1. Data Preparation

***1.1. Definition***

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data, and combining datasets to enrich data.

Data preparation is often a lengthy undertaking for data engineers or business users, but it is essential as a prerequisite to put data in context in order to turn it into insights and eliminate bias resulting from poor data quality.

For example, the data preparation process usually includes standardizing data formats, enriching source data, and/or removing outliers.

**Benefits of data preparation in the cloud**

76% of data scientists say that data preparation is the worst part of their job, but efficient, accurate business decisions can only be made with clean data. Data preparation helps:

Fix errors quickly — Data preparation helps catch errors before processing. After data has been removed from its original source, these errors become more difficult to understand and correct.

Produce top-quality data — Cleaning and reformatting datasets ensures that all data used in analysis will be of high quality.

Make better business decisions — Higher-quality data that can be processed and analyzed more quickly and efficiently leads to more timely, efficient, better-quality business decisions.

Additionally, as data and data processes move to the cloud, data preparation moves with it for even greater benefits, such as:

Superior scalability — Cloud data preparation can grow at the pace of the business. Enterprises don’t have to worry about the underlying infrastructure or try to anticipate their evolutions.

Future proof — Cloud data preparation upgrades automatically so that new capabilities or problem fixes can be turned on as soon as they are released. This allows organizations to stay ahead of the innovation curve without delays and added costs.

Accelerated data usage and collaboration — Doing data prep in the cloud means it is always on, doesn’t require any technical installation, and lets teams collaborate on the work for faster results.

Additionally, a good, cloud-native data preparation tool will offer other benefits (like an intuitive and simple-to-use GUI) for easier and more efficient preparation.

1.2. Software Used:

- Python: is a [high-level](https://en.wikipedia.org/wiki/High-level_programming_language), [general-purpose programming language](https://en.wikipedia.org/wiki/General-purpose_programming_language). Its design philosophy emphasizes [code readability](https://en.wikipedia.org/wiki/Code_readability) with the use of [significant indentation](https://en.wikipedia.org/wiki/Off-side_rule).

Python is [dynamically-typed](https://en.wikipedia.org/wiki/Type_system#DYNAMIC) and [garbage-collected](https://en.wikipedia.org/wiki/Garbage_collection_(computer_science)). It supports multiple [programming paradigms](https://en.wikipedia.org/wiki/Programming_paradigm), including [structured](https://en.wikipedia.org/wiki/Structured_programming) (particularly [procedural](https://en.wikipedia.org/wiki/Procedural_programming)), [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) and [functional programming](https://en.wikipedia.org/wiki/Functional_programming). It is often described as a "batteries included" language due to its comprehensive [standard library](https://en.wikipedia.org/wiki/Standard_library).

[Guido van Rossum](https://en.wikipedia.org/wiki/Guido_van_Rossum) began working on Python in the late 1980s as a successor to the [ABC programming language](https://en.wikipedia.org/wiki/ABC_(programming_language)) and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features such as [list comprehensions](https://en.wikipedia.org/wiki/List_comprehension), [cycle-detecting](https://en.wikipedia.org/wiki/Cycle_detection) garbage collection, [reference counting](https://en.wikipedia.org/wiki/Reference_counting), and [Unicode](https://en.wikipedia.org/wiki/Unicode) support. Python 3.0, released in 2008, was a major revision that is not completely [backward-compatible](https://en.wikipedia.org/wiki/Backward_compatibility) with earlier versions. Python 2 was discontinued with version 2.7.18 in 2020.

- Jupyter: is a project with goals to develop [open-source software](https://en.wikipedia.org/wiki/Open-source_software), [open standards](https://en.wikipedia.org/wiki/Open_standard), and services for [interactive computing](https://en.wikipedia.org/wiki/Interactive_computing) across multiple [programming languages](https://en.wikipedia.org/wiki/Programming_language). It was spun off from [IPython](https://en.wikipedia.org/wiki/IPython" \o "IPython) in 2014 by [Fernando Pérez](https://en.wikipedia.org/wiki/Fernando_P%C3%A9rez_(software_developer)) and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are [Julia](https://en.wikipedia.org/wiki/Julia_(programming_language)), [Python](https://en.wikipedia.org/wiki/Python_(programming_language)) and [R](https://en.wikipedia.org/wiki/R_(programming_language)). Its name and logo are an [homage](https://en.wikipedia.org/wiki/Homage_(arts)) to [Galileo](https://en.wikipedia.org/wiki/Galileo_Galilei)'s discovery of the [moons of Jupiter](https://en.wikipedia.org/wiki/Moons_of_Jupiter), as documented in notebooks attributed to Galileo. Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is financially sponsored by NumFOCUS.

- Numpy: a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Higuaín with contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Numarray into Numeric, with extensive modifications. NumPy is open-source software and has many contributors.

