Analysis of Optimal Growing Conditons of Grapes in Italy

Group 12

December 6th, 2024

- Egbo Joseph (143037470)
- Goudy Joshua (169031329)
- Thambiaiah Melissa (169060509)
- Umar Emaan (169108097)


```
suppressWarnings(suppressPackageStartupMessages({
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
library(tidyverse)
library(knitr)
library(scatterplot3d)
library(corrr)
library(ggridges)
library(arrow)
library(tidymodels)
library(viridis)
library(kableExtra)
library(ggrepel)
library(patchwork)}))
```

Introduction

The Italian Wine Grapes data set provides a comprehensive understanding of the cultivation processes pertaining to grapes alongside the relationship that exists with the production of high-quality wine. With a focus on select regions across the heart of Italy, the 'grapes' data set captures the essence of varying elements associated with grape cultivation including the region, variety, sugar content, quality score, rainfall, and sun exposure over the duration of August to September. This report allows individuals to explore the intersections between agriculture and statistics while offering valuable insights into the optimal conditions that yield premium wines. The link between specific factors and their contributions to grape cultivation can be perceived by examining the trends and correlations that prevail from this study.

Abstract

The fundamental purpose of this analysis is to determine which cultivation processes in the 'grapes' data set can maximize the grape quality. In order to achieve this purpose, the goals are as required; the creation of graphical and predictive models in order to fulfill the research question, "How effective is the specific factor's contribution to the grape quality score?".

```
grapes_0<-read_csv('GRAPE_QUALITY.csv')
```

Data Cleaning:

```
grapes<-grapes_0|>
  select(!(sample_id))
grapes$quality_category<-factor(grapes$quality_category, levels =
                                  c('Low', 'Medium', 'High', 'Premium'))
grapes<-grapes|> mutate(wine_type = case_when(
        variety == 'Riesling' ~ 'White',
        variety == 'Pinot Noir' ~ 'Red',
        variety == 'Sauvignon Blanc'~'White',
        variety == 'Merlot'~'Red',
        variety == 'Zinfandel'~'Red',
        variety == 'Chardonnay'~'White',
        variety == 'Syrah'~'Red',
        variety == 'Cabernet Sauvignon'~'Red',
        .default = variety),
      harvest_week = as.integer(week(harvest_date)) - 30 )
grapes<-grapes|> mutate(size_category = case_when(
        berry_size_mm < 15 ~ 'small',
        berry_size_mm < 20 ~ 'medium',
        berry_size_mm >= 20 ~ 'large',
        .default = 'uncategorized' ))
grapes$size_category<-factor(grapes$size_category, levels = c('small', 'medium', 'large'))
```

In order to clean the data for the purpose of making it more comprehensible and effective for analysis, a variable was created to allow for the categorization of red versus white wine. Similarly, cleaning the data entailed the creation of a scale for berry sizes in an attempt to separate the grapes by size into three categories ranging from small, medium and large. Ultimately, a summary table of the numerical values within the data were organized that corresponding to the relationship that exists between specific variables which possessed the most significant impact in contributing to the quality of the grape.

Exploratory Table 1: Feature Correlation

```
grapes_split <- initial_validation_split(grapes, prop = c(0.6, 0.2))
grapes_train <- training(grapes_split)
```

```
generate_correlation_table <- function(data) {
  continuous_vars <- data %>% select_if(is.numeric)
  correlation_matrix <- cor(continuous_vars, use = "complete.obs")
  correlation_table <- as.data.frame(correlation_matrix)
  return(correlation_table) }
# Function Source: Footnote 2
grapes_woqs<-grapes_train|>select(!(quality_score))
round_table <- function(data, digits = 1) {data %>% mutate(across(everything(),
                                                                  ~ round(.x, digits)))}
corrs <- generate_correlation_table(grapes_woqs)
corrs <- round_table(corrs, 4)
```

```
rownames(corrs)<-c('Sugar Content', 'Acidity(PH)', 'Cluster Weight(g)', 'Berry Size(mm)',
                   'Sun Hours', 'Soil Moisture (%)', 'Rainfal(mm)', 'Harvest Week')
a<-kable(corrs[1:4], col.names=c('Sugar Content','Acidity(PH)', 'Cluster Weight(g)',
                                 'Berry Size(mm)'), booktabs = FALSE)|> kable_styling()|>
  column_spec(1, width = '8em')|> column_spec(c(2,3,4,5), width = '5em')
b<-kable(corrs[4:7], col.names = c('Sun Hours','Soil Moisture(%)', 'Rainfal(mm)',
                                   'harvest Week'), booktabs = FALSE)|>
kable_styling()|>
  column_spec(1, width = '8em')|> column_spec(c(2,3,4,5), width = '5em')
```

	Sugar Content	Acidity(PH)	Cluster Weight(g)	Berry Size(mm)
Sugar Content	1.0000	-0.0088	-0.0273	0.0142
Acidity(PH)	-0.0088	1.0000	-0.0081	-0.0307
Cluster Weight(g)	-0.0273	-0.0081	1.0000	0.0048
Berry Size(mm)	0.0142	-0.0307	0.0048	1.0000
Sun Hours	0.0500	0.0235	0.0394	-0.0845
Soil Moisture (%)	0.0421	0.0199	-0.0281	0.0271
Rainfal(mm)	0.0592	-0.0623	-0.1029	-0.0059
Harvest Week	-0.0685	-0.0035	-0.0340	-0.0238

