Assignment: Team Diversity Analytics

Dunja Novaković

2020-07-20

Instructions

This assignment reviews my *Team Diversity* analytical lecture. You will use the *team_diversity.Rmd* file I reviewed in the video lectures to complete this assignment. You will *copy and paste* relevant code from that file and update it to answer the questions in this assignment. You will respond to questions in each section after executing relevant code to answer a question. You will submit this assignment to its *Submissions* folder on *D2L*. You will submit this *(1)* completed **R Markdown** script and *(2)* a *HTML* or *PDF* rendered version of it to *D2L* by the due date and time. If you installed `TinyTeX` successfully, then I prefer a *PDF* version.

To start:

For any analytical project, you want to create a clear project directory structure.
All materials from this course should exist in one folder on your computer. Inside of that main course folder, you should create folders to store course documentation, lecture analytical projects, assignments analytical projects, etc. Inside of your folder for assignments analytical projects, you should create folder for this assignment named *team_diversity*.

Any analytical project folder should contain inside it at least three additional folders named *scripts*, *data*, and *plots*. Store this script in the *scripts* folder, the data for this assignment in the *data* folder, and any requested plots in the *plots* folder. Each analytical project should also contain a **.Rproj** file in its top-level directory. Go to the *File* menu in *RStudio*, select *New Project...*, choose *Existing Directory*, go to the folder you created to contain this analytical project. Select it as the top-level directory for this **RStudio Project**.

Global Settings

The first code chunk sets the global settings for the remaining code chunks in the document. Do *not* change anything in this code chunk.

Load Packages

In this code chunk, we load three packages we need for this assignment:

1. **here**,
2. **tidyverse**, and
3. **readxl**.

We will use functions from these packages to import the data, examine the data, calculate summaries on the data, and create visualizations from the data. Do *not* change anything in this code chunk.

```
## load libaries for use in current working session
## here for workflow
library(here)

## tidyverse for data manipulation and plotting
# loads eight different libraries simultaneously
library(tidyverse)

## readxl to import Excel data
library(readxl)
```

Task 1: Load Data

As your first task for this assignment, you need to load the data of interest for the assignment. We will use the same data as in the analytical lecture: **team_diversity.xlsx**.

Use the **read_excel** function from the **readxl** package and the **here** function from the **here** package to load the data for this working session. Import the data as the object **team_div_orig_data**. Make a copy of the data named **team_div_work_data**. Use **glimpse** to preview **team_div_work_data**.

Question 1.1: How many observations and variables are there in the data initially?

Response 1.1: *Number of observations: 927. Number of variables: 25.*

Update **team_div_work_data** by converting the **Location** and **Function** numeric variables to factor variables. Add the correct labels for the levels of both factors. Execute both parts in one chained command using the exact code from the analytical lecture. Apply **glimpse** to the updated **team_div_work_data**.

Question 1.2: What is the **Location** of the first team? What is the **Function** of the first team?

Response 1.2: *Location: Central London. Function: Prof. Service.*

Update **team_div_work_data** by using the exact code from the analytical lecture to:

1. compute a new variable named **London** from the existing variable **Location**,
2. overwrite **BAME** by multiplying it by *100*,
3. compute average *engagement*, *integrity*, and *supervisor support*.

Apply **glimpse** to the updated **team_div_work_data**.

Question 1.3: What is the **integrity** score for the second team? What is the **suprvsr_supp** for the the second team?

