Shriansh
PAGE NO.
DATE:

BUSINESS ANALYTICS

(SEPTEMBER 2022 TERM)

BY GAGNEET KAUR

Date
Sept 10, 2022

Shrinidhi
PAGE NO.
DATE:

WEEK 1

INTRODUCTION TO DATA VISUALI-ZATION

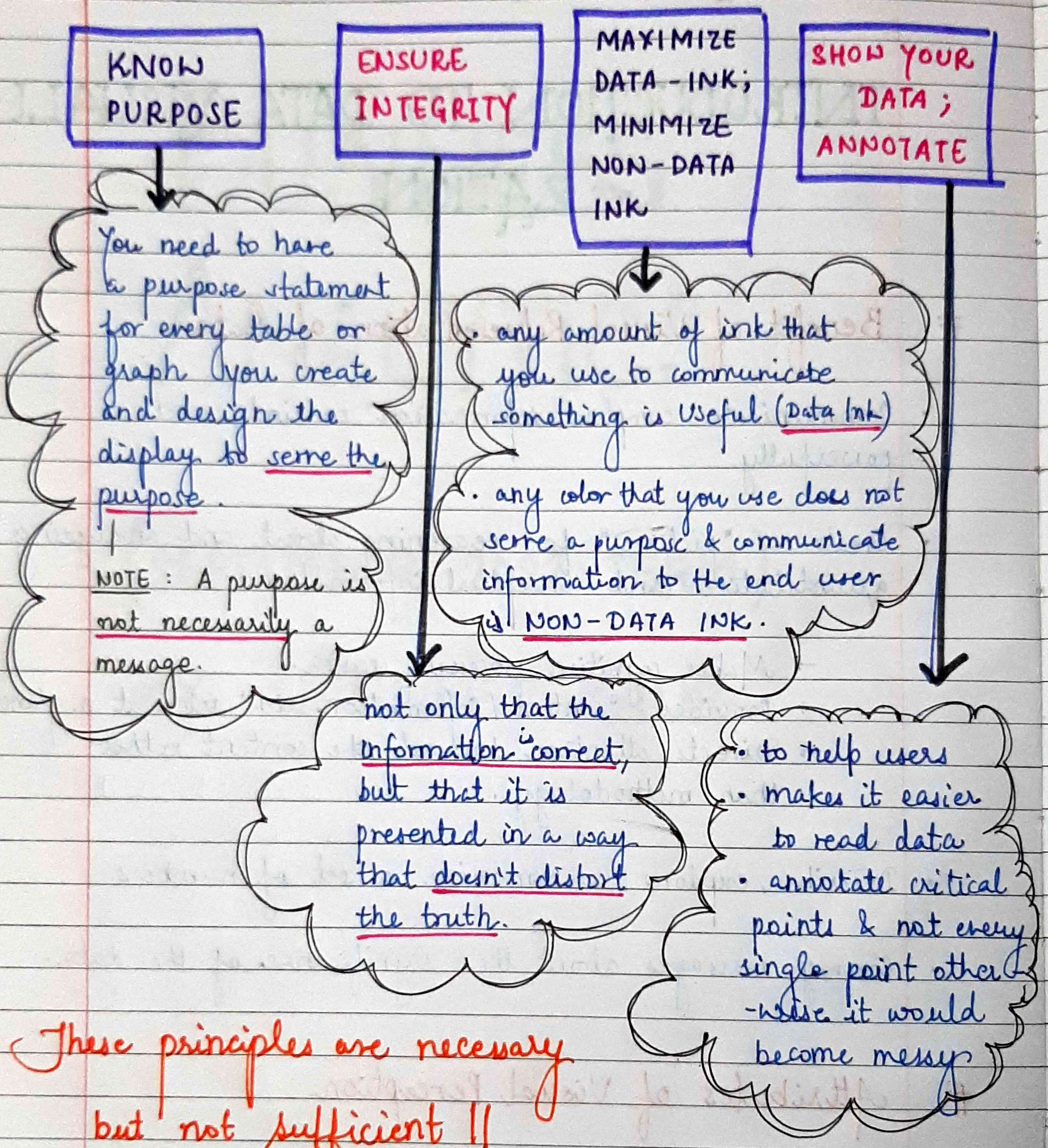
Benefit of Visual Representation of Data

- Communicate complex information concisely and powerfully
- Create a "picture" for reasoning about and analyzing quantitative and conceptual information.
  - → Makes cognitive processing easier.
  - → Provides "content / information rich" view at a glance
  - → Directs attention toward the content rather than methodology.
- Describe, explore, summarize a set of numbers
- Convey messages about the significance of the data.

Attributes of Visual Perception

→ Form → Colour → Spatial

Four "UMBRELLA" principles of effective visualization



These principles are necessary but not sufficient !!

