Predicting Ski Resort Lift Ticket Prices

IBM Applied Data Science Capstone Project

March 17, 2019

Single-day lift ticket prices

- Expensive
- Proxy for the ski resort riding experience
- Depend on the size of the resort
- Might depend on the resort off-mountain infrastructure
- Predicting prices is important for dynamic pricing and might guide the future resort infrastructure development
- Comparing actual and predicted prices might might help find good deals

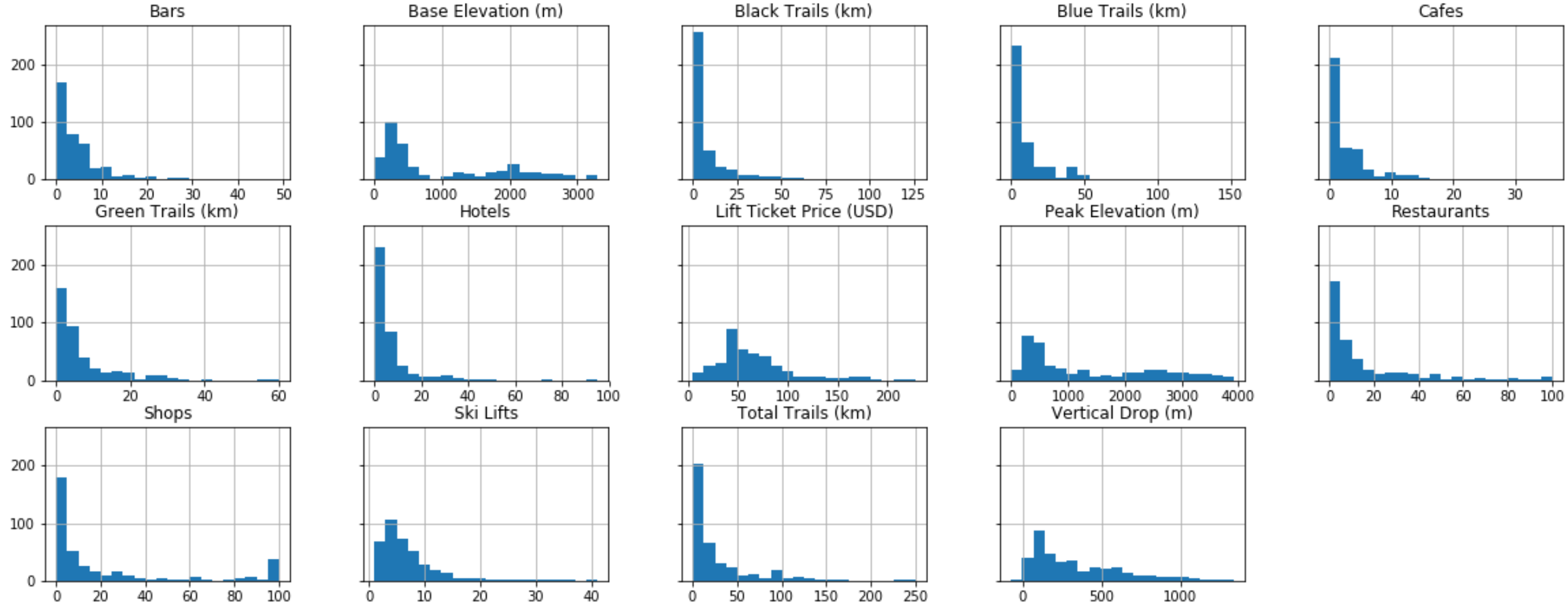
Data acquisition

- www.skiresort.info [number of ski lifts, vertical drop, price, etc.]
- www.onthesnow.com [lift ticket price]
- wikipedia.org [number of ski lifts, vertical drop, price, etc.]
- www.google.com [ski resort latitude and longitude]
- developer.foursquare.com [number of restaurants, hotels, etc.]

