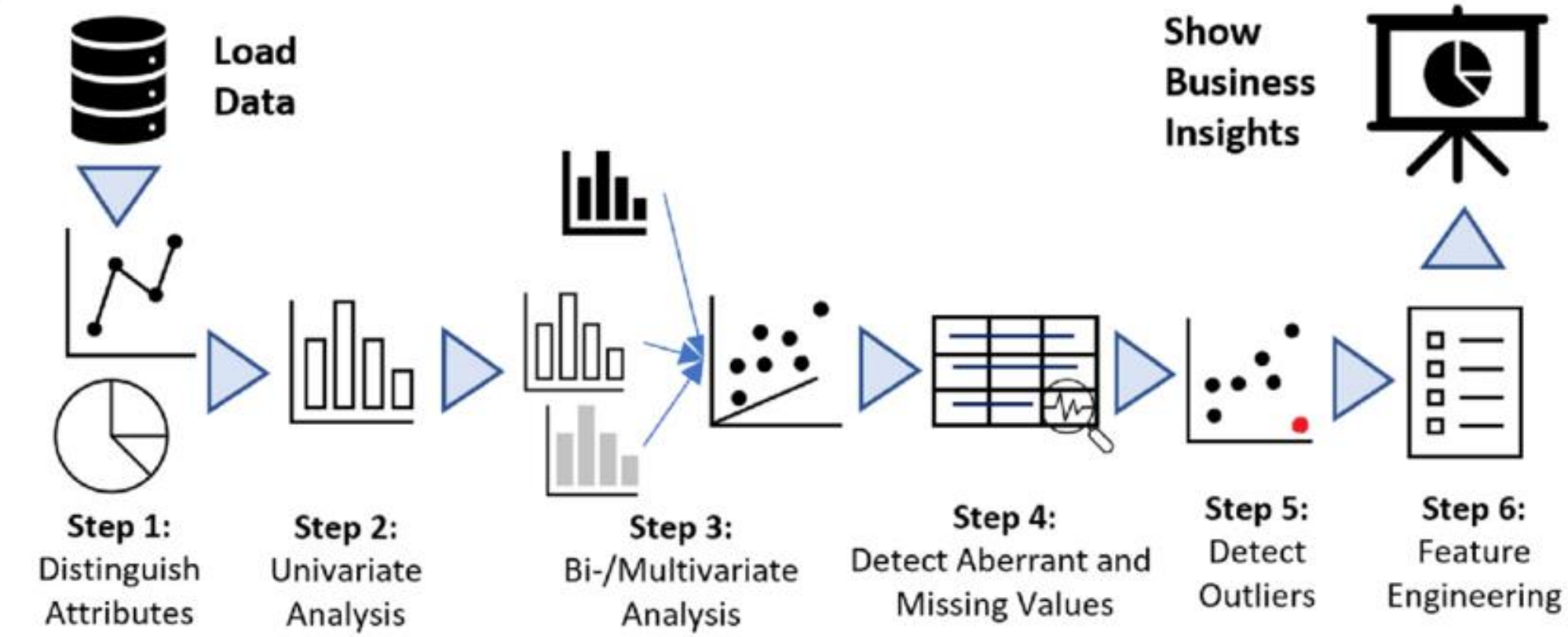
Exploratory Data Analysis

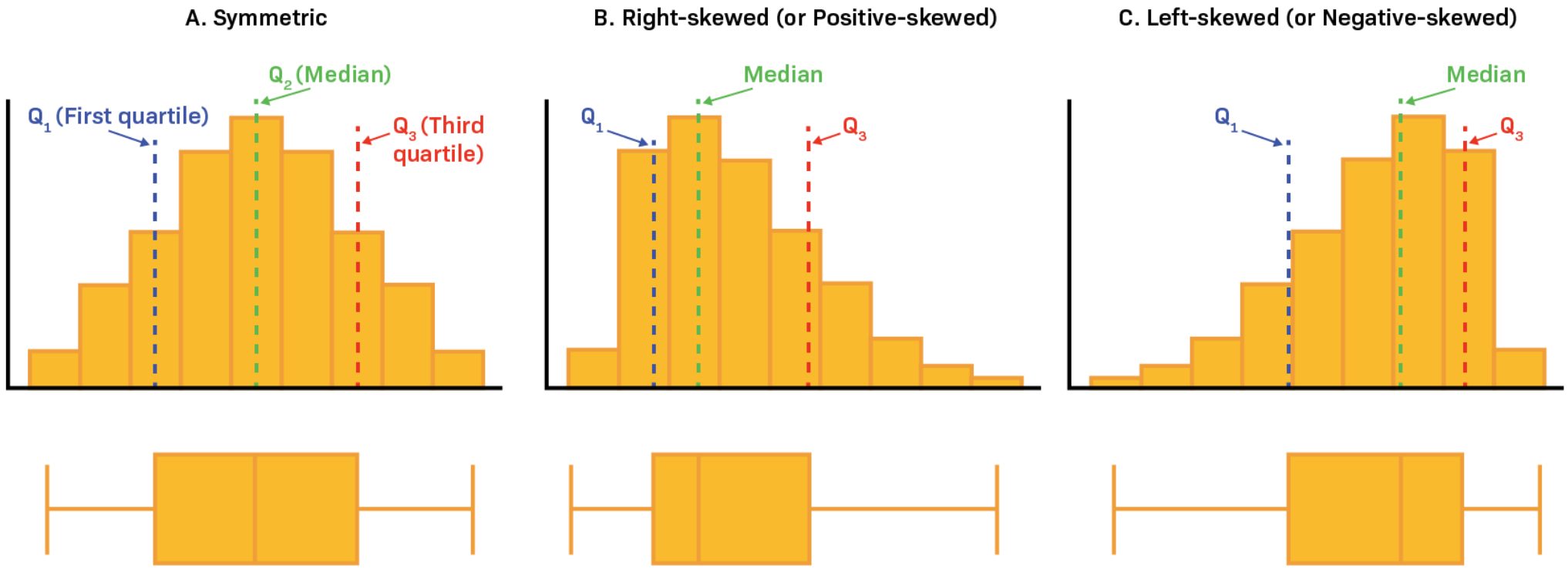
CS 459 Business Intelligence

Summarizing 6-steps of Data Wrangling

EDA

