),
('color', 3927),
('freshness', 3918),
('delivers', 3879),
('vintage', 3874),
('grapefruit', 3850),
('feel', 3789),
('clove', 3775),
('2018', 3674),
('yellow', 3661),
('big', 3639),
('pineapple', 3633),
('tastes', 3626),
('wild', 3624),
('honey', 3606),
('floral', 3573),
('simple', 3563),
('earth', 3563),
('bottling', 3550),
('easy', 3505),
('finishes', 3496),
('tropical', 3481),
('round', 3433),
('refreshing', 3429),
('apricot', 3395),
('core', 3371),
```

```
('intense', 3345),
('solid', 3342),
('scents', 3340),
('cinnamon', 3317),
('show', 3277),
('lightly', 3277),
('showing', 3266),
('complexity', 3243),
('attractive', 3234),
('alcohol', 3227),
('minerality', 3226),
('stone', 3193),
('baked', 3184),
('plenty', 3180),
('noir', 3141),
('generous', 3100),
('cranberry', 3097),
('pretty', 3093),
('make', 3091),
('body', 3086),
('tangy', 3080),
('toasty', 3032),
('glass', 3022),
('edge', 2988),
('crushed', 2958),
('spices', 2946),
('cedar', 2945),
('weight', 2940),
('blueberry', 2937),
('bitter', 2905),
('grapes', 2850),
('smoke', 2842),
('zest', 2827),
('medium', 2804),
('franc', 2789),
('mocha', 2788),
('oaky', 2781),
('strong', 2752),
('flower', 2741),
```

```
('silky', 2730),
('2020', 2726),
('2017', 2711),
('blanc', 2699),
('concentration', 2694),
('toasted', 2682),
('bodied', 2679),
('zesty', 2660),
('nice', 2644),
('barrel', 2639),
('jammy', 2622),
('delicate', 2589),
('wines', 2573),
('medium-bodied', 2567),
('roasted', 2565),
('mix', 2527),
('bottle', 2524),
('anise', 2508),
('almond', 2502),
('perfumed', 2479),
('powerful', 2462),
('lush', 2460),
('layers', 2454),
('deep', 2452),
('mature', 2451),
('high', 2444),
('aromatic', 2444),
('subtle', 2443),
('tones', 2425),
('mint', 2403),
('bold', 2397),
('makes', 2394),
('peppery', 2386),
('dusty', 2381),
('offering', 2377),
('vibrant', 2377),
('depth', 2376),
('aftertaste', 2373),
('riesling', 2373),
```

```
('malbec', 2364),
('rounded', 2353),
('rose', 2353),
('candied', 2347),
('baking', 2317),
('whiff', 2277),
('pure', 2275),
('caramel', 2269),
('warm', 2266),
('jam', 2264),
('hard', 2262),
('brings', 2255),
('grape', 2248),
('textured', 2245),
('sweetness', 2242),
('end', 2231),
('espresso', 2213),
('peel', 2197),
('lean', 2196),
('thick', 2195),
('blackberries', 2192),
('pair', 2169),
('power', 2163),
('exotic', 2163),
('berries', 2160),
('meat', 2158),
('heavy', 2144),
('french', 2141),
('flowers', 2139),
('tangerine', 2123),
('lingering', 2117),
('aged', 2099),
('bouquet', 2094),
('lovely', 2086),
('grenache', 2082),
('fruitiness', 2075),
('sour', 2065),
('brisk', 2060),
('velvety', 2051),
```

```
('vineyards', 2027),
('valley', 2026),
('100', 2012),
('layered', 2010),
('drinking', 1999),
('currants', 1991),
('acids', 1991),
('forward', 1982),
('side', 1976),
('estate', 1961),
('nicely', 1952),
('2016', 1951),
('price', 1947),
('2019', 1946),
('polished', 1938),
('petit', 1937),
('grilled', 1933),
('supple', 1925),
('varietal', 1918),
('enjoy', 1908),
('violet', 1886),
('verdot', 1865),
('chewy', 1860),
('tannin', 1859),
("wine's", 1859),
('10', 1850),
('sugar', 1839),
('sangiovese', 1835),
('taste', 1835),
('quality', 1827),
('sharp', 1782),
('tea', 1773),
('rosé', 1771),
('elegance', 1768),
('develop', 1756),
('vines', 1744),
('intensity', 1743),
('giving', 1742),
('classic', 1716),
```

