Assignment 4

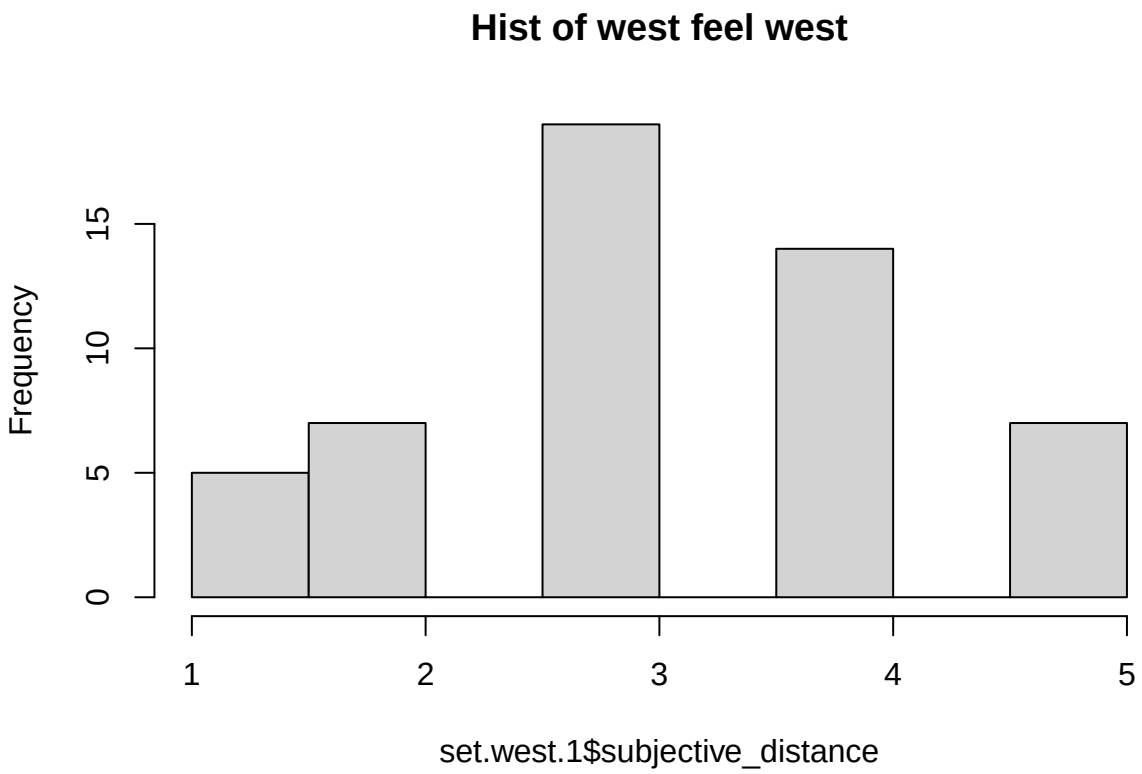
Mingrui Duan

Question 1: Load the data. Look for outliers and remove. Describe 3 checks you should do on the dataset you have loaded. Do them and fix any issues you find.

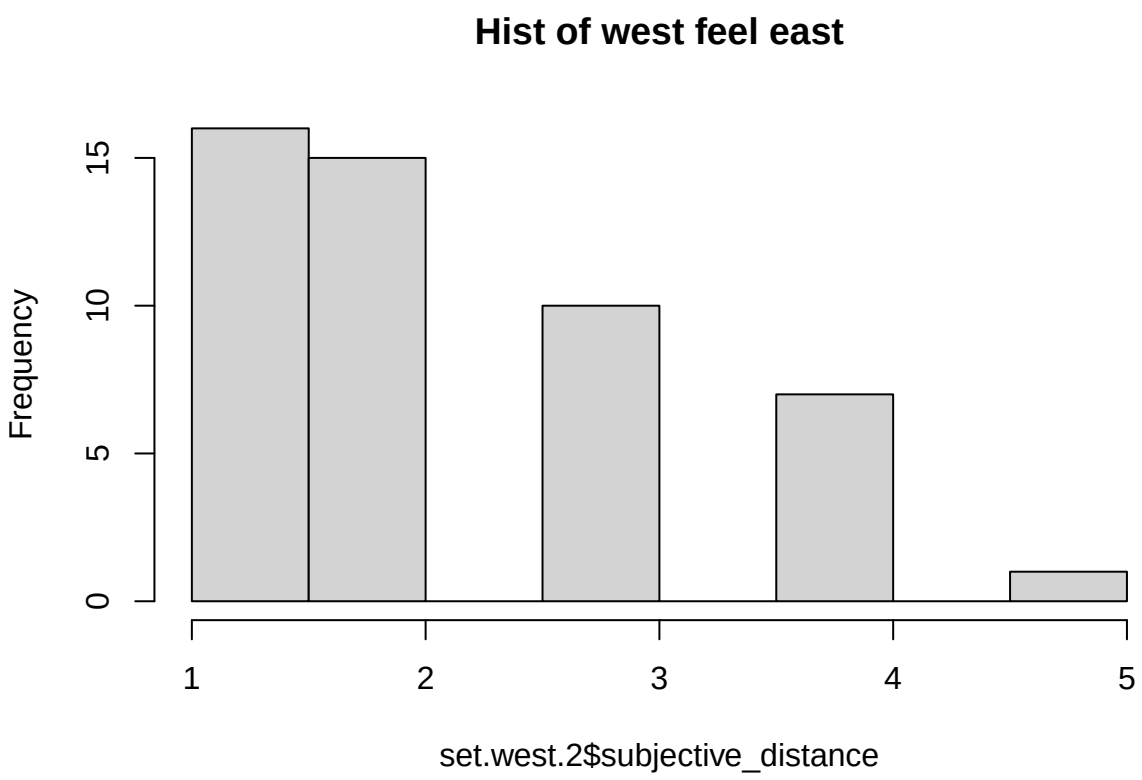
```
d <- read.csv("data.csv", header = TRUE) # Load the dataset
summary(d)
```

```
##   orientation     station      subjective_distance
##  Min.   :1.0   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.0   1st Qu.:1.250   1st Qu.:2.000
##  Median :1.5   Median :2.000   Median :3.000
##  Mean   :1.5   Mean   :2.495   Mean   :2.673
##  3rd Qu.:2.0   3rd Qu.:3.750   3rd Qu.:4.000
##  Max.   :2.0   Max.   :4.000   Max.   :6.000
```

