HW3

2024-09-22

Homework 3 – Due 28 September 2024

The total points on this homework is 175. Five points are reserved in total for clarity of presentation, punctuation and commenting with respect to the code.

1.

Consider the dataset available in the Excel file at **wind.xls** which contains measurements on wind direction taken at Gorleston, England between 11:00 am and noon on Sundays in the year 1968 (Measurements were not recorded for two Sundays). Note that the data are in angular measurements, and also that the file is in Microsoft Excel format. Therefore, we will need for a way to read the file in a different format.

(a)

The R package readxl1is one providing functionality to read in MS Excel files. Install (if needed) and load the library. Then, read in the file, and assign to a dataframe. *5 points*

```
require(readxl)
```

```
## Loading required package: readxl
```

```
windDf <- as.data.frame(readxl::read_excel("C:/Users/samue/OneDrive/Desktop/Iowa_State_PS/STAT 5790/PS/
```

(b)

Provide descriptive summaries of the measurements such as means, standard deviations, medians, quartiles and inter-quartile ranges. *10 points*

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
seasonsSummary <- sapply(windDf, summary)

iqrSeason <- function(season) {
    quartile1 <- quantile(windDf[[season]], 0.75)
    quartile2 <- quantile(windDf[[season]], 0.25)
    iqr <- quartile1 - quartile2
    iqr[[1]]
}

iqrVals <- sapply(c("Spring", "Summer", "Autumn", "Winter"),
    iqrSeason)
endDf <- rbind.data.frame(seasonsSummary, iqrVals)
rownames(endDf)[rownames(endDf) == "7"] <- "IQR"

endDf
```

```
##           Spring    Summer Autumn   Winter
## Min.      0.0000  10.00000     30  50.0000
## 1st Qu.  55.0000  20.00000    155 205.0000
## Median  185.0000  35.00000    215 255.0000
## Mean    176.6667  80.83333    200 238.3333
## 3rd Qu. 275.0000 150.00000    260 297.5000
## Max.    350.0000 190.00000    350 340.0000
## IQR     220.0000 130.00000    105  92.5000
```

(c)

Given that these are angular data, do any of these descriptive measures above make sense? Why/why not? Think about the average between $1\circ$ and $359\circ$. *5 points*

Some (but not most!) of the above summary statistics make sense. Short and sweet the answer is no though, because even the ones that make some sense still require a degree of transformation to understand. For example: Knowing angular measurements are between 1 and 359 does not give us an idea of **angular dispersion** or **circular mean**, which tend to be used when studying angular measurements. So transforming these using a trigometric method would likely aid in interpretability. But I would argue that what we have, i.e. summary statistics of "raw measurements", make sense insomuch as it tells us that the angular measurements between seasons are different (their range of values, minimum, and maximum). However, we cannot surmise much beyond that without undergoing some transformations, either to a particular statistic or to the underlying data.

(d)

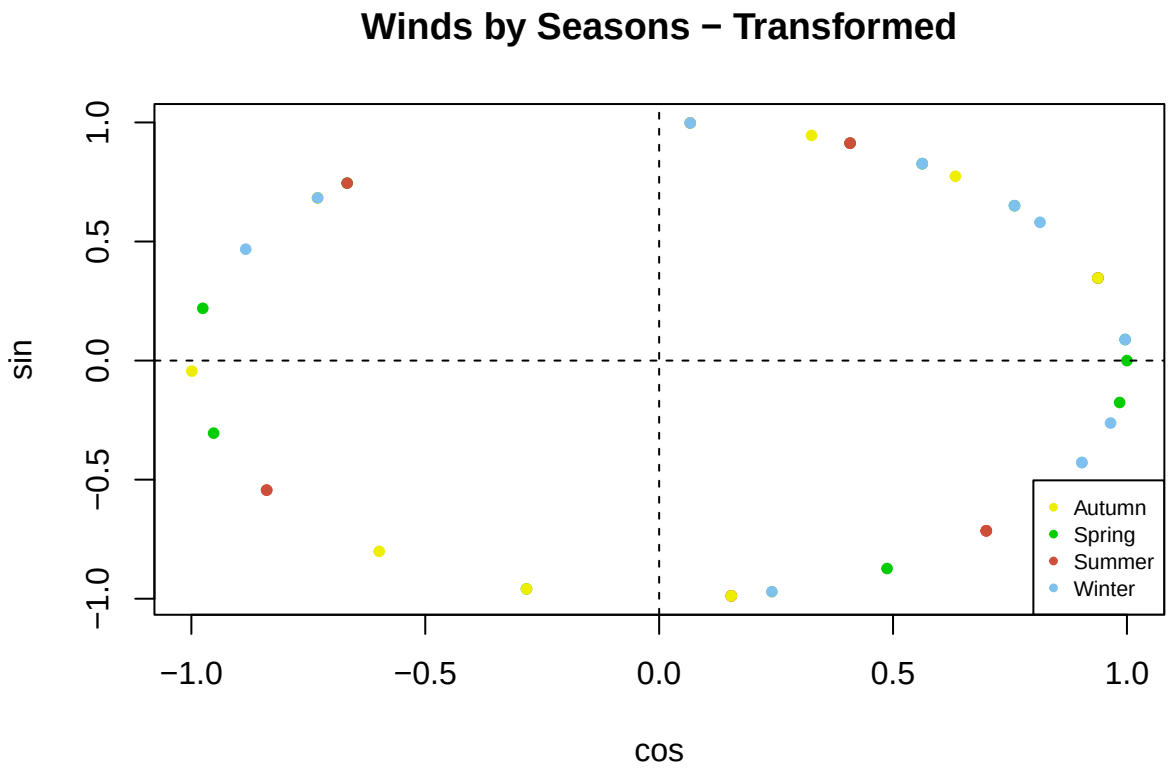
Plot, in one figure, the angular measures, using color for the season. (Note that to obtain a meaningful plot, we need to display angle in terms of a bivariate plot. One way to do so is to use a bivariate direction vector given by (

$$cos(\theta), sin(\theta)$$

) for each angle.) Comment on seasonal differences, if any. *10 points*

```r
springDf <- data.frame(wind = windDf$Spring, season = "Spring")
summerDf <- data.frame(wind = windDf$Summer, season = "Summer")
autumnDf <- data.frame(wind = windDf$Autumn, season = "Autumn")
winterDf <- data.frame(wind = windDf$Winter, season = "Winter")

seasonsDf <- rbind.data.frame(springDf, summerDf, autumnDf, winterDf)
seasonsDf$season <- as.factor(seasonsDf$season)
```

