Popularity of Munros

Filip Balucha, Advaith Sai Maddipatla, Tudor Finaru

15th June 2021

1 Overview

This paper answers a few key questions regarding Munro hiking by applying statistical methods on data that is freely available online. It aims to explore the reasons behind the popularity of specific hills, and then provides an appropriate clustering of Munros according to their features. To achieve this, we use techniques such as linear regression, principal component analysis, and K-Means clustering. These techniques yield good results, enabling us to reach a number of solid conclusions. Firstly, we established the existence of a statistically significant relationship between Munro altitude and number of ascents. Furthermore, we managed to identify several other factors that have an influence on Munro popularity – for example the numbers of neighbouring Munros, hotels, and the distance to the nearest city. Finally, we divided the Munros into four well-defined clusters according to their features.

2 Introduction

Context and motivation We aim to shed light on the patterns of human behaviour and preferences surrounding Munro hiking – an increasingly popular pastime [1]. Many of these hills hold an almost sacred spot in the Scottish psyche, with some locals attaching their very identities to these pieces of timeless, rugged landscape [2]. Considering the extensive data on Munros that is available online – as well as the lack of previous scientific literature on the subject – we deemed this topic worthy of investigation.

Previous work There has not yet been a comprehensive study to investigate the patterns that we wish to observe in this paper. However, there certainly are examples of similar studies carried out about other mountain ranges – a particularly thorough one has been conducted on the Italian Dolomites [3]. Among other things, the way the authors nested peaks according to their features has been a useful source of inspiration for our own clustering and regression models.

Objectives We are setting out to answer the following questions:

- Is there a statistically significant relationship between altitude of Munro and the number of ascents?
- What other factors have a significant impact on Munro popularity?
- Can Munros be clustered according to their features?

3 Data

Data provenance For the purposes of this paper, we used three datasets:

- **WalkHighlands (WH)**, from which we extracted data on the popularity of Munros as well as the accommodation facilities in their vicinity. The data on ascents relies on contributions from registered users – they can select and rate the Munros they have climbed. We retrieved the data by

scraping the main Munro tables and subpages; additionally, we manually copied the number of facilities for each type of accommodation from the accommodation subpages. These methods were chosen after a careful assessment of the Terms and Conditions (T&Cs) [4].

- **The Database of British and Irish Hills (DoBIH)**, from which we extracted geographical data on Munros. The origins of this data can be traced in a 'series of articles in Marhofn and Relative Matters magazine', according to the website. The T&Cs for this dataset impose 'no restrictions on use of the data by third parties', so long as the terms of the Creative Commons license are respected [5]. Retrieving the data was easy, as it only involved downloading a readily available CSV file.
- **Simplemaps Cities Database (SCD)**, from which we extracted data on British cities' location and population. The data comes from the US National Geospatial-Intelligence Agency, and is freely available under the MIT license [6]. Again, data retrieval amounted to downloading a CSV file.

Web scraping was performed in accordance with James Densmore's rules for ethical web scraping [7]. We only scraped data where no API was accessible, we always provided a 'User Agent' string to make our intentions clear to the site owner – as well as to provide a way to contact us – and we requested data at the reasonable rate of at most 1 request per 10 seconds. Finally, we only saved the data we absolutely needed, and did not pass it off as our own.

Table 1: The variables present in the final dataset.

Variable	Type	Description
name	string	Name of the Munro
altitude	integer	Altitude of the Munro [m]
ascent_count	integer	Number of ascents by WH users
region	string	Region in which the Munro is located
county	string	County in which the Munro is located
island	string	Island on which the Munro is located
latitude, longitude	float	Coordinates of the Munro's peak
bb_count	integer	Number of B&Bs in the region
hotel_count	integer	Number of hotels in the region
hostel_count	integer	Number of hostels in the region
cottage_count	integer	Number of cottages in the region
camping_count	integer	Number of camping and glamping sites in the region
nearest_city_dist	float	Distance to the city closest to the Munro [km]
nearest_city_population	integer	Population of the city closest to the Munro
neighbor_count_<r_1>_<r_2>	integer	Number of neighbouring Munros within $(r_1, r_2]$ km
population_<r_1>_<r_2>	integer	Population within $(r_1, r_2]$ km

Data description After processing the data and joining the resulting tables, we obtained a dataset with 282 records and variables as outlined in Table 1.

Data processing For the DoBIH data, we only chose Munros. We kept only the relevant columns, which were name, altitude, island, county, and latitude and longitude.

