**Probability & Statistics**

**IPE 205**

by

Tanmoy Das

<https://github.com/tanmoyie/applied-statistics>

[www.kaggle.com/tanmoyie](https://www.kaggle.com/tanmoyie)

<https://www.youtube.com/playlist?list=PLHyZ7Tamw-fdGRPgIXiYDZuLdtYoeP4CM>

\*\* This document is incomplete & will be updated frequently

**Estimated Contents:**

1. Introduction to Statistics, Data, Data Visualization (Box plot, Normal Probability Plot etc), EDA **project**
2. Basic laws of probability, conditional probability, rando. .m variables, measures of central tendency and dispersion, mathematical exception, probability distributions, transformation of variables, moments and moment generating functions, sampling,
3. Correlation and regression analysis, **project**
4. Analysis of variance, **project, Data Mining**
5. Central limit theorem, chi-square distribution, t-distribution, f-distribution: estimation and confidence interval,
6. Statistical hypothesis and testing, goodness-of-fit tests.
7. Experimental designs, randomized block design, factorial design, introduction to stochastic problems in engineering. **Project**

# Introduction to Data

Introduction to Statistics, Data, Data Visualization (Box plot, Normal Probability Plot etc), EDA **project**

Show 1 table of data & graph side by side

Nominal data vs numeric data

Roll number vs cgpa

Positive correlation & no correlation

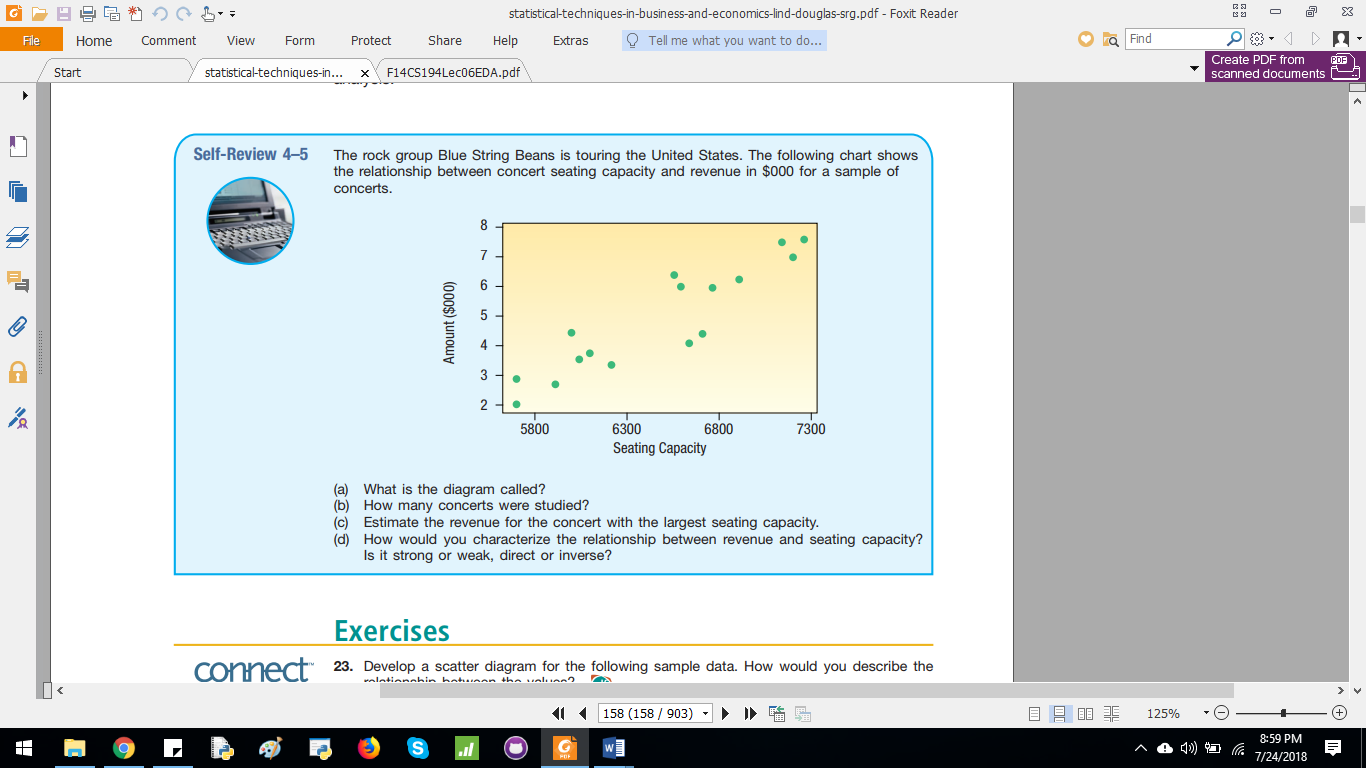
Attendance vs Cgpa ; age of student vs cgpa (find real data of IPE 02, mist)

## Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

Typical [graphical techniques](https://en.wikipedia.org/wiki/Statistical_graphics) used in EDA are:

* [Box plot](https://en.wikipedia.org/wiki/Box_plot)
* [Histogram](https://en.wikipedia.org/wiki/Histogram)
* [Multi-vari chart](https://en.wikipedia.org/wiki/Multi-vari_chart)
* [Run chart](https://en.wikipedia.org/wiki/Run_chart)
* [Pareto chart](https://en.wikipedia.org/wiki/Pareto_chart)
* [Scatter plot](https://en.wikipedia.org/wiki/Scatter_plot)
* [Stem-and-leaf plot](https://en.wikipedia.org/wiki/Stemplot)
* [Principal component analysis](https://en.wikipedia.org/wiki/Principal_component_analysis) (PCA)



## Introduction to Data Science

**What are various steps involved in an Data Science project?**

•           Understand the business problem

•           Explore the data and become familiar with it.

