2023

Mini Hackathon 2023 Preliminary Round Report

TEAM NAME : DATAWONDERS

TEAM MEMBERS:

IDUSHA AMANDI PERERA

IRUSHA ARUNDI PERERA

RAMINDU WALGAMA

OSHANI WICKRAMASINGHE

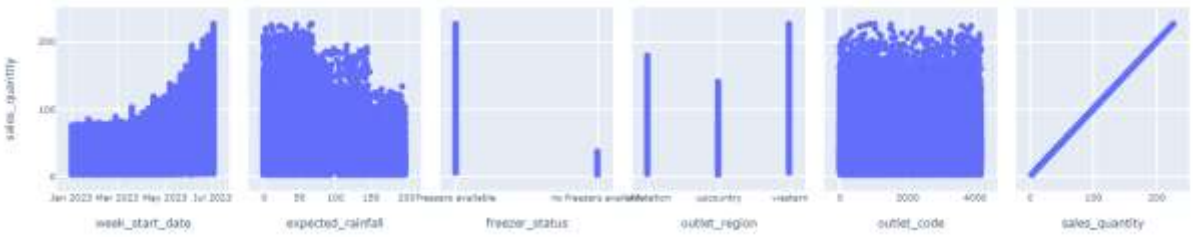
Analysis

This section will explain the feature engineering part of the data set. First, identify the data types of the columns. week_start_date is a date, and it was shown as an object then covert the data type to the datetime datatype for future use. Expected_rainfall was an object data type because of the values with its measurement. Remove the symbols and convert them to int. Split the week_start_date into day, month, and year to calculate the further relation between sales_quantity.

```
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   week_start_date   113400 non-null  object
 1   expected_rainfall 113400 non-null  object
 2   freezer_status    113400 non-null  object
 3   outlet_region     113400 non-null  object
 4   outlet_code       113400 non-null  object
 5   sales_quantity    113400 non-null  int64
dtypes: int64(1), object(5)
```

The freezer_status column is a categorical variables column. Therefore, identify the unique categories of the column. There are some data entry mistakes because of the additional spacing. Convert all values of the column to unique two categories and use label encoding to convert to the numerical values. The categories of the outlet_region convert to the numerical values by using ordinal_mapping. Convert all categorical variables to numerical values because many machine learning algorithms work with numerical data rather than categorical data. As well as in the outlet_code prefix of the data 'outlet_code_' remove and convert the column to the int.

Next, identify the correlation between the columns by using graphical visualization.



Also, use the method to calculate the correlation.

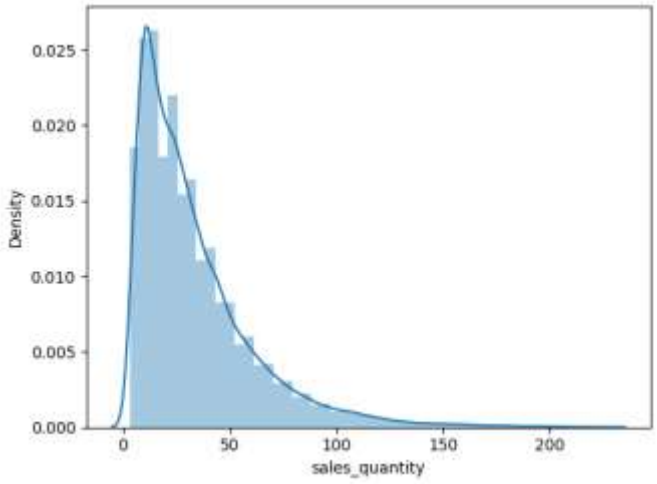
	expected_rainfall	outlet_code	sales_quantity
expected_rainfall	1.000000	0.001255	-0.093707
outlet_code	0.001255	1.000000	0.010736
sales_quantity	-0.093707	0.010736	1.000000

According to the above table, the expected_rainfall and outlet_code have no relation with sales_quantity. Therefore drop the expected_rainfall when using the machine learning model.

