84.51° is a wholly owned subsidiary of The Kroger Company. Our data is sourced from consumer transactions and behavior across **24 banners, 38 states and 62 Million households.** We are the research, development and innovation arm of the Nation’s second largest retailer.

We are a community of analysts – we are researchers in search of the customer story. We are the voice of the customer. Customers share their story with us each and every day, and it is our job to tell their story. Our Analysis team is known for solving client and customer problems, employing the most appropriate statistical and mathematical analytic approaches.

We expect all analysts to be hands on with respects to the analysis of a retailer database, using the full suite of 84.51° technical tools – SAS, SQL, R, Excel, and other specific internal tools – to create timely, relevant, and actionable insights.

**Brands:** Hershey, Nestle, Pepsico, Pifzer, MayBelline, P&G, TimeWarnerCable, Unilever, Colgate-Palmolive, GSK, Dole, Dannon, Del Monte, Kellogg’s, LOREAL, King’s Hawaiian, Lindt and More…

**Introduction –**

I’m basically from an Engineering background

* 1. Worked as a software engineer for 1 year – Sanctum Technologies Pvt Ltd Bangalore (1 Years)

Then did my MBA – PGDM in **Marketing** and **operations research** management, It is equivalent to a Certification at US Standards.

Marketing Management, Marketing Strategy, Consumer Behavior, Business Statistics, Operational Research, Production Operation Management, Management Information Systems, Optimization Models

* 2. After my MBA, I worked for a company **Wings INet Technologies, as an Analyst, Worked on ETL projects**, Used SQL (10 Months)

Integrating the data from multiple source like Excel, Flat files(text), XML file, HTTP, Database (it may be sales database, Customer database, product database, CRM database .. another source)

Used Microsoft Visual Studio BI Modules - SSIS, SSAS, for Creating the data warehouse and Data marts

ETL Tool:

SQL Server Integration Services (SSIS), SQL Server Analytical Services (SSAS) and SQL Server Reporting Services (SSRS)

*Other ETL tools,*

* *OBIEE 11g– Oracle Business Intelligence Enterprise Edition*
* *IMB InfoSphere Datastage – Works on Unix platform*
* *Informatica*
* *Abintio*
* *SAP*

**Database:** Used DB2, Oracle -12c database, SQL server

*Other Databases: DB2 – IBM, Terradata , Oracle, MY SQL, SAP Hana, Greenplum NO SQL Databases: Hadoop, Pig , Hive, Oziee, Sqoop, MongoDB, Cassandra*

* 3. Worked for Analytical firm- Analytics Quotient from 2012 to 2015,

It’s a Marketing Consulting firm, worked on many consulting project for worked for brand like Red Lobster, Coca-Cola, and Essilor – Retail and Service Industry. involved data analysis, (3 year 8 Months)

**Use Database Querying – SQL**

**Statistical analysis – R, SAS, Python and SPSS**

**Used DB2, Oracle -12c database, SQL server**

**Reporting Systems – Tableau, Power BI**

**Used Excel, MS power point for reporting and presentation**

**Used Excel macros for automating tasks and creating excel tools.**

Marketing analytical projects involved, defining a problem, collecting and analyzing the data to gain the marketing insights and delivering the analysis insight with K&I team and helping in taking strategic decision.

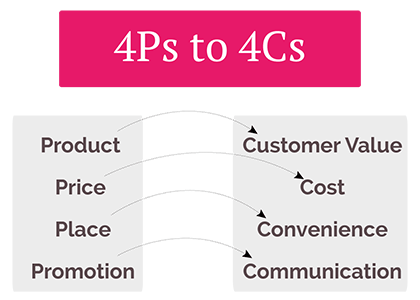
Involved with lot for Ad-Hoc projects, (2 days to 5 days), short term project (2 Week to 4 Week), Long term project (5 Week to 6 Months),

Work involved, creating timely reports - Monthly or Quarterly reporting projects, K&I Dashboards for Managers and helping in strategic decision making.

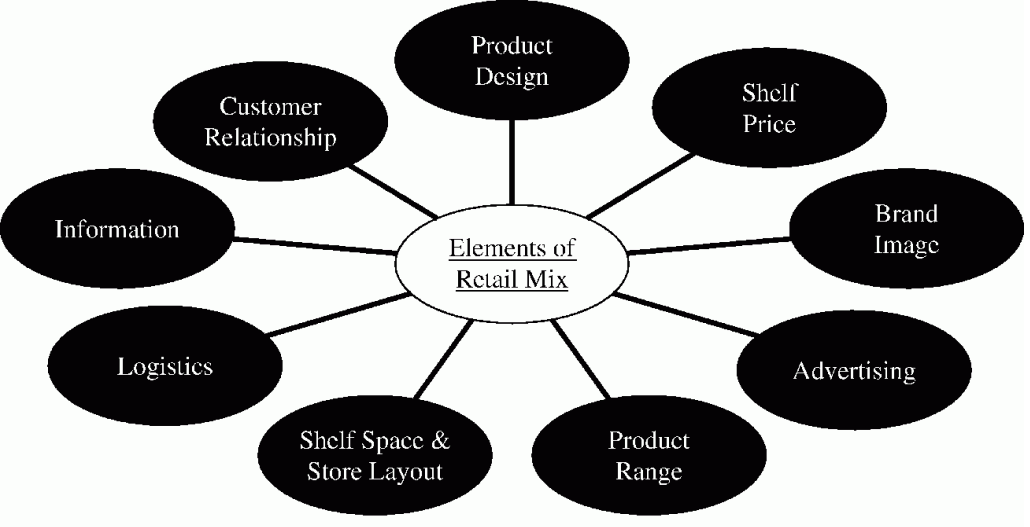
* Used a lot Analytical and Market Research techniques for answering various business questions, in the marketing context, analyzing consumer needs.
* **Analyzing the effectiveness of retail product offerings, price optimization and marketing campaign Vs Competitors for Red Lobster.**
* **Analyzing media channels like TV, Radio, Online Video, Print Ads, Organic Traffic Vs Paid traffic etc.**
* **Identifying the purchase patterns and consumer’s behaviors and making their life easier, by providing what they need.**
* **Used the time series models to find out the variation in the data due to trend, seasonality and cyclicality etc.**
* Identify, analyze and interpret trends or patterns in complex data sets.
* Automation of Models with help of development team (Front end development team) for the production.
* Leading the analytical project with 1 or 2 Junior Analysts, mentoring them to get the project done.
* Reviewing the analytics insight with Domain Consultants and Statisticians before delivery.
* Marketing Frameworks: Building the marketing frameworks, **Marketing framework** is a visual representation or a logical flow of your marketing plan.
* Doing SWOT, Marketing Mix - 4P Analysis, Retail Marketing Strategy
* SWOT Analysis



* Marketing mix is often referred to as the '4 Ps', i.e. product, price, place and promotion

* Retail Marketing Strategy:



This was my role at Analytics Quotient

* As a senior analyst, I have mentored 2 analysts, I was responsible for leading the analytics project from the front. We had domains and business consultant to evaluate the analytical outcomes before the delivery.

**Project 1:**

**Overall Satisfaction Analysis for Red Lobster Client- Used R for Analysis – Machine Learning**

How the **overall satisfaction of the consume could be improved**? What all the factors involved in driving the overall satisfaction of the consume. Conducting descriptive statistical analysis using regression models to analysis the impact of independent variable.

It was a descriptive statistical project for the brand Red lobster. The project involved identifying what factors could influences the overall satisfaction of consumers. The overall satisfaction was rated from 1 to 5 range, with 1 being the least satisfied and 5 being the most satisfied. The independent factors could be the quality of food, ambiance, location, waiting time etc. which could play a vital role in impacting the overall result.

We considered the SMG (Service Management Group) database survey results in analyzing the impact on overall customer satisfaction. We used Ordinal logistic regression methodology in explaining the importance of features. Basically, business equation would look like.

***Overall Satisfaction Rating = food quality + ambiance + time spent + amount spent + cleanliness + service + taste + hot pipping + parking + delivery + location + waiting time + etc.***

The analysis involved predicting the overall satisfaction - ordinal rating, by analyzing the impact of each independent factors in explaining the output.

Packages used: **MASS package**, **polr** function for Ordinal logistics regression model.

Analysis Outcome: I have conducted the same experiment with different location and analyzed what are the driving factor for that location. As expected some restaurant near down town, people are more concern towards waiting time, ambiance. In contrast, if we compare to the suburbs people were more concerned about food taste, delivery time, distance to reach etc.

**Project 2:**

**Sentiment Analysis for Red Lobster Client uisng Natural Language Processing (NLP)**

Sentiment analysis using public comments from SMG database for Red Lobster Client

(It was in a Proof of Concept Stage)

Tool Used: Python – BeautifulSoup and NLTK packages procession text; NumPy and Pandas for working with dataframes.

In this project, we took the corpus – public comments from SMG database, we also considered Yelp.com reviews about the restaurant. Extracted the data in HTML format and Created a Beautifulsoup object to extract the required fields.

* Used the Numpy and Panda packages to work with data frames
* Used CountVectorizer to convert the corpus into bag of words and created the vector space with all the features (unique words)
* Applying the transformation: Modifying minimum document frequency and Maximum document frequency
* Applying by grams and N-grams to capture – Cheddar bay biscuits, Red Lobster, Pina Colada
* Removed the stop words using NLTK
* Applied the stemming and lemmatization, to understand the frequency of the words.
* Used the AFINN sentiment dictionary to analyze the corpus,
* We had a plan to customize the sentiment dictionary for Red Lobster brand, it was in a POC stage.
* The corpus was of very small size, we just had a little over 2000 comments to analyze.

**Project 3:**

**DMI (Direct marketing Investment) allocation for Coca-Cola Client (For USA, China, Philippines, Turkey, European Markets)**

Building the marketing strategy to increase the sales. Used the forecasting techniques for sales projections and Used the constraint optimization algorithm to optimize the marketing budget allocation.

Used Constraint optimization algorithms to **optimize the marketing budget**.

Used Time Series Models for descriptive analysis – decomposed the of time series of revenue, to analyze the trend seasonality and cyclicality impact on overall sales. Used forecasting models like – Simple exponential smoothing, Moving Average, ARIMA models, in predicting the brand sales to allocate the Marketing budget.

I was responsible for Data collection and data preparation and normalizing the data, using R and supporting data consultants in the data modeling phase.

We also worked on Segmentation Analysis, Market Basket Analysis, Correspondence Analysis, Principle component analysis, I don’t have these test cases to discuss now on these concepts.

Ad Hoc requests like - Analyzing how the marketing or advertising campaign is performing,

What is the impact of TV, Print, Online media channels, how much are we spending on these media channels Vs competitors; Using regression models ( Revenue Output, all are Input)

Competitive pricing analysis, comparing the product price vs competitors

**Market Basket Analysis** – Analyzing the **Impulsive behaviors of consumes and designing the** food combo offer for red lobster brand.

