Final Project (Python)

December 3, 2017

```
In [1]: import pandas as pd
        import numpy as np
        import pickle
        import matplotlib.pyplot as py
        from sklearn.model_selection import train_test_split
        from sklearn import metrics
        from sklearn import tree
        import sklearn
        from sklearn import ensemble
        from sklearn import linear_model
```

1 Import Raw Data

```
In [2]: df = pd.read_json('/Users/alexanderdowney/Downloads/winemag-data_first150k.json')
```

```
In [3]: df
```

```
Out[3]:             country                                        description  \
        0                US  This tremendous 100% varietal wine hails from ...
        1             Spain  Ripe aromas of fig, blackberry and cassis are ...
        2                US  Mac Watson honors the memory of a wine once ma...
        3                US  This spent 20 months in 30% new French oak, an...
        4            France  This is the top wine from La Bégude, named aft...
        5             Spain  Deep, dense and pure from the opening bell, th...
        6             Spain  Slightly gritty black-fruit aromas include a s...
        7             Spain  Lush cedary black-fruit aromas are luxe and of...
        8                US  This re-named vineyard was formerly bottled as...
        9                US  The producer sources from two blocks of the vi...
        10            Italy  Elegance, complexity and structure come togeth...
        11               US  From 18-year-old vines, this supple well-balan...
        12               US  A standout even in this terrific lineup of 201...
        13           France  This wine is in peak condition. The tannins an...
        14               US  With its sophisticated mix of mineral, acid an...
        15               US  First made in 2006, this succulent luscious Ch...
        16               US  This blockbuster, powerhouse of a wine suggest...
        17            Spain  Nicely oaked blackberry, licorice, vanilla and...
        18           France  Coming from a seven-acre vineyard named after ...
```

```
19                 US  This fresh and lively medium-bodied wine is be...
20                 US  Heitz has made this stellar rosé from the rare...
21              Spain  Alluring, complex and powerful aromas of grill...
22              Spain  Tarry blackberry and cheesy oak aromas are app...
23                 US  The apogee of this ambitious winery's white wi...
24                 US  San Jose-based producer Adam Comartin heads 1,...
25        New Zealand  Yields were down in 2015, but intensity is up,...
26                 US  Bergström has made a Shea designate since 2003...
27                 US  Focused and dense, this intense wine captures ...
28                 US  Cranberry, baked rhubarb, anise and crushed sl...
29                 US  This standout Rocks District wine brings earth...
...               ...                                                ...
