Data Exploration Part 2

Lesson 2

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Schedule

Part 1	Part 2	Part 3	Part 4
Lesson 1 Data Exploration 1	**Lesson 3** Combinatorics	**Lesson 6** Intro to Regression	**Lesson 9** Näive Bayes
Lesson 2 Data Exploration 2	**Lesson 4** Hypothesis Testing	**Lesson 7** Regularization	**Lesson 10** Basic Text Analysis
Milestone 1 Data Visualization	**Lesson 5** Intro to Bayes	**Lesson 8** Time Series Analysis	**Milestone 4** Independent Project
	Milestone 2 Hypothesis Sim	**Milestone 3** Regression Models	

Housekeeping

- Should have turned in Quiz 1 and HW 1
- Assignment 02 and Quiz 02 are posted- due January 20th
- Milestone 01 is posted - due January 20th

Recap of Lesson 01 - Data Exploration Part 1

- Descriptive statistics:
  - Mean / median
  - Standard deviation / IQR
  - Covariance / correlation
  - Frequency (categorical)
- Visualization (univariate):
  - Bar plot
  - Histogram
  - Boxplot
  - KDE plot
  - Violin plot

What We'll Cover Today

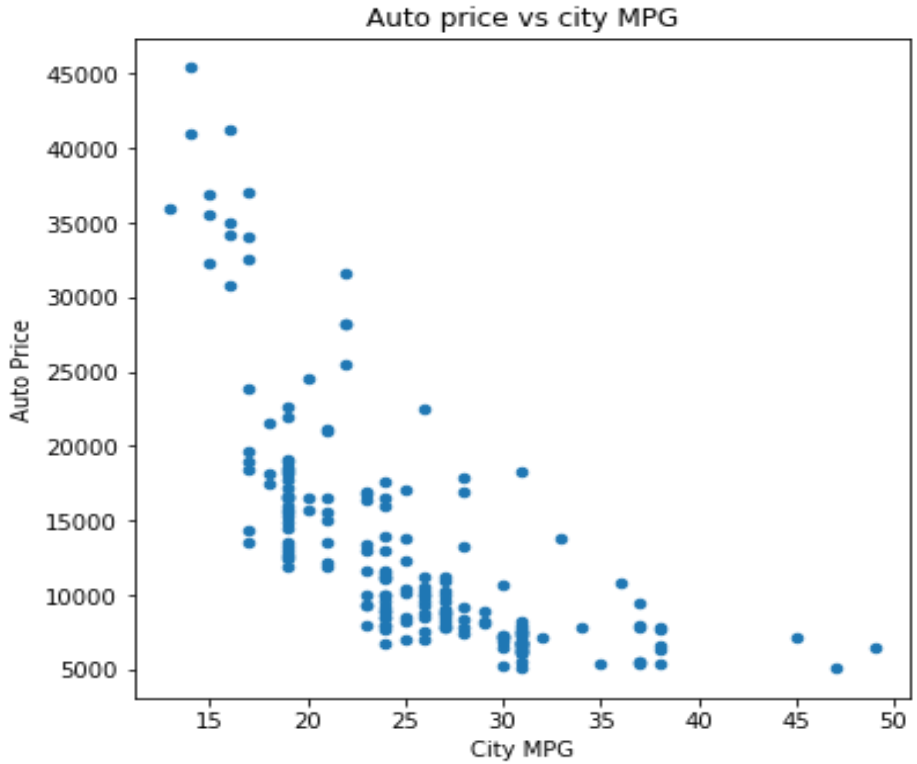
- Review Homework
- Multivariate Visualizations
  - Two dimensional plots
    - Scatter plots
    - 2D KDE plots
    - Hexbin plots
    - Heat maps
    - Line plot
  - Correlation and covariance plots (seaborn)
  - Using Aesthetics
    - Color
    - Shape
    - Size
  - Faceted (conditioned plotting)

