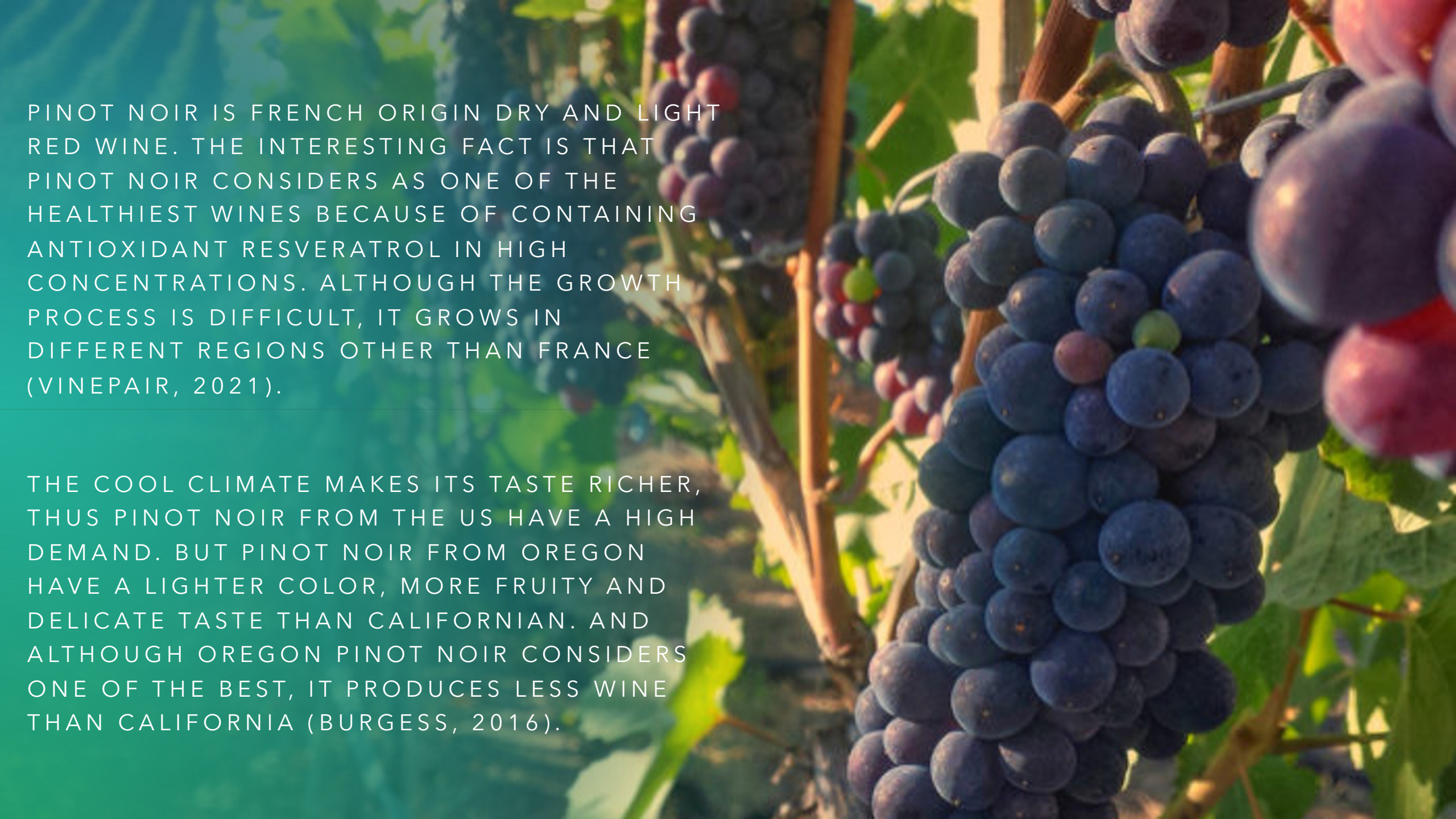
# Week 6: Final Project Pinot Noir from US

