Devang Kankaria
M.Sc. in CS (DS)
23335884

MAIN ASSIGNMENT

1) Executive summary/objective statement followed by central figure

The purpose of this study is to find out what affects the ratings of French wines. 2,500 Wine Enthusiast reviews were taken into consideration for this analysis. According to my classification model, price and variety are the most important determinants of wine quality. My key finding shows this connection – expensive wines are more likely to be rated better while certain types of them are also preferred by reviewers over others. Such understanding could help sellers make their choices right in terms of selection as well as promote it so that they can improve customer contentment which will eventually lead them into making more profits.

2) Introduction

In the wine world, there is nothing more important for buyers and sellers to know than what makes French wines good. To investigate and predict ratings of wine, this report uses a dataset containing 2,500 reviews of French wines from the Wine Enthusiast website. The investigation concentrates on such things as price, description terms used in reviews or grape variety etc., that can affect the rating given to a bottle of wine according to different statistical models designed in this study. The aim here is therefore not only providing useful information but also giving tips which are based on facts discovered through studying data so that merchants may find out high-quality products at lower costs when they analyze their own records.

3) Data description

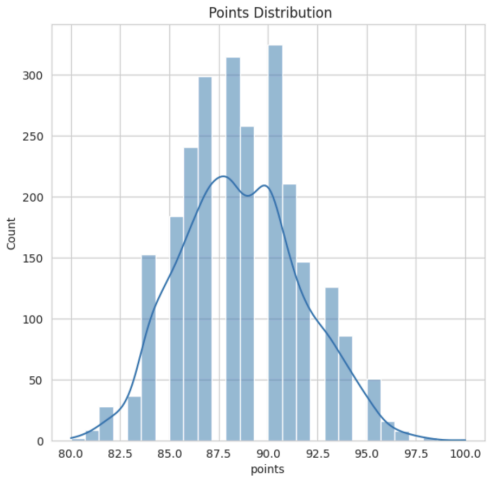
Points: Numeric score given to the wine by reviewers usually on a scale of 100
Superior Rating: Binary indicator (1=superior, 0 otherwise) based on whether points>=90
Price: Cost of wine in dollars
Variety: The type of wine; for example Champagne Blend, Bordeaux-style Red Blend etc.
Indicator Variables: Binary variables indicating presence of specific terms in the description: Crisp, Dry, Finish, Firm, Fresh, Fruit, Full, Rich, Round, Soft, and Sweet.
Description: Wine Description text.

4) Analysis

a) Preliminary Assessment

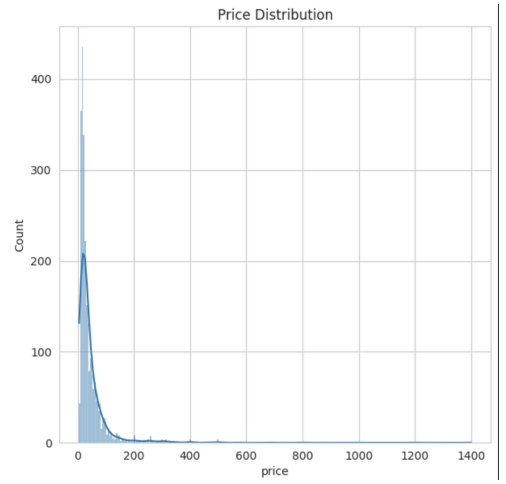
- From a thorough run through of the dataset under consideration, there were no missing values noticed. This facilitates Data preprocessing methods.
- "Description" feature holds description of textual nature that can be employed for more advanced text analysis or feature extraction techniques.
- The "Variety" feature present in the dataset under study holds disparate values. This is, in turn, reasonably important for understanding the wine ratings.