```r
plot(sin(seasonsDf$wind) ~ cos(seasonsDf$wind), xlab = "cos",
    ylab = "sin", main = "Winds by Seasons - Transformed", col = c("yellow2",
        "green3", "tomato3", "skyblue2")[seasonsDf$season], pch = 20)
legend(x = "bottomright", legend = levels(seasonsDf$season),
    col = c("yellow2", "green3", "tomato3", "skyblue2"), pch = 20,
    cex = 0.7)
abline(v = 0, lty = "dashed")
abline(h = 0, lty = "dashed")
```



Observation: The distributions of directions between seasons appear similar upon initial inspection, though some differences become apparent. In particular, if we look at the counts of points in each quadrant of the plot, e.g. upper right quadrant of points between $cos(\theta) \in (0, 1)$ and $sin(\theta) \in (0, 1)$, Autumn and Winter have 5 observations, whereas Summer has only 3 and Spring has at least 6.

Using an alternative formulation, noted below, we can observe these differences more easily, and see there are marked differences between the seasons.

Note: The below code and plot (should be just one!) are included as they aided in analysis of the wind observations. I recognize this is not called for in this question.

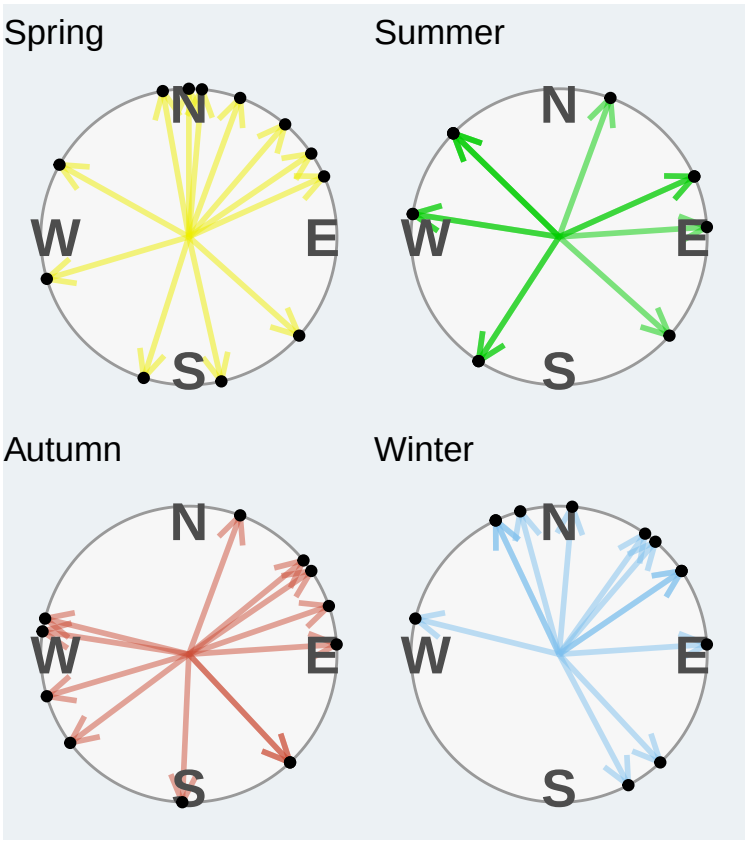
```r
par(mfrow = c(2, 2))
plot(sin(windDf$Spring) ~ cos(windDf$Spring), xlab = "cos", ylab = "sin",
    main = "Spring", col = "yellow2")
abline(v = 0, lty = "dashed")
abline(h = 0, lty = "dashed")
plot(sin(windDf$Summer) ~ cos(windDf$Summer), xlab = "cos", ylab = "sin",
    main = "Summer", col = "green3")
abline(v = 0, lty = "dashed")
abline(h = 0, lty = "dashed")
plot(sin(windDf$Autumn) ~ cos(windDf$Autumn), xlab = "cos", ylab = "sin",
    main = "Autumn", col = "tomato3")
abline(v = 0, lty = "dashed")
abline(h = 0, lty = "dashed")
plot(sin(windDf$Winter) ~ cos(windDf$Winter), xlab = "cos", ylab = "sin",
    main = "Winter", col = "skyblue2")
abline(v = 0, lty = "dashed")
abline(h = 0, lty = "dashed")
```

```r
library(ggplot2)
df <- data.frame(a = seasonsDf$wind%%(2 * pi), b = seasonsDf$season)

ggplot(x) + annotate("rect", xmin = -Inf, xmax = Inf, ymin = 0,
    ymax = 1, fill = "gray97") + geom_hline(yintercept = 1, color = "gray60") +
    geom_segment(aes(x = df$a, xend = df$a, y = 0, yend = 1),
        color = c("yellow2", "green3", "tomato3", "skyblue2")[df$b],
        size = 1, alpha = 0.5, arrow = arrow(angle = 25, length = unit(4,
            "mm"))) + annotate("text", x = 0:3 * pi/2, y = 0.9,
    label = c("N", "E", "S", "W"), size = 7, fontface = 2, color = "gray30") +
    # geom_point(aes(x = a, y = 1)) +
scale_x_continuous(limits = c(0, 2 * pi)) + coord_polar() + theme(legend.position = "bottom",
    plot.background = element_rect(color = NA, fill = "#ecf1f4")) +
    theme_void() + ggtitle("Seasons")
```

```r
library(ggplot2)
library(patchwork)

x1 <- ggplot(data.frame(a = windDf$Spring%%(2 * pi))) + annotate("rect",
    xmin = -Inf, xmax = Inf, ymin = 0, ymax = 1, fill = "gray97") +
    geom_hline(yintercept = 1, color = "gray60") + geom_segment(aes(x = a,
    xend = a, y = 0, yend = 1), color = "yellow2", size = 1,
    alpha = 0.5, arrow = arrow(angle = 25, length = unit(4, "mm"))) +
    annotate("text", x = 0:3 * pi/2, y = 0.9, label = c("N",
        "E", "S", "W"), size = 7, fontface = 2, color = "gray30") +
    geom_point(aes(x = a, y = 1)) + scale_x_continuous(limits = c(0,
    2 * pi)) + coord_polar() + theme_void() + theme(plot.background = element_rect(color = NA,
    fill = "#ecf1f4")) + ggtitle("Spring")
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
x2 <- ggplot(data.frame(a = windDf$Summer%%(2 * pi))) + annotate("rect",
    xmin = -Inf, xmax = Inf, ymin = 0, ymax = 1, fill = "gray97") +
    geom_hline(yintercept = 1, color = "gray60") + geom_segment(aes(x = a,
    xend = a, y = 0, yend = 1), color = "green3", size = 1, alpha = 0.5,
    arrow = arrow(angle = 25, length = unit(4, "mm"))) + annotate("text",
    x = 0:3 * pi/2, y = 0.9, label = c("N", "E", "S", "W"), size = 7,
    fontface = 2, color = "gray30") + geom_point(aes(x = a, y = 1)) +
    scale_x_continuous(limits = c(0, 2 * pi)) + coord_polar() +
    theme_void() + theme(plot.background = element_rect(color = NA,
    fill = "#ecf1f4")) + ggtitle("Summer")

