

Mutate, Merge, and Bind Lab

Ben Goldstone

10/4/2023

Part 1: Mutate Practice

##Load in Baseball Dataset

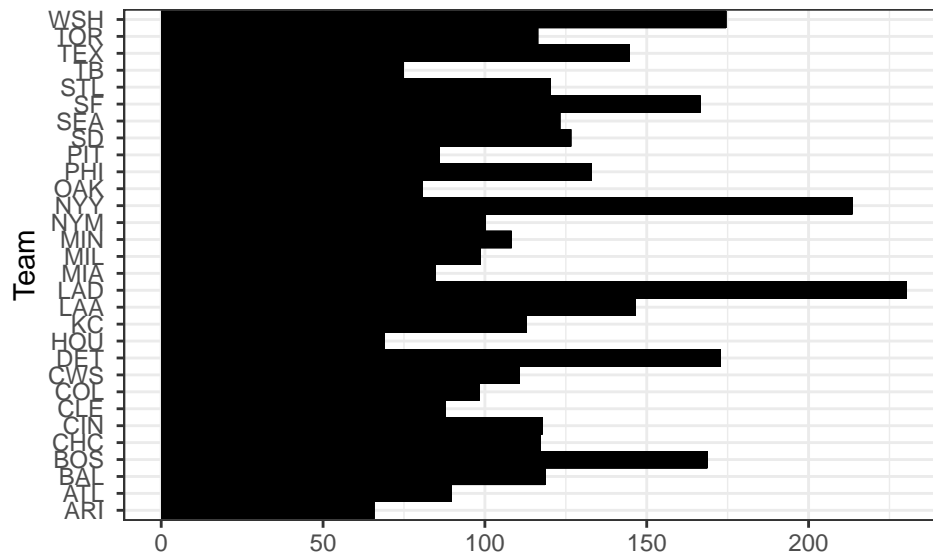
```
data("BaseballSalaries2015")
head(BaseballSalaries2015)
```

```
##           Name Salary Team Position
## 1 Clayton Kershaw    31  LAD      SP
## 2 Justin Verlander    28  DET      SP
## 3 Zack Greinke        27  LAD      SP
## 4 Josh Hamilton       25  LAA      LF
## 5 Ryan Howard          25  PHI      1B
## 6 Cliff Lee           25  PHI      SP
```

```
Baseball12 = BaseballSalaries2015 %>%
  group_by(Team) %>%
  mutate(TotalSalary = sum(Salary), NumPlayers = n())
```

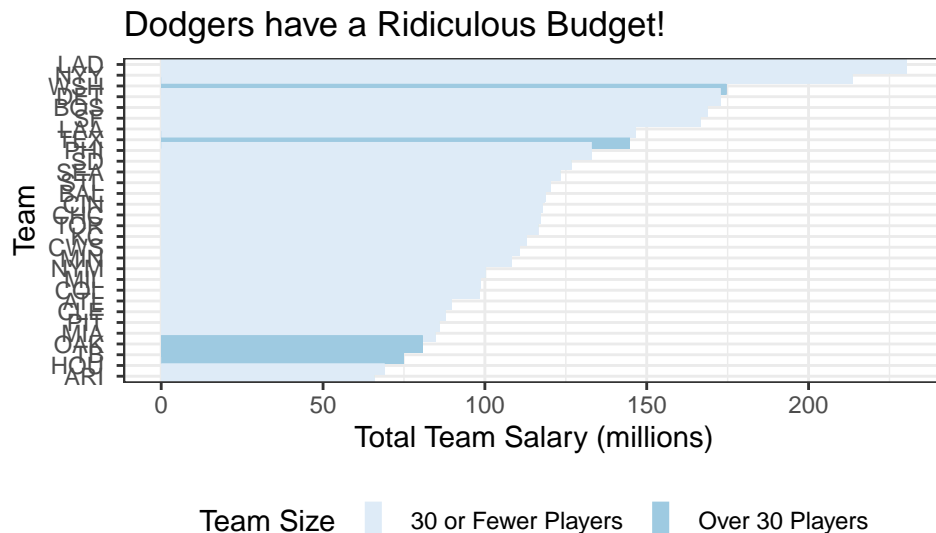
Graph and Graph Refinements

```
gf_linerange(Team~ 0 + TotalSalary, data = Baseball12, size = 3)
```



```
gf_linerange(fct_reorder(Team, TotalSalary)~ 0 + TotalSalary, data=Baseball12, size = 3, color = ~NumPlay)
gf_refine(scale_color_brewer(
```

```
palette = "Blues",
labels = c("30 or Fewer Players", "Over 30 Players")))) %>%
gf_theme(legend.position="bottom") %>%
gf_labs(y = "Team", x = "Total Team Salary (millions)", title = "Dodgers have a Ridiculous Budget!", co
```