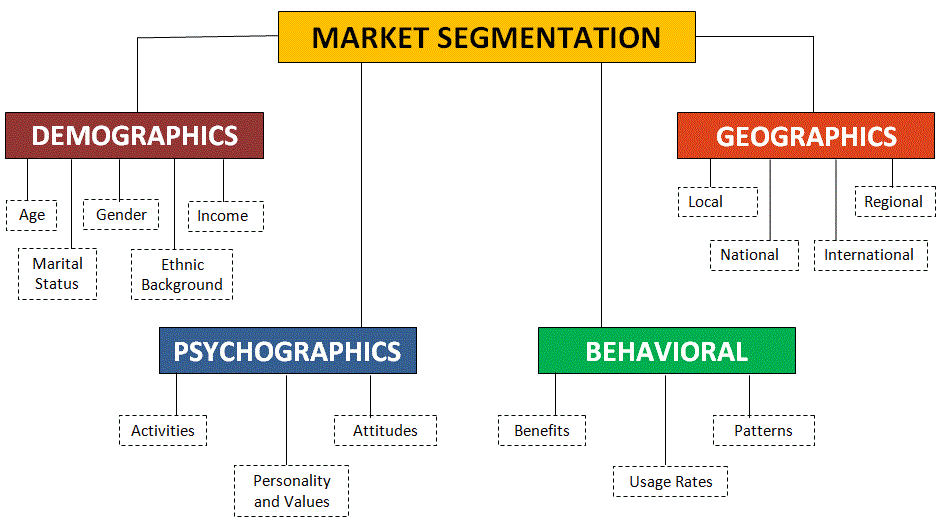
**Segmentation Profiling** – Dividing the target market into subset of consumer and analyzing their preference, choice of food, consumption patterns, Average spends.

* + - Demographic Segmentation – By Age group, gender, Income, Marital status, ethnicity.
    - Geographic Segmentation – regional, State, restaurant location, international etc.
    - Behavioral Segmentation – Benefits, Pattern, Usage Rate

Correspondence Analysis/ Conjoint Analysis

* + - Psychographic Segmentation - Activities, Attitudes, personal values etc.

Experience with variety of statistical methods – clustering, regression, predictive modelling, Market Research methods in answering business questions.



**I can also work on big data platform like HADOOP for data processing, can process unstructured data**

**Can work on SPARK engine for Machine learning,**

**from my Masters I have also learned predictive modeling, Time Series Modeling, Text Mining, Natural Language Processing, Liner Programming and Social Media Analytic and Applied data mining concepts – Linear Discriminant Analysis, Decision Trees, Random Forest, Bagging Boosting, KNN, K Mean, Hierarchical - Agglomerative and Divisive Clustering, Support Vector Machine, Stochastic Gradient Boosting algorithms.**

**Case Study Ideas:**

1. Using the constraint optimization models to optimize the, Revenue, Sales, profit, production units etc.

(LP Simplex problem, Constraint Optimization problems, **Transportation Problem – Logistics**)

1. Define custom **target audiences** based on **consumers purchase behaviors** in specific categories, brands and products.
2. Execute a campaign with display, banner, and **preroll media** served via purchase-based targeting on the open web Or Conduct In-depth interviews, with closed group of people or target audience (based on the historical purchase, brand spend, category spend, total store shopping behavior, online activity, demographic – Age, Income, Sex, Head Counts in the family etc, .…) and researching about the attitude of the consumer towards the product, characteristics of the shopper who purchased the item, In-depth behavioral analysis to understand the important attributes and behaviors on consume. Attributes - price, appearance, value for money
3. Modeling using Market basket analysis, statistical significance test, ANOVA, segmentation analysis, Principle component analysis, Factor Analysis, Correspondence Analysis, Regression Modeling, Logistic regressing, ARIMA, Predictive modeling.
4. Use control methodology to measure incremental uplift based on type of exposure Finding out the buyers uplift – did advertisement drive the shopper to the store and drive the sales uplift? Increase in the sales per household? Return on asset per audience? What were the profiles of the households exposed to the campaign?

**Market Research Methodology:**

1. Problem Definition - Define the problem, consider purpose of the study, the relevant background information, what information is needed, and how it will be used in decision making

**What is the objective?**

1. Development of an Approach to the Problem: **Formulating an objective** or theoretical framework, analytical models, **research questions**, **hypotheses**, and **identifying characteristics** or factors that can influence the research design
2. Research Design Formulation: A research design is a framework or **blueprint for conducting the marketing research** project

Procedures necessary for obtaining the required information

Design a study that will test the hypotheses of interest

Formulating the research design involves the following steps: [1]

Secondary data analysis

Qualitative research

Methods of collecting quantitative data (survey, observation, and experimentation)

Definition of the information needed

Measurement and scaling procedures

Questionnaire design

Sampling process and sample size

Plan of data analysis

1. **Field Work or Data Collection:** personal interviewing
2. **Data Preparation and Analysis -** Data preparation includes the editing, coding, transcription, and verification of data. **Multivariate techniques** are used for analyzing data when there are two or more measurements on each element and the variables are analyzed simultaneously
3. **Report Preparation and Presentation**
   * Report should address the specific research questions identified
   * Describes the approach,
   * The research design,
   * Data collection
   * Data analysis procedures adopted
   * Presents the results and the major findings.

**Case Study 1: TV advertising campaign was driving sales for the brand and each sub-brand within the portfolio. The Brand also wanted to understand the effect of different frequency levels to optimize future media planning.**

**Problem definition:**

1. What is the sales impact of the new TV creative strategy for the brand in total and for each subbrand?
2. What is the effect of different frequency levels on sales?

**Approach to the problem: Research questions, identifying characteristics**

**Research design:**

**Data Collection:**

**Analysis Methods**

**Findings and Presenting:**

**Case 2: Client launched a super-premium laundry detergent with high in-market expectations. Since the launch of product, volume has been lower than expected**

**• For those that have not repeated, what are the barriers to repeating?**

**• For those that do repeat, why do they do so?**

**Marketing Research Analytics Methodologies:**

Market Research Steps Involves:

* Identification: involves defining the marketing research problem (or opportunity) and determining the information that is needed to address it.
* Collection: data must be obtained from relevant sources.
* Analysis: Data are analyzed, interpreted and results are drawn

Dissemination of information: the findings, implications, and recommendations are provided in a format that makes this information actionable and directly useful as an input into decision making.

**Marketing Research Methodologies:**

**Clustering or Segmentation Analysis**

Segmentation is the process of dividing potential markets or consumers into specific groups. Dividing target markets into subsets of consumers based on needs and wants.

There are four main types of segmentation used in market research analysis: a priori, usage, attitudinal and need

1. **a** **priori** (most commonly used) : A subset of frequent item set must be a frequent item set, It’s an iterative approach to find frequent item set; Use the frequent item set to generate association rules.
2. **Segmentation** clustering (also used frequently): Segmentation is completed either by decile or pereto analysis.
3. **Attitudinal (Cluster analysis):** Cluster analysis to create customer psychological profiles. Preference data (scaling) is better suited toward this type of analysis
4. **Needs Based Segmentation**: Needs based segmentation is the concept that the **market can be divided based on customer need**.  This type of analysis is used to develop **products that sell rather than trying to sell** products a business developed. Needs based segmentation **uses conjoint analysis** to separate the groups according to functional performance.

**Market Basket Analysis (Association Analysis):**

Market Basket Analysis: Market Basket Analysis is a modelling technique based on the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items.

The set of items a customer buys is referred to as an itemset, and market basket analysis seeks to find relationships between purchases.

In retailing, most purchases are bought on **impulse**. Market basket analysis gives clues as to what a customer might have bought, if he buys product A and product B ..

**Conjoint Analysis: Conjoint features**

Conjoin Analysis is a marketing insight technique for predicting how products you create and redesign should perform when take into market. It’s a statistical methodology help to **unpack the preference of consumer** with regard to different marketing offers.

Two important dimension: Inferred **utility function** of each attribute 2. **Relative importance** of each attribute to the consumer.

Putting the right features and charging the right price based on consumer need.

What is important in the product and how much they are willing to pay? Conjoint features/ right features, right price to make the product success and beat the market.

**Correspondence Analysis:**

Correspondence analysis is a statistical technique that provides a **graphical representation of cross tabulations** (which are also known as cross tabs, or contingency tables). Simple two-way and multi-way tables containing some measure of correspondence between the rows and columns.

**Multiple Correspondence Analysis (MCA)**

Multiple correspondence analysis (MCA) may be considered to be an extension of simple correspondence analysis to more than two variables. For an introductory overview of simple correspondence analysis, refer to the introductory section . Multiple correspondence analysis is a simple correspondence analysis carried out on an indicator (or design) matrix with cases as rows and categories of variables as columns.

**What is principal components analysis?**

Principal components analysis is a procedure for identifying a smaller number of uncorrelated variables, called "principal components", from a large set of data. The goal of principal components analysis is to explain the maximum amount of variance with the fewest number of principal components. Principal components analysis is commonly used in the social sciences, market research, and other industries that use large data sets.

You can use principal components analysis to reduce the number of variables and avoid multicollinearity, or when you have too many predictors

NOTE:

1. PCA is used when you want to predict the Y variable

2. PCA cannot be used when you want to find out the impact of X variable on Y variables, it doesn’t help in finding out the relationship or Marginal impact between X and Y

**Number of principal components**

**Kaiser method:** Retain components with eigenvalues greater than 1.

**Scree test:** The ideal pattern in a scree plot is a **steep curve**, followed by a bend and then a flat or horizontal line.

**Percentage of variation explained:** Retain components that cumulatively explain a certain percentage of variation. The acceptable level of explained variance depends on how you use Principal Components. For descriptive purposes, you might only need 80% of the variance explained.

Example

A consumer products company wants to analyze customer responses to several characteristics of a new shampoo: color, smell, texture, cleanliness, shine, volume, amount needed to lather, and price. They perform a principal components analysis to determine whether they can form a smaller number of uncorrelated variables that are easier to interpret and analyze. The results identify the following patterns:

* Color, smell, and texture form a "Shampoo quality" component.
* Cleanliness, shine, and volume form an "Effect on hair" component.
* Amount needed to lather and price form a "Value" component.

**What is factor analysis?**

* Help to finds out the relationship between different factors
* Similar to PCA , except FA is used what the causal hypothesis to be tested !!
* Also helps in reduction of variable, you can come up with smaller set of factors

Factor analysis is a method for explaining the structure of data by explaining the correlations between variables. Factor analysis summarizes data into a few dimensions by condensing a large number of variables into a smaller set of latent variables or factors.

**SUMMARY:**

**Segmentation Analysis** – Dividing the target market into subset of consumer and analyzing their preference, choice of food,

**Correspondence Analysis** - Graphical representation of cross tabulation

**Principal components analysis** - Smaller number of uncorrelated variables, called "principal components", from a large set of data. Goal of principal components analysis is to explain the maximum amount of variance with the fewest number of principal components

Example: Helps in designing the consume survey

**INTERVIEW EXPERIENCE:**

How you analyze the data?

