390R EDA1 notes
10/29/2020

EDA is what questions you can ask from your data. You need to ask questions about the data for analysis. Usually at the beginning of analysis it's hard to have any questions of your data. 2 main question you can loosely word these questions as:
1.What type of variation occurs within my variables?
2.What type of covariation occurs between my variables?

To show variation in variable, histogram for continuous bar charts for categorical
Using diamonds dataset, lets explore distributions of variables
Cut variable (Categorical): ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut))
Carat variable (Continuous)
ggplot(data = diamonds) +
 geom_histogram(mapping = aes(x = carat), binwidth = 0.5)

You should always explore a variety of binwidths when working with histograms, as different bin widths can reveal different patterns. For example, here is how the graph above looks when we zoom into just the diamonds with a size of less than three carats and choose a smaller binwidth.

```
diamonds %>%  filter(carat < 3) %>%
  ggplot( mapping = aes(x = carat)) +
  geom_histogram(binwidth = 0.1)
```

Using smaller bin widths can make seeing patterns in the data easier.

If you want to overlay multiple histograms in one plot, you can use geom_freqpoly which is the same as geom_histogram but uses lines instead of bars.

Once we get a histogram/barchart we want to summarize it to find patterns. We want to explain why this happens. Usually we can explore the relationship with other variables as well.

In the diamonds carat histogram example we can calculate the average price of the diamonds for each value of the carat. It clearly shows an increasing pattern. However, at some points there are 'missing' spots that separate different relationships. Let's do some modelling to display this.
Create a new object
```
Diamonds_prince <- diamonds %>%  filter(carat < 3) %>%
  group_by(carat)%>% summarize(price_avg=mean(price))
```

```
ggplot(diamonds_price)+geom_point(aes(x=carat,y=price_avg))+
geom_smooth(data=filter(diamonds_price,carat<1),aes(x=carats,y=price_avg),method="lm")
```
that will create a linear model for the data points 0,1

We can copy this for the next values of data.
ggplot(diamonds_price)+geom_point(aes(x=carat,y=price_avg))+
geom_smooth(data=filter(diamonds_price,carat<1),aes(x=carats,y=price_avg),method="lm")+
geom_smooth(data=filter(diamonds_price,carat>=1,carat<1.5),aes(x=carats,y=price_avg),method="lm")+
geom_smooth(data=filter(diamonds_price,carat>=1.5,carat<2),aes(x=carats,y=price_avg),method="lm")+
geom_smooth(data=filter(diamonds_price,carat>2),aes(x=carats,y=price_avg),method="lm")
Now we have 4 different linear models for this data and we can analyze the data more intensely.

Looking at the eruptions data in faithful dataset
ggplot(faithful)+geom_histogram(aes(x=eruptions),binwidth=0.25)
We can see two means, a smaller one at 2 and a larger one at 4.5. To analyze the data we could say there are two different relationships as there aren't many values between the two.

Unusual values
Outliers are observations that are unusual; data points that do not seem to fit the pattern. Sometimes outliers are data entry errors; other times outliers suggest important new science. When you have a lot of data, outliers are sometimes difficult to see in a histogram. For example, take the distribution of the y variable from the diamonds dataset. The only evidence of outliers is the unusually wide limits on the x-axis.
If we want to zoom in on a plot to ignore the outliers, we can use ylim to set limits on what y values are shown.
ggplot(diamonds)+geom_histogram(aes(x=y),binwidth=0.5)+coord_cartesian(ylim=c(0.50))

Question: 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.
ggplot(diamonds)+geom_histogram(aes(x=x),binwidth=0.1) shows most obs. Between 4 and 9
ggplot(diamonds)+geom_histogram(aes(x=z),binwidth=0.01) all obs. Between 2 and 7
ggplot(diamonds)+geom_histogram(aes(x=y),binwidth=0.01)
We can look at how all the observations compare to each other.
diamonds%>%summarize(mean(x>y),mean(x>z),mean(y>z))
Outputs: 0.434, 1.00, 1.00
So we learn that most x values are smaller than y, and all x and y values are greater than z values.

Question: Explore the distribution of price. Do you discover anything unusual or surprising? The last digits of prices are often not uniformly distributed. Plot the distribution of the last one and two digits of prices.
ggplot(filter(diamonds,price<2500))+geom_histogram(aes(x=price),binwidth=10)
There are no diamonds priced between 1450 to 1550.

diamonds%>%mutate(ending =price %%100)%>% ggplot()+geom_bar(aes(x=ending))
diamonds%>%mutate(ending =price %%10)%>% ggplot()+geom_bar(aes(x=ending))
Both are uniformly distributed.
diamonds%>% mutate(ending =price %%100) %>% group_by(ending) %>%count() %>%
arrange(desc(n))%>% print(n=Inf)
This lists the counts for every 2 digit combination. The print(n=Inf) will display all 100 results instead of just the first 10.