Biodiversity for the National Parks

Valerii Popovych

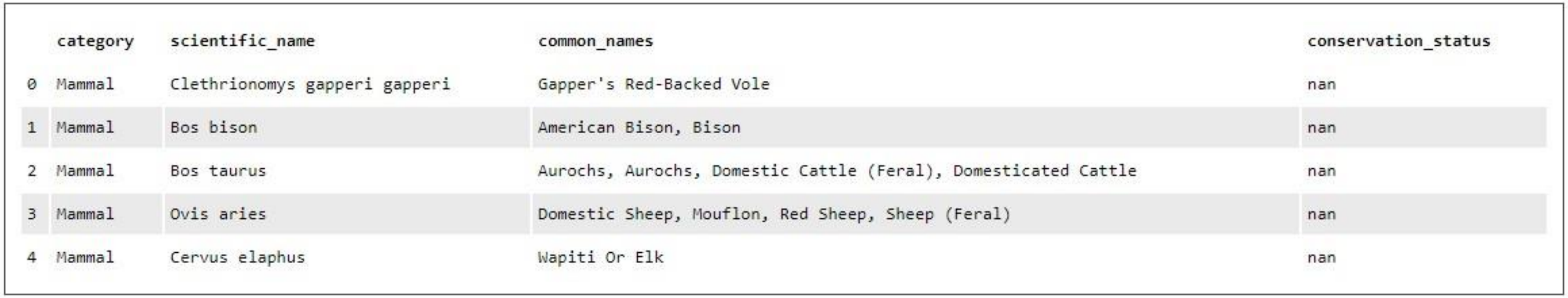
06/05/2017

Import *species_info.csv*

- At first, we upload the “species_info.csv” file, in which was gathered the information about different species in National Parks

- The file contains *“category”*, *“scientific_name”*, *“common_names”*, *“conservation_status”* columns and 5824 rows (incl. row with column names)

- Obviously, we want to take a helicopter view on general situation to start.



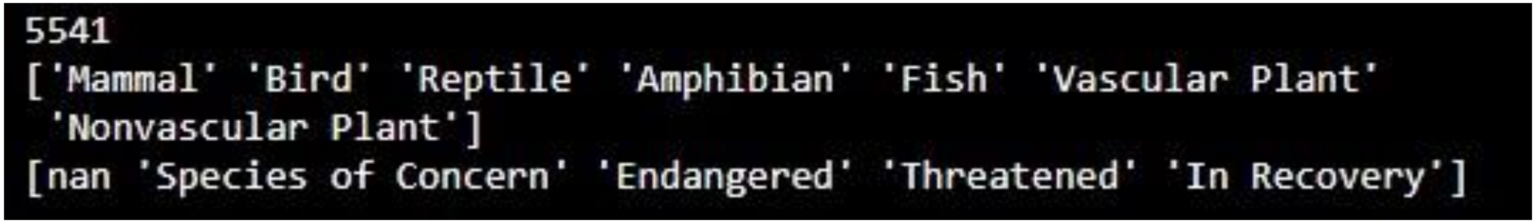
	category	scientific_name	common_names	conservation_status
0	Mammal	Clethrionomys gapperi gapperi	Gapper's Red-Backed Vole	nan
1	Mammal	Bos bison	American Bison, Bison	nan
2	Mammal	Bos taurus	Aurochs, Aurochs, Domestic Cattle (Feral), Domesticated Cattle	nan
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	nan
4	Mammal	Cervus elaphus	Wapiti Or Elk	nan

- Looks like the information couldn’t be analyzed manually – too many rows! Let’s use some simple and helpful comands.

Analyzing *species_info.csv*

- Let’s be short and focused:
  - In National Parks exist 5 541 unique biological species in total;
  - Main biological categories, which can be easily understand by regular human being (unlike the “scientific_names” column) are: "Mammal”, "Bird“, "Reptile“, "Amphibian", "Fish“, "Vascular Plant“ and "Nonvascular Plant“
  - Also, regarding the most important column “conservational_status”. Each species can have one of the next conservational statuses: “Species of Concern”, “Endangered”, “Threatened”, “In Recovery”, “nan”.