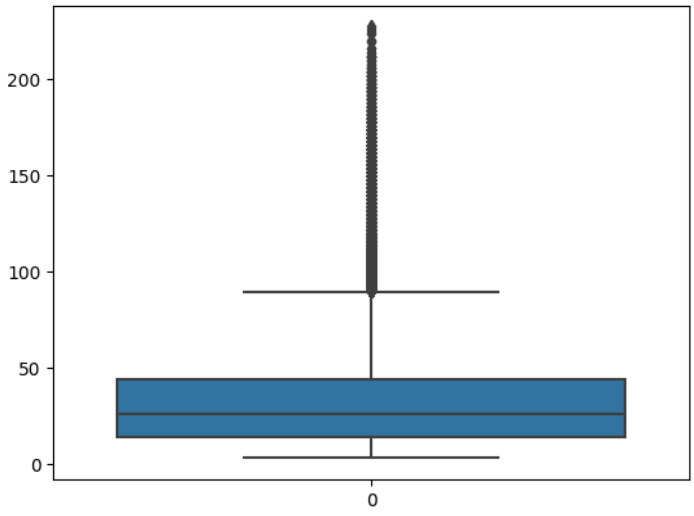
Outliers

This section will consider outliers. Outliers can happen in every dataset. When doesn't consider outliers, it may be given wrong information about the dataset.

In this data set, get the following graph.



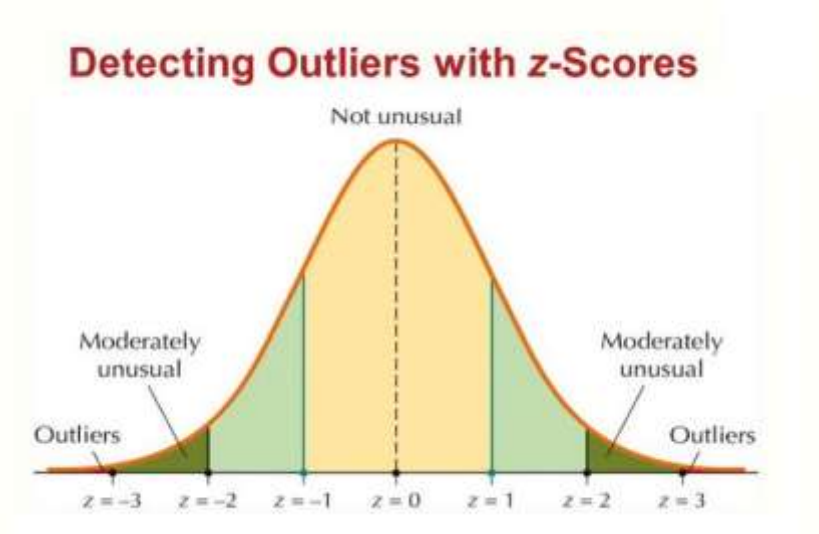
This plot is completely left-hand side. So definitely this dataset has outliers. When plotting the boxplot, it shows more information about outliers.



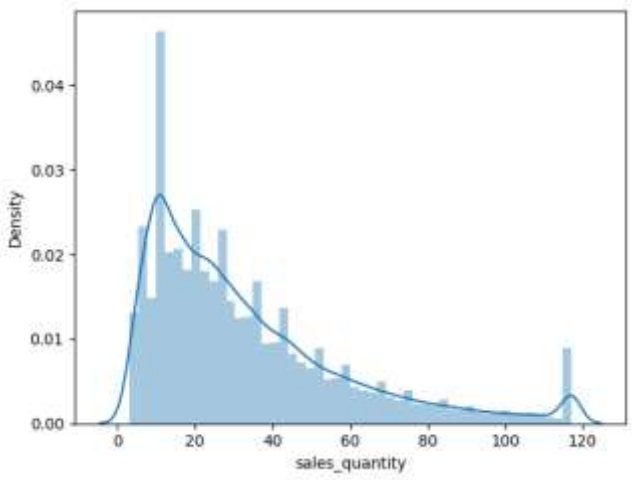
This boxplot shows the dataset has many outliers. Should consider how to decrease the number of outliers.

