Data Visualization

Pooja Chopra

7 June 2019

R Markdown

In this sections we will learn about how to visualize our data in a systematic way. In this we will use DPLY and GGPLOT togethe.

How to create ggplot

ggplot(df, aes(x= variable1 , y = variable2))

Layers

ggplot(df, aes(x=carat, y=price, color=cut)) + geom_point() + geom_smooth() # Adding scatterplot geompoint (layer1) and smoothing geom (layer2).

Aesthetic mappings

we can use third variable in the plot by mapping the aesthetics with term color, size,alpha , Shapapr

ggplot(df, aes(x=variable1, y=variable1, color=variable3)) + geom_point() + labs(title="Scatterplot", x="variable1", y="variable1") # add axis lables and plot title.

Facets

1. different values of cut plotted in the different chart.

ggplot(data = df) + geom_point(mapping = aes(x = variable1, y = variable2)) + # To split plot by single varaible we use facet wrap facet_wrap(~ variable3, nrow = 3)

ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy)) + # when we want put all the plots in a grid by two varaible we use facet grid facet_grid(drv ~ cyl)

Geometric objects

ggplot(df) + geom_smooth(mapping = aes(x = variable1, y = variable2)) #a smooth line fitted to the data.

ggplot(df) + geom_smooth(mapping = aes(x = displ, y = hwy, linetype = drv)) #geom_smooth() will draw a different line, with a different linetype.

ggplot(df) + geom_point(mapping = aes(x = variable1, y = variable2)) + ## Point display geom_smooth(mapping = aes(x = variable11, y = variablel2)) ## Line display

Time series Plot

If we want to Plot time series directly from a time series object use ggfortify package

Plot multiple timeseries on same ggplot

data(economics, package="ggplot2") # init data

economics <- data.frame(economics) # convert to dataframe

I: plot multiple time series using 'geom_line's

ggplot(economics) + geom_line(aes(x=date, y=pce, color="pcs")) + geom_line(aes(x=date, y=unemploy, col="unemploy")) + scale_color_discrete(name="Legend") + labs(title="Economics") # plot multiple time series using 'geom_line's

II: using melting Function library(reshape2) df <- melt(economics[, c("date", "pce", "unemploy")], id="date") ggplot(df) + geom_line(aes(x=date, y=value, color=variable)) + labs(title="Economics")# plot multiple time series by melting

Bar Chart

plot1 <- ggplot(df, aes(x=variable 1)) + geom_bar() + labs(title="Frequency bar chart") # Y axis derived from counts of X item

Custom layout

gridExtra package provides the facility to arrage multiple ggplots in a single grid

grid.arrange(plot1, plot2, ncol=2)

As part of EDA we will be covering 1. Variation, 2. Missing Values 3. Covariation

library(nycflights13) library(arm) library(GGally) library(tidyverse) library(lvplot) library(ggstance)

Variation

Q1 Explore the distribution of each of the x, y, and z variables in diamonds. What do you learn? Think about a diamond and how you might decide which dimension is the length, width, and depth.

diamonds %>% gather(key = dist, vals, x, y, z) %>% ggplot(aes(vals, colour = dist)) + geom_freqpoly(bins = 100)

Distribution of x and y is almost same , as in this graph we can see that it it overlap with X variable.

Q2. How many diamonds are 0.99 carat? How many are 1 carat? What do you think is the cause of the difference?

diamonds %>% filter(carat %in% c(0.99, 1)) %>% count(carat)

It could be that 0.99 is repeated 23 times.

Q3. Compare and contrast coord_cartesian() vs xlim() or ylim() when zooming in on a histogram. What happens if you leave binwidth unset? What happens if you try and zoom so only half a bar shows?

diamonds %>% ggplot(aes(y)) + geom_histogram() + coord_cartesian(ylim = c(0, 50)) #xlim deleted the observations at 0.

diamonds %>% ggplot(aes(y)) + geom_histogram() + xlim(c(0, 60)) + coord_cartesian(y = c(0, 50)) #xlim and ylim inside coord_cartesian don't exclude the data

diamonds %>% ggplot(aes(y)) + geom_histogram(bins = 30) + coord_cartesian(xlim = c(2, 60), ylim = c(0, 50))

Missing value

Q1. What happens to missing values in a histogram? What happens to missing values in a bar chart? Why is there a difference?

diamonds %>% ggplot(aes(price)) + geom_histogram(bins = 1000) # In a histogram, missing values leave a gap in the distribution, as in the gap in the above histogram of price and In the barplot, the function removes the `NA` value.

Q2. What does na.rm = TRUE do in mean() and sum()? Ans: It removes the `NA` from the calculations.

Covariation

Covariation describes the behavior between variables.

1. Use what you've learned to improve the visualisation of the departure times of cancelled vs. non-cancelled flights.

flight <- flights %>% mutate(cancelled = is.na(dep_time), sched_hour = sched_dep_time %/% 100, sched_min = sched_dep_time %% 100, sched_dep_time = sched_hour + sched_min / 60)

flight %>% ggplot(aes(sched_dep_time, colour = cancelled)) + geom_density()

flight %>% ggplot(aes(cancelled, sched_dep_time)) + geom_boxplot()

flight %>% ggplot(aes(sched_dep_time, ..density.., colour = cancelled)) + geom_freqpoly(binwidth = 1/2)

2. What variable in the diamonds dataset is most important for predicting the price of a diamond? How is that variable correlated with cut? Why does the combination of those two relationships lead to lower quality diamonds being more expensive?

Ans: display(lm(price ~ ., diamonds), detail = T)

3. Install the ggstance package, and create a horizontal boxplot. How does this compare to using coord_flip()?

library(ggstance) diamonds %>% ggplot(aes(cut, carat)) + geom_boxplot() + coord_flip()

4. One problem with boxplots is that they were developed in an era of much smaller datasets and tend to display a prohibitively large number of "outlying values". One approach to remedy this problem is the letter value plot. Install the lvplot package, and try using geom_lv() to display the distribution of price vs cut. What do you learn? How do you interpret the plots?

p <- ggplot(diamonds, aes(cut, price, colour = ..LV..)) p + geom_lv() p <- ggplot(diamonds, aes(cut, carat, fill = ..LV..)) p + geom_lv()

This plot is useful for having a more detailed description of the tails in a distribution.

5. Compare and contrast geom_violin() with a facetted geom_histogram(), or a coloured geom_freqpoly(). What are the pros and cons of each method?

diamonds %>% ggplot(aes(cut, price)) + geom_violin() diamonds %>% ggplot(aes(price)) + geom_histogram() + facet_wrap(~ cut, scale = "free_y", nrow = 1) diamonds %>% ggplot(aes(price)) + geom_freqpoly(aes(colour = cut))