



## **SIX WEEKS SUMMER TRAINING REPORT**

On

### **MICROLEARNING IN DATA SCIENCE**

Submitted by

**Sneha Singh**

**Registration No:11808742**

**Programme Name: B.Tech(CSE)**

Under the Guidance of

**Mr.Somu Baura & Ms.Ananya (Board Infinity)**

**School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**(May-June, 2020)**

## DECLARATION

I hereby declare that I have completed my six weeks summer training at **Board Infinity** from **05/08/2020 to 06/15/2020** under the guidance of **Mr.Somu Baura & Ms.Ananya**. I have declare that I have worked with full dedication during these six weeks of training and my learning outcomes fulfill the requirements of training for the award of degree of **B.Tech**, Lovely Professional University, Phagwara.

Sneha Singh

Name of Student :**Sneha Singh**

Registration no :**11808742**

Date: 09/28/2020

## **ACKNOWLEDGEMENT**

Learning from the Board Infinity was interesting. During these 6 weeks of training, I learnt a lot on 'Data Science', especially the analyzing large data .

I have to thank Mr Kunal Naik,Dr.Rishi,Mr.Amit for his invaluable time and effort in teaching and guiding me. Therefore, I am grateful to the people in the Board Infinity for the chance to learn this course and skills.

I especially want to thank Lovely Professional University for giving me this opportunity to learn the knowledge and skills and to use my summertime productively.

Further on, I want to thank the students and interns in the Board Infinity who made this demanding time joyful but always efficient.

# CERTIFICATE

**BOARD**

## CERTIFICATE OF COMPLETION

THIS CERTIFICATE IS AWARDED TO

**Sneha Singh**

for successfully completing two months program in  
Data Science

June 15, 2020

ISSUED DATE



CEO - BOARD INFINITY

BI20A0106009

CERTIFICATE NO.

## **CONTENTS**

<b>S.No.</b>		<b>Page No.</b>
1.	Introduction.....	06
2.	Technology Learnt.....	07-40
3.	Reason for choosing this Technology.....	41
4.	Projects.....	42-43
5.	Future Scope Of Data Science.....	44
6.	Career in Data Science.....	45
7.	Cocclusion.....	46
8.	Bibilography.....	47

## **INTRODUCTION**

### **What Is Data Science?**

Data science provides meaningful information based on large amounts of complex data or big data. Data science, or data-driven science, combines different fields of work in statistics and computation to interpret data for decision-making purposes.

**Data Science is a combination of mathematics, programming, statistics, data analysis, and machine learning.** By combining all these, Data Science uses advanced algorithms and scientific methods to extract information and insights from large datasets – both structured and unstructured. The advent of Big Data and Machine Learning has further fuelled the growth of Data Science. Today, Data Science is being used across all parallels of various industries, including business, healthcare, finance, and education.



