Data Manipulation and Visualization on Milk Data

GulnurUzun

2024-03-29

Calling the Libraries to be Used in the Case

```
library(dplyr)
library(ggplot2)
library(tidyr)
```

Read the Milk.csv File

MilK.CSV file have read with **read.table()** function, **rownames have removed** from this table.

```
milk.data = read.table("Milk.csv", header = TRUE, sep = ",", row.names = 1)
head(milk.data)
```

```
##   protein Time Cow   Diet
## 1    3.63    1 B01 barley
## 2    3.57    2 B01 barley
## 3    3.47    3 B01 barley
## 4    3.65    4 B01 barley
## 5    3.89    5 B01 barley
## 6    3.73    6 B01 barley
```

Assign the Centered Plot Title to centered.plot.title

centered.plot.title will used in next parts for title move the center.

```
centered.plot.title = theme(plot.title = element_text(hjust = 0.5))
```

Part a: Boxplot of Protein Measurements by Feeding Strategy (All Data)

In the feeding strategy versus protein measurements graph **(Figure 1)**, if cows have **barley type of the diet**, **maximum mean protein measurments** about the diet type have observed, however if cows have **both barley and lupins type of diet**, **maximum protein measurment have observed**.

```
boxplot.protein.vs.diet = ggplot(data = milk.data)+
  aes(x = Diet, #categoric attribe in x-axes
      y = protein, #numeric attribute in y-axes
      color = Diet)+
  geom_boxplot()+ #boxplot have added
  geom_point()+ #data points have added
  geom_jitter(alpha=0.4)+ #density of data point, with alpha option have added the transparen
cy
  labs(title = "Boxplot of Protein Measurements by Feeding Strategy (All Data Points)",
       x = "Feeding Strategy",
       y = "Protein Measurements")+
  centered.plot.title #title have set to center
boxplot.protein.vs.diet
```

