OPTIMIZATION FOR DATA SCIENCE

WINE PRICE PREDICTION

Annanya Kannan
Prashanth Asok Kumar
Vasanth Mohan

GOAL:

Our goal is to develop a system that will help wine makers and new people interested in wine business to choose their pricing wisely. We have used a huge dataset with wine reviews from wine tasters, to predict the price of varieties of wine, using the rating of the wine.

THE CONTEXT:

Our prediction is to help new wine makers or new wine brands to price their product effectively.
Our system will generate a function that predicts the prices of the blend/variety based on the rating from wine taster and the country of origin of the wine.

Columns: Measure Names

Rows: variety

Variety with maximum rating and maximum price

variety	Max...	Max...
Pinot Noir	99	2,500
Cabernet Sauvig..	100	625
Chardonnay	100	2,013
Red Blend	99	500
Bordeaux-style ..	100	3,300
Syrah	100	750
Riesling	98	775
Nebbiolo	99	595
Sangiovese	100	800
Sauvignon Blanc	96	135
Merlot	100	625
Champagne Blend	100	600
Zinfandel	97	100
Malbec	97	400
RosÃ©	96	800
Sparkling Blend	98	250
Tempranillo	97	600
Portuguese Red	100	450
White Blend	97	375
RhÃ´ne-style Re..	98	500
Cabernet Franc	97	180
Sangiovese Gros..	100	900
Shiraz	99	850

Cleaned data, the variety of the wine is sorted with maximum price.

QUESTION ASKED:

- What is the correlation between price and rating?
- How much does the origin (country) impact the price regardless of the rating?

HYPOTHESIS:

- Higher the rating (points) of the wine, more the expense.
- The source of the wine may inflate or deflate the price/value.

ASSUMPTIONS:

- From the visualization, the price of wine in all the countries peaks between the points/ratings 89 and 94.
- We can observe that there is a fall in pricing after the same.

PROBLEMS DEALT:

- To identify and use proper visualization tools to recognize patterns in data.
- Data cleaning:
  - ✓ To remove all data which was unnecessary for prediction.
  - ✓ To remove entries with null characters.
- Adjusting for biasing based on country.
- Some data points are not weighted (number of ratings is insufficient) properly when considering rating and source of origin.

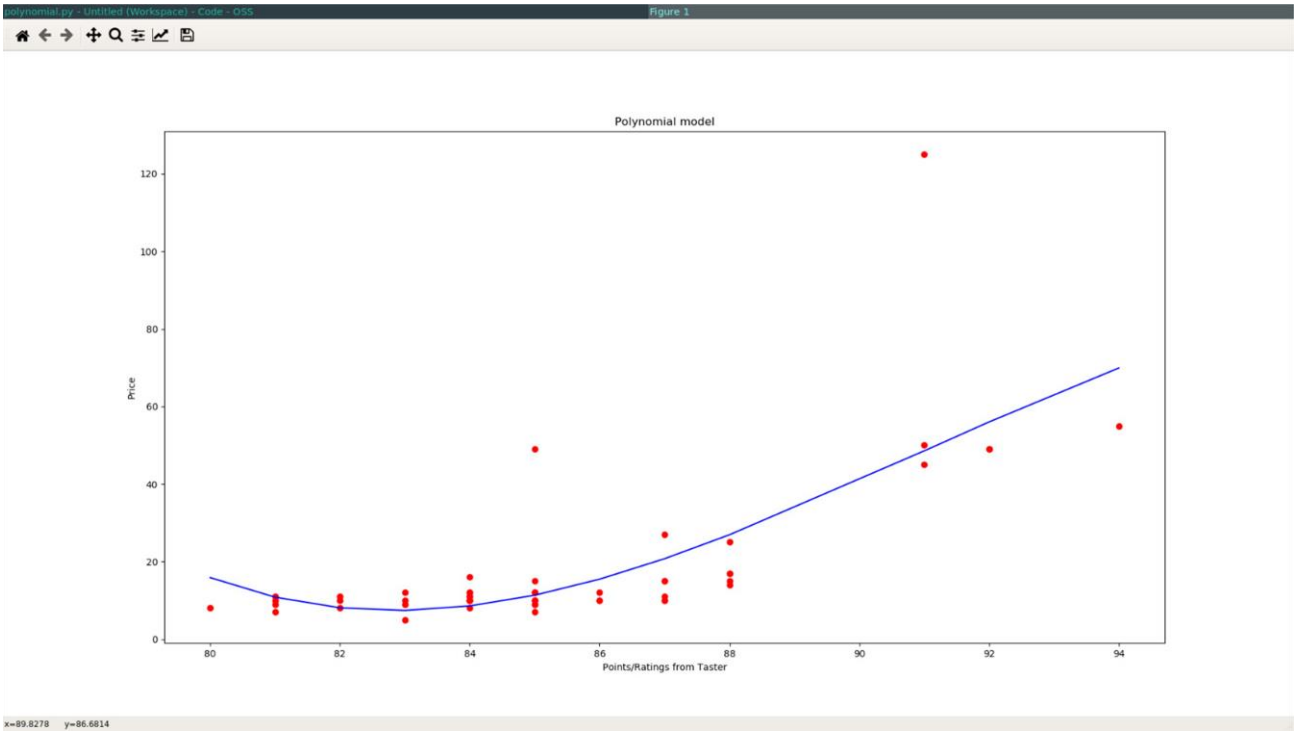
DATA SET USED:

https://www.kaggle.com/zynicide/wine-reviews#winemag-data-130k-v2.csv

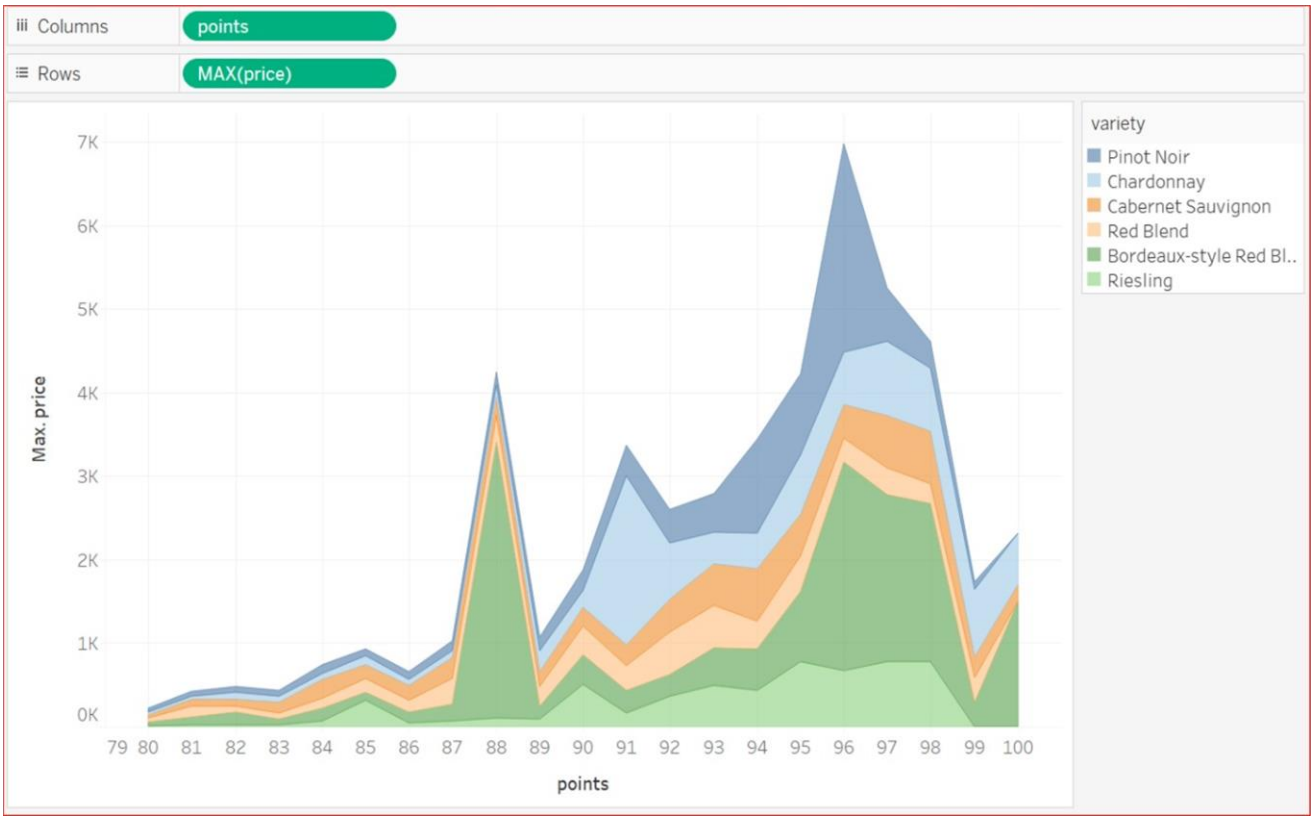
ALGORITHM USED:

- Polynomial regression:

✓ Based on the visualization of multiple varieties we can see that the pricing of the wine, does not grow linearly.
✓ Since the graph is not linear, we chose polynomial regression to train the system for quadratic or higher degrees of growth.



VISUAL REPRESENTATION OF DATA POINTS:



CODE OUTPUT:

```
currDf = df.loc[ (df['variety'] == wineToBePriced) & (df['country'] == origin) ].sort_values(['points','price'])

print("Description of available data\n",currDf.describe())
choice = input("Continue? Y/n : ")
if choice != "Y":
    sys.exit()

# Training the dataset
points_x = currDf[["points"]]
prices_y = currDf["price"]

X_train, X_test, y_train, y_test = train_test_split(points_x, prices_y, test_size=0.25, random_state=0)

poly_reg = PolynomialFeatures(degree=3)
X_poly = poly_reg.fit_transform(points_x)
pol_reg = LinearRegression()
pol_reg.fit(X_poly, prices_y)

# Prediction
predicted_y = pol_reg.predict(poly_reg.fit_transform([[ratingGiven]]))
print('Recommended price: ', predicted_y)

# Viz
plt.scatter(points_x, prices_y, color='red')
```

```
[vazark@vazark-i3 dOpt_asgmt]$ python polynomial.py
Description of available data
          points       price
count  53.000000   53.000000
mean   85.433962   18.037736
std     3.072893   19.645867
min    80.000000    5.000000
25%    84.000000   10.000000
50%    85.000000   11.000000
75%    87.000000   15.000000
max    94.000000  125.000000
Continue? Y/n : Y
Recommended price:  [8.5608955]
[vazark@vazark-i3 dOpt_asgmt]$
```