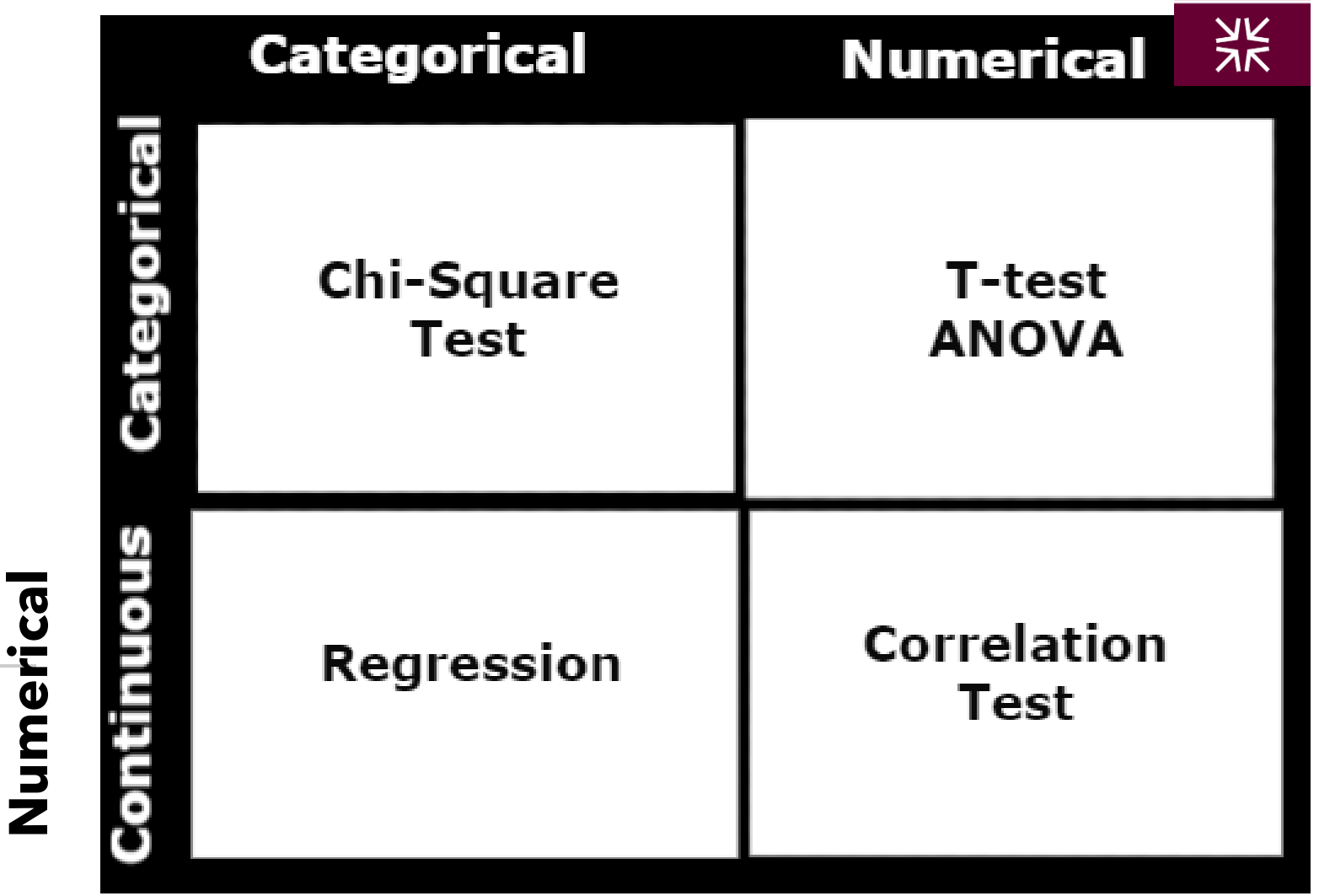

Load Data
Step 1: Distinguish Attributes
Step 2: Univariate Analysis
Step 3: Bi-/Multivariate Analysis
Step 4: Detect Aberrant and Missing Values
Step 5: Detect Outliers