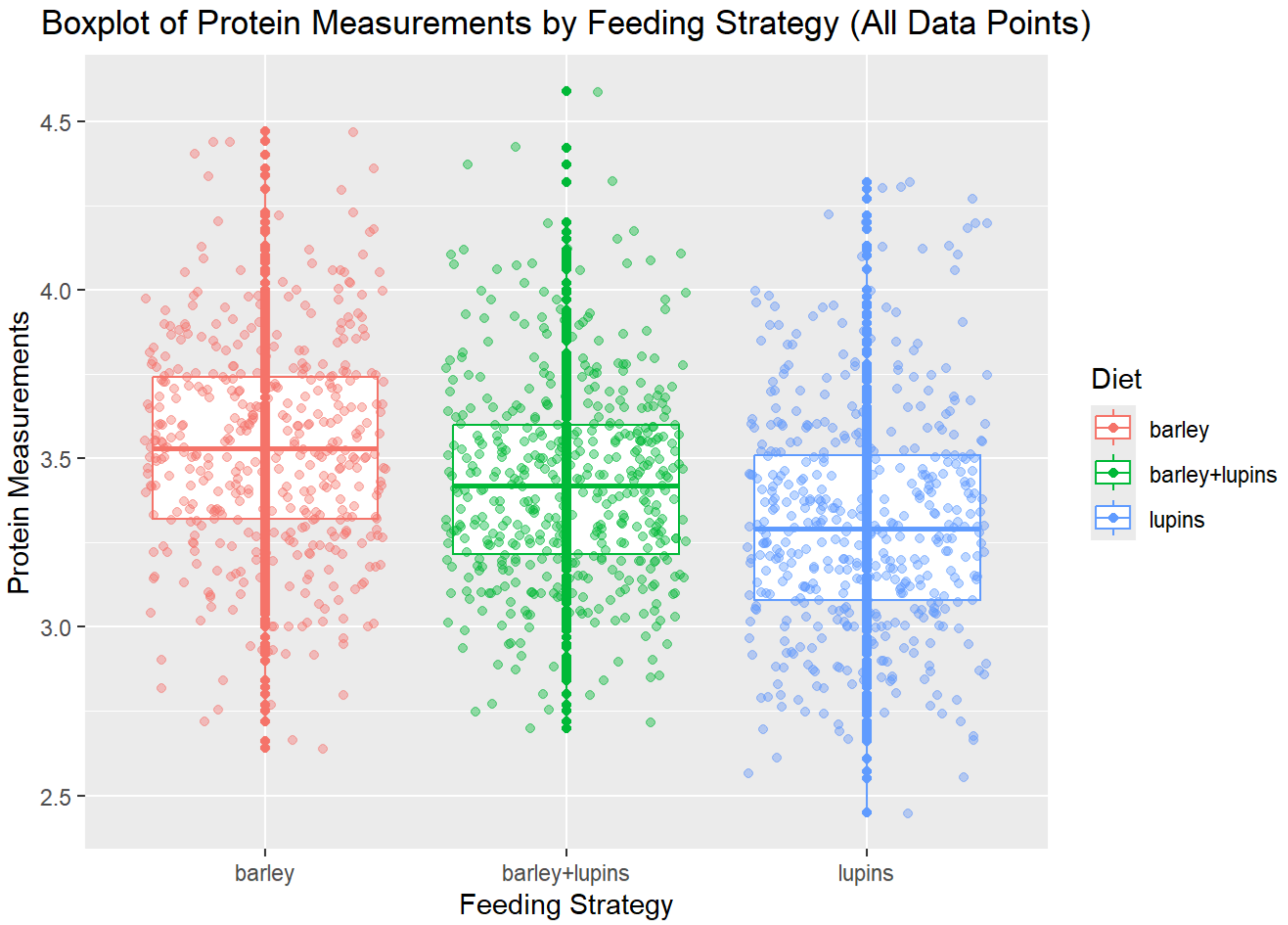

Figure 1. Boxplot of Protein Measurements by Feeding Strategy (All Data Points)

Part b: Boxplot of Protein Measurements by Feeding Strategy (First Week Data)

In the feeding strategy versus protein measurments graph **(Figure 2)**, if cows have **both barley and lupins type of diet**, **maximum mean of protein measurments** and **maximum protein measurment value** have observed **for first week**.

In these part have not used geom_jitter because have visualized only first week information.

```
boxplot.w1.filter = milk.data %>%
  filter(Time == 1) %>% #filtering the first week
  ggplot(aes(x = Diet,
             y = protein,
             color = Diet))+ #color scale by Diet type
  geom_boxplot()+ #boxplot have added
  geom_point()+
  labs(title = "Boxplot of Protein Measurements by Feeding Strategy (First Week Data)",
       x = "Feeding Strategy",
       y = "Protein Measurements")+
  centered.plot.title #title have set to center
boxplot.w1.filter
```

