

# ADVANCE STATISTICS ( Week 2)

# DSBA CURRICULUM DESIGN

## FOUNDATIONS

Python for Data  
Science

Statistical Methods  
for Decision Making

## CORE COURSES

Advanced  
Statistics(Week-2/5)

Data Mining

Predictive Modelling

Machine Learning

Time Series  
Forecasting

Data Visualization

SQL

## DOMAIN APPLICATIONS

Financial Risk  
Analytics

Marketing Retail  
Analytics



# LEARNING OBJECTIVE OF THIS MODULE

- ANOVA
- EDA
- PCA

# LEARNING OBJECTIVES OF THIS SESSION

- Removing Duplicates
- Missing Value Treatment
- Outlier Treatment
- Normalizing and Scaling
- Encoding Categorical Variable
- Bivariate Analysis

## TRY ANSWERING THE FOLLOWING

- What is the syntax to check missing value in a data?
- Which formula is used to check outliers in a data?
- Does extreme values impact Median of a given variable?





## EDA : A Bank Account that increases life expectancy!

### Do you know how your data has been generated?

The way banking sector has evolved over the years is through a lot of mergers i.e. a number of small co-operative bank merged into a large Bank. Obviously these small banks have to adapt to the work culture and operational practices of the larger bank and there is a transition period involved, but there are times when we come across interesting findings when we analyse the data.

Back in 2016 while working on the historical data of a large bank, it was found that ~8% of their customers are more than 115 years old. While this was surprising, it was further found that the date of birth of all the customers was 1-1-1900. Upon investigation it was found that all these customers belonged to the co-operative banks which were merged some 3-4 years back. Now, often these co-operative banks operate out of rural areas and didn't maintain the customer DoB information, because in a lot of rural areas the birth date was not properly captured in any record even by the family. The MIS(management information system) resource collating the data simply entered a default value for DoB.

Reference: We would rather not name the bank, but similar findings are very much possible in any organization.





# EDA-Exploring your Social Media posts

## **It is not only your friends and family members who are interested in your posts/updates**

American Express Company, also known as AMEx, is best known for its charge cards, credit cards and traveller's cheques. Not only that the company has always been class apart when it comes to its services, American Express is one of the first to put its data analytics and capabilities to work. As a global integrated network, it has the data advantage because it can fully view all transactions in real-time, which means working with one of the largest, most valuable and distinct data sets in the entire world.

American Express has developed sophisticated techniques to predict a customers likelihood of default by constantly monitoring his/her social media account activities. Please remember, exploratory data analysis is not limited to structured data, but even the unstructured data such as what you write on twitter or Facebook can be analysed e.g. does your social media account reveal that you are desperately looking out for a job opportunity? Or you are broke because of a massive layoff in your organization which is likely to affect you? Have you recently changed your address/city? All these indicators are meaningful for your bank while commenting on your credit worthiness.





## Let's Learn Together – A Unique Platform for Peer to Peer Learning

### Next Week's Theme:

### PCA and Its Real Time/Industry Applications



### Benefits of Peer to Peer Learning:

- ❖ Active Learning
- ❖ Gain a Deeper Understanding
- ❖ Feel More Comfortable
- ❖ Personalized Learning Experience

### What all can be discussed in a Discussion forum?

- ❖ Analytical Concepts
- ❖ Issues in Code
- ❖ Real Time/Industry Examples



# Car Case Study

Car, an aspirational product for many is a product which everyone wants to own one day. There are various car models from several car manufacturers available in the market. A lot of pre-launch research is done by company before introducing a new car model in a market as this is important because new car launch requires huge amount of investments in the form of R&D, marketing, manufacturing and logistics. A wrong market study could lead to huge dept on company's balance sheet.

A new Automobile company is planning to expand and introduce new models in the US Market. We have a data about the car owners from few US cities. Perform EDA on the data and state the outcomes from the data.



**ANY QUESTIONS**





**HAPPY LEARNING**