Roskilde University

Computer Science and Informatics

How to predict future sales by analyzing a big data set with the help of machine learning algorithms?

Students:

Mohammad Mosiur Rahman
Hunor Vadasz-Perhat
Mohammad Azizul Huq

Supervisor:

Anders Lassen

Introduction **3**
Problem Statement 3
Project of exploration 3

Methodology **4**
CRISP-DM model 4
Project plan 7

Analysis **8**
Project Plan 8
Business understanding 8
Determining Business Objective 8
Assess Situation 8
Data science 8
Math, statistics and domain knowledge 9
Programming language 9
Determine Data Mining Goals 10
Produce Project Plan 10
Data Understanding 10
Collect Initial Data 10
Describe Data 10
Structured versus unstructured data 11
Quantitative versus qualitative data 12
Is our data quantitative or qualitative 12
Level of measurement 13
The nominal level 13
The ordinal level 13
The interval level 14
The ratio level 14
Types of measurement of our data 14
Additional information 14
Explore data 15
Negative values 16
NaN values 17
Number of orders 17
Data Preparation 19
Clean data 20
Data Visualization 20
Select Data 23
Modelling **23**

Select modelling technique 23

Linear Regression 25

Types of Linear Regression Relationships 27

Example of How to find the Linear Regression Equation: 27

Find the Coefficient of Determination 28

Multiple Linear Regression 29

Implementation and Result 30

ARIMA Model 32

Box-Jenkins Methodology 33

Method and Feature 34

ARIMA Model 35

Conclusion 40

Bibliography 42

Introduction

Small business and large enterprises all want to get on the latest train which is machine learning. Big names such as Amazon and Google already use the technology for their own advantage. It might be that smaller companies fear that machine learning is out of their reach but it does not have to be that way at all.

Simply put machine learning (ML) is a process a software application uses to actively learn from the data that is imported into it. It is as the way us humans would use our own past experiences as a part of a particular learning process. But ML is no longer used for specialized research projects by a data science team. Enterprises now make use of ML to get an in-depth business intelligence (BI) and predictive analytics from the ever increasing amounts of data. But business intelligence in itself is a complex field as well. It refers to a process where business related data needs to be acquired, stored and analyzed accordingly. This process has several different aspects such as analytics, predictive modelling, data mining etc.

There is a lot of information companies can get from online customer purchasing behavior. But ML technology brings an important improvement in the process of understanding the target audience and its needs. Every piece of data that is collected from personal profiles are a valuable source of information that can help the company to predict how a new product would be accepted on the market or would tell more about which qualities should be further developed. Furthermore as ML is able to store and use data that has been collected from every possible business aspect it is possible to create automation of various processes.

Problem Statement

How to predict future sales by analyzing a big data set with the help of machine learning algorithms?

In our project we would like to provide a clear-eyed look at some aspects of what machine learning is and how it can be used today from a business perspective. More specifically, the goal of the project is to predict the amount of products that would be sold in a company within a period of time. In order to reach this goal, the group will choose a related data set and analyze the data set by using different related machine learning algorithms.

Project of exploration

The team members did not have previous experience with data science, data analysis and machine learning. Therefore the group decided not to carry out a project that is traditional and carefully planned and followed to its finest details but to look at the process as an exploration.

Because of the exploratory nature of our project in the beginning we spent a lot of time trying to formulate the problem we want to answer and acquiring the necessary theoretical and practical knowledge. In the middle of our project we needed to revisit our goals. We decided to focus on 2 models for prediction namely Multiple Linear Regression and ARIMA model. Multiple linear regression was our primary focus. The reason for choosing ARIMA model was to get to know another model and see the advantages and disadvantages of these models.
The group managed to go through the project plan that included the steps that would be necessary to conduct a data science project and answer our problem formulation to a certain degree which would include the implementation of models and producing results. But we have to state that our original goals needed to be adjusted during the process.

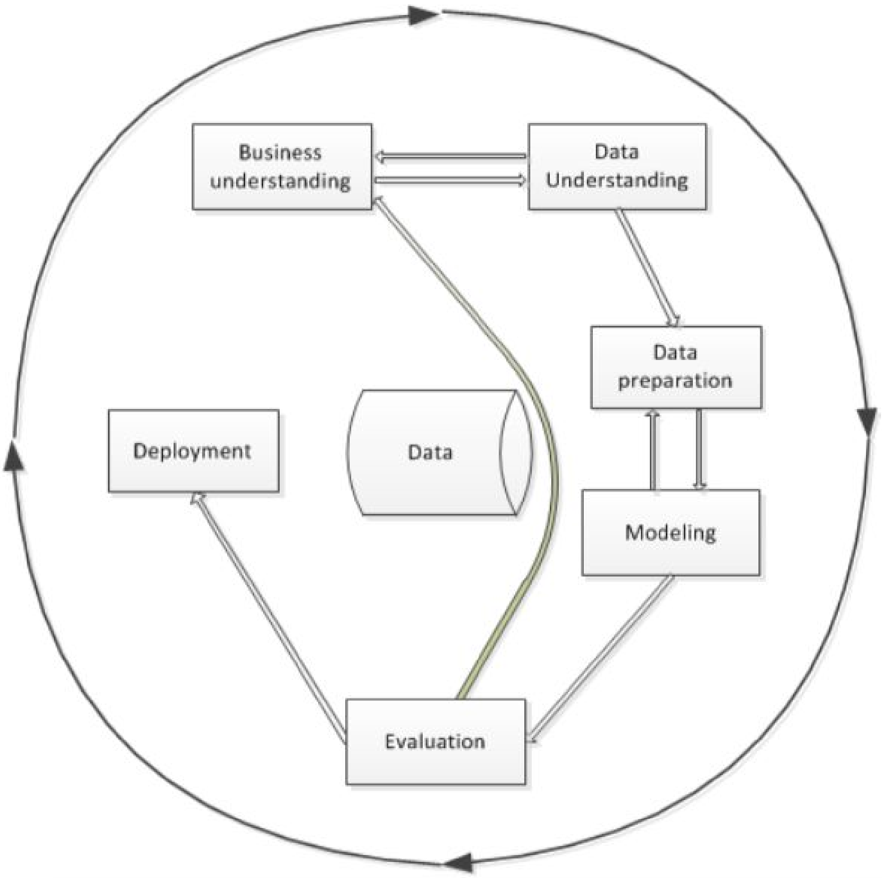
Methodology

We wanted to work out a structured approach to plan and execute our project. As a roadmap and inspiration we used the CRISP-DM model which stands for cross-industry process for data mining. It is said to be a robust and proven methodology. "*...CRISP-DM proposes an iterative process flow, with non-strictly defined loops between phases, and overall iterative cyclical nature of DM project itself...*" [7] The model has an idealised sequence of events but the tasks can be performed in a different order and it will be often necessary to go back to previous tasks and repeat certain actions and to alter them as well. This flexibility is in line with our exploratory approach towards our project.

In the following we will look deeper into the model and the tasks that are related to each phase. Finally we will give our own project plan that is based upon the CRISP-DM model. This project plan will serve as a guide for our analysis later on.

CRISP-DM model

The Cross Industry Standard Process for Data Mining (CRISP-DM) was a project in 1996. The goal was to define and create a DM (data mining) process model that could be used by the industry. The two requirements were to be industry and tool neutral.



Phases of the CRISP-DM reference model [7]

Each phase has its own tasks. And hereby we would like to give a list of the phases and the tasks that belong to each phase. The article provides a list of of phases and its related tasks and we would like to further give our understanding of them.

CRISP-DM model [6]:

Phase 1. Business understanding

The first phase of the model is to understand what the team wants to reach from a business perspective.

- Determine business objectives: describe the primary objectives from a business perspective. For example the goal might be to keep current customers by predicting when they are ready to move to a competitor.
- Assess situation: put together a glossary of terminology, consider areas such as the inventory resources (for example personnel, computing resources, software), risk analysis, costs and benefits
- Determine DM goals: describe the intended outputs that would fulfill the business objectives and define the criteria for a successful outcome
- Produce project plan: describe a plan for reaching the goals. The plan should be specific about steps and include selection of tools

Phase 2. Data understanding

The second stage requires to acquire the data in the project resources for understanding

- Collect initial data: collect data that fits with the goals and the source of the data should be reliable
- Describe data: describe the data including the format, quantity, identities of fields and other features if available. Also should evaluate if the data acquired satisfies requirements
- Explore data: explore the data and describe results including the findings for missing values, null values, contradicting data formats etc. It is a good idea to use graphs and plots to visualize results
- Verify data quality: address if data is complete, correct

Phase 3. Data preparation

The third stage is where team needs to decide which part of data to use

- Select data: list the data to be used or excluded and give a reasoning for that
- Clean data: prepare the data quality to the level that is required by the analysis. This may include selecting a subset of data for example
- Construct data: produce derived attributes or new records
- Integrate data: methods (merge, aggregate) are used to combine multiple databases, tables or records
- Format data: choose the appropriate format for each data entity

Phase 4. Modelling

This step involves the use of statistical and machine learning models and the implementation of mathematical validation metrics in order to quantify the models and their effectiveness.

- Select modelling technique: document the modelling technique and consider the assumptions the model implies
- Generate test design: describe the intended plan for training, testing and evaluating the models
- Build model: run the modelling tool on the data set to create one or more models
- Assess model: rank the models according to the evaluation criteria. Business objectives and success criteria should be considered

Phase 5. Evaluation

At this step it must be possible to conclude the results in a digestible format. The accuracy and generality of the model should be evaluated

- Evaluate results: summarize assessment results according to business success criteria and the models that meet the criteria can be approved
- Review process: summarize the activities that were missed and the ones that should be prepared
- Determine next steps: list further actions and describe how to carry on

Phase 6. Deployment

At the final stage the results should be evaluated and the plans for deployment should be determined

- Plan deployment: summarize the deployment strategy and its steps
- Plan monitoring and maintenance: summarize the monitoring and maintenance strategy including its steps
- Produce final report: write a final report that would include the previous deliverables containing the results in an organized manner
- Review project: assess what went good, what went wrong, what needs to improved

Besides making our own plan it was important to clear out how we looked upon our project. This meant that we spent a considerably crucial time on figuring out in what a project is in general and what a project is not. In order to do so we brainstormed on the various aspects that would describe a project.

First we distinguished between a project and an operation. Our conclusion was that a project would have the following characteristics:

- It is unique by nature and not a routine task
- It has a specific goal
- It delivers a certain type of product at the end of the project
- It contains factors that are not known and therefore risks can show up
- It has a limited amount of period (and usually with a deadline)

As for the project type our project would fit into the category of an analysis project in which the group would analyze the given topic and its various aspects from different angles to get a clearer overview. Therefore we could conclude that our project is rather a development oriented regarding the project type.

The CRISP-DM model was the backbone for our working process and the project management took place accordingly. The frequent supervisions helped us to crystalize our goals, to offer possible solutions to certain upcoming hardships, to question our thought processes, to summarize our thoughts and to put us on the track when we felt lost.

Project plan

The project plan that we made according to the CRISP-DM model would look the following:

- Business Understanding: in this phase our goal is to understand the business and data mining objectives, to give a brief overview of the complexity of the data science field and finally to refer to how we made our project plan for the report
- Data Understanding: in this phase our initial goal is to better understand the data set we have. Our approach is to first provide a theoretical knowledge then apply that knowledge on our data set. Later we would look into specific cases such as missing values for example.
- Data preparation: after understanding the data the next step is the data preparation. The goal is to prepare the data for the future modelling steps. Hereby we would use the

findings from the previous steps and would apply them. An example would be to treat missing values in a way that would be appropriate for the model we want to use.

- Modelling: once the data is prepared and selected the next step is to build the model. In this step we can start training and testing the machine learning models. Hereby it is possible to experiment with different learning algorithms to get to know which algorithm works the best.

Analysis

In this section the goal is to attend to give a solution to the problem statement. The structure of the section is based upon our project plan. The general guideline is the following: the plan consists of phases and each phase has its own tasks to perform. Our aim is to first give a theoretical background that is connected to the particular task then to apply that theoretical knowledge on our data set and showcase our findings. Note that this does not apply to all phases or steps.

Project Plan

Business understanding

Determining Business Objective

The business objective of the project is to analyze the given data set and to be able to predict the daily amounts of sold products of the company in particular for the upcoming year.

Assess Situation

In this section we will give a brief overview of the concepts and their definitions that we think is important to understand. Data science is an interdisciplinary field. Different terms are used and some of them are used interchangeably. Partly due to the fact that the field is growing fast and that new terms are added meanwhile the meaning of old terms are altered. This can cause confusion. In order to avoid this we use this section to clear the confusion.

Data science

Data science is a multidisciplinary field that uses algorithms, scientific methods, processes and systems to get knowledge from data [8]. This data can be structured and unstructured. Data science attempts to connect various fields with each other. These fields are statistics, machine learning, data analysis and the methods that are related. It is also using theories from mathematics, statistics and computer science.

