Data Mining for Business

Data Pre-processing

Dr. Shipra Maurya
Department of Management Studies
IIT (ISM) Dhanbad
Email: shipra@iitism.ac.in

Data Collection

- Data collection is the methodological process of gathering information about a specific subject
- Data should be collected legally and ethically
- Sources of data:
  - Primary
  - Secondary
- Three types of consumer data:
  - First-party data
  - Second-party data
  - Third-party data

Factors to be considered before Data Collection

- Define the business problem that you are trying to solve
- The data subject you need to collect data from
- Collection Timeframe
- The data collection method best suited to your needs – applicable in case of primary data
- Company’s budget

Data Collection Methods

- Survey
- Transactional Tracking
- Interviews and Focus group
- Observation – people interacting with the website
- Online Tracking
- Social Media tracking
- Databases – Example – Bloomberg, Prowess, RBI Statistics database etc.

Data Integration

- Data integration is a data pre-processing technique that combines data from different sources to create a unified view of data for analytical purposes

- **Major approaches of Data Integration:**
  - **Tight coupling** – data warehouse is the ultimate source of collecting data
  - **Loose coupling** – data lies in source databases and an interface is used to fetch the data from source systems

Issues/Challenges in Data Integration

- **Schema Integration** – Integrate data from multiple sources
- **Redundancy** - An attribute may be redundant if it can be derived or obtained from another attribute or set of attributes
- **Detection and Resolution of data value conflicts** - Attribute values from different sources may differ for the same real-world entity

Data Cleaning

"Garbage In, Garbage Out".

- Data cleaning or Data scrubbing or Data cleansing
- Data cleaning is the process of preparing the dataset for analysis by weeding out the incorrect or irrelevant information

Based on a Research study

Image Source: Internet

Characteristics of Quality Data

- Accuracy – checking the feasibility of the data. Eg., location and pin code; height, weight of the person etc.
- Completeness – no missing values
- Consistency - how the data responds to cross-checks with other fields
- Validity - Typical constraints applied on online forms to collect data are:
  - Data-type constraints
  - Range constraints
  - Unique constraints
  - Cross-field validation

Data Cleaning Steps

- **Remove duplicates or irrelevant observations –**
  - Duplicate observations are generally caused by joining data from different tables
  - Irrelevant observations are those that are not required to address the problem at hand

- **Fix structural errors –**
  - typos in the name of features, the same attribute with a different name, mislabeled classes, i.e. separate classes that should really be the same, or inconsistent capitalization.
  - For example, you may find “N/A” and “Not Applicable” both appear, but they should be analyzed as the same category

Data Cleaning Steps

- Manage unwanted outliers
- Handle missing data
- Validate data accuracy

Handling Missing Values

- Why handling missing values is important?
- How to handle missing values?
  - Delete rows with missing values
  - Impute missing values for continuous variables
  - Impute missing values for categorical variables
  - Other Imputation methods such as Last Observation Carried Forward (LOCF), Interpolation of the variable before and after a timestamp
  - Use algorithms that support missing values such as k-nearest neighbors, Naïve Bayes, XGBoost etc.
  - Prediction of missing values using Regression or Classification models
  - Imputation using Deep Learning Library - Datawig

Delete rows with Missing values

```
print(data.isnull().sum())
print(data.shape)
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
(891, 12)
```

With Null values

```
data.dropna(inplace=True)
print(data.isnull().sum())
print(data.shape)
```

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
(183, 12)
```

Without Null values

Note: Titanic Dataset is used

Impute missing values for Continuous variables

```
[10] data["Age"][:20]

0      22.0
1      38.0
2      26.0
3      35.0
4      35.0
5       NaN
6      54.0
7       2.0
8      27.0
9      14.0
10      4.0
11     58.0
12     20.0
13     39.0
14     14.0
15     55.0
16      2.0
17      NaN
18     31.0
19      NaN
Name: Age, dtype: float64
```

With Null values

```
data["Age"] = data["Age"].replace(np.NaN, data["Age"].mean())
print(data["Age"][:20])

0     22.000000
1     38.000000
2     26.000000
3     35.000000
4     35.000000
5     29.699118
6     54.000000
7      2.000000
8     27.000000
9     14.000000
10     4.000000
11    58.000000
12    20.000000
13    39.000000
14    14.000000
15    55.000000
16     2.000000
17    29.699118
18    31.000000
19    29.699118
Name: Age, dtype: float64
```

Without Null values

Impute missing values for Categorical variables

```
[12] data.isnull().sum()

PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

With Null values

```
[13] data["Cabin"] = data["Cabin"].fillna('U')

[14] data.isnull().sum()

PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       2
dtype: int64
```

Without Null values

Target Leakage

- It happens when you train your algorithm on a dataset that includes information that would not be available at the time of prediction when you apply that model to data you collect in the future.

