Project

2025-08-30

Project

In this project, we consider data taken from Distribution, climatic relationships, and status of American pikas (Ochotona princeps) in the Great Basin, USA". The main data set is contained in "Appendix 3: Master Database – Extant Great Basin Pika Sites, Published and Unpublished Sources."

First, we start by reading data from a csv file. This file contains data about location of population of pikas in The Great Basin, USA.The variables are Mountain Range, Location, State, Coordinates (Latitude, Longitude), Elevation site Specific or Mean (in meters), Aspect (measured in degrees), Surveyyear, and citation/Source of the data.

Reading the csv file with data about pikas

Used libraries: #library(readr) #library(pdftools)
#library(dplyr) #library(ggplot2)

```
library(readr)
library(pdftools)           # Load the library

pikas<-read.csv('Appendix_3CSV.csv')
```

We are going to focus on the following qualitative variables:

1) Mountain Range,
2) Location,
3) State,

and the following quatitative data:

1) Coordinates,
2) Elevation,
3) Aspect.

In the following table, we give a summary of the data.

```
summary(pikas)
```

```
##       ID            Mountain_Range       Location            State          
##  Length:2387        Length:2387        Length:2387        Length:2387       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##    Latitude          Longitude         Elevation_Site_Specific
##  Length:2387        Length:2387        Length:2387            
##  Class :character   Class :character   Class :character       
##  Mode  :character   Mode  :character   Mode  :character       
##    elev_low          elev_high            Aspect           surveyyear       
##  Length:2387        Length:2387        Length:2387        Length:2387       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

```
##    citation
##  Length:2387
##  Class :character
##  Mode  :character
```

Next, we select the variables we are interested in and we are going to convert numerical data from “character” to “numeric”. We also remove some data that does not make sense in the context of the problem.

```
library(dplyr)
#pikas_clean is the new data set we are going to use from now on.
#We start by choosing variables of interest.
pikas_clean <- select(pikas,
    Mountain_Range,
    Location,
    State,
    Latitude,
    Longitude,
    Elevation_Site_Specific,
    Aspect,
    surveyyear
)
```

```
#We convert this data into numeric data.
#Converting latitude, longitude, elevation, and aspect into numerical data:

pikas_clean$Elevation_Site_Specific = as.numeric(pikas_clean$Elevation_Site_Specific)

pikas_clean$Latitude = as.numeric(pikas_clean$Latitude)

pikas_clean$Longitude = as.numeric(pikas_clean$Longitude)

pikas_clean$Aspect = as.numeric(pikas_clean$Aspect)

pikas_clean$surveyyear = as.numeric(pikas_clean$surveyyear)

#If we check the data, there are some data in Aspects (in degrees) that are out of the range from 0 to

pikas_clean <- pikas_clean %>% filter(Aspect>0 & Aspect<=360)
#pikas
```

Next, we give a summary of the quantitative variables of the data:

```
summary(pikas_clean)
```

```
##  Mountain_Range       Location            State              Latitude
##  Length:2358        Length:2358        Length:2358        Min.   :36.77
##  Class :character   Class :character   Class :character   1st Qu.:38.05
##  Mode  :character   Mode  :character   Mode  :character   Median :40.51
##                                                           Mean   :39.88
##                                                           3rd Qu.:41.54
##                                                           Max.   :42.77
##
##    Longitude      Elevation_Site_Specific     Aspect          surveyyear
##  Min.   :-120.6   Min.   :1766            Min.   :  1.00   Min.   :2005
##  1st Qu.:-119.6   1st Qu.:1978            1st Qu.: 90.25   1st Qu.:2011
##  Median :-119.3   Median :2678            Median :188.00   Median :2014
```

```
##  Mean   :-118.2   Mean   :2581            Mean   :189.11   Mean   :2013
##  3rd Qu.:-118.3   3rd Qu.:3109            3rd Qu.:288.00   3rd Qu.:2015
##  Max.   : 120.3   Max.   :4009            Max.   :359.00   Max.   :2017
##                                                            NA's   :90
```

From the survey we can see that the range of time of the collected data goes from 2005, until 2017.

We want to see how the min, max, mean elevation of extant pikas has change with respect to time in the Sierra Nevada range in California.