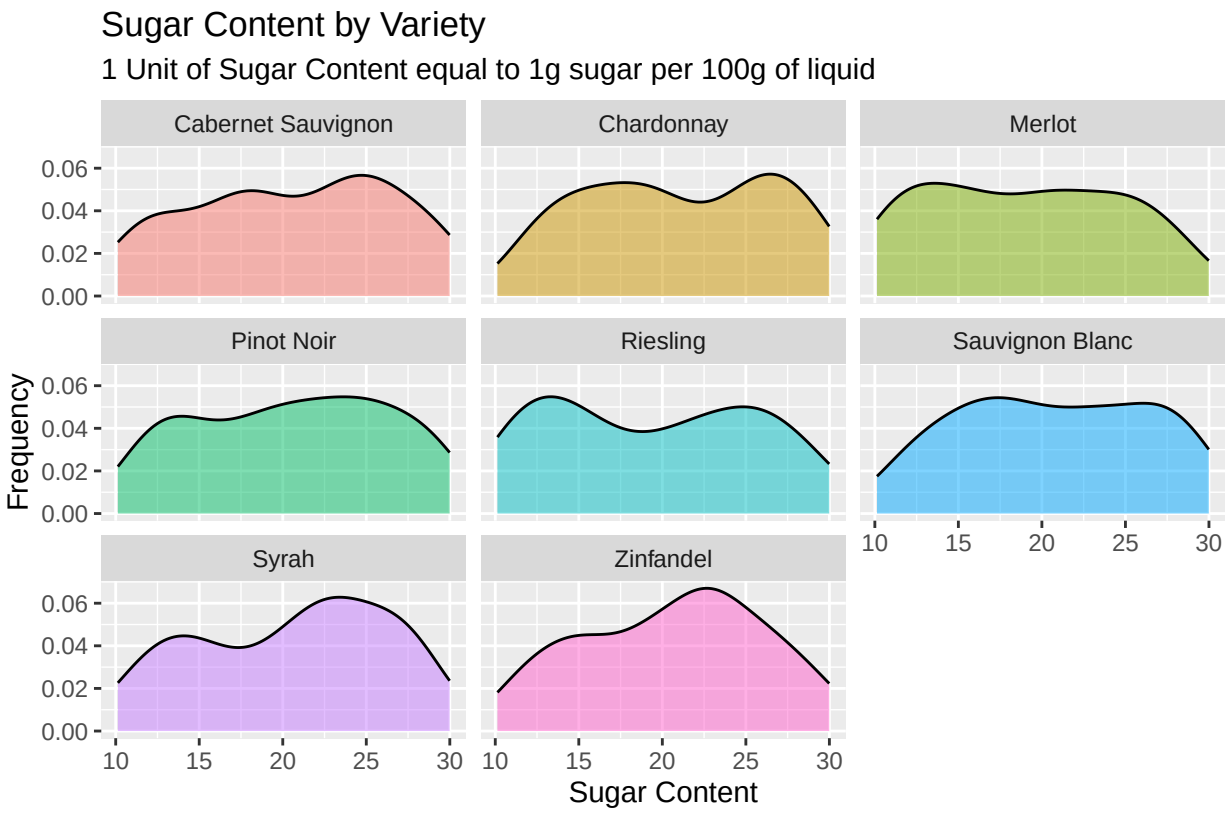
	Sun Hours	Soil Moisture(%)	Rainfal(mm)	harvest Week
Sugar Content	0.0142	0.0500	0.0421	0.0592
Acidity(PH)	-0.0307	0.0235	0.0199	-0.0623
Cluster Weight(g)	0.0048	0.0394	-0.0281	-0.1029
Berry Size(mm)	1.0000	-0.0845	0.0271	-0.0059
Sun Hours	-0.0845	1.0000	0.0502	-0.0019
Soil Moisture (%)	0.0271	0.0502	1.0000	0.0805
Rainfal(mm)	-0.0059	-0.0019	0.0805	1.0000
Harvest Week	-0.0238	-0.1044	-0.0150	0.0341

In the first step of analyzing the data, a table was created to effectively convey the correlation between every individual variable that does not play a role in dictating the quality of the grapes. The reasoning behind the use of correlation relates to the efficiency of determining the strength associated with the relationship between two variables. To build on this, we determine how much A (the target variable) changes concurrently when B (the feature variable) increases by 1, which is how correlation is defined. Conclusively, there is minimal correlation between the variables and this is revealed through the variables possessing a generally weak slope against each other.