STUDENT: FATIMA NURMAKHAMADOVA

INSTRUCTOR: TOM BREUR

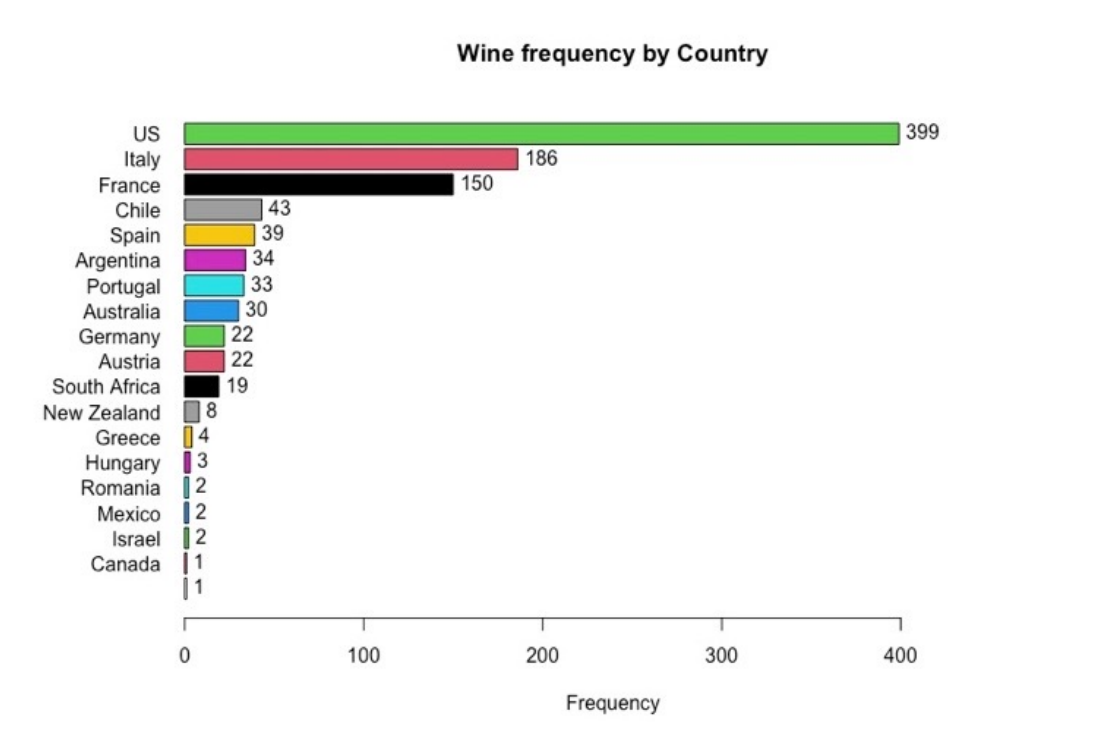
CLASS: ALY6010 12/19/2021



PINOT NOIR IS FRENCH ORIGIN DRY AND LIGHT RED WINE. THE INTERESTING FACT IS THAT PINOT NOIR CONSIDERS AS ONE OF THE HEALTHIEST WINES BECAUSE OF CONTAINING ANTIOXIDANT RESVERATROL IN HIGH CONCENTRATIONS. ALTHOUGH THE GROWTH PROCESS IS DIFFICULT, IT GROWS IN DIFFERENT REGIONS OTHER THAN FRANCE (VINEPAIR, 2021).

