Socio-Informatics 348

Revision Session

Dr Lisa Martin

Department of Information Science
Stellenbosch University

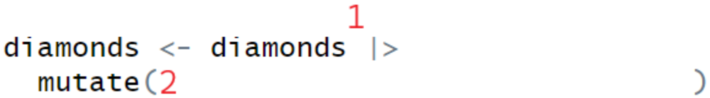
Example Question 1

Using the diamonds dataset from the ggplot2 package, create a new variable that represents the diamond's price divided by carat. Display the first few rows of the updated dataset.

4 marks

Example Question 1

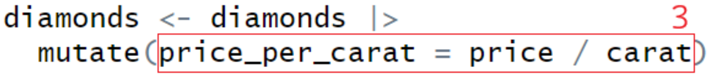
1. ‘Using the diamonds dataset from the ggplot2 package’
2. ‘create a new variable’



```
                      1
diamonds <- diamonds |>
  mutate(2                            )
```

Example Question 1

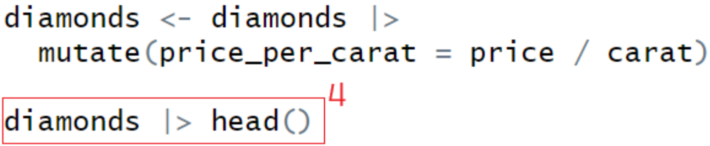
3. ‘new variable that represents price divided by carat’



```
diamonds <- diamonds |>
  mutate(price_per_carat = price / carat)
```

Example Question 1

4. Display the first few rows of the updated dataset



```
diamonds <- diamonds |>
  mutate(price_per_carat = price / carat)

diamonds |> head()
```

4

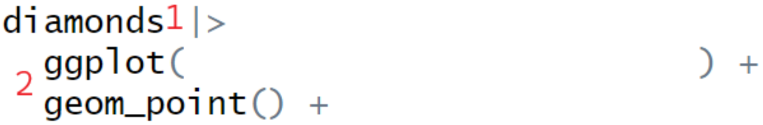
Example Question 2

Using the diamonds dataset from the ggplot2 package, create a scatter plot showing the relationship between price (y-axis) and carat (x-axis). Add a linear regression line to the plot, but remove the shaded confidence interval (standard errors) around the line.

5 marks

Example Question 2

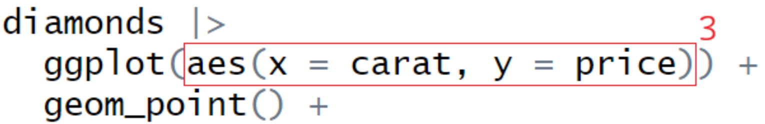
1. ‘Using the diamonds dataset from the ggplot2 package’
2. ‘create a scatterplot’



```
diamonds |>
  ggplot(                                   ) +
  geom_point() +
```

Example Question 2

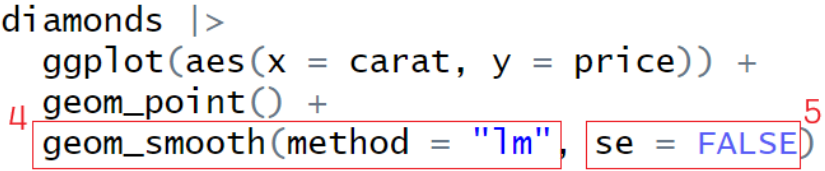
3. ‘...showing the relationship between price (y-axis) and carat (x-axis)’



```
diamonds |>
  ggplot(aes(x = carat, y = price)) +
  geom_point() +
```

