Summary

Write a summary statement that highlights the key processes and findings from this notebook. This should include information such as the **original number of rows in the data**, **whether our own resort was actually present** etc. What **columns, if any, have been removed**? **Any rows**? Summarise the reasons **why**. Were any **other issues found**? What **remedial actions** did you take? **State where you are in the project**. Can you **confirm what the target feature** is **for your desire to predict ticket price**? How many **rows were left in the data**? Hint: this is a great opportunity to reread your notebook, check all cells have been executed in order and from a "blank slate" (restarting the kernel will do this), and that your workflow makes sense and follows a logical pattern. As you do this you can pull out salient information for inclusion in this summary. Thus, this section will provide an important **overview of "what" and "why"** without having to dive into the "how" or any unproductive or inconclusive steps along the way.

Data loaded and displayed to view the records and audited.

Found there are 2 kinds of ticket price (AdultWeekday and AdultWeekend) and other columns with potential features

Raise a question what quantity will I want to model? Which ticket price?

Run ski data and saw fastEight with NaN

After running: Missing values by columns sorted by COUNT AND %

fastEightover 50% with missing values

AdultWeekday 16%with missing values

AdultWeekend 15% with missing values

Big Mountain Resort with no missing values

AdultWeekday price of 81 and AdultWeekend price of 81

Since I am looking for desired target quantity Question is? What overlap is there in these missing values?

Investigate further

Examined numeric features, look into categorical features like resort name, region, state

Select columns dtaypes are object

330 rows and 3 columns so ALL DATA

Region, name, state with no missing values

What? Is name unique? Is region always same as state?

Found out the are 2 Crystal Mountain resorts : SAME NAME, BUT DIFFERENT, Region and State

Above resort is unique in itself, onr iin Michigan and other in Washington

Unique value on each row

Find out the relationship between region and state??

How many cases where Region and state did not have the same answer or differ?

Differ 33 cases

The vast majority of differences are in California(state) and most region are being called Sierra Nevada northern California

Because a few states are split across multiple named regions, there are slightly more unique regions than states.

see New York accounting for the majority of resorts

target resort is in Montana, which comes in at 13th place

problem task includes the contextual insight that the data are for resorts all belonging to the same market share THEREBY INCLUDE all

expect prices to be similar amongst them

A boxplot grouped by State is an ideal way to quickly compare prices

The best approach here definitely would include consulting with the client or other domain experts

They might know of good reasons for treating states equivalently or differently. The data scientist is rarely the final arbiter of such a decision. But here, you'll see if we can find any supporting evidence for treating states the same or differently.

Our primary focus is our Big Mountain resort, in Montana. Does the state give you any clues to help decide what your **primary target response feature should be (weekend or weekday ticket prices)?**

**Average weekend and weekday price by state**



The figure above represents a dataframe with two columns, one for the average prices of each kind of ticket. This tells you how the 

The figure above represents a dataframe with two columns, one for the average **prices** of each kind of ticket. This tells you how the average ticket price varies from state to state. But can you get more insight into the difference in the distributions between states?

Overall tells me that majority of the states weekend prices are higher than the weekday prices are

Aside from some relatively expensive ticket prices in California, Colorado, and Utah, most prices appear to lie in a broad band from around 25 to over 100 dollars. Some States show more variability than others. Montana and South Dakota, for example, both show fairly small variability as well as matching weekend and weekday ticket prices. Nevada and Utah, on the other hand, show the most range in prices. Some States, notably North Carolina and Virginia, have weekend prices far higher than weekday prices. You could be inspired from this exploration to consider a few potential groupings of resorts, those with low spread, those with lower averages, and those that charge a premium for weekend tickets. However, you're told that you are **taking all resorts to be part of the same market share**, you could argue against further segment the resorts. Nevertheless, ways to consider using the State information in your modelling include:

* disregard State completely
* retain all State information
* retain State in the form of Montana vs not Montana, as our target resort is in Montana

You've also noted anothe**r effect above: some States show a marked difference between weekday and weekend ticket prices**. It **may make sense to allow a model to take into account not just State but also weekend vs weekday.**

Thus we currently have two main questions you want to resolve:

* What do you do about the two types of ticket price?
* What do you do about the state information?

Having decided to reserve judgement on how exactly you utilize the State, turn your attention to cleaning the numeric features.

#### Numeric data summary

**Recall you're missing the ticket prices for some 16% of resorts. This is a fundamental problem that means you simply lack the required data for those resorts and will have to drop those records. But you may have a weekend price and not a weekday price, or vice versa. You want to keep any price you have.**

Just over 82% of resorts have no missing ticket price, 3% are missing one value, and 14% are missing both. You will definitely want to drop the records for which you have no price information, however you will not do so just yet. There may still be useful information about the distributions of other features in that 14% of the data.

Note that, although we are still in the 'data wrangling and cleaning' phase rather than exploratory data analysis, **looking at distributions of features is immensely useful in getting a feel for whether the values look sensible and whether there are any obvious outliers to investigate.** Some exploratory data analysis belongs here, and data wrangling will inevitably occur later on. It's more a matter of emphasis. Here, we're interesting in focusing on whether distributions look plausible or wrong. Later on, we're more interested in relationships and patterns.