150900          Chile  Aromas of freshly cut lumber, complete with so...
150901          Chile  Lavishly oaked, the fruit here struggles to ma...
150902          Chile  This medium weight Chardonnay offered aromas o...
150903          Chile  Very light berry and mint aromas open this aus...
150904          Chile  A lot of Chilean Cabernets seem to have a dist...
150905          Chile  There's not much point in making a reserve-sty...
150906         France  This lovely wine, a Monopole, is already showi...
150907         France  Rion holds back on the new oak, letting the pu...
150908         France  Another premier cru from Michel Gros, this one...
150909         France  This is a lovely, fragrant Burgundy, with a sm...
150910         France  Scents of graham cracker and malted milk choco...
150911         France  This needs a good bit of breathing time, then ...
150912         France  The nose is dominated by the attractive scents...
150913         France  Inky and rustic, yet in a refined manner. This...
150914             US  Old-gold in color, and thick and syrupy. The a...
150915             US  Decades ago, Beringers then-winemaker Myron N...
150916             US  An impressive wine that presents a full bouque...
150917         France  Light and elegant, this spicy, lively wine is ...
150918         France  Jacquart makes a full-bodied, ripe style of Ch...
150919         France  This classy example opens with a very floral n...
150920          Italy  Rich and mature aromas of smoke, earth and her...
150921         France  Shows some older notes: a bouquet of toasted w...
150922          Italy  Made by 30-ish Roberta Borghese high above Man...
150923         France  Rich and toasty, with tiny bubbles. The bouque...
150924         France  Really fine for a low-acid vintage, there's an...
150925          Italy  Many people feel Fiano represents southern Ita...
150926         France  Offers an intriguing nose with ginger, lime an...
150927          Italy  This classic example comes from a cru vineyard...
150928         France  A perfect salmon shade, with scents of peaches...
150929          Italy  More Pinot Grigios should taste like this. A r...

                                 designation  points  price  \
0                          Martha's Vineyard      96  235.0
1       Carodorum Selección Especial Reserva      96  110.0
2              Special Selected Late Harvest      96   90.0
3                                    Reserve      96   65.0
```

```
4                                 La Brûlade      95   66.0
5                                  Numanthia      95   73.0
6                                  San Román      95   65.0
7                    Carodorum Único Crianza      95  110.0
8                                     Silice      95   65.0
9                       Gap's Crown Vineyard      95   60.0
10                        Ronco della Chiesa      95   80.0
11           Estate Vineyard Wadensvil Block      95   48.0
12                            Weber Vineyard      95   48.0
13                   Château Montus Prestige      95   90.0
14                            Grace Vineyard      95  185.0
15                                    Sigrid      95   90.0
16                           Rainin Vineyard      95  325.0
17                     6 Años Reserva Premium      95   80.0
18                             Le Pigeonnier      95  290.0
19                      Gap's Crown Vineyard      95   75.0
20                                Grignolino      95   24.0
21                   Prado Enea Gran Reserva      95   79.0
22                                Termanthia      95  220.0
23                             Giallo Solare      95   60.0
24                             R-Bar-R Ranch      95   45.0
25                           Maté's Vineyard      94   57.0
26                             Shea Vineyard      94   62.0
27                                   Abetina      94  105.0
28                           Garys' Vineyard      94   60.0
29                           The Funk Estate      94   60.0
...                                      ...     ...    ...
150900                         Prima Reserva      81   13.0
150901                               Reserva      81   12.0
150902                        Estate Bottled      81   10.0
150903                                   120      81    7.0
150904                                  None      81   10.0
150905                         Prima Reserva      80   13.0
150906                         Clos des Reas      93   65.0
150907                       Les Beaux-Monts      92   52.0
150908                           Aux Brulees      90   65.0
150909                  Clos dea Argillieres      89   52.0
150910                                  None      89   38.0
150911                          Les Chaliots      87   37.0
150912                           Les Charmes      87   65.0
150913                                  None      94   30.0
150914           Late Harvest Cluster Select      94   25.0
150915                           Nightingale      93   30.0
150916                             J. Schram      93   65.0
150917                         Brut Mosaïque      92   30.0
150918                        Cuvée Mosaïque      92   38.0
150919                       Cuvée President      91   37.0
150920                          Brut Riserva      91   19.0
```

```
150921         Blanc de Blancs Brut Mosaïque      91   38.0
150922                             Superiore      91    NaN
150923                              Demi-Sec      91   30.0
150924                          Diamant Bleu      91   70.0
150925                                  None      91   20.0
150926                        Cuvée Prestige      91   27.0
150927                         Terre di Dora      91   20.0
150928                       Grand Brut Rosé      90   52.0
150929                                  None      90   15.0

                  province                     region_1  \
0               California                  Napa Valley
1           Northern Spain                         Toro
2               California               Knights Valley
3                   Oregon            Willamette Valley
4                 Provence                       Bandol
5           Northern Spain                         Toro
6           Northern Spain                         Toro
7           Northern Spain                         Toro
8                   Oregon           Chehalem Mountains
9               California                 Sonoma Coast
10      Northeastern Italy                       Collio
11                  Oregon                 Ribbon Ridge
12                  Oregon                 Dundee Hills
13        Southwest France                      Madiran
14                  Oregon                 Dundee Hills
15                  Oregon            Willamette Valley
16              California   Diamond Mountain District
17          Northern Spain             Ribera del Duero
18        Southwest France                       Cahors
19              California                 Sonoma Coast
20              California                  Napa Valley
21          Northern Spain                        Rioja
22          Northern Spain                         Toro
23              California                  Edna Valley
24              California          Santa Cruz Mountains
25                   Kumeu                         None
26                  Oregon            Willamette Valley
27                  Oregon            Willamette Valley
28              California        Santa Lucia Highlands
29              Washington      Walla Walla Valley (WA)
...                    ...                          ...
150900        Maipo Valley                         None
150901        Maipo Valley                         None
150902        Maipo Valley                         None
150903        Rapel Valley                         None
150904        Maipo Valley                         None
150905        Maipo Valley                         None
```

```
150906            Burgundy               Vosne-Romanée
150907            Burgundy               Vosne-Romanée
150908            Burgundy               Vosne-Romanée
150909            Burgundy           Nuits-St.-Georges
150910            Burgundy           Chambolle-Musigny
150911            Burgundy           Nuits-St.-Georges
150912            Burgundy           Chambolle-Musigny
150913        Rhône Valley         Châteauneuf-du-Pape
150914          California             Anderson Valley
150915          California                 North Coast
150916          California                 Napa Valley
150917           Champagne                   Champagne
150918           Champagne                   Champagne
150919           Champagne                   Champagne
150920  Northeastern Italy                      Trento
150921           Champagne                   Champagne
150922  Northeastern Italy  Colli Orientali del Friuli
150923           Champagne                   Champagne
150924           Champagne                   Champagne
150925      Southern Italy           Fiano di Avellino
150926           Champagne                   Champagne
150927      Southern Italy           Fiano di Avellino
150928           Champagne                   Champagne
150929  Northeastern Italy                  Alto Adige

                       region_2                variety  \
0                          Napa     Cabernet Sauvignon
1                          None          Tinta de Toro
2                        Sonoma        Sauvignon Blanc
3             Willamette Valley             Pinot Noir
4                          None     Provence red blend
5                          None          Tinta de Toro
6                          None          Tinta de Toro
7                          None          Tinta de Toro
8             Willamette Valley             Pinot Noir
9                        Sonoma             Pinot Noir
10                         None               Friulano
11            Willamette Valley             Pinot Noir
12            Willamette Valley             Pinot Noir
13                         None                 Tannat
14            Willamette Valley             Pinot Noir
15            Willamette Valley             Chardonnay
16                         Napa     Cabernet Sauvignon
17                         None            Tempranillo
18                         None                 Malbec
19                       Sonoma             Pinot Noir
20                         Napa                   Rosé
21                         None      Tempranillo Blend
```

```
22                         None          Tinta de Toro
23                Central Coast             Chardonnay
24                Central Coast             Pinot Noir
25                         None             Chardonnay
26                         None             Pinot Noir
27            Willamette Valley             Pinot Noir
28                Central Coast             Pinot Noir
29              Columbia Valley                  Syrah
...                         ...                    ...
