Rock Climbing Logbook (8a.nu) Analysis

or: Does Being Tall *Really* Help?

Steve Bachmeier
2019-03-23

0 Code required for report

In [1]:

```
#-----------------------------------------------------------------
# Run the following to hide the In[] and Out[] margin.
# Doing so will not allow headings to be collapsed.
'''
from IPython.core.display import display,HTML
display(HTML('<style>.prompt{width: 0px; min-width: 0px; visibility: collapse}</style>'))
'''

#-----------------------------------------------------------------
# Run the following to import required libraries
import pandas as pd
import numpy as np
import dill
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.formula.api as smf
from statsmodels.stats.outliers_influence import variance_inflation_factor as vif_

#-----------------------------------------------------------------
# Run the following to load required pickled objects

df_all = dill.load(open("df_all.pkl", "rb"))
df_results_univariate = dill.load(open("df_results_univariate.pkl", "rb"))
df_results = dill.load(open("df_results.pkl", "rb"))
```

1 Synopsis

A large dataset scraped from the online rock climbing log www.8a.nu was analyzed using ordinary least squares regression to determine whether certain characteristics - specifically, a climbers height, weight, age, and years of experience - have a statistically significant influence on overall rock climbing ability. The dataset was cleaned and then organized to include the *maximum* climbing grade attained by each unique user on a scale of 0 to 82 (which corresponds to 0 to 9c+/10a in the commonly used Font rock climbing grade scale).

It was found that height and weight are highly correlated and so weight was dropped as a predictor. Further, the data was broken out into four different cases - two for gender and two for climbing type (bouldering and rope climbing). A summary of the results is as follows:

- **Age is a statistically signficant predictor for all four cases and consistently shows a negative correlation with climbing ability.**
- **Height is not a statistically significant predictor for females but is a negative predictor for males in both bouldering and rope climbing. That is, for a linear regression model, a male's climbing ability is negatively impacted by an increase in height.**
- **Years of experience is a statistically significant predictor for all four cases and consistently shows a positive correlation with climbing ability.**

The final results table for the multivariate analysis is shown below.

In [2]:

```
df_results[df_results['Coefficient'] != 'Intercept']\
        .sort_values(by=['Coefficient','Gender'])\
        [['Coefficient','Gender','Type','Value','Std_error','P_value','Statist
ically_significant']]
```

Out[2]:

	Coefficient	Gender	Type	Value	Std_error	P_value	Statistically_significant
2	age	F	Bouldering	-0.27	0.03	0.000	Yes
10	age	F	Rope	-0.23	0.03	0.000	Yes
6	age	M	Bouldering	-0.35	0.01	0.000	Yes
14	age	M	Rope	-0.37	0.01	0.000	Yes
1	height	F	Bouldering	-0.05	0.03	0.059	No
9	height	F	Rope	-0.02	0.03	0.445	No
5	height	M	Bouldering	-0.04	0.01	0.000	Yes
13	height	M	Rope	-0.07	0.01	0.000	Yes
3	years_climbing	F	Bouldering	0.98	0.05	0.000	Yes
11	years_climbing	F	Rope	1.30	0.04	0.000	Yes
7	years_climbing	M	Bouldering	0.66	0.02	0.000	Yes
15	years_climbing	M	Rope	0.91	0.01	0.000	Yes

Notes:

- There does appear to be some non-random trends in the plot of the maximum climbing grade residuals as a function of fitted values from the final multivariate model. This implies that there are some affects that are not being accounted for (eg a missing predictor, a higher-order effect, etc) and the model can probably be improved.
- 8a.nu users are rather talented rock climbers. As such, the data used for this analysis is not terribly representative of the general population of rock climbers.

2 Overview

This section provides a brief overview of the background, data source, and goal of this analysis.

2.1 Background

Rock climbing is a tough sport. It takes physical strength, extreme endurance in strange muscle and tendon groups that are not typically targeted (finger ligaments, forearms, etc), dancer-like body awareness and balance (although I am failing hard in this arena!), and serious mental grit. With such obstacles to overcome, then, nothing is more infuriating than when I finally - finally! - successfully complete a climb without falling only to hear such grumblings as: "Ugh, it's so easy if you're tall" or "It's not fair - you can just reach past the difficult holds!"

It's true that with a height of six feet and a +2 so-called ape index (that is, my tip-to-tip arm span is my height plus two inches, ie 6'2") I can sometimes do in one big move what a shorter guy or gal might do in two or three. But conversely, and I've argued this adamently for years, there are plenty of times where the next hold is perfectly at arms length for a short person whereas for me it's relatively lower and requires me to scrunch up into all sorts of weird and inefficient positions. The more I crunch, the more my butt sticks out, and the farther from the wall my center of gravity goes - it's physics! That's not even to mention the math behind leverage (where my longer arms are not necessarily an advantage) or the fact that taller people also tend to be heavier.

So what's the deal? Is it better for climbing to be tall or short? And while we are at it, what about other factors like weight, gender, and years of experience?

2.2 Data

The raw data was downloaded from: https://www.kaggle.com/dcohen21/8anu-climbing-logbook (https://www.kaggle.com/dcohen21/8anu-climbing-logbook).

The data used was scraped from an online logbook, https://beta.8a.nu/ (https://beta.8a.nu/), by David Cohen (https://www.kaggle.com/dcohen21 (https://www.kaggle.com/dcohen21)). It should be noted that the user's code to scrape the website is no longer avaialable due to DMCA takedown.

Per the data description, it was collected on 2017-9-13. It is assumed to be accurate.

2.2.1 8a.nu description

8a.nu is a rather popular online logbook used to track rock climbs completed. A user creates an account, searches for a specific climb he/she finished, and logs information about it such as its difficulty; whether it was on-sighted (climbed without falling on the first time), flashed (climbed without falling on the first time but after having watched someone else do it), or redpointed (climbed without falling after previously failing); how much he/she liked it based on a 0-3 star rating; and any other notes.

The website also includes details about each climb such as the consensus difficulty grade and rating, location, and type of climb ('rope' refers to taller climbs that are typically done with a harness/rope and belayer and 'boulder' refers to shorter climbs - typically 10 meters or less - done with no rope).

Finally, user details can be optionally input including things such as gender, height, weight, the year he/she began climbing, etc.

It is important to note that it has been assumed that climbs on 8a.nu are only logged when successfully completed without falling.

2.3 Goal

To determine which - if any - attributes give climbers a statistically significant advantage.

3 Data wrangling

This section outlines the data formatting completed. Refer to the Jupyter Notebook analysis.ipynb for entire analysis and code.

3.1 Data preparation

The downloaded dataset comes in the form of an SQLite database. Using the Python `sqlite3` package, all tables in the database are read into `pandas` dataframes and then saved out as csv files. All csv files are then read back into Python into one dictionary which is finally unpacked into separate raw data dataframes; there are four:

- users (62592 rows, 22 columns)
- ascents (~4.1 million rows, 28 columns)
- climbing methods (5 rows, 4 columns)
- climbing grades (83 rows, 14 columns)

Each row in the users dataframe corresponded to a different unique person and includes such information as user ID, name, location, gender, height, weight, etc.

The ascents dataframe is a log of all of the climbs logged on 8a.nu and includes information like the ascent ID, the user ID of that specific ascent, the method ID, date, notes, etc.

The method dataframe is a small one that includes ony five rows. Its columns consist of 'ID', 'score', 'shorthand', and 'name'. 'Shorthand' and 'name' are redundant (lower case vs capitalized) and are simply the types of finish a climber can get (redpoint, flash, onsight, and toprope). It's unclear what exactly score is. Note that there is no option for failed attempts (finishing but having fallen at least once in the process).

Finally, the grades dataframe lists all of the current climbing grade possibilities from very easy to world class. The grade ID runs from 0 to 83 and *nearly* lines up with the index of the dataframe; IDs 48, 61, and 74 are missing. The grades dataframe *index*, on the other hand, lines up one-to-one with the commonly used Font rating system (0 to 9c+/10a). Other columns are included (including the American Yosemite Decimal System grading system for rope climbing and the V scale for bouldering) but the fact that the Font grades align completely with the index numbers implies that the grade index is a good indicator of difficulty. Essentially, it is assumed that climbs are linearly difficult from a grade index of 0 (Font 0) up to a grade index of 82 (Font 9c+/10a).

3.2 Data cleaning

With the raw data now in usable dataframes, they can be cleaned. Refer to analysis.ipynb for the code itself.

In summary:

1. Drop unnecessary columns from all four dataframes
2. Check for null/empty values in all dataframes [a]
3. For the user dataframe:
   A. Convert the 'sex' variable to dummy variable 'is_female' [b]
   B. For the birth years that exist, extract just the year.
   C. Drop any remaining duplicate user IDs.
4. For the ascent dataframe:
   A. Drop any duplicate ascent IDs.
   B. Convert user ID column to integers.

[a] It should be noted that there is a fair amount of null values for the 'birth' feature in the users dataframe; 55.5% of people did not bother putting their birth date into the website. It was decided not to drop the column nor to impute the values; instead, the data is filtered ad hoc as necessary.