Math, statistics and domain knowledge

Math and statistics are used for formulas and equations to make analysis of different sort. Domain knowledge would refer to the problem domain which could be business, finance or medicine. The range is big and involvement of professional of the particular domain could be necessary. And finally the programming part would mean that the team working on a project has the ability to use a programming language.

Programming language

For programming language we chose Python. But why Python? Data science can be done in multiple languages. Such languages are Julia, R or Python. These are just a few of the languages that are available for us.

The math used in data science is statistics and probability. The domains of mathematics are used to make models. A model is an organized relationship between the elements of the data set. The math can vary between basic algebra and advanced probabilistic and statistical modeling. A probabilistic model uses probability to find relationships between elements of data including randomness. Meanwhile statistical model would use general statistical propositions to make relationships between data.

It is not necessary to be a master a given domain. Although it is important to have a basic understanding of the domain in particular and more importantly to be able to possess good communication skills to represent the results.

But why Python? Data science can be done in multiple languages. Such languages are Julia, R or Python. These are just a few of the languages that are available for us.

Here we want to point at a variety of reasons:

- It is easy to learn Python even with limited primer programming knowledge
- Python is commonly used both in the academic world as an introductory language and also widely used in the industry
- There is a detailed and up-to-date documentation that is available besides the massive amount of tutorials, blogs
- There are libraries and modules that are specifically built for data science
- The libraries and modules are easy to pick up and powerful. Here we can think of:
- Pandas
- Numpy
- Scipy

Python has the power of the other commonly used programming languages such as C++, Java or C# and due to its simple and easy to use syntax it is a valid choice that can satisfy the goal of the project.

Machine learning combines computing power with sophisticated algorithms in order to discover the relationships in the data set and make models accordingly.

The goal of this section was to give a general overview of the sophisticated nature of the data science field and to look at some basic terminologies. As we can conclude data science is a multidisciplinary field and it is crucial to have the proper and necessary exposure to the theoretical background or coding practices and have a basic understanding of the given domain otherwise we risk that oversimplify the task that we are trying to model. This might have big and serious negative consequences for the company, organization or customers of a service.

Determine Data Mining Goals

Data mining is the process of discovering the possible patterns in a big data set [8]. It is an interdisciplinary subfield of computer science and statistics thus it is heavily dependent on machine learning and statistical models.

Besides using data mining to get to know patterns our group is also aiming at using predictive analytics. Predictive analysis is using statistical techniques from predictive modelling, machine learning and data mining. From a business point of view such predictive models can be used to find patterns in different data sources such as transactional and historical.

Produce Project Plan

In order to get a full description of the Project Plan and how it was made please refer to the Methodology section where we described the process and our considerations in great detail.

Data Understanding

In this section we will look at our data from different angles. Our approach is to combine theory with practice. First we dive into the theoretical background where we understand terms and their usage then apply this knowledge on the data in order to better understand and discover our data.

Collect Initial Data

In the process of choosing a data set that is related to online retail, our goal was to base our choice on a set that satisfies the goals of the project. The set should fulfil the basic requirement which was that it should contain transactions of the sold units from a shop. Besides, it should be open source, publicly available and no missing, inconsistent data should be present.

Taking into consideration the various factors that played a role in choosing a data set we used the UCI machine learning repository. It is a website of “collections of databases, domain theories and generators that are used by the machine learning community” [4]. The website is used by many student and teachers since 2007 when the site was established. It is also worth to notice that the site was in the top 100 most cited computer science papers during these years.

Describe Data

Our data set is an online retail data set. It is by nature a transactional data set which contains all the transactions occurring between a period of time for the UK based and registered non-store online retail. The source is the Social Engineering department of the London South Bank

University. Regarding sales the company mainly sells unique gifts for various occasions. It is also worth to mention that many of the customers are wholesalers[1].

In the process of describing the data we looked at the following:

- Is the data structured or unstructured?
- Is our data quantitative or qualitative?
- Level of measurement

Structured versus unstructured data

There are two basic types of data which are structured and unstructured. Whether the data set is structured or unstructured is the first consideration. A structured data is usually organized in tables more precisely columns and rows. Meanwhile unstructured data does not follow any organized manner or hierarchy. For example a database of scientific observations that is recorded by scientists could be considered a structured data format. On the other hand data in text forms that would appear on a social media platform such as posts could be considered unstructured.

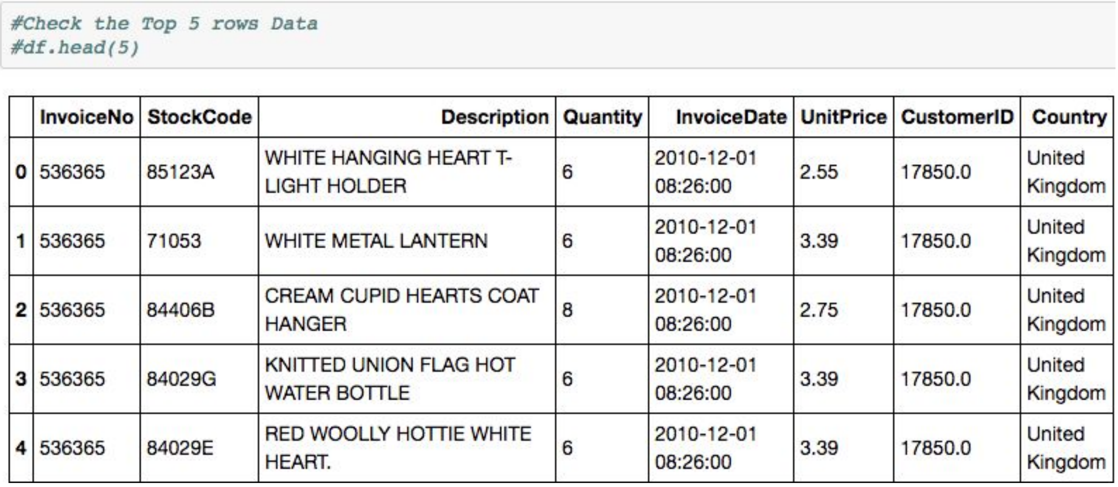
Statistical models were made with structured data in mind. Since majority of the data in the world is not structured it is worth to look at the so called preprocessing where a certain type of structure is applied to the data.

Here we would like to give few examples of data preprocessing:

- Word or phrase count
- Relative length of text
- Special characters

Since the data set we chose is structured into tables and columns we will not go into detail in these steps. We just wanted to specify the fact that most of the data in the world is not structured therefore it is important to keep it in mind and be familiar with some techniques that could help us to structurize our data.

In order to give a visual representation we will display the first five rows of the data set [10]:



```
#Check the Top 5 rows Data
#df.head(5)
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

[1] Definition for “wholesaler”: Person or firm that buys large quantity of goods from various producers or vendors, warehouses them, and resells to retailers.- http://www.businessdictionary.com/definition/wholesaler.html - Accessed: 2019-05-16

As we can see the data is in a table with rows and columns. Rows have an identifier and the columns have a descriptive name so we can conclude that the data set is indeed organized in a structural manner

Quantitative versus qualitative data

After being able to recognize whether the data is structured or unstructured the next step is to describe the characteristics of the data. There are two types:

- Quantitative data which can be described by numbers and basic mathematical calculations can be made
- Qualitative data which is using descriptive language and categories

When we are trying to decide whether or not the data is qualitative or quantitative we can ask the following questions that would help us:

- Can we describe the data by using numbers?
- If we can not do that it would imply that the data is qualitative
- If we can we still need to know if these numbers make any sense when we add them together. For example let us think about zip code. A zip code is always a sequence of numbers. So we could assume that it is quantitative. But what if we add the zip codes together or make an average? Is it going to be a meaningful measurement that would tell us an important aspect of our data? Well the obvious answer is no

Quantitative data can be divided into two further categories:

- Discrete data: it is counted and can take certain values. For example a dice roll has only six values.
- Continuous data: it is measured and has an infinite range of values. For example the height of a person is a continuous number because an infinite scale of decimal is possible

Is our data quantitative or qualitative

Let us look at the data set and recognize whether each attribute is quantitative or qualitative and some of our considerations and realizations:

- InvoiceNo - Qualitative because even though it is expressed in numbers and mathematical calculations can be performed but would not be meaningful
- StockCode - Qualitative because even though it is expressed in numbers and mathematical calculations can be performed but would not be meaningful
- Description - Qualitative because it is not expressed in numbers and we can not perform mathematical calculations on it
- Quantity - Quantitative because it is expressed in numbers. Mathematical calculations can be performed
- UnitPrice - Quantitative because it is expressed in numbers. Mathematical calculations can be performed
- CustomerID - Qualitative because even though it is expressed in numbers and mathematical calculations can be performed but would not be meaningful
- Country - Qualitative because it is described using a name
- InvoiceDate - We left this variable at last for a reason. Most variables can be clear and easy to categorize but there are exceptions and time is one of them. Time can be tricky because

if we look at how much time we have been waiting for the bus then we would count the minutes and seconds but those time units would be rounded because there are also milliseconds, nanoseconds etc. There would be an infinite number of possibilities so actually any variable that is time continuous. But would it make sense to average two dates? It certainly would not. Based on this example it would be more reasonable to classify date as a categorical variable so qualitative (categorical) rather than numerical (quantitative). So the category depends on the usage that needs to be considered for each user case.

Level of measurement

Knowing the level of measurement is important due to the following:

- When we know that the data is nominal then it is clear that the numerical values are just short codes for the names that are longer
- If a measurement is nominal then we know that we would not average the data values

Data can be put into four categories. Different rules apply to them therefore different mathematical calculations can be performed and in order to understand the possibilities we will look at these categories. The four levels are the following:

- The nominal level
- The ordinal level
- The interval level
- The ratio level

The nominal level

The nominal level is the most basic level of measurement. Nominal is also known as categorical or qualitative. Examples of nominal variables are sex, color or preferred type fruit. These are descriptions or labels with no sense of order. Nominal values can be stored as a word or text or given a numerical code. However the numbers do not imply orders. To summarize nominal data we use a frequency or percentage. You can not calculate a mean or average value for nominal data. We can not perform mathematics on the nominal level of data except equality and set membership.

The ordinal level

The next level of measurement is ordinal. Examples of ordinal variables are rank, satisfaction and fanciness. Ordinal variables have a meaningful order but the intervals between the values in the scale may not be equal. For example the gap between first and second runners in a race may be small whereas there is a bigger gap between second and third. Similarly there may be a big difference between satisfied and unsatisfied but a smaller difference between unsatisfied and very unsatisfied. Like nominal ordinal data can be given as frequencies. It is quite a common practice to calculate a mean for ordinal data particularly in research people's behaviour to find mean values for ordinal data. We can also add ordering and comparison to the list. Ordering is the natural order of the data. For example the spectrum of light that is visible for us humans have various names. It could be considered as a natural order. Comparisons would mean to put a “4” on a survey where “4” is worse than “10”.

The interval level

The most precise level of measurement is interval level. This label includes things that can be measured rather than classified, ordered such as number of customers, weight, age and size. Interval ratio data is also known as scale or quantitative. Interval/ratio can be discrete with whole numbers or continuous with fractional numbers. The data is mathematically versatile. We can perform all the operations that are allowed on the lower levels like orderings, comparisons etc. along with two additional operations which are addition and subtraction.

The ratio level

Finally the ratio level makes it possible to define order or difference it is also allows us to multiply and divide. For example if a person has an amount of money in the bank it would make sense to say that €400 is twice as much as €200.

Types of measurement of our data

- InvoiceNo - Nominal
- StockCode - Nominal
- Description - Nominal
- Quantity - Ratio
- UnitPrice - Ratio
- CustomerID - Nominal
- Country - Nominal
- InvoiceDate - We left this variable at last for a reason. Most variables can be clear and easy to categorize but there are exceptions and time is one of them. Well dates themselves are interval but if you do not want to advance any change over time there are only few dates available so it would be nominal. For example if you suggest that the day of the week that makes a difference then you can switch the day into a few dates and it is nominal. You can treat dates as ordinal in a case where you are looking at something and would like to model it in terms of who was present at the given time. In the case where you would treat dates as interval are when the starting point is arbitrary but the units are fixed so 1-4 is not double 6-12.

Additional information

Hereby we will provide some additional information on our data set:

- The transactional data set has all the transactions between 01/12/2010 and 09/12/2011 for a non-store.
- The number of attributes are 8 and the missing values are N/A.
- The attribute characteristics are integer and real.
- The number of instances is 541909.
- The size of the online retail data set is 23.7 MB.

In summary every time we start our work with a new set of data there are certain questions we should ask:

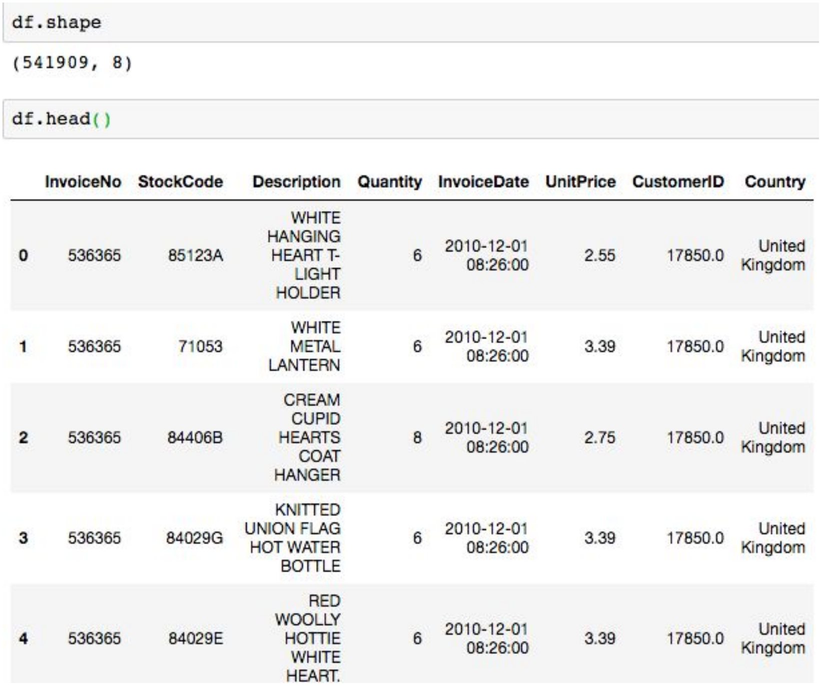
- Is the data organized or not organized? For example is our data structured into a table with rows and columns?
- Is each columns qualitative and quantitative? For example do the values represent qualities?
- What is the level of data each column at? For example nominal, ordinal, interval or ratio level?

But why are so many questions and considerations regarding our data set that we want to work with? Well the answer is that the more we are able to answer these questions the better we understand our data set thus we can interpret the data better and more precisely. Our answers and the conclusions we make will decide how to convert our data from one level to another or which type of graphs we should use.

Explore data

For the exploration we used Microsoft Azure Notebooks which is a useful workspace to develop and run Jupyter notebooks on Azure. It supports multiple languages such as R, Python or F#. After importing the necessary packages and the data it served as a playground for experimentation for us.
(The following screenshots for code and visualization are from our project from Microsoft Azure Notebooks [10].)

Once we have loaded the data it should look the following:



```
df.shape
```

```
(541909, 8)
```

```
df.head()
```

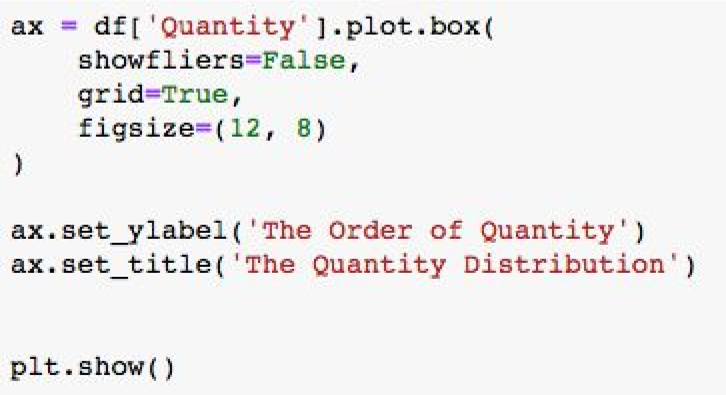
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

As the diagram shows we indeed have 8 columns and as the result of the shape function there is 541909 rows [10]. Each row represents the specific details for an invoice. And as we saw the total number of rows is 541909. Regarding what each columns represents:

- InvoiceNo: a 6-digit number that is uniquely assigned to every transaction.
- StockCode: a 5-digit number that is uniquely assigned to every product
- Description: provides details about the particular product.
- Quantity: provides the quantity of each product for each transaction
- InvoiceDate: provides the date and time of each transaction.
- UnitPrice: provides the price per unit information
- CustomerID: a 5-digit number that is uniquely assigned to every customer.
- Country: provides the name of the country.

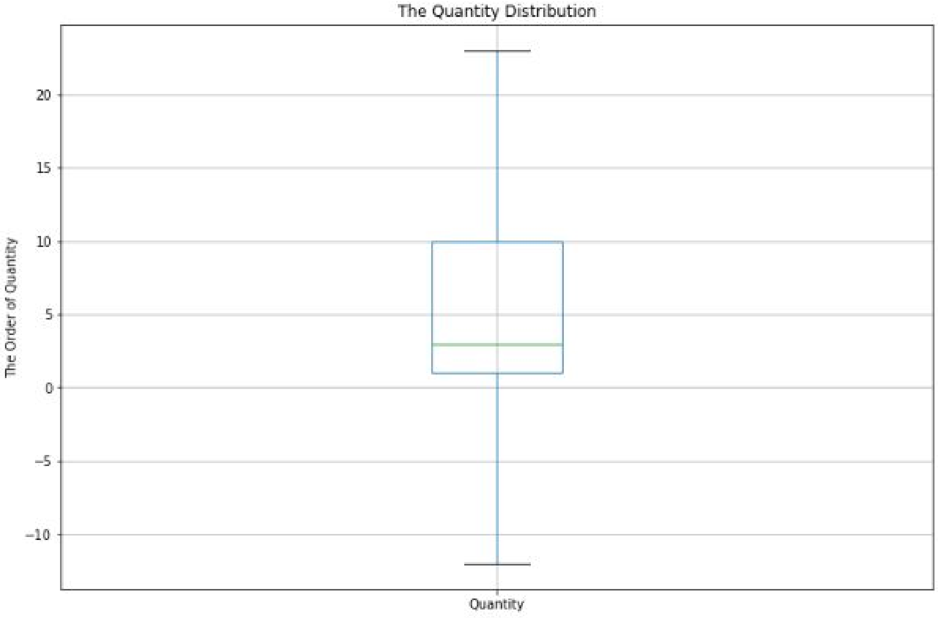
Negative values

In the attribute description we saw that the letter ‘c’ indicates a cancellation. This is considered a negative value and it is important to decide early on how to deal with it. First we will look at the distribution of the Quantity column.

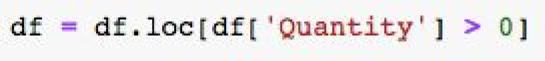


```
ax = df['Quantity'].plot.box(
    showfliers=False,
    grid=True,
    figsize=(12, 8)
)

ax.set_ylabel('The Order of Quantity')
ax.set_title('The Quantity Distribution')


plt.show()
```

We can visualize the distribution in a box plot by using the plot.box() function.



The cancelled orders have negative value in the column. If we decide to leave those values out from further analysis we can use the following code:

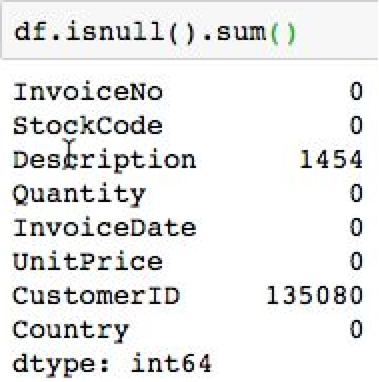


```
df = df.loc[df['Quantity'] > 0]
```

Even though we lose a significant amount of data but we got rid of noise.

NaN values

In order to figure out if we need to deal with missing values we can run the following code:



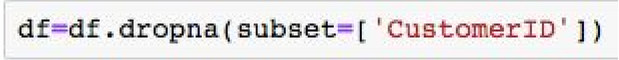
```
df.isnull().sum()
```

```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

So Description and CustomerId are the two columns where there are missing values. There are basically two ways to deal with this:

- Try to fill in the missing data
- To drop the missing values

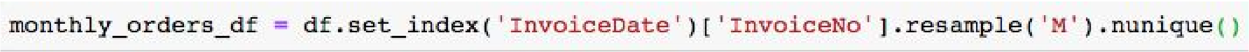
Filling up the missing data with another value might make more sense for the first time because we would not lose valuable data. But the questions raises itself: what can we fill up the missing values with? Getting to know the actual CustomerID is not possible and using some made-up data might be misleading. So in this situation the best decision would be to exclude the missing CustomerID. This action can be performed by the following code:



```
df=df.dropna(subset=['CustomerID'])
```

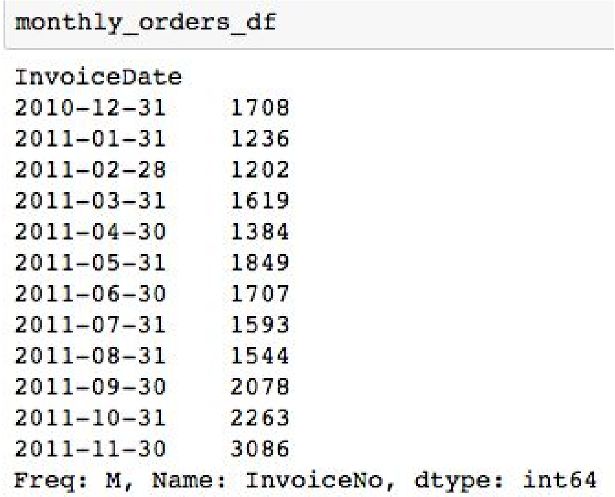
Number of orders

It is important for a business to know the number of orders within a period of time.



```
monthly_orders_df = df.set_index('InvoiceDate')['InvoiceNo'].resample('M').nunique()
```

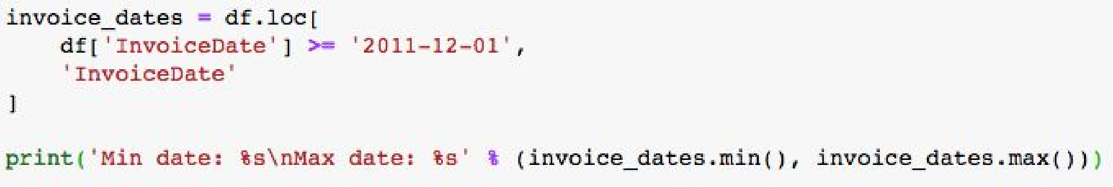
The resample function will convert data into monthly series of data by using ‘M’ and also count the individual invoice numbers. The result looks the following:



```
monthly_orders_df
```

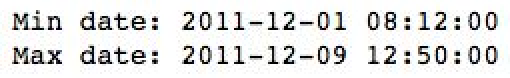
```
InvoiceDate
2010-12-31    1708
2011-01-31    1236
2011-02-28    1202
2011-03-31    1619
2011-04-30    1384
2011-05-31    1849
2011-06-30    1707
2011-07-31    1593
2011-08-31    1544
2011-09-30    2078
2011-10-31    2263
2011-11-30    3086
Freq: M, Name: InvoiceNo, dtype: int64
```

One notable thing is that there is a sudden drop in December 2011. We can look at all invoice dates from 1st of December, 2011 and print the minimum and maximum dates with the following code:



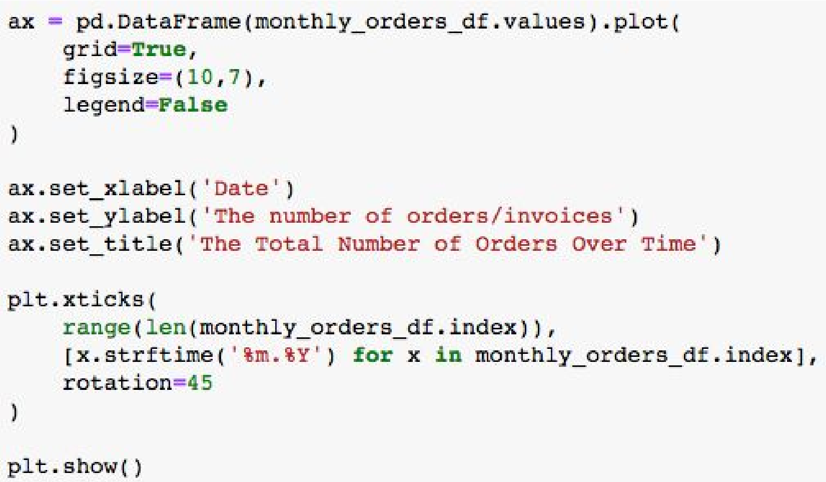
```
invoice_dates = df.loc[
    df['InvoiceDate'] >= '2011-12-01',
    'InvoiceDate'
]

print('Min date: %s\nMax date: %s' % (invoice_dates.min(), invoice_dates.max()))
```

Result is the following:



```
Min date: 2011-12-01 08:12:00
Max date: 2011-12-09 12:50:00
```