150900                     None     Cabernet Sauvignon
150901                     None                 Merlot
150902                     None             Chardonnay
150903                     None     Cabernet Sauvignon
150904                     None     Cabernet Sauvignon
150905                     None                 Merlot
150906                     None             Pinot Noir
150907                     None             Pinot Noir
150908                     None             Pinot Noir
150909                     None             Pinot Noir
150910                     None             Pinot Noir
150911                     None             Pinot Noir
150912                     None             Pinot Noir
150913                     None  Rhône-style Red Blend
150914  Mendocino/Lake Counties         White Riesling
150915              North Coast            White Blend
150916                     Napa        Champagne Blend
150917                     None        Champagne Blend
150918                     None        Champagne Blend
150919                     None        Champagne Blend
150920                     None        Champagne Blend
150921                     None        Champagne Blend
150922                     None                  Tocai
150923                     None        Champagne Blend
150924                     None        Champagne Blend
150925                     None            White Blend
150926                     None        Champagne Blend
150927                     None            White Blend
150928                     None        Champagne Blend
150929                     None           Pinot Grigio

                         winery
0                         Heitz
1       Bodega Carmen Rodríguez
2                      Macauley
3                         Ponzi
4          Domaine de la Bégude
5                     Numanthia
6                      Maurodos
```

```
7       Bodega Carmen Rodríguez
8                     Bergström
9                     Blue Farm
10             Borgo del Tiglio
11       Patricia Green Cellars
12       Patricia Green Cellars
13            Vignobles Brumont
14               Domaine Serene
15                    Bergström
16                         Hall
17                     Valduero
18           Château Lagrézette
19                 Gary Farrell
20                        Heitz
21                         Muga
22                    Numanthia
23             Center of Effort
24                     Comartin
25                  Kumeu River
26                    Bergström
27                        Ponzi
28                         Roar
29                       Saviah
...                         ...
150900               De Martino
150901                Undurraga
150902               De Martino
150903               Santa Rita
150904               De Martino
150905               De Martino
150906              Michel Gros
150907              Daniel Rion
150908              Michel Gros
150909              Daniel Rion
150910              Michel Gros
150911              Michel Gros
150912              Daniel Rion
150913          Le Vieux Donjon
150914                  Navarro
150915                 Beringer
150916              Schramsberg
150917                 Jacquart
150918                 Jacquart
150919                H.Germain
150920                  Letrari
150921                 Jacquart
150922        Ronchi di Manzano
150923                 Jacquart
```

```
150924  Heidsieck & Co Monopole
150925    Feudi di San Gregorio
150926                H.Germain
150927                Terredora
150928                   Gosset
150929            Alois Lageder

[150930 rows x 10 columns]
```

2 Clean the Data

```
In [4]: size = []
        size.append(len(df))
        df = df[np.isfinite(df['price'])]
        size.append(len(df))
        df = df[np.isfinite(df['points'])]
        size.append(len(df))

        countries = df['country'].unique()
        country_count=[]
        countries_kept = []
        for i in range(0,len(countries)):
            country_count.append(len(df.loc[df['country'] == countries[i]]))
        for j in range(0,len(countries)):
            if country_count[j]>150:
                countries_kept.append(countries[j])
        df = df[df['country'].isin(countries_kept)]
        size.append(len(df))

        amount_of_varieties = len(df['variety'].unique())
        varietal = df['variety'].unique()
        varietals = []
        amount_of_each_variety = []
        Tol = 15
        for i in range(0,amount_of_varieties):
            if len(df.loc[df['variety'] == varietal[i]]) > Tol:
                amount_of_each_variety.append(len(df.loc[df['variety'] == varietal[i]]))
                varietals.append(varietal[i])
        df = df[df['variety'].isin(varietals)]
        size.append(len(df))

        regions = df['region_1'].unique()
        region_count=[]
        regions_kept = []
        for i in range(0,len(regions)):
            region_count.append(len(df.loc[df['region_1'] == regions[i]]))
        for j in range(0,len(regions)):
```

```
    if region_count[j]>10:
        regions_kept.append(regions[j])
df = df[df['region_1'].isin(regions_kept)]
size.append(len(df))
size
```

```
Out[4]: [150930, 137235, 137235, 136334, 134711, 111009]
```