## **TECHNOLOGY LEARNT**

### **1. INTRODUCTION TO DATA SCIENCE AND FIELD CLARITY**

1.1BASICS OF DATA SCIENCE

### **2. BUSINESS ANALYTICS WITH EXCEL**

2.1.ANALYZING DATA USING EXCEL TOOLS

2.2. CREATING PIVOT TABLES

2.3.CREATING PIVOT CHARTS

2.4.ADDING SLICERS

2.4 CONNECTIONS

2.5.DASHBOARD

### **3. DATA VISUALIZATION WITH TABLEAU**

3.1.VISUALIZING THE DATA

3.2. STORY TELLING

3.3.WORKING WITH LARGE DATA SET

### **4. MATHS FOR DATA SCIENCE**

4.1.PROBABILTY

4.2.STATISTICS

4.3.LINEAR REGRESSION

### **5. ADVANCED SQL**

5.1.GROUPING THE DATA

5.2.ORDERING THE DATA

## **6. PYTHON FOR DATA SCIENCE**

6.1.NUMPY

6.2.PANDAS

6.3.MATPLOTT

6.4.SEABORN

6.5.BAR GRAPH

## **7. DEEP DRIVE INTO MACHINE LEARNING**

7.1BASICS OF MACHINE LEARNING

7.1.LINEAR REGRESSION

7.2.ALGORITHMS OF MACHINE LEARNING

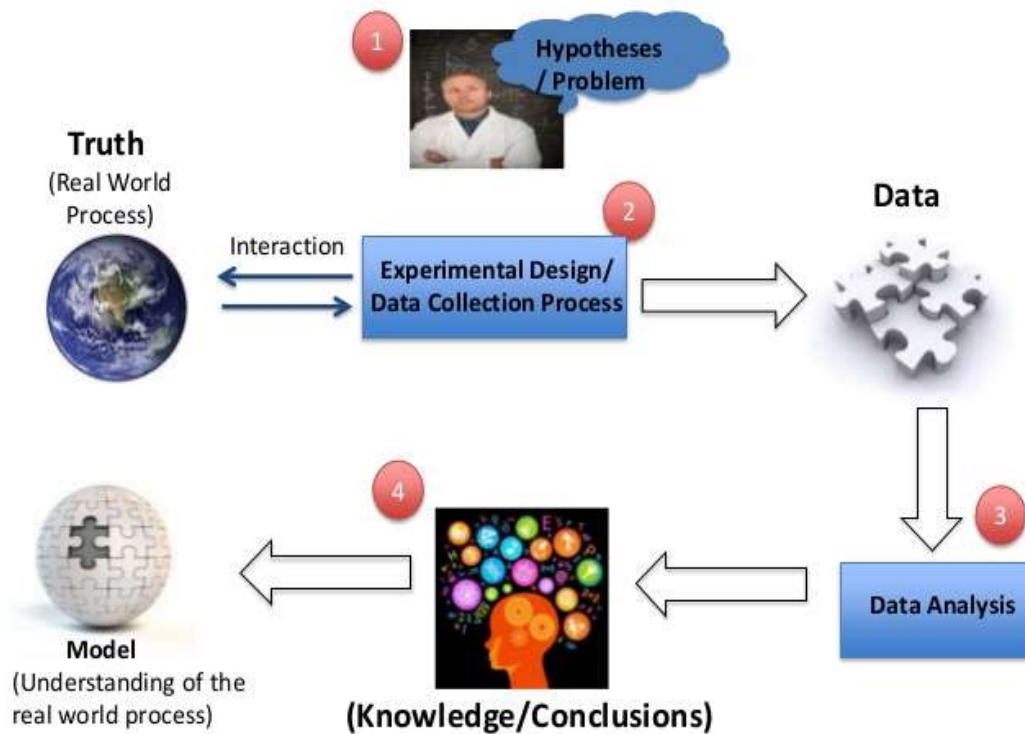


## BASICS OF DATA SCIENCE

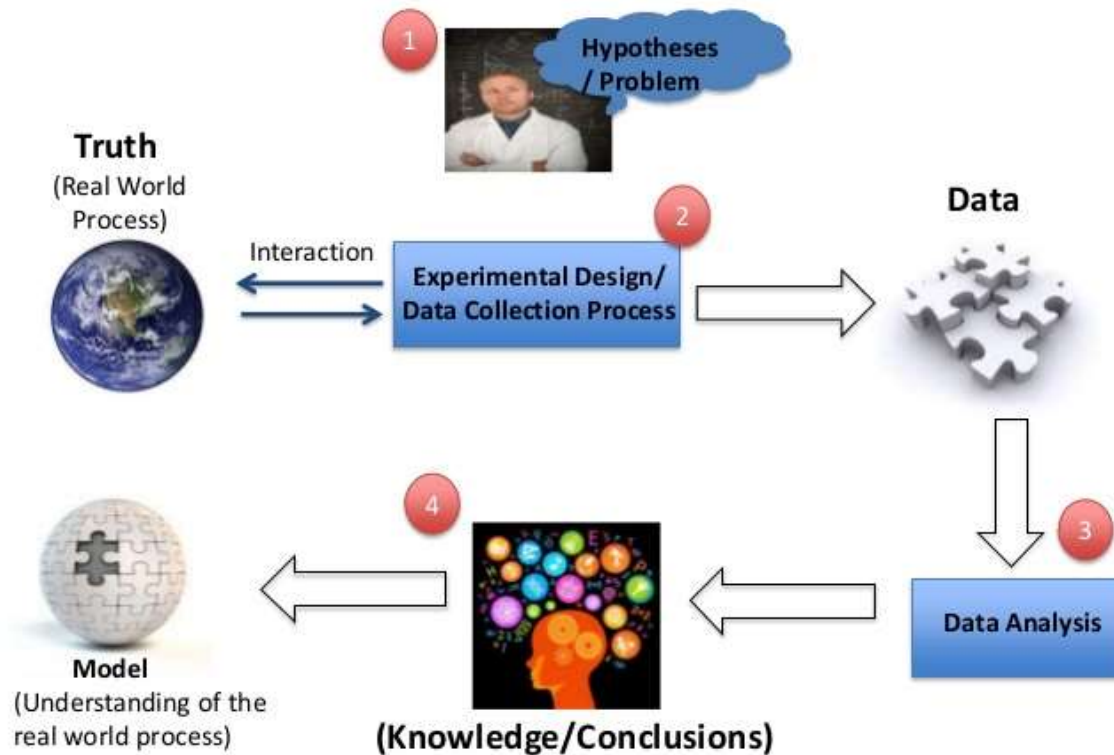


- Data Science = Data Analysis with more rigorous scientific principles

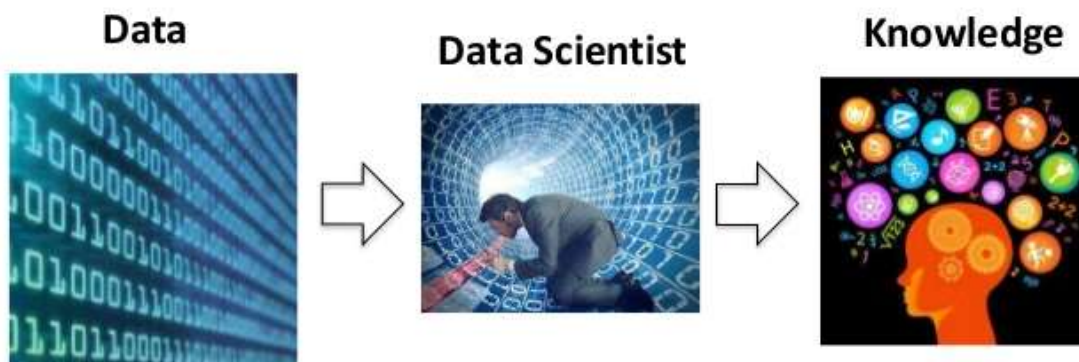
### Process of Data Science



# Process of Data Science

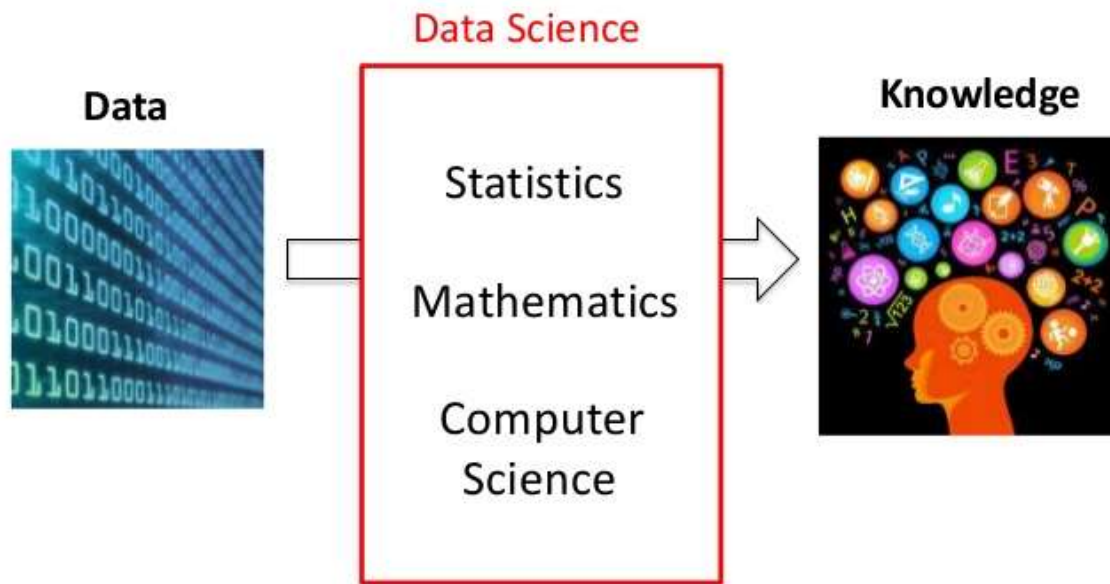


## The Data Scientist?

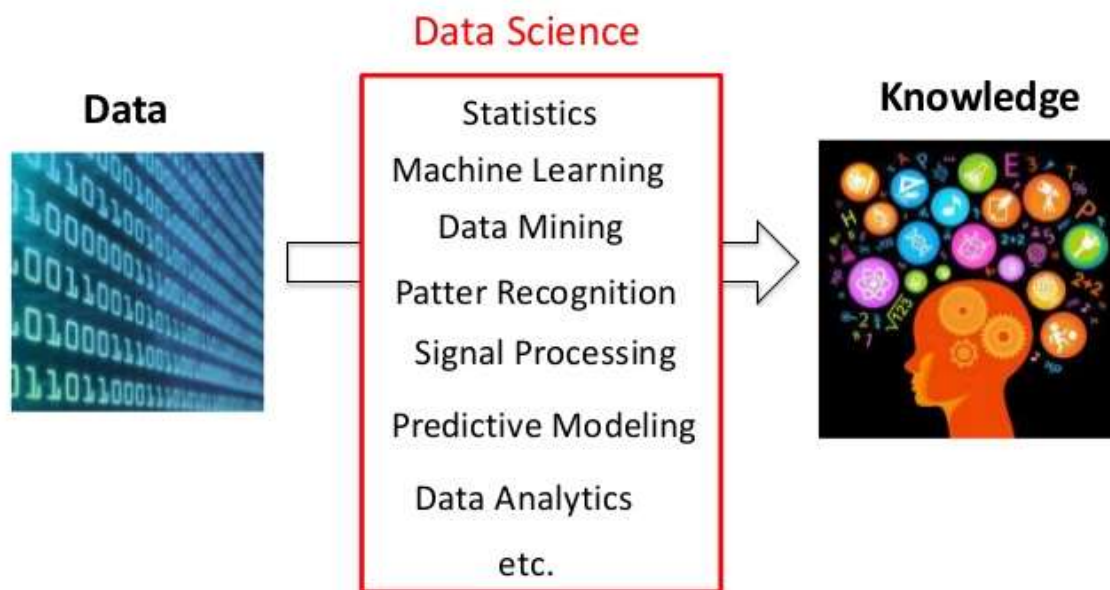


- The person who is involved in the Data Science process

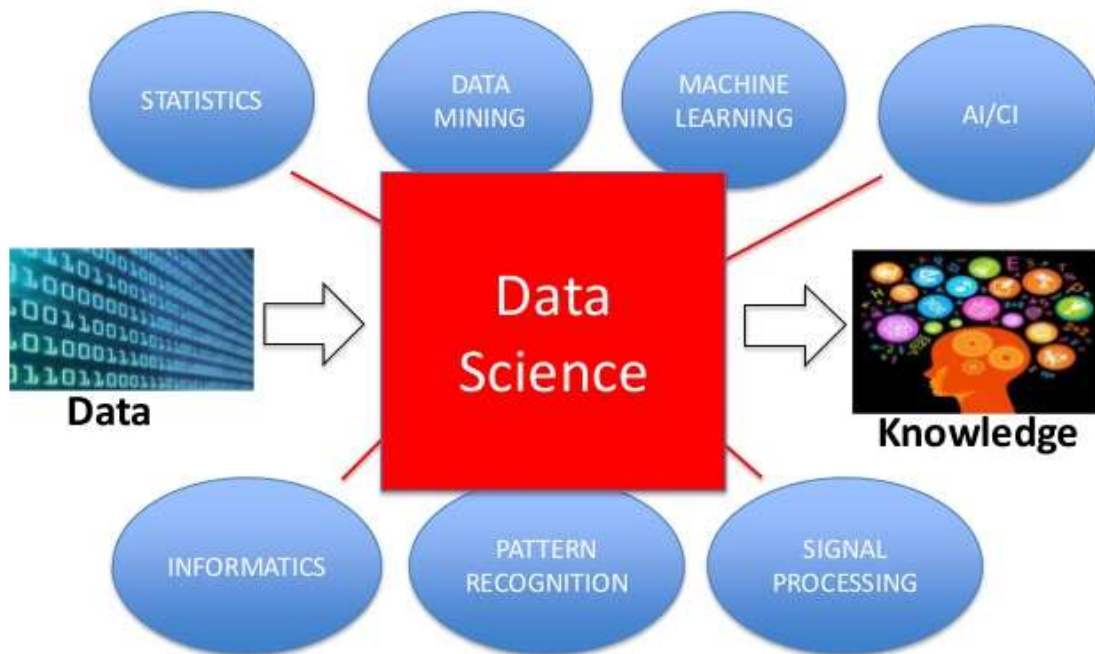
# The Unified Field of Data Analytics



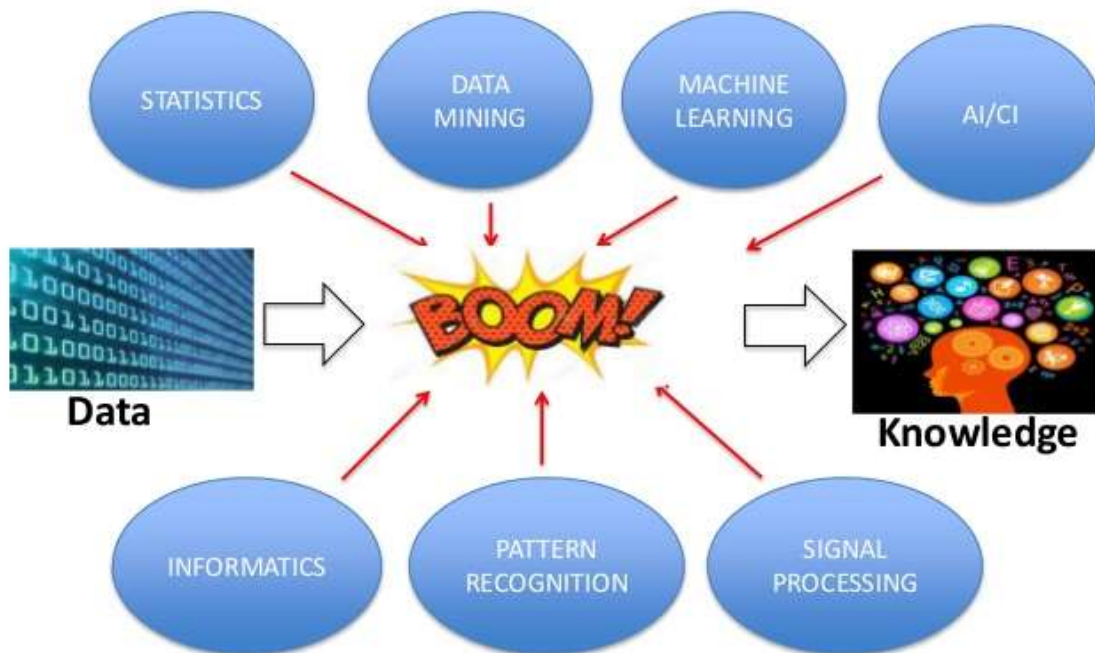
# The Unified Field of Data Analytics



## Birth of Data Science!

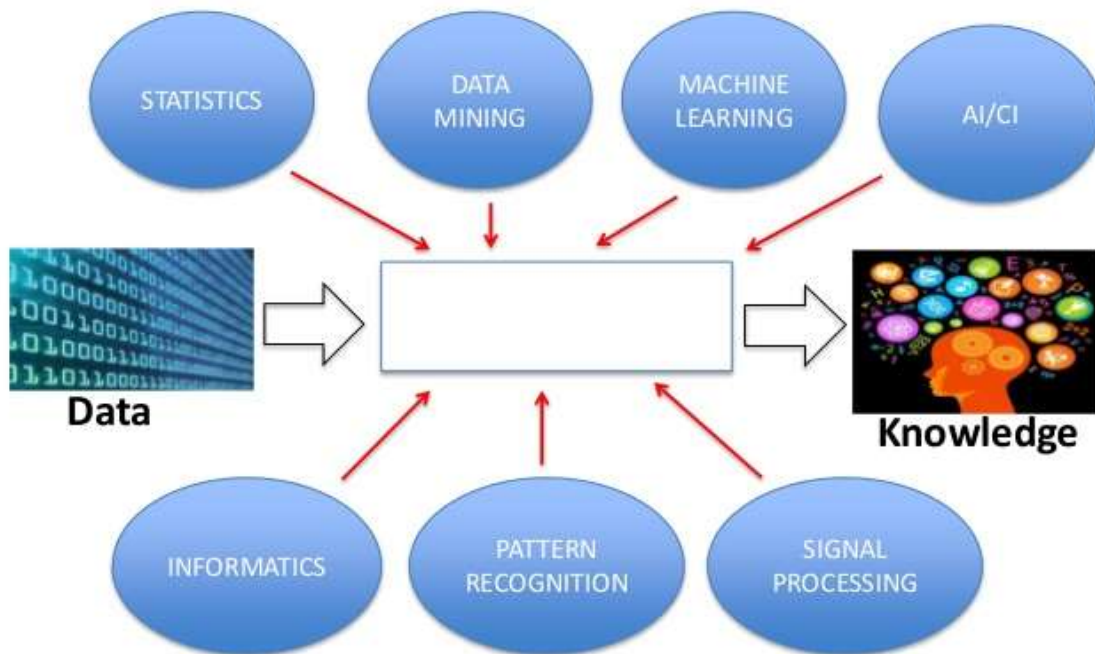


## Several Years Back...

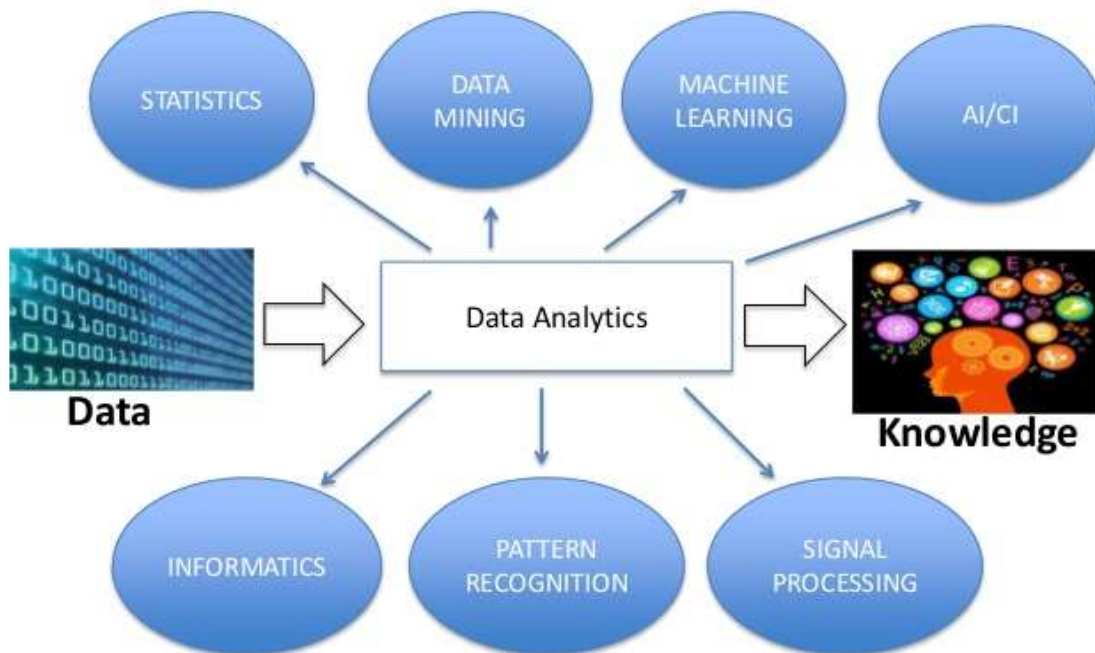




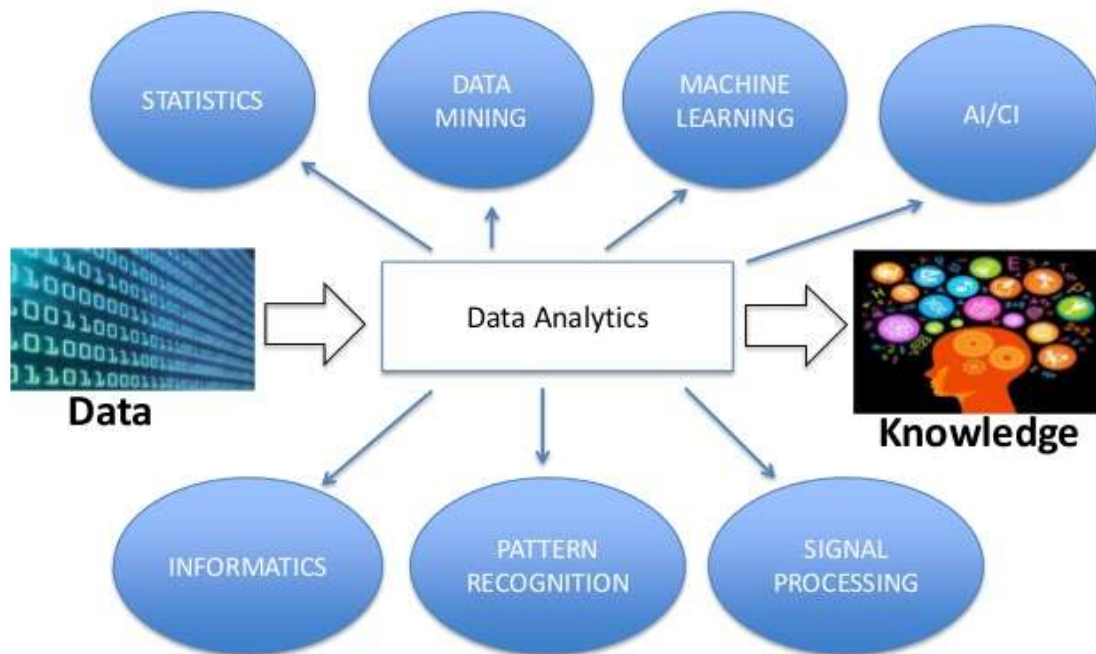
## Several Years Back...



## The existing fields of Data Analytics



## The existing fields of Data Analytics



## The existing fields of Data Analytics



# What is Data Analytics?

- The Process of generating knowledge/insights from data



## A Brief history about Data Science

The term data science has existed for the better part of the last 30 years and was originally used as a substitute for "computer science" in 1960. Approximately 15 years later, the term was used to define the survey of data processing methods used in different applications. In 2001, data science was introduced as an independent discipline. The Harvard Business Review published an [article](#) in 2012 describing the role of the data scientist as the “sexiest job of the 21st century.”

## Role of Data Scientist

Data scientists help companies interpret and manage data and solve complex problems using expertise in a variety of data niches. They generally have a foundation in computer science, modeling, statistics, analytics, and math - coupled with a strong business sense. It's this merging of esoteric intelligence and practical knowledge that makes the data scientist so valuable to a company

## Aspects of data science

### 1. PREDICTIVE ENGINES

To meet the requirements of data scientists, we need to provide a very comprehensive range of analyses and algorithms – but, at the same time, ones that scale to address the huge volumes of data. We can do this by enabling the most comprehensive collection of analyses and algorithms available by using the R integration for the SAP HANA® platform, thereby giving access to the incredibly comprehensive R open source algorithm library and, furthermore, its extensive data visualizations.

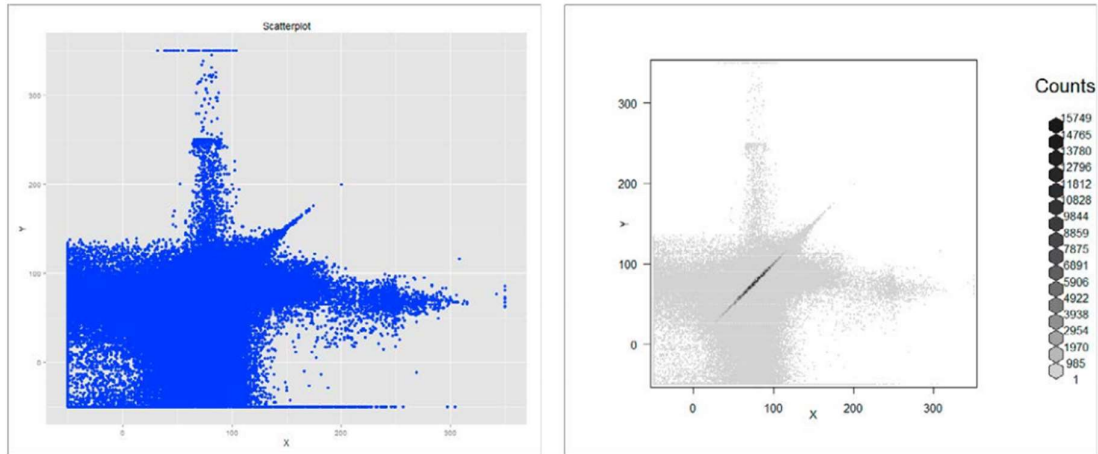
- The R Language for Statistical Computation and Graphics
- Predictive Analysis Library in SAP HANA
- Streaming Analytics with SAP HANA Smart Data Streaming and the PAL
- Automated Data Science with the Automated Predictive Library in SAP HANA
- Comprehensive and Scalable, Automatic and Expert