For the SCD data, we first filtered out cities not in the UK. Then, for each Munro, we computed the distance to and population of the nearest city. Distance was determined using the haversine formula, which gives the distance between two points on the surface of Earth given their longitude and latitude. To represent the various distances from which hikers come, we also computed the total population within increasing ranges of distances from the Munro – $(0, 25]$, $(25, 50]$, $(50, 75]$, and $(75, 100]$ km.

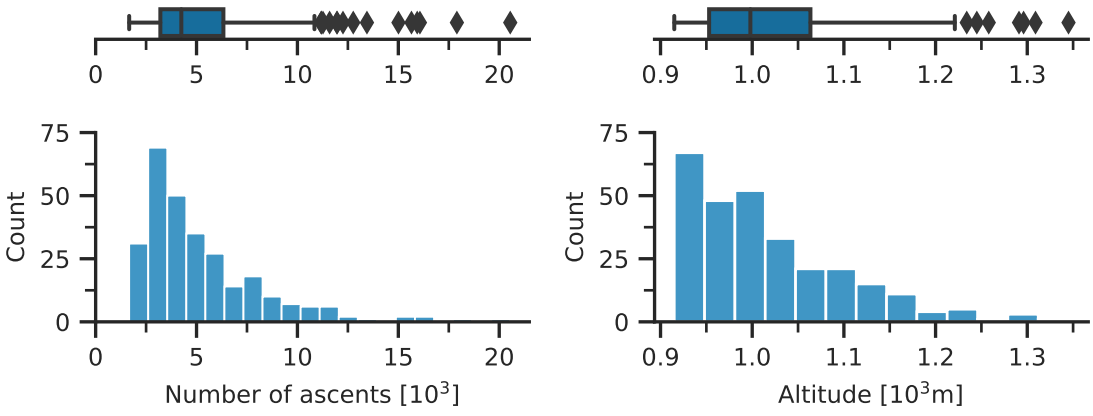
We treated Mull and Skye – the two islands with Munros – separately, since neither Portree nor Tobermory, their corresponding largest settlements, occur in SCD. Considering that Skye is connected to the mainland by road, we considered the impact of mainland cities on the popularity of Skye Munros. However, since Mull is isolated from the mainland, we replaced all city-related values with NaN.

Merging the WH and DoBIH datasets was challenging. First, we created a unique key for each Munro in both tables. The key of choice was a stringified tuple consisting of name and altitude, since only these two fields are available in both datasets. Unfortunately, they did not always match exactly, and some particular cases needed to be handled manually (e.g. the name 'Carn Dearg' appears in DoBIH three times). For the remaining data, we matched each key from WH with the closest key in DoBIH based on string edit-distance, which is a measure of difference between two strings. We verified the matched key pairs by inspecting manually any pair for which the difference in altitudes was greater than 10m (i.e. greater than $\approx 1\%$ of 1,018m – the mean Munro altitude) or if the intersection of their names did not match either name. We also ensured that all keys were unique. We then merged the data and removed unnecessary fields such as the aforementioned keys.

After merging, we computed the number of neighbours for each Munro, i.e. the number of Munros located within a 'ring' comprising areas located between 0 and 5 km, and 5 and 20 km away from the Munro's peak. The former represents Munros that could be reached on foot as part of a single trek, while the latter other Munros that are within driving distance and could be visited during a single trip. We only considered neighbouring Munros located on the same land mass, since we assume that a climber who is to climb multiple Munros in a restricted area will not want to drive to another island to do so (e.g. Skye).

4 Exploration and analysis

This section provides a summary of our data analysis process. It includes key visualisations which helped us reach our conclusions, as well as interpretations of the results (corrected to 3 significant figures) at each step of the process. We split the section into 3 parts – each corresponds to one of the questions that we seek to answer in our paper.



(a) The distribution of number of ascents.

(b) The distribution of altitude.

Figure 1: The respective distributions of number of ascents and altitude.

4.1 Assessment of the statistical significance of the relationship between Munro altitude and the number of ascents

To get an initial view of the data, we visualise the distributions of Munro ascent count and altitude (Figure 1). We observe that most Munros have been ascended 2,500 to 3,000 times. The associated

boxplot shows that a Munro with ascent count above 11,000 is an outlier (Figure 1a). We also notice that most Munros have an altitude just below 1,000m, with a altitude above 1,230m being an outlier (Figure 1b).