Response 1.3: *integrity: 84.75; suprvsr_supp:76.00.*

```
#### Q1.1
### import data
## use here() to locate file in our project directory;
## use read_excel to import the Excel data file
team_div_orig_data <- read_excel(here("data", "team_diversity.xlsx"))

## make a working copy of the data
team_div_work_data <- team_div_orig_data

## glimpse data
glimpse(team_div_work_data)
```

```
## Rows: 927
## Columns: 25
## $ DepartmentGroupNumber <dbl> 18, 19, 29, 30, 35, 37, 44, 45, 47, 50, 51, 63, ~
## $ GroupSize             <dbl> 18, 28, 11, 17, 32, 60, 23, 36, 32, 70, 52, 15, ~
## $ PercentMale           <dbl> 65, 67, 33, 32, 18, 65, 15, 26, 65, 5, 75, 50, 5~
## $ BAME                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ NumberTeamLeads       <dbl> 3, 4, 2, 2, 5, 9, 3, 5, 5, 10, 7, 2, 4, 3, 9, 7,~
## $ NumberFeMaleTeamLeads <dbl> 1, 0, 0, 0, 3, 0, 1, 0, 0, 2, 0, 0, 1, 0, 1, 0, ~
## $ Location              <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3, 2, 2, ~
## $ Function              <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ EMPsurvEngage_1       <dbl> 100, 96, 100, 100, 81, 97, 91, 100, 91, 97, 92, ~
## $ EMPsurvEngage_2       <dbl> 100, 86, 91, 100, 69, 97, 100, 94, 88, 97, 94, 1~
## $ EMPsurvEngage_3       <dbl> 94, 86, 100, 100, 72, 87, 100, 89, 91, 93, 98, 9~
## $ EMPsurvEngage_4       <dbl> 78, 54, 82, 94, 44, 92, 87, 97, 81, 90, 98, 87, ~
## $ EMPsurvEngage_5       <dbl> 100, 96, 100, 100, 81, 97, 100, 97, 97, 99, 94, ~
## $ EMPsurvEngage_6       <dbl> 94, 57, 82, 88, 44, 78, 87, 94, 72, 90, 88, 87, ~
## $ EMPsurvEngage_7       <dbl> 89, 89, 100, 100, 84, 92, 96, 97, 94, 99, 98, 10~
## $ EMPsurvEngage_8       <dbl> 72, 82, 73, 88, 38, 55, 91, 69, 69, 80, 69, 80, ~
## $ EMPsurvEngage_9       <dbl> 78, 82, 91, 94, 47, 83, 96, 78, 81, 93, 85, 93, ~
## $ EMPorgIntegrity1      <dbl> 72, 75, 73, 100, 59, 87, 100, 61, 69, 69, 67, 60~
## $ EMPorgIntegrity2      <dbl> 78, 86, 91, 100, 81, 92, 100, 89, 91, 94, 92, 10~
## $ EMPorgIntegrity3      <dbl> 89, 89, 91, 100, 88, 93, 100, 89, 97, 97, 94, 10~
## $ EMPorgIntegrity4      <dbl> 89, 89, 91, 100, 78, 80, 100, 81, 78, 80, 80, 93~
## $ EMPsurvSUP1           <dbl> 72, 79, 73, 100, 81, 88, 100, 75, 78, 96, 98, 87~
## $ EMPsurvSUP2           <dbl> 89, 79, 91, 100, 78, 92, 100, 83, 94, 96, 94, 93~
## $ EMPsurvSUP3           <dbl> 89, 64, 73, 100, 72, 78, 96, 78, 72, 90, 94, 100~
## $ EMPsurvSUP4           <dbl> 83, 82, 73, 100, 75, 87, 100, 78, 84, 91, 98, 80~
```

```
#### Q1.2
### change particular variables to factors
## overwrite working data
team_div_work_data <- team_div_work_data %>%
  ## change Location and LondonorNot to factor variables
  mutate_at(vars(Location, Function), as_factor) %>%
  ## recode levels for both factors
        # recode levels for Location factor
  mutate(Location = fct_recode(Location,
                     # change 1 to Central London
                     `Central London` = "1",
                     # change 2 to Greater London
                     `Greater London` = "2",
                     # change 3 to Rest of UK
                     `Rest of UK` = "3"),
         # recorde levels for Function factor
         Function = fct_recode(Function,
                     # change 1 to Sales Staff
                     `Sales` = "1",
                     # change 2 to Professional Service
                     `Prof. Service` = "2"))

## Examine data
glimpse(team_div_work_data)
```

```
## Rows: 927
## Columns: 25
## $ DepartmentGroupNumber <dbl> 18, 19, 29, 30, 35, 37, 44, 45, 47, 50, 51, 63, ~
## $ GroupSize             <dbl> 18, 28, 11, 17, 32, 60, 23, 36, 32, 70, 52, 15, ~
## $ PercentMale           <dbl> 65, 67, 33, 32, 18, 65, 15, 26, 65, 5, 75, 50, 5~
## $ BAME                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ NumberTeamLeads       <dbl> 3, 4, 2, 2, 5, 9, 3, 5, 5, 10, 7, 2, 4, 3, 9, 7,~
## $ NumberFeMaleTeamLeads <dbl> 1, 0, 0, 0, 3, 0, 1, 0, 0, 2, 0, 0, 1, 0, 1, 0, ~
## $ Location              <fct> Central London, Central London, Central London, ~
## $ Function              <fct> Prof. Service, Prof. Service, Prof. Service, Pro~
## $ EMPsurvEngage_1       <dbl> 100, 96, 100, 100, 81, 97, 91, 100, 91, 97, 92, ~
## $ EMPsurvEngage_2       <dbl> 100, 86, 91, 100, 69, 97, 100, 94, 88, 97, 94, 1~
## $ EMPsurvEngage_3       <dbl> 94, 86, 100, 100, 72, 87, 100, 89, 91, 93, 98, 9~
## $ EMPsurvEngage_4       <dbl> 78, 54, 82, 94, 44, 92, 87, 97, 81, 90, 98, 87, ~
## $ EMPsurvEngage_5       <dbl> 100, 96, 100, 100, 81, 97, 100, 97, 97, 99, 94, ~
## $ EMPsurvEngage_6       <dbl> 94, 57, 82, 88, 44, 78, 87, 94, 72, 90, 88, 87, ~
## $ EMPsurvEngage_7       <dbl> 89, 89, 100, 100, 84, 92, 96, 97, 94, 99, 98, 10~
## $ EMPsurvEngage_8       <dbl> 72, 82, 73, 88, 38, 55, 91, 69, 69, 80, 69, 80, ~
## $ EMPsurvEngage_9       <dbl> 78, 82, 91, 94, 47, 83, 96, 78, 81, 93, 85, 93, ~
## $ EMPorgIntegrity1      <dbl> 72, 75, 73, 100, 59, 87, 100, 61, 69, 69, 67, 60~
## $ EMPorgIntegrity2      <dbl> 78, 86, 91, 100, 81, 92, 100, 89, 91, 94, 92, 10~
## $ EMPorgIntegrity3      <dbl> 89, 89, 91, 100, 88, 93, 100, 89, 97, 97, 94, 10~
## $ EMPorgIntegrity4      <dbl> 89, 89, 91, 100, 78, 80, 100, 81, 78, 80, 80, 93~
## $ EMPsurvSUP1           <dbl> 72, 79, 73, 100, 81, 88, 100, 75, 78, 96, 98, 87~
## $ EMPsurvSUP2           <dbl> 89, 79, 91, 100, 78, 92, 100, 83, 94, 96, 94, 93~
## $ EMPsurvSUP3           <dbl> 89, 64, 73, 100, 72, 78, 96, 78, 72, 90, 94, 100~
## $ EMPsurvSUP4           <dbl> 83, 82, 73, 100, 75, 87, 100, 78, 84, 91, 98, 80~
```

```
#### Q1.3
### compute new variables from existing variables
## overwrite working data
team_div_work_data <- team_div_work_data %>%
         ## compute new factor variable from existing factor variable
  mutate(London = fct_collapse(Location,
                    # combine two factor levels into one
                    London = c("Central London", "Greater London")),
         ## compute BAME as percentage
         BAME = BAME*100,
         ## compute average engagement score
         engage = rowMeans(select(., EMPsurvEngage_1:EMPsurvEngage_9)),
         ## compute average integrity score
         integrity = rowMeans(select(., EMPorgIntegrity1:EMPorgIntegrity4)),
         ## compute average supervisor support
         suprvsr_supp = rowMeans(select(., EMPsurvSUP1:EMPsurvSUP4)))

## Examine data
glimpse(team_div_work_data)
```

```
## Rows: 927
## Columns: 29
## $ DepartmentGroupNumber <dbl> 18, 19, 29, 30, 35, 37, 44, 45, 47, 50, 51, 63, ~
## $ GroupSize             <dbl> 18, 28, 11, 17, 32, 60, 23, 36, 32, 70, 52, 15, ~
## $ PercentMale           <dbl> 65, 67, 33, 32, 18, 65, 15, 26, 65, 5, 75, 50, 5~
## $ BAME                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ NumberTeamLeads       <dbl> 3, 4, 2, 2, 5, 9, 3, 5, 5, 10, 7, 2, 4, 3, 9, 7,~
## $ NumberFeMaleTeamLeads <dbl> 1, 0, 0, 0, 3, 0, 1, 0, 0, 2, 0, 0, 1, 0, 1, 0, ~
## $ Location              <fct> Central London, Central London, Central London, ~
## $ Function              <fct> Prof. Service, Prof. Service, Prof. Service, Pro~
## $ EMPsurvEngage_1       <dbl> 100, 96, 100, 100, 81, 97, 91, 100, 91, 97, 92, ~
## $ EMPsurvEngage_2       <dbl> 100, 86, 91, 100, 69, 97, 100, 94, 88, 97, 94, 1~
## $ EMPsurvEngage_3       <dbl> 94, 86, 100, 100, 72, 87, 100, 89, 91, 93, 98, 9~
## $ EMPsurvEngage_4       <dbl> 78, 54, 82, 94, 44, 92, 87, 97, 81, 90, 98, 87, ~
## $ EMPsurvEngage_5       <dbl> 100, 96, 100, 100, 81, 97, 100, 97, 97, 99, 94, ~
## $ EMPsurvEngage_6       <dbl> 94, 57, 82, 88, 44, 78, 87, 94, 72, 90, 88, 87, ~
## $ EMPsurvEngage_7       <dbl> 89, 89, 100, 100, 84, 92, 96, 97, 94, 99, 98, 10~
## $ EMPsurvEngage_8       <dbl> 72, 82, 73, 88, 38, 55, 91, 69, 69, 80, 69, 80, ~
## $ EMPsurvEngage_9       <dbl> 78, 82, 91, 94, 47, 83, 96, 78, 81, 93, 85, 93, ~
## $ EMPorgIntegrity1      <dbl> 72, 75, 73, 100, 59, 87, 100, 61, 69, 69, 67, 60~
## $ EMPorgIntegrity2      <dbl> 78, 86, 91, 100, 81, 92, 100, 89, 91, 94, 92, 10~
## $ EMPorgIntegrity3      <dbl> 89, 89, 91, 100, 88, 93, 100, 89, 97, 97, 94, 10~
## $ EMPorgIntegrity4      <dbl> 89, 89, 91, 100, 78, 80, 100, 81, 78, 80, 80, 93~
## $ EMPsurvSUP1           <dbl> 72, 79, 73, 100, 81, 88, 100, 75, 78, 96, 98, 87~
## $ EMPsurvSUP2           <dbl> 89, 79, 91, 100, 78, 92, 100, 83, 94, 96, 94, 93~
## $ EMPsurvSUP3           <dbl> 89, 64, 73, 100, 72, 78, 96, 78, 72, 90, 94, 100~
## $ EMPsurvSUP4           <dbl> 83, 82, 73, 100, 75, 87, 100, 78, 84, 91, 98, 80~
## $ London                <fct> London, London, London, London, London, London, ~
## $ engage                <dbl> 89.44444, 80.88889, 91.00000, 96.00000, 62.22222~
## $ integrity             <dbl> 82.00, 84.75, 86.50, 100.00, 76.50, 88.00, 100.0~
## $ suprvsr_supp          <dbl> 83.25, 76.00, 77.50, 100.00, 76.50, 86.25, 99.00~
```

