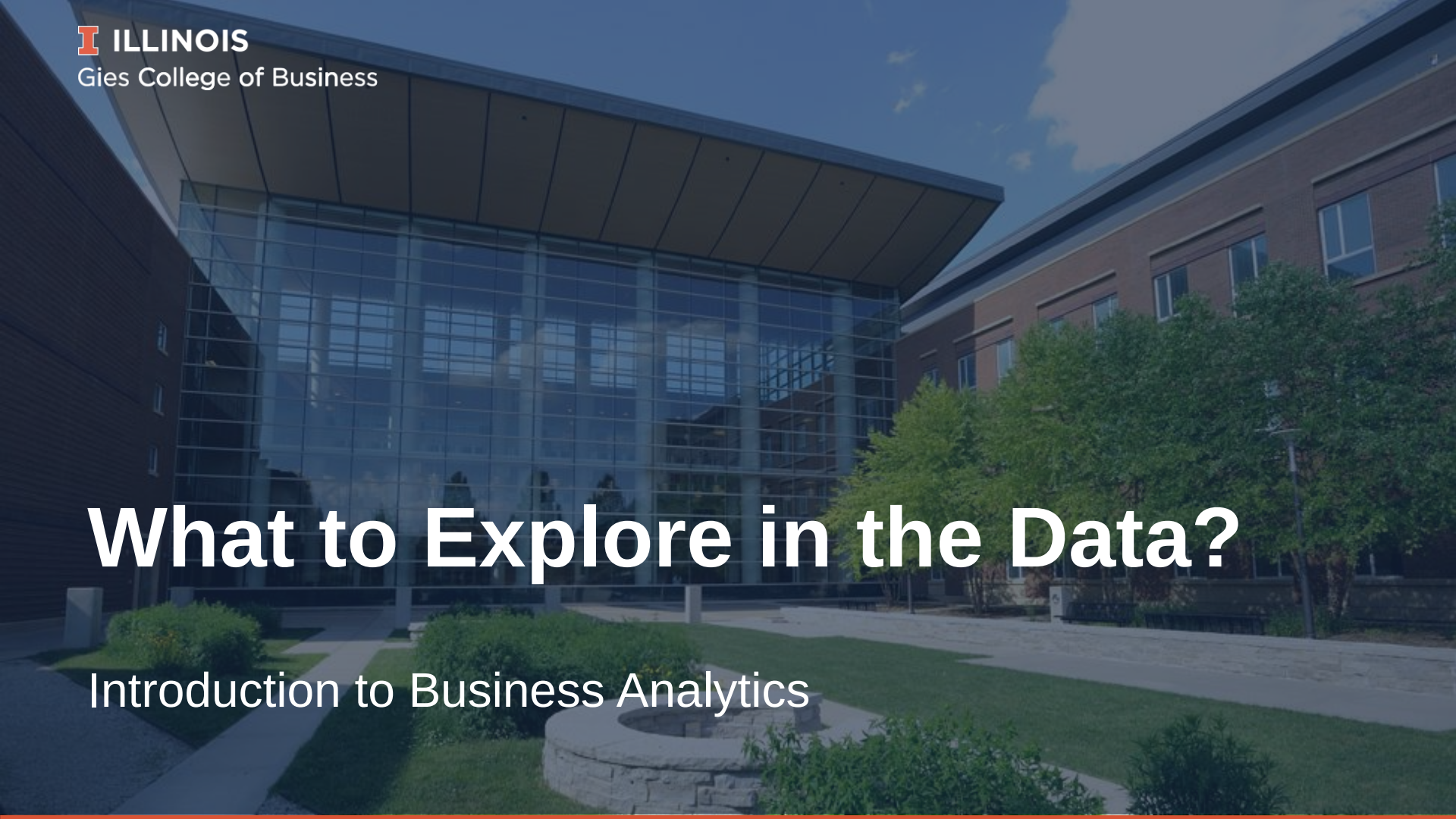
Introduction to Business Analytics

Ashish Khandelwal



What to Explore in the Data?