<b>Association analysis</b> <ul style="list-style-type: none"><li>• Apriori</li><li>• Apriori Lite</li><li>• FP-growth</li><li>• KORD – top K rule discovery</li></ul>	<b>Cluster analysis</b> <ul style="list-style-type: none"><li>• ABC classification</li><li>• DBSCAN</li><li>• K-means</li><li>• K-medoid clustering</li><li>• K-medians</li><li>• Kohonen self-organized maps</li><li>• Agglomerate hierarchical</li><li>• Affinity propagation</li><li>• Latent Dirichlet allocation (LDA)</li><li>• Gaussian mixture model (GMM)</li><li>• Cluster assignment</li></ul>	<b>Probability distribution</b> <ul style="list-style-type: none"><li>• Distribution fit /Weibull analysis</li><li>• Cumulative distribution function</li><li>• Kaplan-Meier survival analysis</li><li>• Quantile function</li></ul>	<b>Statistic functions (univariate)</b> <ul style="list-style-type: none"><li>• Mean, median, variance, standard deviation</li><li>• Kurtosis</li><li>• Skewness</li></ul>
<b>Classification analysis</b> <ul style="list-style-type: none"><li>• CART</li><li>• C4.5 decision tree analysis</li><li>• CHAID decision tree analysis</li><li>• K-nearest neighbor</li><li>• Logistic regression</li><li>• Neural network</li><li>• Naive Bayes</li><li>• Random forest</li><li>• Support vector machine</li><li>• Confusion matrix</li><li>• Area under curve (AUC)</li><li>• Parameter selection/model evaluation</li></ul>	<b>Time-series analysis</b> <ul style="list-style-type: none"><li>• Single exponential smoothing</li><li>• Double exponential smoothing</li><li>• Triple exponential smoothing</li><li>• Forecast smoothing</li><li>• ARIMA/seasonal ARIMA</li><li>• Brown's exponential smoothing</li><li>• Croston method</li><li>• Linear regression with damped trend and seasonal adjust</li><li>• Forecast accuracy measures</li><li>• Test for white noise, trend, seasonality</li></ul>	<b>Outlier detection</b> <ul style="list-style-type: none"><li>• Interquartile range test (Tukey's test)</li><li>• Variance test</li><li>• Anomaly detection</li><li>• Grubbs' outlier test</li></ul>	<b>Statistic functions (multivariate)</b> <ul style="list-style-type: none"><li>• Covariance matrix</li><li>• Pearson's correlations matrix</li><li>• Chi-squared tests</li><li>• Test of quality of fit</li><li>• Test of independence</li><li>• F-test (variance equal test)</li></ul>
<b>Regression</b> <ul style="list-style-type: none"><li>• Multiple linear regression</li><li>• Polynomial regression</li><li>• Exponential regression</li><li>• Bivariate geometric regression</li><li>• Bivariate logarithmic regression</li></ul>		<b>Link prediction</b> <ul style="list-style-type: none"><li>• Common neighbors</li><li>• Jaccard's coefficient</li><li>• Adamic/Adar</li><li>• Katzβ</li></ul>	<b>Other</b> <ul style="list-style-type: none"><li>• Weighted scores table</li><li>• Substitute missing values</li></ul>
		<b>Data preparation</b> <ul style="list-style-type: none"><li>• Sampling</li><li>• Random distribution sampling</li><li>• Binning</li><li>• Scaling</li><li>• Partitioning</li><li>• Principal component analysis (PCA)</li></ul>	

### 2. DATA VISUALIZATION

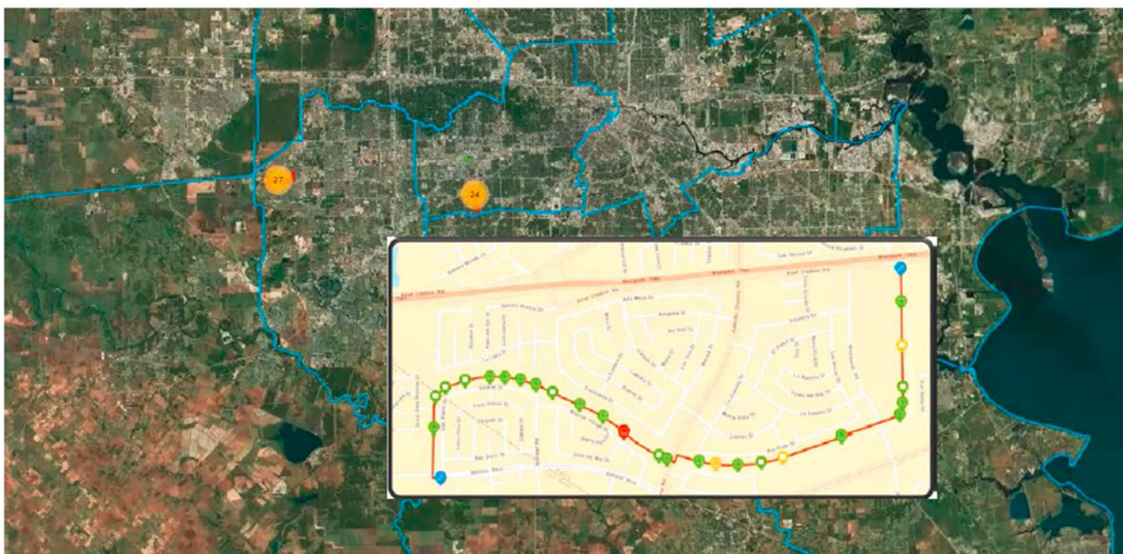
Large data-volume data visualization is a challenge. A scatter plot of a million data points may appear as a solid rectangle. We need visual interactivity and smarter representations of numerous data points. For example, Figure 9 shows a scatter plot of a large data set from an SAP IoT predictive maintenance project, then presented again as a hexbin plot with color graduation representing data volume from which a pattern can be discerned.





### 3. GEOSPATIAL DATA ANALYSIS

Our natural understanding of our world is through spatial analysis – mapping where things are and seeing how they relate. SAP HANA includes a multilayered spatial engine supporting spatial columns, spatial access methods, and spatial reference systems, to deliver high performance and results in everything from modeling and storage, to analysis and presentation of spatial data. With these enhanced geographical information system features, SAP HANA provides a common database for both business and spatial data. The spatial edition of SAP HANA includes spatial clustering using the algorithms – grid, k means, and DBSCAN (density-based spatial clustering of applications with noise). Spatial clustering can be performed on a set of geospatial points in SAP HANA.