Task 2: Team Supervisor Support and Engagement

For your second task, you will examine **suprvsr_supp** (supervisor support).

Plot the empirical density distribution of **suprvsr_supp** for both levels of **Function**. Correctly label all axes and legends.

Question 2.1: Is the most frequent value for *supervisor support* higher for the *professional service* or *sales* function? Do the two distributions overlap a lot or a little?

Response 2.1: *The most frequent value for supervisor support is higher for the professional service function. The two distributions overlap a lot.*

Create a scatterplot to show the relationship between *supervisor support* (i.e., **suprvsr_supp**) and *engagement* (i.e., **engage**). Include all of the following in this one plot:

1. place *supervisor support* on the x-axis and *engagement* on the y-axis,
2. color the points by **Function**,
3. use a *loess* line to fit the data points for each **Function**, and
4. write a *descriptive title* for the plot.

Question 2.2: Do *high* values of *supervisor support* tend to associate with *high* or *low* values of *engagement*? Does the relationship between *supervisor support* and *engagement* differ greatly by **Function**?

Response 2.2: *High values of supervisor support tend to associate with high values of engagement. Judging by the loess line, the relationship between supervisor support and engagement differs by function for lower levels of supervisor support. There is no significant difference by function for higher values of supervisor support.*

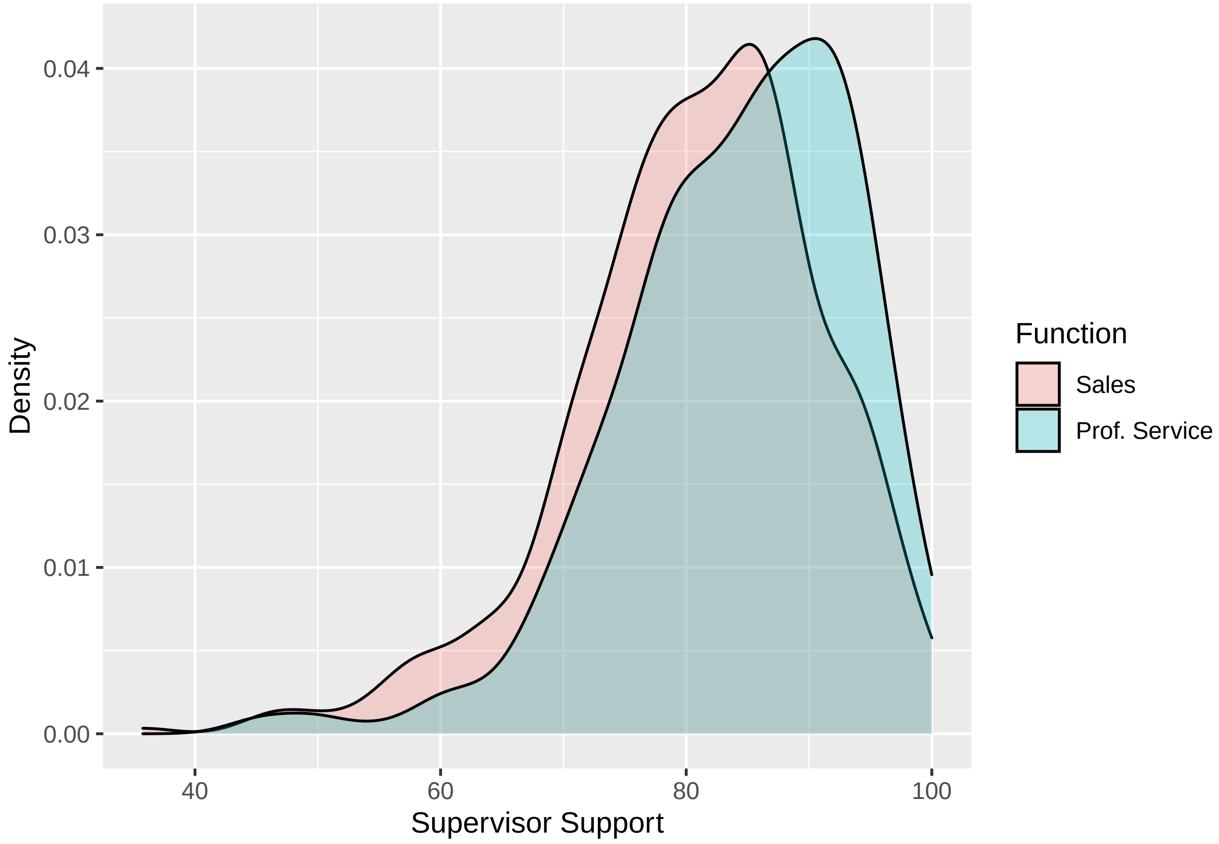
Create another scatterplot to show the relationship between *supervisor support* (i.e., **suprvsr_supp**) and *engagement* (i.e., **engage**). Include all of the following in this one plot:

1. place *supervisor support* on the x-axis and *engagement* on the y-axis,
2. *color* and *shape* the points by **Function**,
3. size the points by the *number of female team leads*, and
4. write a *descriptive title* for the plot.

Question 2.3: Is the overall relationship (ignoring **Function** and *number of female team leads*) between team *supervisor support* and *engagement* the same as in the prior plot? Do you see any differences in the relationship between *supervisor support* and *engagement* for different *number of female team leads*?

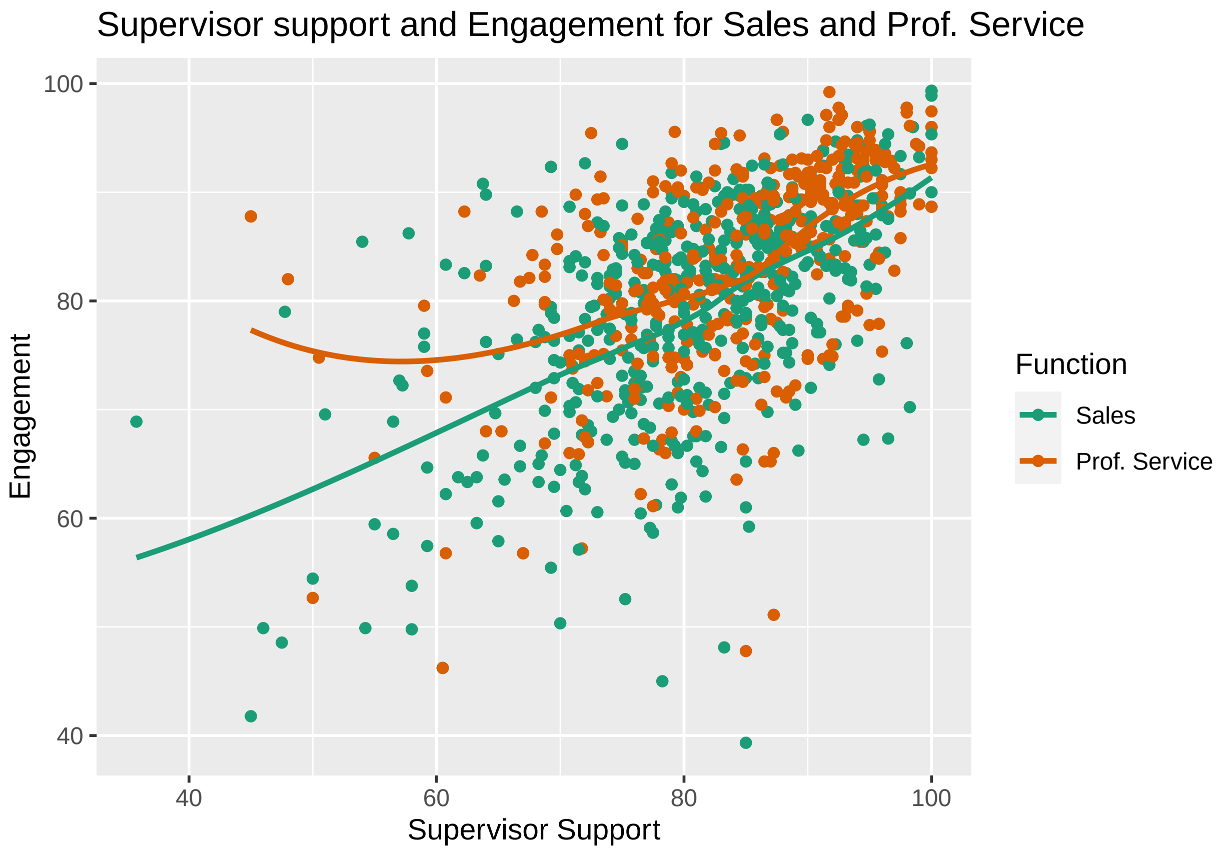
Response 2.3: *The overall relationship between team supervisor support and engagement is the same as in the prior plot. I do not see any differences in the relationship between supervisor support and engagement for different number of female team leads*

```
#### Q2.1
### plot distribution of single continuous/numeric variable
### across a levels of a factor variable
## choose data
ggplot(data = team_div_work_data) +
  ## choose density geometry for numeric variable;
  ## specify numeric on x-axis and factor variable for fill
  geom_density(aes(x = suprvsr_supp, fill = Function), alpha = 0.25) +
  ## label axes
  labs(x = "Supervisor Support", y = "Density", fill = "Function")
```