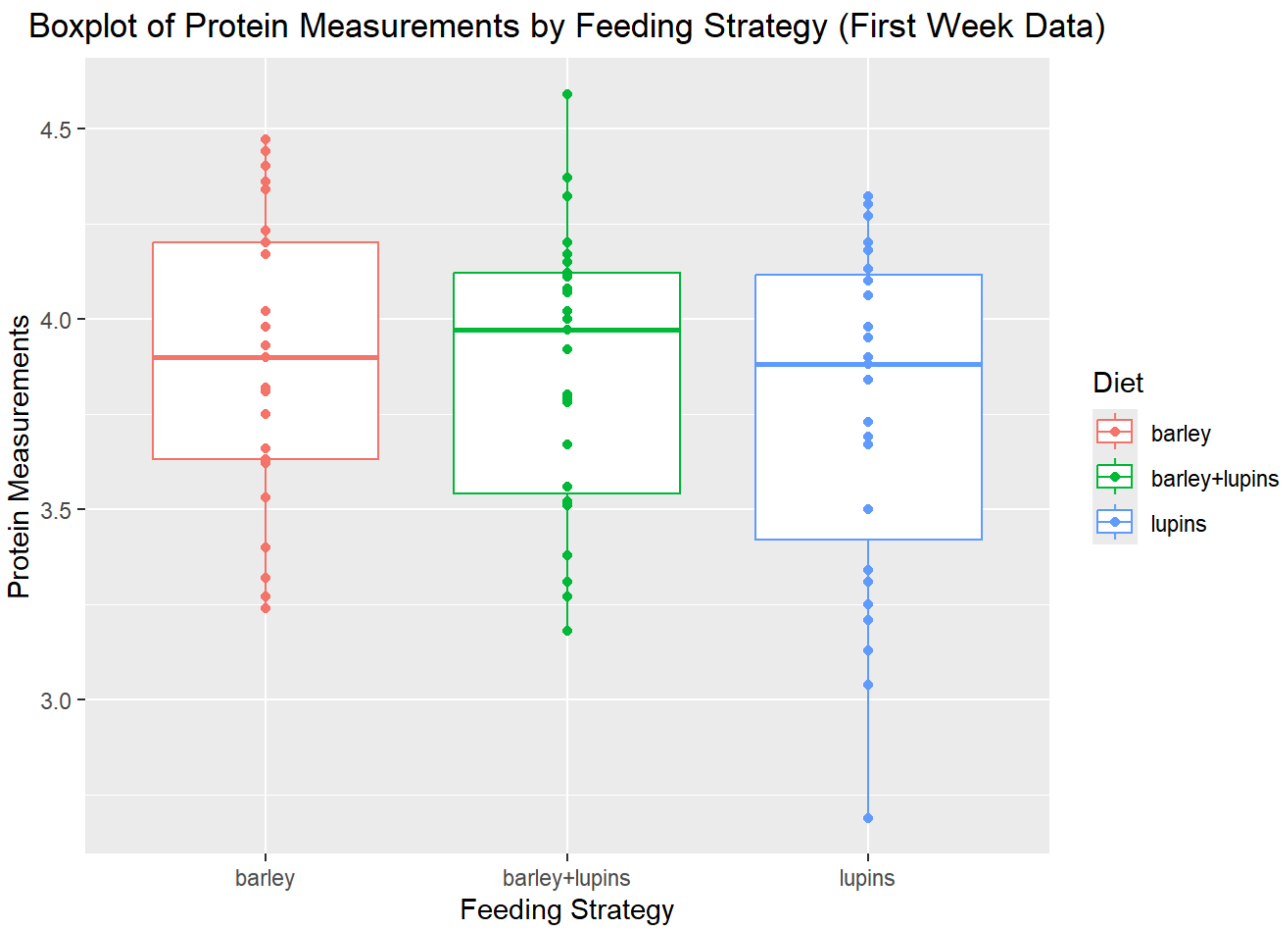

Figure 2. Boxplot of Protein Measurements by Feeding Strategy (First Week Data)

Part c: Protein Content Changes Over Time for Selected Cows

In the protein content changes over time graph for relevant cows **(Figure 3)**, for each cow different protein measurements level have observed. In all 18 weeks, in **B01 cow** have observed **maximum protein measurement** in last of the 18 weeks; however **first week**, in **L02 cow** have observed **maximum protein measurment**.

```
select.cow.line.plot = milk.data %>%
  filter(Cow %in% c("B01", "B02", "BL01", "BL02", "L01", "L02") & Time <= 18) %>% #relevant c
ows and time scale have defined
  ggplot(aes(x = Time,
             y = protein,
             color = Cow))+ #different color for each cow
  geom_line()+ #lines have added
  geom_point()+ #points have added
  labs(title = "Protein Content Changes Over Time for Selected Cows",
       x = "Week",
       y = "Protein Content")+
  centered.plot.title #title have set to center
select.cow.line.plot
```