[b] The 'sex' variable had three unique values. The vast majority were 0s and 1; four users had sex=255. As I highly doubt that the 8a.nu team has implemented non-binary gender assignment functionality into the website, those four users are simply dropped from the dataset.

As for the 0s and 1s, it is found that 0s are on average heavier and taller than 1s. Further, there are significantly more 0s than 1s. As most climbers are male and males tend to be taller and heavier than females, the 'sex' category is renamed to 'is_female'.

3.3 Data merge and new feature creation

With cleaned datasets, they can now be merged into one large working dataframe. This is completed by merging the user dataframe onto the ascent dataframe using the user ID primary keys. The columns of the new working dataframe are then rearranged and two new features added: 'age' ('year' - 'birth_year') and 'years_climbed' ('year' - 'started').

With the new 'age' and 'years_climbed' features created, the 'year', 'birth_year', and 'started' columns are no longer of interest and are dropped.

The grades dataframe index values are then merged onto the working dataframe using the grade ID primary keys. Both grade ID columns are then dropped.

Finally, to clarify some of the column names (and remain consistent), the following labels are renamed:

- 'user_id' to 'id_user'
- 'method_id' to 'id_method'
- 'climb_type' to 'is_bouldering' [a]

Notes

[a] The 'climb_type' variable is binary and consists of 0s and 1s. By comparing the specific climbs with the most 0s and the most 1s (ie the most popular climbs for each 'climb_type') to their descriptions on another popular online log (www.mountainproject.com), it became clear that 'climb_type' = 0 refers to rope climbs and 'climb_type' = 1 refers to bouldering problems. As such, the variable is renamed to 'is_bouldering' to reduce confusion.

3.4 Filter data

Recall that there are four different method types (redpoint, flash, onsight, and toprope). For the purposes of this analysis, toprope ascents are not considered successful [a] and so all ascents with an 'id_method' = 4 are dropped from the working dataframe. Then the 'id_method' column is dropped.

At this point the working dataframe consists of ~4 million rows and 9 columns.

Notes:

[a] Toproping is when a person climbs attached to the rop which goes up to an anchor at the top of the climb and then back down to the belayer.

3.4.1 A note about NULL/bogus values

It should be noted here that, despite having cleaned up the data, NULL and bogus values have been left intact on purpose. Specifically, the date data ('years_climbing' and 'age') is full of NULL values as well as values that simply don't make sense (negatives, 0s, and impossibly large numbers). Likewise, the weight data includes some 0s and the height data includes values ranging from 0 cm to 255 cm (0-8.4 feet).

All of this data is purposefully left in the working dataset because to remove them all at once would drastically reduce the dataset even for variables that are otherwise fine. For example, over 21% of the observations have NULL values for 'age' (as a result of either the original 'year' or 'birth_year' being NULL). However, just because users chose not to input that specific data does not mean that the rest of their inputs are bad! As such, the entirety of the dataset is kept intact and the following filters will be applied as necessary throughout the analysis:

- 'height': 120-240 cm (~4-8 feet)
- 'weight': not 0, which for this dataset happens to be equivalent to 40-100 kg (~88-220 kg)
- 'age': 10-50 years
- 'years_climbing': 0-40 years

Note that these filters are somewhat arbitrary! There certainly may be climbers in the dataset shorter than 120 cm, taller than 240 cm, with more than 40 years of climbing experience, or outside the age range of 10-50. However, it is assumed that most users that have these outlier vales are (a) either just that - outliers - and so should not be included anyway or (b) simply input incorrect data.

3.5 Reduce

To accurately gage a climber's skill level, it is important to not consider every single grade ever climbed but instead to extract each user's personal best. Much like a world-class sprinter does not always run to beat a record, climbers very often - mostly, really - climb at grades that are below their historical high. As such, the total working dataframe is reduced to only include each user's *oldest* maximum 'index_grade' for both rope climbs and boulder problems.

After this reduction, the 'id_ascent' column is no longer needed for such sorting and grouping and so is dropped.

The head of the working dataframe is shown below.

In [3]:
```
df_all.head()
```

Out[3]:
```
   id_user  is_female  height  weight  is_bouldering  index_grade   age  years_climbing
0        1          0     177      73              0           59  25.0             5.0
1        1          0     177      73              1           49  26.0             6.0
2        2          0       0       0              0           47   NaN             1.0
3        2          0       0       0              1           45   NaN             1.0
4        3          0     180      78              0           59  26.0             4.0
```

3.6 Separate data

Recognizing that rope climbing and bouldering are quite different despite both being considered "rock climbing," it was decided to separate the dataset based on 'is_bouldering'. Similarly, the datasets are separated on 'is_female' as well to get different results for each gender. There are thus five different dataframes to be used depending on the analysis at hand:

- df_all (49,598 rows, 8 columns)
- df_bouldering_female (2,396 rows, 6 columns)
- df_bouldering_male (16,817 rows, 6 columns)
- df_rope_female (4,592 rows, 6 columns)
- df_rope_male (25,793 rows, 6 columns)

Note that the four data subsets contain six columns instead of eight like `df_all` ; this is because 'is_female' and 'is_bouldering' are not useful since that information is contained in the dataframe names and so are dropped.

4 Exploratory data analysis

It is always a good idea to do at least a bit of exploratory data analysis before diving too deep into an analysis; it can shed light on trends, potential problems, etc.

4.1 Basic statistics

Basic statistics of the `df_all` dataframe with all filters discussed above are shown below.

In [4]:

```
df_all[
    (df_all['height'] >= 120) & (df_all['height'] <= 240) &
    (df_all['weight'] != 0) &
    (df_all['age'] >= 10) & (df_all['age'] <= 50) &
    (df_all['years_climbing'] >= 0) & (df_all['years_climbing'] <= 40)
].describe()
```

Out[4]:

	id_user	is_female	height	weight	is_bouldering	index_grade	
count	21370.000000	21370.000000	21370.000000	21370.000000	21370.000000	21370.000000	213
mean	28426.796678	0.110014	176.283528	68.000608	0.400796	51.369584	
std	18047.372392	0.312915	8.722877	9.928711	0.490071	9.025313	
min	1.000000	0.000000	120.000000	40.000000	0.000000	0.000000	
25%	13441.000000	0.000000	172.000000	63.000000	0.000000	47.000000	
50%	26109.500000	0.000000	177.000000	68.000000	0.000000	53.000000	
75%	41793.000000	0.000000	182.000000	73.000000	1.000000	59.000000	
max	67022.000000	1.000000	228.000000	100.000000	1.000000	79.000000	

Of particular interest is perhaps the mean values. From the table we see that the average climber in the 8a.nu database (for both genders and both types of climbing) is ~176 cm (~5' 9.3") tall, weighs 68 kg (150 lbs), and climbs at a grade index of ~51 (which corresponds to 7b in the Font grading system, 12b in the Yosemite Decimal System, or V8 in the V scale for bouldering) after 6.8 years of climbing. Also, the average values for 'is_female' and 'is_bouldering', being binary dummy variables, can be interpreted as 11% of the users are female and 40% of the ascents are for boulder problems.

It should be noted that an average best climbing grade if Font 7b is very respectable; it can take many years of consistent climbing to get to that level. It is thus clear that users of 8a.nu tend to be quite accomplished at the sport and this should be kept in mind when interpreting results as the data is not necessarily representative of all climbers.

