Statistical analysis in RStudio

Contents

0.0.1 Abstact

The objective of the report is to perform Statistical analysis using Seal Data from 2007 to 2010.The Seals data is collected every year from mid of June till mid of September.After the disappearance of seals in the early 1930s, for the first time the seals started to colonized near the Tees Estuary.The Seals in the specific year are monitored and the data is collected accordingly.

The objective of this report is to analyze seal population distribution and variance in Tees Estuary and perform analysis. In order to perform the analysis, check for the normality check and perform the correlation and statistical association using Kruskal-wallis test and Pairwise Wilcox test. The performance of different test will help to evaluate the performance in other words it will help to make understand that the occurring data is significant or not. With the basic analysis of the data, however the data needs to be more specifically explored, to understand the analysis better, the comparison of average count using each Year was performed to check the presence of seal absence and presence in specific years. Using Kruskal wallis and Pairwise Wilcox test, In 2009, the spices populations were not significantly common than rest of the years.

After performing all the test using the data, the exploring of data was significantly needed for better understanding of the performance of both the species in different year. The data was explored using ggplot library and performed visual analysis which shows the error bars are reflecting standard error. Finally the last task was to explore the data and perform the visual analysis using ggplot library.

1 INTRODUCTION

For many hundreds of years, Seals in the Northern England have lived near mouth of River Tees but declined rapidly by late 1800s Years. The main reason reason for rapid declining was number of factors such as pollution, environmental changes.As this resulted complete disappearance of seals in the 1930s. In the mid 1950s when the plats nearby were replaced by the newer ones, the pollution in the river started

controlling, due toin 1970s the Harbor seals started to reappearing .The Only estuary where the Harbor seals were re-colonized was in Teesmouth which proved a direct result for improving the environment.Therefore, due to this in late 19th century a central committee was formed in 1992 in order to monitor seals in the Tees Estuary and to compare them to previous years.

Considering the duration of seals, the seals likely to seen in the mid of June till mid of September Each Year.However, Each year , for the same duration a monitoring is carried out when seals give birth until they gater to molt. The following data were recorded each year

- The Number of Seal Species.
- The Number of Seal Pups.
- The Location of Seals
- Instance of disturbance to seals
- Deaths or injuries to Seals
- Abandonment of Seals
- Weather Conditions

2 MATERIALS AND METHODS

2.1 Data Collection

The main objective of statistical analysis is to find the correlation and Trends of seal abundance using the serial dataset from year 2007 to 2010.

Firstly we will download the data and convert it into a csv file and then import data into R studio after importing data We will read the data and convert the data into data frame Which will import data into rows and columns. If we check the head of data we can find There are 5 columns 210 observations.

- The first column represents the year, each row consists of the year from 2007 to 2010.
- The second column represents the month of a specific year; there are 15 unique variables if it presents the month of a particular Year.
- The third column represents a site which consists of 7 unique variables A,B,C,D,Split and Wall.
- The fourth column represents species which have true unique species Harbour and grey.
- The final column represents the average count of each species in a particular year year and month each variable has a unique count value.

Summer	Year.Month	Site	Species	average.count
summer-2007	2007.Jun	A	HARBOUR	14.1
summer-2007	2007.Jun	B	HARBOUR	0.3
summer-2007	2007.Jun	C	HARBOUR	14.7
summer-2007	2007.Jun	Spit	HARBOUR	0.1
summer-2007	2007.Jun	Wall	HARBOUR	0.7
summer-2007	2007.Jun	D	HARBOUR	0.0

```
## 'data.frame':    210 obs. of  5 variables:
##  $ Summer       : chr  "summer-2007" "summer-2007" "summer-2007" "summer-2007" ...
##  $ Year.Month   : chr  "2007.Jun" "2007.Jun" "2007.Jun" "2007.Jun" ...
##  $ Site         : chr  "A" "B" "C" "Spit" ...
##  $ Species      : chr  "HARBOUR" "HARBOUR" "HARBOUR" "HARBOUR" ...
##  $ average.count: num  14.1 0.3 14.7 0.1 0.7 0 0 0.3 0 2.6 ...
```

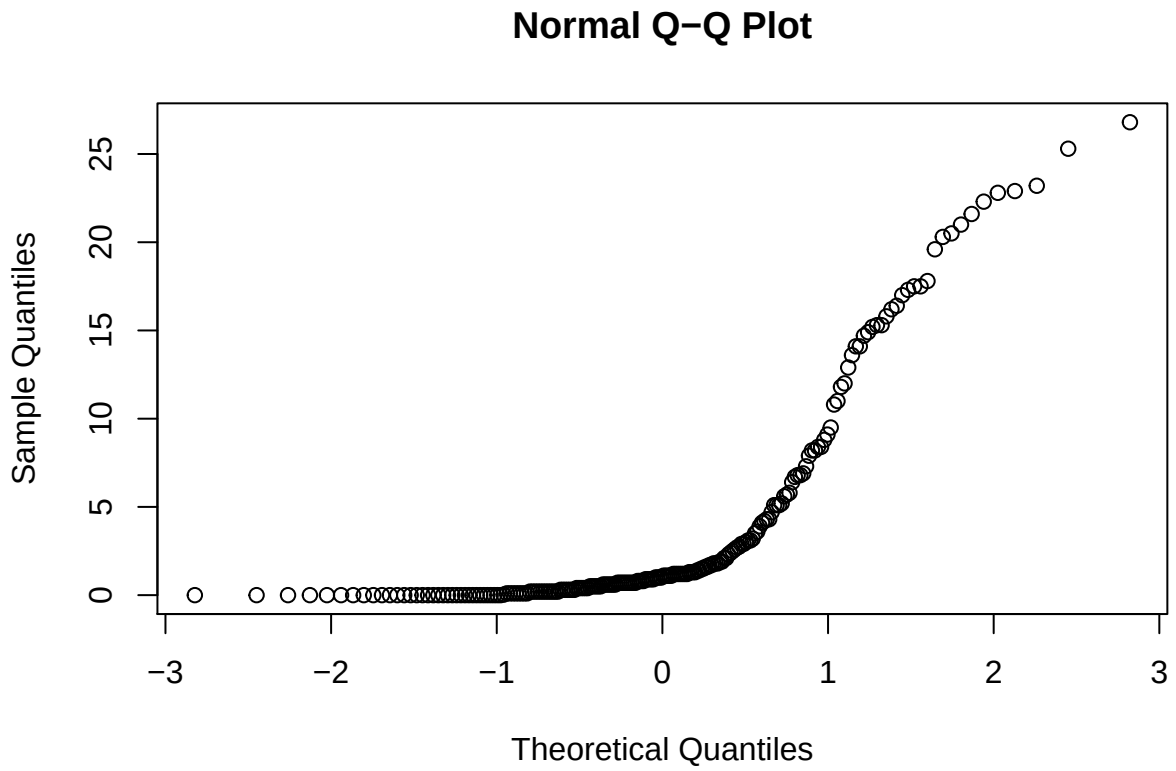
On the next section we will use various methods for the statistical analysis and explore the data which will help to understand the relationship between dependent and independent variables. Before proceeding further if we see the data set we can see that only column has numerical values which we will be considering to use for our analysis, Moving into deep observation, the Average count consist of each numerical value which represents the average count of seals presents in particular year an month. In addition considering the observation it can be identified for two species which is Grey and Harbor. We can say that the average count is quiet dependent for each of rows and though each of the row is quiet responsible of the analysis therefore in order to explore in the detailed way we can run some normality test.