Review Homework

Plot Types

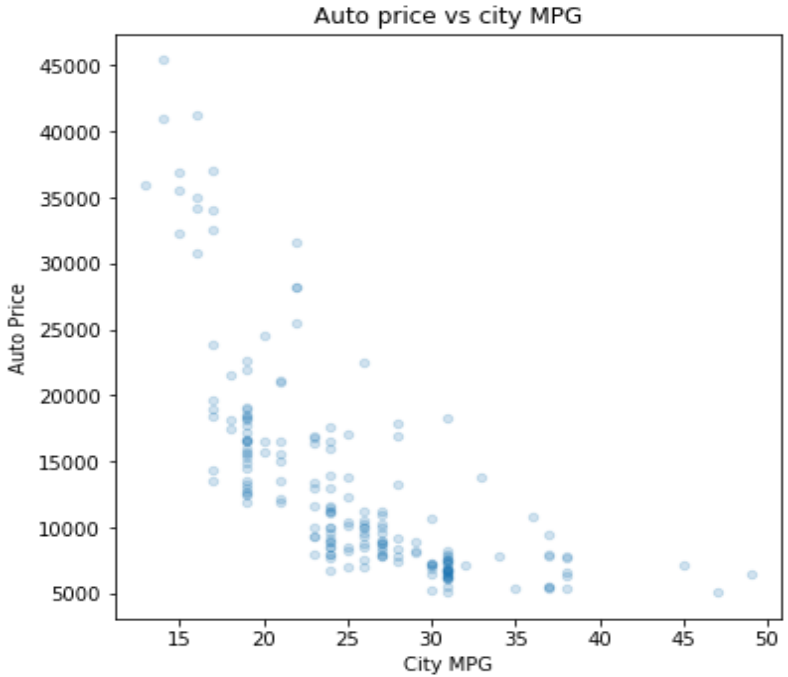
Scatter Plots -Overview

- Show relationship between two variables, one on the x-axis and the other on the y-axis.
- Notice that many points overlap, making it difficult to know how many data points are in each region.



Scatter Plots - Transparency

- To deal with overplotting, one option is to use transparency.
- Use the *alpha* parameter in the plot function:
  - Accepts values between 0-1
  - 0 = completely transparent
  - 1 = not transparent
- Transparency is additive, so not good for large number of overlapping points



