

Lab Visual-Eyes

Abhi Thanvi, Paul Holaway

June 28th, 2022

Contents

Lab Visual-Eyes	2
Welcome	2
The Idea of this Lab	2
Problem 1: Go Big or Go Home...	2
Problem 2	6
Part 1: We are all Scattered...	6
Part 2: We might be wrong :'(.	7
Submission	7

Lab Visual-Eyes

Welcome

Just like learning a new spoken language, you will not learn the language without practice. Labs are an important part of this course. Collaboration on labs is **extremely encouraged**. If you find yourself stuck for more than a few minutes, ask a neighbor or course staff for help. When you are giving help to your neighbor, explain the **idea and approach** to the problem without sharing the answer itself so they can figure it out on their own. This will be better for them and for you. For them because it will stick more and they will have a better understanding of the concept. For you because if you can explain it to other students, that means you understand it better too.

The Idea of this Lab

The idea behind this lab is to allow us to understand the importance and usefulness of data visualizations. As part of being a Data Scientist or Analyst, you will be processing tons of data. Sometimes it is easier to look at a snapshot of the data through visual graphs and plots. Another reason why data visualization is useful is to share your insights. Often times in the real world, you will be providing insights to the business or other stakeholders (people who invest time and resources in your project). As a result, you will be sharing a lot of your insights with them, and many times you need to communicate your analysis effectively. This is where Data Visualizations comes in handy as it allows you to communicate your ideas in a very convincing manner to even some [non technical stakeholders!](#). Therefore, this lab is designed to give you an overview what different plots are and how to interpret them.

“When it comes to data, I am a man of few words...that’s why I plot them.” - Woke Abhi

Problem 1: Go Big or Go Home...

Howdy! This is Abhi, today I am gonna follow the common saying Go Big or Go Home. Unfortunately, I won’t be able to pay for your ride back home...so you are stuck to “Go Big” with this part. So the way you are gonna big is that I am gonna put a lot of responsibility on y’all. I will be giving you a basic problem which is to plot a histogram but you will be spotting some issues, and you will have to figure out how to tackle those. I have complete faith in you that you are more than capable to do this. I am not giving you such a task because I cannot solve it and Paul asked me to...maybe. Good Luck!

Question 1: Import the apple_sales data set and print it

Answer:

```
# Import and Print apple_sales.csv
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.1 --

## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.7      v dplyr   1.0.9
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
apple_sales <- read.csv("~/Desktop/DPI 2022/apple_sales.csv", stringsAsFactors=TRUE)
head(apple_sales,10)
```

##	Time	Period	iPhone	iPad	iPod	Mac
## 1	Q4/98	1	NA	NA	NA	0.944
## 2	Q1/99	2	NA	NA	NA	0.827
## 3	Q2/99	3	NA	NA	NA	0.905
## 4	Q3/99	4	NA	NA	NA	0.772
## 5	Q4/99	5	NA	NA	NA	1.377
## 6	Q1/00	6	NA	NA	NA	1.043
## 7	Q2/00	7	NA	NA	NA	1.016
## 8	Q3/00	8	NA	NA	NA	1.122
## 9	Q4/00	9	NA	NA	NA	0.659
## 10	Q1/01	10	NA	NA	NA	0.751

Question 2: So I want you to create a histogram that shows the sales for each apple product. Do you see any issues? How would this impact your histogram?

Hint: Are values filled out for each period?

Answer: The biggest issue with creating a histogram from this data set is that many values are labelled “NA” in this data set; in other words, there is no value at those points. This would impact our histogram, since we can’t fill our histogram with null values.

Question 3: Go through the data set and note the range of period where all 4 products sales have been recorded. Report it down below, it will help you later for filtering rows.

Hint: Your answer will have a start period and an end period (Ex: Periods 10 - 30)

Answer: Periods 47-64

Question 4: Let's filter the rows out where all 4 products sales are filled in. Let's go!

