Hw 3 Solutions

Grade Question 1 easily – so long as their solutions are reasonable or close, give them full credit

Question 1

```
tweets <- scan("https://raw.githubusercontent.com/ds4stats/case-studies/master/twitter-sentiment/Ghostbu
```

Part A

```
dd <- str_remove_all(tweets, "<.*>")
```

Part B

```
rts_idx <- str_detect(tweets, "^RT @")
rts <- tweets[rts_idx]
sum(rts_idx)
```

```
## [1] 3056
```

Part C

```
norts <- tweets[!rts_idx]

# This includes hate and hated since hate a subset of hated
str_detect(str_to_lower(norts), pattern = "hate") %>% sum()
```

```
## [1] 44
```

```
str_detect(str_to_lower(norts), pattern = "bad") %>% sum()
```

```
## [1] 36
```

```
str_detect(str_to_lower(norts), pattern = "lo+ve") %>% sum()
```

```
## [1] 122
```

Part D

```
ll <- str_extract_all(tweets, pattern = "https://t\\.co/[:alnum:]{10}") %>% unlist()
ll <- str_remove(ll, "https://t\\.co/")

(lla <- str_count(ll, "[:alpha:]") %>% sum() / (26 * 2)) # 52 total capital/lower case letters
```

```
## [1] 525.52
```

```
(lln <- str_count(ll, "[:digit:]") %>% sum() / 10) # 10 numbers 0-9
```

```
## [1] 604.3
```

```
(la <- str_count(ll, "[:lower:]") %>% sum() / 26)
```

```
## [1] 514.81
```

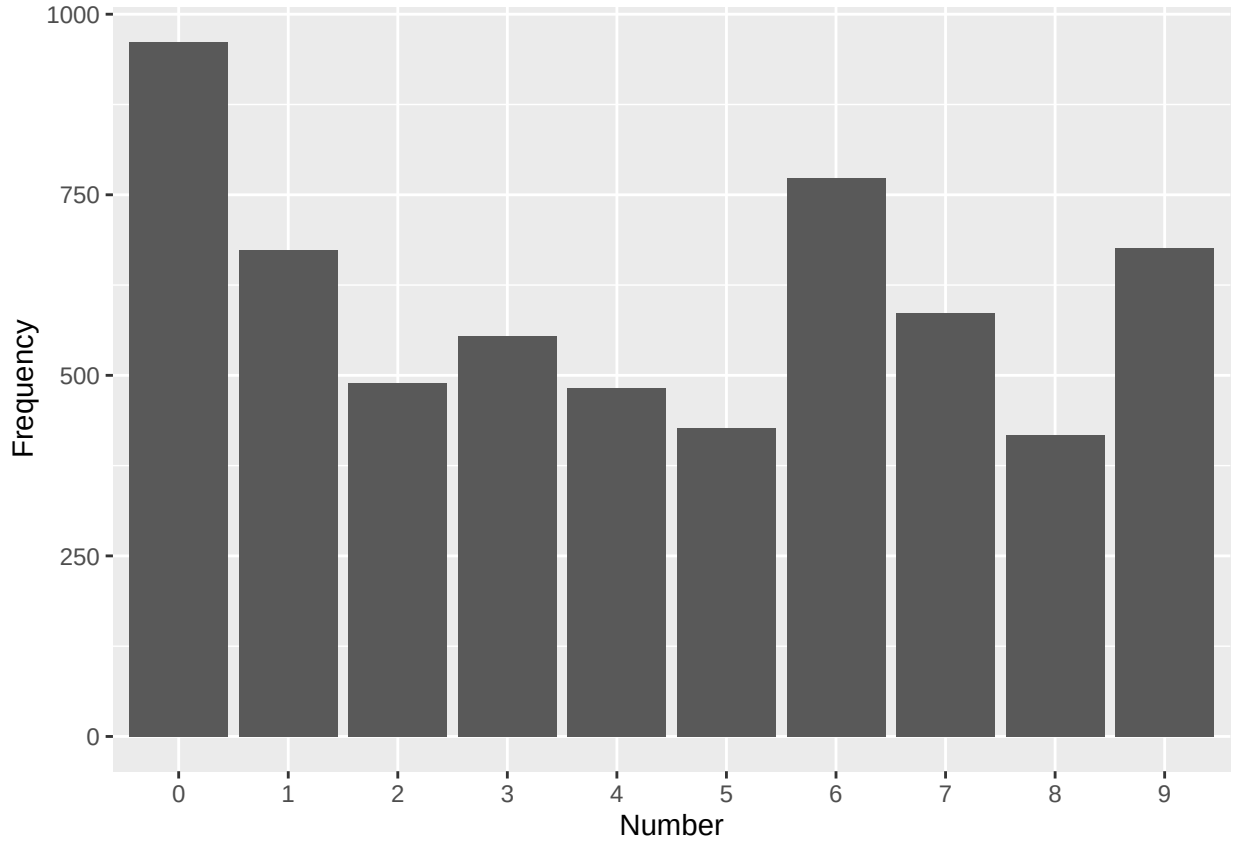
```
(lu <- str_count(ll, "[:upper:]") %>% sum() / 26)
```

```
## [1] 536.23
```

```
(llt <- str_count(ll, "[:alnum:]") %>% sum() / (10 + 26*2))
```

```
## [1] 538.23
```

```
# test
nn <- str_extract_all(ll, "[:digit:]") %>% unlist() %>% table()
df <- data.frame(N = nn)
colnames(df) <- c("Number", "Frequency")
ggplot(df, aes(x = Number, y = Frequency)) + geom_bar(stat = "identity")
```



