STA3115_homework1

Jaehun Shon

2024-09-16

Problem 1: Splitting the "time" variable

```
billboard %>%
  separate(time, into = c("minutes", "seconds", "mm"), sep = ":") %>%
  .[, c("year", "artist.inverted", "track", "minutes", "seconds")] -> billboard.1

billboard.1 %>%
  head(10) %>%
  kable(col.names = c("Year", "Artist", "Track", "Minutes", "Seconds"),
        caption = "Top 10 rows of Billboard Data",
        format = "markdown")
```

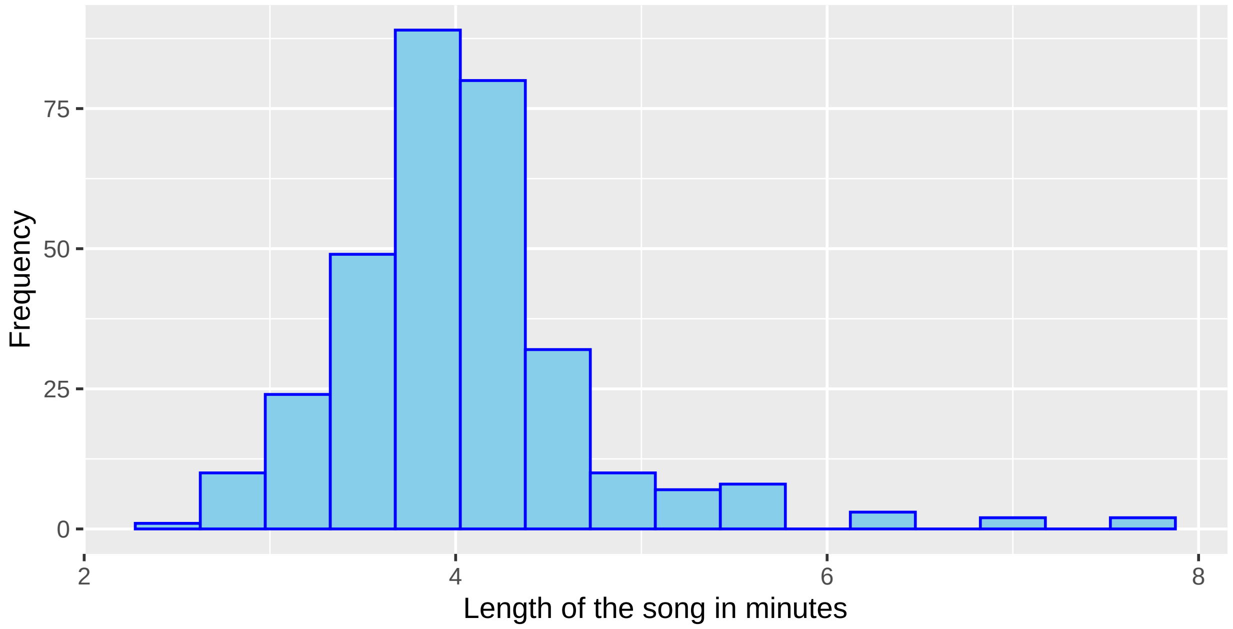
Table 1: Top 10 rows of Billboard Data

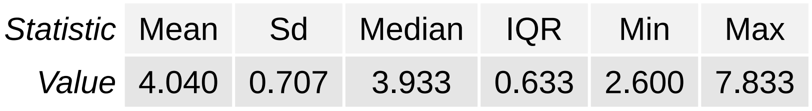
Year	Artist	Track	Minutes	Seconds
2000	Destiny's Child	Independent Women Part I	03	38
2000	Santana	Maria, Maria	04	18
2000	Savage Garden	I Knew I Loved You	04	07
2000	Madonna	Music	03	45
2000	Aguilera, Christina	Come On Over Baby (All I Want Is You)	03	38
2000	Janet	Doesn't Really Matter	04	17
2000	Destiny's Child	Say My Name	04	31
2000	Iglesias, Enrique	Be With You	03	36
2000	Sisqo	Incomplete	03	52
2000	Lonestar	Amazed	04	25

Problem 2: Creating a new time variable and histogram

```
billboard.1 |>
  mutate(time_in_min = as.numeric(minutes) + as.numeric(seconds) / 60) -> timevar

t <- timevar$time_in_min
description <- data.frame(
  Statistic = c("Mean", "Sd", "Median", "IQR", "Min", "Max"),
  Value = c(mean(t), sd(t), median(t), IQR(t), range(t)) %>% round(., 3)
) %>% t
```

```r
histplot <- ggplot(timevar, aes(x = time_in_min)) +
  geom_histogram(fill = "skyblue", color = "blue", binwidth = 0.35) +
  labs(x = "Length of the song in minutes",
       y = "Frequency")

table_grob <- tableGrob(description)
final_plot <- grid.arrange(histplot, table_grob, nrow = 2, heights = c(3/4, 1/4))
```