x3 <- ggplot(data.frame(a = windDf$Autumn%%(2 * pi))) + annotate("rect",
    xmin = -Inf, xmax = Inf, ymin = 0, ymax = 1, fill = "gray97") +
    geom_hline(yintercept = 1, color = "gray60") + geom_segment(aes(x = a,
    xend = a, y = 0, yend = 1), color = "tomato3", size = 1,
    alpha = 0.5, arrow = arrow(angle = 25, length = unit(4, "mm"))) +
    annotate("text", x = 0:3 * pi/2, y = 0.9, label = c("N",
        "E", "S", "W"), size = 7, fontface = 2, color = "gray30") +
    geom_point(aes(x = a, y = 1)) + scale_x_continuous(limits = c(0,
    2 * pi)) + coord_polar() + theme_void() + theme(plot.background = element_rect(color = NA,
    fill = "#ecf1f4")) + ggtitle("Autumn")

x4 <- ggplot(data.frame(a = windDf$Winter%%(2 * pi))) + annotate("rect",
    xmin = -Inf, xmax = Inf, ymin = 0, ymax = 1, fill = "gray97") +
    geom_hline(yintercept = 1, color = "gray60") + geom_segment(aes(x = a,
    xend = a, y = 0, yend = 1), color = "skyblue2", size = 1,
    alpha = 0.5, arrow = arrow(angle = 25, length = unit(4, "mm"))) +
    annotate("text", x = 0:3 * pi/2, y = 0.9, label = c("N",
        "E", "S", "W"), size = 7, fontface = 2, color = "gray30") +
    geom_point(aes(x = a, y = 1)) + scale_x_continuous(limits = c(0,
    2 * pi)) + coord_polar() + theme_void() + theme(plot.background = element_rect(color = NA,
    fill = "#ecf1f4")) + ggtitle("Winter")

x1 + x2 + x3 + x4
```


Spring
N
W
E
S
Summer
N
W
E
S
Autumn
N
W
E
S
Winter
N
W
E
S

2.

The Central Limit Theorem (CLT). CLT is considered to be one of the most important results in statistical theory. It states that means of an arbitrary finite distribution are always distributed according to a normal distribution, provided that the sample size, n, for calculating the mean is large enough. To see how big n needs to be we can use the following simulation idea:

(a)

Generate $m = 1000$ samples of size $n = 2$ from a $Uniform(0, 1)$ distribution and storage the samples in a matrix of dimensions 2×1000. (Hint: runif() generates values from a random uniform distribution between 0 and 1.) *6 points*

```
seeds <- 1:1000  # 10 seeds
count <- 0
UniformNums <- replicate(1000, {
    count <<- count + 1
    set.seed(seeds[count])
    runif(2, min = 0, max = 1)
})
matrixDf <- matrix(UniformNums, nrow = 2, ncol = 1000)
dim(matrixDf)
```

```
## [1]    2 1000
```

(b)

Calculate the mean X {Some Math my Knitr Didn't Like} for each sample. (Hint: consider using matrix operations.) *6 points*

```
# Overall
top <- sum(matrixDf)
bottom <- length(matrixDf)
top/bottom
```

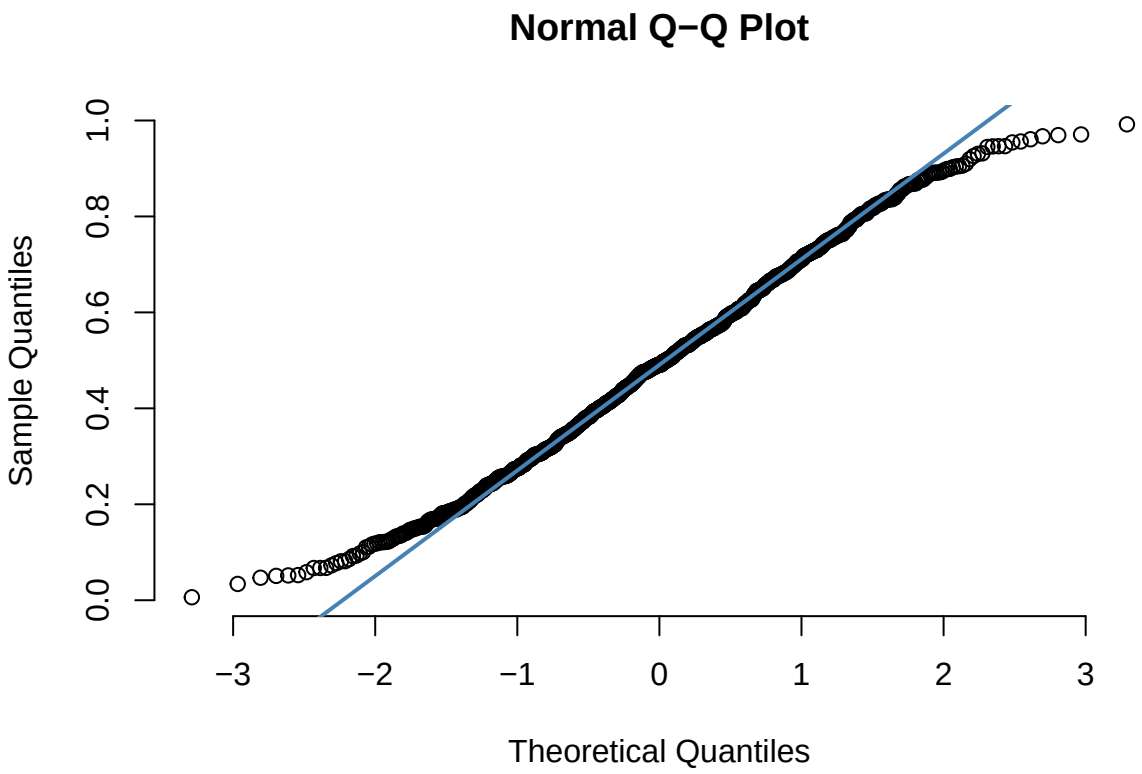
```
## [1] 0.4930478
```

```
# By Sample By column
XsDf <- colMeans(matrixDf)

