1. **Outlier Treatments**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable as Python File (.py) and R file as .r extension files.

Please ensure you update all the details:

**Name: Avishek kumar verma**

**Batch Id: 05012021\_10A.M**

**Topic: Preliminaries for Data Analysis**

**Problem Statement:**

Most of the datasets have extreme values or exceptions in their observations. These values affect the predictions (Accuracy) of the model in one way or the other, removing these values is not a very good option. For these types of scenarios, we have the techniques for treating such values. Explore on various other techniques to treat these values, you can go through this link:

**

**Ans:-**

**Data types: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| crime | Crime | Ratio | Relevant |
| zn | zn | Ratio | Relevant |
| Indus | Indus | Ratio | Relevant |
| nox | nox | Ratio | Relevant |
| rm | rm | Ratio | Relevant |
| age | age | Ratio | Relevant |
| dis | dis | Ratio | Relevant |
| rad | rad | Count | Relevant |
| tax | tax | Ratio | Relevant |

1. **Imputation**

**Problem Statement:**

Majority of the datasets have missing values, that might be because the data collected were not at regular intervals or the breakdown of instruments and so on. It is nearly impossible to build the proper model or in other words, get accurate results. The common techniques are either removing those records completely or substitute those missing values with the logical ones, there are various techniques to treat these types of problems.



**Ans:-**

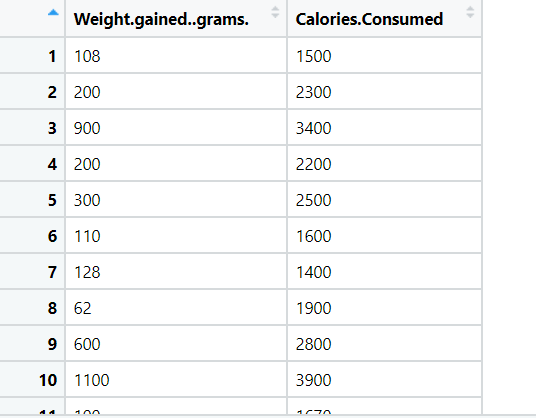
**Data types: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| CASENUM | CASENUM | Ratio | Relevant |
| ATTORNEY | ATTORNEY | Binary | Relevant |
| CLMSEX | CLMSEX | Binary | Relevant |
| CLMINSUR | CLMINSUR | Binary | Relevant |
| SEATBELT | SEATBELT | Binary | Relevant |
| CLMAGE | CLMAGE | Ratio | Relevant |
| LOSS | LOSS | Ratio | Relevant |

1. **TRANSFORMATIONS**

**Problem Statement:**

Everything will revolve around the data in Analytics world. Proper data will help you to make useful predictions which improve your business. Sometimes the usage of original data as it is does not help to have accurate solutions. It is needed to convert the data from one form to another form to have better predictions. Explore on various techniques to transform the data for better model performance. you can go through this link:



**Ans:-**

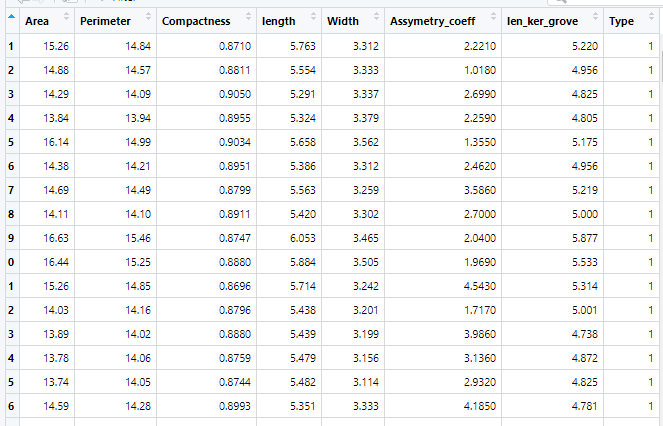
**Data types: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| Weight\_gained\_grams | Weight\_gained\_grams | Ratio | Relevant |
| Calories\_Consumed | Calories\_Consumed | Ratio | Relevant |

1. **STANDARDIZATION & NORMALIZATION**

**Problem Statement:**

Data is one of the most important assets. It is often common that data is stored in distinct systems with different formats and scales. These seemingly small differences in how the data is stored can result in misinterpretations and inconsistencies in your analytics. Inconsistency can make it impossible to deliver reliable information to management for good decision making. We have the preprocessing techniques to make the data uniform. Explore the various techniques to have reliable uniform standard data, you can go through this link:



**Ans:-**

**Data types: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| Area | Area | Ratio | Relevant |
| Perimeter | Perimeter | Ratio | Relevant |
| Compactness | Compactness | Ratio | Relevant |
| length | length | Ratio | Relevant |
| Width | Width | Ratio | Relevant |
| Assymetry\_coeff | Assymetry\_coeff | Ratio | Relevant |
| len\_ker\_grove | len\_ker\_grove | Ratio | Relevant |
| Type | Type | Ratio | Relevant |

1. **Dummy Variables**

**Problem Statement:**

Data is one of the most important assets. It is often common that data is stored in distinct systems with different formats and forms. Non-numeric form of data makes it tricky while developing mathematical equations for prediction models. We have the preprocessing techniques to make the data convert to numeric form. Explore the various techniques to have reliable uniform standard data, you can go through this link:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Animals | Gender | Homly | Types |
| 1 | Cat | Male | Yes | A |
| 2 | Dog | Male | Yes | B |
| 3 | Mouse | Male | Yes | C |
| 4 | Mouse | Male | Yes | C |
| 5 | Dog | Female | Yes | A |
| 6 | Cat | Female | Yes | B |
| 7 | Lion | Female | Yes | D |
| 8 | Goat | Female | Yes | E |
| 9 | Cat | Female | Yes | A |
| 10 | Dog | Male | Yes | B |

**Ans:-**

**Data types: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| Animals | Animals | Nominal | Relevant |
| Gender | Gender | Binary | Relevant |
| Homly | Homly | Binary | Relevant |
| Types | Types | Ordinal | Relevant |

1. **Zero - Variance Features**

**Problem statement:**

Find which columns of the given dataset with zero variance, explore various techniques used to remove the zero variance from the dataset to perform certain analysis.



