## ===START OF FILE: C:\Users\ambyb\Downloads\BA\_W3\_Introduction to Data Analysis.pdf===

Bigdata and Artificial Intelligence  
Introduction to Data Analysis  
Bigdata and Artificial Intelligence  
Introduction to Data Analysis  
TOC  
  
• Fundamentals of data analysis  
• Descriptive statistics  
• Summarizing data  
ㅁ  
• Data Relationships  
• Pitfalls of summary statistics  
• Prediction and Forecasting  
• Inferential Statistics

---Page 3---

Bigdata and Artificial Intelligence  
Fundamentals of data analysis  
▪ Data Collection  
• Every analysis starts with collecting data. We can collect data from a variety of sources, including  
databases, APIs, flat files, and the Internet.  
▪ Data Wrangling  
• After we have our data, we need to prepare it for our analysis. This may involve reshaping it,  
changing data types, handling missing values, and/or aggregating it.  
▪ Exploratory Data Analysis (EDA)  
• We can use visualizations to explore our data and summarize it. During this time, we will also  
begin exploring the data by looking at its structure, format, and summary statistics.  
▪ Drawing Conclusions  
• After we have thoroughly explored our data, we can try to draw conclusions or model it.  
9

---Page 4---

Bigdata and Artificial Intelligence | Introduction to statistics  
Population vs. Sample  
Population Sample  
Random  
Observation  
Selection  
①  
②  
Inference  
Parameters Statistics  
Ex) Population Mean (𝜇),  
Ex) Sample Mean (𝑥ҧ),  
③  
Population Variance (𝜎2) Sample Variance (𝑠2)  
10

---Page 5---

Bigdata and Artificial Intelligence  
Sampling  
▪ Simple random sampling  
• Pick with a random number generator  
▪ Stratified random sampling  
• Randomly pick preserving the proportion of groups in the data  
▪ Bootstrapping  
• Sampling with replacement (more info: YouTube video and Wikipedia article))  
11

---Page 6---

Bigdata and Artificial Intelligence  
Introduction to Data Analysis  
TOC  
  
• Fundamentals of data analysis  
• Descriptive statistics  
• Summarizing data  
ㅁ  
• Data Relationships  
• Pitfalls of summary statistics  
• Prediction and Forecasting  
• Inferential Statistics

---Page 7---

Bigdata and Artificial Intelligence  
Descriptive Statistics  
▪ We use descriptive statistics to describe the data.  
▪ The data we work with is usually a sample taken from the population.  
▪ The statistics we will discuss here are referred to as sample  
statistics because they are calculated on the sample and can be  
used as estimators for the population parameters.  
13

---Page 8---

Bigdata and Artificial Intelligence  
Measures of Center  
▪ Mean  
• The sample mean is an estimator for the population mean (𝜇) and is defined as:  
▪ Median  
• The median represents the 50th percentile of our data; this means that 50% of  
the values are greater than the median and 50% are less than the median.  
• It is calculated by taking the middle value from an ordered list of values.  
14

---Page 9---

Bigdata and Artificial Intelligence  
Measures of Center  
▪ Mode  
• The mode is the most common value in the data. We can use it to describe  
categorical data or, for continuous data, the shape of the distribution:  
15

---Page 10---

Bigdata and Artificial Intelligence  
Measures of Spread  
▪ Measures of spread tell us how the data is dispersed; this will  
indicate how thin (low dispersion) or wide (very spread out) our  
distribution is.  
▪ Range  
• The range is the distance between the smallest value (minimum) and the largest  
value (maximum):  
16

---Page 11---

Bigdata and Artificial Intelligence  
Measures of Spread  
▪ Variance  
• The variance describes how far apart observations are  
spread out from their average value (the mean).  
• When calculating the sample variance, we divide by 𝑛 − 1  
instead of n to account for using the sample mean (𝑥ҧ):  
• This is referred to as Bessel's correction and is applied to  
get an unbiased estimator of the population variance.  
▪ Standard Deviation  
• The standard deviation is the square root of the variance,  
giving us a measure in the same units as our data.  
• The sample standard deviation is calculated as follows:  
17

---Page 12---

Bigdata and Artificial Intelligence  
Measures of Spread  
2  
Note that 𝜎 is the population variance and 𝜎 is the population standard deviation.  
18

---Page 13---

Bigdata and Artificial Intelligence  
Measures of Spread  
▪ Coefficient of Variation  
• The coefficient of variation (CV) gives us a unitless  
ratio of the standard deviation to the mean. Since, it  
has no units we can compare dispersion across  
datasets:  
▪ Interquartile Range  
• The interquartile range (IQR) gives us the spread of  
data around the median and quantifies how much  
dispersion we have in the middle 50% of our  
distribution:  
19

---Page 14---

Bigdata and Artificial Intelligence  
Measures of Spread  
▪ Quartile Coefficient of Dispersion  
• The quartile coefficient of dispersion also is a unitless statistic for comparing  
datasets.  
• However, it uses the median as the measure of center.  
• It is calculated by dividing the semi-quartile range (half the IQR) by the midhinge  
(midpoint between the first and third quartiles):  
20

---Page 15---

Bigdata and Artificial Intelligence  
Introduction to Data Analysis  
TOC  
  
• Fundamentals of data analysis  
• Descriptive statistics  
• Summarizing data  
ㅁ  
• Data Relationships  
• Pitfalls of summary statistics  
• Prediction and Forecasting  
• Inferential Statistics

---Page 16---

Bigdata and Artificial Intelligence  
Summarizing data  
▪ The 5-number summary provides 5 descriptive statistics that  
summarize our data:  
22

---Page 17---

Bigdata and Artificial Intelligence  
Box Plot  
▪ This summary can be visualized using a box  
plot (also called box-and-whisker plot).  
• The box has an upper bound of 𝑄 and a lower bound  
3  
of 𝑄 .  
1  
• The median will be a line somewhere in this box.  
• The whiskers extend from the box towards the  
minimum/maximum.  
• For our purposes, they will extend to 𝑄 + 1.5 × 𝐼𝑄𝑅  
3  
and 𝑄 − 1.5 × 𝐼𝑄𝑅 and anything beyond will be  
1  
represented as individual points for outliers.  
23

---Page 18---

Bigdata and Artificial Intelligence  
Histogram  
▪ The box plot doesn't show us how the data is distributed within the quartiles.  
▪ To get a better sense of the distribution, we can use a histogram, which will  
show us the amount of observations that fall into equal-width bins.  
▪ We can vary the number of bins to use, but be aware that this can change our  
impression of what the distribution appears to be:  
24

---Page 19---

Bigdata and Artificial Intelligence  
Kernel Density Estimate  
▪ We can also visualize the distribution using a kernel density estimate (KDE).  
▪ This will estimate the probability density function (PDF).  
▪ This function shows how probability is distributed over the values.  
▪ Higher values of the PDF mean higher likelihoods:  
25

---Page 20---

Bigdata and Artificial Intelligence  
Kernel Density Estimate  
▪ Note that both the KDE and histogram estimate the distribution:  
26

---Page 21---

Bigdata and Artificial Intelligence  
Skewed Distributions  
▪ Skewed distributions have more observations on one side.  
▪ The mean will be less than the median with negative skew, while the opposite is  
true of positive skew:  
27

---Page 22---

Bigdata and Artificial Intelligence  
Cumulative Distribution Function  
▪ We can use the cumulative distribution function (CDF) to find probabilities of getting  
values within a certain range.  
▪ The CDF is the integral of the PDF:  
∞  
▪ Note that 𝑓(𝑡) is the PDF and ׬ 𝑓 𝑡 𝑑𝑡 = 1  
−∞  
▪ The probability of the random variable 𝑋 being less than or equal to the specific value  
of 𝑥 is denoted as 𝑃 𝑋 ≤ 𝑥 .  
▪ Note that for a continuous random variable the probability of it being exactly 𝑥 is zero.  
28

---Page 23---

Bigdata and Artificial Intelligence  
Cumulative Distribution Function  
▪ Let's look at the estimate of the CDF from the sample data we used for the box  
plot, called the empirical cumulative distribution function (ECDF):  
We can find any range we want if we use some algebra as in the rightmost subplot above.  
29

---Page 24---

Bigdata and Artificial Intelligence  
Introduction to Data Analysis  
TOC  
  
• Fundamentals of data analysis  
• Descriptive statistics  
• Summarizing data  
ㅁ  
• Data Relationships  
• Pitfalls of summary statistics  
• Prediction and Forecasting  
• Inferential Statistics

