**Module 1**

Machine learning is the study of all of the different ways in which models can be built from data.

AI -

Black Box:

Vector

Loss function: square of difference of actual result is y, predicted result is ŷ

The function which takes the actual result y and our predicted result ŷ and scores it as the square of the difference is known as the loss function L.

Probability and statistics

actual result is y, predicted result is ŷ

Y generic, y sample of output. Y ~ py Y has distribution py. [y]10 data set of sample size 10 = [y1, y2,…y10]

Discrete or continuous distribution

*You are correct! The answer “Non-deterministic system*” *is correct because such a system exhibits a different behavior every time it is sampled.*

*You are correct! The terms ‘distribution’ and ‘probability density function’ refer to the same thing.*

*That is correct! The answer “discrete distribution” is correct because in this system, the data can be only certain values such as integers, (0,1,2, and 3, for example).*

*You are correct! The answer “Continuous distribution“ is correct because in continuous distribution, the random variable X can have any value since there are infinite values X can take.*

*You are correct! The answer “True” is correct because the capital letter Y is used to denote the generic output of a system, and the samples of the output are denoted with a small letter y.*

*You are correct! The answer “Y is distributed according to a distribution Fy” is correct because Y is sampled using the function Fy.*

*You are correct! The answer “12” is correct because the subscript number in the question denotes how many unique outputs to return.*

E[Y] expected value, expectation, mean of Y or Py. Mean is avg, median is middle number

Small variance, large variance:graph will be wider and shorter.

stddev

*You are correct! The answer “One” is correct because the sum of the entire area under the curve will always equal one. Any area between two vertical lines running through the distribution will always be less than one.*

*You are correct! The answers “E[Y]” and “*μy*” are correct because these are the correct representations for the expected value of Y.*

*You are correct! The answer “True” is correct because although the mean and median measure different things, they can be the same value.*

*You are correct! The answer “True” is correct because as n grows, there are more data points to base the average (mean) calculations on, and therefore, it becomes a better representation of the expected value.*

*You are correct! The answer “True” is correct because variance is a measure of how widely distributed the values are. The larger the variance grows, the wider the graph will be.*

*You are correct! The answer “The square root of the variance” is correct because standard deviation is the square root of the variance.*

*You are correct! The answer “*𝛔y” *is correct because this is the correct symbol for the standard deviation.*

*The answer “*.loc[slice] *when the index is integers” is incorrect because when the index is integer the function*.loc[]*is preferable.*

Jupyter Notebook, key operations in pandas, including how to use arrays, create dataframes from dictionaries, and change indexes.

The statement .loc[[0:3],[’column1’,’column4’]] will return indexes 0 to 3 and columns from column1 and column4.

locinfo = data.loc[0]

teamdata = data.loc[data.Name == 'LeBron James', 'Name':'Team']

fr = data.iloc[0]

teamdata = data.loc[data.Name == 'LeBron James', 'Name':'Team']

sr = data.iloc[1]

lr = data.iloc[-1]

fc = data.iloc[:,0]

sc = data.iloc[:,1]

lc = data.iloc[:,-1]

firstname = data['Name'].str.split(expand=**True**).iloc[0,0]

firstname1 = data['Name'].str.split().str[1].iloc[0]

lastname = data['Name'].str.split(expand=**True**).iloc[0,1]

lastname1 = data['Name'].str.split().str[1].iloc[0]

**Quiz**

In which scenario is .loc[] not preferable to []? : **Correct Answer :** .loc[slice] when the index is non integers

What is the output of .loc[[5,6,9]]? : The output would be three rows of data: rows with the index of 5, 6, and 9.