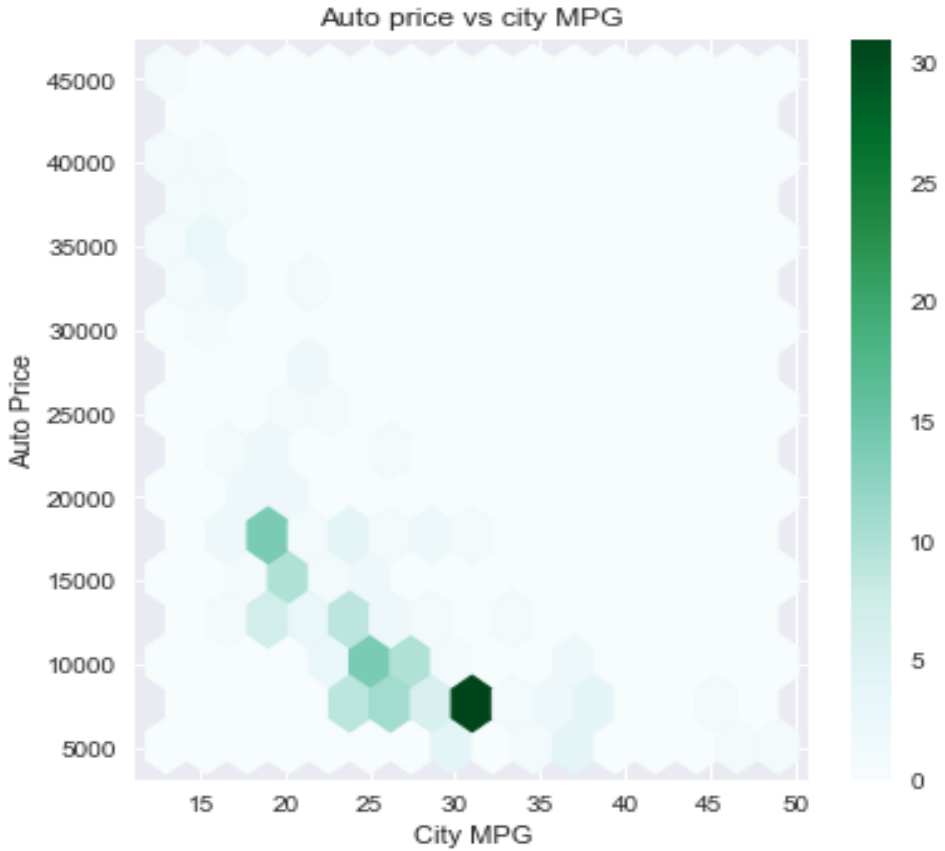
alpha = 0.2

Scatter Plots - Large Number of Overlaps

- Two options for dealing with large number of overlaps:
  - Hexbin plot (discrete)
  - 2D KDE (continuous)

Hexbin Plots

- Use when you have a lot of data points
- Represents relationship between two numerical variables
- Similar to histogram, instead of bar height, color is used.
- Hexbine = number of points per bin
- Number of bins can be set and adjusted as needed



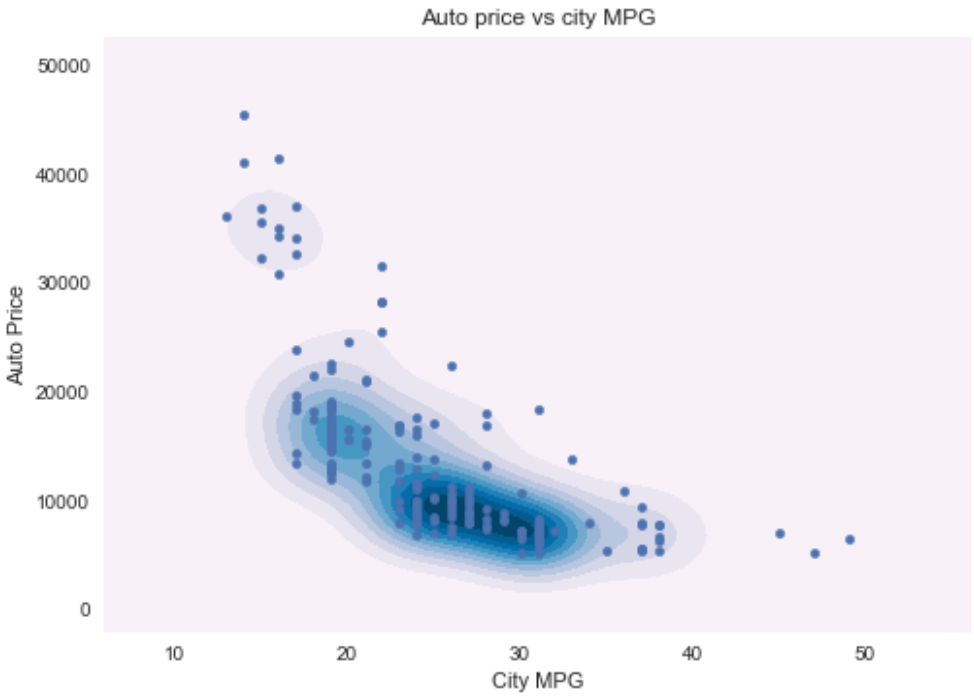
2D Kernel Density Estimation Plots

- Similar to 1D KDE, Gaussian (aka normal distribution) is a common kernel choice.
- Kernel = type of dist at each point
- More accurate representation of distribution
- Gaussian 1D:

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

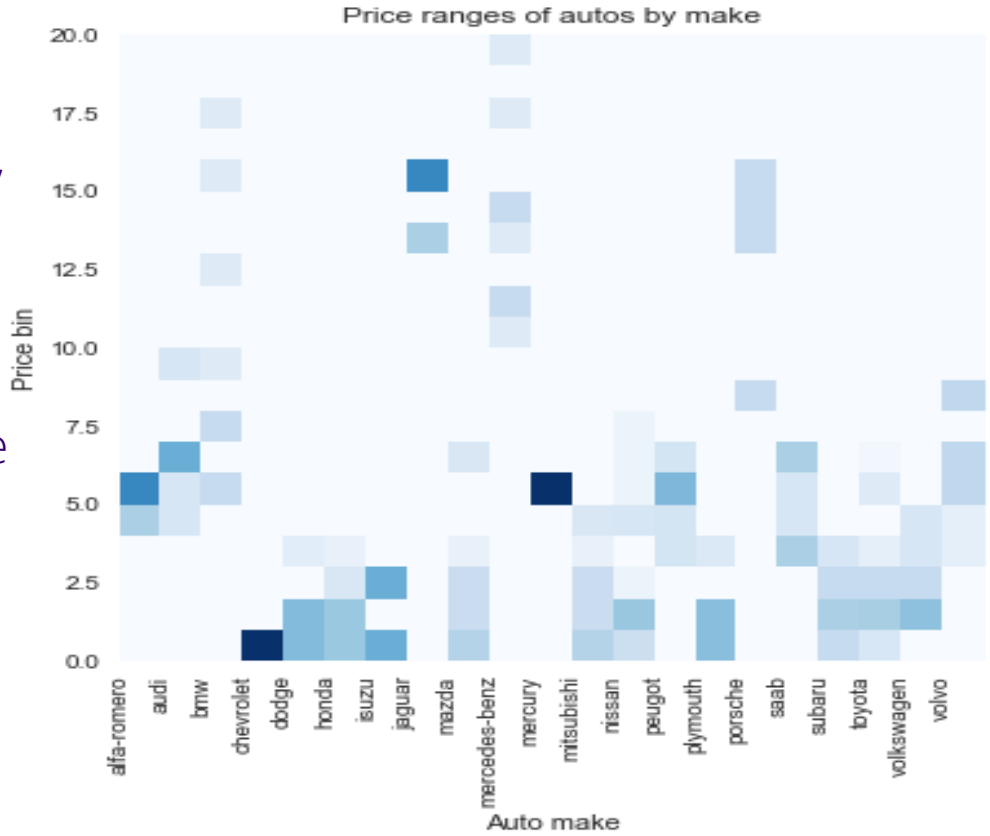
- Gaussian function 2d (one Guassia component per point):

$$f(x,y) = A\exp\left(-\left(\frac{(x-x_o)^2}{2\sigma_x^2} + \frac{(y-y_o)^2}{2\sigma_y^2}\right)\right).$$