---Page 25---

Bigdata and Artificial Intelligence  
Scaling Data  
▪ In order to compare variables from different distributions, we would have to  
scale the data, which we could do with the range by using min-max scaling:  
▪ Another way is to use a Z-score to standardize the data:  
31

---Page 26---

Bigdata and Artificial Intelligence  
Quantifying Relationships between Variables  
▪ The covariance is a statistic for quantifying the relationship between variables  
by showing how one variable changes with respect to another (also referred to  
as their joint variance):  
• E[X] is the expectation of the random variable X (its long-run average).  
▪ The sign of the covariance gives us the direction of the relationship, but we  
need the magnitude as well.  
▪ For that, we calculate the Pearson correlation coefficient (𝜌):  
32

---Page 27---

Bigdata and Artificial Intelligence  
Quantifying Relationships between Variables  
From left to right: no correlation, weak negative correlation, strong positive correlation,  
and nearly perfect negative correlation.  
33

---Page 28---

Bigdata and Artificial Intelligence  
Quantifying Relationships between Variables  
▪ Often, it is more informative to use scatter plots to check for relationships  
between variables.  
▪ This is because the correlation may be strong, but the relationship may not be  
linear:  
34

---Page 29---

Bigdata and Artificial Intelligence  
Quantifying Relationships between Variables  
▪ Remember, correlation does not imply causation.  
▪ While we may find a correlation between X and Y, it does not mean that X  
causes Y or Y causes X.  
▪ It is possible there is some Z that causes both or that X causes some  
intermediary event that causes Y — it could even be a coincidence.  
35

---Page 30---

Bigdata and Artificial Intelligence  
Introduction to Data Analysis  
TOC  
  
• Fundamentals of data analysis  
• Descriptive statistics  
• Summarizing data  
ㅁ  
• Data Relationships  
• Pitfalls of summary statistics  
• Prediction and Forecasting  
• Inferential Statistics

---Page 31---

Bigdata and Artificial Intelligence  
Pitfalls of summary statistics  
▪ Not only can our correlation coefficients be misleading, but so can  
summary statistics.  
▪ Anscombe's quartet is a collection of four different datasets that  
have identical summary statistics and correlation coefficients.  
▪ However, when plotted, it is obvious they are not similar.  
37

---Page 32---

Bigdata and Artificial Intelligence  
Pitfalls of summary statistics  
▪ Not only can our correlation coefficients be misleading, but so can summary  
statistics.  
▪ Anscombe's quartet is a collection of four different datasets that have identical  
summary statistics and correlation coefficients, however, when plotted, it is  
obvious they are not similar:  
38

---Page 33---

Bigdata and Artificial Intelligence  
Pitfalls of summary statistics  
▪ Another example of this is the Datasaurus Dozen:  
39

---Page 34---

Bigdata and Artificial Intelligence  
Introduction to Data Analysis  
TOC  
  
• Fundamentals of data analysis  
• Descriptive statistics  
• Summarizing data  
ㅁ  
• Data Relationships  
• Pitfalls of summary statistics  
• Prediction and Forecasting  
• Inferential Statistics

---Page 35---

Bigdata and Artificial Intelligence  
Prediction  
▪ Say our favorite ice cream shop has asked us to help predict how  
many ice creams they can expect to sell on a given day.  
▪ They are convinced that the temperature outside has strong  
influence on their sales, so they collected data on the number of ice  
creams sold at a given temperature.  
41

---Page 36---

Bigdata and Artificial Intelligence  
Prediction  
▪ We agree to help them, and the first thing we do is make a scatter plot of the data they gave us:  
▪ We can observe an upward trend in the scatter plot: more ice creams are sold at higher temperatures.  
42

---Page 37---

Bigdata and Artificial Intelligence  
Prediction  
▪ In order to help out the ice cream shop, though, we need to find a way to make predictions from  
this data.  
▪ We can use a technique called regression to model the relationship between temperature and  
ice cream sales with an equation:  
43

---Page 38---

Bigdata and Artificial Intelligence  
Prediction  
▪ We can use the resulting equation to make predictions for the number of ice  
creams sold at various temperatures.  
▪ However, we must keep in mind if we are interpolating or extrapolating.  
▪ If the temperature value we are using for prediction is within the range of the  
original data we used to build our regression model, then we are interpolating  
(solid portion of the red line).  
▪ On the other hand, if the temperature is beyond the values in the original data,  
we are extrapolating, which is very dangerous, since we can't assume the  
pattern continues indefinitely in each direction (dotted portion of the line).  
• Extremely hot temperatures may cause people to stay inside, meaning no ice creams will be sold,  
while the equation indicates record-high sales.  
44

---Page 39---

Bigdata and Artificial Intelligence  
Forecasting  
▪ Forecasting is a type of prediction for time series.  
• In a process called time series decomposition, time series is decomposed into a trend component,  
a seasonality component, and a cyclical component.  
45

---Page 40---

Bigdata and Artificial Intelligence  
Trend  
▪ The trend component describes the behavior of the time series in the long-term  
without accounting for the seasonal or cyclical effects.  
▪ Using the trend, we can make broad statements about the time series in the  
long-run, such as: the population of Earth is increasing or the value of a stock is  
stagnating.  
46

---Page 41---

Bigdata and Artificial Intelligence  
Seasonality  
▪ Seasonality of a time series explains the systematic and calendar-related  
movements of a time series.  
▪ For example, the number of ice cream trucks on the streets of New York City is  
high in the summer and drops to nothing in the winter; this pattern repeats  
every year regardless of whether the actual amount each summer is the same.  
47

---Page 42---

Bigdata and Artificial Intelligence  
Cyclical  
▪ Lastly, the cyclical component accounts for anything else unexplained or  
irregular with the time series; this could be something like a hurricane driving  
the number of ice cream trucks down in the short-term because it isn't safe to  
be outside.  
▪ This component is difficult to anticipate with a forecast due to its unexpected  
nature.  
48

---Page 43---

Bigdata and Artificial Intelligence  
Introduction to Data Analysis  
TOC  
  
• Fundamentals of data analysis  
• Descriptive statistics  
• Summarizing data  
ㅁ  
• Data Relationships  
• Pitfalls of summary statistics  
• Prediction and Forecasting  
• Inferential Statistics

---Page 44---

Bigdata and Artificial Intelligence  
Inferential Statistics  
▪ Inferential statistics deals with inferring or deducing things from the  
sample data we have in order to make statements about the  
population as a whole.  
▪ Before doing so, we need to know whether we conducted an  
observational study or an experiment.  
• An observational study can't be used to determine causation because we can't  
control for everything.  
• An experiment on the other hand is controlled.  
50

---Page 45---

Bigdata and Artificial Intelligence | Introduction to statistics  
Statistical Inference  
Population Sample  
Random  
Observation  
Selection  
①  
②  
Inference  
Parameters Statistics  
Ex) Population Mean (𝜇),  
Ex) Sample Mean (𝑥ҧ),  
③  
Population Variance (𝜎2) Sample Variance (𝑠2)  
51

---Page 46---

Bigdata and Artificial Intelligence  
Hypothesis testing  
▪ We also have the option of using hypothesis testing.  
▪ First, we define a null hypothesis (say the true population mean is 0), then we  
determine a significance level (1 - confidence level), which is the probability of  
rejecting the null hypothesis when it is true.  
▪ Our result is statistically significant if the value for the null hypothesis is outside  
the confidence interval.  
52

---Page 47---

Bigdata and Artificial Intelligence  
Introduction to Data Analysis  
  
Q&A  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr

## ===START OF FILE: C:\Users\ambyb\Downloads\BA\_W4\_pandas.pdf===

---Page 1---

Bigdata and Artificial Intelligence  
pandas  
  
Hanyang University  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr

---Page 2---

Bigdata and Artificial Intelligence  
pandas  
TOC  
  
• Introduction to pandas  
• Data Types  
ㅁ  
• Data Exploration  
• Data Selection  
• Data Manipulation  
• Exercises

---Page 3---

Bigdata and Artificial Intelligence  
Pandas  
▪ Pandas is an open-source Python library providing high-performance, easy-to-  
use data structures and data analysis tools.  
▪ Designed for working with structured data (e.g., CSV, Excel files) and time  
series data.  
▪ Essential for data manipulation, preprocessing, and exploratory data analysis in  
Python.  
3

---Page 4---

Bigdata and Artificial Intelligence  
Key Features of Pandas  
▪ DataFrame and Series objects for efficient data manipulation.  
▪ Comprehensive tools for reading and writing data between in-memory data  
structures and different file formats.  
▪ Powerful data cleaning and preparation capabilities.  
▪ Flexible reshaping and pivoting of datasets.  
▪ Time series-specific functionality: date range generation and frequency  
conversion, moving window statistics, etc.  
4

