zekeLabs

Statistics for Data Science

“Goal - Become a Data Scientist”

“A Dream becomes a Goal when action is taken towards its achievement” - Bo Bennett

“The Plan”

“A Goal without a Plan is just a wish”

Overview of Statistics

- Introduction to Statistics
- Importance of Statistics
- Understanding Variables Types
- Descriptive vs Inferential Statistics

Introduction to Statistics

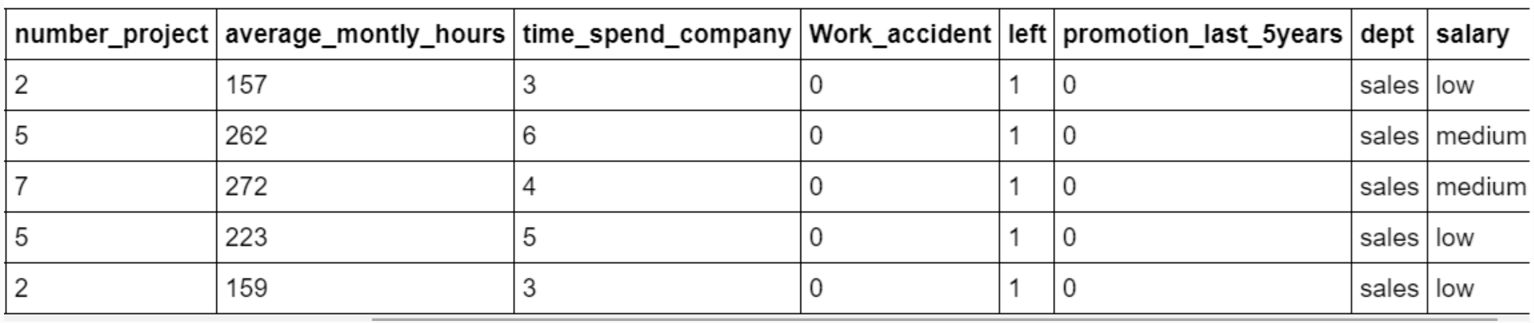
- Science of learning from data.
- Methodical data collection.
- Employ correct data analysis.
- Presenting analysis effectively.
- Opposite to statistics is “Anecdotal Evidence”.

Importance

- Avoid getting biased samples
- Prevent overgeneralization
- Wrong causality
- Incorrect Analysis
- Applied to any domain

Variables

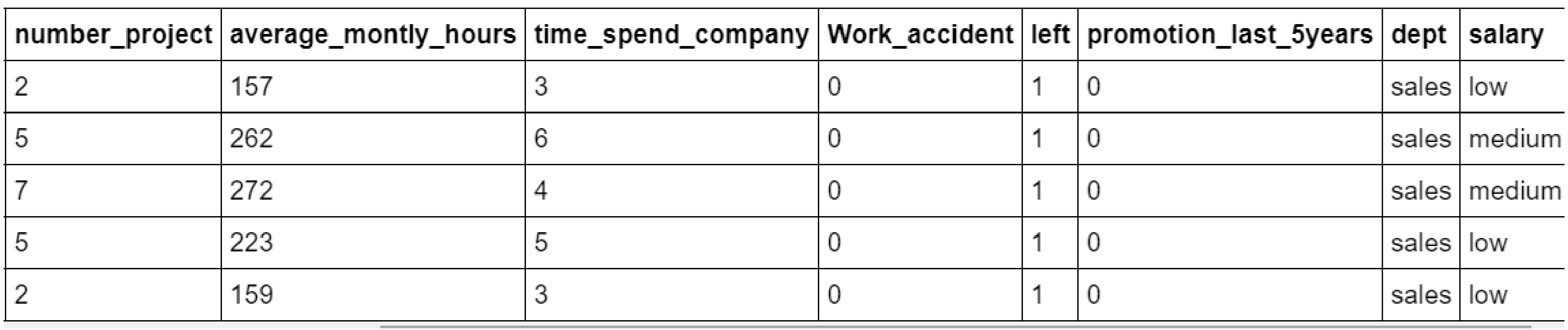
- Explanatory (predictor or independent)
- Response (outcome or dependent)
- A variable can serve as independent in one study and dependent in another



number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	dept	salary
2	157	3	0	1	0	sales	low
5	262	6	0	1	0	sales	medium
7	272	4	0	1	0	sales	medium
5	223	5	0	1	0	sales	low
2	159	3	0	1	0	sales	low