# colMeans(t(matrixDf))
```

(c)

Draw a QQ-plot for the 1000 Xs to judge the normality. Comment. *5 points*

```
qqnorm(XsDf, pch = 1, frame = FALSE)
qqline(XsDf, col = "steelblue", lwd = 2)
```



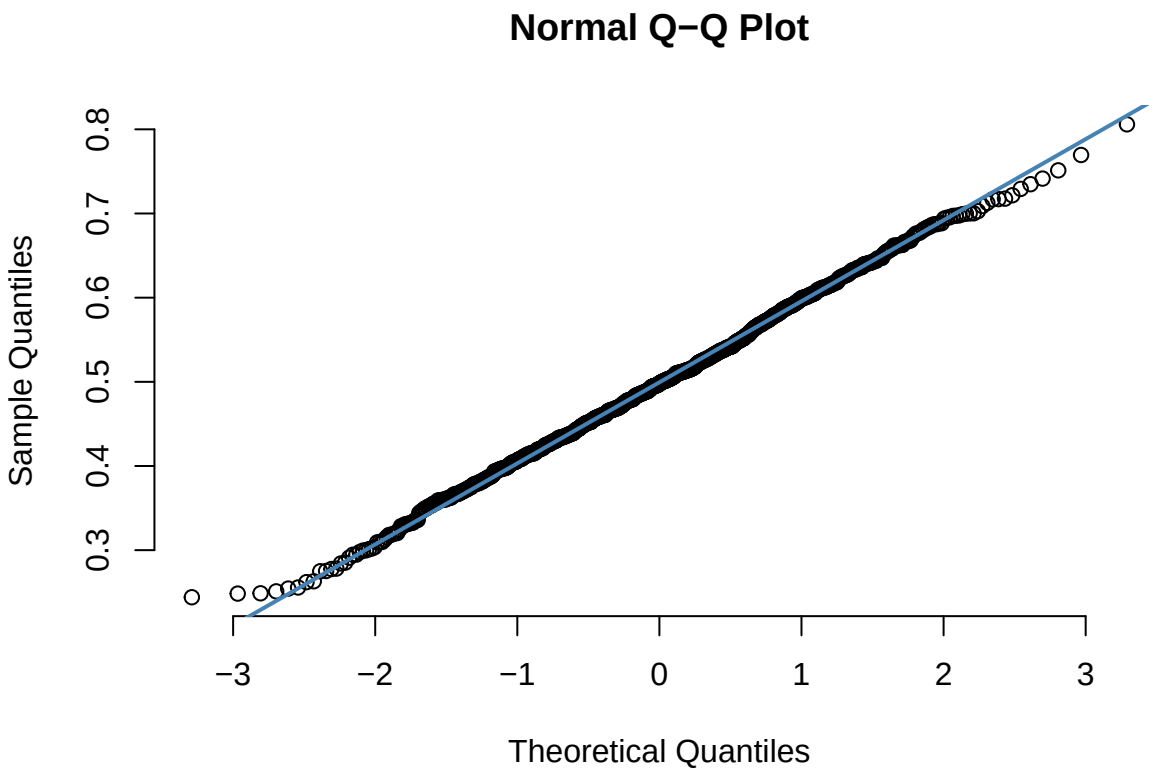

```
# hist(XsDf)
```

We generally have a good fit of normality, but the tails deviate a bit from what we'd empirically/theoretically expect, as in we appear to have the opposite of a heavy tail/potentially negative kurtosis.

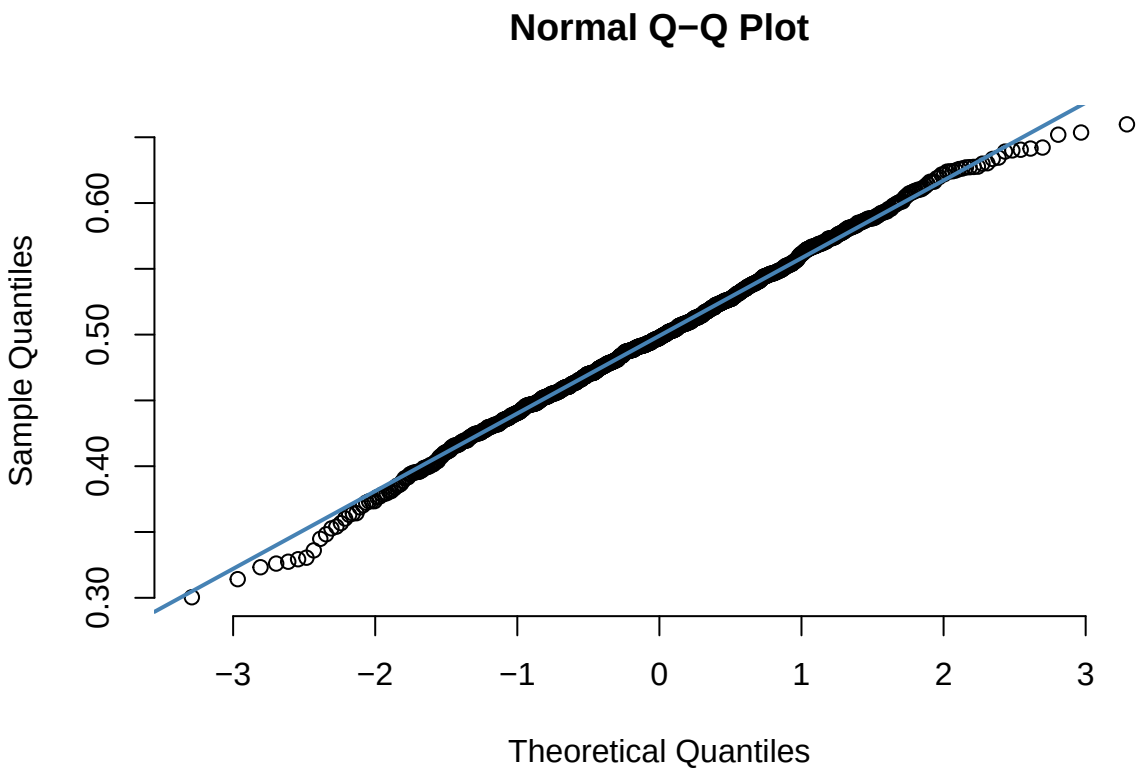
(d)

Repeat the procedure ((a),(b) and (c)) for $n = 10$, 25, and 100 with $m = 1000$. Turn in the QQ-plots and the R code. *10 points*

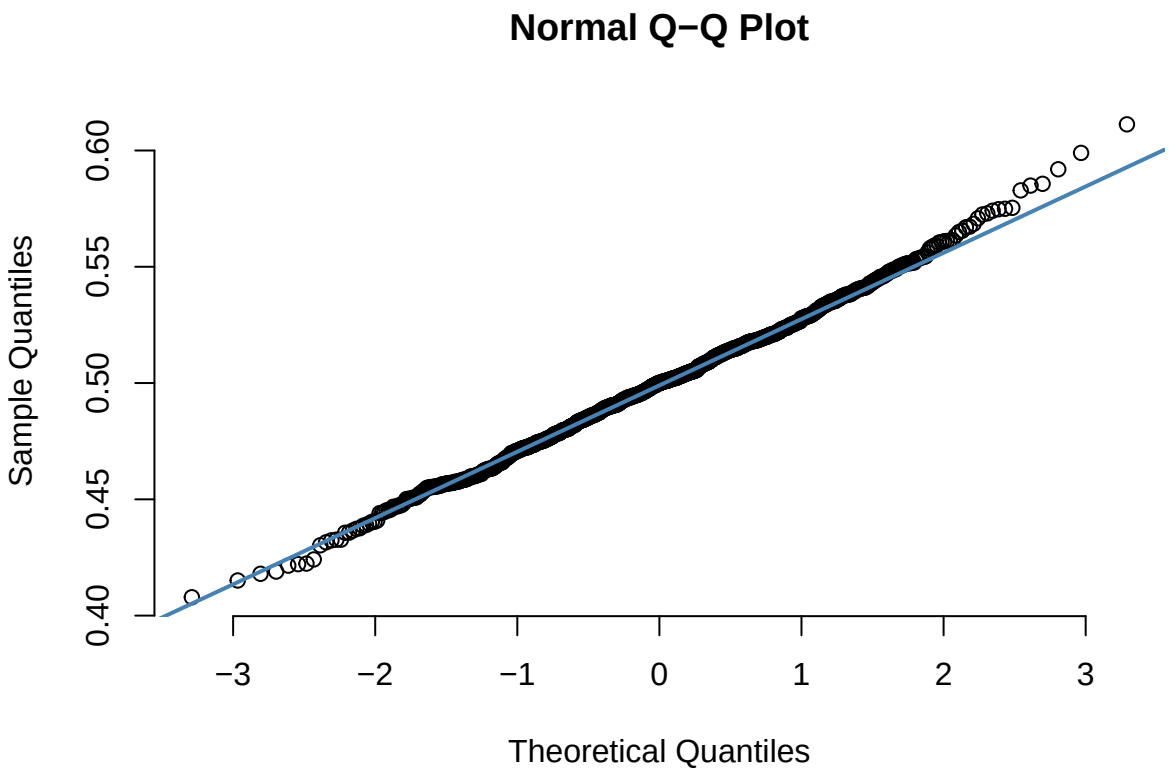
```
seeds <- 1:1000  # 10 seeds
count <- 0
UniformNums <- replicate(1000, {
    count <<- count + 1
    set.seed(seeds[count])
    runif(10, min = 0, max = 1)
})
matrixDf <- matrix(UniformNums, nrow = 10, ncol = 1000)