In order to better understand and identify some key outliers, it is worth having a look at the scatterplot of ascent count and altitude shown in Figure 2 (note the different increment on the two axes). It comes as little surprise that the two most popular Munros are Ben Lomond and Ben Nevis. In fact, Ben Lomond takes the top spot despite being less than 1,000 metres high. This could likely be explained by its extraordinary popularity with the people of Glasgow, Scotland's most populous city. It is within easy reach of said city, and is well-known as an accessible spot of natural beauty for Glaswegians. On the other hand, Ben Nevis' significant popularity was expected given its status as the tallest mountain in Britain. Its relatively isolated location in the North-West of Scotland does little to deter people from all over the UK from attempting the comparatively strenuous hike.

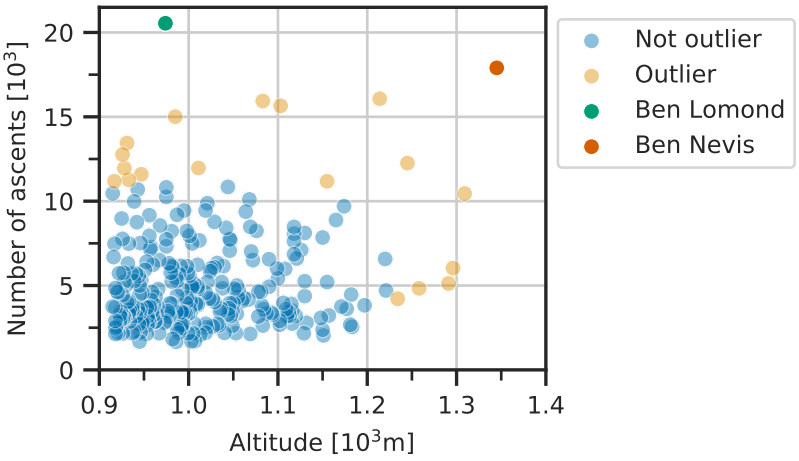


Figure 2: The scatterplot of ascent count and altitude with key outliers highlighted.

We are now in a better position to answer our initial question. There appears to be a positive relationship between Munro altitude and ascent count (Figure 2). Moreover, the outliers (e.g. Ben Nevis) should also contribute to that relationship. To formally explore the relationship, we apply Ordinary Least Squares (OLS) linear regression.

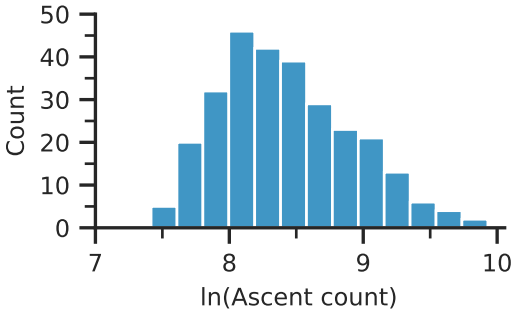


Figure 3: The distribution of number of ascents upon log-transformation.

Before we apply linear regression, we make sure that the dependent variable is not highly skewed. The distribution of ascent count exhibits a right skew (Figure 1a). It is approximately log-normal because the histogram increases to its mode quite quickly and decreases thereafter. Furthermore, the mode is less than the median (from the boxplot). To fix this, we log-transform the dependent variable, which yields a distribution that is less skewed and looks approximately normal (Figure 3). Since the datapoints cluster

far from the origin (Figure 2), we normalise the independent variable to aid numerical stability – i.e. center it around 0. Thus, we arrive at the linear model:

$$\ln(y) = \beta_0 + \beta_1 x^*$$

where x^* is normalised altitude and y is ascent count. To evaluate the statistical significance of the relationship, we define the associated null and alternate hypotheses as follows:

$$H_0 = \text{The coefficient of altitude in the linear model is equal to zero, i.e. } \beta_1 = 0$$
$$H_a = \text{The coefficient of altitude in the linear model is not equal to zero, i.e. } \beta_1 \neq 0$$

We aim to reject the null hypothesis at the conventional 5% level. Applying linear regression on the processed data yields a prediction that exhibits a visible increasing trend (Figure 4).

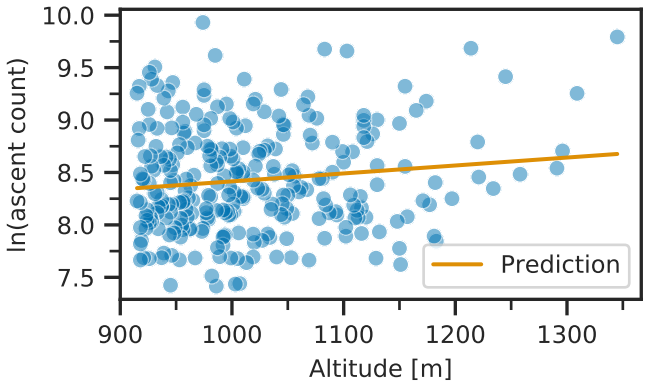


Figure 4: The scatterplot of ln(ascent count) against altitude along with fitted values.

