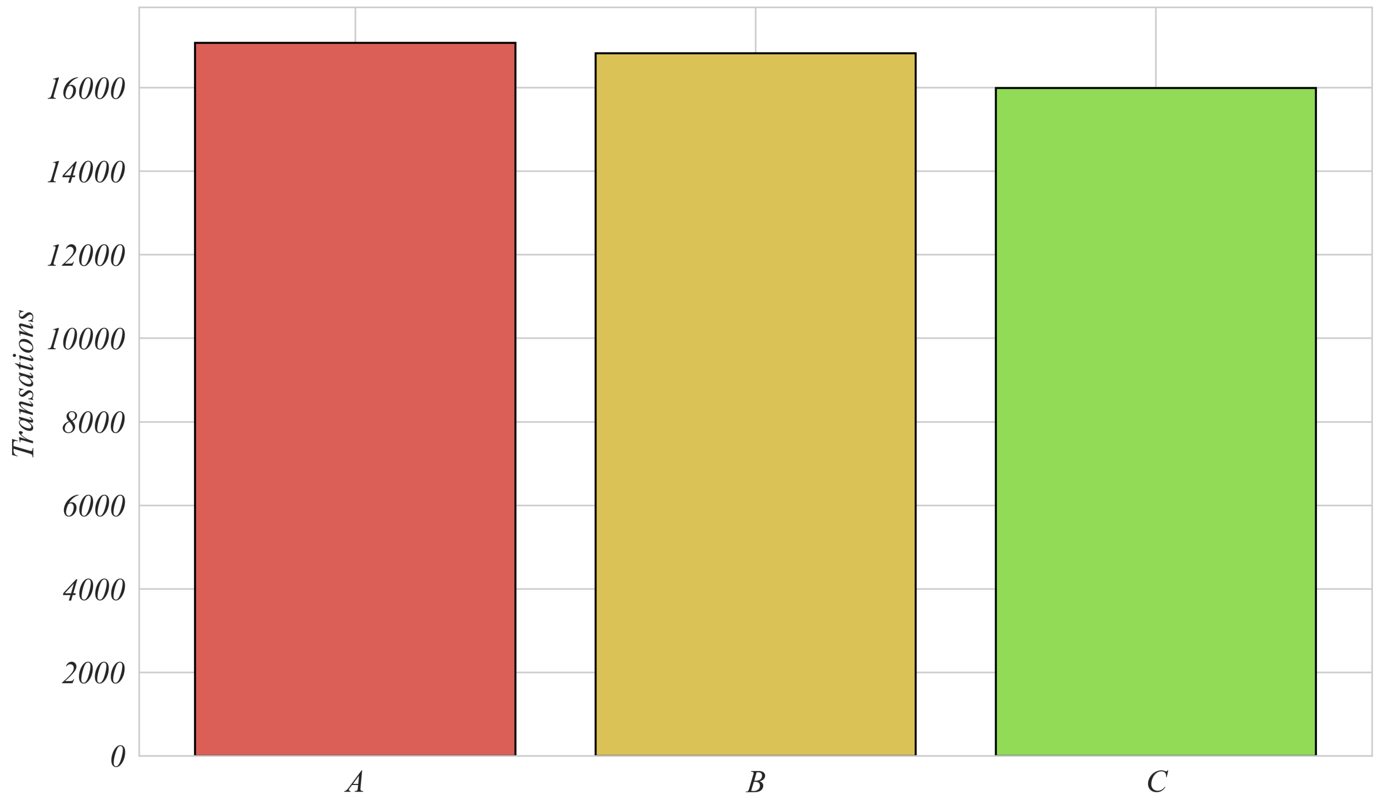
Part 1.3 | Comparing Continous Variables by Category