Shrinath
PAGE NO.
DATE:

Executing Your Information Display Is A Three Step Process

(1) Defining Message

- What am I trying to communicate?
- What is the message?
- How do I make that message clear at a glance?

(2) Choosing Form

- Should I use text, table, graph or a diagram?
- Or a combination?

(3) Creating Designing

- What design principles lead to quick cognitive processing and effective communication of the message?

DaD
Sept 11, 2022

Design Of Visualization Process

DEFINING MESSAGE

- Depending on the objective, that is what needs to be emphasized.
- Depending on what is the message, what do I want to communicate, accordingly you will construct the display.
- Once the message is understood, choosing a form becomes apparent.

CHOOSING FORM

- Form becomes very self-evident because it is a combination of the message and the type of data. When you put these two together, it is very self-evident what type of display you want to choose.

Best Practices While Choosing Form

- one key decision to make is whether to display the data as a table or a chart.
- the whole idea is how do you communicate your message quickly without any loss in transmission.
- Choosing Table OR Chart ?

Shrinath
PAGE NO.
DATE:

Choose a Table

- Display complete data set.
- Focus on specific item (highlighted) in content of complete data set.
- Present wide range of data that is difficult to scale using a graph.
- Explain how numbers a derived - how results were calculated.

Choose A Chart

when you want to compare ; to show patterns

- Compare a slice of information
- Show "cause of change" (y-axis) vs. "effect" (x-axis).
- Show change over time.
- Show patterns of data distribution (normal curve)

Choose appropriate graph type for message :

If your message stresses...	then choose...
→ Components of one item	→ Pie Chart
→ Components of multiple items	→ 100% column / stacked column chart
→ Item comparison	→ Bar Chart
→ Change over time	→ Column / Line Chart
→ Frequency, Distribution	→ Histogram
→ Correlation	→ Paired bar, scatter dot

CREATING DESIGNS

Why is designing important?

→ Designing is important because once you have got a good message, got a good form, and the design can really completely derail the community message that you are trying to convey.

Best Practices - Creating Design

- Avoid 3-D effects (It has zero purpose)
- Avoid legends ; consider using labels
- Avoid contrasting borders around objects
- Use annotations to highlight key data changes or to focus on specific data points.

Graphs : Best Practices Lead to Uncluttered Look

- Design graph to support message ; consider using talking head here
- Use minimal grid lines, ideally none
- Use thin lines, thin axes, thin bars, thin arrows to show trends
- Display subtle but visible data point marks - only enough to show trend

Shrinath
PAGE NO.
DATE:

- Use minimal tick marks to display scale ; usually just min / max on Y-axis.
- Label axes ; label graphic items
- Opt for value labels wherever possible ; delete some to avoid clutter.

DASHBOARDS

Visualizing On Dashboards

→ What are data dashboards ?
→ How do you decide what should go on a dashboard ?
→ Domain Specific ?
→ What are generic principles ?
→ Do only descriptives go on the dashboard ?

Definition

→ A visual display of the most important information needed to achieve one or more objectives that has been consolidated on a single screen so it can be monitored and understood at a glance.

Basic Dashboard Principles

1. Scan the big picture.
2. Zoom in on important specifics.
3. Link to supporting details.

→ TUTORIALS ←

MISLEADING VISUALS

1. Misleading Axes
   ↳ y-axis not starting from 0
2. Wrong Data
   ↳ for example : sum of shares of pie chart doesn't total to 100%
3. Inappropriate chart type
4. Incorrect Bar Chart
5. Misleading 3-D Pie Chart

Date
Sept 12, 2022

Monday

PAGE NO.
DATE:

WEEK 2

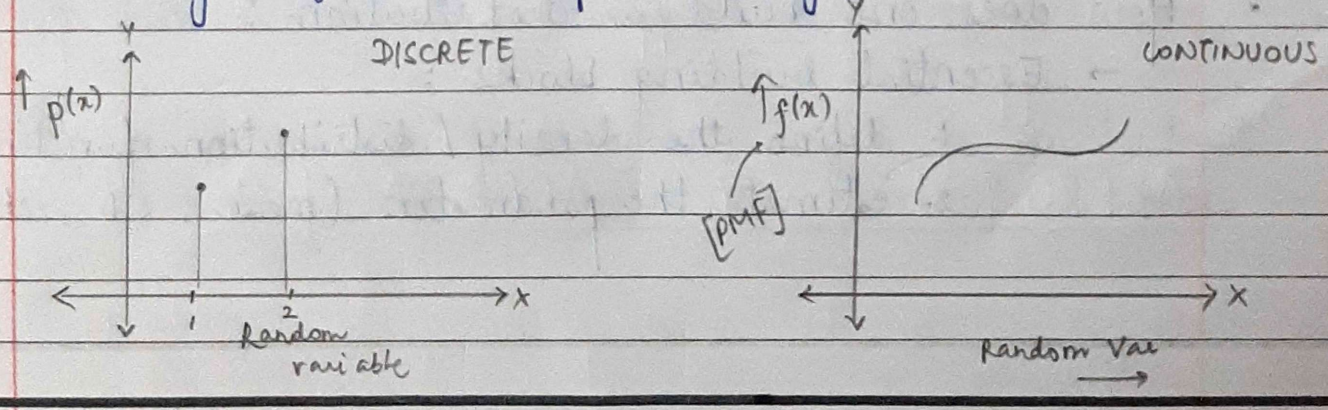
Probability Distributions

PROBABILITY DISTRIBUTIONS

- A statistical model that shows possible outcomes of a particular event or course of action as well as the statistical likelihood of each event.

- Probability distributions for a DISCRETE RANDOM VARIABLE may look like all the possible values of the RV along with the corresponding probabilities that the random variable will take on that particular value.

- For a continuous distributions, it is generally represented by a density function.
For density we need a small interval to actually define some probability.



Theoretical Distributions : normal dis., uniform dist., binomial dist., exponential dist., etc.

PAGE NO.
DATE:

HOW DO WE GO ABOUT USING DATA

→ How do we use the collected business data (sales, volume, loan defaulters, salary hikes in an organization, etc.)?

1. The data values themselves are used directly in the simulation. This is called TRACE-DRIVEN SIMULATION.

2. 'Fit' a theoretical distribution to the data (and check whether that 'fit' is good!)

3. The data values could be used to define an Empirical Distribution function in some way.

WHAT ARE EMPIRICAL DISTRIBUTIONS

- Using the data, we build our own distributions.
  → So here we are not fitting a distribution to the data, we are actually building a distribution from the data the we collected.

- How does one build a distribution?
  → Essential building blocks :
  1. define the density / distribution functions.
  2. estimate the parameters (mean, SD, etc.)

HOW TO BUILD EMP. DISTRIBUTIONS?

EXAMPLES

(1) For Ungrouped Data

Let $X_{(i)}$ denote the i_{th} smallest of the X_j's so that:

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \ldots\ldots \leq X_{(n)}$$

$$F(x) = \begin{cases} 0 & , \text{if } x < X_{(1)} \\ \dfrac{i-1}{n-1} + \dfrac{x - X_{(i)}}{(n-1)\left(X_{(i+1)} - X_{(i)}\right)} & , \text{if } X_{(i)} \leq x \leq X_{(i+1)} \text{ for } i = 1, 2, \ldots, n-1 \\ 1 & , \text{if } X_{(n)} \leq x \end{cases}$$

(2) For Grouped Data

Suppose that n X_j's are grouped in k adjacent intervals $[a_0, a_1), [a_1, a_2), \ldots [a_{k-1}, a_k)$ so that the j^{th} interval contains n_j observations. $n_1 + n_2 + \ldots + n_k = n$

Let a piecewise linear function G be such that $G(a_0) = 0$, $G(a_j) = \dfrac{(n_1 + n_2 + \ldots + n_j)}{n}$, then:

$$G(x) = \begin{cases} 0 & , \text{if } x < a_0 \\ G(a_{j-1}) + \dfrac{x - a_{j-1}}{a_j - a_{j-1}} \left[G(a_j) - G(a_{j-1}) \right] & , \text{if } a_{j-1} \le x < a_j, \\ & \quad j = 1, 2, \ldots, k \\ 1 & , \text{if } a_k \le x \end{cases}$$

The Three Approaches : COMPARISON

→ Approach 1 is used to validate simulation model when comparing model output for an existing system with the corresponding output for the system itself.

→ Drawbacks of Approach 1
- Simulation can only reproduce only what happened historically
- There is seldom enough data to make all simulations runs.

→ Approach 2 and 3 avoid these shortcomings so that any value between minimum & maximum can be generated. So approaches 2 & 3 are preferred over approach 1.

→ If theoretical distributions can be found that fits the observed data (approach 2), then its preferred over approach 3.

APPROACH 3 vs APPROACH 2

- Empirical distribution may have some irregularities if

small no. of data points are available. Approach 2 smoothens out the data and may provide information on the overall underlying distribution.

- In approach 3, it is usually not possible to generate values outside the range of observed data in the simulation.

- If one wants to test the performance of the simulated system under extreme conditions, that cannot be done using approach 3.

- There may be compelling (physical) reasons in some situations for using a particular theoretical distribution. In that case too, it is better to get empirical support for that distribution from the observed data.