Data Types of Variables - Quantitative versus Qualitative

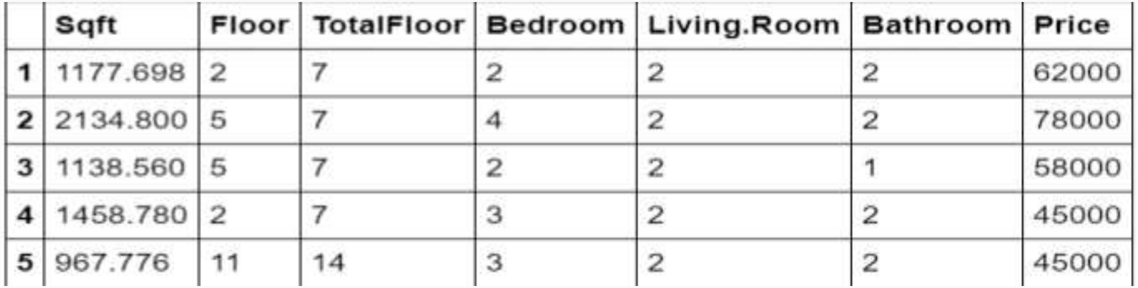
- Quantitative - Numerical data. Eg. weight, temperature, number_project
- Qualitative - Non-numerical data. Eg. dept, salary



number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	dept	salary
2	157	3	0	1	0	sales	low
5	262	6	0	1	0	sales	medium
7	272	4	0	1	0	sales	medium
5	223	5	0	1	0	sales	low
2	159	3	0	1	0	sales	low

Types of Quantitative Variables

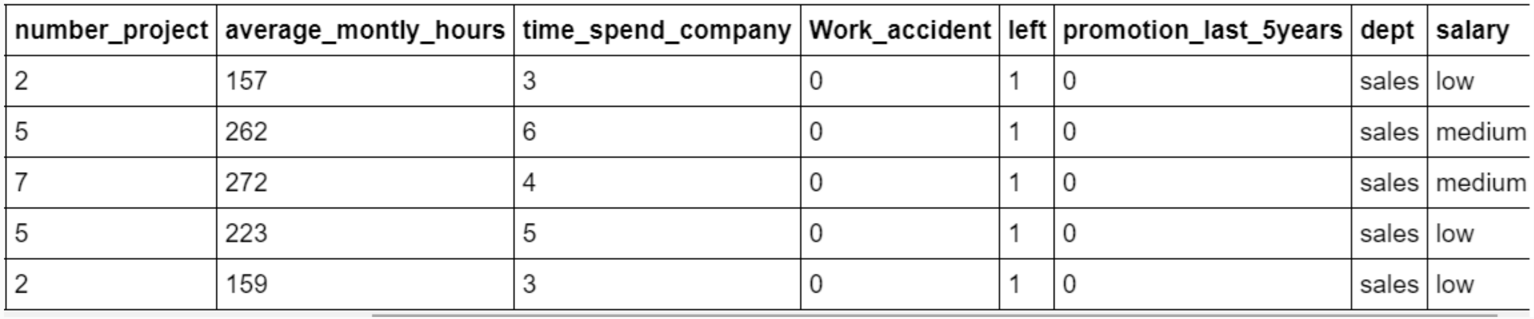
- Continues - any numeric value. Eg. Sqft
- Discrete - count of the presence of a characteristic, result, item, or activity. Eg. Floor



	Sqft	Floor	TotalFloor	Bedroom	Living.Room	Bathroom	Price
1	1177.698	2	7	2	2	2	62000
2	2134.800	5	7	4	2	2	78000
3	1138.560	5	7	2	2	1	58000
4	1458.780	2	7	3	2	2	45000
5	967.776	11	14	3	2	2	45000

Qualitative Data: Categorical, Binary, and Ordinal

- Categorical or Nominal. Eg - dept (sales, RD etc.)
- Binary. Eg. Left (1 or 0)
- Ordinal. Eg. salary (low, medium, high)



number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	dept	salary
2	157	3	0	1	0	sales	low
5	262	6	0	1	0	sales	medium
7	272	4	0	1	0	sales	medium
5	223	5	0	1	0	sales	low
2	159	3	0	1	0	sales	low

Choosing Statistical Analysis based on data type

Types of Statistical Analysis

- Descriptive Statistics - Describes data.
    - Common Tools - Central tendency, Data distribution, skewness

- Inferential Statistics - Draw conclusions from the sample & generalize for entire population
    - Common Tools - Hypothesis Testing, Confidence Intervals, Regression Analysis