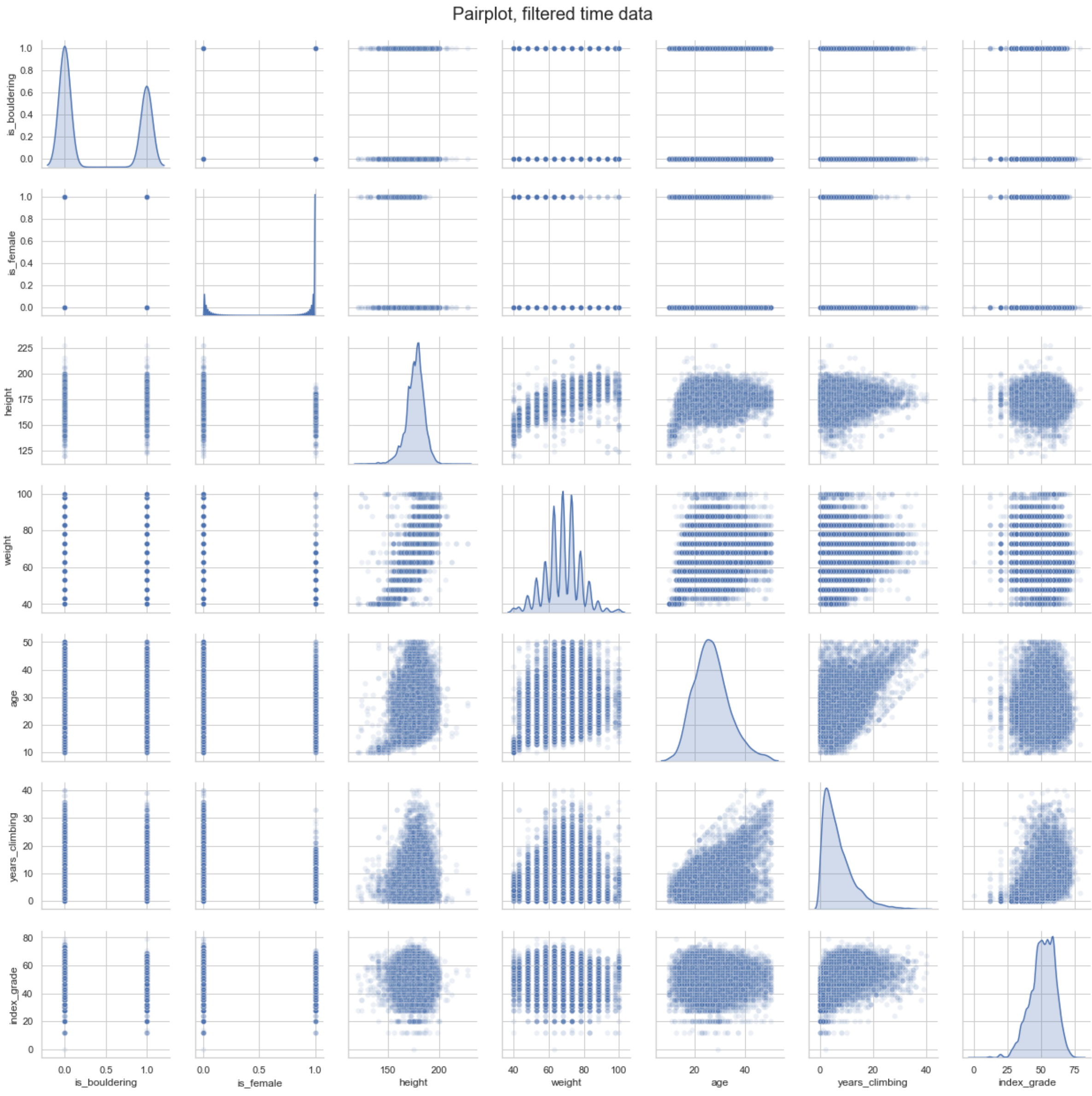
4.1 Pairplot

Pairplots can be useful at shedding light on correlations between the variables of a given dataset.