*You are correct! The answer “The output would be three rows of data: rows with the index of 5, 6, and 9” is correct because the statement has a function*.loc[]*with defined indexes which are to be returned.*

What is the standard alias used when importing a pandas library? : pd

*You are correct! The answer “pd” is correct because this is the standard alias.*

In Python, a list of numbers can be represented with which of the following? *(Check all that apply.) :* Array, Tuple, Vector

*You are correct! In Python, a list of numbers can be represented with a vector, array, or tuple.*

The statement .loc[[0:3],[’column1’,’column4’]] will return indexes 0 to 3 and columns from column1 and column4. : False

*You are correct! The answer “False” is correct because the statement will return rows with indexes from 0 to 3, but will only return column1 and column4, not the columns in-between (column2 and column3).*

What will the output of DF[DF[’meters’]>2] be? : Return all the rows and columns where data in the column ‘meters’ is greater than 2.

*That is correct! The answer “Return all the rows and columns where data in the column ‘meters’ is greater than 2” is correct because the statement has a condition which states meters > 2.*

If you want to select the first four rows of a dataframe, how can you do this using slice indexing? : DF[:4]

*You are correct! The answer “*DF[:4]*” is correct because this code tells pandas to select rows from zero to 4 using slice indexing. If you leave the space before the colon blank, pandas assumes you want to start at the first row.*

What is the symbol to represent the standard deviation of a distribution? : 𝛔y

*You are correct! The answer “*𝛔y” *is correct because this is the correct symbol for the standard deviation.*

How can an index be set in creating a new dataframe? *(Check all that apply.) :* Use set\_index() to assign a column to be the index., Keep the default index., Pass an index to the constructor with the index input argument.

*You are correct! “Pass an index to the constructor with the index input argument”, “Use*set\_index()*to assign a column to be the index”, and “Keep the default index” are all correct ways to set an index when creating a new dataframe.*

What is represented by Y~Fy? : Y is distributed according to a distribution Fy

*You are correct! The answer “Y is distributed according to a distribution Fy” is correct because Y is sampled using the function Fy.*

What do you call a system with the same input which returns a different output each time it is sampled*? :* Non-deterministic system

*You are correct! The answer “Non-deterministic system*” *is correct because such a system exhibits a different behavior every time it is sampled.*

If a distribution has a large variance its graph will be wider and shorter. : True

*You are correct! The answer “True” is correct because variance is a measure of how widely distributed the values are. The larger the variance grows, the wider the graph will be.*

The function which takes the actual result y and our predicted result ŷ and scores it as the square of the difference is known as the loss function L. : True

*You are correct! The loss function (L) is calculated as the square of the difference between the actual result and the predicted result.*

Machine learning is the study of all of the different ways in which models can be built from data. : True

*You are correct! “True” is correct because machine learning encompasses all the ways models can be built to model systems.*

The mean and median of a distribution can be the same. : True

*You are correct! The answer “True” is correct because although the mean and median measure different things, they can be the same value.*

Which of the following functions would show you count, mean, std, min, max? : describe()

*You are correct! The answer “*describe()*” is correct because*describe()*is used to get the statistical data analysis of the dataframe columns, that is, the mean, median, and std.*

With the dataframe df= pd.dataframe({’A’:[0.1,0.2,0.3], ‘B’:[10,20,30], ‘C’:[1,2,3]}), how can you create a bar chart? : df.sum().plot(kind=’bar’)

*You are correct! The answer “*df.sum().plot(kind=’bar’)*” is correct because a bar chart is plotted among the sum of the columns as bar heights.*

If the output of a function can be any real number, the distribution is known as a (blank). : Continuous distribution

*You are correct! The answer “Continuous distribution“ is correct because in continuous distribution, the random variable X can have any value since there are infinite values X can take.*

With the dataframe df, to plot a line chart with columns A and B, what is the correct code? : df.plot(kind=’line’,x=’index’,y=[’A’,’B’])

*You are correct! The answer “*df.plot(kind=’line’,x=’index’,y=[’A’,’B’])*” is correct because it sets the x-axis as the index and plots the lines of column A and B on the y-axis with the correct syntax.*

The generic output of the system is represented with a capital letter Y and the samples of the output as small letter y. : True

*You are correct! The answer “True” is correct because the capital letter Y is used to denote the generic output of a system, and the samples of the output are denoted with a small letter y.*