I was asked to analyze grocery store data.

Experience:

Placed in a room then interviewed in 3 phases. 2 one on one interviews then one case study presentation with 30 mins to prepare. Case Study was much more technical than expected and they asked a lot for **choosing** and **implementing theoretical data sets**.

Data Trends using Kroger Data, making **inventory decisions**

**Things to Consider When Making a Decision in Inventory Management**

**Warehousing** - Reduce the amount of floor space used, without jeopardizing supply for production. Smaller supply shipments more frequently may be one solution. Finished goods may also be shipped frequently or sent to off-site storage to maximize room for raw materials.

**Cost** - The value of raw materials represents stalled cash flow, don't actively contribute to the bottom line

"just in time" delivery, improve profit margin on finished goods, but only if additional warehousing and finance charges are avoided.

**Delivery Time** - Time between order and delivery of raw materials is an important factor. When the time between order and delivery is weeks, an inventory manager must factor this in to maintain sufficient supply, while balancing warehousing space and inventory costs.

**Turnaround** - When considering inventory. Where finished goods are concerned, turnaround refers to how long stock sits before sale. An inventory manager wants this time to be as short as possible. Turnaround refers to how long current stock supplies production. Where **minimum** and **maximum inventory** levels are set, the minimum level considers supply turnaround and delivery time, while maximum values address warehousing and cost factors

**TIME SERIES ANALYSIS:**

Suppose a model is being developed to predict the occupancy of a hotel on a particular day based on historical data. Describe your strategy to evaluate various methods such moving average, exponentially weighted moving average, ARIMA, etc.

Before we evaluate and finalize on which models to uses, time series modeling has to be done based on the type of time series data. If the hotel occupancies are increasing exponentially, I would use logarithmic transformation to convert multiplicative patters to additive pattern. Which helps us to linearizes the exponential growth also **stabilized the variance** in the data. Hence, we should do some mathematical transformation if necessary, the mathematical transformation could be Box-Cox lamda transformation, Calendar adjustment, population adjustments or Inflation adjustments etc.

**Some of the assumptions for forecasting:**

* The variance (the error term) should be normally distributed
* The assumptions of homoscedasticity or homogeneity of variance
* There should not be autocorrelation (remainder is a white noise)

**Before using any of the time series models, the time series must be Stationary with no variance.**

* **Check for Auto Correlation with Box-Ljung Test -**
* **Check for Stationarity with ADF test**

**Which models to use for forecasting Occupancy rate?**

**Moving average:** This method is taking simple average of recent data. When data are in short supply with highly irregular variance, we would prefer moving average method. This model is robust against messy data and with outliers.

**Simple exponential smoothing:** When data are non-seasonal and display time varying mean without a consistent trend, i.e. where there is no Trend, seasonality, Cyclicality we use simple exponential smooting.

**Linear exponential smoothing (Holt’s method)**: When data are non-seasonal and display time varying local trends (usually applicable to data that are “smoother” in appearances –i.e., less noisy than what would be well fitted by simple exponential smoothing)

1. Alpha - is the smoothing parameter for the level; Higher the alpha more weight is given to the more recent observations.

2. Beta - is the smoothing parameter for the trend, Higher the beta it captures more variation in the Time Series. Smaller the beta value the variations in the trend will be smoother.

3. h = No of forecast

**Multiple regression:** When data are correlated with other explanatory variable, In our case there could be some spurious correlation between occupancy rate, with unemployment rate, GDP, Air traffic, disposable income etc. To find out these spurious relations we would use multiple regression.

**ARIMA**

* ARIMA models are designed to squeeze all autocorrelation out of the original time series;
* ARIMA models forecasts for the stationarized dependent variable are a linear function of lags of the dependent variable and/or lags of the errors.
* When to Use: When data are relatively plentiful (4 seasons or more) and can be satisfactorily **stationarized by differencing** and other mathematical transformations; when it is not necessary to explicitly separate out the seasonal component (if any) in the form of seasonal indices

**Random walk:** When there is no seasonality, cyclicality or trend in the time series data. When we cannot identify, the data generating process or when we cannot use the historical data to predict the future, it’s a random or unpredictable data. Usually we see these types of series in stock price change.

**Based on all these factors we could use any of the suitable model in predicting the occupancy rate.**

**Brainteaser Questions:**

https://www.themuse.com/advice/7-insane-brain-teasers-you-could-actually-encounter-in-an-interview

**Puzzle:**

1. There's three bags and all of them are mislabeled. One has all blue marbles, one has all yellow, and one has both blue and yellow marbles. How many marbles would you have to draw before you know which bag is which?

**lets assume:**

box 1 is labelled Oranges (O)

box 2 is labelled Apples (A)

box 3 is labelled Apples and Oranges (A+O)

and that ALL THREE BOXES ARE LABELLED INCORRECTLY"

Solution: The trick is to actually pick a fruit from the A+O labeled box

Pick a fruit from box 3:

1) if you pick an Orange:

- box 3's real label can only be O or A

- box 3's current label is A+O

- since ALL LABELS ARE INCORRECT then box 3's real label can not be A+O

- box 3's new label should then be O by elimination

- since ALL LABELS ARE INCORRECT

- box 1's label is changed to A

- box 2's label is changed to A+O

- SOLVED

2) if you pick an Apple:

- box 3's real label can only be O or A

- box 3's current label is A+O

- since ALL LABELS ARE INCORRECT then box 3's real label can not be A+O

- box 3's new label should then be A by elimination (not O)

- since ALL LABELS ARE INCORRECT

- box 1's label is changed to A+O

- box 2's label is changed to O

- SOLVED

1. A thief was presented in the court of a king. For a judgment, the king asked him to make any statement. If the statement was to be true, he would be burnt alive and if the statement was to be false, he would be thrown in the sea. But after hearing the statement, the king had to let him go.

Solution:

I will be thrown into the sea. Now if the king threw him in the sea, the statement will come as true and thus he will have to be burnt alive. But if the king burnt him, the statement will come out as false. Thus the king had no choice but to leave him.

1. There is a town named Springville. In that town you can find extremely hilarious facts. None of the resident has 456,789 hairs. The number of residents are more than the number of hair in any one of the.

The answer is 456,789.

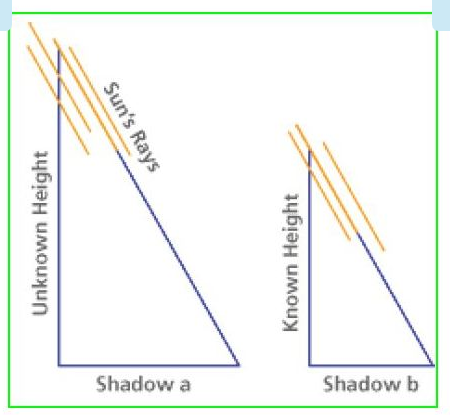
Let us begin with two residents. The number of hairs on their head can be zero and one. If we extrapolate the fact, we will come to know that the number of hairs with 'n' number of residents will always range from zero to (n-1).

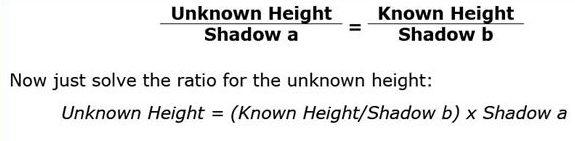
If we go above 456,789 to 456,790 with none of the residents having 456,789 hairs, the number of hairs on them will be different and one of them must have more than 456,789 which will clearly violate the fact that the number of residents are more than hairs on an individual.

1. Tia and Mia are twin sisters.One of them is a liar whereas the other one is a truth teller.   
   I called their home number and one of them picked up.   
   I asked. 'Does Tia lie?' The person replied with a yes.   
   Whom did I talk with? Tia or Mia ?

**Solution**I spoke with Mia.  
  
Let us assume that Tia is a liar and Mia is a truth teller.  
If I asked the question from Mia, the answer will be yes. If I asked the question from Tia, the answer will be no. Thus I must have asked from Mia.  
  
Let us assume that Tia is a truth teller and Mia is a liar.  
If I asked the question from Mia, the answer will be yes. If I asked from Tia, the answer will be no. Thus in this case as well, I must have asked from Mia.

1. How to measure the height of the building?





1. How many veterinarians are there in the US?

There was a packet containing a case study that I had a set amount of time to review before the interview process began. There was a screenshot of a data set contained in this packet, and questions touched on topics such as how would you join these data sets, add this feature to the data set and how would you determine…

Data Trends using Kroger data, making inventory decisions

**Can you tell us about a time you have worked with larger dataset?**

During my masters – Used airline traffic data

The data consists of **flight arrival and departure details for all commercial flights** within the USA, from **October 1987 to April 2008**. This is a large dataset: there are **nearly 120 million records** in total, and takes up **1.6 gigabytes of space compressed** and 12 gigabytes when uncompressed.

**The challenge**

The aim of the data expo is to provide a graphical summary of important features of the data set.

When is the best time of day/day of week/time of year to fly to minimise delays?

Do older planes suffer more delays?

How does the number of people flying between different locations change over time?

How well does weather predict plane delays?

Can you detect cascading failures as delays in one airport create delays in others? Are there critical links in the system?

**Used HDinsight Cluster from Azure to process and analyze the data. Used Hive, PIG, Scoop, Oziee job flows for processing and analyzing the data.**

Brainteasers (especially estimating quantities or durations of time)

How do you handle feedback?

Name an instance when you received negative feedback.

# Facts:

There was a meeting between Offshore AQ team (from India) and Onshore AQ team (from USA) with one of our clients from Florida (Red Lobster). We were supposed to present the final phase of a project for which our manager Kuldeep has travelled to Atlanta. Unfortunately, he did not have the **VPN (Virtual Private Network)** installed in his system, due to which he could not access our internal servers (AQ Servers) from outside the office. We thought of sharing our desktop from India by connecting to our internal servers using Lync meeting. Unfortunately, that plan did not go well because of the network issue we had on that day. It was a late night call at 11:00 PM in INDIA and at that time the IT guy who was supposed to troubleshoot these issues was already back home. We were helpless. We had to wrap-up our important meeting unfinished.

# Options:

We had an internal meeting in our team to come up with a solution. Hopefully, below options can help us overcome these issues in the future.

1. The **VPN installations** should be prioritized, as we know that the IT people are getting overloaded with bug fixes and network issues. The VPN installation tickets raised should be given more priority. Without VPN installations we cannot connect to our internal servers or handle any task due to which there will be wastage of executive hours.
2. We can have some dedicated audio and video conference lines to handle this kind of priority client calls. Even though we have quite a number of conference lines, due to huge amount of projects we are facing a lot of overlap issues in blocking them for project calls. Probably, having some dedicated conference lines for each team would be a suitable solution to address this issue.
3. IT Support team should be available 24X7 so that they can help us resolve these kind of issues.
4. Checklist should be created by our IT department for business travelers. The list should include the type of softwares (team view, Skype etc) to be installed, IDs, VPN installations, network cables etc.