Step 6: Feature Engineering
Show Business Insights

Univariate Analysis
Histograms and Box Plots



Multi-Variate Analysis

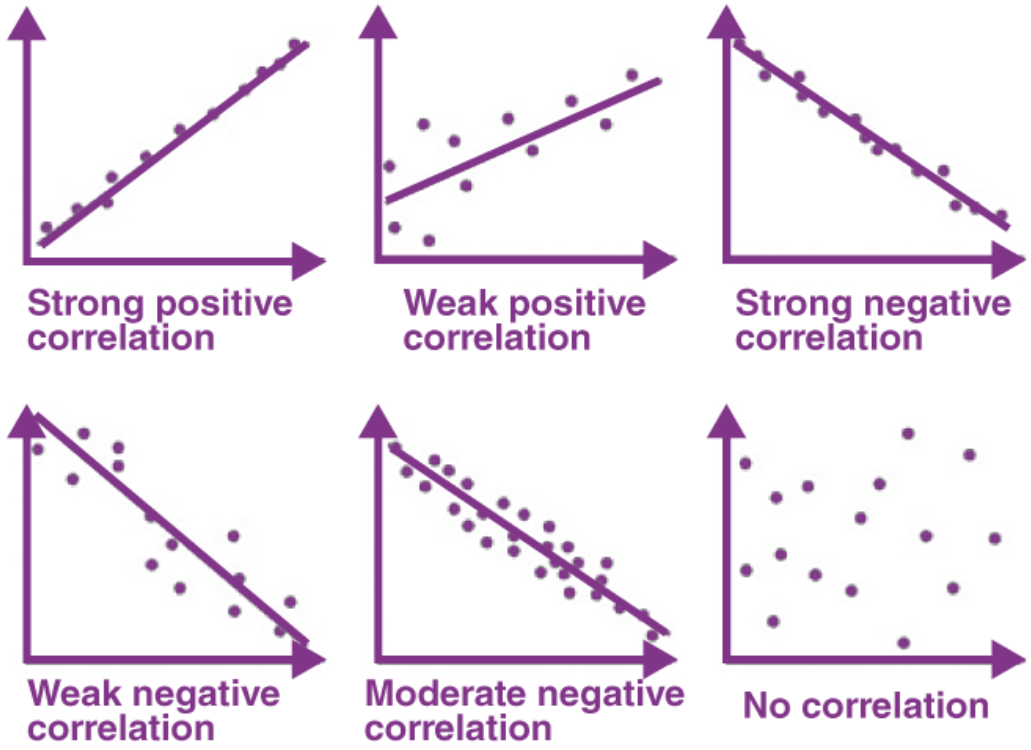
Bi-variate/ Multivariate Statistical Testing