Part 2

States Dataset

```
States1 = read_csv("~/CSVs/States1.csv")
```

```
## Rows: 50 Columns: 6
```

```
## -- Column specification -----
```

```
## Delimiter: ","
```

```
## chr (4): State, PartyAff2016, TimeZones, Climate
```

```
## dbl (2): MinimumWage, Union
```

```
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
States2 = read_csv("~/CSVs/States2.csv")
```

```
## Rows: 50 Columns: 6
```

```
## -- Column specification -----
```

```
## Delimiter: ","
```

```
## chr (2): State, GovernorParty
```

```
## dbl (2): MinimumWage, NumberofCities
```

```
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(States1)
```

```
## # A tibble: 6 x 6
```

```
##   State      MinimumWage PartyAff2016 TimeZones Climate Union
```

```
##   <chr>           <dbl> <chr>           <chr>      <chr>    <dbl>
## 1 Alabama         0    Rep           Cental    South    1819
## 2 Alaska        10.2 LeanRep      Alaska    North    1959
## 3 Arizona         12    Swing        Mountain South    1912
## 4 Arkansas        10    Rep           Central   South    1836
## 5 California      13    Dem           Pacific   Central  1850
## 6 Colorado        12    Swing        Mountain  Central  1876
```

```
head(States2)
```

```
## # A tibble: 6 x 6
##   State      MinimumWage Population BiggestCityPop NumberofCities GovernorParty
##   <chr>      <dbl>      <dbl>      <dbl>      <dbl> <chr>
## 1 Alabama      NA      4921532      1090435          5 Republican
## 2 Alaska      10.3      731158      291826          1 Republican
## 3 Arizona      12.2     7421401     1633017          5 Republican
## 4 Arkansas      11     3030522     198606          0 Republican
## 5 California    13     39368078    3990456         15 Democrat
## 6 Colorado     12.3     5807719     705576          3 Democrat
```

```
TruncatedStates1 = States1[c("State", "Union")]
```

```
AllStates = full_join(TruncatedStates1, States2)
```

```
## Joining, by = "State"
```

```
head(AllStates)
```

```
## # A tibble: 6 x 7
##   State Union MinimumWage Population BiggestCityPop NumberofCities GovernorParty
##   <chr> <dbl>      <dbl>      <dbl>      <dbl>      <dbl> <chr>
## 1 Alab~ 1819      NA      4921532      1090435          5 Republican
## 2 Alas~ 1959     10.3      731158      291826          1 Republican
## 3 Ariz~ 1912     12.2     7421401     1633017          5 Republican
## 4 Arka~ 1836      11     3030522     198606          0 Republican
## 5 Cali~ 1850      13     39368078    3990456         15 Democrat
## 6 Colo~ 1876     12.3     5807719     705576          3 Democrat
```

```
gf_linerange(fct_reorder(State, MinimumWage) ~ 0 + MinimumWage, data = na.omit(AllStates), size = 2, color =
  gf_labs(y = "State", x = "MinimumWage", title = "Washington, Massachusetts, and California Have the Highest Minimum Wage in the US"),
  gf_refine(scale_color_brewer(
    palette = 1,
    labels = c("Below 13", "Above 13"))) %>%
  gf_theme(plot.title = element_text(hjust = 0.5))
```