2.2 Normality Test for Seal population in Tees Estuary from 2007 to 2010.

After checking rows and columns of data, it can be easily identified that only the average.count data column has numerical value. If we check the normality of the average.count column using the qqnorm() function,The Below visualization data represents the relationship between the theoretical and sample quantities which derive the plotted data is distinctly curved and determines that the data is not normal.

2.2.1 QQNORM TEST

qqnorm is generic function in r that produces which helps to generate Quantile plot of the y values. The x plot consists of theoretical values where as the y plot consists of sample quantities. A Quantile-Quantify (Q-Q) plot is a scatter plot comparing the fitted and empirical distributions in terms of the dimensional values of the variable (i.e., empirical quantifies)(Vito Ricco,p.4).



While finding out that data is not normal in normality test using qqnorm function, if we run the same test using shapiro.test() function, The p-value is Greater than 0.01 we can say that it then all hypothesis is not rejected Data is not significantly distributed therefore we have to perform non parametric test.

2.2.2 shapiro Test

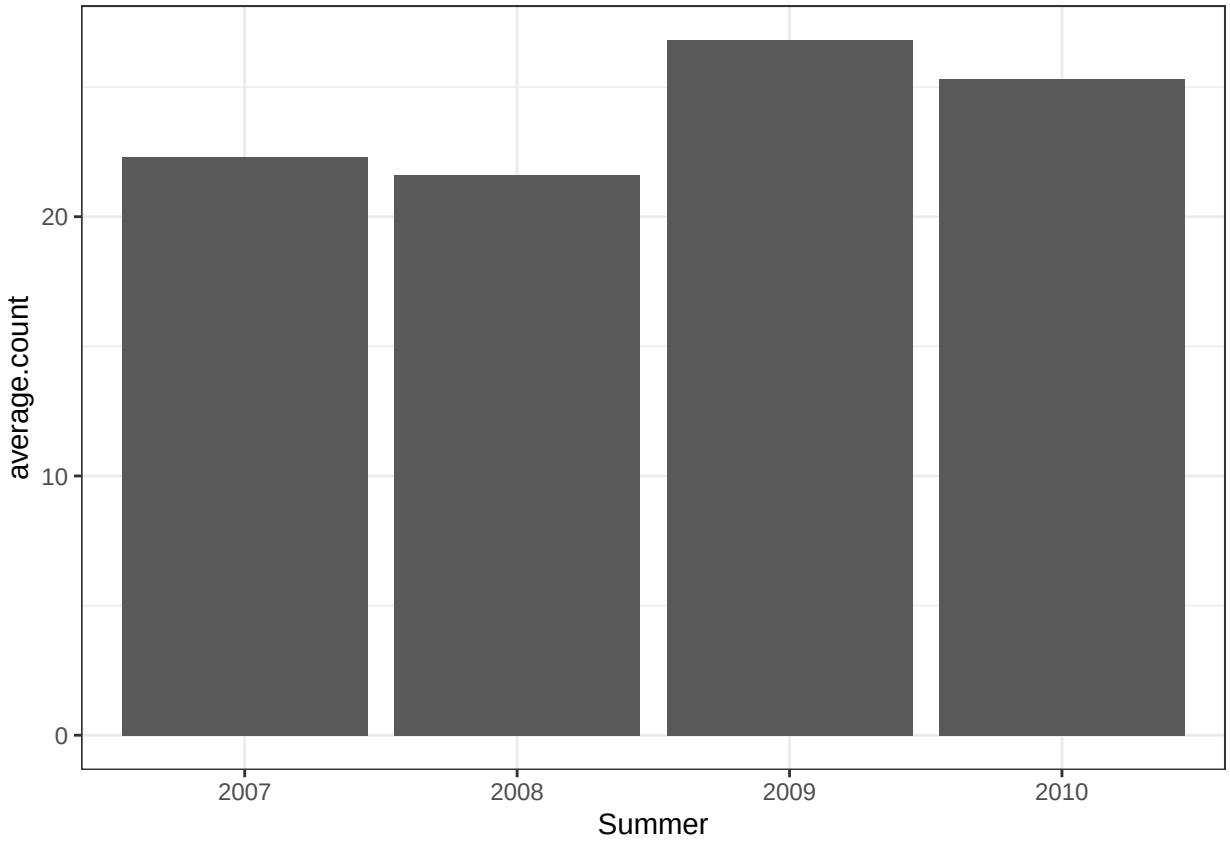
Shaprio Normailty test is used to find whether the data is fit for the normal distribution This test was the first test that was able to detect departures from normality due to either skewness or kurtosis, or both (Althouse et al., 1998)(Nornadiah Mohd Razali and Bee Wah Yap ,p.25).

When we run the shaprio test for average count we can see that the p value is 2.2 and the value of w is 0.67. Thus this test will reject the normality.Since the data is not significant,in order to find out The number of seals are significantly different for each year plot the data of average count of number of seals for each year but before that we have to change the data type of the first four columns into factors.

If we change the data type we get to know that

- The first column consist of four factor i.e 2007 till 2010
- The Second column consists of 15 factor each factor is month of each year.
- The Third column consists of seven factor, each factor has considered as the different sites of the species.
- The Fourth Factor consists of two factor each factor is unique species.
- The Final column is to be left over, as it has numerical data type and represents each unique counts for particular year and month.

As we have seen that, after changing the data type we can find that the summer column has four factor levels, Year. month column has 15 factor levels, the site column as 7 factor level and and the species column two factor levels. Therefore we can check now how much average count of species is there in each year which will be helpful to analyse the normality of data.



In the above graph if we can see the the average count of species is highest in 2009 and followed by 2010

where as the least count can be seen in the year 2008 in Tees Estuary but visualizing the plot we can cannot consider the changes are significant or not.

In order to find out this we can transfer some non parametric test like kruskal Wallis rank sum test Which help us to determine the overall correlation and statistical Association for data by providing a chi-squared interpretation and a p-value.

2.2.3 Kruskal-Wallis Test

Kruskal Wallis Test is Considered as Non-Primitive test and an alternative of Test.It is also called as One way ANNOVA test as the is performed by different individual groups.They are considered as non- primitive test because

the non-parametric tests are less powerful than their parametric counterparts,i.e. a parametric test is more likely to detect a genuine effect in the data , if there is one, than a non-parametric test (TEODORA H. MEHOTCHEVA,p.3).

In order to perform the Krukals-wallis test, we need to use the kruskal.test() function.

```
## [1] 0.1006759
```

Firstly, if we explore the overall difference between all the variables we can see that The p-value is 0.10 and chi-squared value is 6.23 which is greater than the normality distribution, thus we can say that it There is no significant difference between the groups and considering the the data we can also see that the data is not homogeneous.Yet, we know there is difference but we don't know where it is.Therefore, In order to explore this into more detail to find out that the pairs of groups are different or not, will use pairwise willox test For comparing different group levels for multiple testing.