XsDf <- colMeans(matrixDf)

qqnorm(XsDf, pch = 1, frame = FALSE)
qqline(XsDf, col = "steelblue", lwd = 2)
```




```r
seeds <- 1:1000  # 10 seeds
count <- 0
UniformNums <- replicate(1000, {
    count <<- count + 1
    set.seed(seeds[count])
    runif(25, min = 0, max = 1)
})
matrixDf <- matrix(UniformNums, nrow = 25, ncol = 1000)

XsDf <- colMeans(matrixDf)

qqnorm(XsDf, pch = 1, frame = FALSE)
qqline(XsDf, col = "steelblue", lwd = 2)
```




```r
seeds <- 1:1000  # 10 seeds
count <- 0
UniformNums <- replicate(1000, {
    count <<- count + 1
    set.seed(seeds[count])
    runif(100, min = 0, max = 1)
})
matrixDf <- matrix(UniformNums, nrow = 100, ncol = 1000)

XsDf <- colMeans(matrixDf)

qqnorm(XsDf, pch = 1, frame = FALSE)
qqline(XsDf, col = "steelblue", lwd = 2)
```



(e)

What conclusions can you draw? *3 points*

"Boy, R sure can do some fast simulations!"

Pithy comments aside, the above plots don't look very different from one another. To me, this indicates two things: The first being that having a large m may be more of a contributing factor to achieving normality in the observations, which in turn means that we achieve something "sufficiently normal" for relatively small n. Contrapositively, this to me indicates that we will continue to see (or at least expect to see) fairly normal observations nonetheless contain some possible deivations or lack of a perfect fit even for larger and larger n.

3.

Do you ride a bicycle? (If not, you should seriously consider it.) Bike-sharing is the idea that you can rent a bike at one station and ride it to another where you drop it off. Users are charged by the amount of blocks of time that they have the bike. The data set **bikes.csv** in the Datasets section of Canvas contain information about a bike sharing service in Washington DC. Each row in the dataset is a record on one rental/trip.

For all the answers, provide the R code necessary for a complete solution.

```
knitr::include_graphics("memes.png")
```

(a)

Load the data into R. How many trips were there overall?. How many factor variables are in the data, how many variables are of other kinds? *4 points*

```
library(readr)
bikes <- read_csv("C:/Users/samue/OneDrive/Desktop/Iowa_State_PS/STAT 5790/PS/PS3/bikes.csv")
```

```
## Rows: 77186 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): Start.date, wday, Start.Station, End.Station, Subscriber.Type, Bike.
## dbl (2): Duration, hour
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
nrow(bikes)
```

```
## [1] 77186
```

```
summary(bikes)
```

```
##     Duration       Start.date            wday                hour
##  Min.   :     0   Length:77186       Length:77186       Min.   : 0.00
##  1st Qu.:   420   Class :character   Class :character   1st Qu.:10.00
##  Median :   720   Mode  :character   Mode  :character   Median :15.00
##  Mean   :  1094                                         Mean   :14.28
##  3rd Qu.:  1140                                         3rd Qu.:18.00
##  Max.   :775560                                         Max.   :23.00
##  Start.Station      End.Station        Subscriber.Type       Bike.
##  Length:77186       Length:77186       Length:77186       Length:77186
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

There were 77,186 trips contained in the bikes dataset.

There appear to be 6 Factor Variables in the bikes dataset: - wday - hour (arguable, but there are only a set of values from 0 to 23, discrete and modular) - Start.Station - End.Station - Subscriber.Type - Bike.

And there appears to be 2 non-Factor Variables in the bikes dataset: - Duration (all multiples of 10, weirdly enough) - Start.date (Stored as a character, but can be converted to a continuous date and time)

(b)

The variable Duration contains the length of the bike rental in seconds.

i

How long was the longest rental (converted into days)?. *2 points*

```
# 86400 seconds in a day
max(bikes$Duration)/86400
```

```
## [1] 8.976389
```

8.976389 days, or roughly speaking: 8 days, 23 hours, 26 minutes long.

ii

What other information is available in the data on this trip? *5 points*

```
bikes |>
    filter(Duration == max(Duration))
```

```
## # A tibble: 1 x 8
##   Duration Start.date     wday   hour Start.Station  End.Station Subscriber.Type
##      <dbl> <chr>          <chr> <dbl> <chr>          <chr>       <chr>
## 1   775560 6/4/2014 10:19 Wed      10 Metro Center ~ 3rd & H St~ Casual
## # i 1 more variable: Bike. <chr>
```

This trip started at 10am on a Wednesday, using a W21208 while beginning at the Metro Center, and ended at a different station (3rd & H St NW). This was also a "Casual" subscriber to the service.

iii

How many trips (of all trips) lasted more than one day? *2 points*

```
# 86400 seconds in a day
oneDay <- 86400
bikes |>
    filter(Duration > 86400) |>
    nrow()
```

```
## [1] 7
```

7 trips lasted longer than one day.