- Pandas – It is a free software library implemented for the Python language that is dedicated to data analysis. It was initially developed with the aim of managing financial data. However, its development has been escalating over time, leaving us with a multitude of useful functions not only in that sector (rojanala, 2021). This library has a "BSD" free software license and implements a multitude of operations that facilitate data processing. Currently, it is one of the most widely used language extensions in many projects that require efficient information processing. This package contains a series of features that make it one of the most used packages in data analysis (rojanala, 2021). Some of these features are:

* It includes new data structures that make it easier to handle large amounts of data. The most important data structures included for our project are the Data Frame and the Series.
* More efficient versions of structures already implemented for the storage of information (Sriraam, 2022).
* Multilevel indexing that allows powerful control over table groupings and large data collections.
* Simple support for input/output operations from files such as Excel or “.csv”
* Advanced sequence range generator that provides a wide range of possibilities for loops with a greater degree of control (Sriraam, 2022).

Thanks to the characteristics that we have just mentioned, this library adapts perfectly to the present project.

- Scikit-learn : a library implemented for the Python language and has a large community that supports its development. It is currently the most used library in this language for projects that include machine learning techniques. It provides a large number of functionalities in data mining and analysis and is also perfectly structured in its fields of action (Yang, Yuan and Li, 2020). These fields, as we can see on their website, are the following:

* Classification
* Regression
* clustering
* Data Dimension Reduction
* Model selection and generation
* pre-processing

In this way, this library allows us to implement the entire part of the project dedicated to grade prediction and clustering (Yang, Yuan and Li, 2020).

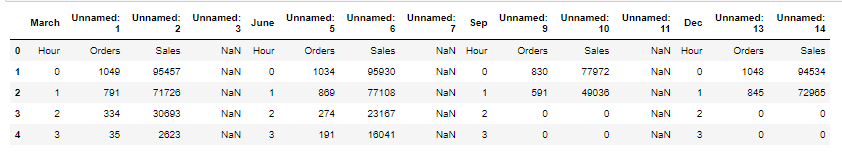
- matplotlib: a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

- Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

- Plotly: a technical computing company headquartered in Montreal, Quebec, that develops online data analytics and visualization tools. Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python, R, MATLAB, Perl, Julia, Arduino, and REST.

**1.3. What does our data look like?**

1.3.1 Let's take a look on a sample of the data:



**1.3. Checking the validity of our data**

#### Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

#### Why is Data preprocessing important?

#### Preprocessing of data is mainly to check the data quality. The quality can be checked by the following:

#### Accuracy: To check whether the data entered is correct or not.

#### Completeness: To check whether the data is available or not recorded.

#### Consistency: To check whether the same data is kept in all the places that do or do not match.

#### Timeliness: The data should be updated correctly.

#### Believability: The data should be trustable.

#### Interpretability: The understandability of the data.

#### Major Tasks in Data Preprocessing:

#### Data cleaning

#### Data integration

#### Data reduction

#### Data transformation

#### 1.3.1 Picking only the important features:

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for several reasons:

simplification of models to make them easier to interpret by researchers/users, [1]

shorter training times, [2]

to avoid the curse of dimensionality, [3]

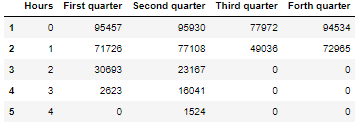
improve data's compatibility with a learning model class, [4]

encode inherent symmetries present in the input space.][8]

The central premise when using a feature selection technique is that the data contains some features that are either redundant or irrelevant and can thus be removed without incurring much loss of information. Redundant and irrelevant are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated.

Feature selection techniques should be distinguished from feature extraction, Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points). Archetypal cases for the application of feature selection include the analysis of written texts and DNA microarray data, where there are many thousands of features, and a few tens to hundreds of samples.

So, let's choose only the features that we're going to need in our research:



#### 1.3.2 Creating a table containing the quarters and their sales per hour:

#### 

#### 1.3.3 Adding total sales over a year:

#### 

#### 1.3.4 Creating a table that has the sales for each quarter:

#### 

2. Analysis:

## **What Is Data Analysis?**

Although many groups, organizations, and experts have different ways to approach data analysis, most of them can be distilled into a one-size-fits-all definition. Data analysis is the process of cleaning, changing, and processing raw data, and extracting actionable, relevant information that helps businesses make informed decisions. The procedure helps reduce the risks inherent in decision-making by providing useful insights and statistics, often presented in charts, images, tables, and graphs.

A simple example of data analysis can be seen whenever we take a decision in our daily lives by evaluating what has happened in the past or what will happen if we make that decision. Basically, this is the process of analyzing the past or future and making a decision based on that analysis.

It’s not uncommon to hear the term “[big data](https://www.simplilearn.com/tutorials/big-data-tutorial/what-is-big-data)” brought up in discussions about data analysis. Data analysis plays a crucial role in processing big data into useful information. Neophyte data analysts who want to dig deeper by revisiting big data fundamentals should go back to the basic question, “[What is data](https://www.simplilearn.com/what-is-data-article)?”

## **Why is Data Analysis Important?**

Here is a list of reasons why data analysis is such a crucial part of doing business today.