---Page 5---

Bigdata and Artificial Intelligence  
Import Packages  
5

---Page 6---

Bigdata and Artificial Intelligence  
Learning Material  
▪ Colab  
• https://colab.research.google.com/drive/1XLYIYRsrHCSNdlRn8oHcVnwmcRxJ1g2P#scroll  
To=SNV72ZSviyVZ&forceEdit=true&sandboxMode=true  
▪ Originally from  
• Complete-guide-to-data-analysis-using-Python---IMDB-movies-data  
6

---Page 7---

Bigdata and Artificial Intelligence  
pandas  
TOC  
  
• Introduction to pandas  
• Data Types  
ㅁ  
• Data Exploration  
• Data Selection  
• Data Manipulation  
• Exercises

---Page 8---

Bigdata and Artificial Intelligence  
Data Types: Series  
▪ A one-dimensional labeled array capable of holding data of any type.  
▪ A Series can be created from a list of values, arrays and dictionaries.  
8

---Page 9---

Bigdata and Artificial Intelligence  
Data Types: Create series from lists  
▪ The index values are generated by default, but we can also define  
custom indexes at the time of creating series.  
9

---Page 10---

Bigdata and Artificial Intelligence  
Data Types: Row-indexes of series  
▪ Below is a Series of marks and the associated subjects can be defined as  
custom row indexes.  
10

---Page 11---

Bigdata and Artificial Intelligence  
Data Types: Name of series values and index  
▪ Both the series values and index can have name of its own and its  
defined as follows:  
11

---Page 12---

Bigdata and Artificial Intelligence  
Data Types: Subsetting the Series  
▪ Values inside the series can be  
accessed using the [ ] square  
bracket slicing operation.  
• Series can be sliced using the  
default index.  
• Series can also be sliced using  
string indices.  
12

---Page 13---

Bigdata and Artificial Intelligence  
Data Types: Creating Series from Dictionary  
▪ A Dictionary is a data structure which maps keys to a set of values.  
▪ A dictionary contains key-value pairs.  
▪ A Series is similar to a dictionary in a way that it maps given indexes  
to a set of values.  
13

---Page 14---

Bigdata and Artificial Intelligence  
Data Types: Creating Series from Dictionary  
▪ I have a dictionary that represents fruits and prices. Let's create a  
series from this dictionary.  
14

---Page 15---

Bigdata and Artificial Intelligence  
Data Types: Creating Series from Dictionary  
▪ (Continued)  
15

---Page 16---

Bigdata and Artificial Intelligence  
Data Types: DataFrame  
▪ A two-dimensional, size-mutable, potentially heterogeneous tabular  
data structure with labeled axes (rows and columns).  
16

---Page 17---

Bigdata and Artificial Intelligence  
Data Types: Create DataFrame from Series object  
▪ Let's create a DataFrame from the Marks series.  
17

---Page 18---

Bigdata and Artificial Intelligence  
Data Types: Create DataFrame from Dictionary  
▪ Let's say we have 2 series of heights and weights of the same set of persons  
and we want to put it together in a table.  
▪ Here we can create dictionary using  
both series and create a DataFrame  
using DataFrame method.  
18

---Page 19---

Bigdata and Artificial Intelligence  
Data Types: Create DataFrame by importing data from file  
▪ Pandas is extremely useful and comes handy when want to load data  
from various file formats like csv, excel, json etc.  
▪ For this analysis we will load data from a csv file.  
▪ We will use 'IMDB-Movie-Data' opensource data to learn more about  
pandas.  
• This dataset contains 7787 rows of data in 12 columns.  
• It contains data about movies and shows on netflix.  
19

---Page 20---

Bigdata and Artificial Intelligence  
Data Types: Create DataFrame by importing data from file  
20

---Page 21---

Bigdata and Artificial Intelligence  
Data Types: Create DataFrame by importing data from file  
▪ There are many options available for read\_csv().  
21

---Page 22---

Bigdata and Artificial Intelligence  
pandas  
TOC  
  
• Introduction to pandas  
• Data Types  
ㅁ  
• Data Exploration  
• Data Selection  
• Data Manipulation  
• Exercises

---Page 23---

Bigdata and Artificial Intelligence  
Data Exploration : info()  
▪ As you can see, the info( ) method gives all the details about this dataframe like  
columns, number of observations and the datatype of these columns.  
23

---Page 24---

Bigdata and Artificial Intelligence  
Data Exploration : head(), tail()  
▪Let's do a quick preview of the data by using head( ) and tail( )  
methods.  
• head( ) returns the top 5 rows in the dataset by default and can  
also take the number of rows to be viewed as a parameter.  
• tail( ) returns the bottom 5 rows in the dataset by default and can  
also take the number of rows as an optional parameter.  
24

---Page 25---

Bigdata and Artificial Intelligence  
Data Exploration : head()  
25

---Page 26---

Bigdata and Artificial Intelligence  
Data Exploration : tail()  
26

---Page 27---

Bigdata and Artificial Intelligence  
Data Exploration : columns and shapes  
▪ The column names in the data can be viewed by using columns.  
▪ The shape of the dataset can be viewed by using shape.  
• This function tells us that there are 1000 rows and 12 columns in the dataset.  
27

---Page 28---

Bigdata and Artificial Intelligence  
Data Exploration : describe()  
▪ Let's use describe( ) method to understand the numerical attributes  
in the data.  
• describe( ) shows the basic statistical summaries of numerical attributes in  
the data.  
28

---Page 29---

Bigdata and Artificial Intelligence  
Data Exploration : describe()  
▪ Some Insights from description of data.  
• The min and max values in 'Year' depict the minimum and maximum release  
years. We can see that the dataset contains movies from 2006 to 2016.  
• The average rating for the movies in this dataset is about 6.7 and the  
mininum rating is 1.9 and the maximum rating is 9.0.  
• The maximum revenue earned by a movie is 936.6 millions.  
29

---Page 30---

Bigdata and Artificial Intelligence  
Data Exploration : isnull()  
▪ isnull() function shows how many missing values are in the dataset  
with respect to different columns.  
• Here 'Revenue' and 'Metascore' columns has null values.  
30

---Page 31---

Bigdata and Artificial Intelligence  
Data Exploration : unique()  
31

---Page 32---

Bigdata and Artificial Intelligence  
Data Exploration : value\_counts()  
32

---Page 33---

Bigdata and Artificial Intelligence  
pandas  
TOC  
  
• Introduction to pandas  
• Data Types  
ㅁ  
• Data Exploration  
• Data Selection  
• Data Manipulation  
• Exercises

---Page 34---

Bigdata and Artificial Intelligence  
Data Selection: Extracting data by columns  
▪ Extracting data from Pandas column is similar to Series. The column  
label is used to extract data from the column.  
• This will return the 'genre' column as a series.  
34

---Page 35---

Bigdata and Artificial Intelligence  
Data Selection: Extracting data by columns  
▪ If we want to retrieve it as a dataframe, then indexing must be done  
using double square brackets.  
35

---Page 36---

Bigdata and Artificial Intelligence  
Data Selection: Extracting data by columns  
▪ If we want to extract multiple columns from the data, simply add the  
column names to the list.  
36

---Page 37---

Bigdata and Artificial Intelligence  
Data Selection: Extract data by row index  
▪ If we want to slice data from specific rows, loc and iloc can be used.  
▪ loc  
• Here indexing is done based on explicit index - locates by name.  
• It can take string indexes to retrive data from the specified rows.  
▪ iloc  
• Here indexing is done based on Python's numerical index - locates by integer index.  
• This works only with integer indexes to retrieve data from specified rows.  
37

---Page 38---

Bigdata and Artificial Intelligence  
Data Selection: Extract data by row index  
38

---Page 39---

Bigdata and Artificial Intelligence  
Data Selection: Extract data by row index  
39

---Page 40---

Bigdata and Artificial Intelligence  
Data Selection: Extract data by row index  
40

---Page 41---

Bigdata and Artificial Intelligence  
Data Selection: Extract data by row index  
▪ Retrieving data with row indices  
41

---Page 42---

Bigdata and Artificial Intelligence  
Data Selection: Extract data by row index  
▪ Retrieving data with row indices  
42

---Page 43---

Bigdata and Artificial Intelligence  
Data Selection: Conditional Filtering  
▪ Let's try to pick the 'Title','Actors','Director' for movies where the  
rating is minimum.  
43