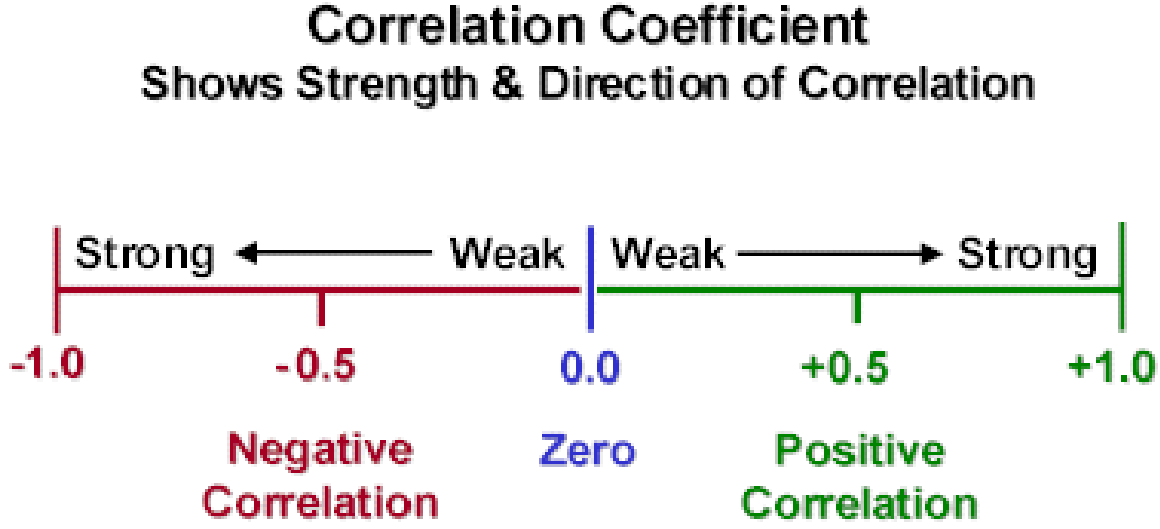


Correlation Testing

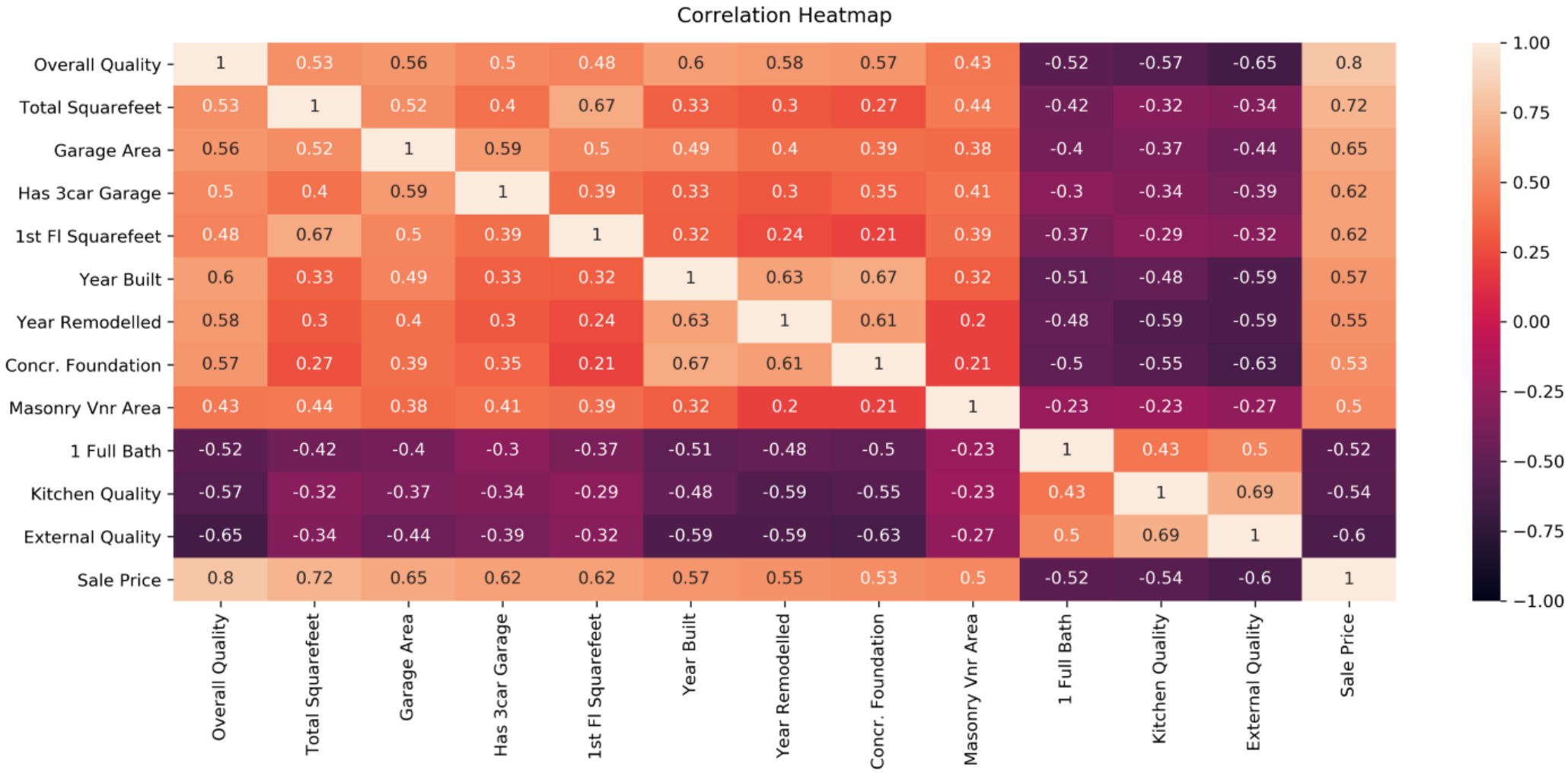
- One way to quantify the relationship between two variables is to use the *Pearson correlation* coefficient, which measures the linear association between two variables.
- *Correlation Coefficient*
  - **-1** indicates a perfectly negative linear correlation
  - **0** indicates no linear correlation
  - **1** indicates a perfectly positive linear correlation