* Better Customer Targeting: You don’t want to waste your business’s precious time, resources, and money putting together advertising campaigns targeted at demographic groups that have little to no interest in the goods and services you offer. Data analysis helps you see where you should be focusing your advertising efforts.
* You Will Know Your Target Customers Better: Data analysis tracks how well your products and campaigns are performing within your target demographic. Through data analysis, your business can get a better idea of your target audience’s spending habits, disposable income, and most likely areas of interest. This data helps businesses set prices, determine the length of ad campaigns, and even help project the quantity of goods needed.
* Reduce Operational Costs: Data analysis shows you which areas in your business need more resources and money, and which areas are not producing and thus should be scaled back or eliminated outright.
* Better Problem-Solving Methods: Informed decisions are more likely to be successful decisions. Data provides businesses with information. You can see where this progression is leading. Data analysis helps businesses make the right choices and avoid costly pitfalls.
* You Get More Accurate Data: If you want to make informed decisions, you need data, but there’s more to it. The data in question must be accurate. Data analysis helps businesses acquire relevant, accurate information, suitable for developing future marketing strategies, business plans, and realigning the company’s vision or mission.

## 2.1. Sales per hour:

##### 

##### *Before jumping to the charts, let's first know What is a line chart:*

A line chart is a graphical representation of an asset's historical price action that connects a series of data points with a continuous line. This is the most basic type of chart used in finance, and it typically only depicts a security's closing prices over time. Line charts can be used for any timeframe, but they most often use day-to-day price changes.

#### A heat map (or heatmap) is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space. There are two fundamentally different categories of heat maps: the cluster heat map and the spatial heat map. In a cluster heat map, magnitudes are laid out into a matrix of fixed cell size whose rows and columns are discrete phenomena and categories, and the sorting of rows and columns is intentional and somewhat arbitrary, with the goal of suggesting clusters or portraying them as discovered via statistical analysis. The size of the cell is arbitrary but large enough to be clearly visible. By contrast, the position of a magnitude in a spatial heat map is forced by the location of the magnitude in that space, and there is no notion of cells; the phenomenon is considered to vary continuously.

#### "Heat map" is a relatively new term, but the practice of shading matrices has existed for over a century.

#### From these three graphs we can come up with the knowledge of:

## **The sale's trend is identical for all quarters so there is no abnormal change concerning sales between quarters.**

## **The sales are null between 4 am and 12 pm which is normally the time during which the stores close.**

## **It starts increasing slowly till 18 pm.**

## **The rate of sales increase rapidly from 6pm till it reaches it peak around 20.**

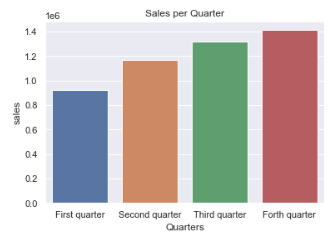
## **Then it decreases late at night around 23 pm and 0 am till the stores close at 2 AM.**

## **The sales at 2 am became null starting from the third quarter, we can presume that the stores started closing at 2 am**

## **The last figure shows the total number of sales per hour during all year. we see the overall trend is the same as the trend of sales per each quarter**

here we'll be showing a bar chart, but let's first know what a bar char is:

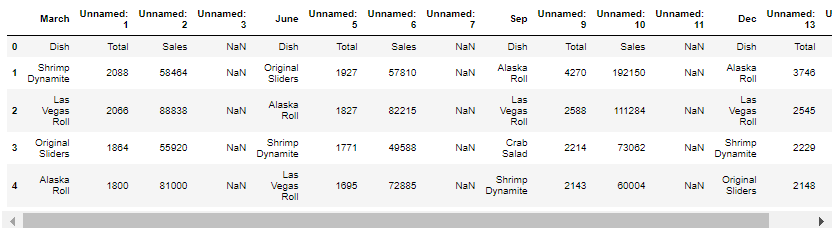
A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.



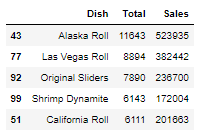
We see a consistent increase in sales per quarter which is a sign of a healthy business.

2.3. Popular Dishes analysis:

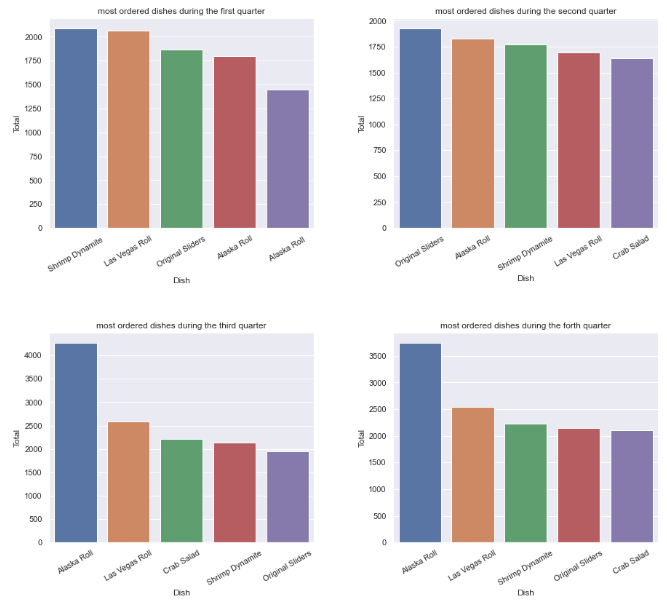
Let's first show a sample of the data before preprocessing:



### Creating a table containing the number of sales over the year.



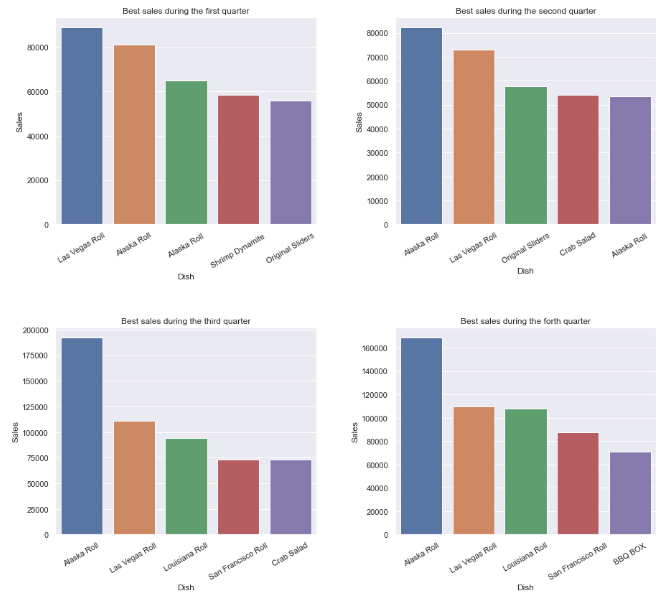
### 3.4. Top dishes per trimester:



The 'Alaska Roll' / 'Las Vegas ROLL' / Original Sliders / Shrimp Dynamite are the dishes that remain in top 5 orders during all quarters.

We see the Alaska Rolls inserted two times during the first and second quarter. The most plausible explanation is an error in entering data because we can see that the sales in the third and fourth quarters equal almost the double of sales during the two firsts quarters.

If the last hypothesis is right, the Alaska Roll is the most popular dish by a large margin in comparison with other dishes during all quarters.

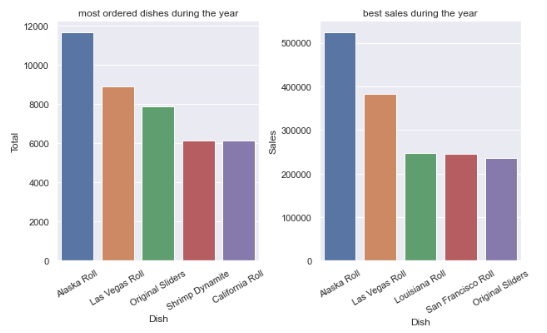


The 'Las Vegas Roll' and 'Alaska Roll' are the dishes that remain in top 5 sales during all quarters.

We see that we have the Alaska Rolls inserted two times during the first and second quarter. The most plausible explanation is an error in entering data because we can see that the sales in the third and fourth quarters equal almost the double of sales during the two firsts quarters.

If the last hypothesis is right, the Alaska Roll is the most popular dish by a large margin in comparison with the dishes during all quarters.

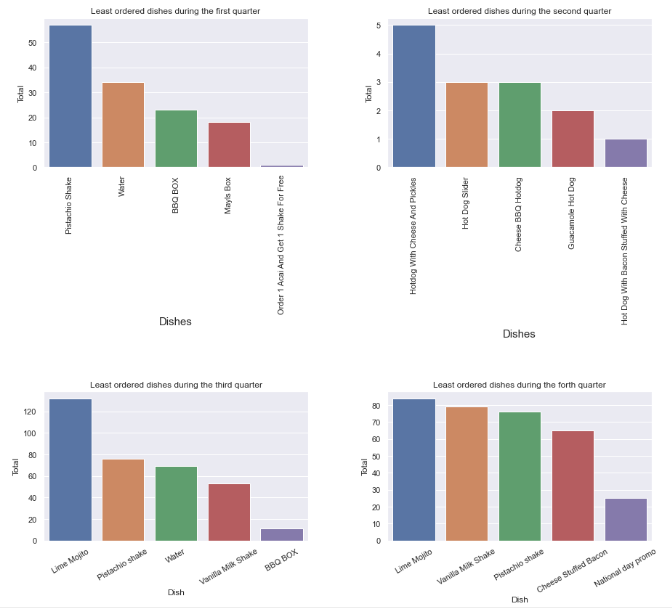
### 3.5. Top dishes over the year:



The Alaska Rolls and the Las Vegas Roll are the most ordered and sold dishes.

The California Rolls/Shrimp Dynamite/California Roll are more often ordered than the Louisiana Rolls but the sales of the Louisiana are superior. It is probably due to the price of the dishes. For a better comparison, the preparation costs of each dish should be included to compare the profit of the dishes rather than just the sales.