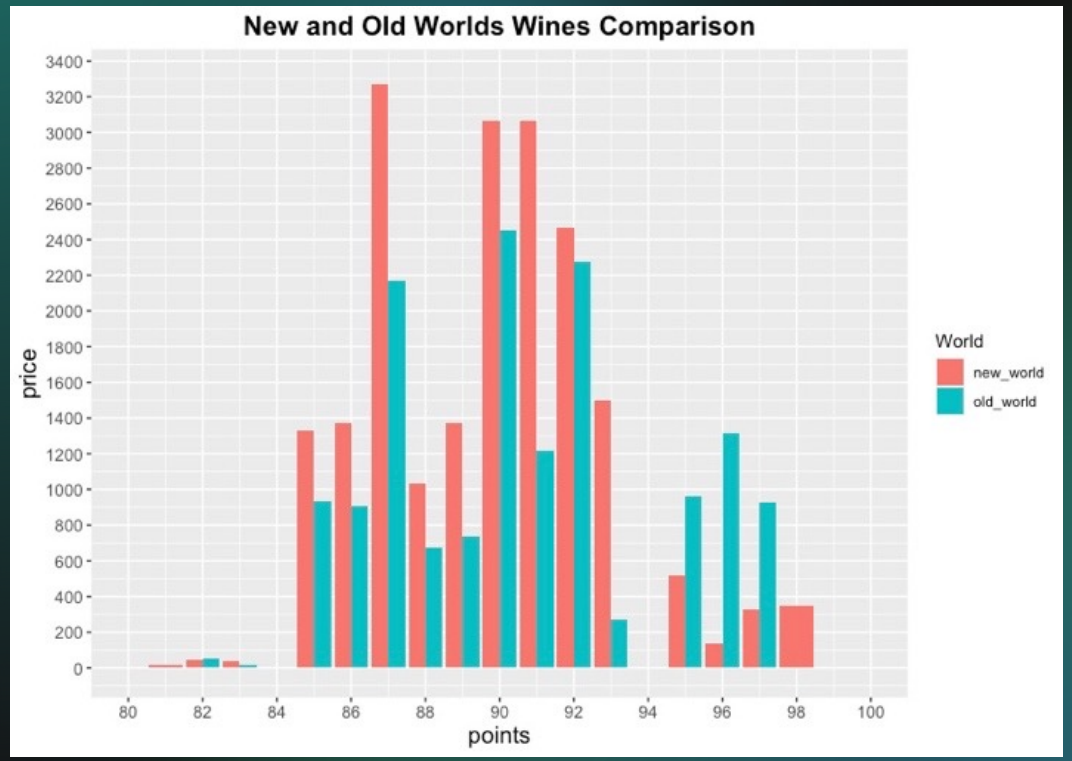
THE COOL CLIMATE MAKES ITS TASTE RICHER, THUS PINOT NOIR FROM THE US HAVE A HIGH DEMAND. BUT PINOT NOIR FROM OREGON HAVE A LIGHTER COLOR, MORE FRUITY AND DELICATE TASTE THAN CALIFORNIAN. AND ALTHOUGH OREGON PINOT NOIR CONSIDERS ONE OF THE BEST, IT PRODUCES LESS WINE THAN CALIFORNIA (BURGESS, 2016).

