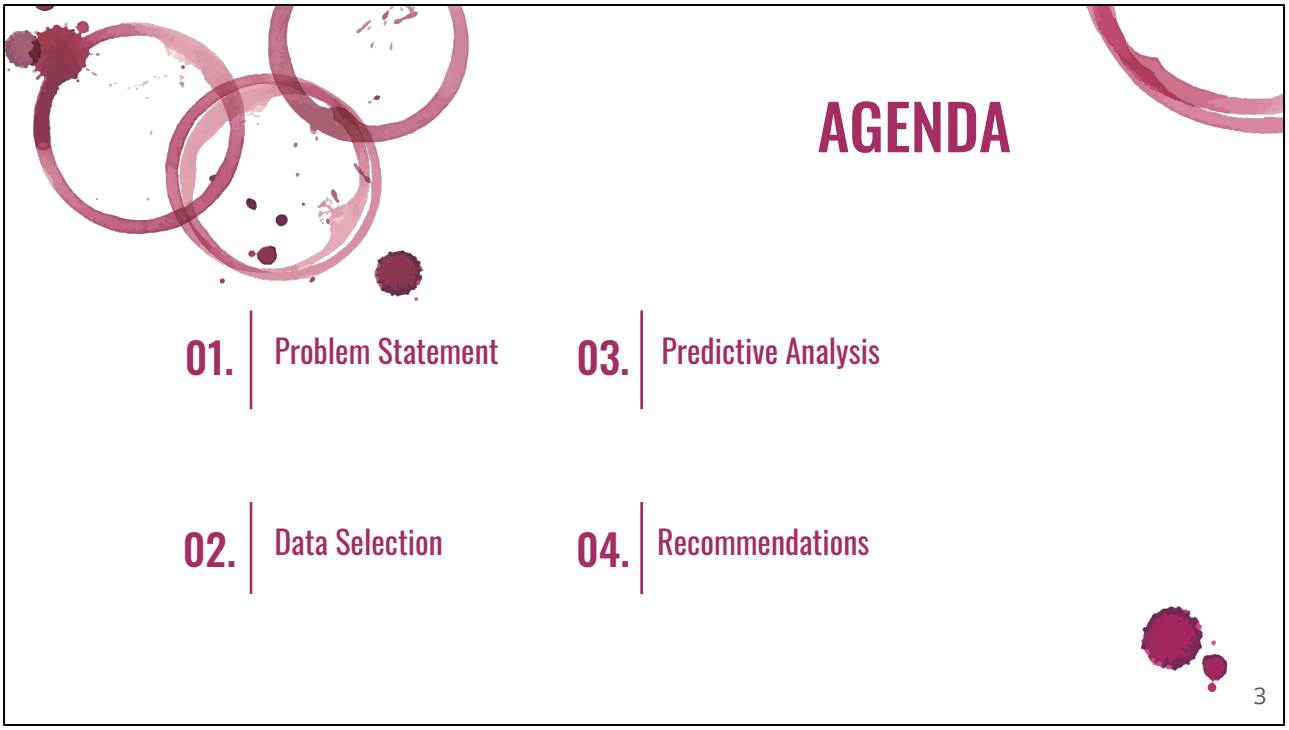
Basic intro:

Good evening, we


Asad Mahmood
Zhonghan Deng
Irene Ramirez
Jared Halterman
Trang Vu
Saba Sajadi
Laura Bathurst
Moxshil Shah
UNIVERSITY OF THE PACIFIC
1851



Today we will be covering our collaboration with Gallo winery and how our efforts ultimately resulted in the expected output.

The four main concepts covered will be:

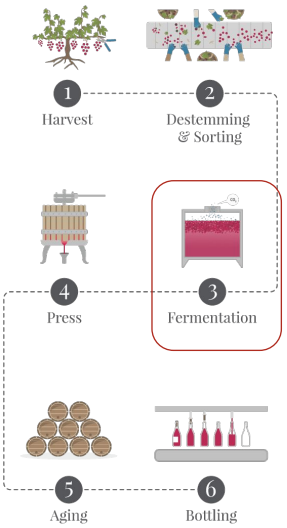
1. Problem Statement
    a. Brief description of the project objective
2. Data Selection
    a. The data extraction and data cleaning process
3. Modeling
    a. What were the different modeling scenarios we ran through and which model did we end up choosing
4. Issues
    a. What were some of the IT issues we had and the limitations we faced
5. Recommendations
    a. What are the results we got and the action items/ next steps. Additional, what are some improvements that can be made.

01. Problem Statement

Project Objective

Problem Statement

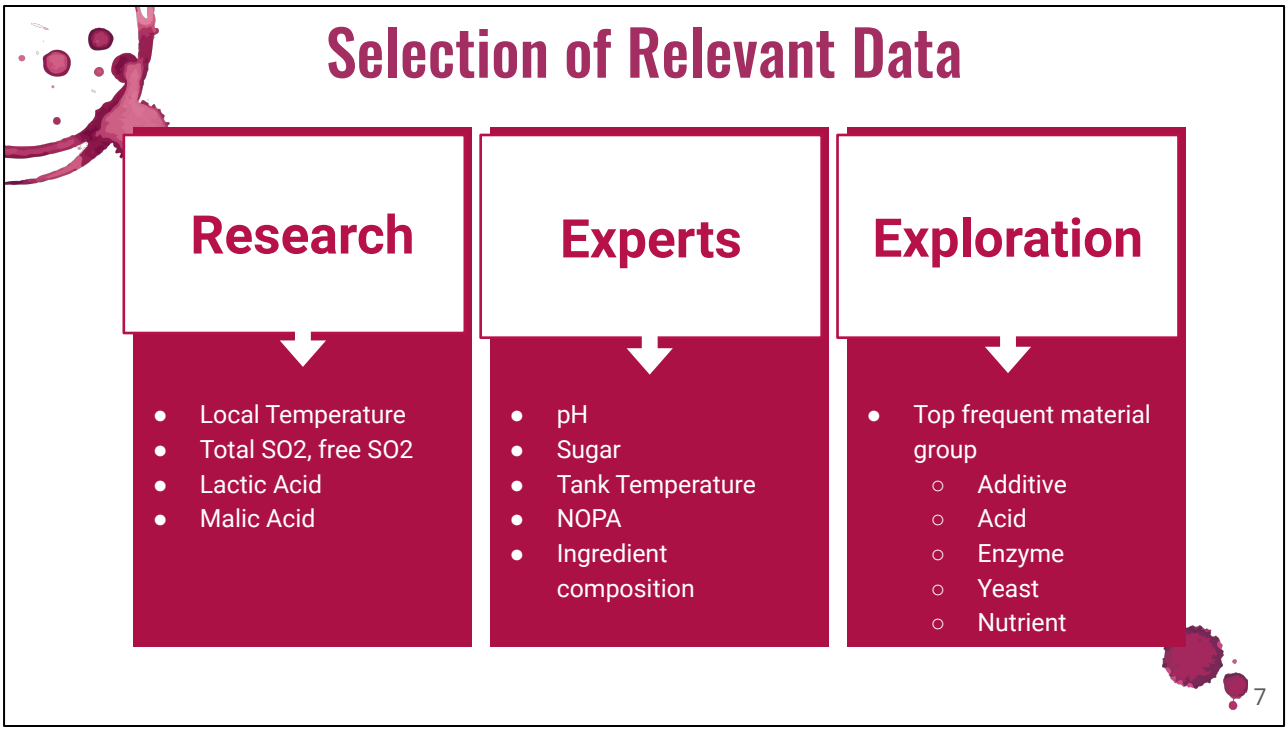
Determining when fermentation is completed and is ready to be bottled based on the type of wine (red or white) and the location of the tank.