---Page 44---

Bigdata and Artificial Intelligence  
Data Selection: Conditional Filtering  
▪ What if we want to pick only movies that are released from 2010 to  
2016, have a rating of less than 6.0 but topped in terms of revenue.  
44

---Page 45---

Bigdata and Artificial Intelligence  
pandas  
TOC  
  
• Introduction to pandas  
• Data Types  
ㅁ  
• Data Exploration  
• Data Selection  
• Data Manipulation  
• Exercises

---Page 46---

Bigdata and Artificial Intelligence  
Data Manipulation: Groupby operations  
▪ groupby is another interesting operation that can be peformed using  
Pandas groupby( ) method.  
▪ This comes in handly when we want to apply aggregation operations  
and functions on a set of grouped data.  
46

---Page 47---

Bigdata and Artificial Intelligence  
Data Manipulation: Groupby operations  
▪ Let's group the data by Director and see what's the revenue earned by each director.  
• Here all of the data is grouped by 'Director' column and then aggregation 'sum' is applied.  
47

---Page 48---

Bigdata and Artificial Intelligence  
Data Manipulation: Groupby operations  
▪ Aggregation is applied to all of the numerical attributes in the data.  
▪ If we want to apply specific aggregation to only selected attribute,  
we can do so selecting the columns as described below.  
48

---Page 49---

Bigdata and Artificial Intelligence  
Data Manipulation: Sorting  
▪ Sorting is one of the most important functions of pandas which is  
heavily used for data analysis.  
▪ In the above example, where we have listed the average rating for  
each Director, if we want to sort them from highly rated to lowest,  
we can perform the sorting operation.  
49

---Page 50---

Bigdata and Artificial Intelligence  
Data Manipulation: Sorting  
▪ We can see that Director 'Nitesh Tiwari' has the highest average  
rating in this dataset.  
50

---Page 51---

Bigdata and Artificial Intelligence  
Data Manipulation: Sorting  
▪ Now if we want to see which movies had the highest revenue and the highest  
rating, we can do so by following this.  
• Here sorting operation is performed on Revenue followed by Rating.  
51

---Page 52---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
▪ Pandas has isnull( ) and notnull( ) for detecting null values in a dataframe.  
52

---Page 53---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
▪ Here we know that 'Revenue (Millions)' and 'Metascore' are two columns  
where there are null values.  
53

---Page 54---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
▪ As we have seen null values in data, we can either choose to drop those or  
impute these values.  
• Dropping can be done either by rows or by columns depending on our need.  
• drop( ), dropna( ) are some functions used to drop null values.  
• When we drop values, either the complete column or row is dropped and not the single  
values.  
54

---Page 55---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
▪ Let's say I want to drop the rows where Metascore is null  
55

---Page 56---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
▪ Let's say I want to drop the rows where Metascore is null  
• Here we can clearly see that 2 columns with missing data are dropped  
56

---Page 57---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
▪ we can use thresh parameter to specify the minimum number of non null  
values for the column/row to be held without dropping  
57

---Page 58---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
▪ we can use thresh parameter to specify the minimum number of non null  
values for the column/row to be held without dropping  
• Here none of the rows are dropped because data is missing only in 2 columns  
58

---Page 59---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
▪ Here we can see that 'Metascore' column is dropped completely.  
59

---Page 60---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
▪ We know that there are some records where the Revenue is null.  
▪ We can impute these null values with mean Revenue,  
• fillna( ) --> function used to fill null values with specified values  
inplace = True signifies that the changes be made permanently in the dataset.  
60

---Page 61---

Bigdata and Artificial Intelligence  
Data Manipulation: Dealing with missing values  
61

---Page 62---

Bigdata and Artificial Intelligence  
Data Manipulation: Apply( ) Functions  
▪ The apply( ) comes in handy when we want to apply any function to  
the dataset  
▪ apply( ) function returns a value after passing each row of the  
dataframe to some function  
▪ The function can be built-in or user defined  
62

---Page 63---

Bigdata and Artificial Intelligence  
Data Manipulation: Apply( ) Functions  
▪ For example, if we want to classify the movies based on their ratings,  
we can define a function to do so and then apply the function to the  
dataframe as shown below  
63

---Page 64---

Bigdata and Artificial Intelligence  
Data Manipulation: Apply( ) Functions  
64

---Page 65---

Bigdata and Artificial Intelligence  
Data Manipulation: Pivot table  
▪ Pivot table is another useful functionality when it comes to data  
analysis  
▪ Remeber the pivot tables in excel? Pandas can be used to create  
such pivot tables as well  
▪ Pivot table takes the column wise data as input and groups the rows  
to create summary of data  
▪ The pivot table should be created based on a particular column,  
index and aggregation function to be applied on top of the data  
65

---Page 66---

Bigdata and Artificial Intelligence  
Data Manipulation: Pivot table  
66

---Page 67---

Bigdata and Artificial Intelligence  
pandas  
TOC  
  
• Introduction to pandas  
• Data Types  
• Data Exploraㅁ tion  
• Data Selection  
• Data Manipulation  
• Exercises

---Page 68---

Bigdata and Artificial Intelligence  
Basic Level Questions  
▪ Q1. Display the number of rows and columns  
in the dataset.  
▪ Q2. Find the maximum, minimum, mean, and  
median values of movie ratings.  
▪ Q3. Identify columns with missing values, and  
show the number of missing values per column.  
68

---Page 69---

Bigdata and Artificial Intelligence  
Intermediate Level Questions  
▪ Q4. Find the title and director of the  
movie that earned the highest revenue.  
▪ Q5. List the movies released after 2015  
with a rating of 8.0 or higher. Display  
only the Title and Rating columns.  
▪ Q6. Calculate the average rating for  
each director, and display the top 3  
directors with the highest average  
ratings.  
69

---Page 70---

Bigdata and Artificial Intelligence  
pandas  
  
Q&A  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr

## ===START OF FILE: C:\Users\ambyb\Downloads\BA\_W5\_visualization.pdf===

---Page 1---

Bigdata and Artificial Intelligence  
pandas  
TOC  
  
• Introduction to pandas  
• Data Types  
• Data Exploraㅁ tion  
• Data Selection  
• Data Manipulation  
• Exercises

---Page 2---

Bigdata and Artificial Intelligence  
Basic Level Questions  
▪ Q1. Display the number of rows and columns  
in the dataset.  
▪ Q2. Find the maximum, minimum, mean, and  
median values of movie ratings.  
▪ Q3. Identify columns with missing values, and  
show the number of missing values per column.  
2

---Page 3---

Bigdata and Artificial Intelligence  
Intermediate Level Questions  
▪ Q4. Find the title and director of the  
movie that earned the highest revenue.  
▪ Q5. List the movies released after 2015  
with a rating of 8.0 or higher. Display  
only the Title and Rating columns.  
▪ Q6. Calculate the average rating for  
each director, and display the top 3  
directors with the highest average  
ratings.  
3

---Page 4---

Bigdata and Artificial Intelligence  
pandas  
  
Q&A  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr

---Page 5---

Bigdata and Artificial Intelligence  
Visualizing Data with Matplotlib  
  
Hanyang University  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr

---Page 6---

Bigdata and Artificial Intelligence  
Visualizing Data with Matplotlib  
TOC  
  
• Matplotlib  
• Exercises  
ㅁ

---Page 7---

Bigdata and Artificial Intelligence  
Setup  
▪ Colab playground  
• https://colab.research.google.com/drive/1zaZ5iCueuhR27gaZKDsNjZQ5IjTMcEJo?usp=sh  
aring  
▪ We will be working with 2 datasets:  
• Facebook's stock price throughout 2018  
(fb\_stocks\_prices\_2018.csv)  
• Earthquake data from September 18, 2018 - October 13, 2018  
(earthquakes.csv)  
• European Centre for Disease Prevention and Control's (ECDC) daily number of new  
reported cases of COVID-19 by country worldwide dataset collected on September 19,  
2020 via this link  
(covid19\_cases.csv)  
7

---Page 8---

Bigdata and Artificial Intelligence  
Import  
▪ We need to import matplotlib.pyplot for plotting.  
8

---Page 9---

Bigdata and Artificial Intelligence  
Plotting lines  
▪ We need to import matplotlib.pyplot for plotting.  
9

---Page 10---

Bigdata and Artificial Intelligence  
Plotting lines  
▪ Since we are working in a Jupyter notebook, we can use the magic  
command %matplotlib inline once and not have to call plt.show() for each plot.  
10

