Reg No. 2205006

MA334 SP
Report on Species Distribution in the UK

Section I
Data Exploration

The given CSV file contains cleaned data which contains columns for different species of flora and fauna like ‘bees’, ‘birds’, ‘Vascular Plants’, etc. There are total 11 species given which are considered as important species for preservation in UK. Other columns include Location- which is a categorical value of name of grid in which areas are divided , Dominant Land class- in which type of land is described , Period- which is from 1970 to 2000 and 2000-2013 , Easting and Northing (coordinates for latitude and longitude), and Ecological Status which contains mean of the population of all Species.

Total rows are 5280 containing 2640 distinct Locations, 45 different types of areas and two periods. The data is collected to generate insights about species distribution among the area and decreasing population of species from P70 to P00.

We create a separate data frame containing 7 given species and other columns and consider that for further analysis. Categorical variables are converted into factors as those can be used into analysis. The population given in this data is normalized as to make it easy to observe and analyse. We add a column with ecological status for the given 7 species by taking mean for that species.

Then we create a Table containing Species Name, their Mean, Standard Deviation, Skewness, Kurtosis, Range Min and Range Max.

```
> print(Analysis_table%>%arrange(Standard_deviation,skewness))
                   Names  Mean Standard_deviation skewness Kurtosis Range_min Range_max
1         Vascular_plants 0.787              0.101   -0.126    3.103      0.42       1.2
2                    Bird 0.887              0.107   -1.507    7.052      0.24      1.17
3              Macromoths 0.849              0.141   -1.139    5.006      0.09      1.26
4 Grasshoppers_._Crickets 0.629              0.209   -0.087    2.638      0.07      1.59
5                Carabids 0.607              0.215   -0.487    2.747      0.01       1.2
6                 Isopods  0.55              0.215    0.048    2.355      0.05      1.26
7                    Bees 0.605              0.311    0.958    6.744      0.03      3.31
```

Here are some insights that can be drawn from this table:

- The mean values range from 0.55 for Isopods to 0.887 for Birds. This indicates that Birds have the highest average value, while Isopods have the lowest.
- The standard deviation values range from 0.101 for Vascular plants to 0.311 for Bees. This suggests that the data for Bees is more spread out than the data for Vascular plants.
- "Skewness is a measure of the symmetry of the distribution." (“3.1. Discuss why you did or did not create a bar chart for...”) A skewness value of 0 indicates a perfectly symmetrical distribution. The skewness values in this table range from -1.507 for

Birds to 0.958 for Bees. This indicates that the distribution of data for Birds is highly skewed to the left, while the distribution for Bees is moderately skewed to the right.

- Kurtosis is a measure of the "peakedness" of the distribution. A kurtosis value of 0 indicates a perfectly normal distribution. The kurtosis values in this table range from -1.139 for Macromoths to 7.052 for Birds. This indicates that the data for Birds has a very high peak, while the data for Macromoths is relatively flat.
- The range of values for each type of animal/plant varies. For example, the range of values for Birds is much larger (7.052) than the range for Isopods (2.355). This suggests that there is more variability in the data for Birds than for Isopods.

The Box-Plots are plotted for each species for observing the distribution of data and the information on the outliers present.