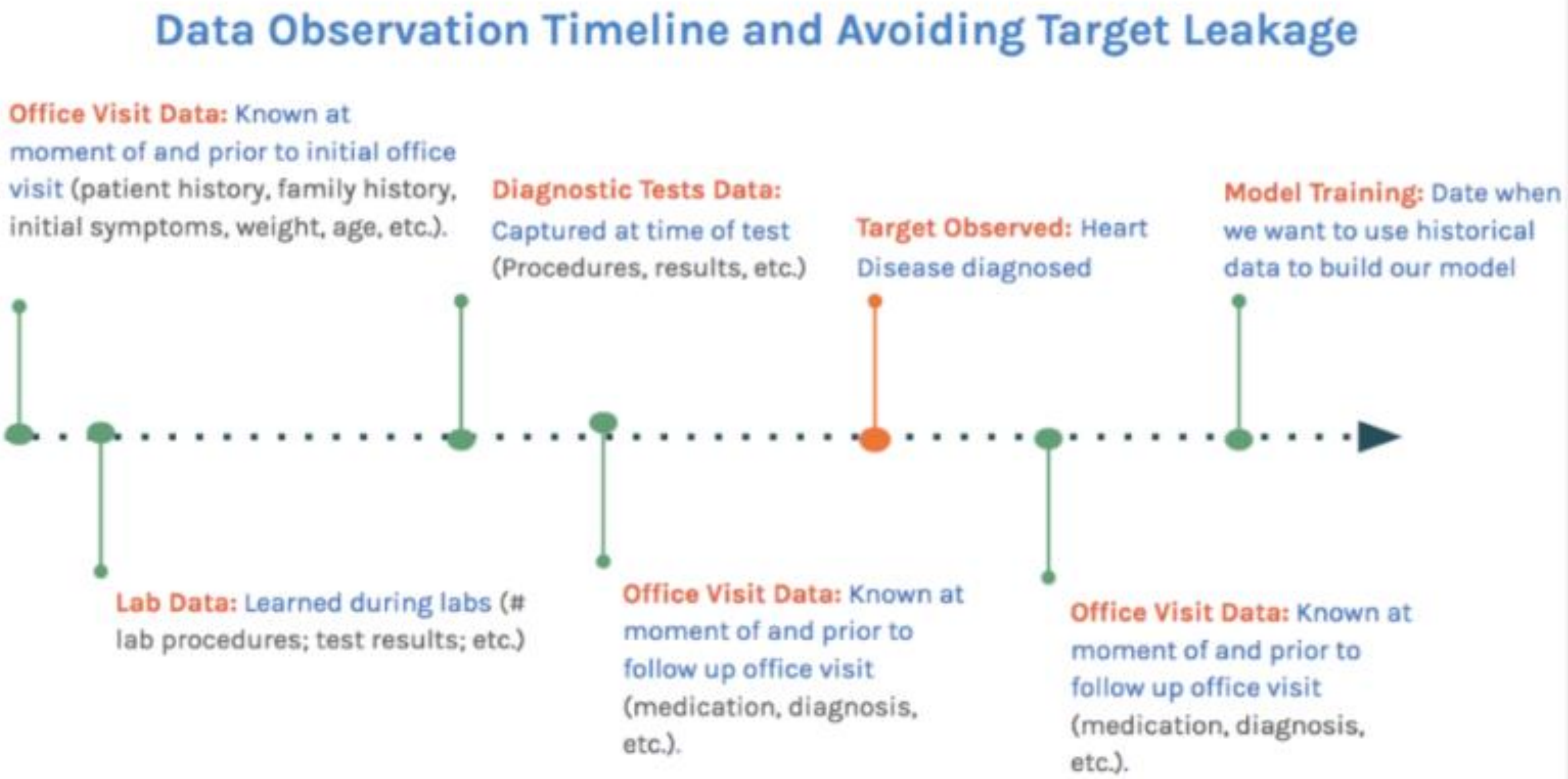


Image Source: Internet

Why is Target leakage important?

- It causes a model to overrepresent its generalization error, which makes it useless for any real-world application

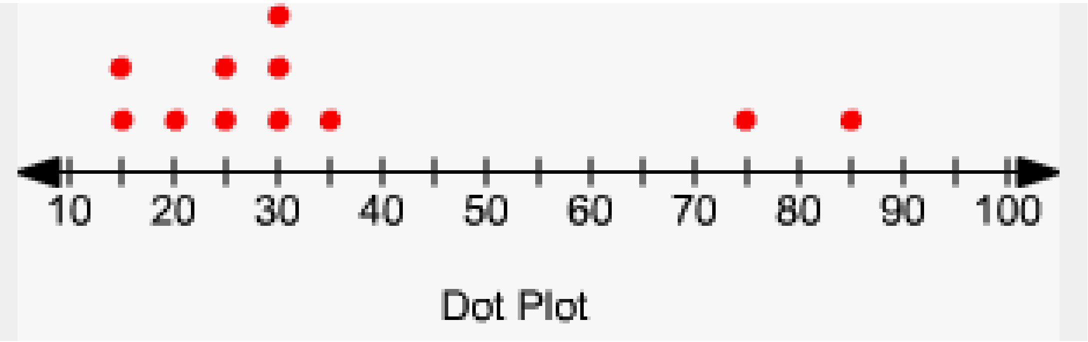
How to identify Target leakage?

