MPA 634 Key  
Data Science for Managers  
Midterm I: Winter 2021

# I. Definitions and Concepts

1. Compare and contrast filter, mutate, and summarize. Your answer should include an explanation of how each one affects data frames or tibbles.  
     
   Filter: chooses a subset of rows

Mutate: creates a new column and puts the result at the end of the tibble

Summarise: collapses the information into a tibble and creates new variables based on the summary function (mean, median, sd, IQR, etc) called in the calculation part of the summarise statement.

1. Identify and explain at least **three** different ways that you can use a violin plot to determine whether or not a distribution is positively or negatively skewed.

Check the position of the median within the box. If it is in the middle, that is an indication that we have a symmetric distribution. If it is closer to the lower or left hinge, then it indicates a positively or right skewed distribution. If it is closer to the upper or right hinge, then it suggests a negatively or left skewed distribution.

Compare the position of the median and the mean. Similar mean and medians correspond to symmetrical distributions. A mean above the median corresponds to a positively skewed distribution and a mean below the median associates with a negatively skewed distribution.

Length of whiskers also can be used to judge symmetry. Equal length means symmetric, right larger than left corresponds to positive skewness, and left longer than right suggests negative skewness.

Number of outliers is a fourth indicator. Balance in the number of outliers in the left and right tails corresponds to symmetry. More outliers in the right tail happens with positive skewness and more outliers in the left tail suggests negative skewness.

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. Explain each of the seven parts of the grammar of graphics by writing a script that illustrates the definition of each part.

|  |  |
| --- | --- |
| **Data**: Identify the data frame or tibble 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 an 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, 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. **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

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

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

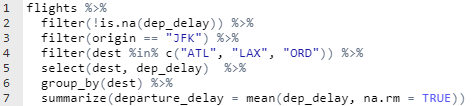
1 cty 16.9 17 4.26 5

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 on the right side. This suggests positive skewness.
  4. Outliers  
       
     Outliers are those points that lay beyond the whiskers. There are very few outliers in this data.

# II. Line by Line Code Interpretation (Don’t interpret the first line)

Code Chunk I



Line 2: This chooses all of the flights that have a recorded departure delay. The is.na function is true for those observations that are missing. The ! turns all of the false values into true values which is what we want, those flights which do have a recorded departure time.

Line 3: Chooses the rows or flights that leave from the JFK airport

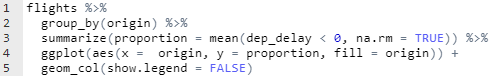
Line 4: Chooses only those flights that go to Atlanta, Los Angeles, or Chicago

Line 5: Chooses only the dest and dep\_delay variables to put into the new tibble.

Line 6: Informs R that we are interested in results for each separate destination

Line 7: Collapses the information into a tibble that has a row for each separate destination and then calculates the mean departure delay after removing any missing observations. The na.rm isn’t actually necessary since that was accomplished in line 2.

Code Chunk II



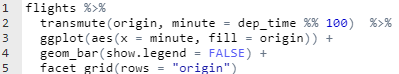
Line 2: Give a summary of the data for each one of the three NYC airports.

Line 3: Calculates the percentage of the flights that leave early. The dep\_delay < 0 creates a logical variable which is 1 when it is true and 0 otherwise. The average then sums these values and divides by the sample size. This gives the proportion.

Line 4: Gives the aesthetics by assigning airport to the x axis and the proportion calculated in line 3 to the y-axis. The bars that come in the next step are filled by mapping fill colors based on origin.

Line 5: We use geom\_col because we already did the calculations needed to draw a bar chart in the summarize step. We don’t need a legend so we suppress it.

Code Chunk III



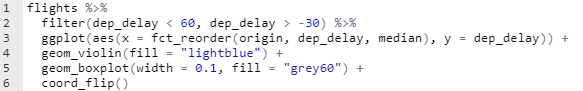
Line 2: Transmute is a combination of select and mutate. In this case it selects origin and then creates a new variable called minute, where this new variable is the result of the modulus function %%. By taking the remainder after dividing by 100, we are able to drop the hour from dep\_time.

Line 3: Informs ggplot that we want a bar chart for the minute variable and we would like to fill our bars with a color scheme that depends on the origin airport.

Line 4: The geo\_bar function counts the number of flights for each minute of the day for each airport. We don’t need to see the legend because it is redundant.

Line 5: Creates a separate bar graph for each of the different airports.

Code Chunk IV



Line 2: Chooses only those flights that have departure delays between -30 and 60. The comma in this case means and.

Line 3: Anticipating the violin plot that follows, we need a categorical variable (origin) and a numerical variable (dep\_delay). We would like to order the categorical variable by the median of the numerical variable. That is what the fct\_recorder() part of this command accomplishes.

Line 4: Adds a violin plot base layer to the graph and fill with the lightblue color

Line 5: Adds a boxplot layer with a smaller width so we can see the violin plot beneath. The boxplot is filled with a grey color.

Line 6: Alters the coordinate system to give a horizontal orientation by switching the x and y axes.

Code Chunk V



Line 2: Communicates that we want to statistics for each airport in the origin variable

Line 3: Calculates the number of distinct destinations for each airport. The result is a tibble with 3 lines.

Line 4: Arranges the resulting tibble from the largest number of destinations for the smallest.