In [5]:

```
%matplotlib inline

sns.set(style="whitegrid")
g = sns.pairplot(df_all[
    (df_all['height'] >= 120) & (df_all['height'] <= 240) &
    (df_all['weight'] != 0) &
    (df_all['age'] >= 10) & (df_all['age'] <= 50) &
    (df_all['years_climbing'] >= 0) & (df_all['years_climbing'] <= 40)]\
                 [['is_bouldering','is_female','height','weight','age','years_
climbing','index_grade']],
                 diag_kind='kde', dropna=True, plot_kws={'alpha':0.1})
g.fig.suptitle("Pairplot, filtered time data", size=20)
plt.subplots_adjust(top=0.95)
```

```
C:\Users\steve\Anaconda3\lib\site-packages\scipy\stats\stats.py:1713: FutureW
arning: Using a non-tuple sequence for multidimensional indexing is deprecate
d; use `arr[tuple(seq)]` instead of `arr[seq]`. In the future this will be in
terpreted as an array index, `arr[np.array(seq)]`, which will result either i
n an error or a different result.
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```



Several interesting points are highlighted by the above pairplot. Regarding the distribtuions of the features (as shown in the KDE plots along the diagonal):

- Height and weight are qualitatively normally distributed.
- Age is qualitatively normally distributed with perhaps a very slight right skew.
- Years climbing is clearly right skewed. This makes sense since most people do not spend many, many years rock climbing (or playing any sport for that matter).
- The max grade distribution appears to be slightly left skewed.

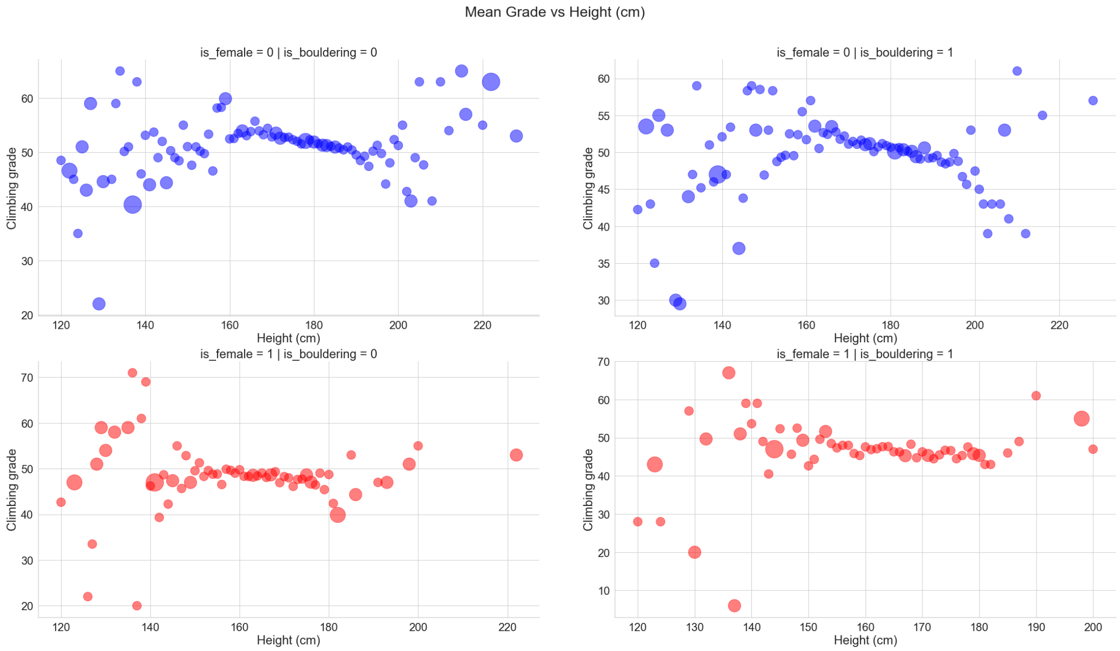
Pairplots are particularly useful for comparing one feature against another. For example, weight and height seem to have a positive correlation with one another (which perhaps can be expected). Also not surprising is that there is an apparent positive correlation with age and the number of years climbing.

Interestingly, aside from 'years_climbing', there does not seem to be any features that are obviously correlated with 'index_grade'.

4.2 Mean 'index_grade' vs 'height'

Scatter plots of the mean 'index_grade' as a function of height for both genders and climbing types are shown below; the sizes of the points correspond to the number of users at a given grade and height. Interestingly, the maximum climbing grades in the dataset are *not* linearly correlated with height. Instead, the data appears somewhat parabolic where the highest grades are obtained by climbers in the middle of the height range; shorter and taller climbers seem unable to reach the same level of ability on average.

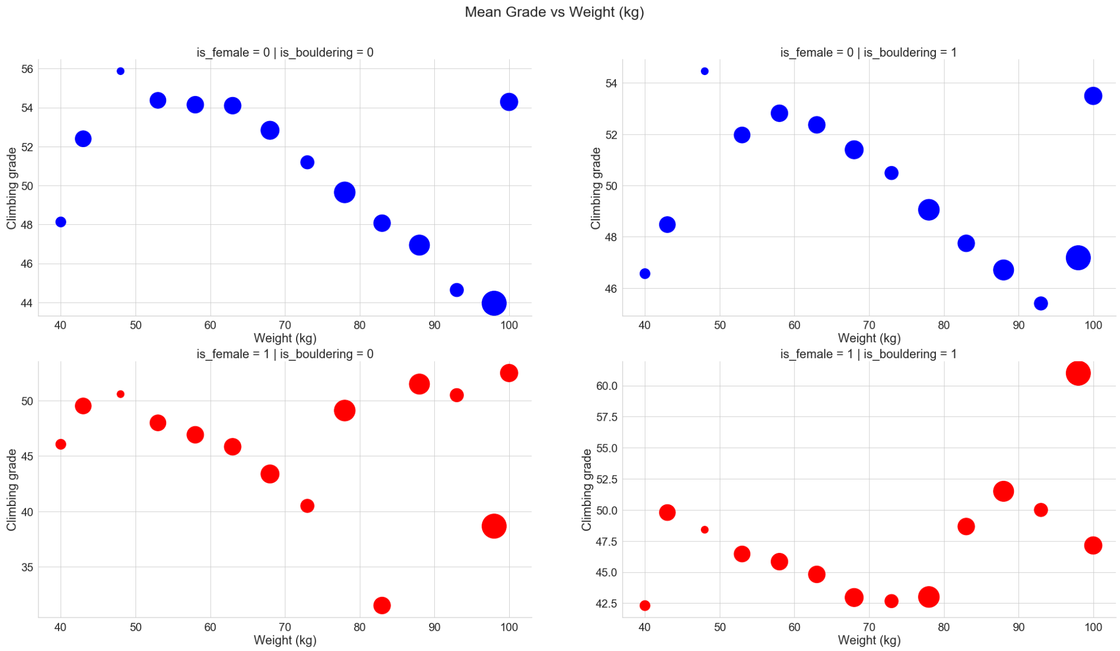
Note that confounding variables and multicollinearity must always be considered before making conclusions. For example, it's entirely possible (and indeed hinted at in the above pairplot) that height and weight are correlated; care must be taken then to not make broad conclusions about the affects of height on climbing ability until further statistical analyses are completed.



4.2 Mean 'index_grade' vs 'weight'

Scatter plots of average 'index_grade' as a function of weight for both genders and climbing types are shown below; the sizes of the points correspond to the number of users at a given grade and weight. For males, weight seems to interact with maximum climbing grade similarly to how heights do (see above section) - there is a sort of parabolic trend. For females, however, the trends are not so clear.

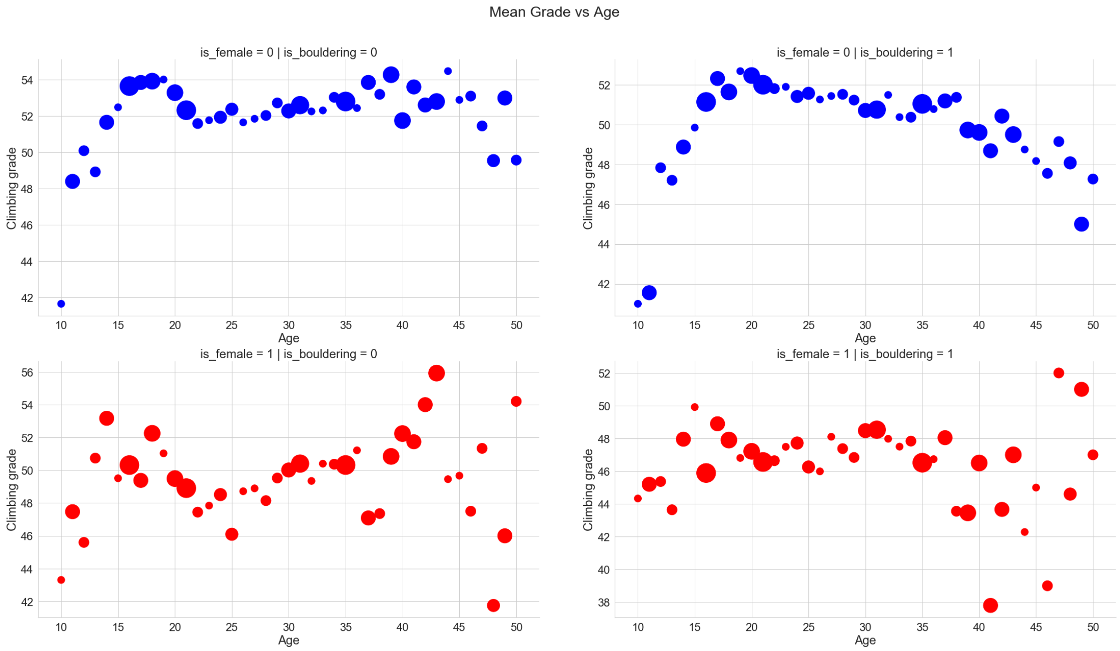
It should be noted that, like height, there may be other confounding or highly correlated variables tied to weight. For example, it might seem like the heavier a climber is, the less successful he/she will become. However, if that increased weight is due to muscle mass instead of fat mass, then it is entirely possible that a heavy climber becomes very accomplished.



4.3 Mean 'index_grade' vs 'age'

Scatter plots of mean 'index_grade' as a function of age for both genders and climbing types are shown below; the sizes of the points correspond to the number of users at a given grade and age. Somewhat similar to the trends above, the affects of age on males seem to be parabolic where the hardest climbing grades occur towards the middle of the age range. The affects on females, however, is not as clear.

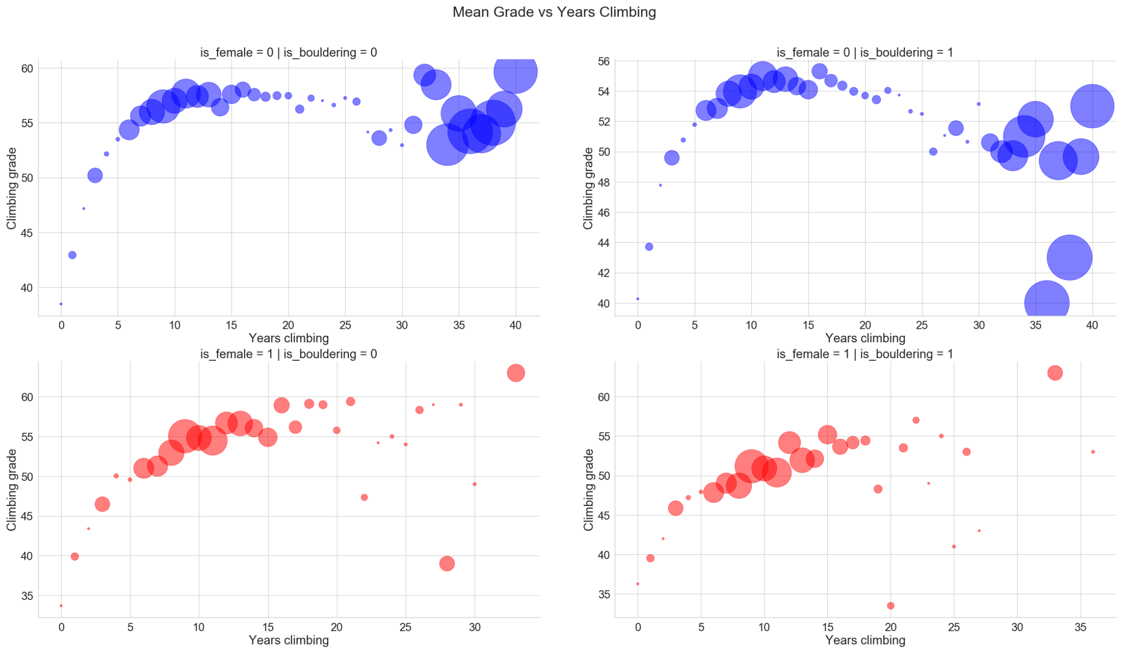
It does make sense that very young climbers are unable to climb as well as older climbers - height, strength, and available time are not necessarily in children's favor. It also makes sense that grades tend to drop off towards the higher ages; this may simply be due to the affects of aging on the body.



4.4 Mean 'index_grade' vs 'years_climbing'

Scatter plots of mean 'index_grade' as a function of years climbing for both genders and climbing types are shown below; the sizes of the points correspond to the number of users at a given grade and years. Unsurprisingly, the maximum grade increases with the number of years - but only up to a certain point. Especially for males, the grades seem to drop as climbers become very experienced. Perhaps this is due to the effects of age since 'years_climbing' and 'age' are more than likely highly correlated.

It is worth noting that there are a large number of males with greater than 30 years of experience but very few females in that category.


Mean Grade vs Years Climbing

is_female = 0 | is_bouldering = 0

is_female = 0 | is_bouldering = 1

is_female = 1 | is_bouldering = 0

is_female = 1 | is_bouldering = 1

5 Analysis

The plots shown above suggest that there are not blatantly obvious correlations between a person's rock climbing ability and his/her height, weight, age, or years of experience. The goal here is then to fit various linear regression models to determine whether inferences can be made. For all analyses, **a significance level of 0.05 is used** and so a P-value <= 0.05 implies that the null hypothesis that no change in the dependent variable ('index_grade') should occur for a unit change in the independent variable (ie the coefficient - or slope - is zero) is rejected. In other words, **the calculated results are considered statistically significant if the corresponding P-values <= 0.05.**

5.1 Discussion - ordinary least squares method

For the analyses below, Python's `statsmodels` package is used. Models are fit using the ordinary least squares `.ols()` method which takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in} + \epsilon_i,$$

where i represents each observation (row in the dataset), y_i are the dependent variables (eg 'index*grade'), $x*{i} $areindependentvariables(eg'height','weight',etc),nisthenumberofindependentvariables,$\epsilon $aretheerrorsateachobservation,and$*\beta*{i}$ are the coefficients to be solved for. The algorithm creates a best fit that minimizes the residual sum of squares.

5.1.1 Required assumptions

The assumptions for this type of method to mathematically work are:

- The model is linear in parameters. That is, the coefficient parameters β_i are linear.
- The sum of the residuals must equal zero.
- There is no multicollinearity. That is, none of the variables are perfectly correlated with each other.
- There is random sampling of the observations.

The first two assumptions are essentially how the algorithm itself works. Regardless of whether the data is linear in parameters, the program will treat them like they are. Likewise, the model will be fit such that the sum of the residuals is zero. Whether accurate or not, a solution will be found! The third assumption (no multicollinearity) is a mathematical requirement for the algorithm to find a solution - the software will crash if a perfect correlation exists due to failed matrix algebra.

As for the fourth assumption: it's not needed for the algorithm to run but it is necessary for the underlying statistics to hold. Unfortunately, the 8a.nu dataset is probably not a random sampling of all climbers. For example, it only sampled from climbers who have access to the internet. Also - and this is just a hunch - it probably samples moreso from younger climbers than older ones. Finally, as mentioned above, 8a.nu climbers are particularly skilled rock climbers. Still, it can be useful to derive results as long as it's understood that they apply to 8a.nu users rather than the climbing population as a whole.

5.1.2 Assumptions for quality results

Once the models are fit, the question becomes whether or not the results trusted. Some of the assumptions made in this regard are:

- The residuals are identically and independently distributed (iid).
    - In addition, *for small samples*, the residuals are normally distributed.
- None of the variables are highly correlated with each other.
- The residuals are not correlated with any of the independent variables.
- The residuals do not predict other residuals, ie there is no pattern to the residual plots.
- The residuals are homoskedastic; that is, they have a constant variance.

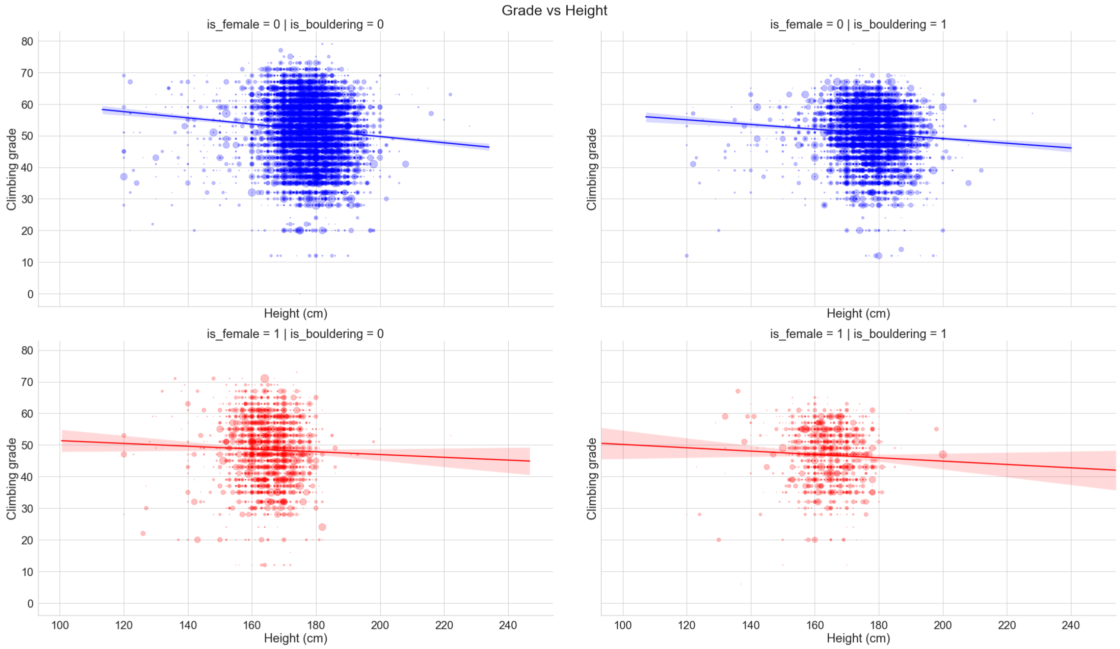
So how can we determine whether the above assumptions hold true?

- This is a rather large dataset so normal distribution of the residuals is probably not necessary. Still, quantile-quantile (QQ) plots are used to qualitatively assess residual normality.
- Consider dropping any independent variables that are highly correlated with each other. **The threshold for a high correlation is set at 0.5 for this analysis**.
    - Also consider the variable inflation factor (VIF) - in addition to correlation factor - before dropping. **The threshold for a high VIF is set at 2 for this analysis**.
- Ensure that the sum of the residuals = 0.
- Ensure that the mean of the residuals = 0.
- Create a scatterplot (and corresponding lowess smooth line) of the residuals as a function of the independent variables.
    - The points should appear random and be scattered about the horizontal axis for residual = 0.
        - If any patterns are present it may be an indication of high variable correlation.
        - If the residuals get larger or smaller as a function of the variable then it indicates heteroskedasticity.

5.2 Univariate analyses

Let's first get a sense of any linear trends found by comparing the dependent variable ('index_grade') against each independent variable separately for all four categories (bouldering female, bouldering male, rope female, and rope male). Note that caution must be used when making conclusions from univariate results since it is often the case that an outcome depends on more than a single independent variable. Further, confounding variables can lead to erroneous conclusions.

As an example, a scatterplot of 'index_grade' vs 'height' is plotted for each category along with the corresponding linear regression lines below. Refer to the analysis notebook for the same plots for 'weight', 'age', and 'years_climbing'.



Interestingly, in all four cases, the linear regression appears to have a negative slope. Notice also that the 95% confidence interval bands are a bit larger for females than for males.

The `ols()` method in the `statsmodels` package is now used to extract the coefficients, standard errors, and P-values; see the summary table below.