### 3.6. Bottom dishes per trimester:



#### 

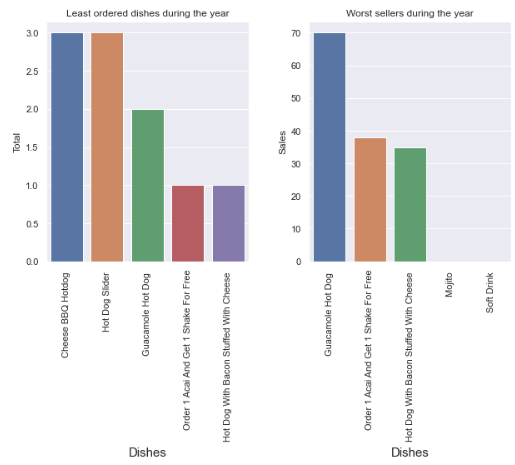
The least ordered/sold dishes change every quarter.

The least ordered/sold food are side dishes/drinks | Promos that remain for a limited time | Hot dogs

The number of orders/sales during the second quarter is much lower than the other quarters

Mojito/Soft Drinks has 0 as a price and they were present only in the second quarter

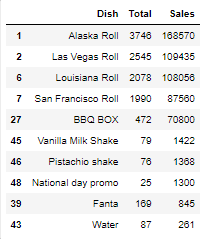
### 3.7. Bottom dishes over the year:

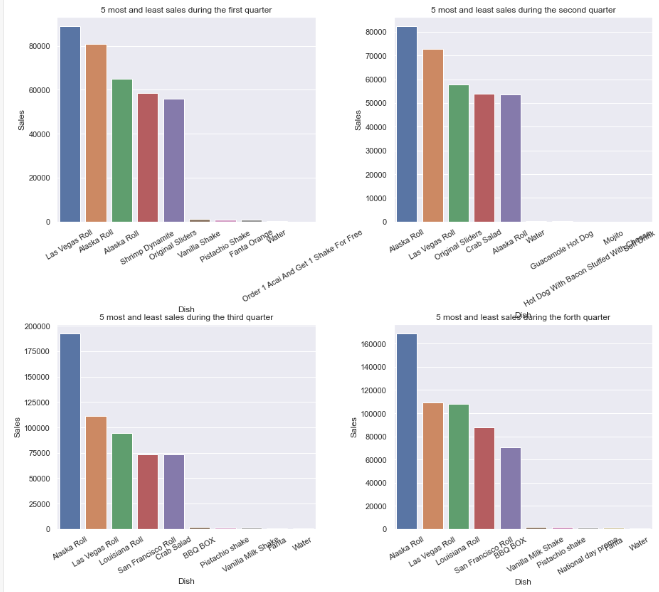


#### The least ordered/sold food are side dishes | drinks | Promos that remain for a limited time | Hot dogs.

### 3.8. Comparison between the dishes with the highest and lowest sales/orders:

#### We first need to create a table containing the dishes with highest and lowest sales:

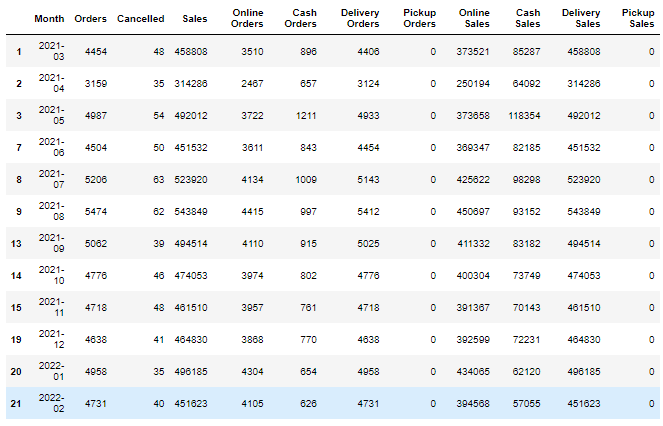




**We can see a huge difference in sales between the top and bottom dishes.**

### 3.9. Orders' distribution:

### Data after preparation:

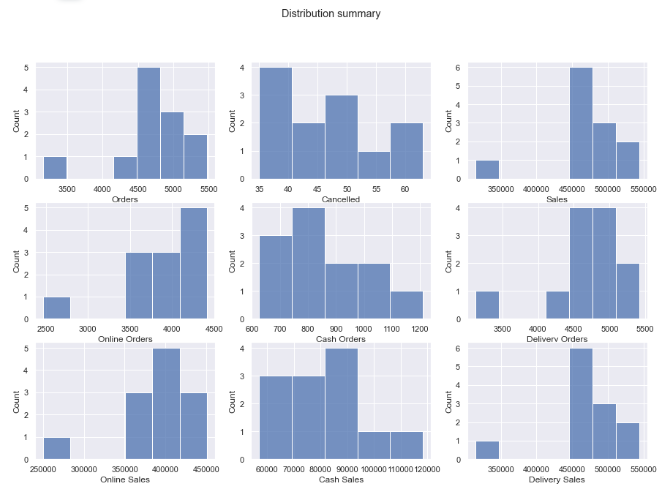


## What Is a Histogram?

A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a [bar graph](https://www.investopedia.com/terms/b/bar-graph.asp), the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