b) Understanding Distributions



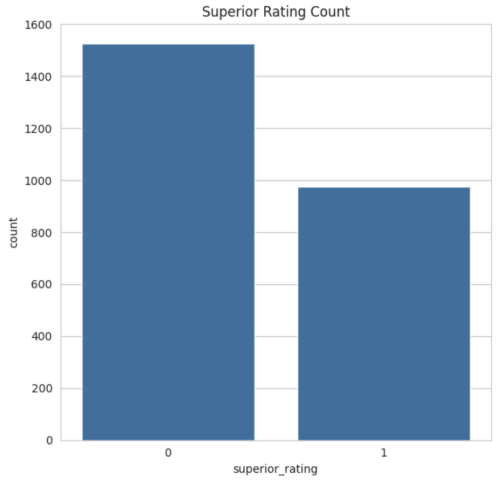
The points reasonably follow a nature of normal distribution with a slight skew towards higher ratings. This follows that most wines have been rated quite highly.

```
sns.histplot(wine_data['points'], kde=True, ax=axes[0, 0])
axes[0, 0].set_title(' Points Distribution')
```



Prices are disseminated in a fashion that manifests right-skewness; this illustrates that while majority of wines fall within lower price ranges, there exist some very expensive outliers.

```
sns.histplot(wine_data['price'], kde=True, ax=axes[0, 1])
axes[0, 1].set_title('Price Distribution')
```



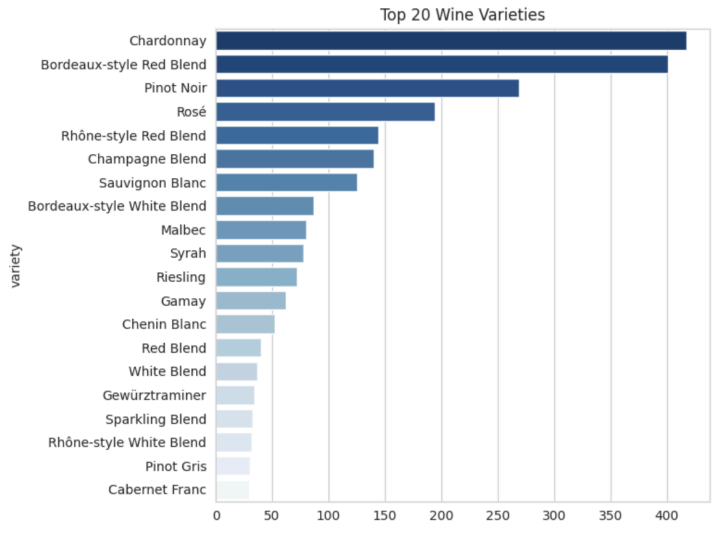
Wines rated as superior (1) and not superior (0) are distinctly distinguishable, therefore it can be used to train models for binary classification tasks

```
sns.countplot(x='superior_rating', data=wine_data, ax=axes[1, 0])
axes[1, 0].set_title('Superior Rating Count')
```

Subplots are created using "plt.subplots()" so that different aspects of the data can be compared side by side to show distribution of
1) points,
2) price,
3) number of superiors ratings, and
4) most common wine varieties

Distribution plots are made with Seaborn's "histplot" for continuous data while categorical data uses "countplot" and "barplot" with blue color scheme being used for the latter.

c) Studying wines' distribution in the dataset undertaken



Some types are present in greater frequency than others; this may indicate popularity or prevalence among records hence influential to wine rating predictions.

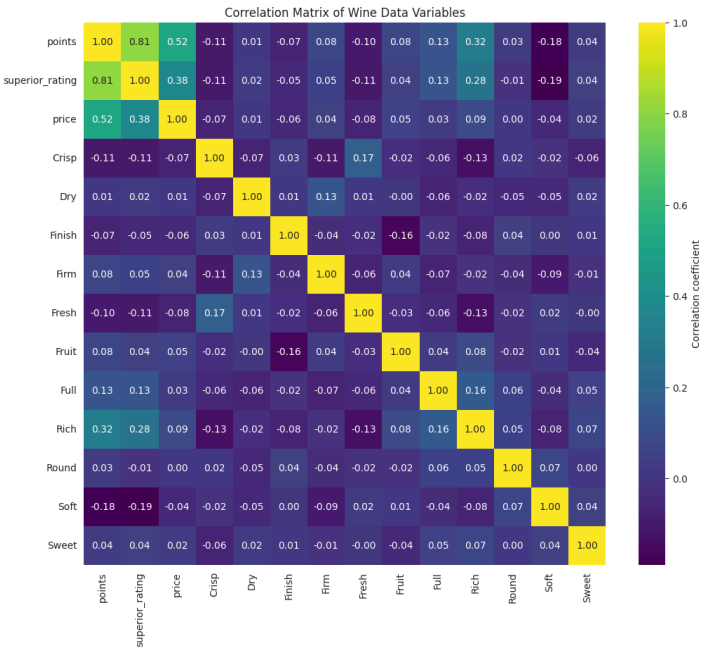
Associated Code:

```
top_varieties = wine_data['variety'].value_counts().head(20)
sns.barplot(x=top_varieties.values, y=top_varieties.index, ax=axes[1, 1],palette='Blues_r')
axes[1, 1].set_title('Top 20 Wine Varieties')
```