```
In [6]: df_results_univariate[df_results_univariate['Variable'] == 'Height (cm)']
```

Out[6]:

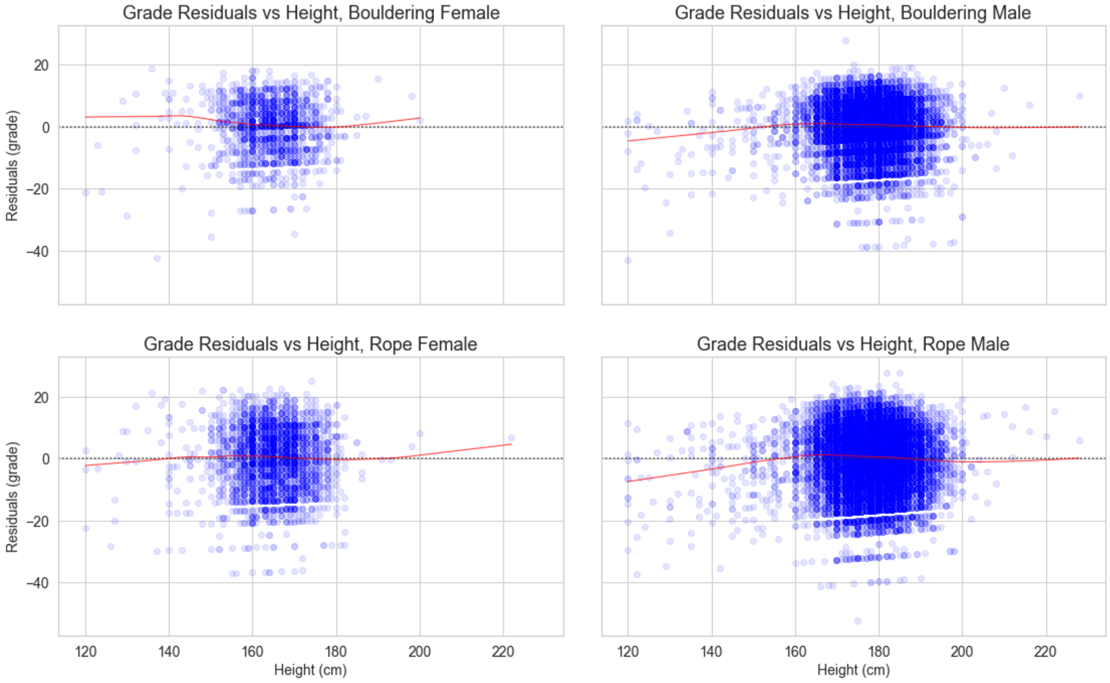
	Variable	Type	Gender	Coefficient	Std_error	P_value	Statistically_significant
0	Height (cm)	Bouldering	F	-0.053	0.028	0.063	No
1	Height (cm)	Bouldering	M	-0.074	0.010	0.000	Yes
2	Height (cm)	Rope	F	-0.044	0.024	0.072	No
3	Height (cm)	Rope	M	-0.098	0.009	0.000	Yes

Indeed, in all cases, the coefficient is negative. However, the P-values for males are both zero while the P-value for females are both above the significance threshold of 0.05. The following statements can thus made:

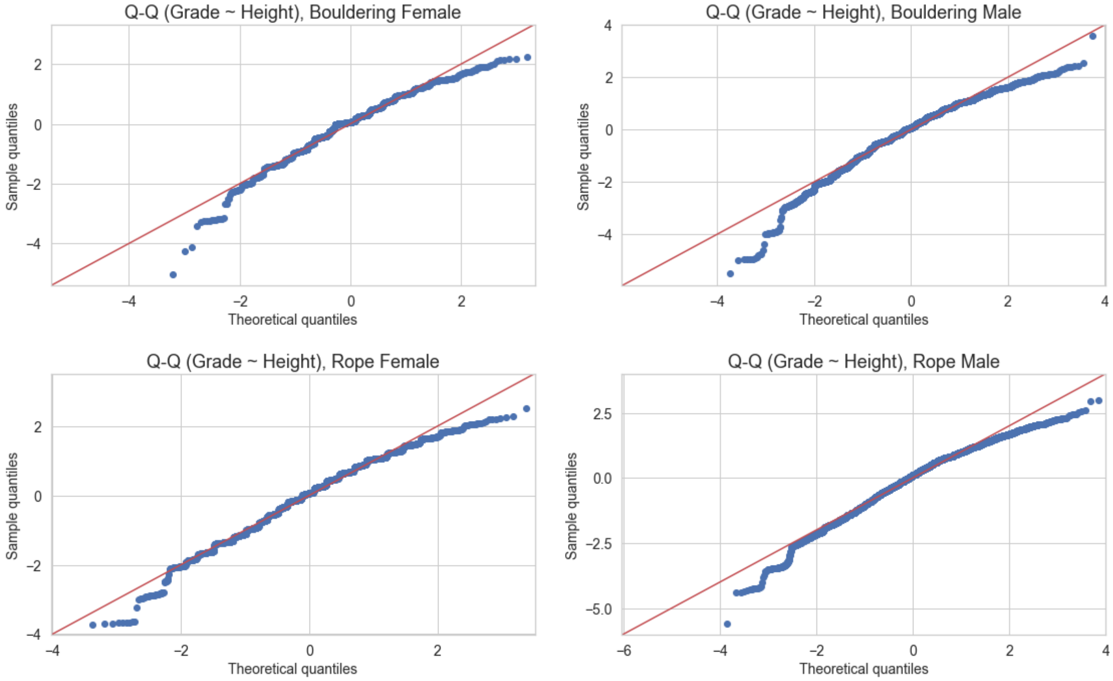
- Height is not a statistically signficant predictor of climbing ability for female.s
- For every cm increase in height, males' best bouldering grade is expected to decrease by 0.074.
- For every cm increase in height, males' best rope climbing grade is expected to decrease by 0.098.

Again, however, confounding variable can be at play and conclusions must wait until multivariate analyses have been considered.

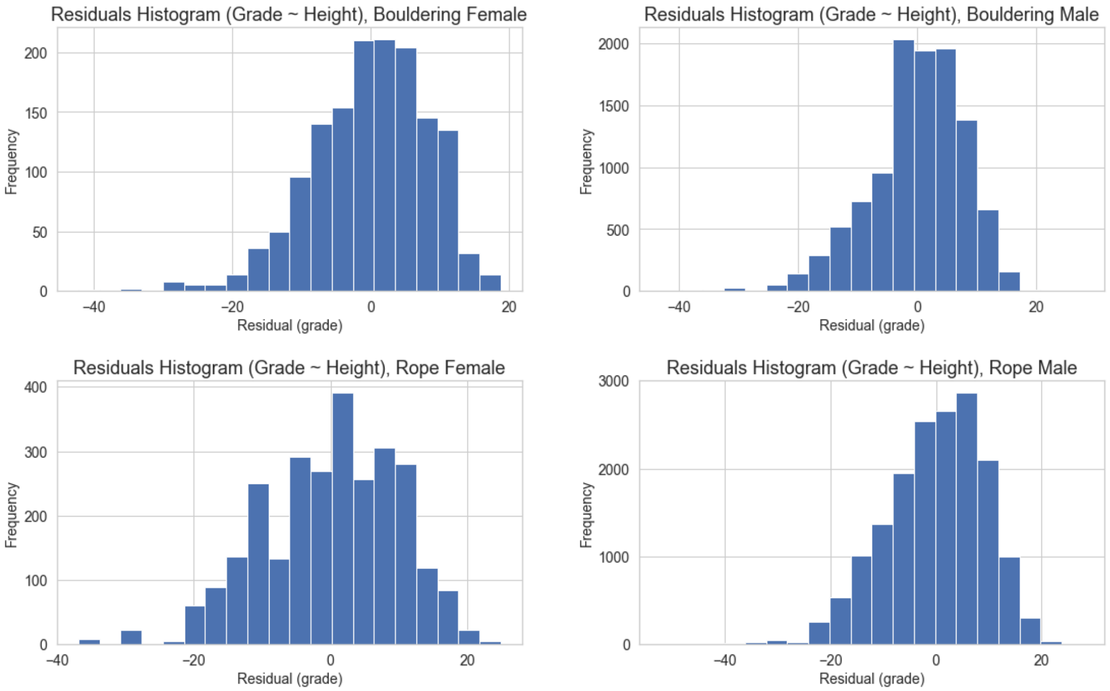
The assumptions made for this analysis are checked by ensuring that the residuals have a mean of 0 and a sum of 0 (they do) and that they are scattered in a random fashion about the x-axis. This can be qualitatively confirmed by inspecting all four residual scatterplots below; indeed, no significant patterns appear to exist as the lowess smooth lines lie approxiamately on the axis at y=0.



To check that the residuals are normally distributed (again, this is not necessarily a requirement for large datasets), one can inspect the so-called quantile-quantile (QQ) plots. These plot the sample quantiles as a function of the theoretical quantiles. The more this scatterplot looks like the ideal 45-degree line (shown in red), the more normally distributed the samples are.

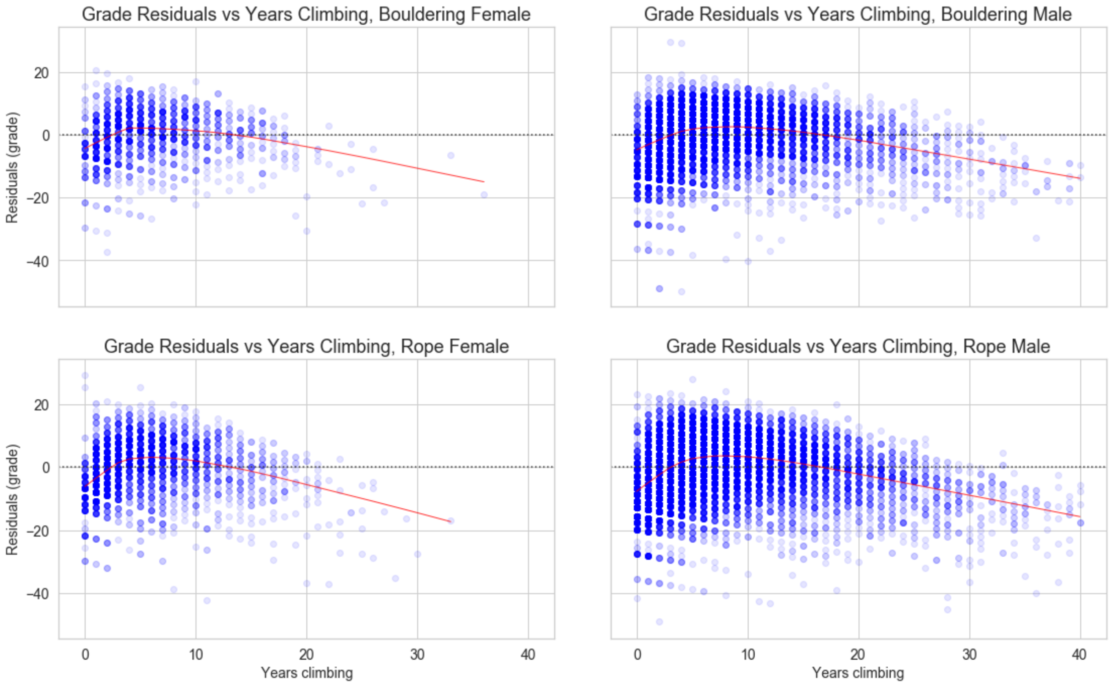


Although this is a strictly qualitative approach, it is a useful one. All four plots show a somewhat normal distribution around the center but then the markers drop off at both ends; this imiplies that the residuals are left-skewed. To check, the residual histograms are shown below.



Indeed, the histograms seem to be mostly normal with a left skew. Although not completely ideal, the datasets used are rather large and so such deviations from ideal are assumed to be acceptable.

For brevity's sake, the plots shown above are not included in this report for the other three independent variables of interest ('weight', 'age', and 'years_climbing'); refer to the analysis notebook if interested. Needless to say, similar results were found with the exception of the residuals plot for the 'years_climbing' variable which appears to have a bit of a pattern to it; see below. This implies that the residuals can be somewhat predicted by 'years_climbing' and may suggest that the linear model fit is not totally appropriate.



The results of all independent variables are shown below.

