IE7275 Final Project Report:Impact of Covid-19 on beer sals

-- Team 22

Problem Setting

Digital technologies, artificial intelligence and machine learning are acce lerating in all aspects of our work and life, and are having a significant impact on business management and operations. They are having a significant impact on business management and operations. Many companies are beginning to apply data science in depth to inspire transformative innovation and dig ital transformation. Advanced technologies can help us extract many feature s and trends from data that cannot be noticed manually, and apply advanced algorithmic models that can be used to improve the company. The application of advanced algorithmic models can be used to improve business statistics a nd forecasting, optimize business decisions, and enhance operational effici ency.

Problem Definition

Demand forecasting is the beginning of the entire sales and logistics suppl y chain planning and is one of the most critical planning operations for an y consumer goods company. Proper demand forecasting can effectively guide s ales planning and supply chain production planning, package procurement pla nning and production planning, which is important for business operation ef ficiency. It is important for the improvement of business efficiency. In th e past, demand forecasting has relied more on manual extrapolation and simp le traditional time series models. Recently, digitalization transformation has taken root in every industry, and it is imperative to apply advanced al

gorithmic models to improve the performance of a company's statistical fore casts.

With lots of types of wine, the company separates its products as many bran ds to aim different customers and areas. Vast information can lead to incor rect sales estimates between manufacturers and dealers. At this time, the p rocessing and prediction of data is very important.With lots of types of wi ne, the company separates its products as many brands to aim different cust omers and areas. Vast information can lead to incorrect sales estimates bet ween manufacturers and dealers. At this time, the processing and prediction of data is very important.

Data Source

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

Data Description

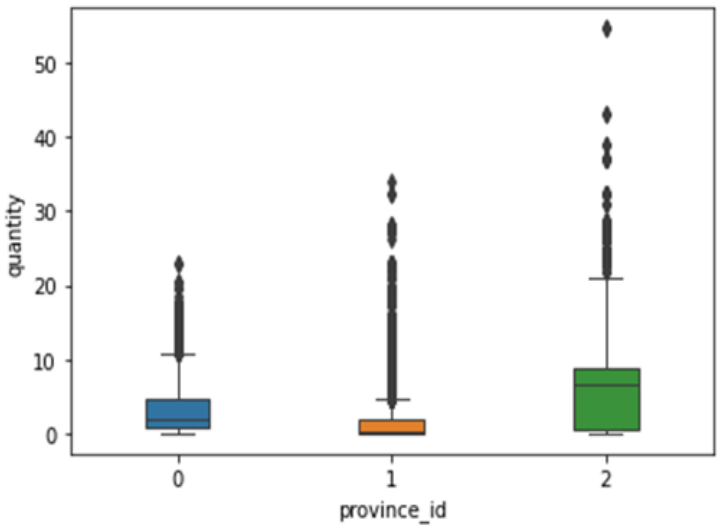
There will be two datasets. The first one will be "item", it consis ts of 118 rows and 5 columns. Its column names include columns like item, brand, brand series. The second one will be sales data named "train", it has 145195 rows and 7 columns. Column names includes columns like date, ite m, order number,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

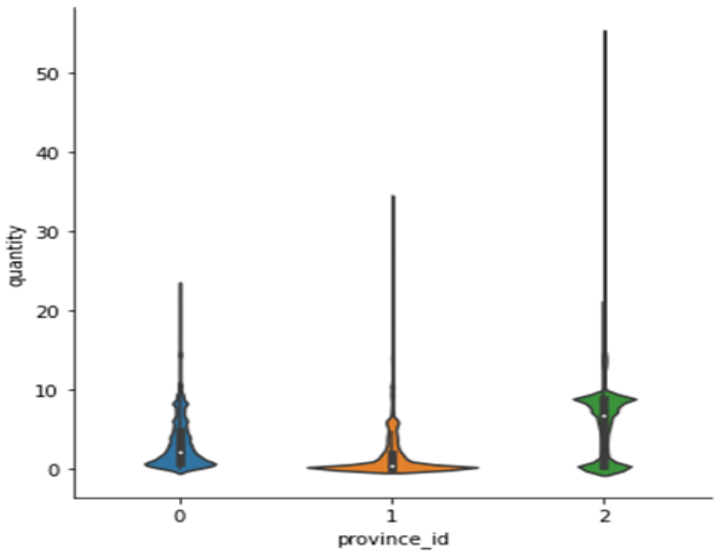
Data Exploration

The first thing to do for EDA is using seaborn to draw the heatmap.

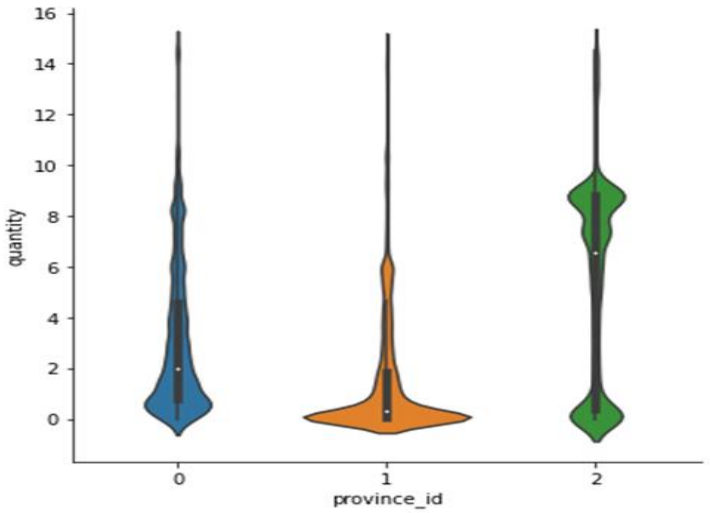
It shows that predictors are almost independent with each other. Most Correlation values are very small and there are correlations greater than 0.7. It is not necessary to do dimension reduction at this stage.