```
#We choose just Sierra Nevada data about elevation and surveyear
pikasSN <- pikas_clean %>%
    select(
    Mountain_Range,
    Elevation_Site_Specific,
    surveyyear) %>%
    filter(Mountain_Range == 'Sierra Nevada')

head(pikasSN)
```

```
##   Mountain_Range Elevation_Site_Specific surveyyear
## 1  Sierra Nevada                    3436       2016
## 2  Sierra Nevada                    3164       2016
## 3  Sierra Nevada                    3284       2016
## 4  Sierra Nevada                    3091       2016
## 5  Sierra Nevada                    3100       2016
## 6  Sierra Nevada                    3287       2016
```

Question 1: Year, Min, Max, and Mean vs. elevation

Now we are going to create a new data set with min, max, and mean elevation per surveyear.

```
#First, we note that information about 2006 is not useful
pikasSNyear <- filter(pikasSN, surveyyear == 2006)
#pikasSNyear
min(pikasSNyear$Elevation_Site_Specific)
```

```
## [1] Inf
```

```
#Hence, we are going to remove it in the next code block (see line: pikas_mMm1 <- pikas_mMm[-2,] below)
```

```
#This creates vectors containing min, max, and mean of elevation_Site_Specific per year in Sierra Nevad

min_ve <- numeric(0) #This vector will have the mininum elevation per year value in each entry.
max_ve <- numeric(0) #This vector will have the maximum elevation per year value in each entry.
mean_ve <- numeric(0) #This vector will have the mean elevation per year value in each entry.

for (year in (2005:2017)){
  #From our data pikasSN, we take an speifici year.
  pikasSNyear <- filter(pikasSN, surveyyear == year)
  #we compute the min, max and mean elevation value for 'year' and we save this information in min_ve, 
  min_ve <-append(min_ve, min(pikasSNyear$Elevation_Site_Specific))
  max_ve <-append(max_ve, max(pikasSNyear$Elevation_Site_Specific))
  mean_ve <-append(mean_ve, mean(pikasSNyear$Elevation_Site_Specific))
}
min_ve
```