To diagnose the fit numerically, we use the output of statsmodels. The p-value of 0.0340 for altitude tells us that there is a $\approx 3.40\%$ probability that the relationship between altitude and ascent count may be due to chance. Since $0.0345 < 0.05$, we reject the null hypothesis that the coefficient of altitude in the model is 0 at the 5% level. Thus, there is a statistically significant relationship between altitude and ascent count. However, we observe that at 0.0159, the R^2 value is quite low. This indicates that the model does not fit the data well. This motivates the use of further predictors to aid our analysis.

Since we normalised altitude, it has mean 0. Thus for a Munro of mean altitude, the expected value of ln(ascent count) is given by the intercept $\beta_0 \approx 8.4286$. The expected ascent count for a Munro of mean altitude is then $e^{\beta_0} = e^{8.4286} \approx 4,580$.

We now interpret the slope $\beta_1 \approx 0.0008$. The regression is of the form $\ln(y) = \beta_0 + \beta_1 x^*$ where $x^* = x - \bar{x}$ is the value of altitude. Then $\ln(y_1) = \beta_0 + \beta_1 x_1^*$ and $\ln(y_2) = \beta_0 + \beta_1 x_2^*$ for two observations. Then

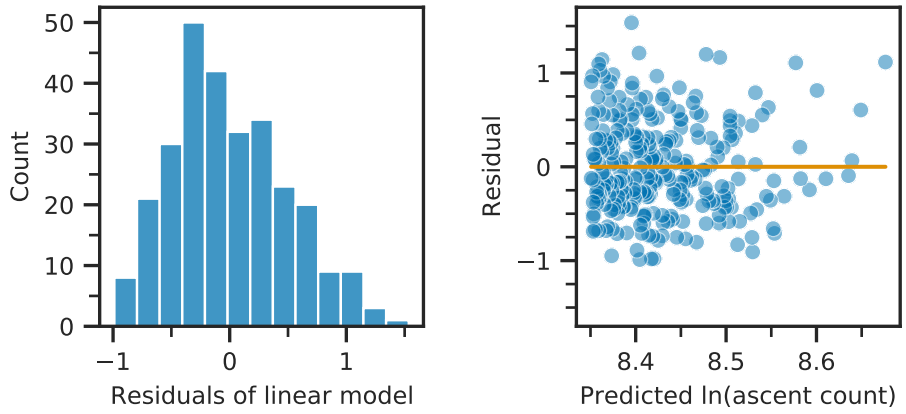
$$\beta_1(x_2 - x_1) = \beta_1((x_2 - \bar{x}) - (x_1 - \bar{x})) = \beta_1(x_2^* - x_1^*) = \ln(y_2) - \ln(y_1) = \ln(y_2/y_1)$$

Consider a unit increase in altitude. Then $x_2 - x_1 = 1$, so that $e^{\beta_1} = y_2/y_1$, which can be rewritten as

$$e^{\beta_1} - 1 = (y_2 - y_1)/y_1$$

A unit increase in altitude thus leads to an $\approx e^{0.0008} - 1 \approx 7.58 \times 10^{-4}\%$ increase in ascent count. The confidence interval of $[5.57 \times 10^{-5}, 1.00 \times 10^{-3}]$ tells us that in 95% of all samples that could be drawn, the increase in ascent count will be within $[5.70 \times 10^{-3}, 1.00 \times 10^{-1}]\%$.

Finally, to validate our approach, we inspect the distribution of residuals (Figure 5a). They are approximately normally distributed, which is an assumption of linear regression. Furthermore, the variance of residuals does not change as a function of the predicted values of the independent variable (Figure 5b). Thus, the dataset is not heteroscedastic. Therefore, a linear model was a reasonable choice for the problem at hand.



(a) The distribution of the residuals.

(b) The scatterplot of residuals against predictions.

Figure 5: The residual plots for the univariate linear regression model.

4.2 Other factors that have a significant impact on Munro popularity

Moving on to our second question, we are motivated by the applicability and interpretability of the linear model in the previous section. To accommodate more features and improve our model, we will be using its generalisation – multivariate OLS linear regression.

We first standardise the dataset – this helps with numerical stability because some features have a different range of values (e.g. population-related values are up to four orders of magnitude greater than hotel_count).