Exploratory Plot 2: Sugar Content by Variety

```
ggplot(grapes_train)+
  geom_density(mapping = aes(x = sugar_content_brix, fill = variety), alpha = 0.5,
               show.legend = FALSE)+
  facet_wrap(~variety)+
  labs(y = 'Frequency',
```

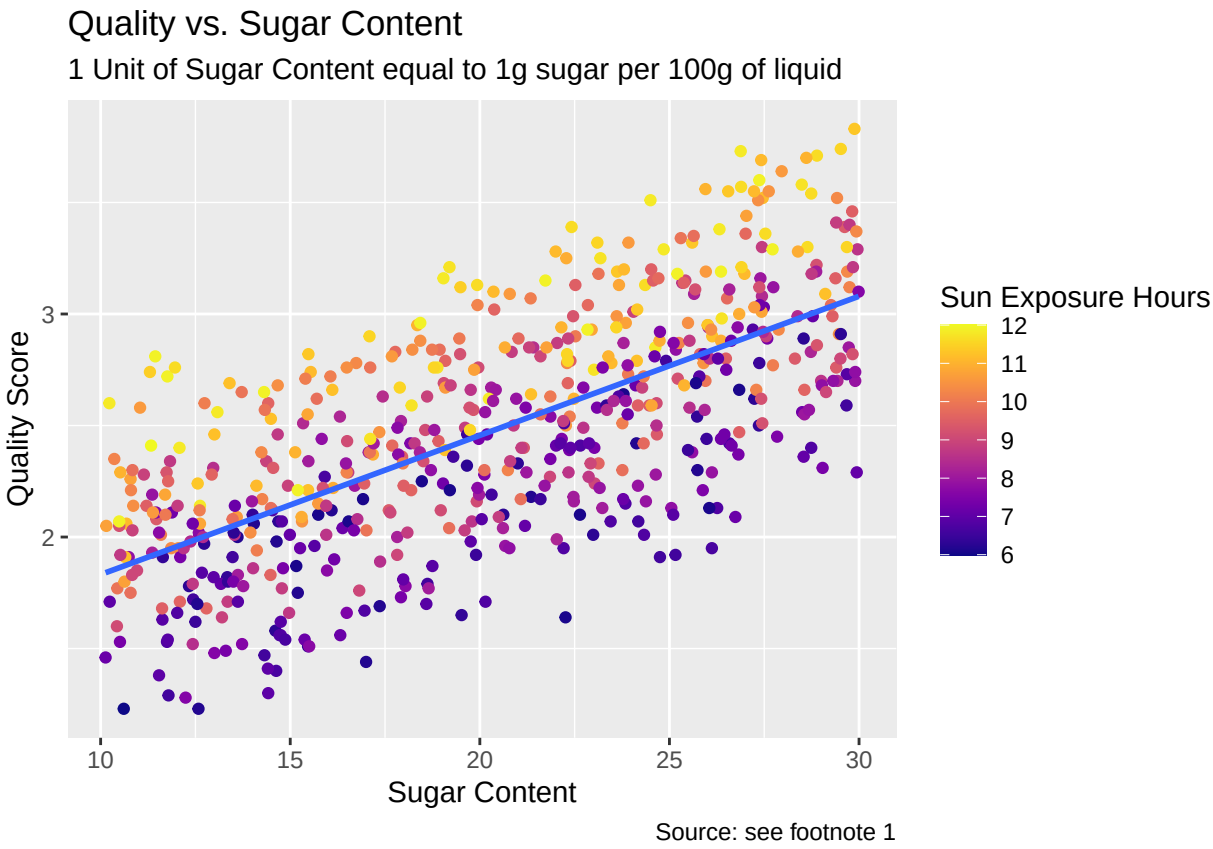
```
       x = 'Sugar Content',
       title = 'Sugar Content by Variety',
       subtitle = '1 Unit of Sugar Content equal to 1g sugar per 100g of liquid',
       caption = 'Source: see footnote 1')
```



In the second step of analyzing the data, the comparison was made between sugar content and the variety of grapes. Fundamentally, sugar is measured in brix, which is defined as one gram of sugar per one hundred gram of grape liquids. From the evident graphical representation of the frequency it can be said that the variety of Riesling, Syrah and Zinfandel tend to have a comparatively higher sugar content, while varieties such as Merlot and Sauvignon act inversely.