```
In [7]: df_results_univariate
```

Out[7]:

	Variable	Type	Gender	Coefficient	Std_error	P_value	Statistically_significant
0	Height (cm)	Bouldering	F	-0.053	0.028	0.063	No
1	Height (cm)	Bouldering	M	-0.074	0.010	0.000	Yes
2	Height (cm)	Rope	F	-0.044	0.024	0.072	No
3	Height (cm)	Rope	M	-0.098	0.009	0.000	Yes
4	Weight (kg)	Bouldering	F	-0.062	0.028	0.028	Yes
5	Weight (kg)	Bouldering	M	-0.129	0.008	0.000	Yes
6	Weight (kg)	Rope	F	-0.103	0.025	0.000	Yes
7	Weight (kg)	Rope	M	-0.188	0.008	0.000	Yes
8	Age	Bouldering	F	-0.022	0.031	0.478	No
9	Age	Bouldering	M	-0.068	0.011	0.000	Yes
10	Age	Rope	F	0.037	0.027	0.168	No
11	Age	Rope	M	0.002	0.010	0.814	No
12	Years climbing	Bouldering	F	0.846	0.039	0.000	Yes
13	Years climbing	Bouldering	M	0.410	0.012	0.000	Yes
14	Years climbing	Rope	F	1.160	0.033	0.000	Yes
15	Years climbing	Rope	M	0.625	0.011	0.000	Yes
16	Gender	Bouldering	-	-4.648	0.175	0.000	Yes
17	Gender	Rope	-	-3.812	0.154	0.000	Yes

Perhaps surprisingly, all of the coefficients for climbing grade as a function of height and weight (independently) are negative; for those cases that are statistically significant (that is, the P-value <= 0.05), this means that on average being taller or heavier results in a worse personal highest climbing grade!

On the other hand - and not surprisingly - climbing grade as a function of years climbing is positive (and statistically significant for all four cases). Age, however, is not a significant influencer on climbing grade except for males bouldering in which case it is negatively correlated.

Finally, for both bouldering and rope climbing, the grade vs. gender variable ('is_female') is negative; this can be interpreted as females on average climb at lower grades than males.

Again, it is incredibly important to remember that these findings are the results of a *univariate* analysis where the maximum climbing grades are fit with a linear model assuming it is a function of only a single variable at a time - conclusions must not be drawn until a full multivariate analysis is completed and fleshed out enough to be trusted.

5.3 Multivariate analyses

The above univariate analysis is useful for understanding how the outcome (ie 'index_grade') behaves as a function of a single independent variable at a time. In reality, however, it may very well be that a better model can be fit - still using ordinary least squares regression - by accounting for more than one feature at a time.

5.3.1 Variable reduction

It's not an uncommon phenomenon that seemingly different variables are strongly related to one another and including both in a model can lead to incorrect results. It's thus important to reduce the variables included in the model to only those that are not strongly correlated with one another. For the sake of this analysis the correlation threshold is 0.5.

Another test for multicollinearity is the variable inflation factor (VIF). The threshold VIF for this analysis is 2.

An aside: Note this is somewhat different than the problems caused by confounding variables which are variables not accounted for in the model that may be the true cause of an effect; they are one reason why it's always important to remember that correlation is not the same as causation. A fun example of confounding variables are the studies that show a correlation between the number of people drowning and ice cream sales. Could eating ice cream really cause a statistically significant increase in drownings? While the old wive's tale may blame it on the now-deceased ice cream eaters not waiting 30 minutes before going for a swim, in reality, the confounding variable - and the presumably *true* reason for the correlation - is nothing more than the outside temperature. As the temperature increases, so does the amount of ice cream consumed as well as the number of people who swim (and, consequently, drown).

The correlation matrix and corresponding heatmap for the data at hand is shown below. Note that all variable filters have been set.

In [8]:

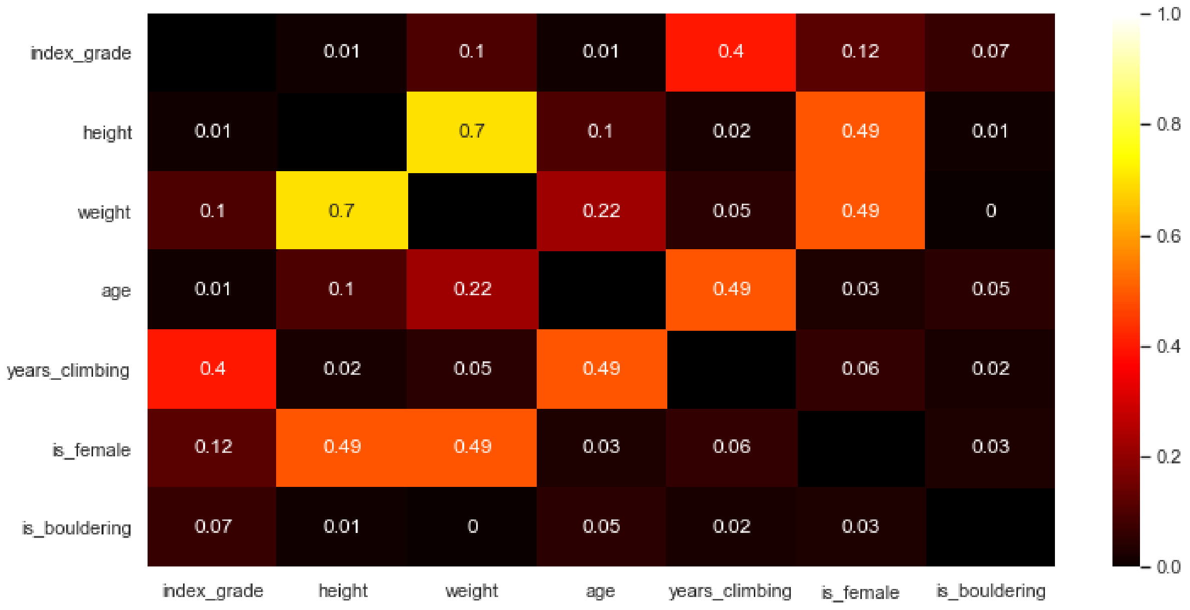
```
# ---- FILTER DATA ----
df_all_filtered = df_all[(df_all['height'] >= 120) &
      (df_all['height'] <= 240) &
      (df_all['weight'] != 0) &
      (df_all['age'] >= 10) &
      (df_all['age'] <= 50) &
      (df_all['years_climbing'] >= 0) &
      (df_all['years_climbing'] <= 40)].drop(columns='id_user')\
[['index_grade','height','weight','age','years_climbing','is_female','is_bould
ering']]