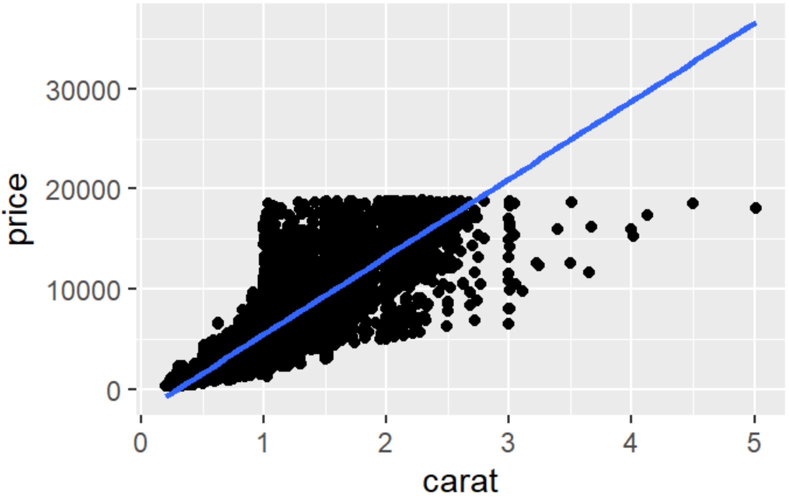
Example Question 2

4. ‘Add a linear regression line to the plot’
5. ‘…, but remove the shaded confidence interval (standard errors) around the line.’



```
diamonds |>
  ggplot(aes(x = carat, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

Example Question 2



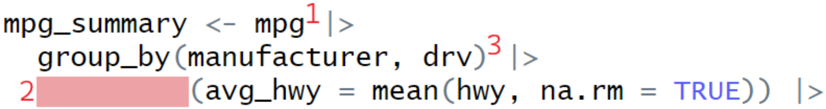
Example Question 3

Using the mpg dataset from the ggplot2 package, transform the dataset to show the average highway miles per gallon (hwy) for each manufacturer across different drive types (drv). After transforming the dataset, it should display a row for each manufacturer, and a column for each drive type. The cells should contain the average highway mpg for the corresponding manufacturer–drive combination (if there is one).

6 marks

Example Question 3

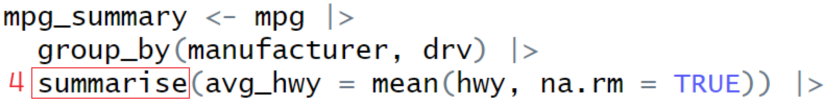
1. ‘Using the mpg dataset from the ggplot2 package’
2. ‘...transform the dataset to show the average highway miles per gallon (hwy)’
3. ‘...for each manufacturer across different drive types (drv)’



```
mpg_summary <- mpg1 |>
  group_by(manufacturer, drv)3 |>
 2          (avg_hwy = mean(hwy, na.rm = TRUE)) |>
```

Example Query 3

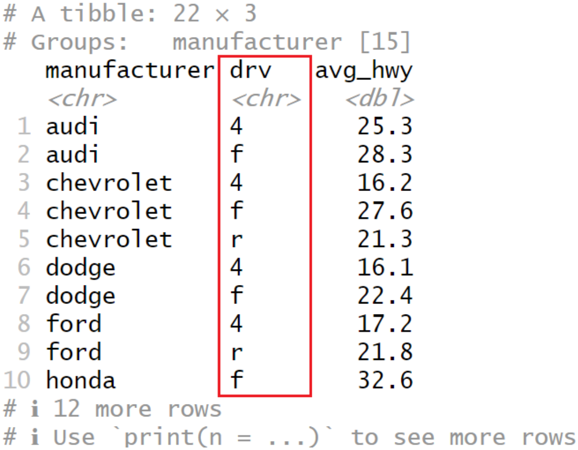
4. '...it should display a row for each manufacturer...'



```
mpg_summary <- mpg |>
  group_by(manufacturer, drv) |>
  summarise(avg_hwy = mean(hwy, na.rm = TRUE)) |>
```

Example Question 3

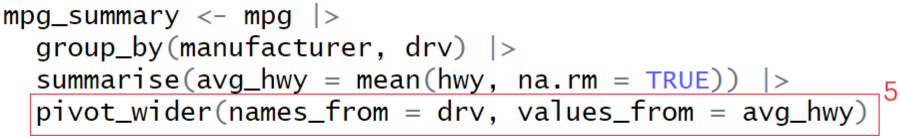
5. ‘..., and a column for each drive type.’



