Midterm Exam

ST 597 | Spring 2017
University of Alabama

Setup

Instructions

- you are allowed three things: 1 page of notes, the Data Visualization cheatsheet, the Data Transformation cheatsheet
- you may not have any other tabs opened besides midterm.R and the datasets
- Do not open any other program besides RStudio. I will consider it a violation of the honor policy and you will be reported for academic violation.

Getting Started

1. Open RStudio
2. Close all tabs in the script pane
3. Clear your Environment: Session -> Restart R
4. Clear your History: Go to History Tab and click on the Broom to clear all
5. Load tidyverse by typing `library(tidyverse)`
    - if this is not working, then you must first install it `install.packages("tidyverse")`
6. After I announce the data transfer do the following:
    a. Open the exam script: `File -> Open File...`, then open `C:/Insight Files/midterm.R`.
    b. Load the data by typing: `load("C:/Insight Files/examdata.RData")` in the R console. See the midterm.R file for the code. You should see 4 datasets in your environment: `offers`, `people`, `scores`, `yelp`.
7. Put your name at the top of midterm.R
8. Do not change the file name or move midterm.R. But do save the file regularly.

After Finishing Exam:

- save midterm.R (ensure it is saved in `C:/Insight Files/midterm.R`)
- save History: go to History Tab and click on save button. Save as: `C:/Insight Files/history`
- Raise your hand to indicate you are finished with the exam and ready to submit.
- After I acknowledge you, you can open a browser and email me the *two* files:
    - `C:/Insight Files/midterm.R`, `C:/Insight Files/history`
    - `mporter@cba.ua.edu`
- **DO NOT LOGOFF!** until I give you permission. I am also trying to retrieve the files remotely.

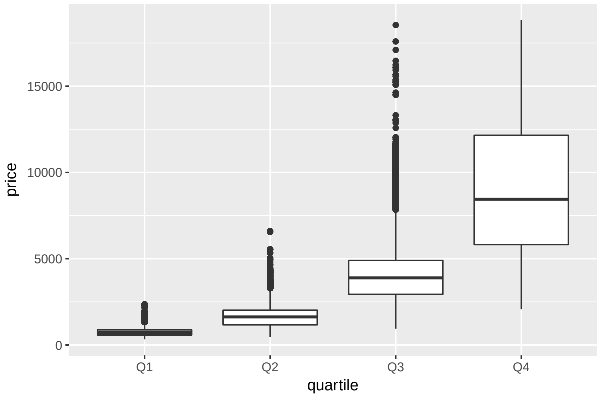
More Diamonds

Use the diamonds data from the ggplot2 package (part of tidyverse):

```r
library(tidyverse)
data(diamonds)
```

---

1. Create a scatterplot to show the relationship between carat and price.
    - put `carat` on x-axis and `price` on the y-axis
    - color all of the points blue
    - set the shape of the points according to `cut`
    - set the size of the points according to `clarity`
    - add a smooth curve fit with line color of `orange` and fill color of `black`
2. Make this boxplot of diamond price for each quartile of carat.



The Perfect Job

You should see three data sets in your environment:

- `offers`: job offers made to applicants
- `people`: applicants and their personalities
- `scores`: score (utility) for jobtype - personality combinations

---

3. How many offers did each person (`name`) receive?
    - Create a tibble (or data frame) that shows the number of offers per person
    - order the table so the person with the most offers is first
    - Resolve any ties by reverse alphabetical order (so Bob would come before Amy if both have same number of offers)
4. Find the best job offer for each person.
    - Create a tibble (or data frame) that shows the best offer for each person
    - The best offer is the offer with the highest score
    - Hint: you need to combine the data so the score for the `jobtype` and `personality` can be determined for each offer
    - some people have multiple offers with same best score. You can return one or all of these.

Yelp

The following problems requre the `yelp` data.

The columns are:

- `review_id`: the id for the review
- `user_id`: the reviewer's id
- `date`: date of review
- `stars`: the star rating (1-worst, 5-best)
- `bus_category`: the type of business being reviewed
- `bus_id`: the id for the business being reviewed

---

5. Create a tibble (or data frame) of all 4-star (`stars`) reviews of restaurants and nightlife (`bus_category`) businesses.

6. Average Star Rating

    a. Calculate the average star rating (`stars`) from all reviews. Report the answer.

    b. Calculate the average stars rating (`stars`) for every business category (`bus_category`) and report the category with the largest average star rating.

7. Which business category (`bus_category`) has the highest proportion of 1-star (`stars`) reviews?

8. Produce a plot that shows the number of reviews in each `bus_category` and `stars` pair. Use any method you want, but the resulting graphic should enable me to see at a glance e.g. the approximate number of *2-star restaurant* reviews.