BADM 372 Applied Analytics

BADM 372

2022-01-20

Contents

Chapter 1

About this course

This website serves as headquarters for **BADM 372 Applied Analytics**.

Content here will be updated with any changes made during the semester, so if at any point you are told there was a change in the schedule or an assignment, you can come here to get the updated version.

Also, this website has benefited greatly from lots of free, readily available resources posted on the web and we leverage these extensively. I would encourage you to review these resources in your analytics journey. Some that we specifically use with great frequency are these (**and I say a loud THANK YOU to the authors!**):

- R for Data Science
- An Introduction to Statistical Learning with Applications in R
- Data Science in a Box
- stackoverflow.com, for example

Chapter 2

Syllabus

Instructor: Tobin Turner

Office Hours: mutually convenient time arranged by email e-mail: jtturner@presby.edu

2.1 Course Objectives and Learning Outcomes

This course is designed to introduce to data science. Students will apply statistical knowledge and techniques to both business and non-business contexts.

At the end of this course students should be able to:

By the end of the course, you will be able to…

- gain insight from data
- gain insight from data, reproducibly
- gain insight from data, reproducibly, using modern programming tools and techniques
- gain insight from data, reproducibly and collaboratively, using modern programming tools and techniques
- gain insight from data, reproducibly (with literate programming and version control) and collaboratively, using modern programming tools and techniques
- communicate results effectively

This course will be focused on both understanding and applying key business analytical concepts. Although the text serves as a useful foundation for the concepts covered in the class, simple memorization of the material in the text will not be sufficient. Class participation, discussion, and application are critical.

2.2 Text and Resources

- This course website (primary resource)
- R for Data Science
- An Introduction to Statistical Learning with Applications in R
- Data Science in a Box
- stackoverflow.com, for example
- Other free, publicly available datasets and publications.

2.3 Performance Evaluation (Grading)

- Quizzes and Assignments - 40%
- Exam 1 - 20%
- Exam 2 - 20%
- Final Exam - 20%

2.3.1 Exams

Exams will cover assigned chapters in the textbook, other assigned readings, lectures, class exercises, class discussions, videos, and guest speakers. I will typically allocate time prior to each exam to clearly identify the body of knowledge each test will cover and to answer questions about the format and objectives of the exam.

2.3.2 Quizzes – DON'T MISS CLASS

- The average of all quizzes and assignments will comprise the Quizzes and Assignments - 40% portion of your final grade
- Quizzes and Assignments are designed to prepare you for your exams and to ensure you stay up with the course material
- **Missed Quizzes and Assignments cannot be made up later. Be present.**

Quizzes rule. **LISTEN. - Missed Quizzes and Assignments cannot be made up later. Be present.**

2.3.3 Final Average

- Final Average Grade
  - 90-100 A

- 88-89 B+
- 82-87 B+
- 80-81 B-
- 78-79 C+
- 72-77 C+
- 70-71 C-
- 60-69 D
- 59 and below F

2.4 Class Participation:

I will frequently give readings or assignments for you to complete prior to the next class meeting. I expect you to fully engage the material: answer questions, pose questions, provide insightful observations. Keep in mind that quality is an important component in "participation." Periodic cold calls will take place. I will also put students in the "hot seat" on occasion. In these class sessions, I may select a random group of students to lead us in the discussion and debate. Because the selection of participants will not be announced until class begins, everyone will be expected to prepare for the discussion. Reading the assigned chapters and articles are the best way to prepare for the discussion. If you have concerns about being called on in class, please see me to discuss. The purpose of the "hot seat" is not to stress or embarrass students, but to encourage students to actively engage the material.

2.5 Phones

Phones are not allowed to be used in class without the instructor's prior consent. If you have a need of a phone during class please let me know before class. Unauthorized use of electronic devices may result in the lowering of the grade or dismissal from the class. **I mean this.**

The phone thing? I mean this.

2.6 Attendance

You are expected to be regular and punctual in your class attendance. Students are responsible for all the material missed and homework assignments made. If class is missed, notes/homework should be obtained from another student. If I am more than 15 minutes late, class is considered cancelled. No more than 4 absences are allowed during a semester. Exceeding the absence policy may result in receiving an F for the course. The professors roll is the official roll and students not present when roll is taken will be counted as absent. If a

student must miss an exam, she or he must work out an agreeable time with the instructor to take the test prior to the exam being given. If a student misses a test due to an emergency, the student must inform the instructor as soon as is possible. In special cases, the instructor may allow the student to take a make-up exam.

2.7 Accommodations

Presbyterian College is committed to providing reasonable accommodations for all students with documented disabilities. If you are seeking academic accommodations under the Americans with Disabilities Act, you must register with the Academic Success Office, located on 5th Avenue (beside Campus Police). To receive these accommodations, please obtain the proper Accommodations Approval Form from that office, and then meet with me at the beginning of the semester to discuss how we may deliver your approved accommodations. I especially encourage you to meet with me well in advance of the actual accommodations being provided, as it may not be feasible to offer immediate accommodations without sufficient advance notice (such as in the case of tests). I can assure you that all discussions will remain confidential. Disability Services information is located at this link http://bit.ly/PCdisabilityservices

Additionally, it is the student's responsibility to give the instructor one week's notice prior to each instance where accommodation will be required.

2.8 Honor Code and Plagiarism:

All assignments/exams must be your own work. Any copying or use of unauthorized assistance will be treated as a violation of PC's Honor Code. If you are unsure of what resources are allowed, please ask. Please note that all text longer than 7 words taken from ANY other source must be placed in quotations and cited. Also, summarizing ANY other source must also be cited. Using ANY other source and showing work to be your own is a violation of plagiarism and the honor code.

2.9 First-Generation Version:

I am a Presby First+ Advocate. I am here to support our current first-generation students. At Presbyterian College, first-generation students are those in which neither parent nor legal guardian graduated from a four-year higher education institution with a bachelor's degree. If you are a first-generation college student, please contact me. For more information about support for first-generation college students on our campus visit our Presby First+ webpage.