I would highly recommend option 2. We thought this could be really effective to avoid these kinds of embarrassing situations. Also setting some priorities in the IT Support team, as suggested in option 1 and option 3, can help us avoiding these minor IT failures which cause big issues. Do let us know for any concerns. Please make some time to take this issue forward and let us know if you have any additional thoughts. We are hoping to get a solution by this Friday.

Explain group work experiences.

If you were to open a store with a friend, how would you divide the responsibilities?

Decision making method when there is only limited data available, explaining research/work I have done and methodologies used to analyze?

Metric that would help you determine whether company's new product is doing well.

* Monthly unique visitors/ buyers (loyalty card)
* Number of support tickets created – Are they facing any issue?
* Active user of the product
* Churn Rate – How many people are leaving? / How many loyal customer are not coming back?
* Average Session time / How frequently the item being purchased?
* Monthly revenue/ Average revenue
* User sentiment / Analyzing comments / Ratings
* Customer acquisition cost (CAC) – How much are you spending to get a new customer?
* Organic traffic vs. paid traffic
* Conversion rate to customer – Free customer to paid customer

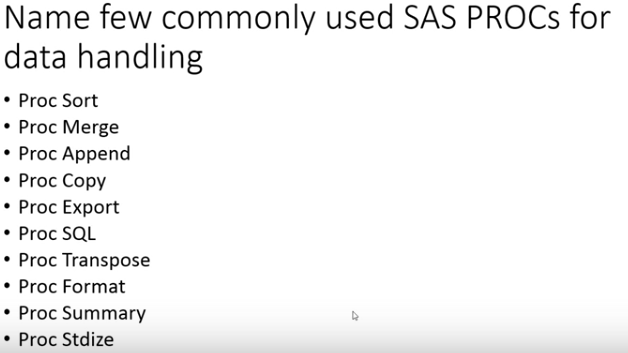
8) Mention what is the difference between data mining and data profiling?

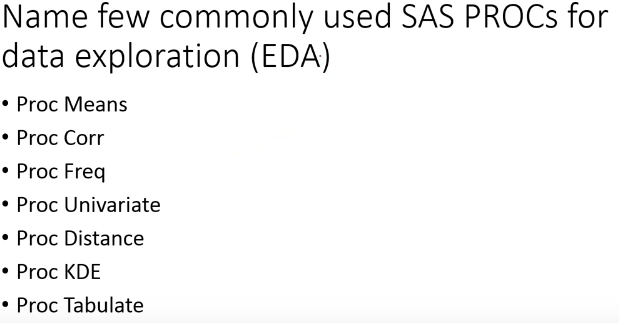
The difference between data mining and data profiling is that

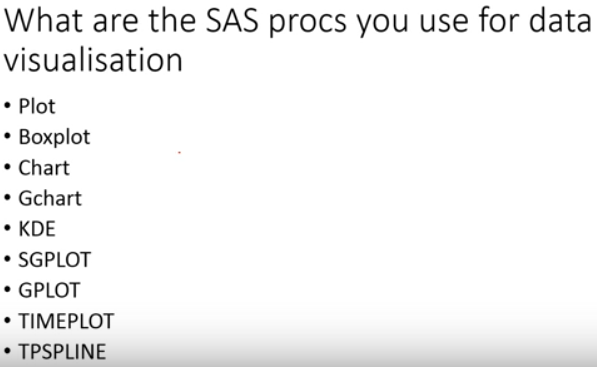
Data profiling: It targets on the instance analysis of individual attributes. It gives information on various attributes like value range, discrete value and their frequency, occurrence of null values, data type, length, etc.

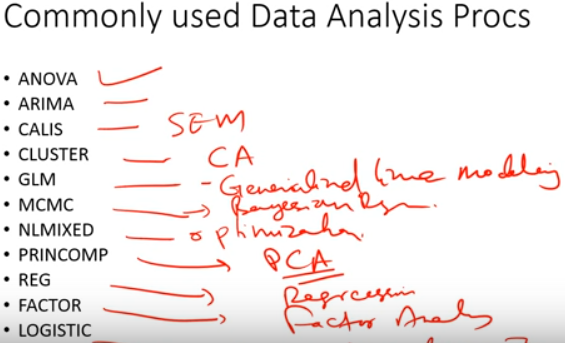
Data mining: It focuses on cluster analysis, detection of unusual records, dependencies, sequence discovery, relation holding between several attributes, etc.

**SAS Programming:**









1. **SQL - SELECT**

SELECT DISTINCT, TOP 10, TOP 10 PERCENT,

FROM

JOIN - INNER JOIN, LEFT JOIN, RIGHT JOIN, CROSS JOIN

ON

WHERE - Name LIKE 'M%' or Name LIKE ’[^R]\_\_F[0-9]%' , IN (5, ‘black’, 7), AND,OR, ISNULL, IS NOT NULL, BETWEEN =<>

GROUP BY

HAVING

ORDER BY Name ASC DESC

OFFSET 0 ROWS FETCH NEXT 10 ROWS ONLY;

**EXAMPLE:**

SELECT P.name as ProductName, SUM(OD.LineTotal) as Revenue,

SUM(OD.LineTotal)/COUNT(OD.LineTotal) as UnitPrice

FROM SalesLT.Product as P

JOIN SalesLT.SalesOrderDetail as OD

ON P.ProductID = OD.ProductID

WHERE OD.LineTotal > 1000

GROUP BY P.Name

HAVING name BETWEEN ‘F’ AND ‘L’

ORDER BY Revenue DESC;

**CONVERTING DATA TYPES:**

CAST - CAST(ProductID AS varchar(3))

CONVERT - CONVERT(varchar(4), ProductID)

TRY\_CAST (Size AS Integer) AS NumericSize - (note incompatible sizes are returned as NULL)

**HANDLING NULL**

ISNULL(TRY\_CAST(Size AS Integer),0) OR ISNULL(Color, '')

IS NOT NULL - WHERE SellEndDate IS NOT NULL;

**STATEMENT WITH PREFERENCE**

COALESCE(DiscontinuedDate, SellEndDate, SellStartDate) AS FirstNonNullDate

**CASE STATEMENTS**

**EXAMPLE1:**

SELECT Name,

CASE

WHEN SellEndDate IS NULL THEN 'On Sale'

ELSE 'Discontinued'

END AS SalesStatus

FROM SalesLT.Product;

**EXAMPLE2:**

SELECT Name,

CASE Size

WHEN 'S' THEN 'Small'

WHEN 'M' THEN 'Medium'

WHEN 'L' THEN 'Large'

WHEN 'XL' THEN 'Extra-Large'

ELSE ISNULL(Size, 'n/a')

END AS ProductSize

FROM SalesLT.Product;

**WHERE STATEMENTS:**

WHERE ProductNumber LIKE 'BK-[^R]%-[0-9][0-9]';

WHERE Color IN ('Black','Red','White') and Size IN ('S','M');

WHERE (Color LIKE 'Black' or Color LIKE 'Red' OR Color LIKE 'White') and (Size LIKE 'S' or Size LIKE 'M')

WHERE ProductCategoryID IN (5, 6, 7) AND SellEndDate IS NULL; -

**# JOIN STATEMENTS: INNER JOIN, LEFT JOIN, RIGHT JOIN, FULL JOIN, CROSS JOIN – Cartesian product, SELF JOIN,**

**JOIN – INNER JOIN**

SELECT SalesLT.Product.Name As ProductName, SalesLT.ProductCategory.Name AS Category

FROM SalesLT.Product

INNER JOIN SalesLT.ProductCategory

ON SalesLT.Product.ProductCategoryID = SalesLT.ProductCategory.ProductCategoryID;

**LEFT JOIN:**

SELECT C.CompanyName, C.FirstName + C.LastName as Contact, OH.SalesOrderID, OH.TotalDue

FROM SalesLT.Customer AS C

Left JOIN SalesLT.SalesOrderHeader as OH

ON C.CustomerID = OH.CustomerID

ORDER BY OH.TotalDue DESC;

**FULL JOIN**

SELECT C.CustomerID, P.ProductID

FROM SalesLT.Customer as C

FULL JOIN SalesLT.SalesOrderHeader as OH

ON C.CustomerID = OH.CustomerID

FULL JOIN SalesLT.SalesOrderDetail as OD

ON OH.SalesOrderID = OD.SalesOrderID

FULL JOIN SalesLT.Product as P

ON OD.ProductID = P.ProductID

WHERE OH.SalesOrderID IS NULL

**CROSS JOIN** --Call each customer once per product

SELECT p.Name, c.FirstName, c.LastName, c.Phone

FROM SalesLT.Product as p

CROSS JOIN SalesLT.Customer as c

ORDER BY FirstName;

**SELF JOINS:**

SELECT e.EmployeeName, m.EmployeeName AS ManagerName

FROM SalesLT.Employee AS e

LEFT JOIN SalesLT.Employee AS m

ON e.ManagerID = m.EmployeeID

ORDER BY e.ManagerID;

**# UNION, UNION ALL, INTESECT, EXCEPT**

UNION: Use UNION to combine the row sets returned by multiple queries.

UNION ALL: Includes common item

INTERSECT: Only common items

EXCEPT: Only not matching items

**# USING FUNCTION AND AGGREGATING DATA – SCALAR FUNCTIONS, WINDOW FUNCTION, LOGICAL FUNCITONS, AGGREGATE FUNCTIONS**

**SCALAR FUNCTIONS**

SELECT YEAR(SellStartDate) AS SellStartYear, DATENAME(mm,SellStartDate) AS SellStartMonth, DAY(SellStartDate) AS SellStartDay, DATENAME(dw, SellStartDate) AS SellStartWeekday, DATEDIFF(yy,SellStartDate, GETDATE()) AS YearsSold, UPPER(Name) AS ProductName, CONCAT(FirstName + ' ', LastName) AS FullName, LEFT(ProductNumber, 2) AS ProductType, ROUND(P.Weight,0) as ApproxWeight

**LOGICAL FUNCTIONS:** ISNUMERIC(Size) = 1

**IF FUNCTION:**

IIF(ProductCategoryID IN (5,6,7), 'Bike', 'Other') ProductType , IIF(ISNUMERIC(Size) = 1, 'Numeric', 'Non-Numeric') SizeType

**CHOOSE** (cat.ParentProductCategoryID, 'Bikes','Components','Clothing','Accessories') AS ProductType

**WINDOW FUNCTIONS & RANK FUNCTION**

**EXAMPLE1:**

SELECT TOP(100) ProductID, Name, ListPrice,

RANK() OVER(ORDER BY ListPrice DESC) AS RankByPrice

FROM SalesLT.Product AS p

ORDER BY RankByPrice;