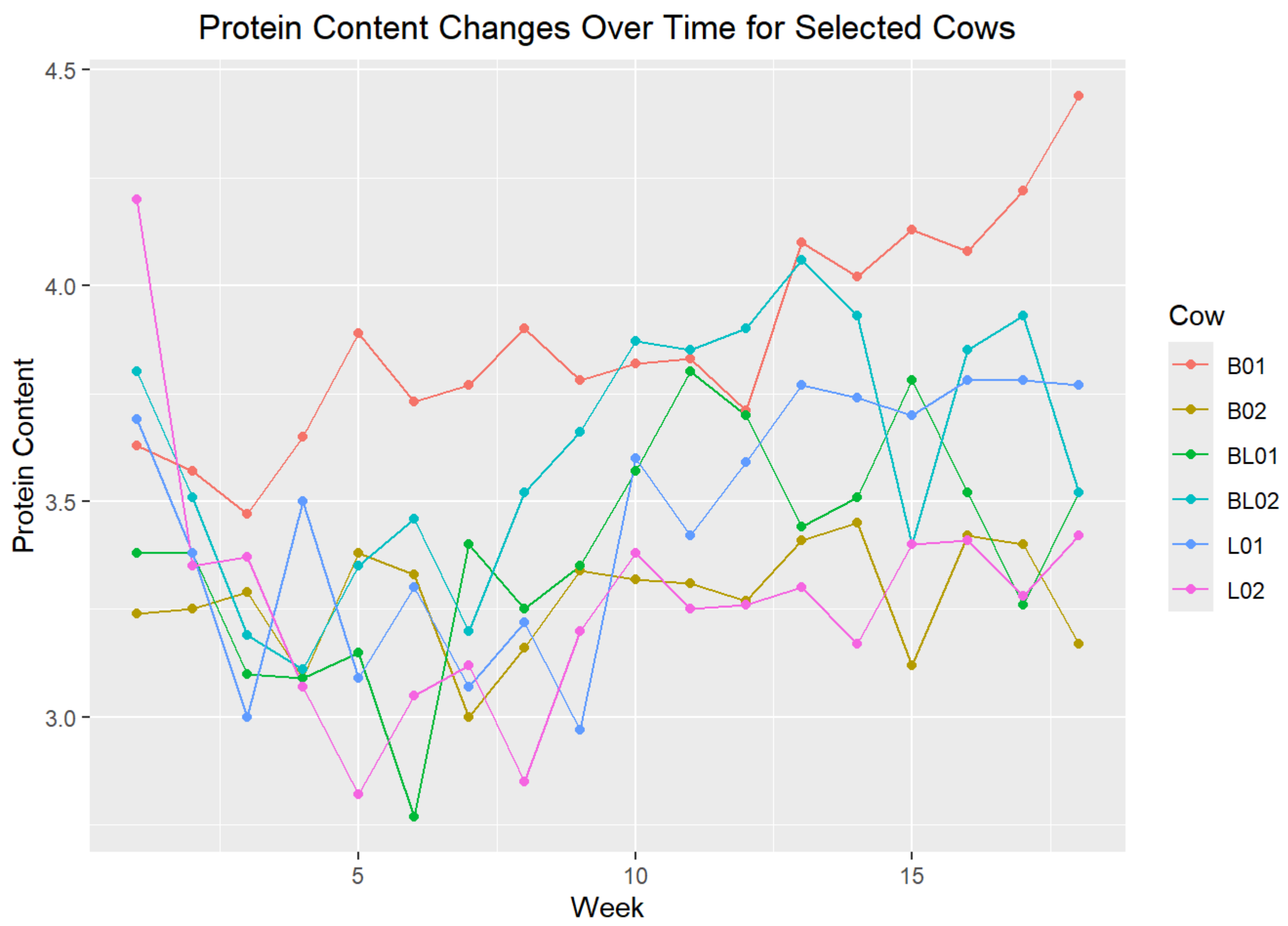

Figure 3. Protein Content Changes Over Time for Selected Cows

Part d: Using select.cow.line.plot, Creating 3 Different Subplots of Diet Strategies

the subplots **(Figure 4)** have splitted by **diet strategy**. **First graph** have consist of **barley** type of diet for **B01** and **B02**, **second graph** have consist of **both barley and lupins** type of diet for **BL01** and **BL02**, and **third graph** have consist of **lupins** type of diet for **L01** and **L02**.

```
subplot.feeding.cow = select.cow.line.plot+
  facet_wrap(~Diet)
subplot.feeding.cow
```