### 4. SERIES DATA PROCESSING

When monitoring machine efficiency, energy consumption, or network flow, the ability to monitor data over time enables you to investigate and act on patterns in the series data. SAP HANA supports series data processing to enable efficient processing of large volumes of series data in conjunction with business data to assess business impact. This is critical

functionality for IoT and predictive maintenance applications in which series data volumes are huge.

## **5. UNSTRUCTURED DATA ANALYSIS**

Data science is mainly associated with structured data analysis – in other words, the analysis of data with a structure to it, usually in the form of variables or columns, by records or rows. However, there is a huge amount of data in unstructured formats, such as documents, e-mails, and blogs, which is generally textual, and hence the term “text analysis” is used when trying to analyze this unstructured content. It is said that up to 80% to 90% of enterprise relevant information originates in unstructured data residing inside or outside an organization, such as in blogs, forum postings, social media, Wikis, e-mails, contact-center notes, surveys, service entries, warranty claims, and so on.

SAP HANA supports text-search, text-analysis, and text-mining functionality for unstructured text sources. It supports full-text and fuzzy search using a full-text index to preprocess text linguistically, using techniques such as normalization, tokenization, word stemming, and part-of-speech tagging.

## **6. SIMULATION – DETERMINISTIC AND PROBABILISTIC, AND OPTIMIZATION**

Simulation can take the form of deterministic modeling, whereby specific data values are used to model processes or operations and sensitivity analysis or what-if analysis is used to explore the inherent uncertainty in the data. Simulation in the form of probabilistic modeling explores the uncertainty through assigning probability distributions to the input data for a model and calculating the probability distributions for the output variables. For example, in a capital investment appraisal, you can estimate finding the probability of achieving specific net present values or discounted cash flow yields of the cash flow.

Optimization may be used to determine the overall optimal capital investment program subject to constraints such as the total amount invested and the required individual project investment levels, for example for maintenance. Both simulation and optimization are supported in SAP HANA through application function libraries.

## **7. DEEP LEARNING ON SENSOR DATA**

The relevance of deep learning to IoT comes from the huge volumes of data generated by sensors. Applications include image recognition, speech recognition, and robotics and motor control. Deep learning has been described as a set of algorithms that “mimics the brain” and is equated to neural networks that “learn in layers.”

## **8. EDGE COMPUTING**

Computing on the edge is very important when a very quick response from the system is required – for instance, in the automotive area, the interaction between a navigation system has to be very quick when the data science component is trying to optimize fuel consumption by taking the driving style into account. It is also required when the data volume generated on-site is so large that the throughput required to process it by a central application, together with the incoming streams from other sources, cannot be provided. This may be the case in scenarios in which high-resolution images or videos need to be analyzed.

## **HOW TO ANALYZE DTA IN EXCEL**

To know how to analyze data in excel, you can instantly create different types of charts, including line and column charts, or add miniature graphs. You can also apply a table style, create PivotTables, quickly insert totals, and apply conditional formatting. Analyzing large data sets with Excel makes work easier if you follow a few simple rules:

- Select the cells that contain the data you want to analyze.
- Click the Quick Analysis button image button that appears to the bottom right of your selected data (or press CTRL + Q).
- Selected data with Quick Analysis Lens button visible
- In the Quick Analysis gallery, select a tab you want. For example, choose Charts to see your data in a chart.
- Pick an option, or just point to each one to see a preview.
- You might notice that the options you can choose are not always the same. That is often because the options change based on the type of data you have selected in your workbook.

To understand the best way to analyze data in excel, you might want to know which analysis option is suitable for you. Here we offer you a basic overview of some of the best options to choose from.

- **Formatting:** Formatting lets you highlight parts of your data by adding things like data bars and colors. This lets you quickly see high and low values, among other things.
- **Charts:** Charts Excel recommends different charts, based on the type of data you have selected. If you do not see the chart you want, click **More Charts**.
- **Totals:** Totals let you calculate the numbers in columns and rows. For example, Running Total inserts a total that grows as you add items to your data. Click the little black arrows on the right and left to see additional options.
- **Tables:** Tables make it easy to filter and sort your data. If you do not see the table style you want, click More.

### **How to Analyze Data in Excel: Data Analysis**

Data Analysis is simpler and faster with Excel analytics. Here, we offer some tips for work:

- **Create auto expandable ranges with Excel tables:** One of the most underused features of MS Excel is Excel Tables. Excel Tables have wonderful properties that allow you to work more efficiently. Some of these features include:
- **Formula Auto Fill:** Once you enter a formula in a table it will be automatically be copied to the rest of the table.
- **Auto Expansion:** New items typed below or at the right of the table become part of the table.
- **Visible headers:** Regardless of your position within the table, your headers will always be visible.
- **Automatic Total Row:** To calculate the total of a row, you just have to select the desired formula.
- **Use Excel Tables as part of a formula:** Like in dropdown lists, if you have a formula that depends on a Table, when you add new items to the Table, the reference in the formula will be automatically updated.
- **Use Excel Tables as a source for a chart:** Charts will be updated automatically as well if you use an Excel Table as a source. As you can see, Excel Tables allow you to create data sources that do not have to be updated when new data is included.

## **PIVOT TABLES**

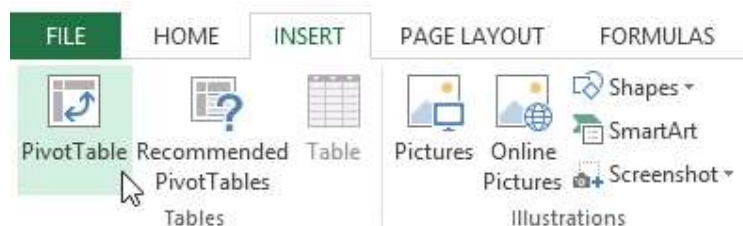
When you have a lot of data, it can sometimes be difficult to analyze all of the information in your worksheet. **PivotTables** can help make your worksheets more manageable by **summarizing** data and allowing you to **manipulate** it in different ways.

### **CREATING PIVOT TABLES**

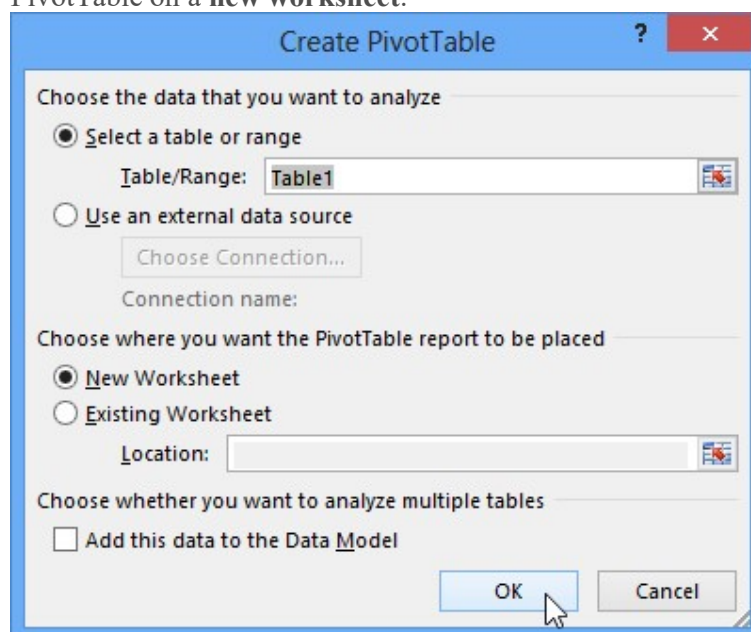
1. Select the **table** or **cells** (including column headers) containing the data you want to use.

	A	B	C	D	E	F
1	Salesperson	Region	Account	Order Amount	Month	
2	Albertson, Kathy	East	29386	\$925.00	January	
3	Albertson, Kathy	East	74830	\$875.00	February	
4	Albertson, Kathy	East	90099	\$500.00	February	
5	Albertson, Kathy	East	74830	\$350.00	March	
6	Brennan, Michael	West	82853	\$400.00	January	
7	Brennan, Michael	West	72949	\$850.00	January	
8	Brennan, Michael	West	90044	\$1,500.00	January	
9	Brennan, Michael	West	82853	\$550.00	February	
10	Brennan, Michael	West	72949	\$400.00	March	
11	Davis, William	South	55223	\$235.00	February	
12	Davis, William	South	10354	\$850.00	January	
13	Davis, William	South	50192	\$600.00	March	
14	Davis, William	South	27589	\$250.00	January	
15	Dumlao, Richard	West	67275	\$400.00	January	

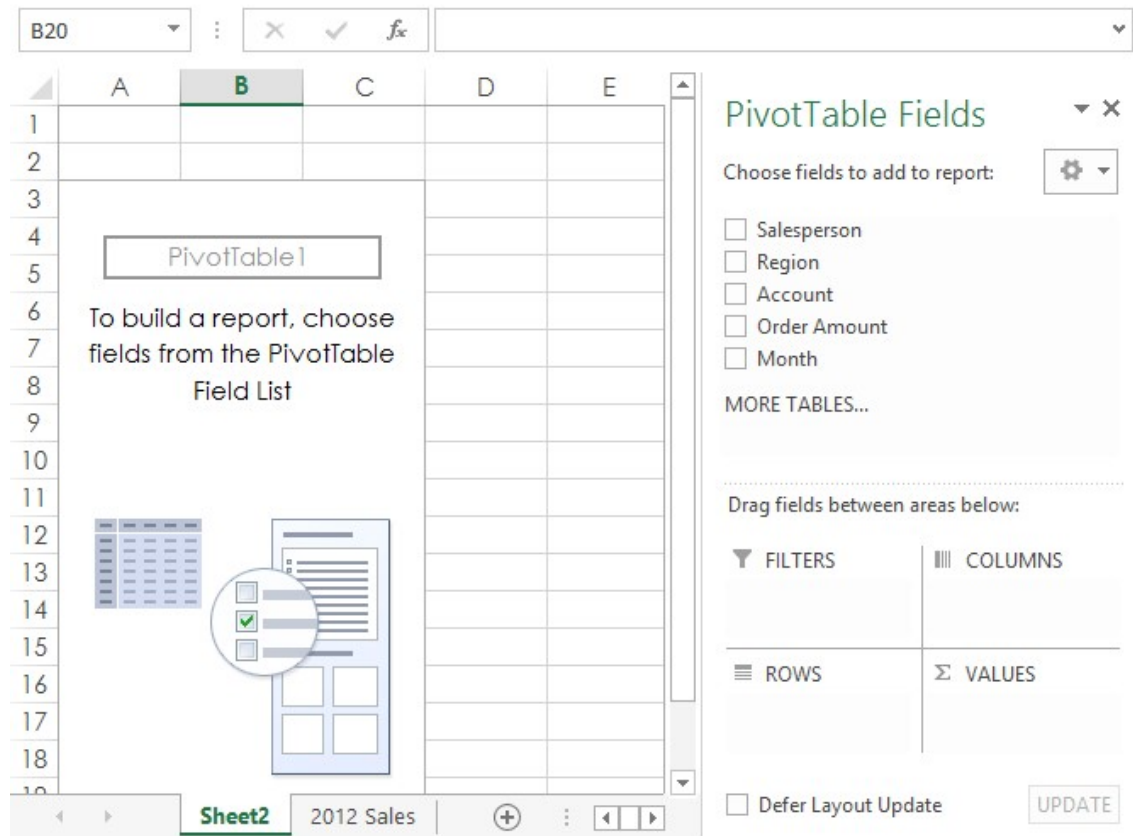
- From the **Insert** tab, click the **PivotTable** command.



- The **Create PivotTable** dialog box will appear. Choose your settings, then click **OK**. In our example, we'll use **Table1** as our source data and place the PivotTable on a **new worksheet**.



4. A blank **PivotTable** and **Field List** will appear on a new worksheet.



5. Once you create a PivotTable, you'll need to decide which **fields** to add. Each field is simply a **column header** from the source data. In the **PivotTable Field List**, check the box for each field you want to add. In our example, we want to know the total **amount** sold by each **salesperson**, so we'll check the **Salesperson** and **Order Amount** fields.