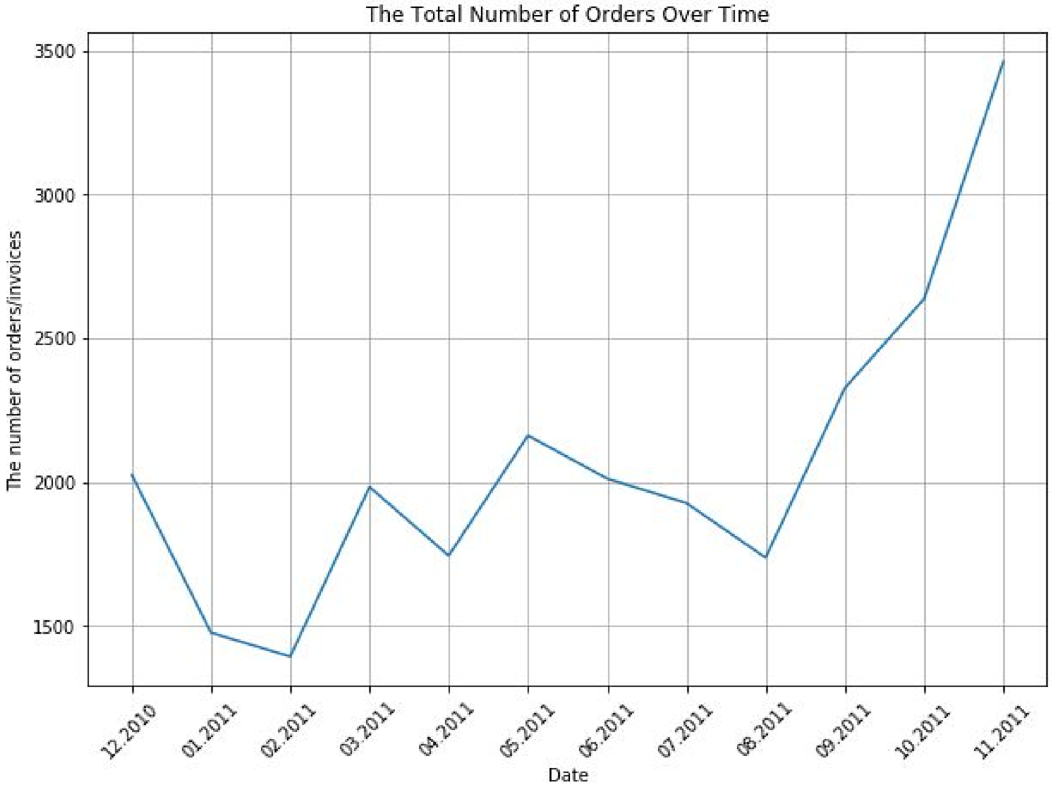
We only have data from 1st of December in 2011 and the use of such a data would be a misrepresentation in the later analysis. If we decided to get rid of the data we could use the following code:



```
ax = pd.DataFrame(monthly_orders_df.values).plot(
    grid=True,
    figsize=(10,7),
    legend=False
)

ax.set_xlabel('Date')
ax.set_ylabel('The number of orders/invoices')
ax.set_title('The Total Number of Orders Over Time')

plt.xticks(
    range(len(monthly_orders_df.index)),
    [x.strftime('%m.%Y') for x in monthly_orders_df.index],
    rotation=45
)

plt.show()
```

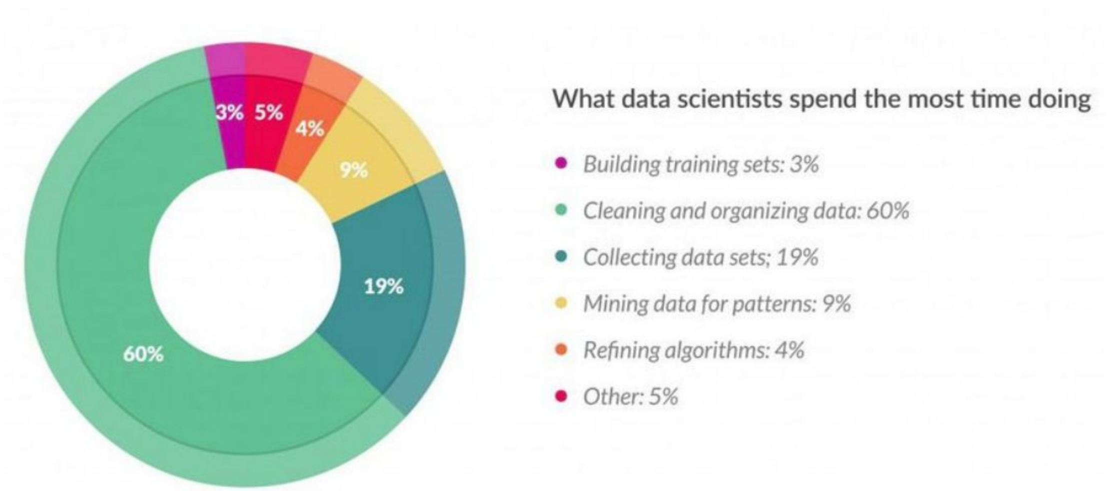
We use the plot function. When we use the xticks function we can customize the labels of the x-ticks. This x axis is formatted by year and month. When we wrote x.strftime('%m.%Y') where x is the object and %m stands for month and %Y stands for year. The strftime function will format date object into the given format.

The result without the month that is missing values for the whole month is the following:



Data Preparation

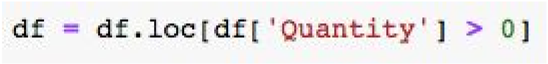
In this section we will dive into the preparation of our data. Getting to know our data set is important because "...*Data scientists spend 60% of their time on cleaning and organizing data. Collecting data sets comes second at 19% of their time, meaning data scientists spend around 80% of their time on preparing and managing data for analysis...*"[9]



What data scientist spend most of their time [9]

Clean data

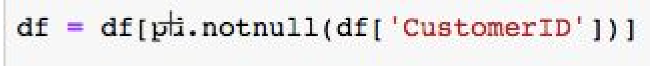
- Negative values: canceled transactions are represented as negative value. In we want to get rid of them we do with the following code:



```
df = df.loc[df['Quantity'] > 0]
```

This way we will take only the positive values and will store them in the variable.

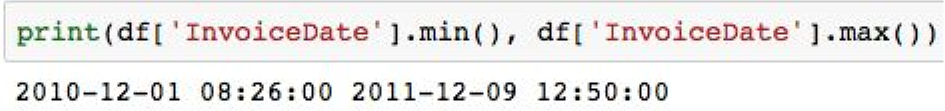
- NaN values: we will drop the records that are missing CustomerID with the following code:



```
df = df[pd.notnull(df['CustomerID'])]
```

The not null function will return a list of array. In this array the true values will tell us that the value in the given index is not null. We will keep these values in the column and in the variable.

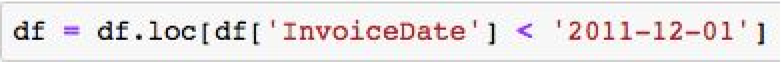
- Incomplete data: the transaction data is not complete. The following code will prove it:



```
print(df['InvoiceDate'].min(), df['InvoiceDate'].max())
```

```
2010-12-01 08:26:00 2011-12-09 12:50:00
```

The data for the last month is not complete. In we want to get rid of them we do with the following code:



```
df = df.loc[df['InvoiceDate'] < '2011-12-01']
```

We will keep transactions before December 01, 2011 and store in variable.

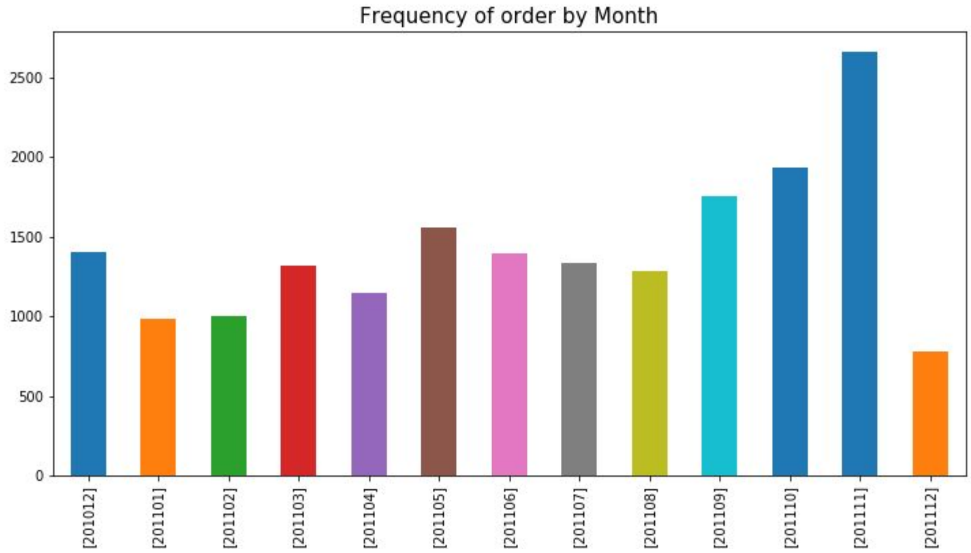
Data Visualization

Data visualization is the process of taking raw data,transforming it into graphs,charts,images and even videos that explain the numbers and allow us to gain insights from it.It changes the way we make sense of the information to create value out of it. Data visualization is quick,easy way to convey concepts in a universal manner-and we can experiment with different scenarios by making slights adjustments[16].

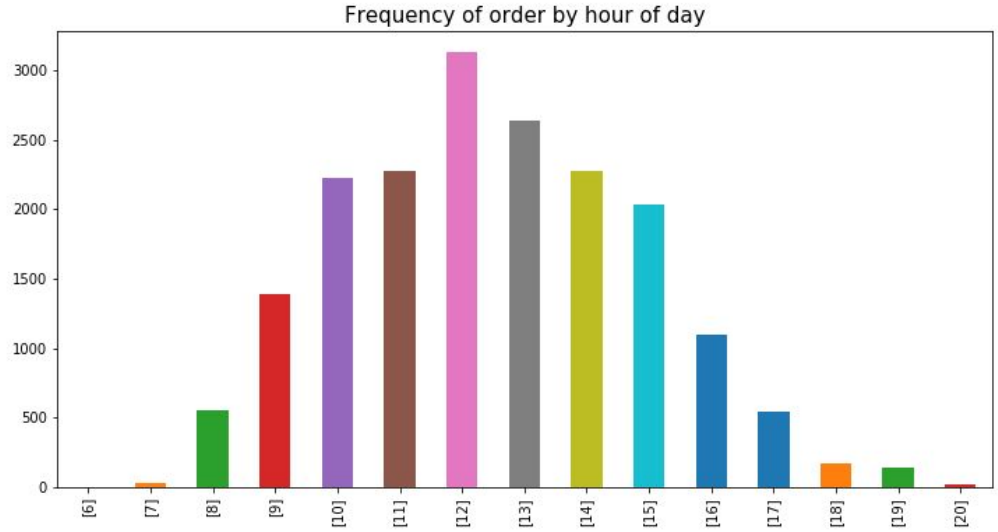
Data visualization can:

- Identify areas that need attention and improvement
- Predict sales volumes
- Clarify which factors influence customers behaviour

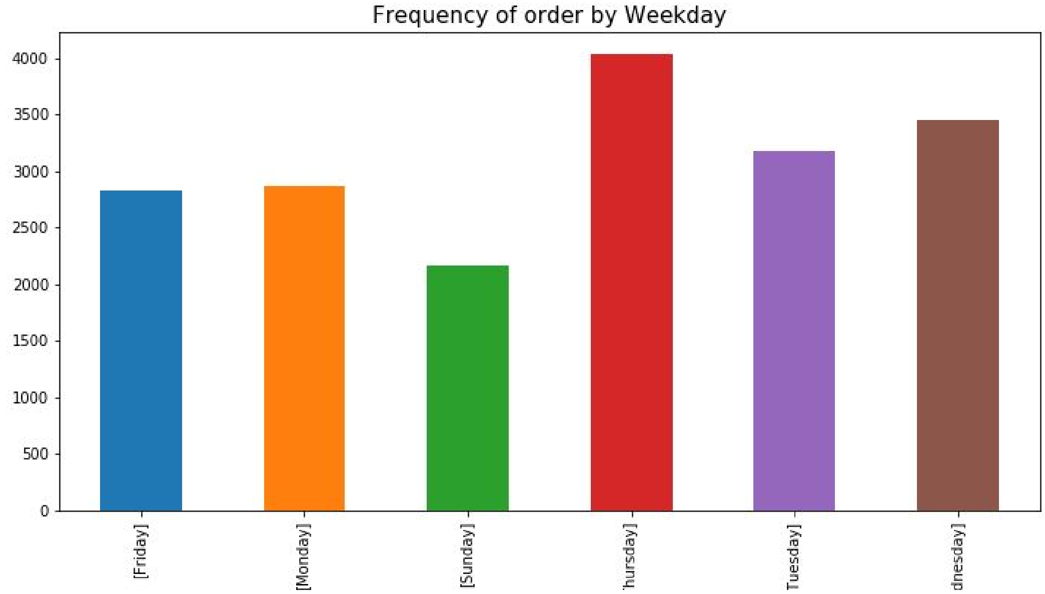
We have used some library of python for visualize our data. In following diagrams represent the specific part of our data visualization.