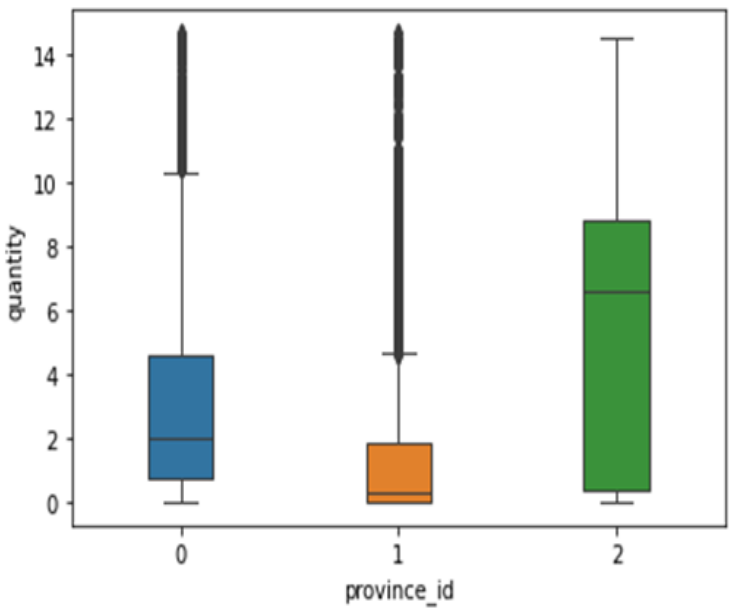
Also, by seaborn, the second step is to draw the boxplot and violin plot.





These two results look very weird and there are a lot of outliers, they nee d to be removed as much as possible.



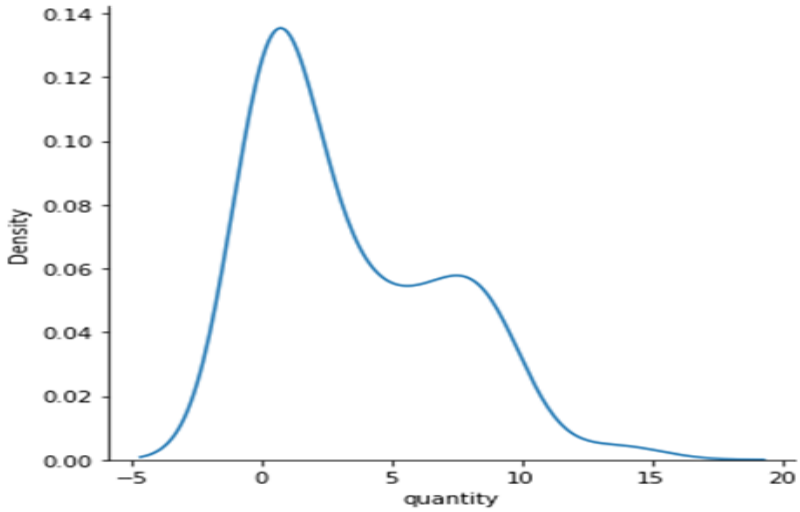


datetime	item_id	brand_id	brandfamily_id	package_id	info_id	order_id	user_id	province_id	city_id	quantity
2017-01-02	1	6	1	1	3	8806	404	2	19	8.298021
2017-01-02	1	14	1	1	3	8806	404	2	19	8.298021
2017-01-02	1	6	1	1	3	22552	404	2	19	8.298021
2017-01-02	1	14	1	1	3	22552	404	2	19	8.298021
2017-01-06	1	6	1	1	3	54066	395	0	10	0.377183
...	...	...	...	...	...	...	...	...	...	...
2020-09-21	12	19	3	4	11	52082	18	0	10	3.643725
2020-09-26	12	19	3	4	11	11453	450	0	10	1.457490
2020-09-11	12	19	3	4	11	58771	262	0	10	2.914980
2020-09-12	12	19	3	4	11	33368	248	0	10	1.457490
2020-09-19	12	19	3	4	11	91984	24	0	10	0.728745

183546 rows × 10 columns

Set Q1 and Q3 equals to quantile(0.25) and quantile(0.75), apply IQR rules and drop from the index, the rows left are 183546. Although the boxplot and violin plot does not look perfect, they are much better and it's necessary to keep some outliers for reality.