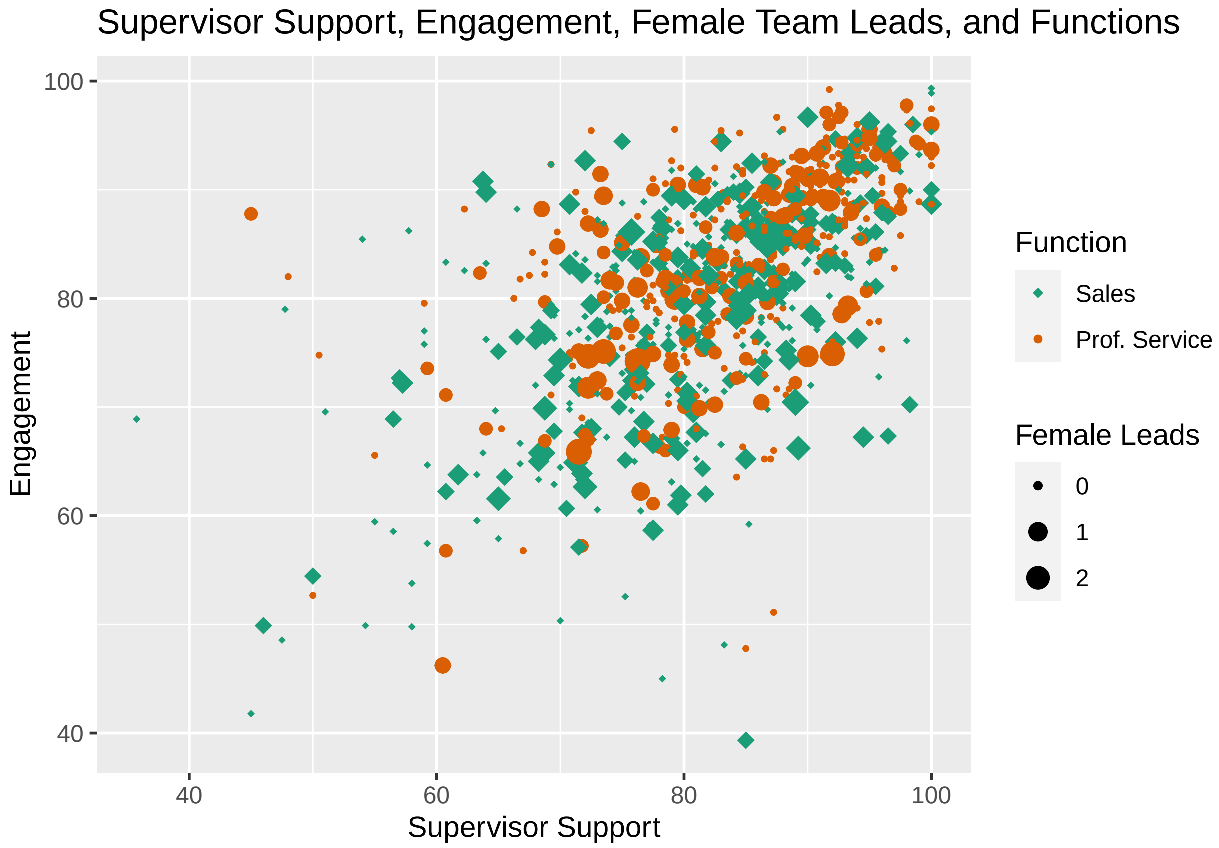
```
#### Q2.2
### examine relationship between two numeric variables;
### use loess line to examine type of relationship;
### use factor variable to color points
## choose data
ggplot(team_div_work_data, aes(x = suprvsr_supp, y = engage,
                            # color data points
                            color = Function)) +
  ## choose point geometry for scatterplot
  geom_point() +
  ## loess line
  geom_smooth(method = "loess", se = FALSE) +
  ## label axes
  labs(x = "Supervisor Support", y = "Engagement") +
  ## change default colors
  scale_color_brewer(palette = "Dark2") +
  ## add descriptive title
  ggtitle("Supervisor support and Engagement for Sales and Prof. Service")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
#### Q2.3
### examine relationship between three numeric variables
## choose data
ggplot(team_div_work_data, aes(x = suprvsr_supp, y = engage,
                         # size data points
                         size = NumberFeMaleTeamLeads,
                         # shape data points
                         shape = Function,
                         # color data points
                         color = Function)) +
  ## choose point geometry for scatterplot
  geom_point() +
  ## scale the size of dots
  scale_size_continuous(breaks = 0:2) +
  ## scale shapes
  scale_shape_manual(values = c(18, 20)) +
  ## change default colors
  scale_color_brewer(palette = "Dark2") +
  ## label axes
  labs(x = "Supervisor Support", y = "Engagement",
       size = "Female Leads", color = "Function") +
  ## add descriptive title
  ggtitle("Supervisor Support, Engagement, Female Team Leads, and Functions")
```

```
## Warning: Removed 19 rows containing missing values (geom_point).
```



Task 3: Transform Data

For your third task, you will transform **team_div_work_data** to long format.