To obtain more interpretable results, we perform feature selection. The tool we are using – scikit-learn's SelectKBest – expects continuous data, so we remove categorical or boolean variables. We score the features using the scoring function f_regression, which computes the correlation coefficient between the regressor and the target [8]. It outputs a p-value, which corresponds to the null hypothesis that there is no linear interaction between between the regressor and the target. We pick the regressors whose p-value falls below a certain threshold. As in the previous section, we choose the threshold to be 0.05, such that the null hypothesis can be rejected at the 5% level. That is, we pick only those regressors for which there is a 5% probability that the linear interaction with the target is due to chance.

Before applying multivariate OLS linear regression, we ensure that there is no significant multicollinearity between our features which could reduce numerical stability, jeopardising the quality of our model. We consider a correlation coefficient higher than 0.8 as indicative of high collinearity. We consult the correlation matrix for the selected independent features – all coefficients are within said threshold, hence we keep all features. We also try reintroducing boolean and categorical features. This leads to numerical issues – perhaps due to said features acting as proxies to the continuous ones. Finally, numerical instability and more features reduce interpretability, so we do not consider these features.

We reapply OLS regression to continuous features. We will use each of the standardised features as regressors. The target will be the log-transformed ascent count, for the same reason as in the previous section. We find that the variables population_25_50 and hostel_count both have quite high p-values of over 0.2. This is likely due to multicollinearity with other population- and accommodation-related features, and it is therefore safe to remove it. After reapplying OLS, we obtain a good fit with R^2_{adj} of 0.525 – much better than in the univariate case. In addition, all the features in the model have a p-value of less than 0.01, which indicates the obtained coefficients are reliable.

We now wish to interpret the output of statsmodels. Since the regressors are standardised and the target is log-transformed, we first need to transform them to an interpretable format. We are interested in the impact of a unit change in a regressor on the target. The linear regression is of the form

$$\ln(y) = \beta_0 + \beta_1 z^{(1)} + \cdots + \beta_n z^{(n)}$$

where $z^{(i)}$ is the standardised i-th regressor. To interpret the response in the target to a unit increase in the i-regressor, we let $\forall j \neq i,\ x^j := 0$. Then we have $\ln(y) = \beta_0 + \beta_i z^{(i)}$. Now

$$\beta_i(z_2^{(i)} - z_1^{(i)}) = \ln(y_2) - \ln(y_1) = \ln(y_2/y_1) \tag{1}$$

Due to standardisation, $z^{(i)} = (x^{(i)} - \bar{x}^{(i)})/\sigma_{x^{(i)}}$. We thus rewrite (1) as: $\beta_i(x_2^{(i)} - x_1^{(i)})/\sigma_{x^{(i)}} = \ln(y_2/y_1)$. By performing steps similar to those in previous section, we arrive at:

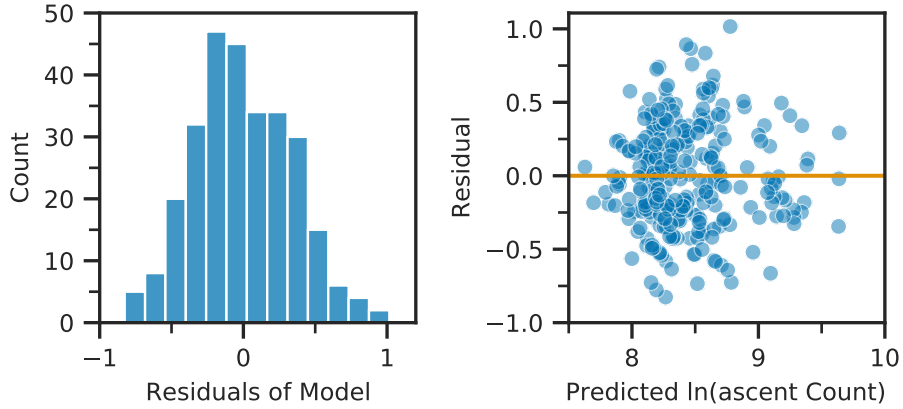
$$\exp(\beta_i/\sigma_{x^{(i)}}) - 1 = (y_2 - y_1)/y_1$$

for a unit increase in an original feature. We now transform all coefficients accordingly (Table 2).

Table 2: The values and confidence intervals of transformed coefficients.