Summarizing Data

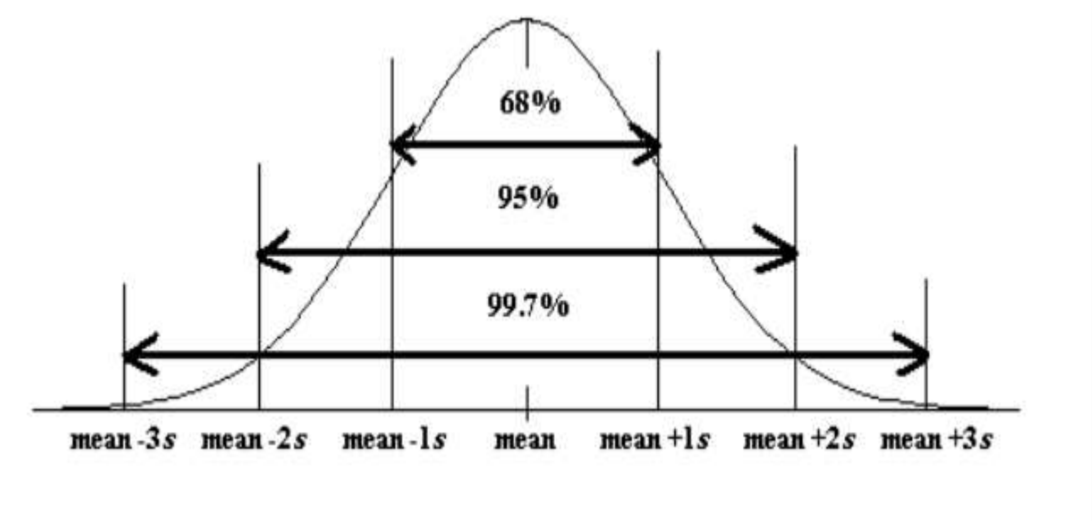
- Measure of Central Tendency
- Measure of Variability
- Visualizing Data

Measure of Central Tendency

- Mean - Average of data, suited for continuous data with no outliers
- Median - Middle value of ordered data, suited for continuous data with outliers
- Mode - Most occuring data, suited for categorical data (both nominal and ordinal)