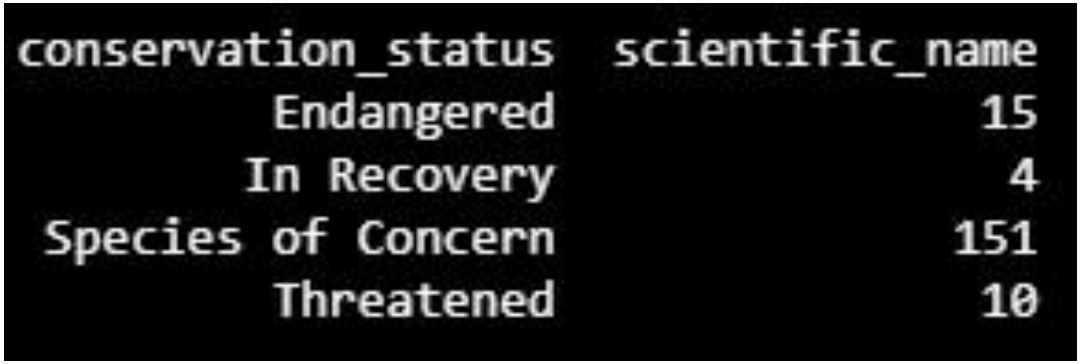
- For now we are not sure, but we can assume that the “nan” conservational status refers to the the species which are relatively safe and are not close to extinction now or in the nearest future.



```
5541
['Mammal' 'Bird' 'Reptile' 'Amphibian' 'Fish' 'Vascular Plant'
 'Nonvascular Plant']
[nan 'Species of Concern' 'Endangered' 'Threatened' 'In Recovery']
```

Digging deeper into Endangered Species - 1

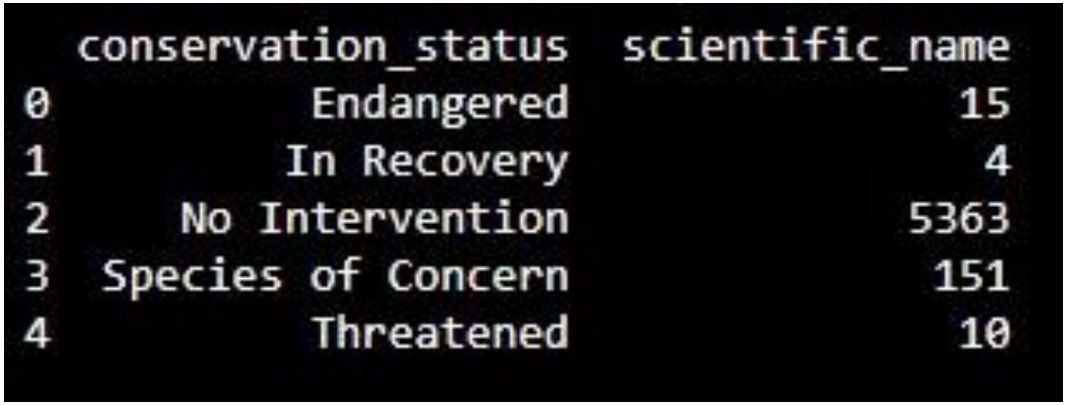
- On the previous slide we received the information, which is really important for our task. Since we do have 4 different types of Endagered Species, let’s look closer on them!
- With simple *groupby* function we want to see, what Conservation Status is the most populated with species.
- Good news - we can see, that among all Endangered Species most of them goes to group “Species of Concern”, which means we have time to save them if we start do something **now.**



```
conservation_status  scientific_name
         Endangered               15
        In Recovery                4
 Species of Concern              151
         Threatened               10
```

Digging deeper into Endangered Species - 2

- Obviously, there is a huge diference between number of total unique species in the table and number of Endangered Species. What is the reason of this difference?
- Oh yes, probably we were right about “nan” Conservation Status. This status corresponds with species that are actually **not** in danger.
- To have a clearer picture, we can provide all non-endangered species with new Conservation Status “No Intervention”. Let’s look on the grouped table again:

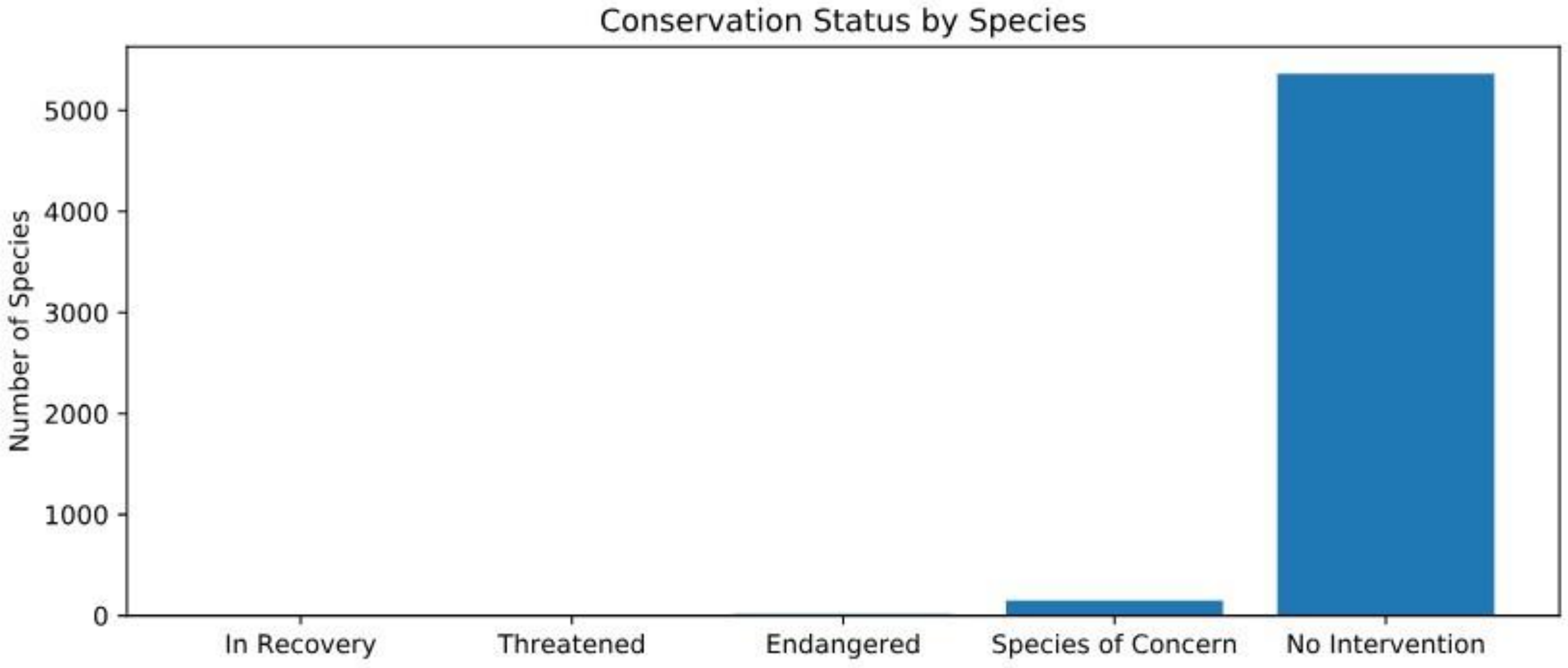


```
  conservation_status  scientific_name
0          Endangered               15
1         In Recovery                4
2     No Intervention             5363
3  Species of Concern              151
4          Threatened               10
```

- Great news! The total number of endangered species is **very small!** Let’s be honest, if they extinct – barely anyone will spot this.