Statistic	Mean	Sd	Median	IQR	Min	Max
Value	4.040	0.707	3.933	0.633	2.600	7.833

Problem 3: Summary statistics by weeks on chart

Table 2: Comparison of Statistics for Below 15 weeks and More than 15 weeks

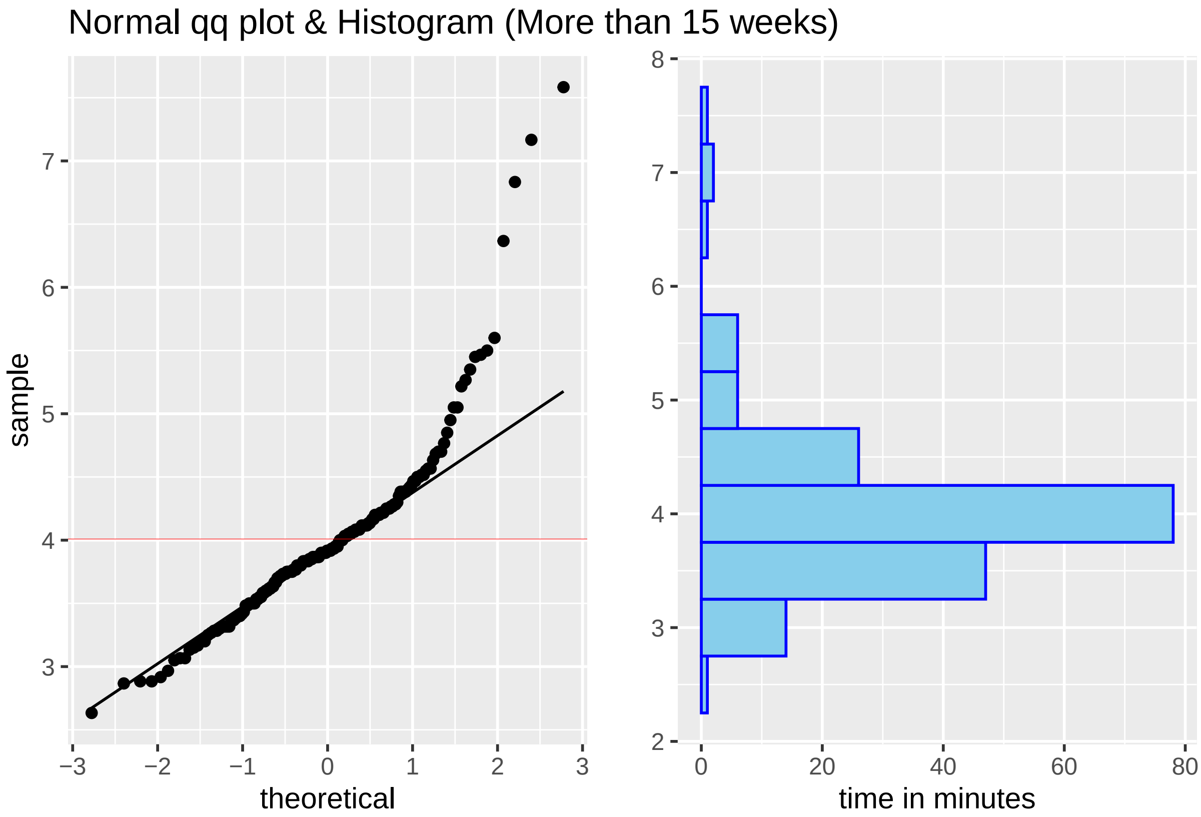
Statistic	Below 15 weeks	More than 15 weeks
Mean	4.0825926	4.0091575
Standard Deviation	0.7193945	0.6974643
Median	4.0166667	3.9166667
IQR	0.6333333	0.6083333