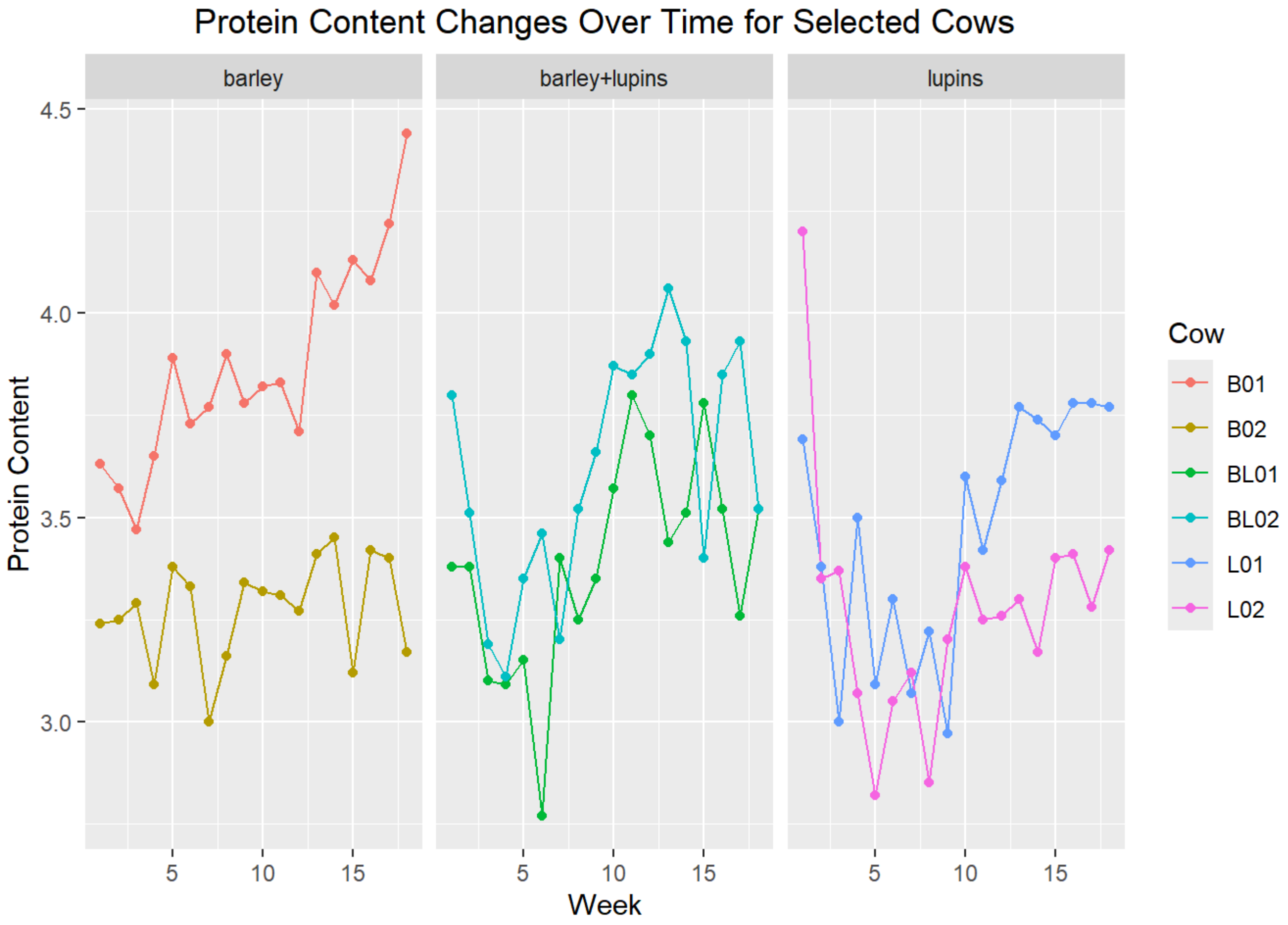

Figure 4. Creating 3 Different Subplots of Diet Strategies

Part e: Week Number vs. Mean Protein Content

In the week number versus mean protein content graph **(Figure 5)**, also have included number of cows for relevant week. the coloring was made according to the change in the number of weeks.

