MPA 634  
Data Science for Managers  
Midterm I: Winter 2019

# I. Definitions and Concepts

1. Define standardized variables, covariance, and correlation. Explain how they are related to each other.  
     
   **standardized variables**  
     
   A standardized variable has a mean of zero and a standard deviation of 1. This is very similar to a z-score. In order to standardize a variable, we subtract the mean and divide by the standard deviation. The scale function in R standardizes a variable and creates a unitless measure. The notation for standardizing a variable is:

   
**covariance**  
  
A deviation is defined as the distance of an observation from its mean. The covariance is the average of the products of the deviations of two variables from their means. This is given by:



The problem with a covariance is that it retains its units of measure so it is very difficult to interpret.  
  
**correlation**  
  
The correlation divides each deviation by the standard deviation and then calculates the average product using the following formula:

   
  
The advantage of correlation is that is doesn’t have any units so we can readily judge the strength of a linear relationship by evaluating the value of the correlation coefficient which has to be in the range of -1 to 1.  
  
**relationship among standardized variables, covariance, and correlation**We recognize the  and  as z-scores. Therefore, the correlation is the covariance between two z-scores. It is a unitless measure.

1. Carefully explain how the whiskers of a boxplot are constructed. How do whiskers help us identify outliers?  
     
   To construct the lower or left hand whisker, we measure is 1.5 times the interquartile range below the lower hinge of the boxplot and then move back towards the box until we encounter a data point. The lower hinge of the box is the first quartile.  
     
   Similarly, to construct the upper or right hand whisker, we measure is 1.5 times the interquartile range above the upper hinge of the boxplot and then move back towards the box until we encounter a data point. The upper hinge is the third quartile.  
     
   Those points that lie to the left of the lower whisker or to the right of the upper whisker are designated as outliers.
2. Define each of the three parts of exploratory data analysis.  
   1. Transformation  
        
      We can transform our data in three different ways.  
        
      i) We can perform arithmetic operations such as logarithm, square root on existing variables or create new variables. We use mutate and transmute to accomplish this.  
        
      ii) We can select subsets of the observations or variables using filter and select.

iii) Calculation of summary statistics. In this case, we use summarize.

* 1. Visualization  
       
     Insightful pictures and graphs of data that help us formulate questions. When what we see doesn’t match with what we expected, then visualization often provokes new questions and directions for our investigation.
  2. Modeling  
       
     Once we have discovered patterns in our data from visualization, then we can formalize them into a mathematical or computational model. The purpose of the model is to extract patterns from the data. We iterate through this process until we have found all of the patterns in the data.  
       
     These three components are not a set of linear steps but combine in a web of interconnections that allows us to formulate new questions and seek new insightful answers.

1. Explain how geoms and stats are related to each other in the layered grammar of graphics. Illustrate your answer with an example.  
     
   Geometric objects or geoms create graph layers in ggplots. Sometimes they can directly map original quantitative and qualitative information into aesthetics without other calculations.  
     
   In ggplot, stats are when further calculations are needed in order to create graphs. Often times geoms are associated with unique stats. For example, geom\_bar includes the calculation of counting frequencies in order to create a bar graph from original data. This contrasts with geom\_col which expects that the calculations have already occurred.  
     
   The geom\_smooth also has inherent underlying calculations such as regression and loess.
2. **Compare** the location, scale, symmetry, and outliers of city and highway mileage using the following information:

A close up of a map

Description automatically generated

# A tibble: 2 x 5

type\_of\_driving Mean Median Standard\_Deviation Interquartile\_Range

*<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

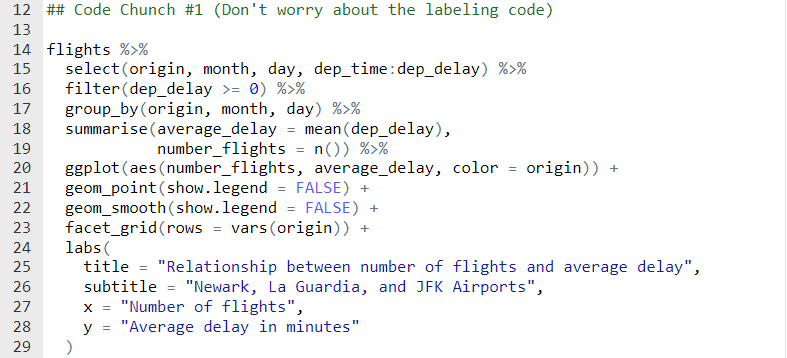
1 cty 16.9 17 4.26 5

2 hwy 23.4 24 5.95 9

* 1. Location  
       
     The center of the distribution is larger for highway than city as indicated by the means and medians. In the boxplot, the median in the line in the middle of the box.
  2. Scale  
       
     More variability exists for highway than city. Both the standard deviation and interquartile range are larger for highway. The length of the box or rectangle in the boxplot corresponds to the interquartile range.
  3. Symmetry  
       
     Visually assessing symmetry gives conflicting results:  
       
     i) The median locates closer to the upper hinge in both cases which would suggest negative skewness.  
       
     ii) The right whisker is longer than the left whisker which suggests positive skewness.   
       
     iii) Both highway and city mileage have large observations or outliers outside of the whiskers. This suggests positive skewness.
  4. Outliers  
       
     Outliers are those points that lay beyond the whiskers.

# II. Line by Line Code Interpretation

Code Chunk 1 (Don’t worry about the labeling code)



Line 15: Choose a subset of variables. The dep\_time:dep\_delay notation means choose all of the variables in the tibble between dep\_time and dep\_delay.

Line 16: Choose only those observations that were delayed. If the flight left early, then it won’t be chosen.

Line 17: Specifies that we want information for every possible combination of origin airport and day in the tibble. The day is identified by the month and day variables.

Lines 18-19: Calculates the average delay and number of flights for each day and airport. This collapses the data into a much smaller tibble.

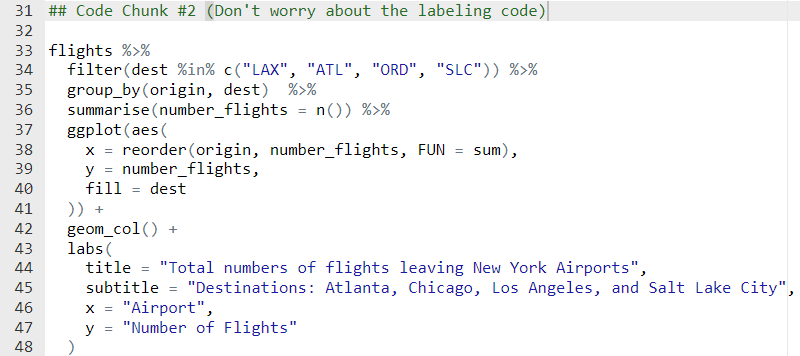
Line 20: Assigns number of flights to the x-axis, average departure delay to the y-axis, and colors the points and lines based on the origin airport.

Line 21: Creates a scatterplot with no legend

Line 22: Creates a locally weighted scatterplot smoothed line on the graph.

Line 23: Creates a series of three graphs that are stacked on top of each other, one for each of the origin airports.

Code Chunk II (Don’t worry about the labeling code)



Line 34: Chooses the observations for flights going to Los Angeles (LAX), Atlanta (ATL), Chicago (ORD), **or** Salt Lake City (SLC).

Line 35: group\_by specifies that we are going to complete calculations for each origin and destination combination.

Line 36: counts the number of flights going from each NYC airport to each of the possible destinations.

Lines 37 – 41: Assigns origin to the x-axis, number of flights to the y-axis, and fills the geometric objects with color based on the destination.

Line 42: Creates a bar chart using geom\_col. We use geom\_col rather than geom\_bar because we already counted using the summarise statement.