This graph represents how many orders were made per month. In November 2011 there was made the highest order compared to others month.



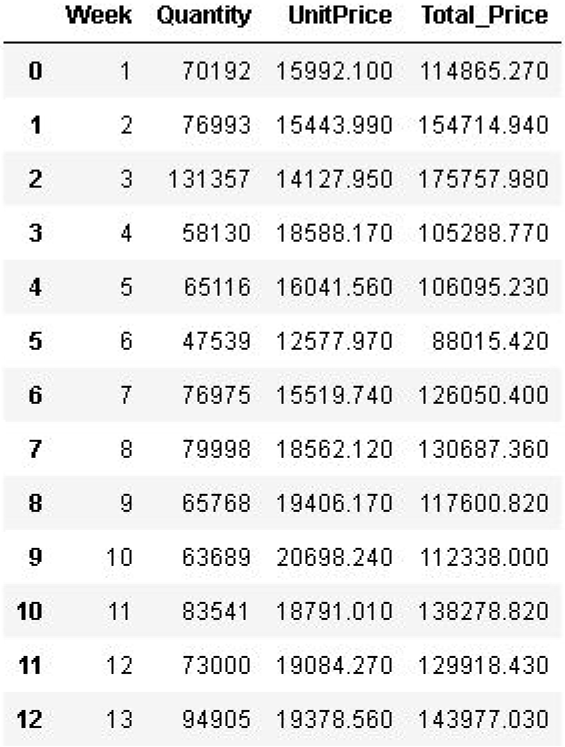
After investigating the frequency of the order by the hour of the day we can conclude that the busiest time for making an order is at 12. The busiest time period is between 9am and 16pm.



Similarly this graph represents the frequency of order by weekday. Here customers made more orders on Thursday compared to others day.

Select Data

Here we cleaned and selected the data that would be used for the Multiple Linear regression model in the next phase. The Total_Price is the dependent variable. Quantity and Unit Price represent the independent variables.



	Week	Quantity	UnitPrice	Total_Price
0	1	70192	15992.100	114865.270
1	2	76993	15443.990	154714.940
2	3	131357	14127.950	175757.980
3	4	58130	18588.170	105288.770
4	5	65116	16041.560	106095.230
5	6	47539	12577.970	88015.420
6	7	76975	15519.740	126050.400
7	8	79998	18562.120	130687.360
8	9	65768	19406.170	117600.820
9	10	63689	20698.240	112338.000
10	11	83541	18791.010	138278.820
11	12	73000	19084.270	129918.430
12	13	94905	19378.560	143977.030

As the diagram shows each row represents the specific Week, the total Quantity of products,the Total_Price which is denoted as total sale.

Modelling

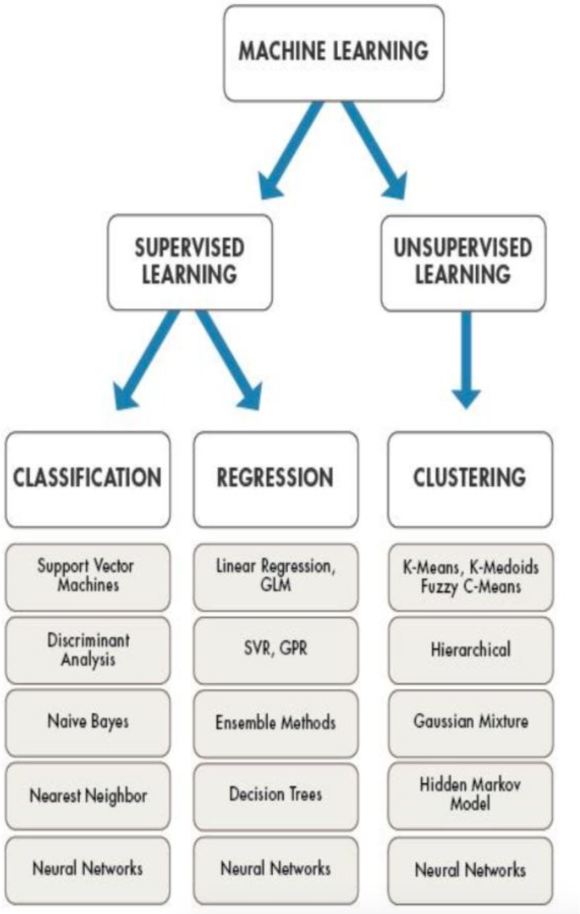
Select modelling technique

Model selection in the field of machine learning can have different meanings. In the machine learning field,the terms hypothesis and model are often used interchangeably. Hypothesis could be the “educated guess” and the model would be the action of this guess to test this hypothesis[2].

The important considerations for choosing a machine learning algorithm were the following:

- Type of problem: algorithms were designed to solve specific problems. It is important to keep in mind what the goal is since it defines what kind of algorithm we should choose.
- Accuracy: the required accuracy can be different depending on the application. An approximation might be adequate but it may lead to a big reduction in processing time.
- Training time: different algorithms have different running time. Training time would depend on the size of the dataset and the target accuracy

- Linearity: algorithms such as linear regression and logistic regression make use of linearity. Depending on the given problem they can bring accuracy down. On the other hand linear algorithms tend to be relatively simple and fast to train
- Number of parameters: parameters affect how the algorithm behaves. If there are a big number of parameters it is harder to find a good combination. On the other hand more parameters give greater flexibility, training time and accuracy
- Number of features: the number of features in datasets can be very big compared to the number of data points. The big number of features can slow down learning algorithms and make the training time too long
- There are generally three types of machine learning algorithms [15]:
    - Supervised learning algorithms: such algorithms are used to predict when the outcome is known. We can think of the situation in which we would like to predict who would buy in the upcoming few days. A supervised learning algorithm can be fed with historical data purchase and various data points such as user`s age, address, purchase date to make features that would predict the target
    - Unsupervised learning algorithm: these algorithms are used when we do not know the specific prediction. We can think of a situation in which we want to cluster customers into subgroups based on the customer behaviour.
    - Reinforcement learning algorithms: these algorithms are used when we want the model learn by itself and train itself based on prior knowledge. We can think of a situation in which a company would like to test its retail strategies and choose upon the strategy that works the best for the company.



Machine learning algorithms [15]

We have selected two Machine learning models for background research e.g. Linear Regression, Multiple Regression Model,ARIMA model. These models are good for prediction. In our project we will going to implement one of these models.
According problem statement we would like to make predictions. Therefore for example Linear Regression is a good choice because “...Linear Regression is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). …we establish relationship between independent and dependent variables by fitting a best line…”[15]

Linear Regression

Linear regression is a machine learning algorithm based on supervised learning.Linear regression uses one or more independent variable (X) to explain or predict the outcome of dependent variable (Y).The value of dependent variable, is a function of the independent variable, value of y is always dependent on the value of x. One variable,denoted x, is regarded as the predictor or independent variable. Another variable,denoted y,is regarded as the response ,outcome or dependent variable.

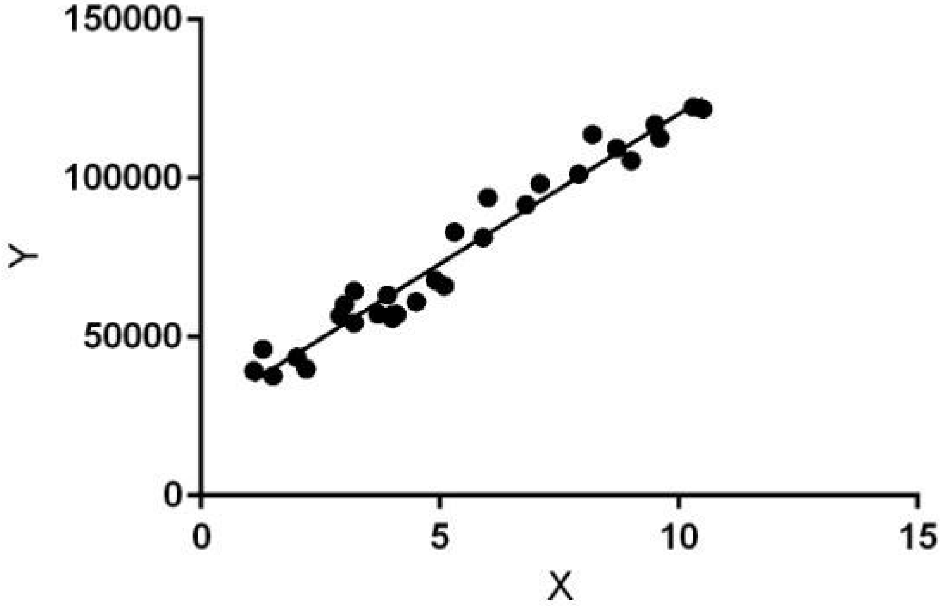


Fig.1[13]

In the figure above X(input) is the work experience and Y(output) is the salary of a person.The regression line is best fit line for our model.

Algebra reviews of simple linear regression equation represented by :

y=mx+c

Where y is the dependent variable, *m* is known as the slope/ coefficient of the line,x is independent variable and *C* is known as the intercept[1].

In statistics world, we use different notation:

$$y = \beta_0 + \beta_1 x$$

β0 is the intercept and β1 is the slope or coefficient of the X.

Linear regression try to draw a line to minimize the error term ε. Linear regression must include this error (ε) term. This is how Linear Regression Model looks like.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

ε is the error term that describes the effect on y of all factors other than x. β0 is the y-intercept and β1 is the slope or coefficient of the X.

Major use cases for Linear regression are:

- Evaluating Trends and sales Estimates
- Analyzing the impact of price Changes
- Predict monthly product sales and improve yearly revenue projection.

- Trend forecasting e.g what will be the price of bitcoin next six month?

Types of Linear Regression Relationships

In linear regression, data is modelled using a straight line. When independent variable increasing on x-axis so dependent variable also increasing on y-axis. So in this scenario, what kind of linear regression line we will get? we will get positive regression line.

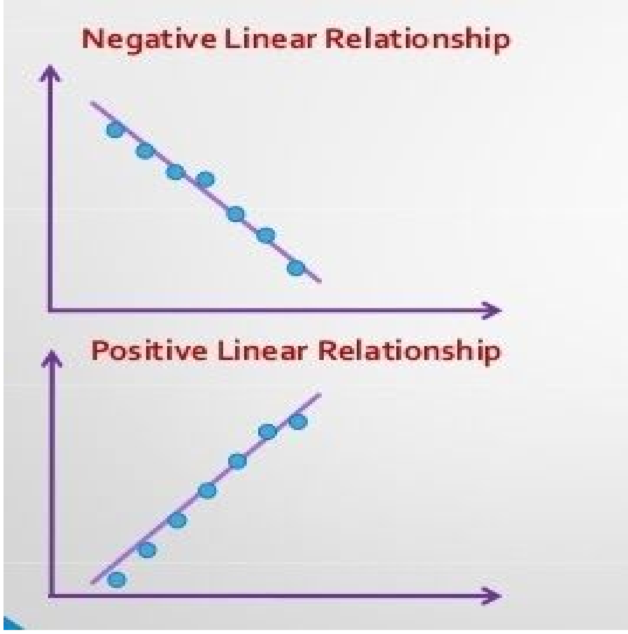


Fig.2[1]. Positive and Negative correlation

In other hand independent variable is increasing on x-axis and dependent variable is decreasing on y-axis. So we will get negative regression line[1].

Example of How to find the Linear Regression Equation:

We will use following custom made table data to find the linear regression equation. The x column shows bill amount and y column shows the tip amount.

Meal	Bill($dollar$)	Tip(dollar)	Bill deviation	Tip deviation	Deviation products	Bill deviation squared	Tip deviation squared
	x	y	$xi - \bar{x}$	$yi - \bar{y}$	$(xi - \bar{x})(yi - \bar{y})$	$(xi - \bar{x})2$	$(yi - \overline{y)2}$
1	34	5	-40	-5	200	1600	25
2	108	17	34	7	238	1156	49

3	64	11	-10	1	-10	100	1
4	88	8	14	-2	-28	196	4
5	99	14	25	4	100	625	16
6	51	5	-23	-5	115	529	25
	$\bar{x}$=74	$\bar{y}$=10			$\Sigma = 615$	$\Sigma = 4206$	$\Sigma = 120$

The linear equation form : $\hat{y} = b_0 + b_1x_0$. To conduct a regression analysis, we need to find $b0$ and $b1$. Calculations are shown below. Notice that all of our inputs for the regression analysis come from above tables.

First, we find regression coefficient($b1$):

$$b_1 = \Sigma [(x_i - \bar{x})(y_i - \bar{y})] / \Sigma [(x_i - \bar{x})^2]$$

$$b1=612/4206$$

$$b1=0.1462$$

We got the slope of our regression coefficient(b1), now we can solve for the y-intercept(bo) :

$$b_0 = \bar{y} - b_1 * \bar{x}$$

$$b0=10-(0.1462)(74)$$

$$b0= -0.8188$$

Therefor, the regression equation is:

$$\hat{y} = -0.8188+0.1462 x$$

$$\hat{y} = 0.1462x - 0.8188$$

a value for the independent variable (*x*), perform the computation, and we have an estimated value (ŷ) for the dependent variable. In our example, the independent variable is the bill amount and the tip amount is the dependent variable. From the regression equation we can quick interpret that for every dollar the bill amount increase, we expect or predict the tip amount to increase by 0.1462 dollar .

