(#) Some things to Read & Consider.

Vo Imp

⇒ [Steps to Data Exploration & Preperation.]

→ Variable Identification
→ Univariate Analysis
→ Bi-variate Analysis
→ Missing Values Treatment
→ Outlier Treatment
→ Variable transformation
→ Variable Creation.

Generalised EDA

] Feature Engineering

[We need to iterate over steps 4-7 times to get refined model]

---

(1) [Predictor_Variables → Input Variables]

Univariate Analysis for -

→ Continuous Variable

We need to understand the central tendency & the spread of the variable

Note - Univariate analysis is also performed to highlight the missing & outlier value.

→ Categorical Variable.

We'll use frequency table to understand distribution of each category.
We can also read as percentage of values under each category. It can be measured using two metrics Count & Count% against each category. Bar chart can be used as Visualization.

Bi-variate - relation b/w two variables.
We look for association & dissociation b/w two variables at pre-significance level.

b/w -

① Continuous & Continuous →
To check relation b/w two continuous variable we use scatter plot.

② Categorical & Categorical
→ Two-way table of count & count%.
The rows represent category of one variable & the column represents category of the other variable.
→ Stacked-Column Chart
Stacked-bar plot, visual of a two way table,
two way table ↔ crosstab, sns.countplot(hue = ?) ✓

→ Chi-Square test
It is used to derive the Statistical significance b/w two variables. It returns probability for the computed chi-Square distribution with the degree of freedom.
Probability → 0 - are dependent, 1 - are independent.

DATE
PAGE

Note

The chi-Square test statistic for a test independence of two categorical variables is founded by

$$\chi^2 = \frac{\sum (O-E)^2}{E}$$

where O - observed frequency
E - expected frequency under the null hypothesis -

$$E = \frac{\text{row total} \times \text{Column total}}{\text{Sample Size}}$$

→ (Most imp)

(3) Categorical & Continuous (Box Plots for smaller levels)

We can draw box-plots for each level of categorical Variable.

If levels are smaller in number it'll not show Statistical Significance.

To look at the Statistical Significance we can perform z-test, t-test or ANOVA.

z-test | t-test

Either test asses whether mean of two groups are statistically different from each other or not.

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

if the probability of z is small then the different b/w the averages is more significant.

The t-test is very similar to z-test but it is used when no. of observation for both categories is less than 30.

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$x_1, x_2 \rightarrow$ averages
$s_1^2, s_2^2 \rightarrow$ variances
$n_1, n_2 \rightarrow$ counts
$t \rightarrow$ has t distribution with $n_1 + n_2 - 2$ degree of freedom.

ANOVA

It assess whether the avg. of more than two groups is statistically different.

(2) # Treating the missing values

(1) Deletion

(2) Mean / Mode / Median Imputation.

(3) Prediction model - predictive model to estimate missing data.

(4) KNN Imputation

details are in PC / mobile.

DATE
PAGE

3 # Outliers Detection & treatment. (Part I)

Types

- → Artificial / Non-natural
- → Natural

Examples –

- → Data Entry Errors (Human)
- → Measurement Errors (Instrumental)
- → Experimental Error &
- → Intentional Outliers (Self reported measures)
- → Data Processing Error (Data mining)
- → Sampling Error
- → Natural Outliers.

Impact due to Outliers.

- → It increases the error variance and reduces the power of Statistical test.
- → If the outliers are non-randomly distributed, they can decrease normality.
- → They can bias or influence estimates that may be of substantive interest.
- → They can also impact Statistical model assumptions.

Detection of Outliers.

Visualization using –

- → Box plots
- → Histograms
- → Scatter plots.

Removing Outliers

→ Deleting Observations
→ Transforming and binning values.
[eg → Natural log of a value reduces the variation caused by extreme values.]
→ Imputing. (If artificial outlier, then we can impute it just like missing values)
→ Treat Separately.
If there are significant number of outliers we need to treat them seperately in the statistical model.

Final Stage – Feature Engineering. (Imp)

We perform feature Engineering after completing the 5 steps –

- Variable Identification
- Univariate, Bivariate Analysis
- Missing values imputation
- Outliers treatment
- feature Engineering

Variable Transformation.
Variable/feature creation.

#. Variable Transformation

– It refers to the replacement of a variable by a function. for instance, replacing a variable by square, cube or logarithm transformation.

#. When to use these transformations?

+ When we want to change the Scale of a variable or Standardize the value of a variable for better understanding

+ When we can transform the complex non-linear relationships into the linear relationships. Log transform is commonly used in this situation.

+ Symmetric distribution is preferred over a skew distribution as it is easier to interpret and generate inferences. It is used to reduce skewness.

Note
- for right skewed, we take square / cube root or logarithm for variable.
- For left skewed, we take square / cube or exponential of variables.

Common Methods of transformation.
- Logarithm
- Square / Cube roots
- Binning - Used to categorize variable. It is performed on original values, percentile or frequency.

(#) Variable / feature Creation.

feature variable generation is a process to generate new variables / features based on existing variables.
for eg, if we have date (dd-mm-yy) as an input variable in a dataset. We can generate new variables like day, month, year, week, weekday that may have better relationships with target variable.

DATE
PAGE

④ Various techniques to create new features.

1) Creating derived Variables. –
This refers to creating new variables from existing variable using set of functions or different methods.

2) Creating dummy Variables. –
One of the most common application of dummy variable is to convert categorical to numerical Variable.
Dummy Variables are also called Indicator variables. It is useful to take categorical variable as a predictor in statistical models.