Answer:

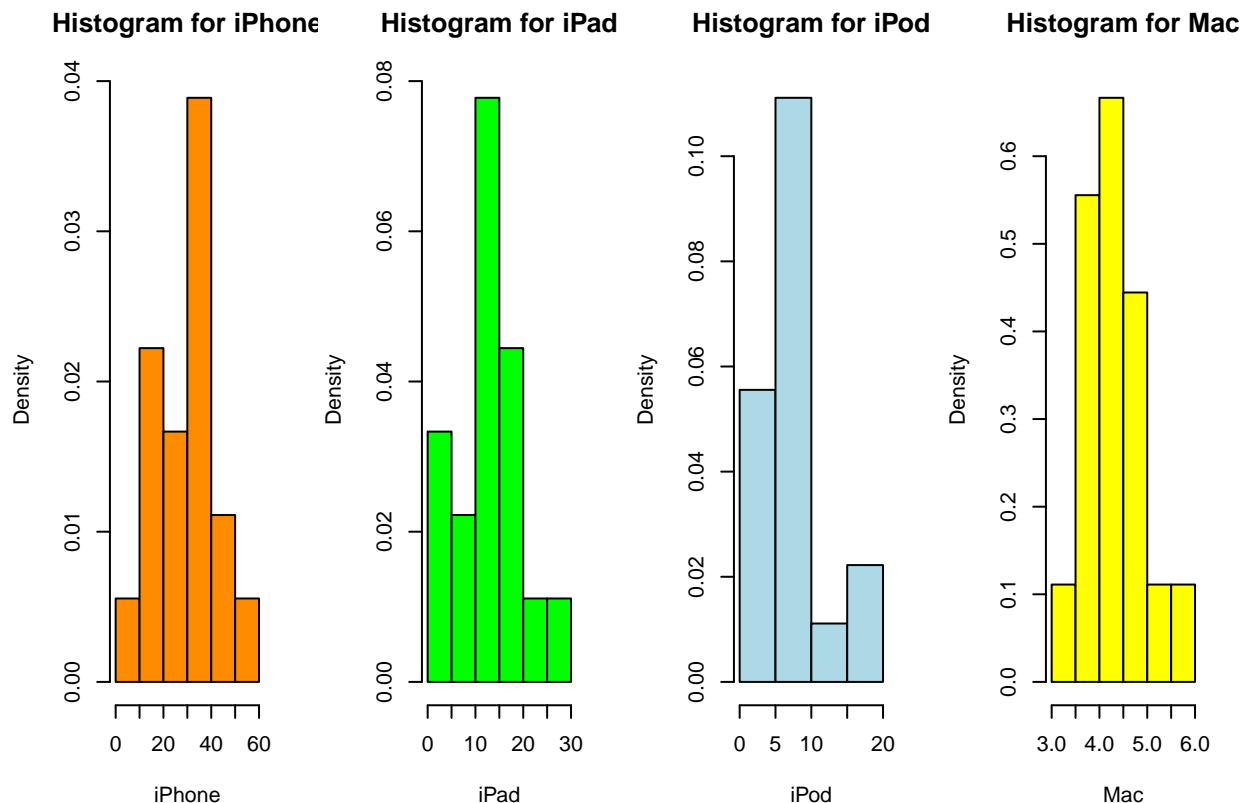
```
filtered_sales = drop_na(apple_sales)
filtered_sales
```