```
# A tibble: 22 × 3
# Groups:   manufacturer [15]
   manufacturer drv   avg_hwy
   <chr>        <chr>   <dbl>
 1 audi         4        25.3
 2 audi         f        28.3
 3 chevrolet    4        16.2
 4 chevrolet    f        27.6
 5 chevrolet    r        21.3
 6 dodge        4        16.1
 7 dodge        f        22.4
 8 ford         4        17.2
 9 ford         r        21.8
10 honda        f        32.6
# ℹ 12 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Example Question 3

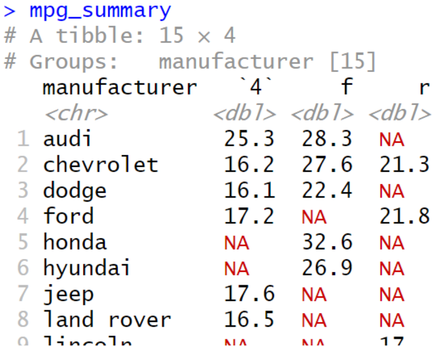
5. '..., and a column for each drive type.'



```
mpg_summary <- mpg |>
  group_by(manufacturer, drv) |>
  summarise(avg_hwy = mean(hwy, na.rm = TRUE)) |>
  pivot_wider(names_from = drv, values_from = avg_hwy)
```

Example Question 3

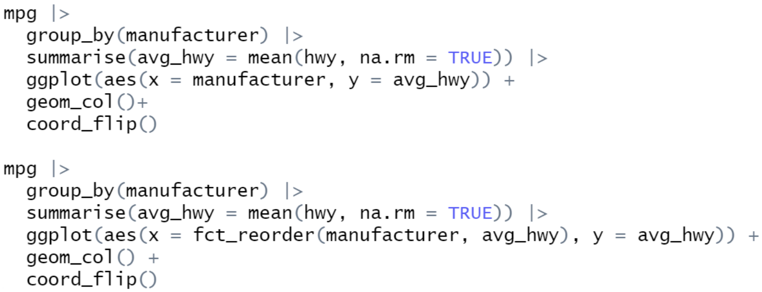
6. ‘The cells should contain the average highway mpg for the corresponding manufacturer–drive combination (if there is one).’



```
> mpg_summary
# A tibble: 15 × 4
# Groups:   manufacturer [15]
   manufacturer   `4`     f     r
   <chr>        <dbl> <dbl> <dbl>
 1 audi          25.3  28.3  NA  
 2 chevrolet     16.2  27.6  21.3
 3 dodge         16.1  22.4  NA  
 4 ford          17.2  NA    21.8
 5 honda         NA    32.6  NA  
 6 hyundai       NA    26.9  NA  
 7 jeep          17.6  NA    NA  
 8 land rover    16.5  NA    NA  
 9 lincoln       NA    NA    17  
```

Example Request

fct_reorder(): Reorder factor levels by values of another variable.

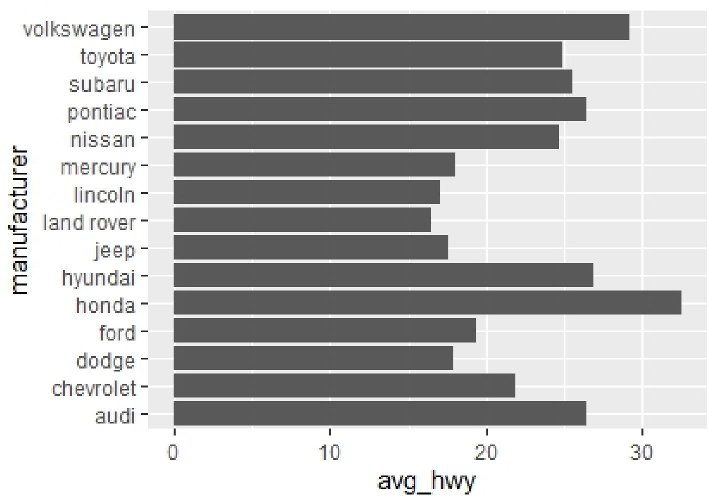


```
mpg |>
  group_by(manufacturer) |>
  summarise(avg_hwy = mean(hwy, na.rm = TRUE)) |>
  ggplot(aes(x = manufacturer, y = avg_hwy)) +
  geom_col()+
  coord_flip()

mpg |>
  group_by(manufacturer) |>
  summarise(avg_hwy = mean(hwy, na.rm = TRUE)) |>
  ggplot(aes(x = fct_reorder(manufacturer, avg_hwy), y = avg_hwy)) +
  geom_col() +
  coord_flip()
```