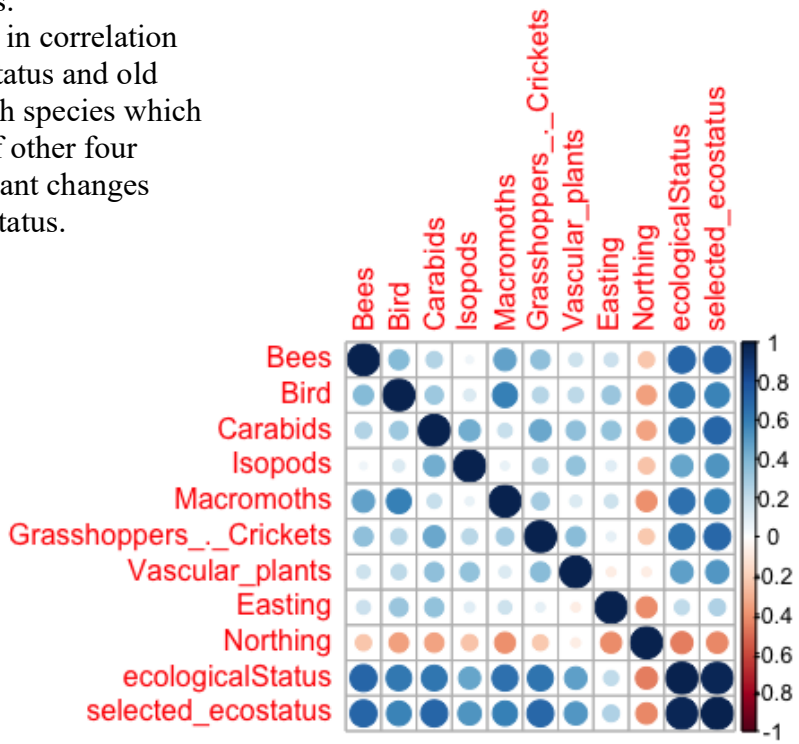
The species Richness in each area can also be calculated by grouping the data on the basis of location and counting the observations.

CORRELATION

Correlation can be calculated on the numerical columns of the dataset and a corrplot can be plotted. The correlation table that shows the correlation coefficients between different variables.

Some Insights from the table-

- Bees, bird, and Macromoths have higher correlations with ecological Status and selected_ecostatus, indicating that these species may be important indicators of ecosystem health.
- Carabids and Crickets have higher correlations with each other, as well as with Isopods, suggesting that these species may share similar habitat requirements.
- Vascular plants, Easting, and Northing have low correlations with most of the other variables, indicating that they may not strongly influence the distribution or abundance of the animal species.
- There is a small difference seen in correlation of new variable of ecological status and old variable of ecological status with species which shows that removing the data of other four species have resulted in significant changes in the value of new ecological status.



Section II
Hypothesis Tests

To perform two distinct types of hypothesis tests, we need to formulate two hypotheses and choose appropriate statistical tests.

1. Here is the First Test-

Hypothesis 1: There is a significant difference in mean ecological status between the two periods (Y70 vs Y00).
Null Hypothesis: There is no significant difference in mean ecological status between the two periods.
Alternative Hypothesis: There is a significant difference in mean ecological status between the two periods.
Statistical Test: Two-sample t-test for independent samples.

```
        Two Sample t-test

data:  period1$selected_ecostatus and period2$selected_ecostatus
t = 9.6309, df = 5278, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.02454204 0.03708694
sample estimates:
mean of x mean of y
0.7174392 0.6866247
```

The first hypothesis test is a two-sample t-test that compares the mean ecological status of the species between two different periods: P70-P00 and P00-P13. The null hypothesis is that there is no significant difference in the mean ecological status between the two periods, while the alternative hypothesis is that there is a significant difference in the mean ecological status between the two periods.
The result of the t-test shows that the t-value is 9.6309 with 5278 degrees of freedom, and the p-value is less than 2.2e-16, which is extremely small. This means that we can reject the null hypothesis and conclude that there is a significant difference in the mean ecological status between the two periods. The sample means of ecological status for period1 and period2 are 0.7174392 and 0.6866247, respectively.
The 95 percent confidence interval for the difference in means between the two periods is [0.02454204, 0.03708694]. This indicates that the mean ecological status in the later period (P00-P13) is lower than that in the earlier period (P70-P00) by an amount ranging from 0.02454204 to 0.03708694.

This result suggests that there has been a significant decline in the ecological status of the species between the two periods, indicating a potential decrease in the population of the species. It could be due to a variety of factors such as habitat loss, climate change, pollution, and other anthropogenic activities. Further investigation is required to determine the exact cause of this decline and to develop appropriate conservation measures to prevent further decline in the population of these species.

2. Here is the Second Test-

Hypothesis test for correlation between two variables:

Null hypothesis: There is no correlation between the number of bees and the number of bird species.
Alternative hypothesis: There is a correlation between the number of bees and the number of bird species.