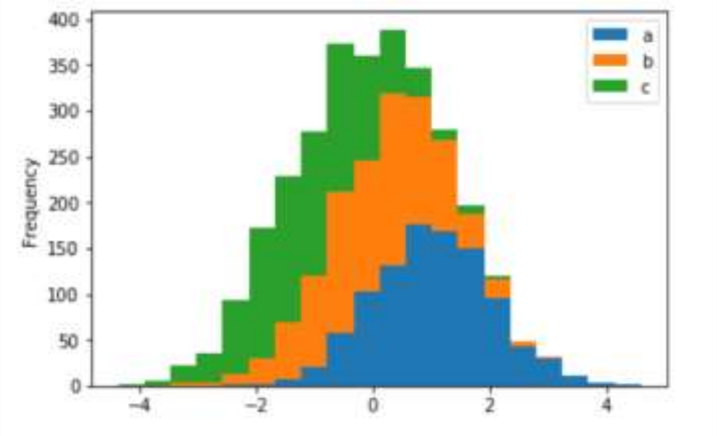
Measure of Variance

- Range
- Interquartile Range
- Variance
- Standard Deviation

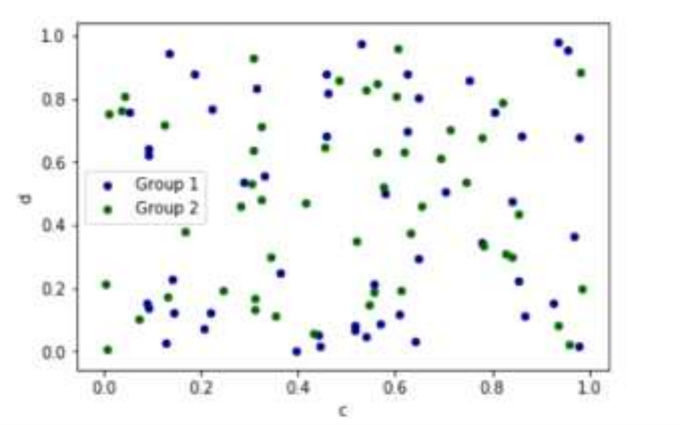


Visualizing Continuous Data

- Histogram

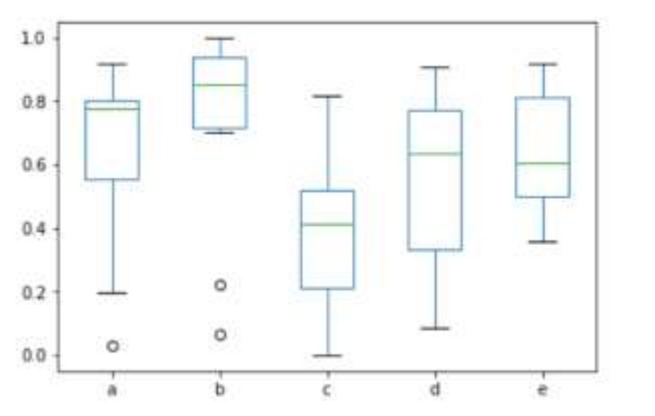


- ScatterPlot



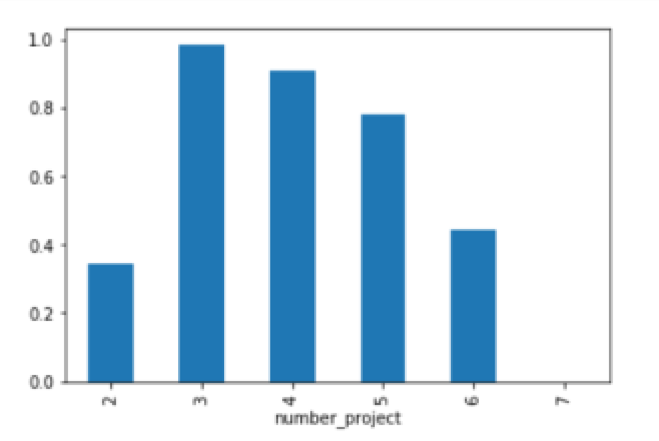
Visualizing Continuous Data - 2

- Box-Plot

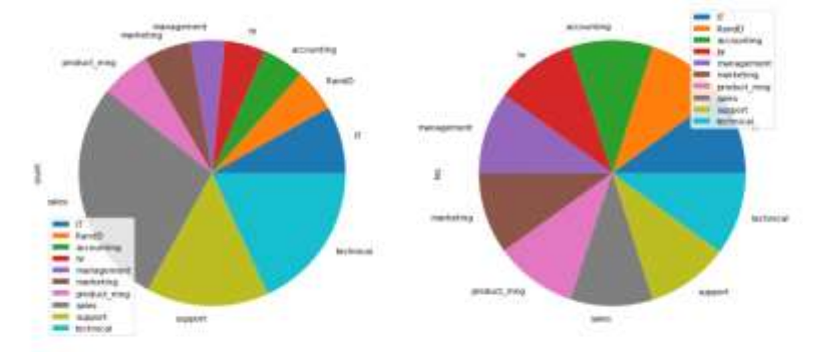


Visualizing Discrete Data

- Histogram



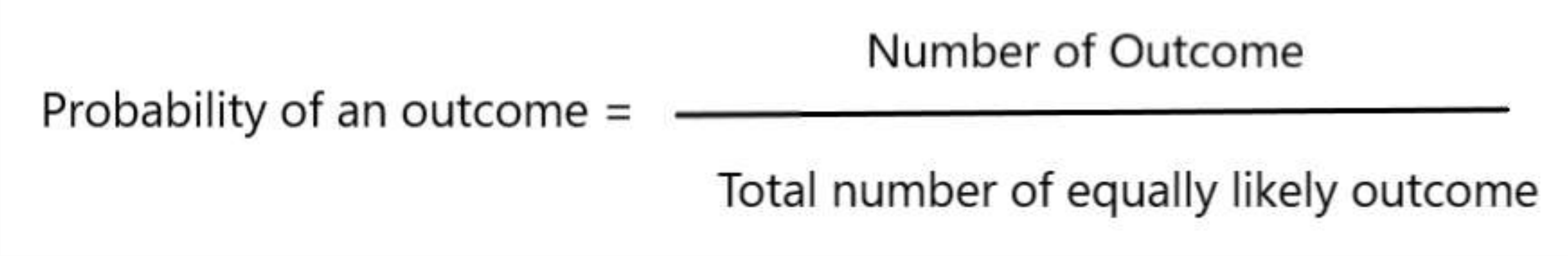
- Pie



Probability Distribution

- Basics of Probability
- Conditional Probability
- Discrete Probability Function
- Continuous Probability Function
- Central Limit Theorem

Probability of Single Event



$$\text{Probability of an outcome} = \frac{\text{Number of Outcome}}{\text{Total number of equally likely outcome}}$$

Probability of Two Independent Events

P(A AND B) = P(A) * P(B)

Probability of heads on tossing of two coins P(A) * P(B) = ½ * ½ = ¼

P(A OR B) = P(A) + P(B) - P(A AND B)

Probability of head in 1st flip or probability of head in 2nd flip or both

½ + ½ - ¼ = ¾

Conditional Probability

Probability of an event given the other event has occurred.