The goal of our project is to determine when fermentation is complete and when wine is ready to be bottled. The time to completion would depend on the type of wine and location of the tank.The benefit of this knowledge would allow Gallo to maximize their production and tank turn over time and minimize the labor of physical wine testing. Keep in mind that this was our goal throughout the process of developing the modeling but as you will see later on, this goal was not fully met due to some limitations.

02. Data Selection

Data Extraction and Cleanup

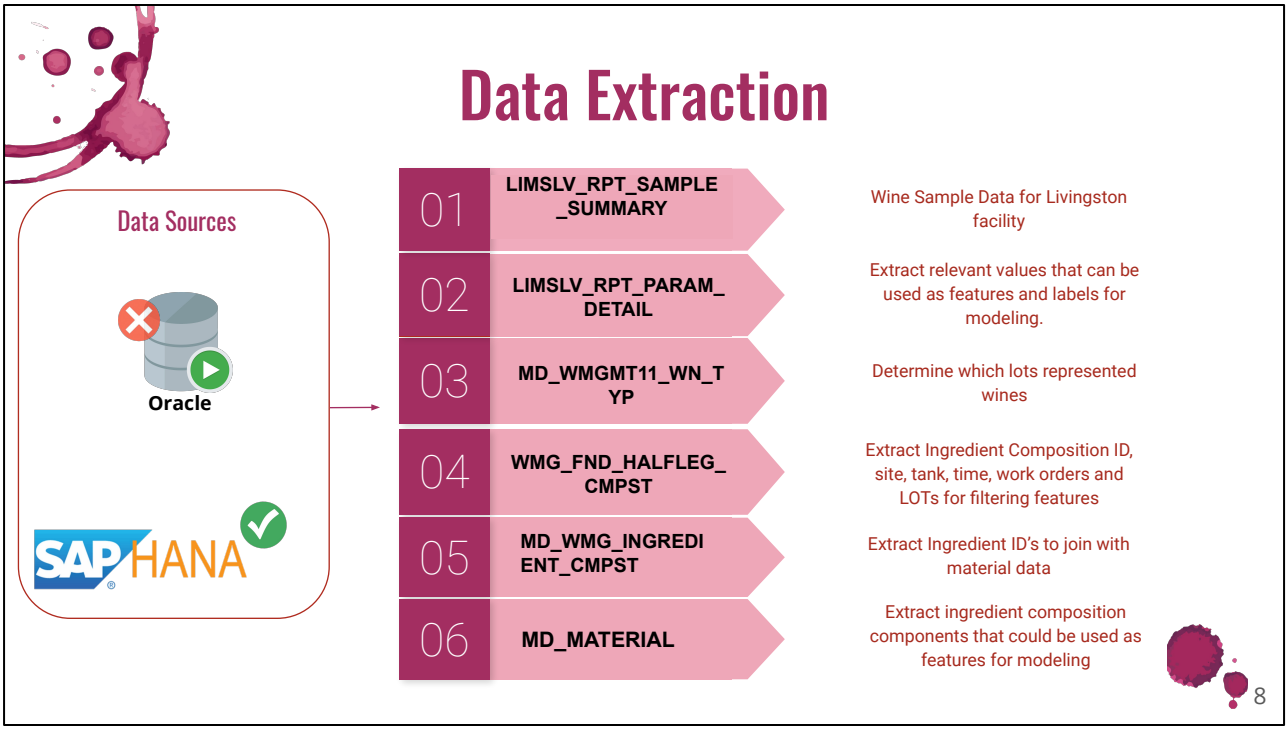


We acquired knowledge and select relevant data by doing personal research (using research papers about fermentation process and a Gallo sales manual from 2002), meetings with Gallo's experts, and data exploration. Our team identified available key components in the provided databases that impacts the fermentation process such as SO2 (sulfur dioxide), lactic acid and malic acid. We also dictated temperature was highly correlated with wine fermentation time. Due to the inconsistency of the internal tank temperature data, we pulled local temperature data in hopes of finding significance with each batch.

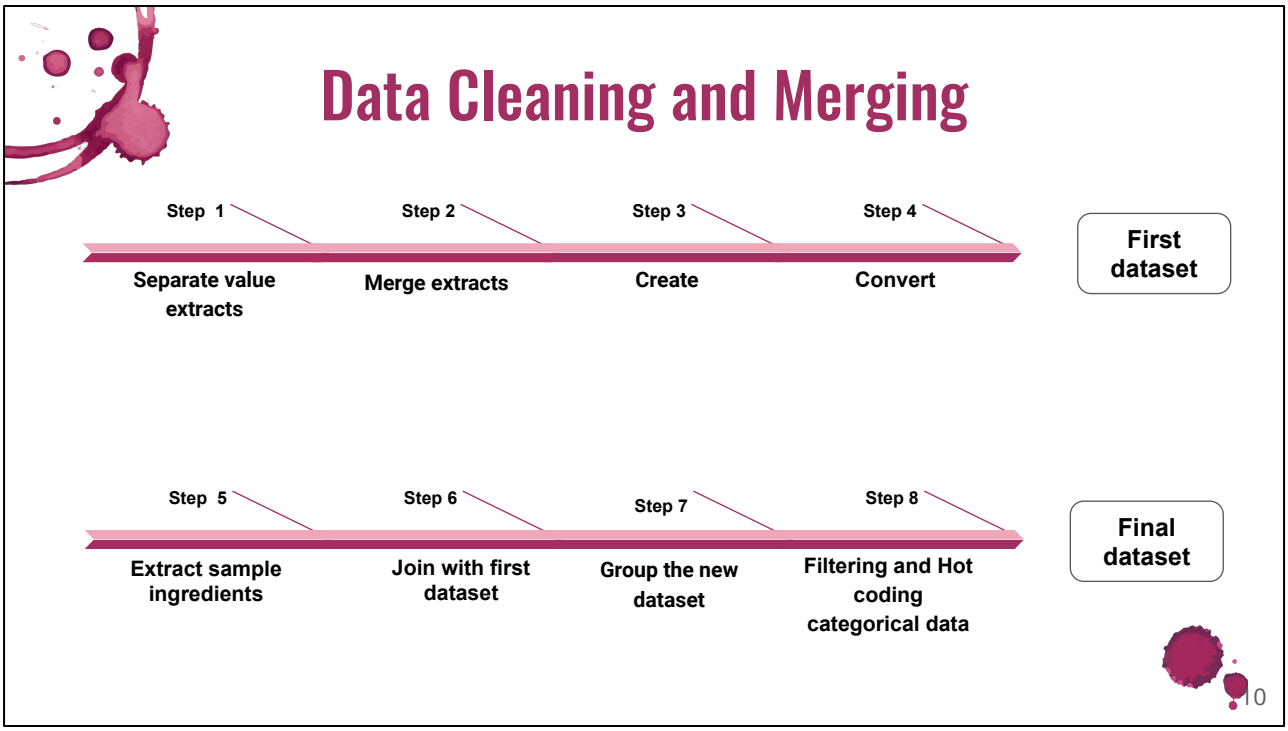
Gallo employees supplied us with the basic knowledge of fermentation including the importance of ph values, glucose fructose, NOPA and tank temperature and how the data sample was collected and logged into LIMS with one of the fermentation process state and different measurements. They also assisted us to specify the relationship among tables and how we could link the tables together to achieve the data that we need. The query formula for ingredient composition is also clarified.