With the dataframe df= pd.dataframe({’A’:[0.1,0.2,0.3], ‘B’:[10,20,30], ‘C’:[1,2,3]}), how would you create a histogram that includes columns B and C? : df.plot(kind = ‘hist’,y=[’B’,’C’])

*That is correct! The answer “*df.plot(kind = ‘hist’,y=[’B’,’C’])*” is correct because this is the right method to build a histogram between column B and C.*

The function tail() is used to show the last rows of the dataframe. : True

*That is correct! The answer “True” is correct because*tail()*is used to show the last rows of the dataframe.*

The more complex the task or problem becomes, (blank). *(Check all that apply.) :* ...the more complicated the model needs to be to represent it. , ...the more data is needed to train the model. , ...the more computational power is needed to process that data.

*You are correct! All of these* *are correct because all the scenarios occur if the task or problem is complex.*

Which of the following is a way to select data from a dataframe? *(Check all that apply.) :* table, table.loc[], table.iloc[]

*You are correct! The answers “*table[]*”, “*table.loc[]*”, and “*table.iloc[]”*are all correct because all of them are statements to select data from a dataframe.*

In the plot() function, what does the parameter figsize=(1,2) specify? : The width and height of the plot

*You are correct! The answer “The width and height of the plot” is correct because the*figsize*parameter has two parts to set: the width and height.*

**Fishing**

Almost any equipment or gear used for fishing can be called fishing tackle. Some examples are [hooks](https://en.wikipedia.org/wiki/Fishing_hook), [lines](https://en.wikipedia.org/wiki/Fishing_line), [sinkers](https://en.wikipedia.org/wiki/Fishing_sinker), [floats](https://en.wikipedia.org/wiki/Fishing_float), [rods](https://en.wikipedia.org/wiki/Fishing_rod), [reels](https://en.wikipedia.org/wiki/Fishing_reel), [baits](https://en.wikipedia.org/wiki/Fishing_bait), [lures](https://en.wikipedia.org/wiki/Fishing_lure), [spears](https://en.wikipedia.org/wiki/Spearfishing), [nets](https://en.wikipedia.org/wiki/Fishing_net), [gaffs](https://en.wikipedia.org/wiki/Fishing_gaff), [traps](https://en.wikipedia.org/wiki/Fishing_trap), [waders](https://en.wikipedia.org/wiki/Waders_(footwear)) and tackle boxes.

Why a fish bites a baited hook or lure involves several factors related to the sensory physiology, behaviour, feeding ecology, and biology of the fish as well as the environment and characteristics of the bait/hook/lure.[28] There is an intricate link between various fishing techniques and knowledge about the fish and their behaviour including [migration](https://en.wikipedia.org/wiki/Fish_migration), [foraging](https://en.wikipedia.org/wiki/Forage_fish) and [habitat](https://en.wikipedia.org/wiki/Oceanic_habitats). The effective use of fishing techniques often depends on this additional knowledge.[29] Some fishers follow [fishing folklores](https://en.wikipedia.org/wiki/Solunar_theory) which claim that fish feeding patterns are influenced by the position of the sun and the moon.

Fishing is another example of “black box” even though even though there are specific fishing tackles are used to target certain fish kinds like lines by weight, hook size, sinkers to adjust how close to bottom, floats to adjust how close to surface, baits and lures to target certain population. There are also other factors like time of day, environmental factor, feeding behavior, fish sensory, as well as the sun and moon positions to affect a fish bite. The end result is still not predictable.

**Datasets**

[**https://student.emeritus.org/courses/4765/discussion\_topics/257482**](https://student.emeritus.org/courses/4765/discussion_topics/257482)

* A short description of the content of the data
* One or two possible business uses of the data
* The location (URL) of the dataset
* A summary of the characteristics of the dataset using the describe() function in pandas

<https://www.kaggle.com/prasertk/internet-broadband-and-mobile-speeds-by-country>

<https://www.kaggle.com/vivovinco/san-francisco-incident-reports-2018present>

**Context**

The Nobel Prize is a set of annual international awards bestowed in several categories by Swedish and Norwegian institutions in recognition of academic, cultural, or scientific advances. The will of the Swedish chemist, engineer, and industrialist Alfred Nobel established the five Nobel prizes in 1895. The prizes in Chemistry, Literature, Peace, Physics, and Physiology or Medicine were first awarded in 1901. The prizes are widely regarded as the most prestigious awards available in their respective fields.