Model Plot 2: Quality vs. Sugar Content

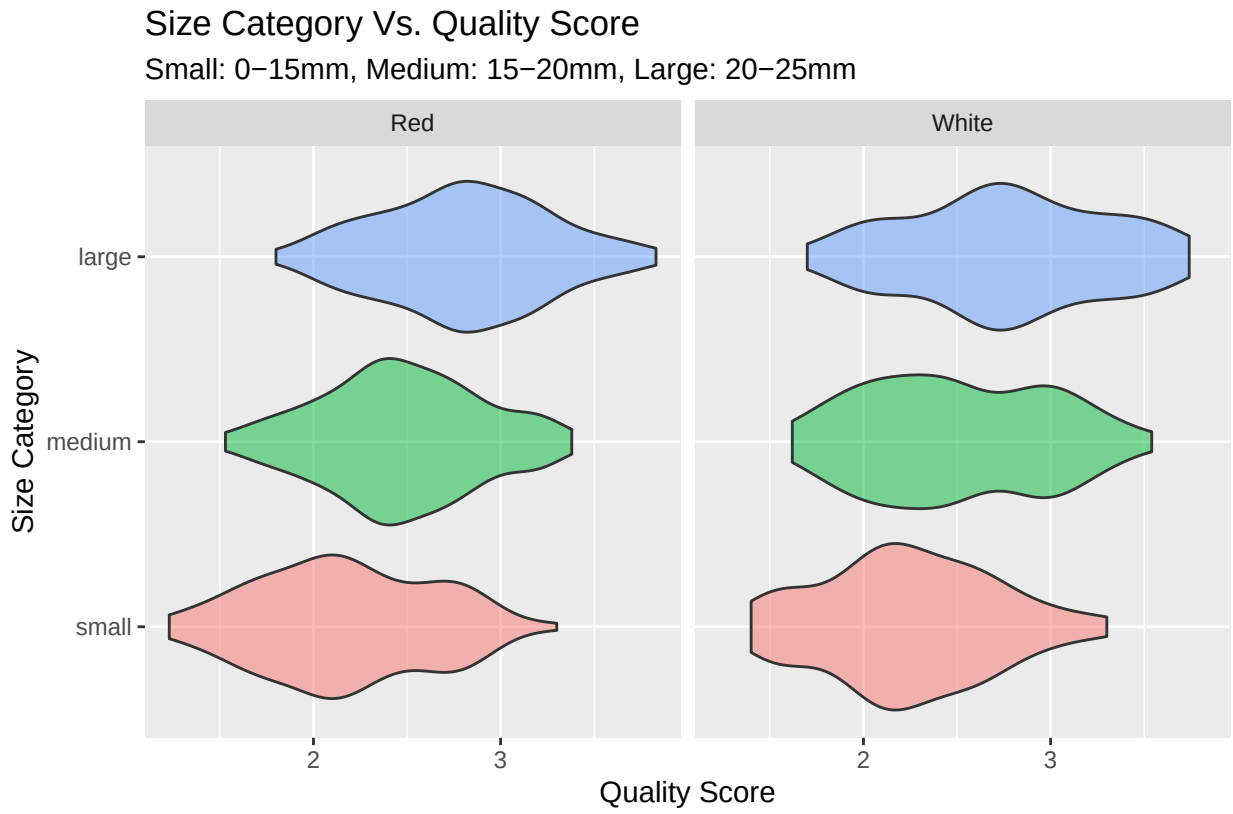
```
ggplot(grapes_train,mapping=aes(x = sugar_content_brix, y = quality_score))+
  geom_point(mapping = aes(colour = sun_exposure_hours))+
  geom_smooth(method = 'lm', formula = 'y~x', se = FALSE)+
  scale_color_viridis(option = "C") +
  labs(x = 'Sugar Content',
       y = 'Quality Score',
       title = 'Quality vs. Sugar Content',
       subtitle = '1 Unit of Sugar Content equal to 1g sugar per 100g of liquid',
       caption = 'Source: see footnote 1',
       colour = 'Sun Exposure Hours')
```



Conducting further analysis between sugar content, sun exposure and the target variable, that being sugar level. From the exploration of the previously mentioned relation, it is conclusive that as sugar content in a grape increases, the quality score simultaneously increases. Additionally, as sunlight hours increase in the cultivation of grapes, the quality score is also inclined to increase as well. Apart from this, sugar content and sunlight hours present little to no correlation between one another.

Model Plot 2: Size Category Vs. Wine Type

```
ggplot(grapes_train, aes(y = size_category, x = quality_score))+
  geom_violin(aes(fill = size_category), alpha = 0.5 ,show.legend=FALSE)+
  facet_wrap(~wine_type) +
  labs(title = 'Size Category Vs. Quality Score',
       subtitle = 'Small: 0-15mm, Medium: 15-20mm, Large: 20-25mm',
       x = 'Quality Score',
       y = 'Size Category',
       caption = 'Source: see footnote 1')
```



Source: see footnote 1

In our final graphical exploration of the data, a comparison is conducted between the sugar content against the size of the grapes and the type of wine that the grapes are inclined to produce. From the extraction of the data, we can derive that larger grapes tend to enhance the quality score of the grape, while red wine grapes have the tendency to be larger and higher in quality. On the contrary, white wine grapes are smaller on average, while remaining low in quality.