For exploration, we use python to go over the material dataset and identify the top frequency of ingredient group added in the latest work orders before sample taken as follows : additive, acid, enzyme, yeast and nutrient.

Notes: Malic Acid provides a strong link to wines tasting 'flat' if there is not enough while SO2 is a chemical compound used by winemakers to help keep their wine protected from the negative effects of oxygen exposure. Temperature plays a huge role that doesn't live within lab data. Tank has gauges and data is sent over to the wine manager tank. Wine makers don't trust the temps in sample table, not a good representation of actual temperature in the tank.



For the data sources offered, we have SQL developer setup for us to extract data.Due to the slowness and data extraction timeouts that happened a lot in SQL developer, we move forward with HANA where we can successfully extract data from. On the left side, these are list of tables we used to extract data. The first one is the wine sample data for livingston facility, the second one is used to extract relevant features and labels for modeling. The third one is used to determine which lots represent which wines. The fourth one is used to extract composition ID information along with tank, work orders and LOTs while the fifth is to extract Ingredient ID. We need these two data tables since they are required for us to retrieve the data of interest in the 6th table MD_Material where contains ingredient additions that could be used as features for modeling.



For data cleaning and merging, our first step was to get **separate extracts** of glucose, NOPA, ammonia, PH values, total So2 (sulfur dioxide), free So2, Lactic acid and Malic acid in csv file format from HANA. Then we **merged** the extracts into one file (using python) and **created** a wine type column based on the LOT column. After that, we **converted** all p_value columns from string(text) to float(numerical digits).

We then **created an Ingredient composition query** to filter important columns for the ingredients composition as stated in the earlier state then **join** with the first datasets (glucose/fructose, NOPA, ammonia, and PH values) using the common columns EQUIPMENT_ID and LOT_I. Lastly, we **group** the new joined dataset by TNK_I, LOT_I, S_CREATED_DATE, CMPLT_T and filter only rows that have latest **completed time** of work orders that happens before sampling time of specific tank and lot.

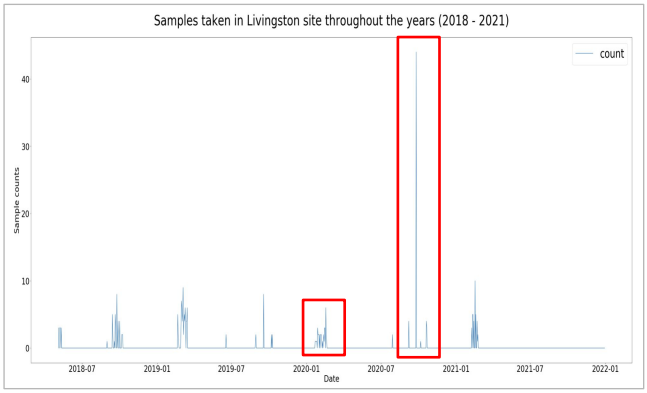
Reference BELOW

Notes: In ingredient data, we have tank, each tank has many lot, each lot has many work-orders, and each work-order is many adjustment of ingredient component completed at certain time. We need to link the ingredient table with our first data by tank and LOT, and also need to compare the **completed date** (in **ingredient data**) with **S_created_date (date of sampling in our data)** to filter only the latest work order right before sample was taken.

Step 1 - glucose/fructose, NOPA, ammonia, and PH values, Total So2, Free So2, Lactic acid, Malic acid
Step 2 - Merge all csv files together
Step 3 - Create wine_type column based on the LOT column
Step 4 - Convert all p_value columns from string to float

Step 5 - Extract ingredient data
Step 6 - Using common columns Equipment_ID and LOT_I
Step 7 - By Tank ID, LOT ID, Sample Created Date, and Complete date.
Step 8 - Filter By latest Completed work orders time before Sampling time and by the top frequency ingredients. Hot code categorical columns