2.2.4 Pairwise wilocx Test

Pairwise wilocx test is also a non-primitive test which uses the multiple distributions or repetitive distributions where two or more than two observation is considered to the test. This test helps to test the significant distibtions of a group to derive the paticluar group lies in normality or not.

Considering the the pairwise solution below for the average count and the year group in order to find the data is significantly distributed or not.

	2007	2008	2009
2008	0.8933132	NA	NA
2009	0.2023815	0.4280243	NA
2010	0.4280243	0.6385207	0.8933132

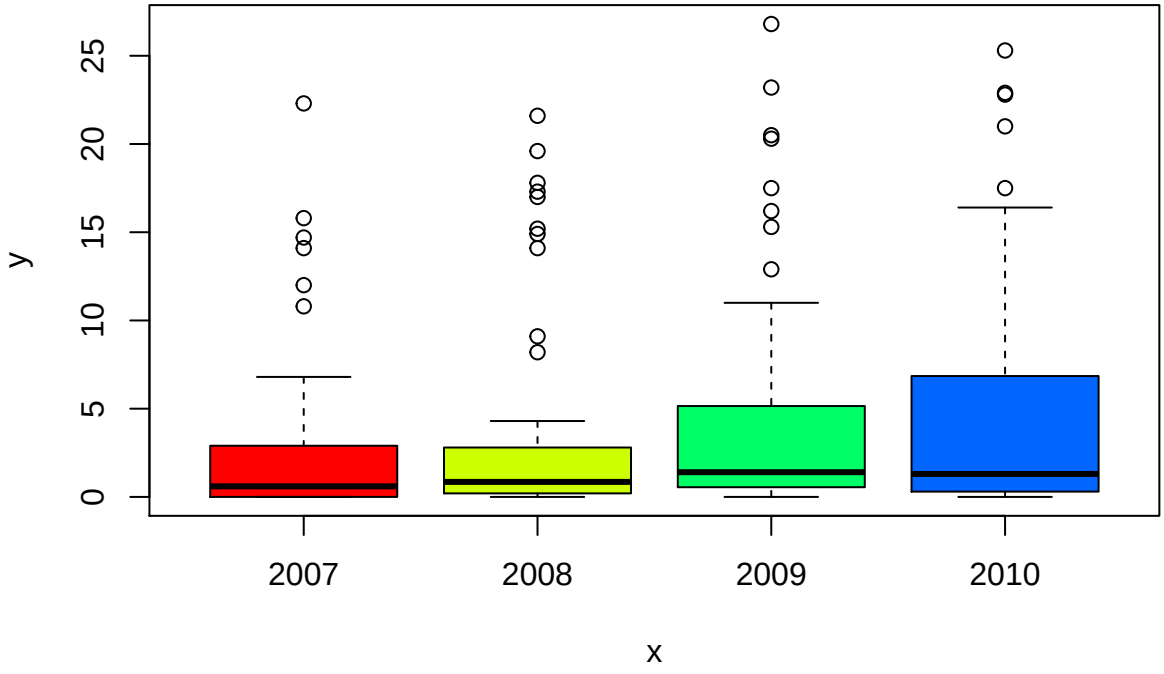
In the above test we can see that the P value for significant years is quite normal. further, providing an adjustment method we can avoid the false positive results in order to see more accuracy.

	2007	2008	2009
2008	0.5359879	NA	NA
2009	0.1730807	0.1730807	NA
2010	0.1730807	0.3192604	0.7529974

In the above test we can see that the adjusted P value is greater than 0.05, this shows that your value is

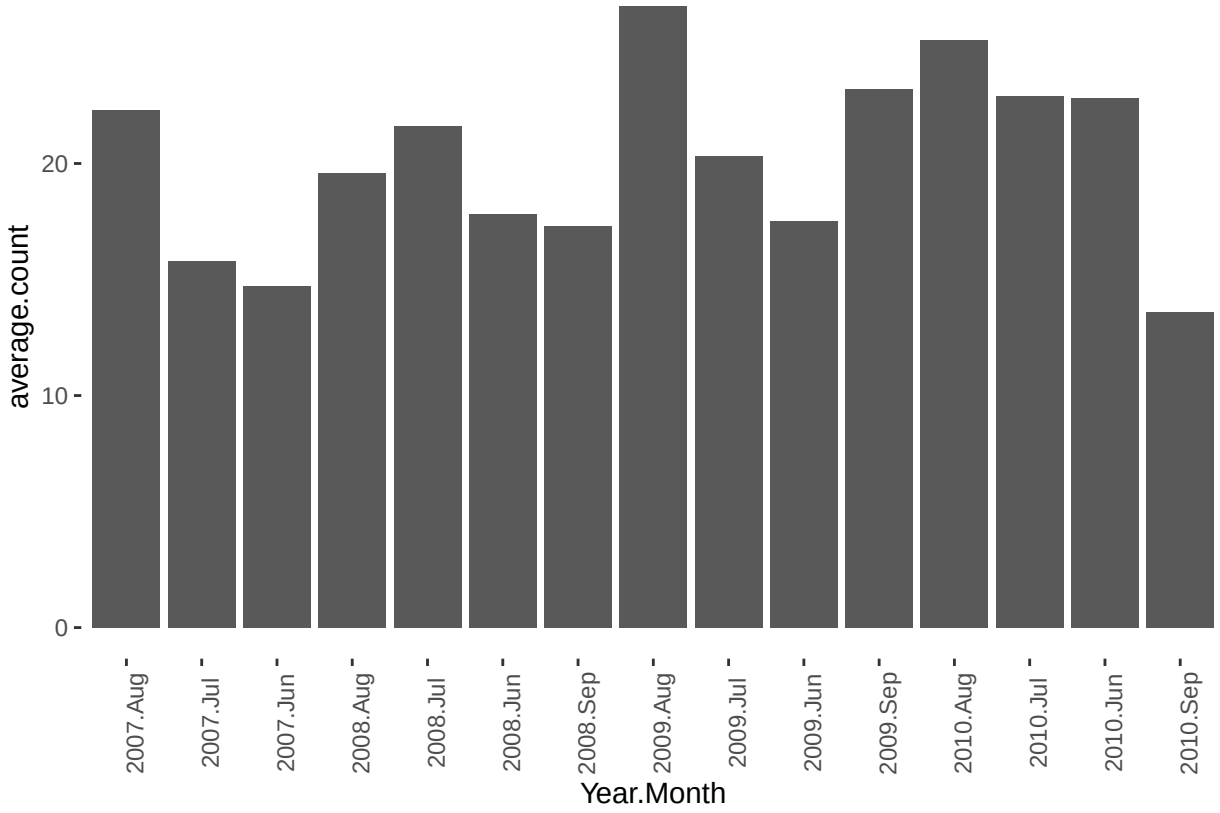
not significant. In order to understand it with more clarity if we plot our data we can see that there was a significant difference between 2010 and 2007, we can see that the highest number of seals were counted in 2010. Also we can see that in 2007.