```
##  [1] 2581  Inf 2067 1827 2191 2653 2038 1837 2620 2453 2217 2305 2513
```

```r
max_ve
```

```
##  [1] 3005 -Inf 3953 3620 3574 3509 3527 3438 3580 3510 3427 3769 2555
```

```r
mean_ve
```

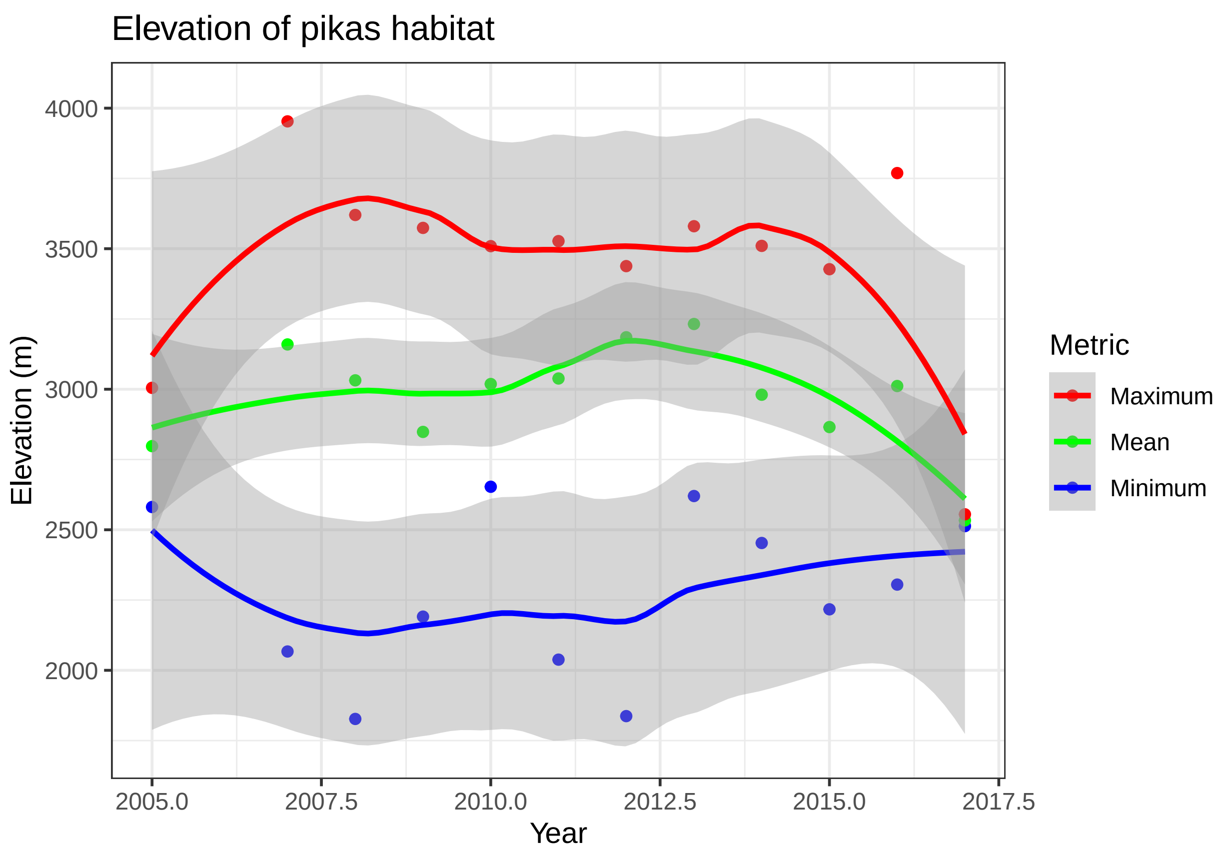
```
##  [1] 2797.667      NaN 3159.364 3031.632 2848.174 3018.571 3038.153 3184.882
##  [9] 3232.050 2980.460 2865.465 3011.437 2533.333
```

```r
#Next, we add a new column in our data called 'year'
year <- c(2005:2017)
#We create a data frame with the vectors above and 'year'
pikas_mMm <- data.frame(year,
  Minimum = min_ve,
  Maximum = max_ve,
  Mean = mean_ve
)
#We remove row two since it does not contain useful data.
pikas_mMm1 <- pikas_mMm[-2,]

head(pikas_mMm1)
```

```
##   year Minimum Maximum     Mean
## 1 2005    2581    3005 2797.667
## 3 2007    2067    3953 3159.364
## 4 2008    1827    3620 3031.632
## 5 2009    2191    3574 2848.174
## 6 2010    2653    3509 3018.571
## 7 2011    2038    3527 3038.153
```

Now we make a plot with all the new data frame created above.

```r
library(ggplot2)