**Ans:-**

**Data types: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| square.length | square.length | Ratio | Relevant |
| square.breadth | square.breadth | Ratio | Relevant |
| rec.Length | rec.Length | Ratio | Relevant |
| rec.breadth | rec.breadth | Ratio | Relevant |
| colour | colour | Nominal | Relevant |

1. **STRING MANIPULATIONS**

**Problem Statement:**

It is obvious that as part of data analysis we encounter a lot of text data which is a collection of strings which in turn is a sequence of characters. Access the text data and manipulate as per our requirements. you can go through this link for further assistance:

1. Create a string “Grow Gratitude”.

Code for the following tasks:

1. How do you access the letter “G” of “Growth”?
2. How do you find the length of the string?
3. Count how many times “G” is in the string.
4. Create a string “Being aware of a single shortcoming within yourself is far more useful than being aware of a thousand in someone else.”

Code for the following

1. Count the number of characters in the string.
2. Create a string "Idealistic as it may sound, altruism should be the driving force in business, not just competition and a desire for wealth"

Code for the following tasks:

1. get one char of the word
2. get the first three char
3. get the last three char
4. create a string "stay positive and optimistic". Now write a code to split on whitespace.

Write a code to find if:

1. The string starts with “H”
2. The string ends with “d”
3. The string ends with “c”
4. Write a code to print " 🪐 " one hundred and eight times. (only in python)
5. Write a code to print " o " one hundred and eight times. (only in R)
6. Create a string “Grow Gratitude” and write a code to replace “Grow” with “Growth of”
7. A story was printed in a pdf, which isn’t making any sense. i.e.:

“.elgnujehtotniffo deps mehtfohtoB .eerfnoilehttesotseporeht no dewangdnanar eh ,ylkciuQ .elbuortninoilehtdecitondnatsapdeklawesuomeht ,nooS .repmihwotdetratsdnatuotegotgnilggurts saw noilehT .eert a tsniagapumihdeityehT .mehthtiwnoilehtkootdnatserofehtotniemacsretnuhwef a ,yad enO .ogmihteldnaecnedifnocs’esuomeht ta dehgualnoilehT ”.emevasuoy fi yademosuoyotplehtaergfo eb lliw I ,uoyesimorp I“ .eerfmihtesotnoilehtdetseuqeryletarepsedesuomehtnehwesuomehttaeottuoba saw eH .yrgnaetiuqpuekow eh dna ,peels s’noilehtdebrutsidsihT .nufroftsujydobsihnwoddnapugninnurdetratsesuom a nehwelgnujehtnignipeelsecno saw noil A”

You have noticed that the story is printed in a reversed order. Rectify the same and write a code to print the same story in a correct order.

1. **DISCRETIZATION**

**Problem Statement:**

Everything will revolve around the data in Analytics world. Proper data will help you to make useful predictions which improve your business. Sometimes the usage of original data as it is does not help to have accurate solutions. It is needed to convert the data from one form to another form to have better predictions. Explore on various techniques to transform the data for better model performance. you can go through this link:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |

**Ans:-**

**Data types: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| Sepal.Length | Sepal.Length | Ratio | Relevant |
| Sepal.Width | Sepal.Width | Ratio | Relevant |
| Petal.Length | Petal.Length | Ratio | Relevant |
| Petal.Width | Petal.Width | Ratio | Relevant |
| Species | Species | Nominal | Relevant |

1. **Duplication Typecasting**

**Problem statement:**

Q1. For the given dataset perform the type casting (convert the datatypes, ex. float to int)

Q2. Check for the duplicate values, and handle the duplicate values (ex. drop)

Q3. Do the data analysis (EDA)?

Such as histogram, boxplot, scatterplot etc

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
| 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 12/1/2010 8:26 | 2.55 | 17850 | United Kingdom |
| 536365 | 71053 | WHITE METAL LANTERN | 6 | 12/1/2010 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 12/1/2010 8:26 | 2.75 | 17850 | United Kingdom |
| 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 12/1/2010 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 12/1/2010 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 22752 | SET 7 BABUSHKA NESTING BOXES | 2 | 12/1/2010 8:26 | 7.65 | 17850 | United Kingdom |
| 536365 | 21730 | GLASS STAR FROSTED T-LIGHT HOLDER | 6 | 12/1/2010 8:26 | 4.25 | 17850 | United Kingdom |
| 536366 | 22633 | HAND WARMER UNION JACK | 6 | 12/1/2010 8:28 | 1.85 | 17850 | United Kingdom |
| 536366 | 22632 | HAND WARMER RED POLKA DOT | 6 | 12/1/2010 8:28 | 1.85 | 17850 | United Kingdom |

**Ans:-**

**Data types: -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| InvoiceNo | InvoiceNo | Nominal | Not Relevant |
| StockCode | StockCode | Nominal | Relevant |
| Description | Description | Nominal | Relevant |
| Quantity | Quantity | Count | Relevant |
| InvoiceDate | InvoiceDate | Count | Relevant |
| UnitPrice | UnitPrice | Ratio | Relevant |
| CustomerID | CustomerID | Nominal | Not Relevant |
| Country | Country | Nominal | Relevant |