2.10 Continuing Advocate Version

I am a Presby First+ Advocate. I am committed to supporting first-generation students at Presbyterian College. At Presbyterian College, first-generation students are those in which neither parent nor legal guardian graduated from a four-year higher education institution with a bachelor's degree. If you are a first-generation college student, please contact me anytime or visit me during my office hours. For more information about support for first-generation college students on our campus visit our Presby First+ webpage.

Chapter 3

Our Class Rhythm

Monday: Wrap up previous topic and introduce what you've pre-read about. Chat. Play. Work some examples. Make sure the topics applies to real-life.

Wednesday: Work more examples. Chat as needed. **Live our best lives. :)**.

Friday: Apply what we've learned – demonstrate your mastery (typically in the form of a quiz, lab, or assignment). Rinse. Repeat.

Chapter 4

End in Mind

Dana Simmons: "Can you predict which students will enroll at PC?"

Christina Miller: ??? Well, can you? ???

Chapter 5

Schedule

This is a tentative schedule, and it will change. **BUT** I will do my very best to review this often so that we all stay on the same page and so that you may plan accordingly!

Spring 2022

Date	Topic
Monday, January 10, 2022	Intro and A1 review
Wednesday, January 12, 2022	Rmarkdown
Friday, January 14, 2022	Lab 1: Rmarkdown
Monday, January 17, 2022	MLK Holiday
Wednesday, January 19, 2022	ggplot
Friday, January 21, 2022	Lab 2: ggplot
Monday, January 24, 2022	EDA & ggplot
Wednesday, January 26, 2022	EDA & ggplot
Friday, January 28, 2022	Lab 3: EDA & ggplot
Monday, January 31, 2022	TIDY SPREAD AND GATHER (R4DS CH 9 DPLYR)
Wednesday, February 2, 2022	RELATIONAL DATA (R4DS CH 10 DPLYR)
Friday, February 4, 2022	QUIZ
Monday, February 7, 2022	STRINGS (R4DS CH 11 stringr)
Wednesday, February 9, 2022	STRINGS (R4DS CH 12 factors)
Friday, February 11, 2022	QUIZ
Monday, February 14, 2022	Dates and Times
Wednesday, February 16, 2022	Dates and Times
Friday, February 18, 2022	QUIZ
Monday, February 21, 2022	Functions
Wednesday, February 23, 2022	Functions

Date	Topic
Friday, February 25, 2022	QUIZ
Monday, February 28, 2022	Itertation
Wednesday, March 2, 2022	Itertation
Friday, March 4, 2022	QUIZ
Monday, March 7, 2022	LAUNCH PROJECT
Wednesday, March 9, 2022	INDEPENDENT PROJECT
Friday, March 11, 2022	INDEPENDENT PROJECT
Monday, March 14, 2022	SPRING BREAK
Wednesday, March 16, 2022	SPRING BREAK
Friday, March 18, 2022	SPRING BREAK
Monday, March 21, 2022	INDEPENDENT PROJECT
Wednesday, March 23, 2022	PRESENTATIONS
Friday, March 25, 2022	PRESENTATIONS
Monday, March 28, 2022	Model Builiding/ADVISING WEEK
Wednesday, March 30, 2022	Model Builiding/ADVISING WEEK
Friday, April 1, 2022	QUIZ
Monday, April 4, 2022	regrssion
Wednesday, April 6, 2022	stepwise addition/deletion
Friday, April 8, 2022	QUIZ
Monday, April 11, 2022	logistic regression
Wednesday, April 13, 2022	trees & forests
Friday, April 15, 2022	Easter Holidays
Monday, April 18, 2022	Easter Holidays
Wednesday, April 20, 2022	Model Builiding
Friday, April 22, 2022	QUIZ
Monday, April 25, 2022	PRESENTATIONS
Wednesday, April 27, 2022	PRESENTATIONS
Friday, April 29, 2022	LAST DAY
Monday, May 2, 2022	Final Exam 8:30 p.m. – F period

Chapter 6

Lab 1 Excercises

Let's make sure we feel good about BADM 371 material.

All open notes/internet/R4DS/etc., **but all work must be your own**.

Use the starwars data (dplyr package) to answer/do:

1. Who is the tallest individual? Shortest?
2. How many homeworlds are there?
3. Which homeworld has the most individuals? Fewest? Average # of idividuals per homeworld?
4. Make a plot of all individuals with mass on the x axis and height on the y axis.
5. Put a best fit line on this plot.
6. Who is the biggest outlier in this dataset?
7. Calculate BMI for all these individuals. What is the average BMI for all individuals?
8. What is the average BMI for each homeworld?
9. Which homeworlds have the greatest percentage of individuals with BMI's greater than the average you found in #8 above?
10. How many individuals have no missing data? Which variables have the most missing data?

Chapter 7

Lab 1 in Rmarkdown

7.1 R Markdown

```
library(dplyr)
```

1. Who is the tallest individual? Shortest?

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>    66.0   167.0   180.0   174.4   191.0   264.0       6
```

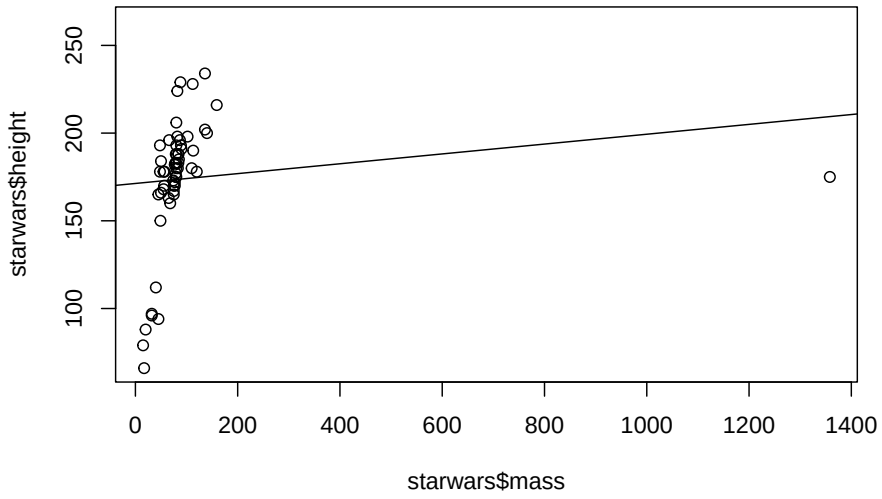
2. How many homeworlds are there?