(c)

Start.Station describes the start of each trip.

i

From which station did most trips originate? How many trips? *4 points*

```
bikes |>
    group_by(Start.Station) |>
    summarize(count = n()) |>
    arrange(desc(count)) |>
    head()
```

```
## # A tibble: 6 x 2
##   Start.Station                          count
##   <chr>                                  <int>
## 1 Massachusetts Ave & Dupont Circle NW  1595
## 2 Columbus Circle / Union Station       1587
## 3 Lincoln Memorial                      1541
## 4 Jefferson Dr & 14th St SW             1312
## 5 15th & P St NW                        1150
## 6 Thomas Circle                         1073
```

Massachusetts Ave & Dupont Circle NW has most trips originate from it.

ii

Is that the same station at which most trips ended (End.Station)? *3 points*

```
bikes |>
    group_by(End.Station) |>
    summarize(count = n()) |>
    arrange(desc(count)) |>
    head()
```

```
## # A tibble: 6 x 2
##   End.Station                            count
##   <chr>                                  <int>
## 1 Massachusetts Ave & Dupont Circle NW  1736
## 2 Columbus Circle / Union Station       1645
## 3 Lincoln Memorial                      1506
## 4 Jefferson Dr & 14th St SW             1331
## 5 15th & P St NW                        1255
## 6 Thomas Circle                         1063
```

Massachusetts Ave & Dupont Circle NW also was the most frequent end station, so yes, it is both the most common start and stop location.

(d)

When a bike is not returned, the End.Station is marked as "". How often do bikes not get returned? What is reported for the duration of those trips? Change the value of Duration to NA for these records. *4 points*

```
bikes$Duration[is.na(bikes$End.Station)] <- NA

bikes |>
    filter(is.na(End.Station))
```

```
## # A tibble: 1 x 8
##   Duration Start.date     wday   hour Start.Station  End.Station Subscriber.Type
##      <dbl> <chr>          <chr> <dbl> <chr>          <chr>       <chr>
## 1       NA 6/4/2014 18:47 Wed      18 Georgia & New~ <NA>        Registered
## # i 1 more variable: Bike. <chr>
```

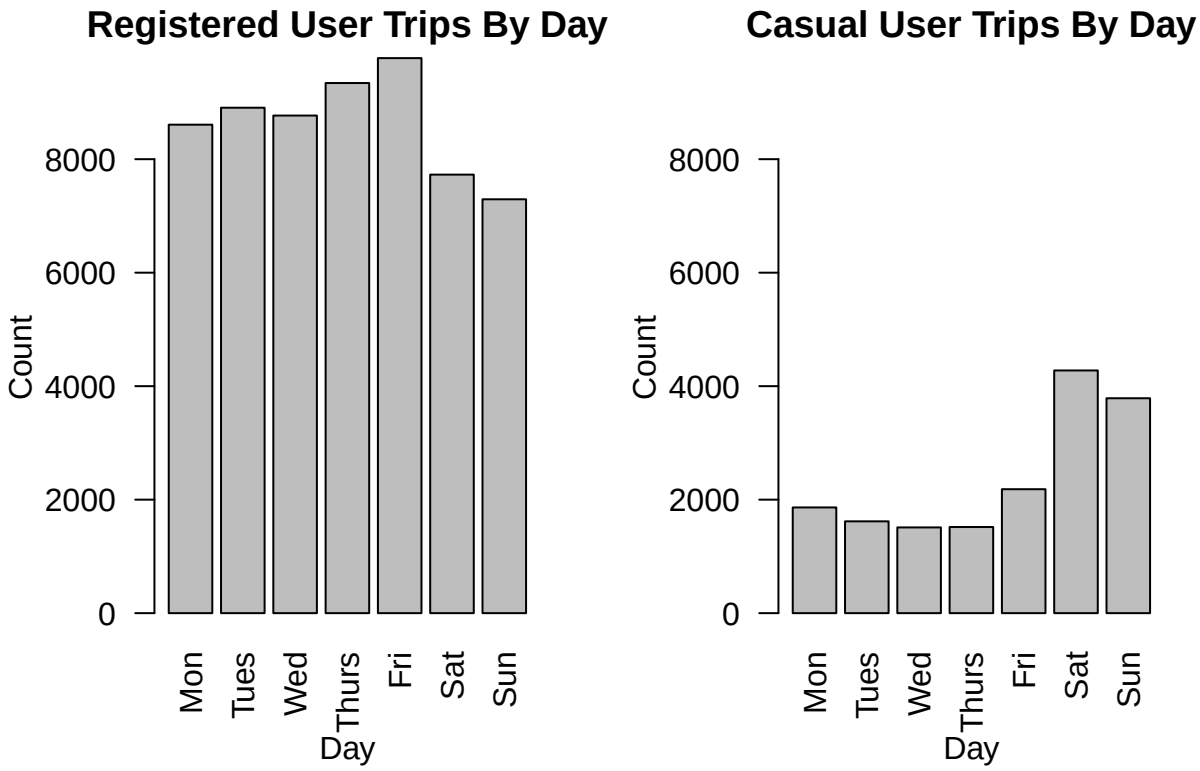
(e)

Plot barcharts of the number of trips on each day of the week, for each Subscriber.Type. Note that one can divide the plotting area using par(mfrow = c(....)) if needed. Make sure that the days of the week (wday) are in the usual order (Start with Mondays). Describe any patterns you see. *6 points*

```
dailySummaries <- bikes |>
    group_by(wday, Subscriber.Type) |>
    summarize(count = n())
```

```
## `summarise()` has grouped output by 'wday'. You can override using the
## `.groups` argument.
```

```
subs <- dailySummaries |>
    filter(Subscriber.Type == "Registered")

casuals <- dailySummaries |>
    filter(Subscriber.Type == "Casual")

days_of_the_week <- c("Mon", "Tues", "Wed", "Thurs", "Fri", "Sat",
    "Sun")
par(mfrow = c(1, 2))

barplot(subs$count ~ factor(subs$wday, days_of_the_week), main = "Registered User Trips By Day",
    ylim = c(0, 9000), xlab = "Day", ylab = "Count", las = 2)
barplot(casuals$count ~ factor(casuals$wday, days_of_the_week),
    main = "Casual User Trips By Day", ylim = c(0, 9000), xlab = "Day",
    ylab = "Count", las = 2)
```



Observations: Between both subscriber types, most Casual subscriber usage spikes during the weekend (Saturday-Sunday), whereas Registered subscriber usage decreases during the weekend. Also, perhaps unsurprisingly, the population of Registered subscribers goes on a trip more often than casual users regardless of the day of the week.

4.

The Titanic. The dataset, available as a comma-separated file on the WWW at **titanic.txt** provides the survival status of passengers on the Titanic, together with their names, age, sex and passenger class. (Note that a good portion of the ages for the 3rd Class passengers is missing.)