Data Overview

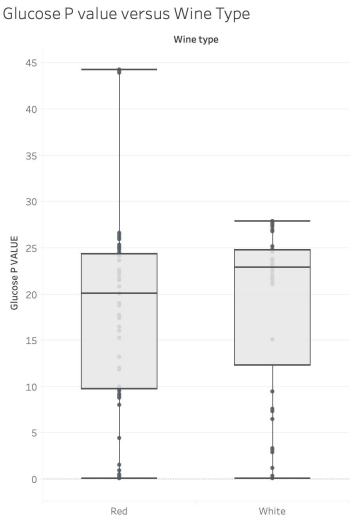


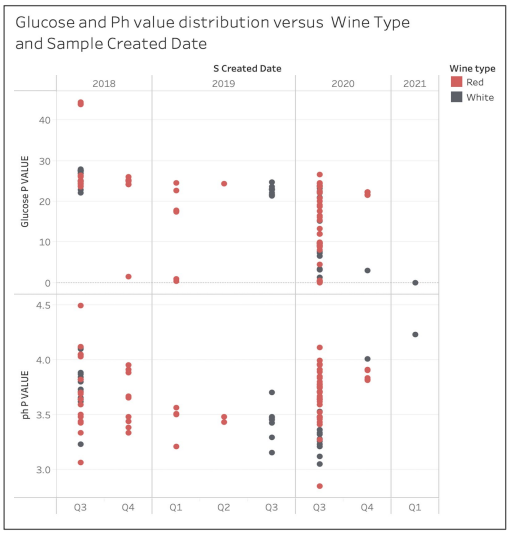

```
RangeIndex: 296 entries, 0 to 295
Data columns (total 36 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Ammonia_P_VALUE                                 296 non-null    float64
 1   Glucose_P_VALUE                                 296 non-null    float64
 2   NOPA_P_VALUE                                    296 non-null    float64
 3   ph_P_VALUE                                      296 non-null    float64
 4   Wine_number                                     296 non-null    int64
 5   Free SO2_P_VALUE                                296 non-null    float64
 6   Lactic Acid_P_VALUE                             296 non-null    float64
 7   Malic Acid_P_VALUE                              296 non-null    float64
 8   Total SO2_P_VALUE                               296 non-null    float64
 9   MaterialName_ACID-TARTARIC ACID 55LB BAG        296 non-null    float64
 10  MaterialName_ACID-TARTARIC ACID SUPERSACKS      296 non-null    float64
 11  MaterialName_ACID-TARTARIC SUPERSACK 2204.6 BULK BAG  296 non-null  float64
 12  MaterialName_ADDITIVE-COPPER SULFATE RGNT 100LB DRUM  296 non-null  float64
 13  MaterialName_ADDITIVE-POT METABI 14.7% LIQ      296 non-null    float64
 14  MaterialName_ADDITIVE-POTASSIUM METABISULF 99% 55LB  296 non-null  float64
 15  MaterialName_ADDITIVE-SULFUR DIOXIDE BULK       296 non-null    float64
 16  MaterialName_CARBON-NUCHAR HD MAX BULK MEADWESTVACO  296 non-null  float64
 17  MaterialName_ENZYME-PECTINEX XXL 25KG           296 non-null    float64
 18  MaterialName_ENZYME-ROHALASE BXL AB ENZYMES     296 non-null    float64
 19  MaterialName_ENZYME-ROHAPECT VR-L 25KG          296 non-null    float64
 20  MaterialName_ENZYME-ROHAVIN L 25KG AB ENZYMES   296 non-null    float64
 21  MaterialName_ENZYME-ROHAVIN MX 25 KG AB ENZYMES  296 non-null  float64
 22  MaterialName_ENZYME-SUMIZYME MHT                296 non-null    float64
 23  MaterialName_FINING AID-PORK GELATIN 175 PS 30  296 non-null    float64
 24  MaterialName_FINING AID-SILICA GEL 2,700 LB  TOTE  296 non-null  float64
 25  MaterialName_FINING AID-U.S. GELATIN 55 LB      296 non-null    float64
 26  MaterialName_FINING AID-VITIBEN BENTONITE       296 non-null    float64
 27  MaterialName_NUTRIENT-DIAMMONIUM PHOSPHATE 55LB  296 non-null  float64
 28  MaterialName_YEAST-RAVAGO MAURIVIN ELEGANCE 10KG  296 non-null  float64
 29  MaterialGroupName_Acid                          296 non-null    float64
 30  MaterialGroupName_Additive                      296 non-null    float64
 31  MaterialGroupName_Carbon                        296 non-null    float64
 32  MaterialGroupName_Enzyme                        296 non-null    float64
 33  MaterialGroupName_Fining Aid                    296 non-null    float64
 34  MaterialGroupName_Nutrient                      296 non-null    float64
 35  MaterialGroupName_Yeast                         296 non-null    float64
```


After merging all data collected from HANA, we got our final dataset with **36 columns and 296 rows**. An overview of the final dataset is on the right side. The graph on the left side is basically showing the count of samples taken in Livingston throughout the year **2018 to 2021** based on Sample Created Date column.The data sample were also taken inconsistency. For example, 2020 witnessed a high peak in term of the amount of samples taken whereas the others period witnessed much smaller amount of samples counts. There is a contradiction between the high amount of features extracted and the lack of sufficient sample taken as data points.

- **2021**: Only February samples found
- **2020:** January - February, July, September - October samples found.
- **2019:** February - March, June, August, September - October samples found.
- **2018:** May, August, September - October samples found.

Data Exploration





- We make some data exploration to understand the characteristics and distribution of data points.
- On the left hand side, this is the distribution of Glucose P value versus wine type. We could see that the distribution of glucose P value in Red Wine is in wider range than white ones. The median of glucose P-value in white wine is higher than red ones. That is reasonable since white wine often contains more sugar than red wine and once the glucose value is down to certain low values then the fermentation is reached. It often takes longer for red wine fermentation to be reached since red wine is often dry and less sweet than white.
- Second plot shows us the distribution of two measurement ph values and glucose versus wine type throughout different quarters when samples were taken. We could see the wider distribution of red wine than white wine data throughout the plot. In 2018 Q4, 2019 Q1 and Q2, we only see the sample of red wines. In Q3 of 2018 and 2020, there are evenly and dense distribution of red and white. We could also see some quarter with only white wines as samples taken.
- Reason for exploration: seasonal aspect of data, understand more about how wine theory correlates with the way data is distributed, level of glucose values and ph value vary by the period that samples were taken. (wider range in third quarter each year).

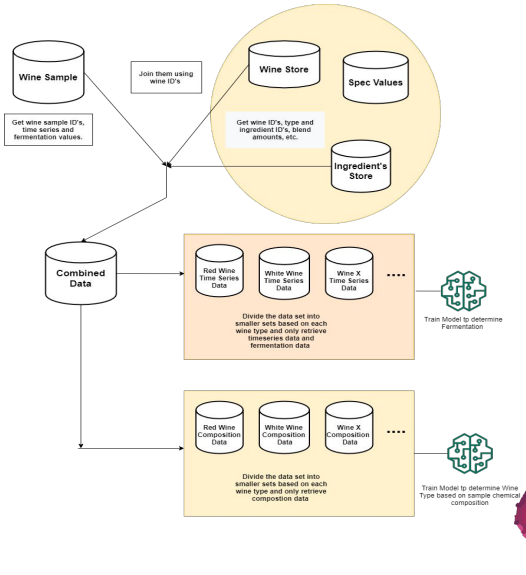
Note (In case of question): We want to see throughout quarter, how data distributed in the perspective of wine types and measurement so we could have a sense of how data perform in different timeline, at the same time, see how important of the consistent data sample affects the comprehensiveness of data and the accurate of the prediction model in the later stage. When enough samples are gathered then time series forecasting could be used to predict fermentation trend and create an accurate production level.
Glucose gets down to **0.2 RS** (dry whites), **0 .6 RS** (sweet wines)

03. Predictive Analysis

Modeling Scenarios and Results

Modeling Strategy

- Initial strategy was to do a time series analysis and forecasting using deep learning techniques.
- Over time, we had to shift to regression analysis because of data on hand.
- We have trained the data on thirty plus different regression algorithms to pick which one works the best.