There were some high counts of seal, but in the previous bar graph it was suggested that there was a big significant difference between 2007 and 2010. From this we can say that the statistical tests were more accurate and help us to find the significant errors which were present in data.



After this we can check the presence of the seals in Tees Estuary for each month in a particular year. We might be able to explore the difference for the presence of seals in each month of the year.

The plot below shows the values of average count for each month of particular year.We can see that in 2009 Aug,the number of species were counted the most and in 2010 Sep the species were counted the least. But in order to find significant relationship we need to explore the data and run test accordingly.



2.3 Normality Test For Seal population in Tees Estuary for Each Month

If we consider the year 2007, we can see that there is no significant difference between different months.The p-value is 0.51 and the adjusted p-value is than 0.62. Considering the normality, we have to explore for the each year in order to find out if the data is significant or not not in every year.

```
## [1] 0.5191072
```

	2007.Aug	2007.Jul
2007.Jul	0.6273101	NA
2007.Jun	0.6273101	0.6273101

If we consider the year 2008, we can see that there is no significant difference between different months.The p-vaule for kruskal test is 0.8 and the adjusted p-values is 0.93.Further, we need to check for 2009 and 2010.

```
## [1] 0.8349213
```

	2008.Aug	2008.Jul	2008.Jun
2008.Jul	0.9263274	NA	NA
2008.Jun	0.9263274	0.9263274	NA
2008.Sep	0.9263274	0.9263274	0.9263274

If we consider the year 2009, we can see that there is no significant difference between different months. The p-value is 0.93 and the adjusted p-value is 1. We can see that until now we didnot find any significant differnce, lets check the the test of 2010.

```
## [1] 0.9359747
```

	2009.Aug	2009.Jul	2009.Jun
2009.Jul	1	NA	NA
2009.Jun	1	1	NA
2009.Sep	1	1	1

If we consider the year 2010, we can see that there is no significant difference between different months. the p-value of the krukal test is 0.4 and the adjusted p-vaues varies from year to year, thus the normality test for each month can be considered as the significant difference therefore we can consider the normailty of each year.

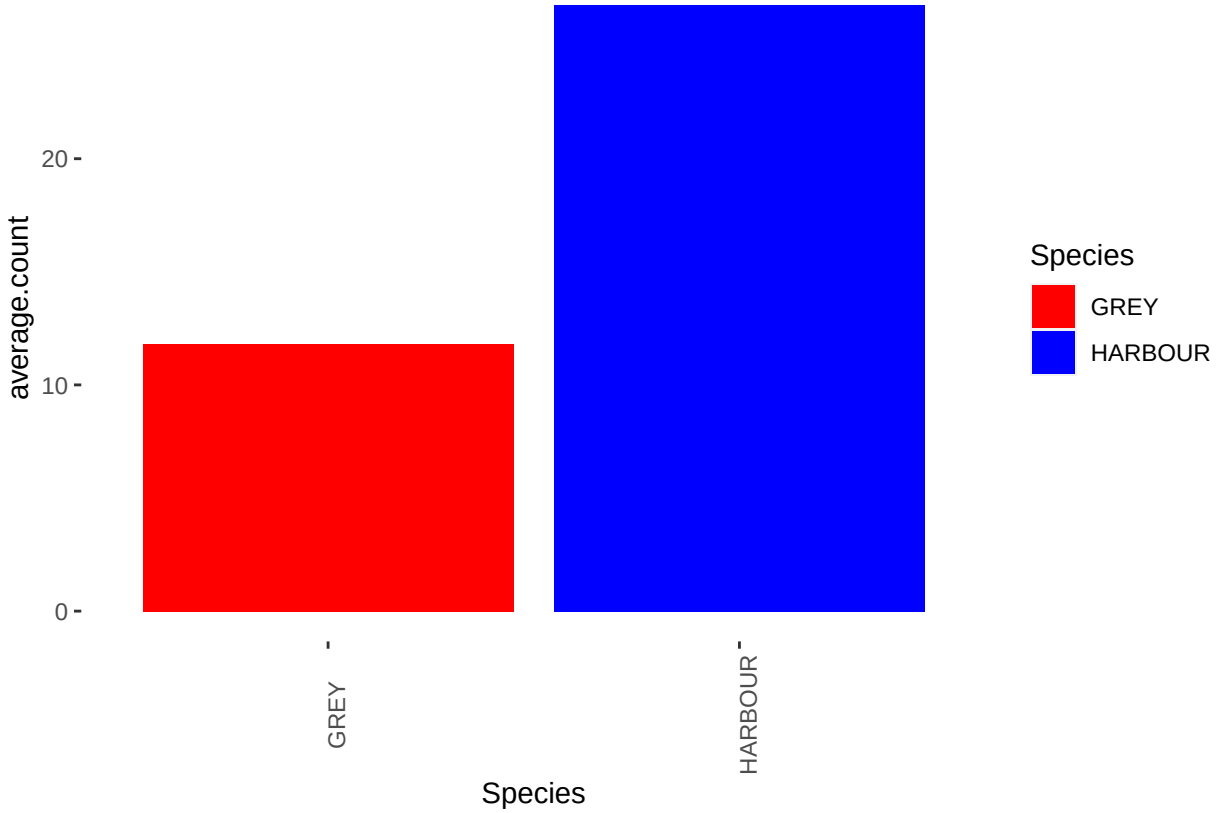
```
## [1] 0.4752711
```

	2010.Aug	2010.Jul	2010.Jun
2010.Jul	0.7546306	NA	NA
2010.Jun	0.5976528	0.9629236	NA
2010.Sep	0.9629236	0.5976528	0.5976528

We have seen above the normality test also did not perform very well for each month but we have seen that there should be a significant difference in the data. So we will try to explore the Species group in order to find the normality of derivatives.

2.4 Normality test for Seal population in Tees Estuary of each species.

Considering the data and previous stats we found that there were no significant difference for each month but in order to get detailed information about the species we can check the count of each type of species and find out the non-significant relationship of the data .