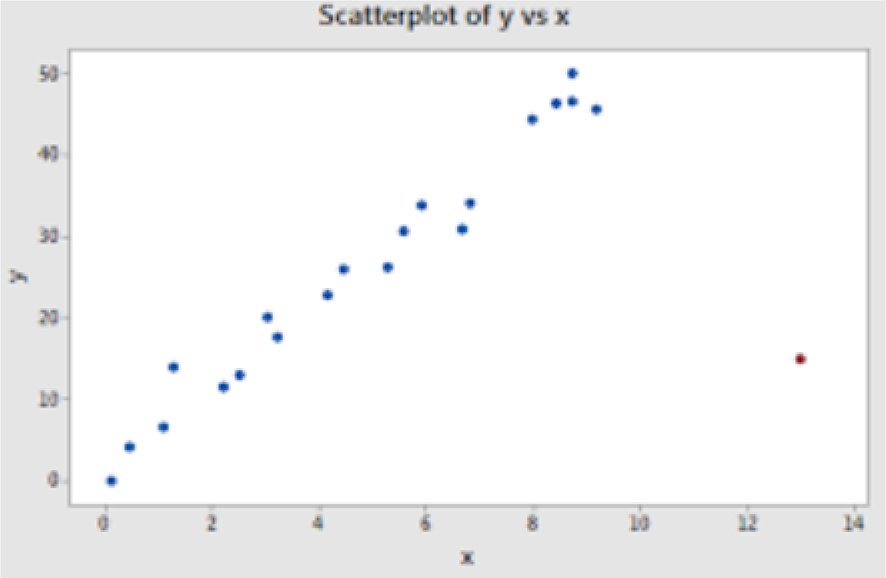
- **Accuracy score of a model** – a near-perfect accuracy score for a model is a red flag
- **Feature Impact –** if a feature has a very high score, i.e., it will have more impact on target variable compared to other features. There is a high probability that target leakage is present in the dataset

Any observations from the Image?

Image Source: Internet

Any observations from the Image?





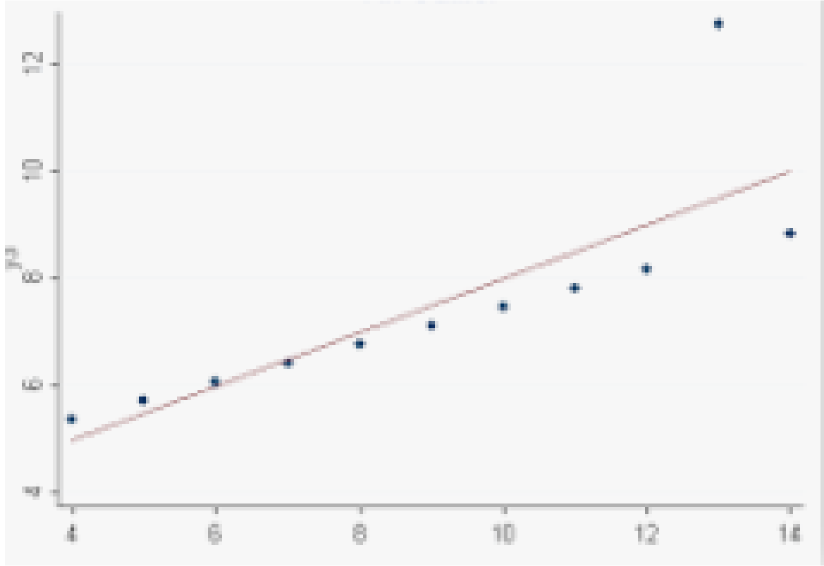


Image Source: Internet

Outliers

- Outliers are extreme values that deviate from other observations in the dataset
- **Causes of Outliers:**
  - Data entry errors (human made errors)
  - Data processing errors (data manipulation related errors)
  - Experimental errors (experiment planning/execution errors)
  - Natural or novelties in data (example – crude oil price decline in May 2020)

Types of Outliers

- Univariate outliers
- Multivariate outliers

How to detect Outliers?

- **Visualization -**
  - Scatterplot
  - Box plot
  - Histogram

- **Data descriptive statistics**

- **Z-score** – helps to understand how far the data point is from the mean

$$z - score = \frac{Data\,point\ - mean}{Standard\ deviation}$$

- **IQR (Inter-Quartile Range) -**

$$IQR = Q3 - Q1$$

Upper = Q3 + 1.5 * IQR

Lower = Q1 – 1.5 * IQR

How to treat Outliers?

- Retain
- Trimming -data points are removed
- IQR Score - data points are removed
- Replacing outliers with Median values – can we use mean? - data points are modified
- Winsorization - Flooring and Ceiling - data points are modified, not trimmed or removed
- Log transformation

Data Visualization

- Data visualization is the process of creating graphical representations of information
- **Data Visualization Techniques:**
  - Pie Chart
  - Bar Chart
  - Histogram
  - Gantt Chart
  - Heat Map
  - Box and Whisker Plot
  - Waterfall Chart
  - Area Chart
  - Scatter Plot
  - Frequency Table
  - Pictogram Chart
  - Timeline
  - Highlight Table
  - Bullet Graph
  - Choropleth Map
  - Word Cloud
  - Network Diagram
  - Correlation Matrices
  - Cross-tab

Assignment

Frequency Distribution Table

- Frequency is how often something has happened
- Frequency distribution indicates how the frequency is distributed over values
- Example - Plot the frequency distribution of Virat Kohli's T20I games' scores