```
#> # A tibble: 49 x 1
#>    homeworld
#>    <chr>
#>  1 Tatooine
#>  2 Naboo
#>  3 Alderaan
#>  4 Stewjon
#>  5 Eriadu
#>  6 Kashyyyk
#>  7 Corellia
#>  8 Rodia
#>  9 Nal Hutta
#> 10 Bestine IV
#> # ... with 39 more rows
```

3. Which homeworld has the most individuals? Fewest? Average # of individuals per homeworld?

```
#> # A tibble: 49 x 2
#>    homeworld     n
#>    <chr>     <int>
#>  1 Naboo        11
#>  2 Tatooine     10
#>  3 <NA>         10
#>  4 Alderaan      3
#>  5 Coruscant     3
#>  6 Kamino        3
#>  7 Corellia      2
#>  8 Kashyyyk      2
#>  9 Mirial        2
#> 10 Ryloth        2
#> # ... with 39 more rows
#> # A tibble: 49 x 2
#>    homeworld          n
#>    <chr>          <int>
#>  1 Aleen Minor        1
#>  2 Bespin             1
#>  3 Bestine IV         1
#>  4 Cato Neimoidia     1
#>  5 Cerea              1
#>  6 Champala           1
#>  7 Chandrila          1
#>  8 Concord Dawn       1
#>  9 Dathomir           1
#> 10 Dorin              1
#> # ... with 39 more rows
```

4-6. Make a plot of all individuals with mass on the x axis and height on the y axis. Put a best fit line on this plot. Who is the biggest outlier in this dataset?




```
#> # A tibble: 1 x 3
#>   name                  mass height
#>   <chr>                <dbl>  <int>
#> 1 Jabba Desilijic Tiure  1358    175
```

7. Calculate BMI for all these individuals. What is the average BMI for all individuals?

   Via google: With the metric system, the formula for BMI is weight in kilograms divided by height in meters squared. Since height is commonly measured in centimeters, an alternate calculation formula, dividing the weight in kilograms by the height in centimeters squared, and then multiplying the result by 10,000, can be used

```
#> # A tibble: 59 x 4
#>    name                BMI height  mass
#>    <chr>             <dbl>  <int> <dbl>
#>  1 Luke Skywalker     26.0    172    77
#>  2 C-3PO              26.9    167    75
#>  3 R2-D2              34.7     96    32
#>  4 Darth Vader        33.3    202   136
#>  5 Leia Organa        21.8    150    49
#>  6 Owen Lars          37.9    178   120
#>  7 Beru Whitesun lars 27.5    165    75
```

```
#>  8 R5-D4              34.0     97    32
#>  9 Biggs Darklighter  25.1    183    84
#> 10 Obi-Wan Kenobi     23.2    182    77
#> # ... with 49 more rows
#> # A tibble: 1 x 1
#>   `mean(BMI)`
#>         <dbl>
#> 1        32.0
```

8. What is the average BMI for each homeworld?

```
#> # A tibble: 40 x 2
#>    homeworld  avg.BMI
#>    <chr>        <dbl>
#>  1 Nal Hutta    443.
#>  2 Vulpter       50.9
#>  3 Kalee         34.1
#>  4 Bestine IV    34.0
#>  5 <NA>          32.6
#>  6 Malastare     31.9
#>  7 Trandosha     31.3
#>  8 Tatooine      29.3
#>  9 Sullust       26.6
#> 10 Dathomir      26.1
#> # ... with 30 more rows
```

9. Which homeworlds have the greatest percentage of individuals with BMI's greater than the average you found in #8 above? How many individuals have no missing data? Which variables have the most missing data?

```
#> # A tibble: 5 x 2
#>   homeworld  avg.BMI
#>   <chr>        <dbl>
#> 1 Nal Hutta    443.
#> 2 Vulpter       50.9
#> 3 Kalee         34.1
#> 4 Bestine IV    34.0
#> 5 <NA>          32.6
```

10. How many individuals have no missing data? Which variables have the most missing data?

    Via google: https://stackoverflow.com/questions/22353633/filter-for-complete-cases-in-data-frame-using-dplyr-case-wise-deletion

```
#> # A tibble: 29 x 14
#>    name        height  mass hair_color  skin_color eye_color
#>    <chr>        <int> <dbl> <chr>       <chr>      <chr>
#>  1 Luke Skywa~    172    77 blond       fair       blue
#>  2 Darth Vader    202   136 none        white      yellow
#>  3 Leia Organa    150    49 brown       light      brown
#>  4 Owen Lars      178   120 brown, grey light      blue
#>  5 Beru White~    165    75 brown       light      blue
#>  6 Biggs Dark~    183    84 black       light      brown
#>  7 Obi-Wan Ke~    182    77 auburn, wh~ fair       blue-gray
#>  8 Anakin Sky~    188    84 blond       fair       blue
#>  9 Chewbacca      228   112 brown       unknown    blue
#> 10 Han Solo       180    80 brown       fair       brown
#> # ... with 19 more rows, and 8 more variables:
#> #   birth_year <dbl>, sex <chr>, gender <chr>,
#> #   homeworld <chr>, species <chr>, films <list>,
#> #   vehicles <list>, starships <list>
#> Warning: `funs()` was deprecated in dplyr 0.8.0.
#> Please use a list of either functions or lambdas:
#>
#>   # Simple named list:
#>   list(mean = mean, median = median)
#>
#>   # Auto named with `tibble::lst()`:
#>   tibble::lst(mean, median)
#>
#>   # Using lambdas
#>   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
#> This warning is displayed once every 8 hours.
#> Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
#> # A tibble: 1 x 14
#>    name height  mass hair_color skin_color eye_color
#>   <int>  <int> <int>      <int>      <int>     <int>
#> 1     0      6    28          5          0         0
#> # ... with 8 more variables: birth_year <int>, sex <int>,
#> #   gender <int>, homeworld <int>, species <int>,
#> #   films <int>, vehicles <int>, starships <int>
```

Chapter 8

Lab 2: Pretty pictures!

Please make sure you have read and understood R4DS Chapter on data visulization. Also check out the *Data Visualization with ggplot2 Cheat Sheet* from RStudio. and

Think DEEPLY: Why is being able to generate good data vislization in R important *even* with awesome tools like PowerBI and Tableau around?

8.1 ggplot package and code