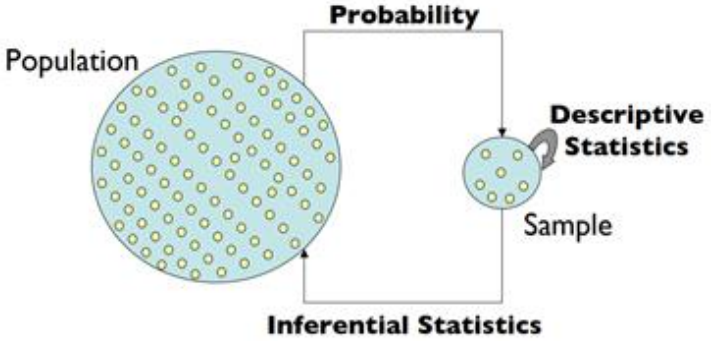
Introduction to Business Analytics

Exploratory Data Analysis (EDA)

An approach to analyzing datasets to summarize their main characteristics, often with visual methods

Exploratory Data Analysis (EDA)

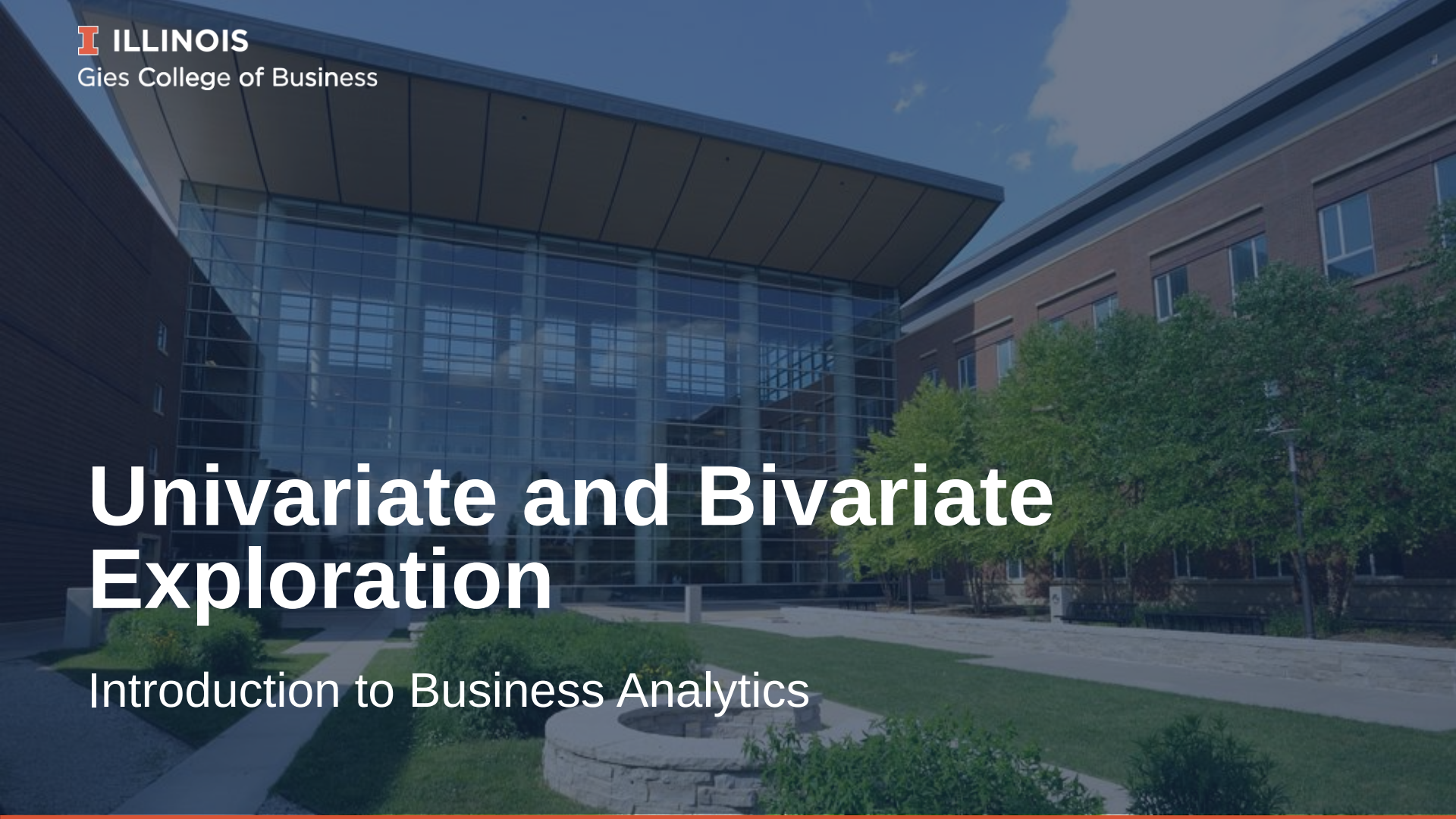
- Trends
- Outliers
- Patterns in data



http://orzo.union.edu/~khetans/Teaching/BNG202/Stats%20Lecture%201.pdf

What do we explore?