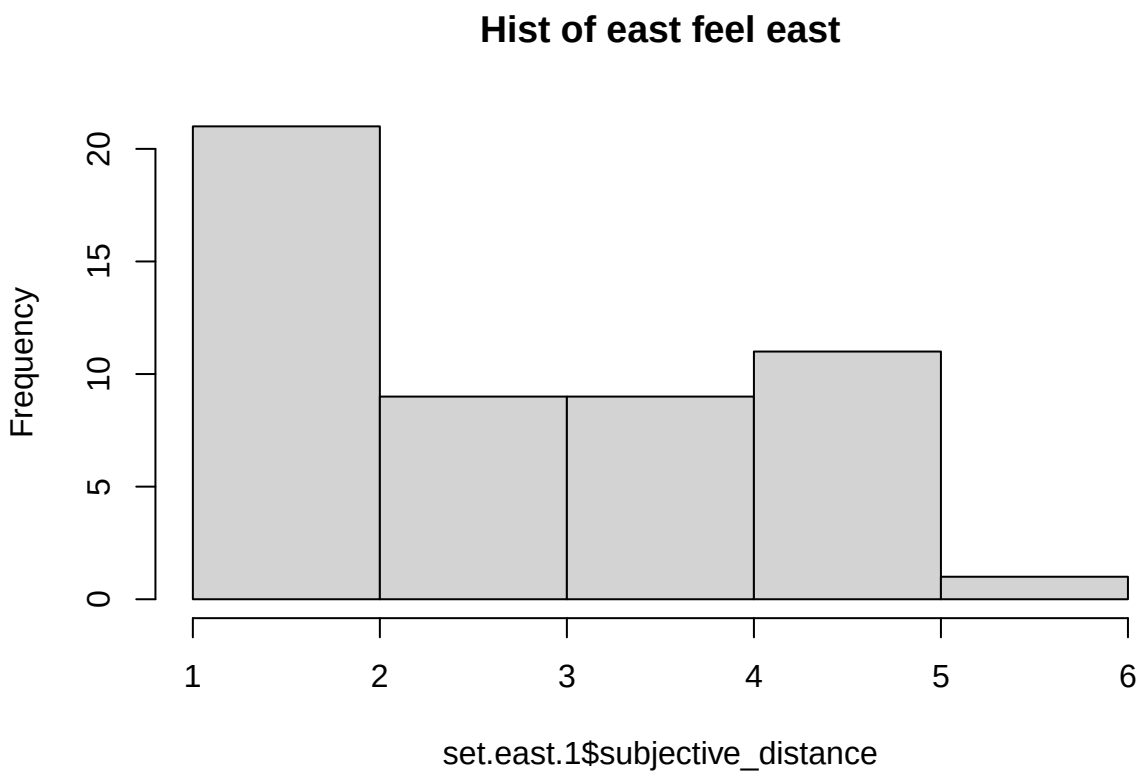
```
# Make hist of each attribute to determine which outlier to be removed
# Set the orientation as "West" and test the west bay station
set.west.1 <- subset(d, d$orientation == 1 & d$station <= 2)
hist(set.west.1$subjective_distance, main = "Hist of west feel west")
```



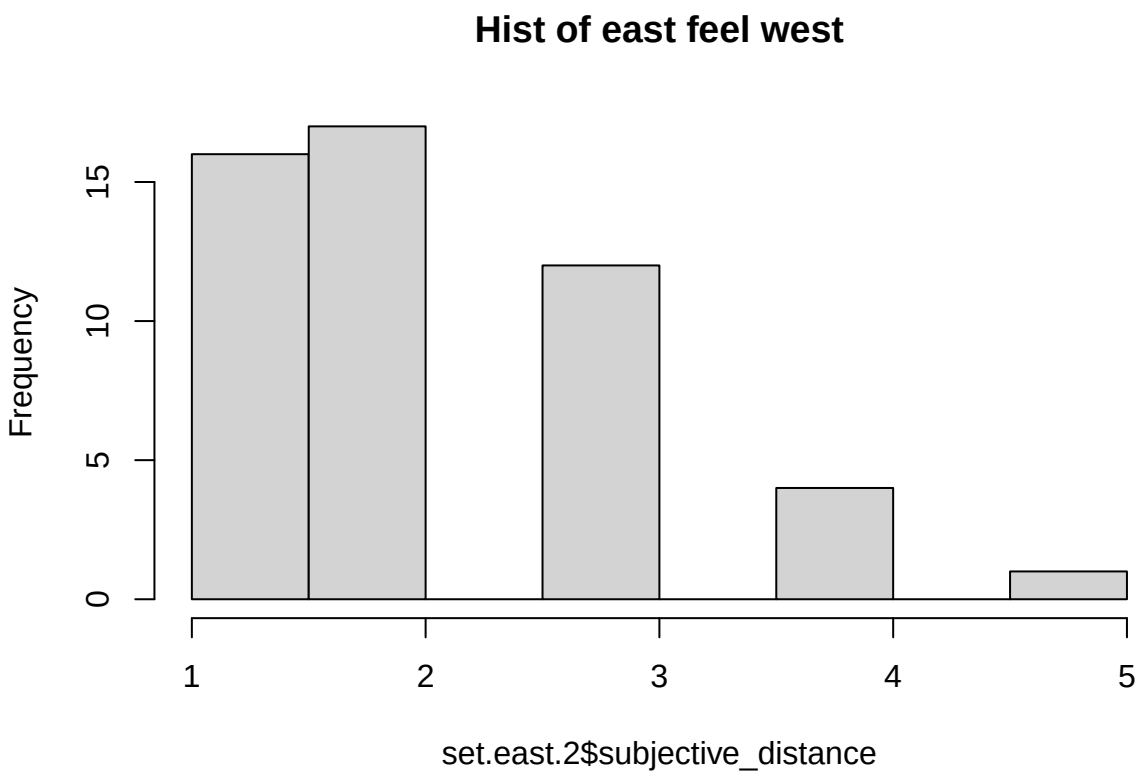

```r
# Set the orientation as "West" and test the east bay station
set.west.2 <- subset(d, d$orientation == 1 & d$station > 2)
hist(set.west.2$subjective_distance, main = "Hist of west feel east")
```




```r
# Get the outlier remove
set.west.2.without <- subset(set.west.2, d$subjective_distance < 5)

# Set the orientation as "East" and test the east bay station
set.east.1 <- subset(d, d$orientation == 2 & d$station > 2)
hist(set.east.1$subjective_distance, main = "Hist of east feel east")
```




```r
# Get the outlier remove
set.east.1.without <- subset(set.east.1, d$subjective_distance < 6)

# Set the orientation as "East" and test the west bay station
set.east.2 <- subset(d, d$orientation == 2 & d$station <= 2)
hist(set.east.2$subjective_distance, main = "Hist of east feel west")
```




```
# Get the outlier remove
set.east.2.without <- subset(set.east.2, d$subjective_distance < 5)

# Combine those new dataframe to a new one
d <- rbind(set.west.1, set.west.2.without, set.east.1.without, set.east.2.without)

# Check the size of each attribute
print(length(d$orientation))
```

```
## [1] 615
```

```
print(length(d$station))
```

```
## [1] 615
```

```
print(length(d$subjective_distance))
```

```
## [1] 615
```

I first plot 4 histograms categorized by orientation and stations, to find out those outliers and remove that. From these 4 plots, we can tell that: - If the orientation is "West" with west bay station, subjective distance >= 5 should be removed - If the orientation is "East" with east bay station, subjective distance > 6 should be removed - If the orientation is "East" with west bay station, subjective distance > 5 should be removed After remove these outliers, we combine them to a new dataframe.