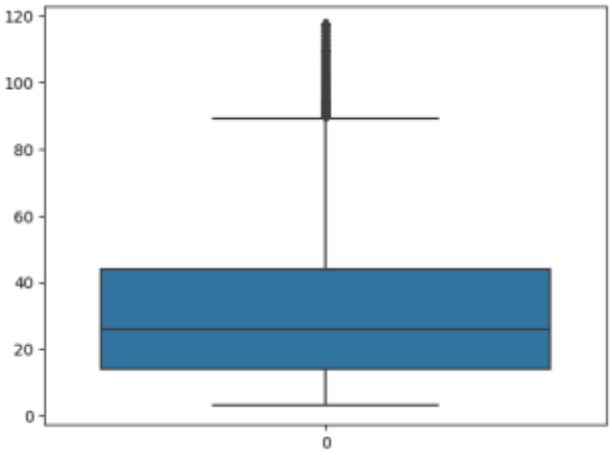
Method 1 – Z-score method

Usually, Z-score = 3 is considered as a cut-off value to set the limit. Therefore, any z-score greater than +3 or less than -3 is considered as outliers.



After using this method





According to this histogram dataset, the spread among this range is much better than the beginning dataset and According to this boxplot, outliers are decreased.

Method 2 – IQR method

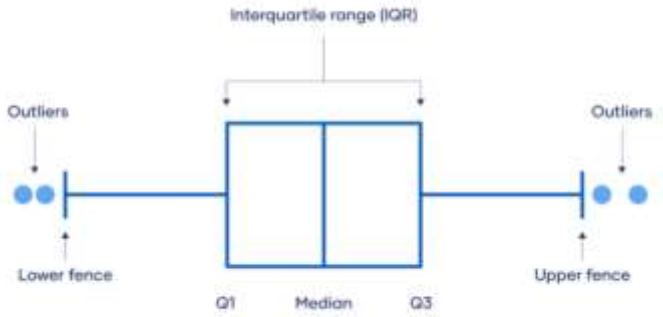
Sort your data from low to high.

Identify the first quartile (Q1), the median, and the third quartile (Q3).

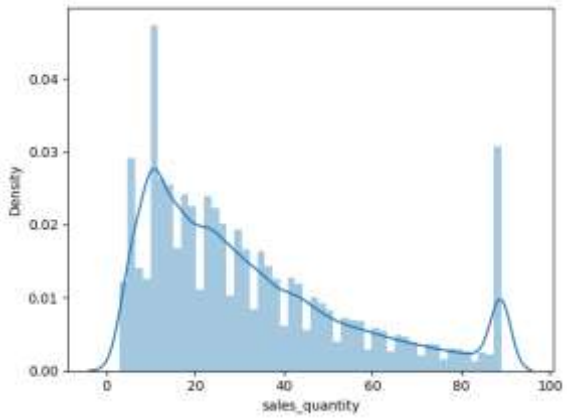
Calculate your IQR = Q3 – Q1.

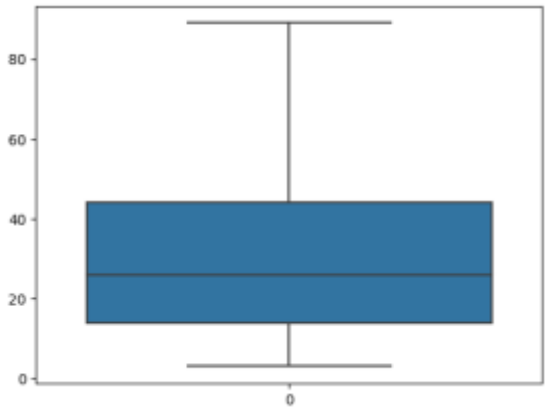
Calculate your upper fence = Q3 + (1.5 * IQR)

Calculate your lower fence = Q1 – (1.5 * IQR)