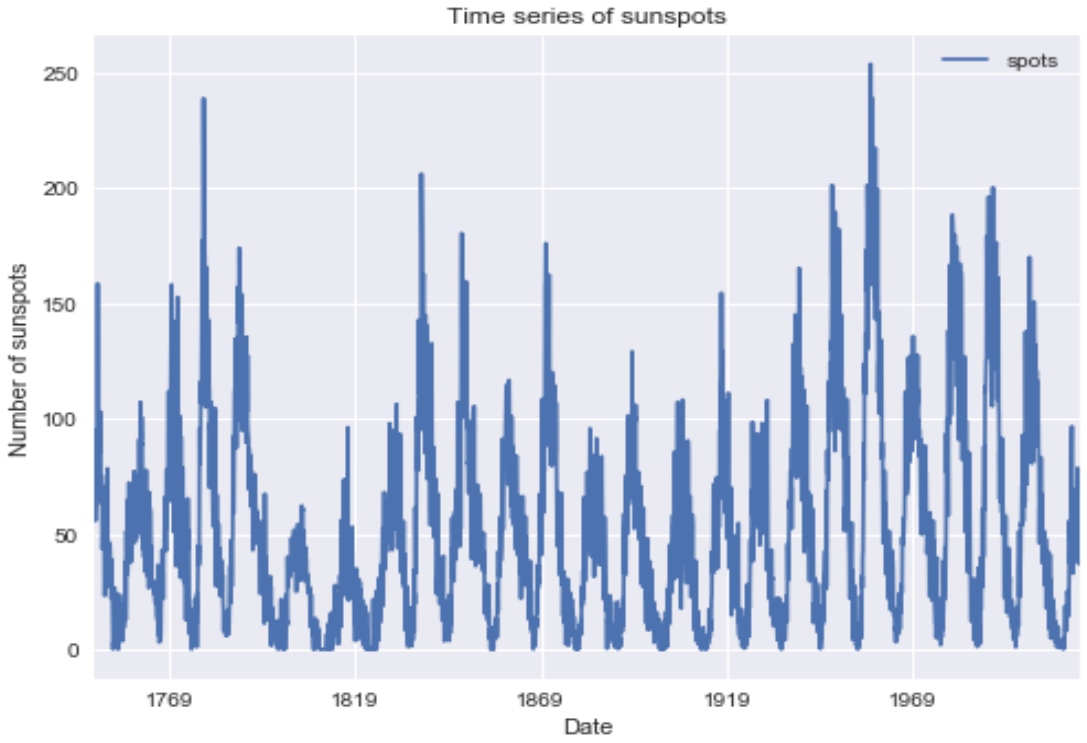
Heat Maps

- Similar to histogram and hexbins, instead of bar height, color is used.
- Here a categorical variable is shown in the x-axis
- Shows relative intensity of a value within an array
- Darker colors = great intensity
- Intensity could be counts, correlation, etc.



Line Plots

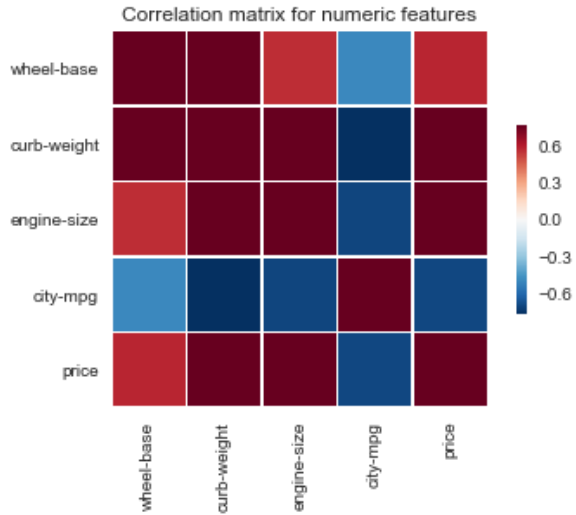
- Shows relationship between dependent variables
- Good for things like time series
- X axis = independent variable
- Y axis = dependent variable
- Line connects individual data points
- Line = “trend line”
- Easy to view high and low points in data