- Missing values
- Outliers
- Univariate distribution
- Bivariate and multivariate distribution



Univariate and Bivariate Exploration

Introduction to Business Analytics

Two Types of Variables

1) Numeric

- Continuous – Age, Income, Weight
- Discreet – Count

2) Categorical

- Race, Gender, Socioeconomic Status

Exploration

- Univariate
- Bivariate
- Multivariate

Univariate Exploration

For numeric variable

1) Measures of central tendency
    Mean, median, quartiles

2) Measure of dispersion
    Variance, range, interquartile range

Visualization – Histogram and boxplot

Univariate Exploration

For categorical variable

1) Counts/frequency
2) Counts proportion

Frequency table

Visualization - Bar chart/column chart

Bivariate Exploration

Assessing relationship between two numeric variables

Correlation captures the degree of association between two numeric variables

Visualization – Scatter plot

Bivariate Exploration

Assessing relationship between a numeric and a categorical variable

Visualization – Grouped boxplot and grouped histogram

Multivariate Exploration

Map other variables using

1. Color *(for numeric and categorical variables)*
2. Shape *(for categorical variables)*
3. Size *(for numeric variables)*
4. Facet grid – show separate chart for various subsets of data

Introduction to the ggplot2 Package

Introduction to Business Analytics

GGplot2

Each chart is a combination of:

1. Data
2. Aesthetic **mappings** (variables) (x axis, y axis, color, shape, size)
3. Layer: geometric objects (histogram, boxplot, violin plot, point, line, smooth, bar, col)
4. Layer: statistical transformation (identity, log)

GGplot2

Each chart is a combination of:

5. Scales: map values from data space to aesthetic space
6. Coordinate system (most frequent is Cartesian, others are polar)
7. Faceting: break up data into subsets for creating separate graphics based on each subset
8. Theme: font size, background color, legend position