Problem 4: Comparing song length distribution

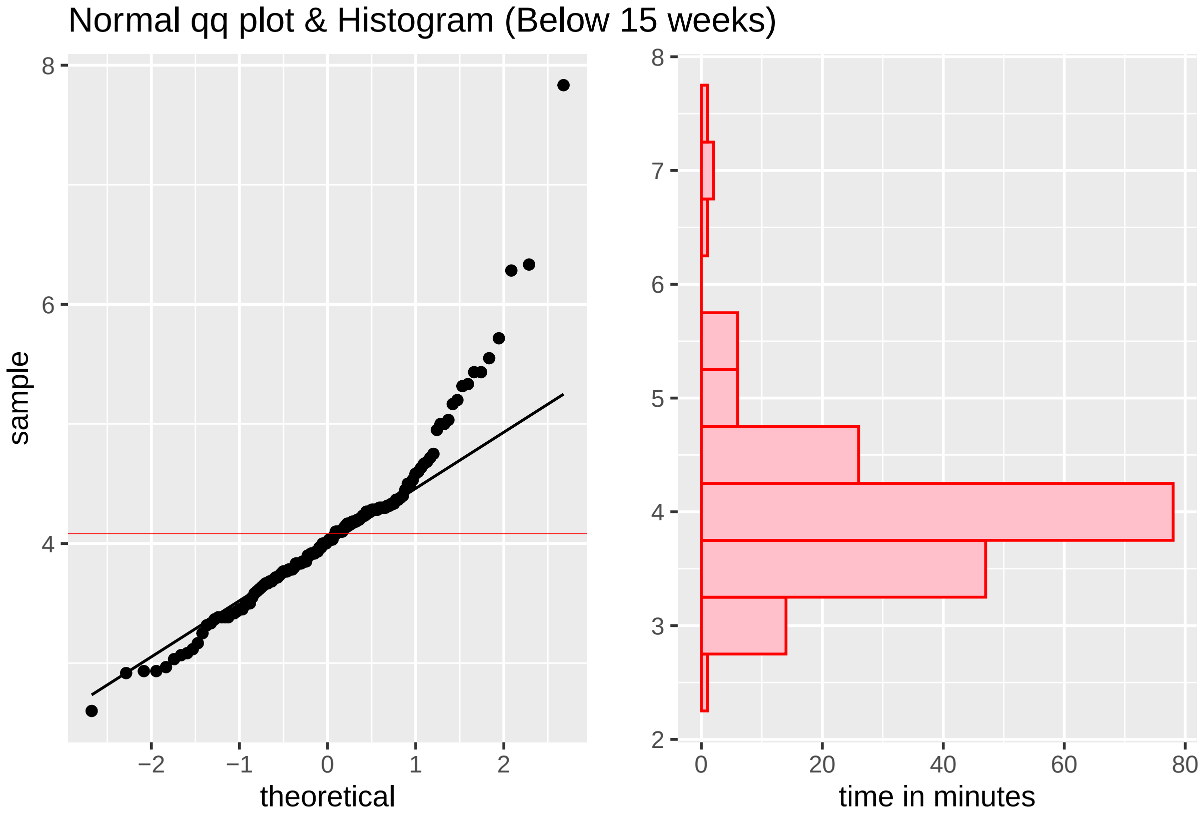
```r
plot1 <- ggplot(billboard.2, aes(x = time_in_min)) +
  geom_density(aes(fill = ifelse(!is.na(x16th.week), "More than 15 weeks", "Below 15 weeks"),
                   color = ifelse(!is.na(x16th.week), "More than 15 weeks", "Below 15 weeks")),
               alpha = 0.7, adjust = 0.7) +
  labs(x = "Length of the song in minutes", y = "Density",
       fill = "", color = "") +
  scale_fill_manual(values = c("More than 15 weeks" = "skyblue", "Below 15 weeks" = "pink")) +
  scale_color_manual(values = c("More than 15 weeks" = "blue", "Below 15 weeks" = "red")) +
  theme_minimal()


qq.df <- as.data.frame(qqplot(a, b, plot.it = FALSE))
plot2 <-ggplot(qq.df, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, color = "black") +
  labs(x = "Below 15 weeks", y = "More than 15 weeks",
       title = "QQ Plot: Below 15 weeks vs. More than 15 weeks") +
  xlim(range(c(a, b))) +
  ylim(range(c(a, b))) +
  theme_minimal()

pp <- plot1 + plot2
```

Problem 5: Comparing to the Normal distribution

```r
qq_plot <- ggplot() +
  stat_qq(data = billboard.2 %>% filter(!is.na(x16th.week)),
          aes(sample = time_in_min)) +
  stat_qq_line(data = billboard.2 %>% filter(!is.na(x16th.week)),
               aes(sample = time_in_min), linewidth = 0.5) +
  geom_abline(aes(intercept = mean(b), slope = 0), color = 'red', linewidth = 0.1) +
  labs(x = "theoretical", y = "sample", title = 'Normal qq plot & Histogram (More than 15 weeks)')

hist_plot <- ggplot() +
  geom_histogram(data = billboard.2 %>% filter(!is.na(x16th.week)),
                 aes(x = time_in_min),
                 binwidth = 0.5, fill = "skyblue", color = "blue") +
  coord_flip() +
  labs(x = "", y = "time in minutes")

qq_plot + hist_plot
```



```
qq_plot <- ggplot() +
  stat_qq(data = billboard.2 %>% filter(is.na(x16th.week)),
          aes(sample = time_in_min)) +
  stat_qq_line(data = billboard.2 %>% filter(is.na(x16th.week)),
               aes(sample = time_in_min), linewidth = 0.5) +
  geom_abline(aes(intercept = mean(a), slope = 0), color = 'red', linewidth = 0.1) +
  labs(x = "theoretical", y = "sample", title = 'Normal qq plot & Histogram (Below 15 weeks)')

hist_plot <- ggplot() +
  geom_histogram(data = billboard.2 %>% filter(!is.na(x16th.week)),
                 aes(x = time_in_min),
                 binwidth = 0.5, fill = "pink", color = "red") +
  coord_flip() +
  labs(x = "", y = "time in minutes")

qq_plot + hist_plot
```


Normal qq plot & Histogram (Below 15 weeks)
sample
8
6
4
−2
−1
0
1
2
theoretical
8
7
6
5
4
3
2
0
20
40
60
80
time in minutes