Crosstabs or Contingency Tables

- special type of frequency distribution tables
- powerful tool for comparing two or more categorical variables
- Example - Titanic Data - find out relationship between survivorship and gender

Exploratory Data Analysis Steps

- Handling missing values
- Handling Outliers
- Data visualization
- Correlation Analysis
- Storytelling from the data

Exploratory Data Analysis

- Multivariate Analysis - Correlation Analysis
  - **Continuous vs. Continuous Variables -**
    - Pearson Correlation Matrix (-1 to +1) - Parametric
    - Spearman Rank Correlation (-1 to +1) - Non-parametric
  - **Nominal vs. Nominal Variables –**
    - Cramer’s V (0 to +1) – uses Chi-square statistics – Non-parametric
  - **Ordinal vs. Ordinal Variables –**
    - Spearman Rank Correlation (-1 to +1) - Non-parametric
    - Cramer’s V (0 to +1) – uses Chi-square statistics -Non-parametric

Exploratory Data Analysis

- Multivariate Analysis - Correlation Analysis…
  - **Continuous vs. Ordinal Variables -**
    - Spearman Rank Correlation - (-1 to +1) – Non-parametric
  - **Nominal vs. Continuous Variables –**
    - Point Biserial (-1 to +1) - Parametric
    - Kruskal-Wallis H Test- Non-parametric
  - **Nominal vs. Ordinal Variables –**
    - Rank Biserial Correlation Coefficient (-1 to +1) – Non-parametric

Correlation Coefficient

- Correlation is a bivariate analysis
- Measures the strength of association and direction of the relationship between the two variables
- It does not indicate the causality i.e., it does not define the variables as dependent and independent and hence correlation is different than regression

Pearson's Correlation Coefficient

- A parametric method that measures the covariance of the two variables divided by the product of their standard deviations. Used for two continuous variables. Ranges between -1 to +1.

$$= \frac{Covariance\ between\ x\ and\ y}{Std\ dev\ of\ x * Std\ dev\ of\ y}$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Image Source: Internet

Pearson's Correlation Coefficient Example

Subject	Age x	Glucose Level y	xy	x^2	y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

$$6(20{,}485) – (247 \times 486) / [\sqrt{[[6(11{,}409) – (247^2)] \times [6(40{,}022) – 486^2]]}$$

$$= 0.5298$$

Image Source: Internet

Spearman Rank Correlation Coefficient

- A non-parametric method that measures the association between two variables (continuous and ordinal). Ranges between -1 to +1.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Physics	Rank	Math	Rank	d	d squared
35	3	30	5	2	4
23	5	33	3	2	4
47	1	45	2	1	1
17	6	23	6	0	0
10	7	8	8	1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0

= 1 – (6*12)/(9(81-1))

= 1 – 72/720

= 1-0.1

= 0.9

Image Source: Internet

Cramer’s V

- Indicates the strength of association between **two categorical variables** (there must be two or more unique values in each variable). Ranges between 0 to +1.

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

- ϕ_c denotes Cramér’s V;*
- χ^2 is the Pearson chi-square statistic from the aforementioned test;
- N is the sample size involved in the test and
- k is the lesser number of categories of either variable.

Image Source: Internet

Chi-Square Test

- For continuous variable significance – z-test and t-test are used
- For categorical variable significance – Chi-square test is used
- It is a test of statistical significance for categorical variables
- Types of Chi-square test
  - Chi-square goodness-of-fit test
  - Chi-square test of association

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$\chi 2$ = Chi-Square value

O_i = Observed frequency

E_i = Expected frequency

Image Source: Internet

Chi-Square Goodness-of-fit Test

- A non-parametric test used to identify statistical difference between observed and expected value
- **Example** – relationship between student CGPA and number of students placed. The researcher will be interested in identifying whether or not observed frequencies of placed students are equally distributed for different categories of CGPA.

Contingency Table

Number of Students Placed	CGPA					
	Below 6	**6-7**	**7-8**	**8-9**	**9-10**	**Total**
Observed Frequency	5	10	20	35	30	100
Expected Frequency	20	20	20	20	20	100

Image Source: Internet

Chi-Square Goodness-of-fit Test Example (1/2)

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Number of Students Placed	CGPA					
	Below 6	**6-7**	**7-8**	**8-9**	**9-10**	**Total**
Observed Frequency (O)	5	10	20	35	30	100
Expected Frequency (E)	20	20	20	20	20	100
(Oi – Ei)^2 :(Step 1)	225	100	0	225	100	
Step 1 / Ei :(Step 2)	11.25	5	0	11.25	5	32.5

Chi-square test statistic is 32.5

Null hypo.: There is no difference between observed and expected value

Alternate hypo.: There is difference between observed and expected value

Image Source: Internet

Chi-Square Goodness-of-fit Test Example(2/2)

- Degree of freedom : (k-1) : (5-1) = 4
- For 4 DOF, find out the 5% level of significance (alpha) in the table
- The critical value is 9.488
- Calculated chi-square test statistic is 32.5
- Since TS (32.5) >CV (9.488) hence reject the null which means that students CGPA are related with their placement