3 This is the final dataset!

```
In [5]: df
```

```
Out[5]:        country                                        description  \
        0           US  This tremendous 100% varietal wine hails from ...
        1        Spain  Ripe aromas of fig, blackberry and cassis are ...
        2           US  Mac Watson honors the memory of a wine once ma...
        3           US  This spent 20 months in 30% new French oak, an...
        4       France  This is the top wine from La Bégude, named aft...
        5        Spain  Deep, dense and pure from the opening bell, th...
        6        Spain  Slightly gritty black-fruit aromas include a s...
        7        Spain  Lush cedary black-fruit aromas are luxe and of...
        8           US  This re-named vineyard was formerly bottled as...
        9           US  The producer sources from two blocks of the vi...
        10       Italy  Elegance, complexity and structure come togeth...
        11          US  From 18-year-old vines, this supple well-balan...
        12          US  A standout even in this terrific lineup of 201...
        13      France  This wine is in peak condition. The tannins an...
        14          US  With its sophisticated mix of mineral, acid an...
        15          US  First made in 2006, this succulent luscious Ch...
        16          US  This blockbuster, powerhouse of a wine suggest...
        17       Spain  Nicely oaked blackberry, licorice, vanilla and...
        18      France  Coming from a seven-acre vineyard named after ...
        19          US  This fresh and lively medium-bodied wine is be...
        20          US  Heitz has made this stellar rosé from the rare...
        21       Spain  Alluring, complex and powerful aromas of grill...
        22       Spain  Tarry blackberry and cheesy oak aromas are app...
        23          US  The apogee of this ambitious winery's white wi...
        24          US  San Jose-based producer Adam Comartin heads 1,...
        26          US  Bergström has made a Shea designate since 2003...
        27          US  Focused and dense, this intense wine captures ...
        28          US  Cranberry, baked rhubarb, anise and crushed sl...
        29          US  This standout Rocks District wine brings earth...
        31          US  Steely and perfumed, this wine sees only 20% n...
        ...        ...                                                ...
        150879      US  A heavy wine, atypical of the appellation, whi...
        150883      US  A coppery colored, off-dry-to-frankly-sweet wi...
        150884      US  Here's a nice everyday drinking wine with some...
        150886      US  A soft, round quaffer filled with warmth. Slig...
```