The Busiest Shop

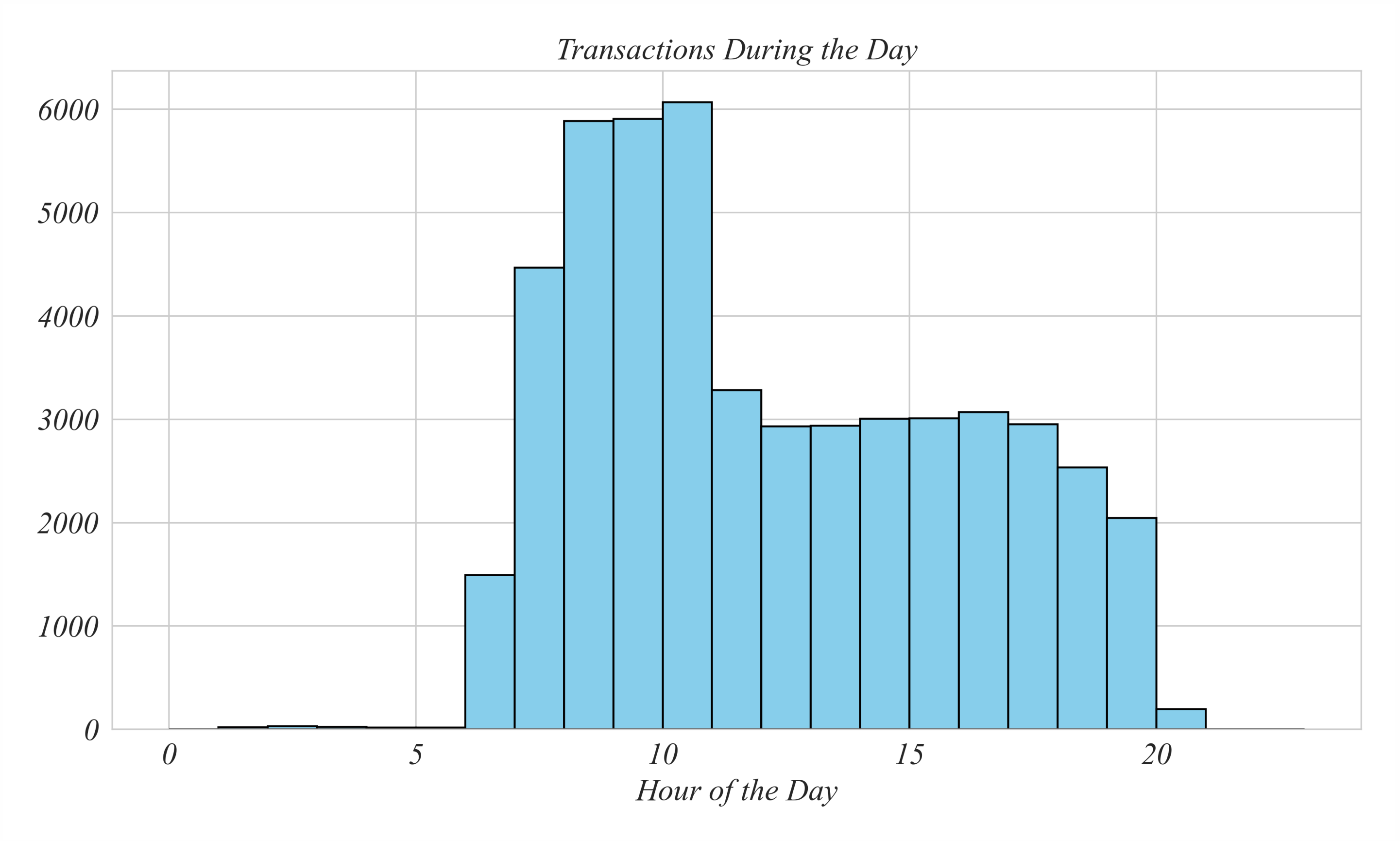
The owners of a chain of coffee shops is trying to respond to their staff feeling stressed and overworked by hiring one more barista, and are trying to use transaction data to decide where to assign the new staffmember. Lets start by visualizing the total number of transactions at each shop. What kind of visualization would be most useful?

Well since this is categorical data, a histogram won't work. This isn't a great application for pie charts, since we'd like to understand how transations are spread across the three locations, not comparing one shop to the whole. A bar chart is a great place to start.