Variable	Description
Name	Recorded name of passenger\
PClass	Passenger class: 1st, 2nd or 3rd\
Age	Age in years\
Sex	male or female\
Survived	1 = Yes, 0 = No

(a)

Read in the dataset, using R. Note that the fields are comma-delimited. *3 points*

```
titantic <- read.table(file = "C:/Users/samue/OneDrive/Desktop/Iowa_State_PS/STAT 5790/PS/PS3/titanic.t
    sep = ",", header = TRUE)
```

(b)

Using for instance the function table, cross-classify the passengers by gender and passenger class. Do a further cross-tabulation, perhaps using the same function additionally stratified by survival status. Comment on your findings. *3 + 3 + 3 points*

```
table(titantic$PClass, titantic$Sex)
```

```
##
##       female male
##   1st    143  179
##   2nd    107  173
##   3rd    212  499
```

```
table(titantic$PClass, titantic$Sex, titantic$Survived)
```

```
## , ,  = 0
##
##
##       female male
##   1st      9  120
##   2nd     13  148
##   3rd    132  441
##
## , ,  = 1
##
##
##       female male
##   1st    134   59
##   2nd     94   25
##   3rd     80   58
```

Females in 1st class tended to survive than not, and likewise for females in the 2nd class. However, this trend is inverted for 3rd class females, which is to say more 3rd class females did not survive than survive. Looking at men, we observe more men tended not to survive than survive regardless of class, though the proportion of 1st class men who survived is greater than the proportion of 3rd class men who didn't survive.

(c)

Is there any preliminary evidence that there is a difference in ages among people who survived and those that did not? To answer this question, calculate the mean difference in ages (separately, for both men and women) and the standard errors of the means. What assumptions do you need to make in order to answer this question? *(2 + 3 + 2)× 2 + 4 points*

```
goneMale <- titantic |>
    filter(Survived == 0) |>
    filter(Sex == "male") |>
    na.omit()

survivedMale <- titantic |>
    filter(Survived == 1) |>
    filter(Sex == "male") |>
    na.omit()

goneFemale <- titantic |>
    filter(Survived == 0) |>
    filter(Sex == "female") |>
    na.omit()

survivedFemale <- titantic |>
    filter(Survived == 1) |>
    filter(Sex == "female") |>
    na.omit()

mean(goneMale$Age) - mean(survivedMale$Age)
```

```
## [1] 6.368905
```

```
mean(goneFemale$Age) - mean(survivedFemale$Age)
```

```
## [1] -5.965734
```

```
# sd(goneMale$Age)/sqrt(length((goneMale$Age)))
# sd(survivedMale$Age)/sqrt(length((survivedMale$Age)))
# sd(goneFemale$Age)/sqrt(length((goneFemale$Age)))
# sd(survivedFemale$Age)/sqrt(length((survivedFemale$Age)))
```

```
sqrt(var(goneMale$Age)/(length((goneMale$Age))) + var(survivedMale$Age)/(length((survivedMale$Age))))
```

```
## [1] 1.720795
```

```r
sqrt(var(goneFemale$Age)/(length((goneFemale$Age))) + var(survivedFemale$Age)/(length((survivedFemale$Ag
```

```
## [1] 1.854056
```

Omitting NAs in the dataset, our summary statistics are:

Average Male Difference in Age: 6.368905 years (Dead - Survived) Male Std Error: 1.720795

Average Female Difference in Age: -5.965734 years (Dead - Survived) Female Std Error: 1.854056

Preliminary evidence indicates that women who survived tended to be older on average, and that men who survived tended to be younger on average.

An assumption one should make when supporting the validity of this claim is that not other variables could possibly upend this presumption. In statistical terms, we suppose that observations are independent between groups, i.e. between men and women, and that each of the variables in question is independent of some other variable. For example, one would need to suppose that no other factors could have caused this phenomenon to occur, which is rather difficult to support given past analysis and general consensus over the fact that people in the upper class were more likely to survive compared to lower class. Combine this information with a potential demographic skew of, say, older women tending to belong to the upper class of passengers, and we start to suspect some other factors playing a role in survival.

5.

This problem is a short exercise to get you into manipulating matrices, columns, entries, and some preliminary approaches when dealing with text data (character strings). The objective is to understand how to use available tools to perform our analysis.

The 109th US Congress, comprising the Senate and the House of Representatives, was the legislative branch of the US government from January 3, 2005 to January 3, 2007. During this period, 542 bills were voted on by the US Senate. Each of 100 Senators either voted in favor or against or failed to record their vote on each of these bills. Details voting preferences of the 100 senators on these bills is provided in the file available on Canvas in **senate-109.txt**.

(a)

Read in the file, *noting that the fields are tab-delimited.* Also, *there are apostrophe quotes in some of the field names.* The first column of the file contains the name of the bill and its type, the second column contains the number of missing votes for each bill and the remaining 100 columns contain the votes of each senator (a vote in favor = 1, a vote against = -1, and a no vote = 0). *5 points*

```
senateVote <- read.table(file = "C:/Users/samue/OneDrive/Desktop/Iowa_State_PS/STAT 5790/PS/PS3/senate-
    sep = "\t", quote = "\"", header = TRUE)
```

(b)

Bill type. The field **bill_type_bill_name_bill_ID** contains details on the bills. (Make sure that this field is a vector of character strings.) The first part of this field (before the " ") contains the type of the bill. We will now proceed with obtaining the bill type only, and in doing so, perform a series of operations on character strings.

i

Our objective is to take the above vector of character strings, and for each element, to only keep the portion that contains the string preceding the first " ". There are a few ways to do this, but you can use the function sub() and its allies (see ?sub for examples) which replaces the first time a desired string matches in each element with our choice. Use this to create a new vector of character strings containing only the bill type. *10 points*

```
require(dplyr)
TypeOnlyDf <- senateVote |>
  mutate(
    bill_type_bill_name_bill_ID = gsub("\\_.*","",bill_type_bill_name_bill_ID)
    # the below will extract using a different method
    # and create a new column
    # BillTypeOnly = str_extract(senateVote$bill_type_bill_name_bill_ID, "^[^_]+(?=_)")
  )

unique(TypeOnlyDf$bill_type_bill_name_bill_ID)
```

```
##  [1] "Appropriations"
##  [2] "Budget, Spending and Taxes"
##  [3] "Civil Liberties"
```

```
##  [4] "Energy Issues"
##  [5] "Executive Branch"
##  [6] "Foreign Aid and Policy Issues"
##  [7] "Gun Issues"
##  [8] "Immigration"
##  [9] "Military Issues"
## [10] "National Security Issues"
## [11] "Science and Medical Research"
## [12] "Abortion Issues"
## [13] "Business and Consumers"
## [14] "Education"
## [15] "Environmental Issues"
## [16] "Family and Children Issues"
## [17] "Health Issues"
## [18] "Senior and Social Security Issues"
## [19] "Trade Issues"
## [20] "Transportation Issues"
## [21] "Veterans Issues"
## [22] "Congressional Affairs"
## [23] "Defense"
## [24] "Labor"
## [25] "Legal Issues"
## [26] "Social Issues"
## [27] "Employment and Affirmative Action"
## [28] "Technology and Communication"
## [29] "Agriculture Issues"
## [30] "Campaign Finance and Election Issues"
## [31] "Government Reform"
## [32] "Crime Issues"
## [33] "Arts and Humanities"
## [34] "Regulatory Issues"
## [35] "Drug Issues"
## [36] "Welfare and Poverty"
```