6. The selected fields will be added to one of the four areas below the Field List. In our example, the **Salesperson** field has been added to the **Rows** area, while the **Order Amount** has been added to the **Values** area. Alternatively, you can click, hold, and drag a field to the desired area.



## PivotTable Fields

Choose fields to add to report:

- ☒ Salesperson
- ☐ Region
- ☐ Account
- ☒ Order Amount
- ☐ Month

MORE TABLES...

Drag fields between areas below:

FILTERS

COLUMNS

ROWS

VALUES

Salesperson

Sum of Order ...

☐ Defer Layout Update

UPDATE

7. The PivotTable will calculate and summarize the selected fields. In our example, the PivotTable shows the amount sold by each salesperson

	A	B	C
1			
2			
3	<b>Row Labels</b>	<b>Sum of Order Amount</b>	
4	Albertson, Kathy	2650	
5	Brennan, Michael	3700	
6	Davis, William	1935	
7	Dumlao, Richard	1490	
8	Flores, Tia	4565	
9	Post, Melissa	1690	
10	Thompson, Shannon	3160	
11	Walters, Chris	4375	
12	<b>Grand Total</b>	<b>23565</b>	
13			
14			
15			
16			
17			
18			
19			

Total amount sold by each salesperson

## PivotTable Fields

Choose fields to add to report:

- ☒ Salesperson
- ☐ Region
- ☐ Account
- ☒ Order Amount
- ☐ Month

MORE TABLES...

Drag fields between areas below:

FILTERS

COLUMNS

ROWS

VALUES

Salesperson

Sum of Order ...

☐ Defer Layout Update

UPDATE

Just like with normal spreadsheet data, you can sort the data in a PivotTable using the **Sort & Filter** command in the Home tab. You can also apply any type of **number formatting** you want. For example, you may want to change the **Number Format** to **Currency**. However, be aware that some types of formatting may disappear when you modify the PivotTable.

Row Labels	Sum of Order Amount
Flores, Tia	\$4,565.00
Walters, Chris	\$4,375.00
Brennan, Michael	\$3,700.00
Thompson, Shannon	\$3,160.00
Albertson, Kathy	\$2,650.00
Davis, William	\$1,935.00
Post, Melissa	\$1,690.00
Dumlao, Richard	\$1,490.00
<b>Grand Total</b>	<b>\$23,565.00</b>

If you change any of the data in your source worksheet, the PivotTable **will not update automatically**. To manually update it, select the PivotTable and then go to **Analyze > Refresh**.

## Pivoting data

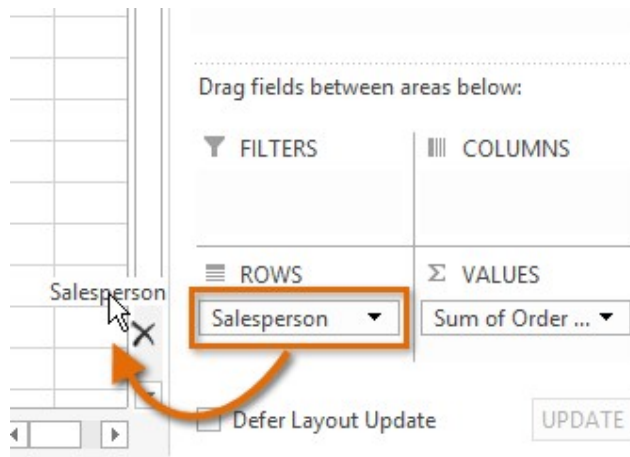
One of the best things about PivotTables is that they can quickly **pivot**—or reorganize—data, allowing you to look at your worksheet data in different ways. Pivoting data can help you answer **different questions** and even **experiment** with the data to discover new trends and patterns.

In our example, we used the PivotTable to answer the question: **What is the total amount sold by each salesperson?** But now we'd like to answer a new question: **What is the total amount sold in each month?** We can do this by simply changing the field in the **Rows** area.

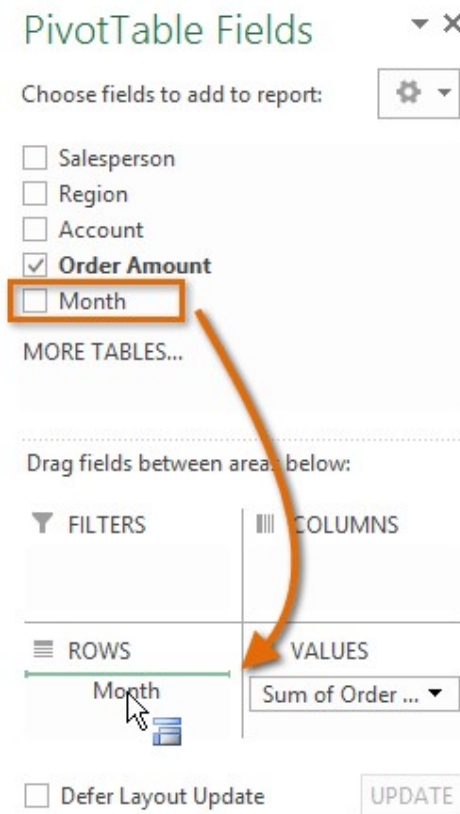
### *To change the row:*

1. Click, hold, and drag any existing **fields** out of the **Rows** area. The field will disappear.





2. Drag a new field from the **Field List** into the **Rows** area. In our example, we'll use the **Month** field.



3. The PivotTable will adjust—or pivot—to show the new data. In our example, it now shows the total order amount for each month.

Row Labels	Sum of Order Amount
January	9090
February	9160
March	5315
<b>Grand Total</b>	<b>23565</b>

**PivotTable Fields**

Choose fields to add to report:

- ☐ Salesperson
- ☐ Region
- ☐ Account
- ☒ Order Amount
- ☒ Month

MORE TABLES...

Drag fields between areas below:

FILTERS	COLUMNS
ROWS	VALUES
Month	Sum of Order ...

☐ Defer Layout Update UPDATE

### *To add columns:*

So far, our PivotTable has only shown **one column** of data at a time. In order to show **multiple columns**, you'll need to add a field to the **Columns** area.

1. Drag a field from the **Field List** into the **Columns** area. In our example, we'll use the **Region** field.

**PivotTable Fields**

Choose fields to add to report:

- ☐ Salesperson
- ☐ Region
- ☐ Account
- ☒ Order Amount
- ☒ Month

MORE TABLES...

Drag fields between areas below:

FILTERS	COLUMNS
	Region
ROWS	VALUES
Month	Sum of Order ...

☐ Defer Layout Update UPDATE

2. The PivotTable will include multiple columns. In our example, there is now a column for each region.

Sum of Order Amount		Column Labels				
Row Labels	East	North	South	West	Grand Total	
January	1690	1140	3110	3150	9090	
February	1950	1720	3975	1515	9160	
March	700	300	3790	525	5315	
<b>Grand Total</b>	<b>4340</b>	<b>3160</b>	<b>10875</b>	<b>5190</b>	<b>23565</b>	

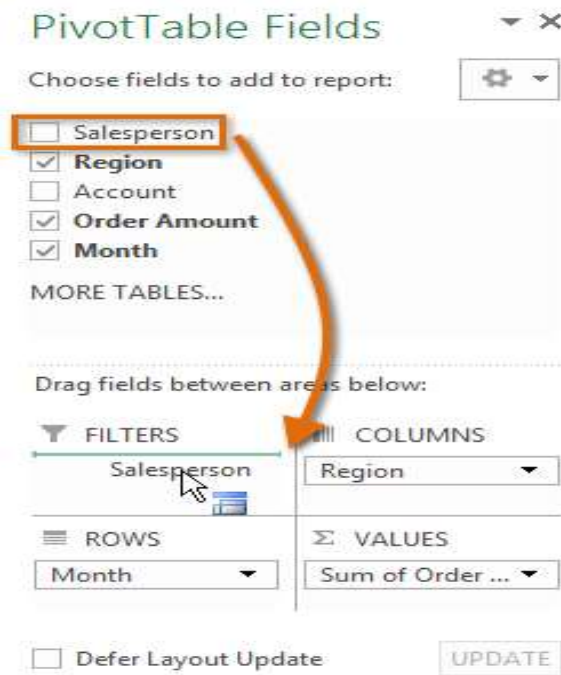
## Filters

Sometimes you may want focus on just a certain section of your data. Filters can be used to narrow down the data in your PivotTable, allowing you to view only the information you need.

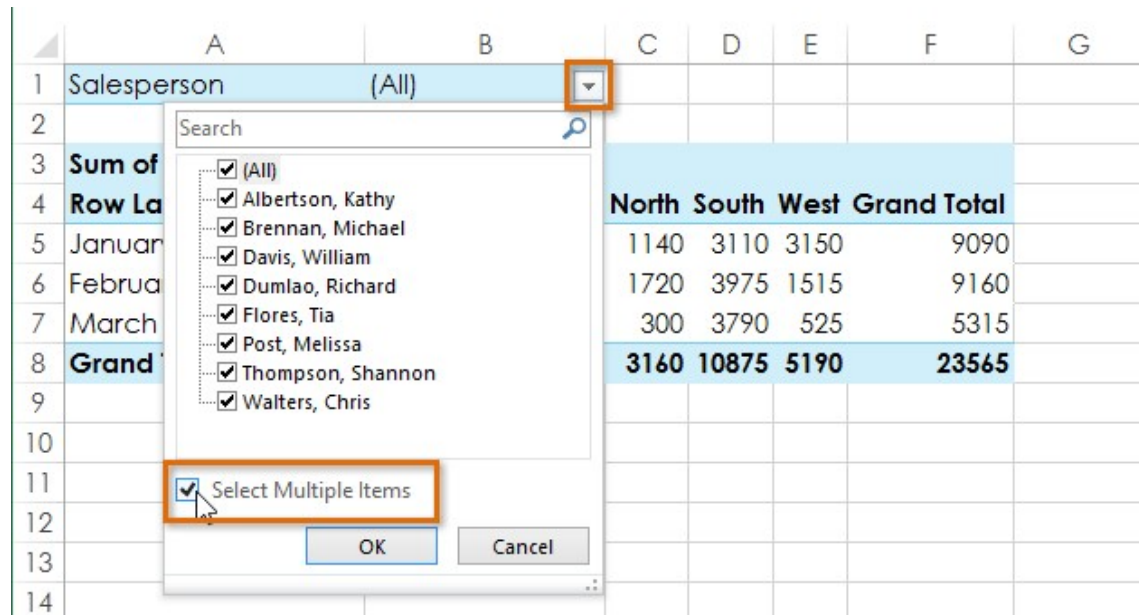
### *To add a filter:*

In our example, we'll filter out certain salespeople to determine how they affect the total sales.

1. Drag a field from the **Field List** to the **Filters** area. In this example, we'll use the **Salesperson** field.

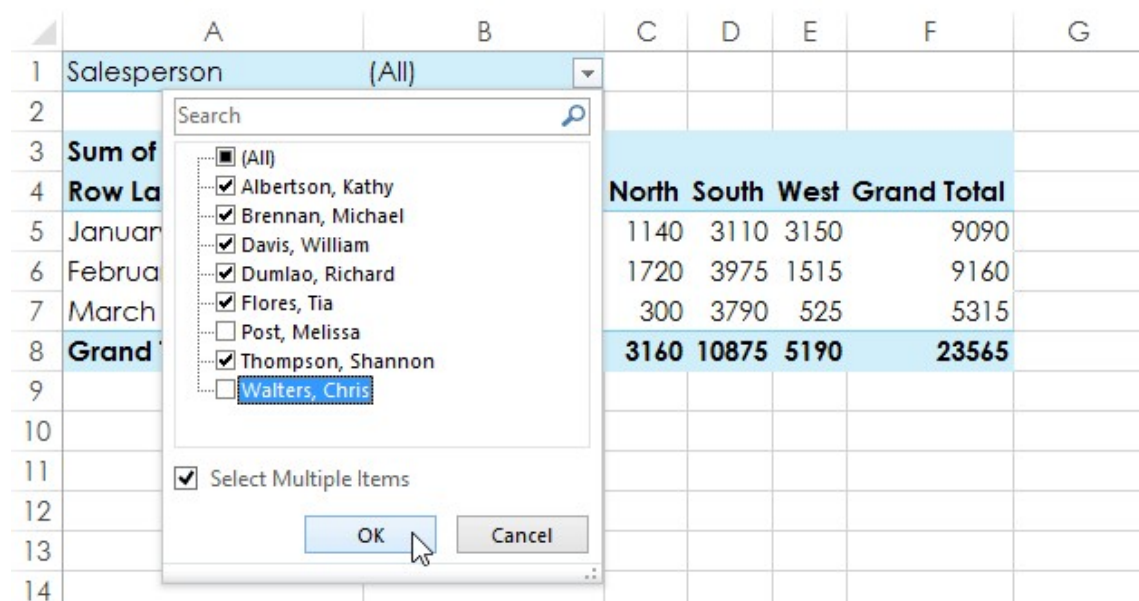


- The **filter** will appear above the PivotTable. Click the **drop-down arrow**, then check the box next to **Select Multiple Items**.