As displayed the graph, maximum protein measurement have observed in first week, while minimum protein measurement have observed in end of the week. the situation can be associated with number of cows, because number of cows have decreased week by week.

```
avg.protein.vs.week.scatter.plot = milk.data %>%
  group_by(Time) %>% #group by the time (week)
  summarise(n_cows = n(), #for each week, number of cows
            average.protein.content = mean(protein)) %>% #for each week, average of the protein measurments
  ggplot(aes(x = Time,
             y = average.protein.content,
             label = n_cows, #number of cows labels have defined
             color = Time))+ #color scales by the time
  geom_point()+ #points have added
  geom_text(hjust = -0.2, nudge_x = 0.2)+ #in relevant week, number of cows

  #The code adjusts text elements on the plot using geom_text().
  #hjust = -0.2 aligns the text horizontally to the left, and
  #nudge_x = 0.2 shifts the text 0.2 units to the right along the x-axis.

  labs(title = "Week Number vs. Mean Protein Content",
       x = "Week Number",
       y = "Mean Protein Content")+
  centered.plot.title #title have set to center
avg.protein.vs.week.scatter.plot
```

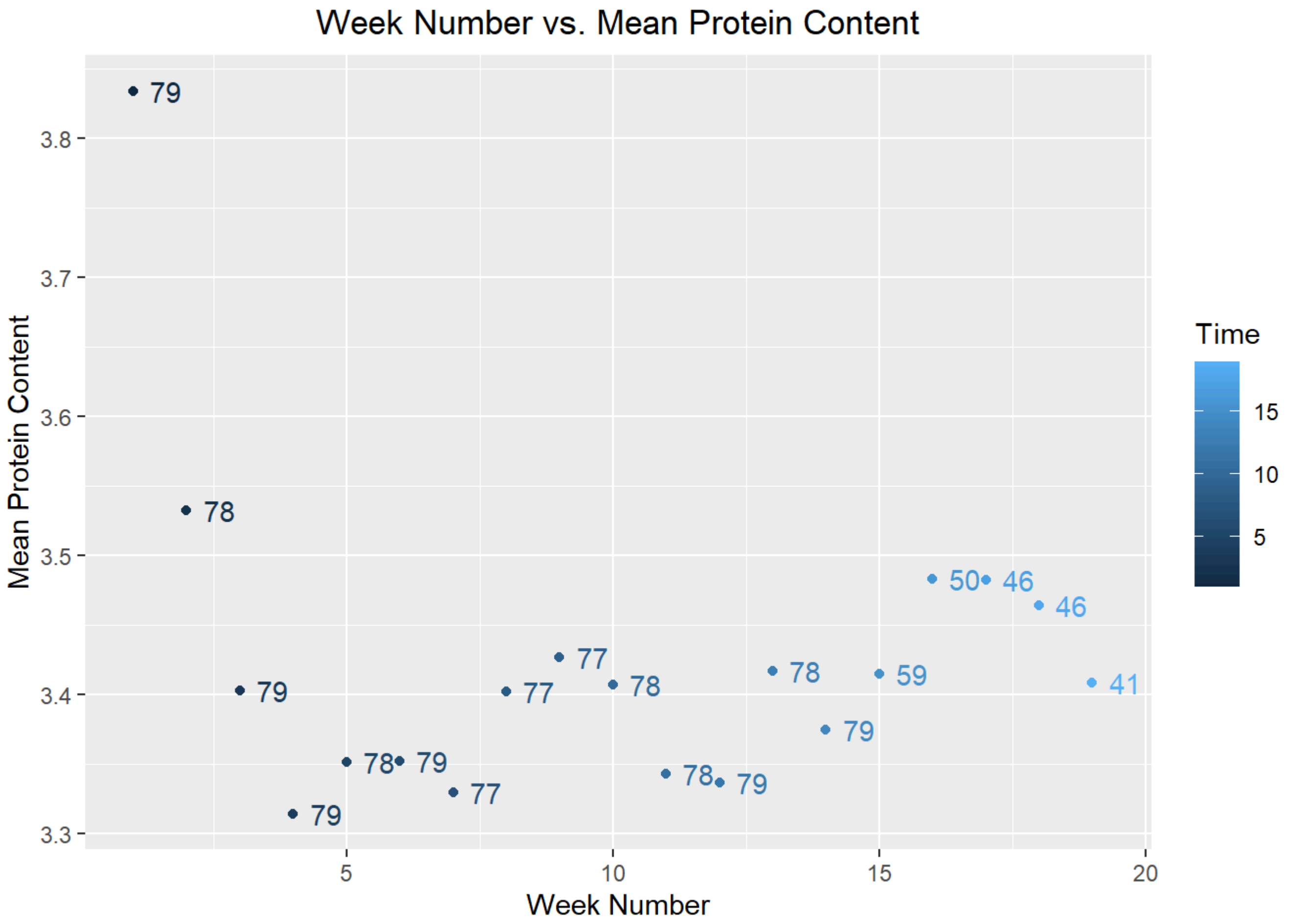

Figure 5. Week Number vs. Mean Protein Content

Part g: By the Week and the Type of Diet Week Number vs. Mean Protein Content

In the by the week and the type of diet week number versus mean protein content graph **(Figure 6)**, **maximum protein measurment** in the **barley** diet type have observed. In addition to, graph have included number of cows . As Figure 6, to prefer the **barley diet** is **suitable for the maximum performance** and **maximum produce the protein content**.

```
time.diet.filter.scatter.plot = milk.data %>%
  group_by(Time, Diet) %>% #grouped by time and diet features
  summarise(n.cow.time.diet = n(), #number of cows by the time and diet
            avg.protein.content.time.diet = mean(protein)) %>% #mean of the protein measurmen
t by the time and diet feature
  ggplot(aes(x = Time,
             y = avg.protein.content.time.diet,
             label = n.cow.time.diet))+ #label have set for the number of cows by diet and we
ek
  geom_point(aes(color = Diet))+ #color scaled by the type of diet
  geom_text(hjust = -0.2, nudge_x = 0.2)+ #in relevant week and diet type number of cows
  labs(title = "Week Number vs. Mean Protein Content",
       x = "Week Number",
       y = "Mean Protein Content")+
  centered.plot.title #title have set to center
```

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

```
time.diet.filter.scatter.plot
```