Initially, (Figure on the right) we planned on designing an app that would predict the fermentation stage of a certain wine type whenever its sample is taken out in the field using two step process. The first step will using a model trained on wine's chemical composition data, which will determine the wine type. The second step will be using the previous prediction to choose that specific wine's fermentation prediction model. This model will be trained on fermentation data procured earlier from samples

However, as the project progressed we saw that there was less data for us to train a time series forecasting model around or employ any of the advanced machine learning techniques. That is why we chose to go with regression analysis for now. We trained it over thirty plus different regression algorithms. The primary reason was to check which algorithm gives us the best results and then spend time on fine tuning its parameters at a later stage.

Prediction Accuracy Measures

- The **Adjusted R-squared** increases only if the new predictor improves the model more than would be expected by chance.
  - The higher the adjusted R-squared, the new predictors add more value.

- The **Root-Mean-Square Error** (RMSE) is a measure of the differences between values predicted by a model and the values observed.
  - The smaller the better.

But before we go through our predictive analysis I want to walk you through the prediction accuracy measures we chose

Adjusted R-Squared increases only if the new predictors improves the model more than would be expected by chance. higher value indicates that the additional input variables are adding value to the model.
Generally we will use adjusted R-squared, because if you add more useless variables into the model, adjusted r-squared will decrease.
RMSE - is a measure of the differences between values predicted by a model and the values observed.

---- Don't say
R - Squared the percentage of the response variable variation that is explained by a linear model. higher the better btw the range 0 to 1

Predictive Analysis (First Dataset)

- This dataset had three features
  - Ammonia
  - PH
  - NOPA
- It had one label i.e. Glucose/Fructose.
- ExtraTreesRegressor is the best model
  - **Adjusted R-Squared:** 0.73
  - **RMSE:** 4.97

Model	Adjusted R-Squared	R-Squared	RMSE
ExtraTreesRegressor	0.73	0.75	4.97
XGBRegressor	0.69	0.71	5.35
RandomForestRegressor	0.66	0.68	5.58
KNeighborsRegressor	0.63	0.66	5.82
AdaBoostRegressor	0.62	0.64	5.92
DecisionTreeRegressor	0.62	0.64	5.95
BaggingRegressor	0.61	0.64	5.98
GradientBoostingRegressor	0.56	0.59	6.35
HistGradientBoostingRegressor	0.54	0.57	6.53
LGBMRegressor	0.51	0.55	6.68
SVR	0.36	0.41	7.64
NuSVR	0.33	0.38	7.82
Lars	0.31	0.36	7.96
TransformedTargetRegressor	0.31	0.36	7.96
LinearRegression	0.31	0.36	7.96
Ridge	0.31	0.36	7.96
SGDRegressor	0.31	0.36	7.96
BayesianRidge	0.31	0.36	7.97
OrthogonalMatchingPursuitCV	0.31	0.36	7.97
RidgeCV	0.31	0.35	7.98
ElasticNetCV	0.30	0.34	8.04
LassoLarsIC	0.29	0.34	8.05
LassoCV	0.29	0.34	8.06
LassoLarsCV	0.29	0.33	8.11
LarsCV	0.29	0.33	8.11
PoissonRegressor	0.27	0.32	8.17
LinearSVR	0.25	0.30	8.31
Lasso	0.25	0.30	8.33
ExtraTreeRegressor	0.24	0.29	8.34
HuberRegressor	0.22	0.27	8.47
GammaRegressor	0.22	0.27	8.49
ElasticNet	0.20	0.26	8.56
MLPRegressor	0.18	0.23	8.70
GeneralizedLinearRegressor	0.17	0.23	8.71
TweedieRegressor	0.17	0.23	8.71
OrthogonalMatchingPursuit	0.17	0.23	8.72
PassiveAggressiveRegressor	0.08	0.14	9.19
LassoLars	-0.08	-0.00	9.95
DummyRegressor	-0.08	-0.00	9.95
RANSACRegressor	-0.09	-0.02	10.03
KernelRidge	-2.08	-1.88	16.84
GaussianProcessRegressor	-17597.29	-16404.19	1272.14

These are the results when building regression models using the first dataset. The main features was used are Ammonia, PH, NOPA to predict Glucose/Fructose. Like we mentioned earlier we using thirty plus regression algorithms to see which one gives us the best result. From the picture on the right side, the top three performing models are, ExtraTreesRegressor, XGBRegressor and Random forest regressor.

Extra tree regressor performed the best, which is due to its extremely randomized nature and is known to work well with noisy data (Noisy data is data with a large amount of additional meaningless information in it called noise). The adjusted R-squared is 0.73 which is a decent value and is setting up a baseline for us to determine if adding more predictors in future would help or not. We are getting an RMSE of 4.97 which for a baseline is a good value.

Predictive Analysis (Final Dataset)

- In order to get a better results, we added more columns to our original dataset.
- Again, ExtraTreesRegressor gave the best result.
  - **Adjusted R-Squared:** 0.83
  - **RMSE:** 2.56

Model	Adjusted R-Squared	R-Squared	RMSE
ExtraTreesRegressor	0.83	0.93	2.56
DecisionTreeRegressor	0.77	0.91	2.98
XGBRegressor	0.73	0.89	3.19
RandomForestRegressor	0.73	0.89	3.22
GradientBoostingRegressor	0.65	0.86	3.66
BaggingRegressor	0.63	0.85	3.76
ExtraTreeRegressor	0.56	0.82	4.09
AdaBoostRegressor	0.47	0.78	4.51
KNeighborsRegressor	0.36	0.74	4.94
LGBMRegressor	0.17	0.66	5.64
HistGradientBoostingRegressor	0.16	0.66	5.66
LassoLarsIC	-0.58	0.36	7.77
SGDRegressor	-0.61	0.34	7.85
BayesianRidge	-0.62	0.34	7.86
ElasticNetCV	-0.64	0.33	7.91
RidgeCV	-0.64	0.33	7.92
LassoCV	-0.65	0.33	7.93
LarsCV	-0.67	0.32	7.98
LassoLarsCV	-0.69	0.31	8.03
Lasso	-0.71	0.30	8.08
OrthogonalMatchingPursuit	-0.72	0.30	8.11
TweedieRegressor	-0.74	0.29	8.16
GeneralizedLinearRegressor	-0.74	0.29	8.16
Ridge	-0.76	0.28	8.21
NuSVR	-0.79	0.27	8.27
ElasticNet	-0.80	0.27	8.28
SVR	-0.80	0.27	8.29
OrthogonalMatchingPursuitCV	-0.81	0.26	8.31
PoissonRegressor	-0.82	0.26	8.34
LinearRegression	-0.89	0.23	8.51
TransformedTargetRegressor	-0.89	0.23	8.51
MLPRegressor	-0.96	0.20	8.65
GammaRegressor	-0.99	0.19	8.72
HuberRegressor	-1.17	0.12	9.11
LinearSVR	-1.21	0.10	9.19
Lars	-1.24	0.09	9.25
RANSACRegressor	-1.32	0.06	9.42
LassoLars	-1.46	-0.00	9.69
DummyRegressor	-1.46	-0.00	9.69
PassiveAggressiveRegressor	-2.80	-0.54	12.04
KernelRidge	-5.76	-1.75	16.07
GaussianProcessRegressor	-1068.80	-434.17	202.14