I will check the following 3 things to make sure the loading process is correct: - Overall data size - Attributes - Check max and min of each attribute to avoid data missing

After checked, the size of the dataset is correct. There's 3 attributes which is proper and each column's min and max is also correct.

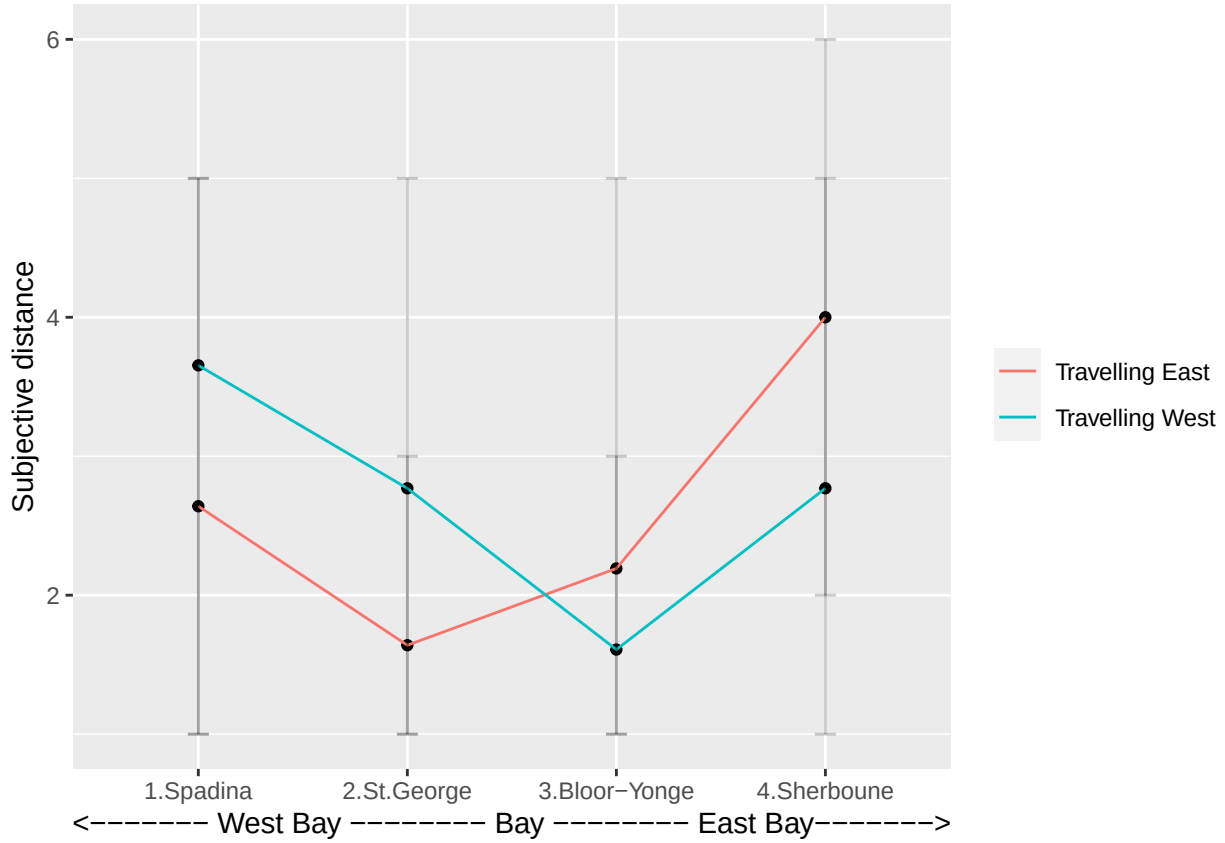
Question 2: You may have noticed that the columns are not given useful names and are numbers. Figure out what each number in each column refers to (e.g., either East/West in the station column or which station in the station column), and then create new columns which have factors for East/West and subway station.

Through the paper, we know that the authors has two orientation values: "toward" & "away from". From Figure 1 we can tell west stations have smaller value in the "station" attribute. So: - 1 in the station column will be "Spadina"; - 2 in the station column will be "St.George"; - 3 in the station column will be "Bloor-Yonge"; - 4 in the station column will be "Sherboune"; Also, - 1 in the orientation column will be "toward"; - 2 in the orientation column will be "away from";

```
# Load the dataset
d <- read.csv("data.csv", header = TRUE)
# Create a new factor of "orientation"
orientation_name <- c()
for (i in d$orientation) {
  if (i == 1) {
   orientation_name[length(orientation_name)+1] <- "toward"
  } else {
    orientation_name[length(orientation_name)+1] <- "away from"
  }
}
# Create a new factor of "station"
station_name <- c()
bay <- c()
for (i in d$station) {
  if (i <= 2){
    if (i == 1) {
      station_name[length(station_name)+1] <- "Spadina"
    } else {
      station_name[length(station_name)+1] <- "St.George"
    }
    bay[length(bay)+1] <- "West"
  } else {
    if (i == 3) {
      station_name[length(station_name)+1] <- "Bloor-Yonge"
    } else {
      station_name[length(station_name)+1] <- "Sherboune"
    }
    bay[length(bay)+1] <- "East"
  }
}
# Add the new factors to the dataframe
d$orientation_name <- orientation_name
d$station_name <- station_name
d$bay <- bay
```

Question 3: Using the data provided, first replicate the above figure. Use stat_summary to compute means, and another stat_summary to compute the error bars.