Result-

```
> cor.test(given_data$Bees, given_data$Bird, method = "pearson")

        Pearson's product-moment correlation

data:  given_data$Bees and given_data$Bird
t = 29.475, df = 5278, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3525459 0.3988744
sample estimates:
      cor
0.3759451
```

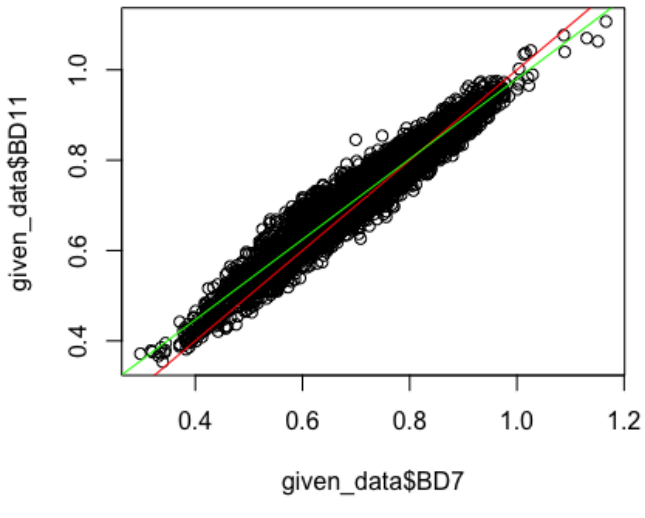
The Pearson's correlation test result shows that there is a significant positive correlation (correlation coefficient, r = 0.3759451) between the population of bees and birds in the given dataset. The p-value obtained is less than the significance level of 0.05, which indicates that the correlation is statistically significant, and we reject the null hypothesis that the true correlation is equal to zero.
P value is less than 2.2e-16.
The 95% confidence interval (0.3525459, 0.3988744) suggests that we can be 95% confident that the true correlation between bees and birds population in the population lies within this interval.

These insights suggest that the population of bees and birds might be positively associated, meaning that an increase in the population of bees may lead to an increase in the population of birds. This information can be useful for conservation efforts, as increasing the population of bees can have a positive impact on the population of birds and other species that depend on them.

Section III
Simple Linear Regression

Applying linear regression on BD7 and BD11.
Results-

- The regression equation is given by:
  BD11 = 0.090938 + 0.889447*BD7



Green Line is Regression line

- The intercept of 0.090938 is the expected value of BD11 when BD7 is equal to zero. The coefficient of 0.889447 indicates that for every unit increase in BD7, BD11 is expected to increase by 0.889447 units.
- The p-value of the F-statistic is less than 2.2e-16, which is highly significant. This indicates that the model as a whole is significant and the predictor variable BD7 has a strong relationship with the response variable BD11.
- The multiple R-squared value of 0.9315 indicates that the model explains 93.15% of the variation in BD11, while the adjusted R-squared value is the same as the multiple R-squared value because there is only one predictor variable.
- The residual standard error of 0.02828 indicates that the model has a good fit to the data, and the residuals have a mean of zero and constant variance.

In conclusion, the model suggests a strong positive relationship between BD11 and BD7, indicating that the ecological status of the selected 7 species has a strong association with the ecological status of all 11 species.

Then we apply regression on the ecological status data separated by periods ,

```
[1] "Regression results for period Y70:"

Call:
lm(formula = BD11 ~ BD7, data = data_by_period[[i]])

Residuals:
      Min        1Q    Median        3Q       Max
-0.086958 -0.017452  0.000274  0.019167  0.067201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.053950   0.003338   16.16   <2e-16 ***
BD7         0.930733   0.004601  202.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02552 on 2638 degrees of freedom
Multiple R-squared:  0.9394,    Adjusted R-squared:  0.9394
F-statistic: 4.092e+04 on 1 and 2638 DF,  p-value: < 2.2e-16
```

```
[1] "Regression results for period Y00:"

Call:
lm(formula = BD11 ~ BD7, data = data_by_period[[i]])

Residuals:
      Min        1Q    Median        3Q       Max
-0.066480 -0.020955 -0.002307  0.019191  0.125196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.109660   0.003131   35.03   <2e-16 ***
BD7         0.872912   0.004487  194.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02857 on 2638 degrees of freedom
Multiple R-squared:  0.9348,    Adjusted R-squared:  0.9348
F-statistic: 3.785e+04 on 1 and 2638 DF,  p-value: < 2.2e-16
```

The results show the regression analysis of the relationship between BD7 and BD11 for two different time periods - Y00 and Y70.