Packages such as
1) seaborn
2) matplotlib.pyplot
were used to style and create visualizations for wine data.

The script sets the aesthetic of plots with Seaborn's "set_style" method which gives a "whitegrid" background for better visibility.

d) Affiliation between their features' correlation



- Rich and Superior Rating:
(0.28: A weak positive correlation)
between the descriptor "Rich" and superior rating. Thus, richer wines are mildly connected to the notion of receiving higher ratings and these 2 descriptors are loosely connected.

- Full and Rich Descriptors:
(0.16: A very weak positive correlation)
implies that the descriptors have an utterly weak and insignificant relationship together when describing wines; i.e., if one is used then its unlikely so too will be the other most times.

- Points and Superior Rating:
(0.81: A strong positive correlation)

This means that higher-rated wines are more likely to be judged as 'superior'. Points awarded for a wine have a robust positive relationship with the judgement 'superior'.

- Points and Price:
(0.52: A moderate positive correlation)

On average, wines that score highly generally tend to cost more. So here, higher points are generally associated with increased prices.

- Superior Rating and Price:
(0.38: A weak to moderate positive correlation)

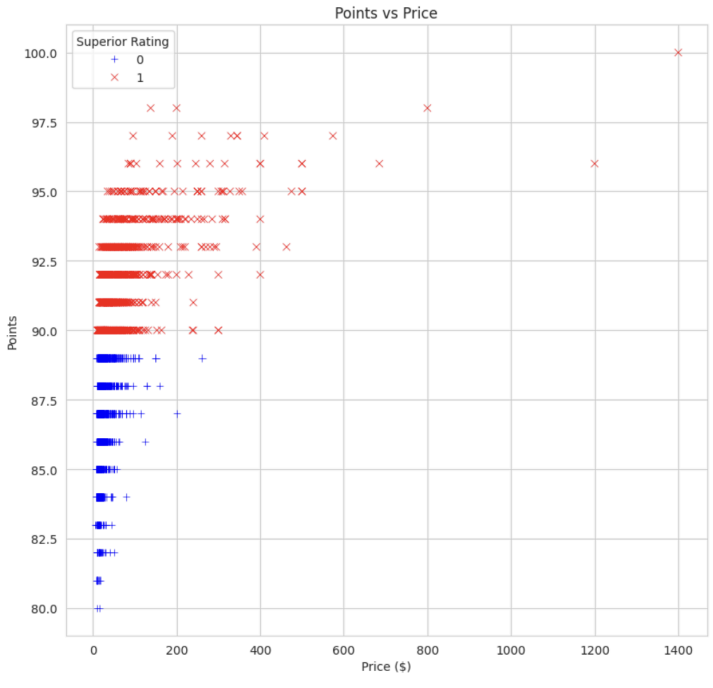
between superior rating and price are loosely tied with one another, where changes in one variable could be responsible for the changes in the other, nevertheless feebly correlated.

- Points and Rich Descriptor:
(0.32: A weak positive correlation)

"points" and the descriptor "rich" have flimsy rapport with each other, where alterations in one could reflect in the amendment of the other, however not strong correlated.

e) Statistical Analysis

1) Scatter Plot:



Axes:
a) x-axis: price($) of wine
b) y-axis: points allotted for each wine (arraying 80 to 100)

Symbols:
a) +: Wines considered - not superior (0), mainly concentrated below 90 points.
b) X: Wines considered - superior (1), primarily concentrated around or above 90 points.

Trends:
Higher priced wines generally have higher ratings, especially those that are considered superior.