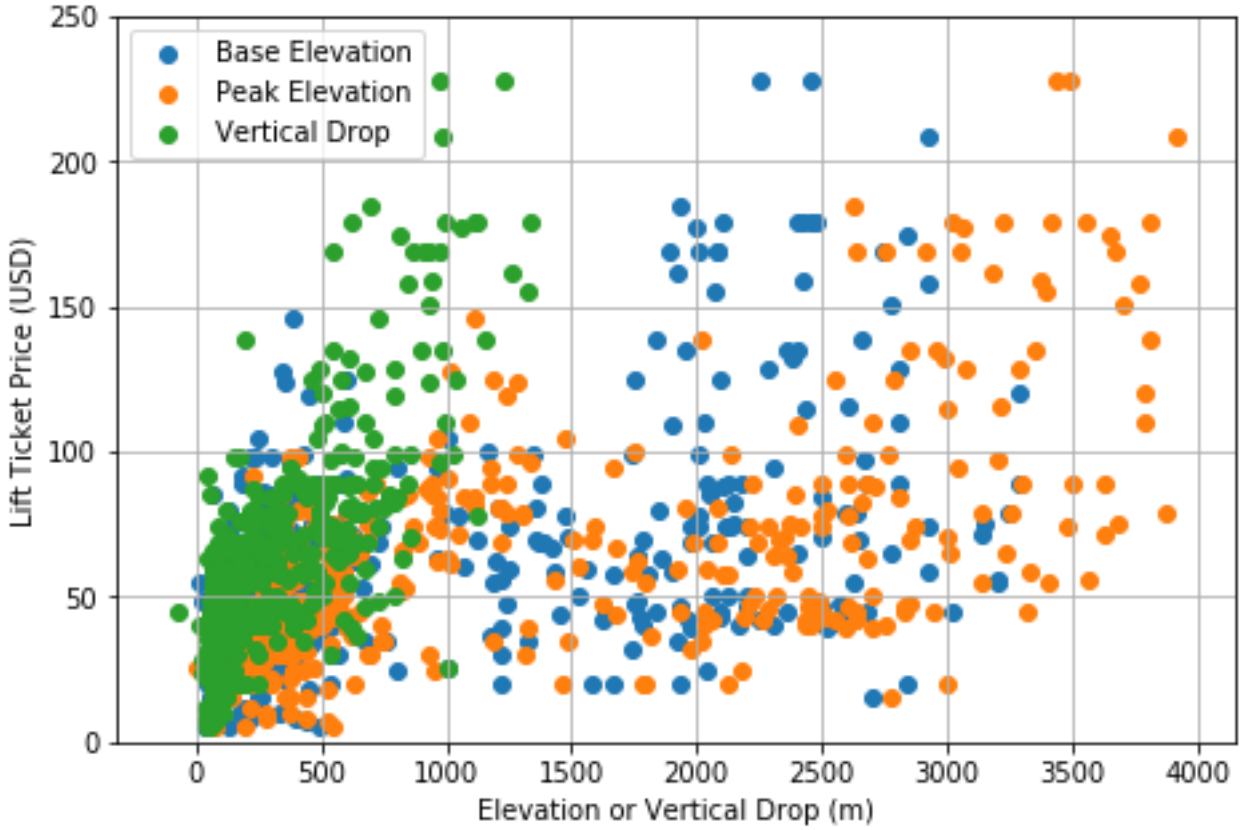
Data cleaning

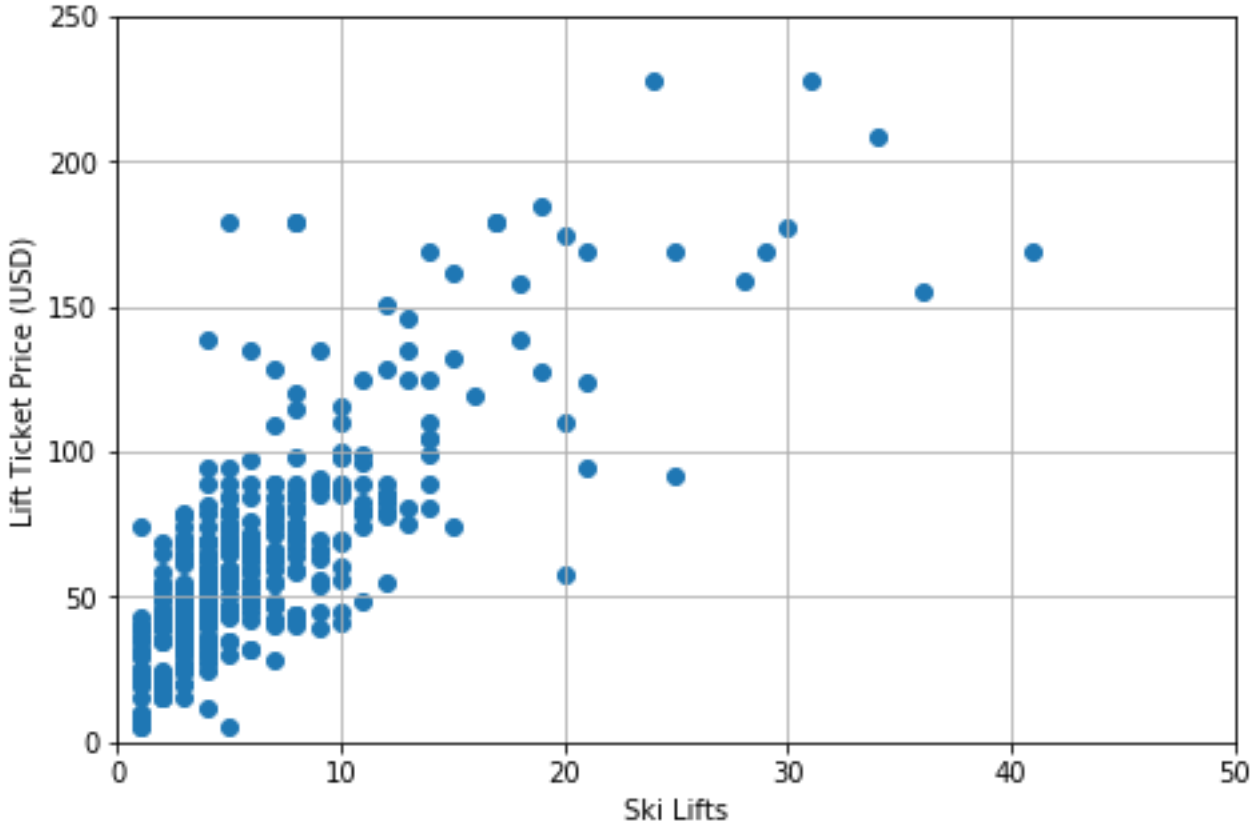
- Remove duplicates
- Fill missing price values/improve price relevance by merging all datasets and keeping the highest price value (assumption for the most recent price)
- Drop irrelevant or mostly unfilled features
- Drop 33 ski resorts and areas that still have missing prices
- Add information on number of venues within a 5 km radius using Foursquare API
- Resulting dataset has 385 entries with price information and 13 additional numerical features