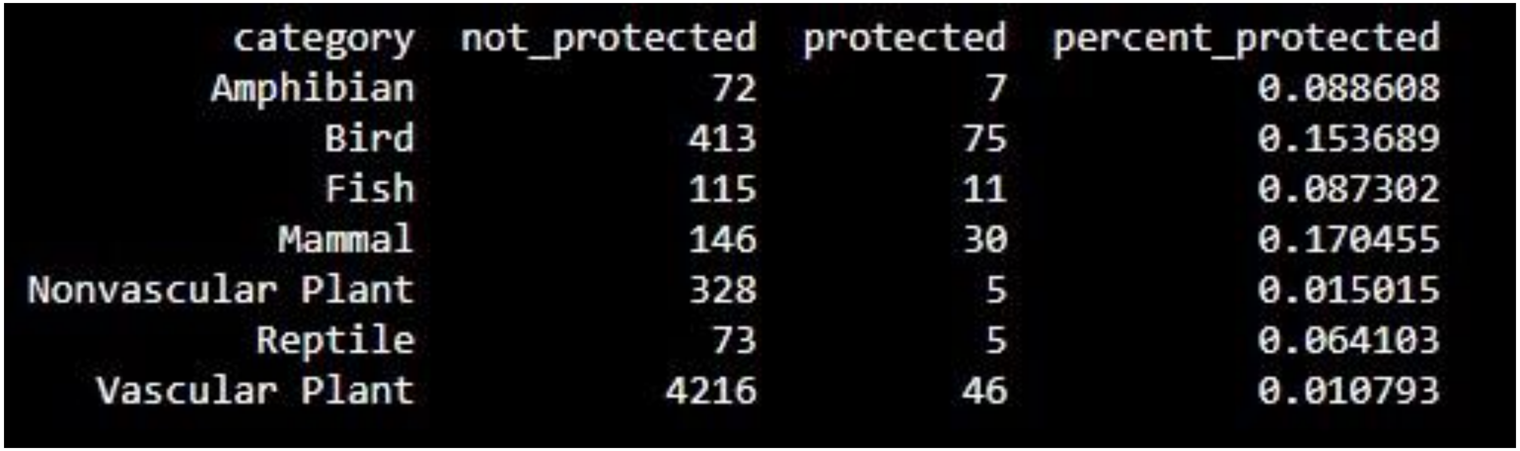
First strong visualisation

- As Data Analysts we understand, that there is no better way to explane complex figures, than to present them in a pretty and clear way.
- So let’s built a graph and show first results of our project. Luckily for us, the graph is **extremely** self-explanatory and shows that for most species in National Parks there is no danger at the moment.



Endangered Species - 1

- Going further – let’s understand how many species are Endangered and how many are not **inside each biological category**
- For this, we create the special column is_protected which helps us to distinguish “No Intervention” and the rest (which are endangered in some way).
- We receive quite simple table but for some conclusions let’s make one additional step. We can count the share of **protected** species in each category. For this we divide the number of protected species in category by **total** number of species.

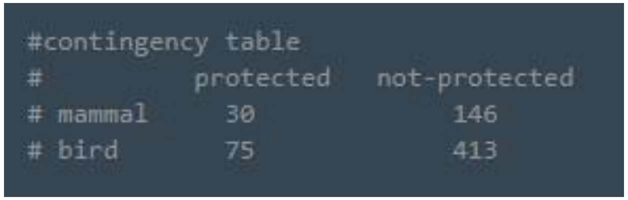


```
          category  not_protected  protected  percent_protected
         Amphibian             72          7           0.088608
              Bird            413         75           0.153689
              Fish            115         11           0.087302
            Mammal            146         30           0.170455
 Nonvascular Plant            328          5           0.015015
           Reptile             73          5           0.064103
    Vascular Plant           4216         46           0.010793
```

- Bird and Mammal looks weird, right?

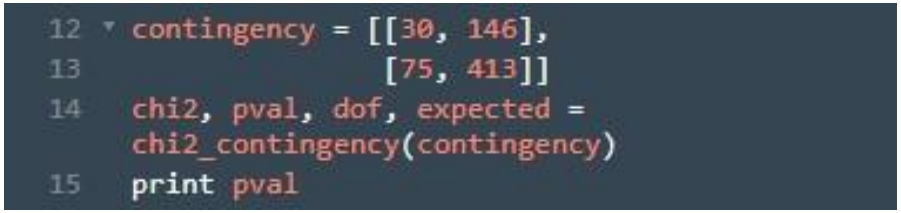
Endangered Species - 2

- Real math, finally! Looks like the Bird and Mammal is a good place to start. On previous slide we mentioned that these categories are both close to be endangered, but does the difference is signifficant? Let’s find out!



```
#contingency table
#           protected    not-protected
# mammal       30             146
# bird         75             413
```

- Using the Chi-Squared Test and received the p-value = 0.687594809666 – which is more than 0.05, so we can **confirm our null hypothesis.** Our null hypothesis was “the difference between Endangered Level of Mammal and Bird is due to chance”