P(B|A) - Probability of event B given A has happened

P(A AND B) = P(A) * P(B|A)

Probability of drawing 2 aces = P(drawing one ace from deck) * P(drawing one ace given already one ace is pulled out)

Probability of drawing 2 aces = 4/52 * 3/51

Probability distribution

- A function describing the likelihood of obtaining possible values that a random variable can assume.
- Consider salary of employee data, we can create distribution of salary.
- Such distribution is useful to know which outcome is more likely.
- Sum of probability of all outcomes is 1, so every outcome has likelihood between 0 & 1
- PDF are divided into two types based on data - Discrete and Continues

Discrete Probability Distribution Function

- Probability mass functions for discrete data
- Binomial Distribution for Binary Data (Yes/No)
- Poisson Distribution for count data (No. of cars per family)
- Uniform Distribution for Data with equal probability (Rolling dice)

Binomial Distribution

Poisson Distribution

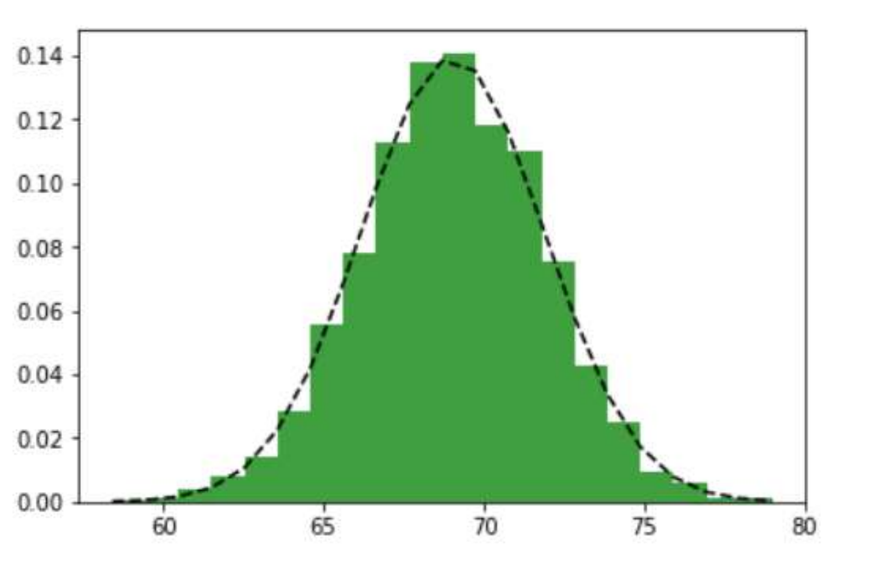
Uniform Distribution

Probability distribution for continuous data

- Probability mass function for continuous data
- Central tendency, variation & skewness important parameters
- Normal Probability Distribution or Gaussian Distribution or Bell curve
- Lognormal Probability Distribution

Normal Distribution

- A probability function that describes how the values of a variable are distributed.
- Symmetric distribution
- Mean = 69, Std = 2.8
- Notation Alert, mu & sigma term used for entire population



Height Distribution

Normal Distribution - 2

- Empirical Rule of Normal Distribution : 68 - 95 - 99
- Standard Normal Distribution : Mean = 0, Std = 1.0
- Z-scores is a great way to understand where a specific observation fall wrt entire population. It is basically number of std far from mean.

Lognormal Distribution

Descriptive Statistics

- Introduction
- Central Tendency
- Data Distribution
- Skewness
- Correlation

Lognormal Distribution

Inferential Statistics

- Introduction
- Hypothesis Testing
- Confidence Intervals
- Regression Analysis

Relationships between Variables

- Chi-square Test of Independence
- Correlation and Linear Regression
- Analysis of Variance or ANOVA

Thank You !!!

Let us know how can we help your organization to Upskill the employees to stay updated in the ever-evolving IT Industry.

www.zekeLabs.com | +91-8095465880 | info@zekeLabs.com