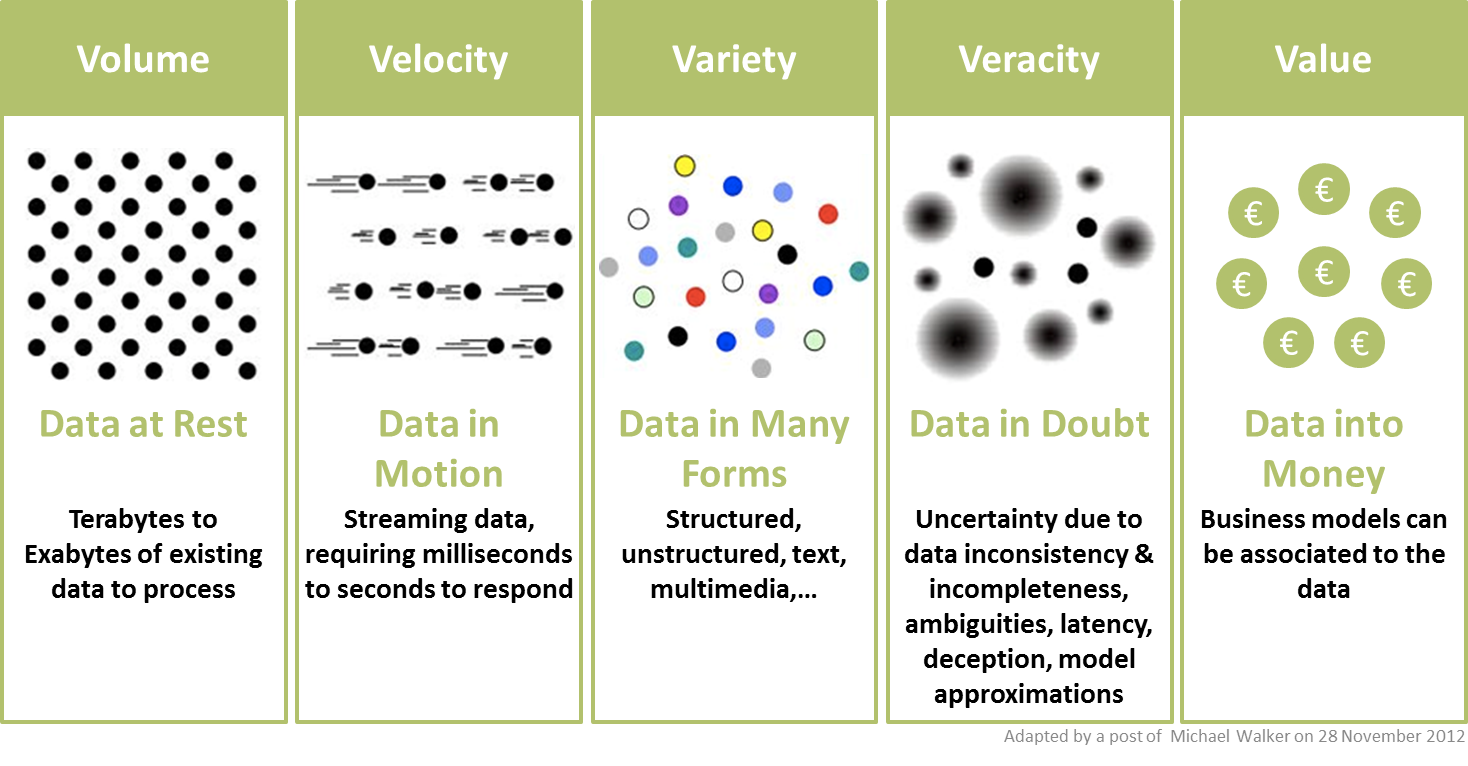
•           Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.

•           After data preparation, start running the model, analyse the result and tweak the approach. This is an iterative step till the best possible outcome is achieved.

•           Validate the model using a new data set.

•           Start implementing the model and track the result to analyse the performance of the model over the period of time.

## Big Data



Veracity, one of the most overlooked Big Data characteristics, is directly related to data quality, as it refers to the inherent biases, noise and abnormality in data. Because of veracity, the data values might not be exact real values, rather they might be approximations. In other words, the data might have some inherent impreciseness and uncertainty.

## Data Visualization

**Relevant questions[[1]](#footnote-1)**

**How do we make a good data visualization?**

To do that, choose the right type of chart for your data:

* Line Charts to track changes or trends over time and show the relationship between two or more variables.
* Bar Charts to compare quantities of different categories.
* Scatter Plots show joint variation of two data items.
* Pie Charts to compare parts of a whole - used them sparingly since people have hard time comparing the area of pie slices
* You can show additional variables on a 2-D plot using color, shape, and size
* Use interactive dashboards to allow experiments with key variables

# Probability & Distribution

Basic laws of probability, conditional probability, rando. .m variables, measures of central tendency and dispersion, mathematical exception, probability distributions, transformation of variables, moments and moment generating functions, sampling,

# Correlation and Regression analysis

Correlation and regression analysis, **project**

# ANOVA

analysis of variance, **project, Data Mining**

# Central Limit Theorem & distributions in its propencity

central limit theorem, chi-square distribution, t-distribution, f-distribution: estimation and confidence interval,

# Hypothesis Testing

Statistical hypothesis and testing, goodness-of-fit tests.

# Engineering Design Experiments

Experimental designs, randomized block design, factorial design, introduction to stochastic problems in engineering. **Project**

Experimental designs

P442, Six sigma handbook

# Data Mining

1. https://www.kdnuggets.com [↑](#footnote-ref-1)