	A	B	C	D	E	F	G
1	Salesperson	(All)					
2							
3	Sum of						
4	Row Labels						
5	January		1140	3110	3150		9090
6	February		1720	3975	1515		9160
7	March		300	3790	525		5315
8	Grand Total		3160	10875	5190		23565

- Uncheck** the box for any items you don't want to include in the PivotTable. In our example, we'll uncheck the boxes for a few different salespeople, then click **OK**.



	A	B	C	D	E	F	G
1	Salesperson	(All)					
2							
3	Sum of						
4	Row Labels						
5	January		1140	3110	3150		9090
6	February		1720	3975	1515		9160
7	March		300	3790	525		5315
8	Grand Total		3160	10875	5190		23565

- The PivotTable will adjust to reflect the changes.

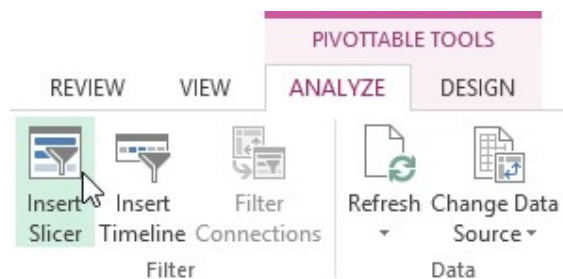
	A	B	C	D	E	F	G
1	Salesperson	(Multiple Items) ▼					
2							
3	Sum of Order Amount	Column Labels ▼					
4	Row Labels ▼	East	North	South	West	Grand Total	
5	January	925	1140	2755	3150	7970	
6	February	1375	1720	1220	1515	5830	
7	March	350	300	2525	525	3700	
8	Grand Total	2650	3160	6500	5190	17500	
9							
10							

## Slicers

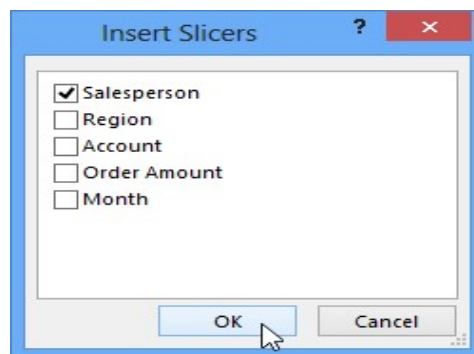
**Slicers** make filtering data in PivotTables even easier. Slicers are basically just **filters**, but they're easier and faster to use, allowing you to instantly pivot your data. If you frequently filter your PivotTables, you may want to consider using slicers instead of filters.

### *To add a slicer:*

1. Select any cell in the PivotTable.
2. From the **Analyze** tab, click the **Insert Slicer** command.



3. A dialog box will appear. Select the desired **field**. In our example, we'll select **Salesperson**, then click **OK**.



4. The slicer will appear next to the PivotTable. Each selected item will be highlighted in **blue**. In the example below, the slicer contains a list of all salespeople, and **six** of them are currently selected.

	A	B	C	D	E	F	G
1	Salesperson	(Multiple Items)					
2							
3	Sum of Order Amount	Column Labels					
4	Row Labels	East	North	South	West	Grand Total	
5	January	925	1140	2755	3150	7970	
6	February	1375	1720	1220	1515	5830	
7	March	350	300	2525	525	3700	
8	Grand Total	2650	3160	6500	5190	17500	
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							

Salesperson
Albertson, Kathy
Brennan, Michael
Davis, William
Dumlao, Richard
Flores, Tia
Post, Melissa
Thompson, Shannon
Walters, Chris

5. Just like **filters**, only **selected** items are used in the PivotTable. When you **select** or **deselect** items, the PivotTable will instantly reflect the changes. Try selecting different items to see how they affect the PivotTable. Press and hold the **Ctrl** key on your keyboard to select multiple items from a slicer.

Salesperson
Albertson, Kathy
Brennan, Michael
Davis, William
Dumlao, Richard
Flores, Tia
Post, Melissa
Thompson, Shannon
Walters, Chris



You can also click the **Filter icon** in the top-right corner to select all items from the slicer at once.

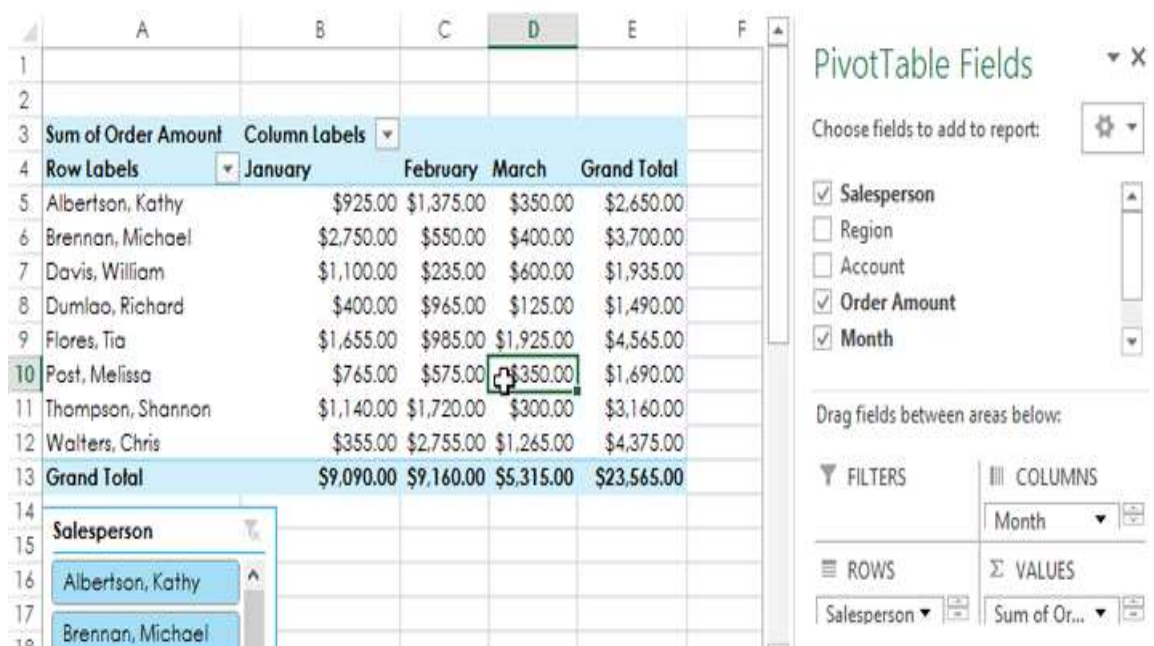
## PivotCharts

**PivotCharts** are like regular charts, except they display data from a **PivotTable**. Just like regular charts, you'll be able to select a **chart type**, **layout**, and **style** that will best represent the data.

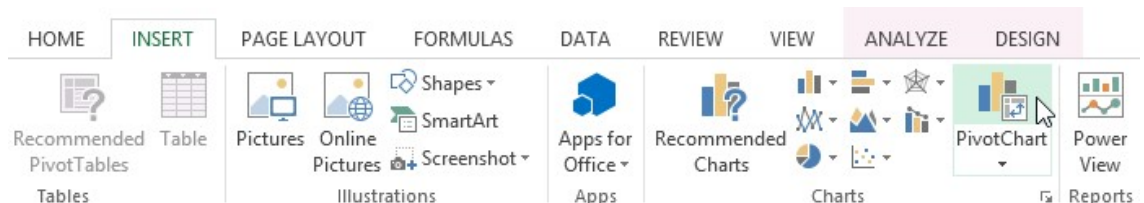
### *To create a PivotChart:*

In this example, our PivotTable is showing each person's total sales per month. We'll use a PivotChart so we can see the information more clearly.

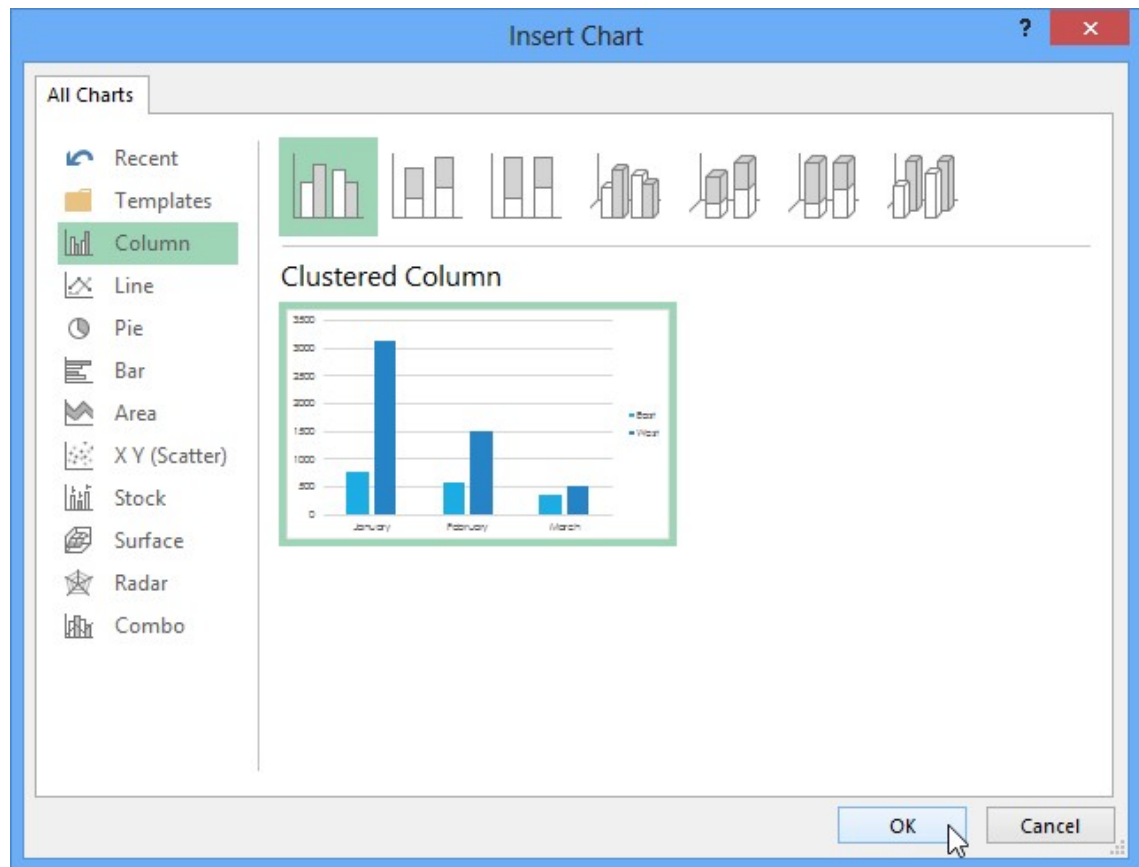
1. Select any cell in your PivotTable.