Feature	Response [%]	Response (CI Lower) [%]	Response (CI Upper) [%]
altitude	8.74×10^{-2}	3.76×10^{-2}	1.37×10^{-1}
hotel_count	-7.08×10^{-1}	-9.84×10^{-1}	-4.32×10^{-1}
neighbor_count_5_20	-1.31	-1.81	-8.05×10^{-1}
nearest_city_dist	-5.12×10^{-1}	-8.88×10^{-1}	-1.35×10^{-1}
pop_0_25	2.55×10^{-3}	1.32×10^{-3}	3.78×10^{-3}
pop_50_75	3.27×10^{-5}	2.42×10^{-5}	4.11×10^{-5}
pop_75_100	3.53×10^{-5}	2.81×10^{-5}	4.25×10^{-5}

The intercept gives the expected value of ascent count on a log scale when all regressors are 0. The intercept is ≈ 8.4286 which implies the expected number of ascents is then $\approx e^{8.4286} \approx 4580$ ascents. The most dominant feature is neighbor_count_5_20. Every extra neighbouring Munro within 5-20km reduces the ascent count by $\approx 1.31\%$. The other dominant features are the number of hotels, wherein each extra hotel reduces ascent count by $\approx 0.708\%$, and the distance to the nearest city, such that each extra kilometer distance leads to a $\approx 0.512\%$ decrease in ascent count. As indicated in the first section, altitude also has an impact on ascent count, albeit smaller. Namely, each meter of altitude increases ascent count by $\approx 0.874\%$. There is also a slight positive impact on ascent count associated with population within 0-25km and 50-100km.



(a) The distribution of the residuals.

(b) The scatterplot of residuals against predictions.

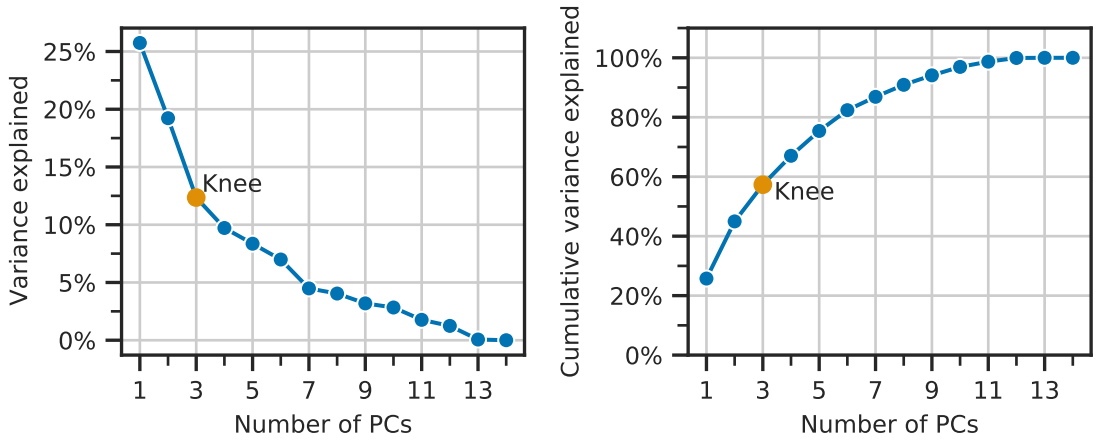
Figure 6: The residual plots for the multivariate linear regression model.

Again, to validate our approach, we inspect the distribution of residuals (Figure 6a). They are approximately normally distributed, which is an assumption of linear regression. Furthermore, the

variance of residuals does not change as a function of the predicted values of the independent variable (Figure 6b). Thus, the dataset is not heteroscedastic. Therefore, we again conclude that a multivariate linear model was a reasonable choice for the problem at hand.

4.3 Clustering Munros according to their features

Moving on to our final question, we aim to cluster Munros according to their features. This may help us discover a sub-structure within the dataset and help us understand it better. To this end, we use K-Means Clustering; unfortunately, as a distance-based method, K-Means suffers from the 'curse of dimensionality' and will perform poorly when applied to a dataset with many variables. Thus, before we apply K-Means, we reduce the number of independent variables using Principal component analysis (PCA).



(a) Scree plot.

(b) Cumulative scree plot.

Figure 7: Visual diagnostics for PCA.

In order to determine the appropriate number of components to be used, we plot a scree plot to find the knee. It represents the point at which adding more principal components would not explain a significant variance in the data. The first 3 principal components help explain a considerable amount of 60% variance (Figure 7b). We thus transform the dataset using the first 3 principal components.