Between 1901 and 2020, the Nobel Prizes and the Prize in Economic Sciences were awarded to 949 people . The Nobel Prize is an international award administered by the Nobel Foundation in Stockholm, Sweden, and based on the fortune of Alfred Nobel, Swedish inventor and entrepreneur. In 1968, Sveriges Riksbank established The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel, founder of the Nobel Prize. Each Prize consists of a medal, a personal diploma, and a cash award.

A person or organization awarded the Nobel Prize is called Nobel Laureate. The word "laureate" refers to being signified by the laurel wreath. In ancient Greece, laurel wreaths were awarded to victors as a sign of honor.

**Dataset Description**

This dataset includes a record for every individual that was awarded the Nobel Prize between 1901 and 2020.

**Business Use Case**

The analysis can be used to identify which country invest in science, subject, and/or what institutions are more credible.

**URL**

<https://www.kaggle.com/bahramjannesarr/nobel-prize-from-1901-till-2020>

**Dataset Characteristic**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | year | share | age | age\_get\_prize |
| count | 923.000000 | 923.000000 | 923.000000 | 923.000000 |
| mean | 1971.102925 | 2.020585 | 79.346696 | 59.823402 |
| std | 33.606425 | 0.943415 | 11.134606 | 12.571109 |
| min | 1901.000000 | 1.000000 | 23.000000 | 17.000000 |
| 25% | 1947.000000 | 1.000000 | 73.000000 | 51.000000 |
| 50% | 1977.000000 | 2.000000 | 80.000000 | 60.000000 |
| 75% | 2000.000000 | 3.000000 | 87.500000 | 69.000000 |
| max | 2019.000000 | 4.000000 | 103.000000 | 97.000000 |

**Content**

+120.000 rows and 10 columns. Columns' description are listed below.

* DateTime : Datetime in "dd.mm.yyyy hh:mm" format
* Temperature : Temperature at 2 m in °C
* Sunshine Duration : Sunshine duration in min
* Shortwave Radiation : Shortwave radiation in W/m²
* Relative Humidity : Relative Humidity at 2 m in %
* Mean Sea Level Pressure : Mean Sea Level Pressure (MSL) in hPa
* Soil Temperature : Soil temperature at 0-10 cm down in °C
* Soil Moisture : Soil moisture at 0-10 cm down in m³/m³
* Wind Speed : Wind speed at 10 m in km/h
* Wind Direction : Wind direction at 10 m in degrees

**Acknowledgements**

Data from [Meteoblue](https://www.meteoblue.com/). Image from [Anadolu Agency](https://www.aa.com.tr/en/asia-pacific/turkeys-gallipoli-anniversary-marked-in-pakistan/1089591).

If you're reading this, please upvote.

**Dataset Description**

This dataset contains hourly weather data between 2008 and 2021 for Gallipoli, Turkey.

**Business Use Case**

Throughout day/week/month/year wind direction and wind speed can be analyzed for possible wind power installations.

This information is also useful for construction business to decide regarding building strength against wind.

