MPA 634  
Data Science for Managers  
Midterm I: Fall 2019

# I. Definitions and Concepts

1. Compare and contrast hypothesis generation and hypothesis confirmation.

**Hypothesis generation** is part of the data exploration process whereby we use transformations, visualization, and modeling to generate questions. We then look for answers to these questions using our data. This investigation in turn spawns more questions and subsequent investigations and models. With hypothesis generation, our data gets recycled and used repeatedly.

**Hypothesis confirmation** is the process whereby we collect a data set with the express purpose of testing an assertion or precise mathematical model. The data can only be used in one analysis.

The **key difference** is how many times you use your data. As soon as you use your data more than once, you are a data explorer.

1. Define and illustrate all seven parts of the grammar of graphics by outlining a script that creates bar graphs from the diamonds data frame. Remember that possible variables in the diamonds data frame are price, carat, cut, color, and clarity.

|  |  |
| --- | --- |
| **Data**: Identify the data frame used in the graphic | diamonds %>% |
| **Aesthetics**: Assignment of values to the elements that comprise a graph. This includes assigning variables to the x-axis, y-axis, color, fill, shape, linetype, and transparency. The assignment can occur using values of a variable within and aes or can be assigned arbitrary values | ggplot(aes(x = cut, fill = clarity) %>% |
| **Geometric Objects**: Creation of layers in graph | geom\_bar(position = “dodge”) |
| **Stats**: calculations needed to create graphs from the data | In order to draw the graph, the we must first count how many diamonds are in each cut\clarity combination. |
| **Position**: jitter in geom\_point and identity, fill, and dodge with geom\_bar and geom\_col | position = “dodge” creates a side by side bar chart |
| **Coordinate System**: switch axes or choose a different coordinate system | coord\_flip() creates a horizontal rather than vertical bar chart |
| **Facet**: Create multiple graphs based on a categorical variable | Facet\_grid(rows = vars(color)) which creates a separate bar chart for each diamond color |

1. Use Venn diagrams to define and illustrate the meaning of "&", "|", and "!".
2. Locate Tukey's five number summary on the following boxplot.



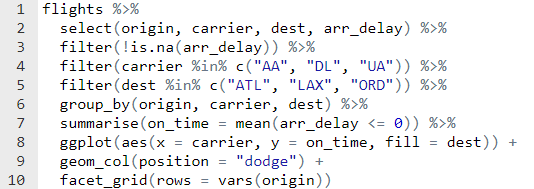
1. Use the following violin plots to compare the shapes of the distributions of highway miles per gallon for suv and subcompact vehicles .



* 1. Location
  2. Scale
  3. Symmetry
  4. Outliers

# II. Line by Line Code Interpretation

## Code Chunk 1



## Code Chunk II