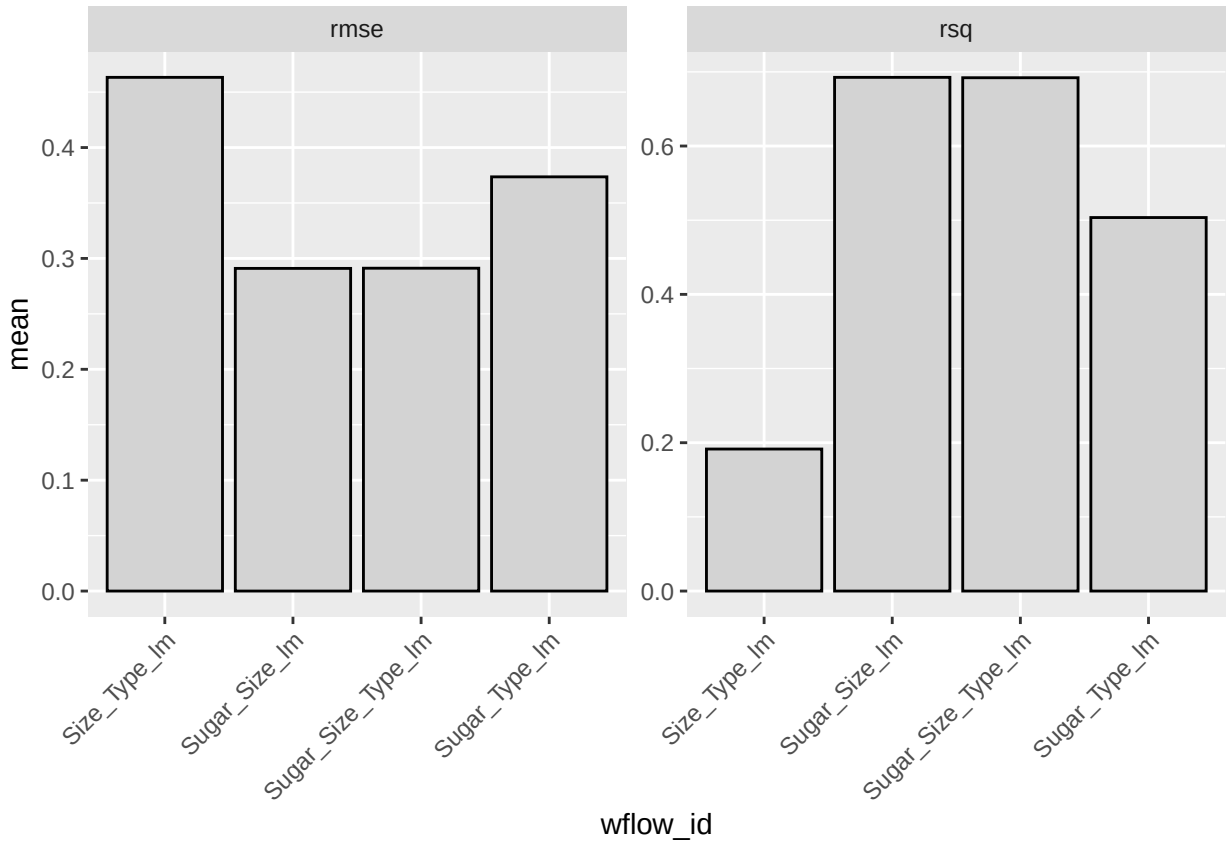
Exploratory Linear Model 1:

```
recipe1<-recipe(quality_score ~ sugar_content_brix + size_category + wine_type,
     data = grapes_train)|>step_dummy(wine_type)|>step_interact(~size_category)
recipe2<-recipe(quality_score ~ sugar_content_brix + size_category,
               data = grapes_train) |>step_interact(~size_category)
recipe3<-recipe(quality_score ~ sugar_content_brix + wine_type,
                  data = grapes_train) |>step_dummy(wine_type)
recipe4<-recipe(quality_score ~ size_category + wine_type,
  data = grapes_train)|>step_dummy(wine_type)|>step_interact(~size_category)
grapes_recipe <- list(Sugar_Size_Type = recipe1,Sugar_Size = recipe2,
                      Sugar_Type = recipe3,Size_Type = recipe4)
```

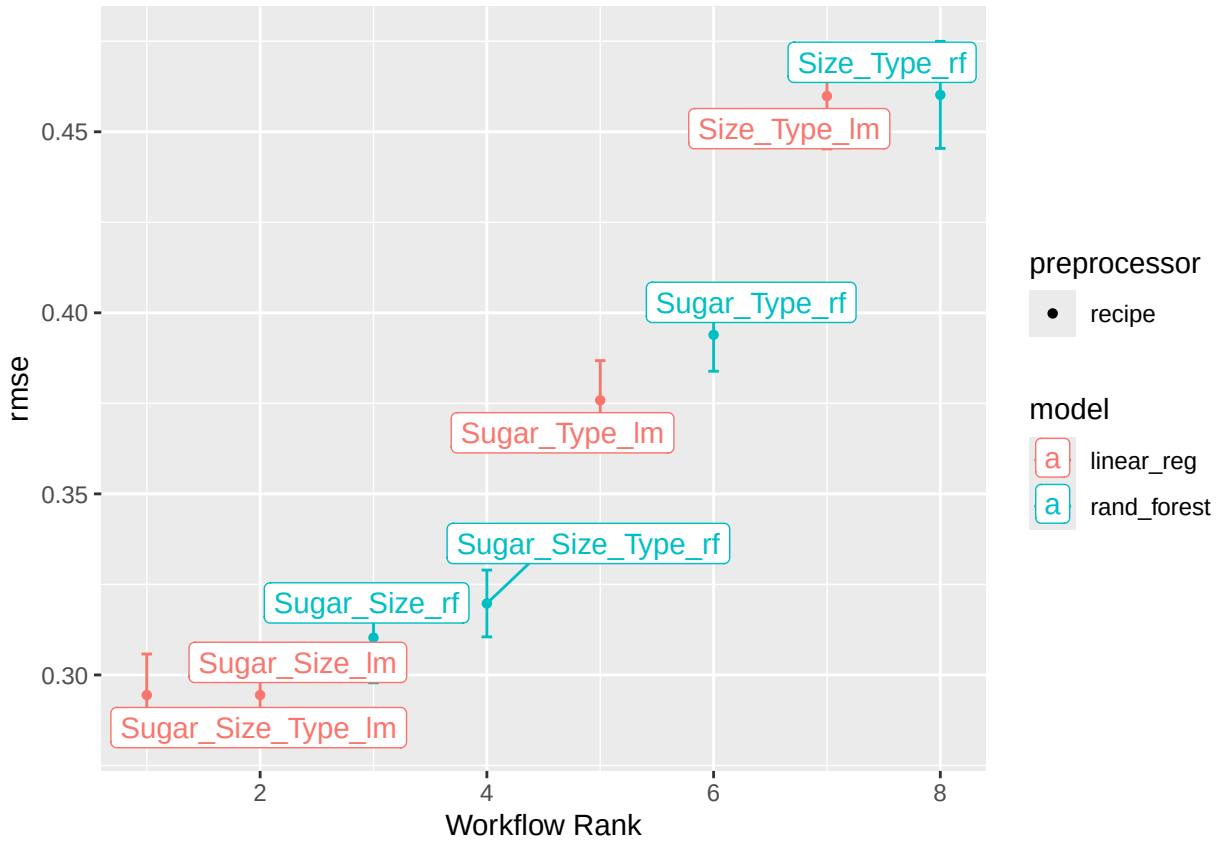
```
grapes_workflow_set <- workflow_set(grapes_recipe, models = list(lm = linear_reg()))
```

```
set.seed(1000)
grapes_fit <- workflow_set(grapes_recipe,
  models = list(lm = linear_reg())) |>
```

```
  workflow_map(fn = "tune_grid",resamples = validation_set(grapes_split))
grapes_fit |>collect_metrics()|>ggplot() +
    aes(x = wflow_id, y = mean) +
    geom_col(fill = "lightgrey", colour = "black") +
    facet_wrap(~ .metric, scales = "free")+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
train_test <- initial_split(grapes, prop = 0.8)
rf <- rand_forest(trees = 1000, mtry = tune(), min_n = 4) |>
    set_engine("randomForest") |>
    set_mode("regression")
tune_grid <- tibble(mtry = 1:4)
lm_versus_rf <- workflow_set(grapes_recipe,
  models = list(lm = linear_reg(), rf = rf)) |>
    workflow_map(fn = "tune_grid",grid = tune_grid,seed = 100,
        control = control_grid(save_pred = TRUE, save_workflow = TRUE),
        resamples = vfold_cv(training(grapes_split), v = 5))
lm_versus_rf |>autoplot(select_best = TRUE, metric = "rmse") +
    geom_label_repel(aes(label = wflow_id))
```