**URL**

<https://www.kaggle.com/vivovinco/hourly-weather-data-in-gallipoli-20082021>

**Dataset Characteristic**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Temperature | Sunshine Duration | Shortwave Radiation | Relative Humidity | Mean Sea Level Pressure | Soil Temperature | Soil Moisture | Wind Speed | Wind Direction |  |
| count | 122733.000000 | 122733.000000 | 122733.000000 | 122733.000000 | 122733.000000 | 122733.000000 | 122733.000000 | 122733.000000 | 122733.000000 |
| mean | 15.670196 | 21.084680 | 202.333411 | 69.683924 | 1015.073141 | 16.336311 | 0.210705 | 19.391254 | 108.071393 |
| std | 8.192027 | 27.030718 | 272.710423 | 19.079325 | 6.795215 | 8.054375 | 0.067094 | 11.141315 | 91.270893 |
| min | -13.800000 | 0.000000 | 0.000000 | 13.000000 | 981.300000 | -5.200000 | 0.102000 | 0.000000 | 0.310000 |
| 25% | 9.700000 | 0.000000 | 0.000000 | 55.000000 | 1010.600000 | 10.100000 | 0.153000 | 10.400000 | 44.090000 |
| 50% | 15.500000 | 0.000000 | 16.020000 | 72.000000 | 1014.400000 | 15.900000 | 0.214000 | 18.300000 | 60.360000 |
| 75% | 21.400000 | 56.000000 | 379.140000 | 86.000000 | 1019.200000 | 22.400000 | 0.263000 | 26.800000 | 195.020000 |
| max | 40.800000 | 60.000000 | 950.520000 | 100.000000 | 1043.600000 | 40.100000 | 0.433000 | 93.300000 | 360.000000 |

<https://www.kaggle.com/andrewmvd/citymapper-mobility-index>

<https://www.kaggle.com/susant4learning/crime-in-los-angeles-data-from-2020-to-present>

**Context**

This Data set contains various information about a set of cars that were manufactured with set of factory parameters like cylinder size, number of cylinders, fuel consumption, Carbon dioxide emissions etc.

**Content**

Feature set includes model, make, vehicle type, engine size,Transmission, fuel consumption etc. The target set is Carbon dioxide Emission.

**Inspiration**

The inspiration was to learn and help understand the concept behind data science's basics.

**Dataset Description**

This dataset contains make, model, vehicle type, engine size, and carbon dioxide emission.

**Business Use Case**

This carbon footprint data can be used for consumer awareness about what car companies, make, model are more environmentally friendly.

**URL**

<https://www.kaggle.com/prathamtripathi/co2-emissions-by-cars-in-canada>

**Dataset Characteristic**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MODEL | ENGINE\_SIZE | CYLINDERS | FUEL\_CONSUMPTION\* | Unnamed: 9 | Unnamed: 10 | Unnamed: 11 | CO2\_EMISSIONS |
| count | 679.0 | 679.000000 | 679.000000 | 679.000000 | 679.000000 | 679.000000 | 679.000000 | 679.000000 |
| mean | 2001.0 | 3.252577 | 5.798233 | 14.591900 | 10.613844 | 12.802798 | 23.107511 | 293.656848 |
| std | 0.0 | 1.203751 | 1.531073 | 3.025654 | 2.357724 | 2.685590 | 5.308083 | 60.372456 |
| min | 2001.0 | 1.000000 | 3.000000 | 4.900000 | 4.000000 | 4.500000 | 14.000000 | 104.000000 |
| 25% | 2001.0 | 2.200000 | 4.000000 | 12.700000 | 9.000000 | 11.000000 | 19.000000 | 253.000000 |
| 50% | 2001.0 | 3.000000 | 6.000000 | 14.300000 | 10.100000 | 12.300000 | 23.000000 | 283.000000 |
| 75% | 2001.0 | 4.200000 | 6.000000 | 16.650000 | 12.500000 | 14.850000 | 26.000000 | 340.000000 |
| max | 2001.0 | 8.000000 | 12.000000 | 23.200000 | 18.000000 | 20.800000 | 63.000000 | 478.000000 |

* A short description of the content of the data
* One or two possible business uses of the data
* The location (URL) of the dataset
* A summary of the characteristics of the dataset using the describe() function in pandas

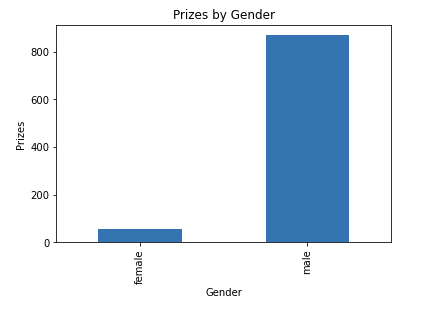
Using personally-sourced data (see Discussion 1.2), create a visualization using the pandas plot() function (scatter, bar chart, line) and post your most interesting results. For the visualization that you selected, please describe why you chose that plot type and any transformations to the data that you had to make in order to generate your visualization. Additionally, please describe any interesting trends (i.e., increased monthly sales) that you observe in your results.