For the Y00 period, the intercept and coefficient estimates are 0.109660 and 0.872912, respectively. The coefficient is statistically significant with a very low p-value of < 2.2e-16, indicating a strong positive linear relationship between BD7 and BD11 during this period. The R-squared value is 0.9348, indicating that BD7 can explain 93.48% of the variation in BD11 during this period.

Similarly, for the Y70 period, the intercept and coefficient estimates are 0.053950 and 0.930733, respectively. The coefficient is also statistically significant with a very low p-value of < 2.2e-16, indicating a strong positive linear relationship between BD7 and BD11 during this period. The R-squared value is 0.9394, indicating that BD7 can explain 93.94% of the variation in BD11 during this period.

Overall, the results suggest that there is a strong positive linear relationship between BD7 and BD11, and this relationship holds for both periods analyzed.

Section IV

Multiple Linear Regression

BD4 is calculated and stored in the 'given_data' data frame which contains seven species and BD7 and BD11 columns.

Multiple linear model is fitted which give the following Result-

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.093259   0.010521   8.864  < 2e-16 ***
Bees                     0.100127   0.003705  27.025  < 2e-16 ***
Bird                     0.213058   0.011803  18.051  < 2e-16 ***
Carabids                 0.036652   0.005650   6.487 9.55e-11 ***
Isopods                  0.062623   0.005123  12.225  < 2e-16 ***
Macromoths               0.257849   0.009195  28.041  < 2e-16 ***
Grasshoppers_._Crickets 0.048916   0.005624   8.698  < 2e-16 ***
Vascular_plants          0.113598   0.010928  10.395  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07037 on 5272 degrees of freedom
Multiple R-squared:  0.6108,    Adjusted R-squared:  0.6103
F-statistic:  1182 on 7 and 5272 DF,  p-value: < 2.2e-16
```

Then feature selection is done by keeping p-value threshold as 6.592555e-34. The remaining species after these are "Bees", "Bird", "Isopods" and "Macromoths".
Regression model is fitted on these species with the following Insights-

```
Call:
lm(formula = BD4 ~ ., data = given_data[, c(significant_vars,
    "BD4")])

Coefficients:
(Intercept)         Bees         Bird      Isopods   Macromoths
     0.1671       0.1174       0.2490       0.1036       0.2622
```

1. All of the selected predictor variables are statistically significant, with p-values much smaller than 0.05.

2. The intercept is also statistically significant, with a p-value smaller than 0.001.

3. The multiple R-squared value of the model is 0.582, indicating that the model explains 58.2% of the variability in `BD4`.

4. The adjusted R-squared value is 0.5817, which is very close to the multiple R-squared value, suggesting that the addition of predictors beyond the selected ones does not significantly improve the model fit.

7. The F-statistic is highly significant with a p-value much smaller than 0.05, indicating that the model as a whole is significant.

8. The estimated coefficients for the selected predictors show that all of them have a positive association with `BD4`.

9. Specifically, an increase in one standard deviation of `Bees` is associated with an increase in `BD4` by 0.1174 units, an increase in one standard deviation of `Bird` is associated with an increase in `BD4` by 0.2490 units, an increase in one standard deviation of `Isopods` is associated with an increase in `BD4` by 0.1036 units, and an increase in one standard deviation of `Macromoths` is associated with an increase in `BD4` by 0.2622 units.

The results suggest that the selected predictors (`Bees`, `Bird`, `Isopods`, and `Macromoths`) are important predictors of `BD4`, and their positive associations with `BD4` are consistent with the hypothesis that higher biodiversity is associated with higher abundance or diversity of certain taxa.

Model Selection Using AIC is performed and following result is obtained-

```
Start:  AIC=-27965.55
BD4 ~ Bees + Bird + Carabids + Isopods + Macromoths + Grasshoppers_._Crickets +
    Vascular_plants