Name the long formatted data **team_div_long**. Select the variables: 1. **DepartmentGroupNumber**, 2. **Function**, 3. **Location**, 4. **PercentMale**, and 5. **BAME** (i.e., Black, Asian, Minority Ethnic). Pivot longer the variables **PercentMale** and **BAME**, name the new variable that identifies the newly created rows **pm_bame**, and name the new variable that holds the numeric values **pm_bame_val**.

Print the top of **team_div_long**.

Question 3.1: What is the *percentage of males* in *department number 30*?

Response 3.1: *32%.*

Use **team_div_long** to calculate the *summary statistics* for the combination of **Function**, **Location**, and **pm_bame**. You will compute the mean, standard deviation, count of teams, and standard error of the mean just like in the analytical script. Note that you will use **Location** and not **London**. Save the results in a new object named **team_div_summ**.

Question 3.2: What is the *average* (i.e., mean) team *BAME* for the *professional service* function in *Central London*? What is the *standard deviation* of the team *percentage of males* for the *sales* function in the *UK excluding London*?

Response 3.2: *19.9; 16.2.*

```r
#### Q3.1
### pivot data to make it longer for plotting goals
## create long version of data
team_div_long <- team_div_work_data %>%
  ## Select variables
  select(DepartmentGroupNumber, Function, Location,
         PercentMale, BAME) %>%
  ## Make data long format
  pivot_longer(cols = c(PercentMale, BAME),
               # set names of two new variables
               names_to = "pm_bame", values_to = "pm_bame_val") %>%
  ## change labels for new factor variable
  mutate(pm_bame = fct_recode(pm_bame,
                              `Percent Male` = "PercentMale"))

## print the data
team_div_long
```

```
## # A tibble: 1,854 x 5
##    DepartmentGroupNumber Function      Location       pm_bame      pm_bame_val
##                    <dbl> <fct>         <fct>          <fct>              <dbl>
##  1                    18 Prof. Service Central London Percent Male          65
##  2                    18 Prof. Service Central London BAME                  NA
##  3                    19 Prof. Service Central London Percent Male          67
##  4                    19 Prof. Service Central London BAME                  NA
##  5                    29 Prof. Service Central London Percent Male          33
##  6                    29 Prof. Service Central London BAME                  NA
##  7                    30 Prof. Service Central London Percent Male          32
##  8                    30 Prof. Service Central London BAME                  NA
##  9                    35 Prof. Service Central London Percent Male          18
## 10                    35 Prof. Service Central London BAME                  NA
## # ... with 1,844 more rows
```

```
#### Q3.2
### calculate group summary statistics
## create summary data object
team_div_summ <- team_div_long %>%
  ## group by Function and pm_bame
  group_by(Function, Location, pm_bame) %>%
  ## calculate summary statistics
              # choose pm_bame_val variable
  summarize_at(vars(pm_bame_val),
                   # calculate average for each group
              list(mean = ~ mean(., na.rm = T),
                   # calculate standard deviation for each group
                   sd = ~ sd(., na.rm = T),
                   # count teams in each group
                   count = ~ n())) %>%
  ## calculate standard error of mean for each group
  mutate(se = sd/sqrt(count)) %>%
  ## remove grouping
  ungroup()

## print summary statistics
team_div_summ
```

```
## # A tibble: 12 x 7
##    Function      Location       pm_bame       mean    sd count    se
##    <fct>         <fct>          <fct>        <dbl> <dbl> <int> <dbl>
##  1 Sales         Central London BAME         14     9.87   133 0.856
##  2 Sales         Central London Percent Male 69.1  15.7    133 1.36
##  3 Sales         Greater London BAME         10.4  10.6    143 0.883
##  4 Sales         Greater London Percent Male 72.8  16.9    143 1.41
##  5 Sales         Rest of UK     BAME          5.40  6.85   226 0.455
##  6 Sales         Rest of UK     Percent Male 71.5  16.2    226 1.07
##  7 Prof. Service Central London BAME         19.9  12.5    149 1.03
##  8 Prof. Service Central London Percent Male 44.8  18.5    149 1.51
##  9 Prof. Service Greater London BAME         15.0  12.1    108 1.16
## 10 Prof. Service Greater London Percent Male 45.8  18.8    108 1.81
## 11 Prof. Service Rest of UK     BAME          7.65  9.11   168 0.703
## 12 Prof. Service Rest of UK     Percent Male 43.1  20.7    168 1.60
```

Task 4: Team Diversity

For this task, you will examine team diversity in the organization.

Use **team_div_long** to create a *boxplot* showing the relationship between **Function**, **Location**, and **pm_bame**. Save the plot as the object **team_div_boxplot**. Include all of the following in the plot: 1. place **Function** on the x-axis and **pm_bame_val** on the y-axis, 2. color outliers *red*, 3. include *jittered* data points, 4. make a facet grid with **Location** in rows and **pm_bame** in the columns, 5. fill the boxplots by **Location** and **pm_bame** simultaneously, and 6. do *not* include a legend. Print the plot for viewing in your working session.

Save the plot to your *plots* folder. Name the plot file **team_div_boxplot.png**.

Question 4.1: Describe the takeaways from the plot.

Response 4.1: *There is a significant difference between the percentage of males in Sales and Prof. Service. There is no significant difference between the percentage of males by function depending on location. While the standard deviation for the percentage of males in Sales is small, it is somewhat larger for Prof. Service (meaning the percentage of males in Prof. Service groups varies more). The sales function teams seem to be consisted mainly out of white team members. More BAME team members seem to be present at Prof. Services teams. Their percentage is consistent across teams outside of London but varies greatly across teams for London areas (as indicated by the IQR). There seems to be a smaller percentage of BAME for teams located outside of London.*

Use **team_div_summ** to create a *bar plot with error bars* showing the relationship between **Function**, **Location**, and **pm_bame**. Save the plot as the object **team_div_summ_plot**. Include all of the following in the plot: 1. place **Function** on the x-axis and **mean** on the y-axis, 2. make a facet grid with **Location** in rows and **pm_bame** in the columns, 3. include text to write the mean plus/minus two standard errors of the mean, 4. fill the boxplots by **Location** and **pm_bame** simultaneously, and 5. do *not* include a legend. Print the plot for viewing in your working session.