**Noble Prizes**

I analyzed Nobel Prize Winners dataset by gender, category, age, institutions and number of institutions in country by using kaggle dataset: <https://www.kaggle.com/bahramjannesarr/nobel-prize-from-1901-till-2020>.

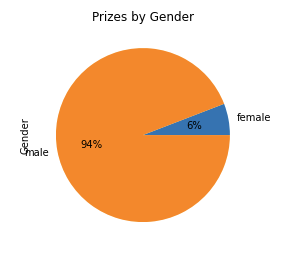
Data reveals there is a huge gap between genders, majority of Noble Prize winners are male by using bar chart:

data.groupby('gender').size().plot(kind='bar',title='Prizes by Gender', xlabel='Gender', ylabel='Prizes')



It is an uneven distribution among gender, only 6% is female by using pie chart for better visibility:

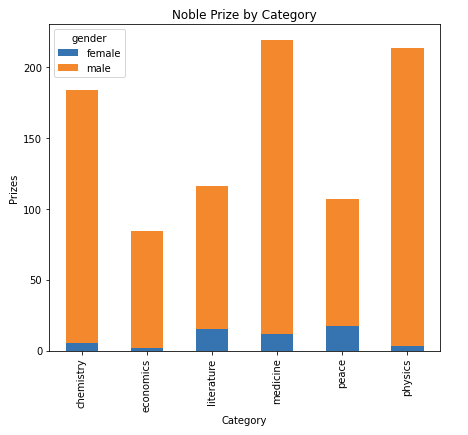
data.groupby('gender').size().plot(kind='pie', title='Prizes by Gender', ylabel='Gender', autopct='%1.0f%%')



Further break gender down by category reveals that females participate more in Peace, Medicine and Literature categories and less in Chemistry, Economics and Physics categories. I chose the stacked bar representation for comparison:

cnt=data.groupby(['category','gender']).size().unstack()

cnt.plot(kind='bar',stacked=True, title='Noble Prize by Category', xlabel='Category', ylabel='Prizes', figsize=(7,6))



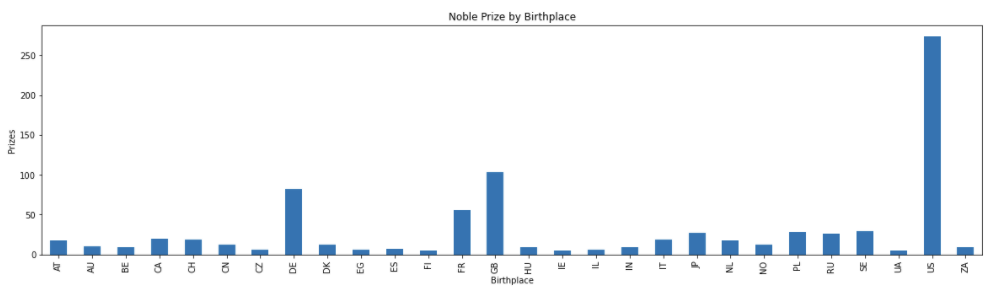
Also, there are more participants in Medicine, Physics and Chemistry as they tend to be team effort than the rest.