                         Df Sum of Sq    RSS    AIC
<none>                                 26.107 -27966
- Carabids                1    0.2084 26.315 -27932
- Grasshoppers_._Crickets 1    0.3747 26.481 -27899
- Vascular_plants         1    0.5351 26.642 -27867
- Isopods                 1    0.7400 26.847 -27826
- Bird                    1    1.6135 27.720 -27658
- Bees                    1    3.6167 29.723 -27289
- Macromoths              1    3.8938 30.000 -27240
```

The result of the `stepAIC` function suggests that the best model for predicting `BD4` includes all seven predictor variables: `Bees`, `Bird`, `Carabids`, `Isopods`, `Macromoths`, `Grasshoppers_._Crickets`, and `Vascular_plants`. This is indicated by the fact that the output of `step.model` shows the same formula as the original model (`model`), which included all seven predictors.

The coefficients of the final model are also shown in the output. They represent the estimated effect of each predictor variable on the response variable (`BD4`) while holding all other predictors constant. For example, the coefficient for `Bird` is 0.21306, which suggests that a one-unit increase in `Bird` is associated with a 0.21306 unit increase in `BD4`, on average, while holding all other predictors constant. Similarly, the coefficient for `Macromoths` is 0.25785, which suggests that a one-unit increase in `Macromoths` is associated with a 0.25785 unit increase in `BD4`, on average, while holding all other predictors constant.

Section V
Open Analysis

We start the analysis of the given data by first pre-processing it. Converting categorical values into factors. First, we have plotted a table showing distribution of species data and calculated skewness , Min range, Max range etc.

Histogram of BD7 is plotted to see the distribution of data which shows that it follows a normal distribution graph with standard deviation of 0.12. Then a percentage interval variable "Qs" is created that is used to separate the areas with most number of species and least number of species.

Box plot of the ecological data for top 10 percent suggest that most of the areas have 10 observations and areas with more than 50 observation can be considered outliers.

Then data is grouped by land type and then new mean for ecological statuses is calculated. Most readings is seen in the area 3e with 173 number of observations.

Then we create a new pivoted table, in which there are separate columns for Y70 and Y00.

```
# A tibble: 45 × 3
# Groups:   dominantLandClass [45]
   dominantLandClass   Y00   Y70
   <fct>             <dbl> <dbl>
 1 10e               0.724 0.746
 2 11e               0.822 0.789
 3 12e               0.809 0.814
 4 13e               0.724 0.771
 5 13s               0.661 0.771
 6 15e               0.793 0.836
 7 15w               0.634 0.786
 8 16e               0.684 0.719
 9 17e               0.760 0.753
10 17w1              0.797 0.772
# … with 35 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Then an new column is created which subtracts ecological status from one period to another , giving values in positive – if the biodiversity is increased and negative if biodiversity is decreased. Box plot is also plotted to check the distribution of data which shows that most of the data is negative and mean is also in negative (-0.038).

Classify on the basis of Country

We filter the data using "grepl" function which identifies the alphabet and added the column country to the data.
After this , a Subset is created to filter and select the data in terms of country , three different data frames are made –