```
ggplot(data = ___, mapping = aes(x = ___)) +
  geom_histogram(binwidth = ___) +
  facet_wrap(~___)

Let's deconstruct this code:
- `ggplot()` is the function we are using to build our plot, in layers.
- In the first layer we always define the data frame as the first argument. Then, we define the m
- In the next layer we represent the data with **geom**etric shapes, in this case with a histogra
- In the final layer we facet the data by neighbourhood.
```

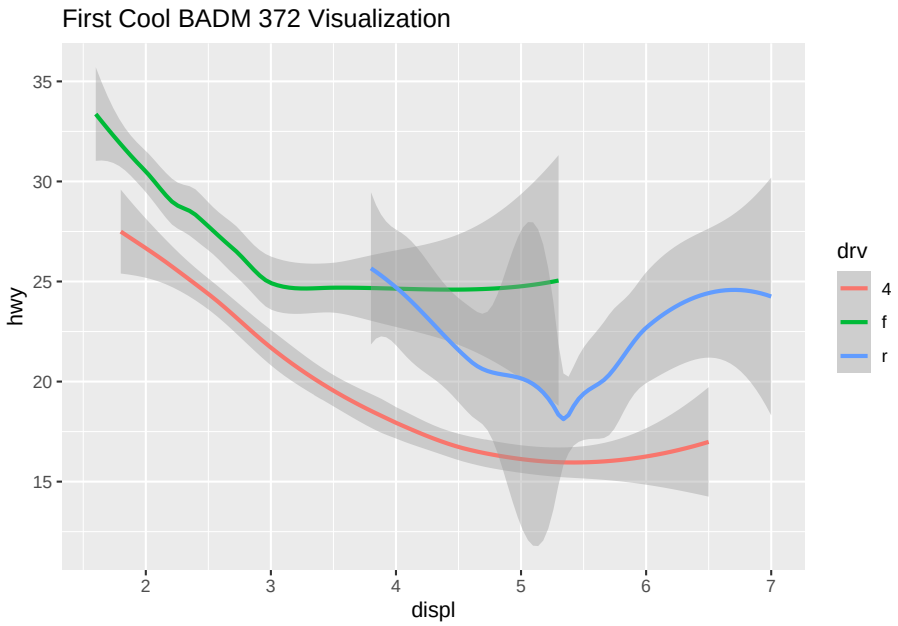
8.2 Packages

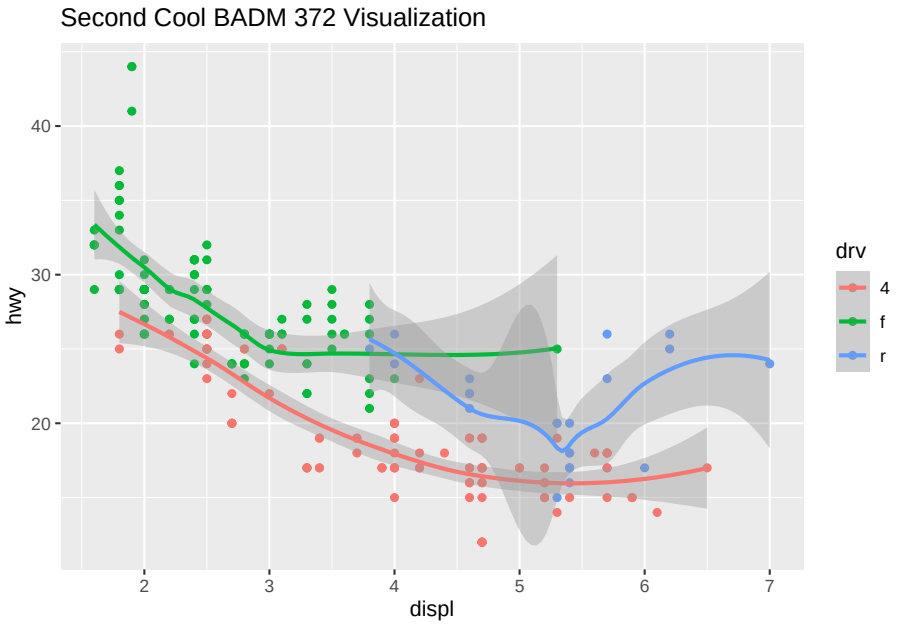
We'll use the **tidyverse** packages for this analysis, and the data is in the **dsbox** package. Run the following code in the Console to load these packages.

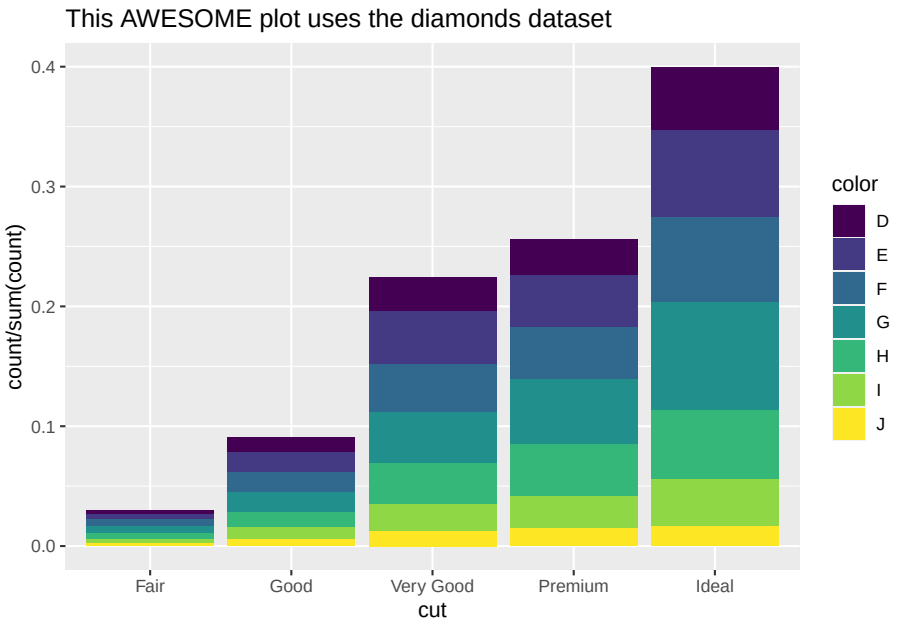
```
library(tidyverse)
library(dsbox)
```

8.3 Excercises

1. Create these figures using the data sets `mpg` or `diamonds` as needed:




Second Cool BADM 372 Visualization
hwy
40
30
20
2
3
4
5
6
7
displ
drv
4
f
r


This AWESOME plot uses the diamonds dataset
count/sum(count)
0.4
0.3
0.2
0.1
0.0
Fair
Good
Very Good
Premium
Ideal
cut
color
D
E
F
G
H
I
J

8.4 Airbnb listings in Edinburgh

This data comes from the `dsbox` package. Recent development in Edinburgh regarding the growth of Airbnb and its impact on the housing market means a better understanding of the Airbnb listings is needed. Using data provided by Airbnb, we can explore how Airbnb availability and prices vary by neighborhood.

The data come from the Kaggle database. It's been modified to better serve the goals of this exploration.

8.4.1 Learning goals

The goal of this assignment is not to conduct a thorough analysis of Airbnb listings in Edinburgh (yet?), but instead to give you a chance to practice your workflow, data visualization, and interpretation skills.

8.4.2 Data

2. The dataset you'll be using is called `edibnb` the data is in the **dsbox** package. Run `View(edibnb)` in your Console to view the data in the data viewer. What does each row in the dataset represent?