##	Time Period	iPhone	iPad	iPod	Mac
## 1	Q2/10	47 8.398	3.270	9.406	3.472
## 2	Q3/10	48 14.102	4.188	9.051	3.885
## 3	Q4/10	49 16.235	7.331	19.446	4.134
## 4	Q1/11	50 18.647	4.694	9.017	3.760
## 5	Q2/11	51 20.338	9.246	7.535	3.947
## 6	Q3/11	52 17.073	11.123	6.622	4.894
## 7	Q4/11	53 37.044	15.434	15.397	5.198
## 8	Q1/12	54 35.064	11.798	7.673	4.017
## 9	Q2/12	55 26.028	17.042	6.751	4.020
## 10	Q3/12	56 26.910	14.036	5.344	4.923
## 11	Q4/12	57 47.789	22.860	12.679	4.061
## 12	Q1/13	58 37.430	19.477	5.633	3.952
## 13	Q2/13	59 31.241	14.617	4.569	3.754
## 14	Q3/13	60 33.797	14.079	3.498	4.574
## 15	Q4/13	61 51.025	26.035	6.049	4.837
## 16	Q1/14	62 43.719	16.350	2.761	4.136
## 17	Q2/14	63 35.203	13.276	2.926	4.413
## 18	Q3/14	64 39.272	12.316	2.641	5.520

Question 5: Nice! Let's finish this up, can you make a dark orange histogram with x-axis for iPhone, green histogram for iPad, light blue histogram iPod, and yellow histogram for Mac. This means I need 4 histograms so you might need to adjust how you show me "4" histograms side-by-side.

Answer:

```
par(mfrow = c(1,4))
hist(filtered_sales$iPhone, xlab = "iPhone", main = "Histogram for iPhone",
      col = "dark orange", probability = TRUE)
hist(filtered_sales$iPad, xlab = "iPad", main = "Histogram for iPad",
      col = "green", probability = TRUE)
hist(filtered_sales$iPod, xlab = "iPod", main = "Histogram for iPod",
      col = "light blue", probability = TRUE)
hist(filtered_sales$Mac, xlab = "Mac", main = "Histogram for Mac",
      col = "yellow", probability = TRUE)
```



Thank you so much for helping me out! You guys rock! You have mastered making nice looking histograms in R!

Problem 2

Part 1: We are all Scattered...

Apparently, there is a reason why iPod sales were not there after a certain period. There were no significant iPod sales. My guess is it's because of iPhone.

Question 1: Could you think of reasons why iPod sales going down could be a reason of iPhone being introduced by apple? If you think it did not, justify why you think that iPhone is not the reason *or* provide an alternative reason for iPod's decline in sales.

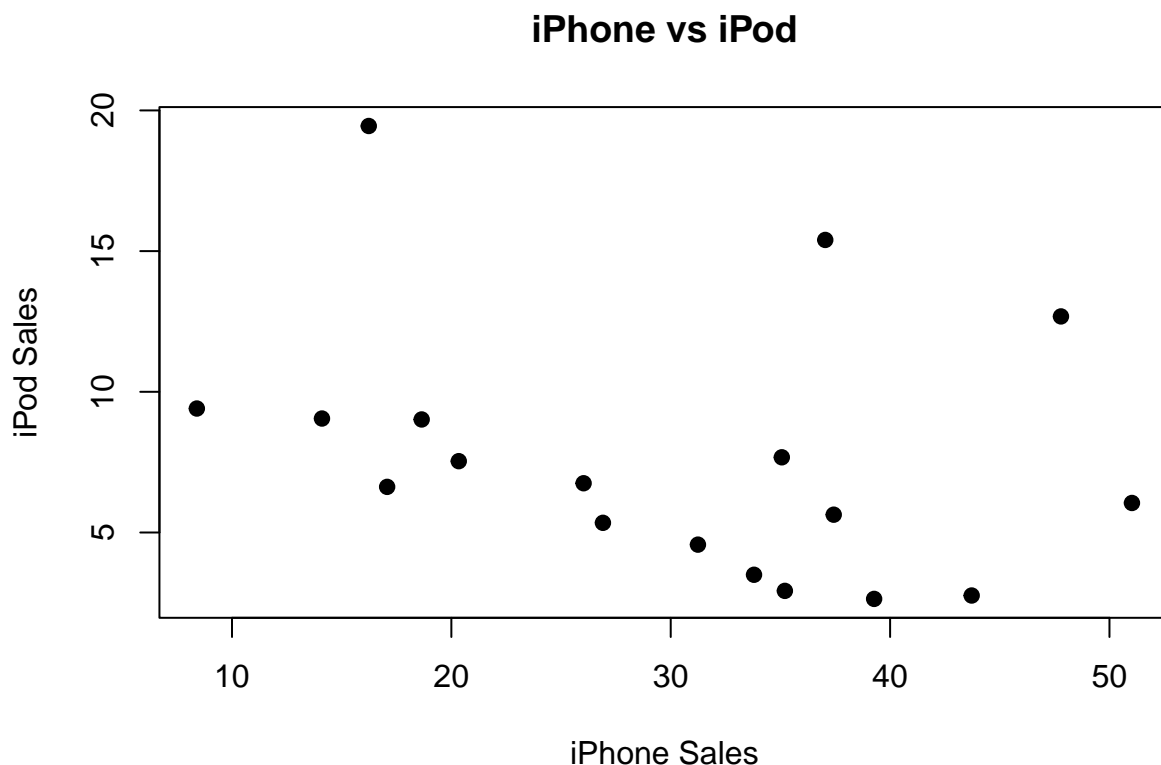
Answer: The reason that iPod sales went down is because of the introduction of the iPhone; since the iPhone can do everything that the iPod does (and more), many consumers saw no point in buying iPods, and sales plummeted.