In order to get better model, we extracted more features from HANA and combined them together to get final dataset. Again, ExtraTreesRegressor gave the best result, there was a decrease in RMSE value to (2.56 from 4.97). This a sign that we are in the right direction because the range of glucose/fructose values we saw are from 0 to 50 and an error of 2.56 is a significant improvement from 4.97. The increase in Adjusted R-Squared to 0.83 from 0.73 which shows that by adding more predictors we actually did increase model's ability to predict.

We needed access to more records and features to get better results.

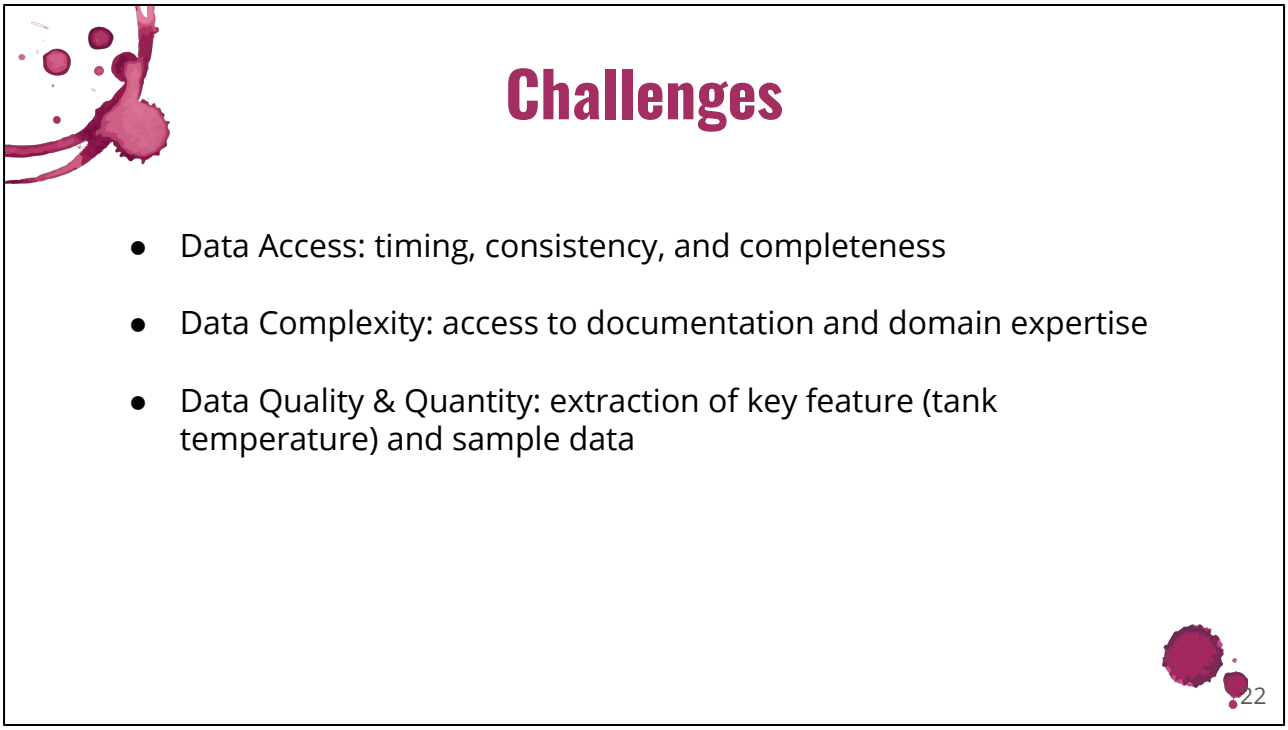
Insights

- Extra Trees Regressor has been the best performing algorithm from start.
- Composition data adds value to the model.
- Impossibile to generate simulated data at the given moment.
- If given sufficient number of records, an accurate business predictive model with lower RMSE can be created with the presented predictors/features earlier.

So to summarize our findings, as seen from results earlier extra trees regressor is the best performing algorithm in both scenarios and should be kept in consideration when there is enough sample data. As seen in the second scenario when we added composition data the adjusted r squared value increased which clearly is an indicator that its important and adds value to the predictive model. Samples are currently taken from Livingston site, but if more samples data are captured from different sites, the model will be more comprehensive since each sites could have different temperature conditions. The state the data is right now in it can't be used to generate actionable simulated data in this project's timeline. If samples are taken consistency, then a statistician could simulate data in short term using statistical modeling. In long term though, when enough samples are gathered then time series forecasting could be used to predict fermentation trend and create an accurate production level predictive model with an even lower RMSE can be created with the presented predictors/features earlier.

04. Recommendations

Challenges and Next Steps



CHALLENGES [02:56]

As requested, we have compiled a list of challenges we faced along with recommendations to assist Gallo in improving the process for future teams. Our most significant challenges arose from issues involving data access, data complexity, as well as the quality and quantity of available data.

1. One of our biggest challenges was achieving full and consistent access to the data we required. It wasn't until five weeks into the project that any on our team were able to explore Gallo data firsthand. At that point some of the team gained access to Oracle data, but firewalls and VPN issues extended the delay into week 6 for Mac users. Then, slow export speeds made it impossible to work with large quantities of Oracle data, so Josh began working on getting us access to the HANA database, which we started working with week eight. We were never able to achieve working HANA credentials for the entire team, however, and retaining access to any of the databases continued to be a challenge throughout the project. An additional problem was with the completeness of the data we were using; on several occasions we discovered important data missing from HANA, further slowing our ability to understand and use the data effectively, especially given its complexity.