---Page 11---

Bigdata and Artificial Intelligence  
Scatter plots  
▪ We can pass in a string specifying the style of the plot. This is of the form  
[marker][linestyle][color].  
▪ For example, we can make a black dashed line with '--k' or a red scatter plot  
with 'or':  
11

---Page 12---

Bigdata and Artificial Intelligence  
Scatter plots  
▪ Here are some examples of how you make a format string:  
▪ Note that we can also use format strings of the form [color][marker][linestyle], but  
the parsing by matplotlib (in rare cases) might not be what we were aiming for.  
12

---Page 13---

Bigdata and Artificial Intelligence  
Histograms  
13

---Page 14---

Bigdata and Artificial Intelligence  
Histograms  
▪ Notice how our assumptions of the distribution of the data can change based  
on the number of bins (look at the drop between the two highest peaks on the  
righthand plot):  
14

---Page 15---

Bigdata and Artificial Intelligence  
Plot components  
▪ Figure : Top-level object that holds the other plot components.  
▪ Axes : Individual plots contained within the Figure.  
15

---Page 16---

Bigdata and Artificial Intelligence  
Creating subplots  
▪ Simply specify the number of rows and columns to create:  
16

---Page 17---

Bigdata and Artificial Intelligence  
Creating subplots  
▪ As an alternative to using plt.subplots() we can add Axes objects to the Figure  
object on our own.  
▪ This allows for some more complex layouts, such as picture in picture:  
17

---Page 18---

Bigdata and Artificial Intelligence  
Creating Plot Layouts with gridspec  
▪ We can create subplots with varying sizes as well:  
18

---Page 19---

Bigdata and Artificial Intelligence  
Saving plots  
▪ Use plt.savefig() to save the last created plot.  
▪ To save a specific Figure object, use its savefig() method.  
19

---Page 20---

Bigdata and Artificial Intelligence  
Cleaning up  
▪ It's important to close resources when we are done with them.  
▪ We use plt.close() to do so.  
▪ If we pass in nothing, it will close the last plot, but we can pass in the  
specific Figure object to close or say 'all' to close all Figure objects  
that are open.  
▪ Let's close all the Figure objects that are open with plt.close():  
20

---Page 21---

Bigdata and Artificial Intelligence  
Additional plotting options: Specifying figure size  
▪ Just pass the figsize argument to plt.figure(). It's a tuple of (width, height):  
▪ This can be specified when creating subplots as well:  
21

---Page 22---

Bigdata and Artificial Intelligence  
Additional plotting options: rcParams  
▪ A small subset of all the available plot settings (shuffling to get a good variation  
of options):  
22

---Page 23---

Bigdata and Artificial Intelligence  
Additional plotting options: rcParams  
▪ We can check the current default figsize using rcParams:  
▪ We can also update this value to change the default (until the kernel is  
restarted):  
23

---Page 24---

Bigdata and Artificial Intelligence  
Additional plotting options: rcParams  
▪ Use rcdefaults() to restore the defaults.  
• Note this is slightly different than before  
because running %matplotlib inline sets a  
different value for figsize.  
• After we reset, we are going back to the  
default value of figsize before that import:  
▪ This can also be done via pyplot:  
24

---Page 25---

Bigdata and Artificial Intelligence  
More about Matplotlib  
https://matplotlib.org/  
25

---Page 26---

Bigdata and Artificial Intelligence  
Visualizing Data with Matplotlib  
TOC  
  
• Matplotlib  
• Exercises  
ㅁ

---Page 27---

Bigdata and Artificial Intelligence  
Exercises  
▪ Q1. Create a histogram showing the distribution of movie ratings.  
plt.figure(figsize=(8, 5))  
plt.hist(df['Rating'], bins=15, color='skyblue', edgecolor='black')  
plt.title('Distribution of Movie Ratings')  
plt.xlabel('Rating')  
plt.ylabel('Number of Movies')  
plt.grid(True)  
plt.show()  
27

---Page 28---

Bigdata and Artificial Intelligence  
Exercises  
▪ Q2. Scatter Plot of Ratings vs Revenue  
plt.figure(figsize=(8, 5))  
plt.scatter(df['Rating'], df['Revenue (Millions)'], alpha=0.6)  
plt.title('Ratings vs Revenue')  
plt.xlabel('Rating')  
plt.ylabel('Revenue (Millions)')  
plt.grid(True)  
plt.show()  
28

---Page 29---

Bigdata and Artificial Intelligence  
Exercises  
▪ Q3. Bar Chart of Top 10 Most Frequent Directors.  
(Hint : value\_counts() )  
top\_directors = df['Director'].value\_counts().head(10)  
plt.figure(figsize=(10, 5))  
Plt.bar(top\_directors.index, top\_directors.values, color=‘orange’)  
Plt.title(‘Top 10 Most Frequent Directors’)  
Plt.xlabel(‘Director’)  
plt.ylabel('Number of Movies')  
plt.xticks(rotation=45, ha='right')  
plt.tight\_layout()  
plt.show()  
29

---Page 30---

Bigdata and Artificial Intelligence  
Exercises  
▪ Q4. Line Plot of Average Ratings per Year (Hint : groupby() )  
avg\_rating\_by\_year = df.groupby('Year')['Rating'].mean()  
plt.figure(figsize=(10, 5))  
plt.plot(avg\_rating\_by\_year.index, avg\_rating\_by\_year.values,  
marker='o')  
plt.title('Average Movie Rating by Year')  
plt.xlabel('Year')  
plt.ylabel('Average Rating')  
plt.grid(True)  
plt.show()  
30

---Page 31---

Bigdata and Artificial Intelligence  
Exercises  
▪ Q5. Box Plot of Ratings by Genre (Top 3 Genres)  
# Only choose the first one from the Genre values  
df['MainGenre'] = df['Genre'].apply(lambda x: x.split(',')[0].strip())  
top\_genres = df['MainGenre'].value\_counts().head(3).index  
plt.figure(figsize=(8, 5))  
data = [df[df['MainGenre'] == genre]['Rating'] for genre in top\_genres]  
plt.boxplot(data, labels=top\_genres)  
plt.title('Rating Distribution by Top 3 Genres')  
plt.ylabel('Rating')  
plt.grid(True)  
plt.show()  
31

---Page 32---

Bigdata and Artificial Intelligence  
Visualizing Data with Matplotlib  
  
Q&A  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr

## ===START OF FILE: C:\Users\ambyb\Downloads\BA\_W6\_Text Data.pdf===

---Page 1---

Bigdata and Artificial Intelligence  
Text Data  
  
Hanyang University  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr

---Page 2---

Bigdata and Artificial Intelligence | Text Data  
Colab Link  
▪ https://colab.research.google.com/drive/1tYJeovkqX1YbrkgEBV5lxCjbKs\_UNpZV?usp=sharing

---Page 3---

Bigdata and Artificial Intelligence | Text Data  
Natural Language Processing (NLP)  
▪ Natural language processing (NLP) is the ability of a computer program to understand  
human language as it is spoken and written.  
▪ Natural language: Human language  
• Speech, Conversation, Sentence, Sign Language…  
▪ Diverse forms of natural language  
• Social media posts  
• Web page  
• Hospital prescription  
• Audio data in voice mail  
• …  
▪ Applications of NLP : Machine translation, AI speaker, chatbot, STT (Speech-To-Text)

---Page 4---

Bigdata and Artificial Intelligence  
Text Data  
TOC  
  
• Text Data  
• Text Preprocessing  
ㅁ  
• Text Feature  
(Document Modeling)

---Page 5---

Bigdata and Artificial Intelligence | Text Data  
Corpus  
▪ Corpus is a refined text dataset.  
• NLTK includes more than 50 large, well-organized text data.  
▪ Examples of Corpus  
• Web Text Corpus, Twitter Corpus, Shakespeare Corpus,  
Polarization of emotions (positive vs. negation),  
Names Corpus, WordNet, Reuters Benchmark Corpus, …

---Page 6---

Bigdata and Artificial Intelligence | Text Data  
Install Corpus  
▪ Install the names corpus from NLTK  
• http://www.nltk.org/nltk\_data/ (Check available corpus)

---Page 7---

Bigdata and Artificial Intelligence | Text Data  
namescorpus  
# import the names corpus to Python code  
# Print the first 5 words  
# Print the total number of words in the names corpus

---Page 8---

Bigdata and Artificial Intelligence | Text Data  
“20 newsgroups” dataset  
▪ Example dataset provided by sklearn  
• 20,000 posts exists for 20 online newsgroups.  
• A newsgroup is an Internet space (Internet Forum) where you can ask and  
answer questions about a particular topic.  
• comp.graphics • rec.sport.baseball • talk.politics.misc  
• comp.os.ms-windows.misc • rec.sport.hockey • talk.politics.guns  
• comp.sys.ibm.pc.hardware • sci.crypt • talk.politics.mideast  
• comp.sys.mac.hardware • sci.electronics • talk.religion.misc  
• comp.windows.x • sci.med • alt.atheism  
• rec.autos • sci.space • soc.religion.christian  
• rec.motorcycles • misc.forsale

---Page 9---

Bigdata and Artificial Intelligence | Text Data  
“20 newsgroups” dataset  
▪ You can download a dataset with sklearn functions.  
• Caching occurs when code is first executed and downloaded.  
• On subsequent runs, cached data is used without re-downloading.  
▪ Downloaded datasets are available as Python Dictionary.