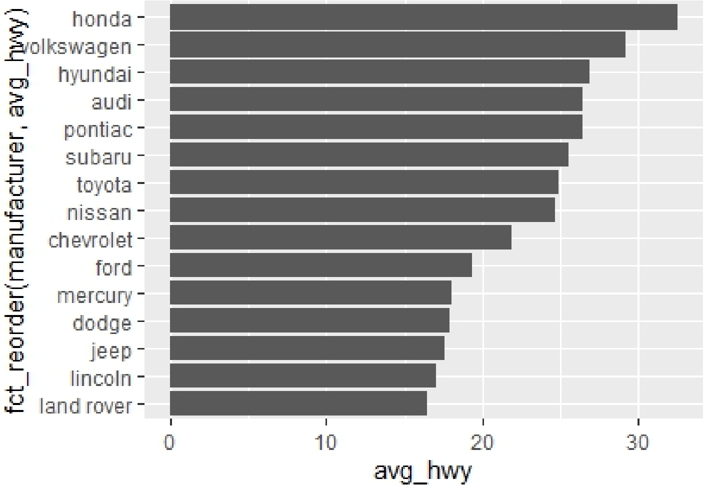
Example Request

fct_reorder(): Reorder factor levels by values of another variable.



Example Request

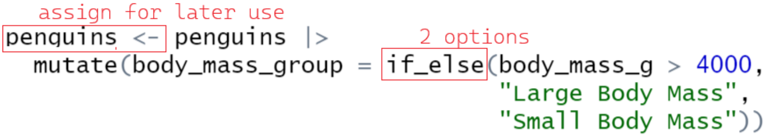
fct_reorder(): Reorder factor levels by values of another variable.



Example Question 4

First, create a new variable that classifies penguins into two groups:

- "Large Body Mass" if body_mass_g is greater than 4000,
- "Small Body Mass" otherwise.



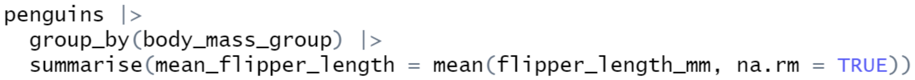
assign for later use

2 options

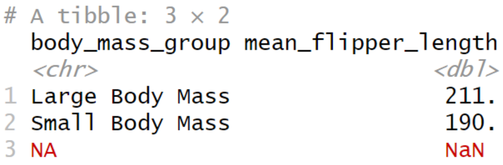
```
penguins <- penguins |>
  mutate(body_mass_group = if_else(body_mass_g > 4000,
                                   "Large Body Mass",
                                   "Small Body Mass"))
```

Example Question 4

Using this classification, calculate the mean flipper length (mm) for each body mass group to compare whether larger-bodied penguins tend to have longer flippers.



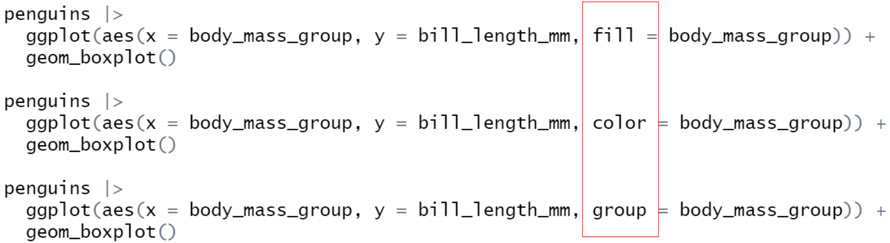
```
penguins |>
  group_by(body_mass_group) |>
  summarise(mean_flipper_length = mean(flipper_length_mm, na.rm = TRUE))
```



```
# A tibble: 3 × 2
  body_mass_group mean_flipper_length
  <chr>                         <dbl>
1 Large Body Mass                211.
2 Small Body Mass                190.
3 NA                             NaN
```

Example Question 4

Investigate how bill length (mm) differs across these body mass groups by creating an appropriate plot (choose a plot other than a density plot).



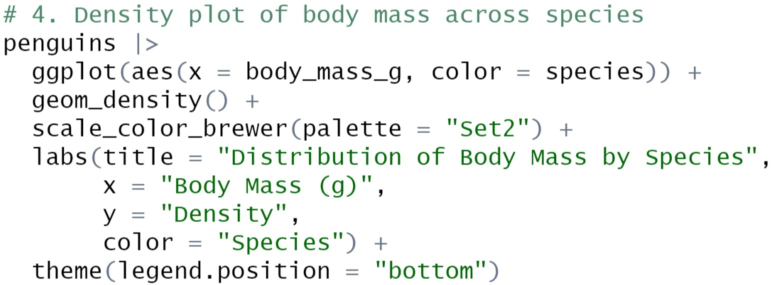
```
penguins |>
  ggplot(aes(x = body_mass_group, y = bill_length_mm, fill = body_mass_group)) +
  geom_boxplot()

penguins |>
  ggplot(aes(x = body_mass_group, y = bill_length_mm, color = body_mass_group)) +
  geom_boxplot()

penguins |>
  ggplot(aes(x = body_mass_group, y = bill_length_mm, group = body_mass_group)) +
  geom_boxplot()
```