***Figure 1:*** *Wine Frequency by Country*



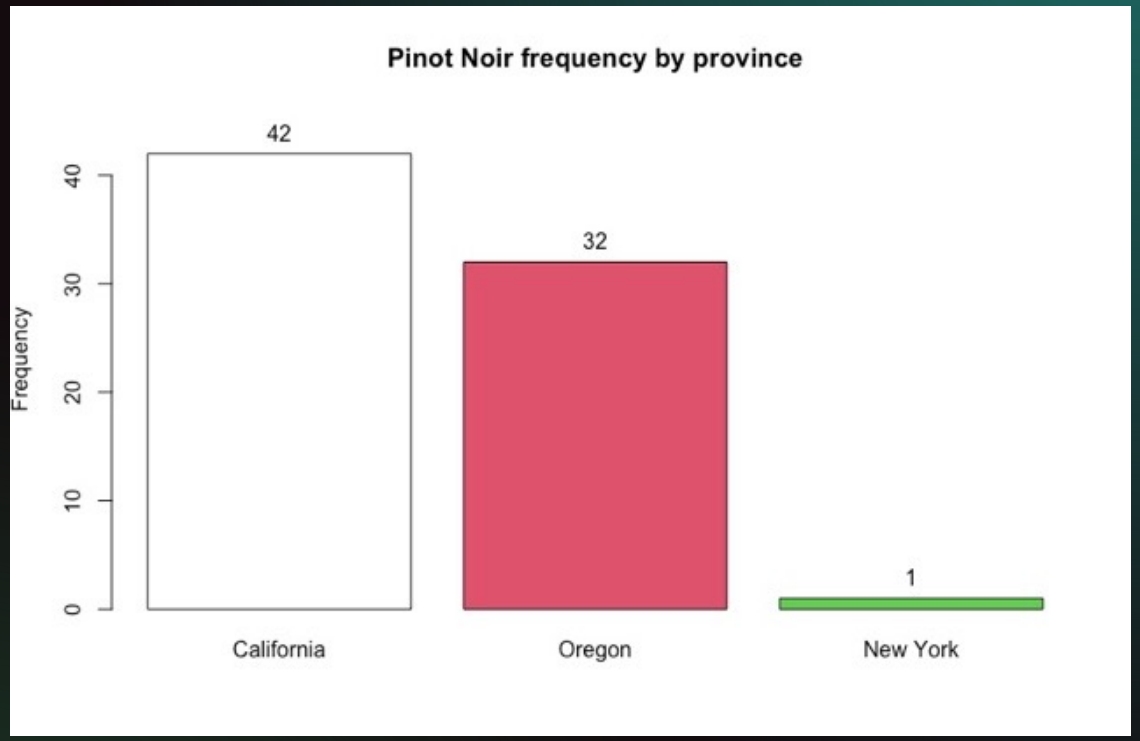

There are 18 countries in wine dataset. The US is dominating with 399 wine types, then Italy with 186 wine types, and in the third place is France with 150 wine types.

***Figure 2:*** *New and Old-World Price Comparison*



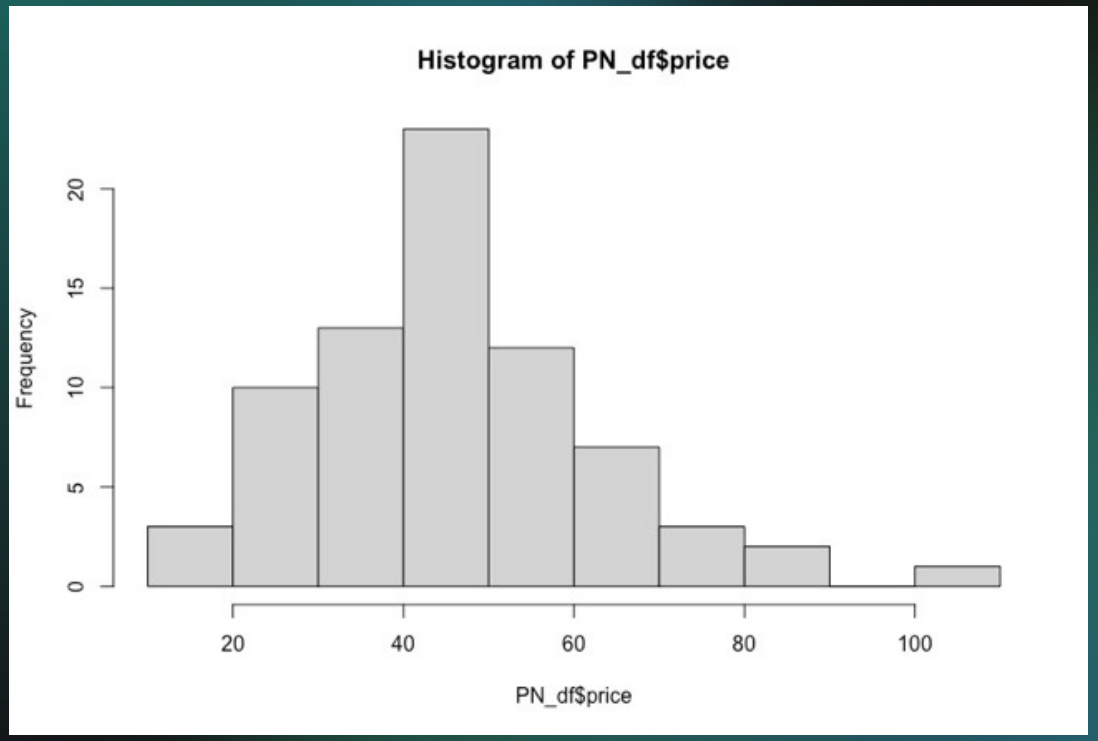

New World wines have the higher price though not higher points than Old World wines. While Old World have more ranked wines with lower prices and in lower quantity.

***Figure 3:*** *Pinot Noir Frequency by Province*




These are two major provinces that produce Pinot Noir in the US, California, and Oregon. But California has 42 wines while Oregon 32 which is for 10 wines less.

***Figure 4:*** *Price distribution of Pinot Noir in the US*




The price of Pinot Noir in the US has normal distribution. The average price is $46.5, and the median is $47.3.