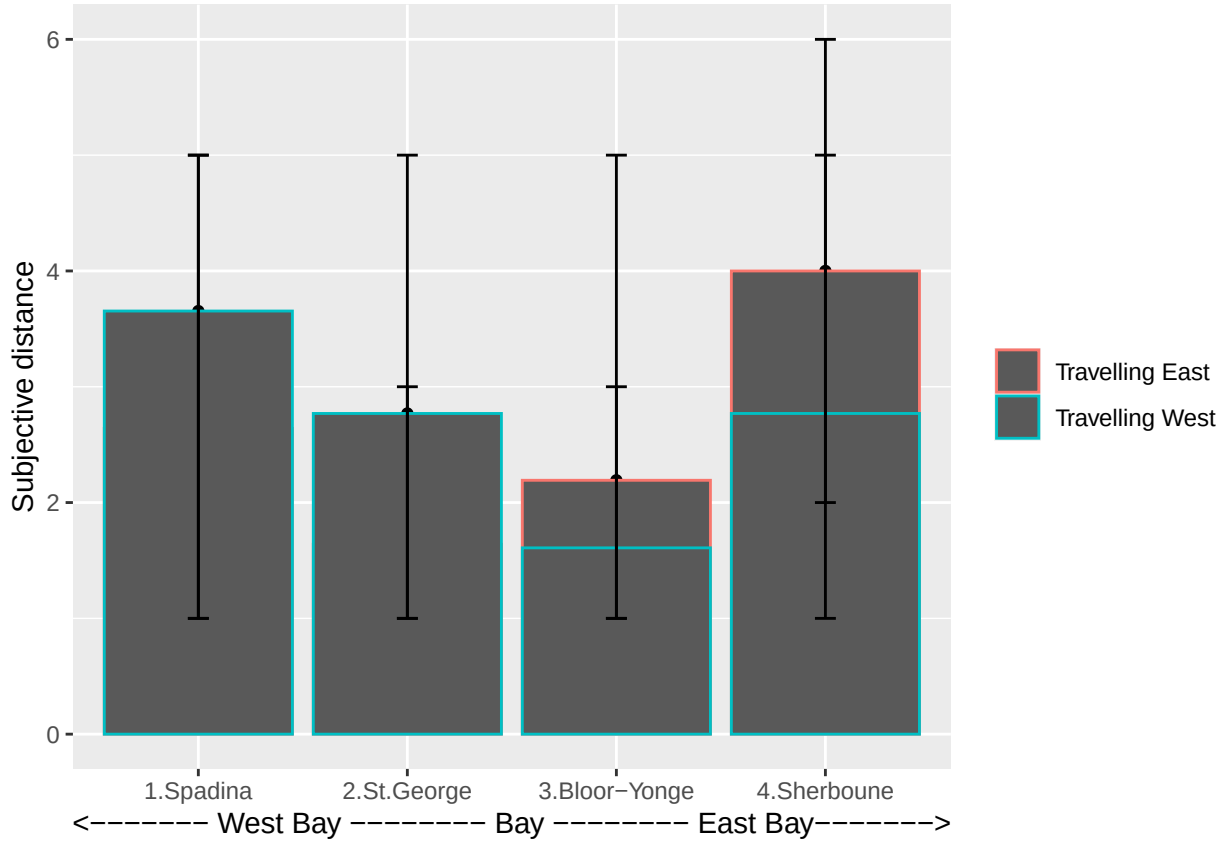
```
# Load the data
d <- read.csv("data.csv", header = TRUE)
# Create a new factor of "orientation"
orientation_name <- c()
for (i in d$orientation) {
  if (i == 1) {
   orientation_name[length(orientation_name)+1] <- "Travelling West"
  } else {
    orientation_name[length(orientation_name)+1] <- "Travelling East"
  }
}
# Create a new factor of "station"
station_name <- c()
bay <- c()
for (i in d$station) {
  if (i <= 2){
    if (i == 1) {
      station_name[length(station_name)+1] <- "1.Spadina"
    } else {
      station_name[length(station_name)+1] <- "2.St.George"
    }
    bay[length(bay)+1] <- "West"
  } else {
    if (i == 3) {
      station_name[length(station_name)+1] <- "3.Bloor-Yonge"
    } else {
      station_name[length(station_name)+1] <- "4.Sherboune"
    }
    bay[length(bay)+1] <- "East"
  }
}

# Add the new factors to the dataframe
d$orientation_name <- orientation_name
d$station_name <- station_name
d$bay <- bay

# Plot
library("ggplot2")
e <- d$station_name
p <- ggplot(d, aes(x=station_name, y=subjective_distance, group = orientation_name)) +
        stat_summary(fun = "mean", geom = "point") +
        stat_summary(fun = "mean", geom = "line", aes(color=orientation_name)) +
        stat_summary(fun.min = min, fun.max = max, geom = "errorbar", width = 0.1, alpha=0.2) +
        ylab("Subjective distance") +
        xlab("<------- West Bay -------- Bay -------- East Bay------->") +
        theme(legend.title = element_blank())
p
```



Question 4: Replot the same figure using a barplot (geom="bar") instead of a line plot. Which is a better visualization, bar plot or line plot? Why?

```
# Load the data
d <- read.csv("data.csv", header = TRUE)
# Create a new factor of "orientation"
orientation_name <- c()
for (i in d$orientation) {
  if (i == 1) {
   orientation_name[length(orientation_name)+1] <- "Travelling West"
  } else {
    orientation_name[length(orientation_name)+1] <- "Travelling East"
  }
}
# Create a new factor of "station"
station_name <- c()
bay <- c()
for (i in d$station) {
  if (i <= 2){
    if (i == 1) {
      station_name[length(station_name)+1] <- "1.Spadina"
    } else {
      station_name[length(station_name)+1] <- "2.St.George"
    }
```

```
    bay[length(bay)+1] <- "West"
  } else {
    if (i == 3) {
      station_name[length(station_name)+1] <- "3.Bloor-Yonge"
    } else {
      station_name[length(station_name)+1] <- "4.Sherboune"
    }
    bay[length(bay)+1] <- "East"
  }
}

# Add the new factors to the dataframe
d$orientation_name <- orientation_name
d$station_name <- station_name
d$bay <- bay

# Plot
library("ggplot2")
e <- d$station_name
p <- ggplot(d, aes(x=station_name, y=subjective_distance, group = orientation_name)) +
        stat_summary(fun = "mean", geom = "point") +
        stat_summary(fun = "mean", geom = "bar", aes(color=orientation_name)) +
        stat_summary(fun.min = min, fun.max = max, geom = "errorbar", width = 0.1) +
        ylab("Subjective distance") +
        xlab("<------- West Bay -------- Bay -------- East Bay------->") +
        theme(legend.title = element_blank())
p
```



Question 5: In a few sentences each, explain what "standard error" bars mean at two different levels: (a) to an advanced undergraduate, and (b) to a five year old child.

(a) "Standard error" is the standard deviation of its sampling distribution, it reflects an estimate of it. It is used to measure the difference between samples' mean and the distribution's mean.
(b) "Standard error" is a criterion to tell you to what extent you are different from other kids like you.

