Joshua Grant

4/18/2024

DSCI 230

**DSCI 230 Final Project**

For this project, I was contacted by a mountain biking company in Bentonville, AR, where there are world-class mountain biking trails. The company has customers who register to go biking before arriving, and customers who walk in without a reservation. For this project, there are **registered** and **causal** users, with registered being those who made reservations and casual users who did not.

For my data analysis, I aimed to answer a few questions, including:

* If weather affects one group of customers more than another
* If there are thresholds in which a group may decide not to come (snow, rain, high temps, etc.)
* Explore a relationship between weather attributes and registered/casual customers

To answer some of these questions, I used a variety of methods. I utilized tools provided in python, such as pandas, NumPy, matplotlib, seaborn, and scikit-learn. After cleaning the data, I performed various operations on the dataset to see what I could uncover. These operations included:

1. Sorting the dataset by registered users to see what conditions brought the most registered customers
2. Computing a boolean array to see which rows have a temperature of at least 200
3. Using join operations to merge two subsets of the data together
4. Using GroupBy methods to view total viewers by month and average users per weekday
5. Slice the dataset to look at a certain month with DateTime
6. Create a Period Index for quarterly analysis
7. Find the distance between a date in the highest attended month vs the lowest attended month

I found some interesting notes here. I found that August was the month that brought the most registered customers, while July brought the most casual customers to the company. I also found that Tuesday, surprisingly was the day of the week that, on average, brought the most registered customers. Friday brought the most casual customers, however. This makes sense, as casual walk-in customers are much more likely to have free time and go out to go mountain biking on the weekends.

After this, I started creating some visualizations to observe the relationship between different weather conditions and customer attendance levels. The first plot I created was a line plot just looking at the relationship between registered and casual customer attendance over the year:

A graph of a graph of a customer

Description automatically generated with medium confidence

Now, keep in mind that throughout all these visualizations, that we have significantly more registered customers than we do casual customers. This plot shows exactly that. To interpret this graph correctly, we need to look at the shape of the line plot. We can see that the two lines move pretty consistently with each other, meaning that the time of year doesn't trend to more registered customers visiting compared to casual, or vice versa. We can also see that the company gets their most customers in the summer and fall months, which isn’t surprising due to warmer weather.

Next, we’ll look at two different bar plots. The vertical bar plot will view the number of customers in relation to wind speed (measured in meters per second), and the horizontal bar plot will view the number of customers in relation to the amount of precipitation:

A graph of a number of customers

Description automatically generatedA graph with numbers and a bar

Description automatically generated with medium confidence

We can see here that the relationship between registered and casual users is essentially the same in both graphs. However, there are other interesting observations to be made here. Interestingly, there’s a huge spike in attendance when the wind is exactly 30 m/s. This is likely either a strange coincidence or we could have a significant amount of data points with a wind speed of 30 m/s compared to all other wind speeds. The most interesting observation here is how attendance absolutely drops off a cliff when there is essentially any rain. This isn’t really surprising because no one wants to bike in the rain, but it’s still an interesting observation to point out.

Next, we’ll look at two different scatter plots fitted with regression lines. Both are overlaid with both groups we’re interested in, casual and registered users. The first one looks at attendance correlated with higher temperatures, while the other looks at attendance correlated with lower temperatures:

A chart of blue and orange dots

Description automatically generatedA graph of blue dots

Description automatically generated

We can see here that these two graphs have an inverse relationship, which is exactly what we’d expect to see. They both follow what we observed earlier in our line plot that higher attendance happens in the warmer months, while the winter months see the least action. It’s also important to remember what we’re analyzing here. At first glance, it may look like there’s some sort of significant difference between registered and casual users because the lines are on such different trajectories, but we must recognize what the y-axis is: number of customers. The reason the slope of the regression lines differ so much is because, as stated earlier, the number of registered customers is greater than that of the casual customers. So with that said, it’s safe to assume that while temperature clearly has an effect on overall attendance, there’s no difference in attendance between the two groups.

The final visualization we will look at is a box plot that splits the dates into the four seasons, and then views customer attendance in relation to the season:

A chart with different colored squares

Description automatically generated

There's definitely some interesting observations to be made here. First, we can view that the median line (black line in the blue box) in the winter box isn't close to the middle of the box, meaning our data occurring in winter months either has a lot of variation in terms of customer attendance. We can also see that the box plot for the summer box is pretty small, meaning the data is clustered around the median. This is a sign of very reliable data. Looking at the big picture, summer and fall months attract the most customers, with spring not far behind, while winter is firmly in last. It's also worth noting that we've got some outliers in the summer and fall boxes. It’s interesting how we have the most outliers in the summer box, even though it’s our tightest grouping of data. The spring box likely has such a large area because, in Arkansas, spring months see the most variation in temperature. They get more winter temperatures than the fall does, however it also warms up significantly towards the end of spring, which is probably why this box ranges so widely.

The final part of our data analysis was building a model that used weather attributes provided in the dataset, and tried to predict whether a potential customer would go biking that day or not.

Since we didn’t have a direct observation of whether a potential customer did or did not go biking, we had to come up with a creative method ourselves. What I decided to do was get the mean and standard deviation values for each weather attribute in our dataset. I then created a threshold value that represents what is within 1 standard deviation of the mean. If the data point is within 1 standard deviation, it passes. Otherwise, it didn’t pass. We then compared all of our attributes together with this threshold, and if it passed for all 4 of our weather attributes, we assumed that a potential customer would attend.

After finally obtaining what would be our ‘predicted’ value, we used linear regression to build the model. We split our data into training and testing splits, fit the model, used the model to predict our y-values, and finally, scored the model. We ended up achieving about a 75% accuracy rating for this model, which means that it can predict with 75% accuracy if a potential customer will go biking that day or not. This could potentially be very useful to the bike store owner. For example, if he knew weather would be rough one week, he could apply this model to adjust prices to be lower or adjust schedules to give some employees some extra time off.

To recap this data analysis, we found that no weather attribute seemed to affect one group of customers more than another. We also found that the store experiences significantly more customer traffic in warm weather than any other environment. We also developed a model that is capable of predicting if potential customers are likely to go mountain biking on a given day depending on the weather that day. This model, along with our visualizations and data aggregation results, could be utilized by the company to make more cost-effective business decisions and ultimately lead to higher profits. That concludes my DSCI 230 final project. I hope you enjoyed and learned as much as I did!