```
**Hint:** The Markdown, ggplot2, and dplyr Quick Reference sheets has an example of in
```

3. How many observations (rows) does the dataset have? What interesting data is present? What was the purpose of this data being collected in the first place? Visit the kaggle site if needed.

Each column represents a variable. We can get a list of the variables in the data frame using the `names()` function. How else can we find out details of about these variables?

```
names(edibnb)
#>  [1] "id"                   "price"
#>  [3] "neighbourhood"        "accommodates"
#>  [5] "bathrooms"            "bedrooms"
#>  [7] "beds"                 "review_scores_rating"
#>  [9] "number_of_reviews"    "listing_url"
```

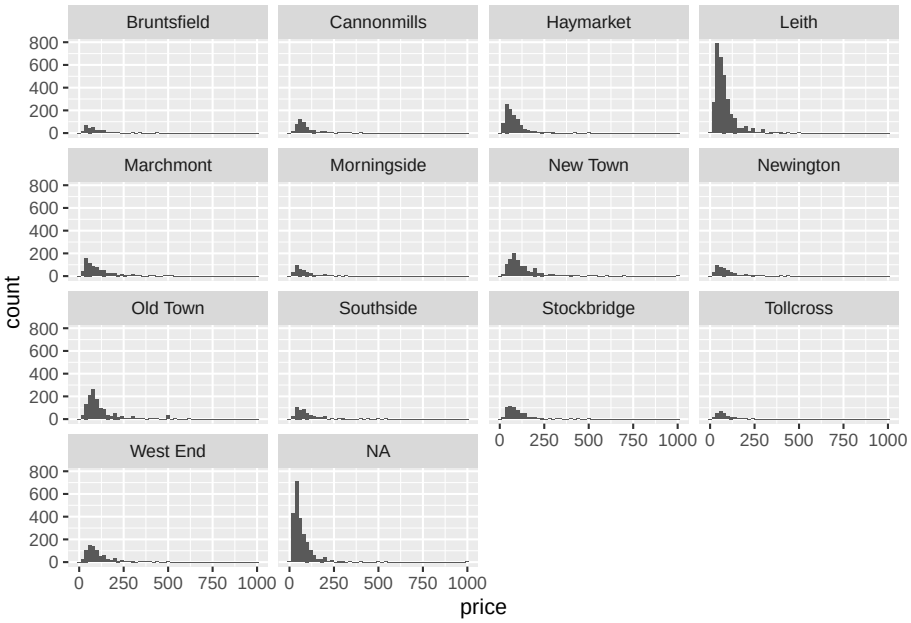
You can find descriptions of each of the variables in the help file for the dataset, which you can access by running `?edibnb` in your Console.

4. Create a faceted histogram where each facet represents a neighborhood and displays the distribution of Airbnb prices in that neighborhood. Your histogram may be similar (or better! than the example below.)

5. Create a faceted histogram where each facet represents a neighborhood and displays the distribution of Airbnb prices in that neighborhood. You histogram may be similar (or better! than the example below.)

```
**Note:** The plot will give a warning about some observations with non-finite values for price b
```