Question 6: Make a plot, including standard error, illustrating whether there is a main effect of East-vs-West.

```
# Load the data
d <- read.csv("data.csv", header = TRUE)
# Create a new factor of "orientation"
orientation_name <- c()
for (i in d$orientation) {
  if (i == 1) {
   orientation_name[length(orientation_name)+1] <- "Travelling West"
  } else {
    orientation_name[length(orientation_name)+1] <- "Travelling East"
  }
}
```

```r
# Create a new factor of "station"
station_name <- c()
bay <- c()
for (i in d$station) {
  if (i <= 2){
    if (i == 1) {
      station_name[length(station_name)+1] <- "1.Spadina"
    } else {
      station_name[length(station_name)+1] <- "2.St.George"
    }
    bay[length(bay)+1] <- "West"
  } else {
    if (i == 3) {
      station_name[length(station_name)+1] <- "3.Bloor-Yonge"
    } else {
      station_name[length(station_name)+1] <- "4.Sherboune"
    }
    bay[length(bay)+1] <- "East"
  }
}

# Add the new factors to the dataframe
d$orientation_name <- orientation_name
d$station_name <- station_name
d$bay <- bay

# Compute the standard error of each station
standard.station.1 <- sd(subset(d, d$station == 1)$subjective_distance)
standard.station.2 <- sd(subset(d, d$station == 2)$subjective_distance)
standard.station.3 <- sd(subset(d, d$station == 3)$subjective_distance)
standard.station.4 <- sd(subset(d, d$station == 4)$subjective_distance)
```

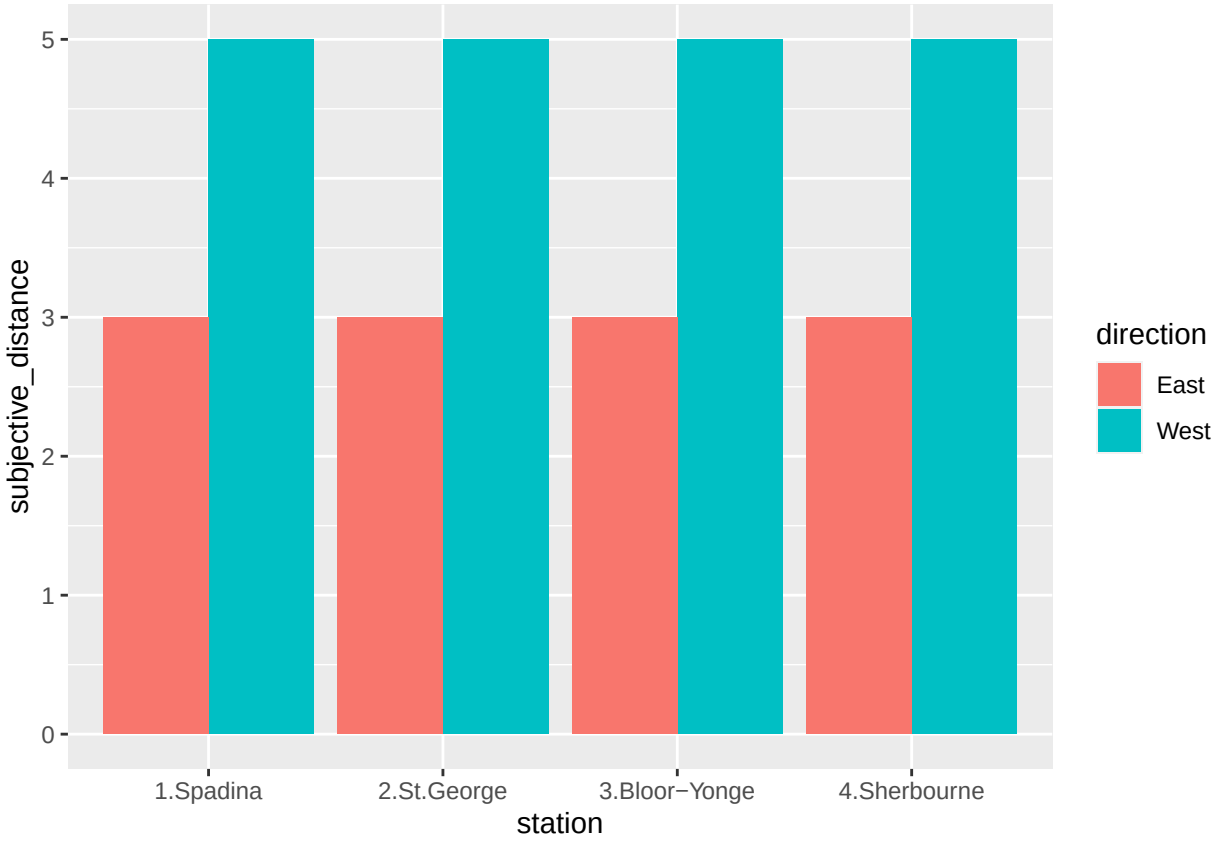
Question 7: Make a plot, including standard error, illustrating whether there are main effects of Station.

```r
# Load the data
d <- read.csv("data.csv", header = TRUE)
# Create a new factor of "orientation"
orientation_name <- c()
for (i in d$orientation) {
  if (i == 1) {
   orientation_name[length(orientation_name)+1] <- "Travelling West"
  } else {
    orientation_name[length(orientation_name)+1] <- "Travelling East"
  }
}
# Create a new factor of "station"
station_name <- c()
bay <- c()
for (i in d$station) {
  if (i <= 2){
```

```
    if (i == 1) {
      station_name[length(station_name)+1] <- "1.Spadina"
    } else {
      station_name[length(station_name)+1] <- "2.St.George"
    }
    bay[length(bay)+1] <- "West"
  } else {
    if (i == 3) {
      station_name[length(station_name)+1] <- "3.Bloor-Yonge"
    } else {
      station_name[length(station_name)+1] <- "4.Sherboune"
    }
    bay[length(bay)+1] <- "East"
  }
}

# Add the new factors to the dataframe
d$orientation_name <- orientation_name
d$station_name <- station_name
d$bay <- bay

# Compute the standard error of each station
standard.station.1 <- sd(subset(d, d$station == 1)$subjective_distance)
standard.station.2 <- sd(subset(d, d$station == 2)$subjective_distance)
standard.station.3 <- sd(subset(d, d$station == 3)$subjective_distance)
standard.station.4 <- sd(subset(d, d$station == 4)$subjective_distance)
```