Birthplace of Noble Prize winners concentrate only in a few countries where majority in US followed by United Kingdom, Germany and France:

bornset=data.groupby('born\_country\_code', as\_index=False).size()

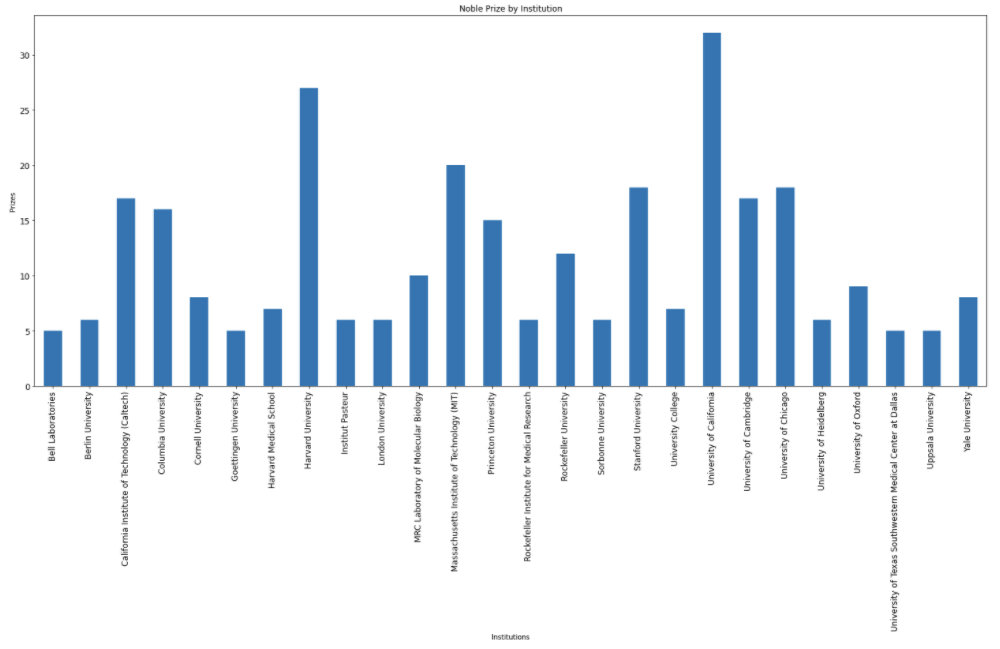
bornset=bornset[bornset['size'] > 4]

bornset.plot(kind='bar', figsize=(20,5), x='born\_country\_code', title='Noble Prize by Birthplace', xlabel='Birthplace', ylabel='Prizes', legend=False)



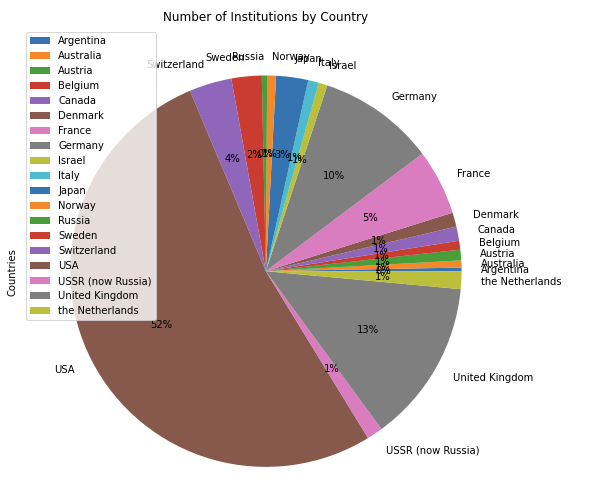
As well as top institutions hosted noble prize winners, University of California is at the top:

subset[subset['size'] > 4].plot(kind='bar', figsize=(25,10), x='name\_of\_university', title='Noble Prize by Institution', xlabel='Institutions', ylabel='Prizes', fontsize=12, legend=False)



These information yield where those institutions concentrated. More than half of prize winners are located in US followed by United Kingdom, Germany and France just like their birthplace which points those brilliant folks do not migrate:

subset.plot(kind='pie', title='Number of Institutions by Country', x='country\_of\_university', y='size', ylabel='Countries', autopct='%1.0f%%', figsize=(9, 9), fontsize=10, rot=180)



Final look at the data is age breakdown by prize, not surprisingly peak at between 54 and 64:

data.groupby('age\_get\_prize').size().plot(kind='bar', figsize=(13,5), title='Noble Prize versus Age', xlabel='Age', ylabel='Prizes'