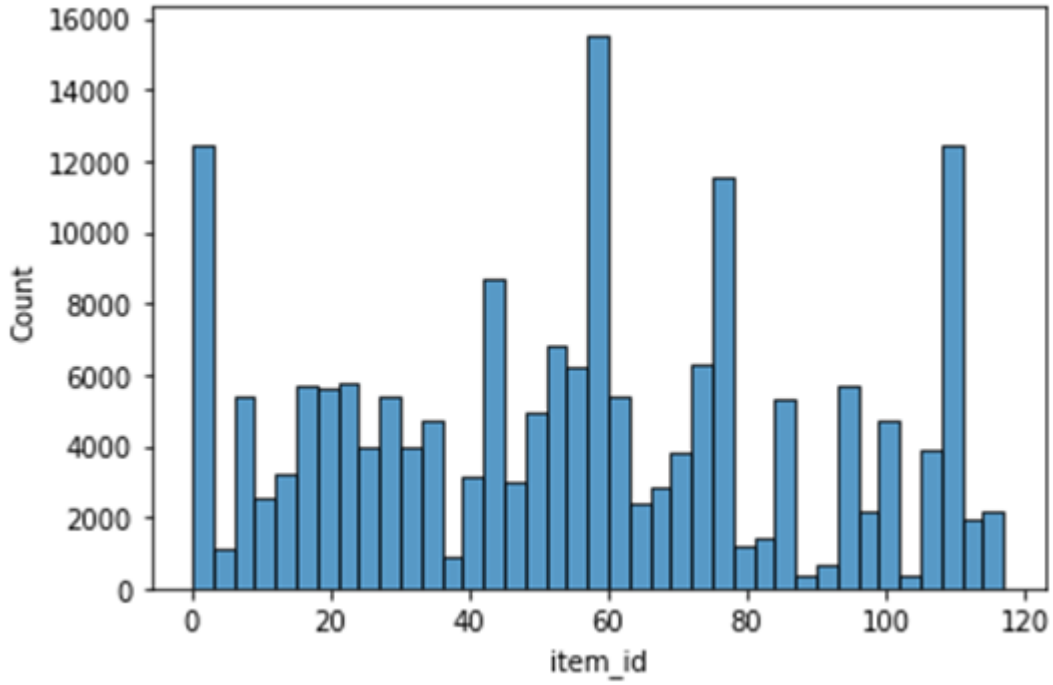
Data visualization



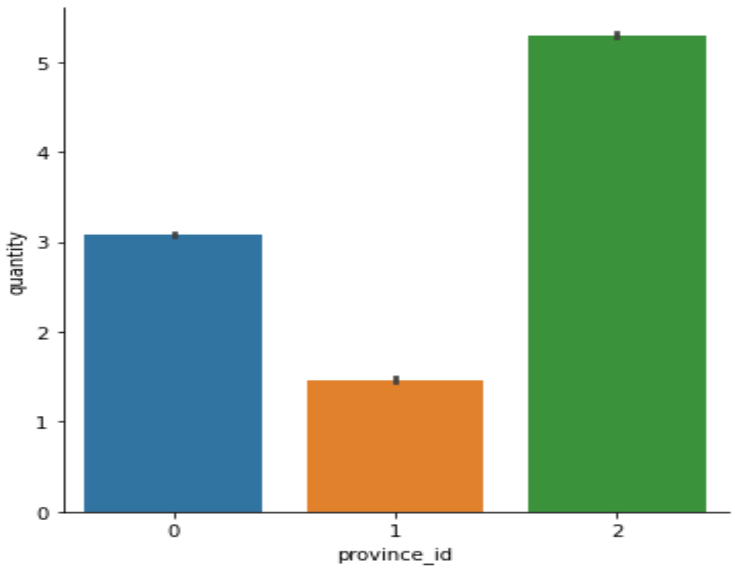
Keep using seaborn package in this part, the first graph is density plot by displot.

From the density plot of quantity, it shows that most quantities are small and near 0, only a small part are large.