2. This data complexity was the second significant challenge we faced. We

were required to form our understanding of the available data primarily through verbal communication and live demos -- without written documentation from a data dictionary or entity relationship diagram -- though answers to specific questions were provided by email or on Microsoft Teams. The same was true for information about Gallo's wine fermentation process. It wasn't until early March that we were given diagrams depicting Gallo's three distinct fermentation processes for white, red, and sparkling wine. One team member was eventually able to locate a Gallo sales manual from 2002 containing over 500 pages of detailed explanations useful for decoding portions of the available data, but the ad hoc nature of our access to this information slowed our progress significantly, as did our inability to meet with winemakers to gain domain-specific knowledge about the processes captured in Gallo's data, despite initial optimism about the possibility of doing so and several meetings scheduled for this purpose. While our Gallo contacts did their best to provide us with the information we needed, a deeper domain understanding would have helped us extract relevant data much more quickly and effectively, and thus contribute to the creation of a more effective model.

3. Because of the delays arising from these challenges, it was not until two weeks ago --week 13-- that we were finally able to extract the necessary data and began quickly merging and cleaning the data for model use. In the end, the data that we had was not sufficient to predict the best time to bottle Gallo wine well. As our team was unable to extract tank temperatures that corresponded to lab analysis of juice undergoing fermentation, we were missing one of the most important features. The small number of samples available for modeling, and the inconsistent timing of sampling, were additional issues. Thus, we only had 296 rows available in our final dataset, and we needed more features to make a better model with higher accuracy.

The next slide has our list of recommendations, ranked in order of importance.

Recommendations

1. Collect additional data on temperature and samples
2. Enable early and consistent access to data
3. Provide written documentation about data and Gallo
4. Increase access to domain expertise through wider stakeholder involvement

RECOMMENDATIONS [1:50]

1. Gallo has deep data going back many, many years. However, based on our work, it appears that additional data of several types is needed. Thus, our most important recommendation is for additional data collection. Specifically, improved collection and storage of tank temperature is important, due to the role of temperature in fermentation. When storing tank temperature, we recommend that local temperature also be stored due to its potential importance as a factor. Further, if samples are taken at a set schedule and consistently, then a forecasting model could be created using deep learning to build an RNN -- or even a simple time series -- to predict fermentation values at a given time for specific wines.

3. Our next recommendation is that Gallo provide earlier, more consistent, and more complete access to the data to allow for faster identification of problems with data quality and quantity, as well as earlier identification of knowledge gaps.

2. Our third recommendation is that written documentation relevant to the project be provided to team members as part of the initial orientation. A data dictionary and entity relationship diagram would be especially helpful, and would be useful to regular Gallo employees involved in data analysis and

manipulation as well. We also recommend including Gallo-specific information relevant to the project (like the diagrams and sales manual mentioned previously). Access to relevant documentation would also reduce the need for extended consultation with Gallo staff throughout the project, thus reducing the strain of integrating an additional project into existing work responsibilities.

4. Our final recommendation is to increase access to domain expertise by involving key stakeholders -- like Gallo winemakers -- in the project. This would help future teams deliver better business value through increased understanding of stakeholders' needs early on in the project. It's especially important to be able to consult internal domain experts when outside help is not possible because of non-disclosure agreements.

The next slide lists the specific next steps we recommend.

Next Steps

1. Collect tank and local temperatures at regular, frequent intervals *(5 minute intervals)*
2. Collect samples of fermenting juice at regular, frequent intervals *(3 times per day for a minimum of one month)*
3. Collect a timestamp when fermentation is complete

SLIDE: NEXT STEPS [00:35]

First, collect tank and local temperatures at regular and frequent intervals, we suggest 5 minute intervals. Second, ensure that samples of fermenting juice are taken on a set schedule and at least three times a day for a minimum of one month for a specific season like summer. And third, collect a "finished" timestamp when fermentation is complete. It's critical that this data be high quality -- that is, be consistent, accurate, and complete, so we recommend that a data quality engineer oversee its collection. Successful collection of this additional data -- even for a limited period of time -- would make it possible to simulate data based on statistical models and mathematical equations in the short term to build predictive models for short term.

CLOSING COMMENTS: [00:35]

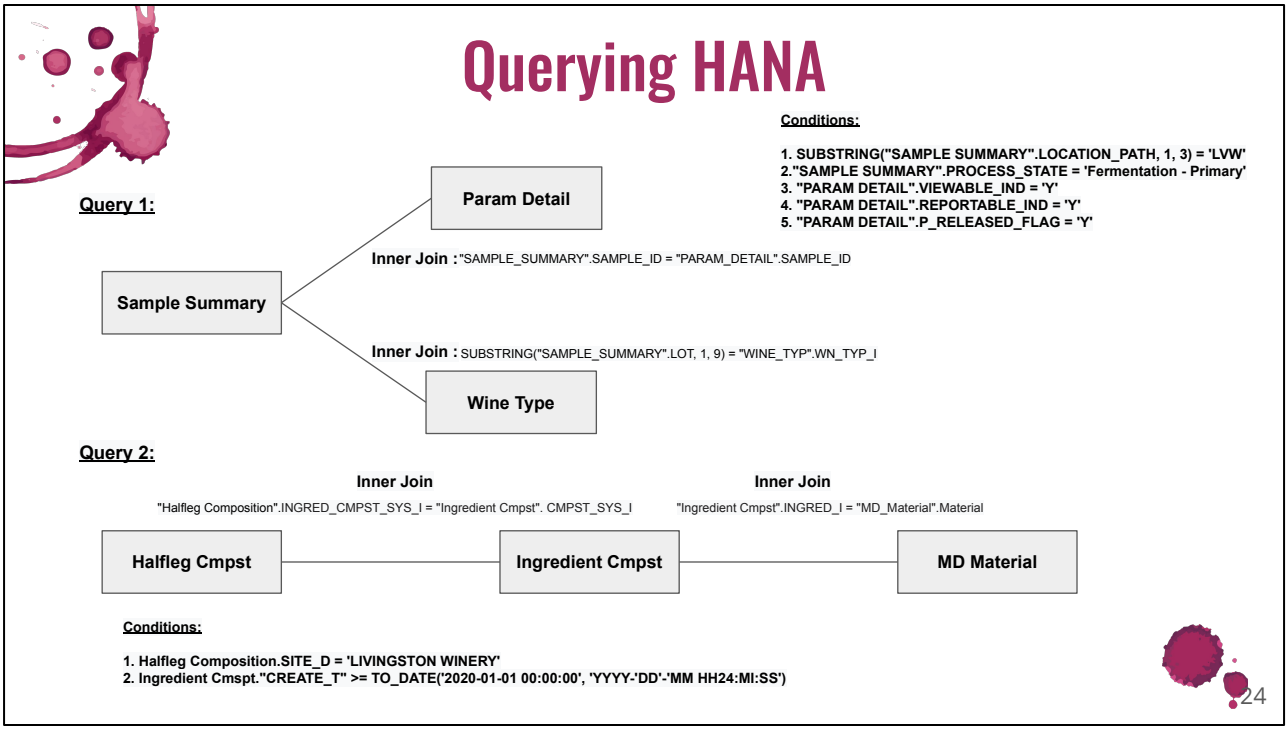
As we close, we want to say that we understand that the ability of any organization (including Gallo) to provide frictionless access for a project of this type can be limited for a wide variety of reasons, ranging from time constraints and security concerns to hacker attacks and a global pandemic, and we are deeply appreciative of the opportunity to work the Gallo to use complex data to attempt to solve a real world problem and create business value. It is our hope that the lessons learned from the challenges we faced over the course of the project can prove valuable to all involved, contributing to increased success for future projects. We're happy to answer any questions you might have. Thank you.