Data exploration: Histograms

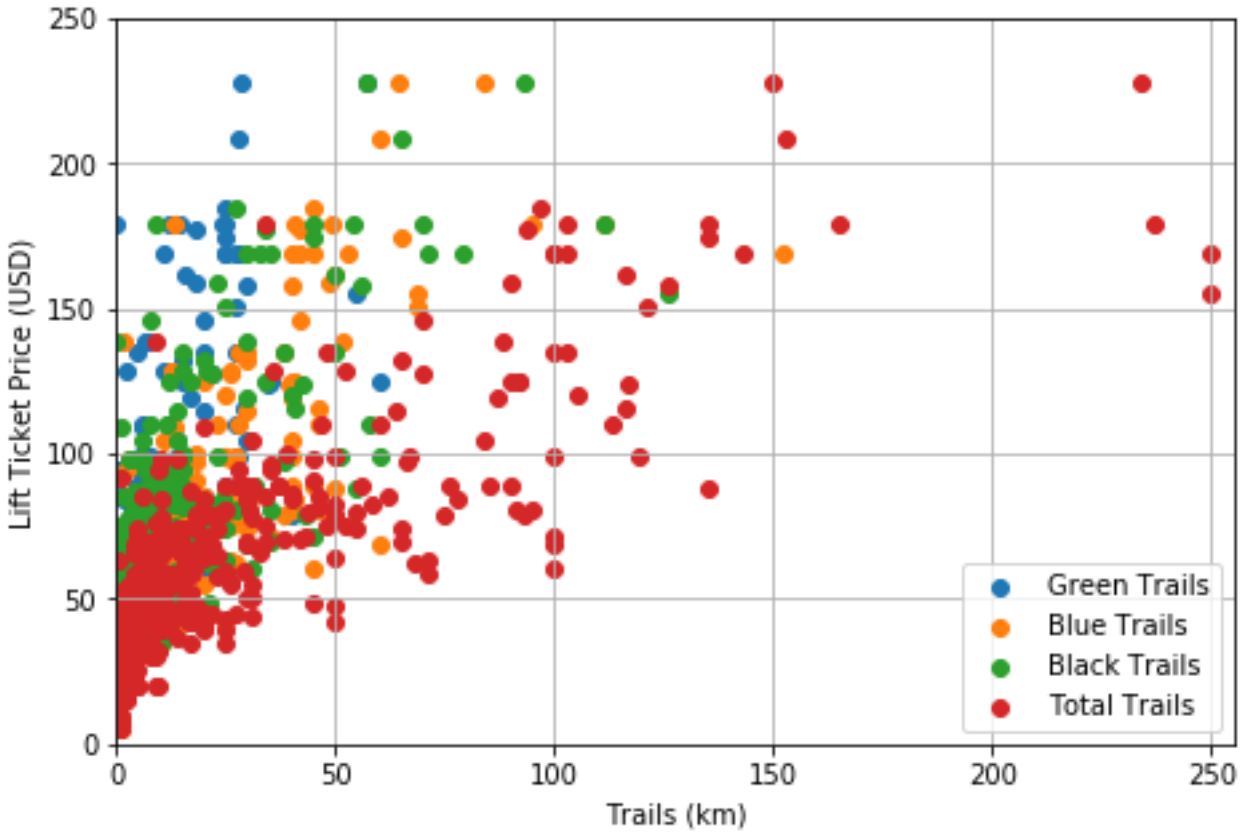


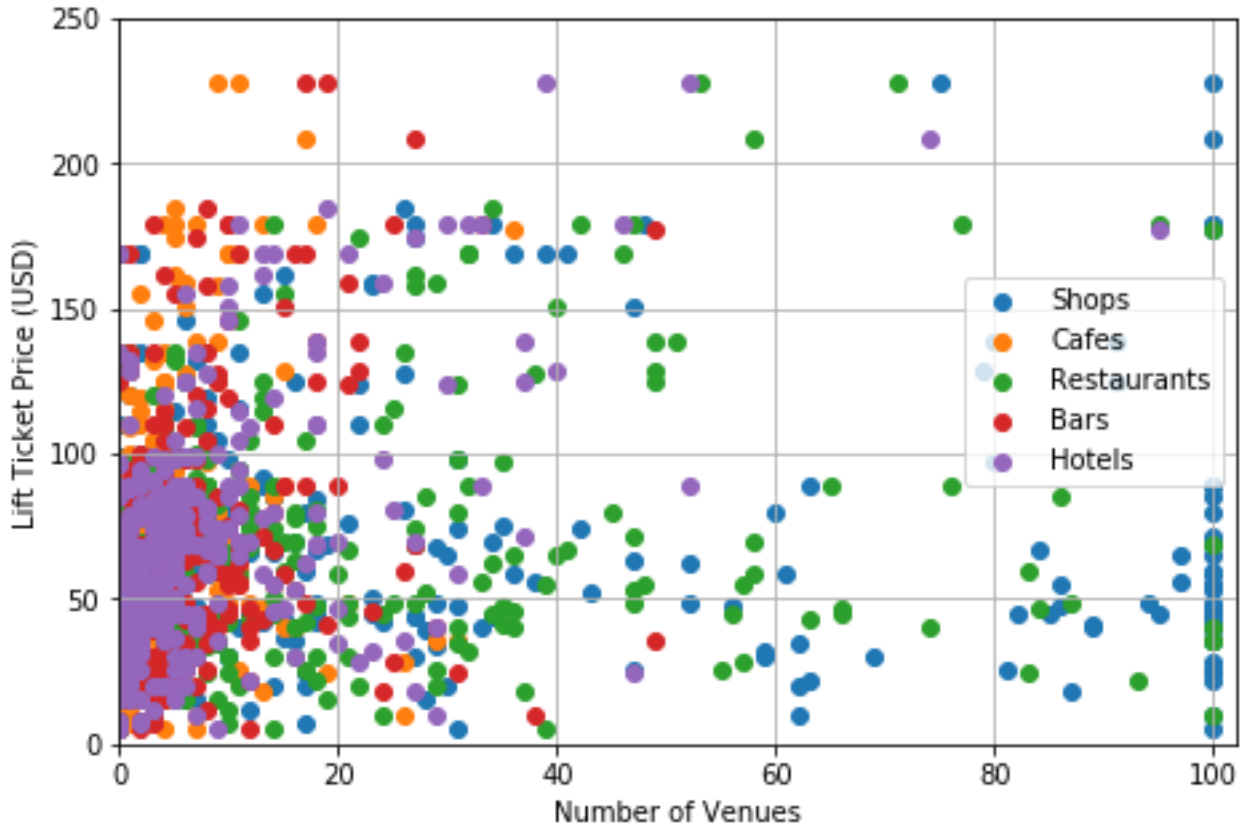
Data exploration: Correlations



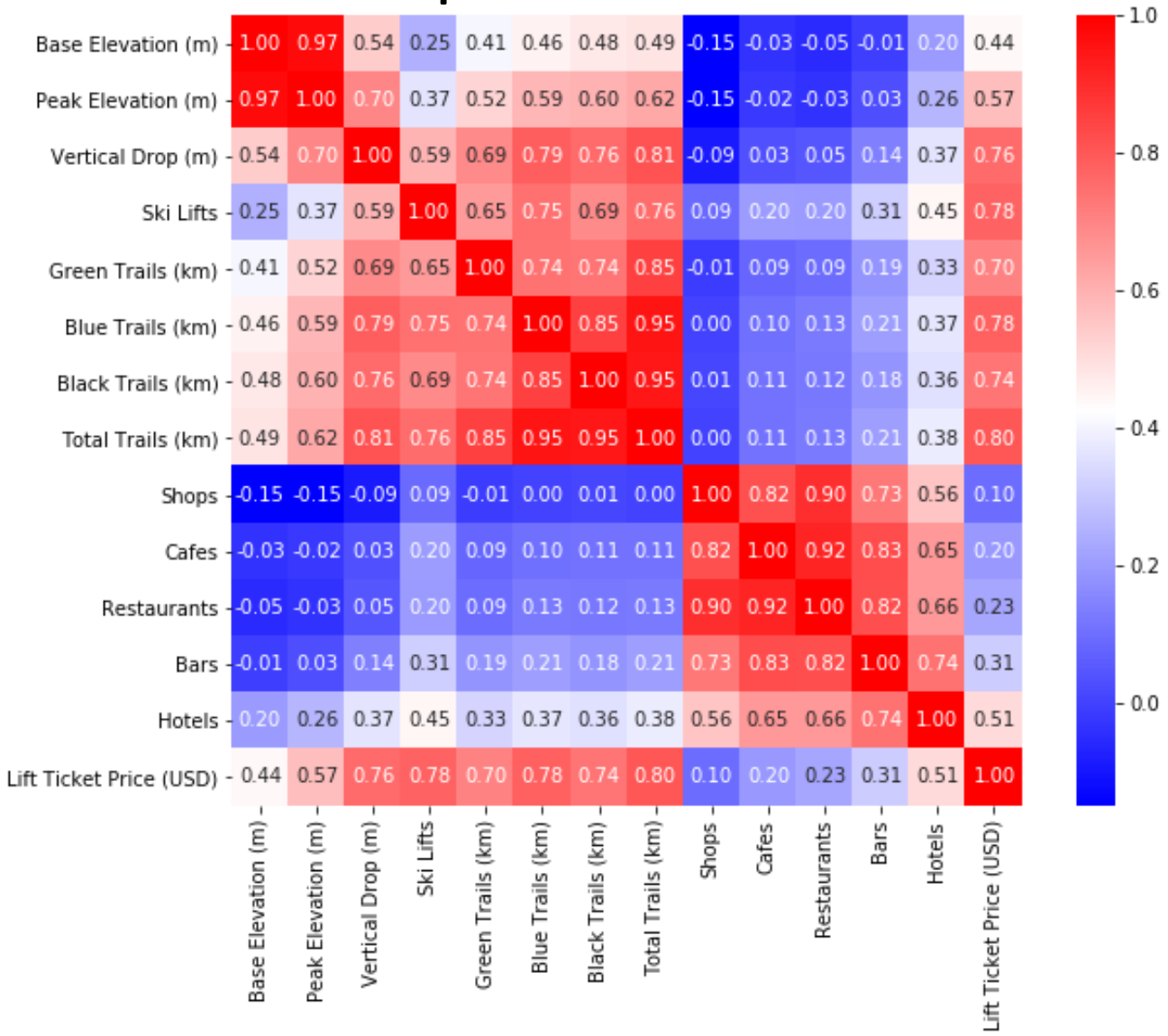


Data exploration: Correlations





Data exploration: Correlations