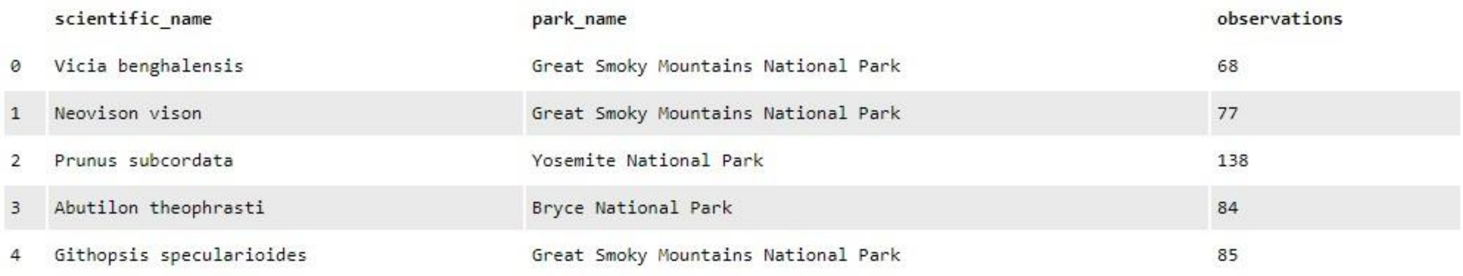


```
contingency = [[30, 146],
               [75, 413]]
chi2, pval, dof, expected = chi2_contingency(contingency)
print pval
```

- Also, let’s make the same exercise for Mammal and Reptile categories. Whoa! Our ***pval_reptile_mammal*** is 0.0383555902297 which is **less** than 0.05
- Conclusion? Strong significant difference shows us that among Reptile and Mammale categories one of them is **really closer to extinction** and that’s not due to a chance, but due to some real reasons which somebody should found out.

New file: observations and sheeps

- Apparently, somebody did a good thing and made some observations in different National Parks. We were asked to help with some analytics job about the sheeps. Here’s how the data we received looks:

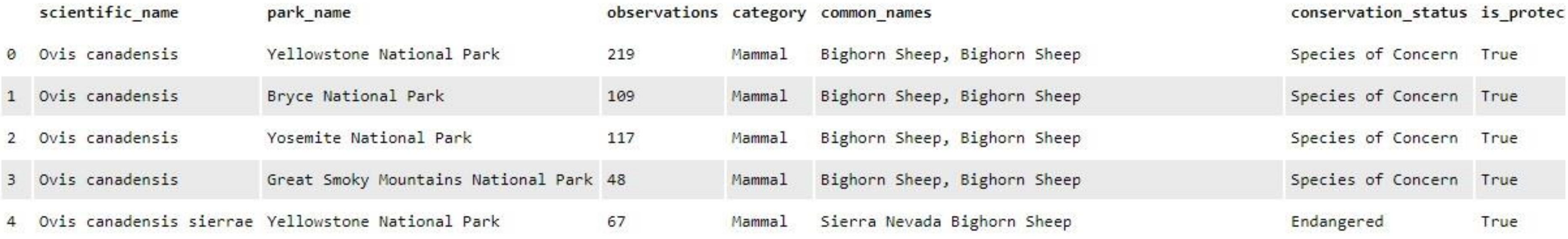


	scientific_name	park_name	observations
0	Vicia benghalensis	Great Smoky Mountains National Park	68
1	Neovison vison	Great Smoky Mountains National Park	77
2	Prunus subcordata	Yosemite National Park	138
3	Abutilon theophrasti	Bryce National Park	84
4	Githopsis specularioides	Great Smoky Mountains National Park	85

- As we can see, the “observations.csv” file contains only scientific names. They are totally irrelevant for us. But now we understand, that the column “scientific_name” wasn’t useless in the file “species_info.csv”. Let’s make a merge by the “scientific_name” column between “species_info.csv” and “observations.csv”
- So, at first we should define all rows in species which contains “sheep” and mark all these rows with a help of a new column “is_sheep”
- The rule is as simple as possible – if “common_names” has the word “sheep” – then the column “is_sheep” returns us **True.** Otherwise – **False.**
- Let’s select all rows, which have “True” in the column “is_sheep”