Question 2: Oh I never thought about it! I think you all are market experts and are starting to build some expertise in using intuition to make sense of business market of Apple. This is an important skill to have, but let's back it up with data. Hopefully we are right about iPhone's introduction is the reason why iPod sales went down. *Create a scatter plot with x-axis being "iPhone Sales" and y-axis being "iPod Sales", and the main title being "iPhone vs iPod"*

Hint: Make sure to use the data frame with the period with all 4 products being present in the market. Also make sure to add `pch=19` at the end to have filled in dots. It just looks better

Answer:

```
plot(filtered_sales$iPhone, filtered_sales$iPod, xlab = "iPhone Sales",  
      ylab = "iPod Sales", main = "iPhone vs iPod", pch = 19)
```



Part 2: We might be wrong :(

Question 1: What kind of correlation do you see between iPhone sales and iPod sales based on your histogram. Make sure to be descriptive in terms of

- 1.) strength (weak, moderately strong, or strong)
- 2.) direction (positive or negative)
- 3.) shape (linear or non-linear)
- 4.) The variables considered for correlation

Hint: Strength refers to how close the points are together following an imaginary line. The closer they are to the imaginary line, the “stronger” we say it is. The farther they are from the imaginary line, the “weaker” it is.

Answer: There is a somewhat weak, negative, linear correlation between iPhone and iPad sales.

Question 2: So based on our previous response and our histogram, do you think iPhone *caused* iPod sales to decline.

Hint: Pay attention to the word cause and what correlation means!

Answer: I don’t think that the iPhone necessarily caused the iPod to decline, because the scatterplot shows only a weak negative correlation between iPhone sales and iPad sales. This means that although the iPhone’s increase in popularity may have contributed somewhat to the iPad’s decline, it probably wasn’t the biggest determining factor.

Question 3: This is sad! Let’s redeem ourselves! Could you think of more reasons or “associations” why iPod sales might have declined other than iPhone’s introduction to market? These are confounding variables we haven’t thought of, so give it some thought.

Answer: Another reason why the iPod may have declined was because the iPod’s main purpose was to play music; however, as time progressed, more people started downloading music or listening to it from apps instead of a speaker system like the iPod. The advent of apps like Spotify and YouTube may have contributed to the decline in need for the iPod, thus resulting in a decline in its sales.

You have achieved a lot today! You explored different plots and graphs, but also the importance of correlation and what it represents! Keep these ideas in mind, because we will revisit these conceptual ideas and build on top of those to do some fancy statistical tools!

Submission

Once you have finished your lab...

1. Go to the top left and click **File** and **Save**.
2. Click on the **Knit** button to convert this file to a PDF.
3. Submit **BOTH** the **.Rmd** file and **.pdf** file to Blackboard by 11:59 PM tonight.