After using this method,



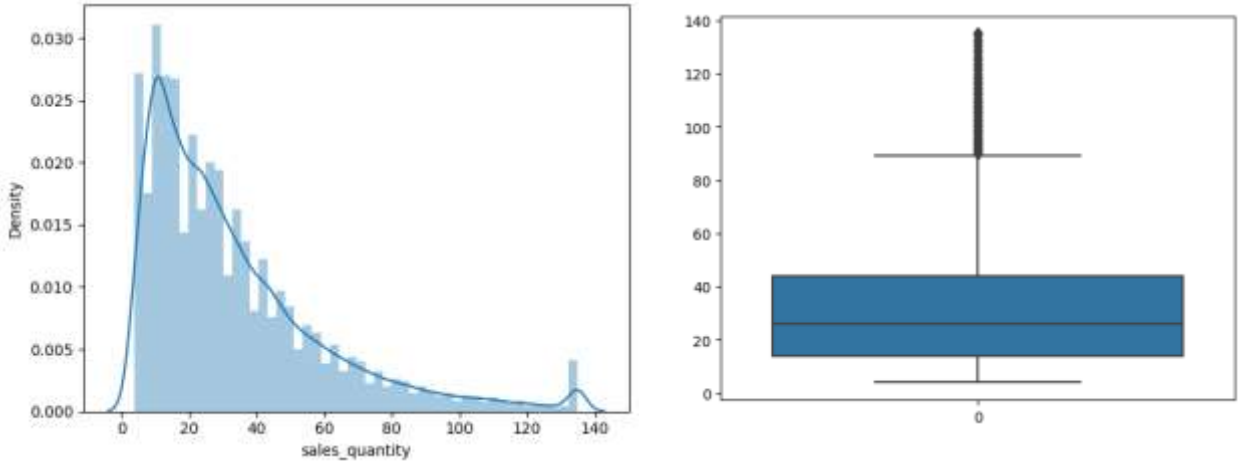


According to this histogram dataset, the spread among this range is much better than the beginning dataset and According to this boxplot, outliers are removed.

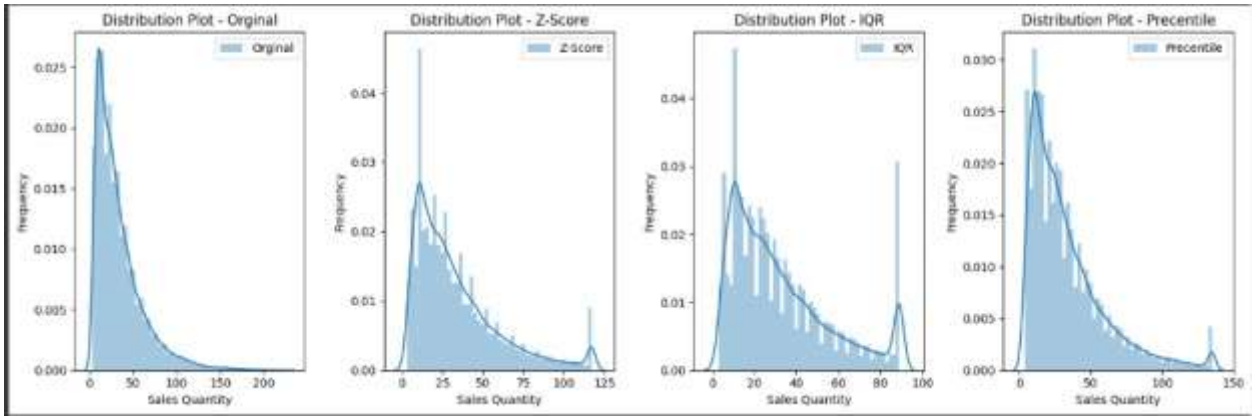
Method 3 - Percentile method

The percentile method is a technique used to treat outliers by identifying and capping extreme values based on a specified percentage threshold. It involves calculating the threshold values based on percentiles and replacing any data points that exceed these thresholds with the corresponding threshold values.

After using this method,



According to this histogram dataset, the spread among this range is much better than the beginning dataset and According to this boxplot, outliers are decreased.



Answering the Questions:

a)

#	Column	Non-Null Count	Dtype
0	week_start_date	113400 non-null	datetime64[ns]
1	expected_rainfall	113400 non-null	int64
2	freezer_status	113400 non-null	int64
3	outlet_region	113400 non-null	int64
4	outlet_code	113400 non-null	int64
5	sales_quantity	113400 non-null	int64

b)