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Image Source: Internet

Chi-Square Test of Association

- Used to understand if there any relationship between two categorical independent variables (features)

	Boys	Girls	Total	Expected Frequency
Pass	17 (0.1216)	20 (0.1216)	**37**	(37*25)/50 = 18.5
Fail	8 (0.3461)	5 (0.3461)	**13**	(13*25)/50 = 6.5
Total	**25**	**25**	**50**	
Note – values in parenthesis are chi-square values				

Sum of Chi-square values =0.9354; DOF = (No. of rows – 1) * (No. of Column – 1) = (1-1)*(1-1) = 1

Critical value : 3.84; TS < CV hence we fail to reject the null hypothesis

Null hypo: The two categorical variables are independent

Alternate hypo: The two categorical variables are not independent

Image Source: Internet

Point Biserial Correlation

- Indicates the strength of association between two variables where one is **continuous** and another is **binary**. Ranges between -1 to +1. It assumes continuous variable follows normal distribution.

$$r_{pb} = \frac{M_1 - M_0}{s_n}\sqrt{pq}$$

- M_1 = mean (for the entire test) of the group that received the positive binary variable (i.e. the "1").
- M_0 = mean (for the entire test) of the group that received the negative binary variable (i.e. the "0").
- S_n = standard deviation for the entire test.
- p = Proportion of cases in the "0" group.
- q = Proportion of cases in the "1" group.

Image Source: Internet

Rank Biserial Correlation

- A non-parametric test that Indicates the strength of association between two variables where one variable is **nominal** and another is **ordinal.**

$$r_{rb} = 2 * (Y_1 - Y_0) / n.$$

Where:

- n = number of data pairs in the sample,
- Y_0 = Y score means for data pairs with $x = 0$,
- Y_1 = Y score means for data pairs with $x = 1$.

Image Source: Internet

Rank Biserial Correlation Example

For example, let's say you had the following data:

Dichotomous variable: 1,1,1,0,1

Ordinal variable: 3,1,5,4,2

Y0 = 4 (only one ordinal variable is paired with 0).

Y1 = 3+1+5+2/4 = 11/4 = 2.75

n = 5

Giving a rank-biserial correlation coefficient of: 2 * (2.75 – 4)/6 = -0.21.

Kruskal Wallis H test

- It is a non-parametric test
- The test determines whether there is any statistically significant difference between the two variables (**Continuous and Nominal**)
- Test statistic called H-statistic is calculated and compared against critical value

H_0: population medians are equal.

H_1: population medians are not equal.

n = sum of sample sizes for all samples,

c = number of samples,

T_j = sum of ranks in the j^{th} sample,

n_j = size of the j^{th} sample.

$$H = \left[\frac{12}{n(n+1)} \sum_{j=1}^{c} \frac{T_j^2}{n_j} \right] - 3(n+1)$$

Image Source: Internet

Kruskal Wallis H test Example

- A shoe company wants to know if three groups of workers have different salaries:

  Women: 23K, 41K, 54K, 66K, 78K.

  Men: 45K, 55K, 60K, 70K, 72K

  Minorities: 20K, 30K, 34K, 40K, 44K.

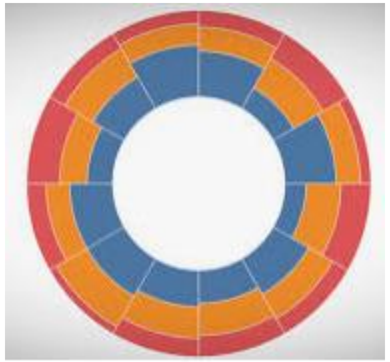
- Step 1 - sort the data for all the groups together and assign rank
- Step 2 – Add the ranks for each group
- Step 3 – Calculate H-Statistic

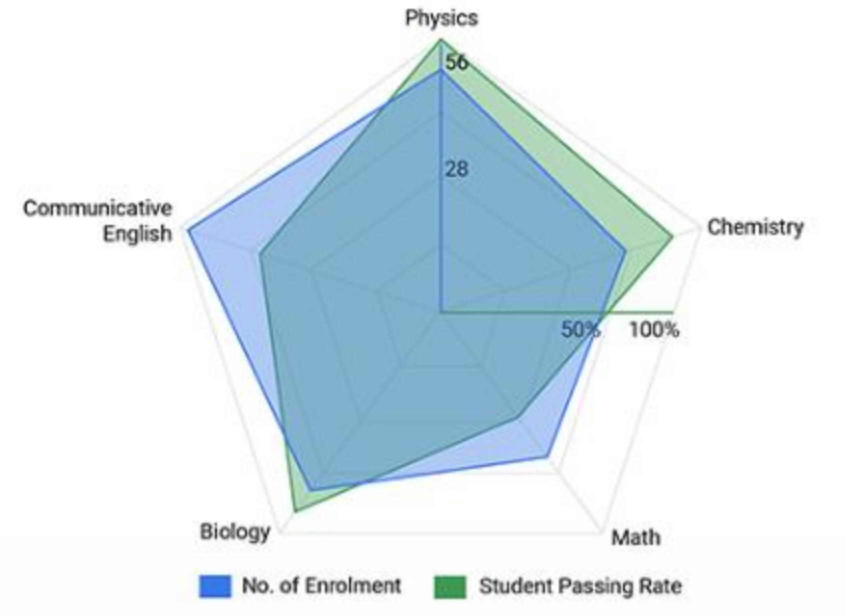
$$H = \left[\frac{12}{15(15+1)}\right]\left[\frac{44^2}{5} + \frac{56^2}{5} + \frac{20^2}{5}\right] - 3(15+1)$$