Introduction to ggplot Syntax

Introduction to Business Analytics

Two Functions That Cover Three Major Components

Components are

Function

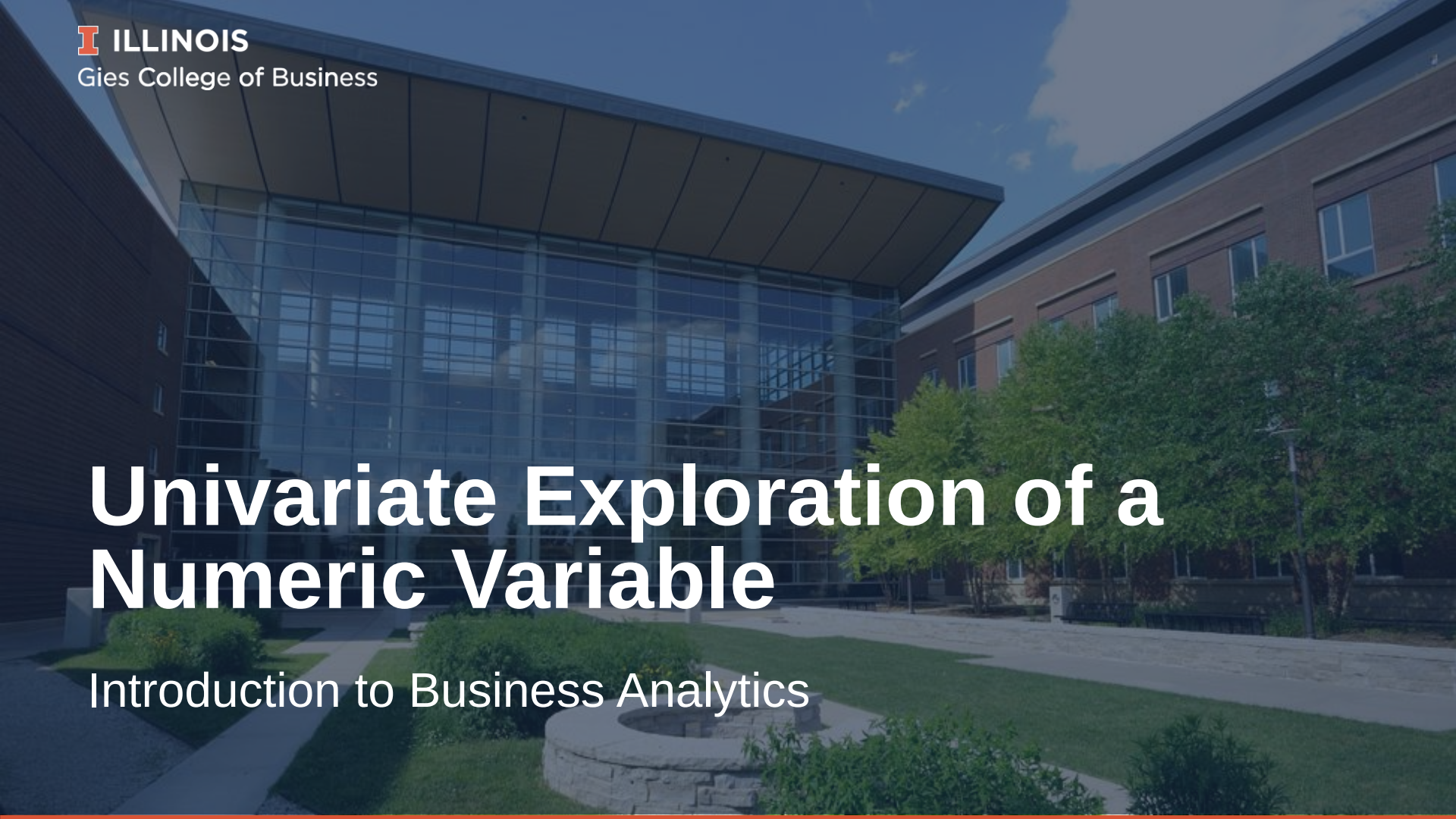
1. Data
2. Aesthetic

ggplot()

3. Geometric Object

geom_OBJECT

OBJECT is NOT the real word that comes here. It has to be replaced by either ***histogram***, ***bar, point, line***, etc., depending on what geometric object you want to create



Univariate Exploration of a Numeric Variable

Introduction to Business Analytics

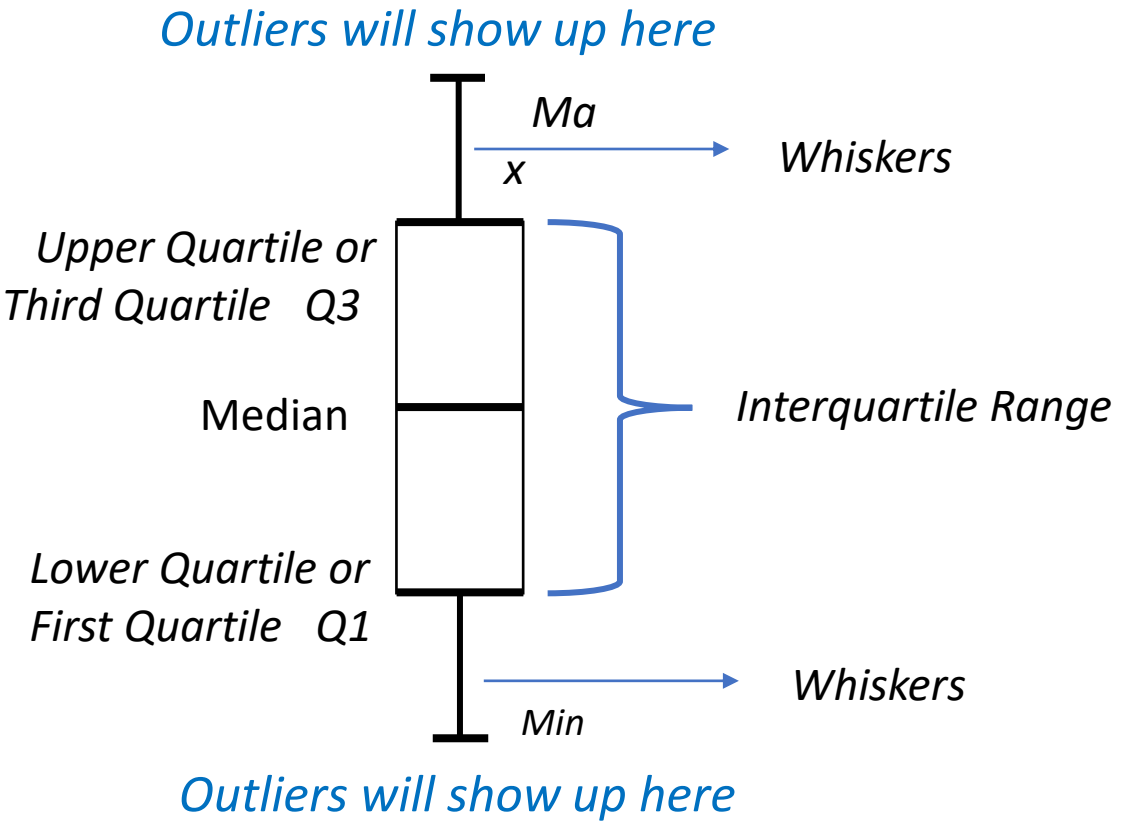
Visualization for Univariate Exploration of Numeric Variable

- Histogram
- Boxplot

Histogram

- Shows a numeric variable grouped in bins/intervals on X axis
- The number of observations that fall in each bucket is represented on Y axis

Boxplot of Price



*Price of **75%** observations < Q3 price*

***50%** observations < median price*

***50%** observations > median price*

*Price of **25%** observations < Q1 price*