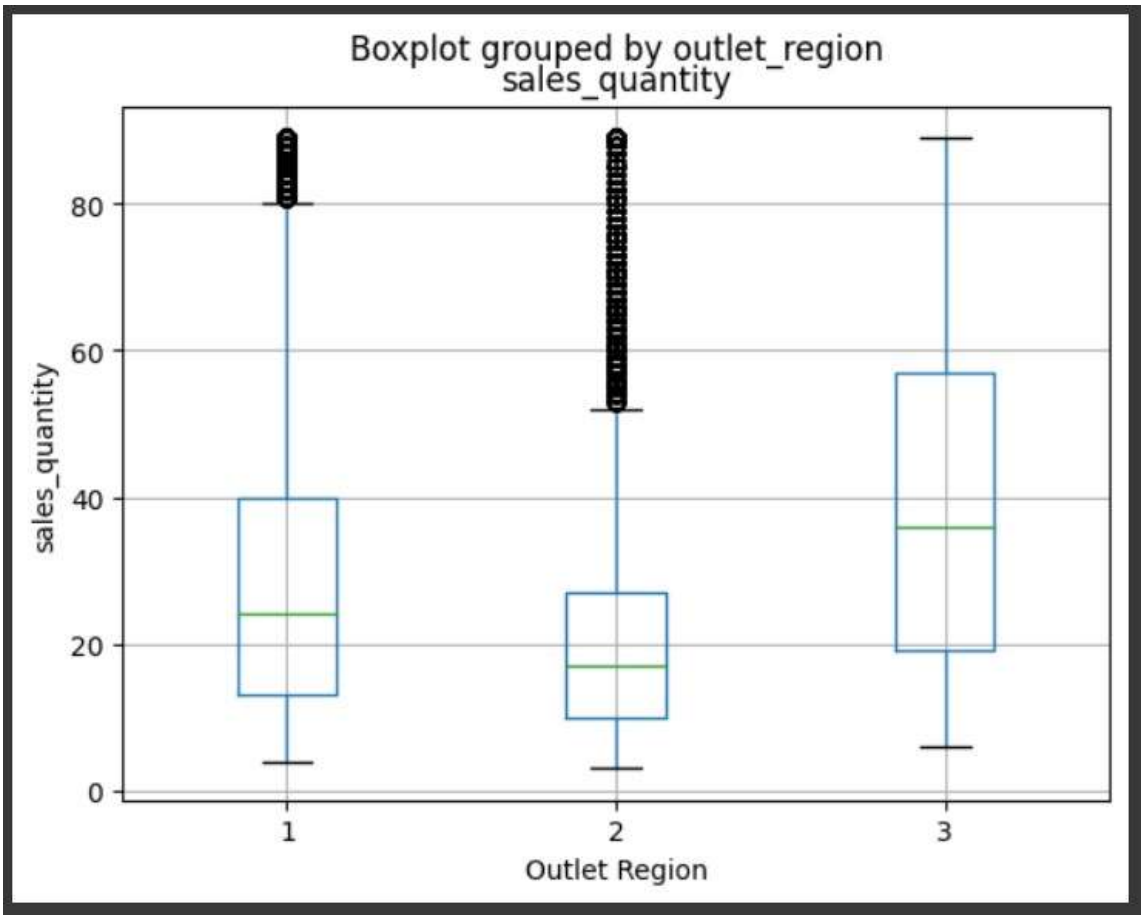
	expected_rainfall	freezer_status	outlet_region	outlet_code	sales_quantity
count	113400.000000	113400.000000	113400.000000	113400.000000	113400.000000
mean	63.893157	0.241667	1.880952	2100.500000	32.232769
std	48.293180	0.428095	0.905079	1212.440877	23.286831
min	0.000000	0.000000	1.000000	1.000000	3.000000
25%	27.000000	0.000000	1.000000	1050.750000	14.000000
50%	54.000000	0.000000	2.000000	2100.500000	26.000000
75%	87.000000	0.000000	3.000000	3150.250000	44.000000
max	199.000000	1.000000	3.000000	4200.000000	89.000000

c) sales quantity

d) Randomforest method.
Because it creates multiple trees and merges them to get accurate and statable prediction values and it reduces overfitting and develops the system or generalizes. There are more outliers in this dataset than more useful because it is more robust to the outliers.

e) 32.232769
No correlation between rainfall and total weekly sales (-0.093707)

	expected_rainfall	outlet_code	sales_quantity
expected_rainfall	1.000000	0.001255	-0.093707
outlet_code	0.001255	1.000000	0.010736
sales_quantity	-0.093707	0.010736	1.000000


Boxplot grouped by outlet_region
sales_quantity
sales_quantity
0
20
40
60
80
1
2
3
Outlet Region