```
#> Warning: Removed 199 rows containing non-finite values
#> (stat_bin).
```



6. Create a similar visualization, this time showing the distribution of review scores (`review_scores_rating`) across neighborhoods. In your answer, include a brief interpretation of how Airbnb guests rate properties in general and how the neighborhoods compare to each other in terms of their ratings.

7. Create another informative visualization of your choosing. Be prepared to share it with the class – although the visualization should need no explaining!

8.5 Instructional staff employment trends

The next dataset is about instructional staff employee hiring trends between 1975 and 2011.

The dataset is called `instructors` found in `dsbox`. You can find descriptions of each of the variables in the help file for the dataset, which you can access by running `?instructors` in your Console.

The American Association of University Professors (AAUP) is a nonprofit membership association of faculty and other academic professionals. This report compiled by the AAUP shows trends in instructional staff employees between 1975 and 2011, and contains an image very similar to the one given below.

8.6 More Excercises

8. Recreate a graph similar to the one above.

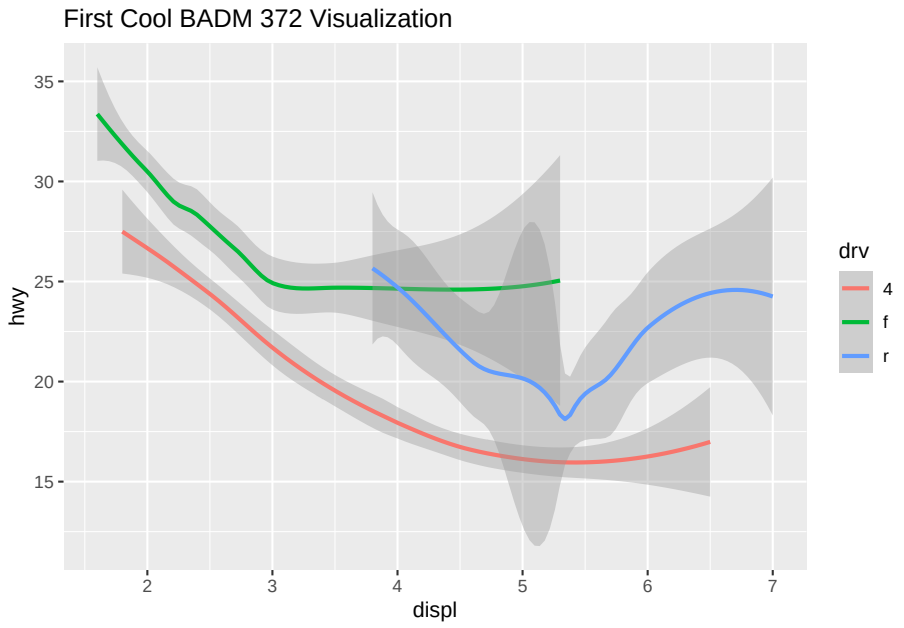
9. Discuss how you would improve upon this visualization if the main objective was to communicate that the proportion of part-time faculty have gone up over time compared to other instructional staff types.Implement the improvements and provide your improved visualization as part of your answer. Also write a few sentences about why you chose to make these improvements and how they address the main goal stated above.

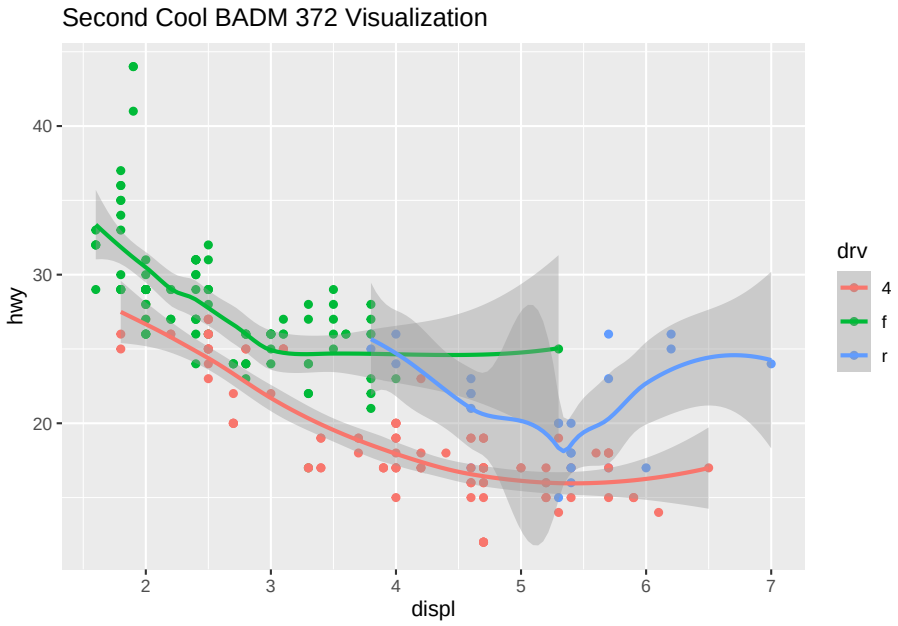
Chapter 9

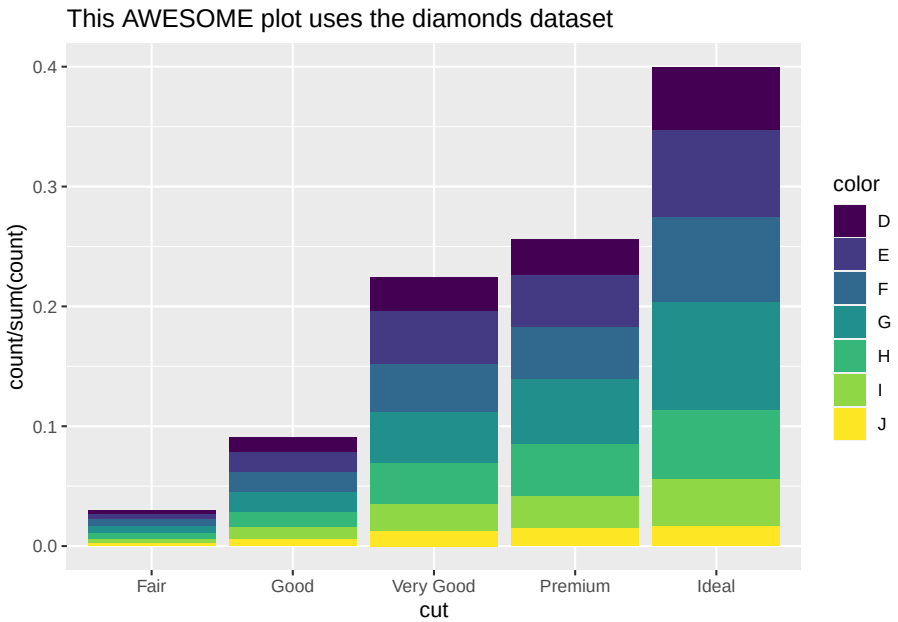
Lab 2 – ggplot without `dsbox`

9.1 Excercises using the data sets `mpg` or `diamonds`

1. Create these figures using the data sets `mpg` or `diamonds` as needed:




Second Cool BADM 372 Visualization
hwy
displ
drv
4
f
r
40
30
20
2
3
4
5
6
7

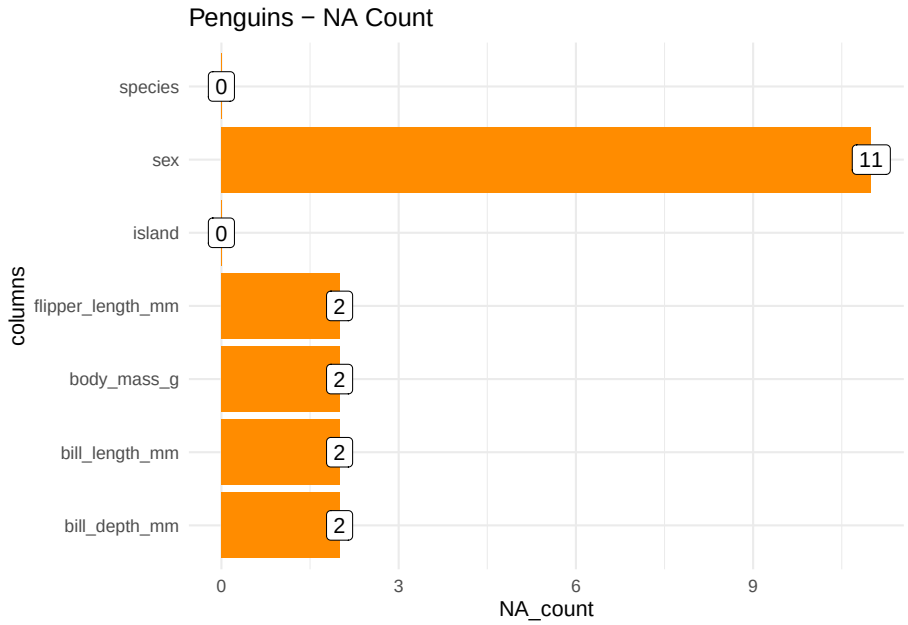

This AWESOME plot uses the diamonds dataset
count/sum(count)
cut
color
D
E
F
G
H
I
J
0.4
0.3
0.2
0.1
0.0
Fair
Good
Very Good
Premium
Ideal

9.2 palmerpenguins

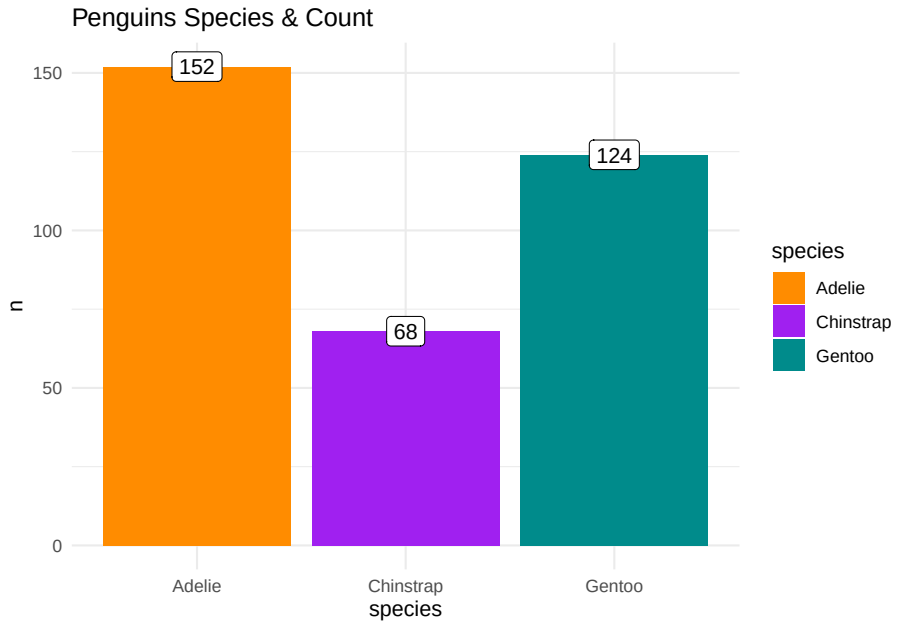
`palmerpenguins` is a realtively new package on CRAN, so you can install it from CRAN instead of Github.