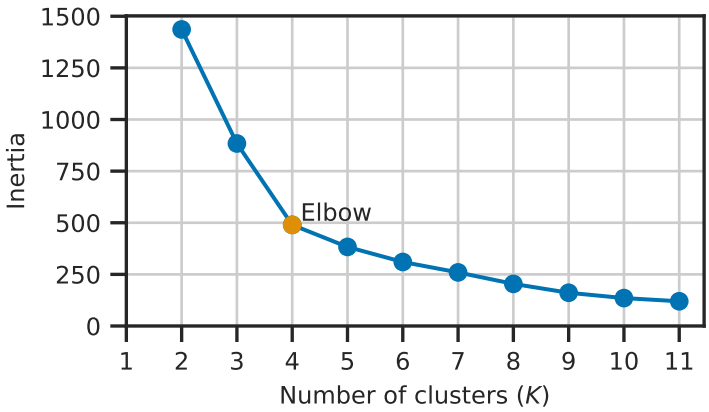


Figure 8: The scree plot for K-Means.

In order to determine the appropriate number of clusters, we inspect the inertia (i.e. the sum of within cluster sum of squared errors) of our K-Means model with increasing number of clusters (Figure 8). To

find the optimal number of clusters and avoid overfitting, we use the elbow method. As the number of clusters increases, they can explain a greater fraction of variance, reducing inertia. However, there will be a point (the elbow) after which the inertia will decrease significantly [9] – resulting in overfitting. There is a clear elbow at $K = 4$ (Figure 8); thus, the optimal number of clusters is 4. We now perform K-Means. Since we cluster in 3 dimensions, we plot the predictions on a 3D plot and indicate the clusters' centroids (Figure 9). We now use the information about clusters to inspect the features that contribute to each cluster. To obtain a summary measure for each feature, we compute its mean for each cluster.

Cluster 1: The average Munro in Cluster 1 has ~2,000 inhabitants within 25km, while the nearest city has ~10,000 inhabitants and is ~35km away. This cluster has the largest population 25-100km away out of all clusters. At about 60, it also has the largest number of hotels. It also has a fair amount of bed and breakfast accomodations (about 50). It also has the lowest the number of neighbouring Munros: slightly above 2. Since there are very few neighbouring Munros and there are a lot of (presumably relatively high-end) hotels near these Munros, we expect this cluster to comprise of fairly 'exclusive' Munros which are suitable for visitors from a larger city (which makes them quite popular, too). Two examples of Munros that fall into this category are Ben Lomond and Ben Lawers.

Cluster 2: The average Munro in Cluster 2 does not have a large city within 100km. The nearest city is about ~45km away with a population slightly higher than ~7,500. The dominant accommodation type for this cluster are cottages and camping sites, and there are relatively few B&Bs, hotels and hostels. It has about 25 neighboring Munros within 20km. We are therefore dealing with a region with a fairly high number of Munros, with relatively few accommodation facilities and also quite far away from any major cities. Munros that belong to this cluster are fairly isolated and slightly less popular – two examples would be Mount Keen and Ben Hope.

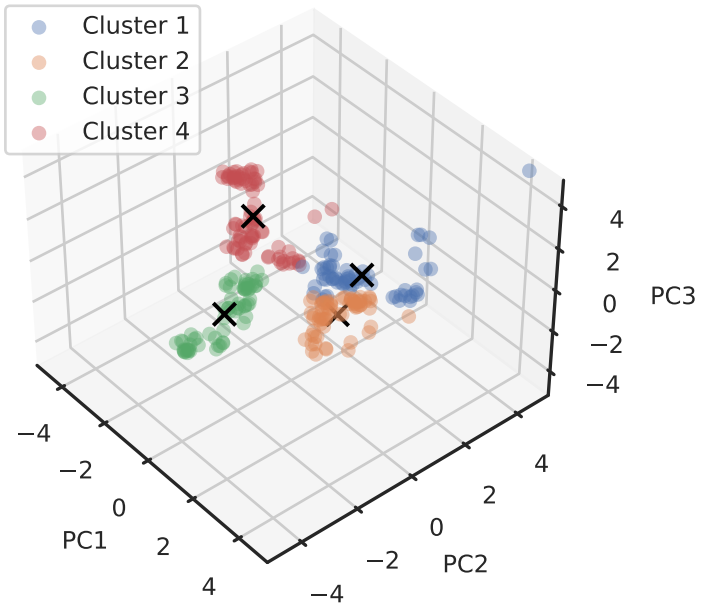


Figure 9: The Munro clusters as identified by K-Means.

Cluster 3: For the average Munro in Cluster 3, the closest city is more than 50km away, but has ~20,000 inhabitants – the most across all clusters. It has the largest number of cottages and campings out of all clusters, with relatively fewer B&Bs, hotels and hostels. At more than 4, it also has the largest number of neighboring Munros within 5km. Visitors to Munros included in this cluster might wish to hike several neighboring peaks during their trip and stay at a cottage or camping site. At ~6,000 ascents, these Munros are fairly popular – perhaps some inhabitants of the nearby large city are regular visitors, or the high concentration of nearby Munros makes the entire area a popular hiking destination. The latter

point is supported by the inclusion of peaks such as Cairn Gorm and The Cairnwell in this cluster; both of these are located in the Cairngorms Natural Park, one of Scotland's prime areas for hiking.