### **KEY TAKEAWAYS**

* A histogram is a bar graph-like representation of data that buckets a range of classes into columns along the horizontal x-axis.
* The vertical y-axis represents the number count or percentage of occurrences in the data for each column
* Columns can be used to visualize patterns of data distributions.
* In trading, the MACD histogram is used by technical analysts to indicate changes in momentum.
* The MACD histogram columns can give earlier buy and sell signals than the accompanying MACD and signal lines.



**Orders**:

Most orders are between 4400 and 5500.

There is an outliner with 3500 orders which was during April.

It is recommended to find the cause of low orders in April

**Cancelled**:

The cancels orders range from 35 to 60 cancels per month

Approximately equally distributed without an outliner. It would still be recommended to find the causes of the higher number of cancels during the second quarter (June: 50 cancels | July: 62 | August: 63)

**Online Orders**:

Most Online orders are between 3500 and 4500

**Cash orders**:

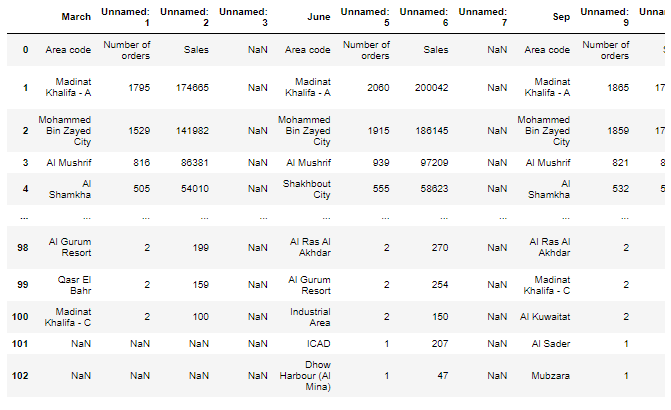
Most Cash orders are between 800 and 1100

We see that most orders are online rather than cash orders

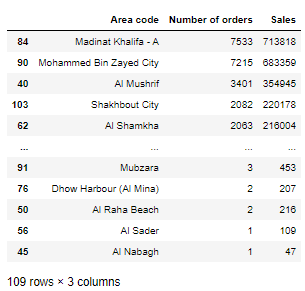
**Same conclusions for Sales because Amount of Sales= 10 x Number of Orders**

## 3.10. Areas' Analysis:

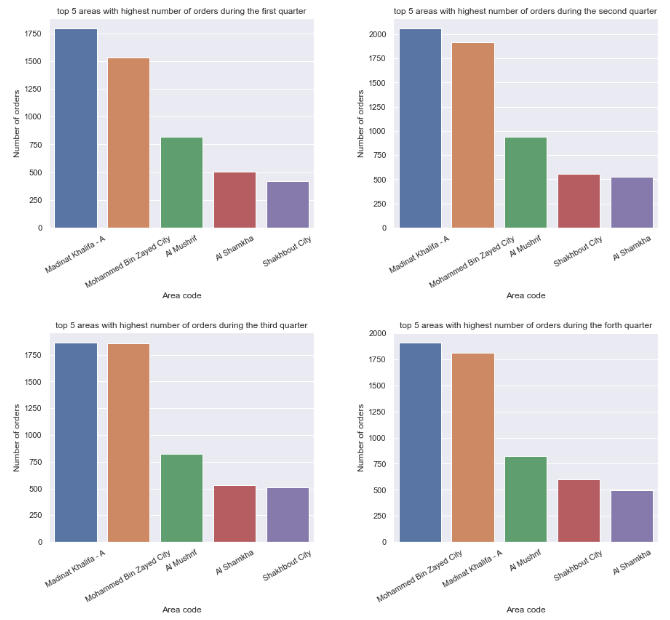
Showing a sample of the data before preprocessing:



Showing a sample of the data after preprocessing:

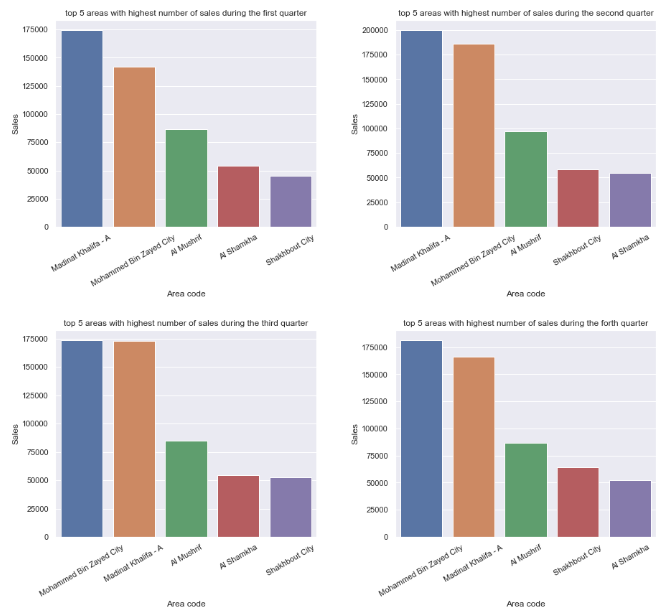


## 3.11. Top areas per trimester:



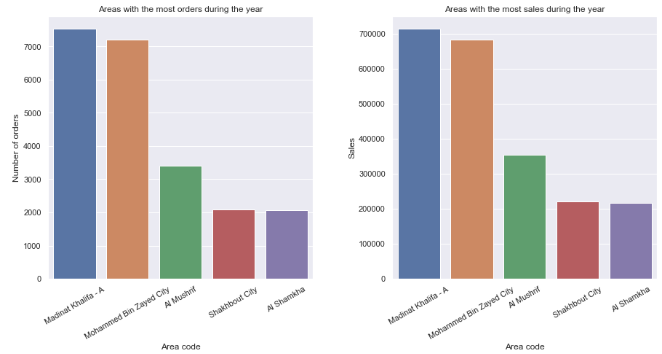
The 'Madinat Khalifal' / 'Mohammed Bin Zayed City' / Al Mushrif / Al Shamkhae / Shakhbout City are the top 5 areas with the most number of orders during all quarters.

'Madinat Khalifal' / 'Mohammed Bin Zayed City' are in the top 2 and 'Al Mushrif' remains in third during all quarters

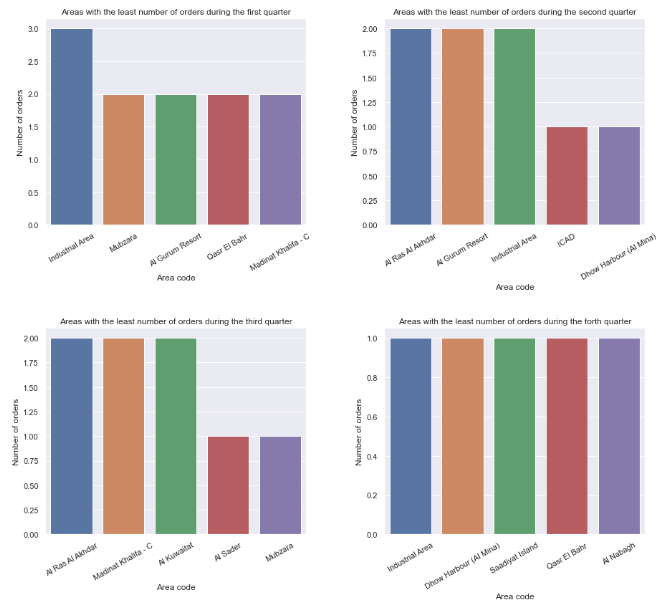


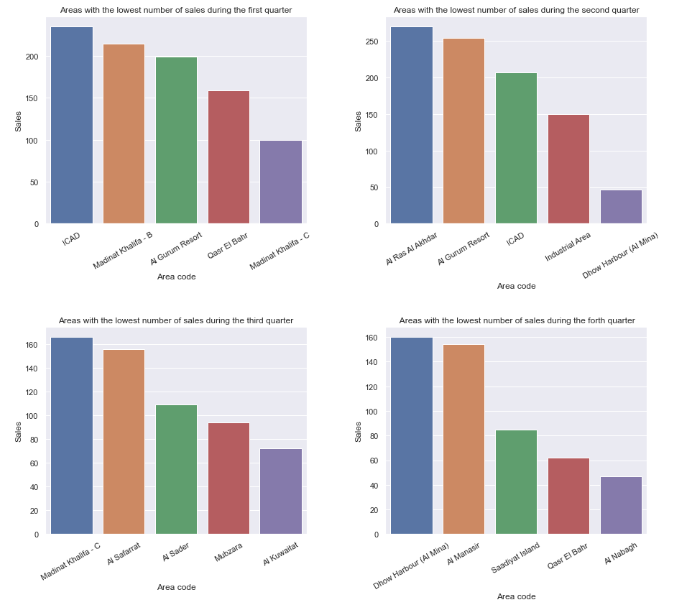
Same conclusions for Sales because there is a big correlation between Sales and Orders (1 order = 10 unit of Sales)