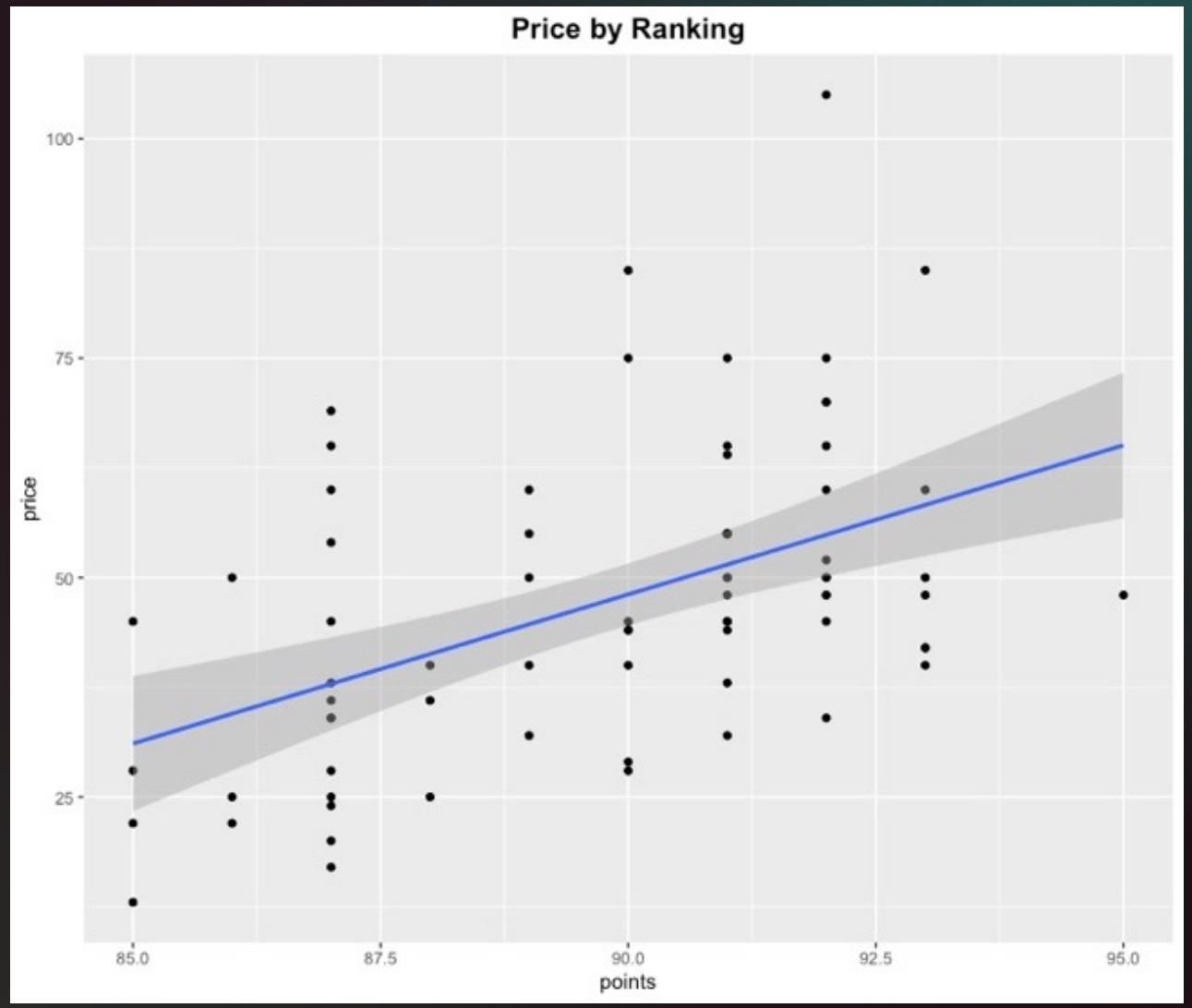



***Figure 5:*** *ggplot and Linear Regression for Price by Ranking in the US*

The ggplot demonstrates the relationship of price and points in the US. There are some outliers, thus the relationship seems roughly linear. For now, the assumption is that there is a positive linear relationship between two variables which will be tested by regression model.