Price Tiers:
Most wines below $200 earn points between 80 and just under 90, which means that they are not guaranteed a very high score by their price alone. On the other hand, among wines in more expensive brackets — notably those priced above $200 — there tends to be a higher frequency of superior rankings.

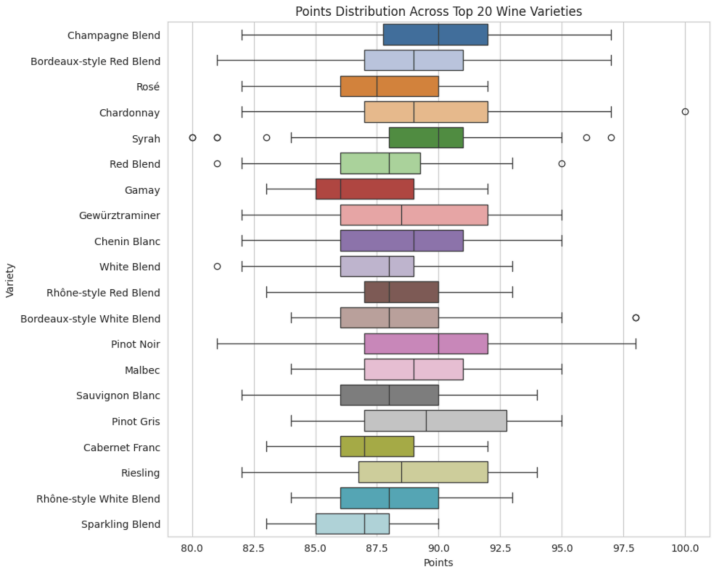
Scatterplot Distribution:
It is noticeable in the graph that around and above 90 points earned (including), there is a steep increase in the number of red crosses (representing superior rated wines). This implies that most of the best

Outliers:
At every price point shown on this chart no wine falls below 80 points but many exceed it; however for all costs shown here no bottle rates lower than approximately eighty marks.

High Priced Consistency:
Every single wine over $300 receives at least an outstanding rating. What it means is that if it is marked "superior" then its cost must be more than three hundred dollars although nothing stops us from saying vice versa i.e., any bottle costing over three hundred bucks should get ranked as such or even better if possible.

quality wines fall within this range of scores as was expected from our observation about prices vis-à-vis perceived quality i.e., higher priced bottles should correspond to better ones according to points awarded.

2) Box Plot:



Champagne Blend:
Champagne Blend ranges from 81.0 to 97.0 points with a median score of 90.0, which implies that they are generally good wines. For wine sellers, it means that they have the potential to charge ~$148.0 for a bottle of Champagne Blend as long as they perform well and receive positive feedback in addition to being of high-quality. (90.0)

Pinot Noir:
Pinot Noir is an expensive wine(~$275.0), but it consistently gets higher ratings; however there is still much variation in rating points awarded – this means while one might expect excellent quality wines from this grape variety. While often priced towards top end among all reds, its scores range widely indeed – from 81.0 up until 98.0 points.

Axes:
a) x-axis: points received associated with the wine varieties (arraying 80-100) ranked from top to bottom
b) y-axis: top 20 wine varieties

First Observation:
On average, people think Champagne Blends and Pinot Noir are better quality than Gewürztraminer because they have higher median points.

Outliers:
There are a few different wines that stick out as uncommon, and they are Chardonnay, Syrah, Red Blend, White Blend and Bordeaux-style White Blend. These wines have outliers denoted by circles.