```
150889      US  A bizarre style of wine. The aromas are Port-l...
150892      US  A light, earthy wine, with violet, berry and t...
150896      US  Some raspberry fruit in the aroma, but things ...
150906  France  This lovely wine, a Monopole, is already showi...
150907  France  Rion holds back on the new oak, letting the pu...
150908  France  Another premier cru from Michel Gros, this one...
150909  France  This is a lovely, fragrant Burgundy, with a sm...
150910  France  Scents of graham cracker and malted milk choco...
150911  France  This needs a good bit of breathing time, then ...
150912  France  The nose is dominated by the attractive scents...
150913  France  Inky and rustic, yet in a refined manner. This...
150914      US  Old-gold in color, and thick and syrupy. The a...
150915      US  Decades ago, Beringers then-winemaker Myron N...
150916      US  An impressive wine that presents a full bouque...
150917  France  Light and elegant, this spicy, lively wine is ...
150918  France  Jacquart makes a full-bodied, ripe style of Ch...
150919  France  This classy example opens with a very floral n...
150920   Italy  Rich and mature aromas of smoke, earth and her...
150921  France  Shows some older notes: a bouquet of toasted w...
150923  France  Rich and toasty, with tiny bubbles. The bouque...
150924  France  Really fine for a low-acid vintage, there's an...
150925   Italy  Many people feel Fiano represents southern Ita...
150926  France  Offers an intriguing nose with ginger, lime an...
150927   Italy  This classic example comes from a cru vineyard...
150928  France  A perfect salmon shade, with scents of peaches...
150929   Italy  More Pinot Grigios should taste like this. A r...

                                 designation  points  price  \
0                          Martha's Vineyard      96  235.0
1       Carodorum Selección Especial Reserva      96  110.0
2              Special Selected Late Harvest      96   90.0
3                                    Reserve      96   65.0
4                                 La Brûlade      95   66.0
5                                  Numanthia      95   73.0
6                                  San Román      95   65.0
7                    Carodorum Único Crianza      95  110.0
8                                     Silice      95   65.0
9                       Gap's Crown Vineyard      95   60.0
10                        Ronco della Chiesa      95   80.0
11           Estate Vineyard Wadensvil Block      95   48.0
12                            Weber Vineyard      95   48.0
13                   Château Montus Prestige      95   90.0
14                            Grace Vineyard      95  185.0
15                                    Sigrid      95   90.0
16                           Rainin Vineyard      95  325.0
17                    6 Años Reserva Premium      95   80.0
18                             Le Pigeonnier      95  290.0
19                      Gap's Crown Vineyard      95   75.0
```

```
20                               Grignolino      95   24.0
21                  Prado Enea Gran Reserva      95   79.0
22                               Termanthia      95  220.0
23                            Giallo Solare      95   60.0
24                            R-Bar-R Ranch      95   45.0
26                            Shea Vineyard      94   62.0
27                                  Abetina      94  105.0
28                          Garys' Vineyard      94   60.0
29                          The Funk Estate      94   60.0
31                                 Babushka      90   37.0
...                                     ...     ...    ...
150879                                 None      83   16.0
150883                        Reserve White      83    7.0
150884                                 None      83   10.0
150886                                 None      82   10.0
150889                      Lafond Vineyard      82   35.0
150892                              Coastal      82   10.0
150896                                 None      82   10.0
150906                        Clos des Reas      93   65.0
150907                      Les Beaux-Monts      92   52.0
150908                          Aux Brulees      90   65.0
150909                 Clos dea Argillieres      89   52.0
150910                                 None      89   38.0
150911                         Les Chaliots      87   37.0
150912                          Les Charmes      87   65.0
150913                                 None      94   30.0
150914          Late Harvest Cluster Select      94   25.0
150915                          Nightingale      93   30.0
150916                            J. Schram      93   65.0
150917                        Brut Mosaïque      92   30.0
150918                       Cuvée Mosaïque      92   38.0
150919                      Cuvée President      91   37.0
150920                         Brut Riserva      91   19.0
150921        Blanc de Blancs Brut Mosaïque      91   38.0
150923                             Demi-Sec      91   30.0
150924                         Diamant Bleu      91   70.0
150925                                 None      91   20.0
150926                       Cuvée Prestige      91   27.0
150927                        Terre di Dora      91   20.0
150928                      Grand Brut Rosé      90   52.0
150929                                 None      90   15.0

                  province                  region_1  \
0               California               Napa Valley
1           Northern Spain                      Toro
2               California            Knights Valley
3                   Oregon         Willamette Valley
4                 Provence                    Bandol
```

```
5           Northern Spain                       Toro
6           Northern Spain                       Toro
7           Northern Spain                       Toro
8                   Oregon         Chehalem Mountains
9               California               Sonoma Coast
10      Northeastern Italy                     Collio
11                  Oregon               Ribbon Ridge
12                  Oregon               Dundee Hills
13        Southwest France                    Madiran
14                  Oregon               Dundee Hills
15                  Oregon          Willamette Valley
16              California  Diamond Mountain District
17          Northern Spain           Ribera del Duero
18        Southwest France                     Cahors
19              California               Sonoma Coast
20              California                Napa Valley
21          Northern Spain                      Rioja
22          Northern Spain                       Toro
23              California                Edna Valley
24              California       Santa Cruz Mountains
26                  Oregon          Willamette Valley
27                  Oregon          Willamette Valley
28              California      Santa Lucia Highlands
29              Washington    Walla Walla Valley (WA)
31              California       Russian River Valley
...                    ...                        ...
150879          California            Anderson Valley
150883          California                 California
150884          California                 California
150886          California                 California
150889          California          Santa Ynez Valley
150892          California                 California
150896          California                 California
150906            Burgundy              Vosne-Romanée
150907            Burgundy              Vosne-Romanée
150908            Burgundy              Vosne-Romanée
150909            Burgundy          Nuits-St.-Georges
150910            Burgundy          Chambolle-Musigny
150911            Burgundy          Nuits-St.-Georges
150912            Burgundy          Chambolle-Musigny
150913        Rhône Valley        Châteauneuf-du-Pape
150914          California            Anderson Valley
150915          California                North Coast
150916          California                Napa Valley
150917           Champagne                  Champagne
150918           Champagne                  Champagne
150919           Champagne                  Champagne
150920  Northeastern Italy                     Trento
```

```
150921           Champagne                  Champagne
150923           Champagne                  Champagne
150924           Champagne                  Champagne
150925      Southern Italy          Fiano di Avellino
150926           Champagne                  Champagne
150927      Southern Italy          Fiano di Avellino
150928           Champagne                  Champagne
150929  Northeastern Italy                 Alto Adige

                       region_2                variety  \
0                          Napa     Cabernet Sauvignon
1                          None          Tinta de Toro
2                        Sonoma        Sauvignon Blanc
3             Willamette Valley             Pinot Noir
4                          None     Provence red blend
5                          None          Tinta de Toro
6                          None          Tinta de Toro
7                          None          Tinta de Toro
8             Willamette Valley             Pinot Noir
9                        Sonoma             Pinot Noir
10                         None               Friulano
11            Willamette Valley             Pinot Noir
12            Willamette Valley             Pinot Noir
13                         None                 Tannat
14            Willamette Valley             Pinot Noir
15            Willamette Valley             Chardonnay
16                         Napa     Cabernet Sauvignon
17                         None            Tempranillo
18                         None                 Malbec
19                       Sonoma             Pinot Noir
20                         Napa                   Rosé
21                         None      Tempranillo Blend
22                         None          Tinta de Toro
23                Central Coast             Chardonnay
24                Central Coast             Pinot Noir
26                         None             Pinot Noir
27            Willamette Valley             Pinot Noir
28                Central Coast             Pinot Noir
29              Columbia Valley                  Syrah
31                       Sonoma             Chardonnay
...                         ...                    ...
