Problem Set 1 Code

2025-11-13

Task 1

In the following a descriptive analysis of the "detailed_fish_market_data", regarding Whiting is conducted. At first the required packages are loaded.

```
## Load packages and dataset
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)

detailed <- read_tsv("../data/detailed_fish_market_data.txt")
```

Data preparation

As the dataset has some observations, that are "NA" or relatively obvious outliers, those are removed first. Especially the two obersvations, that seem to stem from another dealer are removed.

```
# delete those rows that have NA for
# "price","quan", "totr", "tots" and
# filter for whiting (no king)
detailed_whiting <- detailed %>%
  filter(!is.na(pric),
         !is.na(quan),
         !is.na(totr),
         type == "w") %>%
  arrange(date)


# There seem to be two entries in the dataset, where there are two dealer per day.
# Since this is the case only for two out of all days in April and May: drop those two observations
detailed_whiting <- detailed_whiting %>%
  # frequency of the same tots value for different days
  group_by(date, dayw, tots) %>%
  mutate(n_same_tots = n()) %>%

  # number of distinct tots days
  group_by(date, dayw) %>%
  mutate(n_tots_values = n_distinct(tots)) %>%
  ungroup() %>%

  # delete rows for which (there are multiple different tots values
```

```
  #                                     AND
  #                          for which the tot value only appears once)
  filter(!(n_tots_values > 1 & n_same_tots == 1)) %>%

  # delete rows that are not longer needed
  select(-n_same_tots, -n_tots_values)


## two cases, where > 1 dealer is present
tots_inconsistent <- detailed_whiting |>
  group_by(date, dayw) |>
  mutate(
    n_tots = n_distinct(tots)
  ) |>
  filter(n_tots > 1) |>
  arrange(date, dayw, tots, totr)
```

Besides some definitions for plot formatting in the following the dataset for the analysis on daily level is created.

```
# theme for plots
theme_fontsize <- theme(
  plot.title = element_text(size = 14),
  plot.subtitle = element_text(size = 10),
  axis.title = element_text(size = 12),
  axis.text = element_text(size = 11),
  legend.text = element_text(size = 11),
)



# dataset for the daily-level
detailed_whiting_daily <- detailed_whiting %>%
  group_by(date) %>%
  summarise(
    avg_pric = mean(pric),
    totr = first(totr),
    tots = first(tots),
    dayw = first(dayw),
    n_trsact = n(),
    strate = first(tots)/first(totr)
  )

# labels for time series data plots
date_seq <- seq(
  from = as.Date("1992-04-06"),
  to   = as.Date("1992-05-15"),
  by   = "day"
)

# format as "MM-YYYY"
day_labels <- format(date_seq, "%d-%m")
```

```
# Named character vector: names are month_ids
day_lookup_vec <- setNames(day_labels, c(seq(406,430, by = 1),seq(501,515, by=1)))

break_vec_x_axis <- c(seq(406,430, by = 7),seq(504,515, by=7))
all_days_x_axis <- c(seq(406,430, by = 1),seq(501,515, by=1))
```