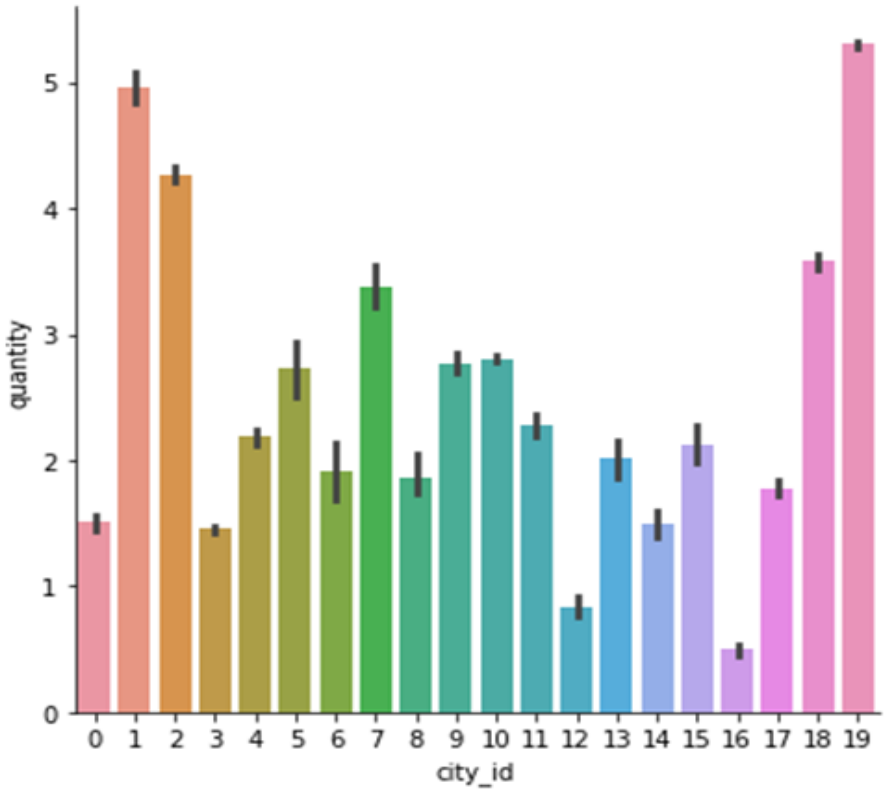
And then histograms by histplot

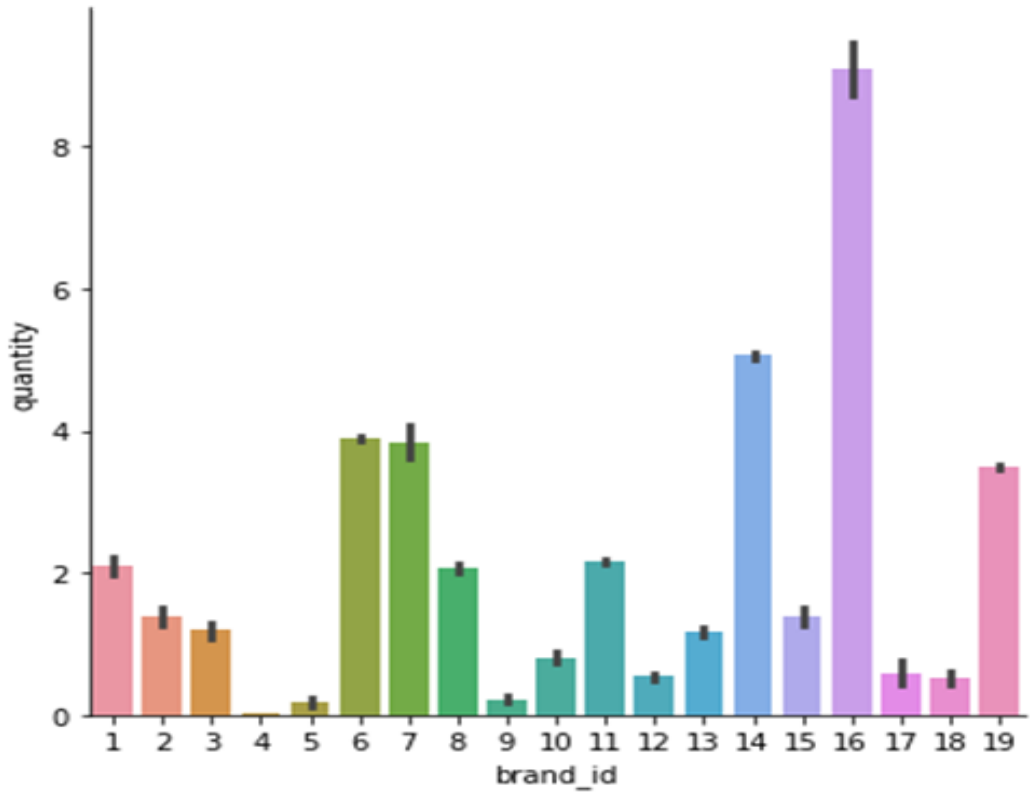


The histograms are through id and there is a huge difference between each i tem.



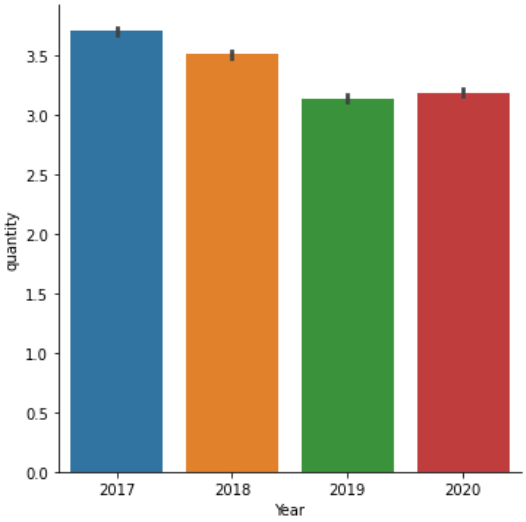
By bar chart it can be found that province 2 has most quantities, followed by province 0. Province 1 has the least quantity.

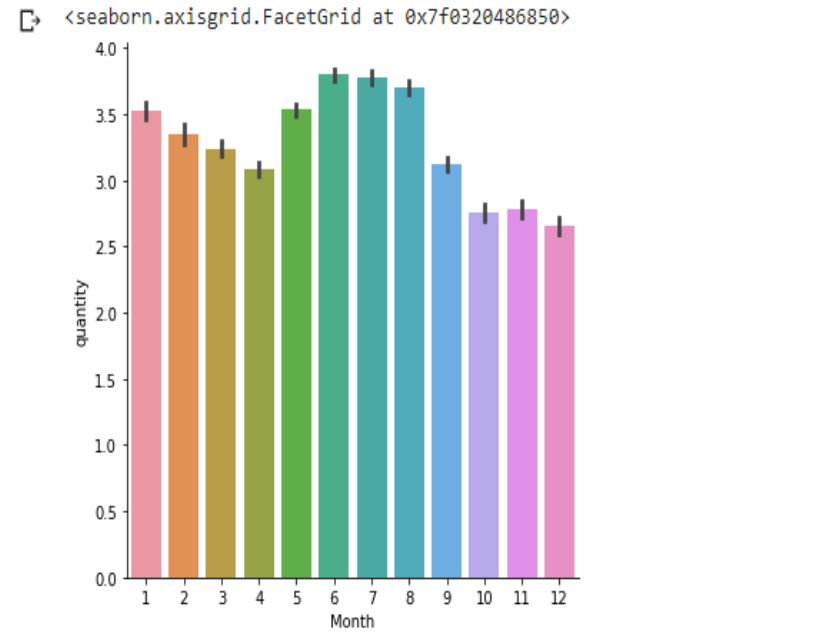




And the sales performance difference among cities and brands are also huge, which laid the base that the further analysis and prediction should focus on items and states.