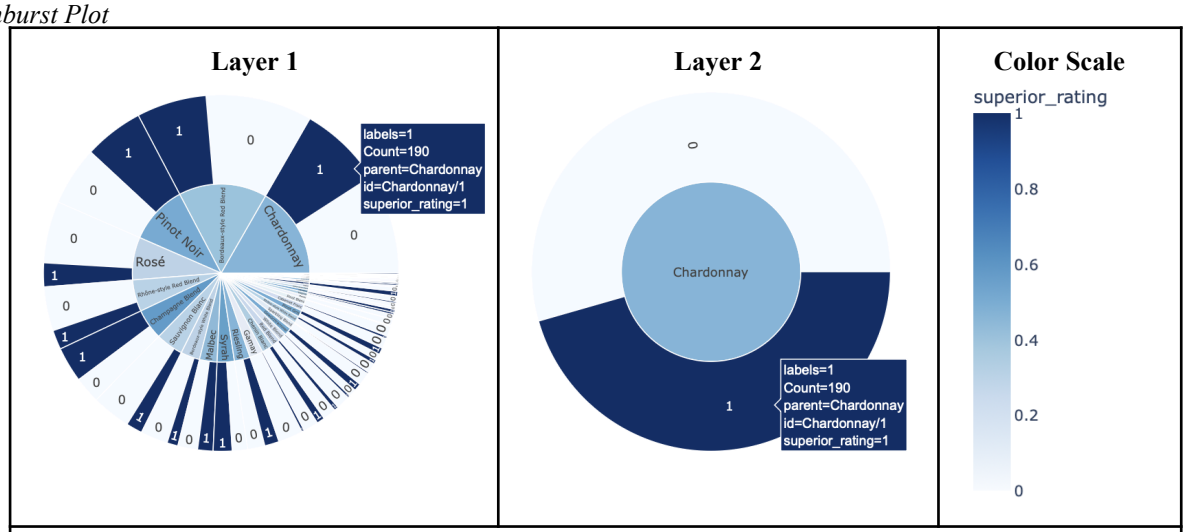
Median:
A line inside each box shows where the score for that wine is in relation to all other scores given; this number represents half way between highest possible mark down low end at 0%. So if you want an idea of how well a certain wine did overall just look at where its box falls vertically (higher=better). For example; Champagne Blends had many more good ratings compared to bad ones so it would be considered a highly rated wine.

Colors:
Each different color represents one type of wine so that it can easily be identified which is which.

Rosé:
The Rosé category has a range from loosely early 80s to loosely early 90s (81.0-91.0) with the middle point being around 87.5 suggesting consistently good quality throughout its offerings. Wine sellers could market Rosé as an approachable and dependable option for those looking for decent wines on a budget (~$190) without experiencing drastic fluctuations in standards met because they will always know what to expect from this particular variety due to its consistently wrapped up decent ratings over time.

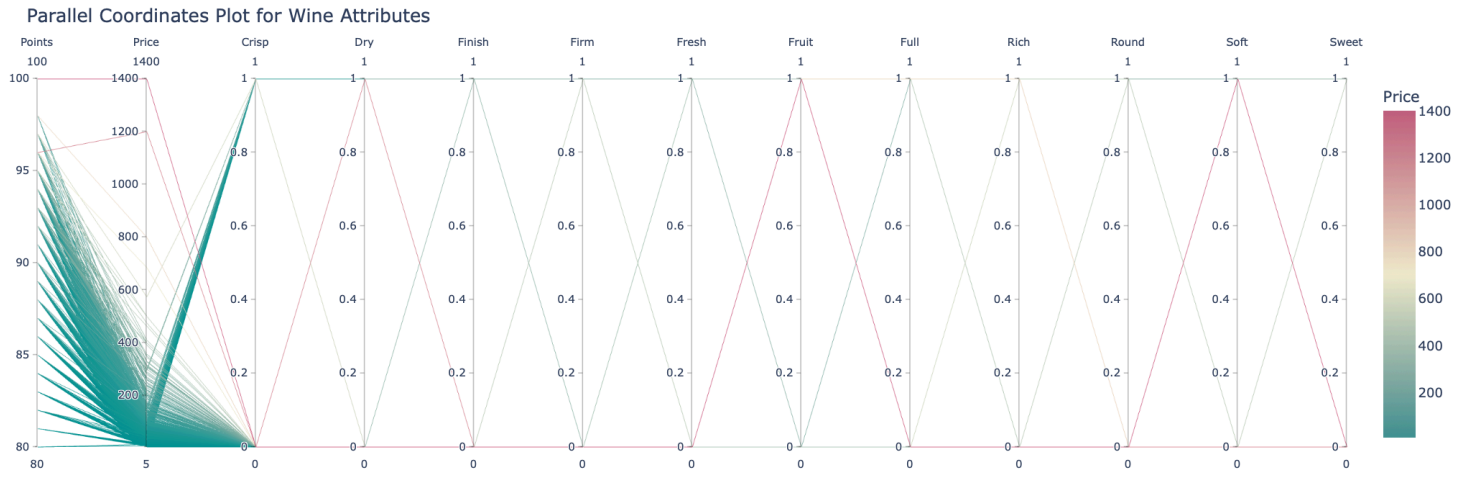
f) Additional Interactive and Creative Distributions

1) Sunburst Plot



The sunburst plot shows the distribution of wine types and their highest ratings, with a detailed look at Chardonnay which has the highest rating of 1 with 190 counts. This representation is good for showing levels of hierarchy as well as quantities because it allows you to see easily which varieties are most commonly considered excellent within the dataset.