Descriptive analysis

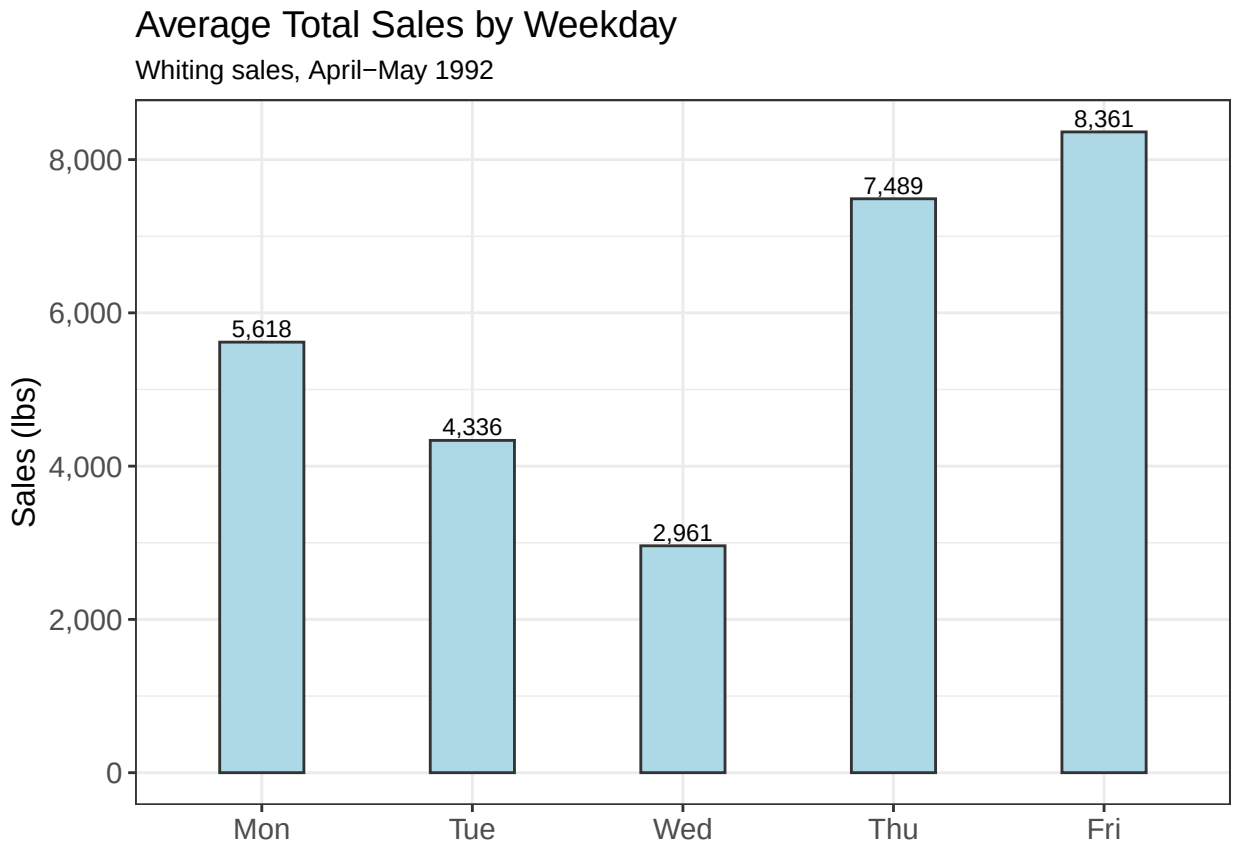
```
####
# summary of the daily datset
####
detailed_whiting_daily %>%
  select(totr, tots, n_trsact) %>%
  summary()
```

```
##       totr            tots          n_trsact
##  Min.   :  200   Min.   :  490   Min.   : 4.00
##  1st Qu.: 1990   1st Qu.: 3360   1st Qu.:18.00
##  Median : 6080   Median : 5535   Median :25.00
##  Mean   : 5881   Mean   : 5760   Mean   :25.05
##  3rd Qu.: 7927   3rd Qu.: 7495   3rd Qu.:32.50
##  Max.   :15940   Max.   :15455   Max.   :57.00
```

As the summary indicates, total sales and therefore the total received amount of Whiting in lbs inherit a large amount of variation. The amount of transaction per day also shows a broad variety of values, with the minimum of four and a maximum of 57.

To gain a first insight in the properties of the total sales of Whiting in the period of April to May 1992 a bar-chart and a time-series plot are used.

```
####
# barchart average sales by dayw (Day of the Week)
####
detailed_whiting_daily %>%
  group_by(dayw) %>%
  summarise(avg_tots = mean(tots, na.rm = TRUE)) %>%
  ggplot(aes(x = factor(dayw), y = avg_tots)) +
  geom_col(fill = "lightblue", colour = "grey20", width = 0.4) +
  labs(title = "Average Total Sales by Weekday",
       subtitle = "Whiting sales, April-May 1992",
       y = "Sales (lbs)",
       x = NULL)+
  geom_text(
    aes(label = scales::comma(round(avg_tots, 0))),
    vjust = -0.3,
    size = 3
  ) +
  scale_x_discrete(breaks = 1:5,
                   labels = c("Mon", "Tue", "Wed", "Thu", "Fri")) +
  scale_y_continuous(labels = scales::comma) +
  theme_bw() +
  theme_fontsize
```




```
####
# time series of tots (total sales)
####
tots_plot_df <- detailed_whiting_daily %>%
  select(date, tots) %>%
  complete(date = 406:515,
           fill = list(tots = 0)) %>%
  arrange(date) %>%
  filter(!between(date, 431, 500)) %>%
  mutate(date_fac = factor(date, levels = date))


ggplot(tots_plot_df, aes(x=date_fac, y = tots, group = 1)) +
  geom_col(width = 0.2,
           colour = "lightblue",
           fill="grey10") +
  labs(title = "Daily Total Sales over Time",
       subtitle = "Whiting sales, April-May 1992",
       y = "Sales (lbs)",
       x = NULL)+
  scale_x_discrete(breaks = as.character(all_days_x_axis),
                   labels = function(x) {
                     lab <- rep("", length(x))
                     sel <- x %in% as.character(break_vec_x_axis)
                     lab[sel] <- day_lookup_vec[x[sel]]
                     lab
```

```
                  }) +
theme_bw() +
theme_fontsize
```