The Busiest Time

Simply looking at which shop is busiest, Shop A might benefit most from the extra hire. But since demand for coffee is not constant throughout the day, it might be most helpful to examine when demand spikes the most during the day. Looking at variation over time like this is a great application for a histogram to visualize transations, and since time is a continuous variable we can't use a pie chart or bar chart. Lets start by combining transations together across all three shops and group into equally sized bins.

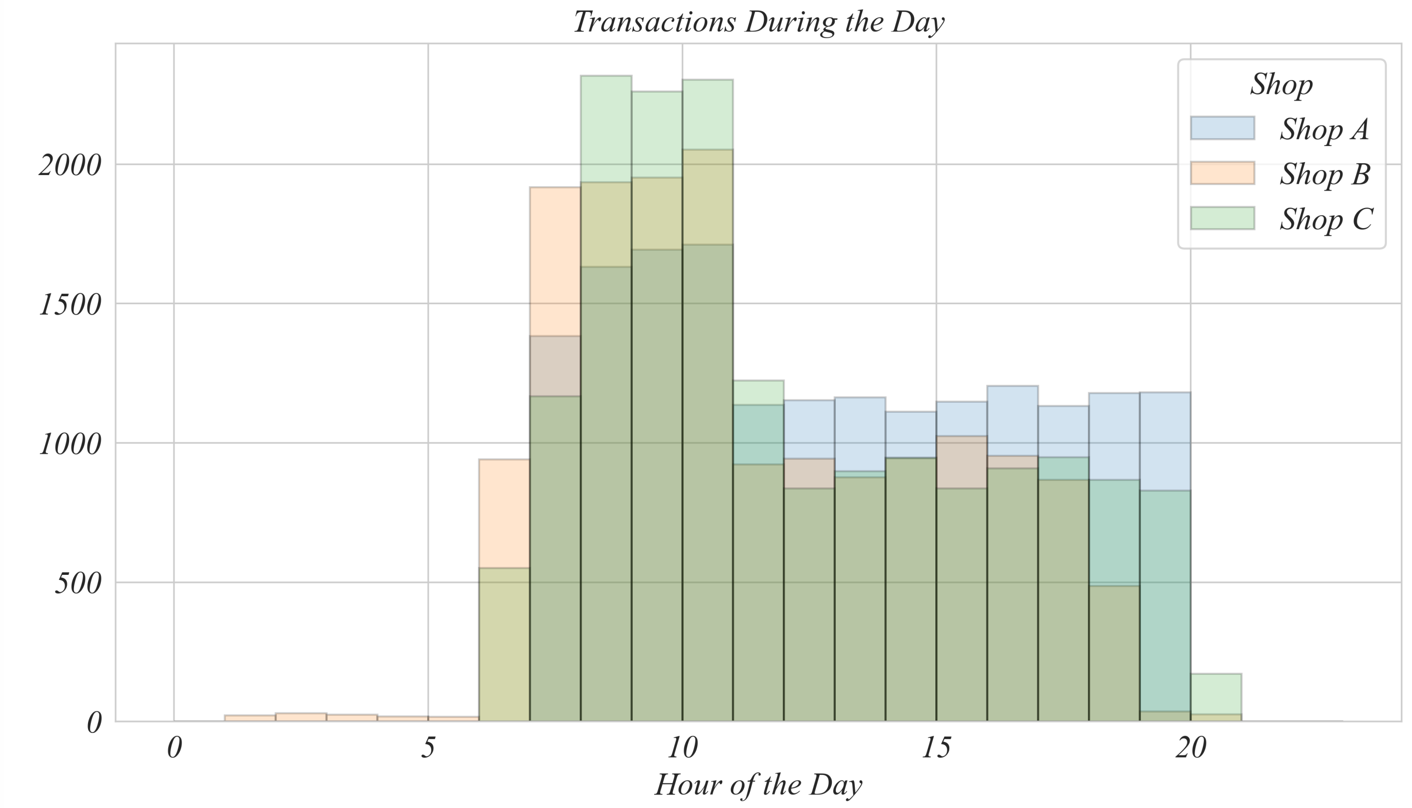


This type of data can tell us which time of day we should be hiring for. From this it looks like the morning shift is most busy, so it might be a good time of day to add another barista.