# ---- CORRELATION MATRIX ----
corr_df_all_filtered = round(df_all_filtered.corr(), 2)
corr_df_all_filtered
```

Out[8]:

	index_grade	height	weight	age	years_climbing	is_female	is_bouldering
index_grade	1.00	-0.01	-0.10	-0.01	0.40	-0.12	-0.07
height	-0.01	1.00	0.70	0.10	0.02	-0.49	0.01
weight	-0.10	0.70	1.00	0.22	0.05	-0.49	0.00
age	-0.01	0.10	0.22	1.00	0.49	-0.03	-0.05
years_climbing	0.40	0.02	0.05	0.49	1.00	-0.06	0.02
is_female	-0.12	-0.49	-0.49	-0.03	-0.06	1.00	-0.03
is_bouldering	-0.07	0.01	0.00	-0.05	0.02	-0.03	1.00

In [9]:

```
# ---- CORRELATION HEATMAP ----
plt.figure(figsize = (12,6))
sns.set(font_scale=1)
g = sns.heatmap(abs(corr_df_all_filtered), annot=True,
                cmap='hot', mask=np.eye(len(corr_df_all_filtered)))
g.set_facecolor('black')
```



The highest correlation factor between two variables is 0.7 and occurs with 'height'/'weight'. This is perhaps not terribly surprising and so 'weight' will be dropped as a predictor of 'index_grade' in favor of 'height' (this was a somewhat arbitrary decision as 'weight' could just as easily be dropped, but height is usually discussed in a rock climbing setting more than weight).

All other variables are below the threshold correlation of 0.5 - but 'is_female'/'height', 'is_female'/'weight', and 'years_climbing'/'age' are all very close with a correlation of 0.49. Let's fit a linear model to the (filtered) dataset - including all predictors - and calculate VIFs:

In [10]:

```
lm_all_filtered = smf.ols(
    'index_grade ~ height + weight + age + years_climbing + is_female + is_bouldering',
    data = df_all_filtered).fit()

vif = pd.DataFrame()
variables = lm_all_filtered.model.exog
vif["VIF"] = [vif_(variables, i) for i in range(variables.shape[1])]
vif["Feature"] = lm_all_filtered.model.exog_names
vif
```

Out[10]:

	VIF	Feature
0	619.927690	Intercept
1	2.090677	height
2	2.208518	weight
3	1.413647	age
4	1.336331	years_climbing
5	1.425297	is_female
6	1.006420	is_bouldering

Ignoring the large VIF for the model intercept - which is meaningless and should hypothetically be infinite for a constant - we see that the only VIFs above the threshold of 2 are for 'height' and 'weight' which have already been shown to have a high correlation with one another. As such, only 'weight' will be dropped from the model.

5.3.2 Model fit

An ordinary least squares linear regression is fit to the model for all four cases (female/male and bouldering/rope). To summarize the above variable reduction section, it is decided to fit 'index_grade' as a linear function of 'height', 'age', and 'years_climbing' ('weight' was dropped from the model as it is highly correlated with 'height'). Note that the data was filtered for realistic values of 'height', 'age', and 'years_climbing' (but not for 'weight' since that variable was dropped anyway).

An example of the applied fit is:

```
lm_final_bouldering_female = \
        smf.ols('index_grade ~ height + age + years_climbing',
                data=df_final[
                        (df_final['is_bouldering'] == 1) &
                        (df_final['is_female'] == 1)]
               ).fit()
```

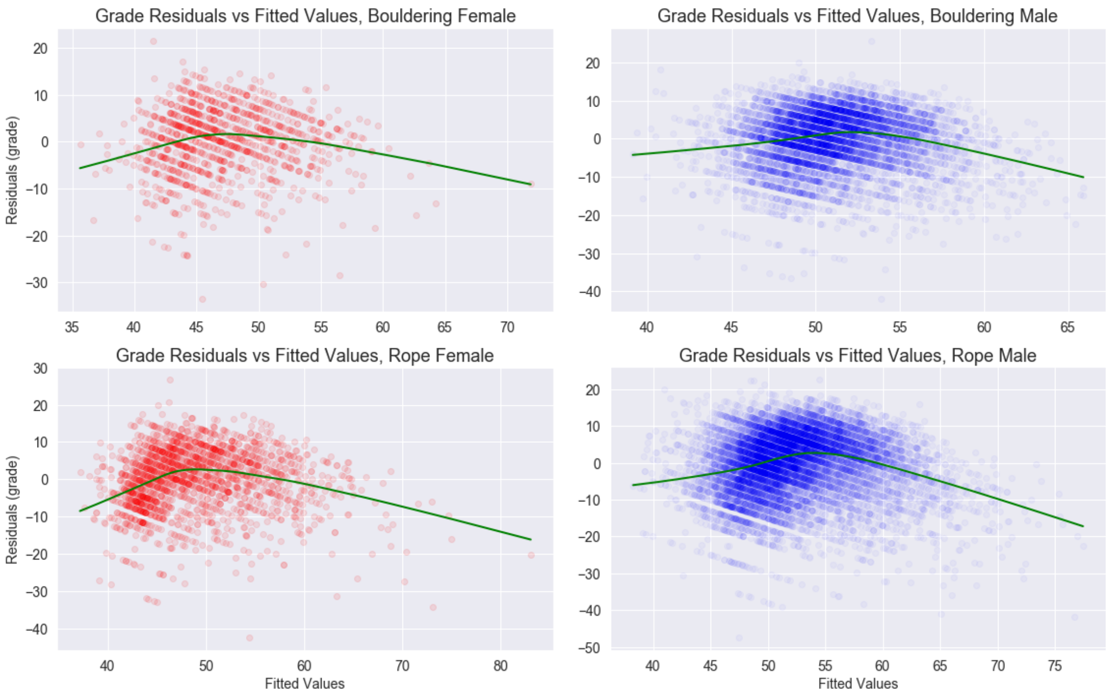
The results dataframe is shown below:

```
In [11]: df_results
```

```
Out[11]:
```

	Type	Gender	Coefficient	Value	Std_error	P_value	Statistically_significant
0	Bouldering	F	Intercept	57.06	4.65	0.000	Yes
1	Bouldering	F	height	-0.05	0.03	0.059	No
2	Bouldering	F	age	-0.27	0.03	0.000	Yes
3	Bouldering	F	years_climbing	0.98	0.05	0.000	Yes
4	Bouldering	M	Intercept	63.02	1.84	0.000	Yes
5	Bouldering	M	height	-0.04	0.01	0.000	Yes
6	Bouldering	M	age	-0.35	0.01	0.000	Yes
7	Bouldering	M	years_climbing	0.66	0.02	0.000	Yes
8	Rope	F	Intercept	50.65	4.11	0.000	Yes
9	Rope	F	height	-0.02	0.03	0.445	No
10	Rope	F	age	-0.23	0.03	0.000	Yes
11	Rope	F	years_climbing	1.30	0.04	0.000	Yes
12	Rope	M	Intercept	68.43	1.75	0.000	Yes
13	Rope	M	height	-0.07	0.01	0.000	Yes
14	Rope	M	age	-0.37	0.01	0.000	Yes
15	Rope	M	years_climbing	0.91	0.01	0.000	Yes

As above, it's always a good idea to plot the residuals to ensure that they are not a function of any variable. In a multivariate model, it's appropriate to plot the residuals as a function of the fitted values. This is done for all four cases and shown below.



The lowess smoothing lines in green show similar trends to the univariate models for 'index_grade' as a function of 'years_climbing'; in all four cases the residuals seem to increase at first and then begin decreasing at fitted values between ~50-60. This non-random pattern in residuals implies that the model is not doing as good of a job as we'd like explaining the dependent variable, ie the predictor variables chosen are not telling the whole story. This may be due to there being a missing variable, a missing higher-order term of one or more variables, or perhaps some interaction between existing variables that is not properly accounted for. Regardless of the cause, the predictors used do not seem to account for all of the non-random error which is what an ideal model would do.

5.4 Discussion of results

As mentioned above, possibly the biggest deviation from the ordinary least squares assumptions for quality is that the residuals do not seem to lie in a random scattering along the fitted values. There is apparently some real-life effect that is not being modeled appropriately, such as a missing variable or some interaction between variables. That being said, if we assume that the model's inadequacies do not drastically change the results obtained, then the following conclusions can be made.

Note:

The values for the coefficient intercept are not terribly useful in this model. The intercept is the value of the dependent variable ('index_grade') when all dependent variables ('height', 'age', and 'years_climbing') are zero. In this model, however, our filtered dataset does not allow heights or ages of zero - not to mention a height of zero is not even possible. As such, the intercept values are really only a product of the model itself but has no real-life meaning here.

5.4.1 Females, bouldering

The subset of results for this category is

```
In [12]: df_results[(df_results['Gender'] == 'F') & (df_results['Type'] == 'Bouldering'
         )]
```

Out[12]:

	Type	Gender	Coefficient	Value	Std_error	P_value	Statistically_significant
0	Bouldering	F	Intercept	57.06	4.65	0.000	Yes
1	Bouldering	F	height	-0.05	0.03	0.059	No
2	Bouldering	F	age	-0.27	0.03	0.000	Yes
3	Bouldering	F	years_climbing	0.98	0.05	0.000	Yes

Height is not a statistically significant factor in climbing ability. Interestingly, age has a negative correlation (a 0.27 reduction in grade for every year gained) whereas the number of years climbing has a positive correlation (a 0.98 increase in grade for every extra year of experience gained).

5.4.1 Males, bouldering

The subset of results for this category is

In [13]:
```
df_results[(df_results['Gender'] == 'M') & (df_results['Type'] == 'Bouldering'
)]
```

Out[13]:

	Type	Gender	Coefficient	Value	Std_error	P_value	Statistically_significant
4	Bouldering	M	Intercept	63.02	1.84	0.0	Yes
5	Bouldering	M	height	-0.04	0.01	0.0	Yes
6	Bouldering	M	age	-0.35	0.01	0.0	Yes
7	Bouldering	M	years_climbing	0.66	0.02	0.0	Yes

In the case of males bouldering, all three coefficients are statistically significant. Height has a negative coefficient (a 0.04 reduction in grade for every cm in height gained). As is the case for females bouldering, age has a negative correlation (a 0.35 reduction in grade for every year gained) whereas the number of years climbing has a positive correlation (a 0.66 increase in grade for every extra year of experience gained).

5.4.1 Females, rope climbing

The subset of results for this category is

In [14]:
```
df_results[(df_results['Gender'] == 'F') & (df_results['Type'] == 'Rope')]
```

Out[14]:

	Type	Gender	Coefficient	Value	Std_error	P_value	Statistically_significant
8	Rope	F	Intercept	50.65	4.11	0.000	Yes
9	Rope	F	height	-0.02	0.03	0.445	No
10	Rope	F	age	-0.23	0.03	0.000	Yes
11	Rope	F	years_climbing	1.30	0.04	0.000	Yes

The results here are very similar to those for females bouldering. Again, height is not a statistically significant predictor of females rope climbing. Meanwhile, age has a negative correlation (a 0.23 reduction in grade for every year older) and experience has a positive correlation (1.39 increase in grade for every year of experience gained).

5.4.1 Males, rope climbing

The subset of results for this category is

```
In [15]: df_results[(df_results['Gender'] == 'M') & (df_results['Type'] == 'Rope')]
```

Out[15]:

	Type	Gender	Coefficient	Value	Std_error	P_value	Statistically_significant
12	Rope	M	Intercept	68.43	1.75	0.0	Yes
13	Rope	M	height	-0.07	0.01	0.0	Yes
14	Rope	M	age	-0.37	0.01	0.0	Yes
15	Rope	M	years_climbing	0.91	0.01	0.0	Yes

The results here are very similar to those for males bouldering. Again, height has a negative correlation with grade (a 0.07 reduction in grade for every cm of height gained). Age has a negative correlation (a 0.37 reduction in grade for every year older) and experience has a positive correlation (0.91 increase in grade for every year of experience gained).

6 Suggestions for improvement

Some recommendations to potentially improve this analysis include:

- Drop 'age' from the model (since it is somewhat correlated with 'years_climbing') and see if that reduces the pattern found in the residual plot.
  - Drop 'years_climbing' instead of 'age' and see how that changes the results.
- Explore and account for any outliers that exist.
- Consider applying linear fits to non-linear variables. This may be especially useful for those variables that seem to result in the maximum climbing grade having a somewhat parabolic shape (ie height and weight).
  - Note that a linear fit only assumes that the model is linear with respect to the coefficients β_{in} - it says nothing about the variables themselves which may indeed be transformed, eg a polynomial linear fit of the form $y_i = \beta_0 + \beta_1 x_{i1}^2 + \beta_2 x_{i2}^2 + \ldots + \beta_n x_{in}^2 + \epsilon_i$.
- It might be interesting to perform a principal component analysis. This would result in loss of interpretability of the variables, but some insight might be gained by looking at the most important variables to the principal components.
- Use datetime information instead of simply extracting the years to get more accurate 'age' and 'years_climbing' data.
- Explore the *rate* at which each user progresses by considering time series data.
- Gather more data! Users of 8a.nu tend to be rather strong climbers and so this analysis leans very much towards those who have made rock climbing a significant part of their life. It therefore may not be appropriate to draw conclusons about more casual rock climbers.