1 Week 1

1. How do you see data ?
   - Good Decisions are based on an accurate understanding of Good data.
   - Present Data in a Precise, Concise and Understandable Way.
   - Two Types of data
     - Categorical
     - Numerical: Discrete and Continuous
   - Core Principle on which visualization of data is done: Nature of data dictates which visualization to use.
2. Benefits of visual representation of data.
   - Communicate complex information concisely and precisely.
   - Create a "picture" for reasoning about and analysing quantitative and conceptual information.
   - Provides "Information Rich View" at a glance.
   - Directs attention towards content rather than methodology.
   - Describe, Explore and Summarize a set of numbers.
   - Convey messages about significance of data.
3. 4 Principles of effective visualization.
   - Know Purpose
   - Ensure Integrity
   - Maximize data ink and Minimize non data ink
   - Show your data, annotate
4. Executing your information display is a 3 step process.
   - Defining the message
     - What am I trying to communicate ?
     - Should I use text, table, graph or a combination ?
     - The message/statistic you want to emphasize
   - Choosing Form
     - What is the message ?
     - What design principles lead to quick cognitive processing & effective communication ?
     - Whether to display the data as a table or a chart
   - Creating Designs
     - How do I make the message clear at a glance ?
     - Avoid 3D effects, Avoid legends(Use Labels), Avoid contrasting borders around objects, Use annotations to highlight key data changes or to focus on specific data points.
5. Dashboard
   - A visual display of the most important information needed to achieve one or more objective that has been consolidated on a single screen so it can be monitored & understood at a glance.
   - Scan the big picture, Zoom in on important specifics, Link to supporting details.

2 Week 2

1. Probability Distributions
   - Trace-Driven Simulation: Data values themselves used directly in simulations.
   - Fit: Use a theoretical distribution for the data.
   - Data values could be used to define empirical distribution.
2. Empirical Distributions
   - Using data we build our own distributions.
   - Define density/Distribution function
   - Estimate Parameters
   - Ungrouped data: $X_1 \leq X_2 \leq X_3 \leq ... \leq X_n$

$$E(x) = \begin{cases} 0 & \text{for } x < X_1 \\ \frac{i-1}{n-1} + \frac{x - X_i}{(n-1)(X_{i+1} - X_i)} & \text{for } X_i \leq x < X_{i+1},\ i = 1, 2, ..., n-1 \\ 1 & \text{for } X_n \leq x \end{cases}$$

   - Grouped Data : $nX_j^{'}$s are grouped in k adjacent intervals so that the jth interval contains nj observations, $n_1 + n_2 + ... + n_k = n$
     Intervals: $(a_0, a_1), (a_1, a_2), ..., (a_{k-1}, a_k)$,
     $G(a_0) = 0, G(a_j) = \frac{n_1 + n_2 + n_3 + ... + n_j}{n}$

$$G(x) = \begin{cases} 0 & \text{for } x < a_0 \\ G(a_{j-1}) + \frac{x - a_{j-1}}{a_j - a_{j-1}}[G(a_j) - G(a_{j-1}] & \text{for } a_{j-1} \leq x < a_j,\ j = 1, 2, 3, ..., k \\ 1 & \text{for } a_k \leq x \end{cases}$$

3. Clues from summary statistics
   - Symmetric distributions: mean $\approx$ median, eg: Normal Distribution
   - Coefficient of Variation(cv): Ratio of Standard deviation & mean, $\frac{\sigma}{\mu}$
     Continuous Distributions: cv ≈ 1, eg: Exponential Distribution
     Right/Positive skewed histogram: cv > 1, eg: log normal distribution
   - Lexi$^{'}$s ratio: Same as cv for Discrete Distributions.
   - Skewness(v): Measure of symmetry of a distribution
     v = 0, Normal Distribution
     v > 0, right skewed(exponential distribution)
     v < 0, left skewed
4. Parameter Estimation
   - Once distribution is guessed, next step is estimating parameters of the distribution.
   - Most common method used is MLE.
5. Goodness of Fit
   - Can be checked by
     - Frequency Comparison(a bit technical)
     - Probability Comparison(Visual tool)
     - Goodness of Fit test(statistical test for goodness)
   - Quantile-Quantile Plot(Q-Q Plot)
     - Graph of q_i quantile of model vs q_i quantile of sample distribution.
     - $x_{q_i}^M = \hat{F}^{-1}(q_i)$
     - $x_{q_i}^S = \tilde{F}^{-1}(q_i) = x_i, i = 1, 2, 3, ...$
     - If our distribution is correct, then we will get a line with slope 1 and intercept 0 (Linear) & $x_{q_i}^M \approx x_{q_i}^S$
     - Amplifies difference between the tails of model distribution.
   - Probability-Probability Plot(P-P Plot)