---Page 10---

Bigdata and Artificial Intelligence | Text Data  
“20 newsgroups” dataset  
▪ Key of dataset Dictionary as follows  
• data : List of post texts  
• filenames : List of filepath for each post  
• target\_names : List of group name text (Total 20)  
· List index represents a group id.  
• target : List of group ids for each post.  
• DESCR : Description of '20 newsgroups' dataset

---Page 11---

Bigdata and Artificial Intelligence | Text Data  
“20 newsgroups” dataset  
▪ Main text is mapped to the "data" key as a list, and each is  
accessible through the index value of the list

---Page 12---

Bigdata and Artificial Intelligence | Text Data  
“20 newsgroups” dataset  
# group id for the first post  
▪ The "target" key contains the  
group id of each post text  
# Group names by group id  
• The first data (Index = 0) has 7  
as its group id.  
• Group id exists from 0 to 19.  
▪ To see the newsgroup name  
for each group id, see the  
"target\_names" key.

---Page 13---

Bigdata and Artificial Intelligence  
Text Data  
TOC  
  
• Text Data  
• Text Preprocessing  
ㅁ  
• Text Feature  
(Document Modeling)

---Page 14---

Bigdata and Artificial Intelligence | Text Data  
Tokenization  
▪ Tokenization breaks a given text sequence into pieces.  
▪ Divided pieces are called tokens and become the basic unit  
of subsequent analysis.

---Page 15---

Bigdata and Artificial Intelligence | Text Data  
n-gram  
▪ Represents the number of fragments that make up the token  
• Unigram (Unigram): One piece  
• Bigram: Two consecutive pieces  
• Trigram: 3 consecutive pieces  
• N-gram (n-gram): nconsecutive pieces of tokens  
Input: Machine learning is awesome, right?  
Unigram: Machine learning is awesome right  
Bigram: Machine learning Learning is Is awesome awesome right

---Page 16---

Bigdata and Artificial Intelligence | Text Data  
POS tagging  
▪ Tagging each token in a given text sequence with the appropriate  
part of speech  
Part of Speech Example  
Noun David, machine  
Pronoun them, her  
Adjective awesome, amazing  
Verb read, write  
Adverb very, quite  
Preposition out, at  
Conjunction and, but  
Interjection unfortunately, luckily  
Article a, the

---Page 17---

Bigdata and Artificial Intelligence | Text Data  
POS tagging  
• NNP: Proper Noun  
• VB: Verb  
• VBP: Verb (Present)  
• TO: to Preposition  
• NN: Noun  
• DT: Determiner  
Token POS

---Page 18---

Bigdata and Artificial Intelligence | Text Data  
Name Entity Recognition (NER)  
▪ NER specifies a word or phrase into a particular category, such as a  
person name, company name, or region name.  
• Hanyang University → University  
• → Person  
• Gyeonggido → Region

---Page 19---

Bigdata and Artificial Intelligence | Text Data  
Name Entity Recognition (NER)

---Page 20---

Bigdata and Artificial Intelligence | Text Data  
Stemming  
▪ Stemming is the process of reducing inflected (or  
sometimes derived) words to their word stem, base or root  
form—generally a written word form.

---Page 21---

Bigdata and Artificial Intelligence | Text Data  
Lemmatization  
▪ Lemmatization is the process of grouping together the  
inflected forms of a word so they can be analyzed as a  
single item, identified by the word's lemma, or dictionary  
form.

---Page 22---

Bigdata and Artificial Intelligence  
Text Data  
TOC  
  
• Text Data  
• Text Preprocessing  
ㅁ  
• Text Feature  
(Document Modeling)

---Page 23---

Bigdata and Artificial Intelligence | Text Data  
Document Model  
▪ The document model is the representation of a document by its  
characteristics and its values.  
▪ A document can be represented as a kind of bag of words if the entire  
order of words is not considered important in a document.  
• Example: Document 1 → Model: {'Word 1': 3, 'Word 2': 4 …}  
• Simple but practically works well  
▪ There are ways to use word order and speech tags  
▪ Word-embedding based on deep learning have been also widely used.  
• Word2Vec, Doc2Vec, etc.

---Page 24---

Bigdata and Artificial Intelligence | Text Data  
Term Frequency (TF)  
▪ One of the key textual characteristics about the document  
is term frequency.  
• Each document is represented by each word’s occurrence.  
Car Train Coffee Cookie …  
Doc 1 3 4 0 0  
Doc 2 3 3 0 1  
Doc 3 1 1 7 6  
… … …

---Page 25---

Bigdata and Artificial Intelligence | Text Data  
Term Frequency (TF)  
▪ CountVectorizer can be used for TF-based Model.  
▪ STEP 1: Modeling  
Option for selecting top 100 terms by the total frequency

---Page 26---

Bigdata and Artificial Intelligence | Text Data  
Term Frequency (TF)  
▪ Step 2: Transforming  
• Transform a document with term frequency for 100 selected words  
# Transform each document  
# Example of a transformed document (TF)

---Page 27---

Bigdata and Artificial Intelligence | Text Data  
Term Frequency (TF)  
▪ Words with little information (called Stopwords) can be  
excluded from the beginning.  
• Ex) a, the, are

---Page 28---

Bigdata and Artificial Intelligence | Text Data  
Term Frequency (TF)  
▪ Options

---Page 29---

Bigdata and Artificial Intelligence | Text Data  
TF-IDF  
▪ Terms that are commonly found in most documents do not  
have the unique characteristics of the document well even  
if their TF value is high.  
▪ TF-IDF (Term Frequency-Inverse Document Frequency) is  
the method of penalizing words occurred in many  
documents.  
𝑛 𝑛 : The number of all the documents  
𝐷 𝐷  
𝑇𝐹\_𝐼𝐷𝐹 = 𝑇𝐹 ∙ log  
1 + 𝑛  
𝑡 𝑛 : The number of the documents containing term t  
𝑡

---Page 30---

Bigdata and Artificial Intelligence | Text Data  
TF-IDF  
TF  
Car Train Coffee Cookie …  
Doc 1 3 4 0 0  
Doc 2 3 3 0 1  
Doc 3 1 1 7 6  
IDF  
Car Train Coffee Cookie …  
log(3/4) log(3/4) log(3/2) log(3/3)  
TF-IDF  
Car Train Coffee Cookie …  
Doc 1 3\*log(3/4) 4\*log(3/4) 0\*log(3/2) 0\*log(3/3)  
Doc 2 3\*log(3/4) 3\*log(3/4) 0\*log(3/2) 1\*log(3/3)  
Doc 3 1\*log(3/4) 1\*log(3/4) 7\*log(3/2) 6\*log(3/3)

---Page 31---

Bigdata and Artificial Intelligence | Text Data  
TF-IDF  
▪ TfidfVectorizer can be used for TFIDF-based Model.  
▪ STEP 1: Modeling

---Page 32---

Bigdata and Artificial Intelligence | Text Data  
TF-IDF  
▪ Step 2: Transforming  
• Transform a document with TF-IDF values for 100 selected words  
# Transform each document  
# Example of a transformed document (TF-IDF)

---Page 33---

Bigdata and Artificial Intelligence | Text Data  
Other Text Features  
▪Frequency-based features  
• Term Existence  
• Term Frequency  
• TF-IDF  
▪ Term appearance order  
▪ Text length  
▪ …

---Page 34---

Bigdata and Artificial Intelligence  
Text Data  
TOC  
  
• Text Data  
• Text Preprocessing  
ㅁ  
• Text Feature  
(Document Modeling)

---Page 35---

Bigdata and Artificial Intelligence | Text Data  
Colab Link  
▪ https://colab.research.google.com/drive/1SpLzYR4\_wornISDdu\_JiL0\_2XEMnJquc?usp=sharing