#Plot year vs Minimum, maximum, and mean
ggplot(data = pikas_mMm1) +
  theme_bw() +
  #we add point corresonding to min, max, mean values
  geom_point(aes(x = year, y = Minimum, color = "Minimum"))+
  geom_point(aes(x = year, y = Mean, color = "Mean"))+
  geom_point(aes(x = year, y = Maximum, color = "Maximum"))+
  #we create curves corresonding to data from min, max, mean values
  geom_smooth(aes(x = year, y = Minimum, color = "Minimum")) +
  geom_smooth(aes(x = year, y = Mean, color = "Mean") ) +
  geom_smooth(aes(x = year, y = Maximum, color = "Maximum")) +
  labs(
    title = "Elevation of pikas habitat",
    x = "Year",
    y = "Elevation (m)",
    color = "Metric"
  ) +
  scale_color_manual(values = c("Minimum" = "blue",
                                "Mean" = "green",
                                "Maximum" = "red"))
```



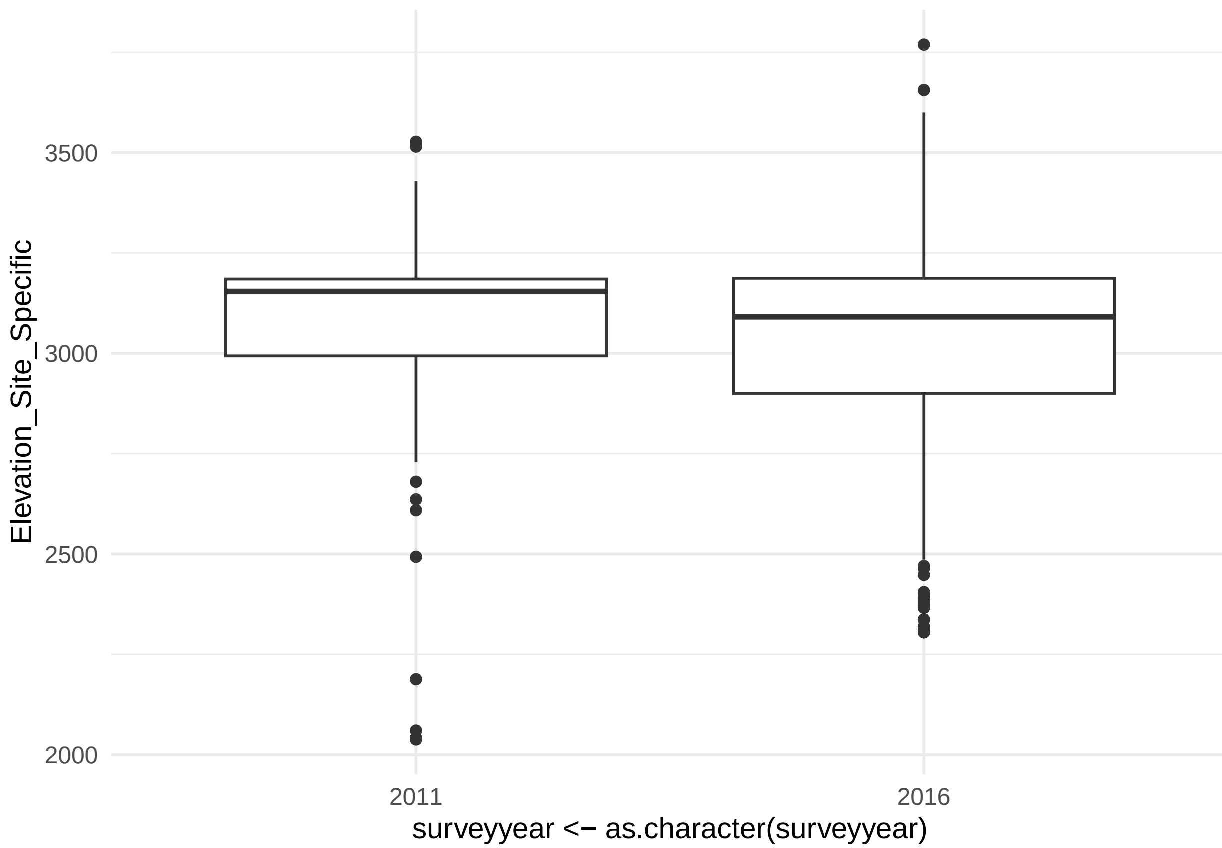
This graph suggest that in 12 year (i.e. from 2005 to 2017), the elevation where pikas have been observed haven't changed considerably since the mean has stayed in the interval [2500 m, 3500 m]. ## Boxplot and t-test

In order to analyze further the how elevation of extant site of pikas has changed over time, we can perform a t-test analysis to compare elevation means. We choose year 2011 and 2016 since we have more data for these two years.