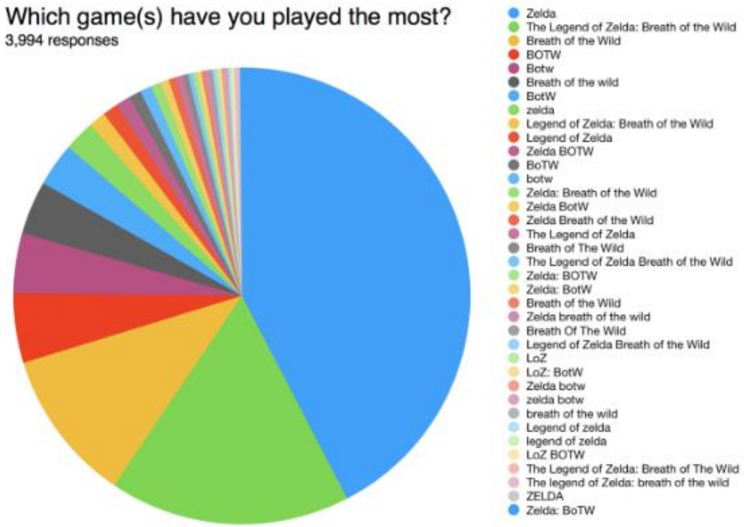
- Step 4 – Find the chi-square value (critical value) using alpha = 0.05 and degree of freedom as k-1
- Step 5 – compare the value of step 3 and 4
- Step 6 - If the critical value (5.9915) < H-Statistic (6.72) – Reject the null and vice-versa