ii

Tabulate the frequency for each type of bill. *5 points*

```
TypeOnlyDf |>
    group_by(bill_type_bill_name_bill_ID) |>
    summarize(count = n()) |>
    arrange(desc(count)) |>
    mutate(prop = paste(round(count/sum(count), 3) * 100, "%"))
```

```
## # A tibble: 36 x 3
##    bill_type_bill_name_bill_ID   count prop
##    <chr>                         <int> <chr>
##  1 Appropriations                  111 20.5 %
##  2 Budget, Spending and Taxes       75 13.8 %
##  3 Executive Branch                 44 8.1 %
##  4 Defense                          29 5.4 %
##  5 Health Issues                    22 4.1 %
##  6 Immigration                      22 4.1 %
```

```
##  7 Abortion Issues                  19 3.5 %
##  8 Foreign Aid and Policy Issues    19 3.5 %
##  9 Energy Issues                    18 3.3 %
## 10 Education                        16 3 %
## # i 26 more rows
```

(c)

Data Quality. The second field contains the number of votes which were not recorded for each senator. We will evaluate if there is any discrepancy. To do so, note that if $\boldsymbol{X}$ is the matrix of votes (only), then the diagonal elements of $\boldsymbol{XX}'$ plus the column of missing votes should match the total number of senators. (Note that there is an "easier"way to do this, using the function apply(), but you are not asked to try that here, since we will be encountering this in detail later.) *5 points*

We expect 100 senators in our dataset.

```
missingMatrix <- as.matrix(TypeOnlyDf$missing_votes, ncol = 1)
voteMatrix <- as.matrix(TypeOnlyDf[, 3:102], nrow = nrow(TypeOnlyDf),
    ncol = 100)
tVote <- t(voteMatrix)

XXt <- voteMatrix %*% tVote
```

```
missingDiag <- diag(XXt) + missingMatrix
```

```
which(missingDiag != 100)
```

```
## integer(0)
```

No discrepancies, as given from the above method. I am suspicious...

Note: The matrix itself was not presented in full because of its dimensions/visual difficulty (a 542 x 1 matrix)

(d)

One issue here is whether we can discern voting trends on different issues. However, every vote here is recorded as a-1/0/1 vote, regardless of whether it is for/neutral/against on a conservative/moderate/liberal issue. For this reason, we will analyze the datasets according to whether senators voted with or against the (majority) leader, Senator Bill Frist (a Republican whose votes are recorded in the last field). Thus, we will convert all the votes of the other senators relative to whether they voted with or against Senator Frist. Do so, using a set of appropriate matrix and vector operations, after eliminating those Bills from consideration where the leader did not record a vote. *15 points*

```
filterDf <- TypeOnlyDf |>
    filter(William.H...Bill..Frist..TN. != 0)

fristOnly <- filterDf |>
    select(William.H...Bill..Frist..TN.) |>
    as.matrix(nrow = 539, ncol = 1)

otherSenators <- filterDf |>
```

```
    select(-c(William.H...Bill..Frist..TN., bill_type_bill_name_bill_ID,
        missing_votes)) |>
    as.matrix(nrow = 539, ncol = 99)
```

```
M <- otherSenators
v <- fristOnly
# positive entries when agreed (-1 * -1 = 1) negative if
# voted against 0 if they voted neutral

result <- M * matrix(v, nrow = nrow(M), ncol = ncol(M), byrow = TRUE)
analysisData <- cbind.data.frame(billType = filterDf$bill_type_bill_name_bill_ID,
    result)
```

(e)

For each bill type identified earlier, tabulate the average number of times senators voted with, against, or indifferently from the Senate majority leader. *10 points*

```
resultsNeutral <- analysisData |>
    group_by(billType) |>
    # Get a summary of a senators votes by billType
summarize(across(everything(), ~sum(. == 0, na.rm = TRUE), .names = "zero_{col}"),
    ) |>
    # mutate so we keep billType for use later now we sum
    # the total of senators across senators
mutate(total_zero = rowSums(across(starts_with("zero_")))) |>
    select(c(billType, total_zero)) |>
    arrange(desc(billType))

resultsFor <- analysisData |>
    group_by(billType) |>
    summarize(across(everything(), ~sum(. == 1, na.rm = TRUE),
        .names = "for_{col}"), ) |>
    mutate(total_for = rowSums(across(starts_with("for_")))) |>
    select(c(billType, total_for)) |>
    arrange(desc(billType))

resultsAgainst <- analysisData |>
    group_by(billType) |>
    summarize(across(everything(), ~sum(. == -1, na.rm = TRUE),
        .names = "against_{col}"), ) |>
    mutate(total_against = rowSums(across(starts_with("against_")))) |>
    select(c(billType, total_against)) |>
    arrange(desc(billType))
```

```
combinedCounts <- mutate(resultsNeutral, resultsAgainst, resultsFor)
```

```
averageVotingProp <- combinedCounts |>
    mutate(totalCount = total_zero + total_against + total_for,
        averageNeutral = round(total_zero/totalCount, 3), averageFor = round(total_against/totalCount,
            3), averageAgainst = round(total_for/totalCount,
```

```
            3)) |>
    select(c(billType, averageFor, averageNeutral, averageAgainst))

averageVotingProp
```

```
averageVotinCounts <- combinedCounts |>
    mutate(totalCount = total_zero + total_against + total_for,
        averageNeutral = round(total_zero/totalCount * 99, 2),
        averageFor = round(total_against/totalCount * 99, 2),
        averageAgainst = round(total_for/totalCount * 99, 2)) |>
    select(c(billType, averageFor, averageNeutral, averageAgainst))

averageVotinCounts
```

```
## # A tibble: 36 x 4
##    billType                          averageFor averageNeutral averageAgainst
##    <chr>                                  <dbl>          <dbl>          <dbl>
##  1 Welfare and Poverty                     18.8           48.8           31.5
##  2 Veterans Issues                         52.7            2             44.3
##  3 Transportation Issues                   20.1           32.4           46.5
##  4 Trade Issues                            35.4           18.8           44.8
##  5 Technology and Communication            18             30.1           50.9
##  6 Social Issues                           29             19.3           50.7
##  7 Senior and Social Security Issues       42.4           12.4           44.1
##  8 Science and Medical Research            44              0             55
##  9 Regulatory Issues                       20.7           46.3           32
## 10 National Security Issues                35.1           15.6           48.2
## # i 26 more rows
```