```
pikasSNy <- pikasSN %>% filter(surveyyear == 2011 | surveyyear == 2016) #this data set contains data ab

head(pikasSNy)
```

```
##   Mountain_Range Elevation_Site_Specific surveyyear
## 1  Sierra Nevada                    3436       2016
## 2  Sierra Nevada                    3164       2016
## 3  Sierra Nevada                    3284       2016
## 4  Sierra Nevada                    3091       2016
## 5  Sierra Nevada                    3100       2016
## 6  Sierra Nevada                    3287       2016
```

```
#We plot boxplots to see graphically the mean of each sample (i.e. one corresponding to 2014 and anothe
ggplot(pikasSNy) +
  aes(x = surveyyear <- as.character(surveyyear), y = Elevation_Site_Specific) +
  geom_boxplot() +
  theme_minimal()
```



The boxplot make us infer that the medians of the two samples are different. By using a t-test we can compact the means

```
#First we compute the variance of each set of data:

pikas2016 <- filter(pikasSN, surveyyear == 2016)
var(pikas2016$Elevation_Site_Specific) #variance of data for 2016
```

```
## [1] 84159.06
```

```
pikas2016 <- filter(pikasSN, surveyyear == 2011)
var(pikas2016$Elevation_Site_Specific) #variance of data for 2011
```

```
## [1] 109507.2
```

```
#we apply t-test to our samples comparing elevation in 2011 vs 2016
diff <- t.test(Elevation_Site_Specific ~ surveyyear,
  data = pikasSNy,
  var.equal = FALSE #we use var.equal = FALSE since variance is different.
)
diff
```

```
##
##  Welch Two Sample t-test
##
## data:  Elevation_Site_Specific by surveyyear
## t = 0.5738, df = 78.638, p-value = 0.5677
## alternative hypothesis: true difference in means between group 2011 and group 2016 is not equal to 0
## 95 percent confidence interval:
```

```
##  -65.96481 119.39582
## sample estimates:
## mean in group 2011 mean in group 2016
##           3038.153           3011.437
```

Since the p-value is greater than 0.05, we cannot reject our null-hypothesis, i.e., that the two mean values of elevation in 2011 and 2016 are different. Hence we cannot conclude that the two means are different. In fact, if we change the value of the INITIAL YEAR to another value different than 2011, we will still get p-values less than 0.05, hence we cannot say that the elevation mean has either increase or decrease during the these two particular years in Sierra Nevada.

Question 2: Elevation vs. Aspect

We want to study if the Elevation depends on the Aspect variable. We can make a linear model between Elevation vs. Aspect in order to see if there exist a linear correlation:

```r
#here our independent variable is aspect and our dependent variable is Elevation_Site_Specific
elevation_aspect <- lm(data = pikas_clean, formula = Elevation_Site_Specific ~ Aspect)
class(elevation_aspect)
```

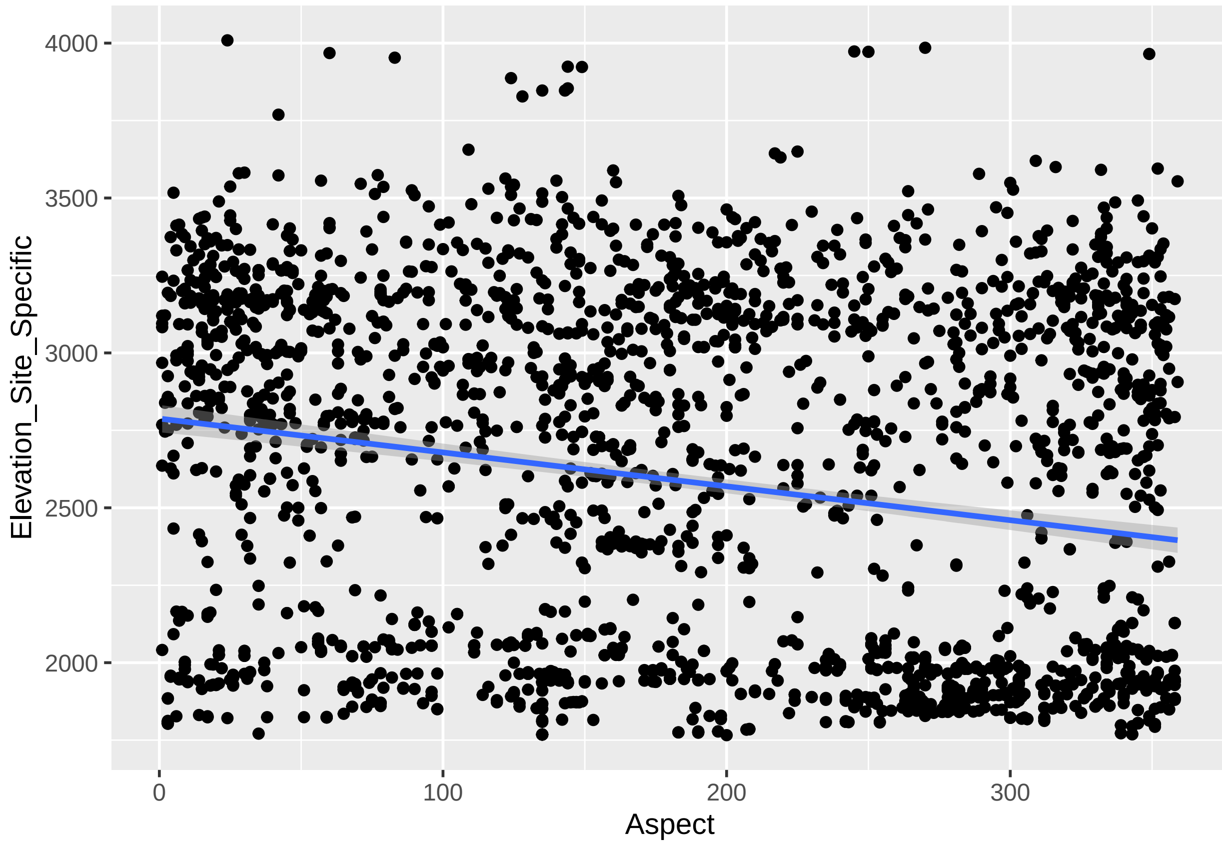
```
## [1] "lm"
```