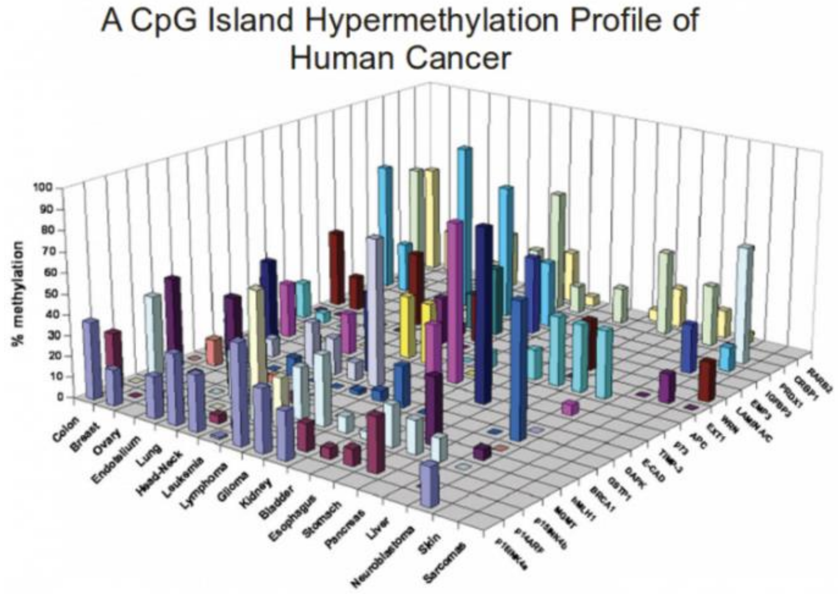
Image Source: Internet

Examples of bad data visualization









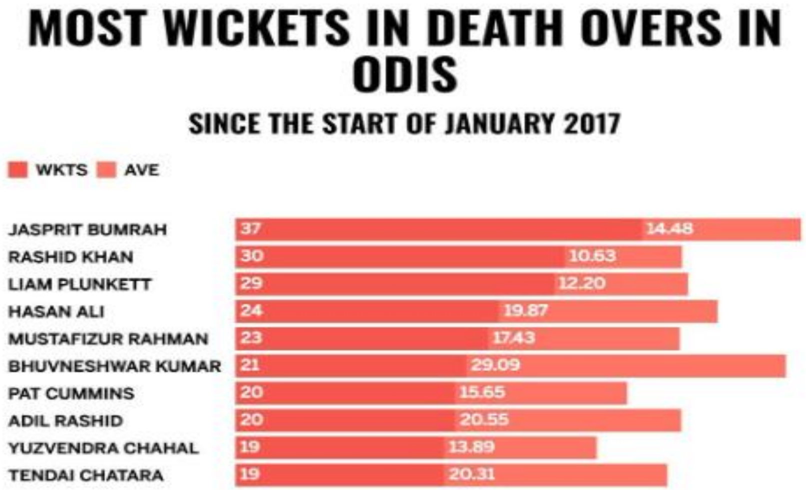
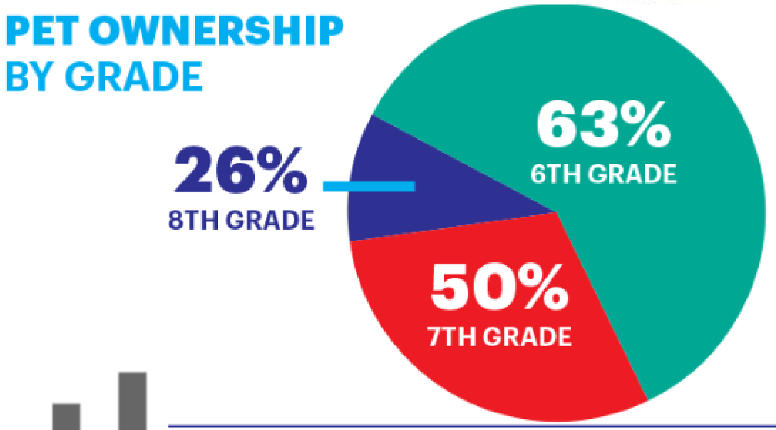
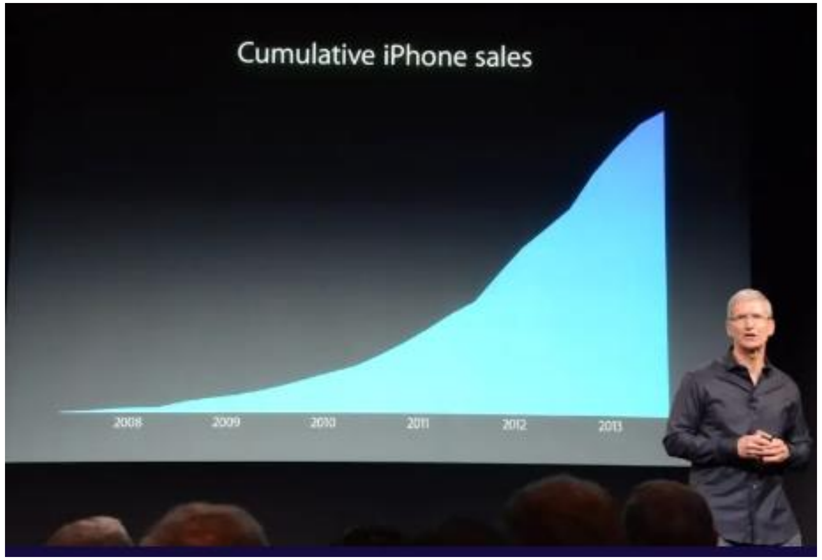
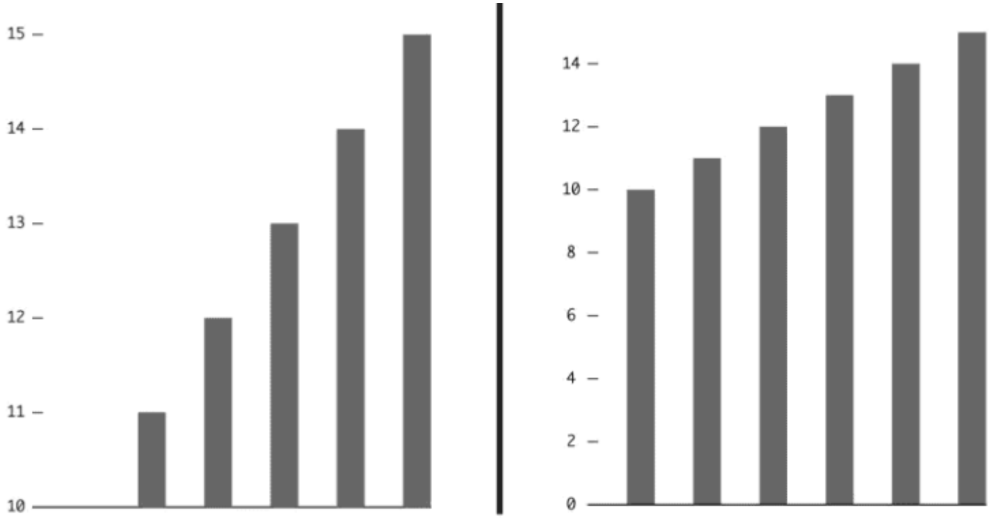