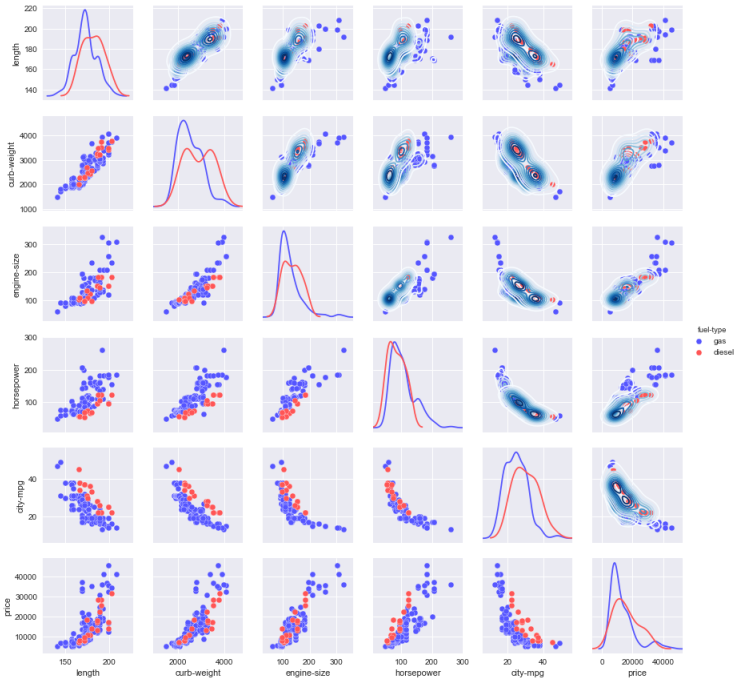


Plotting Relationships

Covariance and Correlation Plots

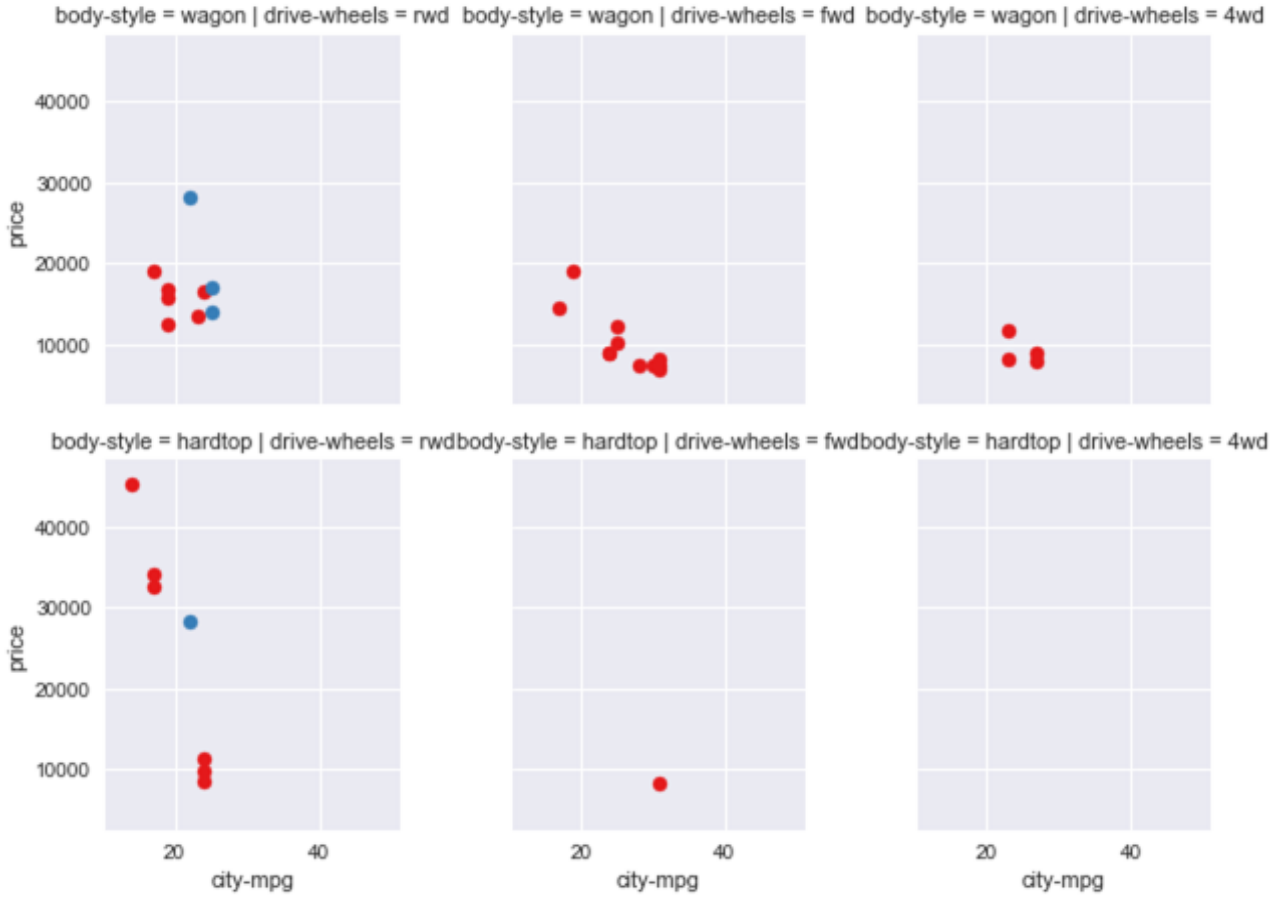
- Heat maps are useful for showing correlation visually.
- Pairwise scatter plot (scatter matrix) is another powerful plot for showing the relationship between variables