150879  Mendocino/Lake Counties             Pinot Noir
150883         California Other              Zinfandel
150884         California Other             Chardonnay
150886         California Other                 Merlot
150889            Central Coast             Pinot Noir
150892         California Other                 Merlot
150896         California Other             Pinot Noir
```

```
150906                     None             Pinot Noir
150907                     None             Pinot Noir
150908                     None             Pinot Noir
150909                     None             Pinot Noir
150910                     None             Pinot Noir
150911                     None             Pinot Noir
150912                     None             Pinot Noir
150913                     None  Rhône-style Red Blend
150914  Mendocino/Lake Counties         White Riesling
150915              North Coast            White Blend
150916                     Napa        Champagne Blend
150917                     None        Champagne Blend
150918                     None        Champagne Blend
150919                     None        Champagne Blend
150920                     None        Champagne Blend
150921                     None        Champagne Blend
150923                     None        Champagne Blend
150924                     None        Champagne Blend
150925                     None            White Blend
150926                     None        Champagne Blend
150927                     None            White Blend
150928                     None        Champagne Blend
150929                     None           Pinot Grigio

                         winery
0                         Heitz
1       Bodega Carmen Rodríguez
2                      Macauley
3                         Ponzi
4          Domaine de la Bégude
5                     Numanthia
6                      Maurodos
7       Bodega Carmen Rodríguez
8                     Bergström
9                     Blue Farm
10             Borgo del Tiglio
11       Patricia Green Cellars
12       Patricia Green Cellars
13            Vignobles Brumont
14               Domaine Serene
15                    Bergström
16                         Hall
17                     Valduero
18           Château Lagrézette
19                 Gary Farrell
20                        Heitz
21                         Muga
22                    Numanthia
```

```
23               Center of Effort
24                       Comartin
26                      Bergström
27                          Ponzi
28                           Roar
29                         Saviah
31                       Zepaltas
...                           ...
150879                   Edmeades
150883                 Glen Ellen
150884                 Hawk Crest
150886                    Camelot
150889                     Lafond
150892                   Callaway
150896                    Camelot
150906                Michel Gros
150907                Daniel Rion
150908                Michel Gros
150909                Daniel Rion
150910                Michel Gros
150911                Michel Gros
150912                Daniel Rion
150913           Le Vieux Donjon
150914                    Navarro
150915                   Beringer
150916                Schramsberg
150917                   Jacquart
150918                   Jacquart
150919                  H.Germain
150920                    Letrari
150921                   Jacquart
150923                   Jacquart
150924  Heidsieck & Co Monopole
150925    Feudi di San Gregorio
150926                  H.Germain
150927                  Terredora
150928                     Gosset
150929              Alois Lageder

[111009 rows x 10 columns]
```

4 Useful Functions

```
In [6]: def R2_score(y_pred,y_true):
            # u is the residual sum of squares
            u = ((y_true - y_pred) ** 2).sum()
            # v is the total sum of squares
            v = ((y_true - y_true.mean()) ** 2).sum()
            return (1-u/v)
```

```
In [7]: def report_metrics(y_pred, y_true):
            m1 = metrics.mean_absolute_error(y_true, y_pred)
            m2 = metrics.median_absolute_error(y_true,y_pred)
            m3 = metrics.explained_variance_score(y_true,y_pred)
            m4 = metrics.r2_score(y_true,y_pred)
            #print("Mean Absolute Error:",m1,"| Median Absolute Error:", m2,"| Explain Varianc