What features do we have possible cause for concern about and why?

* SkiableTerrain\_ac because values are clustered down the low end,
* Snow Making\_ac for the same reason,
* fastEight because all but one value is 0 so it has very little variance, and half the values are missing,
* fastSixes raises an amber flag; it has more variability, but still mostly 0,
* trams also may get an amber flag for the same reason,
* **yearsOpen because most values are low but it has a maximum of 2019, which strongly suggests someone recorded calendar year rather than numbe**r of years.

Silverton Mountain resort has an incredibly large skiable terrain area.

But what can you do when you have one record that seems highly suspicious?

You can see if your data are correct. Search for "silverton mountain skiable area". If you do this, you get some [useful information](https://www.google.com/search?q=silverton+mountain+skiable+area).

You can spot check data. You see your top and base elevation values agree, but the skiable area is very different. Your suspect value is 26819, but the value you've just looked up is 1819. The last three digits agree. This sort of error could have occured in transmission or some editing or transcription stage. You could plausibly replace the suspect value with the one you've just obtained. Another cautionary note to make here is that although you're doing this in order to progress with your analysis, this is most definitely an issue that should have been raised and fed back to the client or data originator as a query. You should view this "data correction" step as a means to continue (documenting it carefully as you do in this notebook) rather than an ultimate decision as to what is correct.

**NB whilst you may become suspicious about your data quality, and you know you have missing values, you will not here dive down the rabbit hole of checking all values or web scraping to replace missing values.**

you now see a rather long tailed distribution. You may wonder about the now most extreme value that is above 8000, but similarly you may also wonder about the value around 7000. If you wanted to spend more time manually checking values you could, but leave this for now. The above distribution is plausible

You can adopt a similar approach as for the suspect skiable area value and do some spot checking. To save time, here is a link to the website for [Heavenly Mountain Resort](https://www.skiheavenly.com/the-mountain/about-the-mountain/mountain-info.aspx). From this you can glean that you have values for skiable terrain that agree. Furthermore, you can read that snowmaking **covers 60%** of the trails.

This is **less than the value of 3379 in your data** so you may have a judgement call to make. However, notice something else. **You have no ticket pricing information at all for this resort. Any further effort spent worrying about values for this resort will be wasted. You'll simply be dropping the entire row**

**Drop the fastEight column in its entirety; half the values are missing and all but the others are the value zero. There is essentially no information in this column.**



What about yearsOpen? How many resorts have purportedly been open for more than 100 years?

Okay, one seems to have been open for 104 years. But beyond that, one is down as having been open for 2019 years. This is wrong! What shall you do about this?

What does the distribution of yearsOpen look like if you exclude just the obviously wrong one?

The above distribution of years seems entirely plausible, including the 104 year value. You can certainly state that no resort will have been open for 2019 years! It likely means the resort opened in 2019. It could also mean the resort is due to open in 2019. You don't know when these data were gathered!

The smallest number of years open otherwise is 6. You can't be sure whether this resort in question has been open zero years or one year and even whether the numbers are projections or actual. In any case, you would be adding a new youngest resort so it feels best to simply drop this row.

The other features you had mild concern over, you will not investigate further. Perhaps take some care when using these features.

## 2.7 Derive State-wide Summary Statistics For Our Market Segment

You have, by this point removed one row, but it was for a resort that may not have opened yet, or perhaps in its first season. Using your business knowledge, you know that state-wide supply and demand of certain skiing resources may well factor into pricing strategies. Does a resort dominate the available night skiing in a state? Or does it account for a large proportion of the total skiable terrain or days open?

If you want to add any features to your data that captures the state-wide market size, you should do this now, before dropping any more rows. In the next section, you'll drop rows with missing price information. Although you don't know what those resorts charge for their tickets, you do know the resorts exists and have been open for at least six years. Thus, you'll now calculate some state-wide summary statistics for later use.

Many features in your data pertain to chairlifts, that is for getting people around each resort. These aren't relevant, nor are the features relating to altitudes. Features that you may be interested in are:

* TerrainParks
* SkiableTerrain\_ac
* daysOpenLastYear
* NightSkiing\_ac

When you think about it, these are features it makes sense to sum: the total number of terrain parks, the total skiable area, the total number of days open, and the total area available for night skiing. You might consider the total number of ski runs, but understand that the skiable area is more informative than just a number of runs.

## 8 Drop Rows With No Price Data

You know there are two columns that refer to price: 'AdultWeekend' and 'AdultWeekday'. You can calculate the number of price values missing per row. This will obviously have to be either 0, 1, or 2, where 0 denotes no price values are missing and 2 denotes that both are missing

**About 14% of the rows have no price data. As the price is your target, these rows are of no use. Time to lose them.**

These distributions are much better. There are clearly some skewed distributions, so keep an eye on fastQuads, fastSixes, and perhaps trams. These lack much variance away from 0 and may have a small number of relatively extreme values. Models failing to rate a feature as important when domain knowledge tells you it should be is an issue to look out for, as is a model being overly influenced by some extreme values. If you build a good machine learning pipeline, hopefully it will be robust to such issues, but you may also wish to consider nonlinear transformations of features.