- *Peak Elevation* and *Total Trails* are redundant
- Resort stats and number of nearby venues features do not correlate as much as between themselves
- *Shops* and *Cafes* unlikely to improve the prediction performance

Models: Feature selection

1. *All Stats* (naïve, baseline model): base elevation, peak elevation, vertical drop, ski lifts, green trails, blue trails, black trails, total trails

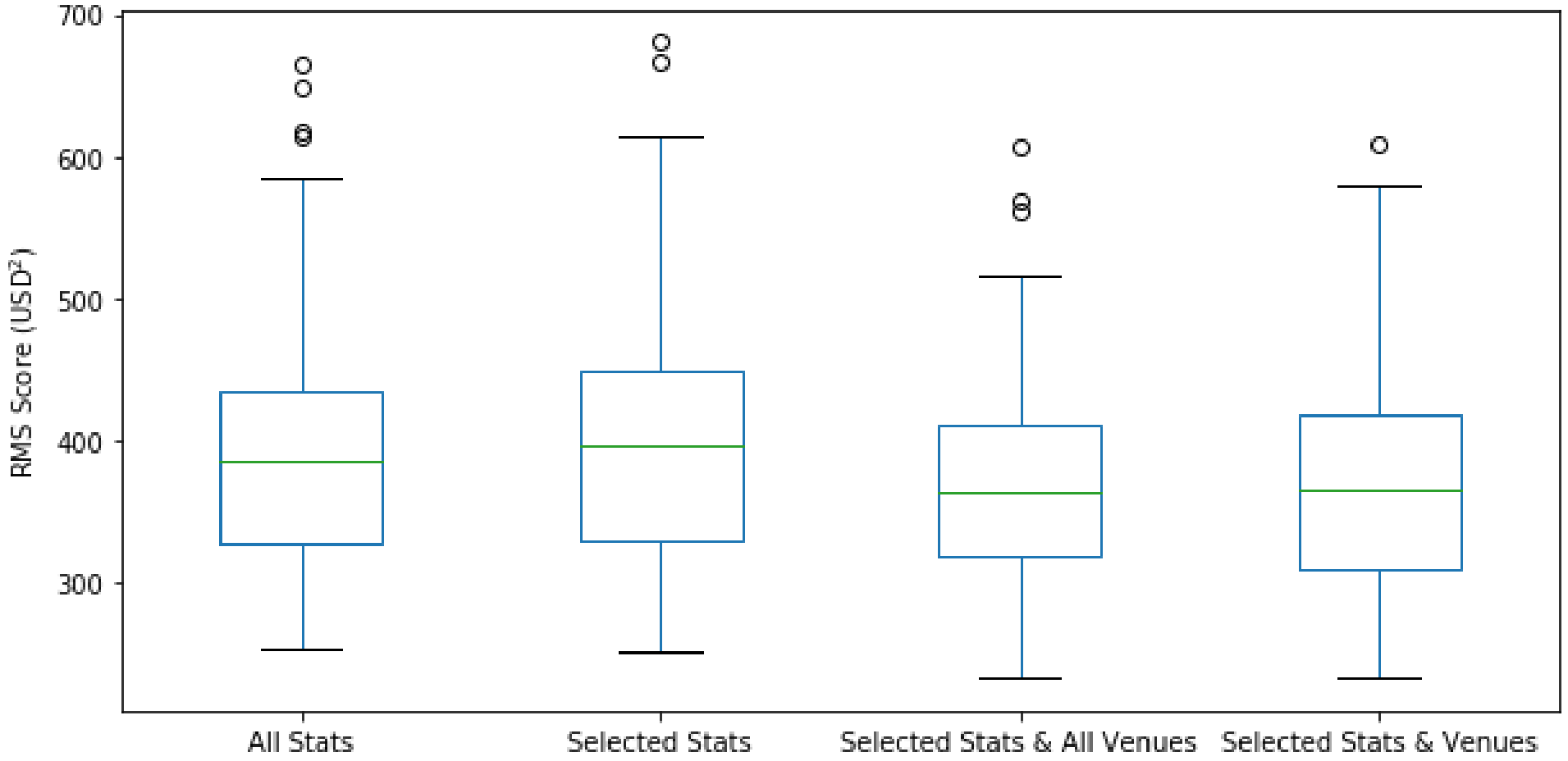
2. *Selected Stats* (less-redundant model): base elevation, vertical drop, ski lifts, green trails, blue trails, black trails