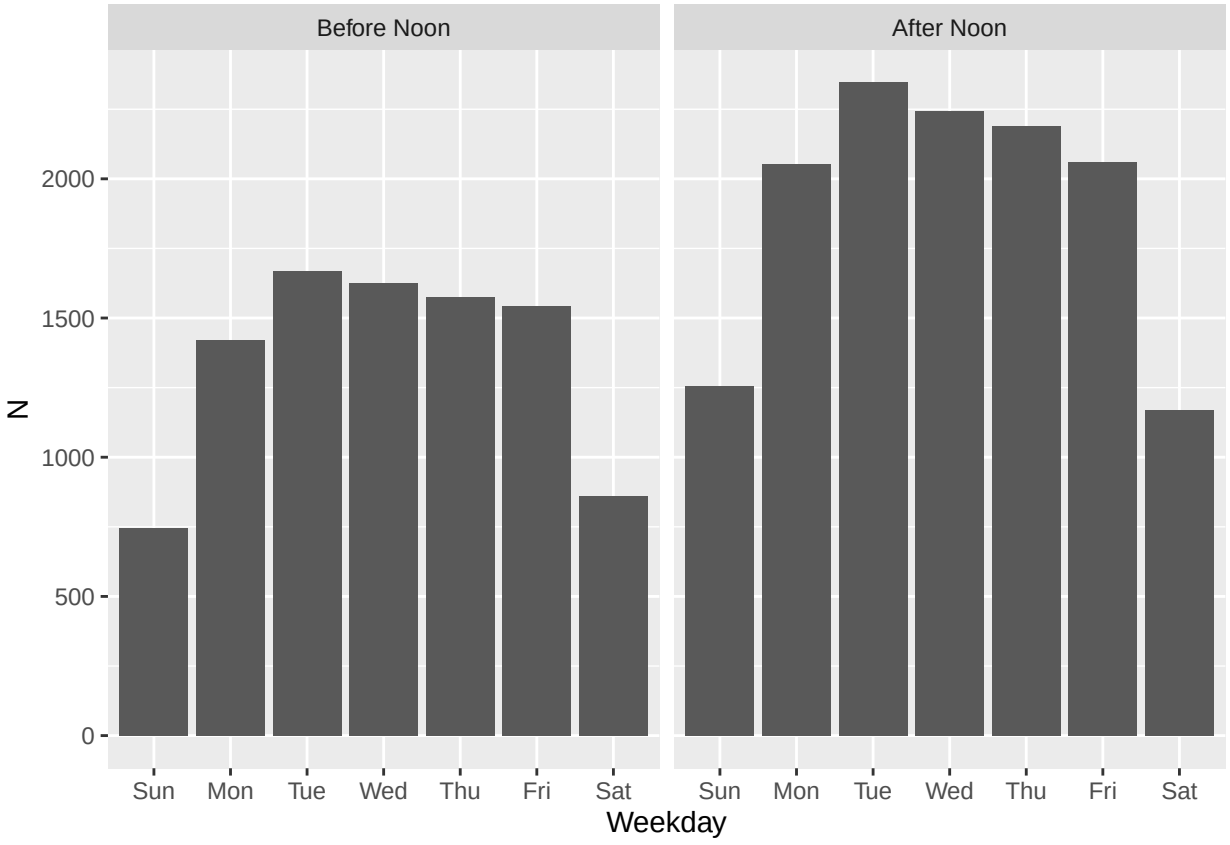
Question 2

```
ny_stories <- read.csv("https://storybench.org/reinventingtv/abc7ny.csv")
ca_stories <- read.csv("https://storybench.org/reinventingtv/kcra.csv")
combined <- rbind(data.frame(ny_stories, location = "NY"), data.frame(ca_stories, location = "CA"))
```

Part A

```
tt <- combined %>%
  mutate(weekday = wday(mdy_hm(datetime), label = TRUE),
         afterNoon = hour(mdy_hm(datetime)) >= 12) %>%
  group_by(weekday, afterNoon) %>%
  summarize(N = n()) %>% na.omit()


ggplot(tt, aes(x = weekday, y = N)) + geom_bar(stat = "identity") + facet_wrap(~afterNoon, labeller = as
```



Part B

```
combined %>%
  mutate(containsChina = str_detect(headline, "China"),
         containsRus = str_detect(headline, "Russia"),
         containsGermany = str_detect(headline, "Germany")) %>%
  group_by(location) %>%
  summarize(China = sum(containsChina),
            Russia = sum(containsRus),
            Germany = sum(containsGermany)) -> tt #%>% knitr::kable()
knitr::kable(tt)
```

location	China	Russia	Germany
CA	7	67	2
NY	24	17	2

Part C

I originally asked for total number of capital words but this ended up being very ambiguous. Anything reasonable should get full credit.

```
combined %>%
  mutate(month = month(mdy_hm(datetime)),
         capitalwords = str_count(teaser, "[:upper:]{1}")) %>% # So that AA, for example, counts as one
  group_by(month) %>%
  summarize(totalCap = sum(capitalwords)) %>% na.omit()
```

```
## # A tibble: 7 x 2
##   month totalCap
##   <dbl>    <int>
## 1     1     8034
## 2     7     7496
## 3     8    17751
## 4     9    19875
## 5    10    20924
## 6    11    15625
## 7    12    16261
```