```
('lots', 1714),
('accents', 1696),
('months', 1691),
('fragrant', 1690),
('cut', 1687),
('variety', 1670),
('add', 1665),
('impressive', 1657),
('year', 1654),
('prune', 1652),
('sense', 1652),
('nectarine', 1648),
('adds', 1646),
('sage', 1646),
('provide', 1645),
('apples', 1644),
('5', 1614),
('cocoa', 1613),
('cellar', 1611),
('packed', 1603),
('olive', 1603),
('blue', 1601),
('close', 1598),
('cab', 1597),
('open', 1585),
('expression', 1577),
('accented', 1570),
('pleasant', 1562),
('sip', 1545),
('raspberries', 1538),
('ripeness', 1511),
('integrated', 1499),
('finishing', 1492),
('blossom', 1489),
('smells', 1486),
('straightforward', 1472),
('acidic', 1460),
('wet', 1459),
('plump', 1438),
```

```
('touches', 1438),
('produced', 1434),
('mild', 1430),
('cool', 1422),
('length', 1420),
('beautiful', 1408),
('midpalate', 1398),
('pie', 1396),
('racy', 1395),
('forest', 1389),
('leaf', 1364),
('aroma', 1360),
('juice', 1357),
('raisin', 1357),
('carry', 1351),
('tightly', 1346),
('tomato', 1338),
('buttered', 1332),
('front', 1329),
('pink', 1317),
('ample', 1316),
('set', 1306),
('fleshy', 1302),
('beautifully', 1301),
('rustic', 1300),
('food', 1292),
('chunky', 1289),
('backed', 1286),
('astringent', 1278),
('back', 1275),
('raw', 1264),
('plums', 1261),
('lot', 1261),
('interesting', 1260),
('focused', 1259),
('graphite', 1258),
('citrusy', 1257),
('honeysuckle', 1257),
('lemony', 1257),
```

```
('potential', 1240),
('viognier', 1221),
('purple', 1221),
('2015', 1220),
('skin', 1215),
('20', 1209),
('appealing', 1209),
('short', 1200),
('streak', 1199),
('making', 1198),
('hot', 1194),
('elements', 1194),
('plush', 1193),
('mango', 1190),
('suggest', 1186),
('find', 1186),
('petite', 1175),
('pomegranate', 1171),
('approachable', 1170),
('underbrush', 1165),
('framed', 1162),
('heat', 1159),
('meaty', 1153),
('beef', 1149),
('excellent', 1149),
('appellation', 1133),
('fine-grained', 1133),
('boysenberry', 1132),
('off-dry', 1129),
('grip', 1124),
('leathery', 1122),
('menthol', 1117),
('robust', 1116),
('soil', 1115),
('winery', 1114),
('fairly', 1112),
('honeyed', 1110),
('coconut', 1106),
('candy', 1105),
```

```
('2022', 1097),
('zinfandel', 1095),
('spiced', 1092),
('lift', 1082),
('smoked', 1081),
('sirah', 1078),
('cream', 1078),
('mourvèdre', 1078),
('top', 1075),
('opulent', 1071),
('minty', 1069),
('black-fruit', 1068),
('15', 1065),
('extra', 1064),
('refined', 1057),
('perfect', 1055),
('thyme', 1052),
('nutmeg', 1051),
('steely', 1049),
('dominate', 1041),
('boasts', 1036),
('gentle', 1032),
('fully', 1030),
('salty', 1028),
('butter', 1021),
('woody', 1017),
('restrained', 1016),
('lend', 1016),
('rubbery', 1013),
('support', 1012),
('broad', 1011),
('fig', 1002),
('features', 1000),
('3', 999),
('linger', 998),
('small', 997),
('tar', 996),
('slight', 995),
('hold', 993),
```