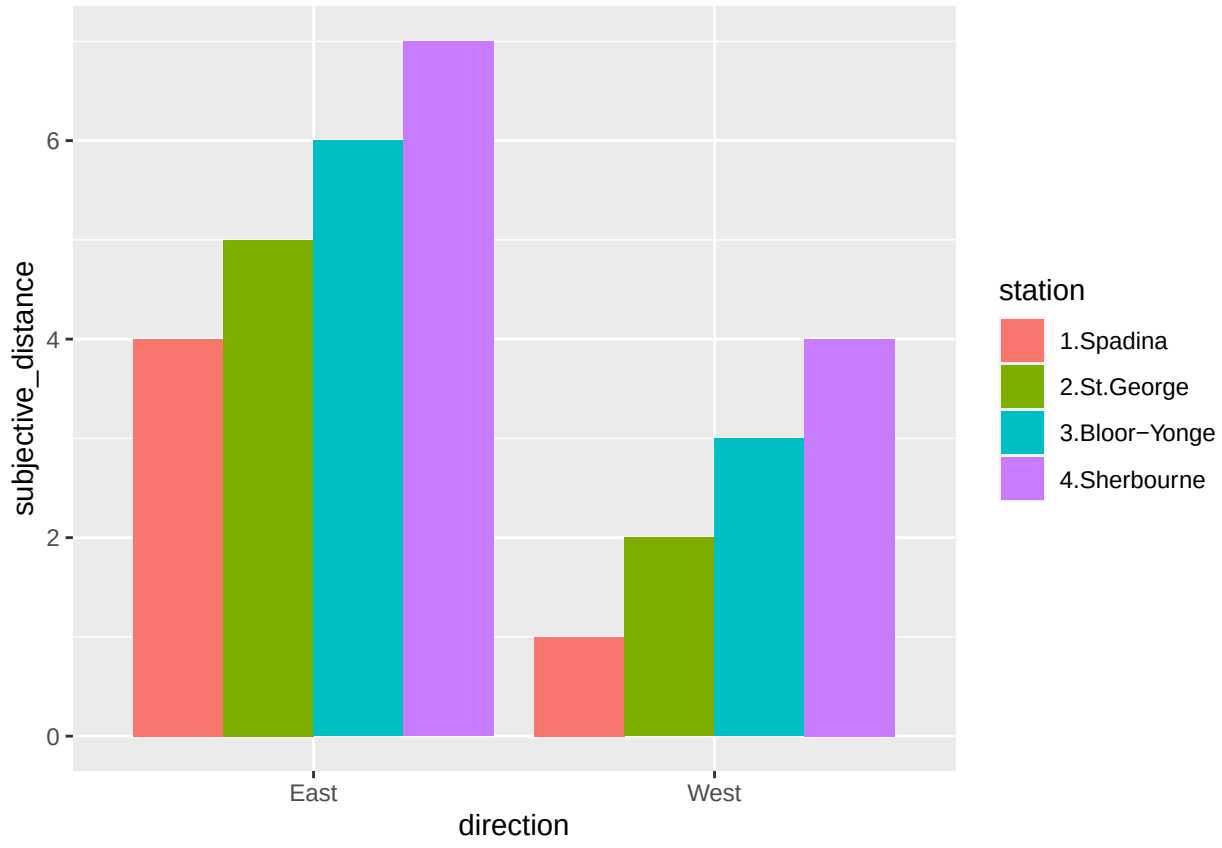
Question 8: main effect of direction, no effect of station, no interactions

```
# Create the subjective distance column
subjective_distance <- c(3, 4, 5, 1, 2, 3,
                         3, 4, 5, 1, 2, 3,
                         3, 4, 5, 1, 2, 3,
                         3, 4, 5, 1, 2, 3)
# Create the station column
station <- c("1.Spadina", "1.Spadina", "1.Spadina", "1.Spadina", "1.Spadina", "1.Spadina",
             "2.St.George", "2.St.George", "2.St.George", "2.St.George", "2.St.George", "2.St.George",
             "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bl
             "4.Sherbourne", "4.Sherbourne", "4.Sherbourne", "4.Sherbourne", "4.Sherbourne", "4.Sherbou
#
direction <- c("West", "West", "West", "East", "East", "East",
               "West", "West", "West", "East", "East", "East",
               "West", "West", "West", "East", "East", "East",
               "West", "West", "West", "East", "East", "East")

d <- data.frame(direction, station, subjective_distance)
p <- ggplot(d, aes(x=station, y=subjective_distance, fill=direction)) +
        geom_bar(stat = "identity", position = position_dodge())
p
```