3. *Selected Stats & All Venues* (model with all venue info): base elevation, vertical drop, ski lifts, green trails, blue trails, black trails, shops, cafes, restaurants, bars, hotels

4. *Selected Stats & Venues* (model with the most relevant venue info): base elevation, vertical drop, ski lifts, green trails, blue trails, black trails, restaurants, hotels

Model evaluation: RMS

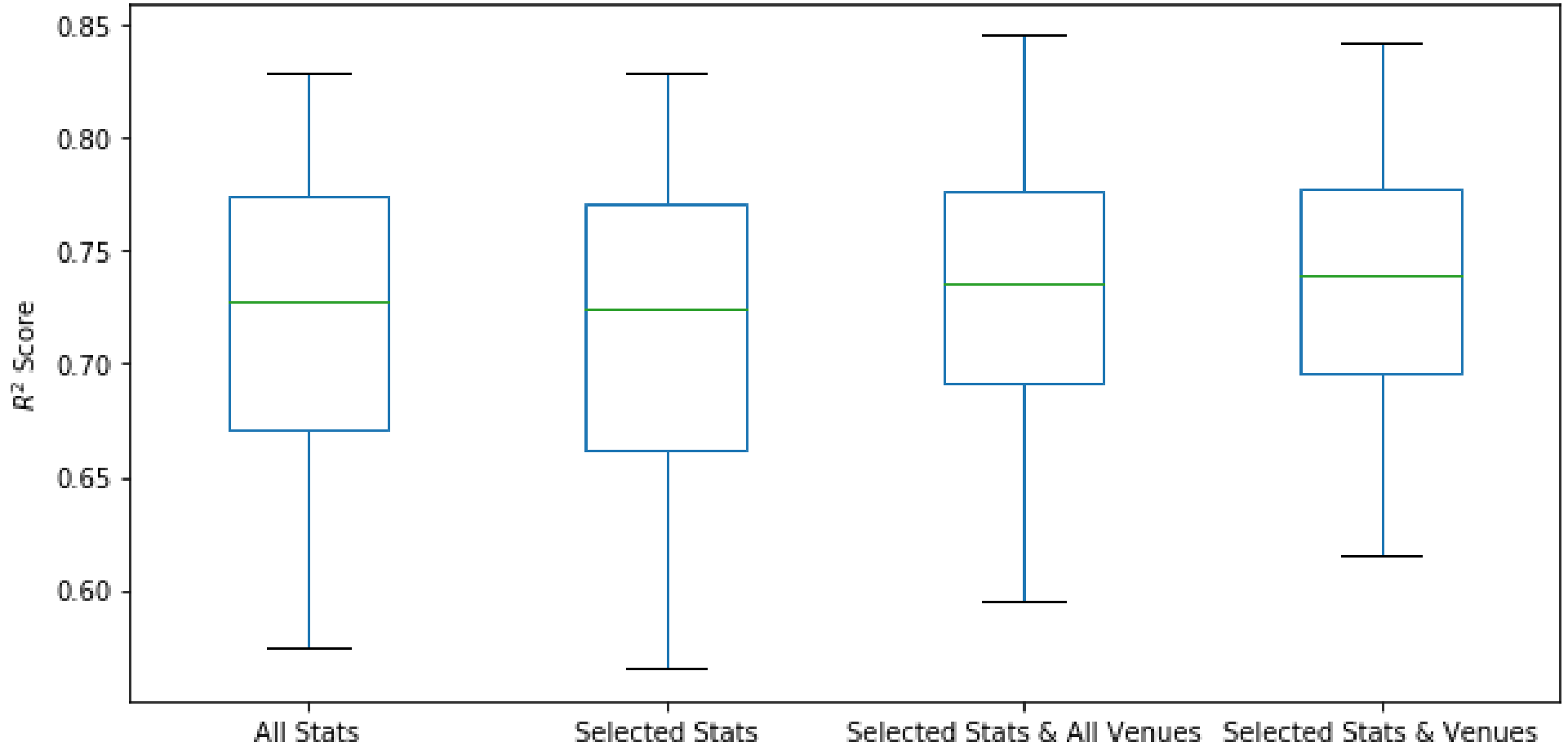
	All Stats	Selected Stats	Selected Stats & All Venues	Selected Stats & Venues
count	100.000000	100.000000	100.000000	100.000000
mean	393.426425	401.158813	373.261989	373.279516
std	93.018727	94.993118	78.437656	78.704967
min	252.369772	250.531713	232.478668	232.805593
25%	327.209253	328.888881	318.170719	310.094670
50%	386.489422	395.957857	363.236319	366.224141
75%	435.562997	449.859424	411.631497	418.904158
max	664.320995	680.741818	606.805815	609.259115