Find the Coefficient of Determination

We can use following formula to find the coefficient of determination

$$R^2 = \{ (1 / N) * \Sigma [(x_i - \bar{x}) * (y_i - \bar{y})] / (\sigma_x * \sigma_y) \}^2$$

where N is the number of observations used to fit the model, Σ is the summation symbol, xi is the x value for observation i, $\bar{x}$ is the mean of x value, yi is the y value for observation i, $\bar{y}$ is the mean of y value, σx is the standard deviation of x, and σy is the standard deviation of y[14].

First we need to find the standard deviation of x (σx):

$$\sigma_x = \text{sqrt} \, [\, \Sigma \, (x_i - \bar{x})^2 \, / \, N \,]$$

where sqrt is the square root and N is the number of samples.

$$\sigma x = \sqrt{4206/6} = 26.4764$$

Next, we find standard deviation of y (σy):

$$\sigma y = \sqrt{120/6} \ = 4.4721$$

Now we compute the coefficient of determination:

$$R^2 = \{ \, (1 / N) * \Sigma \, [\, (x_i - \bar{x}) * (y_i - \bar{y}) \,] \, / \, (\sigma_x * \sigma_y) \, \}^2$$

$$R2 \ = 0.7493 \ \text{ OR } 74.93\%$$

We conclude that 74.93% of the total sum of squares can be explained by using the regression equation to predict the Tip amount. The reminder 25.07% is error which explained the variation of the independent variable. Now we can develop a model that says this fit well or does not fit well. In this case,we can say that it is good fit because of coefficient of determination is higher.

Multiple Linear Regression

Multiple regression is an extension of simple linear regression. It is a statistical technique that uses several independent variable to predict the outcome of a response variable.The idea of multiple linear regression is to model the linear relationship between the independent variables and response(dependent) variable[1,13].

We have seen the concept of simple linear regression where a single predictor variable X was used to model the response variable Y . In many applications, there is more than one factor that influences the response. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables.

Examples:

- The selling price of a house can depend on the location, the number of bedrooms, the number of bathrooms, the year the house was built, the square footage of the lot and a number of other factors.
- The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.

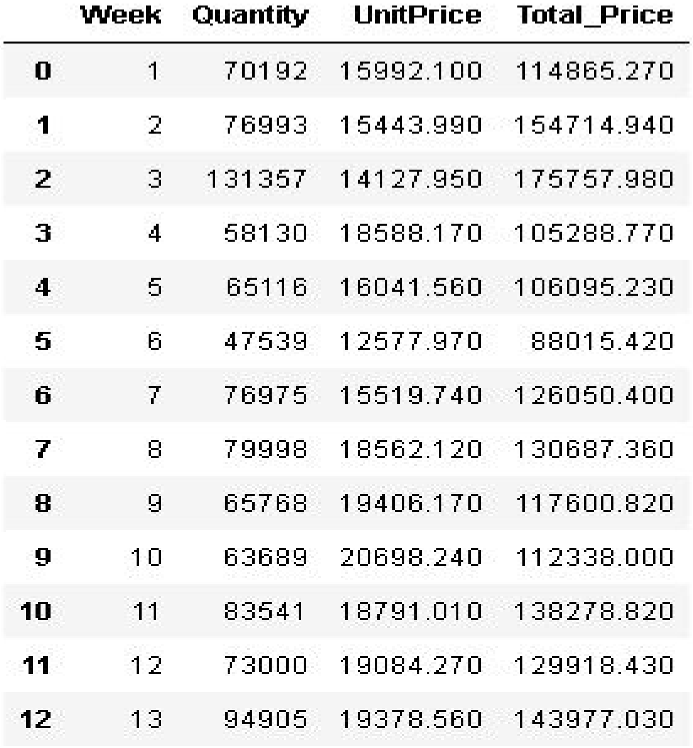
A multiple linear regression model with k predictor variables x1, x2, ..., xk and a response y, can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \epsilon.$$

ε are the residual terms of the model and the distribution assumption we place on the residuals will allow us later to do inference on the remaining model parameters. β0 is the y-intercept and β1, β2, ..., βk is the coefficient for x1,x2….xk predictor variable[1].

Implementation and Result

In this section, we have implemented our chosen model which can predict the total sales of any week of the year. We used the data that was cleaned and selected previously to feed the Multiple Linear Regression model.



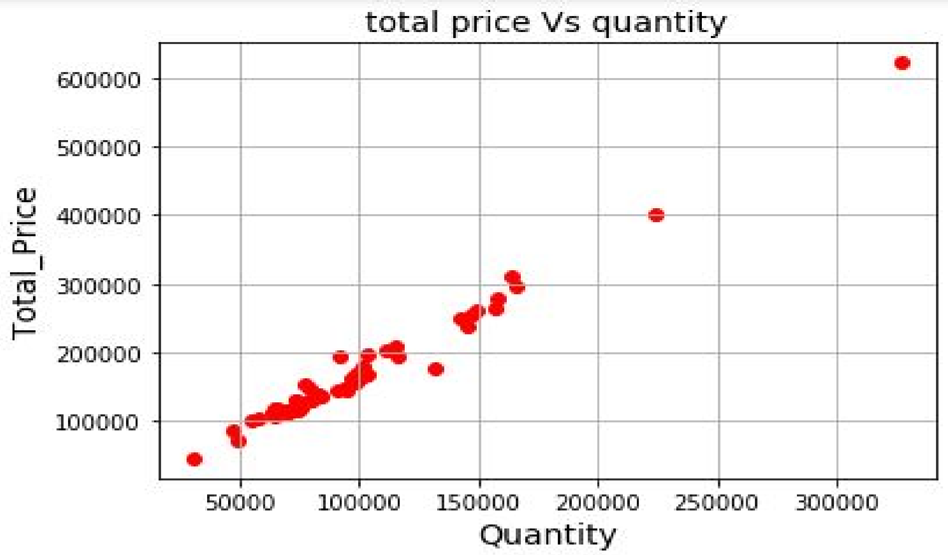
	Week	Quantity	UnitPrice	Total_Price
0	1	70192	15992.100	114865.270
1	2	76993	15443.990	154714.940
2	3	131357	14127.950	175757.980
3	4	58130	18588.170	105288.770
4	5	65116	16041.560	106095.230
5	6	47539	12577.970	88015.420
6	7	76975	15519.740	126050.400
7	8	79998	18562.120	130687.360
8	9	65768	19406.170	117600.820
9	10	63689	20698.240	112338.000
10	11	83541	18791.010	138278.820
11	12	73000	19084.270	129918.430
12	13	94905	19378.560	143977.030

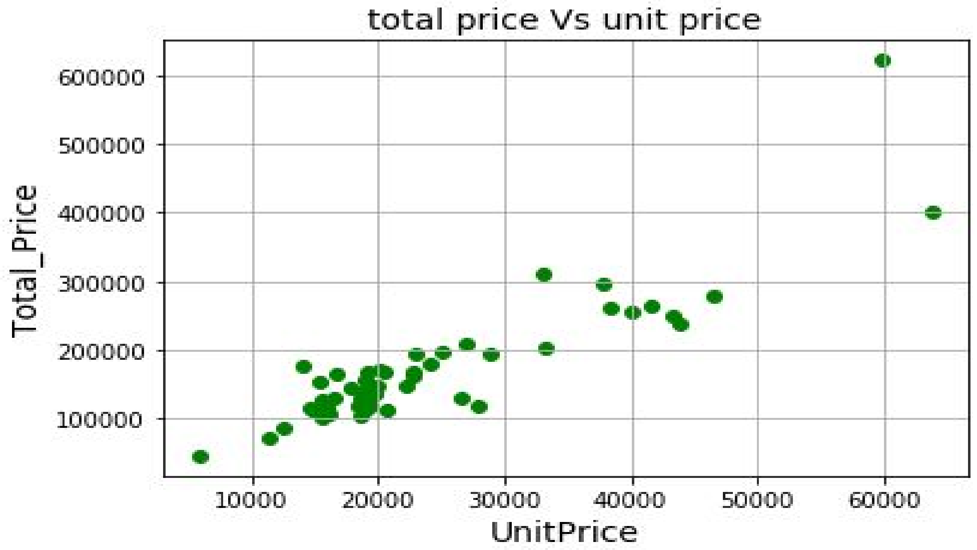
As the diagram shows each row represents the specific Week, the total Quantity of products,total UnitPrice and the Total_Price which is denoted as total sale.

Before we execute a multiple linear regression model, we checked that a linear relationship exists in our data set between:

- The Total_Price(dependent variable) and the Quantity(independent variable)
- The Total_Price(dependent variable) and the UnitPrice(independent variable)

To perform a linearity check, we used scatter diagram with python matplotlib library which provide us following two diagrams:

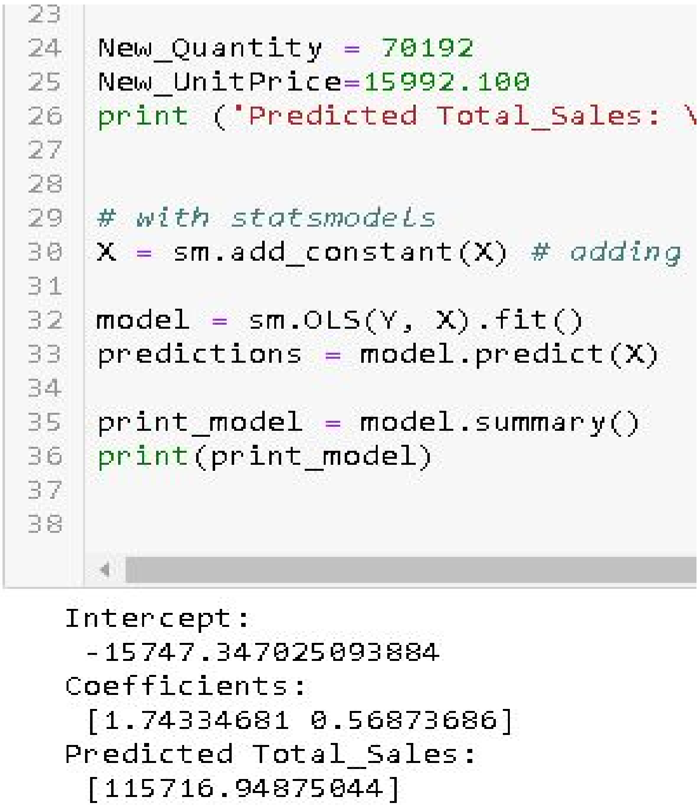




As we can see, a linear relationship exists in both diagrams:

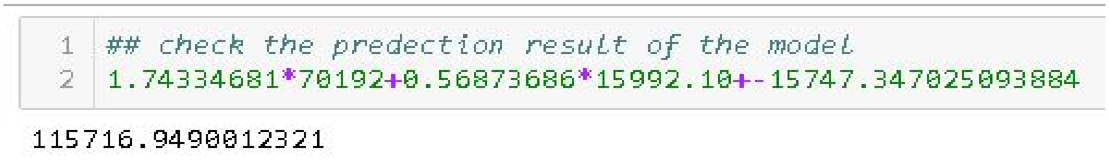
- In the first diagram, when quantity increase, the total_price also increase
- In the second diagram, when unit price increase, the total_price increase

This following code output includes the intercept and coefficients.We used this information to build the multiple linear regression.



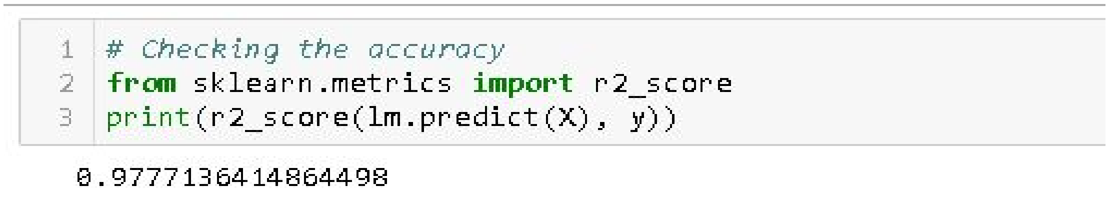
```

New_Quantity = 70192
New_UnitPrice=15992.100
print ('Predicted Total_Sales: \


# with statsmodels
X = sm.add_constant(X) # adding

model = sm.OLS(Y, X).fit()
predictions = model.predict(X)

print_model = model.summary()
print(print_model)


```

```
Intercept:
 -15747.347025093884
Coefficients:
 [1.74334681 0.56873686]
Predicted Total_Sales:
 [115716.94875044]
```



```
## check the predection result of the model
1.74334681*70192+0.56873686*15992.10+-15747.347025093884
```

```
115716.9490012321
```



```
# Checking the accuracy
from sklearn.metrics import r2_score
print(r2_score(lm.predict(X), y))
```

```
0.9777136414864498
```

The result of the model would show us the total sales for the week number 1. In the last screenshot we are checking if the previous result was correct. As we can see there is a minor difference between the two results and it can be because our model accuracy is 0.9777.