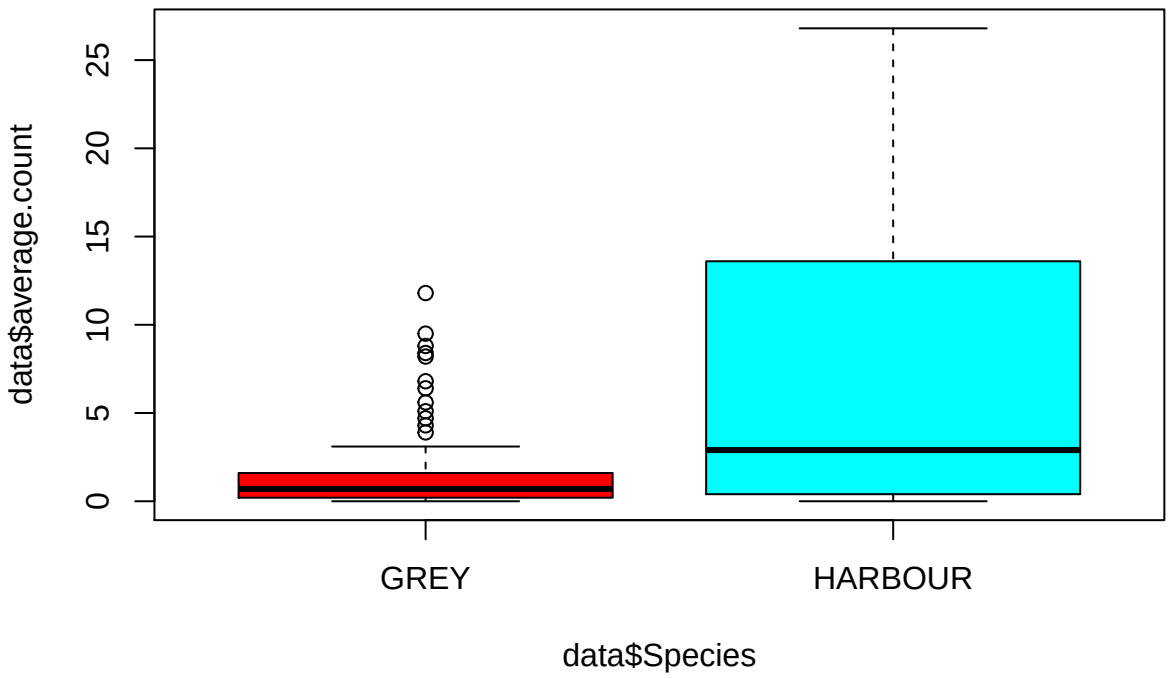
As we had seen the above graph about the average count of the harbor species are very much than the grey species.We can say that the Harbor species are more common than the grey species, using the kruskal test we can find out the difference between each species in Tees Estuary.

```
## [1] 1.561916e-05
```

	GREY
HARBOUR	1.57e-05

After performing the test we can easily see the p values is greater than 0.01 and after plotting the below graph we can say that the Harbour seals population are more comparatively grey. In other way, it can be derived that the harbor seals in Tees Estuary are more significantly common than the grey seals

Also, we can explore data in more detailed way in order to find out significant difference of species for each year.



If we look at the below stats, we can clearly see the p-value for 2007 year is 0.09 and adjusted p-value is 0.07 also greater than the significant value which derives there is slightest significance difference in the Species. However this test will fail considering the significant difference of normality test.

```
## [1] 0.06938051
```

	GREY
HARBOUR	0.0713513

Again, If we look at the below stats, we can clearly see the p-value for 2008 year is 0.09 and adjusted p-value is 0.10 which is greater than significant value which derives the group is not in range of normailty and we have to consider it as a failed test.

```
## [1] 0.09866302
```

	GREY
HARBOUR	0.100351

If we look at the below stats, we can clearly see the p-value for 2009 year is to 0.001 and adjusted p-value is also 0.001 which derives there is significance difference in the Species.Therefore we can say that the data is normally distributed in the year 2009.

```
## [1] 0.001307411
```

	GREY
HARBOUR	0.0013452

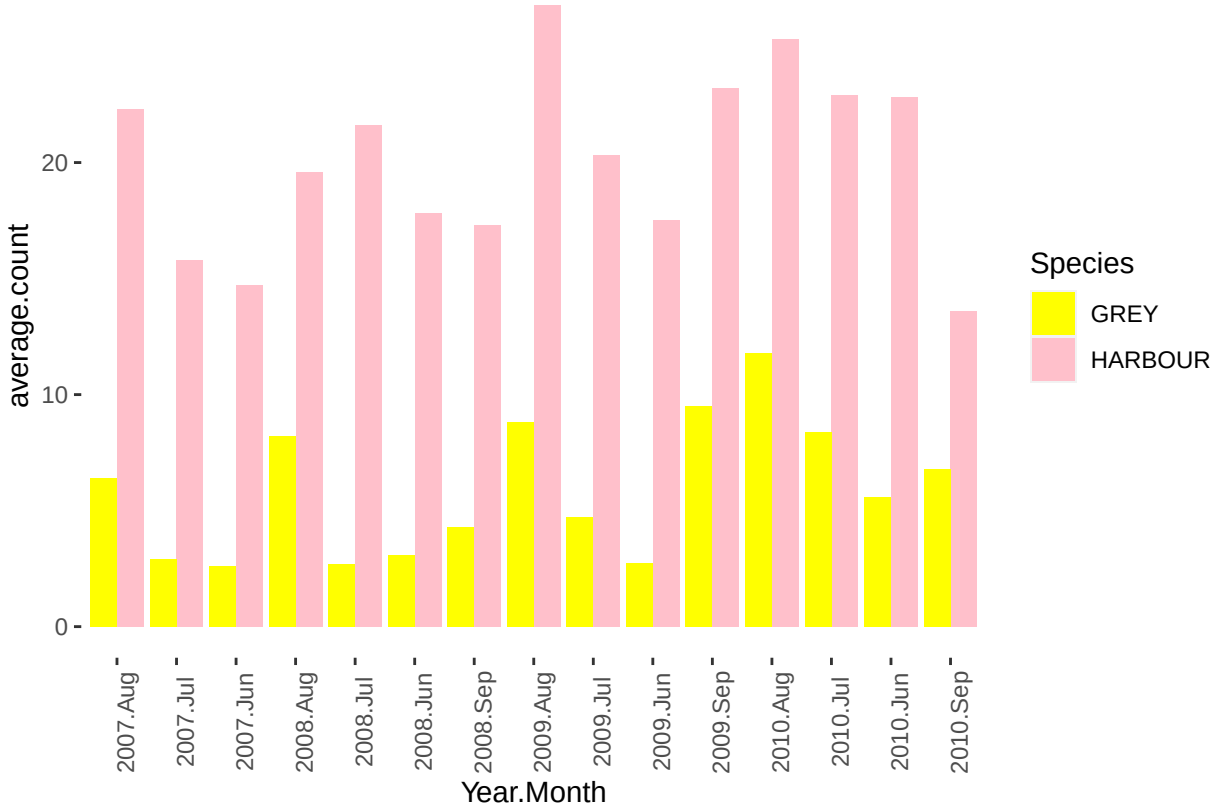
Finally, If we look at the below stats, we can clearly see the p-value for 2010 year is 0.04 and adjusted p-value is is 0.04 which derives there is significance difference in the Species. Therfore we can say that the data is normaly distibuted in the year 2010

```
## [1] 0.04234657
```

	GREY
HARBOUR	0.0431888

From this we can easily say that in 2007 and 2008 the population of species were not significantly different where as in 2009 and 2010 we can say that the population is significantly different from other.By this we can say that in the year 2009 and 2010 the data is significantly distributed and is normalized.

Therefore from here we need to calculate the error in order to find the out better results and improve the performance but before that we can check the population of individual species for particular month and further we can develop the error bars after calculating the error .