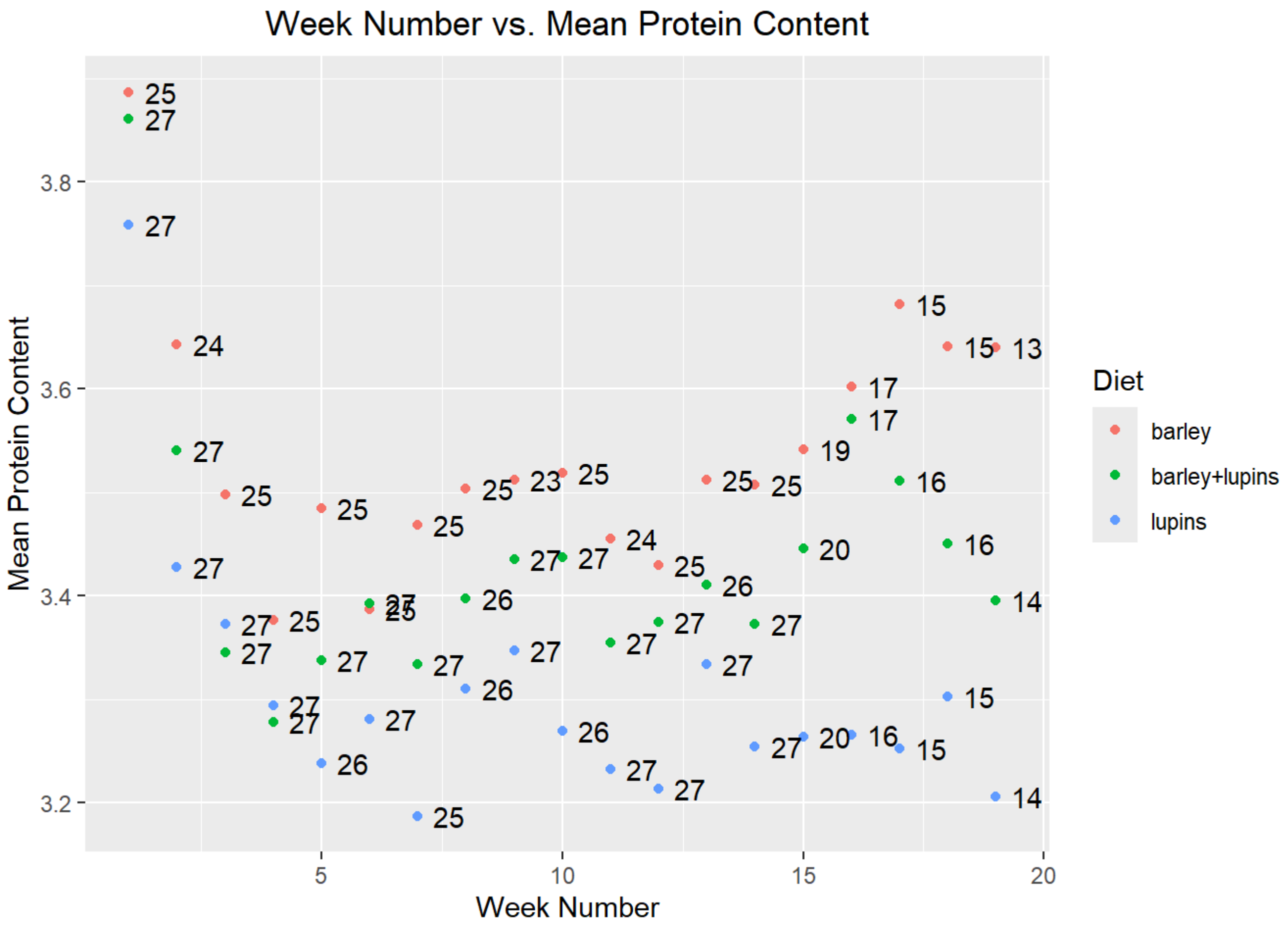

Figure 6. By the Week and the Type of Diet Week Number vs. Mean Protein Content

Part h: Create the Wide Format Data from Tidy Format: Using pivot_wider() Function

Firstly, the reason the data is in tidy format is that there is a row for each observation unit (data point) and each variable (protein content, time(week), cow barcode, feeding strategy (diet)) is in a column. The tidy structure have provided to users, easier preprocessing, analysis, and visualization.

This data created with pivot_wider is more complex than data in tidy format, and it is more difficult to work with this data.

```
new.milk.data <- milk.data %>% #with pivot_wider converted to wide format
  pivot_wider(names_from = Time, #new column information have assigned
              values_from = protein) %>% #for wide format, new values have assigned as the pr
otein measurment info
  as.data.frame() #have converted from tibble to the dataframe
head(new.milk.data)
```

```
##   Cow   Diet    1    2    3    4    5    6    7    8    9   10   11   12   13
## 1 B01 barley 3.63 3.57 3.47 3.65 3.89 3.73 3.77 3.90 3.78 3.82 3.83 3.71 4.10
## 2 B02 barley 3.24 3.25 3.29 3.09 3.38 3.33 3.00 3.16 3.34 3.32 3.31 3.27 3.41
## 3 B03 barley 3.98 3.60 3.43 3.30 3.29 3.25 2.93 3.20 3.27 3.22 2.93 2.92 2.82
## 4 B04 barley 3.66 3.50 3.05 2.90 2.72 3.11 3.05 2.80 3.20 3.18 3.14 3.18 3.24
## 5 B05 barley 4.34 3.76 3.68 3.51 3.45 3.53 3.60 3.77 3.90 3.87 3.61 3.85 3.94
## 6 B06 barley 4.36 3.71 3.42 3.95 4.06 3.73 3.92 3.99 3.70 3.88 3.71 3.62 3.74
##     14   15   16   17   18   19
## 1 4.02 4.13 4.08 4.22 4.44 4.30
## 2 3.45 3.12 3.42 3.40 3.17 3.00
## 3 2.64   NA   NA   NA   NA   NA
## 4 3.37 3.30 3.40 3.35 3.28   NA
## 5 3.87 3.60 3.06 3.47 3.50 3.42
## 6 3.42   NA   NA   NA   NA   NA
```