Then we draw bar graph with different years and months





There is not much difference on the sale between each month. We can see that in the winter the sales are not as good as other seasons. The overall trend for years is decreasing. We find that even if we have Covid-19 in 2020, the overall sale did not decrease a lot from 2019.

Data Mining Tasks

Wine is an alcoholic drink typically made from fermented grapes. Yeast consumes the sugar in the grapes and converts it to ethanol, carbon dioxide and heat. Different varieties of grapes and strains of yeasts are major factors in different styles of wine.

Now, the team received one wine company's desensitization data, which masked an important message of business replaced by numbers. The company hopes to receive predicted data in order to understand its selling information in the next several months.

The team's first objective is to analyze the sales performance of each product in each state, from 2017 to 2020, by statistical and visualization methods. And then the team will figure out how the sales performance changed in the past few years, especially in 2020 under the COVID-19 situation. Finally, the team will make a prediction for the future sales performance based on the past data.

	item_id	brand_id	brandfamily_id	package_id	info_id
0	1	[6, 14]	1	1	BHS
1	6	[6, 19]	2	1	BHM
2	54	[6, 14]	1	1	BHS
3	15	[6, 19]	2	1	BHM
4	108	[19]	3	1	MQB
..	...	...	...	...	...
113	78	[6, 8]	4	4	BZMN
114	5	[11]	5	1	HGO
115	83	[19]	3	1	MQB
116	87	[19]	3	2	MQB
117	12	[19]	3	4	MQB

[118 rows x 5 columns]

There are two datasets that will be used in this project. The first one is the dataset for product information. There are 5 columns in this dataset. T

he column names are item id, brand id, brand family id, package id, and inf o id. And there are 119 rows for unique products.

	datetime	order_id	user_id	item_id	province_id	city_id	quantity
0	2017-01-02	8806	404	1	2	19	8.298021
1	2017-01-02	22552	404	1	2	19	8.298021
2	2017-01-02	494	489	6	1	0	0.376088
3	2017-01-02	62721	28	54	2	19	1.488643
4	2017-01-02	24790	231	15	0	10	0.748268
...	...	...	...	...	...	...	...
145190	2020-12-29	72305	390	18	0	10	0.733838
145191	2020-12-30	55224	321	27	1	3	0.792584
145192	2020-12-31	75987	326	105	0	4	0.009584
145193	2020-12-31	27911	326	105	0	4	0.015935
145194	2020-12-31	55625	326	75	0	4	0.069767

[145195 rows x 7 columns]

The second dataset is the dataset for sales performance. There are 7 column s in this dataset. The column names are date, order id, user id, item id, s tate id, city id, and quantity. This dataset shows that in an unique order, how much quantity of a product is sold to a user, which stands for the liqu id store, also the order date and the location information of this liquid s tore. There are 145196 rows in this dataset.

Data integrition

There are two datasets, so the first step is to combine them together. Afte r using pandas to read these two files, it shows that there is a common col umn called item id. Then these two tables can be joined together by the sam e item id. Also, set date as row key as it is not a feature.

	item_id	brand_id	brandfamily_id	package_id	info_id	datetime	order_id	user_id	province_id	city_id	quantity
0	1	[6, 14]	1	1	BHS	2017-01-02	8806	404	2	19	8.298021
1	1	[6, 14]	1	1	BHS	2017-01-02	22552	404	2	19	8.298021
2	1	[6, 14]	1	1	BHS	2017-01-06	54066	395	0	10	0.377183
3	1	[6, 14]	1	1	BHS	2017-01-07	67828	377	0	4	0.754366
4	1	[6, 14]	1	1	BHS	2017-01-10	4205	349	0	9	3.017462
...	...	...	...	...	...	...	...	...	...	...	...
145190	12	[19]	3	4	MQB	2020-09-21	52082	18	0	10	3.643725
145191	12	[19]	3	4	MQB	2020-09-26	11453	450	0	10	1.457490
145192	12	[19]	3	4	MQB	2020-09-11	58771	262	0	10	2.914980
145193	12	[19]	3	4	MQB	2020-09-12	33368	248	0	10	1.457490
145194	12	[19]	3	4	MQB	2020-09-19	91984	24	0	10	0.728745

145195 rows × 11 columns

After using pandas merge, the join result gets a table with 145195 rows and 11 columns, which is good.

It can be seen that brand id has composite items as a list. In order to cha nge its types to numeric, use the literal_eval module from ast package to e xplode the list.

	item_id	brand_id	brandfamily_id	package_id	info_id	datetime	order_id	user_id	province_id	city_id	quantity
0	1	6	1	1	BHS	2017-01-02	8806	404	2	19	8.298021
0	1	14	1	1	BHS	2017-01-02	8806	404	2	19	8.298021
1	1	6	1	1	BHS	2017-01-02	22552	404	2	19	8.298021
1	1	14	1	1	BHS	2017-01-02	22552	404	2	19	8.298021
2	1	6	1	1	BHS	2017-01-06	54066	395	0	10	0.377183
...	...	...	...	...	...	...	...	...	...	...	...
145190	12	19	3	4	MQB	2020-09-21	52082	18	0	10	3.643725
145191	12	19	3	4	MQB	2020-09-26	11453	450	0	10	1.457490
145192	12	19	3	4	MQB	2020-09-11	58771	262	0	10	2.914980
145193	12	19	3	4	MQB	2020-09-12	33368	248	0	10	1.457490
145194	12	19	3	4	MQB	2020-09-19	91984	24	0	10	0.728745

217723 rows × 11 columns

Which gives 217723 rows as a result.