2) Parallel Coordinates Plot



In parallel coordinates plot you can see how different wine attributes relate one another such as ‘Crisp’, ‘Dry’, ‘Finish’ etc., to ratings (‘Points’) and prices. The color gradient based on price makes it easy to tell apart visually; this draws attention to the fact that wines priced higher often score better in terms of ‘Points’ awarded them by tasters who also describe them using words like “rich” or “full.”

---

5) Model [Construction + Employment], Analysis, and Conclusion

a) Construction

In order to predict superior_rating based on points and price:

Bayesian Model: I created a logistic regression model using PyMC. This involved setting priors for the parameters and explaining how these parameters connect with the likelihood of seeing the data. What we want to do here is predict superior_rating based on points and price. For the regression coefficients, I assigned Normal priors, while observations were given a Bernoulli likelihood.

Logistic Regression Model Formula

For predicting the probability p that a wine is rated as "superior" (y = 1), based on its points and price,
and potentially their interaction, the logistic regression model can be written as:

$$\text{logit}(p) = \beta_0 + \beta_1 \times \text{points} + \beta_2 \times \text{price} + \beta_3 \times \text{points} \times \text{price}$$

where:

$\text{logit}(p) = \log(p/(p-1))$	is the log odds of the probability p that the wine is rated as “superior”.
β_0	The intercept term represents baseline log odds when any other predictor values(points and price) are at their reference levels (usually zero-centered if standardized).
β_1	Coefficient for the points variable.
β_2	Coefficient for 'price' variable.
β_3	Coefficient for interaction between 'points' and ‘price’

Prior Distributions:

$$\beta_i \sim \mathcal{N}(0, \sigma^2)$$

Likelihood Function:

$$y_i \sim \text{Bernoulli}(p_i), \text{ where } p_i = P(y_i = 1)$$

Posterior Distribution:

$$P(\beta|Y) = P(Y|\beta) * P(\beta)/P(Y)$$

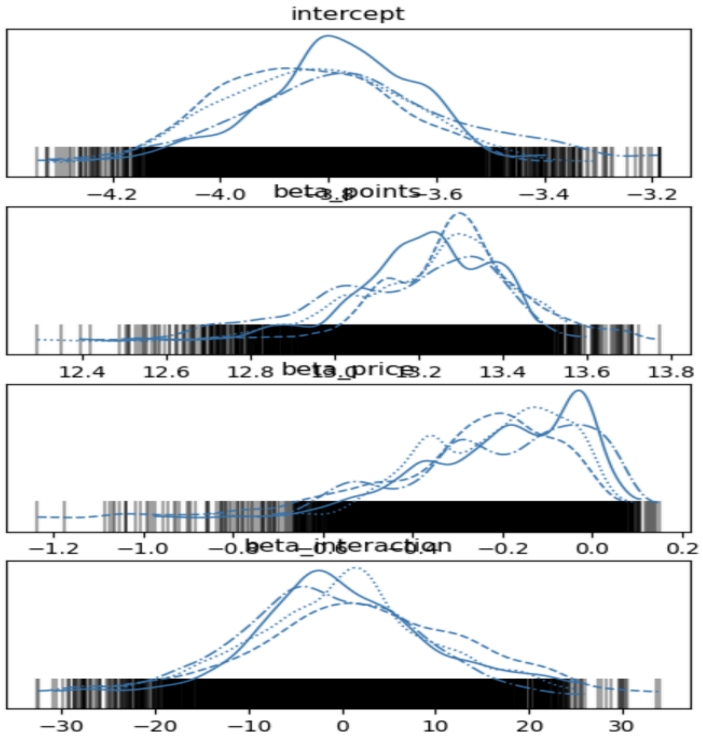
Here,

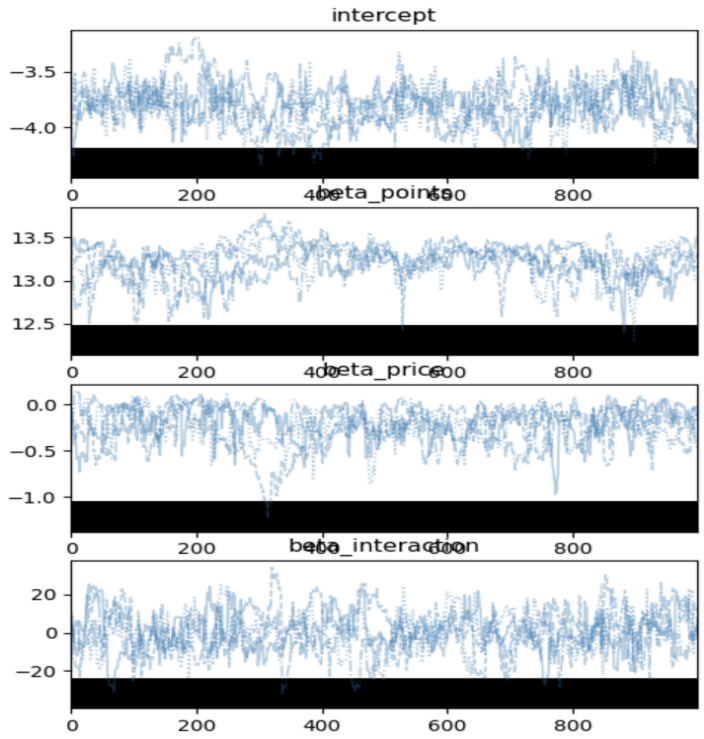
$P(\beta\|Y)$	posterior distribution of the parameters β given the data Y
$P(Y\|\beta)$	likelihood of the data Y given the parameters β
$P(\beta)$	prior distribution of the parameters
$P(Y)$	The data’s marginal likelihood - acts as a normalising constant

Fitting the Model Using MCMC: After that, I performed MCMC to draw samples from a posterior distribution over parameter values.

Diagnostics and Visualization: To analyze my MCMC’s output, which included diagnostics as well as posterior distributions, I used ArviZ.

Chaining: In an attempt to assess convergence and wholeness of my model, I employed multiple (4) chains within each MCMC run.





Posterior Distribution Plots:
These density plots are combined with histograms that reflect the posterior distributions of intercept, beta_points, beta_price and beta_interaction as model parameters. These figures represent the parameter's probability distribution after taking into account both data and priors.
Each diagram presents a curve in the shape of a bell which points to our estimated value's uncertainty. The mean or peak of this distribution identifies where according to evidence from the dataset gives us highest expectation for what an appropriate value might be.

Trace Plots:
The plots display values for each parameter across iterations sampled during MCMC process. In a good converging model, these traces should resemble a "hairy caterpillar" with lots of bunching up all over the chain without any trend visible or drifts over long term.
In terms convergence shown by trace plot, it is not same across them all. For instance beta_points and beta_price look like they want more investigation done on them because there seems some patterns or additional number of iterations / adjustments may be necessary.

Trace plot of intercept:
- This intercept trace shows fluctuation around the means of -3.5 to -4.0 which indicates that it converges steadily around these means. The density plot is narrow reflecting a high belief in estimated values for the intercept.

Trace Plot of Beta_Points:
- The `beta_points` trace oscillates stably around the means lying between 12.5 and 13.5 approximately. A corresponding density plot is sharply peaked showing accurate estimation as well as good convergence.

Trace Plot of Beta_Price:
- Comparatively stable trace appears moving near zero mean with fewer fluctuations for `beta_price` that signals less influence on response variable than `beta_points`. Density plot has a peak near zero with quite narrow spread indicating fine convergence.