ARIMA Model

ARIMA Models is a time series model which is mostly used to predict the uncertain nature of business trends in an effort to help business owners make better decision and plans. A historical data are gathered over the periods and data is accumulated at regular intervals. We can analyze the historical patterns in the data by using statistics and mathematics, and this technique can help for forecasting or predicting the future of the business.

The Arima time series analysis is one of the most critical and complex technique, to produce accurate forecasts based on a description of historical patterns in the data. It uses shifts and lags from the previous data to uncover patterns(e.g. seasonality, moving average) of the data.

The Arima models are a class of linear models which is capable to represent stationary as well as non-stationary time series. The stationary data is the same and the value is constant and that non-stationary process have no natural constant mean level.

The Arima models do not involve any independent variables in their construction rather Arima uses of the information in the series itself to generate forecasts. These models rely heavily on autocorrelation patterns in the data.

Box-Jenkins Methodology

There are different kinds of Time series model but the Box-Jenkins methodology is different from other methods because this methodology does not assume any particular pattern in the previous data of the series to be forecast. It uses the iterative approach to find the best model from its historical data and this process repeated until acquiring a satisfactory model.

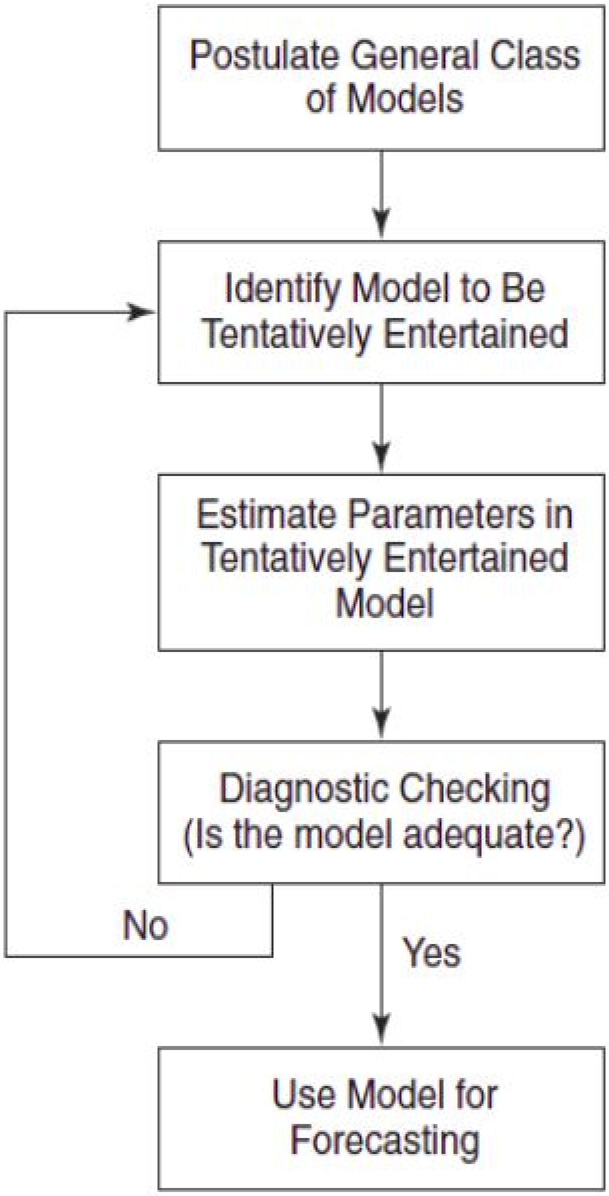

Postulate General Class of Models
Identify Model to Be Tentatively Entertained
Estimate Parameters in Tentatively Entertained Model
Diagnostic Checking (Is the model adequate?)
No
Yes
Use Model for Forecasting

Fig :Box-Jenkins model-building diagram(18)

Method and Feature

A time series data into a more workable format used in Arima model. It has three components which are autoregressive component, moving average and white noise. If the time series data is not stationary then it’s needed to be converted into stationary data where observation will not be constant.

ARIMA Model

The Arima model have three components, an autoregressive(AR) a differencing (I) and a moving average(MA) component. Non-seasonal Arima model is displayed as ARIMA p,d,q. The p,d and q values represent the amount of periods to lag for in ARIMA calculation.

p = Autoregressive order which is allowed to incorporate the effect of past values.

d = order of integration. It includes terms in the model to incorporate the differencing of the amount. This process is called differencing and the d refers to the number of transformations used in the process. (The number of past time points to subtract from the current values)

q = Moving average order and it refers to the lag of the error component. The error component refers to the part of the time series not explained by trend or seasonality. MA models look like linear regression models where the predictive variables are the previous q periods of errors and the error values observed from its previous time points in the past.

P= It is denoted Seasonal Autoregressive Order
D= Seasonal Differencing Order
Q= Seasonal Moving Average Order

The ARIMA Model has four steps which are used to determine the optimal ARIMA Model. The following four steps are :
1.Identification phase
2.Estimation phase
3.Diagnostic phase
4.Forecasting/Predicting phase

Step 1: Model Identification
In the Identification phase the time-series data appears in a line graph. The unit root tests are used to recognize if the data is stationary or not. A correlogram will also be used to decide which of the components have an Autocorrelation Work (ACF) or Partial Autocorrelation Work (PACF). Suppose if two lags are the same in AC and PACF of corregram then it’s tentatively ARIMA Model.

Autocorrelation :
Autocorrelation indicates how correlated a time series data with its past values. It is calculated from the data are subject to sampling variation.

Partial autocorrelation :

The partial autocorrelation is the essence of the relation between an observation in a time series with an observation at prior time steps. After calculating the autocorrelation the intervening observations are removed. “The partial autocorrelation at lag k is the correlation that results after removing the effect of any correlations due to terms at shorter lags.”[17]
In the identification phases if the series is not stationary then it is converted into a stationary series by differencing. However, the original series replaced by a series of differences:

$$\Delta^2 Y_t = \Delta(\Delta Y_t) = \Delta(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t-2}$$

Differencing is done until a plot of data is converted into a series and the autocorrelations are omitted rapidly. “Consequently, from this point on, the ARIMA(p,d,q) notation is used to indicate models for both stationary(d=0) and nonstationary(d>0) time series.”[18]

Step 2: Model Estimation
After selecting the model the parameter for that model needs to be estimated. The parameter of Arima model would be the minimized sum of squares of the fitting errors. To reduce the squared error here we applied the nonlinear least squares procedures.
The residual mean square error is also the estimate of the variance error which is defined as

$$s^2 = \frac{\sum_{t=1}^{n} e_t^2}{n - r} = \frac{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}{n - r} \qquad ($$

where

$$\begin{aligned} e_t = Y_t - \hat{Y}_t &= \text{the residual at time } t \\ n &= \text{the number of residuals} \\ r &= \text{the total number of parameters estimated} \end{aligned}$$

Formula for calculating RMSE [18]

The residual mean square error assessing model fit and compare the other model. It is also to calculate the forecast error limit.

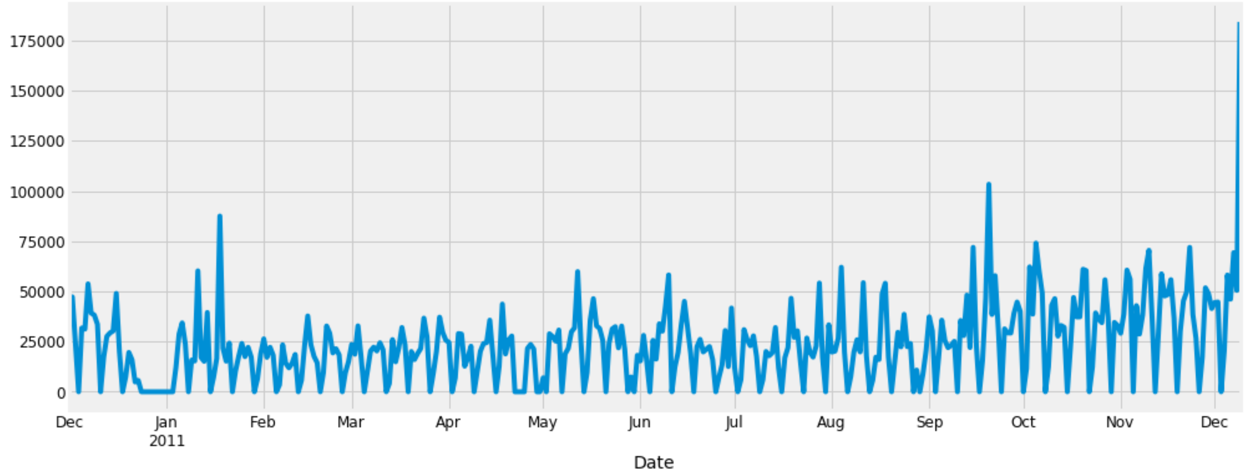
Step 3: Model Diagnostics
The diagnostic phase will determine if the ARIMA model fits the purpose. These characteristics incorporate being tightfisted, stationary, invertible, converging, has high quality parameter estimates and has white noise residual series. If these characteristics are not met the cycle is remade from phase 1 to phase 4.

Step 4:
In the last phase which is the predictive phase provides the estimates under the three scenarios which are most-likely, worst case and best case.

Implementation :
In our E-Commerce dataset allows us to apply time series to forecast the future sales. We used the ARIMA model in our dataset to better understand and predict the next year sells in the time series.

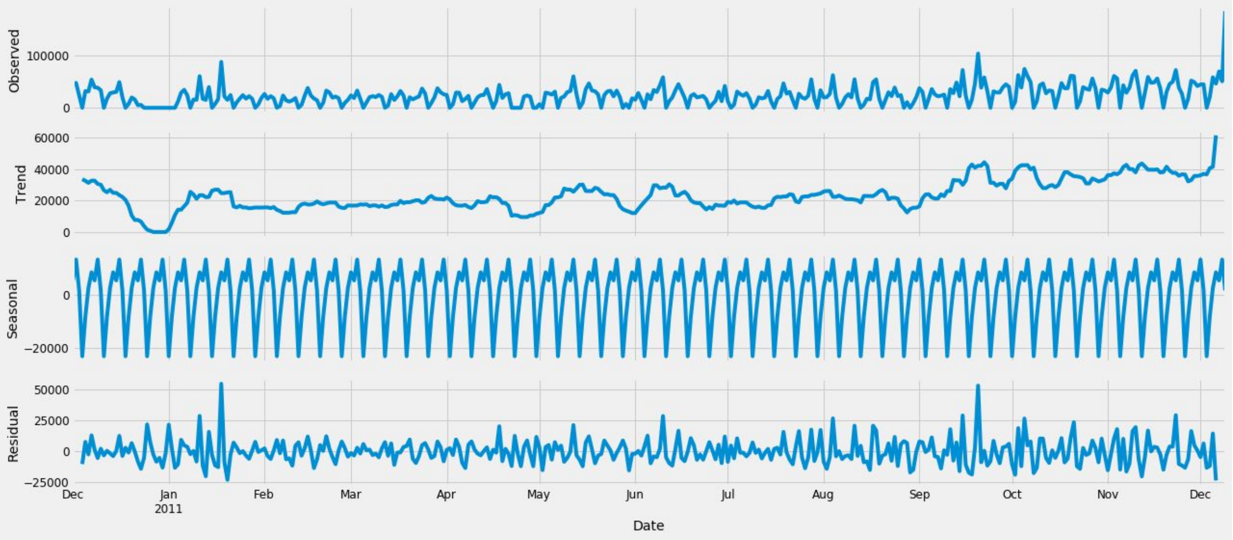
Feature Selection:
After cleaning our dataset we analyzed and visualized the inside relation of data in different points. But for the Arima model we only need time index and total sales. So we will apply only two features in our time series data.

Visualizing the time series data:
Now we will visualize our dataset to understand the stationary or non-stationary as well to follow the trend of our time series data.



The diagram shows the total sales for a period of time on a daily basis.