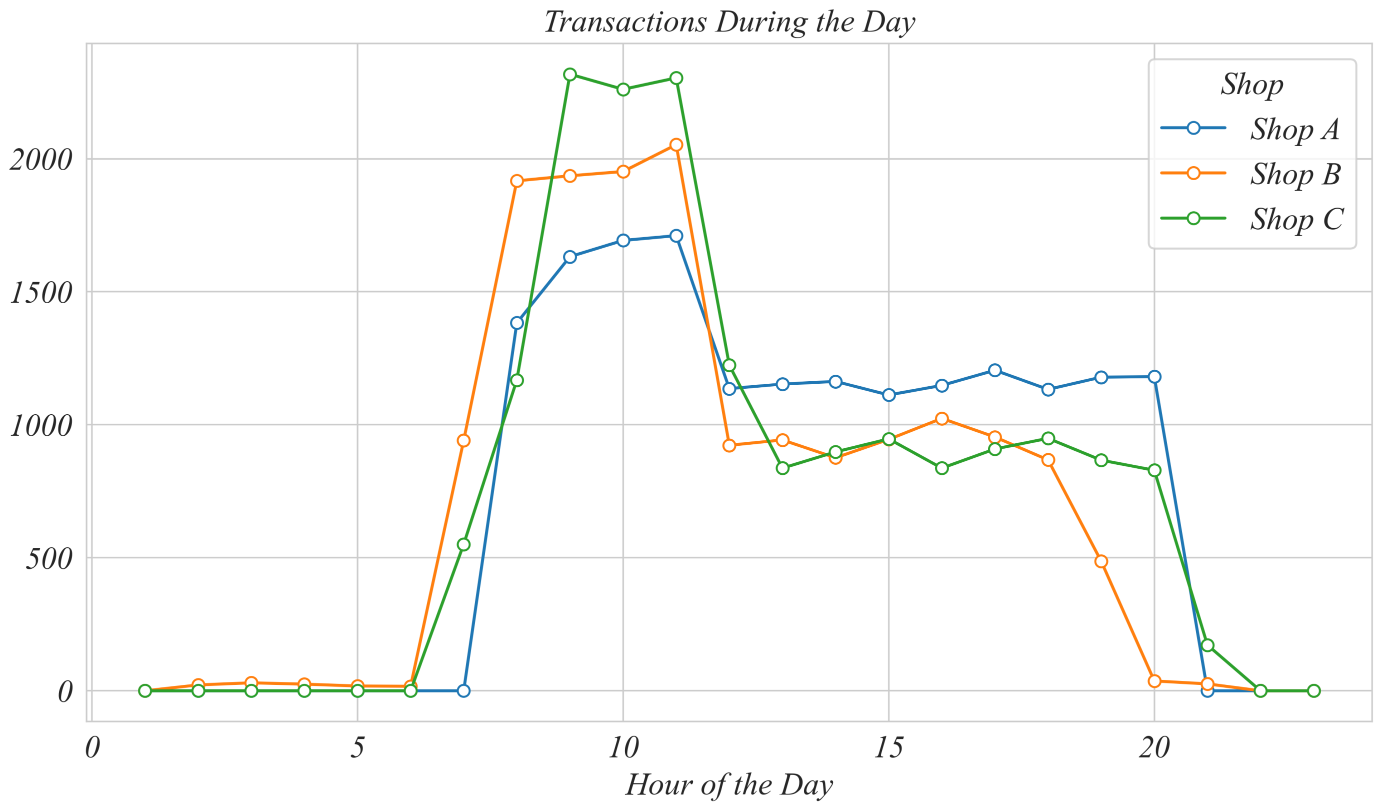
The Busiest Shift

So far we've found that Shop A is the busiest shop and that mornings are the busiest time. Does this mean we should hire an extra barista for the morning shift at Shop A?