---Page 36---

Bigdata and Artificial Intelligence | Text Data  
Import Modules

---Page 37---

Bigdata and Artificial Intelligence | Text Data  
Read the dataset  
Before running the code, download the two dataset files from LMS

---Page 38---

Bigdata and Artificial Intelligence | Text Data  
Check the dataset

---Page 39---

Bigdata and Artificial Intelligence | Text Data  
Combine all the reviews into one per movie

---Page 40---

Bigdata and Artificial Intelligence | Text Data  
Choose a Movie

---Page 41---

Bigdata and Artificial Intelligence | Text Data  
TF Weight

---Page 42---

Bigdata and Artificial Intelligence | Text Data  
Visualize Terms with TF Weights

---Page 43---

Bigdata and Artificial Intelligence | Text Data  
TFIDF Weight

---Page 44---

Bigdata and Artificial Intelligence | Text Data  
Visualize Terms with TFIDF Weights

---Page 45---

Bigdata and Artificial Intelligence | Text Data  
TFIDF weight + Only Noun / Adjectives

---Page 46---

Bigdata and Artificial Intelligence | Text Data  
TFIDF weight + Only Noun / Adjectives

---Page 47---

Bigdata and Artificial Intelligence  
Text Data  
  
Q&A  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr

## ===START OF FILE: C:\Users\ambyb\Downloads\BA\_W2\_Python101.pdf===

---Page 1---

Bigdata and Artificial Intelligence  
Python 101  
  
Hanyang University  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr

---Page 2---

Bigdata and Artificial Intelligence  
Python 101  
TOC  
  
• Colab  
• Basic data types  
• Variables  
ㅁ  
• Collections of items  
• Functions  
• Control Flow Statements  
• Loops  
• Packages  
• Class

---Page 3---

Bigdata and Artificial Intelligence  
Google Colab  
https://colab.research.google.com  
REF: https://www.jcchouinard.com/google-colab-with-python/  
3

---Page 4---

Bigdata and Artificial Intelligence  
Create a Notebook file  
▪ File → New Python3 Note  
• You need your Google account.  
Write your codes here  
4

---Page 5---

Bigdata and Artificial Intelligence  
How to Run Python Code in Google Colab?  
▪ You can run code or markdown instantly in any cell.  
5

---Page 6---

Bigdata and Artificial Intelligence  
Explore Your Colab Environment  
6

---Page 7---

Bigdata and Artificial Intelligence  
Make Your Notebook More Interesting with Markdown  
7

---Page 8---

Bigdata and Artificial Intelligence  
Useful Keyboard Shortcuts in Google Colab  
Command Shortcut  
Create cell Command/Ctrl+M+B  
Select all cells ⌘/Ctrl+Shift+A  
Run all cells ⌘/Ctrl+F9  
Run the selected cell Command/Ctrl+Enter  
Interrupt execution Command or Ctrl+M+I  
8

---Page 9---

Bigdata and Artificial Intelligence  
Other Tips  
9

---Page 10---

Bigdata and Artificial Intelligence  
Python 101  
TOC  
  
• Colab  
• Basic data types  
• Variables  
ㅁ  
• Collections of items  
• Functions  
• Control Flow Statements  
• Loops  
• Packages  
• Class

---Page 11---

Bigdata and Artificial Intelligence  
Basic data types: Numbers  
▪ Numbers in Python can be represented as integers (e.g. 5) or floats  
(e.g. 5.0).  
▪ We can perform operations on them:  
11

---Page 12---

Bigdata and Artificial Intelligence  
Basic data types: Booleans  
▪ We can check for equality giving us a Boolean:  
12

---Page 13---

Bigdata and Artificial Intelligence  
Basic data types: Booleans  
▪ These statements can be combined with logical operators: not, and, or  
13

---Page 14---

Bigdata and Artificial Intelligence  
Basic data types: Strings  
▪ Using strings, we can handle text in Python.  
▪ These values must be surrounded in quotes — single ('...') is the standard, but  
double ("...") works as well:  
▪ We can also perform operations on strings. For example, we can see how long it  
is with len():  
14

---Page 15---

Bigdata and Artificial Intelligence  
Basic data types: Strings  
▪ We can select parts of the string by specifying the index  
(Note that in Python the 1st character is at index 0:)  
▪ We can concatentate strings with +:  
▪ We can check if characters are in the string with the in operator:  
15

---Page 16---

Bigdata and Artificial Intelligence  
Python 101  
TOC  
  
• Colab  
• Basic data types  
• Variables  
ㅁ  
• Collections of items  
• Functions  
• Control Flow Statements  
• Loops  
• Packages  
• Class

---Page 17---

Bigdata and Artificial Intelligence  
Variables  
▪ Variables give us a way to store data types.  
• We define a variable using the variable\_name = value syntax:  
▪ The variable name cannot contain spaces; we usually use \_ instead.  
• The best variable names are descriptive ones:  
17

---Page 18---

Bigdata and Artificial Intelligence  
Variables  
▪ Variables can be any data type.  
▪ We can check which one it is with type(), which is a function (more on that later):  
▪ If we need to see the value of a variable, we can print it using the print()  
function:  
18

---Page 19---

Bigdata and Artificial Intelligence  
Variables  
▪ Notice that just typing text causes an error.  
• Errors in Python attempt to clue us in to what went wrong with our code.  
• In this case, we have a NameError exception which tells us that 'hello' is not  
defined.  
• This means that the Python interpreter looked for a variable named hello, but it  
didn't find one.  
19

---Page 20---

Bigdata and Artificial Intelligence  
Python 101  
TOC  
  
• Colab  
• Basic data types  
• Variables  
ㅁ  
• Collections of items  
• Functions  
• Control Flow Statements  
• Loops  
• Packages  
• Class

---Page 21---

Bigdata and Artificial Intelligence  
Collections of Items: Lists  
▪ We can store a collection of items in a list:  
▪ The list can be stored in a variable. Note that the items in the list can  
be of different types:  
21

---Page 22---

Bigdata and Artificial Intelligence  
Collections of Items: Lists  
▪ We can see how many elements are in the list with len():  
▪ We can also use the in operator to check if a value is in the list:  
22

---Page 23---

Bigdata and Artificial Intelligence  
Collections of Items: Lists  
▪ We can select items in the list just as we did with strings, by  
providing the index to select:  
▪ Python also allows us to use negative values, so we can easily select  
the last one:  
23

---Page 24---

Bigdata and Artificial Intelligence  
Collections of Items: Lists  
▪ Another powerful feature of lists (and strings) is slicing.  
▪ ... or every other one:  
▪ We can even select the list in reverse:  
24

---Page 25---

Bigdata and Artificial Intelligence  
Collections of Items: Lists  
▪ Note: This syntax is [start:stop:step] where the selection is  
inclusive of the start index, but exclusive of the stop index.  
• If start isn't provided, 0 is used.  
• If stop isn't provided, the number of elements is used (4, in our case);  
this works because the stop is exclusive.  
• If step isn't provided, it is 1.  
25

---Page 26---

Bigdata and Artificial Intelligence  
Collections of Items: Lists  
▪ We can use the join() method on a string object to  
concatenate all the items of a list into single string.  
▪ The string we call the join() method on will be used as the  
separator, here we separate with a pipe (|):  
26

---Page 27---

Bigdata and Artificial Intelligence  
Collections of Items: Tuples  
▪ Tuples are similar to lists; however, they can't be modified after  
creation i.e. they are immutable.  
▪ Instead of square brackets, we use parenthesis to create tuples:  
27

---Page 28---

Bigdata and Artificial Intelligence  
Collections of Items: Tuples  
▪ Immutable objects can't be modified:  
28

---Page 29---

Bigdata and Artificial Intelligence  
Collections of Items: Dictionaries  
▪ We can store mappings of key-value pairs using dictionaries:  
29

---Page 30---

Bigdata and Artificial Intelligence  
Collections of Items: Dictionaries  
▪ To access the values associated with a specific key, we use the  
square bracket notation again:  
▪ We can extract all of the keys with keys():  
30

---Page 31---

Bigdata and Artificial Intelligence  
Collections of Items: Dictionaries  
▪ We can extract all of the values with values():  
▪ Finally, we can call items() to get back pairs of (key, value) pairs:  
31

---Page 32---

Bigdata and Artificial Intelligence  
Collections of Items: Sets  
▪ A set is a collection of unique items; a common use is to remove  
duplicates from a list.  
▪ These are written with curly braces also, but notice there is no key-  
value mapping:  
32

