Homework #6

Select and Mutate - Derrik Adams

## Concepts

1. Compare and contrast filter and select.
2. Explain how “:” can be used when selecting variables.

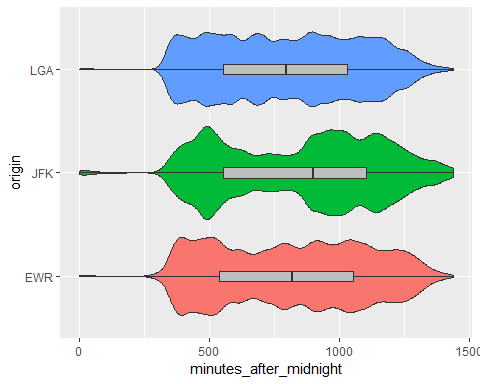
### Concept Answers

1. Filter is used to select observations with certain values (so essentially rows of data), while select is used to select variables, or specific columns of data.
2. The “:” can be used to select ranges of variables. So you could say 1:5 without having to say to select 1, 2, 3, 4, and 5.

## Interpretation of code

Interpret the following R code:

converted\_times <- mutate(  
 flights,  
 hours = dep\_time %/% 100,  
 minutes = dep\_time %% 100,  
 minutes\_after\_midnight = 60 \* hours + minutes  
)  
ggplot(  
 data = converted\_times,  
 mapping = aes(x = origin,  
 y = minutes\_after\_midnight,  
 fill = origin)  
) +  
 geom\_violin(show.legend = FALSE) +  
 geom\_boxplot(width = 0.1,  
 fill = "grey",  
 show.legend = FALSE) +  
 coord\_flip()



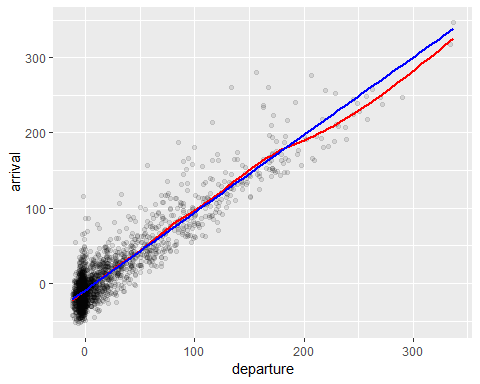
### Code Interpretation

28/29 - Create new data called converted\_times that adds onto flights data  
30 - Converts hours into just the hour part using integer division on dep\_time (e.g. 517 is now just 5)  
31 - Converts minutes into just the minute part using modulo on dep\_time (e.g. 517 is now just 17)  
32 - Calculates the number of minutes after midnight by using hours and minutes variables  
33/34/35 - Creates new plot based on the converted times data  
36 - Maps the origin of the flights to the x-axis aesthetic 37 - Maps the minutes after midnight to the y-axis aesthetic  
38 - Color the plot by origin  
39/40 - Create violin plot without the legend  
41 - Create a boxplot with width of .1  
42 - Fill the boxplot with color grey to standout on the violin plot  
43 - Hide the legend of the boxplot  
44 - Flip the coordinate axis

## Coding Problem 1:

1. Select only those flights going from JFK to Chicago.
2. Create a data frame that includes the following variables: dep\_delay and arr\_delay.
3. Rename the variables dep\_delay to departure and arr\_delay to arrival.
4. Draw a scatterplot of arrival on departure with colored loess and linear regression smoothers without confidence intervals.

#Not sure which Chicago airport is being referenced or if all of them are, so just include all  
JFK\_to\_CH <- filter(flights, (origin == "JFK" & dest == "CGX") | ( origin == "JFK" & dest == "CHI") | ( origin == "JFK" & dest == "DPA") | ( origin == "JFK" & dest == "MDW") | ( origin == "JFK" & dest == "ORD") | ( origin == "JFK" & dest == "PWK") | ( origin == "JFK" & dest == "ZUN"))  
  
JFK\_to\_CH\_delay <- select(JFK\_to\_CH, dep\_delay, arr\_delay)  
  
JFK\_to\_CH\_delay <- rename(JFK\_to\_CH\_delay, departure = dep\_delay, arrival = arr\_delay)  
  
ggplot(data = JFK\_to\_CH\_delay, mapping = aes(x = departure, y = arrival))+  
 geom\_point(alpha = .1)+  
 geom\_smooth(color = "red", se = FALSE)+  
 geom\_smooth(color = "blue", method = lm, se = FALSE)



## Coding Problem 2: Use mutate and the ranking function to find the 5 most delayed flights departing from JFK.

1. Choose the flights departing from JFK
2. Create a new variable called delay which is the rank of the delay with 1 being the longest delay.
3. For the five most delayed flights, print the new variable delay along with year, month, day, dest, carrier, and dep\_delay with the most delayed at the top of the table.

JFK\_flights <- filter(flights, origin == "JFK")  
  
JFK\_flights <- mutate(JFK\_flights,  
 delay = min\_rank(desc(dep\_delay)))  
  
JFK\_flights\_selected <- select(JFK\_flights, delay, year, month, day, dest, carrier, dep\_delay)  
  
top\_flights\_arranged <- arrange(JFK\_flights\_selected, delay)  
  
head(top\_flights\_arranged, 5)

## # A tibble: 5 x 7  
## delay year month day dest carrier dep\_delay  
## <int> <int> <int> <int> <chr> <chr> <dbl>  
## 1 1 2013 1 9 HNL HA 1301  
## 2 2 2013 6 15 CMH MQ 1137  
## 3 3 2013 9 20 SFO AA 1014  
## 4 4 2013 7 22 CVG MQ 1005  
## 5 5 2013 4 10 TPA DL 960