            return m1,m2,m3,m4
```

```
In [8]: def Transform_df_to_X(df):
            # To get variety data via one hot encoding
            varieties_kpt = df['variety'].unique()
            dummy_variety = pd.get_dummies(df['variety'])
            variety = []
            for i in range(0,len(df['variety'].unique())):
                variety.append(dummy_variety[varieties_kpt[i]])

            # To get country data via one hot encoding
            countries_kpt = df['country'].unique()
            dummy = pd.get_dummies(df['country'])
            country = []
            for i in range(0,len(df['country'].unique())):
                country.append(dummy[countries_kpt[i]])

            # To get variety data via one hot encoding
            regions_kpt = df['region_1'].unique()
            dummy_variety = pd.get_dummies(df['region_1'])
            region = []
            for i in range(0,len(df['region_1'].unique())):
                region.append(dummy_variety[regions_kpt[i]])

            X = df[['points']].as_matrix()

            for i in range(0,len(country)):
                X = np.c_[X,country[i]]
            for j in range(0,len(variety)):
                X = np.c_[X,variety[j]]
            for j in range(0,len(region)):
                X = np.c_[X,region[j]]
        #    X = np.c_[X,sent_sums]
        #    X = np.c_[X,sent_prob]
        #    X = np.c_[X,sent_neg]
            X = np.c_[X,np.ones(len(df['points']))]
            return X
```

5 Linear Models

```
In [9]: def Wine_LinLstSq_Regression(Xtrain,Xtest,Ytrain):
            w,resdiuals,rank,singular_vals = np.linalg.lstsq(Xtrain, Ytrain)
            w_matrix = np.transpose(np.asmatrix(w))
            w_array = np.squeeze(np.asarray(Xtest*w_matrix))
            return w_array # Returns the prediction vector
```

```
In [10]: def Wine_Huber_Linear_Regression(Xtrain, Xtest, Ytrain):
             hlr = sklearn.linear_model.HuberRegressor()
             hlr = hlr.fit(Xtrain, Ytrain)
             return hlr.predict(Xtest)
```

6 Trees

```
In [11]: def Wine_Decision_Tree_Regression(Xtrain,Xtest,Ytrain):
             clf = tree.DecisionTreeRegressor()
             # useful code: min_samples_leaf=10,max_depth=3,max_leaf_nodes = 100
             clf = clf.fit(Xtrain, Ytrain)
             return clf.predict(Xtest) # Returns the prediction vector
```

```
In [12]: def Wine_Random_Forest_Regression(Xtrain, Xtest, Ytrain):
             rfr = sklearn.ensemble.RandomForestRegressor()
             rfr = rfr.fit(Xtrain, Ytrain)
             return rfr.predict(Xtest)
```

```
In [13]: def Wine_Huber_Tree_Regression(Xtrain, Xtest, Ytrain):
             htr = sklearn.ensemble.GradientBoostingRegressor(loss='huber')
             htr = htr.fit(Xtrain, Ytrain)
             return htr.predict(Xtest)
```

```
In [14]: def Wine_Ls_Tree_Regression(Xtrain, Xtest, Ytrain):
             htr = sklearn.ensemble.GradientBoostingRegressor(loss='ls')
             htr = htr.fit(Xtrain, Ytrain)
             return htr.predict(Xtest)
```

7 SVM

```
In [15]: def Wine_SVM_Regression(Xtrain, Xtest, Ytrain):
             svr = sklearn.svm.SVR()
             svr = svr.fit(Xtrain, Ytrain)
             return svr.predict(Xtest)
```

8 Train/Test Set Split

```
In [16]: sent_sums = pd.read_json('/Users/alexanderdowney/Downloads/sentiment_sums.json')
         sent_prob = pd.read_json('/Users/alexanderdowney/Downloads/sentiment_probabilities.js
         sent_neg = pd.read_json('/Users/alexanderdowney/Downloads/sentiment_probabilities_neg
In [17]: X = Transform_df_to_X(df)
         data = X
         target = df['price'].as_matrix()
```

```
In [35]: # total examples after data cleaning: 129964
         X_train, X_test, y_train, y_test = train_test_split(data, target, test_size=0.1, rando
```

9 Example on Split data

```
In [36]: report_metrics((Wine_Decision_Tree_Regression(X_train,X_test,y_train)),y_test)
```

```
Out[36]: 0.62524266974376341
```

10 Cross Validation

```
In [37]: def test_cv(model):
             scores_m1 = []
             scores_m2 = []
             scores_m3 = []
             scores_m4 = []
             for k in range(0, 10):
                 X_tr, X_te, y_tr, y_te = train_test_split(X_train, y_train, test_size=0.1, ra
                 train_data_input = X_tr
                 train_data_output = y_tr