Install it like a normal package. After successful installation, you can find out that there are two datasets attached with the package – penguins and penguins_raw. You can check out their help page (?penguins_raw and ?penguins_raw) to understand more about respective datasets.

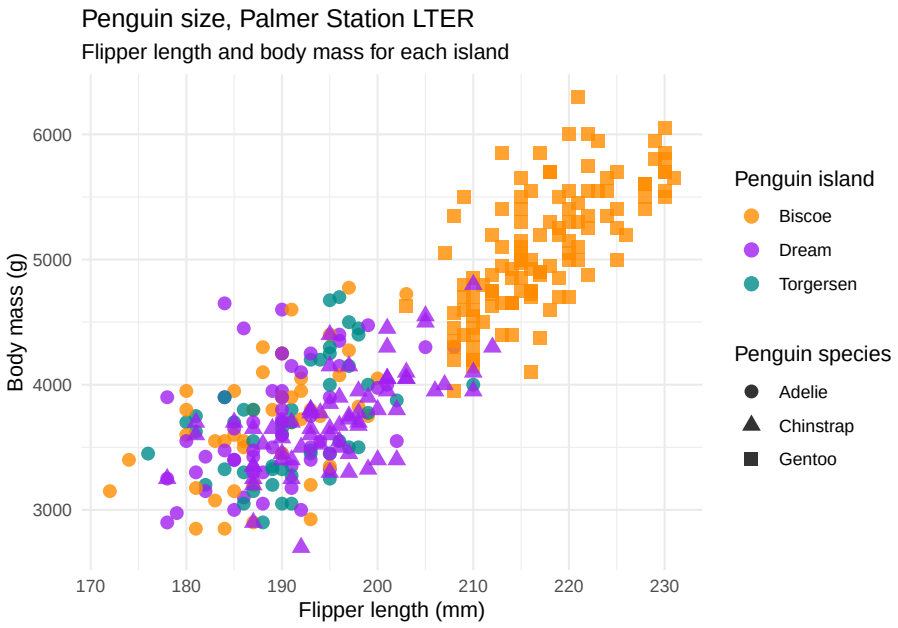
1. Please make a well-labeled, meangingful plot that show how many missing variables there are for each variable in the dataset. Your results shoud look something like this:



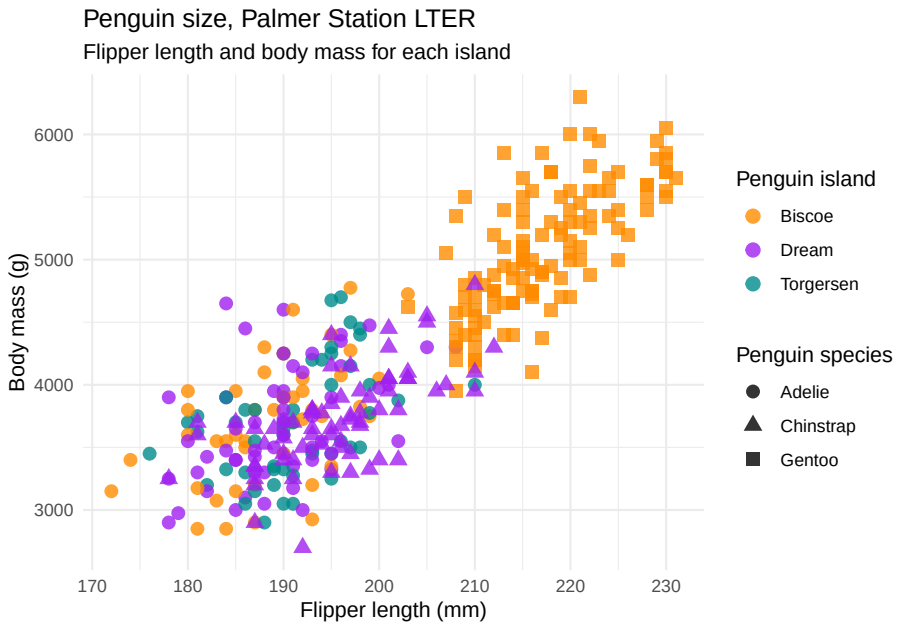
2. Make a plot showing a count of penguins of each species.



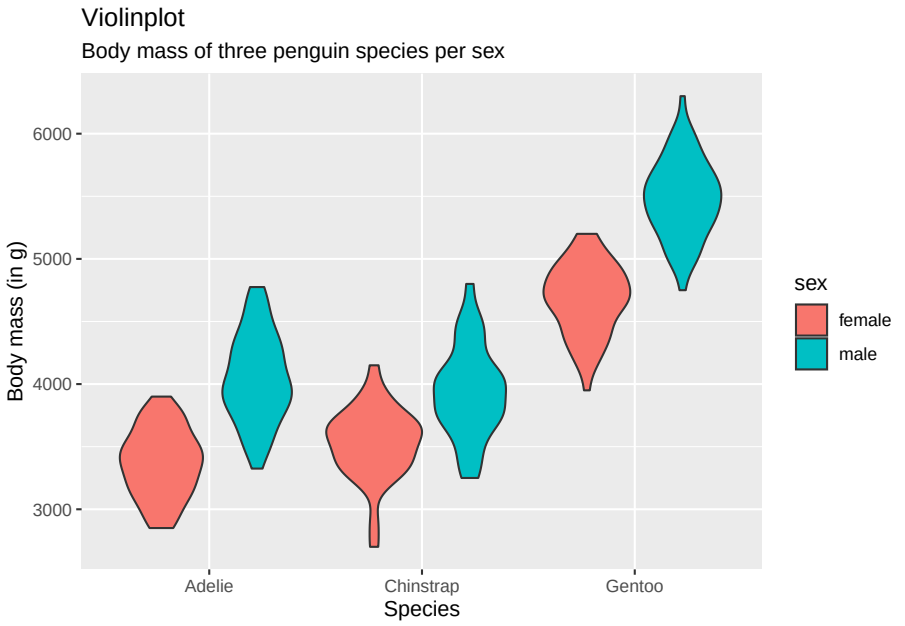
3. Create a plot that illustrates the relationship between flipper_length_mm and body_mass_g with respect to each species.

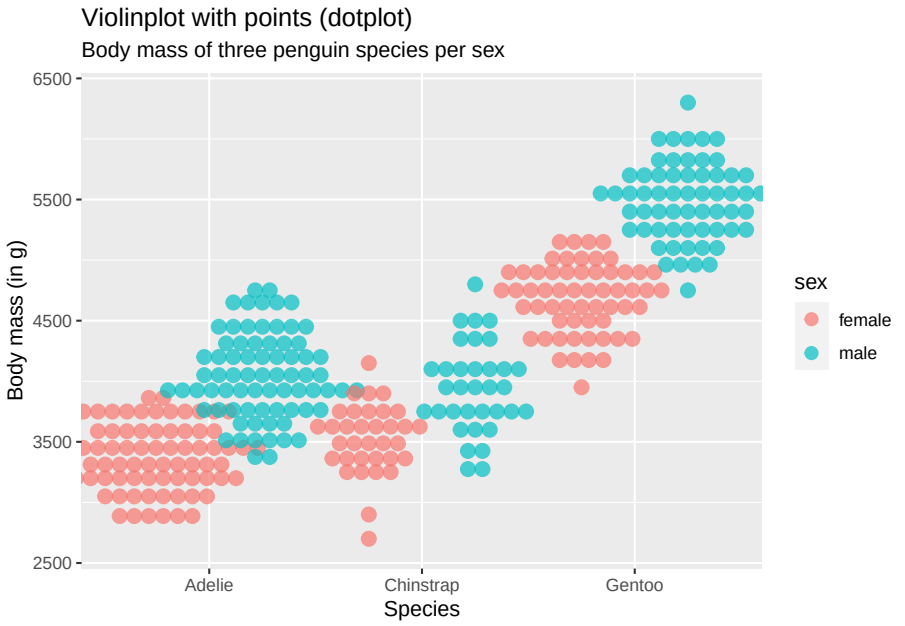


4. Create a plot that illustrates the relationship between flipper_length_mm and body_mass_g with respect to each species for each island.



5. Create a few plots of your own using new/interesting geoms and make sure the plots have meangiful, imprmfative labels, too. For possible examples:


Violinplot
Body mass of three penguin species per sex
Body mass (in g)
6000
5000
4000
3000
Adelie
Chinstrap
Gentoo
Species
sex
female
male


Violinplot with points (dotplot)
Body mass of three penguin species per sex
Body mass (in g)
6500
5500
4500
3500
2500
Adelie
Chinstrap
Gentoo
Species
sex
female
male