Exploratory Data Analysis Summary

There were 3 categorical variables in the data: Name, Region, and State. The rest were numerical, including the ones created during EDA. The notebook looks at some variables grouped by state and then scales the data so that PCA can be performed. PCA, or principal component analysis, looks at the different features of the data in a linear fashion and shows how much variance they explain in the target variable. In this case, our target variable is the price of an adult ticket on the weekend. The scaled data is checked for mean and standard deviation, which reveals that the std function has a bias. This bias is removed with the argument ddos=0.

The seaborn plot doesn’t make sense. There is clearly a densely populated area of the graph and the notebook says there’s no pattern. Montana is not in the densely populated area, so I don’t think there’s a case for Montana to be compared with other states, except in the case of low data counts. I don’t agree that the PCA plot shows ample evidence for comparing Montana to other states, but there aren’t that many data points to begin with. For that reason, Montana has to be compared to other states. There was a pattern between state and ticket price, but low data counts make it so there is no other choice but to keep all the states in the data.

In the heatmap, resort\_night\_skiing\_state\_ratio is the most correlated with ticket price. This supports my hypothesis that night skiing would be a welcome expansion to the resort. Some others that have positive correlations with ticket price include: fastQuads, Runs, and SnowMaking\_ac. Resorts per 100k capita can be quite varied with ticket price and is something to watch out for. The target feature for modelling is still the ticket price, but resort\_night\_skiing\_state\_ratio will be watched closely for patterns. If the modeling shows the state name is too complex to show, then a separate, sorted data frame with shared index of the state number will do. That way, each state is represented by a number of where it comes in the list alphabetically. Precautions must be taken to avoid line plots with this data, due to the arbitrary “number” on the state.