Data Cleaning

Since the dataset is the raw data from the company database, in order to cl ean the data, the first step is to check the data types.

```
item_id             int64
brand_id           object
brandfamily_id      int64
package_id          int64
info_id            object
order_id            int64
user_id             int64
province_id         int64
city_id             int64
quantity          float64
dtype: object
```

By using dtypes, it shows that the brand id and info id is an object, which is hard to handle. For the brand id, it can be converted to int64 by just using astype. But for info id, it should be converted to category and cat.codes it, then it can be converted to int 64 by astype.

```
item_id             int64
brand_id            int64
brandfamily_id      int64
package_id          int64
info_id             int64
order_id            int64
user_id             int64
province_id         int64
city_id             int64
quantity          float64
dtype: object
```

After converting the data type, it is necessary to check if there are any NA values by dropna or negative values in the dataset. Fortunately, there is no such a problem in this dataset.

Feature selections

The team thinks it can not directly use datetime as the column as it has to o many different possible values which are not proper for predictions. it i s converted to datetime as an index, instead of adding 'Year' and 'Mont h' as new columns for predictions.

```
[ ] df_wine['datetime'] = pd.to_datetime(df_wine['datetime'])
    df_wine.insert(0, 'Month',   df_wine['datetime'].dt.month)
    df_wine.insert(0 , 'Year',  df_wine['datetime'].dt.year)
```

```
[ ] df_wine.set_index('datetime')
    df_wine = df_wine.set_index('datetime')
    df_wine
```

datetime	Year	Month	item_id	brand_id	brandfamily_id	package_id	info_id	order_id	user_id	province_id	city_id	quantity
2017-01-02	2017	1	1	6	1	1	BHS	8806	404	2	19	8.298021
2017-01-02	2017	1	1	14	1	1	BHS	8806	404	2	19	8.298021
2017-01-02	2017	1	1	6	1	1	BHS	22552	404	2	19	8.298021
2017-01-02	2017	1	1	14	1	1	BHS	22552	404	2	19	8.298021
2017-01-06	2017	1	1	6	1	1	BHS	54066	395	0	10	0.377183
...	...	...	...	...	...	...	...	...	...	...	...	...
2020-09-21	2020	9	12	19	3	4	MQB	52082	18	0	10	3.643725
2020-09-26	2020	9	12	19	3	4	MQB	11453	450	0	10	1.457490
2020-09-11	2020	9	12	19	3	4	MQB	58771	262	0	10	2.914980
2020-09-12	2020	9	12	19	3	4	MQB	33368	248	0	10	1.457490
2020-09-19	2020	9	12	19	3	4	MQB	91984	24	0	10	0.728745

217723 rows × 12 columns

Check datatype again to make sure they consistent with other data

```
Year                int64
Month               int64
item_id             int64
brand_id            int64
brandfamily_id      int64
package_id          int64
info_id             int64
order_id            int64
user_id             int64
province_id         int64
city_id             int64
quantity          float64
dtype: object
```

Prediction and Forecast

The team will use machine learning methods to predict our target. There are 5 methods (KNN , Random Forest, Linear Regression, Neural Network, LightGBM) and they are all regression methods. The details of our models will be talked about below. The project uses previous years' data to predict recent years' data(all labels are available). After tuning parameters on all models, the team chooses the best model and predict on manually created test sets to forecast future values(both features and target are not available)

Hold Out Method

The team will try Hold Out in the project

The data sets will be splitted into two parts at first. As the project is a forecasting project, the dataset was splitted by year. Data belonging to 2017,2018,2019 is in the training set, data belonging to 2020 is in validation. The train set is used on every model and the performance is evaluated on validation model

Data Mining Models/Methods

KNN method

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. For the certain data sample, it can find out the several train samples that are close to the train set based on the distance metrics. When the team preprocesses the dataset, it shows a good linear shape, so a KNN regression model would naturally fit this data well.

In order to find the value k to decide how many samples are close to the train data, the data standardization needs to be done firstly. Then the team decided to use normalized KNN with grid search to find the best performance model.

Random Forest Method

Random Forest method is one of the most powerful machine learning methods that is used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. Every tree in the random forest is training a subset of the data.The basic idea behind it is to combine multiple decision trees in determining the final output, rather than relying on individual decision trees.Each decision tree has a high variance, but when we combine all decision trees together in parallel, the result variance is very low because each decision tree has been perfectly trained for specific sample data, so the output does not depend on one decision tree but on multiple decision trees. The final output is the average of all the decision tree outputs. Therefore, it can be seen as the fittest model for the current problem.

To improve the performance of the model, tuning hyperparameters is a necessary step. When the team tests the tuning method for the random forest model, it finds that even though the grid search is the better approach, it takes too much time to run the computation. Based on that situation, the team still chooses to find the best estimator: n and also the best max_depth: d.

