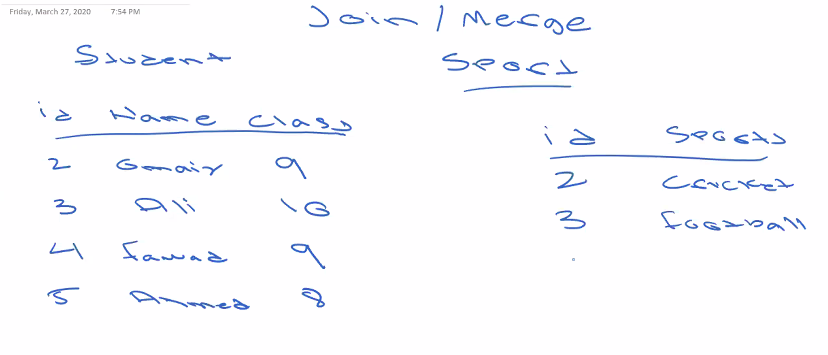
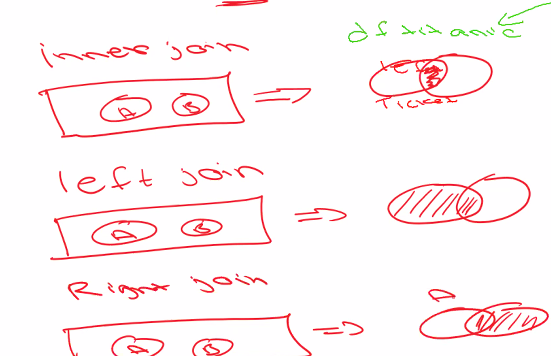
* String is not mutable
* Starting value is included in slicing and ending value is not included or excluded in the slicing.
* Immutable is used for something that requires that it cannot be changed like security, history notebooks.
* NO duplication in set and list can be made in ascending order
* By default string
* Break comes out of closest loop
* Continue goes to beginning of closest loop
* Pandas is used to manipulate large datasets
* Pandas is extensively used for data analysis
* Matplotlib is used for data visualization
* Scikit lib is used for machine learning
* Telcos use Sas
* Feature engineering using and creating new data
* 64 is the bit
* NaN not a number is null
* 30% is usually deleted
* 70% is used by using median or mean
* Body style explanatory variable
* Price response variable
* Continuous= scatterplot
* Discrete=histogram
* Skewness is due to outliers
* Boxplots will remove the outliers
* Line plot is always used for time dependent
* We check mode of 1964 and 2011 graphs
* Median or mean both may work(for exercise)
* Life exp and year is a dependent relation
* Seaborn is better for visualization since it can commute code in less number of lines
* All data descriptions must be known
* Feature engineering using these variables to create more variables
* No survival column in dfcomp and we learning about the survival 1200 passenger divided b/w titanic and comp.800 and 400.
* We are using catplot so we have to tell which aggregation we are using
* Starts with space then search a character that starts with a capital letter and ends with a small letter.
* + is for iteration keep going till the dot comes since Master and Mr count different no of words.(Some are more than two).
* Group(1) means group with expression
* Title is the Drive variable
* For story telling you should know each Feature inside and out.
* Normalize tells percent without it count will be shown
* We passed the name in reset\_index the name of new column in this case perc
* Dodge parameter for separate showing else it will show both together
* One ticket was assigned for multiple people such as group of friends,family,etc
* Join or merge
* 
* We join two tables by applying join conditions on both table where there is a common element.
* In this case we use id to join student and sports data bases which are in separate tables.
* 
* Index is by default
* The table we are using is automatically left
* We will use inner join
* Inner join picks up those which are common in both A and B
* Left join will use dominating table and common of both A and B
* Right join is opposite of left join it uses right dominating table and common of both A and B.
* We cannot create a new column directly we have to use temp since we do not know how many rows contain the ticket and the size of df titanic and dfcomp becomes different.
* 0 is not a null value
* Subplot is of distribution plot
* 
* Outer will be bring all the information.
* In this case inner left right will all produce same result
* We replace 1,2 with categorical values
* 
* We replace homogenous values with mean.
* When data is right skewed mean does not respect true center median does that is why we choose median.
* We have created banding
* Fare improves survival improves
* We wont explore embarkment further since it contains no business knowledge
* Title gives a good estimation of age since Master,Mister etc can be used to discern age
* We look at median values in box plots
* 1 person only so median is only 1 so it cannot be taken for one person so we use median for entire p class.
* We do this median value stuff for 30% less values
* We replace with mode Southampton in embarked since it is categorical variable
* We need to drop cabin since it contains 77% missing values.
* Fmt tells how till how many points
* We need to choose the relation with best relation with target variable and minimum relation to the other variables.
* We need those variables in machine learning which gives the most accurate results and are distinct.
* We can use algorithms for feature selection as well.
* We never give two same features because model can become biased.