# Question 1: How does rank (points) affect the price of Pinot Noir in the US?

***Figure 6:*** *Two-sample t-test for the US Wine Points*

```
> ##QUESTION 1:
> #Is there any significant difference in points between California and Oregon wine?
> ##Step 1 - Two-sample t-test
> t.test(points ~ province, data = PN_df, var.equal = TRUE)

        Two Sample t-test

data:  points by province
t = 0.82188, df = 72, p-value = 0.4139
alternative hypothesis: true difference in means between group California and group Oregon is not equal to 0
95 percent confidence interval:
 -0.678801  1.631182
sample estimates:
mean in group California     mean in group Oregon
                89.97619                 89.50000
```

***Figure 7:*** *Two-sample t-test for the US Wine Price*

```
> ##Step 2 - Two-sample t-test: Is there any significant difference in price between California and Oregon wine?
> t.test(price ~ province, data = PN_df, var.equal = TRUE) #no significant difference

        Two Sample t-test

data:  price by province
t = -0.25007, df = 72, p-value = 0.8032
alternative hypothesis: true difference in means between group California and group Oregon is not equal to 0
95 percent confidence interval:
 -9.131839  7.096125
sample estimates:
mean in group California     mean in group Oregon
                46.85714                 47.87500
```

**Step 1 - Two-sample t-test:** *Is there any significant difference between California and Oregon wine points?*

The results showed that mean points was almost equal in both provinces, in California 89.98 compared to Oregon 89.5, no statistically significant difference of *0.48 (95% CI, -0.68 to 1.63) USD, t(72) = 0.82 p = .4139.* The null hypothesis is not rejected.

**Step 2** - Two-sample t-test: *Is there any significant difference in price between California and Oregon wine?*

The results showed that mean price was almost equal in both provinces, in California $46.85 compared to Oregon $47.87, no statistically significant difference of *1.018 (95% CI, -9.13 to 7.09) USD, t(72) = -0.25, p = .8032.* The null hypothesis is not rejected.

```
> ##Step 3 - Regression testing of relationship between Price and points of Pinot Noir in the US
> lm_points <- lm(PN_df,formula=price ~ points)
> summary(lm_points) #R2 0.2255

Call:
lm(formula = price ~ points, data = PN_df)

Residuals:
    Min      1Q  Median      3Q     Max
-20.881  -9.880  -3.679   9.373  50.124

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -257.8269    64.7095  -3.984  0.00016 ***
points         3.3989     0.7206   4.717 1.14e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.17 on 72 degrees of freedom
Multiple R-squared:  0.2361,    Adjusted R-squared:  0.2255
F-statistic: 22.25 on 1 and 72 DF,  p-value: 1.142e-05
```

## *Figure 8: Regression Test for the Price and Points of wine in the US*

**Step 3** - Regression testing of relationship between Price and points of Pinot Noir in the US

The results show the significant relationship between price and rank ($p < 0.001$, $R2 = 0.2255 \pm 0.0106$, $F(1, 72) = 22.25$), with a 3.39$ increase in price for every point increase in rank. Although only 22.5% of price can be explained by wine points. The null hypothesis is rejected.

# Question 2: How does location (province) affect the price of Pinot Noir in the US?

**Step 3** - Regression testing of relationship between Price and points of Pinot Noir in the US

The results show the significant relationship between price and rank (*p <0.001, R2 = 0.2255 ± 0.0106, F (1, 72) = 22.25),* with a 3.39$ increase in price for every point increase in rank. Although only 22.5% of price can be explained by wine points. The null hypothesis is rejected.

```
> ##QUESTION 2:
> #How does location (province) affect the price of Pinot Noir in the US?
> ##Step 1- Regression testing of relationship between Price and both locations
> lm_province <- lm(PN_df,formula=price ~ province)
> summary(lm_province) #location itself does not have any affect on the price

Call:
lm(formula = price ~ province, data = PN_df)

Residuals:
    Min      1Q  Median      3Q     Max
-34.875 -10.857  -0.866   8.143  57.125

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      46.857      2.677   17.51   <2e-16 ***
provinceOregon    1.018      4.070    0.25    0.803
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.35 on 72 degrees of freedom
Multiple R-squared:  0.0008678, Adjusted R-squared:  -0.01301
F-statistic: 0.06253 on 1 and 72 DF,  p-value: 0.8032
```

***Figure 9:*** *Regression Test for the Price and Location in the US*

**Step 2** - Regression testing of relationship between Price and Points + locations

A multiple regression model was calculated to test if ranking and location taken significantly predicted wine price in the US. The *price* is a dependent variable, and *points* + *province* are independent variables (Figure 10). The results show a significant relationship between price and points + locations (*p=0, R2 = 0.242, F (1, 72) = 11.33),* with a 3.45$ increase in price for every point increase in rank in total. And with a 2.66$ increase in price for every point increase in rank in Oregon. And only 24.2% of price can be explained by wine points and locations which is 2% higher than the regression test made for the points only (Figure 8). The null hypothesis is rejected.

