School of Mathematical Sciences

Tribhuvan University

Program: Data Science

Python Programming: Outline (For Beginners)

- Installation
- Introduction to Jupyter Notebook/ IDE
- Defining variables and assigning values
- Basic Input/Output
- Python Data Types
- Number Data type (int and float)
- Conditions
- Iteration
- List, Set, Tuple, Dictionary
- Python Functions
- Numpy
- Pandas
- Data Visualization

Practice Set – 1: Exploring pandas dataframe and starting basic operations

- **Prerequisites:** Concept of indexing and slicing operations in list and numpy (One dimensional and Multidimensional)

1. Import the necessary libraries and import the dataset. (titanic.csv)
2. See the first 15 entries.
3. See the last 10 entries.
4. What is the shape of dataset?
5. What is the number of observations in the dataset?
6. What is the number of columns in the dataset?
7. Print the name of all the columns.
8. Print only the Pclass column.
9. What are the different values in Pclass column
10. How many different Pclass are in the dataset.
11. Summarize the dataset. (using describe() function)
12. Summarize only the Age column.
13. What is the mean age of passengers?
14. See if there is any relationship between Age column and Fare column. (Use scatter plot and also find correlation)
15. Select only the rows where age is more than 30.

16. Check if null values are present or not in the dataset. (If present display the rows)
17. Drop the column ‘Ticket’.
18. Create a new column ‘Family’ which is defined as:
    Family = SibSp + Parch

Practice Set – 2: Linear Regression

- **Prerequisites:** Successful completion of Practice Set-1, Concepts of Linear Regression

Suppose it is 2001, and you’re working as an analyst for the Oakland Athletics of Major League Baseball (MLB). Your primary goal is to make recommendations on how to develop the team to be able to make the playoffs in the following season (2002).

For this problem, you will be working with the *‘baseball.csv’* data, which contains season-team level data going back to 1962.

1. Initial Data Processing
   - First subset the data to all years prior to 2002. We will build our models on this data.
   - Save the 2002 season-level data for testing.

2. How many wins does it typically take to make the playoffs?
   - What is the typical number of games a team should win in the regular season to expect to make the playoffs?

3. How does a team win?
   Once we know our goal of total regular season wins (from the previous question), we need to determine how we can make this happen.
   At its most basic level, in baseball, a team wins when it scores more runs than its opponent. Let’s extrapolate this idea to the season-level.
   - Create a new column called runs_diff which is defined as:

     Run Differential = Runs Scored −Runs Allowed. diff

     (In words, the season run differential for a team is the difference between the total runs scored for the season and the total runs allowed (by a team’s opponents) for the season.)

   - Since our data is at the season level, what does is mean if the run differential is greater than zero?

- Using **statsmodels and scikit-learn**, create and fit a simple linear regression model (with an intercept) *using the run differential to predict the wins in a given season.* Print the output summary.
- What is the estimated regression model? Write it out in terms of expected wins, run differential, and estimated coefficients. Round the intercept to the nearest integer and the slope to the nearest tenth.
- Interpret the slope within this context.
- Suppose in one year, a team's total run differential is 50. How many wins should that team expect for the given season? All else equal, if this same team wanted its expected wins to increase by 1 next season, what should it aim to increase its run differential by?
- Based on your estimated model and the typical number of wins needed to make the playoffs, what run differential should Oakland aim for next season to make the playoffs?

4. How does a team get runs?
   - Use the .corr() pandas method in python to compute the Pearson correlation between *batting average (BA), on-base percentage (OBP), slugging percentage (SLG), and runs scored (RS).*

     Historically, batting average (BA) has traditionally been used for team and player evaluations. Would you agree or disagree that this should be the primary metric to use going forward? Why?

     Create a linear regression model using *on-base percentage and slugging percentage* to predict runs scored.
   - Suppose that you expect that Oakland's on-base percentage and slugging percentage in the next season (2002) will be the same as in 2001. How many runs scored do you expect to see next season?
   - Based on the run differential goal previously and your answer to the previous question, how many runs allowed are needed to achieve the desired run differential?