Problem Set 1

Steven Cao

10/11/2019

Part 1, Question 1

The paper from which the dataset was derived can be found here: https://peerj.com/articles/2537/

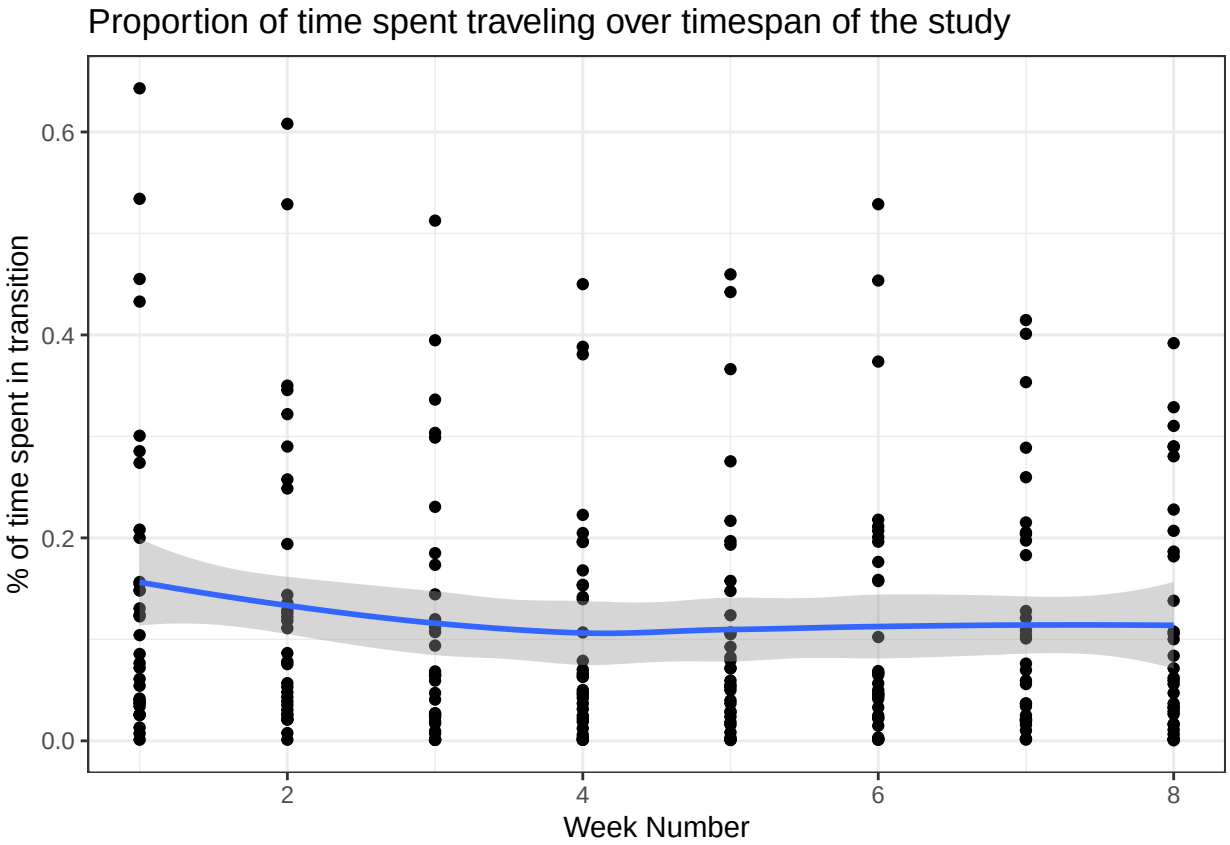
The dataset itself can be found here: https://dfzljdn9uc3pi.cloudfront.net/2016/2537/1/data.zip (Alternatively, it can be found on the page on which the paper is hosted.)

Part 1, Questions 2 & 3

```
# load preprocessed data
trackingData <- read.csv("data.csv")

# extract just two variables of interest, time (week number) and "location variance" (cf. referenced pa
transitionTimes <- trackingData %>%
  dplyr::select(weekNumber, TransitionTime)

# plot our two variables as a scatterplot
ggplot(transitionTimes, aes(x=weekNumber, y=TransitionTime) ) +
  geom_point() +
  geom_smooth(method = loess) +
  labs(x = "Week Number",
       y = "% of time spent in transition",
       title = "Proportion of time spent traveling over timespan of the study") +
  theme_bw()
```



The average percentage of time that the subjects spent each week traveling (i.e. moving from place to place) is fairly constant.