The initial predictive model uses the sugar content of the grapes, their respective size and the type of wine that the grapes are most likely to produce to determine the quality of the grapes. In the analysis, four variations of the features were used to identify which combination had the strongest effect on the overall quality of the grapes. From this, it can be determined that using all three features creates the best overall model, but the size of the grapes has a much greater influence on quality compared to the wine type.

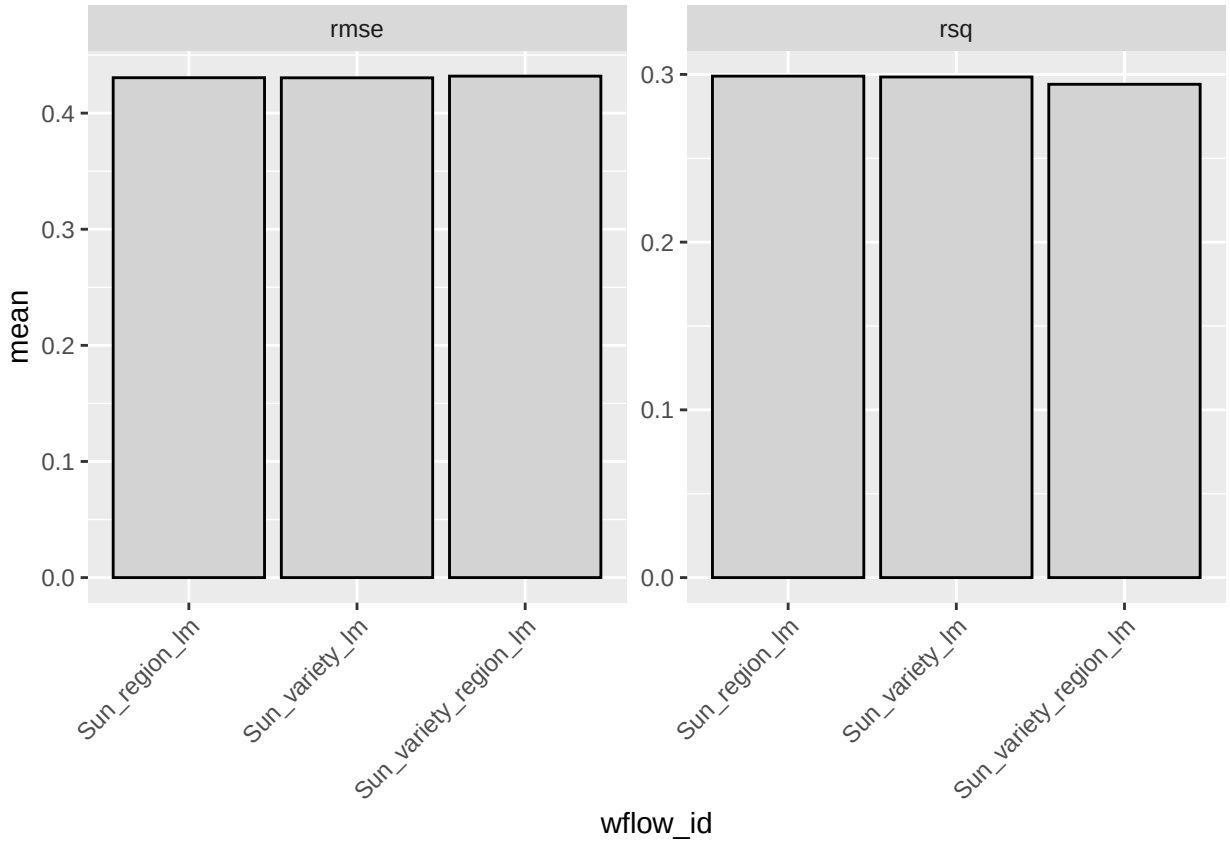
Exploratory Linear Model 2:

```
recipe5<-recipe(quality_score ~ sun_exposure_hours + region + variety,
                data = grapes_train)|>step_interact(~region) |> step_interact(~variety)
recipe6<-recipe(quality_score ~ sun_exposure_hours + region,
                data = grapes_train) |> step_interact(~region)
recipe7<-recipe(quality_score ~ sun_exposure_hours + variety,
                data = grapes_train) |>step_interact(~variety)
grapes_recipe_2 <- list(Sun_variety_region = recipe5,
                        Sun_region = recipe6, Sun_variety = recipe7)
```