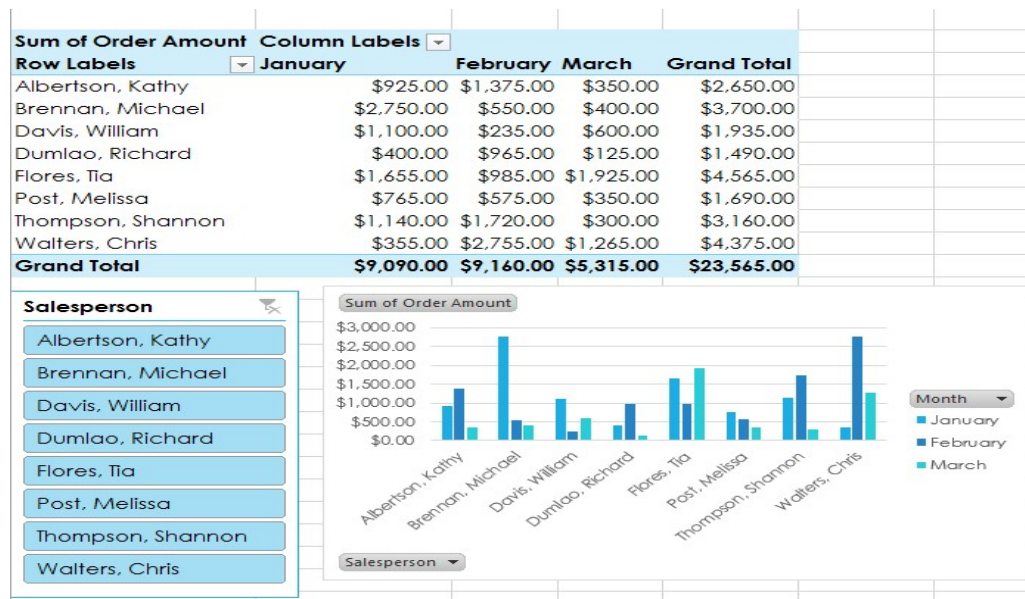
2. From the **Insert** tab, click the **PivotChart** command.



3. The **Insert Chart** dialog box will appear. Select the desired **chart type** and **layout**, then click **OK**.

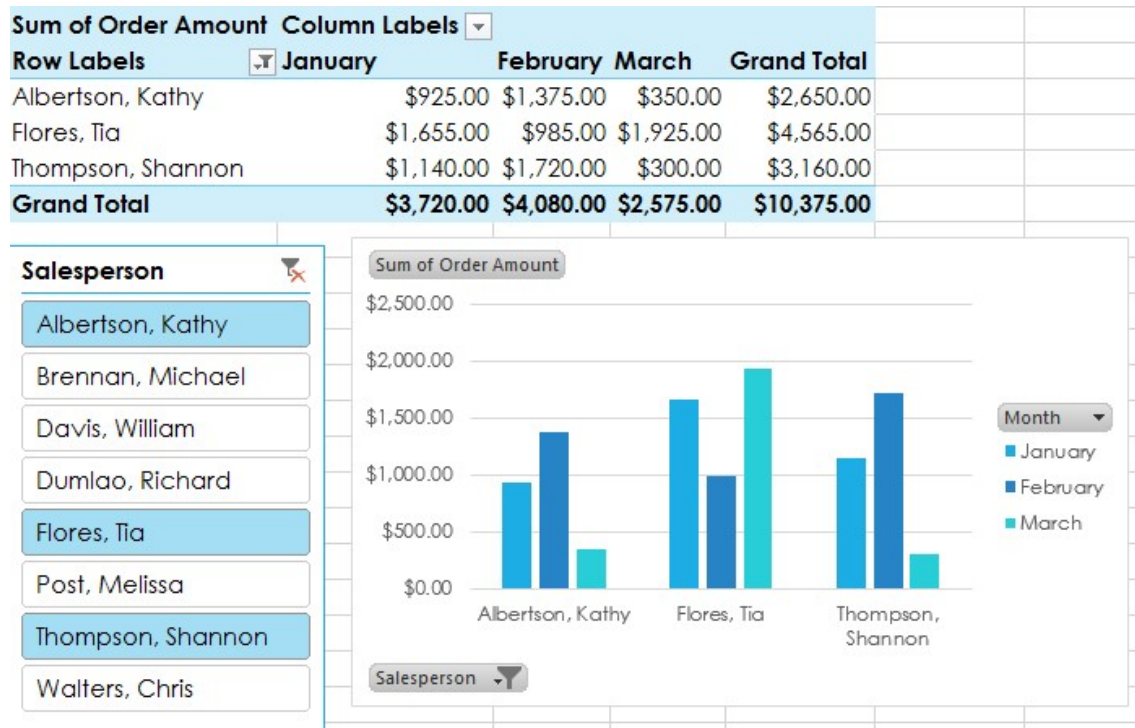


4. The PivotChart will appear.



Try using **slicers** or **filters** to change the data that is displayed. The PivotChart will automatically adjust to show the new data.





## DATA VISUALIZATION WITH TABLEAU

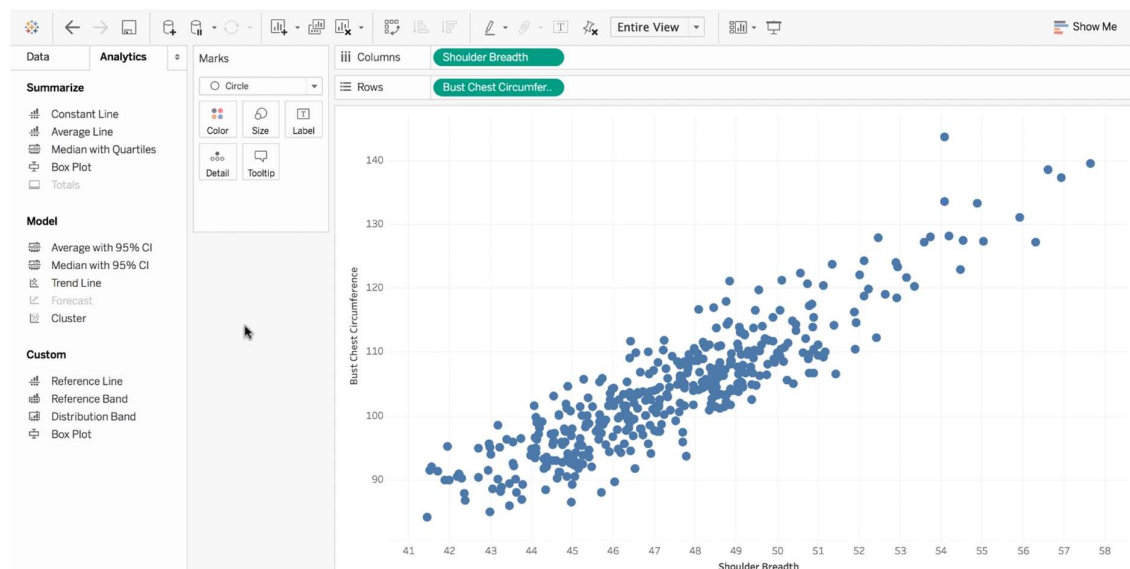
Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Why data visualization is important for any career

It's hard to think of a professional industry that doesn't benefit from making data more understandable. Every STEM field benefits from understanding data—and so do fields in government, finance, marketing, history, consumer goods, service industries, education, sports, and so on.

While we'll always wax poetically about data visualization (you're on the Tableau website, after all) there are practical, real-life applications that are undeniable. And, since visualization is so prolific, it's also one of the most useful professional skills to develop. The better you can convey your points visually, whether in a dashboard or a slide deck, the better you can leverage that information.

Examples of data visualization in action



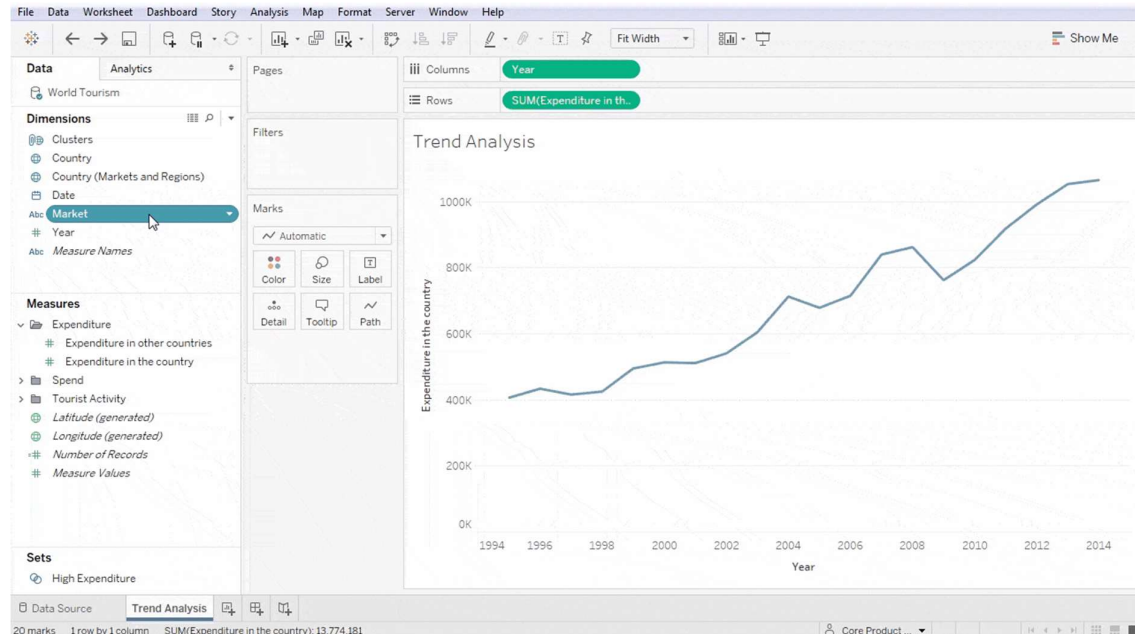
Of course, one of the best ways to understand data visualization is to see it. What a crazy concept!

With public data visualization galleries and data everywhere online, it can be overwhelming to know where to start. We've collected 10 of the best examples of data visualization of all time, with examples that map historical conquests, analyze film scripts, reveal hidden causes of mortality, and more.

Tableau's own public gallery shows off loads of visualizations made with the free Tableau Public tool, we feature some common starter business dashboards as usable templates,

and Viz of the Day collects some of the best community creations. Plus, there are tons of great blogs and books about data visualization containing excellent examples, explanations, and information about best practices.

## The different types of visualizations



When you think of data visualization, your first thought probably immediately goes to simple bar graphs or pie charts. While these may be an integral part of visualizing data and a common baseline for many data graphics, the right visualization must be paired with the right set of information. Simple graphs are only the tip of the iceberg. There's a whole selection of visualization methods to present data in effective and interesting ways.

## Common general types of data visualization:

- Charts
- Tables
- Graphs
- Maps
- Infographics
- Dashboards

## More specific examples of methods to visualize data:

- Area Chart
- Bar Chart
- Box-and-whisker Plots
- Bubble Cloud

- Bullet Graph
- Cartogram
- Circle View
- Dot Distribution Map
- Gantt Chart
- Heat Map
- Highlight Table
- Histogram
- Matrix
- Network
- Polar Area
- Radial Tree
- Scatter Plot (2D or 3D)
- Streamgraph
- Text Tables
- Timeline

## **PYTHON FOR DATA SCIENCE**

Python is the number one choice of programming language for many data scientists and analysts. One of the reasons of this choice is that python is relatively easier to learn and use. More importantly, there is a wide variety of third party libraries that ease lots of tasks in the field of data science. For instance, **numpy** and **pandas** are great data analysis libraries. **Scikit-learn** and **tensorflow** are for machine learning and data preprocessing tasks. We also have python data visualization libraries such as **matplotlib** and **seaborn**. There are many more useful libraries for data science in python ecosystem.

### **10 IN-BUILT FUNCTION**

#### **1. Set**

Set function returns a set object from an iterable (e.g. a list). Set is an unordered sequence of unique values. Thus, unlike lists, sets do not have duplicate values.

#### **2. Enumerate**

Enumerate provides a count when working with iterables. It can be placed in a for loop or directly used on an iterable to create an enumerate object. Enumerate object is basically a list of tuples.

#### **3. Isinstance**

Isinstance function is used to check or compare the classes of objects. It accepts an object and a class as arguments and returns True if the object is an instance of that class.

#### **4. Len**

Len function returns the length of an object or the number of items in an object.