Correlation Analysis





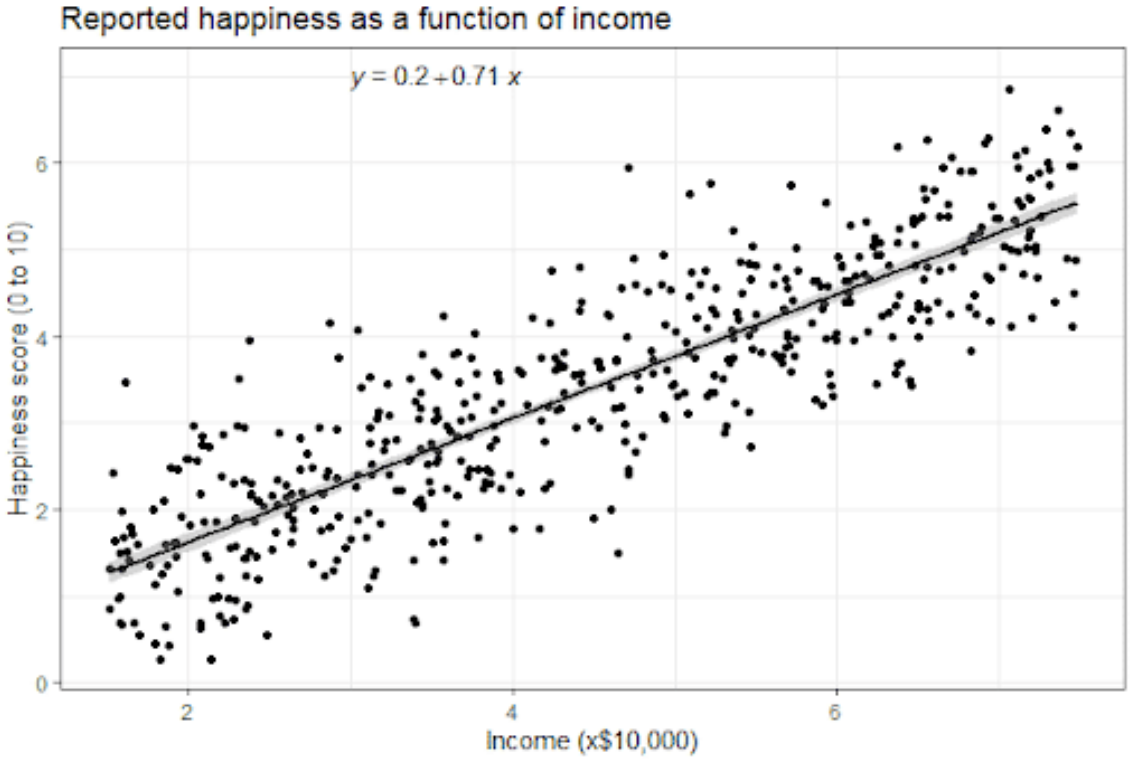
Correlation heat map



P-value in Hypothesis Testing

- The **p-value** indicates the **strength of evidence against a null hypothesis**.
- Smaller p-value → Stronger evidence against the null hypothesis
- The commonly used threshold (**α**) is **0.05 (5%)**.
- **If p-value < 0.05:** Reject the null hypothesis and **accept the alternative hypothesis**.
- **If p-value ≥ 0.05: Fail to reject** the null hypothesis (not enough evidence to accept the alternative hypothesis).