```
('tasty', 990),
('brown', 989),
('accent', 987),
('appeal', 984),
('intensely', 983),
('winemaker', 983),
('chalky', 982),
('charred', 982),
("winery's", 972),
('taut', 967),
('2025', 954),
('richly', 953),
('sandalwood', 953),
('bacon', 951),
('inviting', 948),
('yeasty', 946),
('bean', 946),
('modest', 945),
('fermented', 943),
('mouthwatering', 943),
('bordeaux', 939),
('density', 939),
('doles', 936),
('region', 934),
('background', 931),
('luscious', 931),
('low', 930),
('carries', 929),
('drying', 928),
('producer', 928),
('starts', 925),
('floor', 918),
('suggests', 915),
('rind', 903),
('wrapped', 903),
('sourced', 902),
('blended', 902),
('bracing', 900),
('buttery', 900),
```

```
('easy-drinking', 897),
('intriguing', 896),
('soften', 894),
('star', 892),
('california', 890),
('golden', 888),
('coming', 887),
('basic', 886),
('element', 879),
('4', 876),
('profile', 874),
('easygoing', 871),
('backbone', 870),
('thin', 870),
('30', 868),
('moderate', 867),
('pressed', 866),
('succulent', 866),
('typical', 864),
('balsamic', 862),
('de', 862),
('pears', 858),
('minerally', 858),
('grass', 857),
('softly', 856),
('grown', 854),
('banana', 852),
('residual', 849),
('family', 849),
('lifted', 842),
('game', 841),
('mellow', 837),
('linear', 835),
('bitterness', 830),
('immediately', 830),
('marked', 827),
('milk', 827),
('bring', 826),
('remains', 823),
```

```
('peaches', 822),
('mountain', 822),
('mark', 820),
('highlights', 819),
('tough', 817),
('lingers', 815),
('briny', 812),
('drinkable', 811),
('stage', 811),
('unusual', 807),
('zin', 807),
('leads', 805),
('stalky', 804),
('combine', 803),
('lacks', 802),
('firmly', 802),
('loaded', 799),
('grippy', 798),
('napa', 798),
('leafy', 797),
('mushroom', 797),
('quickly', 796),
('gritty', 795),
('2021', 795),
('austere', 794),
('suggesting', 793),
('subdued', 791),
('burnt', 791),
('6', 783),
('butterscotch', 783),
('varieties', 782),
('dishes', 781),
('brooding', 780),
('2', 780),
('effort', 780),
('sauce', 779),
('chocolaty', 779),
('ground', 776),
('leaves', 773),
```

```
('finely', 771),
('light-bodied', 769),
('flat', 767),
('frame', 767),
('pencil', 763),
('summer', 763),
('acid', 762),
('sparkling', 762),
('red-berry', 761),
('40', 760),
('accessible', 759),
('friendly', 758),
('red-fruit', 758),
('chopped', 756),
('youthful', 755),
('neutral', 753),
('assertive', 749),
('cake', 748),
('offer', 747),
('extracted', 746),
('ginger', 745),
('asian', 742),
('wound', 737),
('late', 736),
('bite', 733),
('50', 732),
('crispness', 732),
('grassy', 731),
('selection', 731),
('nuances', 730),
('lasting', 729),
('overripe', 726),
('lavender', 726),
('start', 725),
('barrels', 725),
('pleasing', 725),
('skins', 722),
('lychee', 720),
('violets', 714),
```

```
('freshly', 711),
('hearty', 709),
('impression', 708),
('natural', 707),
('enjoyable', 706),
('stewed', 706),
('everyday', 705),
('personality', 704),
('dominated', 702),
('char', 700),
('rough', 696),
('planted', 695),
('pithy', 693),
('play', 691),
('warmth', 691),
('brambly', 690),
('enticing', 689),
('turns', 688),
('interest', 686),
('emerge', 685),
('pine', 684),
('tempranillo', 683),
('single-vineyard', 683),
('tasting', 682),
('elegantly', 681),
('flavorful', 681),
('extremely', 681),
('slate', 679),
('high-toned', 679),
('passion', 676),
('rubber', 676),
('spring', 674),
('dessert', 674),
('final', 674),
('nut', 671),
('pith', 667),
('rhubarb', 667),
('grabby', 665),
('generic', 663),
```