## 5. Sorted

As the name clearly explains, sorted function is used to sort an iterable. By default, sorting is done in ascending order but it can be changed with **reverse** parameter.

## 6. Zip

Zip function takes iterables as argument and returns an iterator consists of tuples.

## 7. Any and All

I did not want to separate **any** and **all** functions because they are kind of complement of each other. Any returns True if any element of an iterable is True. All returns True if all elements of an iterable is True

One typical use case of any and all is checking the missing values in a pandas series or dataframe.

## 8. Range

Range is an immutable sequence. It takes 3 arguments which are **start**, **stop**, and **step** size.

## 9. Map

Map function applies a function to every element of an iterator and returns an iterable.

## 10. List

List is a mutable sequence used to store collections of items with same or different types.

## **MACHINE LEARNING IN DATA SCIENCE**

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

### **Why is machine learning important?**

Resurging interest in machine learning is due to the same factors that have made data mining and Bayesian analysis more popular than ever. Things like growing volumes and varieties of available data, computational processing that is cheaper and more powerful, and affordable data storage.

### **What are some popular machine learning methods?**

Two of the most widely adopted machine learning methods are supervised learning and unsupervised learning – but there are also other methods of machine learning. Here's an overview of the most popular types.

**Supervised learning** algorithms are trained using labeled examples, such as an input where the desired output is known. For example, a piece of equipment could have data points labeled either “F” (failed) or “R” (runs). The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly. Through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data. Supervised learning is commonly used in applications where historical data predicts likely future events. For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim.

**Unsupervised learning** is used against data that has no historical labels. The system is not told the "right answer." The algorithm must figure out what is being shown. The goal is to explore the data and find some structure within. Unsupervised learning works well on transactional data. For example, it can identify segments of customers with similar attributes who can then be treated similarly in marketing campaigns. Or it can find the main attributes that separate customer segments from each other. Popular techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering and singular value decomposition.

These algorithms are also used to segment text topics, recommend items and identify data outliers.

**Semisupervised learning** is used for the same applications as supervised learning. But it uses both labeled and unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data (because unlabeled data is less expensive and takes less effort to acquire). This type of learning can be used with methods such as classification, regression and prediction. Semisupervised learning is useful when the cost associated with labeling is too high to allow for a fully labeled training process. Early examples of this include identifying a person's face on a web cam.

**Reinforcement learning** is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards. This type of learning has three primary components: the agent (the learner or decision maker), the environment (everything the agent interacts with) and actions (what the agent can do). The objective is for the agent to choose actions that maximize the expected reward over a given amount of time. The agent will reach the goal much faster by following a good policy. So the goal in reinforcement learning is to learn the best policy.

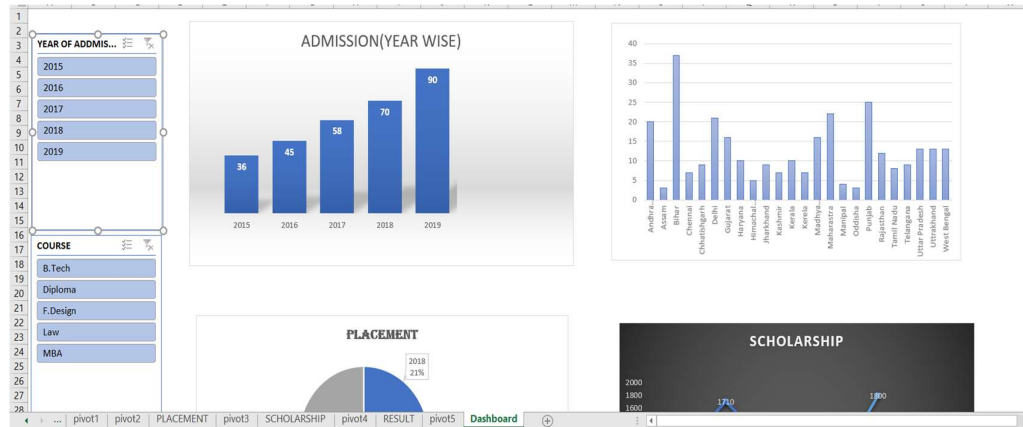


## **REASON WHY CHOOSE THIS COURSE**

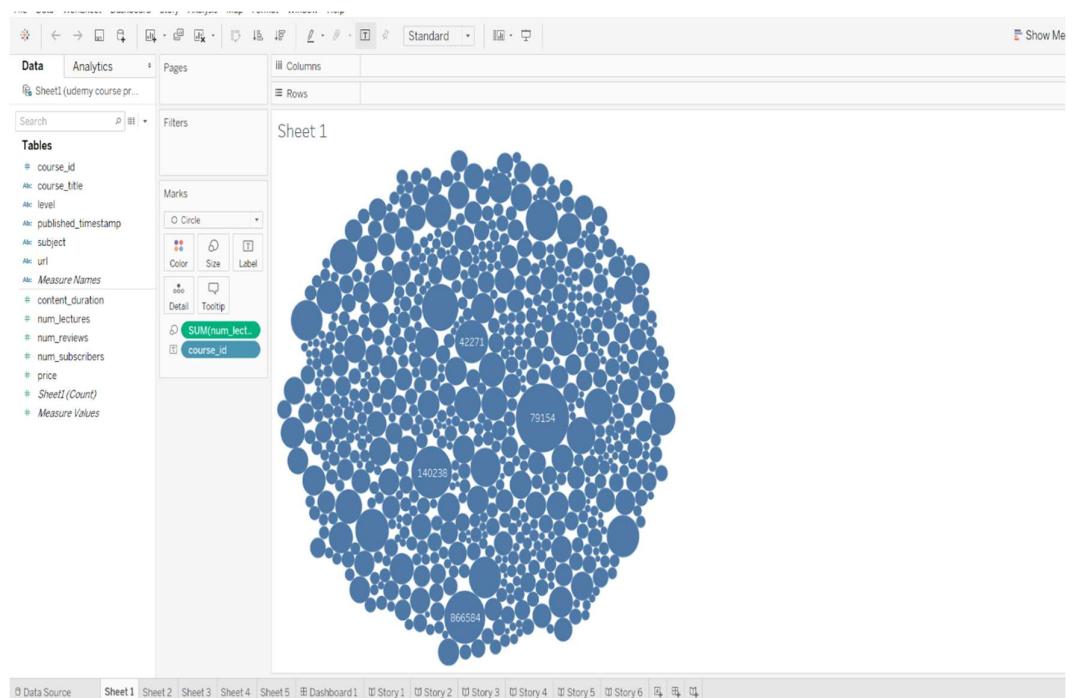
Data science makes me feel powerful! There are basically two reasons for my saying that. One is the fact that data science is a rapidly evolving area, so that a large part of a data scientist's work, at least as I envision it, is updating his or her methods, tools, and workflows. I really like that my work is evolving with the progress of data science, so that I never get bored. Also, I like to think that with the growing amount of data available to analyze, using and designing new tools to gain efficiency is a crucial part of the job. And of course, I find learning about and using such tools particularly .

# PROJECTS

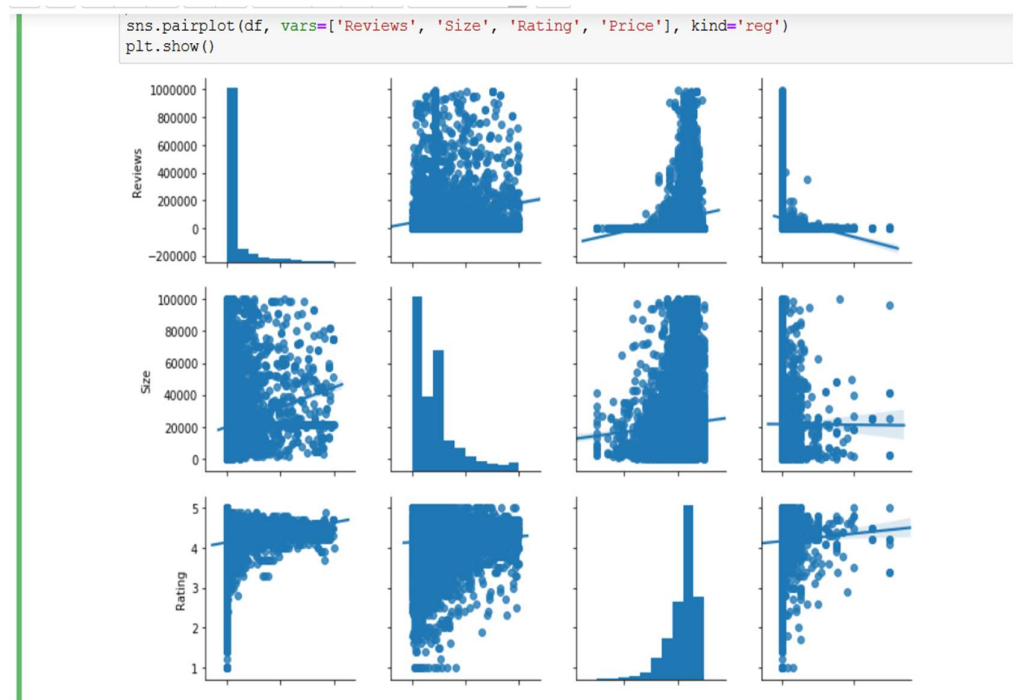
## 1. UNIVERSITY DASHBOARD USING EXCEL



## 2. UDEMY DASHBOARD USING TABLEAU



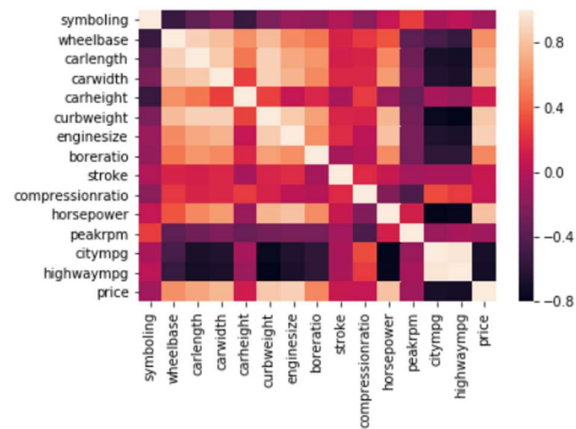
### 3. PYPLOT USING MATPLOTLIB



### 4. DATA CLEANING USING MACHINE LEARNING

```
In [12]: sb.heatmap(df.corr())
```

Out[12]: <matplotlib.axes.\_subplots.AxesSubplot at 0x22c2c31c400>



## **FUTURE SCOPE OF DATA SCIENCE**

Since Data Science is still an evolving field, there's much more to expect from it in the future. Let's look at some of the exciting Data Science trends that may soon become a reality in the upcoming future:

- While the IoT is already a reality that connects smart devices, in the future, we might be looking forward to being a part of an Intelligent Digital Mesh – a connected hub of apps, devices, and people working together in sync.
- Product marketing and customer service will be revolutionized by advanced chatbots, Virtual Reality (VR), and Augmented Reality (AR). We might be looking forward to a time when personalized customer experience will include live simulations, interactive demos, visualization of proposed solutions.
- Blockchain might just go mainstream – it will not only be limited to the finance sector, but blockchain will apply to healthcare, banking, insurance and other industries.
- Automated ML systems and Augmented Analytics together will transform Predictive Analytics and take it to the next level. Predictive Analytics will further help change the face of healthcare.

## **CAREER IN DATA SCIENCE**

1. Business Intelligence Developer
2. Data Architect
3. Applications Architect
4. Infrastructure Architect
5. Enterprise Architect
6. Data Analyst
7. Data Scientist
8. Data Engineer
9. Machine Learning Scientist
10. Machine Learning Engineer
11. Statistician

## **CONCLUSION**

After completing this project, I concluded that this project was the good opportunity to implement my information that I have learnt during my six week training program. This project is more informative and more helpful for understanding the concept of the Analyzing the data with different tools. This project is only a small and easy one but it is enough to implement my concept. I can further try much harder to make much more efficient and useful app that can benefit to other.

## **BIBLIOGRAPHY**

- simplilearn.com
- wikipedia
- ppt provided by Board Infinity