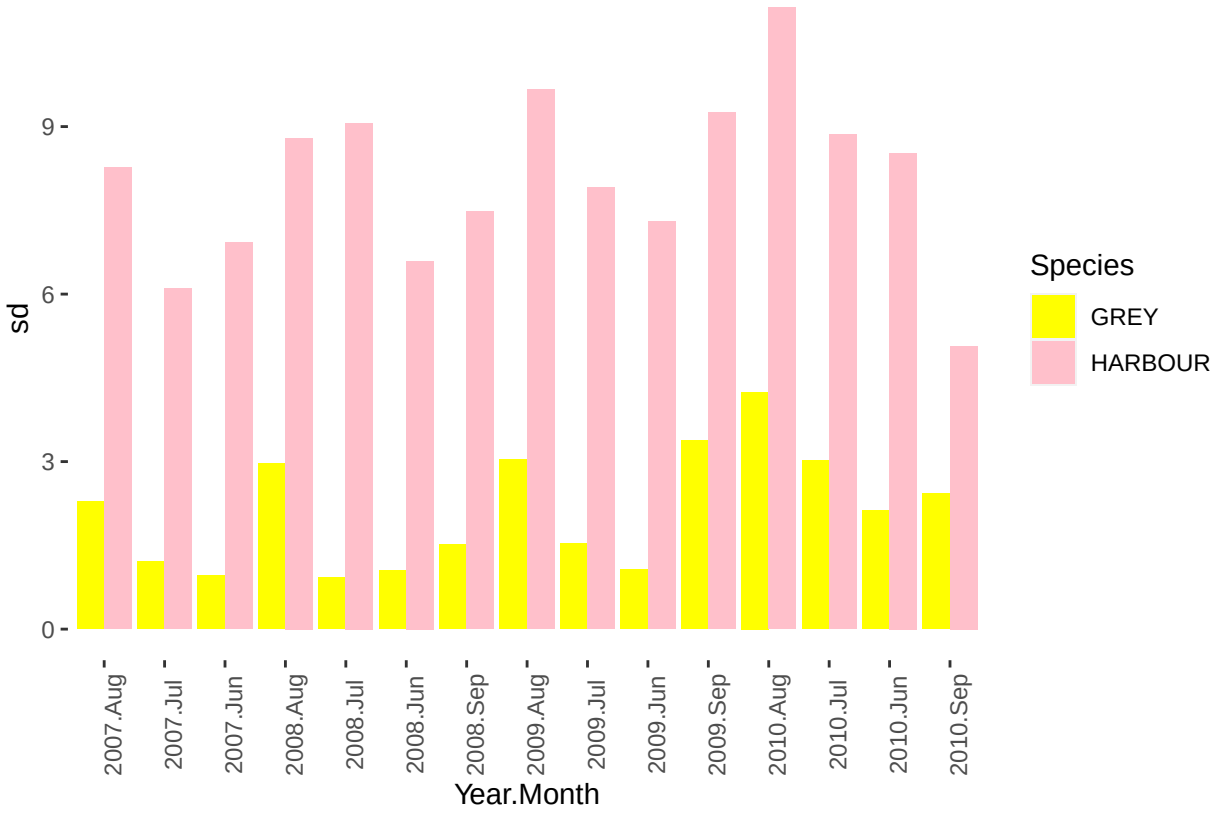


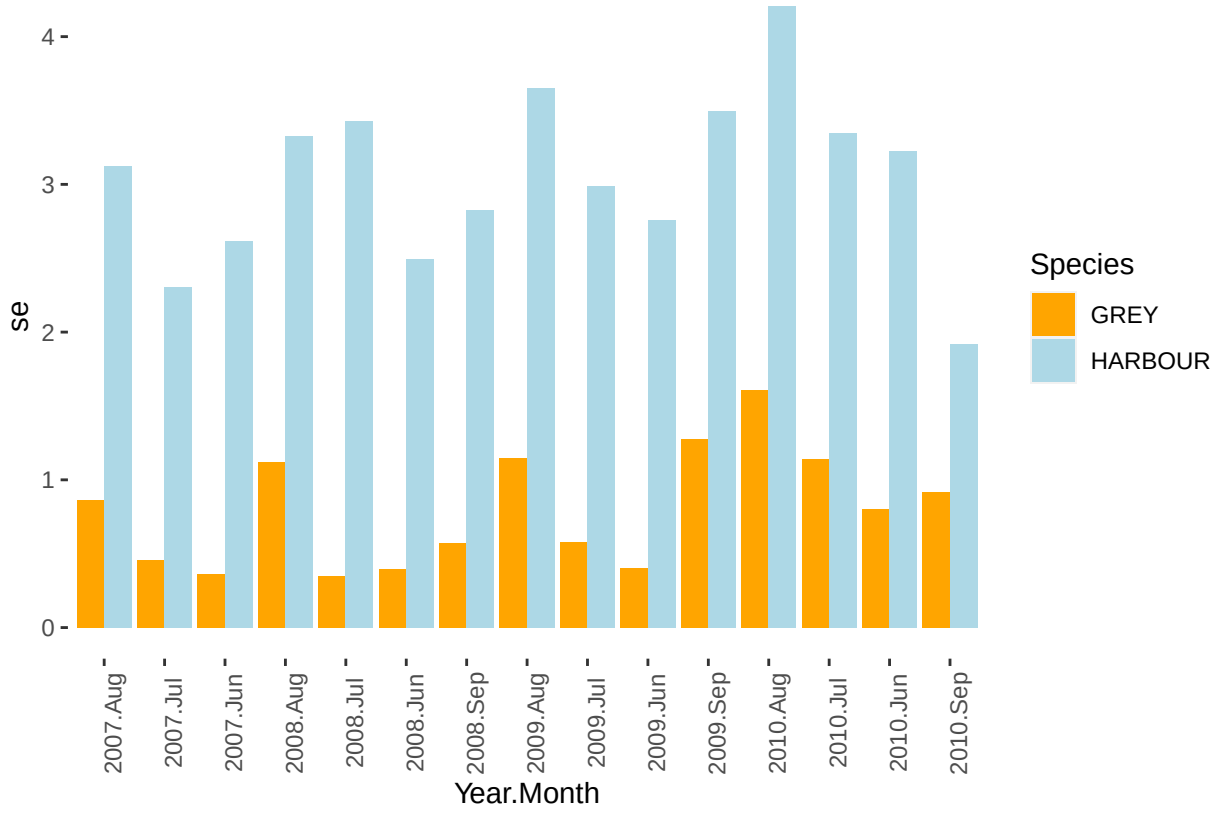
If you look at the graph of average count of each species in a particular month everywhere we can see that