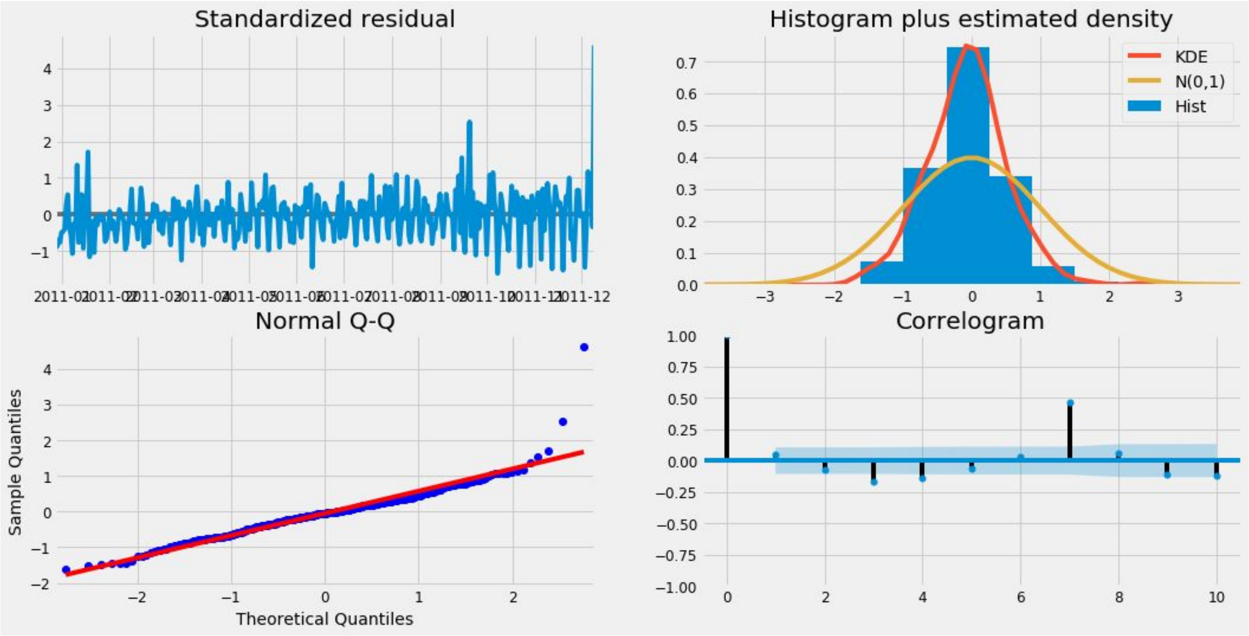
ARIMA Time Series Model



We used the additive model to segment our data. The model that assumes that the difference between the values within a period of time would be approximately the same each year. Looking at it from another perspective we could say that the amplitude of the seasonal effect is the same each year.

Parameter for ARIMA Model
The goal is to find the values of ARIMA that optimize a metric of interest. There are existing best practices to achieve this goal but the correct parametrization requires domain expertise and time. It is also important to consider which programming language we are using since R would provide an automatic solution byt in Python we should do a workaround. In short we should use a so-called grid search to figure out the combinations of the parameters. Once the possible parameters are explored the optimat set will be the one that would have the best performance.

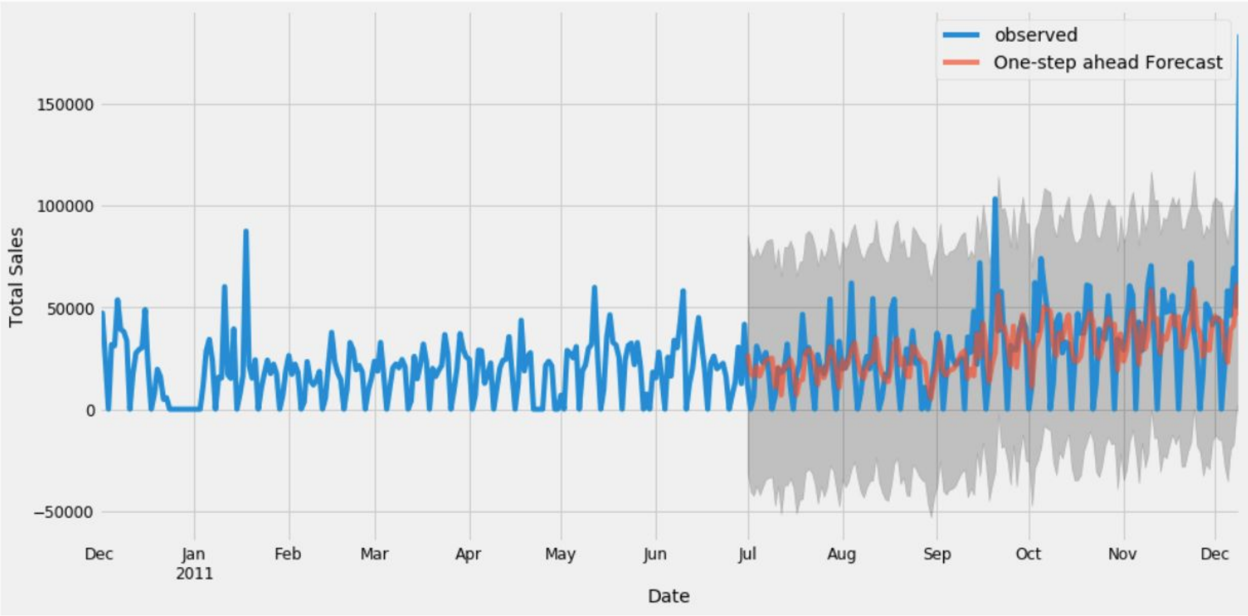
Fitting an ARIMA
The goal is to make sure that the residuals are uncorrelated and normally distributed with zero-mean. If the seasonal model does not live up to these criterias it shows that it can be further developed.



In the right top figure we can see that the red KDE line and the yellow N line look very different from each other. It would tell us that the residuals are not normally distributed.
In the Normal Q-Q the blue dots which would be the ordered distribution of residuals follow a linear trend. This one indicates that the residuals are normally distributed.
In the Standardized residual diagram we cannot make such a clear conclusion and we could say that there is white noise.
The Correlogram (autocorrelation) diagram would tell us that time series residuals have low correlation with lagged version of themselves.

The model produces a fit that might be satisfactory fit and this fit would enable us to get to know our time series data better. Additional steps could be to validate the forecasts and modify them accordingly.

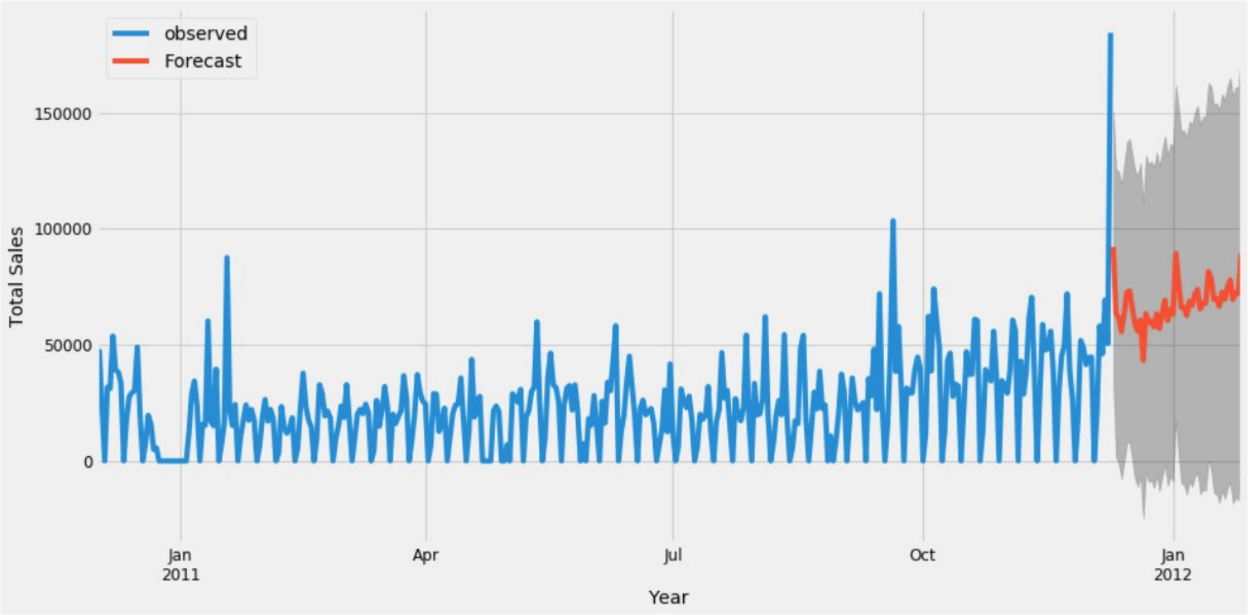
Validating forecast



We made a model for time series that can we use for prediction. If we want to validate it we could do it by comparing the predicted values to the real values of the time series. As we can see our forecast shows an overall increasing trend and that our forecast aligns to this increasing trend.

Producing and visualizing forecasts

Let us say we decide to leverage our seasonal time series model. The goal is to forecast future values. It can be done by the get_forecast() attribute. The get_forecast() attribute of the time series object can calculate values for a number of steps.



Our forecast shows that the time series is expected to continue to increase. This increase takes place at a steady pace. The more we look into the future the more insecure we can become. The confidence interval which is the dark grey shows this insecurity in our values. It gets larger as we move into the future.

Conclusion

The time has come to look back on our report and to draw some conclusions and to highlight the milestones with our findings. First we would look at various parts of our report then later on from a general perspective.

In the business understanding part we focused on laying down some basic terminologies that can cause confusion since as we experienced terms are used interchangeably. Also another important part was the project plan which formed the basic structure for the report. We learned that it is very crucial to have clear and realistic goals and a clear structure otherwise a lot of time needs to be spent unnecessarily on revisiting this stage.

In the upcoming parts such as data understanding and data preparation the primary goal was to provide a combination of theoretical and practical knowledge to better understand our data set. In this part we realized that these processes really boil down to understand the real use of statistics and the interpretation of the results. Conclusion is that the field heavily depends on statistical analysis and also critical thinking. Therefore the group spent a big amount of time to acquire the appropriate knowledge that could be later on applied.

The importance of understanding our data became clear when we reached the next phase which was data preparation. Here we learnt that in order to properly do a good job in this phase we need to have a good knowledge of our data and a good knowledge of the model we want to use since we prepare the data for the particular model we want to use. The data that we cleaned, selected was the parameter for the model we chose. If we missed a few aspects it would be reflected in our results as well. We looked at two models and implemented them in order to get results and test our understanding of them. The results gave us useful insights for interpretation. But our conclusion is that more time and work is needed on these models to get a deeper understanding.

From a more general perspective we can conclude that the group managed to answer the problem formulation. It is important to mention that even though the fulfillment took place as looking back we can say that there is place for further investigation and development regarding the process of a data science project. We have reached far and the process made us indeed humble towards a field that has a lot to offer and a lot to require.

Bibliography

1. Jake VanderPlas. Python Data Science Handbook. O’Reilly Media. 2016

2. Sebastian Raschka. Model evaluation,model selection,and algorithm selection in machine learning. University of Wisconsin–Madison Department of Statistics, November, 2018.

3. Samir Madhavan. Mastering Python for Data Science. Packt Publishing. August, 2015

4. UCI Machine Learning Repository - https://archive.ics.uci.edu/ml/datasets/online+retail - Accessed: 2019-03-02

5. Official Python site - https://www.python.org - Accessed: 2019-04-13

6. Eduardo Rivo, Javier de la Fuente, Ángel Rivo, Eva García-Fontán, Miguel-Ángel Cañizares, Pedro Gil. Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. Received: 30 March 2011 / Accepted: 30 April 2011. DOI: 10.1007/s12094-012-0764-8. Source: PubMed

7. Olegas NIAKŠU. CRISP Data Mining Methodology Extension for Medical Domain. Institute of Mathematics and Informatics Vilnius University. Akademijos g. 4, LT-08663 Vilnius, Lithuania. Page 94.

8. Sinan Ozdemir. Principles of Data Science. Packt Publishing. December, 2016

9. Gil Press. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says - https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#21273e176f63 - Accessed: 2019-04-27

10. Microsoft Azure Notebooks: Data Science Project - https://notebooks.azure.com/azizulhuq/projects/datascienceproject/html/Retail%20Analytics.ipynb - Accessed: 2019-03-05

11. Peter J. Brockwell, Richard A. Davis. Introduction to Time Series and Forecasting. Springer. April, 2010.

12. M. Bremer. Multiple regression. Math 261A. Spring, 2012 - http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement%205%20-%20multiple%20regression.pdf - Accessed: 2019-05-04

13. Statistics How To. Linear Regression: Simple Steps and Video - Find the Equation, Coefficient and Slope-

https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/ - Accessed: 2019-05-07

14. Gareth James , Daniela Witten,Trevor Hastie Robert Tibshirani. An Introduction to Statistical Learning, publisher- Springer New York Heidelberg Dordrecht London,ISBN 978-1-4614-7138-7 (eBook) ,page 59-68 page 71-82

15. Neeraj Kumar. A Review on Machine Learning Algorithms, Tasks and Applications. Chitkara University. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 6, Issue 10, October 2017, ISSN: 2278 – 1323. October, 2017

16. Chun-houh Chen, Wolfgang Härdle, Antony Unwin. Handbook of Data Visualization. Springer. ISBN 978-3-540-33036-3. Page 15-57

17. Paul S.P. Cowpertwait. Introductory Time Series with R (Use R!). Springer. 2009

18. John E. Hanke, Dean Wichern. Business Forecasting. Pearson; 9th edition. 2008