Statistics

Descriptive Statistics

- **Overview**
  - Descriptive statistics allow you to characterize your data based on its properties
- **Measures of Frequency**
  - Count, Percent, Frequency
  - Shows how often something occurs
  - Use this when you want to show how often a response is given
  - Feature Engineering — Create Features from Categorical Variables
- **Measures of Central Tendency**
  - Mean, Median, Mode
  - Locates the distribution by various points
  - Use this when you want to show how an average or most commonly indicated response
  - Feature Engineering
    - Missing Value Treatment
    - Create Features using Categorical and Numerical Variables
- **Measures of Dispersion or Variation**
  - Range, Variance, Standard Deviation
  - Identifies the spread of scores by stating intervals
  - Range = High/Low Points
  - Variance or Standard Deviation = difference between observed score and mean
  - Use this when you want to show how "spread out" the data are. It is helpful to know when your data are so spread out that it affects the mean
  - Feature Selection
    - Select features with variability
    - Outlier Detection and Imputation
- **Measure of Position**
  - Percentile Ranks, Quartile Ranks
  - Describe how scores fall in relation to one another. Relies on standardized scores
  - Use this when you need to compare scores to a normalized score
  - Feature Engineering
    - Ranked Features
    - Outlier Detection using IQR

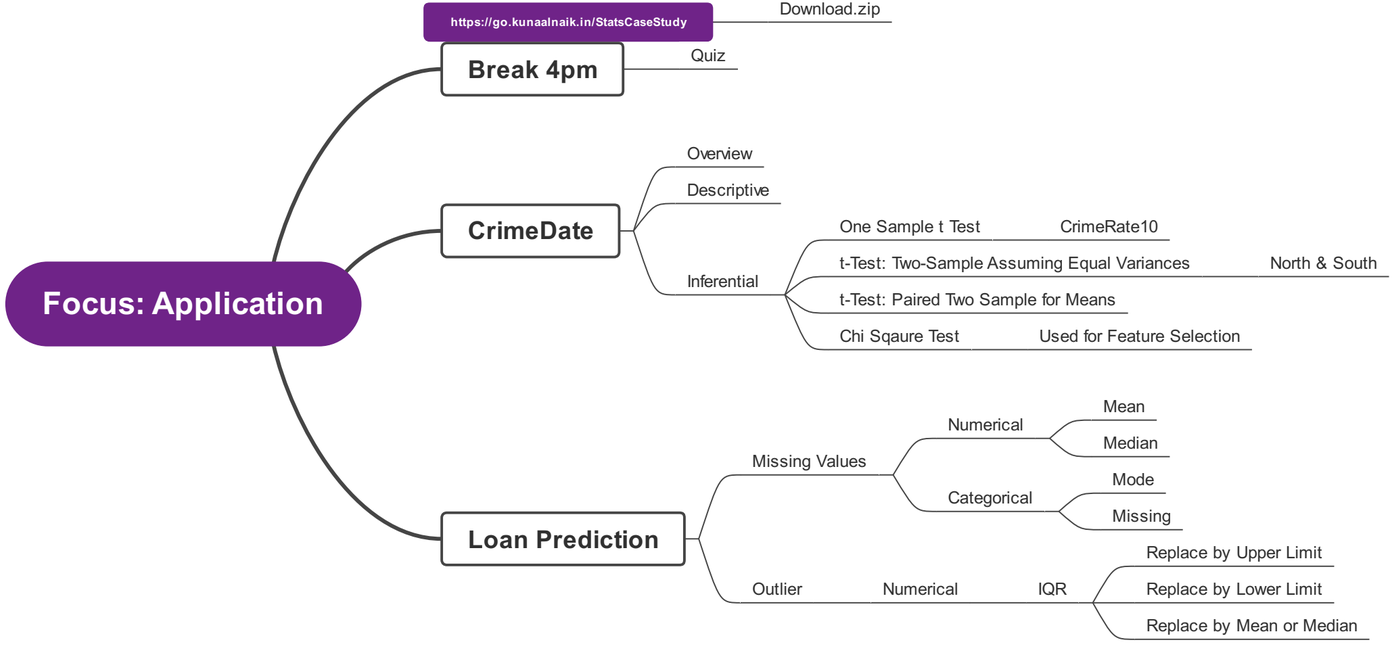
Inferential Statistics

- **Overview**
  - Used when you want to move beyond simple description or characterization of your data and draw conclusions based on your data
  - Making comparisons across time, comparing different groups, or trying to make predictions based on data that has been collected
- **Parametric Statistics**
  - **Overview**
    - Estimate the value of a population parameter from the characteristics of a sample
    - Assumes the values is a sample are normally distributed
    - Interval/Ratio level data required
  - **Tests**
    - **T-Test**
      - One Sample T-Test — Single sample within a population — Compare Salary of Data Scientist compared to known National Average
      - Independent T-Tests — Two independent groups that are mutually exclusive (same feature) — Compare Salaries of Data Scientist and ML Engineers
      - Paired or correlated T-Tests — Same one group with two measures over time — Compare Starting salaries vs Current Salaries of Data Scientists
    - **ANOVA**
      - One Way - One Feature
      - Two Way - Two Features
    - **Mutiple Regression**
- **Nonparametric Statistics**
  - **Overview**
    - No Assumptions about underlying distribution of the sample (do not follow normal distribution)
    - Used when the data do not meet the assumption for a parametric test (ordinal and nominal data)
  - **Tests**
    - Chi Square — Feature Selection - Higher the p-values, better used for model training
    - Mann Whitney U test(Wilcoxon Rank Sum Test)

```python
# Load libraries
from sklearn.datasets import load_iris
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

# Load iris data
iris_dataset = load_iris()

# Create features and target
X = iris_dataset.data
y = iris_dataset.target

# Convert to categorical data by converting data to integers
X = X.astype(int)

# Two features with highest chi-squared statistics are selected
chi2_features = SelectKBest(chi2, k = 2)
X_kbest_features = chi2_features.fit_transform(X, y)

# Reduced features
print('Original feature number:', X.shape[1])
print('Reduced feature number:', X_kbest.shape[1])
```

Output:

```
Original feature number: 4
Reduced feature number : 2
```


Focus: Application
https://go.kunaalnaik.in/StatsCaseStudy
Download.zip
Break 4pm
Quiz
CrimeDate
Overview
Descriptive
Inferential
One Sample t Test
CrimeRate10
t-Test: Two-Sample Assuming Equal Variances
North & South
t-Test: Paired Two Sample for Means
Chi Sqaure Test
Used for Feature Selection
Loan Prediction
Missing Values
Numerical
Mean
Median
Categorical
Mode
Missing
Outlier
Numerical
IQR
Replace by Upper Limit
Replace by Lower Limit
Replace by Mean or Median