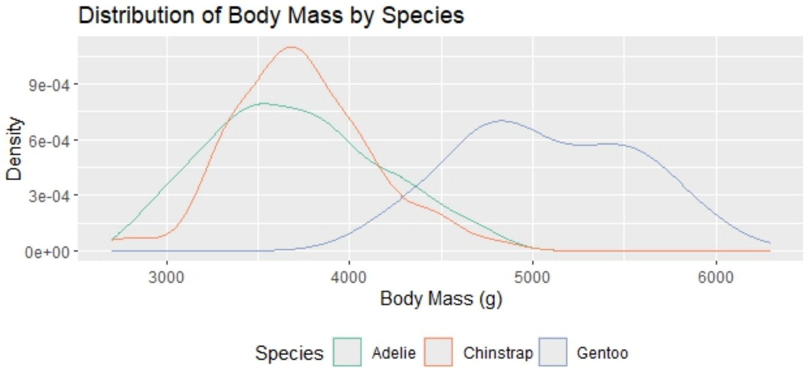
Example Question 4

Create a density plot where each species is shown in a different colour. Apply a ColorBrewer palette for these colours, position the legend at the bottom of the plot, and ensure the legend title is properly formatted.



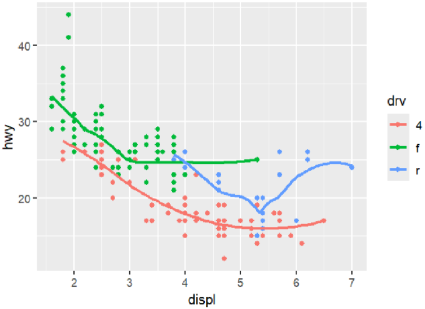
```
# 4. Density plot of body mass across species
penguins |>
  ggplot(aes(x = body_mass_g, color = species)) +
  geom_density() +
  scale_color_brewer(palette = "Set2") +
  labs(title = "Distribution of Body Mass by Species",
       x = "Body Mass (g)",
       y = "Density",
       color = "Species") +
  theme(legend.position = "bottom")
```

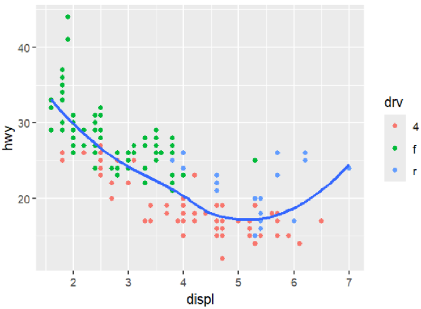
Example Question 4



Example Request 2

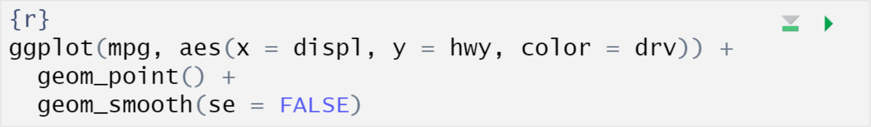
- Global vs Local mappings (example from prac 3)



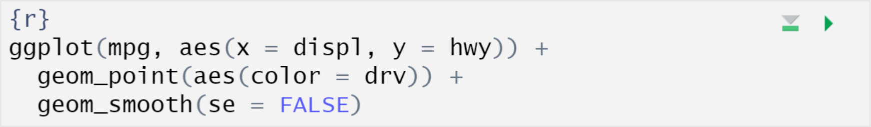


Example Request 2

- Global vs Local mappings (example from prac 3)



```
{r}
ggplot(mpg, aes(x = displ, y = hwy, color = drv)) +
  geom_point() +
  geom_smooth(se = FALSE)
```



```
{r}
ggplot(mpg, aes(x = displ, y = hwy)) +
  geom_point(aes(color = drv)) +
  geom_smooth(se = FALSE)
```