Trace Plot of Beta_Interaction:
- The variability is more reflected by `beta_interaction` trace with some values swinging from +20 to -20. Wider Highest Density Interval (HDI) in summary table is due to wider corresponding density plot also having been indicated by its width.

```
with pm.Model() as logistic_model_reduced:
    intercept = pm.Normal("intercept", mu=0, sigma=10)
    beta_points = pm.Normal("beta_points", mu=0, sigma=10)
    beta_price = pm.Normal("beta_price", mu=0, sigma=10)
    beta_interaction = pm.Normal("beta_interaction", mu=0, sigma=10)

    logits = intercept + beta_points * X_train[:, 0] + beta_price * X_train[:, 1]
    observed = pm.Bernoulli("observed", logit_p=logits, observed=y_train)
    trace_reduced = pm.sample(1000, tune=1000, chains=4, return_inferencedata=True)

az.plot_trace(trace_reduced)
summary = az.summary(trace_reduced)
summary
```

index	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
intercept	-3.8	0.177	-4.135	-3.475	0.02	0.014	78.0	105.0	1.03
beta_points	13.218	0.192	12.805	13.518	0.026	0.018	53.0	73.0	1.06
beta_price	-0.223	0.193	-0.568	0.075	0.028	0.02	53.0	128.0	1.08
beta_interaction	0.256	10.253	-17.848	22.089	0.812	0.575	163.0	340.0	1.02

Intercept:
Mean: -3.8, this means that if both the points and price are at their average values a wine is rated as superior.
HDI (3%, 97%): -4.135 to -3.475 – This shows the majority of the posterior distribution in addition to giving a range for credible values for an intercept.

Beta_Points (Coefficient for points):
Mean: 13.218 – Indicates there is strong positive correlation between log odds of being superior with points above average. Higher points greatly improve the prospect to be rated superior.
HDI (3%, 97%): from 12.805 to 13.518 – In fact it is very narrow so we can say that effect size has been estimated accurately because its confidence interval is tight which implies high certainty in this regard.

Beta_Price (Coefficient for price):
Mean: -0.223; This means there's slightly negative relationship between prices and log odds of being better off although such effect remains small.
HDI (3%, 97%): -0.568, 0.075 – Here zero falls within the range representing the incertitude of reporting the reliability on price for accurately determining superior ratings.

Convergence Assessment in MCMC Analysis:

$\hat{R}$**(R_hat)**	**ESS**
Reasonable convergence is indicated when all the $\hat{R}$ values are nearly 1.0 for each parameter. The $\hat{R}$ values lie between 1.02 and 1.08, which means that they are getting closer to converge but some numbers greater than 1.05 may show they need a better look or more iterations.	Good sampling efficiency is suggested by relatively high ESS values across all parameters. `beta_price` has an ESS of 53.0 while `beta_interaction``s effective sample size equals to 340; hence enough samples have been taken to ensure stable estimation for these quantities.

Conclusion for both Non-Technical and Technical Audience:
My analysis demonstrates that 'points' is a good indicator for superior_ratings. What this means is that the more points a wine has, the more likely it will be rated as superior. 'Price' is relatively less important to note in terms of its effect on wine quality; however this does not mean all expensive wines are better than cheaper ones. This finding should help sellers understand which wines are perceived as being higher value for money by consumers – regardless their prices – thus enabling them make rational decisions while purchasing.

Bayesian logistic regression analysis worked well to identify factors affecting superior_rating of wines. Therefore convergence and reliability of the results were confirmed through MCMC sampling used during estimation of model parameters such as $\hat{R}$ values and effective sample sizes (ESS) obtained from trace plots with density plots indicating stable convergence mainly around 'price' and 'points' together. In case there was ambiguity about an interaction term where its impact on ratings had been assessed comprehensively so that full understanding could be given concerning this factor's influence towards perception about different types of wines were made clear during reporting process.

Further recommendations – would other methods/approaches/different data be more suitable?
To make the study more comprehensive, parameters such as wine regions, grape varieties and vintages can be incorporated as additional information to identify factors that affect the quality and price of wines.
Another consumer oriented approach may involve the use of consumer reviews and ratings as part of the variables.
Also it may be useful in evaluating vintage effects especially if a time series analysis is done which looks at trends over time.
In capturing complex nonlinear relationships and interactions, methodologies in machine learning such as support vector machines (SVM) or random forests can be used for better predictive precision.
These advanced analytic methods coupled with an expanded dataset are likely to produce broader, better, detailed, and comprehensive findings.