Regression



- In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a **dependent variable** and **one or more independent variables**.

Regression

- **Linear Regression:**
  - Used when the dependent variable is numerical.
  - Example: Predicting house prices based on size and location.
- **Logistic Regression:**
  - Used when the dependent variable (target) is categorical.
  - Example:
    - Predicting whether an email is spam (1) or not (0).
    - Determining if a tumor is malignant (1) or benign (0).

Interpreting Results

```
from sklearn.linear_model import LinearRegression

#initiate linear regression model
model = LinearRegression()

#define predictor and response variables
X, y = df[['x1', 'x2']], df.y

#fit regression model
model.fit(X, y)
```

R-square value:

- This means that **76.67%** of the variation in the response variable can be explained by the two predictor variables in the model.

```
#display regression coefficients and R-squared value of model
print(model.intercept_, model.coef_, model.score(X, y))

70.4828205704 [ 5.7945 -1.1576] 0.766742556527
```

$$y = 70.48 + 5.79x_1 - 1.16x_2$$

```python
import statsmodels.api as sm

#define response variable
y = df['y']

#define predictor variables
x = df[['x1', 'x2']]

#add constant to predictor variables
x = sm.add_constant(x)

#fit linear regression model
model = sm.OLS(y, x).fit()

#view model summary
print(model.summary())
```

P-value for each variable can also be computed. This gives the statistical significance of each variable when p value is less than 0.05.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.767
Model:                            OLS   Adj. R-squared:                  0.708
Method:                 Least Squares   F-statistic:                     13.15
Date:                Fri, 01 Apr 2022   Prob (F-statistic):            0.00296
Time:                        11:10:16   Log-Likelihood:                -31.191
No. Observations:                  11   AIC:                             68.38
Df Residuals:                       8   BIC:                             69.57
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         70.4828      3.749     18.803      0.000     61.839      79.127
x1             5.7945      1.132      5.120      0.001      3.185       8.404
x2            -1.1576      1.065     -1.087      0.309     -3.613       1.298
==============================================================================
Omnibus:                        0.198   Durbin-Watson:                   1.240
Prob(Omnibus):                  0.906   Jarque-Bera (JB):                0.296
Skew:                          -0.242   Prob(JB):                        0.862
Kurtosis:                       2.359   Cond. No.                         10.7
==============================================================================
```

ANOVA – Analysis of Variance

- **ANOVA** is a statistical formula used to compare variances across the means (or average) of different groups.
- A range of scenarios use it to determine if there is any difference between the means of different groups


EMPLOYEE TRAINING IMPACT
ON CUSTOMER SATISFACTION
qualtrics.XM
Customer Satisfaction Ratings
10
9
8
7
6
5
4
3
2
1
Beginner
Intermediate
Advanced
Levels of Employee Training

Interpreting Results

- A one-way ANOVA uses the following null and alternative hypotheses:
- **H_0 (null hypothesis):** $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$
  (all the population means are equal)
- **H_1 (alternate hypothesis):** at least one population mean is different from the rest

ANOVA Results

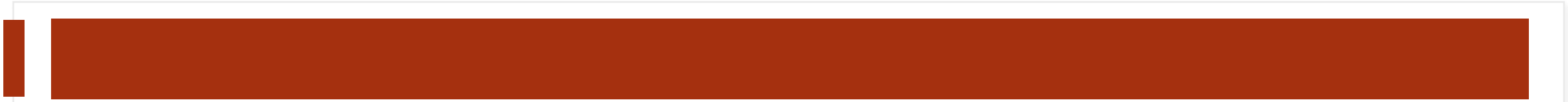
```
from scipy.stats import f_oneway

#perform one-way ANOVA
f_oneway(group1, group2, group3)