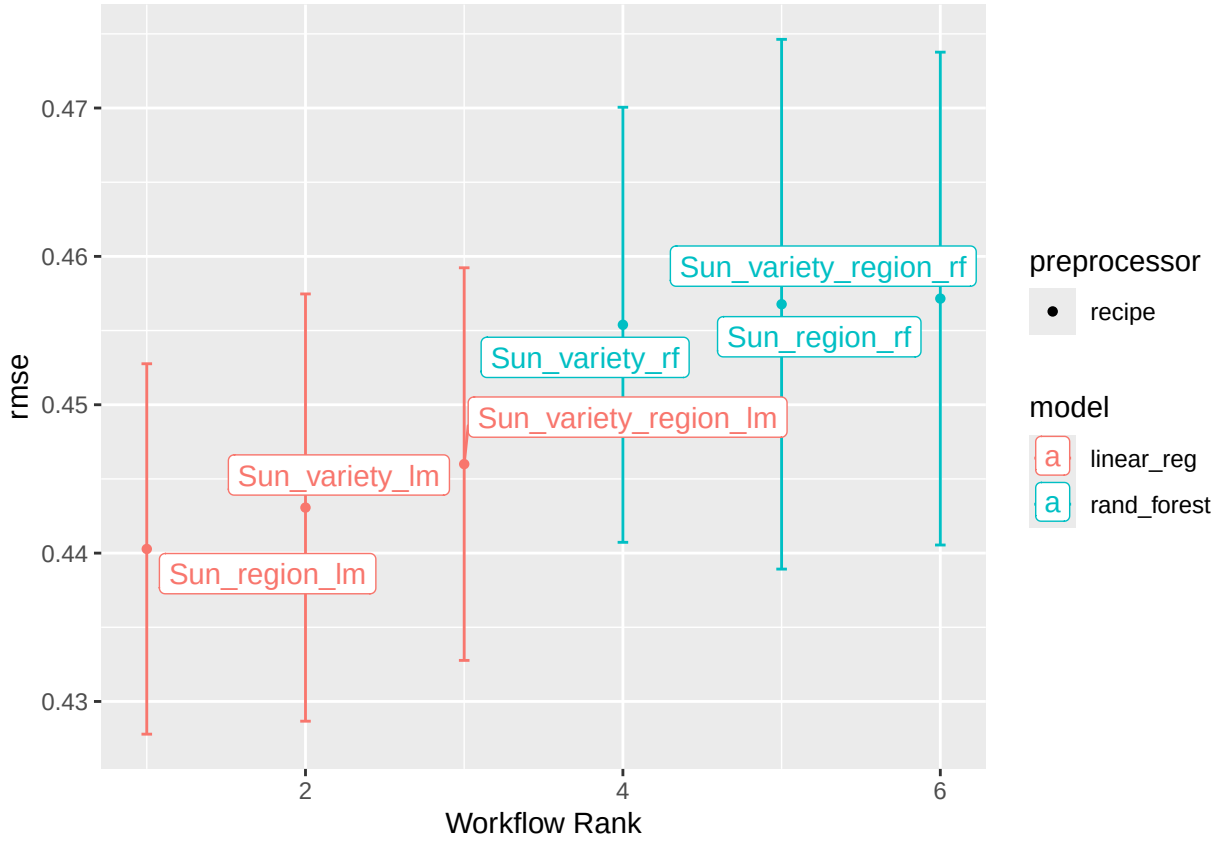
```
grapes_workflow_set_2 <- workflow_set(grapes_recipe_2, models = list(lm = linear_reg()))
```

```
set.seed(1000)
grapes_fit_2 <- workflow_set(grapes_recipe_2,
  models = list(lm = linear_reg())) |>
  workflow_map(fn = "tune_grid",resamples = validation_set(grapes_split))
grapes_fit_2 |>collect_metrics()|>ggplot() +aes(x = wflow_id, y = mean) +
```

```
    geom_col(fill = "lightgrey", colour = "black") +
    facet_wrap(~ .metric, scales = "free")+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
train_test <- initial_split(grapes, prop = 0.8)
rf_2 <- rand_forest(trees = 1000, mtry = tune(), min_n = 3) |>
    set_engine("randomForest") |> set_mode("regression")
tune_grid_2 <- tibble(mtry = 1:3)
lm_versus_rf_2 <- workflow_set(grapes_recipe_2,
  models = list(lm = linear_reg(), rf = rf)) |>
    workflow_map(fn = "tune_grid",grid = tune_grid_2,seed = 100,
        control = control_grid(save_pred = TRUE, save_workflow = TRUE),
        resamples = vfold_cv(training(grapes_split), v = 5))
lm_versus_rf_2 |>
    autoplot(select_best = TRUE, metric = "rmse") +
    geom_label_repel(aes(label = wflow_id))
```



