Week 6:
Final Project

Student: Fatima Nurmakhamadova
Instructor: Tom Breur

Class: ALY6010 12/19/2021

Exploratory Data Analysis of wine dataset

The dataset “Wine_tasting.csv” contains the information about variety of wine, winery, country of origin, tasters, price, and points(ranking). And consists of 1000 rows, and 14 columns (variables). It contains both, text, and numerical data. While text data is categorical, the numerical data is both, continuous (price, points) and discrete (order of wines in a dataset).

The purpose of the dataset is to give detailed information about each wine type all over the world. As for data source, the data in “Wine_tasting.csv” dataset is a result of taste analysis of wines grown in different regions sorted by variety, price, and points.

Figure 1 demonstrates that there are 18 countries in this 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.

Further, thanks to Professor Tom Breur, I decided to compare price and points of wines dividing countries by new and old world. Basically, the New World represents wines origin from post-colonial countries (VinePair, 2018). In this dataset there are 8 such countries: Argentina, Australia, Canada, Chile, Mexico, New Zealand, South Africa, and US. Whereas the Old World represents wines origin from Europe, and Middle East (VinePair, 2018). In this dataset there are 10 such countries: Austria, France, Germany, Greece, Hungary, Israel, Italy, Portugal, Romania, and Spain.

Figure 2 demonstrates that New World wines have the higher price though not higher points than Old World wines. While New World countries also have wines with higher ranking and low price, most of their wines are ranked between 85-93. Whereas Old World have more ranked wines with lower prices and in lower quantity.

The most wines are produced by New World (incl. US with most of its wines) having 87 points with total price of $3273.

Pinot Noir from US

Further, I decided to analyze one of the most popular wine variety called Pinot Noir which is also one of the dominating varieties among New World wines. It 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 3 shows two major provinces that produce Pinot Noir in the US, California, and Oregon. But we should keep in mind that California has 42 wines while Oregon 32 which is for 10 wines less. Thus, the further analysis of the factors that affect the wine price will be focused on these two locations. Figure 4 shows the price of Pinot Noir in the US has normal distribution. The average price is $46.5, and the median is $47.3.

There are several factors that might drive the wine price, which is the dependent variable. One of them is wine ranking which is named as '*points*' in the dataset, another is location (provinces). First part of the analysis is related to the US, thus both provinces together. Second part of analysis is separately by provinces. Thus, first analysis will be related to the price and ranking relationship. The second will be the relationship between price and location. Both t-test and regression test will be applied and explained with corresponding graphs, and hypothesis testing.

The main question is to find out if the high rank results higher price of Pinot Noir in the US since it is commonly perceived that the higher ranked products are more expensive.

Figure 5 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?

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

H 0: $\mu 1 = \mu 2$
H 1: $\mu 1 - \mu 2 \neq d$

An independent t-test was run on two samples of the US provinces with 73 wines, to determine if there was a statistically significant difference in mean points of Pinot Noir between California, and Oregon (Figure 6). There were 42 wines from California, and 32 wines from Oregon. 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?*

H 0: $\mu 1 = \mu 2$
H 1: $\mu 1 - \mu 2 \neq d$

An independent t-test was run on two samples of the US provinces with 73 wines, to determine if there was a statistically significant difference in mean price of Pinot Noir between California, and Oregon (Figure 7). There were 42 wines from California, and 32 wines from Oregon. 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

H0: $\beta 1 = 0$
HA: $\beta 1 \neq 0$

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 the US (Figure 8). 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 provinces

H0: $\beta 1 = 0$
HA: $\beta 1 \neq 0$

A simple regression model was calculated to predict wine price based on location where *price* as a dependent variable, and *province* as independent variable of Pinot Noir in the US (Figure 9). The results show no significant but quite negative relationship between price and location (*p=0, R2 = -0.013, F (1, 72) = 0.062).* The null hypothesis is not rejected. The selected model does not follow the trend of the data.

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

H0: $\beta 1 = 0$
HA: $\beta 1 \neq 0$

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.

Step 3 - 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.

To make an accurate conclusion, further analysis will subset the dataset by provinces.

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

1. **Regression testing of relationship between Price and Points in California**

H0: $\beta 1 = 0$
HA: $\beta 1 \neq 0$

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.

2. **Regression testing of relationship between Price and Points in Oregon**

H0: $\beta 1 = 0$
HA: $\beta 1 \neq 0$

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.

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.