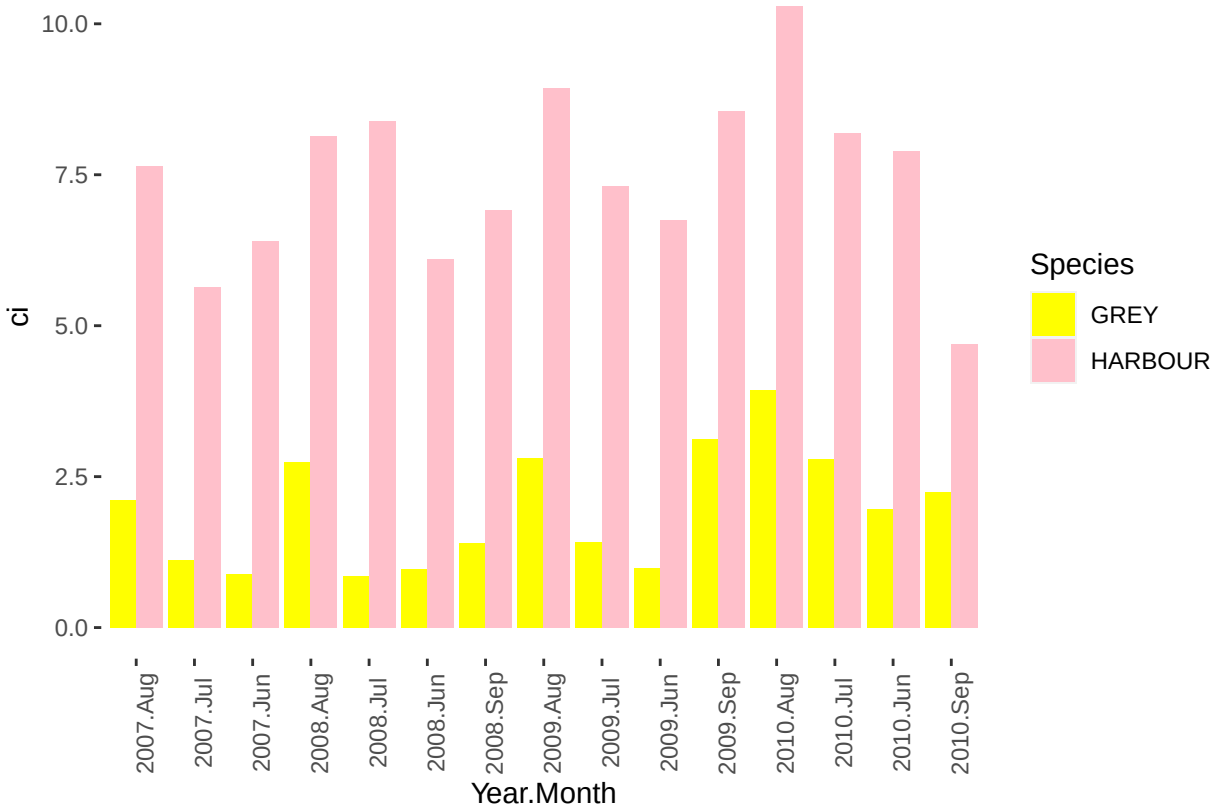
the existence of herbal species are way more than the gray species. Therefore we can say that the herbal species are more common than the gray species.In order to find the significant differences between the species we need to find the error species for particular month in each year using SummarySE() function.

Year.Month	Species	N	average.count	sd	se	ci
2007.Aug	GREY	7	1.2857143	2.2886885	0.8650430	2.1166839
2007.Aug	HARBOUR	7	6.3428571	8.2679818	3.1250034	7.6466079
2007.Jul	GREY	7	0.8714286	1.2051477	0.4555030	1.1145757
2007.Jul	HARBOUR	7	5.1714286	6.0939080	2.3032807	5.6359249
2007.Jun	GREY	7	0.7142857	0.9599107	0.3628121	0.8877693
2007.Jun	HARBOUR	7	4.2714286	6.9254878	2.6175883	6.4050079

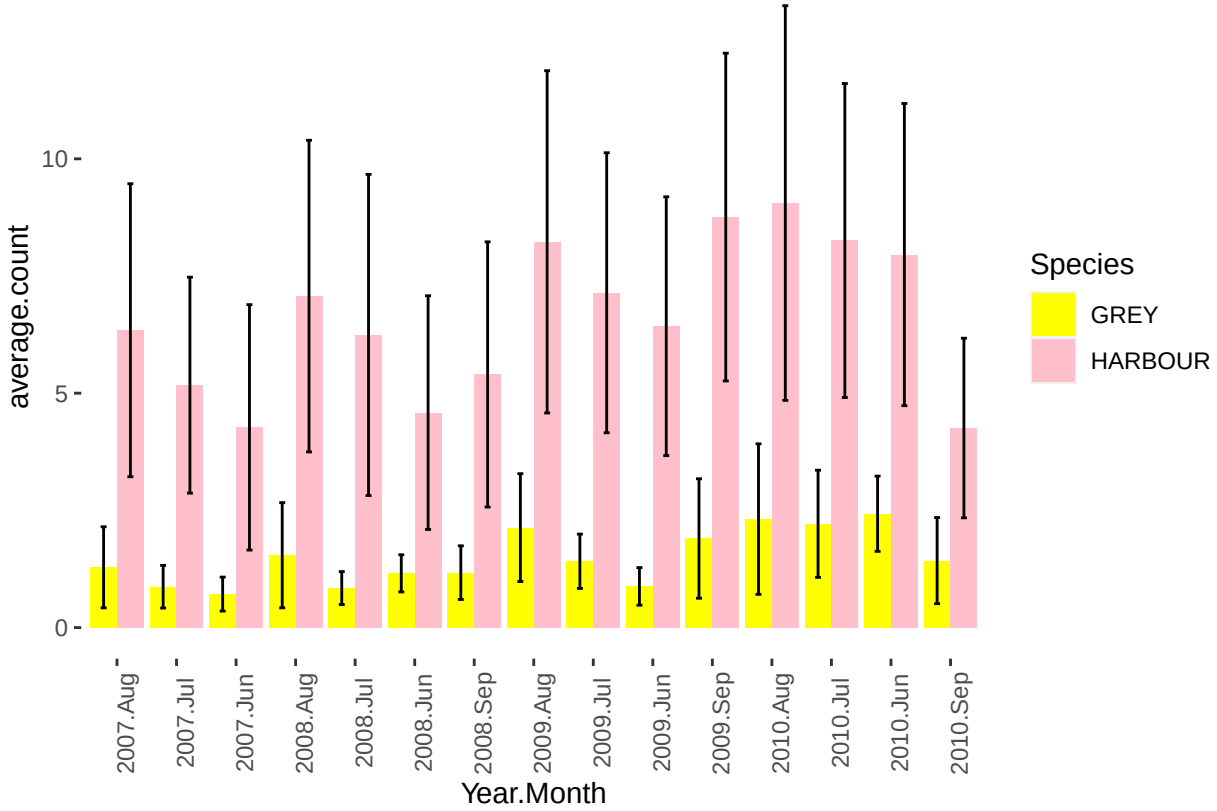
The Below Graph represents the standard deviation,Standard Error and CI of each of the species in particular months in a years. Considering the graph we can see that the In August 2010 the standard deviation seems to be very far from the mean value but in 2009 June and July, The SD values are very close to with mean value.