**EXAMPLE2:**

SELECT c.Name AS Category, p.Name AS Product, ListPrice,

RANK() OVER(PARTITION BY c.Name ORDER BY ListPrice DESC) AS RankByPrice

FROM SalesLT.Product AS p

JOIN SalesLT.ProductCategory AS c

ON p.ProductCategoryID = c.ProductcategoryID

ORDER BY Category, RankByPrice;

**AGGREGATE FUNCTIONS**

SELECT COUNT(\*) AS Products, COUNT(DISTINCT ProductCategoryID) AS Categories,

AVG(ListPrice) AS AveragePrice

FROM SalesLT.Product;

**BY PRODUCT CATEGORY**

SELECT C.Name as ProductCategory, Count(P.ProductID) as TotalTypes,

AVG(P.ListPrice) as Average

FROM SalesLT.Product AS p

JOIN SalesLT.ProductCategory AS c

ON P.ProductCategoryID = C.ProductCategoryID

GROUP BY C.Name

ORDER BY C.Name

**HAVING – HAVING clause to filter based on aggregate**

**Performance of Salesperson**

SELECT Salesperson, COUNT(CustomerID) Customers

FROM SalesLT.Customer

GROUP BY Salesperson

HAVING COUNT(CustomerID) > 100

ORDER BY Salesperson;

**# SUBQUERY– SCALAR SUBQUERY, MULTIVALUED SUBQUERY, CORRELATED SUBQUERY, CROSS APPLY**

**SCALAR SUBQUERY**

List price is higher than the highest unit price of items that have sold

SELECT \* FROM SalesLT.Product

Where ListPrice >

(SELECT Max(UnitPrice) FROM SalesLT.SalesOrderDetail)

**MULTIVALUED SUBQUERY:**

List products that have an order quantity greater than 20

SELECT NAME, ProductID

FROM SalesLT.Product

WHERE ProductID IN

(SELECT ProductID FROM SalesLT.SalesOrderDetail

WHERE OrderQty > 20)

**USING JOIN:**

SELECT Name

FROM SalesLT.Product P

JOIN SalesLT.SalesOrderDetail SOD

ON P.ProductID=SOD.ProductID

WHERE OrderQty>20

**CORRELATED SUBQUERY:**

--For each customer list all sales on the last day that they made a sale

SELECT CustomerID, SalesOrderID, OrderDate

FROM SalesLT.SalesOrderHeader AS SO1

WHERE orderdate =

(SELECT MAX(orderdate)

FROM SalesLT.SalesOrderHeader AS SO2

WHERE SO2.CustomerID = SO1.CustomerID)

ORDER BY CustomerID

**CROSS APPLY:**

The APPLY operator enables you to execute a **table-valued function** for each row in a row set

returned by a SELECT statement. Conceptually, this approach is **similar to a correlated subquery**.

**CROSS APPLY** returns matching rows, like an inner join

**OUTER APPLY** returns all rows in the original SELECT query results with NULL values

for rows where no match was found

CREATE FUNCTION SalesLT.udfMaxUnitPrice (@SalesOrderID int)

RETURNS TABLE

AS

RETURN

SELECT SalesOrderID, Max(UnitPrice) as MaxUnitPrice FROM

SalesLT.SalesOrderDetail

WHERE SalesOrderID=@SalesOrderID

GROUP BY SalesOrderID;

--Display the sales order details for items that are equal to the maximum unit price for that sales order

SELECT \* FROM SalesLT.SalesOrderDetail AS SOH

CROSS APPLY SalesLT.udfMaxUnitPrice(SOH.SalesOrderID) AS MUP

WHERE SOH.UnitPrice=MUP.MaxUnitPrice

ORDER BY SOH.SalesOrderID;

**# VIEW, TEMPORARY TABLES, AND VARIABLES**

**-- Create a view**

CREATE VIEW SalesLT.vCustomerAddress

AS

SELECT C.CustomerID, FirstName, LastName, AddressLine1, City, StateProvince

FROM

SalesLT.Customer C

JOIN SalesLT.CustomerAddress CA

ON C.CustomerID=CA.CustomerID

JOIN SalesLT.Address A

ON CA.AddressID=A.AddressID

**-- Query the view**

SELECT CustomerID, City

FROM SalesLT.vCustomerAddress

CREATE VIEW SalesLT.vProductOrders

AS

SELECT P.ProductID, Name, OD.OrderQty, OD.UnitPrice

FROM SalesLT.Product as P

JOIN SalesLT.SalesOrderDetail as OD

ON P.ProductID = OD.ProductID

**-- Join the view to a table**

SELECT c.StateProvince, c.City, ISNULL(SUM(s.TotalDue), 0.00) AS Revenue

FROM SalesLT.vCustomerAddress AS c

LEFT JOIN SalesLT.SalesOrderHeader AS s

ON s.CustomerID = c.CustomerID

GROUP BY c.StateProvince, c.City

ORDER BY c.StateProvince, Revenue DESC;

**02-Temp Tables and Variables**

**-- Temporary table**

CREATE TABLE #Colors

(Color varchar(15));

INSERT INTO #Colors

SELECT DISTINCT Color FROM SalesLT.Product;

**-- Table variable**

DECLARE @Colors AS TABLE (Color varchar(15));

INSERT INTO @Colors

SELECT DISTINCT Color FROM SalesLT.Product;

SELECT \* FROM @Colors;

**03-TVFs - Table valued function**

CREATE FUNCTION SalesLT.udfCustomersByCity (@City AS VARCHAR(20))

RETURNS TABLE

AS

RETURN

(SELECT C.CustomerID, FirstName, LastName, AddressLine1, City, StateProvince

FROM SalesLT.Customer C

JOIN SalesLT.CustomerAddress CA

ON C.CustomerID=CA.CustomerID

JOIN SalesLT.Address A

ON CA.AddressID=A.AddressID

WHERE City=@City);

SELECT \* FROM SalesLT.udfCustomersByCity('Bellevue')

**04-Derived Tables**

SELECT Category, COUNT(ProductID) AS Products

FROM

(SELECT p.ProductID, p.Name AS Product, c.Name AS Category

FROM SalesLT.Product AS p

JOIN SalesLT.ProductCategory AS c

ON p.ProductCategoryID = c.ProductCategoryID) AS ProdCats

GROUP BY Category

ORDER BY Category;

**05- CTEs : Common table expression**

--Using a CTE

WITH ProductsByCategory (ProductID, ProductName, Category)

AS

(

SELECT p.ProductID, p.Name, c.Name AS Category

FROM SalesLT.Product AS p

JOIN SalesLT.ProductCategory AS c

ON p.ProductCategoryID = c.ProductCategoryID

)

SELECT Category, COUNT(ProductID) AS Products

FROM ProductsByCategory

GROUP BY Category

ORDER BY Category;

**SQL - CREATE**

CREATE TABLE SalesLT.CallLog

(

CallID int IDENTITY PRIMARY KEY NOT NULL (1001,1),

CallTime datetime NOT NULL DEFAULT GETDATE(),

SalesPerson nvarchar(256) NOT NULL,

CustomerID int NOT NULL REFERENCES SalesLT.Customer(CustomerID),

PhoneNumber nvarchar(25) NOT NULL,

Notes nvarchar(max) NULL

);

GO

CREATE TABLE SalesLT.Employee

(EmployeeID int IDENTITY PRIMARY KEY,

EmployeeName nvarchar(256),

ManagerID int);

GO

**SQL - INSERT**

**-- INSERT A ROW**

INSERT INTO SalesLT.CallLog

VALUES

('2015-01-01T12:30:00', 'adventure-works\pamela0', 1, '245-555-0173', 'Enquiry about delivery');

**-- INSERT DEFAULTS AND NULLS**

INSERT INTO SalesLT.CallLog

VALUES

(DEFAULT, 'adventure-works\david8', 2, '170-555-0127', NULL);

**-- INSERT A ROW WITH EXPLICIT COLUMNS**

INSERT INTO SalesLT.CallLog (SalesPerson, CustomerID, PhoneNumber)

VALUES

('adventure-works\jillian0', 3, '279-555-0130');

**-- INSERT MULTIPLE ROWS**

INSERT INTO SalesLT.CallLog

VALUES

(DATEADD(mi,-2, GETDATE()), 'adventure-works\jillian0', 4, '710-555-0173', NULL),

(DEFAULT, 'adventure-works\shu0', 5, '828-555-0186', 'Called to arrange deliver of order 10987');

**-- RETRIEVING INSERTED IDENTITY**

SELECT SCOPE\_IDENTITY();

**--OVERRIDING IDENTITY**

SET IDENTITY\_INSERT SalesLT.CallLog ON;

INSERT INTO SalesLT.CallLog (CallID, SalesPerson, CustomerID, PhoneNumber)

VALUES

(11, 'adventure-works\josé1', 11, '926-555-0159');

SET IDENTITY\_INSERT SalesLT.CallLog OFF;

**SQL – UPDATE**

**UPDATE** SalesLT.Customer

SET EmailAddress = NULL

WHERE CustomerID % 7 = 1;

**-- UPDATE MULTIPLE COLUMNS**

UPDATE SalesLT.CallLog

SET SalesPerson = '', PhoneNumber = ''

**-- UPDATE FROM RESULTS OF A QUERY**

UPDATE SalesLT.CallLog

SET SalesPerson = c.SalesPerson, PhoneNumber = c.Phone

FROM SalesLT.Customer AS c

WHERE c.CustomerID = SalesLT.CallLog.CustomerID;

**SQL – DELETE**

DELETE FROM SalesLT.CallLog

WHERE CallTime < DATEADD(dd, -7, GETDATE());

DELETE From SalesLT.Employee

Where EmployeeName = 'adventure-works\david8'

DELETE FROM SalesLT.Employee

WHERE ManagerID = 3

The TRUNCATE TABLE statement is used to delete the data inside a table, but not the table itself.

TRUNCATE TABLE SalesLT.CallLog;

**SQL - ALTER**

The ALTER TABLE statement is used to add, delete, or modify columns in an existing table.

The ALTER TABLE statement is also used to add and drop various constraints on an existing table.