Cluster 4: The average Munro in Cluster 4 has the largest number of people within 25km across all clusters, but relatively few people beyond that. The nearest city is ~25km away and has a population of about 10,000. Other larger settlements lie beyond 75km away. The dominant accommodation types are B&Bs and hostels. At almost 25, it has the largest number of neighboring Munros within 5-20km. Two examples of peaks that belong to this cluster are Ben Nevis and Stob Dearg; these are both located near Fort William and Glencoe in North-West Scotland, an area famous for its high concentration of dramatic Munros – which certainly ensures a large number of neighboring peaks. Apart from a few minor nearby towns that also present a number of accommodation facilities, Munros in this cluster seem to be quite isolated compared to others – and this definitely applies to the two aforementioned examples.

5 Discussion and conclusions

Summary of findings First of all, we have proven that there exists a statistically significant relationship between Munro altitude and number of ascents – taller Munros can generally be expected to exhibit higher ascent counts. We have also explored the influence of other factors on the number of ascents. Some of our results were perhaps surprising – for instance a higher number of neighboring Munros seems to negatively affect ascent count – however most of them were intuitive e.g. Munros located near densely populated areas can be expected to have a higher ascent count. Finally, we have clustered the peaks into four different categories according to their features. We expanded on this by emphasising the correspondence between specific clusters and various moutaineous regions of Scotland.

Evaluation of own work: strengths and limitations Our models should be quite reliable – owing to the fact that we only use techniques that fit the given context well. We also aim to be rigorous in our working, providing mathematical justifications to further clarify certain transformations we applied to the data, as well as our plotting choices.

As a result of this, our conclusions should mostly be trustworthy; however, the nature of the data might affect their accuracy somewhat. Perhaps the most important thing to note is that our Munro climbing data is limited to the number of ascents by separate individuals – and we are using this as a proxy for Munro popularity when dealing with Question 2. Additionally, it is worth noting that WalkHighlands requires users to register before allowing them to contribute. This means that the group of people who contributed to the data we are using might not be fully representative of all hikers. The contributors can be expected to be active internet users with a remarkable passion for hiking – which is not necessarily the case for every Munro climber.

Comparison with any other related work Scarpa and Thiene also explored the patterns of human behaviour and preferences surrounding hiking in a certain area – namely the Italian Dolomites [3]. Their analysis went slightly deeper than ours, and mostly focused on the manner in which specific mountaineous sites can be substituted for others, provided that some similar features exist between them. There has been no similar study regarding Scottish Munros, and hence our work still stands as a relevant paper, albeit minimal in its scope – and it is certainly worth considering extending it.

Improvements and extensions First of all, we could improve our work by considering more factors that lead to Munro popularity; however, this might be limited by the amount of data that is freely available online. Additionally, we could take a similar approach to Scarpa and Thiene, and look for substitution patterns between different Munros – our clustering might prove extremely useful for this [3]. Finally, we were considering implementing a Munro recommendation system, through which users would receive Munro climbing suggestions based on their recorded preferences.

References

[1] Climbing scottish mountains: why 'munro-bagging' is on the up and up. *Right Vision Media*, 2019.

[2] Tom Mordue David M Brown, Sharon Wilson. Using hike-along ethnographies to explore women's leisure experiences of munro bagging. *Leisure Studies*, 2019.

[3] Riccardo Scarpa Mara Thiene. Hiking in the alps: exploring substitution patterns of hiking destinations. *Tourism Economics*, 2008.

[4] John Barnard Simon Edwardes George Gradwell Jim Bloomer Dave Marshall Graham Jackson, Chris Crocker. Walkhighlands terms and conditions. Retrieved on 7 April 2021.

[5] John Alan Dawson, Alan Holmes and Paddy Dillon Eric Yeaman Clem Clements Simon Stewart Chris Buxton Gwyn Lewis Bill Birkett Timothy Synge Mark Richards Alfred Wainwright Anne Nuttall, Michael Dewey. Dobih terms and conditions. 2021. Retrieved on 7 April 2021.

[6] Simple cities database terms of use. 2021. Retrieved on 7 April 2021.

[7] James Densmore. Ethics in web scraping. *Towards Data Science*, 2017.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] Wikipedia contributors. Elbow method (clustering), 2020. [Online; accessed 08-April-2021].