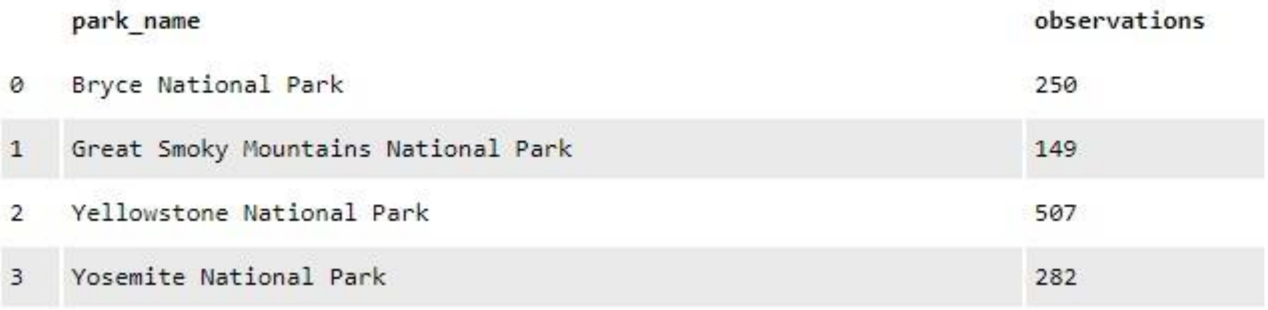
Merging “observations” and “species”

- Lets’ merge two tables with a right join and receive the merged and filtered final table:



	scientific_name	park_name	observations	category	common_names	conservation_status	is_protec
0	Ovis canadensis	Yellowstone National Park	219	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True
1	Ovis canadensis	Bryce National Park	109	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True
2	Ovis canadensis	Yosemite National Park	117	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True
3	Ovis canadensis	Great Smoky Mountains National Park	48	Mammal	Bighorn Sheep, Bighorn Sheep	Species of Concern	True
4	Ovis canadensis sierrae	Yellowstone National Park	67	Mammal	Sierra Nevada Bighorn Sheep	Endangered	True

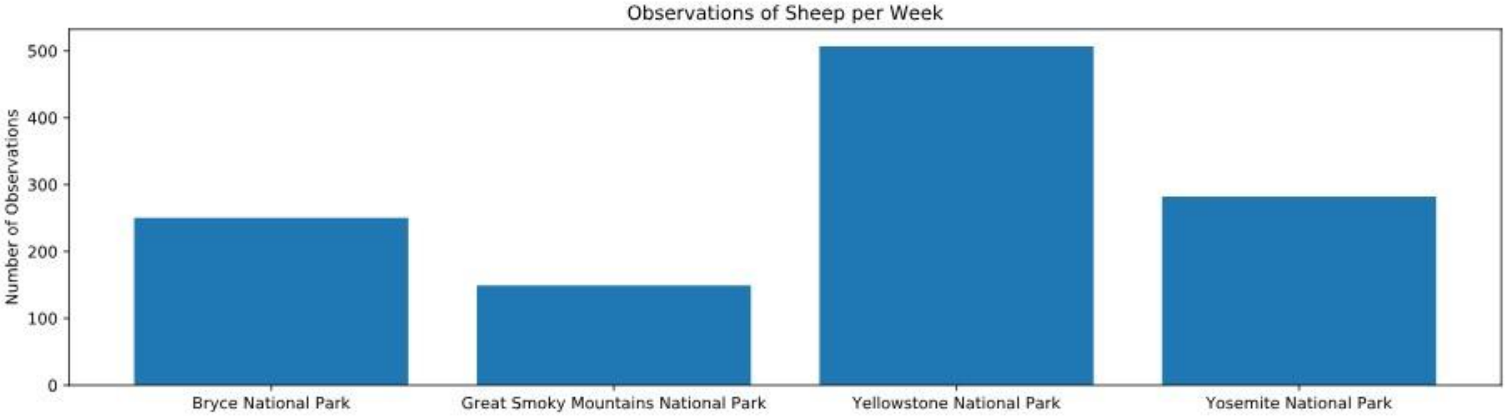
- The data seems to be correct, but not very easy to work with! Since we want to find out the observations made in each National Park – let’s group our table by a National Park creteria and count total number of observations made in each National Park.



