|  |
| --- |
| **The Pipe Operator**  penguins %>%  select(species, island, body\_mass\_g, sex, year) %>%  filter(island == "Torgersen",  body\_mass\_g > 1000) %>%  summarize(mean\_mass = mean(body\_mass\_g),  median\_mass = median(body\_mass\_g),  observations = n()) %>%  arrange(desc(mean\_mass))  Sequences of data wrangling steps together, where each step is connected with “and then”. For example, in the above code, we take the penguins data **and then** select specific columns, **and then** filter out specific observations, **and then** summarize the body mass variable, **and then** arrange the body masses in descending order. |
| **Obtaining Statistics for Different Groups**  <NAME OF DATASET> %>%  group\_by(<NAME OF CATEGORICAL VARIABLE>) %>%  summarize(<NAME OF STAT> = <STAT FUNCTION>(<NAME OF VARIABLE>)  )  For example, to calculate the mean dep\_delay for each airline carrier (from the nycflights dataset) we have:  nycflights %>%  group\_by(carrier) %>%  summarize(mean\_dep\_delay = mean(dep\_delay)  ) |
| **Finding the (Distinct) Levels of a Categorical Variable**  distinct(<NAME OF DATASET>, <NAME OF CATEGORICAL VARIABLE>) |
| **Filtering to Include Specific Values of a Categorical Variable**  filter(nycflights,   <NAME OF CATEGORICAL VARIABLE> %in% c(“<VALUE1>”,“<VALUE2>”,“<VALUE3>”)  )  The %in% is the “inclusion” operator -- it tells R you are interested in keeping all the values included in the provided list.  The c() function creates a list of levels you are interested in keeping in the dataset. |

|  |
| --- |
| **Side-by-Side Boxplots**  ggplot(data = <NAME OF DATASET>,  mapping = aes(x = <NAME OF VARIABLE>)) +  geom\_boxplot() +  labs(x = "<TITLE FOR THE X-AXIS>")  ***Note:*** This stacks the boxplots horizontally. If you want for your boxplots to be stacked vertically, you use **y =** instead of **x =** . Keep in mind you will need to change the location of you axis label, too! |
| **Faceted Histograms**  ggplot(data = <NAME OF DATASET>,  mapping = aes(x = <NAME OF NUMERICAL VARIABLE>)) +  geom\_histogram(binwidth = <WIDTH OF BINS>) +  facet\_wrap(~ <NAME OF CATEGORICAL VARIABLE>)  labs(x = "<TITLE FOR THE X-AXIS>") |
| **Ridge Plots**  ggplot(data = <NAME OF DATASET>,  mapping = aes(x = <NAME OF NUMERICAL VARIABLE>,  y = <NAME OF CATEGORICAL VARIABLE>)) +  geom\_density\_ridges()+  labs(x = "<TITLE FOR THE X-AXIS>",  y = “<TITLE FOR THE Y-AXIS>”) |
| **Adding Colors to a Plot (Demonstrated with a Ridge Plot)**  For any plot, you can add a color by mapping it to a variable (categorical or numerical)!  ggplot(data = <NAME OF DATASET>,  mapping = aes(x = <NAME OF NUMERICAL VARIABLE>,  y = <NAME OF CATEGORICAL VARIABLE>,  fill = <NAME OF CATEGORICAL VARIABLE>)) +  geom\_density\_ridges(alpha = 0.5)+  labs(x = "<TITLE FOR THE X-AXIS>",  y = “<TITLE FOR THE Y-AXIS>”,  fill = “<TITLE FOR LEGEND>“)  ***Note:*** To make the colors more transparent, you can use the alpha argument. Numbers closer to 0 are more transparent and numbers closer to 1 are more opaque. |

|  |
| --- |
| **Adding Facets to a Plot (Demonstrated with a Histogram)**  For any plot, you can add a facet by simply adding (with a +) the following code:  ggplot(data = <NAME OF DATASET>,  mapping = aes(x = <NAME OF NUMERICAL VARIABLE>)) +  geom\_histogram(binwidth = <WIDTH OF BINS>) +  facet\_wrap(~ <NAME OF CATEGORICAL VARIABLE>)  labs(x = "<TITLE FOR THE X-AXIS>") |

|  |
| --- |
| **Adding Labels to a Plot**  For any plot, you can add axis labels by simply adding (with a +) the following code:  labs(x = “<NAME OF X-AXIS VARIABLE>”,  y = “<NAME OF Y-AXIS VARIABLE>”,  fill = “<NAME OF FILL VARIABLE>”)  For example, the code could look like this:  ggplot(data = <NAME OF DATASET>,  mapping = aes(x = <NAME OF NUMERICAL VARIABLE>,  y = <NAME OF CATEGORICAL VARIABLE>,  fill = <NAME OF CATEGORICAL VARIABLE>)) +  geom\_density\_ridges()+  labs(x = "<TITLE FOR THE X-AXIS>",  y = “<TITLE FOR THE Y-AXIS>”,  fill = “<TITLE FOR THE LEGEND>”) |