  - Graph of model Probability $\hat{F}(X_i)$ vs Sample Probability $\tilde{F}_n(X_i)$
  - Valid for both Continuous and Discrete data sets.
  - I chosen distribution is correct then $\hat{F}(X_i) \approx \tilde{F}_n(X_i)$, the plot will be linear with slope 1 and intercept 0.
  - Amplifies differences between middle portion of the model distribution.
- Goodness of fit tests
  - Statistical Hypothesis test that is used to assess formally whether observations are independent samples from a particular distribution.
  - H_0: Observation are independent.
  - Chi-Square Test
    * Require frequency tables: Bins, Object Frequency, Expected frequency.
    * Calculate test statistic $\chi^2 = \frac{\Sigma(O_i - E_i)^2}{E_i}$
    * Compute p-value, if it is less than significant level(α) then reject H_0
    * Compute tabulated $\chi^2_{k-p-1,\alpha}$, if $\chi^2_{tabulated} < \chi^2_{computed}$ then reject H_0.
      p = number of parameters
      k = number of bins

3 Week 3

1. Ordinal Data: Categorical data which can be ordered.
2. Conditional Probability: $\frac{\text{Joint Probility}}{\text{Marginal Probability}}$
3. Conditional Probability can be compared using Joint Probability table(Contingency Table)
4. Baye$'$s Rule
   - Posterior Probability can be found using initial probability and additional information
   - $P(A \cap B) = P(A|B)P(B)$
   - $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$
5. Chi-Squared Test of Independence
6. Null Hypothesis H_0: Categorical Variables are independent.
7. Alternate Hypothesis H_1: Categorical variables are not independent.
8. Example Table:

City\Preferred Brand	Brand A	Brand B	Brand C	Total
Mumbai	279	73	225	577
Chennai	165	47	191	403
Total	444	120	416	980

   - Independent(Explanatory) Variable is City
   - Dependent(Response) Variable is Brand Preference
   - f_o: Observed Frequency(from Samples)
   - f_e: Expected Frequencies, if variables were independent = $\frac{\text{Row Total} \cdot \text{Column Total}}{\text{Overall Total}}$
   - degree of freedom = (number of rows - 1) $\cdot$ (number of columns - 1)

4 Week 4

1. Demand Response Curve
   - Properties: Non-Negative, Continuous & Differentiable and Generally Downwards Slopping
   - Price Sensitivity = $\frac{D(P_2) - D(P_1)}{P_2 - P_1}$
   - Demand Elasticity = $-\frac{\frac{D(P_2) - D(P_1)}{D(P_1)}}{\frac{P_2 - P_1}{P_1}}$
2. Linear Response Curve

- $D(P) = D_o - mP$
- Satiating Price $P_s = \frac{D_o}{m}$; $D(P_s) = 0$
- Demand at $P = 0$ is D_o
- Elasticity $\varepsilon = \frac{mP}{D_o - mP}$

3. Constant Elasticity Curve
   - $D(P) = cP^{-\varepsilon}$
   - $c =$ Demand when $P = 1$
   - Revenue R $= P \cdot D(P)$
4. Elasticity
   - $\varepsilon < 1$: Inelastic Product Demand, Increase Revenue $\implies$ Increase Price
   - $\varepsilon > 1$: Elastic Product Demand, Increase Revenue $\implies$ Decrease Price
5. Simple Linear Regression can be used to fir Linear curves
   - Loss/Error $= \frac{\Sigma(y-\hat{y})^2}{N}$
   - error term $e = (y - \hat{y})^2 \sim N(0, \sigma_e^2)$
   - error terms are independent, have equal variance and are nirmally distributed.