Faceted Plots

- Show relationship between sets of variables
- Split your data into one or more varialbes and plot together
- Variables that vary on x or y axis
- Allows for columns and rows



Aesthetics

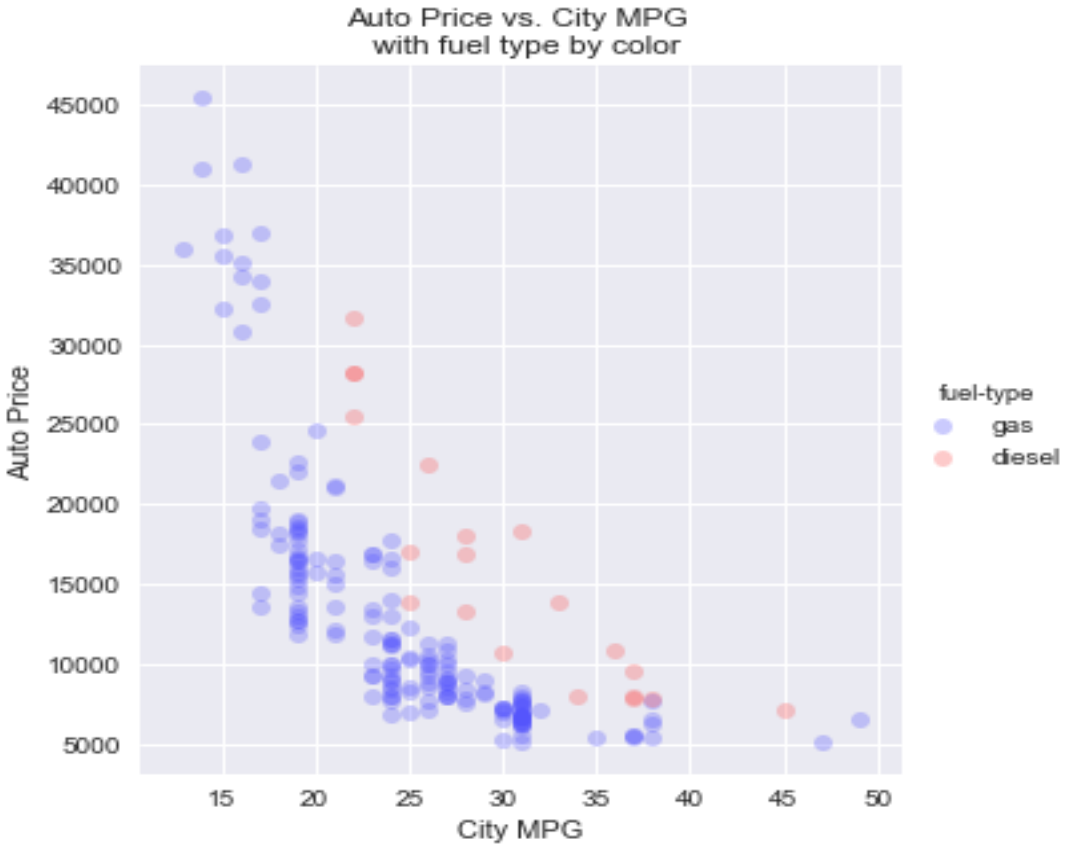
Aesthetics

- To display more than 2 variables on a 2 dimensional screen, aesthetics are one choice.
- We will look at three common plot aesthetics that can achieve this goal:

Asthetic	Data Types
Color	categorical
Size	numeric, ordered categorical
Marker shape	categorical

Aesthetics (Color)

- Color can be used to display an additional categorical variable or intensity of the variable
- Here we're displaying 3 variables: price, city MPG, and fuel type.