It could be, but we don't yet know enough. To learn a little more, we could separate out demand throughout the day separately by shop. Then we can plot all three shops on the same graph.



What do you think of this view? To me it looks a bit hard to read. It's hard to tell where one bar starts and another one ends. Staking bars like this works sometimes but is typically too busy. Fortunately, there's another visualization tool, the **line graph**, which connects the top of each bar.



A line graph is much easier to compare across shops by through time. The morning shift at Shop C has the highest peak sales. Many things could explain this, such as higher demand at Shop C, faster baristas at Shop C, simpler orders at Shop C, or many other reasons. The data available can tell us a lot about where hiring a new barista might be most helpful, but data on it's own often cannot paint a complete picture.

Summary

- Categorical variables and continuous variables can give us different views of the same data.
- Often we can visualize both views on the same graph using visualization techniques for continuous variables within the category.
- Line graphs help simplify the visualization of multiple categories.

Excel Exercise

Lets build some similar figures with this dataset in Excel. We've already worked with bar charts and histograms so I'll include them in the example spreadsheet but won't go into detail on how to construct them. Lets focus on creating a line graph for all sales using the FREQUENCY command and a line graph with the three shops separately using a **Pivot Table**.

First, lets use the FREQUENCY command to create a line graph. Create a new sheet by clicking the "+" button at the bottom. You can rename it "sales_frequency" if you'd like. We need to define the bins the FREQUENCY command will use. Create a column label "BINS" in the top cell (eg. A1). Then fill in the column with bin edges starting with 1:00:00 (a format represeting the hour:minute:section) and ending with 23:00:00. There's a little drag trick you might be able to figure out. :) Then in the next column, create a header "FREQUENCY" and enter the following command in the cell below.

```
=FREQUENCY(Coffee_Sales_Reciepts!C:C,A2:A25)
```

This command is a little more involved, so lets walk through each part. The command "FREQUENCY" takes in a data range and bins it according to the bins it takes in. We have the data range for reciepts in the first sheet in column C. So to tell the function where to look for the data, we have to tell it to look in column C (selecting all values in the column by "C:C") in the sheet labeled "Coffee_Sales_Reciepts" (selected by "Coffee_Sales_Reciepts!"). The Excel syntax for this is the following:

```
Coffee_Sales_Reciepts!C:C
```

To tell the function where to look for the bins, we simply add the bin range in second entry of the function as the following:

```
A2:A25
```

One last note here. You're more than welcome to create the frequency data in the first sheet to skip the need to reference another sheet, creating a second sheet like we've done is best practice for keeping your work readable (both for you at a later time and for others).

Once you've hit enter, you should see the frequencies of transactions show up next to the hour in which they occured. To plot this data, select both columns, go to the Insert tab, and click around until you find the 2-d line graph. This will give you a nice place to start making the graph look nice.

This approach is nice when we have a considerable amount of data bins and only one category. But in our example, we would like to know the frequency of transactions by shop. One approach would be to separate the data by shop and perform the approach we just used on all three shop's data. A less involved approach uses a **pivot table**.

A pivot table is a powerful tool to summarize and change the structure of data. Start by selecting the two columns that contain the data we need: "sales_outlet_id" and "transaction_time". Then click on the tab Insert, then pivot table. This will bring up a menu with the highlighted data as the input range. Make sure the option to create the output in a new sheet, which we'll later label "pivot_table".

Once you hit enter, you'll be taken to the new sheet with the "PivotTable Fields" right side bar. There's a lot going on here. For now, what we need to do is summarize the data by shop: drag the "sales_outlet_id" field name from "FIELD NAME" into the "Columns" box. This organizes the data by shop. Then we can bin the transactions by hour: drag the "transaction_time" field name into the "Rows" box and make sure Minutes and Seconds are not selected. This should produce a frequency table for each shop and a grand total. To plot this on one line graph, select the data and one row of headers, click into Insert, then line graph, then select the 2-d line graph. From here, you can make the graph look as pretty as you'd like. :)