Save the plot to your *plots* folder. Name the plot file **team_div_summ.png**.

Question 4.2: Describe the takeaways from the plot.

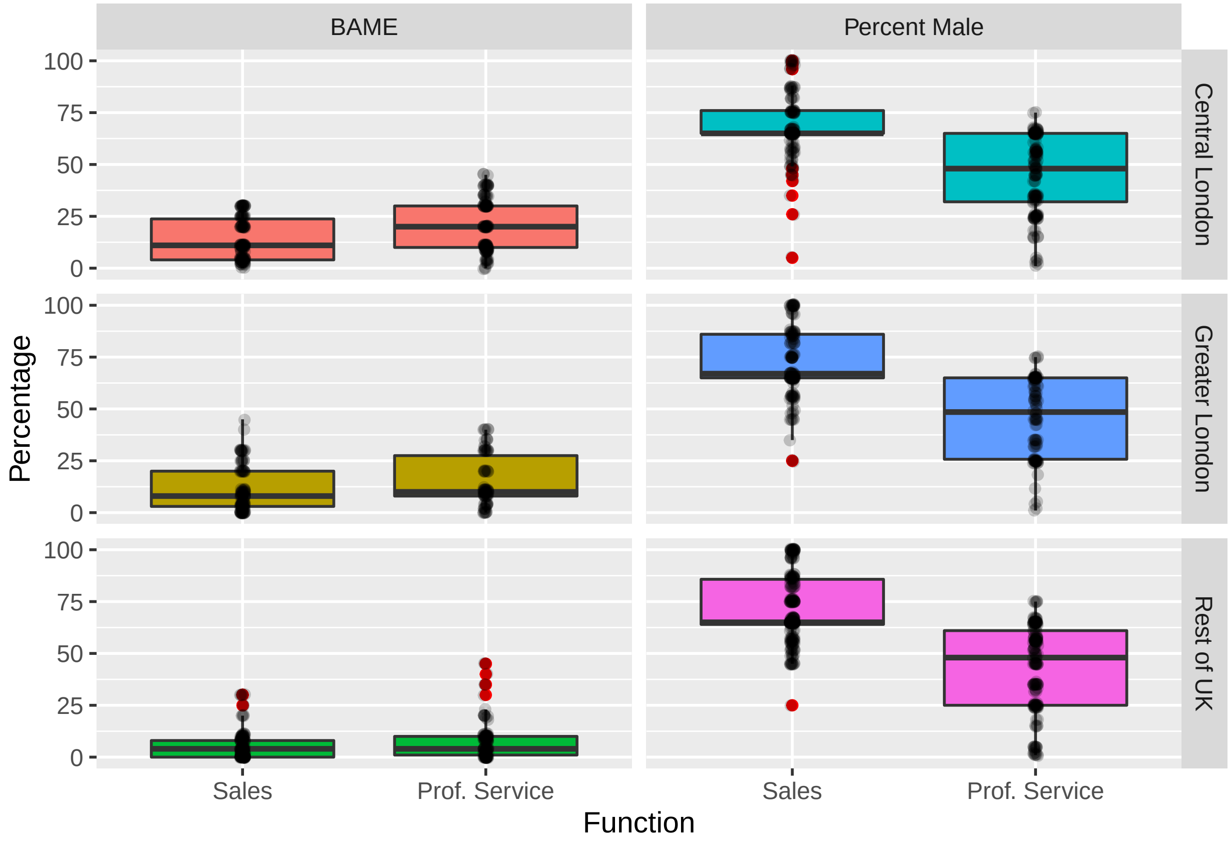
Response 4.2: *The plot leads to the same conclusions as the prior one. The average percentage of males in sales teams tends to be around 70% regardless of the location of the team. Te average percentage of males in Prof. Service tends to be around 44% across all locations. The average BAME score is 14% and 10.4% for Central London and Greater London respectively. The BAME score is significantly lower for teams outside of London and averages 5.4%.*

```r
#### Q4.1
### create boxplot of PercentMale and BAME compared on Function
## choose data and mapping
team_div_boxplot <- ggplot(data = team_div_long,
       # place Function and percentage value on x and y axes
       mapping = aes(x = Function, y = pm_bame_val,
                     # color boxplots by Location and pm_bame
                     # simultaneously
                     fill = interaction(Location, pm_bame))) +
  ## add boxplot
  geom_boxplot(outlier.color = "red") +
  ## add points
  geom_jitter(width = 0.01, alpha = 0.2) +
  ## facet for variable type
  facet_grid(Location ~ pm_bame) +
  ## labels
  labs(y = "Percentage") +
  ## hide legend
  theme(legend.position = "none")

## print plot
team_div_boxplot
```

```
## Warning: Removed 181 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 181 rows containing missing values (geom_point).
```



```
### save plots to folder in project directory
## save a single plot to a file
## specify folder and file name in project directory
ggsave(here("plots", "team_div_boxplot.png"),
       # specify plot to save
       plot = team_div_boxplot,
       # specify units
       units = "in", width = 9, height = 5)
```