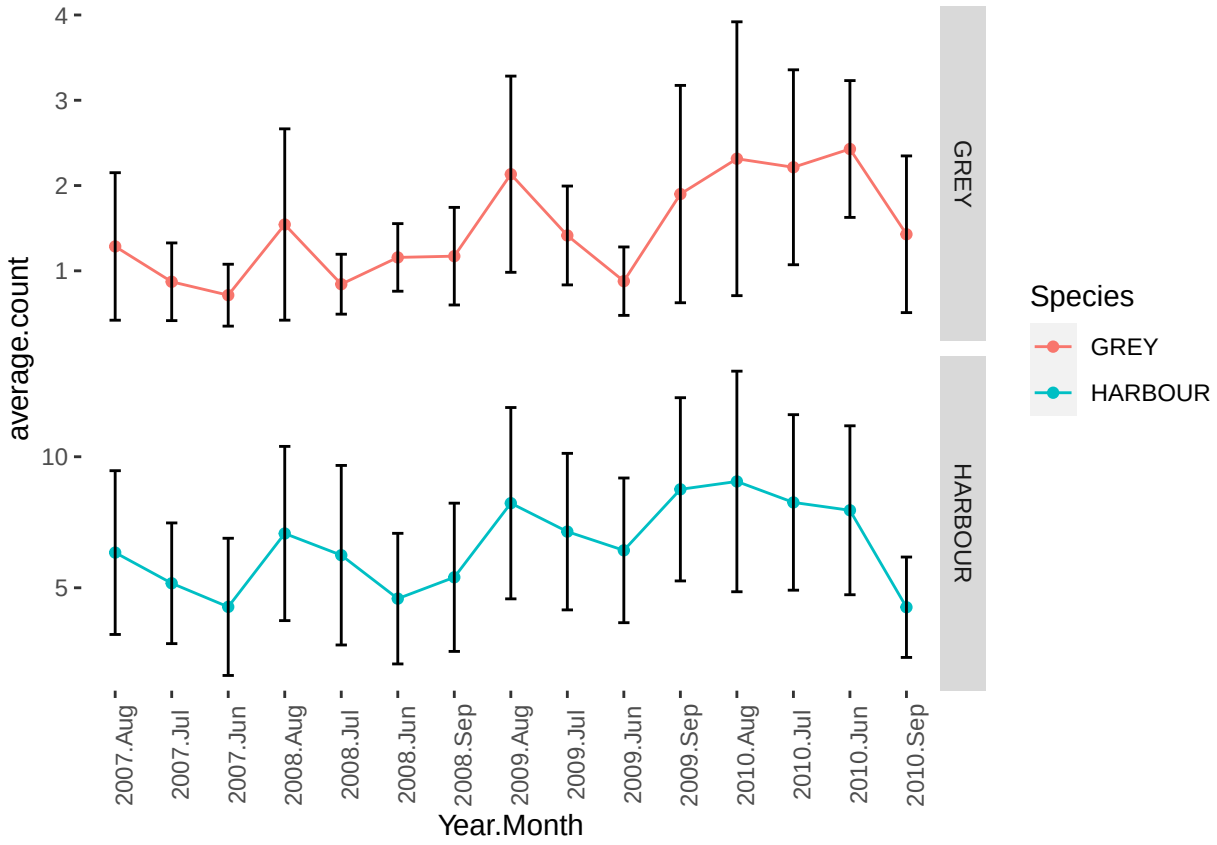

4
3
2
1
0
se
2007.Aug
2007.Jul
2007.Jun
2008.Aug
2008.Jul
2008.Jun
2008.Sep
2009.Aug
2009.Jul
2009.Jun
2009.Sep
2010.Aug
2010.Jul
2010.Jun
2010.Sep
Year.Month
Species
GREY
HARBOUR



If we check summary data we can find that the date of it consists of count, mean, standard deviation, standard error of mean and confidence interval. in order to find more detailed information we can plot the significant errors graph using ggplot2 in order to find significant difference between each species



Considering the above graph clearly shows that the graph provides an error bar which significantly shows how each species is using standard error. This clearly signifies difference species for a particular month.



3 RESULTS

In the above test, we have analyzed the significant difference between the species and the count values. We have different nominal variables and only one measure of variables, however we are able to identify non-significant values of the population of seals over time.
Moreover, when comparing data according to the year wise, we found out that the highest seal was obtained in 2010 and there were high counts in summer 2007. But they also found out that the levels of variance in the data particularly word not explored. Therefore we started exploring data for each month in a year in order to find out the presence of seals, but the test suggested that there is no significant difference and different months for each year. Moreover, Started analyzing particular species for each year. Initially we had seen that the harbor seals are more common than grey seals of the Tees Estuary, after that we found out 2009 and 10 the data appeared to be more significant 2007 and 2008 the data was slightly significant. After that, we calculated the standard error, and visualized the data which represented how the variables are using standard errors.

Considering the data and the above factors, In Tees Estuary after the environmental stabilization the growth of the seals has seen more rapid particularly for the harbor species.In the four year analysis the harbor seals is being colonized within the Tees estuary and is being healthy growing and could be predicted growth rate of 50% over next 5 years where as Grey seals growth rate is sable or slightly increasing each year could be increased by 20% to 30% over next 5 years. However we can say that, The major cause of disappearing of seals was due environmental degradation, if we maintain a eco-friendly environment we can see rapid increment in growth rate of species within the Tees estuary.

By summing up, we can say that the data is well explored, analyzed and ready to be processed for modeling which will help to give better results for prediction.

4 DISCUSSION

In the above data, We were able to successfully explore and analyse and find out the normality distribution of existence for both the species. According to the non parametric test, the most common use of Kruskal Wallis Test is when the data has one nominal variable and one measurement variable but test does not assume that a data han is well distributed and completely align for two parameters which can be mean and standard deviation and also it is called as one way anova(Biostathandbook, One Way Annovas). also this test assumes that the null hypothesis of mean groups are same. therefore if distribution groups are same, the Kruskal wallis test will not show a significant difference in their distribution. Yet, the test does not considered that the data are normally distributed, which can be a big advantage but 800 data has different groups different variants the test will give inaccurate result. Therefore, if the distribution is different and variant is different we can use anova test for accurate results.

4.0.1 REFERENCES

- FITTING DISTRIBUTIONS WITH R by Mr Vito Ricci.
  Avialable from: https://mran.revolutionanalytics.com/snapshot/2015-11-22/doc/contrib/Ricci-distributions-en.pdf [Accessed at 20th December 2020]
- Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests by Nornadiah Mohd Razali and Bee Wah Yap, Avialable from: https://www.researchgate.net/profile/Bee_Yap/publication/267205556_Power_Comparisons_of_Shapiro-Wilk_Kolmogorov-Smirnov_Lilliefors_and_Anderson-Darling_Tests/links/5477245b0cf29afed61446e1/Power-Comparisons-of-Shapiro-Wilk-Kolmogorov-Smirnov-Lilliefors-and-Anderson-Darling-Tests.pdf [Accessed at 21st December 2020]
- THE KRUSKAL–WALLLIS TEST, TEODORA H. MEHOTCHEVA Available from : http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/presentations/Mehotcheva-2008-Kruskal-Wallis.pdf [Accessed at 21st December 2020]
- The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR), Thorsten Pohlert Available from: https://cran.microsoft.com/snapshot/2014-09-08/web/packages/PMCMR/vignettes/PMCMR.pdf [Accessed at 22nd December 2020]
- Handbook of Biological Statistics, John H. McDonald Available from : http://www.biostathandbook.com/kruskalwallis.html [Accessed at 23d December 2020]