```
('compact', 662),
('perfectly', 662),
('purity', 662),
('persistent', 662),
('things', 661),
('sparkler', 661),
('lighter', 659),
('steel', 657),
('black-cherry', 656),
('center', 656),
('deliciously', 656),
('form', 655),
('adding', 655),
('finesse', 654),
('eucalyptus', 654),
('strongly', 652),
('displays', 651),
('flavored', 651),
('oily', 649),
('deeply', 649),
('steak', 647),
('satisfying', 646),
('2023', 646),
('citric', 645),
('closes', 643),
('meet', 642),
('heady', 642),
('2010', 640),
('pale', 640),
('blends', 639),
('site', 638),
('focus', 638),
('range', 630),
('bread', 630),
('papaya', 630),
('lemon-lime', 627),
('modern', 626),
('takes', 625),
('surprisingly', 625),
```

```
('perfume', 625),
('8', 624),
('wait', 622),
('gorgeous', 622),
('richer', 622),
('resiny', 622),
('easily', 621),
('part', 619),
('bone', 619),
('line', 619),
('feeling', 618),
('jasmine', 616),
('apéritif', 616),
('gris', 616),
('nutty', 616),
('include', 614),
('mixed', 614),
('syrup', 614),
('leading', 613),
('mousse', 612),
('reveals', 610),
('25', 609),
('snappy', 608),
('60', 604),
('additional', 604),
('dryness', 603),
('honeydew', 601),
('2014', 601),
('screwcap', 597),
('sugary', 597),
('opening', 596),
('expressive', 596),
('combines', 595),
('yeast', 595),
('point', 594),
('improve', 593),
('well-balanced', 592),
('2012', 591),
('tang', 591),
```

```
('moderately', 590),
('minerals', 589),
('meats', 585),
('half', 585),
('layer', 585),
('stony', 585),
('slender', 585),
('medicinal', 584),
('herbaceous', 583),
('cru', 583),
('edges', 583),
('lees', 583),
('petal', 582),
('fennel', 582),
('real', 582),
('mediterranean', 580),
('amount', 579),
('supported', 577),
('shiraz', 575),
('laced', 575),
('mildly', 574),
('meyer', 574),
('cases', 573),
('bell', 573),
('label', 571),
('named', 571),
('drinks', 570),
('compelling', 569),
('whiffs', 567),
('result', 566),
('true', 565),
('flesh', 564),
('stainless', 563),
('bramble', 562),
('sophisticated', 562),
('wonderful', 560),
('truffle', 559),
('slowly', 558),
('shy', 557),
```

```
('astringency', 557),
('apricots', 553),
('syrupy', 551),
('12', 550),
('vivid', 549),
('combination', 548),
('includes', 548),
('southern', 548),
('cured', 545),
('riper', 544),
('lends', 544),
('funky', 544),
('wonderfully', 544),
('level', 543),
('caramelized', 540),
('substantial', 540),
('petals', 540),
('appetizing', 540),
('reserve', 538),
('developing', 537),
('gooseberry', 531),
('leave', 530),
('loads', 530),
('oregano', 529),
('holds', 528),
('salt', 527),
('waves', 527),
('aromatics', 527),
('brightened', 526),
('mingle', 525),
('drunk', 525),
('softness', 525),
('river', 525),
('sea', 522),
('equal', 522),
('overly', 522),
('barbecue', 520),
('table', 520),
('release', 520),
```

```
('direct', 519),
('turn', 516),
('strawberries', 516),
('2011', 516),
('penetrating', 516),
('decent', 511),
('tilled', 507),
('raisiny', 506),
('dominates', 505),
('2030', 505),
('streaks', 505),
('tone', 504),
('vegetal', 501),
('burst', 501),
('parts', 500),
('cheese', 500),
('present', 500),
('kiwi', 498),
('closed', 498),
('fat', 497),
('bubbles', 497),
('fermentation', 497),
('including', 496),
('pleasantly', 495),
('underlying', 495),
('future', 494),
('terms', 494),
('tension', 493),
('smelling', 492),
('dill', 490),
('saturated', 490),
('delicately', 489),
('balances', 489),
('huge', 487),
('lengthy', 487),
('standard', 486),
('liqueur', 483),
('cigar', 482),
('blossoms', 482),
```