*Figure 10: Regression Test for the Price and Points + Location in the US*

```
> ##Step 2-Regression test: Check how the ranking & locations affect the price
> lm_prov_price <- lm(PN_df,formula=price ~ points+province)
> summary(lm_prov_price)

Call:
lm(formula = price ~ points + province, data = PN_df)

Residuals:
    Min      1Q  Median      3Q     Max
-19.841 -10.160  -2.995   8.477  48.498

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -263.6405    65.3838  -4.032 0.000137 ***
points            3.4509     0.7262   4.752 1.02e-05 ***
provinceOregon    2.6611     3.5870   0.742 0.460600
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.22 on 71 degrees of freedom
Multiple R-squared:  0.242,     Adjusted R-squared:  0.2206
F-statistic: 11.33 on 2 and 71 DF,  p-value: 5.36e-05
```

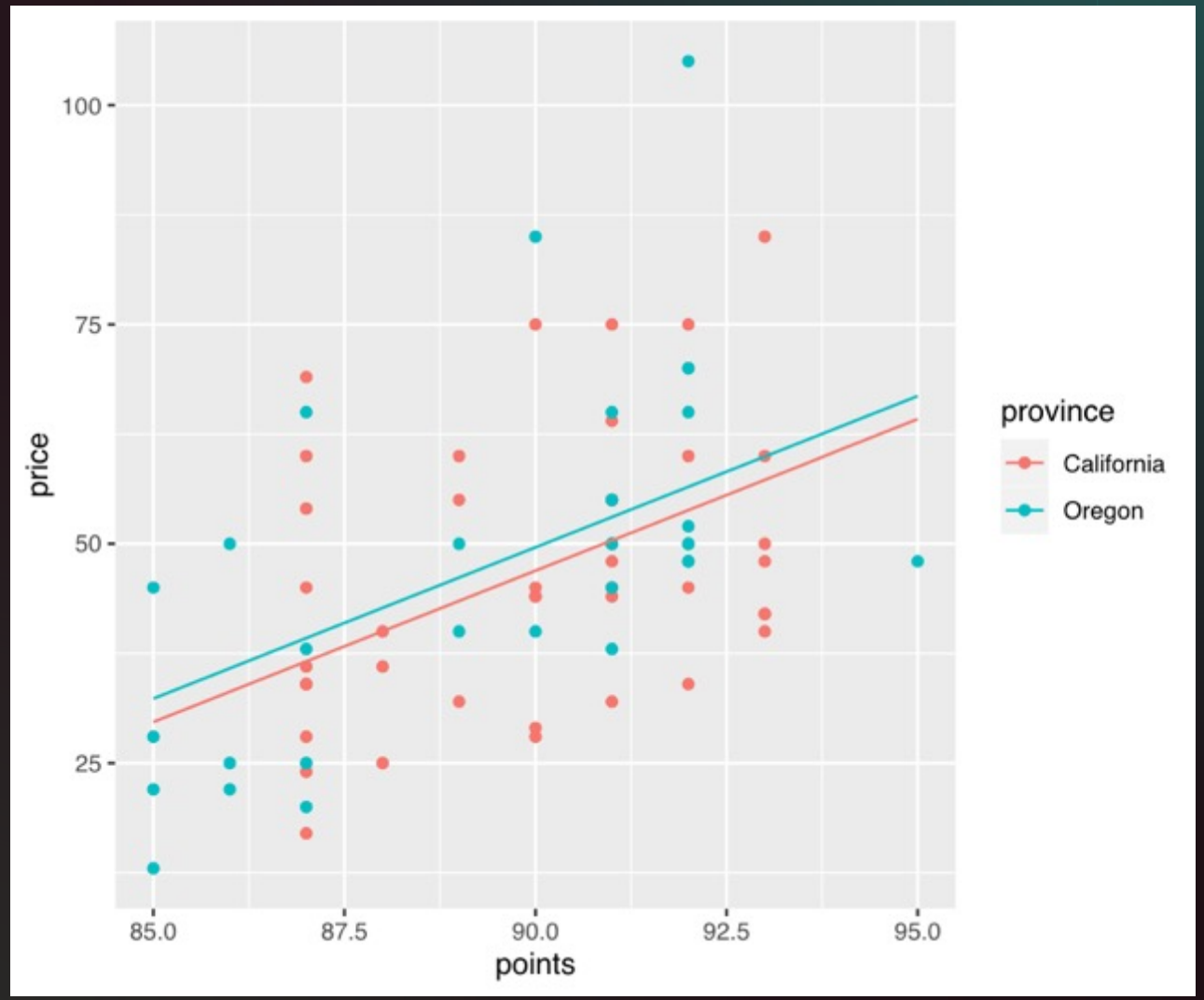



***Figure 11:*** *Multiple Linear Regression of Price and Ranking by Provinces*

Figure 11 demonstrates that Pinot Noir price in Oregon is higher than in California corresponds to the regression models in Figure 8. This is also might be explained by the mean price of wine in Oregon which was a bit higher than in California (Figure 7). Although the relationship between points and price is less clear, the is still a positive linear relationship. Thus, some wines of Pinor Noir variety have higher price for higher rank.

# Question 3: How does ranking affect the price in each location?

```
> #in California
> lm_cal_price <- lm(PN_California,formula=price ~ points)
> summary(lm_cal_price)

Call:
lm(formula = price ~ points, data = PN_California)

Residuals:
    Min      1Q  Median      3Q     Max
-22.194 -12.056  -3.056   7.401  30.357

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -184.805     95.164  -1.942   0.0592 .
points         2.575      1.057   2.435   0.0194 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.69 on 40 degrees of freedom
Multiple R-squared:  0.1291,    Adjusted R-squared:  0.1073
F-statistic: 5.929 on 1 and 40 DF,  p-value: 0.01944
```