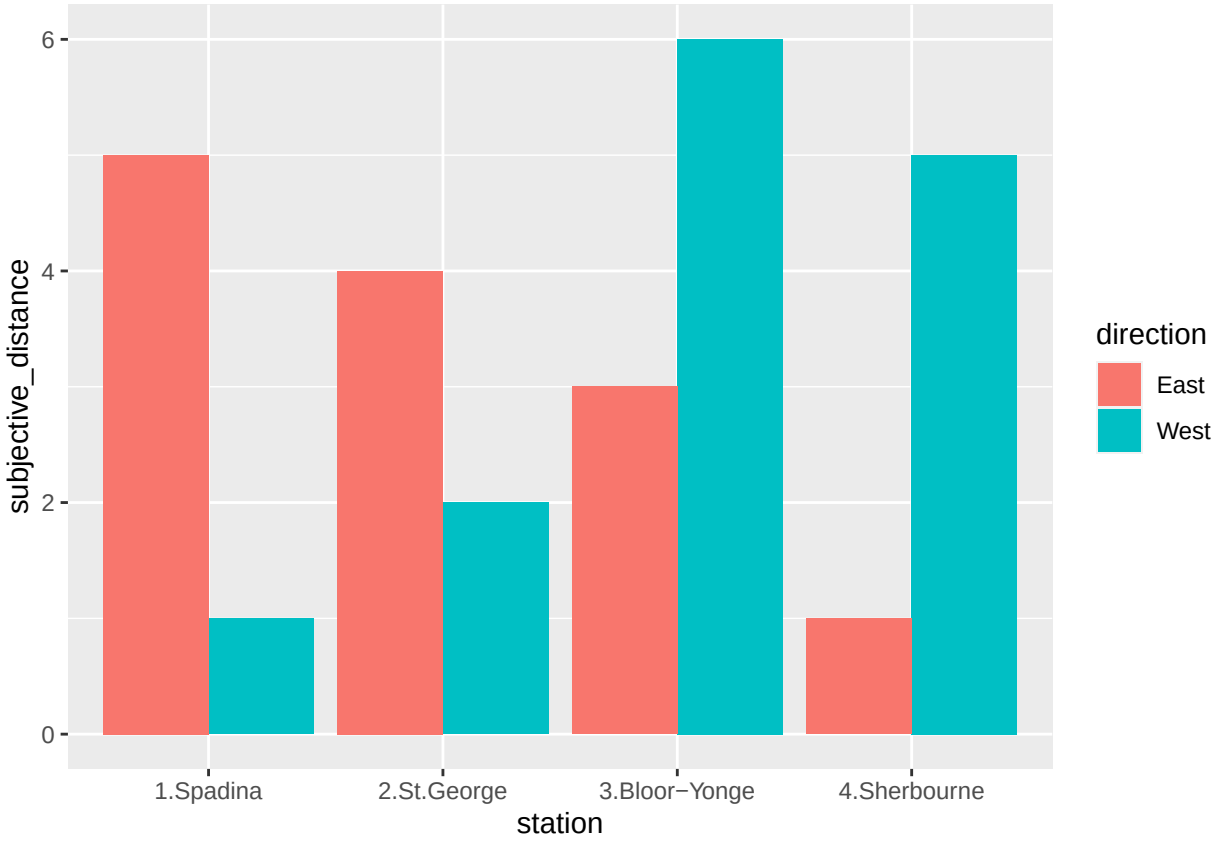
Question 9: main effect of station, main effect of east/west, no interactions

```
# Create the subjective distance column
subjective_distance <- c(1, 1, 1, 2, 2, 2,
                         3, 3, 3, 4, 4, 4,
                         4, 4, 4, 5, 5, 5,
                         6, 6, 6, 7, 7, 7)
# Create the station column
station <- c("1.Spadina", "1.Spadina", "1.Spadina", "2.St.George", "2.St.George", "2.St.George",
             "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "4.Sherbourne", "4.Sherbourne", "4.Sherb
             "1.Spadina", "1.Spadina", "1.Spadina", "2.St.George", "2.St.George", "2.St.George",
             "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "4.Sherbourne", "4.Sherbourne", "4.Sherb
#
direction <- c("West", "West", "West", "West", "West", "West",
               "West", "West", "West", "West", "West", "West",
               "East", "East", "East", "East", "East", "East",
               "East", "East", "East", "East", "East", "East")

d <- data.frame(direction, station, subjective_distance)
p <- ggplot(d, aes(x=direction, y=subjective_distance, fill=station)) +
        geom_bar(stat = "identity", position = position_dodge())
p
```



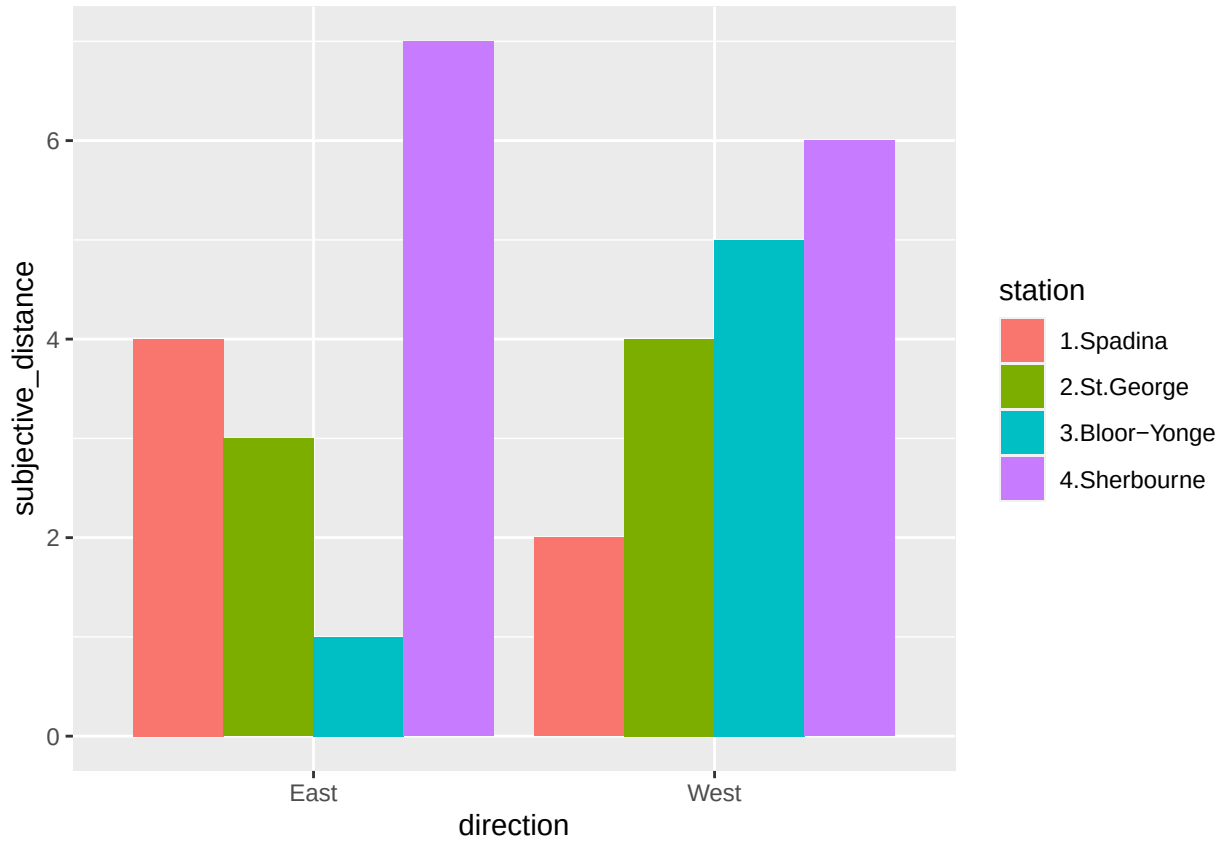
Question 10: main effect of station, main effect of east/west, interaction

```
# Create the subjective distance column
subjective_distance <- c(1, 1, 1, 4, 5, 3,
                         2, 2, 2, 4, 3, 2,
                         6, 6, 5, 2, 3, 2,
                         5, 5, 5, 1, 1, 1)
# Create the station column
station <- c("1.Spadina", "1.Spadina", "1.Spadina", "1.Spadina", "1.Spadina", "1.Spadina",
             "2.St.George", "2.St.George", "2.St.George", "2.St.George", "2.St.George", "2.St.George",
             "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bl
             "4.Sherbourne", "4.Sherbourne", "4.Sherbourne", "4.Sherbourne", "4.Sherbourne", "4.Sherbou
#
direction <- c("West", "West", "West", "East", "East", "East",
               "West", "West", "West", "East", "East", "East",
               "West", "West", "West", "East", "East", "East",
               "West", "West", "West", "East", "East", "East")
d <- data.frame(direction, station, subjective_distance)
p <- ggplot(d, aes(x=station, y=subjective_distance, fill=direction)) +
       geom_bar(stat = "identity", position = position_dodge())
p
```