```
('hazelnut', 482),
('edgy', 481),
('hills', 480),
('fare', 479),
('hay', 478),
('champagne', 475),
('root', 475),
('italy', 474),
('early', 473),
('work', 473),
('influence', 472),
('pleasure', 472),
('stand', 472),
('drop', 472),
('sensations', 472),
('complete', 471),
('nero', 471),
('distinctive', 469),
('nuanced', 469),
('pineapples', 469),
('limes', 469),
('built', 468),
('70', 468),
('fun', 467),
('black-skinned', 466),
('recall', 465),
('orchard', 464),
('touriga', 463),
('reminiscent', 463),
('oaked', 462),
('nuts', 462),
('bay', 462),
('roussanne', 462),
('80', 461),
('peppercorn', 461),
('follow', 460),
('rare', 460),
('continue', 459),
('2009', 458),
```

```
('black-currant', 458),
('fill', 458),
('expect', 457),
('framework', 456),
('vintages', 456),
('worth', 456),
('2024', 455),
('guava', 454),
('pungent', 454),
('totally', 454),
('14', 453),
('rest', 453),
('continues', 452),
('inky', 451),
('providing', 451),
('asphalt', 449),
('hits', 448),
('ends', 448),
('leaving', 448),
('resin', 448),
('santa', 448),
('considerable', 448),
('nature', 447),
('tongue', 447),
('unique', 447),
('foods', 446),
('due', 446),
('prominent', 445),
('benefit', 444),
('american', 444),
('sticky', 444),
('based', 443),
('shape', 443),
('bringing', 443),
('stylish', 442),
('scent', 441),
('2013', 441),
('alluring', 440),
('added', 439),
```

```
('hit', 438),
('grigio', 438),
('primary', 438),
('briary', 437),
('evident', 437),
('characteristics', 436),
('marks', 436),
('earthiness', 436),
('kick', 436),
('cardamom', 433),
('cool-climate', 433),
('waxy', 433),
('promise', 431),
('sleek', 430),
('run', 430),
('match', 430),
('seductive', 430),
('entry', 429),
('reveal', 429),
('oregon', 428),
('healthy', 428),
('fragrance', 428),
('experience', 428),
('pairing', 427),
('watermelon', 426),
('decade', 426),
('scratchy', 425),
('upfront', 425),
('chicken', 424),
('densely', 423),
('plummy', 422),
('chard', 422),
('crème', 422),
('tartness', 422),
('amounts', 420),
('heart', 420),
('presents', 419),
('equally', 418),
('single', 418),
```

```
 ('enjoyment', 417),
 ('cellaring', 416),
 ('keeping', 416),
 ('county', 416),
 ('cuvée', 416),
 ('stone-fruit', 415),
 ('pipe', 415),
 ('lamb', 415),
 ('oil', 415),
 ('wide', 413),
 ('flint', 413),
 ('iron', 412),
 ('sunny', 412),
 ('brawny', 411),
 ('past', 411),
 ('middle', 411),
 ('fruit-forward', 411),
 ('pasta', 410),
 ('pork', 409),
 ('create', 408),
 ('weighty', 408),
 ('muscular', 408),
 ('blueberries', 408),
 ('sémillon', 407),
 ('bordeaux-style', 407),
 ('pastry', 406),
 ('textural', 406),
 ('1', 405),
 ('rosemary', 405),
 ('choppy', 405),
 ('extract', 402),
 ('similar', 402),
 ...]
```

In [0]:
```
# save your table to 'top_common_words.txt'
```

In [0]:
```
LF=open("top_common_words.txt","w")
for line in cmw:
```