**ALTER TABLE - ADD Column**

ALTER TABLE table\_name

ADD column\_name datatype;

**Example:**

ALTER TABLE Persons

ADD DateOfBirth date;

**ALTER TABLE - DROP COLUMN**

ALTER TABLE table\_name

DROP COLUMN column\_name;

**Example:**

ALTER TABLE Persons

DROP COLUMN DateOfBirth;

**ALTER TABLE - ALTER/MODIFY COLUMN**

To change the data type of a column in a table, use the following syntax:

ALTER TABLE table\_name

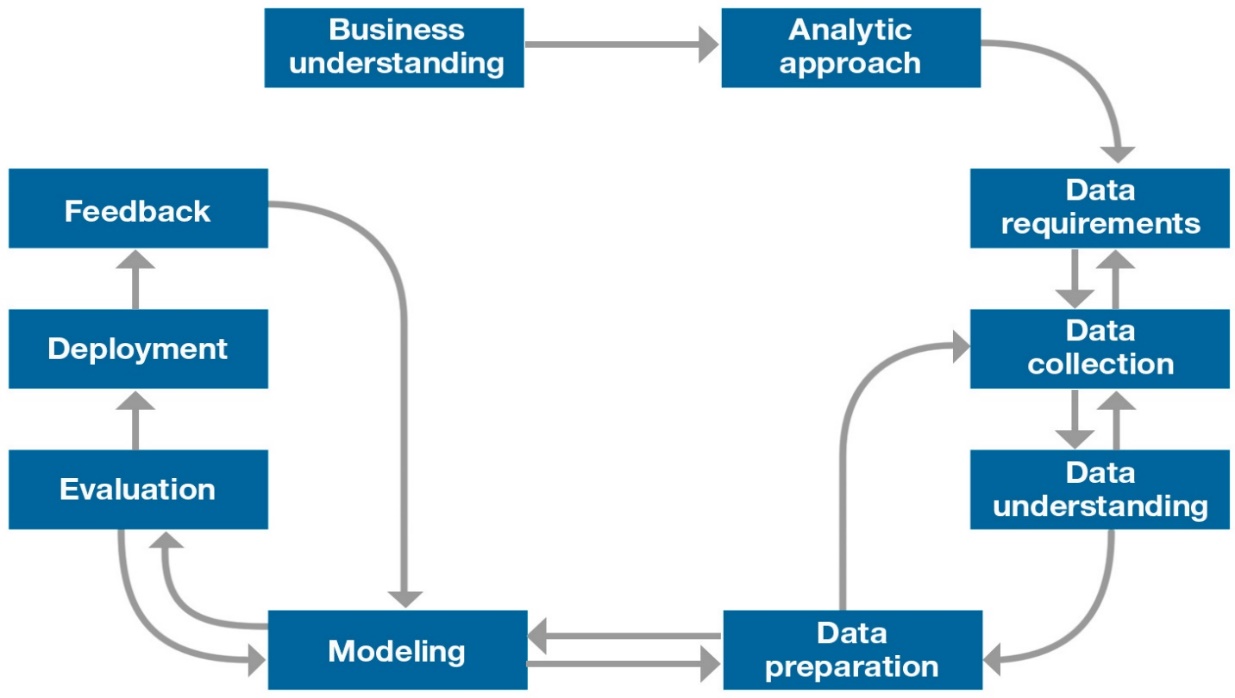
ALTER COLUMN column\_name datatype;

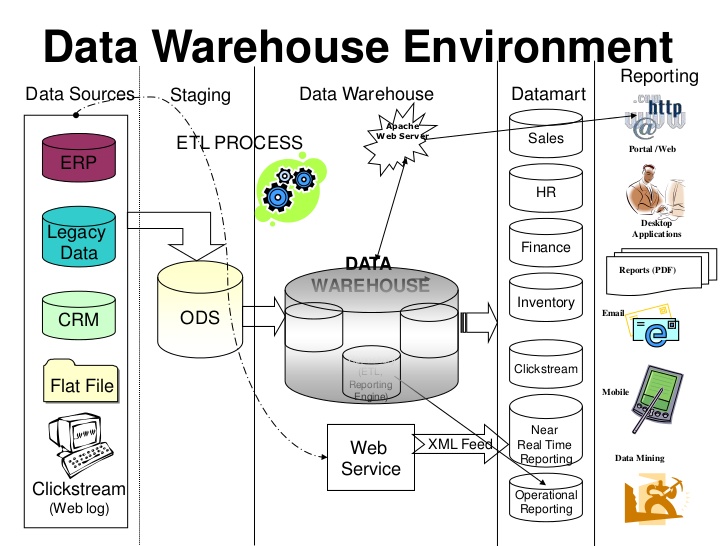
Example:

ALTER TABLE Persons

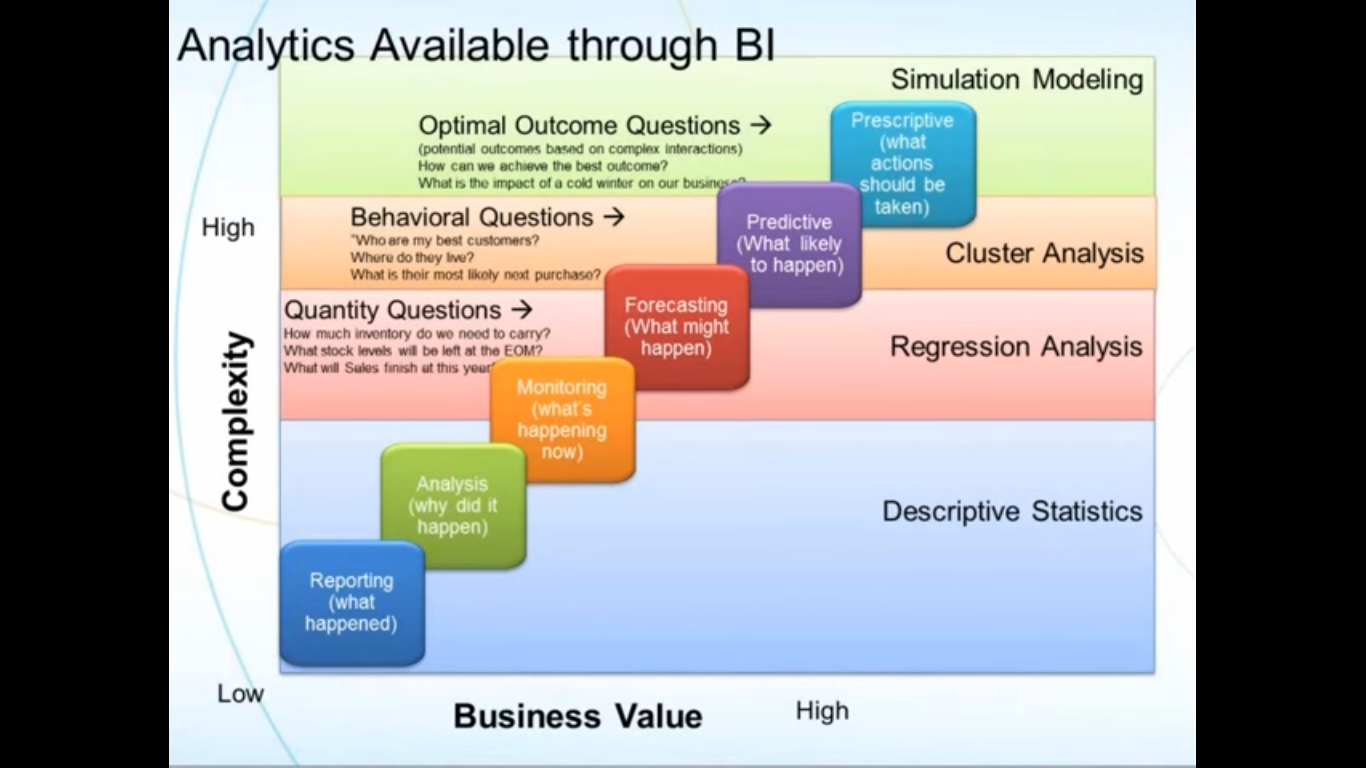
ALTER COLUMN DateOfBirth year;

