Report Format Your report can be about 20 pages long (including cover page,

figures, and tables). Use 1.5 inter-line space, 12pt font size, and

1-inch margins on all sides. Your report should include the

following sections.

Problem Setting Describe the context in which the problem is defined. It can give

background of the domain, application, and challenges

Problem definition State specific problem being addressed within the scope of the

problem setting. It could be the list of questions for which you

are trying to find answers through data analytics.

Data Sources Describe the data source including citation to website and/or

publication reference.

Data Description Describe data available to address the problem. The description

can include names of the datasets, number of columns, number

of rows, and a sample of variable names

Data Exploration Describe statistical and visualization methods used to explore

the data.

Data Mining Tasks Describe data mining tasks performed in the project. They can

include data reduction, data transformation, missing data

imputation, classification, prediction, association, and clustering

Data Mining

Models/Methods

Describe data mining models and methods employed to address

the problem. They include multiple linear regression, Logistic

regression, k-NN, Naïve Bayes classifier, decision trees, neural

networks, linear, neural networks, association rules, etc.

Performance

Evaluation

Discuss the performance evaluation methods used. They include

metrics such as MAE, AE, and RMSE. They can also include tools

like lift charts, AUROC, F1 score, etc.

Project Results Discuss the key finding and deliverables of the project

Project Outcomes

Describe the value created by your data mining effort,

particularly by your predictive models and/or association rules.

**Problem Setting**

Our project problem will be the data analytics and forecast for sales. The domain of our problem is the beer industry, and the challenge will be finding the impact of COVID-19 on sales and forecasting future sales data based on the past data and the impact of COVID-19.

**Problem Definition**

1. What is the sales volume or performance of one beer brand/item in one state/city/liquid store or pub in past few years/months?
2. How does COVID-19 impact the sales volume?
3. How will sales perform in the next year?

**Data Source**

The sales data will be collected from Budweiser and the data will be desensitized.

**Data Description**

There will be two datasets. The first one will be “item”, it consists of 118 rows and 5 columns. Its columns names include columns like item, brand, brand series. The second one will be sales data named “train”, it has 145195 rows and 7 columns. Column names includes columns like date, item, order number, city, and state.

**Data Exploration:**

**Data Mining Tasks**

**Data Mining Models/Methods：** We apply all regression methods learned in this course to our project, including multiple linear regression, KNN regression, Random Forest Regression, Neural Network. We tune hyperparameters and choose the best model for all methods. We also applied grid search with cross validation on multiple linear regression method, pipe with normalization on neural network

**Performance Evaluation:** As we use regression methods, we use MSE (mean squared error), RMSE (root mean squared error), MAE(mean absolute error). The main metric we use to identify our best model is RMSE

**Project Results:**

**Impact of the Project Outcomes:**