***Figure 12:*** *Regression Test for the Price and Points in California*

A simple regression model was calculated to predict wine price based on its ranking where *price* as a dependent variable, and *points* as independent variable of Pinot Noir in California (Figure 12). The results show the significant relationship between price and rank ($p < 0.05$, $R2 = 0.1073 \pm 0.0224$, $F(1, 40) = 5.9$), with a 2.5$ increase in price for every point increase in rank. Although only 10.7% of price can be explained by points. The null hypothesis is rejected.

## ***Figure 13:*** *Regression Test for the Price and Points in Oregon*

A simple regression model was calculated to predict wine price based on its ranking where *price* as a dependent variable, and *points* as independent variable of Pinot Noir in Oregon (Figure 13). The results show the significant relationship between price and rank ($p < 0.001$, $R2 = 0.337 \pm 0.0214$, $F(1, 30) = 16.76$), with a 4.1$ increase in price for every point increase in rank. About 33.7% of price can be explained by points. The null hypothesis is rejected.

```
> #in Oregon
> lm_oreg_price <- lm(PN_Oregon,formula=price ~ points)
> summary(lm_oreg_price)

Call:
lm(formula = price ~ points, data = PN_Oregon)

Residuals:
    Min      1Q  Median      3Q     Max
-22.635 -10.013  -4.944   7.814  46.779

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -322.495     90.525  -3.562 0.001250 **
points         4.138      1.011   4.093 0.000295 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.86 on 30 degrees of freedom
Multiple R-squared:  0.3584,    Adjusted R-squared:  0.337
F-statistic: 16.76 on 1 and 30 DF,  p-value: 0.0002953
```

# Conclusion

- As a result, we can conclude that there is a significant relationship between price and points of Pinot Noir in the US. Although location did not show the significant relationship with the price, the regression model and ggplot shows that Oregon wine price better fits the trend than wine price in California.

- Although both provinces have low and high-cost wines, Oregon has higher ranked thus expensive wines.

# References:


- Vivino. (2020, October 20). *10 unmissable facts about Pinot noir grapes & wines*. Retrieved December 20, 2021, from https://www.vivino.com/wine-news/10-unmissable-facts-about-pinot-noir-grapes-wines
- Enthusiast, W. (2018, July 6). *11 food-friendly California pinot noirs*. Wine Enthusiast. Retrieved December 20, 2021, from https://www.winemag.com/2018/07/06/california-pinot-noirs/