	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282

Second Strong Visualisation

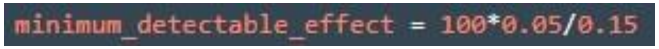
- As in situation before, the simple table often doesn’t express some conclusions clear enough. And as in situation before, let’s ask for a help from a bar chart:



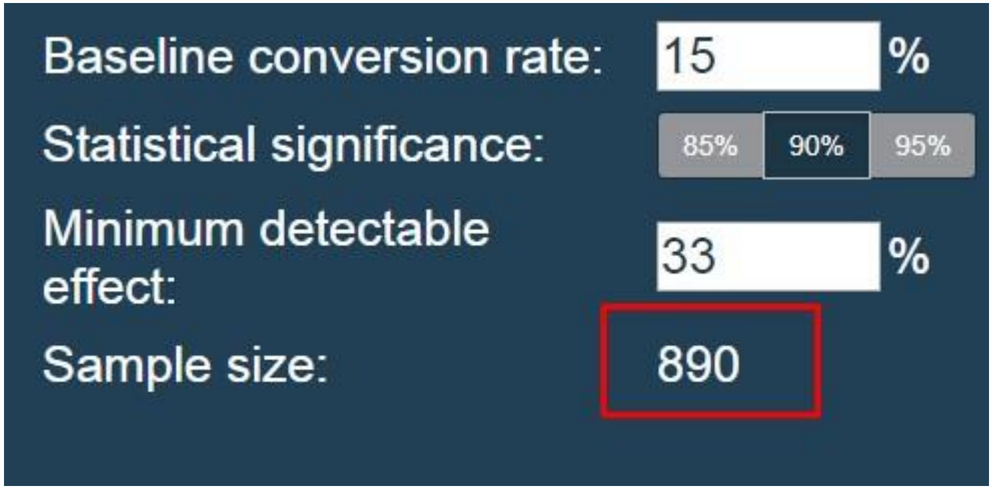
- Sheeps do like Yellowstone National Park, don’t they?

Foot and mouth disease: sample size

- Last thing: we found out, that among sheeps in different National Parks is distributed foot and mouth disease. We were asked to define, does the program of reducing foot and mouth disease working – and how much observations and time do the Park Rangers need to recceive the **reliable** answers?
- We know, that last year around 15% of sheeps were infected by foot and mouth disease. This is our baseline. Counting the MDE: and received 33%. So, the Park Rangers needs to observate **890** sheeps this year to confirm or deny the efficiency of new cure program.



```
minimum_detectable_effect = 100*0.05/0.15
```



Baseline conversion rate: 15 %

Statistical significance: 85% [✓ 90%] 95%

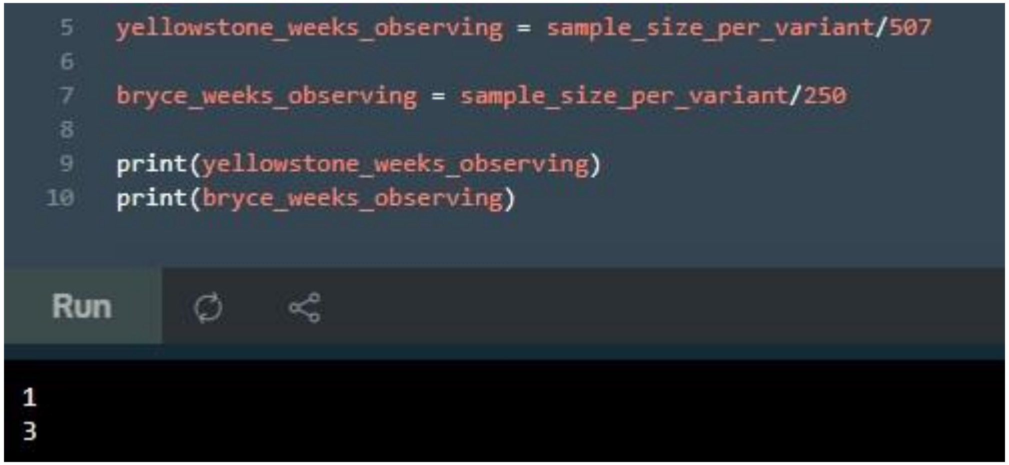
Minimum detectable effect: 33 %

Sample size: 890

Foot and mouth disease: timings

- And really last thing – how much time will new observations take? We can say, that Park Rangers will need:
  - 1 week (rounded) in Yellowstone National Park
  - and 3 weeks (rounded) in Bryce National Park…

- …To gather the enough Sample Size (890 observations) and understand the efficiency of the new cure program for sheeps.



```
yellowstone_weeks_observing = sample_size_per_variant/507

bryce_weeks_observing = sample_size_per_variant/250

print(yellowstone_weeks_observing)
print(bryce_weeks_observing)
```

```
1
3
```