---Page 33---

Bigdata and Artificial Intelligence  
Collections of Items: Sets  
▪ How many items are in this set?  
▪ We put in 4 items but the set only has 3 because duplicates are removed:  
▪ We can check if a value is in the set:  
33

---Page 34---

Bigdata and Artificial Intelligence  
Python 101  
TOC  
  
• Colab  
• Basic data types  
• Variables  
ㅁ  
• Collections of items  
• Functions  
• Control Flow Statements  
• Loops  
• Packages  
• Class

---Page 35---

Bigdata and Artificial Intelligence  
Functions  
▪ We can define functions to package up our code for reuse.  
▪ We have already seen some functions: len(), type(), and print().  
• They are all functions that take arguments.  
▪ Note that functions don't need to accept arguments, in which case they are  
called without passing in anything (e.g. print() versus print(my\_string)).  
▪ Aside: we can also create lists, sets, dictionaries, and tuples with functions:  
list(), set(), dict(), and tuple()  
35

---Page 36---

Bigdata and Artificial Intelligence  
Defining functions  
▪ We use the def keyword to define functions.  
▪ Let's create a function called add() with 2 parameters, x and y, which will be the  
names the code in the function will use to refer to the arguments we pass in  
when calling it:  
36

---Page 37---

Bigdata and Artificial Intelligence  
Defining functions  
▪ Once we run the code above, our function is ready to use:  
▪ Let's add some numbers:  
37

---Page 38---

Bigdata and Artificial Intelligence  
Return values  
▪ We can store the result in a variable for later:  
▪ Notice the print statement wasn't captured in result. This variable will only have what the  
function returns. This is what the return line in the function definition did:  
▪ Note that functions don't have to return anything. Consider print():  
38

---Page 39---

Bigdata and Artificial Intelligence  
Return values  
▪ If we take a look at what we got back, we see it is a NoneType object:  
▪ In Python, the value None represents null values. We can check if our variable is  
None:  
▪ Warning: make sure to use comparison operators (e.g. >, >=, <, <=, ==, !=) to  
compare to values other than None.  
39

---Page 40---

Bigdata and Artificial Intelligence  
Function arguments  
▪ Note that function arguments can be anything, even other functions.  
▪ We will see several examples of this in the text. The function we defined  
requires arguments.  
▪ If we don't provide them all, it will cause an error:  
40

---Page 41---

Bigdata and Artificial Intelligence  
Function arguments  
▪ We can use help() to check what arguments the function needs (notice the  
docstring ends up here):  
41

---Page 42---

Bigdata and Artificial Intelligence  
Function arguments  
▪ We will also get errors if we pass in data types that add() can't work with:  
42

---Page 43---

Bigdata and Artificial Intelligence  
Python 101  
TOC  
  
• Colab  
• Basic data types  
• Variables  
ㅁ  
• Collections of items  
• Functions  
• Control Flow Statements  
• Loops  
• Packages  
• Class

---Page 44---

Bigdata and Artificial Intelligence  
Control Flow Statements  
▪ Sometimes we want to vary the path the code takes based on some criteria.  
▪ For this we have if, elif, and else. We can use if on its own:  
44

---Page 45---

Bigdata and Artificial Intelligence  
Control Flow Statements  
▪ Calling this function with negative input causes the code under the if statement  
to run:  
▪ Calling this function with positive input skips the code under the if statement,  
keeping the number positive:  
45

---Page 46---

Bigdata and Artificial Intelligence  
Control Flow Statements  
▪ Sometimes we need an else statement as well:  
▪ This triggers the code under the if statement:  
▪ Since the Boolean check in the if statement was False, this triggers the code  
under the else statement:  
46

---Page 47---

Bigdata and Artificial Intelligence  
Control Flow Statements  
▪ For more complicated logic, we can also use elif. We can have any number of elif  
statements. Optionally, we can include else.  
47

---Page 48---

Bigdata and Artificial Intelligence  
Control Flow Statements  
▪ The code keeps checking the conditions in the if statements from top to bottom  
until it finds multiply:  
▪ The code keeps checking the conditions in the if statements from top to bottom  
until it hits the else statement:  
48

---Page 49---

Bigdata and Artificial Intelligence  
Python 101  
TOC  
  
• Colab  
• Basic data types  
• Variables  
ㅁ  
• Collections of items  
• Functions  
• Control Flow Statements  
• Loops  
• Packages  
• Class

---Page 50---

Bigdata and Artificial Intelligence  
Loops: while loops  
▪ With while loops, we can keep running code until some stopping condition is  
met:  
50

---Page 51---

Bigdata and Artificial Intelligence  
Loops: while loops  
▪ Note this can also be written as, by moving the condition to the while statement:  
▪ With for loops, we don't have to worry about checking if we have reached the  
stopping condition.  
• Conversely, while loops can cause infinite loops if we don't remember to update variables.  
51

---Page 52---

Bigdata and Artificial Intelligence  
Loops: for loops  
▪ With for loops, we can run our code for each element in a collection:  
▪ We can use for loops with lists, tuples, sets, and dictionaries as well:  
52

---Page 53---

Bigdata and Artificial Intelligence  
Python 101  
TOC  
  
• Colab  
• Basic data types  
• Variables  
ㅁ  
• Collections of items  
• Functions  
• Control Flow Statements  
• Loops  
• Packages  
• Class

---Page 54---

Bigdata and Artificial Intelligence  
Imports  
▪ We have been working with the portion of Python that is available  
without importing additional functionality.  
▪ The Python standard library that comes with the install of Python is  
broken up into several modules, but we often only need a few.  
▪ We can import whatever we need: a module in the standard library,  
a 3rd-party library, or code that we wrote.  
54

---Page 55---

Bigdata and Artificial Intelligence  
Imports  
▪ This is done with an import statement:  
▪ If we only need a small piece from that module, we can do the following instead:  
55

---Page 56---

Bigdata and Artificial Intelligence  
Imports  
▪ Anything you import is added to the namespace, so if you create a  
new variable/function/etc. with the same name it will overwrite the  
previous value.  
▪ For this reason, we have to be careful with variable names e.g. if you  
name something sum, you won't be able to add using the sum()  
built-in function anymore.  
▪ Using notebooks or an IDE will help you avoid these issues with  
syntax highlighting.  
56

---Page 57---

Bigdata and Artificial Intelligence  
Installing 3rd-party Packages  
▪ To install a package, we can use pip3 install <package\_name>.  
▪ Optionally, we can provide a specific version to install  
pip3 install pandas==0.23.4.  
▪ Without that specification, we will get the most stable version.  
▪ When we have many packages to install, we will typically use a  
requirements.txt file: pip3 install -r requirements.txt.  
▪ Note: running pip3 freeze > requirements.txt will send the list of  
packages installed in the activate environment and their respective  
versions to the requirements.txt file.  
57

---Page 58---

Bigdata and Artificial Intelligence  
Python 101  
TOC  
  
• Colab  
• Basic data types  
• Variables  
ㅁ  
• Collections of items  
• Functions  
• Control Flow Statements  
• Loops  
• Packages  
• Class

---Page 59---

Bigdata and Artificial Intelligence  
Class  
▪ So far we have used Python as a functional programming language, but  
we also have the option to use it for object-oriented programming.  
▪ You can think of a class as a way to group similar functionality together.  
▪ Let's create a calculator class which can handle mathematical operations  
for us.  
• For this, we use the class keyword and define methods for taking actions on the  
calculator.  
• These methods are functions that take self as the first argument.  
• When calling them, we don't pass in anything for that argument (example after this):  
59

---Page 60---

Bigdata and Artificial Intelligence  
Class  
60

---Page 61---

Bigdata and Artificial Intelligence  
Class  
▪ In order to use the calculator, we need to instantiate an instance or  
object of type Calculator.  
▪ Since the \_\_init\_\_() method has no parameters other than self, we  
don't need to provide anything:  
61

---Page 62---

Bigdata and Artificial Intelligence  
Class  
▪ Let's try to add some numbers:  
▪ Oops!! The calculator is not on. Let's turn it on:  
▪ Let's try again:  
62

---Page 63---

Bigdata and Artificial Intelligence  
Class  
▪ We can access attributes on object with dot notation. In this  
example, the only attribute is on, and it is set in the \_\_init\_\_()  
method:  
▪ Note that we can also update attributes:  
63

---Page 64---

Bigdata and Artificial Intelligence  
Class Orientation  
  
Q&A  
Human-AI Interaction Lab  
http://hai.hanyang.ac.kr