1 newpage

2 Data Science: Bridging Principles and Practice

2.1 Part 6: Preparing to Model (Bike Sharing case study)

2.2 6b. Exploratory Data Analysis (EDA)

QUESTION: Which of the variables are numerical and which are categorical? Intuitively, which do you think would be useful for predicting the number of riders on a given day? Would you choose different variables depending on if you wanted to predict casual versus registered rider counts?

ANSWER:

- numerical: temp, felt temp, humidity, windspeed, casual, registered, total riders
- categorical: is holiday, is work day, season, all of the date variables, weather

Note that weather can be confusing, since the values are numbers. In fact, these numbers represent different categories of weather (e.g. cloudy, sunny). While the categories can be ordered by severity, they are not really a measurement like a numerical variable would be.

Answers will vary here. Intuitively, we might expect that fewer riders would ride on days with a very cold temperature or very bad weather. Weather conditions also might affect casual riders more than registered, since we might assume that registered riders are more likely to be commuters who cannot put off a bike ride due to inclement weather. We also might expect that more casual riders would ride on weekends or holidays, while the number of registered riders may be higher on workdays if they use their bikes to commute.

2.2.1 Summary Statistics

QUESTION: Looking at these statistics as data scientists, we're interested in a few things in particular: - are there any values missing (e.g. days for which some data was not collected)? - what ranges of values does each variable take? - are there any extreme values that might throw off our analysis?

ANSWER: - no missing values: all counts are 731 - temp, felt temp, humidity, and windspeed all range between 0 and 1 (which makes sense, because according to the data dictionary they were normalized to fall within that scale. Ridership tends to number in the thousands, although casual ridership counts are a bit lower. - not at first glance. The means and medians (50th percentiles) of the variables are all very similar, indicating that they aren't very skewed. There are some very small and very large values for rider counts (e.g. 2 is the minimum number of casual riders). Visualization will help us better understand how the data are distributed.

2.2.2 Visualization Continued: Numerical Data and Widgets

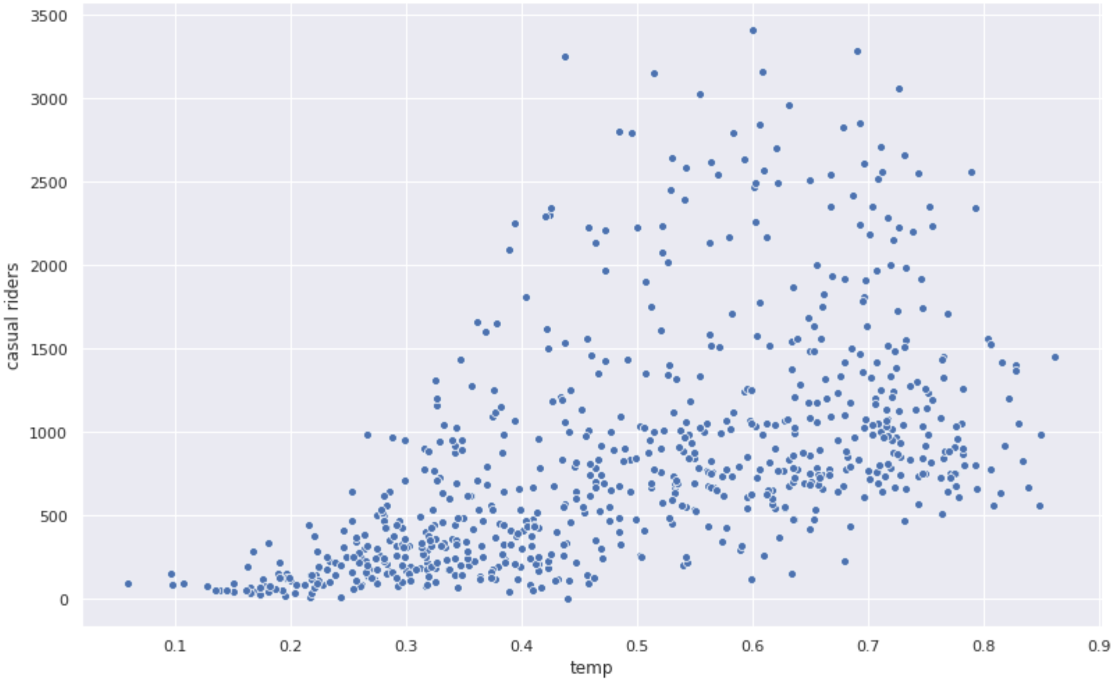
QUESTION: Describe the distributions of the different variables. Are they normally distributed? Are any of them skewed (that is, do any of them have most of their values to the left or right of the histogram's center)? What values do each of them take on?

ANSWER: - temp and felt temp: very similar in shape, with most of the data concentrated in the middle of the distribution and a few rough peaks in the center. All values fall between 0 and 1 - humidity: almost normally distributed, with a slight left skew. All values fall between 0 and 1 - windspeed: almost normally distributed, with a slight right skew. All values fall between 0 and 1

EXERCISE: Try plotting at least one numerical explanatory variable (temp, felt temp, windspeed, or humidity) against a response variable (casual, registered, or total riders). What would you say about the relationship between the two variables based on the scatter plot?

Note: answers will vary here depending on the variables you picked.

```
In [5]: # Fill in the ellipses with your code
        sns.scatterplot(x="temp", y="casual riders",
                        data=bikes);
```



QUESTION: Based on the scatter plots, which variables appear to be linearly correlated with rider counts? Which variables appear to be non-linearly correlated or uncorrelated? Is the apparent correlation different for casual or registered riders?

ANSWER:

Date appears to be non-linearly correlated: the scatter plot follows a "wave" pattern that goes up in summer months and down in winter months for casual and registered riders.

Temp and felt time may have a slight linear correlation: the scatter plots seem to go up slightly as temperature goes up.

Windspeed and humidity may not be correlated- their scatter plot clouds look mostly shapeless.

QUESTION:Many of our categorical variables are related to time (e.g. week day, month, etc). How do usage patterns over time differ for registered and casual users? In what categories do the different user types act similarly?

ANSWER: Ridership (both casual and registered) does seem to go up during warmer months and seasons and fall in cooler times. Casual and registered riders also act similarly in relation to the weather- the worse the weather is, the fewer people ride.

Casual and registered riders tend to act oppositely in relation to the type of work day. Casual ridership goes up on holidays and weekends, while registered ridership goes up on weekdays and non-holidays.

2.3 6c.The Test-Train Split

QUESTION: Data are often expensive to collect, and having a good predictive model can be the difference between success and ruin. Given these factors, the decision of how much data to set aside for testing and validation is very much a matter of opinion.

What are some reasons for putting a larger portion of data into the training set? What are some downsides? Think about monetary and computational costs, as well as the potential risks of having a model that makes inaccurate predictions.

ANSWER: Putting more data in the training set might make sense, because we want our model to generalize as well as possible to data it hasn't been trained on. If the model doesn't see enough examples in its training data, it will be biased towards the few it did see and may give inaccurate predictions. Also, given that data collection is expensive, it can be tempting to put as much of the data we collect as possible into the model, rather than using it to test the model.

However, putting too much data in the training set (and too little in testing and validation sets) can mean that we won't have a measure of how well our model generalizes to data it hasn't been trained on. This can have major consequences- if a model performs very well on training data and we have no testing data, using that model in decision making can have high costs (e.g. not allocating enough bikes in our bike-sharing venture, or buying too many bikes) or huge consequences (e.g. misdiagnosing a patient in a predictive healthcare model). Models with large training sets also may take a lot more time to train or make predictions. For very large data sets, this could be a day or more.