```r
summary(elevation_aspect)
```

```
##
## Call:
## lm(formula = Elevation_Site_Specific ~ Aspect, data = pikas_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -981.71 -550.09   58.72  486.27 1558.65
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2787.9912    22.5397  123.69   <2e-16 ***
## Aspect        -1.0935     0.1028  -10.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 554.6 on 2356 degrees of freedom
## Multiple R-squared:  0.04587,    Adjusted R-squared:  0.04546
## F-statistic: 113.3 on 1 and 2356 DF,  p-value: < 2.2e-16
```

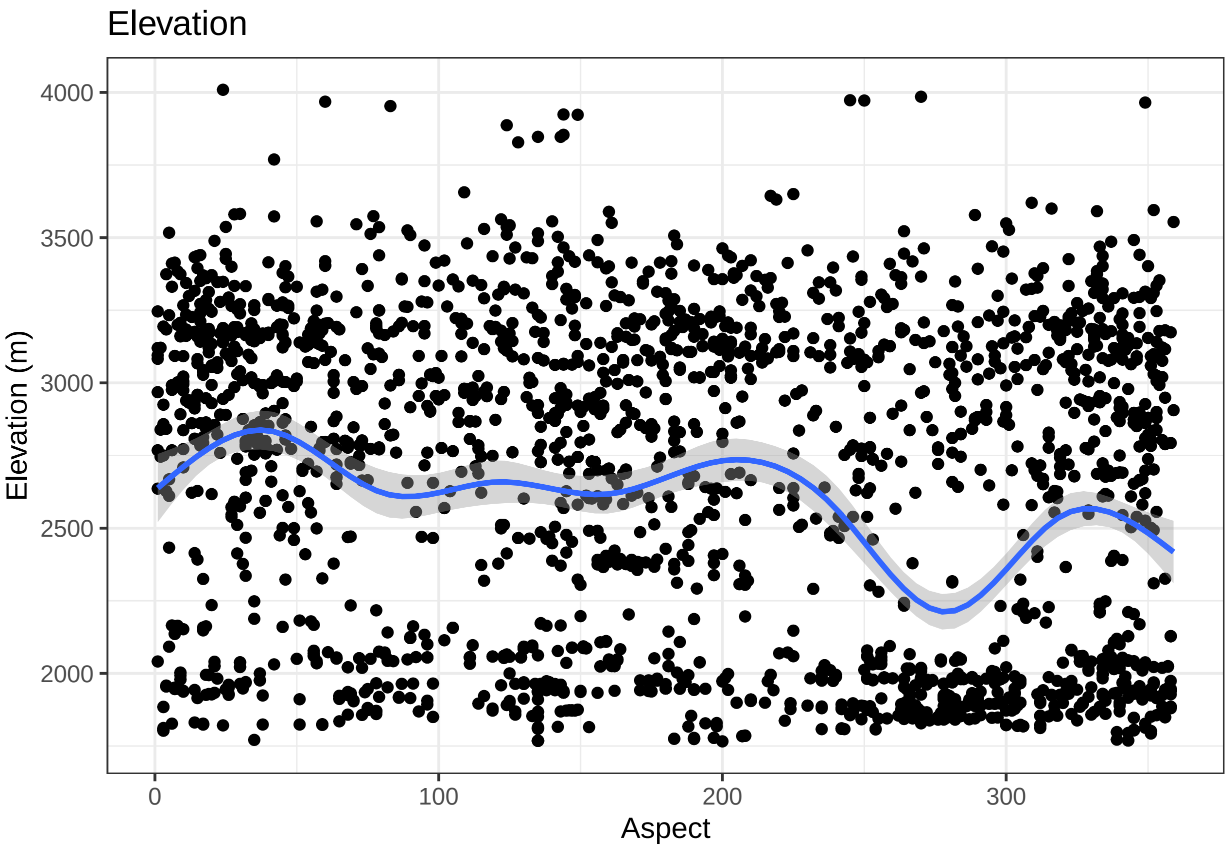
Although the p-value and the adjusted R-squared are both less than 0.05, the Residual standard error is BIG. To see this graphically we make a plot of this data:

```r
#
ggplot(
    data = pikas_clean,
    mapping = aes(x = Aspect, y = Elevation_Site_Specific, color = )
    ) +
    geom_point() +
    geom_smooth(method = "lm", formula = y ~ x)
```



This is reasonable if we take into account that Aspect is a 'periodic variable', since 360 degrees = 0 degrees. Since the linear model does not fit the data, we can use geom_smooth in order to study graphically the relationship between Elevation and Aspect

```
#We plot
ggplot(data = pikas_clean) +
  theme_bw() + geom_point(aes(x = Aspect, y = Elevation_Site_Specific, color=)) +
  geom_smooth(aes(x = Aspect, y = Elevation_Site_Specific, color=)) +
  labs(
    title = "Elevation",
    x = "Aspect",
    y = "Elevation (m)")
```



We can see that the elevation attaches its maximum pick at ~45 degrees and its minimum at ~270 degrees. From this we could infer that the the position of the pikas habitat affects the elevation of its habitat. However, to make a valid conclusion, we suggest to make a nonlineal model for future studies. In this case, due to to the periodicity of 'Aspect', a trigonometric model would be a good choice.