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.



Minimum

First Quartile

Median

Third Quartile

Maximum

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

The location or center is measured by the median or the middle line with the box. This shows that the location of distribution is larger for subcompact cars than suv vehicles.

* 1. Scale

The scale or spread is measures by the Interquartile range which is the length of the rectangle. This shows that the mileage of suv’s is more compact and less spread out than subcompacts.

* 1. Symmetry  
       
     Insights into symmetry come from the position of the median in the box, the length of the comparative length of the whiskers, and the number of outliers in the tails. Since we don’t have the mean to compare with the median, we can’t use this.

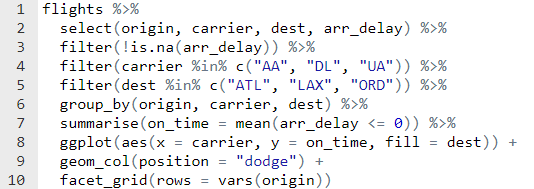
Both suv’s and subcompacts seem to be positively skewed. This is natural since there is a natural lower bound at zero. For both suv’s and subcompacts the medians locate closer to the lower hinge which is evidence of positive skewness. The whiskers are approximately the same length this isn’t evidence of skewness. Both outliers in the right tail, so this is consistent with positive skewness.

* 1. Outliers

Suv’s have four outliers so this could be an indicator of thick tails. Subcompacts have two outliers but they are a large distance from the median, so this might also indicate thick tails.

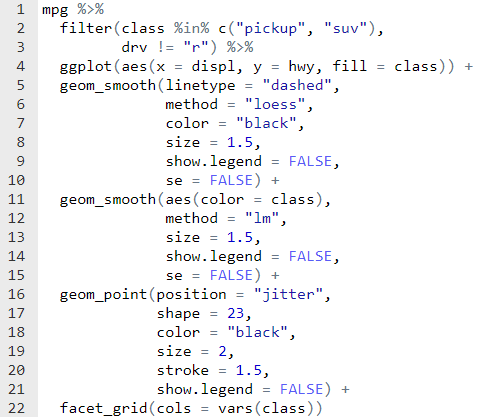
# II. Line by Line Code Interpretation

## Code Chunk 1



1. Specifies that we are using flights as our data source.
2. Creates a new tibble that only contains the columns origin, carrier, dest, and arr\_delay
3. Retains only that rows in the new tibble for which arr\_delay is not missing. The !is.na is interpreted as not missing since is.na means missing.
4. Chooses a further subset of rows there the airline carrier is American, Delta, or United. The %in% means AA or DL or UA
5. Chooses only those observations that fly to Atlanta, Los Angeles, or Chicago. The %in% means ATL or LAX or ORD.
6. Anticipates that we want to summarize our data by all combinations of origin, carrier, and destination. This means three origins (JFK, Newark, and La Guardia), three airlines (American, Delta, and United), and three destinations (Atlanta, Los Angeles, and Chicago). This gives 27 different possibilities.
7. Creates a new variable that is the proportion of flights that arrive on time for each origin, airline, and destination combination. It does that by first making a comparison of arrival delays with 0. The result of this comparison is a new column that is a logical type, where TRUE is 1 and FALSE is 0. The mean function adds up these values and divides by the sample size. This gives a proportion of flights that arrive on time. The result is a tibble that has four columns: origin, carrier, and destination, and the proportion.
8. The tibble from 7 comes into ggplot which assigns carrier (American, Delta, or United) to the x axis and the on\_time proportion to the y-axis. We anticipate a bar chart so we fill our bars with the different categories that correspond to the destination (Atlanta, Los Angeles, or Chicago).
9. We use geom\_col because we already have counts. We don’t need a corresponding stats to complete calculations because we already did that. We choose dodge for the position because we want the bars for each destination to be side by side rather than stacked.
10. Creates a separate graph each origin airport. The rows in the facet\_grid correspond to each airport.

## Code Chunk II



1. Specifies the data frame mpg as the source of our information and data
2. Subsets the data frame so that it only includes the rows for either pickup trucks or suv’s.
3. Excludes the rear wheel drive vehicles. This could also have been accomplished with drv == “f” | drv == “4”. Either way we only want front and four-wheel drive vehicles.
4. Sets up the aesthetics for ggplot by assigning displacement to the x-axis and highway miles per gallon to the y-axis. If we anticipate a hollow plotting shape in a scatterplot, then we choose to color the fill based on the class of the vehicle which is either a pickup or suv.
5. This is the first smoothing layer and will use a dashed linetype. The linetype is an arbitrary value so all smoothed lines from this geom will be the same.
6. The lines for this layer are calculated using the loess smooth.
7. The color of the lines is black
8. The size of the line is specified as 1.5
9. The legend for this layer is suppressed. It is necessary to suppress the legend cause of the fill = class in line 4
10. The loess line doesn’t have a confidence interval  
    In summary, each loess line drawn by this layer is dashed, black, has a width 2.5, and doesn’t have a confidence interval.
11. Create the second smoothing layer. The aes(color = class) means that a separate colored line will be drawn for each class (pickup or suv)
12. Identifies the smoothing method as a simple regression.
13. Specifies that the width of the line be 1.5
14. Suppresses the legend for the same reasons as line 9
15. Suppresses the confidence interval for the regression lines  
    In summary, this geom creates a layer with a simple regression live that is colored by class, has a width of 1.5, and no confidence interval
16. Creates the third layer which is the scatterplot and jitters the points with the position qualifier so that we can see all of the points, even if they overlap
17. Chooses a shape of 23 which is the hollow diamond shown in the result graph. The fill for the hollow diamond comes the the fill = class in line 4
18. Color = “black” outlines the hollow shape in black
19. The size of each diamond is 2
20. The border of each diamond is 1.5  
    In summary, this geom creates a scatterplot with a hollow diamond filled with color according to the class of the vehicle and a black border.
21. Once again we suppress the legend that would have been drawn because of the fill = class in line 4
22. Creates a row of two charts (2 columns), one for pickup trucks and one for suv’s.