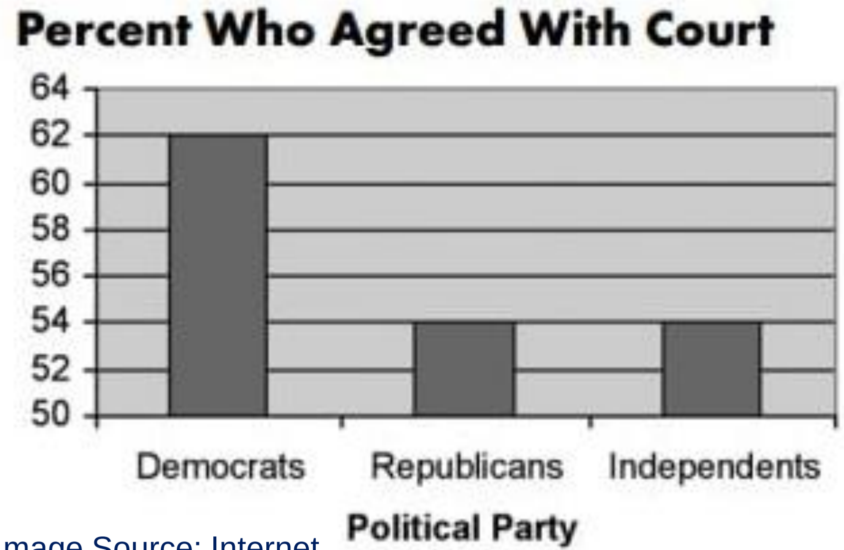
Image Source: Internet

Examples of Misleading data visualization









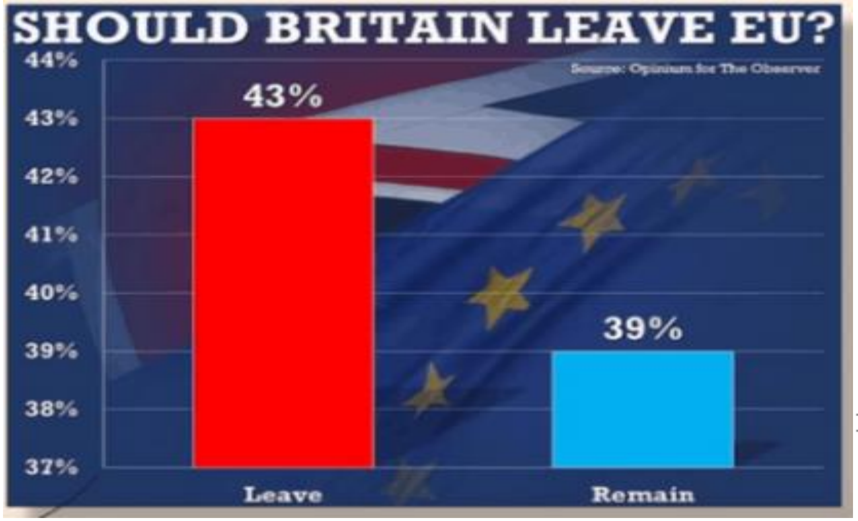


Image Source: Internet

Mistakes to be avoided in Data Visualization

- Using the wrong type of chart
- Including too many variables
- Using inconsistent Scales
- Poor Colour choices
  - Using too many colors
  - Using familiar colors in surprising ways
  - Using colors with little contrast
  - Not accounting for viewers who may be colorblind

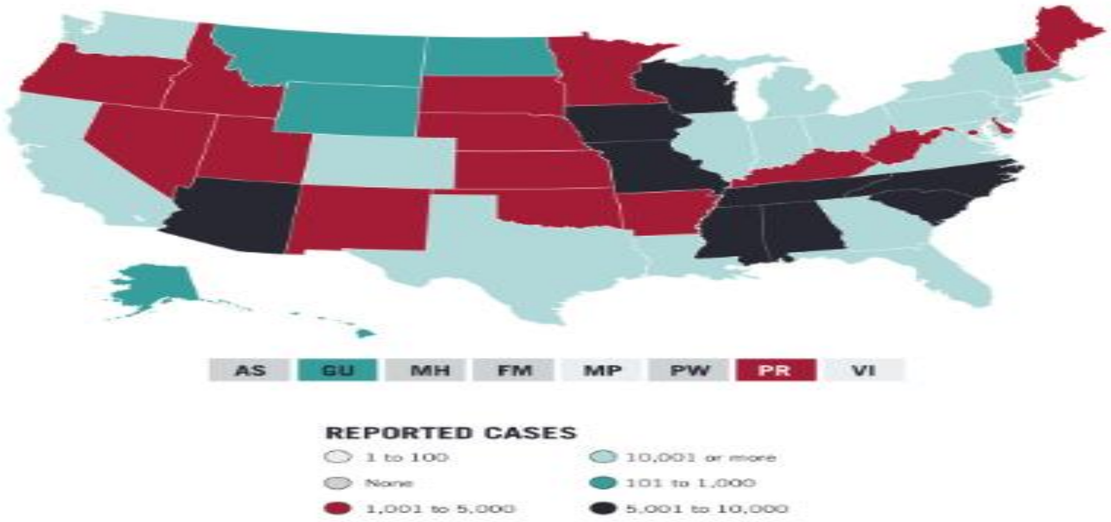


Image Source: Internet

Storytelling with Data

- Understand the problem
    - What are you solving for?
    - Who is your audience?
- Nature of Solution
    - Identify what does a successful solution look like?
- Choose an effective visual
- Eliminate clutter – empathize with the audience
- Think like a designer
- Tell a story

Storytelling with Data Example

- **Business Problem:** The client (bank) wants to know why the Non-Performing Assets (NPAs) are increasing in his/her bank?
- **Analytics team should:**
  - Understand the business problem by breaking it down
  - While breaking down the problem, the team should also visualize the solution
  - In doing the above, the team should always think from client's perspective

Storytelling with Data: Activity

- Business Client's Question – Why our sales is going down?
- Nature of Business –An Indian FMCG which has 10 products in its product portfolio

- Following the steps in storytelling with data, design the wireframe of the solution with synthetic/imaginary data

Thank you!