(statistic=2.3575, pvalue=0.1138)
```

The F test statistic is **2.3575** and the corresponding p-value is **0.1138**.

F-Statistic (Fisher Statistic):
A measure used to test the overall significance of a statistical model by comparing model fit against a baseline model.

- The F-statistic is compared with a critical value or evaluated using the p-value.
- If the **p-value < 0.05**, the model is considered statistically significant (**Reject the null hypothesis**).
- If the **p-value ≥ 0.05**, the model is not statistically significant (**Fail to reject the null hypothesis**).
- **Note:** A large F-statistic generally corresponds to a small p-value, indicating statistical significance.

Head over to Descriptive Statistics Notebook



Create a Summary Table for the ANOVA Test

Column A	Column B	P-Value	Reject / Fail to Reject	Difference exists / No Difference
Sales	Order Priority	0.22		

Complete Table for ANOVA test

Column A	Column B	P-Value	Reject / Fail to Reject	Difference exists / No Difference
Sales	Order Priority	0.22		
Sales	Ship Mode	0.00		
Sales	Region	0.33		
Sales	Customer Segment	0.63		
Sales	Product Category	4.908931e-168		
Sales	Sub Category	0.00		
Sales	Product Container	0.00		

Tukey Test

- Tukey's test determines the individual means which are significantly different from a set of means.
- Tukey's test is a multiple comparison test and is applicable when there are more than two means being compared (for two means, utilize a *t* test).
- Typically, Tukey's test is utilized after ANOVA has shown that significant difference exists and this determines where the difference exists.

Interpret ANOVA

```
#enter data for three groups
a = [85, 86, 88, 75, 78, 94, 98, 79, 71, 80]
b = [91, 92, 93, 90, 97, 94, 82, 88, 95, 96]
c = [79, 78, 88, 94, 92, 85, 83, 85, 82, 81]

#perform one-way ANOVA
f_oneway(a, b, c)

F_onewayResult(statistic=5.167774552944481, pvalue=0.012582197136592609)
```

- H0 : all the population means are equal
- We can see that the overall p-value from the ANOVA table is **0.01258**.

- Since this is less than 0.05, reject null hypothesis which implies that the mean values across each group are not equal.
- Proceed to perform Tukey's Test to determine exactly which group means are different.

Interpreting Tukey

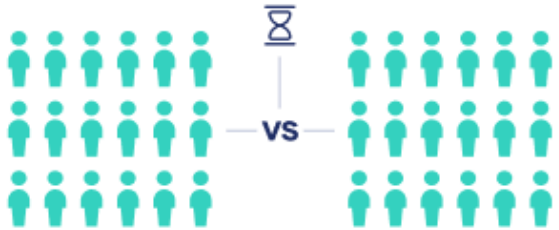
- There is a statistically significant difference between the means of groups *a* and *b* and groups *b* and *c*, but not a statistically significant difference between the means of groups *a* and *c*.