Date: September 13, 2022

TUESDAY

Guessing The Distribution

CLUES FROM SUMMARY STATISTICS

- For the symmetric distributions mean and median should match. In the sample data, if these values are sufficiently close to each other, we can think of a symmetric distribution (e.g. NORMAL)
- Coefficient of Variation (CV) { std dev / mean }
  - → for continuous distributions
  - → The CV = 1 → Exponential Distribution
  - → If CV > 1 (slightly right skewed curve) → Lognormal

Note: for many distributions CV may not even be properly defined. When?

- Lexis Ratio: same as CV for discrete distributions
- Skewness (v): measure of symmetry of a distribution.
  - → v = 0 – Normal Dist.
  - → v > 0 – Distribution skewed towards RIGHT (v=2 exp. dist.)
  - → v < 0 – LEFT skewed

PARAMETER ESTIMATION

- Once distribution is guessed, the next step is estimating the parameters of the distribution.
- Each distribution has a set of parameters.
  - → Normal Distribution has mean & standard deviation
  - → Exponential Distribution has a "λ".

- Most common method of parameter estimation : MLE

GOODNESS - OF - FIT

- For the input data we have, we have assumed a probability distribution.
- We also have estimated the parameters for the same.
- How do we know this fitted distribution is "good enough"?
- It can be checked by several methods :
  1. Frequency comparison (a bit technical)
  2. Probability plots (visual tool)
  3. Goodness-of-fit tests (statistical test of goodness. very widely used)

PROBABILITY PLOTS

Q-Q plot : Quantile - quantile plot

- Graph of the q_i - quantile of a fitted (model) distribution versus the q_i - quantile of the sample dist.

$$x^M_{q_i} = \hat{F}^{-1}(q_i)$$

$$x^S_{q_i} = \tilde{F}_n^{-1}(q_i) = X_{(i)}, \quad i = 1, 2, \ldots, n$$

- If $\hat{F}(x)$ is the correct distribution that is fitted, for a large sample size, then $\hat{F}(x)$ and $F_n(x)$ will be close together and the Q-Q plot will be approximately linear with intercept 0 and slope 1.
- For small sample, even if $\hat{F}(x)$ is the correct distribution, there will some departure from the straight line.

P-P plot : Probability - probability plots

- A graph of the model probability $\hat{F}(X_{(i)})$ against the sample probability $\tilde{F}_n(X_{(i)}) = q_i$, $i = 1, 2, \ldots, n$
- It is valid for both continuous as well as discrete data sets.
- If $\hat{F}(x)$ is the correct distribution that is fitted, for a large sample size, then $\hat{F}(x)$ and $F_n(x)$ will be close together and the P-P plot will be approximately linear with intercept 0 and slope 1.

Conclusion

- The Q-Q plot will amplify the differences between the tails of the model distribution and the sample distribution.

Shrinath
PAGE NO.
DATE:

- Whereas, the P-P plot will amplify the differences at the middle portion of the model & sample dist.

GOODNESS-OF-FIT TESTS

- A goodness-of-fit test is a statistical hypothesis test that is used to assess formally whether the observations $X_1, X_2, X_3, \ldots, X_n$ are an independent sample from a particular distribution with function F^*.

H_0 : The X_i's are IID random variables with distribution function F^*.

- Two famous tests :
  1. Chi-square test
  2. Kolmogorov - ~~Sim~~ Smirnov test

CHI-SQUARE TEST

- Applicable for both, continuous as well as discrete, distributions.

- Method of calculating chi-square test statistic :

(1) Divide the entire range of fitted distribution into k adjacent intervals -- $[a_0, a_1], [a_1, a_2), \ldots, [a_{k-1}, a_k)$, where it could that $a_0 = -\infty$ in which case the first interval is $(-\infty, a_1)$ & $a_k = \infty$.

N_j = # of X_i's in the j^{th} interval $[a_{j-1}, a_j)$, $j = 1, 2, \ldots, n$

(2) Next, we compute the expected proportion of X_i's that would fall in the j^{th} interval if we were sampling from fitted distribution.

- For <u>continuous</u> distributions : $p_j = \int_{a_{j-1}}^{a_j} \hat{f}(x)\, dx$
- For <u>discrete</u> distributions : $p_j = \sum_{a_{j-1} \le x_j < a_j} \hat{p}(x_j)$
- Finally the test statistic is calculated as :

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

- This calculated value of the test statistic is compared with the tabulated value of chi-square distribution with $k-1$ df at $1-\alpha$ level of significance.

If $\chi^2 > \chi^2_{k-1, 1-\alpha}$ Reject H_0

If $\chi^2 \le \chi^2_{k-1, 1-\alpha}$ Do not Reject H_0