```
  LF.write(words+':'+str(cmw)+'\n')
```

2.2 Find key words for describing Shiraz using TF-IDF

In [0]:
```
# select 'description' from 'variety' eqaul to  'Shiraz'
```

In [0]:
```
EDT=DAFA[['description','variety']]
```

In [0]:
```
DAFA3=EDT[EDT['variety']=='Shiraz']
DAFA3
```

Out[0]:

	description	variety
232	Lifted cedar and pine notes interspersed with ...	Shiraz
293	This wine displays ample concentration in its ...	Shiraz
356	Dusty, firm, powerful: just a few apt descript...	Shiraz
365	The Taylor family selected Clare Valley for it...	Shiraz
628	This plush, full-bodied wine is already delici...	Shiraz
636	The Torbreck wines, with the emphasis they pla...	Shiraz
893	This is an unusually complex Shiraz for the pr...	Shiraz
997	Blended from a patchwork of old vineyards thro...	Shiraz
1008	Intense and focused, this good value Coonawarr...	Shiraz
1074	Sourced from vines planted in 1847, this is a ...	Shiraz
1154	This wine boasts admirable complexity, but com...	Shiraz
1715	Raspberry and cherry mingle with essence of mo...	Shiraz
1723	Tart blackberry and vanilla flavors mark this ...	Shiraz
1886	Big, bold and structured, this is a dry yet mo...	Shiraz
1934	Mint marks the nose of this otherwise darkly a...	Shiraz

	description	variety
2256	Not a wine for the faint of heart. Alluring no...	Shiraz
2594	Made in a superripe, dense, almost fudge-like ...	Shiraz
2598	An easy-drinking introduction to Shiraz from W...	Shiraz
2631	A drink-now style of warming Shiraz, with plen...	Shiraz
2981	You've got to like oak to appreciate this wine...	Shiraz
2982	Mouthfilling and round, this wine's savory not...	Shiraz
2993	A solid value, the 2013 Pillar Box offers oodl...	Shiraz
3099	From a cooler region than its Bin 28 stablemat...	Shiraz
3135	Full-bodied and firm, this dark-fruited Shiraz...	Shiraz
3288	Smoky, cedary notes nicely complement the blue...	Shiraz
3554	This luxury cuvée from Thorne Clarke combines ...	Shiraz
3588	Dark ruby in color, this wine offers raspberry...	Shiraz
3598	Really dark, vibrant purple in color, with oak...	Shiraz
3952	Herbal and green peppercornish, with dark berr...	Shiraz
4232	Fleshy black plums and berries burst from nose...	Shiraz
...	...	...
125427	Like Spinal Tap, this wine turns the volume up...	Shiraz
125435	Fully mature, this flavorful, authentic-tastin...	Shiraz
125436	The Dead Arm rarely wows this reviewer in our ...	Shiraz
125438	This Shiraz offers wonderful purity of fruit, ...	Shiraz
125440	Intensely toasty on the nose, along with power...	Shiraz
125441	Dense and fudge-like in consistency, with arom...	Shiraz
125505	Tobacco, cassis and vanilla notes bring the ar...	Shiraz
125514	In the context of 2010, I have to admit to som...	Shiraz
125641	Not simply a jammy mouthful of fruity richness...	Shiraz

	description	variety
125648	Mellow and composed on the nose, with a nice b...	Shiraz
125820	A surprisingly good little Syrah. The nose is ...	Shiraz
125875	Deep red violet in color, this wine has a bouq...	Shiraz
125878	Aromas of blueberry and clove, with a soft hin...	Shiraz
126321	Ripe dark berries meet chocolate, herbs and ch...	Shiraz
126449	Sourced from one of the world's oldest survivi...	Shiraz
126977	Starts out with proprietary aromas of clove, s...	Shiraz
127011	Mint, black cherries and mocha aromas and flav...	Shiraz
127122	Rubber, plum and pepper notes open this light-...	Shiraz
127766	Rocland puts out a number of inexpensive wines...	Shiraz
127816	Dark and somewhat heady, this opens with dense...	Shiraz
127822	A pleasing melange of spicy aromas start the n...	Shiraz
128278	Savvy shoppers may be able to find this for ab...	Shiraz
128742	This is a powerful Shiraz, with heady notes of...	Shiraz
128867	This spicy, peppery Shiraz is elegant and stra...	Shiraz
128918	A meaty Shiraz that pulls no punches, proudly ...	Shiraz
129107	This full-bodied, single-vineyard bottling has...	Shiraz
129389	Banrock Station's 2006 Shiraz is a throwback t...	Shiraz
129446	Black licorice, raisiny fruit and a little swe...	Shiraz
129447	This win is dense and dusty, with ripe ripe re...	Shiraz
129944	Deep garnet in the glass, this has a nose of b...	Shiraz

822 rows × 2 columns

In [0]:
```
# use TfidfVectorizer to calculate TF-IDF score
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
Tfidf = TfidfVectorizer(analyzer='word', stop_words = 'english')
```

In [0]:
```
LA=Tfidf.fit_transform(DAFA3['description'])
print(LA)
```