gton, Massachusetts, and California Have the Highest Minimum



Animals Dataset

```
Animals4 = read_csv("~/CSVs/Animals4.csv")
```

```
## Rows: 50 Columns: 6
```

```
## -- Column specification -----
```

```
## Delimiter: ","
```

```
## chr (3): Animal, TypeofAnimal, Diet
```

```
## dbl (3): Lifespan(years), GestationPeriod(days), Length(in)
```

```
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Animals5 = read_csv("~/CSVs/Animals5.csv")
```

```
## Rows: 99 Columns: 9
```

```
## -- Column specification -----
```

```
## Delimiter: ","
```

```
## chr (4): Animal, Class, Diet, Eggs
```

```
## dbl (5): Lifespan, Gestation, AvgLength, NumberBabies, NumberEggs
```

```
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(Animals4)
```

```
## # A tibble: 6 x 6
```

```
##   Animal   `Lifespan(years)` `GestationPeriod(~ TypeofAnimal Diet   `Length(in)`
##   <chr>         <dbl>         <dbl> <chr>         <chr>         <dbl>
## 1 Mole             4             40 mammal    carni~             6
## 2 Capybara         4            140 mammal    herbi~            51
## 3 Hedgehog         5             40 mammal    insec~            10
## 4 Baboon          30            180 mammal    omniv~            20
## 5 Horse           28            340 mammal    herbi~            96
## 6 Camel           40            410 mammal    carni~           120
```

```
head(Animals5)
```

```
## # A tibble: 6 x 9
##   Animal Lifespan Gestation AvgLength NumberBabies Class Diet NumberEggs Eggs
##   <chr>      <dbl>      <dbl>      <dbl>          <dbl> <chr> <chr>      <dbl> <chr>
## 1 Turtle      80        60       80.4            NA Rept~ Carn~      110 Yes
## 2 Yak         20       270      126             1 Mamm~ Herb~       NA No
## 3 Salama~     20       913       6.4            NA Amph~ Carn~     450 Yes
## 4 BlackH~     20       180      31.2             1 Mamm~ Herb~       NA No
## 5 Vicuna      24       335       60             1 Mamm~ Herb~       NA No
## 6 Cabyba~     10       150       48             5 Mamm~ Herb~       NA No
```

```
# Animal 4 modifications
```

```
colnames(Animals4)[c(2,3,6)] = c("Lifespan", "Gestation", "Length")
```

```
Animals4.keep = Animals4 %>% select(Animal, Lifespan, Gestation, Diet, Length)
```

```
colnames(Animals5)[5] = "Length"
```

```
Animals5.keep = Animals5 %>% select(Animal, Lifespan, Gestation, Diet, Length) %>% mutate(Diet = tolower(Diet))
```

```
Only.Animals4 = anti_join(Animals4.keep, Animals5.keep, by="Animal")
```

```
All.Animals = rbind(Only.Animals4, Animals5.keep)
```

```
All.Animals
```

```
## # A tibble: 139 x 5
##   Animal Lifespan Gestation Diet Length
##   <chr>      <dbl>      <dbl> <chr>      <dbl>
## 1 Mole      4        40 carnivore      6
## 2 Capybara  4       140 herbivore     51
## 3 Hedgehog  5        40 insectivore  10
## 4 Baboon   30       180 omnivore     20
## 5 Camel    40       410 carnivore    120
## 6 Hog      8       120 omnivore     60
## 7 Puma     12        91 carnivore     84
## 8 Parakeet 13        18 herbivore      7
## 9 Bison    15       283 herbivore   112.
## 10 Ox      20       283 herbivore     60
## # ... with 129 more rows
```

```
gf_boxplot(Length~Diet, data=All.Animals, fill=~Diet, outlier.shape = NA) %>%
  gf_refine(coord_cartesian(ylim = c(0, 50))) %>%
  gf_theme(legend.position = "none") %>%
  gf_labs(title = "Carnivores have the largest spread of lengths")
```

```
## Warning: Removed 18 rows containing non-finite values (stat_boxplot).
```