Part 1, Question 4

```
skim(transitionTimes)
```

```
## Skim summary statistics
##  n obs: 264
##  n variables: 2
##
## -- Variable type:integer -------------------------------------------------------------------------
##    variable missing complete   n mean  sd p0  p25 p50  p75 p100     hist
##  weekNumber       0      264 264  4.5 2.3  1 2.75 4.5 6.25    8 <U+2587><U+2587><U+2587><U+2587><U+25
##
## -- Variable type:numeric -------------------------------------------------------------------------
##        variable missing complete   n mean   sd      p0   p25   p50  p75
##  TransitionTime       0      264 264 0.12 0.13 0.00054 0.026 0.067 0.18
##  p100     hist
##  0.64 <U+2587><U+2582><U+2582><U+2581><U+2581><U+2581><U+2581><U+2581>
```

Part 1, Question 5

The average percentage of time spent each week in transition is 12% (very roughly, 3 hours a day), with a standard deviation of 13% (very roughly, also 3 hours a day). Although, standard deviation is not a very good metric in this case because of the leftward skew: people are much more likely to spend less time in transition rather than, say, the majority of their day traveling. (For those who do, it might be part of their job.)

Part 2, Question 1

Visual exploratory data analysis serves both as a great way of not only seeing but easily understanding and comprehending patterns in the data. Additionally, it tends to serve as a "reality anchor" of "what exactly is the nature of the data I'm working with". For example, it tends to be easier to see and notice outliers when data is actually visualised.

Numeric exploratory data analysis, on the other hand, allows for the manipulation of data so as to uncover further possible patterns and trends: for instance, looking at whether the mean and median of a dataset differ greatly from one another (which would suggest either outliers, skewness, or both).

Part 2, Question 2

Example 1: https://upload.wikimedia.org/wikipedia/commons/d/d1/20181204_Warming_stripes_%28global%2C_WMO%2C_1850-2018%29_-_Climate_Lab_Book_%28Ed_Hawkins%29.png

The figure is supposed to represent global warming trends (more specifically, temperature in some part of the globe). This is something of an extreme example of the paucity of information (e.g. lack of a legend for the colour-coding, lack of axes, and so on). To be fair to the author, though, that kind of minimalism was very explicit in the intentions.

Example 2: https://upload.wikimedia.org/wikipedia/commons/e/e5/Opinion_polling_UK_2020_election_short_axis.png

The axes could use some work: the y-axis could use a label to describe what it is exactly, whereas the x-axis could be made a bit tidier by using numerical values instead of listing the name of the months, which lends itself to visual overload. (While the legend is missing from the image itself, it is present within the caption of the embed on Wikipedia, from which this was pulled.)

Part 2, Question 3

Example 1: https://upload.wikimedia.org/wikipedia/commons/3/35/Human_losses_of_world_war_two_by_country.png

The graph's title and legend is descriptive and simple to read, while each categorical entry (i.e. each country) is easy to skim through. Additionally, each country is listed in descending order, based on the absolute number of casualties.

Example 2: https://upload.wikimedia.org/wikipedia/commons/8/8b/Moore%27s_Law_Transistor_Count_1971-2018.png

I'm going to need to qualify what about this I think was done well, because on the one hand, there is a lot of text and information that - while there is good reason to put it there - could be temporarily hidden for the sake of visual ease. But the main point that I want to get to is the decision to use a logarithmic scale for the Y-axis instead of a linear one, because it can be "easier" to look at and gauge a curve which is linear than a curve which is, well, curved (because not all curves are due to being an exponential function).

Part 2, Question 4

EDA is very good to use before doing any kind of confirmatory data analysis because it helps the researchers get a bearing on the nature of the data they are dealing with. Knowing the nature of the data can also help the researchers decide which confirmatory approach is more appropriate to use. As was seen with Anscombe's quartet, it's easy (or at least, easier) to judge graphs than to stare at tables of numbers.

Part 2, Question 5

"Confirmatory data analysis" is where one tests a hypothesis about the data. Colloquially and crudely, one believes that his/her data is of a certain nature: for example, that the behaviour between two variables is linear. Then one performs a test to see how well his/her preconceived model "squares up with" or "explains" the data that he/she has gotten. We are seeing how well something fits and making assumptions and testing assumptions about how the data behaves (which also includes assumptions about where that data came from). An example would be a linear regression between engine size and fuel efficiency.

"Exploratory data analysis" makes less assumptions about the data than confirmatory data analysis, because we are not trying to test a hypothesis - we are only trying to "look at it" for "what it is" as best as we can. It can be said that it is during EDA that we begin to uncover underlying structures/patterns in the data, which then warrants us making additional assumptions. (Meaning that EDA is conducive to confirmatory data analysis.) An example would be making boxplots to compare the effect size of some drug and a placebo.