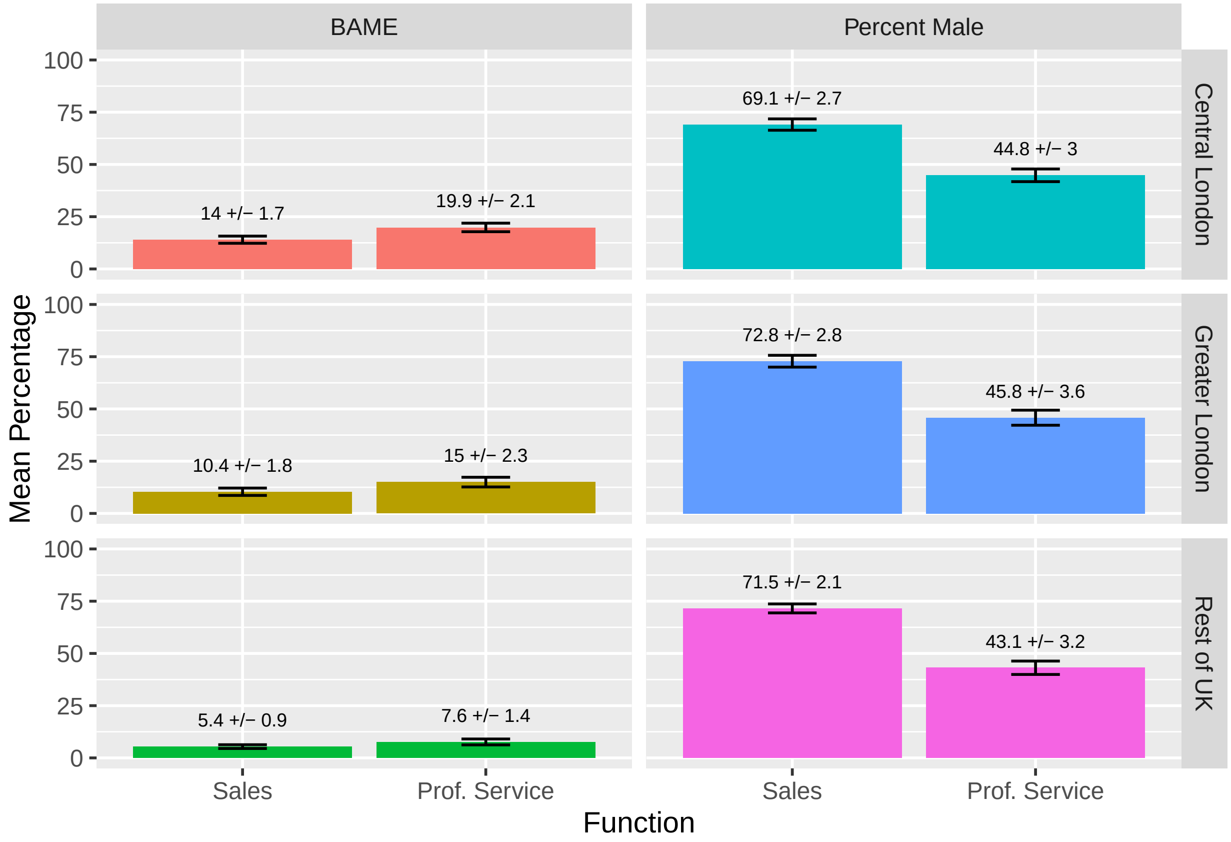
```
## Warning: Removed 181 rows containing non-finite values (stat_boxplot).

## Warning: Removed 181 rows containing missing values (geom_point).
```

```
#### Q4.2
### create bar graph of PercentMale and BAME compared on Function
## choose data and mapping
team_div_summ_plot <- ggplot(data = team_div_summ,
       # place Function and mean value on x and y axes
       mapping = aes(x = Function, y = mean,
                     # color boxplots by Location and pm_bame
                     # simultaneously
                     fill = interaction(Location, pm_bame))) +
  ## add dodged bar graph
  geom_bar(stat = "identity", position = "dodge") +
  ## add error bars
```

```
                    # minimum value
  geom_errorbar(aes(ymin = mean - 2*se,
                    # maximum value
                    ymax = mean + 2*se),
                # width of error bars
                width = 0.2) +
  ## add text
                        # round mean value to one decimal
  geom_text(aes(label = paste0(round(mean, 1),
                               # plus or minus
                               " +/- ",
                               # round standard error
                               round(2*se, 1))),
            # size and justify text
            size = 2.5, vjust = -1.5) +
  ## facet for variable type
  facet_grid(Location ~ pm_bame) +
  ## adjust y limits
  ylim(0, 100) +
  ## labels
  labs(y = "Mean Percentage") +
  ## hide legend
  theme(legend.position = "none")

## print plot for viewing
team_div_summ_plot
```



```r
### save plots to folder in project directory
## save a single plot to a file
## specify folder and file name in project directory
ggsave(here("plots", "team_div_summ.png"),
       # specify plot to save
       plot = team_div_summ_plot,
       # specify units
       units = "in", width = 9, height = 5)
```

Task 5: Diversity Analytics

Watch the conceptual overview video, *Team Diversity Review*, on *D2L* discussing our analytical work the first two weeks of the course. The video reviews our diversity analytics in two organizations. The first week we examined differences in the job roles of men and women in a management consulting firm. This second week we examined ethnic and gender differences in sales and professional service positions in a financial organization. Respond to the following questions after watching the video.

Question 5.1: Why should organizations care about analyzing discrepancies in the positions different gender and ethnicities take on in an organization?

Response 5.1: *Companies should make an effort to create a workplace that accurately reflects our diverse society. Diversity should not be limited to workforce and should extend to managerial and executive positions as well. Neglecting discrepancies in the positions different gender and ethnicities could lead to a series of consequences which include a damaged public image of the company and discrimination lawsuits. In addition, having a diverse workforce increases the amount of different viewpoints which help guide companies in the right direction.*

Question 5.2: Last week, we analyzed the percentage of men and women along the organizational hierarchy in a management consuting firm. What did we learn from our analytics?

Response 5.2: *We learned that executive and top-level business roles are male-dominated.*

Question 5.3: This week, we analyzed the percentage of men and ethnic minorities in sales and professional service positions in a financial organization. What did we learn from our analytics?

Response 5.3: *We learned that the majority of team members were males. In addition, positions that require contact with customers were primarily filled by white people.*

Question 5.4: What can we advise these two organizations with regard to investigating the discrepancies in the roles different gender and ethnicities take on in their organizations?

Response 5.4: *We can advise the companies to reevaluate their HR policies concerning diversity and equality in the company. Moreover, the hiring process should be updated in order to remove biased criteria and allow for a more diverse workplace.*