



Department: Computer Science

Specialization: Software Engineering

Course: Fundamental Of Data Science

Week 4 - Lab Session: Data Preprocessing in Python

Exercise 1: Wildlife Strikes

We shall study the dataset **FAA Wildlife Strikes**. This is a dataset that the U.S. Federal Aviation Administration puts together. Long story short, Wildlife strikes occur when an aircraft and animal collide (usually birds). As you can probably imagine, that isn't good for anyone, especially the birds. When birds collide with aircraft, pilots fill out a report and submit it to the FAA. The dataset contains information such as the kind of damage done and which phase of flight it occurred.

Part 1 – Import and Explore Data

1. Import the dataset strikes.csv with the function read_csv of the library pandas
2. Exploration of the data:
 - a. Shape of the data
 - b. Display the four first rows of the DataFrame, the five last rows
 - c. Display the columns and row names
 - d. Get the number of occurrences of birdstrike for each airline operator using the function `value_count()`. Save this number into the variable `operator_counts`. What is the type of this variable?
 - e. Save the Operator name in a variable
 - f. Save the counts in another variable transforming the dataframe `operator_counts` into a numpy array.

Part 2 – Data Cleaning

Missing Values:

1. Check the number of missing values per columns.
2. Calculated the percentages of missing values.
3. Decide:
 - a. Which columns should be dropped ?

- b. Which columns should be filled ?
 - c. Fill missing values
4. We now want to create barplots using the libraries matplotlib and seaborn to visualize occurrences of birdstrike per airline operator. Comment the following lines


```
paired_palette = sns.color_palette("colorblind")
sns.set_palette(paired_palette, 10)
plt.xticks(rotation=45)
plt.xlabel("x-axis Title: Airline operators", fontsize=20)
plt.ylabel("y-axis Title: Number of birdstrikes", fontsize=20)
plt.title("Main title: Birdstrikes per Airline Operator",
          fontsize=20)
barplot = sns.barplot(x=operators[:10], y=counts[:10])
```

Exercise 2: Survival on Titanic dataset

The **RMS Titanic** was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of **15 April 1912**, after it collided with an **Iceberg** during its maiden voyages from *Southampton* to *New York City*. There were an estimated **2,224 passengers** and crew aboard the ship, and more than **1,500 died**, making it one of the deadliest commercial *peacetime maritime disasters* in modern history.

Women and children first? The aim is to understand how survivors of Titanic were selected...

Part 1 - Importation of the data and description of the dataset

1. In the first practical session, download the dataset and set the dataset as Dataframe.
2. Describe the dataset titanic: features, nature of the features, number of observations.
3. Basic statistics: mean of each variable, quartiles
4. Percentage of missing values for each column. Sort by descending values.

Part 2 – Remove Irrelevant Features

Remove :

- PassengerId
- Ticket
- Name

Explain why these columns are removed. (2 – 3 Paragraph).

Part 3 – Missing values

1. Compute missing value percentages.
2. Handling missing values:
 - Age -> median
 - Embarked -> mode

- Cabin -> drop or justify decision

Part 4 – Encoding

Encode:

- Sex
- Embarked

Use:

- Label Encoding or one-hot Encoding

Part 5 – Scaling

1. Scale numerical features:

- Age
- Fare

Use:

- Min-Max Scaling or Standardization

Part 6 – Outlier Detection

1. Detect outlier in Fare using IQR
2. Visualization using boxplot
3. Explain whether you remove them or not ?

Basic graphic analysis

We want to understand what features could contribute to a high survival rate. It would make sense if everything would be correlated with a high survival rate.

1. We focus on the features 'Age' and 'Sex'
 - a. Separate the dataset into men and women
 - b. Display the distribution of the age survivors and non survivors according to the 'Sex'.
2. At first glance is there some link between 'Embarked' and 'Survival'.
3. At first glance is there some link between 'Pclass' and 'Survival'.

End