```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower    upper  reject
-----------------------------------------------------
     a      b      8.4 0.0158   1.4272 15.3728   True
     a      c      1.3 0.8864  -5.6728  8.2728  False
     b      c     -7.1 0.0453 -14.0728 -0.1272   True
-----------------------------------------------------
```

T-test

- A t-test is a statistical test that **compares the means of two samples**. Null hypothesis: The difference in group means is zero and an alternate hypothesis that the difference in group means is different from zero.

Paired-samples t test

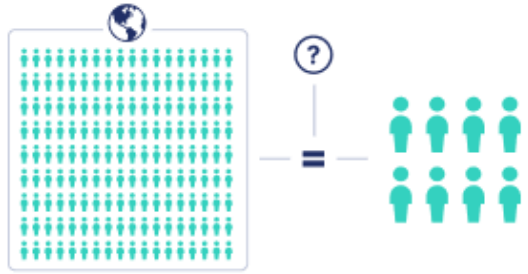


Investigate whether there's a difference within a group between two points in time (within-subjects).

Independent-samples t test

Investigate whether there's a difference between two groups (between-subjects).

One-sample t test



Investigate whether there's a difference between a group and a standard value or whether a subgroup belongs to a population.

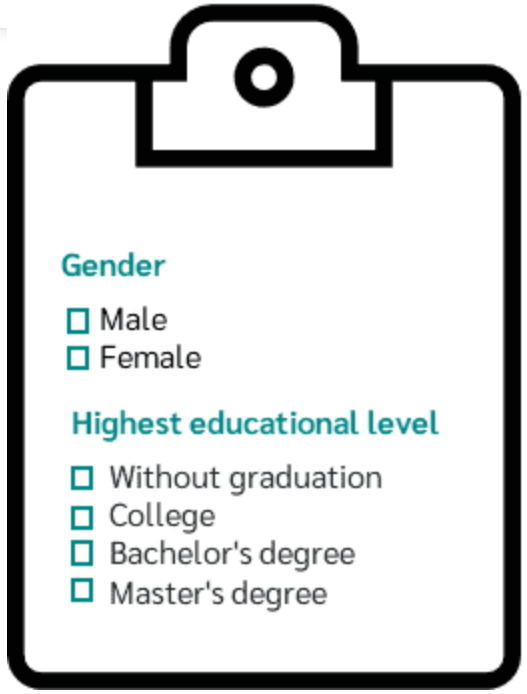
Interpreting Results

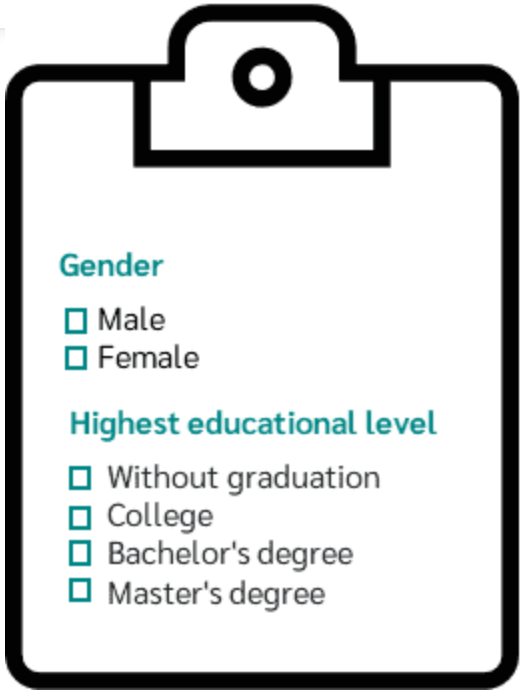
Compare **P-value** in the same way.

The **t-value** measures the size of the difference relative to the variation in your sample data. Put another way, T is simply the calculated difference represented in units of standard error. **The greater the magnitude of T, the greater the evidence against the null hypothesis**.

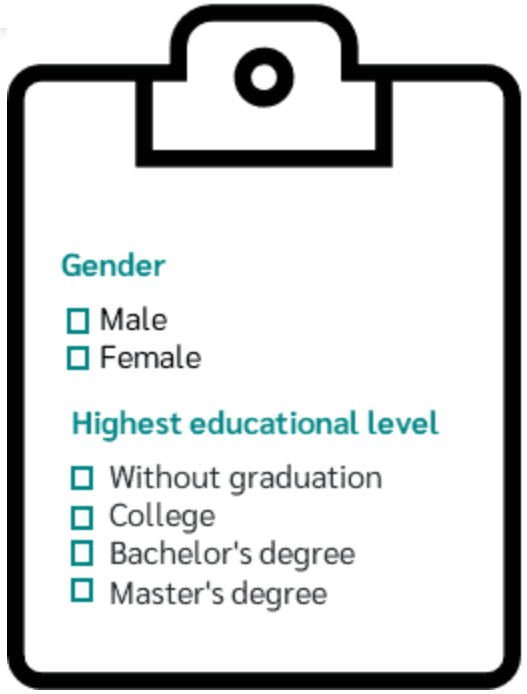
Chi-square

- **Chi-square** is a statistical test used to examine the differences between **categorical variables** from a random sample in order to judge goodness of fit between expected and observed results.
- **Chi-square** is most commonly used by researchers who are studying survey response data because it applies to categorical variables. Demography, consumer and marketing research, political science, and economics are all examples of this type of research
- **Example**: Is gender related to political party preference?


Gender
Male
Female
Highest educational level
Without graduation
College
Bachelor's degree
Master's degree



Fall	Gender	Highest educational level
1	Male	College
2	Female	Without graduation
3	Male	Without graduation
4	Male	Bachelor's degree
5	Female	Master's degree
6	Male	Bachelor's degree
7	Female	Master's degree
...	...	...



Fall	Gender	Highest educational level
1	Male	College
2	Female	Without graduation
3	Male	Without graduation
4	Male	Bachelor's degree
5	Female	Master's degree
6	Male	Bachelor's degree
7	Female	Master's degree
...	...	...

	Female	Male
Without graduation	6	7
College	13	16
Bachelor's degree	16	15
Master's degree	8	11
Total	43	49

Is there a correlation between gender and the highest level of education? → Chi² - Test