Question 11: main effect of station, no main effect of east/west, interaction

```
# Create the subjective distance column
subjective_distance <- c(2, 2, 2, 4, 4, 4,
                         3, 3, 5, 6, 6, 6,
                         4, 4, 4, 3, 3, 3,
                         1, 1, 1, 7, 7, 7)
# Create the station column
station <- c("1.Spadina", "1.Spadina", "1.Spadina", "2.St.George", "2.St.George", "2.St.George",
             "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "4.Sherbourne", "4.Sherbourne", "4.Sherb
             "1.Spadina", "1.Spadina", "1.Spadina", "2.St.George", "2.St.George", "2.St.George",
             "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "4.Sherbourne", "4.Sherbourne", "4.Sherb
#
direction <- c("West", "West", "West", "West", "West", "West",
               "West", "West", "West", "West", "West", "West",
               "East", "East", "East", "East", "East", "East",
               "East", "East", "East", "East", "East", "East")

d <- data.frame(direction, station, subjective_distance)
p <- ggplot(d, aes(x=direction, y=subjective_distance, fill=station)) +
        geom_bar(stat = "identity", position = position_dodge())
p
```

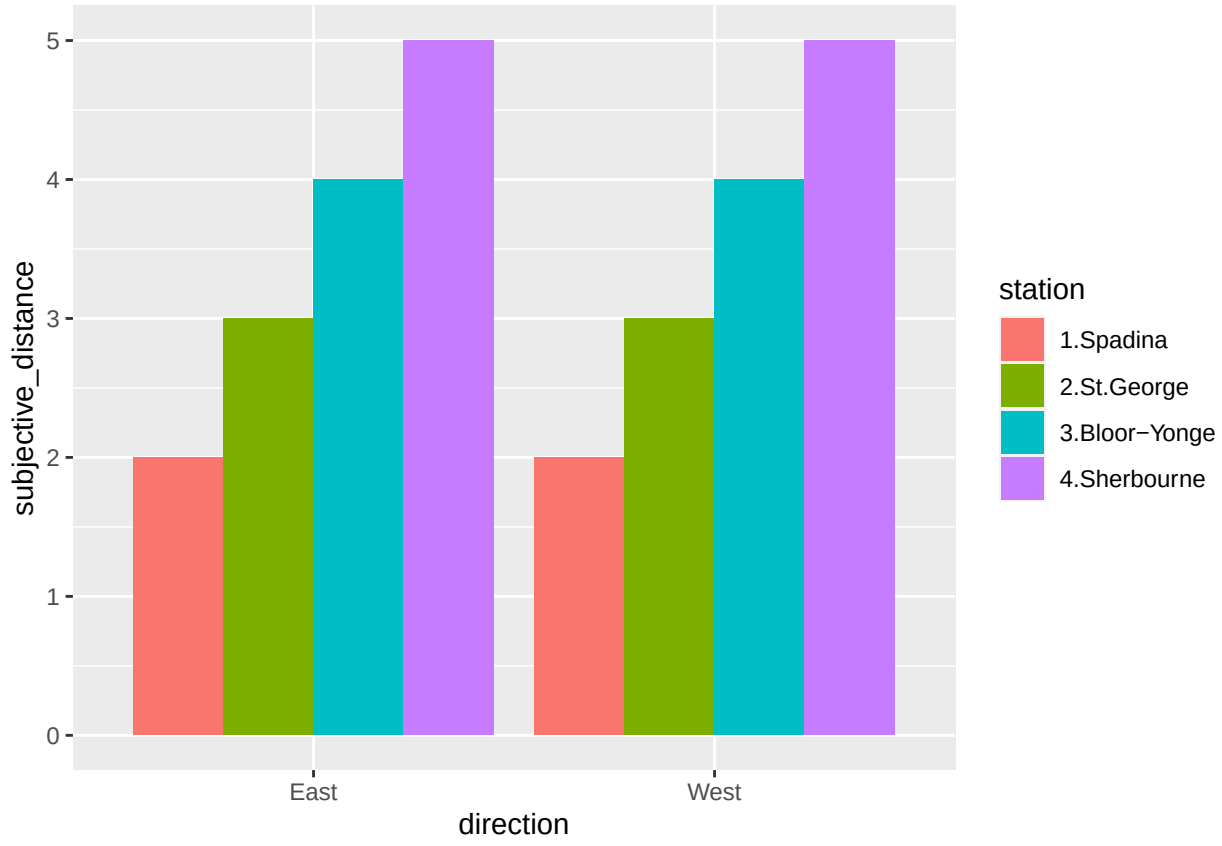



```
# The mean of "1.Spadina" is 3;
# The mean of "2.St.George" is 3.5;
# The mean of "3.Bloor-Yonge" is 2;
# The mean of "4.Sherbourne" is 6.5;

# The mean of "West" is 3.75;
# The mean of "East" is 3.75;
```

Question 12: main effect of station, no main effect of east/west, no interaction

```
# Create the subjective distance column
subjective_distance <- c(1, 1, 2, 2, 2, 3,
                         3, 3, 4, 4, 4, 5,
                         1, 1, 2, 2, 2, 3,
                         3, 3, 4, 4, 4, 5)
# Create the station column
station <- c("1.Spadina", "1.Spadina", "1.Spadina", "2.St.George", "2.St.George", "2.St.George",
             "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "4.Sherbourne", "4.Sherbourne", "4.Sherb
             "1.Spadina", "1.Spadina", "1.Spadina", "2.St.George", "2.St.George", "2.St.George",
             "3.Bloor-Yonge", "3.Bloor-Yonge", "3.Bloor-Yonge", "4.Sherbourne", "4.Sherbourne", "4.Sherb
#
direction <- c("West", "West", "West", "West", "West", "West",
               "West", "West", "West", "West", "West", "West",
               "East", "East", "East", "East", "East", "East",
```

```
                "East", "East", "East", "East", "East", "East")

d <- data.frame(direction, station, subjective_distance)
p <- ggplot(d, aes(x=direction, y=subjective_distance, fill=station)) +
        geom_bar(stat = "identity", position = position_dodge())
p
```



Question 13: In a few sentences, explain what a "main effect" is in this experiment to (a) an advanced undergraduate, and (b) to a five year old.

(a) "main effect" is the effect of one independent variable on the dependent variable. It ignores the effects of any other independent variables.

(b) "main effect" is the thing that can separate you and other kids when you have the other thing same.