## 3.12. Top areas over the year:

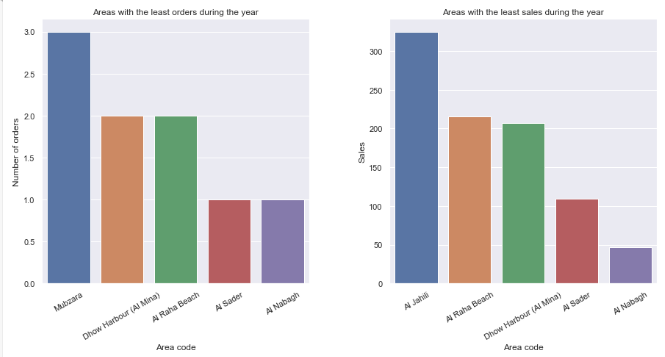


### 3.13. Bottom areas per trimester:





### 3.14. Bottom areas over the year:

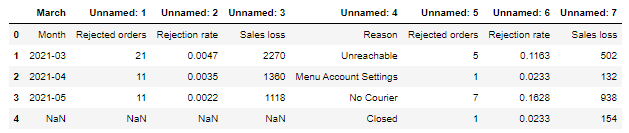


Al Nabagh / Al Sader / Al Raha Beach / Dhow Harbour / Mubzara have the worst number of orders over last year ranging from 1 to 3 orders.

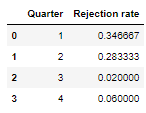
Al Nabagh / Al Sader / Al Raha Beach / Dhow Harbour / Al Jahli have the worst number of sales over last year ranging from 50 to 350 total sales.

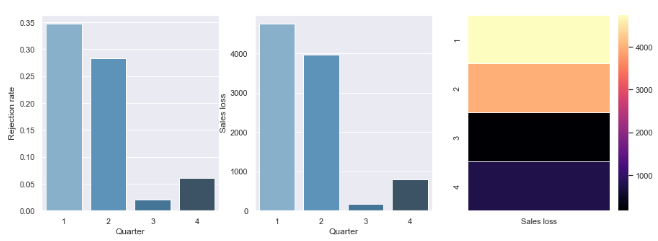
### 3.15. Rejected orders:

Data before preprocessing:



Data after preprocessing:



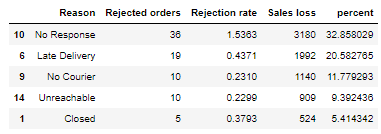


The first and second quarters have the most Sales loss/ Rejection rate followed by the fourth quarter then finally the Third one.

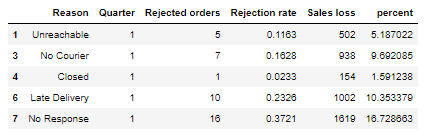
There is a big difference between the losses during the first and second quarters in comparison with the third and fourth quarters.

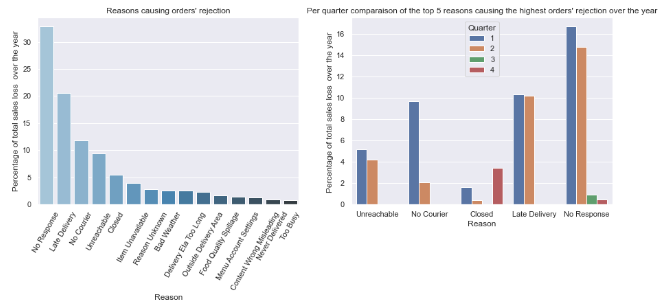
### 3.16. Reasons of rejection:

Data before preprocessing:



Data after preprocessing:





In the first graph, we see the reasons causing the orders' rejections.

We see that the top 5 reasons causing the highest amount of losses are: 'No Response','Late Delivery','No Courier','Unreachable','Closed'

In the second graph, we see the percentage of total loses caused by the top 5 reasons causing the highest amount of losses during the year.

We see that there is a big decrease in the loses in general specially the ones caused by the responses/Delivery/Courier betsween the first and second half of the year.

***Conclusion:***

In this research period, we have covered several important topics that aim to create a prototype of a tool that allows the teacher to monitor his class. This monitoring will be done by viewing statistics, grouping branding and investment with similar behaviors and predicting grades. Although the development will focus on the specific subject of Interactive Systems Development (ISD), the system will consist of a modular architecture that it to be easily expanded (Ding and Guo, 2018). In this way, in the future work section you can adapt to new subjects and add new features. For this, the project will be based on the visualization of the data in a pleasant format for the user, the creation of behavior models of the branding and investment who study a certain subject and the use of automatic learning techniques to try to predict the performance of the restaurant in the UAE

Our steps were:

1- Preparing the data to be read and displayed in a software and research environment at the same time using techniques such as Python, Jupyter and other necessary techniques that helped us complete the project.

2- We also, using those previously mentioned techniques, cleaned and pre-processed the data so that it would be ready for subsequent analyzes

3- Using the skills of analysis and technology, we created many statistical drawings that answered very important topics that our research was based on, some of those topics were:

- Sales per hour

- Popular Dishes analysis

- Top dishes per trimester

- Top dishes over the year

- Bottom dishes per trimester

- Bottom dishes over the year

- Comparison between the dishes with the highest and lowest sales/orders

- Orders' distribution

- Areas' Analysis

- Top areas per trimester

- Top areas over the year

- Bottom areas per trimester

- Bottom areas over the year

- Rejected orders

- Reasons for rejection