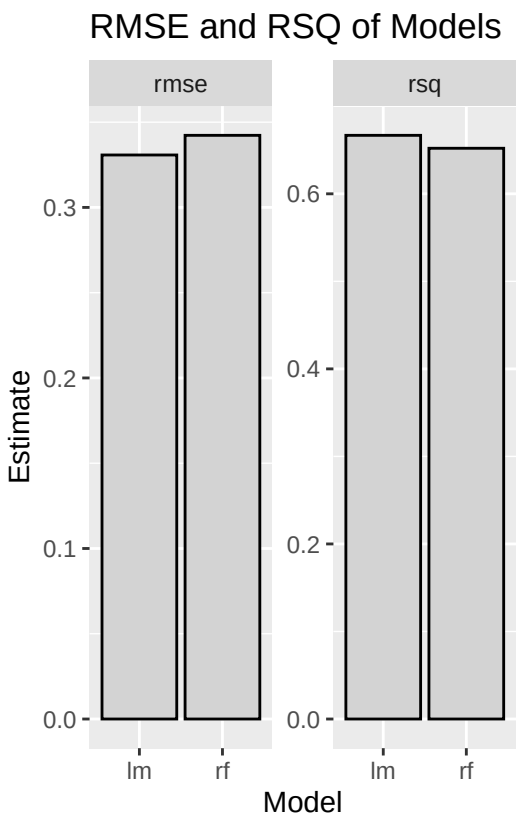
The secondary predictive model, instead uses the amount of sunlight the grapes receive, their respective region and the specific variety to determine the quality of the grapes. In the analysis, three variations of the features were used to identify which combination of respective region and variety had the strongest effect on the overall quality of the grapes. From the secondary analysis, it can be determined that using the amount of sun exposure and variety creates the best overall model, indicating the region of grapes is not necessary in predicting the quality of the grapes.

Final Model:

```
best_lm <- lm_versus_rf |>extract_workflow("Sugar_Size_lm")
best_rf <- lm_versus_rf |>extract_workflow("Sugar_Size_rf")
```

```
sslm<-lm_versus_rf|>extract_workflow_set_result('Sugar_Size_lm')|>
  select_best(metric='rmse')
ssrf<-lm_versus_rf|>extract_workflow_set_result('Sugar_Size_rf')|>
  select_best(metric='rmse')
test_lm<-lm_versus_rf|>
  extract_workflow('Sugar_Size_lm')|>
  finalize_workflow(sslm)|>
  last_fit(split = grapes_split)
test_rf<-lm_versus_rf|>
  extract_workflow('Sugar_Size_rf')|>
  finalize_workflow(ssrf)|>
```

```
  last_fit(split = grapes_split)
point<-bind_rows(
    lm = collect_predictions(test_lm),
    rf = collect_predictions(test_rf),
    .id = "model") |>
    ggplot() +
    aes(x = quality_score, y = .pred, colour = model) +
    geom_point(shape = 1)+
  labs(title = 'Quality Score Model Predictions',
    y = 'Predicted Quality',
    x = 'Actual Quality')
bar<-bind_rows(
    lm = test_lm,
    rf = test_rf,
    .id = "model") |>
    unnest_wider(.metrics) |>
    unnest_longer(c(.metric, .estimator, .estimate, .config)) |>
    ggplot() +
    aes(x = model, y = .estimate) +
    geom_col(fill = "lightgrey", color = 1) +
    facet_wrap(~ .metric, scales = "free_y")+
    labs(title = 'RMSE and RSQ of Models',
         y = 'Estimate',
         x = 'Model')
bar + point
```



From the analysis preformed previously, the quality of grapes has a strong correlation to the sugar amount and the size of the grape. This can be seen in the rank of the model, the large RSME value and the large RSQ value.

Conclusions

Conclusively, this study proficiently presents the influence of grape cultivation measurements on the quality of grapes harvested from the heart of Italy. In the exploratory portion of the analysis, it was discovered that there exists a plausible link between sugar content, as well as variety and size, to the quality pertaining to the grapes. Furthermore, this correlation was expanded upon through the utilization of predictive models, the correlation metric, and the RMSE and RSQ of models. Building on this, the greatest correlation obtained to the target variable was using the sugar content and size of grapes, identifiable in the final model. This can be logically perceived as these grapes are primarily used in the production of wine, in which comparatively sweeter wines can be extracted in great quantity from grapes that are larger in size and possess higher levels of sweetness.

The primary limitation corresponding to this data is associated with the noticeably short time periods within the data set from which the grapes were collected. Moreover, measurements such as pH, soil moisture, and rainfall were ineffectual as they possessed minimal correlation with one another and failed to contribute to the quality of wine grapes. Ultimately, this analysis offers valuable insights into the influential relationships that lie within the cultivation of grapes.

[1]

[2]

[1]Vishal. (2024). Grape Quality [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/9781093
[2]OpenAI. (2024). ChatGPT (Dec 4 version) [Large language model]. https://chat.openai.com/