                 test_data_input = X_te
                 y_pred = model(train_data_input, test_data_input, train_data_output)
                 m1,m2,m3,m4 = report_metrics(y_pred, y_te)
                 scores_m1.append(m1)
                 scores_m2.append(m2)
                 scores_m3.append(m3)
                 scores_m4.append(m4)
                 print("Mean Absolute Error:",m1,"| Median Absolute Error:", m2,"| Explain Var
             print("Average Mean Absolute Error:",np.mean(scores_m1),"| Average Median Absolut
             return np.mean(scores_m1),np.mean(scores_m2),np.mean(scores_m3),np.mean(scores_m4
```

Linear Models

```
In [38]: test_cv(Wine_LinLstSq_Regression)
```

```
Mean Absolute Error: 13.0743198175 | Median Absolute Error: 8.82995605469 | Explain Variance Sc
Mean Absolute Error: 13.3658231007 | Median Absolute Error: 8.8486328125 | Explain Variance Sco
Mean Absolute Error: 13.1649700917 | Median Absolute Error: 8.65719985962 | Explain Variance Sc
Mean Absolute Error: 13.3252047716 | Median Absolute Error: 8.6852312088 | Explain Variance Sco
Mean Absolute Error: 13.3042964399 | Median Absolute Error: 8.81427001953 | Explain Variance Sc
Mean Absolute Error: 13.5289581424 | Median Absolute Error: 8.82426452637 | Explain Variance Sc
Mean Absolute Error: 13.6397256377 | Median Absolute Error: 8.56969833374 | Explain Variance Sc
Mean Absolute Error: 13.1940175113 | Median Absolute Error: 8.728515625 | Explain Variance Scor
Mean Absolute Error: 13.2212549705 | Median Absolute Error: 8.88458251953 | Explain Variance Sc
Mean Absolute Error: 13.5662815513 | Median Absolute Error: 8.88725280762 | Explain Variance Sc
Average Mean Absolute Error: 13.3384852035 | Average Median Absolute Error: 8.77296037674 | Ave
```

```
Out[38]: (13.3384852034613,
          8.7729603767395012,
          0.40019905433246361,
          0.40017730192397583)
```

```
In [39]: test_cv(Wine_Huber_Linear_Regression)
```

```
Mean Absolute Error: 12.8734929394 | Median Absolute Error: 6.99187677706 | Explain Variance Sc
Mean Absolute Error: 13.1681318773 | Median Absolute Error: 7.10912258031 | Explain Variance Sc
Mean Absolute Error: 13.4263534967 | Median Absolute Error: 7.39112147587 | Explain Variance Sc
Mean Absolute Error: 13.0587740486 | Median Absolute Error: 7.05634239574 | Explain Variance Sc
Mean Absolute Error: 13.0092553023 | Median Absolute Error: 6.96977869445 | Explain Variance Sc
Mean Absolute Error: 13.271993787 | Median Absolute Error: 6.86031199339 | Explain Variance Sco
Mean Absolute Error: 14.1359213105 | Median Absolute Error: 7.46938237167 | Explain Variance Sc
Mean Absolute Error: 12.8385265496 | Median Absolute Error: 7.10887097286 | Explain Variance Sc
Mean Absolute Error: 13.0541088492 | Median Absolute Error: 6.9629765475 | Explain Variance Sco
Mean Absolute Error: 13.6625267993 | Median Absolute Error: 7.53113309535 | Explain Variance Sc
Average Mean Absolute Error: 13.249908496 | Average Median Absolute Error: 7.14509169042 | Aver
```

```
Out[39]: (13.249908495993186,
          7.1450916904211139,
          0.21429171688398826,
          0.19629384057766472)
```

Trees

```
In [40]: test_cv(Wine_Decision_Tree_Regression)
```

```
Mean Absolute Error: 9.92979005108 | Median Absolute Error: 5.0 | Explain Variance Score: 0.57
Mean Absolute Error: 9.98005371212 | Median Absolute Error: 5.0 | Explain Variance Score: 0.42
Mean Absolute Error: 10.0050328468 | Median Absolute Error: 5.0 | Explain Variance Score: 0.58
Mean Absolute Error: 10.0502938725 | Median Absolute Error: 5.0 | Explain Variance Score: 0.67
Mean Absolute Error: 9.91460713829 | Median Absolute Error: 5.0 | Explain Variance Score: 0.64
Mean Absolute Error: 10.5718824545 | Median Absolute Error: 5.0 | Explain Variance Score: 0.43
Mean Absolute Error: 10.2826824073 | Median Absolute Error: 5.0 | Explain Variance Score: 0.55
Mean Absolute Error: 9.84443259208 | Median Absolute Error: 5.0 | Explain Variance Score: 0.61
Mean Absolute Error: 10.0098580846 | Median Absolute Error: 5.0 | Explain Variance Score: 0.58
Mean Absolute Error: 10.3186840952 | Median Absolute Error: 5.0 | Explain Variance Score: 0.44
Average Mean Absolute Error: 10.0907317254 | Average Median Absolute Error: 5.0 | Average Expla
```

```
Out[40]: (10.090731725446851, 5.0, 0.55426496138775161, 0.55422389608872114)
```

```
In [41]: test_cv(Wine_Random_Forest_Regression)
```