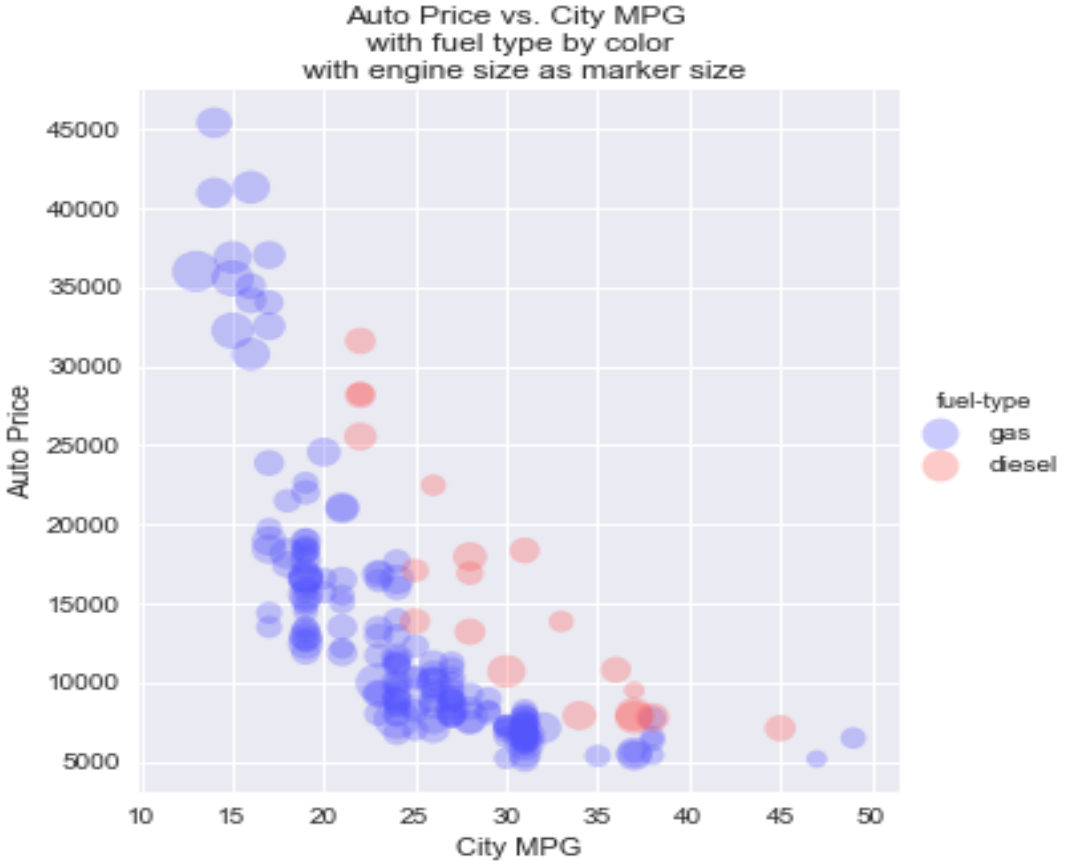


Summary

- Multivariate Visualizations
  - Two dimensional plots
    - Scatter plots
    - 2D KDE plots
    - Hexbin plots
    - Heat maps
    - Line plot
  - Correlation and covariance (seaborn)
  - Using Aesthetics
    - Color
    - Shape
    - Size
  - Faceted (conditioned plotting)

Aesthetics (Marker Size)

- Marker size is primarily used for numeric data
- Here we're displaying 4 variables: price, city MPG, engine size, and fuel type.



Aesthetics (Marker Shape)

- Marker shape is primarily used for categorical data
- This plot displays 5 variables: price, city MPG, engine size, aspiration type, and fuel type.
- Note how this plot uses all three aesthetics and may be getting overly complex for a single plot

Aspect Ratio

- Plot aspect ratio (height vs width of plot area) can be important when displaying data.
- There's no magic aspect ratio formula, try a few and make sure to select one that clearly represents the data