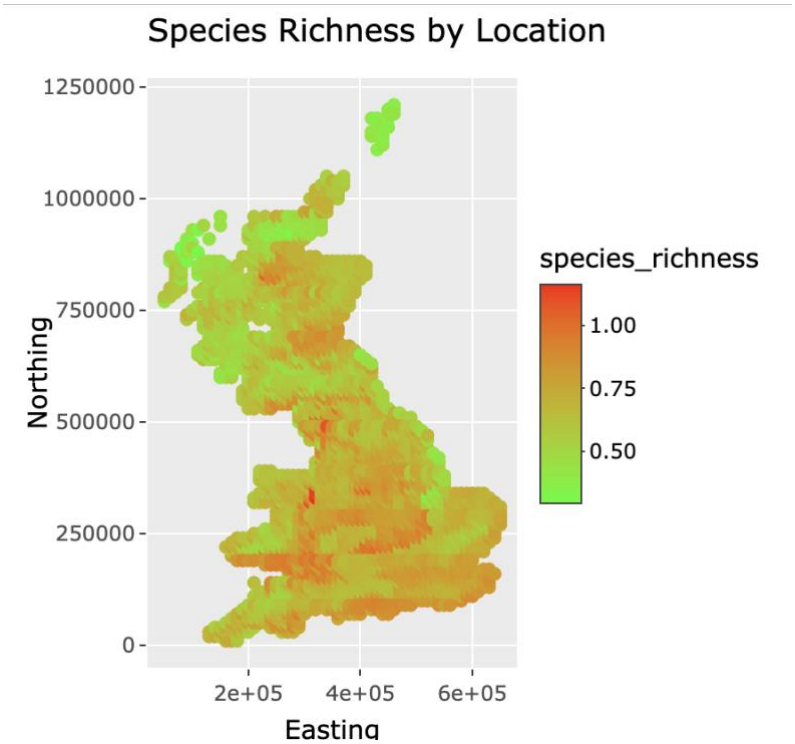
England_data, Scotland_data, and Wales_data. Then these datasets are filtered and pivoted to separate the Periods and add a column to calculate the change in ecological status during that period.
Boxplot is plotted with the variable 'eco_change' for all three countries.

Mean loss of biodiversity for England is -0.022 , for Scotland -0.061 and for Wales is -0.033.

These are the indicators of loss of biodiversity population during the period of 70's to 2010's. The worst performing country in terms of overall mean is Scotland which has a very high loss and this must dealt with actions to reduce this in future .

We can also calculate the percentage change in the areas where there is loss and gain of biodiversity and data shows that 76.19 % of England's Areas have suffered loss in biodiversity , 93.75% for Scotland and 62.50% for Wales. The best performing country in terms of percentage of areas suffered losses is Wales and worst performing is Scotland.

Plotting few graphs to analyse the species richness visually.



A Multiple regression model is fitted in between Ecological Status and Northing + Eastings. Following insights can be drawn from it.

The linear regression model shows that both easting and northing have a significant effect on species richness. The intercept of the model is 0.7477, which means that at a hypothetical location where both easting and northing are zero, the expected species richness would be 0.7477. The coefficient of easting is positive and very small (9.007e-08), indicating that for every unit increase in easting, the expected species richness increases by 9.007e-08 units, holding northing constant. Similarly, the coefficient of northing is negative and small in magnitude (-1.714e-07), indicating that for every unit increase in northing, the expected species richness decreases by 1.714e-07 units, holding easting constant.

The p-values for both easting and northing are less than 0.05, which means that the coefficients are statistically significant and the relationship between each predictor and species richness is not due to chance. The adjusted R-squared value of 0.1963 indicates that the model explains 19.63% of the variance in species richness, which is a moderate amount.

Based on this model, we can conclude that there is a significant relationship between location (easting and northing) and species richness, suggesting that species richness varies spatially. This information can be useful for conservation efforts, as areas with high species richness can be targeted for protection and management.

Conclusion

The report on Species Distribution in the UK has found that there has been a significant decline in the ecological status of the species between the two periods, indicating a potential decrease in the population of the species. It could be due to a variety of factors such as habitat loss, climate change, pollution, and other anthropogenic activities. Further investigation is required to determine the exact cause of this decline and to develop appropriate conservation measures to prevent further decline in the population of these species.

The report has also found that there is a significant relationship between location (easting and northing) and species richness, suggesting that species richness varies spatially. This information can be useful for conservation efforts, as areas with high species richness can be targeted for protection and management.

Overall, the report provides valuable insights into the factors that affect the distribution of species and the challenges that they face. It is a valuable resource for anyone who is interested in learning more about this important topic.

Reference:

- Developing a biodiversity-based indicator for large-scale environmental assessment: a case study of proposed shale gas extraction sites in Britain (Doi: 10.1111/1365-2664.12784)
- Using The National Grid (www.os.uk)
- ITE land classification (Bunce et al. 2007)
- *Peakedness - an overview | ScienceDirect Topics,* https://www.sciencedirect.com/topics/mathematics/peakedness.