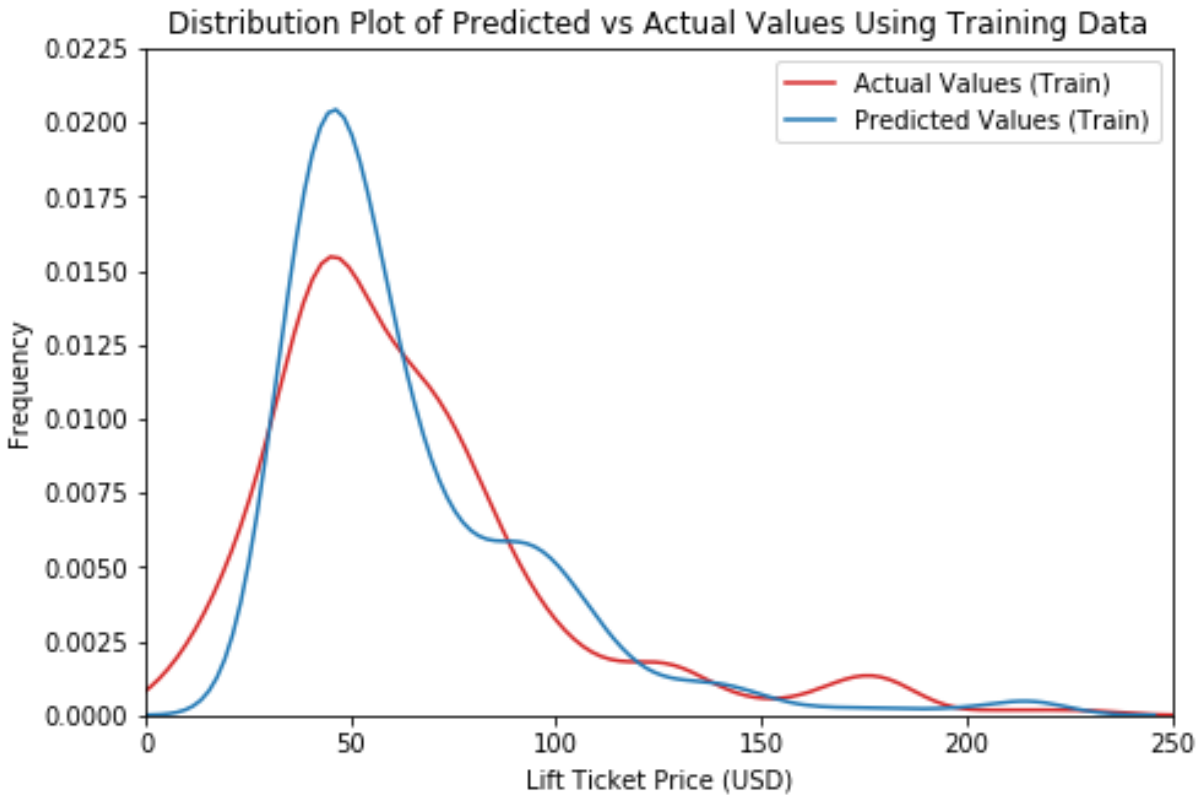
Model evaluation: R^2

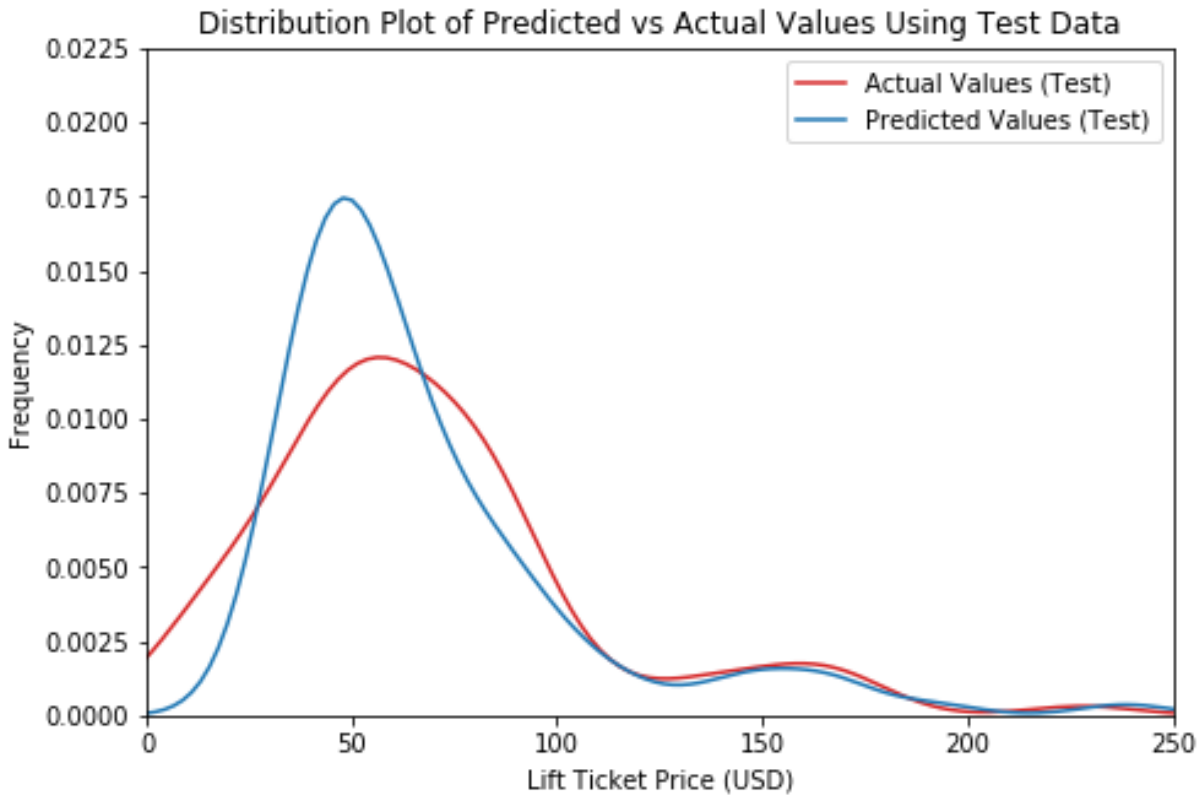
Selected Stats & Venues performs the best!

	All Stats	Selected Stats	Selected Stats & All Venues	Selected Stats & Venues
count	100.000000	100.000000	100.000000	100.000000
mean	0.719842	0.714327	0.733802	0.734135
std	0.066570	0.067899	0.059033	0.057498
min	0.575071	0.565145	0.595133	0.615049
25%	0.670912	0.662536	0.690870	0.695566
50%	0.727849	0.724549	0.735853	0.738808
75%	0.774616	0.771226	0.775955	0.777494
max	0.828915	0.828941	0.845145	0.842283



Best model performance





Best model coefficients

Base Elevation (USD/m)	Vertical Drop (USD/m)	Ski Lifts (USD)	Green Trails (USD/km)	Blue Trails (USD/km)	Black Trails (USD/km)	Restaurants (USD)	Hotels (USD)
0.003009	0.037872	2.512523	0.415106	0.252716	0.025395	0.061857	0.445723

Coefficients suggest that the best predictors for lift ticket price are:

- number of ski lifts
- lengths of the green trails
- number of hotels nearby

While this does not prove causality, this seems to be helpful in finding the most important on- and off-mountain infrastructure factors

Conclusion and future directions

- Including the number of nearby restaurants and, especially, of hotels improves the model performance
- Improvement is relatively small but the effect is robust: off-mountain infrastructure is important
- Model improvement ideas:
  - select more specific venues nearby/filter the venue datasets/adjust the radius
  - include other basic resorts stats as the average annual snowfall
  - include the cost of living at the state
  - include one hot encoding of the resort owners [www.nsaa.org]
  - explore residuals for any non-linear behavior
  - cluster the resorts for more insights