**Foundation Methodology for Data Science**





**Business Intelligence and Data Science Trade off**

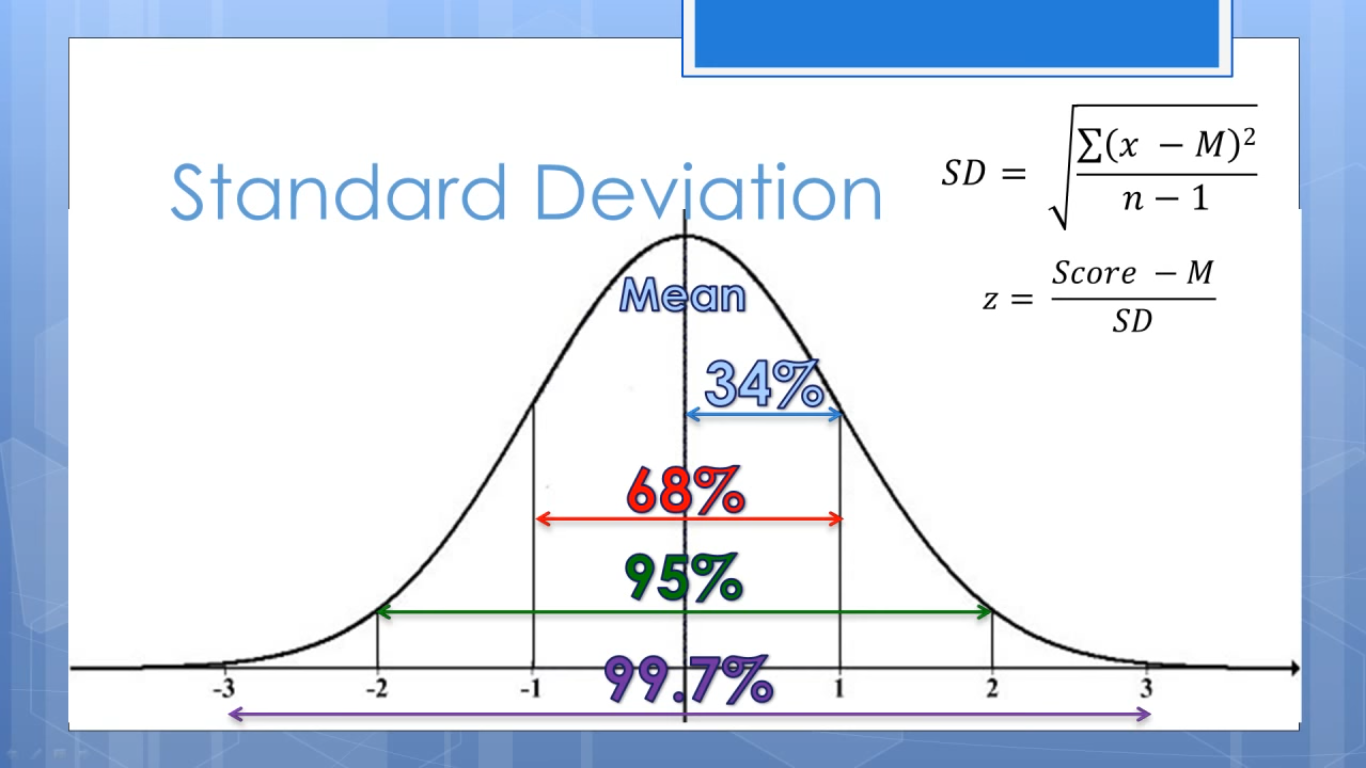


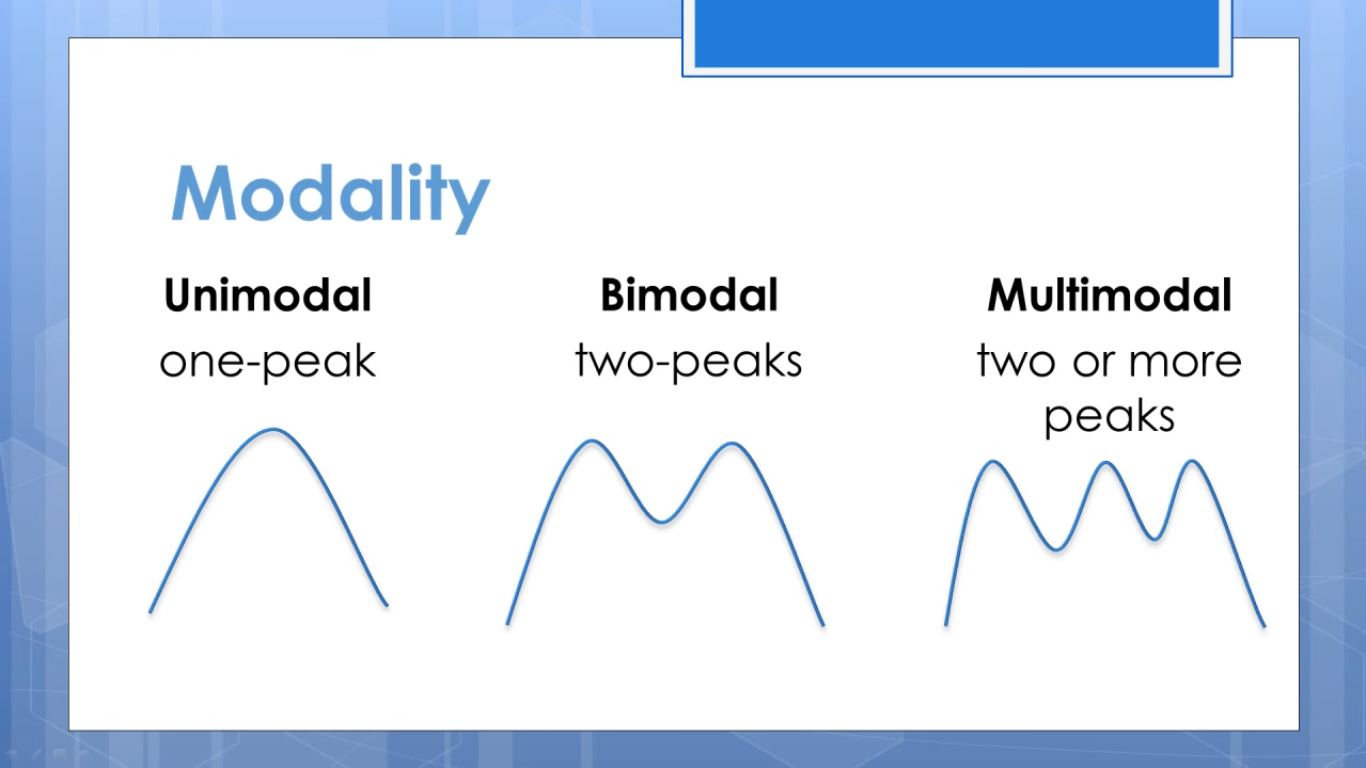
**Normal Distributions, Standard Deviations, Modality, Skewness and Kurtosis: Understanding concepts**

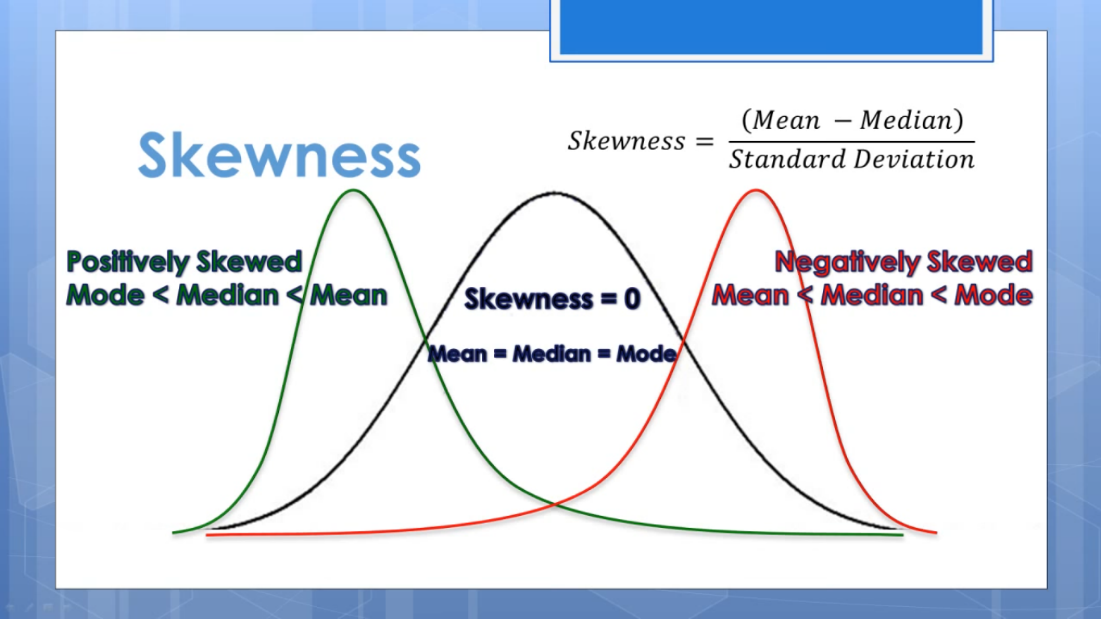
Mean = Median = Mode

**Variance: A measure of data dispersion.**

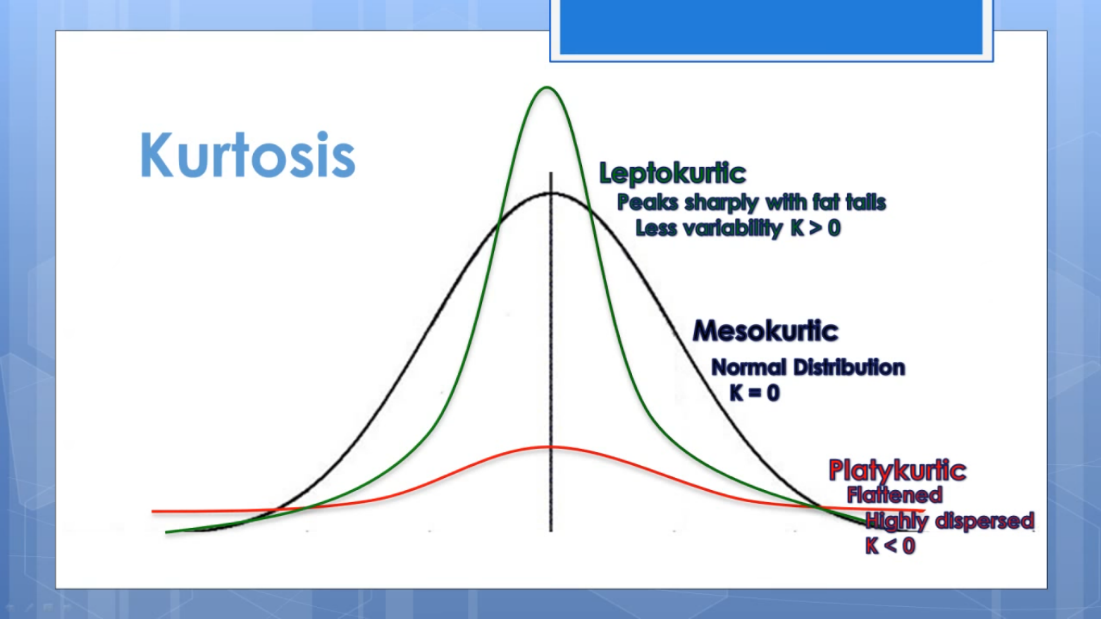
**Standard Deviation: The square root of variance.**







**Skewness: A measure of symmetry or asymmetry in the distribution of data.**

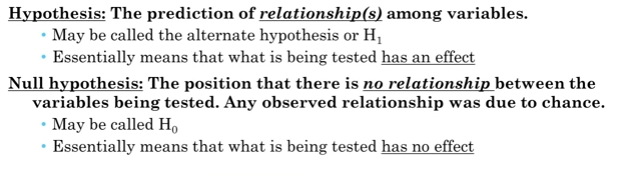


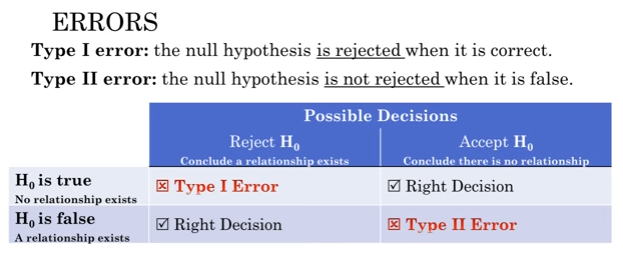
**Kurtosis: A measure of whether the data are peaked or flat relative to a normal distribution.**

**T Tail Test (For 1 independent variable)**

**F Tail test (For more than 1 independent variable or predictors)**

**NULL Hypothesis 🡪 NULL 🡪 NO Relation**





# **What is a chi-square test?**

The chi-square test can be used to determine the **association between categorical variables**. It is based on the difference between the expected frequencies (e) and the observed frequencies (n) in one or more categories in the frequency table.

# **What is Cramer's V2**

Cramer's V2 measures **association between two variables**. A value of z**ero indicates that there is no association**. A value of one indicates that there is a perfect association.

# **What is Fisher's exact test?**

Fisher's exact test to **analyze a 2x2 contingency table** (frequency table) and test whether the **row variable and column variable are independent** (H0: the row variable and column variable are independent).

## What is a contingency table?

A contingency table is table that **tallies observations by multiple categorical variables**. The tables' rows and columns correspond to these categorical variables.

**One Tail Vs 2 Tails test:**

**What is a one-tailed test? -** Testing for the **possibility of the relationship** in one directions. Example: If you are using a significance level of .05, a one-tailed test allots all of your alpha to testing the statistical significance in the one direction of interest.

**What is a two-tailed test? -** Testing for the **possibility of the relationship** in both directions. If you are using a significance level of 0.05, a two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction.

***Example:***

*Imagine you have developed a new drug that you believe is an improvement over an existing drug.  You wish to* ***maximize your ability to detect the improvement****, so you opt for a one-tailed test. In doing so, you fail to test for the* ***possibility that the new drug is less effective than the existing drug****.*

**ANOVA**

ANOVA is used to compare differences of means among more than 2 groups. ANOVA compares the amount of variation between groups with the amount of variation within groups.

Step 1) Variation between groups (Fig 1)

Step 2) Variation within groups  (Fig 1)

Step 3) The F ratio

The F ratio is then calculated as:

https://www.edanzediting.com/sites/default/files/2-stats-anova.png

If the **p value from the ANOVA is significant** (less than 0.05 or your chosen alpha level), then you can conclude that the **groups are not all the same** (because the means varied from each other by too large an amount).