```
  (0, 1480)     0.22052677909080337
  (0, 459)      0.13451531230510097
  (0, 1900)     0.2589218360130864
  (0, 1739)     0.17157438131096184
  (0, 1356)     0.2747939726342262
  (0, 169)      0.20806074680652425
  (0, 2924)     0.2747939726342262
  (0, 2276)     0.1565579136291564
  (0, 2880)     0.1743877334441604
  (0, 2573)     0.19862033618403585
  (0, 1348)     0.16878597858179323
  (0, 1835)     0.09908403291729269
  (0, 2336)     0.14435314264061302
  (0, 1727)     0.1759230480867741
  (0, 1871)     0.14958029650555693
  (0, 2747)     0.2589218360130864
  (0, 977)      0.23178823924137498
  (0, 2329)     0.19102000085209705
  (0, 1019)     0.06613666850245224
  (0, 1718)     0.21179171197464444
  (0, 1314)     0.2747939726342262
  (0, 2612)     0.0841016051738063
  (0, 2544)     0.2046546424696636
  (0, 830)      0.22575393295574728
  (0, 612)      0.23178823924137498
  :     :
  (821, 1739)   0.09095782938176501
  (821, 1019)   0.07012291419686519
  (821, 2612)   0.08917064885424014
  (821, 1038)   0.1632598401054577
  (821, 2807)   0.09780093772987634
  (821, 503)    0.12282824000925395
  (821, 1734)   0.124656946219883
```

```
  (821, 490)    0.21741985421561066
  (821, 2268)   0.1214211589275493
  (821, 2619)   0.3091263088471426
  (821, 2387)   0.17376430521194985
  (821, 2704)   0.16701897056464968
  (821, 308)    0.2508232037213394
  (821, 564)    0.22455700268383658
  (821, 1099)   0.17376430521194985
  (821, 1922)   0.23381855721568545
  (821, 1189)   0.22060116160082402
  (821, 311)    0.1882207455785851
  (821, 723)    0.19578798662424315
  (821, 242)    0.14816514328261116
  (821, 1656)   0.2025333212715433
  (821, 1115)   0.245758777697772
  (821, 910)    0.41009908231218395
  (821, 2899)   0.20772820710632972
  (821, 1135)   0.2025333212715433
```

In [0]:
```
# find words with TF-IDF score >0.4 and sort them
```

In [0]:
```
NW=Tfidf.get_feature_names()

Wds=Tfidf.get_feature_names
for NW2,NW3 in zip(NW,LA.toarray()[0]):
 if NW3>0.04:
  print (NW2,":", NW3)
```

```
appealing : 0.20806074680652425
cedar : 0.13451531230510097
consumption : 0.23178823924137498
early : 0.22575393295574728
falls : 0.23178823924137498
finish : 0.06613666850245224
imperceptible : 0.2747939726342262
intense : 0.16878597858179323
interspersed : 0.2747939726342262
lifted : 0.22052677909080337
nearly : 0.21179171197464444
```

```
nice : 0.1759230480867741
notes : 0.17157438131096184
palate : 0.09908403291729269
peppery : 0.14958029650555693
pine : 0.2589218360130864
scents : 0.1565579136291564
short : 0.19102000085209705
shows : 0.14435314264061302
suggest : 0.2046546424696636
surprisingly : 0.19862033618403585
tannins : 0.0841016051738063
ultimately : 0.2589218360130864
way : 0.1743877334441604
woodsy : 0.2747939726342262
```

In [0]:
```
# save your table to 'key_Shiraz.txt'
```

In [0]:
```
with open('key_Shiraz.txt','w+')as f:
  for NW2,NW3 in zip(NW,LA.toarray()[0]):
    if NW3>0.04:
      f.write(words+':'+str(NW3)+'\n')
```