Appendix:

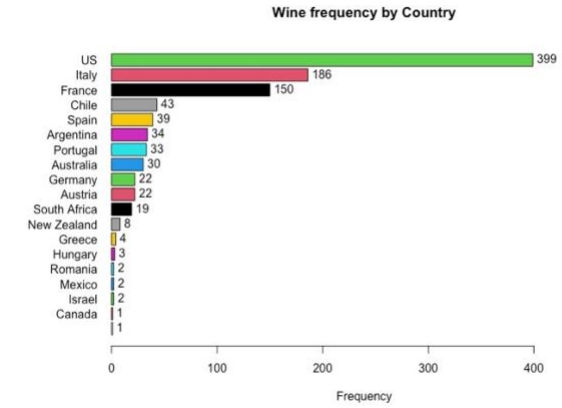

Figure 1: *Wine Frequency by Country*

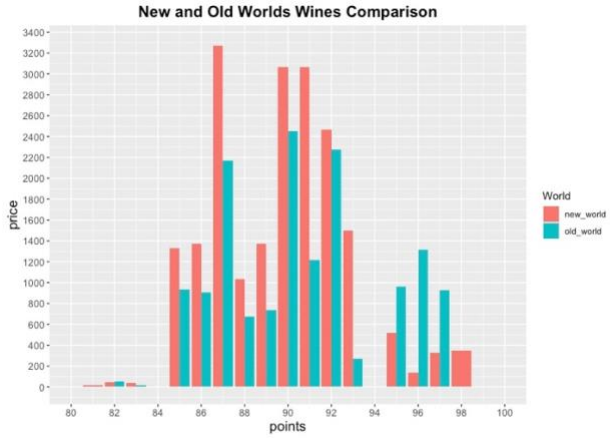

Figure 2: *New and Old World Price Comparison*

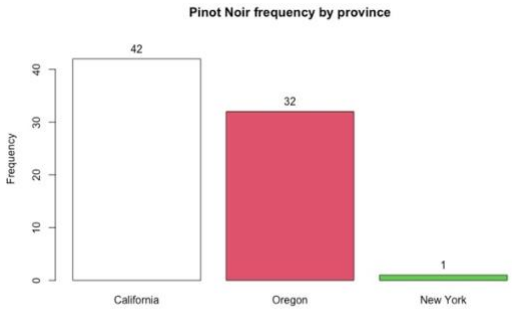

Figure 3: *Pinot Noir Frequency by Province*

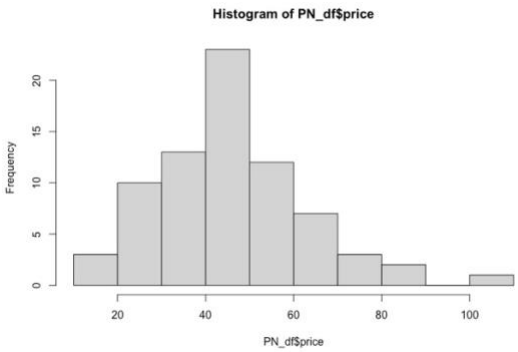

Figure 4: *Price distribution of Pinot Noir in the US*

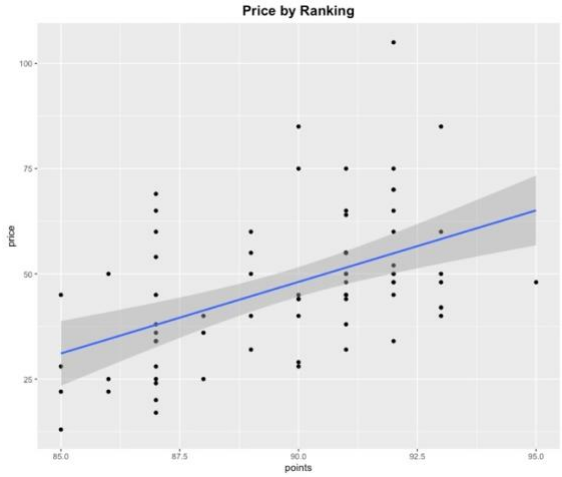

Figure 5: *ggplot and Linear Regression for Price by Ranking in the US*

```
> ##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 6: *Two-sample t-test for the US Wine Points*

```
> ##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
```

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

```
> ##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*

```
> ##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 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 10: Regression Test for the Price and Points + Location in the US

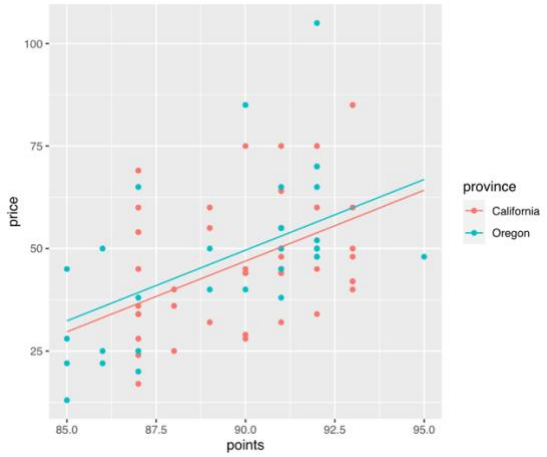

Figure 11: *Multiple Linear Regression of Price and Ranking by Provinces*