THANKS

Do you have any questions?

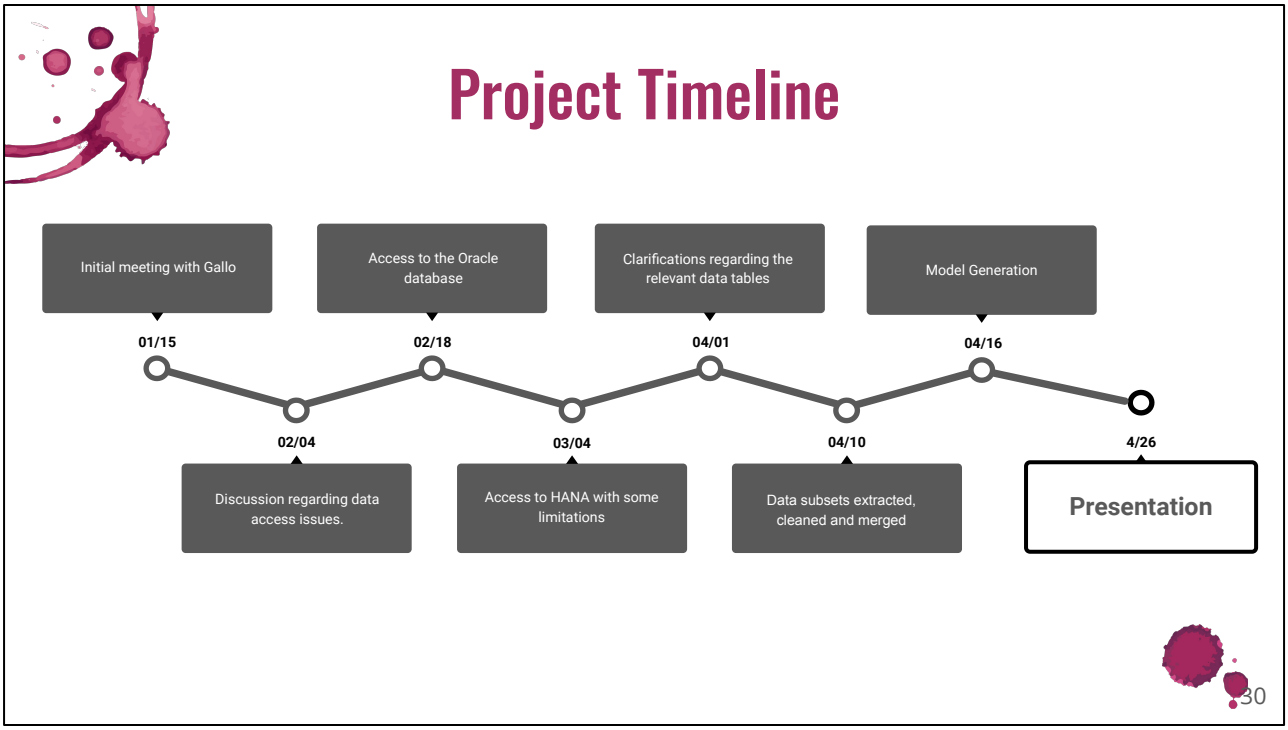
Next Steps

- Ensure that samples are taken on a set schedule and a minimum of three times a day.
- Using this data can be simulated by developing statistical models/ mathematical equations in the short term to build predictive models.
- Hire a data quality engineer to ensure that sample data collected is of high quality. For example
  - Consistency
  - Accuracy
  - Completeness



For the first query, we join sample summary table with param detail with Sample ID as common column. Sample Summary table is also merged with Wine Type table using common column Wine Type. And then we filter the Location to be livingston for the site, and process state would be fermentation-primary. In param detail table, we would also filter the viewable,reportable and released flad columns to be yes.

For the second query, we join Halfleg Cmpst table to Ingridient Cmpst table with common column as CMPST_SYS_I. Ingredient Cmpst is also merged with MD_Material table with common column Ingridient_I = Material.



Our journey essentially began on **January 15** with the first virtual meeting. We met with Joshua and John and we discussed Gallo itself and the goal of the project. Joshua also introduced the general wine making process. We also received basic information regarding what we will need to access Gallo's database such as Oracle SQL Developer, a VPN and etc..

After our first meeting, we ran into technical issues restricting our team from accessing the Oracle database due to VPN issues. **On February 4th,** we had our first technical troubleshooting meeting and began working with Gallo to resolve these problems.

Two weeks later, **on February 18th**, the team was able to gain access to the Oracle Database. Now we were able to inspect the database, however, we were unable to pull large quantities of data due to very slow export speeds. As a result, Josh began working on getting us access to HANA database.

On March 4th, our team was able to gain access to **one** HANA account, which gave us the ability to extract data at a much faster rate. But before we could pull the relevant data required to build our model, we still needed more clarification on the relationships of different data tables in conjunction with the fermentation process

On April 1st, we were able to get clarification from Jessica.

And **on April 10th,** our team was able to extract the necessary data. We then quickly started to merge and clean the data for model use.

On April 16th, the first models were developed and worked on. The insights and predictive modeling was done and the presentation planning begun.

Which brings us to today, the presentation. Although the timeline unveils our substantial struggle with data access and extraction, we were ~~able to produce a passable model that revealed we are~~ on the right track in selecting relevant features to predict fermentation process.

Recommendations

- Consistent samples at different time intervals.
- Proper temperature data.
  - Internal Tank Temperature when sample is taken
  - Local Temperature
- An ERD diagram that could help us understand how tables are linked together.
- More in-depth time series analysis such as using Recurrent Neural Networks could be conducted if
  - there are more samples taken
  - samples are taken in a constant time interval.

1) The samples should be taken at a set schedule and should be collected often
2) Tank temperature is an important factor in determining fermentation process which is not collected accurately. When storing tank temperature local temperature should also be stored as it also a contributing factor in fermentation.
3) An ERD diagram of tables in the database would be really helpful to understand how these databases are linked and will be especially helpful in setting up graph db structure.
4) Now, if samples are taken at a set schedule and consistently then deep learning could be used to build an RNN to predict fermentation values at a given time for a certain wine.