Population and area data for the US states can be obtained from [wikipedia](https://simple.wikipedia.org/wiki/List_of_U.S._states). Listen, you should have a healthy concern about using data you "found on the Internet". Make sure it comes from a reputable source. This table of data is useful because it allows you to easily pull and incorporate an external data set. It also allows you to proceed with an analysis that includes state sizes and populations for your 'first cut' model. Be explicit about your source (we documented it here in this workflow) and ensure it is open to inspection. All steps are subject to review, and it may be that a client has a specific source of data they trust that you should use to rerun the analysis

Extract the state name, population, and total area (square miles) columns

Do you have all the ski data states accounted for?

{'Massachusetts', 'Pennsylvania', 'Rhode Island', 'Virginia'}

No??

Better! You have an empty set for missing states now. You can confidently add the population and state area columns to the ski resort data

Having created this data frame of summary statistics for various states, it would seem obvious to join this with the ski resort data to augment it with this additional data. You will do this, but not now. In the next notebook you will be exploring the data, including the relationships between the states. For that you want a separate row for each state, as you have here, and joining the data this soon means you'd need to separate and eliminate redundances in the state data when you wanted it.

Finally, what will your target be when modelling ticket price? What relationship is there between weekday and weekend prices?

**A couple of observations can be made. Firstly, there is a clear line where weekend and weekday prices are equal. Weekend prices being higher than weekday prices seem restricted to sub $100 resorts**. Recall from the boxplot earlier that the distribution for weekday and weekend prices in Montana seemed equal. Is this confirmed in the actual data for each resort? Big Mountain resort is in Montana, so the relationship between these quantities in this state are particularly relevant.

Is there any reason to prefer weekend or weekday prices? Which is missing the least

**Weekend prices have the least missing values of the two, so** drop the weekday prices and then keep just the rows that have weekend price

**ski\_data.shape**

(277, 25)

Having dropped rows missing the desired target ticket price, what degree of missingness do you have for the remaining rows

Yes, the percentage of missing values per row appear in multiples of 4

This is almost as if values have been removed artificially... Nevertheless, what you don't know is how useful the missing features are in predicting ticket price. You shouldn't just drop rows that are missing several useless features.

There are still some missing values, and it's good to be aware of this, but leave them as is for now.



ski\_data.shape

Save this to your data directory, separately. Note that you were provided with the data in raw\_data and you should saving derived data in a separate location. This guards against overwriting our original data.

After loading and reviewing the ski data, we started with 330 rows and 27 columns and the feature resort, Big Mountain Resort is in the list of resorts and had no missing values. The focus is to find answers to predict the best ticket price that Big Mountain Resort can use for their pricing strategy.

First, it was found there are two ticket price columns (AdultWeekend and AdultWeekday) These columns are the target feature to predict ticket price but could also be affected by other potential features in the dataset.

These were issues encountered:

a) There was a question of uniqueness of the data in names, regions and states. It was found that using the name, combined with either the region or state, provided unique identification for each resort.

b) Looking at the missing data in the two price columns, it was decided that those that were missing both could be dropped as they would provide no guidance for pricing decisions. Missing data in other columns were kept for the time being.

c) Dealing with correcting incorrect data and how it would affect the investigation and should be noted back to the client or data originator.

d) Encountered questionable number of years open. The data was filtered and later excluded the questionable data.

e) Accounted for missing states in ski data.

After making the dataset unique in names, states and regions, the result says the majority of the states weekend prices are higher than the weekday prices. The remaining states showed a marked difference between weekdays and weekend prices.

Findings: 16% of the resorts have missing ticket prices and this means the lack of data for these resorts would likely lead to the dropping of those records. The data showed that 82% of the entries had both prices, 3% had one of the prices, and 14% had no pricing data.

The following rows that were removed:

1. Remove a row because that resort may not have opened yet or perhaps in its first season based on the data gathered
2. Removed 14% of the rows that have no price data. We are interested in data with prices and these did not have any data

Next, is a question to consider: Does statewide supply and demand of certain skiing resources factor into pricing strategies? Investigation included terrain parks at the resort, total skiable areas, days open last year and night skiing total area.

Summary statistics for various states where in population and state area was added to the resort data to find out more about the relationship between the states making the investigation of the data wider in scope to find the answers to the main question. Further investigation on this question is pending.

The final ski data shape is 277 rows and 25 columns.

After the cleaning and wrangling at this point, weekend prices have the least missing values between the two so dropping the weekday prices seems reasonable.

After doing this, the percentage of missing values per row appears in multiple of 4’s. It is as if the values have been removed artificially. It makes one think that missing features would still be useful in predicting ticket prices and dropping rows that are missing several useless features may cause us to lose useful information.

At this point in the project, additional data cleaning is needed, and further analysis to reach the conclusion.