```
> #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*

```
> #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
```

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

References:

Bevans, R. (2020, December 14). *A step-by-step guide to linear regression in R*. Scribbr. Retrieved December 20, 2021, from https://www.scribbr.com/statistics/linear-regression-in-r/

Bluman, A. G. (2018). Chapters 1, 2, 3 & 4 In Elementary statistics: A step by step approach (pp. 1-168). New York, NY: McGraw-Hill Education.

Burgess, words: L. (2016, May 22). *Pinot noir reveals the differences between Oregon and California climates*. VinePair. Retrieved December 20, 2021, from https://vinepair.com/wine-blog/pinot-noir-reveals-the-differences-between-oregon-and-california-climates/

Kabacoff, R. (2015). R in Action, Second Edition (2nd edition). ISBN 978-1-617-29138-8

Independent-samples T-test using R, Excel and RStudio (page 4). Independent-samples t-test using R, Excel and RStudio (page 4) | Interpreting and reporting the results for an independent-samples t-test. (n.d.). Retrieved December 20, 2021, from https://statistics.laerd.com/r-tutorials/independent-samples-t-test-using-r-excel-and-rstudio-4.php

Kassambara. (2018). *Subset data frame rows in R*. Datanovia. Retrieved November 12, 2021, from https://www.datanovia.com/en/lessons/subset-data-frame-rows-in-r/.

kohske. (2011, May 22). *Adding space between bars in GGPLOT2*. Stack Overflow. Retrieved November 12, 2021, from https://stackoverflow.com/questions/6085238/adding-space-between-bars-in-ggplot2.

M, M. (2019, January 4). *How to sum a variable by Group*. Stack Overflow. Retrieved November 12, 2021, from https://stackoverflow.com/questions/1660124/how-to-sum-a-variable-by-group.

Machlis, S. (2017, August 18). *Beginner's Guide to r: Easy ways to do basic data analysis*. Computerworld. Retrieved November 12, 2021, from https://www.computerworld.com/article/2598083/app-development-beginner-s-guide-to-r-easy-ways-to-do-basic-data-analysis.html?page=2.

Malik, S. (2019, April 17). *Data Analysis and visualisations using R*. Medium. Retrieved November 12, 2021, from https://towardsdatascience.com/data-analysis-and-visualisations-using-r-955a7e90f7dd.

Marsja, E. (2020, November 8). *How to add a column to a dataframe in R with tibble & dplyr*. Erik Marsja. Retrieved November 12, 2021, from https://www.marsja.se/how-to-add-a-column-to-dataframe-in-r-with-tibble-dplyr/.

Schork, J. (2020, September 30). *Conditionally remove row from data frame in R (example): Delete rows*. Statistics Globe. Retrieved December 18, 2021, from https://statisticsglobe.com/r-remove-row-from-data-frame-condition

Soetewey, A. (2020, January 22). *Descriptive statistics in R*. Stats and R. Retrieved November 12, 2021, from https://statsandr.com/blog/descriptive-statistics-in-r/.

Thieme, C. (2021, June 16). *Understanding linear regression output in R*. Medium. Retrieved December 20, 2021, from https://towardsdatascience.com/understanding-linear-regression-output-in-r-7a9cbda948b3

timmur. (2016, July 18). *Revise the number of ticks in the x-axis?* Stack Overflow. Retrieved November 12, 2021, from https://stackoverflow.com/questions/38438989/revise-the-number-of-ticks-in-the-x-axis.

Unpaired two-samples T-test in R. STHDA. (n.d.). Retrieved December 18, 2021, from http://www.sthda.com/english/wiki/unpaired-two-samples-t-test-in-r

VinePair. (2018, August 27). *The Guide to Old World Wine vs. New World Wines: Wine 101*. VinePair. Retrieved November 12, 2021, from https://vinepair.com/wine-101/guide-old-world-vs-new-world-wines/.

VinePair. (2021, August 10). *Learn about pinot noir: Wine 101 (updated 2020)*. VinePair. Retrieved December 18, 2021, from https://vinepair.com/wine-101/learn-pinot-noir/

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

Zach. (2021, May 18). *The Complete Guide: How to report regression results*. Statology. Retrieved December 20, 2021, from https://www.statology.org/how-to-report-regression-results/