Multiple Linear Regression Method

Multiple Linear Regression is the most simple and straightforward way for solving a regression problem with multiple predictors. The team will try three different types of linear regression: starting with normal linear regression and trying regularizations on linear regression: lasso regression(regularization with 1st norm) and ridge regression(regularization with 2rd norm). Applying gridsearch on the datasets will give more valid results for this method. The team will try to turn on and turn off "fit-intercept" and "normalize" for all our three linear models. Besides tuning those bool parameters, for lasso and ridge, by trying to tune the alpha values, it can be seen which alpha will generate better scores.

Neural Network Method

Neural Network is one of the most popular and powerful models nowadays. It usually takes a long time to run but generally gives good results as data is handled in detail in hidden layers. The data was scaled at first as it was already known that normalization will give better performance for neural networks generally. Then we tune our parameters. Here we try 4 combinations (0.001, 0.01, 0.1, 1) of learning rate and 3 combinations of transfer functions ('identity', 'logistic', 'relu').

LightGBM Method

LightGBM is a state of the art model and has been very popular in industry in recent years. As it is a model not that familiar compared with other models and also it takes a long time to run. The parameter tuning processing will not be complicated. Num_leaves is increased once (64 to 100) and max_bin is decreased once (255 to 100) to compare performances

Performance Evaluation

As the problem is a regression problem and also forecast problem, the team uses MAE, and RMSE, together with comparing forecast values with actual values in a line graph to evaluate the model performance. Those are regression metrics as our metrics to evaluate our models through the whole project. RMSE is chosen as the main metric to compare and identify which model is better, as it is able to give a fair evaluation about our loss in regression models. The line graph is used to see how close the predicted line and actual are fitting each other.

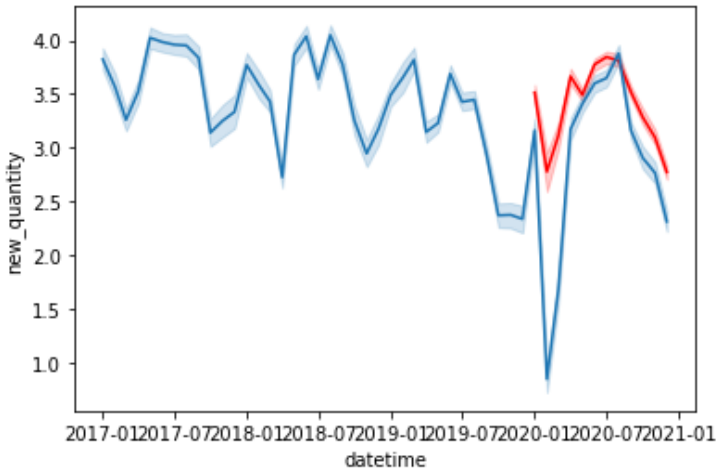
Project Results

Best Model Selection

We check performance by comparing both RMSE and closeness of predicted line and actual line.

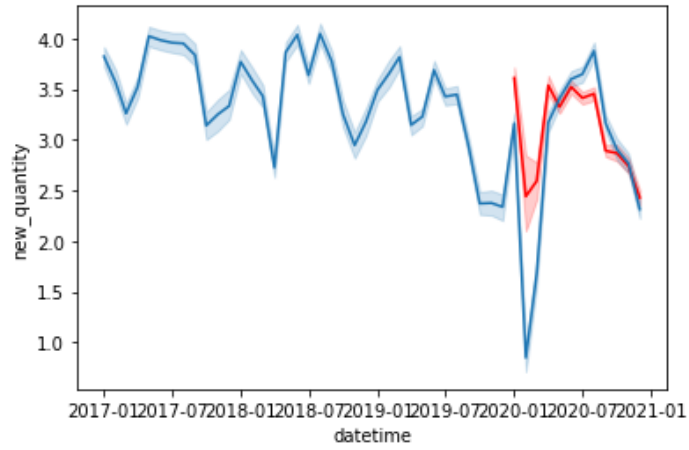
KNN:

	k	weights	alogorithm	mae	rmse
45	20	distance	ball_tree	2.23472	3.15163



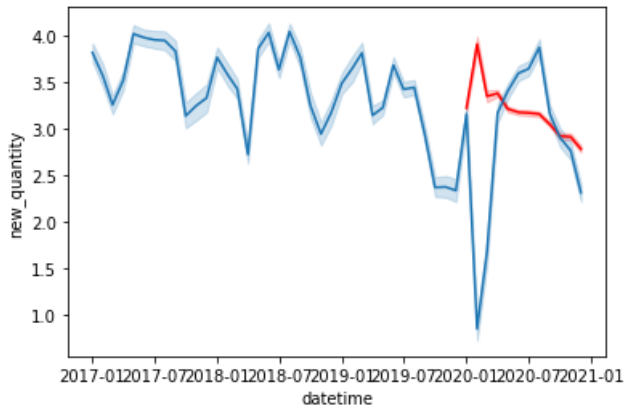
Random Forest:

	n	depth	mae	rmse
4	80	10	1.70245	2.82092



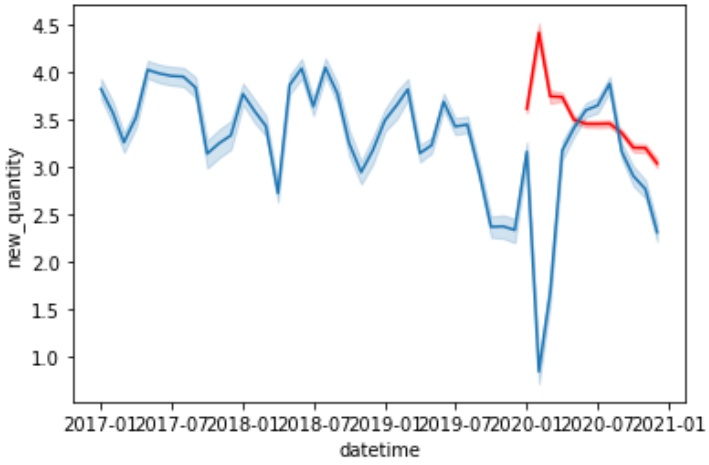
Linear Regression:

	alpha	fit	normalize	mae	rmse
4	0.5	True	True	2.64431	3.28438



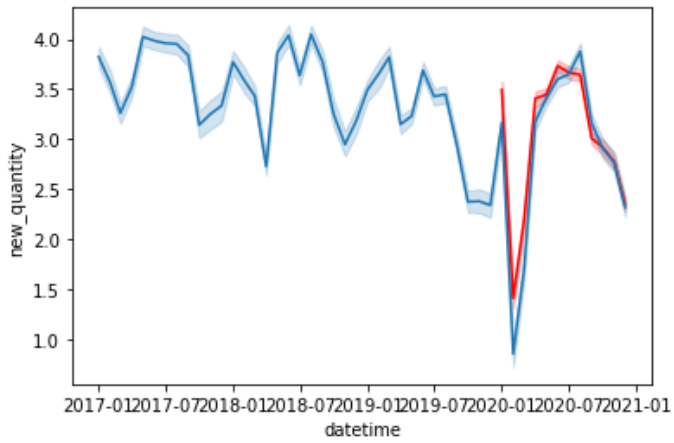
Nerual Network:

	leaning_rate	transfer_function	RMSE	MAE
0	0.001	identity	3.30504	2.66757



LightGBM:

```
Early stopping, best iteration is:
[2319]  valid_0's rmse: 2.18902
```



Rank the performance from best to worst:

RMSE:

LightGBM(2.18)

Random Forest(2.82)

KNN(3.15)

Linear Regression(3.28)

Neural Network(3.30)

Predicted Line:

LightGBM

Random Forest

KNN

Linear Regression

Neural Network

Overall Performance:

LightGBM

Random Forest

KNN

Linear Regression

Neural Network

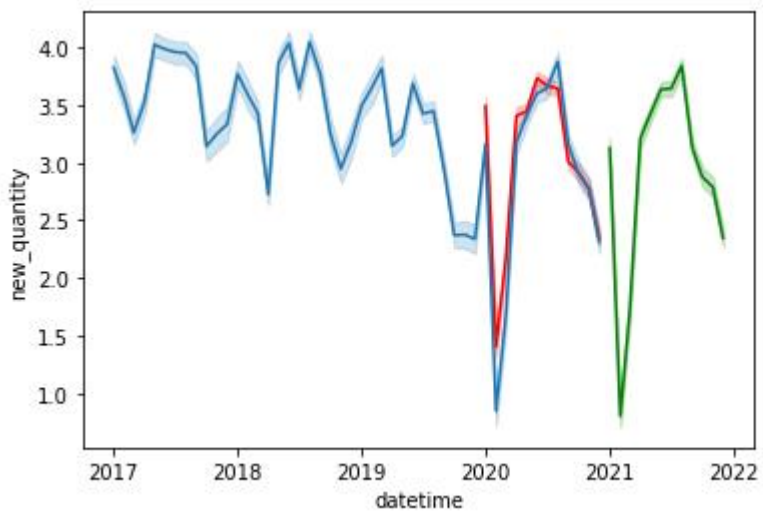
Predict on Test sets

The team chooses the best model: LightGBM with below parameters:

```
params = {
          "objective" : "regression",
          "metric" : "rmse",
          "num_leaves" : 100,
          "learning_rate" : 0.005,
          "bagging_fraction" : 0.7,
          "feature_fraction" : 0.5,
          "bagging_frequency" : 6,
          "bagging_seed" : 42,
          "verbosity" : 1,
          "seed": 42,
          }
```

As we lack an available data set for next year as test data , manually made data was necessary. The data distribution for year 2021 was assumed to be t he same as data in 2020, The only difference is the year was changed to 20 21. The train set will also cover more data to increase the predicting acc uracy, the most recent year 2020 will be included in the training phase ins

tead of in the validation phase. In other words, the best model was trained on the whole dataset and test on the manually made fake data to predict nex t year’s results.



The trend of year 2021 and 2020 was very similar as the data of 2021 mocks data of 2020. The forecasting is successful but this situation is not likel y to happen in real life as the trend was too similar. After training on e xisting data, how to make predictions about future results based on availab le existing data is a tough question and should be considered as a signific ant problem next time.

Impact of the Project Outcomes

With the final prediction based on the wine company’s original selling dat a, it will provide its predicted data to draw the next several month’s sell ing needs. By looking at the predictive data structure, the wine company ca n target certain hot products and reduce production of certain products to avoid unnecessary losses.In addition, in view of different forecast sales i n different regions, the company can deploy transportation capacity and the Sales Department in advance to connect with local dealers, and the Marketin g Department can also make further work arrangements for regions with weak forecast sales. While Covid-19 affects the accuracy of predictive models, i

t does provide a high degree of insight into how the company will perform i
n the coming months.