**Bivariate Analysis**

Bivariate analysis is the **simultaneous analysis of two variables** (attributes). It explores the concept of relationship between two variables, whether there exists an association and the strength of this association, or whether there are differences between two variables and the significance of these differences.

**Predictive Modeling and Machine Learning**

1. **Regression:**

Regression is a data mining task of predicting the value of target (numerical variable) by building a model based on one or more predictors (numerical and categorical variables).

**OLS Regression**

**Watch out for P, Residual Plot, Coefficients and Intercepts, R-Squared and goodness’ of fit**

Regression analysis generates an equation to describe the statistical relationship between one or more predictor variables (independent variables) and the response variable (dependent variable)

# **Why You Need to Check Your Residual Plots for Regression Analysis:**

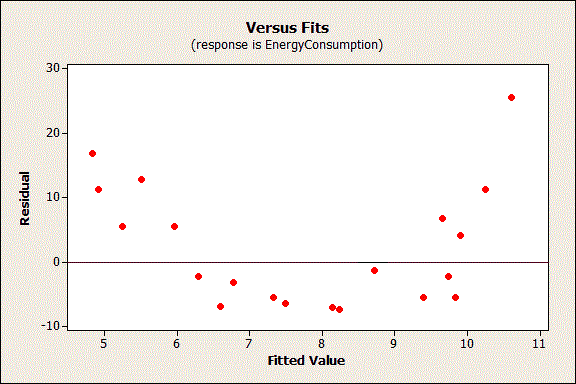
Two basic components of a valid regression model: **Response = (Constant + Predictors) + Error**  (Or) **Response = Deterministic + Stochastic**

**The Deterministic Portion:** This is the part that is explained by the predictor variables in the model.

**The Stochastic Error:** Stochastic is a fancy word that means random and unpredictable. Error is the difference between the expected value and the observed value.

**Assumptions:**

1. **The residuals should not be correlated with another variable**
2. **Adjacent residuals should not be correlated with each other (autocorrelation).**



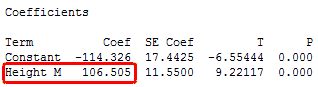
## Interpret the P-Values in Linear Regression Analysis?

## The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis 🡪 There is a significant relationship between X and Y

## Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response; 🡪 Probability of getting Null hypothesis is more, there is no significant relationship between X and Y.

## Interpret the Regression Coefficients for Linear Relationships?

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant



 The coefficient indicates that for every additional meter in height you can expect weight to increase by an average of 106.5 kilograms, keeping other thing constants.

## What Is R-squared?

**R-squared = Explained variation / Total variation**

R-squared is always between 0 and 100%:

* 0% indicates that the model **explains** **none of the variability** of the response data around its mean.
* 100% indicates that the model explains all the **variability of the response data around its mean**.

R-squared does not indicate whether a regression model is adequate. **You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data**!

**It’s just fine to have low R-squared values**.

For example, any **field that attempts to predict human behavior**, such as psychology, typically has R-squared values **lower than 50%**. Humans are simply harder to predict than, say, physical processes.

**Logistic Regression: Nominal regression family. glm function from MASS package**

**Ordinal Regression 🡪 MASS, polr function**

**Nominal Regession🡪**

**Clustering**

1. **KNN Clustering: (Based on Number of Nearest Neighbor) – 3 Near Neighbor etc.…**

**(K D Tree ) KNN Clustering: ( Based on Number, Nearest Neighbor)**

* Key: Nearby points are used as a basis of learning the alogrithms
* **Voronoi tessellation**: Decision boundary separate one class from another class.
* **Monotectic**, Hard boundaries and Hierarchical

In this case, we are given some data points for training and also a new unlabelled data for testing. Our aim is to find the class label for the new point. The algorithm has different behavior based on k.

* Case 1 : k = 1 or Nearest Neighbor Rule
* Case 2 : k = K or k-Nearest Neighbor Rule

Typically k is odd when the number of classes is 2. Lets say k = 5 and there are 3 instances of C1 and 2 instances of C2. In this case , KNN says that new point has to labeled as C1 as it forms the majority.

**Voronoi tessellation:** In mathematics, a **Voronoi diagram** is a [partitioning](https://en.wikipedia.org/wiki/Partition_of_a_set) of a [plane](https://en.wikipedia.org/wiki/Plane_(geometry)) into regions based on distance to points in a specific subset of the plane

Choosing value of K:

Smaller K 🡪 Higher variability, unstable decision.

Larger K 🡪 take all the input variables and assign most frequent class as outcome

1. **K Mean -> Centroid Method**

**K Mean Clustering (Based on Distance)**

One of the most frequently used unsupervised algorithms is K Means. K Means Clustering is exploratory data analysis technique. This is **non-hierarchical method** of grouping objects together.

* *Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.*
* *Assign each object to the group that has the* ***closest centroid****.*
* *When all objects have been assigned, recalculate the* ***positions of the K centroids****.*
* *Repeat Steps 2 and 3 until the* ***centroids no longer move****. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.*
* *Key: No categorical or Ordinal data only Numeric; Distance b/w each point and nearest centroid decreases every time.*

How many cluster:

* Try with different number of K’s (cluster) and optimize the variance.
* Look for elbow method – Scree plot
* Polythetic, Hard boundaries and flat in nature

1. **KNN Parzen window clustering:** Size of the **window remains constant**, ma**jority will be considered** ; if equal – More **weight give for near neighbor** , or flip a coin method, Nearest one
2. **Hierarchical Clustering: Top down & Bottom up**

* Agglomerative Clustering – Bottom UP
* Divisive clustering – Top down approach

**Decision Trees, Random Forest:**

* Decision tree builds regression or classification models in the form of a tree structure.
* It breaks down a dataset into smaller and smaller subsets, final result is a tree with decision nodes and leaf nodes
* It a greedy algorithm keeps on splitting until find mutually exclusive datasets, or Single tons.
* Works on divide and conquer algorithm
* If the inputs are numerical data 🡪 Converts into Categories by Binning

**How it works:**

* 1. Checks for the **best attribute for splitting** base on entropy and information gain

**Entropy** 🡪 Basically tells how pure or impure the subset is. If subset is pure 🡪 0 bits (output of the entropy)

**Information Gain** 🡪 It’s the aggregation of entropy, higher the information gain more certain about the split. Attributes with higher information gain would be considered as root node, followed by branch and leaf node.

Overfitting of Decision Tree: Keeps splitting until each node contain pure subsets

To avoid overfitting: Stop splitting when statistically not significant and pruning using validation set.

**Random Forest: Instead of training dataset, RF train subsets. It pick random subset and learns to classify**

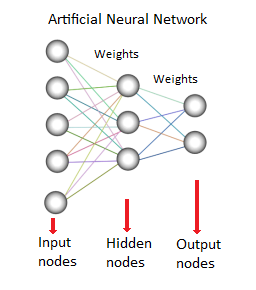
**Given new data, X is classified using each of the trees use majority voting – class predicted most often.**

### **Artificial Neural Network**

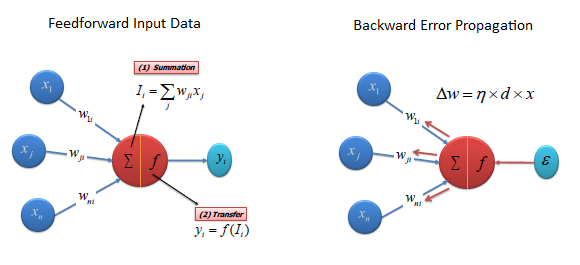
An artificial neutral network (ANN) is a system that is based on the biological neural network, such as the brain. The brain has approximately **100 billion neurons**, which communicate through **electro-chemical signals**. The neurons are connected through junctions called synapses.

Each neuron receives thousands of connections with other neurons, constantly receiving incoming signals to reach the cell body. If the resulting sum of the sig**nals surpasses a certain threshold**, a response is sent through the axon.

An **ANN** is comprised of a **network of artificial neurons** (also known as "nodes"). These nodes are connected to each other, and the strength of their connections to one another is assigned a value based on their strength: inhibition (maximum being -1.0) or excitation (maximum being +1.0). Three types of neurons in an ANN,**input nodes**, **hidden nodes**, and **output nodes**.



The input nodes take in information, in the form which can be numerically expressed, based on the weights. Based on the connection strengths (weights), inhibition or excitation, and transfer functions, the activation value is passed from node to node and signal is sent to output.



Transfer (Activation) Functions

The transfer function translates the input signals to output signals. Four types of transfer functions are commonly used, **Unit step (threshold), sigmoid, piecewise linear, and Gaussian**.

**Algorithm**

There are different types of neural networks, but they are generally classified into feed-forward and feed-back networks.

A **feed-forward network** is a non-recurrent network which contains inputs, outputs, and hidden layers; the signals can only travel in one direction.

A **feed-back network** has feed-back paths meaning they can have signals traveling in both directions using loops. All possible connections between neurons are allowed.

**Anomaly Detection (Outlier Detection):**

Anomaly detection is a technique used to identify **unusual patterns that do not conform to expected behavior**, called outliers.

It has many applications in business, from intrusion detection (**identifying strange patterns in network traffic that could signal a hack**) to system health monitoring (spotting a **malignant tumor in an MRI scan**), and from fraud detection in credit card transactions to **fault detection in operating environments**.

1. **Point anomalies:** A single instance of data is anomalous if it's too far off from the rest. *Business use case:* Detecting credit card fraud based on "amount spent."
2. **Contextual anomalies:** The abnormality is context specific. This type of **anomaly is common in time-series** data. *Business use case:* Spending $100 on food every day during the holiday season is normal, but may be odd otherwise.
3. **Collective anomalies:** A set of data instances collectively helps in detecting anomalies. *Business use case:*Someone is **trying to copy data form a remote machine to a local host** unexpectedly, an anomaly that would be flagged as a potential cyber attack.

**Method 1: Simple Statistical Methods**

The simplest approach to identifying irregularities in data is to flag the data points that deviate from common statistical properties of a distribution, including mean, median, mode, and quantiles. Let's say the definition of an anomalous data point is one that deviates by a certain standard deviation from the mean.

**Method 2: Machine Learning-Based Approaches**

**Density-Based Anomaly Detection**

Density-based anomaly detection is based on the k-nearest neighbors algorithm. Assumption: Normal data points occur around a dense neighborhood and abnormalities are far away.

**Clustering-Based Anomaly Detection**Clustering is one of the most popular concepts in the domain of unsupervised learning. Assumption: Data points that are similar tend to belong to similar groups or clusters, as determined by their distance from local centroids.

**Support Vector Machine-Based Anomaly Detection :** A SVM is typically associated with supervised learning, but there are extensions ([OneClassSVM,](http://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html" \t "_blank) for instance) that can be used to identify anomalies as an unsupervised problems (in which training data are not labeled). The algorithm learns a soft boundary in order to cluster the normal data instances using the training set, and then, using the testing instance